

Valg af Uddannelse

En empirisk analyse af hvilke faktorer,
der påvirker unges valg af uddannelsesniveau

Anne Sofie Bundgaard Kristensen, Kristine Ravn Sørensen, Ninna Grydholt

Økonomi, Aalborg University Business School

Specialeafhandling



Abstract

This project examines which factors affect young people's choice of investing in education. The formal research question to be answered is: Which factors affect young people's choice of level of education? Where affect is defined as the existence of a probable causal connection.

Integrating both neoclassical microeconomic theory and empirical findings the research question is examined. First labour market theory examining individuals' investment decision in education is presented. Limitations of the theory concerning the research question are discussed along with assumptions underlying neoclassical economic theory. To meet the theoretical limitations a broad-spectrum empirical literature study is conducted. In relation to this, it is assumed that a better educational performance increases the level of education.

Based on labour market theory and empirical findings, two groups of hypotheses are formulated. The first group of hypotheses deals with the factors associated with young people's highest completed level of education. The second group of hypotheses looks at whether the same connection exists both when looking at all young people and only young people who have completed higher education. The hypotheses are investigated using a quantitative analysis, which examines the highest level of education in 2019 for young Danish people who finished the 9th grade in 2002. Education is ranked into seven vertical categories and an ordinal variable is created. Due to the practical econometric properties of analyzing an ordinal dependent variable two separate regression models are constructed and examined using an ordered logistic regression method. The robustness of the models is checked using ordinary least squares regression. The first regression model, model 1, examines what factors affect individuals' highest completed level of education. Model 1 examines the choice between completing primary school, gymnasium education, vocational education or higher education. The results from model 1 are used to examine the first group of hypotheses which states that explanatory factors positively correlate with young people's level of education. In model 2 only individuals who have completed higher education are included. In this model, the choice between a short higher education, an academic bachelor's degree, a professional bachelor's degree and a graduate degree or more is examined. Examining the results from both model 1 and model 2, the second group of hypotheses are examined. These hypotheses state that explanatory factors positively correlate with young people's level of education regardless of whether you look at all young people or only young people who have completed higher education. The result of the quantitative analysis is that young people's abilities and their mother's and father's level of education have a significant positive correlation with young people's highest level of education.

After conducting the quantitative analysis the representativity of the data is assessed. Furthermore the operationalization of educational level and underlying assumption of the literature study is discussed. Lastly, it is assessed whether it is probable that a causal connection exists.

As there are both significant correlations in the quantitative analysis and these correlations are supported by neoclassical labour market theory, the project concludes that it is probable that there is a positive causal connection between abilities and young people's choice of education level and parents' education level and young people's choice of education level.

**Titel:**

Valg af Uddannelse:

En empirisk analyse af hvilke faktorer,
der påvirker unges valg af uddannelsesniveau

Tema:

Speciale

Periode:

Forårssemester 2022

Gruppe:

4

Forfattere:

Anne Sofie Bundgaard Kristensen

Studienummer: 20174534

asbk17@student.aau.dk

Kristine Ravn Sørensen

Studienummer: 20174317

krsa17@student.aau.dk

Ninna Grydholt

Studienummer: 20174320

ngrydh17@student.aau.dk

Vejleder:

Ina Drejer

id@business.aau.dk

Normalsider: 88

Afleveringsdato: 3. juni 2022

Aalborg University Business School

Aalborg Universitet

<https://www.business.aau.dk/>

Indholdsfortegnelse

Abstract	ii
Figurer	vii
Tabeller	vii
Kapitel 1 Introduktion	1
1.1 Problemformulering	2
1.1.1 Problemfelt	2
1.2 Projektstruktur	2
Kapitel 2 Metode	4
2.1 Videnskabsteoretiske Overvejelser	4
2.2 Afgrænsning	4
2.3 Data	7
2.4 Det Danske Uddannelsessystem	8
2.4.1 Opbygning af Uddannelsessystemet	8
2.4.2 Rangering af Uddannelsesniveauer	8
Kapitel 3 Investeringsbeslutningen i Uddannelse	10
3.1 Investeringsbeslutninger	10
3.2 Schooling Modellen	10
3.2.1 Wage-Schooling Locus	11
3.3 Begrænsninger i Teorien	12
3.4 Opsummering	14
Kapitel 4 Empirisk Litteraturstudie	15
4.1 Metodiske Afgrænsninger	15
4.2 Faktorer	16
4.2.1 Socioøkonomiske Faktorer	16
4.2.2 Peer-Gruppe	18
4.2.3 Indkomst	18
4.2.4 Evner	19
4.2.5 Grundskole	19
4.2.6 Bopæl	20
4.2.7 Bløde Faktorer	20
4.2.8 Køn	21
4.2.9 Etnicitet	21
4.2.10 Alder	21
4.3 Opsummering	21
Kapitel 5 Hypoteser	22
5.1 Opstilling af Hypoteser	22
5.2 Deskriptiv Statistik	24
5.3 Opsummering	28
Kapitel 6 Økonometrisk Metode	29
6.1 Paneldata og Tværsnitsdata	29

6.2	Modelovervejelser	29
6.3	Opsummering	31
Kapitel 7	Modelopbygning	32
7.1	Opstilling af Regressionsmodeller	32
7.2	Afhængig Variabel	33
7.3	Uafhængige Variable	34
7.3.1	Forklarende Variable	35
7.3.2	Kontrolvariable	38
7.3.3	Udeladte variable	38
7.4	Opsummering	38
Kapitel 8	Kvantitativ Analyse	40
8.1	Beslutningsregel for Hypoteser	40
8.2	Modellernes Resultater	40
8.2.1	Undersøgelse af Hypotesegruppe 1	43
8.2.2	Undersøgelse af Hypotesegruppe 2	44
8.3	Robusthedstjek ved OLS-estimering	45
8.4	Opsummering af Resultater for Forklarende Variable	46
8.5	Resultater for Kontrolvariable	46
8.6	Vurdering af Regressionsmodellerne	48
8.6.1	Test af Komprimerede Modeller	50
8.7	Opsummering	50
Kapitel 9	Unges Uddannelsesvalg	51
9.1	Repræsentativitet i Datagrundlaget	51
9.2	Rangering af Uddannelsesniveauer	53
9.3	Bedre Uddannelsespræstation øger Uddannelsesniveauet	53
9.4	Sociologiske Aspekter	54
9.5	Kausale Sammenhænge	54
9.6	Opsummering	55
Kapitel 10	Konklusion	56
Kapitel 11	Perspektivering	57
Litteratur		58
Bilag A	Dataoversigt	65
A.1	Oversigt	65
A.2	Grundskolekarakter	66
A.3	Alder & Køn	66
A.4	Etnicitet	66
A.5	Kommunetype	67
A.6	Søskende	67
A.7	Forældre	67
A.8	Indkomst	68
A.9	Uddannelsesniveau	69
A.10	Endeligt Datasæt	69
Bilag B	Tabeller Deskriptiv Statistik	73
Bilag C	Variable	75
C.1	Kontrolvariable	75
C.2	Udeladte Variable	82

Bilag D Teori om Ordered Logistisk Regression	85
Bilag E T-tests	87
E.1 Forklarende Variable - Model 1	88
E.2 Kontrolvariable - Model 1	90
E.3 Forklarende Variable - Model 2	96
E.4 Kontrolvariable - Model 2	98
Bilag F Robusthedsundersøgelse	104
Bilag G Korrelationsmatricer	106
G.1 Model 1	107
G.2 Model 2	108
Bilag H Odds Ratios og Konfidensintervaller	109
H.1 Odds Ratios og Konfidensintervaller - Model 1	109
H.2 Odds Ratios og Konfidensintervaller - Model 2	110
Bilag I Antagelser for Regressionsmodeller	111
I.1 Ordered Logistisk Regression	111
I.1.1 Undersøgelser af Antagelserne for Ordered Logistisk Regression	112
I.1.2 AIC	116
I.2 Ordinary Least Squares Regression	117
I.2.1 Undersøgelser af Antagelserne for OLS Regression	117
Bilag J OLS modeller	122
J.1 Model 1	122
J.2 Model 2	123
Bilag K Modeller med Indkomst	124
K.1 Model 1	124
K.2 Model 2	127
K.3 T-tests af Indkomst	130
Bilag L Komprimerede Modeller	131
L.1 Model 1	131
L.2 Model 2	132

Figurer

2.1	Oversigt over Rangering af Uddannelsesniveauer	9
3.1	Wage-Schooling Locus	12
5.1	Højeste Fuldførte Uddannelsesniveau	25
5.2	Højeste Fuldførte Uddannelsesniveau Opdelt efter Køn	26
5.3	Højeste Fuldførte Uddannelsesniveau Opdelt efter Kommunetype	27
5.4	Højeste Fuldførte Uddannelsesniveau Opdelt efter Etnicitet	28
B.1	Boxplot over Indkomst	74
G.1	Korrelationstest af Uafhængige Variable - Model 1	107
G.2	Korrelationstest af Uafhængige Variable - Model 2	108
I.1	Grafisk Regressionsdiagnostik - OLS Model 1	118
I.2	Grafisk Regressionsdiagnostik - OLS Model 2	118
I.3	Histogram over Fejlld - OLS Model 1	120
I.4	Histogram over Fejlld - OLS Model 2	121

Tabeller

7.2	Karakteristika for Variablen Grundskolekarakterer - Model 1 og Model 2	35
7.3	Karakteristika for Variablen Mor Indkomst*, Opgjort i Kr. (2002-Priser) - Model 1 og Model 2	36
7.4	Oversigt over Procentvis Fordeling af Mor Uddannelsesniveau - Model 1 og Model 2	37
7.5	Karakteristika for Variablen Far Indkomst*, Opgjort i Kr. (2002-Priser) - Model 1 og Model 2	37
7.6	Oversigt over Procentvis Fordeling af Far Uddannelsesniveau - Model 1 og Model 2	38
8.1	Ordered Logistisk Regression - Model 1	41
8.2	Ordered Logistisk Regression - Model 2	42
9.1	Oversigt over Uddannelsesniveau i 2019 for 2002-Kohorten og Aldersgrupperne 25-29 År, 30-34 År og 35-39 År	52
A.1	Oversigt over Anvendte Datasæt fra Danmarks Statistiks Forskerservice	65
A.2	Antal Observationer Opdelt efter Uddannelsesniveau	69
A.3	Antal Observationer Opdelt efter Alder	69
A.4	Antal Observationer Opdelt efter Køn	70
A.5	Antal Observationer Opdelt efter Etnicitet	70
A.6	Antal Observationer Opdelt efter Kommunetype	70
A.7	Antal Observationer Opdelt efter Antal Søskende	70
A.8	Antal Observationer Opdelt efter Mor Alder	70
A.9	Antal Observationer Opdelt efter Mor Socialklasse	71
A.10	Antal Observationer Opdelt efter Mor Uddannelsesniveau	71
A.11	Antal Observationer Opdelt efter Mor Familietype	71

A.12 Antal Observationer Opdelt efter Far Alder	71
A.13 Antal Observationer Opdelt efter Far Socialklasse	72
A.14 Antal Observationer Opdelt efter Far Uddannelsesniveau	72
A.15 Antal Observationer Opdelt efter Far Familietype	72
B.1 Højeste fuldførte Uddannelsesniveau	73
B.2 Højeste Fuldførte Uddannelsesniveau Opdelt efter Køn	73
B.3 Højeste Fuldførte Uddannelsesniveau Opdelt efter Kommunitype	73
B.4 Højeste Fuldførte Uddannelsesniveau Opdelt efter Etnicitet	74
B.5 Karakteristika for Grundskolekarakterer, Standpunkt, Afgangsprøve og Gennemsnit	74
C.1 Oversigt over Procentvis Fordeling af Alder - Model 1 og Model 2	76
C.2 Oversigt over Procentvis Fordeling af Køn - Model 1 og Model 2	76
C.3 Oversigt over Procentvis Fordeling af Etnicitet - Model 1 og Model 2	76
C.4 Oversigt over Procentvis Fordeling mellem Kommunityper - Model 1 og Model 2	77
C.5 Karakteristika for Variablen Søkende - Model 1 og Model 2	78
C.6 Karakteristika for Variablen Mor Alder, Opgjort i År - Model 1 og Model 2	79
C.7 Oversigt over Procentvis Fordeling mellem Mor Familietype - Model 1 og Model 2	79
C.8 Oversigt over Procentvis Fordeling mellem Mor Socialklasse - Model 1 og Model 2	80
C.9 Karakteristika for Variablen Fars Alder, Opgjort i År - for Model 1 og Model 2	81
C.10 Oversigt over Procentvis Fordeling mellem Far Familietype - Model 1 og Model 2	81
C.11 Oversigt over Procentvis Fordeling mellem Far Socialklasse - Model 1 og Model 2	82
E.1 T-test: Uddannelsesniveau & Karaktergennemsnit - Model 1	88
E.2 T-test: Uddannelsesniveau & Mor Indkomst - Model 1	88
E.3 T-test: Uddannelsesniveau & Mor Uddannelsesniveau - Model 1	89
E.4 T-test: Uddannelsesniveau & Far Indkomst - Model 1	89
E.5 T-test: Uddannelsesniveau & Far Uddannelsesniveau - Model 1	90
E.6 T-test: Uddannelsesniveau & Alder - Model 1	90
E.7 T-test: Uddannelsesniveau & Køn - Model 1	91
E.8 T-test: Uddannelsesniveau & Etnicitet - Model 1	91
E.9 T-test: Uddannelsesniveau & Kommunitype - Model 1	92
E.10 T-test: Uddannelsesniveau & Søkende - Model 1	92
E.11 T-test: Uddannelsesniveau & Mor Alder - Model 1	93
E.12 T-test: Uddannelsesniveau & Mor Familietype - Model 1	93
E.13 T-test: Uddannelsesniveau & Mor Socialklasse - Model 1	94
E.14 T-test: Uddannelsesniveau & Far Alder - Model 1	94
E.15 T-test: Uddannelsesniveau & Far Familietype - Model 1	95
E.16 T-test: Uddannelsesniveau & Far socialklasse - Model 1	95
E.17 T-test: Uddannelsesniveau & Karaktergennemsnit - Model 2	96
E.18 T-test: Uddannelsesniveau & Mor Indkomst - Model 2	96
E.19 T-test: Uddannelsesniveau & Mor Uddannelsesniveau - Model 2	97
E.20 T-test: Uddannelsesniveau & Far Indkomst - Model 2	97
E.21 T-test: Uddannelsesniveau & Far Uddannelsesniveau - Model 2	98
E.22 T-test: Uddannelsesniveau & Alder - Model 2	98
E.23 T-test: Uddannelsesniveau & Køn - Model 2	99
E.24 T-test: Uddannelsesniveau & Etnicitet - Model 2	99
E.25 T-test: Uddannelsesniveau & Kommunitype - Model 2	100
E.26 T-test: Uddannelsesniveau & Søkende - Model 2	100
E.27 T-test: Uddannelsesniveau & Mor Alder - Model 2	101
E.28 T-test: Uddannelsesniveau & Mor Familietype - Model 2	101
E.29 T-test: Uddannelsesniveau & Mor Socialklasse - Model 2	102
E.30 T-test: Uddannelsesniveau & Far Alder - Model 2	102
E.31 T-test: Uddannelsesniveau & Far Familietype - Model 2	103

E.32	T-test: Uddannelsesniveau & Far Socialklasse - Model 2	103
H.1	Odds Ratios og Konfidensintervaller - Model 1	109
H.2	Odds Ratios og Konfidensintervaller - Model 2	110
I.1	Brant-test - Model 1	113
I.2	Test for Proportional Odds Assumption - Model 1	114
I.3	Brant-test - Model 2	115
I.4	Test for Proportional Odds Assumption - Model 2	116
I.5	Variance Inflation Factors - OLS Model 1	119
I.6	Variance Inflation Factors - OLS Model 2	119
I.7	Middelværdi af Fejllad - OLS Model 1	119
I.8	Middelværdi af Fejllad - OLS Model 2	119
I.9	Breusch-Pagan Test for Homoskedasticitet - OLS Model 1	120
I.10	Breusch-Pagan Test for Homoskedasticitet - OLS Model 2	120
J.1	OLS Model med Robuste Standardfejl - Model 1	122
J.2	OLS Model med Robuste Standardfejl - Model 2	123
K.1	Ordered Logistisk Regression med Indkomst - Model 1	124
K.2	Odds Ratios og Konfidensintervaller - Model 1 med Indkomst	125
K.3	OLS Model med Indkomst med Robuste Standardfejl - Model 1	126
K.4	Ordered Logistisk Regression med Indkomst - Model 2	127
K.5	Odds Ratios og Konfidensintervaller - Model 2 med Indkomst	128
K.6	OLS Model med Indkomst med Robuste Standardfejl - Model 2	129
K.7	T-test: Uddannelsesniveau & Indkomst - Model 1	130
K.8	T-test: Uddannelsesniveau & Indkomst - Model 2	130
L.1	Ordered Logistisk Regression - Model 1 Komprimeret	131
L.2	Odds Ratios og Konfidensintervaller - Model 1 Komprimeret	131
L.3	OLS Model Komprimeret med Robuste Standardfejl - Model 1	132
L.4	Ordered Logistisk Regression - Model 2 Komprimeret	132
L.5	Odds Ratios og Konfidensintervaller - Model 2 Komprimeret	133
L.6	OLS Model Komprimeret med Robuste Standardfejl - Model 2	133

I en tale i oktober 2021 udtalte den nuværende uddannelses- og forskningsminister Jesper Petersen at *"uddannelse og forskning er vigtige politiske emner"* (Petersen, 2021). Den politiske vigtighed af uddannelse afspejles i, at der eksisterer mange politiske målsætninger og tiltag både indenfor uddannelses- og undervisningsområdet. I 2012 havde den daværende regering en målsætning om, at 60% af en ungdomsårgang skulle gennemføre en videregående uddannelse i 2020. Derudover skulle mindst 25% af en ungdomsårgang gennemføre en lang videregående uddannelse. Dengang var sigtet, at de videregående uddannelser i højere grad skulle fremme væksten og produktiviteten i samfundet (Ministeriet for Forskning, Innovation og Videregående Uddannelser, Økonomi- og Indenrigsministeriet, Finansministeriet, Erhvervs- og Vækstministeriet, Beskæftigelsesministeriet, Ministeriet for Børn og Undervisning, 2012). Regeringen satsede, ifølge den daværende uddannelsesminister Morten Østergaard, på at *"have den bedst uddannede generation nogensinde"* (Børne og Undervisningsministeriet, 2013). For at opnå målsætningerne, er der blevet foretaget mange reformer på området. På universitetsområdet alene har der de sidste 20 år været 29 reformer (PAULSEN, 2022). I dag eksisterer der blandt andet mål for, at flere skal fuldføre en erhvervsuddannelse og at mindst 90% af 25-årige i 2030 skal have gennemført en ungdomsuddannelse (Børne- og Undervisningsministeriet, 2022*b,a*). Set i lyset af en udtalelse fra Jesper Petersen i talen i 2021, synes fokuset at være skiftet de sidste 10 år og vi skal, ifølge Jesper Petersen, *"uddanne til samfundets behov"* (Petersen, 2021).

Uanset om regeringens mål er, hvor mange der skal gennemføre ungdomsuddannelser, erhvervsuddannelser eller videregående uddannelser, handler de uddannelsespolitiske mål ultimativt om unges valg af uddannelse. I Danmark siges der at være frit valg af uddannelse og at alle, der har kvalifikationerne og motivationen, kan studere og få den uddannelse de ønsker. Netop begrebet frit uddannelsesvalg refererer til, at man i Danmark har fri, lige og gratis adgang til uddannelse (Petersen, 2021). Det frie valg af uddannelse er dog ikke ensbetydende med, at der ikke eksisterer faktorer, der driver unges uddannelsesvalg. I en analyse fra 2018 finder Nielsen (2018), at unges uddannelsesvalg påvirkes af deres nære relationer gennem eksempelvis familiestruktur og forældres indkomstgrundlag. Dette betegnes ofte som en persons socioøkonomiske baggrund. Hvis unges uddannelsesvalg i høj grad drives af deres socioøkonomiske baggrund, er det urealistisk at forvente, at målene kan opfyldes ved brug af uddannelsespolitik alene. I dette tilfælde bør blikket måske i højere grad rettes mod socialpolitik.

Danmark investerer massivt i uddannelse (Børne og Undervisningsministeriet, 2013; Petersen, 2021). Investeringerne i uddannelse er med til at skabe vækst og velstand i samfundet (Gottfries, 2013). Uddannelse er dermed både et politisk og samfundsøkonomisk vigtigt emne. Uddannelse er også en investering for den enkelte dansker (Arbejderbevægelsens Erhvervsråd, 2009). Indenfor den mikroøkonomiske disciplin betragtes individers valg af uddannelse som en investeringsbeslutning. Det betyder, at når unge træffer et valg af uddannelse, træffer de dermed et valg om, hvor meget de vil investere i uddannelse. I dette perspektiv er det dermed teoretiske fænomener som diskonteringsrente og evner, der driver beslutningen (Borjas, 2016).

Uanset hvilket perspektiv uddannelse og unges uddannelsesvalg betragtes i, er det et vigtigt emne. For at opnå de uddannelsespolitiske mål, omend disse handler om at få flere til at tage en videregående uddannelse eller uddanne til samfundets behov, er det nødvendigt at forstå unges uddannelsesvalg. Er uddannelsesvalget drevet af socioøkonomiske forhold? Er det teoretiske fænomener som individers diskonteringsrente, der afgør, hvorvidt de investerer i uddannelse eller er det noget helt tredje?

Svarene på disse spørgsmål har en betydning for de samfundsøkonomiske investeringer. Hvad der driver uddannelsesvalg, er en problemstilling, der er undersøgt før, eksempelvis af Nielsen (2018), som undersøger valget mellem erhvervsuddannelse og gymnasial uddannelse. Der synes dog empirisk at være en tendens til, at uddannelsesvalg simplificeres til et valg mellem to uddannelser og dermed ikke afspejler kompleksiteten af det valg unge står overfor.

På grund af vigtigheden af unges uddannelsesvalg, både for uddannelsespolitiske mål og samfundsøkonomiske forhold, findes det interessant at undersøge dette dybere. Ud fra et uddannelsespolitisk perspektiv findes det særligt interessant at undersøge valget af uddannelsesniveau, da det ofte er dette, der formuleres specifikke mål for. Denne interesse omfatter alle niveauer af uddannelse. Ud fra et økonomisk teoretisk aspekt, empiriske undersøgelser samt kvantitative bidrag, beskæftiger afhandlingen sig derfor med at undersøge, hvilke faktorer der påvirker unges uddannelsesvalg.

1.1 Problemformulering

Ovenstående overvejelser har ført til udarbejdelsen af følgende problemformulering:

Hvilke faktorer påvirker unges valg af uddannelsesniveau?

Påvirker defineres her, som at der kan sandsynliggøres en kausal sammenhæng.

1.1.1 Problemfelt

I nærværende afsnit uddybes problemformuleringen ved en eksplicit definition af begreberne, hvilket har til formål at indsnævre projektets problemfelt. Efterfølgende redegøres der for den proces, der ligger til grund for besvarelsen af problemformuleringen.

Unge defineres som danske unge. I denne sammenhæng defineres begrebet *uddannelsesniveau* som den højeste fuldførte uddannelse. Det betyder, at det i denne forbindelse kun er fuldført uddannelse, der betragtes og dermed ikke forventet fuldført uddannelse. Endvidere betyder det, at det er irrelevant på hvilket uddannelsesniveau, den senest fuldførte uddannelse er, hvis ikke denne er den højeste. Disse afgrænsninger forklares nærmere i kapitel 2.2. Betragtes formuleringen *valg af uddannelsesniveau*, betyder dette, at uddannelsesbeslutningen i projektet udelukkende undersøges i et vertikalt perspektiv. I relation til begrebet *valg*, er det væsentligt at definere, at valget ikke er fuldstændigt frit, men er begrænset af blandt andet gennemførelse af uddannelse på et lavere niveau.

Problemformuleringen besvares ved anvendelse af en kvantitativ metode. I besvarelsen lægges der vægt på neoklassisk økonomisk teori og empiriske resultater. Det todelte fokus kommer til udtryk ved, at relevant neoklassisk arbejdsmarkedsteori præsenteres, hvorefter der foretages et litteraturstudie, der undersøger empiriske fund på området. På baggrund af resultaterne fra litteraturstudiet og den økonomiske teori, opstilles to grupper af hypoteser, der omhandler hvorvidt særligt interessante faktorer har en sammenhæng med unges uddannelsesniveau. Hypoteserne undersøges empirisk i en kvantitativ analyse, hvor der gøres brug af en kohorte. Kohorten består af unge i alderen 14-18 år, der færdiggjorde grundskolens 9. klasse i 2002. Mere information herom fremgår af kapitel 2.2. Resultaterne fra den kvantitative analyse vurderes i henhold til teorien samt empiriske fund. Det diskuteres endvidere, hvorvidt datagrundlaget og dermed resultaterne fra den kvantitative analyse, er repræsentative og generaliserbare. Afslutningsvis vurderes det, hvorvidt der kan sandsynliggøres kausale sammenhænge.

1.2 Projektstruktur

Med udgangspunkt i det introducerende afsnit, problemformuleringen samt problemfeltet er følgende projektstruktur udarbejdet. Projektstrukturen fremlægges med det formål at give læseren en kort og præcis oversigt over indholdet i projektets kapitler.

Kapitel 2

I kapitel 2 fremlægges de metodiske overvejelser samt beslutninger der er foretaget, herunder centrale afgrænsninger. Derudover rangeres de danske uddannelsesniveauer. Overvejelserne, der fremgår af kapitlet, er nødvendige for at kunne udarbejde en fyldestgørende besvarelse af problemformuleringen.

Kapitel 3

I kapitel 3 redegøres der for neoklassisk teori vedrørende investeringsbeslutningen i uddannelse. Eksplicit redegøres der for *schooling* modellen. Til sidst diskuteres teoriramens begrænsninger.

Kapitel 4

I kapitel 4 foretages der et empirisk litteraturstudie, hvilket bidrager til at opstille to grupper af hypoteser, der præsenteres i kapitel 5. Litteraturstudiet bidrager desuden til udarbejdelsen af den kvantitative analyse.

Kapitel 5

På baggrund af teorien og de empiriske resultater fra kapitel 3 og 4 opstilles der i kapitel 5 to grupper af hypoteser. Formålet hermed er at udarbejde konkrete arbejdsspørgsmål, der kan undersøges ved brug af empirisk data. I relation hertil beskrives dataet, der anvendes i den kvantitative analyse, og der foretages deskriptiv statistik af specifikke faktorer.

Kapitel 6

I forlængelse af kapitel 5 vurderes det i kapitel 6, hvordan man på bedst mulig vis kan undersøge hypoteserne empirisk. Denne vurdering foretages på baggrund af en præsentation og gennemgang af fordele og ulemper ved forskellige typer af regressionsmodeller.

Kapitel 7

I kapitel 7 præsenteres regressionsmodellerne, der er udarbejdet på baggrund af de foregående kapitler. Regressionsligningen præsenteres og der redegøres for de forklarende variable, hvilket inkluderer en beskrivelse af operationaliseringen af variablene samt forventningerne hertil baseret på teorien og de empiriske fund.

Kapitel 8

Kapitel 8 indeholder den omtalte kvantitative analyse. Estimeringen af regressionsmodellerne opstillet i kapitel 7 foretages på baggrund af de økonometriske vurderinger, der fremgår af kapitel 6. Resultatet af analysen fortolkes med udgangspunkt i forventninger baseret på teorien og litteraturstudiet.

Kapitel 9

I kapitel 9 diskuteres repræsentativiteten af datagrundlaget, der ligger til grund for den kvantitative analyse. Derudover diskuteres valg og antagelser, der er taget i forbindelse med besvarelsen af problemstillingen. Det diskuteres endvidere, hvorvidt det er muligt at sandsynliggøre kausale forhold.

Kapitel 10

I kapitel 10 konkluderes det, hvilke faktorer der påvirker unges valg af uddannelsesniveau. Dette gøres ved at inddrage informationer, overvejelser og resultater fra de foregående kapitler.

Kapitel 11

I kapitel 11 reflekteres der over, hvad resultaterne fra konklusionen i kapitel 10 kan anvendes til. Derudover overvejes det, hvordan et videre arbejde med problemstillingen kan gribes an.

Metode 2

I kapitlet beskrives de metodiske overvejelser og afgrænsninger, der er foretaget, samt dataet der benyttes til deskriptiv statistik og i den kvantitative analyse. Til sidst i kapitlet beskrives opbygningen af det danske uddannelsesystem og der foretages en rangering af uddannelsesniveauer.

2.1 Videnskabsteoretiske Overvejelser

Besvarelsen af problemstillingen er inspireret af flere videnskabsteoretiske tilgange. Det videnskabsteoretiske afsæt er dermed ikke entydigt præget af ét epistemologisk videnssyn. I det følgende redegøres der for de videnskabsteoretiske overvejelser og hvilken betydning disse har for besvarelsen af problemformuleringen.

Der gøres i projektet overvejende brug af en deduktiv tilgang, hvilken ofte benyttes indenfor rationalisme. Ved brug af en udelukkende deduktiv tilgang, foretages en undersøgelse med udgangspunkt i teori (Sonne-Ragans, 2012). I projektet gøres der dog også brug af empiriske fund, hvilket medfører at tilgangen ikke alene er deduktiv. Dermed foretages undersøgelsen både med udgangspunkt i teori og empiri, hvilket blander tilgangene fra rationalisme og empirisme.

Undersøgelsen af hypoteser håndteres forskelligt i forskellige videnskabsteoretiske tilgange. I nærværende tilfælde undersøges hypoteserne i overvejende grad ud fra en positivistisk tilgang, hvilket betyder, at det undersøges, hvorvidt de kan understøttes af data. Undersøgelsen af hypoteserne bidrager dog ikke alene til besvarelsen af problemstillingen, hvorfor der i projektet generelt også findes elementer af kritisk realisme. Det betyder dermed, at der i projektet ikke sandsynliggøres en kausal sammenhæng når der eksisterer empirisk belæg og at hypoteserne er understøttet, uden at der også eksisterer en teoretisk sammenhæng (Sonne-Ragans, 2012).

Inspiration fra positivisme afspejles også i at der undersøges *sandsynliggørelse* af kausale sammenhænge. Indenfor den positivistiske tilgang betegnes evidens som noget der er indlysende tydeligt uden at være bevist (Ahrenkiel, n.d.). I forhold til indeværende problemstilling betyder det, at kausale sammenhænge kan sandsynliggøres, hvis og kun hvis de bakkes op af det økonomisk-teoretiske grundlag og samtidig understøttes af en eller flere hypoteser.

Sammenfattende besvares problemformuleringen dermed ikke med udgangspunkt i én videnskabsteoretisk tilgang, men der findes inspiration i flere videnskabsteoretiske grene.

2.2 Afgrænsning

I dette afsnit redegøres der for de til- og fravalg, der er foretaget i forbindelse med besvarelsen af problemformuleringen.

Kohorten

Problemstillingen undersøges i en kvantitativ analyse ved at følge en udvalgt kohorte. Ved at designe et studie, hvor man følger en kohorte, er det muligt at undersøge sammenhængen mellem adskillige uafhængige variable og en række udfald i den afhængige variabel. Dette design er valgt, da den afhængige variabel i den kvantitative analyse er forskellige uddannelsesniveauer. Uddannelsesystemets opbygning og

hvordan niveauerne er opgjort beskrives længere nede i dette kapitel. Der er valgt et retrospektivt design, hvor kohorten, der fuldførte grundskolens¹ 9. klasse i 2002 følges. Et retrospektivt design er nødvendigt for at kunne følge individers uddannelsesbeslutninger. En ulempe ved retrospektivt design er, at det afhænger af, hvad der er målt og registreret tidligere og dette ofte er registreret til andre formål end det, der undersøges (Euser et al., 2009). For at følge personerne over den længst mulige periode tages der udgangspunkt i kohorten fra 2002, da det er den tidligste kohorte, hvor der er, i det tilgængelig data, er registreret karakterer fra grundskolen. Karakterer vurderes, på baggrund af teoretiske argumenter og empiriske undersøgelser, at være en betydelig faktor for analysen. Som nævnt tages der udgangspunkt i det tidligste data, der har været til rådighed for at kunne følge personerne i den længst mulige periode. Af samme årsag følges personerne indtil 2019, hvor det seneste data for alle individer er tilgængeligt. Der tages kun udgangspunkt i én kohorte, da yderligere modellering ligger udenfor projektets tidsramme. Ved kun at tage udgangspunkt i én kohorte er analysen i risiko for at finde korrelationer, der er begrænset til denne ene kohorte. Dette diskuteres i kapitel 9. Kohorten afgrænses til individer, der var mellem 14-18 år, da de afsluttede 9. klasse. Derved frasortes vokselever og enkeltfagskursister, samt unge, der har gennemført grundskolen på under normeret tid eller startede tidligere end minimumsalderen på 4 år og 10 måneder (Jensen og Nielsen, 2010; Skole og forældre, 2020). Denne afgrænsning er foretaget, da personer, der er yngre eller ældre end den valgte aldersgruppe, kan forårsage bias i dataet, da deres forudsætninger antages, at være anderledes end den valgte aldersgruppe.

Efter afgrænsningen er individerne i denne kohorte mellem 31-35 år gamle i 2019. Gennemsnitsalderen for gennemførelsen af en erhvervsuddannelse er 28 år, mens den er 27 år for en nyuddannet bachelor og lidt over 28 år for en nyuddannet kandidat (Produktivitetskommissionen, 2014; Uddannelses- og Forskningsministeriet, 2013). På dette tidspunkt forventes størstedelen af kohorten at have færdiggjort deres højeste fuldførte uddannelse. På baggrund af dette antages det i projektet, at personerne i kohorten har fuldført deres højeste uddannelsesniveau. Der er 7,3% af personer, der fuldførte grundskoles 9. klasse i 2002, der stadigvæk er under uddannelse i 2019. Disse individer inddrages ikke i populationen, da de kan forårsage støj, idet de sandsynligvis ikke har fuldført deres højeste uddannelsesniveau.

Der tages udgangspunkt i individer i 2002, der i løbet af året har fuldført grundskolens 9. klasse, da de her står overfor deres første uddannelsesbeslutning. Ved at tage udgangspunkt i kohorten fra 2002 er der dermed ingen, som ikke mindst har grundskolen, som højeste fuldførte uddannelse. I skoleåret 2001-2002 var der 4,4% af dem, der påbegyndte 9. klasse, som ikke fuldførte i samme skoleår. (Danmarks Statistik, n.d.w; Danmarks Statistik, n.d.h)

I det videre arbejde menes der med "individer der har færdiggjort grundskolen", personer der har færdiggjort grundskolens 9. klasse i 2002.

Vertikalt Uddannelsesniveau

Der afgrænses i det videre arbejde fra at undersøge, hvilke faktorer der påvirker valget af uddannelse på et horisontalt niveau. Der undersøges dermed kun uddannelse i et vertikalt perspektiv. Dette betyder, at det i projektet undersøges, hvilke faktorer der påvirker længden af uddannelse og ikke hvilke faktorer, der gør at man vælger eksempelvis én bestemt kort videregående uddannelse fremfor en anden.

Det undersøges som nævnt, om der eksisterer en sammenhæng mellem faktorerne og individernes højeste fuldførte uddannelsesniveau (*educational attainment*) og ikke hvilket uddannelsesniveau, individerne forventer at fuldføre (*educational expectations*). Da fokuset er på det højeste fuldførte uddannelsesniveau, er det her ubetydeligt, hvad det seneste opnåede uddannelsesniveau er, hvis det ikke er det højeste niveau individerne har fuldført en uddannelse på. Fokuset er på fuldført uddannelse, da dette sikrer at et individs uddannelsesniveau er opnået. Et fokus på forventet fuldført uddannelse, ville medføre en usikkerhed om hvorvidt uddannelsen gennemføres.

¹Grundskolen dækker over efterskoler, folkeskoler, friskoler og private grundskoler, kommunalene ungdomsskoler og ungdomskostskoler, specialskoler for børn, dagbehandlingstilbud og behandlingshjem samt andre skoler (Danmarks Statistik, n.d.i).

Formel Uddannelse

I besvarelsen af problemformuleringen afgrænses der fra at inddrage ikke-formel uddannelse. Ikke-formel uddannelse er for eksempel kurser, seminarer, workshops og sidemandsoplæring. Formel uddannelse giver formelt anerkendte kompetencer, hvilket betyder, at de anerkendes ved en offentlig godkendt eksamen (Danmarks Statistik, n.d.y). Denne afgrænsning er foretaget, da ikke-formel uddannelse ikke er registreret systematisk og ikke er inddelt efter en skala som formel uddannelse er. Disse karakteristika medfører, at det ikke er muligt at inddele ikke-formel uddannelse i niveauer. Foruden dette investeres der oftest først i ikke-formel uddannelse senere i livet (Danmarks Statistik, n.d.p). På grund af begrænsninger i dataet observeres individerne i populationen kun til de er i starten af 30'erne. Denne begrænsning medfører, at ikke-formel uddannelse taget senere i livet ikke kan inddrages. Ikke-formel uddannelse tages ofte i forbindelse med arbejde og er typisk selvfinansieret eller finansieret af arbejdspladsen. Derved afhænger mængden af ikke-formel uddannelse af tilknytningen til arbejdsmarkedet af den enkelte arbejdsplads og af muligheder for finansiering (Danmarks Statistik, n.d.y). Derudover stiger deltagelse i efteruddannelse efter personers uddannelsesniveau (DØRS, 2003). Det er dermed ikke de samme forhold, som påvirker investeringsbeslutningen i ikke-formel og formel uddannelse. Af disse årsager inddrages kun formel uddannelse.

Teori

Der tages udgangspunkt i neoklassisk arbejdsmarkedsteori. Hvilket i projektet betyder, at valget af uddannelsesniveau anses som en investeringsbeslutning. Ud over neoklassisk arbejdsmarkedsteori, inddrages der resultater fra empiriske studier, der er baseret på teori fra sociologi og adfærdsøkonomi. Sociologiske og adfærdsøkonomiske teorier anvendes ikke til at sandsynliggøre eksistensen af kausale sammenhænge i det videre arbejde. Dermed sandsynliggøres eksistensen af kausale sammenhænge kun ud fra neoklassisk arbejdsmarkedsteori.

Empiri

Som beskrevet i problemfeltet inddrages der i den kvantitative analyse resultater fra empiriske studier, hvorfor litteraturstudiet undersøger en bred vifte af studietyper. Kapitel 4 beskriver både empiriske studier med observerede og selvrapporterede variable samt teoretiske undersøgelser.

På trods af, at det fremgår ovenfor, at besvarelsen undersøger uddannelse på vertikalt niveau, er litteraturstudiet foretaget i et bredt perspektiv. I litteraturstudiet inddrages der dermed studier, der omhandler uddannelsesbeslutninger på både vertikalt og horisontalt niveau. Herudover inddrages der studier om uddannelsespræstation, *educational expectations*, *educational attainment*. Litteratursøgningen har ikke været begrænset til danske studier, dermed inddrages studier foretaget i lande fra hele verden. Disse studier kan ikke sammenlignes direkte med resultater fra dette projekt, da strukturen i uddannelsessystemet er forskellig fra land til land og finansiering af uddannelse i Danmark er anderledes end i de fleste andre lande i verden. Uddannelse i Danmark er skattefinansieret og afhænger derfor ikke af den enkeltes finansieringsmuligheder. Dog kan økonomiske forhold stadig have betydning for uddannelsesbeslutningen. Årsagen til at litteraturstudiet foretages i så bredt et perspektiv er, at der ikke er mange studier, som undersøger netop valget af vertikalt uddannelse i dansk kontekst. Det vurderes også, at resultaterne fra de empiriske studier kan anvendes til at sige noget om det vertikale uddannelsesniveau, da de fokuserer på nogle af de samme faktorer.

Nogle af studierne, der inddrages i litteraturstudiet, fokuserer endvidere på specifikke uddannelsesvalg, eksempelvis valget mellem erhvervsuddannelse, gymnasiale uddannelse og ingen yderligere uddannelse. Studierne findes relevante at inddrage, da uddannelse er en proces, hvor man skal gennemføre et uddannelsesniveau, før man kan forsætte til det næste. Derfor vurderes det, at faktorer, hvor der eksisterer en sammenhæng med specifikke uddannelsesvalg, også er relevante at inkludere, når der fokuseres på det højeste fuldførte uddannelsesniveau. Der er udelukkende medtaget faktorer, der ligger indenfor rammen af neoklassisk og adfærdsøkonomisk teori samt socioøkonomiske faktorer.

Økonometrisk Metode

Problemstillingen undersøges ved hjælp af regressionsmodeller i en kvantitativ analyse. Regressionsmodellerne estimeres som ordered logistisk regressioner på baggrund af dataet fra kohorten. Det er med den valgte modeltype og dataafgrænsning ikke muligt at teste for kausale effekter (Euser et al., 2009). Den kvantitative analyse undersøger derved kun om der eksisterer signifikante sammenhænge mellem det højeste fuldførte uddannelsesniveau og de udvalgte uafhængige variable. Sammenhængene, der findes i den empiriske analyse, kan derfor skyldes andre forklarende variable, der er forskellige mellem individer, men som også påvirker det højeste fuldførte uddannelsesniveau. Det er derudover muligt, at der findes spuriøse sammenhænge, fordi der er udeladt variable, som har en kausal sammenhæng med valget af højeste fuldførte uddannelsesniveau. En detaljeret redegørelse af de økonometriske overvejelser fremgår af kapitel 6.

2.3 Data

I indeværende afsnit redegøres der for, hvor dataet, der benyttes i projektet, stammer fra. Derefter beskrives fordele og ulemper ved den valgte datakilde.

Anvendelighed af Dataet

Dataet, der benyttes i projektet, er registerdata fra Danmarks Statistiks Forskerservice. Adgangen til Danmarks Statistiks Forskerservice er opnået gennem Aalborg University Business Schools (AAUBS) adgang hertil. Dette betyder, at adgangen til data fra forskerservicen er begrænset af AAUBSs adgang til data på ét specifikt forskningsprojekt. Begrænsningen medfører, at der kan eksistere data, som ikke er muligt at anvende i projektet. Ultimativt kan datagrundlaget have en betydning for den kvantitative analyses resultat. Det vides dog ikke hvilke data, der ellers er tilgængeligt på Danmarks Statistiks Forskerservice. Det vurderes, at det data, der er tilgængelig gennem AAUBSs adgang, er tilstrækkelig til at besvare problemformuleringen. Dette vurderes, da det har været muligt at operationalisere størstedelen af de faktorer, neoklassisk arbejdsmarkedsteori samt empiriske studier finder relevant, jf. kapitel 3 og kapitel 4. Operationaliseringen af variablene samt udfordringerne i forbindelse med besvarelsen af de opstillede hypoteser præsenteres i kapitel 7 og bilag C.

Registerdata er indsamlet til administrativt brug på individniveau over individernes livsperiode. Med dataet er det muligt at samle informationer om uddannelse, indkomst, socioøkonomisk baggrund og andre typer af information på individniveau i ét datasæt. Dette betyder, at det er muligt at følge både individer, familier og grupper af individer, såsom uddannelsesårge. Registerdata er indsamlet til andre formål end de anvendes til i projektet. Dette kan betyde, at det ikke er alle faktorer, som findes relevante, der er mulige at medtage. Derudover er det nødvendigt at benytte proxies for andre faktorer. En detaljeret gennemgang af de anvendte datasæt kan findes i bilag A.

Fordelen ved registerdata er, at der er færre målefejl end i eksempelvis spørgeskemaundersøgelser. Der er for eksempel ingen recall-bias, da oplysninger bliver registreret, så personer ikke husker forkert. Derudover er der ingen systematiske bortfald i registreringen, da personer kun kan forlade dataet, hvis de udvandrer eller dør (Møberg, 2021). I projektets dataarbejde frasorteres bestemte personer på baggrund af manglende informationer omkring dem. Umiddelbart vurderes dette ikke at have en betydning for populationens repræsentativitet. Det anerkendes dog at systematisk frasortering af observationer kan forårsage bias i dataet, hvorfor betydningen heraf diskuteres i kapitel 9. Det endelige datasæt, der benyttes i projektet, indeholder 39.090 observationer. Til sammenligning var der ifølge Danmarks Statistik 55.595 personer, der færdiggjorde grundskolens 9. klasse i 2002 (Danmarks Statistik, n.d.h). Det lavere antal personer i datasættet skyldes blandt andet, at der er foretaget en begrænsning på baggrund af alder, at personer ikke må være under uddannelse i 2019 og at alle oplysninger for et individ skal være tilgængelige. Nærmere beskrivelse af hvordan det endelige datasæt er udarbejdet findes i bilag A.

Grafer og Tabeller

Da dataet fra Danmarks Statistiks Forskerservice indeholder personfølsomme oplysninger og da der samtidig er risiko for at kunne identificere personer, er der nogle regler, der skal overholdes ved præsentation heraf. For at sikre, at det ikke er muligt at identificere specifikke individer, er der en restriktion for et minimum af antal observationer, der må vises i tabeller, grafer og lignende. Derfor vil der, i tilfælde, hvor der er mindre end 5 observationer i tabeller eller grafer, blive enten fjernet eller sammenlagt enkelte grupper. Det fremgår tydeligt, når dette har været nødvendigt. Det vil eksempelvis fremgå af en tabel, at en værdi er fjernet på grund af risiko for identificering.

2.4 Det Danske Uddannelsessystem

Det beskrives i nærværende afsnit kort, hvordan det danske uddannelsessystem er opbygget. Beskrivelsen af uddannelsessystemet er vigtig for det videre arbejde, da opbygningen af uddannelsessystemet ligger til grund for rangeringen af uddannelsesniveauerne. Rangeringen af uddannelsesniveauerne er central for besvarelsen af problemformuleringen. På baggrund af rangeringen operationaliseres uddannelsesniveauerne i kapitel 6 til brug i den kvantitative analyse.

2.4.1 Opbygning af Uddannelsessystemet

Uddannelsessystemet er inddelt således, at de forskellige uddannelsesniveauer er opdelt efter læringsniveauer, som angivet i ISCED skalaen. ISCED skalaen er en international anerkendt skala, der blandt andet bruges til at sammenligne uddannelsesniveauer på tværs af lande (Eriksen, 2021; Børne- og Undervisningsministeriet, 2021). På skalaen er det laveste niveau grundskolen og det højeste niveau en forskeruddannelse. Efter grundskolen kan de unge vælge mellem at tage 10. klasse, begynde på en gymnasial uddannelse, en erhvervsuddannelse eller ikke at tage yderligere uddannelse. Hvis individernes højeste fuldførte uddannelse er 9. klasse eller 10. klasse defineres det i det videre arbejde, som at deres højeste fuldførte uddannelse er grundskolen. Årsagen hertil er, at 10. klasse er en frivillig forlængelse af grundskolen. En gymnasial uddannelse varer typisk 3 år og er en forbedrende uddannelse til en videregående uddannelse. For at blive optaget på en gymnasial uddannelse skal man erklæres uddannelsesparat. En erhvervsuddannelse varer typisk 4,5 år og forbereder de studerende til at deltage på arbejdsmarkedet efter endt uddannelse. I 2015 blev der indført karakterkrav på erhvervsuddannelser. Før 2015 blev alle der søgte ind på en erhvervsuddannelse optaget (Eriksen, 2021). Videregående uddannelse inddeles i tre niveauer: kort-, mellemlang- og lang videregående uddannelse. En kort videregående uddannelse tager 2 år. En mellemlang videregående uddannelse er enten en professionsbachelor, som typisk varer 3,5 år eller en akademisk bachelor, som er normeret til 3 år. En kandidatuddannelse er en overbygning på en mellemlang videregående uddannelse og er normeret til 2 år (Børne- og Undervisningsministeriet, n.d.c). Efter en kandidat er det muligt at tage en ph.d, som er normeret til 3 år (Børne- og Undervisningsministeriet, n.d.c). Der er mange veje gennem uddannelsessystemet. Det er muligt, at blive optaget på både en kort videregående uddannelse eller en professionsbachelor, med både en erhvervsuddannelse og en gymnasial uddannelse. Optagelse på en akademisk bachelor kræver derimod en gymnasial uddannelse (Børne- og Undervisningsministeriet, 2020). Ud over kravene om fuldførelse af uddannelse på et lavere niveau, er der på nogle videregående uddannelser specifikke krav til for eksempel karakterniveau, gennemførelse af specifikke fag eller at individer består en optagelsesprøve.

2.4.2 Rangering af Uddannelsesniveauer

Formålet med den kvantitative analyse er at undersøge sammenhængen mellem flere faktorer og individers højeste fuldførte uddannelsesniveau. Derfor er individernes højeste fuldførte uddannelse kategoriseret ud fra Danmarks Statistiks Forskerservice datasæt *udda2019*. Hensigten er, at uddannelsesniveauer skal anvendes som afhængig variabel i den forestående regressionsanalyse. Til at klassificere de forskellige uddannelser, anvendes *n_audd2015t_l1l5* formatet fra Danmarks Statistiks Forskerservice. Formatet indeholder 4.131 observationer, der hver repræsenterer en uddannelse i Danmark. Hver af de 4.131 uddannelser kobles på en af følgende kategorier: *95 Afbrudt*, *10 Grundskole*, *20 Almengymnasiale*

uddannelser, 25 Erhvervsgymnasiale uddannelser, 35 Erhvervsfaglige praktik- og hovedforløb, 40 Korte videregående uddannelser, 50 Mellemlange videregående uddannelser, 60 Bachelor, 65 Lange videregående uddannelser samt 70 Forskeruddannelser. Formatet gør det altså muligt at kategorisere de 4.131 forskellige uddannelser i nogle overordnede grupper. Da flere uddannelser er på det samme niveau på ISCED skalaen, er der foretaget en rangering for at skabe en ordinal diskret variabel, der kan benyttes som den afhængige variabel i arbejdet med at besvare problemstillingen.

Uddannelserne er i projektet følgende, rangeret fra lavest til højest: grundskole, gymnasial uddannelse, erhvervsuddannelse, kort videregående uddannelse, akademisk bachelor, professionsbachelor samt lang videregående uddannelse. Rangeringen er illustreret i figur 2.1. Denne rangering giver ikke mulighed for at et individs uddannelsesniveau er lavere end grundskolen, da disse personer er sorteret fra i kohorten. Rangeringen er den samme, som i ISCED skalaen bortset fra gymnasial uddannelse og erhvervsuddannelse, hvor erhvervsuddannelser her er defineret til at være på et højere niveau, da det er en erhvervskompetencegivende uddannelse. En erhvervskompetencegivende uddannelse gør en i stand til at varetage arbejdsopgaver indenfor et bestemt erhverv (Børne- og Undervisningsministeriet, n.d.b). Mellemlang uddannelse er opdelt i akademisk bachelor og professionsbachelor. Dette er gjort for at adskille dem, da det kun er de færreste af de akademiske bacheloruddannelser, der giver reelle erhvervskompetencer, hvilket er nødvendigt for at gå direkte ud på arbejdsmarkedet (Houliind, 2018; DØRS, 2003). Professionsbachelor er derfor defineret som at være på et højere niveau end en akademisk bachelor, da det antages at personer med en professionsbachelor har nemmere ved at få et job (Danmarks Statistik, 2017). Desuden defineres en forskeruddannelse, som en lang videregående uddannelse. Denne definition er blandt andet foretaget, da det kun er et mindretal af befolkningen, der har en forskeruddannelse. Derudover vurderes det også, at en forskeruddannelse er en anden type uddannelse. Det er for eksempel samtidig med at være en uddannelse også en ansættelse, der er rettet mod en speciel jobtype. For at sikre, at der er nok observationer til at fortolke på resultater er disse slået sammen. Desuden er det kun en mindre andel af forskningsartikler, der foretager en opdeling mellem disse uddannelsesniveauer, hvilket medfører, at resultaterne fra den kvantitative analyse er mere sammenlignelige med andre studier. I projektet defineres de fire højeste uddannelsesniveauer som videregående uddannelse (Børne- og Undervisningsministeriet, n.d.c).

Rangeringen er endvidere lavet ud fra, at individer i neoklassisk økonomi antages at være rationelle agenter. Ud fra denne tankegang vil individerne sætte de forskellige uddannelsesniveauer i en hierarkisk orden efter deres forventede økonomiske afkast (Jaeger, 2007).

Figur 2.1: Oversigt over Rangering af Uddannelsesniveauer



Investeringsbeslutningen i Uddannelse 3

I dette kapitel redegøres der for neoklassisk mikroøkonomisk teori, der omhandler, hvordan individer træffer beslutningen om at investere i uddannelse. Herunder beskrives *schooling* modellen og den underliggende *wage-schooling locus*. Formålet med kapitlet er at redegøre for økonomisk teori, der ligger til grund for investeringsbeslutningen i uddannelse. Til sidst i kapitlet diskuteres begrænsningerne i teorien.

3.1 Investeringsbeslutninger

Humankapital er defineret som akkumuleringen af investeringer i individer, såsom uddannelse eller jobtræning (Mankiw og Taylor, 2017). Denne investering giver individer kompetencer. Hvordan et individ vælger at sammensætte sine kompetencer kaldes indenfor arbejdsmarkedsteori investeringsbeslutningen i uddannelse. *Schooling* modellen beskriver, hvordan individer tager denne beslutning og hvilke faktorer, der påvirker beslutningen. Mere specifikt undersøger modellen, hvornår individer vælger at stoppe med at tage mere uddannelse (Borjas, 2016).

For at foretage en investeringsbeslutning skal alle omkostninger og indtægter forbundet hermed inkluderes. Dette gælder uanset om det er en investering i uddannelse eller fysisk kapital. Værdien af investeringen udregnes som nutidsværdien, der er lig den tilbagediskonterede værdi af de omkostninger og indtægter, der teoretisk set er forbundet med investeringsbeslutningen (Borjas, 2016). Ved at opgøre værdien som en nutidsværdi, er det muligt at sammenligne indtægter og udgifter fra forskellige perioder. Diskonteringsrenten er et udtryk for, hvordan man vægter fremtidige indtægter og omkostninger. Jo højere diskonteringsrenten er, jo mindre vægt tillægger man fremtidige omkostninger og indtægter. Diskonteringsrenten er derved en afvejning af, hvor meget af ens nuværende forbrugsevne man er villig til at give op for at kunne forbruge mere i fremtiden (Borjas, 2016).

3.2 Schooling Modellen

I *schooling* modellen antages det, at individer vælger at tage den mængde uddannelse, der maksimerer nutidsværdien af deres livsindtjening. Ifølge teorien vælger et individ dermed at tage mere uddannelse, da det vil øge vedkommendes fremtidige indtjening (Borjas, 2016).

I *schooling* modellen afgrænses der fra at tage andre faktorer end indkomst i betragtning og det antages dermed, at det kun er den fremtidige indtjening som følge af uddannelse, der spiller en rolle i forhold til beslutningen om at investere i uddannelse. Der afgrænses derfor i modellen fra at undersøge, hvordan et individs nytte påvirkes på andre områder. Det antages derudover, at der ikke er *on-the-job training* samt, at allerede tilegnede evner ikke deprecierer over tid. Sammenfattende, afhænger et individs beslutning om at tage mere uddannelse i modellen udelukkende af det økonomiske afkast herved. I forbindelse hermed antages det i *schooling* modellen, at lønnen stiger i takt med, at mere uddannelse optages (Borjas, 2016). Der er ifølge modellen to forskellige typer omkostninger. Der er de reelle omkostninger forbundet med at tage ekstra uddannelse, det kan eksempelvis være studieafgift og indkøb af bøger eller andet udstyr. Derudover er der også offeromkostninger forbundet med at tage uddannelse. Offeromkostningerne er den indtjening man går glip af ved ikke at deltage aktivt på arbejdsmarkedet. Derudover er der ikke-finansielle omkostninger, som dækker over individerens indsats og anstrengelser for at gennemføre uddannelse og eventuel misfornøjelse ved at studere. Det antages, at uddannelse ikke har nogen indeværende effekt

for individet, men i højere grad er et signal for en højere produktivitet. Dette betyder, at uddannelse ikke har nogen egentlig påvirkning på individets velbefindende eller livskvalitet. I forlængelse heraf vil arbejdsgiverne være nødt til at tilbyde højere løn, hvis en person har en længere uddannelse ellers vil ingen tage ekstra uddannelse. Hvis arbejdsgiverne ikke kompenserer for den tid, som arbejdstagerne bruger på at uddanne sig, vil arbejdstagerne ikke have incitament til at tage yderligere uddannelse. Incitamentet forsvinder, da arbejdstagerne ikke vil blive kompenseret for de omkostninger, de har i forbindelse med ekstra uddannelse. Derfor vil de ikke maksimere deres livstidsindkomst ved at tage ekstra uddannelse (Borjas, 2016).

Formelt kan nutidsværdien af den fremtidige indtjening for et individ, der eksempelvis har en grundskole som højeste afsluttede uddannelse, skrives på følgende måde:

$$PV_{GRUND} = w_{GRUND} + \frac{w_{GRUND}}{(1+r)} + \frac{w_{GRUND}}{(1+r)^2} + \dots + \frac{w_{GRUND}}{(1+r)^t} \quad (3.1)$$

hvor PV angiver nutidsværdien af livsindkomsten, w er løn, r angiver diskonteringsrenten og t er tidsperioden (Borjas, 2016).

Hvis den højeste afsluttede eksamen eksempelvis er en gymnasial uddannelse, der antages at være 3 år, kan nutidsværdien af den fremtidige indtjening for et individ skrives på følgende måde:

$$PV_{GYM} = -H - \frac{H}{(1+r)} - \frac{H}{(1+r)^2} + \frac{w_{GYM}}{(1+r)^3} + \frac{w_{GYM}}{(1+r)^4} + \dots + \frac{w_{GYM}}{(1+r)^t} \quad (3.2)$$

Hvor H angiver de årlige omkostninger forbundet med uddannelse. De første tre termer angiver nutidsværdien af omkostningerne ved at tage en gymnasial uddannelse. De efterfølgende termer angiver nutidsværdien af den fremtidige indtjening efter uddannelsen. Det antages, at antallet af tidsperioder t er det samme, uagtet om ens højeste uddannelse er en grundskole eller en gymnasial uddannelse. Derudover antages det, at omkostninger til uddannelse er ens for alle, og at lønnen kun afhænger af antal års uddannelse (Borjas, 2016). Ud fra ligning 3.1 og ligning 3.2 ses det, at nutidsværdien af et individs fremtidige indtjening afhænger af løn, omkostningerne til uddannelse, antal tidsperioder og diskonteringsrenten. I modellen antages det som tidligere nævnt, at et individ vil maksimere nutidsværdien af sin livsindkomst. Det er altså nutidsværdien af livsindkomsten, der bestemmer, hvor lang en uddannelse individet tager. Individer vil vælge at tage en gymnasial uddannelse, hvis nutidsværdien af indtjeningen er højere end nutidsværdien af indtjeningen ved udelukkende at tage grundskolen (Borjas, 2016).

$$PV_{GRUND} > PV_{GYM}$$

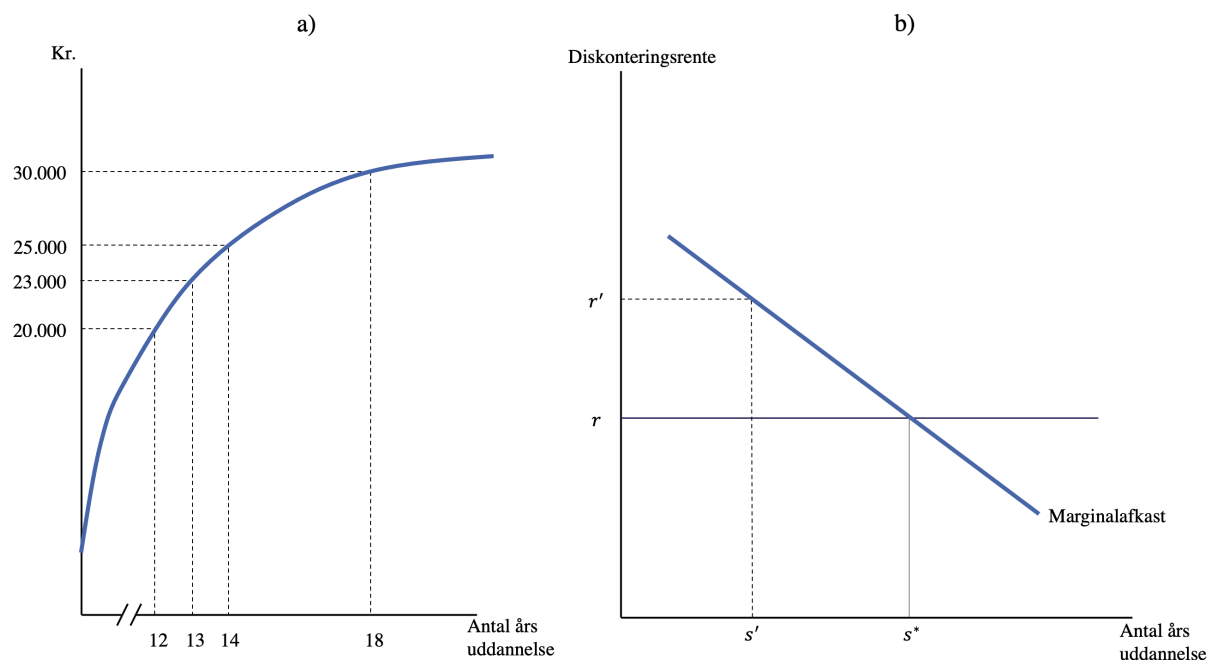
Det optimale uddannelsesvalg for individer er derved det antal år, der maksimerer nutidsværdien af deres samlede indtjening gennem hele livet. Forskellen i personers uddannelsesvalg afhænger derved af diskonteringsrenten, som er den eneste faktor, der kan variere mellem individer. Diskonteringsrenten antages at være konstant og perfekt elastisk for et individ (Borjas, 2016).

I tilfælde af at to individer udelukkende er forskellige i deres diskonteringsrente, vil individet med en højere diskonteringsrente stoppe med uddanne sig tidligere end det individ, der har en lavere diskonteringsrente. Dette skyldes, at personen er mere nutidsorienteret (Borjas, 2016).

3.2.1 Wage-Schooling Locus

Et individs uddannelsesbeslutning kan generaliseres til en situation med to eller flere valgmuligheder. I sammenhæng hermed er det relevant at se på *wage-schooling locus* modellen. Denne model er en grafisk repræsentation af, hvor meget løn en arbejdsgiver er villig til at betale for X antal års uddannelse, jf. figur 3.1.a. *Wage-schooling locus* er en del af *schooling* modellen. Ifølge *wage-schooling locus* er lønnen markedsbestemt og afhænger dermed af udbuddet af arbejdstagere med det pågældende antal års uddannelse og efterspørgslen heraf (Borjas, 2016). Fra arbejdstagerens synspunkt er lønnen, der associeres med et uddannelsesniveau, konstant og et individ ved derfor, at 12 års uddannelse vil give en løn på eksempelvis 20.000 kr. årligt.

Figur 3.1: Wage-Schooling Locus



Kilde: Tilpasset fra Borjas (2016)

Hældningen på kurven i figur 3.1.a er positiv, hvilket betyder, at længere uddannelse medfører højere løn. Arbejdsgiverne skal tilbyde en højere løn til højere uddannede medarbejdere for at kompensere for de omkostninger, der er forbundet med den længere uddannelse. Overordnet indikerer hældningen i *wage-schooling locus* modellen, hvor meget lønnen stiger, hvis et individ bliver i skole i yderligere et år. Hældningen er ifølge teorien tæt relateret med et empirisk mål for *rate of return* ved uddannelse. Rate of return ved uddannelse kaldes også afkastraten af uddannelse og er den procentvise ændring i lønnen ved et ekstra års uddannelse (Borjas, 2016).

Som det fremgår af figur 3.1.a, har kurven en konkav form. Dette skyldes, at den økonomiske gevinst ved at gennemføre et års ekstra uddannelse falder i takt med, at mere uddannelse tilegnes. Dette betyder, at marginalafkastet af uddannelse ifølge modellen er aftagende (Borjas, 2016). Marginalafkastet af uddannelse (MMR) kan dermed illustreres ved en lineær nedadgående kurve, jf. figur 3.1.b (Borjas, 2016).

Ifølge *wage-schooling locus* stopper man med at tage yderligere uddannelse, når marginalafkastet af uddannelse er lig med diskonteringsrenten. Derved afhænger uddannelsesbeslutningen af to faktorer, der kan variere fra individ til individ; diskonteringsrenten r og marginalafkastet af uddannelse (Borjas, 2016). I tilfælde, hvor individer, der har den samme diskonteringsrente har forskellige *wage-schooling locus*, skyldes det, at de har forskellig afkastrate af et ekstra års uddannelse. Forskellen skyldes, at personer har forskellige evner. Personer med et højere niveau af evner har et større afkast af ekstra uddannelse, hvorfor deres kurve for marginalafkast er længere til højre end hos personer med et lavere niveau af evner. Det betyder, at kurven for marginalafkast skærer senere med diskonteringsrenten for folk med et højere niveau af evner, hvorfor de vil gennemføre flere antal års uddannelse.

3.3 Begrænsninger i Teorien

I det følgende diskuteres anvendelsen af den præsenterede teori i en kvantitativ analyse. Efterfølgende diskuteres begrænsningerne i *schooling* modellen og den neoklassiske teori generelt.

Investeringsbeslutningen i humankapital handler i den præsenterede model om at maksimere livsind-

komst. Forskellen i hvor meget personer investerer i uddannelse forklares i modellen af personernes diskonteringsrente og evner. Empirisk er der fundet belæg for, at en lang række andre faktorer hænger sammen med investeringsbeslutningen i uddannelse. Disse faktorer er baseret på teori udenfor neoklassisk økonomi, som ikke præsenteres her. Et eksempel er socioøkonomiske faktorer (Benders, 2013). Indenfor nogle grene af økonomi argumenteres der dog for, at socioøkonomiske faktorer er underliggende i den neoklassiske teoris antagelser. Der argumenteres for at individer tager højde for disse faktorer når de fastsætter diskonteringsrenten. Dermed er argumentet, at teori fra andre discipliner, herunder sociologi, underbygger den økonomiske teori og at man dermed alene ud fra økonomisk teori kan forklare uddannelsesvalg (Rose, 1952). Der er evidens for, at lavindkomstfamilier har en højere diskonteringsrente end højindkomstfamilier, hvilket er et eksempel på, at der er forskel i diskonteringsrenten på baggrund af socioøkonomiske faktorer (Borjas, 2016).

Herudfra kan det diskuteres, hvorvidt socioøkonomiske faktorer i sig selv er forklarende, eller om de er underliggende i økonomisk teori. Hvis faktorerne er underliggende i den neoklassiske teori, kan der stilles spørgsmålstejn til, om det er nødvendigt at inddrage dem, når investeringsbeslutningen i uddannelse undersøges. Problemet med udelukkende at tage udgangspunkt i diskonteringsrenten er, at det ikke vides, hvordan den fastsættes og at den ikke kan observeres. Hvis man vil inddrage den, er det derfor nødvendigt at benytte en proxy. Nogle gange benyttes markedsrenten som proxy, da det er den rente, man kan få på markedet ved at investere penge i andre investeringsgoder end uddannelse (Borjas, 2016). Det antages dog i *schooling* modellen, at diskonteringsrenten er konstant for individer. Markedsrenten ændrer sig over tid og derfor er det ved brug af denne som proxy nødvendigt at vide, hvornår individer fastsætter deres diskonteringsrente. Overordnet er diskonteringsrenten dermed vanskelig at estimere.

Vender man tilbage til argumentet om, at socioøkonomiske faktorer er underliggende i diskonteringsrenten, er det dog muligt at inddrage faktorer, der argumenteres for at påvirke fastsættelsen af diskonteringsrenten. Ved inddragelse af faktorer, der påvirker diskonteringsrenten, er det altså muligt indirekte at tage højde for den.

Udover at det ikke er muligt at måle diskonteringsrenten direkte, er der også nogle andre begrænsninger i anvendelsen af teorien. En forudsætning i teorien er, at det er muligt at differentiere uddannelse efter små ændringer i længden på uddannelse. Dette kan eksempelvis være ét års ekstra uddannelse. I virkeligheden er dette ikke muligt, da uddannelsessystemet er inddelt med diskrete hop, jf. kapitel 2.4.1. Arbejdsgivere er ikke villige til at betale mere for at en person, har gennemført en del af en uddannelse. Det er desuden svært at blive ansat uden et eksamensbevis eller blive aflønnet for en uddannelse uden dokumentation. Den manglende aflønning for et ekstra års uddannelse kan skyldes, at uddannelse anses som en signalværdi for allerede iboende evner. Et eksamensbevis er derved en måde at signalere, at man er mere produktiv end individer uden et eksamensbevis. Ved ikke at gennemføre uddannelse signalerer individet, at vedkommende er mindre produktiv end individer, der gennemfører uddannelsen (Borjas, 2016).

Der antages i *schooling* modellen, at viden fra uddannelse ikke deprecierer over tid, hvilket er urealistisk, da nye teknologier opstår og viden går tabt, hvis det ikke holdes ved lige (DØRS, 2003). Det antages derudover, at der ikke er on-the-job træning. Dette er til trods for, at danske virksomheder brugte 15 mia. kr. på efteruddannelse og interne kurser, i 2017 (Dansk Arbejdsgiverforening, 2019). Lønnen i teorien afspejler, at personer med længere uddannelse er mere produktive end personer med en kortere uddannelse. Dermed aflønnes personer efter deres produktivitet, så lønforskellen udelukkende skyldes produktivitetsgab mellem individer på grund af uddannelse. Der kan dog være andre forklaringer på lønforskellen, eksempelvis et købermonopol. Købermonopolet kan holde efterspørgslen nede for at presse lønnen ned. Det kan f.eks. være tilfældet for sygeplejersker og folkeskolelærere. Tilgængelighed kan en høj grad af organisering blandt lønmodtagere medvirke til en højere løn end den markedsbestemte (DØRS, 2003).

I teorien antages det, at den eneste grund til at tage yderligere uddannelse er for at maksimere ens livsindkomst, men der findes evidens for, at uddannelse også har andre positive effekter. Det er blandt andet fundet, at personer, der har en længere uddannelse, lever længere og er sundere. Derudover oplever de lavere arbejdsløshedsrate, højere jobsikkerhed og lavere arbejdsbelastning på arbejdsmarkedet. Uden for arbejdet oplever de større social tilknytning, højere personlig tryghed og subjektiv trivsel, stærkere civilt engagement og er mere bekymrede for miljøet og menneskerettigheder. Disse positive effekter er

ikke bare et resultat af højere indkomst (Llena-Nozal et al., 2019). De positive effekter ved uddannelse bidrager også i sig selv til en højere livsindkomst. Der kan dermed være andre faktorer, der er relevante at overveje, når en person foretager investeringsbeslutningen i humankapital.

Overordnet er der nogle antagelser i *schooling* modellen, der ikke stemmer overens med virkeligheden. Antagelserne i den neoklassiske teori, der er underliggende modellen, har også nogle begrænsninger.

I neoklassisk teori antages det, at individer har perfekt information. Det gælder også perfekt information om den fremtidige indkomst, kurven for *wage-schooling locus*, og derved afkastraten for uddannelse ved de forskellige uddannelseslængder (Borjas, 2016). Individer har dog i realiteten imperfekt viden om nytten for alle de forskellige uddannelser både finansielt og personligt. Derved overholdes denne antagelse ikke i praksis. Personer har ifølge adfærdøkonomi *bounded rationality*, da der er begrænsede mængder af tilgængelig information, tid og hukommelse. På grund af disse begrænsninger er personer ikke i stand til at sammenligne information om alle de uddannelser, der findes (Samson, 2019). I 2017 var der 880 uddannelsesmuligheder i Danmark (Thomsen, 2017). Dermed er der mange uddannelser en person skal samle information om og forholde sig til. De mange muligheder kan føre til *choice overload*. Choice overload betyder, at der er psykologiske omkostninger ved at tage beslutningen, da valget kan være svært og kræve mange kræfter. I tilfælde, hvor der opstår choice overload, kan personer opleve fænomenet *decision fatigue*. Decision fatigue er, hvor individer tager beslutninger på et meget begrænset grundlag eller helt vælger ikke at tage en beslutning, da de ikke kan overskue de mange valgmuligheder (Samson, 2019). Samtidig ændrer økonomien og de sociale forhold sig, så det ikke er muligt at forudsætte hvilke uddannelser, der er behov for i fremtiden og hvordan ændringer påvirker afkastet af de forskellige uddannelsesniveauer. Denne usikkerhed gør det sværere at foretage investeringsbeslutningen i virkeligheden.

Den neoklassiske antagelse om perfekt rationalitet, der underligger teorierne, er desuden urealistisk. For at være perfekt rationelle skal individer være tidskonsistente, hvilket betyder, at deres præferencer skal være ens over tid. Der findes dog forskellige tilfælde af myopia, hvor kortsigtede hensyn tillægges større værdi end den langsigtede værdi af uddannelse. I disse tilfælde maksimerer individer ikke deres livstidsindkomst (Andersen et al., 2018).

Overordnet har *schooling* modellen og den underliggende neoklassiske teori, nogle antagelser, der ikke stemmer overens med virkeligheden. Det har andre teorier, der ligger bag faktorerne, hvor der empirisk er fundet en sammenhæng, dog også. På trods af at antagelserne ikke stemmer overens med virkeligheden, er de nødvendige for, at teorien kan forklare den generelle sammenhæng. Alternativt er der for mange specialtilfælde, der skal tages højde for. *Schooling* modellen er også begrænset af, at det er vanskeligt at operationalisere eksempelvis diskonteringsrenten i praksis. Da disse begrænsninger eksisterer, vurderes det ikke at være tilstrækkeligt alene at anvende teorierne i undersøgelsen af hvilke faktorer, der påvirker unges valg af uddannelsesniveau. Derfor undersøges det hvilke faktorer, der empirisk er fundet belæg for at hænge sammen med uddannelsesvalg.

3.4 Opsummering

I *schooling* modellen er der to faktorer, der forklarer forskellen mellem personers uddannelsesbeslutninger. Disse to faktorer er diskonteringsrenten og marginalafkastet af uddannelse, som er forskellige alt efter personers evner. En lavere diskonteringsrente antages i modellen at øge sandsynligheden for at individer tager længere uddannelse. Mens et lavere niveau af evner antages at mindske sandsynligheden for at individer tager længere uddannelse. I modellen afhænger beslutningen dermed udelukkende af det økonomiske afkast. Evner og diskonteringsrenten er faktorer, der er svære at operationalisere i virkeligheden. Som følge af udfordringerne med operationalisering af teorien, findes det interessant at undersøge, hvordan empiriske studier undersøger unges valg af uddannelsesniveau.

Empirisk Litteraturstudie 4

I kapitlet undersøges det, hvilke faktorer der empirisk er fundet at hænge sammen med unges valg af uddannelse. Først beskrives de metodiske afgrænsninger i forhold til teori, studietype og litteratursøgning. Derefter beskrives de empiriske resultater for hver enkelt faktor. Det empiriske litteraturstudie har til formål at indkredse hvilke faktorer, der er relevante at inddrage i den forestående kvantitative analyse.

4.1 Metodiske Afgrænsninger

Den bagvedliggende teori for faktorerne, der inddrages i litteraturstudiet, kommer både fra den økonomiske disciplin og fra andre samfundsvidenskabelige discipliner. Den underliggende teori for faktorerne i dette kapitel er udover økonomisk teori eksempelvis *Sociological Rational Choice Theory*, som kombinerer økonomisk og sociologisk teori. Ifølge denne teori maksimerer individer både økonomiske mål men også sociale mål, da uddannelsesbeslutninger i teorien altid tages i en social kontekst. Der er i teorien et socialt afkast ved at vedligeholde sociale netværk, anerkendelse fra peer-gruppe og familie samt intergenerationel fastholdelse af familiestatus (Jaeger, 2007). En anden underliggende teori for faktorerne er relativ risikoaversion. I denne sociologiske teori træffes uddannelsesvalg ud fra en individuel og rationel vurdering af det forventede afkast, omkostninger samt sandsynligheden for succes ved de forskellige uddannelsesvalg. I teorien har individer en stærk præference for at minimere risikoen for nedadgående mobilitet og opnå et lavere uddannelsesniveau end deres forældre. Den marginale nytte ved ekstra uddannelse aftager i modellen, når individerne har opnået samme uddannelsesniveau, som deres forældre (Munk og Mattsson, 2008). Ud over teorierne er der flere begreber, der understøtter de empiriske fund, herunder intergenerationel mobilitet, tidsinkonsistente præferencer og *self-evaluation bias*. Teorier fra andre samfundsvidenskabelige discipliner samt andre teoretiske retninger end neoklassisk teori uddybes ikke nærmere, da projektet har et neoklassisk økonomisk fokus, jf. kapitel 2.

Studierne, der inddrages i dette afsnit, er, som beskrevet i kapitel 2, en blanding af kvalitative og kvantitative studier. Det er kun en lille andel af studierne, der undersøger, om der eksisterer kausale sammenhænge mellem den specifikke faktor og uddannelsesvalget. Derved kan resultaterne fra studierne forårsages af en anden underliggende faktor, der ikke er taget højde for i studiet. Det kan derfor heller ikke udelukkes, at der ikke eksisterer omvendt kausalitet mellem faktoren og uddannelsesvalget. De udvalgte faktorer er dog funderet i teori, som understøtter en påvirkning fra faktoren til uddannelsesvalget og ikke omvendt.

Litteraturstudiet er som nævnt foretaget i et bredt perspektiv. Der er derfor inddraget studier og faktorer, hvor det er fundet, at der eksisterer en sammenhæng med individers uddannelsespræstation. I projektet antages det, at en bedre uddannelsespræstation medfører, at et individ er mere tilbøjeligt til at tage uddannelse på et højere niveau. Denne antagelse foretages, da flere empiriske studier finder en sammenhæng mellem forskellige mål for præstation i uddannelse og sandsynligheden for at fortsætte med yderligere uddannelse (Produktivitetskommissionen, 2014; Jaeger, 2007). Endvidere er det en antagelse i *schooling* modellen, at ens evner påvirker ens investeringsbeslutning i uddannelse. Både studier, der undersøger *educational attainment* og *educational expectations*, er inddraget, da mange af de samme faktorer inddrages i undersøgelser af, hvad der påvirker gennemførelsen af højeste uddannelsesniveau og forventninger om hvad ens højeste fuldførte uddannelsesniveau bliver (Thomsen og Bom, 2019). Endvidere inddrages der studier, som hovedsageligt fokuserer på horisontal uddannelse. Der er dog kun resultater med fra disse studier, som forklarer forskellen i uddannelsesniveau. I gennemgangen af litteraturen er

Benders (2013), Koch et al. (2015), Jaeger (2007) og Nielsen og Vissing-Jorgensen (2006) inkluderet. Disse kilder indeholder litteraturreviews af andre studier, der undersøger det samme interessefelt. Derfor kan resultater fra disse kilder godt stamme fra flere undersøgelser.

4.2 Faktorer

I indeværende afsnit beskrives hver enkelt faktor, som ifølge empiriske studier har en sammenhæng med personers uddannelsesvalg, samt i hvilken retning sammenhængen er.

4.2.1 Socioøkonomiske Faktorer

Forældres socioøkonomiske status er og har ifølge Reardon et al. (2012) altid været en af de stærkeste indikatorer for unges uddannelsesmæssige præstation og opnåelse (Benders, 2013). Socioøkonomiske faktorer og socioøkonomisk status er dog forholdsvis komplekse begreber og udgøres af mere end blot forældrenes indkomst. Faktorerne der inddrages i dette afsnit defineres i andre studier som socioøkonomiske faktorer. Socioøkonomiske faktorer og status kan dermed i højere grad defineres som en gruppe af indikatorer (Benders, 2013). Det er derfor svært at vide hvilke af de socioøkonomiske faktorer, der driver de unges uddannelsesvalg. Det medfører at der kan opstå en spuriøs sammenhæng mellem de socioøkonomiske faktorer og unges uddannelsesvalg.

Forældres Indkomst

I mange tilfælde forbindes socioøkonomisk status med forældres indkomst og flere studier undersøger sammenhængen mellem forældres indkomst og uddannelsesmæssige resultater. Adskillige større studier har fundet evidens for, at elever fra familier med højere indkomst har højere karaktergennemsnit (Benders, 2013). Sean F. Reardon, sociolog fra Stanford University, har fundet ud af, at der eksisterer et såkaldt *achievement gap* mellem elever fra lavindkomstfamilier og velhavende familier. Det amerikanske undervisningsministerium definerer achievement gap som forskellen i faglige præstationer mellem grupper af studerende (Benders, 2013). Ifølge Reardon var forskellen mellem testscorer fra studerende fra lavindkomst- og velhavende familier steget med ca. 40% fra 1960'erne til 2007 (Reardon et al., 2012). Resultaterne fra det amerikanske studie skal ses i lyset af, at indkomstuligheden er større i USA end i Danmark. Derfor kan det ikke forventes, at forskellen i faglige præstationer for elever, der kommer fra en lavindkomstfamilie kontra en velhavende familie, er i samme størrelsesorden i Danmark, som Reardon et al. (2012) har fundet i USA. Andersen et al. (2018) finder, at forældre med en beskeden indkomst har færre økonomiske ressourcer til at finansiere deres børns uddannelse. Da uddannelse er finansieret gennem skatten i Danmark, forventes denne effekt være mindre dominerende i Danmark end i lande, hvor uddannelse ikke er finansieret gennem skatten.

Familietype

I relation til forældrenes indkomst, findes der også evidens for, at der er en sammenhæng mellem familietype og de unges faglige præstationer (Benders, 2013). Med familietype refereres der til, hvorvidt et individs forældre er samboende eller enlige.

Ifølge Benders (2013) lever flere elever end tidligere i familier med én forældre, hvilket kan have en betydning for elevens faglige resultater. Argumentet er, at en enlig forældre er eneforsøger og derfor har mindre tid og ikke optimale muligheder for at yde støtte til lektielæsning eller vigtige beslutninger. Endvidere påpeges det, at husholdninger med enlige forældre som oftest har en lavere indkomst (Benders, 2013). I en rapport udarbejdet af Lacour og Tissington (2011) konkluderes det, at forældres involvering spiller en stor rolle for børnenes faglige resultater. Samme forfattere har desuden konkluderet, at børn fra hjem med enlige forældre ikke klarer sig ligeså godt i skolen som børn fra hjem med to forældre (Benders, 2013; Lacour og Tissington, 2011).

Forældres Uddannelse

En rapport udarbejdet af tænketanken DEA peger på, at en del af forskellen på unges uddannelsevalg er forklaret af deres forældres, bedsteforældres samt søskendes uddannelsesniveau (Østergaard-Tygesen, 2018). Ifølge adskillige studier findes der en sammenhæng mellem unges uddannelsesniveau og deres forældres uddannelsesniveau. Denne tendens findes i Danmark på trods af, at der eksisterer en højere

social mobilitet end eksempelvis i USA og at yngre årgange bliver mere uddannede end ældre. Tendensen eksisterer, selvom der tages højde for karakterniveauet på ungdomsuddannelser¹ (Andersen et al., 2018). I 2007 var det næsten hver anden 30-årig i Danmark, der havde samme uddannelsesniveau, som mindst en af deres forældre (Pihl, 2007). Ifølge Østergaard-Tygesen (2018) er forældrenes uddannelsesniveau den mest betydningsfulde enkeltfaktor i forhold til de unges uddannelsesvalg. Argumentet herfor er, at forældrenes kendskab til uddannelsessystemet og arbejdsmarkedet har en afgørende betydning, da et begrænset kendskab hertil kan være en udfordring. Dette kan være en udfordring, da forældre ikke kan guide de unge gennem uddannelsessystemet. De har et begrænset kendskab til ansøgningsprocessen, kravene til de forskellige uddannelsesniveauer og hvor mange gange man kan tage grundforløbet på en erhvervsuddannelse. Resultaterne fra et pakistansk studie understøtter, at der eksisterer en sammenhæng mellem det hjemlige miljø og elevers faglige bedrifter samt at uddannede forældre i højere grad kan skabe et miljø, der skaber grobund for akademisk succes for deres børn (Benders, 2013; Farooq et al., 2011). Ifølge Nielsen (2012) er kausaliteten mellem forældres uddannelse og de unges uddannelse begrænset og der argumenteres for, at sammenhængen hovedsageligt skyldes underliggende evner ved forældre, der påvirker den mængde tid og penge forældrene investerer i børnene. Dette understøttes af adfærdsøkonomisk forskning, der dog finder en stærk korrelation mellem generationer, der er relateret til et bredere perspektiv af familiemiljøet end kun uddannelse og indkomst. Adfærdsøkonomisk evidens om overførsel af bløde værdier og præferencer mellem generationer viser, at familiemiljøet spiller en stor rolle når det kommer til formning af præferencer og bløde værdier, hvilket videre påvirker investering i uddannelse (Koch et al., 2015).

Socialklasse

Individets socialklasse anses ifølge Benders (2013) generelt for at være en indikation på økonomiske og uddannelsesmæssige muligheder eller privilegier. Ifølge Benders (2013) reflekterer det dog et bredere aspekt af menneskelige kvaliteter. Ifølge ham er der et kulturmæssigt aspekt, hvilket på grund af vigtigheden af uddannelsesmæssige færdigheder for beskæftigelsesmæssig status, har nogle naturlige associationer med uddannelsesmæssige resultater, der går ud over økonomiske- og uddannelsesmuligheder (Benders, 2013). De studier, der er inkluderet i indeværende kapitel, klassificerer socialklasse ud fra tilknytningen til uddannelsessystemet, indkomst og beskæftigelse (Becker og Hecken, 2009; Østergaard-Tygesen, 2018). Danmarks Statistik klassificerer socialklasse ud fra personens væsentligste indkomstkilde eller beskæftigelse (Danmarks Statistik, n.d.v). Ud fra disse definitioner af socialklasse forventes der at være en stærk korrelation mellem denne og forældres uddannelsesniveau og indkomst. Danmarks Statistiks definition benyttes til operationaliseringen af socialklasse i bilag C.

Becker og Hecken (2009) har i overensstemmelse med andre studier fundet, at der er større sandsynlighed for, at personer fra en højere socialklasse vælger ekstra uddannelse efter en ungdomsuddannelse. Østergaard-Tygesen (2018) har også fundet at en høj socialklasse i udlandet er forbundet med en højere grad af tilknytning til uddannelsessystemet samt at unge fra mindre privilegerede hjem er mindre tilbøjelige til at påbegynde uddannelsesforløb. Becker og Hecken (2009) finder, at en væsentlig motivationsfaktor for uddannelsesbeslutninger på tværs af socialklasser er, at fastholde status ved minimum at have samme uddannelsesniveau som ens forældre. De forskellige socialklasser værdisætter muligheden for opadgående mobilitet og frygten for social degradering forskelligt. Arbejdsklassen er mere tilbøjelig til at fravælge ekstra uddannelse, da risiko for fiasko og omkostninger ved uddannelse er højere end for de højere socialklasser.

Forældres Alder

Flere undersøgelser finder, at børn af yngre mødre² klarer sig dårligere i grundskolen. En af forklaringerne hertil er, at forældrene, på grund af deres lave alder, ikke er klar til at give børnene den nødvendige støtte. En anden årsag er, at børnene hyppigere har en lav fødselsvægt og er født for tidligt. Der er fundet en kausal sammenhæng mellem lav fødselsvægt og præstationer i grundskolen (Jensen og Nielsen, 2010).

Søskende

¹Begrebet ungdomsuddannelser dækker både over erhvervsuddannelser og gymnasiale uddannelser (Undervisningsudvalget 2016-2017, 2017)

²Yngre mødre defineres af Jensen og Nielsen (2010) som kvinder der får et barn når de er under 24 år.

At være en del af en mindre familie er i empiriske studier blevet koblet sammen med bedre uddannelsesmæssige resultater. Elever med færre søskende, får mere opmærksomhed fra forældrene og har flere ressourcer tilgængelige end børn fra større familier. Ekstra opmærksomhed og støtte fra forældre medfører ifølge empiriske undersøgelser bedre faglige resultater (Benders, 2013). Lacour og Tissington (2011) finder også, at en dårlig faglig præstation er korreleret med mangel på ressourcer. Dette understøttes af, at Jensen og Nielsen (2010) finder, at personer med søskende har lavere sandsynlighed for at gennemføre en ungdomsuddannelse end enebørn. Sandsynligheden for at gennemføre en ungdomsuddannelse er ifølge Jensen og Nielsen (2010) lavest for individer med 2 eller flere søskende sammenlignet med et individ, der er enebarn.

4.2.2 Peer-Gruppe

Jaeger (2007) finder, at jo højere værdi de unge tillægger peer-gruppe relationer, jo mere sandsynligt er det, at de vælger den uddannelse som størstedelen af deres peer-gruppe vælger. I studiet er peer-gruppen defineret som personer i den samme grundskole, som individet. Jaeger (2007) finder, at de unge maksimerer både deres økonomiske og sociale afkast, når de foretager uddannelsesbeslutninger. Undersøgelsen finder dog, at det økonomiske afkast tillægges større vægt i beslutningen end det sociale afkast. Det er fundet at have negative effekter på uddannelsespræstation, når flere personer i ens peer-gruppe har karakterer i den lave ende af karakterskalaen. Peer-gruppen er her personer på samme gymnasiale uddannelse. En mulig forklaring af denne effekt er, at peer-gruppen forstyrrer læringsmiljøet, da de ikke lægger vægt på at klare sig godt i skolen (Eriksen, 2021).

4.2.3 Indkomst

I *schooling* modellen vælger personer det antal års uddannelse, der maksimerer nutidsværdien af deres livsindkomst (Borjas, 2016). Flere studier understøtter denne teori ved at finde, at individer er mere tilbøjelige til at vælge en uddannelse, hvor nutidsværdien af indkomsten er højere (Nielsen og Vissing-Jorgensen, 2006). Studier måler indkomsten ved de forskellige uddannelsesniveauer forskelligt. Jaeger (2007) måler, om der er en sammenhæng mellem det økonomiske afkast og personers uddannelsesbeslutninger ved at spørge individer, hvor stor vægt de lægger på indkomst, når de vælger uddannelse. Nielsen og Vissing-Jorgensen (2006) måler forskellen i indkomsten for personer, der har gennemført forskellige uddannelsesniveauer og estimerer forskellen i indkomsten ved de forskellige uddannelsesniveauer herudfra. Der bliver altså ikke taget højde for, at personer kan have forskellige evner. Jaeger (2007) finder en positiv sammenhæng mellem det gennemsnitlige afkast af uddannelse og det forventede antal års uddannelse i USA. *Choice* modellen fra Nielsen og Vissing-Jorgensen (2006) er baseret på dansk registerdata. Resultater herfra finder, at der er præference for uddannelse med en højere gennemsnitlig startindkomst og hvor der er lav risiko for udsving i indkomsten. Mens uddannelse, hvor der er en høj varians af permanente indkomstchok, fravælges. Nielsen og Vissing-Jorgensen (2006) opstiller et optimeringsproblem for individer, der er 19 år. Her antager de, at individerne tager den endelige uddannelsesbeslutning. Optimeringsproblemet sammenligner nutidsværdien af indkomsten ved de mulige uddannelsesmuligheder. Indkomstprocessen, der benyttes til at estimere den forventede indkomst, består af fem parametre: den gennemsnitlige log startløn, den gennemsnitlige vækstrate for lønnen indtil individerne er 40 år, den gennemsnitlige vækstrate for lønnen for når de er over 40 år samt variansen af midlertidige chok til lønnen og variansen af permanente chok til lønnen. Estimeringen tager højde for forskelle i karakterer og uddannelsestyper. Hartog og Diaz-Serrano (2007) finder, at indkomstrisiko har en negativ effekt på sandsynligheden for at forsætte på universitet i Spanien. Carneiro et al. (2003) finder, at indkomst spiller en lille rolle i valget mellem ungdomsuddannelse og universitetet i USA. Dette finder de, da valget mellem dem kun har en moderat påvirkning på livstidsnytte målt som forbrug. Deres model antager, at nytte er logget og tillader ikke, at personer udligner deres forbrug gennem deres livscyklus, som det ellers antages i neoklassisk teori.

Modsat de tidligere resultater finder en registerbaseret undersøgelse foretaget af DEA, at unge ikke tager valget mellem gymnasial uddannelse og erhvervsuddannelse ud fra deres ønsker om jobsikkerhed eller løn. Konklusionen fra undersøgelsen er, at unge er upåvirket af afkast og risiko, når de vælger

uddannelse (Nielsen, 2018). DEA måler i undersøgelsen alle i den erhvervsaktive³ alders forventede indkomst, differentieret på uddannelsesniveau. Risikoen måles i undersøgelsen som den andel, der er arbejdsløs indenfor de forskellige uddannelsesniveauer og variationen i indkomsten på de enkelte niveauer (Nielsen, 2018).

4.2.4 Evner

Ifølge den præsenterede teori i kapitel 3 har personer med et højere niveau af iboende evner et større marginalafkast af uddannelse. Derved vil personer med den samme diskonteringsrente gennemføre forskellige mængder uddannelse, hvis de ikke har samme niveau af evner. Personer med et højere niveau af evner har et større marginalafkast og vil dermed tage mere uddannelse end personer med et lavere niveau af evner. Adskillige studier finder resultater, der understøtter, at der er en positiv selvselektion til højere niveauer af uddannelse (Nielsen og Vissing-Jorgensen, 2006).

Evner er i den præsenterede teori defineret, som en persons iboende produktivetskapalet (Borjas, 2016). På grund af den brede definition er det svært at måle evner i praksis og der benyttes derfor proxies i empiriske studier. Karakterer i grundskolen, kan benyttes som proxy, da de til en vis grad er adgangsgivende til ungdomsuddannelser. Desuden er der en positiv sammenhæng mellem karakterer i matematik og dansk i grundskolen og sandsynligheden for at gennemføre en uddannelse (Produktivitetskommissionen, 2014). Karakterer i grundskolen afhænger af børnenes iboende evner, deres socioøkonomiske baggrund og skolens løftehøjde (Jensen og Nielsen, 2010). Koch et al. (2015) peger på, at empiriske studier ofte gør brug af en proxy eller antager, at *ability* er en enkelt, tidsinvariant faktor, der kan differentieres ud af paneldata. Jaeger (2007) har benyttet læsefærdigheder som proxy for evner og finder, at dette er en stærk indikator for uddannelsesvalg. Personer med højere testscore i læsning har en større sandsynlighed for at vælge gymnasial uddannelse relativt til ingen uddannelse og en lavere sandsynlighed for at vælge en erhvervsuddannelse.

4.2.5 Grundskole

I litteraturen findes der mange studier omkring, hvordan skolens ressourcer har en sammenhæng med udbyttet af uddannelse. Herunder har der været et særligt fokus på klassestørrelse og lærerens kvalitet (Koch et al., 2015). Krueger (2003) og Heinesen (2010) argumenterer for, at klassestørrelse har en væsentlig effekt på de uddannelsesmæssige resultater. Det samme gør Nielsen (2012), som finder evidens for, at større klasser mindsker læringen og derved sænker karaktergennemsnittet i klassen. Jensen og Nielsen (2010) finder, at målbare skolekarakteristika som klassestørrelse og lærernes aldersfordeling kun kan forklare en mindre del af hvor mange af skolens elever, der gennemfører en ungdomsuddannelse. Dette kan skyldes, at elever bliver påvirket af andre institutioner og får mere livserfaring, der har større indflydelse på, om de gennemfører en ungdomsuddannelse. Tilgængelighed har socioøkonomiske faktorer ifølge undersøgelsen over tid en mere konstant effekt på gennemførelsen af en ungdomsuddannelse. Forskningsresultater, der omhandler, hvorvidt klassestørrelse har en påvirkning på faglige og sociale færdigheder, er tvetydige (Nielsen, 2012).

I forhold til lærerens kvalitet er det svært at definere, hvad der karakteriserer "en god lærer". Dette vurderes på baggrund af, at formelle kvalifikationer, certifikater og erfaring ikke har en stor forklaringskraft (Koch et al., 2015). Slottved (2021) finder dog, at lærerkvalifikationer er en af de vigtigste faktorer til at løfte niveauet i grundskolen. Det er blandt andet vist, at eleverne opnår 0,2 karakter højere i eksamensresultater, hvis de har en lærer, der fik én karakter højere end gennemsnittet i gymnasiet (Produktivitetskommissionen, 2014). Et dansk studie fra 2009 finder, at antallet af undervisningstimer i matematik i 9.klasse har en betydning på elevernes præstation ved eksamen (Jensen og Nielsen, 2010). Derudover finder (Jensen og Nielsen, 2010), at andre skoleforhold end klassestørrelse og lærerens kvalitet har en mindre effekt på uddannelsesbeslutningen, herunder elevsammensætning, skolestørrelse og udgifter pr. elev. Resultaterne for disse variable er tvetydige. Dermed findes der eksempler på studier, hvori der

³Undersøgelsen inkluderer personer i den erhvervsaktive alder, der har en tæt tilknytning til arbejdsmarkedet. Det vil sige, at personer på overførselsindkomst ekskluderes (Nielsen, 2018).

ikke findes et konsistent forhold mellem skolernes tilgængelige ressourcer og de studerendes præstation (Hanushek, 2006).

Empiriske studier har ikke alene undersøgt skoleressourcernes påvirkning på uddannelsesbeslutningen. Østergaard-Tygesen (2018) har igennem interviews fundet, at unges oplevelse af grundskolen har en betydning for deres valg mellem gymnasial uddannelse og erhvervsuddannelse. Negative oplevelser som mobning og skoletræthed gør, at flere unge vælger en erhvervsuddannelse.

4.2.6 Bopæl

Afstand til uddannelsesinstitutioner kan ifølge empiriske studier have en indflydelse på uddannelsesbeslutninger. Afstanden fra forældres bopæl og forskellige uddannelsesinstitutioner benyttes i Nielsen og Vissing-Jørgensen (2006) som kontrolvariabel for økonomiske fordele ved at bo tæt ved ens forældre. At medtage afstanden til uddannelsesinstitution har en lille effekt på estimerne i deres model for valg af uddannelsestype. Jensen og Nielsen (2010) finder, at antallet af ungdomsuddannelser i nærområdet har en lille påvirkning på gennemførelsen af en ungdomsuddannelse. Derudover finder de, at kommunale forhold som ungdomsarbejdsløsheden også kun har en lille effekt på gennemførelse af en uddannelse. Der er en skæv fordeling af optaget på de videregående uddannelser opdelt efter kommunetype. I 2020 var 82% af de optagede på videregående uddannelser optaget på en uddannelse placeret i en storby, mens kun 1% var optaget på en uddannelse i en landkommune. Derved står de unge fra kommuner med få uddannelsespladser overfor valget mellem at flytte hjemmefra eller at blive i nærområdet, hvor der er færre uddannelsestyper at vælge imellem og hvor størstedelen af uddannelserne er på niveauet under videregående uddannelse (Hallmann og Carstensen, 2020; Thomsen og Bom, 2019). Forskellen i optagelsen på videregående uddannelser er der, til trods for, at der tages højde for indbyggertal i kommunerne. Studiet tager dog ikke højde for befolknings sammensætningen (Hallmann og Carstensen, 2020).

4.2.7 Bløde Faktorer

Der findes evidens for, at bløde værdier spiller en lige så vigtig rolle for investeringer i uddannelse som kognitive evner gør (Koch et al., 2015). Indenfor den økonomiske disciplin er det blevet mere normalt at inddrage et bredere sæt af værdier, men i mange tilfælde behandles de bløde faktorer dog stadig som en form for *black box*. Indenfor det adfærdøkonomiske felt har man dog forsøgt at undersøge, hvilke bløde faktorer der påvirker investering i uddannelse og effekten heraf (Koch et al., 2015). I litteraturreviewet i Koch et al. (2015) finder studier evidens for, at villighed til at konkurrere, problemer med selvkontrol i form af tidsinkonsistente præferencer, indre motivation og selvtillid alle er bløde faktorer, der har en sammenhæng med individers uddannelsesvalg.

Et valg om at tage en uddannelse involverer ifølge neoklassisk teori at bytte omkostninger, der forekommer lige nu, med fordele, der ligger længere fremme i tiden. I adfærdøkonomi findes det, at personer, der står overfor intertemporale trade-offs som dette, ofte udviser tidsinkonsistente præferencer (Koch et al., 2015). Flere studier undersøger forholdet mellem karakterer og tidspræferencer, eksempelvis finder Golsteyn et al. (2014) en negativ sammenhæng mellem utålmodighed og skolepræstation (Koch et al., 2015). Tidspræferencer kan ifølge Oreopoulos (2007) også hjælpe med at forklare store uddannelsesmæssige beslutninger, såsom at droppe ud. Generelt set findes der evidens, der sammenkæder utålmodighed og problemer med selvkontrol til uddannelsesmæssige beslutninger og præstationer (Koch et al., 2015).

Bénabou og Tirole (2002) finder, at et over-positivt syn på egne evner kan være en vigtig motivationsfaktor, fordi evner og indsats er komplementære faktorer i uddannelsesmæssig præstation (Koch et al., 2015). Et tysk tværsnitsstudie fra Becker og Hecken (2009) finder, at der er korrelation mellem en række bløde faktorer og unges socioøkonomiske baggrund. Studiet finder blandt andet, at den socialklasse forældrene tilhører har en sammenhæng med de unges selvvalgte evner i skolen, deres opfattelse af omkostninger og fordele ved ekstra uddannelse. Studiet finder, at disse elementer er af afgørende betydning for investeringsbeslutningen i uddannelse.

Bløde faktorer kan ud fra studierne resultater være med til at øge forståelsen for kompleksiteten, der ligger i at træffe et uddannelsesvalg (Koch et al., 2015). Ikke desto mindre er mange af de parametre, som studierne peger på, svære at operationalisere i praksis. Bløde faktorer måles ofte subjektivt, da det

er svært at finde objektive mål, der måler værdierne. De selvangivne mål kan være biased, da personer ikke nødvendigvis giver korrekte oplysninger. De kan eksempelvis ubevist overvurdere deres selvkontrol, da det er mere social acceptabelt at have en høj selvkontrol (Wooldridge, 2010).

4.2.8 Køn

Ifølge Duckworth og Seligman (2006) er der forskel i piger og drenges akademiske præstation, hvor der er en tendens til, at piger får bedre karakterer og drenge klarer sig bedre i standardiserede tests. I en dansk kontekst svarer det til, at piger får bedre karakterer både i standpunktskarakterer og til eksaminer, mens drenge klarer sig bedre i standardiserede tests, som PISA eller IQ-tests. Ifølge Koch et al. (2015) kan dette skyldes en forskel i reaktionen på konkurrencepres, da der findes evidens for, at der er en forskel i resultaterne blandt køn ved eksempelvis en turnering, men ikke i en neutral situation. Drenge synes at klare sig bedre i konkurrencebetonede situationer end piger. I det omfang, at uddannelsesmæssige aktiviteter og uddannelsesprogrammer opfattes som konkurrenceprægede miljøer, kan kønsforskelle i konkurrenceevner eller -villighed være en vigtig faktor bag kønsforskelle i uddannelsesvalg (Koch et al., 2015).

Jensen og Nielsen (2010) finder, at drenge har en lavere sandsynlighed for at gennemføre en gymnasial uddannelse. Jaeger (2007) finder, at mænd har en større sandsynlighed for at tage en erhvervsuddannelse.

4.2.9 Etnicitet

Jaeger (2007) finder, at personer med ikke-dansk etnicitet er mindre tilbøjelige til at tage en erhvervsuddannelse end personer med dansk etnicitet. Til gengæld har personer med anden etnisk baggrund ifølge Jensen og Nielsen (2010) en større sandsynlighed for at gennemføre en ungdomsuddannelse, når der tages højde for socioøkonomiske baggrundsvariable. DØRS (2003) finder, at forskellen i andelen med en gymnasial uddannelse alt efter etnicitet kan forklares af socioøkonomiske faktorer, men at forskelle i disse faktorer ikke kan forklare forskellen i andelen, der har en erhvervsuddannelse. Dermed er der flere tvetydige resultater for unges etnicitets påvirkning alt efter hvilke baggrundsvariable, der tages højde for og hvilket uddannelsesniveau, der undersøges.

4.2.10 Alder

Jensen og Nielsen (2010) finder en sammenhæng mellem de unges alder ved afslutningen af grundskolens 9. klasse og deres præstation til afgangsprøverne i Danmark. De elever, der er ældre end 16 år ved afgangsprøverne, klarer sig generelt lidt bedre sammenlignet med de elever, der var under 16 år ved afgangsprøverne. En forskel i præstation på grund af alder kan være en indikator på forskel i modenhed eller læringsparathed.

4.3 Opsummering

Sammenfattende er der i empirisk sammenhæng en bred enighed om, at forældres socioøkonomiske faktorer har en sammenhæng med de unges uddannelsesvalg. Specifikt finder nogle studier, at der er en positiv sammenhæng mellem forældres indkomst og uddannelsesniveau samt unges valg af uddannelsesniveau. Adskillige studier finder, at evner er en forklarende faktor for uddannelsesvalg. Evner har ifølge studierne en positiv påvirkning på højeste fuldførte uddannelsesniveau. Derudover findes det, at en række af kontrolvariablene som køn, etnicitet og alder har en sammenhæng med individets uddannelsesbeslutning. Derfor vurderes det nødvendigt at tage højde for disse.

På baggrund af den præsterede økonomiske teori fra forrige kapitel og de fundne sammenhænge i dette kapitel opstilles i næste kapitel to hypotesegrupper.

Hypoteser 5

På baggrund af de metodiske overvejelser, der fremgår af kapitel 2, samt de teoretiske og empiriske resultater fra henholdsvis kapitel 3 og kapitel 4, opstilles i nærværende kapitel to grupper af hypoteser. Formålet med opstillingen af hypoteserne er at formulere nogle relaterbare og konkrete arbejdsspørgsmål, der kan undersøges empirisk i en kvantitativ analyse. Undersøgelsen af hypoteserne skal i sammenhæng med teorien bidrage til besvarelsen af problemformuleringen.

Da hypoteserne opstilles med henblik på at undersøge disse empirisk, findes det relevant efterfølgende at betragte dataet, der ligger til grund for undersøgelsen heraf. Dele af dataet betragtes ved brug af en deskriptiv tilgang, hvilket gør det muligt at undersøge, hvor godt dataets, og dermed kohortens, karakteristika stemmer overens med hele befolkningen.

5.1 Opstilling af Hypoteser

I indeværende afsnit opstilles to hypotesegrupper, der hver består af fem hypoteser. Hypoteserne benyttes til at undersøge, hvorvidt nogle af de faktorer, der fremgår af litteraturstudiet samt den økonomiske teori, har en signifikant sammenhæng med niveauet af unges højeste fuldførte uddannelse.

I det empiriske litteraturstudie undersøges individers uddannelsesbeslutning i et bredt perspektiv, jf. kapitel 4. Af kapitlet fremgår resultater fra studier, der undersøger uddannelsespræstation, specifikke uddannelsesvalg, uddannelsesbeslutningen i et internationalt perspektiv, *educational expectations* samt uddannelse på et horisontalt niveau. Resultaterne fra de empiriske studier kan derfor ikke benyttes direkte i besvarelsen af problemformuleringen, men de kan bidrage til at forstå hvilke faktorer, der påvirker uddannelsesbeslutninger generelt. Endvidere kan de bidrage med forventninger til hvilke faktorer, der påvirker valget af uddannelsesniveau generelt og hvordan.

Ifølge neoklassisk økonomisk teori er det individers evner og diskonteringsrente, der påvirker deres uddannelsesbeslutning. Det teoretiske grundlag tager udgangspunkt i hele uddannelsessystemet, mens størstedelen af de empiriske studier i kapitel 4 enten fokuserer på valget mellem gymnasial uddannelse og erhvervsuddannelse eller mellem ingen videregående uddannelse og videregående uddannelse. Derfor findes det interessant at undersøge, om det er de samme faktorer, der påvirker valget mellem alle niveauer af uddannelse, som der påvirker valget mellem eksempelvis en gymnasial uddannelse og en erhvervsuddannelse. De studier fra litteraturstudiet, der undersøger videregående uddannelse, er udenlandske studier. I udlandet er uddannelsessystemet dog ikke opbygget på samme måde, som i Danmark. Derfor findes det interessant at undersøge dette nærmere i en dansk kontekst.

Der eksisterer som nævnt empirisk belæg for, at der er en sammenhæng mellem flere økonomiske og socioøkonomiske faktorer og unges uddannelsesvalg, jf. kapitel 4. Det er dog ikke alle faktorer, hvor der teoretisk eller empirisk findes en ensrettet sammenhæng med uddannelsesniveau. I det videre arbejde benyttes begrebet *forklarende faktorer* om de faktorer, hvor der teoretisk findes en klar sammenhæng mellem den pågældende faktor og uddannelsesniveau. Det er kun disse faktorer, hvor det, på baggrund af både litteraturstudiet og teorien, er muligt at opstille klare forventninger til sammenhængen med uddannelsesniveau, herunder retningen heraf. Endvidere defineres en faktor kun som forklarende, hvis den i praksis er operationaliserbar på en måde, der gør det muligt at sammenligne den med operationaliseringen af faktoren i den empiriske litteratur. Under denne definition er de forklarende faktorer som følger: unges evner, deres mødres og fædres indkomst samt uddannelsesniveau. Som det

fremgår af kapitel 3.3 kan socioøkonomiske faktorer benyttes som en proxy for diskonteringsrenten. Derfor findes der teoretisk en sammenhæng mellem de socioøkonomiske faktorer, forældres indkomst og uddannelsesniveau, og unges uddannelsesniveau.

Information, om hvorfor det netop er de forklarende faktorer, der har særlig interesse, fremgår efter en præsentation af hypotesegruppe 1. Hypotesegruppe 1 består af fem hypoteser, 1a-1e, en for hver af de forklarende faktorer.

Hypotesegruppe 1

Hypotese 1a

Der eksisterer en positiv sammenhæng mellem unges evner og deres højeste fuldførte uddannelsesniveau.

Hypotese 1b

Der eksisterer en positiv sammenhæng mellem mødres indkomst og unges højeste fuldførte uddannelsesniveau.

Hypotese 1c

Der eksisterer en positiv sammenhæng mellem mødres uddannelsesniveau og unges højeste fuldførte uddannelsesniveau.

Hypotese 1d

Der eksisterer en positiv sammenhæng mellem fædres indkomst og unges højeste fuldførte uddannelsesniveau.

Hypotese 1e

Der eksisterer en positiv sammenhæng mellem fædres uddannelsesniveau og unges højeste fuldførte uddannelsesniveau.

De forklarende faktorer er af særlig interesse, da problemformuleringen er; hvilke faktorer der påvirker unges valg af uddannelse, hvor påvirker defineres, som at der eksisterer en kausal sammenhæng. I den kvantitative analyse er det muligt at undersøge sammenhænge men ikke *kausale* sammenhænge. For at kunne sandsynliggøre, hvorvidt der eksisterer en kausal sammenhæng, er det dermed ikke nok, at der eksisterer en signifikant effekt i en regressionsmodel, men det er nødvendigt at argumentere for kausalitet ud fra et teoretisk perspektiv. Dermed er det kun disse faktorer, der kan benyttes i besvarelsen af problemformuleringen.

Det fremgår ovenfor at de studier, der undersøger valget om videregående uddannelse, er udenlandske studier. Det findes derfor interessant at undersøge, hvorvidt der eksisterer de samme sammenhænge, når man kun undersøger sammenhængen for unge, der har fuldført en videregående uddannelse. På baggrund af dette opstilles hypotesegruppe 2, der består af fem hypoteser.

Hypotesegruppe 2

Hypotese 2a

Der eksisterer en positiv sammenhæng mellem unges evner og deres højeste fuldførte uddannelsesniveau, også når der kun ses på unge, som har fuldført en videregående uddannelse.

Hypotese 2b

Der eksisterer en positiv sammenhæng mellem mødres indkomst og unges højeste fuldførte uddannelsesniveau, også når der kun ses på unge, som har fuldført en videregående uddannelse.

Hypotese 2c

Der eksisterer en positiv sammenhæng mellem mødres uddannelsesniveau og unges højeste fuldførte uddannelsesniveau, også når der kun ses på unge, som har fuldført en videregående uddannelse.

Hypotese 2d

Der eksisterer en positiv sammenhæng mellem fædres indkomst og unges højeste fuldførte uddannelsesniveau, også når der kun ses på unge, som har fuldført en videregående uddannelse.

Hypotese 2e

Der eksisterer en positiv sammenhæng mellem fædres uddannelsesniveau og unges højeste fuldførte uddannelsesniveau, også når der kun ses på unge, som har fuldført en videregående uddannelse.

Hypoteserne i hypotesegruppe 2 undersøges i forlængelse af sammenhængene, der undersøges i hypotese 1a-1e. De resterende operationaliserbare faktorer fra kapitel 3 og 4 benyttes i den videre analyse som kontrolvariable. Det har for nogle af kontrolvariablene ikke været muligt at opstille klare forventninger, fordi der empirisk ikke er fundet en klar sammenhæng mellem faktoren og uddannelsesniveau. I nogle tilfælde har det ikke været muligt at operationalisere faktoren på en måde, der gør den tilstrækkeligt sammenlignelig med teorien og resultaterne fra litteraturstudiet. Det findes dog stadig relevant, at inddrage variablene i den kvantitative analyse, da de kan have en sammenhæng med valg af uddannelsesniveau. Det er endvidere nødvendigt at inddrage dem, da en eventuel effekt heraf i praksis ellers vil blive opfanget af de andre variable eller i fejleddet (Wooldridge, 2016). Derudover er variable som alder, køn samt etnicitet inkluderet i regressionsmodellerne, så det er muligt at kontrollere for disse. Kontrolvariablene er ikke af primær interesse, da det ikke er muligt at vurdere, hvorvidt der muligt at sandsynliggøre en kausal sammenhæng mellem dem og unges valg af uddannelsesniveau.

5.2 Deskriptiv Statistik

I dette afsnit betragtes dataet, der benyttes i den kvantitative analyse, ved brug af deskriptiv statistik. Først redegøres der for datasættet, der benyttes, hvorefter fordelingen af højeste fuldførte uddannelse betragtes på tværs af centrale kontrolvariable som køn, etnicitet og kommune. Populationens karakteristika sammenholdes med informationer fra Danmarks Statistik og oplysninger fra Undervisningsudvalget for 2016-2017 og Ministiet for Ligestillingen for at sikre, at der ikke er betydelige afvigelser i populationen.

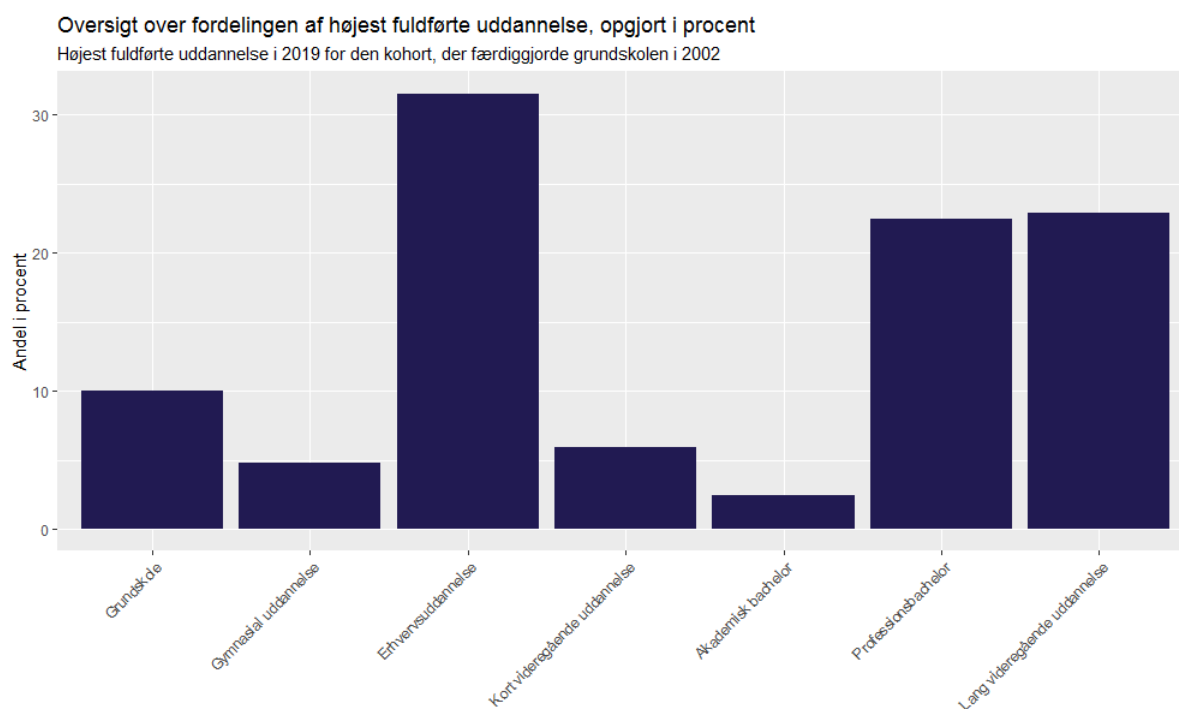
Beskrivelse af Dataet

Til udarbejdelsen af den kvantitative analyse, tages der, som nævnt i kapitel 2.3, udgangspunkt i data fra Danmarks Statistiks Forskerservice. Det endelige datasæt indeholder 39.090 observationer. Information om fordelingen af observationer i det endelige datasæt fremgår af bilag A.10.

Højeste Fuldførte Uddannelsesniveau

Af figur 5.1 fremgår en oversigt over fordelingen af højeste fuldførte uddannelse for populationen. Den største enkeltstående gruppe er personer med en erhvervsuddannelse, hvor 31,5% af personerne har en erhvervsuddannelse som højeste fuldførte uddannelsesniveau. Dette er også den største uddannelsesgruppe

Figur 5.1: Højeste Fuldførte Uddannelsesniveau



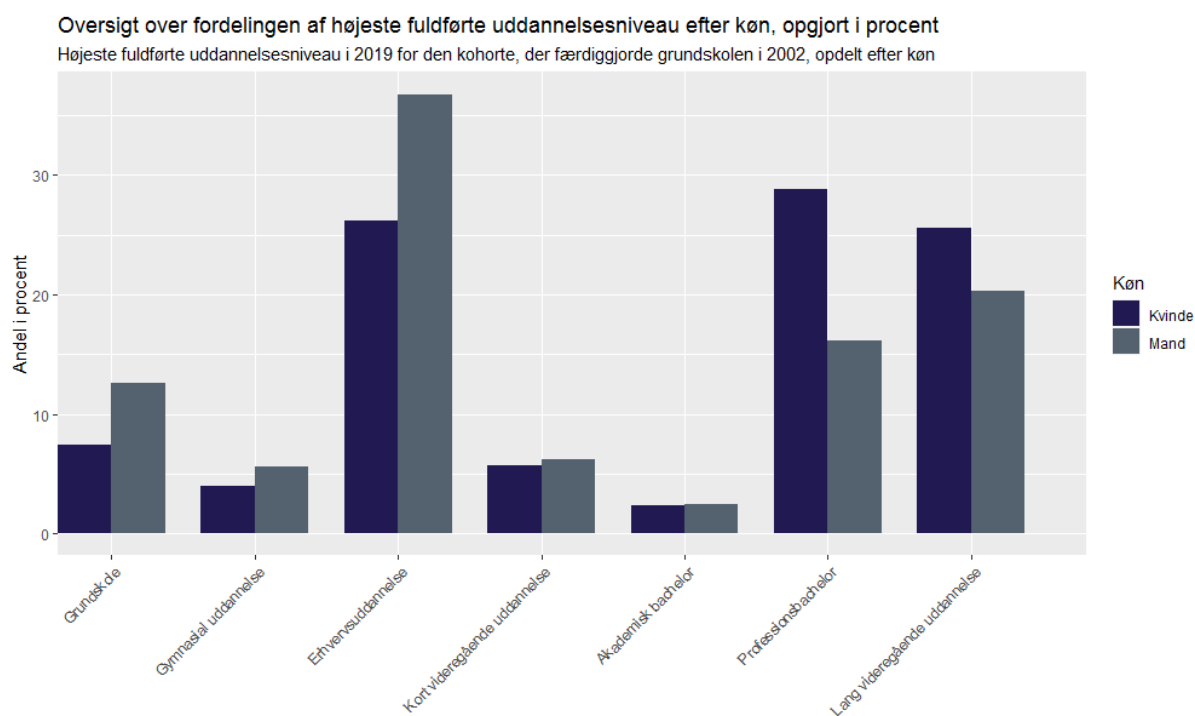
Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice, jf. tabel B.1 i bilag B

for alle i den erhvervsaktive alder (Danmarks Statistik, n.d.k). Niveauerne lang videregående uddannelse og professionsbachelor er de næst største enkeltstående uddannelsesniveauer. Henholdsvis 22,9% og 22,4% af personerne i populationen har en lang videregående uddannelse eller en professionsbachelor. Den mindste gruppe i populationen er personer, hvis højeste fuldførte uddannelsesniveau er en akademisk bachelor. Denne gruppe udgør 2,4%. En forklaringen på at, der er så få personer med en akademisk bachelor kan være, at størstedelen forsætter direkte videre til en kandidatuddannelse og derved får en lang videregående uddannelse, da de færreste akademiske bacheloruddannelser er erhvervskompetencegivende (Danmarks Statistik, 2017; DØRS, 2003). Samlet har 53,7% af populationen en videregående uddannelse som højeste fuldførte uddannelsesniveau. Dette er en del over niveauet for alle i den erhvervsaktive alder, hvor ca. 34,0% af befolkningen i 2019 havde en videregående uddannelse (Danmarks Statistik, n.d.k).

Højeste Fuldførte Uddannelsesniveau på Tværs af Køn

I figur 5.2 differentieres de unges højeste fuldførte uddannelsesniveau på køn. Empiriske resultater fra litteraturstudiet indikerer, at der er forskel på drenge og pigers akademiske præstation, hvilket antages at have betydning for valget af uddannelsesniveau. I grafen er det tydeligt, at der er en forskel i højeste fuldførte uddannelsesniveau fordelt på køn. Af figuren fremgår det, at 36,8% af mændene har en erhvervsuddannelse som højeste fuldførte uddannelsesniveau, hvor det kun er 26,2% af kvinderne. Endvidere fremgår det, at 12,6% af mænd har en grundskole som højeste fuldførte uddannelse, mens den tilsvarende andel er på 7,4% for kvinderne. Dette stemmer overens med at Undervisningsudvalget 2016-2017 (2017) har fundet, at der er en overvægt af drenge med grundskolen eller mindre, som højeste fuldførte uddannelsesniveau. Der er derimod en større andel af kvinderne, der har en professionsbachelor, 28,8%, eller lang videregående uddannelse, 25,5%. For mændene er det henholdsvis 16,1% samt 20,3%, jf. figur 5.2. Denne fordeling stemmer overens med data fra 2004, hvor der var en overvægt af kvinder på professionsbachelor- og lange videregående uddannelser (Minister for Ligestilling, 2006). Fokuseres der på videregående uddannelse generelt, så har 62,4% af kvinderne en videregående uddannelse, mens det er 45,1% for mændene. Den store forskel i kønnene kan skyldes, at de videregående uddannelser oftest ligger i de større byer, hvor der er en overvægt af kvinder (Danmarks Statistik, n.d.g). Overvægten af

Figur 5.2: Højeste Fuldførte Uddannelsesniveau Opdelt efter Køn



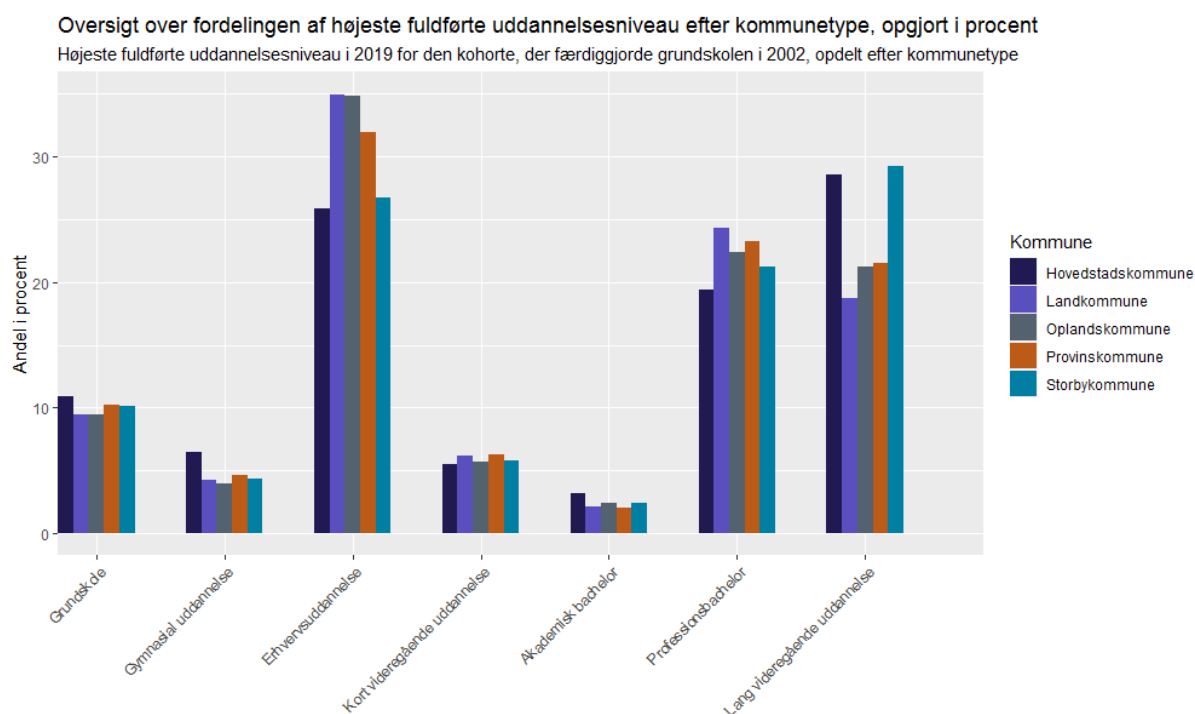
Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice, jf. tabel B.2 i bilag B

kvinder kan forklares af, at kvinder har en tendens til at flytte fra yderområder, mens mænd i højere grad vælger at blive boende i udkantsområderne (Nielsen og Faber, 2016).

Højeste Fuldførte Uddannelsesniveau på Tværs af Kommunetype

Det fremgår af figur 5.3 at der er forskel på højeste gennemførte uddannelsesniveau, hvis der kontrolleres for, hvilken kommunetype den unge boede i, da vedkommende færdiggjorde grundskolen. Kommunetyperne defineres og forklares nærmere i bilag A. Det fremgår tydeligt af figuren, at der er geografiske forskelle for højeste fuldførte uddannelsesniveau. De største forskelle observeres ved erhvervsuddannelse samt lang videregående uddannelse. En mindre andel af personerne fra hovedstadskommuner samt storbykommuner har en erhvervsuddannelse som højeste fuldførte uddannelsesniveau, omvendt har en større andel en lang videregående uddannelse. De geografiske forskelle i højeste fuldførte uddannelsesniveau kan skyldes, at unge tager uddannelser, der ligger i deres nærområde (Thomsen og Bom, 2019). I storbyerne og hovedstadskommunerne er der nemmere adgang til mellemlange- og lange videregående uddannelser. Samtidig er der en lidt højere andel af personer fra hovedstadskommuner, der ikke har en erhvervskompetencegivende uddannelse. Ca. 18,7% har en ikke-erhvervskompetencegivende uddannelse som højeste fuldførte uddannelse, sammenlignet med hele populationen er det ca. 15,5%. Den højere andel kan skyldes, at der er en større andel af personer med anden etnisk oprindelse end dansk, der er bosat i hovedstadskommunerne og blandt dem er der en højere andel uden erhvervskompetencegivende uddannelse (Danmarks Statistik, 2019b).

Figur 5.3: Højeste Fuldførte Uddannelsesniveau Opdelt efter Kommune

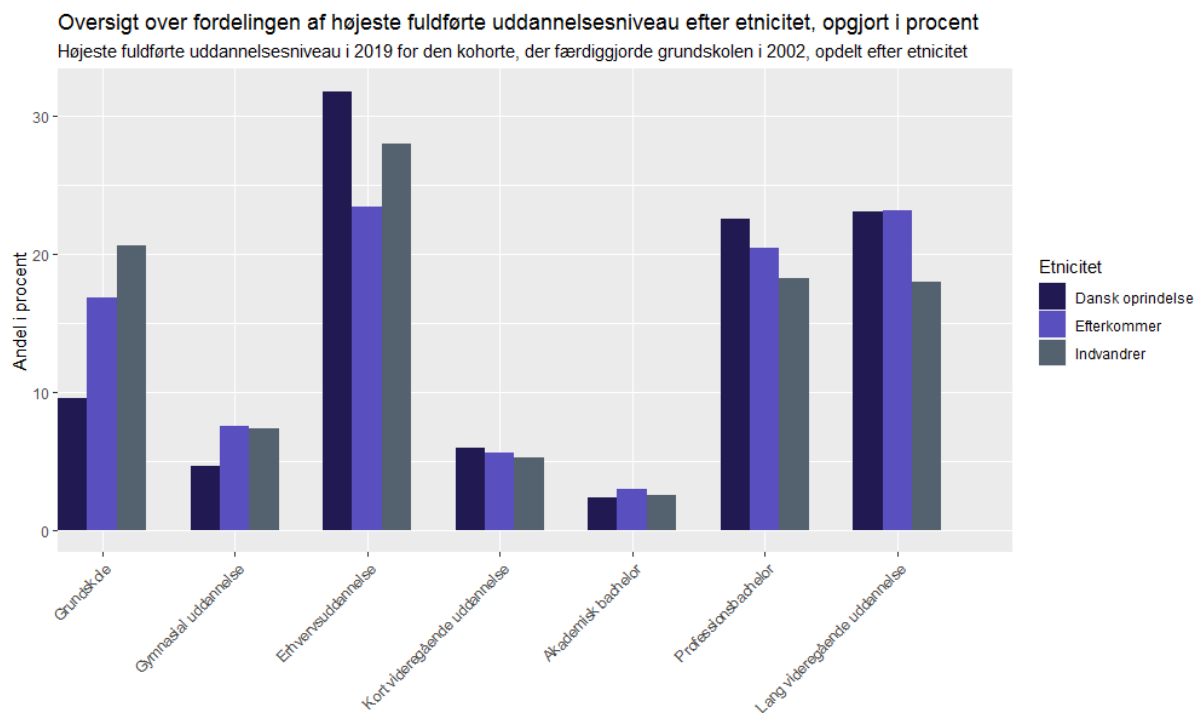


Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice, jf. tabel B.3 i bilag B

Højeste Fuldførte Uddannelsesniveau på Tværs af Etnicitet

Det fremgår tydeligt af figur 5.4, at der også er en forskel på personernes højeste fuldførte uddannelsesniveau på tværs af etnicitet. Der er en større andel af indvandrere og efterkommere, hvis højeste fuldførte uddannelse er en grundskole eller en gymnasial uddannelse, jf. figur 5.4. For indvandrere er det ca. 28%, der ikke har taget yderligere uddannelse. 24,4% af efterkommerne har en ikke-erhvervskompetencegivende uddannelse, hvilket er 10 procentpoint mere end for personer med dansk oprindelse, hvor andelen er 14,2%. Dette stemmer overens med at Undervisningsudvalget 2016-2017 (2017) har fundet, at der er en overvægt af indvandrere og efterkommere med grundskolen eller mindre, som højeste fuldførte uddannelsesniveau. Der ses endvidere en forskel i andelen med en erhvervsuddannelse som højeste fuldførte uddannelse fordelt på etnicitet. 31,8% af personer af dansk oprindelse har en erhvervsuddannelse som højeste fuldførte uddannelse, hvorimod det kun er 23,4% af efterkommere og 28,0% af indvandrere. Derudover har en lavere andel af indvandrere en professionsbachelor og lang videregående uddannelse som højeste fuldførte uddannelse, jf. figur 5.4. Det er ca. samme andel af personer med dansk oprindelse, indvandrere og efterkommere, der har en kort videregående uddannelse og en akademisk bachelor som højeste fuldførte uddannelse.

Figur 5.4: Højeste Fuldførte Uddannelsesniveau Opdelt efter Etnicitet



Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice, jf. tabel B.4 i bilag B

5.3 Opsummering

I kapitlet opstilles to hypotesegrupper, hver med fem hypoteser. Hypoteserne omhandler de forklarende faktorer. De forklarende faktorer er unges evner, deres forældres indkomst samt deres forældres uddannelsesniveau. Endvidere beskrives datagrundlaget for den kvantitative analyse, hvori hypoteserne undersøges. Herudfra er det klart, at kvinder i populationen har et højere uddannelsesniveau end mænd. Derudover er der en tendens til at unge fra storbyerne og hovedstaden har lange videregående uddannelser, hvorimod unge fra land-, oplands- samt provinskommuner i højere grad er repræsenteret ved erhvervsuddannelse. I forhold til etniske forskelle, indikerer populationen at en større andel af indvandrere og efterkommere har en ikke-kompetencegivende uddannelse, sammenlignet med unge af dansk oprindelse. I det følgende kapitel redegøres der for overvejelserne, der ligger til grund for valget af metode til undersøgelse af hypoteserne.

Økonometrisk Metode 6

I nærværende kapitel redegøres der for de metodiske overvejelser, der ligger til grund for valget af modeltype og generelt i forbindelse med udarbejdelsen af den kvantitative analyse. Det findes relevant at præsentere de teoretiske overvejelser, da valget af modeltype har en betydning for udarbejdelsen og fortolkningen af resultaterne.

6.1 Paneldata og Tværsnitsdata

For at undersøge om de forklarende faktorer empirisk har en positiv sammenhæng med unges valg af uddannelsesniveau, udarbejdes en regressionsanalyse. Det tilgængelige data fra Danmarks Statistiks Forskerservice er paneldata. Paneldata indeholder information om individers adfærd på tværs af individer men også over tid. I nærværende tilfælde undersøges det, hvordan individers valg af uddannelsesniveau varierer. Det antages generelt for paneldata, at der eksisterer korrelation over tid for et givet individ, men at der ikke eksisterer korrelation individer imellem. Dette betyder eksempelvis, at indkomsten for et individ er korreleret over tid, men er uafhængigt på tværs af individer.

I nærværende tilfælde behandles paneldataet som tværsnitsdata. Årsagen til dette er, at der ikke er behov for at følge individerne kontinuerligt over tid, men at der anvendes variable, som måles på forskellige tidspunkter. Yderligere overvejelser omkring det konkrete dataarbejde, herunder de udvalgte år, fremgår af kapitel 2 samt bilag A.

6.2 Modelovervejelser

I udarbejdelsen af regressionsmodellerne, er flere forskellige modeltyper overvejet. På baggrund af rangeringen af uddannelsesniveauer, er den afhængige variabel en diskret variabel, hvilket betyder, at den kan antage et endeligt antal værdier (Wooldridge, 2016). Da den afhængige variabel er diskret, er muligheden for at benytte en multinomial logistisk model eller en ordered logistisk model undersøgt. Begge disse modeltyper behandler den afhængige variabel som en diskret variabel. En ordered logistisk regression kaldes også *ordinal logistisk regression*, *ordered response model* eller *proportional odds logistic regression* (McNulty, 2021). Både en multinomial- og en ordred logistisk regression gør brug af *maximum likelihood* estimering, men er forskellige på andre punkter. Forskellen ligger blandt andet i, hvordan den afhængige variabel, y , behandles. Som beskrevet ovenfor består den afhængige variabel af et endeligt antal værdier i både en multinomial og ordered logistisk regression. Hver enkelt af disse værdier repræsenterer en bestemt kategori, j . I en multinomial logistisk regression har de numeriske værdier ingen betydning, det har de derimod i en ordered logistisk regression (Parry, 2016). Det kan eksempelvis være når en bog vurderes på en skala fra et til fem. Et mere relevant eksempel kan være uddannelsesniveau på en skala fra et til syv, hvor $y = 7$ er den længste uddannelse, $y = 1$ er grundskole og $y = 2$ til $y = 6$ er uddannelseslængder derimellem. Det har dermed en betydning at syv er et højere tal end seks, selvom at tallet i sig selv kun har en betydning for rangeringen. Det betyder eksempelvis, at det ikke er muligt at konkludere, hvorvidt forskellen mellem fire og to er dobbelt så stor eller vigtig som forskellen mellem et og to (Wooldridge, 2010).

En ordered logistisk regression er en udvidelse af en logistisk regression, hvor den afhængige variabel er binær og hvor logaritmen deraf har et lineært forhold til de uafhængige variable. I en logistisk regression er der én regressionsfunktion. Hvis den afhængige variabel i stedet har j kategorier, så vil der være $j - 1$ logistiske funktioner (Parry, 2016). I en model med 7 kategorier i den afhængige variabel,

korresponderende til antallet af uddannelsesniveauer, vil der være syv kategorier ($j = 7$), hvorfor der vil være $j - 1 = 6$ funktioner ved brug af en ordered logistisk regression. At estimere en ordered logistisk regression med statistisk software er forholdsvis simpelt, men fortolkningen deraf kan være vanskelig. Fortolkningen bliver sværere jo flere kategorier, der er i den afhængige variabel som følge af, at antallet af funktioner, der skal fortolkes på, stiger (Parry, 2016). Modellens konstruktion betyder endvidere, at der fortolkes på kumulative sandsynligheder. Dette betyder, at man undersøger sandsynligheden for at y er mindre end eller lig med j , altså $P(Y \leq j)$. En formel gennemgang af en ordered logistisk regression fremgår af bilag D.

Udover den numeriske betydning af den afhængige variabel, har en ordered logistisk regression også en bestemt underliggende antagelse, som skal overholdes. Denne antagelse kaldes *proportional odds assumption* eller den parallelle regressionsantagelse. Antagelsen består i, at koefficienterne, der beskriver sandsynligheden for at tilhøre den laveste kategori i forhold til alle højere kategorier af den afhængige variabel, er den samme som de, der beskriver sandsynligheden for at tilhøre den næst-laveste kategori og alle højere kategorier og så fremdeles (Greene, 2018; Wooldridge, 2010). Dette betyder, at outputtet af en ordered logistisk regression indeholder et skæringspunkt for hvert niveau af den afhængige variabel, med undtagelse af en, og én enkelt hældning for hver af de uafhængige variable (Parry, 2016). I tilfælde af at denne antagelse ikke overholdes, betyder det, at effekten af en eller flere af de uafhængige variable varierer signifikant mellem kategorier (McNulty, 2021). Antagelsen er en konsekvens af modellens opbygning og det er muligt at teste, om den overholdes ved brug af en Brant-test (Greene, 2018; Wooldridge, 2010). De statistiske tests, der kan udføres i forskellige software programmer, eksempelvis Brant-testen, er kritiseret for at have en tendens til at forkaste nulhypotesen og dermed indikere, at proportional odds assumption ikke overholdes (Harrell Jr, 2001; UCLA: Statistical Consulting Group, n.d.a). Som følge af kritikken af de statistiske tests, findes det relevant at undersøge, hvorvidt modellerne overholder proportional odds assumption på anden vis. I sammenhæng hermed er det relevant at nævne, at proportional odds assumption i mange tilfælde generelt ikke overholdes (Harrell Jr, 2001; McNulty, 2021; UCLA: Statistical Consulting Group, n.d.a). Denne tendens eksisterer i høj grad i mere komplekse modeller. Med komplekse modeller menes der modeller, som inkluderer flere uafhængige variable, herunder både kategoriske variable, dummy variable og kontinuerlige variable. Long (1997) og Long og Freese (2006) skriver, at det ikke er unormalt at antagelsen ikke overholdes. I litteraturen argumenteres der også for, at antagelsen ofte ikke overholdes ved store stikprøvestørrelser (O'Connell, 2006). Det vurderes derfor, at der er en betydelig sandsynlighed for, at den heller ikke overholdes i nærværende tilfælde.

I tilfælde af at proportional odds assumption ikke overholdes, er det muligt at benytte en multinomial logistisk regression. Ifølge Glen (2014) skal man dog være opmærksom på, at man risikerer at overestimere antallet af parametre, hvis man ignorerer den rangerede struktur i dataet. Dermed kan man risikere at overse statistisk signifikante resultater (Glen, 2014). En multinomial logistisk regression, synes derfor heller ikke at være problemfri.

Som nævnt tidligere kan det være svært at fortolke en ordered logistisk regression, da den består af $j - 1$ funktioner. Denne udfordring, der i høj grad er relevant i nærværende tilfælde, imødekommes ikke ved brug af en multinomial logistisk regression, da denne type model estimerer samme antal funktioner, som en ordered logistisk regression. I tilfælde af at man arbejder med flere end 5 kategorier i den afhængige variabel, foreslår Glen (2014), at man behandler den afhængige variabel som en kontinuerlig variabel. Dette vil muliggøre en estimering ved brug af *ordinary least squares* (OLS). OLS er en af de mest anvendte metoder til estimering af regressionsmodeller og er betydeligt lettere at fortolke (Glen, 2014; Wooldridge, 2010). Indenfor moderne mikroøkonometri er det accepteret, at en OLS-regression fungerer som en god approksimering for en regression med en diskret afhængig variabel. Der foreligger et argument om, at en ikke-lineær estimering sandsynligvis følger *conditional expectation function* bedre end en lineær model, men at det i forhold til marginale effekter har en lille betydning (Angrist og Pischke, 2008).

Flere forskellige kilder peger på, at vurderingen af hvilken model der er den optimale i de fleste tilfælde, er en subjektiv og ikke mindst udfordrende vurdering. Årsagen til dette skal findes i at grænserne mellem dem er slørede samt at der er fordele og ulemper ved alle modeltyper (Glen, 2014; Robitzsch, 2020). I mange tilfælde er der dermed ikke én type model, der entydigt er bedre end de andre modeller, hvorfor en detaljeret vurdering af den enkelte situation er nødvendig.

Resultat af overvejelserne

I nærværende tilfælde vurderes det at være mest optimalt at gøre brug af en ordered logistisk regression. Denne modeltype er valgt, da rangeringen af uddannelsesniveauer vurderes at være essentiel for undersøgelsen af hypoteserne og ultimativt besvarelsen af problemformuleringen. At kunne behandle den afhængige variabel som ordinal, er derfor en central del af denne vurdering. Flere af studierne, der inddrages i litteraturstudiet, benytter også en ordered logistisk regression til at undersøge uddannelsesvalg (Jaeger, 2007; Nielsen og Vissing-Jorgensen, 2006).

Som det fremgår ovenfor, er det vanskeligt at fortolke en ordered logistisk regression, hvis den afhængige variabel har mange kategorier (Parry, 2016). Denne egenskab vanskeliggør arbejdet med modeltypen i nærværende tilfælde, hvor uddannelsesniveauet er inddelt i 7 niveauer, jf. kapitel 2.4.1. Egenskaben vurderes at vanskeliggøre fortolkningen på et sådant plan, at resultaterne af den kvantitative analyse ikke kan benyttes til at undersøge hypoteserne og besvare problemstillingen (Greene, 2018; Wooldridge, 2010). Det er derfor nødvendigt, at finde et kompromis, hvor antallet af kategorier i den afhængige variabel mindskes, men hvor modellens evne til at behandle den komplekse problemstilling bibeholdes på bedste vis. Den mest hensigtsmæssige måde at imødekomme udfordringen på, vurderes i dette tilfælde at være en etablering af to forskellige ordered logistiske regressionsmodeller, henholdsvis **model 1** og **model 2**. Årsagen til denne vurdering følger efter en kort præsentation af modellerne.

I hver af modellerne består den afhængige variabel af 4 kategorier, baseret på rangeringen af uddannelsesniveauerne, jf. figur 2.1. I model 1 er den afhængige variabel opdelt i 4 forskellige kategorier; grundskole, gymnasial uddannelse, erhvervsuddannelse og videregående uddannelse. Kategorien videregående uddannelse er en sammenlægning af alle fire typer af videregående uddannelse. I model 2 er den afhængige variabel ligeledes opdelt i 4 forskellige kategorier, disse kategorier er de fire typer af videregående uddannelse; kort videregående, akademisk bachelor, professionsbachelor og lang videregående uddannelse. En mere formel og detaljeret præsentation af model 1 og model 2 fremgår af kapitel 7.

Ved at konstruere to modeller mindskes antallet af kategorier i den afhængige variabel fra 7 til 4. Dette gør, som nævnt tidligere, fortolkningen af resultaterne mindre vanskelig, da der estimeres 3 funktioner for hver model i stedet for 6 (Parry, 2016). Regressionsanalysen foretages med det formål at undersøge hypoteserne opstillet i kapitel 5.

Proportional odds assumption er en streng antagelse, der i mange tilfælde ikke overholdes (Harrell Jr, 2001; McNulty, 2021; UCLA: Statistical Consulting Group, n.d.a). At antage, at den overholdes på tværs af alle uddannelsesniveauer, vurderes derfor at være urealistisk. Det er derfor også en af de største udfordringer ved en ordered logistisk regression. Antagelsen er desuden svær at teste, da statistiske tests kritiseres for at have en tendens til at forkaste nulhypotesen (at antagelsen overholdes). Foruden dette er der en generel tendens til at antagelsen ikke overholdes i komplekse modeller (Harrell Jr, 2001; McNulty, 2021; UCLA: Statistical Consulting Group, n.d.a). På grund af de mange observationer og typer af uafhængige variable må nærværende modeller antages at være komplekse. For at robusthedstjekke resultaterne af model 1 og model 2, findes det relevant at estimere modellerne ved brug af OLS. Som nævnt ovenfor er det generelt accepteret i moderne mikroøkonometri, at en OLS-regression er en god approksimering for en funktion med en diskret afhængig variabel (Angrist og Pischke, 2008). Endvidere er fortolkningen af en OLS regression meget intuitiv, da det er muligt at fortolke estimaterne direkte (Wooldridge, 2016). Af disse årsager vurderes det, at estimering af modellerne ved brug af OLS kan fungere som et robusthedstjek af estimaterne fra de ordered logistisk regressioner.

6.3 Opsummering

Det vurderes, ud fra ovenstående økonometriske overvejelser, at hypoteserne opstillet i kapitel 5 bedst undersøges ved brug af to ordered logistisk regressionsmodeller. Ved at gøre brug af denne modeltype er det muligt at opretholde rangeringen af uddannelsesniveauer, hvilken vurderes at være central for det videre arbejde. Robustheden af resultaterne af de to ordered logistisk regressionsmodeller, testes ved brug af OLS-estimering. Ud fra overvejelserne i indeværende kapitel opstilles regressionsligningen i det følgende kapitel.

Modelopbygning 7

I indeværende kapitel opstilles regressionsmodellerne, der skal bidrage til undersøgelsen af hypoteserne. På baggrund af overvejelserne i kapitel 6, gøres der brug af to ordered logistiske regressioner. Efter opstilling af regressionsligningen, præsenteres variablene, der inddrages i modellerne. Variablene er udvalgt på baggrund af teorien og resultaterne fra det empiriske litteraturstudie. Til slut redegøres der endvidere kort for variable, der ikke inddrages i den kvantitative analyse.

7.1 Opstilling af Regressionsmodeller

For at undersøge hypotese 1a-1e og hypotese 2a-2e opstilles følgende regressionsligning:

$$U_{it_1} = \beta_{j0} + \beta_1 KG_{it_0} + \beta_2 MI_{it_0} + \beta_3 MU_{it_0} + \beta_4 FI_{it_0} + \beta_5 FU_{it_0} + \beta_6 A_{it_0} + \beta_7 K_{it_0} + \beta_8 E_{it_0} + \beta_9 KO1_{it_0} + \beta_{10} KO2_{it_0} + \beta_{11} KO3_{it_0} + \beta_{12} KO4_{it_0} + \beta_{13} S_{it_0} + \beta_{14} MA_{it_0} + \beta_{15} MF_{it_0} + \beta_{16} MS1_{it_0} + \beta_{17} MS2_{it_0} + \beta_{18} MS3_{it_0} + \beta_{19} MS4_{it_0} + \beta_{20} FA_{it_0} + \beta_{21} FF_{it_0} + \beta_{22} FS1_{it_0} + \beta_{23} FS2_{it_0} + \beta_{24} FS3_{it_0} + \beta_{25} FS4_{it_0} + \varepsilon_{it} \quad (7.1)$$

Hvor β_{j0} er skæringspunkterne og ε_i er det idiosynkratiske fejld. Da det er ordered logistiske regressioner, estimeres $J - 1$ ligninger, jf. bilag D. Det eneste, der varierer mellem dem, er dog skæringspunktet, hvorfor skæringspunkterne angives β_{j0} . Konstanterne $\beta_1, \beta_2, \dots, \beta_{25}$ er modellernes parametre. Endvidere er $t_0 = 2002$, hvilket betyder variablen er opgjort i 2002. $t_1 = 2019$ og dermed er variablen opgjort i 2019. Den afhængige og de uafhængige variable er som følger:

Ligningsforklaring		
Symbol	Forklaring	Type
U_{it_1}	Uddannelsesniveau	Ordinal
KG_{it_0}	Karaktergennemsnit	Kontinuerlig
MI_{it_0}	Mor indkomst	Kontinuerlig
MU_{it_0}	Mor uddannelsesniveau	Diskret*
FI_{it_0}	Far indkomst	Kontinuerlig
FU_{it_0}	Far uddannelsesniveau	Diskret*
A_{it_0}	Alder	Diskret*
K_{it_0}	Køn	Dummy
E_{it_0}	Etnicitet, ikke dansk oprindelse	Dummy
$KO1_{it_0}$	Kommunetype, opland	Dummy
$KO2_{it_0}$	Kommunetype, provins	Dummy
$KO3_{it_0}$	Kommunetype, storby	Dummy
$KO4_{it_0}$	Kommunetype, hovedstad	Dummy
S_{it_0}	Søskende	Diskret*
MA_{it_0}	Mor alder	Diskret*

MF_{it_0}	Mor familietype, enlig	Dummy
$MS1_{it_0}$	Mor socialklasse, selvstændig	Dummy
$MS2_{it_0}$	Mor socialklasse, pension	Dummy
$MS3_{it_0}$	Mor socialklasse, overførselsindkomst	Dummy
$MS4_{it_0}$	Mor socialklasse, andre	Dummy
FA_{it_0}	Far alder	Diskret*
FF_{it_0}	Far familietype, enlig	Dummy
$FS1_{it_0}$	Far socialklasse, selvstændig	Dummy
$FS2_{it_0}$	Far socialklasse, pension	Dummy
$FS3_{it_0}$	Far socialklasse, overførselsindkomst	Dummy
$FS4_{it_0}$	Far socialklasse, andre	Dummy

*Variablen er opgjort diskret men operationaliseret som en kontinuerlig variabel

Regressionsligningen er den samme for både model 1 og model 2. Forskellen i modellerne opstår, da operationaliseringen af den afhængige variabel er forskellig. I model 1 inddrages alle unge i populationen uanset hvilket uddannelsesniveau, der er deres højeste gennemførte. I model 2 inddrages der derimod kun unge, som har gennemført en videregående uddannelse. Dette betyder, at der er færre observationer i model 2 end i model 1. Denne konstruktion medfører, at karakteristikaene for de forskellige variable kan variere fra model til model. Medianen og gennemsnit af den "samme" uafhængige variabel er dermed nødvendigvis ikke lig den samme værdi i model 2 som i model 1. For at sikre, at de uafhængige variable, der inddrages i modellen hver især, har en signifikant effekt på unges uddannelsesniveau udføres t-tests for hver af dem. T-testene udføres ved brug af en OLS-estimering, hvilket betyder, at den afhængige variabel behandles som en kontinuerlig variabel. På trods af, at den afhængige variabel er en ordinal variabel, vurderes det, at resultaterne af t-testsene kan give en indikation af, om der eksisterer en sammenhæng mellem den afhængige og uafhængige variabel. Resultaterne af t-testsene for både model 1 og model 2 fremgår af bilag E. I tilfælde af, at t-testen for en uafhængig variabel viser, at der er en signifikant sammenhæng med den afhængige variabel, fremgår der ikke yderligere herom. Hvis t-testen derimod viser, at der ikke eksisterer en signifikant sammenhæng, fremgår det under den pågældende variabel. For yderligere at sikre robustheden af de uafhængige variable, er en række andre operationaliseringer undersøgt. Information om disse robusthedsundersøgelser fremgår af bilag F.

Som beskrevet i kapitel 2, udarbejdes den kvantitative analyse på baggrund af data fra Danmarks Statistiks Forskerservice. En detaljeret beskrivelse af datasættet, der ligger til grund for analysen og dermed variablene i modellerne, fremgår af bilag A. I model 1 er der 39.090 observationer, hvilket er de personer, der færdiggjorde grundskolens 9. klasse i 2002 og som inddrages i den kvantitative analyse. Af de 39.090 personer havde 20.988 i 2019 gennemført en videregående uddannelse. Dermed er der 20.988 observationer i model 2.

I det følgende beskrives variablene, der indgår i regressionsligningen og dermed i den kvantitative analyse. Denne beskrivelse inkluderer desuden en redegørelse for forventningen til variablen, der tager udgangspunkt i den præsenterede teori og resultaterne fra studierne præsenteret i kapitel 4. Der redegøres endvidere kort for variablenes karakteristika i model 1 og model 2.

7.2 Afhængig Variabel

Den afhængige variabel, U_{it_1} , er både i model 1 og model 2 defineret som et individs højeste fuldførte uddannelse målt i 2019. U_{it_1} er en operationalisering af rangeringen af uddannelsesniveauerne i Danmark, jf. kapitel 2.4.2. Variablen er diskret og behandles som en ordinal variabel, hvilket betyder, at dens værdi har en betydning for rangeringen af den pågældende kategori. De tilgængelige værdier for U_{it_1} varierer mellem model 1 og model 2, hvorfor variablen beskrives enkeltvis for hver model.

Model 1, $U_{it_1} = 1, \dots, 4$

I model 1 er den afhængige variabel opdelt i 4 kategorier; grundskole ($U_{it_1} = 1$), gymnasial uddannelse ($U_{it_1} = 2$), erhvervsuddannelse ($U_{it_1} = 3$) og videregående uddannelse ($U_{it_1} = 4$). Kategorien videregående uddannelse er en sammenlægning af alle fire niveauer af videregående uddannelse, jf. kapitel 2.4.2.

Model 2, $U_{it_1} = 4, \dots, 7$

I model 2 er den afhængige variabel ligeledes opdelt i 4 kategorier, disse kategorier er de fire niveauer af videregående uddannelse; kort videregående ($U_{it_1} = 4$), akademisk bachelor ($U_{it_1} = 5$), professionsbachelor ($U_{it_1} = 6$) og lang videregående uddannelse ($U_{it_1} = 7$).

På grund af begrænsninger i det tilgængelige data, er det som nævnt kun muligt at observere personerne til de er mellem 31-35 år, jf. kapitel 2.2. Derfor er det ikke muligt at observere, hvis de fuldfører et højere uddannelsesniveau senere i livet. Operationaliseringen af variabelen kan dermed potentielt undervurdere personernes højeste fuldførte uddannelsesniveau, da der i virkeligheden ikke eksisterer et tydeligt *cut-off* på tilegnelsen af uddannelse. Det vurderes, at denne potentielle undervurdering er lille, da gennemsnitsalderen for gennemførelse af forskellige uddannelser er under alderen for populationen i 2019 (Produktivitetskommissionen, 2014; Uddannelses- og Forskningsministeriet, 2013).

7.3 Uafhængige Variable

I det følgende beskrives de uafhængige variable. Alle uafhængige variable er opgjort i år 2002, hvor personerne i populationen færdiggjorde grundskolens 9. klasse. Som tidligere nævnt er det efter grundskolen, at individerne står overfor deres første uddannelsesbeslutning. Det antages i dette projekt, at de mest nylige værdier, inden dette valg træffes, har den største påvirkning på individers uddannelsesbeslutning. Kun at måle faktorerne på et enkelt tidspunkt har nogle begrænsninger, da de fleste af dem kan ændre sig over tid. Ifølge Nielsen (2012) har socioøkonomiske faktorer som forældres uddannelse, en mere konstant effekt på individers uddannelsesvalg over livsperioden end for eksempel grundskoleforhold, jf. kapitel 4. De udfordringer, der beskrives i forhold til operationaliseringen af variablene for individers mor er de samme, som er gældende for individers far og nævnes derfor ikke ved beskrivelse af variablene for faren.

Da faktorerne kan variere over tid, er det ideelt set mere hensigtsmæssigt at måle dem over en længere periode. Denne fremgangsmåde er dog vanskelig at gøre brug af i praksis, da det er vanskeligt at vurdere, hvornår mødres uddannelsesniveau har størst påvirkning på unges uddannelsesbeslutning. Specifikt er det altså svært at vurdere, hvorvidt mødres uddannelsesniveau har størst betydning for unges uddannelsesbeslutning, når de er 15 år eller 25 år. Det forventes dog, at værdier, der er observeret senere i livet, sandsynligvis påvirker individers uddannelsesvalg i mindre grad. De fleste er eksempelvis flyttet hjemmefra, har eventuelt været igennem flere uddannelsesinstitutioner og måske været på arbejdsmarkedet (Danmarks Statistik, 2020).

At antage at individer tager en beslutning om hvilket niveau af uddannelse, de højest vil gennemføre, allerede når de færdiggør grundskolen, er en streng antagelse. Dette er strengt, taget i betragtning af, hvor mange der tager dobbeltuddannelse. Ifølge Børne- og Undervisningsministeriet (n.d.a), forventes 11,4% af 2002 årgangen at gennemføre både en gymnasial uddannelse og en erhvervsuddannelse indenfor 25 år efter de færdiggjorde grundskolen. Foruden dette er der mange, der senere i livet tager en anden uddannelse, der både kan være på samme-, højere- eller lavere niveau som den højeste fuldførte uddannelse, de har. Dette taler for, at personer ændrer beslutning i forhold til deres uddannelsesniveau gennem livet.

Værdier, der er observeret før de unge står overfor deres første uddannelsesvalg, har samme udfordring som værdier, der måles efter eller på samme tidspunkt. Selvom værdierne for 2002 ikke er repræsentative for en hel periode, vurderes det dog stadig, at dette er det bedste alternativ i nærværende tilfælde.

7.3.1 Forklarende Variable

Karaktergennemsnit, KG_{it_0}

Karaktergennemsnit inddrages i den kvantitative analyse som en forklarende variabel. Variablen er en proxy for individernes evner, der ifølge *schooling* modellen påvirker uddannelseslængde. Karaktergennemsnit, KG_{it_0} , er defineret som et individs karaktergennemsnit fra grundskolen. Gennemsnittet er udregnet ud fra både standpunktskarakterer og karakterer fra 9. klasses afgangsprøver. Der er taget udgangspunkt i gennemsnittet for populationen, da grundskolekaraktererne herfor er normalfordelte. Derfor er der ikke er nogle ekstreme værdier til at trække gennemsnittet op eller ned (Newbold et al., 2013). Karaktergennemsnit er en kontinuerlig variabel. Årsagen til, at der ikke differentieres mellem typer af karakterer, er operationel. I 2002, hvor de unge gik ud af grundskolen, var folkeskolens afgangsprøver ikke obligatoriske, hvorfor nogle af personerne ikke har opgjort karakterer herfor (Møller, 2019). For at bibeholde flest mulige observationer udregnes der derfor et samlet gennemsnit. Medianen og gennemsnittet for henholdsvis standpunktskarakterer og karakterer fra afgangsprøver er på samme niveau, jf. tabel B.5 i bilag B. Det forventes derfor ikke at ændre resultaterne, hvis kun den ene type karakterer inddrages. Det antages, at karakterskalaen er retvisende i forhold til forskelle mellem karakterer, således at der er et større spring fra karaktererne 4 til 7 end 02 til 4. En detaljeret beskrivelse af, hvordan operationaliseringen er foretaget i praksis, fremgår af bilag A.

Evner er vanskelige at definere som en entydig størrelse, hvorfor det også er vanskeligt at måle og operationalisere evner i praksis. I andre kvantitative analyser benyttes proxies som eksempelvis læsefærdigheder og grundskolekarakterer (Jaeger, 2007; Jensen og Nielsen, 2010; Nielsen og Vissing-Jorgensen, 2006). De fleste proxies er blevet kritiseret for ikke at være gode mål for evner, da de måler en bestemt type af evner. Grundskolekarakterer måler, hvordan individer klarer sig indenfor boglige fag og kan derved ikke opfange, om personer er for eksempel socialt begavede eller dygtige håndværkere. Derudover kan lave karakterer skyldes, at individer ikke lægger en indsats og dermed er lave karakterer ikke nødvendigvis på grund af manglende evner.

Der er ikke et entydigt svar på hvilken proxy, der bedst beskriver et individs evner. Grundskolekarakterer er valgt, da der er oplysninger herom for alle personerne i kohorten. Endvidere benyttes karakterer i flere studier, der undersøger samme problemstilling.

I den præsenterede teori antages det, at personer med flere evner har et større marginalafkast af uddannelse og derved gennemfører mere uddannelse (Borjas, 2016). Dette understøttes af empiriske studier, der finder, at der er en positiv selvseleksion til yderligere uddannelse (Nielsen og Vissing-Jorgensen, 2006). Dermed forventes det, at der eksisterer en positiv sammenhæng mellem evner, her målt som karaktergennemsnit, og unges højeste fuldførte uddannelsesniveau.

Af nedenstående tabel 7.2 fremgår en oversigt over karakteristikaene for variabelen karaktergennemsnit i model 1 og model 2. Af tabellen fremgår det, at medianen og karaktergennemsnittet er højere for model 2 end model 1, hvor alle individer har gennemført en videregående uddannelse. Dette understøtter forventningen til variabelen.

Tabel 7.2: Karakteristika for Variablen Grundskolekarakterer - Model 1 og Model 2

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Model 1	-3,00	4,85	6,38	6,17	7,57	12,00
Model 2	0,00	6,29	7,24	7,18	8,19	11,67

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

Ved at benytte gennemsnittet er det ikke muligt, at se variationen i et individs karakterer. Derved kan en person klare sig godt i en række fag og dårligt i andre og derved have et middelmådigt karaktergennemsnit. Hvis der er stor variation i, hvilke karakterer individer får i grundskolen, er denne ikke en god proxy variabel for individers iboende evner.

Mor Indkomst, MI_{it_0}

MI_{it_0} defineres som logaritmen af et individ i 's mors skattepligtige indkomst plus en kr.¹ i 2002, hvor individet færdiggjorde grundskolen. Morens skattepligtige indkomst er opgjort i kr. i 2002-priser. Variablen er en kontinuierlig variabel og benyttes som en forklarende variabel.

Mødres indkomst benyttes, som nævnt sammen med de andre forklarende socioøkonomiske faktorer, som proxies for diskonteringsrenten. Ud fra *schooling* modellen forventes der derfor en positiv sammenhæng mellem morens indkomst og individers højeste fuldførte uddannelsesniveau, da højere familieindkomst medfører at den unge har en lavere diskonteringsrente. Desuden forventes dette, da flere af de empiriske studier fra kapitel 4 finder en positiv sammenhæng mellem unges uddannelsespræstationer og deres forældres indkomst (Benders, 2013; Reardon et al., 2012; Andersen et al., 2018).

Af tabel 7.3 fremgår det, at minimumværdien for mødres indkomst både i model 1 og model 2 er 0, hvilket indikerer, at moren ikke har opgjort nogen skattepligtig indkomst for 2002. Det fremgår endvidere, at både 1. kvartil, medianen, gennemsnittet og 3. kvartil for mødres indkomst er højere i model 2 end i model 1. Forskellene fra model 1 til model 2 understøtter forventningerne om, at der er en positiv sammenhæng mellem forældres indkomst og sandsynligheden for at gennemføre en uddannelse på et højere niveau.

Tabel 7.3: Karakteristika for Variablen Mor Indkomst*, Opgjort i Kr. (2002-Priser) - Model 1 og Model 2

	Min.	1st Qu.	Median	Mean	3rd Qu. ²
Model 1	0,00	147.839,00	197.929,00	206.433,00	254.699,00
Model 2	0,00	161.884,00	215.170,00	224.128,00	272.909,00

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice. *I regressionsmodellerne benyttes $\log(\text{skattepligtig indkomst} + 1)$

Da variabelen *mor indkomst* er angivet som en enkelt værdi opgjort i 2002, da individ i færdiggjorde grundskolen, er det ikke muligt at tage højde for variationer i indkomsten over tid. Indkomsten varierer i forskellige rater igennem livet (Eriksen, 2021). Morens indkomst er dermed et øjebliksbillede for 2002, hvilket medfører, at den er sårbar overfor midlertidige stød. Det vil være mere ideelt at gøre brug af flere år til at måle, hvordan morens indkomst har været gennem individets opvækst. Dermed ville sårbarheden overfor irregulære situationer såsom midlertidig arbejdsløshed, barsel eller bonusudbetalinger minimeres. Operationaliseringen af variabelen kan både medføre, at morens gennemsnitlige indkomst under- eller overvurderes. Ovenfor argumenteres der dog for, at det er relevant at benytte informationer fra netop 2002, da det er på dette tidspunkt, individet træffer dets første uddannelsesbeslutning. Derfor vurderes indkomsten i 2019 også at være en tilstrækkelig god indikation i nærværende tilfælde.

Der eksisterer en svag positiv korrelation mellem *mor indkomst* og *mor socialklasse*, *lønmodtager*, samt en svag negativ korrelation mellem *mor indkomst* og *mor socialklasse*, *overførselsindkomst* samt *mor socialklasse*, *andre*, jf. bilag G.

Mor Uddannelsesniveau, MU_{it_0}

Variablen, MU_{it_0} , defineres som individ i 's mors højeste gennemførte uddannelsesniveau i 2002. Rangeringen af uddannelsesniveauerne følger rangeringen, der fremgår af kapitel 2.4.2. MU_{it_0} kan dermed tage værdierne 1 til 7, hvor $MU_{it_0} = 1$ er grundskole, $MU_{it_0} = 2$ er gymnasial uddannelse og så videre. Variablen er derved diskret og benyttes som en forklarende variabel i regressionsmodellerne.

Som beskrevet i hypotese 1c forventes der at være en positiv sammenhæng mellem mødres uddannelsesniveau og unges højeste fuldførte uddannelsesniveau. Dette er forventningen, da *mor uddannelse* fungerer som en proxy for individ i 's diskonteringsrente. I sammenhæng med mødres uddannelsesniveau betyder

¹Da $\log(0)$ ikke er matematisk defineret lægges der 1 kr. til morens skattepligtige indkomst, for at kunne inkludere mødre uden indkomst.

²Max. værdier fjernet pga. risiko for identificering

det, at hvis mødres uddannelsesniveau er højere, er individets diskonteringsrente lavere. En lavere diskonteringsrente medfører ifølge teorien, at man tilegner sig mere uddannelse (Borjas, 2016). Forventningen bakkes op af empiriske studier, der finder evidens for, at der eksisterer en sammenhæng mellem forældres og unges uddannelsesniveau, jf. kapitel 4 (Østergaard-Tygesen, 2018; Andersen et al., 2018; Undervisningsudvalget 2016-2017, 2017; Farooq et al., 2011).

Af nedenstående tabel fremgår en oversigt over fordelingen af mødres uddannelsesniveau for model 1 og model 2. Som det ses i tabellen, er der i model 2 en mindre andel, hvis mødres højeste fuldførte uddannelsesniveau er grundskolen, mens der er en større andel, hvis højeste fuldførte uddannelse er en videregående uddannelse. Det understøtter forventningen om, at unge, hvis mødre har et højere uddannelsesniveau, har større sandsynlighed for at fuldføre et højere uddannelsesniveau.

Tabel 7.4: Oversigt over Procentvis Fordeling af Mor Uddannelsesniveau - Model 1 og Model 2

	Grund-skole	Gymnasial uddannelse	Erhvervs-uddannelse	Kort videre-gående	Akademisk bachelor	Professions-bachelor	Lang videre-gående
Model 1	27,76 %	4,00 %	36,53 %	3,78 %	0,39 %	23,28 %	4,27 %
Model 2	18,60 %	4,54 %	32,59 %	4,79 %	0,48 %	32,17 %	6,84 %

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

Der eksisterer en korrelation mellem mødres uddannelsesniveau og fædres uddannelsesniveau både i model 1 og model 2, jf. figur G.1 og figur G.2 i bilag G. Korrelationsmatricerne i bilag G viser endvidere, at der i både model 1 og model 2 er en svag negativ korrelation mellem *mor uddannelsesniveau* og *mor socialklasse*, *overførelsesindkomst* men en svag positiv korrelation mellem *mor uddannelsesniveau* og *mor socialklasse*, *lønmodtager*.

Da indkomsten ifølge økonomisk teori er markedsbestemt efter uddannelsesniveau, forventes der at være en positiv korrelationen mellem mors højeste uddannelsesniveau og hendes indkomst. Der er en svag positiv korrelation mellem dem i begge regressionsmodellerne, jf. figur G.1 og figur G.2 i bilag G.

Far Indkomst, FI_{it_0}

FI_{it_0} defineres som logaritmen af individ i 's fars skattepligtige indkomst plus en kr.³ i 2002, hvor individet færdiggjorde grundskolen. Farens skattepligtige indkomst er opgjort i kr. i 2002-priser. FI_{it_0} er en kontinuierlig variabel og benyttes i nærværende tilfælde som en forklarende variabel. I lighed med *mor indkomst*, forventes der at være en positiv sammenhæng mellem *farens indkomst* og individers højeste fuldførte uddannelsesniveau, da den er en proxy for diskonteringsrenten. Karakteristika for *far indkomst* fremgår af tabel 7.5. FI_{it_0} er for både 1. kvartil, medianen, gennemsnittet og 3. kvartil højere i model 2 end i model 1. Dette understøtter forventningen til variabelen.

Tabel 7.5: Karakteristika for Variablen Far Indkomst*, Opgjort i Kr. (2002-Priser) - Model 1 og Model 2

	Min.	1st Qu.	Median	Mean	3rd Qu. ⁴
Model 1	0	185.186,00	250.347,00	284.044,00	327.518,00
Model 2	0	202.294,00	270.472,00	315.931,00	360.800,00

Kilde: Egne beregninger på data fra Danmarks Statistiks Forskerservice. *I regressionsmodellerne benyttes $\log(\text{skattepligtig indkomst} + 1)$

³Da $\log(0)$ ikke er matematisk defineret lægges der 1 kr. til farens skattepligtige indkomst, for at kunne inkludere fædre uden indkomst.

⁴Max. værdier fjernet pga. risiko for identificering

I model 1 og model 2 er der en svag positiv korrelation mellem *far indkomst* og dummy variablen *far socialklasse*, *lønmodtager*. Der er en svag negativ korrelation mellem *far indkomst* og *far socialklasse*, *overførselsindkomst*, jf. bilag G.

Far Uddannelsesniveau, FU_{it_0}

Variablen, FU_{it_0} , defineres som individ i 's fars højeste gennemførte uddannelsesniveau i 2002. Rangeringen af uddannelsesniveau følger rangeringen, der fremgår af kapitel 2.4.2. FU_{it_0} kan dermed tage værdierne 1 til 7, hvor $FU_{it_0} = 1$ er grundskole, $FU_{it_0} = 2$ er gymnasial uddannelse og så videre. Variablen er en forklarende variabel, og det forventes, at der er en positiv sammenhæng mellem fædres uddannelsesniveau og sandsynligheden for at gennemføre uddannelse på et højere niveau, jf. hypotese 1e. Ligesom for mødres uddannelsesniveau er der i model 2 en mindre andel, hvis fædres højeste fuldførte uddannelsesniveau er grundskole, mens der er en større andel, hvis højeste fuldførte uddannelsesniveau er en videregående uddannelse, jf. tabel 7.6. Det understøtter forventningen, om at unge, hvis fædre har et højere uddannelsesniveau, har større sandsynlighed for at gennemføre en uddannelse på højere niveau.

Tabel 7.6: Oversigt over Procentvis Fordeling af Far Uddannelsesniveau - Model 1 og Model 2

	Grund- skole	Gymnasial uddannelse	Erhvervs- uddannelse	Kort videre- gående	Akademisk bachelor	Professions- bachelor	Lang videre- gående
Model 1	24,11 %	3,77 %	46,07 %	4,07 %	0,60 %	12,66 %	8,72 %
Model 2	16,42 %	4,59 %	41,79 %	4,74 %	0,89 %	17,84 %	13,74 %

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

Der forventes, ligesom for mors indkomst, at være en positiv korrelationen mellem mellem fars højeste uddannelsesniveau og hans indkomst. Der eksisterer en meget svag korrelation mellem fars uddannelsesniveau og hans indkomst, jf. figur G.1 og figur G.2 i bilag G. Korrelationsmatricerne i bilag G viser, at der i model 1 er en svag negativ korrelation mellem *far uddannelsesniveau* og *far socialklasse*, *overførselsindkomst* samt en svag positiv korrelation mellem *far uddannelsesniveau* og *far socialklasse*, *lønmodtager*.

7.3.2 Kontrolvariable

Alle ovennævnte variable inddrages i regressionsmodellerne som forklarende variable og herudfra undersøges hypotese 1a-1e og hypotese 2a-2e. I modellerne inddrages der også en række kontrolvariable, jf. kapitel 5. Disse fremgår af ligningsforklaringen. En beskrivelse af hver af kontrolvariablene og forventningerne hertil samt variablenes operationalisering og karakteristika fremgår af bilag C.1. Alle kontrolvariable er inddraget på baggrund af kapitel 3 og kapitel 4.

7.3.3 Udeladte variable

Variablene, der inddrages i regressionsmodellerne, er som nævnt udvalgt på baggrund af empiriske studier, som er beskrevet i kapitel 4 og den præsenterede teori i kapitel 3. Der er dog fire faktorer der ikke inddrages i modellerne. Det drejer sig om faktorerne for peer-gruppe, den forventede indkomst, forholdene i grundskolen og bløde faktorer. Disse er fravalgt, da de ikke kan måles direkte og desuden er det ikke muligt, at finde proxies i det tilgængelige data, der måler faktorerne tilstrækkeligt. Derudover er der for nogle af faktorerne fundet tvetydige resultater. For yderligere information om faktorerne samt en uddybning af, hvorfor de er fravalgt, henvises til bilag C.2.

7.4 Opsummering

For at undersøge de formulerede hypoteser er der opstillet to regressionsmodeller. De afhængige variable i regressionsmodellerne er det højeste fuldførte uddannelsesniveau for den kohorte, der færdiggjorde

grundskolens 9. klasse i 2002. I model 1 er uddannelsesniveauerne: grundskole, gymnasial uddannelse, erhvervsuddannelse og videregående uddannelse. I model 2 er uddannelsesniveauerne: kort videregående uddannelse, akademisk bachelor, professionsbachelor og lang videregående uddannelse. Begge modeller indeholder det samme antal uafhængige variable. De forklarende variable i modellerne er mødres indkomst og højeste fuldførte uddannelsesniveau i 2002, samt det tilsvarende for individernes fædre. Derudover er evner, målt som karaktergennemsnit, en forklarende variabel. De resterende 20 variable inddrages som kontrolvariable. Der er en række udfordringer med operationaliseringerne af variablene. Det vurderes dog, at inddragelsen af variablene er nødvendig for at undgå misspecifikation, da de inddrages på baggrund af økonomisk teori og resultater fra empiriske studier, jf. kapitel 3 og 4. Variablene anvendes i næste kapitel i de to regressionsmodeller.

Kvantitativ Analyse 8

I indeværende kapitel undersøges det, om de opstillede hypoteser understøttes. Endvidere præsenteres estimerne af kontrolvariablene og disse sammenholdes med forventningerne fra bilag C. Der foretages robusthedsundersøgelser af modellerne og afslutningsvis foretages en samlet vurdering af modellerne.

8.1 Beslutningsregel for Hypoteser

I det følgende beskrives det, hvordan hypoteserne undersøges. Til at undersøge hypoteserne opstilles to modeller, model 1 og model 2, jf. kapitel 6 og kapitel 7. I model 1 inddrages individer fra datasættet uanset hvilket uddannelsesniveau, der er deres højeste gennemførte, hvorimod det i model 2 kun er individer, der har gennemført en videregående uddannelse, der inddrages.

Hypotese 1a-1e, er at der eksisterer en positiv sammenhæng mellem en bestemt forklarende faktor og unges højeste fuldførte uddannelsesniveau. I sammenhæng med dette defineres de forklarende faktorer som de faktorer der, på baggrund teorien og det empiriske litteraturstudie, kan opstilles klare forventninger til, jf. kapitel 5. Disse faktorer undersøges i hver sin hypotese 1a-1e.

For at undersøge, hvorvidt hypoteserne i gruppe 1 understøttes, udarbejdes en ordered logistisk regression for model 1. Som beskrevet i kapitel 6 vurderes det, at en ordered logistisk regression er den mest optimale model, da denne både kan håndtere, at den afhængige variabel er diskret og rangeret. Estimerne og odds ratios i modellerne samt deres p-værdier undersøges og herudfra vurderes det, hvorvidt hver af hypoteserne understøttes eller ikke. En hypotese i gruppe 1 understøttes, hvis estimatet for den forklarende faktor er positivt og signifikant i model 1. Et estimat betegnes her som signifikant, hvis det er signifikant på maksimalt et 10% signifikansniveau (Wooldridge, 2016).

Hypotese 2a-2e er, at der eksisterer en positiv sammenhæng mellem en bestemt af de forklarende faktor og unges højeste fuldført uddannelsesniveau, også når der udelukkende ses på unge, som har fuldført en videregående uddannelse. For at undersøge hver af hypoteserne for hypotese 2, 2a-2e, sammenlignes resultaterne for model 2 med resultaterne for model 1. Det undersøges, om estimerne i modellerne er signifikante og har et positivt fortegn. Hypoteserne, 2a-2e, betegnes hver især som understøttet, hvis den forklarende faktor i hypotesen har en signifikant positiv effekt i både model 1 og model 2.

8.2 Modellernes Resultater

Model 1 og model 2 er ordered logistiske regressioner, hvilket betyder, at variablenes koefficienter estimeres som log odds eller logit. Dette kan matematisk skrives på følgende måde $\text{logit}(P(Y \leq j)) = \beta_{j0} + \beta_{j1}x_1 + \dots + \beta_{jp}x_p$, jf. bilag D. På baggrund af dette er fortolkningen af modellernes koefficienter ikke intuitiv. Det er dog muligt at udregne odds ratios, hvilket gør fortolkningen af estimerne mere gennemskuelig. Odds ratios findes ved at eksponentiere koefficienterne, jf. bilag D (UCLA: Statistical Consulting Group, n.d.b).

I nedenstående tabel 8.1 ses resultaterne af model 1, der estimerer, hvilke faktorer der påvirker individers højeste fuldførte uddannelsesniveau. Som det fremgår af kapitel 7.2, består den afhængige variabel i denne model af uddannelsesniveauerne grundskole, gymnasial uddannelse, erhvervsuddannelse og videregående uddannelse.

Umiddelbart efter resultaterne af model 1, præsenteres resultaterne af model 2. I model 2 består den afhængige variabel af uddannelsesniveauerne kort videregående uddannelse, akademisk bachelor,

professionsbachelor og lang videregående uddannelse. Resultaterne af model 2 ses i tabel 8.2.

Af tabel 8.1 og tabel 8.2 fremgår både estimer, standardfejl, t-værdier og p-værdier. Det er muligt at undersøge estimaternes signifikans ved brug af t-værdierne og p-værdierne, i det følgende undersøges dog kun p-værdierne (Wooldridge, 2016). Af bilag H fremgår endvidere odds ratios og 95% konfidensintervaller for estimerne i model 1 og model 2.

Resultater for Model 1

Tabel 8.1*

Tabel 8.1: Ordered Logistisk Regression - Model 1

	<i>Estimat</i>	<i>Std. Fejl</i>	<i>t-værdi</i>	<i>p-værdi</i>	<i>Odds Ratios</i>	
KG_{it_0}	0.555	0.007	79.957	0.000	***	1.742
MI_{it_0}	0.024	0.007	3.334	0.001	***	1.024 (1)
MU_{it_0}	0.122	0.007	18.416	0.000	***	1.130
FI_{it_0}	0.005	0.006	0.821	0.412		1.005 (1)
FU_{it_0}	0.133	0.007	18.594	0.000	***	1.142
A_{it_0}	-0.195	0.026	-7.628	0.000	***	0.823
K_{it_0}	-0.324	0.022	-14.444	0.000	***	0.723
$E1_{it_0}$	0.322	0.057	5.645	0.000	***	1.380
$KO1_{it_0}$	-0.034	0.032	-1.061	0.289		0.967
$KO2_{it_0}$	-0.051	0.030	-1.719	0.086		0.950
$KO3_{it_0}$	0.054	0.040	1.363	0.173		1.056
$KO4_{it_0}$	-0.173	0.033	-5.209	0.000	***	0.841
S_{it_0}	0.044	0.013	3.487	0.001	***	1.045
MA_{it_0}	0.024	0.003	6.850	0.000	***	1.024
MF_{it_0}	-0.121	0.036	-3.403	0.001	***	0.886
$MS1_{it_0}$	-0.044	0.061	-0.719	0.472		0.957
$MS2_{it_0}$	-0.132	0.215	-0.611	0.541		0.877
$MS3_{it_0}$	-0.218	0.033	-6.534	0.000	***	0.804
$MS4_{it_0}$	-0.057	0.106	-0.536	0.592		0.945
FA_{it_0}	0.005	0.003	1.828	0.068		1.005
FF_{it_0}	-0.056	0.036	-1.547	0.122		0.946
$FS1_{it_0}$	0.030	0.037	0.802	0.422		1.030
$FS2_{it_0}$	-0.196	0.092	-2.135	0.033	*	0.822
$FS3_{it_0}$	-0.268	0.039	-6.850	0.000	***	0.765
$FS4_{it_0}$	-0.101	0.102	-0.990	0.322		0.904
1 2 ($\beta_{1,0}$)	0.192	0.417	0.461	0.645		
2 3 ($\beta_{2,0}$)	0.318	0.417	0.764	0.445		
3 4 ($\beta_{3,0}$)	2.437	0.417	5.842	0.000	***	
Antal Obs.						39,090
Residual Deviance						68808.92
AIC						68864.92

Note: *p<0.1 **p<0.05 ***p<0.01

*I Tabel 8.1 benyttes punktum som decimal separator.

(1) Odds ratios for MI_{it_0} og FI_{it_0} er $\exp(\log(\log(odds)))$ da MI_{it_0} og FI_{it_0} er loggede variable.

Resultater for Model 2

Tabel 8.2*

Tabel 8.2: Ordered Logistisk Regression - Model 2

	<i>Estimat</i>	<i>Std. Fejl</i>	<i>t-værdi</i>	<i>p-værdi</i>	<i>Odds Ratios</i>	
KG_{it_0}	0.474	0.010	46.721	0.000	***	1.607
MI_{it_0}	0.007	0.010	0.719	0.472		1.007 (1)
MU_{it_0}	0.092	0.008	11.914	0.000	***	1.097
FI_{it_0}	0.007	0.007	0.940	0.347		1.007 (1)
FU_{it_0}	0.105	0.008	13.258	0.000	***	1.111
A_{it_0}	-0.078	0.037	-2.142	0.032	*	0.925
K_{it_0}	0.123	0.028	4.332	0.000	***	1.130
$E1_{it_0}$	0.434	0.077	5.613	0.000	***	1.544
$KO1_{it_0}$	0.103	0.041	2.491	0.013	*	1.109
$KO2_{it_0}$	0.085	0.038	2.230	0.026	*	1.089
$KO3_{it_0}$	0.284	0.048	5.879	0.000	***	1.329
$KO4_{it_0}$	0.214	0.042	5.152	0.000	***	1.238
S_{it_0}	0.025	0.017	1.481	0.139		1.025
MA_{it_0}	0.015	0.004	3.312	0.001	***	1.015
MF_{it_0}	0.045	0.050	0.910	0.363		1.047
$MS1_{it_0}$	0.182	0.074	2.466	0.014	***	1.199
$MS2_{it_0}$	-0.441	0.264	-1.671	0.095		0.643
$MS3_{it_0}$	0.007	0.050	0.142	0.887		1.007
$MS4_{it_0}$	0.106	0.140	0.759	0.448		1.112
FA_{it_0}	0.009	0.004	2.315	0.021	*	1.009
FF_{it_0}	0.006	0.052	0.117	0.907		1.006
$FS1_{it_0}$	0.062	0.045	1.374	0.169		1.064
$FS2_{it_0}$	-0.227	0.115	-1.971	0.049	*	0.797
$FS3_{it_0}$	0.064	0.059	1.087	0.277		1.066
$FS4_{it_0}$	0.257	0.131	1.966	0.049	*	1.293
4 5 ($\beta_{1,0}$)	2.064	0.589	3.501	0.001	***	
5 6 ($\beta_{2,0}$)	2.473	0.589	4.196	0.000	***	
6 7 ($\beta_{3,0}$)	4.743	0.590	8.038	0.000	***	
Antal Obs.						20,988
Residual Deviance						42580.32
AIC						42636.32

Note: *p<0.1 **p<0.05 ***p<0.01

*I Tabel 8.2 benyttes punktum som decimal separator.

(1) Odds ratios for MI_{it_0} og FI_{it_0} er $\exp(\log(\log(odds)))$ da MI_{it_0} og FI_{it_0} er loggede variable.

Det er undersøgt, om model 1 og model 2 overholder antagelserne for en ordered logistisk regression. En beskrivelse af antagelserne samt undersøgelse af, hvorvidt model 1 og model 2 overholder dem, fremgår af bilag I. Både model 1 og model 2 overholder antagelse 1, som er, at den afhængige variabel er ordinal. Derudover er en eller flere af de uafhængige variable kontinuerlige eller ordinale, jf. tabel 7.1, hvilket betyder, at antagelse 2 også overholdes i begge modeller. Det er ved brug af korrelationsmatricerne i bilag G undersøgt, hvorvidt der eksisterer multikollinearitet i modellerne. Dette er ikke tilfældet, hvorfor antagelse 3 også overholdes, jf. bilag I.

Både model 1 og model 2 har problemer med at overholde proportional odds assumption, jf. bilag I. Som det fremgår af kapitel 6, er denne antagelse streng og overholdes i mange tilfælde ikke. Der eksisterer endvidere en generel tendens til at antagelsen ikke overholdes i komplekse modeller, hvilket både model 1 og model 2 antages at være, jf. kapitel 6 (Harrell Jr, 2001; McNulty, 2021; UCLA:

Statistical Consulting Group, n.d.b). Af de forklarende variable er det især variabelen karaktergennemsnit, KG_{it_0} , der er problematisk, jf. bilag I. På baggrund af overvejelserne foretaget i kapitel 6 findes det dog stadig mest optimalt at gøre brug af en ordered logistisk regression. For at sikre robustheden, estimeres regressionsligning 7.1 med OLS-estimering for både model 1 og model 2, jf. kapitel 6. Robusthedstjek af modellerne fremgår af kapitel 8.3.

8.2.1 Undersøgelse af Hypotesegruppe 1

I det følgende undersøges det først, hvorvidt hver af hypoteserne i gruppe 1 understøttes. Hvorvidt hypoteserne er understøttet vurderes ud fra kriterierne opstillet i kapitel 8.1. Det undersøges om hypoteserne understøttes ved brug af resultaterne fra model 1.

Undersøgelse af Hypotese 1a: Evner

Hypotese 1a er, at der eksisterer en positiv sammenhæng mellem unges evner og deres højeste fuldførte uddannelsesniveau. Det fremgår af kapitel 7, at unges karaktergennemsnit fra grundskolen benyttes som en proxy for deres evner. Derfor er det nødvendigt at undersøge estimatet og odds ratios for variabelen *karaktergennemsnit*, KG_{it_0} , i model 1 for at vurdere, hvorvidt hypotese 1a understøttes.

Estimatet for variabelen KG_{it_0} er positivt og signifikant i model 1. Odds ratios for variabelen er lig 1,742, jf. tabel 8.1. Dette betyder, at ved en enheds stigning i et individs karaktergennemsnit er vedkommende 1,742 gange mere sandsynlig til at gennemføre en uddannelse på niveau j eller derunder, når alle andre variable holdes konstant (UCLA: Statistical Consulting Group, n.d.b). Dette kan også skrives som, at et individ har 74,2% større sandsynlighed for at gennemføre en uddannelse på niveau j eller derunder, når alle andre variable holdes konstant. Da der i modellen antages proportional odds, er forholdet mellem eksempelvis det laveste og alle højere niveauer af uddannelse, det samme som forholdet mellem det næst laveste niveau og alle højere niveauer af uddannelse (UCLA: Statistical Consulting Group, n.d.b). Dermed betyder estimatet for KG_{it_0} , at hvis et individs karaktergennemsnit stiger med en, så er sandsynligheden for at gennemføre en gymnasial uddannelse eller derunder 74,2% højere. Det betyder også, at sandsynligheden for at gennemføre en erhvervsuddannelse eller derunder er 74,2% højere og at sandsynligheden for at gennemføre en videregående uddannelse eller derunder er 74,2% højere. Dette er når alle andre variable holdes konstant. Fremover skrives dette som at *individet gennemfører en uddannelse, der er ét niveau højere*.

Da estimatet for karaktergennemsnit er positivt og signifikant, eksisterer der en positiv sammenhæng mellem unges evner og deres højeste fuldførte uddannelsesniveau. Dermed understøttes hypotese 1a i model 1. Det fremgår ovenfor, at variabelen KG_{it_0} har problemer med overholdelse af proportional odds, dette behandles nærmere i kapitel 8.3.

Undersøgelse af Hypotese 1b: Mødres Indkomst

Hypotese 1b er, at der eksisterer en positiv sammenhæng mellem mødres indkomst og unges højeste fuldførte uddannelsesniveau. For at undersøge hvorvidt denne hypotese understøttes, undersøges variabelen *mor indkomst*, MI_{it_0} . Det fremgår af kapitel 7.3, at operationaliseringen af denne variabel er logaritmen af individ i 's mors indkomst, hvilket betyder, at fortolkningen heraf er lidt anderledes. Estimatet kan derfor ikke aflæses direkte af tabel 8.1.

For at finde ud af hvilken sammenhæng en stigning i mors indkomst på 1% har med unges sandsynlighed for at gennemføre en uddannelse, der er ét niveau højere, ganges estimatet (0,024) med 0,01 ($0,024 \cdot 0,01 = 0,00024$). For at finde odds ratios er det dermed denne værdi (0,00024), der skal eksponentieres, hvilket giver ca. 1,00024. Alternativt er det også muligt at udregne det ved værdien, der fremgår af odds ratios (1,024) i tabel 8.1: $\exp((\log(1,024) \cdot \log(1,01))) \approx 1,00024$.

Dette betyder, at hvis et individs mors indkomst stiger med 1% er der 0,024% større sandsynlighed for, at individet gennemfører en uddannelse, der er ét niveau højere, når alle andre faktorer holdes konstant. Dermed eksisterer der en positiv, enddog meget lille, signifikant effekt mellem mødres indkomst og unges højeste fuldførte uddannelsesniveau. Dette betyder, at hypotese 1b understøttes.

Undersøgelse af Hypotese 1c: Mødres Uddannelse

Hypotese 1c er, at der eksisterer en positiv sammenhæng mellem mødres uddannelsesniveau og unges højeste fuldførte uddannelsesniveau. Det undersøges, hvorvidt denne hypotese understøttes, ved at undersøge variablen *mor uddannelse*, MU_{it_0} , i model 1. Det fremgår af tabel 8.1, at estimatet for MU_{it_0} er positivt og signifikant. Odds ratios er lig 1,130, hvilket betyder følgende: hvis et individs mors uddannelse stiger med et niveau, stiger sandsynligheden for at individet fuldfører en uddannelse, der er ét niveau højere, med 13,0%. Dermed understøttes hypotese 1c og der eksisterer en positiv signifikant sammenhæng mellem mødres uddannelsesniveau og unges højeste fuldførte uddannelsesniveau.

Undersøgelse af Hypotese 1d: Fædres Indkomst

Hypotese 1d er, at der eksisterer en positiv sammenhæng mellem fædres indkomst og unges højeste fuldførte uddannelsesniveau. For at undersøge, hvorvidt denne hypotese understøttes, undersøges variablen *far indkomst*, FI_{it_0} , i model 1. Det fremgår af tabel 8.1, at estimatet for FI_{it_0} i model 1 ikke er signifikant. Dette betyder, at hypotese 1d ikke understøttes. En årsag til, at der ikke findes en signifikant sammenhæng kan skyldes, at uddannelse i Danmark er finansieret gennem skatten, jf. kapitel 4. Dermed er der ifølge modellen forskel mellem mødres og fædres indflydelse. Det tyder ud fra estimaterne på, at det kun er mødre, der har en indflydelse på unges uddannelsesbeslutning. Det kan potentielt forklares af, at mødre ofte er den primære omsorgsperson.

Undersøgelse af Hypotese 1e: Fædres Uddannelse

Hypotese 1e er, at der eksisterer en positiv sammenhæng mellem fædres uddannelsesniveau og unges højeste fuldførte uddannelsesniveau. Det undersøges hvorvidt hypotesen understøttes, ved at undersøge variablen *far uddannelse*, FU_{it_0} , i model 1. Det fremgår af tabel 8.1, at estimatet for model 1 er lig 0,133 samt at det er signifikant. Fortolkes der på odds ratios for FU_{it_0} , som er 1,142, betyder dette; at hvis et individs fars uddannelsesniveau stiger med et niveau, stiger sandsynligheden for at individet fuldfører en uddannelse, der er ét niveau højere, med 14%, når alle andre variable holdes konstant. Ud fra resultaterne i model 1 betegnes hypotese 1e som understøttet.

8.2.2 Undersøgelse af Hypotesegruppe 2

I det følgende undersøges det, om hypoteserne i gruppe 2 understøttes. Hypoteserne i gruppe 2 undersøges ved brug af resultaterne fra model 1 og model 2, der fremgår af tabel 8.1 og tabel 8.2.

Hypotese 2a: Evner

Hypotese 2a er, at der eksisterer en positiv sammenhæng mellem unges evner og deres højeste fuldførte uddannelsesniveau, også når der udelukkende ses på unge, som har fuldført en videregående uddannelse. For at undersøge, om hypotese 2a understøttes, ses der først om variablen KG_{it_0} er positiv og signifikant i model 2. Det fremgår af tabel 8.2, at dette er tilfældet. Odds ratios for variablen er lig med 1,607. Dette betyder, at et individ har 60,7% større sandsynlighed for at fuldføre en uddannelse, der er ét niveau højere, når vedkommendes karaktergennemsnit fra grundskolen stiger med en, hvis alle andre variable holdes konstant. Dette er for individer, der har gennemført en videregående uddannelse.

Af ovenstående afsnit 8.2.1 fremgår det, at hypotese 1a understøttes. Da der både i model 1 og i model 2 eksisterer en signifikant positiv sammenhæng mellem unges evner og deres højeste fuldførte uddannelsesniveau, betegnes hypotese 2a dermed som understøttet.

Hypotese 2b: Mødres Indkomst

Hypotese 2b er, at der eksisterer en positiv sammenhæng mellem mødres indkomst og unges højeste fuldførte uddannelsesniveau, også når der udelukkende ses på unge, som har fuldført en videregående uddannelse. Det fremgår af tabel 8.2, at estimatet for variablen *mor indkomst*, MI_{it_0} ikke er signifikant. Dette betyder, at der ikke kan påvises en positiv signifikant sammenhæng mellem mødres indkomst og

unges højeste fuldførte uddannelse, når der kun ses på unge, som har fuldført en videregående uddannelse. Det er dermed ikke muligt at understøtte hypotese 2b. Ud fra modellens resultater tyder det på, at unge ikke er afhængige af, hvorvidt deres forældre kan finansiere uddannelse. Derudover kan de manglende signifikante estimerater skyldes, at der er mindre ulighed i Danmark end i USA, hvor der er fundet signifikante resultater. Dermed er forældres økonomiske forhold i ifølge modellerne ikke vigtige for unges uddannelsesbeslutning. Dette tyder desuden på, at forældres indkomst ikke er en god proxy for individers diskonteringsrente.

Hypotese 2c: Mødres Uddannelse

Hypotese 2c er, at der eksisterer en positiv sammenhæng mellem mødres uddannelsesniveau og unges højeste fuldførte uddannelsesniveau, også når der udelukkende ses på unge, som har fuldført en videregående uddannelse. I model 1 er variabelen *mor uddannelse*, MU_{it_0} , positiv og signifikant. Hypotese 2c betegnes som understøttet, hvis dette også er tilfældet i model 2. I tabel 8.2 ses det, at estimatet for MU_{it_0} er positivt og signifikant. Ud fra resultaterne i model 1 og model 2 understøttes hypotese 2c. I model 2 er odds ratios for MU_{it_0} lig 1,097. Dette betyder, at hvis mødres uddannelse stiger med et niveau, stiger sandsynligheden for at unge fuldfører en uddannelse, der er ét niveau højere, med 9,7%, når vedkommende har fuldført en videregående uddannelse og når alle andre faktorer holdes konstant.

Hypotese 2d: Fædres Indkomst

Hypotese 2d er, at der eksisterer en positiv sammenhæng mellem fædres indkomst og unges højeste fuldførte uddannelsesniveau, også når der udelukkende ses på unge, som har fuldført en videregående uddannelse. Det fremgår af kapitel 8.2.1, at der ikke eksisterer en signifikant sammenhæng mellem fædres indkomst og unges højeste fuldførte uddannelsesniveau i model 1. Dermed kan hypotese 2d allerede herudfra betegnes som ikke understøttet. For god ordens skyld, fremgår det dog også af estimatet for variabelen FI_{it_0} i tabel 8.2, at der ikke eksisterer en signifikant sammenhæng, hvis kun unge, der har taget en videregående uddannelse, tages i betragtning. Dermed understøttes hypotese 2d ikke. Dette betyder, at fædres indkomst ikke har en signifikant sammenhæng med unges højeste fuldførte uddannelsesniveau.

Hypotese 2e: Fædres Uddannelse

Hypotese 2e er, at der eksisterer en positiv sammenhæng mellem fædres uddannelsesniveau og unges højeste fuldførte uddannelsesniveau, også når der udelukkende ses på unge, som har fuldført en videregående uddannelse. I undersøgelsen af hypotese 1e ses det, at der eksisterer en positiv sammenhæng mellem variabelen *far uddannelse*, FU_{it_0} , i model 1 hvor alle unge inddrages. I tabel 8.2 fremgår det, at der også eksisterer en positiv signifikant effekt mellem FU_{it_0} og unges højeste fuldførte uddannelsesniveau, når der kun ses på unge der har gennemført en videregående uddannelse. Da der eksisterer en signifikant positiv sammenhæng i både model 1 og model 2 er hypotese 2e understøttet.

Odds ratios for FU_{it_0} er lig med 1,111 i model 2, jf. tabel 8.2. For hver enheds stigning i fædres uddannelsesniveau stiger sandsynligheden for, at unge gennemfører en uddannelse, der er ét niveau højere, med 11,1% givet, at vedkommende har gennemført en videregående uddannelse og når alle andre variable holdes konstant.

8.3 Robusthedstjek ved OLS-estimering

For at sikre robustheden af resultaterne fra henholdsvis model 1 og model 2, estimeres to tilsvarende modeller med OLS-estimering. OLS regressionen for model 1, er navngivet **OLS Model 1** og OLS regressionen for model 2, er navngivet **OLS Model 2**. Et estimat af en variabel vurderes at være robust, hvis estimatet er signifikant og går i samme retning i både den ordered logistiske model og den dertilhørende OLS model. At estimatet går i samme retning betyder, at det enten er positivt eller negativt i begge modeller. Resultaterne af OLS model 1 og OLS model 2 fremgår af tabel J.1 og tabel J.2 i bilag J. Det er undersøgt, hvorvidt OLS modellerne overholder Gauss-Markov antagelserne (Wooldridge, 2016). Umiddelbart har begge modeller problemer med heteroskedasticitet, hvorfor det er robuste standardfejl,

der fremgår af tabellerne i bilag J. Antagelserne samt testsene heraf fremgår af bilag I.

Sammenlignes resultaterne for de forklarende variable i OLS model 1 med dem fra model 1, ses det, at estimaterne for KG_{it_0} , MI_{it_0} , MU_{it_0} og FU_{it_0} er positive og signifikante i begge modeller. Som det fremgår ovenfor, er estimatet for *far indkomst*, FI_{it_0} , ikke signifikant i model 1, hvilket heller ikke er tilfældet i OLS model 1. Dermed stemmer resultaterne for de forklarende variable overens i model 1 og OLS model 1. Dette betyder, at hypotese 1a-1c og 1e understøttes i både model 1 og OLS model 1. Herudfra vurderes resultaterne for hypoteserne i gruppe 1 at være robuste. Ses der på resultaterne for OLS model 2 i forhold til model 2 for de forklarende variable, stemmer disse overens med hinanden for fire af de forklarende variable. Estimaterne for KG_{it_0} , MU_{it_0} og FU_{it_0} er positive og signifikante i begge modeller, mens at estimatet for FI_{it_0} ikke er signifikant i hverken model 2 eller OLS model 2. I model 2 er estimatet for MI_{it_0} ikke signifikant, det er det dog i OLS model 2 på et 10% signifikansniveau. Dette er dog den eneste afvigelse der er mellem henholdsvis model 2 og OLS model 2. Overordnet understøttes hypotese 2a, 2c samt 2e dermed også i OLS modellerne, mens at hypotese 2d ikke understøttes i OLS modellerne, hvilket heller ikke er tilfældet i de ordered logistiske regressioner. Den eneste afvigelse er som nævnt signifikansen for estimatet for MI_{it_0} . Et signifikant estimat for MI_{it_0} i OLS model 2, vurderes dog ikke, at være tilstrækkeligt til at understøtte hypotese 2b, når resultaterne fra model 2 og OLS model 2 er tvetydige.

Det fremgår ovenfor og af bilag I, at variablen KG_{it_0} har problemer med proportional odds assumption i både model 1 og model 2. I OLS modellerne er estimatet for KG_i henholdsvis 0,202 og 0,186 og disse er signifikante på et 1% signifikansniveau. Det vurderes ud fra robusthedstjekket, at der eksisterer en positiv signifikant sammenhæng mellem karakterer og individers højeste fuldførte uddannelsesniveau. Da variablen ikke overholder proportional odds assumption er størrelsen af odds ratio dog usikkert, da det varierer mellem cut-offs. Sammenfattende vurderes det ud fra robusthedstjekket ved brug af OLS modeller, at resultaterne for de forklarende variable i model 1 og model 2 er robuste i forhold til undersøgelse af hypotese 1a-1e og 2a-2e.

8.4 Opsummering af Resultater for Forklarende Variable

Det fremgår af de foregående afsnit, at henholdsvis hypotese 1a-1c samt 1e overholdes og at hypotese 2a, 2c og 2e overholdes. Dette betyder, at der på tværs af alle uddannelsesniveauer findes en robust signifikant positiv sammenhæng mellem unges evner og deres højeste fuldførte uddannelsesniveau, samt forældres uddannelsesniveau og unges højeste fuldførte uddannelsesniveau. Når der ses på mødres indkomst, eksisterer der ikke en signifikant sammenhæng, når der kun ses på unge, som har fuldført en videregående uddannelse. Derfor findes der ikke en signifikant positiv sammenhæng mellem mødres indkomst og unges højeste fuldførte uddannelsesniveau. Der er heller ikke en signifikant sammenhæng mellem fædres uddannelsesniveau og unges højeste fuldførte uddannelsesniveau, både når der ses på alle unge og kun unge, der har gennemført en videregående uddannelse.

8.5 Resultater for Kontrolvariable

I det følgende undersøges resultaterne for kontrolvariablene i henholdsvis model 1 og model 2, i sammenhæng med resultaterne fra OLS modellerne. I model 1 har 10 ud af 20 kontrolvariable et signifikant estimat, ud af disse har 8 variable et signifikant estimat, der også går i samme retning i OLS model 1. Samlet er der 12 kontrolvariable, der har signifikante estimater i OLS model 1. I model 2 har 12 kontrolvariable et signifikant estimat og ud af disse har 7 variable et signifikant estimat, der har samme retning i OLS model 2. Samlet er der 9 kontrolvariable, der har signifikante estimater i OLS model 2. Overordnet er der altså nogle afvigelser mellem resultaterne fra henholdsvis model 1 og OLS model 1 og resultaterne fra model 2 og OLS model 2. I det følgende præsenteres resultater for variable, der er signifikante og har samme retning i både den ordered logistiske regression og den dertilhørende OLS model. Resultaterne for variablene, hvis estimater ikke er signifikante og robuste, præsenteres ikke, da det i modellerne ikke er muligt at påvise signifikante sammenhæng med de uafhængige variable. Estimaterne og odds ratios for alle uafhængige variable, der præsenteres i det følgende, er sammenhængen med den afhængige variabel, når alle andre variable holdes konstant.

Variablen alder, A_{it_0} , har i model 1 et signifikant negativt estimat. Odds ratios for variabelen er lig 0,823. Det betyder, at hvis et individ er 1 år ældre i begyndelsen af det år vedkommende færdiggør grundskolen, der 17,7% $((1 - 0,823) \cdot 100 = 17,7)$ lavere sandsynlighed for, at vedkommende gennemfører en uddannelse, der er ét niveau højere. Dette er umiddelbart overraskende, da Jensen og Nielsen (2010) finder, at der eksisterer en positiv sammenhæng mellem alder og uddannelsespræstation. Ses der kun på unge, der har fuldført en videregående uddannelse, findes der også et negativt forhold. Dette resultat er dog ikke robust. En årsag til, at der findes en negativ sammenhæng, kan være, at de personer, der er ældre har gået et år om eller er kommet et år senere i skole. Dette kan eksempelvis være på grund af sociale problemer eller af hensyn til barnets udvikling, som er vedholdende igennem hele barnets liv (Børne og Undervisningsministeriet, 2022).

Køn, K_{it_0} , har et negativt signifikant estimat i model 1 og odds ratios er lig 0,723. Hvis et individ er dreng/mand, er der dermed 27,7% $((1 - 0,723) \cdot 100 = 27,7)$ lavere sandsynlighed for, at vedkommende gennemfører en uddannelse, der er ét niveau højere. Dette resultat stemmer overens med forventningerne formet på baggrund af den deskriptive statistik, jf. bilag C.1. Der er ikke et robust signifikant estimat for K_{it_0} i model 2, hvorfor det ikke kan vurderes, om der eksisterer samme tendens, når man kun ser på unge, der har gennemført en videregående uddannelse. I model 1 har variabelen problemer med overholdelse af proportional odds assumption, jf. bilag I. Dette betyder, at estimatet varierer på tværs af cut-offs og derfor bør størrelsen heraf ikke tillægges værdi.

Dummy variabelen for etnicitet, E_{it_0} , har et robust positivt signifikant estimat i både model 1 og model 2, jf. tabel 8.1 og tabel 8.2. Dermed eksisterer der en positiv sammenhæng mellem at være af ikke dansk oprindelse og ens højeste fuldførte uddannelse. I model 1 er odds ratios lig 1,380, mens at det i model 2 er 1,544. Det betyder, at unge af ikke dansk oprindelse er 38,0% mere sandsynlige til at fuldføre en uddannelse, der er ét niveau højere. Hvis der kun ses på unge, der har gennemført en videregående uddannelse, er unge af ikke dansk oprindelse 54,4% mere sandsynlige til at gennemføre en uddannelse, der er ét niveau højere. Det fremgår af bilag I, at E_{it_0} har problemer med at overholde proportional odds assumption, hvorfor størrelsen af estimatet ikke bør tillægges værdi. I forhold til den deskriptive statistik er det positive resultat overraskende, jf. kapitel 5.2. Flere empiriske studier, finder dog en positiv sammenhæng mellem at være af ikke dansk oprindelse og unges uddannelsesniveau, når der tages højde for socioøkonomiske faktorer. Herudfra er resultatet som forventet, da modellerne inkluderer socioøkonomiske faktorer.

Hvis man kun ser på unge, der har gennemført en videregående uddannelse, eksisterer der ifølge resultaterne fra model 2 en positiv sammenhæng mellem at bo i en storby- eller hovedstadskommune og unges højeste fuldførte uddannelsesniveau. Odds ratios for dummy variablene storbykommune, $KO3_{it_0}$, og hovedstadskommune, $KO4_{it_0}$, er henholdsvis 1,329 og 1,238. Dermed er individer, der boede i en hovedstadskommune sammenlignet med alle andre kommunetyper 23,8% mere sandsynlige til at fuldføre en uddannelse, der er ét niveau højere, givet, at de har taget en videregående uddannelse. Det fremgår af resultaterne for model 1 og OLS model 1, at der ikke eksisterer en robust signifikant sammenhæng mellem nogen af dummy variablene for kommunetype ($KO1_{it_0}$, $KO2_{it_0}$, $KO3_{it_0}$, $KO4_{it_0}$) når alle unge i populationen tages i betragtning. Den positive sammenhæng mellem at bo i en storby- eller hovedstadskommune er som forventet, jf. bilag C.1.

Uanset om man ser på alle unge eller kun unge, der har gennemført en videregående uddannelse, er der et positivt signifikant estimat for variabelen *mor alder*, MA_{it_0} . Odds ratios for MA_{it_0} er 1,024 i model 1 og 1,015 i model 2, jf. tabel 8.1 og tabel 8.2. Når et individs mor er 1 år ældre, er der dermed 2,4% større sandsynlighed for, at et individ gennemfører en uddannelse, der er ét niveau højere. Hvis der kun ses på individer, der har gennemført en videregående uddannelse, er sandsynligheden for, at de gennemfører en uddannelse, der er ét niveau højere, 1,5% højere, når morens alder stiger med 1 år. Den positive sammenhæng stemmer overens med forventningerne til variabelen, jf. bilag C.1

Ifølge resultaterne fra model 1 eksisterer der en negativ sammenhæng mellem mødres familietype, MF_{it_0} , og unges højeste fuldførte uddannelsesniveau. Hvis mødre er *enlige* er der 11,4% $((1 - 0,886) \cdot 100 = 11,4)$ lavere sandsynlighed for, at unge gennemfører en uddannelse, der er ét niveau højere. Dette resultatet stemmer overens med forventningerne, jf. bilag C.1. Der eksisterer dog ikke en signifikant sammenhæng, hvis man kun ser på unge, der har gennemført en videregående uddannelse, jf. tabel 8.2.

Unge, hvis mødre tilhører socialklassen *overførselsindkomst*, $MS3_{it_0}$, har 19,6% $((1 - 0,804) \cdot 100 = 19,6)$

lavere sandsynlighed for at fuldføre en uddannelse, der er ét niveau højere, sammenlignet med unge, hvis mødre tilhører en af de andre socialklasser. Sammenhængen stemmer overens med forventningen fra bilag C.1. Der eksisterer ikke en signifikant sammenhæng for de andre typer af socialklasser, hverken når der ses på alle unge eller kun unge, der har gennemført en videregående uddannelse, jf. tabel 8.1 og 8.2.

Ifølge model 2 har fædres alder en signifikant sammenhæng med unges uddannelsesniveau, når man kun ser på unge, der har gennemført en videregående uddannelse, jf. tabel 8.2. Den positive sammenhæng er dog forholdsvis lille, da odds ratios for FA_{it_0} er lig med 1,009, jf. tabel 8.2. Når et individs far er 1 år ældre, er sandsynligheden for at vedkommende gennemfører en uddannelse, der er ét niveau højere, 0,9% større givet, at vedkommende har taget en videregående uddannelse.

Ses der på farens socialklasse, eksisterer der i flere tilfælde en sammenhæng mellem denne og unges højeste niveau af uddannelse, jf. tabel 8.1 og tabel 8.2. Der eksisterer en signifikant negativ sammenhæng mellem unge, der har en far hvis socialklasse er *pension*, $FS2_{it_0}$, og de unges højeste uddannelsesniveau. I model 1 og i model 2 er odds ratios henholdsvis 0,822 og 0,797. Dette er sammenlignet med unge, hvis fædre tilhører de andre socialklasser. Dette gælder, uanset om man ser på alle unge eller kun unge, der har gennemført en videregående uddannelse. I forhold til forventningerne, stemmer resultaterne overens hermed, jf. bilag C.1.

Hvis der ses på alle unge, eksisterer der en negativ sammenhæng mellem unge, hvis fædre tilhører socialklassen *overførselsindkomst*, $FS3_{it_0}$, og unges højeste fuldførte uddannelsesniveau. Hvis et individs far tilhører socialklassen *overførselsindkomst*, i forhold til alle andre socialklasser, er der 23,5% lavere sandsynlighed for, at individet gennemfører en uddannelse, der er ét niveau højere. Dette resultat stemmer overens med forventningerne. Der eksisterer ikke en signifikant sammenhæng, når der kun ses på unge, der har gennemført en videregående uddannelse.

Når man kun ser på unge, der har gennemført en videregående uddannelse, er der en positiv sammenhæng mellem far socialklasse, *andre*, $FS4_{it_0}$, og de unges højeste uddannelsesniveau. Det betyder, at hvis et individs far tilhører socialklassen *andre* er sandsynligheden for, at individet fuldfører en uddannelse, der er ét niveau højere, 29,3% større, givet at individet har gennemført en videregående uddannelse. Sammenlignet med forventningerne er dette et overraskende resultat. Gruppen er dog en blandet gruppe, hvorfor den kan være svær at sammenligne med andre socialklasser. Der eksisterer ikke en signifikant sammenhæng, når der ses på alle unge. I hverken model 1 eller model 2 eksisterer der en signifikant sammenhæng mellem unge, der har en far hvis socialklasse er *selvstændig*, $FS1_{it_0}$, og de unges højeste fuldførte uddannelsesniveau.

Sammenfattende eksisterer der sammenhænge mellem flere af kontrolvariablene i både model 1 og model 2 og unges højeste fuldførte uddannelsesniveau. Generelt varierer det, afhængig af om man ser på alle unge eller kun unge, der har gennemført en videregående uddannelse, om dummy variablene *kommunetype*, *mor socialklasse* og *far socialklasse* er signifikante. Der er i regressionsmodellerne ikke en konsistent sammenhæng med, om det er mødre eller fædre, der har signifikante estimater. Derfor er det ud fra den kvantitative analyse ikke muligt at vurdere hvem af forældrene, der har den største indflydelse på unges uddannelsesvalg. Det fremgår af bilag C.1 at disse kontrolvariable har været vanskelige at operationalisere i praksis. Hvorvidt den manglende signifikans skyldes, at operationaliseringerne ikke i tilstrækkelig grad afspejler de underliggende faktorer eller det skyldes, at der simpelthen ikke eksisterer en reel sammenhæng er uklart. På tværs af alle uddannelsesniveauer findes der signifikante og robuste sammenhænge mellem etnicitet, mødres alder samt at have en far, der tilhører socialklassen *pension*. Der er en positiv sammenhæng mellem at være af ikke dansk oprindelse samt mødres alder og unges højeste fuldførte niveau af uddannelse, mens der er en negativ sammenhæng mellem at have en far, der tilhører socialklassen *pension*.

8.6 Vurdering af Regressionsmodellerne

I kapitel 8.3 er der foretaget et robusthedstjek af model 1 og model 2 ved brug af OLS modeller. I indeværende afsnit foretages der yderligere en vurdering af modellerne, der ligger udover hvilken metode,

der bruges til estimering heraf.

For at foretage en overordnet vurdering af modellerne, undersøges forklaringsgraden heraf samt deres goodness-of-fit. For ordered logistiske regressioner eksisterer der et mål for goodness-of-fit, der kaldes AIC, jf. bilag I. For model 1 er dette lig 68864,92 mens at det for model 2 er 42636,32. AIC kan dog ikke sige meget om forklaringsgraden, men kan kun benyttes til sammenligning af modeller (Dziak et al., 2012). Da det ikke er muligt at sige noget ud fra AIC alene, ses der i stedet på $Adj.R^2$ for OLS modellerne. I OLS model 1 er $Adj.R^2 = 27,8$, mens den i OLS model 2 er $Adj.R^2 = 14,1$. Dette betyder, at modellerne kan forklare henholdsvis 27,8% og 14,1% af variationen i den afhængige variabel. Dermed er der en stor del af variationen i unges højeste uddannelsesniveau, der ikke kan forklares af de faktorer, der inddrages i modellen. Det fremgår af kapitel 7, at nogle faktorer, hvor der empirisk er fundet en sammenhæng med uddannelsesvalg, ikke er inddraget i modellen. Herunder bløde faktorer, forhold i grundskolen og individers forventede indkomst.

Både i *schooling* modellen og i flere empiriske studier er der fundet en positiv sammenhæng mellem den forventede indkomst og antal års uddannelse, jf. kapitel 4. Derfor er det nøje overvejet, hvorvidt et mål herfor skulle inddrages i regressionsmodellerne. Det er fravalgt at inddrage et mål herfor, da det ikke har været muligt, at finde en god proxy. Med en *god* proxy, menes der en som måler både den forventede livsindkomst og variationen i indkomsten på uddannelsesniveau, men hvor der ikke er stor risiko for omvendt kausalitet, jf. bilag C.2. En af de overvejede proxies for livsindkomst, er individernes egen indkomst i 2019, hvor der dog forventes at eksistere omvendt kausalitet. Af bilag K fremgår resultaterne fra en model, hvor denne proxy for forventet indkomst inkluderes. Modellernes AIC, der er henholdsvis 68422,55 og 42565,63 er næsten identiske med model 1 og model 2. Sammenlignes $Adj.R^2$ er de henholdsvis 28,9% og 14,4% for model 1 og model 2 hvis indkomst inkluderes, jf. bilag K. Variablen bidrager dermed ikke betydeligt til modellernes forklaringsgrad og da der potentielt eksisterer problemer med omvendt kausalitet udelades den.

I forbindelse med udarbejdelsen af modellerne, er det undersøgt, hvorvidt det har en betydning for resultaterne, hvis nogle af de uafhængige variable havde en anden operationalisering. Det fremgår af bilag F, hvilke andre operationaliseringer, der er overvejet. På baggrund af overvejelserne i bilag F er variablene operationaliseret så mest mulig information bevares.

Der kan eksistere en spuriøs sammenhæng, hvis en faktor, der inddrages i en model, ikke påvirker den afhængige variabel. Den spuriøse sammenhæng kan opstå, fordi en model mangler at tage højde for en underliggende faktor. At inkludere mange uafhængige variable i en regressionsmodel kan også forårsage, at man finder spuriøse sammenhænge. Dette kan medføre, at ægte sammenhænge udvandes og at estimaterne har store standardfejl (National Library of Medicine, 2017). Både model 1 og model 2 indeholder 25 uafhængige variable, hvilket vurderes at være mange. For at minimere risikoen for spuriøse sammenhænge er der foretaget t-tests for hver af de uafhængige variable i forbindelse med opstillingen af regressionsmodellerne, jf. kapitel 7. T-testsene fremgår af bilag E. Alle faktorer, der inddrages i modellerne, er signifikante i t-testsene i model 1. Variablene E_{it_0} og S_{it_0} er dog ikke signifikante i model 2 og t-testene viser derudover, at de har en negativ forklaringsgrad, jf. tabel E.24 og E.26 i bilag E. De inddrages dog i model 2 for at have to sammenlignelige modeller, hvilket er nødvendigt for at kunne undersøge hypoteserne i gruppe 2. Sandsynligheden for at finde spuriøse sammenhænge vurderes at være større ved inddragelse af flere socioøkonomiske faktorer, såsom søskende og forældres socialklasse, indkomst og uddannelsesniveau. De underliggende forklaringer for sammenhængen mellem disse faktorer og unges uddannelsesniveau er, at de hver har indflydelse på mængden af forældres ressourcer i form af tid og penge.

Alle faktorer, der inddrages i regressionsmodellerne, er udvalgt på baggrund af teori og empiriske studier. I forbindelse med vurderingen af, om der kan eksistere spuriøse sammenhænge, er der udarbejdet komprimerede versioner af model 1 og model 2. I det følgende undersøges resultaterne heraf i sammenligning med model 1 og model 2.

8.6.1 Test af Komprimerede Modeller

I de komprimerede modeller er kun de forklarende faktorer, KG_{it_0} , MI_{it_0} , MU_{it_0} , FI_{it_0} , FU_{it_0} , inddraget som uafhængige variable. Af bilag L fremgår resultaterne af de komprimerede modeller, herunder også OLS estimeringer samt odds-ratios og konfidensintervaller. I tilfælde af, at der eksisterer spuriøse sammenhænge mellem de socioøkonomiske faktorer og unges uddannelsesniveau, mindskes risikoen for dette ved ekskluderingen af kontrolvariablene.

Der er også andre fordele ved at undersøge komprimerede modeller. I de komprimerede modeller fjernes nogle af de variable, der ikke overholder proportional odds assumption, jf. bilag I. Komprimerede modeller er også interessante i forhold til parsimoni-princippet eller *Occams Razor*. Dette foreskriver, at "*entities should not be multiplied unnecessarily*", hvilket betyder, at enkelthed foretrækkes (Gibbs, 1996).

Resultaterne af de komprimerede modeller viser, at der eksisterer de samme positive sammenhænge mellem de forklarende faktorer som i model 1 og model 2. Det fremgår af tabel L.1 og tabel L.4 i bilag L, at AIC er lig 69389,69 for den komprimerede model 1 og lig 42774,68 for den komprimerede model 2. Sammenlignes forklaringsgraderne for OLS estimeringerne er de 26,5% og 13,7% for de komprimerede modeller. Dermed forklarer de komprimerede modeller en lille smule mindre end model 1 og model 2.

Selvom inddragelsen af de mange variable i model 1 og model 2 ikke bidrager meget til forklaringsgraden, risikerer man ved udeladelsen heraf, at modellerne bliver misspecificeret. Endvidere er alle variablene i modellerne udvalgt på baggrund af teori og empiriske studier, hvorfor der forventes at eksistere en sammenhæng med unges uddannelsesniveau. På baggrund heraf vurderes det, at hypoteserne på bedst vis undersøges i model 1 og 2, da der kontrolleres for flere faktorer.

8.7 Opsummering

Resultaterne af regressionsmodellerne finder, at der eksisterer positive sammenhænge mellem unges evner, mødres samt fædres uddannelsesniveau og unges højeste fuldførte uddannelsesniveau på tværs af alle uddannelsesniveauer. Til gengæld findes der ikke signifikante estimater for forældres indkomst på tværs af modellerne. Ud fra den kvantitative analyse kan det dermed konkluderes, at der for 2002-kohorten eksisterer en sammenhæng mellem deres evner og højeste fuldførte uddannelsesniveau, samt forældres uddannelsesniveau og deres højeste fuldførte uddannelsesniveau. I det følgende diskuteres modellernes repræsentativitet samt projektets underliggende antagelser. Endvidere vurderes det, hvorvidt resultaterne af analysen kan bidrage til at sandsynliggøre, at der eksisterer en kausal sammenhæng.

Unge Uddannelsesvalg 9

I kapitlet diskuteres det, hvorvidt datagrundlaget for den kvalitative analyse er repræsentativt. Det diskuteres både i forhold til alle, der færdiggjorde grundskolen i 2002, men også i forhold til efterfølgende årganges valg af uddannelsesniveau. Derefter diskuteres rangeringen af uddannelsesniveauerne og antagelsen om at "en bedre uddannelsespræstation øger sandsynligheden for at gennemføre en uddannelse på et højere niveau". Efterfølgende diskuteres det, om de inkluderede socioøkonomiske faktorer allerede er underliggende i økonomiske antagelser eller om forklaringen findes i sociologisk teori. Til sidst vurderes det, om det kan sandsynliggøres, at der er kausale sammenhænge mellem de forklarende variable og de unges højeste fuldførte uddannelsesniveau.

9.1 Repræsentativitet i Datagrundlaget

Som nævnt i kapitel 2.3 er datagrundlaget begrænset til et specifikt forskningsprojekt ved Danmarks Statistiks Forskerservice.

I den kvantitative analyse er der fundet statistiske sammenhænge mellem nogle af de forklarende faktorer og unges højeste fuldførte uddannelsesniveau for 2002-kohorten. De forklarende faktorer, hvor der findes statistiske sammenhænge, er mødres og fædres uddannelsesniveau samt unges evner. For at kunne vurdere om disse sammenhænge også eksisterer for andre kohorter, findes det nødvendigt at sikre, at 2002 kohorten er repræsentativ. Der er i datamodelleringen foretaget systematiske fravalg for at sikre, at alle oplysninger er der for alle personer i den undersøgte population. Alle personer, der var yngre end 14 år og ældre end 18 år ved udgangen af 2001, er systematisk frasorteret i dataet. Det vides ikke, hvad karakteristika for disse personer er, hvorfor konsekvenserne af fravalget er ukendte. Denne frasortering er foretaget, da gennemsnitsalderen ved færdiggørelsen af grundskolen er 15 år, hvorfor karakteristika for individer, der er relativt meget yngre eller ældre end det, forventes at være anderledes (Danmarks Statistik, n.d.h). Forskellige årsager til, at personer kan afvige fra gennemsnitsalderen, er beskrevet i kapitel 2.2. Afvigende karakteristika for denne gruppe kan potentielt medføre, at estimerne er upræcise. Ud over alder er der systematisk frasorteret individer, som kun har en eller ingen juridiske forældre og individer, der har deres eget familie ID og derved ikke deler folkeregisteradresse med deres forældre, da de færdiggjorte grundskolens 9. klasse.

For at vurdere, hvorvidt den anvendte population er repræsentativ og resultaterne dermed kan generaliseres til andre årgange, sammenlignes karakteristika af dataet for 2002-kohorten med senere årgange. Dette inkluderer populationsstørrelse, køn, etnicitet samt karaktergennemsnit. Populationsstørrelsen på afgangselever i 9. klasse var i 2002 55.595 personer. Frem til 2014 har der årligt været en gennemsnitlig stigning på 1.041 personer, der har gennemført 9. klasse. Fordelingen mellem drenge og piger har stort set været uændret, med en gennemsnitlig stigning i 0,05% per år i drenge og et tilsvarende fald i piger. Andelen af personer med dansk oprindelse er i gennemsnit faldet med 0,19%, indvandrere i gennemsnit faldet med 0,16%, hvorimod andelen af efterkommere i gennemsnit er vokset med 0,34% om året (Danmarks Statistik, n.d.x). Udregnes ændringen i karaktergennemsnittet for perioden 2002-2014 på samme måde, som i bilag A er karaktergennemsnittet i gennemsnit steget med 0,07 om året. Dermed er der ikke store forskelle årgangene imellem.

Udover at sammenligne dataet med senere årgange, sammenholdes noget af dataet også med datasæt fra Statistikbanken hos Danmarks Statistik. Sammenlignes det højeste fuldførte uddannelsesniveau for

den population, der er anvendt i projektet, med det højeste fuldførte uddannelsesniveau i 2019 for aldersgrupperne 25-29 år, 30-34 år samt 35-39 år, er det overordnet set repræsentativt, jf. tabel 9.1. Personer fra kohorten, der færdiggjorde grundskolens 9. klasse i 2002, befinder sig hovedsageligt i aldersgruppen 30-34 år, hvorfor der sammenlignes med denne gruppe og omkringliggende aldersgrupper. Den største forskel mellem populationen og det generelle data fra Statistikbanken er i forhold til aldersgruppen 25-29 år. Her er der en større andel, der har grundskole, gymnasial uddannelse samt akademisk bachelor. Derimod er der færre med en erhvervsuddannelse, professionsbachelor samt lang videregående uddannelse. Forskellen kan forklares af, at der er personer i denne aldersgruppe, der stadigvæk er ved at færdiggøre uddannelse på et højere niveau. Ifølge kapitel 2.2 stemmer dette overens med, at mange unge i aldersgruppen 25-29 år stadig er i gang med at færdiggøre uddannelser. Aldersgruppen 30-34 år stemmer i højere grad overens med 2002-kohorten, hvilket indikerer, at kohorten er repræsentativ. Den største afvigelse mellem 2002-kohorten og de undersøgte aldersgrupper er andelen af personer, der har grundskolen som den højeste fuldførte uddannelse. Denne andel er ca. 5 procentpoint højere i tallene fra Danmarks Statistiks Statistikbank. En mulig forklaring er, at personer med grundskole som højeste fuldførte uddannelsesniveau fra Statistikbanken også inkluderer personer, der ikke har nogen uddannelse og samtidig personer, der kun har fuldført grundskolen op til 6. klasse. De grupper udgør til sammen 2,5%. Derudover er klassetrinene 7.-9. sammenlagt, hvorfor det ikke er muligt at tilskrive, hvor stor en del, der har gennemført 9. klasse (?). Derudover inkluderer dataet fra Statistikbanken personer, der på tidspunktet for registreringen er indskrevet på en uddannelse. De selv samme personer er bevidst fravalgt i datasættet for 2002-kohorten.

Tabel 9.1: Oversigt over Uddannelsesniveau i 2019 for 2002-Kohorten og Aldersgrupperne 25-29 År, 30-34 År og 35-39 År

Uddannelsesniveau	Fordeling Kohorten	25-29 år DST	30-34 år DST	35-39 år DST
Grundskole	10,02 %	16,55 %	15,23 %	15,17 %
Gymnasial uddannelse	4,77 %	13,80 %	6,45 %	5,31 %
Erhvervsuddannelse	31,52 %	22,64 %	26,46 %	30,32 %
Kort videregående uddannelse	5,96 %	5,59 %	5,70 %	6,26 %
Akademisk bachelor	2,41 %	7,93 %	2,79 %	2,17 %
Professionsbachelor	22,40 %	18,38 %	20,48 %	19,71 %
Lang videregående uddannelse	22,91 %	16,11 %	22,89 %	21,05 %

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice samt ?

På baggrund af datasæmmensætningen i forhold til køn, etnicitet og størrelse, vurderes det overordnet, at det anvendte datasæt er repræsentativt for den pågældende årgang. Dermed vurderes det, at sammenhængen, der er fundet for de forklarende faktorer i den kvantitative analyse, kan generaliseres til uddannelsesbeslutningen for andre årgange end 2002-kohorten.

Uagtet af, hvor repræsentativ den anvendte population er for fremtidige årgange, er det væsentligt at overveje forklaringsgraden i model 1 og model 2. Som nævnt i kapitel 8.6 er der en stor del af variationen i unges valg af uddannelsesniveau, som ikke forklares af de faktorer, der er inkluderet i den kvantitative analyse. I modellerne tages der eksempelvis ikke højde for uddannelsesreformer eller skift i uddannelsesmæssige tendenser i samfundet. En eventuel videre diskussion vil være meget relevant for et videre arbejde med emnet. Dog som nævnt i kapitel 2.2, afgrænses der fra at kigge på andre faktorer end de neoklassiske, adfærdøkonomiske og socioøkonomiske faktorer. Derfor er denne diskussion udenfor rammerne af dette projekt.

I Danmark siges der at være frit valg af uddannelse, jf. kapitel 1. Det forholder sig dog sådan, at mange uddannelser har krav, der skal opfyldes for at kunne blive optaget. Det kan eksempelvis være krav om gennemførelse af et specifikt uddannelsesniveau, karaktergennemsnit og lignende. Disse restriktioner kan have konsekvenser for unges valg af uddannelsesniveau. Krav om gennemførelse af uddannelsesniveau er primært noget, der rammer det vertikale uddannelsesniveau. Krav om specifikke fag samt karakterkrav er i højere grad noget, der rammer det horisontale niveau, eksempelvis hvorfor nogle kan blive optaget

på økonomiuddannelsen og ikke jura, eller omvendt.

9.2 Rangering af Uddannelsesniveauer

Operationaliseringerne af de afhængige variable i den kvantitative analyse tager, som nævnt i kapitel 2.4.2, udgangspunkt i ISCED skalaen, der benyttes til at rangere uddannelserne. Her er det fundet nødvendigt at afvige fra skalaen og adskille uddannelser, som ISCED skalaen vurderer at være på samme niveau, og opdele dem i separate niveauer. Eksempelvis placeres en gymnasial uddannelse og erhvervsuddannelse i ISCED skalaen på samme niveau, hvor de i projektet placeres på forskellige niveauer. En anden fremgangsmåde end denne kunne være at tage fuldkommen udgangspunkt i ISCED skalaen i rangeringen af uddannelsesniveauer til den kvantitative analyse. ISCED har udarbejdet skalaen i forhold til sværhedsgraden af undervisningen. De vurderer altså, at en gymnasial uddannelse og en erhvervsuddannelse har den samme sværhedsgrad (Børne- og Undervisningsministeriet, 2021). Udelukkende at kigge på sværhedsgraden, vurderes at være et snævert syn på rangeringen af uddannelsesniveauerne. En erhvervsuddannelse er erhvervskompetencegivende i modsætning til en gymnasial uddannelse (Børne- og Undervisningsministeriet, n.d.c). Rangeringen er i projektet foretaget på baggrund af, hvorvidt en uddannelse er erhvervskompetencegivende. Benyttes lønnen som indikation for rangeringen af uddannelsesniveauerne, er det interessant at betragte medianlønnen for de forskellige uddannelsesniveauer. Medianlønnen er ud fra en investeringsbeslutningstankegang relevant at betragte, når niveauer rangeringen. Ifølge *schooling* modellen skal de individer, der gennemfører længere uddannelser kompenseres for den tid og de offeromkostninger, der er forbundet med erhvervsen af det pågældende uddannelsesniveau. Hvis uddannelsesniveauerne betragtes ud fra en sammenlægning af *bef2001*, *udda2001* samt *ind2001* fra Danmarks Statistiks Forskerservice, er medianlønnen for en erhvervsuddannelse højere end for en gymnasial uddannelse, hvilket understøtter den valgte rangering, jf. figur B.1.

ISCED skalaen skelner ikke mellem, hvorvidt en bacheloruddannelse er en akademisk eller en professionsbachelor, de betegnes derimod begge som en mellemlang videregående uddannelse på samme sværhedsniveau (Børne- og Undervisningsministeriet, 2021). De fleste akademiske bacheloruddannelser er ikke erhvervskompetencegivende, hvilket er årsagen til den valgte rangering (Børne- og Undervisningsministeriet, n.d.c). Medianlønnen, der er udregnet på samme vilkår som ovenfor, er højere for personer med en professionsbachelor end for personer, hvor den højeste fuldførte uddannelse er en akademisk bachelor, jf. figur B.1. Ud fra et datateknisk synspunkt, kan der argumenteres for, at professionsbachelor og akademisk bachelor kan lægges sammen. Andelen af personer, der har en akademisk bachelor som højeste fuldførte uddannelsesniveau er kun på 2,4%, jf. figur 5.1. Det betyder altså, at datagrundlaget for dette uddannelsesniveau er meget begrænset, hvorfor udsving i nogle af personernes karakteristika, kan føre til en skævvridning af estimaterne. Dette, sammen med det faktum at akademisk bachelor og professionsbachelor normalt omtales som et niveau, taler for at rangere uddannelserne på samme niveau. Betragtes uddannelsesniveauerne ud fra et tidsmæssigt aspekt, er en gymnasial uddannelse 1,5 år kortere end en erhvervsuddannelse og en akademisk bachelor er normeret et halvt år kortere end en professionsbachelor.

Hvis uddannelsesniveauerne betragtes ud fra medianlønnen og uddannelsestypernes længde, eksisterer samme rangering, som der er foretaget i projektet.

9.3 Bedre Uddannelsespræstation øger Uddannelsesniveauet

En af grundantagelserne i besvarelsen af problemstillingen er, at en bedre uddannelsespræstation øger sandsynligheden for at gennemføre en uddannelse på et højere niveau. Karaktergennemsnit er her inkluderet som en indikation på den unges uddannelsespræstation. Resultaterne fra den kvantitative analyse i kapitel 8 indikerer, at der eksisterer en sammenhæng mellem andre faktorer og unges uddannelsesbeslutning, selvom at der tages højde for unges præstation i grundskolen. Hvis de faktorer, der er inddraget i det empiriske litteraturstudie kun påvirker uddannelsesniveauet gennem uddannelsespræstationen, ville de ikke have en signifikant sammenhæng med uddannelsesniveauet. Det peger på, at der er andre underliggende forklaringer på, hvorfor faktorerne har en signifikant sammenhæng

med de unges uddannelsesniveau. Omvendt kan der argumenteres for, at variablene *mor alder*, *køn* samt individets *alder* er signifikante i begge modeller, fordi at de opfanger noget af forklaringskraften fra *karaktergennemsnit* og det derfor er spuriøse sammenhænge, der er fundet. Det virker dog ikke til at være tilfældet, da estimatet for *karaktergennemsnit* er relativt ens, når model 1 og model 2 sammenlignes med de komprimerede modeller, hvor disse faktorer er udeladt. Det kan argumenteres for at disse faktorer i stedet påvirker de unges selvkontrol og tålmodighed, hvilket kan påvirke fastsættelsen af diskonteringsrenten. Det vides ikke, hvordan diskonteringsrenten fastsættes, men der er fundet sammenhænge mellem den og socioøkonomiske faktorer. Den underliggende forklaring er, at disse påvirker de unges tålmodighed (Becker og Mulligan, 1997).

9.4 Sociologiske Aspekter

I sociologi undersøges samfundet og/eller sociale relationer (Aalborg Universitet, n.d.). Fra et sociologisk synspunkt er der en sammenhæng mellem en persons socioøkonomiske baggrund og vedkommendes uddannelsesbeslutning. Personer forsøger at maksimere deres sociale afkast eller minimere deres risiko. Dette gør de ved at tage lignende uddannelsesvalg, som deres familie og omgangskreds. Derved er det ikke underliggende værdier, der tilegnes gennem opvæksten, men beslutningerne træffes i højere grad ud fra den kontekst den enkelte befinder sig i. Ud fra denne tilgang er de sociologiske faktorer ikke en proxy for diskonteringsrenten, men en forklarende faktorer i sig selv.

Det er i arbejdsmarkedsøkonomi ikke uset at benytte sociologiske aspekter, som socioøkonomiske faktorer til at studere økonomiske fænomener (Rose, 1952). Spørgsmålet er, om de sociologiske aspekter er underliggende i økonomisk teori, som antaget i projektet, eller om det er nødvendigt at inddrage teorierne fra sociologi for at kunne forklare sammenhængen mellem faktorerne og unges højeste fuldførte uddannelsesniveau.

I projektet benyttes forskellige socioøkonomiske faktorer som proxies for diskonteringsrenten. Dette er gjort, da Becker og Mulligan (1997) finder evidens for, at der eksisterer en sammenhæng mellem socioøkonomiske faktorer, som families indkomst og hvor nutidsorienterede individerne er. En højere familieindkomst indikerer en højere tålmodighed, hvilket medfører at individer er mindre nutidsorienterede. Det antages dermed, at tidspræferencen påvirkes af tålmodigheden.

Becker og Mulligan (1997) argumenterer for, at indkomst og kultur påvirker individers diskonteringsrente. Derved argumenteres der for, at det ikke kun er indkomst, der har en påvirkning på fastsættelsen af diskonteringsrenten, men også kulturelle forhold. Nielsen (2012) finder en begrænset kausalitet mellem forældres og unges uddannelsesniveau og argumenterer for, at det er forældrenes underliggende evner, der påvirker de unges uddannelsesniveau og derved ikke det faktum, at forældrene har fuldført en uddannelse på et bestemt niveau. Det taler for, at det er kulturelle forhold, der påvirker værdier eller egenskaber ved individer. Det antages i projektet, at disse værdier er formet i barndommen. Holder argumentationen og antagelsen, vurderes det, at de socioøkonomiske faktorer i den kvantitative analyse fungerer som proxies for diskonteringsrenten og derved kan resultaterne fra den kvantitative analyse understøttes af neoklassisk teori.

Uanset hvad der er den underliggende forklaring bag de socioøkonomiske faktorer, er der empirisk fundet en signifikant sammenhæng mellem forældres uddannelsesniveau og unges uddannelsesniveau, hvor der er teori, der understøtter sammenhængen. For at sandsynliggøre at der eksisterer en kausal sammenhæng, er det dog vigtig at vide, ud fra hvilken teoriramme, der argumenteres, for at den kausale sammenhæng eksisterer.

9.5 Kausale Sammenhænge

I projektet undersøges det hvilke faktorer, der påvirker unges valg af uddannelsesniveau. Resultaterne fra den kvantitative analyse i kapitel 8 er, at der eksisterer en signifikant positiv sammenhæng mellem evner samt forældres uddannelsesniveau og unges højeste fuldførte uddannelsesniveau. Det er ikke muligt at konkludere, hvorvidt denne sammenhæng er kausal ud fra den kvantitative analyse alene. For at gøre det

muligt at konkludere et kausalt forhold, skal alle andre faktorer holdes konstant, hvilket ikke er muligt ved brug af den anvendte økonometriske metode (Wooldridge, 2016). Som tidligere nævnt er der en stor del af variationen i unges højeste fuldførte uddannelsesniveau, der ikke kan forklares i den kvantitative analyse. Der er altså nogle faktorer, der ikke kan holdes konstante, hvorfor kausale forhold ikke kan konkluderes alene på baggrund af disse resultater.

Ifølge teorien eksisterer der en positiv sammenhæng mellem unges evner og deres højeste fuldførte uddannelsesniveau (Borjas, 2016). Derudover er der også empirisk belæg for, at der eksisterer en positiv sammenhæng (Nielsen og Vissing-Jorgensen, 2006; Jensen og Nielsen, 2010). Det vurderes dog i indeværende tilfælde, at empirisk belæg ikke er tilstrækkeligt til at sandsynliggøre en kausal sammenhæng, da resultaterne fra disse studier kan have de samme udfordringer, som den kvantitative analyse foretaget i dette projekt. Der kan dermed være spuriøse sammenhænge eller mangle at blive taget højde for underliggende faktorer. Det vurderes, at det er muligt at sandsynliggøre kausale sammenhænge mellem unges evner og deres højeste fuldførte uddannelse, da der findes en signifikant sammenhæng i den kvantitative analyse, og fordi der ifølge neoklassisk arbejdsmarkedsteori eksisterer en positiv sammenhæng.

I den kvantitative analyse findes en signifikant positiv sammenhæng mellem forældres uddannelsesniveau og unges højeste fuldførte uddannelse. Samtidig argumenteres der indenfor nogle grene af økonomien for, at socioøkonomiske faktorer er underliggende i den neoklassiske teoris antagelser (Rose, 1952). Dette betyder dermed, at forældres uddannelsesniveau kan fungere som en proxy for diskonteringsrenten. Dermed eksisterer der også ifølge teorien en positiv sammenhæng mellem forældres uddannelsesniveau, i form af diskonteringsrenten, og unges valg af uddannelsesniveau. Da der både i den kvantitative analyse findes en signifikant sammenhæng og at denne også eksisterer ifølge neoklassisk teori, vurderes det, at det kan sandsynliggøres, at der eksisterer en positiv kausal sammenhæng mellem forældres uddannelse og unges valg af uddannelsesniveau.

Den kvantitative analyse finder ikke, at der er en signifikant positiv sammenhæng mellem forældres indkomst og unges højeste fuldførte uddannelsesniveau. Derfor vurderes det, at en kausal sammenhæng ikke kan sandsynliggøres. Dette er overraskende, taget i betragtning af, at det kan argumenteres for, at forældrenes indkomst kan bruges som en proxy for diskonteringsrenten på samme vis som deres uddannelsesniveau. Det er dog ikke muligt i dette projekt, at finde empirisk belæg for sammenhængen mellem forældres indkomst og unges uddannelsesniveau. At det ikke er muligt, er ikke ensbetydende med, at der ikke eksisterer en kausal sammenhæng. Den kausale sammenhæng kan dog ikke påvises i dette projekt.

9.6 Opsummering

Det vurderes at 2002-kohorten er repræsentativ og generaliserbar for andre årganges uddannelsevalg. Denne vurdering er foretaget på baggrund af populationsstørrelsen, sammensætningen af køn, etnicitet og karaktergennemsnit. Dermed vurderes det, at der eksisterer en positiv sammenhæng mellem unges evner, samt deres mødres og fædres uddannelsesniveau og de unges højeste fuldførte uddannelsesniveau for efterfølgende årgange. Sammenhængene der er fundet i den kvantitative analyse er understøttet af neoklassisk teori og andre empiriske studier, der har undersøgt uddannelsevalget i både Danmark og andre lande. At der teoretisk er samme positive sammenhænge sandsynliggør, at der eksisterer kausale sammenhænge.

I projektet undersøges det hvilke faktorer, der påvirker unges valg af uddannelsesniveau, hvor påvirker defineres som, at der kan sandsynliggøres en kausal sammenhæng.

Ud fra neoklassisk teori og den kvantitative analyse sandsynliggøres en positiv kausal sammenhæng mellem unges evner og deres valg af uddannelsesniveau. Derudover sandsynliggøres en positiv kausal sammenhæng mellem forældres uddannelsesniveau og unges valg af højeste fuldførte uddannelsesniveau.

Den kvantitative analyse består af to ordered logistisk regressioner, model 1 og model 2. Analysen tager endvidere udgangspunkt i en kohorte, der gennemførte grundskolens 9. klasse i 2002. Resultaterne af analysen vurderes at være generaliserbare, da målbare faktorer for de omkringliggende årgange er relativt identiske med 2002-kohorten. I regressionsmodellerne inddrages forklarende faktorer samt kontrolvariable, udvalgt på baggrund af neoklassisk teori og empiriske resultater. Ifølge den neoklassiske arbejdsmarkedsteori, *schooling* modellen, bestemmes uddannelsesbeslutninger af individers iboende evner og diskonteringsrente. I teorien medfører et højere niveau af evner, at man tilegner sig mere uddannelse. Omvendt medfører en højere diskonteringsrente, at man tilegner sig mindre uddannelse.

Der er empirisk belæg for, at der eksisterer en positiv sammenhæng mellem evner målt ved proxies og gennemførelse af uddannelse. I den kvantitative analyse operationaliseres evner som unges karaktergennemsnit fra grundskolen og der findes en positiv sammenhæng mellem variabelen og unges højeste fuldførte niveau af uddannelse. Den positive sammenhæng er signifikant, uanset om man ser på alle unge eller kun unge, der har gennemført en videregående uddannelse.

Der er empirisk belæg for, at børn gennemfører uddannelse på samme niveau som deres forældre. Der findes i den kvantitative analyse en positiv signifikant sammenhæng mellem både mødres og fædres uddannelsesniveau og unges højeste fuldførte uddannelsesniveau, både når man ser på alle unge og kun unge, der har gennemført en videregående uddannelse. Socioøkonomiske faktorer kan anvendes som proxy variable for diskonteringsrenten. I sammenhæng med forældres uddannelse betyder dette, at hvis forældres uddannelse er højere, er individets diskonteringsrente lavere. En lavere diskonteringsrente betyder, at man tilegner sig mere uddannelse. Dermed er resultaterne fra den kvantitative analyse, for både unges evner og deres forældres uddannelsesniveau, understøttet af neoklassisk arbejdsmarkedsteori.

Da der både findes signifikante sammenhænge i den kvantitative analyse og at disse sammenhænge understøttes af neoklassisk arbejdsmarkedsteori, konkluderes det i projektet, at det kan sandsynliggøres, at der eksisterer kausale sammenhænge mellem evner og unges valg af uddannelsesniveau samt forældres uddannelsesniveau og unges valg af uddannelsesniveau. Dermed påvirker evner og deres forældres uddannelsesniveau, unges valg af uddannelsesniveau.

Ud fra den kvantitative analyse kan det også konkluderes, at der eksisterer signifikante og robuste sammenhænge mellem etnicitet, mødres alder samt at have en far, der tilhører socialklassen *pension*, både når der ses på alle unge og når der kun ses på unge, der har gennemført en videregående uddannelse. Der er en positiv sammenhæng mellem at være af ikke dansk oprindelse samt mødres alder og unges højeste fuldførte niveau af uddannelse, mens der er en negativ sammenhæng mellem at have en far, der tilhører socialklassen *pension*. Det er dog ikke muligt ud fra projektet at sandsynliggøre, at disse sammenhænge er kausale.

Det følgende kapitel giver et indblik i forfatterens refleksioner over, hvad resultaterne kan anvendes til. Derudover beskrives der eventuelt videre arbejde med problemstillingen.

I projektet sandsynliggøres det, at der eksisterer kausale sammenhænge mellem unges evner og deres valg af uddannelsesniveau, men også mellem forældres uddannelse og unges valg af uddannelsesniveau. Det konkluderes dermed, at disse faktorer påvirker unges valg af uddannelsesniveau. Ses resultaterne i et samfundsøkonomisk perspektiv, er de interessante, da viden om hvilke faktorer, der påvirker unges valg af uddannelse, er værdifuld, når det skal besluttes, hvor investeringerne i uddannelse skal placeres. Resultaterne fra projektet er primært interessante i sammenhæng med uddannelsespolitiske mål, der omhandler niveauer, da uddannelsesvalget i projektet undersøges i et vertikalt perspektiv. Beslutningen om, hvor investeringerne i uddannelse skal placeres, tages ultimativt ud fra hvilket mål man ønsker at opnå. Resultaterne fra projektet taler dog for, at det er relevant at tage socialpolitik i betragtning i forhold til uddannelsesbeslutninger.

Selvom at der i projektet konkluderes, at unges valg af uddannelse påvirkes af deres evner og forældres uddannelsesniveau, viser resultaterne også, at der eksisterer andre signifikante sammenhænge. Derfor er et videre arbejde med problemstillingen interessant.

Ud fra et moderne mikroøkonometrisk perspektiv er det interessant at undersøge resultaterne af en OLS-estimering, hvor alle syv niveauer af uddannelse undersøges i en model. I et sådant perspektiv er det muligt at undersøge alle niveauer af uddannelse på en gang, uden opdeling af modellerne. Som det fremgår af projektet, er dette dog ikke en korrekt metode til estimering af en ordinal variabel, men kan dog fungere som en approksimering.

Det er endvidere også interessant at arbejde videre med at undersøge robustheden af modellerne der inddrages i projektet. En tilgang hertil kan være at undersøge, hvorvidt det påvirker resultaterne at ekskludere variablene fra modellerne enkeltvis. Det har ikke været muligt indenfor rammerne af projektet, men vil være relevant i et videre arbejde.

I projektet argumenteres der for, at socioøkonomiske faktorer er underliggende faktorer, der bestemmer diskonteringsrenten. Et videre arbejde med dette perspektiv kan være at undersøge, hvorvidt man ud fra socioøkonomiske faktorer, kan bestemme individets diskonteringsrente. Dette forventes, at være et matematisk kompliceret arbejde. Endvidere kræver det, at det er muligt at vægte forskellige socioøkonomiske faktorer betydning for fastsættelsen af diskonteringsrenten. Hvorvidt det er muligt i praksis er uklart, men i et teoretisk perspektiv er det interessant.

Litteratur

Aalborg Universitet (n.d.), 'Bachelor sociologi'. lokaliseret d. 31.05.2022.

URL: <https://www.aau.dk/uddannelser/bachelor/sociologidet-s%C3%A6rlige-ved-sociologi>

Adam Petrie (n.d.), 'VIF'. lokaliseret d. 09.05.2022.

URL: www.rdocumentation.org/packages/regclass/versions/1.6/topics/VIF

Ahrenkiel, L. (n.d.), 'Positivism'. lokaliseret d. 01.06.2022.

URL: <https://laeremiddel.dk/viden-og-vaerktoejer/videnskabsteori/videnskabsteoretiske-retninger/positivisme/>

Andersen, T. M., Bentsen, K. H., Kromann, L., Munch, J. R., Schaur, G., Sørensen, A. og Walldorf, J. (2018), *Afkast af uddannelse. Det samfundsmæssige og individuelle rationale*, Syddansk universitetsforlag.

Angrist, J. D. og Pischke, J. (2008), 'Mostly Harmless Econometrics: An Empiricist's Companion'.

Arbejderbevægelsens Erhvervsråd (2009), 'Store Samfundsøkonomiske Gevinster af Uddannelse'.

Asri, M. N. M., Hashim, N. H., Desa, W. og Ismail, D. (2016), 'Pearson Product Moment Correlation (PPMC) and Principal Component Analysis (PCA) for objective comparison and source determination of unbranded black ballpoint pen inks', *Australian Journal of Forensic Sciences* 1, 1–18.

Becker, G. S. og Mulligan, C. B. (1997), 'The endogenous determination of time preference', *The Quarterly Journal of Economics* 112(3), 729–758.

Becker, R. og Hecken, A. E. (2009), 'Higher education or vocational training? An empirical test of the rational action model of educational choices suggested by Breen and Goldthorpe and Esser', *Acta Sociologica* 52(1), 25–45.

Bénabou, R. og Tirole, J. (2002), 'Self-confidence and personal motivation', *The quarterly journal of economics* 117(3), 871–915.

Benders, D. (2013), 'Role of socioeconomics in education', *International Journal of Social* .

Borjas, G. (2016), *Labor economics*, McGraw-Hill Education, New York, NY.

Bosworth, D. og Ford, J. (1985), 'Income expectations and the decision to enter higher education', *Studies in Higher Education* 10(1), 21–31.

Børne- og Undervisningsministeriet (2020), 'Muligheder for videregående uddannelse'. lokaliseret d. 04.05.2022.

URL: <https://www.uvm.dk/erhvervsuddannelser/uddannelser/overblik/muligheder-for-videregaende-uddannelse>

Børne- og Undervisningsministeriet (2021), 'Det ordinære uddannelsessystem'. lokaliseret d. 26.05.2022.

URL: <https://www.uvm.dk/uddannelsessystemet/overblik-over-det-danske-uddannelsessystem/det-ordinaere-uddannelsessystem>

Børne- og Undervisningsministeriet (2022a), 'Den Uddannelsespolitiske Målsætning'. lokaliseret d. 26.05.2022.

URL: <https://www.uvm.dk/forberedende-grunduddannelse/skoleudvikling/baggrund-for-fgu/den-uddannelsespolitiske-maalsaetning>

Børne- og Undervisningsministeriet (2022b), 'Klare Mål'. lokaliseret d. 26.05.2022.

URL: <https://www.uvm.dk/erhvervsuddannelser/skoleudvikling/klare-maal/om-klare-maal>

Børne- og Undervisningsministeriet (n.d.a), 'Andel af en ungdomsårgang der forventes at få en uddannelse'. lokaliseret d. 20.04.2022.

URL: <https://uddannelsesstatistik.dk/Pages/Reports/1865.aspx>

Børne- og Undervisningsministeriet (n.d.b), 'Ord og begreber'. lokaliseret d. 06.05.2022.

URL: <https://www.ug.dk/efteruddannelse/ord-og-begreber>

Børne- og Undervisningsministeriet (n.d.c), 'UddannelsesGuiden'. lokaliseret d. 04.05.2022.

URL: <https://www.ug.dk/>

Børne og Undervisningsministeriet (2013), 'OECD: Danmark Investerer mest i Uddannelse'. lokaliseret d. 26.05.2022.

URL: <https://ufm.dk/aktuelt/pressemeddelelser/2013/oecd-danmark-investerer-mest-i-uddannelse>

Børne og Undervisningsministeriet (2022), 'Skolestart og Undervisningspligt'. lokaliseret d. 27.05.2022.

URL: <https://www.uvm.dk/folkeskolen/fag-timetale-og-overgange/skolestart-og-boernehaveklassen/skolestart>

Carneiro, P., Hansen, K. T. og Heckman, J. J. (2003), 'Estimating distributions of treatment effects with an application to the returns to schooling and measurement of the effects of uncertainty on college'.

Danmarks Statistik (2017), '83 pct. af bachelorer går direkte videre til kandidat'. lokaliseret d. 25.04.2022.

URL: <https://www.dst.dk/da/Statistik/nyheder-analyser-publ/nyt/NytHtml?cid=24238>

Danmarks Statistik (2018a), 'Inddeling af Danmarks kommuner'.

Danmarks Statistik (2018b), 'Indkomster for personer 2017'. lokaliseret d. 19.05.2022.

URL: <https://www.dst.dk/Site/Dst/Udgivelser/nyt/GetPdf.aspx?cid=28214>

Danmarks Statistik (2019a), 'Højt uddannelsesniveau blandt kinesere'. lokaliseret d. 19.05.2022.

URL: <https://www.dst.dk/da/Statistik/nyheder-analyser-publ/nyt/NytHtml?cid=30123>

Danmarks Statistik (2019b), 'Indvandrere i Danmark'. lokaliseret d. 05.05.2022.

URL: <https://www.dst.dk/Site/Dst/Udgivelser/GetPubFile.aspx?id=29446sid=indv2019>

Danmarks Statistik (2020), 'Unge flytter senest hjemmefra i de store byer'. lokaliseret d. 19.05.2022.

URL: <https://www.dst.dk/da/Statistik/nyheder-analyser-publ/nyt/NytHtml>

Danmarks Statistik (n.d.a), 'ALDER'. lokaliseret d. 06.05.2022.

URL: <https://www.dst.dk/da/Statistik/dokumentation/Times/moduldata-for-befolkning-og-valg/alder>

Danmarks Statistik (n.d.b), 'ANTB00'. lokaliseret d. 15.05.2022.

URL: <https://www.dst.dk/extranet/staticsites/TIMES3/html/dec9db1b-1d33-4ee9-b14c-ae3bd980401c.htm>

Danmarks Statistik (n.d.c), 'BEDOEMMELSESFORM'. lokaliseret d. 13.05.2022.

URL: <https://www.dst.dk/da/Statistik/dokumentation/Times/uddannelsesdata/karakterer-fag-i-grundskolen-/bedoemmelsesform>

Danmarks Statistik (n.d.d), 'FAMILIE_ID'. lokaliseret d. 12.05.2022.

URL: <https://www.dst.dk/da/Statistik/dokumentation/Times/familieindkomst/familie-id>

Danmarks Statistik (n.d.e), 'FAMILIE_TYPE'. lokaliseret d. 04.05.2022.

URL: <https://www.dst.dk/da/Statistik/dokumentation/Times/cpr-oplysninger/familier-og-husstande/familie-type>

Danmarks Statistik (n.d.f), 'FAR_ID'. lokaliseret d. 04.05.2022.

URL: <https://www.dst.dk/da/Statistik/dokumentation/Times/moduldata-for-befolkning-og-valg/far-id>

- Danmarks Statistik (n.d.g), 'FOLK1A: Folketal den 1. i kvartalet efter område, køn, alder og civilstand'. lokaliseret d. 05.05.2022.
URL: <https://www.statistikbanken.dk/FOLK1A>
- Danmarks Statistik (n.d.h), 'Fuldførte i grundskolen efter alder, herkomst, national oprindelse, køn, institutionstype og område'. lokaliseret d. 25.04.2022.
URL: <https://www.statistikbanken.dk/U39>
- Danmarks Statistik (n.d.i), 'Grundskole'. lokaliseret d. 15.03.2022.
URL: <https://www.dst.dk/da/TilSalg/Forskningsservice/Dokumentation/hoejkvalitetsvariable/>
- Danmarks Statistik (n.d.j), 'GRUNDSKOLEKARAKTER'. lokaliseret d. 13.05.2022.
URL: <https://www.dst.dk/da/Statistik/dokumentation/Times/uddannelsesdata/karakterer-fag-i-grundskolen-/grundskolekarakter>
- Danmarks Statistik (n.d.k), 'HFAUDD'. lokaliseret d. 15.05.2022.
URL: <https://www.dst.dk/da/Statistik/dokumentation/Times/moduldata-for-uddannelse-og-kultur/hfaudd>
- Danmarks Statistik (n.d.l), 'IE_TYPE'. lokaliseret d. 09.05.2022.
URL: <https://www.dst.dk/da/Statistik/dokumentation/Times/cpr-oplysninger/ie-type>
- Danmarks Statistik (n.d.m), 'KOEN'. lokaliseret d. 09.05.2022.
URL: <https://www.dst.dk/da/Statistik/dokumentation/Times/CPR-oplysninger/BEF-PERSONNUMMER/KOEN>
- Danmarks Statistik (n.d.n), 'KOM'. lokaliseret d. 28.04.2022.
URL: <https://www.dst.dk/da/Statistik/dokumentation/Times/moduldata-for-befolkning-og-valg/kom>
- Danmarks Statistik (n.d.o), 'Kommunegrupper, v1:2018-'. lokaliseret d. 26.04.2022.
URL: <https://www.dst.dk/da/Statistik/dokumentation/nomenklaturer/kommunegrupper>
- Danmarks Statistik (n.d.p), 'Kursusdeltagelse ved voksen og efteruddannelse efter enhed, tidsangivelse, køn, bopælsområde, uddannelsesområde, tid og alder'. lokaliseret d. 06.05.2022.
URL: <https://www.statistikbanken.dk/VEU20>
- Danmarks Statistik (n.d.q), 'MOR_ID'. lokaliseret d. 04.05.2022.
URL: <https://www.dst.dk/da/Statistik/dokumentation/Times/moduldata-for-befolkning-og-valg/mor-id>
- Danmarks Statistik (n.d.r), 'PNR'. lokaliseret d. 12.05.2022.
URL: <https://www.dst.dk/da/Statistik/dokumentation/Times/forskningsservice/pnr>
- Danmarks Statistik (n.d.s), 'QSPLINDK'. lokaliseret d. 04.05.2022.
URL: <https://www.dst.dk/da/Statistik/dokumentation/Times/personindkomst/qsplindk>
- Danmarks Statistik (n.d.t), 'SKALA'. lokaliseret d. 13.05.2022.
URL: <https://www.dst.dk/da/Statistik/dokumentation/Times/uddannelsesdata/karakterer-hele-uddannelser/skala>
- Danmarks Statistik (n.d.u), 'SKOLEAAR'. lokaliseret d. 13.05.2022.
URL: <https://www.dst.dk/da/Statistik/dokumentation/Times/uddannelsesdata/karakterer-fag-i-grundskolen-/skoleaar>
- Danmarks Statistik (n.d.v), 'SOCIO13'. lokaliseret d. 04.05.2022.
URL: <https://www.dst.dk/da/Statistik/dokumentation/Times/Personindkomst/SOCIO13>
- Danmarks Statistik (n.d.w), 'Tilgang af elever, grundskolen efter uddannelse, alder, herkomst, national oprindelse, køn, institutionstype og område'. lokaliseret d. 25.04.2022.
URL: <https://www.statistikbanken.dk/U29>

- Danmarks Statistik (n.d.x), 'U3907: Fuldførte elever, grundskolen efter uddannelse, alder, herkomst, national oprindelse, køn, institutionstype og område (AFSLUTTET)'. lokaliseret d. 26.05.2022.
URL: <https://www.statistikbanken.dk/statbank5a/SelectVarVal/Define.asp?MainTable=U3907>
- Danmarks Statistik (n.d.y), 'Vi efteruddanner os mindre end for fem år siden'. lokaliseret d. 06.05.2022.
URL: <https://www.dst.dk/da/Statistik/nyheder-analyser-publ/nyt/NytHtml?cid=18449>
- Dansk Arbejdsgiverforening (2019), 'Danske virksomheder efteruddanner for 15 mia. kr. om året'. lokaliseret d. 06.05.2022.
URL: <https://www.da.dk/politik-og-analyser/uddannelse-og-kompetencer/2019/danske-virksomheder-efteruddanner-for-15-mia.-om-aaret/>
- Duckworth, A. L. og Seligman, M. E. (2006), 'Self-discipline gives girls the edge: Gender in self-discipline, grades, and achievement test scores.', *Journal of educational psychology* 98(1), 198.
- Dziak, J. J., Coffman, D. L., Lanza, S. T., Li, R. og Jermiin, L. S. (2012), *Sensitivity and Specificity of Information Criteria*, Vol. 12, The Pennsylvania State University.
- DØRS (2003), 'Dansk Økonomi efterår 2003. KAPITEL II - UDDANNELSE'.
- Eriksen, J. (2021), Opportunities Education: Intergenerational Mobility, and Vocational and Continuing Education, PhD thesis. Dissertation not published.
- Euser, A. M., Zoccali, C., Jager, K. J. og Dekker, F. W. (2009), 'Cohort studies: prospective versus retrospective', *Nephron Clinical Practice* 113(3), c214–c217.
- Farooq, M. S., Chaudhry, A. H., Shafiq, M. og Berhanu, G. (2011), 'Factors affecting students' quality of academic performance: a case of secondary school level', *Journal of quality and technology management* 7(2), 1–14.
- Gibbs, P. (1996), 'What is Occam's Razor?'. lokaliseret d. 26.05.2022.
URL: www.math.ucr.edu/home/baez/physics/General/occam.html
- Glen, S. (2014), 'Ordered logistic regression; ordinal/ordered logit'. lokaliseret d. 26.04.2022.
URL: <https://www.statisticshowto.com/ordered-logistic-regression/>
- Golsteyn, B. H., Grönqvist, H. og Lindahl, L. (2014), 'Adolescent time preferences predict lifetime outcomes', *The Economic Journal* 124(580), F739–F761.
- Gottfries, N. (2013), *Macroeconomics*, Red Globe Press, London, England.
- Greene, W. H. (2018), 'Econometric Analysis', p. 1126.
- Hallmann, C. og Carstensen, M. (2020), 'Optag 2020: Antal uddannelsespladser på de videregående uddannelser'.
- Hanushek, E. A. (2006), 'School resources', *Handbook of the Economics of Education* 2, 865–908.
- Harrell Jr, F. (2001), 'Regression modeling strategies: with application to linear models, performance measure used in the evaluation of the discriminative logistic regression, and survival analysis'.
- Hartog, J. og Diaz-Serrano, L. (2007), 'Earnings risk and demand for higher education: A cross-section test for Spain', *Journal of Applied Economics* 10(1), 1–28.
- Heinesen, E. (2010), 'Estimating class-size effects using within-school variation in subject-specific classes', *The Economic Journal* 120(545), 737–760.
- Houliand, C. (2018), 'Djof tror ikke på mange job til bachelorer'. lokaliseret d. 26.04.2022.
URL: <https://www.altinet.dk/artikel/djoef-tror-ikke-paa-mange-job-til-bachelorer>

- Indenrigs- og boligministeriet (n.d.a), 'Kommunalreformen i 2007'. lokaliseret d. 15.05.2022.
URL: <https://im.dk/arbejdsomraader/kommunal-og-regionaloekonomi/kommunale-opgaver-og-struktur/kommunalreformen-i-2007>
- Indenrigs- og boligministeriet (n.d.b), 'Oversigt over de fem regioner og 98 nye kommuner'.
- Jaeger, M. M. (2007), 'Economic and social returns to educational choices: Extending the utility function', *Rationality and Society* 19(4), 451–483.
- Jensen, V. og Nielsen, L. (2010), 'Veje til ungdomsuddannelse 1–statistiske analyser af folkeskolens betydning for unges påbegyndelse og gennemførelse af en ungdomsuddannelse', *København: SFI–Det Nationale Forskningscenter for Velfærd* 10, 24.
- Koch, A., Nafziger, J. og Nielsen, H. S. (2015), 'Behavioral economics of education', *Journal of Economic Behavior & Organization* 115, 3–17.
- Krueger, A. B. (2003), 'Economic considerations and class size', *The economic journal* 113(485), F34–F63.
- Lacour, M. og Tissington, L. D. (2011), 'The effects of poverty on academic achievement', *Educational Research and Reviews* 6(7), 522–527.
- Llena-Nozal, A., Martin, N. og Murtin, F. (2019), 'The economy of well-being: Creating opportunities for people's well-being and economic growth'.
- Long, J. S. (1997), 'Regression models for categorical and limited dependent variables: Advanced quantitative techniques in the social sciences series'.
- Long, J. S. og Freese, J. (2006), *Regression models for categorical dependent variables using Stata*, Vol. 7, Stata press.
- Lærd Statistics (n.d.), 'Pearson Product-Moment Correlation'. lokaliseret d. 19.05.2022.
URL: <https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php>
- Mankiw, N. G. og Taylor, M. P. (2017), *Microeconomics*, Cengage Learning.
- McNulty, K. (2021), *Handbook of Regression Modeling in People Analytics: With Examples in R and Python*, Chapman and Hall/CRC.
- Minister for Ligestilling (2006), 'Kvindes og mænds uddannelser og job - Hvordan bløder vi op på det kønsopdelte uddannelses- og erhvervsvalg og det kønsopdelte arbejdsmarked?'.
- Ministeriet for Forskning, Innovation og Videregående Uddannelser, Økonomi- og Indenrigsministeriet, Finansministeriet, Erhvervs- og Vækstministeriet, Beskæftigelsesministeriet, Ministeriet for Børn og Undervisning (2012), 'Kommissorium for Ministerudvalget og Videregående Uddannelser'.
- Munk, M. D. og Mattsson, C. (2008), 'Social uddannelsesmobilitet på kandidat- og forskeruddannelser'.
- Møberg, R. J. (2021), 'Lektion 2: Introduktion til forløbsanalyse', *Aalborg Universitet* . lokaliseret d. 12.05.2022.
- Møller, E. B. (2019), 'Minister: Uacceptabelt at så mange ikke tager alle 9.-klasse-prøver'. lokaliseret d. 16.05.2022.
URL: <https://www.folkeskolen.dk/afgangsprover-karakterer-skolen-i-samfundet/minister-uacceptabelt-at-sa-mange-ikke-tager-alle-9-klasse-prover/352121>
- National Library of Medicine (2017), 'Common pitfalls in statistical analysis: Logistic regression'. lokaliseret d. 19.05.2022.
URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5543767/>
- Newbold, P., Carlson, W. L. og Thorne, B. M. (2013), *Statistics for business and economics*, Pearson.

- Nielsen, H. P. og Faber, S. T. (2016), Udkantsdanmark: Kampen om stedet og ordene, in 'Den globale udkant', Aalborg Universitetsforlag.
- Nielsen, H. S. (2012), 'Et økonomisk perspektiv på uddannelsesindsatser'.
- Nielsen, H. S. og Vissing-Jørgensen, A. (2006), 'The impact of labor income risk on educational choices: Estimates and implied risk aversion', *Unpublished Manuscript, Department of Economics, University of Aarhus*.
- Nielsen, K. M. R. (2018), 'Afkast og risiko ved at tage en uddannelse. Vælger unge uddannelse i et investeringsperspektiv?'
- O'Connell, A. A. (2006), *Logistic regression models for ordinal response variables*, Vol. 146, sage.
- Olesen, F. og Pedersen, K. (2007), 'Den Økonomiske Teoris Tødder fra Aristoteles til Lucas', 1.
- Oreopoulos, P. (2007), 'Do dropouts drop out too soon? wealth, health and happiness from compulsory schooling', *Journal of public Economics* 91(11-12), 2213–2229.
- Parry, S. (2016), 'Ordinal logistic regression models and statistical software: What you need to know', *Cornell University Statistical Consulting Unit*.
- PAULSEN, L. S. (2022), '29 reformer på 20 år: Universiteterne mangler en reformpause'. lokaliseret d. 29.05.2022.
URL: <https://www.akademikerbladet.dk/aktuelt/2022/april/29-reformer-paa-20-aar-universiteterne-mangler-en-reformpause>
- Petersen, J. (2021), 'Det Samfund Vi Ønsker Os'. lokaliseret d. 26.05.2022.
URL: <https://ufm.dk/ministeriet/ministeren/taler/det-samfund-vi-onsker-os>
- Pihl, M. D. (2007), 'Hver anden ung går i fars eller mors fodspor'. lokaliseret d. 18.03.2022.
URL: <https://www.ae.dk/analyse/2017-01-hver-anden-ung-gaar-i-fars-eller-mors-fodspor/>
- Produktivitetskommissionen (2014), *Uddannelse og innovation. Analyserapport 4*.
- Raza, H. (2019a), 'Forelæsningslides fra Økonometri 1, lektion 5+6'.
- Raza, H. (2019b), 'Forelæsningslides fra Økonometri 1, lektion 8'.
- Reardon, S. et al. (2012), 'The widening academic achievement gap between the rich and the poor', *Community Investments* 24(Sum), 19–2438.
- Robitzsch, A. (2020), Why ordinal variables can (almost) always be treated as continuous variables: Clarifying assumptions of robust continuous and ordinal factor analysis estimation methods, in 'Frontiers in Education', Vol. 5, Frontiers, p. 177.
- Rose, A. M. (1952), 'The potential contribution of sociological theory and research to economics', *The American Journal of Economics and Sociology* 12(1), 23–34.
- Samson, A. (2019), *The Behavioral Economics Guide 2019 (with an Introduction by Uri Gneezy)*, Behavioral Science Solutions Ltd. lokaliseret d. 29.03.2022.
URL: <https://www.behavioraleconomics.com>
- Skole og forældre (2020), 'Børnehaveklasse'. lokaliseret d. 20.04.2022.
URL: <https://www.skole-foraeldre.dk/artikel/b%C3%B8rnehaveklasse>
- Slottved, M. (2021), 'Styrk de mest udfordrede folkeskoler og få flere faglærte'. lokaliseret d. 21.03.2022.
URL: <https://www.vive.dk/da/udgivelser/styrk-de-mest-udfordrede-folkeskoler-og-faa-flere-faglaerte-16765//>
- Sonne-Ragans, V. (2012), *Anvendt videnskabsteori*, Samfundslitteratur.

- Thomsen, A. T. (2017), 'Unge kan vælge mellem 880 uddannelser'. lokaliseret d. 29.03.2022.
URL: <https://www.dr.dk/ligetil/unge-kan-vaelge-mellem-880-uddannelser>
- Thomsen, J.-P. og Bom, L. H. (2019), 'Danske unges uddannelsesforventninger: Med særligt fokus på unge fra kortuddannede hjem', *VIVE-Det Nationale Forsknings-og Analysecenter for Velfærd*.
- UCLA: Statistical Consulting Group (n.d.a), 'Ordered Logistic Regression | R Data Analysis Example'. lokaliseret d. 01.05.2022.
URL: <https://stats.oarc.ucla.edu/r/dae/ordinal-logistic-regression/>
- UCLA: Statistical Consulting Group (n.d.b), 'Ordered Logistic Regression | Stata Annotated Output'. lokaliseret d. 28. april 2022.
URL: www.stats.idre.ucla.edu/stata/output/ordered-logistic-regression/
- Uddannelses- og Forskningsministeriet (2013), 'Studerendes alder som færdige kandidater'. lokaliseret d. 25.04.2022.
URL: <https://ufm.dk/aktuelt/pressemeddelelser/arkiv/2009/ugens-tal-dtu-har-landets-yngeste-kandidater/studerendes-alder-som-faerdige-kandidater>
- Undervisningsudvalget 2016-2017 (2017), 'Bedre veje til en ungdomsuddannelse - anbefalinger til regeringen'.
- University of St Andrews (n.d.), 'Ordinal regression'. lokaliseret d. 11.05.2022.
- Wooldridge, J. M. (2010), *Econometric analysis of cross section and panel data*, MIT press.
- Wooldridge, J. M. (2016), *Introductory Econometrics*, 6 edn, Cengage Learning, Boston, MA, USA.
- Østergaard-Tygesen, S. F. (2018), *Hvad Driver Unges Uddannelsesvalg?*, Tænk tanken DEA.

Dataoversigt A

Til udarbejdelsen af den kvantitative analyse, er der anvendt data fra Danmarks Statistiks Forskerservice, hvor udgangspunktet har været individerne, der færdiggjorde grundskolen i 2002. Informationer fra en række forskellige datasæt fra Danmarks Statistik er samlet til ét datasæt. Bilaget er opbygget slavisk i forhold til sammensætningen af dataet. Løbende bliver der redegjort for de observationer der fjernes eller mistes, herudover reflekteres det over, hvad der kan ligge til grund for de mistede observationer. Til sidst i bilaget præsenteres den endelige sammensætning af datasættet og hvordan variablene heri er fordelt. En oversigt over de anvendte datasæt fra Danmarks Statistiks Forskerservice fremgår af tabel A.1.

Tabel A.1: Oversigt over Anvendte Datasæt fra Danmarks Statistiks Forskerservice

	Observationer	Variable
udda2001	5.330.607	14
ind2001	5.415.908	12
bef2001	5.326.164	15
bef2002	5.280.854	15
udda2002	5.285.753	14
ind2002	5.212.035	12
akm2002	5.212.038	8
fam2002	2.712.809	26
udfk2018	34.581.001	24
udda2019	4.889.175	15
ind2019	5.113.515	11
n_audd2015t_l1l5	4131	2

Kilde: Danmarks Statistiks Forskerservice

Formålet med dataarbejdet er at udarbejde et datafundament, der kan anvendes i undersøgelsen af de opstillede hypoteser og besvarelsen af problemformuleringen. For at gøre dette bedst muligt, inkluderes højeste fuldførte uddannelsesniveau i 2019 for de individer, der færdiggjorde grundskolen i 2002. Specifikt de individer, der færdiggjorde 9. klasse. Foruden dette inkluderes en række variable, der, ud fra de teoretiske og empiriske fund, forventes at bidrage til forklaringen af individers højeste fuldførte uddannelsesniveau samt andre variable, der har til formål at kontrollere for forskellige faktorer.

A.1 Oversigt

Der tages udgangspunkt i datasættet *bef2002*, der indeholder information om alle registrerede herboende danskere. I datasættet findes der variable som *pnr*, *MOR_ID*, *FAR_ID*, *ALDER*, *FAMILIE_ID*, *IE_TYPE*, *KOEN* og *KOM*, der alle findes relevante for empirisk at kunne svare på den valgte problemstilling. *pnr* er en unik identifikation af det enkelte individ som optræder i alle datasættene. *pnr* gør det dermed muligt, at kombinere datasæt og opnå yderligere information omkring individerne (Danmarks Statistik, n.d.r). *MOR_ID* samt *FAR_ID* angiver p-nummeret på de, der er registreret som juridisk "mor" og "far" for hvert enkelt individ. (Danmarks Statistik, n.d.g; Danmarks Statistik, n.d.f) *FAMILIE_ID* indeholder specifikke ID for familier, hvor en familie kan bestå af en enlig eller et samlevende par og eventuelle hjemmboende børn (Danmarks Statistik, n.d.d). ID'et anvendes senere hen til at koble antallet af søskende på datasættet. Variablen *ALDER* angiver individernes alder i

udgangen af året før. Udover at bruge variabelen til at afgrænse aldersgrupper, anvendes den også som en kontrolvariabel i den kvantitative analyse. Variabelen *IE_TYPE* angiver individernes etnicitetstype. Der differentieres mellem indvandrere, efterkommere og personer med dansk oprindelse (Danmarks Statistik, n.d.l). *KOEN* variabelen anvendes i analysen til at kontrollere for, hvilket køn individet er registreret som (Danmarks Statistik, n.d.m). Til sidst angiver variabelen *KOM* i hvilken kommunkode individet har folkeregisteradresse (Danmarks Statistik, n.d.n).

A.2 Grundskolekarakter

De udvalgte variable fra *bef2002* er sammensat med data fra *udfk2018*, der indeholder information omkring individer i grundskolen for årgangene 2001/2002-2017/2018. Det grundlæggende datasæt er altså *udfk2018*-datasættet. I datasættet omkring grundskolen, er der taget udgangspunkt i variablene *PNR*, *GRUNDSKOLEKARAKTER*, *BEDOEMMELSESFORM*, *SKALA*, *KLASSETRIN* samt *SKOLEAAR*. Sidst nævnte angiver, hvilken skoleårgang eleverne tilhører (Danmarks Statistik, n.d.u). Som redegjort for i kapitel 2.2, er udgangspunktet den kohorte, der færdiggjorde 9. klasse i 2002. Ved at isolere disse observationer, reduceres datasættet til 1.032.203 observationer, for 53.321 unikke individer, der i gennemsnit har 19 individuelle karakterer. Ved nærmere undersøgelse med dataet, har det vist sig, at der er nogle personer, der er opgjort fejlagtigt. Dette kommer til udtryk ved, at der er 1.310 personer, der er registreret til at være mellem 6-11 år gamle, når de færdiggør grundskolen. Fælles for disse individer er, at de kun optræder én gang i *udfk2018* datasættet, og altså kun har én karakter angivet. Ved at fjerne disse, er der 1.030.893 observationer og 52.011 unikke individer i datasættet. Til sammenligning var der ifølge Danmarks Statistik 55.595 personer, der færdiggjorde 9. klasse i 2002 (Danmarks Statistik, n.d.h). Folkeskolens afgangsprøver var i 2002 ikke obligatoriske, hvilket kan være en af årsagerne til, at tallet afviger (Møller, 2019). Variabelen *GRUNDSKOLEKARAKTER* angiver, hvilken karakter eleven har fået i en given bedømmelsessituation. I nærværende tilfælde er karaktererne givet på den daværende anvendte 13-skala. Danmarks Statistik har omregnet disse til den aktuelle 7-trins skala (Danmarks Statistik, n.d.j; Danmarks Statistik, n.d.t; Danmarks Statistik, n.d.c). For at kunne anvende karaktererne som en proxy for de iboende evner i projektet, udarbejdes et gennemsnit for hver af individerne, med udgangspunkt i karaktererne givet både ved standpunkt og afgangsprøve. Der er udarbejdet et samlet karaktergennemsnit, der inkluderer begge bedømmelsesformer. I processen er der både udarbejdet et karaktergennemsnit for standpunkt og afgangsprøve individuelt. Som nævnt i kapitel 7, var det i 2002 ikke obligatorisk at gå til afgangsprøver, hvorfor nogle af personerne ikke har karakterer for dette. Det endelige valg er taget med et ønske om at bibeholde så stort et datasæt som muligt, men samtidig også for at tage højde for, at individerne kan præstere forskelligt, i forskellige situationer, jf. kapitel 4. Overordnet set er karaktererne inkluderet som en indikation på individernes evner i den kvantitative analyse, på baggrund af fundene i litteraturstudiet og teorien, der påpeger, at der er en sammenhæng mellem evner og uddannelseniveau (Nielsen og Vissing-Jorgensen, 2006; Produktivitetskommissionen, 2014; Borjas, 2016).

A.3 Alder & Køn

Når de to datasæt *udfk2018* samt *bef2002* kobles sammen, er der 51.957 observationer. Der er nogle observationer, der tabes i sammenlægningsprocessen, da personer ikke optræder begge steder. Dette er en nødvendighed for at have fuld information om individerne. Som nævnt i kapitel 2.2, inddrages kun de individer, der er i alderen 14-18 år. Ved at foretage denne afgrænsning, reduceres datasættet med 104 observationer, og indeholder dermed 51.853 observationer. Individernes køn er defineret i 2002 ud fra den førnævnte *KOEN* variabel, der angiver individernes juridiske køn (Danmarks Statistik, n.d.m). Det antages altså i datasættet, at det køn, individerne har i 2002 er det samme som de har i 2019.

A.4 Etnicitet

Individernes etnicitet er opgjort i *bef2002* datasættet, hvor individer opdeles i enten at være af dansk oprindelse, indvandrer eller efterkommer (Danmarks Statistik, n.d.l). Definitionen af individers etnicitet følger Danmarks Statistik definition heraf (Danmarks Statistik, n.d.l). Indvandrere er individer der er

født i udlandet og hvor ingen af individets forældre er både danske statsborgere og født i Danmark. Efterkommere er personer der er født i Danmark, men ingen af forældrene er både danske statsborgere og født i Danmark. Hvis et individ, uanset fødested, har mindst én forælder der er både dansk statsborger og født i Danmark, er individet af dansk oprindelse. En mere detaljeret forklaring fremgår af Danmarks Statistiks hjemmeside (Danmarks Statistik, n.d.l).

A.5 Kommunetype

KOM variabelen fra *bef2002* datasættet er inkluderet, da den indeholder en kommunekode, der repræsenterer, i hvilken kommune individets folkeregisteradresse er registreret (Danmarks Statistik, n.d.n). Variablen benyttes i nærværende tilfælde til at finde individers folkeregisteradresse i 2002, hvor de færdiggjorde grundskolen. Litteraturstudiets fund tyder på, at der kan være en sammenhæng mellem individets uddannelsesbeslutninger samt vedkommendes geografiske afstand til uddannelsesinstitutioner (Undervisningsudvalget 2016-2017, 2017). Danmarks Statistik har udarbejdet en inddeling af de danske kommuner (Danmarks Statistik, n.d.o). De er inddelt i 5 grupper, der er baseret på indbyggertal samt tilgængelighed til arbejdsplader. Jobtilgængelighed er ikke alene de direkte jobmuligheder i kommunen, men reflekterer også indirekte økonomisk aktivitet forbundet med jobskabelsen. De 5 kommunegrupper er: Landkommune, Oplandskommune, Provinsbykommune, Storbykommune samt Hovedstadskommune. Grupperingerne muliggøre en nem sammenligning af kommunerne (Danmarks Statistik, 2018a).

Kommunekoderne i *bef2002* datasættet tager udgangspunkt i de 271 kommuner i 2002, hvorimod Danmarks Statistiks kommunegrupper tager udgangspunkt i de 98 kommuner, der er efter kommunesammenlægningen i 2007 (Danmarks Statistik, n.d.o; Indenrigs- og boligministeriet, n.d.a). Kommunesammenlægningen betød, udover at der blev færre kommuner, at nogle kommunegrænser blev rykket. Det er i indeværende tilfælde ikke muligt at identificere individernes adresse mere konkret end den pågældende kommune pga. begrænset data på området. Det er i projektet derfor ikke muligt at tage højde for kommunegrænser, der blev flyttet. Sammenlægningen er dermed udelukkende lavet på baggrund af de overordnede kommuner før og efter (Indenrigs- og boligministeriet, n.d.b).

A.6 Søskende

Datasættet *fam2002* indeholder 2.717.809 observationer, der hver repræsenterer en familie, jf. tabel A.1. En familie kan som nævnt bestå af en enlig eller et samlevende par og eventuelle hjemmeboende børn. Hver familie er identificeret ved et unikt ID, *FAMILIE_ID*. ID'et gør det muligt at koble information fra datasættet sammen med individerne i den valgte kohorte. *fam2002* indeholder i alt 26 variable, hvoraf en er det unikke *FAMILIE_ID* og de resterende 25 variable angiver, hvor mange børn i alderen 0-24, der er i familien (Danmarks Statistik, n.d.b).

$$S\ddot{o}sken\ddot{e} = \left(\sum_{i=0}^{24} x_i \right) - 1 \quad (\text{A.1})$$

Ligning A.1 angiver udregningen af søskende, hvor i er alder og x er antallet. Der kan være søskende, der ikke kommer med i opgørelsen, da de på ikke bor på samme folkeregisteradresse. Ved at koble *søskende* variabelen på datasættet til den kvantitative analyse, er der 50.873 observationer. De 980 observationer, der tabes i denne proces kan være individer, der er flyttet hjemmefra eller på anden vis har skabt sit eget familie ID. Det vurderes, at disse individer kan skabe en skævvridning af estimer, hvorfor tabet af observationer accepteres. Hvis det er tilfældet, at de unge er flyttet hjemmefra i en tidlig alder, kan det skyldes ustabilitet i hjemmet. At inddrage disse personer i en kvantitativ analyse, der tager udgangspunkt i, at de socioøkonomiske faktorer omkring den unge har en sammenhæng med de unges uddannelsesbeslutning, kan føre til upræcise estimer.

A.7 Forældre

For at inddrage information om socioøkonomiske faktorer, jf. kapitel 4.2.1, inkluderes variable om individernes forældre. For at gøre dette, tages der på ny udgangspunkt i *bef2002* datasættet. Denne

gang fokuseres der på *MOR_ID* samt *FAR_ID* variablene. "mor" og "far" er to forskellige datasæt, men det forestående arbejde er identisk, uanset om det er for "mor- eller "far"-dataet. Efter at have fjernet duplikater, er der 1.610.662 observationer for "mor" og 1.563.745 observationer for "far". Duplikater kan opstå, hvis "mor" eller "far" har mere end det ene barn, der blev færdig med grundskolen i 2002. Datasættet kobles sammen med *udda2002* for at finde forældrenes højeste fuldførte uddannelse, *HFAUDD* (Danmarks Statistik, n.d.k). Registreringen af højeste fuldførte uddannelse sker efter rangeringen i kapitel 2.4.2 ved hjælp af *n_audd2015t_l1l5*-formatet fra Danmarks Statistik. Variablen for den højeste fuldførte uddannelse, *HFAUDD* fra *udda2002* datasættet, indeholder et ID, der er unikt for den pågældende uddannelse. Formatet fra Danmarks Statistik indeholder 4.131 observationer, der hver repræsenterer en uddannelse i Danmark. Hver af de 4.131 observationer kobles på en uddannelsesstype, som er følgende: *95 Afbrudt, 10 Grundskole, 20 Almengymnasiale uddannelser, 25 Erhvervsgymnasiale uddannelser, 35 Erhvervsfaglige praktik- og hovedforløb, 40 Korte videregående uddannelser, 50 Mellemlange videregående uddannelser, 60 Bachelor, 65 Lange videregående uddannelser* samt *70 Forskeruddannelser*. Formatet gør det altså muligt at kategorisere de 4.131 forskellige uddannelser i nogle overordnede grupper, der så kan bruges til rangeringen fra kapitel 2.4.2 Herefter tages der udgangspunkt i *ind2002*-datasættet for at koble forældrenes indkomst, *QSPLINDK*, på. Indkomsten er registreret som den skattepligtige indkomst, der er angivet på selvangivelsen. Der er her taget højde for almindelig beskatning og diverse fradrag. Der ses bort fra aktieindkomster i den skattepligtige indkomst (Danmarks Statistik, n.d.s). Der er inkluderet en variabel for socialklasse fra *akm2002* datasættet, som indeholder en socioøkonomisk klassifikation, der er lavet på baggrund af forældrenes indkomstkilde. Denne opgørelse er baseret på Danmarks Statistik (n.d.v). Opgørelsen består af 22 forskellige statusser. For ikke at øge kompleksiteten, sammenlægges nogle af statusserne. Disse er sammenlagt til seks statusser: *selvstændig, lønmodtager, pension, overførselsindkomst, børn under 15 år eller andet*. De 6 statusser indeholder som følger *Selvstændig* inkluderer statusserne *selvstændig, selvstændig 10 eller flere ansatte, selvstændig 5-9 ansatte, selvstændig 1-4 ansatte* og *selvstændig ingen ansatte*. Statussen *lønmodtager* inkluderer *medarbejdende ægtefælle, lønmodtager med ledelsesarbejde, lønmodtager i arbejde der forudsætter færdigheder på højeste niveau, lønmodtager i arbejde der forudsætter færdigheder på mellem niveau, lønmodtager i arbejde der forudsætter færdigheder på grundniveau, andre lønmodtagere* og *lønmodtager; stillingsangivelse ikke oplyst*. *Pension* udgøres af grupperne *folkepensionister* og *efterlønsmodtager, mv*. Statussen *overførselsindkomst* udgøres af grupperne *arbejdsløs mindst halvdelen af året, modtager af sygedagpenge; uddannelsesgodtgørelse; orlovsydelser mm., under uddannelse; inkl. skoleelever min. 15 år, førtidspensionister* og *kontanthjælpsmodtagere*. Andre udgøres af gruppen *andre*. Personer der indgår i gruppen *andre* er karakteriseret ved, at de ingen eller lille tilknytning til arbejdsmarkedet (Danmarks Statistik, n.d.v). Den sidste gruppe er *børn under 15 år ultimo året*, der er dog ikke nogen i den kategori i populationen, hvorfor den ikke inddrages. Sammenlægningen betyder, at der eksempelvis ikke differentieres mellem, hvor mange ansatte en selvstændig måtte have eller hvorvidt lønmodtageren har ledelsesansvar eller færdigheder i højere eller lavere grad. *ALDER*, der angiver forældrenes alder ultimo året før inkluderes (Danmarks Statistik, n.d.a). Til sidst inkluderes den enkelte forældres *FAMILIE_TYPE*, hvor der skelnes mellem, hvorvidt forældren bor som enlig eller som i et par (Danmarks Statistik, n.d.e).

Ved at koble datasættet for "mor" til *MOR_ID* i datasættet i den kvantitative analyse, reduceres antallet af observationer til 50.790 og yderligere til 49.822, når "far" datasættet kobles på *FAR_ID*. De observationer, der forsvinder, er hos individer, hvor der ikke er information om deres "mor" og/eller "far". Det er nødvendigt at hvert individ i den kvantitative analyse har registreret både en juridisk mor og far i *bef2002* datasættet. Det vurderes at være et acceptabelt tab af observationer, i nærværende tilfælde, hvor information om forældrene i litteraturen ofte vurderes at spille en stor rolle i individernes uddannelsesvalg.

A.8 Indkomst

Fra *ind2019* inkluderes individernes skattepligtige indkomst i 2019. Ved at koble 2019 indkomsten på dataet, er der 48.565 observationer i datasættet. Det er altså 1.257 observationer, der mistes på den 17 årige periode. Tabet af observationer kan repræsentere individer, der ikke længere bor i Danmark eller i

mellemtiden er afgået ved døden.

A.9 Uddannelsesniveau

Det højeste gennemførte uddannelsesniveau er inkluderet ved hjælp af *udda2019* datasættet fra Danmarks Statistik, hvor *HFAUDD* variabelen er inddelt i uddannelsesniveauer efter formatet *n_audd2015t_l1l5* med samme fremgangsmåde som hos forældrene. Derefter er de blevet klassificeret efter rangeringen i kapitel 2.4.2. Højeste uddannelsesniveau kobles på det samlede datasæt, der anvendes i den kvantitative analyse, hvilket er årsagen til at der er 47.744 observationer tilbage.

For at kunne bearbejde datasættet i den kvantitative analyse, er observationer, hvor der er manglende værdier, blevet fjernet. Dette reducerer datasættet til 41.926 observationer. De observationer, hvor de manglende værdier gør sig gældende, er hovedsageligt hos forældre-variablene. Det vurderes dog ud fra litteraturstudiets fund herom, at forældre-variablene er relevante forklarende faktorer for valget af uddannelsesniveau, hvorfor tabet accepteres.

Til sidst foretages der et valg om at frasortere de individer, der i 2019 er indskrevet på en uddannelse. Dette gøres for ikke at skævvride dataet. Derudover sikrer det også, at indkomstvariablen i 2019 ikke påvirkes negativt af individer, der som resultat af igangværende uddannelse modtager overførselsindkomst. 7.3% af individerne er i 2019 indskrevet på en uddannelse, hvorfor det endelige datasæt er på 39.090 observationer. Hvordan fordelingen af variablene er for de 39.090 observationer, præsenteres nedenfor.

A.10 Endeligt Datasæt

Følgende sektion giver et overblik over fordelingen af observationer i det endelige datasæt, der anvendes i den kvantitative analyse.

Uddannelsesniveau

Tabel A.2: Antal Observationer Opdelt efter Uddannelsesniveau

Udd.	Grundskole	Gymnasial	Erhvervs- uddannelse	Kort videre- gående	Akademisk bachelor	Professions- bachelor	Lang videre- gående
Obs.	3.918	1.864	12.320	2.331	943	8.758	8.956

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

Alder

Tabel A.3: Antal Observationer Opdelt efter Alder

Alder	14	15	16	17	18
Obs.	1.331	33.349	4.005	351	54

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

Køn

Tabel A.4: Antal Observationer Opdelt efter Køn

Køn	Pige	Dreng
Obs.	19.405	19.685

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

Etnicitet

Tabel A.5: Antal Observationer Opdelt efter Etnicitet

Etnicitet	Dansk oprindelse	Indvandrere	Efterkommer
Obs.	37.365	1.119	606

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

Kommunetype

Tabel A.6: Antal Observationer Opdelt efter Kommunetype

Kommunetype	Land-kommune	Oplands-kommune	Provins-kommune	Storby-kommune	Hovedstads-kommune
Obs.	10.522	7.164	9.495	4.322	7.587

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

Søskende

Tabel A.7: Antal Observationer Opdelt efter Antal Søskende

Søskende	0	1	2	3	4	5	6	7	8	<8 ¹
Obs.	9.065	19.154	8.182	2.074	417	110	52	23	5	8

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

Mor Alder

Tabel A.8: Antal Observationer Opdelt efter Mor Alder

Mor alder	>32 ²	32	33	34	35	36	37	38	39	40	41	42
Obs.	27	66	175	424	783	1.266	1.812	2.362	2.607	3.138	3.372	3.288

Mor alder	43	44	45	46	47	48	49	50	51	52	53	54
Obs.	3.229	3.021	2.723	2.410	1.990	1.603	1.267	981	759	551	410	303

Mor alder	55	56	57	58	59	60	<60 ³
Obs.	219	130	89	45	22	13	5

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

¹Max. værdier fjernet pga. risiko for identificering

²Min. værdier fjernet pga. risiko for identificering

³Max. værdier fjernet pga. risiko for identificering

Mor Socialklasse

Tabel A.9: Antal Observationer Opdelt efter Mor Socialklasse

Mor socialklasse	Lønmodtager	Selvstændig	Pension	Overførselsindkomst	Andre
Obs.	31.917	1.439	105	4.978	651

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

Mor Uddannelsesniveau

Tabel A.10: Antal Observationer Opdelt efter Mor Uddannelsesniveau

Mor udd.	Grundskole	Gymnasial	Erhvervs-uddannelse	Kort videre-gående	Akademisk bachelor	Professions-bachelor	Lang videre-gående
Obs.	10.853	1.562	14.279	1.477	152	9.099	1.668

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

Mor Familietype

Tabel A.11: Antal Observationer Opdelt efter Mor Familietype

Mor familietype	Sammen	Enlig
Obs.	33.456	5.634

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

Far Alder

Tabel A.12: Antal Observationer Opdelt efter Far Alder

Far alder	>32 ⁴	32	33	34	35	36	37	38	39	40	41	42	43
Obs.	7	9	32	108	225	471	746	1.152	1.444	1.952	2.437	2.686	2.900

Far alder	44	45	46	47	48	49	50	51	52	53	54	55
Obs.	3.004	2.564	2.460	2.087	1.717	1.464	1.170	2.968	3.089	978	839	706

Far alder	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71
Obs.	474	398	258	217	130	101	82	59	38	19	20	23	13	13	10	5

Far alder	72	73	<73 ⁵
Obs.	5	5	5

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

Far Socialklasse

⁴Min. værdier fjernet pga. risiko for identificering

⁵Max. værdier fjernet pga. risiko for identificering

Tabel A.13: Antal Observationer Opdelt efter Far Socialklasse

Far socialklasse	Lønmodtager	Selvstændig	Pension	Overførselsindkomst	Andre
Obs.	30.008	4.195	669	3.624	594

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

Far Uddannelsesniveau

Tabel A.14: Antal Observationer Opdelt efter Far Uddannelsesniveau

Far udd.	Grundskole	Gymnasial	Erhvervs- uddannelse	Kort videre- gående	Akademisk bachelor	Professions- bachelor	Lang videre- gående
Obs.	9.425	1.473	18.009	1.590	233	4.950	3.410

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

Far Familietype

Tabel A.15: Antal Observationer Opdelt efter Far Familietype

Far familietype	Sammen	Enlig
Obs.	33.571	5.519

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

Tabeller Deskriptiv Statistik

B

Uddannelsesniveau

Tabel B.1: Højeste fuldførte Uddannelsesniveau

Uddannelsesniveau	Fordeling
Grundskole	10,02 %
Gymnasial uddannelse	4,77 %
Erhvervsuddannelse	31,52 %
Kort videregående uddannelse	5,96 %
Akademisk bachelor	2,41 %
Professionsbachelor	22,40 %
Lang videregående uddannelse	22,91 %

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

Køn

Tabel B.2: Højeste Fuldførte Uddannelsesniveau Opdelt efter Køn

Uddannelsesniveau	Kvinder	Mænd
Grundskole	7,39 %	12,62 %
Gymnasial uddannelse	3,97 %	5,55 %
Erhvervsuddannelse	26,20 %	36,75 %
Kort videregående uddannelse	5,69 %	6,23 %
Akademisk bachelor	2,40 %	2,42 %
Professionsbachelor	28,81 %	16,09 %
Lang videregående uddannelse	25,53 %	20,33 %

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

Kommunetype

Tabel B.3: Højeste Fuldførte Uddannelsesniveau Opdelt efter Kommunetype

Uddannelsesniveau	Land-kommune	Oplands-kommune	Provins-kommune	Storby-kommune	Hovedstads-kommune
Grundskole	9,45 %	9,46 %	10,28 %	10,18 %	10,94 %
Gymnasial uddannelse	4,30 %	4,02 %	4,66 %	4,40 %	6,48 %
Erhvervsuddannelse	34,95 %	34,81 %	31,95 %	26,72 %	25,83 %
Kort videregående uddannelse	6,17 %	5,71 %	6,33 %	5,81 %	5,55 %
Akademisk bachelor	2,19 %	2,40 %	2,02 %	2,43 %	3,22 %
Professionsbachelor	24,27 %	22,39 %	23,28 %	21,24 %	19,40 %
Lang videregående uddannelse	18,68 %	21,20 %	21,48 %	29,22 %	28,58 %

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

Etnicitet

Tabel B.4: Højeste Fuldførte Uddannelsesniveau Opdelt efter Etnicitet

Uddannelsesniveau	Dansk	Indvandrere	Efterkommer
Grundskole	9,60 %	20,55 %	16,83 %
Gymnasial uddannelse	4,64 %	7,42 %	7,59 %
Erhvervsuddannelse	31,75 %	27,97 %	23,43 %
Kort videregående uddannelse	5,99 %	5,27 %	5,61 %
Akademisk bachelor	2,40 %	2,59 %	2,97 %
Professionsbachelor	22,56 %	18,23 %	20,46 %
Lang videregående uddannelse	23,06 %	17,96 %	23,10 %

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

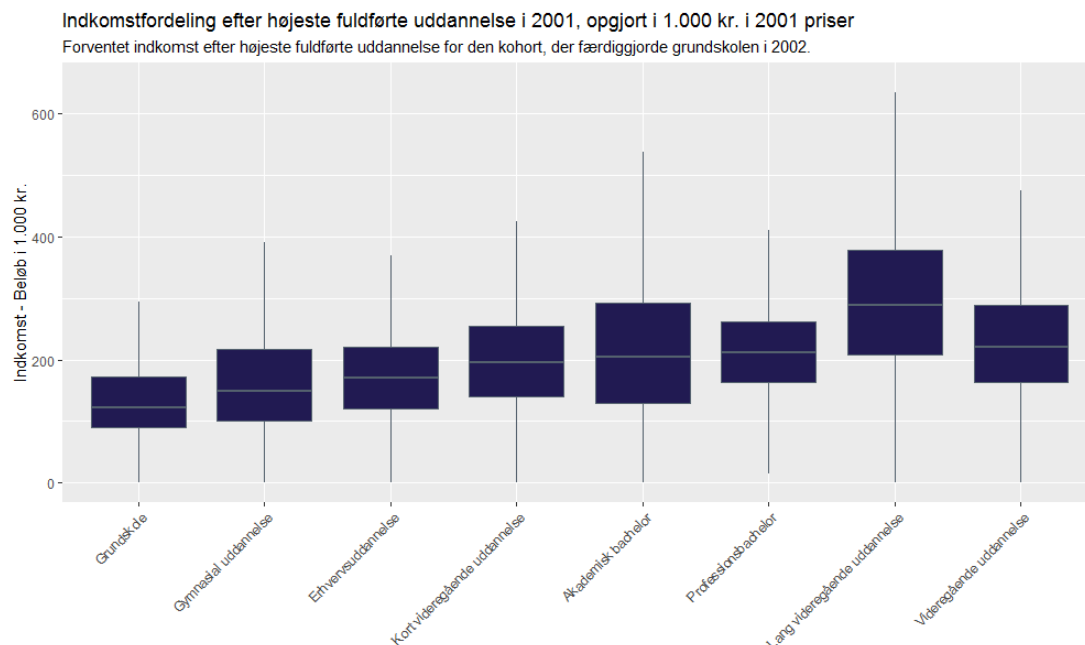
Karaktergennemsnit

Tabel B.5: Karakteristika for Grundskolekarakterer, Standpunkt, Afgangsprøve og Gennemsnit

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Standpunkt	-3,00	4,82	6,46	6,18	7,55	12,00
Afgangsprøve	2,00	4,60	6,11	6,05	7,50	12,00
Gennemsnit*	-3,00	4,85	6,38	6,17	7,57	12,00

Kilde: Egne beregninger på data fra Danmarks Statistiks Forskerservice. *Gennemsnit er udregnet ud fra karaktererne fra standpunkt og afgangsprøver

Figur B.1: Boxplot over Indkomst



Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

Indkomsten er fjernet for alle personer, der tjener under/over 1,5 gange af interkvartil, pga. risiko for identificering

Af dette bilag fremgår en præsentation af kontrolvariablene der inddrages i regressionsmodellerne. Herunder redegøres der for definitionen og operationaliseringen af kontrolvariablene, samt hvilke forventninger der er til dem. Foruden dette præsenteres de variable hvor der er indraget resultater fra andre studier i kapitel 4, men som ikke inddrages i den kvantitative analyse. Der redegøres for, hvorfor det kunne have været relevant at inddrage disse variable, samt hvorfor de er fravalgt.

C.1 Kontrolvariable

I det følgende redegøres der for kontrolvariablene, der inddrages i regressionsmodellerne. Det fremgår af kapitel 5, at det ikke, på baggrund af kapitel 3 og kapitel 4, har været muligt at opstille klare forventninger til kontrolvariablene. Dette kan være fordi, at der empirisk ikke er fundet en klar sammenhæng mellem faktoren og uddannelsesniveau eller fordi at operationalisering af faktoren ikke gør det muligt at sammenligne direkte med resultaterne fra kapitel 3 og kapitel 4. Variable som alder, køn samt etnicitet er inkluderet i regressionsmodellerne, så det er muligt at kontrollere for disse. Kontrolvariablene er ikke af primær interesse, da det ikke er muligt at sandsynliggøre, hvorvidt der eksisterer en kausal sammenhæng mellem dem og unges valg af uddannelsesniveau. I det følgende opstilles der forventninger til variablene på baggrund af den deskriptive statistik, jf. kapitel 5.2, samt kapitel 3 og kapitel 4.

Der er foretaget t-tests for hver af kontrolvariablene i både model 1 og model 2, resultaterne heraf fremgår af bilag E. I tilfælde af, at t-testen for en uafhængig variabel viser, at der er en signifikant sammenhæng med den afhængige variabel, fremgår der ikke yderligere herom. Hvis t-testen derimod viser, at der ikke eksisterer en signifikant sammenhæng, fremgår det under den pågældende variabel. For yderligere at sikre robustheden af kontrolvariablene, er en række andre operationaliseringer undersøgt. Information om disse robusthedundersøgelser fremgår af bilag F.

Alder, A_{it_0}

Variablen alder, A_{it_0} , defineres som individernes alder i begyndelsen af det år, de færdiggjorde grundskolen, jf. bilag A. Alder er defineret i hele år. Som beskrevet i kapitel 2, inkluderes der kun individer mellem 14-18 år, hvorfor variablen kan tage værdierne $A_{it_0} = 14, \dots, 18$. Variablen er en diskret variabel med et endeligt antal værdier og benyttes som kontrolvariabel. Danmarks Statistiks opgørelse af alder betyder, at variablen A_{it_0} ikke nødvendigvis afspejler alderen på det tidspunkt et individ færdiggjorde grundskolen, da grundskolen færdiggøres i juni måned. Dermed kan et individ, hvor $A_{it_0} = 14$, potentielt være 15 år ved udgangen af grundskolen. Alderen afspejler dermed i højere grad, hvilket år individet er født.

Af nedenstående tabel C.1 fremgår en oversigt over den procentvise fordeling af alder for henholdsvis model 1 og model 2.

Som det fremgår af tabel C.1, er fordelingen af alder rimelig ens for de to modeller. Der er en lidt højere andel af personer på 14 år i model 2 og en lidt mindre andel over 15 år.

Ud fra sammenhængene fundet i det empiriske litteraturstudie forventes det, at der er en positiv sammenhæng mellem alder og uddannelsespræstation og derved højeste fuldførte uddannelsesniveau (Jensen og Nielsen, 2010). Som det kan ses af tabel C.1 er der en mindre andel af personer over

Tabel C.1: Oversigt over Procentvis Fordeling af Alder - Model 1 og Model 2

	14 år	15 år	16 år	17 år	18 år
Model 1	3,40 %	85,31 %	10,25 %	0,90 %	0,14 %
Model 2	4,35 %	88,23 %	6,74 %	0,59 %	0,09 %

Kilde: Egne beregninger på data fra Danmarks Statistiks Forskerservice

medianalderen, i model 2, hvilket taler for at der er en negativ sammenhæng mellem alder og højeste fuldførte uddannelsesniveau.

Køn, K_{it_0}

Variablen køn, K_{it_0} , benyttes som en kontrolvariabel og defineres som en dummy variabel, hvor $K_{it_0} = 0$ hvis individet er en kvinde/pige og $K_{it_0} = 1$ hvis individet er en mand/dreng. Definition af køn som en binær variabel sker på baggrund af, at der i Danmark kun er to juridiske køn. Dette betyder dermed ikke, at forfatterne ikke anerkender eksistensen af andre køn. Det forventes på baggrund af den deskriptive statistik i kapitel 5.2, at der er en negativ sammenhæng mellem køn og højeste fuldførte uddannelsesniveau. Der kan også ses i tabel C.2, hvor der er en større andel af kvinder i model 2 end model 1.

Tabel C.2: Oversigt over Procentvis Fordeling af Køn - Model 1 og Model 2

	Kvinde	Mand
Model 1	49,64 %	50,36 %
Model 2	57,72 %	42,28 %

Kilde: Egne beregninger på data fra Danmarks Statistik Forskerservice

Etnicitet, E_{it_0}

Etnicitet er i regressionsmodellerne inddraget som en kontrolvariabel. I regressionsmodellerne er den operationaliseret, som en dummy variabel, der angiver hvorvidt individerne er af dansk oprindelse eller af ikke dansk oprindelse. Definitionen af individers oprindelse følger definitionen fra Danmarks Statistik (Danmarks Statistik, n.d.). $E_{it_0} = 1$ hvis en person er af ikke dansk oprindelse. Personer med dansk oprindelse er angivet som benchmark, da det er den største gruppe i populationen. Dette fremgår af nedenstående tabel C.3, hvor over 95% af personerne i modellerne er af dansk oprindelse.

Tabel C.3: Oversigt over Procentvis Fordeling af Etnicitet - Model 1 og Model 2

	Dansk oprindelse	Ikke dansk oprindelse
Model 1	95,59 %	4,41 %
Model 2	96,15 %	3,85 %

Kilde: Egne beregninger på data fra Danmarks Statistiks Forskerservice

Det er ud fra operationaliseringen ikke muligt at differentiere mellem indvandrere og efterkommere. Som det fremgår af figur 5.4 i kapitel 5.2, er der forskel i andelen med de forskellige uddannelsesniveauer for efterkommere og indvandrere. Det er dog kun ca. 1,5% af observationerne i populationen, der er efterkommere, jf. tabel A.5 i bilag A. For at sikre tilstrækkeligt store grupper sammenlægges indvandrere og efterkommere. Det er desuden ikke muligt at differentiere mellem, hvilket land personer med anden oprindelse end dansk kommer fra. Der er stor forskel i, hvad fordelingen af højeste fuldførte uddannelsesniveau er, alt efter oprindelsesland (Danmarks Statistik, 2019a). Det er dog under 5% af populationen, der er af anden oprindelse end dansk. At differentiere efter deres oprindelsesland vil medføre flere dummy variable og færre frihedsgrader i modellerne. Derudover vil opdelingen resultere

i små grupper, hvorfor der er mindre variation indenfor etniciteterne. Dette medfører, at det er mere vanskeligt at finde signifikante sammenhænge. Endvidere er det i praksis heller ikke muligt at differentiere efter oprindelsesland i det tilgængelige data.

Ud fra resultaterne fra studierne gennemgået i kapitel 4 er det tvetydigt, hvordan sammenhængen mellem etnicitet og højeste fuldførte uddannelsesniveau er alt efter hvilke socioøkonomiske faktorer, der er taget højde for i undersøgelsen (Jaeger, 2007; Jensen og Nielsen, 2010; DØRS, 2003). På baggrund af den deskriptive statistik i kapitel 5.2 forventes det, at der er en negativ sammenhæng mellem etnicitet og højeste fuldførte uddannelsesniveau, så personer med dansk oprindelse har større sandsynlighed for at gennemføre en højere rangeret uddannelse.

Kommunetype, $KO1_{it_0}$, $KO2_{it_0}$, $KO3_{it_0}$, $KO4_{it_0}$

I regressionsmodellerne inddrages dummy variable, der angiver, hvilken kommunetype individet havde adresse i, da vedkommende i 2002 færdiggjorde grundskolen. Disse inddrages som kontrolvariable. Kommunetypen defineres efter Danmarks Statistik, som inddeler kommuner i følgende fem kategorier; landkommuner, oplandskommuner, provinskommuner, storbykommuner og hovedstadskommuner. Denne inddeling er af Danmarks Statistik foretaget med udgangspunkt i tilgængeligheden af arbejdspladser og antallet af indbyggere i kommunens største by (Danmarks Statistik, n.d.o). En oplandskommune har flere indbyggere og arbejdspladser end en landkommune, en provinskommune har flere indbyggere og arbejdspladser end en oplandskommune, og så fremdeles. Variablen er en proxy for afstanden til uddannelsesinstitutioner. Af individerne i populationen havde den største gruppe i 2002 adresse i en landkommune, hvorfor denne er brugt som benchmark. Hvis $KO1_{it_0} = 1$ havde individet adresse i en oplandskommune, $KO2_{it_0} = 1$ i en provinskommune, $KO3_{it_0} = 1$ i en storbykommune og $KO4_{it_0} = 1$ i en hovedstadskommune. En detaljeret beskrivelse af dataarbejdet, herunder hvordan der er taget højde for kommunesammenlægningen i 2007, fremgår af bilag A.

Af nedenstående tabel C.4 fremgår den procentvise fordeling mellem kommunetyperne for henholdsvis model 1 og model 2. Af tabellen fremgår det, at fordelingen mellem kommunetyper er forholdsvis ens i modellerne. Andelen af individer fra grupperne storby- og hovedstadskommune er dog ca. 1 procentpoint højere i model 2 end i model 1. Af t-testene for model 1, fremgår det, at oplands-, provins- og hovedstadskommune ikke er signifikante. Dermed eksisterer der ikke et linenært forhold mellem disse dummy variable og den afhængige variabel, jf. tabel E.9. I model 1 er dummy variablen for storbykommune signifikant. I model 2 er alle af kommunetyperne signifikante.

Tabel C.4: Oversigt over Procentvis Fordeling mellem Kommunetyper - Model 1 og Model 2

	Land	Opland	Provins	Storby	Hovedstad
Model 1	26,92 %	18,33 %	24,29 %	11,06 %	19,41 %
Model 2	25,72 %	17,65 %	24,03 %	12,09 %	20,51 %

Kilde: Egne beregninger på data fra Danmarks Statistiks Forskerservice

Det forventes, at der er en positiv sammenhæng mellem kommunetype, for hovedstadskommuner og storbykommuner, og højeste fuldførte uddannelsesniveau. Det forventes, at der er en negativ sammenhæng for de restende kommunetyper. Disse forventninger foretages på baggrund af resultaterne fra de empiriske studier i kapitel 4 samt på baggrund af der, i kapitel 5.2 er en lavere andel af individer, som var bosat i land-, opland- og provinskommuner, da de færdiggjorde grundskolen, og som har gennemført en lang videregående uddannelse (Hallmann og Carstensen, 2020; Nielsen og Vissing-Jørgensen, 2006; Jensen og Nielsen, 2010).

Kontrolvariablen kommunetype fungerer som en proxy for afstand til uddannelsesinstitutioner. Et bedre mål for afstanden vil være den præcise afstand fra individets folkeregisteradresse, da personen færdiggjorde grundskolen, og til uddannelsesinstitutionen hvor personen gennemførte sin højeste fuldførte uddannelse. Afstanden til uddannelsesinstitutioner afhænger af, hvilken type uddannelse individet vælger. Uddannelsesinstitutioner på de forskellige niveauer i en vis udstrækning placeret i hele landet, selvom

der er færre pladser udenfor storbyerne (Hallmann og Carstensen, 2020). Den præcise afstand ville være bedre, da kommunerne dækker et stort areal, så det er forskelligt, hvor langt personer, der bor i samme kommune, har til samme uddannelsesinstitution.

Variablen for kommunetype er medtaget, for at kontrollere for de geografiske afstande til de forskellige uddannelsesniveauer og ikke til de specifikke uddannelsesinstitutioner personer tager uddannelse på. I nærværende tilfælde vurderes det, at grupperingerne af kommuner foretaget af Danmarks Statistik til en hvis grad kan afspejle afstanden et individ har til uddannelsesinstitutioner og derved er en tilstrækkelig måleenhed.

Netop som følge af at operationaliseringen er vanskelig samt at resultaterne ikke er helt sammenlignelige med empiriske fund, inddrages variablen, som en kontrolvariabel.

Søskende, S_{it_0}

Variablen søskende, S_{it_0} , defineres som antallet af individ i 's søskende der boede på samme folkeregisteradresse, da individet færdiggjorde grundskolen. Mere information omkring operationaliseringen af variablen, S_{it_0} , fremgår af bilag A. Variablen er diskret og benyttes som en kontrolvariabel. Det forventes ud fra studierne i kapitel 4, at der eksisterer en negativ sammenhæng mellem antallet af søskende og højeste fuldførte uddannelsesniveauer.

Antallet af søskende har ifølge studier inddraget i det empiriske litteraturstudie en sammenhæng med uddannelsespræstation, da familier med flere børn har færre ressourcer (Benders, 2013). De socioøkonomiske faktorer beskriver mange af de samme tendenser, hvorfor de forventes at være korreleret. Herudfra kan det forventes, at der er en korrelation mellem antal søskende og forældres indkomst. Der eksisterer i dataet dog ikke korrelation mellem mors ellers fars indkomst og antallet af søskende, jf. bilag G.

Af nedenstående tabel C.5 fremgår det, at karakteristika for variablen søskende er næsten identiske i model 1 og model 2. Dette taler for, at der ikke er en sammenhæng mellem antallet af søskende og sandsynligheden for at gennemføre en videregående uddannelse. T-testen for model 2 viser, at der ikke eksisterer et lineært forhold mellem antallet af søskende og den afhængige variabel.

Tabel C.5: Karakteristika for Variablen Søskende - Model 1 og Model 2

	Min.	1st Qu.	Median	Mean	3rd Qu. ¹
Model 1	0,00	1,00	1,00	1,14	2,00
Model 2	0,00	1,00	1,00	1,13	2,00

Kilde: Egne beregninger på data fra Danmarks Statistiks Forskerservice

S_{it_0} operationaliseres ud fra data fra Danmarks Statistiks Forskerservice, hvor antallet af børn mellem 0-24 år registreres. I tilfælde af, at individ i har en hjemmeboende søskende på 30 år, inddrages denne dermed ikke i variablen. Endvidere registreres udeboende søskende heller ikke (Danmarks Statistik, n.d. d). Dette betyder, at antallet af søskende for et individ potentielt undervurderes. Det optimale ville være, at have en variabel, der tog højde for alle de søskende individet har og ikke bare dem, som bor på samme folkeregisteradresse og som er under 24 år, når individet er mellem 14 og 18 år. Operationaliseringen af variablen betyder, at det ikke er muligt at differentiere mellem søskendes alder. Det antages derfor, at det kun er antallet af søskende, og ikke deres alder, der påvirker individers valg af uddannelsesniveau. Denne antagelse findes rimelig, da det ifølge empiriske fund er forældres ressourcer, der bliver påvirket af antal af børn og ikke hvor gamle børnene er (Benders, 2013; Lacour og Tissington, 2011).

Mor Alder, MA_{it_0}

Variablen *mor alder*, MA_{it_0} , defineres som individ i 's mors alder, i begyndelsen af det år, individet færdiggjorde grundskolen. Variablen opgøres i hele år og er derfor en diskret variabel. Det forventes ud fra resultater fra Jensen og Nielsen (2010), at der er en positiv sammenhæng mellem unges mødres alder og unges sandsynlighed for at gennemføre en uddannelse på et højere niveau.

¹Max. værdier fjernet pga. risiko for identificering

Af tabel C.6 fremgår en oversigt over karakteristika for MA_{it_0} i model 1 og model 2. Medianen i begge modeller er 43 år.

Tabel C.6: Karakteristika for Variablen Mor Alder, Opgjort i År - Model 1 og Model 2

	1st Qu.	Median	Mean	3rd Qu. ²
Model 1	40,00	43,00	42,93	46,00
Model 2	40,00	43,00	43,66	47,00

Kilde: Egne beregninger på data fra Danmarks Statistiks Forskerservice

På baggrund af resultaterne fra Jensen og Nielsen (2010) forventes det, at individer med yngre mødre klarer sig dårligere i grundskolen, da yngre mødre ikke er klar til at give individerne den støtte, de har behov for (Jensen og Nielsen, 2010). At individer har en yngre mor, er dog ikke nødvendigvis ensbetydende med, at moren ikke er i stand til at opfylde alle barnets behov. Et bedre mål set i dette perspektiv vil derfor være mødres forældreegenskaber. Det er dog ikke muligt at opnå information herom i det tilgængelige data.

Mor Familietype, MF_{it_0}

MF_{it_0} er en dummy variabel, der angiver individ i 's mors familietype. Hvis $MF_{it_0} = 1$ betegnes morens familietype som *enlig*. Hvis $MF_{it_0} = 0$ betegnes morens familietype som *sammen*, hvilket betyder, at hun bor sammen med et andet voksent menneske. Variablen indeholder dermed ikke direkte information om, hvorvidt individ i 's juridiske forældre er samlevende/gift med hinanden eller om forældrene bor med andre voksne mennesker (Danmarks Statistik, n.d.e). Der er dog en korrelation mellem mors familietype og fars familietype, jf. bilag G. Variablen benyttes, som en kontrolvariabel i den kvantitative analyse. T-testsene viser at der eksisterer et lineært forhold mellem MF_{it_0} og den afhængige variabel i model 1 men ikke i model 2. Af nedenstående tabel C.7 fremgår det, at 85,59% og 87,61% af individernes mødre har familietypen *sammen* i henholdsvis model 1 og model 2.

Tabel C.7: Oversigt over Procentvis Fordeling mellem Mor Familietype - Model 1 og Model 2

	Sammen	Enlig
Model 1	85,59 %	14,41 %
Model 2	87,61 %	12,39 %

Kilde: Egne beregninger på data fra Danmarks Statistiks Forskerservice

Det forventes, at der er en negativ sammenhæng mellem dummy variablen og sandsynligheden for at gennemføre en uddannelse, der er rangeret på et højere niveau. Ligesom for søskende er forklaringen på denne forventede sammenhæng, at enlige forældre har mindre tid og færre ressourcer (Benders, 2013; Lacour og Tissington, 2011). Ifølge korrelationsmatricerne i bilag G, er der dog ikke nogen korrelation mellem morens indkomst og hvorvidt hun er enlig.

I studierne inddraget i kapitel 4 er der fokus på, hvorvidt forældrene er eneforsøgere eller ikke. Med operationaliseringen af mors familietype er det muligt at tage højde for dette. Det er dog ikke muligt at differentiere mellem eksempelvis gifte forældre og fraskilte forældre, der begge bor sammen med andre.

Mor socialklasse, $MS1_{it_0}, MS2_{it_0}, MS3_{it_0}, MS4_{it_0}$

Der inddrages dummy variable, der angiver morens socialklasse i regressionsmodellerne. Morens socialklasse defineres som morens væsentligste indkomstkilde eller beskæftigelse. Denne definition er foretaget med udgangspunkt i Danmarks Statistiks definition af socioøkonomisk klassifikation (Danmarks Statistik, n.d.v). I den kvantitative analyse defineres 5 forskellige socialklasser, henholdsvis; lønmodtager, selvstændig, pension, overførselsindkomst og andre. Den største gruppe er lønmodtagere, jf. tabel

²Max. værdier fjernet pga. risiko for identificering

C.8, hvorfor denne benyttes som benchmark. Et individs mor tilhører denne gruppe, hvis hverken $MS1_{it_0}$, $MS2_{it_0}$, $MS3_{it_0}$ eller $MS4_{it_0}$ er lig med 1. Hvis $MS1_{it_0} = 1$ tilhører individ i 's mor gruppen selvstændig. Hvis $MS2_{it_0} = 1$ tilhører individ i 's mor gruppen pension, der inkluderer folkepensionister og efterlønsmodtagere mv. (Danmarks Statistik, n.d.v). Hvis $MS3_{it_0} = 1$ er individ i 's mors primære indkomstkilde overførselsindkomst. Mor tilhører derimod gruppen andre, hvis $MS4_{it_0} = 1$. Denne gruppe er karakteriseret ved, at personerne deri har ingen eller lille tilknytning til arbejdsmarkedet (Danmarks Statistik, n.d.v). En detaljeret beskrivelse af operationaliseringen af dummy variablene fremgår af bilag A. Variablene benyttes, som kontrolvariable i regressionsmodellerne. Af t-testene for model 1, fremgår det, at det ikke har en signifikant effekt, at mor er selvstændig eller i gruppen pension. Dermed eksisterer der ikke et linenært forhold mellem dummy variablene og den afhængige variabel, jf. tabel E.13. I model 2 er der ikke en signifikant sammenhæng, hvis moren er i gruppen pension eller i gruppen andre, jf. tabel E.29.

Af nedenstående tabel C.8 fremgår en oversigt over den procentvis fordeling mellem kategorierne i model 1 og model 2. Som det fremgår af tabellen, er fordelingen i modellerne rimelig ens. Der er dog en større andel af lønmodtagere og en mindre andel af mødre på overførselsindkomst i model 2.

Tabel C.8: Oversigt over Procentvis Fordeling mellem Mor Socialklasse - Model 1 og Model 2

	Lønmodtager	Selvstændig	Pension	Overførselsindkomst	Andre
Model 1	81,65 %	3,68 %	0,27 %	12,73 %	1,67 %
Model 2	85,25 %	3,98 %	0,28 %	8,95 %	1,53 %

Kilde: Egne beregninger på data fra Danmarks Statistiks Forskerservice

Studier præsenteret i det empiriske litteraturstudie finder, at individer fra en højere socialklasse har større sandsynlighed for at begynde på ekstra uddannelse efter en ungdomsuddannelse (Becker og Hecken, 2009; Østergaard-Tygesen, 2018). Dermed forventes det, at en højere socialklasse medfører større sandsynlighed for at gennemføre en uddannelse på et højere niveau. I nærværende tilfælde er det dog ikke umiddelbart tydeligt, hvilke kategorier der kan defineres som høj eller lav socialklasse. Derfor er det svært at fortolke på estimaterne for denne faktor. Socialklasse bliver af andre defineret ud fra individers arbejde, uddannelse og økonomi (Becker og Hecken, 2009; Østergaard-Tygesen, 2018; Danmarks Statistik, n.d.v). For at kunne sammenligne estimaterne med forventninger hertil rangeres dummy variablene for socialklasse også ud fra disse parametre.

En lønmodtager kan umiddelbart have alle niveauer af uddannelse. Dette gælder også de andre socialklasser, hvilket gør det svært at rangere grupperne ud fra denne parameter. Tilknytning til arbejdsmarkedet og arbejdsopgaver er til en vis udstrækning nemmere at rangere ud fra. Personer på overførselsindkomst, pension og i kategorien andre har en lille tilknytning til arbejdsmarkedet og forventes derfor, ud fra denne parameter, at være en lavere socialklasse end lønmodtagere. Selvstændige kan udføre de samme opgaver, som lønmodtagere, hvorfor det er svært at bestemme en rangering af socialklasse heraf. Ud fra det økonomiske parameter er det dog muligt at undersøge henholdsvis median- og gennemsnitsindkomst for grupperne. Ifølge data fra Danmarks Statistik er gennemsnitslønnen for selvstændige højere end gennemsnitslønnen for lønmodtagere. Omvendt er medianlønnen for selvstændige lavere end den er for lønmodtagere. Generelt er der dog stor variation i selvstændiges indkomst (Danmarks Statistik, 2018b). Ud fra et indkomstmæssigt perspektiv er det dermed heller ikke muligt at vurdere, hvorvidt selvstændige kan defineres som en højere socialklasse end lønmodtagere. Ud fra dette parameter er personer på overførselsindkomst, pension og i kategorien andre også rangeret lavere end lønmodtagere. Det forventes, derfor at individer, hvis mor er lønmodtager, har en større sandsynlighed for at gennemføre en uddannelse på et højere niveau end individer, hvis mor er på overførselsindkomst, pension og i kategorien andet. Forventningerne til selvstændige i forhold til lønmodtagere er tvetydige, da det er svært at rangere denne gruppe. I t-testen for socialklasse i bilag E er gruppen med selvstændige positiv, men er kun signifikant i model 2, jf. tabel E.12 og E.28. Dette betyder, at individer, hvis mor var selvstændig i 2002, har en højere sandsynlighed for at have gennemført en uddannelse på et højere niveau end individer hvis mor er i anden socialklasse. Dette taler for at selvstændige tilhører en højere socialklasse end de andre grupper. Operationaliseringen af socialklasse er udfordret af, at det er svært at rangere ud fra

definitionen fra Danmarks Statistik. Derudover er definitionen af socialklasse bred, hvilket gør, at der er flere parametre, socialklasser kan rangeres ud fra. Uddannelse, arbejde og økonomi er ifølge Benders (2013) bare en del af, hvad der definerer en socialklasse. Det er dog ikke muligt at tage udgangspunkt i eksempelvis menneskelige kvaliteter, da de ikke er tilgængelige i det anvendte data.

Der er i hypoteserne valgt, at tage udgangspunkt i forældres indkomst og uddannelsesniveau, da der er færre udfordringer med operationaliseringen af disse variable. Dette er valgt, selvom socialklasse er defineret til at tage højde for både indtægtskilde og tilknytning til arbejdsmarkedet. Korrelationen mellem mødres socialklasse og indkomst, samt mellem mødres socialklasse og uddannelsesniveau er beskrevet under de tilhørende forklarende faktorer i kapitel 7.

Far Alder, FA_{it_0}

Variablen *far alder*, FA_{it_0} , defineres som individ i 's fars alder, i begyndelsen af det år individet færdiggjorde grundskolen. Variablen opgøres i hele år og er derfor en diskret variable. Af tabel C.9 fremgår en oversigt over karakteristika for FA_{it_0} i model 1 og model 2. Medianen i model 1 er 45 år, mens den er i model 2 er 46 år.

Tabel C.9: Karakteristika for Variablen Fars Alder, Opgjort i År - for Model 1 og Model 2

	1st Qu.	Median	Mean	3rd Qu. ³
Model 1	42,00	45,00	45,78	49,00
Model 2	43,00	46,00	46,39	49,00

Kilde: Egne beregninger på data fra Danmarks Statistiks Forskerservice

Der er ikke fundet empiriske studier, der har undersøgt sammenhængen mellem fars alder og uddannelsespræstation eller individers højeste fuldførte uddannelse. Derfor er ikke nogen forventninger til, hvilken sammenhæng variabelen har med den afhængige variabel. Farens alder inkluderes derfor i regressionsmodeller, som en kontrolvariabel.

Far Familietype, FF_{it_0}

FF_{it_0} er en dummy variabel, der angiver individ i 's fars familietype. Hvis $FF_{it_0} = 1$ betegnes farens familietype som *enlig*. Hvis $FF_{it_0} = 0$ betegnes farens familietype som *sammen*. Mere information omkring definitionen heraf fremgår af afsnittet, mor familietype, MF_{it_0} . Variablen benyttes, som kontrolvariabel i regressionsmodellerne. Af nedenstående tabel C.10 fremgår det, at over 85% af individernes fædre har familietype *sammen* i både model 1 og model 2.

Tabel C.10: Oversigt over Procentvis Fordeling mellem Far Familietype - Model 1 og Model 2

	Sammen	Enlig
Model 1	85,88 %	14,12 %
Model 2	88,35 %	11,65 %

Kilde: Egne beregninger på data fra Danmarks Statistiks Forskerservice

Det forventes ligesom for individers mors familietype, at der en negativ sammenhæng mellem dummy variabelen og sandsynligheden for at gennemføre en uddannelse på et højere niveau. I lighed med korrelation mellem mors familietype og hendes indkomst, eksisterer der ikke nogen korrelation mellem farens familietype og hans indkomst, jf. G.

³Max. værdier fjernet pga. risiko for identificering

Far Socialklasse, $FS1_{it_0}$, $FS2_{it_0}$, $FS3_{it_0}$, $FS4_{it_0}$

I den kvantitative analyse inddrages der dummy variable, der angiver farens socialklasse. Farens socialklasse defineres som individ i s fars væsentligste indkomstkilde eller beskæftigelse, og definitionen er foretaget med udgangspunkt i Danmarks Statistiks definition (Danmarks Statistik, n.d.v). Der er ligesom for moren defineret 5 forskellige socialklasser. Hvis $FS1_{it_0} = 1$ tilhører individ i s far gruppen selvstændig. Hvis $FS2_{it_0} = 1$ tilhører individ i s far gruppen pension (Danmarks Statistik, n.d.v). Hvis $FS3_{it_0} = 1$ er individ i s fars primære indkomstkilde overførselsindkomst. Hvis $FS4_{it_0} = 1$ tilhører faren gruppen andre. Hvis ingen af dummy variablene er lig med 1, tilhører individets far referencegruppen lønmodtagere. Variablen benyttes som en kontrolvariabel i regressionsmodellerne. Af t-testene for model 2, fremgår det, at *far socialklasse*, *selvstændig* og *far socialklasse, pension*, ikke er signifikante. Dermed eksisterer der ikke et linenært forhold mellem disse dummy variable og den afhængige variabel, jf. tabel E.32. I model 1 er alle dummy variablene for *far socialklasse* signifikante, jf. tabel E.16.

Det fremgår af tabel C.11 at over 75% af fædrene er lønmodtagere i både model 1 og model 2. Der er større andel af fædrene, der er lønmodtagere og selvstændige i model 2, mens at en mindre andel tilhører gruppen, der er på overførselsindkomst.

Tabel C.11: Oversigt over Procentvis Fordeling mellem Far Socialklasse - Model 1 og Model 2

	Lønmodtager	Selvstændig	Pension	Overførselsindkomst	Andre
Model 1	76,77	10,73	1,71	9,27	1,52
Model 2	79,06	11,29	1,69	6,52	1,44

Kilde: Egne beregninger på data fra Danmarks Statistiks Forskerservice

Forventningen til fars socialklasse er den samme, som for mors socialklasse. Der eksisterer også de samme problemer med definition af en højere socialklasse. Yderligere argumentation for forventninger og rangering fremgår af afsnittet om mors socialklasse.

C.2 Udeladte Variable

I det følgende præsenteres de udeladte variable.

Peer-gruppe

Årsagen til at peer-gruppe er fravalgt er, at det rent datateknisk ville kræve meget datamodellering at finde hver individs peer-gruppe. I studierne der inddrages i kapitel 4 defineres peer-gruppe som elever på samme skole eller uddannelse som individet (Eriksen, 2021; Jaeger, 2007). I det tilgængelige data er det muligt at finde alle elever, der færdiggjorde grundskolen i 2002 på samme skole, som individet. Det datatekniske arbejde er meget omfattende og der forventes endvidere en korrelation med andre socioøkonomiske faktorer, hvilket vil opfange noget af forklaringen peer-gruppe vil bidrage med. Af disse årsager inddrages variablen ikke i regressionsmodellerne.

Indkomst

Indkomst er ikke inddraget i regressionsmodellerne, da det er svært at finde en proxy variabel der måler den forventede livstidsindkomst og samtidig måler variationen, der er i indkomsten på de forskellige uddannelsesniveauer. For at kunne måle den forventede indkomst ved de forskellige uddannelsesniveauer præcist kræver det at efterspørgslen efter de forskellige uddannelsesniveauer kendes for hele kohortens liv. Endvidere skal der tages højde for variationen mellem uddannelser på det samme uddannelsesniveau. Overvejede mulige proxies ud fra det tilgængelige data er individers egen indkomst og medianindkomsten i 2002 for alle i befolkningen over 14, der ikke var studerende. At benytte individets egen indkomst i 2019, som en proxy for livsindkomst medfører kausalitetsproblemer. Kausalitetsproblemerne opstår da det ikke vides om det er uddannelsesniveauet der påvirker indkomsten eller om indkomsten påvirker hvilket uddannelsesniveau individerne har valgt. Derudover er det ikke muligt at tage højde for variationer

i indkomsten over livscyklussen eller midlertidige stød hertil, da individer kun observeres i et år på arbejdsmarkedet (Eriksen, 2021). Det forventes dog at individers indkomst kan forklare en stor del af variationen i de afhængige variable. I tabel K.7 og tabel K.8 i bilag K fremgår det individers indkomst ved OLS-estimering kan forklare henholdsvis 59,7% i model 1 og 69,2% af variationen i individers højeste fuldførte uddannelsesniveau. Det er testet hvordan det ville påvirke estimeringen af regressionsmodellerne, der benyttes til at undersøge de opstillede hypoteser. Resultater af modellerne fremgår af bilag K. I model 1 er der to estimater mere, der er robuste og signifikante, når indkomst inkluderes i modellen på 10% signifikansniveau. AIC værdierne er næsten identiske og alle variable har det samme fortegn i begge modeller. I model 2 er der en variabel mere, der er robust og signifikant, når indkomst inkluderes. AIC værdien er næsten identisk med model 2 og alle estimater har samme fortegn i begge modeller. Indviders indkomst er dog fravalgt, som proxy for forventet livsindkomst, på grund af kausalitetsproblemet og fordi den bidrager begrænset til at forbedre regressionsmodellernes forklaringsgrad.

Medianindkomsten for alle individer, der var på arbejdsmarkedet, da kohorten færdiggjorde grundskolen, er overvejet som en proxy. Denne er overvejet, da det er den indkomst individerne kunne observere, at individer med forskellige uddannelsesniveauer tjente, da de skulle tage deres uddannelsesbeslutning. Dermed kunne individerne se, hvad de selv kunne forvente at tjene ved at tage samme uddannelsesniveau. Problemet med denne operationalisering er, at den er perfekt korreleret med den afhængige variabel. Derudover er det kun en approksimation af, hvad individer kan forvente at tjene, da den ikke viser variationen i indkomsten. Oplysninger om medianindkomsten er nødvendigvis ikke tilgængelige for individerne, hvilket betyder, at de bygger deres forventninger ud fra en række kilder, som måske ikke afspejler medianindkomsten (Bosworth og Ford, 1985).

Flere af studierne i det empiriske litteraturstudie har estimeret den forventede indkomst, som forskelle i indkomsten ved de forskellige uddannelsesniveauer, hvor der tages højde for forskelle i karakterer og/eller uddannelsestype (Nielsen og Vissing-Jorgensen, 2006).

Grundskole

Regressionsmodellerne indeholder ikke proxy variable for grundskole forhold, såsom klassestørrelse, elev-sammensætning eller lærerens kvalifikationer. Disse er fravalgt, da tidligere forskning har vist tvetydige effekter på valg af uddannelsesniveau (Krueger, 2003; Heinesen, 2010; Nielsen, 2012; Jensen og Nielsen, 2010; Koch et al., 2015; Produktivitetskommissionen, 2014; Hanushek, 2006). Disse forhold kan endvidere ændre sig gennem elevernes skoleforløb og elever kan have skiftet skole. Dette gør det svært at måle variablene, der både kan ændre sig for år til år, men også efter hvilket fag, der undervises i. Mange af forholdene er desuden svære at måle, som f.eks. individets oplevelse af grundskolen eller lærerens kvalifikationer. Der er i det tilgængelige data ikke proxies, der i tilstrækkelig grad vurderes at kunne måle disse faktorer.

Ved ikke at inddrage variable, der er vigtige for forklaringen af uddannelsesniveau bliver regressionsmodellerne misspecificeret. Effekterne af den udeladte variabel vil blive indfanget af de medtagne variable, som både kan forårsage et opadgående eller nedadgående bias, samt bias mod 0 (Wooldridge, 2016). Det vurderes, at regressionsmodellerne bliver misspecificeret ved at udlade variable for grundskole forhold, da estimaterne, der er fundet, typisk er små og tvetydige.

Bløde Faktorer

Som beskrevet i kapitel 4 er det blevet mere normalt at inddrage bløde faktorer i regressionsmodeller (Koch et al., 2015). Disse værdier er dog sværere at operationalisere, hvilket er årsagen til, at de stadigvæk behandles som en *black box*, hvor der ikke vides, hvilke værdier der tages højde for. At inddrage variable for bløde værdi kan være med til at forklare nogle af de begivenheder, hvor individer ikke opfører sig perfekt rationelt, som det forventes i neoklassisk økonomi. I flere af de sociologiske teorier, der ligger til grund for faktorerne i litteraturstudiet antages det også, at individer opfører sig rationelt, dog ikke perfekt rationelt. Ved ikke at inddrage variable for disse bløde faktorer, er der risiko for at regressionsmodellerne bliver misspecificeret. Da der er fundet korrelation mellem nogle bløde faktorer og individers socioøkonomiske faktorer kan det forventes, at disse variable vil blive biased ved ikke at medtage bløde faktorer i den kvantitative analyse. Selvom det er blevet mere normalt, at inddrage bløde faktorer,

har de fleste økonomiske modeller stadigvæk ikke det med. I nærværende tilfælde inddrages bløde faktorer ikke, da det ikke har været muligt at finde retvisende proxies for bløde faktorer i det tilgængelige data fra Danmarks Statistik. For at kunne foretage en kvantitativ undersøgelse, der inddrager bløde faktorer, vurderes det, at yderligere information om individerne og dermed data er nødvendig. Indsamlingen heraf har ikke været mulig indenfor rammerne af dette projekt.

Teori om Ordered Logistisk Regression



I dette bilag opstilles notationen for en ordered logistisk regression. En ordered logistisk regression kaldes også en *ordinal logistisk regression*, *ordered response model* eller *proportional odds logistic regression*.

Før at regressionen opstilles er det relevant at foretage nogle definitioner. Y er en ordinal, afhængig variabel med J kategorier. $P(Y \leq j)$ er den kumulative sandsynlighed for at Y er mindre end eller lig med en specifik kategori $j = 1, \dots, J - 1$. Endvidere er $P(Y \leq J) = 1$ (UCLA: Statistical Consulting Group, n.d.a). Sandsynligheden for at tilhøre en bestemt kategori eller en kategori derunder kan defineres som:

$$\frac{P(Y \leq j)}{P(Y > j)} \quad (\text{D.1})$$

for $j = 1, \dots, J - 1$, da $P(Y > J) = 0$ og det ikke er muligt at dividere med 0. Logaritmen af sandsynligheden, log-odds, kaldes også logit, således at

$$\log \frac{P(Y \leq j)}{P(Y > j)} = \text{logit}(P(Y \leq j)) \quad (\text{D.2})$$

Hvilket betyder at logaritmen af sandsynligheden for at tilhøre en bestemt kategori er lig med logit af den af den kumulative sandsynlighed for at tilhøre kategorien j .

En ordered logistisk regression kan defineres på følgende måde:

$$\text{logit}(P(Y \leq j)) = \beta_{j0} + \beta_{j1}x_1 + \dots + \beta_{jp}x_p \quad (\text{D.3})$$

hvor $\beta_{j0} + \beta_{j1}x_1 + \dots + \beta_{jp}x_p$ er modellens koefficient parametre (skæringspunkt og hældninger) med p uafhængige variable for $j = 1, \dots, J - 1$.

Når *parallel lines assumption* (referes også til som *proportional odds assumption*) overholdes, er skæringspunktet for hver kategori forskellig, men hældningen er konstant på tværs af kategorier (UCLA: Statistical Consulting Group, n.d.a). Dette betyder, at ovenstående ligning D.3 kan reduceres til:

$$\text{logit}(P(Y \leq j)) = \beta_{j0} + \beta_1x_1 + \dots + \beta_px_p \quad (\text{D.4})$$

I det statistiske software program RStudio parametriseres en ordered logistisk regression som:

$$\text{logit}(P(Y \leq j)) = \beta_{j0} - \eta_1x_1 - \dots - \eta_px_p \quad (\text{D.5})$$

hvor $\eta_i = -\beta_i$ Hvis man eksempelvis har en binær uafhængig variabel, vil koefficienten for denne, på grund af *proportional odds assumption*, være den samme for alle j kategorier. De to funktioner for henholdsvis $x_1 = 1$ og $x_1 = 0$ er:

$$\begin{aligned} \text{logit}(P(Y \leq j) \mid x_1 = 1) &= \beta_{j0} - \eta_1 \\ \text{logit}(P(Y \leq j) \mid x_1 = 0) &= \beta_{j0} \end{aligned} \quad (\text{D.6})$$

Det betyder endvidere at:

$$\text{logit}(P(Y \leq j) \mid x_1 = 1) - \text{logit}(P(Y \leq j) \mid x_1 = 0) = -\eta_1 \quad (\text{D.7})$$

Odds ratios er relevante, da udregningen heraf kan gøre fortolkningen af modellen mere gennemskuelig. Af denne årsag, præsenteres den formelle proces til udregning af odds ratios. Da eksponenten er den inverse funktion af logaritmen, er det muligt at eksponentiere begge sider af ligning D.7 og ved at gøre brug af regnereglen ($\log(b) - \log(a) = \log(b/a)$), får man:

$$\frac{(P(Y \leq j) \mid x_1 = 1) / (P(Y > j) \mid x_1 = 1)}{(P(Y \leq j) \mid x_1 = 0) / (P(Y > j) \mid x_1 = 0)} = \exp(-\eta_1)$$

For at gøre denne ligning mere simpel er det muligt at anvende substitution, hvor $\frac{P(Y \leq j) \mid x_1 = 1}{P(Y > j) \mid x_1 = 1} = p_1 / (1 - p_1)$ og $\frac{P(Y \leq j) \mid x_1 = 0}{P(Y > j) \mid x_1 = 0} = p_0 / (1 - p_0)$ (UCLA: Statistical Consulting Group, n.d.a). Dette betyder at odds ratio kan defineres som:

$$\frac{p_1 / (1 - p_1)}{p_0 / (1 - p_0)} = \exp(-\eta_1).$$

Hvilket, pga. regnereglen ($\exp(-\eta_1) = \frac{1}{\exp(\eta_1)}$), er lig:

$$\exp(\eta_1) = \frac{p_0 / (1 - p_0)}{p_1 / (1 - p_1)} \quad (\text{D.8})$$

Dette kan ligeledes skrives som følgende, hvilket kan være mere intuitiv i forhold til fortolkning.

$$\frac{(P(Y > j) \mid x_1 = 1) / (P(Y \geq j) \mid x_1 = 1)}{(P(Y > j) \mid x_1 = 0) / (P(Y \geq j) \mid x_1 = 0)} = \exp(\eta_1)$$

(UCLA: Statistical Consulting Group, n.d.a)

T-tests E

Dette bilag indeholder de *t-tests*, der er udarbejdet mellem den afhængige variabel og hver af de uafhængige variable i både model 1 og model 2. Kapitlet er opdelt i forhold til de to modeller. Først præsenteres resultatet af *t-tests*ene for variablene i model 1, hvorefter resultaterne for variablene i model 2 præsenteres.

T-tests er brugbare til at undersøge forholdet mellem to variable. I nærværende tilfælde er det interessant at undersøge, hvorvidt de uafhængige variable isoleret set har en signifikant sammenhæng med den afhængige variabel i de to modeller. Et eksempel på en lineær regressionsmodel med én uafhængig variabel:

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

β_1 måler den partielle effekt af x_i på y (Wooldridge, 2016). Hvis $\beta_1 = 0$, eksisterer der ikke noget lineært forhold mellem x_1 og y . Ændringer i x_1 vil altså ikke have nogen effekt på y (Newbold et al., 2013). For at undersøge, hvorvidt et lineært forhold eksisterer, er det muligt at opstille en hypotese:

$$H_0 : \beta_1 = 0$$

hvor alternativ hypotesen er:

$$H_1 : \beta_1 \neq 0$$

Hvis $\beta_1 \neq 0$, så har x_i en effekt på y , hvilket betyder at nulhypotesen kan forkastes (Wooldridge, 2016).

Ved at antage at hældningskoefficienten, b_1 , er normalfordelt, er det muligt at teste hypotesen:

$$t = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{b_1 - 0}{s_{b_1}} = \frac{b_1}{s_{b_1}}$$

Dette er en *Student's t*-fordeling med $(n - 2)$ frihedsgrader (Newbold et al., 2013). Hvis værdierne af t er tilstrækkeligt større end 0, forkastes H_0 . Når H_0 forkastes er det ikke muligt at afvise et forhold mellem x_1 og y (Wooldridge, 2016).

For at kunne foretage *t-tests* på dataet, skal følgende antagelser være overholdt: Dataet skal være uafhængigt, være nogenlunde normalfordelt og have homogene varianser (Wooldridge, 2016).

I det statistiske softwareprogram R, er det muligt at udarbejde *t*-statistikken for to variable ved at anvende en lineær regressionsfunktion. Resultaterne skal tolkes ud fra koefficientens, x_i , signifikansniveau, hvor stjernerne, ***, **, *, afspejler hvor sandsynligt effekten af ændringer i y_1 sker som et resultat af ændringer i x_1 . Signifikansniveauerne for koefficienterne tager udgangspunkt i *p*-værdien (Wooldridge, 2016).

Som nævnt ovenfor anvendes *t*-testen til at undersøge et lineært forhold mellem den afhængige og en uafhængig variabel. I nærværende tilfælde er den afhængige variabel uddannelsesniveauet i 2019 for den kohorte, der blev færdig med grundskolen i 2002. I operationaliseringen af den afhængige variabel i kapitel 6, inddeles uddannelsesniveauerne som en ordered afhængig variabel. For at udarbejde de kommende *t-test*, har det været nødvendigt at ændre uddannelsesniveauerne til en numerisk variabel fra 1-7 korresponderende med uddannelsesniveauerne opgjort i rangeringen. Model 1 indeholder niveauerne 1-4, hvor 4 omfatter alle de videregående uddannelser. Model 2 omfatter kun de videregående uddannelser, hvorfor den indeholder niveauerne 4-7. Dette kan medføre udfordringer for estimeringen, da sammenhængen mellem den afhængige og de uafhængige variable ikke nødvendigvis er lineært.

I nyere mikroøkonomisk tænkning, vurderes OLS som en god indikator til at estimere modeller, der ikke nødvendigvis er oplagte til en OLS fremgangsmåde (Angrist og Pischke, 2008). På denne baggrund er det vurderet, at fremgangsmåden med t-statistikkerne er tilstrækkelig til at undersøge forholdet mellem uddannelsesniveauerne og forskellige uafhængige variable, til trods for at t-testen er beregnet til at teste kontinuerlige variable.

E.1 Forklarende Variable - Model 1

Karaktergennemsnit, KG_{it_0}

Tabel E.1: T-test: Uddannelsesniveau & Karaktergennemsnit - Model 1

	<i>Dependent variable:</i>
	Uddannelsesniveau, U_{it_1}
Karaktergennemsnit, KG_{it_0}	0.244*** (0.002)
Constant	1.779*** (0.014)
Observations	39,090
R ²	0.251
Adjusted R ²	0.251
Residual Std. Error	0.822 (df = 39088)
F Statistic	13,087.440*** (df = 1; 39088)
<i>Note:</i>	*p<0.1; **p<0.05; ***p<0.01

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

Mor Indkomst, MI_{it_0}

Tabel E.2: T-test: Uddannelsesniveau & Mor Indkomst - Model 1

	<i>Dependent variable:</i>
	Uddannelsesniveau, U_{it_1}
Mor indkomst, MI_{it_0}	0.00000*** (0.00000)
Constant	3.003*** (0.010)
Observations	39,090
R ²	0.024
Adjusted R ²	0.024
Residual Std. Error	0.938 (df = 39088)
F Statistic	966.664*** (df = 1; 39088)
<i>Note:</i>	*p<0.1; **p<0.05; ***p<0.01

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

Mor Uddannelsesniveau, MU_{it_0}

Tabel E.3: T-test: Uddannelsesniveau & Mor Uddannelsesniveau - Model 1

	<i>Dependent variable:</i>
	Uddannelsesniveau, U_{it_1}
Mor uddannelsesniveau, MU_{it_0}	0.124*** (0.002)
Constant	2.877*** (0.009)
Observations	39,090
R ²	0.066
Adjusted R ²	0.066
Residual Std. Error	0.918 (df = 39088)
F Statistic	2,771.331*** (df = 1; 39088)
<i>Note:</i> *p<0.1; **p<0.05; ***p<0.01	

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

Far Indkomst, FI_{it_0}

Tabel E.4: T-test: Uddannelsesniveau & Far Indkomst - Model 1

	<i>Dependent variable:</i>
	Uddannelsesniveau, U_{it_1}
Far indkomst, FI_{it_0}	0.00000*** (0.00000)
Constant	3.152*** (0.007)
Observations	39,090
R ²	0.015
Adjusted R ²	0.015
Residual Std. Error	0.942 (df = 39088)
F Statistic	611.105*** (df = 1; 39088)
<i>Note:</i> *p<0.1; **p<0.05; ***p<0.01	

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

Far Uddannelsesniveau, FU_{it_0}

Tabel E.5: T-test: Uddannelsesniveau & Far Uddannelsesniveau - Model 1

	<i>Dependent variable:</i>
	Uddannelsesniveau, U_{it_1}
Far uddannelsesniveau, FU_{it_0}	0.125*** (0.002)
Constant	2.882*** (0.009)
Observations	39,090
R ²	0.061
Adjusted R ²	0.061
Residual Std. Error	0.920 (df = 39088)
F Statistic	2,532.746*** (df = 1; 39088)
<i>Note:</i>	*p<0.1; **p<0.05; ***p<0.01

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

E.2 Kontrolvariable - Model 1**Alder, A_{it_0}**

Tabel E.6: T-test: Uddannelsesniveau & Alder - Model 1

	<i>Dependent variable:</i>
	Uddannelsesniveau, U_{it_1}
Alder, A_{it_0}	-0.306*** (0.011)
Constant	7.903*** (0.171)
Observations	39,090
R ²	0.018
Adjusted R ²	0.018
Residual Std. Error	0.941 (df = 39088)
F Statistic	729.073*** (df = 1; 39088)
<i>Note:</i>	*p<0.1; **p<0.05; ***p<0.01

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

Køn, K_{it_0}

Tabel E.7: T-test: Uddannelsesniveau & Køn - Model 1

<i>Dependent variable:</i>	
Uddannelsesniveau, U_{it_1}	
Køn, K_{it_0} <i>dummy</i>	-0.294*** (0.009)
Constant	3.437*** (0.007)
Observations	39,090
R ²	0.024
Adjusted R ²	0.024
Residual Std. Error	0.938 (df = 39088)
F Statistic	958.842*** (df = 1; 39088)
<i>Note:</i> *p<0.1; **p<0.05; ***p<0.01	

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

Etnicitet, E_{it_0}

Tabel E.8: T-test: Uddannelsesniveau & Etnicitet - Model 1

<i>Dependent variable:</i>	
Uddannelsesniveau, U_{it_1}	
Ikke dansk oprindelse, E_{it_0} <i>dummy</i>	-0.292*** (0.023)
Constant	3.302*** (0.005)
Observations	39,090
R ²	0.004
Adjusted R ²	0.004
Residual Std. Error	0.948 (df = 39088)
F Statistic	156.885*** (df = 1; 39088)
<i>Note:</i> *p<0.1; **p<0.05; ***p<0.01	

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

Kommunetype, $KO1_{it_0}$, $KO2_{it_0}$, $KO3_{it_0}$ & $KO4_{it_0}$

Tabel E.9: T-test: Uddannelsesniveau & Kommunetype - Model 1

	<i>Dependent variable:</i>
	Uddannelsesniveau, U_{it_1}
Oplandskommune, $KO1_{it_0}$ <i>dummy</i>	0.006 (0.015)
Provinskommune, $KO2_{it_0}$ <i>dummy</i>	-0.002 (0.013)
Storbykommune, $KO3_{it_0}$ <i>dummy</i>	0.058*** (0.017)
Hovedstadskommune, $KO4_{it_0}$ <i>dummy</i>	0.003 (0.014)
Constant	3.281*** (0.009)
Observations	39,090
R ²	0.0004
Adjusted R ²	0.0003
Residual Std. Error	0.950 (df = 39085)
F Statistic	3.548*** (df = 4; 39085)

Note: *p<0.1; **p<0.05; ***p<0.01

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

Søskende, S_{it_0}

Tabel E.10: T-test: Uddannelsesniveau & Søskende - Model 1

	<i>Dependent variable:</i>
	Uddannelse, U_{it_1}
Søskende, S_{it_0}	-0.024*** (0.005)
Constant	3.316*** (0.008)
Observations	39,090
R ²	0.001
Adjusted R ²	0.0005
Residual Std. Error	0.950 (df = 39088)
F Statistic	20.507*** (df = 1; 39088)

Note: *p<0.1; **p<0.05; ***p<0.01

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

Mor Alder, MA_{it_0}

Tabel E.11: T-test: Uddannelsesniveau & Mor Alder - Model 1

	<i>Dependent variable:</i>
	Uddannelse, U_{it_1}
Mor alder, MA_{it_0}	0.032*** (0.001)
Constant	1.936*** (0.044)
Observations	39,090
R ²	0.024
Adjusted R ²	0.024
Residual Std. Error	0.938 (df = 39088)
F Statistic	959.944*** (df = 1; 39088)
<i>Note:</i> *p<0.1; **p<0.05; ***p<0.01	

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

Mor Familietype, MF_{it_0}

Tabel E.12: T-test: Uddannelsesniveau & Mor Familietype - Model 1

	<i>Dependent variable:</i>
	Uddannelsesniveau U_{it_1}
Mor enig, MF_{it_0} <i>dummy</i>	-0.232*** (0.014)
Constant	3.322*** (0.005)
Observations	39,090
R ²	0.007
Adjusted R ²	0.007
Residual Std. Error	0.946 (df = 39088)
F Statistic	289.931*** (df = 1; 39088)
<i>Note:</i> *p<0.1; **p<0.05; ***p<0.01	

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

Mor Socialklasse, $MS1_{it_0}$, $MS2_{it_0}$, $MS3_{it_0}$ & $MS4_{it_0}$

Tabel E.13: T-test: Uddannelsesniveau & Mor Socialklasse - Model 1

	<i>Dependent variable:</i>
	Uddannelsesniveau, U_{it_1}
Mor selvstændig, $MS1_{it_0}$ <i>dummy</i>	0.005 (0.025)
Mor pension, $MS2_{it_0}$ <i>dummy</i>	0.024 (0.092)
Mor overførselsindkomst, $MS3_{it_0}$ <i>dummy</i>	-0.442*** (0.014)
Mor andre, $MS4_{it_0}$ <i>dummy</i>	-0.190*** (0.037)
Constant	3.348*** (0.005)
Observations	39,090
R ²	0.024
Adjusted R ²	0.024
Residual Std. Error	0.938 (df = 39085)
F Statistic	243.441*** (df = 4; 39085)

Note: *p<0.1; **p<0.05; ***p<0.01

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

Far Alder, FA_{it_0}

Tabel E.14: T-test: Uddannelsesniveau & Far Alder - Model 1

	<i>Dependent variable:</i>
	Uddannelsesniveau, U_{it_1}
Far alder, FA_{it_0}	0.018*** (0.001)
Constant	2.467*** (0.041)
Observations	39,090
R ²	0.010
Adjusted R ²	0.010
Residual Std. Error	0.945 (df = 39088)
F Statistic	411.151*** (df = 1; 39088)

Note: *p<0.1; **p<0.05; ***p<0.01

Kilde: Egne beregninger på data fra Danmarks Statistiks Forskerservice

Far Familietype, FF_{it_0}

Tabel E.15: T-test: Uddannelsesniveau & Far Familietype - Model 1

<i>Dependent variable:</i>	
Uddannelsesniveau, U_{it_1}	
Far enig, FF_{it_0} <i>dummy</i>	-0.266*** (0.014)
Constant	3.326*** (0.005)
Observations	39,090
R ²	0.010
Adjusted R ²	0.009
Residual Std. Error	0.945 (df = 39088)
F Statistic	375.863*** (df = 1; 39088)
<i>Note:</i> *p<0.1; **p<0.05; ***p<0.01	

Kilde: Egne beregninger på data fra Danmarks Statistiks Forskerservice

Far socialklasse, $FS1_{it_0}$, $FS2_{it_0}$, $FS3_{it_0}$ & $FS4_{it_0}$

Tabel E.16: T-test: Uddannelsesniveau & Far socialklasse - Model 1

<i>Dependent variable:</i>	
Uddannelsesniveau, U_{it_1}	
Far selvstændig, $FS1_{it_0}$ <i>dummy</i>	0.026* (0.016)
Far pension el. efterløn, $FS2_{it_0}$ <i>dummy</i>	-0.092** (0.037)
Far overførselsindkomst, $FS3_{it_0}$ <i>dummy</i>	-0.428*** (0.017)
Far andre, $FS4_{it_0}$ <i>dummy</i>	-0.153*** (0.039)
Constant	3.330*** (0.005)
Observations	39,090
R ²	0.018
Adjusted R ²	0.018
Residual Std. Error	0.941 (df = 39085)
F Statistic	175.287*** (df = 4; 39085)
<i>Note:</i> *p<0.1; **p<0.05; ***p<0.01	

Kilde: Egne beregninger på data fra Danmarks Statistiks Forskerservice

E.3 Forklarende Variable - Model 2

Karaktergennemsnit, KG_{it_0}

Tabel E.17: T-test: Uddannelsesniveau & Karaktergennemsnit - Model 2

	<i>Dependent variable:</i>
	Uddannelsesniveau, U_{it_1}
Karaktergennemsnit, KG_{it_0}	0.212*** (0.004)
Constant	4.640*** (0.030)
Observations	20,988
R ²	0.113
Adjusted R ²	0.113
Residual Std. Error	0.889 (df = 20986)
F Statistic	2,679.592*** (df = 1; 20986)
<i>Note:</i>	*p<0.1; **p<0.05; ***p<0.01

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

Mor Indkomst, MI_{it_0}

Tabel E.18: T-test: Uddannelsesniveau & Mor Indkomst - Model 2

	<i>Dependent variable:</i>
	Uddannelsesniveau, U_{it_1}
Mor indkomst, MI_{it_0}	0.00000*** (0.00000)
Constant	5.939*** (0.014)
Observations	20,988
R ²	0.015
Adjusted R ²	0.015
Residual Std. Error	0.937 (df = 20986)
F Statistic	312.998*** (df = 1; 20986)
<i>Note:</i>	*p<0.1; **p<0.05; ***p<0.01

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

Mor Uddannelsesniveau, MU_{it_0}

Tabel E.19: T-test: Uddannelsesniveau & Mor Uddannelsesniveau - Model 2

	<i>Dependent variable:</i>
	Uddannelsesniveau, U_{it_1}
Mor uddannelsesniveau, MU_{it_0}	0.094*** (0.003)
Constant	5.795*** (0.014)
Observations	20,988
R ²	0.040
Adjusted R ²	0.040
Residual Std. Error	0.924 (df = 20986)
F Statistic	881.903*** (df = 1; 20986)
<i>Note:</i> *p<0.1; **p<0.05; ***p<0.01	

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

Far Indkomst, FI_{it_0}

Tabel E.20: T-test: Uddannelsesniveau & Far Indkomst - Model 2

	<i>Dependent variable:</i>
	Uddannelsesniveau, U_{it_1}
Far indkomst, FI_{it_0}	0.00000*** (0.00000)
Constant	6.080*** (0.010)
Observations	20,988
R ²	0.006
Adjusted R ²	0.006
Residual Std. Error	0.941 (df = 20986)
F Statistic	129.697*** (df = 1; 20986)
<i>Note:</i> *p<0.1; **p<0.05; ***p<0.01	

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

Far Uddannelsesniveau, FU_{it_0}

Tabel E.21: T-test: Uddannelsesniveau & Far Uddannelsesniveau - Model 2

	<i>Dependent variable:</i>
	Uddannelsesniveau, U_{it_1}
Far uddannelsesniveau, FU_{it_0}	0.095*** (0.003)
Constant	5.803*** (0.014)
Observations	20,988
R ²	0.040
Adjusted R ²	0.040
Residual Std. Error	0.925 (df = 20986)
F Statistic	875.933*** (df = 1; 20986)
<i>Note:</i>	*p<0.1; **p<0.05; ***p<0.01

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

E.4 Kontrolvariable - Model 2**Alder, A_{it_0}**

Tabel E.22: T-test: Uddannelsesniveau & Alder - Model 2

	<i>Dependent variable:</i>
	Uddannelsesniveau, U_{it_1}
Alder, A_{it_0}	-0.095*** (0.017)
Constant	7.585*** (0.261)
Observations	20,988
R ²	0.001
Adjusted R ²	0.001
Residual Std. Error	0.943 (df = 20986)
F Statistic	29.926*** (df = 1; 20986)
<i>Note:</i>	*p<0.1; **p<0.05; ***p<0.01

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

Køn, K_{it_0}

Tabel E.23: T-test: Uddannelsesniveau & Køn - Model 2

<i>Dependent variable:</i>	
Uddannelsesniveau, U_{it_1}	
Køn, K_{it_0} <i>dummy</i>	-0.068*** (0.013)
Constant	6.188*** (0.009)
Observations	20,988
R ²	0.001
Adjusted R ²	0.001
Residual Std. Error	0.943 (df = 20986)
F Statistic	26.385*** (df = 1; 20986)
<i>Note:</i> *p<0.1; **p<0.05; ***p<0.01	

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

Etnicitet, E_{it_0}

Tabel E.24: T-test: Uddannelsesniveau & Etnicitet - Model 2

<i>Dependent variable:</i>	
Uddannelsesniveau, U_{it_1}	
Ikke dansk oprindelse, E_{it_0} <i>dummy</i>	-0.027 (0.034)
Constant	6.161*** (0.007)
Observations	20,988
R ²	0.00003
Adjusted R ²	-0.00002
Residual Std. Error	0.944 (df = 20986)
F Statistic	0.647 (df = 1; 20986)
<i>Note:</i> *p<0.1; **p<0.05; ***p<0.01	

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

Kommunetype, $K1_{it_0}$, $K2_{it_0}$, $K3_{it_0}$ & $K4_{it_0}$

Tabel E.25: T-test: Uddannelsesniveau & Kommunetype - Model 2

	<i>Dependent variable:</i>
	Uddannelsesniveau, U_{it_1}
Oplandskommune, $KO1_{it_0}$	0.062*** (0.020)
Provinskommune, $KO2_{it_0}$	0.047** (0.018)
Storbykommune, $KO3_{it_0}$	0.177*** (0.023)
Hovedstadskommune, $KO4_{it_0}$	0.170*** (0.019)
Constant	6.081*** (0.013)
Observations	20,988
R ²	0.005
Adjusted R ²	0.005
Residual Std. Error	0.941 (df = 20983)
F Statistic	28.328*** (df = 4; 20983)
<i>Note:</i> *p<0.1; **p<0.05; ***p<0.01	

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

Søskende, S_{it_0}

Tabel E.26: T-test: Uddannelsesniveau & Søskende - Model 2

	<i>Dependent variable:</i>
	Uddannelsesniveau, U_{it_1}
Søskende, S_{it_0}	-0.001 (0.007)
Constant	6.161*** (0.011)
Observations	20,988
R ²	0.00000
Adjusted R ²	-0.00005
Residual Std. Error	0.944 (df = 20986)
F Statistic	0.017 (df = 1; 20986)
<i>Note:</i> *p<0.1; **p<0.05; ***p<0.01	

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

Mor Alder, MA_{it_0}

Tabel E.27: T-test: Uddannelsesniveau & Mor Alder - Model 2

<i>Dependent variable:</i>	
Uddannelsesniveau, U_{it_1}	
Mor alder, MA_{it_0}	0.023*** (0.001)
Constant	5.168*** (0.063)
Observations	20,988
R ²	0.012
Adjusted R ²	0.012
Residual Std. Error	0.938 (df = 20986)
F Statistic	251.041*** (df = 1; 20986)
<i>Note:</i> *p<0.1; **p<0.05; ***p<0.01	

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

Mor Familietype, MF_{it_0}

Tabel E.28: T-test: Uddannelsesniveau & Mor Familietype - Model 2

<i>Dependent variable:</i>	
Uddannelsesniveau, U_{it_1}	
Mor enlig, MF_{it_0} <i>dummy</i>	-0.031 (0.020)
Constant	6.164*** (0.007)
Observations	20,988
R ²	0.0001
Adjusted R ²	0.0001
Residual Std. Error	0.944 (df = 20986)
F Statistic	2.504 (df = 1; 20986)
<i>Note:</i> *p<0.1; **p<0.05; ***p<0.01	

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

Mor Socialklasse, $MS1_{it_0}$, $MS2_{it_0}$, $MS3_{it_0}$ & $MS4_{it_0}$

Tabel E.29: T-test: Uddannelsesniveau & Mor Socialklasse - Model 2

	<i>Dependent variable:</i>
	Uddannelsesniveau, U_{it_1}
Mor selvstændig, $MS1_{it_0}$ <i>dummy</i>	0.083** (0.033)
Mor pension, $MS2_{it_0}$ <i>dummy</i>	-0.103 (0.123)
Mor overførselsindkomst, $MS3_{it_0}$ <i>dummy</i>	-0.161*** (0.023)
Mor andre, $MS4_{it_0}$ <i>dummy</i>	0.016 (0.053)
Constant	6.171*** (0.007)
Observations	20,988
R ²	0.003
Adjusted R ²	0.003
Residual Std. Error	0.942 (df = 20983)
F Statistic	14.745*** (df = 4; 20983)

Note: *p<0.1; **p<0.05; ***p<0.01

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

Far Alder, FA_{it_0}

Tabel E.30: T-test: Uddannelsesniveau & Far Alder - Model 2

	<i>Dependent variable:</i>
	Uddannelsesniveau U_{it_1}
Far alder, FA_{it_0}	0.014*** (0.001)
Constant	5.502*** (0.058)
Observations	20,988
R ²	0.006
Adjusted R ²	0.006
Residual Std. Error	0.941 (df = 20986)
F Statistic	130.299*** (df = 1; 20986)

Note: *p<0.1; **p<0.05; ***p<0.01

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

Far Familietype, FF_{it_0}

Tabel E.31: T-test: Uddannelsesniveau & Far Familietype - Model 2

<i>Dependent variable:</i>	
Uddannelsesniveau, U_{it_1}	
Far enig, FF_{it_0} <i>dummy</i>	-0.059*** (0.020)
Constant	6.166*** (0.007)
Observations	20,988
R ²	0.0004
Adjusted R ²	0.0003
Residual Std. Error	0.943 (df = 20986)
F Statistic	8.306*** (df = 1; 20986)
<i>Note:</i> *p<0.1; **p<0.05; ***p<0.01	

Kilde: Egne beregninger på data fra Danmarks Statistiks Forskerservice

Far Socialklasse, $FS1_{it_0}$, $FS2_{it_0}$, $FS3_{it_0}$ & $FS4_{it_0}$

Tabel E.32: T-test: Uddannelsesniveau & Far Socialklasse - Model 2

<i>Dependent variable:</i>	
Uddannelsesniveau, U_{it_1}	
Far selvstændig, $FS1_{it_0}$	0.014 (0.021)
Far pension, $FS2_{it_0}$	-0.061 (0.051)
Far overførselsindkomst, $FS3_{it_0}$	-0.076*** (0.027)
Far andre, $FS4_{it_0}$	0.101* (0.055)
Constant	6.163*** (0.007)
Observations	20,988
R ²	0.001
Adjusted R ²	0.0005
Residual Std. Error	0.943 (df = 20983)
F Statistic	3.527*** (df = 4; 20983)
<i>Note:</i> *p<0.1; **p<0.05; ***p<0.01	

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

Robusthedsundersøgelse



I dette bilag gennemgås forskellige operationaliseringer af de uafhængige variable. Der redegøres for de forskellige overvejede og undersøgte operationaliseringer, samt hvorvidt disse operationaliseringer ændrer resultaterne. Resultater for de forskellige robusthedstest fremgår ikke af projektet.

Alder, A_{it_0}

Størstedelen af individer er 15 år i starten af det år, de færdiggør grundskolen. Det er derfor undersøgt, hvorvidt det har en betydning for resultatet, hvis individers alder opdeles efter, om de er over eller under medianen. Robusthedstesten viser, at det ikke påvirker retningen eller signifikansen af estimatet i hverken model 1 eller model 2.

Karakter, KG_{it_0}

I forhold til karakterer er der overvejet flere forskellige operationaliseringer af variablen. Herunder udarbejdelse af en ordinal variabel, hvor unges karaktergennemsnit fra grundskolen deles op i kategorierne *under 02*, *mellem 02 og 7* samt *over 7*. Endvidere er det overvejet at operationalisere karaktergennemsnit som en dummy variabel, hvor man differentierer mellem karakterer *under eller lig 4* og *over 4* eller *under 02* og *02 og over*. Det er undersøgt, hvordan alle disse operationaliseringer påvirker estimatet. Uanset, hvordan variablen operationaliseres, eksisterer der en signifikant positiv sammenhæng mellem karakter og den afhængige variabel.

For kontinuerlige variable er det dog ikke god praksis at dele variablen op i grupper, da denne opdeling ofte er arbitrær og man mister en del af informationen ved at dele den op (National Library of Medicine, 2017). Derfor operationaliseres variablen karakter derfor som en kontinuerlig variabel i model 1 og model 2.

Etnicitet, E_{it_0}

Der er under 5% af stikprøven, der er indvandrere eller efterkommere, jf. tabel C.3. Derfor er disse to kategorier i modellen defineret som én kategori af individer med *ikke dansk oprindelse*. Det er dog undersøgt, hvorvidt det har en betydning, hvis indvandrere og efterkommere defineres ved hver sin dummy variabel. Resultaterne viser, at estimaterne er signifikante og går i samme retning, som estimatet når de betegnes som én kategori. Af hensyn til sammenlignelighed med individer af dansk oprindelse, samles disse i en kategori. Dermed er det muligt at fortolke resultaterne af at være en person af ikke dansk oprindelse, sammenlignet med en person af dansk oprindelse. Hvis der var to dummy-variable, ville man i stedet sammenligne indvandrere med personer af dansk oprindelse og efterkommere.

Kommunetype, $KO1_{it_0}$, $KO2_{it_0}$, $KO3_{it_0}$, $KO4_{it_0}$

Kommunetype er en kontrolvariabel, der fungerer som en proxy for den geografiske afstand til uddannelsesinstitutioner. En anden operationalisering heraf kunne være at inddrage en dummy variabel for alle kommuner i Danmark undtagen én. Denne operationalisering er ikke foretaget i praksis af flere årsager. Når man arbejder med dummy-variable og aflæser estimaterne sammenlignes der med alle andre kategorier. Eksempelvis undersøges Aalborg Kommune overfor alle andre kommuner. Det vurderes at være for omfattende et arbejde at fortolke på 97 dummy-variable for kommuner. Endvidere vurderes det ikke, at det vil fungere som en bedre proxy for afstanden til uddannelsesinstitutioner.

Det er overvejet at benytte en anden kommunetype end *landkommune* som benchmark. Hvis denne operationalisering foretages, er det dermed muligt at undersøge, hvorvidt landkommune er forskellig fra de andre kommunetyper.

Søskende, S_{it_0}

I forhold til operationaliseringen af variabelen S_{it_0} er det undersøgt, hvorvidt det har en betydning at operationalisere denne som *under median* eller *median og over*. Retningen af estimatorne eller signifikansen heraf ændres ikke ved denne operationalisering i hverken model 1 eller model 2. Af denne årsag bevares variabelen som en diskret variabel, da der derved bibeholdes mest mulig information herfra.

Forældres alder, MA_{it_0}, FA_{it_0}

Det er undersøgt, hvilken betydning det har at operationalisere henholdsvis *mor alder* og *far alder* på anden vis. Det er overvejet at udarbejde en dummy variabel, hvor der differentieres mellem, om individers mødre og fædre er henholdsvis over eller under/lig med median. Hvis variablene operationaliseres på denne måde, findes en signifikant negativ sammenhæng mellem forældres alder og unges højeste fuldførte uddannelse. Dette er forskelligt fra resultaterne i model 1 og model 2. Det vurderes dog, at denne operationalisering vil medføre at meget af variabelens information går tabt. Operationaliseringen, der fremgår af bilag C, bibeholdes for at bevare så meget af variabelens information som muligt.

Forældres indkomst, MI_{it_0}, FI_{it_0}

Det er overvejet, hvorvidt det vil have en betydning for variablene *mor indkomst* og *far indkomst* at operationalisere dem som over/under median. Det er dog, som nævnt ovenfor ikke god praksis at dele kontinuerlige variable op i grupper, da man mister en del af informationen (National Library of Medicine, 2017). Da MI_{it_0} og FI_{it_0} er forklarende faktorer, vurderes det at være mest hensigtsmæssigt at bevare så meget information som muligt. Af denne årsag operationaliseres variablene derfor som kontinuerlige variable i model 1 og model 2.

Forældres socialklasse, $MS1_{it_0}, MS2_{it_0}, MS3_{it_0}, MS4_{it_0}, FS1_{it_0}, FS2_{it_0}, FS3_{it_0}, FS4_{it_0}$

Operationaliseringen af forældres socialklasse er foretaget på baggrund af Danmarks Statistiks definition af socioøkonomisk klassifikation, jf. bilag A. Som det fremgår af bilaget, udgør eksempelvis gruppen af lønmodtagere flere forskellige grupper for lønmodtagere. Andre operationaliseringer er overvejet i form af differentiering mellem lønmodtagere. Et andet alternativ er kun at differentiere mellem personer, der er tilknyttet til arbejdsmarkedet og personer, der ikke er tilknyttet til arbejdsmarkedet. I dette tilfælde vil *lønmodtagere* og *selvstændige* udgøre den ene gruppe, mens *pension*, *overførselsindkomst* og *andre* vil udgøre den anden gruppe.

Overordnet er formålet med operationaliseringen at foretage en rangering heraf. Som det fremgår af bilag C har det dog været vanskeligt at rangere klasserne på en skala fra lav til høj. Det vurderes i denne sammenhæng, at ovennævnte operationaliseringer vil besværliggøre rangeringen yderligere. Af denne årsag operationaliseres forældres socialklasse som 4 dummy-variable.

Forældres Uddannelse, MI_{it_0}, FI_{it_0}

Der er overvejet flere operationaliseringer af variablene *mor uddannelse* og *far uddannelse*. En af operationaliseringerne, der er overvejet, er henholdsvis opdeling af ikke erhvervskompetencegivende og erhvervskompetencegivende uddannelser, hvor der dermed differentieres mellem henholdsvis *grundskole* og *gymnasial uddannelse* og alle resterende uddannelser. Da MI_{it_0} og FI_{it_0} defineres som forklarende faktorer, vurderes det dog at være væsentlig for analysen, at så meget information bevares som muligt. Operationaliseringen er derfor foretaget således, at forældres uddannelse opdeles efter samme rangering som de unges uddannelse.

Korrelationsmatricer



Følgende bilag indeholder en gennemgang af korrelation og kovarians. Derudover er der en korrelationmatrice for model 1 og model 2, hvor korrelationen mellem de uafhængige variable vises.

Kovarians måler den lineære afhængighed mellem to variable. En positiv kovarians indikerer at to variable bevæger sig i den samme retning og omvendt med en negativ kovarians. Kovariansestimateret afhænger af måleenheden, hvilket korrelationskoefficienten ikke gør. Korrelationen mellem x_i og z_i kan skrives som:

$$\text{Corr}(x_i, z_i) \equiv \frac{\text{Cov}(x_i, z_i)}{\text{sd}(x_i) \cdot \text{sd}(z_i)} = \frac{\sigma_{xz}}{\sigma_x \sigma_z}$$

Hvis x_i og z_i er uafhængige af hinanden betyder det $\text{Corr}(x_i, z_i) = 0$. Korrelationen er, lige som kovariansen, et mål for en lineær afhængighed mellem to variable (Newbold et al., 2013; Wooldridge, 2010). Dog betyder $\text{Corr}(x_i, z_i) = 0$ ikke nødvendigvis at x_i og z_i er uafhængige af hinanden, men derimod blot at der ikke eksisterer et lineært forhold mellem dem.

En korrelationsmatrice giver et estimat for, hvordan to variable i gennemsnit varierer i forhold til hinanden.

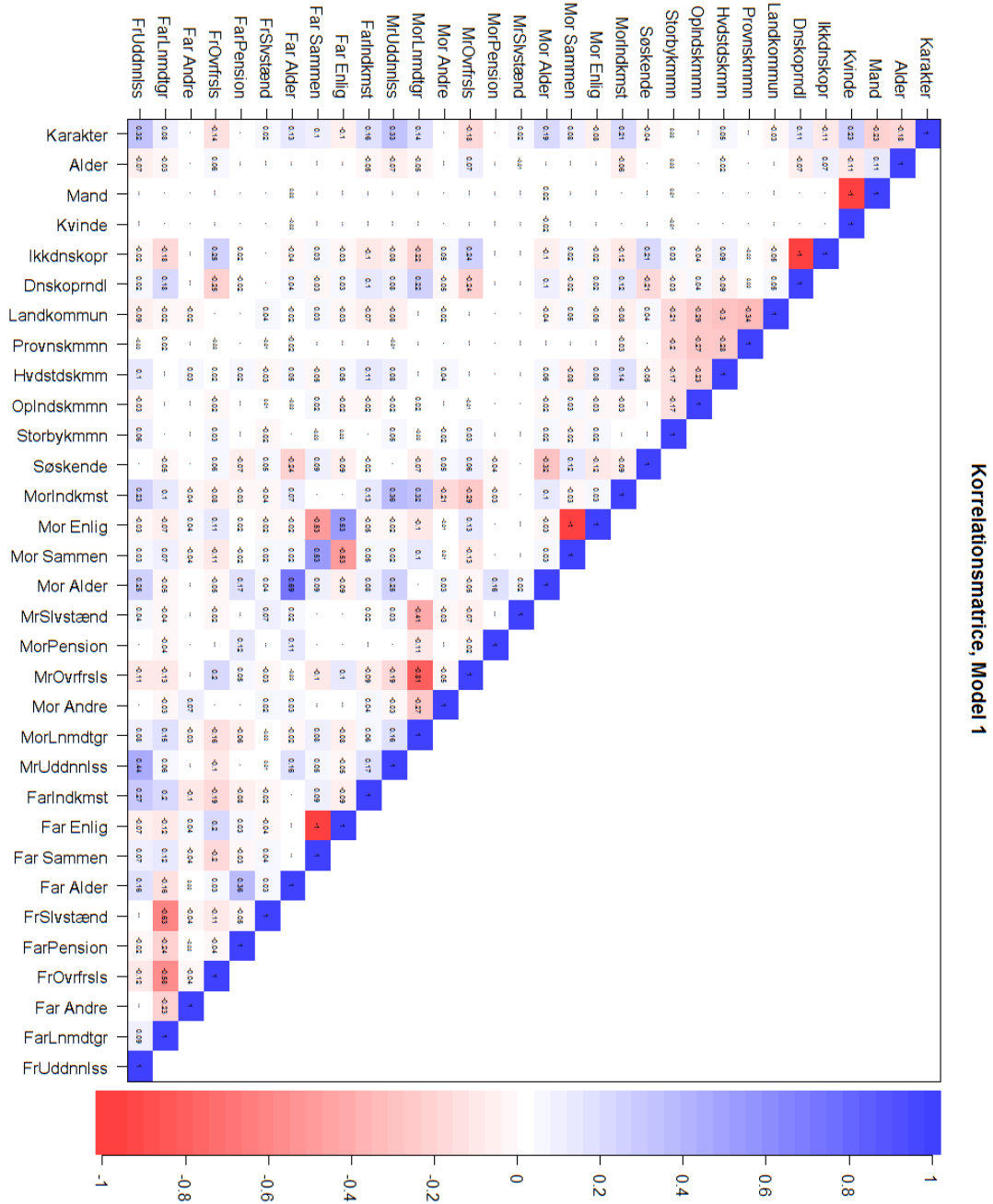
Nedenstående korrelationsmatricer er udarbejdet som *Pearson Product Moment Correlation* (PPMC). Denne bliver ofte brugt til at estimere forholdet mellem variable (Asri et al., 2016). Det gør den ved at forsøge at tegne en lige linje mellem de to variable, der på bedst mulig måde passer til forholdet mellem de to. Korrelationskoefficienten angiver, hvor godt datapunkterne passer til linjen. Korrelationskoefficienterne kan tage værdierne $-1 \leq \text{Corr}(x_i, z_i) \leq 1$. $\text{Corr}(x_i, z_i) = 1$ indikerer et perfekt positivt lineært forhold. $\text{Corr}(x_i, z_i) = -1$ indikerer et perfekt negativt lineært forhold. De ekstreme forhold med -1 samt 1 sker sjældent i praksis. Dog indikerer værdier tæt på -1 samt 1 et stærkt lineært forhold (Newbold et al., 2013). Ifølge Lærd Statistics (n.d.), vurderes det ofte, at en korrelation mellem 0,8 og 1 indikerer en stærk korrelation mellem variablene, hvor en korrelation mellem 0,5 og 0,8 indikerer en medium korrelation. Hvis korrelationskoefficienterne er mindre end 0,5 tolkes det som en svag korrelation, hvor koefficienter tæt på 0 signalerer mangel på korrelation. I denne type tests tages der ikke højde for, hvorvidt variablene er afhængige eller uafhængige, hvorfor nedenstående matricer ikke indeholder den afhængige variabel, uddannelsesniveau.

For at estimere Pearson's korrelationskoefficienter korrekt, er det blandt andet nødvendigt at variablene er kontinuerlige og normalfordelte. I nærværende tilfælde er dette dog ikke gældende for alle de uafhængige variable, hvorfor det ikke er muligt at vurdere, hvor korrekte estimaterne er. Den største udfordring med korrelationsmatricen i nærværende tilfælde, er de mange variable opgjort som dummies. I praksis vil dummies i nogle tilfælde have en perfekt negativ korrelation ud fra deres operationaliseringen. Det kan ses for korrelationskoefficienten mellem *Far enlig* og *Far sammen*, som er -1 i de nedenstående matricer. Udover udfordringerne med dummies, er der i nærværende tilfælde også gjort brug af diskrete variable, hvilke ikke overholder antagelsen om at være kontinuerlige og normalfordelte.

En anden måde, hvorpå det er muligt at udregne korrelationskoefficienter er *Spearman's Rank-Order Correlation*. Denne metode tager højde for at variablene kan være ordinale. For at Spearman's kan tage højde for det ordinale forhold kræver det dog, at variablene er rangeret, hvorfor det ikke vurderes at være anvendeligt i forhold til dummy variablene i indeværende problemstilling (Lærd Statistics, n.d.).

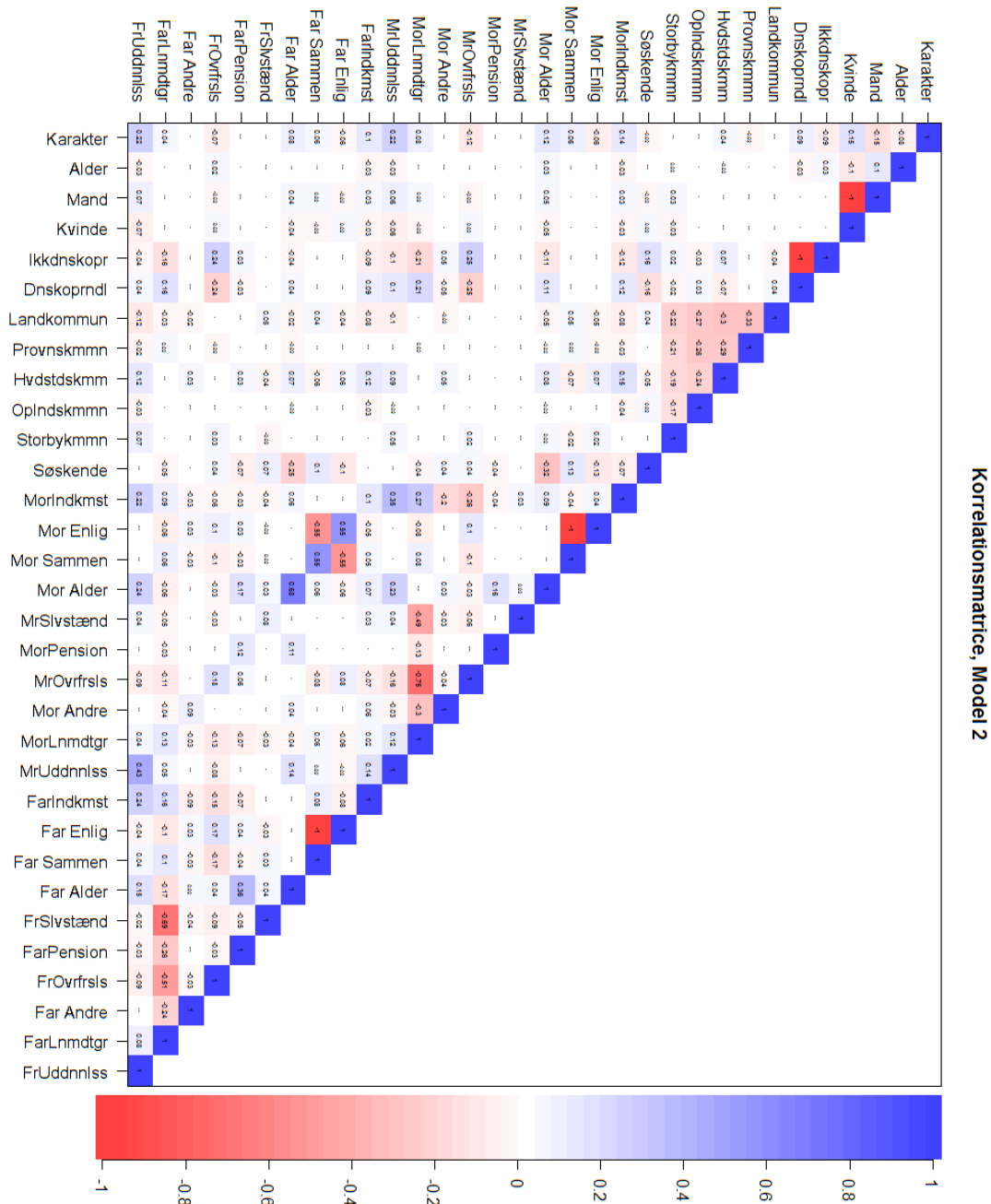
G.1 Model 1

Figur G.1: Korrelationstest af Uafhængige Variable - Model 1



G.2 Model 2

Figur G.2: Korrelationstest af Uafhængige Variable - Model 2



Odds Ratios og Konfidensintervaller



Konfidensintervaller kaldes også interval estimer, da de giver en række sandsynlige værdier for parameteren. Intervallet inkluderer punkt-estimatet, i dette tilfælde odds ratios, indenfor en på forhånd defineret sandsynlighedsramme (Wooldridge, 2016). Af nedenstående tabeller fremgår et 95% konfidensinterval, udarbejdet ved brug af standardfejl og hvor der antages en normalfordeling. Hvis intervallet ikke krydser 1, er odds ratios statistisk signifikant (UCLA: Statistical Consulting Group, n.d.b). Et eksempel på et statistisk signifikant resultat er A_{it_0} i model 1, mens et eksempel på et ikke statistisk signifikant estimat er $KO1_{it_0}$ i model 1, jf. tabel H.1 I nedenstående tabeller, tabel H.1 og tabel H.2, benyttes punktum som decimal separator.

H.1 Odds Ratios og Konfidensintervaller - Model 1

Tabel H.1: Odds Ratios og Konfidensintervaller - Model 1

	Odds ratios	2.5 %	97.5 %
KG_{it_0}	1.742	1.718	1.765
MI_{it_0}	1.024	1.010	1.039
MU_{it_0}	1.130	1.115	1.145
FI_{it_0}	1.005	0.994	1.016
FU_{it_0}	1.142	1.126	1.158
A_{it_0}	0.823	0.782	0.865
K_{it_0}	0.723	0.692	0.756
E_{it_0}	1.380	1.234	1.543
$KO1_{it_0}$	0.967	0.908	1.029
$KO2_{it_0}$	0.950	0.896	1.007
$KO3_{it_0}$	1.056	0.977	1.141
$KO4_{it_0}$	0.841	0.789	0.898
S_{it_0}	1.045	1.020	1.072
MA_{it_0}	1.024	1.017	1.031
MF_{it_0}	0.886	0.826	0.950
$MS1_{it_0}$	0.957	0.850	1.078
$MS2_{it_0}$	0.877	0.575	1.337
$MS3_{it_0}$	0.804	0.753	0.858
$MS4_{it_0}$	0.945	0.768	1.162
FA_{it_0}	1.005	1.000	1.011
FF_{it_0}	0.946	0.881	1.015
$FS1_{it_0}$	1.030	0.958	1.108
$FS2_{it_0}$	0.822	0.687	0.984
$FS3_{it_0}$	0.765	0.709	0.826
$FS4_{it_0}$	0.904	0.741	1.104

H.2 Odds Ratios og Konfidensintervaller - Model 2

Tabel H.2: Odds Ratios og Konfidensintervaller - Model 2

	Odds ratios	2.5 %	97.5 %
KG_{it_0}	1.607	1.575	1.639
MI_{it_0}	1.007	0.988	1.026
MU_{it_0}	1.097	1.080	1.113
FI_{it_0}	1.007	0.993	1.021
FU_{it_0}	1.111	1.094	1.128
A_{it_0}	0.925	0.861	0.993
K_{it_0}	1.130	1.069	1.195
E_{it_0}	1.544	1.326	1.796
$KO1_{it_0}$	1.109	1.022	1.202
$KO2_{it_0}$	1.089	1.010	1.173
$KO3_{it_0}$	1.329	1.209	1.461
$KO4_{it_0}$	1.238	1.142	1.343
S_{it_0}	1.025	0.992	1.060
MA_{it_0}	1.015	1.006	1.024
MF_{it_0}	1.047	0.949	1.154
$MS1_{it_0}$	1.199	1.038	1.386
$MS2_{it_0}$	0.643	0.383	1.079
$MS3_{it_0}$	1.007	0.913	1.111
$MS4_{it_0}$	1.112	0.845	1.463
FA_{it_0}	1.009	1.001	1.016
FF_{it_0}	1.006	0.909	1.113
$FS1_{it_0}$	1.064	0.974	1.163
$FS2_{it_0}$	0.797	0.635	0.999
$FS3_{it_0}$	1.066	0.950	1.197
$FS4_{it_0}$	1.293	1.001	1.670

Antagelser for Regressionsmodeller I

I dette bilag redegøres der for de antagelser der ligger til grund for en ordered logistisk regression og en ordinary least squares regression (OLS). Herudover undersøges det hvorvidt model 1 og model 2 samt OLS model 1 og OLS model 2 overholder deres respektive antagelser.

I.1 Ordered Logistisk Regression

Den formelle opstilling af en ordered logistisk regression fremgår af bilag D. I dette afsnit præsenteres antagelserne for en ordered logistisk regression. Det undersøges endvidere om model 1 og model 2, der præsenteres i kapitel 7 overholder antagelserne.

Proportional Odds Assumption

Som nævnt i kapitel 6, er antagelsen om proportional odds nødvendig for at kunne opstille en ordered logistisk regression. Antagelsen er, at koefficienterne, der beskriver sandsynligheden for at tilhøre den laveste kategori i forhold til alle højere kategorier af den afhængige variabel, er den samme som de, der beskriver sandsynligheden for at være mellem den næst-laveste kategori og alle højere kategorier og så fremdeles (Greene, 2018; Wooldridge, 2010).

Som nævnt i bilag D refereres proportional odds assumption og parallel lines assumption til som én og samme antagelse. Parallel lines assumption henviser til at hældningen for de forskellige funktioner er ens, hvilket grafisk kan ses ved parallelle linjer (Parry, 2016; UCLA: Statistical Consulting Group, n.d. a). Navnet proportional odds assumption, henviser til at *odds ratios* for alle $J - 1$ kategorier er den samme (UCLA: Statistical Consulting Group, n.d. a).

Antagelser for en Ordered Logistisk Regression*

Antagelse 1

Den afhængige variabel er en ordinal variabel.

Antagelse 2

En eller flere af de uafhængige variable er enten kontinuertlige, kategoriske eller ordinale.

Antagelse 3

Der eksisterer ingen multikollinearitet, hvilket betyder at to eller flere variable ikke er stærkt korreleret.

Antagelse 4

Der eksisterer proportional odds, hvilket betyder at hver af de uafhængige variable har en identisk effekt på hvert kumulative trin for den ordinale afhængige variabel.

*Kilde: University of St Andrews (n.d.)

I.1.1 Undersøgelser af Antagelserne for Ordered Logistisk Regression

Antagelse 1 er at den afhængige variabel er en ordinal variabel. Denne antagelse overholdes for både model 1 og model 2, da den afhængige variabel i begge tilfælde er en rangering af individers højeste fuldførte uddannelsesniveau, jf. kapitel 7.

Antagelse 2, der er, at en eller flere af de uafhængige variable er enten kontinuerlig, kategorisk eller ordinal, overholdes også i begge model. Af kapitel 7 fremgår det at der er kontinuerlige uafhængige variable i form af *mor indkomst* og *far indkomst* og ordinale variable i form af *mor uddannelse* og *far uddannelse*.

Antagelse 3 er at der ikke eksisterer multikollinearitet i modellen. I bilag G defineres korrelations og korrelationsmatricer for henholdsvis model 1 og model 2 fremgår af figur G.1 og figur G.2. Korrelationsmatricerne viser at der eksisterer en forholdsvis høj korrelation mellem variablene *mor alder* og *far alder* i både model 1 og model 2. Korrelationen er henholdsvis 0,69 og 0,68. Dette betyder, at individernes juridiske forældre er nogenlunde jævnaldrene. Endvidere er der en korrelation mellem *mor sammen* og *far sammen*, på 0,53 i model 1. I model 2 er korrelationen på 0,55. Det fremgår af bilag G at en korrelation mellem 0,8 og 1 vurderes at være *stærk* (Lærd Statistics, n.d.). Da korrelationen mellem *mor alder* og *far alder* samt *mor sammen* og *far sammen* er under 0,8 vurderes korrelationen derfor ikke at være stærk. Der vurderes derfor at der ikke eksisterer multikollinearitet i model 1 eller model 2 og at antagelsen dermed overholdes.

Antagelse 4 er den førnævnte proportional odds antagelse. Denne antagelse er for begge modeller undersøgt ved en Brant-test og ved brug af en såkaldt binær metode.

Antagelsen undersøges ved brug af begge metoder, da statistiske tests som Brant-testen, generelt er kritiseret for at have en tendens til indikere at proportional odds assumption ikke overholdes (Harrell Jr, 2001; UCLA: Statistical Consulting Group, n.d.a).

Nulhypotesen i Brant-testen er at proportional odds antagelsen overholdes og dermed at der eksisterer proportional odds i modellen. I tilfælde af, at værdierne der fremgår af kolonnen *probability* er under 0,05 forkastes nulhypotesen. Resultaterne af Brant-testen for model 1 fremgår af nedenstående tabel I.1.

Omnibus viser resultatet af Brant-testen for hele modellen. Som det fremgår af tabel I.1 viser Brant-testen for model 1 som helhed, at proportional odds assumption ikke overholdes. Dette resultat er ikke overraskende, da statistiske tests for antagelsen, har tendens til indikere at den ikke overholdes. citepharrell2001regression,UCLAorderedR. Det fremgår af tabellen at specielt værdien for variabelen KG_i har en særlig stor værdi, hvilket tyder på, at det er denne der driver resultatet af brant-testen. For at undersøge hvilke uafhængige variable der har problemer med proportional odds, undersøges antagelsen ved brug af en binær metode (O'Connell, 2006).

Af tabel I.2 fremgår odds ratios for tre binære modeller. I hver af modellerne behandles et cut-off, hvilket betyder at man i den første model undersøger grundskole i forhold til alle højere niveauer af uddannelse. I den anden binære model undersøges grundskole og gymnasial uddannelse i forhold til alle højere niveauer af uddannelse og så fremdeles. Odds ratios for de tre modeller sammenlignes for hver af de uafhængige variable. Hvis odds ratios er forholdsvis lig hinanden for alle tre binære modeller, forventes antagelsen at være overholdt for den pågældende uafhængige variabel.

Det fremgår af tabel I.2 at odds ratios for de tre binære modeller er forholdsvis ens for langt de fleste af variablene. Eksempelvis er odds ratios for variablene $KO1_{it_0}$, MI_{it_0} , MA_{it_0} , MF_{it_0} , $MU_{it_0t_0}$, meget lig hinanden på tværs af alle modeller. Det ser ud til at der opstår problemer med proportional odds for variablene KG_{it_0} og E_{it_0} . Odds ratios for KG_{it_0} varierer fra 1,358 til 2,031, mens at odds ratios for E_{it_0} varierer fra 0,937 til 1,910. På baggrund af tabel I.2 vurderes det dermed at være variablene KG_{it_0} og E_{it_0} der har problemer med proportional odds antagelsen.

Tabel I.1: Brant-test - Model 1

	X2	df	probability ¹
Omnibus	5,798.310	50	0.000
KG_{it_0}	3,140.111	2	0.000
MI_{it_0}	2.967	2	0.227
MU_{it_0}	139.309	2	0.000
FI_{it_0}	1.822	2	0.402
FU_{it_0}	95.941	2	0.000
A_{it_0}	27.630	2	0.000
K_{it_0}	39.378	2	0.000
E_{it_0}	148.594	2	0.000
$KO1_{it_0}$	2.344	2	0.310
$KO2_{it_0}$	11.115	2	0.004
$KO3_{it_0}$	26.190	2	0.000
$KO4_{it_0}$	54.784	2	0.000
S_{it_0}	16.377	2	0.000
MA_{it_0}	9.969	2	0.007
MF_{it_0}	14.565	2	0.001
$MS1_{it_0}$	4.350	2	0.114
$MS2_{it_0}$	0.630	2	0.730
$MS3_{it_0}$	29.412	2	0.000
$MS4_{it_0}$	3.500	2	0.174
FA_{it_0}	16.806	2	0.000
FF_{it_0}	4.560	2	0.102
$FS1_{it_0}$	3.463	2	0.177
$FS2_{it_0}$	4.221	2	0.121
$FS3_{it_0}$	6.259	2	0.044
$FS4_{it_0}$	3.970	2	0.137

¹I tabel I.1 benyttes punktum som decimal separator

Tabel I.2: Test for Proportional Odds Assumption - Model 1

	Odds Ratios Binær 1	Odds Ratios Binær 2	Odds Ratios Binær 3 ²
(Intercept)	3.751	4.667	0.034
KG_{it_0}	1.713	1.358	2.031
MI_{it_0}	1.025	1.016	1.027
MU_{it_0}	1.083	1.048	1.159
FI_{it_0}	0.995	1.001	1.009
FU_{it_0}	1.111	1.069	1.160
A_{it_0}	0.781	0.856	0.792
K_{it_0}	0.849	0.735	0.671
E_{it_0}	1.275	0.937	1.910
$KO1_{it_0}$	0.943	0.987	0.956
$KO2_{it_0}$	0.845	0.869	1.006
$KO3_{it_0}$	0.837	0.887	1.177
$KO4_{it_0}$	0.676	0.663	0.929
S_{it_0}	1.001	1.000	1.075
MA_{it_0}	1.035	1.023	1.023
MF_{it_0}	0.792	0.793	0.957
$MS1_{it_0}$	0.911	0.855	0.996
$MS2_{it_0}$	0.889	1.012	0.834
$MS3_{it_0}$	0.655	0.692	0.883
$MS4_{it_0}$	0.779	0.767	0.993
FA_{it_0}	1.000	0.996	1.011
FF_{it_0}	0.863	0.866	0.965
$FS1_{it_0}$	1.128	1.034	1.027
$FS2_{it_0}$	0.710	0.853	0.826
$FS3_{it_0}$	0.678	0.723	0.790
$FS4_{it_0}$	0.801	0.757	0.975

²I tabel I.2 benyttes punktum som decimal separator

Af følgende tabel I.3 fremgår resultatet af Brant-testen for model 2.

Resultatet af Brant-testen for model 2 er, at proportional odds antagelsen ikke overholdes i modellen, da værdien for *Omnibus* i kolonnen *probability* er mindre end 0,05. Dette er dog ikke overraskende, da antagelsen ofte ikke overholdes i komplekse modeller med mange uafhængige variable (O'Connell, 2006). Det undersøges endvidere ved en binær metode hvorvidt model 2 overholder proportional odds assumption. Af tabel I.4 fremgår odds ratios for tre binære modeller, hvor der i hver af modellerne behandles ét cutoff.

Det ses i tabel I.4 at odds ratios for langt de fleste variable er forholdsvis ens på tværs af de tre cut-offs. I model 2 varierer variablene K_i , KG_i og E_i dog på tværs af de tre modeller. Odds ratios for K_i varierer fra 0,679 til 1,534 mens at odds ratios for KG_i varierer fra 1,241 til 1,900. Ligesom i model 1 er der også i model 2 stor variation mellem odds ratios for variabelen E_i . I model 2 varierer odds ratios for E_i mellem 1,156 til 1,770, jf. tabel I.4. Endvidere er der en forholdsvis stor variation på tværs af cut-offs for $FS4_i$. Dette betyder at der disse variable har problemer med overholdelsen af proportional odds.

På baggrund af resultaterne fra Brant-testene og binær model metoden overholdes antagelse 4 ikke for hverken model 1 eller model 2.

Tabel I.3: Brant-test - Model 2

	X2	df	probability ³
Omnibus	2, 284.052	50	0.000
KG_{it_0}	1, 200.235	2	0.000
MI_{it_0}	0.954	2	0.621
MU_{it_0}	53.118	2	0.000
FI_{it_0}	0.837	2	0.658
FU_{it_0}	53.834	2	0.000
A_{it_0}	18.432	2	0.000
K_{it_0}	428.350	2	0.000
E_{it_0}	28.116	2	0.000
$KO1_{it_0}$	5.507	2	0.064
$KO2_{it_0}$	6.532	2	0.038
$KO3_{it_0}$	21.692	2	0.000
$KO4_{it_0}$	52.719	2	0.000
S_{it_0}	3.355	2	0.187
MA_{it_0}	3.228	2	0.199
MF_{it_0}	4.863	2	0.088
$MS1_{it_0}$	10.495	2	0.005
$MS2_{it_0}$	5.283	2	0.071
$MS3_{it_0}$	1.975	2	0.373
$MS4_{it_0}$	2.664	2	0.264
FA_{it_0}	7.369	2	0.025
FF_{it_0}	1.215	2	0.545
$FS1_{it_0}$	0.613	2	0.736
$FS2_{it_0}$	1.096	2	0.578
$FS3_{it_0}$	0.356	2	0.837
$FS4_{it_0}$	2.748	2	0.253

³I tabel I.3 benyttes punktum som decimal separator

Tabel I.4: Test for Proportional Odds Assumption - Model 2

	Odds Ratios Binær 1	Odds Ratios Binær 2	Odds Ratios Binær 3 ⁴
(Intercept)	0.054	0.146	0.006
KG_{it_0}	1.439	1.241	1.900
MI_{it_0}	1.017	1.011	1.002
MU_{it_0}	1.144	1.081	1.103
FI_{it_0}	1.002	1.009	1.006
FU_{it_0}	1.121	1.065	1.121
A_{it_0}	1.029	1.067	0.848
K_{it_0}	0.679	0.655	1.534
E_{it_0}	1.542	1.156	1.770
$KO1_{it_0}$	1.050	1.014	1.164
$KO2_{it_0}$	0.957	1.010	1.137
$KO3_{it_0}$	1.052	1.087	1.515
$KO4_{it_0}$	0.997	0.932	1.430
S_{it_0}	1.058	1.059	1.011
MA_{it_0}	1.013	1.009	1.019
MF_{it_0}	1.188	1.074	1.013
$MS1_{it_0}$	1.058	0.957	1.298
$MS2_{it_0}$	0.410	0.621	0.650
$MS3_{it_0}$	1.003	0.952	1.000
$MS4_{it_0}$	0.836	0.947	1.210
FA_{it_0}	1.011	1.003	1.012
FF_{it_0}	0.991	0.956	1.016
$FS1_{it_0}$	1.021	1.023	1.078
$FS2_{it_0}$	0.923	0.815	0.756
$FS3_{it_0}$	1.072	1.039	1.054
$FS4_{it_0}$	1.723	1.298	1.197

I.1.2 AIC

Akaike Information Criterion (AIC) er et mål for goodness-of-fit. Den nominelle værdi af AIC siger ikke meget i sig selv, men AIC kan benyttes til at sammenligne modeller. AIC skrives på følgende form:

$$[-2 \log L + kp]$$

Hvor L er likelihood funktionen, p er antallet af parametre i modellen og $k = 2$. (Dziak et al., 2012) AIC defineres som estimatet af en konstant plus det relative forhold mellem den ukendte likelihood-funktion af dataet samt modellens fittede likelihood funktion. Jo lavere AIC er, jo tættere er modellen på den sande model. (Dziak et al., 2012) AIC fremgår for de respektive modeller i sammenhæng med resultaternes modeller.

⁴I tabel I.4 benyttes punktum som decimal separator

I.2 Ordinary Least Squares Regression

I det følgende redegøres der for antagelserne for en ordinary least squares (OLS). OLS er en almindeligt brugt metode til estimeringen af en multipel lineær regressionsmodel (Wooldridge, 2016). Ved brug af OLS estimering behandles den afhængige variabel som en kontinuer variabel.

Antagelser for en Ordinary Least Squares Regression*

MLR 1 - Lineær i Parametrene

Forholdet mellem den afhængige variabel og de uafhængige variable er lineært, hvilket udtrykkes ved følgende:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + u$$

MLR 2 - Stokastisk Stikprøve

Der eksisterer en stokastisk stikprøve bestående af n observationer, $(x_1, x_2, \dots, x_k, y_i) : i = 1, 2, \dots, n$, der følger populationsmodellen i antagelse MLR 1.

MLR 3 - Ingen Perfekt Multikollinearitet

I stikprøven (og populationen), er ingen af de uafhængige variable konstante. Der eksisterer endvidere ikke noget perfekt lineært forhold mellem de uafhængige variable.

MLR 4 - Betinget Middelværdi lig nul

Givet hvilken som helst værdi af de uafhængige variable har fejleddet, u , en forventet værdi lig 0.

$$E(u|x_1, x_2, \dots, x_k) = 0$$

Under antagelse MLR 1 til MLR 4 er estimaterne i modellen unbiased.

MLR 5 - Homoskedasticitet

Givet enhver værdi af de uafhængige variable har fejleddet, u , samme varians.

$$Var(u|x_1, x_2, \dots, x_k) = \sigma^2$$

Under Gauss-Markov antagelserne, MLR 1 til MLR 5, er modellen defineret som BLUE; best linear unbiased estimator.

MLR 6 - Normalitet i Fejleddet

Populationsfejleddet, u , skal, foruden at være uafhængig af de uafhængige variable, have et gennemsnit på 0 og en varians på σ^2 , også være normalfordelt.

$$u \sim N(0, \sigma^2)$$

Under antagelse MLR 1 til MLR 6 kan modellen defineres som en klassisk lineær model (CML).

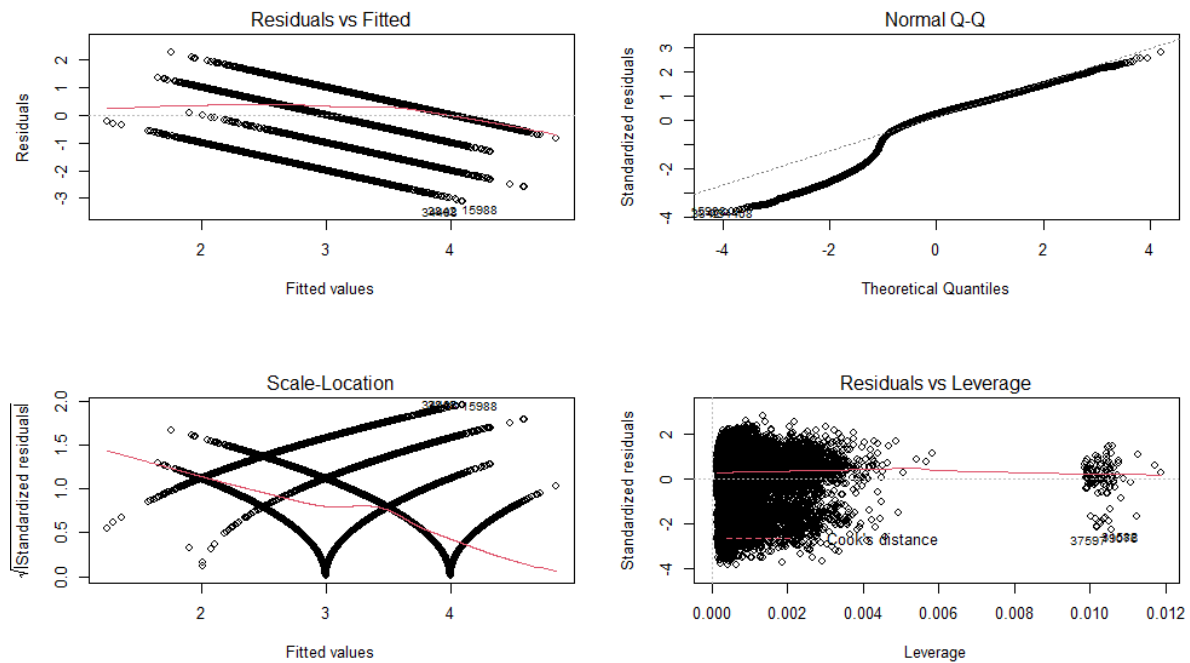
* Kilde: Wooldridge (2016), side 74-82

I.2.1 Undersøgelser af Antagelserne for OLS Regression

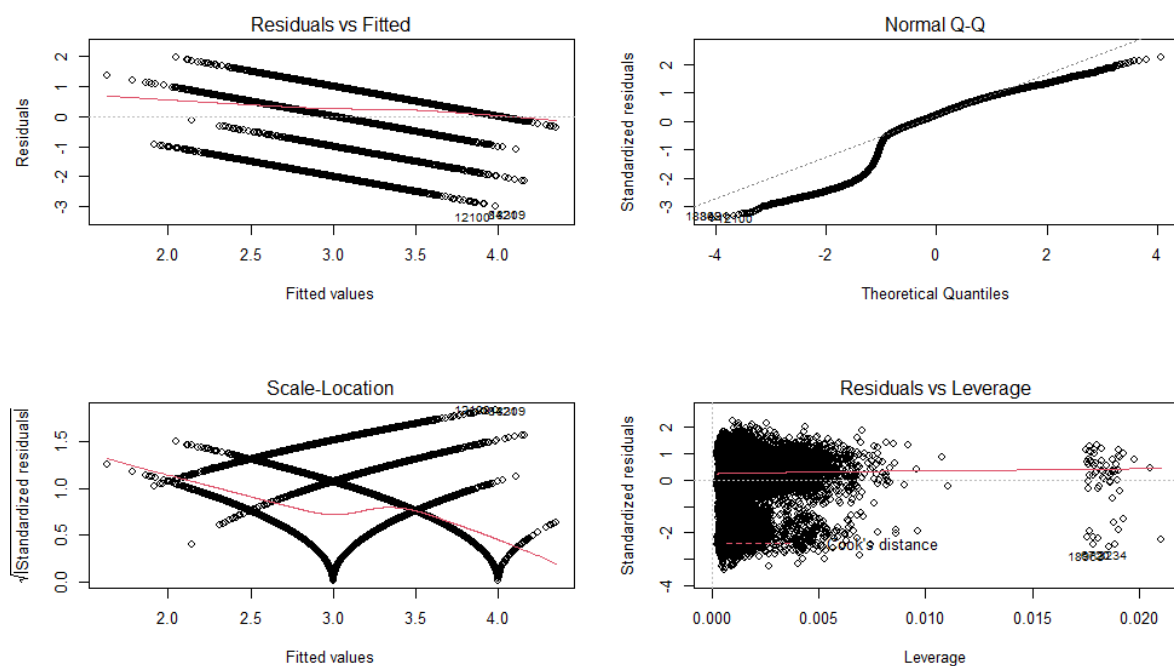
Efter opstilling af en OLS regression, er det nødvendigt at sikre at antagelserne (MLR 1 til MLR 6) overholdes. Dette gøres ved at udføres tests, undersøge modellerne grafisk og ved at vurdere modellernes karakteristika. I det følgende undersøges det både om OLS model 1 og OLS model 2 overholder antagelserne.

I MLR 1 antages det at modellen er lineær i parametrene. Det undersøges om denne antagelse overholdes i model 1 og model 2 ved brug af grafisk modelkontrol.

Figur I.1: Grafisk Regressionsdiagnostik - OLS Model 1



Figur I.2: Grafisk Regressionsdiagnostik - OLS Model 2



Af den første graf i figur I.1 og figur I.2 fremgår *Residuals vs. Fitted*, som er forholdet mellem fejlleddene

og de estimeres værdier. Des fladere den røde tendenslinje er, des mere lineært er forholdet mellem de estimerede værdier og fejlleddene. Som det fremgår af figurerne er tendenslinjen næsten helt flad for både model 1 og model 2. Derfor vurderes det, at der eksisterer et lineært forhold mellem de estimerede værdier og fejlleddene (Raza, 2019b).

Det er ikke muligt at teste MLR antagelse 2 eller undersøge den grafisk. Det vurderes dog, på baggrund af stikprøvernes størrelse på henholdsvis 39.090 og 20.988 observationer, at disse er stokastiske og at antagelse MLR 2 overholdes i både OLS model 1 og OLS model 2 (Olesen og Pedersen, 2007).

I MLR 3 antages det, at der ikke eksisterer perfekt multikollinearitet i modellen. Det er muligt at undersøge hvorvidt denne antagelse overholdes ved brug af *Variance Inflation Factor* (VIF). VIF viser hvor meget af variansen for en estimeret parameter, $\hat{\beta}_j$, er højere end i tilfælde af ingen multikollinearitet. En generel guideline er som følger; hvis VIF er større end 5 har modellen problemer med at estimere den pågældende koefficient (Adam Petrie, n.d.).

Tabel I.5: Variance Inflator Factors - OLS Model 1

Variabel	VIF	Variabel	VIF	Variabel	VIF	Variabel	VIF ⁵
KG_{it_0}	1.325	A_{it_0}	1.049	S_{it_0}	1.212	FA_{it_0}	2.203
MI_{it_0}	1.704	K_{it_0}	1.080	MA_{it_0}	2.231	FF_{it_0}	1.462
MU_{it_0}	1.388	E_{it_0}	1.202	MF_{it_0}	1.422	$FS1_{it_0}$	1.106
FI_{it_0}	1.443	$KO1_{it_0}$	1.376	$MS1_{it_0}$	1.050	$FS2_{it_0}$	1.211
FU_{it_0}	1.366	$KO2_{it_0}$	1.451	$MS2_{it_0}$	1.042	$FS3_{it_0}$	1.209
		$KO3_{it_0}$	1.280	$MS3_{it_0}$	1.179	$FS4_{it_0}$	1.305
		$KO4_{it_0}$	1.453	$MS4_{it_0}$	1.569	FU_{it_0}	1.366

Tabel I.6: Variance Inflator Factors - OLS Model 2

Variabel	VIF	Variabel	VIF	Variabel	VIF	Variabel	VIF ⁶
KG_{it_0}	1.131	A_{it_0}	1.020	S_{it_0}	1.208	FA_{it_0}	2.149
MI_{it_0}	1.663	K_{it_0}	1.051	MA_{it_0}	2.172	FF_{it_0}	1.501
MU_{it_0}	1.334	E_{it_0}	1.178	MF_{it_0}	1.478	$FS1_{it_0}$	1.098
FI_{it_0}	1.399	$KO1_{it_0}$	1.394	$MS1_{it_0}$	1.048	$FS2_{it_0}$	1.208
FU_{it_0}	1.332	$KO2_{it_0}$	1.482	$MS2_{it_0}$	1.040	$FS3_{it_0}$	1.162
		$KO3_{it_0}$	1.325	$MS3_{it_0}$	1.147	$FS4_{it_0}$	1.269
		$KO4_{it_0}$	1.507	$MS4_{it_0}$	1.527	FU_{it_0}	1.332

Det fremgår af både tabel I.5 og tabel I.6 at VIF ikke nærmer sig 5 eller 10 for nogle af variablene. Det vurderes derfor at der ikke eksisterer perfekt multikollinearitet i modellen.

Det er muligt at undersøge hvorvidt antagelse MLR 4 overholdes, ved at udregne middelværdien af fejlleddet. Disse udregninger foretages i RStudio og resultatet heraf fremgår af tabel I.7 og tabel I.8. Udregningen viser, at middelværdien er lig 0 i både OLS model 1 og OLS model 2, hvorfor antagelse MLR 4 vurderes at være overholdt i begge modeller.

Tabel I.7: Middelværdi af Fejllad - OLS Model 1

6.248987e-18⁷

Tabel I.8: Middelværdi af Fejllad - OLS Model 2

4.729536e-18⁸

⁵I tabel I.5 benyttes punktum som decimal separator

⁶I tabel I.5 benyttes punktum som decimal separator

⁷I tabel I.7 benyttes punktum som decimal separator

⁸I tabel I.8 benyttes punktum som decimal separator

Som det næste undersøges det, om modellen udviser homoskedasticitet. Af graferne *Scale-Location* i figur I.1 og figur I.2 kan det udledes hvorvidt der eksisterer en tendens til heteroskedasticitet i modellen. For at modellen opfylder antagelsen om homoskedasticitet, skal variansen af fejlleddene være forholdsvis konstant. I graferne illustreres dette ved en horisontal rød tendenslinje (Raza, 2019b). Da tendenslinjerne for både model 1 og model 2 er faldende, er der mistanke om at der eksisterer heteroskedasticitet i modellen. Dette kan testes ved brug af en *Breusch-Pagan* (BP) test. Nulhypotesen for testen er, at der eksisterer homoskedasticitet i modellen. BP-testen er foretaget i R-studio og resultatet heraf fremgår af nedenstående tabel I.9 og tabel I.10.

Tabel I.9: Breusch-Pagan Test for Homoskedasticitet - OLS Model 1

BP = 2459.6	df = 25	p-value < 2.2e-16 ⁹
-------------	---------	--------------------------------

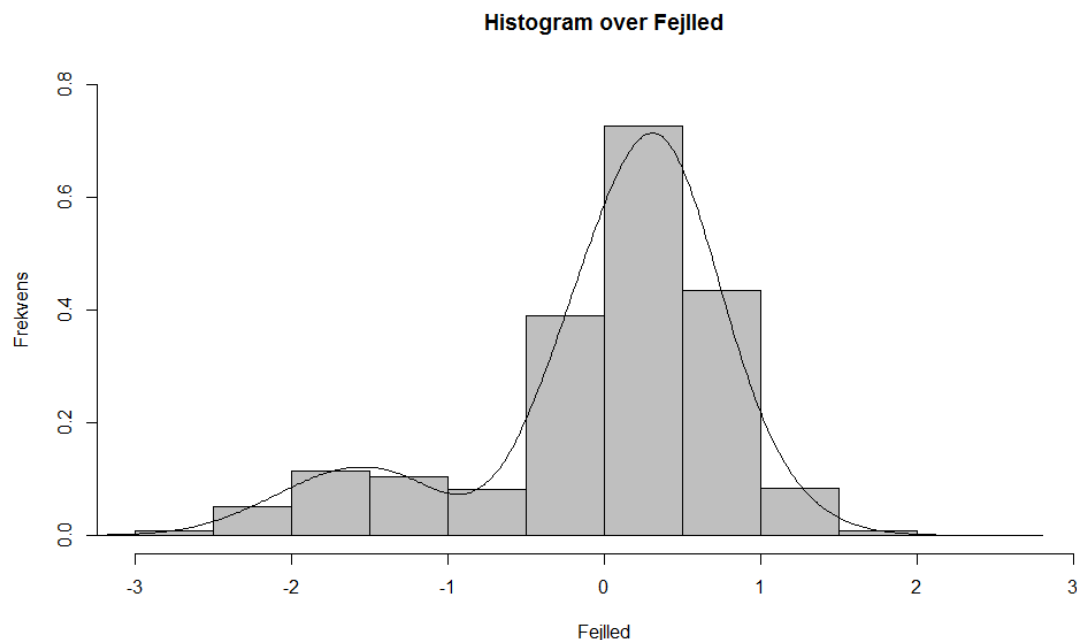
Tabel I.10: Breusch-Pagan Test for Homoskedasticitet - OLS Model 2

BP = 579.79	df = 25	p-value < 2.2e-16 ¹⁰
-------------	---------	---------------------------------

Da p-værdierne for BP-testen er meget tæt på nul for både model 1 og model 2, forkastes nulhypotesen på et 1% signifikansniveau. Der eksisterer dermed heteroskedasticitet i begge modeller. I tilfælde hvor der eksisterer heteroskedasticitet i en model er t- og F-statistikker ikke pålidelige (Wooldridge, 2016). Det er muligt at imødekomme dette problem ved at beregne robuste standardfejl og gøre brug af disse i stedet. De robuste standardfejl er udregnet for OLS model 1 og OLS model 2 og fremgår af tabel J.1 og tabel J.2 i bilag J.

Antagelse MLR 6 er normalitetsantagelsen. Denne skal overholdes for at modellen kan defineres som en klassisk lineær model (Wooldridge, 2016). Ved at udarbejde en grafisk repræsentation af fejlleddene i form af et histogram er det muligt at undersøge om denne antagelse overholdes. Histogrammerne for fejlleddene for OLS model 1 og OLS model 2 fremgår af henholdsvis figur I.3 og figur I.4.

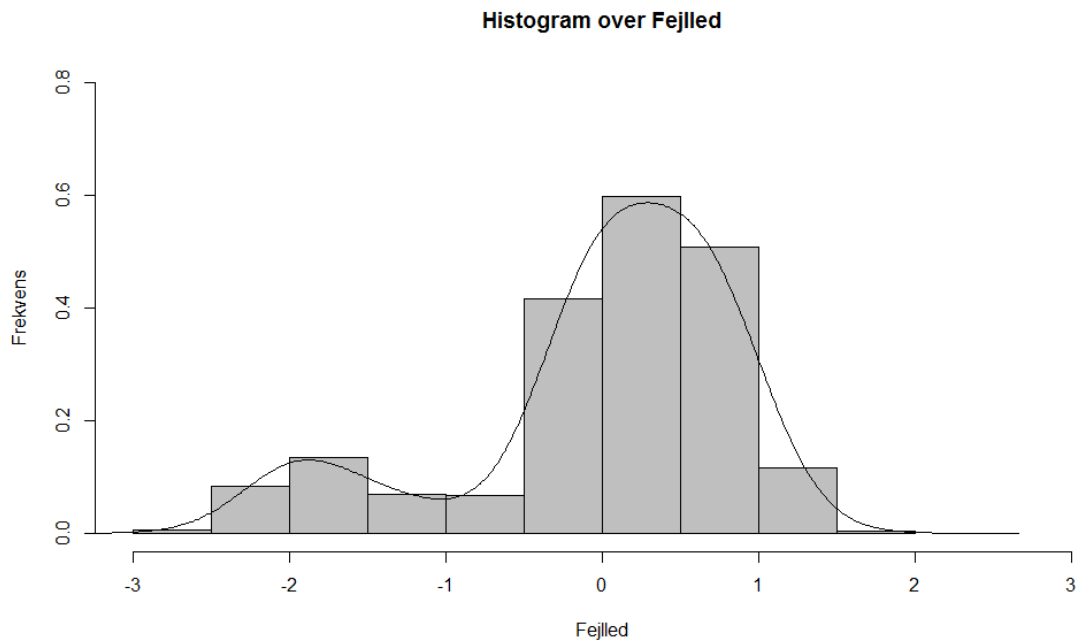
Figur I.3: Histogram over Fejled - OLS Model 1



⁹I tabel I.9 benyttes punktum som decimal separator

¹⁰I tabel I.10 benyttes punktum som decimal separator

Figur I.4: Histogram over Fejled - OLS Model 2



Af figurerne fremgår det, at fejleddene for model 1 og model 2 ikke følger en normalfordeling. Det fremgår også af graferne *Normal Q-Q* i figur I.1 og figur I.2, at fejleddene i modellerne ikke følger en normalfordeling. Hvis fejleddene følger den prikkede linje i *Normal Q-Q* grafen perfekt er de normalfordelt, hvilket ikke er tilfældet for hverken OLS model 1 eller OLS model 2 (Raza, 2019b). I små stikprøver er dette problematisk da t-statistikker og F-statistikker i en sådan situation ikke følger henholdsvis en t-fordeling og en F-fordeling. I indeværende tilfælde, hvor stikprøverne består af henholdsvis 39.090 og 20.988 observationer, vurderes det dog ikke at være problematisk da der eksisterer asymptotisk normalitet (Raza, 2019a). Dermed følger t-statistikkerne og F-statistikkerne tilnærmelsesvis en t-fordeling og F-fordeling.

OLS modeller



J.1 Model 1

Resultater for OLS Model 1 med Robuste Standardfejl

Tabel J.1*

Tabel J.1: OLS Model med Robuste Standardfejl - Model 1

	<i>Estimat</i>	<i>Std. Fejl</i>	<i>t-værdi</i>	<i>p-værdi</i>	
(Intercept)	2.805	0.176	15.940	0.000	***
KG_{it_0}	0.202	0.003	79.813	0.000	***
MI_{it_0}	0.007	0.003	2.277	0.023	**
MU_{it_0}	0.032	0.002	13.171	0.000	***
FI_{it_0}	0.001	0.002	0.372	0.710	
FU_{it_0}	0.032	0.002	12.758	0.000	***
A_{it_0}	-0.092	0.011	-8.473	0.000	***
K_{it_0}	-0.111	0.009	-12.923	0.000	***
E_{it_0}	0.103	0.025	4.088	0.000	***
$KO1_{it_0}$	-0.016	0.012	-1.289	0.197	
$KO2_{it_0}$	-0.029	0.011	-2.554	0.012	**
$KO3_{it_0}$	-0.008	0.015	-0.560	0.576	
$KO4_{it_0}$	0.096	0.013	-7.514	0.000	***
S_{it_0}	0.006	0.005	1.219	0.223	
MA_{it_0}	0.009	0.001	6.374	0.000	***
MF_{it_0}	-0.061	0.016	-3.913	0.000	***
$MS1_{it_0}$	-0.033	0.022	-1.518	0.129	
$MS2_{it_0}$	-0.070	0.067	-1.054	0.292	
$MS3_{it_0}$	-0.137	0.015	-8.929	0.000	***
$MS4_{it_0}$	-0.077	0.044	-1.752	0.080	
FA_{it_0}	0.002	0.001	1.329	0.184	
FF_{it_0}	-0.042	0.016	-2.665	0.008	**
$FS1_{it_0}$	0.014	0.013	1.057	0.290	
$FS2_{it_0}$	-0.082	0.037	-2.244	0.025	*
$FS3_{it_0}$	-0.150	0.018	-8.368	0.000	***
$FS4_{it_0}$	-0.072	0.042	-1.728	0.084	*
Antal Obs.	39,090				
R²	0.279				
Adj. R²	0.278				
F-statistik	604.2 ***	(df = 25; 39064)			
Note:	*p<0.1 **p<0.05 ***p<0.01				

*I Tabel J.1 benyttes punktum som decimal separator.

J.2 Model 2

Resultater for OLS Model 2 med Robuste Standardfejl

Tabel J.2*

Tabel J.2: OLS Model med Robuste Standardfejl - Model 2

	<i>Estimat</i>	<i>Std. Fejl</i>	<i>t-værdi</i>	<i>p-værdi</i>	
(Intercept)	1.271	0.256	4.962	0.000	***
KG_{it_0}	0.186	0.004	44.030	0.000	***
MI_{it_0}	0.003	0.004	0.762	0.446	*
MU_{it_0}	0.041	0.003	11.889	0.000	***
FI_{it_0}	0.002	0.003	0.685	0.493	
FU_{it_0}	0.040	0.003	11.600	0.000	***
A_{it_0}	-0.021	0.016	-1.293	0.196	
K_{it_0}	-0.008	0.013	-0.654	0.513	
E_{it_0}	0.169	0.035	4.838	0.000	***
$KO1_{it_0}$	0.035	0.019	1.879	0.060	
$KO2_{it_0}$	0.023	0.017	1.306	0.191	
$KO3_{it_0}$	0.095	0.021	4.509	0.000	***
$KO4_{it_0}$	0.060	0.018	3.246	0.001	**
S_{it_0}	0.014	0.008	1.868	0.062	*
MA_{it_0}	0.006	0.002	2.966	0.003	**
MF_{it_0}	0.026	0.028	1.129	0.259	
$MS1_{it_0}$	0.043	0.031	1.378	0.168	
$MS2_{it_0}$	-0.248	0.127	-1.945	0.052	*
$MS3_{it_0}$	-0.006	0.024	-0.239	0.811	
$MS4_{it_0}$	0.009	0.062	0.150	0.880	
FA_{it_0}	0.004	0.002	2.195	0.028	**
FF_{it_0}	-0.002	0.024	-0.082	0.935	
$FS1_{it_0}$	0.019	0.020	0.959	0.338	
$FS2_{it_0}$	-0.094	0.053	-1.796	0.073	*
$FS3_{it_0}$	0.021	0.027	0.761	0.447	
$FS4_{it_0}$	0.108	0.054	1.983	0.047	**
Antal Obs.	20,988				
R²	0.142				
Adj. R²	0.141				
F-statistik	138.509 *** (df = 25; 20962)				
Note: *p<0.1 **p<0.05 ***p<0.01					

*I Tabel J.2 benyttes punktum som decimal separator.

Modeller med Indkomst



K.1 Model 1

Resultater for Model 1 med Indkomst

Tabel K.1*

Tabel K.1: Ordered Logistisk Regression med Indkomst - Model 1

	<i>Estimat</i>	<i>Std. Fejl</i>	<i>t-værdi</i>	<i>p-værdi</i>	<i>Odds Ratios</i>
KG_{it_0}	0.547	0.007	78.741	0.000 ***	1.729
MI_{it_0}	0.022	0.007	3.024	0.002 ***	1.022
MU_{it_0}	0.122	0.007	18.412	0.000 ***	1.130
FI_{it_0}	0.002	0.006	0.390	0.697	1.002
FU_{it_0}	0.135	0.007	18.843	0.000 ***	1.146
A_{it_0}	-0.196	0.026	-7.616	0.000 ***	0.822
K_{it_0}	-0.354	0.023	-15.705	0.000 ***	0.702
E_{it_0}	0.286	0.069	4.152	0.000 ***	1.331
$KO1_{it_0}$	-0.022	0.032	-0.686	0.492	0.978
$KO2_{it_0}$	-0.042	0.030	-1.401	0.161	0.959
$KO3_{it_0}$	0.069	0.040	1.738	0.082 *	1.071
$KO4_{it_0}$	-0.147	0.033	-4.463	0.000 ***	0.863
S_{it_0}	0.051	0.013	3.995	0.000 ***	1.052
MA_{it_0}	0.022	0.003	6.501	0.000 ***	1.023
MF_{it_0}	-0.119	0.036	-3.315	0.001 ***	0.888
$MS1_{it_0}$	-0.013	0.061	-0.219	0.827	0.987
$MS2_{it_0}$	-0.116	0.216	-0.538	0.591	0.890
$MS3_{it_0}$	-0.198	0.033	-5.933	0.000 ***	0.820
$MS4_{it_0}$	-0.007	0.106	-0.068	0.946	0.993
FA_{it_0}	0.006	0.003	2.020	0.043 *	1.006
FF_{it_0}	-0.065	0.036	-1.802	0.072 *	0.937
$FS1_{it_0}$	0.048	0.037	1.285	0.199	1.049
$FS2_{it_0}$	-0.165	0.092	-1.795	0.073 *	0.848
$FS3_{it_0}$	-0.244	0.039	-6.230	0.000 ***	0.784
$FS4_{it_0}$	-0.102	0.102	-1.002	0.316	0.903
I_{it_1}	0.130	0.006	21.477	0.000 ***	1.145
1 2 ($\beta_{1,0}$)	1.260	0.425	2.965	0.003 ***	
2 3 ($\beta_{2,0}$)	1.776	0.425	4.181	0.000 ***	
3 4 ($\beta_{3,0}$)	3.915	0.425	9.211	0.000 ***	
Antal Obs.					39,090
Residual Deviance					68364.55
AIC					68422.55

Note: *p<0.1 **p<0.05 ***p<0.01

*I Tabel K.1 benyttes punktum som decimal separator.

Odds Ratios og Konfidensintervaller

Tabel K.2: Odds Ratios og Konfidensintervaller - Model 1 med Indkomst

	Odds ratios	2.5 %	97.5 %
KG_{it_0}	1.729	1.705	1.753
MI_{it_0}	1.022	1.008	1.037
MU_{it_0}	1.130	1.116	1.145
FI_{it_0}	1.002	0.991	1.013
FU_{it_0}	1.145	1.129	1.161
A_{it_0}	0.822	0.781	0.864
K_{it_0}	0.702	0.672	0.734
E_{it_0}	1.331	1.163	1.523
$KO1_{it_0}$	0.978	0.918	1.042
$KO2_{it_0}$	0.959	0.904	1.017
$KO3_{it_0}$	1.071	0.991	1.158
$KO4_{it_0}$	0.863	0.809	0.921
S_{it_0}	1.052	1.026	1.079
MA_{it_0}	1.023	1.016	1.030
MF_{it_0}	0.888	0.828	0.953
$MS1_{it_0}$	0.987	0.876	1.112
$MS2_{it_0}$	0.890	0.583	1.360
$MS3_{it_0}$	0.820	0.768	0.876
$MS4_{it_0}$	0.993	0.807	1.222
FA_{it_0}	1.006	1.000	1.012
FF_{it_0}	0.937	0.873	1.006
$FS1_{it_0}$	1.049	0.975	1.128
$FS2_{it_0}$	0.848	0.708	1.015
$FS3_{it_0}$	0.784	0.726	0.846
$FS4_{it_0}$	0.903	0.739	1.103
I_{it_1}	1.139	1.125	1.153

OLS model

Resultater for OLS Model 1 med Indkomst

Tabel K.3*

Tabel K.3: OLS Model med Indkomst med Robuste Standardfejl - Model 1

	<i>Estimat</i>	<i>Std. Fejl</i>	<i>t-værdi</i>	<i>p-værdi</i>	
(Intercept)	2.173	0.178	12.226	0.000	***
KG_{it0}	0.121	0.003	82.108	0.000	***
MI_{it0}	0.006	0.003	2.286	0.022	**
MU_{it0}	0.032	0.002	13.215	0.000	***
FI_{it0}	0.000	0.002	0.022	0.982	
FU_{it0}	0.032	0.002	12.810	0.000	***
A_{it0}	-0.091	0.011	-9.200	0.000	***
K_{it0}	-0.121	0.009	-14.353	0.000	***
E_{it0}	0.117	0.025	5.423	0.000	***
$KO1_{it0}$	-0.012	0.012	-0.996	0.319	
$KO2_{it0}$	-0.028	0.011	-2.443	0.015	**
$KO3_{it0}$	-0.007	0.014	-0.478	0.633	
$KO4_{it0}$	-0.094	0.013	-7.637	0.000	***
S_{it0}	0.006	0.005	1.278	0.201	
MA_{it0}	0.008	0.001	6.430	0.000	***
MF_{it0}	-0.059	0.016	-4.259	0.000	***
$MS1_{it0}$	-0.022	0.022	-1.007	0.314	
$MS2_{it0}$	-0.063	0.066	-0.792	0.428	
$MS3_{it0}$	-0.134	0.015	-10.164	0.000	***
$MS4_{it0}$	-0.060	0.043	-1.510	0.131	
FA_{it0}	0.002	0.001	1.541	0.123	
FF_{it0}	-0.044	0.016	-3.099	0.002	***
$FS1_{it0}$	0.018	0.013	1.295	0.195	
$FS2_{it0}$	-0.077	0.037	-2.242	0.025	*
$FS3_{it0}$	-0.146	0.018	-9.531	0.000	***
$FS4_{it0}$	-0.069	0.041	-1.822	0.069	*
I_{it1}	0.055	0.003	23.438	0.000	***
Antal Obs.	39,090				
R²	0.289				
Adj. R²	0.288				
F-statistik	610.200 *** (df = 26; 39063)				
Note: *p<0.1 **p<0.05 ***p<0.01					

*I Tabel K.3 benyttes punktum som decimal separator.

K.2 Model 2

Resultater for Model 2 med Indkomst

Tabel K.4*

Tabel K.4: Ordered Logistisk Regression med Indkomst - Model 2

	<i>Estimat</i>	<i>Std. Fejl</i>	<i>t-værdi</i>	<i>p-værdi</i>		<i>Odds Ratios</i>
KG_{it_0}	0.469	0.010	46.191	0.000	***	1.599
MI_{it_0}	0.008	0.010	0.801	0.423		1.008
MU_{it_0}	0.090	0.008	11.688	0.000	***	1.095
FI_{it_0}	0.005	0.007	0.745	0.456		1.005
FU_{it_0}	0.106	0.008	13.309	0.000	***	1.112
A_{it_0}	-0.082	0.037	-2.230	0.026	**	0.922
K_{it_0}	0.100	0.028	3.508	0.000	***	1.105
E_{it_0}	0.376	0.096	3.913	0.000	***	1.456
$KO1_{it_0}$	0.109	0.041	2.638	0.008	***	1.116
$KO2_{it_0}$	0.088	0.038	2.302	0.021	**	1.092
$KO3_{it_0}$	0.297	0.048	6.151	0.000	***	1.346
$KO4_{it_0}$	0.231	0.041	5.583	0.000	***	1.260
S_{it_0}	0.031	0.017	1.809	0.070	*	1.031
MA_{it_0}	0.014	0.004	3.111	0.002	***	1.014
MF_{it_0}	0.046	0.050	0.925	0.355		1.047
$MS1_{it_0}$	0.199	0.074	2.691	0.007	***	1.220
$MS2_{it_0}$	-0.428	0.264	-1.620	0.105		0.652
$MS3_{it_0}$	0.034	0.050	0.685	0.494		1.035
$MS4_{it_0}$	0.150	0.140	1.067	0.286		1.161
FA_{it_0}	0.009	0.004	2.429	0.015	**	1.009
FF_{it_0}	-0.003	0.052	-0.050	0.960		0.997
$FS1_{it_0}$	0.063	0.045	1.385	0.166		1.065
$FS2_{it_0}$	-0.195	0.115	-1.692	0.091	*	0.823
$FS3_{it_0}$	0.094	0.059	1.585	0.113		1.098
$FS4_{it_0}$	0.247	0.131	1.894	0.058	*	1.281
I_{it_1}	0.093	0.010	9.517	0.000	***	1.097
4 5 ($\beta_{1,0}$)	3.093	0.602	5.142	0.000	***	
5 6 ($\beta_{2,0}$)	3.504	0.602	5.825	0.000	***	
6 7 ($\beta_{3,0}$)	5.780	0.602	9.595	0.000	***	
Antal Obs.						20,988
Residual Deviance						42507.63
AIC						42565.63

Note: *p<0.1 **p<0.05 ***p<0.01

*I Tabel K.4 benyttes punktum som decimal separator.

Odds Ratios og Konfidensintervaller

Tabel K.5: Odds Ratios og Konfidensintervaller - Model 2 med Indkomst

	Odds ratios	2.5 %	97.5 %
KG_{it_0}	1.599	1.567	1.631
MI_{it_0}	1.008	0.989	1.027
MU_{it_0}	1.095	1.078	1.111
FI_{it_0}	1.005	0.991	1.020
FU_{it_0}	1.112	1.094	1.129
A_{it_0}	0.922	0.858	0.990
K_{it_0}	1.105	1.045	1.168
E_{it_0}	1.456	1.206	1.758
$KO1_{it_0}$	1.116	1.028	1.210
$KO2_{it_0}$	1.092	1.013	1.176
$KO3_{it_0}$	1.346	1.225	1.480
$KO4_{it_0}$	1.260	1.162	1.366
S_{it_0}	1.031	0.997	1.066
MA_{it_0}	1.014	1.005	1.023
MF_{it_0}	1.047	0.950	1.155
$MS1_{it_0}$	1.220	1.055	1.410
$MS2_{it_0}$	0.652	0.389	1.094
$MS3_{it_0}$	1.035	0.938	1.141
$MS4_{it_0}$	1.161	0.882	1.528
FU_{it_0}	1.009	1.002	1.017
FF_{it_0}	0.997	0.902	1.103
$FS1_{it_0}$	1.065	0.974	1.164
$FS2_{it_0}$	0.823	0.656	1.031
$FS3_{it_0}$	1.098	0.978	1.233
$FS4_{it_0}$	1.281	0.991	1.654
I_{it_1}	1.097	1.076	1.118

OLS model

Resultater for OLS Model 2 med Indkomst

Tabel K.6*

Tabel K.6: OLS Model med Indkomst med Robuste Standardfejl - Model 2

	<i>Estimat</i>	<i>Std. Fejl</i>	<i>t-værdi</i>	<i>p-værdi</i>	
(Intercept)	0.835	0.262	3.187	0.001	***
KG_{it_0}	0.184	0.004	43.482	0.000	***
MI_{it_0}	0.004	0.004	0.832	0.405	*
MU_{it_0}	0.041	0.003	11.838	0.000	***
FI_{it_0}	0.001	0.003	0.505	0.613	
FU_{it_0}	0.040	0.003	11.737	0.000	***
A_{it_0}	-0.021	0.016	-1.298	0.194	
K_{it_0}	-0.016	0.013	-1.331	0.183	
E_{it_0}	0.173	0.035	4.957	0.000	***
$KO1_{it_0}$	0.036	0.019	1.925	0.054	
$KO2_{it_0}$	0.022	0.017	1.264	0.206	
$KO3_{it_0}$	0.096	0.021	4.554	0.000	***
$KO4_{it_0}$	0.058	0.018	3.162	0.002	**
S_{it_0}	0.014	0.008	1.816	0.069	*
MA_{it_0}	0.006	0.002	2.903	0.003	***
MF_{it_0}	0.027	0.023	1.192	0.233	
$MS1_{it_0}$	0.048	0.031	1.544	0.122	
$MS2_{it_0}$	-0.242	0.128	-1.898	0.057	*
$MS3_{it_0}$	-0.004	0.024	-0.167	0.867	
$MS4_{it_0}$	0.020	0.062	0.325	0.744	
FA_{it_0}	0.003	0.002	2.256	0.024	**
FF_{it_0}	-0.004	0.024	-0.127	0.898	
$FS1_{it_0}$	0.018	0.020	0.895	0.371	
$FS2_{it_0}$	-0.089	0.054	-1.705	0.088	*
$FS3_{it_0}$	0.022	0.027	0.819	0.412	
$FS4_{it_0}$	0.105	0.054	1.946	0.051	*
I_{it_1}	0.036	0.005	7.632	0.000	***
Antal Obs.	20,988				
R2	0.145				
Adj. R2	0.143				
F-statistik	136.500 *** (df = 26; 20961)				

Note: *p<0.1 **p<0.05 ***p<0.01

*I Tabel K.6 benyttes punktum som decimal separator.

K.3 T-tests af Indkomst

Tabel K.7: T-test: Uddannelsesniveau & Indkomst - Model 1

	<i>Dependent variable:</i>
	Uddannelsesniveau, U_{it_1}
Forventet indkomst, I_{it_1}	0.00001*** (0.00000)
Constant	0.631*** (0.011)
Observations	39,090
R ²	0.597
Adjusted R ²	0.597
Residual Std. Error	0.603 (df = 39088)
F Statistic	57,802.090*** (df = 1; 39088)
<i>Note:</i>	*p<0.1; **p<0.05; ***p<0.01

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

Tabel K.8: T-test: Uddannelsesniveau & Indkomst - Model 2

	<i>Dependent variable:</i>
	Uddannelsesniveau, U_{it_1}
Forventet indkomst, I_{it_1}	0.00002*** (0.00000)
Constant	1.773*** (0.021)
Observations	20,988
R ²	0.692
Adjusted R ²	0.692
Residual Std. Error	0.524 (df = 20986)
F Statistic	47,170.560*** (df = 1; 20986)
<i>Note:</i>	*p<0.1; **p<0.05; ***p<0.01

Kilde: Egne beregninger baseret på data fra Danmarks Statistiks Forskerservice

Komprimerede Modeller



L.1 Model 1

Resultater for Komprimeret Model 1

Tabel L.1*

Tabel L.1: Ordered Logistisk Regression - Model 1 Komprimeret

	<i>Estimat</i>	<i>Std. Fejl</i>	<i>t-værdi</i>	<i>p-værdi</i>	<i>Odds Ratios</i>
KG_{it_0}	0.593	0.007	89.091	0.000 ***	1.810
MI_{it_0}	0.023	0.006	3.970	0.000 ***	1.023
MU_{it_0}	0.130	0.006	20.156	0.000 ***	1.139
FI_{it_0}	0.008	0.005	1.611	0.107	1.008
FU_{it_0}	0.141	0.007	20.145	0.000 ***	1.151
1 2 ($\beta_{1,0}$)	2.058	0.084	24.632	0.000 ***	
2 3 ($\beta_{2,0}$)	2.563	0.083	30.731	0.000 ***	
3 4 ($\beta_{3,0}$)	4.656	0.086	54.192	0.000 ***	
Antal Obs.					39,090
Residual Deviance					69373.69
AIC					69389.69

Note: *p<0.1 **p<0.05 ***p<0.01

*I Tabel L.1 benyttes punktum som decimal separator.

Odds Ratios og Konfidensintervaller

I nedenstående tabel L.2 benyttes punktum som decimal separator.

Tabel L.2: Odds Ratios og Konfidensintervaller - Model 1 Komprimeret

	Odds ratios	2.5 %	97.5 %
K_{it_0}	1.810	1.787	1.834
MI_{it_0}	1.023	1.011	1.034
MU_{it_0}	1.139	1.125	1.153
FI_{it_0}	1.008	0.998	1.017
FU_{it_0}	1.151	1.136	1.167

OLS model

Resultater for Komprimeret OLS Model 1 med Robuste Standardfejl

Tabel L.3*

Tabel L.3: OLS Model Komprimeret med Robuste Standardfejl - Model 1

	<i>Estimat</i>	<i>Std. Fejl</i>	<i>t-værdi</i>	<i>p-værdi</i>	
(Intercept)	1.538	0.035	44.559	0.000	***
KG_{it_0}	0.221	0.002	92.171	0.000	***
MI_{it_0}	0.009	0.002	3.909	0.000	***
MU_{it_0}	0.036	0.002	15.109	0.000	***
FI_{it_0}	0.003	0.002	1.716	0.086	*
FU_{it_0}	0.034	0.002	14.115	0.000	***
Antal Obs.	39,090				
R²	0.265				
Adj. R²	0.265				
F-statistik	2813 *** (df = 5; 39084)				
Note:	*p<0.1 **p<0.05 ***p<0.01				

*I Tabel L.3 benyttes punktum som decimal separator.

L.2 Model 2

Resultater for Komprimeret Model 2

Tabel L.4*

Tabel L.4: Ordered Logistisk Regression - Model 2 Komprimeret

	<i>Estimat</i>	<i>Std. Fejl</i>	<i>t-værdi</i>	<i>p-værdi</i>	<i>Odds Ratios</i>
KG_{it_0}	0.463	0.010	46.975	0.000 ***	1.590
MI_{it_0}	-0.005	0.008	-0.582	0.561	0.995
MU_{it_0}	0.102	0.008	13.576	0.000 ***	1.108
FI_{it_0}	-0.001	0.006	-0.200	0.841	0.999
FU_{it_0}	0.123	0.008	15.979	0.000 ***	1.131
4 5 ($\beta_{1,0}$)	1.761	0.122	14.477	0.000 ***	
5 6 ($\beta_{2,0}$)	2.171	0.121	17.888	0.000 ***	
6 7 ($\beta_{3,0}$)	4.429	0.124	35.609	0.000 ***	
Antal Obs.					20,988
Residual Deviance					42758.68
AIC					42774.68
Note:	*p<0.1 **p<0.05 ***p<0.01				

*I Tabel L.4 benyttes punktum som decimal separator.

Odds Ratios og Konfidensintervaller

I nedenstående tabel L.5, benyttes punktum som decimal separator.

Tabel L.5: Odds Ratios og Konfidensintervaller - Model 2 Komprimeret

	Odds ratios	2.5 %	97.5 %
K_{it_0}	1.590	1.559	1.621
MI_{it_0}	0.995	0.980	1.011
MU_{it_0}	1.108	1.091	1.124
FI_{it_0}	0.999	0.986	1.011
FU_{it_0}	1.131	1.114	1.148

OLS model

Resultater for Komprimeret OLS Model 2 med Robuste Standardfejl

Tabel L.6*

Tabel L.6: OLS Model Komprimeret med Robuste Standardfejl - Model 2

	<i>Estimat</i>	<i>Std. Fejl</i>	<i>t-værdi</i>	<i>p-værdi</i>	
(Intercept)	1.499	0.055	27.412	0.000	***
KG_{it_0}	0.185	0.004	44.404	0.000	***
MI_{it_0}	0.000	0.004	0.137	0.891	
MU_{it_0}	0.044	0.003	13.096	0.000	***
FI_{it_0}	-0.002	0.003	-0.5656	0.572	
FU_{it_0}	0.045	0.003	13.675	0.000	***
Antal Obs.	20,988				
R²	0.137				
Adj. R²	0.137				
F-statistik	668.400 *** (df = 5; 20982)				
Note:	*p<0.1 **p<0.05 ***p<0.01				

*I Tabel L.6 benyttes punktum som decimal separator.