Statistisk analyse af big data

Elena K. Gerassimovitch

Aalborg Universitet

AALBORG UNIVERSITET
STUDENTERRAPPORT

School of Engineering and Science
Skjernvej 4a
9220 Aalborg Øst
www.ses.aau.dk

Titel:

Statistisk analyse af big data

Projektperiode:
Speciale, Matematik
Vintersemesteret 2018

Deltagere:
Elena K. Gerassimovitch

Vejleder:
Christophe A.N. Biscio

Oplagsantal: 3

Sidetal: 69

Afsluttet: 4. januar 2019

Synopsis:
In this project we have worked with different approaches for data analyses in order to apply them to "big data". We have looked at regression analysis and hypothesis test, applying the theory on a big data set "*Gene*" that consist of 6658 measurements from 15 healthy subjects and 15 subjects with a skin disease. In the project we have defined the restrictions of the approaches used and have looked at the possible solutions. We have found out that both logistic lasso and elastic net regressions, with a higher weight on the lasso, give more sparse models with less variables, while the ridge regression works less optimally.
Hypothesis test was only useful on a big data set when corrected for p-values. We have used different p-value adjusting algorithms and found out, that the procedures which controls false discovery rate, is more effective, than the procedures which controls family-wise error rate. In our case, the Benjamini Hochberg algorithm results in expected improvement, while the Holm procedure was less useful.
We have identified features in the data set "Gene", which can be useful to consider to improve diagnostic methods. Both lasso regression and the Benjamini Hochberg algorithm reduces the model from 6658 features to respectively 4 and 5, using specific control parameters. Moreover, both methods allow us to expand the number of features, if needed. In this project we have not been able to validate results obtained from a medical point of view and further study is needed in order to do it.

Indhold

Indledning **1**

1 Generelle- og generaliserede lineære modeller **3**
1.1 Generelle lineære modeller . . . 4
1.1.1 Estimering af middelværdi . . . 4
1.2 Eksempler på generelle lineære modeller . . . 7
1.3 Generaliserede lineære modeller . . . 12
1.3.1 Eksponentielle familier . . . 12
1.3.2 Den generaliserede lineære model og dens egenskaber . . . 16
1.3.3 Logistisk regression . . . 17

2 Regressionsanalyse **18**
2.1 Mindste kvadraters løsning . . . 18
2.2 Regulering . . . 20
2.2.1 K- foldet krydsvalidering . . . 22
2.2.2 De signifikante variabler . . . 24
2.2.3 Fordele og ulemper ved lasso og ridge regression . . . 25
2.3 Elastisk net . . . 26

3 Inferens for store tal **29**
3.1 T-test . . . 29
3.2 Sande og falske opdagelser . . . 31
3.3 Den bayesianske tilnærmelse . . . 32
3.4 P-værdier . . . 34
3.5 FWER justering og FDR kontrol . . . 34
3.5.1 Bonferroni korrektion . . . 34
3.5.2 Holms Procedure . . . 36
3.5.3 Benjamini Hochbergs algoritme . . . 37
3.5.4 Bayes fortolkning . . . 39

Konklusion **40**

Litteraturliste **42**

Appendiks **44**

A **45**
A.1 Lagrange multiplier . . . 45
A.2 Konveksitet . . . 46

B **49**
B.1 Datasættet . . . 49
B.2 Lineær regression . . . 50
B.3 ANOVA . . . 53
B.4 Lasso og ridge regression . . . 55
B.5 Elastisk net regression . . . 59

B.6 T-test 61
B.7 Holms procedure og Benjamini Hochbergs algoritme 64
B.8 Diverse mindre koder, som blev brugt under dataanalyse 66

Indledning

Forebyggelse og præcise diagnoser spiller en vigtig rolle i patientbehandling. Ved patientundersøgelser giver moderne teknologier muligheden for at registrere tusindvis af parametre, og derved eksploderer mængden af de samlede data drastisk. De store datasæt skal kunne analyseres, og der er flere variabler, der skal sættes i sammenhæng med hinanden. Når man analyserer store datasæt, kan man ikke altid være sikker på, at man har fundet sammenhængen mellem "årsag" og "konsekvens". Det typiske problem er, at der ikke er nogen garanti for, at man tager alle faktorer, som har betydning, i betragtning samtidig med, at man risikerer at forkaste signifikante variabler.
Målet med dette projekt er at belyse de forskellige problemstillinger, som opstår, når man arbejder med store datasæt og at komme ind på nogle af de metoder, der bruges inden for statistisk analyse for de store datasæt.
I dette projekt arbejder vi med regressionsanalyse og hypotesetest og undersøger, hvordan metoderne kan bruges til analyse af større datasæt. Vi tager udgangspunkt i de generelle lineære modeller (GLM) og kommer ind på det grundlæggende bag analyse af varians (ANOVA) og lineær regression. Vi viser, at lineær regression ikke kan bruges på de store datasæt uden modifikationer og beskæftiger os med Least absolute shrinkage and selection operator (lasso) regression, ridge regression og elastisk net regression for at afgøre, hvilke af disse metoder med fordel kan anvendes på de store datasæt.
I projektet vil vi derudover arbejde med t-test for at udpege og overkomme de udfordringer, som opstår, når man anvender hypotesetest på de store datasæt.
Vi illustrerer teorien med dataanalyse af datasættet "*Gene*", som indeholder reel genetisk materiale for patienter med en hudsygdom og en kontrolgruppe. Målingerne er udtryksniveauet for 6658 forskellige gener hos 30 personer, og vi er interesserede i at identificere de gener, der er relevante for fastsættelse af diagnosen. I projektet vil vi referere til gen i som Prob$i$. For eksempel er Prob40 stikprøven for gen 40. Dataanalysen er udført i R.
Projektet er opbygget i tre kapitler og Appendiks.

- Kapitel 1 omhandler de generelle og generaliserede lineære modeller (GGLM).
  I dette kapitel kommer vi ind på det grundlæggende indenfor GLM og GGLM og giver forskellige eksempler på, hvordan man anvender GLM og GGLM i praksis. Kapitlet er baseret på [1], [2] og [3].

- Kapitel 2 omhandler regressionsanalyse.
  I dette kapitel kigger vi nærmere på lasso, ridge og elastisk net regression. Kapitlet er baseret på [4], [5] og [6]. I disse kilder gøres der rede for de forskellige regressionsmodeller, og der diskuteres fordele og ulemper ved deres anvendelse og modifikationer.

- Kapitel 3 omhandler hypotesetest.
  I dette kapitel diskuterer vi de udfordringer, som man står med, ved anvendelse af hypotesetest for analyse af de store datasæt. Vi vil i dette kapitel derfor komme ind på de forskellige algoritmer, der giver en justering af p-værdier, når man arbejder med de store

datasæt. Kapitlet er primært baseret på [7], som blandt andet diskuterer family wise error rate (FWER) og falsk opdagelses rate (FDR) samt metoder til at teste multiple hypoteser. I kapitlet bruges der også de sekundære kilder [8] og [9], som begge kommer kort ind på FWER og FDR.

- Appendiks er todelt. Appendiks A er en mindre teoretisk del, bestående af Appendiks A.1 og A.2. I Appendiks A.1 gør vi kort rede for Lagrange multiplier metoden, og i Appendiks A.2 kommer vi ind på konvekse optimeringsproblemer. I Appendiks B.1 til B.8 findes der detaljer af dataanalysen.

Kapitel 1 | Generelle- og generaliserede lineære modeller

I dette kapitel kommer vi til at bruge definitionerne for den flerdimensionale normalfordeling og likelihoodfunktionen.

Definition 1.1. k-flerdimensional normalfordeling
Vektoren $\boldsymbol{Z}$ er k-flerdimensionalt normalfordelt, hvis den har samme fordeling som $\boldsymbol{AY}+\boldsymbol{b}$, hvor $\boldsymbol{Y} = (Y_1, \ldots, Y_N)^T$ er en vektor, og variablerne $Y_1, \ldots, Y_N$ er uafhængig identisk fordelte, $\boldsymbol{A}$ er en $k \times N$ matrix, mens $\boldsymbol{b}$ er en k-dimensional vektor.
Hvis $\boldsymbol{Z}$ er flerdimensional normalfordelt, udtrykkes det ved

$$\boldsymbol{Z} \sim N(\boldsymbol{b}, \boldsymbol{AA}^T).$$

For en normalfordelt vektor $\boldsymbol{Z}$, hvor middelværdien $E[\boldsymbol{Z}] = \boldsymbol{b} = \boldsymbol{\mu}$ og kovariansen $Cov[\boldsymbol{Z}] = \sigma\boldsymbol{\Sigma}$, betegner man $\boldsymbol{Z} \sim N_k(\boldsymbol{\mu}, \sigma^2\boldsymbol{\Sigma})$ med tæthedsfunktionen givet ved

$$f(\boldsymbol{z}) = \frac{1}{(2\pi)^{k/2}\sigma^k\sqrt{|\boldsymbol{\Sigma}|}} \exp\Big\{ -\frac{1}{2\sigma^2}(\boldsymbol{z}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{z}-\boldsymbol{\mu})\Big\}. \tag{1.1}$$

Her er $\boldsymbol{\Sigma}$ kovariansmatricen, med indgang $\boldsymbol{\Sigma}_{i,j} = Cov(Z_i, Z_j)$.
Vi vil referere til likelihoodfunktionen for at estimere parametrene, givet tæthedsparametre og en tæthedsfunktion.
Likelihooden for at få en værdi $\boldsymbol{\theta}$ af parameterene i en model, er sandsynligheden for at få de forskellige udfald $Y_1 = y_1$, $Y_2 = y_2$, ..., $Y_N = y_N$, udregnet for værdien $\boldsymbol{\theta}$. Likelihoodfunktionen opnås som en funktion af $\boldsymbol{\theta} \in \Theta^n$.

Definition 1.2. Likelihoodfunktionen
Givet tæthedsfunktionen $f_Y(\boldsymbol{y}, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta^n$ for observationerne $\boldsymbol{y} = (y_1, \ldots, y_N)$, er likelihoodfunktionen for $\boldsymbol{\theta}$, defineret ved

$$L(\boldsymbol{\theta}; \boldsymbol{y}) = c(y_1, \ldots, y_N) f_Y(y_1, \ldots, y_N; \boldsymbol{\theta}), \tag{1.2}$$

hvor $c(y_1, \ldots, y_N)$ er en konstant.

I en statistisk model giver likelihoodfunktionen et mål for præference af forskellige parameterværdier.
Det, man typisk er interesseret i, er den maksimale værdi af likelihoodfunktionen. Givet observationerne $\boldsymbol{y} = (y_1, \ldots, y_N)$ er maksimum likelihood estimatet (MLE) en funktion $\hat{\boldsymbol{\theta}}(\boldsymbol{y})$, som opfylder

$$L(\hat{\boldsymbol{\theta}}; \boldsymbol{y}) = \sup_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta}; \boldsymbol{y}).$$

Ofte, især hvis man har de eksponentielle led i tæthedsfunktionen, tager man udgangspunkt i log-likelihoodfunktion

$$\ell(\boldsymbol{\theta}; \boldsymbol{y}) = Log(L(\boldsymbol{\theta}; \boldsymbol{y})). \tag{1.3}$$

Maksimum likelihood estimatet fås som løsning af ligning

$$S(\boldsymbol{\theta}; \boldsymbol{y}) = \mathbf{0}, \tag{1.4}$$

hvor $S(\boldsymbol{\theta}; \boldsymbol{y}) = \frac{\delta}{\delta \boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \boldsymbol{y})$ er scorefunktionen.

1.1 Generelle lineære modeller

Statistiske modeller bruges til at forklare variationer i et observationssæt, og man kan, for eksempel, beskrive det forventede udfald Y_i ved brug af middelværdi μ_i og fejlledet ϵ_i, der dækker over den statistiske usikkerhed. Man antager, at middelværdien har en lineær struktur, mens fejleddet er normalfordelt. Hvis man har N observationer $\boldsymbol{Y} = (Y_1, \ldots, Y_N)$, kan responsvariablen beskrives på matrixform

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}. \tag{1.5}$$

Definition 1.3. Generelle lineære modeller
Vi betragter N observationer $Y_1, \ldots, Y_N$ givet ved $\boldsymbol{Y} = (Y_1, \ldots, Y_N)^T$, som er normalfordelt. En general lineær model for $Y_1, \ldots, Y_N$ er en statistisk model på formen $\boldsymbol{Y} \sim N_N(\boldsymbol{\mu}, \sigma^2 \boldsymbol{\Sigma})$, hvor der gælder følgende hypotese for $\boldsymbol{\mu}$

$$\mathcal{H}_0 : \boldsymbol{\mu} - \boldsymbol{\mu}_0 \in U_0. \tag{1.6}$$

Her er U_0 et lineært underrum af $\mathbb{R}^n$ med dimension n, og $\boldsymbol{\mu}_0$ er en vektor bestående af a-priori middelværdier, som ikke hører til mængden U_0.
Vi lader det lineære underrum $U_0 = \text{span}\{x_1, \ldots, x_n\}$ og betragter GLM, hvor der gælder

$$\mathcal{H}_0 : \boldsymbol{\mu} - \boldsymbol{\mu_0} = \boldsymbol{X}\boldsymbol{\beta}, \tag{1.7}$$

hvor $\boldsymbol{\beta} \in \mathbb{R}^n$, og $\boldsymbol{X}$ har fuld rang.
Matricen $\boldsymbol{X}$ defineres som designmatricen, da den bestemmer, hvorledes de ukendte parametre i $\boldsymbol{\beta}$ er forbundet med responsværdierne i $\boldsymbol{Y}$. Den i'te række af designmatricen er givet ved modelvektoren $\boldsymbol{x_i} = [x_{i1}, \ldots, x_{in}]$ for den i'te observation.

1.1.1 Estimering af middelværdi

I dette afsnit tager vi udgangspunkt i en model $\boldsymbol{Y} \sim N_N(\boldsymbol{\mu}, \sigma^2 \boldsymbol{\Sigma})$, hvor vi ønsker at estimere middelværdien $\boldsymbol{\mu}$ ved hjælp af likelihoodfunktionen

Lad $\boldsymbol{y}_1$ og $\boldsymbol{y}_2$, være to vektorer på $\mathbb{R}^N$. Det indreprodukt er defineret ved:

$$\delta_\Sigma(\boldsymbol{y_1}, \boldsymbol{y_2}) = \boldsymbol{y}_1^T \boldsymbol{\Sigma}^{-1} \boldsymbol{y}_2,$$

hvor $\boldsymbol{y}_1$ og $\boldsymbol{y}_2$ er ortogonale, når $\delta_\Sigma(\boldsymbol{y_1}, \boldsymbol{y_2}) = 0$. Normen er givet ved:

$$||\boldsymbol{y}||_\Sigma = \sqrt{\delta_\Sigma(\boldsymbol{y}, \boldsymbol{y})}.$$

Deviansen $D(\boldsymbol{y}; \boldsymbol{\mu})$ betegner den kvadratiske norm af vektoren $\boldsymbol{y} - \boldsymbol{\mu}$, som svarer til det indreprodukt $\delta_\Sigma(\boldsymbol{y} - \boldsymbol{\mu}, \boldsymbol{y} - \boldsymbol{\mu})$:

$$D(\boldsymbol{y}; \boldsymbol{\mu}) = \delta_\Sigma(\boldsymbol{y} - \boldsymbol{\mu}, \boldsymbol{y} - \boldsymbol{\mu}) = \boldsymbol{(y - \mu)^T \Sigma^{-1} (y - \mu).} \tag{1.8}$$

Ved hjælp af deviansen kan man udtrykke tæthedsfunktionen for normalfordelingen (1.1) som:

$$f(\boldsymbol{y}; \boldsymbol{\mu}, \sigma^2) = \frac{1}{(\sqrt{2\pi})^N \sigma^N \sqrt{|\boldsymbol{\Sigma}|}} \exp\Big\{ -\frac{1}{2\sigma^2} D(\boldsymbol{y}; \boldsymbol{\mu}) \Big\}.$$

For at finde den maksimale likelihood estimat for middelværdien $\boldsymbol{\mu}$, kan vi bruge (1.2), hvor konstanten $c(y_1, \ldots, y_N)$ sættes til 1:

$$L(\boldsymbol{\mu}, \sigma^2; \boldsymbol{y}) = \frac{1}{(\sqrt{2\pi})^N \sigma^N \sqrt{|\boldsymbol{\Sigma}|}} \exp\Big\{ -\frac{1}{2\sigma^2} D(\boldsymbol{y}; \boldsymbol{\mu}) \Big\}. \tag{1.9}$$

Vi udlader konstanterne $(\sqrt{2\pi})^N$ og $\sqrt{|\boldsymbol{\Sigma}|}$, da likelihoodfunktionen ikke udtrykker sandsynlighedstætheden for parametrene, men i stedet for er et mål for, hvad parametrene er.
I vores tilfælde giver det god mening at bruge log-likelihoodfunktionen:

$$\ell(\boldsymbol{\mu}, \sigma^2; \boldsymbol{y}) = -\frac{N}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} D(\boldsymbol{y}; \boldsymbol{\mu}).$$

Heraf kan vi se, at jo mindre deviansen er, jo større er log-likelihoodfunktionen.
Ud fra (1.4) har vi maksimum af log-likelihoodfunktionen som løsning af ligning:

$$S(\boldsymbol{\mu}, \sigma^2; \boldsymbol{y}) = -\frac{1}{2\sigma^2} \cdot \frac{d}{d\boldsymbol{\mu}} (D(\boldsymbol{y}; \boldsymbol{\mu})) = \frac{1}{\sigma^2} [\boldsymbol{\Sigma}^{-1} \boldsymbol{y} - \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}] = \frac{1}{\sigma^2} \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}) = \boldsymbol{0}. \tag{1.10}$$

Siden vektorrummet er udstyret med det indreprodukt δ_Σ, finder man estimatet af $\boldsymbol{\mu}$ ved projektion ind i spændingsrummet $\boldsymbol{\mu}$. Fra ovenstående scorefunktion (1.10) kan man se, at maksimum likelihood for $\boldsymbol{\mu}$ for modellen $\boldsymbol{Y} \sim N_N(\boldsymbol{\mu}, \sigma^2 \boldsymbol{\Sigma})$, er observationen givet ved $\boldsymbol{y}$.

Sætning 1.4.
For en hypotese givet ved

$$\mathcal{H}_0 : \boldsymbol{\mu}(\boldsymbol{\beta}) = \boldsymbol{X\beta} \tag{1.11}$$

finder man maksimum likelihood estimatet for $\boldsymbol{\beta}$ som løsning af ligningen

$$\boldsymbol{X}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{y} = \boldsymbol{X}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{X} \hat{\boldsymbol{\beta}}. \tag{1.12}$$

Hvis $\boldsymbol{X}$ har fuld rang vil estimatet $\boldsymbol{\beta}$ være givet ved

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y}. \tag{1.13}$$

Bevis

Af udtryk (1.9) ses der, at maksimum med hensyn til $\boldsymbol{\mu}$ fås ved at minimere deviansen $D(\boldsymbol{y};\boldsymbol{\mu})$, for $\boldsymbol{\mu} \in U_0$. Idet $D(\boldsymbol{y};\boldsymbol{\mu})$ er en norm med indre produktet, δ_Σ, opnås minimum ved at projicere ned på U_0.
Antaget, at (1.11) er gældende, har vi :

$$\frac{\delta}{\delta\boldsymbol{\beta}}\boldsymbol{\mu}(\boldsymbol{\beta}) = \frac{\delta}{\delta\boldsymbol{\beta}}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}.$$

Ved at udnytte parametriseringen af $\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\mu}(\boldsymbol{\beta})$ vil scorefunktionen $S(\boldsymbol{\beta}, \sigma^2; \boldsymbol{y})$ være

$$S(\boldsymbol{\beta}, \sigma^2; \boldsymbol{y}) = \frac{\delta}{\delta\boldsymbol{\beta}}\ell(\boldsymbol{\beta}, \sigma^2; \boldsymbol{y}) = \boldsymbol{J}^T\frac{\delta}{\delta\boldsymbol{\mu}}\ell(\boldsymbol{\mu}(\boldsymbol{\beta}), \sigma^2; \boldsymbol{y}),$$

hvor $\boldsymbol{J} = \frac{\delta\boldsymbol{\mu}}{\delta\boldsymbol{\beta}}$ er Jacobi matricen.
Fra (1.10) har vi scorefunktionen givet med hensyn til $\boldsymbol{\mu}$. Bruger vi denne, med $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$, fås

$$S(\boldsymbol{\beta}, \sigma^2; \boldsymbol{y}) = \frac{1}{\sigma^2}\boldsymbol{X}^T[\boldsymbol{\Sigma}^{-1}\boldsymbol{y} - \boldsymbol{\Sigma}^{-1}\boldsymbol{X}\boldsymbol{\beta}].$$

For at finde maksimum likelihood estimatet $\boldsymbol{\beta}$ løser vi ligningen

$$S(\hat{\boldsymbol{\beta}}, \sigma^2; \boldsymbol{y}) = \boldsymbol{0}.$$

Heraf vil

$$\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y} = \boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\hat{\boldsymbol{\beta}},$$

hvilket stemmer overens med, at maksimum likelihood for $\boldsymbol{\beta}$ findes som løsning af (1.12).
Hvis $\boldsymbol{X}$ har fuld rang, er $\boldsymbol{X}$ også invertibel, hvilket medfører (1.13). □

1.2 Eksempler på generelle lineære modeller

Som typiske eksempler på GLM kigger vi på lineær regression og ANOVA og illustrerer den praktiske anvendelse af disse metoder på vores datasæt.

Eksempel 1.5. Regressionsanalyse som GLM
For en multipel lineær regression antager man, at der er n valgte forklarende variabler $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ og udfaldet $Y_1, \ldots, Y_N$ opfylder

$$y_i = \sum_{j=0}^{n} x_{ij}\beta_j + \epsilon_i \quad \text{for} \quad i = 1, \ldots, N$$

med

$$\boldsymbol{x}_i = [1, x_{i1}, \ldots, x_{in}]^T \quad \text{og} \quad \boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_n]^T.$$

For en simpel lineær regression antages der, at $n = 1$ og der gælder for den i'te række, at

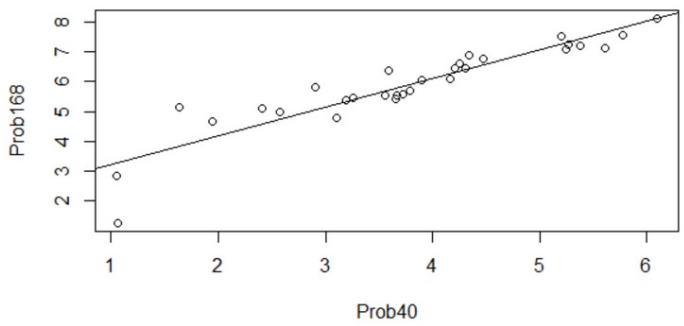
$$\boldsymbol{x}_i = [1, x_{i1}]^T \quad \text{og} \quad \boldsymbol{\beta} = [\beta_0, \beta_1]^T.$$

I forhold til vores datasæt kan vi for eksempel finde modellen, som binder nogle korrelerede gener sammen.

Dataanalyse 1.6. Lineær regressionsmodel for de korrelerede gener
Lad os forholde os til to gener, Prob40 og Prob168. Vi lader værdier for det ene gen, Prob40, være den uafhængige variabel $\boldsymbol{X}$ og værdierne for Prob168 være responsvariablen $\boldsymbol{Y}$. Vi kan skrive sammenhængen mellem de to gener:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta},$$

hvor $\boldsymbol{\beta} = [2.27 \;\; 0.9581]^T$ (se Appendiks B.1 og B.2). Figur 1.1 viser sammenhængen mellem værdierne for de valgte gener, samt regressionslinjen.

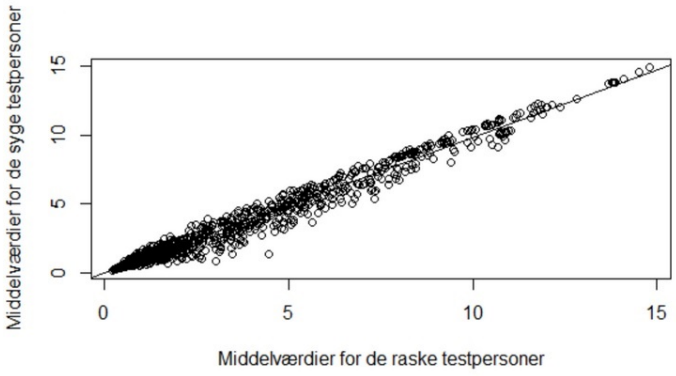


Figur 1.1: XY-plot for udtryksniveauer for generne Prob168 og Prob40 samt regressionslinje.

Dataanalyse 1.7. Lineær regressionsmodel for middelværdier
Vi fremstiller en model, som binder middelværdierne for de syge og de raske testpersoner sammen. I dette tilfælde ville $Y_i = \mu_i(2)$, middelværdier for de forskellige gener for de syge testpersoner, og $X_i = \mu_i(1)$, middelværdier for de forskellige gener for de raske testpersoner. Den lineære regressionsmodel vil være

$$\boldsymbol{\mu}(2) = \boldsymbol{\mu}(1)\boldsymbol{\beta}.$$

Resultatet af dataanalysen giver os $\boldsymbol{\beta} = [0.02 \;\; 0.976]^T$ (se Appendiks B.2). Interceptleddet $\beta_0 = 0.02$ er tæt på 0, og hældningen $\beta_1 = 0.976$ er tæt på 1, hvilket egentlig betyder, at, hvis vi kigger på de middelværdier for alle genernes udtryksniveau samtidigt, er det svært at se nogle forskelle mellem de syge og de raske testpersoner.
Figur 1.2 illustrerer sammenhængen mellem middelværdierne for generne for de raske og de syge testpersoner. Selvom man fra denne model ikke er i stand til at komme med nogen konklusion om diagnose, er modellen stadigvæk brugbar. Man kan, for eksempel, identificere outliers, de gener, man eventuelt skal tage i betragtning ved fastsættelsen af diagnose.



Figur 1.2: XY-plot af middelværdierne for genernes udtryksniveauer for testpersoner med og uden sygdom, samt regressionslinje.

Et andet eksempel på GLM, vi kigger på, er ANOVA.
Ved hjælp af ANOVA afgører man, om middelværdierne for 2 eller flere stikprøver er signifikant forskellige. Det vil sige, at man tester nulhypotesen $\mathcal{H}_0$, om middelværdierne for de forskellige stikprøver er ens.

Eksempel 1.8. Oneway ANOVA som GLM
Lad $X_{i1}, \ldots, X_{iN}$ være n uafhængige stikprøver, hvor der gælder at $X_{ij} \sim N(\mu_i, \sigma^2)$ med $i = 1, \ldots, n$ og $j = 1, \ldots, N$. For ANOVA gælder følgende hypoteser

$$\mathcal{H}_0 : \mu_1 = \mu_2 = \ldots = \mu_n \tag{1.14}$$

og

$$\mathcal{H}_1 : \mu_{i_1} \neq \mu_{i_2} \tag{1.15}$$

for nogen i_1 og i_2.
At X_{ij} er normalfordelt, $X_{ij} \sim N(\mu_i, \sigma^2)$, kan også udtrykkes som

$$x_{ij} = \mu_i + \epsilon_{ij}, \tag{1.16}$$

hvor $i = 1, \ldots, n$, og $j = 1, \ldots, N$.
Vi ser, at (1.16) allerede minder om formen (1.5).
Vi sætter først

$$\boldsymbol{Y} = [x_{11}, \ldots, x_{1N}, \ldots, x_{i1}, \ldots, x_{iN}, \ldots, x_{n1}, \ldots, x_{nN}]^T$$

og

$$\boldsymbol{\epsilon} = [\epsilon_{11}, \ldots, \epsilon_{1N}, \ldots, \epsilon_{i1}, \ldots, \epsilon_{iN}, \ldots, \epsilon_{n1}, \ldots, \epsilon_{nN}]^T.$$

Vi reparametriserer gruppen af middelværdier, μ_i, til

$$\mu_i = \mu + \alpha_i.$$

Således $\beta_0 = \mu$. Vi får dog samtidig en ekstra parameter α_i, som typisk opfylder $\sum_{i=1}^{n} \alpha_i = 0$. Designmatricen for GLM'en er givet ved $\boldsymbol{x_i} = [1, x_{i1}, \ldots, x_{in}]$ samt $\boldsymbol{\beta} = [\mu, \alpha_1, \ldots, \alpha_n]$, hvor x_{ij} er en dummy variabel som opfylder $x_{ij} = 1$, hvis den i'te observation tilhører j, og $x_{ij} = 0$, hvis ikke. Hvis $n = N$, vil $\boldsymbol{X}$ være en kvadratisk design matrix, og estimatet af $\boldsymbol{\beta}$ findes som løsning af

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}.$$

For at afgøre om hypotesen om ens middelværdier kan accepteres, opnår man n uafhængige estimater af variansen

$$s_i^2 = \frac{1}{N-1} \sum_{j=1}^{N} (X_{ij} - \bar{X}_i')^2, \quad \text{for} \quad i = 1, \ldots, n, \tag{1.17}$$

hvor $\bar{X}_i'$ betegner det empiriske gennemsnit for stikprøverne X_{ij}:

$$\bar{X}_i' = \frac{1}{N} \sum_{j=1}^{N} X_{ij}.$$

Idet alle stikprøverne forholder sig til testtørrelsen N, beregner man gennemsnit for estimatet s_i^2 i (1.17), det vil sige variansen indenfor de forskellige grupper af uafhængige variabler[1] ved:

$$s_{SSE} = \frac{1}{N} \sum_{i=1}^{n} s_i^2 = \frac{1}{n(N-1)} \sum_{i=1}^{n} \sum_{j=1}^{N} (X_{ij} - \bar{X}_i')^2.$$

Hvis $\mathcal{H}_0$ er sand, antages der, at middelværdierne for stikprøverne er normalfordelt, $\bar{X}_i' \sim N(\mu_i, \sigma^2/N)$. På denne måde får vi et andet estimat[2] for variansen σ^2, variansen mellem grupper af uafhængige variabler:

$$s_{SST} = \frac{N}{n-1} \sum_{i=1}^{n} (\bar{X}_i' - \bar{X})^2.$$

Her er $\bar{X}$ er middelværdien af $\bar{X}_i'$. Hvor s_{SSE} altid er et unbias estimat for σ^2, er s_{SST} kun unbias, hvis $\mathcal{H}_0$ er sand. Hvis det omvendte gælder, og den alternative hypotese $\mathcal{H}_1$ er sand, kan man vise, at $E[s_{SST}] > \sigma^2$, hvilket betyder, at, hvis s_{SST} er signifikant større end s_{SSE}, forkaster man $\mathcal{H}_0$.
Som de praktiske eksempler på multivariabel ANOVA kigger vi på, om der er forskel i genernes middelværdier for de forskellige gener, og mellem de syge og raske testpersoner for vores datasæt.

[1] SSE står for sum of squares error og er givet ved $SSE = \sum_{i=1}^{n} \sum_{j=1}^{N} (X_{ij} - \bar{X}_i')^2$

[2] SST står for sum of squares of treatments, og er givet ved $SST = \sum_{i=1}^{n} \sum_{j=1}^{N} (\bar{X}_i' - \bar{X})^2$

Dataanalyse 1.9. Multivariabel ANOVA
Vi har valgt 4 tilfældige gener i vores datasæt og testede ved hjælp af ANOVA, om der er forskel i deres middelværdier.
Nulhypotesen er dermed

$$\mathcal{H}_0 : \ \mu_{257} = \mu_{634} = \mu_{1949} = \mu_{2391},$$

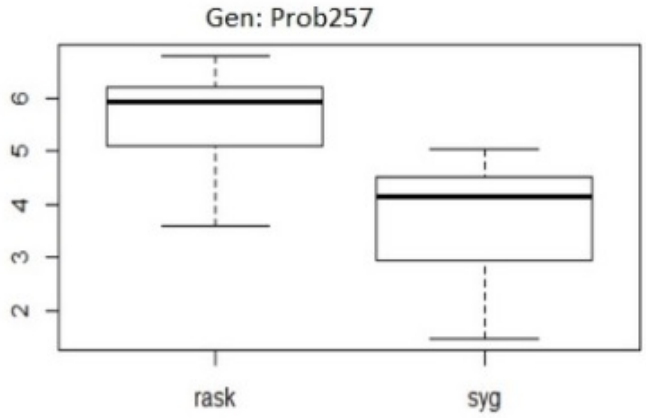
og vi tester den med en signifikansniveau på $\alpha = 0.05$.
Denne hypotese forkastes (se Appendiks B.3), hvilket ikke er overraskende.

Ulempen med one-way ANOVA er, at den kun kan afgøre, at mindst to stikprøver er signifikant forskellige, men ikke kan fortælle, hvilke stikprøver det er.
I forhold til dataanalysen ville det være mere interessant for vores projekt at identificere de gener og grupper af gener, som er forskellige fra hinanden.

Dataanalyse 1.10. Sammenligning af gener for de syge og raske testpersoner
Vi anvender ANOVA på vores datasæt for at sammenligne generne hver for sig for de raske og de syge testpersoner. Vi tester nulhypoteser for n stikprøver

$$\mathcal{H}_{0i} : \mu_i(1) = \mu_i(2),$$

for $i = 0, \ldots, n$, hvor $\mu_i(1)$ og $\mu_i(2)$ henholdsvis er middelværdien for det i'te gen for de raske og de syge testpersoner.
For datasættet "$MindreData$" resulterede analysen i, at 4 nulhypoteser ud af 8 testede blev forkastet med en signifikansniveau på $\alpha = 0.05$ (se Appendiks B.3). Dette betyder, at der for de gener er tale om signifikante forskelle i middelværdier for de syge og de raske testpersoner. Figur 1.3 viser et boksplot for et af disse gener.



Figur 1.3: Boksplot for de syge og raske testpersoner, for gen Prob257.

For datasættet "*Gene*" resulterede dataanalysen i 890 forkastede nulhypoteser. Dette giver os 890 gener, som eventuelt har betydning for, om en person hører til de raske eller til de syge.

I Kapitel 3 "Inferens for store tal " kommer vi ind på de problemer, man møder ved anvendelsen af hypotesetest på de store datasæt.

1.3 Generaliserede lineære modeller

Generaliserede modeller bruges, når målingernes fordeling ikke har en lineær middelværdistruktur. Hvor de generelle lineære modeller anvendes til at beskrive kontinuert data, bruges generaliserede lineære modeller til at beskrive diskrete data, som, for eksempel, kan have en poisson eller en binomial fordeling. I vores projekt arbejdes der med data, hvor responsvariablen antager værdien 1 eller 0, det vil sige binomial fordelte data.
For at definere den generaliserede lineære model starter vi med at kigge på den eksponentielle dispersionsfamilie, den eksponentielle dispersionsmodel, variansfunktionen og linkfunktionen.

1.3.1 Eksponentielle familier

Den naturlige eksponentielle familie består af fordelinger, hvis tæthedsfunktion kan skrives på formen

$$f_Y(y;\theta) = c(y)\exp\Big\{\theta y - \kappa(\theta)\Big\}. \tag{1.18}$$

Funktionen $\kappa(\theta)$ kaldes kumulantfrembringeren, $\theta \in U$ kaldes den kanoniske parameter, og parameterrummet $U \subset \mathbb{R}$.
Man kan faktorisere tæthedsfunktionen (1.18) som

$$f_Y(y;\theta) = c(y)b(\theta)\exp\Big\{\theta y\Big\},$$

hvor den ene faktor kun afhænger af y, den anden kun afhænger af θ, mens den sidste faktor $\exp\{\theta y\}$ både afhænger af y og θ. Ved at indføre en ekstra præsisionsparameter $\lambda > 0$ kan man fra (1.18) definere den eksponentielle dispersionsfamilie som

$$f_Y(y;\theta,\lambda) = c(y,\lambda)\exp\Big\{\lambda(\theta y - \kappa(\theta))\Big\}. \tag{1.19}$$

Formålet med dispersionsfamilien er at adskille de middelværdi relaterede egenskaber, som er karakteriseret gennem kumulantfrembringeren, fra dispersionsegenskaber. Man kan vise, at både normal og binomial fordelingen hører til den eksponentielle dispersionsfamilie

Eksempel 1.11. Den normale fordeling som eksponentiel dispersionsfamilie

Lad $Y \sim N(\mu, \sigma^2)$. Tæthedsfunktionen for Y:

$$f_Y(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big\{ -\frac{(y-\mu)^2}{2\sigma^2} \Big\} = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{\frac{-y^2}{2\sigma^2}\} \exp\{\frac{1}{\sigma^2}(y\mu - \frac{\mu^2}{2}))\}.$$

Sætter man $\lambda = 1/\sigma^2$, $\mu = \theta$, $c(y,\lambda) = \sqrt{\frac{\lambda}{2\pi}} \exp\{\frac{-\lambda y^2}{2}\}$ og $\kappa(\theta) = \theta^2/2$, opfylder $f_Y(y;\mu,\sigma^2)$ (1.19), og dermed tilhører den normale fordeling de eksponentielle dispersionsfamilier.

Eksempel 1.12. Den binomiale fordeling som eksponentiel dispersionsfamilie
Lad $X \sim Bin(n,p)$. Tæthedsfunktionen for X er givet ved

$$f_X(x;n,p) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{n}{x} (1-p)^n (\frac{p}{1-p})^x \qquad \text{for } x = 0,\ldots,n.$$

Af dette kan vi udlede tæthedsfunktionen for fordelingen $Y = X/n$

$$f_Y(y;n,p) = \binom{n}{ny} (1-p)^n (\frac{p}{1-p})^{ny} \qquad \text{for } y = 0, 1/n, \ldots, 1 \text{ og } 0 \le p \le 1,$$

hvilket kan omskrives til

$$f_Y(y;n,\theta) = \binom{n}{ny} \exp\{n(\theta y - \log(1+\exp(\theta)))\},$$

hvor $\theta = \log(\frac{p}{1-p})$, for $0 < p < 1$.
Dette opfylder (1.19) for $\lambda = n$, $c(y,\lambda) = \binom{\lambda}{\lambda y}$ og $\kappa(\theta) = \log(1+\exp(\theta))$.

Særligt gælder der, at middelværdi relaterede egenskaber for den eksponentielle dispersionsfamilie er bestemt ved kumulantfrembringeren $\kappa(\theta)$. For både den normale og binomial fordelingen er

$$E[Y] = \mu = \kappa'(\theta) \qquad \text{og} \qquad Var[Y] = \sigma^2 = \frac{\kappa''(\theta)}{\lambda}.$$

Lemma 1.13.
Hvis Y har en fordeling, givet ved (1.19), gælder der følgende for middelværdien og variansen

$$E[Y] = \kappa'(\theta) \qquad \text{og} \qquad Var[Y] = \frac{\kappa''(\theta)}{\lambda}. \tag{1.20}$$

Bevis
Lad Y have en fordeling på formen (1.19), omskrevet til

$$f_Y(y;\lambda,\theta) = \exp\Big\{ c(y,\lambda) + \lambda(\theta y - \kappa(\theta)) \Big\}.$$

Idet $f_Y(y;\lambda,\theta)$ er en tæthedsfunktion, gælder der, at

$$\int f_Y(y;\lambda,\theta)dy = \int \exp\Big\{c(y,\lambda) + \lambda(\theta y - \kappa(\theta))\Big\}dy = 1. \tag{1.21}$$

Differentierer man på begge sider med hensyn til θ, fås

$$\int \frac{d}{d\theta}\Big[\exp\Big\{c(y,\lambda) + \lambda(\theta y - \kappa(\theta))\Big\}\Big]dy = 0. \tag{1.22}$$

Ligningen (1.22) kan yderligere omskrives til:

$$\int \frac{d}{d\theta}\Big[\exp\Big\{c(y,\lambda) + \lambda(\theta y - \kappa(\theta))\Big\}\Big]dy =$$

$$= \int \exp\Big\{c(y,\lambda) + \lambda(\theta y - \kappa(\theta))\Big\} \cdot \frac{d}{d\theta}\Big[\lambda(\theta y - \kappa(\theta))\Big]dy =$$

$$= \int \exp\Big\{c(y,\lambda) + \lambda(\theta y - \kappa(\theta))\Big\} \cdot \lambda(y - \kappa'(\theta))dy =$$

$$= \lambda\Big[\int y\exp\Big\{c(y,\lambda) + \lambda(\theta y - \kappa(\theta))\Big\}dy - \kappa'(\theta)\int \exp\Big\{\lambda(\theta y - \kappa(\theta)) + c(y,\lambda)\Big\}dy\Big] =$$

$$= \lambda\Big[\int y f_Y(y;\lambda,\theta)dy - \kappa'(\theta)\int f_Y(y;\lambda,\theta)dy\Big] = 0. \tag{1.23}$$

Fra definitionen af middelværdien μ har vi

$$E[Y] = \mu = \int y f_Y(y;\lambda,\theta)dy,$$

og ved brug af normalisering (1.21) får vi ud fra (1.23):

$$\mu - \kappa'(\theta) = 0 \Longleftrightarrow \kappa'(\theta) = \mu = E[Y].$$

I det følgende beviser vi, at $Var[Y] = \kappa''(\theta)/\lambda$.
Vi differentierer ligning (1.22) med hensyn til θ en gang til.

$$\frac{d^2}{d\theta^2}\int \exp\Big\{c(y,\lambda) + \lambda(\theta y - \kappa(\theta))\Big\}dy = 0,$$

og bruger tidligere resultat af differentiering (1.22):

$$\frac{d}{d\theta}\int \exp\Big\{c(y,\lambda) + \lambda(\theta y - \kappa(\theta))\Big\} \cdot \lambda(y - \kappa'(\theta))dy =$$

$$= \int \left[\exp\left\{ c(y,\lambda) + \lambda(\theta y - \kappa(\theta)) \right\} \cdot (\lambda(y - \kappa'(\theta)))^2 - \kappa''(\theta)\lambda \exp\left\{ c(y,\lambda)\lambda(\theta y - \kappa(\theta)) \right\} \right] dy =$$

$$= \lambda \Big[\lambda \int (y - \kappa'(\theta))^2 f_Y(y;\lambda,\theta) dy - \kappa''(\theta) \int f_Y(y;\lambda,\theta) dy \Big] =$$

$$= \lambda \Big[\lambda \int (y - \mu)^2 f_Y(y;\lambda,\theta) dy - \kappa''(\theta) \Big] = 0.$$

Hvilket resulterer i

$$\lambda \int (y - \mu)^2 f_Y(y;\lambda,\theta) dy = \kappa''(\theta). \tag{1.24}$$

Per definition er variansen af Y givet ved

$$Var[Y] = \int (y - \mu)^2 f_Y(y;\lambda,\theta) dy,$$

og derved omskrives (1.24) til

$$\lambda Var[Y] = \kappa''(\theta) \Longleftrightarrow Var[Y] = \frac{\kappa''(\theta)}{\lambda}.$$

□

Funktionen $\tau(\theta) = \kappa'(\theta)$ definerer en injektiv afbildning $\mu = \tau(\theta)$ af parameterrummet U for den kanoniske parameter θ ind på en delmængde $\mathcal{M}$ af den reelle akse. Mængden $\mathcal{M}$ kaldes middelværdirummet. Funktionen $\tau(\theta)$ er monoton voksende, da der gælder

$$\tau'(\theta) = \kappa''(\theta) = Var[Y]\lambda > 0.$$

Ved hjælp af $\tau(\theta)$ kan man udtrykke sammenhængen mellem variansen og middelværdien:

$$\tau(\theta) = \mu = \kappa'(\theta).$$

Definition 1.14. Kanonisk linkfunktion
Relationen mellem parametrene μ og θ er givet ved $\mu = \tau(\theta) = \kappa'(\theta)$, og den inverse afbildning

$$\theta = \tau^{-1}(\mu)$$

kaldes den kanoniske linkfunktion.

En linkfunktion er en vilkårlig monoton afbildning, $\eta = g(\mu)$, af middelværdirummet ind på en delmængde af den reelle akse. Formålet med linkfunktionen $\eta = g(\cdot)$ er at danne den forklarende variabel, det vil sige den lineære prædiktor.
Som følge af $\kappa''(\theta) = Var[Y]$ kan man udtrykke fordelingens variansen som funktion af middelværdien:

$$V(\mu) = \kappa''(\tau^{-1}(\mu)). \tag{1.25}$$

Funktionen $V(\mu)$ beskriver således variansen af en givet familie af fordelinger i term af middelværdien.

1.3.2 Den generaliserede lineære model og dens egenskaber

For GGLM gælder der, at tæthedsfunktionen for variablerne $Y_1, \ldots, Y_N$ tilhører den eksponentielle dispersionsfamilie. Omvendt GLM er GGLM derved 'mere' generaliseret. GGLM er defineret ved linkfunktionen $g(\cdot)$, som beskriver relationen mellem den lineære prædiktor η_i og middelværdien μ_i:

$$\eta_i = g(\mu_i),$$

og er en model på formen

$$g(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\eta}. \tag{1.26}$$

Definition 1.15. Den generaliserede lineære model

Lad os have en fordeling for $Y_1, \ldots, Y_N$, hvor $Y_1, \ldots, Y_N$ er indbyrdes uafhængige variabler, som kan beskrives ved en eksponentiel dispersionsmodel med variansfunktionen $V(\mu)$ på formen (1.25).
Den generaliserede lineære model for $Y_1, \ldots, Y_N$ beskriver en affin hypotese for værdierne $\eta_1, \ldots, \eta_N$, på følgende form:

$$\mathcal{H}_0 : \boldsymbol{\eta} - \boldsymbol{\eta}_0 \in L, \tag{1.27}$$

hvor L er et lineært underrum af $\mathbb{R}^n$ af dimension n og $\boldsymbol{\eta}_0$ betegner en vektor af kendte værdier, som ikke hører til mængden L, og hvor

$$\eta_i = g(\mu_i)$$

er en transformation af middelværdierne $\mu_1, \ldots, \mu_N$.
Lader man underrummet $L = \text{span}\{x_1, \ldots, x_n\}$ være valgt således der gælder

$$\boldsymbol{\eta} - \boldsymbol{\eta}_0 = \boldsymbol{X}\boldsymbol{\beta},$$

hvor $\boldsymbol{\beta} \in \mathbb{R}^n$ og $\boldsymbol{X}$ har fuld rang, defineres $\boldsymbol{X}$ som designmatricen for GGLM.

Modeller, der har en normalfordeling, er både GLM og GGLM. Dette følger af, at normalfordelingen tilhører den eksponentielle dispersionsfamilie (se Eksempel 1.11). Som et andet eksempel på GGLM kigger vi på logistisk regression.

1.3.3 Logistisk regression

Logistisk regression bruges til analyse af de kategoriske datasæt blandt andet, hvis responsvariabel kun antager værdien 0 eller 1. I dette tilfælde modelleres log-likelihood brøken som den lineære kombination:

$$\log\left(\frac{Pr[Y=1|\boldsymbol{X}=\boldsymbol{x}]}{Pr[Y=0|\boldsymbol{X}=\boldsymbol{x}]}\right) = \log\left(\frac{Pr[Y=1|\boldsymbol{X}=\boldsymbol{x}]}{1-Pr[Y=1|\boldsymbol{X}=\boldsymbol{x}]}\right) = \beta_0 + \boldsymbol{\beta}^T\boldsymbol{x}, \tag{1.28}$$

hvor $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ er en vektor bestående af n prædiktore, $\beta_0 \in \mathbb{R}$, er interceptttermen, mens $\boldsymbol{\beta} \in \mathbb{R}^n$ er en vektor af regressionskoefficienter.
Ud fra (1.28) kan man få den betingede sandsynlighed:

$$Pr[Y=1|\boldsymbol{X}=\boldsymbol{x}] = \frac{\exp\{\beta_0 + \boldsymbol{\beta}^T\boldsymbol{x}\}}{1+\exp\{\beta_0 + \boldsymbol{\beta}^T\boldsymbol{x}\}}. \tag{1.29}$$

Hvis sandsynligheden for at $Y = 1$, givet en stokastisk vektor $\boldsymbol{X} = \boldsymbol{x}$, er større end 0.5, vil responsvariablen være 1, mens, hvis det omvendte er tilfældet, vil $Y = 0$.

Eksempel 1.16. Logistisk regression som GGLM
Logistisk regression håndterer blandt andet tilfældet, hvor responsvariablen er binær, mens den betingede fordeling er binomial. Således er (1.28) en linkfunktion og har en binomial tæthedsfunktion.
I eksempel 1.12, viser vi, at den binomiale fordeling tilhører den eksponentielle dispersionsfamilie med $\theta = \log(\frac{p}{1-p})$. Dette betyder, at (1.28) er en GGLM med linkfunktionen

$$\eta = g(\mu) = \log(\frac{\mu}{1-\mu}).$$

I det følgende kapitel går vi i detaljer med regressionsanalyse.

Kapitel 2 | Regressionsanalyse

2.1 Mindste kvadraters løsning

Mindste kvadraters løsning (MKL) bruges i regressionsanalyse til at finde den bedste model for et givet datasæt. Metoden går ud på at minimere summen af de kvadrerede risidualer, og vi starter med en simpel lineær regression for N observationer, hvor vi betragter

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \text{for} \quad i = 1, \ldots, N.$$

Her er $\epsilon_i \sim N(0, \sigma^2)$, således, at $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$.
Man kan påvise, at MKL følger fra maksimum likelihood løsningen.
Tæthedsfunktionen for y_i er givet ved:

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(y_i - (\beta_0 + \beta_1 x_i))^2\}.$$

For de uafhængige variabler $y_1, \ldots, y_N$ har vi

$$f(y_1, \ldots, y_N) = \prod_{i=1}^{N} f(y_i) = (\frac{1}{\sigma\sqrt{2\pi}})^N \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - (\beta_0 + \beta_1 x_i))^2\}.$$

Antager man, at σ er kendt, vil likelihoodfunktionen opfylde

$$L(\beta_0, \beta_1) \propto \exp\Big\{ -\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - (\beta_0 + \beta_1 x_i))^2 \Big\}.$$

Vi kan se, at likelihoodfunktionen $L(\beta_0, \beta_1)$ maksimeres ved at minimere

$$\min_{\beta_0, \beta_1} \Big\{ \sum_{i=1}^{N} \epsilon_i^2 \Big\},$$

hvor $\epsilon_i = y_i - (\beta_0 + \beta_1 x_i)$ er risidualen for y_i.
Lad os have et datasæt med N observationer $\{y_i, x_i\}_{i=1}^N$, hvor hver $\boldsymbol{x_i} = (x_{i1}, \ldots, x_{in})$ er en n-dimensional vektor af forklarende variabler og hver $y_i \in \mathbb{R}$ er tilhørende responsvariabel.
For multiple lineære regressionsmodeller er responsvariablen y_i en lineær funktion, givet ved

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + x_{in}\beta_n + \epsilon_i.$$

Vi kan udtrykke responsen $\boldsymbol{Y} = (y_1, \ldots, y_N)$ på matrixform således:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.1}$$

hvilket er ækvivalent med

$$||\boldsymbol{\epsilon}|| = \sqrt{\epsilon_1^2 + \ldots + \epsilon_N^2} = ||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||.$$

Minimering af $\sum_{i=1}^{N} \epsilon_i^2 = ||\boldsymbol{\epsilon}||^2$ med hensyn til $\boldsymbol{\beta}$ gøres ved at finde minimum

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^n} \left\{ ||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 \right\} = \min_{\beta_0,\beta} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{n} x_{ij}\beta_j)^2 \right\}. \tag{2.2}$$

MKL virker efter hensigten, når responsvariablen y_i kun er afhængig af en variabel $\boldsymbol{x_i}$, men den er ikke optimal, når y_i er afhængig af flere variabler. Der er generalt følgende ulemper med MKL:

- MKL kan ikke skelne mellem variabler med ringe eller ingen indflydelse på responsvariablen.
- MKL er brugbar, så længe antallet af observationer N er større end antallet af forklarende variabler n. Hvis $n \approx N$, risikerer man at overse observationer, hvilket medfører, at estimatet ved MKL ikke nødvendigvis er sand. Hvis $N < n$ vil MKL ikke være entydig, og derfor vil denne metode i dette tilfælde fejle.

2.2 Regulering

For de store datasæt gælder der tit og ofte, at antallet af observationer N er betydeligt mindre end antallet af forklarende variabler n. Dette gælder blandt andet også for det datasæt, vi arbejder med. I disse tilfælde vil MKL ikke kunne bruges, og det er nødvendigt at regulere estimeringsprocessen. Dette kan gøres ved at indføre en afgrænsning $C \subseteq \mathbb{R}^n$ således, at man estimerer parametrene β_0 og $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)$ ved at minimere:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^n} \left\{ ||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 \right\}, \text{ afgrænset af } \boldsymbol{\beta} \in C. \tag{2.3}$$

Valget af C kommer til at have en betydning for estimaterne $\hat{\boldsymbol{\beta}}$.
Det typiske valg, er at vælge C enten med hensyn til ℓ_1-normen, som er defineret ved:

$$||\boldsymbol{\beta}||_1 = \sum_{i=1}^{n} |\beta_i|$$

eller ℓ_2-normen, som er defineret ved

$$||\boldsymbol{\beta}||_2 = \sqrt{\sum_{i=1}^{n} \beta_i^2}.$$

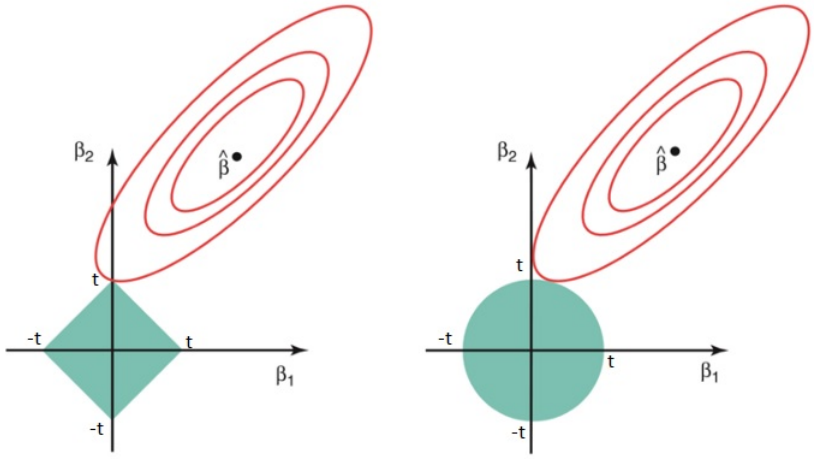
Når man bruger ℓ_1 normen refereres der til lasso regression, og minimeringsproblemet er givet ved

$$\hat{\boldsymbol{\beta}}_{lasso} = \min_{\boldsymbol{\beta} \in \mathbb{R}^n} \left\{ ||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 \right\}, \text{ afgrænset af } ||\boldsymbol{\beta}||_1 \leq t. \tag{2.4}$$

I tilfælde af ℓ_2 normen refererer man til ridge regression:

$$\hat{\boldsymbol{\beta}}_{ridge} = \min_{\boldsymbol{\beta} \in \mathbb{R}^n} \left\{ ||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 \right\}, \text{ afgrænset af } ||\boldsymbol{\beta}||_2 \leq t. \tag{2.5}$$

I ovenstående problemer kontrollerer t radiussen af det afgrænsede område og derved størrelsen af afgrænsningsmængden, som bliver dannet. Figur 2.1 illustrerer, hvordan koefficienterne β_1 og β_2 estimeres under lasso regression (til venstre) og ridge regression (til højre). De grønne områder præsenterer afgrænsningsmængder, der reguleres for at minimere $||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2$, afbildet ved de elliptiske niveaukurver.
Hvor man for lasso regression kan finde et t således, at koefficienten β_i er 0, er dette ikke muligt for ridge regression, da niveaukurverne ikke rører afgrænsningsområdet, når β_i er 0. Fordelen ved at have en koefficient β_i til at være 0 er, at den tilhørende forklarende variabel $\boldsymbol{x_i}$ derved ikke kommer til at have betydning for responsvariablen.



Figur 2.1: Illustration af estimering af lasso regression (til venstre) og ridge regression (til højre). De grønne områder er afgrænsningsområderne, med $|\beta_1| + |\beta_2| \leq t$ for lasso og $\beta_1^2 + \beta_2^2 \leq t^2$ for ridge. De røde ellipser er niveaukurver for $||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2$. Illustrationen er taget fra [4].

Reguleringsproblemet med hensyn til ℓ_q normen, hvor $q > 0$, udtrykkes ved

$$\min_{\beta_0,\beta} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{n} x_{ij}\beta_j)^2 \right\} \text{ afgrænset af } \beta_0^q + \beta_1^q + \ldots + \beta_n^q \leq t^q. \tag{2.6}$$

I stedet for at formulere reguleringsproblemet i term af radiussen t, kan man omskrive udtryk (2.6) på Lagrange form (se Appendiks A.1) :

$$\min_{\beta_0,\beta} \mathcal{L}(\beta_0, \beta, \lambda) = \min_{\beta_0,\beta} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{n} x_{ij}\beta_j)^2 + \lambda(\beta_0^q + \beta_1^q + \ldots + \beta_n^q - t^q) \right\} \tag{2.7}$$

og derved indføre parameteren λ, der bestemmer styrken af 'straffen' og kaldes en vendingsparameter. Da leddet $-\lambda t^q$ i (2.7) ikke er relevant med hensyn til minimeringsproblemet, kan vi udtrykke ℓ_q reguleringen på Lagrange form:

$$\hat{\boldsymbol{\beta}}_q = \min_{\boldsymbol{\beta} \in \mathbb{R}^n} \left\{ ||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_q \right\}, \tag{2.8}$$

hvor $||\boldsymbol{\beta}||_q$ kaldes ℓ_q-straffen.
På denne måde vil man i stedet for at kontrollere radiussen t af afgrænsningsmængden, kontrollere styrken af straffen λ, hvilket i nogle tilfælde kan være en fordel.
Indtil videre har vi kigget på lineær regression for normalfordelte responsvariabler, men de samme principper gælder også for den logistiske regression.
Når test størrelsen N er mindre end antallet af forklarende variabler n kan logistisk regression ikke bruges uden modifikation, og for den generaliserede model minimerer man den negative

log-likelihood sammen med en ℓ_q straf

$$\min_{\beta_0,\boldsymbol{\beta}}\{-\frac{1}{N}\ell(\beta_0,\boldsymbol{\beta};\boldsymbol{Y},\boldsymbol{X})+\lambda||\boldsymbol{\beta}||_q\} \tag{2.9}$$

Her er $\ell(\beta_0,\boldsymbol{\beta};\boldsymbol{Y},\boldsymbol{X})$ log-likelihoodfunktionen, som er givet i (1.3), udfaldet er en N-dimensional vektor $\boldsymbol{Y}$, og prædiktorene er givet ved matricen $\boldsymbol{X}$ med størrelsen $N \times n$.

Givet den logistiske model (1.28), vil den negative log-likelihood med ℓ_1 regulering være givet ved

$$-\frac{1}{N}\sum_{i=1}^{N}\Big[y_i\log(Pr[Y=1|\boldsymbol{x_i}])+(1-y_i)\log(Pr[Y=0|\boldsymbol{x_i}])\Big]+\lambda||\boldsymbol{\beta}||_1 =$$

$$=-\frac{1}{N}\sum_{i=1}^{N}\Big[y_i(\beta_0+\boldsymbol{\beta}^T\boldsymbol{x_i})-\log(1+\exp\{\beta_0+\boldsymbol{\beta}^T\boldsymbol{x_i}\})\Big]+\lambda||\boldsymbol{\beta}||_1,$$

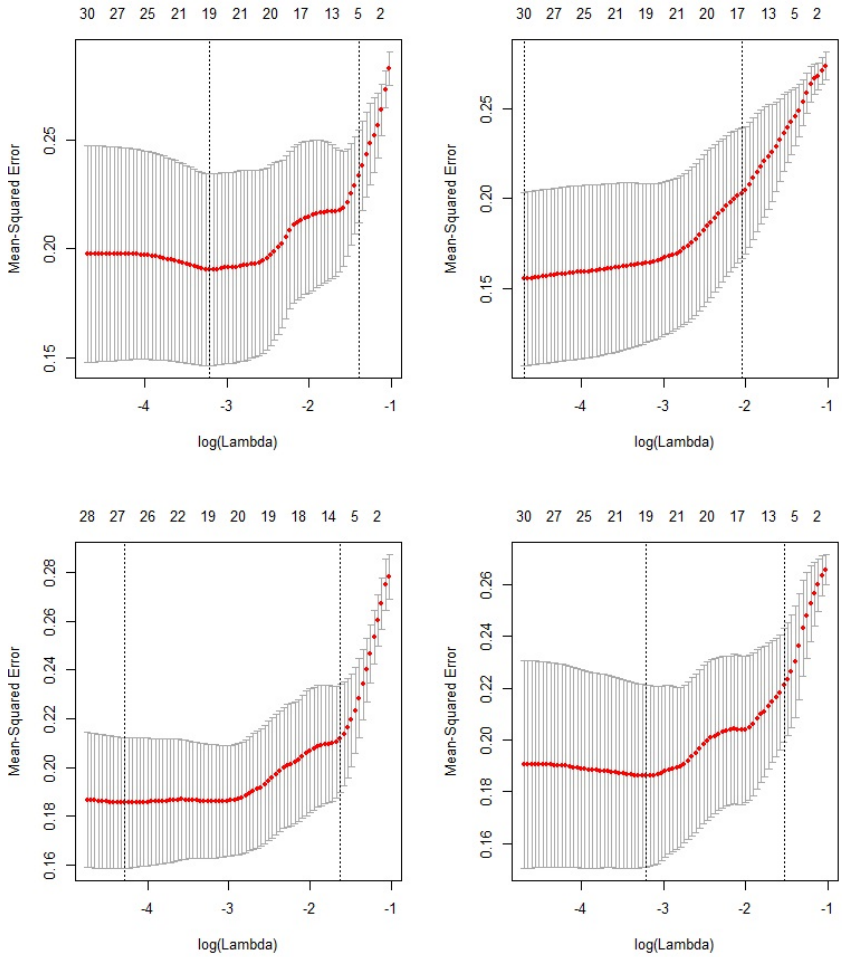
hvor y_i er udfaldsværdierne.

2.2.1 K- foldet krydsvalidering

I dette afsnit kigger vi primært på lasso regression. Afgrænsningen t styrer kompleksiteten af lasso regressionen (2.4). Store værdier af t gør, at modellen kommer til at fitte dataene tæt, men for store værdier af t kan føre til overfitting. Det betyder, at usignifikante variabler ifølge modellen kan komme til at have større betydning, end de har i virkeligheden. Omvendt vil for små værdier af t medføre, at modellerne ikke kommer til at fitte dataene tæt, og man risikerer at miste signifikante variabler. Det samme er tilfældet for λ i Lagrange formen. For store værdier af λ risikerer man at miste signifikante koefficienter og for små værdier, at koefficienterne ikke får den 'bedste' forklarende værdi med hensyn til variablerne.
Vi ønsker at finde det t eller λ, som balancerer mellem disse to problemstillinger således, at vi hverken kommer til at overfitte dataene eller til at miste nogle signifikante variabler. En løsning er at bruge K-foldet krydsvalidering. I det følgende forklares metoden med hensyn til λ, men samme fremgangsmåde er gældende for t.
K-foldet krydsvalidering går ud på at dele datasættet op i K grupper, hvor K typisk er 5 eller 10. Observationerne placeres tilfældigt i de K forskellige grupper, og grupperne analyseres derefter med hensyn til de forskellige λ værdier. Vi lader en af de K grupper være testsættet/valideringssættet og lader de resterende $K-1$ grupper være træningssættet. Antager man, at der er N observationer, vil der være $\frac{N}{K}$ observationer i hver gruppe. Man udfører lasso regression på træningsdataene med forskellige λ værdier og bruger hver fittede model til at forudsige responsen i testsættet. Vi registrerer prædiktionsfejlen for hver værdi af λ. Denne procedure gentages K gange således, at hver af de K grupper har muligheden for at være testdata, hvor resterende $K-1$ grupper fungerer som træningsdata. På denne måde opnås der K forskellige estimater af prædiktions fejl over en række λ værdier. Den overordnede prædiktionsfejl, krydsvalideringsfejlen af λ, udregnes ved at tage gennemsnittet af prædiktionsfejlene i hver

gruppe K. Krydsvalideringsfejlkurven (Figur 2.2) for λ viser den estimerede middelkvadratfejl (mean squared error) afhængig af de forskellige λ værdier.
Da det er tilfældigt, hvordan observationerne placeres i de forskellige K grupper, er der flere mulige krydsvalideringer for et givet datasæt.

Dataanalyse 2.1. Krydsvalidering på datasættet "*Gene*"
Figur 2.2 viser fire mulige grafer efter krydsvalidering af datasættet "*Gene*" (se Appendiks B.4). Graferne viser sammenhængen mellem den middelkvadratfejl, man får fra modellen, og den tilhørende λ værdi.



Figur 2.2: Krydsvalideringer af datasættet "*Gene*".

Spørgsmålet er nu, hvilken λ man skal vælge for at opnå den bedst mulige model for sit datasæt. Man ønsker typisk at minimere middelkvadratfejlen, og den første tanke er, at den λ, som svarer til den mindste middelkvadratfejl, λ_{min}, er den bedste med hensyn til modellering af datasæt. Samtidig skal man tage højde for, at fejlkurverne er tilfældige. Dette betyder, at man for hver

krydsvalidering får forskellige λ, som svarer til den mindste krydsvalideringsfejl i hver enkelte krydsvalidering.
Et alternativ valg af λ kaldes "*one standard error rule*" og handler om at vælge den største λ indenfor 1 standard afvigelse fra den mindste krydsvalideringsfejl $\lambda_{1.se}$. Vi har således to mulige λ valg, hvor den anden nævnte har større værdi, og dermed mere regulering. Mere regulering er typisk det, man har brug for, når man har mange variabler. Valget af λ påvirker desuden antallet af signifikante forklarende variabler. Lavere værdi af λ resulterer i flere signifikante variabler, mens større λ værdier resulterer i færre forklarende variabler.
Da fejlkurverne er tilfældige, kan man reducere tilfældigheden og finde λ, ved at udføre et passende stor antal af krydsvalideringer og tage gennemsnittet af de fundne λ værdier.

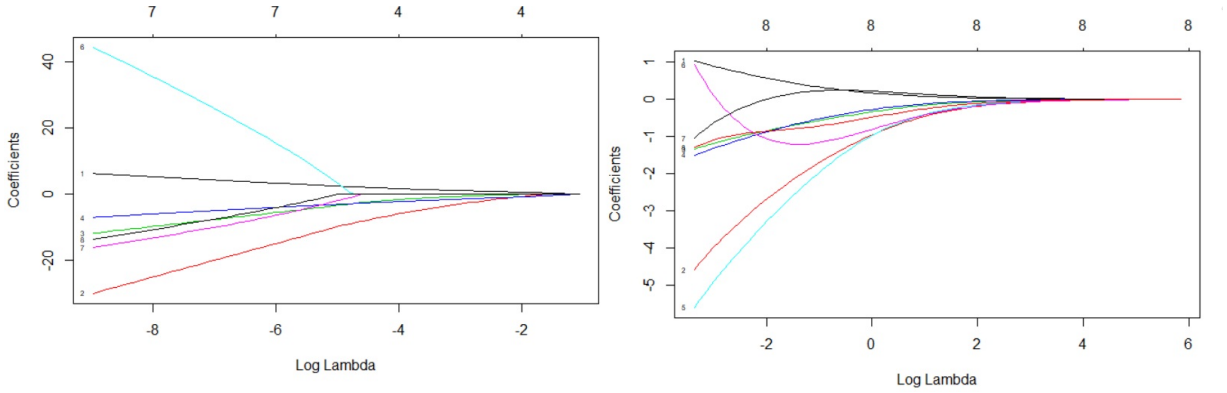
Dataanalyse 2.2. Antal krydsvalideringer
For datasættet "*Gene*" har vi udført et antal krydsvalideringer og har taget gennemsnittet af λ_{min} og $\lambda_{1.se}$, for at se, om λ_{min} og/eller $\lambda_{1.se}$ bliver stabiliseret.
Vi har undersøgt, hvordan antallet af krydsvalidering på datasættet "*Gene*" påvirker λ_{min} og $\lambda_{1.se}$, ved at foretage 10, 100, 500, 1000 og 2000 krydsvalideringer med lasso regression. Hvert antal antal af krydsvalideringer blev gentaget 3 gange.
For datasættet "*Gene*" viste det sig, at omkring 2000 krydsvalideringer ville være et passende antal for, at $\lambda_{min} \approx 0.12$ og $\lambda_{1.se} \approx 0.28$ stabiliserer sig (se Appendiks B.4). Når vi bruger lasso regression med $\lambda_{min} \approx 0.12$, får vi 17 signifikante gener, mens $\lambda_{1.se} \approx 0.28$ resulterer i fire signifikante gener: Prob2136, Prob2391, Prob6080 og Prob6185.

2.2.2 De signifikante variabler

Jo flere forklarende variabler responsvariablen $\boldsymbol{Y}$ afhænger af, jo besværligere kan det være at finde ud af, hvilke der er signifikante. Graferne på Figur 2.3 afspejler regressionskoefficienterne β_0 og $\boldsymbol{\beta}$ som funktion af λ og repræsenterer forskellige stier for, hvordan koefficienterne opfører sig for de forskellige λ værdier. Stierne skal læses fra højre til venstre, og ved at se på de stier, som hører til de forskellige regressionskoefficienter, kan man afgøre, hvilke koefficienter er signifikante. De stier, som starter først, er dem, hvis tilhørende variabler er signifikante og derved også har betydning for responsvariablen.

Dataanalyse 2.3. Lasso og ridge regression på datasættet "$MindreData$".
Figur 2.3 (til venstre) viser resultat af lasso regression på datasættet "$MindreData$" (se Appendiks B.4). For eksempel svarer den blå graf til gen Prob3272. Resultatet af ridge regression kan ses til højre. Som man kan se ud fra graferene, kan man fra ridge stien ikke afklare, hvilke koefficienter, der skiller sig ud. Derfor tager vi udgangspunkt i lasso stien. Stierne, som starter først (såsom 1,2,3 og 4), svarer til de signifikante variabler, mens dem, der kommer til sidst (såsom 6,7 og 8), ikke gør. Dataanalyse på datasættet "$MindreData$" viser, at uanset, om man vælger λ_{min} eller $\lambda_{1.se}$, vil den logistiske lasso regression skrumpe 4 ud af 8 regressionskoefficienter til 0, og de koefficienter, som er tilbage, hører til generne:

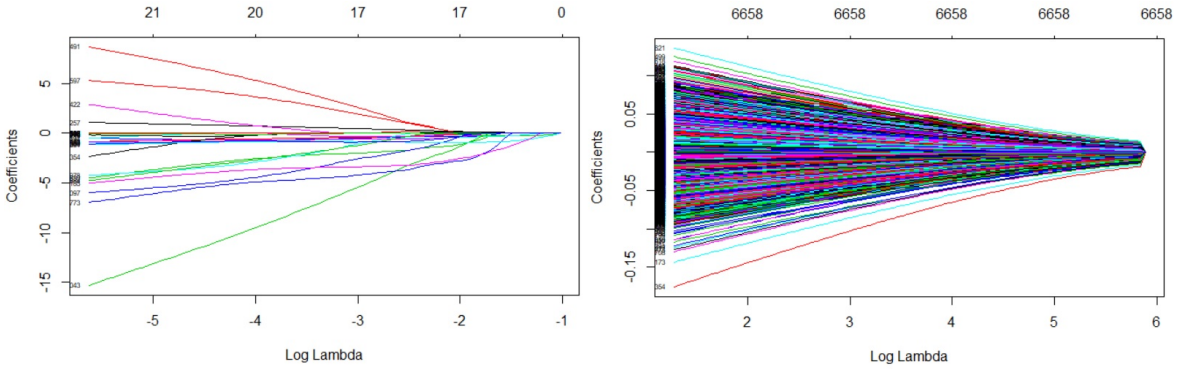
Prob257, Prob634, Prob1949 og Prob2391.



Figur 2.3: Lasso stien for datasættet "*MindreData*" til venstre og ridge stien til højre.

Hvor ovenstående tager udgangspunkt i et mindre datasæt med 8 forklarende variabler, kan det være besværligt at arbejde med stierne med mange flere variabler, se Figur 2.4.

Dataanalyse 2.4. Lasso og ridge regression på datasættet "*Gene*"
Figur 2.4 viser lasso og ridge stier for datasættet "*Gene*". Som man kan se, er det vanskeligt at afgøre noget fra hverken lasso eller ridge stierne.



Figur 2.4: Lasso stien for datasættet "*Gene*" til venstre og ridge stien til højre.

Et af problemerne med regulering er, at vi ikke kan være sikre på, om vi har for meget eller for lidt regulering. Ønsker man mere regulering, er $\lambda_{1.se}$ et godt valg, og for mindre regulering er λ_{min} foretrukket.

2.2.3 Fordele og ulemper ved lasso og ridge regression

Idet lasso regression sætter usignifikante regressionskoefficienter til at være 0, må man spørge sig selv om, hvornår det giver mening at bruge den på et datasæt. Hvis der er få forklarende variabler, kunne det tænkes, at lasso regression tvinger koefficienter til 0, selvom den tilhørende

forklarende variabel har betydning for responsvariablen. Derudover, hvis vi omvendt har mange forklarende variabler, der har nogenlunde lige stor betydning for responsen, risikerer man også at få en fejlmodel. Hvis der er mange variabler, og man kun er interesseret i dem med signifikant betydning for responsen, er lasso regression en løsning. Ulempen ved lasso regression er, at hvis to signifikante variabler er højt korrelerede, vil en tilfældig af disse eventuelt blive sat til 0.
Ridge regressionen vil ligesom lasso også skrympe de forskellige regressionskoefficienter, men på grund af ℓ_2 normen, vil disse aldrig kunne blive 0. Fordelen ved, at koefficienter ikke bliver skrympet til 0, er, at de tilhørende variabler ikke mister signifikans. Derved vil ridge regressionen kunne være en bedre løsning end lasso, når man arbejder med et datasæt, hvor der enten er få variabler eller mange variabler med lige stor betydning, og/eller nogle af variablerne er højt korrelerede.

2.3 Elastisk net

I tilfælde med mange forklarende variabler kan man være udsat for, at der er nogle grupper af variabler, som korrelerer med hinanden. Hvis det er tilfældet, vil lasso regression højst sandsynligt fejle. For at løse dette bruger man elastisk net regression, hvor der tilføjes en kvadratisk del $||\boldsymbol{\beta}||_2$ til straffen i lasso regressionen.
Elastisk net er en kompromis mellem ridge og lasso straffene og giver estimatet $\hat{\boldsymbol{\beta}}_{enet}$ defineret ved:

$$\hat{\boldsymbol{\beta}}_{enet} = \min_{\boldsymbol{\beta}\in\mathbb{R}^n} \left\{ ||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda_2||\boldsymbol{\beta}||_2^2 + \lambda_1||\boldsymbol{\beta}||_1 \right\}, \tag{2.10}$$

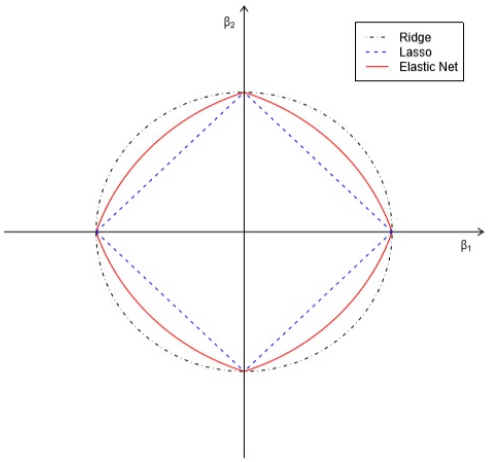
hvor ridge fås ved at lade $\lambda_1 = 0$ og $\lambda_2 = \lambda$, og lasso fås ved at lade $\lambda_1 = \lambda$ og $\lambda_2 = 0$.
Ved at definere elastisk net straffen, $J(\boldsymbol{\beta})$:

$$J(\boldsymbol{\beta}) = \frac{1}{2}(1-\alpha)||\boldsymbol{\beta}||_2^2 + \alpha||\beta||_1, \tag{2.11}$$

hvor $\alpha = \frac{\lambda_2}{\lambda_1+\lambda_2}$, kan man også udtrykke elastisk net optimeringsproblem således:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^n} \left\{ ||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 \right\} \text{ afgrænset af } J(\boldsymbol{\beta}) \leq t. \tag{2.12}$$

Funktionen $J(\boldsymbol{\beta})$ er en konveks kombination af lasso og ridge straffen (se Appendiks A.2), og α bestemmer vægten af lasso og ridge straffen. Når $\alpha = 1$ fås lasso afgrænsningen, mens $\alpha = 0$ resulterer i ridge afgrænsningen.
Hvor ℓ_1 delen af straffen frembringer en udtyndet model, resulterer den kvadratiske del $||\boldsymbol{\beta}||_2^2$ i, at der ikke er begrænsninger med hensyn til antallet af valgte variabler.
Figur 2.5 illustrerer straffen fra den elastiske net med $\alpha = 0.5$, sammenlignet med straffen fra lasso og ridge regression. Området af afgrænsningen viser, at straffen for den elastiske net med $\alpha = 0.5$ er strengt konveks. Specielt gælder der, at når $\alpha < 1$, vil minimeringsproblemet (2.12) være strengt konveks, hvilket betyder, at vi har en entydig løsning, og at α styrer graden af konveksiteten.



Figur 2.5: Konturplot af straffen for den elastiske net sammenlignet med konturplot af straffen for ridge og lasso regression. Illustrationen er taget fra [6].

Regulering af α giver mulighed for, at flere variabler bliver signifikante.

Dataanalyse 2.5. Elastisk net regression for datasættet "*Gene*"
Hvor vi med lasso regression på datasættet "*Gene*" fik 17 signifikante gener ved brug af den gennemsnitlige λ_{min} og 4 ved brug af den gennemsnitlige $\lambda_{1.se}$, giver elastisk net regression mulighed for flere signifikante gener, da den, som nævnt, også tager hensyn til, de korrelerede grupper af gener.
Jo tættere α er på 0, jo flere gener vil kvalificeres som signifikante. Tabel 2.1 viser antallet af signifikante gener, vi får som resultat af elastisk net regression på datasættet "*Gene*" med forskellige α værdier.
Valget af α afhænger blandt andet af, hvor mange gener der er behov for at tage udgangspunkt i. Hvor 294 gener kunne tænkes at være for mange gener at tage stilling til, når man skal afgøre, om en person er syg, kan for få gener også være misledende. I forhold til datasættet "*Gene*" vides der ikke på forhånd, hvilke gener korrelerer med hinanden, derfor ville $\alpha = 0.5$, som giver ligevægt af lasso og ridge regression, være et fornuftigt valg. Med $\alpha = 0.5$ har vi 45 gener at tage stilling til. Sammenlignet med lasso regression vil vi desuden komme til at tage hensyn til de gener, som enten korrelerer og/eller næsten er lige så signifikante, som de gener, vi fik fra lasso.
Dataanalysen (se Appendiks B.5) bekræfter, at der er grupper af de korrelerede gener. Vi har blandt andet fundet ud af, at der for $\alpha = 0.5$ er én gruppe på 5 korrelerede gener: Prob21, Prob40, Prob53, Prob290, Prob1542 og to grupper på to korrelerede gener: Prob2269 og Prob2391, samt Prob6542 og Prob6555. Det viste sig, at den sidste gruppe består af to identiske gener, og dette er sandsynligvis resultat af indtastningsfejl.

α	λ_{min}	Antal signifikante variabler λ_{min}	$\lambda_{1.se}$	Antal signifikante variabler $\lambda_{1.se}$
0	≈ 4.10	6658	≈ 20.16	6658
0.1	≈ 0.045	501	≈ 0.54	294
0.3	≈ 0.020	195	≈ 0.29	84
0.5	≈ 0.018	96	≈ 0.22	45
0.7	≈ 0.028	55	≈ 0.23	24
0.9	≈ 0.078	26	≈ 0.27	7
1	≈ 0.12	17	≈ 0.28	4

Tabel 2.1: Antal forklarende variabler efter anvendelse af elastisk net regression på datasættet "*Gene*".

Selvom elastisk net regression fungerer som en god kompromis mellem lasso og ridge regression, er der stadigvæk en del spørgsmål, som er relateret til de konkrete situationer og datasæt.

Kapitel 3 | Inferens for store tal

3.1 T-test

I Dataanalyse 1.10 har vi udført ANOVA på datasættet "*Gene*", dels for at demonstrere, hvordan metoden anvendes i praksis, dels for at påpege, at metoden, som den er, ikke helt er brugbar for de store datasæt. I dette afsnit kigger vi nærmere på, hvilke udfordringer man møder, når man arbejder med hypotesetest på de store datasæt, og på de modifikationer, som kan tilføjes for at opnå tilstrækkelige resultater.
Hvor vi med ANOVA sammenligner flere stikprøver, fungerer ANOVA som t-test, når man kun afgører om middelværdier for to stikprøver er ens. Vi vil for simpelheds skyld holde os til "to-tailed" parret t-test. For de store datasæt kommer vi til at arbejde med n grupper af stikprøver og beregne t-værdien:

$$t_i = \frac{\mu_i(2) - \mu_i(1)}{s_i} \qquad \text{for} \qquad i = 1, \ldots, n,$$

hvor s_i er et estimat for standardfejlen af $\mu_i(2) - \mu_i(1)$, givet ved

$$s_i^2 = \frac{\sum_{j=1}^{N_1}(x_{ji} - \mu_i(1))^2 + \sum_{j=1}^{N_2}(x_{(j+N_1)i} - \mu_i(2))^2}{N-2} \cdot (\frac{1}{N_1} + \frac{1}{N_2}), \tag{3.1}$$

hvor N_1 og N_2 er størrelsen på de to stikprøver, $N = N_1 + N_2$, er samlet antal observationer i gruppen, x_{ji} er værdien for j'te observation i stikprøverne i gruppe i, og $\mu_i(1)$ er middelværdien for den ene stikprøve i gruppe i, mens $\mu_i(2)$ er middelværdien for den anden stikprøve i gruppe i.
Nulhypotesen, $\mathcal{H}_{0i}$, formuleres som:

$$\mathcal{H}_{0i} : \mu_i(1) = \mu_i(2),$$

med den alternative hypotese H_{1i} :

$$\mathcal{H}_{1i} : \mu_i(1) \neq \mu_i(2).$$

Hvis t-værdien for en givet signifikansniveau er større end den kritiske værdi, beregnet fra den kumulative fordelingsfunktion for t_{N-2} fordelingen, forkaster man $\mathcal{H}_{0i}$.

Dataanalyse 3.1. T-fordeling
I datasættet "*Gene*" tager vi udgangspunkt i $N = 30$, med $N_1 = 15$ raske og $N_2 = 15$ syge testpersoner. Vi observerer $n = 6658$ gener. Datasættet er en 30×6658 matrix X, hvor elementet

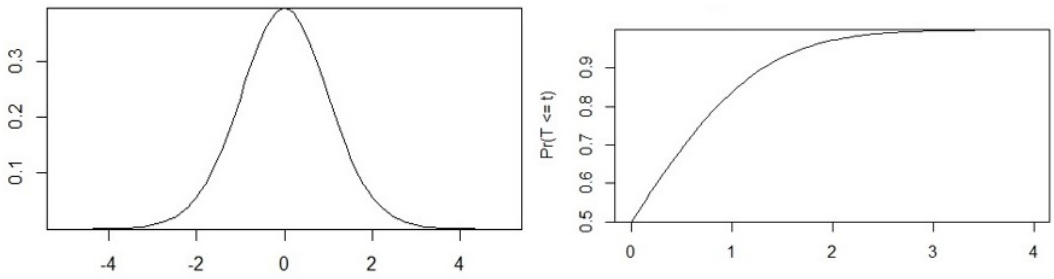
$$x_{ji} = \text{gen } i's \text{ udtryksniveau for testperson } j,$$

hvor $i = 1, 2, \ldots, 6658$ og $j = 1, 2, \ldots, 30$; med $j = 1, 2, \ldots, 15$ for de raske testpersoner og

$j = 16, 17 \ldots, 30$ for testpersonerne med sygdom.
Vi vil teste, om middelværdier for et bestemt gen i er ens. For vores analyse, når der ikke er forskel i middelværdier mellem de raske og de syge testpersoner for et givet gen i, betyder det, at gen i ikke er signifikant for at afgøre om testpersonen er syg. Vi formulerer nulhypotesen:

$$\mathcal{H}_{0i} : \textit{gen i er usignifikant}$$

Ud fra t_i-værdien i formlen (3.1) undersøger man, om x_{ji} er ens fordelt for stikprøverne for de raske og de syge testpersoner. Vi forkaster H_{0i}, hvis den absolutte værdi for t_i er for stor i forhold til den kritiske værdi. For vores datasæt har vi $N - 2 = 28$ frihedsgrader, og vi har valgt en signifikansniveau på 5% ($\alpha = 0.05$) for at forkaste $\mathcal{H}_{0i}$ hypotesen. Figur 3.1 viser tæthedsfunktionen og den kumulative fordelingsfunktion for t-fordelingen med 28 frihedsgrader, og vi forkaster $\mathcal{H}_{0i}$, når $t > 2.05$. (se Appendiks B.6).



Figur 3.1: Tæthedsfunktion og kumulativ fordelingsfunktion for t-fordelingen med 28 frihedsgrader.

I stedet for t-værdier for en enkelt nulhypotese $\mathcal{H}_{0i}$ kan man tage udgangspunkt i z-værdier, som er givet ved

$$z_i = \Phi^{-1}(F_{N-2}(t_i)), \tag{3.2}$$

hvor Φ og F_{N-2} er funktioner, som er fordelt kumulativt for standard normalfordelingen og t_{N-2} fordelingen. Nulhypotesen $\mathcal{H}_{0i}$ kan omskrives til:

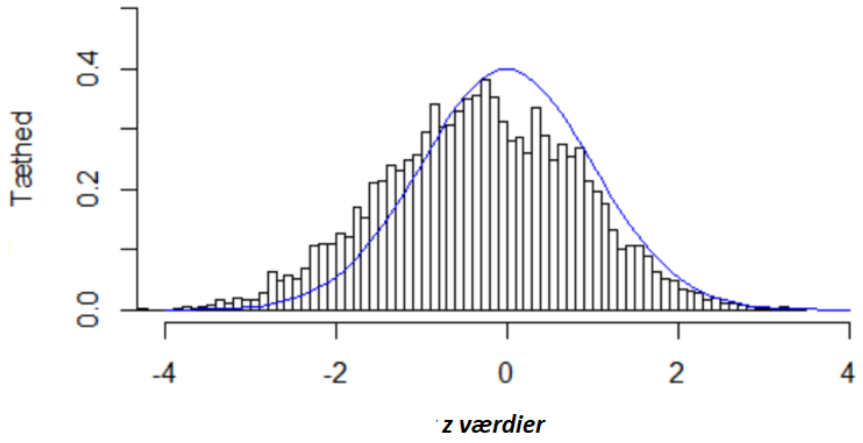
$$H_{0i} : z_i \sim N(0, 1).$$

På denne måde kan man sammenligne fordelingen af z-værdierne for variablerne med tæthedsfunktionen for $\mathcal{N}(0, 1)$:

$$f(z) = c \cdot \frac{\exp\{\frac{-z^2}{2}\}}{\sqrt{2\pi}}, \tag{3.3}$$

hvor man bestemmer parameteren c således, at tæthedskurven er i samme område som histogrammet for de forskellige z-værdier. Som regel justerer man parameter c i (3.3) således, at histogrammet og tæthedsfunktionen er ens normaliseret.
Hvis vores $\mathcal{H}_{0i}$ er sand for alle i, ville alle z_i værdier for de n forskellige variabler komme til at følge normalfordelingskurven.

Dataanalyse 3.2. z-værdier
For at illustrere anvendelse af z-værdier har vi beregnet dem ud fra t-værdierne. (se Appendiks B.6)
Figur 3.2 viser histogrammet for z-værdierne for 6658 forskellige gener. Som man kan, se følger z-værdierne ikke helt normalfordelingskurven, hvilket er et godt resultat, da vi i dette tilfælde forkaster nulhypotesen, at alle gener, er usignifikante. Det bekræftes også i Dataanalyse 1.10, hvor 890 nulhypoteser blev forkastet.



Figur 3.2: Histogrammet for z-værdier for 6658 gener samt normalfordelingskurven $N(0,1)$.

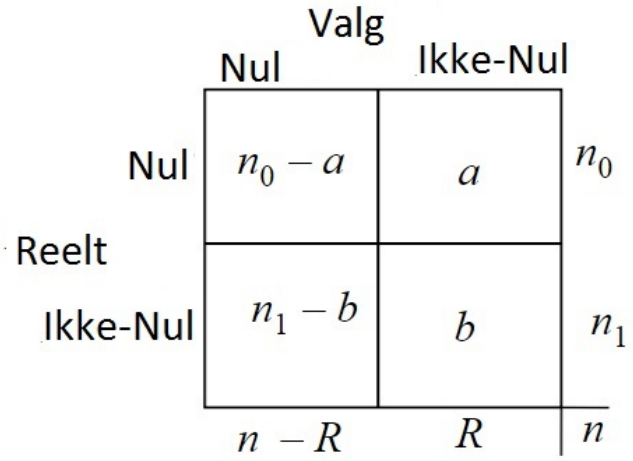
3.2 Sande og falske opdagelser

Når man tester hypoteser, skelner man mellem to typer fejl:

1. Type-I fejl: at forkaste $\mathcal{H}_{0i}$, selvom $\mathcal{H}_{0i}$ er sand.
2. Type-II fejl: at acceptere $\mathcal{H}_{0i}$, selvom $\mathcal{H}_{0i}$ er falsk.

Man skelner mere præcist mellem positive og falsk positive opdagelser. En positiv opdagelse er en signifikant opdagelse, hvor p-værdien er under den valgte signifikansniveau. En falsk positiv er en opdagelse, som ifølge p-værdien er signifikant, men i virkeligheden ikke er det (Type I-fejl). Med en signifikansniveau på $\alpha = 0.05$ er der 5% sandsynlighed for at have en falsk positiv opdagelse. For eksempel, for datasættet "*Gene*", hvor vi har forkastet 890 nulhypoteser, betyder det, at ca. 45 af dem kunne være falsk positive.
Vi sætter $\mathcal{V}$ til at være en valg-regel, som enten producerer valget "nul" for at acceptere nulhypotesen eller "$ikke-nul$", hvis nulhypotesen forkastes for hver af de n testede nulhypoteser $\mathcal{H}_{0i}$. Figur 3.3 viser de valg, valg-reglen $\mathcal{V}$ kunne foretage sig. Lad os sige, at n_0 ud af de n

tilfælde i virkeligheden er nul, og n_1 tilfælde er ikke-nul. Valg-reglen $\mathcal{V}$ har valgt at forkaste $R = a + b$, ud af n nulhypoteser.



Figur 3.3: En valg-regl $\mathcal{V}$, som har forkastet R ud af n nulhypoteser. Af de egentlige nul tilfælde har $\mathcal{V}$ valgt at forkaste a ikke-nul tilfælde, og af de ikke-nul tilfælde har $\mathcal{V}$ valgt at forkaste b ikke-nul tilfælde. Illustrationen er inspireret af [7].

Som man kan se på Figur 3.3, er vi kommet til at forkaste a opdagelser forkert og derved begået en Type-I fejl. Samtidig beholder vi $n_1 - b$ falske opdagelser og således også begår en Type-II fejl.
Den falske opdagelsesrate (FDR) er det forventede antal af sande tilfælde a ud af alle de forkastede tilfælde R og er givet ved:

$$FDR = E[\frac{a}{R}|R > 0]Pr[R > 0].$$

Den falske opdagelsesproportion (FDP) er delen af de forkastede tilfælde, som i virkeligheden er sande nulhypoteser. Dette betyder, at den falske opdagelsesproportion for ovenstående er:

$$FDP = \frac{a}{R}.$$

3.3 Den bayesianske tilnærmelse

I dette afsnit vil vi redegøre for den bayesianske falske opdagelsesrate samt se nærmere på den bayesianske tilnærmelse for at estimere denne.
Lad $\boldsymbol{\theta}$ være en ukendt vektor med a-priori sandsynlighed $g(\boldsymbol{\theta})$. Lad $\boldsymbol{\omega}$ være observerede data, som har tæthedsfunktion $f_{\boldsymbol{\theta}}(\boldsymbol{\omega})$. Bayes formel for den betingede tæthed af $\boldsymbol{\theta}$, givet $\boldsymbol{\omega}$, er

$$g(\boldsymbol{\theta}|\boldsymbol{\omega}) = \frac{g(\boldsymbol{\theta})f_{\boldsymbol{\theta}}(\boldsymbol{\omega})}{f(\boldsymbol{\omega})}, \tag{3.4}$$

hvor $f(\boldsymbol{\omega})$ er den marginale tæthedsfunktionen for $\boldsymbol{\omega}$ givet ved

$$f(\boldsymbol{\omega}) = \int g(\boldsymbol{\theta}) f_{\boldsymbol{\theta}}(\boldsymbol{\omega}) d\boldsymbol{\omega}. \tag{3.5}$$

Lad n være antal af nulhypoteser, hvor enten $\mathcal{H}_{0i}$ gælder med sandsynlighed π_0, eller $\mathcal{H}_{1i}$ gælder med sandsynlighed $\pi_1 = 1 - \pi_0$, og lad z-værdierne have tæthedsfunktionen $f_0(z)$ eller $f_1(z)$ således, at

$\pi_0 = Pr[H_{0i} \text{ gælder}]$	$f_0(z) = \text{tæthedsfunktion, hvis } H_{0i} \text{ gælder}$
$\pi_1 = Pr[H_{1i} \text{ gælder}]$	$f_1(z) = \text{tæthedsfunktion, hvis } H_{1i} \text{ gælder}$

Lad F_0 og F_1 betegne kumulative fordelingsfunktioner svarende til tæthedsfunktionen f_0 og f_1 således, at der for en vilkårlig mængde $\mathcal{Z} \subseteq \mathbb{R}$ gælder, at

$$F_0(\mathcal{Z}) = \int_{\mathcal{Z}} f_0(z)dz \text{ og } F_1(\mathcal{Z}) = \int_{\mathcal{Z}} f_1(z)dz.$$

Med udgangspunkt i begge tæthedsfunktioner $f_0(z)$ og $f_1(z)$ kan den blandede tæthedsfunktion skrives som

$$f(z) = \pi_0 f_0(z) + \pi_1 f_1(z),$$

og den kumulative fordelingsfunktion som

$$F(\mathcal{Z}) = \pi_0 F_0(\mathcal{Z}) + \pi_1 F_1(\mathcal{Z}).$$

Lad os observere $z \in \mathcal{Z}$ værdier, og undersøge, om disse passer til $\mathcal{H}_{1i}$, eller om de passer til $\mathcal{H}_{0i}$. Ud fra Bayes formel (3.4) kan vi udtrykke den betingede sandsynlighed for $\mathcal{H}_{0i}$, givet z-værdier:

$$FDR(\mathcal{Z}) \equiv Pr[\mathcal{H}_{0i} \text{ gælder} | z_i \in \mathcal{Z}] = \frac{\pi_0 F_0(\mathcal{Z})}{F(\mathcal{Z})} \tag{3.6}$$

Vi kalder sandsynligheden $FDR(\mathcal{Z})$ for den bayesianske falske opdagelsesrate. Navnet kommer af, at hvis $z \in \mathcal{Z}$ er givet til at være en sand $\mathcal{H}_{1i}$, vil $FDR(\mathcal{Z})$ give sandsynligheden for, at vi har lavet en falsk opdagelse.
Hvis vi betegner $\bar{F}(\mathcal{Z})$ til at være den empiriske fordelingsfunktion af de n z-værdier

$$\overline{F}(\mathcal{Z}) = \frac{\#\{z_i \in \mathcal{Z}\}}{n},$$

hvor $\#\{z_i \in \mathcal{Z}\}$ betegner antallet af $z_i \in \mathcal{Z}$, beskriver $\overline{F}(\mathcal{Z})$ andelen af observerede z_i værdier i mængden $\mathcal{Z}$. Ved at sætte denne ind i (3.6) kan vi udtrykke den estimerede bayesianske falske opdagelsesrate $\overline{FDR}(\mathcal{Z})$ ved:

$$\overline{FDR}(\mathcal{Z}) \equiv \frac{\pi_0 F_0(\mathcal{Z})}{\overline{F}(\mathcal{Z})} \tag{3.7}$$

Når antallet af tilfælde n er stort nok, vil det forventes, at $\overline{F}(\mathcal{Z})$ er tæt på $F(\mathcal{Z})$, og at $\overline{FDR}(\mathcal{Z})$ vil være en god tilnærmelse for $FDR(\mathcal{Z})$.

3.4 P-værdier

P-værdien af en teststatistik er den laveste signifikansniveau for, hvornår vi kan forkaste en antaget nulhypotese for et givet datasæt. Det er således værdien

$$p = Pr[t(x) \geq t_{obs}],$$

hvilket er sandsynligheden for at observere en værdi af teststørrelsen $t(x)$, der er større end den t-værdi, vi aktuelt står med. Man vælger typisk signifikansniveauet α til at være 0.05, hvor H_{0i} forkastes, når p-værdien p opfylder $p \leq \alpha$.
Når vi tester en mængde af hypoteser simultant, givet et signifikansniveau α, resulterer dette i, at sandsynligheden for tilfældigt at observere en sand $\mathcal{H}_{1i}$ vil være

$$Pr[\textit{mindst et } \mathcal{H}_{1i} \textit{ gælder}] = 1 - Pr[\mathcal{H}_{0i} \textit{ gælder for alle } i] = 1 - (1-\alpha)^n.$$

I tilfælde af store antal testede hypoteser ville der næsten være 100% sandsynlighed for at observere en sand $\mathcal{H}_{1i}$. Det vil sige, at vi altid finder noget uanset, om det er der eller ej.

Eksempel 3.3. Tilfældig observation af mindst ét signifikant gen
Tager vi udgangspunkt i vores mindre datasæt, hvor vi tester 8 hypoteser med en $\alpha = 0.05$, er sandsynligheden for at tilfældigvis observere mindst ét signifikant gen givet ved

$$Pr[\textit{mindst ét signifikant gen}] = 1 - Pr[\textit{ingen signifikante gener}] = 1 - (1-0.05)^8 \approx 0.34.$$

Det vil sige, at der er 34% chance for at observere et signifikant gen, selvom det kunne være tilfældet, at ingen af generne er signifikante.
Situationen for datasættet "*Gene*" er endnu værre. Her vil der næsten være 100% sandsynlighed for at observere mindst et signifikant gen:

$$Pr[\textit{mindst ét signifikant gen}] = 1 - (1-0.05)^{6658} \approx 1.$$

Vi har brug for en justering af p-værdier, og i det næste afsnit kommer vi til at kigge på de forskellige metoder og algoritmer for at opnå denne.

3.5 FWER justering og FDR kontrol

3.5.1 Bonferroni korrektion

Ved Bonferroni korrektionen korrigeres signifikansniveauet til

$$\alpha_{just} = \frac{\alpha}{n},$$

hvor n er antal af tests. Således får man en ny α, som ikke overstiger en kritisk værdi

$$\alpha_{kritisk} = 1 - (1 - \alpha_{just})^n.$$

Eksempel 3.4. Bonferroni korrektion
Med 8 tests og $\alpha = 0.05$ vil vi først forkaste nulhypotesen $\mathcal{H}_{0i}$, hvis p-værdien er under $\alpha_{just} = 0.00625$. Sandsynligheden for, at man mindst har ét signifikant resultat:

$$Pr[\textit{mindst ét signifikant gen}] = 1 - (1 - 0.00625)^8 \approx 0.049$$

hvilket betyder, at sandsynligheden for at tilfældigt forkaste $\mathcal{H}_{0i}$ kommer under de ønskede 5%.
Vælger vi $n = 6658$ og $\alpha = 0.05$, vil $\alpha_{just} \approx 7.51 \cdot 10^{-6}$, og

$$Pr[\textit{mindst ét signifikant gen}] = 1 - (1 - \alpha_{just})^{6658} \approx 0.049.$$

Idet α_{just} tilnærmer sig 0, gør Bonferroni korrektionen det svært at forkaste nulhypotesen $\mathcal{H}_{0i}$.

Bonferroni korrektionen er således beregnet til at minimere risikoen for at få mindst ét signifikant resultat, som ikke er signifikant, men for store n værdier begynder det at være svært at forkaste nulhypoteserne.
Familie wise error rate (FWER) er defineret som sandsynligheden for at lave mindst én falsk forkastelse i en familie af hypotesetest, hvilket betyder, det er sandsynligheden for at lave mindst én Type-I fejl.
Tager vi udgangspunkt i betingelserne fra Figur 3.3, kan vi udtrykke FWER som

$$FWER = Pr[a \geq 1].$$

Man kan præsentere FWER kontrolprocedure som en algoritme, hvor input er p-værdier $p_1, \ldots, p_n$, og output er en liste af accepterede og forkastede nulhypoteser. Man kontrollerer FWER ved at lave en begrænsning

$$FWER \leq \alpha. \tag{3.8}$$

Bonferroni korrektionen udvider den klassiske FWER kontrol metode til at forkaste de nulhypoteser, som opfylder

$$p_i \leq \frac{\alpha}{n}. \tag{3.9}$$

Ved at sætte I_0 til at være mængden af indeks for de sande nulhypoteser med n_0 medlemmer, fås der ud fra begrænsningen (3.9) og Booles ulighed[1], at

$$FWER = Pr[\bigcup_{I_0}(p_i \leq \frac{\alpha}{n})] \leq \sum_{I_0} Pr[p_i \leq \frac{\alpha}{n}] = n_0 \cdot \frac{\alpha}{n} \leq \alpha.$$

[1] Booles ulighed: $Pr[\bigcup A_i] \leq \sum Pr[A_i]$

Bonferroni begrænsningen opfylder derved FWER kontrolproceduren.
Problemet med Bonferroni korrektionen er, at jo flere nulhypoteser vi tester, jo svære bliver det at forkaste en falsk $\mathcal{H}_{0i}$. Vi har derfor brug for en forbedring, som gør det lettere at forkaste en falsk $\mathcal{H}_{0i}$.

3.5.2 Holms Procedure

Holms procedure giver et større område til at forkaste en falsk $\mathcal{H}_{0i}$, sammenlignet med Bonferroni korrektionen. Proceduren går ud på:

1. At sortere p-værdierne således:
$$p_1 \leq p_2 \ldots \leq p_i \leq \ldots \leq p_n.$$
2. For trin $i = 1, ..., n$. Hvis der gælder
$$p_i \leq \frac{\alpha}{n - i + 1},$$
forkaster man $\mathcal{H}_{0i}$ og går videre til næste trin. Ellers beholder man $\mathcal{H}_{0i}, \ldots, \mathcal{H}_{0(n-1)}$ og stopper.

Proceduren starter således med $i = 1$, og nulhypotesen $\mathcal{H}_{0i}$ forskastes indtil første i, hvor
$$p_i > \frac{\alpha}{n - i + 1}.$$
Holms procedure kontrollerer, ligesom Bonferroni korrektionen, FWER.
Lad I_0 være mængden af indeks for de sande nulhypoteser med n_0 medlemmer og j være det mindste indeks, som opfylder $p_j = \min_{i \in I_0} p_i$.
Idet $j \leq n - n_0 + 1$, gælder der
$$\frac{\alpha}{n - 1 + j} \leq \frac{\alpha}{n_0},$$
og således
$$FWER = Pr[\min_{i \in I_0} p_i \leq \frac{\alpha}{n_0}] \leq \sum_{i \in I_0} Pr[p_i \leq \frac{\alpha}{n_0}] \leq \alpha, \tag{3.10}$$
hvilket betyder, at Holms procedure kontrollerer FWER.
Holms procedure accepterer hypoteserne med de mindste p-værdier, og i modsætning til Bonferroni korrektionen gives der også mulighed for at acceptere hypoteser med højere p-værdi end $p = \frac{\alpha}{n}$.

Dataanalyse 3.5. Holms procedure
For datasættet "*MindreData*" havde vi fra ANOVA 4 forkastede nulhypoteser. Ved anvendelse af Holms procedure på dette datasæt blev antallet af de forkastede hypoteser ikke reduceret yderligere (se Appendiks B.7).
Anvendelsen af Holms procedure på datasættet "*Gene*" medførte, at antallet af forkaste-

de hypoteser blev reduceret fra 890 hypoteser til 1. Det signifikante gen er ifølge Holms procedure Prob6080. Vi kan konstatere, at i tilfældet med vores datasæt og $n = 6658$ virker Holms procedure ikke efter hensigten, og har de samme begrænsninger som Bonferroni korrektionen. Tabel 3.1 viser korrigerede p-værdier for de første tre trin for Holms procedure.

i	Holms kriterium $\frac{\alpha}{n-i+1}$	Bonferronis kriterium $\frac{\alpha}{n}$
1	$7.510 \cdot 10^{-6}$	$7.510 \cdot 10^{-6}$
2	$7.511 \cdot 10^{-6}$	$7.510 \cdot 10^{-6}$
3	$7.512 \cdot 10^{-6}$	$7.510 \cdot 10^{-6}$

Tabel 3.1: Kritiske p-værdier for Holms procedure og Bonferroni korrektion for datasættet "*Gene*".

3.5.3 Benjamini Hochbergs algoritme

Som man kan se fra Dataanalyse 3.5, er der behov for nogle andre kontrolprocedurer for de større datasæt.
I stedet for at kontrollere sandsynligheden for at forkaste en sand nulhypotese (FWER), vil vi kontrollere antallet af de forkastede nulhypoteser, som reelt er sande (FDR). Vi vil gerne have minimeret andelen af de falsk positive opdagelser. Dette kan gøres ved hjælp af Benjamini Hochbergs algoritme.
Lad valg-reglen $\mathcal{V}$ producerer p-værdier, p_i, for hvert tilfælde i således, at p_i er uniformt fordelt, hvis $\mathcal{H}_{0i}$ er sand.
Benjamini Hochbergs (BH) algoritme går ud på:

1. At sortere p-værdierne således:

$$p_1 \leq p_2 \ldots \leq p_i \leq \ldots \leq p_n.$$

2. At udregne de individuelle p-værdiers BH kriterium, som

$$p_{kri} = \frac{i}{n} q,$$

hvor n er antal af tests, mens q er en fastsat værdi i $[0, 1]$.

3. At sammenligne de originale p-værdier med de kritiske BH værdier fundet i trin 2. Her lader man i_{max} være det største indeks, hvor der gælder:

$$p_i \leq p_{kri}.$$

Man forkaster $\mathcal{H}_{0i}$, for

$$i \leq i_{max},$$

med andre ord, når den BH justerede p-værdi er større eller lig med den originale p-værdi. Man accepterer $\mathcal{H}_{0i}$ for

$$i > i_{max}.$$

Man kan også starte BH algoritmen med $i = n$ og bevæge sig mod $\mathcal{H}_{01}$, hvor nulhypotesen $\mathcal{H}_{0i}$ accepteres, indtil første gang

$$p_i \leq \frac{i}{n} q,$$

hvorefter vi forkaster $\mathcal{H}_{0i}$.

Dataanalyse 3.6 (Benjamini Hochbergs algoritme)**.**
BH algoritmen på datasættet "*Gene*" med $q = 0.05$ resulterede i 5 forkastede nulhypoteser, som svarer til følgende signifikante gener: Prob21, Prob53, Prob257, Prob2136 og Prob6080 (se Appendiks B.7).
Tabel 3.2 viser antallet af forkastede nulhypoteser efter anvendelse af BH algoritmen med forskellige q-værdier. Ved at justere q-værdien kan man justere antallet af de forkastede hypoteser, og jo større q-værdien er, jo flere nulhypoteser forkaster man.

q	0.01	0.05	0.1	0.2
Antal forkastede hypoteser	1	5	24	270

Tabel 3.2: Sammenhængen mellem q-værdier og antallet af de forkastede nulhypoteser efter anvendelse af BH algoritmen på datasættet "*Gene*" .

Hvor Bonferroni korrektionen og Holms procedure kontrollerer FWER, viser vi i det følgende, at BH-algoritmen kontrollerer FDP og således FDR. Vi tager udgangspunkt i n nulhypoteser $\mathcal{H}_{0i}$, hvor n_0 ud af de n tilfælde er sande $\mathcal{H}_{0i}$ og n_1 er falske. Lad $R_\mathcal{V}$ være antallet af de forkastede nulhypoteser og $a_\mathcal{V}$ være antallet af dem, som faktisk er sande $\mathcal{H}_{0i}$. Den falske opdagelsesproportion er således givet ved:

$$FDP_\mathcal{V} = \frac{a_\mathcal{V}}{R_\mathcal{V}},$$

hvor $FDP_\mathcal{V} = 0$, hvis $R_\mathcal{V} = 0$.

Sætning 3.7.
Hvis p-værdierne, som svarer til de sande $\mathcal{H}_{0i}$, er uafhængige af hinanden, vil valg-reglen $BH(q)$, baseret på BH algoritmen, kontrollere den forventede falske opdagelsesproportion således

$$E[FDP_{BH(q)}] = \pi_0 q \leq q, \tag{3.11}$$

hvor $\pi_0 = n_0/n$.

Et bevis for Sætning 3.7 findes i [7] på s. 51.
Delen π_0 er typisk ikke kendt, men kan estimeres, og q betegnes typisk for kontrolraten af $BH(q)$. $BH(q)$ er stærkere til at identificere de falske tilfælde af $\mathcal{H}_{0i}$ end FWER kontrollerede algoritmer.
I Sætning 3.7 vil man i tilfældet, hvor der ikke forkastes nogle hypoteser R, være tvunget til at tage $a/R = 0$ for $R = 0$, ved at definere $0/0 = 0$. Dette kan undgås ved Storey's "positiv falske opdagelsesrate" kriterium, hvor vi kun forholder os til situationer, når hypoteser forkastes, således $R > 0$. Denne metode har dog en ulempe, når $n_0 = n$. Dette betyder, at der ikke er nogen sande tilfælde af $\mathcal{H}_{1i}$ hypoteser, og at alle forkastelser vil være falske opdagelser.
Hvor man typisk vælger $\alpha = 0.05$, er det populære valg $q = 0.1$. Ved at se på de empiriske bayesianske termer vil vi i følgende afsnit klargøre betydningen af q.

3.5.4 Bayes fortolkning

Vi ser i det følgende på, hvorledes $BH(q)$ kan fortolkes i forhold til den bayesianske falske opdagelsesrate.
Vi antager, at værdierne p_i svarer til test statistikkerne for z_i (3.2), således at:

$$p_i = F_0(z_i), \tag{3.12}$$

for $i = 1, 2, \ldots, n$, hvor $F_0(z)$ er den kumulative fordelingsfunktion for normalfordelingen.
Lad z_i betegne den i'te værdi i

$$z_1 \leq z_2 \ldots \leq z_i \leq \ldots \leq z_n.$$

Vi har således vores p-værdier givet ved $p_i = F_0(z_i)$ for den venstre hale af den kumulative fordelingsfunktion for z_i, og $p_i = 1 - F_0(z_i)$ for den højre hale.
Ved at betegne $\#\{z_i \leq z\}$ som antallet af $z_i \leq z$ vil andelen af de observerede z_i værdier, som er mindre end z, $\overline{F}(z)$ kunne skrives som

$$\overline{F}(z) = \frac{\#\{z_i \leq z\}}{n}.$$

Ved at tage udgangspunkt i $\overline{F}(z_i)$, fås

$$\overline{F}(z_i) = \frac{i}{n}. \tag{3.13}$$

Vi kan omskrive $BH(q)$ ved hjælp (3.12) og (3.13) til

$$\frac{F_0(z_i)}{\overline{F}(z_i)} \leq q,$$

som er ækvivalent med

$$\frac{\pi_0 F_0(z_i)}{\overline{F}(z_i)} \leq \pi_0 q. \tag{3.14}$$

Vi kan se, at det venstre udtryk i (3.14) er den estimerede bayesianske falske opdagelsesrate $\overline{FDR}(\mathcal{Z})$, (se (3.7)), og vi kan omskrive (3.7) på følgende måde:

$$\overline{FDR}(z_i) \leq q. \tag{3.15}$$

Hvis i_{max} er det største indeks for (3.15), forkaster man nulhypotesen $\mathcal{H}_{0i}$ for alle $i \leq i_{max}$ og accepterer resterende nulhypoteser.
Værdien q er et estimat af Bayes sandsynlighed for, at man forkaster en sand nulhypotese $\mathcal{H}_{0i}$. Ved at ændre q kontrollerer man således FDR. Vælger man $q = 0.1$, betyder det, at ud af de forkastede tilfælde er der mindst 90% sande opdagelser.

Konklusion

I dette projekt har vi arbejdet med de forskellige tilgange inden for GLM og GGLM og beskæftiget os med anvendelsen af hypotesetest og regressionsanalyse for de store datasæt.
Et af de problemer, man møder, når man foretager hypotesetest på de store datasæt, er, at man står tilbage med en stor mængde af forkastede hypoteser. Det er næsten garanteret, at mindst en af de forkastede hypoteser er en sand nulhypotese for den givne problemstilling. For at formindske sandsynligheden for at forkaste en sand nulhypotese har vi kigget på forskellige algoritmer, der kan kontrollere antallet af de forkastede hypoteser. Man kan vælge at kontrollere p-værdier ved at kontrollere FWER eller FDR, og vi har i forbindelse med dette kigget på henholdsvis Holms procedure og Benjamini Hochbergs algoritme. Anvendt på datasættet "*Gene*" viste det sig, at Holms procedure var mindre effektiv, da antallet af forkastede hypoteser blev reduceret fra 890 til 1. BH-algoritmen gav bedre muligheder for at justere kritiske p-værdier. Ved at vælge en kontrolparameter q i $FDR \leq q$ kunne man kontrollere antallet af de forkastede hypoteser. Fastsættelsen $q = 0.05$ resulterede i 5 signifikante gener: Prob21, Prob53, Prob257, Prob2136 og Prob6080. Ønsker man flere parametre for fastsættelse af diagnose, kan man vælge $q = 0.1$ og ende med 24 gener at tage stilling til.
Som alternativ til hypotesetest brugte vi regressionsanalyse. Vi startede med at kigge på MKL, og da metoden ikke var optimal for at arbejde med multivariable sammenhænge, anvendte vi ridge og lasso regression. Ved ridge regression skrympes regressionskoefficienterne β, men man får aldrig dem skrympet til 0, og alle forklarende variabler vil have en betydning for responsen. I tilfældet af lasso regression kommer flere regressionskoefficienter til at være 0, og de tilhørende forklarende variabler ville være usignifikante for responsen.
En af udfordringerne er valget af afgrænsningsradius for det givne optimeringsproblem. Ved at omformulere regulering af radiussen for ℓ_q-normen til Lagrange formen kunne vi regulere styrken af straffen λ frem for afgrænsningsområdet. Anvendelsen af krydsvalidering gav os mulighed for at vælge de optimale λ-værdier, som svarer til den mindste middelkvadratfejl og den største λ indenfor 1 standard afvigelse fra den mindste krydsvalideringsfejl. Større λ-værdier resulterer i mere regulering, mens mindre λ-værdier giver flere forklarende variabler. Vi har bekræftet dette på vores datasæt, hvor lasso regression resulterede i 4 signifikante gener for $\lambda \approx 0.28$: Prob2136, Prob2391, Prob6080 og Prob6185, mens en λ-værdi på $\lambda \approx 0.12$ resulterede i 17 signifikante gener.
Det viste sig også, at der er risiko for, at man ved lasso regression kan miste korrelerede variabler. Elastisk net regression er en kompromis mellem lasso og ridge regression og resulterer i større antal af de forklarende variabler end ren lasso regression. Elastisk net straffen er en vægtet blanding af lasso og ridge straffen, og vi har undersøgt, hvordan den påvirker resultaterne for vores datasæt. For vægten på 0.9 med $\lambda \approx 0.27$ blev der modelleret med 7 signifikante gener, mens vægten på 0.1 med $\lambda \approx 0.54$ resulterede i 294 signifikante gener.
Uden nærmere kendskab til de enkelte problemstillinger indenfor medicinske områder, er det svært at afgøre, om de fundne resultater giver mening for diagnosefastsættelse, og hvor mange variabler man skal fokusere på. Både for hypotesetest og regressionsanalyse kan man ved at

regulere α i elastisk net regression og q i BH algoritmen kontrollere, hvor mange variabler vi ønsker at tage stilling til. Vi vil alligevel bemærke at, både Prob6080 og Prob2136, som vi fik ved begge tilgange, kunne være gener, man skulle kigge nærmere på. Et af de andre problemer, som er forbundet med det givne datamateriale, er, at antallet af testpersoner er betydeligt mindre end antallet af variabler. Dette er ellers typisk i, for eksempel, tilfælde af de sjældne sygdomme. Man kan undersøge mange faktorer hos en enkelt patient, men, hvis sygdommens forekomst er 1 ud af 1.000.000, ville det være tæt på umuligt at samle en repræsentativ testgruppe.
I rammerne af dette projekt kan vi konkludere, at både hypotesetest og regressionsanalyse skal modificeres for at være anvendelig på de store datasæt. Både BH algoritmen for hypotesetest og lasso regressionen frembringer brugbare resultater for at forbedre prædiktions muligheder.
Udover de tilgange, vi har kigget på, findes der også en række andre metoder, der bruges indenfor analyse af store medicinske datasæt, som for eksempel cluster analyse, random forest og det bayesianske netværk [10]. Ved cluster analyse inddeles data i grupper med specifikke egenskaber for at skabe et overblik over, om der er tendenser indenfor de forskellige grupper. Med hensyn til genetisk materiale søger man efter grupper af gener, som har lignende egenskaber, for at kunne finde de gener, der skiller sig ud og dermed har betydning for en patients sygdom. Under random forest udfører man en stor mængde af bootstrap på det givne datasæt og derefter for hvert bootstraped datasæt skaber et klassifikationstræ. Man får heraf en bred variation af træer og tester et givet datasæt ved at køre det igennem de forskellige klassifikationstræer.
Udover at undersøge, om en sygdom er betinget af de forskellige geners udtryksniveau, kunne man også i vores projekt undersøge ved hjælp af det bayesianske netværk, om generne er betingede af hinanden. Det bayesianske netværk er baseret på sandsynlighedsregning, kombineret med grafteoretiske algoritmer. Ved at undersøge relationer mellem generne, ville man have mere viden om de enkelte gener og i sammensætning med for eksempel elastisk net regression, kunne man få en bedre ide om, hvad vægten af lasso regression skulle være. Dermed bliver man mere sikker på, at de fundne gener har betydning for klassifikation af sygdom.

Litteratur

[1] Henrik Madsen og Poul Thyregod. CRC Press, 2011.
Introduction to General and Generalized Linear Models.

[2] Poul Thyregod, 2005.
Noter til Specialkursus i videregående statistik.

[3] Peter Olofsson og Mikael Andersson. Wiley, 2012.
Probability statitcs and stochastic processes.

[4] Trevor Hastie, Robert Tibshirani og Martin Wainwright, 2015.
Statistical learning with sparcity. The lasso and generalizations.

[5] Ryan Tibshirani (with Larry Wasserman), 2015
Sparsity and the Lasso, Statistical Machine Learning.

[6] Hui Zou and Trevor Hastie, Department of Statistics Stanford University, 2005
Regularization and Variable Selection via the Elastic Net

[7] Bradley Efron. Cambridge University Press, 2010.
Large - Scale Inference. Emperical Bayes Methods for Estimation, Testing, and Prediction.

[8] John D. Storey. Ann. Statist. Volume 31, no. 6, 2003
The positive false discovery rate: A Bayesian Interpretation and the q-value.

[9] Megan Goldman, 2008.
Why is multiple testing a problem? - Bioeng C141 - Statistics for Bioinformatics.

[10] Changwon Yoo, Luis Ramirez, and Juan Liuzzi, 2014
Big Data Analysis Using Modern Statistical and Machine Learning Methods in Medicine

[11] Stephanie Glen, 2015
Benjamini-Hochberg Procedure
`https://www.statisticshowto.datasciencecentral.com/benjamini-hochberg-procedure/`.

[12] Stephanie Glen, 2016
Holm-Bonferroni Method: Step by Step
`https://www.statisticshowto.datasciencecentral.com/holm-bonferroni-method/`.

[13] Trevor Hastie and Junyang Qian, 2014
Glmnet Vignette
`https://www.statisticshowto.datasciencecentral.com/`

holm-bonferroni-methodhttps://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html/.

[14] Marvin L. Bittinger, David J. Ellenbogen and Scott A. Surgent. Addison Wesley, 2012 *Calculus and its applications, 10th edition.*

[15] Cedric E. Ginestet, Boston University. 2013. *Regularization: Ridge Regression and Lasso.*

Appendiks A

A.1 Lagrange multiplier

Dette afsnit er baseret på [14].
Lad $f(x, y)$ være afgrænset af en multivariabel funktion $g(x, y) = 0$. Vi kan minimere/maksimere $f(x, y)$ ved hjælp af Lagrange multiplier metoden:

1. Vi indfører en ny parameter λ og definerer en ny funktion $\mathcal{L}$ som:

   $$\mathcal{L}(x, y, \lambda) = f(x, y) - \lambda g(x, y)$$

   funktionen $\mathcal{L}$ kaldes Lagrange funktionen, mens parameteren λ refereres til som Lagrange multiplier.

2. Vi finder de kritiske punkter af $\mathcal{L}$ ved at løse ligningen:

   $$\nabla \mathcal{L}(x, y, \lambda) = \mathbf{0} \tag{A.1}$$

3. Lad (a, b, λ) repræsentere løsningen til (A.1). Vi kan nu afgøre, om (a, b) giver et maksimum eller minimum for $f(x, y)$.

Hvor ovenstående viser Lagrange multiplier metoden for to variabler, kan metoden udvides til funktioner af tre eller flere variabler.

A.2 Konveksitet

Dette afsnit er baseret på [15].
Lasso og ridge regression er begge konvekse optimeringsproblemer. Vi starter med at definere, hvad et konvekst optimeringsproblem er, og derefter forklarer, hvorfor lasso regressionen er et konvekst problem, mens ridge er et strengt konvekst problem.

Definition A.1. Konveks funktion
Lad $f : I \to R$ være en funktion, hvor $I \subseteq \mathbb{R}$. Hvis der for hvert $x_1, x_2 \in I$ og for hvert t, hvor $0 \leq t \leq 1$, gælder, at

$$f((1-t)x_1 + tx_2) \leq (t-1)f(x_1) + tf(x_2), \tag{A.2}$$

er funktionen f konveks i I.
Funktionen f er strengt konveks på I, hvis der for hvert $x_1 \neq x_2 \in I$ og hvert t, hvor $0 < t < 1$, gælder, at

$$f((1-t)x_1 + tx_2) < (t-1)f(x_1) + tf(x_2). \tag{A.3}$$

Strengt konvekse funktioner har ét globalt minimum. Ikke-konvekse funktioner kan omvendt have flere lokale minimummer, hvilket betyder, at der er flere punkter, som opfylder, at de er bedst i et område, men ikke globalt optimale. Hvis man har et ikke-konvekst problem, kan man ikke teste om en løsning, man har fundet, også er den bedste.

Lemma A.2.
Lad f være en differentiabel funktion på et interval I. Funktionen f er (strengt) konveks på I, hvis dens dobbelte afledte funktion er (strengt) voksende på I.

Både lasso og ridge regression er konvekse optimeringsproblemer, og dette vises ved differentiering af ledene i optimeringsudtrykket (2.8), hvor $q = 1$ svarer til lasso regression, og $q = 2$ svarer til ridge regression.
Vi ser først på udtrykket for MKL, som er et fælles led for både lasso og ridge regression.
Vi kan udtrykke MKL ved brug af summen af de kvadrerede risidualer (se (2.2)) således:

$$||\boldsymbol{\epsilon}||^2 = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}).$$

Af Lemma A.2 ved vi, at en funktion er strengt konveks, hvis dens dobbelte afledte er strengt voksende. Dette er tilfældet for MKL delen, da der gælder, at $\boldsymbol{X} > \boldsymbol{0}$ og

$$\frac{\delta^2}{\delta \boldsymbol{\beta}^2}||\boldsymbol{\epsilon}||^2 = 2(\boldsymbol{X}^T\boldsymbol{X}) > \boldsymbol{0}.$$

Ligeledes ses der for ℓ_2 normen i ridge regression, at

$$\lambda||\boldsymbol{\beta}||_2 = \lambda\boldsymbol{\beta}^T\boldsymbol{\beta}$$

er strengt konveks, da

$$\frac{\delta^2 \lambda \boldsymbol{\beta}^T \boldsymbol{\beta}}{\delta \boldsymbol{\beta}} = 2\lambda > \mathbf{0},$$

og for ℓ_1 normen i lasso regression, at

$$\lambda ||\boldsymbol{\beta}||_1 = \lambda |\boldsymbol{\beta}|$$

er konveks, da

$$\frac{\delta^2 \lambda ||\boldsymbol{\beta}||_1}{\delta \boldsymbol{\beta}} = \mathbf{0} \geq \mathbf{0}.$$

Da begge led i ridge regressionen er strengt konvekse, betyder det, at ridge er et strengt konvekst optimeringsproblem, mens lasso er et konvekst optimeringsproblem, da den består af en konveks og en streng konveks del.

Sætning A.3.
Løsningen til en ridge regression er entydig, og er givet ved:

$$\hat{\boldsymbol{\beta}}_{ridge} = (\boldsymbol{X}^T \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^T \boldsymbol{Y}.$$

Bevis
Vi kan omskrive (2.5), til:

$$||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_2 = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda (\boldsymbol{\beta}^T \boldsymbol{\beta}),$$

og finder $\hat{\boldsymbol{\beta}}_{ridge}$ som løsning af ligning

$$\frac{\delta}{\delta \boldsymbol{\beta}} (||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_2) = \mathbf{0}.$$

Dette giver at

$$\frac{\delta}{\delta \boldsymbol{\beta}} (||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_2) =$$

$$= \frac{\delta}{\delta \boldsymbol{\beta}} \Big((\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^T \boldsymbol{\beta} \Big) = \mathbf{0}.$$

$$\Updownarrow$$

$$2(\boldsymbol{X}^T \boldsymbol{X})\boldsymbol{\beta} - 2\boldsymbol{X}^T \boldsymbol{Y} + 2\lambda \boldsymbol{\beta} = \mathbf{0}$$

$$\Updownarrow$$

$$2(\boldsymbol{X^T X})\beta + 2\lambda\boldsymbol{\beta} = 2\boldsymbol{X^T Y}$$

$$\Updownarrow$$

$$(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})\boldsymbol{\beta} = \boldsymbol{X}^T\boldsymbol{Y}$$

$$\Updownarrow$$

$$\hat{\boldsymbol{\beta}}_{ridge} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

□

Appendiks B

B.1 Datasættet

	Prob1	Prob2	Prob3	Prob4	Prob5	Prob6	Prob7	Prob8
Subject1	9.172434	0.8282310	0.5484337	11.703450	0.6265843	10.504830	0.6805805	7.473903
Subject2	9.841630	0.8118601	0.6519188	12.171130	0.6283513	10.862100	0.6161459	7.984346
Subject3	8.526931	0.7115715	0.6477221	11.394790	0.6535096	10.236100	0.6385523	6.245332
Subject4	9.790411	0.7286916	0.5401568	11.778870	0.5585985	10.864430	0.5833724	7.725852
Subject5	10.439280	0.7884122	0.5197105	12.867760	0.7934539	11.479950	0.5222204	8.527100
Subject6	9.631492	0.7663501	0.6294500	12.095660	1.3623930	10.794470	0.6998680	7.770390
Subject7	9.694419	0.8603156	0.6852528	11.938610	0.6636261	10.611370	0.7312123	7.306466
Subject8	9.542788	0.7601823	0.6746542	11.592320	0.7349983	10.678570	0.5948305	7.179664
Subject9	10.066970	0.7350588	0.5641429	11.812510	0.4821154	10.568110	0.6375680	8.177573
Subject10	10.772420	0.7598031	0.4895325	12.128380	0.6350042	11.105750	0.6334863	8.491671

Figur B.1: Udklip af datasættet "*Gene*" .

Til analyse bruges datasættet "*Gene*", som indeholder målinger for patienter med en hudsygdom og en kontrolgruppe. Målingerne er udtryksniveauet for $n = 6658$ forskellige gener hos $N = 30$ personer, heraf er 15 raske testpersoner og 15 testpersoner med hudsygdom. Datasættet er dermed en 30×6658 matrix $\boldsymbol{X}$, med, elementet

$$x_{ij} = \text{gen } j's \text{ udtryksniveau for testperson } i,$$

hvor

$$i = 1, 2, \ldots, N, \ j = 1, 2, \ldots, n;$$

således, at $i = 1, 2, \ldots, 15$ for de raske testpersoner, mens $i = 16, 17, \ldots, 30$ for testpersonerne med hudsygdomme.
Responsvariablen $\boldsymbol{Y}$ er en 1×30 vektor med
$y_i = 1$, når $i = 1, \ldots, 15$ for de raske testpersoner og
$y_i = 0$, når $i = 16, \ldots, 30$ for de syge personer.
Genet i refereres til som Prob$i$.
Datasættet "*MindreData*" er et udsnit af datasættet "*Gene*" bestående af 8 tilfældig valgte gener: Prob257, Prob634, Prob1949, Prob2391, Prob3272, Prob4829, Prob5344 og Prob5623.

B.2 Lineær regression

Vi anvender lineær regression for at finde sammenhængen mellem de korrelerede gener. Vi vælger Prob168 som afhængig variabel og Prob40, som uafhængig variabel i:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}.$$

Følgende kommandoer i R bruges til lineær regression:

```
data_x=data.frame(Prob168, Prob40)

regmodel=lm(Prob168~Prob40, data=data_x)

plot(Prob40, Prob168)

abline(regmodel)
```

Ud fra kørsel har vi fået $\beta_0 = 2.27$ og $\beta_1 = 0.9581$

```
Call:
lm(formula = Prob168 ~ Prob40, data = data_x)

Residuals:
     Min       1Q   Median       3Q      Max
-2.01242 -0.24614  0.01481  0.25721  1.31104

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.2700     0.3167   7.167 8.45e-08 ***
Prob40        0.9581     0.0792  12.096 1.23e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5672 on 28 degrees of freedom
Multiple R-squared:  0.8394,    Adjusted R-squared:  0.8336
F-statistic: 146.3 on 1 and 28 DF,  p-value: 1.233e-12
```

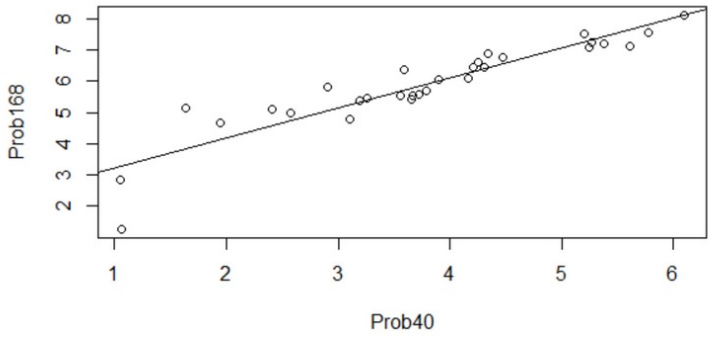
Sammenhængen mellem værdierne for de valgte gener, samt regressionslinjen, kan ses på Figur B.2.
En anden anvendelse af simpel lineær regression i forhold til vores datasæt er at sammenligne de gennemsnitlige udtryksniveauer for de forskellige gener for de syge og de raske testpersoner. Den lineære regressionsmodel er i dette tilfælde:

$$\boldsymbol{\mu}(2) = \boldsymbol{\mu}(1)\boldsymbol{\beta},$$

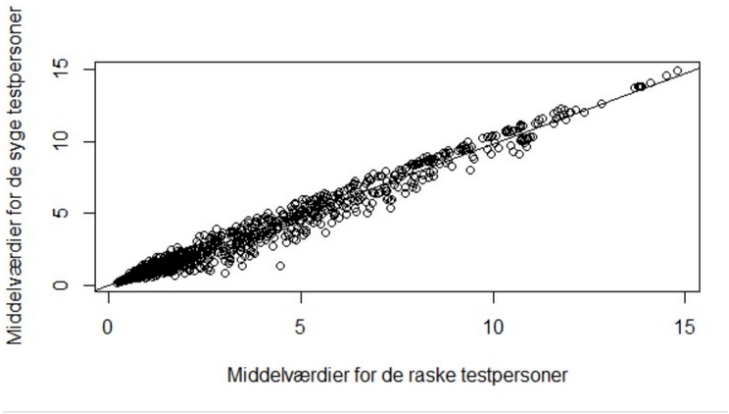
hvor $\boldsymbol{\mu}(1)$ og $\boldsymbol{\mu}(2)$ er vektorer, som indeholder middelværdier for udtryksniveauer for de forskellige gener, forholdsvis for de raske og de syge testpersoner.
Vi har fået følgende resultat:

```
Call:
lm(formula = syg_mean ~ rask_mean, data = data_x)

Residuals:
     Min       1Q   Median       3Q      Max
-3.01154 -0.02194  0.00212  0.02814  1.14492

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.021314   0.003035   7.023 2.39e-12 ***
rask_mean   0.976273   0.001528 638.983  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2008 on 6656 degrees of freedom
Multiple R-squared:  0.984,     Adjusted R-squared:  0.984
F-statistic: 4.083e+05 on 1 and 6656 DF,  p-value: < 2.2e-16
```

Interceptleddet $\beta_0 = 0.02$ er tæt på 0, og hældningen $\beta_1 = 0.976$ er tæt på 1, hvilket peger i retning af, at, hvis vi kigger på alle gener samtidigt, er det svært, nærmest umuligt, at se nogle forskelle mellem de syge og de raske testpersoner. På Figur B.3 ses sammenhængen mellem middelværdier for de raske og de syge testpersoner.



Figur B.2: XY-plot af udtryksniveauer for generne Prob168 og Prob40, samt regressionslinje.



Figur B.3: XY-plot af middelværdierne for genernes udtryksniveauer for testpersoner med og uden sygdom, samt regressionslinje.

B.3 ANOVA

For at demonstrere anvendelsen af ANOVA har vi brugt 4 udvalgte gener: Prob257, Prob634, Prob1949, Prob2391 og testet nulhypotesen:

$$\mathcal{H}_0 : \mu_{257} = \mu_{634} = \mu_{1949} = \mu_{2391}.$$

I R udføres ovenstående ved hjælp af følgende kommandoer:

```
x_fire=c(mindreData$Prob257, mindreData$Prob634, mindreData$Prob1949, mindreData$Prob2391)
group = factor(rep(letters[1:4], each = 30))
anova_model=lm(x_fire~group)
print(anova(anova_model))
```

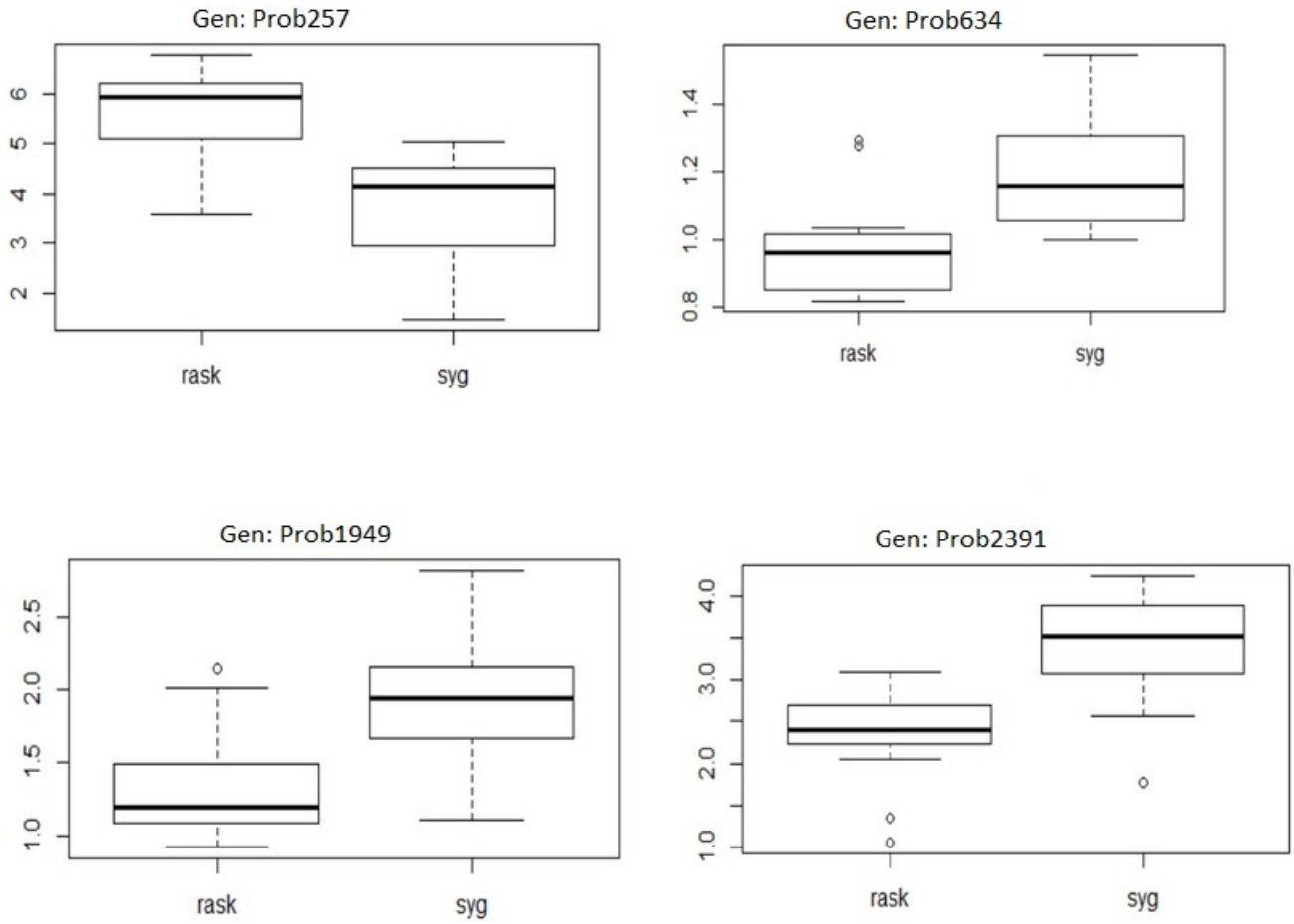
Det er ikke overraskende, at nulhypotesen forkastes

```
Analysis of Variance Table

Response: x_fire
           Df  Sum Sq Mean Sq F value    Pr(>F)
group       3 225.198  75.066  98.477 < 2.2e-16 ***
Residuals 116  88.423   0.762
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ved hjælp af multivariabel ANOVA er vi ikke i stand til at afgøre, hvilke variabler skiller sig ud, men kun forkaste nulhypotesen, hvis der mindst er to variabler med forskellige middelværdier. I forhold til vores datasæt giver det mening at sammenligne generne for de syge testpersoner med tilsvarende gener for de raske testpersoner parvis. I bilag B.8 findes der en mindre kode, som blev brugt til analysen:

```
res=an_tt_zc(x_rask, x_syg, number_gen)
```

En kørsel på datasættet "*Mindredata*" resulterede i, at 4 nulhypoteser ud af 8 testede blev forkastet med en signifikans niveau $\alpha = 0.05$. For generne: Prob257, Prob634, Prob1949 og Prob2391 er der dermed tale om signifikante forskelle i middelværdier for de syge og de raske testpersoner (se Figur B.4).



Figur B.4: Boksplot af udtryksniveauet for gener for de syge og raske testpersoner. De 4 gener er: Prob257, Prob634, Prob1949 og Prob2391.

For datasættet "*Gene*" var resultatet 890 forkastede nulhypoteser.

B.4 Lasso og ridge regression

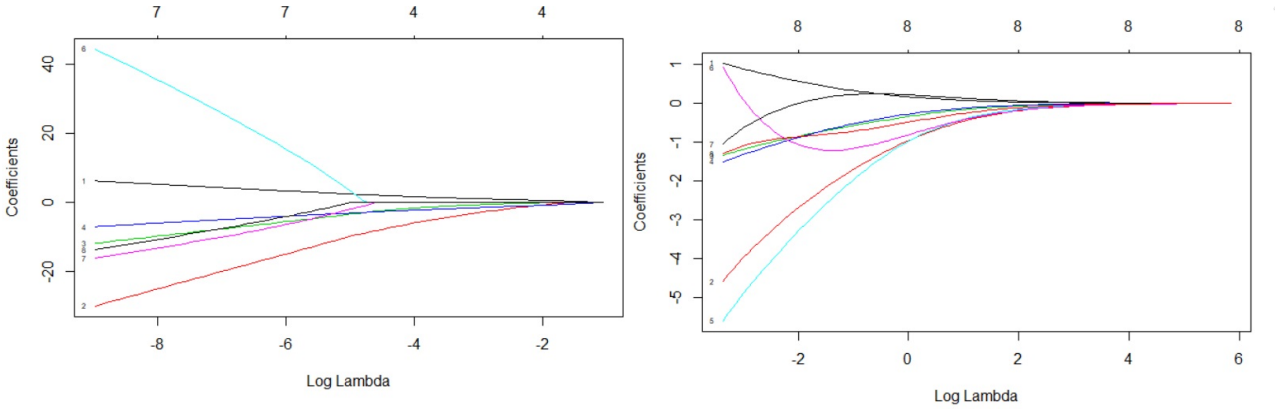
Lasso og ridge regression i R ved hjælp af følgende kommandoer:

```
model1<-glmnet(x, y, alpha=0, family = "binomial")
model2<-glmnet(x, y, alpha=1, family = "binomial")
```

Resultat af regressionen er blandt andet matricerne af regressionskoefficienterne β for de forskellige λ værdier, som kan præsenteret i form af stier ved hjælp af:

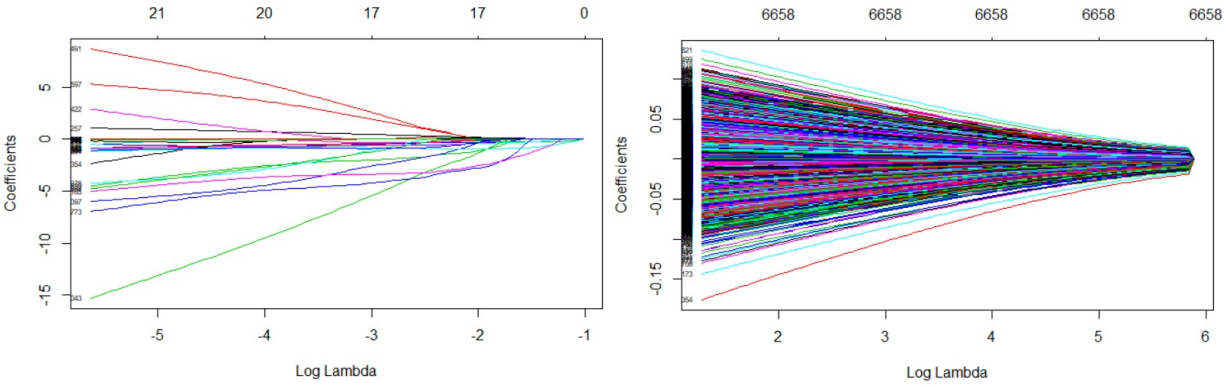
```
plot(model1,xvar = "lambda", label = TRUE)
```

På Figur B.5 ses lasso stien for datasættet "*MindreData*" til venstre og ridge stien til højre.



Figur B.5: Lasso stien og ridge stien for datasættet "*MindreData*".

For det store datasæt ser det mindre overskueligt ud:



For det mindre datasæt har vi udført krydsvalidering med lasso regression og har beregnet λ_{min} som, svarer til den mindste middelkvadratfejl, og $\lambda_{1.se}$, indenfor 1 standard afvigelse fra den mindste krydsvalideringsfejl:

```
modelcv<-cv.glmnet(x, y, alpha=1, family = "binomial",type.measure =  "mse")

print(modelcv$lambda.min)
print(coef(modelcv, s =modelcv$lambda.min))

print(modelcv$lambda.1se)
print(coef(modelcv, s =modelcv$lambda.1se))
```

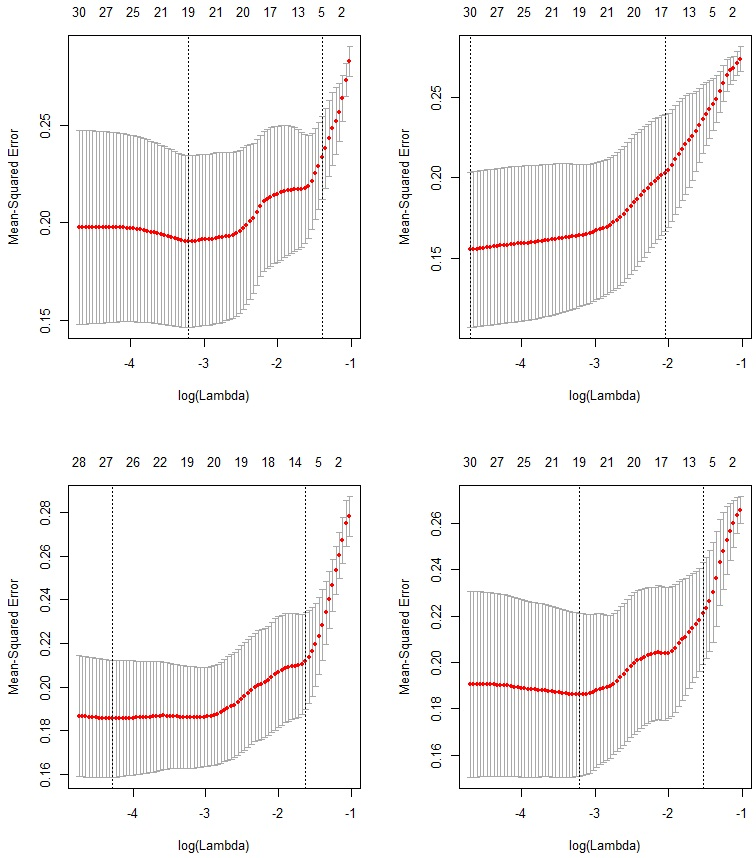
Begge valg af λ resulterer i de samme signifikante variabler, mens selve modellerne ser forskellige ud.
For λ_{min}:

```
[1] 0.03370447
9 x 1 sparse Matrix of class "dgCMatrix"
                    1
(Intercept)  5.0312399
Prob257      1.2879377
Prob634     -4.0467213
Prob1949    -0.9766317
Prob2391    -1.7860080
Prob4829     .
Prob3272     .
Prob5344     .
Prob5623     .
```

og for $\lambda_{1.se}$:

```
[1] 0.0937847
9 x 1 sparse Matrix of class "dgCMatrix"
                    1
(Intercept)  1.7187338
Prob257      0.7483454
Prob634     -1.4858206
Prob1949    -0.2923020
Prob2391    -1.0896180
Prob4829     .
Prob3272     .
Prob5344     .
Prob5623     .
```

Krydsvalideringerne resulterer i de forskellige λ værdier, som det kan ses på Figur B.6 nedenfor:



Figur B.6: Krydsvalideringer af datasættet "*Gene*".

For datasættet "*Gene*" har vi udført et antal krydsvalideringer og har taget gennemsnittet af λ_{min} og $\lambda_{1.se}$, for at se, om λ_{min} og/eller $\lambda_{1.se}$ bliver stabiliseret.
Vi har undersøgt, hvordan antallet af krydsvalideringer på datasættet "*Gene*" påvirker λ_{min} og $\lambda_{1.se}$ ved at foretage 10, 100, 500, 1000 og 2000 krydsvalideringer, hver udført tre gange.
Det viste sig, at for datasættet "*Gene*" ville omkring 2000 krydsvalideringer være et passende antal for, at $\lambda_{min} \approx 0.12$ og $\lambda_{1.se} \approx 0.28$ stabiliserer sig. Nedenfor er der givet et eksempel på, hvordan resultaterne ser ud efter 2000 krydsvalideringer.

Første kørsel:

```
"Alpha er"
1
"Antal krydsvalederinger"
2000
"lambda_min"
0.1190565
"Signifikante gener og tilsvarende beta-verdier for lambda_min "
.]     0  257  597  630  634 1949 2136 2391 2502 4043 4097 5193 5491 5773 6080 6185 6300 6566
.] 12.30139189  0.21004109  0.28398334 -1.26455114 -0.35140316 -0.18206091 -0.06376984 -0.4532
i5
)] -0.05634957 -1.85459907 -0.80051432 -0.45754560  0.28426237 -2.90471912 -0.94510386 -2.7559
i2
']  0.12662557 -0.48373622
"Antal signifikante gener"
17
"lambda_1se"
0.2771294
"Signifikante gener og tilsvarende beta-verdier for lambda_1SE "
   0 2136 2391 6080 6185
 3.75294895 -0.22539860 -0.07372885 -0.52093913 -0.36265406
"Antal signifikante gener"
4
```

Anden kørsel:

```
] "Alpha er"
] 1
] "Antal krydsvalederinger"
] 2000
] "lambda_min"
] 0.1197635
] "Signifikante gener og tilsvarende beta-verdier for lambda_min "
L]     0  257  597  630  634 1949 2136 2391 2502 4043 4097 5193 5491 5773 6080 6185 6300 6566
L] 12.27435447  0.20821957  0.27024243 -1.25991019 -0.35214763 -0.17985536 -0.06559267 -0.4519
99
9] -0.04994951 -1.83204656 -0.78002900 -0.45291427  0.27159508 -2.89638930 -0.94555649 -2.7479
91
7]  0.12659113 -0.47675905
] "Antal signifikante gener"
] 17
] "lambda_1se"
] 0.2785698
] "Signifikante gener og tilsvarende beta-verdier for lambda_1SE "
]    0 2136 2391 6080 6185
]  3.67660329 -0.22393413 -0.07012209 -0.51187681 -0.33551347
] "Antal signifikante gener"
] 4
```

B.5 Elastisk net regression

Elastisk net regression er en kompromis mellem lasso og ridge regression. Vi har udført den på datasættet "*Gene*" for de forskellige α værdier i (2.12).
Resultaterne er samlet i Tabel B.1.

α	λ_{min}	Antal signifikante variabler λ_{min}	$\lambda_{1.se}$	Antal signifikante variabler $\lambda_{1.se}$
0	≈ 4.10	6658	≈ 20.16	6658
0.1	≈ 0.045	501	≈ 0.54	294
0.3	≈ 0.020	195	≈ 0.29	84
0.5	≈ 0.018	96	≈ 0.22	45
0.7	≈ 0.028	55	≈ 0.23	24
0.9	≈ 0.078	26	≈ 0.27	7
1	≈ 0.12	17	≈ 0.28	4

Tabel B.1: Antal forklarende variabler efter anvendelse af elastisk net regression på datasættet "*Gene*".

Vi kan se, at man for $\alpha = 0.5$ kommer til at have 41 ekstra variabler, sammenlignet med lasso regression for $\lambda_{1.se}$. Med udgangspunkt i λ_{min} resulterede elastisk net regression i 79 ekstra variabler. I forhold til lasso tager elastisk net forbehold til de korrelerede variabler og dermed resulterer i, at flere grupper af variabler er indbyrdes korrelerede.
Vi har undersøgt, hvilke gener korrelerer med hinanden for $\alpha = 0.5$, ved at beregne Pearson korrelationskoefficient for de 45 signifikante variabler, og vi betragtede variabler til at være korrelerede, når korrelationskoefficienten $r \geq 0.8$. (se Appendiks B.8).

```
indeks_matrix=c(21,40,53,257,290,302,597,630,634,1121, 1542, 1550, 1713, 1949, 2136,
                2146, 2223, 2225, 2242, 2269, 2391, 2502, 2635, 3173, 3417, 3986,
                4043, 4097, 4236, 4978, 5193, 5354, 5491, 5684, 5773, 5911, 6018, 6080,
                6117, 6185, 6212, 6300, 6542, 6555, 6566)

INTER_CORRELATION(x, 1, indeks_matrix)
```

Resultat viste sig at give én gruppe på 5 korrelerede gener: Prob21, Prob40, Prob53, Prob290, Prob1542 og to grupper på to gener: Prob2269 og Prob2391, samt Prob6542 og 6555
Det viste sig også, at den sidste gruppe består af to identiske gener Prob6542 og Prob6555, og er sandsynligvis resultat af fejlindtastning.

B.6 T-test

T-fordeling med n frihedsgrader har følgende tæthedsfunktion:

$$f_t(y) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{\pi n}\Gamma(\frac{n}{2})}\left(1+\frac{y^2}{n}\right)^{-\frac{n+1}{2}},$$

hvor n er antallet af frihedsgrader: $n = N-2 = 28$ for vores datasæt, og $\Gamma(n)$ er Eulers gamma funktion.
For lige n gælder der:

$$\frac{\Gamma(\frac{n+1}{2})}{\sqrt{\pi n}\Gamma(\frac{n}{2})} = \frac{(n-1)(n-3)\cdots 5\cdot 3}{2\sqrt{n}(n-2)(n-4)\cdots 4\cdot 2}.$$

Da vi har $y^2 < n$, kan kumulativ fordelingsfunktion skrives som:

$$F(t) = \int_{-\infty}^{t} f(u)du = \frac{1}{2} + t\frac{\Gamma(\frac{1}{2}(n+1))}{\sqrt{\pi n}\Gamma(\frac{n}{2})} F *_1 (\frac{1}{2}, \frac{1}{2}(n+1); \frac{3}{2}; -\frac{t^2}{n}),$$

hvor

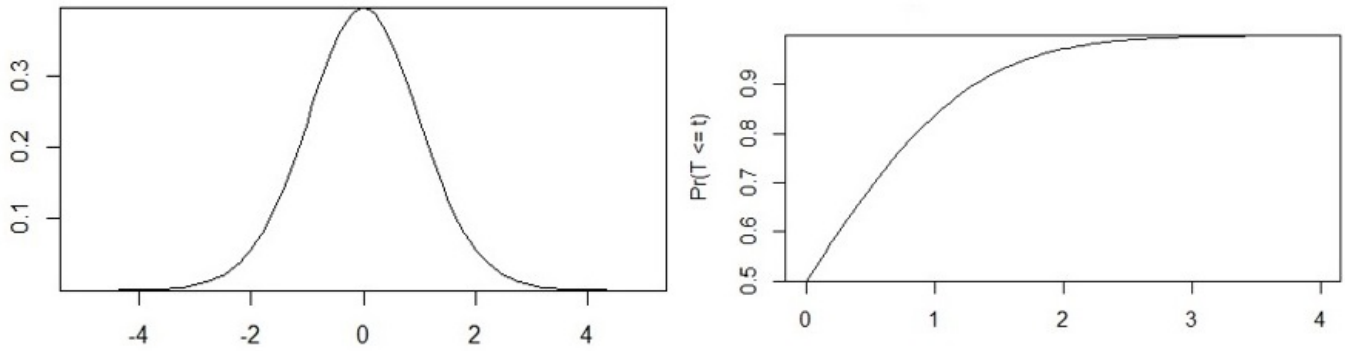
$$F *_1 (a,b;c;z) \equiv F_1(a,b;c;z) = 1 + \frac{ab}{c}\frac{z}{1!} + \frac{a(a+1)b(b+1)}{c(c+1)}\frac{z^2}{2!} + \cdots$$

er hypergeometrisk række.
Vi har fremstillet graferne for $f(t)$ og $F(t)$ (Figur B.7), samt beregnet den kritiske t-værdi for $\alpha = 0.05$ ved hjælp af følgende linjer i R:

```
plot(function(x) dt(x, df = 28), -5, 5,
     main = "Tæthedsfunktion for t-fordeling", yaxs = "i")

plot(function(x) pt(x, df = 28), 0, 4,
     main = "Cummulativ tæthedsfunktion for t-fordeling med
     28 frighedsgrader", yaxs = "i", xlab = "t", ylab="Pr(T <= t)")
```

```
print("KRITISK VÆRDI FOR alpha=0,05:")
print(qt(0.975, 28))
```

```
[1] "KRITISK VÆRDI FOR alpha=0,05:"
[1] 2.048407
```



Figur B.7: Tæthedsfunktion og kumulativ fordelingsfunktion for t-fordelingen med 28 frihedsgrader..

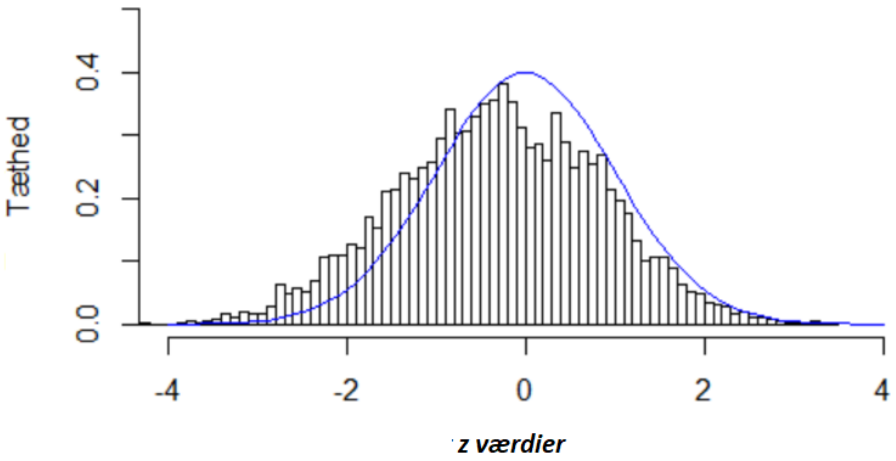
Beregning af z-værdier ifølge (3.2) udføres i R ved hjælp af følgende kommandoer:

```
ttest=t.test(x1, x2, paired = TRUE)
z_score=qnorm(ttest$p.value, mean=0, sd=1, log.p=FALSE)
```

For datasæt "Gene" blev t-statistik omformuleret til z-statistik (se Appendiks B.8):

```
res=an_tt_zc(x_rask, x_syg, number_gen)
```

og histogrammet af z-værdier sammen med normalfordelingskurve kan ses på Figur B.8



Figur B.8: Histogrammet for z-værdier for 6658 gener samt normalfordelingskurven $N(0,1)$.

Som man kan se på ovenstående histogram, følger z-fordeling ikke helt normalfordelingskurven med $\mu = 0$ og $\sigma^2 = 1$, hvilket egentlig er et godt resultat, da vi i dette tilfælde forkaster nulhypotesen, at der ikke er nogle signifikante gener.

B.7 Holms procedure og Benjamini Hochbergs algoritme

Anvendelsen af Holms procedure (se Appendiks B.8) på det store datasæt medførte, at der kun blev forkastet en enkelt $\mathcal{H}_{0i}$ hypotese, som svarer til gen Prob6080. På datasættet "*MindreData*" blev antallet af af forkastede nulhypoteser ikke ændret yderligere sammenlignet med ANOVA. Selvom Holms procedure burde give mere frihed til at forkaste hypoteserne end Bonferroni korrektionen, er det ikke tilfældet for vores datasæt "*Gene*". For hvert skridt burde de sorterede p-værdier være mindre eller lig med $\frac{\alpha}{n-i+1}$. De kritiske p-værdier for de første 3 skridt i algoritmen for datasættet "*Gene*" er samlet i Tabel B.2, sammen med Bonferroni korrektionen.

i	Holms kriterium $\frac{\alpha}{n-i+1}$	Bonferronis kriterium $\frac{\alpha}{n}$
1	$7.510 \cdot 10^{-6}$	$7.510 \cdot 10^{-6}$
2	$7.511 \cdot 10^{-6}$	$7.510 \cdot 10^{-6}$
3	$7.512 \cdot 10^{-6}$	$7.510 \cdot 10^{-6}$

Tabel B.2: Kritiske p-værdier for Holms procedure og Bonferroni korrektion for datasættet "*Gene*" .

Som man kan se, er der for de første skridt i Holms procedure ikke rigtigt forskel mellem Holms justering og Bonferroni korrektionen for $n = 6658$.
Anvendelsen af Benjamini Hochbergs algoritme (se Appendiks B.8) med $q = 0.05$ på datasættet "*Gene*" gav en reduktion fra 890 nulhypoteser til 5:

```
"Antal forkastede Nulhypoteser efter BH"
5
"tilsvarende variabler er PROB...."
6080
2136
53
257
21
```

Dette svarer til Prob21, Prob53, Prob257, Prob2136 og Prob6080.
Vi har kørt algoritmen med de forskellige q-værdier og samlet resultaterne i Tabel B.3.

q	0.01	0.05	0.1	0.2
Antal forkastede hypoteser	1	5	24	270

Tabel B.3: Sammenhængen mellem q-værdier og antallet af de forkastede nulhypoteser efter anvendelse af BH algoritmen på datasættet "*Gene*" .

B.8 Diverse mindre koder, som blev brugt under dataanalyse

```
#Løkken til at udføre hypotesetest for et antal hypoteser
#number_gen er antal hypoteser
#x og y er matriserne med data for de raske og de syge testpersoner

an_tt_zc<-function(x, y, number_gen){

for(j in 1:number_gen){

  for(i in 1:15){
    one_x[i]=x[i,j]   #samler lige dataer for hver gen for sig, for at beholder overblik
    one_y[i]=y[i,j]

  }
  ttest=t.test(one_x, one_y, paired=TRUE)
  t_value[j]=ttest$statistic
  p_value[j]=ttest$p.value
  z_score[j]=qnorm(p_value[j], mean=0, sd=1, log.p=FALSE)
  group=c(one_x, one_y)
  data = data.frame(y = group, group = factor(group))
  fit = lm(y ~ group, data)
  anova_res=anova(fit)
  p=anova_res$`Pr(>F)`
  p_value_an[j]=as.vector(p)[1]
  if(p_value_an[j]<0.05) {
    counter_forkast=counter_forkast+1
  }


}

print("Antal forkastede Nulhypoteser")
print(counter_forkast)
}
```

```
#Procedure for krydsvalederinger
#x er matricen med forklarende variabler
#antal_cv er antal krydsvalederinger
#alpha er i intervallet [0,1], hvor 0 svarer til ridge, 1 -til lasso
#middelværdier for lambdamin og lambda1se, sammen med tilsvarende beta
#og forklarende variabler udskrives.

kryds_valid<-function(x, antal_cv, alpha){
n=rep(15,2)
y=rep(1:0, n)
lambda_min=c(length(antal_cv))
lambda_se=c(length(antal_cv))
for(i in 1:antal_cv) {
  x2=cv.glmnet(x,y, alpha=alpha,nfolds=5, family="binomial")
  lambda_min[i]=x2$lambda.min
  lambda_se[i]=x2$lambda.1se
}
```

```
lambda1=mean(lambda_min)
lambda2=mean(lambda_se)
model<-glmnet(x,y, alpha=alpha, family="binomial", lambda=lambda1)
coef_min1=coef(model)
index_coef=which(coef_min1!=0)
print("Alpha er")
print(alpha)
print("Antal krydsvalederinger")
print(antal_cv)
print("lambda_min")
print(lambda1)
print("Signifikante gener og tilsvarende beta-verdier for lambda_min ")
print(index_coef-1)
print(coef_min1[index_coef])
print("Antal signifikante gener")
print(length(index_coef)-1)
model<-glmnet(x,y, alpha=alpha, family="binomial", lambda=lambda2)
coef_min1=coef(model)
index_coef=which(coef_min1!=0)
print("lambda_1se")
print(lambda2)
print("Signifikante gener og tilsvarende beta-verdier for lambda_1SE ")
print(index_coef-1)
print(coef_min1[index_coef])
print("Antal signifikante gener")
print(length(index_coef)-1)
```

```
#Beregning af de korrelerede variabler.'
#x er matricen med datasæt; start_index=0,
#indeks_matrix indeholder indeks for de variabler i x, som ønskes
#at teste korrelation for
#Resulat udprintes

INTER_CORRELATION<-function(x, start_index, indeks_matrix){

  antal_gen=length(indeks_matrix)
  one_x=vector(length = 15)
  to_x=vector(length=15)
  if (start_index==antal_gen) {
    return(1)
  }
  else {
      for(i in 1:15) one_x[i]=x[i,indeks_matrix[start_index]]

      print('gen')
      print(indeks_matrix[start_index])

      for(j in (start_index+1):antal_gen){

                for(i in 1:15) to_x[i]=x[i,indeks_matrix[j]]
                if(cor(one_x, to_x)>0.8){
                    print("korrelerer med")
                    print(indeks_matrix[j])
                }

      }
      INTER_CORRELATION(x, start_index+1, indeks_matrix)
   }

}
```

```
#HOLM procedure.
#alpha er signifikans for den enkelte H0
#p_value er vektoer af p værdier for N nulhypoteser

holm_procedure<-function(alpha, p_value){

N=length(p_value)
counter_forkast=0
sort_p=sort(p_value, index.return=TRUE)
p_value_sort=sort_p$x
indeks_p=sort_p$ix

while (i <=N) {

  if(p_value_sort[i]<alpha/(number_gen-i+1)) {

      j=indeks_p[i]
      i=i+1
      counter_forkast+1
    }
  else {i=N+1
  print("Antal forkastede Nulhypoteser efter Holm")
  print(counter_forkast)
  for(j in 1:counter_forkast) print(indeks_p[j])

  }
}

}
```

```
#for(i in 1:imaks) print(indeks_p[i])
#BH algoritme
#q er intervallet [0,1]
#p_value er vektoer af p værdier for n nulhypoteser

BH_algoritme<-function(q, p_value){

N=length(p_value)
sort_p=sort(p_value, index.return=TRUE)
p_value_sort=sort_p$x
indeks_p=sort_p$ix
imaks=0
for (i in 1:N) {

  if(p_value_sort[i]<=q*i/N)
  {
    j=indeks_p[i]
    imaks=i
  }
  else {
      i=N+1
      print("Antal forkastede Nulhypoteser efter BH")
      print(imaks)
      for(j in 1:imaks) print(indeks_p[j])
  }
}
}
```