COPENHAGEN BUSINESS SCHOOL
HANDELSHØJSKOLEN

Kursusevalueringer på CBS

Hvilke studerende besvarer og hvordan svarer de?

Julie Kargaard Lindegaard (102575)

Sofie Nicoline Rostgaard Petersen (101134)

***

Kandidatuddannelsen i Erhvervsøkonomi og Matematik

Vejleder: Dorte Kronborg

ECTS: 30

Afleveringsdato: d. 17. maj 2021

194.224 anslag / 120 (133) sider

Abstract

This paper addresses the student evaluations at Copenhagen Business School (CBS) from the academic year 2018/2019. Our research question arises from our curiosity of why we as students are asked to evaluate every course we attend.

As an education is a societal investment, we find it of high relevance to help CBS secure the quality of the study programmes supported by the student evaluations. This leads to the first part of this paper which addresses the response rates and gives an overview of the data applied. This is covered by our first research question regarding which students submit their student evaluations, and which do not. To examine this, we use logistic regression. Further we seek to find a method for predicting the respondents for the future. We find that it is not possible to predict which students submit their student evaluation, but we find that the effects "gender" and "nationality" are decisive.
In the second part of the analysis, we investigate how the students respond to the question "The course has extensively increased my knowledge of the subject". For this analysis we apply cumulative logit models with and without a random effect. We find that "gender", "nationality", "the level of complexity of the syllabus", "time spent preparing for the course" and "the use of lecturer's office hours", are decisive effects.
We reach the conclusion that with the use of logistic regression and cumulative logits with a random effect, statistical methods for analysis of the student evaluations from the academic year 2018/2019 at CBS, are found.

The data applied is extensive and we have only been able to cover parts of it, due to the limitations of the formal requirements for this paper. This leads to our discussion of the abilities of the results to cover for CBS in general or only for the study programmes covered in the analysis.

God læselyst

-

Julie & Sofie

Indholdsfortegnelse

1 Indledning **5**

1.1 Problemformulering 6

1.2 Afgrænsning 6

1.3 Opgavens struktur 8

2 Vores data **9**

2.1 Historisk overblik over evalueringsmetode samt svarprocenter 9

2.2 Datas oprindelse og indhold 11

2.3 Databearbejdning på det samlede datasæt 15

2.3.1 Spørgeskemaet 16

2.3.2 Variabelliste 18

2.4 Krydstabeller på det endelige datasæt 26

3 Teoretisk analyse af kategorisk data **35**

3.1 Kategorisk Data 35

3.2 Generaliserede Lineære Modeller 36

3.2.1 Estimation 43

3.2.2 Statistisk inferens 44

3.2.3 Model selektion 48

3.3 Generaliseret Lineær Mixed Modeller 49

3.3.1 Estimation 50

3.3.2 Statistisk inferens 51

3.3.3 Modelselektion 52

3.4 Logistisk regression med binær responsvariabel 52

3.4.1 Estimation 54

3.4.2 Statistisk inferens og modelselektion 57

3.4.3 Prædiktion 57

3.5 Kumulativ logit model 60

3.5.1 Estimation 65

3.5.2 Statistisk inferens og modelselektion 66
3.6 Kumulativ logit model med tilfældig effekt 69
3.6.1 Estimation 71
3.6.2 Statistisk inferens og modelselektion 72

4 Hvilke studerende besvarer, og hvilke besvarer ikke kursusevalueringen 74
4.1 Anvendt data 74
4.2 Logistisk regression for Kandidatuddannelsen i Erhvervsøkonomi 78
4.3 Logistisk regression for Bacheloruddannelsen i Erhvervsøkonomi 83
4.4 Logistisk regression for Bacheloruddannelsen i Erhvervsøkonomi og Erhvervsjura 86
4.5 De 3 logistiske regressionsmodeller 89
4.6 Prædiktion: Hvilke studerende besvarer og hvilke besvarer ikke kursusevalueringen 92

5 Hvordan de studerende svarer på kursusevalueringen 97
5.1 Anvendt data 97
5.2 Kumulativ logit model for Master of Science i Business Language and Culture 100
5.2.1 Test af proportional odds antagelsen 103
5.2.2 Goodness-of-fit test 104
5.2.3 Opsummering af kumulativ logit model (CCDCO2002U) 105
5.3 Kumulativ logit model med tilfældig effekt for Master of Science in Business Language and Culture 105
5.3.1 Test af model med mixed effekt 109
5.3.2 Det "gennemsnitlige" kursus 110
5.3.3 Opsummering af kumulativ logit model med tilfældig effekt (MScBLC) 113

6 Opsamling af opgaven 115
6.1 Resultater 115
6.2 Diskussion 117
6.3 Perspektivering 119

7 Bilag **124**

A Opgørelse af CBS's bestand 124

B Quasibinomial startmodel output for BScHA 125

C Quasibinomial startmodel output for BScJUR 126

D Kumulativ logit model output for startmodel for CCDCO2002U_E18 127

E Kumulativ logit model med tilfældig effekt output for startmodel for MScBLC 128

F R-kode 129

F.1 Kode vedrørende den logistisk regressionsmodel 129

F.2 Kode vedrørende den kumulative logit model 131

1 Indledning

Kursusevalueringer har været en almindelig del af vores studieliv på Copenhagen Business School (CBS) de seneste seks år. Hen imod slutningen af hvert kursus har vi fået tilsendt en kursusevaluering, som vi er blevet bedt om at besvare. Flere gange har vi undret os over, hvordan det kan være, at vi skal evaluere alle vores kurser, men også hvordan resultaterne fra kursusevalueringerne egentlig bliver anvendt.

CBS stræber efter høj uddannelseskvalitet for at sikre, at vi som studerende og dimittender lever op til det højeste internationale niveau. En af grundende til de mange kursusevalueringer er således, at CBS anvender disse til at kvalitetssikre undervisningen (Copenhagen Business School, 2021). En anden årsag til brugen af kursusevalueringer er, at kvalitetssikring af uddannelserne på uddannelsesinstitutioner er et lovkrav: "Universitetet skal sørge for, at dets uddannelser efter stk. 1 og 2 er omfattet af en kvalitetssikring enten i Danmark efter reglerne i lov om akkreditering af videregående uddannelsesinstitutioner eller i udlandet efter udenlandske kvalitetssikringssystemer" jf. universitetslovens § 3 a, stk. 3.

En uddannelse er en samfundsmæssig investering, og netop derfor synes vi, at dét at kvalitetssikre undervisningen, er af høj relevans. Desværre er vores opfattelse at det langt fra er alle studerende, der får besvaret deres kursusevalueringer. Dette er en skam, for det betyder, at CBS har et mindre grundlag at måle kvaliteten af undervisningen på, og dermed også et ringere grundlag for at forbedre uddannelserne, der hvor det er nødvendigt. Derfor vil vi med denne opgave forsøge at skabe en forståelse af, hvilke studerende der besvarer, og hvilke studerende der ikke besvarer deres kursusevalueringer, og om der er nogle forhold, der er bestemmende herfor.
På den måde håber vi på, at kunne hjælpe CBS til at lave en strategi for, hvilke studerende de skal henvende sig til, for eksempelvis at øge svarprocenten. Ved at øge svarprocenten vil CBS få et mere præcist billede af den samlede population af studerende, og på den måde få et mere fuldstændigt billede af kvaliteten af undervisningen på skolen.
Ved at have højkvalitets uddannelser bidrager CBS positivt til erhvervslivet, med dimittender

af høj kvalitet. Dette er til gavn for både de studerende, som bliver udfordret på deres uddannelse, men også for den danske samfundsøkonomi.

Efter seks lærerige, udfordrende og sjove år på CBS ønsker vi nu at give lidt igen, og håber på blandt andet at kunne bidrage med en metode til, hvordan CBS kan undersøge, hvilke studerende der besvarer, og hvilke der ikke besvarer deres kursusevalueringer.

1.1 Problemformulering

Formålet er at undersøge, hvorvidt det er muligt at opstille statistiske modeller til brug for analyse af data fra kursusevalueringer. Dette gøres med henblik på at bestemme en metode til dels at forudsige, hvilke studerende der besvarer kursusevalueringen, og dels til at bestemme en metode, der kan anvendes til at undersøge hvilke forhold, der er bestemmende for de studerendes faglige udbytte af undervisningen. De anvendte metoder er dels en logistisk regressionsmodel med overspredning og dels modeller for kumulativ logit med og uden tilfældig effekt. Udgangspunktet for undersøgelsen er de studerendes kursusevalueringer for studieåret 2018/2019 på Copenhagen Business School (CBS). Information om de studerende, som anvendes i analysen, er de demografiske data herunder køn, nationalitet, fødselsår og opnåede karakterer.

1.2 Afgrænsning

For at kunne lave en dybdegående analyse, der besvarer problemformuleringen, er det nødvendigt at lave nogle afgræsninger, særligt i forbindelse med valg af data. Først og fremmest ønsker vi at undgå de potentielle effekter, som COVID-19 har haft på CBS's kursusevaluerings resultater, og derfor tager vi udgangspunktet i det seneste studieår, inden COVID-19 begyndte at have en effekt på undervisningen på CBS. Dette er årsagen til, at denne opgave er afgrænset til at bruge data for studieåret 2018/2019.

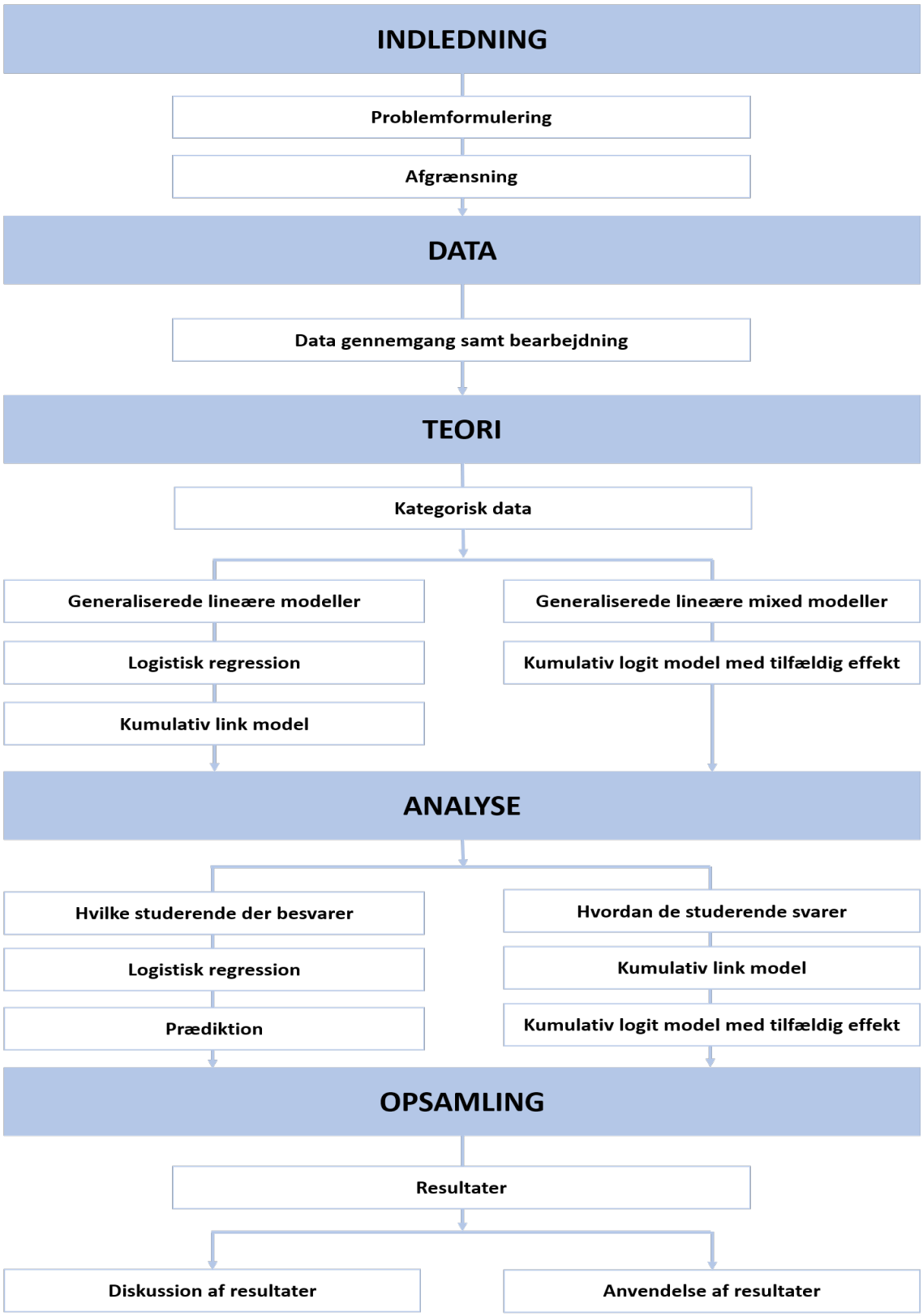
Det modul i CBS's evalueringssystem, som datasættet stammer fra, indeholder kursuseva-

lueringerne for både alle dagsuddannelserne[1] samt alle HD-uddannelserne[2]. Der indgår dermed tre uddannelsesniveauer i det datasæt, som er modtaget i forbindelse med denne opgave. Vi har truffet et valg om at afgrænse datasættet til kun at indeholde dagsuddannelserne blandt andet fordi, at HD-uddannelserne skiller sig ud ved at være en betalingsuddannelse. At det er en betalingsuddannelse er også årsagen til, at denne afgrænsning er lavet, da det kan have en betydning for, hvad en typisk studerende er på en HD-uddannelse versus en typisk studerende på en bachelor- eller kandidatuddannelse. Vi ønsker derfor at undgå, at disse studerende blandes sammen. Vi har i forvejen et stort datasæt, så selvom det afgrænses til kun at indeholde studerende tilknyttet bachelor- og kandidatuddannelserne, har vi 80.431 observationer.

Da vi stadigvæk har at gøre med et stort datasæt, er det derfor nødvendigt at afgrænse yderligere. Gennem opgaven træffer vi løbende nogle valg, hvor det kun er dele af data, som analyseres, da det ikke vil være muligt at analysere alle observationerne inden for omfanget af denne opgave. Løbende gør vi os nogle erfaringer, som er med til at præge udvælgelsen af hvilke dele af datasættet, der fokuseres på undervejs i opgaven.
Datasættet indeholder hovedsageligt kategorisk data, hvilket er årsagen til, at der anvendes statistiske metoder, som er særligt egnet til at håndtere dette. Vi ser bort fra de dele af spørgerammen for kursusevalueringen, som ikke anvender en målbar skala, det er for eksempel kommentarfelter. Opgaven afgrænses yderligere til kun at se på de spørgsmål i kursusevalueringen, som vedrører selve kurset. Det betyder at de spørgsmål, der relaterer sig til underviserne er frasorteret.

[1] Bachelor- og kandidatuddannelserne
[2] Diplomuddannelse

1.3 Opgavens struktur



2 Vores data

2.1 Historisk overblik over evalueringsmetode samt svarprocenter

Dette afsnit er inkluderet i opgaven for at give læseren en forståelse for, hvordan kursusevalueringsprocessen forløber på CBS samt en forståelse af, hvordan både svarprocenten, bestanddelen af studerende[3] samt antallet af kurser historisk set har fordelt sig siden foråret 2012 til og med efteråret 2020. Dette har til formål at understrege, at det anvendte datasæt i opgaven, som er for studieåret 2018/2019, afspejler et typisk studieår for CBS's kursusevalueringer.

CBS har evalueret efter samme metode siden 2008. Kursusevalueringerne for henholdsvis bachelor- og kandidatuddannelserne på CBS bliver håndteret efter samme metode, og de studerende bliver mødt af den samme spørgeramme, når de modtager deres kursusevalueringer. Der bliver udsendt en kursusevaluering for alle de kurser, som indgår i hver enkelt uddannelses studieordning. Dog håndteres praktikforløb, bachelorprojekter og kandidatafhandlinger særskilt og er ikke en del af kursusevalueringerne på CBS. Det er dermed alle ECTS[4] givende kurser, der bliver evalueret i det samme modul i evalueringssystemet, bortset fra førnævnte praktikforløb, bachelorprojekter og kandidatafhandlinger, som også er ECTS-givende.

Evalueringsprocessen forløber sådan, at en kursusevaluering bliver sendt ud til den studerende dagen før sidste undervisningsgang, og den studerende vil have mulighed for at besvare spørgeskemaet frem til og med dagen før eksamen. Der vil derfor altid blive evalueret før en eksamen. Måden den studerende kan evaluere deres kursusevaluering på er enten via et link fra en invitations e-mail eller via de to efterfølgende påmindelses e-mails, som sendes til den studerendes CBS e-mail via evalueringssystemet. Påmindelses e-mails bliver kun sendt til den studerende, hvis den studerende ikke har besvaret kursusevalueringen. Alternativt kan den studerende altid tilgå kursusevalueringen via sit Learning Management System (LMS), som på nuværende tidspunkt er Canvas, hvor underviserne også kommunikerer til de

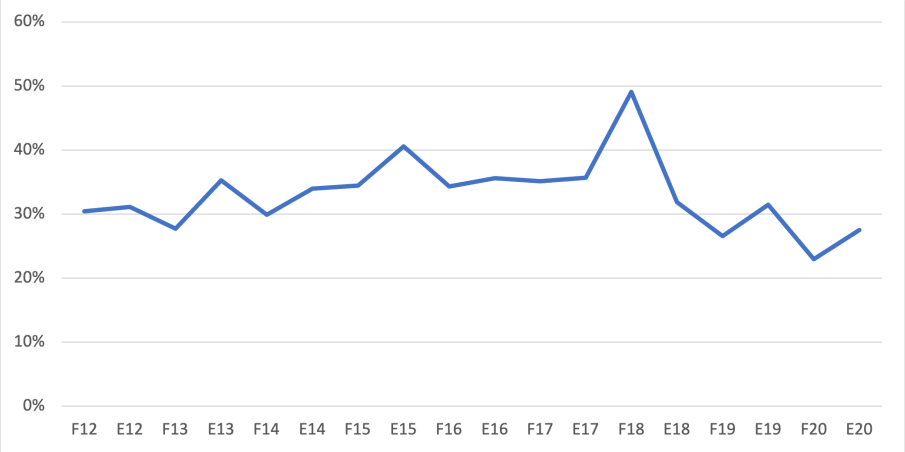
[3] Se Bilag A

[4] ECTS-systemet er et internationalt pointsystem der er konstrueret, så et års fuldtidsuddannelse udgør 60 ECTS-point(Uddannelses- og Forskningsministeriet, 2020)

studerende. CBS anbefaler underviserne at afsætte 10 minutter ved sidste undervisningsgang til at evaluere, men det er op til underviserne selv, hvorvidt det gøres eller ej.

Ser man på svarprocenterne for dagsuddannelserne over en tidsrække, der spænder fra forårsemestret 2012 (F12) til og med efteråret 2020 (E20) ses det, at svarprocenten har varieret over tid, hvilket fremgår af Figur 1. Det største udsving ses i foråret 2018 (F18), hvor svarprocenten har været helt oppe på 49%. Foråret 2020 er det semester, hvor der har været den laveste svarprocent, her var den 23%.

Figur 1: Svarprocent per semester fra Forår 2012 til Efterår 2020

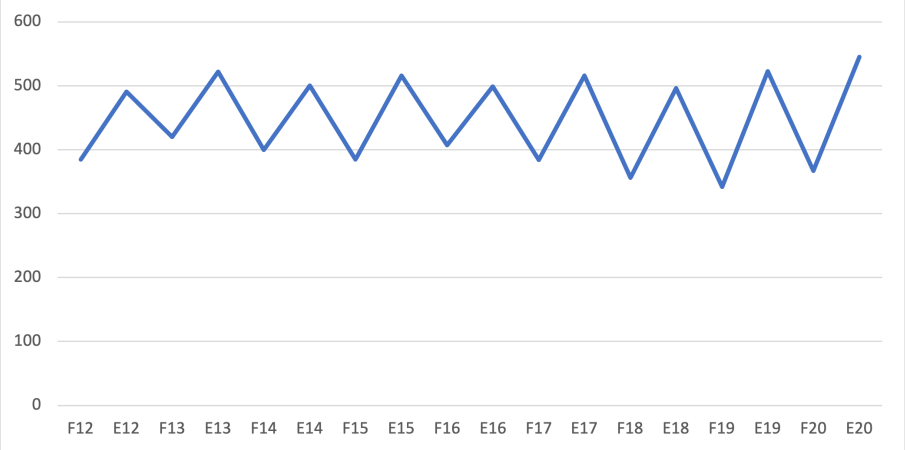


Bestanden af studerende på CBS's dagsuddannelser har været en smule aftagende siden 2012. I 2017 skete der et markant fald i bestanddelen af studerende, hvilket kan forklares med at fremdriftsreformen trådte i kraft. Dette har blandt andet betydet, at "evighedsstuderende" blev smidt ud, samt at de studerende blev presset til hurtigere at færdiggøre deres uddannelse (Copenhagen Business School, 2017). Om dette kan have en sammenhæng med svarprocentens stigning i F18 er ikke til at sige. Siden fremdriftsreformen er trådt i kraft, er bestanddelen steget en smule og ser nu ud til at ligge stabilt omkring 16.000 indskrevne studerende på dagsuddannelserne.

Ses der på antallet af kurser fordelt på semestrene fra F12 til E20, er det tydeligt at se en klar

tendens med udsving i antal kurser imellem forårs- og efterårssemestrene. Det ses af figur 2, at antallet af kurser i efteråret er steget med tiden. Efteråret indeholder mange valgfag, det kan derfor godt tænkes, at det øgede antal kurser skyldes et større udbud af valgfag.

Figur 2: Antal kurser per semester fra Forår 2012 til Efterår 2020



Som tidligere nævnt, så er der blevet evalueret på samme måde siden efteråret 2008 på CBS, og efter samme metode, altså samme spørgeramme og samme udsendelsesmetode, men i efteråret 2018, blev evalueringsystemet skiftet ud. Den studerende har derfor ikke bemærket en forskel, kun eventuelt at udseendet af evalueringen er blevet ændret en smule.

Efter at have set på svarprocenten, bestanden af studerende på dagsuddannelserne, antallet af kurser samt evalueringsprocessen over tid, mener vi at data for studieåret 2018/2019 afspejler et gennemsnitligt studieår på CBS.

2.2 Datas oprindelse og indhold

Datasættet i denne opgave består af to dataudtræk, som er blevet flettet sammen. Det ene dataudtræk er fra CBS's evalueringssystem og det andet dataudtræk er fra STADS, et studieadministrativ system, som blandt andet indeholder de studerendes opnåede karakterer. Dataudtrækket fra evalueringssystemet indeholder alle resultater for bachelor- og kandidatuddannelsernes kursusevalueringer for studieåret 2018/2019, hvilket vil sige, at dataudtrækket

består af de to semestre efterår 2018 (E18) og forår 2019 (F19). Dataudtrækket fra STADS indeholder alle de studerendes opnåede eksamenskarakterer for kurserne i E18 og F19.
Nedenfor beskriver vi, hvad de to dataudtræk indeholder og herefter beskriver vi, hvordan de to dataudtræk er flettet sammen, og hvilke komplikationer det har medført. Dette leder til vores samlede datasæt, som er det endelige datasæt, der anvendes i opgaven.

Indhold af dataudtrækket med kursusevaluering resultaterne

Dataudtrækket fra evalueringssystemet indeholder alle de udsendte kursusevalueringer, der har været udsendt for dagsuddannelserne i E18 og F19. Dataudtrækket indeholder dermed både de udsendte kursusevalueringer, som ikke er besvaret, samt dem der er besvaret.
På CBS indeholder kursusevalueringen forskellige spørgerammer, da der findes forskellige typer af undervisningsformer for kurserne. De forskellige typer af undervisningsformer, som kan anvendes er: tilstedeværende undervisning, online undervisning, blended undervisning[5] og projektbaseret undervisning. Der findes en spørgeramme til hver af de forskellige undervisningsformer, som kaldes for henholdsvis "standard", "online", "blended" og "projekt". I alle spørgerammerne, bortset fra "projekt", indgår hele "standard" spørgerammen. I "online" og "blended" spørgerammerne bliver "standard" spørgerammen suppleret med spørgsmål tilknyttet de online og blended elementer kurset må have.

På nogle kurser kan den studerende være tilmeldt flere hold, som alle har med det samme kursus at gøre, da der kan være forskellige behov i undervisningen. Et eksempel kan være, at et kursus, for eksempel Mikroøkonomi, indeholder både forelæsninger og øvelser. Til forelæsningerne vil alle de studerende typisk være samlet på et forelæsningshold, og til øvelserne vil de studerende typisk være opdelt på forskellige øvelseshold. Her vil den studerende altså både være tilmeldt forelæsningsholdet og et af øvelsesholdene, som begge er tilknyttet kurset Mikroøkonomi. For nogle kurser kan det være, at de studerende på de forskellige øvelseshold møder forskellige undervisere. Der vil dog altid blive evalueret på mindste enhed, hvilket betyder, at man følger de studerende ned til det mindste hold, som de er tilmeldt,

[5]En blanding af tilstedeværende undervisning og online undervisning

og her bliver de bedt om at evaluere de undervisere, som de har mødt på deres øvelseshold, de undervisere de har mødt til forelæsningerne, og derudover bliver de bedt om at evaluere kurset. Der kan derfor godt være scenarier, hvor de studerende er fra samme uddannelse og har samme kursus, men har mødt forskellige undervisere.

Evalueringsystemet indeholder udover evalueringsresultaterne også demografiske data om underviserne og de studerende, som er med i dataudtrækket. De demografiske data vedrørende den studerende er informationer om den studerendes køn, nationalitet og fødselsår.
Dette dataudtræk indeholder studerende, som er tilmeldt et kursus helt frem til den sidste undervisningsgang. Hvis en studerende af en eller anden årsag har frameldt sig kursets eksamen eller efterfølgende er blevet afmeldt kurset, vil den studerende stadig have haft mulighed for at evaluere kurset og vil stadig fremgå som en, der har været tilmeldt kurset.

Indhold af dataudtrækket med de studerendes opnåede karakterer
Dataudtrækket fra STADS, indeholder alle karakterer, der er givet for dagsudannelserne for semestrene E18 og F19. Nogle kurser har delprøver undervejs i kursusforløbet, og disse karakterer indgår også i dataudtrækket. Derudover indeholder dataudtrækket både eksamensresultater fra de ordinære eksamener samt reeksamenerne. Dette betyder, at hvis den studerende er udeblevet eller har dumpet den ordinære eksamen og efterfølgende har været til reeksamen, da vil den seneste opnåede karakter fremgå i dataudtrækket.
Dette dataudtræk indeholder kun studerende, som har været tilmeldt en eksamen i et kursus.

Sammenfletning af de to dataudtræk
Kursusevalueringerne bliver evalueret ud fra en kursuskode, som er unik for hvert enkelt kursus, disse koder kaldes for UVA-koder. Når der gives karakterer til de studerende, bliver karaktererne givet på en unik eksamenskode. Hver enkelt eksamen har sin egen unikke eksamenskode, som kaldes for en EKA-kode. Det vil sige, at alle kurser både har en UVA-kode og en eller flere EKA-koder tilknyttet. Et kursus kan have flere EKA-koder, hvis der for eksempel har været delprøver, altså flere eksamener, tilknyttet kurset. Karakterudtrækket

fra STADS er det eneste der indeholder både UVA-koder og EKA-koder og er derfor det eneste der kan hjælpe os med at sammenflette dataudtrækkene.
Desværre kan der være uoverensstemmelser i UVA-koderne mellem karakterudtrækket og dataudtrækket fra evalueringssystemet, hvilket har besværliggjort sammenfletningen for os. Årsagen til at der kan være uoverensstemmelser er blandt andet, at nogle uddannelser har sammenlæsning. Sammenlæsning vil sige, at der er nogle studerende fra forskellige uddannelser, som følger samme kursus. Et eksempel på sammenlæsning er Bacheloruddannelsen i Erhvervsøkonomi og Projektledelse og Bachelor of Science in International Shipping and Trade, som har kurset Mikroøkonomi sammen. I evalueringssystemet vil kurset blive evalueret på én UVA-kode, grundet fælles undervisning, men i dataudtrækket med karakterer, vil de studerende være opdelt på hver deres UVA-kode, som angiver hver deres uddannelse, altså har de to forskellige UVA-koder. For at tage højde for dette, har vi lavet en manuel fletning, der får de studerendes opnåede karakterer tilknyttet deres evalueringsresultater.
For at flette dataudtrækkene rigtigt sammen, laver vi et overblik over, hvilke uddannelser der har sammenlæsning og hvilke kurser det omhandler. Når vi har dette overblik er det muligt at slå op i kursuskataloget[6], hvilke UVA-koder de to uddannelser har. Ved at rette dette til i dataudtrækket med evalueringsresultaterne, kan de studerendes opnåede karakterer flettes på dataudtrækket. Vi vil i resten af opgaven omtale UVA-koderne som "kursuskoder".

En anden udfordring der opstår under sammenfletningen af de to dataudtræk er, at der findes kurser, som forløber over hele studieåret, altså over begge semestre: E18 og F19. Nogle af disse kurser vil først have en eksamen i F19, når studieåret slutter. Der vil dog være en evaluering af kurset for hvert semester. Ved sammenfletningen er det kun kurset for F19, som får karakterer. For at få tilknyttet karakterer i E18 har vi lavet et overblik over, hvilke uddannelser der har kurser, som forløber over to semestre. Ved at have dette overblik har vi ved manuelt arbejde, tildelt den studerendes opnåede karakter til både evalueringsresultaterne for kurset i E18 samt F19, selvom der kun har været én eksamen og kun er givet én karakter i kurset.

[6]CBS har et kursuskatalog over alle kurser (https://kursuskatalog.cbs.dk/search.aspx)

Derudover findes der nogle kurser, som har flere delprøver undervejs i kursusforløbet, og dermed flere eksamener tilknyttet samme kursus. Ved at skabe et overblik over, hvilke kurser dette omhandler, har vi ved manuelt arbejde, fået tildelt den sidste opnåede eksamenskarakter til den studerendes evalueringsresultater.

Som beskrevet tidligere, holder evalueringsssytemet øje med, hvilke studerende der har været tilmeldt kurset frem til evalueringen starter, men herefter tager den ikke højde for, om den studerende er blevet afmeldt kurset. Da eksamen i kurset foregår til sidst, vil det her fremgå, hvis nogle studerende er blevet afmeldt kurset inden eksamen. Der er derfor nogle studerende fra evalueringsudtrækket, som ikke har fået en karakter ved sammenfletningen, og her vil "NA" fremgå som opnået karakter.

Vi har nu et stort fælles datasæt, som vi anvender i denne opgave. Datasættet består af en række for hver udsendt kursusevaluering for semestrene E18 og F19, hvor den studerendes opnåede karakter i kurset er tilknyttet.

2.3 Databearbejdning på det samlede datasæt

Vi har nu et stort datasæt, som kræver lidt yderligere databearbejdning, før at vi står tilbage med et datasæt, der indeholder de oplysninger, som vi ønsker at bruge til vores analyse. Vi ønsker også at reducere datasættet, da det indeholder en masse data, som vi ved, at vi ikke får brug for, og således med fordel kan fjerne på forhånd.
Grundet de forskellige spørgerammer har vi først og fremmest ekskluderet de kurser fra datasættet, som er blevet evalueret med en "projekt" spørgeramme, da vi ved, at de ikke har noget til fælles med de andre spørgerammer, og derfor kan resultaterne ikke bruges. Dette omhandler 0 kursusevalueringer fra E18 og 13 kursusevalueringer for F19. De resterende spørgerammer er beholdt, dog er de supplerende spørgsmål i "online" og "blended" spørgerammen blevet fjernet. Der indgår dermed kun spørgsmål, som er en del "standard" spørgerammen, som alle studerende har besvaret, hvis de har besvaret deres kursusevaluering.
"Standard" spørgerammen indeholder spørgsmål samt kommentarfelter, hvor kommentar-

felterne bliver ekskluderet fra datasættet, da der vil være behov for tekstanalyse til at analysere disse, og det går vi ikke videre med i denne opgave. Vi har derfor kun spørgsmål med i datasættet, som har tilknyttet en målbar skala.
Denne opgave er afgrænset til kun at fokusere på kursuset og ikke underviseren, derfor er spørgsmålene der omhandler underviserne samt deres demografiske data ekskluderet fra datasættet. Fordi vi vælger at slette underviserspørgsmålene kan vores datasæt reduceres en del. Det kan det fordi der i datasættet er en række per udsendte evaluering per underviser. Det vil sige, at hvis en studerende på et kursus har mødt fire forskellige undervisere, da vil der være fire rækker i datasættet. Dette gør det dermed muligt at reducere datasættet en del, således at vi nu kun har en række per studerende per kursus.
Grundet personfølsomme oplysninger ekskluderes alle de kurser, hvor der er under 5 besvarelser, hvilket er den samme tærskelværdi, som CBS bruger, når der laves rapportering på evaluering resultaterne. Dette er for at opretholde anonymitet.
Uddannelser som har været under udfasning, eller som er helt nye er blevet ekskluderet, da der ikke er fyldestgørende data til at kunne sige noget om disse uddannelser. Dette omhandler 4 uddannelser; Bachelor i Europæisk Business (BAEUB), Bachelor of Arts in Information Management, (BAIM), Bachelor i Erhvervsøkonomi i markeds- og kulturanalyse (BScMAK) og Master of Science in Business Administration and Bioentrepreneurship (MScBIO).
Vi har nu et endeligt datasæt, som er klar til brug.

2.3.1 Spørgeskemaet

Dette leder os frem til følgende spørgsmål som indgår i vores endelige datasæt:

- **Angiv i hvilken grad du er enig:**
  - Kursus_S1: Jeg har fået et stort fagligt udbytte af undervisningen
  - Kursus_S2: Der var god sammenhæng i fagets faglige og pædagogiske opbygning
  - Kursus_S3: Fagets mål er klare for mig
  - Kursus_S4: Jeg oplevede at der var feedback elementer integreret i undervisningen (fx skriftlige afleveringer, quizzer og peer-2-peer)

  - Kursus_S5: Alt i alt havde jeg en positiv oplevelse af faget

  Svarmuligheder:

  - 1: Helt uenig
  - 2: Overvejende uenig
  - 3: Hverken enig eller uenig
  - 4: Overvejende enig
  - 5: Helt enig
  - N/A: Uafklaret/ikke relevant

- **Pensums sværhedsgrad var...**

  - 1: Alt for lav
  - 2: For lav
  - 3: Passende
  - 4: For høj
  - 5: Alt for høj

- **Hvor stort er dit tidsforbrug til forberedelse og efterbearbejdning pr. lektion (á 45 minutter)? - fremmøde tælles ikke med**

  - 1: Ca. 1/2 time (pr. lektion)
  - 2: Ca. 1 time (pr. lektion)
  - 3: Ca. 1 1/2 time (pr. lektion)
  - 4: Ca. 2 timer (pr. lektion)
  - 5: Over 2 timer (pr. lektion)

- **Hvor mange af lektionerne til dette fag er du mødt op til?**
  - 1: Ca. 20 procent
  - 2: Ca. 40 procent
  - 3: Ca. 60 procent
  - 4: Ca. 80 procent
  - 5: Ca. 100 procent
- **Har du gjort brug af din undervisers kontortid?**
  - 1: Aldrig
  - 2: 1-2 gange
  - 3: 3-4 gange
  - 4: Mere end 4 gange

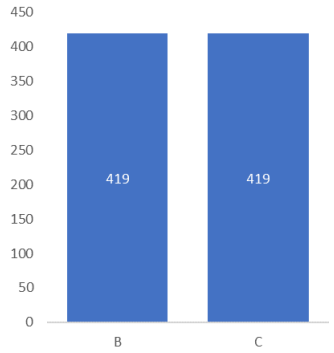
2.3.2 Variabelliste

Efter databearbejdningen indgår 23 variable i datasættet. Nedenfor beskrives variablene enkeltvis, og de fleste har et tilhørende histogram, som giver et overblik over størrelsen af variablene. Variablene er udvalgt med det formål at kunne anvendes til at analysere om der er nogle tendenser for, hvilke studerende der får besvaret deres kursusevalueringer. Derudover kan variablene også anvendes til at analysere, om der er nogle tendenser omkring, hvordan de studerende der besvarer kursusevalueringen så svarer på spørgsmålene.

Niveau

Denne variabel beskriver, om det enkelte kursus tilhører en bachelor- eller kandidatuddannelse. Variablen kan enten være "B" for en bacheloruddannelse eller "C" for kandidatuddannelse. Af histogrammet på figur 3 nedenfor fremgår fordelingen af henholdsvis "B" og "C".

Figur 3: Histogram af variablen Niveau



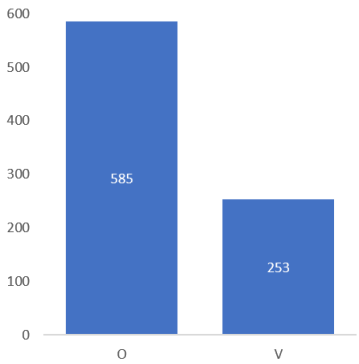
Kursus_Kode

Denne variabel beskriver, hvilket kursus det er, der bliver evalueret. Der er i alt 809 forskellige kursuskoder i datasættet. En kursuskode kan godt være identisk på tværs af to semestre, da et kursus kan forløbe over to semestre.

Kursus_Type

Denne variabel beskriver, hvilken type kurset er, da et kursus kan være enten et obligatorisk kursus eller et valgfag. Så variablen kan enten være "O" for obligatorisk kursus eller "V" for valgfag. Af histogrammet på figur 4 nedenfor fremgår fordelingen af henholdsvis "O" og "V".

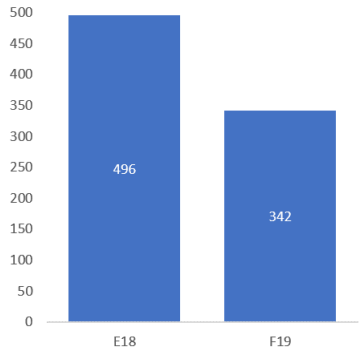
Figur 4: Histogram af variablen Kursus_Type



Eval_Semester

Denne variabel beskriver, hvilket semester den enkelte kursusevaluering tilhører. Variablen kan være enten "E18" for efterårssemestret 2018 eller "F19" for forårssemesteret 2019. Af histogrammet nedenfor i figur 5 fremgår det, hvor mange kursusevalueringer der har været sendt ud i henholdsvis "E18" og "F19". Der er normalt færre kurser i foråret, da flere studerende er igang med deres bachelorprojekt eller kandidatafhandling.

Figur 5: Histogram af variablen Eval_Semester



Kursus_Kode_Eval_Semester

Denne variabel er en kombination af kursuskoden, og hvilket semester kurset er blevet evalueret på. Der er i alt 838 kurser, når et kursus som løber over to semestre ses på som to kurser, da der er blevet udsendt to kursusevalueringer.

Udd_Kode

Denne variabel beskriver moderuddannelsen til det kursus, der evalueres og er en forkortelse for uddannelseskode. Ved moderuddannelse forståes den overordnede uddannelse. Dette betyder, at nogle af uddannelserne indeholder flere retninger eller linjer, som for eksempel bacheloruddannelsen BSc In Business Language and Culture, som har en tysk, en fransk og en spansk linje. Der er i alt 39 forskellige uddannelseskoder.

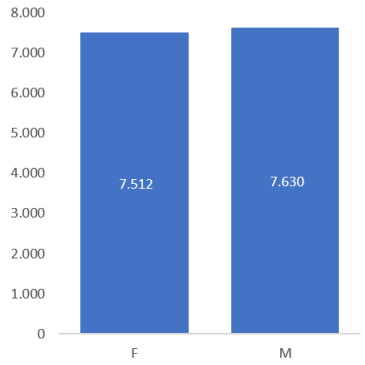
Student_ID

Denne variabel viser hver enkelte studerendes unikke ID. Der er i alt 15.142 forskellige studenter ID'er. Alle studerende er ikke nødvendigvis tilmeldt et kursus på begge semestre, dette kan skyldes, at den studerende er på udveksling det ene semester, eller at en studerende mangler nogle enkelte kurser. Der kan være flere årsager end disse to.

Køn

Denne variabel beskriver den studerendes køn. Variablen kan enten være "F" for female, altså en kvindelig studerende, eller "M" for man, altså en mandlig studerende. Af histogrammet nedenfor i figur 6 fremgår det, hvor mange studerende der er henholdsvis "F" og "M".

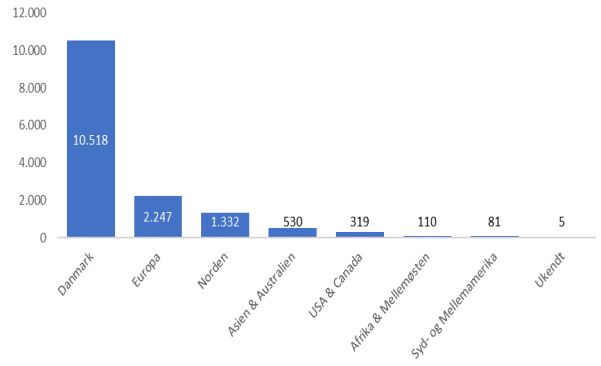
Figur 6: Histogram af variablen Køn



Nationalitet

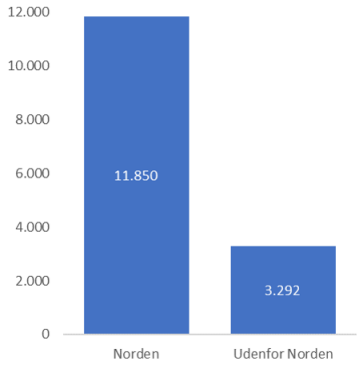
Denne variabel beskriver den studerendes nationalitet. Der er i alt 104 forskellige nationaliteter i datasættet. For overskuelighedens skyld er det besluttet at inddele nationaliteterne i 8 grupper, som er en opdeling CBS også anvender: Europa, Danmark, Asien & Australien, Norden, USA & Canada, Afrika & Mellemøsten, Syd- og Mellemamerika, Ukendt Nationalitet. I histogrammet nedenfor i figur 7 ses fordelingen af de studerendes nationaliteter, inddelt i de 8 grupper.

Figur 7: Histogram af variablen Nationalitet og dets 8 grupperinger



Da det fremgår af ovenstående histogram, at det er begrænset, hvor mange internationale studerende der er, inddeles de 8 grupper i stedet i to nye 2 grupper. Nationalitet er dermed opdelt i de to grupper "Norden" og "Udenfor Norden", således at variablen Nationalitet kun har to niveauer. Herfra og i resten af opgaven vil Nationalitet således referere til "Norden" og "Udenfor Norden" og histogrammet af figur 8 viser fordelingen af disse.

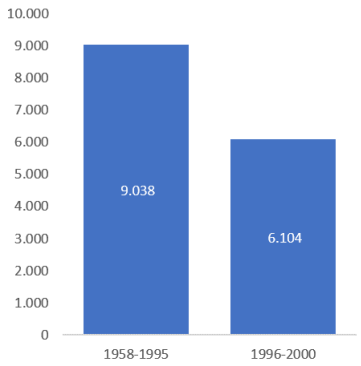
Figur 8: Histogram af variablen Nationalitet fordelt på "Norden" og "Udenfor Norden"



Fødselsår

Denne variabel beskriver den studerendes fødselsår. Der er i alt 44 forskellige fødselsår, som spænder fra år 1958 til år 2000. Fødselsårene er inddelt i 2 intervaller: "1958-1995" og "1996-2000". Af histogrammet i figur 9 fremgår de studerendes fødselsår i de 2 intervaller.

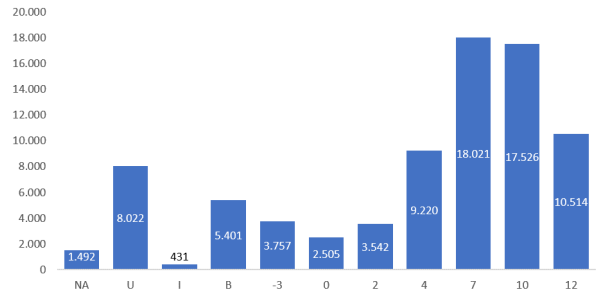
Figur 9: Histogram af variablen Fødselsår



Karakter

Denne variabel beskriver den studerendes karakter i det kursus, der evalueres. Variablen har 11 værdier, hvoraf de 7 er den danske 7-trins skala: "-3", "00", "02", "4", "7", "10", "12". Derudover er der en bestået/ikke-bestået skala med "B" for bestået og "I" for ikke bestået. For begge karakterskalaer findes der et "U", som står for udeblevet. Yderligere er der nogle studerende uden karakter, og her betegner "NA" den manglende bedømmelse. Fordelingen af karakterer fremgår af histogrammet i figur 10.

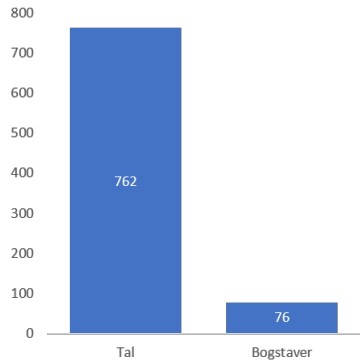
Figur 10: Histogram af variablen Karakter



Skalatype

Denne variabel beskriver hvilken karakterskala kurset anvender. Det vil sige, at hvis kursets eksamen har anvendt 7-trins skalaen, da vil skalatypen fremgå som "Tal" og hvis det har været bestået/ikke-bestået skalatypen der er anvendt, så vil der stå "Bogstaver". Af histogrammet i figur 11 fremgår fordelingen af "Tal" og "Bogstaver".

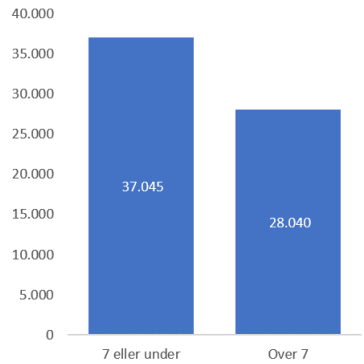
Figur 11: Histogram af variablen Skalatype



Karakteropdeling

Denne variabel beskriver, om den givne karakter har været "7 eller under" eller "Over 7". Det er kun kurser, hvor skalatypen "Tal" er anvendt. I figur 12 nedenfor fremgå et histogram af fordelingen af de to karakteropdelinger.

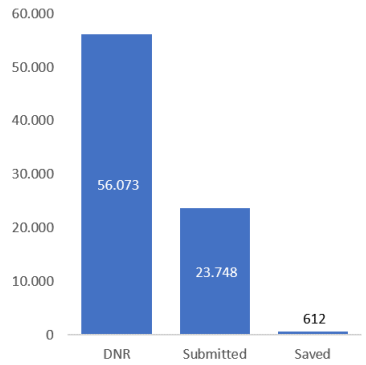
Figur 12: Histogram af variablen Karakteropdeling



Resp_Status

Denne variabel beskriver, om den studerende har besvaret kursusevalueringen og er en forkortelse for respondent status. Variablen har tre værdier; "DNR", som er en forkortelse for "Did Not Reply", altså at den studerende ikke har besvaret spørgeskemaet, "Saved", altså den studerende har påbegyndt sin besvarelse af kursusevalueringen, men har ikke færdiggjort den, eller "Submitted" som betyder, at den studerende har besvaret spørgeskemaet. Af histogrammet i figur 13 nedenfor fremgår fordelingen af de 3 typer af respondent status.

Figur 13: Histogram af variablen Resp_Status



Kursus_S1, Kursus_S2, Kursus_S3, Kursus_S4, Kursus_S5

Disse variable er alle tilknyttet kursusspørgsmålene fra kursusevalueringen. Disse variable har alle samme form og er derfor beskrevet sammen. Variablene kan have værdierne 1, 2, 3, 4, 5, "NA" eller D/A. Hvor værdien 1 er "Helt uenig", 2 er "Overvejende uenig", 3 er "Hverken enig eller uenig", 4 er "Overvejende enig" og 5 er "Helt enig", "NA" er "Uafklaret/ikke relevant" og ved de studerende, der ikke har besvaret spørgeskemaet vil "D/A" fremgå.

Pensum

Denne variabel kan have værdierne: 1, 2, 3, 4, 5 eller "D/A", og beskriver den studerendes egen vurdering af pensums sværhedsgrad. Hvor 1 er "Alt for lav", 2 er "For lav", 3 er "Passende", 4 er "For høj" og 5 er "Alt for høj" og for de studerende, der ikke har besvaret spørgeskemaet, vil "D/A" fremgå.

Tidsforbrug

Denne variabel kan have værdierne 1, 2, 3, 4, 5 samt D/A, og beskriver det tidsforbrug den studerende selv mener at have brugt på forberedelse og efterbearbejdning pr. lektion. Hvor 1 er "Ca. 1/2 time" (pr. lektion), 2 er "Ca. 1 time (pr. lektion)", 3 er "Ca. 1 1/2 time (pr . lektion), 4 er "2 timer (pr. lektion) og 5 er "Over 2 timer" (pr. lektion). For de studerende, der ikke har besvaret spørgeskemaet, vil D/A fremgå.

Fremmøde

Denne variabel kan have værdierne 1, 2, 3, 4, 5 samt D/A og beskriver, hvor mange lektioner den studerende selv vurderer at have mødt op til i det kursus der evalueres. Hvor 1 er "Ca. 20%", 2 er "Ca. 40%", 3 er "Ca. 60%", 4 er "Ca. 80%" og 5 er "Ca. 100%". For de studerende, der ikke har besvaret spørgeskemaet, vil D/A fremgå.

Kontortid

Denne variabel kan have værdierne 1, 2, 3, 4 samt D/A og beskriver den studerendes vurdering af, hvor mange gange den studerende selv er mødt op til undervisers kontortid. Hvor 1 er "Aldrig", 2 er "1-2 gange", 3 er "3-4 gange" og 4 er "Mere end 4 gange". For de studerende, der ikke har besvaret spørgeskemaet, vil D/A fremgå.

2.4 Krydstabeller på det endelige datasæt

Med kendskab til datasættes endelige indhold er det oplagt at sammenligne variablene på tværs i form af krydstabeller. Eftersom det er et spørgeskema vi har med at gøre, er det oplagt først at undersøge, hvordan svarprocenten fordeler sig. Svarprocenten kan undersøges ud fra krydstabeller på de forskellige variable, for at få en indikation af, om der er nogle tendenser i data. Efterfølgende er det interessant at undersøge, hvordan besvarelserne for de forskellige spørgsmål fordeler sig. Eksempelvis om de studerende generelt er kritisk eller mindre kritiske, og om det for eksempel er normalt at anvende kontortid.

I tabel 1 fremgår svarprocenten fordelt ud på de to semestre samt den overordnede svarprocent for studieåret 2018/2019.

Tabel 1: Overordnet svarprocenter samt fordelt per semester

	Submitted	N	Svarprocent	Antal kurser	Min kursus	Maks kursus
E18	14.278	44.791	31,88%	496	6,13%	93,33%
F19	9.470	35.640	26,57%	342	4,17%	87,50%
Total	23.748	80.431	29,53%	838	4,17%	93,33%

For E18 er den gennemsnitlige svarprocent $31,88\%$, og for F19 er den lidt lavere nemlig på $26,57\%$. Det er umiddelbart ikke en imponerende svarprocent, da det kun er omkring 1/4, der besvarer deres kursusevalueringer i F19. Da vi på CBS evaluerer på kursusniveau, er det interessant at se, hvordan svarprocenten fordeler sig blandt kurserne. Af tabel 1 fremgår det, at der er et stort spænd mellem minimum og maksimum svarprocenten for de enkelte kurser. Svarprocenten blandt kurserne varierer helt fra $4,17\%$ og op til $93,33\%$. Dette gør sig gældende for begge semestre. Det ses også af tabel 1, at antallet af kurser er noget mindre for F19, hvilket giver god mening, da der i forårssemesteret skrives bachelorprojekter og kandidatafhandlinger. Dog kan det ses, at de studerende er lidt bedre til at få evalueret deres kursusevalueringer i efteråret end i foråret. At der er et stort spænd i svarprocenterne på kursusniveau indikerer, at det er muligt at få stort set alle studerende på de forskellige kurser, til at besvare deres kursusevalueringer.

Da datasættet indeholder både bachelor- og kandidatuddannelser, er det interessant at se, om der er forskel på svarprocenten på de to uddannelsesniveauer. De studerende på en kandidatuddannelse kan enten have taget en bachelor på CBS eller en bachelor fra et andet universitet. De studerende kan derfor have forskellige forudsætninger, hvilket potentielt kan have en indflydelse på, om de får besvaret deres kursusevalueringer og hvordan de svarer. Det fremgår af tabel 2, at der ikke er den store variation i svarprocenten iblandt bachelor-

og kandidatuddannelserne, da den gennemsnitlige svarprocent for bacheloruddannelserne er 28, 71%, og den gennemsnitlige svarprocent for kandidatuddannelserne er 30, 66%. Det ser derfor ikke ud til at have nogen betydning for, om den studerende får besvaret deres kursusevalueringer, om den studerende er en bachelor- eller kandidatstuderende.

Tabel 2: Svarprocent fordelt på bachelor og kandidatuddannelser

	Submitted	N	Svarprocent	Antal kurser	Min kursus	Maks kursus
B	13.398	46.670	28,71%	419	4,17%	85,19%
C	10.350	33.761	30,66%	419	4,55%	93,33%
Total	23.748	80.431	29,53%	838	4,17%	93,33%

Det er ret pudsigt, at der er udbudt præcis samme antal kurser for både bachelor- og kandidatuddannelserne, nemlig 419. Det er igen tydeligt, at der er en stor variation i svarprocenterne imellem kurserne. Med den relativt lille forskel i svarprocenten blandt bachelor- og kandidatstuderende, kan vi ikke sige noget om, at bachelorstuderende er bedre til at få besvaret deres kursusevalueringer end kandidatstuderende og omvendt.

På alle uddannelserne vil de studerende have en blanding af to typer af kurser; obligatoriske kurser og valgfag. De obligatoriske kurser er forudbestemt for den studerende, da disse kurser er en del af den valgte uddannelse. Valgfagene kan vælges på tværs af uddannelser, dog kan bachelorstuderende kun vælge bachelor-valgfag, og kandidatstuderende kun vælge kandidat-valgfag. Selvom der er valgfrihed for at vælge imellem disse valgfag, vil hver uddannelse have en fastlagt ramme for, hvilke valgfag der kan vælges imellem. På valgfagene kan de tilmeldte studerende derfor have forskellige forudsætninger for at tage kurset, da det vil være en blanding af studerende fra forskellige uddannelser. Dette forventes at have en indflydelse på, hvordan de studerende svarer på kursusevalueringen, og derfor skal der eventuelt tages højde for dette. Endnu en måde at anskue svarprocenten på kan dermed være at se på om det gør en forskel, at det er et obligatorisk kursus eller et valgfag, som den studerende er tilmeldt. Af tabel 3, fremgår det dog, at svarprocenten næsten er identisk, for obligatoriske

kurser og valgfag, og at kursustypen dermed ikke har en indflydelse på, om de studerende besvarer deres kursusevaluering. Der er dog en del færre valgfag, end der er obligatoriske kurser. Det ses igen, at svarprocenten blandt kurserne også varierer meget fordelt ud på de obligatoriske kurser og valgfag.

Tabel 3: Svarprocent fordelt på obligatoriske fag og valgfag

	Submitted	N	Svarprocent	Antal kurser	Min kursus	Maks kursus
O	18.981	64.294	29,52%	585	4,17%	85,71%
V	4.767	16.137	29,54%	253	5,50%	93,33%
Total	23.748	80.431	29,53%	838	4,17%	93,33%

Ud fra de krydstabeller som vi har set på frem til nu, er det tydeligt, at svarprocenten varierer meget fra kursus til kursus.
En anden vinkel er at se på uddannelserne, om der er nogle uddannelser, hvor de studerende generelt er bedre til at besvare deres kursusevalueringer end andre. Størrelserne på uddannelserne kan også variere meget, hvilket potentielt kan have en indflydelse på, hvor gode de studerende er til at få besvaret deres kursusevalueringer. I tabel 4 fremgår de fem uddannelser med den højeste gennemsnitlige svarprocent og de fem uddannelser med den laveste gennemsnitlige svarprocent.

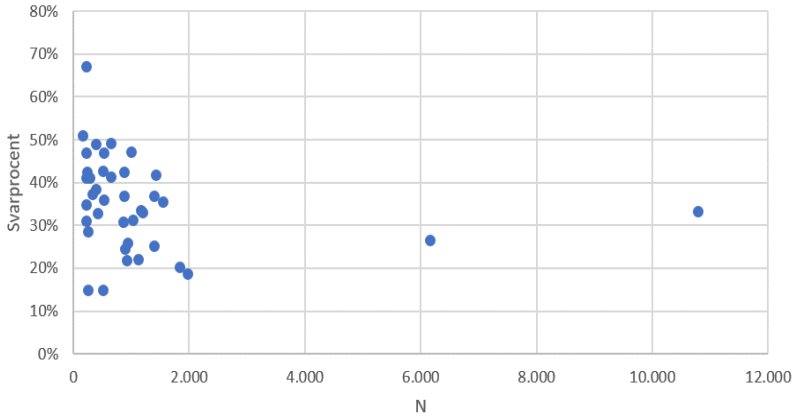
De fire uddannelser som befinder sig i toppen har en gennemsnitlig svarprocent på over 80%, hvilket er en flot svarprocent. Tabel 4 viser, at der er nogle uddannelser, som er bedre til at få besvaret deres kursusevalueringer end andre. Det kan derfor godt tænkes, at kulturen blandt uddannelserne har en indflydelse på, hvor gode de studerende er til at få besvaret deres kursusevalueringer. Ved at kigge på kolonnen "N" i tabel 4, som er antal udsendte kursusevalueringer for uddannelsen, kan det ses, at det ikke er de største uddannelser, som har den gennemsnitlige højeste svarprocent. Det fremgår nærmere, at de større uddannelser har tendens til at have lavere svarprocent. Det kan derfor også være, at størrelsen af uddannelsen har noget at gøre med, om de studerende får besvaret deres kursusevalueringer eller ej.

Tabel 4: De 5 uddannelser med højeste svarprocenter og de 5 med laveste

5 højeste svarprocent	Submitted	N	Svarprocent	Minimum	Maksimum
MSocScSEM	245	451	54,32%	26,32%	85,71%
MScBLC	810	1.674	48,39%	20,51%	87,50%
BScDMA	202	446	45,29%	20,00%	76,71%
BScISH	385	858	44,87%	19,57%	81,82%
MSocScPKL	207	485	42,68%	18,64%	62,12%
5 laveste svarprocent					
MScJUR	321	1.450	22,14%	12,50%	58,06%
BScJUR	699	3.879	18,02%	4,17%	59,87%
MScAUD	487	2.817	17,29%	7,94%	59,46%
MScMEC	60	424	14,15%	10,61%	31,82%
BAEOK	141	1.076	13,10%	6,12%	41,38%

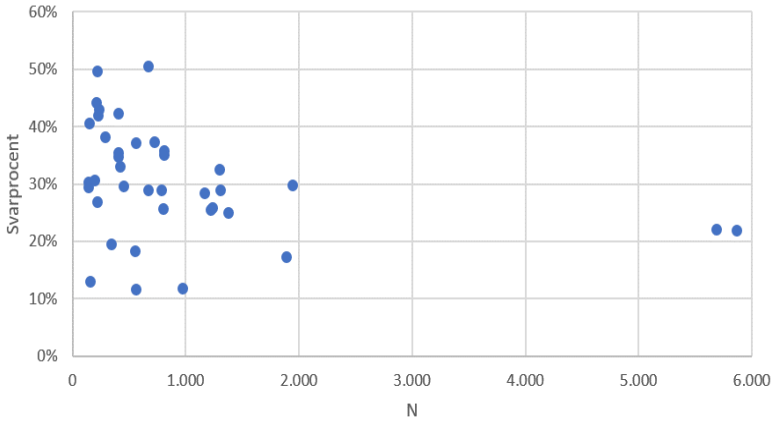
Nedenfor i figur 14 undersøger vi, om der er en sammenhæng mellem svarprocenten og antal udsendte kursusevalueringer per uddannelse, for E18. Det fremgår af figur 14, at der ikke ser ud til at være en sammenhæng mellem svarprocenten og antal udsendte evalueringer for E18. Det ses, at uddannelser med et mindre antal udsendte kursusevalueringer lige så vel kan have en svarprocent i den lave ende af skalaen på eksempelvis 10%, som i den høje ende af skalaen, på eksempelvis 50%. Det ligner dog, at de uddannelser som har haft mange udsendte kursusevalueringer har en svarprocent, der minder om den gennemsnitlige svarprocent omkring de 30%.

Figur 14: Svarprocent versus antal udsendte kursusevalueringer per uddannelse for E18



I figur 15 undersøger vi, om der eventuelt er en sammenhæng mellem antal udsendte kursusevalueringer og svarprocenten for F19. Begge figurer minder meget om hinanden, og det ser derfor ud til, at der generelt ikke er en sammenhæng mellem svarprocenten og antallet af udsendte kursusevalueringer.

Figur 15: Svarprocent versus antal udsendte kursusevalueringer per uddannelse for F19



En anden måde at anskue svarprocenten på er ved at se på de demografiske data omkring de studerende for at se, om der end tendens for, om der er nogle studerende, der generelt er bedre til at få besvaret deres kursusevalueringer end andre. Først og fremmest er det oplagt

at se på køn for at undersøge, om de kvindelige studerende er bedre til at få besvaret deres kursusevalueringer, end de mandlige studerende er eller analogt. I tabel 5 fremgår det, at kvinder generelt er lidt bedre til at få besvaret deres kursusevalueringer end de mandlige studerende. Ved at lave en chi-i-anden test kan vi teste nulhypotesen der siger at der ikke

Tabel 5: Svarprocent fordelt på kvinder og mænd

Køn	Submitted	N	Svarprocent
F	13.213	39.467	33,48%
M	10.535	40.964	25,72%
Total	23.748	80.431	29,53%

er afhængighed mellem, om kursusevalueringen bliver besvaret eller ej, og køn. Testen giver en p-værdi på $< 2,2e-16$, hvilket er $< 0,05$, og betyder at vi kan afvise nulhypotesen. Det betyder, at den studerendes køn har en betydning for, om den studerende får besvaret sin kursusevaluering eller ej.

Nogle af dagsuddannelserne på CBS er danske og bliver derfor primært undervist på dansk, mens andre uddannelser er engelske og bliver derfor kun undervist på engelsk. Der vil derfor være nogle uddannelser, hvor det primært er danske studerende samt studerende fra Norden, der er i stand til at følge med i dansk undervisning, der er tilmeldt. På de engelske uddannelser vil det i større grad være en blanding af danske og internationale studerende, som er tilmeldte, da undervisningssproget er på engelsk. Det er derfor også interessant at se, om nationaliteten har en indflydelse på, hvor gode de studerende er til at få besvaret deres kursusevalueringer. Variablen *Nationalitet* er opdelt i "Norden" og "Udenfor Norden" og i tabel 6 ses det, at de studerende "Udenfor Norden" er bedre til at få besvaret deres kursusevalueringer end dem fra "Norden". Ved at lave en chi-i-anden test, kan vi teste nulhypotesen der siger at der ikke er afhængighed mellem, om kursusevalueringen bliver besvaret eller ej, og den studeredes nationalitet. Testen giver en p-værdi på $< 2,2e-16$, og nulhypotesen kan dermed afvises. Det betyder at, om den studerende får besvaret sin kursusevaluering eller ej, afhænger af

om den studerende har nationalitet "Udenfor Norden" eller "Norden". Dette kan eventuelt hænge sammen med, at de studerende "Udenfor Norden" har betalt for deres uddannelse, og derfor er mere opmærksomme på at sørge for at få fortalt, hvordan de synes, kurset har været.

Tabel 6: Svarprocent fordelt på nationalitet

Nationalitet	Submitted	Hovedtotal	Svarprocent
Norden	17.889	64.755	27,63%
Udenfor Norden	5.859	15.676	37,38%
Total	23.748	80.431	29,53%

Det er også interessant at se, om der er en sammenhæng mellem de studerendes fødselsår og hvor hvor gode de studerende er til at få besvaret kursusevalueringen. Ved at kigge på tabel 7 kan det ses, at de studerende der er lidt yngre, er de studerende der er bedst til at få besvaret deres kursusevalueringer. Ved en chi-i-anden får vi en p-værdi på 0,034 og vi afviser derfor nulhypotesen om, at om den studerende får besvaret sin kursusevalueringer eller ej er uafhængigt af den studerendes fødselsår. Vores p-værdi er en del højere end for de to tidligere chi-i-anden test, og det er derfor mere diskutabelt om nulhypotesen kan afvises.

Tabel 7: Svarprocent og Fødselsår

Fødselsår	Submitted	N	Svarprocent
1958-1995	12.141	42.639	28,47%
1996-2018	11.607	37.792	30,72%
Total	23.748	80.431	29,53%

Efter at have skabt et overblik over hvordan svarprocenterne fordeler sig ud på de forskellige variable i vores i datasæt, er det nu interessant at se, hvordan de studerende så svarer på spørgsmålene, når de besvarer deres kursusevalueringer. For at få et overordnet billede af, hvordan de studerende svarer bestemmes middelværdien, medianen, standardafvigelsen og

variansen per spørgsmål og fremgår af tabel 8.

Tabel 8: Overordnet svarfordeling af spørgsmålene

Spørgsmål	Gennemsnit	Median	Std. Afv.	N
Kursus_S1	4,06	4	1,04	23.617
Kursus_S2	3,89	4	1,09	23.325
Kursus_S3	3,82	4	1,11	23.569
Kursus_S4	3,68	4	1,22	22.799
Kursus_S5	3,87	4	1,16	23.590
Pensum	3,14	3	0,61	23.748
Tidsforbrug	2,39	2	1,29	23.748
Fremmøde	4,11	4	0,98	23.559
Kontortid	1,10	1	0,35	23.559

Ved at se på besvarelserne til kursusspørgsmålene, ses et generelt højt gennemsnit, da svarskalaen går fra 1 til 5. Jo højere niveau på svarskalaen den studerende vælger, i jo højere grad er den studerende enig i udsagnet. Dette indikerer at de studerende er forholdsvis positive i forhold til de kursusrelaterede spørgsmål. Medianen for alle kursusspørgsmålene fremgår også at være 4, som er den anden højeste svarmulighed. Spørgsmålet omkring kontortid, er den variabel der stikker mest ud, da den har den laveste gennemsnitlige værdi, $1,10$, men dette indikerer blot, at de studerende generelt set ikke anvender tilbudet om kontortider, hvilket den tilhørende standardafvigelse på $0,35$ også bekræfter.
Variansen indikerer hvor store udsving der er i svarene. Variansen er størst på tidsforbrug, $1,66$, hvilket giver god mening, da det er op til den studerende, hvordan han/hun vælger at gribe uddannelsen an.
En generel tanke, når det er en svarskala der behandles, er at svarskalaens fortolkning er individuel. Nogle studerende kan eksempelvis opfatte værdien 5 som en oplevelse der er nærmest uopnåelig, hvorimens andre studerende kan opfatte værdien 5, som en oplevelse der blot er over gennemsnittet.

3 Teoretisk analyse af kategorisk data

I dette afsnit vil vi gennemgå teori vedrørende kategorisk data. Vi vil starte med at definere, hvad kategorisk data er. Efterfølgende gennemgår vi to forskellige klasser af regressionsmodeller, nemlig klassen af Generaliseret Lineære Modeller (GLM) og klassen af Generaliseret Lineære Mixed Modeller (GLMM), som kan anvendes til håndtering af kategorisk data. Den generelle teori vedrørende GLM og GLMM er byggestenene for de regressionsmodeller, som vi ønsker at anvende. Derefter vil vi dykke ned i, hvad en logistisk regressionsmodel med en binær responsvariabel er. Vi vil også se på, hvordan man estimerer, tester og prædikterer med den logistiske regressionsmodel. Efterfølgende vil vi se på en kumulative logit model, som er en model, der indeholder flere end to responsvariable. Først ser vi på en kumulativ logit model, som kun indeholder faste effekter, og efterfølgende ser vi på en kumulativ logit model, der også indeholder en tilfældig effekt. For begge modeller vil vi se på, hvordan der estimeres og testes.

3.1 Kategorisk Data

Denne opgave består af spørgeskemadata for kursusevalueringerne på CBS. Spørgeskemadata er en type af data, der ofte indeholder kategoriske variable. En kategorisk variabel består af en målbar skala, et eksempel i denne opgave er variablen $Resp_Status$, som angiver, hvorvidt kursusevalueringen er blevet besvaret, "Submitted", eller ej, "DNR"[7]. Et andet eksempel i denne opgave er svarskalaerne, der knytter sig til hvert enkelt spørgsmål, som indeholder 4 eller 5 kategorier, hvor følgende er et eksempel på 5 kategorier, hvor værdierne fra 1-5 betegner: "Helt uenig", "Overvejende uenig", "Hverken enig eller uenig", "Overvejende enig" og "Helt enig".

En kategorisk responsvariabel med to kategorier kaldes en binær responsvariabel, og en kategorisk responsvariabel med flere end to kategorier kaldes for en multinomial responsvariabel. Når vi har at gøre med en multinomial responsvariabel, skelnes der mellem to typer af

[7]DNR = Did not reply, hvilket vil sige at kursusevalueringen ikke er besvaret

kategoriske skalaer; ordinal og nominal. I det ordinale tilfælde har kategoriernes rækkefølge en betydning, det har de ikke i det nominale tilfælde (Agresti, 2015, s. 9).
Denne opgave beskæftiger sig udelukkende med den ordinale skala, da svarskalaerne er baseret på en "Likert-skala"[8], som måler styrken af holdninger til et udsagn i en bestemt rangorden og dermed har kategoriernes rækkefølge en betydning.

Hovedformålet med at analysere kategorisk responsdata er at modellere, hvad sandsynligheden er for at den ordinale responsvariabel Y er lig med kategori j ud af de J mulige kategorier, givet de forklarende variable x, som kan skrives som:

$$P(Y = j),\ j = 1, 2, ..., J$$

Den vigtigste model for kategorisk responsdata er en logistisk regressionsmodel. En logistisk regressionsmodel med faste effekter, er en af mange regressionsmodeller i klassen af Generaliserede Lineære Modeller, (GLM) (Agresti, 2013, s. 163).

3.2 Generaliserede Lineære Modeller

En Generaliseret Lineær Model (GLM) er en model, der kan tage højde for ikke-normale respons fordelinger og mulige ikke-lineære funktioner af middelværdien og er således en udvidelse af de Lineære Normal Modeller [9]. En GLM kan dermed tage højde for kategorisk data, som ikke har en normal respons fordeling. Fælles for alle modeller i klassen af GLM'er er, at de består af tre komponenter. De tre komponenter er følgende:

- Den første komponent er den tilfældige komponent, som specificerer responsvariablen, Y, og dens sandsynlighedensfordeling. Her er Y en variabel, der indeholder n uafhængige observationer, $Y = (y_1, ..., y_n)^T$.

[8] Definition af en likert skala: "En sociologisk skala til måling af holdninger. Forsøgspersoners grader af samtykke/modvilje i forhold til en række påstande gives hver en værdi, som udtrykker en holdnings styrke." (Wulff, 2009)

[9] Den lineære regressionsmodel beskriver forholdet mellem middelværdien af responsvariablen og et sæt af forklarende variable, med inferens der antager normalfordeling af responsvariablen (Agresti, 2007, s. 2)

- Komponent nummer to er den systematiske komponent, η, som specificerer parameteret, β, og de forklarende variable x, gennem deres linearkombination ved deres lineære prædiktor. Her er $\eta = (\eta_1, ..., \eta_n)^T$, β indeholder p forklarende variable, hvor $j = 1, ..., p$, så $\beta = (\beta_1, ..., \beta_p)^T$, og de forklarende variable, X, indeholder værdierne x_{ij}. X er en designmatrice, som er (n x p), så designmatricen har n rækker, en for hver observation, og p søjler, en for hvert parameter i β. Den systematiske komponent på vektorform er da givet ved:

$$\eta = \beta^T X$$

- Den tredje komponent er link funktionen g, som er linket mellem den tilfældige komponent og den systematiske komponent. Link funktionen fortæller, hvordan den forventede værdi af responsvariablen relaterer sig til den lineære prædiktor for de forklarende variable:

$$g[E(Y)] = \eta$$

De ovenstående tre komponenter medfører nogle forskellige forhold, som vi vil komme ind på nu. Elementerne $(y_i, ..., y_n)$ i Y antages at være uafhængige og identisk fordelte med en fordeling, der tilhører en eksponentiel familie. De mest anvendte fordelinger for en eksponentiel familie er; normalfordelingen, binomialfordelingen og poissonfordelingen. Ved at begrænse GLM'er til de eksponentielle familiers fordelinger opnås et generelt udtryk for likelihood ligningerne, den asymptotiske fordeling af estimaterne for model parametrene og en algoritme til at fitte modellerne.
Den systematiske komponent medfører, at det ikke er nødvendigt at antage en lineær sammenhæng mellem responsvariablen, Y, og de forklarende variable, X, da den lineære prædiktor er et udtryk for, at de er lineære i parametrene. GLM'er opfatter Y som tilfældig og X som fast, og netop derfor bliver den lineære prædiktor også kaldet for den systematiske komponent. Det gælder for det tredje komponent, link funktionen, at det er en monoton, differentiabel funktion.

I de eksponentielle familiers repræsentation af fordelingen, fungerer et bestemt parameter som det *naturlige parameter*. For normalfordelingen er denne parameter middelværdien, for binomialfordelingen er det "log af oddset" og for poissonfordelingen er det "log af middelværdien". Link funktionen g som transformerer $E[Y]$ til dets *naturlige parameter*, kaldes for det kanoniske link. Denne link funktion, som er lig med det naturlige parameter med den lineære prædiktor, genererer de mest anvendte GLM'er. Når en GLM bruger den kanoniske link funktion gælder der simple egenskaber, for eksempel haves en konkav log-likelihood funktion og likelihood ligningerne (Agresti, 2015, s. 3).

De tre komponenter af GLM'en er nu på plads, og vi ser nu nærmere på den tilfældige komponent og fordelingen af denne, som følger de eksponentielle familiers fordelinger, og disse har et generelt udtryk for momenter og for likelihood ligningerne.

Eksponentielle dispersionsfamilier

Følgende afsnit følger kapitel 4 i bogen "Foundations of Linear and Generalized Linear Models" af Alan Agresti. Den tilfældige komponent i GLM'en specificerer, som tidligere nævnt, responsvariablen, Y, som har uafhængige og identisk fordelte observationer, fra en fordeling, der har tæthedsfunktion for y_i på formen:

$$f(y_i;\theta_i,\phi) = exp\left\{\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)}\right] + c(y_i,\phi)\right\} \tag{3.2.1}$$

Ovenstående tæthedsfunktion kaldes for den eksponentielle dispersionsfamilie. Her er θ_i det *naturlige parameter* og ϕ er *dispersionsparameteren*. Ofte er dispersionsparameteren $a(\phi) = 1$, hvilket medfører at det sidste led $c(y_i,\phi) = c(y_i)$. Hvis dette er tilfældet er det en *naturlige eksponentiel familie* og har da formen $f(y_i;\theta_i) = h(y_i)exp[y_i\theta_i - b(\theta_i)]$. Ellers har $a(\phi)$ ofte formen $a(\phi) = \phi$ eller $a(\phi) = \frac{\phi}{\omega_i}$ for $\phi > 0$ og en kendt vægt, ω_i. For eksempel er $\omega_i = n_i$ når y_i er middelværdi for n_i uafhængige observationer.
Forskellige valg af $a(\cdot)$ og $b(\cdot)$ giver anledning til forskellige fordelinger, som eksempelvis binomialfordelingen, som benyttes i denne opgave. Udtrykkene for $E(y_i)$ og $var(y_i)$ bruger

mængder fra ligning 3.2.1.

Ved at lade $L_i = log\ f(y_i; \theta_i, \phi)$ være bidraget fra y_i til log-likelihood funktionen, $L = \sum_i\ L_i$, da er:

$$L_i = \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi), \tag{3.2.2}$$

Ved at bestemme den første afledte og den anden afledte af log-likelihood funktionen 3.2.2 fås:

$$\frac{\partial L_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a(\phi)}, \qquad \frac{\partial^2 L_i}{\partial \theta_i^2} = \frac{-b''(\theta_i)}{a(\phi)}$$

Hvor $b'(\theta_i)$ og $b''(\theta_i)$ henholdvis betegner den første og den anden afledte af $b(\cdot)$ evalueret i θ_i. Nu anvendes de generelle likelihood resultater til at bestemme:

$$E\left(\frac{\partial L}{\partial \theta}\right) = 0 \quad \text{og} \quad -E\left(\frac{\partial^2 L}{\partial \theta^2}\right) = E\left(\frac{\partial L}{\partial \theta}\right)^2 \tag{3.2.3}$$

Som holder under regularitetsbetingelserne, som er opfyldt af den eksponentielle dispersionsfamilie.
Fra den første ligning i ovenstående ligning 3.2.3 med en enkelt observation fås:

$$\frac{E[y_i - b'(\theta_i)]}{a(\phi)} = 0, \quad \text{således at } \ \mu_i = E(y_i) = b'(\theta_i). \tag{3.2.4}$$

Fra den anden ligningen i ligning 3.2.3 fås:

$$\frac{b''(\theta_i)}{a(\phi)} = E\left[\frac{(y_i - b'(\theta_i))}{a(\phi)}\right]^2 = \frac{var(y_i)}{[a(\phi)]^2},$$

Dermed er:

$$var(y_i) = b''(\theta_i)a(\phi). \tag{3.2.5}$$

Således bestemmer funktionen $b(\cdot)$ i ligning 3.2.1 momenterne af y_i og kaldes for en kumulant funktion, fordi når $a(\phi) = 1$, giver dens afledte kumulanterne af fordelingen.

GLM med binomialfordelingen

I denne opgave er det binomialfordelingen, fra de eksponentielle dispersionsfamilier, der fokuseres på, da vi blandt andet er interesseret i at lave en logsitisk regressionsmodel. Derfor vil vi nu se nærmere på udtryk for middelværdien og variansen for binomialfordelingen.

Når vi har at $n_i y_i$ har en $bin(n_i, \pi_i)$-fordeling, hvor y_i betegner andelen af succes, så kan vi nøjes med at se på $E(y_i) = \pi_i$, som dermed ikke afhænger af n_i. Lad $\theta_i = log[\frac{\pi_i}{1-\pi_i}]$. Så er $\pi_i = c\frac{exp(\theta_i)}{1+exp(\theta_i)}$ og $log(1-\pi_i) = -log[1+exp(\theta_i)]$. Tætheden kan nu skrives op:

$$\begin{aligned} f(y_i; \pi_i, n_i) &= \binom{n_i}{n_i y_i} \pi_i^{n_i y_i}(1-\pi_i)^{n_i - n_i y_i}, \qquad y_i = 0, \frac{1}{n_i}, \frac{2}{n_i}, ..., 1, \\ &= exp\left[\frac{y_i\theta_i - log[1+exp(\theta_i)]}{\frac{1}{n_i}} + log\binom{n_i}{n_i y_i}\right]. \end{aligned} \tag{3.2.6}$$

Denne tæthed har samme form, som den eksponentielle dispersionsfamilie i ligning 3.2.1 med $b(\theta_i) = log[1+exp(\theta_i)]$, $a(\phi) = \frac{1}{n_i}$ og $c(y_i, \phi) = log\binom{n_i}{n_i y_i}$. Det naturlige parameter er

$\theta_i = log\left[\frac{\pi_i}{1-\pi_i}\right]$, som er *logit*.

Ved at anvende ligning 3.2.4 og ligning 3.2.5 fås da for binomialfordelingen:

$$E(y_i) = b'(\theta_i) = \frac{exp(\theta_i)}{[1 + exp(\theta_i)]} = \pi_i, \tag{3.2.7}$$

$$var(y_i) = b''(\theta_i)a(\phi) = \frac{exp(\theta_i)}{[1 + exp(\theta_i)]^2 n_i} = \frac{\pi_i(1-\pi_i)}{n_i}. \tag{3.2.8}$$

Den eksponentielle dispersionsfamilie er således vist ved binomialfordelingen og et udtryk for middelværdien og variansen, er fundet (Agresti, 2015, s.122).

Som tidligere nævnt forbinder link funktionen i en GLM den tilfældige komponent med den lineære prædiktor. Link funktionen g, som tranformerer middelværdien til det *naturlige parameter* θ_i i ligning 3.2.1, kaldes for det kanoniske link. Fra den netop udledte eksponentielle dispersionsfamilie, er den kanoniske link funktion *logit link* for binomialfordelingen. Her vil det for poissonfordelingen være "*log link*" og for normalfordelingen "*identity link*".
Den kanonise link funktion er oftest den, der bliver valgt som standard, hvilket er at foretrække, da der gælder fordelagtige statistiske egenskaber ved brug af det kanoniske link, herunder brug af Newton Raphson og Fisher scoring til at Maksimum Likelihood Estimaterne (MLE) (Agresti, 2015, s.123).

Likelihood ligningerne for GLM

Det ønskes nu at opnå et generelt udtryk for likelihood ligningerne for en GLM. For n uafhængige observationer fra ligning 3.2.2 haves det, at log-likelihood er:

$$L(\beta) = \sum_{i=1}^{n} L_i = \sum_{i=1}^{n} logf(y_i; \theta_i, \phi) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + \sum_{i=1}^{n} c(y_i, \phi). \tag{3.2.9}$$

Notationen $L(\beta)$ reflekterer afhængigheden af θ på modelparametrene β. Til at differentiere log-likelihood, ligning 3.2.9, bruges kædereglen, og det leder os til likelihood ligningerne for en GLM:

$$\frac{\partial L(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(y_i - \mu_i) xij}{var(y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0, \qquad j = 1, 2, ..., p, \tag{3.2.10}$$

her er $\eta_i = \sum_{j=1}^{p} \beta_j x_{ij} = g(\mu_i)$ for link funktionen g. For de forskellige link funktioner, for de forskellige eksponentielle familier, leder link funktionen til forskellige sæt af ligninger.

Overdispersion for binomialfordelingen i GLM

Likelihood ligningen 3.2.10 for en GLM afhænger af en antaget fordeling af Y_i kun gennem μ_i og $v(\mu_i)$. Valget af fordelingen bestemmer da middelværdi-varians forholdet; $v(\mu_i)$. Quasi-likelihood estimation antager kun et middelværdi-varians forhold, i stedet for en specifik fordeling af Y_i. Den har den samme link funktion og samme lineære prædiktor som en GLM, men i stedet for at antage en fordelingstype for Y_i antager den kun:

$$var(Y_i) = v(\mu_i)$$

for en valgt variansfunktion v. Dermed har vi, at quasi-likelihood funktionen kan håndtere overdispersion. Den simple quasi-likelihood tilgang anvender dermed en alternativ variansfunktion:

$$var(\pi_i) = \frac{\phi \pi_i (1 - \pi_i)}{n_i}$$

Overdispersion sker, når $\phi > 0$. Quasi-likelihood estimaterne er det samme som maksimum likelihood estimatet (MLE) for binomialfordelingen, siden ϕ ryger ud af ligning 3.2.10 (Agresti, 2013, s. 151).

3.2.1 Estimation

Likelihood ligningerne er oftest ikke-lineære i $\hat{\beta}$. Dette afsnit beskriver en generel iterativ metode til at løse ikke-lineære ligninger, for på den måde at finde maksimum likelihood estimatet (MLE). Der findes to iterative metoder, Newton Raphson og Fischer Scoring. I denne opgave anvender vi Newton Rapshon, og derfor lægger vi mest vægt på denne.

Newton Raphson metoden

Newton Raphson metoden løser ikke-lineære ligninger iterativt, og kan anvendes til at bestemme det punkt, hvor funktionen tager sit maksimum. Newton Raphson metoden starter med en initial approksimation for løsningen. Den opnår en anden approksimation ved at approksimere funktionen i nærheden af den initiale approksimation ved et andengradspolynomium og finder så punktet for polynomiets maksimum værdi. Metoden gentager så dette step for at generere en sekvens af approksimationer, som konvergerer mod punktet af maksimum, når funktionen er passende, og/eller den initiale approksimation er god. Matematisk bestemmer Newton Raphson metoden værdien $\hat{\beta}$, for hvilken, en funktion $L(\beta)$ er maksimeret. Lad:

$$u = \left(\frac{\partial L(\beta)}{\partial \beta_1}, \frac{\partial L(\beta)}{\partial \beta_2}, ..., \frac{\partial L(\beta)}{\partial \beta_p} \right)^T \tag{3.2.11}$$

Her betegner H hessematricen, som har indgang $h_{ab} = \frac{\partial^2 L(\beta)}{\partial \beta_a \partial \beta_b}$. Lad $u^{(t)}$ og $H^{(t)}$ være u og H evalueret ved $\beta^{(t)}$, approksimation t for $\hat{\beta}$. Step t i den iterative proces ($t = 0, 1, 2, ...$) approksimerer $L(\beta)$ nær $\beta^{(t)}$ for udtrykket op til anden orden af dens Taylor serie udvikling:

$$L(\beta) \approx L(\beta^{(t)}) + u^{(t)T}(\beta - \beta^{(t)}) + \left(\frac{1}{2}\right)(\beta - \beta^{(t)})^T H^{(t)}(\beta - \beta^{(t)})$$

Ved at løse $\frac{\partial L(\beta)}{\partial \beta} \approx u^{(t)} + H^{(t)}(\beta - \beta^{(t)}) = 0$ for β, giver den næste approksimation:

$$\beta^{(t+1)} = \beta^{(t)} - (H^{(t)})^{-1}u^{(t)} \tag{3.2.12}$$

under antagelse af, at $H^{(t)}$ er ikke-singulær. Iterationerne fortsætter indtil ændringerne $L(\beta^{(t)})$ mellem successive cyklusser er tilstrækkelig små. MLE er grænsen af $\beta^{(t)}$, da $t \to \infty$, dog er dette ikke nødvendigt, hvis $L(\beta)$ har andre lokale maksima hvor $u(\beta) = 0$. Hvis der er andre lokale maksima, er en god initial aproksimation afgørende.

For mange GLM'er, inklusiv den binomiale logistiske regression, med en fuld-rang model matrice, er hessematricen negativ definit, og log-likelihood funktionen er en streng konkav funktion. Da vil MLE af modellens parametre eksistere og være unikke under generelle betingelser (Agresti, 2015, s. 139).

Fisher Scoring

Den anden iterative metode til at finde maksimum af likelihood ligningerne er Fisher scoring, som vi ikke vil gå i dybden med i denne opgave. Dog er det værd at nævne, at forskellen fra Newton Raphson ligger i den måde, hessematricen bruges. Fisher Scoring bruger den forventede værdi af hessematricen, kaldet den forventede information, hvor Newton Raphson bruger selve hessematricen kaldet den observerede information (Agresti, 2015, s. 139).

3.2.2 Statistisk inferens

Likelihood funktionen kan bruges til inferens for en GLM, og der er tre standardmetoder til at gøre dette. Generelt testes der på parameteret β, hvor nulhypotesen $H_0 : \beta = \beta_0$[10] testes op i mod den alternative hypotese $H_1 : \beta \neq \beta_0$. Der kan dertil bestemmes et konfidensinterval, som kan hjælpe med at visualisere usikkerheden for testen, i form af et interval (Agresti 2015,

[10]Her er β_0 en bestemt nul-værdi og altså ikke intercept parameteret, som det til forkveksling også kunne være

s. 128). Nedenfor gennemgåes de tre forskellige standardmetoder en efter en.

Likelihood Ratio Test

Likelihood ratio test er en testmetode, der bruger likelihood funktionen ved hjælp af ratioen af dens værdi ℓ_0 ved β_0 og dens maksimum ℓ_1 over β værdierne, som tillader enten H_0 eller H_1 at være sand. Ratioen $\Lambda = \ell_0/\ell_1 \leq 1$, da ℓ_0 kommer fra at maksimere en begrænset β værdi. Likelihood ratio test statistikken skrives på formen:

$$-2log\Lambda \;=\; -2log(\ell_0/\ell_1) \;=\; -2(L_0 - L_1) \tag{3.2.13}$$

Her betegner L_0 og L_1 de maksimerede likelihood funktioner. Under regularitetsbetingelserne har den en begrænset $nul - \chi^2$ fordeling, da $n \to \infty$, med 1 frihedsgrad.

Likelihood ratio testen kan udvides til at indeholde flere parametre. De χ^2-fordelte frihedsgrader er lig med differencen i dimensionerne af parameterrummet under $H_0 \cup H_1$ og under H_0, som har dimension $dim\beta_0$, når modellen er parametriseret til at være identificerbar. Likelihood ratio testen tester således goodness-of-fit for to konkurrerende statistiske modeller baseret på ratioen af hver deres likelihood ligning (Agresti, 2015, s. 128).

Wald test

Wald testen er en testmetode, der er baseret på opførslen af log-likelihood funktionen ved ML estimatet $\hat{\beta}$.

Fra den inverse informationsmatrix fås standardfejlen af de ukendte parameterværdier. Når det ubegrænsede ML estimat erstattes opnåes et estimatet på standardfejlen (SE) af $\hat{\beta}$. For nulhypotesen $H_0 : \beta = \beta_0$ er teststatistikken, der bruger ikke-nul estimerede standardfejl:

$$z = \frac{(\hat{\beta} - \beta_0)}{SE} \tag{3.2.14}$$

som kaldes for "*Wald statistikken*". z har en approksimeret standardfordeling, når $\beta = \beta_0$ og z^2 har en approksimeret χ^2-fordeling med 1 frihedsgrad. Hvis der haves flere parametre; $\beta = (\beta_0, \beta_1)$ til at teste nulhypotesen $H_0 : \beta_0 = 0$ er Wald χ^2 statistikken følgende:

$$\hat{\beta_0^T}[\hat{var}(\hat{\beta}_0)]^{-1}\hat{\beta}_0, \tag{3.2.15}$$

hvor $\hat{\beta}_0$ er det ubegrænsede ML estimat for β_0, og $\hat{var}(\hat{\beta}_0)$ er en samling af den ubegrænsede estimerede kovariansmatrice af $\hat{\beta}$ (Agresti, 2015, s. 129).

Score test

Score test er en tredje inferens metode, som anvender "score statistikken". Score testen bruger hældningen, score funktionen, og den forventede krumning af log-likelihood funktionen, evalueret ved nul-værdien β_0. χ^2 formen af score statistikken er følgende:

$$\frac{[\frac{\partial L(\beta)}{\partial \beta_0}]^2}{-E[\frac{\partial^2 L(\beta)}{\partial \beta_0^2}]}, \tag{3.2.16}$$

hvor notationen reflekterer de afledte i forhold β, som er evalueret i β_0. I tilfældet med flere parametre er score statistikken på kvadratisk form, baseret på vektoren af de partielt afledte af log-likelihood ligningen og den inverse informations matrice begge evalueret ved H_0 estimaterne (Agresti, 2015, s. 130).

Konfidensintervaller

For hver af de tre standardmetoder ovenfor er det muligt at konstruere et konfidensinterval ved at invertere testen. I tilfældet med et enkelt parameter er et 95% konfidensinterval for β, sættet af β_0 for hvilken, testen af nulypotesen $H_0 : \beta = \beta_0$ har en p-værdi $> 0,05$. Her betegner z_a $(1-a)$-kvartilen af normalfordelingen. Et $100(1-a)\%$ konfidensinterval baseret på asymptotisk normalitet bruger $z_{\alpha/2}$.

Wald konfidensintervallet er sættet af β_0 for hvilken $\frac{|\hat{\beta}-\beta_0|}{SE} < z_{\alpha/2}$, hvilket giver intervallet $\hat{\beta} \pm z_{\alpha/2}(SE)$. Konfidensintervallet for score testen simplificerer ofte intervallet af sættet af β_0 for hvilken $\frac{|\hat{\beta}-\beta_0|}{SE_0} < z_{\alpha/2}$, hvor SE_0 er standardfejlen estimeret under restriktionen at $\beta = \beta_0$. Lad $\chi^2_d(a)$ betegne den $(1-a)$-kvartil af χ^2-fordelingen med d frihedsgrader, da er likelihood ratio test konfidensintervallet et sæt af β_0 for hvilken $-2[L(\beta_0) - L(\hat{\beta})] < \chi^2_1(\alpha)$.

Når $\hat{\beta}$ har en normalfordeling, da er log-likelihood funktionen en andengradspolynomium. For små samples eller ikke-normal data eller når β er nær grænsen af parameterrummet, da kan fordelingen af $\hat{\beta}$ være langt fra normal. En betydelige forskel i resultaterne af wald og likelihood ratio test angiver, at fordelingen af $\hat{\beta}$ er langt fra normal. Det vil dermed være at foretrække eksempelvis at anvende likelihood ratio test metoden (Agresti, 2015, s. 131).

Profil likelihood konfidensintervaller er at foretrække, når vi har en model med flere parametre. Disse konfidensintervaller er baseret på at invertere likelihood ratio testen for forskellige mulige nul-værdier af β og betragter de andre parametre $\mathbf{\Psi}$ i modellen som nuisance parametre, og er derfor ikke af interesse. Når man inverterer en likelihood ratio test for nulhypotesen $H_0 : \beta = \beta_0$ for at tjekke, om β_0 tilhører konfidensintervallet, da varierer ML estimatet $\hat{\mathbf{\Psi}}(\beta_0)$ af $\mathbf{\Psi}$, som maksimerer likelihood under nul-værdien, på samme måde som β_0 gør. Profil likelihood funktionen er $L(\beta_0, \hat{\mathbf{\Psi}}(\beta_0))$, vist som en funktion af β_0. For hver β_0 giver denne funktion maksimum, af den ordinære log-likelihood, i forhold til betingelsen $\beta = \beta_0$. Evalueret ved $\beta_0 = \hat{\beta}$ er den maksimerede log-likelihood $L(\hat{\beta}, \mathbf{\Psi})$, som forekommer for det ubegrænsede ML estimat. Profil likelihood konfidensintervallet for β er da sættet af β_0 for hvor:

$$-2[L(\beta_0, \hat{\mathbf{\Psi}}(\beta_0)) - L(\hat{\beta}, \hat{\mathbf{\Psi}})] < \chi^2_1(\alpha).$$

Intervallet indeholder alle β_0, som ikke er afvist i likelihood ratio testen af den nominale størrelse α (Agresti, 2015, s. 132).

3.2.3 Model selektion

Backwards elimination versus forward selection

Modelvalg for en GLM bliver mere kompliceret i takt med at antallet af forklarende variable øges, hvilket skyldes den hurtige stigning i mulige effekter og interaktioner. Modelvalgsprocessen har to konkurrerende formål; modellen skal være kompleks nok til at fitte data ordentligt. Men på den anden side skal modellen udglatte, frem for at overfitte data og ideelt set også være relativt simpel at fortolke. Der findes to trinvise procedure til modelvalg, "backwards elimination" og "forward selection". Backwards elimination starter ud med en kompleks model og fjerner sekventielt de forklarende variable en efter en. Ved hvert stadie vælges den forklarende variable, der vil være mindst skadende for modellen at udelade. Dette kan eksempelvis vurderes ud fra den højeste p-værdi ved test af signifikans. Backwards elimination processen stopper, når yderligere fjernelse af forklarende variable vil lede til en forværring af modellen, målt ud fra dens fit, eksempelvis ved AIC.

Forward selection starter med en simpel model og tilføjer så de forklarende variable sekventielt. For hvert stadie vælges den forklarende variabel der giver den bedste forbedring af modellens fit. Processen stopper, når yderligere tilføjelse af variable ikke forbedrer modellens fit, målt eksempelvis på AIC (Agresti, 2015, s. 143).

AIC

Til at udvælge den bedst mulige model er den mest velkendte metode Akaike information citerion (AIC). AIC vurderer en model på, hvor tæt dens fittede værdier er på de sande forventede værdier, summeret ved en given distance mellem de to. Den optimale model er den model, der har sine fittede værdier tættest på de sande udfaldssandsynligheder. Den optimale model er således den, der minimerer:

$$AIC = -2(Loglikelihood - antallet\ af\ parametre\ i\ modellen) \tag{3.2.17}$$

Selvom formålet med at trække antallet af parametre fra i modellen er at korrigere for bias, ender AIC med at "straffe" en model for at have mange parametre. Selvom en simpel model

er længere fra den sande tilstand end en mere kompleks model, kan den simple model muligvis give bedre estimater for det sande forventede output. Når en model har mange forklarende variable, kan AIC hjælpe til udvælgelsen af variable. Ud af et sæt af potentielle modeller identificerer vi modellen med lavest AIC, eller alternativt en eller flere af de simple modeller med AIC tæt på minimum, som vores endelige model (Agresti, 2015, s. 146).

3.3 Generaliseret Lineær Mixed Modeller

Indtil nu har vi set på modelklassen af GLM'er, som indeholder faste effekter, β. En fast effekt betegner alle kategorier, vi ønsker at se på, såsom køn, nationalitet og fødselsår. Vi vil nu se på modelklassen Generaliseret Lineær Mixed Modeller (GLMM), som indeholder både faste effekter, β, og tilfældige effekter, u_i. Ved at tilføje en tilfældig effekt bliver det muligt at tage højde for klyngedannelser i data. Tillægsordet "mixed" i GLMM refererer til tilstedeværelsen af både faste effekter og tilfældige effekter i den lineære prædiktor.
De tilfældige effekter gælder oftest for en sample, hvor u_i refererer til en klynge i i en sample. I tilfælde hvor populationen, n, er stor, vil det at behandle u_i som en fast effekt øge antallet af parametre i modellen. Ved i stedet at tilføje den som en tilfældig effekt, behandler vi den i stedet som en random sample fra populationen, og dette finder vi interessant. Observationer i en klynge er normalt positivt korreleret og tenderer til at være mere ens end observationer fra andre klynger (Agresti, 2007, s. 298).

En GLMM indeholder ligesom en GLM 3 komponenter, dog er der en forskel i den lineære prædiktor, som er komponent nummer 2, da en GLMM tillader både faste og tilfældige effekter i modellen.
Vi lader y_{it} betegne observation t i klynge i. Da lader vi x_{it} være værdien af de forklarende variable for den pågældende observation. Det ligner således en GLM bortset fra betingelsen u_i. Lad $\mu_{it} = E(Y_{it}|u_i)$, altså middelværdien af responsvariablen for den givne værdi af den tilfældige effekt. Med link funktionen $g(\cdot)$ gør følgende sig gældende for en GLMM:

$$g(\mu_{it}) = x_{it}^T\beta + z_{it}^T u_i \tag{3.3.1}$$

Den tilfældige effekt behandles som en tilfældig variabel og antages at have en normal $N(0, \sigma)$-fordeling, med ukendte parametre. Dens varians, σ^2, refererer vi til som varianskomponenten (Agresti, 2013, s. 519).

3.3.1 Estimation

At fitte en GLMM er ikke lige til, da likelihood funktionen ikke har en lukket form. En GLMM er derfor en to-trins model. I første trin, betinget af den tilfældige effekt u_i, antages observationerne at følge en GLM. Det betyder, at observationerne er uafhængige med y_{it} i klynge i, som har en fordeling der følger den eksponentielle familie med den forventede værdi μ_{it} linket til en lineær prædiktor som i ligning 3.3.1. Ved første trin er $z_{it}u_i$ et kendt udgangspunkt. Dernæst i andet trin antages u_i at være uafhængigt fra en $N(0, \Sigma_u)$-fordeling.

For en diskret variabel betegnes vektoren af observationer med $\mathbf{y}$ og vektoren med de tilfældige effekter med $\mathbf{u}$. Lad da $f(\mathbf{y}|\mathbf{u}; \boldsymbol{\beta})$ antage den betingede tæthedsfunktion af $\mathbf{y}$ givet $\mathbf{u}$. Det haves at $f(\mathbf{u}; \boldsymbol{\Sigma})$ betegner den normale sandsynlighedstæthedsfunktion for $\mathbf{u}$. Likelihood funktionen $\ell(\boldsymbol{\beta}, \boldsymbol{\Sigma}; \mathbf{y})$ for en GLMM er sandsynlighedsmassefunktionenen $f(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\Sigma})$ af $\mathbf{y}$ vist som en funktion af $\boldsymbol{\beta}$ og $\boldsymbol{\Sigma}$. Denne tæthedsfunktion refererer til den marginale fordeling af $\mathbf{y}$ efter at have integreret den tilfældige effekt ud:

$$\ell(\boldsymbol{\beta}, \boldsymbol{\Sigma}; \mathbf{y}) = f(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \int f(\mathbf{y}|\mathbf{u}; \boldsymbol{\beta}) f(\mathbf{u}; \boldsymbol{\Sigma}) d\mathbf{u} \tag{3.3.2}$$

Ligning 3.3.2 kaldes ofte for marginal likelihood. Flere metoder kan evaluere den marginale likelihood numerisk og maksimere den som en funktion af $\boldsymbol{\beta}$ og $\boldsymbol{\Sigma}$. En af metoderne som kan anvendes er Gauss-Hermite quadrature metoden (Agresti, 2013, s. 519).

Gauss-Hermite quadrature metoden

Metoden Gauss-Hermite quadrature kan anvendes til at fitte Maksimum likelihood funktionen for en GLMM. Integralet der bestemmer likelihood funktionens dimension afhænger af den tilfældige effekts struktur. Når dimensionen af den tilfældige effekt, u_i, er lav, som i det en-dimensionale integrale ovenfor i ligning 3.3.2 kan numeriske standard integrationsmetoder bruges til at approksimere likelihood funktionen.

Gauss-Hermite quadrature er en metode, der kan approksimere integralet af en funktion $h(\cdot)$ multipliceret med en normal tæthedsfunktion. Approksimationen er en endelig vægtet sum, som evaluerer funktionen ved bestemte punkter, i bund og grund approksimeres arealet under en kurve ved et diskret histogram. I et tilfælde med en univariat normal tilfældig effekt da vil approksimationen have følgende form:

$$\int_{-\infty}^{\infty} f(u)exp(-u^2)du \approx \sum_{k=1}^{q} c_k f(s_k), \tag{3.3.3}$$

som har vægte c_k og quadrature punkter (s_k), som er tabuleret. Approksimationen forbedres i takt med q, antallet af quadrature punkter, øges. Den approksimerede likelihood kan da maksimeres ved eksempelvis Newton Raphson, se ligning 3.2.12, som giver ML estimaterne for $\hat{\boldsymbol{\beta}}$ og $\hat{\boldsymbol{\Sigma}}$ (Agresti, 2013, s. 520).

3.3.2 Statistisk inferens

Når en GLMM er fittet gælder de samme 3 standardmetoder for inferens, som står beskrevet i afsnit 3.2.2 for de faste effekter i GLMM'en, som for en GLM. Hovedfokus i GLMM inferens er vedrørende de faste effekter. Inferens for de tilfældige effekter, altså inferens om den tilfældige effekts varianskomponenter, er mere kompleks. Det kan dog være relevant af lave inferens omkring varianskomponenten σ_u af u_i, som beskriver variationen blandt klyngerne. En type inferens kan være at sammenligne modellen med særtilfældet hvor $\sigma_u = 0$. Den mere simple model vil falde i grænsen af parameterrummet relativ til den mere komplekse model, siden σ_u ikke kan være negativ. Når dette sker, er den normale likelihood ratio χ^2-test

til at sammenligne modeller ikke valid. På samme vis har Wald statistikken heller ikke en approksimeret standard normal nul fordeling. For at teste nulhypotesen $H_0 : \sigma_u = 0$ op imod den alternative hypotese $H_1 : \sigma_u > 0$ for en tilfældig effekt model, har den asymptotiske nul fordeling af likelihood ratio testen sandsynlighed $1/2$ ved 0 og sandsynlighed $1/2$ for at følge formen af en χ^2-fordeling med 1 frihedsgrad. Test statistik værdien 0 opstår når $\hat{\sigma_u}$, og i det tilfælde vil maksimum af likelihood funktionen være identisk under H_0 og H_1. Når $\hat{\sigma}_u > 0$ og den observeret test statistik er lig med t, da vil p-værdien for denne large-sample test være $\frac{1}{2}P(\chi_1^2 > t)$, som er halvdelen af p-værdien, der anvendes for en χ_1^2 asymptotisk test (Agresti, 2010, s. 288).

3.3.3 Modelselektion

Samme modelselektionsmetoder gør sig gældende for en GLMM, som for en GLM, som står beskrevet i afsnit 3.2.3. Der findes dermed "Backwards elimination" og "Forward selection", som kan anvendes til udvælgelse af den bedste mulige model. Ligeledes gør det, der står beskrevet i GLM afsnittet 3.2.3 om AIC sig også gældende for en GLMM.

3.4 Logistisk regression med binær responsvariabel

Ved gennemgang af modelklassen GLM blev det understreget at i denne opgave, er det binomialfordelingen fra de eksponentielle dispersionsfamilier, vi fokuserer på. Dette skyldes, at vi ønsker at lave en logistisk regressionsmodel. Vi vil nu se på, hvordan en logistisk regressionsmodel med binær responsvariabel ser ud.

En model for en logistisk regression, også kaldet for en logit-model er en statistisk model, der modellerer sandsynligheden eller log-odds for, at en bestemt begivenhed indtræffer. En logistisk regression kan beskrive sammenhængen mellem en kategorisk responsvariabel, Y, og et antal forklarende variable, x. Ved en binær logistisk regression har responsvariablen, Y, to kategorier, hvilket betyder, at der er to mulige udfald, som ofte bliver omtalt som "succes" og "fiasko". Det ønskes at finde sandsynligheden for, at en af de to udfald finder sted på

baggrund af de forklarende variable, x. For at opstille en model for en logistisk regression sættes x til at være et sæt af forklarende variable, og de associerede sandsynligheder sættes til at være $\pi(x)$, hvor:

$$\pi(x) = P(Y = 1|X = x) = 1 - P(Y = 0|X = x)$$

Modellen for en logistisk regression er da:

$$\pi(x) = \frac{exp(\alpha + \beta x)}{1 + exp(\alpha + \beta x)}$$

hvilket svarer til log-odds, som er det samme som logit, og har følgende lineære sammenhæng:

$$logit[\pi(x)] = log\frac{\pi(x)}{1 - \pi(x)} = \alpha + \beta x$$

Ved at tage exp på begge sider af lighedstegnet kan det ses, at oddset er en eksponentiel funktion af x. Dette giver en simpel fortolkning af graden af β, for da er oddset ganget med $exp(\beta)$ for hver 1-enheds øgning i x. Så $exp(\beta)$ er en odds ratio. En odds ratio for modellen angiver dermed et forhold mellem oddset af det ene sker ved tilstedeværelse af noget andet.

Modellen indeholder parametrene α og β. Parameteren α også kaldet interceptet er ikke af særlig interesse, men parameteren β's fortegn bestemmer, hvorvidt $\pi(x)$ stiger eller falder, når x øges. Når $\beta = 0$ haves det, at responsvariablen Y er uafhængig af de forklarende variable x. (Agresti, side 164).

Hvis der er kvalitative prædiktorer, også omtalt som faktorer, bruger vi indikatorvariable til at håndtere dette. En måde at opstille en model, hvor indikator variable repræsenterer en

faktor er ved at lade variablen $x_i = 1$ for observationerne i række i og $x_i = 0$ for resten, for $i = 1, ..., I - 1$. Modellen er da:

$$logit(\pi_i) = \alpha + \beta_1 x_1 + \beta_2 x_2 + .. + \beta_{I-1} x_{I-1} \tag{3.4.1}$$

Her haves parameter redundans ved ikke at forme en indikator variabel for kategori I. Den kategori der ekskluderes for indikatorvariablen er vilkårlig og omtales som værende den kategori, der sættes som baseline (Agresti, 2013, s. 176).

En logistisk regressionsmodel kan også indeholde flere forklarende variable, hvor de forklarende variable kan være en blanding af kvalitative og kvantitative variable. Modellen for $\pi(x) = P(Y = 1)$ ved værdierne $x = (x_1, ..., x_p)$ for p forklarende variable er da:

$$logit(\pi(x)) = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$$

Som også kan skrives som:

$$\pi(x) = \frac{exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)}{exp(1 + \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)}$$

For de kvalitative forklarende variable anvendes indikator variable for kategorierne. Her refererer parameteren β_j til effekten af x_j på log odds for at $Y = 1$ ved at holde de resterende forklarende variable konstant. (Agresti, 2013, s. 182)

3.4.1 Estimation

For at estimere parametrene i den logistiske regressionsmodel anvendes et specialtilfælde af GLM estimationsresultaterne fra afsnit 3.2.1. Med n objekter behandles de n binære svar som uafhængige. Lad $x_i = (x_{i0}, x_{i1}, ..., x_{ip})$ betegne i af værdierne for de p forklarende variable

og lad $x_{i0} = 1$ betegne interceptet, $i = 1, .., N$. En logistisk regressionsmodel, ligning 3.4.1, hvor interceptet α behandles som en regressions parameter, β_0, for en forklarende variabel der altid summer til 1, er:

$$\pi(x_i) = \frac{exp(\sum_{j=0}^{p} \beta_j x_{ij})}{1 + exp(\sum_{j=0}^{p} \beta_j x_{ij})} \tag{3.4.2}$$

Når vi har flere observationer for en fast x_i-værdi, da er det tilstrækkeligt kun at se på antallet af observationerne, n_i, og antallet af "succeser". Nu betegner y_i dermed kun mængden af "succes". Da haves det, at $\{Y_i, ..., Y_N\}$ er uafhængige binomier med $E(Y_i) = n_i\pi(x_i)$, hvor $n_1 + ... + n_N = n$. Den samlede tæthedsfunktion er proportional med produktet af N binomial funktioner:

$$\begin{aligned}
&\prod_{i=1}^{N} \pi(x_i)^{y_i}[1 - \pi(x_i)]^{n_i - y_i} \\
&= \left\{ \prod_{i=1}^{N} exp\left[log\left(\frac{\pi(x_i)}{1 - \pi(x_i)} \right)^{y_i} \right] \right\} \left\{ \prod_{i=1}^{N} [1 - \pi(x_i)]^{n_i} \right\} \\
&= \left\{ exp\left[\sum_{i=1}^{N} y_i log \frac{\pi(x_i)}{1 - \pi(x_i)} \right] \right\} \left\{ \prod_{i=1}^{N} [1 - \pi(x_i)]^{n_i} \right\}
\end{aligned}$$

For model 3.4.2 er den i'te logit lig med $\sum_j \beta_j x_{ij}$ og dermed er det eksponentielle led lig med $exp[\sum_i y_i(\sum_j \beta_j x_{ij})] = exp[\sum_j (\sum_i y_j x_{ij})\beta_j]$. Siden at $[1 - \pi(\mathbf{x}_i)] = [1 + exp(\sum_j \beta_j x_{ij})]^{-}1$, da er log-likelihood ligningen lig med:

$$L(\beta) = \sum_j \left(\sum_i y_i x_{ij} \right) \beta_j - \sum_i n_i log \left[1 + exp\left(\sum_j \beta_j x_{ij} \right) \right] \tag{3.4.3}$$

Ved at differentiere likelihood ligningerne i forhold til β og sætte ligningen lig med 0 da fås:

$$\sum_i y_i x_{ij} - \sum_i n_i \hat{\pi}_i x_{ij} = 0 \ for \ j = 0, 1, ..., p \tag{3.4.4}$$

Siden

$$\frac{L(\beta)}{\beta_j} = \sum_i y_i x_{ij} - \sum_i n_i x_{ij} \frac{exp(\sum_k \beta_k x_{ik})}{1 + exp(\sum_k \beta_k x_{ik})}$$

Her er $\hat{\pi}_i = \frac{exp(\sum_k \hat{\beta} x_{ik})}{[1+exp(\sum_k \hat{\beta} x_{ik})]}$ MLE for $\pi(x_i)$. Disse ligninger er ikke-lineære og kræver iterative løsninger. En af de iterative løsninger som vil kunne klare dette er Newton Raphson metoden (Agresti, 2013, s. 193).

Newton Raphson metoden

Newton Raphson metoden er beskrevet i afsnit 3.2.1 for den generelle GLM og anvendes nu for en logistisk regression. Lad:

$$u_j^{(t)} = \left.\frac{\partial L(\beta)}{\partial \beta_j}\right|_{\beta^{(t)}} = \sum_i (y_i - n_i \pi_i^{(t)}) x_{ij}$$

$$h_{ab}^{(t)} = \left.\frac{\partial^2 L(\beta)}{\partial \beta_a \partial \beta_b}\right|_{\beta^{(t)}} = -\sum_i x_{ia} x_{ib} n_i \pi_i^{(t)} (1 - \pi_i^{(t)})$$

Her er $\pi^{(t)}$, approksimationen t for $\hat{(}\pi)$, som er opnået fra $\beta^{(t)}$ gennem:

$$\pi_i^{(t)} = \frac{exp(\sum_{j=1}^p \beta_j^{(t)} x_{ij})}{1 + exp(\sum_{j=1}^p \beta_j^{(t)} x_{ij})} \tag{3.4.5}$$

Her bruges $u^{(t)}$ og $H^{(t)}$ med ligning 3.2.12 for at opnå næste værdi $\beta^{(t+1)}$, som i denne kontekst er:

$$\beta^{(t+1)} = \beta^{(t)} + \{X^T Diag[n_i \pi_i^{(t)}(1 - \pi_i^{(t)})] X\}^{-1} X^T (y - \mu^{(t)}) \tag{3.4.6}$$

Hvor $\mu_i^{(t)} = n_i \pi_i^{(t)}$. Dette bruges til at opnå π^{t+1} og så videre. Med et initialt gæt $\beta^{(0)}$ giver ligning 3.4.5 $\pi^{(0)}$ og, for $t > 0$ fortsætter iterationerne som beskrevet ved at bruge

ligning 3.4.6 og ligning 3.4.5. Ved grænsen konvergerer $\pi^{(t)}$ og $\beta^{(t)}$ mod ML estimatet $\hat{\pi}$ og $\hat{\beta}$ (Agresti, 2013, s. 194).

3.4.2 Statistisk inferens og modelselektion

Efter at have fittet modellen ønskes det at teste de faste effekter, altså parameteren β. Til dette er det muligt at anvende de 3 standardmetoder likelihood ratio test, Wald test og score testen, som vi har været omkring i GLM afsnittet 3.2.2.

Som det også står beskrevet i afsnit 3.2.2, er det muligt at bestemme et tilhørende konfidensinterval for β ved at invertere en test ved nulhypotesen $H_0 : \beta = \beta_0$ (Agresti, 2013, s.170).

Der er i praksis ingen garanti for, at en logistisk regressionsmodel vil passe data perfekt. For en hvilken som helst type af binær data er en måde at tjekke manglen af fit på, at anvende likelihood ratio test, som kan sammenligne en mere simpel model med en mere kompleks model, som det også står beskrevet i afsnit 3.2.2. En mere kompleks model kan være en, som indeholder ikke-lineære effekter, flere prædiktorer eller eventuelt interaktion blandt prædiktorer. Hvis de mere komplekse modeller ikke fitter bedre, da giver det en sikring om, at den simple model er rimelig. (Agresti, 2013, s. 172). Til modelselektion, altså til at vælge den bedste mulige model, da kan backwards eliminaiton eller forward selection anvendes samt AIC, som det står beskrevet i afsnit 3.2.3.

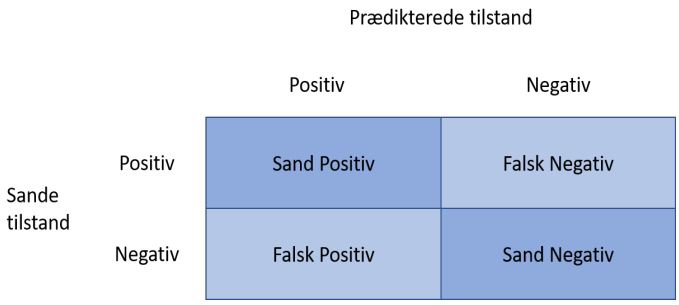
3.4.3 Prædiktion

Ved prædiktion kan en models performance vurderes. Et typisk scenarie er, at en model har to mulige udfald, hvilket også står beskrevet i afsnittet om den logistiske regressionsmodel, hvor udfaldene oftest omstales som "succes" og "fiasko". Til at prædiktere anvendes et datasæt, hvor udfaldene er observeret med de valgte variable, og ved at bygge en model baseret på dette, kan modellen forsøge at prædiktere udfaldene for nye, uobservererede objekter (Hastie et al., 2017, s. 2). Efterfølgende kan en ROC[11]-kurve analyse anvendes til at bestemme

[11]ROC står for "Receiver operating characteristic"

modellens evne til at prædiktere nye uobserverede observationer.
Ved at udføre en ROC-kurve analyse kan der laves en diagnostik over, hvor meget der prædikteres positivt, og hvor meget der prædikteres negativt, for en given cutoff-værdi. En cutoff-værdi angiver, hvor sikker prædiktionen skal være. Hvis en måling af et objekt er <cutoff-værdien, da er prædiktionen for objektet negativ, og hvis målingen for et objekt er $\geq$ cutoff-værdien, da er prædiktionen for dette objekt positiv.
Det gælder dog ikke nødvendigvis, at prædiktionen for det pågældende objekt er sandt, blot fordi den er positiv. Det kan derfor siges, at der findes fire mulige udfald af prædiktionen for et objekt: "Sand Positiv", "Sand Negativ", "Falsk Positiv" og "Falsk Negativ". Dette danner en klassifikationstabel, som fremgår af figur 16.

Figur 16: Klassifikationstabel



Hver enkel prædiktion for hver enkelt objekt falder ind i én af de fire mulige udfald, som fremgår af tabel 16. Forskellige rater kan da bruges til at beskrive klassifikationstabellen mere overordnet. Der findes både rater, som er baseret på de sande tilstande, på de prædikterede tilstande samt rater der er for hele tabellen. Følgende rater er for den sande tilstand:

- "True Positive" Raten (TPR) er andelen af objekter, som vi ved er positive, der bliver prædikteret positive, dermed er: TPR = $\frac{Sand\ Positiv}{Sand\ Positiv\ +\ Falsk\ Negativ}$
- "True Negative" Raten (TNR) er andelen af objekter, som vi ved er negative, der bliver prædikteret negative, dermed er: TNR = $\frac{Sand\ Negativ}{Falsk\ Positiv+Sand\ Negativ}$
- "False Negative" Raten (FNR) er andelen af objekter, som vi ved er positive, men som bliver prædikteret negative, dermed er: FNR = $\frac{Falsk\ Negativ}{Sand\ Positiv+Falsk\ Negativ}$

- "False Positive" Raten (FPR) er andelen af objekter, som vi ved er negative, men som bliver prædikteret positive, dermed er: FPR = $\frac{Falsk\ Positiv}{Falsk\ Positiv+Sand\ Negativ}$

ROC-kurven har TPR op af y-aksen og FPR ud af x-aksen og plottes for alle, mulige cutoff-værdier, så hvert enkelt punkt på ROC-kurven repræsenterer en mulig cutoff-værdi. Punkterne i ROC-kurven forbindes, så de former en kurve. Cutoff-værdier der resulterer i en lav FPR tenderer til et resultat mod TPR, altså langs y-aksen. Som TPR øges, vil FPR øges. Jo bedre diagnostik af prædiktionen, jo hurtigere vil TPR nærme sig 1, altså 100%. En næsten perfekt prædiktion vil da have en ROC-kurve, der er mere eller mindre vertikal fra (0,0) til (0,1) og så horisontal til (1,1), altså en kurve der går mod det øvre venstre hjørne. Diagonallinjen er en reference linje, som indikerer tilfældigt klassificeret objekter.

En måde at angive hvor præcis prædiktionen er, er ved at se på værdien af "Area under the ROC curve" (AUC), som altså er området under ROC-kurven. Jo højere AUC, jo bedre performance. De mulige værdier for AUC går fra 0,5, som indikerer, at det er helt tilfældigt diagnosticeret, og op til 1, som indikerer, at det er perfekt diagnosticeret. AUC'en angiver dermed den gennemsnitlige TPR for alle mulige FPR.
Værdien af AUC beregnes ved at summere områderne, som er formet under de forbunde cutoff-punkter, der danner ROC kurven (NCSS Statistical Software, 2021, s. 3). Ved at definere det T_1 komponent af det i'te objekt altså $V(T_{1i})$ som:

$$V(T_{1i}) = \frac{1}{n_0} \sum_{j=1}^{n_0} \Psi(T_{1i}, T_{0j})$$

og definer det T_0'te komponent af det j'te objekt, altså $V(T_{0i})$, som:

$$V(T_{0j}) = \frac{1}{n_1} \sum_{i=1}^{n_1} \Psi(T_{1i}, T_{0j}),$$

hvor det haves at:

$$\Psi(X,Y) = 0 \; hvis \; Y > X$$
$$\Psi(X,Y) = \frac{1}{2} \; hvis \; Y = X$$
$$\Psi(X,Y) = 1 \; hvis \; Y < X$$

da kan den empiriske AUC estimeres som (DeLong et al., 1988, s. 839):

$$AUC = \sum_{i=1}^{n_1} \frac{V(T_{1i})}{n_1} = \sum_{j=1}^{n_0} \frac{V(T_{0j})}{n_0} \tag{3.4.7}$$

3.5 Kumulativ logit model

Frem til nu har vi kun set på en responsvariabel, som indeholder to kategorier, altså en binær responsvariablen. Hvis vi i stedet har en multinomial responsvariabel, er der behov for at udvide den logistiske regressionsmodel til at kunne tage højde for dette. I vores opgave ser vi kun på regressionsmodeller med multinomial responsvariable, hvor skalaen er ordinal. En populær type af regressionsmodeller specielt egnet til ordinal data er de kumulative link modeller (CLM'er). En kumulativ link model er en regressionsmodel, hvis observationer har en ordinal skala, altså observationer der falder i et ordnet, endeligt, sæt af kategorier. En kumulativ link model kan dermed udnytte, at der er en rangorden i responsvariablen. Der findes 5 standard link funktioner; "logit", "probit", "log-log", "clog-log" og "cauchit" (Chrstiensen, 2015, s. 5).

De ordinale observationer repræsenteres af en tilfældig variabel Y, hvor den antager værdi j, hvis den i'te ordinale observation falder i den j'te kategori, da haves at $j = 1, ..., J$ og $J \geq 2$. Hvis $J = 2$ vil sige, at der er to kategorier, og da haves en binær logistisk regressionsmodel. Y følger en multinomial fordeling med parameter π, hvor π_j betegner sandsynligheden for at den

i'te observation falder i svarkategori j. Dermed haves det, at de kumulative sandsynligheder er givet ved:

$$P(Y \leq j|\mathbf{x}) = \pi_1(\mathbf{x}) + ... + \pi_j(\mathbf{x}) \ for \ j = 1, ..., J.$$

Denne opgave fokuserer på "logit" link funktionen, og den er defineret som: $logit(\pi(\mathbf{x})) = log(\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})})$ og da er de kumulative logits defineret som:

$$logit[P(Y \leq j|\mathbf{x})] = log\frac{P(Y \leq j|\mathbf{x})}{1 - P(Y \leq j|\mathbf{x})} = log\frac{\pi_1(\mathbf{x}) + ... + \pi_j(\mathbf{x})}{\pi_{j+1}(\mathbf{x}) + ... + \pi_J(\mathbf{x})}, \ j = 1, ..., J-1,$$

hvor $P(Y \leq j|\mathbf{x})$ er den kumulative sandsynlighed for Y er mindre end eller lig med en specifik kategori $j = 1, ..., J-1$, i stedet for at modellere responskategoriernes sandsynlighed enkeltvis. Generelt kan responsvariablen Y have værdierne fra 1 til J, så hver kumulative logit bruger alle J kategorier. Der kan dermed formes $J-1$ logits, og den ordinale model inkorporerer alle $J-1$ logits i en model. Modellen er dermed defineret for alle kategorier bortset fra den sidste kategori (Agresti, 2013, s. 301).

Den kumulative logit model er en udvidelse af den logistiske regressionsmodel, og dens form minder derfor meget om den logistiske regressionsmodels, som står beskrevet i afsnit 3.4.
En kumulativ logit model som bruger alle (J-1) kumulative logits, har følgende form:

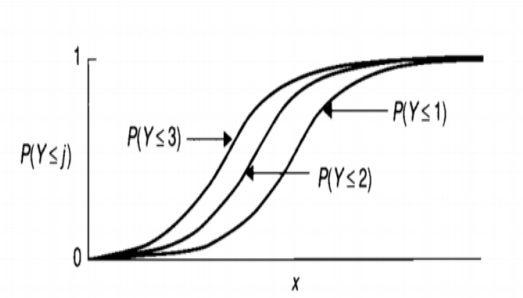
$$logit[P(Y \leq j|\mathbf{x})] = \alpha_j + \boldsymbol{\beta}^T\mathbf{x}, \ j = 1, ..., J-1. \tag{3.5.1}$$

I modellen er $\mathbf{x}$ en vektor med forklarende variable, og $\boldsymbol{\beta}$ er det tilhørende regressionsparameter. Parameteren α_j angiver hver enkelte kumulative logit med dets eget intercept, for hvert j. Det haves dermed at regressionsdelen, altså $\boldsymbol{\beta}^T\mathbf{x}$, er uafhængig af j og derfor har $\boldsymbol{\beta}$ den

samme effekt på hver af de $J-1$ logits, og interceptet er for sig selv i den lineære prædiktor. Hældningen er dermed den samme, men horisontalt starter de forskellige steder. Det haves at α_j stiger, når j stiger, idet at sandsynligheden $[P(Y \leq j|\mathbf{x})]$ vokser i j for et fast $\mathbf{x}$, og logit er en stigende funktion af $[P(Y \leq j|\mathbf{x})]$.
Parameteren $\boldsymbol{\beta}$ siger noget om i hvor høj grad den tilhørende variabel $\mathbf{x}$ påvirker responsvariablen. Det haves, at parameteren $\boldsymbol{\beta}$ angiver den partielle effekt på responsen, når variablen øges med 1, og alle andre variable holdes konstant. Hvis $\mathbf{x}$ stiger med 1 enhed, vil responsvariablen alt andet lige stige eller falde med $\boldsymbol{\beta}$. Jo større $\boldsymbol{\beta}$, desto større betydning har den tilhørende forklarende variabel for kategorien af responsvariablen. Da β har den samme effekt for hver af de kumulative logits, vil det sige, at estimatet $\boldsymbol{\beta}$ er det samme, uanset hvilken responskategori respondenten falder i. Denne antagelse kaldes for "proportional odds antagelsen" (Agresti, 2013, s. 301).

På figur 17 ses en model med 4 kategorier i responsvariablen og $\mathbf{x}$. Det ses, at hver kumulative sandsynlighed har sin egen kurve for $j = 1,\ 2,\ 3\ og\ 4$, som beskriver ændringen som en funktion af $\mathbf{x}$. Kurven for $P(Y \leq j)$ ligner en logistisk kurve for en binær respons med udfaldet $(Y \leq j)$ og $(Y > j)$. Den fælles effekt $\boldsymbol{\beta}$ for hvert j gør, at kurverne har den samme form. Kurverne er dermed identiske men blot forskudte til enten venstre eller højre. Jo større forskellen $\alpha_k - \alpha_j$ er, for en given værdi af $\boldsymbol{\beta}$, jo større vil den horisontale distance mellem kurverne for de to kumulative sandsynligheder være. Det er $|\boldsymbol{\beta}|$ som bestemmer, hvor hurtigt kurven stiger eller falder. For et fast $\mathbf{x}$ har kurverne den samme ordning, som de kumulative sandsynligheder, hvor den for $P(Y \leq 1)$ er lavest (Agresti, 2015, s. 210).

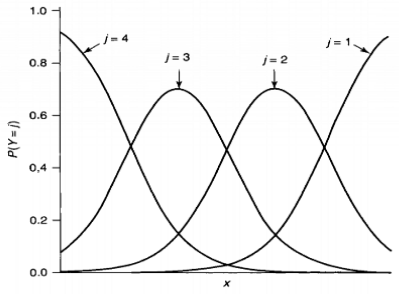
Figur 17: Plot af kumulative sandsynligheder



Note: Figuren er taget fra (Agresti, 2015, s. 210)

Figur 18 viser de tilhørende kurver for sandsynligheden for at falde i de forskellige kategorier, $P(Y = j) = P(Y \leq j) - P(Y \leq j - 1)$. Som $\mathbf{x}$ øges, er sandsynligheden for at responsen Y falder i en lavere ende af den ordinale skala mere sandsynlig. Den identiske form for de forskellige kurver viser, at fordelingen af Y til forskellige værdier af $\mathbf{x}$ er stokastisk ordnede. Hvis $\boldsymbol{\beta} > 0$ vil hver kumulative logit øges, som $\mathbf{x}$ øges, hvilket betyder, at $P(Y \leq j)$ også øges. Dette indikerer, at den betingede Y's fordeling er stokastisk mindre ved højere niveauer af $\mathbf{x}$. Hvis $\boldsymbol{\beta} < 0$ vil den betingede Y's fordeling være stokastisk højere ved højere niveauer af $\mathbf{x}$. Hvis $\boldsymbol{\beta} = 0$ vil grafen have en horisontal linje for hver kumulative sandsynlighed og da vil X og Y være statistisk uafhængige (Agresti, 2013, s. 301).

Figur 18: Sandsynlighed for kategorierne i en proportional odds version af en kumulativ logit model



Note: Figuren er taget fra (Agresti, 2007,s. 181)

Nogle vil opskrive modellen med et $'-'$ i stedet for en $'+'$ på højre side af lighedstegnet, som der er gjort i ligning 3.5. Ved at anvende et $'-'$ vil det betyde, at jo højere værdien er af $\boldsymbol{\beta}^T\mathbf{x}$, jo højere er sandsynligheden for at svaret falder i en kategori af den øvre del af svarkategorien (Agresti, 2010, s.49).
De forklarende variable i den kumulative logit model kan være kvantitative, kategoriske med en indikator variabel, eller en blanding af begge typer (Agresti, 2007, s. 182).

Kumulativ odds ratio

Odds ratioen (OR) for at begivenheden $Y \leq j$ ved x_1 relativ til den samme begivenheden ved x_2 er:

$$OR = logit[P(Y \leq j|\mathbf{x}_1) - logit[P(Y \leq j|\mathbf{x}_2) = log\frac{P(Y \leq j|\mathbf{x}_1)/P(Y > j|\mathbf{x}_1)}{P(Y \leq j|\mathbf{x}_2)/P(Y > j|\mathbf{x}_2)} = \boldsymbol{\beta}^T(\mathbf{x}_1 - \mathbf{x}_2)],$$

som er uafhængig af j. En odds ratio af kumulative sandsynligheder kaldes for en kumulativ odds ratio. Det haves at $\mathbf{x}_1$ og $\mathbf{x}_2$ er to niveauer af en given forklarende variabel. Odds ratioen for en begivenhed $\mathbf{x}$ er, når $Y \leq j$, $exp(\boldsymbol{\beta})$. Odds ratio er proportional med differensen mellem $\mathbf{x}_1$ og $\mathbf{x}_2$ med $\boldsymbol{\beta}$ som konstant, og idet at $\boldsymbol{\beta}$ knytter sig til hvert enkelt kumulative logit, betegnes modellen som "proportional odds" modellen. Fordelen ved en proportional odds model er, at med den samme effekt for forskellige kumulative logits, er effekten nem at opsummere og fortolke (Agresti, 2013, 302). Det er muligt at finde et konfidensinterval for odds ratioen ved at tage grænserne af konfidensintervallet for $\boldsymbol{\beta}$, hvilket vil lede til asymmetriske konfidensinterval for odds ratio (Christensen, 2015, s. 8).

Den kumulativ logit model med en latent variabel

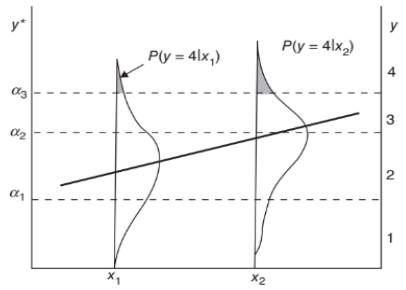
Den kumulative logit model kan udvides til at indeholde en underliggende kontinuerlig latent variabel, S, med en kumulativ fordelingsfunktion. En latent variabel er en uobserveret variabel, der antages at underligge, hvad der faktisk bliver observeret. Den underliggende variabel motiverer den samme effekt $\boldsymbol{\beta}$ for forskellige j i proportional odds modellen. Den

ordinale responsvariabel Y er da observeret i kategori j, hvis S_i er i mellem skæringskoefficienterne $\alpha_{(j-1)} < S_i < \alpha_j$, hvor det antages at skæringskoefficienterne er af den kontinuerlige skala:

$$-\infty \equiv \alpha_0 \leq \alpha_1 \leq ... \leq \alpha_{J-1} \leq \theta_J \equiv \infty$$

Det fremgår på figur 19, at når S_i falder i kategori j, så falder den latente variabel i det j'te interval af værdier.

Figur 19: Ordinal model med latent variabel



Note: Figur er taget fra (Agresti, 2015, s. 211)

Når $J = 4$ vil der være 3 skæringskoefficienter, α_1, α_2, og α_3, som vil opdele området under kurven i 4 dele, hvor hver del repræsenterer sandsynligheden for, at et svar falder i en af de 4 svarkategorier. Selve skæringskoefficienten er fast på skalaen, men lokationen af den latente fordeling, og dermed også de 4 dele under kurven, ændres med x_i (Agresti, 2015, s. 211).

3.5.1 Estimation

Koefficienterne i den lineære kombination kan estimeres ved Maximum Likelihood Estimation (MLE). For objekt i lad $(Y_{i1}, ..., Y_{iJ})$ være binære indikatorer for svaret, hvor $y_{ij} = 1$ for kategori j for hvilken svaret falder i, når $Y_i = j$ når $y_{ij} = 1$ og $y_{ik} = 0$ for $k \neq j$. Det haves at x_i betegner værdien for de forklarende variable for objekt i. Lad $\pi_j(x_i)$ betegne $P(Y_i =$

$j|X = x_i)$. For uafhængige observationer er likelihood funktionen baseret på produktet af de multinomiale tæthedsfunktioner for de n objekter. Da er den multinomiale likelihood funktion for den kumulative logit model følgende:

$$\prod_{i=1}^{n}\left[\prod_{j=1}^{J}\pi_j(x_i)^{y_{ij}}\right] = \prod_{i=1}^{n}\prod_{j=1}^{J}[P(Y \leq j|x_i) - P(Y \leq j-1|x_i)]^{y_{ij}} =$$
$$\prod_{i=1}^{n}\prod_{j=1}^{J}\left[\frac{exp(\alpha_j + \beta^T x_i)}{1 + exp(\alpha_j + \beta^T x_i)} - \frac{exp(\alpha_{j-1} + \beta^T x_i)}{1 + exp(\alpha_{j-1} + \beta^T x_i)}\right]^{y_{ij}}$$

Likelihood funktionen er en funktion af (α_j, β). Log-likelihood funktionen betegnes med $L(\alpha_j, \beta)$. Hver likelihood ligning opnås ved at differentiere L i forhold til parametrene og sætte ligningen lig med 0. For at simplificere dette betegnes:

$$G(z) = \frac{exp(z)}{1 + exp(z)},\ g(z) = \frac{exp(z)}{[1 + exp(z)]^2}$$

Da vil likelihood ligningen for en effekt parameter β_k være lig med:

$$\sum_{i=1}^{n}\sum_{j=1}^{c} y_{ij}x_{ik}\frac{g(\alpha_j + \beta^T x_i) - g(\alpha_{j-1} + \beta^T x_i)}{G(\alpha_j + \beta^T x_i) - G(\alpha_{j-1} + \beta^T x_i)} = 0$$

Disse ligninger er ikke-lineære i parametrene og har derfor ikke en lukket form. Likelihood funktionerne maksimeres for at opnå MLE ved brug af iterative metoder, hvor Newton Raphson metoden kan anvendes (Agresti, 2010, s. 59). Denne står beskrevet i afsnit 3.2.1.

3.5.2 Statistisk inferens og modelselektion

Efter at have fittet modellen ønskes det at teste modelparametrene, altså de faste effekter β. Til dette er det muligt at anvende de 3 standardmetoder likelihood ratio test, Wald test og

score testen, som vi har været omkring i GLM afsnittet 3.2.2.

Det er også muligt at bestemme et tilhørende konfidensinterval for odds ratio af de kumulative sandsynligheder. Når sample størrelsen er lille, eller når en stor procentdel af observationerne falder i den højeste eller laveste kategori af responsvariablen, da er det bedre at bruge likelihood ratio test af konfidensintervaller baseret på profil likelihood funktionen (Agresti, 2010, s. 61), som står beskrevet i afsnit 3.2.2.

Test af proportional odds antagelsen For den kumulative logit model antages det, at den har form som en proportional odds model. En mulig test er da at teste, hvorvidt denne antagelse holder. For at teste dette, er det muligt at lave en score test af proportional odds antagelsen. Denne test vil da teste, om effekterne er de samme for hver kumulative sandsynlighed. Dette kan gøres ved at sammenligne en model, som har samme β for hvert j, med en mere kompleks model der har separate β_j for hvert j. Hvilket betyder, at proportional odds modellen:

$$logit[P(Y \leq j)] = \alpha + \beta^T x, \ j = 1, ..., J-1, \tag{3.5.2}$$

som har en parameter for hver prædiktor, sammenlignes med den mere komplekse model:

$$logit[P(Y \leq j)] = \alpha_j + \beta^T x, \ j = 1, ..., J-1. \tag{3.5.3}$$

Med en enkelt prædiktor har proportional odds versionen, ligning 3.5.2 et β, og den mere komplekse model, ligning 3.5.3 har $J-1$ parametre; β_j.
Når en model generaliseres til at have separate parametre, da haves det, at kurverne ikke er parallelle for de forskellige logits, de kan derfor godt krydse for nogle x-værdier. Denne model vil dermed overtræde rangordenen blandt de kumulative sandsynligheder. Modellen

med proportional odds formen antyder, at fordelingen af Y for en prædiktor tenderer til at være enten ens, højere, lavere end fordelingen af Y for en anden prædikteret værdi (Agresti, 2007, s. 185). Hvis det viser sig, at sådan en model ikke fitter godt, da kan der anvendes en mere generel model med separate effekter for forskellige kumulative sandsynligheder. Her vil det så være muligt, at kurvene for forskellige kumulative sandsynligheder kan stige eller falde ved forskellige rater, men så vil disse kurver krydse ved bestemte prædiktor værdier. Dette er dog strengt taget upassende for modellen, da det overtræder ordningen af de kumulative sandsynligheder, som de må have, altså $P(Y \leq 2) \leq P(Y \leq 3)$. Så ved at bruge proportional odds modellen sikres det, at de kumulative sandsynligheder har den passende ordning for alle prædiktor værdierne (Agresti, 2007, s. 184)

Ud fra test statistikken fra sammenligning af de to modeller kan det vurderes, om der er evidens for "lack of fit" eller ej. Score testen kan bruges til at teste, om effekten er den samme for hver kumulative logit, i mod den alternative med seperate effekter. Testen sammenligner modellen med en parameter for x_1 og en for x_2 til den mere komplekse model med to parametre for hver, hvor det tillades at have forskellige effekter for $logit[P(Y \leq 1)]$ og $logit[P(Y \leq 2)]$. Hvis score statistikkens tilhørende p-værdi er under $0,05$, vil den mere komplekse model være signifikant bedre, og hvis den er over $0,05$, da vil den mere komplekse model ikke passe signifikant bedre. Hvis score testen har en lille p-værdi, kan det være en idé at tjekke, om bruddet på proportional odds antagelsen er væsentlig. Dette kan tjekkes ved at sammenligne estimater opnået fra det separate logistiske fit til den binære sammenlægning af responsen. Hvis en proportional odds model passer dårligt, findes der andre strategier for at forsøge at få modellen til at passe bedre. En strategi kan være at tilføje et led yderligere, såsom interaktioner til den lineære prædiktor, eller det kan være at prøve en link funktion, hvor responskurven er ikke-symmetrisk, altså en komplementær eller log-log link-funktion. (Agresti, 2013, s. 307)

Goodness-of-fit test

En anden mulig test for den kumulative logit model er at tjekke tilstrækkeligheden af modellens fit. Dette kan gøres ved en global goodness-of-fit test. Det er muligt at lave en

goodness-of-fit test af nulhypotesen: at modellen holder, i mod den alternative hypotese: at modellen ikke holder. Den alternative hypotese siger, at modellen fitter data perfekt. Test statistikken sammenligner det observerede antal i mod det forventede frekvens estimater, baseret på modellens fit. For en særlig opsætning af x_i af de forklarende variable, hvor den observerede multinomiale sample har n_i observationer, lad da $n_{ij},\ j = 1, ..., J$ betegne de observerede optællinger for de J kategorier. Under nulhypotesen at modellen holder, er de korresponderende forventede frekvens estimater baseret på model-estimaterne for $P(Y = j|x_i)$ lig (Agresti, 2010, s. 67):

$$\hat{\mu}_{ij} = n_i \hat{P}(Y = j|x_i),\ j = 1, ..., J$$

Til modelselektion, altså til at vælge den bedste mulige model, kan backwards elimination, forward selection samt AIC anvendes, alle tre står beskrevet i afsnit 3.2.3.

3.6 Kumulativ logit model med tilfældig effekt

Frem til nu har vi kun set på regressionsmodeller af modelklassen GLM. Den kumulative logit model ønsker vi at udvide til også at indeholde en tilfældig effekt, og vi lander dermed i modeklassen GLMM.

En kumulativ logit model med tilfældig effekt er en modeltype for ordinal data med klynger, som indeholder en tilfældig effekt i den lineære prædiktor for hver klynge. Denne modeltype tillader dermed heterogenitet i de forskellige klynger, som har en særlig form af de forklarende variable. En model med tilfældig effekt er derfor en type af en latent variabel model, som blev berørt i afsnit 3.5. De observerede klynger betragtes som tilfældige stikprøver fra alle sæt af mulige klynger, og den tilfældige effekt for en klynge antages at være en uobserveret tilfældig variabel. Betinget af den tilfældige effekt behandles observationer i en klynge som uafhængige.

En kumulativ logit model med tilfældig effekt er af modelklassen GLMM, som er en udvidelse af GLM, som står beskrevet i afsnit 3.3. I følgende afsnit ser vi på en GLMM for en ordinal

responssvariabel, som antages at have en multinomial fordeling for hver værdi af den faste og tilfældige effekt (Agresti, 2010, s. 281).

Lad Y_{it} betegne responsen for observation t i klynge i. Lad $x_{1it}, x_{2it}, ..., x_{kit}$ betegne værdierne af de k forklarende variable, for hver observation. Den tilfældige effekt for klynge i betegnes med u_i. I det mest simple og mest almindelige tilfælde er u_i et intercept led i modellen. Modellen er da en tilfældig intercept model. For de mulige kategorierne $j = 1, 2, ..., J-1$, er den kumulative logit model på proportional odds formen med et tilfældigt intercept følgende:

$$logit[P(Y_{it} \leq j|u_i)] = u_i + \alpha_j + \beta_1 x_{1it} + \beta_2 + x_{2it} + ... + \beta_k x_{kit} \tag{3.6.1}$$

Modellen bruger den samme tilfældige effekt for hver kumulative sandsynlighed. Ved at bruge et overordnet intercept led af formen "$u_i + \alpha_j$" tillades subjektivitet på tværs af de tilfældige effekter på responsen. Dette tillader den ordinale skala's skæringskoefficienter at variere på tværs af de tilfældige effekter. Et subjekt med relativ høj positiv u_i har relativ stor kumulativ sandsynlighed, og dermed en relativ høj chance for at forekomme ved den lave ende af den ordinale skala. Et subjekt med relativ høj negativ u_i har relativ lille kumulativ sandsynlighed, og dermed en relativ lav chance for at forekomme ved den lave ende af den ordinale skala. Modellen kan også udtrykkes som:

$$logit[P(Y_{it} \leq j|u_i)] = \alpha_j - (u_i + \beta_1 x_{1it} + \beta_2 + x_{2it} + ... + \beta_k x_{kit}), \tag{3.6.2}$$

hvilket er et naturligt resultat af, at det er en underliggende latent variabel model. Ved at skrive modellen på denne måde, vil høje værdier af en tilfældig og fast effekt, svare til en høj sandsynlighed for at være i den høje ende af den ordinale skala. Ved opstilling af modellen efter ligning 3.6.1 kan samme fortolkning laves ved blot at indsætte $P(Y_{it} > j)$ fremfor $P(Y_{it} \leq j)$. Intercept parametrene α_j tilfredstiller $\alpha_1 < \alpha_2 < ... < \alpha_{J-1}$, hvilket

understreger ordningen af de kumulative sandsynligheder for hvert i.

Som allerede nævnt antages det, at den tilfældige effekt, u_i, er uobserveret, og værdien er derfor ukendt. Det er normalt antaget, at den varierer fra klynge til klynge ifølge en $N(0, \sigma_u^2)$-fordeling. Variansparametret σ_u^2 er et parameter, som estimeres sammen med de faste effekter. Som det står beskrevet i GLMM afsnittet, afsnit 3.3, er grunden til at den tilfældige effekt ikke bliver behandlet som en fast effekt, at modellen vil komme til at indeholde for mange parametre. Ved at behandle den tilfældige effekt, som en tilfældig effekt, da kan vi nøjes med at tilføje en ekstra parameter til modellen, nemlig σ_u. En anden grund til dette er, at hvis den tilfældige effekt blev behandlet som en fast effekt, da ville inferensen kun være omkring klyngerne og ikke den samlede population af klynger (Agresti, 2010, s. 283).

3.6.1 Estimation

Modelfitting processen estimerer de faste effekter og standardafvigelsen σ_u af den tilfældige effekt. Som beskrevet i afsnit 3.3.1 er en GLMM en to-trins model, hvilket også gør sig gældende for den kumulative logit model med tilfældig effekt.
Ved første trin har vi at betingelsen af den tilfældige effekt og de faste effekter's observationer, antages at være uafhængige. Ved andet trin antages den tilfældige effekt at være uafhængig af normalfordelingen. Ved at integrere den tilfældige effekt ud, opnås en marginal fordeling for de mulige udfald, og der haves en marginal likelihood funktion. Dette er en funktion af de faste effekt parametre; $(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \{\alpha_1, ..., \alpha_{J-1}, \beta_1, ..., \beta_k\}$ og parameteren σ_u har den tilfældige effekts fordeling $N(0, \sigma_u^2)$.
For en kumulativ logit model med tilfældig effekt med link funktion g lader vi for hver af de T_i observationer i klynge i, y_{it} være responsen for observation t, som vi repræsenterer ved en vektor $\mathbf{y}_{it}$ af de J binære indikatorer. Det er, lad $y_{ijt} = 1$ hvis observationen falder i kategori j og lad $y_{ijt} = 0$ ellers, $i = 1, ..., n,\ j = 1, ..., J,\ t = 1, ..., T_i$. Givet u_i, lad $\pi_j(\mathbf{x}_{it}; u_i) = P(y_{ijt} = 1)$. Vi behandler $\mathbf{y}_{it}$ som en multinomial observation med tæthedsfunktionen:

$$\pi_1(\mathbf{x}_{it}; u_i)^{y_{i1t}} \pi_2(\mathbf{x}_{it}; u_i)^{y_{i2t}} ... \pi_J(\mathbf{x}_{it}; u_i)^{y_{iJt}}.$$

Hvert led i dette produkt er differensen af kumulative sandsynligeheder med den inverse link funktion:

$$\pi_j(\mathbf{x}_{it}; u_i) = g^{-1}(u_i + \alpha_j + \boldsymbol{\beta}^T \mathbf{x}_{it}) - g^{-1}(u_i + \alpha_{j-1} + \boldsymbol{\beta}^t \mathbf{x}_{it}),$$

med $\alpha_0 = -\infty$ og $\alpha_J = \infty$. Med en $N(0, \sigma_u^2)$ tæthedsfunktion for u_i, har den marginale likelihood funktion følgende form:

$$\prod_{i=1}^{n} \left\{ \int_{-\infty}^{\infty} \left[\prod_{t=1}^{T_i} \prod_{j=1}^{J} (g^{-1}(u_i + \alpha_j + \mathbf{x}_{it}^T \boldsymbol{\beta}) - g^{-1}(u_i + \alpha_{j-1} + \mathbf{x}_{it}^T \boldsymbol{\beta}))^{y_{ijt}} \right] \left(\frac{1}{\sqrt{2\pi\sigma_u}} e^{\frac{-u_i^2}{2\sigma_u^2}} \right) du_i \right\}$$

Det beregningstunge ved at fitte en GLMM er behovet for at evaluere integralet for at opnå den marginale likelihood funktion, da den ikke har en lukket form. Når først marginal likelihood funktionen er approksimeret, kan Gauss-Hermite metoden anvendes til at maksimere og dermed give MLE for parametrene. Gauss-Hermite quadrature approksimerer integralet, som bestemmer den marginale likelihood funktion ved en begrænset vægtet sum, som evalueres ved quadrature punkter(Agresti, 2010, s. 287), hvilket også står beskrevet i afsnit 3.3.1.

3.6.2 Statistisk inferens og modelselektion

Efter at have fittet modellen ønskes det at teste de faste effekter, β. Til dette er det muligt at anvende de 3 standardmetoder likelihood ratio test, Wald test og Score testen, som er berørt i afsnit 3.2.2. Inferens omkring den tilfældige effekt, dens varianskomponent, er mere kompleks (Agresti, 2013, s. 522), hvilket også er understreget i afsnit 3.3.2.

Det er også muligt at bestemme et tilhørende konfidensinterval for odds ratio af de kumulative sandsynligheder. Enten kan de klassiske symmetriske intervaller bestemmes, eller også kan bedre konfidensintervaller opnås ved at anvende profile likelihood, som står beskrevet i afsnit 3.2.2.

Der er ingen garanti for at modellen vil passe perfekt. Ved manglende fit i modellen kan det skyldes flere ting, såsom udeladte prædiktorer, udeladte interaktionen mellem prædiktorer, brud på proportional odds antagelsen og/eller forkert link funktion. Her kan likelihood ratio testen anvendes til at sammenligne mere komplekse modeller med mere simple modeller. Til modelselektion, altså til at udvælge den bedst mulige model, da kan backwards elimination, forward selection samt AIC anvendes, som står beskrevet i afsnit 3.2.2.

4 Hvilke studerende besvarer, og hvilke besvarer ikke kursusevalueringen

Frem til nu er der gennemgået teori, som kan anvendes til at undersøge, om det er muligt at sige noget om evalueringsresultaterne for studieåret 2018/2019 for CBS. I dette afsnit vil vi undersøge, om det er muligt at opstille en eller flere modeller til at sige noget om, hvilke studerende der besvarer, og hvilke studerende der ikke besvarer kursusevalueringen.
På baggrund af disse modeller vil vi undersøge, om det er muligt at bruge dem til at prædiktere på andre uddannelser, hvilke studerende der besvarer, og hvilke der ikke besvarer deres kursusevalueringer. På baggrund af denne analyse vil det være muligt at hjælpe CBS til at forstå, hvilke studerende de skal have fat i for at forsøge at øge svarprocenten.

4.1 Anvendt data

Fra afsnit 2.4, som indeholder krydstabeller, er det tydeligt at se, at det er langt fra alle studerende på CBS, som besvarer deres kursusevalueringer. Det fremgår af tabel 1, at den gennemsnitlige svarprocent for studieåret 2018/2019 er $29,53\%$, ligeledes fremgår det, at svarprocenten på tværs af kurser varierer meget, da minimum svarprocenten for et kursus er $4,17\%$, og maksimum svarprocenten for et kursus er $93,33\%$. Variationen mellem svarprocenterne blandt kurserne er med til at indikere, at det bør være muligt at få svarprocenten op. Fra tabel 4 ses det, at svarprocenten blandt uddannelserne varierer meget, den varierer med en gennemsnitlig svarprocent fra $13,10\%$ og op til $54,32\%$. Det er svært at sige, hvad det er der gør, at svarprocenten varierer så meget fra uddannelse til uddannelse, det kunne dog godt tænkes at have noget at gøre med kulturen på de forskellige uddannelsen og/eller kulturen blandt de studerende.

Til analysen af hvilke studerende der besvarer, og hvilke det ikke besvarer kursusevalueringen, anvendes nogle udvalgte uddannelser. Vi har udvalgt 3 forskellige uddannelser for at tage højde for eventuelle uddannelsesspecifikke tendenser, som der godt kan tænkes at være, blandt andet på grund af den meget varierende svarprocent blandt uddannelserne. En anden grund

til at vi er nødt til at lave analysen for et udpluk af data er af tids- og pladsmæssige årsager, som gør, at det ikke er muligt at analysere hver enkelte uddannelse.
Til at opstille de statistiske modeller har vi udvalgt de tre største uddannelser på CBS, som er; Kandidatuddannelsen i Erhvervsøkonomi (MScCM), Bacheloruddannelsen i Erhvervsøkonomi (BScHA) og Bacheloruddannelsen i Erhvervsøkonomi og Erhvervsjura (BScJUR).

Når vi opstiller en statistisk model, der skal hjælpe med at analysere, hvilke studerende der besvarer, og hvilke der ikke besvarer deres kursusevalueringer, har vi brug for at inkludere nogle forklarende variable, som vi tænker kan have en effekt. Vi så i afsnit 2.4 på krydstabeller vedrørende svarprocenter fordelt ud på informationer omkring de studerende. Af tabel 5 fremgår det, at cirka hver tredje kvindelige studerende får besvaret kursusevalueringerne, hvorimod det kun er cirka hver fjerde mandlige studerende, der får besvaret. Derudover fremgår det af tabel 6, at de studerende med nationalitet "Udenfor Norden" er bedre til at få besvaret deres kursusevalueringer end de studerende fra "Norden". En forklaring på dette kan være, at de studerende med nationalitet "Udenfor Norden" typisk selv skal betale for deres uddannelse, og derfor lægger mere vægt på at få evalueret det forløb, som de har købt sig til.
Af tabel 7 fremgår det, at jo yngre de studerende er, jo bedre er de til at få besvaret deres kursusevalueringer, dog er variationen i svarprocenten ikke ret stor aldersgrupperne imellem. En årsag til dette kan være, at de yngste studerende, er de "nyeste" på CBS, og derfor er mere pligtopfyldende, hvorimod de fleste af de ældre studerende har gået på CBS i længere tid og derfor kan være trætte af at evaluere hvert et kursus.
Informationer omkring de studerende såsom køn, nationalitet og fødselsår kan dermed godt vise sig at have en effekt på, hvilke studerende der besvarer og ikke besvarer deres kursusevalueringer. Selvom de studerendes opnåede karakterer i kurset først er noget de studerende får efter kursusevalueringen, er det interessant at se, om karakteren har en effekt, da den kan være med til at beskrive den studerendes faglige niveau.
Yderligere så vi i afsnit 2.4, at vi skelner imellem to typer af kurser, da der både findes obligatoriske kurser samt valgfag. Af tabel 3 fremgik det, at det ikke er af stor betydning om

kurset er obligatorisk eller et valgfag, da svarprocenten er henholdsvis $29,52\%$ og $29,54\%$. Af tabel 1 fremgår det, at den gennemsnitlige svarprocenten er højere om efteråret, $31,88\%$, end om foråret, $26,57\%$.
De forklarende variable vi anvender til at fitte modellerne for, hvilke studerende der besvarer, og hvilke der ikke besvarer kursusevalueringen, bliver da: *Køn*, *Nationalitet*, *Fødselsår*, *Karakter*, *Kursus_Type* og *Eval_Semester*.

Opgavens datasæt indeholder to typer af karakterskalaer, hvor den ene er den klassiske 7-trinsskala[12], og den anden er en bestået/ikke-bestået skala[13]. Det er primært 7-trins karakterskalaen, der anvendes på CBS, og derfor ser vi kun på 7-trinsskalaen i denne opgave. Det betyder at skalaen bestået/ikke-bestået filteres ud[14]. Derudover filtrerer vi alle studerende med karakteren "U" i variablen *Karakter*[15] ud. Vi filtrerer også alle "NA" i variablen *Karakter* ud, så der kun er studerende, som har fået tildelt en karakter[16].
Variablen *Resp_Status* indeholder tre niveauer; "Saved", "DNR" og "Submitted"[17]. Alle besvarelser der er "Saved" filtreres ud, da besvarelserne ikke er færdiggjort og derfor opfattes som ufuldstændige,[18] men også fordi vi kun ønsker at se på, hvilke studerende der besvarer, og hvilke der ikke besvarer kursusevalueringen.

Når vi ser på data for en hel uddannelse, vil størstedelen af de studerende være tilmeldt flere kurser, typisk 4 kurser per semester, og dermed 8 kurser for et helt studieår. Den studerende modtager en kursusevaluering for hvert enkelt tilmeldt kursus. Det betyder, at den samme studerende kan optræde flere gange i data, og observationerne kan derfor ikke siges at være uafhængige. Der er derfor afhængighed i data, og det er vigtigt at tage højde for dette i den statistiske model. Afhængighed i data kan have indflydelse på variansen og dermed forårsage

[12]Indeholder 8 niveauer; "U" (udeblevet) og 7-trinsskalaen: -3, 00, 02, 4, 7, 10 og 12.
[13]Indeholder 3 niveauer; "U" for udeblevet, "B" for bestået og "I" for ikke-bestået.
[14]Samlet set er det 76 kurser ud af de i alt 838, som anvender bestået/ikke-bestået.
[15]Dette omhandler 8.022 observationer ud af de i alt 80.433
[16]Dette omhandler 1.494 observationer ud af de i alt 80.433
[17]Saved: påbegyndt besvarelsen, DNR: ikke besvaret, Submitted: besvaret
[18]Dette omhandler 612 kursusevalueringer, altså 0,76%, af datasættet

overspredning. Det vides dog ikke på forhånd, hvor stor en betydning afhængigheden i data har for modellen.

For at undersøge hvilke studerende der besvarer, og hvilke der ikke besvarer, opstiller vi en logistisk regressionsmodel med variablen *Resp_Status* som responsvariabel. Vi har dermed en binær logistisk regression, hvor den binære responsvariabel indeholder de to udfald; "DNR", der har værdien 0, og "Submitted", som har værdien 1.

Til at udvælge den bedst mulige model anvendes backwards elimination og AIC. Der opstilles først en startmodel for den pågældende uddannelse, som indeholder alle de mulige forklarende variable, og herefter fjernes de insignifikante variable en efter en, indtil vi har en endelig, reduceret model, som i denne opgave kaldes for en slutmodel.
Når vi opstiller en statistisk model, vurderes den løbende udviklede model på baggrund af testssandsynlighederne for effekten af de forklarende variable. Når der er en signifikant effekt i en af de forklarende variable, betyder det, at den forklarende variabel har en effekt for det der måles på, som i dette tilfælde er *Resp_Status*. Når en forklarende variabel har en effekt, kan det udledes, at respondenterne der tilhører denne forklarende variabel, adskiller sig i den forklarende variabels niveauer, for det der måles på. Insignifikante variable bliver udeladt af modellerne, da de ikke har noget at bidrage med til forklaringen af det, der måles på. Hver gang en forklarende variabel fjernes fra en model, tester vi med likelihood ratio test, om det er acceptabelt at fjerne den. Nulhypotesen der testes er som følgende: H_0: Den reducerede model fitter mere præcist. Nulhypotesen bliver som hovedregel afvist, hvis likelihood ratio testens $p - værdi < 0,05$, da det betyder, at den forklarende variabel, som vi forsøger at fjerne fra modellen, er signifikant, og dermed ikke kan fjernes.

Anvendt R-kode

Til følgende afsnit anvendes pakkerne i R: "stats", "aod", "ROCR" og "pROC". I bilag F.1 fremgår et eksempel på anvendelse af R-kode, som bruges til at udarbejde denne analyse.

4.2 Logistisk regression for Kandidatuddannelsen i Erhvervsøkonomi

Til at begynde med opstiller vi en model for Kandidatuddannelsen i Erhvervsøkonomi (MScCM), som er den uddannelse, der har udsendt flest kursusevalueringer i studieåret 2018/2019, hvilket må betyde, at det er den største uddannelse[19]. Startmodellen for MScCM indeholder alle de forklarende variable. Modellen der beskriver sandsynligheden for, at de studerende besvarer deres kursus-evalueringer er da:

$$logit[P(Resp_Status = 1)] = \alpha + \beta_{Køn} + \beta_{Nationalitet} + \beta_{Fødselsår} + \beta_{Karakter} + \beta_{Kursus_Type} + \beta_{Eval_Semester} \tag{4.2.1}$$

Startmodellen giver en AIC på 14.696. De forklarende variable *Fødselsår* og $Karakter$ er de eneste, som er insignifikante, hvilket fremgår af tabel 9. Inden vi forsøger at fjerne de insignifikante variable, undersøger vi afhængigheden i datas betydning for modellen.

Tabel 9: P-værdier for de forklarende variable i startmodellen for MScCM

Variabel	KønM	Nationalitet	Fødselsår	Karakter	Kursustype	Semester
P-værdi	5,3e-08	<2e-16	0,23	0,25	6,7e-07	<2e-16

For at undersøge afhængigheden i datas betydning for modellen beregnes dispersionsparameteren for startmodellen 4.2.1. Dispersionsparameteren bestemmes ved at dividere residual deviansen med antallet af frihedsgrader, hvilket giver modellen en dispersionsparameter på $1,23$. Hvis dispersionsparameteren er omkring 1, indikerer det, at afhængigheden i data ikke har nogen særlig betydning for modellen. Vi undersøger dog for en god ordens skyld, hvilken indflydelse det har på modellen.

[19]Der udsendes en kursusevaluering til hver tilmeldte studerende

Ved at lave en logistisk regressionsmodel tvinger modellen dispersionsparameteren til at være 1. Det er i stedet muligt at opstille en quasibinomial model, som er en model, der kan tage højde for eventuel overspredning. Ved at opstille en quasibinomial model vil modellen korrigere i de forklarende variables standardafvigelse for at opnå en dispersionsparameter så tæt på 1 som muligt.
Ved at opstille både en logistisk regressionsmodel og en quasibinomial model vil parameter estimaterne være de samme, men standardafvigelsen og deraf p-værdien kan være forskellige. Vi opstiller nu en quasibinomial model, og af tabel 10 fremgår det, at der ikke er nogen bemærkelsesværdig variation i standardafvigelserne ved sammenligning af dem for den logistiske regressionsmodel og quasibinomial modellen. Derudover fremgår det også, at det er de samme forklarende variable, der er signifikante.

Tabel 10: Estimater og p-værdier for startmodellen for en logistisk og quasibinomial model for MScCM

	Logistisk			Quasibinomial		
	Estimat	Std.afv.	P-værdi	Estimat	Std.afv	P-værdi
KønM	-0,218	0,040	5,26e-08 ***	-0,218	0,040	5.60e-08 ***
NatioUdenfor Norden	0,401	0,042	<2e-16 ***	0,401	0,042	<2e-16 ***
Fødselsår1996-2018	0,075	0,063	0,233	0,075	0,063	0,234
Karakter0	-0,160	0,075	0,032 *	-0,160	0,075	0,033 *
Karakter2	-0.142	0,074	0,056 .	-0,142	0,074	0,056 .
Karakter4	-0,039	0,074	0,595	-0,039	0,074	0,595
Karakter7	-0,047	0,074	0,523	-0,047	0,074	0,524
Karakter10	-0,024	0,074	0,747	-0,024	0,074	0,747
Karakter12	-0,073	0,074	0,323	-0,073	0,074	0,323
Kursus_TypeV	-0,231	0,046	6,70e-07 ***	-0,231	0,047	7,05e-07 ***
SemesterF19	-0,665	0,048	<2e-16 ***	-0,665	0,048	<2e-16 ***

Det betyder at afhængigheden i data ikke har en særlig stor indflydelse på modellen, hvilket også var forventeligt, da startmodellens dispersionsparameter blev beregnet til at være tæt på 1. Det er derfor muligt at fortsætte med at bruge den logistiske regressionsmodel, da afhængigheden i data ikke ser ud til at have den store indflydelse på variansen. Vi foretrækker at gå videre med den logistiske regressionsmodel, da AIC er tilgængelig for denne type af modeller. AIC er, som bekendt, et godt mål til at udvælge den bedste model.

For at optimere startmodellen 4.2.1 fjerner vi den forklarende variabel $Karakter$ fra modellen, hvilket giver en ny model, som har en AIC på 14.692. Ud fra AIC'en er den nye model en anelse bedre end startmodellen. For at sikre at den forklarende variable $Karakter$ kan udelades fra modellen, laver vi en likelihood ratio test for at sammenligne de to modeller. Testen giver en p-værdi på $0,251$, hvilket betyder at nulhypotesen[20] ikke kan afvises og at variablen $Karakter$ dermed kan udelades fra modellen. Vi følger nu samme procedure for variablen *Fødselsår*, som med en p-værdi på 0,23, også er en insignifikant variabel. En likelihood ratio test giver en p-værdi på $0,243$, og variablen udelades dermed fra modellen. De tilbageværende variable i modellen; *Køn*, $Nationalitet$, $Kursus_Type$ og $Eval_Semester$ er signifikante og har dermed en effekt på, hvilke studerende der besvarer, og hvilke studerende der ikke besvarer deres kursusevalueringer på uddannelsen MScCM. Det er derfor ikke muligt at fjerne flere variable fra modellen uden at modellens fit forværres. Slutmodellen for MScCM er således:

$$logit[P(Resp_Status = 1)] = \alpha + \beta_{Køn} + \beta_{Nationalitet} + \beta_{Kursus_Type} + \beta_{Eval_Semester} \tag{4.2.2}$$

Ovenstående slutmodel 4.2.2 har en AIC på 14.691, hvilket er lidt lavere end startmodellen, og dette indikerer, at denne model er at foretrække. Da slutmodellen, ligning 4.2.2, er forskellig fra startmodellen, ligning 4.2.1, kan dispersionsparameteren have ændret sig. Derfor beregner vi slutmodellens dispersionsparameter på ny og får $1,231$, som igen er tæt på 1. Derudover opstiller vi slutmodellen som en quasibinomial model, hvor det igen fremgår, at der ikke er

[20] H_0: "Den reducerede model fitter mere præcist"

nogen bemærkelsesværdig forskel i standafvigelserne, og det er dermed de samme forklarende variable, der er signifikante.

Hvert enkelte estimerede parameter angiver den forventede ændring i log odds for at besvare, med 1-enheds øgning i den korresponderende forklarende variabel, når de resterende forklarende variable holdes konstant. Ved at tage 'exp' af de estimerede parametre opnås ratioen af to odds eller ændringen i odds for en multikategorisk skala for 1-enheds øgning i den korresponderende variabel, ved at holde de resterende forklarende variable konstante ved en bestemt værdi. Odds ratioer og tilhørende konfidensintervaller for slutmodellen, ligning 4.2.2, fremgår af tabel 11, ud fra dem kan effekten af de forklarende variable tolkes.

Tabel 11: Odds ratioer og konfidensintervaller for slutmodellen for MscCM

	OR	2,5 %	97,5 %
KønM	0,617	0,569	0,668
NationalitetUdenfor Norden	1,504	1,384	1,632
Kursus_TypeV	0,786	0,719	0,859
Eval_SemesterF19	0,515	0,469	0,566

Ved at se på den forklarende variabel *KønM* imens de resterende forklarende variable holdes konstant, ses det, at odds ratioen for at en mandlig studerende besvarer kursusevalueringen er $0,62$ gange lavere end for kvindelige studerende. Sagt på en anden måde, odds ratioen er 38% mindre for at en mandlig studerende besvarer kursusevalueringen, end at en kvindelig studerende besvarer. Det kan aflæses af konfidensintervallet, at med 95% sandsynlighed vil odds ratioen for *KønM* ligge i intervallet $[0,569; 0,668]$.

Når vi ser på den studerendes nationalitet[21] er odds ratioen for, at en studerende med nationalitet "Udenfor Norden" besvarer kursusevalueringen 1,50 gange større, end hvis den

[21]Inddelt i de to grupper "Norden" og "Udenfor Norden"

studerende har nationalitet "Norden", når alle andre variable holdes konstant. Dette vil sige, at odds ratioen er 50% større for, at de studerende med nationalitet "Udenfor Norden" besvarer deres kursusevalueringer end de studerende fra "Norden". Konfidensintervallet indikerer, at det er med 95% sandsynlighed at odds ratioen for $NationalitetUdenforNorden$ ligger i intervallet $[1,384; 1,632]$.

Ser vi på variablen $Kursus_Type$ kan vi aflæse af tabel 11, at odds ratioen for at kursusevalueringen for et valgfag bliver besvaret, er 0,79 gange (eller 21%) mindre end for at et obligatorisk kursus, når alle andre variable holdes konstant. Konfidensintervallet angiver, at det er med 95% sandsynlighed at odds ratioen for $Kursus_TypeV$ ligger i intervallet $[0,719; 0,859]$. Sidst men ikke mindst ser vi på variablen $Eval_Semester$, hvor vi af tabel 11 kan aflæse, at odds ratioen for at kursusevalueringen bliver besvaret for et kursus i forårssemesteret er 0,52 gange (eller 48%) mindre end for et kursus i efterårssemesteret, når alle andre variable holdes konstant. Konfidensintervallet angiver, at det er med 95% sandsynlighed at odds ratioen for $Eval_SemesterF19$ ligger i intervallet [0,469;0,566].

Opsummering af Kandidatuddannelsen i Erhvervsøkonomi

Først og fremmest kom vi frem til, at det var muligt at gå videre med en logistisk regressionsmodel frem for en quasibinomial model, da afhængigheden i data ikke har en bemærkelsesværdig indflydelse på variansen. Den statistiske model som beskriver, hvilke forklarende variable der har en effekt på, hvilke studerende der besvarer kursusevalueringer for MScCM, er slutmodellen, 4.2.2. Det fremgår af slutmodellen, at de forklarende variable der har størst effekt på responsvariablen, er de to forklarende variable der siger noget om den studerende; *Køn* og $Nationalitet$, samt de to forklarende variable der siger noget om kurset; $Kursus_Type$ samt $Eval_Semester$.

Ud fra odds ratioer fra tabel 11 viste det sig, at odds ratioen for at de mandlige studerende samt studerende fra "Norden" besvarer deres kursusevaluering, er mindre end for henholdsvis kvindelige studerende og studerende med nationalitet "Udenfor Norden". Ligeledes så vi for de to forklarende variable, der beskriver kurset, at odds ratioen for at et valgfag samt et

kursus fra forårssemesteret bliver besvaret i kursusevalueringen er mindre end for henholdsvis de obligatorisk kursuser samt kurser placeret i efterårssemesteret.
Denne viden kan CBS bruge til at motivere de mandlige studerende samt de studerende fra Norden på MScCM til at få besvaret kursusevalueringen. CBS kan med fordel også lægge flere kræfter i at motivere de studerende til at få besvaret kursusevalueringerne i foråret. Ud fra hvad vi har set i krydstabellerne i afsnit 2.4 kommer intet af ovenstående bag på os.

4.3 Logistisk regression for Bacheloruddannelsen i Erhvervsøkonomi

Den næste uddannelse vi analyserer er Bacheloruddannelsen i Erhvervsøkonomi (BScHA), som er den næststørste uddannelse på CBS målt i antal udsendte kursusevalueringer. Vi følger nu samme fremgangsmåde som for MScCM. Vi opstiller først en startmodel, som indeholder alle forklarende variable. Modellen med sandsynligheden for at den studerende besvarer deres kursusevalueringer er da:

$$logit[P(Resp_Status = 1)] = \alpha + \beta_{Køn} + \beta_{Nationalitet} + \beta_{Fødselsår} + \beta_{Karakter} + \beta_{Kursus_Type} + \beta_{Eval_Semester} \tag{4.3.1}$$

Model 4.3.1 kalder vi for startmodellen og den har en AIC på $9.079,6$. Af de forklarende variable, er det *Karakter* og *Fødselsår*, der er insignifikante, hvilket fremgår af tabel 12. Vi ønsker først at undersøge afhængigheden i data for at se, om det har en indflydelse på modellen, før vi begynder at fjerne de insignifikante variable en efter en.

Tabel 12: P-værdier for de forklarende variable i startmodellen for BScHA

Variabel	KønM	Nationalitet	Fødselsår	Karakter	Kursustype	Semester
P-værdi	7,8e-10	2,4e-15	0,920	0,470	0,0023	4.3e-10

Dispersionsparameteren for startmodellen 4.3.1 beregnes til at være $1,130$, hvilket må siges at være tæt på 1. Vi opstiller startmodellen for BScHA som en quasibinomial model, og det ses,

at der ikke er nogen bemærkelsesværdig variation i standardafvigelserne med sammenligning til standardafvigelserne fra den logiske regressionsmodel. Dermed fremgår det også, at det er de samme variable, som er signifikante. Igen ses det, at afhængigheden i data ikke har en særlig stor indflydelse på variansen, hvilket heller ikke var forventeligt, da startmodellens dispersionsparameter blevet beregnet til at være tæt på 1. Vi fortsætter derfor med en logistisk regressionsmodel.

For at optimere startmodellen 4.3.1 fjerner vi de insignifikante variable en af gangen. Vi tester løbende med likelihood ratio test, hvorvidt variablen kan fjernes fra modellen eller ej. Først fjerner vi variablen $Karakter$ fra startmodellen, hvilket giver os en reduceret model, som altså er mere simpel end startmodellen. Ved en likelihood ratio test af startmodellen og den reducerede model opnås en p-værdi på $0,464$. Det betyder, at nulhypotesen ikke kan afvises og den reducerede model er dermed at foretrække. Efterfølgende forsøger vi at fjerne variablen *Fødselsår* fra den reducerede model. Dette er en endnu mere simpel model, som ved en likelihood ratio test testes med modellen, hvor variablen indgår, og med en p-værdi på $0,903$ kan nulhypotesen afvises, og vi står tilbage med en endnu mere reduceret model. BScHA ender således med en slutmodel, der indeholder de forklarende variable: $Køn$, $Nationalitet$, $Kursus_Type$ og $Eval_Semester$. Slutmodellen der angiver sandsynligheden for, at den studerende besvarer kursusevalueringen har dermed følgende form:

$$logit[P(Resp_Status = 1)] = \alpha + \beta_{Køn} + \beta_{Nationalitet} + \beta_{Kursus_Type} + \beta_{Eval_Semester} \tag{4.3.2}$$

Slutmodellen 4.3.2 har en AIC på $9.071,3$, som er lidt lavere end startmodellen 4.3.1, hvilket indikerer, at den reducerede model er at foretrække. Da slutmodellen 4.3.2 er forskellig fra startmodellen 4.3.1 kan dispersionsparameteren have ændret sig. Slutmodellens dispersionsparameter er $1,129$, som igen er tæt på 1. Derudover opstiller vi slutmodellen som en quasibinomial model, hvor det igen fremgår, at der ikke er den store forskel at se i standardafvigelserne, og det er derfor også de samme forklarende variable, der er signifikante. Nedenfor i tabel 13 fremgår slutmodellens odds ratioer og tilhørende konfidensintervaller. Når vi ser på

Tabel 13: Odds ratio og konfidensinterval for slutmodellen for BScHA

	OR	2,5 %	97,5 %
KønM	0,721	0,650	0,80
NationalitetUdenfor Norden	1,912	1,636	2,234
Kursus_TypeV	1,225	1,082	1,386
Eval_SemesterF19	0,717	0,646	0,796

variablen *Køn*, fremgår det af tabel 13, at odds ratioen for at en mandlig studerende besvarer kursusevlueringen er $0,72$ gange (eller 28%) mindre, end det er for kvindelige studerende, når alle andre variable holdes konstant. Ud fra konfidensintervallet kan vi sige, at med 95% sandsynlighed ligger odds ratioen for mænd i intervallet $[0,650; 0,800]$.

På samme måde kan vi aflæse af tabel 13 at odds ratioen for at studerende med nationalitet "Udenfor Norden" besvarer kursusevalueringen er $1,91$ gange (eller 91%) større, end for at en studerende fra "Norden" besvarer kursusevalueringen, når alle andre variable holdes konstant. Konfidensintervallet indikerer, at det er med 95% sandsynlighed at odds ratioen for $Nationalitet\ Udenfor\ Norden$ ligger i intervallet [1,636;2,234]. Ved at se på variablen $Kursus_Type$ fremgår det af tabel 13, at odds ratioen for at en kursusevaluering for et valgfag bliver besvaret er $1,23$ gange (eller 23%) større end for et obligatorisk kursus, når alle andre variable holdes konstant. Konfidensintervallet indikerer, at det er med 95% sandsynlighed, at odds ratioen for $Kursus_TypeV$ ligger i intervallet [0,646;0,796].

Slutteligt kan vi aflæse af tabel 13, at hvis kurset der evalueres ligger i F19, så er oddset for, at den studerende får besvaret kursusevalueringen $0,72$ (eller 28%) mindre end, hvis kurset ligger i E18, når alle andre variable holdes konstant. Ud fra konfidensintervallet kan vi sige, at med 95% sandsynlighed ligger odds ratioen for at F19 kurser bliver besvaret i intervallet $[0,646; 0,796]$.

Opsummering af Bacheloruddannelsen i Erhvervsøkonomi

Først og fremmest kom vi frem til, at det var muligt at gå videre med en logistisk regressionsmodel frem for en quasibinomial model, da afhængigheden i data ikke havde en bemærkelsesværdig indflydelse på variansen. Den statistiske model som beskriver, hvilke forklarende variable der har en effekt på, hvilke studerende der besvarer kursusevalueringen for BScHA, er slutmodellen, 4.3.2. Det fremgår af slutmodellen, at de forklarende variable der har størst effekt på responsvariablen, er de to forklarende variable der siger noget om den studerende: *Køn* og $Nationalitet$, samt de to forklarende variable der siger noget om kurset: $Kursus_Type$ samt $Eval_Semester$.

Ud fra odds ratioerne i tabel 13 viste det sig, at odds ratioen for at de mandlige studerende, samt studerende fra "Norden", besvarer deres kursusevaluering, er mindre end for henholdsvis kvindelige studerende og studerende med nationalitet "Udenfor Norden". Ligeledes så vi for den forklarende variabel $Kursus_Type$, der beskriver kurset, at odds ratioen for at et valgfag bliver besvaret i kursusevalueringen, er større end for et obligatorisk kursus. Slutteligt så vi at odds ratioen for at et kursus fra forårssemesteret F19 bliver evalueret i kursusevalueringen, er lavere end for at en kursusevaluering fra efterårssemesteret E18 bliver besvaret.

Denne viden kan CBS bruge til at motivere de mandlige studerende, samt de studerende fra "Norden" på BScHA til at få besvaret kursusevalueringen. CBS kan med fordel også lægge flere kræfter i at motivere de studerende til at få besvaret kursusevalueringerne i foråret.

4.4 Logistisk regression for Bacheloruddannelsen i Erhvervsøkonomi og Erhvervsjura

Den tredje største uddannelse målt på antal udsendte kursusevaluering er Bacheloruddannelsen i Erhvervsøkonomi og Erhvervsjura (BScJUR). Først opstiller vi en startmodel, som indeholder alle de forklarende variable. Modellen med sandsynligheden for at den studerende besvarer deres kursusevalueringer er da:

$$logit[P(Resp_Status = 1)] = \alpha + \beta_{Køn} + \beta_{Nationalitet} + \beta_{Fødselsår} + \beta_{Karakter} + \beta_{Kursus_Type} + \beta_{Eval_Semester} \tag{4.4.1}$$

Startmodellen 4.4.1 har en AIC på 3.558, og af tabel 14 fremgår det, at de forklarende variable $Nationalitet$, $Karakter$, $Kursus_Type$ og $Eval_Semester$ er insignifikante. Igen undersøges afhængigheden i data, og hvilken betydning det har for modellen, før vi begynder at forsøge at fjerne de insignifikante variable en efter en.

Startmodellens dispersionsparameter beregnes til at være $0,993$, hvilket må siges at være tæt på 1. Vi opstiller startmodellen som en quasibinomial model med alle forklarende variable. Ved at sammenligne standardafvigelserne for de to modeller med hinanden ses det, at der ikke er nogen bemærkelsesværdig forskel i standardafvigelserne og ej heller i p-værdierne[22]. Det er derfor de samme variable, som der er signifikante. Endnu engang leder dette frem til, at afhængigheden i data ikke har en stor betydning for modellen, da variansen næsten ikke ændrer sig, når vi tager højde for afhængigheden. Vi fortsætter derfor med en logistisk regressionsmodel.

Tabel 14: P-værdier for de forklarende variable i startmodellen for BScJUR

Variabel	KønM	Nationalitet	Fødselsår	Karakter	Kursustype	Semester
P-værdi	1.4e-10	0,41	0,0031	0,37	0,17	0,56

Vi fjerner nu de insignifikante variable en efter en. Først fjerner vi variablen $Eval_Semester$, og en likelihood ratio test af startmodellen og denne reducerede model giver en p-værdi på $0,557$, hvilket betyder, at vi ikke kan afvise nulhypotesen, og vi fortsætter derfor med den reducerede model. Det samme gentages, først for variablen $Kursus_Type$, dernæst for variablen $Karakter$ og slutteligt for variablen $Nationalitet$. Ved løbende at lave likelihood ratio test afvises nulhypotesen gang på gang for de tre forklarende variable. Slutmodellen for

[22]Se bilag B

BScJUR ender således med at indeholde de forklarende variable *Køn* og *Fødselsår*. Modellen med sandsynlighed for at den studerende besvarer kursusevalueringen har dermed formen:

$$logit[P(Resp_Status = 1)] = \alpha + \beta_{Køn} + \beta_{Fødselsår} \tag{4.4.2}$$

Slutmodellen har en AIC på $2.870,8$, hvilket er lavere end startmodellen, og det indikerer at slutmodellen er at foretrække. Slutmodellens dispersionsparameter beregnes til at være $0,993$, hvilket er tæt på 1. Vi opstiller slutmodellen som en quasibinomial model, og det viser sig, at standardafvigelserne fra den logiske regressionsmodel sammenlignet med quasibinomial modellen stort set er ens, og dermed er det også de samme forklarende variable, som er signifikante[23].

Slutmodellens odds ratioer og tilhørende konfidensintervaller fremgår nedenfor i tabel 15. Først ser vi på den forklarende variabel *Køn*, hvor det fremgår, at odds ratioen for at en mandelig studerende besvarer sin kursusevaluering er $0,54$ gange (eller 46%) mindre, end for at en kvindelig studerende besvarer, når alle andre variable holdes konstant. Af konfidensintervallet kan vi aflæse, at odds ratioen for *KønM* med 95% sandsynlighed ligger i intervallet [0,447;0,650]. Den anden forklarende variabel i slutmodellen for BScJUR er

Tabel 15: Odds ratioer og tilhørende konfidensintervaller for BScJUR

	OR	2,5 %	97,5 %
KønM	0,539	0,447	0,650
Fødselsår1996-2000	1,381	1,126	1,701

Fødselsår. Af tabel 15 fremgår det, at odds ratioen for at de studerende der er født i intervallet "1996-2000", besvarer deres kursusevalueringer er $1,38$ gange (eller 38%) højere, end for studerende der er født i intervallet "1958-1995", når alle andre variable holdes konstant. Det

[23]Se bilag C

vil altså sige, at de yngste studerende er bedst til at få besvaret deres kursusevalueringer. En årsag til dette kan være, at de yngste studerende sandsynligvis også er de nyeste studerende på CBS. Som ny studerende kan det godt tænkes, at man er mere pligtopfyldende, end når man er nået længere i uddannelsen og måske er blevet træt af at evaluere hvert enkelt kursus. Konfidensintervallet indikerer, at det er med 95% sandsynlighed, at odds ratioen for *Fødselsår1996-2000* ligger i intervallet [1,126;1,701].

Opsummering af Bacheloruddannelsen i Erhvervsøkonomi og Erhvervsjura

Først og fremmest kom vi frem til, at det var muligt at gå videre med en logistisk regressionsmodel frem for en quasibinomial model, da afhængigheden i data ikke har en bemærkelsesværdig indflydelse på variansen. Den statistiske model som beskriver, hvilke forklarende variable der har en effekt på, hvilke studerende der besvarer deres kursusevalueringer for BScJUR, er slutmodellen, 4.4.2. Det fremgår af slutmodellen, at de forklarende variable, der har størst effekt på responsvariablen, er de to forklarende variable *Køn* og *Fødselsår*. Begge de forklarende variable i slutmodellen er variable, der beskriver den studerende.

Ud fra odds ratioerne i tabel 15, viste det sig, at odds ratioen for at de mandlige studerende besvarer deres kursusevaluering er mindre end for de kvindelige studerende. Ligeledes så vi at odds ratioen for at de yngste studerende, dem med fødselsår i intervallet "1996-2000", besvarer deres kursusevaluering, er højere end for de ældste studerende, som har fødselsår i intervallet "1958-1995". Denne viden kan CBS bruge til at forsøge at motivere de mandlige studerende samt de ældste studerende på BScJUR til at få besvaret kursusevalueringen.

4.5 De 3 logistiske regressionsmodeller

Vi har nu gennemgået de tre modeller, der bedst beskriver, hvilke studerende der besvarer deres kursusevalueringer på de tre største uddannelser på CBS. Fælles for alle tre modeller er, at afhængigheden i data ikke har nogen bemærkelsesværdig indflydelse på variansen. Dette gør det muligt for os at træffe et valg om at gå videre med en logistisk regressionsmodel i alle tre tilfælde, og det indikerer, at det formentlig er en generel tendens for alle uddannelserne på CBS for studieåret 2018/2019.

Den forklarende variabel *Køn* har vist sig at være signifikant i alle tre modeller. Vi har samtidigt set i alle modellerne, at odds ratioen for at de mandlige studerende får besvaret deres kursusevalueringer er mindre end for at de kvindelige studerende får besvaret, når de resterende variable holdes konstant. Dette kommer ikke bag på os, da vi i afsnittet 2.4 har en krydstabel, der viser, at der er flere kvindelige studerende end mandlige studerende der har besvaret en kursusevaluering for studieåret 2018/2019.

Slutmodellen for MScCM og BScHA har flere fælles forklarende variable end *Køn*. Begge uddannelser har også de forklarende variable $Nationalitet$, $Kursus_Type$ og $Eval_Semester$ til fælles. Både $Nationalitet$ og $Eval_Semester$ viser for begge uddannelser, at det er de studerende med nationalitet "Norden" samt kurserne i F19, hvor odds ratioen er større for at de studerende får besvaret deres kursusevalueringer, i forhold til henholdsvis de studerende med nationalitet "Udenfor Norden" samt kurserne i E18.

De studerende på bacheloruddannelsen BScHA er typisk de studerende, som vælger kandidatuddannelsen MScCM. Det giver derfor god mening, at de forklarende variable og deres odds ratioer i høj grad viser det samme. BScHA er den største bacheloruddannelse, og MScCM er den største kandidatuddannelse, og de to uddannelser udbyder flest valgfag, som studerende fra alle andre uddannelser på CBS også kan tilmelde sig. Det giver derfor god mening, at kursets type har en effekt på, om der bliver besvaret på kursusevalueringen eller ej, da det potentielt set er en blanding af forskellige studerende, som er tilmeldt disse kurser. Odds ratioerne for variablen $Kursus_Type$ varierer dog for de to uddannelser. Odds ratioen for at besvare en kursusevaluering for et obligatorisk kursus er større end for et valgfag på MScCM, mens det på BScHA er modsat. Dette kan forklares med, at på BScHA er det første gang, at man som studerende får mulighed for at vælge valgfag. Den studerende kan dermed synes at det er spændende og nyt, pludselig at have indflydelse på kurserne, hvorimod de studerende på kandidaten har prøvet at vælge valgfag før.

Slutmodellen for BScJUR indeholder ikke variablen *Nationalitet*, hvilket ikke er overraskende, da det er en dansk uddannelse og der vil derfor typisk kun være studerende tilmeldt fra "Norden". Det kan derfor godt tænkes, at variablen *Nationalitet* typisk vil indgå i en model for de engelske uddannelser og ikke de danske uddannelser. BScHA er dog også en dansk uddannelse, men i og med at det er en stor uddannelse, og de udbyder mange valgfag, og nogle af valgfagene udbydes på engelsk, giver det mening at *Nationalitet* har en effekt.
Køn er altså den eneste forklarende variabel, som er fælles for alle tre modeller. Det vil sige at for at kunne beskrive, hvilke studerende der besvarer, og hvilke studerende der ikke besvarer kursusevalueringerne, er den forklarende variable *Køn* essentiel. *Nationalitet*, *Kursus_Type* og *Eval_Semester* går igen for de to modeller for MScCM og BScHA og er derfor også af relevans, når vi ser på, hvilke studerende der besvarer, og hvilke der ikke besvarer deres kursusevalueringer. Slutteligt er det værd at nævne den forklarende variabel *Fødselsår*, som vi kun så for BScJUR, men som kan have en effekt på, om de studerende besvarer kursusevalueringen eller ej.

Ved sammenligning af start- og slutmodellernes AIC for MScCM og BscHA ser vi, at AIC'en er blevet lidt mindre. Ved at sammenligne AIC'en for BScJUR's start- og slutmodel kan vi se, at der en noget større forskel i AIC'en. Dette er opsummeret i tabel 16. BScJUR er den

Tabel 16: Overblik over modeller og tilhørende AIC

Uddannelse	Startmodel AIC	Slutmodel AIC
MScCM	14.696	14.691
BScHA	9.079,6	9.071,3
BScJUR	3.558	2.870,8

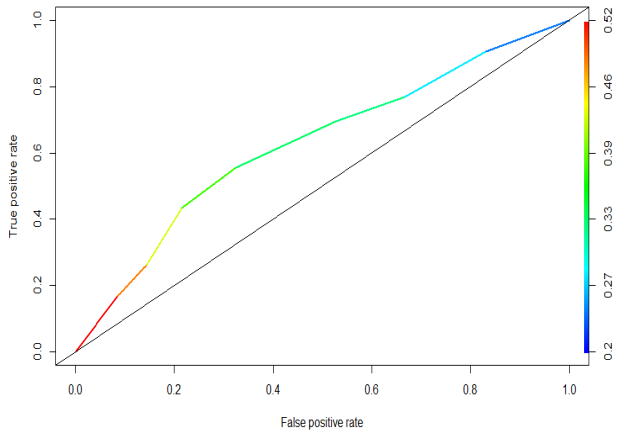
model, der bliver reduceret mest i forhold til startmodellen og indeholder dermed færrest parametre. Antallet af parametre i modellen indgår i udregningen af AIC'en, og det er netop derfor, vi ser et større fald i AIC'en på BScJUR, end for BScHA og MScCM.

4.6 Prædiktion: Hvilke studerende besvarer og hvilke besvarer ikke kursusevalueringen

I afsnittet ovenfor blev der opstillet 3 statistiske modeller for, hvilke studerende der besvarer deres kursusevalueringer. Vi kom frem til en slutmodel for MScCM, ligning 4.2.2, en slutmodel for BScHA, ligning 4.3.2, og en slutmodel for BScJUR, ligning 4.4.2. Disse modeller anvender vi nu til at forsøge at prædiktere på de andre uddannelser, hvilke studerende der besvarer, og hvilke der ikke besvarer deres kursusevalueringer. Til at opsummere de forskellige modellers prædiktive værdi bruger vi en ROC kurve, som plotter "True Positive Rate" (TPR) op imod "False Positive Rate" (FPR). TPR er andelen af observationer, som er blevet prædikteret korrekt, altså der hvor de studerende er korrekt identificeret til at besvare kursusevalueringen. Hvorimod FPR er andelen af observationer, som er blevet prædikteret til at være besvaret, men som ikke er det, altså det er studerende, der er prædikteret til at have besvaret kursusevalueringen, men som ikke har besvaret den. Det er efterfølgende muligt at vurdere, hvor godt modellen prædikterer, ved at beregne "Area under curve" (AUC).

Til at begynde med undersøger vi, hvor godt slutmodellen for MScCM, ligning 4.2.2, prædikterer, hvilke studerende der besvarer deres kursusevaluering på alle andre kandidatuddannelser på CBS. Efter at have forsøgt at prædiktere alle andre kandidatuddannelser er den højeste opnåede AUC på $0,626$. Denne AUC er opnået ved prædiktion af, hvilke studerende der besvarer deres kursusevalueringer på kandidatuddannelsen Master of Science in Social Sciences in Organizational Innovation and Entrepreneurship, MSocScOIE. Det vil sige, at i $62,6\%$ af tilfældene rammer slutmodellen for MScCM, ligning 4.2.2, rigtig på hvilke studerende fra uddannelsen MSocScOIE, der besvarer deres kursusevalueringer. Den tilhø-rende ROC kurve ses nedenfor i figur 20.

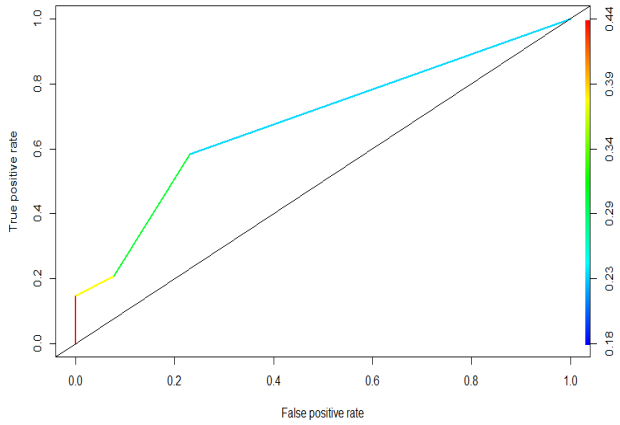
Figur 20: ROC kurve for prædiktion af MSocSCOIE med MScCM slutmodel



For en model med høj prædiktiv værdi vil kurven nærme sig det øvre venstre hjørne, hvilket ikke er tilfældet her. Det ses, at ROC kurven ikke ligger særlig langt fra diagonalen, hvilket stemmer godt overens med, at vi har en AUC på $0,626$. Præcisionen i prædiktionen er dermed ikke ret høj.

Samme fremgangsmåde følges nu for de to resterende modeller BScHA og BScJUR, som blev konstrueret for BScHA i afsnit 4.2 og for BScJUR i afsnit 4.4. Igen ønsker vi at undersøge, om modellerne kan bruges til at prædiktere på andre uddannelser. I og med at begge uddannelser er bacheloruddannelser forsøger vi nu at prædiktere på de resterende bacheloruddannelser. Først forsøger vi med slutmodellen for BScHA fra ligning 4.3.2, og ser hvor godt den prædikterer, hvilke studerende der besvarer deres kursusevalueringer på de resterende bacheloruddannelser. Den højeste AUC vi opnår er på $0,683$, hvilket er, når vi prædikterer på bacheloruddannelsen Bachelor of Science in Business Administration and Digital Management, BScDMA. Den tilhørende ROCkurve ses nedenfor i figur 21.

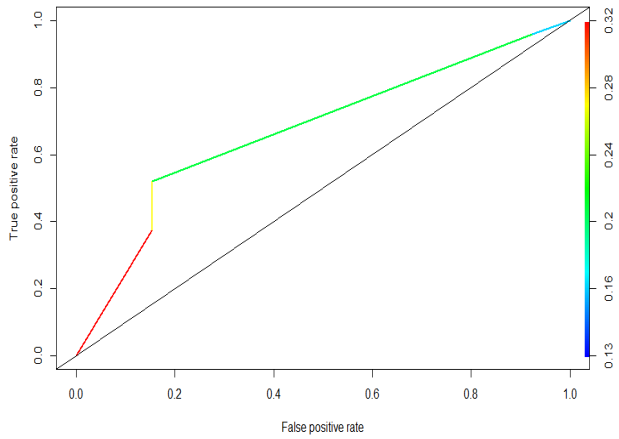
Figur 21: ROC kurve for prædiktion af BScDMA med BScHA slutmodel



Her ser vi, at ROC kurven har et knæk op mod øvre venstre hjørne, altså bevæger sig væk fra diagonalen. Dog er ROC kurven trods knækket tæt op ad diagonalen, og med en AUC på $0,683$ betyder det, at der ikke er nogen høj prædiktiv værdi i modellen. Det er kun i $68,3\%$ af tilfældende, at slutmodellen for BScHA prædikterer korrekt for, hvilke studerende fra uddannelsen BScDMA der besvarer deres kursusevalueringer.

Vi følger nu samme procedure for slutmodellen for BScJUR fra ligning 4.4.2. Den højeste AUC vi opnår er $0,673$. Denne AUC opnår vi på den samme uddannelse, som BScHA også opnår sin højeste AUC på, nemlig bacheloruddannelsen Bachelor of Science in Business Administration and Digital Management, BScDMA. Det er således kun i $67,3\%$ af tilfældene, at slutmodellen for BScJUR prædikterer korrekt for, hvilke studerende fra BScDMA der besvarer deres kursusevalueringer. Den tilhørende ROC kurve ses nedenfor i figur 22.

Figur 22: ROC kurve for prædiktion af BScDMA med BScJUR slutmodel

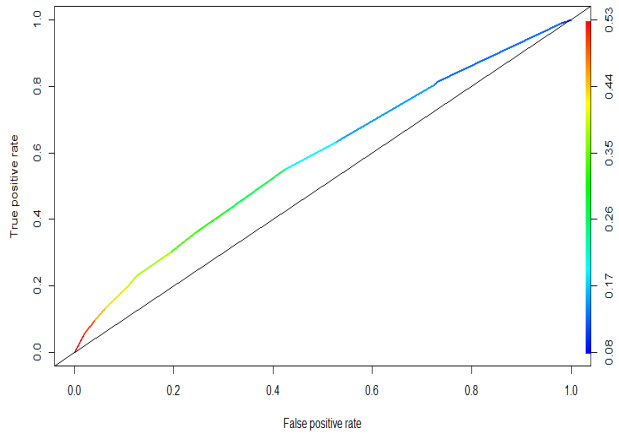


Denne ROC kurve har de samme tendenser som den forrige ROC kurve. Vores AUC er en anelse mindre for denne kurve, end da vi brugte slutmodellen for BScHA til at prædiktere på uddannelsen BScDMA. Uddannelsen BScDMA er den bacheloruddannelsen med færrest antal besvarede kursusevalueringer men dog med den højeste svarprocent.

Generelt set er en AUC omkring $0,65$ ikke særlig høj, da det er meget tæt på diagonalen, som altså angiver, at prædiktionen ikke har bedre odds for at klassificere korrekt, end at slå plat eller krone. Dette betyder, at ingen af de tre modeller har en særlig høj prædiktiv værdi, når vi bruger dem til at se, om de studerende på de andre uddannelser besvarer deres kursusevalueringer.

Vi er nysgerrige på, om modellerne i det mindste kan bruges til at prædiktere besvarelserne på sin egen uddannelse. Dette forsøger vi nu at gøre for BscHA og opdeler derfor data for BScHA i to datasæt, for at se, om BScHA i det mindste kan prædiktere sig selv. Af figur 23 ses ROC kurven, som igen viser sig at ligge tæt op ad diagonalen. Den tilhørende AUC er $0,585$ og dermed den laveste AUC af dem alle. Det betyder, at ikke engang BScHA kan prædiktere "sig selv", da vi ikke opnår en særlig høj prædiktiv værdi. Det er kun er i $58,5\%$ af tilfældende, at BScHA kan prædiktere korrekt på "sig selv", hvor mange studerende der besvarer deres kursusevalueringer.

Figur 23: ROC kurve for prædiktion af BScHA på "sig selv"



Opsummering af prædiktion

Ved at anvende slutmodellen for MScCM til at prædiktere de andre kandidatuddannelser på CBS, blev det uddannelsen MSocScOIE, som slutmodellen bedst kunne prædiktere. Ved at anvende slutmodellen for de to bacheloruddannelser BScHA og BScJUR til at prædiktere de andre bacheloruddannelser på CBS, blev det i begge tilfælde uddannelsen BScDMA, som de to slutmodeller bedst kunne prædiktere. Dog er de opnåede AUC værdier for de tre slutmodeller alle $< 0,7$, og derfor anses vores 3 logistiske regressionsmodeller for at have en lav prædiktiv værdi. Det betyder, at det ikke er muligt at bruge modellerne til at prædiktere, hvilke studerende der besvarer deres kursusevalueringer på de andre uddannelser på CBS. Slutteligt så vi, at ikke engang BScHA kan prædiktere sig selv, da den opnåede AUC siger at kun i $58,5\%$ af tilfældende, rammer modellen rigtigt på, hvilke studerende der besvarer deres kursusevalueringer.

Hvis det havde vist sig at være muligt at anvende modellerne til at prædiktere med en høj prædiktiv værdi, ville det blandt andet betyde, at det for CBS ville være lettere at forudsige, hvilke studerende der besvarer deres kursusevalueringer. På den måde ville det være muligt for CBS at fokusere på at få de studerende til at besvare kursusevalueringen, som sandsynligvis ikke ville besvare den, for på den måde at opnå en højere svarprocent.

5 Hvordan de studerende svarer på kursusevalueringen

Vi vil nu ved anvendelse af statistiske modeller undersøge, hvordan de studerende der besvarer kursusevalueringen svarer på spørgsmålene. Dette ønsker vi at undersøge, da det er interessant at se, om der er nogle effekter der går igen for, hvordan de studerende svarer. Opgavens datasæt indeholder 9 spørgsmål fra CBS's "standard" spørgeramme[24], hvilket bliver for omfattende at analysere. Af både tids- og pladsmæssige årsager udvælger vi derfor ét spørgsmål, som denne analyse laves på.

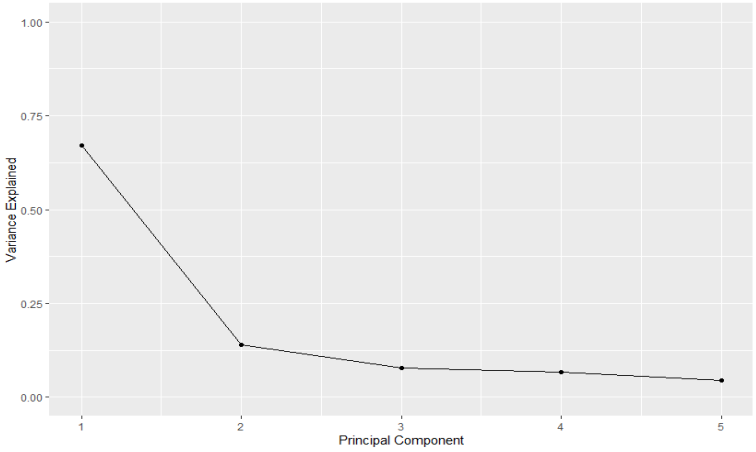
5.1 Anvendt data

Ved at kigge i afsnit 2.3.1, som indeholder spørgeskemaet for kursusevalueringen, fremgår det, at spørgsmålene ret naturligt opdeler sig i to typer af spørgsmål. Den ene type af spørgsmål omhandler selve kurset. Den anden type af spørgsmål indeholder de studerendes egen vurdering af sværhedsgraden af kurset, hvor meget de har forberedt sig til undervisningen, deres deltagelse i undervisning, samt hvorvidt de har anvendt tilbuddet om kontortid. Typen af spørgsmål der omhandler kurset, består af fem spørgsmål, som er interessante at lave denne analyse på. Den anden type af spørgsmål er interessante at anvende som forklarende variable, her er fire spørgsmål.

For at udvælge ét ud af de 5 kursusspørgsmål, som vi vil gå i dybden med i denne analyse, har vi lavet en principal komponent analyse, for at se hvilket af spørgsmålene, der beskriver den største del af variansen i datasættet (James et al, 2013, side 374). Af figur 24 fremgår et screeplot, som beskriver andelen af varians beskrevet af hver principal komponent i analysen. Af screeplottet fremgår det, at den første principale komponent, altså $Kursus_S1$, beskriver den største del af variansen nemlig $0,67$. Vi vælger derfor at fortsætte vores analyse med det første kursusspørgsmål; $Kursus_S1$. Vi vil derfor ikke se nærmere på de andre kursusspørgs-

[24]Spørgerammen for tilstedeværende undervisning

Figur 24: Screeplot af de 5 kursusspørgsmål



mål, altså $Kursus_S2$ som beskriver $0,14$ af variansen, $Kursus_S3$, som beskriver $0,08$ af variansen, $Kursus_S4$, som beskriver $0,07$ af variansen og $Kursus_S5$, som beskriver $0,04$ af variansen. Dette skyldes, at deres principal komponenter ikke beskriver en særlig stor del af variansen i datasættet.

I dette afsnit er det således spørgsmålet $Kursus_S1$, der udvælges til analysen af, hvordan de studerende svarer. $Kursus_S1$ er det spørgsmål, hvor den studerende bliver spurgt indtil i hvilken grad, han/hun er enig i udsagnet: "Jeg har fået et stort fagligt udbytte af undervisningen". Til at analysere hvordan de studerende svarer på dette spørgsmål, har vi flere forskellige variable som er interessante at undersøge, om de har en effekt på, hvordan der bliver svaret.

Til analysen af hvordan de studerende svarer, vil vi anvende følgende variable, som forklarende variable. Fire af de forklarende variable indeholder de demografiske data fra evalueringssystemet vedrørende den studerende: *Køn*, *Nationalitet*, *Fødselsår* og *Karakter*, som alle indeholder to niveauer, og som alle indgår i analysen af, hvilke studerende der besvarer, og hvilke der ikke besvarer kursusevalueringen. Yderligere anvender vi de fire forklarende variable, som er spørgsmål den studerende har besvaret i kursusevalueringen; *Pensum*, $Tidsforbrug$, *Fremmøde* og *Kontortid* som indeholder to eller flere niveauer.

I forbindelse med analysen af hvilke studerende der besvarer og ikke besvarer kursusevalueringerne, i afsnit 4, lavede vi en del filtreringer i datasættet. Samme filtreringer er anvendt i denne analyse omkring hvordan de studerende svarer. Dog foretages der yderligere filtreringer. Til denne analyse filtreres af variablen $Resp_Status$ niveauet "DNR" ud, så datasættet kun indeholder besvarelser, som er "Submitted". Karaktervariablen inddeles i to niveauer "7 eller under" og "Over 7".

For at undersøge hvordan de studerende besvarer $Kursus_S1$, opstiller vi en kumulativ logit regressionsmodel med $Kursus_S1$ som responsvariabel, som er en faktor med 5 niveauer, J=5[25]. Vi har således med en ordinal multikategorisk logistisk regression at gøre, da en ordinal regressionsmodel bruges til at beskrive forholdet mellem ordnede kategoriske responsvariable og en eller flere forklarende variable. Da vi har med den ordinale klasse at gøre, anvendes en kumulativ logit model. Vi anvender likelihood ratio test til at vurdere effekterne af de forklarende variable.

Til at analysere hvordan de studerende der besvarer deres kursusevalueringer svarer, er der nu udvalgt et specifikt spørgsmål samt en række forklarende variable, som antages at have en effekt på, hvordan de studerende svarer. Vi har valgt at fokusere på uddannelsen Master of Science in Business, Language and Culture, MScBLC, da det er en uddannelse, af mellem størrelse på CBS, med $n = 1674$, hvoraf der er 810 "Submitted" kursusevalueringer. Dette fører til en svarprocent på $48,39\%$, som er den næst højeste svarprocent på CBS.
Den høje svarprocent kombineret med at uddannelsen er en international uddannelse, og dermed har studerende fra både "Norden" og "Udenfor Norden" har ført til, at vi har valgt at lave vores analyse på denne uddannelse.

[25] 1 = Helt uenig, 2 = Overvejende uenig, 3 = Hverken enig eller uenig, 4 = Overvejende enig, 5 = Helt enig

Anvendt R-kode

Til følgende afsnit anvendes pakkerne i R: "ordinal", "stats" og "ggplot2". I bilag F.2 fremgår et eksempel på anvendelse af R-kode, som bruges til at udarbejde denne analyse.

5.2 Kumulativ logit model for Master of Science i Business Language and Culture

Inden vi laver en model for hele uddannelsen MScBLC, undersøger vi først nogle forskellige kurser fra uddannelsen enkeltvis. Til at undersøge disse kurser har vi valgt kun at inkludere de fire forklarende variable, som indeholder de demografiske data om de studerende. Her er følgende kurser for MScBLC blevet valgt tilfældigt: "CCBDO1002U_E18", "CCBDO1007U-_E18", "CCDCO2006U_E18" og "CCDCO2002U_F19". Parameter estimater (P.E.) og standardafvigelserne (Std.afv.) for de fire udvalgte kurser fremgår i tabel 17. Tabellen indeholder kun parameter estimater og standardafvigelser for fire af de forklarende variable, for hver af de fire kurser, da det ikke er lykkedes at inkludere alle otte forklarende variable grundet et begrænset datasæt.

Tabel 17: Parameter estimater og standard afvigelser for fire kurser på MScBLC

	CCBDO1002U		CCBDO1007U		CCDCO2006U		CCDCO2002U	
	P.E.	Std.afv.	P.E.	Std.afv.	P.E.	Std.afv.	P.E.	Std.afv.
KønM	-0,625	0,625	-0,989	0,766	-0,551	0,750	-0,566	0,950
Udenfor Norden	-0,652	0,601	-0,826	0,710	-0,161	0,636	-2,521	0,968
Fødselsår1996-2000	-0,486	0,842	-0,471	0,898	-1,409	1,053	-0,208	1,544
KarakterOver7	0,825	0,610	0,238	0,683	1,417	0,886	1,559	1,126

Positive parameter estimater indikerer større sandsynlighed for højere valg af svarkategori sammenlignet med baseline, når alle andre variable holdes konstant. Vi kan se, at den forklarende variabel *Karakter* med niveauet "Over 7" har positivt fortegn og indikerer

således, at de studerende med en opnået karakter "Over 7" er mere tilbøjelige til at svare højere på $Kursus_S1$, end de studerende med opnået karakter fra niveauet "7 og under". På samme vis kan vi se, at alle andre parameter estimater har negativt fortegn, hvilket betyder, at de studerende fra disse kategorier, er mere tilbøjelige til at svare lavere på $Kursus_S1$, end den tilhørende baseline. Dette er et godt første indblik i, hvordan de forklarende variable opfører sig, når de enkelte kurser analyseres. Det fremgår af tabel 17 at parameter estimaterne for de fire udvalgte kurser har samme fortegn for de samme forklarende variable.

Vi ønsker at lave en statistisk model for én af de fire kurser, og til dette er kurset CCDCO2002-U_E18 udvalgt. Til at udvælge den bedst mulige model anvendes backwards elimination. Der opstilles derfor først en startmodel, som indeholder alle de mulige forklarende variable og herefter fjernes de insignifikante variable en efter en, indtil vi har en endelig, reduceret model, som i denne opgave kaldes for en slutmodel.

$Kursus_S1$'s svarskala er af den ordinale skala repræsenteret af Y_i, som har værdi j, hvis den i'te observation falder i den j'te kategori, hvor j = 1, 2, 3, 4, 5. Vores udvalgte kursus, CCDCO2002U_E18, har dog kun besvarelser i kategorierne j = 3, 4, 5, hvilket giver os en model med en responsvariabel, der har 3 kategorier. Startmodellen indeholder alle forklarende variable og opstilles nu:

$$logit[P(Y_i \leq j)] = \alpha_j - \beta_{Køn} - \beta_{Nationalitet} - \beta_{Fødselsår} - \beta_{Karakter} - \beta_{Pensum} - \beta_{Tidsforbrug} - \beta_{Fremmøde} - \beta_{Kontortid} \tag{5.2.1}$$

Da vi ser på kurserne enkeltvis, er afhængigheden i data ikke et problem, som vi så det i afsnit 4, da de studerende kun kan være tilmeldt et kursus én gang.

Startmodellen fra ligning 5.2.1 har en AIC på 59, 38. De forklarende variable *Køn*, *Fødselsår*, $Karakter$, $Tidsforbrug$, *Fremmøde* og $Kontortid$ viser sig alle at være insignifikante[26], og

[26]Se Bilag D for p-værdier

vi fjerner dem derfor en af gangen ved at følge backwards elimination metoden. Der testes løbende med likelihood ratio test (LRT), hvor den simple model testes op imod den mere komplekse model.

Ved en LRT for startmodellen og en model hvor variablen *Køn* udelades, opnås en p-værdi på $0,127$ og på baggrund af denne, udelades variablen fra modellen. Efterfølgende forsøger vi yderligere at udelade variablen *Fødselsår*, hvor vi ved en LRT får en p-værdi på $0,279$, som indikerer, at også denne variabel kan udelades fra modellen. Efterfølgende gøres det samme for variablen $Karakter$, hvor LRT giver en p-værdi på $0,240$, og dermed udelades den også fra modellen. Herefter forsøger vi at fjerne $Tidsforbrug$, hvor LRT giver en p-værdi på $0,266$, der angiver, at denne variabel også kan udelades. Vi forsøger nu også at udelade *Fremmøde* fra modellen, hvor LRT giver en p-værdi på $0,9695$ og derfor udelades den fra modellen. Slutteligt forsøger vi at fjerne variablen $Kontortid$, hvor LRT giver en p-værdi på $0,365$, som afgør, at den også udelades fra modellen. Alle de variable som var insignifikante i startmodellen er dermed udeladt fra modellen. Slutmodellen indeholder dermed de forklarende variable $Nationalitet$ og $Pensum$ og har en AIC på $51,47$, som er mindre end startmodellen. Slutmodellen har følgende form:

$$logit[P(Y_i \leq j)] = \alpha_j - \beta_{Nationalitet} - \beta_{Pensum} \tag{5.2.2}$$

For at lette fortolkningen af de forklarende variables effekt på responsvariablen og altså dermed, hvordan den studerende besvarer $Kursus_S1$, ser vi i tabel 18 på de kumulative odds ratioer og deres tilhørende konfidensintervaller for slutmodellen.

Tabel 18: Odds ratio og tilhørende konfidensinterval for parametrene i slutmodellen 5.2.2 for kursus CCDCO2002U_F19

	Odds ratio	2,5%	97,5%
NationalitetUdenfor Norden	0,106	0,016	0,531
PensumOver middel	0,102	0,014	0,561

Vi ser at for de studerende med nationalitet "Udenfor Norden", er odds ratioen for at svare højt på svarskalaen på $Kursus_S1$, 89,4%[27] lavere, end det er for de studerende med nationalitet "Norden", når alle andre variable holdes konstant.
Den forklarende variabel *Pensum* har tre svarkategorier: "Under middel", "Middel" og "Over middel". Af tabel 18 fremgår kun niveauet "Over middel", hvilket skyldes, at der ikke er nogen studerende, der har svaret at kurset har være "Under middel", og dermed er niveauet "Middel" baseline. De studerende der vurderer *Pensum* til at være "Over middel" har 89,8% lavere odds ratio for at svare højt på svarskalaen til $Kursus_S1$, end de studerende der har svaret at *Pensum* var "Middel". Dette giver umiddelbart god mening, da $Kursus_S1$ spørger til, om den studerende har fået stort fagligt udbytte, og hvis den studerende har vurderet at *Pensum* har været "Over middel", betyder det, at han/hun synes kurset har været svært. En studerende der har fundet kurset svært har derfor formentlig ikke fået lige så stort fagligt udbytte, som en studerende der har fundet pensums niveau "Middel". Det fremgår af konfidensintervallerne i tabel 18 at det er med 95% sandsynlighed at odds ratioen for $Nationalitet\ UdenforNorden$ er i intervallet $[0,016; 0,531]$, og at odds ratioen for *Pensum Over middel* er i intervallet $[0,014; 0,561]$.

5.2.1 Test af proportional odds antagelsen

Vi vil nu teste, om proportional odds antagelsen holder for de forklarende variable i slutmodellen. En måde at teste dette på er ved at lade threshold koefficienterne afhænge af β'erne, altså regressionskoefficienterne tilhørende de forklarende variable. Da slutmodellen, fra ligning 5.2.2, indeholder to forklarende variable, undersøger vi om proportional odds antagelsen holder for dem begge, ved at opstille en proportional odds model, for dem hver. Hver af disse to modeller testes op imod slutmodellen med en LRT for at se, om antagelsen holder eller ej. Nedenfor viser vi, hvordan det er gjort med *Pensum* som eksempel.

[27] $((1 - 0,106)100\%)$

Vi lader threshold koefficienterne afhænge af *Pensum*, som leder til en såkaldt partial proportional odds (PPO) model for *Pensum*, som har følgende form:

$$logit[P(Y_i \leq j)] = \alpha_j - \beta_{Nationalitet} + \tilde{\beta}_{j,Pensum} \tag{5.2.3}$$

Vi tester nu ovenstående model, ligning 5.2.3, op i mod vores slutmodel fra ligning 5.2.2, ved at udføre en LRT, for på den måde at undersøge, om proportional odds antagelsen bliver overholdt for *Pensum*. Det samme er gjort for *Nationalitet*, og begge resultater fremgår af tabel 19.

Tabel 19: LRT af PPO modellerne mod slutmodellen, ligning (5.2.2)

	Log-likelihood	P-værdi
Slutmodel	-21,73	
PPO model, Pensum	-21,68	0,740
PPO model, Nationalitet	-21,18	0,291

Det ses, at der kun er en lille forskel i log-likelihood for begge proportional odds modellerne op imod slutmodellen, og begge LRT er derfor insignifikante. Der er derfor ingen evidens for, at proportional odds antagelsen bliver brudt for hverken *Pensum* eller *Nationalitet*.

5.2.2 Goodness-of-fit test

Til at slutte af med laver vi en Hosmer-Lemeshows goodness-of-fit test på vores slutmodel fra ligning 5.2.2, for at se, hvor godt modellen fitter vores datasæt. Nulhypotesen for goodness-of-fit testen er, at modellen fitter datasættet godt. Den alternative hypotese er, at der er et uspecifikt problem med fittet, også kaldet "lack of fit" (Fagerland, 2017). Vi får en p-værdi på $0,159$, og vi kan derfor ikke afvise vores nulhypotese. Dermed er der evidens for, at vores model fitter datasættet godt.

5.2.3 Opsummering af kumulativ logit model (CCDCO2002U)

Vi har i dette afsnit undersøgt, om der er nogle af de forklarende variable, der har effekt på, hvordan de studerende svarer på spørgsmålet $Kursus_S1$. Først så vi på fire forskellige kurser for MScBLC, at parameter estimaterne havde samme fortegn, for de demografiske data om den studerende. Efterfølgende opbyggede vi en kumulativ logit model for kurset CCDCO2002U_E18, hvor det viste sig at være de forklarende variable $Nationalitet$ og $Pensum$, som var signifikante for, hvordan de studerende svarer på $Kursus_S1$.
Vi så, at de studerende med $Nationalitet$ "Udenfor Norden" er tilbøjelige til at svare, at de har fået et lavere fagligt udbytte af undervisningen end de studerende fra "Norden". Det samme så vi om de studerende, der i $Pensum$ spørgsmålet har vurderet at pensum var "Over Middel" og dermed tilkendegivet, at de synes pensum har været svært. Også disse studerende viste sig at være mere tilbøjelige til at sige, at de har fået et mindre fagligt udbytte af kurset, end de studerende der svarede på pensum spørgsmålet, at de synes $Pensum$ var "Middel", altså passende sværhedsgrad.

Efter at have tolket på de forklarende variable i slutmodellen testede vi med en LRT, om proportional odds antagelsen bliver overholdt for de forklarende variable, der er indeholdt i slutmodellen; $Nationalitet$ og $Pensum$, hvilket vi så at den gør. Slutteligt udførte vi en goodness-of-fit test på slutmodellen for at se, om vores model fitter data godt. Med en p-værdi på $0,159$ så vi, at slutmodellen ikke strider mod modellens antagelse om proportionalitet.

5.3 Kumulativ logit model med tilfældig effekt for Master of Science in Business Language and Culture

I forrige afsnit, afsnit 5.2, sammenlignede vi parameter estimaterne for fire tilfældige kurser på MScBLC, hvor vi så, at parameter estimaterne alle har samme fortegn. Da alle estimaterne er nogenlunde ens for kurserne for MScBLC giver det mening med en mixed model, hvor de faste effekter ikke afhænger af kurset. Når vi senere vil sammenligne kurserne, kan vi gøre

dette ved at analysere vores mixed effekt model.

Vi opstiller nu en kumulativ logit model med tilfældig effekt for uddannelsen MScBLC, hvor kursuskoden sættes til at være den tilfældige effekt og med $Kursus_S1$ som responsvariablen. Startmodellen indeholder følgende forklarende variable: *Køn*, $Nationalitet$, $Karakter$, *Fød−selsår*, $Pensum$, $Tidsforbrug$, *Fremmøde* og $Kontortid$. Formålet er igen at se på, hvordan de studerende svarer på $Kursus_S1$, men denne gang på et større datasæt, da vi ser på hele uddannelsen MScBLC. Ved at se nærmere på datasættet ser vi, at hele svarskalaen for $Kursus_S1$ bruges, dog er der næsten ingen svar i kategori 1, her er kun 5 besvarelser, og kategori 2, her er kun 29 besvarelser, derfor slår vi dem sammen under "2"[28]. Responsvariablen $Kursus_S1$ indeholder nu således 4 niveauer. Opdelingen af $Kontortid$ har vi også ændret, så den nu indeholder to niveauer, som beskriver, om den studerende har anvendt kontortid, niveau "Ja" eller ej, niveau "Aldrig"[29]. Vi har dermed følgende startmodel:

$$\begin{aligned} logit[P(Y_i \leq j|u_{Kursus_Kode})] = \alpha_j - \beta_{Køn} - \beta_{Nationalitet} - \beta_{Karakter} - \beta_{Fødselsår} \\ -\beta_{Pensum} - \beta_{Tidsforbrug} - \beta_{Fremmøde} - \beta_{Kontortid} - u_{Kursus_Kode} \end{aligned} \tag{5.3.1}$$

Modellen fittes med Gauss-Hermite quadrature method, som beskrevet i afsnit 3.3.1, og vi ser, at *Køn*, $Nationalitet$, $Pensum$, $Tidsforbrug$ og $Kontortid$ er signifikante[30]. Startmodellen har en AIC på $1.143,26$. Da vi er startet med en fuld model og ønsker at fjerne de insigifikante variable, følger vi backwards elimination. De insignifikante forklarende variable forsøges nu fjernet en efter en. Vi forsøger først at fjerne $Karakter$, og det testes med LRT, at variablen godt kan udelades, da p-værdien er $0,328$. På samme måde forsøger vi nu at fjerne *Fødselsår*, hvor en LRT giver en p-værdi på $0,478$, og af denne grund udelades variablen fra modellen. Slutteligt forsøger vi at fjerne variablen *Fremmøde*, som viser sig også at kunne fjernes fra modellen, da en LRT giver p-værdien er $0,388$.
Slutmodellen indeholder de forklarende variable: *Køn*, $Nationalitet$, $Pensum$, $Tidsforbrug$

[28] Antallet af studerende der har svaret: 3 = 51 besvarelser, 4 = 251 besvarelser, 5 = 291 besvarelser
[29] Antallet af studerende der har svaret: 1 = 568 studerende, 2 = 56, 3 = 4 studerende, 4 = 0 studerende
[30] Se Bilag E for p-værdier

og *Kontortid* og har en AIC på: 1.136, 86, som er lavere end vores startmodel og beskriver således, at vores slutmodel fitter bedre. Slutmodellen har følgende form:

$$logit[P(Y_i \leq j|u_{Kursus_Kode})] = \alpha_j - \beta_{Køn} - \beta_{Nationalitet} - \beta_{Pensum} - \beta_{Tidsforbrug} - \beta_{Kontortid} - u_{Kursus_Kode} \tag{5.3.2}$$

For at tolke på effekten af de forklarende variable ser vi nu på de kumulative odds ratioer. Den kumulative odds ratio bestemmes ved at se på begivenheden $Y \leq j$, som bestemmes ved $exp(-\beta)$. Nedenfor i tabel 20 fremgår odds ratioerne for slutmodellen, ligning 5.3.2's, forklarende variable.

Tabel 20: Odds ratioer og tilhørende konfidensintervaller for slutmodellen ligning 5.3.2

	Odds ratio	2,5%	97,5%
KønM	0,628	0,427	0,925
NationalitetUdenfor Norden	0,566	0,398	0,807
PensumOver middel	0,680	0,423	1,093
PensumUnder middel	0,230	0,091	0,486
Tidsforbrug2	1,113	0,655	1,889
Tidsforbrug3	1,927	1,084	3,428
Tidsforbrug4	2,385	1,301	4,372
Tidsforbrug5	2,102	1,104	4,000
KontortidJa	3,842	1,875	7,874

Vi ser, at odds ratioen for *KønM* er 0, 628. Det betyder, at odds ratioen for at en mandlig studerende svarer højt på svarskalaen på $Kursus_S1$, er 37,2% lavere end for en kvindelig studerende, når alle andre variable holdes konstant. Dette betyder, at de mandlige studerende er tilbøjelige til at svare, at de har fået et lavere fagligt udbytte af kurset end kvinderne.
Når vi ser på den studerendes nationalitet, kan vi se, at odds ratioen for at en studerende "Udenfor Norden" svarer højt på svarskalaen på $Kursus_S1$, er 43,4% lavere, end det er

for en studerende fra "Norden", når alle andre variable holdes konstant. Det vil sige, at de studerende fra "Norden" er mere tilbøjelige til at sige, at de har fået et større fagligt udbytte af kurset end de studerende fra "Udenfor Norden".

Variablen $Pensum$ beskriver den studerendes egen vurdering af, hvordan pensums sværhedsgrad var for kurset. Her kan vi se at både de studerende der svarer "Over middel", altså de studerende der synes at kurset har været svært, samt de studerende der svarer "Under middel", altså de studerende der synes kurset har været let, har lavere odds ratio for at svare højt på svarskalaen for $Kursus_S1$, end de studerende der synes at pensums sværhedsgrad var "Middel". De studerende der har svaret at pensums sværhedsgrad var "Under middel" er langt mindre tilbøjelige til at svare højt på svarskalaen for $Kursus_S1$. En forklaring på dette kan være, at hvis man som studerende har fundet kurset for let, så er det begrænset, hvor stort et fagligt udbytte man har fået. Omvendt, hvis kurset har været svært, kan det også være begrænset, hvor stort et fagligt udbytte man har fået og derfor giver det også mening, at disse studerende er mindre tilbøjelige til at svare højt på svarskalaen for $Kursus_S1$, end de studerende der har fundet pensums sværhedsgrad "Middel" og dermed passende. Konfidensintervallet for odds ratioen for "Over middel", som er et niveau i $Pensum$, er markeret med rødt, da det spænder over et interval, hvor 1 indgår. Selvom variablen $Pensum$ er signifikant, da er selve niveauet "Over middel", ikke signifikant og netop dette angiver konfidensintervallet.

Variablen $Tidsforbrug$ beskriver den studerendes egen vurdering af, hvor lang tid han/hun har brugt på at forberede sig i gennemsnit til hver undervisningslektion i kurset. Her kan vi se, at odds ratioerne er voksende i takt med, at den studerendes tidforbrug er voksende, dog ser vi et lille fald i odds ratioen fra niveau "4" til niveau "5". Ser vi på den højeste odds ratio for variablen $Tidsforbrug$, som er niveau "4", er odds ratioen $2,39$ gange højere for, at en studerende i denne kategori svarer højt på svarskalaen for $Kursus_S1$, end det er for en studerende der har svaret niveau "1". Umiddelbart giver det god mening, at de studerende der har brugt længere tid på at forberede sig, har fået et større fagligt udbytte

end de studerende, der nærmest ingen tid har brugt på forberedelse til undervisningen. Konfidensintervallet for odds ratioen for variablen $Tidsforbrug$ for niveauet "2" er markeret med rødt, da intervallet spænder over 1. Dette indikerer, at selvom variablen $Tidsforbrug$ er signifikant, da er niveauet "2" det ikke, hvilket konfidensintervallet understreger. Det fremgår også af tabellen, at odds ratioen for $Tidsforbrug$ ved niveau "2" er tæt på 1, hvilket igen viser, at niveauet ikke er signifikant.
Afslutningsvis kan vi se, at odds ratioen for at de studerende, der har benyttet sig af tilbuddet om kontortid hos underviseren, har svaret højt på $Kursus_S1$ er $3,842$ gange større, end det er for studerende der ikke har benyttet sig af kontortiden. Dette er ret interessant, da kontortid generelt ikke er noget de studerende anvender ret meget, hvilket vi så i krydstabel 8. Dog har det vist sig at de studerende på uddannelsen MScBLC benytter sig af tilbuddet om kontortider, og de studerende har været mere tilbøjelige til at svare, at de har fået stort fagligt udbytte af kurset, hvis de har anvendt tilbuddet.

5.3.1 Test af model med mixed effekt

Nu da vi har en slutmodel, ønsker vi at teste, om det gør fittet af vores model bedre at have kursuskode som tilfældig effekt, eller om vores model ville være lige så god uden. Til at undersøge dette kan vi se på standardafvigelsen for den tilfældige effekt σ_u. Af tabel 21 fremgår det, at standardafvigelsen er $0,616$. Hvis standardafvigelsen var 0, ville det indikere at den tilfældige effekt ikke har en effekt. Vi opstiller nu en model magen til vores slutmodel, ligning 5.3.2, blot uden en tilfældig effekt, og tester de to modeller med en LRT. Testen giver en p-værdi på $6,167e-05$, og derfor mener vi, at $Kursus_Kode$ er signifikant. Vi foretrækker derfor modellen med $Kursus_Kode$ som tilfældig effekt.

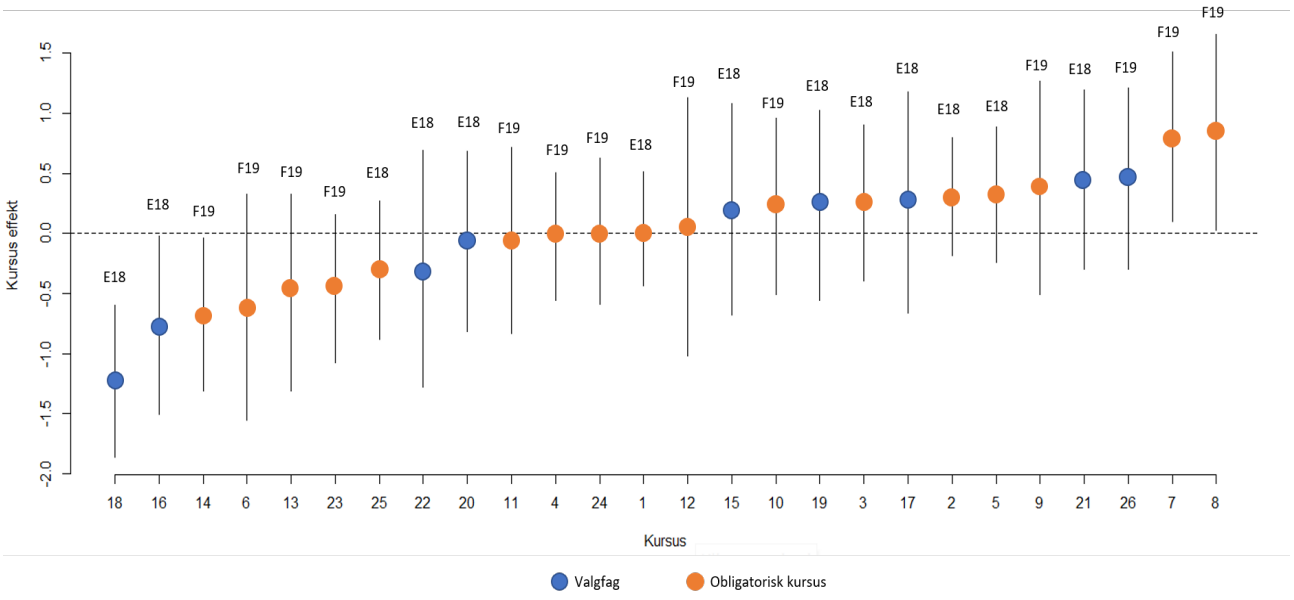
Tabel 21: Profil likelihood konfidensinterval for standardafvigelsen for den tilfældige effekt

	Værdi	2,5%	97,5%
Standardafvigelse	0,616	0,357	0,967

5.3.2 Det "gennemsnitlige" kursus

Den tilfældige effekt, $Kursus_Kode$, kan ikke estimeres på almindeligvis, da det ikke er en parameter, men vi kan give et "bedste gæt" ved at se på den betingede modus. Den betingede modus er forskellen mellem den gennemsnitlige prædikterede respons for et givent sæt af faste effekter, og responsen prædikteret for et bestemt individ. Det er dermed hvor meget ét kursus adskiller sig fra alle kurserne samlet. Den betingede varians giver et usikkerhedsmål for den betingede modus.
Nedenstående figur 25 illustrerer effekterne af $Kursus_Kode$ via den betingede modus med 95% konfidensinterval baseret på den betingede varians.

Figur 25: Kursuseffekt givet den betingede modus med 95% konfidensinterval baseret på den betingede varians



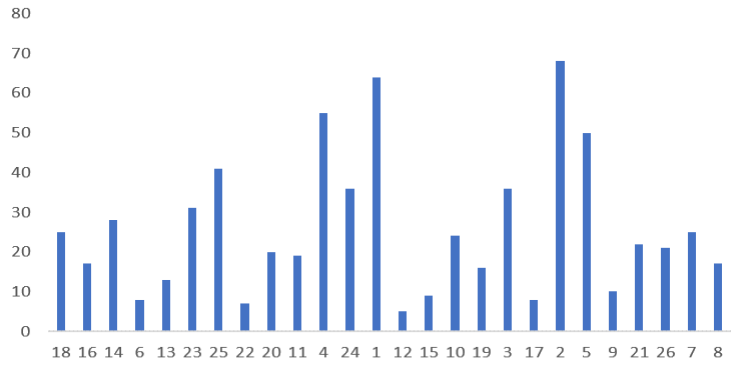
Af figur 25 kan vi se, at kurserne har forskellige niveau, så vi får ordnet kurserne efter deres effekter for, hvor højt de studerende opfatter deres "faglige udbytte" korrigeret for de øvrige effekter. Den horisontale linje viser den gennemsnitlige kursuseffekt.

Det fremgår, at kursus nummer 18 i datasættet er det kursus, som svarer lavest på $Kursus_S1$,

hvor kursus nummer 8 svarer højest på $Kursus_S1$. Den signifikante kursuseffekt indikerer, at de studerende vurderer kurserne forskelligt i forhold til, hvor "stort fagligt udbytte" de har fået. Dette kan tolkes på to måder: Enten har "stort fagligt udbytte" forskellig betydning for de forskellige studerende, eller også er der forskel på, hvor "stort fagligt udbytte", der er opnået i de forskellige kurser. Det er meget sandsynligt at begge dele er korrekt, og derfor spiller en rolle i modellen. Vi ved, ud fra egen erfaring, at forskellen på kurserne kan være stor, og at der derfor kan være forskel på, hvor "stort fagligt udbytte" man som studerende opnår.

Ved at se lidt nærmere på de informationer vi har om de 26 forskellige kurser, der indgår i figur 25, ser vi efter, om der er noget, der kan forklare den forskel, vi ser i kursuseffekten. Ved at se på $Kursus_Type$[31] ser vi, at der ikke umiddelbart er en sammenhæng mellem, om kurset er et valgfag (blå) eller et obligatorisk kursus (orange), og om kurset ligger i toppen eller i bunden. Ser vi på om kurset er udbudt i efteråret 2018 (E18) eller i foråret 2019 (F19), er der heller ikke her noget mønster at spotte, både top og bund er en fin blanding af kurser fra E18 og F19.

Figur 26: Antal studerende per kursus



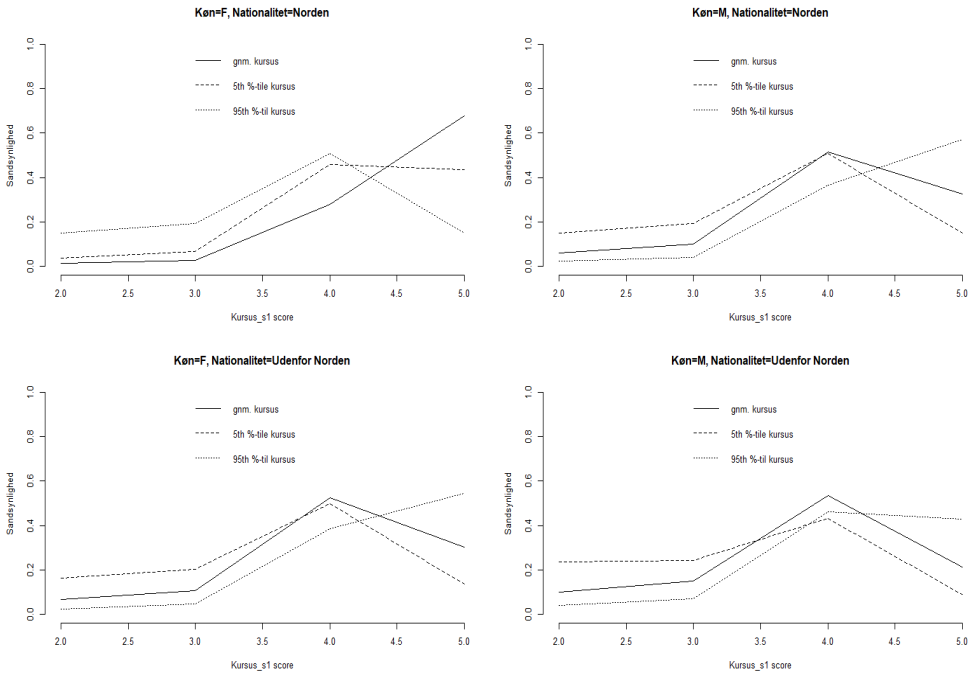
Vi har også undersøgt antallet af studerende på de forskellige kurser, som fremgår af figur 26, men heller ikke her ser vi et mønster. Ud fra den fakta vi har om kurserne, er der således

[31] $Kursus_Type$ angiver om kurset er et obligatorisk kursus, eller et valgfag

ikke nogen klar fortolkning af, hvilke kurser der har høj kursuseffekt, og hvilket der har lav kursuseffekt.

Vi kan nu beregne sandsynlighederne $P(Y = J)$ for j = 1, 2, 3, 4, 5, for det gennemsnitlige, det 5'te og det 95'te percentile kursus, for slutmodellens, 5.3.2, forskellige betingelser. På figur 27 ses der på det gennemsnitlige, det 5'te og det 95'te percentile kursus, hvor de forklarende variable *Køn* og *Nationalitet* holdes konstant, for hver af de mulige kombinationer af variablene. Det fremgår, at de kvindelige studerende fra "Norden" har en langt større sandsynlighed for at vælge svarkategorien 5 til spørgsmålet $Kursus_S1$ end den kvindelige studerende med nationalitet "Udenfor Norden", når vi ser på det gennemsnitlige kursus. Dette mønster gentages ikke for de mandlige studerende for henholdsvis "Norden" og "Udenfor Norden". Det fremgår, at sandsynlighederne for at vælge diverse svarkategorier ikke ændres ved at se på henholdsvis de mandlige studerende fra "Norden" og "Udenfor Norden". Det ligner dermed, at de mandlige studerende har nogenlunde samme mønster i, hvordan de svarer på $Kursus_S1$.

Figur 27: Svarskala sandsynlighed for gennemsnitligt kursus og yderkanterne



Dette stemmer overens med, hvad vi så, da vi så på de kumulative odds ratioer. Her fremgik det, at de kvindelige studerende havde tendens til at svare højere på svarskalaen for $Kursus_S1$ end de mandlige studerende. Det samme gjorde sig gældende for de studerende fra "Norden", da vi så, at de har højere odds ratio for at svare højt på svarskalaen end de studerende "Udenfor Norden". Så kombinationen af, at de kvindelige studerende fra "Norden" ser ud til at have større sandsynlighed for at svare højt på svarskalaen end både de kvindelige studerende "Udenfor Norden" samt for de mandlige studerende, kommer ikke som nogen overraskelse.

5.3.3 Opsummering af kumulativ logit model med tilfældig effekt (MScBLC)

Vi har i dette afsnit lavet en kumulativ logit model med $Kursus_Kode$ som tilfældig effekt. Slutmodellen 5.3.2 indeholder de forklarende variable; $Køn$, $Nationalitet$, $Pensum$, $Tidsforbrug$ og $Kontortid$. Det fremgår af den kumulative logit model tilfældig effekt, at den studerendes vurdering af sin egen indsats i kurset har en effekt på, hvordan den studerende svarer på $Kursus_S1$.

Fra de kumulative odds ratioer fremgik det, at de mandlige studerende er mindre tilbøjelige til at svare, at de har fået højt fagligt udbytte af undervisningen end de kvindelige studerende. Vi så også fra odds ratioerne, at de studerende med nationalitet "Udenfor Norden" er mindre tilbøjelige til at svare, at de har fået højt fagligt udbytte af undervisningen end de studerende fra "Norden". De studerende der svarede, at pensums sværhedsgrad var "Middel", er mere tilbøjelige til at svare, at de har fået højt fagligt udbytte af undervisningen end de studerende, der har svaret enten "Over middel" eller "Under middel". $Tidsforbrug$ indikerer, at jo mere tid den studerende har brugt på at forberede sig til hver enkelt lektion, jo mere tilbøjelig er de studerende til at svare, at de har fået højt fagligt udbytte af undervisningen. Slutteligt så vi, at de studerende der har benyttet sig af kontortid er mere tilbøjelige til at svare, at de har fået stort fagligt udbytte af undervisningen, end de studerende der ikke har benyttet sig af kontortiden.

Efter at have set på odds ratioerne for de forskellige forklarende variable i slutmodellen testede vi, om modellen med tilfældig effekt er at foretrække fremfor en model uden tilfældig effekt. Det viste sig, med en p-værdi på $6,167e-05$ fra en LRT at vores model med tilfældig effekt er at foretrække. Efterfølgende så vi, hvordan et gennemsnitlig kursus ser ud, og hvordan kursuseffekterne ser ud på tværs af kurserne. Vi kunne ikke komme frem til nogen fortolkning af, hvilke kurser der har høj kursuseffekt, og hvilke der har lav kursuseffekt.
Slutteligt undersøgte vi det gennemsnitlige, 5'te og 95'te percentile kursus, hvor de forklarende variable *Køn* og *Nationalitet* holdes konstant for deres kombinationer. Det der sprang mest i øjnene var, at de kvindelige studerende fra "Norden" har størst sandsynlighed for at svare, at de har fået højt fagligt udbytte af kurset. Dette var dog ikke nogen ny information, da dette hænger godt sammen med, hvad der blev tolket ud fra slutmodellens kumulative odds ratioer.

6 Opsamling af opgaven

I dette afsnit fremstiller vi først vores resultater, og efterfølgende diskuterer vi dem. Afslutningsvis belyser vi, hvad CBS kan anvende vores resultater til, samt reflekterer over vores læring i forbindelse med denne specialeproces.

6.1 Resultater

Vi er kommet frem til, at det er muligt at opstille statistiske modeller til brug for analyse af data fra kursusevalueringer. Gennem vores første analyse er vi kommet frem til, at den logistiske regressionsmodel kan anvendes til brug for analyse af, hvilke studerende der besvarer, og hvilke studerende der ikke besvarer kursusevalueringen. Selvom der er afhængighed i data, som kan forårsage overspredning, er det muligt at anvende en logistisk regressionsmodel, fremfor en quasibinomial model. Afhængigheden i data har dermed ikke stor indvirkning på variansen. Yderligere er vi kommet frem til, at til brug for analyse af hvilke forhold der er bestemmende for de studerendes "faglige udbytte" af kurset, kan en kumulativ logit model med tilfældig effekt anvendes.

Ud fra de logiske regressionsmodeller, som er opstillet i analysen, kom vi frem til, at det ikke er muligt at forudsige hvilke studerende der besvarer, og hvilke der ikke besvarer kursusevalueringen. I forsøget på at forudsige, hvilke studernede der besvarer, og hvilke der ikke besvarer kursusevalueringen, kom vi frem til, at vores slutmodeller har en lav prædiktiv værdi. Dette viste sig, at gøre sig gældende både ved prædiktion på andre uddannelser, men også ved prædiktion af en uddannelse "på sig selv".
Dog angiver vores logistiske regressionsmodel at effekterne køn, nationalitet, fødselsår, kursustype og semester, er bestemmende for hvilke studererende der besvarer, og hvilke der ikke besvarer deres kursusevalueringer. Køn, nationalitet og semester er de tre forklarende variable der uden tvivl er bestemmende for, om de studerende besvarer deres kursusevalueringer eller ej. Fra variablen køn så vi, at odds ratioen for at de mandlige studerende besvarer kursusevalueringen er lavere end for de kvindelige studerende. Af variablen nationalitet så

vi, at odds ratioen for at de studerende fra "Norden" besvarer deres kursusevalueringer er lavere end for de studerende med nationalitet "Udenfor Norden". Af variablen semester så vi, at odds ratioen for at kurserne i forårssemesteret 2019 bliver besvaret i kursusevalueringen er lavere end for kurserne i efteråret 2018. De resterende forklarende variable fødselsår og kursustype er også bestemmende men ikke i samme grad.

Ud fra den kumulative logit model med tilfældig effekt, som vi opstillede i analysen, så vi, at det er muligt at bestemme, hvilke forhold der er bestemmende for de studerendes "faglige udbytte" af undervisningen. Til at undersøge dette udvalgte vi kursusspørgsmål 1: "Jeg har fået et stort fagligt udbytte af undervisningen", som responsvariabel. Vi fandt, at de forhold der er bestemmende for de studerendes faglige udbytte af undervisningen er variablene køn, nationalitet, pensum, tidsforbrug og kontortid. Fra variablen køn så vi, at odds ratioen for at de mandlige studerende svarer, at de har fået "højt fagligt udbytte" er lavere end for de kvindelige studerende. Af variablen nationalitet så vi, at odds ratioen for at de studerende med nationalitet "Udenfor Norden" svarer, at de har fået "højt fagligt udbytte" er lavere end for de studerende fra "Norden". Af variablen pensum så vi, at odds ratioen for at de studerende der har angivet pensums sværhedsgrad til "Middel" svarer, at de har fået "højt fagligt udbytte" er højere end for henholdsvis de studerende der har angivet "Under middel" og "Over middel". Af variablen tidsforbrug så vi, at odds ratioen for "højt fagligt udbytte" stiger i takt med, at den studerendes tidsforbrug stiger. Yderligere så vi, at odds ratioen for at en studerende der har anvendt kontortid svarer, at de har fået "højt fagligt udbytte" er højere end de studerende, der ikke har anvendt kontortid.

Alt i alt har vi hermed vist, at det er muligt at opstille statistiske modeller til brug for analyse af kursusevalueringerne for CBS for studieåret 2018/2019. Vi har vist, at det ikke er muligt at forudsige, hvilke studerende der besvarer, og hvilke der ikke besvarer kursusevlaueringen, men vi har fundet nogle forhold, der er bestemmende for hvilke studerende det er. Yderligere har vi vist, at der er bestemmende forhold for de studerendes faglige udbytte af undervisningen.

6.2 Diskussion

I afsnittet ovenfor har vi belyst de resultater, vi er kommet frem til ud fra vores analyse i denne opgave. Disse resultater vil vi gerne dele med CBS for blandt andet at give dem en bedre forståelse, af hvilke effekter der gør sig gældende for, hvilke studerende der besvarer, og hvilke der ikke besvarer deres kurssuevalueringer. Samt give dem indblik i de forhold der er bestemmende for den studerendes faglige udbytte af undervisningen.

Det kan nu diskuteres, hvordan vi kan præsentere vores resultater for CBS. Enten kan vi præsentere det som resultater, der gør sig gældende for hele CBS, eller som resultater der kun gør sig gældende for de udvalgte uddannelser, som analyserne baserer sig på.

Vores resultater er baseret på i alt 4 ud af 39 uddannelser, da omfanget af denne opgave desværre ikke rækker til flere. I analysen af hvilke studerende der besvarer, og hvilke der ikke besvarer deres kursusevalueringer, udfører vi samme analyse på de tre uddannelser; MScCM, BScHA og BScJUR. Årsagen til valget af netop de tre uddannelser er, at det er de tre største uddannelser på CBS for studieåret 2018/2019, målt på antal udsendte kursusevalueringer. Uddannelsen MScCM udgør 20%, uddannelsen BScHA udgør 15% og uddannelsen BScJUR udgør 5% af det samlede antal udsendte kursusevalueringer på CBS. Det betyder, at vores analyse i afsnit 4 er baseret på 40% af vores endelige datasæt. Det vil sige, at analysen er baseret på en forholdsvis stor del af data, men samtidig er det interessant at uddannelsen BScJUR, som er den tredje største uddannelse kun udgør 5% af de udsendte kursusevalueringer. Dermed udgør de resterende 36 uddannelser 60% af vores endelige datasæt. Af de resterende uddannelser anser vi en andel af dem for at være "niche-uddannelser", hvilket vil sige, at de i høj grad adskiller sig fra vores tre udvalgte uddannelser. Et eksempel er uddannelsen Master of Science in Business Administration and Innovation in Health Care, som adskiller sig ved at tage studerende ind fra helt andre bacheloruddannelser end en "typisk CBS uddannelse" såsom bacheloruddannelsen i idræt samt sygeplejeske uddannelsen. Vi har dog erfaret gennem vores studietid på CBS, at det er svært at komme uden om undtagelser. Dog kan der argumenteres for, at det ville være en god

ide også at undersøge et par af "niche-uddannelserne" for at se, om nogle af de bestemmende forhold går igen. Når det så er sagt, er det værd at nævne, at uddannelsen MScCM indeholder i alt 14 forskellige linjer, og dermed bliver en bredere skare af de studerenede naturligt inkluderet i analysen. Derudover indeholder BScHA og MScCM langt størstedelen af udbudte valgfag på CBS, hvilket ligeledes bidrager til den brede skare af studerende, som naturligt inkluderes.

Den fjerde uddannelse der bliver analyseret i opgaven er MScBLC, som anvendes i analysen vedrørende hvilke forhold, der er bestemmende for den studerendes faglige udbytte af undervisningen. Resultaterne er dermed baseret på en lille del af datasættet, nemlig kun 1 ud af de 39 uddannelser som udgør 3,4% af besvarelserne. Argumentet for at vælge netop MScBLC var, at det er en uddannelse med høj svarprocent samt et højt antal besvarelser. Derudover er det en international uddannelse, som gør det muligt at undersøge, om nationalitet er et bestemmende forhold for den studerendes faglige udbytte af undervisningen. Vi har dermed helt bevidst udvalgt en uddannelse, hvor der er mulighed for, at nationalitet kan være bestemmende. Det kan derfor igen diskuteres, hvorvidt resultater fra denne analyse kan præsenteres som resultater, der gør sig gældende for hele CBS, eller som resultater der kun gør sig gældende for den udvalgte uddannelse. For at kunne gennemføre analysen, tog vi også et valg om at slå nogle af niveauerne sammen for de forklarende variable. Dette gjorde vi, fordi det ikke er alle niveauer for alle variable, der er velrepræsenteret i data. Dette gjorde vi eksempelvis for karakter, hvor vi gik fra at have 7 niveauer til kun at have 2 niveauer. Det kan diskuteres, om der ville være lignende behov på de andre uddannelser på CBS, og om vores resultater dermed kan siges at være retvisende for hele CBS, eller kun for MScBLC.

Vi mener derfor, at vi er kommet bedst muligt omkring, og at det må være op til den enkelte at vurdere, om vores resultater kan anvendes til at give et overordnet billede af hele CBS. Det vigtigste er, at man ved anvendelse af vores resultater har kendskab til de til og fravalg, vi har gjort os undervejs samt kender til CBS som uddannelsesinstitution.

6.3 Perspektivering

Vi vil på baggrund af vores resultater anbefale CBS at lave en målerettet strategi for, at få de mandlige studernede samt de studerende fra norden til at besvare deres kursusevalueringer. En sådan strategi kan lede til en højere svarprocent, hvilket giver CBS et større grundlag for deres viden om kvaliteten af undervisningen og dermed kvaliteten af uddannelserne. Større og bedre viden om uddannelsernes kvalitet kan hjælpe CBS med at få optimeret de uddannelser, hvor behovet er.
Højere kvalitet kan tiltrække talentfulde og ambitiøse studerende fra hele verden. Dette kan hjælpe CBS med deres ønske om at levere dimittender med en stærk grundlæggende businessfaglighed og transformative kompetencer til erhvervslivet. Netop dimittender af høj kvalitet og med disse kompetencer, er noget af det aftagerne efterspørger (Copenhagen Business School, 2021). Ved at tilgodse aftagerne kreerer CBS arbejdskraft, der omgående kan komme i arbejde og dermed bidrage til samfundsøkonomien direkte efter endt uddannelse.

En anden anbefaling til CBS er at fortsætte det arbejde, vi har startet, og se nærmere på besvarelserne i de andre kursusspørgsmål. Vi vil også foreslå dem at åbne op for analyser af den del af kursusevalueringen, der berører underviserne, for at belyse kvaliteten af uddannelserne fra denne vinkel.

For os kan hele opgaven og processen omkring den i den grad også bruges. Vi har lært enormt meget af, at skulle strukturere og opbygge en opgave af denne størrelse. Først og fremmest har vi fundet hinandens styrker og hjulpet hinanden til at blive bedre, der hvor vi var svage. Ligeledes har opgaven affødt nogle gode diskussioner omkring det teoretiske indhold. Processen har lært os, hvordan man arbejder med store opgaver, og at et afsnit ikke nødvendigvis kan blive skrevet færdig før de andre, men at specialeskrivning er en levende proces. Slutteligt synes vi, at det har været enormt relevant (endelig), at skulle anvende teori i praksis. Det har ikke været en særlig stor del af vores uddannelse at anvende den opnåede teori, samt nye teorier i praksis, og derfor synes vi, at det er en god måde at afrunde seks års studie.

Vores opgave har ledt til at vi har fået et "højt fagligt udbytte" af vores kandidat, da vi har haft en sjov og lærerig proces med udarbejdelsen af dette speciale.
Vores vejleder Dorte Kronborg har været en uundværlig brik i processen, da hun både har motiveret os, udfordret os og hjulpet os med opbygningen af opgaven. Så tak for det, Dorte.

Tak fordi du læste med

-

Julie & Sofie

Litteraturliste

[1] Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models*, New Jersey, John Wiley &Sons, Inc.

[2] Agresti, A. (2013). *Categorical Data Analysis*, New Jersey, John Wiley &Sons, Inc.

[3] Agresti, A. (2007). *An Introduction to Categorial Data Analysis*, New Jersey, John Wiley &Sons, Inc.

[4] Agresti, A. (2010). *Analysis of Ordinal Categorial Data*, New Jersey, John Wiley &Sons, Inc.

[5] Madsen, H. og P. Thyregod. (2011). *An Introduction to General and Generalized Linear Models*, Florida, CRC Press.

[6] James et al., (2013). *An Introduction to statistical learning*, New York, Springer.

[7] Hastie et al., (2017). *The element of statistical learning*, New York, Springer.

[8] DeLong et al., (1988).*Comparing areas under two or more correlated Receiver Operating Characteristic curves: a nonparametric approach*, Biometrics, 44(3):837-845.

[9] Lovbekendtgørelse nr. 778 af 07-08-209 om universiteter (universitetsloven)

Online kilder

[10] Copenhagen Business School (2020) *CBS årsrapport 2020*.
https://www.cbs.dk/files/cbs.dk/cbs_-_aarsrapport-2020.pdf.
Sidst besøgt d. 6. maj, 2021.

[11] Copenhagen Business School (2017) *CBS årsrapport 2017*.
https://services-webdav.cbs.dk/doc/CBS.dk/Om%20CBS/cbs-aarsrapport-2017-web.pdf.
Sidst besøgt d. 6. maj, 2021.

[12] Copenhagen Business School (2013) *CBS årsrapport 2013*.
https://www.cbs.dk/files/cbs.dk/arsrapport_2013_1.pdf
Sidst besøgt d. 6. maj, 2021.

[13] Uddannelses- og Forskningsministeriet (2020). *ECTS*.
https://ufm.dk/uddannelse/anerkendelse-og-dokumentation/dokumentation/ects.
Sidst besøgt d. 4. maj, 2021.

[14] Fagerland, M. (2017). *How to test for goodness of fit in ordinal logistic regression models*. The Stata Journal, 17(3), side 668–686.
https://journals.sagepub.com/doi/pdf/10.1177/1536867X1701700308.
Sidst besøgt d. 10. maj, 2021

[15] Christensen, RHB (2019) *A Tutorial on fitting Cumulative Link Mixed Models with clmm2 from the ordinal Package*.The Comprehensive R Archive Network.
https://cran.r-project.org/web/packages/ordinal/vignettes/clmm2_tutorial.pdf.
Sidst besøgt d. 12. maj, 2021

[16] Christensen, RHB. (2021) *Cumulative Link Models for Ordinal Regression with the R Package ordinal*. The Comprehensive R Archive Network.
https://cran.r-project.org/web/packages/ordinal/vignettes/clm_article.pdf.
Sidst besøgt d. 14. maj, 2021

[17] Christensen, RHB. (2015) *Analysis of ordinal data with cumulative link models — estimation with the R-package ordinal.* https://cran.microsoft.com/snapshot/2015-07-02/web/packages/ordinal/vignettes/clm_intro.pdf.
Sidst besøgt d. 14. maj, 2021

[18] Jay, M (2019) *logitgof: Hosmer-Lemeshow Tests for Logistic Regression Models* https://www.rdocumentation.org/packages/generalhoslem/versions/1.3.4/topics/logitgof
Sidst besøgt d. 28. april, 2021

[19] Wulff, C (2009) *Likert-skala.* Den Store Danske https://denstoredanske.lex.dk/Likert-skala
Sidst besøgt d. 14. maj, 2021

[20] NCSS Statistical Software (2021) *One ROC curve and Cutoff Analysis.* https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/-One_ROC_Curve_and_Cutoff_Analysis.pdf
Sidst besøgt d. 10. april, 2021

[21] Copenhagen Business School (2021) *Håndbog om uddannelseskvalitet.* cbs.dk https://www.cbs.dk/files/cbs.dk/call_to_action/programme_quality_handbook_da.pdf
Sidst besøgt d. 15. maj, 2021

7 Bilag

A Opgørelse af CBS's bestand

	Opgørelse af CBS's bestand								
Årstal	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**
Bachelor	9.094	7.333	7.423	7.442	7.172	7.167	7.272	7.606	7.940
Kandidat	8.813	9.326	10.324	10.071	8.396	7.744	7.582	7.774	7.658
Samlet	**17.907**	**16.659**	**17.747**	**17.513**	**15.568**	**14.911***	**14.854**	**15.380**	**15.598**

*Studenterbestand faldet pga fremdriftsreformen forventet stabiliseret omkring 2018-2019

Tal for 2012 er taget fra CBS Årsrapport 2013

Tallene for 2013-2015 er taget fra CBS årsrapport 2017

Tallene for 2016-2020 er taget fra CBS's årsrapport 2020

B Quasibinomial startmodel output for BScHA

```
> summary(BScHA1.1 <- glm(RESP_STATUS ~ S_KON + SS_NATIONALITET + S_FODSELSAR + KARAKTER + KURSUS_TYPE + EVAL_SEMESTER, family = quasibinomial,data=Data2))

Call:
glm(formula = RESP_STATUS ~ S_KON + SS_NATIONALITET + S_FODSELSAR +
    KARAKTER + KURSUS_TYPE + EVAL_SEMESTER, family = quasibinomial,
    data = Data2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2228  -0.7922  -0.7038   1.1870   1.8895

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 -0.866003   0.099793  -8.678  < 2e-16 ***
S_KONM                      -0.327877   0.053448  -6.134 8.95e-10 ***
SS_NATIONALITETUdenfor Norden  0.647411   0.081953   7.900 3.16e-15 ***
S_FODSELSAR1996-2018        -0.006574   0.064642  -0.102  0.91900
KARAKTER0                    0.068634   0.096223   0.713  0.47570
KARAKTER2                    0.122237   0.096372   1.268  0.20470
KARAKTER4                    0.012158   0.097017   0.125  0.90027
KARAKTER7                   -0.068961   0.097518  -0.707  0.47949
KARAKTER10                   0.056022   0.096852   0.578  0.56299
KARAKTER12                   0.103800   0.096233   1.079  0.28078
KURSUS_TYPEV                 0.202512   0.066638   3.039  0.00238 **
EVAL_SEMESTERF19            -0.331944   0.053277  -6.231 4.88e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasibinomial family taken to be 1.004403)

    Null deviance: 9327.0  on 8026  degrees of freedom
Residual deviance: 9055.6  on 8015  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
```

C Quasibinomial startmodel output for BScJUR

```
> summary(BScJUR1.1 <- glm(RESP_STATUS ~ S_KON + SS_NATIONALITET + S_FODSELSAR + KARAKTER + KURSUS_TYPE + EVAL_SEMESTER, family = quasibinomial,data=Data3))

Call:
glm(formula = RESP_STATUS ~ S_KON + SS_NATIONALITET + S_FODSELSAR +
    KARAKTER + KURSUS_TYPE + EVAL_SEMESTER, family = quasibinomial,
    data = Data3)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9415  -0.7365  -0.6105  -0.4985   2.1619

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                  -1.09536    0.15730  -6.963 4.10e-12 ***
S_KONM                       -0.61511    0.09622  -6.393 1.89e-10 ***
SS_NATIONALITETUdenfor Norden  0.25081    0.30828   0.814  0.41596
S_FODSELSAR1996-2018           0.31702    0.10751   2.949  0.00322 **
KARAKTER0                    -0.22683    0.17359  -1.307  0.19142
KARAKTER2                    -0.33146    0.17576  -1.886  0.05942 .
KARAKTER4                    -0.11313    0.17056  -0.663  0.50720
KARAKTER7                    -0.14440    0.16970  -0.851  0.39489
KARAKTER10                   -0.36925    0.17634  -2.094  0.03635 *
KARAKTER12                   -0.13545    0.17073  -0.793  0.42761
KURSUS_TYPEV                 -0.29795    0.21774  -1.368  0.17131
EVAL_SEMESTERF19             -0.05646    0.09657  -0.585  0.55883
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasibinomial family taken to be 1.008332)

    Null deviance: 2918.9  on 2887  degrees of freedom
Residual deviance: 2855.8  on 2876  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
```

D Kumulativ logit model output for startmodel for CCDCO2002U_E18

```
formula: KURSUS_S1 ~ S_KON + SS_NATIONALITET + S_FODSELSAR + KKARAKTER + PPESNSUM_S1 + TIDSFORBRUG_S1 + FREMMODE_S1 + KONTORTID_S1
data:    ClmData1

 link  threshold nobs logLik AIC   niter max.grad cond.H
 logit flexible  31   -16.25 60.49 6(0)  3.91e-08 3.3e+02

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
S_KONM                        -2.0302     1.3892  -1.461   0.1439
SS_NATIONALITETUdenfor Norden -4.8684     2.0390  -2.388   0.0170 *
S_FODSELSAR1996-2018          -3.9465     2.8613  -1.379   0.1678
KKARAKTEROver 7                1.6781     1.7084   0.982   0.3260
PPESNSUM_S1Over middel        -4.5460     1.9956  -2.278   0.0227 *
TIDSFORBRUG_S12               -2.6653     1.8103  -1.472   0.1410
TIDSFORBRUG_S13               -0.5564     1.8031  -0.309   0.7577
TIDSFORBRUG_S14               -0.8547     1.9397  -0.441   0.6595
TIDSFORBRUG_S15                1.8015     2.0803   0.866   0.3865
FREMMODE_S14                  -0.7840     1.6718  -0.469   0.6391
FREMMODE_S15                  -0.6043     1.8818  -0.321   0.7481
KONTORTID_S1Ja                -0.2076     1.4689  -0.141   0.8876
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Threshold coefficients:
    Estimate Std. Error z value
3|4   -8.622      3.031  -2.844
4|5   -4.870      2.590  -1.880
```

E Kumulativ logit model med tilfældig effekt output for startmodel for MScBLC

```
Cumulative Link Mixed Model fitted with the adaptive Gauss-Hermite
quadrature approximation with 10 quadrature points

Call:
clmm2(location = KURSUS_S1 ~ S_KON + SS_NATIONALITET + KKARAKTER +
    S_FODSELSAR + PPESNSUM_S1 + TIDSFORBRUG_S1 + FREMMODE_S1 +
    KONTORTID_S1, random = KURSUS_KODE_EVAL_SEMESTER, data = clmmData1,
    Hess = TRUE, nAGQ = 10)

Random effects:
                               Var   Std.Dev
KURSUS_KODE_EVAL_SEMESTER 0.3682162 0.6068082

Location coefficients:
                            Estimate Std. Error z value Pr(>|z|)
S_KONM                      -0.4579   0.1962    -2.3342 0.01958400
SS_NATIONALITETUdenfor Norden -0.5960   0.1778    -3.3518 0.00080293
KKARAKTEROver 7              0.1820   0.1860     0.9786 0.32778662
S_FODSELSAR1996-2018        -0.2263   0.3110    -0.7277 0.46677115
PPESNSUM_S1Over middel      -0.3680   0.2382    -1.5450 0.12234699
PPESNSUM_S1Under middel     -1.6494   0.4230    -3.8995 9.6399e-05
TIDSFORBRUG_S12             -0.0034   0.2717    -0.0125 0.99000652
TIDSFORBRUG_S13              0.5549   0.3030     1.8312 0.06707216
TIDSFORBRUG_S14              0.7830   0.3158     2.4796 0.01315319
TIDSFORBRUG_S15              0.5679   0.3400     1.6705 0.09481263
FREMMODE_S12                -0.3627   0.6571    -0.5520 0.58093793
FREMMODE_S13                 0.1317   0.5431     0.2426 0.80834141
FREMMODE_S14                 0.2078   0.5204     0.3993 0.68968988
FREMMODE_S15                 0.4764   0.5287     0.9012 0.36748933
KONTORTID_S1Ja               1.3438   0.3598     3.7346 0.00018799

No scale coefficients

Threshold coefficients:
    Estimate Std. Error z value
2|3 -2.9997    0.5712    -5.2511
3|4 -1.8954    0.5531    -3.4269
4|5  0.5159    0.5435     0.9493

log-likelihood: -552.6303
AIC: 1143.261
Condition number of Hessian: 225.6311
(1 observation deleted due to missingness)
```

F R-kode

F.1 Kode vedrørende den logistisk regressionsmodel

Nedenfor fremgår den kode der er kørt i R, for at opstille modellerne for MScCM, BScHA og BScJUR i afsnit 4.

På følgende måde opstilles alle startmodellerne for de logistiske regressionsmodeller. Her er koden med uddannelsen MScCM som eksempel:

$$summary(MScCM1 < -glm(RESP_STATUS\ S_KON + SS_NATIONALITET + S_FODSELSAR + KARAKTER + KURSUS_TYPE + EVAL_SEMESTER, family = binomial, data = Data1))$$

På følgende måde køres startmodellen også som en qusibinomial model:

$$summary(MScCM1 < -glm(RESP_STATUS\ S_KON + SS_NATIONALITET + S_FODSELSAR + KARAKTER + KURSUS_TYPE + EVAL_SEMESTER, family = quasibinomial, data = Data1))$$

Til at udføre en Wald-test for hvert enkelt forklarende variable i startmodellen anvendes følgende kode. I koden gøres det for uddannelsen MScCM og der laves en Wald-test af variablen køn som eksempel:

$$wald.test(b = coef(MScCM1), Sigma = vcov(MScCM1), Terms = 2:2)$$

Til løbende at lave likelihood ratio test for at undersøge, om de insignitikante variable kan udelades fra modellen, anvendes følgende kode. Her i koden gøres det for MScCM hvor startmodellen (MScCM1) og en reduceret model (MScCM2) testes for MScCM:

$$anova(MScCM2, MScCM1, test = "LRT")$$

Til at beregne odds ratioer og tilhørende konfidensintervaller for slutmodellerne anvendes følgende kode. Nedenfor er det gjort for slutmodellen (MScCM3) for MScCM:

$$exp(cbind(OR = coef(MScCM3), confint(MScCM3)))$$

Prædiktion

Til lave prædiktion anvendes følgende kode. Først udvælges en slutmodel, her er det slutmodellen for BScJUR (BScJUR5), hvor der prædikteres på datasættet der indeholder data for uddannelsen BScHA (Data2):

$$pred = predict(BScJUR5, Data2, type = "response")$$

$$pred = prediction(pred, Data2\$RESP_STATUS)$$

Med følgende kode er der efterfølgende plottet en ROC-kurve af ovenstående prædiktion:

$$roc = performance(pred, "tpr", "fpr")$$

$$plot(roc, colorize = T, lwd = 2)$$

$$abline(a = 0, b = 1)$$

Med følgende kode er der beregnet en tilhørende AUC-værdi for prædiktionen:

$$auc = performance(pred, measure = "auc")$$

$$print(auc@y.values)$$

F.2 Kode vedrørende den kumulative logit model

Til at opstille en startmodel for kurset CCDCO2002U_E18 er følgende kode anvendt i R:

$$summary(BLC1 < -clm(KURSUS_S1\ S_KON + SS_NATIONALITET +$$
$$S_FODSELSAR + KKARAKTER + PPESNSUM_S1 + TIDSFORBRUG_S1$$
$$+FREMMODE_S1 + KONTORTID_S1,$$
$$data = ClmData1, Hess = TRUE, link = "logit"))$$

Til at lave en goodness-of-fit test (Hosmer og Lemeshow test) for slutmodellen for kurset CCDCO2002U_E18, anvendes følgende kode:

$$predprob < -data.frame(SS_NATIONALITET = ClmData1\$SS_NATIONALITET,$$
$$PPESNSUM_S1 = ClmData1\$PPESNSUM_S1)$$

$$fv < -predict(BLC7, newdata = predprob, type = "prob")\$fit$$

$$logitgof(ClmData1\$KURSUS_S1, fv, g = 5, ord = TRUE)$$

Principal komponent analyse

Til at udføre en principal komponent analyse af de 5 kursusspørgsmål er følgende kode brugt:

$$pca.Sample.2 <- prcomp(PCA.2, center = TRUE, scale = TRUE)$$

$$var_explained = pca.Sample.2\$sdev^2 / sum(pca.Sample.2\$sdev^2)$$

Til at plotte et screeplot af principal komponent analysens 5 principalkomponenter anvendes følgende kode:

$$qplot(c(1:5), var_explained) + geom_line() +$$
$$xlab("PrincipalComponent") +$$
$$ylab("VarianceExplained") +$$
$$ggtitle("ScreePlot") + ylim(0, 1)$$

Generel kode for kumulativ logit model med tilfældig effekt

Til at opstille startmodellen for MScBLC er følgende kode anvendt:

$$summary(clmm21 <- clmm2(KURSUS_S1\ S_KON + SS_NATIONALITET$$
$$+ KKARAKTER + S_FODSELSAR + PPESNSUM_S1 + TIDSFORBRUG_S1$$
$$+ FREMMODE_S1 + KONTORTID_S1,$$
$$random = KURSUS_KODE_EVAL_SEMESTER,$$
$$data = clmmData1, Hess = TRUE, nAGQ = 10))$$

Til at lave løbende likelihood ratio test, for at teste om de insignifikante variable kan udelades fra modellen, anvendes følgende kode. Her gøres det for startmodellen (clmm21) og en reduceret model (clmm22):

$$anova(clmm22, clmm21)$$

For at bestemme profil likelihood konfidensintervallet for varianskomponenten anvendes følgende kode:

$$pr2 <- profile(clmm24, range = c(.1, 2), nSteps = 30, trace = 0)$$
$$confint(pr2)$$

Til at bestemme det "gennemsnitlige kursus" og efterfølgende at plotte dem anvendes følgende kode for slutmodellen (clmm24):

$$ci <- clmm24\$ranef + qnorm(0.975) * sqrt(clmm24\$condVar) \%o\% c(-1, 1)$$
$$ord.re <- order(clmm24\$ranef)$$
$$ci <- ci[order(clmm24\$ranef),]$$
$$plot(1 : 26, clmm24\$ranef[ord.re], axes = FALSE, ylim = range(ci), xlab = "Kursus",$$
$$ylab = "Kursuseffekt")$$
$$axis(1, at = 1 : 26, labels = ord.re)$$
$$axis(2)$$
$$for(iin1 : 26)segments(i, ci[i, 1], i, ci[i, 2])$$
$$abline(h = 0, lty = 2)$$