Kapitel 3. Logistisk regression med binært udfald

Erik Gahner Larsen
egl@sam.sdu.dk

Institut for Statskundskab
Syddansk Universitet

3.1. Indledning

Talrige fænomener inden for samfundsvidenskaberne inddeles i binære kategorier. Som eksempler kan nævnes om en vælger stemte ved det seneste folketingsvalg eller ej, om en politiker stemte for eller imod en lovgivning, om en person er arbejdsløs eller ej, om et land er ramt af borgerkrig eller ej og så videre. I tilfælde hvor man har en binær afhængig variabel, vil en ordinær lineær regression (OLS), jf. Kristensen og Hussain (2016, kapitel 17 og 18) og denne bogs kapitel 2, ikke være den bedste løsning, og en logistisk regressionsmodel vil være et bedre analysevalg.

Nærværende kapitel vil give en anvendelsesorienteret introduktion til den logistiske regressionsmodel. Kapitlet fordrer forudgående kendskab til multivariat OLS-regressioner (jf. kapitel 2), hvormed der også vil være et eksplicit fokus på, hvordan logistiske regressioner afviger fra dette, og hvilke faldgruber, man skal være opmærksom på. I springet fra OLS til den logistiske regression, bevæger vi os ind i en verden, hvor man bliver præsenteret for mange af de samme informationer. Der er stadig regressionskoefficienter, standardfejl, p-værdier og andre informationer, man kender, men det der foregår under kølerhjelmen af den statistiske model, altså estimationsmetoden, er forskellig, og tolkningen af, om en uafhængig variabel påvirker den afhængige variabel, er mere kompliceret end i en OLS regression.

En væsentlig pointe, som vil blive fremhævet i nedenstående, er, at man i det omfang det er muligt, bør visualisere resultaterne fra den logistiske regression (Kastellec og Leoni 2007). Dette af to grunde. For det første er det svært at lave meningsfyldte tolkninger på baggrund af koefficienterne i det output, man får fra den logistiske regression. Det er med andre ord ikke lige så intuitivt som i en OLS regressionsanalyse bl.a. fordi koefficienterne nu ikke længere er målt i samme enheder som responsvariablen i udgangspunktet (sandsynligheder). For det andet muliggør visualiseringen, at

man ikke alene fremhæver konfidensintervallerne for resultaterne, men også fordelingen af de uafhængige variable, man ønsker at udtale sig om.

Der er overordnet tre trin, man bør gennemføre, for at få mest muligt ud af en logistisk regressionsanalyse:

- For det første skal man gennemføre *en specifikation og estimering af modellen*. Dette trin er i praksis at sammenligne med det man gør, når man gennemfører en OLS regression, hvor man specificerer den afhængige variabel, de uafhængige variable og eventuelle ekstra muligheder.
- For det andet skal man bruge resultaterne fra analysen til at kalkulere såkaldt forudsagte sandsynligheder, dvs. sandsynligheder som baseres på beregninger fra den estimerede model. Man skal her udregne marginale effekter af de uafhængige variable.
- For det tredje skal man visualisere disse sandsynligheder. Dette trin er især vigtigt, hvis der er mange forudsagte sandsynligheder, eksempelvis hvis man ønsker at udregne forudsagte sandsynligheder for en uafhængig variabel med mange kategorier.

De tre trin beskæftiger sig med forskellige procedurer, og de valg der træffes på det første trin, har implikationer for det andet og tredje. Tabel 3.1 opsummerer de tre trin, og giver desuden information om, hvilken kommando man kan bruge i Stata, for hvert af de tre trin (se mere om dette i afsnit 3.5).

Tabel 3.1: De tre trin i en logistisk regressionsanalyse

Trin	Procedure	Beskrivelse	Fokus	Stata-kommando
1	Estimering	Tolkning af output fra den logistiske regressionsanalyse	Funktionel form Forudsætninger	. logit
2	Kalkulation	Sandsynlighed for at den afhængige variabel er 1 givet bestemte værdier af den uafhængige variabel	Udregning af forudsagte sandsynligheder	. margins
3	Visualisering	Visualisering af forudsagte sandsynligheder	Usikkerhed Fordeling	. marginsplot

I det følgende vil jeg starte ud med at introducere den lineære sandsynlighedsmodel, der estimeres med den meget udbredte OLS-regressionsmetode. Herefter introduceres den binære logistiske regression med fokus på, hvordan den afviger fra OLS regressioner. Med udgangspunkt i dette gives der et empirisk eksempel, hvor partidentifikation (afhængig variabel) modelleres som en funktion af en række uafhængige variable. De tre forskellige trin anvendes på det empiriske eksempel, og det illustreres, hvordan man implementerer og fortolker disse trin i Stata. Afslutningsvis gives der en konklusion med en tjekliste til, hvad man bør huske, når man gennemfører og rapporterer sin logistiske regressionsanalyse.

I dette fokus på den logistiske regression ligger der også et fravalg af probitregressioner, der også anvendes ved binære afhængige variable. Substantielt set er resultaterne oftest ret ens, og de væsentlige pointer, der formidles i nærværende kapitel i forhold til gennemførelsen af logistiske regressioner, lader sig også anvende på probitregressioner. I Stata, eksempelvis, kan probit-kommandoen bruges på akkurat samme måde som logit-kommandoen.

3.2. Den lineære sandsynlighedsmodel

Det første man skal have styr på er, hvorfor der skal bruges en logistisk regression i stedet for en velkendt ordinær lineær regression. I en OLS-regression modelleres sandsynligheden for, at den afhængige variabel har værdien 1 som en lineær funktion af x:

$$\Pr(y = 1|x) = x\beta$$

hvor β er en vektor af koefficienter og x en vektor af uafhængige variable, eventuelt indeholdende et konstantled. I praksis er der intet, der forhindrer estimering af en lineær sandsynlighedsmodel a la ovenstående. Den primære forskel på en OLS regression og en logistisk regression er, at OLS beskæftiger sig med en kontinuerlig afhængig variabel men en logistisk regression beskæftiger sig med en binær afhængig variabel. Dette ændrer dog ikke på, at man eventuelt kan anvende en OLS regression, når man har en binær afhængig variabel, og i mange tilfælde vil der fås stort set identiske resultater (Hellevik 2009) om end der skal foretages nødvendige tiltag for at tage højde for manglende opfyldelse af forudsætningerne for OLS. Men som tidligere sagt er logistisk regression det optimale valg.

Når man har en afhængig variabel, der antager værdien 0 eller 1, fås en regressionskoefficient, β, der tolkes som ændringen i sandsynligheden for at y=1, når den uafhængige variabel, x, ændres med én enhed. Fordelen i den lineære sandsynlighedsmodel (altså i dette tilfælde OLS regression på en binær afhængig variabel) er, at den er intuitiv og nem at tolke.

Der er dog tre grundlæggende problemer forbundet med at bruge OLS-regressioner på binære afhængige variable. For det første har man med OLS-regressionen risikoen for at få forudsagte sandsynligheder mindre end 0 eller større end 1. Det giver som bekendt ingen mening at tale om en sandsynlighed for, at en afhængig variabel har værdier over 1, altså større end 100%. For det andet fås ikke-normalfordelte fejlled og heteroskedasticitet, der fører til forkerte standardfejl for estimerede parametre. For det tredje den funktionelle form, hvor det antages, at effekten af *x* (dvs. koefficienten β) er konstant, hvilket som oftest vil være forkert.

3.3. Den logistiske sandsynlighedsmodel

I den logistiske model tager man, i modsætning til den lineære model, højde for, at der arbejdes med sandsynligheder, der per definition ikke kan være mindre end 0 eller større end 1. Dette gøres ved at transformere sandsynligheder til odds, altså sandsynligheden for at den afhængige variabel har værdien 1 divideret med sandsynligheden for at den afhængige variabel har værdien 0:

$$odds = \frac{\Pr(y = 1|x)}{1 - \Pr(y = 1|x)} = \frac{\Pr(y = 1|x)}{\Pr(y = 0|x)}$$

Disse odds giver den relative sandsynlighed. Hvis odds er lig 1, betyder det, at sandsynligheden for at $y = 1$ er den samme som sandsynligheden for at $y = 0$, altså 50/50. Hvis odds er større end 1, er sandsynligheden for at $y = 1$ større end sandsynligheden for at $y = 0$ og omvendt, hvis odds er mindre end 1, er sandsynligheden for at $y = 0$ større end sandsynligheden for at $y = 1$.

Den logistiske model er så givet ved en logistisk transformation (kaldet *logit*), som giver log af odds. Det er den naturlige logaritme af ovenstående. Det er dette (logit=log odds), man vil få ud som koefficienter, når man estimerer den logistiske regressionsanalyse:

$$\ln\left(\frac{\Pr(y = 1|x)}{1 - \Pr(y = 1|x)}\right) = log_e(odds) = logit = x\beta$$

Med udgangspunkt i dette kan man isolere $\Pr(y = 1|x)$, altså sandsynligheden for at den afhængige variabel antager værdien 1 givet den uafhængige variabel, ved at bruge den inverse logit-funktion. Denne giver forudsagte sandsynligheder, der aldrig går under 0 eller overskrider 1 (hvor exp er eksponentialfunktionen):

$$\Pr(y = 1|x) = logit^{-1}(x\beta) = \frac{\exp(x\beta)}{1 + \exp(x\beta)}$$

Med denne funktion løses udfordringen omkring definitionsområdet for sandsynligheder, som man har i den lineære sandsynlighedsmodel. Til gengæld fås en model, der ikke kan tolkes lige så intuitivt. Det er af denne grund, at der som regel skal tages yderligere trin, efter estimering af regressionen i Stata og andre programmer, for at man kan tolke på resultaterne. Modsat i en OLS-regression er effekten af x ikke konstant i en logit-regression. Den varierer derimod alt efter, hvor på den uafhængige variabel, vi ændrer med én enhed. Effekten vil således til- eller aftage, hvorfor den kan være større eller mindre ved nogle værdier af x end ved andre.

Tilføj dertil, at effektstørrelsen kan ændre sig alt efter, hvilke værdier der er på de andre uafhængige variable, når man inkluderer sådanne. Implikationen af dette er, at effekten af en uafhængig variabel kan være sensitiv i forhold til, hvilke værdier der er for andre variable. Derfor ser man ofte, når effekten af en variabel skal tolkes, at man eksplicit forholder sig til værdierne på de andre variable, og ikke laver de samme alt andet lige-tolkninger (*ceteris paribus*), som i en OLS regression.

Dette leder os frem til, at modsat en OLS-regression, der anvender mindste kvadraters metode, som er forklaret i kapitel 2, anvender logistisk regression en *maximum likelihood*-estimationsmetode. Med maximum likelihood-estimeringen søges de parametre (givet antaget sandsynlighedsfordeling, her logistisk), gennem *iterationer* (systematiske gæt), der passer bedst til vores data - altså de parametre, der maksimerer sandsynligheden (likelihood) for at observere de data, vi har. I modsætning til OLS-estimaterne er maximum likelihood-estimaterne ikke udregnet med henblik på at minimere summen af de kvadrerede fejl. Dette har den implikation, at man bygger på nogle andre antagelser og skal rapportere andre modelparametre end i en OLS-regression.

I en logistisk regression fås derfor heller ikke en R^2-værdi, men der er dog mulighed for at få såkaldte pseudo R^2-værdier. Pseudo fordi de forsøger at opnå det samme som R^2, hvor værdierne ligger mellem 0 og 1. Der er ikke ét overordnet goodness-of-fit mål for en logistisk regression

(Hagle og Mitchell 1992), og i rapporteringen af den logistiske regressionsanalyse ser man ofte log-likelihood for modellen rapporteret såvel som en pseudo R^2-værdi - i Stata rapporteres fx McFaddens R^2 som standard, men der findes forskellige beregninger af pseudo R^2.

For at opsummere er der tre pointer, der er vigtige at huske på:

- Man går fra en lineær til en ikke-lineær effekt af den uafhængige variabel.
- Man kan ikke lave de samme umiddelbare tolkninger af koefficienterne.
- Man skal eksplicit forholde sig til andre modelparametre.

Overordnet ændrer dette dog ikke på, at man substantielt set er interesseret i det samme som i en OLS, nemlig at specificere en funktion, der kan estimere effekten af x på y. Tabel 3.2 opsummerer de gennemgåede ligheder og forskelle mellem OLS- og logitestimation.

Tabel 3.2: Udvalgte ligheder og forskelle mellem OLS og logistisk regression

	OLS	Logistisk
Afhængig variabel (y)	Intervalskaleret	Binær (0 eller 1)
Udfaldsrum for y	Fra -∞ til +∞	Fra 0 til 1
Uafhængige variable (x)	Alle typer af variable	Alle typer af variable
Funktionel form, effekt af x	Konstant	Varierer
Koefficienter for x	Intuitiv tolkning	Mindre intuitiv tolkning
Determinationskoefficient	R^2	Pseudo R^2
Funktion	Lineær	Ikke-lineær
Estimation	Mindste kvadraters metode	Maximum likelihood

3.4. Logistisk regression i praksis: Tillid til Folketinget og partitilknytning

For at illustrere, hvordan en logistisk regressionsanalyse gennemføres, anvendes spørgeskemadata fra European Social Survey (ESS). Nærmere bestemt bruges den danske del af ESS fra 2014 (European Social Survey, Round 7, Data 2014). Som den afhængige variabel bruger jeg partitilknytning, hvor respondenterne har svaret ja eller nej til, om de føler sig mere knyttet til et parti end andre partier. Denne variabel hedder ”parti” i vores datasæt.

Som uafhængig variabel bruger jeg en variabel med 11 værdier, hvor respondenterne har udtrykt deres tillid til Folketinget, der går fra 0 (ingen tillid til Folketinget) til 10 (høj tillid til Folketinget). Denne variabel hedder ”folketing” i vores datasæt. Foruden dette har jeg fire kontrolvariable, mere specifikt køn (”kvinde”), alder (”alder”), indkomst (”indkomst”) og uddannelse (”uddannelse”). Tabel 3.3 viser deskriptiv statistik for alle variable

Tabel 3.3: Deskriptiv statistik, Danmark, ESS

	Gennemsnit	Standardafvigelse	Minimum	Maksimum
Partitilknytning	0,7	0,46	0	1
Tillid (Folketinget)	5,9	2,44	0	10
Køn (kvinde)	0,5	0,50	0	1
Alder	48,1	18,94	15	100
Indkomst	5,8	2,91	1	10
Uddannelse	4,1	1,88	1	7

I tabel 3.4 estimeres to regressioner, hvor der fokuseres på konstanten (α) og koefficienten for tillid (β). Den første kolonne viser den lineære sandsynlighedsmodel (OLS), og den anden viser den logistiske regression (logit).

Tabel 3.4: Effekten af tillid til Folketinget på partitilknytning, OLS og logit

Partitilknytning	OLS	Logit
Tillid (Folketing)	0,025	0,116
Konstant	0,553	0,173

I OLS-regressionen er konstanten 0,553, hvilket betyder, at sandsynligheden for at føle sig knyttet til et parti er 55,3 procent for en person med ingen tillid til Folketinget (altså når tillid er 0, hvorfor 0,025 ikke indregnes). For enhver positiv ændring i tillid med én enhed, stiger sandsynligheden for at føle sig knyttet til et parti med 0,025, altså 2,5 procentpoint. Hvis man eksempelvis har maksimal tillid til Folketinget, er sandsynligheden for, at man føler sig knyttet til et parti lig med 80,1 procent ($\alpha + 10 \cdot \beta$).

I den logistiske regression (logit-kolonnen) kan man ikke på samme måde tolke konstanten og koefficienten for tillid. Man skal, for at få resultater, der kan sammenlignes med resultaterne fra OLS-regressionen, kalkulere de forudsagte sandsynligheder. Her vil jeg først give et eksempel på udregning af den forudsagte sandsynlighed, når tilliden til Folketinget er 0:

$$\Pr(Partitilknytning = 1|Tillid = 0) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} = \frac{e^{0{,}173+0{,}116\cdot 0}}{1+e^{0{,}173+0{,}116\cdot 0}} = 0{,}543$$

Den forudsagte sandsynlighed fra den logistiske model afviger ikke betydeligt fra resultatet fra den lineære regression. Jeg kan tilsvarende udregne de forudsagte sandsynligheder for alle andre værdier af tillid til Folketinget, eksempelvis også når tilliden til Folketinget er 5:

$$\Pr(Partitilknytning = 1|Tillid = 5) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} = \frac{e^{0{,}173+0{,}116\cdot 5}}{1+e^{0{,}173+0{,}116\cdot 5}} = 0{,}680$$

I tabel 3.5 er alle de forudsagte sandsynligheder vist for både OLS- og logit-regressionen. Desuden er forskellen i de forudsagte sandsynligheder, når den uafhængige variabel stiger med én enhed, vist. Som det kan ses, er effekten i OLS regressionen konstant (0,025), hvor effekten i den logistiske regression varierer, alt efter hvilket niveau for tillid der fokuseres på. OLS-genererede sandsynligheder fungerer her altså nogenlunde fint, når tilliden er omtrent gennemsnitlig, men afviger fra det korrekte (logit) når man bevæger sig ud i yderværdierne af tilliden.

Tabel 3.5: Forudsagte sandsynligheder, OLS og logit

	OLS		logit	
Tillid (x)	$\Pr(y=1)$	$\Delta\Pr(y=1)$	$\Pr(y=1)$	$\Delta\Pr(y=1)$
0	0,553	-	0,543	-
1	0,578	0,025	0,572	0,028
2	0,603	0,025	0,600	0,028
3	0,628	0,025	0,628	0,028
4	0,652	0,025	0,654	0,027
5	0,677	0,025	0,680	0,026
6	0,702	0,025	0,705	0,025
7	0,727	0,025	0,729	0,024
8	0,751	0,025	0,751	0,022

9	0,776	0,025	0,772	0,021
10	0,801	0,025	0,792	0,020

Note: Differensmålene er kalkuleret med flere decimaler end de angivne.

3.5. Logistisk regression i Stata

Til trods for, at forudsigelserne i tabel 3.5 kan kalkuleres uden de store vanskeligheder, er det betydeligt nemmere at få et statistikprogram til at gøre det (og mindre sandsynligt, at man laver fejl).

Her vil jeg tage afsæt i, hvordan det kan estimeres i Stata, mens det i afsnit 3.6 er også kort nævnes, hvordan det kan gøres i SPSS. Estimeringen består af tre trin, der gennemføres med tre kommandoer i Stata, hvor man:

- først estimerer regressionen
- dernæst udregner de forudsagte sandsynligheder
- og til sidst visualiserer dem.

I det følgende beskrives de forskellige faser nærmere.

Trin 1: Estimering af logistisk regression

Det første trin er estimeringen af logit-modellen. Her vælger man, hvilke variable, der skal være uafhængige variable og specificerer den funktionelle form, lige som det kendes fra OLS regressioner. Det vil sige, at interaktionsled m.v. også specificeres her. På samme måde som man bruger kommandoen *regression* til at gennemføre en OLS-regression i Stata, bruger man blot kommandoen *logit*, når man skal gennemføre en logistisk regression. Det er desuden muligt at gennemføre den logistiske regression med kommandoen *logistic*. Denne kommando rapporterer odds ratios, hvilket også kan fås ved at tilføje *or* som en mulighed til *logit*, eksempelvis *logit y x, or*.

Hvis man i Stata bruger kommandoen *logit*, kan man angive den afhængige variabel samt rækken af uafhængige variable, fx:

```
. logit parti folketing kvinde alder indkomst uddannelse
```

Dette giver det output, der er vist i tabel 3.6. Øverst i tabellen ses det, at man gennemfører en maksimum likelihood-estimering, hvor en iterationslog med log-likelihood værdier fås. Hver iteration forsøger at maksimere log-likelihood funktionen.

Tabel 3.6: Output fra logistisk regression i Stata

```
Iteration 0:   log likelihood = -784.28215
Iteration 1:   log likelihood = -765.16804
Iteration 2:   log likelihood = -765.04431
Iteration 3:   log likelihood =  -765.0443

Logistic regression                                     Number of obs   =       1307
                                                        LR chi2(5)      =      38.48
                                                        Prob > chi2     =     0.0000
Log likelihood =  -765.0443                             Pseudo R2       =     0.0245

------------------------------------------------------------------------------
       parti |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   folketing |   .1335212   .0265926     5.02   0.000     .0814007    .1856417
      kvinde |  -.0131942   .1257239    -0.10   0.916    -.2596086    .2332201
       alder |    .012563   .0035406     3.55   0.000     .0056236    .0195025
    indkomst |   .0011582   .0230719     0.05   0.960    -.0440618    .0463782
  uddannelse |  -.0217914   .0380212    -0.57   0.567    -.0963117    .0527289
       _cons |   -.383571   .2791773    -1.37   0.169    -.9307485    .1636066
------------------------------------------------------------------------------
```

Øverst til højre i tabellen fås informationer om modellen. Først antallet af observationer i regressionsmodellen, efterfulgt af et par goodness-of-fit mål. Her gives en likelihood-ratio test (LR chi2(5) samt Prob>chi2), der i dette tilfælde viser, at modellen som helhed er statistisk signifikant (simultan test af alle fem parametre, dvs. undtagen konstantleddet _cons). Derefter McFaddens pseudo R^2, der ukorrekt men populært sagt bliver fortolket som, at modellen samlet set kan forklare 2,45 procent af variationen i partitilknytning (den afhængige variabel). Men i og med at der er tale om en pseudo R^2 er det altså en forkert fortolkning. Målet angivet i stedet forbedringen af vores maximum likelihood-værdi når man medtager de forklarende variable.

Bunden af tabellen ligner til forveksling, hvad man får ud af en lineær regression. Der er regressionskoefficienter og standardfejl præcis som i en OLS-regression. En mindre detalje er de rapporterede signifikanstests. I outputtet kan det ses, at der rapporteres z-værdier i stedet for t-værdier. Ligesom ved en OLS-regression bygger de på nulhypotesen om en koefficient med

værdien 0, men det antages i denne forbindelse, at de følger en normalfordeling i stedet for en t-fordeling.

Koefficienterne er, som i eksemplet ovenfor, log odds (altså den naturlige logaritme af odds ratios). I tabel 3.6. ses, at der er en statistisk signifikant effekt af tillid til Folketinget på respondenternes partitilknytning ($p<0,05$). Det kan ligeledes ses, at koefficienten er positiv (0,1335), hvilket viser, at jo større tilliden er til Folketinget, desto større er sandsynligheden for, at respondenten føler sig knyttet til et politisk parti.

Forskellige mål kan fås ved hjælp af postestimationskommandoen *fitstat*. Kommandoen er ikke en del af standardinstallationen i Stata, men kan findes og installeres gennem *findit fitstat*. Her får man diverse mål fra modellen, herunder forskellige pseudo R^2-mål. Det er vigtigt at huske på, at *fitstat* er en postestimeringskommando, hvilket betyder, at den tager udgangspunkt i den senest estimerede model.

Trin 2: Udregning af forudsagte sandsynligheder

Det næste trin går ud på at udregne de forudsagte sandsynligheder. Når man skal udregne effekten af en uafhængig variabel i en multipel logistisk regression, skal man – modsat i en multipel OLS regression – forholde sig til, hvad værdierne er på de andre uafhængige variable i modellen. Det vil med andre ord sige, at de forudsagte sandsynligheder, man får, påvirkes af værdierne på de forskellige uafhængige variable.

Der er to forskellige tilgange til dette. For det første kan man fokusere på de observerede værdier på de uafhængige variable, udregne forudsagte sandsynligheder for hver af disse observationer/respondenter og så udregne gennemsnittet af disse (Hanmer og Kalkan 2013). For det andet kan man udregne de forudsagte sandsynligheder for en illustrativ, gennemsnitlig eller repræsentativ observation.

Vi kan begynde med at udregne ændringen i sandsynligheden for de forskellige variable. De forudsagte ændringer i sandsynligheder ved de observerede værdier kan bestilles med postestimeringskommandoen *margins*, der giver effekten af alle variable i modellen, og optionen *dydx()* angiver, at det er ændringen i sandsynligheden ved en ændring på én enhed i variablen, vi er interesseret i:

```
margins, dydx(*)
```

Tabel 3.7 viser resultaterne i form af gennemsnitlige marginale effekter, der er udregnet med udgangspunkt i de observerede værdier for de andre variable (hvis man i stedet ønsker at tage udgangspunkt i gennemsnittet på de andre variable, kan man tilføje optionen *atmeans*). Her ses det, at værdien for folketing er 0,0266, hvilket betyder, at sandsynligheden for at være knyttet til et parti stiger med 2,66 procentpoint, for hver stigning i tillid til Folketinget med én enhed.

Tabel 3.7: Output med de forudsagte sandsynligheder i Stata ved observerede værdier

```
Average marginal effects                          Number of obs   =       1307
Model VCE    : OIM
Expression   : Pr(parti), predict()
dy/dx w.r.t. : folketing kvinde alder indkomst uddannelse

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   folketing |   .0265502   .0051193     5.19   0.000     .0165166    .0365839
      kvinde |  -.0026236   .0249996    -0.10   0.916     -.051622    .0463747
       alder |   .0024981   .0006933     3.60   0.000     .0011392     .003857
    indkomst |   .0002303   .0045877     0.05   0.960    -.0087615    .0092221
  uddannelse |  -.0043331   .0075576    -0.57   0.566    -.0191457    .0104794
------------------------------------------------------------------------------
```

Som det næste ønskes de forudsagte sandsynligheder ved forskellige værdier af tillid til Folketinget. Dette kan eksempelvis gøres ved at sætte kontinuerlige variable til gennemsnittet og kategoriske variable til medianen, da det for kategoriske variable ikke giver nogen mening at tale om, at man kan være eksempelvis 0,5 kvinde. Igen bruger man *margins* kommandoen, hvor det tilføjes, hvilken variabel man gerne vil have de forudsagte sandsynligheder for. Her kigger vi på effekten af én variabel, tillid til Folketinget, som specificeres med muligheden *over()* og de andre variable ved bestemte værdier, specificeret med *at()*:

```
margins, over(folketing) at(kvinde=0 alder=48 uddannelse=4 indkomst=6)
```

Tabel 3.8 viser resultaterne. Her ses det, at den forudsagte sandsynlighed for en person med ingen tillid til Folketinget er 0,5347, altså 53,47%. For en person med maksimal tillid til Folketinget er

den forudsagte sandsynlighed 0,8137, altså 81,37%. Man kan derfor bruge beregningerne til at fastslå, hvordan partitilknytning varierer med folks tillid til Folketinget.

Tabel 3.8: Output med de forudsagte sandsynligheder i Stata ved bestemte værdier af x'erne

```
. margins, over(folketing) at(kvinde=0 alder=48 uddannelse=4 indkomst=6)

Predictive margins                                 Number of obs   =       1307
Model VCE    : OIM
Expression   : Pr(parti), predict()
over         : folketing
(del af output fjernet)
------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   folketing |
           0 |   .5347534   .0436976    12.24   0.000     .4491076    .6203992
           1 |   .5677702   .0375421    15.12   0.000     .4941891    .6413513
           2 |   .6001957   .0316259    18.98   0.000     .5382101    .6621813
           3 |   .6317663   .0262708    24.05   0.000     .5802764    .6832562
           4 |   .6622468    .021857    30.30   0.000     .6194078    .7050858
           5 |   .6914366   .0187937    36.79   0.000     .6546017    .7282715
           6 |   .7191735   .0173553    41.44   0.000     .6851577    .7531892
           7 |   .7453353   .0174211    42.78   0.000     .7111906    .7794801
           8 |     .76984   .0184798    41.66   0.000     .7336203    .8060596
           9 |   .7926427   .0199585    39.71   0.000     .7535247    .8317606
          10 |   .8137329   .0214505    37.94   0.000     .7716907     .855775
------------------------------------------------------------------------------
```

Trin 3: Grafisk visualisering

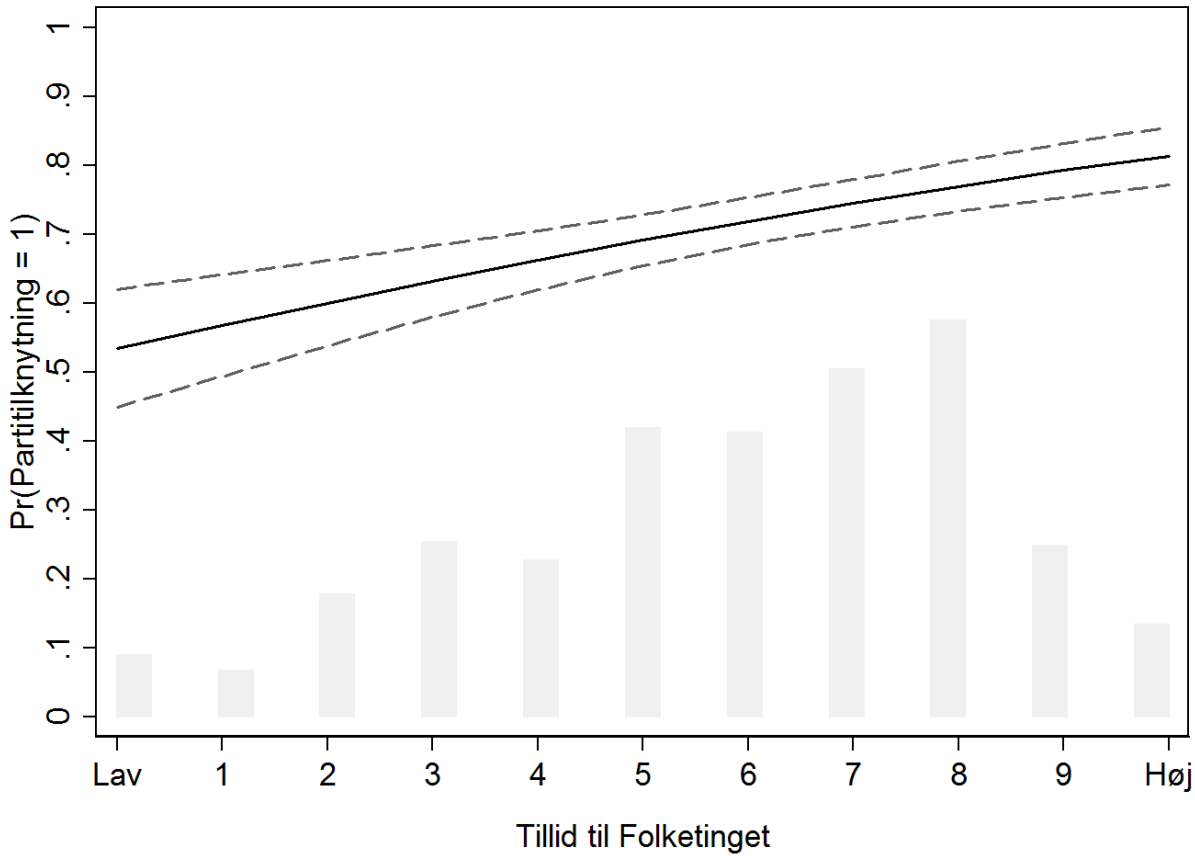
Det tredje og sidste trin er grafisk visualisering af resultaterne. Der kan være tilfælde, hvor det at visualisere sine resultater grafisk giver bedre mening end i andre tilfælde, men ikke desto mindre anbefales det, at man som udgangspunkt visualiserer resultaterne fra en logistisk regression, og kun afviger fra dette, såfremt der er praktiske grunde herfor - eksempelvis at man kun skal formidle to forudsagte sandsynligheder på begrænset plads.

Der er flere fordele ved at visualisere mange forudsagte sandsynligheder. For det første kan man også visualisere usikkerheden i form af 95% konfidensintervaller. Denne slags information kan også formidles i en tabel, men det kan blive til for mange informationer samt gøre det unødigt besværligt at skulle aflæse udviklingen i konfidensintervallerne. For det andet kan man, som det vil blive gjort nedenfor, også visualisere fordelingen af den uafhængige variabel, man gerne vil udtale sig om. Dette er nyttigt, når man arbejder med kontinuerlige variable, hvor der ved nogle værdier

kan være et begrænset antal – eller i nogle tilfælde ingen – observationer. Her bliver læseren gjort bekendt med, hvor det giver bedst mening at tolke resultaterne, og hvor man skal være mere forsigtig.

Kommandoen der anvendes her er *marginsplot*, der ligesom margins er en postestimerings-kommando, hvorfor den tager udgangspunkt i de senest estimerede forudsagte sandsynligheder. I princippet kan man blot nøjes med at køre marginsplot, men i nedenstående eksempel har jeg tilføjet et histogram med *addplot()* samt foretaget andre æstetiske modifikationer.

```
. marginsplot, recast(line) recastci(rline) xlabel(0 "Lav" 1 "" 2 "" 3 "" 4 "" 5 "" 6 "" 7 "" 8 "" 9 "" 10 "Høj") ciopts(lpattern(dash)) scheme(s1mono) title("") ytitle(" " "Pr(Partitilknytning = 1)") xtitle(" " "Tillid til Folketinget") legend(off) addplot(hist folketing, bcolor(gs15) ylabel(0(.1)1) below)
```



Figur 3.1. Forudsagte sandsynligheder med konfidensintervaller

I figuren ses de forudsagte sandsynligheder for alle værdierne på den uafhængige variabel (fuldt optrukne linje), 95% konfidensintervaller for alle forudsagte sandsynligheder (de to stiplede linjer) samt fordelingen af den uafhængige tillidsvariabel (søjlerne på den horisontale akse). Sidstnævnte giver læseren mulighed for at få et indblik i fordelingen af variablen, og dermed hvor mange observationer der er i de kategorier, man udtaler sig om.

3.6. Logistisk regression i SPSS

Man kan også udføre logistiske regressioner i SPSS, I SPSS er indgangen til logistisk regression via undermenuen *Regression* til menuen *Analyze*. Her findes en række valgmuligheder, hvoraf to knytter sig til binære udfaldsvariable:

- *Binary Logistic* udfører en logit regression som den ovenfor beskrevne. Proceduren har en række tilføjede funktioner og tests. Se hjælpemenuen for nærmere detaljer.
- *Probit* udfører ligeledes en logit regression, men kan også omstilles til at estimere med probit, som baseres på normalfordelte sandsynligheder. Se hjælpemenuen for nærmere detaljer.

SPSS er imidlertid mindre velegnet til at beregne de uafhængige variables marginale effekter.

3.7. Afrunding

Nærværende kapitel har givet en anvendelsesorienteret introduktion til den logistiske regressionsmodel. De væsentligste forskelle på en OLS-regression og en logistisk regression er blevet gennemgået, og den substantielle tolkning af resultaterne fra en logistisk regression har været hovedfokus. I denne proces er der opstillet tre trin som man bør følge ved logistisk regressionsanalyse. Der er mange forhold, man skal tage højde for, og nedenstående tjekliste giver et par gode råd til, hvad man skal huske:

1) Inspicer den afhængige variabel, herunder sørg for at have styr på kodningen, så den kun har værdien enten 0 eller 1.
2) Er de essentielle informationer rapporteret i regressionstabellen?
    a. Koefficienter for de uafhængige variable

b. Standardfejl for de uafhængige variables koefficienter
c. Antal observationer i modellen
d. Log-likelihood for modellen
e. Pseudo R^2 for modellen.

3) Kalkuler forudsagte sandsynligheder.
    a. Rapporter hvordan de er kalkuleret, herunder forudsætninger vedr. de uafhængige variables niveau.
4) Lav en visualisering af resultaterne.
    a. Undersøg om der er tilstrækkeligt med information inkluderet i visualiseringen af de forudsagte sandsynligheder (Eksempelvis konfidensintervaller for forudsagte sandsynligheder og evt. fordelingen af den uafhængige variabel).

Listen er ikke udtømmende. Det er vigtigt at holde sig for øje, at alt efter hvordan man specificerer sin logistiske regressionsanalyse, kan der være yderligere aspekter, man skal forholde sig til. Dette gælder som nævnt ovenfor i forhold til tolkningen af koefficienterne i analysen, men også for tolkningen af eksempels interaktionseffekter (Ai og Norton 2003; Berry, DeMeritt og Esarey 2010; Rainey 2015).

Med hensyn til den praktiske udførelse i statistikprogrammet skal man også være opmærksom på hvilken hændelse der estimeres sandsynligheder for, dvs. om det er nul – eller ethændelsen. Ligeledes skal man holde for øje, at den iterative procedure bag maximum likelihood estimeringen stiller større krav til antallet af observationer for at kunne gennemføres stabilt, hvilket dog ikke er specifikt for logit-regression. Det er således risikabelt at bruge maximum likelihood på små stikprøver. Jo flere uafhængige variable der er, desto flere observationer skal der endvidere også være (Long 1997, 53f). Undlad derfor også at inkludere for mange uafhængige variable, hvis du har relativt få observationer at gøre godt med.

I en logistisk regression gør man sig ingen antagelser om fordelingen af de uafhængige variable, men hvis de korrelerer betydeligt, kan det være vanskeligt at estimere den selvstændige effekt af hver variabel. Dette multikollinearitetsproblem er identisk med problemet i en OLS-regression (kapitel 2).

I den ovenfor præsenterede gennemgang af binær logistisk regression er taget udgangspunkt i at $P(y = 1)$ følger en logistisk fordeling. Som alternativ kan bruges andre fordelinger, herunder

normalfordeling, hvormed man opnår den såkaldte probit-estimation, se evt. Long (1997) for en gennemgang. For alle praktiske formål fås sammenlignelige resultater fra probit og logit. Forskellen er at logit beregningsmæssigt er simplere end probit, hvilket historisk var en fordel da man havde begrænset regnekapacitet. Endvidere kan valget af specifikation have betydning for avancerede tilgange til binær logistisk regression, herunder generalisering til paneldata. Disse er belyst i Stata-manualerne og kan for en matematisk funderet læser også findes beskrevet hos Wooldridge (2010).

For de læsere der har interesse i at lære mere om den logistiske regression, gives der en introduktion til flere praktiske funktioner i Stata i kapitel 11 i Sønderskov (2014), og for de læsere der ønsker en pædagogisk introduktion til logistisk regression i programmet R, henvises der til kapitel 5 i Gelman og Hill (2007).

Referencer

Ai, C. R. og E. C. Norton. (2003). Interaction terms in logit and probit models. *Economics Letters*, 80 (1): 123-

Berry, W. D., J. H. R. DeMeritt og J. Esarey. (2010). Testing for Interaction in Binary Logit and Probit Models: Is a Product Term Essential? *American Journal of Political Science*, 54 (1): 248–266.

European Social Survey Round 7 Data (2014). *Data file edition 1.0. Norwegian Social Science Data Services, Norway – Data Archive and distributor of ESS data for ESS ERIC*.

Gelman, A. og J. Hill. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York: Cambridge University Press.

Hagle, T. M. og G. E. Mitchell. (1992). Goodness-of-Fit Measures for Probit and Logit. *American Journal of Political Science*, 36 (3): 762-784.

Hanmer, M. J. og K. O. Kalkan. (2013). Behind the Curve: Clarifying the Best Approach to Calculating Predicted Probabilities and Marginal Effects from Limited Dependent Variable Models. *American Journal of Political Science*, 57 (1): 263–277

Hellevik, O. (2009). Linear versus logistic regression when the dependent variable is a dichotomy. *Quality & Quantity*, 43 (1): 59-74.

Kastellec, J. P. og E. L. Leoni. (2007). Using Graphs Instead of Tables in Political Science. *Perspectives on Politics*, 5 (4): 755-771.

Kristensen, C. J. og M. A. Hussain. (2016). *Metoder i samfundsvidenskaberne.* København: Samfundslitteratur.

Long, J. S. (1997). *Regression Models for Categorical and Limited Dependent Variables*. Sage Publications.

Rainey, C. (2015). Compression and Conditional Effects: A Product Term Is Essential When Using Logistic Regression to Test for Interaction. *Political Science Research and Methods*. doi: 10.1017/psrm.2015.59.

Sønderskov, K. M. (2014). *Stata – en praktisk introduktion*. København: Hans Reitzels Forlag.

Wooldridge, J.M. (2010). *Econometric Analysis of Cross Section and Panel Data.* Cambridge: MIT Press.