

Logistisk regression

Metode 3

Link til slides: kortlink.dk/2ab58

Søren Damsbo-Svendsen
sdas@ifs.ku.dk

Institut for Statskundskab
Københavns Universitet

Uge 10

Oversigt

Uge	Holdtime	Emne	Øvelsesopgave
6		Kvantitativ indholdsanalyse	1
7		Diskursanalyse I	2
8	25-26 februar 2021	Diskursanalyse II	3
9	04-05 marts 2021	Interaktioner og modelspecifikation	4
10	11-12 marts 2021	Logistisk regression	5
11	18-19 marts 2021	Multilevel analyse	6
12		Kausal inferens I: Kausalitet og instrumentvariable (IV)	7
13		Påskeferie	
14	08-09 april 2021	Kausal inferens II: Paneldata	8
15	15-16 april 2021	Kausal inferens III: Eksperimentelle designs	9
16		Kausal inferens IIII: Regression Discontinuity (RD)	10
17		Social Data Science: Big Data	11
18		Process tracing	12
19		Kriterier for god videnskab	
20		Opsamling/spørgetime	
23		Aflevering af skriftlig hjemmeopgave (07 juni 2021)	

Recap fra sidste gang

- Interaktion er, når effekten af X på Y afhænger af, hvilket niveau Z har
 - X og Z interagerer; Z modererer $X \rightarrow Y$
- Super smart redskab til at få flere nuancer og mere teoretisk dybde frem i sin analyse
- Eksempel i Stata: **reg y c.x##i.z**
- Interaktionsleddet er *forskellen i effekt eller noget ekstra, der lægges til effekten*
- Marginale effekter er de *forskellige effekter, X har på Y* for hvert af de respektive niveauer af Z:
 - **margins, dydx(x) at(z = (1 2 3 4 5))** → **marginsplot**

Dagens program

1. Eksamensopgaven

2. Logistisk regression

- Baggrund
- OLS vs. logistisk regression
- Fortolkning/præsentation

3. Øvelsesopgaver undervejs - med/uden break-out

Eksamensnotat

Hjælpenotat til Metode 3 eksamensopgave

Carolin Hjort Rapp

Forår 2021

Dette notat beskriver de formelle rammer for eksamensopgaven i Metode 3 og præsenterer nogle mulige koncepter for opgaven som kan tjene som inspiration for opgaveskrivere.

Formalia for eksamensopgaven

Eksamensopgaven i Metode 3 består af en fri skriftlig hjemmeopgave. Som det fremgår af studieordningen består en fri skriftlig hjemmeopgave af "en analyse af en selvstændigt formulert problemstilling". Modsat i Metode 2, hvor der ønskes en analyse med afsæt i én specifik dataindsamlingsmetode i form af et selvstændigt gennemført survey, er der ikke specifikke krav til hvilken metode opgaven skal tage afsæt i.

Opgaven er først og fremmest en metodeopgave, så det er ikke tanken, at der skal bruges mange kræfter på den teoretiske motivation, men det er selvfølgelig oplagt at trække på relevante teorier fra f. eks. Dansk og Komparativ Politik, Offentlig forvaltning, International Politik eller andre fag som kan bruges til at motivere problemstillingen. Opgaven vurderes på i hvilket omfang I mestrer at analysere en samfundsvidskabelig problemstilling med anvendelse af de redskaber, I har lært i metodeforløbet.

Omfang

Formkrav for skriftlige opgaver bestemmes i https://samf.ku.dk/uddannelser/studenter/service/regleroglove/studieordninger/rammestudieordning_SAMF.pdf. I den d.d. seneste version fremgår det således bl.a. (afsnit 4.5):

Normalsidetallet og antallet af typeenheder skal fremgå af opgavens forside. I optælling af typeenheder inkl. mellemrum inkluderes al tekst i den skriftlige fremstillings hovedtekst. Dvs. inklusive fodnoter, slutnoter, tabeller, ligninger og formler. Følgende tæller ikke med i optællingen af antal typeenheder inkl. mellemrum: Forside, indholdsfortegnelse, evt. resumé eller abstract, litteraturliste, figurer, grafer etc.

Det maksimale omfang for opgaven er jf. studieordningen som følger:

- For én studerende: 19.200 typeenheder (8 normalsider)
- For to studerende: 24.000 typeenheder (10 normalsider)
- For tre studerende: 28.800 typeenheder (12 normalsider)

1. **Litteraturliste:** Der skal være en samlet litteraturliste for opgaven. Litteraturlisten udformes efter normal dansk samfundsvidskabelig stil. Dvs. løbende litteraturhenvisninger angives i brødteksten med parentes (Svensson 1981: 13). Litteraturlisten angiver forfatter (hvor fornavnene ikke skrives helt ud), udgiveset (i parentes), titel, udgivessted: forlag. Ved artikler og bogkapitler angives tillige sidetal. Det er tilladt at henviser til forelæsningsslides såfremt den relevante pointe ikke fremgår af pensumteksten.

2. Layout:

- Foucault, M. (2008). *Sikkerhed, territorium, befolkning*. København: Hans Reitzels Forlag.
- Svendsen, S. (2008) "Politik og vanvid." *Politiske studier*, 17:4, 24-35.
- Hansen, P.K. (2004) "England" i Karstensen, N. (red.) *Alle verdens lande*. København: Gyldendal, 45-75.
- Ved flere forfattere og redaktører skrives navnene i normal rækkefølge, undtagen det første, dvs. det efternavn, der bruges til at alfabetisere efter.

3. **Tabeller og figurer:** Tabeller opstilles med enkelt linjeafstand, ingen lodrette streger, og normalt ingen vandrette streger i selve tabellen. Noter placeres under tabellen i mindre skrift. Tabeloverskrift i fed og fortløbende nummerering af tabeller. Kolonneoverskrifter centrerer, tekst i forspalten (og i teksceller i øvrigt) venstrestilles. Der laves separat kolonne til signifikans tegn angivet med asterisk, som venstre stilles. (se *Politica* hvis I er i tvivl om layout).

4. **Nyt afsnit:** Laves med ét linjeskift og indryk (dvs. tabulator) og ikke med en hel linjes mellemrum.

5. **Formatering:** Anvend en læsevenlig formatering, fx Times New Roman med punkstørrelse 12 og halvanden linjeafstand. Undgå orddelinger, skiftende skriftstørrelser og skriftypen, sidehoved og sidefod etc. Brug aldrig flere mellemrum efter hinanden, men i stedet tabulator.

Koncepter for opgaven

Notatet her præsenterer nedenfor tre forskellige 'koncepter' for eksamensopgaven. De er ikke tænkt som en udtømmende liste, men kan tjene som inspiration til måder at gå til opgaven på. Som koncepterne illustrerer behovet for opgaven, modsat Skriveøvelse 1 i Metode 1 og eksamensopgaven i Metode 2, ikke omfatte selvstændig dataindsamling. Det er således kun det første af de tre koncepter der lægger op til selvstændig dataindsamling, mens de to øvrige trækker på eksisterende datakilder.

Koncepterne er idealtyper som ikke behøver være gensidigt udelukkende. Man kunne fx. oplagt kombinere en kritisk diskussion af et eksisterende studie inden for et af fagets emner med en selvstændig analyse på et andet datasæt.

Som det fremgår kan en opgave dermed også bygges op omkring en enkelt af de metoder, der præsenteres i Metode 3. En sådan opgave bør dog sikre sig samtidig at perspektivere med begreber fra nogle andre dele af faget – fx. efter samme model som den kvalitative perspektivering i Metode 2-opgaven – for at undgå at opgaven bliver for enstregnet.

Eksamens I

- Læs eksamensnotatet på Absalon
- Omfang (typeenheder): 19.200 for én studerende, 24.000 for to studerende, 28.800 for tre studerende
- Kraftig anbefaling: skriv sammen i **grupper** (2-3)
- Eksempler på opgaver kommer på Absalon før påske
 - inspiration, ikke en skabelon
- Aflevering **7. juni 2021** (kl. ??)
- Brug **mindst én metode** fra faget og perspektivér til mindst én anden
 - flere metoder ≠ bedre opgave

Eksamens II

- "analyse af en selvstændigt formuleret problemstilling" → vigtigt!
 - find et substantielt spørgsmål - og en metode - der interesserer jer
 - nogle gange kan det også hjælpe at starte med at finde data/materiale
- Tre gode muligheder fra eksamensnotatet:
 - egen undersøgelse (nyt - mindre skala)
 - arbejd videre med tidligere projekt (fx metode 2-survey data)
 - replikation/kritisk diskussion af eksisterende undersøgelser (smart!) - skal *anvende* metoderne

Vejledning

- Ca. 15 min. per gruppe ad to omgange
- På Zoom
- Lav et kort skriv om tankerne bag opgaven
- Første gang efter påske (muligvis uge 16) og anden gang nær slutningen (muligvis uge 20)
- I hører nærmere!

Dagens formål

- At forstå hvornår og hvorfor, vi bruger logistisk regression (frem for OLS)
- At vide hvordan man laver logistisk regression i Stata
- At vide hvordan man fortolker (og præsenterer) sine resultater

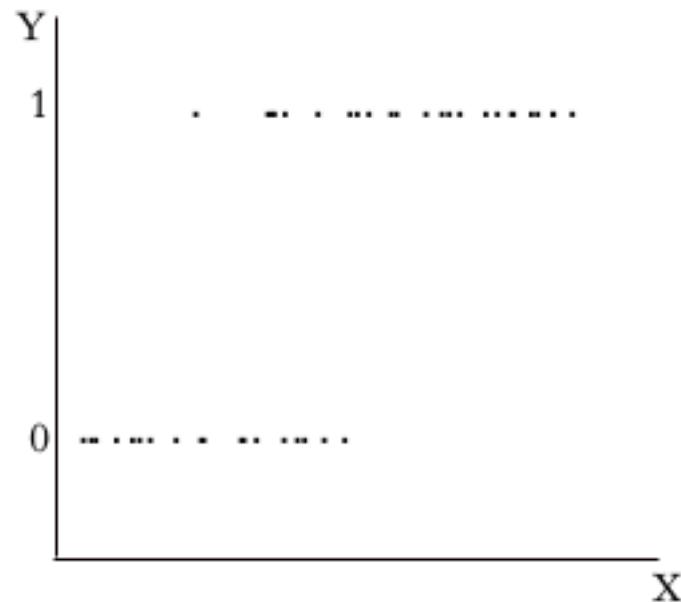
Pensum

- Kellstedt, P. M., Whitten, G. D. (2018). The Fundamentals of Political Science Research. 3rd edition. Cambridge University Press. Kapitel 12.1-12.2.
- Sønderskov, Kim Mannemar (2014). Stata - en praktisk introduktion (2. udg.). København: Hans Reitzels forlag. Kap. 11

Recap fra forelæsningen

Logistisk regression bruges når vores afhængige variabel har et dikotomt udfaldsrum (0/1)

- Visuelt: Alle observationer ligger vandret enten ud for enten $y = 0$ eller $y = 1$
- Når Y er 0 eller 1, kan Y i vores modeller fortolkes som den forudsagte **sandsynligheden for Y fra 0-1**



- Hvad kunne et eksempel på en dikotom afhængig variabel (Y) være?

Hvad vil vi med regressioner?

Husk at med regression forsøger vi at finde den model, som bedst beskriver forholdet mellem X og Y:

- Den bedste linje
- OLS estimerer den rette linje, som minimerer de kvadrerede residualer
 - minimerer hvor meget vi med modellen skyder over/under de sande værdier af Y
 - ret linje = konstant hældning = linearitet
 - hældning = β

(Bl.a.) derfor er OLS fedt:

- Det er **intuitivt** - koefficienten som beskriver forholdet mellem X og Y har en intuitiv fortolkning:
 - når X vokser med +1 udtrykker koefficienten den tilsvarende ændring i Y
- Det er **effektivt** - forholdet mellem X og Y er opsummeret i **ét tal**

Udfordring med OLS

Den store udfordring, når Y er binær:

- Data beskrives åbenlyst *ikke* bedst af en lige linje med konstant hældning
- Det komplicerer vores tolkning markant:
 - Hvordan en ændring i X påvirker sandsynligheden for, at $Y=1$ **afhænger af, hvor på x-aksen vi står**

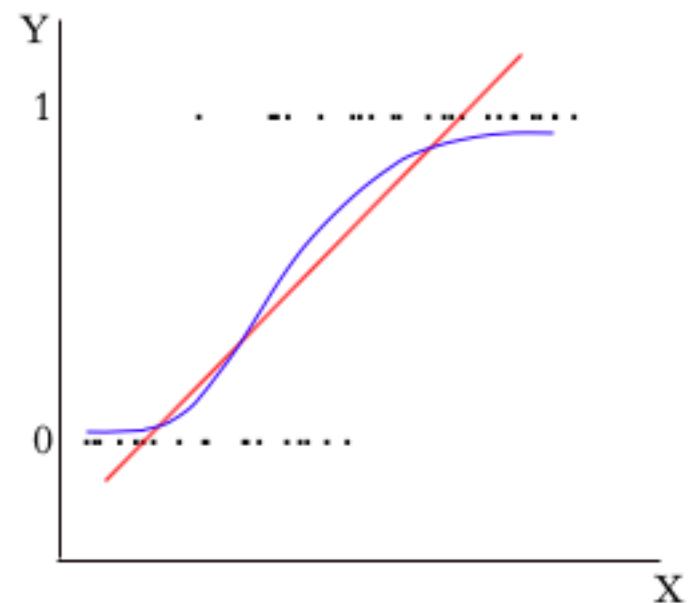
Kan vi alligevel udtrykke det med *ét tal* som i OLS?

- Det kan vi, men dette tal ("logit") er ikke intuitivt
- Heldigvis findes der løsninger, når vi kommer til fortolkning

Hvordan ser logistisk regression ud grafisk?

Her er plottet fra før. Der er blot indsat to linjer til at beskrive data

Hvordan og hvor godt beskriver de data
(dvs. sammenhængen mellem X og Y)?



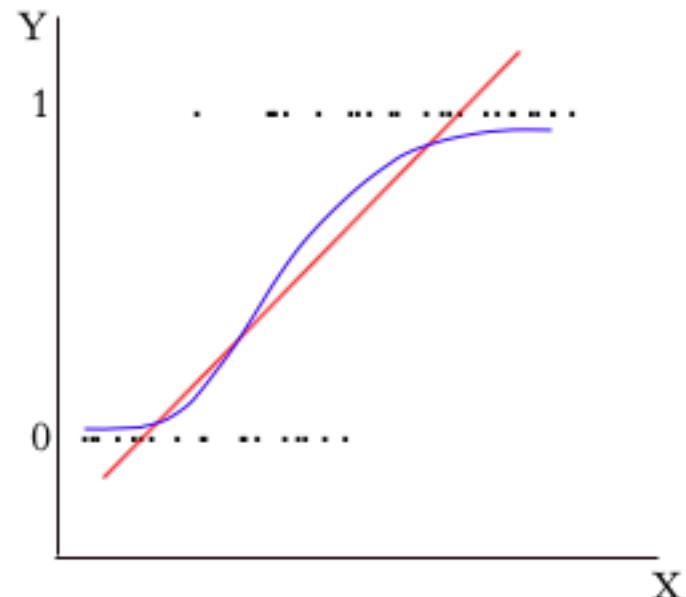
Hvordan ser logistisk regression ud grafisk?

Rød linje:

- en ret linje - passer ikke så godt på punkterne
- = OLS som I kender det → kaldes også en "lineær sandsynlighedsmodel" (LPM), når Y er dikotom

Blå linje:

- en S-formet linje, som smyger sig langs punkterne - passer bedre
- = **logistisk regression**



OLS vs. logit

Vi vil lave en model, der har et binært Y:

1. Først ser vi på en "lineær sandsynlighedsmodel" → OLS med binært Y
2. Så "udvider" vi til **logistisk regression** og sammenligner de to

OLS når Y er binær

Vi starter med lineær regression (OLS)

Her er sandsynligheden for, at $Y=1$ en lineær funktion af X:

$$P(Y = 1|X) = \alpha + \beta X$$

... helt som I kender det (med lidt ny notation)

Den lille forskel ligger i fortolkningen af β :

- "Normal OLS": *Ændringen i den forudsagte værdi af Y, når X vokser med +1*
- OLS med binært Y: *Ændringen i sandsynligheden for, at Y=1, når X vokser med +1*

Øvelsesopgave 5.6

Eksempel på OLS med binært Y

Hvordan vil I fortolke
outputtet?

Afhængig variabel
(*stempers*):

Stemte personligt
(1 = "Ja", 0 = "Nej")

Uafhængige variable:
kvinde (ref. mand), *alder*
(17-100), *indkomst* (16
niveauer), *ungdomsudd* (5
niveauer)

reg stempers i.kvinde alder indkomst i.ungdomsudd, vsquish

Source	SS	df	MS	Number of obs	=	1,813
Model	11.3252239	7	1.61788912	F(7, 1805)	=	6.65
Residual	439.336663	1,805	.243399813	Prob > F	=	0.0000
Total	450.661886	1,812	.24870965	R-squared	=	0.0251

stempers	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
kvinde					
Kvinde	-.0122195	.023733	-0.51	0.607	-.0587665 .0343275
alder	.0040614	.0008276	4.91	0.000	.0024382 .0056846
indkomst	.0107962	.0028663	3.77	0.000	.0051745 .0164178
ungdomsudd					
2	-.0251733	.042096	-0.60	0.550	-.1077352 .0573887
3	-.0526356	.0386431	-1.36	0.173	-.1284255 .0231544
4	-.0373645	.0411475	-0.91	0.364	-.1180662 .0433372
5	-.0231555	.0646459	-0.36	0.720	-.1499441 .1036332
_cons	.2886389	.0645602	4.47	0.000	.1620184 .4152595

Det fede ved OLS

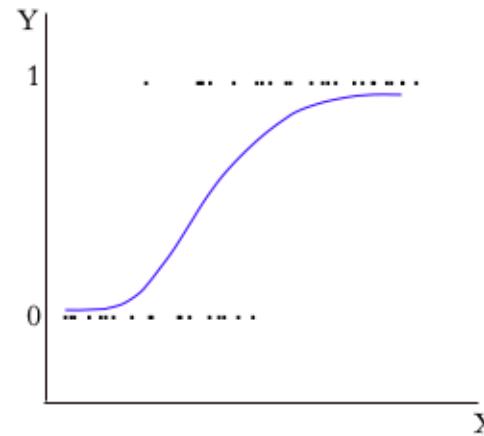
- Fortolkningen er lige til - og vi kender det i forvejen
- (At undersøge interaktioner er ligeud ad landevejen - vender vi tilbage til senere)

Det ufede ved OLS

Der er en række **grunde til, at OLS er problematisk, når Y er en dummy**

- Problemer med fejlled → altid heteroskedastiske og aldrig normalfordelte
- Den funktionelle form (linearitet) er problematisk
- Den rette linje løber i principippet fra $-\infty$ til ∞
 - betyder at vores model kan forudsige (absurde) sandsynligheder større end 100 % og mindre end 0 %

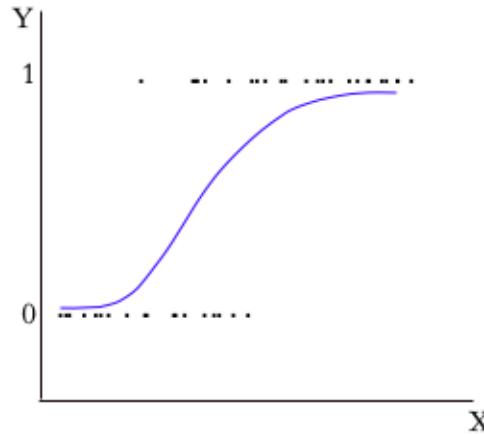
Logistisk regression



Logistisk regression er alternativet til OLS, når vi har en **binær afhængig variabel**

- I logistisk regression begrænses de forudsagte Y-værdier til intervallet [0,1]
- Forholdet mellem X og Y er S-formet (ikke lineært)

Logistisk regression



Logistisk regression fungerer ved at lave en matematisk transformation

- Den generelle logistiske funktion ser således ud:

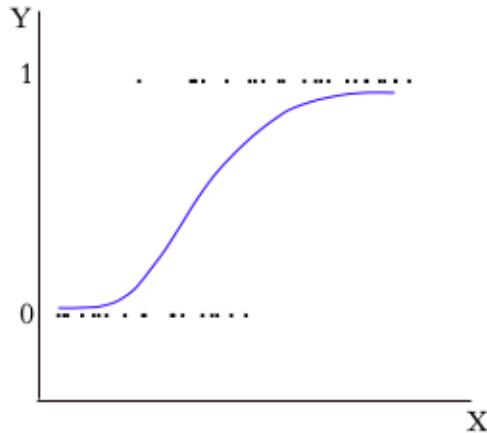
$$F(x) = \frac{e^x}{1+e^x}$$

- Vi indsætter vores lineære sandsynlighedsmodel på X's plads og får:

$$Y = P(Y = 1|X) = F(\alpha + \beta X + \epsilon) = \frac{e^{\alpha+\beta X+\epsilon}}{1+e^{\alpha+\beta X+\epsilon}}$$

- Modellen giver os sandsynligheden for $Y = 1$ som en *logistisk funktion af X*
 - dette trick giver os en meget bedre funktionel form

Logistisk regression



- Vi modellerer stadig effekten af en uafhængig variabel på sandsynligheden for Y - blot med en bedre model
- Fordi den logistiske funktion ikke er lineær, varierer effekten af X på Y med størrelsen af X

Tænk på det sådan her:

- Spørgsmål: "Hvad er hældningen på den S-formede kurve?"
- Svar: "Det kommer an på, hvor på X-variablen, du kigger!"
 - fx stejlest på midten

→ Ja, bedre funktionel form, men **tolkningen bliver sværere**

OLS vs. logit - opsummering

- Tænk på logistisk regression som **en udvidelse af lineær regression** (det er dog en ikke-lineær model), som anvendes, når Y er binær
- Det kan give god mening at starte med at lave OLS - om ikke andet for ens egen skyld
- OLS på binært Y medfører problemer, men har et lidt dårligere ry end hvad rimeligt er
 - Man skal ikke kimse af den værdi, der ligger i lettere fortolkning
 - Når man laver kausalstudier i praksis, er eksogenitetsantagelsen (fravær af spuriøsitet m.m.) langt vigtigere end valget mellem OLS vs. logistisk regression
- **I en metode 3-opgave skal man dog naturligvis bruge logistisk regression**

Fortolkning, eksempler og øvelsesopgaver

Logit-koefficienten

Når vi kører en logistisk regression i Stata

```
logit Y X1 X2 X3
```

... giver Stata os **logit-koefficienten**

- Den udtrykker **ændringen i den naturlige logaritme til oddset, når X stiger med +1**
 - ikke spor intuitivt!
- **P-værdi** kan tolkes som normalt
 - er der en **statistisk signifikant sammenhæng** mellem X og Y?
- **Fortegn**
 - **Positiv (+)**: sandsynligheden for $Y = 1$ *stiger*, når X vokser
 - **Negativ (-)**: sandsynligheden for $Y = 1$ *falder*, når X vokser

Øvelsesopgave 5.1

Eksempel med logit-koefficient

Hvad kan vi sige om
effekten af alder på
sandsynligheden for at
stemme personligt?

```
. logit stempers i.kvinde alder indkomst i.ungdomsudd, nolog
```

	Logistic regression	Number of obs	=	1,813
		LR chi2(7)	=	45.94
		Prob > chi2	=	0.0000
	Log likelihood = -1228.5243	Pseudo R2	=	0.0184

Andet?

stempers	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
kvinde					
Kvinde	-.0503625	.0976871	-0.52	0.606	-.2418257 .1411007
alder	.0166559	.0034484	4.83	0.000	.0098972 .0234147
indkomst	.044231	.0118825	3.72	0.000	.0209417 .0675203
ungdomsudd					
2	-.1034898	.1743209	-0.59	0.553	-.4451525 .2381729
3	-.2157564	.1598628	-1.35	0.177	-.5290816 .0975689
4	-.1523873	.1706846	-0.89	0.372	-.4869229 .1821483
5	-.0929997	.2664117	-0.35	0.727	-.615157 .4291576
_cons	-.8676345	.2689866	-3.23	0.001	-1.394839 -.3404306

Odds ratio

- Hvis vi ønsker logit-koefficienten udtrykt som **odds ratio**, tilføjer vi blot optionen "**or**":
 - logit Y X1 X2 X3, or
- Den udtrykker, hvor mange gange større odds ($Y=1$) bliver, når X vokser med +1
 - faktor som oddset for "succes" ganges med
 - eksempel: "*oddsene for at stemme på blå blok er ca. 1,18 gange større for mænd end for kvinder*"
- **P-værdi** kan tolkes som normalt
 - er der en **statistisk signifikant sammenhæng** mellem X og Y ?
- "**Fortegn**":
 - **OR > 1**: sandsynligheden for $Y = 1$ *stiger*, når X vokser
 - **OR < 1**: sandsynligheden for $Y = 1$ *falder*, når X vokser

Øvelsesopgave 5.2

Eksempel med odds ratio - samme model med koefficienter udtrykt som odds ratio

Fortæller modellen noget,
som den "almindelige"
logit-model ikke fortalte
os?

Fokusér evt. på **alder**

Hvad er uændret?

Øvelsesopgave 5.2

Eksempel med odds ratio - samme model med koefficienter udtrykt som odds ratio

Samme model →
koefficienterne er bare
udtrykt anderledes

Alders-OR på 1.017: Når
alder stiger med ét år,
stiger **oddsne** for at
stemme personligt med
1,7%

Indkomst-OR på 1.045:
Når indkomst stiger med
ét niveau, stiger **oddsne**
for at stemme personligt
med 4%

Mere intuitive måder at rapportere effekter

- Logit-koefficienten og odds ratio er forsøg på at opsummere heterogene effekter med ét tal
 - ikke synderligt informativt eller intuitivt
- De fleste er nok enige i, at det mest intuitive er:
 - **Sandsynlighed**: "der er 70 % sandsynlighed for, at hun stemmer personligt"
 - **Ændringer i sandsynlighed**: "når X vokser med +1, stiger sandsynligheden for, at hun stemmer personligt med 5 procentpoint"
- **Effekten udtrykt som ændringer i sandsynligheder**
 - Average Marginal Effect (AME): Gennemsnittet af den marginale effekt *for alle niveauer af X*
 - Marginal Effect at the Mean (MEM): Den marginale effekt *ved gennemsnittet af X*

Gennemsnitlig Marginal Effekt (AME)

- Gennemsnittet af den marginale effekt ved alle niveauer af X
 - gennemsnitlig ændring i $P(Y = 1)$, når X ændres vokser med +1
 - gennemsnitlig hældning af S-kurven
 - ikke effekten i et bestemt punkt, men på tværs af alle punkter
 - bedste bud på den marginale effekt, *hvis man ikke vidste, hvor på X-aksen man stod*
- Fortolkning:
 - parallel til tolkningen i OLS
 - den gennemsnitlige effekt på $P(Y=1)$, når X stiger med +1
- Smart:
 - tillader os at opsummere resultatet fra en logistisk regression i et enkelt meningsfuldt tal
 - letter fortolkningen enormt meget
- Usmart:
 - særligt når der er stærk ikke-linearitet (S-form) kan den gennemsnitlige effekt være vildledende
- AME er good practice

Øvelsesopgave 5.3

Eksempel med Average Marginal Effect (AME)

Den gennemsnitlige
marginale effekt kan
(pyha, endelig!) fortolkes
som *ændringen i
sandsynlighed*

**Tolk på de gennemsnitlige
marginale effekter (dy/dx)**

Fokusér evt. på *alder* og
kvinde

Øvelsesopgave 5.3

Eksempel med Average Marginal Effect (AME)

Når alder stiger med +1,
stiger sandsynligheden
for at stemme personligt i
gennemsnit med ca. 0,4
procentpoint

Over 50 år svarer det til
en stigning på
 $0.004 * 50 = 0.2 \approx 20$
procentpoint

Sandsynligheden er ca. 1,2
procentpoint lavere for
kvinder end for mænd
(*insignifikant!*)

Marginal Effekt ved Gennemsnittet (MEM)

- Hvor stor er effekten af X på $P(Y = 1)$ i et bestemt punkt på grafen, dvs. for en bestemt værdi af X
 - ofte vælger man gennemsnittet på X , deraf navnet MEM
 - argumentet er, at dette niveau af X er *repræsentativt* i en eller anden forstand
 - men man kan ligeså godt vælge andre repræsentative/illustrative niveauer af teoretiske eller andre grunde
 - hældningen af S-kurven i et bestemt punkt
- Fortolkning:
 - Hvordan ændres sandsynligheden for $Y=1$, når X ændres med +1 - *for observationer som er gennemsnitlige på X*
- Kan også være mere eller mindre smart - *mere eller mindre repræsentativ for den samlede effekt*
 - afhænger bl.a. af S-kurvens hældning og den konkrete X -variabel

Øvelsesopgave 5.3

Eksempel med Marginal Effect at the Mean (MEM)

Tolk igen på de marginale effekter (dy/dx)

Husk at vi her har sat alle uafhængige variable på deres gennemsnit

Fokusér evt. på *alder* og *indkomst*

Øvelsesopgave 5.3

Eksempel med Marginal Effect at the Mean (MEM)

Med udgangspunkt i en person med gennemsnitlig alder, køn, uddannelse og indkomst estimeres følgende:

Når alder stiger med +1, øges sandsynligheden for at stemme personligt med ca. 0,4 procentpoint

Når indkomstniveau stiger med +1, øges sandsynligheden med ca. 1,1 procentpoint

Forudsagte sandsynligheder

Tidligere blev vi enige om, at de mest intuitive størrelser er **sandsynligheder** og **ændringer i sandsynligheder**

- Vi har kigget på *ændringer i sandsynligheder* i form af AME og MEM
- Men hvorfor forsøge så hårdt at opsummere i ét tal, når vi bare kan **plotte den forudsagte sandsynlighed for $Y = 1$ for alle værdier af X ?**
 - giver et hurtigt **overblik** over hele sammenhængen
 - forudsagt Y (sandsynlighed) på Y-aksen, forklarende X på X-aksen

Øvelsesopgave 5.5

Eksempel på forudsagte sandsynligheder

Ved hjælp af kommandoen

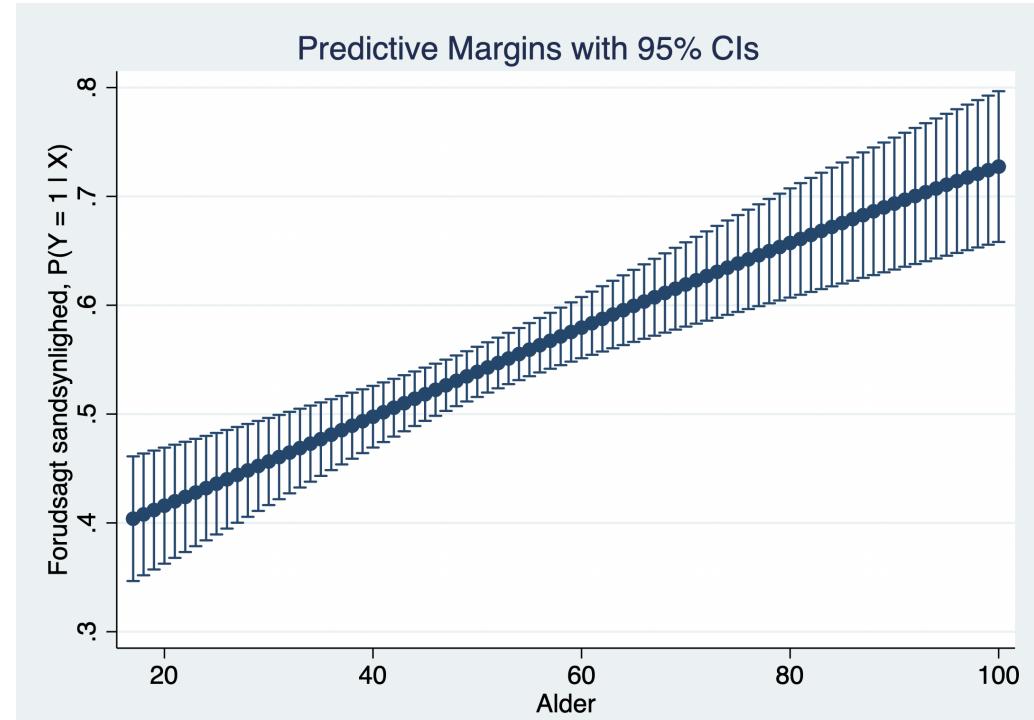
```
margins, at(alder=(17(1)100))
```

efterfulgt af

```
marginsplot
```

kan vi plotte de forudsagte sandsynligheder for at stemme personligt (Y) for alle niveauer af alder (X)

Hvad ser vi?



Øvelsesopgave 5.5

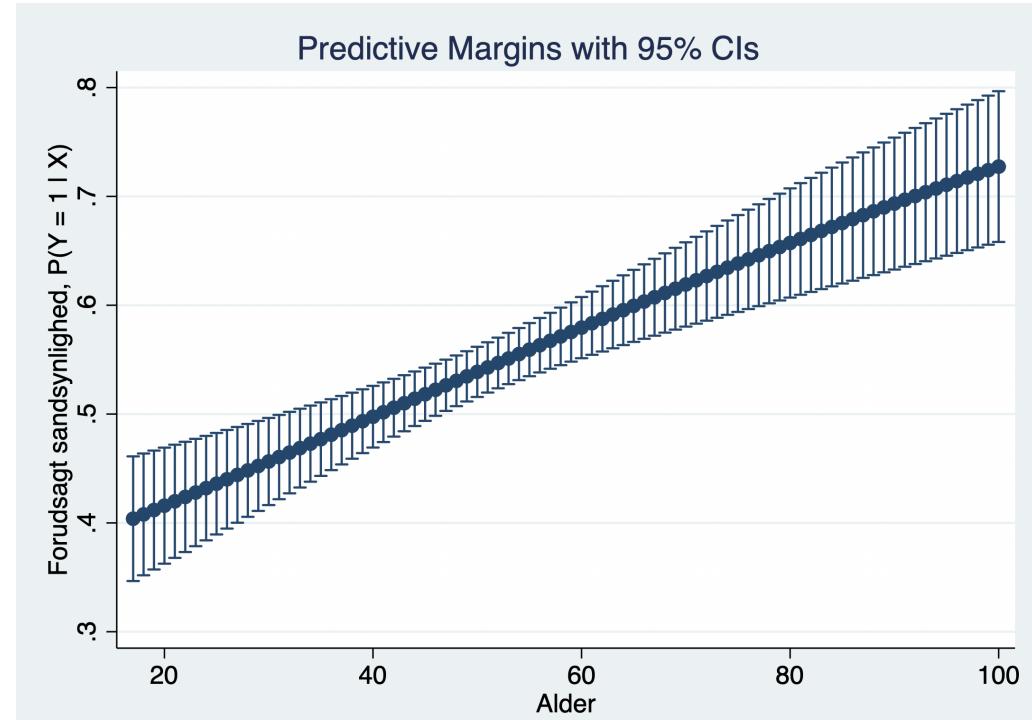
Eksempel på forudsagte sandsynligheder

Fra (fx) 20 til 70 år stiger den forudsagte sandsynlighed for at stemme personligt med ca. 20 procentpoint, hvilket svarer til vores AME, som vi gangede op til 50 år.

Desuden næsten en ret linje:

S-formen er ikke tydelig → mao. er linearitet ikke en problematisk funktionel form i dette tilfælde

Hvad ser vi ikke? Signifikanstest!



Interaktion i logistisk regression

Interaktion i en logit-model er "dobbelt-kompliceret" og kan være **problematisk**:

- Det er komplettest at beregne korrekte standardfejl for interaktioner i ikke-lineære modeller → Stata giver en **P-værdi**, men den er ikke til at stole på
- "På eget ansvar", men bestemt muligt

To praktiske muligheder

- (A) Hvis den modererende variabel (Z) er **binær** → ingen problemer med interaktion i logit
- (B) Hvis den modererende variabel (Z) er **kontinuert eller kategorisk** → lav interaktionsanalysen med OLS
 - tag forbehold og vær opmærksom på forudsætningsbrud
- Også andre muligheder ved kontinuit/kategorisk Z:
 - (C) find et alternativt, intervalskaleret Y og lav OLS-interaktion efter bogen
 - (D) fortsæt med interaktion i logit, men fortolk/præsentér kun marginale effekter (husk på at marginale effekter ikke fortæller om interaktionens signifikans)

Antagelser bag logistisk regression

Mange af de samme model-antagelser som i OLS, men lidt sværere at tjekke (se Sønderskov, 2014, pp. 274ff)

- Modellen er korrekt specifieret (ingen udeladt variabel bias etc.) - vigtigt!
- Fravær af alvorlig multikollinearitet (tjek med VIF i OLS-version af modellen)
- Fravær af meget indflydelsesrige observationer (ikke pensum men se evt. do-fil til øvelsesopgaven)
- Linearitet ("linear in the logits") → der skal være en lineær sammenhæng mellem Y udtrykt som log-odds og den uafhængige variabel X (tjek jf. Sønderskov (2014), pp. 274ff)
- Til gengæld antager vi *ikke* homoskedasticitet og normalfordelte fejlled

Det gælder i øvrigt også for **interaktion** (sidste uge) - at der overordnet gøres de samme antagelser som i OLS uden interaktion (se Sønderskov, 2014, p. 239)

To teknikaliteter

Maximum likelihood estimation

- Logit finder den bedste model (og dermed også "effektstørrelserne") med **maximum likelihood estimation (MLE)** i stedet for ordinary least squares (OLS)
- Ikke så substantielt vigtigt her og nu, men vigtigt at vide

Asymptotisk unbiased estimator

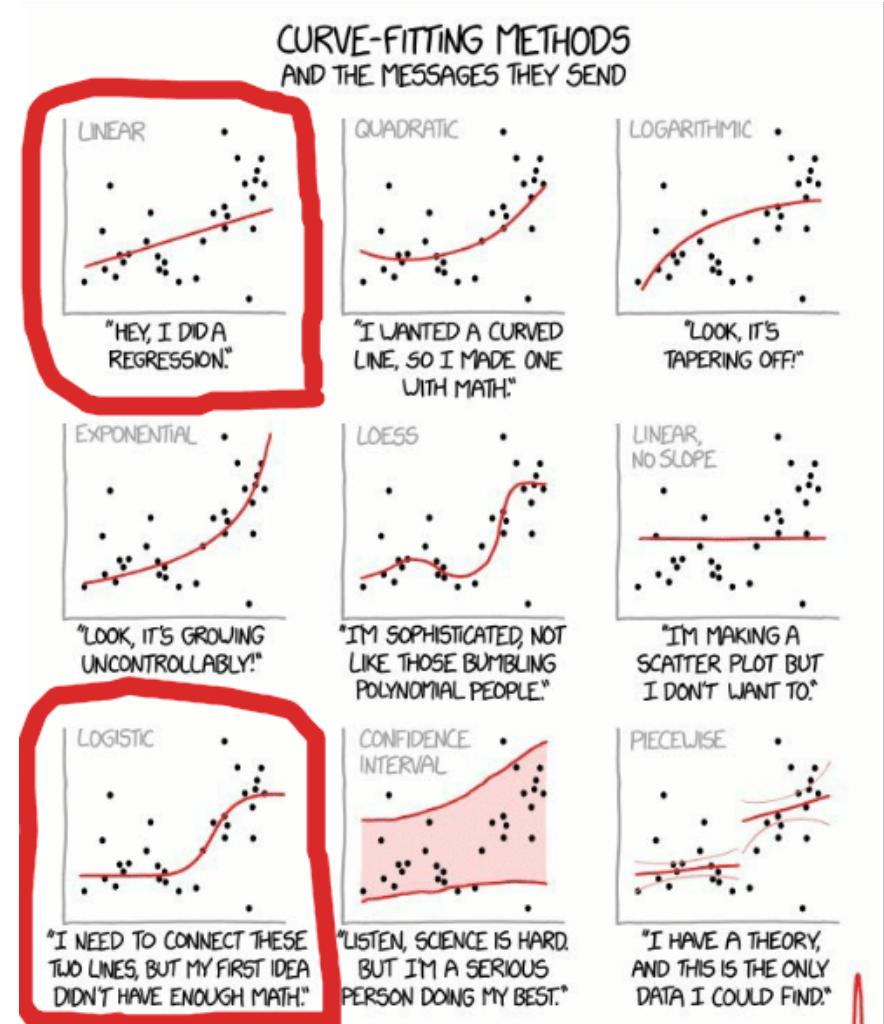
- Logistisk regression er **asymptotisk unbiased**, hvilket betyder, at logit først bliver unbiased i takt med, at N stiger
- Betyder større krav til **antal observationer** (N) end ved normal OLS

Et bud på guidelines ville være, at 100 observationer er for lidt, 500 observationer vil være tilstrækkeligt i langt de fleste tilfælde, og **200 er formentligt nok**, såfremt modellen ikke indeholder mere end 20 uafhængige variable

(Sønderskov, 2014, p. 279)

Dagens pointer

- Logistisk regression er et vigtigt redskab, når vi arbejder med **dikotome afhængige variable**, fordi modellen tager højde for mange af OLS' shortcomings
- Det kommer dog med den ulempe, at vores resultater bliver **langt mindre intuitive**
- I sidste ende laver vi typisk "korrekt" logistisk regression, men forsimpler informationen, når vi fortolker og formidler resultaterne
 - fx vha. AME, MEM, OLS



Næste gang

- Multilevel analyse
- Bruges når vi har uafhængige variable på flere niveauer
 - fx lande-BNP og individ-alder
- Helt nyt emne og ny teknik på pensum → I kan blive de første eksperter

Tak for i dag!