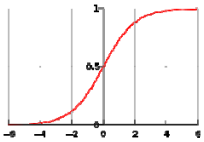
Multippel logistisk regresjon (Rosner 13.7)

Inger Johanne Bakken

Enhet for anvendt klinisk forskning, NTNU
og
Avdeling for forebyggende helsearbeid, SINTEF



Regresjonsmodeller

Hvorfor bruke regresjonsmodeller?

1. **Predikere et utfall** (sykdom, død, blodtrykk e.l.) basert på et sett av forklaringsvariabler (kovariater)
   - Lager statistiske modeller og vurderer modellens Goodness of fit
   - Inklusjon av variabler er basert på statistisk signifikans alene
2. **Måle en sammenheng** (assosiasjonen) mellom et utfall og en forklaringsvariabel (eksponering) og samtidig kontrollere for effekten av andre variable (konfundering og/eller interaksjon)
   - Vurdere om sammenhengen endres når man justerer for potensielle konfundere
   - Vurdere om sammenhengen er forskjellige innenfor ulike nivå av en annen faktor (interaksjon)
   - Inklusjon av andre variable ikke basert på statistisk signifikans alene

Prediksjon versus assosiasjon

- Forskjellig analysestrategi for prediksjon og assosiasjon
- Ofte følges en prediksjonsstrategi selv om hensikten med studien er å måle en/flere assosiasjoner
- "Stepwise regression"? Kan ikke brukes (ukritisk) i assosiasjonsstudier (kommer tilbake til dette).

Sitat Rosner (s 689)

"This is an extremely important technique used to control for one or more continuous or categorical covariates (independent variables) when the outcome (dependent) is binary."

Den binære avhengige variabelen er ofte Sykdom (Ja/nei)

Sammenheng sykdom (D) og eksponering (E)

Forutsetter at vi har uavhengige observasjoner:

- Ingen kovariater, eksponering som kategorisk variabel?
  - Kontigenstabeller og khikvadrat-test
- Eksponering som kategorisk variabel og én kategorisk kovariat?
  - Mantel-Haenszel test (ikke forelest)
- Kontinuerlig eksponering? Én kontinuerlig kovariat? Flere kovariater?
  - Logistisk regresjon

Modell i logistisk regresjon (1)

- Fra multippel lineær regresjon har vi:

$$y = \alpha + \sum_{j=1}^{k} \beta_j x_j + e$$

- Kan vi bruke tilsvarende modell i logistisk regresjon, hvor p er sannsynlighet for sykdom?

$$\mathrm{p} = \alpha + \sum_{j=1}^{k} \beta_j x_j + e$$

Modell i logistisk regresjon (2)

$$p = \alpha + \sum_{j=1}^{k} \beta_j x_j + e \quad ?$$

Ikke forenlig med begrepet ”sannsynlighet”.

Uttrykket kan ha verdier mindre enn null og større enn 1.

Trenger en annen modell.

Odds versus sannsynlighet

- **Sannsynlighet** for terningkast 1: p=1/6=0,17
- **Odds** for terningkast 1 er én til fem eller

  Odds = (1/6) / (5/6) = 1/5
- Odds er sannsynligheten for en hendelse dividert med sannsynlighet for at hendelsen ikke inntreffer

Odds og logistisk regresjon

Anta at Y kan ta verdiene 0 og 1.

La sannsynligheten for at Y=1 være p.

Da er sannsynligheten for at Y=0 gitt ved1-p.

Oddsen for at Y=1 gitt ved p/(1-p).

I logistisk regresjon modelleres ln(odds)=ln(p/1-p)

Modell i logistisk regresjon (3)

logit(p) = ln [p/(1-p)]

$$= \alpha + \sum_{j=1}^{k} \beta_j x_j + e$$

Medlem i familien generaliserte lineære modeller (lineær i logit)

Den logistiske funksjonen (1)

logit(p) = ln [p/(1-p)] = $\alpha + \beta_1 x_1 + \ldots + \beta_k x_k$

Da er p gitt av:

$$p = \frac{e^{\alpha+\beta_1 X_1+\ldots+\beta_k}}{1+e^{\alpha+\beta_1 X_1+\ldots+\beta_k}} = \frac{1}{1+e^{-(\alpha+\beta_1 X_1+\ldots+\beta_k)}}$$

Den logistiske funksjonen (2)

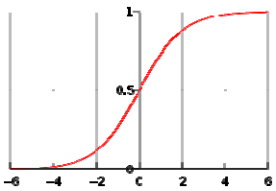
Den logistiske funksjonen er gitt av uttrykket

$$f(z) = \frac{1}{1+e^{-z}}$$

hvor z er gitt av de uavhengige variablene (X’ene)

Den logistiske funksjonen (3)

$$f(z) = \frac{1}{1+e^{-z}}$$

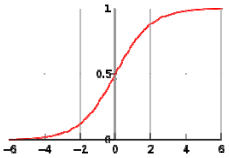


$z \to \infty \Rightarrow f(z) \to 1$ $\quad$ $z \to -\infty \Rightarrow f(z) \to 0$

Den logistiske funksjonen (4)

Logistisk regresjon er basert på den logistiske funksjonen. Brukes til å modellere sannsynligheten for at et individ skal ha en sykdom.

Aldri større enn 1 eller mindre enn 0.



S-form:
Ved lave verdier av z er sannsynligheten lav, ved en viss grense stiger sannsynligheten raskt.
Dette illustrerer hvordan man ofte tenker at eksponering kan påvirke sannsynlighet for sykdom.

Modell for sannsynlighet for sykdom:

$$p = \frac{e^{\alpha+\beta_1 X_1+\ldots+\beta_k}}{1+e^{\alpha+\beta_1 X_1+\ldots+\beta_k}} = \frac{1}{1+e^{-(\alpha+\beta_1 X_1+\ldots+\beta_k)}}$$

Parametrene må estimeres.

Ved MLR benyttes minste kvadraters metode.

Ved logistisk regresjon benyttes "Complex numeric algorithms" (Rosner s 670) (Maximum Likelihood Estimator)

Tolking av regresjonsparametrene (1)

To individer A og B.
Helt lik eksponering, med unntak av x_j.

For individ A er x_j=1 (eksponert).
For individ B er x_j=0 (ikke eksponert).

Da er logit(p_A) og logit(p_B) gitt av:

logit(p_A) = $\alpha + \beta_1 x_1 + \ldots + \beta_j{}^*(1) + \ldots + \beta_k x_k$
logit(p_B) = $\alpha + \beta_1 x_1 + \ldots + \beta_j{}^*(0) + \ldots + \beta_k x_k$

Tolking av regresjonsparametrene (2)

logit(p_A) = $\alpha + \beta_1 x_1 + \ldots + \beta_j{}^*(1) + \ldots + \beta_k x_k$
logit(p_B) = $\alpha + \beta_1 x_1 + \ldots + \beta_j{}^*(0) + \ldots + \beta_k x_k$

logit(p_A) - logit(p_B) = β_j

Dermed er $\ln[p_A/(1-p_A)] - \ln[p_B/(1-p_B)] = \beta_j$

$$\Downarrow$$

$$\ln \frac{p_A(1-p_A)}{p_B(1-p_B)} = \beta_j$$

Tolking av regresjonsparametrene (3)

$$\ln \frac{p_A(1-p_A)}{p_B(1-p_B)} = \beta_j$$

$$\Downarrow$$

$$\frac{p_A/(1-p_A)}{p_B/(1-p_B)} = e^{\beta_j}$$

$$\boxed{e^{\beta_j} = \frac{Odds_A}{Odds_B}}$$

Tolking av regresjonsparametrene (4)

$$e^{\beta_j} = \frac{Odds_A}{Odds_B}$$ = Odds ratio

Den estimerte regresjonskoeffisienten får dermed tolkingen:

$$\hat{O}R = e^{\hat{\beta}_j}$$

Odds ratio ved en enhets økning i x_j når alle de andre x'ene holdes konstant (justert for alle de andre x'ene).

Eksempel 1: Lungekreft og høyt alkoholforbruk (1)

- Tabell 13.3 Rosner

alkohol * cancer Crosstabulation

Count

		cancer		
		1	0	Total
alkohol	1	33	1667	1700
	0	27	2273	2300
	Total	60	3940	4000

Alkohol=1 står for høyt forbruk

- Estimat for Odds ratio: (a*b)/(c*d) = (33 * 2273) / (27 * 1667) =1,67

Eksempel 1: Lungekreft og alkoholforbruk (2)

- Fra logistisk regresjon (Analyze - > Regression -> Bivariate logistic) får vi:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1	alkohol	.511	.261	3.815	1	.051	1.667
	Constant	-4.433	.194	524.366	1	.000	.012

- Kan vi konkludere at høyt alkoholforbruk fører til lungekreft?

Eksempel 1: Lungekreft og alkoholforbruk (3)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1	alkohol	.511	.261	3.815	1	.051	1.667
	Constant	-4.433	.194	524.366	1	.000	.012

- Dette er **CRUDE = UJUSTERT** odds ratio

Eksempel 1: Lungekreft og alkoholforbruk (4)

- Ser nå på forekomsten av lungekreft i de to strataene røyk/ikke-røyk og beregner to estimater for Odds ratio = ab/cd:

Count

			cancer		
røyk			1	0	Total
0	alkohol	1	9	891	900
		0	21	2079	2100
		Total	30	2970	3000
1	alkohol	1	24	776	800
		0	6	194	200
		Total	30	970	1000

Risk Estimate

			95% Confidence Interval	
røyk		Value	Lower	Upper
0	Odds Ratio for cancer (0 / 1)	1.000	.456	2.192
1	Odds Ratio for cancer (0 / 1)	1.000	.403	2.480

Eksempel 1: Lungekreft og alkoholforbruk (5)

- Altså *ingen* sammenheng i hvert stratum separat.
- Tilsvarende i logistisk regresjon, justerer for røyk ved å legge variabelen inn i modellen, og får:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95,0% C.I.for EXP(B)	
								Lower	Upper
Step 1	alkohol	.000	.303	.000	1	1.000	1.000	.552	1.811
	røyk	1.119	.302	13.763	1	.000	3.062	1.695	5.530
	Constant	-4.595	.205	503.554	1	.000	.010		

$$\hat{O}R_{alk} = Exp(B_{alk}) = 1$$

Eksempel 2
Røyking og hjertekardød (1)

- Ønsker å studere sammenhengen mellom røyking og risiko for å dø av hjertekarsykdom basert på data fra HUNT 1 (1984-86)[1]
- 1502 døde (case), selektert 2 kontroller per case
- Data i 2X2 tabell:

Count

		død		
		1	0	Total
røyk	1	321	832	1153
	0	1181	2172	3353
	Total	1502	3004	4506

- Odds for å være case blant røykere: a/b=321/832=0,39
- Odds for å være case blant ikke-røykere: c/d=1181/2172=0,54
- OR=0,39/0,54=0,7
- Konklusjon: "Røykere har lavere risiko for å dø av hjertekarsykdom sammenlignet med ikke-røykere"

[1] Fra Tom Ivar Lunds forelesning 2006

Eksempel 2:
Vurdering av resultatet (1)

- "Røykere har lavere risiko for å dø av hjertekarsykdom sammenlignet med ikke-røykere"
- "Det er en negativ assosiasjon mellom røyking og risiko for hjertekardød"
- Er dette et pålitelig resultat?
- Ser på konfidensintervall og/eller p-verdier (er sammenhengen statistisk signifikant?)

Eksempel 2:
Vurdering av resultatet (2)

- Som i hypotesetesting ellers: p-verdien er sannsynligheten for å få det observerte resultatet (OR=0,7) gitt at det egentlig ikke er noen sammenheng (OR=1,0)
- Her er p<0,001
- Kan vi dermed konkludere at sammenhengen er pålitelig?

Konfundering

- På samme måte som i det forrige eksempelet må vi vurdere effekten av eventuelle konfundere
- Er resultatet påvirket av konfundering med andre variable?
- Dette er variabler som er forbundet både med røyking og risiko for hjertekardød:
  - alder
  - kjønn
  - fysisk aktivitet
  - kosthold
  - osv

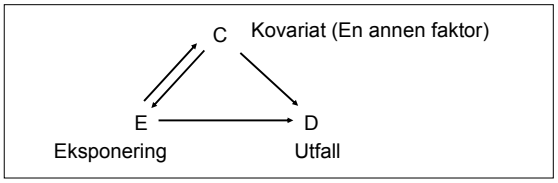
To confound

1. To cause to become confused or perplexed
2. To fail to distinguish; mix up: *confound fiction and fact*
3. To make (something bad) worse
4. To cause to be ashamed
5. To damn
6. To frustrate
7. *Archaic* To bring to ruination

"Konfundering" og "konfunder" finnes ikke i norske (standard) ordbøker, men brukes i statistisk og epidemiologisk litteratur

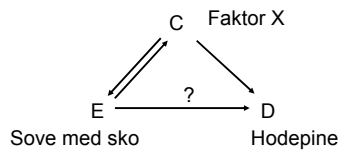
Konfundering

"… en utenforstående tredjefaktor som hører med blant årsakene til sykdommen og ikke påvirkes *årsaksmessig* av den eksponeringsfaktoren vi studerer, men som likevel står i et bestemt forhold til den"
(Geir Jacobsen, Tidsskr Nor Legeforen 27, 2002; 122)



Finn konfunderen[1]

"Jeg er sannsynligvis allergisk mot lær fordi hver gang jeg har sovnet med skoene på, så våkner jeg med dundrende hodepine morgenen etter" (Marve Fleksnes)



[1]Lånt av P Romundstad, ISM

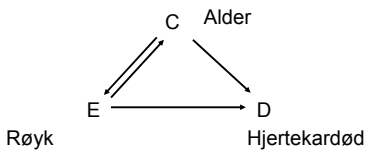
Spuriøse sammenhenger

Vi finner en sammenheng som i virkeligheten kan forklares i sin helhet av en tredje variabel (en konfunder).

Ved hjelp av statistiske analyser fant at sammenhengen mellom alkohol og lungekreft i eksempel 1 var spuriøs.

Kan ikke kjøre logistisk regresjon med tilfellet Fleksnes selv om vi hadde hatt flere observasjoner. En forutsetning for logistisk regresjon er jo at vi har uavhengige observasjoner…

Konfundering i eksempel 2



Alder

Økende alder gir økt risiko for hjertekardød
Røyking er ulikt fordelt i aldersgruppene

Alder fyller dermed begge kriteriene for konfundering

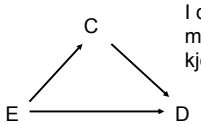
Å justere for konfundering

Informasjon om potensielle konfundere? Kan justere for effekten med statistiske analyser (i all hovedsak regresjonsanalyser)

I eksempel 1 (alkohol og lungekreft) fant vi at assosiasjonen *ikke* var signifikant i de to strataene røyk/ikke-røyk. Dette svarte til at alkohol ikke var signifikant i logistisk regresjon etter justering for røyk.

I eksempel 2 (røyk og hjertekardød) skal vi se om røyk "beskytter mot hjertekardød" etter justering for alder (vær generelt forsiktig med uttrykket "beskytter mot" og lignende, bruk heller uttrykk som inneholder "assosiert med")

Konfunder eller mellomliggende faktor



I dette tilfellet er C er en mellomliggende faktor (er i årsaks-kjeden fra E til D)

Eksempel:
Assosiasjon fedme og hjertekardød, justere for hypertensjon i analysen?

Eksempel 2: Logistisk regresjon (crude OR)

Først kjører vi analysen med bare røykevariabelen:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95,0% C.I.for EXP(B) Lower	Upper
Step 1	røyk	-.343	.075	20.931	1	.000	.710	.613	.822
	Constant	-.609	.036	284.001	1	.000	.544		

Samme resultat for estimatet for OR (= Exp(B)) som da vi regnet for hånd

Eksempel 2:
Logistisk regresjon med justering for alder

Legger til en variabel til i modellen: Alder (som kontinuerlig variabel, forutsetter linearitet i logit)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95,0% C.I.for EXP(B) Lower	Upper
Step 1	røyk	.723	.104	48.474	1	.000	2.061	1.681	2.526
	alder	.147	.005	1031.024	1	.000	1.158	1.148	1.169
	Constant	-11.474	.386	882.044	1	.000	.000		

Dramatisk endring i estimatet for OR etter justering for alder!

Fra ”beskyttende faktor” til ”risikofaktor”
(fra negativ til positiv assosiasjon)

Fortolkning av resultatene

Den dramatiske endringen i estimatet for odds ratio viser viktigheten av å **justere for potensielle konfundere**

Lave p-verdier og smale konfidensintervall er (i beste fall) verdiløse hvis modellen er feil.

Kan ikke justere for ukjente konfundere eller konfundere vi ikke har informasjon om. Informasjonen må også være tilstrekkelig detaljert:

Eksempel 2:
Detaljert informasjon

Hva hvis vi bare hadde informasjon om alder i to kategorier, over og under 50 år?

Analysen med en slik aldersvariabel gir oss:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95,0% C.I.for EXP(B) Lower	Upper
Step 1	røyk	.723	.104	48.474	1	.000	2.061	1.681	2.526
	alder	.147	.005	1031.024	1	.000	1.158	1.148	1.169
	Constant	-11.474	.386	882.044	1	.000	.000		

Vi har gått fra
negativ til *positiv* til *ingen* sammenheng
mellom røyk og hjertekardød…

”There are three kinds of lies: lies, damned lies, and statistics"

Er vi enige med Mark Twain (eller var det kanskje Benjamin Disraeli)?

Helt avgjørende for en pålitelig analyse og presentasjon:

Valg av modell
Tolkning av resultater i sammenheng (sunn fornuft, klinisk skjønn, tidligere funn)

Det holder ikke med lave p-verdier og smale konfidensintervaller!

p-verdier alene har liten verdi

En lav p-verdi alene ser ingenting om effektstørrelsen/klinisk signifikans:

Stort utvalg: kan få lav p-verdi selv om effekten er liten
Stor effekt: kan få lav p-verdi selv om utvalget er lite

Eksempel:
- 1500 kasus, 3000 kontroller: OR=0,7, p<0,001
- 150 kasus, 300 kontroller: OR=0,2, p<0,001
- 3000 kasus, 6000 kontroller: OR=0,9, p<0,001

Rapporter estimatet for OR med tilhørende konfidensintervall.

Fortolkning av estimatene (2)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95,0% C.I.for EXP(B) Lower	Upper
Step 1	røyk	.723	.104	48.474	1	.000	2.061	1.681	2.526
	alder	.147	.005	1031.024	1	.000	1.158	1.148	1.169
	Constant	-11.474	.386	882.044	1	.000	.000		

- Vi har sett at estimatet for odds ratio for dikotom variabel tolkes som odds ratio = odds(eksponert) / odds(ikke eksponert)
- Hvordan tolke koeffisienten for en kontinuerlig variabel?
- Her har vi at estimatet for odds ratio for alder er 1,158
  - Hva betyr dette?
  - Kan vi estimere odds ratio for en økning i alder på 5 år fra denne modellen (slik vi gjorde i MLR)?

Endring i estimatet for odds ratio med 5-års økning i alder?

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95,0% C.I.for EXP(B) Lower	Upper
Step 1	røyk	.723	.104	48.474	1	.000	2.061	1.681	2.526
	alder	.147	.005	1031.024	1	.000	1.158	1.148	1.169
	Constant	-11.474	.386	882.044	1	.000	.000		

- Det er Logit som er lineær!
- Økning i alder med 5 år svarer til en endring i B-verdien til 5 * 0,147 = 0,737
- Estimatet for OR blir da Exp(0,737) = 2,1

Simpson's paradox

- Den observerte sammenhengen er den motsatte av den sanne sammenhengen
- Eks "Røyking beskytter mot hjertekardød" i vårt eksempel 2

Eks: The LBW Paradox: Barn med lav fødselsvekt av røykende mødre har bedre overlevelse enn barn med tilsvarende fødselsvekt født av mødre som ikke røyker

Allen Wilcox: "On the importance – and the unimportance – of birthweight" Int J Epidemiol 2001; 30:1233 - 1241

- "Røyking øker overlevelsen" (Yerushalmy 1960-årene)
- "Røyking skifter fødselsvektkurven til venstre" (MacMahon 1966)
- "Røyking skifter fødselsvektkurven til venstre, økt dødelighet ved alle fødselsvekter" (Wilcox 1993)

Gangen i en analyse: Hvilke variable skal/bør inn i modellen?

- Det finnes ingen klar "oppskrift"
- Mange potensielle forklaringsvariable? Velg ut de mest aktuelle fra univariable analyser og antatt klinisk relevans. Mildt kriterium for utvelgelse: p < 0,25
- Tilpass modellen ved å ta inn og ut variable etter "standard" kriterier
- Sjekk funksjonell form (linearitet osv)
- Vurder interaksjonsledd. Interaksjonsledd bør ha en plausible forklaring. Kun statistisk signifikante interaksjonsledd inkluderes
- Sjekk goodness-of-fit

(Se Hosmer & Lemeshow)

Automatiske prosedyrer for seleksjon av variable

- Forward stepwise
- Backward stepwise
- (Best subset selection)

- "have been criticized because they can yield a biologically implausible model … The problem is not the fact that the computer can select such models but rather that the analyst fails to scrutinize the resulting model carefully … The analyst, not the computer, is ultimately responsible for the review and evaluation of the model"

(Hosmer and Lemeshow, Applied logistic regression, second edition, p 96)

Goodness-of-fit (1)

I MLR startet vi analysene med QQ plott og brukte residualene til å se etter mønster og uteliggende observasjoner

I logistisk regresjon kan Pearson residualer brukes til å se etter uteliggende observasjoner:

$$\boxed{r_i = \frac{y_i - \hat{p}_i}{se(\hat{p}_i)}}$$

y_i er observert verdi, 1 eller 0

$$\hat{p}_i = \frac{e^{\hat{L}_i}}{1 - e^{\hat{L}_i}} \qquad \hat{L}_i = \hat{\alpha} + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k$$

$$se(\hat{p}_i) = \sqrt{\hat{p}_i(1 - \hat{p}_i)}$$

Goodness-of-fit (2)

- I SPSS kan vi for hver enkelt observasjon be om lagring av:

- PRE: Predikert verdi
- PGR: Predikert gruppe
- COO: Cook's distance (endring i regresjonskoeffisienter hvis et case slettes)
- Lev: Leverage (mål for innflytelse på modellen)
- Res: Residual (som på forrige side)
- LRE: Logit av residualet
- SRE: Studentisert residual
- ZRE: Standardisert residual
- DEV: Deviance (disagreement between the maxima of the observed and the fitted log likelihood functions)
- DFB: DFBeta for hver koeffisient inkludert konstanten

Goodness-of-fit (tommelfingerregler)

- COO:
  - Bør være mindre enn 1
- Lev:
  - Ligger mellom 0 (ingen påvirkning) og 1 (fullstendig påvirkning).
  - Forventet Lev er (k+1)/N (k, antall parametre og N antall observasjoner)
- SRE, ZRE, DEV
  - Minst 95 % skal ligge mellom ± 1,96 og minst 99 % skal ligge mellom ± 2,58
- DFB:
  - Bør være mindre enn 1

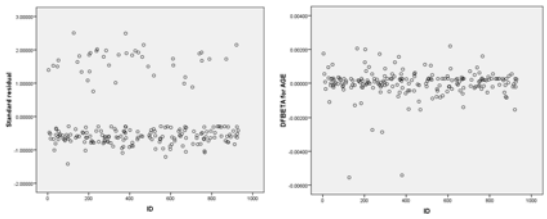
Eksempel fra Hosmer og Lemeshow

Datasett med informasjon om overlevelse etter sykehusinnleggelse (ICU)

- STA: skrevet ut som død (1) eller levende (0)
- Age: alder
- CAN: cancer (1/0)
- INF: infeksjon ved innleggelse (1/0)
- RACE: rase (hvit/svart/annen)
- CPE: CPR ved innleggelse

Av 200 pasienter ble 40 ”utskrevet som død”.
Modellerer først en univariabelt, STA og AGE.
I MLR (kontinuerlig Y) ville vi begynt med et scatterplott mellom Y og AGE

Gjennomgang av SPSS Output

- Datasettet finnes på
  - http://www.umass.edu/statdata/statdata/
  - (Klikk på ”Logistic regression”, datasettet heter ICU)
  - Lagre som Excel på egen maskin og spesifiser at det er en Excel fil når du åpner den via SPSS

Plot av lagrede verdier mot ID, eksempler