SOS3003

Anvendt statistisk dataanalyse i samfunnsvitenskap

Forelesingsnotat 13

Erling Berge
Institutt for sosiologi og statsvitenskap
NTNU

Om manglande data

- Manglande data på **Y** skaper **problem** for analysen
- Manglande data på ein eller fleire x-variablar skaper **ikkje problem** i seg sjølv, men **kan bli eit problem** dersom
  1. Utelating fører til for få case i analysen
  2. Mangel på merksemd fører til eit variabelt tal av case i analysane
- Listevis utelating er beste metoden der problema er små

Manglande data for Y

- MCAR - missing completely at random

  tyder at sjansen for at det skal mangle data på Y ikkje er påverka verken av kva verdi y har eller kva verdi x har
- MAR – missing at random

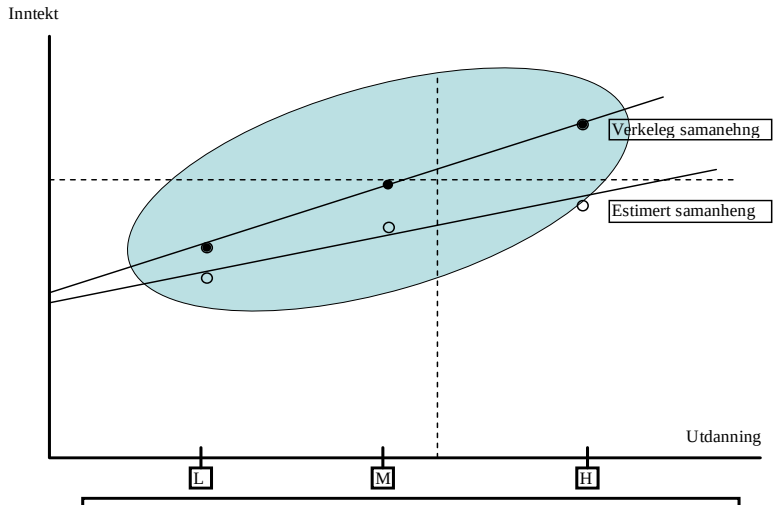
  tyder at sjansen for at det skal mangle data på Y ikkje er påverka av kva verdi x har

Ved MAR kan case fjernast dersom …

- Dersom manglande opplysningar på Y er MCAR eller MAR kan case fjernast dersom det er igjen nok case til å gjennomføre analysane
- Dersom listevis utelating gir for få case kan ein bruke ”Multiple imputation” eller Maximum Likelihood estimering tilpassa manglande data

Når manglande data ikkje er MAR

- Allment kan det seiast at det **ikkje er MAR** data dersom dei som manglar y-verdi eigentleg har systematisk høgare eller lågare verdi enn dei som har ein observert verdi etter kontroll for X
- Dersom manglande opplysningar på Y **ikkje** er **MAR** kan ein ikkje fjerne case. Ein må i staden inkludere i analysen ein eigen modell av den mekanismen som fører til at data manglar



Kjelde: Winship, Christopher, and Robert D. Mare 1992 «Models for sample selection bias», Annual Review of Sociology, 18:327-350

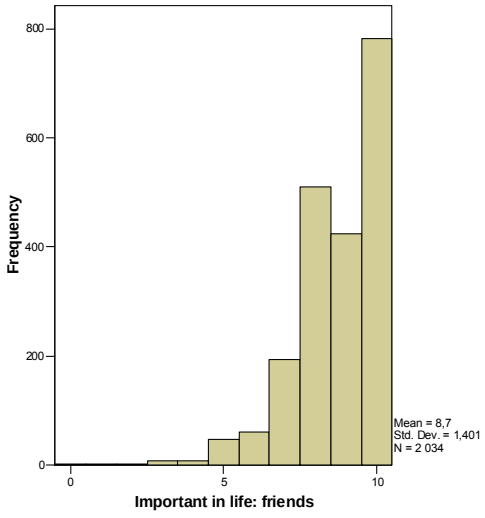
Regresjon og måleskala

- Regresjon krev intervall eller forholdstalskala
- Nominalskala kan brukast ved dummykoding
- Ordinalskala kan brukast ved dummykoding eller ved argument om at det eigentleg er ein underliggande intervallskala. Argumentet er da at avstandane ar dei same mellom alle kategoriane
- Dette argumentet er ikkje gyldig når ein vanlege intervallskala av tabelltekniske årsaker eg gjort om til nokre få rangerte grupper. T.d. inntekt:

Ordinalskala I

Important in life: friends

		Frequency
Valid	Extremely unimportant	2
	1	2
	2	2
	3	7
	4	7
	5	47
	6	60
	7	193
	8	509
	9	423
	Extremely important	782
	Total	2034
Missing	Don't know	2
Total		2036



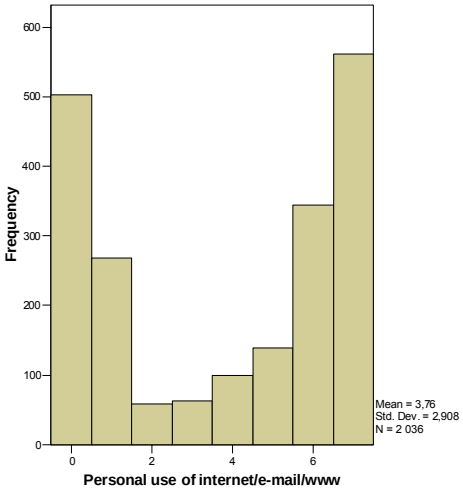
Ordinalskala II

Ikkje-lineær måleskala

Personal use of internet/e-mail/www

		Frequency
Valid	No access at home or work	503
	Never use	268
	Less than once a month	58
	Once a month	63
	Several times a week	100
	Once a week	139
	Several times a week	344
	Every day	561
	Total	2036

Frekvens pr tidseining varier frå gruppe til gruppe



Inntektsdata i ESS

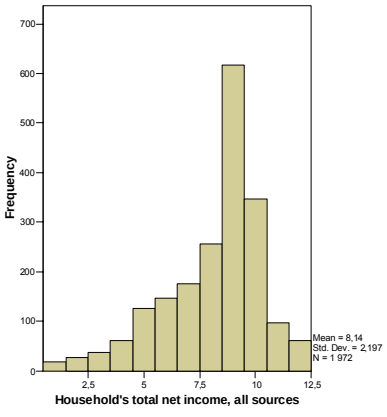
HH income class	**Approximate WEEKLY**	**Approximate MONTHLY**	**Approximate ANNUAL**	**HH income NOK per year** 1€ = 8,50	
J	Less than €40	Less than €150	Less than €1800	**J**	<=15300
R	€40 to under €70	€150 to under €300	€1800 to under €3600	**R**	15300-
C	€70 to under €120	€300 to under €500	€3600 to under €6000	**C**	30600-
M	€120 to under €230	€500 to under €1000	€6000 to under €12000	**M**	51000-
F	€230 to under €350	€1000 to under €1500	€12000 to under €18000	**F**	102000-
S	€350 to under €460	€1500 to under €2000	€18000 to under €24000	**S**	153000-
K	€460 to under €580	€2000 to under €2500	€24000 to under €30000	**K**	204000-
P	€580 to under €690	€2500 to under €3000	€30000 to under €36000	**P**	255000-
D	€690 to under €1150	€3000 to under €5000	€36000 to under €60000	**D**	306000-
H	€1150 to under €1730	€5000 to under €7500	€60000 to under €90000	**H**	510000-
U	€1730 to under €2310	€7500 to under €10000	€90000 to under €120000	**U**	765000-
N	€2310 or more	€10000 or more	€120000 or more	**N**	1020000+

Set inn eit kronebeløp som kode

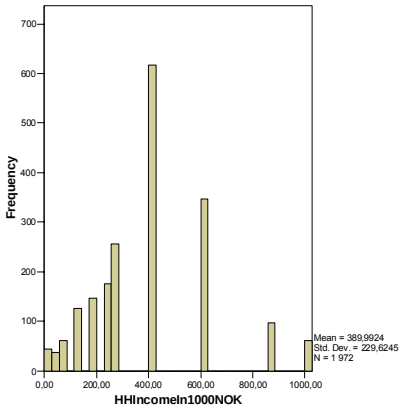
HH income NOK per year	**1€ = 8,50**	**SPSS code**	**A reasonable code of income in 1000 NOK might be**
J	<=15300	1	15
R	15300-	2	25
C	30600-	3	45
M	51000-	4	80
F	102000-	5	130
S	153000-	6	180
K	204000-	7	230
P	255000-	8	280
D	306000-	9	390
H	510000-	10	600
U	765000-	11	800
N	1020000+	12	1020

Resultat av omkoding

Før omkoding



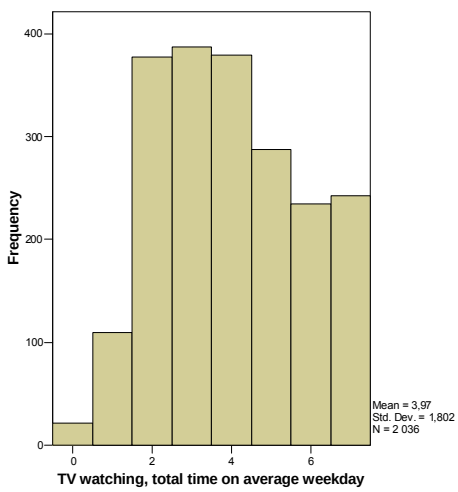
Etter omkoding



TV watching

TV watching, total time on average weekda

		Frequency
Valid	No time at all	21
	Less than 0,5 hour	109
	0,5 hour to 1 hour	377
	More than 1 hour, up to 1,5 hours	387
	More than 1,5 hours, up to 2 hours	379
	More than 2 hours, up to 2,5 hours	287
	More than 2,5 hours, up to 3 hours	234
	More than 3 hours	242
	Total	2036



Dikotomisering og dummykoding (1)

Spørsmål: *hva er forskjellen på dikotomisering og dummykoding?*

Svar: Dikotomisering av ein variabel tyder å dele skalaen i 2 deler (eventuelt samle kategoriane i 2 grupper) slik at variabelen berre får 2 verdiar (høg/ lav, mye/ lite, pluss/ minus, etc.). Alle variablar kan dikotomiserast. Når ein gjer det vil ein sjølvsagt tape mye informasjon. Håpet er at det ein taper ikkje er vesentleg i høve til det ein drøftar.

Dikotomisering og dummykoding (2)

- Dummykoding er ein spesiell måte å kode/ lage hjelpevariable på. For kategoriar i nominalskalavariable eller intervall i ordinal- eller intervallskalavariable lagar ein hjelpevariable som får verdien 1 dersom caset har ein verdi lik kategoriverdien eller ein verdi som ligg i intervallet på ordinal- eller intervallskalaen og 0 dersom det ikkje har ein verdi som spesifisert
- Ved hjelp av dummykoding av kvar kategori på ein nominalskalavariabel treng ein ikkje miste noko av informasjonen i variabelen

Dummykoding har mange namn

SOS3003:	**Dummy**	**Effect**	**Contrast**	
Hamilton	Dummy	Effect		
Hardy	Dummy	Effect	Contrast	
Menard/ **SPSS**	**Indicator**/ Simple	Deviation		
Weisberg	Dummy/ Indicator			
Hosmer& Lemeshow	Reference Partial	Deviation Marginal		

SPSS har mange måtar å kode (1)

- **Indicator**
  - Contrasts indicate the presence or absence of category membership. The reference category is represented in the contrast matrix as a row of zeros.
- **Simple**
  - Each category of the predictor variable (except the reference category) is compared to the reference category.
- **Difference**
  - Each category of the predictor variable except the first category is compared to the average effect of previous categories. Also known as **reverse Helmert contrasts**.

SPSS har mange måtar å kode (2)

- **Helmert**
  - Each category of the predictor variable except the last category is compared to the average effect of subsequent categories.
- **Repeated**
  - Each category of the predictor variable except the first category is compared to the category that precedes it.
- **Deviation**
  - Each category of the predictor variable except the reference category is compared to the overall effect.
- **Polynomial**
  - Orthogonal polynomial contrasts. Categories are assumed to be equally spaced. Polynomial contrasts are available for numeric variables only.

Effekten av måleskala …

- Jeg holder på å teste ulike variabler i logistisk regresjon. De fleste er blitt signifikantet etter at jeg har lagt til samspillsvariabler og kvadratledd, problemet er at på en av variablene blir exp(B) veldig stor, over 15000 på det meste.
- Dette kan jo umulig stemme, men hvordan kan jeg gjøre noe med det. Den blir noe lavere når jeg fjerner kvadratleddet til variabelen, men kvadratleddet er sig så jeg må vel i grunnen ha det med??

Måleskalaen er viktig

- for tolkinga av resultatet
- for storleiken til koeffisientar

- Multiplikasjon med 1000 eller divisjon med 1000 kan gi dramatiske utslag i storleiken på estimerte parametrar

- Eksempel frå logistisk regresjon:

Skalering med 1/1000

	B	S.E.	Wald	df	Sig.	Exp(B)
Educ	-,168	,080	4,414	1	,036	,845
Lived	-,052	,013	14,619	1	,000	,950
Constant	2,833	1,166	5,907	1	,015	17,004

	B	S.E.	Wald	df	Sig.	Exp(B)
Educ1000del	-167,984	79,955	4,414	1	,036	,000
Lived1000del	-51,615	13,500	14,619	1	,000	,000
Constant	2,833	1,166	5,907	1	,015	17,004

Bruk av kvadratledd

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	educ	,993	,614	2,616	1	,106	2,700
	EducSQ	-,044	,023	3,558	1	,059	,957
	lived	-,073	,036	4,200	1	,040	,930
	LivedSQ	,000	,001	,459	1	,498	1,000
	Constant	-4,428	3,993	1,230	1	,268	,012

a. Variable(s) entered on step 1: educ, EducSQ, lived, LivedSQ.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Educ1000del	993,076	613,943	2,616	1	,106	.
	Educ1000delSQ	-44048,2	23351,521	3,558	1	,059	,000
	Lived1000del	-73,092	35,665	4,200	1	,040	,000
	Lived1000delSQ	427,124	630,466	,459	1	,498	3,14+185
	Constant	-4,428	3,993	1,230	1	,268	,012

a. Variable(s) entered on step 1: Educ1000del, Educ1000delSQ, Lived1000del, Lived1000delSQ.

Hva betyr "Stokastisk variasjon"?

- Stokastisk refererer til sannsyn
- I uttrykket ”stokastisk variasjon” tenkjer ein seg variasjon t.d. i form av avvik frå eit gjennomsnitt som det kan knytast visse sannsyn til: t.d. til større avvik til mindre sannsyn
- I uttrykket ”stokastisk variabel” vil det kunne knytast sannsyn til ulike variabelverdiar.

Hva som menes med "null-hypotese" og at denne forkastes?

- Nullhypotese er ein påstand vi gjer om ein parameter for å kunne teste om det er ein rimeleg korrekt påstand
- Hvis nullhypotesa leier til ein urimeleg konklusjon (t.d. stor t-verdi) forkastar vi den som urimeleg
- Uttrykket kan kanskje ha samanheng med at vi vanlegvis tar utgangspunkt i at parameteren er 0, men generelt vil det i dag bli brukt om
- ”Et utsagn hvis feilaktige benektelse vi fortrinnsvis ønsker å unngå” (Sverdrup 1964:175 ”Lov og tilfeldighet I”)
- Synspunktet er at det er best å gjere feil på rett side, t.d. i kvalitetskontroll av fly, bilar, klatretau, … , osv

1. Skal man KUN se på signifikansen når man vurderer riktigheten av hypoteser?

- Ved sida av signifikansnivå må ein alltid vurdere tolkinga substansielt. Dersom det vi finn strir mot sunn fornuft, eller det andre forskarar har funne bør vi sjekke etter kva vi har gjort ein gong til …
- Det gjeld sjølvsagt både for signifikante og usignifikante variablar

Multikollinearitet og autokorrelasjon. Korrelasjon

- Korrelasjon dreiar seg om samvariasjon mellom 2 variablar (kollinearitet)
- Autokorrelasjon dreiar seg om tilhøvet mellom verdiar på ulike case på same variabel (t.d. prosent yrkesaktive fiskarar i kvar kommune). Geografisk nærliggjande case har ofte verdiar som ligg nærmare kvarandre enn to tilfeldig valde kommunar
- Multikollinearitet dreiar seg om korrelasjon mellom fleire variable (t.d. ”tilltit til Stortinget” og ”tillit til politikarar”). Variablar som måler om lag same fenomenet **skal** korrelere med kvarandre og er til ein viss grad gjensidig overflødige
- Teknisk ?

Multikollinearitet

- svært høge korrelasjonar mellom x-variablar
- sjekk korrelasjonar mellom parameterestimat
- sjekk om toleransen (den delen av variasjonen i x som ikkje er felles med andre variablar) er mindre enn t.d. 0,1
- VIF= variansinflasjonsfaktor= 1/toleranse
- dersom multikollinearitet skuldast kvadrering av variablar eller interaksjonsledd er det ikkje problematisk

Toleranse

- Mengda av variasjon i ein variabel x_k som er unik for variabelen vert kalla toleransen til variabelen
- La R^2_k vere determinasjonskoeffisienten i regresjonen av x_k på dei andre x-variablane. Dei andre x-variablane forklarer andelen R^2_k av variasjonen i x_k.
- Da er 1- R^2_k den unike variasjonen, dvs. Toleransen = 1- R^2_k
- Ved perfekt multikollinearitet vil R^2_k = 1 og toleransen = 0
- Låge verdiar av toleransen gjer regresjonsresultata mindre presise (større standardfeil)

VariansInflasjonsFaktoren (VIF)

- standardfeilen til regresjonskoeffisienten b_k kan skrivast

$$SE_{b_k} = \frac{s_e}{\sqrt{RSS_k}} = \frac{s_e}{\sqrt{\left(1 - R_k^2\right) TSS_k}} = \sqrt{VIF}\,\frac{s_e}{\sqrt{TSS_k}}$$

- Her er 1/toleransen = $1/(1\text{-}R^2_k)$ = VIF
- Om alt anna er likt vil lågare toleranse (større VIF) hos x_k gi høgare standardfeil for b_k [den aukar med ein faktor lik kvadratrota av (VIF)]

Indikatorar på multikollinearitet

- Beste indikatoren er toleransen eller VIF (denne er basert på R^2_k)
- Andre indikatorar er
  - Korrelasjon mellom einskildvariable (upåliteleg)
  - Inklusjon / eksklusjon av einskildvariablar gir store endringar i effektane til andre variablar
  - Uventa forteikn til effekten av ein variabel
  - Standardiserte regresjonskoeffisientar større enn 1 eller mindre enn -1
  - Korrelasjon mellom parameterestimat

Kva er for låg toleranse?

Når R^2_k > 0,9 er toleransen < 0,1 og VIF > 10

Multiplikatoren for standardfeilen er da kvadratrota av VIF (ca 3.2)

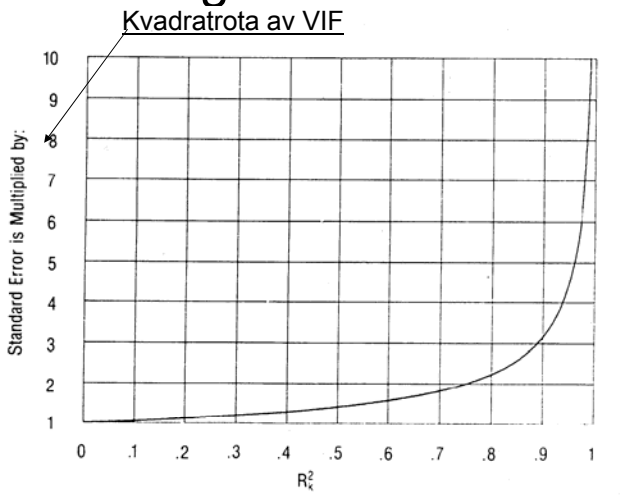


Figure 4.15 Effect of multicollinearity on standard errors (simplified).

Når er multikollinearitet eit problem?

- Det er ikkje eit problem dersom årsaka er kurvelinearitet eller interaksjonsledd i modellen. Men vi må i testinga ta omsyn til at parameterestimat for variablar med høg VIF er upresise. Vi testar dei som gruppe med F-testen
- Når det skuldast at to variablar måler same omgrep kan den eine droppast eller dei kan kombinerast til ein indeks
- Det er eit problem dersom vi treng estimat av variablane sine separate effektar (når kunnskap om deira samla effekt ikkje er nok)

Case med stor verknad på resultatet

- Eit case (eller ein observasjon) har påverknad dersom regresjonsresultatet endrar seg når caset blir utelate
- Somme case har uvanleg stor påverknad på grunn av
  - Uvanleg stor y-verdi (utliggjar)
  - Uvanleg stor verdi på ein x-variabel
  - Uvanlege kombinasjonar av variabelverdiar

Uvanlege Y-verdiar: utliggjarar

Casewise Diagnostics[a]

Case Number	Std. Residual	SYSTRUST Trust in system, mean of b7-b10	Predicted Value	Residual
11883	-3.046	.00	5.2601	-5.2601
28014	-3.125	.75	6.1475	-5.3975
28963	-3.303	.00	5.7047	-5.7047
29216	-3.190	.00	5.5097	-5.5097
29438	-3.029	.75	5.9811	-5.2311
29651	3.280	9.25	3.5858	5.6642
29781	3.657	10.00	3.6848	6.3152
30854	3.313	9.50	3.7784	5.7216

a. Dependent Variable: SYSTRUST Trust in system, mean of b7-b10

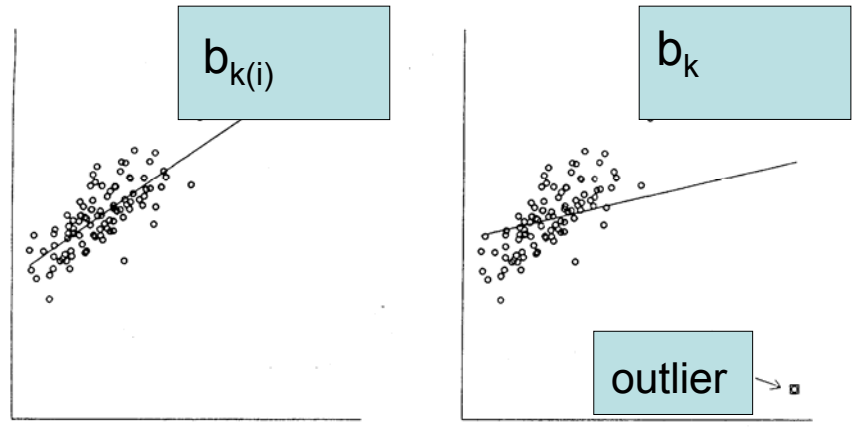
"casewise diagnostics" tabellen

- Bestiller vi "casewise diagnostics" får vi ein tabell med variabelverdiar for dei case som møter kriteriet som er gitt ("outliers outside n standard deviations").
- Default-verdien av n er 3.
- Variablane som er skrive ut er y (avhengig variabel), predikert y, residual, og standardisert residual (seleksjonsvariabel)
- Tabellen vil seie oss om utliggjarar er eit stort (mange) eller lite problem (få)

Stor verknad på einskildkoeffisientar

- Vi ser om eit case har påverknad ved å samanlikne regresjonar med og utan eit bestemt case. Ein kan t.d.
- Sjå på skilnaden mellom b_k og $b_{k(i)}$ der case nr i er utelate i estimeringa av den siste koeffisienten.
- Denne skilnaden målt relativt til standardfeilen til $b_{k(i)}$ vert kalla $DFBETAS_{ik}$

$DFBETAS_{ik}$:



One case may make a lot of difference

$DFBETAS_{ik}$

$$DFBETAS_{ik} = \frac{b_k - b_{k(i)}}{\frac{s_{e(i)}}{\sqrt{RSS_k}}}$$

$s_{e(i)}$ er residualen sitt standardavvik når case nr i er utelate frå regresjonen

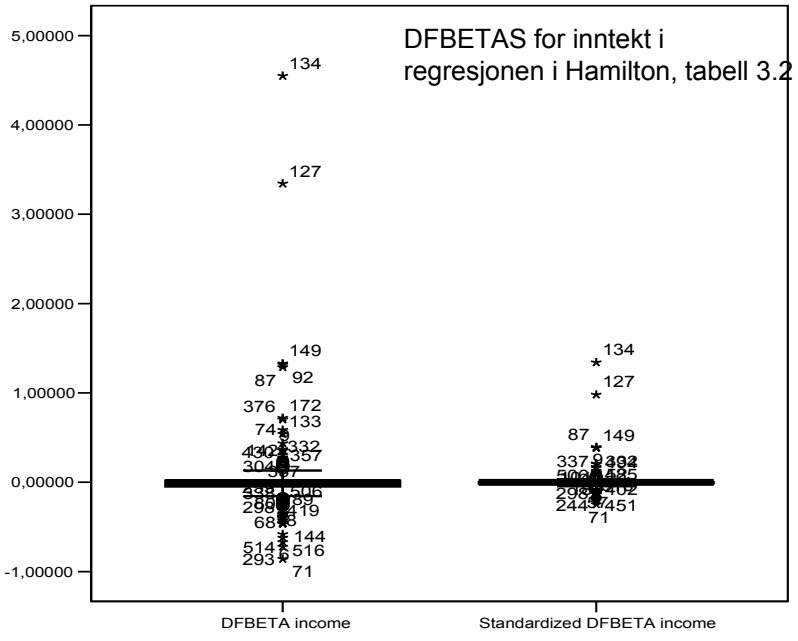
RSS_k er Residual Sum of Squares frå regresjonen av x_k på alle dei andre x-variablane

Kva er ein stor DFBETAS?

- $DFBETAS_{ik}$ vert rekna ut for kvar uavhengig variabel og kvart einaste case. Vi kan ikkje inspisere alle verdiane
- Tre kriterium for å finne dei store verdiane vi treng sjå på (ingen av dei treng vere problematiske)
  - Ekstern skalering: $|DFBETAS_{ik}| > 2/\sqrt{n}$
  - Intern skalering:

    Q_1-1.5IQR < $|DFBETAS_{ik}|$ < Q_3 + 1.5IQR

    (alvorleg utliggjar i box-plott av $DFBETAS_{ik}$)
  - Gap i fordelinga av $DFBETAS_{ik}$



Rekkjefølgje i datafila og case nr er ikkje det same. Case nr er fast.

Rekkje nr	Case nr	water81	water80	water 79	educat	retire	peop 81	cpeop
91	98	1500	1300	1500	16	0	2	0
92	99	3500	6500	5100	14	0	6	0
93	100	1000	1000	2700	12	1	1	0
94	101	3800	12700	4800	20	0	5	0
95	102	4100	4500	2600	20	0	5	0
96	103	4200	5600	5400	16	0	5	-1
97	104	2400	2700	800	16	0	6	0
98	105	1600	2300	2200	14	0	4	0
99	107	2300	2300	3100	16	0	4	-2

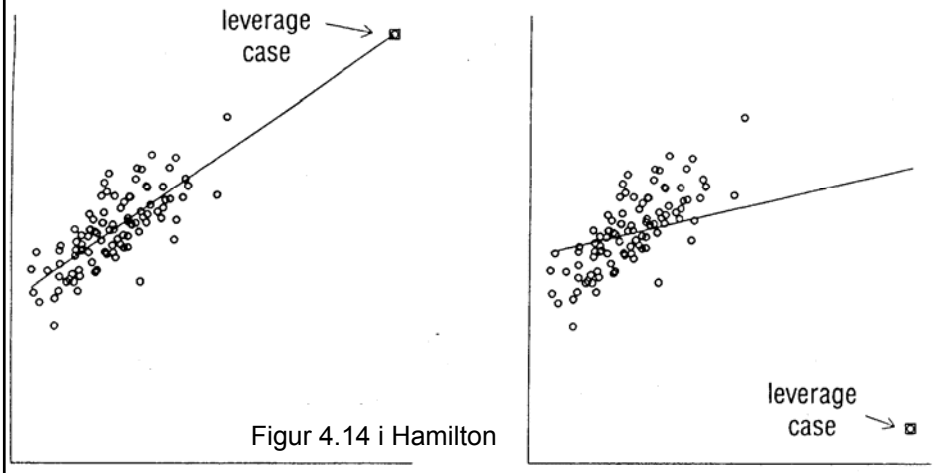
Konsekvensar av case med stor påverknad

- Om vi oppdagar påverknadsrike case skal vi ikkje nødvendigvis ta dei ut av analysen
- Rapporter resultat med og utan casa
- Sjekk påverknadsrike case nøye, kanskje er der målefeil
- Når påverknadsrike case er utliggjarar kan ein minske innverknaden ved transformasjon
- Bruk robust regresjon som ikkje er så lett påverkeleg som OLS regresjon

Potensiell påverknad: leverage

- Den samla påverknaden frå ein bestemt kombinasjon av x-verdiar på eit case måler vi med h_i ”hatt-observatoren”
- h_i varierer frå 1/n til 1. Den har eit gjennomsnitt på K/n (K = # parametrar)
- SPSS rapporterer den sentrerte h_i dvs. (h_i – K/n), vi kan kalle denne for h^c_i

The difference between influence and leverage



Figur 4.14 i Hamilton

High Leverage, Low Influence

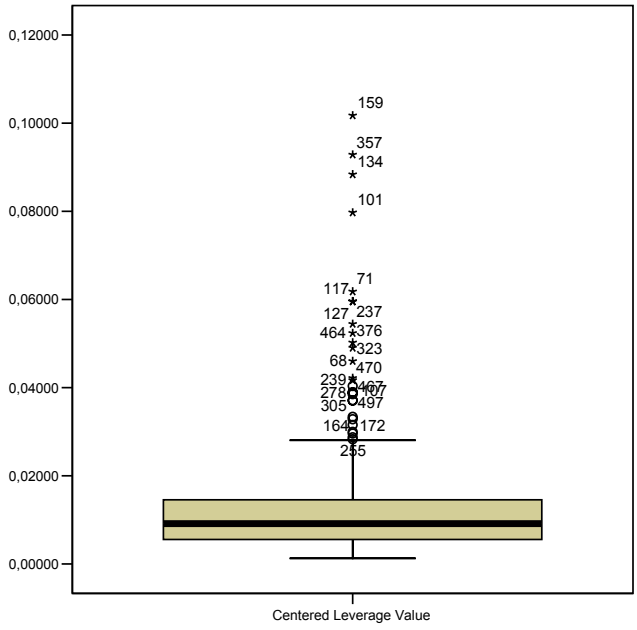
High Leverage, High Influence

Kva er stor verdi av leverage?

- Slik som med DFBETAS kan det stillast opp alternative kriterium. Dei er alle avhengig av utvalsstorleiken n.
  - Dersom $h_i > 2K/n$ (eller $h^c_i > K/n$) finn vi dei ca 5% største h_i ; alternativt
    - Dersom max (h_i) ≤ 0.2 har vi ikkje problem
    - Dersom 0.2 ≤ max (h_i) ≤ 0.5 er der ein viss risiko for problem
    - Dersom 0.5 ≤ max (h_i) har vi truleg eit problem

Sentrert leverage (h^c_i) frå regresjonen i tabell 3.2 i Hamilton

Max av h^c_i er 0.102



Leverage observatoren finst i mange andre case observatorar

- Variansen til den i-te residualen $\text{var}[e_i] = s_e^2[1-h_i]$
- Standardisert residual (*ZRESID i SPSS) $z_i = \frac{e_i}{s_e\sqrt{1-h_i}}$
- Studentifisert residual (*SRESID i SPSS) $t_i = \frac{e_i}{s_{e(i)}\sqrt{1-h_i}}$
- og hugs at standardavviket til residualen er $s_e = \sqrt{RSS/(n-K)}$

Total påverknad: Cook's D_i

- Cook's distanse D_i måler påverknad på heile modellen, ikkje på dei einskilde koeffisientane slik som $DFBETAS_{ik}$

$$D_i = \frac{z_i^2 h_i}{K(1-h_i)}$$

der z_i er den standardiserte residualen

og h_i er hatt observatoren (leverage)

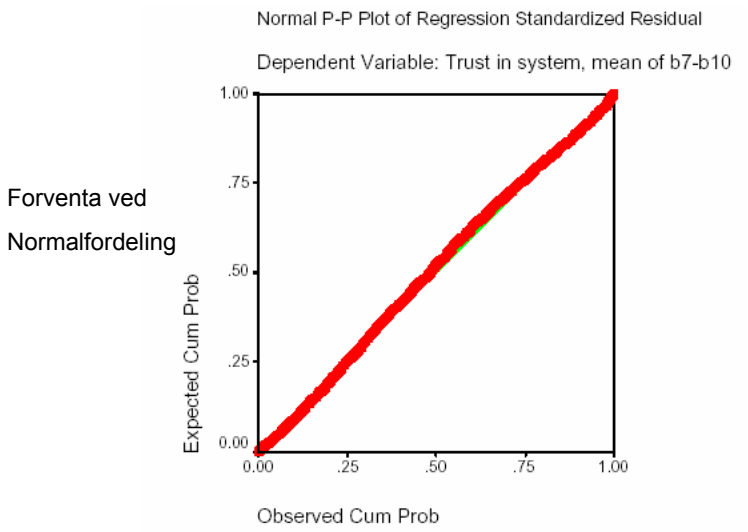
Kva er ein stor D_i ?

- Det kan vere verd å sjå på alle
  - $D_i > 1$ alternativt
  - $D_i > 4/n$ gir dei ca 5% største D_i
- Sjølv om eit case har låg D_i kan det likevel vere slik at det verkar inn på storleiken til einskildkoeffisientar (har stor $DFBETAS_{ik}$)

Kva er eit P-P Plot?

- I regresjonsprosedyren er det noko som heiter ”Normal P-P Plot” og i følgje hjelpemenyen vil dette
- “Displays a normal probability plot of the standardized residuals. Used to check for normality. If the variable is normally distributed, the plotted points form a straight diagonal line.”
- Dette er **ikkje** det same som det Hamilton kallar ”Quantil-Normal Plots” (sjå side 15 i Hamilton) men har same formål: å teste om ei observert fordeling er lik normalfordelinga

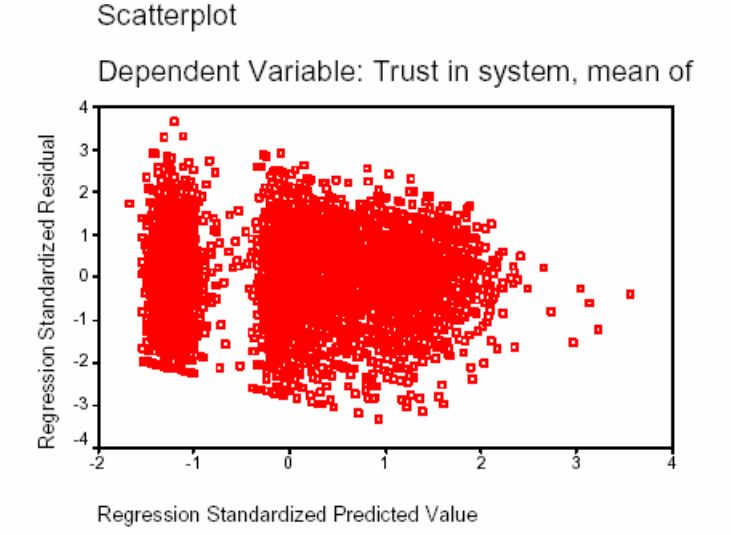
Normal P-P Plot i regresjon: normal probability plots comparing the distribution of standardized residuals to a normal distribution



P-P og Q-Q plott i ”Graphs” menyen

- P-P Plot
- Plots a variable's **cumulative proportions** against the cumulative proportions of any of a number of test distributions. Probability plots are generally used to determine whether the distribution of a variable matches a given distribution. If the selected variable matches the test distribution, the points cluster around a straight line.
- Q-Q Plot
- Plots the **quantiles of a variable's distribution** against the quantiles of any of a number of test distributions. Probability plots are generally used to determine whether the distribution of a variable matches a given distribution. If the selected variable matches the test distribution, the points cluster around a straight line.

Om å ”lese” scatterplott: eksempel frå eksamen Haust 2003



Merkjeleg spreiingsdiagram?

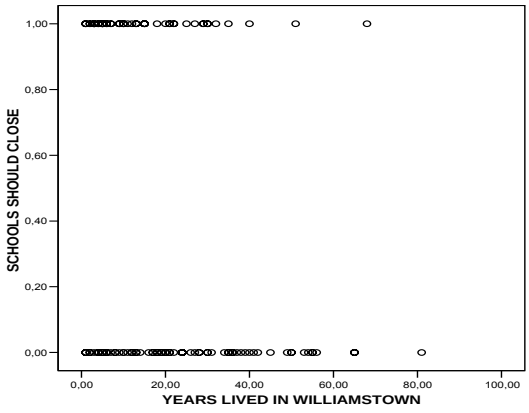
- Det finst i datamaterialet grupper som vi ventar skal ha systematisk lågare predikert verdi av ”Trust in system” (tre land: Polen, Norge, Storbritannia)
- Avhengig variabel har avgrensingar i variasjonsområdet sitt. Den kan t.d. ikkje bli mindre enn 0 eller større enn 10. Dette fører til systematikk i korleis residualane varierer, særleg der observert verdi ligg nær yttergrensene. Vi får ”rette linjer” i utkanten av variasjonsområdet
- Blokka som ligg for seg sjølv er nok Polen

Testing av linearitet i logiten

- **Spørsmål: Hvordan teste for linearitet i logiten? Det hadde vært fint med en ny gjennomgang, gjerne med en litt nærmere spesifisering av hvordan dette testes i SPSS.**
- **Svar**: Kurvelinearitet i logiten kan gi skeive parameterestimat. For å teste om Logiten er lineær i ein x-variabel kan vi gjere følgjande
  - Gruppere x-variabelen
  - For kvar gruppe finne y-gjennomsnitt og rekne det om til logit
  - Lag ein graf av logitane mot gruppert x

Statistiske problem: linearitet i logiten?

- Spreiingsplott for y-x er lite informative sidan y berre har to verdiar
- Y= Lukke skolen mot
- X= år budd i byen



Logiten

- L= naturleg logaritme (Oddsen for y=1) = **ln(p/(1-p))** der p=Pr{y=1}
- For å estimere p treng vi ei gruppe case der nokre har y=1 andre har y=0
- Dersom vi deler opp x-variabelen i intervall vil vi normalt for kvart intervall finne ei gruppe case som har y=1. For kvar gruppe slik definert kan vi rekne ut ein logit
- p = y-gjennomsnitt for dummykoda variable = (prosenten med verdien 1 på y)/100

Eksempel

SCHOOLS SHOULD CLOSE

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	OPEN	87	56,9	56,9	56,9
	CLOSE	66	43,1	43,1	100,0
	Total	**153**	**100,0**	**100,0**	

Descriptive Statistics

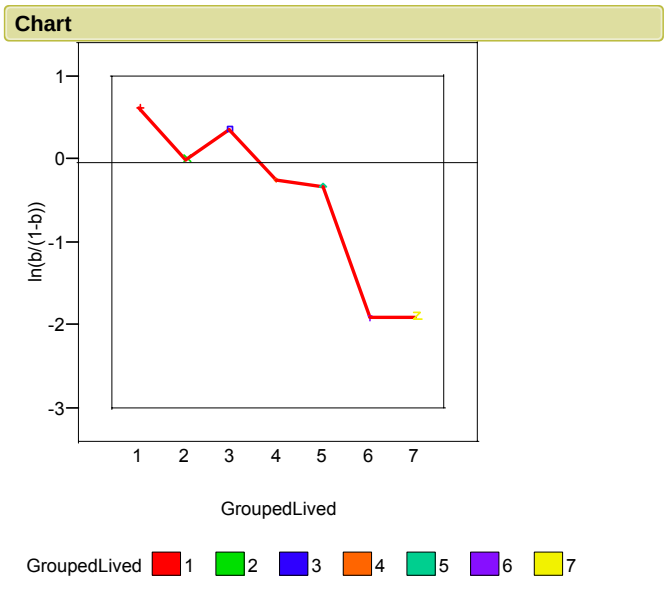
	YEARS LIVED IN WILLIAMSTOWN	SCHOOLS SHOULD CLOSE	Valid N (listwise)
N	153	153	153
Minimum	1,00	,00	
Maximum	81,00	1,00	
Mean	19,2680	,4314	
Std. Deviation	16,95466	,49689	

Linearitet i logiten: eksempel 1

SCHOOLS SHOULD CLOSE		YEARS LIVED IN WILLIAMSTOWN (Banded)						
		<= 3	4-6	7-11	12-22	23-33	34-44	45+
N	OPEN	7	14	7	22	11	13	13
N	CLOSE	13	14	10	17	8	2	2
Within group	Mean (=p)	,65	,50	,59	,44	,42	,13	,13
Logit	Ln(p/(1-p))	0,619	0	0,364	-0,241	-0,323	-1,901	-1,901

Er logiten lineær i "år budd i byen"?

Tja, kanskje det.



Count

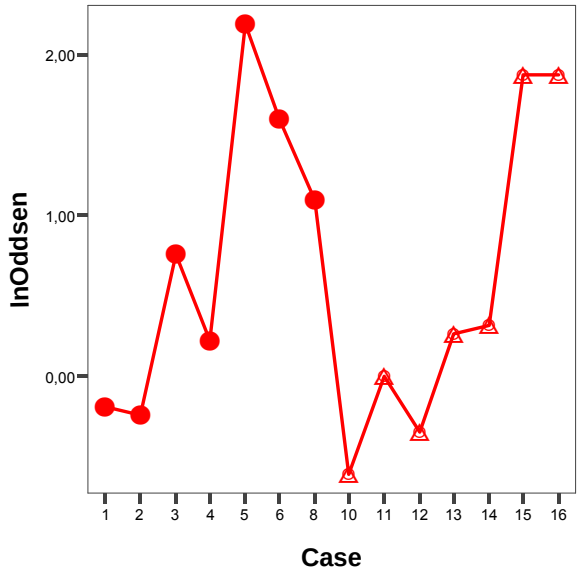
		SCHOOLS SHOULD CLOSE	
		OPEN	CLOSE
YEARS LIVED IN WILLIAMSTOWN (Banded)	<= 3,00	7	13
	4,00 - 6,00	14	14
	7,00 - 11,00	7	10
	12,00 - 22,00	22	17
	23,00 - 33,00	11	8
	34,00 - 44,00	13	2
	45,00+	13	2
YEARS LIVED IN WILLIAMSTOWN (Banded)	1	24	29
	2	14	18
	3	17	8
	4	10	8
	5	9	1
	6	5	1
	7	4	
	8	3	1
	9	1	

+ eksempel 2

Eksempel 1 + 2		SCHOOLS SHOULD CLOSE				
		OPEN		CLOSE		Total
		Count	Row %	Count	Row %	Count
lived(Banded)	<= 3,00	**7**	35,0%	13	65,0%	**20**
	4,00 - 6,00	**14**	50,0%	14	50,0%	**28**
	7,00 - 11,00	**7**	41,2%	10	58,8%	**17**
	12,00 - 22,00	**22**	56,4%	17	43,6%	**39**
	23,00 - 33,00	**11**	57,9%	8	42,1%	**19**
	34,00 - 44,00	**13**	86,7%	2	13,3%	**15**
	45,00+	**13**	86,7%	2	13,3%	**15**
Group Total		**87**	56,9%	66	43,1%	153
Lived9	1-9	**24**	45,3%	29	54,7%	**53**
	10-18	**14**	43,8%	18	56,3%	**32**
	19-27	**17**	68,0%	8	32,0%	**25**
	28-36	**10**	55,6%	8	44,4%	**18**
	37-45	**9**	90,0%	1	10,0%	**10**
	46-54	**5**	83,3%	1	16,7%	**6**
	55-63	**4**	100,0%	0	0%	**4**
	64-72	**3**	75,0%	1	25,0%	**4**
	73-81	**1**	100,0%	0	0%	**1**
Group Total		**87**	56,9%	66	43,1%	153

Eksem-pel 2

Prosent med y=1 i gruppa av lived	ln(p/(100-p))
45,30	**-,19**
43,80	**-,25**
68,00	**,75**
55,60	**,22**
90,00	**2,20**
83,30	**1,61**
100,00	**missing**
75,00	**1,10**
100,00	**missing**
Eksempel 1	
35,00	**-,62**
50,00	**,00**
41,20	**-,36**
56,40	**,26**
57,90	**,32**
86,70	**1,87**
86,70	**1,87**



Til venstre (7 første punkt) gir oppdeling av lived. (2 punkt forsvinn sidan p=100)

Til høgre er alternativ opp-deling basert på

$p^F = 100-p$

Formulering av modell

- **Spørsmål:** Formulering av en modell: Hvor mye skal egentlig være med? Trodde det kun var populasjonsligningen, "La Y være... Sett X til... etc", og forutsetningen for OLS/ Logistisk regresjon. Det virker som det er mye mer omfattende …
- **Svar:** Vel, ikkje så mye meir …

Formulering av modellar

- Definisjon av elementa i modellen
  - **variablar**, feilledd, populasjon og utval
- Definisjon av relasjonar mellom elementa
  - **likninga som bind elementa saman**, utvalsprosedyre, tidsrekkefølgje av hendingar og observasjonar,
- Presisering av føresetnader for bruk av gitt estimeringsmetode
  - tilhøve til substansteori (**spesifikasjon**)
  - **fordeling og eigenskapar ved feilledd**

Elementa i modellen

- **Variablar:** fenomenet vi ønskjer å studere må kunne observerast og seiast å ha ulike tilstandar eller uttrykksformer i ulike einingar i den populasjonen vi observerer. Vi må finne variasjon.
- **Feilledd:** feilleddet er ein abstrakt sekk som inneheld alle dei mange aspekta av populasjonen som vi ikkje er i stand til å observere og inkludere i modellen.
- **Populasjon:** kven eller kva er det vi ønskje å seie noko om?
- **Utval:** idealet er eit reint tilfeldig utval, om vi ikkje kan få det må vi vite nøyaktig korleis utvalsmetoden er knytt opp mot den avhengige variabelen (fenomenet) vi ønskjer å studere

Relasjonar mellom elementa

- **Likninga**: relasjonar mellom variablar
- **Utvalsprosedyre**: skeive (biased) utval pga seleksjon og manglande data
- Tidsrekkjefølgje av hendingar og observasjonar: kausal retning
- Samvariasjon, genuin/ spuriøs samvariasjon
  - Konklusjonar om kausalsamband krev genuin samvariasjon

Når man skal definere relasjonen mellom variablene i en modell, skal "e" være inkludert også for at det skal være korrekt?

Svaret er ja! Men legg merke til:

- Observert Y_i = Predikert Y_i + e_i

(Predikert Y_i = Forventa verdi av Y_i = $E[Y_i]$)

Modellen kan omtalast på to måtar

- $E[Y_i] = f(x_i)$ da er residualen underforstått

eller

- $Y_i = f(x_i) + e_i = E[Y_i] + e_i$ da er residualen inkludert eksplisitt

Og hugs skiljet mellom populasjon og utval

Føresetnader for bruk av gitt estimeringsmetode

For å nytte OLS metoden til å estimere ein lineær modell må følgjande føresetnader gjerast:

- I. Modellen er korrekt, dvs.:
  - alle relevante variablar er med
  - ingen irrelevante er med
  - modellen er lineær i parametrane
- II. Gauss-Markov krava for «Best Linear Unbiased Estimates» (BLUE) er oppfylt
- III. Feilleddet er normalfordelt

Forutsetningen for at en gitt modell er korrekt at modellen må være "lineær i parametrene". Hva betyr det? Er det et problem med f.e. AgeSquare?

Legg merke til at kravet ikkje går på om modellen er korrekt eller ikkje. Kravet går på om vi skal kunne nytte OLS metoden for å estimere parametrane

- **Parametrane** i lineære modellar er β_k eller ”b_k”-ar som vi estimerer
- ”Lineær i parametrane” tyder at vi finn ein og berre ein ”b_k” mellom kvart ”+” i modellen
- Sidan det står berre ein ”b_k” framfor ”AgeSquare” er modellen lineær i parameteren

Er Gauss Markov kravene oppfylt selv om det ikke er normalfordelte restledd? Kan f- og t-test brukes selv om restleddene ikke er normalfordelte?

- Gauss-Markov krava kan godt vere oppfylt sjølv om vi ikkje har normalfordelte restledd, men dei treng ikkje vere det
- F-test og t-test kan nyttast **berre** dersom restleddet er normalfordelt

Krav til semesteroppgåve (1)

- **FORORD**
- Dersom du nyttar **data frå SSB sine granskingar** skal følgjande med i ein fotnote eller i eit forord:
  - "(En del av) de data som er benyttet i denne publikasjonen er hentet fra .....undersøkelsen (årstall). Data i anonymisert form er stilt til disposisjon gjennom Norsk samfunnsvitenskapelig datatjeneste (NSD). Innsamling og tilrettelegging av data ble opprinnelig utført av Statistisk Sentralbyrå. Hverken Statistisk Sentralbyrå eller NSD er ansvarlige for analysen av dataene eller de tolkninger som er gjort her."
- For kommunedata skriv ein ikkje «i anonymisert form».

Krav til semesteroppgåve (2)

- **Innbinding** er ikkje nødvendig. Men om de ønskjer å binde inn oppgåva skal det nyttast innbinding i **A4 format**.
- **Tittelsida skal** minimum innehalde studentnummer og tittel som indikerer avhengig variabel. For somme formål kan namn vere nyttig, men det er frivillig om namnet skal stå på oppgåva.

Krav til semesteroppgåve (3)

- **KRAVSPESIFIKASJONAR**
- a) Med utgangspunkt i deskriptiv statistikk for variablane som skal inkluderast i modellen, skal fordelinga deira beskrivast og mogelege transformasjonar vurderast. **Transformasjonar skal takast i bruk dersom dette kan forbetre analysen substansielt** (dvs. det er teoretiske grunnar til å tru at det marginale sambandet mellom forklaringsvariabel og avhengig variabel er kurvelineært, jfr. pkt d) eller dersom det kan gjere testprosedyrane meir truverdige (residualen kjem nærmare opp mot normalfordelinga)

Krav til semesteroppgåve (4)

- b) Modellen skal innehalde minst ein kategorisk forklaringsvariabel med meir enn to kategoriar (MÅLENIVÅ: nominalskala).
- c) Det skal gjennomførast ei drøfting av mogelege interaksjonar og minst eitt interaksjonsledd skal testast.
- d) Mogelege kurvelineære samanhengar skal drøftast og minst ein kurvelineære samanheng skal testast.

Krav til semesteroppgåve (5)

Med utgangspunkt i den første modellen skal følgjande drøftast

- e) I OLS-regresjon skal normalfordelinga av feilleddet vurderast.
- f) I OLS-regresjon skal det testast for heteroskedastisitet.
- g) I OLS regresjon skal effekten av autokorrelasjon vurderast.
- h) I logit regresjon skal diskrimineringsproblem vurderast.

Krav til semesteroppgåve (6)

- i) I både OLS og logit-regresjon skal multikollinearitetsproblem vurderast
- j) I både OLS og logit-regresjon skal effekten av utliggarar og innflytelsesrike case vurderast og eventuelt illustrerast.
- k) I både OLS og logit-regresjon skal modellspesifikasjonen vurderast.