Ref.:
http://www.svt.ntnu.no/iss/Erling.Berge/

SOS3003

Anvendt statistisk dataanalyse i samfunnsvitenskap

Forelesingsnotat 03

Erling Berge
Institutt for sosiologi og statsvitenskap
NTNU

Forelesing III

- Multivariat regresjon II Hamilton Kap 3 s72-84
- **Om å skrivesemesteroppgåve**
  - Variable og variasjon.
  - Måleteori og målenivå
  - Koding og omkoding
- **Val/ tildeling av avhengig variabel**
  - Eigne data, data frå tidlegare oppgåver
  - Data frå European Social Survey (ESS)

Multippel regresjon: modell

La K = talet på parametrar i modellen

Populasjonsmodellen

- $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + ... + \beta_{K-1} x_{i,K-1} + \varepsilon_i$
  $i = 1, ... ,N$
  der N = talet på case i utvalet eller populasjonen

Utvalsmodellen

- $y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + ... + b_{K-1} x_{i,K-1} + e_i$
  $i = 1, ... ,n$
  der n = talet på case i utvalet eller populasjonen

Estimeringsmetodar

- OLS: Vi finn parametrane ved å minimere RSS.

Dette er imidlertid ikkje den einaste metoden:

1) WLS: Vekta minste kvadraters metode
2) ML: maksimum sannsyn (maximum likelihood)

Partielle effektar

Leverage plott for y og x_k er eit plott der

- y-aksen er residualen frå regresjonen av y på alle x-variablane utanom x_k , og
- X-aksen er residualen frå regresjonen av x_k på alle dei andre x-variablane

Regresjonslina i eit slikt plott vil alltid gå gjennom y=0 og ha ein vinkelkoeffesient lik b_k

Eit eksempel, to uavhengige variablar

Table 2.2 Dependent: **Summer 1981 Water Use**	**B**	Std. Error	t	Sig.
(Constant)	**1201.124**	123.325	9.740	.000
Income in Thousands	**47.549**	4.652	10.221	.000

Table 3.1 Dependent: **Summer 1981 Water Use**	**B**	Std. Error	t	Sig.
(Constant)	**203.822**	94.361	2.160	.031
Income in Thousands	**20.545**	3.383	6.072	.000
Summer 1980 Water Use	**.593**	.025	23.679	.000

Frå tabellane 2.2 (side 46) og 3.1 (side 68) hos Hamilton
Tabellutforming: i tabellen i boka er konstanten på siste linje, SPSS set den på første.
Kva tyder det at koeffisienten til inntekt minkar når vi legg til ein ny variabel?

Om val av uavhengige variablar

- Det er sjeldan eksisterande teori gir oss presise råd om kva for variablar vi skal inkludere i ein modell. Det vil som regel vere eit element av prøving og feiling i arbeidet med å utvikle ein modell.
- Når vi legg til nye variablar skjer fleire ting:
  - Forklaringskrafta aukar: R^2 vert større, men er auken signifikant?
  - Regresjonskoeffisienten viser effekten på y. Er effekten signifikant ulik 0 og så stor at den har substansiell interesse?
  - Spuriøse koeffisientar kan minke. Endrar dei nye variablane tolkinga av dei andre variablane sine effektar?

Parsimonitet

- Parsimonitet er det vi kan kalle eit estetisk kriterium på ein god modell. Vi ønskjer å forklare mest mogeleg av variasjonen i y ved hjelp av færrast mogelege variablar
- Den justerte determinasjonskoeffisienten Adjusted R^2 er basert på parsimonitet i den forstand at den tar omsyn til kompleksiteten i data relativt til modellen gjennom differansen
  n-K (residualen sine fridomsgrader)
  (n = talet på observasjonar, K = talet på estimerte parametrar)

Irrelevant variabel

- Inkludere ein irrelevant variabel
  - Ein variabel er irrelevant dersom den verkelege effekten (β) ikkje er signifikant ulik 0, eller meir pragmatisk, dersom effekten av variabelen er for liten til å ha substansiell interesse.
  - **Inklusjon av ein irrelevant variabel** gjer modellen unødig kompleks og vil føre til at koeffisientestimata på alle variablane får større varians (varierer meir frå utval til utval)

Relevant variabel

- Ein variabel er relevant dersom den
  1. verkelege effekten (β) er signifikant ulik 0, og stor nok til å ha substansiell interesse og
  2. er **korrelert med andre inkluderte** x-variablar
- Dersom vi **utelet ein relevant variabel** vil alle resultat frå regresjonen bli upåliteleg. Modellen er ei urealistisk forenkling

Utvalsspesifikke resultat?

- Å velje variablar er ei avveging mellom ulike riskar. Kva for ein risk som er verst er avhengig av formålet med studien og styrken i relasjonane.
- Gitt testnivå på 0.05 vil vi godt kunne finne utvalsspesifikke resultat. I omlag 5% av alle utval vil ein koeffisient som ikkje er signifikant ulik null "eigentleg" vere signifikant ($\beta \neq 0$) (og tilsvarande for dei som er signifikant ulik null)

Hamilton (s74) eksempel

y_i	vassforbruk etter krisa (1981)
x_{i1}	hushaldsinntekt i tusen dollar
x_{i2}	vassforbruk før krisa (1980)
x_{i3}	utdanning i år for hovudpersonen i hushaldet
x_{i4}	pensjonist (koda 1 dersom hovudpersonen er pensjonist, 0 elles)
x_{i5}	talet på personar i hushaldet under krisa (s81)
x_{i6}	endring i talet på personar 1980 til 1981

Tabell 3.2 (Hamilton s.74)

Dependent Variable: Summer 1981 Water Use	B	Std. Error	t	Sig.	Beta
(Constant)	242.220	206.864	1.171	.242	
Income in Thousands	20.967	3.464	6.053	.000	.184
Summer 1980 Water Use	.492	.026	18.671	.000	.584
Education in Years	-41.866	13.220	-3.167	.002	-.087
Head of house retired?	189.184	95.021	1.991	.047	.058
# of People Resident, 1981	248.197	28.725	8.641	.000	.277
Increase in # of People	96.454	80.519	1.198	.232	.031

Korleis skal vi tolke koeffisienten for "Increase in # of People" ?

Kva vil føre til minke i vassforbruket etter krisa?

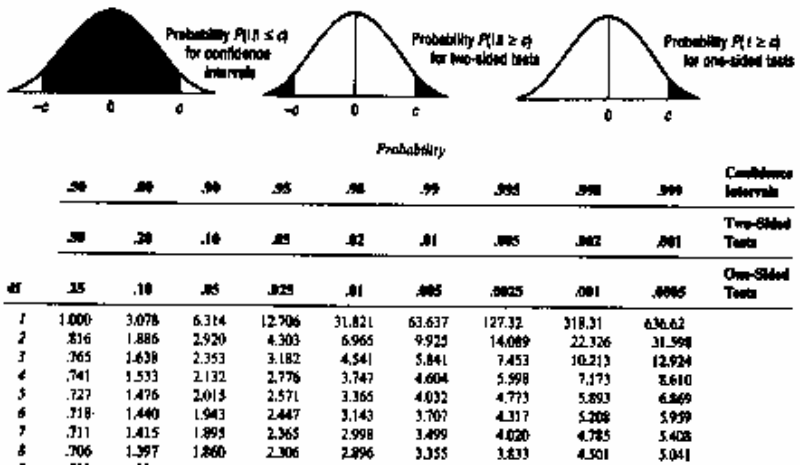
Standardiserte koeffisientar

- Standardiserte variablar (z-skårar)
  $z_{ix} = (x_i - \bar{x})/s_x$ (måleeininga er standardavvik)
- Standardiserte regresjonskoeffisientar (beta-vekter, eller stikoeffisientar)
  $b_k^s = b_k(s_k/s_y)$ (varierer mellom -1 og +1)
- Predikert standardskåre av y_i (z_{iy} med hatt) = $0.18z_{i1} + 0.58z_{i2} - 0.09z_{i3} + 0.06z_{i4} + 0.28z_{i5} + 0.03z_{i6}$

Ref.:
http://www.svt.ntnu.no/iss/Erling.Berge/

t-test

- Skilnaden mellom observert koeffisient (b_k) og uobserverte koeffisient (β_k) standardisert med standardavviket til den observerte koeffisienten (SE_{b_k}) vil normalt vere svært nær null dersom den observerte b_k ligg nær populasjonsverdien. Dette tyder at dersom vi i formelen
- $t = (b_k - \beta_k)/ SE_{b_k}$ set inn H_0: $\beta_k = 0$ og finn at "t" er liten vil vi tru at populasjonsverdien β_k eigentleg er lik 0. Kor stor "t" må vere for at vi skal slutte å tru at $\beta_k = 0$ kan vi finne ut frå kunnskap om samplingfordlingane til b_k og SE_{b_k}



350 Appendix 4 Statistical Tables

Table A4.1 Critical values for student's *t*-distribution

Probability $P(|t| \le c)$ for confidence intervals

Probability $P(|t| \ge c)$ for two-sided tests

Probability $P(t \ge c)$ for one-sided tests

Probability

	Confidence Intervals								
	.50	.80	.90	.95	.98	.99	.995	.998	.999
	Two-Sided Tests								
	.50	.20	.10	.05	.02	.01	.005	.002	.001
df	One-Sided Tests								
	.25	.10	.05	.025	.01	.005	.0025	.001	.0005
1	1.000	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	.816	1.886	2.920	4.303	6.965	9.925	14.089	22.326	31.598
3	.765	1.638	2.353	3.182	4.541	5.841	7.453	10.213	12.924
4	.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041

"t" har ei samplingfordeling kalla t-fordelinga, t-fordelinga varier med talet på fridomsgrader (n-K) og er lista etter signifikansnivået α

Konfidensintervall for β

- Vel ein t_α-verdi frå tabellen over t-fordelinga med n-K fridomsgrader slik at dersom $H_0 : \beta_k = b_k$ er rett vil ein tosidig test ha eit sannsyn på α for å forkaste H_0 når H_0 eigentleg er rett (feil av type I)
  dvs det er eit sannsyn α for at β_k eigentleg ligg utanfor
  $< b_k - t_\alpha(SE_{b_k}) , b_k + t_\alpha(SE_{b_k}) >$
- Dette er det same som at påstanden
  $b_k - t_\alpha(SE_{b_k}) \leq \beta_k \leq b_k + t_\alpha(SE_{b_k})$
  er rett med sannsyn $1 - \alpha$

F-testen: stor mot liten modell

RSS = residual sum of squares med indeks {*}:

RSS{K-H} = RSS i modellen med K-H parametrar

(H er lik skilnaden i talet på parametrar i to modellar)

RSS{K} = RSS i modellen med K parametrar

$$F^{H}_{n-K} = \frac{(RSS\{K-H\} - RSS\{K\})/H}{(RSS\{K\})/(n-K)}$$

er da F-fordelt med H og n-K fridomsgrader

Eksempel (Hamilton tabell 3.1 mot 3.2)

Liten modell Table 3.1	**Sum of Squares**	df	Mean Square	F	Sig.
Regression (Model) (Explained)	671025350.237	2	335512675.119	391.763	.000(a)
Residual	422213359.440	493	856416.551		
Total	1093238709.677	495			

Stor modell Tabell 3.2	Sum of Squares	df	Mean Square	F	Sig.
Regression	740477522.059	**K - 1 = 6**	123412920.343	171.076	.000(a)
Residual	352761187.618	**n - K = 489**	721393.022		
Total	1093238709.677	**n - 1 = 495**			

Test om den store modellen (7 parametrar) er betre enn den vesle (3 parametar)

Noter til tabelleksempel

- K = talet av parametrar i stor modell (6 variablar pluss konstant) = 7
- H = K – talet av parametrar i liten modell (2 variablar pluss konstant) = 7 – 3 = 4
- RSS{K-H} = 422213359.440
- RSS{K} = 352761187.618
- n = 496
- n – K = 496 – 7 = 489
- (RSS{K-H} – RSS{K})/H = (422213359.440 - 352761187.618)/4 = 17363042.9555
- RSS{K}/(n-K) = 352761187.618/489 = 721393.0217

Test av alle parametrane under eitt

- Dersom den store modellen har K parametrar og vi lar den lille modellen vere så liten som mogeleg med berre 1 parameter (gjennomsnittet) vil testen vår ha

  H = K-1. Sett inn i formelen ovanfor får vi

$$F_{\{K-1,\ n-K\}} = \frac{ESS/(K-1)}{RSS/(n-K)}$$

Dette er F-verdien vi finn i ANOVA tabellane frå SPSS

Multikollinearitet (1)

- Multikollinearitet involverer berre x-variablane, ikkje y, og dreiar seg om lineære samband mellom to eller fleir x-variablar
- Dersom det er perfekt korrelasjon mellom to forklaringsvariablar t.d. x og w (dvs. $r_{xw} = 1$) vil den multiple regresjonsmodellen bryte saman
- Tilsvarande skjer dersom der er perfekt korrelasjon mellom to grupper av forklaringsvariable

Multikollinearitet (2)

- Perfekt multikollinearitet er svært sjeldan eit praktisk problem
- Men høge korrelasjonar mellom ulike x-variablar eller ulike grupper av x-variablar vil gjere at estimata av effekten deira blir svært usikker. Dvs. regresjonskoeffisienten vil ha svært stor standardfeil og t-testane blir i praksis uinteressante
- F-testen av ei gruppe variablar er ikkje påverka

Søkjestrategiar

- Det finst metodar for automatisk søk etter forklaringsvariablar i eit større datasett
- Det er vårt råd å ikkje nytte desse strategiane
- Eitt av problema er at p-verdiane i testane vi får ut i slike søk er feil og alt for ”snille” (problemet med multiple samanlikningar)
- Eitt anna problem er at slike søk fungerer dårleg når variablane er høgt korrelert

Dummyvariablar: gruppeskilnader

- Dikotome variable som tar verdiane 0 eller 1 kallast dummy-variable
- I eksempelet i tabell 3.2 (s74) er x_{i4} (hovudperson pensjonist eller ikkje) ein dummyvariabel
- Sett først inn $x_{i4} = 1$ så $x_{i4} = 0$ i likninga
  $y_i = 242 + 21x_{i1} + 0.49x_{i2} - 42x_{i3} + 189x_{i4} + 248x_{i5} + 96x_{i6}$ og
- Forklar kva dei to likningane viser

Interaksjon

- Det er interaksjon mellom to variablar dersom effekten av den eine variabelen varierer etter kva verdi den andre variabelen har.

Interaksjonseffektar i regresjon

- Dersom vi foretar ein ikkje-lineær transformasjon av y vil alle estimerte effektar implisitt vere interaksjonseffektar
- Enkle additive interaksjonseffektar kan vi inkludere i ein lineær modell ved hjelp av produktledd der vi multipliserer to x-variablar med kvarandre
- $\hat{y}_i = b_0 + b_1x_i + b_2w_i + b_3x_iw_i$
- Betinga effekt plott vil kunne illustrere kva interaksjon tyder

Semesteroppgåve: avhengig variabel

Krav til avhengig variabel

- For å kunne gjennomføre ein regresjonsanalyse bør det veljast ein variabel med høveleg skalanivå og nok variasjon.
- **Ordinær OLS-regresjon** krev at den avhengige variabelen er ein intervallskala eller forholdstalskala
- **Logistisk regresjon** krev at den avhengige variabelen har nett 2 verdiar (dikotom)

Skalanivå

Skalanivå	Nominal	Ordinal	Intervall	Forholdstal
nominal	**grupperer**			
ordinal	grupperer	+ **rangerer**		
intervall	grupperer	+ rangerer	+ **avstand**	
forholdstal	grupperer	+ rangerer	+ avstand	+ **absolutt nullpunkt**
eksempel	• Bostads-kommune	• # i sjuke-huskø	• Temperatur i C^0	•Alder •Temperatur i K^0

Haust 2004 © Erling Berge 2004 29

Vanlege variablar

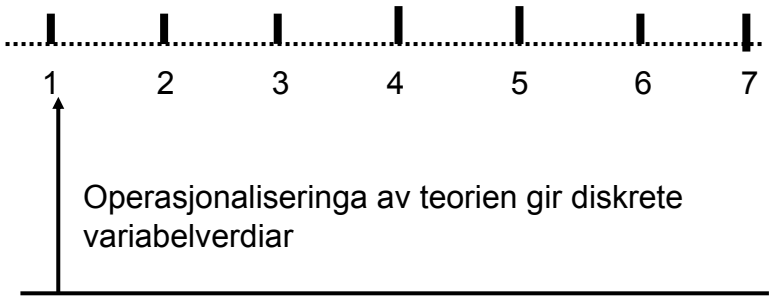
- Svært mange variablar i sosiologi og statsvitskap har formelt sett ordinalskala
- Dei kan med visse føresetnader (og for våre formål) handsamast som intervallskala:
  - Talet av kategoriar som er rangert er ”stort nok” (minst 7)
  - Fordeling på nesten alle kategoriane (nok personar utanfor dei 2-3 modale kategoriane)
  - Kan føresetje ein teoretisk underliggjande skala med avstandsmål

Haust 2004 © Erling Berge 2004 30

Ref.:
http://www.svt.ntnu.no/iss/Erling.Berge/

Måling av variablar

Observasjonen klarer berre å skilje mellom 7 punkt



Teoretisk underliggjande kontinuerleg skala

Typisk: retning og styrke av meiningar eller kjensler

Dikotome variablar

- Har to verdiar og kan nyttast i all slags regresjon
- Alle variablar kan kodast om til å ha to verdiar (dikotom variabel)
- Dersom ein kodar dei to kategoriane 0 og 1 vil tolkinga av effekten deira når dei ernytta som uavhengige variablar bli lettare enn om ein vel andre kodeverdiar
- Talet av observasjonar i den minste kategorien må vere ”stort nok”

Semesteroppgåve

- Det er eit mål at alle skal ha ulike avhengige variablar
- Datamateriale
  - Eigne data (krava til avhengig variabel må stettast)
  - Data frå tidlegare oppgåver (kan nyttast om igjen om ein ønskjer)

  For alle andre er
  - Nye variablar tilgjengeleg frå European Social Survey (see http://www.sv.ntnu.no/iss/Erling.Berge/semesteroppgave.html)

Om å finne avhengig variabel

- Er temaet som variabelen fortel om interessant?
- Er det nok variasjon i variabelen? Kjør ut frekvenstabell, eller deskriptiv statistikk
- Sjekk spesielt kor mange missing det er. Det bør ikkje vere ”for mange” (under 10%?)
- Dersom variabelen ikkje kan nyttast i OLS regresjon: kan den kodast om til dikotomi?

Ref.:
http://www.svt.ntnu.no/iss/Erling.Berge/

Random Numbers

1368	9621	9151	2066	1208	2664	9822	6599	6911	5112
5953	5936	2541	4011	0408	3593	3679	1378	5936	2651
7226	9466	9553	7671	8599	2119	5337	5953	6355	6889
8883	3454	6773	8207	5576	6386	7487	0190	0867	1298
7022	5281	1168	4099	8069	8721	8353	9952	8006	9045
4576	1853	7884	2451	3488	1286	4842	7719	5795	3953
8715	1416	7028	4616	3470	9938	5703	0196	3465	0034
4011	0408	2224	7626	0643	1149	8834	6429	8691	0143
1400	3694	4482	3608	1238	8221	5129	6105	5314	8385
6370	1884	0820	4854	9161	6509	7123	4070	6759	6113
4522	5749	8084	3932	7678	3549	0051	6761	6952	7041
7195	6234	6426	7148	9945	0358	3242	0519	6550	1327
0054	0810	2937	2040	2299	4198	0846	3937	3986	1019
5166	5433	0381	9686	5670	5129	2103	1125	3404	8785
1247	3793	7415	7819	1783	0506	4878	7673	9840	6629
8529	7842	7203	1844	8619	7404	4215	9969	6948	5643
8973	3440	4366	9242	2151	0244	0922	5887	4883	1177
9307	2959	5904	9012	4951	3695	4529	7197	7179	3239
2923	4276	9467	9868	2257	1925	3382	7244	1781	8037
6372	2808	1238	8098	5509	4617	4099	6705	2386	2830