

# SOS3003

## Anvendt statistisk dataanalyse i samfunnsvitenskap

Forelesingsnotat 04

Erling Berge  
Institutt for sosiologi og statsvitenskap  
NTNU

Fall 2004

© Erling Berge 2004

1

## Forelesing IV

Multivariat regresjon II

- Hamilton Kap 3 s84-101,
- Hardy, Melissa A. 1993 "Regression with dummy variables" Sage University Paper: QASS 93, London, Sage,

Fall 2004

© Erling Berge 2004

2

## Interaksjonseffektar i regresjon

- Ein interaksjonseffekt mellom  $x$  og  $w$  kan inkluderast i ein regresjonsmodell ved å ta inn ein hjelpevariabel lik produktet av dei to, dvs. Hjelpevariabel  $H=x*w$
- $y_i = b_0 + b_1*x_i + b_2*w_i + b_3*H_i + e_i$
- $y_i = b_0 + b_1*x_i + b_2*w_i + b_3*x_i*w_i + e_i$

Fall 2004

© Erling Berge 2004

3

## Eksempel frå Hamilton(s85-91)

Sett

- $y$  = naturleg logaritme av klorid konsentrasjon
- $x$  = djupna av brønnen (1=djup, 0=grunn)
- $w$  = naturleg logaritme av avstand frå vei
- $xw$  = interaksjonsledd mellom avstand og djupn (produktet  $x*w$ ). Da er
- $\hat{y}_i = b_0 + b_1x_i + b_2w_i + b_3x_iw_i$

**Vi ser først på dei enkle modellane utan interaksjon**

Fall 2004

© Erling Berge 2004

4

## Figures 3.3 and 3.4 (Hamilton p85-86)

Figure 3.3 is based on

Dependent Variable: lnChlorideConcentra	B	Std. Error	Beta	t	Sig.
(Constant)	3.775	.429		8.801	.000
x= BEDROCK OR SHALLOW WELL?	-.706	.477	-.205	-1.479	.145

Figure 3.4 is based on

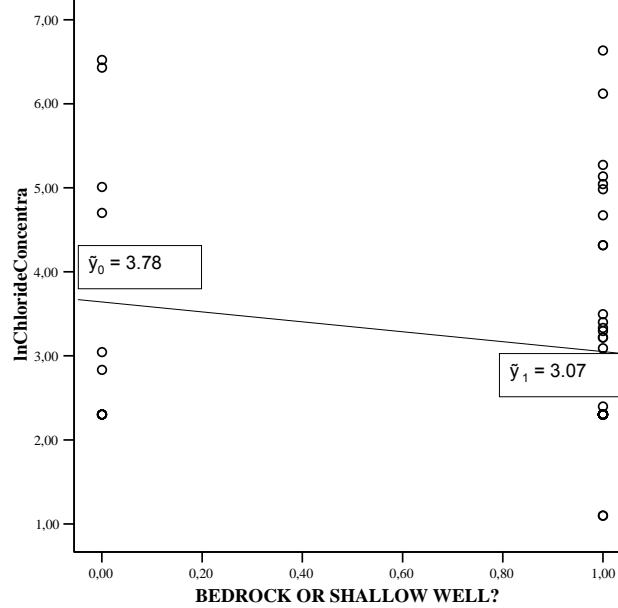
Dependent Variable: lnChlorideConcentra	B	Std. Error	Beta	t	Sig.
(Constant)	4.210	.961		4.381	.000
w= lnDistanceFromRoad	-.091	.180	-.071	-.506	.615
x= BEDROCK OR SHALLOW WELL?	-.697	.481	-.202	-1.449	.154

Fall 2004

© Erling Berge 2004

5

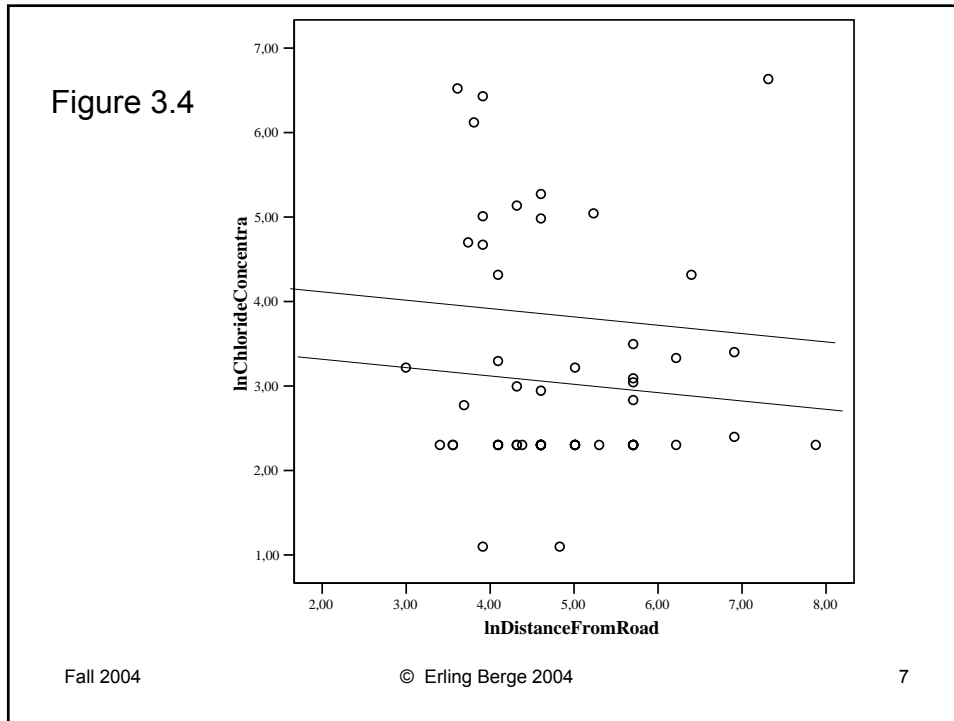
Figure 3.3



Fall 2004

© Erling Berge 2004

6



## Figures 3.5 and 3.6 (Hamilton p89-91)

Figure 3.5 is based on

Dependent Variable: lnChlorideConcentra	B	Std. Error	Beta	t	Sig.
(Constant)	3.666	.905		4.050	.000
w= lnDistanceFromRoad	-.029	.202	-.022	-.144	.886
x*w= lnDroadDeep	-.081	.099	-.128	-.819	.417

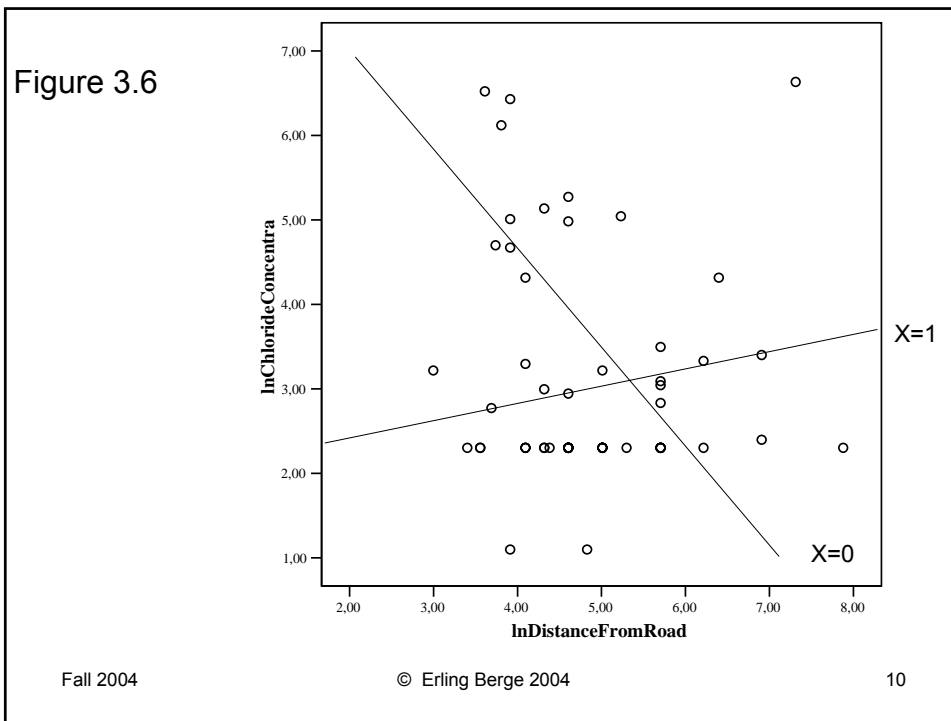
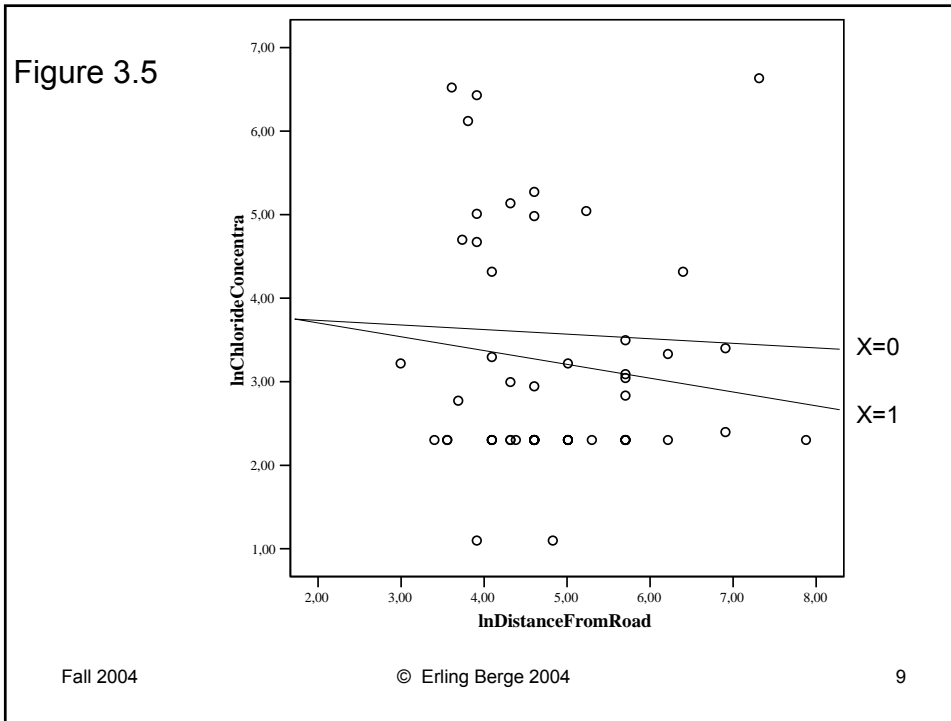
Figure 3.6 is based on

Also see Table 3.4 in Hamilton p90 Dependent Variable: lnChlorideConcentra	B	Std. Error	Beta	t	Sig.
(Constant)	9.073	1.879		4.828	.000
w= lnDistanceFromRoad	-1.109	.384	-.862	-2.886	.006
x= BEDROCK OR SHALLOW WELL?	-6.717	2.095	-1.948	-3.207	.002
x*w= lnDroadDeep	1.256	.427	1.979	2.942	.005

Fall 2004

© Erling Berge 2004

8



## Multikollinearitet

- Når vi tar med alle tre variablane,  $x$ ,  $w$  og  $x*w$  vil vi også introdusere eit visst element av multikollinearitet. Vi kan ikkje stole på testane av einskildkoeffisientar
- Som førre eksempel viser kan vi imidlertid ikkje droppe ein av dei utan fare for å droppe ein relevant variabel
- F-test av t.d.  $w$  og  $z*w$  under eitt unngår testproblemet, og litt eksperimentering med ulike modellar vil vise om utelating av  $w$  eller  $x*w$  endrar samanhengane substansielt

Fall 2004

© Erling Berge 2004

11

## Nominalskalavariabel

- Kan inkluderast i regresjonsmodellar ved å lage nye hjelpevariablar: ein for kvar kategori i nominalskalavariabelen. Med  $J$  kategoriar får vi  $H(j)$ ,  $j=1, \dots, J$  hjelpevariablar
- Dersom vi har ein intervallskala avhengig variabel og ein nominalskala uavhengig variabel vil ein ofte analysere den ved hjelp av variansanalyse (ANOVA)
- Ved introduksjon av hjelpevariable kan vi utføre same analysane i regresjonsmodellen

Fall 2004

© Erling Berge 2004

12

## Variansanalyse -ANOVA

- Analyse av ein intervallskala avhengig variabel og ein eller fleire nominalskalavariabel, ofte kalla faktorar
  - Einvegs ANOVA nyttar ein nominalskalavariabel
  - Tovegs ANOVA nyttar to nominalskalavariabel
  - OSV
- Testar av skilnader mellom grupper baserer seg på vurderingar av om variasjonen innan ei gruppe (definert av "faktorane") er stor eller liten relativt til variasjonen mellom gruppene

Fall 2004

© Erling Berge 2004

13

## Nominalskalavariabel i regresjon (1)

- Dersom den kategoriske variabelen har  $J$  kategoriar kan vi maksimalt ta med  $J-1$  hjelpevariablar i regresjonen  
(  $H(j)$ ,  $j=1, \dots, J-1$  vert inkludert;  $H(J)$  må utelatast)
- Den utelatne hjelpevariabelen kallar vi referansekategori og er den viktigaste for tolkinga av analyseresultatet

Fall 2004

© Erling Berge 2004

14

## Nominalskalavariabel i regresjon (2)

### **Dummy-koding** av nominalskalavariabel

- Hjelpevariabelen  $H(j)$  vert koda 1 for person nr  $i$  dersom person nr  $i$  er å finne i kategori  $j$  på nominalskalavariabelen, den vert koda 0 dersom person  $i$  ikkje er i kategori  $j$ .
- Gjennomsnittet for ein dummykoda variabel er proporsjonen med verdi 1.

Fall 2004

© Erling Berge 2004

15

## Nominalskalavariabel i regresjon (3)

Referansekategori (den utelatne hjelpevariabelen)

- Det bør velgast ein stor og eintydig definert kategori som referansekategori
- Den estimerte effekten av inkluderte hjelpevariablar måler effekten av å vere i den inkluderte kategorien relativt til å vere i referansekategorien

Fall 2004

© Erling Berge 2004

16



## Nominalskalavariabel i regresjon (4)

Dette tyder at

- Regresjonsparameteren for ein inkludert dummykoda hjelpevariabel fortel om tillegg eller frådrag i forventta Y-verdi som personen i får ved å vere i denne kategorien heller enn i referansekategorien

## Nominalskalavariabel i regresjon (5)

Testing I

- Test av om regresjonskoeffisienten for ein inkludert hjelpevariabel er ulik 0 gir svar på om dei personane som er i denne gruppa har ein gjennomsnittleg Y verdi som er ulik gjennomsnittet til personane som er i referansekategorien

## Nominalskalavariabel i regresjon (6)

### Testing II

- Test av om Nominalskalavariabelen yter signifikant til regresjonsmodellen samla sett må skje ved å teste om alle hjelpevariablane under eitt yter til regresjonen
- Vi kan da bruke F-testen, enten etter formelen 3.28 hos Hamilton (s.80) eller 3.3 hos Hardy (s.24). Desse er formelt identiske

## Nominalskalavariabel i regresjon (7)

### Interaksjon

- Når dummy-koda nominalskalavariablar inngår i interaksjonar må alle inkluderte hjelpevariablar multipliserast med den variabelen ein trur den har interaksjon med

## Litt terminologi (1)

- **Dummykoding** av nominalskalavariabeler vert gitt litt ulike namn i ulike lærebøker. Det vert kalla
  1. Dummykoding av Hamilton, Hardy og Weisberg
  2. Indikatorkoding av Menard (og Weisberg)
  3. Referansekoding eller partialmetoden hos Hosmer&Lemeshow

Fall 2004

© Erling Berge 2004

21

## Litt terminologi (2)

- For å reprodusere resultat frå **variansanalyse** ved hjelp av regresjon introduserer Hamilton ei koding av hjelpevariablane han kallar **effektkoding**. Hardy (kap5) kallar det også effektkoding. Hos andre vert det kalla
  1. Avvikskoding av Menard, eller
  2. Marginalmetoden eller avviksmetoden av Hosmer&Lemeshow
- For å få fram særlege gruppesamanlikningar introduserer Hardy (kap5) ein kodemåte ho kallar **kontrastkoding**

Fall 2004

© Erling Berge 2004

22

## Ordinalskala variable

- kan inkluderast som intervallskala dersom den underliggjande teoretiske dimensjonen er kontinuerleg og avstandsmål er fornuftige
- elles kan dei nyttast direkte som y-variabel dersom estimeringsprogrammet er konstruert for det
  - parametrar vert da estimert for kvart nivå over det lågaste som kumulative effektar i høve til lågaste nivå

Fall 2004

© Erling Berge 2004

23

## Nominalskalavariabel

<b>POPULATION TYPE</b>	Frequency	Percent	Valid Percent	Cumulative Percent
POL	48	12.6	12.6	12.6
FARMER	132	34.7	34.7	47.4
POP	200	52.6	52.6	100.0
<b>Total</b>	<b>380</b>	<b>100.0</b>	<b>100.0</b>	

Fall 2004

© Erling Berge 2004

24

## Eksempel på dummykoding

Nominalskala			Hjelpe- variable			
Befolkningsgruppe	Kode	N	H(1)= Pol	H(2)= Farmer	H(3)= People	
Politikarar	1	48	1	0	0	
Bønder	2	132	0	1	0	
Vanleg folk	3	200	0	0	1	Referansekategori

Ein variabel med 3 kategoriar gir opphav til 2 dummykoda variablar i regresjonen med den tredje som referanse

Fall 2004

© Erling Berge 2004

25

## Eksempel på effektkoding

Nominalskala			Hjelpe- variable			
Befolkningsgruppe	Kode	N	H(1)= Pol	H(2)= Farmer		
Politikarar	1	48	1	0		
Bønder	2	132	0	1		
Vanleg folk	3	200	-1	-1		Referansekategori

I effektkodinga får referansekategori den koden -1.  
 Effektkodinga gjer det mogeleg å duplisere alle F-testar i vanleg ANOVA analysar

Fall 2004

© Erling Berge 2004

26

## Kontrastkoding

- Blir nytta til å få fram nett dei samanlikningane som har størst teoretisk interesse.
- Kontrastkoding krev
  - Med J kategoriar må det lagast J-1 kontrastar
  - Kodeverdiane på kvar hjelpevariabel må summere seg til 0
  - Kodeverdiane på to vilkårlige hjelpevariablar må vere ortogonale (vektorproduktet er lik 0)

Fall 2004

© Erling Berge 2004

27

## Bruk av dummykoda variablar (1)

Dependent Variable: l. of political contr. of sales of agric. est.	B	Std. Error	Beta	t	Sig.
(Constant)	4.106	.152		26.991	.000
Pol	.914	.337	.147	2.711	.007
Farmer	.421	.240	.096	1.758	.080

- Konstanten viser her gjennomsnitt på avhengig variabel for referansekategorien
- Gjennomsnittet for politikarar er 0.91 meiningsskårepoeng over gjennomsnittet for referansekategorien
- Gjennomsnittet for bønder er 0.42 meiningsskårepoeng over gjennomsnittet for referansekategorien

Fall 2004

© Erling Berge 2004

28

## Bruk av dummykoda variablar (2)

Dependent Variable: I. of political control of sales of agricultural estates	B	Std. Error	t	Sig.
(Constant)	4.264	.186	22.954	.000
Number of dekar land Owned	.000	.000	2.176	.030
Pol	.566	.382	1.482	.139
Farmer	-.309	.338	-.913	.362

Samanlikn denne tabellen med den førre. Kva har endra seg?

Korleis tolkar vi koeffisientane for "Pol" og "Farmer"?

## Konklusjon fra kapittel 3 (1)

- Lineær regresjon kan lett utvidast til å nytte 2 eller fleire forklaringsvariablar.
- Dersom føresetnadene for regresjonen (at feilledda er uavhengige og identisk normalfordelte – "normal i.i.d. errors) er oppfylt, er regresjon eit allsidig og kraftig analyseverktøy for å studere samanhengen mellom fleire uavhengige variablar og ein avhengig

## Konklusjon fra kapittel 3 (2)

- Den vanlegaste metoden for å estimere koeffisientar kallast OLS (minste kvadratsum metoden)
- Koeffisientar rekna ut i utvalet er estimat av tilsvarande populasjonsverdiar.
- Vi kan gjennom t-testen vurdere kor sikker eit koeffisient estimat er.
- Gjennom F-testen kan vi vurdere fleire koeffisientestimat under eitt

Fall 2004

© Erling Berge 2004

31

## Konklusjon fra kapittel 3 (3)

- Dummyvariablar er nyttige på fleire måtar
  - Ein einsleg dummykoda x variabel vil gi oss ein test på skilnad i gjennomsnitt for to grupper (0 og 1 gruppene)
  - Nominalskalavariablar med fleire enn 2 kategoriar kan omkodast ved hjelp av dummykoding og inkluderast i regresjonsanalysar
  - Ved effektcoding vil vi kunne gjere variansanalyse av ANOVA-typen

Fall 2004

© Erling Berge 2004

32



## Hamilton vs Hardy 1993 (1)

1. Merk at symbolbruken hos Hardy er litt annleis enn hos Hamilton, t.d. vert feilledda kalla "u" og ikkje "e", det vert nytta store latinske bokstavar for utvalsparametrane og k står for talet av variablar. Hos Hamilton er K talet av parametarar
2. I lista over føresetnader for regresjonen (s5) har Hardy sitt krav nr 4 formelt same funksjon som Hamilton sitt krav nr 1. Det må formulerast slik når x-variablane ikkje er faste ("fixed X") men kan vere stokastiske variablar. Det sikrar at feilleddet og inkluderte x-variablar er ukorrelerte (dvs at modellen teknisk sett er korrekt)

Fall 2004

© Erling Berge 2004

33

## Hamilton vs Hardy 1993 (2)

- Formel 3.1 på side 23 viser ein t-test for skilnaden mellom to regresjonskoeffisientar.
- Vi finn der uttrykket

$$t = \frac{(B_j - B_k)}{[\text{var}(B_j) + \text{var}(B_k) - 2\text{cov}(B_j, B_k)]^{1/2}}$$

Fall 2004

© Erling Berge 2004

34

## Hamilton vs Hardy 1993 (3)

- $[\text{var}(B_j) + \text{var}(B_k) + \text{cov}(B_j, B_k)]^{1/2}$  dvs kvadratrota av det som står i parentesen [ ]
- $\text{var}(B_j) = [\text{SE}(b_j)]^2$  dvs kvadratet av standardfeilen til  $b_j$  i Hamilton
- $\text{cov}(B_j, B_k) =$  er lik kovariansen mellom to parametrar. Vi får fram denne i SPSS ved å be om "Covariance matrix" i "Statistics"-opsjonen i regresjonsprosedyren
- Formel 4.5 på side 40 viser ein analog test for samla effekt av to variablar

Fall 2004

© Erling Berge 2004

35

## Hamilton vs Hardy 1993 (4)

- Formel 3.3 side 24 viser ein F-test for skilnaden mellom to modellar basert på determinasjonskoeffisienten og talet på variablar i modellane. Formelt er testen identisk med den Hamilton har på side 80. I valet mellom dei to formlane kan effekten av avrundingsfeil i Determinasjonskoeffisientane vere eit viktig moment

$$F = \frac{(R_3^2 - R_1^2) / (k_3 - k_1)}{(1 - R_3^2) / (N - k_3 - 1)}$$

Fall 2004

© Erling Berge 2004

36

## Nye moment hos Hardy 1993 (1)

- På sidene 53-56 og 60-61 vert det presentert testar for heteroskedastisitet
  - Goldfeld-Quandt testen
  - Levene's test
  - Gleijser's test
    - består i regresjon av residualen sin absoluttverdi  $|e_i|$  på kvar einskild forklaringsvariabel (kvar einskild  $x$ )
    - **NB:** For nominalskalavariabel måtte då alle hjelpevariablane inkluderst samtidig
    - Den er nært beslekta med Levene's test

Fall 2004

© Erling Berge 2004

37

## Nye moment hos Hardy 1993 (2)

- På siden 56-59 drøftast tolkinga av regresjonskoeffisientane til dummyvariablar når den avhengige variabelen er transformert med naturleg logaritme dvs  $y^* = \ln [\text{opphaveleg } y]$
- Den prosentvise endringa i opphaveleg  $y$  som kan seiast å komme av at ein er i kategori 1 heller enn i kategori 0 er gitt ved formel 4.10
- $100[\exp\{b_j\} - 1]$
- Lineær regresjon med  $y^* = \ln [\text{opphaveleg } y]$  som avhengig variabel vert kalla ein **semi-logaritmisk modell**

Fall 2004

© Erling Berge 2004

38

## Nye moment hos Hardy 1993 (3)

- På s61 drøftast problemet med val av signifikansnivå i **multiple samanlikningar**. Om Bonferroni ulikskapen sjå også note 16 til kapittel 4 i Hamilton.
- På s78 drøftast **testing av kurvelinearitet** Ein metode er å bruke dummyvariable (td definert over utdanningsgrupper i staden for å tru på linearitet over år utdanning)
- På s80 visest korleis dummyvariable kan nyttast til å lage "**bitvis**" **regresjon**. Dette er nyttig når vi har visse terskelverdiar for x der vinkelkoeffisienten for regresjonslinja endrar seg drastisk (td i studiet av viktige historiske hendingar i tidsseriedata)