

SOS3003

Eksamensoppgåver

Gjennomgang våren 2004

Erling Berge

Gjennomgang av

- Oppgåve 2 gitt hausten 2001

Haust 2001

OPPGÅVE 2 tabellvedlegget til oppgave 2 er det estimert 16 modellar av eiga inntekt (E.inntekt).

- a) Ta utgangspunkt i modell 1 og forklar med ord kva effekten av "E.utdanning" tyder. Finn så i same modellen eit konfidensintervall for effekten av "E.utdanning". Finn ut frå modell 4 forventa inntekt for ei 40 år gammal kvinne med 12 års utdanning og heiltidsarbeid i Kværner Oil&Gas. Test om det er ein lineær eller kurvelineær samanheng mellom alder og inntekt.
- b) Formuler den modellen som er estimert som Modell 8. Vurder om testane i modell 8 er truverdige. Drøft mogelege forbetringar av modellspesifikasjonen.
- c) Test om "Kjelde til livsopphald" gir ei signifikant yting til å forklare variasjonen i inntekt. Finn ved hjelp av modell 12 eit estimat av skilnaden i inntekt mellom menn og kvinner når ein føreset at dei er funksjonærar i heiltidsarbeid, har 15 års utdanning og arbeider i Direktoratet for naturforvaltning. Ta utgangspunkt i modell 16 og skriv ut formelen for eit betinga effekt plott for samanhengen mellom alder og inntekt for personar med heiltidsarbeid, 15 års utdanning og arbeid i offentleg sektor.
- d) Modell 15 er identisk med modell 8, men er estimert utan dei 2 personane som har størst verknad på estimatet av modell 8. Kva kan seiast om desse personane? Kva konsekvensar har det for regresjonsresultatet at personane vert utelatne?

Vår 2004

3

Haust 2001 Oppgave 2a

- **Ta utgangspunkt i modell 1 og forklar med ord kva effekten av E.utdanning (E.utd) tyder. Finn så i same modellen eit konfidensintervall for effekten av E.utdanning. Finn ut frå modell 4 forventa inntekt for ei 40 år gammal kvinne med 12 års utdanning og heiltidsarbeid i Kværner Oil&Gas. Test om det er ein lineær eller kurvelineær samanheng mellom alder og inntekt.**

Vår 2004

4

Haust 2001 Oppgåve 2a-1

- Modell 1 viser at for kvart år ekstra utdanning ein person får vil inntekta auke med $5,25 \cdot \text{kr } 1000 = \text{kr } 5.250$ om alt anna er likt

Haust 2001 Oppgåve 2a-2

- Dersom vi kan gå ut frå at feilledda er normalfordelte vil eit 95% konfidensintervall (5% signifikansnivå) vere gitt ved

$$b_{E. utd} - SE_{E. utd} \cdot t_{5\%} < \mathbf{b_{E. utd}} < b_{E. utd} + SE_{E. utd} \cdot t_{5\%}$$

Set vi inn i formelen finn vi at

- $5,25 - 0,43 \cdot 1,96 < b_{E. utd} < 5,25 + 0,43 \cdot 1,96$
 $5,25 - 0,85 < b_{E. utd} < 5,25 + 0,85$
 $4,398 < \mathbf{b_{E. utd}} < 6,102$

Haust 2001 Oppg ve 2a-3

Variabel	Variabelverdi	Multiplisert med	Parameter Estimert	Gir resultatet	
Konstant			-87,86921	-87,86921	
Alder	40	*	4,8479342	193,91737	
Alder*Alder	40*40	*	-0,045279	-72,44675	
Mann	0	*	-71,52876	0	
E,utdanning	12	*	6,9131003	82,95720	
Heiltidsarbeid	1	*	75,334491	75,33449	
Offentleg sektor	0	*	-6,776133	0	
Alder*Mann	40*0	*	5,0725196	0	
Alder*Alder*Mann	40*40*0	*	-0,04956	0	
Forventa verdi av E,inntekt				191,89310	

V r 2004

7

Haust 2001 Oppg ve 2a-4 (1)

- I modell 1 er ikkje alder inkludert som variabel, i modell 2 er alder inkludert line rt og i modell 3 er alder inkludert som kurveline r med eit andregradspolynom.
- Gir alder inkludert line rt eit signifikant bidrag til modellen?
- Er alder inkludert kurveline rt signifikant betre enn den line re modellen?

V r 2004

8

Haust 2001 Oppgåve 2a-4 (2)

- F-test av kvadratleddet aleine gir ikkje meir kunnskap enn det t-testen av koeffisienten for kvadratleddet gir
- Begge aldersparametrane er signifikant ulik null
- Generelt kan ein seie at dersom begge parametrane i eit alderspolynom er signifikant ulik 0, vil heile polynomet yte signifikant til forklaringa av variasjonen i den avhengige variabelen. Meir omfattande testing trengst berre dersom ein av parametrane ikkje er signifikant ulik 0.

Haust 2001 Oppgåve 2b

- **Formuler den modellen som er estimert som Modell 8. Vurder om testane i modell 8 er truverdige. Drøft moglege forbetringar av modellspesifikasjonen.**

Haust 2001 Oppgåve 2b-1 (1)

Når vi skal formulere ein modell må vi

- definere elementa som inngår i modellen (variablar og datamaterialet)
- definere relasjonane mellom elementa (regresjonslikninga), og
- presisere kva føresetnader som ein må gjere for å bruke modellen (spesifikasjon av modell, Gauss-Markov krav, normalfordelte feilledd)

Haust 2001 Oppgåve 2b-1 (2)

definere relasjonane mellom elementa i modellen

- I populasjonen føreset vi at det er eit lineært eller kurvelineært samband mellom den avhengige variabelen, Y , og dei uavhengige X -variablane. Dette tyder i vårt høve at når vi lar i gå over heile populasjonen vil

- $$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \beta_6 X_{6i} + \beta_7 X_{7i} + \beta_8 X_{8i} + \beta_9 X_{9i} + \beta_{10} X_{10i} + \beta_{11} X_{11i} + \beta_{12} X_{12i} + \beta_{13} X_{13i} + \beta_{14} X_{14i} + \varepsilon_i$$

Haust 2001 Oppgåve 2b-2

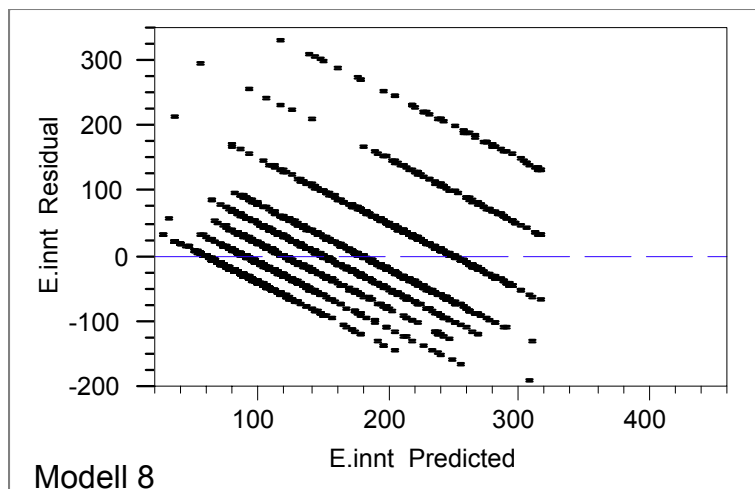
Kva påverkar truverdet av testar?

- Autokorrelasjon?
 - Heteroskedastisitet?
 - Ikkje-normal feilledd?
-
- Autokorrelasjon er lite sannsynleg i ein landsomfattande survey

Vår 2004

13

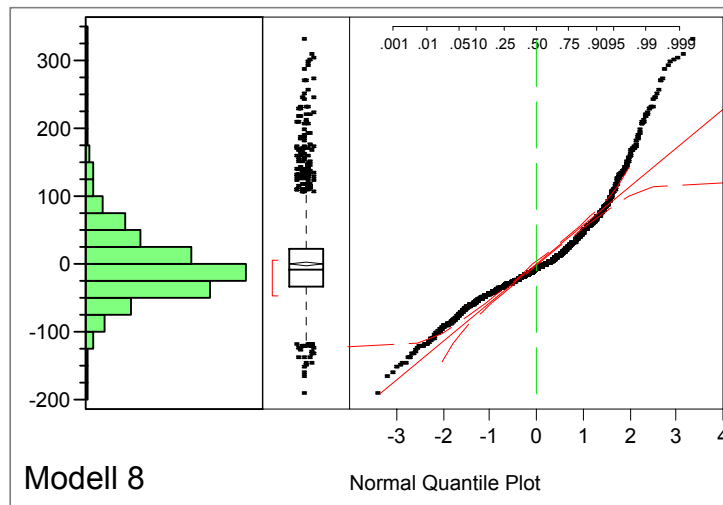
Haust 2001 Oppgåve 2b-2 (2) Heteroskedastisitet?



Vår 2004

14

Haust 2001 Oppgåve 2b-2 (3) Normalfordeling av residualen?



Vår 2004

15

Haust 2001 Oppgåve 2b-2 (4) Konklusjon

- Samla sett vil ein vel seie at tendensen til heteroskedastisitet saman med avvika frå normalfordelinga (dette kan godt vere to sider av same fenomen) tilseier stor varsemd med å tru på testane i modellen.

Vår 2004

16

Haust 2001 Oppg ve 2b-3 (1)

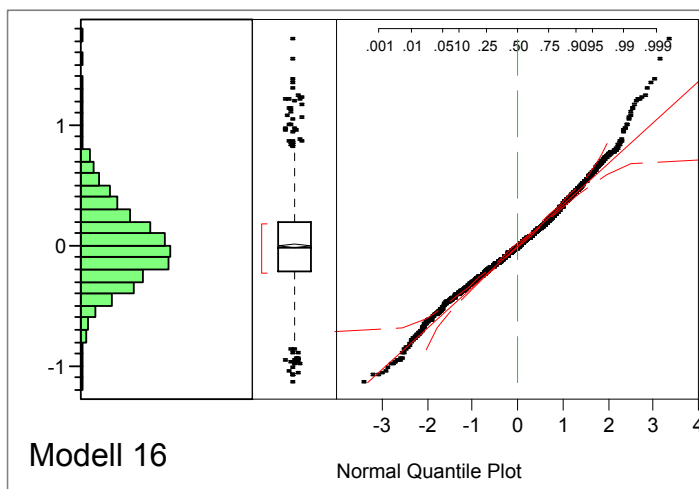
Betre modell gjennom

- Fleire variablar eller transformasjon av avhengig variabel?
- Inspeksjon av ulike modellar tyder p  at modell 16, der Y er transformert, er ei forbedring. Fordelinga av residualane er n rmare normalfordelinga

V r 2004

17

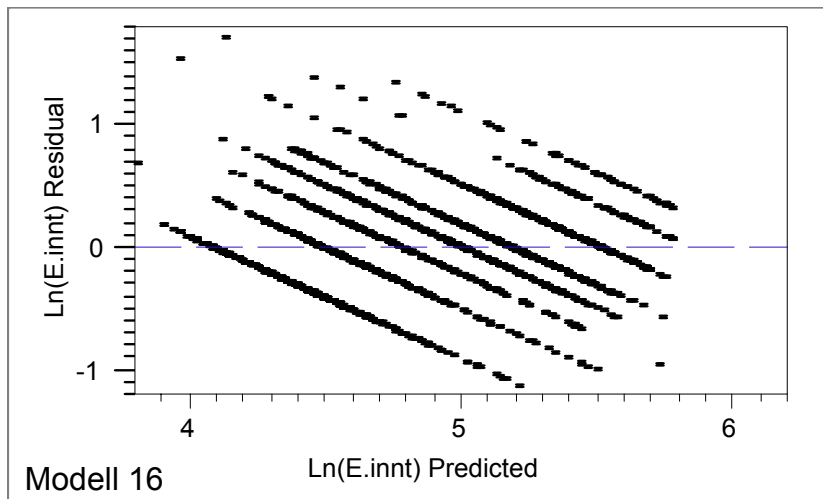
Haust 2001 Oppg ve 2b-3 (2) Kvantil-normal plottet av residualen



V r 2004

18

Haust 2001 Oppg ve 2b-3 (3) Residual plotta mot predikert y



V r 2004

19

Haust 2001 Oppg ve 2d

- Modell 15 er identisk med modell 8, men er estimert utan dei 2 personane som har st rst verknad p  estimatet av modell 8. Kva kan seiast om desse personane? Kva konsekvensar har det for regresjonsresultatet at personane vert utelatne?

V r 2004

20

Haust 2001 Oppgåve 2d-1 2 personar med påverknad i modell 8 (1)

Variabelnamn	Case No	1756	1803	Kommentar
Univ/ høyskole utd		1	0	1 = har akademisk tittel
Næring		6	2	6 = undervisning, forskning, 2 = varehandel, butikk
Ekteskapeleg status		1	2	1 = gift, 2 = aldri gift
Heiltidsarbeid		0	1	1=ja
Offentleg sektor		1	0	1=ja
Talet på arbeidstakarar i husst		2	1	#
Barn i hushaldet		0	0	#
Mors utdanning		7	7	1 år
Fars utdanning		7	15	1 år
E.utdanning		17	9	1 år
Alder		57	19	1 år

Vår 2004

21

Haust 2001 Oppgåve 2d-1 2 personar med påverknad i modell 8 (2)

Variabelnamn	Case No	1756	1803	Kommentar
Mann		0	1	1=ja
E.innt		450	450	1 1000 kroner
HH.inntekt		450	450	1 1000 kroner
Busadtype		2	1	1 = sentrum storby, 2 = forstad storby
Kjelde til livsopphald		4	2	4 = funksjonær (ikkje leiande), 2 = arbeidar faglært
Residual E.innt Model 8		305.481	331.047	
h(i) E.innt Model 8		0.01132	0.00969	
Cook's D(i) Influence E.innt Model 8		0.02193	0.02198	

Vår 2004

22

Haust 2001 Oppg ve 2d-2 Konsekvensane av   utelate 2 case (1)

Variabelnamn	Modell 8 Estimate	Modell 15 Estimate	Modell 8 parameter relativt til model 15
Konstant	-24.48258	-17.43017	1.404609
Alder	3.7200401	3.5891408	1.036471
Alder*Alder	-0.035721	-0.034789	1.026790
Mann	-73.53077	-81.07991	0.906893
E.utdanning	3.2611476	2.9084876	1.121252
Heiltidsarbeid	-105.2539	-115.5403	0.910971
Offentleg sektor	17.576729	16.916805	1.039010
Alder*Mann	3.3566592	3.5562293	0.943882

V r 2004

23

Haust 2001 Oppg ve 2d-2 Konsekvensane av   utelate 2 case (2)

Variabelnamn	Modell 8 Estimate	Modell 15 Estimate	Modell 8 parameter relativt til model 15
Alder*Alder*Mann	-0.027637	-0.029315	0.942760
E.utdanning*Mann	1.5145129	1.8116149	0.836002
Heiltidsarbeid*Mann	23.898241	22.476676	1.063246
Alder*Heiltidsarbeid	6.6503477	7.0310364	0.945856
Alder*Alder*Heiltidsarbeid	-0.07877	-0.082802	0.951306
E.utdanning*Heiltidsarbeid	5.0927931	5.3319909	0.955139
Off sektor*Heiltidsarbeid	-41.23729	-40.44984	1.019467

V r 2004

24

Haust 2001 Oppgåve 2d-2 Konklusjon

- Den substansielle forskjellen er stor både for Mann og for E.utdanning. Det synest rimeleg å sjå dette i lys av at dei to utelatte casa var ein 19 år gammal mann med låg utdanning og ei 57 år gammal kvinne med høg utdanning. Kombinasjonen av variabelverdiar gjer at dei to casa får relativt stor innverknad på regresjonsresultatet. Fordelinga av residualen og spreininga til residualen etter verdien av predikert Y har likevel ikkje endra seg påviseleg.