Linköpngs Universitet
IDA/Statistik
LH

TENTAMEN I SAMBANDSMODELLER, 2014-03-29

Skrivtid: kl: 8-13

Hjälpmedel: Räknedosa. Läroboken: *Applied linear statistical models* av Kutner, Nachtsheim m fl som inte får innehålla anteckningar men får ha markeringar och flärpar. Flärpar får ha en liten anteckning.

Jourhavande lärare: Lotta Hallberg och Tommy Schyman

Redovisa och motivera kort alla dina lösningar

Tolka (om möjligt) alla dina resultat!

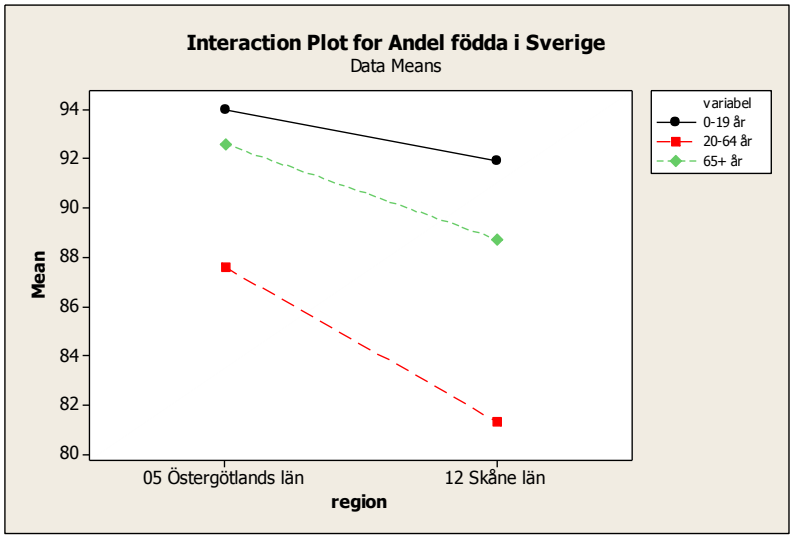
1

I denna uppgift ska andelen födda i Sverige i Sveriges befolkning studeras. Vi ska undersöka om det finns en skillnad i andelen födda i Sverige mellan Skånes län och Östergötlands län. Andra variabler som ska beaktas är ålderskategorier, och tid. Data är hämtade från SCBs hemsida.

Variabeldeklaration:

Variabelnamn	*Anm*	*Värden*
tid	år 1997 till 2011	1997; 1998; 1999; 2000; 2001; 2002; 2003; 2004; 2005; 2006; 2007; 2008; 2009; 2010; 2011
Dummy- variabel	Tre åldersgrupper	0-19 år; 20-64 år; 65+ år
region	länen	05 Östergötlands län; 12 Skåne län
0-19	dummy	1 om 0-19 år, 0 annars
20-24	dummy	1 om 20-24 år, 0 annars
Östergötland	dummy	1 om Östergötlands län, 0 om Skånes län
(0-19)*österg	interaktion	
(20-24)*österg	interaktion	
Andelen födda i Sverige		värden i procentenheter mellan 0 och 100

Följande graf kan vara till hjälp för att förstå data-materialet.



Först har två regressionsmodeller anpassats:

Modell 1

Regression Analysis:

```
The regression equation is
Andel födda i Sverige = 87,3 + 4,08 Östergötland

Predictor        Coef  SE Coef       T      P
Constant      87,3133   0,5987  145,84  0,000
Östergötland   4,0778   0,8467    4,82  0,000

S = 4,01620   R-Sq = 20,9%   R-Sq(adj) = 20,0%
```

Modell 2

Regression Analysis:

```
The regression equation is
Andel födda i Sverige = 88,6 + 4,08 Östergötland + 2,29 0-19 - 6,19 20-24

Predictor        Coef  SE Coef       T      P
Constant      88,6144   0,3704  239,26  0,000
Östergötland   4,0778   0,3704   11,01  0,000
0-19           2,2867   0,4536    5,04  0,000
20-24         -6,1900   0,4536  -13,65  0,000

S = 1,75685   R-Sq = 85,2%   R-Sq(adj) = 84,7%

Analysis of Variance
Source          DF       SS      MS       F      P
Regression       3  1528,12  509,37  165,03  0,000
Residual Error  86   265,44    3,09
Total           89  1793,56
```

a) Pröva om det är skillnad i andelen födda i Sverige mellan Skånes län och Östergötlands län med hjälp av modell 1. 1p
b) Pröva med ett partiellt F-test om dummy-variablerna för åldersgrupper kan läggas till modell 1. 1p
c) Tolka regressionskoefficienten för variabeln Östergötland i modell 2. 1p

Nedan följer en modell utökad med två interaktionstermer samt tidsvariabeln.

Modell 3

Regression Analysis:

```
The regression equation is
Andel födda i Sverige = 627 + 3,85 Östergötland + 3,16 0-19 - 7,40 20-24 +
                        - 1,75 (0-19)*österg + 2,42 (20-24)*österg - 0,269 tid

Predictor          Coef  SE Coef       T      P    VIF
Constant         627,42    47,51   13,21  0,000
Östergötland      3,8533   0,3548   10,86  0,000  3,000
0-19              3,1600   0,3548    8,91  0,000  2,667
20-24            -7,4000   0,3548  -20,86  0,000  2,667
(0-19)*österg    -1,7467   0,5018   -3,48  0,001  3,333
(20-24)*österg    2,4200   0,5018    4,82  0,000  3,333
tid            -0,26881  0,02371  -11,34  0,000  1,000
```

```
S = 0,971739   R-Sq = 95,6%   R-Sq(adj) = 95,3%

Analysis of Variance
Source          DF       SS      MS       F      P
Regression       6  1715,19  285,86  302,73  0,000
Residual Error  83    78,37    0,94
Total           89  1793,56
```

För observation nummer 75 i modell 3 har minitab beräknat

$h_{75,75} = 0{,}096$

Studentized deleted residul$_{75}$= -3,252

$DFFITS_{75} = -1{,}059$

Cook's Distance = 0,1435

Frågorna nedan gäller modell 3.

d) Hur många procentenheter sjunker andelen födda i Sverige varje år enligt modell 3? 1p

e) Prediktera andelen födda i Sverige i Skåne län i åldersgruppen 65+ år 2012. 1p

f) Prediktera med ett 95% prediktionsintervall 'Andelen födda i Sverige' i Östergötland i åldersgruppen 20-24 år 2012. Du får använda direkt att medelfelet för prediktionen $\hat{y}$ är 0,315. 2p

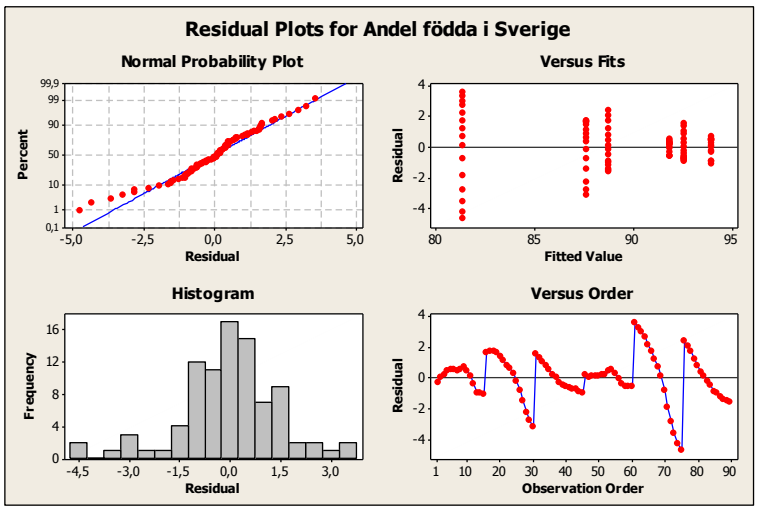
g) De fyra graferna nedan gäller för residualerna från modell 3. Utför analys av residualerna. Är modellen godkänd? 1p

h) Förklara riskerna med multikollinjäritet. Finns det någon risk för multikollinjäritet i modell 3. 1p

i) Ta hjälp av de fyra måtten ovan för att undersöka om, och vilken typ av outlier observation nummer 75 är. 2p

j) Om variabeln tid tas bort från modell 3 så kan modell 3 skrivas om till en tvåvägs ANOVA. Skriv upp modell 3 i denna form och beskriv faktorerna och dess nivåer. Kan man anta att alla antaganden för ANOVA-modellen är uppfyllda? 2p

k) ANOVA-modellen i uppgift j) är inte ett planerat experiment utan ett observerat försök. Förklara varför detta inte är (och heller inte kan göras som) ett planerat experiment. 1p



2

Anta att du har ett stickprov av storlek 15 från en slumpvariabel X som är $N(,1)$. Härled maximumlikelihood-skattningen av μ. 2p

3

Man är intresserad av attityden i en viss fråga från fyra olika politiska partier. Man valde därför 5 politiker slumpmässigt från vartdera partiet. De fick svara på en skala från 1-100 , där 100 är positivt inställd. Följande resultat erhölls:

Parti nr:	1	2	3	4
$\bar{y}_i$	85	80	95	50
s_i	6	7	4	10

a) Sätt upp en envägs-variansanalysmodell och pröva om det är skillnad mellan partierna. Visa att MSE=50.25. 2p

b) Använd Tukey's metod för att pröva vilka partier som skiljer sig åt. Simultan (familje) signifikansnivå är 5%. 2p