Linköpings Universitet
IDA/STIMA-Statistik
LH



# TENTAMEN I SAMBANDSMODELLER, 2017-03-30

**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

**Betygsgränser:** För godkänt krävs minst 12 av 20 poäng och för väl godkänt krävs minst 16 av 20 poäng.

Redovisa och motivera kort alla dina lösningar

---

Tolka (om möjligt) alla dina resultat!

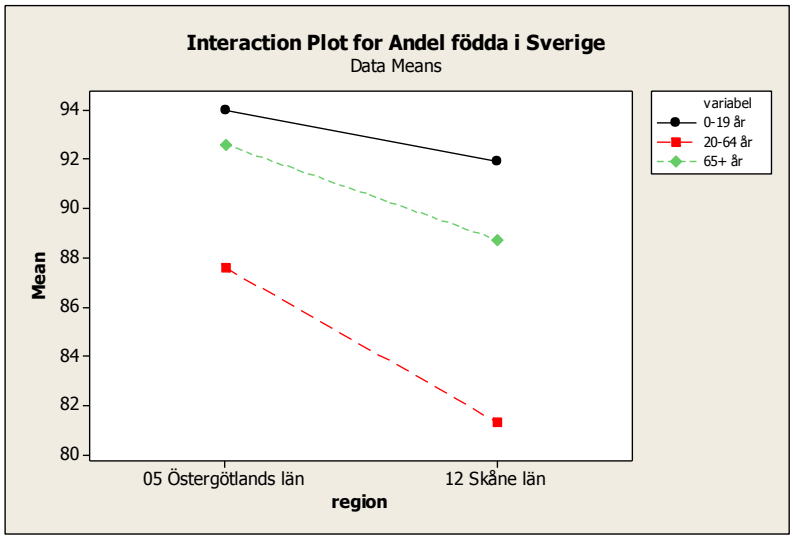
## 1

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

Term             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

Term             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
```

Frågorna nedan gäller modell 3.

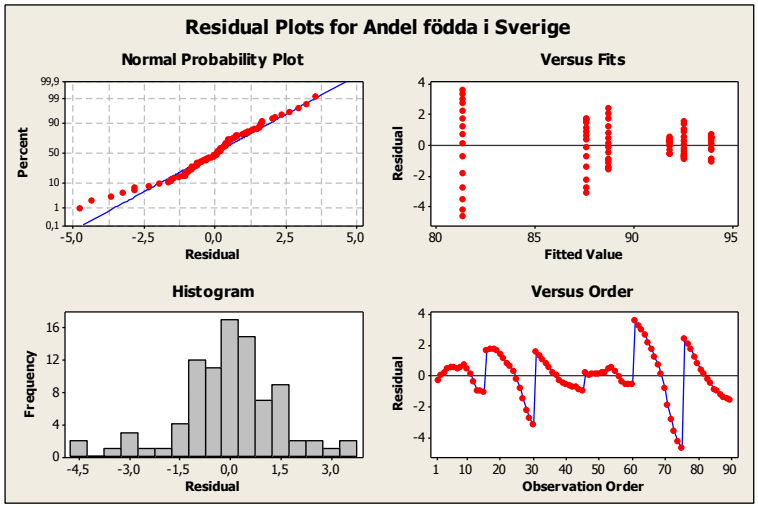
d) Hur många procentenheter sjunker andelen födda i Sverige varje år enligt modell 3? 1p

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 0-19 år 2013. 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? 2p

h) Det finns ju sex tydliga grupper. Vilken grupp har störst spridning? 1p




# 2

Vid ett löpande band produceras en viss typ av kretskort. Vid en kvalitetskontroll förfar man på följande vis: Vid en viss tidpunkt så undersöker man kretskort och räknar antalet kort tills man får en med någon defekt. Den defekta inkluderad i antalet. Detta görs 15 gånger. Resultat:

2 1 35 43 51 46 3 36 38 5 51 7 17 65 25

Låt $X$ vara detta antal som är beskrivet ovan. $X$ är då en geometriskt fördelad slumpvariabel med parameter $\pi > 0$. Sannolikhetfunktionen är given av

$$f(x) = (1-\pi)^{x-1}\pi, \quad x = 1,2,\ldots$$

Maximum-Likelihood skatta $\pi$. Sätt även in observationerna i skattningen. 2p

# 3

Vid ett laboratorium har man många kemister anställda. Man vill undersöka om det de mäter märkbart olika. Därför väljer man slumpmässigt ut fyra kemister och de har fått i uppgift att bestämma procentuella halten metyl-alkohol i en viss kemisk sammansättning. Varje kemist gör tre mätningar.

Resultat:

| Kemist | | | |
|---|---|---|---|
| 1 | 85,0 | 84,0 | 84,4 |
| 2 | 85,1 | 85,2 | 84,9 |
| 3 | 84,7 | 84,5 | 85,2 |
| 4 | 84,2 | 84,1 | 84,6 |

SSE= 0,9533

a) Sätt upp en lämplig modell och testa på lämpligt sätt med ett test om alla anställda kemister på laboratoriet mäter olika. Signifikansnivå 10%. Visa hypoteserna. 3p

b) Beräkna ett 95% konfidensintervall för det förväntade värdet på procentuella halten metyl-alkohol bland alla mätningar som görs. 2p

# 4

Anta att du är en forskare som är intresserad av att undersöka effekten rökning och vikt hos en person har på vilopulsen. Vilopulsen är kategoriserad låg och hög.
Följande anpassning har gjorts i Minitab.

Binary Logistic Regression: Vilopuls versus Rökning, Vikt

```
Link Function: Logit

Response Information
Variable      Value  Count
Vilopuls      Låg       70  (Event)
              Hög       22
              Total     92

Factor Information
Factor  Levels  Values
Rökning      2  Ja, Nej

Logistic Regression Table
                                                 Odds     95% CI
Predictor       Coef    SE Coef      Z      P  Ratio  Lower  Upper
Constant    -1.98717    1.67930  -1.18  0.237
Rökning
 Ja         -1.19297   0.552980  -2.16  0.031   0.30   0.10   0.90
Vikt       0.0250226  0.0122551   2.04  0.041   1.03   1.00   1.05
```

a) Tolka de båda oddskvoterna i utskriften. 2p

b) Prediktera sannolikheten för låg vilopuls för en person som är rökare och väger 120 pounds med hjälp av modellen ovan. 1p