Linköpings Universitet
MAI/Statistik
LH/OE

Tentamen i Sambandsmodeller, 2008-04-01

Skrivtid: kl: 8-12

Hjälpmedel: Räknedosa. Läroboken: *Applied linear statistical models* av Kutner, Nachtsheim m fl som får innehålla anteckningar. Alla upplagor tillåtna. Tre formelblad

Jourhavande lärare: Lotta Hallberg kl 8-10, Olle Eriksson, Stig Danielsson

Redovisa och motivera kort alla dina lösningar

Obs! Skriv namn och personnummer på varje papper du lämnar in.

1

Man vill undersöka om det finns ett linjärt samband mellan antalet sjukdagar Y i ett företag och antalet anställningsår x.
Ett stickprov av storleken 6 drogs bland de anställda och följande resultat erhölls

$$\mathbf{Y}=\begin{pmatrix}10\\8\\2\\0\\7\\2\end{pmatrix} \qquad \mathbf{X}=\begin{pmatrix}1&1\\1&4\\1&4\\1&8\\1&2\\1&5\end{pmatrix}$$

där X är designmatrisen.
Sätt upp den enkla linjära regressionsmodellen $\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$ där alla ε kan antas vara normalfördelade med väntevärde 0 och med känd varians $\sigma^2=5$. $\boldsymbol{\beta}=(\beta_1,\beta_2)'$ är parametervektorn. Efter uppgifterna hittar du uträknat **H, HY** och $(\mathbf{X}'\mathbf{X})^{-1}$.

a) Skatta parametrarna i $\boldsymbol{\beta}$ med minsta kvadratmetoden samt bilda 95% KI för dem. 2p

b) Visa att variansen på residualvektorn **e** är $\sigma^2(\mathbf{I}-\mathbf{H})$, där $\mathbf{H}=\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ 1p

c) Beräkna de studentiserade residualerna. 2p

d) Hur stor är korrelationen mellan parameterskattningarna? 1p

$$\mathbf{H}=\begin{pmatrix}0{,}47&0{,}17&0{,}17&-0{,}23&0{,}37&0{,}07\\0{,}17&0{,}17&0{,}17&0{,}17&0{,}17&0{,}17\\0{,}17&0{,}17&0{,}17&0{,}17&0{,}17&0{,}17\\-0{,}23&0{,}17&0{,}17&0{,}70&-0{,}10&0{,}30\\0{,}37&0{,}17&0{,}17&-0{,}10&0{,}30&0{,}10\\0{,}07&0{,}17&0{,}17&0{,}30&0{,}10&0{,}20\end{pmatrix} \qquad \mathbf{HY}=\begin{pmatrix}9{,}03\\4{,}83\\4{,}83\\-0{,}77\\7{,}63\\3{,}43\end{pmatrix}$$

$$(\mathbf{X}'\mathbf{X})^{-1}=\begin{pmatrix}0{,}70&-0{,}13\\-0{,}13&0{,}03\end{pmatrix}$$

2

I problem 14.11 sid 626 i boken kan du läsa om en undersökning där man vill skatta sambandet mellan pantens storlek och antalet burkar som lämnas tillbaka. Följande modell kördes i SAS.

```
proc logistic descending;
  model y/n=x ;
run;
```

Default är logitlänk.

Programmet genererade bl a utskriften:

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.0763	0.0848	599.0316	<.0001
x	1	0.1358	0.00477	810.3483	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
x	1.145	1.135	1.156

a) Sätt upp modellen och med de skattade parametrarna inskrivna samt ange vilka krav man ska ställa på modellen. 1p

b) Skatta sannolikheten för återlämning av burkar då panten är 6 cent (motsvarar ungefär vår pant på 50öre). 1p

c) Tolka oddskvoten samt dess konfidensintervall. 1p

d) Om en enkel linjär regressionsmodell anpassa direkt till de 6 värden så fås resultatet. Går detta samband att förstå och hur skulle man i så fall tolka det. Utred. 1p

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	35.77098	11.20629	3.19	0.0332
x	1	13.92798	0.60567	23.00	<.0001

3. (5 poäng) Man vill studera miljön i bottensediment i en sjö. Sjön kan delas upp i många olika områden. I den här studien har man inte resurser att kolla alla sådana områden och därför väljs ett slumpmässigt urval om 3 områden. Från vart och ett av dessa områden tar man upp ett stort prov av bottensediment och lägger i var sitt akvarium. I varje akvarium placeras sedan 4 maskar av en viss sort, och man är noga med att alla 12 maskar är lika gamla och på alla sätt lika behandlade fram till nu. Det man till slut observerar är tiden tills maskarna dör. Snabbare död tyder på att miljön är mer ogästvänlig för den här sortens maskar. Datamaterialet visas här, där tiden är maskarnas livslängd i dygn efter att de placeras i akvarierna:

Område nr:	1	2	3
Observationer:	20 20 29 23	33 25 24 26	36 30 40 30

(a) Testa på 5% risknivå
H_0: Förväntad livslängd är lika stor i alla områden.
H_1: Förväntad livslängd är inte lika stor i alla områden.

(b) Beräkna en punktskattning av σ_μ^2 och en punktskattning av σ^2 .

(c) Uppdragsgivaren var inte nöjd med bara den punktskattninga av σ^2 som beräknades i den föregående deluppgiften. Beräkna istället ett 95% konfidensintervall för σ^2 .

(d) Skatta μ. med 95% konfidensintervall.

4. (5 poäng) I en annan sjö har man gjort en liknande studie, men förutsättningarna där är annorlunda. Även där har man använt bottensediment från 3 områden, men den här sjön består grovt sett av just dessa tre områden. Man betraktar därför inte de områden som ingår i studien som något slumpmässigt stickprov. Man är också intresserad av att studera effekten av 2 olika temperaturer som valts så att de ska representera sommar- respektive vinterhalvår. Från varje område tar man därför ett stort prov av bottensediment som fördelas till två olika akvarier med olika temperatur. Även här studeras livslängd hos maskar och liksom i den förra uppgiften använder man 4 maskar i varje akvarium.

   Alla data visas här. De siffror som ges med fetstil är cell- eller marginalmedelvärden (avrundade till 2 decimaler) medan övriga siffror är de individuella maskarnas livslängd. De tre numrerade raderna markerar olika områden och de två numrerade kolumnerna markerar de två olika temperaturerna.

	1	2	
1	**19.75** 15 17 26 21	**12.25** 13 14 10 12	**16.00**
2	**12.25** 13 10 14 12	**11.50** 17 10 6 13	**11.88**
3	**19.25** 24 22 19 12	**13.50** 21 14 9 10	**16.38**
	17.08	**12.42**	**14.75**

   I fortsättningen kallas områdesfaktorn A och temperaturfaktorn B. Man tänker sig att en observation kan beskrivas som en summa av komponenter enligt
   $Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$.
   För att testa nollhypoteser om att effekterna är 0 kör man en vanlig variansanalys, och en utskrift visas här där delar av utskriften ersts med "--".

```
Analysis of Variance for Y

Source  DF      SS      MS      F       P
A       --   99.75      --      --      --
B       --  130.67      --      --      --
A*B     --   49.08      --      --      --
Error   --  325.00      --
Total   --  604.50
```

   Frågorna följer på nästa sida.

(a) Testa på 5% risknivå $H_0 : \sum_{i=1}^{3} \alpha_i^2 = 0 \quad , \quad H_1 : \sum_{i=1}^{3} \alpha_i^2 > 0$

(b) Man vill jämföra alla cellers väntevärden mot varandra i par på en gemensam 5% risknivå. Du ska visa att du kan utföra testen. Eftersom beräkningarna blir väldigt lika varandra så räcker det med att du genomför beräkningarna bara för testet $H_0 : \mu_{11} = \mu_{12} \quad , \quad H_1 : \mu_{11} \neq \mu_{12}$.

(c) Hur stor är styrkan i testet i a om $\alpha_1 = 1, \alpha_2 = -3, \alpha_3 = 2$ och $\sigma^2 = 16$?

(d) Skatta $\mu_{1.} - \mu_{2.}$ med 95% konfidensintervall.

(e) Beräkna svar på samma fråga som i deluppgift a, men nu under antagandet att en observation kan beskrivas som en summa av komponenter enligt
$Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$
d.v.s. att man har en modell utan interaktion.