STOCKHOLMS UNIVERSITET,
MATEMATISKA INSTITUTIONEN,
Avd. Matematisk statistik

Tentamen: Linjära statistiska modeller (MT5001), 2021-12-02

Kristoffer Lindensjö
E-post: kristoffer.lindensjo@math.su.se
Telefonnummer: 070 444 10 07

Tillåtna hjälpmedel: Miniräknare och formelblad (tillhandahålles av institutionen).

Återlämning: information meddelas via kursforum.

Tentamen består av 5 uppgifter. Varje korrekt löst uppgift ger 10 poäng.

- Resonemang ska vara klara, tydliga och kortfattade.
- Svar ska motiveras om inte annat framgår.
- Börja varje uppgift på nytt papper.
- Numrera tydligt varje blad med uppgift och bladordning.
- Skriv ditt kodnummer på varje blad du lämnar in (men inget namn).

Preliminära betygsgränser:

A	B	C	D	E
45	40	35	30	25

Vissa av följande kvantiler kan komma att bli användbara

$$
\begin{aligned}
t_{0.025}(200) &= 1.97190 \\
t_{0.025}(199) &= 1.97196 \\
t_{0.025}(198) &= 1.97202 \\
t_{0.025}(6) &= 2.44691 \\
t_{0.025}(5) &= 2.57058 \\
F_{0.05}(2, 200) &= 3.04106 \\
F_{0.05}(2, 199) &= 3.04129 \\
F_{0.05}(2, 198) &= 3.04152.
\end{aligned}
$$

Lycka till!

Uppgift 1

Lisa och Peter har data gällande omsättning mätt i miljoner SEK (Y_i) och antal anställda (x_i) för 200 företag i Sundsvall för året 2021. De analyserar data med hjälp av enkel linjär regression

$$Y_i = \widetilde{\alpha} + \beta(x_i - \bar{x}) + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2) \ \ iid.$$

Följande gäller

$$\sum_{i=1}^{200} x_i = 23900.00 \qquad \sum_{i=1}^{200} Y_i = 20043.52$$

$$\sum_{i=1}^{200} (x_i - \bar{x})^2 = 666650.00 \qquad \sum_{i=1}^{200} (x_i - \bar{x})(Y_i - \bar{Y}) = 335186.62.$$

Notera att kvantiler finns överst i tentamen.

(A) Beräkna MK-skattningarna $\widehat{\widetilde{\alpha}}$ (för $\widetilde{\alpha}$) och $\widehat{\beta}$ (för β). (2 p)

(B) Tolka kortfattat vad du kommit fram till i **(A)** utifrån den aktuella tillämpningen. (2 p)

Antag från och med nu att

$$\sum_{i=1}^{200} (Y_i - \widehat{\widetilde{\alpha}} - \widehat{\beta}(x_i - \bar{x}))^2 = 1818.10.$$

(C) Ange ett uttryck (icke-simultant) konfidensintervall (95%) för $\alpha + \beta x_0$, där vi som vanligt definierar $\alpha = \widetilde{\alpha} - \beta\bar{x}$. Förenkla uttrycket användandes information om data ovan. (3 p)

(D) Ange simultana konfidensgränser (95%) för linjen $\alpha + \beta x$. Förenkla på samma sätt som ovan. *Ledning: se definition av α ovan.* (3 p)

Uppgift 2

Badmintonrack klassifieras i tre olika viktfördelningskategorier och tre styvhetskategorier. Styvhetskategorierna är *styvt*, *halvstyvt* och *flexibelt*, och kategorierna för viktfördelning är *huvudtungt*, *balanserat* och *huvudlätt*. En badmintonklubb vill förstå hur rackets egenskaper påverkar slagkraften, och designar därför följande experiment: en erfaren spelare slår en fjäderboll så hårt som möjligt förbi två lasersensorer, som mäter hastigheten på fjäderbollen. Detta görs 10 gånger i rad för varje kombination av styvhet och viktfördelning. Totalt har man alltså 90 observationer.

Styvhet	Viktfördelning	Hastighet (km/h)
Styvt	Huvudtungt	389
Styvt	Huvudtungt	365
⋮	⋮	⋮
Halvstyvt	Balanserat	340
⋮	⋮	⋮

Table 1: Exempel på observationer i data.

(A) Formulera en tvåsidig variansanalysmodell av typ I utan samspel, för effekten av styvhet och viktfördelning på slagkraften. (3 p)

(B) Motivera varför en variansanalysmodell av typ I är lämplig. Du ska alltså motivera varför de båda faktorerna bör betraktas som systematiska. (1 p)

(C) Vi vill nu inkludera en samspelseffekt i modellen. Ange vad som krävs gällande antalet observationer i en cell för att vi ska kunna inkludera en samspelseffekt, samt hur många observationer vi har i varje cell för modellen i **(A)**. (2 p)

(D) Utveckla modellen i **(A)** till att inkludera en samspelseffekt. (1 p)

(E) Nedan ses två plottar av residualerna från den anpassade modellen med samspel, den övre mot predikterad respons och den undre mot observationsindex (löper från 1 till 90). Granska dessa och dra slutsatser utifrån vad som observeras. (3 p)

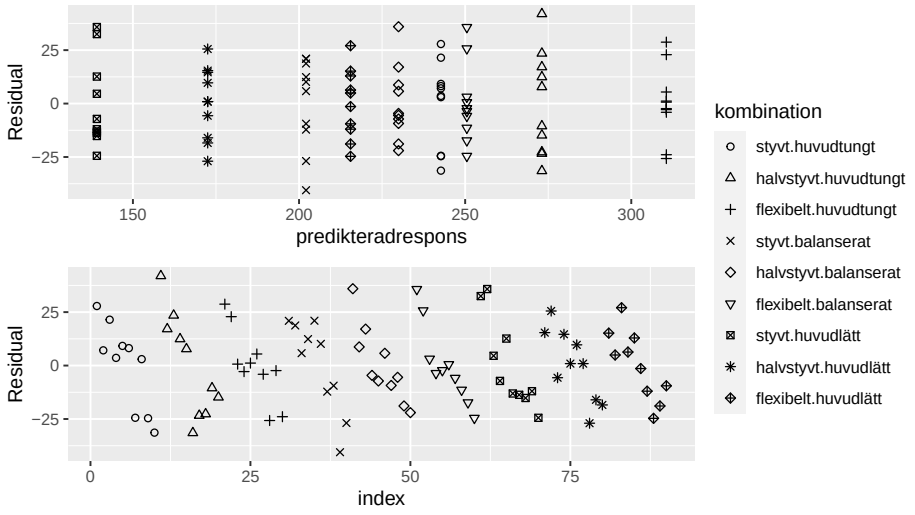

Figure 1: Residualplottar från den anpassade modellen med samspel.

Uppgift 3

Ett företag som säljer en mobilapp vill bestämma hur dess omsättning beror på priset på appen och företagets annonskostnader. Datan som finns tillgänglig visas i Tabell 2.

Kvartal	Omsättning y	Pris x_1	Annonskostnader x_2
1	8	39	6
2	4	45	0
3	5	45	4
4	6	42	5
5	4	46	4
6	4	47	3
7	7	42	5
8	2	48	3

Table 2: Omsättning, pris och annonskostnader för appföretaget.

Omsättningen anges i miljoner kronor och annonskostnaderna i 100 000-tal kronor. Man beräknar följande storheter:

$$\sum_{i=1}^{8}(y_i - \bar{y})^2 = 26,$$

$$\sum_{i=1}^{8}(x_{i1} - \bar{x}_{\cdot 1})^2 = 63.5, \qquad \sum_{i=1}^{8}(x_{i2} - \bar{x}_{\cdot 2})^2 = 23.5,$$

$$\sum_{i=1}^{8}(y_i - \bar{y})(x_{i1} - \bar{x}_{\cdot 1}) = -39, \qquad \sum_{i=1}^{8}(y_i - \bar{y})(x_{i2} - \bar{x}_{\cdot 2}) = 17,$$

$$\sum_{i=1}^{8}(x_{i1} - \bar{x}_{\cdot 1})(x_{i2} - \bar{x}_{\cdot 2}) = -24.5.$$

(A) Ansätt en multipel linjär regressionsmodell för omsättningen med pris och annonskostnader som förklarande variabler. (1 p)

(B) Skatta parametrarna i regressionsmodellen. (4 p)

(C) Ställ upp ett uttryck för att skatta residualtermens varians σ^2 (notera att beräkningarna inte behöver utföras). (1 p)

(D) Man räknar fram att $\hat{\sigma}^2 = 0.355$. Testa nollhypotesen att priset inte har någon signifikant effekt på omsättningen på signifikansnivån 95 procent. Gör även samma test för annonskostnaderna. (3 p)

(E) Nämn ett antagande för linjär regression samt beskriv kortfattad hur det antagandet skulle kunna verifieras. (1 p)

Uppgift 4

Betrakta den allmänna lineära modellen under normalfördelningsantagande, dvs

$$\mathbf{Y} = \mathbf{A}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_N)$$

där $\dim(\boldsymbol{\theta}) = k$. Låt $\boldsymbol{c}$ vara en fix vektor av dimension k.

(A) Vilken fördelning har $\boldsymbol{c}^T\widehat{\boldsymbol{\theta}}$ (*ditt svar behöver inte motiveras*)?
Härled följande konfidensintervall för $\boldsymbol{c}^T\boldsymbol{\theta}$,

$$\boldsymbol{c}^T\widehat{\boldsymbol{\theta}} \pm t_{p/2}(N-k)\hat{\sigma}\sqrt{\boldsymbol{c}^T(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{c}} \tag{1}$$

Ledning för den första delen av uppgiften: använd resultat ifrån formelbladet.
Ledning för den andra delen av uppgiften: använd ditt svar i första delen av uppgiften och resultat ifrån formelbladet.

(5 p)

(B) Härled ett konfidensintervall för interceptet $\widetilde{\alpha}$ i enkel linjär regression

$$Y_i = \widetilde{\alpha} + \beta(x_i - \bar{x}) + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2) \ \ iid, \qquad i = 1, ..., N.$$

Ledning: specificera $\boldsymbol{A}$, $\boldsymbol{\theta}$ och $\boldsymbol{c}$ på lämpligt vis, och använd konfidensintervallet (1) ovan; notera att det är möjligt att besvara den här uppgiften även om du inte löst uppgifterna ovan.

(5 p)

Uppgift 5

(A) Ange definitionen av en AR(2)-process och definitionen av en MA(2)-process. (4 p)

(B) Den här delen av uppgiften gäller Ljung-Box-testet som görs i samband med anpassning av exempelvis AR-, MA- eller ARMA-modeller, och vars test-statistika är:

$$Q = T(T+2)\sum_{1}^{K}\frac{r_k^2}{T-k}.$$

Ange för Ljung-Box-testet (i) test-statistikans fördelning och vad beteckningarna (dvs de olika bokstäverna) i test-statistikan står för, (ii) noll-hypotes och regeln för förkastande av denna, och (iii) normalfördelningsantagandets roll (ingen härledning eller exakta detaljer behövs).

Ledning: (1) börja med att fundera på vad du kan säga om fördelningen gällande alla r_k, och (2) om du känner till motsvarande Box–Pierce-test, så räcker det med den intuitiva förklaringen till detta. (6 p)