STOCKHOLMS UNIVERSITET,
MATEMATISKA INSTITUTIONEN,
Avd. Matematisk statistik

Tentamen: Linjära statistiska modeller (MT5001), 2021-10-28

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

$$
\begin{aligned}
t_{0.025}(50) &= 2.00856 \\
t_{0.025}(49) &= 2.00958 \\
t_{0.025}(47) &= 2.01174 \\
t_{0.025}(46) &= 2.01290 \\
t_{0.025}(44) &= 2.01537 \\
t_{0.025}(42) &= 2.01808 \\
t_{0.05}(44) &= 1.68023 \\
t_{0.05}(42) &= 1.68195.
\end{aligned}
$$

Lycka till!

Uppgift 1

Du har data gällande pris mätt i miljoner SEK (Y_i) och storlek mätt i kvadratmeter (x_i) för 44 lägenhetsförsäljningar i ett visst område i Göteborg under september 2021. Du bestämmer dig för att analysera data med hjälp av enkel linjär regression

$$Y_i = \widetilde{\alpha} + \beta(x_i - \bar{x}) + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2) \ \ iid, \qquad i = 1, ..., 44.$$

Enkla beräkningar ger

$$\sum_{i=1}^{44} x_i = 4092, \qquad \sum_{i=1}^{44} Y_i = 399,$$

$$\sum_{i=1}^{44} (x_i - \bar{x})^2 = 28380, \qquad \sum_{i=1}^{44} (x_i - \bar{x}) Y_i = 2602.$$

Notera att t-kvantiler finns överst i tentamen.

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

Antag från och med nu att

$$\sum_{i=1}^{44} (Y_i - \widehat{\widetilde{\alpha}} - \widehat{\beta}(x_i - \bar{x}))^2 = 50.54.$$

(B) Ange formeln för en väntevärdesriktig estimator $\widehat{\sigma}^2$ för σ^2 som är fördelad enligt $42\widehat{\sigma}^2/\sigma^2 \sim \chi^2(42)$. Beräkna motsvarande skattning (med hjälp av informationen om data ovan). (2 p)

(C) Bilda ett konfidensintervall (95%) för $\widetilde{\alpha} + \beta(x_0 - \bar{x})$ givet $x_0 = 100$. (2 p)

(D) Du har en lägenhet i området som är 100 kvadratmeter stor och vill använda modellen för att säga något om vad dess försäljningspris skulle kunna tänkas bli. Bilda ett prediktionsintervall (95%) för försäljningspriset. (2 p)

(E) Ge en kortfattad tolkning av intervallen i **(C)** och **(D)** utifrån den aktuella tillämpningen. Var noga med att inkludera varför intervallen skiljer sig. (2 p)

Uppgift 2

Carl-Magnus (C-M) på Start-up^2 AB kontaktar dig och vill att du bygger en AI som predikterar hur mycket pengar personer lägger på appar som kör hem snabbmat. Start-up^2 har data gällande 10 unika personer vardera från de fem appar som leverar mat. Av de 10 personerna per app är 5 i gruppen "20 år eller

yngre" och resterande 5 i gruppen "21 år eller äldre". Vi kallar dessa grupper "unga" och "äldre". Varje persons 15 senaste betalningar finns i data. Ett utdrag av data finns i tabellen nedan.

Sammanfattningsvis gäller alltså att: *App* har fem nivåer, *Åldersgrupp* har två nivåer, *Person* har 5 nivåer för varje kombination av app och åldersgrupp, *Belopp* är responsvariabel, och du har 15 observationer per person. Antag att det inte finns något beroende i tiden mellan betalningar, och att endast faktorerna *App*, *Åldersgrupp* och *Person* påverkar betalat belopp.

App	Åldersgrupp	Person	Belopp
App 1	Unga	Person 1	160
App 1	Unga	Person 1	220
⋮	⋮	⋮	⋮
App 3	Äldre	Person 5	272
App 3	Unga	Person 6	130
⋮	⋮	⋮	⋮

Table 1: Exempel på observationer i data.

(A) Bortse inledningsvis från *Person*. Formulera en tvåsidig variansanalysmodell av typ I utan samspel, där alltså *Åldersgrupp* och *App* är de enda faktorer som ingår och betraktas som systematiska, och *Belopp* är din responsvariabel. (3 p)

(B) Du anpassar den tvåsidiga variansanalysmodellen i **(A)** och utför ett F-test vardera för effekten av faktorerna *App* och *Åldersgrupp*. Formulera lämpliga nollhypoteser för dessa test. Antag att dessa nollhypoteser förkastas och förklara vilka slutsatser man kan dra av det. (2 p)

(C) Låt nu även *Person* ingå i variansanalysmodellen, som en slumpmässig faktor som är underordnad *App* och *Åldersgrupp* (kom ihåg från introduktionen att varje persons 15 betalningar hör till samma app). Formulera om modellen så att *Person* inkorporeras, och specificera antaganden du gör om faktorns effekter. (2 p)

(D) Du anpassar modellen i vilken *Person* ingår som en slumpmässig faktor, och får följande estimat: residualvarianskomponenten $\hat{\sigma} = 10.1$ och variankomponenten för *Person* $\hat{\sigma}_P = 23.9$. Dra en slutsats om vad som påverkar responsens variation från dessa estimat. (1 p)

(E) Du granskar de antaganden du presenterade om modellen i **(C)** . Nedanstående figur visar de estimerade slumpmässiga effekterna för person, uppdelat på unga och äldre. Ett antagande kanske inte är uppfyllt i tillfredsställande utsträckning; vilket och vad i figuren antyder detta? (2 p)

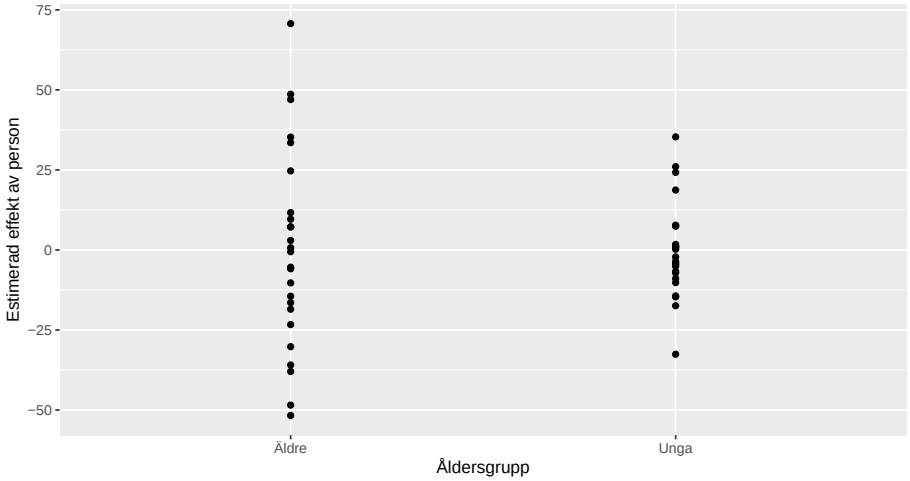

Figure 1: Estimat av de slumpmässiga effekterna av *Person* uppdelat per åldersgrupp.

Uppgift 3

En multipel linjär regressionsmodell på formen

$$Y_i = \theta_0 + x_{i1}\theta_1 + x_{i2}\theta_2 + x_{i3}\theta_3 + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2) \text{ iid}$$

anpassas till ett dataset genom att använda statistisk programvara. En sammanställning av resultatet som erhålls visas i följande tabell.

Parameter	Skattat värde	Medelfel	t-värde
θ_0	-401.22	162.05	-2.476
θ_1	211.74	42.04	5.036
θ_2	-57.84	36.87	-1.569
θ_3	65.88	44.07	1.495
- - -			
$\hat{\sigma}$: 149.7			
R^2: 0.3933			
F-värde ($H_0 : Y_i$ är konstant): 9.939 med 3 och 46 frihetsgrader			

Table 2: Sammanställning av resultat från multipel linjär regression.

(A) Härled antalet observationer i datasetet. (1 p)

(B) Beräkna den justerade förklaringsgraden R^2_{adj}. (2 p)

(C) Ange ett 95-procentigt konfidensintervall för θ_3 och avgör om vi kan förkasta nollhypotesen $H_0 : \theta_3 = 0$.

Ledning: medelfelet beräknas som $\sqrt{\hat{\sigma}^2(\boldsymbol{S}^{-1})_{jj}}$ *och t-kvantiler finns överst i tentamen.* (3 p)

(D) Studera residualplotten i Figure 2. Ange vilket antagande för linjär regression som inte verkar vara uppfyllt. Vad innebär detta för konfidensintervallets pålitlighet? (2 p)

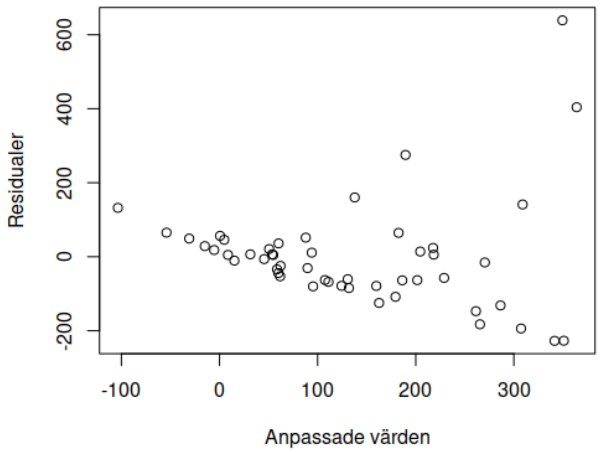

Figure 2: Residualplott från multipel linjär regression.

(E) Motivera varför en log-transformation av Y_i, d.v.s.

$$\log Y_i = \theta_0 + x_{i1}\theta_1 + x_{i2}\theta_2 + x_{i3}\theta_3 + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2) \text{ iid},$$

skulle kunna hantera problemet som noterades i **(D)**. (2 p)

Uppgift 4

(A) Beskiv kortfattat i punktform metoden *Forward selection* ("Framåt-metoder") för stegvis variabelselektion.

Ledning: utgå ifrån en modell helt utan x-variabler. (5 p)

(B) Beskiv kortfattat i punktform metoden *Backward elimination* ("Bakåt-metoder") för stegvis variabelselektion. (5 p)

Uppgift 5

Betrakta en AR(1)-process

$$Y_t = \phi Y_{t-1} + \epsilon_t, \qquad \epsilon_t \sim N(0, \sigma^2) \text{ iid}$$

med $|\phi| < 1$.

(A) Definierna begreppet kovarians-stationär. (4 p)

(B) Härled autokorrelationsfunktionen $\rho_k = Corr(Y_t, Y_{t+k})$ för $k = 1$ och $k = 2$. *Ledning: uppgiften är att skriva om ρ_1 och ρ_2 på så enkel form som möjligt användandes bla. stationäritet och att alla ϵ_t är oberoende.* (6 p)