STOCKHOLMS UNIVERSITET
Statistiska institutionen
Karl Sigfrid

# Tentamen i Dataanalys och Regression, 4.5 hp

**Kurs**: ST1101, Statistik och Dataanalys I, 15 hp
**Tentamensdatum**: 2024-09-27

Skrivtid: 08.00-12.00 (4 timmar).
Godkända hjälpmedel: Miniräknare utan lagrade formler och text. Lexikon. Linjal.
Bifogade hjälpmedel: Formelsamling och statistiska tabeller.

---

Tentamen består av 5 uppgifter, uppdelade i deluppgifter. Maximalt antal poäng anges per deluppgift.

Svar med fullständiga redovisningar ska lämnas. Svar ges på svenska eller engelska.

- Använd endast skrivpapper som tillhandahålls i skrivsalen.
- För full poäng på en uppgift krävs tydliga och väl motiverade lösningar.
- Om du inte lyckas lösa en deluppgift och behöver det svaret för en senare deluppgift så kan du hitta på värdet för att kunna göra beräkningarna i de efterföljande uppgifterna.

Tentamen kan maximalt ge 100 poäng, och för godkänt resultat krävs minst 50.

Betygsgränser:

| | |
|---|---|
| A: | 90-100 |
| B: | 80-89 |
| C: | 70-79 |
| D: | 60-69 |
| E: | 50-59 |
| Fx: | 40-49 |
| F: | 0-39 |

OBS! Fx och F är underkända betyg som kräver omexamination. Studenter som får betyget Fx kan alltså inte komplettera för högre betyg.

`Lycka till!`

---

**Uppgift 1.** (20 poäng)

(a.) Förklara skillnaden mellan deskriptiv statistik och inferens. (5p)

(b.) En elektronikbutik delar in sina varor i fem kategorier. Följande är en frekvenstabell som visar antalet sålda varor per kategori under en dag:

| Kategori | Antal sålda varor |
|---|---|
| Telefoner | 56 |
| Datorer | 32 |
| TV-apparater | 12 |
| Hushållsmaskiner | 7 |
| Övrigt | 4 |

Omvandla först frekvenstabellen till en relativ frekvenstabell. Rita sedan ett lämpligt diagram som illustrerar försäljningen baserat på den relativa frekvenstabellen. Diagrammet behöver inte vara snyggt ritat, men det bör vara korrekt på ett ungefär. Tänk på att välja en typ av diagram som används för den typ av variabel som det handlar om här. (5p)

(c.) Sju studenter gör ett prov, med följande resultat:

| Student | Antal poäng |
|---|---|
| Student 1 | 56 |
| Student 2 | 72 |
| Student 3 | 82 |
| Student 4 | 17 |
| Student 5 | 43 |
| Student 6 | 43 |
| Student 7 | 51 |

Ange poängfördelningens medelvärde (mean), typvärde (mode) och median. (5p)

(d.) En av studenterna i deluppgift (c) som fick 43 poäng på provet säger sig ha fått ett provresultat som ligger ungefär vid den tredje kvartilen i poängfördelningen. Är det korrekt? Motivera ditt svar. Du behöver inte göra någon uträkning för att besvara frågan. (5p)

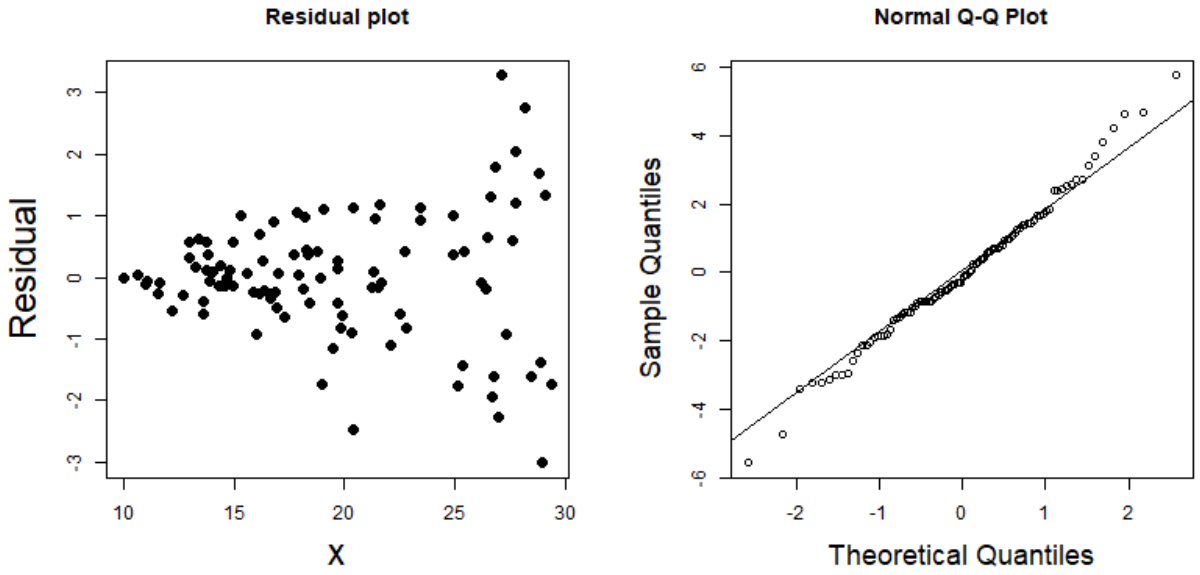
**Uppgift 2.** (20 poäng)

(a.) Vad är en residual i en regressionsmodell, och på vilket sätt är minsta kvadrat-metoden kopplad till residualerna? (5p)

(b.) När du ska anpassa (fit) och utvärdera (evaluate) en regressionsmodell, vad är syftet med att dela in ditt dataset i träningsdata och testdata? (5p)

(c.) Du vill göra en enkel linjär regressionsmodell där responsvariabeln är priset på en bostad och förklaringsvariabeln storleken på bostaden. Ett spridningsdiagram visar att sambandet inte är linjärt. Kan det ändå fungera att använda en enkel linjär regressionsmodell? Om ja, hur kan du gå tillväga? (4p)

(d.) Bilden nedan visar en residualgraf och en normalfördelningsgraf från en enkel linjär regressionsmodell. Vad säger graferna om de modellantaganden som gäller för en regressionsmodell? Finns det någon antagande som ser ut att vara uppfyllt? Finns det något som ser ut att inte vara uppfyllt? (6p)




Figur 1: Residualgraf och normalfördelningsgraf

**Uppgift 3.** (20 poäng)
En myndighet har fått i uppdrag att utvärdera tre olika högskolor. Framför allt är de intresserade av att följa om studenterna på respektive skola har en anställning inom ett år efter examen. Ett antal studenter som tog examen för ett år sedan har därför svarat på en enkät, där de angett om de har fått en anställning eller inte efter sin utbildning. Resultatet visas i korstabellen nedan.

| | | Anställning | |
|---|---|---|---|
| | | Ja | Nej |
| Högskola | Högskola 1 | 31500 | 2533 |
| | Högskola 2 | 33300 | 3362 |
| | Högskola 3 | 28800 | 2277 |

Variabeln *Högskola* anger från vilken av de tre högskolorna som en student fick sin examen. Variabeln *Anställning* anger om studenten har en anställning ett år efter examen.

(a.) Hur många studenter svarade på enkätfrågan? (2p)

(b.) Räkna ut marginalfördelningarna (marginal distributions) för båda variablerna. Uttryck marginalfördelningarna i procent. Tolka marginalfördelningarna. (4p)

(c.) Myndigheten vill jämföra hur bra de olika högskolorna är på att erbjuda utbildningar som leder till jobb. För att jämföra högskolorna med varandra, bör vi räkna ut fördelningen av variabeln *Anställning* betingad på variabeln *Högskola*, eller bör vi räkna ut fördelningen av variabeln *Högskola* betingad på variabeln *Anställning*? Motivera ditt svar. (4p)

(d.) Räkna ut den betingade fördelning som du föreslog i deluppgift **c**. Uttryck svaren i procent. (8p)

(e.) Tolka resultatet från deluppgift **d**. (2p)

**Uppgift 4.** (20 poäng)

Ett företag säljer matvaror till restaurangkedjor, och ger ofta större procentuella rabatter till stora kunder än till små kunder. En säljare på företaget har samlat data och tagit fram regressionsmodellen

$$\hat{y} = 0.05 + 0.2x,$$

där $x$ är den summa i miljoner kronor som en kund handlar för under ett år, och $\hat{y}$ är den estimerade rabatten i procent som kunden får.

(a.) Tolka modellens koefficienter (coefficients), dvs interceptet och lutningskoefficienten (slope coefficient). (4p)

(b.) Vi har följande information: $r_{xy} = 0.25, s_x = 2, s_y = 1.6, \bar{x} = 4.8, \bar{y} = 1.01$.

Förklara notationen, dvs förklara vad $r_{xy}$, $s_x$, $s_y$, $\bar{x}$ och $\bar{y}$ står för.

Verifiera sedan med hjälp av denna information att $b_0 = 0.05$ och att $b_1 = 0.2$. (4p)

(c.) Företagets nuvarnade kunder handlar för mellan 1.2 miljoner och 42 miljoner kronor per år. Om de fick en ny jättekund som handlade för 240 miljoner kronor om året, hur stor rabatt i procent skulle den kunden få enligt modellen? Är det klokt att använda modellen för att estimera rabatten för en ny kund av det här slaget? Motivera ditt svar. (4p)

(d.) Vi lägger till en dummy-variabel i modellen som har värdet 1 för nya kunder som tillkommit senaste året, och 0 för äldre kunder. Modellen blir nu

$$\hat{y} = -0.01 + 0.15x_1 + 0.09x_2,$$

där $x_1$ är den summa i miljoner kronor som en kund handlar för under ett år, $x_2$ är dummyvariabeln för nya kunder, och $\hat{y}$ är den estimerade rabatten i procent som kunden får.

Tolka modellens samtliga tre koefficienter. Kom ihåg att vi nu har en multipel linjär regressionsmodell. (4p)

(e.) För en annan regressionsmodell har vi följande information:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = 22300$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 19251$$

Räkna ut $R^2$ (R-kvadrat) för modellen. (4p)

**Uppgift 5.** (20 poäng)
Ett elbolag med ett stort antal kunder fakturerar kunderna en gång per kvartal. Elkostnaderna för bolagets kunder för det senaste kvartalet följde en normalfördelning med medelvärdet 570 kronor och standardavvikelsen 120 kronor.

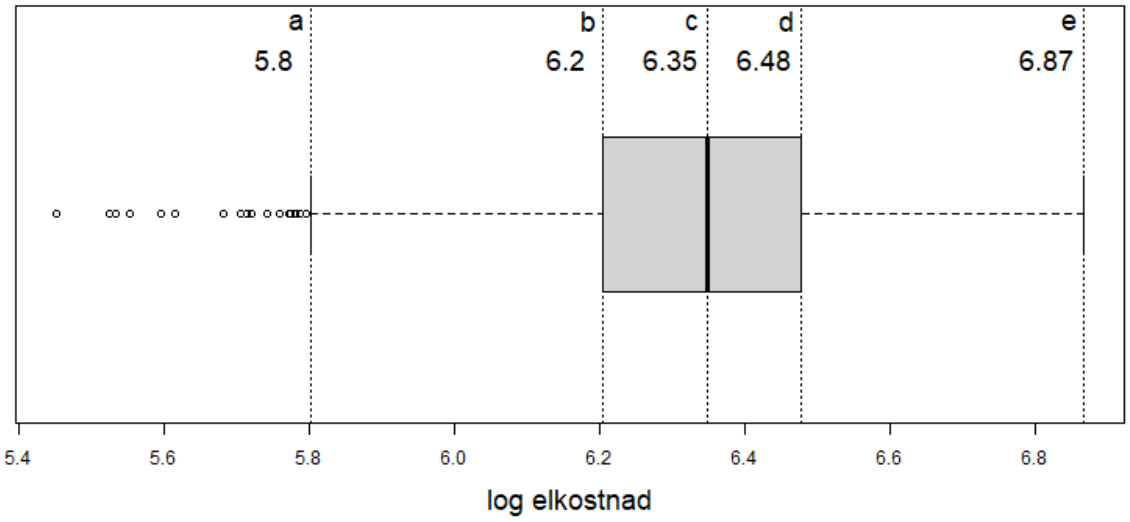
(a.) Jonas, som är en av bolagets kunder, hade för det senaste kvartalet en elkostnad på 700 kronor. Hur stor andel av kunderna hade en *lägre* elkostnad än Jonas senaste kvartalet? (5p)

(b.) Hur hög elkostnad hade en kund vars elkostnad befann sig vid den 10:e percentilen i fördelningen? (5p)

(c.) Beräkna kvartilavståndet (IQR) för fördelningen av kundernas elkostnader. (5p)

(d.) Figuren nedan visar fördelningen av kundernas logaritmerade elkostnader i form av ett låddiagram (boxplot). Använd informationen i låddiagrammet för att räkna ut kvartilavståndet (IQR) för fördelningen av elkostnader. Kvartilavståndet ska rapporteras i kronor. På grund av avrundade värden i låddiagrammet kan resultatet skilja sig något från resultatet i deluppgift c.

Notera att de logaritmerade värden som motsvarar punkterna a, b, c, d och e finns utskrivna i diagrammet. Logaritmerade kostnader är logaritmerade med basen $e$. (5p)




Figur 2: Fördelning av kundernas elkostnader