STOCKHOLMS UNIVERSITET
Statistiska institutionen
Matias Quiroz

Tentamen i Dataanalys och Regression, 4.5 hp

Kurs: ST1101, Statistik och Dataanalys I, 15 hp
Tentamensdatum: 2023-03-27

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

Tentamen består av 4 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.
- Skriv helst endast på en sida per blad.
- För full poäng på en uppgift krävs tydliga, utförliga 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.

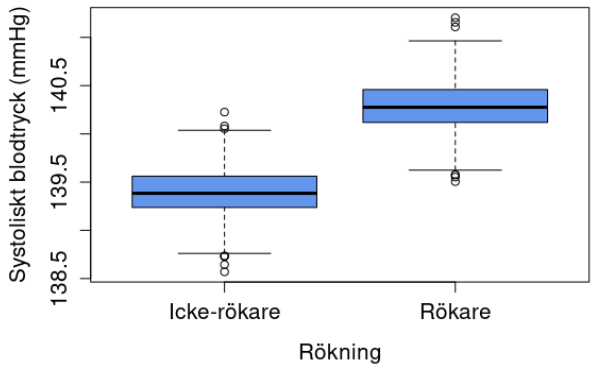
Tentamen kan maximalt ge 100 poäng och för godkänt resultat krävs minst 50.

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. Lösningsförslag läggs ut på Athena efter tentamen i samband med rättningen.

`Lycka till!`



Figur 1: Systoliskt blodtryck för rökare och icke-rökare i Uppgift 1.

Uppgift 1. (25 poäng)

Det systoliska blodtrycket (i millimeter kvicksilver, mmHg), även kallat det övre trycket, mättes hos 504 personer som delades in i två grupper: rökare (171 personer) och icke-rökare (333 personer). Låddiagram för varje grupp presenteras i Figur 1. Vidare är $\overline{x} = 140.2971$ och $\overline{y} = 139.3844$, där x och y är de systoliska blodtrycken i gruppen rökare respektive icke-rökare.

(a.) Beräkna marginalfördelningen för den kategoriska variabeln rökning. (3p)

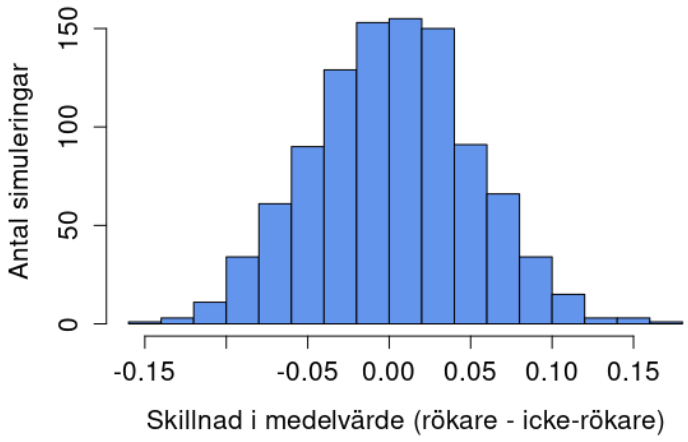
(b.) Jämför fördelningen för det systoliska blodtrycket mellan grupperna rökare och icke-rökare. (5p)

(c.) Det systoliska blodtrycket för ett urval av 6 personer visas i Tabell 1. Räkna ut medelvärdet, medianen och standardavvikelsen baserat på dessa observationer. Tolka standardavvikelsen. (9p)

Person	1	2	3	4	5	6
Systoliskt blodtryck (mmHg)	140.44	141.16	139.67	139.36	139.68	139.29

Tabell 1: Tabell för Uppgift 1 (c.).

(d.) Finns det ett statistiskt signifikant samband mellan rökning och genomsnittligt systoliskt blodtryck och hur ser det sambandet i sådana fall ut? Antag att det inte finns ett samband, dvs att rökning och systoliskt blodtryck är oberoende. Då skulle de 504 blodtrycken fördela sig slumpmässigt bland de två grupperna (rökare och icke-rökare). Låt oss med hjälp av R simulera 1000 stickprov enligt dessa förutsättningar. För varje simulerat stickprov, räknas skillnaden i gruppernas medelvärden, dvs $\overline{x} - \overline{y}$, och resultatet visas i form av ett histogram i Figur 2. Vad kan du dra för slutsats? (8p)



Figur 2: Simuleringen i Uppgift 1 (d.).

Uppgift 2. (25 poäng)

1200 universitetsstudenter blev tillfrågade om sin kroppsuppfattning. Kroppsuppfattning antas ha tre utfall: underviktig, överviktig, hälsosam vikt. Resultaten delades in i kön som antas ha två utfall: man, kvinna. Resultaten sammanställdes i form av en korsstabell som visas i Tabell 2.

	Underviktig	Överviktig	Hälsosam vikt	
Man	295	72	73	
Kvinna	560	163	37	
				Totalt: 1200

Tabell 2: Korstabell för Uppgift 2.

(a.) Beräkna marginalfördelningen för variabeln kroppsuppfattning. (3p)

(b.) Hur stor andel av alla tillfrågade var män som ansåg sig vara överviktiga? (3p)

(c.) Beräkna simultanfördelningen för kön och kroppsuppfattning. (5p)

(d.) Jämför kroppsuppfattning mellan män och kvinnor. (6p)

(e.) Bland de tillfrågade fanns Elsa som ansåg sig själv vara överviktig. Antag att viktfördelningen för kvinnor är normalfördelad med medelvärde 68 kg och standardavvikelse 10 kg. Om z-värdet för Elsas vikt är -0.2, hur mycket väger Elsa? Hur stor andel kvinnor i populationen väger mer än Elsa? (8p)

Uppgift 3. (25 poäng)

Föreligger det ett samband mellan inkomst och studietid? Datasetet vi använder här består av $n = 102$ observationer, där varje observation motsvarar en sysselsättning. Stickprovet består av (x_i, y_i) för $i = 1, 2, \ldots, 102$, där y_i är medelinkomst (USD) och x_i är medelstudietid (år) för sysselsättning i.

För att besvara frågeställningen ovan ska du använda en linjär regression med medelinkomst som responsvariabel och medelstudietid som förklarande variabel . Följande kvantiter har beräknats utifrån stickprovet samt minsta kvadratanpassningen $\hat{y} = b_0 + b_1 x$.

- $\sum_{i=1}^{102} x_i = 1095.28$ och $\sum_{i=1}^{102} y_i = 693386$.
- $\sum_{i=1}^{102} (x_i - \overline{x})^2 = 751.8852$ och $\sum_{i=1}^{102} (y_i - \overline{y})^2 = 1820813411$.
- $\sum_{i=1}^{102} (x_i - \overline{x})(y_i - \overline{y}) = 675804.1$ och $\sum_{i=1}^{102} (y_i - \hat{y}_i)^2 = 1213392025$.

(a.) Anpassa en linjär regression för att undersöka om det finns ett samband mellan medelinkomst och medelstudietid. Tolka sambandet med hjälp av lämplig koefficient i den skattade modellen. (8p)

(b.) Beräkna och tolka R^2. (3p)

(c.) Beräkna och tolka residualstandardavvikelsen s_e. (3p)

(d.) Sysselsättningen kemist har medelinkomst 8403 och medelstudietid 14.62. Beräkna dess residual. (3p)

(e.) Antag att vi får tillgång till en dummy-variabel (indikatorvariabel) `univ`, där kodningen är 1 för "minst 80% har en universitetsutbildning" och 0 för "mindre än 80% har en universitetsutbildning". Den nya anpassade modellen är

$$\hat{y} = b_0 + b_1 x + b_2 \cdot univ.$$

Tolka b_2. Förväntar du dig att b_2 är positiv eller negativ? (8p)

Uppgift 4. (25 poäng)

Datasetet `possum_data` i R innehåller variablerna `age` (ålder i år), `belly` (midjeomkrets i cm) och `hdlngth` (huvudlängd i cm) för 100 australiensiska pungråttor.

(a.) Modellen

$$\widehat{age} = b_0 + b_1 \cdot belly + b_2 \cdot hdlngth$$

har anpassats i R och resultaten visas i Figur 3. Ange skattningarna b_0, b_1 och b_2. Tolka b_1 och b_2. Vad är residualernas medelvärde? (8p)

(b.) Modellen

$$\widehat{\log(age)} = b_0 + b_1 \cdot belly + b_2 \cdot hdlngth$$

har anpassats i R och resultaten visas i Figur 4. Tolka R^2. (5p)

(c.) Använd modellen i (b.) för att prediktera åldern för en pungråtta med en midjeomkrets på 34.5 cm och huvudlängd 90.7 cm. (5p)

(d.) Använd korsvalidering med $K = 4$ folds för att välja mellan modellerna i (a.) och (b.). Tabell 3 innehåller sum of squares error (SSE) för alla folds uppdelat efter modell. Inom en given fold, hur många observationer tillhör träningsdata respektive testdata? (7p)

Modell	**SSE**			
	Fold 1	Fold 2	Fold 3	Fold 4
(a.)	63.3363	83.0180	67.4278	145.2334
(b.)	63.8065	93.0105	87.0939	218.9601

Tabell 3: Tabell för Uppgift 4 (d.).

```
> summary(lm(age ~ belly + hdlngth, data = possum_data))

Call:
lm(formula = age ~ belly + hdlngth, data = possum_data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4402 -1.2321 -0.2368  1.0822  5.0829

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -11.41659    4.64914  -2.456   0.0158 *
belly         0.16225    0.07722   2.101   0.0382 *
hdlngth       0.10685    0.05944   1.797   0.0754 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.752 on 97 degrees of freedom
Multiple R-squared:  0.1472,    Adjusted R-squared:  0.1296
F-statistic:  8.37 on 2 and 97 DF,  p-value: 0.0004432
```

Figur 3: R utskrift för Uppgift 4 (a.).

```
> summary(lm(log(age) ~ belly + hdlngth, data = possum_data))

Call:
lm(formula = log(age) ~ belly + hdlngth, data = possum_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.18202 -0.29884  0.04891  0.37179  0.95731

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.42863    1.33747  -3.311  0.00131 **
belly        0.05562    0.02221   2.504  0.01395 *
hdlngth      0.04104    0.01710   2.400  0.01831 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.504 on 97 degrees of freedom
Multiple R-squared:  0.2142,    Adjusted R-squared:  0.198
F-statistic: 13.22 on 2 and 97 DF,  p-value: 8.369e-06
```

Figur 4: R utskrift för Uppgift 4 (b.).