STOCKHOLMS UNIVERSITET
Statistiska institutionen
Matias Quiroz

Tentamen i Dataanalys och Regression, 4.5 hp

Kurs: ST1101, Statistik och Dataanalys I, 15 hp
Tentamensdatum: 2023-02-10

Skrivtid: 14.00-18.00 (4 timmar).
Godkända hjälpmedel: Miniräknare utan lagrade formler och text. Lexikon.
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.

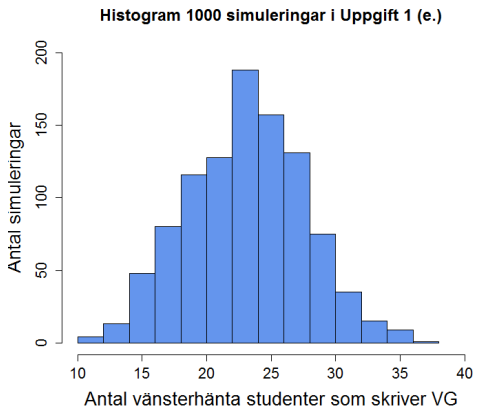
Tentamen kan maximalt ge 100 poäng och för godkänt resultat krävs minst 50.

Betygsgränser:

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

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: Lärarinnans simulering i Uppgift 1 (e.).

Uppgift 1. (25 poäng)

En lärarinna ville utreda om det fanns ett samband mellan god matematikförmåga och handpreferens (hänthet, dvs vilken hand man föredrar att skriva med). Matematikförmågan bedömdes enligt variabeln betyg, där utfallen var underkänd (U), godkänd (G) eller välgodkänd (VG). Handpreferensvariabeln har utfallen högerhänt eller vänsterhänt. Hon samlade ihop data från 1000 studenter och sammanställde korstabellen i Tabell 1.

	U	G	VG	
Vänsterhänt	16	45	34	
Högerhänt	149	542	214	
				Totalt: 1000

Tabell 1: Korstabell för Uppgift 1.

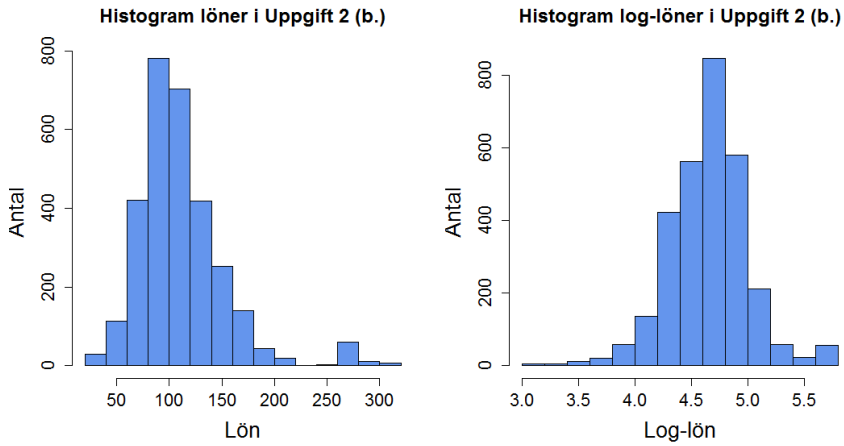
(a.) Beräkna marginalfördelningen för variabeln betyg. Hur stor andel av alla studenterna fick betyget underkänd? (5p)

(b.) Hur stor andel av alla studenterna var högerhänta och fick betyget VG? (4p)

(c.) Hur stor andel av de vänsterhänta fick betyget VG? (4p)

(d.) Hur stor andel av de underkända var vänsterhänta? (4p)

(e.) Lärarinnan misstänkte att vänsterhänta studenter får VG i högre utsträckning än högerhänta studenter men undrade om detta kunde bero på slumpen. För att utreda detta resonerade hon enligt följande. Om hänthet och förmågan att få betyget VG inte skulle ha något samband, då hade de 248 VG betygen fördelat sig slumpmässigt bland de högerhänta och vänsterhänta. Med hjälp av R simulerade hon 1000 stickprov enligt dessa förutsättningar och skapade histogrammet i Figur 1. Vad kan lärarinnan dra för slutsats? (8p)



Figur 2: Histogram för löner och log-löner i Uppgift 2 (b.).

Uppgift 2. (25 poäng)

Den här uppgiften består av oberoende deluppgifter.

(a.) Sofia föddes på BB Sös och vägde 2.4 kg. Antag att populationen nyfödda följer en normalfördelning med medelvärde 3.5 kg och standardavvikelse 1.1. Hur stor andel av de nyfödda väger mer än Sofia? (7p)

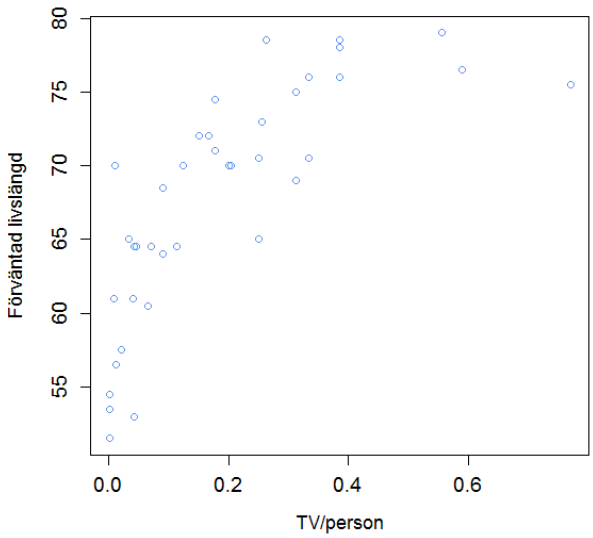
(b.) Figur 2 visar två histogram, löner och log-löner (naturliga logaritmen av löner), för 3000 arbetare i USA (dagslön i USD). Beskriv bägge fördelningarna. Nämn en fördel med att använda log-löner framför löner för en statistisk analys i det här fallet. (4p)

(c.) Timlönen för 6 arbetare visas i Tabell 2. Räkna ut medelvärdet, standardavvikelsen och variansen för variabeln timlön baserat på dessa observationer. (6p)

Arbetare	1	2	3	4	5	6
Timlön	130	125	150	170	120	135

Tabell 2: Tabell för Uppgift 2 (c.).

(d.) Figur 3 visar förhållandet mellan förväntad livslängd och antal TV per person för 38 länder. Föreslå två icke-linjära regressionsmodeller som kan användas för att modellera dessa data. Du behöver inte bekräfta att dina föreslagna transformationer resulterar i ett mer linjärt samband. Vad kan sägas om kausalitet i det här exemplet? (8p)



Figur 3: Punktdiagram för förväntad livslängd och antal TV per person i Uppgift 2 (d.).

Uppgift 3. (25 poäng)

Ett familjeföretag har 5 anställda med olika arbetslivserfarenhet. Företaget vill kunna förutsäga vilken månadslön de ska erbjuda framtida nyanställda baserat på deras arbetslivserfarenhet och vill att du ska ta fram en linjär regressionsmodell de kan använda för detta ändamål. Data finns i Tabell 3.

Tabell 3: Data för familjeföretaget i Uppgift 3.

Arbetslivserfarenhet (år)	5	7	3	6	8
Månadslön (kr)	30000	32000	25000	33000	40000

(a.) Använd minsta kvadratmetoden för att anpassa en linjär regression med hjälp av data i Tabell 3. Korrelationskoefficienten mellan månadslön och arbetslivserfarenhet är 0.9332. Ange minsta kvadratanpassningen (linjens ekvation). (7p)

(b.) Tolka koefficienterna b_0 och b_1 i (a.). (5p)

(c.) Beräkna residualen för den anställde som har 7 års arbetslivserfarenhet. (3p)

(d.) Beräkna och tolka R^2. (4p)

(e.) Företaget ska intervjua två personer imorgon, en med 4 års arbetslivserfarenhet samt en med 25 års arbetslivserfarenhet. Använd din anpassade modell för att ge företaget råd om lämpliga månadslöner. (6p)

Uppgift 4. (25 poäng)

Datasetet `mtcars` i R innehåller variablerna `mpg` (bensinförbrukning mätt i miles per gallon), `hp` (hästkrafter) och `vs` (motortyp) för olika bilar. Variabeln `vs` är en dummy-variabel (indikatorvariabel), där kodningen är 1 för "rak motortyp" och 0 för "V-formad motortyp". Datasetet består av 32 observationer.

(a.) Modellen

$$\widehat{mpg} = b_0 + b_1 \cdot hp + b_2 \cdot vs$$

har anpassats i R och resultaten visas i Figur 4. Ange skattningarna b_0, b_1 och b_2. Tolka b_1 och b_2. (8p)

(b.) Använd modellen i (a.) för att prediktera bensinförbrukningen för en bil med 160 hästkrafter och som har en rak motortyp. (5p)

(c.) Modellen

$$\widehat{mpg^{1/2}} = b_0 + b_1 \cdot hp^{1/2} + b_2 \cdot vs$$

har anpassats i R och resultaten visas i Figur 5. Ange skattningarna b_0, b_1 och b_2. Tolka R^2. (5p)

(d.) Använd modellen i (c.) för att prediktera bensinförbrukningen för en bil med 160 hästkrafter och som har en V-formad motortyp. (7p)

```
> summary(lm(mpg ~ hp + vs, data = mtcars))

Call:
lm(formula = mpg ~ hp + vs, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7131 -2.3336 -0.1332  1.9055  7.9055

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 26.96300    2.89069   9.328 3.13e-10 ***
hp          -0.05453    0.01448  -3.766 0.000752 ***
vs           2.57622    1.96966   1.308 0.201163
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.818 on 29 degrees of freedom
Multiple R-squared:  0.6246,    Adjusted R-squared:  0.5987
F-statistic: 24.12 on 2 and 29 DF,  p-value: 6.768e-07
```

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

```
> summary(lm(sqrt(mpg) ~ I(sqrt(hp)) + vs, data = mtcars))

Call:
lm(formula = sqrt(mpg) ~ I(sqrt(hp)) + vs, data = mtcars)

Residuals:
     Min       1Q   Median       3Q      Max
-0.70013 -0.23399 -0.02781  0.22081  0.78996

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.46054    0.51530  12.538 3.10e-13 ***
I(sqrt(hp)) -0.17709    0.03729  -4.749 5.09e-05 ***
vs           0.14558    0.20543   0.709    0.484
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3809 on 29 degrees of freedom
Multiple R-squared:  0.6894,    Adjusted R-squared:  0.668
F-statistic: 32.19 on 2 and 29 DF,  p-value: 4.324e-08
```

Figur 5: R utskrift för Uppgift 4 (c.).