Lösningsförslag, SDA1, 20241107

Uppgift 1

a

Lämpliga typer diagram för att visa fördelningen av en numerisk variabel kan vara histogram, låddiagram (box plot) eller spridningsdiagram (scatter plot).

Lämpliga typer av diagram för att visa fördelningen av en kategorisk variabel kan vara stapeldiagram eller pajdiagram.

b

En outlier är ett extremvärde som avviker från det övergripande mönstret i en fördelning.

Om vi noterar en outlier i datamaterialet bör vi till att börja med undersöka orsaken till att den finns. Sedan får vi ta ställning till om observationen bör inkluderas i analysen.

c

Bias är ett systematiskt fel som uppstår när vi tar ett stickprov. Det kan till exempel uppkomma i en survey om personer med vissa åsikter är mindre svarsbenägna och därför underrepresenterade i stickprovet. Det kan också uppstå om observationerna i stickprovet kommer ur en grupp som inte representerar hela populationen med avseende på det vi vill mäta, exempelvis om vi vill mäta befolkningens medelinkomst och väljer ut enbart personer som har högskoleutbildning.

d

En dold variabel är en variabel som vi inte har i vår data men som orsakar ett samband mellan variabler som vi studerar. Exempel: Om vi ser ett samband mellan antalet tv-apparater per hushåll och kvalitén på sjukvården i ett land så kan landets ekonomiskt välstånd vara en dold variabel som förklarar varför variablerna samvarierar. Högt ekonomisk välstånd har ett positivt samband både med antalet tv-apparater och med bättre sjukvård.

Uppgift 2

a

Responsvariabeln är variabeln som vi vill estimera med hjälp av en eller flera förklaringsvariabler. Vi brukar kalla responsvariabeln y och förklaringsvariablerna x.

b

Avståndet som e representerar kallas residual.

Antagandet om konstant varians är ett antagande om att storleken på residualerna är oberoende av $\hat{y}$. Om vi har en enkel linjär regression kan vi också säga att storleken på residualerna ska vara oberoende av x.

Antagandet om normalfördelning är ett antagande om att residualerna följer en normalfördelning.

c

Att modell 1 har lägre RMSE när den utvärderas på testdata talar för att den ger bättre prediktioner än modell 2 när den utvärderas på ny data.

Att modell 1 samtidigt har lägre R^2 visar att den presterar något sämre när den utvärderas på samma data som använts för att anpassa modellen.

Vanligtvis vill vi ha en modell som är generaliserbar, och därmed ger bra prediktioner när vi använder den tillsammans med ny data, vilket är ett starkt argument för att välja modell 1.

d

I ett experiment bestämmer den som gör studien vilka individer som ska ingå i vardera grupp som ska jämföras. Genom att göra grupperna lika, med avseende på alla faktorer utom den som vi vill undersöka, kan vi utifrån resultatet dra slutsatser om orsakssamband. Det kan vi normalt sett inte göra om vi har en observationsstudie, där vi jämför grupper som vi inte själva har satt samman.

Uppgift 3

Vår korstabell (också kallad simultanfördelningstabell):

```
m <- c(108, 201, 39, 290, 447, 60) |> matrix(ncol=2)
colnames(m) <- c("Ja", "Nej")
rownames(m) <- c("Olivolja", "Chokladask", "Biobiljett")
m
```

```
            Ja Nej
Olivolja   108 290
Chokladask 201 447
Biobiljett  39  60
```

a

```
sum(m)
```

```
[1] 1145
```

1145 personer fick en belöning.

b

39 pensionärer valde en biobiljett.

c

```
rowSums(m) / sum(m)
```

```
  Olivolja Chokladask Biobiljett
0.34759825 0.56593886 0.08646288
```

```
colSums(m) / sum(m)
```

```
       Ja       Nej
0.3039301 0.6960699
```

Av alla som fick en belöning valde ungefär 35 procent olivolja, 57 procent en chokladask och 9 procent en biobiljett.

Av alla som fick en belöning kategoriserades 30 procent som pensionärer, och 70 procent kategoriserades inte som pensionärer.

d

```
t(m) / colSums(m)
```

```
     Olivolja Chokladask Biobiljett
Ja  0.3103448  0.5775862 0.11206897
Nej 0.3638645  0.5608532 0.07528231
```

e

Vi tittar på pensionärerna för sig och de som inte är pensionärer för sig.

Bland pensionärerna valde 31 procent olivolja, 58 procent en chokladask och 11 procent en biobiljett.

Bland de som inte var pensionärer valde 36 procent olivolja, 56 procent en chokladask och 8 procent en biobiljett.

Uppgift 4

a

Interceptet: En fisk som äter noll kalorier per dag förväntas väga 134 gram vid två månaders ålder. Bör inte tolkas bokstavligt.

Lutningskoefficienten: För varje ytterligare kalori som en fisk äter per dag förväntar vi oss att en fisk väger ytterligare 0.2 gram vid två månaders ålder.

b

Vi byter ut x mot 21 i vår formel och räknar ut $\hat{y}$.

```
134 + 0.2 * 21
```

```
[1] 138.2
```

Vi estimerar att fisken väger 138.2 gram.

c

r_{xy} är korrelationskoefficienten för x och y. s_y, s_y är standardavvikelsen för y respektive x. $\bar{y}$ och $\bar{x}$ är medelvärdet för y respektive x.

```
b1 <- 0.16 * 5 / 4
b1
```

```
[1] 0.2
```

```
b0 <- 137.6 - 0.2 * 18
b0
```

```
[1] 134
```

d

b_0: en fisk som har ätit noll kalorier per dag och som inte har fått vitamintillskott förväntas enligt modellen ha en vikt på 126 gram.

b_1: Givet värdet på x_2 förväntas vikten vid två månaders ålder vara 0.19 gram högre för varje ytterligare kalori per dag.

b_2: Givet värdet på x_1 förväntas en fisk som har fått vitamintillskott att väga 24 gram mer.

e

Med kodningen $x_2 = 1$ för fiskar som får vitamintillskott har vi den ursprungliga formeln:

$\hat{y} = 126 + 0.19x_1 + 24x_2$

Om vi kodar om dummyvariabeln så att $x_2 = 1$ för fiskar som **inte** får vitamintillskott år vi en formel där termen $24x_2$ blir negativ, och där interceptet därför ökar med 24:

$\hat{y} = 150 + 0.19x_1 - 24x_2$

Exempel: Om vi har en fisk som ätit 10 kalorier om dagen och som fick kosttillskott får vi med den ursprungliga formeln:

$\hat{y} = 126 + 0.19x_1 + 24x_2 = 126 + 0.19 \cdot 10 + 24 \cdot 1 = 151.9$

För samma fisk får vi med den omkodade dummyvariabeln:

$\hat{y} = 150 + 0.19x_1 - 24x_2 = 150 + 0.19 \cdot 10 - 24 \cdot 0 = 151.9$

I både fallen skattar vi alltså fiskens vikt till 151.9 gram.

Uppgift 5

a

```
z <- (459 - 340) / 120
z
```

```
[1] 0.9916667
```

```
pnorm(z)
```

```
[1] 0.8393199
```

Boken har fler sidor än 84 procent av bibliotekets böcker.

b

```
z <- qnorm(0.9)
z
```

```
[1] 1.281552
```

```
340 + 120 * z
```

```
[1] 493.7862
```

En bok behöver ha 494 sidor för att sidantalet ska ligga över den 90:e percentilen.

c

Vi hittar inte exakt 0.75 i normalfördelningstabellen, men det närmaste värdet motsvarar Z-värdet 0.67. Z-värdet för Q3 är alltså 0.67, och Z-värdet för Q1 blir då -0.67 på grund av att normalfördelningen är symmetrisk.

```
Q3 <- 340 + 0.67 * 120
Q3
```

```
[1] 420.4
```

```
Q1 <- 340 - 0.67 * 120
Q1
```

```
[1] 259.6
```

```
IQR <- Q3 - Q1
IQR
```

```
[1] 160.8
```

Kvartilavståndet i fördelningen av sidantal är ungefär 161 sidor.

d

```
0.9 * Q1
```

```
[1] 233.64
```

I deluppgift *c* räknade vi ut att den första kvartilen i fördelningen är 259,6 sidor. Om vi minskar antalet sidor i varje bok med 10 procent kommer samma bok som tidigare låg vid första kvartilen fortfarande att ligga vid första kvartilen.

Antalet sidor i boken är 90 procent av det ursprungliga antalet, dvs

$0.9 \cdot 259.6 \approx 234$ sidor