ØVINGER 2017

Løsninger til oppgaver

Øving 12

7.1. Med utgangspunkt i de $n = 5$ observasjonsparene $(x_1, y_1), (x_2, y_2), \ldots, (x_5, y_5)$ beregner vi først middelverdiene

$$\bar{x} = \frac{1}{5}\sum_{i=1}^{5} x_i = 3.2 \qquad \text{og} \qquad \bar{y} = \frac{1}{5}\sum_{i=1}^{5} y_i = 2.4.$$

Estimert kovarians blir

$$s_{XY} = \frac{1}{5-1}\sum_{i=1}^{5}(x_i - \bar{x})(y_i - \bar{y}) = 1.4,$$

mens estimatene av variansene til X og Y er

$$s_X^2 = \frac{1}{5-1}\sum_{i=1}^{5}(x_i - \bar{x})^2 = 1.7 \qquad \text{og} \qquad s_Y^2 = \frac{1}{5-1}\sum_{i=1}^{5}(y_i - \bar{y})^2 = 1.3.$$

Korrelasjonen til observasjonsparene er dermed

$$r = \frac{s_{XY}}{\sqrt{s_X^2 s_Y^2}} = \frac{1.4}{\sqrt{1.7 \cdot 1.3}} = 0.9417.$$

La $\hat{y} = \hat{\alpha} + \hat{\beta}x$ være likningen som beskriver minste kvadraters rette linje. De estimerte koeffisientene er

$$\hat{\beta} = r \cdot \frac{s_Y}{s_X} = r \cdot \sqrt{\frac{s_Y^2}{s_X^2}} = 0.9417 \cdot \sqrt{\frac{1.3}{1.7}} = 0.8235$$

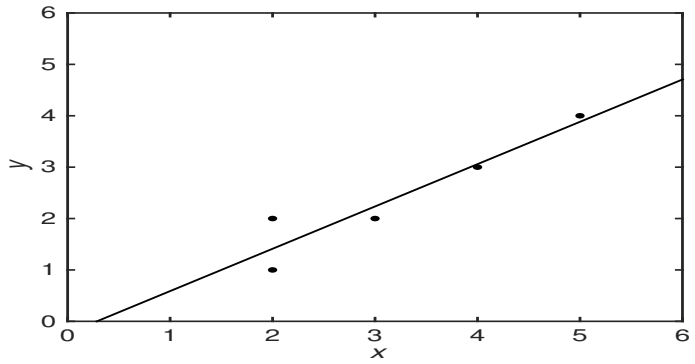
og

$$\hat{\alpha} = \bar{y} - \hat{\beta} \cdot \bar{x} = 2.4 - 0.8235 \cdot 3.2 = -0.2352.$$

Med disse tallverdiene er uttrykket for minste kvadraters rette linje

$$\hat{y} = -0.2352 + 0.8235x.$$

Denne linja er tegnet inn i spredningsplottet.



7.4. Vi har følgende korrigerte datasett.

Motorstørrelse x (hk)	75	145	55	88	122
Bensinforbruk y (liter/mil)	0.48	**0.52**	0.53	0.97	0.78

Korrelasjonen til disse dataene er

$$r = \frac{s_{XY}}{\sqrt{s_X^2 s_Y^2}} = \frac{\sum_{i=1}^5 (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^5 (x_i - \bar{x})^2 \cdot \sum_{i=1}^5 (y_i - \bar{y})^2}} = \frac{2.91}{\sqrt{5258 \cdot 0.1793}} = 0.0948.$$

Koeffisientene til den nye regresjonskurven blir

$$\hat{\beta} = r \cdot \frac{s_Y}{s_X} = r \cdot \sqrt{\frac{(n-1)s_Y^2}{(n-1)s_X^2}} = 0.0948 \cdot \sqrt{\frac{0.1793}{5258}} = 0.00055$$

og

$$\hat{\alpha} = \bar{y} - \hat{\beta} \cdot \bar{x} = 0.656 - 0.00055 \cdot 97.0 = 0.603.$$

Regresjonskurven for det korrigerte datasettet er altsågitt ved

$$\hat{y} = 0.603 + 0.00055x.$$

Utfra dette ser vi at den ene feilregistreringen utgjør forskjellen mellom en signifikant sammenheng og ingen sammenheng. Effekten av å endre en observasjon avtar ettersom antall observasjonspar n øker, så med $n = 500$ vil en enkelt feilregistrering ha liten innvirkning på resultatet.

7.5. Med det nye datasettet hvor motorstørrelsen måles i kilowatt, får vi empirisk korrelasjon

$$r = \frac{\sum_{i=1}^5 (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^5 (x_i - \bar{x})^2 \cdot \sum_{i=1}^5 (y_i - \bar{y})^2}} = \frac{22.131}{\sqrt{2808.5 \cdot 0.2842}} = 0.7833,$$

altså samme korrelasjon som i eksempel 7.1. Dette er forventet, siden det bare er skalaen til observasjonene som er endret i forhold til datasettet i eksempelet, og nevneren i uttrykket for r kompenserer for skalaen. Estimert kovarians blir imidlertid ikke lik (5.53 i stedet for 7.57).

Regresjonslinjen for kilowatt-dataene har estimert stigningstall

$$\hat{\beta} = r \cdot \frac{s_Y}{s_X} = 0.7833 \cdot \sqrt{\frac{0.2842}{2808.5}} = 0.0079,$$

og estimert skjæringspunkt

$$\hat{\alpha} = \bar{y} - \hat{\beta} \cdot \bar{x} = 0.77 - 0.0079 \cdot 70.88 = 0.211.$$

Skjæringspunktet er med andre ord det samme som før, mens stigningstallet er større. Sammenligner vi stigningstallene i denne oppgaven og i eksempel 7.1, ser vi at forholdet mellom det nye og gamle stigningstallet er $0.0079/0.00576 = 1/0.73$, altså forholdet mellom enhetene kW og hk.

7.8 (6). Bruker vi relasjonen $Z = 32 + X \cdot 9/5$ til å transformere temperaturene $x_1, x_2, \ldots, x_n$, får vi Fahrenheit-temperaturene

$$z_i = 32 + \frac{9}{5}x_i, \qquad \text{for} \qquad i = 1, 2, \ldots, n.$$

Middelverdien til de transformerte temperaturene blir

$$\bar{z} = \frac{1}{n}\sum_i z_i = \frac{1}{n}\sum_i \left(32 + \frac{9}{5}x_i\right) = \frac{1}{n}\left(32n + \frac{5}{9}\sum_i x_i\right) =$$

$$= 32 + \frac{9}{5} \cdot \frac{1}{n}\sum_i x_i = 32 + \frac{9}{5}\bar{x}.$$

Den estimerte kovariansen mellom Y og Z blir

$$s_{ZY} = \frac{1}{n-1}\sum_i (z_i - \bar{z})(y_i - \bar{y}) =$$

$$= \frac{1}{n-1}\sum_i \left(\left[32 + \frac{9}{5}x_i\right] - \left[32 + \frac{9}{5}\bar{x}\right]\right)(y_i - \bar{y}) =$$

$$= \frac{1}{n-1}\sum_i \frac{9}{5}(x_i - \bar{x})(y_i - \bar{y}) = \frac{9}{5}s_{XY},$$

og den estimerte variansen til Z blir

$$s_Z^2 = \frac{1}{n-1}\sum_i (z_i - \bar{z})^2 = \frac{1}{n-1}\sum_i \left(\left[32 + \frac{9}{5}x_i\right] - \left[32 + \frac{9}{5}\bar{x}\right]\right)^2 =$$

$$= \frac{1}{n-1}\sum_i \left(\frac{9}{5}\left[x_i - \bar{x}\right]\right)^2 = \left(\frac{9}{5}\right)^2 \frac{1}{n-1}\sum_i (x_i - \bar{x})^2 = \left(\frac{9}{5}\right)^2 s_X^2.$$

Setter vi dette inn i uttrykket for korrelasjonen mellom Z og Y, får vi

$$r_{ZY} = \frac{s_{ZY}}{\sqrt{s_Z^2 \cdot s_Y^2}} = \frac{(9/5)s_{XY}}{\sqrt{(9/5)^2 s_X^2 \cdot s_Y^2}} = \frac{s_{XY}}{\sqrt{s_X^2 \cdot s_Y^2}} = r_{XY}.$$

Korrelasjonen endres altså ikke av transformasjonen fra Celsius til Fahrenheit. Samme argument gjelder for andre lineære transformasjoner, for eksempel den i oppgave 7.5.

La regresjonslinjen for X være gitt ved

$$\hat{y} = \hat{\alpha}_x + \hat{\beta}_x x = 9.5 + 0.5x.$$

Relasjonen mellom X og Z gir

$$\begin{aligned} z &= 32 + \frac{9}{5}x \\ \frac{9}{5}x &= z - 32 \\ x &= \frac{5}{9}(z - 32), \end{aligned}$$

slik at likningen for regresjonslinjen kan skrives om til

$$\hat{y} = \hat{\alpha}_x + \hat{\beta}_x \cdot \frac{5}{9}(z - 32) = \left(\hat{\alpha}_x - \frac{5}{9} \cdot 32\hat{\beta}_x\right) + \left(\frac{5}{9} \cdot \hat{\beta}_x\right) z = \hat{\alpha}_z + \hat{\beta}_z z.$$

Regresjonslinjen for Z har dermed skjæringspunkt

$$\hat{\alpha}_z = \hat{\alpha}_x - \frac{5}{9} \cdot \hat{\beta}_x \cdot 32 = 9.5 - \frac{5}{9} \cdot 0.5 \cdot 32 = 0.6111$$

og stigningstall

$$\hat{\beta}_z = \frac{5}{9} \cdot \hat{\beta}_x = \frac{5}{9} \cdot 0.5 = 0.2778.$$

Regresjonslinjen for Z beskrives altså av likningen

$$\hat{y} = 0.6111 + 0.2778z.$$

Vi kan oppsummere oppgave 7.5 og 7.8 (6) ved å si at en ren skalering av dataene (multiplikasjon med en konstant) kun endrer stigningstallet til regresjonslinjen, mens en generell lineær transformasjon (skalering og translasjon) påvirker både stigningstallet og skjæringspunktet. Korrelasjonen vil være uendret i begge tilfeller.

7.7. Kaller de seks observasjonsparene av kroppshøyde og vekt (x_i, y_i), $i = 1, 2, \ldots, n$, hvor $n = 6$. Regresjonslinjens koeffisienter estimeres til

$$\hat{\beta} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{751}{875.3333} = 0.8580$$

og

$$\hat{\alpha} = \bar{y} - \hat{\beta} \cdot \bar{x} = 76 - 0.8580 \cdot 173.6667 = -72.9989.$$

Den estimerte variansen til feilleddet er

$$s^2 = \frac{1}{n-2} \sum_i (y_i - \hat{y})^2 = \frac{1}{n-2} \sum_i \left(y_i - \hat{\alpha} - \hat{\beta} x_i \right)^2 = \frac{125.6729}{6-2} = 31.4182 =$$

slik at estimert standardavvik blir $s = 5.6052$.

Vår beste gjetning på standardfeilen til koeffisientene $\hat{\alpha}$ og $\hat{\beta}$ er

$$SE(\hat{\alpha}) = \sqrt{\frac{s^2 \sum_i x_i^2}{n \sum_i (x_i - \bar{x})^2}} = 32.9813 \qquad \text{og} \qquad SE(\hat{\beta}) = \sqrt{\frac{s^2}{\sum_i (x_i - \bar{x})^2}} = 0.1895,$$

slik at variablene

$$\frac{\hat{\alpha} - \alpha}{SE(\hat{\alpha})} \qquad \text{og} \qquad \frac{\hat{\beta} - \beta}{SE(\hat{\beta})}$$

er tilnærmet t-fordelte med $n - 2 = 4$ frihetsgrader. Vi kan derfor lage 95% konfidensintervaller for α og β ved å finne endepunktene

$$\hat{\alpha} \pm t_{4,0.025} \cdot SE(\hat{\alpha}) = -72.9989 \pm 2.776 \cdot 32.9813$$

og

$$\hat{\beta} \pm t_{4,0.025} \cdot SE(\hat{\beta}) = 0.8580 \pm 2.776 \cdot 0.1895.$$

Dette gir intervallene $(-164.555, 18.557)$ for α, og $(0.332, 1.384)$ for β. Siden 95% konfidensintervallet for β ikke inneholder null, kan vi si med 95% konfidens at vekten øker når kroppshøyden øker. For å finne ut hvor sikre vi egentlig er på dette, kan vi finne p-verdien til en ensidig test av hypotesene

$$H_0 : \ \beta \leq 0, \qquad H_1 : \ \beta > 0.$$

Bruker $\hat{\beta}/SE(\hat{\beta})$ som testobservator, og antar at denne er t-fordelt med fire frihetsgrader. Forkaster H_0 på signifikansnivå α hvis observert verdi av testobservatoren er

større enn den kritiske verdien $t_{4,\alpha}$. Utfra estimatene av $\hat{\beta}$ og $SE(\hat{\beta})$ har vi observert verdi

$$\frac{\hat{\beta}}{SE(\hat{\beta})} = \frac{0.8580}{0.1895} = 4.5277.$$

I kvantiltabellen til t-fordelingen finner vi at $t_{4,0.01} = 3.747$ og $t_{4,0.005} = 4.604$. Stigningstallet er dermed signifikant større enn null på signifikansnivå 0.01, men ikke på signifikansnivå 0.005. Testens p-verdi ligger et sted mellom disse nivåene. Det er altså mer enn 99% sikkert at vekten øker når kroppshøyden øker.

Forventet vekt for person med høyde $x = 175$ cm er

$$\hat{y} = \hat{\alpha} + \hat{\beta}x = -72.9989 \text{ kg} + 0.8580 \text{ kg/cm} \cdot 175 \text{ cm} = 77.1439 \text{ kg}.$$

Et 95% prediksjonsintervall for $\hat{y}$, i kg, har endepunktene

$$\hat{y} \pm t_{n-2,0.025} \cdot s\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_i (x_i - \bar{x})^2}} = \hat{y} \pm t_{4,0.025} \cdot s\sqrt{1 + \frac{1}{6} + \frac{(x-\bar{x})^2}{\sum_i (x_i - \bar{x})^2}} =$$

$$= 77.1439 \pm 2.776 \cdot 5.6052\sqrt{1 + \frac{1}{6} + \frac{(175 - 173.6667)^2}{875.3333}} =$$

$$= 77.1439 \pm 16.8214,$$

som svarer til intervallet $(60.3225, 93.9653)$.

Forventet vekt for basketballspilleren med høyde $x = 220$ cm er

$$\hat{y} = \hat{\alpha} + \hat{\beta}x = -72.9989 \text{ kg} + 0.8580 \text{ kg/cm} \cdot 220 \text{ cm} = 115.7521 \text{ kg}$$

og grensene for det tilhørende 95% prediksjonsintervallet er

$$\hat{y} \pm t_{4,0.025} \cdot s\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_i (x_i - \bar{x})^2}} =$$

$$= 115.7521 \pm 2.776 \cdot 5.6052\sqrt{1 + \frac{1}{6} + \frac{(220 - 173.667)^2}{875.3333}} =$$

$$= 115.7521 \pm 29.6017$$

som gir intervallet $(86.1504, 145.3538)$.

Når høyden er $x = 100$ cm får vi forventet vekt

$$\hat{y} = \hat{\alpha} + \hat{\beta}x = -72.9989 \text{ kg} + 0.8580 \text{ kg/cm} \cdot 100 \text{ cm} = 12.7970 \text{ kg},$$

med tilhørende 95% prediksjonsintervall $(-29.4345, 55.0285)$, funnet på samme måte som før. De målte høydene $x_1, x_2, \ldots, x_n$ som inngår i datasettet ligger alle mellom

156 cm og 191 cm. Når vi beregner forventet vekt for høydene $x = 220$ cm og $x = 100$ cm bruker vi modellen utenfor området hvor vi har observasjoner, og vi kan derfor ikke stole like mye på disse prediksjonene som på den første, for $x = 175$ cm. Dette gjenspeiles av bredden på prediksjonsintervallene. Når $x = 175$ cm er prediksjonsintervallet 34 kg langt, mot 59 kg når $x = 220$ cm. For $x = 100$ cm er den nedre grensen til prediksjonsintervallet sågar negativ. Siden en negativ vekt ikke gir fysisk mening, er dette en indikasjon på at vi har beveget oss utenfor modellens gyldighetsområde.

Tilleggsoppgave:

De tre kvadratsummene blir

$$SS_T = \sum_i (y_i - \bar{y})^2 = 770.00,$$

$$SS_R = \sum_i (\hat{y}_i - \bar{y})^2 = \sum_i (\hat{\alpha} + \hat{\beta} x_i - \bar{y})^2 = 644.33$$

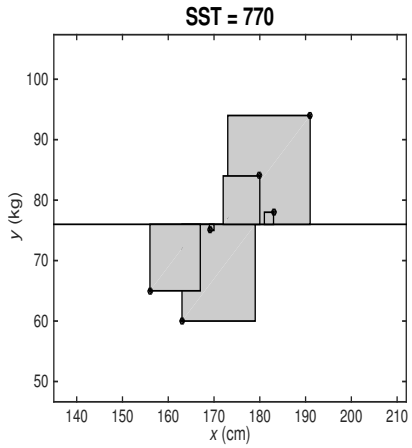
og

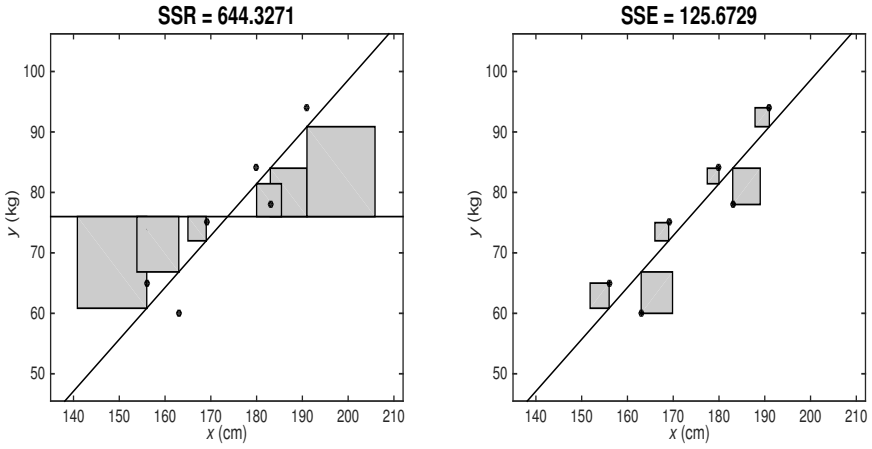
$$SS_E = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - \hat{\alpha} - \hat{\beta} x_i)^2 = 125.67.$$

Vi ser at $SS_T = SS_E + SS_R$, som alltid. Andelen av variasjonen i vekt som forklares av modellen er

$$r^2 = \frac{SS_R}{SS_T} = \frac{644.33}{770.0} = 0.8368.$$

Spredningsplott med inntegnede avvikskvadrater er vist nedenfor.





8.3. Lar X og Y være alder for gjenstander fra henholdvis vest og øst. Antar at X og Y er normalfordelt med forventninger μ_1 og μ_2 og samme varians σ^2. Vi skal teste $H_0 : \mu_1 = \mu_2$ (lik alder) mot $H_1 : \mu_1 \neq \mu_2$ (ulik alder). Beregner interpolert varians

$$S_p^2 = \frac{14 \cdot 177^2 + 11 \cdot 160^2}{25} = 28808$$

slik at $S_p = 169.7$.

Beregner

$$T = \frac{1499 - 1536}{169.7\sqrt{1/15 + 1/12}} = -0.563$$

etter ligning 8.3.

Kan ikke forkaste H_0 siden $|T| \not\geq t_{0.025} = 2.060$ ($15 + 12 - 2 = 25$ frihetsgrader).