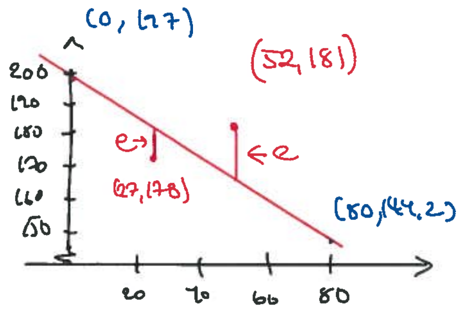
ST1201 - H2018

Oppgave 1

ST1201 H2018 Enkel lineær regresjon

a)



$\hat{\beta}_0 = 197$

$\hat{\beta}_1 = -0.66$

$\hat{\beta}_0 + \hat{\beta}_1 \cdot 80 = 197 - 0.66 \cdot 80 = 144.2$

Når x øker med en enhet vil i gjennomsnitt y øke med $\hat{\beta}_1$ enheter.

Residual:

$$e_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

Residualplottet viser at modellen passer godt. Punktene er tilfeldig plassert rundt 0, og det er like stor spredning av residualene for alle verdier av x.

Vi kunne også sjekket om residualene var tilnærmet normalfordelte, siden residualene er være prediksjoner av feilleddene og vi har antatt at feilleddene er normal fordelte. Normalfordelingen vi feks sjekke med et q-q-plott.

c) $H_0: \beta_1 = \beta_1^*$ mot $H_1: \beta_1 \neq \beta_1^*$

Det er kjent (Tabeller & formler i statistikk, side 36) at

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})Y_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\sum(x_i-\bar{x})Y_i}{SS_x}$$

og vi husker – eller kan finne at

$$E(\hat{\beta}_1) = \frac{\sum(x_i-\bar{x})}{SS_x}\cdot \overset{\beta_0+\beta_1 x_i}{\overset{\parallel}{E(Y_i)}} = \frac{1}{SS_x}\Big(\underbrace{\sum(x_i-\bar{x})}_{0}\cdot\beta_0 + \sum(x_i-\bar{x})x_i\beta_1\Big)$$

$$= \beta_1\cdot\frac{\overbrace{\sum(x_i-\bar{x})^2}^{SS_x}}{SS_x} = \underline{\beta_1}$$

siden $\sum(x_i-\bar{x})^2 = \sum(x_i-\bar{x})x_i$

$$Var(\hat{\beta}_1) = \frac{\overbrace{\sum(x_i-\bar{x})^2}^{SS_x}}{SS_x^2}\cdot\overbrace{Var(Y_i)}^{\sigma^2} = \underline{\frac{\sigma^2}{SS_x}}$$

og tilsvarende for $\hat{\beta}_1^*$.

$E(\hat{\beta}_1-\hat{\beta}_1^*) = E(\hat{\beta}_1)-E(\hat{\beta}_1^*) = \beta_1-\beta_1^*$ Når H_0 er sann er dette 0.

$$Var(\hat{\beta}_1-\hat{\beta}_1^*) \underset{\substack{\uparrow\\ \text{uavh.}\\ \text{utvalg}}}{=} Var(\hat{\beta}_1)+Var(\hat{\beta}_1^*) = \frac{\sigma^2}{SS_x}+\frac{\sigma^2}{SS_x^*}$$

$$= \sigma^2\underbrace{\left(\frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}+\frac{1}{\sum_{j=1}^{n^*}(x_j^*-\bar{x}^*)^2}\right)}_{SSX}$$

Estimator for σ^2: Inspirert fra Spørsmål - og Løsningsforslag s 572

$$S^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 + \sum_{j=1}^{n^*}(Y_j^* - \hat{\beta}_0^* - \hat{\beta}_1^* x_j^*)^2}{n + n^* - 4} = \frac{\sum e_i^2 + \sum e_j^{*2}}{n + n^* - 4}$$

Denne vil ha $E(S^2) = \sigma^2$ og $\frac{(n+n^*-4) \cdot S^2}{\sigma^2} \sim \chi^2_{n+n^*-4}$

Testobservator:

$$T = \frac{\hat{\beta}_1 - \hat{\beta}_1^*}{S \cdot \sqrt{SSX}} \sim t_{n+n^*-4}$$

$$\frac{Z}{\sqrt{\frac{V}{\nu}}} = \frac{\frac{\hat{\beta}_1 - \hat{\beta}_1^*}{\sigma\, SSX} \sim N(0,1)}{\frac{\sqrt{\frac{(n+n^*-4) S^2}{\sigma^2}} \sim \chi^2_{n+n^*-4}}{\sqrt{n+n^*-4}}}$$

Forkast H_0 når $|t_{obs}| > t_{\frac{\alpha}{2}, n+n^*-4}$.

Numeriske verdier

$$S = \sqrt{\frac{5581 + 4669}{\underbrace{52 + 52 - 4}_{100}}} = 10.12$$

$$\sqrt{SSX} = \sqrt{\frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2} + \frac{1}{\sum_{j=1}^{n^*}(x_j^*-\bar{x}^*)^2}}$$

$$= \sqrt{\frac{1}{6969} + \frac{1}{8409}} = 0.016$$

$$\hat{\beta}_1 - \hat{\beta}_1^* = -0.66 - (-0.70) = 0.04$$

$$t_{obs} = \frac{0.04}{10.12 \cdot 0.016} = 0.247$$

$$t_{0.1/2,\,52+52-4} = t_{0.05,\,100} = 1.658$$

$t_{obs} < 1.66 \Rightarrow$ $\underline{\underline{\text{ikke forkast } H_0}}$

Kommentar: dataene i denne oppgaven er simulerte, men basert på artikkelen Nes, Janszky, Wisløff, Støylen & Karlsen (2013) "Age-predicted maximal heart rate in healthy subjects: The HUNT Fitness Study" i Scand J Med Sci Sports 23: 697-704.

Oppgåve 2

a)

μ_j beskriv forventing i gruppe/populasjon evt. med behandling j.

σ^2 er variansen til enkeltobservasjonane

$SS_T = \sum_{i=1}^{10} \sum_{j=1}^{2} (Y_{ij} - \bar{Y}_{..})^2$ er total variasjon og beskriv variasjon om gjennomsnittet for alle observasjonane.

$SS_E = \sum_{i=1}^{10} \sum_{j=1}^{2} (Y_{ij} - \bar{Y}_{.j})^2$ er kvadratsum for feil og representerer variasjon innad i grupper/populasjon eller observasjonar med same behandling.

$SS_{TR} = \sum_{i=1}^{10} \sum_{j=1}^{2} (\bar{Y}_{.j} - \bar{Y}_{..})^2$ er kvadratsum for grupper/populasjonar/materiale/behandlingar og representerer variasjon mellom slike.

Ein kan samanlikne variasjon mellom grupper med variasjon innad i gruppert skalert med fridomsgrader

∂: $\frac{SS_{TR}}{SS_E/18}$. Er denne stor tyder det på at det er forskjellar.

b)

Kilder	Fr. grader	Kvadratsummar	Gj. kvadratsum	F
Materiale	1	0.841	0.841	0.14
Feil	18	111.164	6.18	
Total	19	112.005		

$H_0: \mu_A = \mu_B$ $\qquad H_1: \mu_A \neq \mu_B$

$\frac{SS_{TR}}{SS_E/18} = 0.14 \ll f_{1,18,0.05} = 4.41 \Rightarrow$ ingen grunn til å forkaste H_0

c) Vi pröver ut solane på 10 ~~gutar~~ forsökspersonar som brukar skoa heilt forskjellig, det vil sei vi har stor inhomogenitet imellom forsöks einingane. Dette kan vi löyse ved å la kvar forsökspersonen vere ei blokk. Solane blir randomisert innad kvar blokk. Forsöket er difor lagt opp som eit randomisert blokk design.

Vi har $SS_E^B + SS_B = SS_E \Rightarrow SS_E^B = 111.165 - 110.491 = \underline{0.674}$

SS_B har 9 fridomsgrader $\Rightarrow SS_E^B$ har $19-9-1=9$ fridomsgrader

$$\frac{SS_{TR}}{\frac{SS_E^B}{9}} = \frac{0.845}{\frac{0.675}{9}} = 11.21 > f_{1,9,0.05} = 5.12 \Rightarrow \text{vi forkaster}$$

H_0 i 2b) og konkluderer med at det er forskjellar mellom solane.

d)

$H_0: \mu_B \leq \mu_A \qquad H_1: \mu_B > \mu_A$

$\Leftrightarrow \quad H_0: \mu_B - \mu_A \leq 0 \qquad H_1: \mu_B - \mu_A > 0$

$\mu_B - \mu_A$ er ein kontrast. Estimator for $\mu_B - \mu_A = \bar{Y}_B - \bar{Y}_A = \bar{Y}_2 - \bar{Y}_1$

~~Variansen ($\bar{Y}_A - \bar{Y}_B$)~~ $Var(\bar{Y}_A - \bar{Y}_B) = \frac{\sigma^2}{10} + \frac{\sigma^2}{10} = \frac{\sigma^2}{5}$

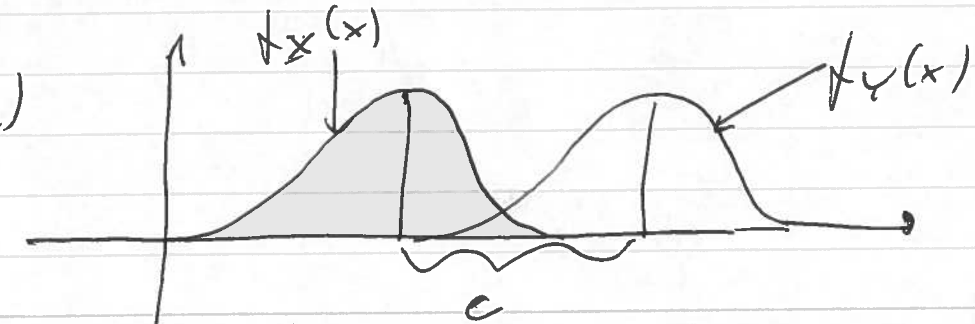
Under H_0: $\dfrac{\bar{Y}_B - \bar{Y}_A}{S\sqrt{\frac{1}{5}}} \sim t_9$. $\quad S^2 = \dfrac{0.675}{9} = 0.0749$

$$\frac{\bar{y}_B - \bar{y}_A}{s\sqrt{\frac{1}{5}}} = \frac{0.41}{\sqrt{\frac{0.0749}{5}}} = 3.35 > t_{9,0.01} = 2.82$$

og vi forkaster H_0 på 1% nivå og konkluderer med at slitasjen med sole B er större enn med sole A

Oppgave 3 (Wilcoxon)

a)



Under H_0 er $c = 0$, og tetthetene dekke da hverandre.

Test: Ranger alle de $n+m$ observasjonene i voksende rekkefølge, og la W_1 være summen av rangene for X-observasjonene.

	3	6	2	4	
X_i:	13.1	16.6	8.8	14.1	
Y_i:	15.7	19.1	16.9	18.9	8.2
	5	9	7	8	1

$$W_1 = 3+6+2+4 = 15$$

$$W_2 = 5+9+7+8+1 = 30$$

$$\left(\text{sum } 45 = \frac{9 \cdot 10}{2}\right)$$

$$\text{p-verdi} = 2 \cdot \min\left(P(W_1 \leq 15),\ P(W_1 \geq 15)\right)$$

* Tabell gir (vi har $\frac{n(n+1)}{2} = \frac{4 \cdot 5}{2} = 10$

$$P(W_1 \leq 15) = P(W_1 - 10 \leq 5)$$

$$= P(U \leq 5) = 0.143$$

$$P(W_1 \geq 15) = P(U_1 \geq 5) = 1 - P(U_1 \leq 4)$$

$$= 1 - 0.095 = 0.905$$

så p-verdi $= 2 \cdot 0.143 = \underline{0.286}$

b) I datasettet:

Par med $(Y_j < X_i)$:

$(8.2,\ 13.1)$, $(8.2,\ 16.6)$, $(15.7,\ 16.6)$,

$(8.2,\ 8.8)$, $(8.2,\ 14.1)$

dvs. $U_1 = 5$.

Skal vise at

$$U_1 = W - \frac{n_1(n_1+1)}{2}$$

~~(Stemmer for eksempelet: 5 =~~

Det er $R_1 - 1$ Y-obs. mindre enn X_1'.
Det er nemlig de som har rang $1, 2, \dots, R_1 - 1$

Tilsvarende er det $R_2 - 2$ Y-obs som er mindre enn X_2'. Det er de som har rang $1, 2, \dots, R_2 - 1$, mens X_1' må tas bort.

Dermed blir

$$U_1 = (R_1 - 1) + (R_2 - 2) + \dots + (R_{n_1} - n_1)$$
$$= W - \frac{n_1(n_1+1)}{2} \quad \text{Q.E.D.}$$

c) Hvis H_0 gjelder har X og Y samme fordeling, så

$$P(Y < X) = P(X < Y)$$

Siden

$$P(Y<X) + P(X<Y) = 1$$

(kontinuerlige fordelinger) må da

$$\underline{P(Y<X) = \frac{1}{2}}$$

$$\hat{p} = \frac{U_1}{n_1 n_2}$$

Skriv

$$U_1 = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} Z_{ij}$$

der $Z_{ij} = \begin{cases} 1 & \text{hvis } Y_j < X_i \\ 0 & \text{ellers} \end{cases}$

Da er

$$E(U_1) = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} E(Z_{ij})$$

$$= \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \overbrace{P(Z_{ij} = 1)}^{=p}$$

$$= n_1 n_2 p$$

Dermed: $\underline{\underline{E\left(\frac{U_1}{n_1 n_2}\right)}} = \underline{\underline{p}}$

Under H_0 er $p = \frac{1}{2}$

dvs $E(U_1) = \frac{n_1 n_2}{2}$

Men da er

$$E(W_1) = E(U_1) + \frac{n_1(n_1+1)}{2}$$

$$= \frac{n_1 n_2}{2} + \frac{n_1(n_1+1)}{2}$$

$$= \frac{n_1(n_1+n_2+1)}{2} \quad \text{QED}$$