Løsningsforslag: STK2120-v15.

Oppgave 1

a) Den statistiske modellen er: $X_{ij} = \mu_i + \epsilon_{ij},\ j = 1, \ldots, J,\ i = 1, \ldots, I$. Her indekserer $i = 1, \ldots, I$ gruppene og $j = 1, \ldots, J$ observasjone innen hver gruppe. Feilleddene $\epsilon_{ij},\ j = 1, \ldots, J,\ i = 1, \ldots, I$ antas å være uavhegige $N(0, \sigma^2)$ med ukjent varians σ^2.

Den viktigste nullhypotesen er at det ikke er noen forskjell mellom gruppene, dvs. $H_0 : \mu_1 = \cdots = \mu_I$.

La x_{ij} være realisasjonen av den tilfeldige variabelen $X_{ij}\ j = 1, \ldots, J,\ i = 1, \ldots, I$. Da viser utskiften

- Summen av kvadratavvik innen grupper:$SSR = \sum_{i=1}^{I} \sum_{j=1}^{J} (x_{ij} - \bar{x}_{i.})^2$ der $\bar{x}_{i.} = \frac{1}{J} \sum_{j=1}^{J} x_{ij}$. I utskriften svarer det til 111601.
- Summen av kvadratavvik mellom grupper:$SSTr = J \sum_{i=I} (\bar{x}_{i.} - \bar{x}_{..})^2$ der $\bar{x}_{..} = \frac{1}{IJ} \sum_{j=1}^{J} x_{ij}$. I utskriften svarer det til 246726.
- Antallet frihetsgrader angir frihetsgradene i χ^2 fordelingen til $SSTr/\sigma^2$ (I-1=3 her) og til SSR/σ^2 (I(J-1)=36 i dette tilfellet).
- Gjenomsnittlig kvadratsum $SSR/I(J-1)$ og $SSTr/(I-1)$ og forholdet mellom dem $F = \frac{SSTr/(I-1)}{SSR/I(J-1)}$.

Under H_0 er F Fisher-fordelt med $I-1$ og $I(J-1)$ frihetsgrader og hypotesen forkastes for store verdier av F. I dette tilfellet er den realiserte verdien av F 26.529, som gir en p-verdi på langt mindre enn et nivå på 0.0001, slik at nullhypotesen må forkastes.

b) $H_0 : (\mu_A + \mu_B + \mu_C)/3 - \mu_D = 0$.
$(\mu_A + \mu_B + \mu_C)/3 - \mu_D$ kan estimeres med $Contr = (\bar{X}_A + \bar{X}_B + \bar{X}_C)/3 - \bar{X}_D$ som er normalfordelt. Under H_0 er forventingen null og variansen $Var(Contr) = (1/3)^2(\sigma^2/10 + \sigma^2/10 + \sigma^2/10) + \sigma^2/10) = \sigma^2(1/3+1)/10 = 2\sigma^2/15$. Da er under H_0, $Contr/\sqrt{2\sigma^2/15}$ normalfordelt og $t = Contr/\sqrt{2s^2/15}$ t-fordelt med 36 frihetsgrader der $s^2 = SSR/36$. Med de oppgitte tallene gir dette: $t = [600.1 + 688.6 + 667.4)/3 - 487.1]/\sqrt{(3100 * 2/15)} = 8.11$ som gir klar forkastning både for ensidig og tosidig hypotese. Den største tabullerte kritiske verdien i en t(36) fordeling er 3.582 svarende til nivå 0.0005.

c) Betydningen av kjønn kan vise seg på to måter. Enten ved at det er samspill mellom kjønn og skole og ved at det er hovedeffekt av kjønn.

Samspill er en høyere ordens effekt, og først når en har funnet ut at denne effekten ikke er til stede gir det mening å undersøke hovedeffektene.

Samspill testes ved opplysningene i tredje rad. Vi ser at p-verdien for F , 0.33, er så stor at hypotesen om intet samspill ikke kan forkastes.

Hovedeffekten av kjønn kan fastslås ved rad 2. Nullhypotesen om ingen hovedeffekt forkastes heller ikke.

Så hverken samspill mellom kjønn og skole eller hovedeffekt av kjønn ser ut til å være til stede.

En annen måte å teste om en hovedeffekt er til stede hvis det er fastslått at det ikke er samspill, er å tilpasse en modell uten samspill. Da vil SSE/σ^2 og $(I-1)(SSE/\sigma^2$ være uavhengige og henholdsvis $\chi^2_{IJ(K-1)}$ of $\chi^2_{(I-1)(J-1)}$

fordelt slik at $(SSAB + SSE)/(IJK - I - J + 1)$ er en forventningsrett estimator for σ^2. Dette gir en F-obsetrvator for hovedeffekt av kjønn (faktor B) lik $[SSB/(J-1)]/[(SSB + SSAB)/(IJK - I - J + 1)]$. I dette tilfellet $(6200/1)/((10755 + 94826)/35) = 2.06$.

d) At linjene er forholdsvis paralelle tyder på at det ikke er noe samspill, og at avstanden mellom linjene er forenelig med at det ikke er noen hovedeffekt av kjønn, selv om det må testes om forskjellen er signifikant.

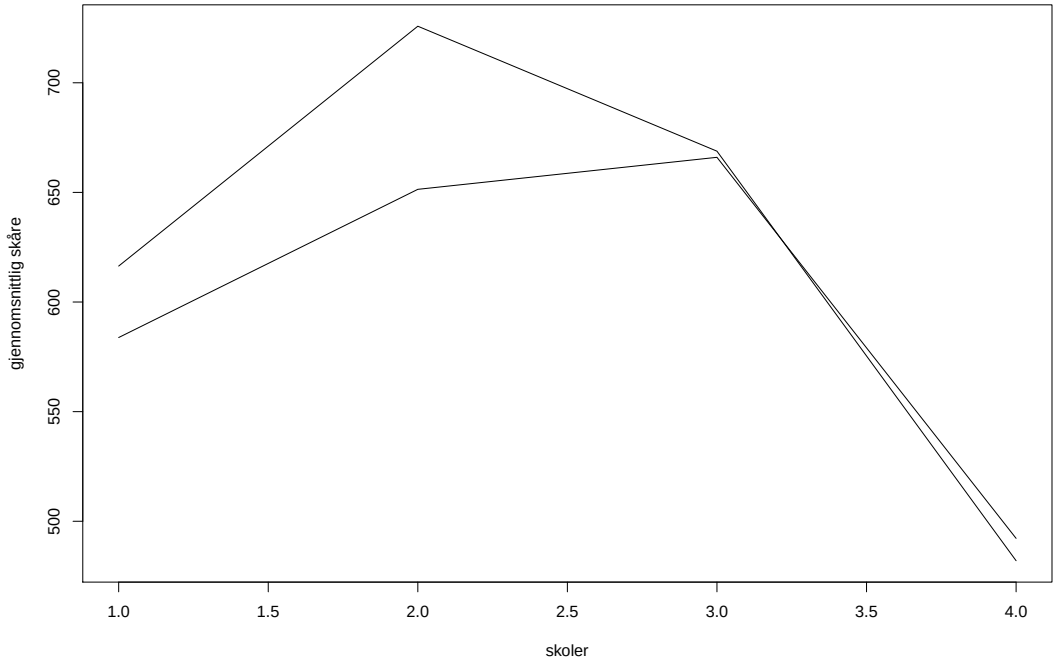

FIGURE 1. Interaksjonsplot. Nederste linje er gjennomsnittene for jentene. Skole A er 1, skole B 2 osv.

Oppgave 2

a) La indeksene $1, \cdots, 7$ angi målingene for kvinnene og indeksene $8, \cdots, 14$ angi målingene for menn.

Den første modellen er en enkel lineær regresjonsmodell

$$Y_i = \alpha_0 + \alpha_1 x_i + \epsilon_i$$

der feilleddene $\epsilon_i,\ 1, \cdots, 14$ antas å være ukorrelerte tilfeldige variable med forventing 0 og konstant varians σ^2, eller den sterkere forutsetningen uavhendige normalfordelte $N(0, \sigma^2)$.

Den andre modellen kan formuleres som

$$Y_i = \beta_0 + \beta_1 kjønn_i + \beta_2 x_i + \epsilon_i$$

der $kjønn$ er en faktor. Med cornerpoint parameterisering har den verdiene 0 for en av gruppene og 1 for den andre. Siden den nederste linja svarer til kvinner og $\hat{\beta}_1$ er positiv slik at linja for menn blir høyere, må $kjønn_i = 0,\ i = 1, \cdots, 7$ og $kjønn_i = 1,\ i = 8, \cdots, 14$.

b) Den midterste linja: a= 1.71166 og b= 0.07946.
Den nederste linja (kvinner): a= 1.641660 og b=0.079457.
Den øverste linja: (menn): a=1.641660 +0.140000= 1.781660 og b=0.079457.

c) For den enkle lineære regresjonsmodellen:

$$\mathbf{X}_1 = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_7 \\ 1 & x_1 \\ \vdots & \vdots \\ 1 & x_7 \end{pmatrix}.$$

For den andre modellen med cornerpoint parameterisering

$$\mathbf{X}_2 = \begin{pmatrix} 1 & 0 & x_1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & x_7 \\ 1 & 1 & x_1 \\ \vdots & \vdots & \vdots \\ 1 & 1 & x_7 \end{pmatrix}.$$

d) En alternativ måte å uttrykke konstantleddet og faktoren kjønn er ved to ortogonale kollonner, dvs.

$$\mathbf{X}_3 = \begin{pmatrix} 1 & 0 & x_1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & x_7 \\ 0 & 1 & x_1 \\ \vdots & \vdots & \vdots \\ 0 & 1 & x_7 \end{pmatrix}.$$

Da blir

$$\mathbf{X}_3'\mathbf{X}_3 = \begin{pmatrix} 1 & \dots & 1 & 0 & \dots & 0 \\ 0 & \dots & 0 & 1 & \dots & 1 \\ x_1 & \dots & x_7 & x_1 & \dots & x_7 \end{pmatrix} \begin{pmatrix} 1 & 0 & x_1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & x_7 \\ 0 & 1 & x_1 \\ \vdots & \vdots & \vdots \\ 0 & 1 & x_7 \end{pmatrix} = \begin{pmatrix} 7 & 0 & 7\bar{x} \\ 0 & 7 & 7\bar{x} \\ 7\bar{x} & 7\bar{x} & 2\sum_{j=1}^{7} x_j^2 \end{pmatrix}$$

e) Stigningsforholdet er tredje element i vektoren $(\mathbf{X}_3'\mathbf{X}_3)^{-1}\mathbf{X}_3'\mathbf{y}$. Fra den oppgitte formelen er det produktet av tredje rad i $(\mathbf{X}_3'\mathbf{X}_3)^{-1}$ som er

$$(7(-7\bar{x}), 7(-7\bar{x}), 49)/(49\times 2\sum_{j=1}^{7} x_j^2 - 2\times 7(7\bar{x})^2) = (-49\bar{x}, -49\bar{x}, 49)/49\times 2(\sum_{j=1}^{7}(x_j - \bar{x})^2,$$

med vektoren $\mathbf{X}_3'\mathbf{y}$. Men

$$\mathbf{X}_3'\mathbf{y} = \begin{pmatrix} 7\bar{y}_1 \\ 7\bar{y}_2 \\ \sum_{j=1}^{7} y_j x_j + \sum_{j=8}^{14} y_j x_j \end{pmatrix}$$

der $\bar{y}_1 = \frac{1}{7}\sum_{j=1}^{7} y_j$, $\bar{y}_2 = \frac{1}{7}\sum_{j=8}^{14} y_j$ og $x_1 = x_8, \cdots, x_7 = x_{14}$. Derfor blir produktet

$$[(-49\bar{x})(7\bar{y}_1)+(-49\bar{x})(7\bar{y}_2)+49\sum_{j=1}^{14} y_j x_j]/[49\times 2\sum_{j=1}^{7}(x_j-\bar{x})^2] = \sum_{j=1}^{14} y_j(x_j-\bar{x})/\sum_{j=1}^{14}(x_j-\bar{x})^2$$

som er det samme som stigningskoeffesienten i den enkle lineære regresjonsmodellen svarende til den midterste linja i figuren.

Oppgave 3

a) Likelihood funksjonen er

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\theta}\exp(-x_i/\theta) = \frac{1}{\theta^n}\exp(-\sum_{i=1}^{n} x_i/\theta) = \frac{1}{\theta^n}\exp(-n\bar{x}/\theta)$$

og log likelihood

$$l(\theta) = \log(L(\theta)) = -n\log(\theta) - n\bar{x}/\theta$$

der $\bar{x} = \sum_{i=1}^{n} x_i$.

Scorefunksjonen, $s(\theta) = \frac{\partial l}{\partial \theta}$, er derfor

$$s(\theta) = \frac{-n}{\theta} + \frac{n\bar{x}}{\theta^2}.$$

SME $\hat{\theta}$ finnes derfor ved å løse $s(\theta) = 0$, dvs. $\hat{\theta} = \bar{x}$.

Fisher informasjonen for de n observasjonene, $I(\theta) = -E[\frac{\partial^2 l}{\partial \theta^2}]$, er

$$I(\theta) = -\frac{n}{\theta^2} + \frac{2n\theta}{\theta^3} = \frac{n}{\theta^2}$$

siden fra formelsamling STK1100/STK1110 er $E(X_i) = \theta$.

b) Fra Cramer-Rao's ulikhet følger at for alle forventningsrette estimatorer $\tilde{\theta}$ for θ, er $Var(\tilde{\theta}) \geq I(\theta)^{-1} = \theta^2/n$. Spesielt siden $E(X_i) = \theta$, er $E(\hat{\theta}) = E(\bar{X}) = \theta$, så $\hat{\theta}$ er forventningsrett. Men $Var(\hat{\theta}) = Var(\bar{X}) = Var(X_i)/n = \theta^2/n$, siden fra formelsamling STK1100/STK1110 er $Var(X_i) = \theta^2$. Variansen til $\hat{\theta}$ er derfor lik den nedre grensen, slik at det ikke finnes noen forventningsrett estimator med mindre varians enn $\hat{\theta}$.

c) Fra de generelle resultatene for sannsynlighetsmaksimeringsestimatoren følger at i dette tilfellet er $(\hat{\theta} - \theta)$ er tilnærmet $N(0, 1/I(\theta))$, dvs. $\sqrt{n}(\hat{\theta} - \theta)$ er tilnærmet $N(0, \theta^2)$. Dette følger også direkte fra sentralgrenseteoremet siden $\hat{\theta} = \bar{X}$.

Et konfidensintervall med tilnærmet konfidensgrad 95% har derfor grenser $\hat{\theta} \pm 1.96\hat{\theta}/\sqrt{n}$.

En test med tilnærmet nivå 0.05 for $H_0 : \theta = \theta_0$ mot alternativet $H_0 : \theta \neq \theta_0$ forkaster hvis $\frac{\sqrt{n}(\hat{\theta}-\theta_0)}{\theta_0} > 1.96$ eller $\frac{\sqrt{n}(\hat{\theta}-\theta_0)}{\theta_0} < -1.96$. Alternativt kan $\hat{\theta}$ benyttes i nevneren.

d) Maksimum av likelihooden under de apriori betingensene er $L(\hat{\theta}) = \frac{1}{\bar{x}^n}\exp(-n)$. Maksimum av likelihooden under nullhypotesen er $L(\theta_0) = \frac{1}{\theta_0^n}\exp(-\frac{n\bar{x}}{\theta_0})$. Da er sannsynlighetskvoten

$$LR = (\frac{\bar{x}}{\theta_0})^n \exp(-n\hat{\theta}/\theta_0 + n) = (\frac{\hat{\theta}}{\theta_0})^n \exp(-n(\hat{\theta} - \theta_0)/\theta_0).$$

Fra generelle resultater om sannsynlighetskvotetesten følger det at $-2\log(LR) = -2n\log(\hat{\theta}/\theta_0) + 2n(\hat{\theta} - \theta_0)/\theta_0$ under nullhypotesen er tilnærmet χ^2 fordelt når n er stor. Antallet frihetsgrader er differansen mellom antall parametre når ingen restriksjoner er pålagt og antallet parametre under nullhypotesen. I dette tilfellet er det henholdsvis 1 og 0. Sannsynlighetskvotetesten består derfor i å forkaste hvis $-2\log(LR) > \chi^2_{0.05,1}$ der $\chi^2_{\alpha,1}$ er $1-\alpha$-fraktilen i χ^2 fordelingen med en frihetsgrad.

Det følgende spørres det ikke om i oppgaven, men det kan være av interesse. Ved en Taylorutvikling av $\log(u)$ om $u = 1$ får man $\log(u) = (u-1) - \frac{1}{2}(u-1)^2 + \cdots$. Anvendt på $-2\log(LR)$ gir dette $-2\log(LR) \approx$

$$-2n(\hat{\theta}/\theta_0 - 1) + \frac{1}{2}2n(\hat{\theta}/\theta_0 - 1)^2 + 2n(\hat{\theta} - \theta_0)/\theta_0 = +n[(\hat{\theta} - \theta_0)/\theta_0]^2.$$

som gir tilnærmet samme resultat som den tosidige testen i punkt c).