## Opgave 1: Regressionsanalyse

Lad $(u_1, x_1), \ldots, (u_n, x_n)$ være $n$ par af reelle tal. Vi skal nu bestemme den rette linie, der passer bedst med disse talpar i den forstand at summen

$$\sum_{s=1}^{n}(x_s - \alpha - \beta u_s)^2$$

minimeres. Man finder altså den linie, $x = \hat{\alpha} + \hat{\beta}u$, for hvilken summen af kvadraterne på punkternes lodrette afstand til linien er mindst mulig. Dette kaldes mindste kvadraters metode.

### (a)

Vi antager at ikke alle $u$'erne er ens, og at $\sum_{s=1}^{n} u_s = 0$. Vi skal vise, at linien, der bestemmes ved mindste kvadraters metode, er givet ved

$$\hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \hat{\beta} = \frac{\sum_{s=1}^{n} x_s u_s}{\sum_{s=1}^{n} u_s^2}$$

$$\sum_{s=1}^{n}(x_s - \alpha - \beta u_s)^2 = \sum_{s=1}^{n}(x_s^2 - 2\alpha x_s + \alpha^2 - 2\beta u_s x_s + 2\alpha\beta u_s + \beta^2 u_s^2)$$

$$\sum_{s=1}^{n}(x_s - \alpha)^2 + \sum_{s=1}^{n}(-2\beta u_s x_s + \beta^2 u_s^2) + 2\alpha\beta\sum_{s=1}^{n} u_s,$$

hvor det bemærkes, at det sidste led er nul. Vi skal prøve at minimere de to summer. Det bemærkes, at begge er "glade"parabler, altså har minimum, hvor den afledte er nul. Vi differentierer altså mht. hhv. $\alpha$ og $\beta$.

$$\frac{d}{d\alpha}\sum_{s=1}^{n}(x_s - \alpha)^2 = \sum_{s=1}^{n}\frac{d}{d\alpha}(x_s - \alpha)^2 = \sum_{s=1}^{n}(2\alpha - 2x_s) = 2n\alpha - 2\sum_{s=1}^{n} x_s$$

$$\frac{d}{d\beta}\sum_{s=1}^{n}(-2\beta u_s x_s + \beta^2 u_s^2) = \sum_{s=1}^{n}(-2u_s x_s + 2\beta u_s^2) = 2\beta\sum_{s=1}^{n} u_s^2 - 2\sum_{s=1}^{n} u_s x_s$$

Det vil sige:

$$0 = 2n\hat{\alpha} - 2\sum_{s=1}^{n} x_s \Rightarrow \hat{\alpha} = \frac{\sum_{s=1}^{n} x_s}{n}$$

$$0 = 2\hat{\beta}\sum_{s=1}^{n} u_s^2 - 2\sum_{s=1}^{n} u_s x_s \Rightarrow \hat{\beta} = \frac{\sum_{s=1}^{n} u_s x_s}{\sum_{s=1}^{n} u_s^2}$$

Lad $X_1, \ldots, X_n$ være uafhængige, normalfordelte stokastiske variable, hvor $X_s \sim N(\alpha + \beta(t_s - \bar{t}), \sigma^2)$ med $(\alpha, \beta) \in \mathbb{R}^2$ og $\sigma^2 > 0$. Antag, at alle t'erne ikke er ens, og definer

$$\bar{X} = \frac{1}{n}\sum_{s=1}^{n} X_s, \qquad \text{og} \qquad \hat{\beta} = \frac{\sum_{s=1}^{n} X_s(t_s - \bar{t})}{\mathrm{SSD}_t},$$

hvor

$$\bar{t} = \frac{1}{n}\sum_{s=1}^{n} t_s, \qquad \text{og} \qquad \text{SSD}_t = \sum_{s=1}^{n}(t_s - \bar{t})^2.$$

## (b)

Vis, at

$$\bar{X} \sim N(\alpha, \sigma^2/n) \qquad \text{og} \qquad \hat{\beta} \sim N(\beta, \sigma^2/\text{SSD}_t)$$

samt at $\bar{X}$ og $\hat{\beta}$ er ukorrelerede:

Det ses, at der gælder

$$\sum_{s=1}^{n}(t_s - \bar{t}) = \sum_{s=1}^{n} t_s + \sum_{s=1}^{n} \bar{t} = \sum_{s=1}^{n} t_s + \sum_{s=1}^{n} \frac{1}{n}\sum_{i=1}^{n} t_i = \sum_{s=1}^{n} t_s + \sum_{i=1}^{n} t_i = 0$$

Da $X_1, \ldots, X_n$ er uafhængige, normalfordelte stokastiske variable, følger det af sætning 6.3.12, at $\bar{X} = \frac{1}{n}\sum_{s=1}^{n} X_s$ er normaltfordelt med middelværdi

$$\begin{aligned} E(\bar{X}) &= E\left(\frac{1}{n}\sum_{s=1}^{n} X_s\right) = \frac{1}{n}E\left(\sum_{s=1}^{n} X_s\right) \\ &= \frac{1}{n}\sum_{s=1}^{n} E(X_s) = \frac{1}{n}\sum_{s=1}^{n}(\alpha + \beta(t_s - \bar{t})) \\ &= \frac{1}{n}\left(n\alpha + \beta\sum_{s=1}^{n}(t_s - \bar{t})\right) = \alpha, \end{aligned}$$

hvor de almindelige regneregler for middelværdi blev benyttet, og sætning 6.3.12 blev brugt ved det tredje lighedstegn.
Ved brug af regneregler for varians og 6.3.12 igen fås ligeså:

$$\begin{aligned} \text{Var}(\bar{X}) &= \text{Var}\left(\frac{1}{n}\sum_{s=1}^{n} X_s\right) = \frac{1}{n^2}\text{Var}\left(\sum_{s=1}^{n} X_s\right) \\ &= \frac{1}{n^2}\sum_{s=1}^{n}\text{Var}(X_s) = \frac{1}{n^2}\sum_{s=1}^{n}\sigma^2 = \frac{\sigma^2}{n}. \end{aligned}$$

Altså haves $\bar{X} \sim N(\alpha, \sigma^2/n)$.

Som ovenfor følger ligledes af sætning 6.3.12 at $\hat{\beta}$ er normaltfordelt med middelværdi:

$$
\begin{aligned}
E(\hat{\beta}) &= E\left(\frac{\sum_{s=1}^{n} X_s(t_s-\bar{t})}{\mathrm{SSD}_t}\right) = \frac{\sum_{s=1}^{n} E(X_s)(t_s-\bar{t})}{\mathrm{SSD}_t} \\
&= \frac{\sum_{s=1}^{n}(\alpha+\beta(t_s-\bar{t}))(t_s-\bar{t})}{\mathrm{SSD}_t} = \frac{\sum_{s=1}^{n}\alpha(t_s-\bar{t})+\beta(t_s-\bar{t})^2}{\mathrm{SSD}_t} \\
&= \frac{\sum_{s=1}^{n}\alpha(t_s-\bar{t})+\beta\mathrm{SSD}_t}{\mathrm{SSD}_t} = \beta + \frac{\alpha\sum_{s=1}^{n}(t_s-\bar{t})}{\mathrm{SSD}_t} = \beta,
\end{aligned}
$$

og varians:

$$
\begin{aligned}
\mathrm{Var}(\hat{\beta}) &= \mathrm{Var}\left(\frac{\sum_{s=1}^{n} X_s(t_s-\bar{t})}{\mathrm{SSD}_t}\right) = \frac{1}{\mathrm{SSD}_t^2}\sum_{s=1}^{n}\mathrm{Var}(X_s(t_s-\bar{t})) \\
&= \frac{1}{\mathrm{SSD}_t^2}\sum_{s=1}^{n}(t_s-\bar{t})^2\mathrm{Var}(X_s) = \frac{1}{\mathrm{SSD}_t^2}\sum_{s=1}^{n}(t_s-\bar{t})^2\sigma^2 \\
&= \frac{\sigma^2}{\mathrm{SSD}_t^2}\sum_{s=1}^{n}(t_s-\bar{t})^2 = \frac{\sigma^2}{\mathrm{SSD}_t^2}\mathrm{SSD}_t = \frac{\sigma^2}{\mathrm{SSD}_t}.
\end{aligned}
$$

Altså haves $\hat{\beta} \sim N(\beta, \sigma^2/\mathrm{SSD}_t)$.

Fra definition 3.8.6 har vi, at

$$
\mathrm{corr}(\bar{X}, \hat{\beta}) = \frac{\mathrm{Cov}(\bar{X}, \hat{\beta})}{\sqrt{\mathrm{Var}(\bar{X})\mathrm{Var}(\hat{\beta})}}
$$

så for at vise at $\bar{X}$ og $\hat{\beta}$ er ukorrelerede, skal vi blot vise, at $\mathrm{Cov}(\bar{X}, \hat{\beta}) = 0$.

Ved brug af (3.8.3), (3.8.4), (3.8.5) og sætning 3.8.3 ses det at

$$
\begin{aligned}
\mathrm{Cov}(\bar{X}, \hat{\beta}) &= \mathrm{Cov}\left(\frac{1}{n}\sum_{s=1}^{n} X_s, \frac{\sum_{s=1}^{n} X_s(t_s-\bar{t})}{\mathrm{SSD}_t}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathrm{Cov}\left(X_i, \frac{\sum_{s=1}^{n} X_s(t_s-\bar{t})}{\mathrm{SSD}_t}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathrm{Cov}\left(\frac{\sum_{s=1}^{n} X_s(t_s-\bar{t})}{\mathrm{SSD}_t}, X_i\right) \\
&= \frac{1}{n\mathrm{SSD}_t}\sum_{i=1}^{n}\sum_{s=1}^{n}(t_s-\bar{t})\mathrm{Cov}\left(X_s, X_i\right) \\
&= \frac{1}{n\mathrm{SSD}_t}\sum_{s=1}^{n}(t_s-\bar{t})\mathrm{Var}(X_s) \\
&= \frac{\sigma^2}{n\mathrm{SSD}_t}\sum_{s=1}^{n}(t_s-\bar{t}) = 0.
\end{aligned}
$$

## Opgave 2: t-test for to stikprøver

### (c)

I følge sætning 8.3.3 er $\bar{X}_1$ normalfordelt med middelværdi $\mu_1$ og varians $\frac{\sigma^2}{n_1}$, og $(n_1-1)s_1^2$ er $\sigma^2\chi^2$-fordelt med $n_1-1$ frihedsgrader, og dermed er $s_1^2$ $\frac{\sigma^2}{n_1-1}\chi^2$-fordelt med $n_1-1$ frihedsgrader. Endvidere er $\bar{X}_1$ og $(n_1-1)s_1^2$ uafhængige ifølge sætning 8.3.3, og dermed er $\bar{X}_1$ og $s_1^2$ uafhængige. Tilsvarende ses det, at $\bar{X}_2$ er normalfordelt med middelværdi $\mu_2$ og varians $\frac{\sigma^2}{n_2}$, $s_2^2$ er $\frac{\sigma^2}{n_2-1}\chi^2$-fordelt med $n_2-1$ frihedsgrader, og $\bar{X}_2$ og $s_2^2$ er uafhængige.

Ved at sætte $k=n_1$, $n=n_1+n_2$, og $X_{11}=X_1, X_{12}=X_2,\ldots,X_{1k}=X_k, X_{21}=X_{k+1}, X_{22}=X_{k+2},\ldots,X_{2n_2}=X_n$ og ved at lade $\varphi$ være funktionen fra $\mathbb{R}^k$ ind i $\mathbb{R}$ der sender $(x_1,\ldots,x_k)$ ind i $\frac{1}{k}\sum_{i=1}^k x_i$ og lade $\psi$ være funktionen fra $\mathbb{R}^{n-k}$ ind i $\mathbb{R}$, der sender $(x_1,\ldots,x_{n-k})$ ind i $\frac{1}{n-k-1}\sum_{s=1}^{n-k}(x_s-\frac{1}{n-k}\sum_{s=1}^{n-k}x_s)^2$ ses det ved brug af sætning 6.2.3 3) [3) står først nævnt efter beviset for sætningen] at $\bar{X}_1$ og $s_2^2$ er uafhængige. Tilsvarende ses det at $\bar{X}_1$ og $\bar{X}_2$, $s_1^2$ og $\bar{X}_2$ samt $s_1^2$ og $s_2^2$ er uafhængige.

### (d)

Vi skal finde fordelingen af

$$s^2=\frac{1}{n-2}\left((n_1-1)s_1^2+(n_2-1)s_2^2\right)\ ,\ n=n_1+n_2.$$

Vi så i (c), at

$$s_1^2\sim\frac{\sigma^2}{n_1-1}\chi^2(n_1-1)\text{ og }s_2^2\sim\frac{\sigma^2}{n_2-1}\chi^2(n_2-1)$$

Men så er

$$\sigma^2S_1=(n_1-1)s_1^2\sim\sigma^2\chi^2(n_1-1)\text{ og }\sigma^2S_2=(n_2-1)s_2^2\sim\sigma^2\chi^2(n_2-1)$$

Vi kan nu trække konstanten $\sigma^2$ uden for parentesen og omskrive de to stokastiske variable $S_1, S_2$ til $\Gamma$-fordelinger, jævnfør MS s.222m, og vi får, at

$$s^2=\frac{\sigma^2}{n-2}(S_1+S_2),$$

hvor $S_i$ er $\Gamma$-fordelt med formparameter $\alpha_i=\frac{n_i-1}{2}$, $i=1,2$ og skalaparameter $\beta_i=2$. Ved hjælp af MS Sætning 8.1.3 kan vi sammenskrive $S_1+S_2$ til en ny $\Gamma$-fordeling, $S$, med formparameter $\alpha=\frac{n_1-1+n_2-1}{2}=\frac{n-2}{2}$ og skalaparameter $\beta=2$. Men så er $S\sim\chi^2(n-2)$ med $f=n-2$ frihedsgrader, igen jævnfør MS s.222m. Altså er

$$Z=s^2\sim\frac{\sigma^2}{n-2}\chi^2(n-2)=\frac{\sigma^2\chi^2(f)}{f}$$

ifølge MS s.225m. $s^2$ er altså en $\Gamma$-fordeling med formparameter $k/2$ og skalaparameter $2\sigma^2/f$, *eller* $s^2\sim\frac{\sigma^2}{n-2}\chi^2(n-2)$.

## (e)

Ifølge (c) er $\bar{X}_1$ og $\bar{X}_2$ uafhængige normalfordelinger med h.h.v. middelværdi $\mu_1$ og $\mu_2$ og varians $\frac{\sigma^2}{n_1}$ og $\frac{\sigma^2}{n_2}$. Dermed er $-\bar{X}_2$ normalfordelt med middelværdi $-\mu_2$ ifølge sætning 5.2.5 og varians $\frac{\sigma^2}{n_2}$, da Var($bX$)=$b^2$Var($X$), og ifølge sætning 6.3.12 er $\bar{X}_1 - \bar{X}_2$ så normalfordelt med middelværdi $\mu_1 - \mu_2 = 0$ (da det antages at $\mu_1 = \mu_2$) og varians $\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)$.
Ved at sætte $k = 2$, $n = 4$, $\bar{X}_1 = X_1, \bar{X}_2 = X_2, s_1^2 = X_3$ og $s_2^2 = X_4$ og ved at lade $\varphi$ være funktionen fra $\mathbb{R}^2$ ind i $\mathbb{R}$ der sender $(x, y)$ ind i $x - y$, og lade $\psi$ være funktionen fra $\mathbb{R}^2$ ind i $\mathbb{R}$ der sender $(x, y)$ ind i $\frac{1}{n-2}((n_1 - 1)x^2 + (n_2 - 1)y^2)$ ses det ved brug af sætning 6.2.3 3) [3) står først nævnt efter beviset for sætningen] at $\bar{X}_1 - \bar{X}_2$ og $s^2$ er uafhængige.

## (f)

Da $\bar{X}_1 - \bar{X}_2$ er normalfordelt med middelværdi 0 og varians $\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}$, kan den skrives som $|\sigma|\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}X$, hvor $X$ er en standard normalfordeling, og da $Y = \frac{n-2}{\sigma^2}s^2 \sim \chi^2(n-2)$, gælder der (idet vi antager $s > 0$), at

$$T_2 = \frac{\bar{X}_1 - \bar{X}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{|\sigma|\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}X}{\sqrt{\frac{\sigma^2}{n-2}\chi^2(n-2)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{X}{\sqrt{Y/(n-2)}}$$

så $T_2$ er $t$-fordelt med $n - 2$ frihedsgrader, da $\bar{X}_1 - \bar{X}_2$ og $s^2$ er uafhængige (ifølge (e)), og $X$ og $Y$ dermed er uafhængige.

## (g)

$T_2$ kan anvendes for at sammenligne to sæt observationer, som man antager kommer fra hver sin normalfordeling ($X_1$ og $X_2$) med samme varians. Vi har inden opgave (e) antaget, at de også har samme middelværdi ($\mu_1 = \mu_2$). $T_2$ kan så bruges som teststørrelse til at afgøre, om antagelsen om normalfordelingernes samme middelværdi kan verificeres.