Matematisk Modellering 1
Hjælpeark

Kaare B. Mikkelsen
20050910

13. september 2007

Indhold

1 Formler

$$SP_{xt} = \sum_{i=1}^{n} x_i t_i$$

(s. 122)

$$SSD = USS - \frac{S^2}{n}$$

$$s^2 = \frac{1}{n-1}(USS - \frac{S^2}{n})$$

(begge s. 64)

$$SSD_2 = \sum_{i=1}^{k} n_i(\bar{x}_i. - \bar{x}..)^2 = \left(\sum_{i=1}^{k} \frac{S_i^2}{n_i}\right) - \frac{S.^2}{n.}$$

(s.103, eller 116)

$$s_1^2 = \frac{\sum_{i=1}^{k} f_{(i)} s_{(i)}^2}{f_1} = \frac{\sum_{i=1}^{k} SSD_{(i)}}{\sum_{i=1}^{k} f_{(i)}}$$

(s. 115)

2 Analyse af k normalfordelte prøver [1]

2.1 Modelcheck

Hvis man vil undersøge om et datasæt er normalfordelt, plottes det i et fraktildiagram (se s. 31 for skabelon).
Hvis dataene ligger på omtrent en ret linje, bliver modellen:

$$M_0 : x_{ij} \sim\sim N(\mu_i, \sigma_i^2), j = 1, ..., n_i, i = 1, ..., k$$

$$\mu_i \leftarrow \bar{x}_i. \sim\sim N(\mu_i, \frac{\sigma_i^2}{n_i})$$

$$\sigma_i^2 \leftarrow s_{(i)}^2 \sim\sim \sigma_i^2 \chi^2(f_{(i)})/f_{(i)}$$

[1] Som afskrevet s. 101-103

2.2 Test af $H_{0\sigma^2} : \sigma_1 = \sigma_2 = ... = \sigma_k$

Teststørrelse:

$$Ba = \frac{-2\ln Q(\mathbf{x})}{C} \overset{*}{\sim\approx} \chi^2(k-1)$$

hvor

$$-2\ln Q(\mathbf{x}) = f_1 \ln s_1^2 - \sum_{i=1}^{k} f_{(i)} \ln s_{(i)}^2 \qquad C = 1 + \frac{1}{3(k-1)}\left[\left(\sum_{i=1}^{k} \frac{1}{f_{(i)}}\right) - \frac{1}{f_1}\right]$$

$$\text{hvor } f_1 = \sum_i f_{(i)}$$

* gælder (approksimativt) hvis $f_{(i)} \geq 2$

Testsandsynlighed:

$$p_{obs}(\mathbf{x}) = 1 - F_{\chi^2(k-1)}(Ba)$$

Hvis en udregning i hånden giver $-2\ln Q(\mathbf{x}) < 0$, har man begået en fehler!
Hvis $H_{0\sigma^2}$ holder vand, står man med:

$$M_1 : x_{ij} \sim\sim N(\mu_i, \sigma^2), j = 1, ..., n_i, i = 1, ..., k$$

$$\mu_i \leftarrow \bar{x}_i. \sim\sim N(\mu_i, \frac{\sigma^2}{n_i})$$

$$\sigma^2 \leftarrow s_1^2 = \frac{SSD_1}{f_1} = \frac{\sum_{i=1}^{k} SSD_{(i)}}{f_1} \sim\sim \sigma^2\chi^2(f_1)/f_1$$

2.3 Test af $H_{0\mu} : \mu_1 = \mu_2 = ... = \mu_k$

Teststørrelse:

$$F(\mathbf{x}) = \frac{s_2^2}{s_1^2} \sim\sim F(k-1, n. - k)$$

$$\text{hvor } s_2^2 = \frac{SSD_2}{k-1}$$

Testsandsynlighed:

$$p_{obs}(\mathbf{x}) = 1 - F_{F(k-1,n.-k)}(F(\mathbf{x}))$$

bemark at hvis hypotesen godkendes, ændres middelværdistrukturen, og dermed varians-estimatet.
Det korte og det lange:

$$M_2 : x_i \sim\sim N(\mu, \sigma^2), i = 1, \ldots, n$$

$$\mu \leftarrow \bar{x}. = \frac{S}{n} \sim\sim N(\mu, \frac{\sigma^2}{n})$$

$$\sigma^2 \leftarrow s^2 \sim\sim \sigma^2\chi^2(f)/f$$

2.4 Konfidensintervaller [2]

Generelt:
Man tager variansen af estimatet, udskifter σ^2 med dets estimat, tager en kvadratrod, og ganger med $t_{1-\alpha/2}(f)$. Også kendt som

$$\left[\text{est. } \pm \text{Std.error} \times t_{1-\alpha/2}(f)\right]$$

Ellers:

For σ^2:

$$\left[\frac{f_1 s_1^2}{\chi^2_{1-\alpha/2}(f_1)}, \frac{f_1 s_1^2}{\chi^2_{\alpha/2}(f_1)}\right]$$

hvor $\chi^2_{1-\alpha/2}(f_1)$ og $\chi^2_{\alpha/2}(f_1)$ betegner fraktiler.

For μ_i:

$$\left[\bar{x}_{i\cdot} - \sqrt{\frac{s_1^2}{n_i}} t_{1-\alpha/2}(f_1), \bar{x}_{i\cdot} + \sqrt{\frac{s_1^2}{n_i}} t_{1-\alpha/2}(f_1)\right]$$

hvor $t_{1-\alpha/2}(f_1)$ betegner en fraktil.

2.5 k = 2 [3]

Test af $H_{0\sigma^2} : \sigma_1^2 = \sigma_2^2$

Lad $s^2_{numerator} = \max\{s^2_{(1)}, s^2_{(2)}\}$, og lad $f_{numerator}$ være de tilhørende frihedsgrader, og lad $s^2_{denominator} = \min\{s^2_{(1)}, s^2_{(2)}\}$, med tilhørende frihedsgrader $f_{denominator}$.

Testørrelse:

$$F(\mathbf{x}) = \frac{s^2_{numerator}}{s^2_{denominator}} \sim\sim F(f_{numerator}, f_{denominator})$$

Testsandsynlighed:

$$p_{obs}(\mathbf{x}) = 2\left[1 - F_{F(f_{numerator}, f_{denominator})}(F(\mathbf{x}))\right]$$

Hvor $F(f_{numerator}, f_{denominator})$ er F-fordelingen med $f_{numerator}$ frihedsgrader i tælleren, og $f_{denominator}$ frihedsgrader i nævneren.

[2] s. 104
[3] s. 94-95

Skulle hypotesen blive accepteret, er modellen så:

$$x_{ij} \sim\sim N(\mu_i, \sigma^2), i = 1, 2, j = 1, \ldots, n_i$$
$$\mu_i \leftarrow \bar{x}. \sim\sim N(\mu_i, \frac{\sigma^2}{n_i})$$
$$\sigma^2 \leftarrow s_1^2 = \frac{f_{(1)} s_{(1)}^2 + f_{(2)} s_{(2)}^2}{f_{(1)} + f_{(2)}} = \frac{SSD_{(1)} + SSD_{(2)}}{f_{(1)} + f_{(2)}} \sim\sim \sigma^2 \chi^2(f_1)/f_1$$

$$\text{hvor } f_1 = f_{(1)} + f_{(2)} = n. - 2$$

Test af $H_0 : \mu_1 = \mu_2$, for samme varians
Teststørrelse:[4]

$$t(\mathbf{x}) = \frac{\bar{x}_1. - \bar{x}_2.}{\sqrt{s_1^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim\sim t(f_1)$$

Testsandsynlighed[5]:

$$p_{obs}(\mathbf{x}) = 2 \left[1 - F_{t(f_1)}(|t(\mathbf{x})|) \right]$$
$$\text{hvor } f_1 = f_{(1)} + f_{(2)} = n_1 + n_2 - 2$$

Test af $H_0 : \mu_1 = \mu_2$, for forskellig varians
Teststørrelse:

$$t(\mathbf{x}) = \frac{\bar{x}_1. - \bar{x}_2.}{\sqrt{s_{(1)}^2/n_1 + s_{(2)}^2/n_2}} \sim\sim t(\bar{f})$$

hvor $t(\bar{f})$ er t-fordelingen med $\bar{f}$ frihedsgrader. $\bar{f}$ kan regnes som

$$\bar{f} = \frac{\frac{s_{(1)}^2}{n_1} + \frac{s_{(2)}^2}{n_1}}{\frac{\left(\frac{s_{(1)}^2}{n_1}\right)^2}{f_{(1)}} + \frac{\left(\frac{s_{(2)}^2}{n_2}\right)^2}{f_{(2)}}}$$

Testsandsynlighed:

$$p_{obs}(\mathbf{x}) = 2 \left[1 - F_{t(\bar{f})}(|t(\mathbf{x})|) \right]$$

2.6 Test af $\mu = \mu_0$ og $\sigma^2 = \sigma_0^2$, for k = 1 [6]

Test af $\mu = \mu_0, \sigma^2$ kendt
Teststørrelse:

$$u(\mathbf{x}) = \frac{\bar{x}. - \mu_0}{\sqrt{\sigma_0^2/n}} \sim\sim N(0, 1)$$

[4] s_1^2 fundet s. 94
[5] f_1 fundet s. 95
[6] s. 78-79

Testsandsynlighed:

$$p_{obs}(\mathbf{x}) = 2\left[1 - \Phi(|u(\mathbf{x})|)\right]$$

Test af $\mu = \mu_0, \sigma^2$ *ukendt*

Teststørrelse:

$$u(\mathbf{x}) = \frac{\bar{x}. - \mu_0}{\sqrt{s^2/n}} \sim\sim t(f)$$

Testsandsynlighed:

$$p_{obs}(\mathbf{x}) = 2\left[1 - F_{t(f)}(|t(\mathbf{x})|)\right]$$

Test af $\sigma^2 = \sigma_0^2$

Teststørrelse:

$$\frac{fs}{\sigma_0^2} \sim\sim \chi^2(f)$$

Testsandsynlighed:

$$\begin{cases} 2F_{\chi^2(f)}(\frac{fs}{\sigma_0^2}) & \text{hvis } \frac{fs}{\sigma_0^2} \leq f \\ 2(1 - F_{\chi^2(f)}(\frac{fs}{\sigma_0^2})) & \text{hvis } \frac{fs}{\sigma_0^2} \geq f \end{cases}$$

3 Multinomialfordelingen

3.1 Proposition 7.1 [7]

For $x_j > 0, j = 1, \ldots, k$ og $\sum^k x_j = n$:

$$\text{Funktionen } g : \Pi^{(k)} \to \mathbb{R}$$
$$\boldsymbol{\pi} \to \pi_1^{x_1} \ldots \pi_j^{x_j} \ldots \pi_k^{x_k}$$
$$\text{har maximum for}$$
$$\hat{\boldsymbol{\pi}} = (\frac{x_1}{n}, \ldots, \frac{x_k}{n}, \ldots, \frac{x_k}{n})$$

Model:

$$M_0 : \mathbf{X} \sim m(n, \boldsymbol{\pi}), \boldsymbol{\pi} \in \boldsymbol{\Pi}^{(k)}$$

det skal bemærkes at

$$\boldsymbol{\Pi}^{(k)} = \{\boldsymbol{\pi} \in \mathbb{R}^k \mid \pi_j > 0, j = 1, \ldots, k, \sum_{j=1}^{k} \pi_j = 1\}$$

3.2 Test af Hypotese $\boldsymbol{\pi} = \boldsymbol{\pi}_0$ (simpel hypotese) [8]

Teststørrelse:

$$-2\ln Q(\mathbf{x}) = 2\sum_{j=1}^{k} x_j \ln(\frac{x_j}{e_j}) \sim\sim \chi^2(k - 1 - d)$$

i det simple tilfælde er $\mathbf{e} = n\boldsymbol{\pi_0}$

[7] s. 305

[8] s. 307-308

Testsandsynlighed:

$$p_{obs}(\mathbf{x}) = 1 - F_{\chi^2(k-1-d)}(-2\ln Q(\mathbf{x}))$$

d er i denne sammenhæng antal frie parametre under hypotesen.

3.3 Generelt tilfælde

I det generelle tilfælde findes $\mathbf{e}$ ved at indsætte $\boldsymbol{\pi(\theta)}$ i

$$L(\boldsymbol{\theta}) = \frac{n!}{x_1! \dots x_j! \dots x_k!} \pi_1(\boldsymbol{\theta})^{x_1} \dots \pi_j(\boldsymbol{\theta})^{x_j} \dots \pi_k(\boldsymbol{\theta})^{x_k}$$

og få dette på en form så man kan bruge Prop. 7.1 på den.

Udkommet af dette, $\hat{\boldsymbol{\theta}}$, kan så puttes ind i

$$\mathbf{e} = (e_1, \dots, e_j, \dots, e_k)^* = n(\pi_1(\hat{\boldsymbol{\theta}}), \dots, \pi_j(\hat{\boldsymbol{\theta}}), \dots, \pi_k(\hat{\boldsymbol{\theta}}))^*$$

Testsandsynlighed:

$$p_{obs} = 1 - F_{\chi^2(k-1-d)}(-2\ln Q(\mathbf{x}))$$

d er i denne sammenhæng antal frie parametre under hypotesen.

4 Lineær Regression

4.1 Test af lineær regression

I det generelle tilfælde er den eneste måde at plotte dataene i en graf, og se hvorvidt placere sig ligger lineært.
Skulle det være tilfældet at der er flere målinger (over 10) for hver værdi af den forklarende variabel (t), således at serien kan deles op i k målinger, hvor alle $n_i(> 10)$ målinger i den i'te gruppe har samme t-værdi, kan man gøre følgende[9]:
Udgangspunktet er

$$M_0 : X_{ij} \sim N(\mu_i, \sigma_i^2)$$

Denne model kan f.eks. tjekkes vha. k fraktilplot's.
Holder den vand, testes:

$$M_1 : X_{ij} \sim N(\mu_i, \sigma^2)$$

vha. Bartlett (se ovenfor).
Næste skridt er

$$M_2 : X_{ij} \sim N(\alpha + \beta t_i, \sigma^2)$$

Reduktionen fra M_1 til M_2 kan testes vha. *til-og-fra-formlen*:

$$F(\mathbf{x}) = \frac{\frac{SSD_{02}-SSD_1}{f_{02}-f_1}}{s_1^2} = \frac{s_2^2}{s_1^2} \sim F(k-2, n-k)$$

[9] s.133-134

hvor $f_1 = n - k$
Teststørrelse:

$$p_{obs}(\mathbf{x}) = 1 - F_{F(k-2,n-k)}(F(\mathbf{x}))$$

Går det godt, står man altså med:

$$\begin{gathered}
M_2 : X_{ij} \sim N(\alpha + \beta t_i, \sigma^2) \\
\beta \leftarrow \hat{\beta} = \frac{SPD_{xt}}{SSD_t} \sim\sim N\left(\beta, \frac{\sigma^2}{SSD_t}\right) \\
\alpha \leftarrow \hat{\alpha} = \frac{1}{n}\left[S_x - \hat{\beta}S_t\right] \sim\sim N(\alpha, \left(\frac{1}{n} + \frac{\bar{t}^2.}{SSD_t}\right) \\
\sigma^2 \leftarrow s_{02}^2 = \frac{1}{n-2}SSD_{02} \sim\sim \sigma^2\chi^2(f_{02})/f_{02} \\
\text{hvor } f_{02} = n - 2
\end{gathered}$$

4.2 Modeloversigt [10]

$$\begin{array}{ccccc}
 & & M_3 : X_i \sim N(\alpha + \beta_0 t_i, \sigma^2) & & \\
 & H_{03} : \beta = \beta_0 \nearrow & & \searrow H_{04} : \alpha = \alpha_0 & \\
M_2 : X_i \sim N(\alpha + \beta t_i, \sigma^2) & & & & M_4 : X_i \sim N(\alpha_0 + \beta_0 t_i, \sigma^2) \\
 & H_{03}^* : \alpha = \alpha_0 \searrow & & \nearrow H_{04} : \beta = \beta_0 & \\
 & & M_3^* : X_i \sim N(\alpha_0 + \beta t_i, \sigma^2) & &
\end{array}$$

4.3 $M_2 \rightarrow M_3$

Test af $H_{03} : \beta = \beta_0$:
Teststørrelse:

$$t(\mathbf{x}) = \frac{\hat{\beta} - \beta_0}{\sqrt{s_{02}^2/SSD_t}} \sim\sim t(n-2)$$

Testsandsynlighed:

$$p_{obs}(\mathbf{x}) = 2\left[1 - F_{t(n-2)}(|t(\mathbf{x})|)\right]$$

Holder den vand, er modellen som beskrevet i modeloversigten, og estimaterne er:

$$\begin{gathered}
\alpha \leftarrow \hat{\alpha}_{M_3} = \bar{x}. - \beta_0\bar{t}. \sim\sim N(\alpha, \frac{\sigma^2}{n}) \\
\sigma^2 \leftarrow s_{03}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\{x_i - (\hat{\alpha}_{M_3} + \beta_0 t_i)\}^2 = \\
\frac{1}{n-1}\left[SSD_{02} + (\hat{\beta} - \beta_0)^2 SSD_t\right] \sim\sim \sigma^2\chi^2(n-1)/(n-1)
\end{gathered}$$

[10] s. 156, alt følgende er taget fra siderne 156-158

1-α-konfidensinterval for α:

$$\left[\hat{\alpha} \pm \sqrt{s_{03}^2(\frac{1}{n})}t_{1-\alpha/2}(f_{03})\right]$$

4.4 $M_2 \rightarrow M_3^*$

Test af $H_{03}^* : \alpha = \alpha_0$
Teststørrelse:

$$t(\mathbf{x}) = \frac{\hat{\alpha} - \alpha_0}{\sqrt{s_{02}^2\left(\frac{1}{n} + \frac{\bar{t}_\cdot^2}{SSD_t}\right)}} \sim\sim t(n-2)$$

Testsandsynlighed:

$$p_{obs}(\mathbf{x}) = 2\left[1 - F_{t(n-2)}(|t(\mathbf{x})|)\right]$$

Holder den vand, er modellen som beskrevet i modeloversigten, og estimaterne er:

$$\beta \leftarrow \hat{\beta}_{M_3^*} = \frac{\sum_{i=1}^n t_i(x_i - \alpha_0)}{\sum_{i=1}^n t_i^2} = \frac{Sp_{xt} - \alpha_0 S_t}{USS_t} \sim\sim N(\beta, \frac{\sigma^2}{USS_t})$$

$$\sigma^2 \leftarrow s_{03}^{*2} = \frac{1}{n-1}\sum_{i=1}^n \{x_i - (\alpha_0 + \hat{\beta}_{M_3^*}^2 t_i)\}^2 =$$

$$\frac{1}{n-1}\left[USS_x + n\alpha_0^2 - 2\alpha_0 S_x - \hat{\beta}_{M_3^*}^2 USS_t\right] \sim\sim \sigma^2\chi^2(n-1)/(n-1)$$

1-α-konfidensinterval for β:

$$\left[\hat{\beta} \pm \sqrt{\frac{s_{03}^{*2}}{USS_t}}t_{1-\alpha/2}(f_{03}^*)\right]$$

4.5 $M_3^* \rightarrow M_4$

Test af $H_{04}^* : \beta = \beta_0$
Teststørrelse:

$$t(\mathbf{x}) = \frac{\hat{\beta}_{M_3^*} - \beta_0}{\sqrt{s_{03}^{*2}/USS_t}} = \frac{SP_{xt} - \alpha_0 S_t - \beta_0 USS_t}{\sqrt{s_{03}^{*2}/USS_t}} \sim\sim t(n-1)$$

Testsandsynlighed:

$$p_{obs}(\mathbf{x}) = 2\left[1 - F_{t(n-1)}(|t(\mathbf{x})|)\right]$$

Holder den vand, er modellen som beskrevet i modeloversigten, og estimaterne er:

$$\sigma^2 \leftarrow s_{04}^2 = \frac{1}{n}\sum_{i=1}^n \{x_i - (\alpha_0 + \beta_0 t_i)\}^2 =$$

$$\frac{1}{n}\left[USS_x + n\alpha_0^2 + \beta_0^2 USS_t - 2\alpha_0 S_x - 2\beta_0 SP_{xt} + 2\alpha_0\beta_0 S_t\right] \sim\sim \sigma^2\chi^2(n)/n$$

4.6 $M_3 \to M_4$

Test af $H_{04} : \alpha = \alpha_0$
Teststørrelse:

$$t(\mathbf{x}) = \frac{\hat{\alpha}_{M_3} - \alpha_0}{\sqrt{s_{03}^2 n}} \sim\sim t(n-1)$$

Testsandsynlighed:

$$p_{obs}(\mathbf{x}) = 2\left[1 - F_{t(n-1)}(|t((x))|)\right]$$

Holder den vand, er modellen som beskrevet i modeloversigten, og estimaterne som beskrevet ovenfor.

5 Lineære normalfordelinger - generelt

5.1 Til-og-fra-formel

Omdrejningspunktet er *til-og-fra-formlen*[11]. Her regnende på overgangen $M_l \to M_m$

$$F(\mathbf{x}) = \frac{\frac{\|\mathbf{x}-P_l(\mathrm{x})\|^2 - \|\mathbf{x}-P_m(\mathrm{x})\|^2}{(n-d_l)-(n-d_m)}}{s_{01}^2} \sim\sim F(d_m - d_l, n - d_1)$$

Med testsandsynlighed

$$p_{obs}(\mathbf{x}) = 1 - F_{F(d_m - d_l, n-d_1)}(F(\mathbf{x}))$$

Det er vist s. 187 hvor man finder størrelserne i et GLM-print.
Det skal iøvrigt bemærkes, at skal man regne på en overgang fra gruppe-afhængig spredning, til ens spredning, så gøres det nemt og smertefrit vha. standardmetoden, se afsnit 2.2 ovenfor.

[11] s. 187

6 Beregningsskemaer

Taget s. 159:
Udregningsskema for model M_q

	x		t
n		n	
S	S_x		S_t
SP		SP_{xt}	
SSD	$USS_x - \frac{S_x^2}{n}$		$USS_t - \frac{S_t^2}{n}$
SPD		$SP_{xt} - \frac{S_x S_t}{n}$	
$\hat{\boldsymbol{\beta}}$		$\frac{SPD_{xt}}{SSD_t}$	
$\hat{\boldsymbol{\alpha}}$		$\frac{1}{n}\left[S_x - \hat{\beta} S_t\right]$	
$\boldsymbol{SSD_{0q}}$		$SSD_x - \frac{SPD_{xt}^2}{SSD_t}$	
$\boldsymbol{s_{0q}^2}$		$\frac{1}{n-2} SSD_{0q}$	

Taget s. 159:
Udregningsskema for model M_q

	x	t
n	n	
S	S_x	S_t
SP	SP_{xt}	
SSD	$USS_x - \frac{S_x^2}{n}$	$USS_t - \frac{S_t^2}{n}$
SPD	$SP_{xt} - \frac{S_x S_t}{n}$	
$\hat{\boldsymbol{\beta}}$	$\frac{SPD_{xt}}{SSD_t}$	
$\hat{\boldsymbol{\alpha}}$	$\frac{1}{n}\left[S_x - \hat{\beta} S_t\right]$	
$\boldsymbol{SSD_{0q}}$	$SSD_x - \frac{SPD_{xt}^2}{SSD_t}$	
$\boldsymbol{s_{0q}^2}$	$\frac{1}{n-2} SSD_{0q}$	