

FORMELSAMLING TIL STK2120

(Versjon av 30. mai 2012)

1 Enveis variansanalyse

Anta at $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$; $j = 1, 2, \dots, J_i$; $i = 1, 2, \dots, I$; der ϵ_{ij} -ene er uavhengige og $N(0, \sigma^2)$ -fordelte. Da har vi at:

(a) Den totale kvadratsummen $SST = \sum_{i=1}^I \sum_{j=1}^{J_i} (Y_{ij} - \bar{Y}_{..})^2$ kan skrives som $SST = SSE + SSTr$ der

$SSE = \sum_{i=1}^I \sum_{j=1}^{J_i} (Y_{ij} - \bar{Y}_{i.})^2$ er kvadratsummen for feil eller kvadratsummen innen ("within") grupper

$SSTr = \sum_{i=1}^I J_i (\bar{Y}_{i.} - \bar{Y}_{..})^2$ er kvadratsummen for behandling eller kvadratsummen mellom ("between") grupper

(b) SSE og $SSTr$ er uavhengige

(c) $MSE = SSE / [\sum_{i=1}^I (J_i - 1)]$ er en forventningsrett estimator for σ^2 .
 SSE / σ^2 er kji -kvadratfordelt med $\sum_{i=1}^I (J_i - 1)$ frihetsgrader.

(d) Hvis alle α_i -ene er lik null, er $SSTr / \sigma^2$ kji -kvadratfordelt med $I - 1$ frihetsgrader

(e) Hvis $J_i = J$ for $i = 1, \dots, I$, så er
 $\max_{i_1, i_2} |(\bar{Y}_{i_1.} - \mu_{i_1}) - (\bar{Y}_{i_2.} - \mu_{i_2})| / \sqrt{MSE / J}$
fordelt som den studentifiserte variasjonsbredde med parametere I og $I(J - 1)$.

2 Toveis variansanalyse

Anta at $Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$; $k = 1, \dots, K$; $j = 1, \dots, J$; $i = 1, \dots, I$; der ϵ_{ijk} -ene er uavhengige og $N(0, \sigma^2)$ -fordelte. Da har vi at:

(a) Den totale kvadratsummen $SST = \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K (Y_{ijk} - \bar{Y}_{...})^2$ kan skrives som $SST = SSA + SSB + SSAB + SSE$ der

$$SSA = JK \sum_{i=1}^I (\bar{Y}_{i..} - \bar{Y}_{...})^2$$

$$SSB = IK \sum_{j=1}^J (\bar{Y}_{.j.} - \bar{Y}_{...})^2$$

$$SSAB = K \sum_{i=1}^I \sum_{j=1}^J (\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...})^2$$

$$SSE = \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K (Y_{ijk} - \bar{Y}_{ij.})^2$$

(b) SSA , SSB , $SSAB$ og SSE er uavhengige

(c) $MSE = SSE / IJ(K - 1)$ er en forventningsrett estimator for σ^2 .
 SSE / σ^2 er kji -kvadratfordelt med $IJ(K - 1)$ frihetsgrader.

- (d) Hvis alle α_i -ene er lik null, er SSA/σ^2 kji-kvadratfordelt med $I - 1$ frihetsgrader
- (e) Hvis alle β_j -ene er lik null, er SSB/σ^2 kji-kvadratfordelt med $J - 1$ frihetsgrader
- (f) Hvis alle γ_{ij} -ene er lik null, er $SSAB/\sigma^2$ kji-kvadratfordelt med $(I - 1)(J - 1)$ frihetsgrader

3 Blokkforsøk (toveis variansanalyse uten gjentak)

Anta at $Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$; $j = 1, \dots, J$; $i = 1, \dots, I$; der ϵ_{ij} -ene er uavhengige og $N(0, \sigma^2)$ -fordelte. Da har vi at:

- (a) Den totale kvadratsummen $SST = \sum_{i=1}^I \sum_{j=1}^J (Y_{ij} - \bar{Y}_{..})^2$ kan skrives som $SST = SSA + SSB + SSE$ der

$$SSA = J \sum_{i=1}^I (\bar{Y}_{i.} - \bar{Y}_{..})^2$$

$$SSB = I \sum_{j=1}^J (\bar{Y}_{.j} - \bar{Y}_{..})^2$$

$$SSE = \sum_{i=1}^I \sum_{j=1}^J (Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..})^2$$

- (b) SSA , SSB og SSE er uavhengige
- (c) $MSE = SSE/[(I - 1)(J - 1)]$ er en forventningsrett estimator for σ^2 .
 SSE/σ^2 er kji-kvadratfordelt med $(I - 1)(J - 1)$ frihetsgrader.
- (d) Hvis alle α_i -ene er lik null, er SSA/σ^2 kji-kvadratfordelt med $I - 1$ frihetsgrader
- (e) Hvis alle β_j -ene er lik null, er SSB/σ^2 kji-kvadratfordelt med $J - 1$ frihetsgrader

4 Tabelldata og kji-kvadrattester

- (a) Anta at (N_1, \dots, N_k) er multinomisk fordelt med sannsynligheter p_i , der $\sum_{i=1}^k N_i = n$ og $\sum_{i=1}^k p_i = 1$.
Hvis $p_i = \pi_i(\boldsymbol{\theta})$, der $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m)$, og $\hat{\boldsymbol{\theta}}$ er maksimum likelihood estimatoren for $\boldsymbol{\theta}$, så er

$$\chi^2 = \sum_{i=1}^k \frac{(N_i - E_i)^2}{E_i}$$

tilnærmet kji-kvadratfordelt med $k - 1 - m$ frihetsgrader når $E_i = n\pi_i(\hat{\boldsymbol{\theta}}) \geq 5$ for (nesten) alle i

- (b) Homogenitetstesting: Anta at for $i = 1, \dots, I$ er (N_{i1}, \dots, N_{iJ}) uavhengige og multinomisk fordelt med sannsynligheter p_{ij} , der $\sum_{j=1}^J p_{ij} = 1$.
Hvis $p_{1j} = \dots = p_{Ij}$, så er

$$\chi^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{(N_{ij} - E_{ij})^2}{E_{ij}}$$

tilnærmet kji-kvadratfordelt med $(I - 1)(J - 1)$ frihetsgrader når $E_{ij} = (N_{i.}N_{.j})/N_{..} \geq 5$ for (nesten) alle i, j

- (c) Uavhengighetstesting: Anta at $(N_{11}, \dots, N_{1J}, N_{21}, \dots, N_{2J}, \dots, N_{I1}, \dots, N_{IJ})$ er multinomisk fordelt med sannsynligheter p_{ij} , der $\sum_{i=1}^I \sum_{j=1}^J p_{ij} = 1$. Hvis $p_{ij} = p_{i.}p_{.j}$ for alle i, j , så er

$$\chi^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{(N_{ij} - E_{ij})^2}{E_{ij}}$$

tilnærmet kji-kvadratfordelt med $(I - 1)(J - 1)$ frihetsgrader når $E_{ij} = (N_{i.}N_{.j})/N_{..} \geq 5$ for (nesten) alle i, j

5 Multippel lineær regresjon

Anta at $Y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{i,k} + \epsilon_i$; $i = 1, 2, \dots, n$; der x_{ij} -ene er kjente tall og ϵ_i -ene er uavhengige og $N(0, \sigma^2)$ -fordelte.

På matriseform kan vi skrive modellen som $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$, der $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ og $\boldsymbol{\beta} = (\beta_0, \dots, \beta_k)^T$ er henholdsvis n - og $k + 1$ -dimensjonale vektorer, og $\mathbf{X} = \{x_{ij}\}$ er en $n \times (k + 1)$ -dimensjonal matrise. Vi har at:

- (a) Minste kvadraters estimator for $\boldsymbol{\beta}$ er $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$.

- (b) La $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \dots, \hat{\beta}_k)^T$. Da er $\hat{\beta}_j$ -ene normalfordelte og forventningsrette, og

$$\text{Var}(\hat{\beta}_i) = \sigma^2 c_{ii} \quad \text{og} \quad \text{Cov}(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2 c_{ij}$$

der c_{ij} er element (i, j) i $(k + 1) \times (k + 1)$ matrisen $\mathbf{C} = (\mathbf{X}^T \mathbf{X})^{-1}$.

- (c) La $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_k x_{i,k}$, og sett $SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$. Da er $S^2 = SSE / (n - (k + 1))$ en forventningsrett estimator for σ^2 , og $(n - (k + 1))S^2 / \sigma^2 \sim \chi_{n-(k+1)}^2$. Videre er S^2 og $\hat{\boldsymbol{\beta}}$ uavhengige.

- (d) La $S_{\hat{\beta}_i}^2$ være den variansestimatorene for $\hat{\beta}_i$ vi får ved å erstatte σ^2 med S^2 i formelen for $\text{Var}(\hat{\beta}_i)$ i punkt b). Da er $(\hat{\beta}_i - \beta_i) / S_{\hat{\beta}_i} \sim t_{n-(k+1)}$.

6 Bootstrapping

Anta fordelingen til data \mathbf{X} er beskrevet ved en fordelingsfunksjon F . La $\theta = \theta(F)$ være en egenskap ved F som estimeres ved $\hat{\theta} = \hat{\theta}(\mathbf{X})$.

- (a) Bootstrapping-idéen er å tilnærme egenskapene til $\hat{\theta}$ ved å anta at et estimat \hat{F} på F er den sanne fordelingsfunksjonen.

(b) Bootstrap estimering av skjevhet til $\hat{\theta}$:

$$b_{\hat{\theta}} = \frac{1}{B} \sum_{b=1}^B \theta_b^* - \theta(\hat{F})$$

(c) Bootstrap estimering av standardavvik til $\hat{\theta}$:

$$\sqrt{\mathbf{E}^{\hat{F}}[(\hat{\theta}(\mathbf{X}^*) - \mathbf{E}^{\hat{F}}[\hat{\theta}(\mathbf{X}^*)])^2]}$$

(d) Standard bootstrap konfidensintervall:

$$(\hat{\theta} - \bar{\delta}, \hat{\theta} - \underline{\delta})$$

der $\underline{\delta}$ og $\bar{\delta}$ er nedre og øvre $\alpha/2$ kvantil i bootstrap fordelingen til $\Delta = \hat{\theta} - \theta$.

7 Maksimum likelihood metoden

Anta at X_1, X_2, \dots, X_n har simultan punktsannsynlighet/sannsynlighetstetthet $f(x_1, x_2, \dots, x_n | \boldsymbol{\theta})$, der $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)$ er en parametervektor (skalar hvis $p = 1$). Vi antar at $f(x_1, x_2, \dots, x_n | \boldsymbol{\theta})$ tilfredsstiller visse deriverbarhetsbetingelser.

- (a) Gitt observerte verdier $X_i = x_i$; $i = 1, \dots, n$; er likelihood-funksjonen $\text{lik}(\boldsymbol{\theta}) = f(x_1, x_2, \dots, x_n | \boldsymbol{\theta})$ og loglikelihood-funksjonen $l(\boldsymbol{\theta}) = \log(\text{lik}(\boldsymbol{\theta}))$.
- (b) Maksimum likelihood *estimatet* er den verdien av $\boldsymbol{\theta}$ som maksimerer $\text{lik}(\boldsymbol{\theta})$ eller ekvivalent maksimerer $l(\boldsymbol{\theta})$. Hvis vi erstatter de observerte x_i -ene med de stokastiske X_i -ene, får vi maksimum likelihood *estimatoren*.
- (c) Maksimum likelihood estimatet $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \dots, \hat{\theta}_p)$ er en løsning av ligningene $s_j(\boldsymbol{\theta}) = 0$; $j = 1, \dots, p$; der $s_j(\boldsymbol{\theta}) = (\partial/\partial\theta_j)l(\boldsymbol{\theta})$ er score-funksjonene. Vektoren av scorefunksjoner er $\mathbf{s}(\boldsymbol{\theta}) = (s_1(\boldsymbol{\theta}), \dots, s_p(\boldsymbol{\theta}))^T$.
- (d) Den observerte informasjonsmatrisen $\bar{\mathbf{J}}(\boldsymbol{\theta})$ er $p \times p$ matrisen med element (i, j) gitt ved $\bar{J}_{ij}(\boldsymbol{\theta}) = -\frac{\partial^2}{\partial\theta_i\partial\theta_j}l(\boldsymbol{\theta})$. Den forventede informasjonsmatrisen (eller Fishers informasjonsmatrise) $\bar{\mathbf{I}}(\boldsymbol{\theta})$ er $p \times p$ matrisen med element (i, j) gitt ved $\bar{I}_{ij}(\boldsymbol{\theta}) = \mathbf{E}[\bar{J}_{ij}(\boldsymbol{\theta})]$. For uavhengige og identisk fordelte observasjoner har vi at $\bar{\mathbf{I}}(\boldsymbol{\theta}) = n\mathbf{I}(\boldsymbol{\theta})$ der $\mathbf{I}(\boldsymbol{\theta})$ er forventet informasjon til en observasjon.
- (e) Når ligningene i punkt (c) ikke har en eksplisitt løsning, kan vi finne maksimum likelihood estimatet ved å bruke Newton-Raphsons metode:

$$\boldsymbol{\theta}^{(s+1)} = \boldsymbol{\theta}^{(s)} + \bar{\mathbf{J}}^{-1}(\boldsymbol{\theta}^{(s)})\mathbf{s}(\boldsymbol{\theta}^{(s)}),$$

ved å bruke Fishers scoringsalgoritme:

$$\boldsymbol{\theta}^{(s+1)} = \boldsymbol{\theta}^{(s)} + \bar{\mathbf{I}}^{-1}(\boldsymbol{\theta}^{(s)})\mathbf{s}(\boldsymbol{\theta}^{(s)}),$$

eller ved passende modifikasjoner av disse.

(f) Når vi har “tilstrekkelig mye” data, er $\hat{\theta}_i$ tilnærmet normalfordelt med forventning θ_i og med varians lik det i -te diagonalelementet til $\bar{\mathbf{I}}^{-1}(\boldsymbol{\theta})$. Kovariansen mellom $\hat{\theta}_i$ og $\hat{\theta}_j$ er tilnærmet lik element (i, j) i $\bar{\mathbf{I}}^{-1}(\boldsymbol{\theta})$. Vi kan estimere varianser/kovarianser ved å sette inn $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ i $\bar{\mathbf{I}}^{-1}(\boldsymbol{\theta})$ eller i $\bar{\mathbf{J}}^{-1}(\boldsymbol{\theta})$.