



Lineární modely

Václav Adamec
vadamec@mendelu.cz

Lineární model



- Maticové vyjádření: $y = X\beta + \varepsilon$

y je $n \times 1$ vektor odezvy; $y \sim N(\mu, \sigma_e^2)$

X je $n \times p$ matice regresorů, důležitá je hodnota matice

β je $p \times 1$ vektor neznámých koeficientů

ε je $n \times 1$ vektor residuálních odchylek $\varepsilon \sim N(0, \sigma_e^2)$

- Střední hodnota modelu: $E(y) = E(y|X) = X\beta \rightarrow E(y_i) = x_i\beta$
- Disperse modelu: $Var(y) = V = I\sigma_e^2$
- Kovariance residuů: $Cov(\varepsilon_i, \varepsilon_j) = 0$

Metoda nejmenších čtverců



- Běžné nejmenší čtverce: Předpoklad stejnorodosti ko - variační struktury $V = I\sigma_e^2$
- Princip: simultánní volba hodnot β , které minimalizují výraz

$$\sum \varepsilon_i^2 = \varepsilon^T \varepsilon = (y - X\beta)^T (y - X\beta)$$

- První derivace výrazu $(y - X\beta)^T (y - X\beta)$ podle parametrického vektoru β ,

$$\frac{\partial (y - X\beta)^T (y - X\beta)}{\partial \beta} = 0$$

- Výstup postaven rovno nule (= soustava normálních rovnic): $X^T X\beta - X^T y = 0$

- Vektor β je pak roven $\hat{\beta} = (X^T X)^{-1} X^T y$

- Potom: $\hat{y} = X\hat{\beta}$ $\hat{e} = y - \hat{y}$ $\bar{y} = \bar{X}\hat{\beta}$

- Vlastnosti: $(y - X\beta)^T 1 = 0$ $(y - X\beta)^T X = 0$ $(y - \mu_y)^T 1 = 0$ $(\hat{y} - \mu_y)^T 1 = 0$

Generalizované nejmenší čtverce



- Předpoklad homogenity je porušen, V je strukturovaná, $V \neq I\sigma_e^2$!
- Řešení: simultánní volba hodnot β , které minimalizují výraz

$$\sum \varepsilon_i^2 = \varepsilon^T V^{-1} \varepsilon = (y - X\beta)^T V^{-1} (y - X\beta)$$

- Postup:
- První derivace výrazu $(y - X\beta)^T V^{-1} (y - X\beta)$ podle parametrického vektoru β ,
- Výstup postaven rovno nule a následuje výpočet řešení: $\frac{\partial (y - X\beta)^T V^{-1} (y - X\beta)}{\partial \beta} = 0$

- Vektor β je pak roven $\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y$

$$\hat{y} = X\hat{\beta}$$

- Residuální variance: $\hat{\sigma}_e^2 = \hat{e}^T \hat{e} / (n - p)$ $SST = y^T V^{-1} y - n\bar{y}^2$

- Determinace: $r^2 = 1 - \frac{\hat{e}^T \hat{e}}{SST}$ $s_e^2 = (1 - r^2) \cdot s_y^2$ $r_{adj}^2 = 1 - \frac{(1 - r^2)(n - 1)}{n - p}$

Testování hypotéz o β



- Odhadované parametry β vykazují střední hodnotu a rozptyl

$$\hat{\beta} \sim N(\beta, \sigma^2 (X^T X)^{-1}) \quad \hat{\beta}_j \sim N(\beta_j, \sigma^2 (X^T V^{-1} X)^{-1})$$

- T-testy koeficientů jsou pak: $H_0: \beta_j = c \quad H_1: \beta_j \neq c$

$$t_{obs} = \frac{\hat{\beta}_j - c}{\sqrt{\sigma^2 (X^T X)^{-1}_{jj}}} \sim t_{n-p}$$

- Modelové F-testy využití rozkladu součtu čtverců: $H_0: \beta_1 = \beta_2 = \dots = \beta_j = 0$

$$F_{obs} = \frac{SSM/p}{SSE/(n-p)} \sim F_{p, n-p}$$

- F-testy hodnot parametrického vektoru jsou: $H_0: \beta_j = 0 \quad H_1: \beta_j \neq 0$

$$F_{obs} = \frac{R(\beta_j | \beta_{j-1}, \beta_{j-2}, \dots, \beta_0)}{SSE/(n-p)} \sim F_{1, n-p}$$

Vícenásobný LM



- Typy statistických testů:

Sekvenční: vyjadřují krokovou sekvenci stavby úplného modelu

Parciální: vyjadřují variabilitu popsanou j-tým parametrem za podmínky ostatních $p-1$ regresorů v úplném modelu

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

Sekvenc

$$R(\beta_1 | \beta_0)$$

$$R(\beta_2 | \beta_0, \beta_1)$$

$$R(\beta_3 | \beta_0, \beta_1, \beta_2)$$

$$\Sigma = R(\beta_3, \beta_2, \beta_1 | \beta_0)$$

ModelSS

$$R(\beta_1 | \beta_0)$$

$$R(\beta_2, \beta_1 | \beta_0)$$

$$R(\beta_3, \beta_2, \beta_1 | \beta_0)$$

Parcial

$$R(\beta_1 | \beta_0, \beta_2, \beta_3)$$

$$R(\beta_2 | \beta_0, \beta_1, \beta_3)$$

$$R(\beta_3 | \beta_0, \beta_1, \beta_2)$$

$$\text{Sekvenční test průkaznosti } \beta_2: F_{obs} = \frac{R(\beta_2, \beta_1 | \beta_0) - R(\beta_1 | \beta_0)}{SSE_{kompl} / (n-p)} \sim F_{1, n-p}$$

Více o modelových F-testech



Parciální test průkaznosti β_3 v přítomnosti β_2, β_1 a β_0 :

$$F_{obs} = \frac{R(\beta_3, \beta_2, \beta_1 | \beta_0) - R(\beta_2, \beta_1 | \beta_0)}{SSE_{kompl} / (n-p)} \sim F_{1, n-p}$$

Simultánní test průkaznosti β_3 a β_2 v přítomnosti β_1 a β_0 :

$$F_{obs} = \frac{(R(\beta_3, \beta_2, \beta_1 | \beta_0) - R(\beta_1 | \beta_0)) / 2}{SSE_{kompl} / (n-p)} \sim F_{2, n-p}$$

Parciální test průkaznosti β_2 v přítomnosti β_3, β_1 a β_0 :

$$F_{obs} = \frac{(R(\beta_3, \beta_2 | \beta_1, \beta_0) - R(\beta_3 | \beta_1, \beta_0))}{SSE_{kompl} / (n-p)} \sim F_{1, n-p}$$

Sekvenční test průkaznosti β_1 v přítomnosti β_0 :

$$F_{obs} = \frac{R(\beta_1, \beta_0) - R(\beta_0)}{SSE_{kompl} / (n-p)} \sim F_{1, n-p}$$

ANOVA jedno-faktorová



- Zvláštní případ lineárního modelu s pevnými efekty
- Regresní model s kvalitativním, nespojitým regresorem
- Obecný princip: součty čtverců pro faktory se rovnají svým středním hodnotám
- Modelová rovnice:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

y_{ij} je závisle proměnná

μ je celkový průměr

α_i je odchylka průměru i - té skupiny podle jednoho třídícího faktoru

ε_{ij} jsou náhodné disturbanční odchylky

$$E(y_{ij}) = \mu + \alpha_i \quad Var(y_{ij}) = Var(\varepsilon_{ij}) = \sigma_\varepsilon^2$$

Předpoklady Anovy:

Homogenita variancí

$$\sigma_1^2 = \sigma_2^2 = \dots = \sigma_p^2$$

Normalita a nezávislost chyb

$$I.I.D., \varepsilon \sim N(0, \sigma_\varepsilon^2)$$

Skupiny se snad liší jen v lokálním parametru

Rozklad součtu čtverců



- Obecný rozklad: $[y^T y - n\bar{y}^2] = [y^T H y] + [y^T y - y^T H y]$

$$SST = SSM + SSE$$

SST = upravená suma čtverců (celková variabilita)

SSM = suma čtverců pro model (mezi skupinami)

SSE = suma čtverců pro chybu (uvnitř skupin)

- Projekční matice $H = X(X^T X)^{-1} X^T$
- Korekční faktor $n\bar{y}^2 = y^T H_0 y = y^T 1(1^T 1)^{-1} 1^T y$
- Parametrizace modelu: p skupin a p+1 parametrů
- Řeší se aplikací úpravou parametrických f-cí aplikací omezení
- Možná omezení: $\sum \alpha_i = 0; \mu = 0; \alpha_i = 0$

Testování průkaznosti



- Globální modelová hypotéza: $H_0: \beta_j = 0 \quad H_1: \beta_j \neq 0$

$$F_{obs} = \frac{SSM / v_A}{\hat{\sigma}_e^2} = \frac{R(\beta_j | \beta_0) / v_A}{SSE / (n - v_A)} \sim F_{v_A, n - v_A}$$

- Hypotéza průkaznosti efektu:

$$F_{obs} = \frac{SSM v_A}{\hat{\sigma}_e^2} = \frac{R(\beta_j | \beta_0) / v_A}{SSE / (n - v_A)} \sim F_{v_A, n - v_A}$$

- Hypotézy modelová a efekťová jsou u jednocestné Anovy shodné
- Následné hypotézy o rozdílech mezi úrovněmi efektu:
- Metoda kontrastu: vektor kontrastu $c = [c_1, c_2, \dots, c_t] ; \sum c_i = 0$

$$H_0: \sum c_i \bar{y}_i = 0$$

$$H_1: \sum c_i \bar{y}_i \neq 0$$

$$F_{obs} = t_{obs}^2 = \frac{(\sum c_i \bar{y}_i)^2}{\hat{\sigma}_e^2 \sum \frac{c_i^2}{n_i}} \sim F_{1, n - v_A}$$

ANOVA dvou - faktorová



- Modelová rovnice s pevnými efekty:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

y_{ijk} je závisle proměnná

μ je celkový průměr

α_i je odchylka průměru i - té skupiny podle třídícího faktoru A

β_j je odchylka průměru j - té skupiny podle třídícího faktoru B

$\alpha\beta_{ij}$ je odchylka průměru ij - té skupiny podle třídících faktorů A a B

ε_{ijk} jsou náhodné disturbanční odchylky

Střední hodnota modelu: $E(y_{ijk}) = \mu + \alpha_i + \beta_j + \alpha\beta_{ij}$

Variance modelu: $Var(y_{ijk}) = Var(\varepsilon_{ijk}) = \sigma_e^2$

Interakce 2. řádu: Závisí efekt úrovně faktoru A na úrovni faktoru B ?

Rozklad součtu čtverců



- Obecný rozklad: $[y^T y - n\bar{y}^2] = [y^T H y] + [y^T y - y^T H y]$

$$SST = SSM + SSE$$

$$SSM = SSA + SSB + SSAB$$

SST = upravená suma čtverců (celková variabilita)

SSA = suma čtverců pro rozdíly mezi skupinami podle faktoru A

SSB = suma čtverců pro rozdíly mezi skupinami podle faktoru B

SSAB = suma čtverců pro interakci efektů A a B

SSE = suma čtverců pro chybu (uvnitř podskupin AB)

- Projekční matice: $H_0 = 1(1^T 1)^{-1} 1^T$

$$H_M = X_M (X_M^T X_M)^{-1} X_M^T$$

$$H_A = X_A (X_A^T X_A)^{-1} X_A^T$$

$$H_B = X_B (X_B^T X_B)^{-1} X_B^T$$

$$H_I = H_M - H_A - H_B + H_0$$

Průkaznost efektů



- Modelový test:

$$F_{obs} = \frac{SSM/\nu}{\hat{\sigma}_e^2} = \frac{y^T(H-H_0)y/\nu}{SSE/(n-\nu)} \sim F_{\nu, n-\nu}$$

- Test faktoru A:

$$F_{obs} = \frac{SSA/\nu_A}{\hat{\sigma}_e^2} = \frac{y^T(H_A-H_0)y/\nu_A}{SSE/(n-\nu)} \sim F_{\nu_A, n-\nu}$$

- Test faktoru B:

$$F_{obs} = \frac{SSB/\nu_B}{\hat{\sigma}_e^2} = \frac{y^T(H_B-H_0)y/\nu_B}{SSE/(n-\nu)} \sim F_{\nu_B, n-\nu}$$

- Test interakce AB:

$$F_{obs} = \frac{SSAB(\nu_A\nu_B)}{\hat{\sigma}_e^2} = \frac{y^T(H_I)y/(\nu_A\nu_B)}{SSE/(n-\nu)} \sim F_{\nu_A\nu_B, n-\nu}$$

Typy efektů



- Jednoduché efekty: Test rozdílu mezi úrovněmi jednoho efektu, fixovaný na některé z úrovní druhého efektu

$$\bar{y}_{11} - \bar{y}_{12} \quad \bar{y}_{21} - \bar{y}_{22}$$

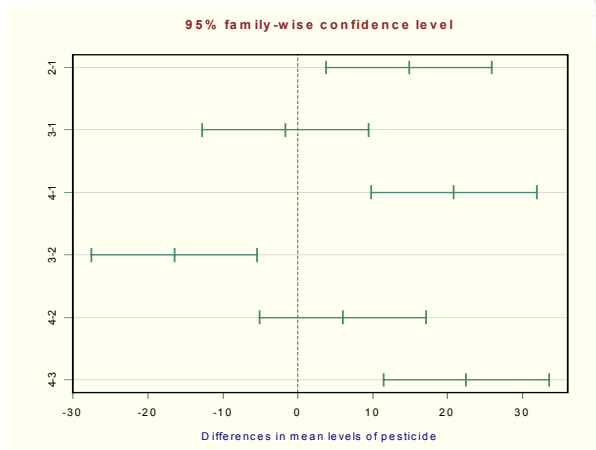
- Hlavní (marginální) efekty: Test rozdílu mezi úrovněmi jednoho efektu, průměrovanou přes úrovně druhého efektu

$$\bar{y}_{1.} - \bar{y}_{2.} \quad \bar{y}_{.1} - \bar{y}_{.2}$$

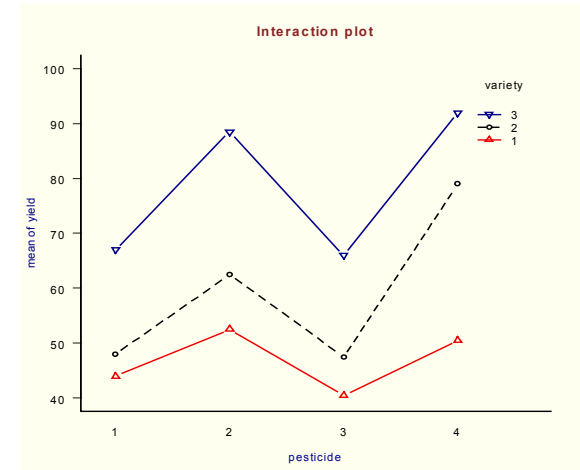
- Interakce: test rozdílu mezi jednoduchými efekty, paralelnosti

$$(\bar{y}_{11} - \bar{y}_{12}) - (\bar{y}_{21} - \bar{y}_{22})$$

Konf. interval hlavních efektů



Graf interakce



Rozklad SSA



- Ortogonalní kontrasty:** dva kontrasty s vektory koeficientů
 $c = [c_1, c_2, \dots, c_i] ; \sum c_i = 0$ $d = [d_1, d_2, \dots, d_i] ; \sum d_i = 0$
 jsou považovány za vzájemně ortogonální, jestliže platí $\sum_i c_i d_i = 0$
- Pro faktor A s $p = r(X_A)$ úrovněmi představuje soubor p-1 ortogonálních kontrastů kompletní soubor kontrastů. Součet čtverců náležejícím těmto ortogonálním kontrastům se sčítá do SSA
- Příklad pro první faktor, $p = 3$:

$$F_{obs} = \frac{SSA/2}{\hat{\sigma}_e^2} = \frac{\frac{1}{2} \left(\frac{(\sum c_i \bar{y}_i)^2}{\sum c_i^2 / n_i} + \frac{(\sum d_i \bar{y}_i)^2}{\sum d_i^2 / n_i} \right)}{\hat{\sigma}_e^2} \approx F_{2, n-v}$$

$\begin{bmatrix} -1, 0, 1 \\ -1, 2, -1 \end{bmatrix}$

Příklad redukce SS



- Jsou dány hodnoty

$R(\beta_0) = 90774$	$R(\beta_1\beta_0) = 94803,7$	$R(\beta_3\beta_1\beta_0) = 97275,6$
$R(\beta_2\beta_0) = 92981$	$R(\beta_2\beta_1\beta_0) = 97010,8$	$R(\beta_3\beta_2\beta_0) = 96859$
$R(\beta_3\beta_2\beta_1\beta_0) = 97463$		

Parciální SS pro úplný model s β_1 a β_2 ,
 $R(\beta_2|\beta_1\beta_0) = R(\beta_2\beta_1\beta_0) - R(\beta_1\beta_0) = 2207,1$
 $R(\beta_1|\beta_2\beta_0) = R(\beta_2\beta_1\beta_0) - R(\beta_2\beta_0) = 4029,8$

Sekvenční SS pro model s β_1 , β_2 a β_3 ,
 $R(\beta_1|\beta_0) = R(\beta_1\beta_0) - R(\beta_0) = 4029,8$
 $R(\beta_2|\beta_1\beta_0) = R(\beta_2\beta_1\beta_0) - R(\beta_1\beta_0) = 2207,1$
 $R(\beta_3|\beta_2\beta_1\beta_0) = R(\beta_3\beta_2\beta_1\beta_0) - R(\beta_2\beta_1\beta_0) = 452,2$

Příklad redukce SS



Modelové SS pro modely s β_1 , β_2 a β_3 ,
 $R(\beta_1|\beta_0) = R(\beta_1\beta_0) - R(\beta_0) = 4029,8$
 $R(\beta_2\beta_1|\beta_0) = R(\beta_2\beta_1\beta_0) - R(\beta_0) = 6236,8$
 $R(\beta_3\beta_2\beta_1|\beta_0) = R(\beta_3\beta_2\beta_1\beta_0) - R(\beta_0) = 6689$

Parciální SS pro modely s β_1 , β_2 a β_3 (nemusí mít smysl)
 $R(\beta_3|\beta_2\beta_1\beta_0) = R(\beta_3\beta_2\beta_1\beta_0) - R(\beta_2\beta_1\beta_0) = 452,2$
 $R(\beta_2|\beta_3\beta_1\beta_0) = R(\beta_3\beta_2\beta_1\beta_0) - R(\beta_3\beta_1\beta_0) = 604$
 $R(\beta_1|\beta_3\beta_2\beta_0) = R(\beta_3\beta_2\beta_1\beta_0) - R(\beta_3\beta_2\beta_0) = 187,38$

Je možné určit SSE ?

Pevné a náhodné efekty



- Pevné a náhodné efekty rozděluje „filosofický pohled“,
- Pevné efekty:
 - Úrovní je „omezený počet“
 - Úrovně jsou dány experimentálním designem
 - Cílem je především estimace střední hodnoty empirických úrovní a event. výběrové chyby
- Náhodné efekty:
 - Úrovní je „vysoký počet“
 - Úrovně jsou dány f-cí pravděpodobnostní hustoty s předpokládanou střední hodnotou ($\mu = 0$) a rozptylem, event. kovarianční strukturou (σ_a^2),
 - Cílem je především předikce úrovní (i nepozorovaných) a komponenty variance
- Hlavní otázka: „Lze tvrdit, že úrovně efektu pocházejí z jistého náhodného procesu?“
 - Pokud ano, je efekt náhodný
 - Pokud ne, efekt je fixní

Náhodný model



- Obsahuje jen náhodné efekty (mimo intercept)

- Modelová rovnice:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

y_{ij} je závisle proměnná

μ je celkový průměr (pevný)

α_i je odchylka průměru i - té skupiny podle jednoho třídícího faktoru (náhodný)

ε_{ij} jsou náhodné disturbanční odchylky

Modelové momenty: $E(y_{ij}) = \mu$ $Var(y_{ij}) = Var(\alpha_i) + Var(\varepsilon_{ij}) = \sigma_a^2 + \sigma_e^2$

Předpoklady: $\alpha \sim N(0, I\sigma_a^2)$ $\varepsilon \sim N(0, I\sigma_e^2)$

Nezávislost odchylek: $Cov(\varepsilon_{ijk}, \varepsilon_{ijk'}) = 0$ $Cov(\alpha_i, \alpha_{i'}) = 0$

$$Cov(\alpha_i, \varepsilon_{ijk}) = 0$$

Hypotéza: $H_0: \sigma_a^2 = 0$ $H_1: \sigma_a^2 \neq 0$

Náhodný model



- Estimace variance σ_a^2 z tabulky Anovy:

- EMS:

$$E(MSE) = \sigma_e^2$$

$$E(MSA) = \sigma_e^2 + n\sigma_a^2$$

$$\sigma_e^2 = MSE = SSE / (n - p)$$

$$\sigma_a^2 = (MSA - MSE) / n$$

- Test σ_a^2 :

$$F_{obs} = \frac{MSA}{MSE} \sim F_{p-1, n-p}$$

- Odhady: Anova umožňuje σ_a^2 mimo parametrický prostor

- Řešení: úprava modelu, rozsáhlejší soubor, korekce na nulu

- Intraklasní korelace r :

$$r = \frac{\sigma_a^2}{\sigma_e^2 + \sigma_a^2} = \frac{MSA - MSE}{MSA + (p - 1)MSE}$$

- $(1 - \alpha)100\%$ konfidenční interval:

$$\frac{F_{obs} - F_{1-\alpha/2, p-1, n-p}}{F_{obs} + (p-1)F_{1-\alpha/2, p-1, n-p}} \leq \rho \leq \frac{F_{obs} - F_{\alpha/2, p-1, n-p}}{F_{obs} + (p-1)F_{\alpha/2, p-1, n-p}}$$

Smíšený lineární model



- Obsahuje pevné i náhodné efekty

- Maticové vyjádření:

$$y = X\beta + Za + \varepsilon$$

y je $n \times 1$ vektor odezvy; $y \sim N(X\beta, \sigma_a^2 + \sigma_e^2)$

X je $n \times p$ incidenční matice regresorů

β je $p \times 1$ vektor neznámých koeficientů

Z je $n \times q$ designová matice náhodných efektů

a je $q \times 1$ vektor náhodných odchylek $a \sim N(0, \sigma_a^2)$

ε je $n \times 1$ vektor residuálních odchylek $\varepsilon \sim N(0, \sigma_e^2)$

Momenty modelu: $E \begin{bmatrix} y \\ a \\ e \end{bmatrix} = \begin{bmatrix} X\beta \\ 0 \\ 0 \end{bmatrix}$ $Var \begin{bmatrix} y \\ a \\ e \end{bmatrix} = \begin{bmatrix} ZGZ^T + R & ZG & R \\ GZ^T & G & 0 \\ R & 0 & R \end{bmatrix}$

$$Cov(\varepsilon_i, \varepsilon_j) = 0$$

$$Cov(\alpha_i, \varepsilon_{ijk}) = 0$$

$$R = I\sigma_e^2$$

$$G = A\sigma_a^2$$

Smíšený lineární model



- Rovnice smíšeného modelu (OLMME):

$$\begin{bmatrix} X^T X & X^T Z \\ Z^T X & Z^T Z + \sigma_e^2 G^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X^T Y \\ Z^T Y \end{bmatrix}$$

- Řešení OLMME:

$$\begin{bmatrix} \hat{\beta} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X^T X & X^T Z \\ Z^T X & Z^T Z + \sigma_e^2 G^{-1} \end{bmatrix}^{-1} \begin{bmatrix} X^T Y \\ Z^T Y \end{bmatrix}$$

- V části $X^T X$ pravděpodobná singularita $\rightarrow$ řešení β nejsou jedinečná

- V části $Z^T Z$ nepravděpodobná singularita $\rightarrow$ řešení a jsou jedinečná

- Inverze matice C : regulární inverze, generalizovaná (pseudo-) inverze

Smíšený lineární model



- Generalizovaná forma GLMM pro řešení:

$$\begin{bmatrix} \hat{\beta} \\ \hat{\alpha} \end{bmatrix} = \begin{bmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{bmatrix}^{-1} \begin{bmatrix} X^T R^{-1} Y \\ Z^T R^{-1} Y \end{bmatrix}$$

- Matrice C** : slouží k výpočtu variance odhadů, nebo jejich lineárních kombinací $k^T \hat{\beta}$ nebo $k^T \hat{\alpha}$, tedy kontrastů,

$$C = \begin{bmatrix} C^{XX} & C^{XZ} \\ C^{ZX} & C^{ZZ} \end{bmatrix}^{-1} = \begin{bmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{bmatrix}^{-1}$$

$$Var(\hat{\beta}) = Var(\hat{\beta} - \beta) = C^{XX} \sigma_e^2 \quad Var(\hat{\alpha}) = Var(\hat{\alpha} - \alpha) = C^{ZZ} \sigma_e^2$$

$$Var(k^T \hat{\beta}) = k^T C^{XX} k \sigma_e^2 \quad Var(k^T \hat{\alpha}) = k^T C^{ZZ} k \sigma_e^2$$

- Kontrasty lze testovat jednotlivě (t-test) nebo skupinově (F-test):

$$t_{obs} = \frac{k^T \hat{\beta}}{\sqrt{Var(k^T \hat{\beta})}} \sim t_{n-p} \quad F_{obs} = \frac{(K\hat{\beta})^T (KCK^T)^{-1} (K\hat{\beta}) / p}{\sigma_e^2} \sim F_{p, n-p}$$

Testování efektů



- Požadavky:

Model je interpretovatelný Průkazný model i dílčí efekty
Modelový popis max. podílu variability Minimální variance chyby

- Testování efektů: F-test redukce residuálního součtu čtverců
- Vhodný k testování jednotlivých efektů a zahnížděných submodelů
- Princip: přidavek efektu do modelu SSE nevyšší
- Příklad: dva modely: kompletní a redukovaný

$$\begin{array}{ll} \text{kompletní} & \text{redukovaný} \\ R(\beta_1, \beta_2, \dots, \beta_p \mid \beta_0) & R(\beta_1, \beta_2, \dots, \beta_{p-1} \mid \beta_0) \end{array}$$

- Testovací statistika:

$$F_{obs} = \frac{(SSE_{red} - SSE_{kompl}) / (v_{red} - v_{kompl})}{SSE_{kompl} / v_{kompl}} \sim F_{v_{red} - v_{kompl}, v_{kompl}}$$

Komponenty variance



- Platí, že $\sum_q \hat{a} = 0$ $proG = I\sigma_a^2$
 $1^T A^{-1} \hat{a} = 0$ $proG = A\sigma_a^2$

- Residuální rozptyl :

$$\sigma_e^2 = \frac{SSE}{n - r(X)}$$

- Rozptyl náhodného efektu a :

$$\sigma_a^2 = \frac{\hat{a}^T A^{-1} \hat{a} + tr(C^{ZZ} A^{-1})}{r(Z)}$$

Matice příbuzenství



- Symetrická čtvercová matice A (q x q) charakterizující kovarianční strukturu náhodných efektů a_i
- Jsou-li zvířata nepříbuzná, matice A je jednotková
- Koeficienty F_x se nalézají na hlavní diagonále po úpravě
- Aditivní vztah jedinců X a Y je dán 1 / 2 aditivního vztahu matky X a Y a 1 / 2 aditivního vztahu otce X a Y
- Výpočetně se řeší přes přímé sestavení A^{-1} (Hendersonova pravidla)

Maximální věrohodnost (ML)



- Aplikovatelná za podmínek normality a nezávislosti residuí
- Využití p.d.f. , (popř. i p.m.f.)

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-X\beta)^T(y-X\beta)}{2\sigma^2}}$$

- Kalkulace věrohodnostní f-ce:

$$L(y, \beta) = \prod_i f(y_i) = \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-X\beta)^T(y-X\beta)}{2\sigma^2}}$$

- Logaritmus věrohodnostní f-ce (skórovací funkce):

$$S(\theta) = \ln L(y, \beta) = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} (y-X\beta)^T (y-X\beta)$$

- $E(S(\theta)) = 0$, $\text{Var}(S(\theta)) = I(\theta)$

- Maximalizací log věrohodnostní f-ce: $\frac{\partial \ln L(\varepsilon, \sigma^2)}{\partial \beta} = 0$

- Vyřešením se získá: $-2X^T y + 2X^T X\beta = 0$ $\tilde{\beta} = (X^T X)^{-1} X^T y$

Maximální věrohodnost (ML)



- Odhad parametrů β roven odhadu MNČ, avšak odhad σ^2 je odlišný:

$$\tilde{\sigma}^2 = \frac{SSE}{n}$$

- ML odhady mají dobré vlastnosti za podmínky asymptoty ($n \gg p$ a $n \rightarrow \infty$), MNČ odhady a ML odhady jsou pak stejné
- ML odhady rozptylu jsou asymptoticky nevychýlené $E(\tilde{\sigma}^2_{n \rightarrow \infty}) = \sigma^2$
- ML odhady rozptylu za podmínky saturevaného modelu ($n = p$) nejsou vůbec možné
- ML odhady variancí vždy v parametrickém prostoru
- Invariance odhadu: $f(\theta) = f(\hat{\theta}_{ML})$

Maximální věrohodnost (ML)



- Asymptotická kovarianční matice parametrů je invertovaná Fisherova informační matice $I(\theta)^{-1}$

$$I(\theta) = \left[\frac{\partial^2 \ln L(y, \beta)}{\partial \beta \partial \beta} \right] = E \left[\left(\frac{\partial \ln L(y, \beta)}{\partial \beta} \right) \left(\frac{\partial \ln L(y, \beta)}{\partial \beta} \right)^T \right]$$

- V případě OL-MME rovno

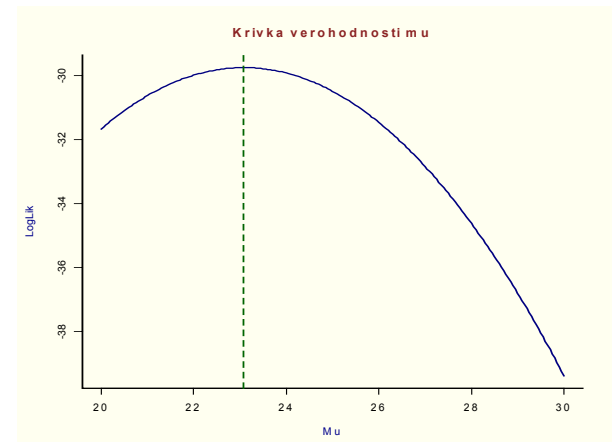
$$I(\theta) = -[X^T X]$$

- V případě GL-MME rovno

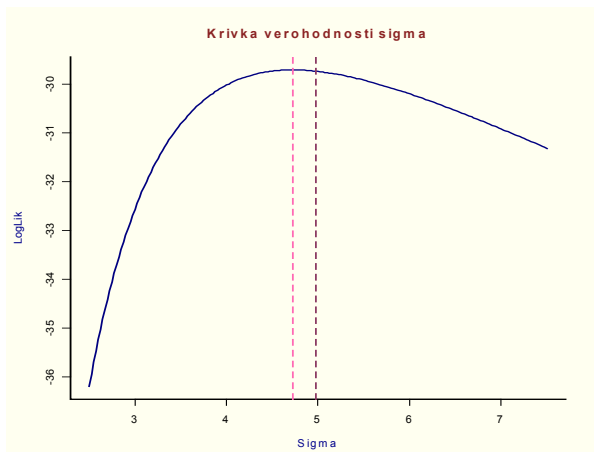
$$I(\theta) = -[X^T V^{-1} X]$$

- Variance závisí na průběhu log věrohodnostní funkce $L(y, \beta)$

Křivka log věrohodnostní f-ce



Křivka log věrohodnostní f-ce



REML



- Metoda odhadu komponentů variance využívající věrohodnosti residuální složky po odhadu pevných efektů metodou OLS popř. GLS
- Princip: Maximalizace věrohodnosti kontrastů residuí
- Definice kontrastů $K^T y$, přičemž $K^T X \beta = 0$ a $K^T X = 0$, $r(K) = n - r(X)$
- Podmínka: $K^T y = K^T (Z\alpha + e)$; $K^T y \sim N(0, K^T V K)$

$$f(K^T y) = \frac{1}{(2\pi)^{n/2} |K^T V^{-1} K|^{-1/2}} e^{-\frac{1}{2} [(K^T y)^T (K^T V K)^{-1} (K^T y)]}$$

$$\ln L(K^T y, \sigma) = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln |K^T V K| - \frac{1}{2} [(K^T y)^T (K^T V K)^{-1} (K^T y)]$$

REML



- Maximalizací log restringované věrohodnostní f-ce podle σ^2 ...:

$$\frac{\partial \ln L(K^T y, \sigma)}{\partial \sigma^2} = 0$$

- Balancovaná Anova a REML shodné
- REML provádí restrikcí, aby odhad splnil podmínky parametrického prostoru ($\sigma^2 \geq 0$)
- Odhady REML jsou „méně“ vychýlené
- Podtypy REML:
Vyžadující derivace: EM-REML, AI-REML
Bez derivací: DF-REML

Věrohodnostní testy



- K testování jednotlivých efektů
- Pracuje s hodnotami věrohodnosti modelů $L(\beta, y)$
- Věrohodnost úplného modelu $\geq$ věrohodnost redukovaného modelu
- Test věrohodnostního poměru:

$$LR = 2(\ln(L(\beta_{\text{kompl}}, y)) - \ln(L(\beta_{\text{red}}, y))) \sim \chi^2_{v=\text{redukce}}$$

- Akaikeho informační kritérium (žádoucí je minimum)
 $AIC = -2 \ln(L(\beta, y)) + 2p$
- Schwarzovo informační kritérium (žádoucí je minimum)
 $BIC = -2 \ln(L(\beta, y)) + p \log(n)$
- C_p – statistika (žádoucí $C_p \rightarrow p$) $C_p = \frac{SSE}{\sigma^2} + 2p - n$

Sire model



- Původní metoda odhadu PBVs
- Matice A obsahuje známá příbuzenství po paternální (GP) linii
- Příbuzenství po maternální linii ignorována

$$y = X\beta + Zs + \varepsilon$$

- s je ½ PBV otce, tzv. PTA
- Platí, že: $V(s) = A_s \sigma_s^2 = \frac{1}{4} A_s \sigma_a^2$ $\sigma_e^2 = (1 - h^2 / 4) \sigma_y^2$ $\sigma_s^2 = (h^2 / 4) \sigma_y^2$

$$\begin{bmatrix} X^T X & X^T Z \\ Z^T X & Z^T Z + A_s^{-1} \frac{\sigma_e^2}{\sigma_s^2} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{s} \end{bmatrix} = \begin{bmatrix} X^T Y \\ Z^T Y \end{bmatrix}$$

- Problémy: samice považovány za „nepříbuzné“
- Období S-MGS model:

$$y = X\beta + Zs + Ws + \varepsilon$$

Animal model



- Současná metoda odhadu PBVs, vývin ze Sire modelu
- Matice A obsahuje všechna známá příbuzenství po linii maternální i paternální
- Existuje v mnoha variantách: M, PE, C, atd.
- Příklad: s efekty jednotlivých genů (QTL efekty)
- Maticové vyjádření:

$$y = X\beta + Za + \varepsilon$$

$$y = X\beta + (Za + ZTv) + \varepsilon$$

Y je n x 1 vektor odezvy; $y \sim N(X\beta, \sigma_m^2 + \sigma_a^2 + \sigma_e^2)$

X je n x p matice regresorů

β je p x 1 vektor neznámých koeficientů

Z je n x q designová matice polygenních náhodných efektů

a je q x 1 vektor náhodných odchylek $a \sim N(0, A\sigma_a^2)$

T je q x 2q designová matice náhodných efektů alel

v je 2q x 1 vektor náhodných odchylek $v \sim N(0, V\sigma_m^2)$

ε je n x 1 vektor residuálních odchylek $\varepsilon \sim N(0, I\sigma_e^2)$

Animal model



- Princip: plemenná hodnota je souhrnem efektů párů alel na k lokusech

$$a_i = \sum_k (v_{ik}^1 + v_{ik}^2) \quad \sigma_a^2 = \sum_k \sigma_k^2$$

- Pak: $\sigma_k^2 = 2\sigma_{vk}^2$
- Vlastnosti: $Var(a) = A\sigma_a^2$ $Var(v) = V\sigma_v^2$ $Cov(a, e) = 0$
- Aspekty: $Cov(v, e) = 0$ $Cov(a, v) = 0$

Výběr markerů pro analýzu: kolik a které (= výpočetní zjednodušení)

Informativnost markerů: původ alely potomka musí být znám u rodičů

Používané designy: Back-crossing, F2 design

- LMME (náhod. část):
- $$\begin{bmatrix} Z^T Z + A^{-1} \frac{\sigma_e^2}{\sigma_a^2} & Z^T ZT \\ T^T Z^T Z & T^T Z^T ZT + V^{-1} \frac{\sigma_e^2}{\sigma_v^2} \end{bmatrix} \begin{bmatrix} \hat{a} \\ \hat{v} \end{bmatrix} = \begin{bmatrix} Z^T Y \\ T^T Z^T Y \end{bmatrix}$$

Animal Model ukázka



8 příbuzných jedinců ve 3 skupinách managementu (např. HYS efekt), výskyt inbreedingu, vlastní užitkovost

Y = 10, 11, 10, 7, 9, 11, 12, 13

X = 2, 1, 3, 1, 2, 3, 1, 2

J = 1, 2, 3, 4, 5, 6, 7, 8

O = 0, 0, 1, 1, 1, 4, 5, 7

M = 0, 0, 2, 2, 3, 2, 6, 2

A =

1,00	0,00	0,5000	0,5000	0,75000	0,25000	0,50000	0,25000
0,00	1,00	0,5000	0,5000	0,25000	0,75000	0,50000	0,75000
0,50	0,50	1,0000	0,5000	0,75000	0,50000	0,62500	0,56250
0,50	0,50	0,5000	1,0000	0,50000	0,75000	0,62500	0,56250
0,75	0,25	0,7500	0,5000	1,25000	0,37500	0,81250	0,53125
0,25	0,75	0,5000	0,7500	0,37500	1,25000	0,81250	0,78125
0,50	0,50	0,6250	0,6250	0,81250	0,81250	1,18750	0,84375
0,25	0,75	0,5625	0,5625	0,53125	0,78125	0,84375	1,25000

Animal Model ukázka



- C_{jj} diagonální prvky pro zvířata neinbrední a inbrední (červeně):
0,226; 0,226; 0,241; 0,240; 0,286; 0,290; 0,277; 0,283;
- C_{jj} diagonální prvky pro všechna zvířata neinbrední (rodíče, vrh):
0,232; 0,232; 0,243; 0,240; 0,240; 0,243; 0,240; 0,240;