Lineární statistické modely II

1 Úvod

Prednášky z predmetu *Lineární statistické modely II* nadväzujú na predmety , *Pravděpodobnost a statistika I, II* a *Lineární statistické modely I*. Predpokladajú sa znalosti získané v týchto predmetoch. Odporúčaná literatúra k štúdiu je

Anděl, J., *Matematická statistika*, SNTL, Praha, 1985.

Rao, C., R., *Lineární metódy statistické indukce a jejich aplikace*, ACADEMIA, Praha, 1978.

Zvára, K., *Regresní analýza*, ACADEMIA, Praha, 1989.

2 Testy dobrej zhody

2.1 Multinomické rozdelenie

Majme urnu a v nej guľky k farieb. Pravdpodobnsť vytiahnutia guľky $i-$tej farby je $\theta_i,\ i = 1, 2, ..., k$, $0 < \theta_i < 1,\ \theta_1 + \theta_2 + ... + \theta_k = 1$. $n-$krát nezávisle ťaháme vždy jednu guľku s vrátením. Označíme si jej farbu. Nech

náhodná premenná $X_1^{(n)}$ je počet vytiahnutých guliek 1. farby v n ťahoch

náhodná premenná $X_2^{(n)}$ je počet vytiahnutých guliek 2. farby v n ťahoch

$\vdots$

náhodná premenná $X_k^{(n)}$ je počet vytiahnutých guliek k. farby v n ťahoch.

Teda máme náhodný vektor $\mathbf{X}_{k,1}^{(n)} = \left(X_1^{(n)}, X_2^{(n)}, ..., X_k^{(n)}\right)',\quad X_i^{(n)}$ sú diskrétne náhodné premenné, ktoré nadobúdajú hodnoty z $\{0, 1, ..., n\}$. Počítajme

$$P\left\{X_1^{(n)} = x_1, X_2^{(n)} = x_2, ..., X_k^{(n)} = x_k\right\}.$$

Zrejme táto pravdepodobnosť je nenulová len pre $x_i \in \{0, 1, ..., n\}$, pričom $x_1 + x_2 + ... + x_k = n$, inde je nulová.

Pravdepodobnosť postupnosti vytiahnutých guliek, ktorá (postupnosť) obsahuje x_1 guliek 1.farby, x_2 guliek 2.farby,...,x_k guliek k.farby je $\theta_1^{x_1}\theta_2^{x_2}...\theta_k^{x_k}$. Počet možných "vytiahnutých postupností" guliek, ktoré obsahujú x_1 guliek 1.farby, x_2 guliek 2.farby,...,x_k guliek k.farby je

$$\binom{n}{x_1}\binom{n-x_1}{x_2}...\binom{n-x_1-...-x_{k-2}}{x_{k-1}} =$$

$$= \frac{n!}{(n-x_1)!\ x_1!}\ \frac{(n-x_1)!}{(n-x_1-x_2)!\ x_2!}\ ...\ \frac{(n-x_1-...-x_{k-2})!}{(n-x_1-...-x_{k-2}-x_{k-1})!\ x_{k-1}!}\ = \frac{n!}{x_1!\ x_2!...x_k!},$$

teda

$$P\left(\mathbf{X}^{(n)} = \mathbf{x}\right) = \frac{n!}{x_1!\ x_2!...x_k!}\ \theta_1^{x_1}\theta_2^{x_2}...\theta_k^{x_k} \tag{1}$$

pre $x_i \in \{0,1,...,n\}$, $x_1+x_2+...+x_n = n$, inde je rovná 0. Rozdelenie pravdepodobnosti dané pravdepodobnostnou funkciou (1) sa volá *multinomické* s parametrami $n, \theta_1, ..., \theta_k$ a značíme $\mathbf{X}^{(n)} \sim Mu(n, \theta_1, ..., \theta_k)$.

Poznámka 2.1: V $Mu(n, \theta_1, ..., \theta_k)$ rozdelení je $k-1$ "nezávislých" parametrov.
Označme X_{ij} náhodnú veličinu

$$X_{ij} = \begin{cases} 1, & \text{ak v } i-\text{tom ťahu vytiahneme guľku } j-\text{tej farby}, \\ 0, & \text{inak}, \end{cases}$$

$i = 1,2,...,n$, $j = 1,2,...,k$. Platí

$$P\{X_{ij} = 1\} = \theta_j, \qquad P\{X_{ij} = 0\} = 1 - \theta_j,$$

$$\mathcal{E}(X_{ij}) = 0(1-\theta_j)+1\theta_j = \theta_j, \ \mathcal{D}(X_{ij}) = \mathcal{E}(X_{ij}-\theta_j)^2 = \mathcal{E}(X_{ij}^2)-\mathcal{E}^2(X_{ij}) = 0^2(1-\theta_j)+1^2\theta_j-\theta_j^2 = \theta_j(1-\theta_j),$$

$$cov(X_{ij}, X_{is}) = \mathcal{E}((X_{ij}-\theta_j)(X_{is}-\theta_s)) = \mathcal{E}(X_{ij}X_{is}) - \theta_j\theta_s = -\theta_j\theta_s$$

pre $j \neq s$, lebo náhodná veličina $X_{ij}X_{is}$ nadobúda len hodnotu 0.

Teda výsledok $i-$teho ťahu popisuje náhodný vektor

$$\mathbf{X}_i = \begin{pmatrix} X_{i1} \\ X_{i2} \\ \vdots \\ X_{ik} \end{pmatrix}, \quad \mathcal{E}(\mathbf{X}_i) = \boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_k \end{pmatrix}, \quad cov\mathbf{X}_i = \begin{pmatrix} \theta_1(1-\theta_1) & -\theta_1\theta_2 & \dots & -\theta_1\theta_k \\ -\theta_2\theta_1 & \theta_2(1-\theta_2) & \dots & -\theta_2\theta_k \\ \vdots & & \ddots & \vdots \\ -\theta_k\theta_1 & & \dots & \theta_k(1-\theta_k) \end{pmatrix}.$$

Zrejme platí $\mathbf{X}_i \sim Mu(1, \theta_1, ..., \theta_k)$ (dokážte ako cvičenie) a tiež $\mathbf{X}^{(n)} = \sum_{i=1}^n \mathbf{X}_i$, pričom $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n$ sú nezávislé.

V ďalšom označme

$$\mathbf{D} = \begin{pmatrix} \sqrt{\theta_1} & 0 & \dots & 0 \\ 0 & \sqrt{\theta_2} & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & & \dots & \sqrt{\theta_k} \end{pmatrix}, \quad \mathbf{U} = \begin{pmatrix} \sqrt{\theta_1} \\ \sqrt{\theta_2} \\ \vdots \\ \sqrt{\theta_k} \end{pmatrix}, \quad \mathbf{Q} = \mathbf{I} - \mathbf{U}\mathbf{U}'.$$

Veta 2.2: (Vlastnosti $Mu(n, \theta_1, ..., \theta_k)$.) Nech $\mathbf{X}^{(n)} \sim Mu(n, \theta_1, ..., \theta_k)$, potom platí

(i) $\mathcal{E}(\mathbf{X}^{(n)}) = n\boldsymbol{\theta}$,

(ii) $cov\mathbf{X}^{(n)} = \begin{pmatrix} n\theta_1(1-\theta_1) & -n\theta_1\theta_2 & \dots & -n\theta_1\theta_k \\ -n\theta_2\theta_1 & n\theta_2(1-\theta_2) & \dots & -n\theta_2\theta_k \\ \vdots & & \ddots & \vdots \\ -n\theta_k\theta_1 & & \dots & n\theta_k(1-\theta_k) \end{pmatrix}$,

(iii) hodnosť $h(cov\mathbf{X}^{(n)}) = k-1$.

(iv) jedna zovšeobecnená inverzia matice $cov\mathbf{X}^{(n)}$ je

$$\left(cov\mathbf{X}^{(n)}\right)^- = \begin{pmatrix} \frac{1}{n\theta_1} & 0 & \dots & 0 \\ 0 & \frac{1}{n\theta_2} & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & & \dots & \frac{1}{n\theta_k} \end{pmatrix}.$$

Dôkaz:

(i) pretože $\mathbf{X}^{(n)} = \sum_{i=1}^{n} \mathbf{X}_i$, je $\mathcal{E}(\mathbf{X}^{(n)}) = \mathcal{E} \sum_{i=1}^{n} \mathbf{X}_i = \sum_{i=1}^{n} \mathcal{E}(\mathbf{X}_i) = n\boldsymbol{\theta}$,

(ii) $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n$ sú nezávislé, preto

$$cov(\mathbf{X}^{(n)}) = cov(\sum_{i=1}^{n} \mathbf{X}_i) = \sum_{i=1}^{n} cov\mathbf{X}_i = n \begin{pmatrix} \theta_1(1-\theta_1) & -\theta_1\theta_2 & \dots & -\theta_1\theta_k \\ -\theta_2\theta_1 & \theta_2(1-\theta_2) & \dots & -\theta_2\theta_k \\ \vdots & & \ddots & \vdots \\ -\theta_k\theta_1 & & \dots & \theta_k(1-\theta_k) \end{pmatrix},$$

(iii) platí

$$cov(\mathbf{X}^{(n)}) = n \left[\begin{pmatrix} \theta_1 & 0 & \dots & 0 \\ 0 & \theta_2 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & & \dots & \theta_k \end{pmatrix} - \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_k \end{pmatrix} (\theta_1, \theta_2, ..., \theta_k) \right] =$$

$$= n[\mathbf{DD} - \mathbf{DUU'D}] = n\mathbf{D}(\mathbf{I} - \mathbf{UU'})\mathbf{D} = n\mathbf{DQD},$$

(poznamenajme len, že $\mathbf{D}$ je regulárna matica a $\mathbf{U'U} = 1$)

$$h(\mathbf{Q}) \geq h(\mathbf{DQD}) \geq h(\mathbf{D}^{-1}\mathbf{DQDD}^{-1}) = h(\mathbf{Q}),$$

čiže

$$h(\mathbf{Q}) = h(\mathbf{DQD}) = h(n\mathbf{DQD}) = h\left(cov\mathbf{X}^{(n)}\right),$$

($\mathbf{Q}$ je idempotentná)

$$h(\mathbf{Q}) = tr\mathbf{Q} = tr(\mathbf{I} - \mathbf{UU'}) = tr\mathbf{I}_{k,k} - tr\mathbf{UU'} = k - tr\mathbf{U'U} = k - 1,$$

(iv)

$$\left(cov\mathbf{X}^{(n)}\right) \left[\frac{1}{n}\mathbf{D}^{-1}\mathbf{D}^{-1}\right] \left(cov\mathbf{X}^{(n)}\right) = n\mathbf{DQD} \left[\frac{1}{n}\mathbf{D}^{-1}\mathbf{D}^{-1}\right] n\mathbf{DQD} = n\mathbf{DQD} = \left(cov\mathbf{X}^{(n)}\right),$$

teda

$$\frac{1}{n}\mathbf{D}^{-1}\mathbf{D}^{-1} = \frac{1}{n} \begin{pmatrix} \frac{1}{\sqrt{\theta_1}} & 0 & \dots & 0 \\ 0 & \frac{1}{\sqrt{\theta_2}} & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & & \dots & \frac{1}{\sqrt{\theta_k}} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{\theta_1}} & 0 & \dots & 0 \\ 0 & \frac{1}{\sqrt{\theta_2}} & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & & \dots & \frac{1}{\sqrt{\theta_k}} \end{pmatrix} = \begin{pmatrix} \frac{1}{n\theta_1} & 0 & \dots & 0 \\ 0 & \frac{1}{n\theta_2} & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & & \dots & \frac{1}{n\theta_k} \end{pmatrix}$$

je jedna zovšeobecnená inverzia matice $cov\mathbf{X}^{(n)}$. Q.E.D.

Poznámka 2.3: Ľahko vidíme, že $cov\mathbf{X}_i = \mathbf{DQD}$.

Veľmi dôležité asmptotické vlastnosti náhodného vektora s $Mu(n, \theta_1, ..., \theta_k)$ rozdelením pravdepodobnosti dostaneme použitím mnohorozmernej centrálnej limitnej vety a mnohorozmernej Sverdrupovej vety.

Veta 2.4: (Mnohorozmerná centrálna limitná veta.) Nech $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, ...$ sú nezávislé, rovnako rozdelené náhodné vektory z $\mathcal{R}^k$, ktoré majú strednú hodnotu $\boldsymbol{\mu}$ a kovariančnú maticu $\mathbf{V}$ (s konečnými prvkami). Ak označíme $\mathbf{Z}_n = \frac{1}{\sqrt{n}}(\boldsymbol{\xi}_1 - \boldsymbol{\mu}) + ... + \frac{1}{\sqrt{n}}(\boldsymbol{\xi}_n - \boldsymbol{\mu})$, tak

$$\mathbf{Z}_n \xrightarrow[n]{\mathcal{D}} N_k(\mathbf{0}, \mathbf{V})$$

(konvergencia v distribúcii).

Dôkaz: nájdeme v knižke Anděl, J., *Matematická statistika*, SNTL, Praha, 1985, str. 185.

Veta 2.5: (Mnohorozmerná Sverdrupova veta.) Ak $b: \mathcal{R}^k \to \mathcal{R}^1$ je spojitá reálna funkcia a $\boldsymbol{\xi}_n \xrightarrow[n]{\mathcal{D}} \boldsymbol{\xi}$, tak $b(\boldsymbol{\xi}_n) \xrightarrow[n]{\mathcal{D}} b(\boldsymbol{\xi})$.

Dôkaz: jednorozmernej Sverdrupovej vety nájdeme v knižke Anděl, J., *Matematická statistika*, SNTL, Praha, 1985, str. 185.

Veta 2.6: (Asymptotické vlastnosti náhodného vektora $\mathbf{X}^{(n)}$.) Pre náhodný vektor $\mathbf{X}^{(n)}$ platí:

(i) $\mathbf{Y}_n = \frac{1}{\sqrt{n}}\mathbf{D}^{-1}(\mathbf{X}^{(n)} - n\boldsymbol{\theta}) \xrightarrow[n]{\mathcal{D}} N_k(\mathbf{0}, \mathbf{Q})$,

(ii) ${}_{(n)}\chi^2 = \sum_{j=1}^k \dfrac{(X_j^{(n)} - n\theta_j)^2}{n\theta_j} \xrightarrow[n]{\mathcal{D}} \chi^2_{k-1}$.

Dôkaz:

(i) Platí

$$\mathbf{Y}_n = \frac{1}{\sqrt{n}}\mathbf{D}^{-1}(\mathbf{X}_1 + \mathbf{X}_2 + ... + \mathbf{X}_n - n\boldsymbol{\theta}) = \frac{1}{\sqrt{n}}\mathbf{D}^{-1}(\mathbf{X}_1 - \boldsymbol{\theta} + \mathbf{X}_2 - \boldsymbol{\theta} + ... + \mathbf{X}_n - \boldsymbol{\theta}) =$$

$$= \frac{1}{\sqrt{n}}(\mathbf{D}^{-1}\mathbf{X}_1 - \mathbf{D}^{-1}\boldsymbol{\theta}) + \frac{1}{\sqrt{n}}(\mathbf{D}^{-1}\mathbf{X}_2 - \mathbf{D}^{-1}\boldsymbol{\theta}) + ... + \frac{1}{\sqrt{n}}(\mathbf{D}^{-1}\mathbf{X}_n - \mathbf{D}^{-1}\boldsymbol{\theta}).$$

Zrejme $\mathbf{D}^{-1}\mathbf{X}_1, \mathbf{D}^{-1}\mathbf{X}_2, ...$ sú nezávislé $k-$rozmerné náhodné vektory, rovnako rozdelené, so strednou hodnotou $\mathbf{D}^{-1}\boldsymbol{\theta}$ a kovariančnou maticou $\mathbf{D}^{-1}(cov\mathbf{X}_i)\mathbf{D}^{-1} = \mathbf{D}^{-1}\mathbf{DQDD}^{-1} = \mathbf{Q}$. Preto podľa mnohorozmernej centrálnej limitnej vety (Veta 2.4) platí

$$\mathbf{Y}_n = \frac{1}{\sqrt{n}}\mathbf{D}^{-1}(\mathbf{X}^{(n)} - n\boldsymbol{\theta}) + ... + \frac{1}{\sqrt{n}}(\mathbf{D}^{-1}\mathbf{X}_n - \mathbf{D}^{-1}\boldsymbol{\theta}) = \frac{1}{\sqrt{n}}\mathbf{D}^{-1}(\mathbf{X}^{(n)} - n\boldsymbol{\theta}) \xrightarrow[n]{\mathcal{D}} N_k(\mathbf{0}, \mathbf{Q}).$$

(ii) Funkcia $b: \mathcal{R}^k \to \mathcal{R}^1$ daná predpisom $b(\mathbf{z}) = \mathbf{z}'\mathbf{z}$ je spojitá, $\mathbf{Y}_n = \frac{1}{\sqrt{n}}\mathbf{D}^{-1}(\mathbf{X}^{(n)} - n\boldsymbol{\theta}) \xrightarrow[n]{\mathcal{D}} N_k(\mathbf{0}, \mathbf{Q})$ (podľa (i)), teda podľa mnohorozmernej Sverdrupovej vety (Veta 2.5)

$$b(\mathbf{Y}_n) = \mathbf{Y}_n'\mathbf{Y}_n = \frac{1}{n}\left(\mathbf{X}^{(n)} - n\boldsymbol{\theta}\right)'\mathbf{D}^{-1}\mathbf{D}^{-1}\left(\mathbf{X}^{(n)} - n\boldsymbol{\theta}\right) =$$

$$= \frac{1}{n}(X_1^{(n)} - n\theta_1, X_2^{(n)} - n\theta_2, ..., X_k^{(n)} - n\theta_k)\begin{pmatrix} \frac{1}{\sqrt{\theta_1}} & 0 & \dots & 0 \\ 0 & \frac{1}{\sqrt{\theta_2}} & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & & \dots & \frac{1}{\sqrt{\theta_k}} \end{pmatrix}\begin{pmatrix} \frac{1}{\sqrt{\theta_1}} & 0 & \dots & 0 \\ 0 & \frac{1}{\sqrt{\theta_2}} & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & & \dots & \frac{1}{\sqrt{\theta_k}} \end{pmatrix}\begin{pmatrix} X_1^{(n)} - n\theta_1 \\ X_2^{(n)} - n\theta_2 \\ \vdots \\ X_k^{(n)} - n\theta_k \end{pmatrix} =$$

$$= \frac{1}{n}\left(\frac{X_1^{(n)} - n\theta_1}{\sqrt{\theta_1}}, \frac{X_2^{(n)} - n\theta_2}{\sqrt{\theta_2}}, ..., \frac{X_k^{(n)} - n\theta_k}{\sqrt{\theta_k}}\right)\begin{pmatrix} \frac{X_1^{(n)} - n\theta_1}{\sqrt{\theta_1}} \\ \frac{X_2^{(n)} - n\theta_2}{\sqrt{\theta_2}} \\ \vdots \\ \frac{X_k^{(n)} - n\theta_k}{\sqrt{\theta_k}} \end{pmatrix} =$$

$$= \sum_{j=1}^{k} \frac{(X_j^{(n)} - n\theta_j)^2}{n\theta_j} \xrightarrow[n]{\mathcal{D}} \mathbf{Y}'\mathbf{Y},$$

kde $\mathbf{Y} \sim N_k(\mathbf{0}, \mathbf{Q})$. Pretože $\mathbf{Q}$ je idempotentná matica, je $\mathbf{I}$ jej jedna zovšeobecnená inverzia a $\mathbf{Y}'\mathbf{Y} = (\mathbf{Y}-\mathbf{0})'\mathbf{Q}^{-}(\mathbf{Y}-\mathbf{0}) \sim \chi^2_{h(\mathbf{Q})}$. Čiže

$$\sum_{j=1}^{k} \frac{(X_j^{(n)} - n\theta_j)^2}{n\theta_j} \xrightarrow[n]{\mathcal{D}} \chi^2_{k-1}. \qquad \text{Q.E.D.}$$

Poznámka 2.7:

(i)

$$_{(n)}\chi^2 = \sum_{j=1}^{k} \frac{(X_j^{(n)} - n\theta_j)^2}{n\theta_j} = \sum_{j=1}^{k} \frac{\left[X_j^{(n)}\right]^2}{n\theta_j} - 2\sum_{j=1}^{k} \frac{X_j^{(n)} n\theta_j}{n\theta_j} + \sum_{j=1}^{k} \frac{n^2\theta_j^2}{n\theta_j} =$$

$$= \sum_{j=1}^{k} \frac{\left[X_j^{(n)}\right]^2}{n\theta_j} - 2\sum_{j=1}^{k} X_j^{(n)} + n\sum_{j=1}^{k} \theta_j = \sum_{j=1}^{k} \frac{\left[X_j^{(n)}\right]^2}{n\theta_j} - n,$$

lebo $\sum_{j=1}^{k} X_j^{(n)} = n$ a $\sum_{j=1}^{k} \theta_j = 1$.

(ii) V praxi sa aproximácia χ^2 použije ak $n\theta_i > 5$ pre všetky $i = 1, 2, ..., k$.

(iii) Realizácie náhodných veličín $X_1^{(n)}, X_2^{(n)}, ..., X_j^{(n)}$ sa volajú *empirické četnosti* a $n\theta_1, n\theta_2, ..., n\theta_k$ sa volajú *teoretické četnosti.*

2.2 Testy dobrej zhody pri známych (niekedy rušivých) parametroch

Majme náhodný pokus, pri ktorom môže nastať k rôznych výsledkov $A_1, ..., A_k$, pričom pre $i \neq j$ je $A_i \cap A_j = \emptyset$ a $\cup_{i=1}^{k} A_i = \Omega$ (istá udalosť), $P(A_i) = \theta_i \in (0,1),\ i = 1, 2, ..., k,\ \ \sum_{i=1}^{k} P(A_i) = \sum_{i=1}^{k} \theta_i = 1$. Tento pokus nezávisle opakujeme $n-$krát a označme

$X_j^{(n)}$ - počet výskytov výsledku A_j (realizáciou tejto náhodnej veličiny je empirická četnosť A_j).

Zrejme $\mathbf{X}^{(n)} = (X_1^{(n)}, ..., X_k^{(n)})' \sim Mu(n, \theta_1, ..., \theta_k)$ (dokážte ako cvičenie).

Majme (hypotetické) hodnoty $\theta_{10}, \theta_{20}, ..., \theta_{k0},\ 0 < \theta_{i0} < 1,\ i = 1, 2, ..., k,\ \sum_{i=1}^{k} \theta i0 = 1$ a testujeme hypotézu

$$H_0: \quad \theta_1 = \theta_{10}, ..., \theta_k = \theta_{k0} \qquad \asymp \qquad H_1: \quad \text{neplatí } H_0$$

Za platnosti H_0 má testovacia štatistika $_{(n)}\chi^2 = \sum_{j=1}^{k} \dfrac{(X_j^{(n)} - n\theta_{j0})^2}{n\theta_{j0}} \sim \chi^2_{k-1}$ rozdelenie (asymptoticky). Ak realizácia tejto štatistiky je väčšia ako $\chi^2_{k-1}(1-\alpha)$ ($(1-\alpha)-$kvantil chi^2 rozdelenia s $k-1$ stupňami voľnosti), tak H_0 zamietame (na hladine významnosti α).

Príklad 2.8: Pri pokuse - hod mincou označme (výsledok) A_1-padne číslo a A_2-padne znak ($k = 2$). $100-$krát hodíme mincou ($n = 100$), pričom $51-$krát sa objavilo číslo a $49-$krát znak. Náhodná veličina $X_1^{(100)}$ je počet padnutí v císla a $X_2^{(100)}$ je počet padnutí znaku pri týchto 100 hodoch, $P(A_i) = \theta_i,\ i = 1, 2$. Testujeme

$$H_0: \quad \theta_1 = \frac{1}{2},\ \theta_2 = \frac{1}{2} \qquad \asymp \qquad H_1: \quad \text{neplatí } H_0$$

(teda testujeme hypotézu, že minca je homogénna, $\theta_{10} = \theta_{20} = \frac{1}{2}$).

Realizácia testovacej štatistiky ${}_{(100)}\chi^2$ je $\dfrac{(51-100\frac{1}{2})^2}{100\frac{1}{2}} + \dfrac{(49-100\frac{1}{2})^2}{100\frac{1}{2}} = \frac{1}{50} + \frac{1}{50} = 0,04$. Pretože $0,95-$kvantil χ_1^2 rozdelenia je $\chi_1^2(0,95) = 3,84$ a realizácia testovacej štatistiky je $0,04 < 3,84$, nezamietame H_0 (nezamietame hypotézu, že minca je homogénna) na hladine významnosti $\alpha = 0,05$.

Príklad 2.9: (Anděl, J., *Matematická statistika*, SNTL, Praha, 1985, str.195.) Chceme testovať hypotézu, že deti v Československu v roku 1957 sa rodili rovnomerne. Označme p_i pravdepodobnosť, že dieťa sa narodí v $i-$tom mesiaci (prirodzene pre i=1 je to "leden", pre $i = 2$ je to "únor", atď.) a náhodnú veličinu X_i-počet narodených detí v i-tom mesiaci. Test založíme na údajoch z nasledujúcej tabuľky (udáva počet narodených detí v Československu v roku 1957 v jednotlivých mesiacoch).

mesiac i	realizácia X_i	počet dní	p_i	np_i	$\frac{(X_i-np_i)^2}{np_i}$
1.	21 182	31	0,08493	21 465	3,731
2.	19 960	28	0,07673	19 393	16,578
3.	22 787	31	0,08493	21 465	81,420
4.	22 805	30	0,08219	20 773	198,769
5.	23 120	31	0,08493	21 465	127,604
6.	21 859	30	0,08219	20 773	56,775
7.	21 367	31	0,08493	21 465	0,447
8.	20 357	31	0,08493	21 465	57,194
9.	20 946	30	0,08219	20 773	1,441
10.	20 037	31	0,08493	21 465	95,000
11.	18 728	30	0,08219	20 773	201,320
12.	19 592	31	0,08493	21 465	163,435
Σ	252 740	365	1,00000	252 740	1 003,744

Keby bol počet narodených detí nezávislý na ročnej dobe, bola by pravdepodobnosť narodenia dieťaťa v danom mesiaci úmerná počtu dní v tomto mesiaci (napr. pre "leden" $p_1 = \frac{31}{365} = 0,08493$, pozri 4. stĺpec tabuľky). Vzhľadom k zaokrúhľovacím chybám sa upravovali tieto pravdepodobnosti tak, aby ich výsledný súčet bol 1. V tomto prípade n=252 740 a $k = 12$. Realizácia testovacej štatistiky ${}_{(252740)}\chi^2$ je 1003,744 a to je viac ako 0.95-kvantil χ_{11}^2 rozdelenia ($\chi_{11}^2(0,95) = 19,7$). Preto zamietame H_0, že sa deti rodili rovnomerne v Českoslvensku v priebehu roka 1957 (na hladine významnosti 0.05).

2.3 Testy dobrej zhody pri neznámych parametroch

Často sa stáva, že $\theta_1, \theta_2, ..., \theta_k$ multinomickéh rozdelenia nepoznáme, alebo sú funkciami ných parametrov $\alpha_1, \alpha_2, ..., \alpha_m, \quad (m < k-1)$. Ilustrujme to na nasledujúcej situácii:

Nech $Y_1, Y_2, ..., Y_n$ je náhodný výber z rozdelenia s distribučnou funkciou $F(x, \boldsymbol{\alpha})$ (teda závisí na $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_m)')$. Rozdeľme os x na k intervalov $I_1, I_2, ..., I_k$, aby $\cup_{j=1}^k I_j = (-\infty, \infty), \quad I_j \cap I_k = \emptyset$ pre $i \neq k$. Označme náhodnú veličinu

$$X_i^{(n)} \ - \ \text{počet realizácií z } \{y_1, y_2, ..., y_n\}, \text{ ktoré padnú do } I_i \text{ (empirická početnosť).}$$

Nech Y má distribučnú funkciu $F(x, \boldsymbol{\alpha})$,

$$\theta_i = P\{Y \in I_i\} = \int_{I_i} dF(x, \boldsymbol{\alpha}) = \theta_i(\boldsymbol{\alpha}), \quad i = 1, 2, ..., k.$$

V tomto prípade má $\mathbf{X}^{(n)} = \left(X_1^{(n)}, X_2^{(n)}, ..., X_k^{(n)}\right)' \sim Mu(n, \theta_1(\boldsymbol{\alpha}), \theta_2(\boldsymbol{\alpha}), ..., \theta_k(\boldsymbol{\alpha}))$ (presvedčte sa ako cvičenie), teda (pre dané $\boldsymbol{\alpha}$)

$$_{(n)}\chi^2(\boldsymbol{\alpha}) = \sum_{i=1}^{k} \frac{(X_i^{(n)} - n\theta_i(\boldsymbol{\alpha}))^2}{n\theta_i(\boldsymbol{\alpha})} \sim \chi^2_{k-1}$$

(asymptoticky, t.j. pre veľké n).

$$\boldsymbol{\alpha}^*_{(n)} = \boldsymbol{\alpha}^*_{(n)}(X_1^{(n)}, X_2^{(n)}, ..., X_k^{(n)}) = argmin\ _{(n)}\chi^2(\boldsymbol{\alpha})$$

nazývame odhadom $\boldsymbol{\alpha}$ *metódou minimálneho* χ^2.

Odhad $\boldsymbol{\alpha}^*_{(n)}$ rieši rovnice

$$\left.\frac{\partial\ _{(n)}\chi^2(\boldsymbol{\alpha})}{\partial \alpha_j}\right|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^*_{(n)}} = 0, \quad j = 1, 2, ..., m.$$

Počítajme

$$\frac{\partial\ _{(n)}\chi^2(\boldsymbol{\alpha})}{\partial \alpha_j} = \frac{\partial}{\partial \alpha_j} \sum_{i=1}^{k} \frac{(X_i^{(n)} - n\theta_i(\boldsymbol{\alpha}))^2}{n\theta_i(\boldsymbol{\alpha})} =$$

$$= \sum_{i=1}^{k} \frac{-2(X_i^{(n)} - n\theta_i(\boldsymbol{\alpha}))n\frac{\partial\theta_i(\boldsymbol{\alpha})}{\partial\alpha_j}n\theta_i(\boldsymbol{\alpha}) - (X_i^{(n)} - n\theta_i(\boldsymbol{\alpha}))^2 n\frac{\partial\theta_i(\boldsymbol{\alpha})}{\partial\alpha_j}}{n^2\theta_i^2(\boldsymbol{\alpha})} =$$

$$= \sum_{i=1}^{k} \left\{ -2\frac{X_i^{(n)} - n\theta_i(\boldsymbol{\alpha})}{\theta_i(\boldsymbol{\alpha})} - \frac{(X_i^{(n)} - n\theta_i(\boldsymbol{\alpha}))^2}{n\theta_i^2(\boldsymbol{\alpha})} \right\} \frac{\partial\theta_i(\boldsymbol{\alpha})}{\partial\alpha_j} = 0, \quad j = 1, 2, ..., m,$$

čiže

$$\sum_{i=1}^{k} \left\{ \frac{X_i^{(n)} - n\theta_i(\boldsymbol{\alpha})}{\theta_i(\boldsymbol{\alpha})} + \frac{(X_i^{(n)} - n\theta_i(\boldsymbol{\alpha}))^2}{2n\theta_i^2(\boldsymbol{\alpha})} \right\} \frac{\partial\theta_i(\boldsymbol{\alpha})}{\partial\alpha_j} = 0, \quad j = 1, 2, ..., m.$$

Ukazuje sa, že vplyv člena $\dfrac{(X_i^{(n)} - n\theta_i(\boldsymbol{\alpha}))^2}{2n\theta_i^2(\boldsymbol{\alpha})}$ pri dostatočne veľkom n nie je podstatný. Zanedbáme tento člen a dostávame sústavu rovníc

$$\sum_{i=1}^{k} \frac{X_i^{(n)} - n\theta_i(\boldsymbol{\alpha})}{\theta_i(\boldsymbol{\alpha})} \frac{\partial\theta_i(\boldsymbol{\alpha})}{\partial\alpha_j} = 0, \quad j = 1, 2, ..., m.$$

Pretože

$$\sum_{i=1}^{k} \frac{n\theta_i(\boldsymbol{\alpha})}{\theta_i(\boldsymbol{\alpha})} \frac{\partial\theta_i(\boldsymbol{\alpha})}{\partial\alpha_j} = n\sum_{i=1}^{k} \frac{\partial\theta_i(\boldsymbol{\alpha})}{\partial\alpha_j} = n\frac{\partial}{\partial\alpha_j}\sum_{i=1}^{k}\theta_i(\boldsymbol{\alpha}) = 0,$$

konečne dostávame rovnice, ktorých riešenie $\widehat{\boldsymbol{\alpha}}_{(n)} = \widehat{\boldsymbol{\alpha}}_{(n)}(X_1^{(n)}, X_2^{(n)}, ..., X_k^{(n)})$ je odhadom $\boldsymbol{\alpha}$ *metódou modifikovaného minimálneho* χ^2. Ich tvar je

$$\sum_{i=1}^{k} \left[\frac{X_i^{(n)}}{\theta_i(\boldsymbol{\alpha})} \frac{\partial\theta_i(\boldsymbol{\alpha})}{\partial\alpha_j} \right]_{\boldsymbol{\alpha}=\widehat{\boldsymbol{\alpha}}_{(n)}} = 0, \quad j = 1, 2, ..., m.$$

Veta 2.10: Nech $\theta_1(\boldsymbol{\alpha}), \theta_2(\boldsymbol{\alpha}), ..., \theta_k(\boldsymbol{\alpha})$ sú funkcie parametra $\boldsymbol{\alpha} \in \mathcal{R}^m$, $m < k-1$ a nech pre všetky body $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_m)'$ nedegenerovaného konečného uzavretého intervalu $A \subset \mathcal{R}^m$ platí

1. $\theta_1(\boldsymbol{\alpha}) + ... + \theta_k(\boldsymbol{\alpha}) = 1$,
2. $\exists c > 0$, že $\theta_i(\boldsymbol{\alpha}) > c, \quad i = 1, 2, ..., k$,
3. pre každé $i \in \{1, 2, ..., k\}$ existujú spojité derivácie $\dfrac{\partial \theta_i(\boldsymbol{\alpha})}{\partial \alpha_j}$, $j \in \{1, 2, ..., m\}$ a tiež $\dfrac{\partial^2 \theta_i(\boldsymbol{\alpha})}{\partial \alpha_j \partial \alpha_l}$, $j, l \in \{1, 2, ..., m\}$,
4. matica

$$\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\alpha}'} = \begin{pmatrix} \frac{\partial \theta_1(\boldsymbol{\alpha})}{\partial \alpha_1} & \cdots & \frac{\partial \theta_1(\boldsymbol{\alpha})}{\partial \alpha_m} \\ \frac{\partial \theta_2(\boldsymbol{\alpha})}{\partial \alpha_1} & \cdots & \frac{\partial \theta_2(\boldsymbol{\alpha})}{\partial \alpha_m} \\ \vdots & \vdots & \vdots \\ \frac{\partial \theta_k(\boldsymbol{\alpha})}{\partial \alpha_1} & \cdots & \frac{\partial \theta_k(\boldsymbol{\alpha})}{\partial \alpha_m} \end{pmatrix}$$

má hodnosť m.
Nech $\boldsymbol{\alpha}^\circ$ je vnútorným bodom A. Označme $\theta_{i0} = \theta_i(\boldsymbol{\alpha}^\circ)$. Nech $\mathbf{X}^{(n)} = \left(X_1^{(n)}, X_2^{(n)}, ..., X_k^{(n)}\right)' \sim Mu(n, \theta_{10}, \theta_{20}, ..., \theta_{k0})$. Potom sústava

$$\sum_{i=1}^{k} \left[\frac{X_i^{(n)}}{\theta_i(\boldsymbol{\alpha})} \frac{\partial \theta_i(\boldsymbol{\alpha})}{\partial \alpha_j} \right]_{\boldsymbol{\alpha} = \widehat{\boldsymbol{\alpha}}_{(n)}} = 0, \quad j = 1, 2, ..., m$$

má práve jedno riešenie $\widehat{\boldsymbol{\alpha}}_{(n)} = (\hat{\alpha}_1^{(n)}, \hat{\alpha}_2^{(n)}, ..., \hat{\alpha}_m^{(n)})'$, ktoré je konzistentným odhadom $\boldsymbol{\alpha}^\circ$ (teda $\widehat{\boldsymbol{\alpha}}_{(n)} \xrightarrow[n]{P} \boldsymbol{\alpha}^\circ$) a platí (asymptoticky)

$${}_{(n)}\chi^2 = \sum_{i=1}^{k} \frac{(X_i^{(n)} - n\theta_i(\widehat{\boldsymbol{\alpha}}_{(n)})^2)}{n\theta_i(\widehat{\boldsymbol{\alpha}}_{(n)})} \sim \chi^2_{k-m-1}.$$

Dôkaz: pozri napríklad v knižke Anděl, J., *Matematická statistika*, SNTL, Praha, 1985, str. 197. Poznamenávame len, že $\widehat{\boldsymbol{\alpha}}_{(n)} \xrightarrow[n]{P} \boldsymbol{\alpha}^\circ \iff \forall \varepsilon > 0 \; P\{\omega : \; ||\widehat{\boldsymbol{\alpha}}_{(n)}(\omega) - \boldsymbol{\alpha}^\circ||^2 > \varepsilon\} \xrightarrow[n]{} 0$.

Poznámka 2.11: Veta 2.10 hovorí, ako dostať konzistentný odhad $\boldsymbol{\alpha}^\circ$ a ako testovať, či náš model "je dobrý", t.j. napr. či údaje pochádzajú z náhodného výberu s daným rozdelením pravdepodobnosti (typom rozdelenia), ktorého distribučná funkcia je $F(x, \boldsymbol{\alpha})$.

2.4 Overenie normálneho rozdelenia

Majme náhodný výber $Y_1, Y_2, ..., Y_n$ z rozdelenia s (nejakou) distribučnou funkciou $F(y)$. Testujeme hypotézu

$$H_0: \text{ náhodný výber je z normálneho rozdelenia} \qquad \asymp \qquad H_1: \text{ neplatí } H_0.$$

Reálnu os rozdelíme na k disjunktných intervalov

$$I_1 = (-\infty, b_1\rangle, \; I_2 = (b_1, b_2\rangle, \; ... \; I_{k-1} = (b_{k-2}, b_{k-1}\rangle, \; I_k = (b_{k-1}, \infty)$$

($b_0 = -\infty < b_1 < b_2 < ... < b_{k-1} < b_k = \infty$, $b_1, b_2, ..., b_{k-1}$ sú vhodne zvolené reálne čísla, o ich voľbe si povieme v Poznámke 2.12). Označme náhodná veličinu $X_i^{(n)}$

$X_i^{(n)}$ – počet realizácií z množiny $\{y_1, y_2, ..., y_n\}$, ktoré padli do I_i, $\quad i = 1, 2, ..., k$,
$\boldsymbol{\alpha} = (\mu, \sigma)'$, kde $\mu \in (-\infty, \infty)$, $\sigma > 0$. Za platnosti H_0 je

$$\theta_i(\boldsymbol{\alpha}) = \theta_i(\mu, \sigma) = P\{Y_j \in I_i\} = P\{b_{i-1} < Y_j < b_i\} = \int_{b_{i-1}}^{b_i} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\dfrac{(x-\mu)^2}{2\sigma^2}} dx, \quad i = 1, 2, ..., k$$

(j je ľubovoľné z $\{1, 2, ..., n\}$). Teraz určíme modifikovaný minimálny χ^2 odhad $\hat{\boldsymbol{\alpha}}_{(n)}$ ako riešenie rovníc

$$\sum_{i=1}^{k} \frac{X_i^{(n)}}{\theta_i(\mu,\sigma)} \frac{\partial \theta_i(\mu,\sigma)}{\partial \mu} = 0,$$

$$\sum_{i=1}^{k} \frac{X_i^{(n)}}{\theta_i(\mu,\sigma)} \frac{\partial \theta_i(\mu,\sigma)}{\partial \sigma} = 0,$$

pričom

$$\frac{\partial \theta_i}{\partial \mu} = \frac{\partial}{\partial \mu} \int_{b_{i-1}}^{b_i} \frac{1}{\sqrt{2\pi}\ \sigma} e^{-\dfrac{(x-\mu)^2}{2\sigma^2}} dx = \frac{1}{\sigma^2} \int_{b_{i-1}}^{b_i} \frac{x-\mu}{\sqrt{2\pi}\ \sigma} e^{-\dfrac{(x-\mu)^2}{2\sigma^2}} dx$$

a

$$\frac{\partial \theta_i}{\partial \sigma} = \frac{\partial}{\partial \sigma} \int_{b_{i-1}}^{b_i} \frac{1}{\sqrt{2\pi}\ \sigma} e^{-\dfrac{(x-\mu)^2}{2\sigma^2}} dx = \int_{b_{i-1}}^{b_i} \left(-\frac{1}{\sqrt{2\pi}\ \sigma^2} e^{-\dfrac{(x-\mu)^2}{2\sigma^2}} + \frac{1}{\sqrt{2\pi}\ \sigma} e^{-\dfrac{(x-\mu)^2}{2\sigma^2}} \frac{(x-\mu)^2}{2\sigma^3} \right) dx =$$

$$= \frac{1}{\sigma^3} \int_{b_{i-1}}^{b_i} \frac{(x-\mu)^2}{\sqrt{2\pi}\ \sigma} e^{-\dfrac{(x-\mu)^2}{2\sigma^2}} dx - \frac{1}{\sigma} \int_{b_{i-1}}^{b_i} \frac{1}{\sqrt{2\pi}\ \sigma} e^{-\dfrac{(x-\mu)^2}{2\sigma^2}} dx.$$

Označme $f(x;\mu,\sigma)$ hustotu rozdelenia $N(\mu,\sigma^2)$. Potom rovnice pre určenie modifikovaného χ^2 odhadu $\hat{\mu}_{(n)}, \hat{\sigma}_{(n)}$ sú

$$\sum_{i=1}^{k} \frac{X_i^{(n)}}{\theta_i(\hat{\mu}_{(n)},\hat{\sigma}_{(n)})} \frac{1}{(\hat{\sigma}_{(n)})^2} \int_{I_i} (x-\hat{\mu}_{(n)}) f(x;\hat{\mu}_{(n)},\hat{\sigma}_{(n)}) dx = 0, \tag{2}$$

$$\sum_{i=1}^{k} \frac{X_i^{(n)}}{\theta_i(\hat{\mu}_{(n)},\hat{\sigma}_{(n)})} \left(\frac{1}{(\hat{\sigma}_{(n)})^3} \int_{I_i} (x-\hat{\mu}_{(n)})^2 f(x;\hat{\mu}_{(n)},\hat{\sigma}_{(n)}) dx - \frac{1}{(\hat{\sigma}_{(n)})} \int_{I_i} f(x;\hat{\mu}_{(n)},\hat{\sigma}_{(n)}) dx \right) = 0. \tag{3}$$

Rovnica (2) sa dá upraviť

$$\sum_{i=1}^{k} \frac{X_i^{(n)}}{\theta_i(\hat{\mu}_{(n)},\hat{\sigma}_{(n)})} \left[\int_{I_i} x f(x;\hat{\mu}_{(n)},\hat{\sigma}_{(n)}) dx - \hat{\mu}_{(n)} \theta_i(\hat{\mu}_{(n)},\hat{\sigma}_{(n)}) \right] = 0,$$

teda

$$\sum_{i=1}^{k} \frac{X_i^{(n)}}{\theta_i(\hat{\mu}_{(n)},\hat{\sigma}_{(n)})} \int_{I_i} x f(x;\hat{\mu}_{(n)},\hat{\sigma}_{(n)}) dx = \hat{\mu}_{(n)} \sum_{i=1}^{k} X_i^{(n)}$$

a keďže $\sum_{i=1}^{k} X_i^{(n)} = n$, dostávame

$$\hat{\mu}_{(n)} = \frac{1}{n} \sum_{i=1}^{k} \frac{X_i^{(n)}}{\theta_i(\hat{\mu}_{(n)},\hat{\sigma}_{(n)})} \int_{I_i} x f(x;\hat{\mu}_{(n)},\hat{\sigma}_{(n)}) dx. \tag{4}$$

Z rovnice (3) zase dostávame

$$\sum_{i=1}^{k} \frac{X_i^{(n)}}{\theta_i(\hat{\mu}_{(n)},\hat{\sigma}_{(n)})} \int_{I_i} (x-\hat{\mu}_{(n)})^2 f(x;\hat{\mu}_{(n)},\hat{\sigma}_{(n)}) dx = (\hat{\sigma}_{(n)})^2 \sum_{i=1}^{k} \frac{X_i^{(n)}}{\theta_i(\hat{\mu}_{(n)},\hat{\sigma}_{(n)})} \theta_i(\hat{\mu}_{(n)},\hat{\sigma}_{(n)}),$$

čiže

$$(\hat{\sigma}_{(n)})^2 = \frac{1}{n} \sum_{i=1}^{k} \frac{X_i^{(n)}}{\theta_i(\hat{\mu}_{(n)},\hat{\sigma}_{(n)})} \int_{I_i} (x-\hat{\mu}_{(n)})^2 f(x;\hat{\mu}_{(n)},\hat{\sigma}_{(n)}) dx. \tag{5}$$

Rovnice (4) a (5) sa riešia (pre dané n) iteračne. Určíme čísla

$$c_2 = \frac{b_2 + b_1}{2},\ c_3 = \frac{b_3 + b_2}{2},\ ...\ c_{k-1} = \frac{b_{k-1} + b_{k-2}}{2},$$

$$c_1 = b_1 - (c_2 - b_1) = \frac{3b_1 - b_2}{2},\ c_k = b_{k-1} + (b_{k-1} - c_{k-1}) = \frac{3b_{k-1} - b_{k-2}}{2}$$

a nultý odhad

$${}_0\hat{\mu}_{(n)} = \frac{1}{n}\sum_{i=1}^{k} X_i^{(n)} c_i$$

$$({}_0\hat{\sigma}_{(n)})^2 = \frac{1}{n}\sum_{i=1}^{k} X_i^{(n)} c_i^2 - ({}_0\hat{\mu}_{(n)})^2$$

(pozor, neodporúča sa použiť $\overline{Y}$ namiesto ${}_0\hat{\mu}_{(n)}$ a $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$ namiesto $({}_0\hat{\sigma}_{(n)})^2$). Teraz sa spočíta zo vzťahu (4)

$${}_1\hat{\mu}_{(n)} = \frac{1}{n}\sum_{i=1}^{k} \frac{X_i^{(n)}}{\theta_i({}_0\hat{\mu}_{(n)}, {}_0\hat{\sigma}_{(n)})} \int_{I_i} x f(x; {}_0\hat{\mu}_{(n)}, {}_0\hat{\sigma}_{(n)}) dx.$$

a zo vzťahu (5)

$$({}_1\hat{\sigma}_{(n)})^2 = \frac{1}{n}\sum_{i=1}^{k} \frac{X_i^{(n)}}{\theta_i({}_0\hat{\mu}_{(n)}, {}_0\hat{\sigma}_{(n)})} \int_{I_i} (x - {}_0\hat{\mu}_{(n)})^2 f(x; {}_0\hat{\mu}_{(n)}, {}_0\hat{\sigma}_{(n)}) dx.$$

Znovu spočítame

$${}_2\hat{\mu}_{(n)} = \frac{1}{n}\sum_{i=1}^{k} \frac{X_i^{(n)}}{\theta_i({}_1\hat{\mu}_{(n)}, {}_1\hat{\sigma}_{(n)})} \int_{I_i} x f(x; {}_1\hat{\mu}_{(n)}, {}_1\hat{\sigma}_{(n)}) dx.$$

a

$$({}_2\hat{\sigma}_{(n)})^2 = \frac{1}{n}\sum_{i=1}^{k} \frac{X_i^{(n)}}{\theta_i({}_1\hat{\mu}_{(n)}, {}_1\hat{\sigma}_{(n)})} \int_{I_i} (x - {}_1\hat{\mu}_{(n)})^2 f(x; {}_1\hat{\mu}_{(n)}, {}_1\hat{\sigma}_{(n)}) dx,$$

a tento iteračný proces opakujeme, kým nedostaneme výsledné odhady $\hat{\mu}_{(n)}, \hat{\sigma}_{(n)}$. Teraz spočítame realizáciu testovacej štatistiky

$${}_{(n)}\chi^2 = \sum_{i=1}^{k} \frac{(X_i^{(n)} - n\theta_i(\hat{\mu}_{(n)}, \hat{\sigma}_{(n)}))^2}{n\theta_i(\hat{\mu}_{(n)}, \hat{\sigma}_{(n)})}.$$

Ak realizácia tejto štatistiky ${}_{(n)}\chi^2_{(real)} \geq \chi^2_{k-3}(1-\alpha)$, tak H_0: *výber pochádza z normálneho rozdelenia* zamietame na hladine významnosti α. Poznamenávame len, že $\chi^2_{k-3}(1-\alpha)$ je $(1-\alpha)-$kvantil χ^2_{k-3} rozdelenia, m použité vo Vete 2.10 je v tomto prípade rovné 2.

Poznámka 2.12: Čísla $b_1, b_2, ..., b_{k-1}$ treba voliť tak, aby pre každé i platilo $n\theta_i > 5$, teda aby v každom intervale I_i bolo aspoň 5 realizácií z $\{y_1, y_2, ..., y_n\}$.

Ako sa overuje, či náhodný výber pochádza z exponenciálneho rozdelenia alebo z Poissonovho rozdelenia pozri v knižke Anděl, J., *Matematická statistika*, SNTL, Praha, 1985, na stranách 201, 207.

Pretože testovanie normality je veľmi dôležité, ukážeme si ešte iný spôsob testovania, či náhodný výber pochádza z normálneho rozdelenia. Definujme si *výberový obecný moment $k-$teho rádu*

$$M'_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k, \quad k = 1, 2, 3, ...,$$

keď $X_1, X_2, ..., X_n$ je náhodný výber. (Poznamenávame len, že $x_1, x_2, ..., x_n$ je realizácia náhodného výberu a $\frac{1}{n}\sum_{i=1}^n x_i$ je realizácia M'_k.) *Výberový centrálny moment $k-$teho rádu* je

$$M_k = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^k, \quad k = 1, 2, ...,$$

výberová šikmosť je

$$A_3 = \frac{M_3}{\sqrt{M_2^3}}$$

a *výberová špicatosť* je

$$A_4 = \frac{M_4}{M_2^2}.$$

Tieto posledné dve náhodné veličiny sú výberovými "protajškami" parametrov šikmosti $\frac{\mu_3}{\sqrt{\mu_2^3}}$ a špicatosti $\frac{\mu_4}{\sqrt{\mu_2^4}}$ (μ_i je (teoretický) centrálny moment $i-$teho rádu uvažovaného rozdelenia). U normálneho rozdelenia je šikmosť rovná 0 a špicatosť rovná 3. Preto v prípade, že náhodný výber pochadza z normálneho rozdelenia, pre veľké n by malo platiť $A_3 \approx 0$ a $A_4 \approx 3$. Dá sa ukázať, že ak náhodný výber pochádza z normálneho rozdelenia $N(\mu, \sigma^2)$, tak

$$\mathcal{E}(A_3) = 0 \qquad \mathcal{D}(A_3) = \frac{6(n-2)}{(n+1)(n+3)}$$

$$\mathcal{E}(A_4) = 3 - \frac{6}{n+1} \qquad \mathcal{D}(A_4) = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}$$

a tiež, že pre $n \to \infty$ majú $\sqrt{n}A_3$ a $\sqrt{n}A_4$ asymptoticky normálne rozdelenie, čiže

$$A_3 \approx N\left(0, \frac{6(n-2)}{(n+1)(n+3)}\right), \quad A_4 \approx N\left(3 - \frac{6}{n+1}, \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}\right).$$

Teda

$$\frac{A_3}{\sqrt{\mathcal{D}(A_3)}} = \frac{A_3\sqrt{(n+1)(n+3)}}{\sqrt{6(n-2)}} \sim N(0,1), \qquad \frac{A_4 - 3 + \frac{6}{n+1}}{\sqrt{\frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}}} \sim N(0,1).$$

Ak realiácia náhodnej veličiny (testovacej štatistiky) $\frac{|A_3|}{\sqrt{\mathcal{D}(A_3)}}$ je väčšia alebo rovná $u(\frac{\alpha}{2})$ ($\frac{\alpha}{2}-$kritická hodnota $N(0,1)$ rozdelenia), tak na hladine významnosti α zamietame hypotézu H_0: *výber pochádza z normálneho rozdelenia.* Toto je *test normality založený na šikmosti.*

Ak realiácia náhodnej veličiny (testovacej štatistiky) $\frac{|A_4 - \mathcal{E}(A_4)|}{\sqrt{\mathcal{D}(A_4)}}$ je väčšia alebo rovná $u(\frac{\alpha}{2})$, tak na hladine významnosti α zamietame hypotézu H_0: *výber pochádza z normálneho rozdelenia.* Toto je *test normality založený na špicatosti.*

Poznámka 2.13: Pre malé výbery $n \in \langle 7, 30\rangle$ sa odporúča Shapirov-Wilkov test normality (pozri napr. Kubáčková, L., *Metódy spracovania experimetálnych údajov*, VEDA, Bratislava, 1979). Pre $n \in \langle 31, 50\rangle$ sa odporúča D'Agostiniho test. Pre $n > 50$ sa už odporúča Pearsonov χ^2 test dobrej zhody.

2.5 Overenie Poissonovho rozdelenia

Teraz si overíme, či výber pochádza z Poissonovho rozdelenia iným spôsobom než Pearsonovym χ^2 testom dobrej zhody. Pri odvodzovaní budeme potrebovať tzv. polynomickú vetu.

Veta 2.14: (Polynomická veta.) Nech $a_1, ..., a_n$ sú reálne čísla, j celé nezáporné číslo. Platí

$$(a_1 + ... + a_n)^j = \sum_{\{\nu_1,\nu_2,...,\nu_n \in \{0,1,...,j\}:\ \sum_{i=1}^n \nu_i = j\}} \frac{j!}{\nu_1!\nu_2!...\nu_n!}\ a_1^{\nu_1} a_2^{\nu_2} ... a_n^{\nu_n}.$$

Dôkaz: Keď roznásobíme $(a_1 + ... + a_n)^j = \overbrace{(a_1 + ... + a_n)(a_1 + ... + a_n)...(a_1 + ... + a_n)}^{j-\text{krát}}$, dostaneme súčet súčinov $a_1^{\nu_1} a_2^{\nu_2} ... a_n^{\nu_n}$, kde $\nu_1, ..., \nu_n$ sú nezáporné celé čísla a $\nu_1 + \nu_2 + ... + \nu_n = j$. Vlastne $\nu_1, \nu_2, ..., \nu_n \in \{0, 1, 2, ..., j\}$. Pre jednu (ľubovoľnú vhodnú) $n-$ticu $\nu_1, \nu_2, ..., \nu_n$ dostaneme práve $\dfrac{j!}{\nu_1!\nu_2!...\nu_n!}$ súčinov $a_1^{\nu_1} a_2^{\nu_2} ... a_n^{\nu_n}$, lebo

a_1 "vyberieme" z $\binom{j}{\nu_1}$ "zátvoriek",

a_2 "vyberieme" z $\binom{j-\nu_1}{\nu_2}$ "zátvoriek",

$\vdots$

a_{n-1} "vyberieme" z $\binom{j-\nu_1-...-\nu_{n-2}}{\nu_{n-1}}$ "zátvoriek",

a_n "vyberieme" už len z tých zátvoriek, ktoré "zostali".

Takto pre (ľubovoľnú vhodnú) $n-$ticu $\nu_1, \nu_2, ..., \nu_n$ dostaneme

$$\binom{j}{\nu_1}\binom{j-\nu_1}{\nu_2}...\binom{j-\nu_1-...-\nu_{n-2}}{\nu_{n-1}} = \frac{j!}{\nu_1!(j-\nu_1)!}\ \frac{(j-\nu_1)!}{\nu_2!(j-\nu_1-\nu_2)!} ... \frac{(j-\nu_1-...-\nu_{n-2})!}{\nu_{n-1}!\nu_n!} = \frac{j!}{\nu_1!\nu_2!...\nu_n!}$$

súčinov $a_1^{\nu_1} a_2^{\nu_2} ... a_n^{\nu_n}$. Q.E.D.

Ak náhodná veličina $X \sim Po(\lambda)$, tak jej pravdepodobnostná funkcia je $P\{X = i\} = e^{-\lambda}\dfrac{\lambda^i}{i!}$, $i = 0, 1, ...,\quad \lambda > 0,\quad \mathcal{E}(X) = \lambda,\quad \mathcal{D}(X) = \lambda$. Nech $X_1, X_2, ..., X_n$ je náhodný výber, pričom $X_i \sim Po(\lambda)$. Združené rozdelenie $X_1, X_2, ..., X_n$ má pravdepodobnostnú funkciu

$$P\{X_1 = x_1, ..., X_n = x_n\} = e^{-n\lambda}\ \frac{\lambda^{\sum_{i=1}^n x_i}}{x_1!x_2!...x_n!}$$

ak $x_1, x_2, ..., x_n \in \{0, 1, ...\}$ (inak $P\{X_1 = x_1, ..., X_n = x_n\} = 0$).

Pre dané (fixné) nezáporné celé číslo t má podmienené rozdelenie $X_1, X_2, ..., X_n / \sum_{i=1}^n X_i = t$ pravdepodobnostnú funkciu

$$P\{X_1 = x_1, X_2 = x_2, ..., X_n = x_n / \sum_{i=1}^n X_i = t\} =$$

$$= \begin{cases} \dfrac{P\{X_1 = x_1, X_2 = x_2, ..., X_n = x_n, \sum_{i=1}^n X_i = t\}}{P\{\sum_{i=1}^n X_i = t\}} & \text{ak } P\{\sum_{i=1}^n X_i = t\} \neq 0, \\ 0 & \text{v inom prípade.} \end{cases} \tag{6}$$

Kedy $P\{\sum_{i=1}^n X_i = t\} \neq 0$? V prípade, že náhodný výber je z $Po(\lambda)$ rozdelenia, je $P\{\sum_{i=1}^n X_i = t\} \neq 0$ práve vtedy ak $x_1, x_2, ..., x_n \in \{0, 1, ..., t\}$ a súčasne $\sum_{i=1}^n x_i = t$. Preto

$$P\{X_1 = x_1, X_2 = x_2, ..., X_n = x_n / \sum_{i=1}^n X_i = t\} =$$

$$= \begin{cases} \dfrac{P\{X_1 = x_1, X_2 = x_2, ..., X_n = x_n, \sum_{i=1}^n X_i = t\}}{P\{\sum_{i=1}^n X_i = t\}} & \text{ak } \ x_1, ..., x_n \in \{0, 1, ..., t\}, \ \sum_{i=1}^n x_i = t, \\ 0 & \text{v inom prípade.} \end{cases} \tag{7}$$

Pre $n-$ticu $x_1, x_2, ..., x_n$ nezáporných celých čísel, pre ktoré platí $\sum_{i=1}^n x_i = t$ je

$$P\{X_1 = x_1, X_2 = x_2, ..., X_n = x_n, \sum_{i=1}^{n} X_i = t\} = P\{X_1 = x_1, X_2 = x_2, ..., X_n = x_n\} =$$

$$= e^{-\lambda}\frac{\lambda^{x_1}}{x_1!}e^{-\lambda}\frac{\lambda^{x_2}}{x_2!}...e^{-\lambda}\frac{\lambda^{x_n}}{x_n!} = e^{-n\lambda}\frac{\lambda^{\sum_{i=1}^n x_i = t}}{x_1!...x_n!} = e^{-n\lambda}\frac{\lambda^t}{x_1!...x_n!} \tag{8}$$

a

$$P\{\sum_{i=1}^{n} X_i = t\} = \sum_{\{x_1,x_2,...,x_n \in \{0,1,...,t\}: \ \sum_{i=1}^n x_i = t\}} P\{X_1 = x_1, ..., X_n = x_n\} =$$

$$= \sum_{\{x_1,x_2,...,x_n \in \{0,1,...,t\}: \ \sum_{i=1}^n x_i = t\}} e^{-\lambda}\frac{\lambda^{x_1}}{x_1!}e^{-\lambda}\frac{\lambda^{x_2}}{x_2!}...e^{-\lambda}\frac{\lambda^{x_n}}{x_n!} =$$

$$= \sum_{\{x_1,x_2,...,x_n \in \{0,1,...,t\}: \ \sum_{i=1}^n x_i = t\}} e^{-n\lambda}\frac{\lambda^t}{x_1!...x_n!} = e^{-n\lambda} \sum_{\{x_1,x_2,...,x_n \in \{0,1,...,t\}: \ \sum_{i=1}^n x_i = t\}} \frac{1}{x_1!...x_n!}. \tag{9}$$

Ak vo Vete 2.14 zvolíme $a_1 = a_2 = ... = a_n = 1$, dostávame

$$\sum_{\{x_1,x_2,...,x_n \in \{0,1,...,t\}: \ \sum_{i=1}^n x_i = t\}} \frac{1}{x_1!...x_n!} = \frac{n^t}{t!}, \tag{10}$$

teda dosadením (10) do (9) dostávame

$$P\{\sum_{i=1}^{n} X_i = t\} = \sum_{\{x_1,x_2,...,x_n \in \{0,1,...,t\}: \ \sum_{i=1}^n x_i = t\}} P\{X_1 = x_1, ..., X_n = x_n\} = e^{-n\lambda}\lambda^t\frac{n^t}{t!}. \tag{11}$$

Ak (8) a (11) dosadíme do (7) dostávame

$$P\{X_1 = x_1, X_2 = x_2, ..., X_n = x_n / \sum_{i=1}^{n} X_i = t\} =$$

$$= \begin{cases} \dfrac{e^{-n\lambda}\dfrac{\lambda^t}{x_1!...x_n!}}{e^{-n\lambda}\lambda^t\dfrac{n^t}{t!}} = \dfrac{t!}{x_1!...x_n!}\left(\dfrac{1}{n}\right)^t & \text{ak } \ x_1, ..., x_n \in \{0, 1, ..., t\}, \ \sum_{i=1}^n x_i = t, \\ 0 & \text{v inom prípade.} \end{cases} \tag{12}$$

Pretože $t = x_1 + x_2 + ... + x_n$, konečne dostávame

$$P\{X_1 = x_1, X_2 = x_2, ..., X_n = x_n / \sum_{i=1}^{n} X_i = t\} =$$

$$= \begin{cases} \dfrac{t!}{x_1!...x_n!}\left(\dfrac{1}{n}\right)^{x_1}...\left(\dfrac{1}{n}\right)^{x_n} & \text{ak } \ x_1, ..., x_n \in \{0, 1, ..., t\}, \ \sum_{i=1}^n x_i = t, \\ 0 & \text{v inom prípade.} \end{cases} \tag{13}$$

Vidíme, že za platnosti $H_0 :$ $X_1, X_2, ..., X_n$ *je náhodný výber z* $Po(\lambda)$ *rozdelenia* má $X_1, X_2, ..., X_n / \sum_{i=1}^n X_i = t$ multinomické $Mu(t, \underbrace{\frac{1}{n}, \frac{1}{n}, ..., \frac{1}{n}}_{n-\text{krát}})$ rozdelenie.

Zrealizujeme náhodný výber a zistíme, že $x_1 + x_2 + ... + x_n = t$ (tentokrát n je pevné číslo - rozsah výberu). Vieme, že ak $\sum_{i=1}^n X_i = t$, tak $X_1, X_2, ..., X_n$ majú multinomické rozdelenie $Mu(t, \frac{1}{n}, \frac{1}{n}, ..., \frac{1}{n})$. Ak t je dostatočne veľké, má

$$\sum_{i=1}^{n} \frac{\left(X_i - t\frac{1}{n}\right)^2}{t\frac{1}{n}} \approx \chi^2_{n-1} \text{ rozdelenie.}$$

V skutočnosti vlastne

$$\sum_{i=1}^{n} \frac{\left(X_i - \frac{1}{n}\sum_{j=1}^n X_j\right)^2}{\frac{1}{n}\sum_{j=1}^n X_j} = \sum_{i=1}^{n} \frac{\left(X_i - \overline{X}\right)^2}{\overline{X}} = \frac{nM_2}{M_1} = Q \approx \chi^2_{n-1}.$$

Ak realizácia $Q_{(real)} \geq \chi^2_{n-1}(1 - \frac{\alpha}{2})$ alebo $Q_{(real)} \leq \chi^2_{n-1}(\frac{\alpha}{2})$, tak na hladine významnosti α zamietame $H_0:\quad X_1, X_2, ..., X_n$ *je náhodný výber z* $Po(\lambda)$ *rozdelenia* ($\chi^2_{n-1}(1-\frac{\alpha}{2})$ je $(1-\frac{\alpha}{2})-$kvantil χ^2 rozdelenia s $n-1$ stupňami voľnosti). Upozorňujeme len, že aproximácia χ^2 rozdelením je možná len ak $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i > 5$.

2.6 Test nezávislosti v kontingenčných tabuľkách

Nech X, Y sú diskrétne náhodnéveličiny, $X \in \{1, 2, ..., r\}$, $Y \in \{1, 2, ..., s\}$, $p_{ij} = P\{X = 1, Y = j\}$, $i = 1, 2, ..., r$, $j = 1, 2, ..., s$ je pravdepodobostná funkcia náhodného vektora $(X,\ Y)'$. Označme

$$p_{i\bullet} = P\{X = i\} = \sum_{j=1}^{s} p_{ij}, \quad p_{\bullet j} = P\{Y = j\} = \sum_{i=1}^{r} p_{ij}.$$

Predpokladajme, že $p_{ij} > 0$ pre všetky $i = 1, 2, ..., r$, $j = 1, 2, ..., s$. Majme náhodný výber $\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \begin{pmatrix} X_2 \\ Y_2 \end{pmatrix}, \dots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix}$ o rozsahu n z rozdelenia rovnakého ako má $\begin{pmatrix} X \\ Y \end{pmatrix}$. Možné výsledky (realizácie) sú

$$\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \dots, \begin{pmatrix} r \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \dots, \begin{pmatrix} r \\ 2 \end{pmatrix}, ..., \begin{pmatrix} 1 \\ s \end{pmatrix}, \begin{pmatrix} 2 \\ s \end{pmatrix}, \dots, \begin{pmatrix} r \\ s \end{pmatrix},$$

teda rs "tried". Nech náhodná veličina

$\xi^{(n)}_{11}$ je počet tých $\begin{pmatrix} X_i \\ Y_i \end{pmatrix}$, ktoré nadobudli $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$,

$\xi^{(n)}_{12}$ je počet tých $\begin{pmatrix} X_i \\ Y_i \end{pmatrix}$, ktoré nadobudli $\begin{pmatrix} 1 \\ 2 \end{pmatrix}$,

$\vdots$

$\xi^{(n)}_{rs}$ je počet tých $\begin{pmatrix} X_i \\ Y_i \end{pmatrix}$, ktoré nadobudli $\begin{pmatrix} r \\ s \end{pmatrix}$.

Náhodné veličiny $\xi^{(n)}_{11}, \xi^{(n)}_{12}, ..., \xi^{(n)}_{rs}$ majú multinomické rozdelenie $Mu(n, p_{11}, p_{12}, ..., p_{rs})$ s pravdepodobnostnou funkciou

$$P\{\xi^{(n)}_{11} = x_{11}, \xi^{(n)}_{12} = x_{12}, ..., \xi^{(n)}_{rs} = x_{rs}\} = \frac{n!}{x_{11}!...x_{rs}!}\, p_{11}^{x_{11}}...p_{rs}^{x_{rs}}, \qquad x_{ij} \in \{0, 1, 2, ..., n\},\ \sum_{i=1}^{r}\sum_{j=1}^{s} x_{ij} = n. \tag{14}$$

Ak označíme n_{ij} realizáciu náhodnej veličiny $\xi^{(n)}_{ij}$ (počet tých členov náhodného výberu, ktoré nadobudli hodnotu $\begin{pmatrix} i \\ j \end{pmatrix}$), môžeme všetky výsledky (realizácie) zapísať do *kontingenčnej* $r \times s$ *tabuľky*:

	Y				
X	1	2	$\dots$	s	Σ
1	n_{11}	n_{12}	$\dots$	n_{1s}	$n_{1\bullet}$
2	n_{21}	n_{22}	$\dots$	n_{2s}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
r	n_{r1}	n_{r2}	$\dots$	n_{rs}	$n_{r\bullet}$
Σ	$n_{\bullet 1}$	$n_{\bullet 2}$	$\dots$	$n_{\bullet s1}$	n

V kontingenčnej tabuľke $n_{i\bullet} = \sum_{j=1}^{s} n_{ij}$, $n_{\bullet j} = \sum_{i=1}^{r} n_{ij}$, $n = \sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}$; $n_{i\bullet}$, $n_{\bullet j}$ sú *marginálne početnosti.*

Lema 2.15: X a Y sú nezávislé $\iff$ $p_{ij} = p_{i\bullet}\, p_{\bullet j}$ pre každé $i \in \{1,2,...,r\}$, $j \in \{1,2,...,s\}$.

Dôkaz: X a Y sú nezávislé práve vtedy, ak pre každé dve borelovské množiny A, B je $P\{X \in A, Y \in B\} = P\{X \in A\}P\{Y \in B\}$. Ak X, Y sú diskrétne a $X \in \{1,2,...,r\}$, $Y \in \{1,2,...,s\}$, tak je nutná a postačujúca podmienka nezávislosti X a Y jednoduchšia, a síce

$$P\{X \in A, Y \in B\} = P\{X \in A\}P\{Y \in B\} \text{ pre každú } A \subset \{1,2,...,r\},\ B \subset \{1,2,...,s\}.$$

Nech sú X a Y nezávislé. Potom pre $A = \{i\}$, $B = \{j\}$, $i \in \{1,2,...,r\}$, $j \in \{1,2,...,s\}$ dostávame

$$p_{ij} = P\{X = i, Y = j\} = P\{X = i\}P\{Y = j\} = p_{ij}.$$

Naopak, nech platí $p_{ij} = p_{i\bullet}\, p_{\bullet j}$ pre každé $i \in \{1,2,...,r\}$, $j \in \{1,2,...,s\}$. Vezmime ľubovoľnú $A = \{i_1,...,i_a\} \subset \{1,2,...,r\}$, ($i_1,...,i_a$ sú rôzne), $B = \{j_1,...,j_b\} \subset \{1,2,...,s\}$, ($j_1,...,j_b$ sú rôzne). Potom

$$P\{X \in A, Y \in B\} = P\{X \in \{i_1,...,i_a\}, Y \in \{j_1,...,j_b\}\} =$$

$$= P\{X = i_1, Y = j_1 \cup X = i_1, Y = j_2 \cup ...\ X = i_1, Y = j_b \cup$$

$$\cup X = i_2, Y = j_1 \cup X = i_2, Y = j_2 \cup ...\ X = i_2, Y = j_b \cup$$

$$\vdots$$

$$\cup X = i_a, Y = j_1 \cup X = i_a, Y = j_2 \cup ...\ X = i_a, Y = j_b\} =$$

$$= \sum_{u=1}^{a}\sum_{t=1}^{b} p_{i_u\ j_t} \overset{\text{predpoklad}}{=} \sum_{u=1}^{a}\sum_{t=1}^{b} p_{i_u\bullet} p_{\bullet j_t} =$$

$$= p_{i_1\bullet}p_{\bullet j_1} + p_{i_1\bullet}p_{\bullet j_2} + ... + p_{i_1\bullet}p_{\bullet j_b} +$$

$$+ p_{i_2\bullet}p_{\bullet j_1} + p_{i_2\bullet}p_{\bullet j_2} + ... + p_{i_2\bullet}p_{\bullet j_b} +$$

$$\vdots$$

$$+ p_{i_a\bullet}p_{\bullet j_1} + p_{i_a\bullet}p_{\bullet j_2} + ... + p_{i_a\bullet}p_{\bullet j_b} =$$

$$= (p_{i_1\bullet} + p_{i_2\bullet} + ... + p_{i_a\bullet})p_{\bullet j_1} + ... + (p_{i_1\bullet} + p_{i_2\bullet} + ... + p_{i_a\bullet})p_{\bullet j_b} =$$

$$= (p_{i_1\bullet} + p_{i_2\bullet} + ... + p_{i_a\bullet})(p_{\bullet j_1} + p_{\bullet j_2} + ... + p_{\bullet j_b}) =$$

$$= (P\{X = i_1\} + P\{X = i_2\} + ... + P\{X = i_a\})\,(P\{Y = j_1\} + P\{Y = j_2\} + ... + P\{Y = j_b\}) =$$

$$= P\{X \in A\}P\{Y \in B\}. \qquad \text{Q.E.D.}$$

Teda hypotéza $\quad H_0: \quad X$ *a* Y *sú nezávislé*

je ekvivalentná s $\quad H_0: \quad p_{ij} = p_{i\bullet} p_{\bullet j}$ *pre každé* $i \in \{1,2,...,r\},\ j \in \{1,2,...,s\}$.

Za platnosti H_0 májú náhodné veličiny $\xi_{11}^{(n)}, \xi_{12}^{(n)}, ..., \xi_{rs}^{(n)}$ multinomické rozdelenie, pričom parametre p_{ij} sú funkciami parametrov $p_{1\bullet}, p_{2\bullet}, ..., p_{r\bullet}, p_{\bullet 1}, p_{\bullet 2}, ..., p_{\bullet s}$ (p_{ij} vyjadríme pomocou nich). Je tu ešte malá komplikácia, a síce to, že "nezávislých " parametrov je iba $r+s-2$. Sú to napríklad $p_{1\bullet}, p_{2\bullet}, ..., p_{r-1\bullet}, p_{\bullet 1}, p_{\bullet 2}, ..., p_{\bullet s-1}$ (v tomto prípade je $p_{r\bullet} = 1 - \sum_{i=1}^{r-1} p_{i\bullet},\ p_{\bullet s} = 1 - \sum_{j=1}^{s-1} p_{\bullet j}$). Teda (za platnosti H_0)

$$p_{ij} = p_{ij}(p_{1\bullet}, p_{2\bullet}, ..., p_{r-1\bullet}, p_{\bullet 1}, p_{\bullet 2}, ..., p_{\bullet s-1}) \quad i = 1,2,...,r,\ j = 1,2,...,s$$

sú pravdepodobosti. Menovite

$p_{11} = p_{1\bullet}p_{\bullet 1}, ..., p_{1s} = p_{1\bullet}(1 - p_{\bullet 1} - p_{\bullet 2} - ... - p_{\bullet s-1}), ...,$

$p_{21} = p_{2\bullet}p_{\bullet 1}, ..., p_{2s} = p_{2\bullet}(1 - p_{\bullet 1} - p_{\bullet 2} - ... - p_{\bullet s-1}), ...,$

$\vdots$

$p_{r1} = (1 - p_{1\bullet} - p_{2\bullet} - ... - p_{r-1\bullet})p_{\bullet 1}, ..., p_{rs} = (1 - p_{1\bullet} - p_{2\bullet} - ... - p_{r-1\bullet})(1 - p_{\bullet 1} - p_{\bullet 2} - ... - p_{\bullet s-1}).$

Sú dostatočne hladké funkcie "neznámych" parametrov $p_{1\bullet}, p_{2\bullet}, ..., p_{r-1\bullet}, p_{\bullet 1}, p_{\bullet 2}, ..., p_{\bullet s-1}$. Tieto parametre odhadujeme metódou modifikovaného minimálneho χ^2.

Hľadáme riešenie sústavy

$$\sum_{a=1}^{r}\sum_{b=1}^{s} \frac{\xi_{ab}^{(n)}}{p_{ab}} \frac{\partial p_{ab}}{\partial p_{k\bullet}} = 0 \quad k = 1,2,...,r-1, \tag{15}$$

$$\sum_{a=1}^{r}\sum_{b=1}^{s} \frac{\xi_{ab}^{(n)}}{p_{ab}} \frac{\partial p_{ab}}{\partial p_{\bullet l}} = 0 \quad l = 1,2,...,s-1. \tag{16}$$

Napíšme si (15) pre $k = 1,2,...,r-1$ (naschvál niekde ponechávame $p_{r\bullet}, p_{\bullet s}$)

$$\frac{\xi_{11}^{(n)}}{p_{1\bullet}p_{\bullet 1}} \frac{\partial[p_{1\bullet}p_{\bullet 1}]}{\partial p_{k\bullet}} + \frac{\xi_{12}^{(n)}}{p_{1\bullet}p_{\bullet 2}} \frac{\partial[p_{1\bullet}p_{\bullet 2}]}{\partial p_{k\bullet}} + ... + \frac{\xi_{1s}^{(n)}}{p_{1\bullet}(1 - p_{\bullet 1} - ... - p_{\bullet s-1})} \frac{\partial[p_{1\bullet}(1 - p_{\bullet 1} - ... - p_{\bullet s-1})]}{\partial p_{k\bullet}} +$$

$$\frac{\xi_{21}^{(n)}}{p_{2\bullet}p_{\bullet 1}} \frac{\partial[p_{2\bullet}p_{\bullet 1}]}{\partial p_{k\bullet}} + \frac{\xi_{22}^{(n)}}{p_{2\bullet}p_{\bullet 2}} \frac{\partial[p_{2\bullet}p_{\bullet 2}]}{\partial p_{k\bullet}} + ... + \frac{\xi_{2s}^{(n)}}{(1 - p_{\bullet 1} - ... - p_{\bullet s-1})} \frac{\partial[p_{2\bullet}(1 - p_{\bullet 1} - ... - p_{\bullet s-1})]}{\partial p_{k\bullet}} +$$

$$\vdots$$

$$\frac{\xi_{r1}^{(n)}}{p_{r\bullet}p_{\bullet 1}} \frac{\partial[(1 - p_{1\bullet} - p_{2\bullet} - ... - p_{r-1\bullet})p_{\bullet 1}]}{\partial p_{k\bullet}} + ... + \frac{\xi_{rs}^{(n)}}{p_{r\bullet}p_{\bullet s}} \frac{\partial[(1 - p_{1\bullet} - p_{2\bullet} - ... - p_{r-1\bullet})(1 - p_{\bullet 1} - ... - p_{\bullet s-1})]}{\partial p_{k\bullet}} = 0 \tag{17}$$

Ak (17) derivujeme podľa $p_{1\bullet}$, dostaneme

$$\frac{\xi_{11}^{(n)}}{p_{1\bullet}p_{\bullet 1}} p_{\bullet 1} + \frac{\xi_{12}^{(n)}}{p_{1\bullet}p_{\bullet 2}} p_{\bullet 2} + ... + \frac{\xi_{1s}^{(n)}}{p_{1\bullet}p_{\bullet s}} p_{\bullet s} - \frac{\xi_{r1}^{(n)}}{p_{r\bullet}p_{\bullet 1}} p_{\bullet 1} - ... - \frac{\xi_{rs}^{(n)}}{p_{r\bullet}p_{\bullet s}} p_{\bullet s} = 0,$$

čiže

$$\frac{\xi_{11}^{(n)} + \xi_{12}^{(n)} + ... + \xi_{1s}^{(n)}}{p_{1\bullet}} - \frac{\xi_{r1}^{(n)} + \xi_{r2}^{(n)} + ... + \xi_{rs}^{(n)}}{p_{r\bullet}} = 0. \tag{18}$$

Označme $\xi_{l\bullet}^{(n)}$ náhodnú veličinu - počet tých $\begin{pmatrix} X_i \\ Y_i \end{pmatrix}$, ktoré nadobudli hodnotu $\begin{pmatrix} l \\ \text{ľubovoľné číslo} \end{pmatrix}$ a

$\xi_{\bullet k}^{(n)}$ náhodnú veličinu - počet tých $\begin{pmatrix} X_i \\ Y_i \end{pmatrix}$, ktoré nadobudli hodnotu $\begin{pmatrix} \text{ľubovoľné číslo} \\ k \end{pmatrix}$, tak (18) môžeme písať

$$\frac{\xi_{1\bullet}^{(n)}}{p_{1\bullet}} - \frac{\xi_{r\bullet}^{(n)}}{p_{r\bullet}} = 0.$$

Ak (17) derivujeme podľa $p_{k\bullet}$, $k = 2, 3, ..., r-1$, dostaneme analogicky

$$\frac{\xi_{k\bullet}^{(n)}}{p_{k\bullet}} - \frac{\xi_{r\bullet}^{(n)}}{p_{r\bullet}} = 0$$

a triviálne aj

$$\frac{\xi_{r\bullet}^{(n)}}{p_{r\bullet}} - \frac{\xi_{r\bullet}^{(n)}}{p_{r\bullet}} = 0.$$

Teda ak $\hat{p}_{1\bullet}, \hat{p}_{2\bullet}, ..., \hat{p}_{r\bullet}$ sú riešenia (15), tak

$$\frac{\xi_{k\bullet}^{(n)}}{\hat{p}_{k\bullet}} - \frac{\xi_{r\bullet}^{(n)}}{\hat{p}_{r\bullet}} = 0, \quad k = 1, 2, ..., r, \tag{19}$$

čiže

$$\xi_{k\bullet}^{(n)} = \frac{\xi_{r\bullet}^{(n)}}{\hat{p}_{r\bullet}} \hat{p}_{k\bullet}, \quad k = 1, 2, ..., r. \tag{20}$$

Rovnice (20) spočítame a dostávame

$$n = \sum_{k=1}^{r} \xi_{k\bullet}^{(n)} = \frac{\xi_{r\bullet}^{(n)}}{\hat{p}_{r\bullet}} \sum_{k=1}^{r} \hat{p}_{k\bullet} = \frac{\xi_{r\bullet}^{(n)}}{\hat{p}_{r\bullet}} 1,$$

teda

$$n = \frac{\xi_{r\bullet}^{(n)}}{\hat{p}_{r\bullet}}$$

a z (19) máme odhady získané metódou modifikovaného minimálneho χ^2 (v prípade platnosti H_0 o nezávislosti X a Y)

$$\hat{p}_{k\bullet} = \frac{1}{n}\, \xi_{k\bullet}^{(n)}, \quad k = 1, 2, ..., r.$$

Úplne analogickou cestou z rovníc (16) získame

$$\hat{p}_{\bullet l} = \frac{1}{n}\, \xi_{\bullet l}^{(n)}, \quad l = 1, 2, ..., s.$$

Podľa Vety 2.10 má za platnosti H_0 o nezávislosti X a Y

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(\xi_{ij}^{(n)} - n\hat{p}_{i\bullet}\hat{p}_{\bullet j})^2}{n\hat{p}_{i\bullet}\hat{p}_{\bullet j}} = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(\xi_{ij}^{(n)} - \frac{\xi_{i\bullet}^{(n)}\xi_{\bullet j}^{(n)}}{n}\right)^2}{\frac{\xi_{i\bullet}^{(n)}\xi_{\bullet j}^{(n)}}{n}} \xrightarrow{\text{asymptoticky}} \chi_f^2,$$

pričom $f = \overbrace{rs}^{\text{počet tried}} - \overbrace{(r-1+s-1)}^{\text{počet parametrov}} - 1 = rs - r - s + 1 = (r-1)(s-1)$.

V kontingenčnej tabuľke máme $n_{ij}, n_{i\bullet}, n_{\bullet j}-$ realizácie náhodných veličín $\xi_{ij}^{(n)}, \xi_{i\bullet}^{(n)}, \xi_{\bullet j}^{(n)}$. Ak teda

$$\sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(n_{ij} - \frac{n_{i\bullet}n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet}n_{\bullet j}}{n}} = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{n n_{ij}^2}{n_{i\bullet}n_{\bullet j}} - 2\sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\frac{n_{i\bullet}n_{\bullet j}}{n}}{\frac{n_{i\bullet}n_{\bullet j}}{n}} n_{ij} + \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\frac{n_{i\bullet}^2 n_{\bullet j}^2}{n^2}}{\frac{n_{i\bullet}n_{\bullet j}}{n}} n_{ij} =$$

$$= n \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{n_{ij}^2}{n_{i\bullet} n_{\bullet j}} - 2n + \frac{1}{n} \underbrace{(n_{1\bullet} + ... + n_{r\bullet})}_{n} \underbrace{(n_{\bullet 1} + ... + n_{\bullet s})}_{n} =$$

$$= n \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{n_{ij}^2}{n_{i\bullet} n_{\bullet j}} - n \geq \chi^2_{(r-1)(s-1)}(1-\alpha),$$

tak na hladine významnosti α zamietame hypotézu o nezávislosti X a Y ($\chi^2_{(r-1)(s-1)}(1-\alpha)$ je $(1-\alpha)-$kvantil $\chi^2_{(r-1)(s-1)}$ rozdelenia). Veličina

$$\sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(n_{ij} - \frac{n_{i\bullet} n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet} n_{\bullet j}}{n}}$$

je (len) testovacou charakteristikou, nie je to miera závislosti medzi X a Y. Výhodnejšie je počítať tento vzťah ako $n \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{n_{ij}^2}{n_{i\bullet} n_{\bullet j}} - n$, lebo vidíme, kde (v ktorom políčku kontingenčnej tabuľky) nadobúda χ^2 veľkej hodnoty a je (najviac) porušená nezávislosť medzi X a Y. Test možno použiť len ak pre každú dvojicu (i,j) je $\frac{n_{i\bullet} n_{\bullet j}}{n} > 5$. Čísla n_{ij} sú empirické četnosti a $\frac{n_{i\bullet} n_{\bullet j}}{n}$ očakávané četnosti (pri nezávislosti X a Y).

Príklad 2.16: (Anděl, J., *Matematická statistika*, SNTL, Praha, 1985, str. 212.) Rodinný stav ženícha resp. nevesty môže byť "slobodný(á)", "ovdovelý(á)" a "rozvedený(á)". V nasledujúcej tabuľke sú údaje o (pôvodnom) rodinnom stave ženícha a nevesty v Československu v roku 1957. Treba zistiť, či rodinný stav ženícha a nevesty sú nezávislé.

	rod. stav nevesty			
rod. stav ženícha	slobodná	ovdovelá	rozvedená	celkom
slobodný	75 564 (71 501)	824 (2 033)	3 463 (6 317)	79 851
ovdovelý	1 370 (2 751)	904 (78)	798 (243)	3 072
rozvedený	4 603 (7 285)	590 (207)	2 943 (644)	8 136
celkom	81 537	2 318	7 204	91 059

Čísla v zátvorkách sú teoretické četnosti (v prípade nezávislosti), napr. 71 501 = (79 851.81 537)/91 059. Hodnota chi^2 kritéria je

$$\chi^2 = \frac{(75\ 564 - 71\ 501)^2}{71\ 501} + ... + \frac{(2\ 943 - 644)^2}{644} = 22\ 850,4 > \chi^2_4(0,95) = 9,488.$$

Preto zamietame hypotézu o nezávislosti (pôvodného) rodinného satvu ženícha a nevesty na hladine významnosti 0,05.

2.7 Test homogenity

Teraz riešme inú úlohu. Pozorujeme (meráme) diskrétnu náhodnú veličinu X, ktorej hodnoty môžu byť $x_1, x_2, ..., x_r$ (napr. $X-$ známka z matematiky, $x_1 = 1, ..., x_5 = 5$). Majme $s \geq 2$ nezávislých výberov (napr. 8. trieda, ktorú učí učiteľ A_1, 8. trieda, ktorú učí učiteľ A_2,...,8. trieda, ktorú učí učiteľ A_s). Rozsahy týchto výberov nech sú $n_{\bullet 1}, n_{\bullet 2}, ..., n_{\bullet s}$ (známe, fixné čísla). Označme náhodné veličiny

$X^{(i)}$ – známka z matematiky u učiteľa A_i $\quad i = 1, 2, ..., s$.

Nech

$X_1^{(1)}, X_2^{(1)}, ..., X_{n_{\bullet 1}}^{(1)}$ je náhodný výber z $X^{(1)}$ rozsahu $n_{\bullet 1}$,

$\vdots$

$X_1^{(s)}, X_2^{(s)}, ..., X_{n_{\bullet s}}^{(s)}$ je náhodný výber z $X^{(s)}$ rozsahu $n_{\bullet s}$.

Všeobecne pre $i = 1, 2, ..., s$ je

$$P\{X^{(i)} = x_1\} =^{(i)} \theta_1,\ P\{X^{(i)} = x_2\} =^{(i)} \theta_2,\ \dots,\ P\{X^{(i)} = x_r\} =^{(i)} \theta_r.$$

Sú rozdelenia pravdepodobnosti náhodných veličín $X^{(1)}, X^{(2)}, ..., X^{(s)}$ rovnaké ? Ide vlastne o test

$$H_0: \begin{pmatrix} {}^{(1)}\theta_1 \\ \vdots \\ {}^{(1)}\theta_r \end{pmatrix} = \begin{pmatrix} {}^{(2)}\theta_1 \\ \vdots \\ {}^{(2)}\theta_r \end{pmatrix} = ... = \begin{pmatrix} {}^{(s)}\theta_1 \\ \vdots \\ {}^{(s)}\theta_r \end{pmatrix} \quad \asymp \quad H_1: \text{ existuje } i \neq j \quad \text{že} \quad \begin{pmatrix} {}^{(i)}\theta_1 \\ \vdots \\ {}^{(i)}\theta_r \end{pmatrix} \neq \begin{pmatrix} {}^{(j)}\theta_1 \\ \vdots \\ {}^{(j)}\theta_r \end{pmatrix}.$$

Označme náhodnú veličinu

ξ_{ij} – počet realizácií (hodnôt) x_i v $j-$tom výbere $j = 1, 2, ..., s,\ i = 1, 2, ..., r$

a n_{ij} realizáciu náhodnej veličiny ξ_{ij}. Dostaneme $r \times s$ kontingenčnú tabuľku

	výber				
hodnota znaku	1	2	...	s	
x_1	n_{11}	n_{12}	...	n_{1s}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	...	n_{2s}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
x_r	n_{r1}	n_{r2}	...	n_{rs}	$n_{r\bullet}$
Σ	$n_{\bullet 1}$	$n_{\bullet 2}$	...	$n_{\bullet s1}$	n

Dá sa ukázať (pozri napr. Cramer, H., *Mathematical Methods of Statistics*, Princeton University Press, 1946), že ak

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(n_{ij} - \dfrac{n_{i\bullet} n_{\bullet j}}{n}\right)^2}{\dfrac{n_{i\bullet} n_{\bullet j}}{n}} \geq \chi^2_{(r-1)(s-1)}(1-\alpha),$$

tak zamietame H_0 na hladine významnosti α. Tento test sa nazýva *test homogenity*.

Priližný $100(1-\alpha)\%-$ný konfidečný interval pre ${}^{(a)}\theta_j -^{(b)} \theta_j$ (teda pre $P\{X^{(a)} = x_j\} - P\{X^{(b)} = x_j\}$) je

$$\left(\frac{n_{ja}}{n_{\bullet a}} - \frac{n_{jb}}{n_{\bullet b}} - \sqrt{\chi^2_{(r-1)(s-1)}(1-\alpha)} \sqrt{\frac{\dfrac{n_{ja}}{n_{\bullet a}}\left(1 - \dfrac{n_{ja}}{n_{\bullet a}}\right)}{n_{\bullet a}} + \frac{\dfrac{n_{jb}}{n_{\bullet b}}\left(1 - \dfrac{n_{jb}}{n_{\bullet b}}\right)}{n_{\bullet b}}}, \right.$$

$$\left. \frac{n_{ja}}{n_{\bullet a}} - \frac{n_{jb}}{n_{\bullet b}} + \sqrt{\chi^2_{(r-1)(s-1)}(1-\alpha)} \sqrt{\frac{\dfrac{n_{ja}}{n_{\bullet a}}\left(1 - \dfrac{n_{ja}}{n_{\bullet a}}\right)}{n_{\bullet a}} + \frac{\dfrac{n_{jb}}{n_{\bullet b}}\left(1 - \dfrac{n_{jb}}{n_{\bullet b}}\right)}{n_{\bullet b}}}, \right).$$

Ak tento interval neobsahuje 0, zamietame na hladine významnosti α hypotézu $H_0:\ {}^{(a)}\theta_j =^{(b)} \theta_j$.

2.8 Štvorpoľné tabuľky (čtyřpolní tabulky)

Ak X a Y sú dichotomické znaky (dvojhodnotové), teda ak v kontingenčnej tabuľke $r = s = 2$, dostávame 2×2 tabuľku (čtyřpolní tabulku)

	Y		
X	1	2	Σ
1	n_{11}	n_{12}	$n_{1\bullet}$
2	n_{21}	n_{22}	$n_{2\bullet}$
Σ	$n_{\bullet 1}$	$n_{\bullet 2}$	n

V tomto prípade ($r = s = 2$) sa (veľmi) zjednoduší výpočet $\chi^2 = \sum_{i=1}^r \sum_{j=1}^s \frac{\left(n_{ij} - \frac{n_{i\bullet}n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet}n_{\bullet j}}{n}}$. Počítajme

$$\left(n_{ij} - \frac{n_{i\bullet}n_{\bullet j}}{n}\right)^2 = \left[\frac{n_{ij}(n_{11}+n_{12}+n_{21}+n_{22}) - (n_{i1}+n_{i2})(n_{1j}+n_{2j})}{n}\right]^2 =$$

$$= \left[\frac{n_{ij}n_{11} + n_{ij}n_{12} + n_{ij}n_{21} + n_{ij}n_{22} - n_{i1}n_{1j} - n_{i1}n_{2j} - n_{i2}n_{1j} - n_{i2}n_{21}}{n}\right]^2. \tag{21}$$

Pre $i = 1, j = 1$ je čitateľ v (21) rovný

$$n_{11}n_{11} + n_{11}n_{12} + n_{11}n_{21} + n_{11}n_{22} - n_{11}n_{11} - n_{11}n_{21} - n_{12}n_{11} - n_{12}n_{21} = (n_{11}n_{22} - n_{12}n_{21})^2,$$

pre $i = 1, j = 2$ je čitateľ v (21) rovný

$$n_{12}n_{11} + n_{12}n_{12} + n_{12}n_{21} + n_{12}n_{22} - n_{11}n_{12} - n_{11}n_{22} - n_{12}n_{12} - n_{12}n_{22} = (-n_{11}n_{22} + n_{12}n_{21})^2.$$

Aj pre $i = 2, j = 1$ a $i = 2, j = 2$ je čitateľ v (21) vždy rovný $(n_{11}n_{22} - n_{12}n_{21})^2$. Teda

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{\left(n_{ij} - \frac{n_{i\bullet}n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet}n_{\bullet j}}{n}} = n\frac{(n_{11}n_{22} - n_{12}n_{21})^2}{n^2} \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{1}{n_{i\bullet}n_{\bullet j}} =$$

$$= \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n^2} \left[\frac{n_{1\bullet}n_{\bullet 1} + n_{1\bullet}n_{\bullet 2} + n_{2\bullet}n_{\bullet 1} + n_{2\bullet}n_{\bullet 2}}{n_{1\bullet}n_{\bullet 1}n_{2\bullet}n_{\bullet 2}}\right] =$$

$$= \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n^2} \left[\frac{(n_{1\bullet} + n_{2\bullet})(n_{\bullet 1} + n_{\bullet 2})}{n_{1\bullet}n_{\bullet 1}n_{2\bullet}n_{\bullet 2}}\right] = n\frac{(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1\bullet}n_{\bullet 1}n_{2\bullet}n_{\bullet 2}}.$$

Ak $\chi^2 \geq \chi^2_1(1-\alpha)$, tak (asymptoticky) na hladine významnosti α zamietame hypotézu o nezávislsti X a Y. Pozor, pre každé $i, j \in \{1, 2\}$ musí byť $\frac{n_{i\bullet}n_{\bullet j}}{n} \geq 5$.

Teraz si ukážeme ešte iný (asymptotický) spôsob testovania nezávislosti dvoch diskrétnych náhodných veličín X a Y, ktoré môžu nadobúdať len dve hodnoty.

Veta 2.17: Nech X a Y sú diskrétne náhodné veličiny, ktoré môžu nadobúdať len dve hodnoty. X a Y sú nezávislé $\iff$ $\delta = \frac{p_{11}p_{22}}{p_{21}p_{12}} = 1$ (označenie z kapitoly 2.6).

Dôkaz: Ak X a Y sú nezávislé, tak $p_{ij} = p_{i\bullet}p_{\bullet j}$ pre každé $i, j \in \{1, 2\}$, čiže

$$\delta = \frac{p_{11}p_{22}}{p_{21}p_{12}} = \frac{p_{1\bullet}p_{\bullet 1}p_{2\bullet}p_{\bullet 2}}{p_{2\bullet}p_{\bullet 1}p_{1\bullet}p_{\bullet 2}} = 1.$$

Naopak, ak $\delta = 1$, tak $p_{21}p_{12} = p_{11}p_{22}$. Preto

$$p_{1\bullet}p_{\bullet 1} = (p_{11} + p_{12})(p_{11} + p_{21}) = p_{11}p_{11} + p_{11}p_{21} + p_{12}p_{11} + \underbrace{p_{12}p_{21}}_{p_{11}p_{22}} = p_{11}(p_{11} + p_{21} + p_{12} + p_{22}) = p_{11}.$$

Podobne dostaneme $p_{ij} = p_{i\bullet}p_{\bullet j}$ pre každé $i, j \in \{1, 2\}$. Q.E.D.

Poznámka 2.18: $\delta = \dfrac{p_{11}p_{22}}{p_{21}p_{12}} = \dfrac{\frac{p_{11}}{p_{12}}}{\frac{p_{21}}{p_{22}}}$ sa volá aj *teoretická interakcia* resp. *pomer šancí (odds ratio)*. Jej odhad je $\hat{\delta} = \dfrac{\xi_{11}\xi_{22}}{\xi_{21}\xi_{12}}$. Realizácia tohto odhadu je $\dfrac{n_{11}n_{22}}{n_{21}n_{12}}$.

Bez dôkazu uvedieme nasledujúcu vetu. Jej dôkaz pozrite napríklad v knižke Anděl, J., *Základy matematické statistiky*, MFF UK, Praha, 2005.

Veta 2.19: Nech X a Y sú diskrétne náhodné veličiny, ktoré môžu nadobúdať len dve hodnoty a $\delta = \dfrac{p_{11}p_{22}}{p_{21}p_{12}}$, $\hat{\delta} = \dfrac{\xi_{11}\xi_{22}}{\xi_{21}\xi_{12}}$. Náhodná veličina

$$\frac{\ln \hat{\delta} - \ln \delta}{\sqrt{\dfrac{1}{n_{11}} + \dfrac{1}{n_{12}} + \dfrac{1}{n_{21}} + \dfrac{1}{n_{22}}}}$$

má asymptoticky $N(0, 1)$ rozdelenie.

Je zrejmé ako budeme testovať

$$H_0 : \quad \delta = \delta_0 \qquad \asymp \qquad H_1 : \quad \delta \neq \delta_0$$

a teda aj nezávislosť X a Y (špeciálny prípad tejto hypotézy ak $\delta_0 = 1$).

2.9 Fisherov exaktný test pre štvorpoľnú tabuľku (Fisherov faktoriálový test)

V predáškach Pravděpodobnost a statistika I sme si dokázali Cauchyho kombinatorický vzorec.

Lema 2.20: (Cauchyho kombinatorický vzorec.) Pre ľubovoľné reálne čísla x, y a celé nezáporné číslo n platí

$$\sum_{k=0}^{n} \binom{x}{k}\binom{y}{n-k} = \binom{x+y}{n}.$$

Ak x, y sú tiež celé nezáporné čísla, tvar predchádzajúceho vzorca je

$$\sum_{k=\max\{0,n-y\}}^{\min\{x,n\}} \binom{x}{k}\binom{y}{n-k} = \binom{x+y}{n}.$$

Budeme ho v nasledujúcom potrebovať.

X a Y sú dichotomické znaky (dvojhodnotové), $\binom{X_1}{Y_1}, \binom{X_2}{Y_2}, \dots, \binom{X_n}{Y_n}$ je náhodný výber o rozsahu n z rozdelenia rovnakého ako má $\binom{X}{Y}$. Pri označení z kapitoly 2.6 majú náhodné veličiny $\xi_{11}^{(n)}, \xi_{12}^{(n)}, \xi_{21}^{(n)}, \xi_{22}^{(n)}$ multinomické rozdelenie $Mu(n, p_{11}, p_{12}, p_{21}, p_{22})$ s pravdepodobnostnou funkciou

$$P\{\xi_{11}^{(n)} = n_{11}, \xi_{12}^{(n)} = n_{12}, \xi_{21}^{(n)} = n_{21}, \xi_{22}^{(n)} = n_{22}\} = \frac{n!}{n_{11}! n_{12}! n_{21}! n_{22}!}\, p_{11}^{n_{11}} p_{12}^{n_{12}} p_{21}^{n_{21}} p_{22}^{n_{22}}, \tag{22}$$

ak $n_{ij} \in \{0, 1, 2\}$, $\sum_{i=1}^{2} \sum_{j=1}^{2} n_{ij} = n$. To znamená, že (22) je pravdepodobnosť, že dostaneme štvorpoľnú tabuľku s hodnotami $n_{11}, n_{12}, n_{21}, n_{22}$ (pri náhodnom výbere rozsahu n).
Za predpokladu, že X a Y sú nezávislé, je

$$p_{11}^{n_{11}} p_{12}^{n_{12}} p_{21}^{n_{21}} p_{22}^{n_{22}} = p_{1\bullet}^{n_{11}} p_{\bullet 1}^{n_{11}} p_{1\bullet}^{n_{12}} p_{\bullet 2}^{n_{12}} p_{2\bullet}^{n_{21}} p_{\bullet 2}^{n_{21}} p_{2\bullet}^{n_{22}} p_{\bullet 2}^{n_{22}} = p_{1\bullet}^{n_{1\bullet}} p_{2\bullet}^{n_{2\bullet}} p_{\bullet 1}^{n_{\bullet 1}} p_{\bullet 2}^{n_{\bullet 2}} = Q.$$

Teda pravdepodobnosť, že dostaneme štvorpoľnú tabuľku s hodnotami $n_{11}, n_{12}, n_{21}, n_{22}$ (pri náhodnom výbere rozsahu n) a za predpokladu, že X a Y sú nezávislé, je

$$P\{\xi_{11}^{(n)} = n_{11}, \xi_{12}^{(n)} = n_{12}, \xi_{21}^{(n)} = n_{21}, \xi_{22}^{(n)} = n_{22}\} = \frac{n!}{n_{11}! n_{12}! n_{21}! n_{22}!}\, Q. \tag{23}$$

Pravdepodobnosť, že pri danom rozsahu výberu n a nezávislosti X a Y vznikne štvorpoľná tabuľka s (vhodnými) marginálnymi početnosťami $n_{1\bullet}, n_{2\bullet}, n_{\bullet 1}, n_{\bullet 2}$ (vlastne stačí mať určené dve z nich ostatné sa už dopočítajú, napr. ak máme $n_{1\bullet}, n_{\bullet 1}$, tak $n_{2\bullet} = n - n_{1\bullet}$, $n_{\bullet 2} = n - n_{\bullet 1}$) je vlasne pravdepodobnosť, že vznikne jedna z tabuliek typu (tvaru)

	i	$n_{1\bullet} - i$	$n_{1\bullet}$
	$n_{\bullet 1} - i$	$n_{2\bullet} - n_{\bullet 1} + i =$ $= n - (n_{1\bullet} + n_{\bullet 1}) + i$	$n - n_{1\bullet}$
	$n_{\bullet 1}$	$n - n_{\bullet 1}$	n

Samozrejme $n_{1\bullet} \geq 0$, $n_{\bullet 1} \geq 0$, $n_{1\bullet} \leq n$, $n_{\bullet 1} \leq n$

$$\left.\begin{array}{c} i \geq 0 \\ i \geq (n_{1\bullet} + n_{\bullet 1}) - n = n_{\bullet 1} - n_{2\bullet} \end{array}\right\} \Rightarrow i \geq \max\{0, n_{\bullet 1} - n_{2\bullet}\} \tag{24}$$

a tiež

$$\left.\begin{array}{c} i \leq n_{1\bullet} \\ i \leq n_{\bullet 1} \end{array}\right\} \Rightarrow i \leq \min\{n_{1\bullet}, n_{\bullet 1}\}. \tag{25}$$

Pravdepodobnosť každej takejto vhodnej tabuľky je daná vzťahom (23).

Preto pravdepodobnosť, že (pri danom rozsahu výberu n a nezávislosti X a Y) vznikne štvorpoľná tabuľka s marginálnymi početnosťami $n_{1\bullet}$, $n_{2\bullet}$ $(= n - n_{1\bullet})$, $n_{\bullet 1}$, $n_{\bullet 2}$ $(= n - n_{\bullet 1})$ (pričom $n_{1\bullet}, n_{\bullet 1} \geq 0$, $n_{1\bullet}, n_{\bullet 1} \leq n$) je

$$\sum_{i=\max\{0, n_{\bullet 1} - n_{2\bullet}\}}^{\min\{n_{1\bullet}, n_{\bullet 1}\}} P\{\xi_{11}^{(n)} = i, \xi_{12}^{(n)} = n_{1\bullet} - i, \xi_{21}^{(n)} = n_{\bullet 1} - i, \xi_{22}^{(n)} = \underbrace{n - n_{1\bullet}}_{n_{2\bullet}} - n_{\bullet 1} + i\} =$$

$$= \sum_{i=\max\{0,n_{\bullet 1}-n_{2\bullet}\}}^{\min\{n_{1\bullet},n_{\bullet 1}\}} \frac{n!}{i!(n_{1\bullet}-i)!(n_{\bullet 1}-i)!(n_{2\bullet}-n_{\bullet 1}+i)!}\, Q =$$

$$= \sum_{i=\max\{0,n_{\bullet 1}-n_{2\bullet}\}}^{\min\{n_{1\bullet},n_{\bullet 1}\}} \frac{n!n_{1\bullet}!n_{2\bullet}!}{n_{1\bullet}!n_{2\bullet}!i!(n_{1\bullet}-i)!(n_{\bullet 2}-(n_{\bullet 1}-i))!(n_{\bullet 1}-i)!}\, Q =$$

$$= Q\frac{n!}{n_{1\bullet}!n_{2\bullet}!} \sum_{i=\max\{0,n_{\bullet 1}-n_{2\bullet}\}}^{\min\{n_{1\bullet},n_{\bullet 1}\}} \binom{n_{1\bullet}}{i}\binom{n_{2\bullet}}{n_{\bullet 1}-i} = Q\frac{n!}{n_{1\bullet}!n_{2\bullet}!}\binom{n_{1\bullet}+n_{2\bullet}}{n_{\bullet 1}} =$$

$$= Q\frac{(n!)^2}{n_{1\bullet}!n_{2\bullet}!n_{\bullet 1}!n_{\bullet 2}!} = P\{\xi_{1\bullet}^{(n)} = n_{1\bullet}, \xi_{\bullet 2}^{(n)} = n_{\bullet 2}, \xi_{\bullet 1}^{(n)} = n_{\bullet 1}, \xi_{\bullet 2}^{(n)} = n_{\bullet 2}\}. \tag{26}$$

pričom sme použili Cauchyho kombinatorický vzorec (Lemu 2.20).

Ak $n_{1\bullet} = n_{11} + n_{12},\ n_{2\bullet} = n_{21} + n_{22},\ n_{\bullet 1} = n_{11} + n_{21},\ n_{\bullet 2} = n_{12} + n_{22}$, tak zrejme platí

$$P\{\xi_{11}^{(n)} = n_{11}, \xi_{12}^{(n)} = n_{12}, \xi_{21}^{(n)} = n_{21}, \xi_{22}^{(n)} = n_{22}, \xi_{1\bullet}^{(n)} = n_{1\bullet}, \xi_{\bullet 2}^{(n)} = n_{\bullet 2}, \xi_{\bullet 1}^{(n)} = n_{\bullet 1}, \xi_{\bullet 2}^{(n)} = n_{\bullet 2}\} =$$

$$= P\{\xi_{11}^{(n)} = n_{11}, \xi_{12}^{(n)} = n_{12}, \xi_{21}^{(n)} = n_{21}, \xi_{22}^{(n)} = n_{22}\},$$

lebo náhodná udalosť

$$\{\omega : \xi_{11}^{(n)}(\omega) = n_{11}, \xi_{12}^{(n)}(\omega) = n_{12}, \xi_{21}^{(n)}(\omega) = n_{21}, \xi_{22}^{(n)}(\omega) = n_{22},$$

$$\xi_{1\bullet}^{(n)}(\omega) = n_{1\bullet}, \xi_{\bullet 2}^{(n)}(\omega) = n_{\bullet 2}, \xi_{\bullet 1}^{(n)}(\omega) = n_{\bullet 1}, \xi_{\bullet 2}^{(n)}(\omega) = n_{\bullet 2}\} =$$

$$\{\omega : \xi_{11}^{(n)}(\omega) = n_{11}, \xi_{12}^{(n)}(\omega) = n_{12}, \xi_{21}^{(n)}(\omega) = n_{21}, \xi_{22}^{(n)}(\omega) = n_{22}\}.$$

Ak teda rozsah náhodného výberu je n a X, Y sú nezávislé, je podľa (23)

$$P\{\xi_{11}^{(n)} = n_{11}, \xi_{12}^{(n)} = n_{12}, \xi_{21}^{(n)} = n_{21}, \xi_{22}^{(n)} = n_{22}, \xi_{1\bullet}^{(n)} = n_{1\bullet}, \xi_{\bullet 2}^{(n)} = n_{\bullet 2}, \xi_{\bullet 1}^{(n)} = n_{\bullet 1}, \xi_{\bullet 2}^{(n)} = n_{\bullet 2}\} =$$

$$= \begin{cases} P\{\xi_{11}^{(n)} = n_{11}, \xi_{12}^{(n)} = n_{12}, \xi_{21}^{(n)} = n_{21}, \xi_{22}^{(n)} = n_{22}\}, & \text{ak } \ n_{1\bullet} = n_{11} + n_{12}, \\ & \quad n_{2\bullet} = n_{21} + n_{22}, \\ & \quad n_{\bullet 1} = n_{11} + n_{21}, \\ & \quad n_{\bullet 2} = n_{12} + n_{22}, \\ 0, & \text{inak} \end{cases}$$

$$= \begin{cases} \dfrac{n!}{n_{11}!n_{12}!n_{21}!n_{22}!}\, Q, & \text{ak } \ n_{1\bullet} = n_{11} + n_{12}, \\ & \quad n_{2\bullet} = n_{21} + n_{22}, \\ & \quad n_{\bullet 1} = n_{11} + n_{21}, \\ & \quad n_{\bullet 2} = n_{12} + n_{22}, \\ 0, & \text{inak.} \end{cases}$$

Podmienená pravdepodobnosť

$$P\{\xi_{11}^{(n)} = n_{11}, \xi_{12}^{(n)} = n_{12}, \xi_{21}^{(n)} = n_{21}, \xi_{22}^{(n)} = n_{22} / \xi_{1\bullet}^{(n)} = n_{1\bullet}, \xi_{\bullet 2}^{(n)} = n_{\bullet 2}, \xi_{\bullet 1}^{(n)} = n_{\bullet 1}, \xi_{\bullet 2}^{(n)} = n_{\bullet 2}\} =$$

$$= \begin{cases} \dfrac{P\{\xi_{11}^{(n)}=n_{11},\xi_{12}^{(n)}=n_{12},\xi_{21}^{(n)}=n_{21},\xi_{22}^{(n)}=n_{22},\xi_{1\bullet}^{(n)}=n_{1\bullet},\xi_{\bullet 2}^{(n)}=n_{\bullet 2},\xi_{\bullet 1}^{(n)}=n_{\bullet 1},\xi_{\bullet 2}^{(n)}=n_{\bullet 2}\}}{P\{\xi_{1\bullet}^{(n)}=n_{1\bullet},\xi_{\bullet 2}^{(n)}=n_{\bullet 2},\xi_{\bullet 1}^{(n)}=n_{\bullet 1},\xi_{\bullet 2}^{(n)}=n_{\bullet 2}\}}, & \text{ak } n_{1\bullet} = n_{11} + n_{12}, \\ & n_{2\bullet} = n_{21} + n_{22}, \\ & n_{\bullet 1} = n_{11} + n_{21}, \\ & n_{\bullet 2} = n_{12} + n_{22}, \\ & \text{a "menovateľ"} \neq 0, \\ 0, & \text{inak} \end{cases}$$

(podľa (23) a (26))

$$= \begin{cases} \dfrac{\dfrac{n!}{n_{11}!n_{12}!n_{21}!n_{22}!}\,Q}{\dfrac{(n!)^2}{n_{1\bullet}!n_{2\bullet}!n_{\bullet 1}!n_{\bullet 2}!}\,Q} = \dfrac{n_{1\bullet}!n_{2\bullet}!n_{\bullet 1}!n_{\bullet 2}!}{n!n_{11}!n_{12}!n_{21}!n_{22}!}, & \text{ak } n_{1\bullet} = n_{11} + n_{12}, \\ & n_{2\bullet} = n_{21} + n_{22}, \\ & n_{\bullet 1} = n_{11} + n_{21}, \\ & n_{\bullet 2} = n_{12} + n_{22}, \\ & \text{a "menovateľ"} \neq 0, \\ 0, & \text{inak.} \end{cases} \tag{27}$$

Jednostranný test

H_0 : X, Y sú nezávislé $\asymp$ H_1 : čím sú "väčšie" hodnoty X, tým sú "väčšie" hodnoty Y
(alebo H_1 : čím sú "menšie" hodnoty X, tým sú "menšie" hodnoty Y)

na hladine významnosti najviac α sa realizuje tak, že sa spočíta (27) pre danú tabuľku, ale aj pre všetky ďalšie, ktoré z nej vzniknú postupným znižovaním najmenšej početnosti o jedničku pri zachovaní marginálnych početností (všetky takéto tabuľky totiž svedčia ešte viac proti hypotéze nezávislosti X a Y).

Príklad 2.21: (Anděl, J., *Matematická statistika*, SNTL, Praha, 1985, str. 215.) U 27 náhodne vybraných pacientiek trpiacich chrípkou sa zisťovalo, či boli proti nej očkované a aký priebeh má choroba. Výsledky sú v nasledujúcej tabuľke.

	priebeh choroby		
	ľahký	ťažký	
očkovaná	10	2	12
neočkovaná	4	11	15
	14	13	27

Zaujíma nás,či súvisí očkovanie s priebehom choroby, teda testujeme

H_0 : očkovanie nesúvisí s priebehom choroby $\asymp$ H_1 : (jednostranná) očkovanie "pozitívne" ovplyvní priebeh choroby

Najmenšia početnosť v tabuľke je $n_{12} = 2$. Spočítame (27) pre danú tabuľku (P=0,004491) ako aj pre tabuľky

	priebeh choroby		
	ľahký	ťažký	
očkovaná	11	1	12
neočkovaná	3	12	15
	14	13	27

	priebeh choroby		
	ľahký	ťažký	
očkovaná	12	0	12
neočkovaná	2	13	15
	14	13	27

(P=0,000272 a P=0,000005). Súčet všetkých týchto pravdepodobností je 0,004768. Je menší ako $\alpha = 0,05$ Preto H_0 zamietame v prospech H_1 na hladine významnosti menšej alebo rovnej 0,05.

Poznámka 2.22: Dá sa ukázať (pozri napr. Anděl, J., *Matematická statistika*, SNTL, Praha, 1985, str. 215), že hladina významnosti tohto testu je menšia alebo rovná α.

Obojstranný test, ktorý má alternatívu

H_1: čím sú "väčšie" hodnoty X, tým sú "väčšie" hodnoty Y alebo čím sú "väčšie" hodnoty X, tým sú "menšie" hodnoty Y

sa konštruuje omoho ťažšie. Je ale veľmi často používaný v praxi, preto ho popíšeme. Má 4 základné varianty (pozri Anděl, J., *Matematická statistika*, SNTL, Praha, 1985, str. 215), my si uvedieme jeden z nich.

V nameranej štvorpoľnej tabuľke nájdeme najmenšiu početnosť. Pozrieme na marginálne početnosti riadka a stĺpca, ktoré sa pretínajú v tejto najmenšej početnosti. Vyberieme si tú marginálnu početnosť z nich, ktorá je menšia a urobíme *líniu* (spojnicu menšej marginálnej početnosti a najmenšej početnosti). Pre nameranú štvorpoľnú tabuľku spočítame P z (27). Teraz vytvárame nové stvorpoľné tabuľky tak, že vždy znížime v danej tabuľke minimálnu početnosť o jednu (pri zachovaní marginálnych početností) a v obdržanej tabľke spočítame P z (27) (ako u jednostranného testu). Potom v línii v pôvodnej tabuľke zameníme početnosti. Vytvárame opäť nové tabuľky tak, že najmenšiu početnosť znižujeme o jednu (pri zachovaní marginálnych početností). Vždy spočítame P z (27). Spočítame všetky takto získané P a keď ich súčet neprekročí α, zamietame H_0 o nezávislosti X a Y oproti obojstrannej alternatíve.

Príklad 2.23: V Príklade 2.21 nech alternatívna hyptéza je H_1 : očkovanie môže pozitívne aj negatívne ovplyvniť priebeh choroby. V pôvodnej tabuľke

	priebeh choroby		
	ľahký	ťažký	
očkovaná	10	2	12
neočkovaná	4	11	15
	14	13	27

je minimálna početnosť $n_{12} = 2$, marginálne početnosť $n_{\bullet 1} = 12 < n_{2\bullet} = 13$, preto línia je n_{11} n_{12} $n_{\bullet 1}$ (čiže 10 2 12). Pravdepodobnosť P z (27) pre túto tabuľku je P=0,004491. Postupne spočítame P z (27) pre tabuľky

	priebeh choroby		
	ľahký	ťažký	
očkovaná	11	1	12
neočkovaná	3	12	15
	14	13	27

	priebeh choroby		
	ľahký	ťažký	
očkovaná	12	0	12
neočkovaná	2	13	15
	14	13	27

(P=0,000272 a P=0,000005). Teraz "vymeníme" v línii v pôvodnej tabuľke n_{11} a n_{12} (pri zachovaní marginálnych početností) a dostávame tabuľku

	priebeh choroby		
	ľahký	ťažký	
očkovaná	2	10	12
neočkovaná	12	3	15
	14	13	27

(pre túto tabuľku je P=0,001497) a postupne tabuľky

	priebeh choroby		
	ľahký	ťažký	
očkovaná	1	11	12
neočkovaná	13	2	15
	14	13	27

	priebeh choroby		
	ľahký	ťažký	
očkovaná	0	12	12
neočkovaná	14	1	15
	14	13	27

(P=0,000062 a P=0,000001). Súčet P zo všetkých šiestich tabuliek je 0,006328, je menší ako $\alpha = 0,05$, teda obojstranný test zamieta H_0 (priebeh choroby nezávisí s očkovaním) na hladine významnosti menšej alebo rovnej 0,05.

3 Všeobecný lineárny model

3.1 Najlepšie nevychýlené (nestranné) lineárne odhady parametra $\boldsymbol{\beta}$ v regulárnom lineárnom modeli

$(\mathbf{Y}_{n,1}, \mathbf{X}\boldsymbol{\beta}_{k,1}, \mathbf{V})$ je všeobecný lineárny (regresný) model, ak náhodný vektor $\mathbf{Y} \in \mathcal{R}^n$ (ktorého realizácie meráme, pozorujeme, observujeme, nazývame ho aj observačným vektorom) má strednú hodnotu $\mathcal{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}_{k,1}$, $\mathbf{X}_{n,k}$ je známa, "pevná" matica (tzv. matica plánu experimentu), $\boldsymbol{\beta} \in \mathcal{R}^k$ je vektor neznámych (ale "pevných", t.j. nenáhodných) parametrov, $cov\mathbf{Y} = \mathbf{V}$ je známa pozitívne semidefinitná (p.s.d.) matica, alebo $cov\mathbf{Y} = \sigma^2\mathbf{H}$, kde σ^2 je známy alebo neznámy parameter (skalárny faktor kovariančnej matice, jednotková disperzia) a $\mathbf{H}$ je známa matica, $cov\mathbf{Y}$ nezávisí od $\boldsymbol{\beta}$, $n > k$. Všeobecný lineárny model $(\mathbf{Y}_{n,1}, \mathbf{X}\boldsymbol{\beta}_{k,1}, \mathbf{V})$ je *regulárny*, ak hodnosť $h(\mathbf{X}) = k$ a $cov\mathbf{Y}$ je pozitívne definitná (p.d.) matica (teda regulárna).

Definícia 3.1:

(i) Nech $\mathbf{g}: \ (\mathcal{R}^n, \mathcal{B}^n) \to (\mathcal{R}^k, \mathcal{B}^k)$ je merateľné zobrazanie. $\mathbf{b} = \mathbf{g}(\mathbf{Y})$ je *lineárnym odhadom* vektora parametrov $\boldsymbol{\beta}_{k,1}$, ak $\mathbf{b} = \mathbf{u} + \mathbf{U}\mathbf{Y}$, kde $\mathbf{u} \in \mathcal{R}^k$ a $\mathbf{U}$ je reálna $k \times n$ matica.

(ii) $\mathbf{b}$ je *nevychýleným (nestranným)* odhadom $\boldsymbol{\beta}$, ak $\mathcal{E}_{\boldsymbol{\beta}}(\mathbf{b}) = \boldsymbol{\beta}$ pre každé $\boldsymbol{\beta} \in \mathcal{R}^k$. (Zamená to toľko, že ak vektor neznámych parametrov je $\boldsymbol{\beta}$ (ľubovoľný $k-$rozmerný vektor), tak stredná hodnota odhadu je práve tento vektor $\boldsymbol{\beta}$.)

(iii) $\mathbf{b}$ je *najlepším nevychýleným (nestranným) lineárnym* odhadom (NNLO) vektora parametrov $\boldsymbol{\beta}$, ak pre každý iný nevychýlený lineárny odhad $\mathbf{b}^*$ parametrov $\boldsymbol{\beta}$ patí, že $cov(\mathbf{b}^*) - cov(\mathbf{b})$ je p.s.d. matica.

Veta 3.2: Lineárny odhad $\mathbf{b} = \mathbf{u} + \mathbf{U}\mathbf{Y}$ je nestranným odhadom $\boldsymbol{\beta}$ práve ak $\mathbf{u} = \mathbf{0}$ a $\mathbf{U}\mathbf{X} = \mathbf{I}_{k,k}$.

Dôkaz: $\mathcal{E}(\mathbf{b}) = \mathcal{E}(\mathbf{u} + \mathbf{U}\mathbf{Y}) = \mathbf{u} + \mathbf{U}\mathcal{E}(\mathbf{Y}) = \mathbf{u} + \mathbf{U}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta} \ \ \forall\, \boldsymbol{\beta} \in \mathcal{R}^k \quad \Longleftrightarrow \quad \mathbf{u} = \mathbf{0}$ a $\mathbf{U}\mathbf{X} = \mathbf{I}_{k,k}$ Q.E.D.

V ďalšom budeme potrebovať nasledujúcu lemu.

Lema 3.3: Pre každú maticu $\mathbf{D}_{k,l}$ platí $\mathcal{M}(\mathbf{D}) = \mathcal{M}(\mathbf{D}\mathbf{D}')$, kde $\mathcal{M}(\mathbf{D}) = \{\mathbf{D}\mathbf{u} : \ \mathbf{u} \in \mathcal{R}^l\}$ je vektorový priestor generovaný stĺpcami matice $\mathbf{D}$ (podpriestor priestora $\mathcal{R}^k$).

Dôkaz: Označme $[\mathcal{M}(\mathbf{D})]^\perp$ ortogonálny doplnok priestora $\mathcal{M}(\mathbf{D})$ v (celom) priestore $\mathcal{R}^k$. Platí $\mathcal{M}(\mathbf{D}) = \mathcal{M}(\mathbf{D}\mathbf{D}') \Longleftrightarrow [\mathcal{M}(\mathbf{D})]^\perp = [\mathcal{M}(\mathbf{D}\mathbf{D}')]^\perp$. Budeme dokazovať rovnosť priestorov $[\mathcal{M}(\mathbf{D})]^\perp$ a $[\mathcal{M}(\mathbf{D}\mathbf{D}')]^\perp$.

Ak $\mathbf{z} \in [\mathcal{M}(\mathbf{D}\mathbf{D}')]^\perp \Longrightarrow \mathbf{z}'\mathbf{D}\mathbf{D}' = \mathbf{0} \Longrightarrow \mathbf{z}'\mathbf{D}\mathbf{D}'\mathbf{z} = 0 \Longrightarrow (\mathbf{D}'\mathbf{z})(\mathbf{D}'\mathbf{z})' = 0 \Longrightarrow \mathbf{D}'\mathbf{z} = \mathbf{0} \Longrightarrow \mathbf{z}'\mathbf{D} = \mathbf{0} \Longrightarrow \mathbf{z} \in [\mathcal{M}(\mathbf{D})]^\perp$, teda $[\mathcal{M}(\mathbf{D}\mathbf{D}')]^\perp \subset [\mathcal{M}(\mathbf{D})]^\perp$.

Ak $\mathbf{z} \in [\mathcal{M}(\mathbf{D})]^\perp \Longrightarrow \mathbf{z}'\mathbf{D} = \mathbf{0} \Longrightarrow \mathbf{z}'\mathbf{D}\mathbf{D}' = \mathbf{O} \Longrightarrow \mathbf{z} \in [\mathcal{M}(\mathbf{D}\mathbf{D}')]^\perp$, teda $[\mathcal{M}(\mathbf{D})]^\perp \subset [\mathcal{M}(\mathbf{D}\mathbf{D}')]^\perp$

Lema 3.4: V regulárnom lineárnom modeli je $\mathbf{X'V}^{-1}\mathbf{X}$ regulárna matica.

Dôkaz: V regulárnom lineárnom modeli je $h(\mathbf{X}_{n.k}) = k$. Využijúc tvrdenie Lemy 3.3 môžeme písať $h(\mathbf{X}) \geq h(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}) = h(\mathbf{X}'\mathbf{V}^{-\frac{1}{2}}\mathbf{V}^{-\frac{1}{2}}\mathbf{X}) = h(\mathbf{X}'\mathbf{V}^{-\frac{1}{2}}) \geq h(\mathbf{X}'\mathbf{V}^{-\frac{1}{2}}\mathbf{V}^{\frac{1}{2}}) = h(\mathbf{X})$, preto $h(\mathbf{X'V}^{-1}\mathbf{X}) = h(\mathbf{X}) = k$, pričom rozmer matice $\mathbf{X'V}^{-1}\mathbf{X}$ je $k \times k$. (Matice $\mathbf{V}^{-\frac{1}{2}}$ a $\mathbf{V}^{-\frac{1}{2}}$ sme si definovali v predáške Lineární statistické modely I.) Q.E.D.

Veta 3.5: V regulárnom lineárnom modeli je $(\mathbf{Y}_{n,1}, \mathbf{X}\boldsymbol{\beta}_{k,1}, \mathbf{V})$ je $\mathbf{b} = (\mathbf{X'V}^{-1}\mathbf{X})^{-1}\mathbf{X'V}^{-1}\mathbf{Y}$ NNLO parametra $\boldsymbol{\beta}$.

Dôkaz:

(i) $\mathbf{b} = (\mathbf{X'V}^{-1}\mathbf{X})^{-1}\mathbf{X'V}^{-1}\mathbf{Y}$ je lineárnym odhadom (matica $\mathbf{U}$ je v tomto prípade $(\mathbf{X'V}^{-1}\mathbf{X})^{-1}\mathbf{X'V}^{-1}$).

(ii) $\mathbf{U}\mathbf{X} = (\mathbf{X'V}^{-1}\mathbf{X})^{-1}\mathbf{X'V}^{-1}\mathbf{X} = \mathbf{I}_{k,k}$, teda podľa Vety 3.2 je $\mathbf{b}$ nestranným odhadom.

(iii) nech $\mathbf{b}^* = \mathbf{W}\mathbf{Y}$ je iný nestranný odhad, teda $\mathbf{W}\mathbf{X} = \mathbf{I}_{k,k}$, $\ cov(\mathbf{b}^*) = \mathbf{WVW'}$, potom

$$cov(\mathbf{b}^*) - cov(\mathbf{b}) = \mathbf{WVW'} - (\mathbf{X'V}^{-1}\mathbf{X})^{-1}\mathbf{X'V}^{-1}\mathbf{V}\mathbf{V}^{-1}\mathbf{X}(\mathbf{X'V}^{-1}\mathbf{X})^{-1} =$$

$$= \mathbf{WVW'} - \mathbf{I}(\mathbf{X'V}^{-1}\mathbf{X})^{-1}\mathbf{I} = \mathbf{WVW'} - \mathbf{W}\mathbf{X}(\mathbf{X'V}^{-1}\mathbf{X})^{-1}\mathbf{X'W'} =$$

$$= \mathbf{W}(\mathbf{V}^{\frac{1}{2}}\mathbf{V}^{\frac{1}{2}} - \mathbf{V}^{\frac{1}{2}}\mathbf{V}^{-\frac{1}{2}}\mathbf{X}(\mathbf{X'V}^{-1}\mathbf{X})^{-1}\mathbf{X'V}^{-\frac{1}{2}}\mathbf{V}^{\frac{1}{2}})\mathbf{W'} = \mathbf{W}\mathbf{V}^{\frac{1}{2}}\underbrace{(\mathbf{I} - \mathbf{V}^{-\frac{1}{2}}\mathbf{X}(\mathbf{X'V}^{-1}\mathbf{X})^{-1}\mathbf{X'V}^{-\frac{1}{2}})}_{\text{symetrická a idempotetná matica } \mathbf{A}=\mathbf{AA}}\mathbf{V}^{\frac{1}{2}}\mathbf{W'} =$$

$$= (\mathbf{W}\mathbf{V}^{\frac{1}{2}}\mathbf{A})(\mathbf{W}\mathbf{V}^{\frac{1}{2}}\mathbf{A})'$$

je pozitívne semidefinitná matica, lebo

$$\forall \mathbf{x} \in \mathcal{R}^k \quad \underbrace{\mathbf{x}'(\mathbf{W}\mathbf{V}^{\frac{1}{2}}\mathbf{A})}_{\mathbf{y}'}(\mathbf{W}\mathbf{V}^{\frac{1}{2}}\mathbf{A})'\mathbf{x} = \mathbf{y}'\mathbf{y} \geq 0. \qquad \text{Q.E.D.}$$

Príklad 3.6: (Vážený priemer.) Nech $Y_1, Y, ...Y_n$ sú nezávislé, $\mathcal{E}(Y_i) = \mu,\ i = 1, 2, ..., n$ a $\mathcal{D}(Y_i) = \sigma_i^2,\ i = 1, 2, ..., n$, všetky $\sigma_i^2 > 0$ a všetky poznáme. Potom $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)'$ sa riadi obecným lineárnym modelom, pričom

$$\mathcal{E}(\mathbf{Y}) = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \mu = \mathbf{1}_{n,1}\mu, \qquad cov(\mathbf{Y}) = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n^2 \end{pmatrix} = \mathbf{V}.$$

NNLO parametra μ (neznáma spoločná stredná hodnota všetkých Y_i) je podľa Vety 3.5

$$b = (\mathbf{1}'\mathbf{V}^{-1}\mathbf{1})^{-1}\mathbf{1}'\mathbf{V}^{-1}\mathbf{Y} =$$

$$= \left[(1, 1, ..., 1) \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{\sigma_n^2} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \right]^{-1} (1, 1, ..., 1) \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{\sigma_n^2} \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} =$$

$$= \left(\sum_{i=1}^n \frac{1}{\sigma_i^2} \right)^{-1} \sum_{j=1}^n \frac{Y_j}{\sigma_j^2}.$$

Tento odhad (túto náhodnú veličinu) nazývame *vážený priemer* $Y_1, Y_2, ..., Y_n$.

Veta 3.7: Nech $Y_1, Y_2, ...Y_n$ sú nezávislé, $Y_i \sim N(\mu, \sigma_i^2),\ i = 1, 2, ..., n$, všetky $\sigma_i^2 > 0$ a všetky poznáme. Potom $b = \left(\sum_{i=1}^n \frac{1}{\sigma_i^2}\right)^{-1} \sum_{j=1}^n \frac{Y_j}{\sigma_j^2}$ je NNLO parametra μ. Štatistika $\chi^2 = \sum_{i=1}^n \frac{(Y_i - b)^2}{\sigma_i^2} \sim \chi_{n-1}^2$ a $b \sim N\left(\mu, \left(\sum_{i=1}^n \frac{1}{\sigma_i^2}\right)^{-1}\right)$.

Dôkaz: Pretože $\mathbf{Y} = (Y_1, ..., Y_n)' \sim N(\mathbf{1}_{n,1}\mu, cov(\mathbf{Y}) = diag(\sigma_1^2, ..., \sigma_n^2))$, je (podľa Príkladu 3.6)

$$b = (\mathbf{1}'(cov(\mathbf{Y}))^{-1}\mathbf{1})^{-1}\mathbf{1}'(cov(\mathbf{Y}))^{-1}\mathbf{Y} = \left(\sum_{i=1}^n \frac{1}{\sigma_i^2} \right)^{-1} \sum_{j=1}^n \frac{Y_j}{\sigma_j^2}$$

a teda $b \sim N\left(\mu, \left(\sum_{i=1}^n \frac{1}{\sigma_i^2}\right)^{-1}\right)$. Ďalej platí

$$\chi^2 = \sum_{i=1}^n \frac{(Y_i - b)^2}{\sigma_i^2} = \sum_{i=1}^n \left(\frac{Y_i^2}{\sigma_i^2} - 2\frac{Y_i^2 b}{\sigma_i^2} + \frac{b^2}{\sigma_i^2} \right) =$$

$$= \sum_{i=1}^n \frac{Y_i^2}{\sigma_i^2} - 2 \underbrace{\left(\sum_{i=1}^n \frac{1}{\sigma_i^2} \right)^{-1} \sum_{j=1}^n \frac{Y_j}{\sigma_j^2}}_{b} \sum_{i=1}^n \frac{Y_i}{\sigma_i^2} + \underbrace{\left(\sum_{i=1}^n \frac{1}{\sigma_i^2} \right)^{-2} \left(\sum_{j=1}^n \frac{Y_j}{\sigma_j^2} \right)^2}_{b^2} \sum_{i=1}^n \frac{1}{\sigma_i^2} =$$

$$= \sum_{i=1}^{n} \frac{Y_i^2}{\sigma_i^2} - \left(\sum_{k=1}^{n} \frac{1}{\sigma_k^2} \right)^{-1} \sum_{k=1}^{n} \frac{Y_k}{\sigma_k^2} \sum_{j=1}^{n} \frac{Y_j}{\sigma_j^2} =$$

$$= \mathbf{Y'} \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{\sigma_n^2} \end{pmatrix} \mathbf{Y} - \left(\sum_{k=1}^{n} \frac{1}{\sigma_k^2} \right)^{-1} \mathbf{Y'} \begin{pmatrix} \frac{1}{\sigma_1^2} \\ \frac{1}{\sigma_2^2} \\ \vdots \\ \frac{1}{\sigma_n^2} \end{pmatrix} \left(\frac{1}{\sigma_1^2}, \frac{1}{\sigma_2^2}, ..., \frac{1}{\sigma_n^2} \right) \mathbf{Y} =$$

$$= \mathbf{Y'} \underbrace{\left\{ \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{\sigma_n^2} \end{pmatrix} - \left(\sum_{k=1}^{n} \frac{1}{\sigma_k^2} \right)^{-1} \begin{pmatrix} \frac{1}{\sigma_1^2} \\ \frac{1}{\sigma_2^2} \\ \vdots \\ \frac{1}{\sigma_n^2} \end{pmatrix} \left(\frac{1}{\sigma_1^2}, \frac{1}{\sigma_2^2}, ..., \frac{1}{\sigma_n^2} \right) \right\}}_{\mathbf{A}} \mathbf{Y}. \tag{28}$$

Pretože

$$\mathbf{A}\mathcal{E}(\mathbf{Y}) = \left[\begin{pmatrix} \frac{1}{\sigma_1^2} & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{\sigma_n^2} \end{pmatrix} - \left(\sum_{k=1}^{n} \frac{1}{\sigma_k^2} \right)^{-1} \begin{pmatrix} \frac{1}{\sigma_1^2} \\ \frac{1}{\sigma_2^2} \\ \vdots \\ \frac{1}{\sigma_n^2} \end{pmatrix} \left(\frac{1}{\sigma_1^2}, \frac{1}{\sigma_2^2}, ..., \frac{1}{\sigma_n^2} \right) \right] \mathbf{1}\mu = \mathbf{0},$$

môžeme písať

$$\chi^2 = \mathbf{Y}'\mathbf{A}\mathbf{Y} = (\mathbf{Y} - \mathcal{E}(\mathbf{Y}))'\mathbf{A}(\mathbf{Y} - \mathcal{E}(\mathbf{Y})).$$

Podľa Vety 13 kapitoly V. knihy Anděl, J., *Matematická statistika*, SNTL, Praha, 1985 (dokazovali sme ju aj na prednáške Lineární statistické modely I), v prípade, že $\mathbf{A}$ je symetrická, p.s.d. matica, $\mathbf{A}cov(\mathbf{Y}) \neq \mathbf{0}$ a idempotentná, tak $(\mathbf{Y} - \mathcal{E}(\mathbf{Y}))'\mathbf{A}(\mathbf{Y} - \mathcal{E}(\mathbf{Y}) \sim \chi^2_{tr[\mathbf{A}cov(\mathbf{Y})]}$. Overme si všetky predpoklady Vety 13.

$\mathbf{A} = \mathbf{A'}$ (zrejmé)

$\forall \mathbf{x} \in \mathcal{R}^n \quad \mathbf{x'Ax} = \sum_{i=1}^n \frac{x_i^2}{\sigma_i^2} - \left(\sum_{k=1}^n \frac{1}{\sigma_k^2} \right)^{-1} \left(\sum_{i=1}^n \frac{x_i}{\sigma_k^2} \right)^2 = C$. Ak vo Schwarzovej nerovnosti $(\sum_{i=1}^n \alpha_i\beta_i)^2 \leq \sum_{i=1}^n \alpha_i^2 \sum_{j=1}^n \beta_j^2$ zvolíme $\alpha_i = \frac{x_i}{\sigma_i}$, $\beta_j = \frac{1}{\sigma_j}$, dostávame, že $\sum_{i=1}^n \frac{x_i^2}{\sigma_i^2} \sum_{k=1}^n \frac{1}{\sigma_k^2} - \left(\sum_{i=1}^n \frac{x_i}{\sigma_k^2} \right)^2 = \left(\sum_{k=1}^n \frac{1}{\sigma_k^2} \right) C \geq 0$, teda $C \geq 0$ a preto $\forall \mathbf{x} \in \mathcal{R}^n \quad \mathbf{x'Ax} \geq 0$. $\mathbf{A}$ je p.s.d. matica.

$$\mathbf{A}cov(\mathbf{Y}) = \left[\begin{pmatrix} \frac{1}{\sigma_1^2} & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{\sigma_n^2} \end{pmatrix} - \left(\sum_{k=1}^{n} \frac{1}{\sigma_k^2} \right)^{-1} \begin{pmatrix} \frac{1}{\sigma_1^2} \\ \frac{1}{\sigma_2^2} \\ \vdots \\ \frac{1}{\sigma_n^2} \end{pmatrix} \left(\frac{1}{\sigma_1^2}, \frac{1}{\sigma_2^2}, ..., \frac{1}{\sigma_n^2} \right) \right] \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n^2 \end{pmatrix} =$$

$= \mathbf{I}_{n,n} - \left(\sum_{k=1}^{n} \frac{1}{\sigma_k^2} \right)^{-1} \begin{pmatrix} \frac{1}{\sigma_1^2} \\ \frac{1}{\sigma_2^2} \\ \vdots \\ \frac{1}{\sigma_n^2} \end{pmatrix} (1, 1, ..., 1) \neq \mathbf{0}$ (mimodiagonálne prvky tejto matice sú nenulové).

$\mathbf{A}cov(\mathbf{Y})\mathbf{A}cov(\mathbf{Y}) =$

$$= \left[\mathbf{I}_{n,n} - \left(\sum_{k=1}^{n} \frac{1}{\sigma_k^2} \right)^{-1} \begin{pmatrix} \frac{1}{\sigma_1^2} \\ \frac{1}{\sigma_2^2} \\ \vdots \\ \frac{1}{\sigma_n^2} \end{pmatrix} (1,1,...,1) \right] \left[\mathbf{I}_{n,n} - \left(\sum_{k=1}^{n} \frac{1}{\sigma_k^2} \right)^{-1} \begin{pmatrix} \frac{1}{\sigma_1^2} \\ \frac{1}{\sigma_2^2} \\ \vdots \\ \frac{1}{\sigma_n^2} \end{pmatrix} (1,1,...,1) \right] =$$

$$= \mathbf{I}_{n,n} - 2 \left(\sum_{k=1}^{n} \frac{1}{\sigma_k^2} \right)^{-1} \begin{pmatrix} \frac{1}{\sigma_1^2} \\ \frac{1}{\sigma_2^2} \\ \vdots \\ \frac{1}{\sigma_n^2} \end{pmatrix} (1,1,...,1) + \left(\sum_{k=1}^{n} \frac{1}{\sigma_k^2} \right)^{-2} \begin{pmatrix} \frac{1}{\sigma_1^2} \\ \frac{1}{\sigma_2^2} \\ \vdots \\ \frac{1}{\sigma_n^2} \end{pmatrix} \left(\sum_{k=1}^{n} \frac{1}{\sigma_k^2} \right)^{-1} (1,1,...,1) = \mathbf{A}cov\mathbf{Y}.$$

Platí $tr[\mathbf{A}cov\mathbf{Y}] = tr\mathbf{I}_{n,n} - \left(\sum_{k=1}^{n} \frac{1}{\sigma_k^2} \right)^{-1} tr \begin{pmatrix} \frac{1}{\sigma_1^2} \\ \frac{1}{\sigma_2^2} \\ \vdots \\ \frac{1}{\sigma_n^2} \end{pmatrix} (1,1,...,1) = n - \left(\sum_{k=1}^{n} \frac{1}{\sigma_k^2} \right)^{-1} tr(1,1,...,1) \begin{pmatrix} \frac{1}{\sigma_1^2} \\ \frac{1}{\sigma_2^2} \\ \vdots \\ \frac{1}{\sigma_n^2} \end{pmatrix} =$
$n - 1.$ Q.E.D.

Poznámka 3.8: Štatistiku $\chi^2 = \sum_{i=1}^{n} \frac{(Y_i - b)^2}{\sigma_i^2}$ z Vety 3.7 využijeme pri testovaní rovnosti stredných hodnôt. Majme $Y_1, Y_2, ..., Y_n$ nezávislé a $Y_i \sim N(\mu_i, \sigma_i^2),\ i = 1, 2, ..., n$, (všetky σ_i^2 poznáme).

$$H_0: \quad \mu_1 = \mu_2 = ... = \mu_n \qquad \asymp \qquad H_1: \ \exists i \neq j \quad \mu_i \neq \mu_j$$

testujeme pomocou štatistiky $\chi^2 = \sum_{i=1}^{n} \frac{(Y_i - b)^2}{\sigma_i^2}$, pričom $b = \left(\sum_{i=1}^{n} \frac{1}{\sigma_i^2} \right)^{-1} \sum_{j=1}^{n} \frac{Y_j}{\sigma_j^2}$. Za platnosti H_0 má $\chi^2 \sim \chi^2_{n-1}$ rozdelenie. Teda ak realizácia $\chi^2_{real} \geq \chi^2_{n-1}(1-\alpha)$, tak H_0 na hladine významnosti α zamietame.

Príklad 3.9: Test rovnosti korelačných koeficientov.
Majme n nezávislých náhodných výberov z dvojrozmerného regulárneho normálneho rozdelenia, teda

$$\begin{pmatrix} ^{(1)}X_1 \\ ^{(1)}Y_1 \end{pmatrix}, ..., \begin{pmatrix} ^{(1)}X_{k_1} \\ ^{(1)}Y_{k_1} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \nu_1 \end{pmatrix}, \begin{pmatrix} ^{(1)}\sigma_1^2 & ^{(1)}\rho\ ^{(1)}\sigma_1\ ^{(1)}\sigma_2 \\ ^{(1)}\rho\ ^{(1)}\sigma_1\ ^{(1)}\sigma_2 & ^{(1)}\sigma_2^2 \end{pmatrix} \right),$$

$$\vdots$$

$$\begin{pmatrix} ^{(n)}X_1 \\ ^{(n)}Y_1 \end{pmatrix}, ..., \begin{pmatrix} ^{(n)}X_{k_n} \\ ^{(n)}Y_{k_n} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_n \\ \nu_n \end{pmatrix}, \begin{pmatrix} ^{(n)}\sigma_1^2 & ^{(n)}\rho\ ^{(n)}\sigma_1\ ^{(n)}\sigma_2 \\ ^{(n)}\rho\ ^{(n)}\sigma_1\ ^{(n)}\sigma_2 & ^{(n)}\sigma_2^2 \end{pmatrix} \right),$$

pričom nech $k_1, k_2, ..., k_n \geq 10$.

Vieme, že pre Fisherovu $Z-$transformáciu platí

$$Z_i = \frac{1}{2} \ln \frac{1 + R_i}{1 - R_i} \approx N\left(\frac{1}{2} \ln \frac{1 +^{(i)} \rho}{1 -^{(i)} \rho}, \frac{1}{k_i - 3} \right), \quad i = 1, 2, ..., n,$$

kde

$$R_i = \frac{\sum_{j=1}^{k_i} (^{(i)}X_j -^{(i)} \overline{X})(^{(i)}Y_j -^{(i)} \overline{Y})}{\sqrt{\sum_{j=1}^{k_i} (^{(i)}X_j -^{(i)} \overline{X})^2 \sum_{j=1}^{k_i} (^{(i)}Y_j -^{(i)} \overline{Y})^2}}$$

je výberový korelačný koeficient na $i-$tom výbere, ${}^{(i)}\overline{X} = \frac{1}{k_i}\sum_{j=1}^{k_i} {}^{(i)}X_j$, ${}^{(i)}\overline{Y} = \frac{1}{k_i}\sum_{j=1}^{k_i} {}^{(i)}Y_j$.

$$H_0: \quad {}^{(1)}\rho = {}^{(2)}\rho = ... = {}^{(n)}\rho \qquad \asymp \qquad H_1: \quad \exists i \neq j \quad {}^{(i)}\rho \neq {}^{(j)}\rho$$

testujeme pomocou testovacej štatistiky

$$\chi^2 = \sum_{i=1}^{n} \frac{(Z_i - b)^2}{\frac{1}{k_i - 3}} = \sum_{i=1}^{n}(k_i - 3)(Z_i - b)^2,$$

kde $b = \frac{1}{\sum_{i=1}^{n}(k_i-3)} \sum_{j=1}^{n} \frac{Z_j}{\frac{1}{k_j-3}} = \frac{1}{\sum_{i=1}^{n} k_i - 3n} \sum_{j=1}^{n}(k_j - 3)Z_j$. Podľa Vety 3.7 má za platnosti H_0 štatistika χ^2 rozdelenie χ^2_{n-1}. Ak teda jej realizácia $\chi^2_{real} \geq \chi^2_{n-1}(1-\alpha)$, tak H_0 zamietame na hladine významnosti α.

3.2 Poznámky k pseudoinverzným maticiam

Lema 3.10: Nech $\mathbf{A}, \mathbf{B}$ sú $k \times l$ reálne matice. $\mathbf{AA^-B} = \mathbf{B} \iff \mathcal{M}(\mathbf{B}) \subset \mathcal{M}(\mathbf{A})$.

Dôkaz: $\mathcal{M}(\mathbf{B}) \subset \mathcal{M}(\mathbf{A}) \iff$ stĺpce matice $\mathbf{B}$ sa dajú napísať ako lineárne kombinácie stĺpcov matice $\mathbf{A} \iff \{\mathbf{B}\}_{\bullet j} = \mathbf{A}\mathbf{d}_j,\ \mathbf{d}_j \in \mathcal{R}^l,\ j = 1, 2, ..., l \iff \exists \mathbf{D}_{l,l} = (\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_l)$, že $\mathbf{B} = \mathbf{AD}$.

Nech $\mathbf{AA^-B} = \mathbf{B} \implies \exists \mathbf{D}\ (= \mathbf{A^-B})$, že $\mathbf{AD} = \mathbf{B} \implies \mathcal{M}(\mathbf{B}) \subset \mathcal{M}(\mathbf{A})$.

Naopak ak $\mathcal{M}(\mathbf{B}) \subset \mathcal{M}(\mathbf{A}) \implies \exists \mathbf{D}$, že $\mathbf{B} = \mathbf{AD} \implies \mathbf{AA^-B} = \mathbf{AA^-AD} = \mathbf{AD} = \mathbf{B}$ Q.E.D.

Úplne analogicky dokážeme nasledujúcu lemu (dôkaz spravte ako cvičenie).

Lema 3.11: Nech $\mathbf{A}, \mathbf{B}$ sú $k \times l$ reálne matice. $\mathbf{BA^-A} = \mathbf{B} \iff \mathcal{M}(\mathbf{B'}) \subset \mathcal{M}(\mathbf{A'})$.

Lema 3.12: $\mathbf{A}_{m,n}$ je reálna matica. Potom

(i) $[(\mathbf{A'A})^-]'$ je g-inverzná matica k matici $\mathbf{A'A}$;

(ii) $\mathbf{A}(\mathbf{A'A})^-\mathbf{A'A} = \mathbf{A}$ (teda $(\mathbf{A'A})^-\mathbf{A'}$ je jedna g-inverzia $\mathbf{A}^-$;

(iii) $\mathbf{A}(\mathbf{A'A})^-\mathbf{A'}$ nezávisí na voľbe $(\mathbf{A'A})^-$ a je vždy symetrická (a jediná).

Dôkaz: (i) Platí $(\mathbf{A'A})(\mathbf{A'A})^-(\mathbf{A'A}) = (\mathbf{A'A})$. Keď túto rovnicu transponujeme (ľavú aj pravú stranu), dostávame $(\mathbf{A'A})[(\mathbf{A'A})^-]'(\mathbf{A'A}) = (\mathbf{A'A})$.

(ii) Pretože podľa Lemy 3.3 $\mathcal{M}(\mathbf{A'}) = \mathcal{M}(\mathbf{A'A})$, priamo z Lemy 3.11 dostávame $\mathbf{A}(\mathbf{A'A})^-\mathbf{A'A} = \mathbf{A}$.

(iii) Nech $(\mathbf{A'A})^-_*$ a $(\mathbf{A'A})^-_{**}$ sú dve g-inverzie matice $(\mathbf{A'A})$. Potom pomocou Lemy 3.10 a Lemy 3.11 dostávame

$$[\mathbf{A}(\mathbf{A'A})^-_*\mathbf{A'} - \mathbf{A}(\mathbf{A'A})^-_{**}\mathbf{A'}][\mathbf{A}(\mathbf{A'A})^-_*\mathbf{A'} - \mathbf{A}(\mathbf{A'A})^-_{**}\mathbf{A'}]' =$$

$$= \mathbf{A}(\mathbf{A'A})^-_* \underbrace{\mathbf{A'A}[(\mathbf{A'A})^-_*]'\mathbf{A'}}_{\mathbf{A'}} - \mathbf{A}(\mathbf{A'A})^-_* \underbrace{\mathbf{A'A}[(\mathbf{A'A})^-_{**}]'\mathbf{A'}}_{\mathbf{A'}} -$$

$$- \mathbf{A}(\mathbf{A'A})^-_{**} \underbrace{\mathbf{A'A}[(\mathbf{A'A})^-_*]'\mathbf{A'}}_{\mathbf{A'}} + \mathbf{A}(\mathbf{A'A})^-_{**} \underbrace{\mathbf{A'A}[(\mathbf{A'A})^-_{**}]'\mathbf{A'}}_{\mathbf{A'}} = \mathbf{0}.$$

Podľa Lemy 3.3 je $\mathcal{M}(\mathbf{0}) = \mathcal{M}\{[\mathbf{A}(\mathbf{A'A})^-_*\mathbf{A'} - \mathbf{A}(\mathbf{A'A})^-_{**}\mathbf{A'}][\mathbf{A}(\mathbf{A'A})^-_*\mathbf{A'} - \mathbf{A}(\mathbf{A'A})^-_{**}\mathbf{A'}]'\} =$
$= \mathcal{M}(\mathbf{A}(\mathbf{A'A})^-_*\mathbf{A'} - \mathbf{A}(\mathbf{A'A})^-_{**}\mathbf{A'})$, teda $\mathbf{A}(\mathbf{A'A})^-_*\mathbf{A'} - \mathbf{A}(\mathbf{A'A})^-_{**}\mathbf{A'} = \mathbf{0}$ a $\mathbf{A}(\mathbf{A'A})^-_*\mathbf{A'} = \mathbf{A}(\mathbf{A'A})^-_{**}\mathbf{A'}$.

Vidíme, že $\mathbf{A}(\mathbf{A'A})^-\mathbf{A'}$ nezávisí od voľby $(\mathbf{A'A})^-$ a je jediná. Zoberme si ľubovoľnú maticu $(\mathbf{A'A})^-$. Potom matica $\frac{1}{2}\{(\mathbf{A'A})^- + [(\mathbf{A'A})^-]'\}$ je symetrická g-inverzia matice $\mathbf{A'A}$ (dokážte). Matica $\mathbf{A}\frac{1}{2}\{(\mathbf{A'A})^- + [(\mathbf{A'A})^-]'\}\mathbf{A'}$ je ale symetrická a preto matica $\mathbf{A}(\mathbf{A'A})^-\mathbf{A'}$ je symetrická pre ľubovoľnú voľbu $(\mathbf{A'A})^-$. Q.E.D.

Veta 3.13: $\mathbf{A}_{k,n}\mathbf{x} = \mathbf{y}_{n,1}$ nech je konzistentný systém (t.j. má riešenie). $\mathbf{A}^-$ nech je ľubovoľná (ale pevná) g-inverzia matice $\mathbf{A}$. Práve všetky riešenia systému $\mathbf{Ax} = \mathbf{y}$ sú z množiny $\mathcal{A} = \{\mathbf{x} \in \mathcal{R}^n : \ \ \mathbf{x} = \mathbf{A}^-\mathbf{y} + (\mathbf{I} - \mathbf{A}^-\mathbf{A})\mathbf{z}, \ \ \mathbf{z} \in \mathcal{R}^n\}$.

Dôkaz: Systém $\mathbf{Ax} = \mathbf{y}$ je konzistentný (má riešenie) $\iff \ \exists\mathbf{x} : \ \ \mathbf{Ax} = \mathbf{y} \iff \mathbf{y} \in \mathcal{M}(\mathbf{A}) \iff \exists\mathbf{w} : \ \mathbf{y} = \mathbf{Aw}$.

Nech $\mathbf{x}_0 \in \mathcal{A}$, teda $\mathbf{x}_0 = \mathbf{A}^-\mathbf{y} + (\mathbf{I} - \mathbf{A}^-\mathbf{A})\mathbf{z}_0 \implies \mathbf{Ax}_0 = \mathbf{AA}^-\mathbf{y} + \underbrace{\mathbf{A}(\mathbf{I} - \mathbf{A}^-\mathbf{A})\mathbf{z}_0}_{\mathbf{0}} = \mathbf{AA}^-\mathbf{Aw} = \mathbf{Aw} = \mathbf{y} \implies \ \mathbf{x}_0$ je riešením systému $\mathbf{Ax} = \mathbf{y}$.

Naopak ak $\mathbf{x}^*$ je riešením konzistentného systému $\mathbf{Ax} = \mathbf{y}$, tak položme $\mathbf{z} = \mathbf{x}^*$ a $\mathbf{A}^-\mathbf{y} + (\mathbf{I} - \mathbf{A}^-\mathbf{A})\mathbf{z} = \mathbf{A}^-\mathbf{y} + (\mathbf{I} - \mathbf{A}^-\mathbf{A})\mathbf{x}^* = \mathbf{A}^-\mathbf{y} + \mathbf{x}^* - \mathbf{A}^-\underbrace{\mathbf{Ax}^*}_{\mathbf{y}} = \mathbf{x}^*$. Teda $\mathbf{x}^* \in \mathcal{A}$. Q.E.D.

3.3 Model s neúplnou hodnosťou, $cov(\mathbf{Y}) = \sigma^2\mathbf{I}$

Majme LRM - lineárny regresný model $(\mathbf{Y}_{n,1}, \mathbf{X}_{n,k}\boldsymbol{\beta}_{k,1}, \sigma^2\mathbf{I})$, v ktorom $h(\mathbf{X}) < k \le n$. Voláme ho *modelom s neúplnou hodnosťou (matice plánu)*. Našim cieľom bude odhad $\mathbf{c'}\boldsymbol{\beta}$, odhad σ^2 a štatistická inferencia o neznámych parameroch (inferencia - proces logického odvodzovania výrokov z iných výrokov).

Príklad 3.14: Majme tri nezávislé skupiny pozorovaní

$Y_{11}, Y_{12}, ..., Y_{1n_1}$ náhodný výber z $N(\mu_1, \sigma^2)$ rozsahu n_1,

$Y_{21}, Y_{22}, ..., Y_{2n_2}$ náhodný výber z $N(\mu_2, \sigma^2)$ rozsahu n_2,

$Y_{31}, Y_{32}, ..., Y_{3n_3}$ náhodný výber z $N(\mu_3, \sigma^2)$ rozsahu n_3.

Model (celého) merania je

$$Y_{ij} = \mu_i + \varepsilon_{ij} \quad i = 1,2,3, \quad j = 1,2,...,n_i,$$

$$\mathbf{Y} = \begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \\ Y_{31} \\ \vdots \\ Y_{3n_3} \end{pmatrix} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \mathbf{Y}_3 \end{pmatrix}, \quad \mathcal{E}(\mathbf{Y}) = \begin{pmatrix} 1 & 0 & 0 \\ & \vdots & \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ & \vdots & \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ & \vdots & \\ 0 & 0 & 1 \end{pmatrix}_{\sum_{i=1}^3 n_i, 3} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} = \mathbf{X}^*\boldsymbol{\gamma}, \quad cov(\mathbf{Y}) = \sigma^2\mathbf{I}_{\sum_{i=1}^3 n_i, \sum_{i=1}^3 n_i}.$$

Je to LRM $(\mathbf{Y}, \mathbf{X}^*\boldsymbol{\gamma}, \sigma^2\mathbf{I})$ plnej hodnosti. Ak tento model preparametrizujeme tak, že položíme $\mu_i = \mu + \alpha_i, \quad i = 1,2,3,$ máme 4 parametre strednej hodnoty $\mu, \alpha_1, \alpha_2, \alpha_3$. Parametre α_i voláme *efekty*.

Preparametrizovaný model je

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad i = 1, 2, 3, \quad j = 1, 2, ..., n_i,$$

$$\mathbf{Y} = \begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \\ Y_{31} \\ \vdots \\ Y_{3n_3} \end{pmatrix} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \mathbf{Y}_3 \end{pmatrix}, \quad \mathcal{E}(\mathbf{Y}) = \begin{pmatrix} 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 1 \end{pmatrix}_{\sum_{i=1}^3 n_i, 4} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} = \mathbf{X}\boldsymbol{\beta}, \quad cov(\mathbf{Y}) = \sigma^2 \mathbf{I}_{\sum_{i=1}^3 n_i, \sum_{i=1}^3 n_i}.$$

Tento model je LRM $(\mathbf{Y}, \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ s neúplnou hodnosťou, lebo $h(\mathbf{X}) = 3$ (prvý stĺpec matice plánu je súčtom druhého, tretieho a štvrtého stĺpca), teda $h(\mathbf{X}) = 3 < k$ $(= 4 < \sum_{i=1}^3 n_i = n)$.

Veta 3.15: Majme LRM $(\mathbf{Y}_{n,1}, \mathbf{X}_{n,k}\boldsymbol{\beta}_{k,1}, \sigma^2\mathbf{I})$ neúplnej hodnosti, teda $h(\mathbf{X}) < k \leq n$. Ak $\mathbf{b}$ je riešením normálnych rovníc $\mathbf{X'Xb} = \mathbf{X'Y}$, tak $\min_{\boldsymbol{\gamma} \in \mathcal{R}^k} (\mathbf{Y} - \mathbf{X}\boldsymbol{\gamma})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\gamma}) = (\mathbf{Y} - \mathbf{Xb})'(\mathbf{Y} - \mathbf{Xb})$, teda $\mathbf{b} = argmin_{\boldsymbol{\gamma} \in \mathcal{R}^k} (\mathbf{Y} - \mathbf{X}\boldsymbol{\gamma})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\gamma})$. Pre každé riešenie $\mathbf{b}$ normálnych rovníc má výraz $(\mathbf{Y} - \mathbf{Xb})'(\mathbf{Y} - \mathbf{Xb})$ rovnakú hodnotu.

Dôkaz: Najprv ukážeme, že systém $\mathbf{X'Xb} = \mathbf{X'Y}$ (pre neznámu $\mathbf{b}$) je konzistentný.
$\mathbf{X'Xb} = \mathbf{X'Y}$ je konzistentný $\iff$ $\mathbf{X'Y} \in \mathcal{M}(\mathbf{X'X})$. Pretože $\mathbf{X'Y} \in \mathcal{M}(\mathbf{X'}) = \mathcal{M}(\mathbf{X'X})$ (Lema 3.3), je $\mathbf{X'Xb} = \mathbf{X'Y}$ konzistentný systém (pre akékoľvek $\mathbf{Y}$).

Nech $\mathbf{b}$ je (ľubovoľné) riešenie $\mathbf{X'Xb} = \mathbf{X'Y}$, čiže $\mathbf{X'}(\mathbf{Y} - \mathbf{Xb}) = \mathbf{0}$, ale aj $(\mathbf{Y} - \mathbf{Xb})'\mathbf{X} = \mathbf{0'}$. Potom pre ľubovoľné $\boldsymbol{\gamma} \in \mathcal{R}^k$ platí

$$(\mathbf{Y} - \mathbf{X}\boldsymbol{\gamma})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\gamma}) = (\mathbf{Y} - \mathbf{Xb} - (\mathbf{X}\boldsymbol{\gamma} - \mathbf{Xb}))'(\mathbf{Y} - \mathbf{Xb} - (\mathbf{X}\boldsymbol{\gamma} - \mathbf{Xb})) =$$

$$= (\mathbf{Y} - \mathbf{Xb})'(\mathbf{Y} - \mathbf{Xb}) - (\mathbf{X}\boldsymbol{\gamma} - \mathbf{Xb})'(\mathbf{Y} - \mathbf{Xb}) - (\mathbf{Y} - \mathbf{Xb})'(\mathbf{X}\boldsymbol{\gamma} - \mathbf{Xb}) + (\mathbf{X}\boldsymbol{\gamma} - \mathbf{Xb})'(\mathbf{X}\boldsymbol{\gamma} - \mathbf{Xb}) =$$

$$= (\mathbf{Y} - \mathbf{Xb})'(\mathbf{Y} - \mathbf{Xb}) - (\mathbf{X}(\boldsymbol{\gamma} - \mathbf{b}))'(\mathbf{Y} - \mathbf{Xb}) - (\mathbf{Y} - \mathbf{Xb})'\mathbf{X}(\boldsymbol{\gamma} - \mathbf{b}) + (\mathbf{X}(\boldsymbol{\gamma} - \mathbf{b}))'\mathbf{X}(\boldsymbol{\gamma} - \mathbf{b}) =$$

$$= (\mathbf{Y} - \mathbf{Xb})'(\mathbf{Y} - \mathbf{Xb}) - (\boldsymbol{\gamma} - \mathbf{b})' \underbrace{\mathbf{X'}(\mathbf{Y} - \mathbf{Xb})}_{\mathbf{0}} - \underbrace{(\mathbf{Y} - \mathbf{Xb})'\mathbf{X}}_{\mathbf{0'}}(\boldsymbol{\gamma} - \mathbf{b}) + \underbrace{(\boldsymbol{\gamma} - \mathbf{b})'\mathbf{X'X}(\boldsymbol{\gamma} - \mathbf{b})}_{\geq 0} \geq (\mathbf{Y} - \mathbf{Xb})'(\mathbf{Y} - \mathbf{Xb}).$$

Zostáva ešte dokázať, že $(\mathbf{Y} - \mathbf{Xb})'(\mathbf{Y} - \mathbf{Xb})$ je rovnaké pre každé riešenie $\mathbf{b}$ normálnych rovníc $\mathbf{X'Xb} = \mathbf{X'Y}$. Podľa Vety 3.13 práve všetky riešenia normálnych rovníc sú $\{(\mathbf{X'X})^-_* \mathbf{X'Y} + (\mathbf{I} - (\mathbf{X'X})^-_* \mathbf{X'X})\mathbf{z} : \ \mathbf{z} \in \mathcal{R}^k\}$ ($(\mathbf{X'X})^-_*$ je ľubovoľná, ale pevná g-inverzia matice $\mathbf{X'X}$). Pre každé riešenie $\mathbf{b}$ normálnych rovníc je

$$(\mathbf{Y} - \mathbf{Xb})'(\mathbf{Y} - \mathbf{Xb}) = \mathbf{Y'Y} - \mathbf{Y'Xb} - \ \mathbf{b'X'Y} + \mathbf{b'} \underbrace{\mathbf{X'Xb}}_{\mathbf{X'Y}} = \mathbf{Y'Y} - \mathbf{Y'Xb} =$$

$$= \mathbf{Y'Y} - \mathbf{Y'X}\{(\mathbf{X'X})^-_* \mathbf{X'Y} + (\mathbf{I} - (\mathbf{X'X})^-_* \mathbf{X'X})\mathbf{z}\} = \mathbf{Y'Y} - \mathbf{Y'X}(\mathbf{X'X})^-_* \mathbf{X'Y}, \tag{29}$$

lebo podľa Lemy 3.11 je $-\mathbf{Y'X} + \mathbf{Y'} \underbrace{\mathbf{X}(\mathbf{X'X})^-_* \mathbf{X'X}}_{\mathbf{X}} = \mathbf{0}$. Podľa Lemy 3.12 (iii) výraz (29) nezáleží od voľby $(\mathbf{X'X})^-_*$ a preto $(\mathbf{Y} - \mathbf{Xb})'(\mathbf{Y} - \mathbf{Xb})$ je rovnaké pre každé riešenie $\mathbf{b}$ normálnych rovníc. Q.E.D.

Veta 3.16: Sústava normálnych rovníc **X'Xb** = **X'Y** je ekvivalentná sústave

$$\left.\frac{\partial(\mathbf{Y}-\mathbf{X}\boldsymbol{\gamma})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\gamma})}{\partial\gamma_l}\right|_{\boldsymbol{\gamma}=\mathbf{b}} = 0, \qquad l = 1,2,...,k.$$

Dôkaz: Platí

$$(\mathbf{Y}-\mathbf{X}\boldsymbol{\gamma})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\gamma}) = \sum_{i=1}^{n}(Y_i - \sum_{j=1}^{k}\{\mathbf{X}\}_{ij}\gamma_j)^2,$$

čiže

$$\left.\frac{\partial(\mathbf{Y}-\mathbf{X}\boldsymbol{\gamma})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\gamma})}{\partial\gamma_l}\right|_{\boldsymbol{\gamma}=\mathbf{b}} = \left.\frac{\partial}{\partial\gamma_l}\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{k}\{\mathbf{X}\}_{ij}\gamma_j)^2\right|_{\boldsymbol{\gamma}=\mathbf{b}} =$$

$$= \left.\frac{\partial}{\partial\gamma_l}\sum_{i=1}^{n}(Y_i - (\{\mathbf{X}\}_{i1}\gamma_1 + \{\mathbf{X}\}_{i2}\gamma_2 + ... + \{\mathbf{X}\}_{ik}\gamma_k))^2\right|_{\boldsymbol{\gamma}=\mathbf{b}} = 0, \qquad l = 1,2,...,k,$$

teda

$$\left.2\sum_{i=1}^{n}(Y_i - (\{\mathbf{X}\}_{i1}\gamma_1 + \{\mathbf{X}\}_{i2}\gamma_2 + ... + \{\mathbf{X}\}_{ik}\gamma_k))(-\{\mathbf{X}\}_{il})\right|_{\boldsymbol{\gamma}=\mathbf{b}} = 0, \qquad l = 1,2,...,k,$$

$$\left.\sum_{i=1}^{n}(Y_i - (\{\mathbf{X}\}_{i1}\gamma_1 + \{\mathbf{X}\}_{i2}\gamma_2 + ... + \{\mathbf{X}\}_{ik}\gamma_k))\{\mathbf{X}\}_{il}\right|_{\boldsymbol{\gamma}=\mathbf{b}} = 0, \qquad l = 1,2,...,k,$$

$$(\mathbf{Y}-\mathbf{X}\mathbf{b})'\mathbf{X} = \mathbf{0},$$

$$\mathbf{X'}(\mathbf{Y}-\mathbf{X}\mathbf{b}) = \mathbf{0}$$

$$\mathbf{X'Y} = \mathbf{X'Xb},$$

čo sú normálne rovnice. Q.E.D.

Z Vety 3.15 a Vety 3.16 dostávame

Dôsledok 3.17: V LRM $(\mathbf{Y}_{n,1}, \mathbf{X}_{n,k}\boldsymbol{\beta}_{k,1}, \sigma^2\mathbf{I})$ neúplnej hodnosti, teda $h(\mathbf{X}) < k \leq n$ platí

$$\mathbf{b} = argmin_{\boldsymbol{\gamma}\in\mathcal{R}^k}(\mathbf{Y}-\mathbf{X}\boldsymbol{\gamma})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\gamma}) = argmin_{\boldsymbol{\gamma}\in\mathcal{R}^k}||\mathbf{Y}-\mathbf{X}\boldsymbol{\gamma}||^2 \iff \mathbf{X'Y} = \mathbf{X'Xb}.$$

Označme $\hat{\mathbf{Y}} = \mathbf{Xb}$, kde $\mathbf{b}$ je ľubovoľné riešenie normálnych rovníc $\mathbf{X'Y} = \mathbf{X'Xb}$. Potom

$$\hat{\mathbf{Y}} = \mathbf{Xb} = \mathbf{X}\left\{(\mathbf{X'X})^-_*\mathbf{X'Y} + (\mathbf{I} - (\mathbf{X'X})^-_*\mathbf{X'X})\mathbf{z}\right\} = \mathbf{X}(\mathbf{X'X})^-_*\mathbf{X'Y} \tag{30}$$

pre jednu pevne vybratú $(\mathbf{X'X})^-_*$. Podľa Lemy 3.11 (iii) vzťah (30), teda $\hat{\mathbf{Y}}$ nezávisí od voľby g-inverzie $(\mathbf{X'X})^-$ a je jediné. $\hat{\mathbf{Y}}$ je *ortogonálnou projekciou* $\mathbf{Y}$ na priestor $\mathcal{M}(\mathbf{X})$. Platí totiž, že $\hat{\mathbf{Y}} \in \mathcal{M}(\mathbf{X})$ a

$$\forall \mathbf{x} \in \mathcal{M}(\mathbf{X}) \quad \text{je} \quad (\mathbf{Y}-\hat{\mathbf{Y}})'\mathbf{x} = (\mathbf{Y}-\mathbf{Xb})'\mathbf{Xu} = \underbrace{(\mathbf{Y'X} - \mathbf{b'X'X})}_{\mathbf{0}}\mathbf{u} = 0,$$

teda $\mathbf{Y} - \hat{\mathbf{Y}} \perp \mathcal{M}(\mathbf{X})$ a $\hat{\mathbf{Y}}$ je taký prvok z $\mathcal{M}(\mathbf{X})$, ktorý je najbližšie k $\mathbf{Y}$.

Definícia 3.18: Nech $\mathbf{b}$ je ľubovoľné riešenie normálnych rovníc.

$$S_e = (\mathbf{Y} - \mathbf{Xb})'(\mathbf{Y} - \mathbf{Xb}) = ||\mathbf{Y} - \mathbf{Xb}||^2 = \mathbf{Y'Y} - \mathbf{Y'Xb} = \mathbf{Y'(I} - \mathbf{X(X'X)^{-}X')Y}$$

je *reziduálny súčet štvorcov (reziduální součet čtverců) RSČ.*

3.4 Odhad skalárnej parametrickej funkcie parametra $\boldsymbol{\beta}$ vo všeobecnom lineárnom modeli

Definícia 3.19: Majme LRM $(\mathbf{Y}_{n,1}, \mathbf{X}_{n,k}\boldsymbol{\beta}_{k,1}, \mathbf{V})$. Povieme, že (skalárna) parametrická funkcia $\theta = \theta(\boldsymbol{\beta}) = \mathbf{c'}\boldsymbol{\beta}$, $(\mathbf{c} \in \mathcal{R}^k$ pevne daný vektor) je *nestranne lineárne odhadnuteľná*, ak existuje jej lineárny nestranný odhad $a + \mathbf{u'Y}$, t.j. ak

$$\exists a \in \mathcal{R}, \mathbf{u} \in \mathcal{R}^n, \text{ že } \mathcal{E}_{\boldsymbol{\beta}}(a + \mathbf{u'Y}) = \mathbf{c'}\boldsymbol{\beta} \;\; \forall \boldsymbol{\beta} \in \mathcal{R}^k.$$

Podotýkame len, že odhadujeme $\mathbf{c'}\boldsymbol{\beta}$ pomocou observačného vektora $\mathbf{Y}$ lineárne, t.j. lineárnou funkciou náhodného vektora $\mathbf{Y}$, teda lineárny odhad je tvaru $a + \mathbf{u'Y}$.

Veta 3.20: V LRM $(\mathbf{Y}_{n,1}, \mathbf{X}_{n,k}\boldsymbol{\beta}_{k,1}, \mathbf{V})$ je parametrická funkcia $\theta = \mathbf{c'}\boldsymbol{\beta}$ lineárne nestranne odhadnuteľná práve vtedy ak $\mathbf{c} \in \mathcal{M}(\mathbf{X'})$ (t.j. ak $\exists\ \mathbf{w} \in \mathcal{R}^n$, že $\mathbf{c} = \mathbf{X'w}$).

Dôkaz: $\theta = \mathbf{c'}\boldsymbol{\beta}$ je lineárne nestranne odhadnuteľná $\iff$ $\exists\ a \in \mathcal{R}, \mathbf{u} \in \mathcal{R}^n$, že $\mathcal{E}_{\boldsymbol{\beta}}(a + \mathbf{u'Y}) = \mathbf{c'}\boldsymbol{\beta}\ \ \forall \boldsymbol{\beta} \in \mathcal{R}^k$ $\iff$ $\exists\ a \in \mathcal{R}, \mathbf{u} \in \mathcal{R}^n$, že $a + \mathbf{u'X}\boldsymbol{\beta} = \mathbf{c'}\boldsymbol{\beta}\ \ \forall \boldsymbol{\beta} \in \mathcal{R}^k$ $\iff$ $a = 0$ a $\exists\ \mathbf{u} \in \mathcal{R}^n$, že $\mathbf{u'X} = \mathbf{c'}$ $\iff$ $\exists \mathbf{u} \in \mathcal{R}^n$, že $\mathbf{X'u} = \mathbf{c}$. Q.E.D.

Dôsledok 3.21: V LRM $(\mathbf{Y}_{n,1}, \mathbf{X}_{n,k}\boldsymbol{\beta}_{k,1}, \mathbf{V})$ je $\hat{\theta} = a + \mathbf{u'Y}$ nestranný lineárny odhad nestranne odhadnuteľnej parametrickej funkcie $\theta = \mathbf{c'}\boldsymbol{\beta}$ (t.j. $\mathbf{c} \in \mathcal{M}(\mathbf{X'})$), práve vtedy ak $a = 0$ a $\mathbf{X'u} = \mathbf{c}$.

Dôsledok 3.22: V LRM plnej hodnosti je každá funkcia $\theta = \theta(\boldsymbol{\beta}) = \mathbf{c'}\boldsymbol{\beta}$ lineárne nestranne odhadnuteľná, lebo $\mathcal{M}(\mathbf{X'}_{k,n})$ je podpriestor $\mathcal{R}^k$, pričom $h(\mathbf{X'}) = k$, teda $\mathcal{M}(\mathbf{X'}) = \mathcal{R}^k$.

Definícia 3.23: Majme LRM $(\mathbf{Y}_{n,1}, \mathbf{X}_{n,k}\boldsymbol{\beta}_{k,1}, \mathbf{V})$. Povieme, že $\hat{\theta} = \mathbf{u'Y}$ je *najlepší nestranný lineárny odhad (NNLO)* lineárne odhadnuteľnej skalárnej parametrickej funkcie $\theta = \theta(\boldsymbol{\beta}) = \mathbf{c'}\boldsymbol{\beta}$, $(\mathbf{c} \in \mathcal{M}(\mathbf{X'})$ je pevne daný vektor), ak $\hat{\theta}$ je lineárny nestranný odhad $\mathbf{c'}\boldsymbol{\beta}$ a pre každý iný lineárny nestranný odhad θ^* funkcie $\mathbf{c'}\boldsymbol{\beta}$ platí $\mathcal{D}(\theta^*) \geq \mathcal{D}(\hat{\theta})$.

Veta 3.24: V LRM $(\mathbf{Y}_{n,1}, \mathbf{X}_{n,k}\boldsymbol{\beta}_{k,1}, \sigma^2\mathbf{I})$, v ktorom $h(\mathbf{X}) < k \leq n$, nech $\theta = \mathbf{c'}\boldsymbol{\beta}$ $(\mathbf{c} \in \mathcal{R}^k$ je pevný vektor) je lineárne nestranne odhadnuteľná parametrická funkcia (t.j. $\mathbf{c} \in \mathcal{M}(\mathbf{X'})$). NNLO tejto funkcie je $\hat{\theta} = \mathbf{c'b}$, kde $\mathbf{b}$ je ľubovoľné riešenie normálnych rovníc $\mathbf{X'Xb} = \mathbf{X'Y}$. Odhad $\hat{\theta} = \mathbf{c'b}$ nezáleží na voľbe riešenia normálnych rovníc.

Dôkaz: Podľa Vety 3.20 je $\theta = \mathbf{c'}\boldsymbol{\beta}$ lineárne nestranne odhadnuteľná práve vtedy ak $\mathbf{c} \in \mathcal{M}(\mathbf{X'})$. Pretože podľa Lemy 3.3 je $\mathcal{M}(\mathbf{X'}) = \mathcal{M}(\mathbf{X'X})$, $\exists\ \mathbf{w} \in \mathcal{R}^k$, že $\mathbf{c} = \mathbf{X'Xw}$.

Odhad $\mathbf{c'b} = \mathbf{w'}\underbrace{\mathbf{X'Xb}} = \mathbf{w'}\underbrace{\mathbf{X'Y}} = \mathbf{u'Y}$ (pomocou normálnych rovníc) je lineárny odhad.

$\mathcal{E}(\mathbf{c'b}) = \mathcal{E}(\mathbf{w'X'Xb}) = \mathcal{E}(\mathbf{w'X'Y}) = \mathbf{w'X'X}\boldsymbol{\beta} = \mathbf{c'}\boldsymbol{\beta} \quad \forall\ \boldsymbol{\beta}$, teda $\mathbf{c'b}$ je lineárny nestranný odhad parametrickej funkcie $\theta = \mathbf{c'}\boldsymbol{\beta}$.

Ak sú $\mathbf{b}^*_{(1)}$ a $\mathbf{b}^*_{(2)}$ ľubovoľné dve riešenia normálnych rovníc $\mathbf{X'Xb} = \mathbf{X'Y}$, potom

$$\mathbf{c'b}^*_{(1)} - \mathbf{c'b}^*_{(2)} = \mathbf{w'X'X}(\mathbf{b}^*_{(1)} - \mathbf{b}^*_{(2)}) = \mathbf{w'}\{\mathbf{X'Y} - \mathbf{X'Y}\} = 0.$$

Teda $\hat{\theta} = \mathbf{c'b}$ nezáleží na voľbe riešenia normálnych rovníc $\mathbf{X'Xb} = \mathbf{X'Y}$.

Vezmime si ľubovoľný iný (iný) nestranný lineárny odhad funkcie $\mathbf{c'}\boldsymbol{\beta}$, a síce $\theta^* = \mathbf{v'Y}$. Z nestrannosti vyplýva

$$\mathcal{E}(\mathbf{v'Y}) = \mathbf{v'X}\boldsymbol{\beta} = \mathbf{c'}\boldsymbol{\beta} \quad \forall\ \boldsymbol{\beta} \in \mathcal{R}^k,$$

čiže

$$\mathbf{X'}\ \mathbf{v}\ = \mathbf{c}. \tag{31}$$

Počítajme

$$\mathcal{D}(\theta^*) - \mathcal{D}(\mathbf{c'b}) = \mathcal{D}(\mathbf{v'Y}) - \mathcal{D}(\underbrace{\mathbf{c'b}}_{\mathbf{w'X'Xb}=\mathbf{w'X'Y}}) = \sigma^2\mathbf{v'v} - \sigma^2\mathbf{w'X'Xw} = \sigma^2\mathbf{v'v} - \sigma^2\underbrace{\mathbf{w'X'X}}_{\mathbf{c'}}(\mathbf{X'X})^-\underbrace{\mathbf{X'Xw}}_{\mathbf{c}} =$$

$$= \sigma^2\{\mathbf{v'v} - \underbrace{\mathbf{v'X}}_{=\mathbf{c'}\ \text{z (31)}}(\mathbf{X'X})^-\mathbf{X'v}\} = \sigma^2\mathbf{v'}\{\underbrace{\mathbf{I} - \mathbf{X}(\mathbf{X'X})^-\mathbf{X'}}_{\mathbf{A}}\}\mathbf{v} \geq 0,$$

lebo $\mathbf{A}$ nezáleží na voľbe $(\mathbf{X'X})^-$, je jediná, symetrická a idempotentná. Preto $\mathbf{c'b}$ je NNLO nestranne lineárne odhadnuteľnej parametrickej funkcie $\theta = \mathbf{c'}\boldsymbol{\beta}$. Q.E.D.

3.5 Odhad vektorovej parametrickej funkcie parametra $\boldsymbol{\beta}$ vo všeobecnom lineárnom modeli

$\boldsymbol{\theta}_{p,1} = \mathbf{C'}\boldsymbol{\beta}$, kde $\mathbf{C'}_{p,k}$ je matica reálnych čísel, nazývame ($p-$rozmernou) vektorovou parametrickou funkciou.

Definícia 3.25: Majme LRM $(\mathbf{Y}_{n,1}, \mathbf{X}_{n,k}\boldsymbol{\beta}_{k,1}, \mathbf{V})$. Povieme, že vektorová parametrická funkcia $\boldsymbol{\theta}_{p,1} = \mathbf{C'}\boldsymbol{\beta}$, ($\mathbf{C'}$ je $p \times k$ reálna matica je *nestranne lineárne odhadnuteľná*, ak existuje jej lineárny nestranný odhad $\mathbf{a} + \mathbf{UY}$, t.j. ak

$$\exists \mathbf{a} \in \mathcal{R}^p, \mathbf{U}_{p,n}, \quad \text{že} \quad \mathcal{E}_{\boldsymbol{\beta}}(\mathbf{a} + \mathbf{UY}) = \mathbf{C'}\boldsymbol{\beta} \quad \forall \boldsymbol{\beta} \in \mathcal{R}^k.$$

Podotýkame len, že odhadujeme $\mathbf{C'}\boldsymbol{\beta}$ pomocou observačného vektora $\mathbf{Y}$ lineárne, t.j. $p-$ rozmernou lineárnou funkciou náhodného vektora $\mathbf{Y}$, teda odhad je tvaru $\mathbf{a} + \mathbf{UY}$.

Veta 3.26: V LRM $(\mathbf{Y}_{n,1}, \mathbf{X}_{n,k}\boldsymbol{\beta}_{k,1}, \mathbf{V})$ je vektorová parametrická funkcia $\boldsymbol{\theta}_{p,1} = \mathbf{C'}\boldsymbol{\beta}$ lineárne nestranne odhadnuteľná práve vtedy ak existuje $p \times n$ matica $\mathbf{U}$, že $\mathbf{C} = \mathbf{X'U'}$. Ak navyše platí $h(\mathbf{C'}) = p$, tak $h(\mathbf{U}) = p$.

Dôkaz: $\boldsymbol{\theta}_{p,1} = \mathbf{C'}\boldsymbol{\beta}$ je lineárne nestranne odhadnuteľná $\iff$ $\exists\ \mathbf{a} \in \mathcal{R}^p, \mathbf{U}_{p,n}$ že $\mathcal{E}_{\boldsymbol{\beta}}(\mathbf{a} + \mathbf{UY}) = \mathbf{C'}\boldsymbol{\beta} \quad \forall \boldsymbol{\beta} \in \mathcal{R}^k \iff \exists\ \mathbf{a} \in \mathcal{R}^p, \mathbf{U}_{p,n}$ že $\mathbf{a} + \mathbf{UX}\boldsymbol{\beta} = \mathbf{C'}\boldsymbol{\beta} \quad \forall \boldsymbol{\beta} \in \mathcal{R}^k \iff \mathbf{a} = \mathbf{0}$ a $\exists\ \mathbf{U}_{p,n}$, že $\mathbf{UX} = \mathbf{C'} \iff \exists \mathbf{U}_{p,n}$, že $\mathbf{X'U'} = \mathbf{C}$.

Nech navyše platí $h(\mathbf{C'}) = p$. Potom $p = h(\mathbf{C}) = h(\mathbf{X'U'}) \leq \min\{h(\mathbf{X'}), h(\mathbf{U'})\} \leq h(\mathbf{U}_{p,n}) \leq \min\{p, n\} \leq p$, teda $h(\mathbf{U}) = p$. Q.E.D.

Dôsledok 3.27: V LRM $(\mathbf{Y}_{n,1}, \mathbf{X}_{n,k}\boldsymbol{\beta}_{k,1}, \mathbf{V})$ je $\hat{\boldsymbol{\theta}} = \mathbf{a}_{p,1} + \mathbf{U}_{p,n}\mathbf{Y}$ nestranný lineárny odhad nestranne odhadnuteľnej parametrickej funkcie $\boldsymbol{\theta}_{p,1} = \mathbf{C'}\boldsymbol{\beta}$ (t.j. $\mathbf{C} = \mathbf{X'U'}$), práve vtedy ak $\mathbf{a} = \mathbf{0}$ a $\mathbf{X'U'} = \mathbf{C}$.

Lema 3.28: V LRM $(\mathbf{Y}_{n,1}, \mathbf{X}_{n,k}\boldsymbol{\beta}_{k,1}, \mathbf{V})$ je vektorová parametrická funkcia $\boldsymbol{\theta}_{p,1} = \mathbf{C'}\boldsymbol{\beta}$ lineárne nestranne odhadnuteľná práve vtedy ak existuje $k \times p$ matica $\mathbf{S}$, že $\mathbf{C} = \mathbf{X'XS}$. Ak navyše platí $h(\mathbf{C'}) = p$, tak $h(\mathbf{S}) = p$.

Dôkaz: Podľa Vety 3.26 $\boldsymbol{\theta}_{p,1} = \mathbf{C'}\boldsymbol{\beta}$ je nestranne lineárne odhadnuteľná $\iff$ $\exists\, \mathbf{U}_{p,n}$, že $\mathbf{C} = \mathbf{X'U'}$, teda $\{\mathbf{C}\}_{\bullet i} = \mathbf{X'}\{\mathbf{U'}\}_{\bullet,i}$, $\ i = 1,2,...,p$. Pretože podľa Lemy 3.3 je $\mathcal{M}(\mathbf{X'}) = \mathcal{M}(\mathbf{X'X})$ a $\{\mathbf{C}\}_{\bullet i} \in \mathcal{M}(\mathbf{X'})$, $\ i = 1,2,...,p$, $\ \exists\ \mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_p$, $\ \mathbf{s}_i \in \mathcal{R}^k$, že $\{\mathbf{C}\}_{\bullet i} = \mathbf{X'X}\mathbf{s}_i$, $\ i = 1,2,...,p$. Preto existuje matica $\mathbf{S}_{k,p} = (\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_p)$, že $\mathbf{C} = \mathbf{X'XS}$.

Ak navyše $h(\mathbf{C'}) = p$, potom $p = h(\mathbf{C}) = h(\mathbf{X'XS}) \le \min\{h(\mathbf{X'X}), h(\mathbf{S})\} \le h(\mathbf{S}_{k,p}) \le \min\{k,p\} \le p$, teda $h(\mathbf{S}) = p$. Q.E.D.

Poznámka 3.29: Je zrejmé (dokážte si), že vektorová parametrická funkcia $\boldsymbol{\theta}_{p,1} = (\theta_1, ..., \theta_p)'$ je nestranne lineárne odhadnuteľná práve vtedy, ak je nestranne lineárne odhadnuteľná každá jej zložka θ_i, $\ i = 1,2,...,p$.

Definícia 3.30: Majme LRM $(\mathbf{Y}_{n,1}, \mathbf{X}_{n,k}\boldsymbol{\beta}_{k,1}, \mathbf{V})$. Povieme, že $\hat{\boldsymbol{\theta}}_{p,1} = \mathbf{UY}$ je *najlepší nestranný lineárny odhad (NNLO)* lineárne odhadnuteľnej vektorovej parametrickej funkcie $\boldsymbol{\theta}_{p,1}(\boldsymbol{\beta}) = \mathbf{C'}\boldsymbol{\beta}$, $\ (\mathbf{C'}_{p,k}$ je pevne daná matica), ak $\hat{\boldsymbol{\theta}}$ je lineárny nestranný odhad $\mathbf{C'}\boldsymbol{\beta}$ a pre každý iný lineárny nestranný odhad $\boldsymbol{\theta}^*$ funkcie $\mathbf{C'}\boldsymbol{\beta}$ platí, že $cov(\boldsymbol{\theta}^*) - cov(\hat{\boldsymbol{\theta}})$ je pozitívne semidefinitná matica.

Veta 3.31: V LRM $(\mathbf{Y}_{n,1}, \mathbf{X}_{n,k}\boldsymbol{\beta}_{k,1}, \sigma^2\mathbf{I})$, v ktorom $h(\mathbf{X}) < k \le n$, nech $\boldsymbol{\theta}_{p,1}(\boldsymbol{\beta}) = \mathbf{C'}\boldsymbol{\beta}$ $\ (\mathbf{C'}_{p,k}$ je pevne daná matica) je lineárne nestranne odhadnuteľná vektorová parametrická funkcia (t.j. $\mathbf{C} = \mathbf{X'XS}$). NNLO tejto funkcie je $\hat{\boldsymbol{\theta}} = \mathbf{C'b}$, kde $\mathbf{b}$ je ľubovoľné riešenie normálnych rovníc $\mathbf{X'Xb} = \mathbf{X'Y}$. Odhad $\hat{\boldsymbol{\theta}} = \mathbf{C'b}$ nezáleží na voľbe riešenia normálnych rovníc.

Dôkaz: Podľa Lemy 3.28 je $\boldsymbol{\theta} = \mathbf{C'}\boldsymbol{\beta}$ lineárne nestranne odhadnuteľná práve vtedy ak $\mathbf{C} = \mathbf{X'XS}$.

Odhad $\mathbf{C'b} = \mathbf{S'}\underbrace{\mathbf{X'Xb}} = \mathbf{S'}\underbrace{\mathbf{X'Y}} = (\mathbf{S'X'})\mathbf{Y}$ (pomocou normálnych rovníc) je lineárny odhad.

$\mathcal{E}(\mathbf{C'b}) = \mathcal{E}(\mathbf{S'X'Xb}) = \mathcal{E}(\mathbf{S'X'Y}) = \mathbf{S'X'X}\boldsymbol{\beta} = \mathbf{C'}\boldsymbol{\beta}\ \ \forall\ \boldsymbol{\beta}$, teda $\mathbf{C'b}$ je lineárny nestranný odhad parametrickej funkcie $\boldsymbol{\theta} = \mathbf{C'}\boldsymbol{\beta}$.

Ak sú $\mathbf{b}^*_{(1)}$ a $\mathbf{b}^*_{(2)}$ ľubovoľné dve riešenia normálnych rovníc $\mathbf{X'Xb} = \mathbf{X'Y}$, potom

$$\mathbf{C'b}^*_{(1)} - \mathbf{C'b}^*_{(2)} = \mathbf{S'X'X}(\mathbf{b}^*_{(1)} - \mathbf{b}^*_{(2)}) = \mathbf{S'}\{\mathbf{X'Y} - \mathbf{X'Y}\} = \mathbf{0}.$$

Teda $\hat{\boldsymbol{\theta}} = \mathbf{C'b}$ nezáleží na voľbe riešenia normálnych rovníc $\mathbf{X'Xb} = \mathbf{X'Y}$.

Vezmime si ľubovoľný iný (iný) nestranný lineárny odhad funkcie $\mathbf{C'}\boldsymbol{\beta}$, a síce $\boldsymbol{\theta}^* = \mathbf{V}_{p,n}\mathbf{Y}$. Z nestrannosti vyplýva

$$\mathcal{E}(\mathbf{VY}) = \mathbf{VX}\boldsymbol{\beta} = \mathbf{C'}\boldsymbol{\beta}\quad \forall\ \boldsymbol{\beta} \in \mathcal{R}^k,$$

čiže

$$\mathbf{V\ X} = \mathbf{C'}. \tag{32}$$

Matica

$$cov(\boldsymbol{\theta}^*)-cov(\hat{\boldsymbol{\theta}}) = cov(\mathbf{VY})-cov(\underbrace{\mathbf{C'b}}_{\mathbf{S'X'Xb=S'X'Y}}) = \sigma^2\mathbf{VV'}-\sigma^2\mathbf{S'X'XS} = \sigma^2\mathbf{VV'}-\sigma^2\underbrace{\mathbf{S'X'X}}_{\mathbf{C'}}(\mathbf{X'X})^-\underbrace{\mathbf{X'Xw}}_{\mathbf{C}} =$$

$$= \sigma^2\{\mathbf{VV'} - \underbrace{\mathbf{VX}}_{=\mathbf{C'}\text{ z (32)}}(\mathbf{X'X})^-\mathbf{X'V'}\} = \sigma^2\mathbf{V}\{\underbrace{\mathbf{I}-\mathbf{X}(\mathbf{X'X})^-\mathbf{X'}}_{\mathbf{A}}\}\mathbf{V'} = \sigma^2\mathbf{VA}(\mathbf{VA})'$$

je pozitívne semidefinitná, lebo $\mathbf{A}$ nezáleží na voľbe $(\mathbf{X'X})^-$, je jediná, symetrická a idempotentná. Preto $\mathbf{C'b}$ je NNLO nestranne lineárne odhadnuteľnej parametrickej funkcie $\boldsymbol{\theta} = \mathbf{C'}\boldsymbol{\beta}$. Q.E.D.

Poznámka 3.32: Je zrejmé, že $\mathcal{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ je nestranne lineárne odhadnuteľná vektorová parametrická funkcia. Ak $cov(\mathbf{Y}) = \sigma^2\mathbf{I}$, tak NNLO $\widehat{\mathcal{E}(\mathbf{Y})} \overset{\text{ozn.}}{=} \hat{\mathbf{Y}} = \mathbf{Xb}$, kde $\mathbf{b}$ je ľubovoľné riešenie normálnych rovníc ($\mathbf{b} = (\mathbf{X'X})^-_*\mathbf{X'Y} + (\mathbf{I} - (\mathbf{X'X})^-_*\mathbf{X'X})\mathbf{z}$), teda $\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X'X})^-\mathbf{X'Y}$ (nezáleží na voľbe $(\mathbf{X'X})^-$) je jediný.

3.6 Testy hypotéz v LRM s neúplnou hodnosťou a $cov(\mathbf{Y}) = \sigma^2\mathbf{I}$

Majme LRM $(\mathbf{Y}_{n,1}, \mathbf{X}_{n,k}\boldsymbol{\beta}_{k,1}, \sigma^2\mathbf{I})$, v ktorom $h(\mathbf{X}) = r < k \le n$. Uvažujme vektorovú parametrickú funkciu $\boldsymbol{\theta}_{p,1}(\boldsymbol{\beta}) = \mathbf{C'}_{p,k}\boldsymbol{\beta}$, $\mathbf{C}_{k,p}$ je reálna matica, $h(\mathbf{C}) = p$. $\boldsymbol{\theta}$ nech je lineárne nestranne odhadnuteľná, t.j.

$$\exists\ \mathbf{U}_{p,n}, \quad \mathbf{C} = \mathbf{X'U'}, \quad h(\mathbf{U}) = p$$

alebo

$$\exists\ \mathbf{S}_{k,p}, \quad \mathbf{C} = \mathbf{X'XS}, \quad h(\mathbf{S}) = p.$$

Pre maticu $\mathbf{XS}$ typu $n \times p$ platí

$$p = h(\mathbf{C}) \le \min\{h(\mathbf{X'}), h(\mathbf{XS})\} \le h(\mathbf{XS}) \le min\{n, p\} \le p,$$

čiže

$$h(\mathbf{XS}) = p. \tag{33}$$

Nech $\mathbf{b}$ je ľubovoľné riešenie normálnych rovníc, tak NNLO lineárne nestranne odhadnuteľnej vektorovej parametrickej funkcie $\boldsymbol{\theta}(\boldsymbol{\beta}) = \mathbf{C'}\boldsymbol{\beta}$ je

$$\hat{\boldsymbol{\theta}} = \mathbf{C'b} = \mathbf{C'}\{(\mathbf{X'X})^-_*\mathbf{X'Y} + (\mathbf{I} - (\mathbf{X'X})^-_*\mathbf{X'X})\mathbf{z}\} = \mathbf{UX}\{(\mathbf{X'X})^-_*\mathbf{X'Y} + (\mathbf{I} - (\mathbf{X'X})^-_*\mathbf{X'X})\mathbf{z}\} =$$

$$= \mathbf{UX}(\mathbf{X'X})^-_*\mathbf{X'Y} = \mathbf{UX}(\mathbf{X'X})^-\mathbf{X'Y} = \mathbf{C'}(\mathbf{X'X})^-\mathbf{X'Y}, \tag{34}$$

pričom tento odhad nezávisí od voľby $(\mathbf{X'X})^-$ a je jediný (pozri Vetu 3.31).

Veta 3.33: Majme LRM $(\mathbf{Y}_{n,1}, \mathbf{X}_{n,k}\boldsymbol{\beta}_{k,1}, \sigma^2\mathbf{I})$, v ktorom $h(\mathbf{X}) = r < k \le n$, pričom $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$. $\mathbf{C'}_{p,k}$ je reálna matica s hodnosťou $h(\mathbf{C'}) = p$ a $\boldsymbol{\theta}(\boldsymbol{\beta}) = \mathbf{C'}_{p,k}\boldsymbol{\beta}$ je lineárne nestranne odhadnuteľná vektorová parametrická funkcia. Platí

(i) $\mathbf{b} = (\mathbf{X'X})^-\mathbf{X'Y} \sim N((\mathbf{X'X})^-\mathbf{X'X}\boldsymbol{\beta}, \sigma^2(\mathbf{X'X})^-\mathbf{X'X}[(\mathbf{X'X})^-]')$,

(ii) ak $\mathbf{b}$ je ľubovoľné riešenie normálnych rovníc, tak $\hat{\boldsymbol{\theta}} = \mathbf{C'b} \sim N(\mathbf{C'}\boldsymbol{\beta}, \sigma^2\mathbf{C'}(\mathbf{X'X})^-\mathbf{C})$,
pričom $\mathbf{C'}(\mathbf{X'X})^-\mathbf{C}$ nezáleží na voľbe $(\mathbf{X'X})^-$ a je regulárna, t.j. $h(\mathbf{C'}(\mathbf{X'X})^-\mathbf{C}) = p$.

Dôkaz: (i) je zrejmé,

(ii) $\exists\, \mathbf{S}_{k,p}$, že $\mathbf{C} = \mathbf{X'XS}$, pričom $h(\mathbf{XS}) = p$. Preto $\hat{\boldsymbol{\theta}} = \mathbf{C'b} = \mathbf{S'}\underbrace{\mathbf{X'Xb}} =$

$= \mathbf{S'}\underbrace{\mathbf{X'Y}} \sim N(\underbrace{\mathbf{S'X'X}}_{\mathbf{C'}}\boldsymbol{\beta}, \sigma^2\mathbf{S'X'XS})$, kde $\mathbf{S'X'XS} = \mathbf{S'X'X}\underbrace{(\mathbf{X'X})^-}_{\text{nezáleží na výbere}}\mathbf{X'XS} = \mathbf{C'}(\mathbf{X'X})^-\mathbf{C}$ je $p \times p$ matica a platí $h(\mathbf{C'}(\mathbf{X'X})^-\mathbf{C}) = h(\mathbf{S'X'XS}) = h(\mathbf{XS}) = p$, teda $\mathbf{C'}(\mathbf{X'X})^-\mathbf{C}$ je regulárna. Q.E.D.

Veta 3.34: Reziduálny súčet štvorcov je

$$S_e = (\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}}) = (\mathbf{Y} - \mathbf{Xb})'(\mathbf{Y} - \mathbf{Xb}) = \mathbf{Y'}(\mathbf{I} - \mathbf{X}(\mathbf{X'X})^-\mathbf{X'})\mathbf{Y}$$

($\mathbf{Xb}$ je jediné pre akékoľvek riešenie $\mathbf{b}$ normálnych rovníc). Ak navyše $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, $h(\mathbf{X}) = r < n$, tak

$$\frac{S_e}{\sigma^2} \sim \chi^2_{n-r}.$$

Dôkaz: $S_e = (\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}}) = (\mathbf{Y} - \mathbf{Xb})'(\mathbf{Y} - \mathbf{Xb}) = \mathbf{Y'Y} - \mathbf{Y'Xb} - \mathbf{b'X'Y} + \mathbf{b'}\underbrace{\mathbf{X'Xb}}_{\mathbf{X'Y}} = \mathbf{Y'Y} - \mathbf{Y'X}(\mathbf{X'X})^-\mathbf{X'Y} = \mathbf{Y'}(\mathbf{I} - \mathbf{X}(\mathbf{X'X})^-\mathbf{X'})\mathbf{Y}$. Pretože $(\mathbf{I} - \mathbf{X}(\mathbf{X'X})^-\mathbf{X'})\mathbf{X}\boldsymbol{\beta} = \mathbf{0}$, môžeme písať

$$\frac{S_e}{\sigma^2} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'\frac{\mathbf{I} - \mathbf{X}(\mathbf{X'X})^-\mathbf{X'}}{\sigma^2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}).$$

Opäť podľa Vety 13 kapitoly V. knihy Anděl, J., *Matematická statistika*, SNTL, Praha, 1985 (dokazovali sme ju aj na prednáške Lineární statistické modely I), v prípade, že $\frac{\mathbf{I} - \mathbf{X}(\mathbf{X'X})^-\mathbf{X'}}{\sigma^2}$ je symetrická, p.s.d. matica, $\frac{\mathbf{I} - \mathbf{X}(\mathbf{X'X})^-\mathbf{X'}}{\sigma^2}cov(\mathbf{Y}) \neq \mathbf{0}$ a idempotentná, tak $(\mathbf{Y} - \mathcal{E}(\mathbf{Y}))'\frac{\mathbf{I} - \mathbf{X}(\mathbf{X'X})^-\mathbf{X'}}{\sigma^2}(\mathbf{Y} - \mathcal{E}(\mathbf{Y})) \sim \chi^2_{tr[\frac{\mathbf{I} - \mathbf{X}(\mathbf{X'X})^-\mathbf{X'}}{\sigma^2}cov(\mathbf{Y})]}$. Overme si všetky predpoklady Vety 13.

Matica $\mathbf{I} - \mathbf{X}(\mathbf{X'X})^-\mathbf{X'}$ je podľa Lemy 3.12 (iii) symetrická. Ľahko vidíme, že je idempotentná, teda to je p.s.d matica. Potom ale aj matica $\frac{\mathbf{I} - \mathbf{X}(\mathbf{X'X})^-\mathbf{X'}}{\sigma^2}$ je symetrická a p.s.d. matica. Ďalej

$$\frac{\mathbf{I} - \mathbf{X}(\mathbf{X'X})^-\mathbf{X'}}{\sigma^2}cov(\mathbf{Y}) = \frac{\mathbf{I} - \mathbf{X}(\mathbf{X'X})^-\mathbf{X'}}{\sigma^2}\sigma^2\mathbf{I} = \mathbf{I} - \mathbf{X}(\mathbf{X'X})^-\mathbf{X'}$$

je idempotentná a nenulová, lebo

$$h(\mathbf{I} - \mathbf{X}(\mathbf{X'X})^-\mathbf{X'}) = tr(\mathbf{I} - \mathbf{X}(\mathbf{X'X})^-\mathbf{X'}) = tr\mathbf{I}_{n,n} - tr[\mathbf{X'X}(\mathbf{X'X})^-] = n - h(\mathbf{X}) = n - r > 0 \qquad (35)$$

(z predpokladov). Ináč pre ľubovoľnú maticu $\mathbf{D}$ je hodnosť $h(\mathbf{D}) = tr\mathbf{DD}^-$, lebo $\mathbf{DD}^-$ je idempotentná, teda $h(\mathbf{DD}^-) = tr[\mathbf{DD}^-]$ a $h(\mathbf{D}) \geq h(\mathbf{DD}^-) \geq h(\mathbf{DD}^-\mathbf{D}) = h(\mathbf{D})$, čiže $h(\mathbf{D}) = h(\mathbf{DD}^-) = tr[\mathbf{DD}^-]$. Okrem toho z Lemy 3.3 vyplýva, že $\mathcal{M}(\mathbf{X'X}) = \mathcal{M}(\mathbf{X'})$, teda $h(\mathbf{X'X}) = h(\mathbf{X'}) = h(\mathbf{X})$. Q.E.D.

Veta 3.35: V LRM $(\mathbf{Y}_{n,1}, \mathbf{X}_{n,k}\boldsymbol{\beta}_{k,1}, \sigma^2\mathbf{I})$, v ktorom $h(\mathbf{X}) = r < k \leq n$ je

$$s^2 = \frac{S_e}{n-r}$$

nestranným odhadom σ^2. (Nevyžaduje sa normalita $\mathbf{Y}$.)

Dôkaz: $S_e = \mathbf{Y'}(\mathbf{I} - \mathbf{X}(\mathbf{X'X})^-\mathbf{X'})\mathbf{Y}$ je kvadratická forma náhodného vektora $\mathbf{Y}$. Platí

$$\mathcal{E}\left(\frac{S_e}{n-r}\right) = \frac{1}{n-r}\mathcal{E}(S_e) = \frac{1}{n-r}\mathcal{E}(\mathbf{Y'}(\mathbf{I} - \mathbf{X}(\mathbf{X'X})^-\mathbf{X'})\mathbf{Y}) =$$

$$= \frac{1}{n-r}\{\boldsymbol{\beta}'\mathbf{X'}\underbrace{(\mathbf{I}-\mathbf{X}(\mathbf{X'X})^{-}\mathbf{X'})\mathbf{X}}_{\mathbf{0}}\boldsymbol{\beta} + tr[(\mathbf{I}-\mathbf{X}(\mathbf{X'X})^{-}\mathbf{X'})\sigma^2\mathbf{I}]\} = \frac{\sigma^2}{n-r}tr(\mathbf{I}-\mathbf{X}(\mathbf{X'X})^{-}\mathbf{X'}) = \sigma^2$$

(z (35)). Q.E.D.

Veta 3.36: Majme LRM $(\mathbf{Y}_{n,1}, \mathbf{X}_{n,k}\boldsymbol{\beta}_{k,1}, \sigma^2\mathbf{I})$, $h(\mathbf{X}) < k$. Ak $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, tak s^2 a $\mathbf{b} = (\mathbf{X'X})^{-}\mathbf{X'Y}$ sú nezávislé.

Dôkaz: $\mathbf{b} = (\mathbf{X'X})^{-}\mathbf{X'Y}$ a $s^2 = \dfrac{S_e}{n-h(\mathbf{X})} = \mathbf{Y'}\dfrac{\mathbf{I}-\mathbf{X}(\mathbf{X'X})^{-}\mathbf{X'}}{n-h(\mathbf{X})}\mathbf{Y}$. Podľa Vety 16, str. 81 knihy Anděl, J., *Matematická statistika*, SNTL, Praha, 1985 (dokazovali sme si ju na prednáške Lineární statistické modely I), sú s^2 a $\mathbf{b}$ nezávislé ak

$$(\mathbf{X'X})^{-}\mathbf{X'}[\sigma^2\mathbf{I}]\frac{\mathbf{I}-\mathbf{X}(\mathbf{X'X})^{-}\mathbf{X'}}{n-h(\mathbf{X})} = \mathbf{0},$$

o čom sa ľahko presvedčíme. Q.E.D.

Veta 3.37: Majme LRM $(\mathbf{Y}_{n,1}, \mathbf{X}_{n,k}\boldsymbol{\beta}_{k,1}, \sigma^2\mathbf{I})$, v ktorom $h(\mathbf{X}) = r < k \leq n$, pričom $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$. $\mathbf{C'}_{p,k}$ je reálna matica s hodnosťou $h(\mathbf{C'}) = p$, $\boldsymbol{\theta}(\boldsymbol{\beta}) = \mathbf{C'}_{p,k}\boldsymbol{\beta}$ je lineárne nestranne odhadnuteľná vektorová parametrická funkcia, $\mathbf{b} = (\mathbf{X'X})^{-}\mathbf{X'Y}$ a $s^2 = \dfrac{S_e}{n-r}$. Ak platí $\mathbf{C'}\boldsymbol{\beta} = \mathbf{0}$, tak

$$F = \frac{(\mathbf{C'b})'[\mathbf{C'}(\mathbf{X'X})^{-}\mathbf{C}]^{-1}\mathbf{C'b}}{ps^2} \sim F_{p,n-r}.$$

Dôkaz: Podľa Vety 3.36 sú $\mathbf{b} = (\mathbf{X'X})^{-}\mathbf{X'Y}$ a s^2 nezávislé, teda aj $\mathbf{C'b}$ a s^2 sú nezávislé, pričom podľa Vety 3.33 (ii) je $\mathbf{C'b} \sim N(\mathbf{C'}\boldsymbol{\beta}, \sigma^2\mathbf{C'}(\mathbf{X'X})^{-}\mathbf{C})$ a $\mathbf{C'}(\mathbf{X'X})^{-}\mathbf{C}$ je regulárna $p \times p$ matica. Ale potom aj $(\mathbf{C'b}-\mathbf{C'}\boldsymbol{\beta})'\frac{1}{\sigma^2}[\mathbf{C'}(\mathbf{X'X})^{-}\mathbf{C}]^{-1}(\mathbf{C'b}-\mathbf{C'}\boldsymbol{\beta})$ (ako funkcia $\mathbf{C'b}$) je nezávislá so s^2, pričom $(\mathbf{C'b}-\mathbf{C'}\boldsymbol{\beta})'\frac{1}{\sigma^2}[\mathbf{C'}(\mathbf{X'X})^{-}\mathbf{C}]^{-1}(\mathbf{C'b}-\mathbf{C'}\boldsymbol{\beta}) \sim \chi^2_p$ (podľa Vety 12, str. 79 v knihe Anděl, J., *Matematická statistika*, SNTL, Praha, 1985, dokazovali sme si ju na prednáške Lineární statistické modely I).

Podľa Vety 3.34 je $\dfrac{S_e}{\sigma^2} \sim \chi^2_{n-r}$ a preto $\dfrac{n-r}{\sigma^2}s^2 = \dfrac{Se}{\sigma^2} \sim \chi^2_{n-r}$ a je nezávislá s náhodnou veličinou $(\mathbf{C'b}-\mathbf{C'}\boldsymbol{\beta})'\frac{1}{\sigma^2}[\mathbf{C'}(\mathbf{X'X})^{-}\mathbf{C}]^{-1}(\mathbf{C'b}-\mathbf{C'}\boldsymbol{\beta}) \sim \chi^2_p$. Preto ak $\mathbf{C'}\boldsymbol{\beta} = \mathbf{0}$, tak

$$F = \frac{\dfrac{(\mathbf{C'b})'\frac{1}{\sigma^2}[\mathbf{C'}(\mathbf{X'X})^{-}\mathbf{C}]^{-1}\mathbf{C'b}}{p}}{\dfrac{\frac{n-r}{\sigma^2}s^2}{n-r}} = \frac{(\mathbf{C'b})'[\mathbf{C'}(\mathbf{X'X})^{-}\mathbf{C}]^{-1}\mathbf{C'b}}{ps^2} \sim F_{p,n-r}.$$

Q.E.D.

Poznámka 3.38: Vetu 3.37 použijeme pri testovaní

$$H_0: \quad \mathbf{C'}\boldsymbol{\beta} = \mathbf{0} \qquad \times \qquad H_1: \quad \mathbf{C'}\boldsymbol{\beta} \neq \mathbf{0}$$

Ak realizácia $F_{real} > F_{p,n-r}(1-\alpha)$, tak H_0 zamietame na hladine významnosti α (porušenie hypotézy má za následok veľké hodnoty F).

4 Testovanie submodelov

Majme LRM

$$\mathbf{Y}_{n,1} = \mathbf{X}_{n,k}\boldsymbol{\beta}_{k,1} + \boldsymbol{\epsilon}_{n,1}.$$

Je možné, že stredná hodnota náhodného observačného vektora nezáleží od jedného alebo niekoľkých parametrov $\beta_{j_1}, ..., \beta_{j_s}$, teda závisí len od vektora $\boldsymbol{\beta}^* = (\beta_{s_1}, ..., \beta_{s_{k_1}})'$, $k_1 < k$. Prechádzame k "zjednodušenému" modelu

$$\mathbf{Y}_{n,1} = \mathbf{X}^*_{n,k_1}\boldsymbol{\beta}^*_{k_1,1} + \boldsymbol{\epsilon}^*_{n,1}.$$

Všeobecnejšie majme

Model M: $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\alpha}, \sigma^2\mathbf{I})$, $\boldsymbol{\alpha} \in \mathcal{R}^k$ (neznáme), σ^2 neznámy parameter, $\mathbf{X}_{n,k}$ známa matica plánu, $h(\mathbf{X}) = r \leq k$,

Model M_1: $\mathbf{Y} \sim N(\mathbf{U}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, $\boldsymbol{\beta} \in \mathcal{R}^{k_1}$ (neznáme), σ^2 neznámy parameter, $\mathbf{U}_{n,k_1}$ známa matica plánu, $h(\mathbf{U}) = r_1 \leq k_1$,

Model M_2: $\mathbf{Y} \sim N(\mathbf{T}\boldsymbol{\gamma}, \sigma^2\mathbf{I})$, $\boldsymbol{\gamma} \in \mathcal{R}^{k_2}$ (neznáme), σ^2 neznámy parameter, $\mathbf{T}_{n,k_2}$ známa matica plánu, $h(\mathbf{U}) = r_2 \leq k_2$,

$1 \leq k_2 < k_1 < k$.

Definícia 4.1: Povieme, že model M_1 je *submodelom (podmodelom)* modelu M, ak stĺpce matica $\mathbf{U}_{n,k_1}$ sa dajú vyjadriť ako lineárne kombinácie stĺpcov matica $\mathbf{X}_{n,k}$, pričom $h(\mathbf{U}) = r_1 \leq k_1 < k$ t.j. ak $\exists \mathbf{K}_{k,k_1}$ $\mathbf{U}_{n,k_1} = \mathbf{X}_{n,k}\mathbf{K}_{k,k_1}$, $k_1 < k$.

V ďalšom predpokladajme, že M_1 je submodelom modelu M a M_2 je submodelom modelu M_1 (teda aj modelu M).

M : $\mathcal{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\alpha}$

M_1 : $\mathcal{E}(\mathbf{Y}) = \mathbf{U}\boldsymbol{\beta} = \mathbf{X}\,\underbrace{\mathbf{K}\boldsymbol{\beta}}_{\boldsymbol{\alpha}}$.

Keď pre $\mathbf{Y}$ platí model M_1, platí pre $\mathbf{Y}$ tiež model M ($\mathbf{Y}$ "sa podriaďuje" aj modelu M) a parameter $\boldsymbol{\alpha}$ je totožný s $\mathbf{K}\boldsymbol{\beta}$, $\mathcal{E}(\mathbf{Y})$ v modeli M je ten istý ako $\mathcal{E}(\mathbf{Y})$ v modeli M_1.

M_2 : $\mathcal{E}(\mathbf{Y}) = \mathbf{T}\boldsymbol{\gamma} = \mathbf{U}\,\underbrace{\mathbf{L}\boldsymbol{\gamma}}_{\boldsymbol{\beta}} = \mathbf{X}\,\underbrace{\mathbf{K}\mathbf{L}\boldsymbol{\gamma}}_{\boldsymbol{\alpha}}$.

Ak M_2 je submodelom modelu M_1, tak $\exists\ \mathbf{L}_{k_1,k_2}$, že $\mathbf{T} = \mathbf{U}\mathbf{L}$. Keď pre $\mathbf{Y}$ platí model M_2, platí pre $\mathbf{Y}$ tiež model M ($\mathbf{Y}$ "sa podriaďuje" aj modelu M) a parameter $\boldsymbol{\alpha}$ je totožný s $\mathbf{K}\mathbf{L}\boldsymbol{\gamma}$, $\mathcal{E}(\mathbf{Y})$ v modeli M je ten istý ako $\mathcal{E}(\mathbf{Y})$ v modeli M_2. Označme odhady $\widehat{\mathcal{E}(\mathbf{Y})}$ v jednotlivých modeloch:

M : $\mathcal{E}(\mathbf{Y})$ odhadneme pomocou $\hat{\boldsymbol{\mu}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}$,

M_1 : $\mathcal{E}(\mathbf{Y})$ odhadneme pomocou $\hat{\boldsymbol{\nu}} = \mathbf{U}(\mathbf{U}'\mathbf{U})^{-}\mathbf{U}'\mathbf{Y}$,

M_2 : $\mathcal{E}(\mathbf{Y})$ odhadneme pomocou $\hat{\boldsymbol{\tau}} = \mathbf{T}(\mathbf{T}'\mathbf{T})^{-}\mathbf{T}'\mathbf{Y}$.

Cieľom je testovať možnú "redukciu" M_1 na M_2 a M na M_1. Ak sa dá redukovať M na M_1, tak $\hat{\boldsymbol{\mu}} \approx \hat{\boldsymbol{\nu}}$. Ak sa dá redukovať M_1 na M_2, tak $\hat{\boldsymbol{\nu}} \approx \hat{\boldsymbol{\tau}}$.

Veta 4.2: Nech reziduálny súčet štvorcov $S_e = (\mathbf{Y} - \hat{\boldsymbol{\mu}})'(\mathbf{Y} - \hat{\boldsymbol{\mu}})$, M_1 je submodelom modelu M a M_2 je submodelom modelu M_1, $1 \leq h(\mathbf{T})(= r_2) < h(\mathbf{U})(= r_1) < h(\mathbf{X})(= r) < n$. Ak pre $\mathbf{Y}$ platí M_1, tak

$$F = \frac{(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\nu}})'(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\nu}})}{S_e}\ \frac{n-r}{r-r_1} \sim F_{r-r_1,n-r}. \tag{36}$$

Ak pre $\mathbf{Y}$ platí M_2, tak

$$F = \frac{(\hat{\boldsymbol{\nu}} - \hat{\boldsymbol{\tau}})'(\hat{\boldsymbol{\nu}} - \hat{\boldsymbol{\tau}})}{S_e} \ \frac{n-r}{r_1 - r_2} \sim F_{r_1 - r_2, n-r}. \tag{37}$$

Dôkaz: Najprv si ukážeme, že $\mathbf{X'Y}$ a $S_e = (\mathbf{Y} - \hat{\boldsymbol{\mu}})'(\mathbf{Y} - \hat{\boldsymbol{\mu}}) = \mathbf{Y'(I - X(X'X)^{-}X')Y}$ sú nezávislé. Podľa Vety 16, str. 81 knihy Anděl, J., *Matematická statistika*, SNTL, Praha, 1985 (dokazovali sme si ju na prednáške Lineární statistické modely I), sú $\mathbf{X'Y}$ a $S_e = (\mathbf{Y} - \hat{\boldsymbol{\mu}})'(\mathbf{Y} - \hat{\boldsymbol{\mu}}) = \mathbf{Y'(I - X(X'X)^{-}X')Y}$ nezávislé ak $\mathbf{X'}\sigma^2\mathbf{I(I - X(X'X)^{-}X') = 0}$, čo je pravda (podotýkame, že $\mathbf{I - X(X'X)^{-}X'}$ je p.s.d. matica). Pretože

$$\hat{\boldsymbol{\mu}} = \mathbf{X(X'X)^{-}X'Y}, \quad \hat{\boldsymbol{\nu}} = \mathbf{U(U'U)^{-}U'Y = U(U'U)^{-}K'X'Y}$$

sú funkcie $\mathbf{X'Y}$, teda aj

$$\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\nu}} = [\mathbf{X(X'X)^{-} - U(U'U)^{-}K'}]\mathbf{X'Y} \quad \text{a} \quad S_e$$

sú nezávislé. Teda čitateľ a menovateľ v (36) sú nezávislé náhodné veličiny.

Nech pre $\mathbf{Y}$ platí model M_1 (čiže aj model M). Podľa Vety 3.34 je $\frac{S_e}{\sigma^2} \sim \chi^2_{n-r}$. Ďalej

$$\mathcal{E}(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\nu}}) = \mathcal{E}[(\mathbf{X(X'X)^{-}X' - U(U'U)^{-}U'})\mathbf{Y}] = \mathbf{X(X'X)^{-}X'U}\boldsymbol{\beta} - \mathbf{U(U'U)^{-}U'U}\boldsymbol{\beta} =$$

$$= \underbrace{\mathbf{X(X'X)^{-}X'X}}_{\mathbf{X}}\mathbf{K}\boldsymbol{\beta} - \mathbf{U}\boldsymbol{\beta} = \underbrace{\mathbf{XK}}_{\mathbf{U}}\boldsymbol{\beta} - \mathbf{U}\boldsymbol{\beta} = \mathbf{0},$$

$$cov(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\nu}}) = [\mathbf{X(X'X)^{-}X' - U(U'U)^{-}U'}]\sigma^2\mathbf{I}[\mathbf{X(X'X)^{-}X' - U(U'U)^{-}U'}] =$$

(lebo nezáleží na voľbe $g-$inverzií)

$$= \sigma^2\{\mathbf{X(X'X)^{-}X'} - \overbrace{\mathbf{X(X'X)^{-}X'}\underbrace{\mathbf{U}}_{\mathbf{XK}}}^{\mathbf{XK=U}}\mathbf{(U'U)^{-}U'} - \mathbf{U(U'U)^{-}}\overbrace{\underbrace{\mathbf{U'}}_{\mathbf{K'X'}}\mathbf{X(X'X)^{-}X'}}^{\mathbf{K'X'=U'}}\mathbf{(U'U)^{-}U'}\} =$$

$$= \sigma^2\{\mathbf{X(X'X)^{-}X' - U(U'U)^{-}U'}\}. \tag{38}$$

Podľa Vety 12, str. 79 v knihe Anděl, J., *Matematická statistika*, SNTL, Praha, 1985, dokazovali sme si ju na prednáške Lineární statistické modely I) platí, že

$$(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\nu}})'[cov(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\nu}})]^{-}(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\nu}}) \sim \chi^2_{h[cov(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\nu}})]},$$

(bez ohľadu na voľbu $g-$inverzie $(cov(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\nu}}))^{-}$), teda

$$(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\nu}})'\frac{1}{\sigma^2}\{\mathbf{X(X'X)^{-}X' - U(U'U)^{-}U'}\}^{-}(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\nu}}) \sim \chi^2_{h[cov(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\nu}})]}.$$

Pretože $\mathbf{X(X'X)^{-}X' - U(U'U)^{-}U'}$ je idempotentná matica (pozri (38)), je jednotková matica $\mathbf{I}$ jej jedna $g-$inverzia a dostávame, že

$$\frac{1}{\sigma^2}(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\nu}})'(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\nu}}) \sim \chi^2_{r-r_1},$$

lebo

$$h[cov(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\nu}})] = h(\sigma^2[\mathbf{X(X'X)^{-}X' - U(U'U)^{-}U'}]) = h[\mathbf{X(X'X)^{-}X' - U(U'U)^{-}U'}] =$$

$$= tr[\mathbf{X(X'X)^{-}X' - U(U'U)^{-}U'}] = tr(\mathbf{X(X'X)^{-}X'}) - tr(\mathbf{U(U'U)^{-}U'}) =$$

$$= tr(\mathbf{X'X(X'X)^{-}}) - tr(\mathbf{U'U(U'U)^{-}}) = h(\mathbf{X}) - h(\mathbf{U}) = r - r_1$$

(podľa záveru dôkazu Vety 3.34). Preto

$$F = \frac{\dfrac{(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\nu}})'(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\nu}})}{\sigma^2(r - r_1)}}{\dfrac{S_e}{\sigma^2(n - r)}} = \frac{(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\nu}})'(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\nu}})}{S_e} \; \frac{n - r}{r - r_1} \sim F_{r-r_1,n-r}.$$

Vzťah (37) dokážeme úplne analogicky. Q.E.D.

Hypotézu

$$H_0: \quad \text{platí submodel } M_1 \qquad \asymp \qquad H_1: \quad \text{neplatí } H_0$$

testujeme nasledovne. Ak realizácia $F_{real} > F_{r-r_1,n-r}(1-\alpha)$ (F dané vzťahom (36)), tak H_0 zamietame na hladine významnosti α.

Veta 4.3: Nech M_1 je submodelom modelu M a M_2 je submodelom modelu M_1. Odhady $\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\nu}}, \hat{\boldsymbol{\tau}}, S_e$ splňujú identity

(i) $S_e + (\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\nu}})'(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\nu}}) + (\hat{\boldsymbol{\nu}} - \hat{\boldsymbol{\tau}})'(\hat{\boldsymbol{\nu}} - \hat{\boldsymbol{\tau}}) = \mathbf{Y'Y} - \hat{\boldsymbol{\tau}}'\hat{\boldsymbol{\tau}}$,

(ii) $S_e + (\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\nu}})'(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\nu}}) = \mathbf{Y'Y} - \hat{\boldsymbol{\nu}}'\hat{\boldsymbol{\nu}}$,

(iii) $S_e = \mathbf{Y'Y} - \hat{\boldsymbol{\mu}}'\hat{\boldsymbol{\mu}}$.

Dôkaz:

$$S_e = (\mathbf{Y} - \hat{\boldsymbol{\mu}})'(\mathbf{Y} - \hat{\boldsymbol{\mu}}) = \mathbf{Y'(I - (X(X'X)^{-}X')Y} = \mathbf{Y'Y} - \mathbf{Y'}\underbrace{\mathbf{X(X'X)^{-}X'X}}_{\mathbf{X}}\mathbf{(X'X)^{-}X'Y} = \mathbf{Y'Y} - \hat{\boldsymbol{\mu}}'\hat{\boldsymbol{\mu}}$$

(lebo nezáleží na výbere $g-$inverzie $(\mathbf{X'X})^{-}$). Dokázali sme (iii).
Pretože $\mathbf{X(X'X)^{-}X'} - \mathbf{U(U'U)^{-}U'}$ je symetrická a idempotentná matica (pozri (38)), je

$$(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\nu}})'(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\nu}}) = \mathbf{Y'[X(X'X)^{-}X' - U(U'U)^{-}U']'[X(X'X)^{-}X' - U(U'U)^{-}U']Y} =$$

$$= \mathbf{Y'[X(X'X)^{-}X' - U(U'U)^{-}U']Y} = \mathbf{Y'}\underbrace{\mathbf{X(X'X)^{-}X'X}}_{\mathbf{X}}\mathbf{(X'X)^{-}X'Y} - \mathbf{Y'}\underbrace{\mathbf{U(U'U)^{-}U'U}}_{\mathbf{U}}\mathbf{(U'U)^{-}U'Y} =$$

$$= \hat{\boldsymbol{\mu}}'\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\nu}}'\hat{\boldsymbol{\nu}}. \tag{39}$$

Analogicky ľahko sa dokáže, že aj $\mathbf{U(U'U)^{-}U'} - \mathbf{T(T'T)^{-}T'}$ je symetrická a idempotentná matica a preto

$$(\hat{\boldsymbol{\nu}} - \hat{\boldsymbol{\tau}})'(\hat{\boldsymbol{\nu}} - \hat{\boldsymbol{\tau}}) = \mathbf{Y'[U(U'U)^{-}U' - T(T'T)^{-}T']'[U(U'U)^{-}U' - T(T'T)^{-}T']Y} =$$

$$= \mathbf{Y'[U(U'U)^{-}U' - T(T'T)^{-}T']Y} = \mathbf{Y'}\underbrace{\mathbf{U(U'U)^{-}U'U}}_{\mathbf{U}}\mathbf{(U'U)^{-}U'Y} - \mathbf{Y'}\underbrace{\mathbf{T(T'T)^{-}T'T}}_{\mathbf{T}}\mathbf{(T'T)^{-}T'Y} =$$

$$= \hat{\boldsymbol{\nu}}'\hat{\boldsymbol{\nu}} - \hat{\boldsymbol{\tau}}'\hat{\boldsymbol{\tau}}. \tag{40}$$

Pomocou (39) a (40) je evidentné (i) a (ii). Q.E.D.

5 Analýza rozptylu

5.1 Analýza rozptylu jednoduchého triedenia

Majme I nezávislých náhodných výberov z normálneho rozdelenia s rovnakými disperziami, teda

1. výber $\quad Y_{1,1}, Y_{1,2}, ..., Y_{1,n_1}$ $\quad$ rozsahu n_1 z $N(\mu_1, \sigma^2)$,
2. výber $\quad Y_{2,1}, Y_{2,2}, ..., Y_{2,n_2}$ $\quad$ rozsahu n_2 z $N(\mu_2, \sigma^2)$,

$\vdots$

$I-$ty výber $\quad Y_{I,1}, Y_{I,2}, ..., Y_{I,n_I}$ $\quad$ rozsahu n_I z $N(\mu_I, \sigma^2)$.

Cieľom je overiť hypotézu

$$H_0: \quad \mu_1 = \mu_2 = ... = \mu_I \qquad \asymp \qquad H_1: \quad \exists\, i \neq j \;\; \mu_i \neq \mu_j$$

Ide o lineárny regresný model

$$Y_{ij} = \mu_i + \epsilon_{ij} \qquad i = 1, 2, ., , , I, \quad j = 1, 2, ..., n_i,$$

ktorý môžeme zapísať maticovo

$$\mathbf{Y}_{n,1} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \mathbf{Y}_I \end{pmatrix} = \begin{pmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ Y_{22} \\ \vdots \\ Y_{2n_2} \\ \vdots \\ Y_{I1} \\ Y_{I2} \\ \vdots \\ Y_{In_I} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ \vdots & & & \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & & & \\ 0 & 1 & \dots & 0 \\ \vdots & & & \\ 0 & 0 & \dots & 1 \\ \vdots & & & \\ 0 & 0 & \dots & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_I \end{pmatrix} = \mathbf{X}^*_{n,I} \boldsymbol{\beta}^*_{I,1} + \boldsymbol{\epsilon}^*_{n,1},$$

kde $n = \sum_{i=1}^{I} n_i$, $\boldsymbol{\epsilon}^* \sim N(\mathbf{0}_{n,1}, \sigma^2 \mathbf{I}_{n,n})$ a $h(\mathbf{X}^*) = I$, teda ide o model plnej hodnosti.

Reparametrizujeme model, teda zavedieme nové parametre $\mu, \alpha_1, ..., \alpha_I$, a síce $\mu_i = \mu + \alpha_i$, $\; i = 1, 2, ..., I$

a teraz $Y_{ij} \sim N(\mu+\alpha_i,\sigma^2), \quad i=1,2,...,I, \quad j=1,2,...,n_i$. Maticový zápis reparametrizovaného modelu je

$$\mathbf{Y}_{n,1} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \mathbf{Y}_I \end{pmatrix} = \begin{pmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ Y_{22} \\ \vdots \\ Y_{2n_2} \\ \vdots \\ Y_{I1} \\ Y_{I2} \\ \vdots \\ Y_{In_I} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ \vdots \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ \vdots \\ 1 & 0 & 1 & \dots & 0 \\ \vdots \\ 1 & 0 & 0 & \dots & 1 \\ \vdots \\ 1 & 0 & 0 & \dots & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_I \end{pmatrix} = \mathbf{X}_{n,I+1}\boldsymbol{\beta}_{I+1,1} + \boldsymbol{\epsilon}_{n,1},$$

kde $\boldsymbol{\epsilon} \sim N(\mathbf{0}_{n,1}, \sigma^2\mathbf{I}_{n,n})$ a $h(\mathbf{X}) = I$ (prvý stĺpec matice $\mathbf{X}$ je súčtom ostatných stĺpcov), teda ide o model neúplnej hodnosti. μ voláme *celkový efekt ošetření* a α_i je *efekt $i-$teho ošetření.*

Cieľom je odhad vektora $\mathcal{E}(\mathbf{Y}) = (\overbrace{\mu+\alpha_1,...,\mu+\alpha_1}^{n_1 \text{ krát}},...,\overbrace{\mu+\alpha_I,...,\mu+\alpha_I}^{n_I \text{ krát}})' = \mathbf{X}\boldsymbol{\beta}$. Podľa Poznámky 3.32 tento odhad vždy existuje a NNLO vektora $\mathcal{E}(\mathbf{Y})$ je $\widehat{\mathcal{E}(\mathbf{Y})} = \widehat{\mathbf{X}\boldsymbol{\beta}} = \hat{\mathbf{Y}} = \mathbf{Xb}$, kde $\mathbf{b}$ je ľubovoľné riešenie normálnych rovníc **X'Xb** = **X'Y**. Podľa Vety 3.16 je sústava normálnych rovníc ekvivalentná sústave

$$\frac{\partial S}{\partial \mu} = 0, \tag{41}$$

$$\frac{\partial S}{\partial \alpha_i} = 0, \quad i=1,2,...,I, \tag{42}$$

kde $S(\mu,\alpha_1,...,\alpha_I) = \sum_{i=1}^I \sum_{j=1}^{n_i} (Y_{ij}-\mu-\alpha_i)^2$.

Označme

$$\sum_{j=1}^{n_i} Y_{ij} = Y_{i\bullet}, \quad \sum_{i=1}^{I}\sum_{j=1}^{n_i} Y_{ij} = \sum_{i=1}^{I} Y_{i\bullet} = Y_{\bullet\bullet}, \quad \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} = \frac{1}{n_i}Y_{i\bullet} = y_{i\bullet},$$

$$n = \sum_{i=1}^{I} n_i, \quad \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{n_i} Y_{ij} = \frac{1}{n}Y_{\bullet\bullet} = y_{\bullet\bullet}$$

(zaužívané označenie). Rovnice (41),(42) sú

$$n\mu + \sum_{i=1}^{I} n_i\alpha_i = Y_{\bullet\bullet}, \tag{43}$$

$$n_t\mu + n_t\alpha_t = Y_{t\bullet}, \quad t=1,2,...,I \tag{44}$$

(preverte). Ako vyzerajú normálne rovnice $\mathbf{X'Xb} = \mathbf{X'Y}$?

$$\mathbf{X'Xb} = \begin{pmatrix} 1 & \dots & 1 & 1 & \dots & 1 & \dots & 1 & \dots & 1 \\ 1 & \dots & 1 & 0 & \dots & 0 & \dots & 0 & \dots & 0 \\ 0 & \dots & 0 & 1 & \dots & 1 & \dots & 0 & \dots & 0 \\ \vdots & & \vdots & & & & & & & \\ 0 & \dots & 0 & 0 & \dots & 0 & \dots & 1 & \dots & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ \vdots & & & & \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ \vdots & & & & \\ 1 & 0 & 1 & \dots & 0 \\ \vdots & & & & \\ 1 & 0 & 0 & \dots & 1 \\ \vdots & & & & \\ 1 & 0 & 0 & \dots & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_I \end{pmatrix} =$$

$$= \begin{pmatrix} n & n_1 & n_2 & \dots & n_I \\ n_1 & n_1 & 0 & \dots & 0 \\ n_2 & 0 & n_2 & \dots & 0 \\ \vdots & \vdots & & & \\ n_I & 0 & & \dots & n_I \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_I \end{pmatrix}$$

a

$$\mathbf{X'Y} = \begin{pmatrix} \sum_{i=1}^{I} \sum_{j=1}^{n_i} Y_{ij} \\ Y_{1\bullet} \\ \vdots \\ Y_{I\bullet} \end{pmatrix},$$

teda

$$n\mu + \sum_{i=1}^{I} n_i \alpha_i = Y_{\bullet\bullet},$$

$$n_t \mu + n_t \alpha_t = Y_{t\bullet}, \quad t = 1, 2, ..., I,$$

čo sú (pochopiteľne) tie isté rovnice ako (43) a (44). Zrejme sústava (43) a (44) má singulárnu maticu, lebo súčet rovníc (44) dáva rovnicu (43). Stačí nám ale nájsť ľubovoľné riešenie tejto sústavy. Môžeme napríklad zvoliť riešenie $\mu^* = 0$ a $\alpha_i^* = y_{i\bullet}$, $\;i = 1, 2, ..., I$. Budeme postupovať ináč (osvedčilo sa to aj pri iných modeloch analýzy rozptylu), a síce pridáme ďalšiu rovnicu (podmienku)

$$\sum_{i=1}^{I} n_i \alpha_i = 0$$

tzv. reparametrizačnú rovnicu. Takto dostávame sústavu

$$\sum_{i=1}^{I} n_i \alpha_i = 0$$

$$n\mu + \sum_{i=1}^{I} n_i \alpha_i = Y_{\bullet\bullet}$$

$$n_t\mu + n_t\alpha_t = Y_{t\bullet}, \quad t = 1, 2, ..., I,$$

ktorej (jediné) riešenie je

$$\mu^\circ = y_{\bullet\bullet}, \quad \alpha_t^\circ = y_{t\bullet} - y_{\bullet\bullet}, \quad t = 1, 2, ..., I.$$

Teda NNLO

$$\hat{\boldsymbol{\mu}} = \widehat{\mathcal{E}(\mathbf{Y})} = \mathbf{X}\begin{pmatrix} \mu^\circ \\ \alpha_1^\circ \\ \vdots \\ \alpha_I^\circ \end{pmatrix} = \begin{pmatrix} \mu^\circ + \alpha_1^\circ \\ \mu^\circ + \alpha_1^\circ \\ \vdots \\ \mu^\circ + \alpha_1^\circ \\ \mu^\circ + \alpha_2^\circ \\ \mu^\circ + \alpha_2^\circ \\ \vdots \\ \mu^\circ + \alpha_2^\circ \\ \vdots \\ \mu^\circ + \alpha_I^\circ \\ \mu^\circ + \alpha_I^\circ \\ \vdots \\ \mu^\circ + \alpha_I^\circ \end{pmatrix} = \begin{pmatrix} y_{1\bullet} \\ y_{1\bullet} \\ \vdots \\ y_{1\bullet} \\ y_{2\bullet} \\ y_{2\bullet} \\ \vdots \\ y_{2\bullet} \\ \vdots \\ y_{I\bullet} \\ y_{I\bullet} \\ \vdots \\ y_{I\bullet} \end{pmatrix}.$$

Ak máme testovať v pôvodnom modeli (s plnou hodnosťou, str. 44)

$$H_0: \quad \mu_1 = \mu_2 = ... = \mu_I \qquad \asymp \qquad H_1: \quad \exists\, i \neq j \;\; \mu_i \neq \mu_j,$$

tak v reparametrizovanom modeli (s neúplnou hodnosťou, str. 45) je to ekvivalentné testovaniu

$$H_0: \quad \alpha_1 = \alpha_2 = ... = \alpha_I \qquad \asymp \qquad H_1: \quad \exists\, i \neq j \;\; \alpha_i \neq \alpha_j.$$

Za platnosti H_0 máme namiesto modelu

$M: \quad \mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$

submodel

$M_1: \quad Y_{ij} = \gamma + \epsilon_{ij}, \quad i = 1, 2, ..., I, \; j = 1, 2, ..., n_i, \;\; \epsilon_{ij} \sim N(0, \sigma^2)$ (všetky nezávislé)

s maticou plánu $\mathbf{U}_{n,1} = (1, 1, ..., 1)'$. Zrejme $\mathcal{M}(\mathbf{U}) \subset \mathcal{M}(\mathbf{X})$, lebo $\mathbf{U} = \mathbf{X}(1, 0, ..., 0)'$. Submodel M_1 je plnej hodnosti. Odhadneme v ňom γ. NNLO parametra γ je $\hat{\gamma} = (\mathbf{U'U})^{-1}\mathbf{U'Y} = \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{n_i} Y_{ij} = \frac{1}{n}Y_{\bullet\bullet} = y_{\bullet\bullet}$. Preto

$$\hat{\boldsymbol{\nu}} = \widehat{\mathcal{E}(\mathbf{Y})} = \begin{pmatrix} y_{\bullet\bullet} \\ y_{\bullet\bullet} \\ \vdots \\ y_{\bullet\bullet} \end{pmatrix} = \mathbf{1}_{n,1} y_{\bullet\bullet}.$$

Podľa Vety 4.3 je reziduálny súčet štvorcov (RSČ) $S_e = \mathbf{Y'Y} - \hat{\boldsymbol{\mu}}'\hat{\boldsymbol{\mu}}$ a $S_e + (\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\nu}})'(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\nu}}) = \mathbf{Y'Y} - \hat{\boldsymbol{\nu}}'\hat{\boldsymbol{\nu}}$. Náhodnú veličinu $\mathbf{Y'Y} - \hat{\boldsymbol{\nu}}'\hat{\boldsymbol{\nu}}$ nazývame S_T - *totálny (celkový) súčet štvorcov* a náhodnú veličinu $(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\nu}})'(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\nu}})$ voláme S_A - *súčet štvorcov medzi triedami*, alebo *súčet štvorcov medzi riadkami*, alebo *súčet* *štvorcov ak je zdroj menlivosti "A"* (t.j. $\hat{\boldsymbol{\mu}}$ uvažuje "A" a $\hat{\boldsymbol{\nu}}$ neuvažuje "A" (zdroj menlivosti)). Platí

$$\hat{\boldsymbol{\mu}} = \begin{pmatrix} y_{1\bullet} \\ y_{1\bullet} \\ \vdots \\ y_{1\bullet} \\ y_{2\bullet} \\ y_{2\bullet} \\ \vdots \\ y_{2\bullet} \\ \vdots \\ y_{I\bullet} \\ y_{I\bullet} \\ \vdots \\ y_{I\bullet} \end{pmatrix}, \qquad \hat{\boldsymbol{\nu}} = \begin{pmatrix} y_{\bullet\bullet} \\ y_{\bullet\bullet} \\ \vdots \\ y_{\bullet\bullet} \\ y_{\bullet\bullet} \\ y_{\bullet\bullet} \\ \vdots \\ y_{\bullet\bullet} \\ \vdots \\ y_{\bullet\bullet} \\ y_{\bullet\bullet} \\ \vdots \\ y_{\bullet\bullet} \end{pmatrix},$$

teda

$$S_T = \sum_{i=1}^{I}\sum_{j=1}^{n_i} Y_{ij}^2 - n y_{\bullet\bullet}^2 = \sum_{i=1}^{I}\sum_{j=1}^{n_i} Y_{ij}^2 - \frac{Y_{\bullet\bullet}^2}{n},$$

$$S_A = (\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\nu}})'(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\nu}}) = \hat{\boldsymbol{\mu}}'\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\nu}}'\hat{\boldsymbol{\nu}}' = \sum_{i=1}^{I} n_i y_{i\bullet}^2 - n y_{\bullet\bullet}^2 = \sum_{i=1}^{I} \frac{Y_{i\bullet}^2}{n_i} - \frac{Y_{\bullet\bullet}^2}{n},$$

$$S_e = S_T - S_A = \sum_{i=1}^{I}\sum_{j=1}^{n_i} Y_{ij}^2 - \sum_{i=1}^{I} \frac{Y_{i\bullet}^2}{n_i}.$$

Za platnosti $H_0: \quad \alpha_1 = \alpha_2 = ... = \alpha_I$ platí podľa Vety 4.2

$$F = \frac{(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\nu}})'(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\nu}})}{(\mathbf{Y} - \hat{\boldsymbol{\mu}})'(\mathbf{Y} - \hat{\boldsymbol{\mu}})} \; \frac{n - h(\mathbf{X})}{h(\mathbf{X}) - h(\mathbf{U})} = \frac{S_A}{S_e} \; \frac{n - I}{I - 1} \sim F_{I-1,n-I}.$$

Ak realizácia $F_{real} \geq F_{I-1,n-I}(1-\alpha)$, tak na hladine významnosti α zamietame H_0.

Tabuľka analýzy rozptylu jednoduchého triedenia.

zdroj variability	súčet štvorcov	stupne voľnosti	S/f	$F = \dfrac{S/f}{S_e/f_e}$
skupiny (riadky, "typ pôdy",...)	S_A	$f_A = I - 1$	$\dfrac{S_A}{f_A}$	$\dfrac{S_A/f_A}{S_e/f_e}$ * **
reziduály	S_e	$f_e = n - I$	$\dfrac{S_e}{f_e}$	–
celkový	S_T	$f_T = n - 1$		

V poslednom stĺpci tabuľky je hodnota testovacieho kritéria. Ak táto hodnota je taká, že test zamieta nulovú hypotézu na hladine významnosti $\alpha = 0,05$, hodnota testovacieho kritéria sa označí jednou hviezdičkou. Ak táto hodnota je taká, že test zamieta nulovú hypotézu na hladine významnosti $\alpha = 0,01$, je zaužívané hodnotu testovacieho kritéria sa označiť dvomi hviezdičkami.

Poznámka 5.1: (Bartlettov test) Aby sme mohli realizovať analýzu rozptylu, musíme overiť, či disperzie v triedach sú rovnaké (predpokladáme normalitu pozorovaní, táto normalita sa tiež overuje testami normality).

Bartlettov test:

$$s_i^2 = \frac{1}{n_i - 1}\left(\sum_{j=1}^{n_i} Y_{ij}^2 - n_i y_{i\bullet}^2\right), \quad i = 1, 2, ..., I,$$

$$s^2 = \frac{1}{n-I}\sum_{i=1}^{I}(n_i - 1)s_i^2, \qquad C = 1 + \frac{1}{3(I-1)}\left(\sum_{i=1}^{I}\frac{1}{n_i - 1} - \frac{1}{n-I}\right).$$

Za platnosti hypotézy $H_0 \quad \sigma_1^2 = \sigma_2^2 = ... = \sigma_I^2$ (rovnaké disperzie v triedach) platí

$$B = \frac{1}{C}\left[(n-I)\ln s^2 - \sum_{i=1}^{I}(n_i - 1)\ln s_i^2\right] \approx \chi_{I-1}^2.$$

Ak realizácia $B_{real} \geq \chi_{I-1}^2(1-\alpha)$, zamietame hypotézu o rovnosti disperzií na hladine významnosti α. Test sa dá aplikovať ak $n_i > 6$ pre všetky $i = 1, 2, ..., I$.

Poznámka 5.2: Ak $H_0 : \quad \alpha_1 = \alpha_2 = ... = \alpha_I$ zamietame, potom sa pýtame pre ktoré $i \neq j$ je $\mu_i \neq \mu_j \quad (\alpha_i \neq \alpha_j)$. Teda vlastne pre každé $i \neq j$ testujeme

$$H_0 : \quad \mu_i = \mu_j \qquad \asymp \qquad H_1 : \quad \mu_i \neq \mu_j.$$

H_0 zamietame na hladine významnosti α, ak

$$|y_{i\bullet} - y_{j\bullet}| > \sqrt{(I-1)\frac{S_e}{n-I}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)F_{I-1,n-I}(1-\alpha)}. \tag{45}$$

Táto metóda sa volá *Schefférho metóda*, jej analýza a odvodenie nájdete v knižke Anděl, J., *Matematická statistika*, SNTL, Praha, 1985, str. 147. Ak $n_1 = n_2 = ... = n_I$, tak sa á použiť aj metóda Tukeyho (pozrite tiež v knihe Anděl, J., *Matematická statistika*, SNTL, Praha, 1985 na str. 150).

Príklad 5.3: (Anděl, J., *Matematická statistika*, SNTL, Praha, 1985, str. 156.) U štyroch odrodách zemiakov (A,B,C,D) sa zisťovala celková motnosť zemiakov, ktoré vyrástli v jednom trse. Výsledky sú v nasledujúcej tabuľke:

odroda	hodnoty (realizácie) Y_{ij}	početnosť n_i	súčet $Y_{i\bullet}$	priemer $y_{i\bullet}$	súčet štvorcov $\sum_j Y_{ij}^2$
A	0,9 0,8 0,6 0,9	4	3,2	0,8	2,62
B	1,3 1,0 1,3	3	3,6	1,2	4,38
C	1,3 1,5 1,6 1,1 1,5	5	7,0	1,4	9,96
D	1,1 1,2 1,0	3	3,3	1,1	3,65
celkom		$n = 15$	$Y_{\bullet\bullet} = 17,1$		$\sum_i \sum_j Y_{ij}^2 = 20,61$

Vzhľadom k malým početnostiam n_i nerealizujeme test normality ani Bartlettov test.

$$S_A = \sum_{i=1}^{I} \frac{Y_{i\bullet}^2}{n_i} - \frac{Y_{\bullet\bullet}^2}{n} = \frac{3,2^2}{4} + \frac{3,6^2}{3} + \frac{7,0^2}{5} + \frac{3,3^2}{3} - \frac{17,1^2}{15} = 0,8160,$$

$$S_T = \sum_{i=1}^{I} \sum_{j=1}^{n_i} Y_{ij}^2 - \frac{Y_{\bullet\bullet}^2}{n} = 20,61 - \frac{17,1^2}{15} = 1,1160,$$

$$S_e = S_T - S_A = 1,1160 - 0,8160 = 0,3.$$

Tabuľka analýzy rozptylu.

zdroj variability	súčet štvorcov	stupne voľnosti	S/f	$F = \frac{S/f}{S_e/f_e}$
odrody	$S_A = 0,8160$	$3(= I - 1)$	$0,2720$	$9,97^*$
reziduály	$S_e = 0,300$	$11(= n - I)$	$0,02727$	$-$
celkový	$S_T = 1,1160$	14		

Pretože $9,97 \geq F_{3,11}(0,95) = 3,59$, zamietame na hladine významnosti $0,05$ (5%) hypotézu, že stredná hodnota hmotnosti trsu zemiakov nezávisí na odrode. Scheffého metódou chceme odhaliť, ktoré odrody sú významne odlišné medzi sebou.

$$\frac{S_e}{n-I} = 0,02727, \quad F_{3,11}(0,95) = 3,59, \quad (I-1)\frac{S_e}{n-I} F_{3,11}(0,95) = 0,29370,$$

preto tabuľka pre porovnávanie dvojíc Scheffého metodou vyzerá nasledovne

zrovnávané odrody	absolútna hodnota rozdielov $\lvert y_{i\bullet} - y_{j\bullet}\rvert$	$\sqrt{(I-1)\frac{S_e}{n-I}\left(\frac{1}{n_i} + \frac{1}{n_j}\right) F_{I-1,n-I}(0,95)}$
A,B	$\lvert y_{1\bullet} - y_{2\bullet}\rvert = 0,4$	0,41
A,C	$\lvert y_{1\bullet} - y_{3\bullet}\rvert = 0,6^*$	0,36
A,D	$\lvert y_{1\bullet} - y_{4\bullet}\rvert = 0,3$	0,41
B,C	$\lvert y_{2\bullet} - y_{3\bullet}\rvert = 0,2$	0,40
B,D	$\lvert y_{2\bullet} - y_{4\bullet}\rvert = 0,1$	0,44
C,D	$\lvert y_{3\bullet} - y_{4\bullet}\rvert = 0,3$	0,40

Len pri porovnávaní odrôd A a C možno na hladine významnosti 0.05 prehlásiť, že tieto dve odrody sú (štatisticky) významne odlišné.

5.2 Analýza rozptylu dvojného triedenia bez interakcií

Príklad 5.4: (Anděl, J., *Matematická statistika*, SNTL, Praha, 1985, str. 167.) Skúmali sa výnosy sena v q/ha v závislosti na $A-$typ pôdy $i = 1, 2$ a $B-$spôsob hnojenia $j = 1, 2, 3$. Každá kombinácia typu pôdy (normálna, kyselá) bola realizovaná s každým spôsobom hnojenia (bez hnojenia, chlievska mrva, vápenaté hnojivo) vždy štyrikrát (na štyroch pozemkoch) nezávisle. Výsledky sú v nasledujúcich tabuľkách:

	spôsob hnojenia j				
typ pôdy A i	bez hnojenia 1	chlievska mrva 2	vápenaté hnojivo 3	súčet $Y_{i\bullet\bullet}$	$\sum_j \sum_k Y_{ijk}^2$
normálna 1	28 32 30 30	37 36 39 36	34 38 37 36	413	14 355
kyselá 2	31, 27 30 29	34 34 30 38	42 40 41 39	415	14 653
súčet $Y_{\bullet j \bullet}$	237	284	307	$Y_{\bullet\bullet\bullet}$=828	
$\sum_i \sum_k^4 Y_{ijk}^2$	7039	10 138	11 831		$\sum_i \sum_j \sum_k Y_{ijk}^2$ =29 008

$Y_{ij\bullet}$	$j=1$	$j=2$	$j=3$
$i=1$	120	148	145
$i=2$	117	136	162

$$\sum_{i=1}^{2}\sum_{j=1}^{3} Y_{ij\bullet}^2 = 115\ 758, \qquad n_{ij} = 4, \quad i = 1,2, \quad j = 1,2,3.$$

V tomto prípade máme dva triediace znaky (typ pôdy $-A$, spôsob hnojenia $-B$). Máme n_{ij} pokusov takých, že u nich je A na $i-$tej a B na $j-$tej úrovni (v tomto prípade $i = 1,2,\ j = 1,2,3$). Výsledky (v tomto prípade výnosy) týchto n_{ij} pokusov sú realizácie náhodných veličín $Y_{ij1}, Y_{ij2}, ..., Y_{ijn_{ij}}$ (v tomto prípade napr. realizácia Y_{111} je 28, realizácia Y_{112} je 32, realizácia Y_{232} je 40, atď.). Základná úloha je rozhodnúť, či všetky úrovne B (spôsob hnojenia) majú na výnosy rovnaký vplyv, alebo nejaký spôsob hnojenia je "signifikantne iný" (lepší, horší). Niekedy treba naviac rozhodnúť, či výnosy závisia od typu pôdy. Model je

$$Y_{ijk} = \mu_{ij} + \epsilon_{ijk}, \tag{46}$$

$\epsilon_{ijk} \sim N(0,\sigma^2)$, $i = 1,2,...,I$, $j = 1,2,...,J$, $k = 1,2,...,n_{ij}$ a všetky ϵ_{ijk} sú nezávislé. Ak označíme

$$\overline{\mu}_{i\bullet} = \frac{\sum_{j=1}^{J}\mu_{ij}}{J}, \quad \overline{\mu}_{\bullet j} = \frac{\sum_{i=1}^{I}\mu_{ij}}{I}, \quad \overline{\mu}_{\bullet\bullet} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\mu_{ij}}{IJ}, \quad (\alpha\beta)_{ij} = \mu_{ij} - \overline{\mu}_{i\bullet} - \overline{\mu}_{\bullet j} + \overline{\mu}_{\bullet\bullet},$$

potom

$$\mu_{ij} = \overbrace{\overline{\mu}_{\bullet\bullet}}^{\mu} + \overbrace{(\overline{\mu}_{i\bullet} - \overline{\mu}_{\bullet\bullet})}^{\alpha_i} + \overbrace{(\overline{\mu}_{\bullet j} - \overline{\mu}_{\bullet\bullet})}^{\beta_j} + \overbrace{(\mu_{ij} - \overline{\mu}_{i\bullet} - \overline{\mu}_{\bullet j} + \overline{\mu}_{\bullet\bullet})}^{(\alpha\beta)_{ij}},$$

a dostávame preparametrizovaný model (46) v tvare

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}, \tag{47}$$

$\epsilon_{ijk} \sim N(0,\sigma^2)$, $i = 1,2,...,I$, $j = 1,2,...,J$, $k = 1,2,...,n_{ij}$ a všetky ϵ_{ijk} sú nezávislé. Parameter μ voláme *celkový efekt*, parameter α_i je *efekt $i-$teho riadku ($i-$tej úrovne faktora A)*, parameter β_j je *efekt $j-$teho stĺpca ($j-$tej úrovne faktora B)* a $(\alpha\beta)_{ij}$ je *interakcia*. V nasledujúcom budeme predpokladať, že interakcia je rovná 0 (pre všetky i,j), teda

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}, \tag{48}$$

$\epsilon_{ijk} \sim N(0,\sigma^2)$, $i = 1,2,...,I$, $j = 1,2,...,J$, $k = 1,2,...,n_{ij}$ a všetky ϵ_{ijk} sú nezávislé. Model (48) voláme *modelom dvojného triedenia bez interakcií.*

Poznámka 5.5: Ak je v každej triede rovnaký počet pozorovaní, t.j. $n_{ij} = K$, $i = 1, 2, ..., I$, $j = 1, 2, ..., J$, tak model (alebo triedenie) voláme *vyvážený (vybalancovaný)*, inak *nevyvážený*.
V nasledujúcom uvažujme vyvážený model (rovnako je tomu aj v Príklade 5.4). Model (48) sa dá maticovo zapísať ako

$M: \quad \mathbf{Y} = \mathbf{X}\boldsymbol{\alpha} + \boldsymbol{\epsilon}, \quad \mathbf{Y} = (Y_{111}, Y_{112}, ..., Y_{IJK})'$,

kde matica $\mathbf{X}$ typu $n(=\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} n_{ijk}) \times (I+J+1)$, ktorej prvý "blok" je matica $\mathbf{X}_{n,I+1}$ zo str. 45 a druhý "blok" si napíšte ako cvičenie, $\boldsymbol{\alpha} = (\mu, \alpha_1, ..., \alpha_I, \beta_1, ..., \beta_J)'$, pričom hodnosť $h(\mathbf{X}) = I + J - 1 = r$ (1. stĺpec matice $\mathbf{X}$ je súčtom stĺpcov druhého až $(I+1)-$vého, a takisto je súčtom stĺpcov druhého "bloku").
V modeli M chceme testovať hypotézu

$$H_{B0}: \quad \beta_1 = \beta_2 = ... = \beta_J \qquad \lessgtr \qquad H_{B1}: \quad \exists\, s \neq t \;\; \beta_s \neq \beta_t$$

(nulovosť efektov ošetrenia B). Za platnosti H_{B0} dostávame submodel

$M_1: \quad \mathbf{Y}_{ijk} = \mu + \alpha_i + \epsilon_{ijk}$,

ktorý môžeme maticovo zapísať ako

$M_1: \quad \mathbf{Y} = \mathbf{U}\boldsymbol{\delta} + \boldsymbol{\epsilon}, \quad \mathbf{Y} = (Y_{111}, Y_{112}, ..., Y_{IJK})'$,

kde matica $\mathbf{U}_{n,I+1}$ je tá istá ako matica $\mathbf{X}$ zo str. 45 a má vždy JK rovnakých riadkov, vektor parametrov $\boldsymbol{\delta} = (\mu, \alpha_1, \alpha_2, ..., \alpha_I)'$. Ľahko vidíme, že $\mathbf{U} = \mathbf{X}\begin{pmatrix} \mathbf{I}_{I+1,I+1} \\ \mathbf{0}_{J,I+1} \end{pmatrix}$, $h(\mathbf{U}) = I$ ($I \leq I + 1$ (-poče stĺpcov matice $\mathbf{U}$) a $I + 1 < I + J + 1$ (-počet stĺpcov matice $\mathbf{X}$)). Preto M_1 je submodelom modelu M (pozri Definíciu 4.1).
V modeli M_1 chceme testovať hypotézu

$$H_{A0}: \quad \alpha_1 = \alpha_2 = ... = \alpha_I \qquad \lessgtr \qquad H_{A1}: \quad \exists\, s \neq t \;\; \alpha_s \neq \alpha_t$$

(nulovosť efektov ošetrenia A). Za platnosti H_{A0} dostávame submodel

$M_2: \quad \mathbf{Y}_{ijk} = \mu + \epsilon_{ijk}$,

ktorý môžeme maticovo zapísať ako

$M_2: \quad \mathbf{Y} = \mathbf{T}\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \quad \mathbf{Y} = (Y_{111}, Y_{112}, ..., Y_{IJK})'$,

kde matica $\mathbf{T}_{n,1}$ je matica samých jedničiek a vektor parametrov $\boldsymbol{\gamma} = \mu$ (skalár). Lahko vidíme, že M_2 je submodelom modelu M_1.

V modeli M odhadneme $\mathcal{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\alpha}$, odhad je $\hat{\boldsymbol{\mu}}$ (kvôli rovnakému značeniu ako v kapitole 4. Normálne rovnice sú

$$\frac{\partial S}{\partial \mu} = 0,$$
$$\frac{\partial S}{\partial \alpha_i} = 0, \quad i = 1, 2, ..., I,$$
$$\frac{\partial S}{\partial \beta_j} = 0, \quad j = 1, 2, ..., J,$$

kde $S = S(\mu, \alpha_1, ..., \alpha_I, \beta_1, ..., \beta_J) = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk} - \mu - \alpha_i - \beta_j)^2$. Po derivovaní dostávame

$$IJK\mu + JK\sum_{i=1}^{I}\alpha_i + IK\sum_{j=1}^{J}\beta_j = Y_{\bullet\bullet\bullet},$$

$$JK\mu + JK\alpha_i + K\sum_{j=1}^{J}\beta_j = Y_{i\bullet\bullet}, \quad i = 1, 2, ..., I$$

$$IK\mu + K\sum_{i=1}^{I}\alpha_i + IK\beta_j = Y_{\bullet j\bullet}, \quad j = 1, 2, ..., J.$$

Pridáme reparametrizačné rovnice

$$\sum_{i=1}^{I}\alpha_i = 0, \qquad \sum_{j=1}^{J}\beta_j = 0$$

(aby sme dostali jednoznačné riešenie). Toto riešenie je

$$\hat{\mu} = y_{\bullet\bullet\bullet}\ ,$$

$$\hat{\alpha}_i = y_{i\bullet\bullet} - y_{\bullet\bullet\bullet}\ , \quad i = 1, 2, ..., I,$$

$$\hat{\beta}_j = y_{\bullet j\bullet} - y_{\bullet\bullet\bullet}\ , \quad j = 1, 2, ..., J.$$

Urobme ešte niekoľko pomocných výpočtov:

$$\sum_{j=1}^{J} y_{\bullet j\bullet} = y_{\bullet 1\bullet} + y_{\bullet 2\bullet} + ... + y_{\bullet J\bullet} = \frac{1}{IK}(Y_{\bullet 1\bullet} + Y_{\bullet 2\bullet} + ... + Y_{\bullet J\bullet}) = \frac{1}{IK}Y_{\bullet\bullet\bullet} = J\frac{1}{IJK}Y_{\bullet\bullet\bullet} = Jy_{\bullet\bullet\bullet} \tag{49}$$

a analogicky

$$\sum_{i=1}^{I} y_{i\bullet\bullet} = y_{1\bullet\bullet} + y_{2\bullet\bullet} + ... + y_{I\bullet\bullet} = \frac{1}{JK}(Y_{1\bullet\bullet} + Y_{2\bullet\bullet} + ... + Y_{I\bullet\bullet}) = \frac{1}{JK}Y_{\bullet\bullet\bullet} = I\frac{1}{IJK}Y_{\bullet\bullet\bullet} = Iy_{\bullet\bullet\bullet}, \tag{50}$$

ďalej

$$\sum_{j=1}^{J}(y_{\bullet j\bullet} - y_{\bullet\bullet\bullet})^2 = \sum_{j=1}^{J}(y^2_{\bullet j\bullet} - 2y_{\bullet j\bullet}y_{\bullet\bullet\bullet} + y^2_{\bullet\bullet\bullet}) = \sum_{j=1}^{J} y^2_{\bullet j\bullet} + Jy^2_{\bullet\bullet\bullet} - 2y_{\bullet\bullet\bullet}\underbrace{\left(\sum_{j=1}^{J} y_{\bullet j\bullet}\right)}_{Jy_{\bullet\bullet\bullet}\ \text{podľa (49)}} = \sum_{j=1}^{J} y^2_{\bullet j\bullet} - Jy^2_{\bullet\bullet\bullet} \tag{51}$$

a analogicky

$$\sum_{i=1}^{I}(y_{i\bullet\bullet} - y_{\bullet\bullet\bullet})^2 = \sum_{i=1}^{I}(y^2_{i\bullet\bullet} - 2y_{i\bullet\bullet}y_{\bullet\bullet\bullet} + y^2_{\bullet\bullet\bullet}) = \sum_{i=1}^{I} y^2_{i\bullet\bullet} + Iy^2_{\bullet\bullet\bullet} - 2y_{\bullet\bullet\bullet}\underbrace{\left(\sum_{i=1}^{I} y_{i\bullet\bullet}\right)}_{Iy_{\bullet\bullet\bullet}\ \text{podľa (50)}} = \sum_{i=1}^{I} y^2_{i\bullet\bullet} - Iy^2_{\bullet\bullet\bullet}. \tag{52}$$

Preto v modeli $M: \quad \mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\alpha}, \sigma^2\mathbf{I})$ je

$$\hat{\boldsymbol{\mu}} = \widehat{\mathbf{X}\boldsymbol{\alpha}} = (\overbrace{\hat{\mu} + \hat{\alpha}_1 + \hat{\beta}_1, ..., \hat{\mu} + \hat{\alpha}_1 + \hat{\beta}_1}^{K\ \text{krát}}, \overbrace{\hat{\mu} + \hat{\alpha}_1 + \hat{\beta}_2, ..., \hat{\mu} + \hat{\alpha}_1 + \hat{\beta}_2}^{K\ \text{krát}}, ..., \overbrace{\hat{\mu} + \hat{\alpha}_I + \hat{\beta}_J, ..., \hat{\mu} + \hat{\alpha}_I + \hat{\beta}_J}^{K\ \text{krát}})',$$

a

$$\hat{\boldsymbol{\mu}}'\hat{\boldsymbol{\mu}} = K\sum_{i=1}^{I}\sum_{j=1}^{J}(\hat{\mu}+\hat{\alpha}_i+\hat{\beta}_j)^2 = K\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{\bullet\bullet\bullet}+y_{i\bullet\bullet}-y_{\bullet\bullet\bullet}+y_{\bullet j\bullet}-y_{\bullet\bullet\bullet})^2 = K\sum_{i=1}^{I}\left\{\sum_{j=1}^{J}(y_{\bullet j\bullet}-y_{\bullet\bullet\bullet}+y_{i\bullet\bullet})^2\right\} =$$

$$= K\sum_{i=1}^{I}\left\{\sum_{j=1}^{J}(y_{\bullet j\bullet}-y_{\bullet\bullet\bullet})^2 + Jy_{i\bullet\bullet}^2 + 2y_{i\bullet\bullet}\overbrace{\sum_{j=1}^{J}(y_{\bullet j\bullet}-y_{\bullet\bullet\bullet})}^{0 \text{ podľa } (49)}\right\} = KI\overbrace{\left[\sum_{j=1}^{J}y_{\bullet j\bullet}^2 - Jy_{\bullet\bullet\bullet}^2\right]}^{\text{podľa } (51)} + KJ\sum_{i=1}^{I}y_{i\bullet\bullet}^2 + 0 =$$

$$= JK\sum_{i=1}^{I}y_{i\bullet\bullet}^2 + IK\sum_{j=1}^{J}y_{\bullet j\bullet}^2 - IJKy_{\bullet\bullet\bullet}^2 \tag{53}$$

a $h(\mathbf{X}) = I + J - 1$.

V submodeli $M_1: \quad \mathbf{Y} \sim N(\mathbf{U}\boldsymbol{\delta}, \sigma^2\mathbf{I})$ je

$$\hat{\boldsymbol{\nu}} = \widehat{\mathbf{U}\boldsymbol{\delta}} = (\overbrace{\mu^\circ+\alpha_1^\circ,...,\mu^\circ+\alpha_1^\circ}^{JK \text{ krát}},\overbrace{\mu^\circ+\alpha_2^\circ,...,\mu^\circ+\alpha_2^\circ}^{JK \text{ krát}},...,\overbrace{\mu^\circ+\alpha_I^\circ,...,\mu^\circ+\alpha_I^\circ}^{JK \text{ krát}})',$$

a

$$\hat{\boldsymbol{\nu}}'\hat{\boldsymbol{\nu}} = JK\sum_{i=1}^{I}(\mu^\circ+\alpha_i^\circ)^2 = JK\sum_{i=1}^{I}(y_{\bullet\bullet\bullet}+y_{i\bullet\bullet}-y_{\bullet\bullet\bullet})^2 = JK\sum_{i=1}^{I}y_{i\bullet\bullet}^2, \tag{54}$$

pričom $h(\mathbf{U}) = I$.

V submodeli $M_2: \quad \mathbf{Y} \sim N(\mathbf{T}\boldsymbol{\gamma}, \sigma^2\mathbf{I})$ je

$$\hat{\boldsymbol{\tau}} = \widehat{\mathbf{T}\boldsymbol{\gamma}} = (y_{\bullet\bullet\bullet}, y_{\bullet\bullet\bullet}, ..., y_{\bullet\bullet\bullet})'$$

a

$$\hat{\boldsymbol{\tau}}\hat{\boldsymbol{\tau}} = ny_{\bullet\bullet\bullet}^2 = IJKy_{\bullet\bullet\bullet}^2, \tag{55}$$

pričom $h(\mathbf{T}) = 1$.

Hypotéze $H_{B0}: \quad \beta_1 = \beta_2 = ... = \beta_J$ "zodpovedá" súčet štvorcov $S_B = (\hat{\boldsymbol{\mu}}-\hat{\boldsymbol{\nu}})'(\hat{\boldsymbol{\mu}}-\hat{\boldsymbol{\nu}})$

– súčet štvorcov, keď zdroj menlivosti sú stĺpce

– súčet štvorcov, keď zdroj menlivosti je spôsob hnojenia (v $\hat{\boldsymbol{\mu}}$ je uvažovaný, v $\hat{\boldsymbol{\nu}}$ nie je uvažovaný).

Z (53),(54) a (51) dostávame

$$S_B = (\hat{\boldsymbol{\mu}}-\hat{\boldsymbol{\nu}})'(\hat{\boldsymbol{\mu}}-\hat{\boldsymbol{\nu}}) = \hat{\boldsymbol{\mu}}'\hat{\boldsymbol{\mu}}-\hat{\boldsymbol{\nu}}'\hat{\boldsymbol{\nu}} = IK\sum_{j=1}^{J}y_{\bullet j\bullet}^2 - IJKy_{\bullet\bullet\bullet}^2 = IK\left(\sum_{j=1}^{J}y_{\bullet j\bullet}^2 - Jy_{\bullet\bullet\bullet}^2\right) = IK\overbrace{\sum_{j=1}^{J}(y_{\bullet j\bullet}-y_{\bullet\bullet\bullet})^2}^{\text{podľa } (51)}.$$

Hypotéze $H_{A0}: \quad \alpha_1 = \alpha_2 = ... = \alpha_I$ "zodpovedá" súčet štvorcov $S_A = (\hat{\boldsymbol{\nu}}-\hat{\boldsymbol{\tau}})'(\hat{\boldsymbol{\nu}}-\hat{\boldsymbol{\tau}})$

– súčet štvorcov, keď zdroj menlivosti sú riadky

– súčet štvorcov, keď zdroj menlivosti je typ pôdy (v $\hat{\boldsymbol{\nu}}$ je uvažovaný, v $\hat{\boldsymbol{\tau}}$ nie je uvažovaný).

$$S_A = (\hat{\boldsymbol{\nu}}-\hat{\boldsymbol{\tau}})'(\hat{\boldsymbol{\nu}}-\hat{\boldsymbol{\tau}}) = \hat{\boldsymbol{\nu}}'\hat{\boldsymbol{\nu}} - \hat{\boldsymbol{\tau}}'\hat{\boldsymbol{\tau}} = JK\sum_{i=1}^{I}y_{i\bullet\bullet}^2 - IJKy_{\bullet\bullet\bullet}^2 = JK\overbrace{\sum_{i=1}^{I}(y_{i\bullet\bullet}-y_{\bullet\bullet\bullet})^2}^{\text{podľa } (52)}.$$

Tiež platí

$$S_T = \mathbf{Y'Y} - \hat{\boldsymbol{\tau}}'\hat{\boldsymbol{\tau}} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} Y_{ijk}^2 - ny_{\bullet\bullet\bullet}^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} Y_{ijk}^2 - IJKy_{\bullet\bullet\bullet}^2$$

a

$$S_e = S_T - S_A - S_B.$$

Podľa Vety 4.2 $H_{B0}: \quad \beta_1 = \beta_2 = ... = \beta_J$ sa testuje pomocou (testovacej) štatistiky

$$F_B = \frac{(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\nu}})'(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\nu}})}{S_e} \frac{\overbrace{n - h(\mathbf{X})}^{IJK-I-J+1=f_e}}{\underbrace{h(\mathbf{X}) - h(\mathbf{U})}_{J-1=f_B}} = \frac{S_B/(J-1)}{S_e/(IJK - I - J + 1)} \sim F_{J-1,IJK-I-J+1} \tag{56}$$

(za platnosti H_{B0}). Hypotéza $H_{A0}: \quad \alpha_1 = \alpha_2 = ... = \alpha_I$ sa testuje pomocou (testovacej) štatistiky

$$F_A = \frac{(\hat{\boldsymbol{\nu}} - \hat{\boldsymbol{\tau}})'(\hat{\boldsymbol{\nu}} - \hat{\boldsymbol{\tau}})}{S_e} \frac{\overbrace{n - h(\mathbf{X})}^{IJK-I-J+1=f_e}}{\underbrace{h(\mathbf{U}) - h(\mathbf{T})}_{I-1=f_A}} = \frac{S_A/(I-1)}{S_e/(IJK - I - J + 1)} \sim F_{I-1,IJK-I-J+1} \tag{57}$$

(za platnosti H_{A0}).

Ak realizácia $F_B^{real} \geq F_{J-1,IJK-I-J+1}(1-\alpha)$, zamietame H_{B0} na hladine významnosti α. Ak H_{B0} nezamietame, môžeme pristúpiť k testovanie H_{A0}. V prípade, že realizácia $F_A^{real} \geq F_{I-1,IJK-I-J+1}(1-\alpha)$, zamietame H_{A0} na hladine významnosti α.

Tabuľka (vyváženej) analýzy rozptylu dvojného triedenia bez interakcií.

zdroj variability	súčet štvorcov	stupne voľnosti	S/f	$F = \frac{S/f}{S_e/f_e}$
riadky (typ pôdy)	S_A	$f_A = h(\mathbf{U}) - h(\mathbf{T}) = I - 1$	$\frac{S_A}{f_A}$	$\frac{S_A/f_A}{S_e/f_e} = F_{I-1,IJK-I-J+1}$
stĺpce (spôsob hnojenia)	S_B	$f_B = h(\mathbf{X}) - h(\mathbf{U}) = J - 1$	$\frac{S_B}{f_B}$	$\frac{S_B/f_B}{S_e/f_e} = F_{J-1,IJK-I-J+1}$
reziduály	S_e	$f_e = n - h(\mathbf{X}) = IJK - I - J + 1$	$\frac{S_e}{f_e}$	–
celkový	S_T	$f_T = f_e + f_A + f_B = IJK - 1$		

Poznámka 5.6: Ak ide o vyvážený model, je jedno, či najprv testujeme H_{B0} a potom H_{A0}, alebo naopak (testovacie štatistiky vyjdú rovnako). Ak je model nevyvážený, sú to rozdielne cesty a interpretácia je ťažká. Podrobnejšie pozri v knižke Anděl, J., *Matematická statistika*, SNTL, Praha, 1985, str. 160.

Poznámka 5.7: Ak zamietame H_{B0} alebo H_{A0}, Scheffého (alebo niekedy aj Tukeyovou) metódou sa zisťuje, medzi ktorými úrovňami faktorov sú rozdiely (pozri v knižke Anděl, J., *Matematická statistika*, SNTL, Praha, 1985, str. 161).

5.3 Dvojné triedenie s interakciami

Môže sa stať u dvojného triedenia, že efekty riadkov a stĺpcov sa jednoducho nesčítajú. Napr. v Príklade 5.4 by mohlo dôjsť k tomu, že niektorý druh hnojiva má špecifický účinok s istým druhom pôdy. Preto sa uvažuje (vo všeobecnosti) realistickejší model (47) teda

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk},$$

$\epsilon_{ijk} \sim N(0, \sigma^2)$, $i = 1, 2, ..., I$, $j = 1, 2, ..., J$, $k = 1, 2, ..., n_{ij}$ a všetky ϵ_{ijk} sú nezávislé. Je to *model dvojného triedenie s interakciami*. Testy v tomto modeli pozri v knižke Anděl, J., *Matematická statistika*, SNTL, Praha, 1985, str. 164.

Samozrejme uvažujú sa aj modely trojného a vyšších triedení aj s interakciami (aj vyšších rádov), pozri tiež napr. v IX. kapitole knižky Anděl, J., *Matematická statistika*, SNTL, Praha, 1985.