KOEFICIENT DETERMINACE V REGRESI S CHYBAMI V OBOU PROMĚNNÝCH

KAREL ZVÁRA

ABSTRACT. **Koeficient determinace lze v lineární regresi s absolutním členem definovat řadou způsobů. Který z nich použít, když chceme koeficient determinace použít v modelu jednoduché lineární funkční závislosti?**

0. ÚVOD

Kvalitu vysvětlení dat vyšetřovanou *regresní* závislostí měříme často pomocí koeficientu determinace. Když chceme tento pojem rozšířit i na jiný model závislosti (ne nutně regresní), můžeme se pokusit k tomu účelu použít některé z dostupných zobecnění koeficientu determinace. V tomto článku se budeme věnovat modelům funkční a strukturní závislosti.

První z těchto úloh lze formulovat následovně. Nechť

$$\text{(0.1)} \qquad \begin{pmatrix} x_i \\ y_i \end{pmatrix} \sim \mathsf{N}\left(\begin{pmatrix} \xi_i \\ \alpha + \beta\xi_i \end{pmatrix}, \sigma^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right), \quad 1 \le i \le n,$$

jsou nezávislé náhodné vektory, v nichž $\alpha, \beta, \xi_1, \ldots, \xi_n, \sigma^2$ jsou neznámé parametry. V literatuře se v této souvislosti hovoří o *funkční* závislosti nebo o regresní přímce s chybami v obou proměnných. V případě, že bychom $\xi_1, \ldots, \xi_n$ chápali jako nezávislé realizace náhodné veličiny ξ, šlo by o *strukturní* model a v (0.1) by bylo rozdělení měřených veličin podmíněné neznámými hodnotami $\xi_1, \ldots, \xi_n$. Standardní předpoklad $\xi \sim \mathsf{N}(\delta, \omega^2)$ by pak vedl k nepodmíněnému rozdělení

$$\text{(0.2)} \qquad \begin{pmatrix} x_i \\ y_i \end{pmatrix} \sim \mathsf{N}\left(\begin{pmatrix} \delta \\ \alpha + \beta\delta \end{pmatrix}, \begin{pmatrix} \omega^2 + \sigma^2 & \beta\omega^2 \\ \beta\omega^2 & \beta^2\omega^2 + \sigma^2 \end{pmatrix} \right), \quad 1 \le i \le n,$$

V tomto případě mají všechny uvažované vektory stejné normální rozdělení s korelačním koeficientem

$$\rho = \frac{\beta\omega^2}{\sqrt{(\omega^2 + \sigma^2)(\beta^2\omega^2 + \sigma^2)}}.$$

Při asymptotických úvahách musíme ve funkčním modelu předpokládat cosi o asymptotickém chování parametrů $\xi_1, \ldots, \xi_n$. Předpoklady

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \xi_i = \delta,$$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (\xi_i - \delta)^2 = \omega^2$$

způsobí, že mnohé asymptotické výrazy budou ve funkčním i strukturním modelu formálně stejné.

Typeset by $\mathcal{A}\mathcal{M}\mathcal{S}$-TEX

1. Definice koeficientu determinace

Připomeňme několik možných definic koeficientu determinace. (Podrobný přehled je uveden např. v [2].) Všechny definice vycházejí z předpokladu, že

$$y_i \sim \mathsf{N}\left(\alpha + \boldsymbol{x}_i'\beta, \sigma^2\right), \quad 1 \le i \le n,$$

jsou nezávislé náhodné veličiny, α, β a σ^2 jsou neznámé parametry a $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ jsou dané vektory. Označíme-li jako a a b odhady parametrů α a β metodou nejmenších čtverců, lze vyrovnané hodnoty počítat pomocí

$$\hat{y}_i = a + \boldsymbol{x}_i'b, \quad 1 \le i \le n.$$

Rezidua u_i se pak definují jako $u_i = y_i - \hat{y}_i$. Součet čtverců reziduí označíme symbolem RSS.

Kdykoliv budeme v dalším textu psát explicitní vzorce pro jednoduchou regresní přímku nebo pro model z Úvodu, bude užitečné označení

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2, \quad S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}), \quad S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2,$$
$$s_{xx} = S_{xx}/n, \quad s_{xy} = S_{xy}/n, \quad s_{yy} = S_{yy}/n.$$

1.1. Víme, že lineární regresní funkce maximalizuje koeficient mnohonásobné korelace mezi náhodnou veličinou (*závisle* proměnnou) Y a *nezávisle* proměnnými $\boldsymbol{X}$. V případě jednoduché lineární regrese jde o klasický Pearsonův korelační koeficient. Čtverec maximální hodnoty výběrového koeficientu mnohonásobné korelace se pak nazývá **koeficient determinace** a značí R^2. V případě mnohorozměrného normálního rozdělení (Y i $\boldsymbol{X}$ chápeme jako náhodné) je R_1^2 konzistentním odhadem čtverce koeficientu mnohonásobné korelace $\rho^2_{Y,\boldsymbol{X}}$. Je známo, že tento koeficient mnohonásobné korelace je roven Pearsonovu koeficientu korelace mezi náhodnou veličinou Y a její nejlepší lineární predikcí pomocí $\boldsymbol{X}$, kterou můžeme označit symbolem $\hat{Y}(\boldsymbol{X})$. Tento vztah platí i pro výběrové protějšky, takže je také

$$R_1^2 = r^2_{y,\hat{y}} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}. \tag{1.1}$$

Speciálně pro jednoduchou regresní přímku dostaneme

$$R_1^2 = \frac{Sxy^2}{SxxSyy}. \tag{1.2}$$

1.2. Jiná interpretace koeficientu determinace má původ v analýze rozptylu. Když se snažíme vysvětlit variabilitu náhodné veličiny Y pomocí funkce konkrétní realizace $\boldsymbol{x}$ náhodného vektoru $\boldsymbol{X}$, zjistíme, že nejlepším prediktorem $\hat{Y}(\boldsymbol{x})$ je podmíněná střední hodnota $\mathsf{E}(Y|\boldsymbol{X} = \boldsymbol{x})$. V případě sdruženého normálního rozdělení je tato střední hodnota lineární funkcí vektoru $\boldsymbol{x}$. Bez ohledu na předpoklad normality platí pro rozptyly vztah

$$\mathsf{var}(Y) = \mathsf{var}(Y.\boldsymbol{x}) + \mathsf{var}(\hat{Y}(\boldsymbol{x})). \tag{1.3}$$

Veličina

$$1 - \frac{\mathsf{var}(Y.x)}{\mathsf{var}(Y)}$$

ukazuje, jak velký díl variability náhodné veličiny Y vysvětluje lineární závislost. Výběrovým protějškem posledního vztahu je alternativní definice koeficientu determinace

$$R_2^2 = 1 - \frac{RSS}{Syy}. \tag{1.4}$$

Není těžké dokázat, že v doposud uvažovaném modelu platí $R_1^2 = R_2^2$.

1.3. Další dvě vyjádření koeficientu determinace odvodíme z testů hypotézy $\mathsf{H}_0 : \boldsymbol{\beta} = \mathbf{0}$. Inspiraci lze najít v [3]. Klasickým postupem je test poměrem věrohodnosti. Sdružená hustota náhodného vektoru $\boldsymbol{y}$ je rovna

$$h(\boldsymbol{y}) = (2\pi\sigma^2)^{n/2} \exp\left(-\sum_{i=1}^{n} (y_i - \alpha - \boldsymbol{x}_i'\boldsymbol{\beta})^2/(2\sigma^2) \right),$$

takže logaritmus věrohodnostní funkce je roven

$$l(\alpha, \boldsymbol{\beta}, \sigma^2) = c - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \boldsymbol{x}_i'\boldsymbol{\beta})^2.$$

Maximalizace dá pro $\alpha, \boldsymbol{\beta}$ odhady totožné s odhady metodou nejmenších čtverců a $\widehat{\sigma^2} = RSS/n$. Výsledná hodnota logaritmu věrohodnostní funkce je rovna $\widehat{l} = c_1 - \frac{n}{2}\ln\widehat{\sigma^2}$. Podobně za platnosti hypotézy $\mathsf{H}_0 : \boldsymbol{\beta} = \mathbf{0}$ dostaneme odhady $\widetilde{\alpha} = \bar{y}$ a $\widetilde{\sigma^2} = S_{yy}/n$. Proto je $\widetilde{l} = c_1 - \frac{n}{2}\ln\widetilde{\sigma^2}$.

Známý asymptotický test poměrem věrohodnosti je založen na statistice $2(\widehat{l} - \widetilde{l})$, která má v regulárním případě za platnosti testované hypotézy asymptoticky rozdělení χ^2. V našem případě dostaneme

$$\begin{aligned} \widehat{l} - \widetilde{l} &= -\frac{n}{2}(\ln\widehat{\sigma^2} - \ln\widetilde{\sigma^2}) \\ &= -\frac{n}{2}\ln\frac{RSS}{Syy} \\ &= -\frac{n}{2}\ln(1 - R_2^2). \end{aligned}$$

Odtud snadno dostaneme novou definici koeficientu determinace

$$R_3^2 = 1 - \exp\left(-\frac{2}{n}(\widehat{l} - \widetilde{l}) \right). \tag{1.5}$$

1.4. Waldův test hypotézy, která tvrdí, že $g(\boldsymbol{\theta}) = \mathbf{0}$ (kde $g(\boldsymbol{\theta})$ je vhodná hladká funkce a $\boldsymbol{\theta}$ je vektor neznámých parametrů), spočívá v tom, že do této funkce zkusíme dosadit maximálně věrohodný odhad vektoru parametrů a vezmeme v úvahu jeho asymptotickou varianční matici:

$$W = g(\widehat{\boldsymbol{\theta}})' \left(\widehat{\mathsf{var}(g(\widehat{\boldsymbol{\theta}}))} \right)^{-1} g(\widehat{\boldsymbol{\theta}}).$$

Za standardních předpokladů regularity je asymptotická varianční matice odhadu θ dána inverzní maticí k Fisherově informační matici. V našem konkrétním případě lineární regrese je maximálně věrohodným odhadem vektoru β odpovídající odhad b metodou nejmenších čtverců s varianční maticí

$$\begin{aligned}\mathrm{var}(b) &= \sigma^2 \left(\sum_{i1}^{n} x_i x_i' \right)^{-1} \\ &= \sigma^2 (X'X)^{-1}.\end{aligned}$$

Použijme opět odhad $\widehat{\sigma^2} = RSS/n$. Protože v uvažovaném případě platí $RSS = S_{yy} - b'(X'X)b$, po malých úpravách ověříme, že i následující definice koeficientu determinace je ekvivalentní v případě lineární regrese s definicí (1.4) :

$$R_4^2 = \frac{W}{W+n}. \tag{1.6}$$

1.5. S posledními dvěma testy (poměrem věrohodnosti a Waldův test) se v učebnicích uvádí často ještě další, s nimi asymptoticky ekvivalentní test, založený na ověření, zda je příslušný Lagrangeův multiplikátor dostatečně blízko nule (viz např. [3] nebo (6e.3.6) v [4]). Označme jako $v(\theta)$ vektor parciálních derivací logaritmu věrohodnostní funkce a jako $F(\theta)$ odpovídající Fisherovu informační matici příslušnou n pozorováním. Označíme-li jako $\tilde{\theta}$ maximálně věrohodný odhad vektoru parametrů za hypotézy, pak má Raova statistika tvar

$$S = v(\tilde{\theta})' \left(F(\tilde{\theta}) \right)^{-1} v(\tilde{\theta}).$$

Po určité námaze dostaneme

$$v(\alpha, \beta, \sigma^2) = \frac{1}{\sigma^2} \begin{pmatrix} 1'(y - \alpha 1 - X\beta) \\ X'(y - \alpha 1 - X\beta) \\ -(1/n) + (y - \alpha 1 - X\beta)'(y - \alpha 1 - X\beta)/(2\sigma^2) \end{pmatrix},$$

$$F(\alpha, \beta, \sigma^2) = \frac{1}{\sigma^2} \begin{pmatrix} 1'1 & 1'X & 0 \\ X'1 & X'X & 0 \\ 0 & 0' & n/(2\sigma^2) \end{pmatrix}.$$

Použijeme-li toho, že za platnosti hypotézy $\mathsf{H}_0 : \beta = 0$ je

$$\widetilde{\sigma^2} = s_{yy},$$

dostaneme po úpravách

$$S = n \left(1 - \frac{RSS}{S_{yy}} \right).$$

Zřejmě je tedy souvislost s koeficientem determinace dána vztahem

$$R_5^2 = \frac{1}{n} S. \tag{1.7}$$

2. Funkční model

Vraťme se k funkčnímu modelu z úvodu. Zlogaritmujeme-li sdruženou hustotu náhodného vektoru x, y, dostaneme logaritmus věrohodnostní funkce

$$l(\alpha,\beta,\sigma^2,\xi) = -n\ln(2\pi) - n\ln(\sigma^2) - \frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(x_i-\xi_i)^2 + \sum_{i=1}^{n}(y_i-\alpha-\beta\xi_i)^2\right].$$

Při hledání odhadů metodou maximální věrohodnosti dostaneme dostaneme soustavu rovnic

$$\sum_{i=1}^{n}(y_i-\alpha-\beta\xi_i) = 0, \tag{2.1}$$

$$\sum_{i=1}^{n}\xi_i(y_i-\alpha-\beta\xi_i) = 0, \tag{2.2}$$

$$(x_i-\xi_i) + (y_i-\alpha-\beta\xi_i) = 0, \quad 1 \le i \le n, \tag{2.3}$$

$$-\frac{n}{\sigma^2} + \frac{1}{2\sigma^4}\left[\sum_{i=1}^{n}(x_i-\xi_i)^2 + \sum_{i=1}^{n}(y_i-\alpha-\beta\xi_i)^2\right] = 0. \tag{2.4}$$

Ze vztahů (2.1) a (2.3) dostaneme po úpravách

$$\frac{1}{n}\sum_{i=1}^{n}\widehat{\xi}_i = \bar{x}, \tag{2.5}$$

$$\widehat{\alpha} = \bar{y} - \beta\bar{x},$$

$$\widehat{\xi}_i = x_i + \frac{\beta}{1+\beta^2}\left((y_i-\bar{y}) - (\beta(x_i-\bar{x})\right), \quad 1 \le i \le n, \tag{2.6}$$

$$\widehat{\xi}_i = \bar{x} + \frac{1}{1+\beta^2}\left((x_i-\bar{x}) + (\beta(y_i-\bar{y})\right), \quad 1 \le i \le n. \tag{2.7}$$

Když odtud dosadíme do (2.2), pak s využitím (2.1) dojdeme ke kvadratické rovnici pro β:

$$S_{xy}\beta^2 - (S_{yy}-S_{xx})\beta - S_{xy} = 0. \tag{2.8}$$

Lze dokázat, že věrohodnostní funkci minimalizuje pouze jeden kořen této kvadratické rovnice (předpokládáme $S_{xy} \neq 0$):

$$\widehat{\beta} = \frac{(S_{yy}-S_{xx}) + \sqrt{(S_{yy}-S_{xx})^2 + 4S_{xy}^2}}{2S_{xy}}. \tag{2.9}$$

Při dalších úpravách je užitečné vědět, že z (2.8) plyne vztah

$$S_{yy} - 2\widehat{\beta}S_{xy} + \widehat{\beta}^2 S_{xx} = (1+\widehat{\beta}^2)(S_{yy} - \widehat{\beta}S_{xx}). \tag{2.10}$$

Odtud je mimo jiné

$$RSS_y = \sum_{i=1}^{n} \left(y_i - \widehat{\alpha} - \widehat{\beta}\widehat{\xi}_i\right)^2$$

$$= \frac{1}{1+\widehat{\beta}^2}(S_{yy} - \widehat{\beta}S_{xy}), \tag{2.11}$$

$$RSS_x = \sum_{i=1}^{n} \left(x_i - \widehat{\xi}_i\right)^2$$

$$= \frac{\widehat{\beta}^2}{1+\widehat{\beta}^2}(S_{yy} - \widehat{\beta}S_{xy}), \tag{2.12}$$

takže je dohromady

$$RSS_{x+y} = RSS_x + RSS_y$$

$$= S_{yy} - \widehat{\beta}S_{xy}. \tag{2.13}$$

V této souvislosti je zajímavé si připomenout, že u klasické regresní přímky platí

$$RSS = S_{yy} - bS_{xy}.$$

Dosadíme-li z (2.13) do (2.4), dostaneme odhad parametru σ^2:

$$s^2 = \frac{RSS_{x+y}}{2n}$$

$$= \frac{s_{yy} - \widehat{\beta}s_{xy}}{2}.$$

Uvážíme-li, že za předpokladů uvedených v úvodu konvergují veličiny s_{xx}, s_{xy}, s_{yy} podle pravděpodobnosti k prvkům varianční matice, totiž po řadě k výrazům

$$\omega^2 + \sigma^2, \quad \beta\omega^2, \quad \beta^2\omega^2 + \sigma^2,$$

zjistíme konzistenci odhadů $\widehat{\alpha}, \widehat{\beta}$. Ovšem dále zjistíme, že je

$$\operatorname{plim} s^2 = \operatorname{plim} \frac{s_{yy} - \widehat{\beta}s_{xy}}{2}$$

$$= \frac{\beta^2\omega^2 + \sigma^2 - \beta\beta\omega^2}{2}$$

$$= \frac{\sigma^2}{2}.$$

Konzistentním odhadem parametru σ^2 tedy bude nikoliv s^2, ale

$$\widehat{\sigma^2} = \frac{RSS_{x+y}}{n} \tag{2.14}$$

$$= s_{yy} - \widehat{\beta}s_{xy}$$

Nyní nalezneme výrazy pro R^2 podle jednotlivých vztahů z 1. kapitoly.

2.1. Protože je podle (1.5) průměrná hodnota odhadů $\widehat{\xi}_i$ totožná s $\bar{x}$, platí nutně $\bar{\widehat{y}} = \bar{y}$. Dále je

$$\begin{aligned}\sum_{i=1}^{n}(y_i-\bar{y})(\widehat{y}_i-\bar{\widehat{y}}) &= \sum_{i=1}^{n}(y_i-\bar{y})(\widehat{\alpha}+\widehat{\beta}\widehat{\xi}_i-\bar{y})\\ &= \sum_{i=1}^{n}(y_i-\bar{y})\widehat{\beta}(\widehat{\xi}_i-\bar{x})\\ &= \frac{\widehat{\beta}}{1+\widehat{\beta}^2}\sum_{i=1}^{n}(y_i-\bar{y})\Big((x_i-\bar{x})+\widehat{\beta}(y_i-\bar{y})\Big)\\ &= \frac{\widehat{\beta}}{1+\widehat{\beta}^2}\Big(S_{xy}+\widehat{\beta}S_{yy}\Big),\end{aligned}$$

takže po dosazení do definice (1.1) dostaneme po malé úpravě

$$R^2_{1y} = \frac{(S_{xy}+\widehat{\beta}S_{yy})^2}{S_{yy}(S_{xx}+2\widehat{\beta}S_{xy}+\widehat{\beta}^2S_{yy})}.$$

S využitím skutečnosti, že $\widehat{\beta}$ je řešením kvadratické rovnice (2.8), lze poslední vztah zjednodušit na

$$R^2_{1y} = \frac{\widehat{\beta}}{1+\widehat{\beta}^2}\frac{\widehat{\beta}s_{yy}+s_{xy}}{s_{yy}}. \tag{2.15}$$

2.2. Do vztahu (1.4) snadno dosadíme místo reziduálního součtu čtverců výraz RSS_y, což vede k

$$\begin{aligned}R^2_{2y} &= 1-\frac{RSS_y}{S_{yy}},\\ &= 1-\frac{s_{yy}-\widehat{\beta}s_{xy}}{(1+\widehat{\beta}^2)s_{yy}},\\ &= R^2_{1y}.\end{aligned}$$

Symboly y v indexech již naznačily následující "kacířskou" myšlenku: místo vysvětlování pouze hodnot veličiny Y se pokusit vysvětlit variabilitu obou měřených veličin, protože obě měříme s náhodnou chybou. To vede k výrazu

$$\begin{aligned}R^2_{2x+y} &= 1-\frac{RSS_{x+y}}{S_{xx}+S_{yy}},\\ &= 1-\frac{S_{yy}-\widehat{\beta}S_{xy}}{S_{xx}+S_{yy}},\\ &= \frac{s_{xx}+\widehat{\beta}s_{xy}}{s_{xx}+s_{yy}}.\end{aligned} \tag{2.16}$$

Jako cvičení nechávám čtenáři ověření tvrzení, že stejný výraz dostaneme, když vyjdeme z definice (1.1).

Je zajímavé porovnat uvedené dvě verse koeficientu determinace R_2^2. Po snadné úpravě, s využitím skutečnosti, že s pravděpodobností 1 je $RSS_{x+y} > 0$, zjistíme, že s pravděpodobností 1 platí ekvivalence

$$R_{2y}^2 \leq R_{2x+y}^2 \Leftrightarrow \widehat{\beta}^2 \leq \frac{s_{xx}}{s_{yy}}. \tag{2.17}$$

2.3. Odhady parametrů metodou maximální věrohodnosti jsou uvedeny v úvodu této kapitoly. Dosadíme-li tyto odhady do logaritmu věrohodnostní funkce, dostaneme

$$\widehat{l} = -n \ln(s_{yy} - \widehat{\beta} s_{xy}).$$

Za platnosti hypotézy $H_0 : \beta = 0$ vyjde zřejmě

$$\widetilde{\alpha} = \bar{y}, \quad \widetilde{\xi}_i = x_i, \quad 1 \leq i \leq n.$$

Podobně jako dříve dostaneme, že odhad

$$\widetilde{\sigma^2} = s_{yy}$$

je konzistentním odhadem rozptylu σ^2. Je tedy (s uvážením, že tentokrát máme k dispozici $2n$ pozorování)

$$\begin{aligned} R_3^2 &= 1 - \exp\left(-\frac{1}{n}(\widehat{l} - \widetilde{l})\right), \\ &= 1 - \frac{s_{yy} - \widehat{\beta} s_{xy}}{s_{yy}}, \\ &= \widehat{\beta}\frac{s_{xy}}{s_{yy}}. \end{aligned} \tag{2.18}$$

V této souvislosti je zajímavé, že v klasické situaci lze koeficient determinace (1.2) upravit do podobného tvaru

$$R^2 = b\frac{s_{xy}}{s_{yy}}.$$

Protože je $R^2 \geq 0$, platí ekvivalence

$$R_3^2 \geq r_{x,y}^2 \Leftrightarrow |\widehat{\beta}| \geq |b|.$$

Lze ovšem snadno dokázat, že nerovnost na pravé straně platí vždy, kdy je $s_{xx} > 0$. Můžeme tedy s pravděpodobností 1 tvrdit, že je

$$R_3^2 \geq r_{x,y}^2. \tag{2.19}$$

S využitím skutečnosti, že s pravděpodobností 1 musí být $s_{yy} - \widehat{\beta} s_{xy} > 0$, snadno zjistíme, že je

$$R_3^2 \leq R_{2y}^2, \tag{2.20}$$

$$R_3^2 \leq R_{2x+y}^2. \tag{2.21}$$

2.4. Abychom spočítali Fisherovu informační matici F, vyjádříme nejprve jednotlivé derivace logaritmu věrohodnostní funkce pomocí náhodných veličin

$$\begin{pmatrix} e_i \\ f_i \end{pmatrix} = \begin{pmatrix} y_i - \alpha - \beta\xi_i \\ x_i - \xi_i \end{pmatrix} \sim \mathsf{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right), \quad 1 \le i \le n.$$

Snadno tak dostaneme

$$\frac{\partial l}{\partial \alpha} = \frac{1}{\sigma^2} \sum_{i=1}^{n} e_i,$$

$$\frac{\partial l}{\partial \beta} = \frac{1}{\sigma^2} \sum_{i=1}^{n} \xi_i e_i,$$

$$\frac{\partial l}{\partial \xi_i} = \frac{1}{\sigma^2}(f_i + e_i), \quad 1 \le i \le n,$$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{1}{\sigma^2} + \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(e_i^2 + f_i^2)\right).$$

Odtud postupně vyjde

$$\mathsf{E}\left(\frac{\partial l}{\partial \alpha}\right)^2 = \frac{n}{\sigma^2},$$

$$\mathsf{E}\left(\frac{\partial l}{\partial \alpha}\right)\left(\frac{\partial l}{\partial \beta}\right) = \frac{\sum_{i=1}^{n} \xi_i}{\sigma^2},$$

$$\mathsf{E}\left(\frac{\partial l}{\partial \alpha}\right)\left(\frac{\partial l}{\partial \xi_i}\right) = \frac{\beta}{\sigma^2}, \quad 1 \le i \le n,$$

$$\mathsf{E}\left(\frac{\partial l}{\partial \alpha}\right)\left(\frac{\partial l}{\partial \sigma^2}\right) = 0,$$

$$\mathsf{E}\left(\frac{\partial l}{\partial \beta}\right)^2 = \frac{\sum_{i=1}^{n} \xi_i^2}{\sigma^2},$$

$$\mathsf{E}\left(\frac{\partial l}{\partial \beta}\right)\left(\frac{\partial l}{\partial \xi_i}\right) = \frac{\beta\xi_i}{\sigma^2}, \quad 1 \le i \le n,$$

$$\mathsf{E}\left(\frac{\partial l}{\partial \beta}\right)\left(\frac{\partial l}{\partial \sigma^2}\right) = 0,$$

$$\mathsf{E}\left(\frac{\partial l}{\partial \xi_i}\right)\left(\frac{\partial l}{\partial \xi_j}\right) = \frac{1+\beta^2}{\sigma^2}\delta_{ij}, \quad 1 \le i, j \le n,$$

$$\mathsf{E}\left(\frac{\partial l}{\partial \xi_i}\right)\left(\frac{\partial l}{\partial \sigma^2}\right) = 0, \quad 1 \le i \le n,$$

$$\mathsf{E}\left(\frac{\partial l}{\partial \sigma^2}\right)^2 = \frac{n}{\sigma^4}.$$

Informační matici vztaženou k vektoru parametrů $(\alpha, \beta, \xi_1, \ldots, \xi_n, \sigma^2)$ tedy lze zapsat jako

$$F = \frac{1}{\sigma^2}\begin{pmatrix} n & \sum\xi_i & \beta\mathbf{1}' & 0 \\ \sum\xi_i & \sum\xi^2 & \beta\boldsymbol{\xi}' & 0 \\ \beta\mathbf{1} & \beta\boldsymbol{\xi} & (1+\beta^2)\mathbf{I} & 0 \\ 0 & 0 & 0' & n/\sigma^2 \end{pmatrix}. \tag{2.22}$$

Protože se hypotéza vztahuje k jedinému parametru β, stačí, když nalezneme prvek na místě $(2,2)$ matice F^{-1}. Označme levý horní roh této inverzní matice rozměru 2×2 symbolem Q. Vzhledem k speciálnímu tvaru invertované matice platí podle často používaného vzorce pro inverzi matice rozdělené na 4 podmatice (viz např. [5], str. 33)

$$\begin{aligned}\sigma^2 Q^{-1} &= \begin{pmatrix} n & \sum\xi_i \\ \sum\xi_i & \sum\xi_i^2 \end{pmatrix} - \frac{\beta^2}{1+\beta^2}\begin{pmatrix} n & \sum\xi_i \\ \sum\xi_i & \sum\xi_i^2 \end{pmatrix}, \\ &= \frac{1}{1+\beta^2}\begin{pmatrix} n & \sum\xi_i \\ \sum\xi_i & \sum\xi_i^2 \end{pmatrix}.\end{aligned}$$

Odtud je hledaný prvek inverze Fisherovy informační matice roven

$$\begin{aligned} q_{22} &= \sigma^2(1+\beta^2)\frac{n}{\det Q}, \\ &= \sigma^2\frac{1+\beta^2}{\sum(\xi_i - \bar{\xi})^2}. \end{aligned}$$

Opět si připomeňme klasickou regresní přímku, kde platí

$$\operatorname{var}(b) = \sigma^2\frac{1}{\sum(x_i - \bar{x})^2}.$$

Z (2.7) plyne, že je

$$\sum_{i=1}^{n}(\hat{\xi}_i - \bar{\xi})^2 = \frac{S_{xx} + 2\hat{\beta}S_{xy} + \hat{\beta}^2 S_{yy}}{(1+\hat{\beta}^2)^2}.$$

Waldova statistika W je tedy rovna

$$\begin{aligned} W &= \frac{\hat{\beta}^2}{\operatorname{var}(\hat{\beta})}, \\ &= \frac{\hat{\beta}^2}{\hat{\sigma}^2(1+\hat{\beta}^2)^3}(S_{xx} + 2\hat{\beta}S_{xy} + \hat{\beta}^2 S_{yy}). \end{aligned} \tag{2.23}$$

Zbývá zapsat koeficient determinace podle (1.6), kam dosadíme skutečně n (jmenovatel z použitého konzistentního odhadu rozptylu σ^2) a nikoliv počet pozorování $2n$. Po úpravách vyjde

$$\begin{aligned} R_4^2 &= \frac{W}{W+n} \\ &= \frac{\hat{\beta}^2}{1+\hat{\beta}^2}\frac{s_{xx} + 2\hat{\beta}s_{xy} + \hat{\beta}^2 s_{yy}}{(1+\hat{\beta}^2+\hat{\beta}^4)(s_{yy} - \hat{\beta}s_{xy}) + \hat{\beta}^2(s_{xx} + s_{yy})}. \end{aligned} \tag{2.24}$$

2.5. Za platnosti hypotézy dostáváme konzistentní odhady ve tvaru

$$\tilde{\alpha} = \bar{y}, \quad \tilde{\beta} = 0, \quad \tilde{\xi} = x, \quad \widetilde{\sigma^2} = s_{yy},$$

což vede k

$$v(\tilde{\theta}) = \frac{1}{s_{yy}} \begin{pmatrix} 0 \\ S_{xy} \\ 0 \\ 0 \end{pmatrix},$$

$$F(\tilde{\theta})^{-1} = \frac{1}{s_{yy}} \begin{pmatrix} * & * & 0' & 0 \\ * & S_{xx}^{-1} & 0' & 0 \\ 0 & 0 & \mathbf{I} & 0 \\ 0 & 0 & 0' & s_{yy}/n \end{pmatrix},$$

kde jsem označili hvězdičkou ty výrazy, které v dalším výpočtu nepotřebujeme. Po úpravě tedy dostaneme

$$S = n \frac{S_{xy}^2}{S_{xx} S_{yy}},$$

což po dosazení do (1.7) dá možná poněkud překvapivý výsledek

(2.25)
$$R_5^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}} = r_{xy}^2.$$

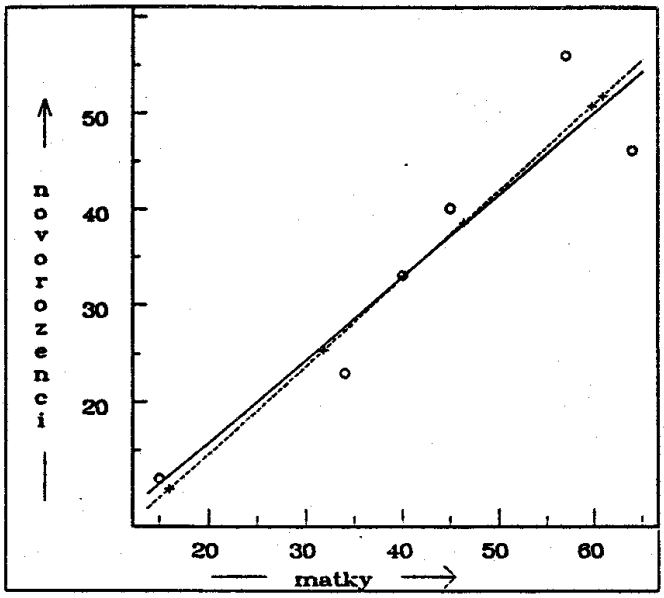
3. Příklad

Numerickou ilustraci provedeme na známých datech o koncentraci kyseliny mléčné v krvi u matek (x) a u jejich novorozenců (y) (viz [1]). Na obrázku je vyznačeno výchozích 6 párů pozorování, souvislou čarou je vyznačena standardní regresní přímka pro závislost y na x, čárkovaně je vyznačena přímka spočítaná podle 2. kapitoly i s odhadnutými body o souřadnicích $(\hat{\xi}_i, \hat{y}_i)$ na této přímce. Na silnou závislost ukazují hodnoty koeficientů spočítaných podle (2.15), (2.16), (2.18), (2.24) a (2.25):

$$R_{1y}^2 = 0{,}961,$$
$$R_{2x+y}^2 = 0{,}968,$$
$$R_3^2 = 0{,}929,$$
$$R_4^2 = 0{,}881,$$
$$R_5^2 = 0{,}874.$$

Pro zajímavost, statistiky W, S a $2(\hat{l} - \tilde{l})$, které asymptoticky mají za platnosti hypotézy stejné ξ^2 rozdělení s jedním stupněm volnosti dají

$$W = 44{,}5,$$
$$S = 5{,}2$$
$$2(\hat{l} - \tilde{l}) = 15{,}9.$$



OBR. 1. Koncentrace kyseliny mléčné [mg/ml]

Článek vznikl za finanční podpory grantu číslo 313/93/0616 Grantové agentury České republiky.

REFERENCES

1. Anděl J., *Matematická statistika*, SNTL, Praha, 1978.
2. Kvålseth T.O., *Cautionary note about R^2*, The American Statistician 39 (1985), 279–285.
3. Magge L., *R^2 measures based on Wald and likelihood ratio joint significance tests*, The American Statistician 44 (1990), 251–253.
4. Rao, C.R., *Lineární metody statistické indukce a jejich aplikace*, Academia, Praha, 1978.
5. Zvára K., *Regresní analýza*, Academia, Praha, 1989.