7 Následky nesplnění předpokladů

V lineárním modelu jsme předpokládali, že známe prostor možných středních hodnot, že všechna pozorování mají stejný rozptyl, že jsou nekorelovaná (resp. nezávislá) a že mají normální rozdělení. Nyní se pokusíme popsat následky, které má nesplnění některého z uvedených předpokladů.

7.1 Prostor středních hodnot

Předpokládejme, že platí

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{e}, \qquad \mathbf{e} \sim (\mathbf{0}, \sigma^2\mathbf{I}), \tag{7.1}$$

přestože my předpokládáme platnost modelu $\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$.

7.1.1 Vlastnosti odhadu $\mathsf{E}\,\mathbf{Y}$

Označme $\mathbf{G} = (\mathbf{X}, \mathbf{Z})$ a $\boldsymbol{\delta} = (\boldsymbol{\beta}', \boldsymbol{\gamma}')'$ a veškeré statistiky vztažené k modelu $\mathbf{Y} \sim (\mathbf{G}\boldsymbol{\delta}, \sigma^2\mathbf{I})$ označíme dolním indexem g. Běžný odhad vektoru $\mathsf{E}\,\mathbf{Y}$ je tedy

$$\hat{\mathbf{Y}}_g = \mathbf{G}(\mathbf{G}'\mathbf{G})^-\mathbf{G}'\mathbf{Y}, \tag{7.2}$$

což je, jak víme např. z (3.13), průmět $\mathbf{Y}$ do $\mathcal{M}(\mathbf{X}, \mathbf{Z}) = \mathcal{M}(\mathbf{X}, \mathbf{MZ})$. S použitím druhého vyjádření dostaneme

$$\begin{aligned}
\hat{\mathbf{Y}}_g &= (\mathbf{X}, \mathbf{MZ}) \begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{O} \\ \mathbf{O} & \mathbf{Z}'\mathbf{MZ} \end{pmatrix}^- \begin{pmatrix} \mathbf{X}' \\ \mathbf{Z}'\mathbf{M} \end{pmatrix} \mathbf{Y} \\
&= \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{Y} + \mathbf{MZ}(\mathbf{Z}'\mathbf{MZ})^-\mathbf{Z}'\mathbf{MY} \\
&= \hat{\mathbf{Y}} + \mathbf{MZ}(\mathbf{Z}'\mathbf{MZ})^-\mathbf{Z}'\mathbf{u} \qquad (7.3) \\
&= \mathbf{X}\mathbf{b}_g + \mathbf{Z}\mathbf{c}_g, \qquad (7.4)
\end{aligned}$$

kde $\mathbf{b}_g$ a $\mathbf{c}_g$ jsou obecně nějaká řešení příslušné normální rovnice.

Když přepíšeme (7.3) tak, aby bylo patrné jakou lineární kombinací sloupců matic $\mathbf{X}, \mathbf{Z}$ je vektor $\hat{\mathbf{Y}}_g$ (co mohou být vektory $\mathbf{b}_g, \mathbf{c}_g$), dostaneme po úpravě (vyjádříme $\mathbf{M}$ pomocí $\mathbf{X}$)

$$\hat{\mathbf{Y}}_g = \mathbf{X}(\mathbf{b} - (\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{Z}\mathbf{c}_g) + \mathbf{Z}\mathbf{c}_g, \tag{7.5}$$

když jsme označili

$$\mathbf{c}_g = (\mathbf{Z}'\mathbf{M}\mathbf{Z})^-\mathbf{Z}'\mathbf{u}. \tag{7.6}$$

Můžeme tedy psát

$$\mathbf{b}_g = \mathbf{b} - (\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{Z}\mathbf{c}_g, \tag{7.7}$$

odkud je zřetelný zejména vztah mezi $\mathbf{b}$ a $\mathbf{b}_g$.

Z (7.3) plyne, že rozdíl reziduálních součtů čtverců mezi uvažovaným modelem $\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ a skutečně platným modelem $\mathbf{Y} \sim (\mathbf{G}\boldsymbol{\delta}, \sigma^2\mathbf{I})$ je

$$\begin{aligned} RSS - RSS_g &= \|\mathbf{M}\mathbf{Z}(\mathbf{Z}'\mathbf{M}\mathbf{Z})^-\mathbf{Z}'\mathbf{u}\|^2 \\ &= \|\mathbf{M}\mathbf{Z}\mathbf{c}_g\|^2. \end{aligned} \tag{7.8}$$

Porovnejme ještě střední hodnoty obou reziduálních součtů čtverců. Protože platí model (7.1), je zřejmě $\mathsf{E}\, RSS_g = (n - \mathsf{h}(\mathbf{X}, \mathbf{Z}))\sigma^2$. Jinak to dopadne u reziduálního součtu čtverců RSS z (nesprávně) předpokládaného modelu. Postupnými úpravami dostaneme

$$\mathsf{E}\, RSS = \mathsf{E}\,||\mathbf{M}\mathbf{Y}||^2 = \mathsf{E}\,||\mathbf{M}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{e})||^2 = \mathsf{E}\,||\mathbf{M}\mathbf{Z}\boldsymbol{\gamma} + \mathbf{M}\mathbf{e}||^2,$$

tedy (s ohledem na $\mathsf{E}\,\mathbf{e} = \mathbf{0}$)

$$\begin{aligned} \mathsf{E}\, RSS &= ||\mathbf{M}\mathbf{Z}\boldsymbol{\gamma}||^2 + \mathsf{E}\,||\mathbf{M}\mathbf{e}||^2 \\ &= ||\mathbf{M}\mathbf{Z}\boldsymbol{\gamma}||^2 + (n - \mathsf{h}(\mathbf{X}))\sigma^2. \end{aligned} \tag{7.9}$$

Vraťme se k odhadu $\hat{\mathbf{Y}}$. Jeho střední hodnota je rovna

$$\mathsf{E}\,\hat{\mathbf{Y}} = \mathbf{H}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{H}\mathbf{Z}\boldsymbol{\gamma}.$$

Obecně tedy není nestranným odhadem pro $\mathsf{E}\,\mathbf{Y}$, má vychýlení

$$\mathsf{bias}\,\hat{\mathbf{Y}} = \mathsf{E}\,\hat{\mathbf{Y}} - \mathsf{E}\,\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{H}\mathbf{Z}\boldsymbol{\gamma} - (\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}) \;=\; -\mathbf{M}\mathbf{Z}\boldsymbol{\gamma}. \tag{7.10}$$

Shrňme vlastnosti odhadů klasického modelu.

Věta 7.1. (Vychýlení odhadů, platí-li širší model) Nechť platí $\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}, \sigma^2\mathbf{I})$. Pro statistiky odvozené z modelu $\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ platí

$$\mathsf{bias}\,\hat{\mathbf{Y}} = -\mathbf{M}\mathbf{Z}\boldsymbol{\gamma}, \tag{7.11}$$

$$\mathsf{bias}\, S^2 = \frac{||\mathbf{M}\mathbf{Z}\boldsymbol{\gamma}||^2}{n - \mathsf{h}(\mathbf{X})}, . \tag{7.12}$$

Porovnejme nyní varianční matice odhadů $\hat{\mathbf{Y}}$ a $\hat{\mathbf{Y}}_g$. Snadno dostaneme

$$\begin{aligned}\operatorname{var}\hat{\mathbf{Y}}_g &= \sigma^2(\mathbf{X},\mathbf{MZ})\begin{pmatrix}\mathbf{X}'\mathbf{X} & \mathbf{O}\\ \mathbf{O} & \mathbf{Z}'\mathbf{MZ}\end{pmatrix}^-\begin{pmatrix}\mathbf{X}'\\ \mathbf{Z}'\mathbf{M}\end{pmatrix}\\ &= \sigma^2\left(\mathbf{H}+\mathbf{MZ}(\mathbf{Z}'\mathbf{MZ})^-\mathbf{Z}'\mathbf{M}\right).\end{aligned} \tag{7.13}$$

Je-li matice $\mathbf{MZ}$ nenulová, bude matice $\operatorname{var}\hat{\mathbf{Y}}_g - \operatorname{var}\hat{\mathbf{Y}}$ pozitivně definitní, takže vychýlený odhad je co do rozptylu lepší. Vychýlené odhady však neporovnáváme pomocí jejich rozptylu či varianční matice, ale pomocí střední čtvercové chyby. *Střední čtvercová chyba* odhadu $\mathbf{T}$ parametru $\boldsymbol{\theta}$ je definována jako

$$\begin{aligned}\operatorname{MSE}(\mathbf{T}) &= \mathsf{E}\,(\mathbf{T}-\boldsymbol{\theta})(\mathbf{T}-\boldsymbol{\theta})'\\ &= \operatorname{var}(\mathbf{T}) + \operatorname{bias}(\mathbf{T})\operatorname{bias}(\mathbf{T})'.\end{aligned}$$

Střední čtvercovou chybu $\hat{\mathbf{Y}}$ jako odhadu pro $\mathsf{E}\,\mathbf{Y}$ lze tedy psát

$$\operatorname{MSE}\hat{\mathbf{Y}} = \operatorname{var}\hat{\mathbf{Y}} + (\operatorname{bias}\hat{\mathbf{Y}})(\operatorname{bias}\hat{\mathbf{Y}})' = \sigma^2\mathbf{H} + \mathbf{MZ}\boldsymbol{\gamma}\boldsymbol{\gamma}'\mathbf{Z}'\mathbf{M}. \tag{7.14}$$

Protože $\hat{\mathbf{Y}}_g$ je nestranným odhadem $\mathsf{E}\,\mathbf{Y}$, platí $\operatorname{MSE}\hat{\mathbf{Y}}_g = \operatorname{var}\hat{\mathbf{Y}}_g$.

Porovnejme střední čtvercové chyby $\hat{\mathbf{Y}}_g$ a $\hat{\mathbf{Y}}$ jako odhadů vektoru $\mathsf{E}\,\mathbf{Y}$:

$$\operatorname{MSE}\hat{\mathbf{Y}}_g - \operatorname{MSE}\hat{\mathbf{Y}} = \sigma^2\left(\mathbf{MZ}(\mathbf{Z}'\mathbf{MZ})^-\mathbf{Z}'\mathbf{M} - \mathbf{MZ}\boldsymbol{\gamma}\boldsymbol{\gamma}'\mathbf{Z}'\mathbf{M}/\sigma^2\right).$$

Nyní stačí použít tvrzení věty A.9 pro $\mathbf{A} = \mathbf{MZ}$ a $\mathbf{c} = \boldsymbol{\gamma}/\sigma$, abychom zjistili, že rozdíl středních čtvercových chyb dá pozitivně semidefinitní matici, právě když je $\|\mathbf{Ac}\|^2 = \|\mathbf{MZ}\boldsymbol{\gamma}/\sigma\|^2 \le 1$. Došli jsme tak k tvrzení následující věty.

Věta 7.2. (Kdy je vychýlení malé) Nechť platí $\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}, \sigma^2\mathbf{I})$. Pro $\hat{\mathbf{Y}}_g$ z tohoto modelu a pro $\hat{\mathbf{Y}}$ z modelu $\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ platí ekvivalence

$$\operatorname{MSE}\hat{\mathbf{Y}}_g \ge \operatorname{MSE}\hat{\mathbf{Y}} \iff \|\operatorname{bias}\hat{\mathbf{Y}}\|^2 \le \sigma^2. \tag{7.15}$$

Při předpovědi budoucího pozorování tedy je výhodnější použít menší model, když je vychýlení způsobené touto volbou dostatečně malé.

Věta 7.3. (Důsledek) Nechť platí $\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}, \sigma^2\mathbf{I})$, nechť $\theta = \mathbf{p}'\boldsymbol{\beta} + \mathbf{s}'\boldsymbol{\gamma}$ je odhadnutelný parametr v tomto modelu. Nechť $\mathbf{b}$ je libovolné řešení normální rovnice $\mathbf{X}'\mathbf{Xb} = \mathbf{X}'\mathbf{Y}$. Potom je parametr $\tau = \mathbf{p}'\boldsymbol{\beta}$ odhadnutelný také v modelu $\mathbf{Y} \sim (\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ a platí

$$\operatorname{MSE}\hat{\theta} \ge \operatorname{MSE}\hat{\tau} \iff \|\mathbf{MZ}\boldsymbol{\gamma}\|^2 \le \sigma^2.$$

D ů k a z: Především je třeba dokázat, že τ je odhadnutelný parametr. Odhadnutelnost θ je podle věty 2.4 ekvivalentní s existencí vektoru $\mathbf{q} \in \mathbb{R}^n$, pro který platí $\mathbf{q}'(\mathbf{X},\mathbf{Z}) = (\mathbf{p}',\mathbf{s}')$. Speciálně to tedy znamená existenci $\mathbf{q}$, pro který platí $\mathbf{q}'\mathbf{X} = \mathbf{p}'$,