9 Zobecnění klasického lineárního modelu

Průvodce studiem

V této kapitole jsou ukázány některé postupy, které rozšiřují oblast aplikace lineárního modelu, zejména za okolností, kdy předpoklady klasického modelu nejsou splněny. Prostudování kapitoly a pochopení souvislostí vyžaduje nejméně tři až čtyři hodiny.

9.1 Transformace původních regresorů

Prozatím jsme se zabývali lineárním modelem, který obsahoval přímo hodnoty regresorů $x_{\cdot,1}, x_{\cdot,2}, \ldots, x_{\cdot,k}$,

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i \tag{19}$$

Jedním z mnoha možných zobecnění je model ve tvaru

$$y_i = \beta_0 + \beta_1 Z_{i,1} + \cdots + \beta_{p-1} Z_{i,p-1} + \varepsilon_i, \tag{20}$$

kde každé $Z_{\cdot,j}$ je nějakou funkcí původních regresorů $x_{\cdot,1}, x_{\cdot,2}, \ldots, x_{\cdot,k}$.

Jako příklady takových modelů můžeme uvést (pro přehlednost zápisu jsou řádkové indexy vynechány):

1. $k = 1$, polynom stupně $p - 1$

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_{p-1} x^{p-1} + \varepsilon$$

2. $k = 2$, $p = 6$, tzv. model 2. řádu

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \varepsilon$$

3. $k = 2$, $p = 10$, tzv. model 3. řádu

$$\begin{aligned} y = \quad & \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \\ & + \beta_{11} x_1^2 + \beta_{12} x_1 x_2 + \beta_{22} x_2^2 + \\ & + \beta_{111} x_1^3 + \beta_{112} x_1^2 x_2 + \beta_{122} x_1 x_2^2 + \beta_{222} x_2^3 + \varepsilon \end{aligned}$$

Další použitelné transformace jsou

- reciproká transformace, tj. $z_j = 1/x_j$, když $\forall x_j > 0$
- logaritmická transformace, tj. $z_j = \ln(x_j)$, když $\forall x_j > 0$
- odmocninová transformace, tj. $z_j = \sqrt{x_j}$, když $\forall x_j \geq 0$

a mnoho podobných dalších transformací a kombinace jejich užití v jednom modelu. Důležité však je, že modely (20) jsou lineární v parametrech, tj. soustava normálních rovnic je soustava p lineárních rovnic pro p odhadovaných parametrů. Splňuje-li model (20) předpoklady klasického lineárního modelu, můžeme pro analýzu a interpretaci takového modelu užít všechny techniky, které jsme dosud užívali pro klasický lineární model ve tvaru (19), tedy pro situaci, kdy jsme pracovali přímo s regresory, nikoliv s jejich funkcemi.

Na tvar lineárního modelu je možno někdy převést i modely, které na první pohled lineární nejsou, např. když závislost vysvětlované veličiny na regresorech x_1, x_2, x_3 může být proložena funkcí

$$\eta = \alpha x_1^{\beta} x_2^{\gamma} x_3^{\delta} \tag{21}$$

Po zlogaritmování dostaneme

$$\ln \eta = \ln \alpha + \beta \ln x_1 + \gamma \ln x_2 + \delta \ln x_3 \tag{22}$$

Pokud hodnoty vysvětlované náhodné veličiny y lze popsat modelem

$$\ln y = \ln \eta + \varepsilon \tag{23}$$

a platí, že $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$, pak k odhadům parametrů $\alpha, \beta, \gamma, \delta$ a jejich interpretaci opět můžeme užít postupů známých z klasického lineárního modelu. Při linearizaci vztahů typu (21) však musíme být opatrní v tom, jakou roli má náhodné kolísání ε (tzv. chybový člen, error). Představa (23) znamená, že náhodné kolísání je *multiplikativní*, nikoliv aditivní, tj. hodnota náhodné veličiny y je vyjádřena modelem

$$y = \eta \exp(\varepsilon) = \alpha x_1^{\beta} x_2^{\gamma} x_3^{\delta} \exp(\varepsilon),$$

nikoliv modelem s aditivní chybou ve tvaru $y = \eta + \text{error}$. To, zda je oprávněné užít multiplikativní model, může vyplynout z věcné analýzy úlohy, ale často i v situacích, kdy model chyb je jiný, může být výsledek být výsledek získaný linearizací a aplikací lineárního modelu užitečným prvním přiblížením k řešení problému.

9.2 Aitkenův odhad

Tento odhad 5eší problém, kdy není splněn předpoklad (2) klasického model, že náhodné složky mají konstantní rozptyl a jsou nekorelované. Připust'me, že náhodné složky mohou být korelované a nemusí mít konstantní rozptyl:

$$\operatorname{cov}(\boldsymbol{\varepsilon}) = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T) = \sigma^2 \boldsymbol{\Omega}, \quad \sigma^2 > 0, \tag{24}$$

kde $\mathbf{\Omega}$ je pozitivně definitní matice. Pak existuje regulární matice $\mathbf{P}$, pro kterou platí

$$\mathbf{P\Omega P}^T = \mathbf{I} \quad \text{a} \quad \mathbf{P}^T\mathbf{P} = \mathbf{\Omega}^{-1} \tag{25}$$

Vynásobíme-li rov.(**??**), tj. klasický lineární model, maticí $\mathbf{P}$ zleva (transformujeme veličiny), dostaneme

$$\mathbf{Py} = \mathbf{PX}\boldsymbol{\beta} + \mathbf{P}\boldsymbol{\varepsilon} \tag{26}$$

Označíme-li $\mathbf{y}^* = \mathbf{Py}$, $\mathbf{X}^* = \mathbf{PX}$ a $\boldsymbol{\varepsilon}^* = \mathbf{P}\boldsymbol{\varepsilon}$, pak rov.(26) můžeme přepsat

$$\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}^* \tag{27}$$

Kovarianční matice náhodných složek v rov.(27) je pak

$$\mathrm{cov}(\boldsymbol{\varepsilon}^*) = E(\boldsymbol{\varepsilon}^*\boldsymbol{\varepsilon}^{*T}) = E(\mathbf{P}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\mathbf{P}^T) = \sigma^2\mathbf{P\Omega P^T} = \sigma^2\mathbf{I}$$

tzn., že pro hvězdičkované veličiny je rov.(27) klasický lineární model. Vyjádříme rovnice pro odhady parametrů v modelu (27) pomocí původních netransformovaných veličin a dostaneme vztahy pro odhady parametrů modelu

$$\mathbf{b} = (\mathbf{X^T\Omega^{-1}X})^{-1}\mathbf{X^T\Omega^{-1}y}, \tag{28}$$

o kterých víme, že to jsou BLU-odhady s kovarianční maticí

$$\mathrm{cov}(\mathbf{b}) = \sigma^2(\mathbf{X^T\Omega^{-1}X})^{-1} \tag{29}$$

Nestranný odhad parametru σ^2 je

$$s^2 = (\mathbf{y} - \mathbf{Xb})^T\mathbf{\Omega}^{-1}(\mathbf{y} - \mathbf{Xb}) \tag{30}$$

který pak můžeme užít k odhadu kovarianční matice a tedy i rozptylů odhadů b_i.

Odhady získané tímto postupem jsou nestranné, některé jsou dokonce BLU-odhady, avšak k jejich výpočtu potřebujeme znát matici $\mathbf{\Omega}$. Tu bohužel v analýze dat v naprosté většině případů neznáme. Nemůžeme tuto matici z dat ani konsistentně odhadnout, v datech máme n nezávislých pozorování a potřebujeme odhadnout $(n^2 + n)/2$ jejích prvků (diagonálu a polovinu nediagonálních prvků matice, matice $\mathbf{\Omega}$ je symetrická). Většinou nezbývá, než na místo předpokladu (24) přijmout nějaké větší omezení.

9.3 Heteroskedascita

Jedna z možností je řešit tzv. heteroskedastickou regresi, tj. připustit, že rozptyly náhodné složky nejsou konstantní, ale náhodné složky v lineárním regresním modelu (3) jsou nekorelované:

$$\mathrm{cov}(\boldsymbol{\varepsilon}) = \sigma^2 \mathrm{diag}(w_1^2, w_2^2, \cdots, w_n^2) \tag{31}$$

kde $w_i^2 > 0$ je váha rozptylu i-tého pozorování. Pak matice $\boldsymbol{\Omega}$ je také diagonální,
$\boldsymbol{\Omega} = \mathrm{diag}(w_1^2, w_2^2, \cdots, w_n^2)$,
inverzní matice je $\boldsymbol{\Omega}^{-1} = \mathrm{diag}(w_1^{-2}, w_2^{-2}, \cdots, w_n^{-2})$
a $\mathbf{P} = \mathrm{diag}(1/w_1, 1/w_2, \cdots, 1/w_n)$.

Pak v modelu (3), tj. v datové matici vydělíme řádek vahou rovnou směrodatné odchylce pozorování

$$y_i/w_i = \beta_0/w_i + \beta_1 x_{i1}/w_i + \cdots + \beta_k x_{ik}/w_i + \varepsilon_i/w_i,$$

můžeme užít OLS-odhady, které budou mít dobré vlastnosti jako v klasickém modelu.

Otázkou je, jak určit váhu pozorování, w_i. Máme několik možností, záleží na řešené úloze:

(1) V modelu máme je jeden regresor x_{i1} a předpokládáme, že pozorování závislé veličiny y_i mají konstantní *relativní* chybu. Pak můžeme položit $w_i = x_{i1}$.

(2) Pozorování závislé veličiny y_i mají konstantní *relativní* chybu a v modelu je více regresorů. Pak nezbývá, než vybrat jeden podle subjektivního rozhodnutí, možná ten, který nejvíce koreluje se závisle proměnnou y.

(3) Nejdříve spočítat $\hat{y}_i$ jako OLS-odhad podle modelu (3) a pak ve druhém kroku položit $w_i = \hat{y}_i$.

(4) Postupovat jako ve variantě (9.3) a dále pokračovat v iteracích, dokud dva po sobě následující odhady nejsou dostatečně blízké. Tomuto postupu se říká metoda *iterovaných vážených čtverců*, iterated WLS (Weighted Least Squares).

9.4 Stochastické regresory

Také předpoklad v klasickém modelu, že matice $\mathbf{X}$ obsahuje pevné hodnoty, u kterých není třeba uvažovat s jejich rozptylem a korelací, je v mnoha aplikacích nerealistický. Pro tzv. *nezávislou stochastickou regresi*, kdy předpokládáme, že matice $\mathbf{X}$ je stochastická, tvoří $(k+1)$ rozměrný náhodný proces a náhodná složka $\boldsymbol{\varepsilon}$ nezávisí na $\mathbf{X}$, uvedeme stručně důležité výsledky, podrobněji viz např. [7].

OLS-odhady $\mathbf{b}$, spočítané podle rov.(10), s^2 podle rov.(13) a

$$\mathbf{S}_{bb} = s^2(\mathbf{X^T X})^{-1}$$

jsou nestranné. Ale nejsou to lineární odhady, neboť jsou stochastickou funkcí náhodného vektoru $\mathbf{y}$ a nejsou to BLU-odhady. Za předpokladu, že v pravděpodobnosti konverguje $\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}/n \to \sigma^2$ a $\mathbf{X^T X}/n \to \boldsymbol{\Sigma}_{XX}$ (kde $\boldsymbol{\Sigma}_{XX}$ je kovarianční matice regresorů) však platí:

- $\mathbf{b} = (\mathbf{X^T X})^{-1}\mathbf{X^T y}$ je konzistentním odhadem vektoru regresních koeficientů $\boldsymbol{\beta}$
- s^2 podle rov.(13) je konzistentním odhadem parametru σ^2
- $\mathbf{S}_{bb} = s^2(\mathbf{X^T X})^{-1}$ lze vzít za konsistentní odhad asymptotické kovarianční matice odhadovaných parametrů $\mathbf{b}$.

To znamená, že OLS-odhady lze užít k běžným testům a určení intervalů spolehlivosti pro parametry.

Pokud jsou náhodné složky modelu normálně rozděleny, jsou OLS-odhady podle rov.(10) také ML-odhady, takže mají dobré asymptotické vlastnosti, jsou konsistentní a asymptoticky eficientní.

9.5 Diskrétní regresory, umělé proměnné

Dosud jsme se zabývali úlohami, ve kterých vysvětlující veličiny byly spojité. Docela často se v analýze dat stává, že data pocházejí ze dvou nebo více populací, vzpomeňme např. na dvouvýběrové testy či analýzu rozptylu. I na taková data můžeme aplikovat lineární regresi. Uvažujme nyní nejjednodušší případ – lineární regresní model s jedním regresorem

$$EY_i = \beta_0 + \beta_1 x_i.$$

Parametr β_1 je směrnice přímky, tzn. vyjadřuje změnu střední hodnoty náhodné veličiny Y_i, změní-li se hodnota regresoru o jedničku. Uvažujme, že regresor x je diskrétní a má hodnoty $\{0, 1\}$, jinými slovy jen rozděluje data do dvou skupin (výběrů) ze dvou populací 0 a 1. Pak test hypotézy $\beta_1 = 0$ znamená totéž jako test hypotézy $\mu_0 = \mu_1$ (shoda středních hodnot obou populací), tj. dvouvýběrový t-test při shodných rozptylech.

Příklad 9.1 Máme otestovat hypotézu, že střední hodnoty veličiny Y ze dvou populací jsou shodné. Výběrové charakteristiky pro oba nezávislé výběry jsou v následující tabulce a obrázku.

výběr	n	průměr	sm. odchylka
0	30	5,06	1,04
1	20	5,96	2,09

K testu si můžeme vybrat několik metod, které nám dají shodné výsledky:

metoda	H_0	předpoklad	statistika	p
t-test(shodné rozptyly)	$\mu_0 = \mu_1$	$\sigma_0^2 = \sigma_1^2$	2,01	0,05
lineární regrese	$\beta_1 = 0$	$\sigma_0^2 = \sigma_1^2$	2,01	0,05
ANOVA	$\mu_0 = \mu_1$	$\sigma_0^2 = \sigma_1^2$	4,03	0,05

Jak vidíme, ve všech třech případech nám vyšla stejná hodnota p, pro t-test a lineární regresi i stejná hodnota statistiky, ačkoliv testujeme různé hypotézy, v analýze rozptylu je hodnota F-statistiky rovna druhé mocnině t-statistiky u ostatních dvou metod. V případě lineární regrese je odhad $b_1 = \bar{y}_1 - \bar{y}_0$ a $b_0 = \bar{y}_0$ a testujeme, zda rozdíl průměrů je dostatečně veliký k zamítnutí hypotézy $\mu_0 = \mu_1$ (připomeňme, že směrnice přímky je změna veličiny y při změně veličiny x o jedničku).

To, že uvedené statistiky vyšly stejně, není žádné překvapení, neboť se vyčíslují ze stejných formulí. Také předpoklady pro všechny uvedené testy jsou shodné, normálně rozdělená residua a shodné rozptyly v obou populacích.

Uvedený příklad ilustruje možnost podobného pohledu na analýzu rozptylu a lineární regresi, ukazuje, že diskrétní regresory mohou být docela snadno interpretovány a naznačuje směry dalšího zobecnění lineárního modelu.

Pokud diskrétní regresor nabývá více než dvou hodnot, lze k rozlišení užít tzv. *umělé proměnné, dummy variables*. Obvykle se jedna z r kategorií vybere jako referenční a $r-1$ dummy proměnných s hodnotami $\{0, 1\}$ pak kóduje kategorie. Odhad směrnice u konkrétní dummy proměnné znamená odhad změny střední hodnoty vysvětlované veličiny oproti referenční kategorii. Podrobněji viz kapitola Logistická regrese.

Při více diskrétních regresorech pomocné proměnné dovolují zkoumat regresní analýzou i velmi komplikované struktury závislosti, případně i v kombinaci s dalšími spojitými regresory tyto závislosti „očišťovat“ od vlivu jiných veličin. Podrobnější výklad takových postupů přesahuje rozsah tohoto kursu, v případě potřeby se obraťte na literaturu, např. Draper a Smith nebo Anděl atd. Tam najdete i další možnosti zavedení pomocných proměnných.

Shrnutí

- *tranformace regresorů, polynom, model druhého řádu, linearizace*
- *heteroskedascita, metoda vážených nejmenších čtverců*
- *diskrétní regresory, umělé proměnné*

Kontrolní otázky

1. *Jaké zjednodušení představuje metoda vážených nejmenších čtverců proti Aitkenovu odhadu?*
2. *Jak interpretovat směrnici regresní přímky v případě spojitých regresorů a jak v případě diskrétních regresorů?*
3. *Co jsou umělé (dummy) proměnné?*