Bodové a intervalové odhady parametrů v regresním modelu

Ekonometrie

Jiří Neubauer

Katedra kvantitativních metod FVL UO Brno
kancelář 69a, tel. 973 442029
email:Jiri.Neubauer@unob.cz

Lineární regresní model

Mějme lineární regresní model (LRM)

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

kde

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}.$$

Odhady neznámých parametrů metodou nejmenších čtverců jsou dány

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y},$$

reziduální součet čtverců je

$$S_e = (\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = \mathbf{Y}'\mathbf{Y} - \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y}.$$

Lineární regresní model – odhady

Odhady parametrů $\widehat{\beta}$ jsou nevychýlené,

$$E(\hat{\beta}) = E\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\right] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{Y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta},$$

reziduální součet čtverců normovaný konstantou $n-k$ nevychýleným odhadem rozptylu σ^2

$$\widehat{\sigma^2} = s_e^2 = \frac{1}{n-k}S_e = \frac{1}{n-k}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2.$$

Za předpokladů normality lze provádět testy hypotéz o parametrech uvažovaného modelu. Dále lze na základě uvedených výsledků konstruovat intervaly spolehlivosti pro neznámé parametry a také konstruovat intervaly spolehlivosti pro predikované hodnoty odezvy Y při daných hodnotách regresorů.

Lineární regresní model – odhady

Předpokládejme nyní, že náhodné chyby $\epsilon_i, i = 1, 2, \ldots, n$ v lineárním regresním modelu mají normální rozdělení s nulovou střední hodnotou a rozptylem σ^2. Potom mají odhady $\widehat{\beta}_j$, $j = 1, 2, \ldots, k$ regresní koeficientů β_j normální rozdělení, tedy platí $\widehat{\beta}_j \sim N(\beta_j, D(\widehat{\beta}_j))$, kde rozptyly $D(\widehat{\beta}_j)$ jsou dány:

$$D(\widehat{\beta}_1) = \sigma^2 v_{11}, D(\widehat{\beta}_2) = \sigma^2 v_{22}, \ldots, D(\widehat{\beta}_k) = \sigma^2 v_{kk},$$

přičemž $v_{11}, v_{22}, \ldots, v_{kk}$ jsou prvky na hlavní diagonále matice $(\mathbf{X}'\mathbf{X})^{-1}$. Rozptyly odhadů regresních parametrů odhadneme $\widehat{D}(\widehat{\beta}_j) = s_e^2 v_{jj}$, druhé odmocniny těchto odhadů

$$s(\widehat{\beta}_j) = \sqrt{s_e^2 v_{jj}}$$

se nazývají **směrodatné chyby** odhadů regresních parametrů.

Lineární regresní model – odhady

Východiskem pro konstrukci intervalů spolehlivosti pro parametry β_j regresního modelu jsou statistiky

$$t = \frac{\widehat{\beta}_j - \beta_j}{s(\widehat{\beta}_j)},$$

které mají Studentovo rozdělení s $n - k$ stupni volnosti. Oboustranný interval spolehlivosti při riziku odhadu α má potom tvar

$$\widehat{\beta}_j - t_{1-\alpha/2}(n-k) \cdot s(\widehat{\beta}_j) < \beta_j < \widehat{\beta}_j + t_{1-\alpha/2}(n-k) \cdot s(\widehat{\beta}_j),$$

kde $t_{1-\alpha/2}(n-k)$ označuje kvantil Studentova rozdělení.

Lineární regresní model – odhady, regresní přímka

Příklad. Následující tabulka udává informaci o teplotě (ve stupních Celsia) v jednom městě a množství zmrzliny (v kilogramech) prodaných v osmi náhodně vybraných cukrárnách.

teplota	34	30	25	32	37	39	31	26
zmrzlina	94	79	56	90	105	126	72	53

Odhad regresní přímky je

$$\widehat{y} = -71{,}789 + 4{,}918x,$$

$s(\widehat{\beta}_1) = 14{,}4079$, $s(\widehat{\beta}_2) = 0{,}4492$, pro $\alpha = 0,05$ je
$t_{1-\alpha/2}(n-k) = t_{0,975}(8-2) = 2{,}44691$, potom 95% intervaly spolehlivosti odhady pro parametry regresní přímky jsou

$$-107{,}02355 < \beta_1 < -36{,}51376,$$

$$3{,}81888 < \beta_2 < 6{,}01695.$$

Lineární regresní model – odhady, regresní parabola

Příklad. U automobilu Trabant se měřila spotřeba paliva v litrech na 100 km (Y) v závislosti na jeho rychlosti (X).

Rychlost	40	50	60	70	80	90	100
Spotřeba	6,1	5,8	6,0	6,5	6,8	8,1	10,0

Odhadnutá parabolická regresní funkce má tvar

$$\widehat{y} = 11{,}392857 - 0{,}207262x + 0{,}001917x^2.$$

$s(\widehat{\beta}_1) = 1{,}1630215$, $s(\widehat{\beta}_2) = 0{,}0351065$, $s(\widehat{\beta}_3) = 0{,}0002489$ pro $\alpha = 0{,}05$ je $t_{1-\alpha/2}(n-k) = t_{0{,}975}(7-3) = 2{,}776445$, potom 95% intervaly spolehlivosti odhady pro parametry parabolické regresní funkce jsou

$$8{,}163792 < \beta_1 < 14{,}6219225,$$

$$-0{,}304733 < \beta_2 < -0{,}1097905,$$

$$0{,}001226 < \beta_3 < 0{,}0026076.$$

Lineární regresní model – odhady, dva lineární regresory

Příklad. Výrobce nealkoholických nápojů má zájem analyzovat potřebný čas k servisu (doplnění lahví případně malý servis zařízení) automatů na výdej lahví s těmito nápoji. Celkovou dobu doplnění lahví je třeba predikovat pomocí dvou dostupných proměnných: počet lahví, které je třeba doplnit do automatu, a vzdálenost, kterou musí údržbář ujít. Vysvětlovanou proměnnou je v tomto případě celkový čas, vysvětlující proměnné jsou počet doplněných lahví a vzdálenost.

čas	16,68	11,5	12,03	14,88	13,75	18,11	8	17,83	79,24	21,5
počet lahví	7	3	3	4	6	7	2	7	30	5
vzdálenost	560	220	340	80	150	330	110	210	1460	605
čas	40,33	21	13,5	19,75	24	29	15,35	19	9,5	35,1
počet lahví	16	10	4	6	9	10	6	7	3	17
vzdálenost	688	215	255	462	448	776	200	132	36	770
čas	17,9	52,32	18,75	19,83	10,75					
počet lahví	10	26	9	8	4					
vzdálenost	140	810	450	635	150					

Lineární regresní model – odhady, dva lineární regresory

Metodou nejmenších čtverců získáme odhad regresní funkce

$$\widehat{y} = 2,34123 + 1,61591x + 0,01438z.$$

$s(\widehat{\beta}_1) = 1,096730$, $s(\widehat{\beta}_2) = 0,170735$, $s(\widehat{\beta}_3) = 0,003613$ pro $\alpha = 0,05$ je $t_{1-\alpha/2}(n-k) = t_{0,975}(25-3) = 2,073873$, potom 95% intervaly spolehlivosti odhady pro parametry parabolické regresní funkce jsou

$$0,066752 < \beta_1 < 4,615710,$$

$$1,261825 < \beta_2 < 1,969990,$$

$$0,006892 < \beta_3 < 0,021878.$$

Predikce

Nalezené odhady $\widehat{\beta}_1, \ldots, \widehat{\beta}_k$ parametrů $\beta_1, \ldots, \beta_k$ regresního modelu lze použít k odhadu regresní funkce y v daném bodě $\mathbf{x}_0 = (x_{01}, x_{02}, \ldots, x_{0k})'$, tedy při hodnotách regresorů $X_1 = x_{01}, X_2 = x_{02}, \ldots, X_k = x_{0k}$. Odhad regresní funkce $y = y(\mathbf{x})$ v bodě $\mathbf{x} = \mathbf{x}_0$ pak získáme ze vztahu

$$\widehat{y} = \widehat{y}(\mathbf{x}_0) = \mathbf{x}_0' \widehat{\boldsymbol{\beta}} = \widehat{\beta}_1 x_{01} + \widehat{\beta}_2 x_{02} + \cdots + \widehat{\beta}_k x_{0k}.$$

Nalezení odhadu $\widehat{y}(\mathbf{x}_0)$ regresní funkce $y(\mathbf{x}_0)$ v bodě $\mathbf{x}_0$ je jednou z nejčastějších úloh regresní analýzy. Odpovídá nalezení střední („průměrné“) hodnoty vysvětlované proměnné Y při daných hodnotách regresorů $X_1 = x_{01}, X_2 = x_{02}, \ldots, X_k = x_{0k}$.

Predikce

Pro konstrukci intervalu spolehlivosti pro regresní funkci se použije statistika

$$t = \frac{\widehat{y}(\mathbf{x}_0) - y(\mathbf{x}_0)}{s(\widehat{y}(\mathbf{x}_0))},$$

kde $s(\widehat{y}(\mathbf{x}_0)) = s_e\sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}$ je směrodatná chyba (odchylka) bodového odhadu $\widehat{y}(\mathbf{x}_0)$. Statistika t má Studentovo rozdělení s $n-k$ stupni volnosti, . Odtud lze odvodit vztah pro oboustranný intervalový odhad $y(\mathbf{x}_0)$ regresní funkce $y(\mathbf{x})$ v bodě $\mathbf{x}_0$

$$\widehat{y}(\mathbf{x}_0) - t_{1-\alpha/2}(n-k)\cdot s(\widehat{y}(\mathbf{x}_0)) < y(\mathbf{x}_0) < \widehat{y}(\mathbf{x}_0) + t_{1-\alpha/2}(n-k)\cdot s(\widehat{y}(\mathbf{x}_0)).$$

Predikce

Zajímá-li nás interval spolehlivosti pro **predikci (předpověď)** vysvětlované veličiny Y v bodě $\mathbf{x}_0 = (x_{01}, x_{02}, \ldots, x_{0k})$, tedy interval spolehlivosti pro pozorování $Y_0 = y(\mathbf{x}_0) + \epsilon_0$, kde ϵ_0 je náhodná chyba tohoto pozorování v bodě $\mathbf{x}_0$, dostaneme s využitím uvedeného modelu

$$\widehat{y}(\mathbf{x}_0) - t_{1-\alpha/2}(n-k) \cdot s_0 < Y_0 < \widehat{y}(\mathbf{x}_0) + t_{1-\alpha/2}(n-k) \cdot s_0,$$

kde s_0 je směrodatná chyba odhadu Y_0, tedy směrodatná chyba veličiny $\widehat{y}(\mathbf{x}_0) + \epsilon_0$, která je rovna $s_0 = s_e \sqrt{1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}$.

Predikce – přímková regrese

Určíme množství prodané zmrzliny pro teplotu 33 °, které lze očekávat na základě spočítané přímkové regresní funkce

$$\widehat{y} = -71{,}789 + 4{,}918x.$$

Bodový odhad je $\widehat{y}(30) = -71{,}789 + 4{,}918 \cdot 33 = 90{,}522$. Označme

$$\mathbf{x_0} = \begin{pmatrix} 1 \\ 33 \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1 & 34 \\ 1 & 30 \\ 1 & 25 \\ 1 & 32 \\ 1 & 37 \\ 1 & 39 \\ 1 & 31 \\ 1 & 26 \end{pmatrix}.$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 6{,}1432836 & -0{,}189552239 \\ -0{,}1895522 & 0{,}005970149 \end{pmatrix}, s = 5{,}813007.$$

Predikce – přímková regrese

Směrodatná chyba bodového odhadu regresní funkce je

$$s(\widehat{y}(\mathbf{x}_0)) = s_e\sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} = 2,130515$$

Intervalový odhad je

$$85{,}30920 < y(\mathbf{x}_0) < 95{,}73557,$$

$t_{0,975}(6) = 2{,}446912$. Směrodatná chyba pro jedno pozorování Y_0 je

$$s_0 = s_e\sqrt{1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} = 6{,}191134$$

Intervalový odhad je

$$75{,}37323 < Y_0 < 105{,}67155.$$