

Bodové a intervalové odhady parametrů v regresním modelu

1 Odhady parametrů

1.1 Bodové odhady

Mějme lineární regresní model (LRM)

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

kde

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}.$$

Odhady neznámých parametrů metodou nejmenších čtverců jsou dány

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

reziduální součet čtverců je

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

Odhady parametrů $\hat{\boldsymbol{\beta}}$ jsou nevychýlené,

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

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

$$\widehat{\sigma^2} = s^2 = \frac{1}{n-k} S_e^2 = \frac{1}{n-k} \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

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.

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

$$D(\hat{\beta}_1) = \sigma^2 h_{11}, D(\hat{\beta}_2) = \sigma^2 h_{22}, \dots, D(\hat{\beta}_k) = \sigma^2 h_{kk},$$

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

$$s(\hat{\beta}_j) = \sqrt{s^2 h_{jj}}$$

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

1.2 Intervalové odhady

Východiskem pro konstrukci intervalů spolehlivosti pro parametry β_j lineárního regresního modelu jsou statistiky $T = (\hat{\beta}_j - \beta_j)/s(\hat{\beta}_j)$, které mají Studentovo rozdělení s $n - k$ stupni volnosti. Oboustranný interval spolehlivosti při riziku odhadu α má potom tvar

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

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

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

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

$s(\hat{\beta}_1) = 14,4079$, $s(\hat{\beta}_2) = 0,4492$, 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

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

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

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

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

$s(\hat{\beta}_1) = 1,1630215$, $s(\hat{\beta}_2) = 0,0351065$, $s(\hat{\beta}_3) = 0,0002489$ 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

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

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

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

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.

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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					

Metodou nejmenších čtverců získáme odhad regresní funkce

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

$s(\hat{\beta}_1) = 1,096730$, $s(\hat{\beta}_2) = 0,170735$, $s(\hat{\beta}_3) = 0,003613$ 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

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

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

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

2 Predikce

Hlavní využití regresního modelu je odhad hodnoty vysvětlované proměnné (resp. její průměrné hodnoty) pro danou hodnotu vysvětlující proměnné, příp. proměnných. Označíme $\mathbf{x}_0 = (x_{01}, x_{02}, \dots, x_{0k})$ vektor hodnot vysvětlujících proměnných, pro něž chceme odhadnout průměrnou (střední) hodnotu vysvětlované proměnné Y (jedná se o tzv. podmíněnou střední hodnotu $E(Y|\mathbf{x}_0)$). Budeme ji dále značit $\mu_{Y|\mathbf{x}_0}$. Tedy $\mu_{Y|\mathbf{x}_0} = E(Y|\mathbf{x}_0)$. Nestranným odhadem $\mu_{Y|\mathbf{x}_0}$ je

$$\hat{\mu}_{Y|\mathbf{x}_0} = \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02} + \dots + \hat{\beta}_k x_{0k}.$$

Pro konstrukci intervalu spolehlivosti pro regresní funkci se použije statistika

$$t = \frac{\hat{\mu}_{Y|\mathbf{x}_0} - \mu_{Y|\mathbf{x}_0}}{s(\hat{\mu}_{Y|\mathbf{x}_0})},$$

která má Studentovo rozdělení s $n-k$ stupni volnosti, $s(\hat{\mu}_{Y|\mathbf{x}_0}) = s\sqrt{\mathbf{x}_0' \mathbf{H} \mathbf{x}_0}$ je směrodatná chyba (odchylka) bodového odhadu $\hat{\mu}_{Y|\mathbf{x}_0}$. Odtud lze odvodit vztah pro oboustranný intervalový odhad.

$$\hat{\mu}_{Y|\mathbf{x}_0} - t_{1-\alpha/2}(n-k) \cdot s(\hat{\mu}_{Y|\mathbf{x}_0}) < \mu_{Y|\mathbf{x}_0} < \hat{\mu}_{Y|\mathbf{x}_0} + t_{1-\alpha/2}(n-k) \cdot s(\hat{\mu}_{Y|\mathbf{x}_0}).$$

Zajímá-li nás interval spolehlivosti pro predikci veličiny Y v bodě $\mathbf{x}_0 = (x_{01}, x_{02}, \dots, x_{0k})$, tedy interval spolehlivosti pro pozorování $Y_0 = \mu_{Y|\mathbf{x}_0} + e_0$, kde e_0 je náhodná chyba tohoto pozorování v bodě $\mathbf{x}_0$, dostaneme s využitím uvedeného modelu

$$\hat{Y}_0 - t_{1-\alpha/2}(n-k) \cdot s(\hat{Y}_0) < Y_0 < \hat{Y}_0 + t_{1-\alpha/2}(n-k) \cdot s(\hat{Y}_0),$$

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

kde $\hat{Y}_0 = \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02} + \dots + \hat{\beta}_k x_{0k}$ a směrodatná chyba odhadu $\hat{Y}_0$ je rovna

$$s(\hat{Y}_0) = s \sqrt{1 + \mathbf{x}_0' \mathbf{H} \mathbf{x}_0}$$

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

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

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

$$\mathbf{x}_0 = \begin{pmatrix} 1 \\ 33 \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1 & 34 \\ 1 & 30 \\ 1 & 25 \\ 1 & 32 \\ 1 & 37 \\ 1 & 39 \\ 1 & 31 \\ 1 & 26 \end{pmatrix}.$$

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

Směrodatná chyba bodového odhadu regresní funkce je

$$s(\hat{\mu}_{Y|\mathbf{x}_0}) = s \sqrt{\mathbf{x}_0' \mathbf{H} \mathbf{x}_0} = 2,130515$$

Intervalový odhad je

$$85,30920 < \mu_{Y|\mathbf{x}_0} < 95,73557,$$

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

$$s(\hat{\mu}_{Y|\mathbf{x}_0}) = s \sqrt{1 + \mathbf{x}_0' \mathbf{H} \mathbf{x}_0} = 6,191134$$

Intervalový odhad je

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

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Příklady k procvičení

1. Byla zjištěna výška otců a výška jejich nejstarších synů [v cm].

otec	165	178	158	170	180	160	170	167	185	165	173	175
syn	162	184	163	170	189	165	177	170	187	176	171	183

- a) Určete bodové odhady parametrů regresní přímky, reziduální rozptyl, směrodatné chyby těchto odhadů a zkonstruujte 95% intervaly spolehlivosti pro parametry regresní přímky.
b) Odhadněte výšku syna při výšce otce 180 cm, určete interval 95% interval spolehlivosti pro regresní funkci i pro predikci v tomto bodě.

[Datový soubor: vyska.otec.syn.txt]

2. O 7 vybraných strojích v určitém podniku máme informace o jejich stáří (v letech) a týdenních nákladech na jejich údržbu (v Kč):

stáří stroje	1	1	3	3	5	6	7
náklady	35	52	81	105	100	125	120

- a) Určete bodové odhady parametrů regresní přímky, reziduální rozptyl, směrodatné chyby těchto odhadů a zkonstruujte 95% intervaly spolehlivosti pro parametry regresní přímky.
b) Určete bodové odhady regresní logaritmické křivky, reziduální rozptyl, směrodatné chyby těchto odhadů a zkonstruujte 95% intervaly spolehlivosti pro parametry regresní přímky.
c) Pro oba modely určete interval 95% interval spolehlivosti pro regresní funkci i pro predikci pro stáří stroje 4 roky.

[Datový soubor: stari.stroje.naklady.txt]

3. Zajímáme se o brzdnou dráhu 63 automobilů v závislosti na výchozí rychlosti. K dispozici je celkem $n = 63$ měření. Proměnná rychlost udává výchozí rychlost (míle/hod.) před začátkem brzdění, proměnná dráha pak udává odpovídající brzdnou dráhu uvedenou ve stopách.

Rychlost	4	5	5	5	5	7	7	8	8	8
Dráha	4	2	8	8	4	6	7	9	8	13
Rychlost	8	9	9	9	10	10	10	12	12	12
Dráha	11	5	13	5	8	17	14	11	21	19
Rychlost	13	13	13	14	14	15	16	16	16	17
Dráha	18	27	15	14	16	16	19	14	34	29
Rychlost	17	18	18	18	19	20	21	21	21	22
Dráha	22	47	29	34	30	48	55	39	42	35
Rychlost	24	25	25	25	25	26	26	27	27	28
Dráha	56	33	59	48	56	39	41	78	57	64
Rychlost	28	29	29	30	30	30	31	35	35	36
Dráha	84	68	54	60	101	67	77	85	107	79
Rychlost	39	40	40							
Dráha	138	110	134							

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

- Určete bodové odhady parametrů regresní přímky, reziduální rozptyl, směrodatné chyby těchto odhadů a zkonstruujte 95% intervaly spolehlivosti pro parametry regresní přímky.
- Určete bodové odhady parametrů kvadratické regresní funkce, reziduální rozptyl, směrodatné chyby těchto odhadů a zkonstruujte 95% intervaly spolehlivosti pro parametry této regresní funkce.
- Odhadněte brzdnou dráhu pro rychlost 20 mil/hod pomocí obou modelů. Pro oba modely určete interval 95% interval spolehlivosti pro regresní funkci i pro predikci pro rychlost 20 mil/hod.

[Datový soubor: brzdna_draha.txt]

- V padesátých letech došlo k úniku radioaktivního odpadu ze skládky v Hanfordu ve státě Washington do řeky Columbia River. V devíti okrscích níže po proudu ve státě Oregon bylo počítáno vystavení radioaktivitě X (na základě vzdálenosti od Hanfordu a vzdálenosti průměrného obyvatele od řeky apod.). Současně se sledovala úmrtnost na rakovinu Y (úmrtnost na 100 000 lidí za rok v letech 1959–64). Získané údaje jsou shrnuty v následující tabulce.

okrsek	radioaktivní vystavení X	úmrtnost na rakovinu Y
Clatsop	8,3	210
Columbia	6,4	180
Cilliam	3,4	130
Hood River	3,8	170
Morrow	2,6	130
Portland	11,6	210
Sherman	1,2	120
Umatilla	2,5	150
Wasco	1,6	140

Pro daný datový soubor odhadněte parametry těchto modelů:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, \dots, n$$

$$Y_i = \beta_0 + \beta_1 \ln x_i + \epsilon_i, i = 1, \dots, n.$$

- Pro regresní přímku určete reziduální rozptyl, směrodatné chyby odhadů regresních parametrů a zkonstruujte 95% intervaly spolehlivosti pro tyto parametry.
- Pro logaritmickou regresní křivku určete reziduální rozptyl, směrodatné chyby odhadů regresních parametrů a zkonstruujte 95% intervaly spolehlivosti pro tyto parametry.
- Odhadněte úmrtnost pro hodnotu radioaktivní vystavení 8 pomocí obou modelů. Pro oba modely určete interval 95% interval spolehlivosti pro regresní funkci i pro predikci pro hodnotu radioaktivní vystavení 8.
- Pro oba modely zkonstruujte pásy spolehlivosti pro regresní funkce a predikované hodnoty, výsledky zobrazte graficky.

[Datový soubor: radiace_umrtnost.txt]

- Cílem studie bylo nalézt závislost mezi tělesným tukem lehkých atletů-běžců y , kteří trénují asi 12 hodin, a zkonsumovaným tukem v jejich každodenní stravě x . U náhodného vzorku 18 běžců byl měřen jejich tělesný podkožní tuk y [%] a sledován v závislosti na zkonsumovaném tuku ve stravě x [%].

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

x	22,0	30,0	24,0	22,0	21,0	36,0	14,0	17,0	20,0
y	9,80	9,70	12,00	11,70	11,60	11,60	8,00	8,60	10,40
x	21,0	35,0	37,0	32,0	35,0	35,0	26,0	24,0	14,0
y	9,70	11,20	10,80	10,90	12,30	11,50	7,80	10,20	7,90

Pro daný datový soubor odhadněte parametry těchto modelů:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, \dots, n$$

$$Y_i = \beta_0 + \frac{\beta_1}{x_i} + \epsilon_i, i = 1, \dots, n.$$

- Pro regresní přímku určete reziduální rozptyl, směrodatné chyby odhadů regresních parametrů a zkonstruujte 95% intervaly spolehlivosti pro tyto parametry.
- Pro hyperbolickou regresní křivku určete reziduální rozptyl, směrodatné chyby odhadů regresních parametrů a zkonstruujte 95% intervaly spolehlivosti pro tyto parametry.
- Odhadněte množství podkožního tuku pro hodnotu tuku ve stravě 30 % pomocí obou modelů. Pro oba modely určete interval 95% interval spolehlivosti pro regresní funkci i pro predikci hodnotu tuku ve stravě 30 %.
- Pro oba modely zkonstruujte pásy spolehlivosti pro regresní funkce a predikované hodnoty, výsledky zobrazte graficky.

[Datový soubor: tuk_sportovci.txt]

6. U automobilu byla změřena spotřeba Y v závislosti na rychlosti X . Údaje jsou v tabulce

Rychlost [km/hod.]	40	50	60	70	80	90	100	110
Spotřeba [l/100 km]	5,7	5,4	5,2	5,2	5,8	6,0	7,5	8,1

Pro daný datový soubor odhadněte parametry těchto modelů:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, \dots, n$$

$$Y_i = \beta_0 + \beta_1 \ln x_i + \beta_2 x_i^2 + \epsilon_i, i = 1, \dots, n.$$

- Pro regresní přímku určete reziduální rozptyl, směrodatné chyby odhadů regresních parametrů a zkonstruujte 95% intervaly spolehlivosti pro tyto parametry.
- Pro logaritmickou regresní křivku určete reziduální rozptyl, směrodatné chyby odhadů regresních parametrů a zkonstruujte 95% intervaly spolehlivosti pro tyto parametry.
- Odhadněte spotřebu pro hodnotu rychlosti 75 km/hod. pomocí obou modelů. Pro oba modely určete interval 95% interval spolehlivosti pro regresní funkci i pro predikci pro hodnotu rychlosti 75 km/hod.
- Pro oba modely zkonstruujte pásy spolehlivosti pro regresní funkce a predikované hodnoty, výsledky zobrazte graficky.

[Datový soubor: rychlost_spotreba.txt]

7. Data popisují výsledky vstupních zdravotních testů uchazečů o službu u policie.

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Tlak	66	87	85	59	76	77	70	66	75	66
Hmotnost	87,36	117,6	82,85	62,32	82	102	70,12	88,07	77,96	74,33
Tuk	16,98	27,6	6,61	3,26	19	27	6,88	18,8	18,87	8,15
Tlak	74	68	72	76	94	63	80	67	77	78
Hmotnost	56,2	81,75	80,24	74,81	61,98	95,23	72,48	92,45	104,56	66,2
Tuk	3,44	20,31	12,96	12,42	3,58	12,91	11,34	17,5	18,93	10,94
Tlak	77	67	78	78	80	95	76	78	73	80
Hmotnost	87,16	82,42	64,11	81,57	99,85	78,49	87,13	65,64	51,76	67,14
Tuk	17,72	9,55	9,54	13,1	17,75	9,57	18,52	6,4	2,86	4,31
Tlak	81	61	65	69	66	75	72	66	93	77
Hmotnost	78,74	86,83	70,48	72,67	85,86	84,86	66,97	68,33	63,34	85,72
Tuk	16,26	9,72	6,29	4,37	14,43	17	5,8	8,14	3,63	23,61
Tlak	68	71	84	81	74	79	89	79	80	67
Hmotnost	89	95,17	84,19	63,12	70,01	82,11	71	94,56	70,91	79,19
Tuk	18,83	19,16	15,83	8,77	6,61	22,22	8,29	26,82	9,32	19,9

Popište vhodným regresním modelem (pokud to lze) závislost tlaku na hmotnosti a procenta tuku v těle. Najděte vhodný model pro popis závislosti hmotnosti na procentech tuku v těle.

- Určete reziduální rozptyl, směrodatné chyby odhadů regresních parametrů a zkonstruujte 95% intervaly spolehlivosti pro oba modely.
- Pro model popisující závislost hmotnosti na procentu tuku v těle odhadně hmotnost pro hodnotu tuku 20 %. Spočtete 95% intervaly spolehlivosti pro hodnotu regresní funkce i predikovanou hodnotu v tomto bodě.

[Datový soubor: vstupni_testy.txt]