6

Регресиона анализа

Задача 6.1. За време на едно истражување за врската меѓу приходот и писменоста на жителите во афричките држави, забележани се следните податоци за пет афрички држави:

годишен приход по жител (во долари)	110	370	380	500	500
неписменост (во проценти)	85	75	73	63	61

а) Врз основа на дадените податоци најди ги оценките на параметрите на моделот на проста линеарна регресија кој ја опишува зависноста на неписменоста од годишниот приход по жител.

б) Предвиди го процентот на неписменост при годишен приход од 200 и од 1000 долари.

в) Најди 95% интервал на доверба за стапката на промена на процентот на неписменост, при единица промена на годишниот приход по жител.

г) Со 5% ниво на значајност, тестирај ја хипотезата дека националниот приход по жител не влијае на процентот на неписменост.

Решение. а) Годишниот приход по жител е независната променлива x, а процентот на неписменост е зависната променлива Y. За дадените податоци имаме

$$\overline{x} = \frac{1}{5}\sum_{i=1}^{5} x_i = \frac{1}{5}\cdot 1860 = 372,$$

$$\overline{y} = \frac{1}{5}\sum_{i=1}^{5} y_i = \frac{1}{5}\cdot 357 = 71,4,$$

$$s_x^2 = \frac{1}{5}\sum_{i=1}^{5} x_i^2 - \overline{x}^2 = \frac{1}{5}\cdot 793400 - 372^2 = 20296,$$

$$s_{xy} = \frac{1}{5}\sum_{i=1}^{5} x_i y_i - \overline{xy} = \frac{1}{5}\cdot 126840 - 372\cdot 71,4 = -1192,8.$$

Тогаш, оценките на параметрите на моделот на проста линеарна регресија, кој ја опишува зависноста на неписменоста од годишниот приход по жител т.е. $Y_i = ax_i + b + \varepsilon_i$, каде $E(\varepsilon_i) = 0$ и $D(\varepsilon_i) = \sigma^2$ и ε_i се независни, се

$$a = \frac{s_{xy}}{s_x^2} = \frac{-1192,8}{20296} = -0,0587702,$$

$$b = \overline{y} - a\overline{x} = 71,4 - (-0,0587702)\cdot 372 = 93,2625,$$

од каде предвидувањата за процентот на неписменост y за познат годишен приход x се прават според формулата $y = ax + b = -0,0587702x + 93,2625$. Оценката на дисперзијата на случајната компонента, однсно средноквадратното отстапување на предвидените вредности според моделот од точните вредности, е

$$\begin{aligned}
\sigma^2 &= \frac{1}{5}\sum_{i=1}^{5}(y_i - (ax_i + b))^2 = \\
&= \frac{1}{5}\sum_{i=1}^{5} y_i^2 + a^2\frac{1}{5}\sum_{i=1}^{5} x_i^2 + b^2 - 2a\frac{1}{5}\sum_{i=1}^{5} x_i y_i - 2b\frac{1}{5}\sum_{i=1}^{5} y_i + 2ab\frac{1}{5}\sum_{i=1}^{5} x_i = \\
&= \frac{1}{5}\cdot 25869 + (-0,0587702)^2\cdot\frac{1}{5}\cdot 793400 + 93,2625^2 - \\
&\quad -2\cdot(-0,0587702)\cdot\frac{1}{5}\cdot 126840 - 2\cdot 93,2625\cdot\frac{1}{5}\cdot 357 + \\
&\quad +2\cdot(-0,0587702)\cdot 93,2625\cdot\frac{1}{5}\cdot 1860 = 5,7389.
\end{aligned}$$

б) Според моделот, предвидениот процент на неписменост при годишен приход од 200 долари е

$$a\cdot 200 + b = -0,0587702\cdot 200 + 93,2625 = 81,5085\%,$$

додека пак предвидениот процент на неписменост при годишен приход од 1000 долари е

$$a\cdot 1000 + b = -0,0587702\cdot 1000 + 93,2625 = 34,4923\%.$$

Да забележиме дека првото предвидување е во рамките на моделот (моделот е изграден врз основа на податоци за годишен приход од 110 до 500 долари), додека второто предвидување излегува од рамките на моделот, и може да биде подложно на други непознати фактори, па не е препорачливо да се користи.

Ирена Стојковска

в) Се бара 95% интервал на доверба за стапката на промена на процентот на неписменост, при единица промена на годишниот приход по жител т.е. 95% интервал на доверба за параметарот a од моделот.

При дополнителна претпоставка за нормално распределени случајни компоненти $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, ја користиме статистиката

$$T_1 = \frac{(\hat{a} - a)\sqrt{(n-2)s_x^2}}{\hat{\sigma}} \sim t_{n-2},$$

како централна статистика, каде $\hat{a}$ и $\hat{\sigma}^2$ се оценувачи на параметрите a и σ^2 соодветно.

За да го одредиме $(1-\alpha)100\%$ интервал на доверба за a, бараме $c, d \in \mathbb{R}$ така што $P\{c < T_1 < d\} = 1 - \alpha$. Статистиката T_1 има студентова распределба која е симетрична во однос на 0, па минимален интервал на доверба се добива за $c = -d$ и тогаш посленото равенство преминува во $P\{|T_1| < d\} = 1 - \alpha$, од каде $P\{|T_1| \geq d\} = 1 - (1-\alpha) = \alpha$, па $d = t_{n-2,\alpha}$ е број кој се чита од таблицата за студентова распределба. Тогаш $c = -d = -t_{n-2,\alpha}$.

Сега, од $P\{c < T_1 < d\} = 1 - \alpha$, имаме

$$P\{c < \frac{(\hat{a} - a)\sqrt{(n-2)s_x^2}}{\hat{\sigma}} < d\} = 1 - \alpha,$$

$$P\{\frac{c\hat{\sigma}}{\sqrt{(n-2)s_x^2}} < \hat{a} - a < \frac{d\hat{\sigma}}{\sqrt{(n-2)s_x^2}}\} = 1 - \alpha,$$

$$P\{-\hat{a} + \frac{c\hat{\sigma}}{\sqrt{(n-2)s_x^2}} < -a < -\hat{a} + \frac{d\hat{\sigma}}{\sqrt{(n-2)s_x^2}}\} = 1 - \alpha,$$

$$P\{\hat{a} - \frac{d\hat{\sigma}}{\sqrt{(n-2)s_x^2}} < a < \hat{a} - \frac{c\hat{\sigma}}{\sqrt{(n-2)s_x^2}}\} = 1 - \alpha,$$

$$P\{\hat{a} - \frac{t_{n-2,\alpha}\,\hat{\sigma}}{\sqrt{(n-2)s_x^2}} < a < \hat{a} + \frac{t_{n-2,\alpha}\,\hat{\sigma}}{\sqrt{(n-2)s_x^2}}\} = 1 - \alpha,$$

односно

$$I_a = \left(\hat{a} - \frac{t_{n-2,\alpha}\,\hat{\sigma}}{\sqrt{(n-2)s_x^2}}, \hat{a} + \frac{t_{n-2,\alpha}\,\hat{\sigma}}{\sqrt{(n-2)s_x^2}}\right)$$

е $(1-\alpha)100\%$ интервал на доверба за a. За дадените податоци $n = 5$ и $\alpha = 0,05$, па од таблицата се наоѓа $t_{n-2,\alpha} = t_{3;0,05} = 3,182$. Врз основа на дадените податоци, реализациите на оценувачите $\hat{a}$ и $\hat{\sigma}^2$ се најдени под а) и изнесуваат $a = -0,0587702$ и $\sigma^2 = 5,7389$ соодветно, од каде $\sigma = 2,3956$, а $s_x^2 = 20296$ е исто така пресметано под а). Овие вредности ги заменуваме во интервалот на добверба и добиваме дека 95% интервал на доверба за a е

$$I_a = (-0,0896624; -0,027878).$$

Ирена Стојковска

г) Се бара да се тестира хипотезата $H_0 : a = 0$ против алтернативната $H_1 : a \neq 0$, со ниво на значајност $\alpha = 5\% = 0,05$. Од в) најдовме дека $(1-\alpha)100\%$ интервал на доверба за a е $I_a = \left(\hat{a} - \frac{t_{n-2,\alpha}\,\hat{\sigma}}{\sqrt{(n-2)s_x^2}}, \hat{a} + \frac{t_{n-2,\alpha}\,\hat{\sigma}}{\sqrt{(n-2)s_x^2}}\right)$, тогаш критичната област за тестирање на $H_0 : a = a_0$, против $H_1 : a \neq a_0$, со ниво на значајност α, е

$$C = \{(x,y) \mid a - \frac{t_{n-2,\alpha}\,\sigma}{\sqrt{(n-2)s_x^2}} \geq a_0 \text{ или } a + \frac{t_{n-2,\alpha}\,\sigma}{\sqrt{(n-2)s_x^2}} \leq a_0\},$$

каде a и σ^2 се реализации на оценувачите $\hat{a}$ и $\hat{\sigma}^2$ соодветно. За $a_0 = 0$, добиваме дека критичната област за тестирање на $H_0 : a = 0$ против $H_1 : a \neq 0$, со ниво на значајност α, е

$$C = \{(x,y) \mid a - \frac{t_{n-2,\alpha}\,\sigma}{\sqrt{(n-2)s_x^2}} \geq 0 \text{ или } a + \frac{t_{n-2,\alpha}\,\sigma}{\sqrt{(n-2)s_x^2}} \leq 0\},$$

односно

$$C = \{(x,y) \mid \frac{a\sqrt{(n-2)s_x^2}}{\sigma} \geq t_{n-2,\alpha} \text{ или } \frac{a\sqrt{(n-2)s_x^2}}{\sigma} \leq -t_{n-2,\alpha}\},$$

што преминува во

$$C = \{(x,y) \mid \left|\frac{a\sqrt{(n-2)s_x^2}}{\sigma}\right| \geq t_{n-2,\alpha}\},$$

затоа што $t_{n-2,\alpha} > 0$. За дадените податоци имаме $n = 5$, $\alpha = 0,05$, од каде $t_{n-2,\alpha} = t_{3;0,05} = 3,182$. Од а) имаме дека $a = -0,0587702$ и $\sigma^2 = 5,7389$, од каде $\sigma = 2,3956$, и $s_x^2 = 20296$. Тогаш,

$$\left|\frac{a\sqrt{(n-2)s_x^2}}{\sigma}\right| = \left|\frac{-0,0587702 \cdot \sqrt{(5-2)\cdot 20296}}{2,3956}\right| = 6,05353 > 3,182 = t_{n-2,\alpha},$$

од каде $(x,y) \in C$, па хипотезата H_0 се отфрла, односно се отфрла хипотезата дека националниот проход по жител не влијае на процентот на писменост.

Забелешка. Интерпретацијата на параметрите кај моделот на проста линеарна регресија $Y = ax + b + \varepsilon$ е следна: Параметарот a е стапката на промена на зависната променлива Y, при единица промена на независната променлива x. Параметарот b е фиксниот дел од Y кој не зависи од x. Вредноста $ax + b$ е очекуваната вредност на Y за дадена вредност на x.

Ирена Стојковска

Задачи за самостојна работа

Задача 6.2. На 10 студенти измерени им се следните вредности за масата (во kg) и висината (во cm):

маса (во kg)	90	65	76	49	85	58	64	73	83	93
висина (во cm)	194	164	162	155	174	164	170	184	185	183

За дадените податоци важи $\sum x_i = 736$, $\sum x_i^2 = 56054$, $\sum y_i = 1735$, $\sum y_i^2 = 302443$, $\sum x_i y_i = 129015$.

а) Врз основа на дадените податоци најди ги оценките на параметрите на моделот на простата линеарна регресија кој ја опишува зависноста на масата од висината.

б) Најди ја очекуваната маса на лице кое има висина од 160 cm.

в) Најди 95% интервал на доверба за очекуваната маса на лице кое има висина од 160 cm.

г) Со 5% ниво на значајност тестирај ја хипотезата дека очекуваната маса на лице кое има висина од 160 cm е 70 kg, против алтернативната дека помала од 70 kg.

Задача 6.3. Се претпосатува дека меѓу процентот на неотплатени кредити и висината на камтната стапка постои линеарна стохастичка врска. Од евиденцијата на кредитното одделение извадени се следните податоци:

висина на кам. стапка (во %)	3,50	3,60	8,75	9,50	10,00	11,50	12,00	18,00	20,00
неотплатени кредити (во %)	4,0	3,5	6,2	6,8	6,7	7,4	7,9	9,2	10,2

За дадените податоци важи $\sum x_i = 96,85$, $\sum x_i^2 = 1292,27$, $\sum y_i = 61,9$, $\sum y_i^2 = 463,67$, $\sum x_i y_i = 769,95$.

а) Врз основа на дадените податоци најди ги оценките на параметрите на моделот на простата линеарна регресија кој ја опишува зависноста на процентот неотплатени кредити од висината на каматната стапка.

б) Предвиди го процентот на неотплатени кредити при висина на каматна стапка од 7%.

в) Најди 95% интервал на доверба за фиксниот дел од неотплатените кредити кој не зависи од висината на каматната стапка.

г) Со 5% ниво на значајност тестирај ја хипотезата дека фиксниот дел од неотплатените кредити кој не зависи од висината на каматната стапка е 2%.

Ирена Стојковска