6

Регресиона анализа

При мерење на повеќе обележја, често се поставува прашањето како тие зависат едно од друго. На пример, како зависи нивото на холестеролот во крвта од возраста на индивидуата, како зависи потрошувачката на електричната енергија во домаќинството од надворешната температура и месечните примања на членовите на домаќинството и слично. При тоа променливата која е предмет на истражувањето се нарекува **зависна променлива** и се означува со Y и таа е случајна променлива (на пример, нивото на холестеролот во крвта, односно потрошувачката на електричната енергија во домаќинството), додека променливата која може да се контролира и која влијае на промените на зависната променлива се нарекува **независна променлива** и во општ случај нив може да ги има повеќе од една и се означуваат со $x^{(1)}, ..., x^{(r)}$ и тие не се случајни променливи (на пример, возраста на индивидуата, односно надворешната температура и месечните примања на членовите на домаќинството).

Одредувањето на зависноста на променливата Y од независните променливи $x^{(1)}, ..., x^{(r)}$ е преку формирање на математички модел кој ќе ја изразува таа зависност. Бидејќи во реалните ситуации на зависноста често влијаат и некои случајни фактори (на пример, грешки при мерењата), треба и тие да бидат земени во предвид при формирањето на моделот. Општиот облик на овој модел наречен **модел на регресија** е

$$Y = f_t(x^{(1)}, ..., x^{(r)}) + \varepsilon, \tag{6.1}$$

каде функцијата f_t се нарекува **регресиона функција** и зависи од параметарот t кој многу често е векторски параметар, а ε е **случајна компонента** со $E(\varepsilon) = 0$ и $D(\varepsilon) = \sigma^2$.

Кога имаме само една независна променлива x, станува збор за **модел на еднодимензионална регресија** т.е.

$$Y = f_t(x) + \varepsilon, \tag{6.2}$$

а кога се повеќе од една независна променлива, имаме **модел на повеќедимензионална регресија**.

Изборот на соодветниот модел на регресија зависи од конкретната ситуација. Затоа, во случај на еднодимензионална регресија, најнапред се прикажуваат графички статистичките податоци $(x_1, y_1), ..., (x_n, y_n)$, каде $x_1, ..., x_n$ се вредности на независната променлива x, а $y_1, ..., y_n$ се соодветните вредности на случајната променлива Y добиени како резултат на набљудувања или мерења, а потоа се бара општиот облик и отстапувањата од овој облик, со цел да се најде математичкиот модел кој најдобро би го опишал обликот.

Основната задача на регресионата анализа е да после изборот на моделот, врз основа на низата статистички податоци, да ги процени непознатите параметри t и σ^2. Попрецизно кажано, во случај на еднодимензионална регресија, треба да се дефинираат оценувачи за непознатите параметри t и σ^2, врз основа на примерокот $(x_1, Y_1), ..., (x_n, Y_n)$, каде $x_1, ..., x_n$ се вредности на независната променлива x и $Y_1, ..., Y_n$ се соодветните случајни променливи дефинирани со

$$Y_i = f_t(x_i) + \varepsilon_i,\ i = 1, ..., n, \tag{6.3}$$

при што за случајните променливи $\varepsilon_1, ..., \varepsilon_n$ претпоставуваме дека се независни со $E(\varepsilon_i) = 0$ и $D(\varepsilon_i) = \sigma^2$, $i = 1, ..., n$.

6.1 Линеарна регресија

Ако регресионата функција f_t во (6.1) е линеарна, станува збор за **линеарен модел на регресија**. Ако кај линеарниот модел имаме само една независна променлива x, тогаш таквиот модел е **прост линеарен модел на регресија**, и тогаш $t = (a, b)$, па моделот е

$$Y = ax + b + \varepsilon, \tag{6.4}$$

каде a и b се параметри и ε е случајна променлива со $E(\varepsilon) = 0$ и $D(\varepsilon) = \sigma^2$. Да забележиме дека тогаш,

$$E(Y) = E(ax + b + \varepsilon) = ax + b + E(\varepsilon) = ax + b,$$

$$D(Y) = D(ax + b + \varepsilon) = D(\varepsilon) = \sigma^2,$$

затоа што $ax + b$ не е случајна променлива.

Смислата на моделот (6.4) е следната. Случајната променлива Y зависи од променливата x, на тој начин што за секој $x = x_i$, $Y_i = ax_i + b + \varepsilon_i$ има еден детерминистички собирок кој е линеарна функција од x_i т.е. $ax_i + b$ и

стохастички собирок ε_i кој претставува случајна осцилација околу детерминистичкиот собирок затоа што $E(\varepsilon_i) = 0$. При едно реализирано мерење, добиваме една реализација на примерокот $(x_1, Y_1), ..., (x_n, Y_n)$, која ја означуваме со $(x_1, y_1), ..., (x_n, y_n)$. Нејзиниот графички приказ има тенденција на групитање околу правата $y = ax + b$. Групирањето зависи од распределбата на $\varepsilon_1, ..., \varepsilon_n$, имено колку нивната дисперзија $D(\varepsilon_i) = \sigma^2$, $i = 1, ..., n$ е поголема, толку растурањето на тие точки околу правата $y = ax + b$ е поголемо.

6.1.1 Оценување на параметрите на регресија

Врз основа на примерокот $(x_1, Y_1), ..., (x_n, Y_n)$, треба да најдеме оценувачи $\widehat{a}$, $\widehat{b}$, $\widehat{\sigma}^2$ за непознатите параметри a, b, σ^2 на линеарниот модел на регресија (6.4). Без воведување на дополнителни претпоставки за распределбата на случајните променливи ε_i, со **методот на најмали квадрати** може да се најдат оценувачите $\widehat{a}$ и $\widehat{b}$.

Нека $(x_1, y_1), ..., (x_n, y_n)$ е реализација на примерокот $(x_1, Y_1), ..., (x_n, Y_n)$, тогаш за реализациите $\widehat{a}'$ и $\widehat{b}'$ на оценувачите $\widehat{a}$ и $\widehat{b}$, според методот на најмали квадрати треба да важи

$$\min_{(a,b)\in\mathbb{R}^2} \sum_{i=1}^{n} (y_i - ax_i - b)^2 = \sum_{i=1}^{n} (y_i - \widehat{a}'x_i - \widehat{b}')^2.$$

Па, слично како претходно (види Дескриптивна статистика, Дводимензионални обележја, оредување на коефициентите на правата на регресија), со решавање на системот

$$\frac{\partial S(a,b)}{\partial a} = 0, \ \frac{\partial S(a,b)}{\partial b} = 0,$$

каде

$$S(a,b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2,$$

се добиваат решенијата (оценките на a и b добиени со методот на најмали квадрати)

$$\widehat{a}' = \frac{s_{xy}}{s_x^2}, \ \widehat{b}' = \overline{y} - \widehat{a}'\overline{x},$$

каде

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \ \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \ s_x^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2,$$

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}),$$

(покажи!). Ако означиме со

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

тогаш оценувачите добиени со метод на најмали квдрати, статистики кои зависат од примерокот $(x_1, Y_1), ..., (x_n, Y_n)$ се

$$\widehat{a} = \frac{1}{ns_x^2}\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y}) = \frac{1}{ns_x^2}\sum_{i=1}^{n}(x_i - \overline{x})Y_i, \tag{6.5}$$

$$\widehat{b} = \overline{Y} - \widehat{a}\overline{x} = \frac{1}{n}\sum_{i=1}^{n}\Big(1 - \frac{\overline{x}}{s_x^2}(x_i - \overline{x})\Big)Y_i, \tag{6.6}$$

(покажи!). Исто така, според методот на најмали квадрати, како оценувач за σ^2 може да се земе

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - (\widehat{a}x_i + \widehat{b}))^2. \tag{6.7}$$

Бидејќи $E(Y_i) = ax_i + b$ и $D(Y_i) = \sigma^2$, за оценувачите $\widehat{a}$ и $\widehat{b}$ имаме

$$E(\widehat{a}) = E\Big(\frac{1}{ns_x^2}\sum_{i=1}^{n}(x_i - \overline{x})Y_i\Big) = \frac{1}{ns_x^2}\sum_{i=1}^{n}(x_i - \overline{x})E(Y_i) = \frac{1}{ns_x^2}\sum_{i=1}^{n}(x_i - \overline{x})(ax_i + b) =$$

$$= \frac{1}{s_x^2}\Big(\frac{a}{n}\sum_{i=1}^{n}(x_i^2 - x_i\overline{x}) + \frac{b}{n}\sum_{i=1}^{n}(x_i - \overline{x})\Big) = \frac{1}{s_x^2}(as_x^2 + b\cdot 0) = a,$$

$$E(\widehat{b}) = E(\overline{Y} - \widehat{a}\overline{x}) = E(\overline{Y}) - E(\widehat{a})\overline{x} = \frac{1}{n}\sum_{i=1}^{n}E(Y_i) - a\overline{x} =$$

$$= \frac{1}{n}\sum_{i=1}^{n}(ax_i + b) - a\overline{x} = a\overline{x} + b - a\overline{x} = b,$$

што значи дека $\widehat{a}$ и $\widehat{b}$ се непристрасни оценувачи за a и b соодветно. Понатаму, од независноста на ε_i, $i = 1, ..., n$ следи и независноста на $Y_i = ax_i + b + \varepsilon_i$, $i = 1, ..., n$ (покажи!). Затоа, за дисперзиите на $\widehat{a}$ и $\widehat{b}$ имаме

$$D(\widehat{a}) = \frac{\sigma^2}{ns_x^2} \to 0,\ D(\widehat{b}) = \frac{\sigma^2}{n}\Big(1 + \frac{\overline{x}^2}{s_x^2}\Big) \to 0,$$

кога $n \to \infty$ (покажи!), од каде заклучуваме дека $\widehat{a}$ и $\widehat{b}$ се конзистентни оценувачи за a и b соодветно.

Ирена Стојковска

Ако сега ја разгледаме статистиката $\widehat{a}x+\widehat{b}$ како оценувач за $ax+b$, бидејќи

$$E(\widehat{a}x+\widehat{b})=ax+b,\ D(\widehat{a}x+\widehat{b})=\frac{\sigma^2}{n}\Big(1+\frac{(x-\overline{x})^2}{s_x^2}\Big)\to 0,$$

кога $n\to\infty$ (покажи!), заклучуваме дека $\widehat{a}x+\widehat{b}$ е непристрасен и конзистентен оценувач за $ax+b$.

Но, за оценувачот $\widehat{\sigma}^2$ дефиниран со (6.7) имаме

$$E(\widehat{\sigma}^2)=E\Big(\frac{1}{n}\sum_{i=1}^{n}(Y_i-(\widehat{a}x_i+\widehat{b}))^2\Big)=\frac{n-2}{n}\sigma^2$$

(покажи!), од каде заклучуваме дека тој не е непристрасен оценувач за σ^2. Затоа, неговата корекција, оценувачот

$$\breve{\sigma}^2=\frac{1}{n-2}\sum_{i=1}^{n}(Y_i-(\widehat{a}x_i+\widehat{b}))^2 \tag{6.8}$$

е непристрасен оценувач за σ^2.

За да ги користиме најдените оценувачи при барање на **интервали на доверба** и **тесирање на хипотези** за параметрите на регресија, потребно е да ги знаеме и нивните распределби, точни или асимптотски. Така, **во случај на голем примерок**, од централната гранична теорема, следи дека оценувачите $\widehat{a}$ и $\widehat{b}$ се асимптотски нормални оценувачи за a и b соодветно, односно

$$\widehat{a}\xrightarrow{dist}\mathcal{N}\Big(a,\ \frac{\sigma^2}{ns_x^2}\Big),\ \widehat{b}\xrightarrow{dist}\mathcal{N}\Big(b,\ \frac{\sigma^2}{n}\Big(1+\frac{\overline{x}^2}{s_x^2}\Big)\Big).$$

Истото важи и за оценувачот $\widehat{a}x+\widehat{b}$, за кој се покажува дека е асимптоцки нормален оценувач за $ax+b$, односно

$$\widehat{a}x+\widehat{b}\xrightarrow{dist}\mathcal{N}\Big(ax+b,\ \frac{\sigma^2}{n}\Big(1+\frac{(x-\overline{x})^2}{s_x^2}\Big)\Big).$$

Па, бри барање на интервали на доверба и тестирање на хипотези за a, b и $ax+b$ врз основа на голем примерок, се користиме со гореспоменатите распределби, при што непознатиот параметар σ^2 го заменуваме со реализација на оценувачот дефиниран со (6.7) или (6.8), што во пракса нема големо значење која оценка ќе се одбере затоа што за големи вредности на n разликата меѓу двете оценки е незначителна.

За да ги најдеме точните распределби на оценувачите $\widehat{a}$ и $\widehat{b}$, **во случај на мал примерок** потребни се дополнителни претпоставки за случајните променливи ε_i, $i=1,...,n$ во линеарниот модел на регресија $Y_i=ax_i+b+\varepsilon_i$, $i=1,...,n$.

Имено, покрај независноста се претпоставува и дека $\varepsilon_i \sim \mathcal{N}(0,\ \sigma^2),\ i = 1, ..., n$. Од тука следи дека и случајните променливи $Y_i,\ i = 1, ..., n$ се нзависни и за нивните распределби важи

$$Y_i \sim \mathcal{N}(ax_i + b,\ \sigma^2),\ i = 1, ..., n.$$

Па тогаш, за точните распределби на оценувачите $\widehat{a}$ и $\widehat{b}$ дефинирани со (6.5) и (6.6) соодветно, имаме

$$\widehat{a} \sim \mathcal{N}\Big(a,\ \frac{\sigma^2}{ns_x^2}\Big),\ \widehat{b} \sim \mathcal{N}\Big(b,\ \frac{\sigma^2}{n}\Big(1 + \frac{\overline{x}^2}{s_x^2}\Big)\Big),$$

бидејќи претставуваат линеарна комбинација од случајните променливи Y_i, $i = 1, ..., n$. Во овој случај, случајната променлива $\widehat{\sigma}^2$ дефинирана со (6.7), а исто така и корегираниот оценувач $\breve{\sigma}^2$ дефиниран со (6.8), претставуваааат збир од квадрати на одредени линеарни комбинации од нормално распределени случајни променливи Y_i, $\widehat{a}$ и $\widehat{b}$, кои не се независни. Но, се покажува дека, и во тој случај, распределбата на така претставената случајна променлива е хи-квадрат распределба. Имено, важи

$$\frac{n}{\sigma^2}\,\widehat{\sigma}^2 \sim \chi^2_{n-2},\ \text{односно}\ \frac{n-2}{\sigma^2}\,\breve{\sigma}^2 \sim \chi^2_{n-2}.$$

Исто така, се покажува дека $\widehat{a}$ и $\widehat{\sigma}^2$, но и $\widehat{b}$ и $\widehat{\sigma}^2$ се независни, што помага при одредување на точните распределби на статистиките кои се користат при наоѓање на интервалите на доверба и тестирање на хипотезите за параметрите a, b и $ax + b$, имено важи

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

$$T_2 = \frac{(\widehat{b} - b)\sqrt{(n-2)s_x^2}}{\widehat{\sigma}\sqrt{s_x^2 + \overline{x}^2}} = \frac{(\widehat{b} - b)\sqrt{ns_x^2}}{\breve{\sigma}\sqrt{s_x^2 + \overline{x}^2}} \sim t_{n-2},$$

$$T_3 = \frac{(\widehat{a}x + \widehat{b} - ax - b)\sqrt{(n-2)s_x^2}}{\widehat{\sigma}\sqrt{s_x^2 + (x - \overline{x})^2}} = \frac{(\widehat{a}x + \widehat{b} - ax - b)\sqrt{ns_x^2}}{\breve{\sigma}\sqrt{s_x^2 + (x - \overline{x})^2}} \sim t_{n-2}$$

(покажи!).

Оценувачи за параметрите на регресија a, b и σ^2 може да се најдат и со **метод на максимална подобност**. Но, и во тој случај потребни се дополнителните претпоставки за нормална распределеност на случајните променливи ε_i, $i = 1, ..., n$ во линеарниот модел на регресија $Y_i = ax_i + b + \varepsilon_i$, $i = 1, ..., n$, имено покрај независноста се претпоставува и дека $\varepsilon_i \sim \mathcal{N}(0,\ \sigma^2),\ i = 1, ..., n$.

Ирена Стојковска

Го означуваме со $t = (a, b, \sigma^2)$ векторот од непознати параметри кој припаѓа на просторот од параметри

$$\Theta = \{(a, b, \sigma^2) \mid a \in \mathbb{R},\ b \in \mathbb{R},\ \sigma^2 > 0\}.$$

Тогаш, функцијата на подобност е

$$L(x, y, t) = (2\pi\sigma^2)^{-n/2} \exp\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - ax_i - b)^2\}.$$

Со решавање на системот равенки на подобност

$$\frac{\partial \ln L(x, y, t)}{\partial a} = 0,\ \frac{\partial \ln L(x, y, t)}{\partial b} = 0,\ \frac{\partial \ln L(x, y, t)}{\partial \sigma^2} = 0,$$

по a, b и σ^2 се добиваат решенијата

$$a = \frac{s_{xy}}{s_x^2},\ b = \overline{y} - a\overline{x},\ \sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (ax_i + b - y_i)^2$$

(покажи!), од каде заклучуваме дека максимално подобни (ML) оценувачи за a, b и σ^2, се $\widehat{a}$, $\widehat{b}$ и $\widehat{\sigma}^2$, дефинирани со (6.5), (6.6) и (6.7) соодветно.