МЕТОД НА НАЈМАЛИ КВАДРАТИ

Регресионата анализа е една од најкористените методи во статистиката. Регресионата анализа овозможува креирање на *модел* за предвидување на вредности на една нумеричка променлива, базирани на вредности на други променливи (или само на една променлива).

Во регресионата анализа, променливата што сакаме да ја предвидиме се нарекува *зависна* променлива, а променливите што се користат за предвидување се нарекуваат *независни* променливи. За да може да се предвиди некоја нумеричка вредност на независната променлива, потребно е да се воспостави математички модел преку кој е дадена поврзаноста помеѓу независната променлива и зависните променливи.

На вредностите на статичкиот белег што се испитува како независна променлива влијаат повеќе фактори, односно независни променливи. Поединечните независни променливи (фактори) ги изразуваме со $X_1, X_2, \cdots, X_k$, а независната променлива со Y. Претпоставката е дека постои врска во облик

$$Y = f(X_1, X_2, \cdots, X_k) + \varepsilon,$$

при што *f*(.) е функција со одреден облик што зависи само од незвисните променливи $X_1, X_2, \cdots, X_k$. Променливата ε е случајна променлива што ги вклучува сите останати немерливи фактори што влијаат на зависната променлива *Y*.

Ако обликот на функцијата *f*(.) е линеарен, тогаш се работи за *линеарен статистички регресионен модел*. Линеарните регресиони модели се користат во бизнисот и економијата. На пример, во бизнисот рутински се користи линеарната поврзаност за оценување на аутпутот, како функција на инпутите. Исто така, продажбата на прозводите или услугите како функција од цената и расположливиот приход во определени региони се важни за планирање на производството и дистрибуцијата. Линеарната регресија овозможува два важни резултата – предвидување на вредноста на зависната променлива како линеарна функција од независните променливи и утврдување на маргиналната (дополнителната) промена на зависната променлива на единечна промена на независните променливи.

Ако се работи за модел каде што се претпоставува дека влијанието на еден фактор, односно независна променлива е најизразено, тогаш станува збор на *прост линеарен регресионен модел*. Целта на регресионата анализа е , користејќи парови на набљудувања (x_i, y_i) , да се одреди „најдобрата“ равенка за линеарната врска меѓу X_1 и *Y*. Всушност, целата е да се најде линија што најдобро се вклопува во набљудуваните податоци.

Линеарниот модел, вклучувајќи ја мерката за необјаснет варијабилитет ε е

$$y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i \text{ , за } i = 1,2, \dots, n \qquad (1)$$

Во моделот β_0 е константа што ја прилагодува линеарната равенка кон податоците, а β_1 е промена на *Y* за единечна промена на X_1 а ε е мерка на случајниот, необјаснетиот варијабилитет, за кој се претпоставува дека има средна вредност 0 и варијанса σ^2.

Методот на најмали квадрати е всушност избор на права линија што го минимизира збирот на оквадратените отстапувања на емпириските набљудувања од точките на линијата. Со методот на најмали квадрати се наоѓаат оценки на коефициентите со одлични статистички својства. Комбинацијата на логичност, едноставност и корисни резултати ја објаснуваат корисноста и масовната употреба на овој метод за изведување на линеарни равенки за деловни и економски апликации.

Оценувањето и заклучувањето по методот на најмали квадрати ги користи следниве претпоставки:

1. Независната променлива X_1 е со познати вредности без случајна грешка
2. Варијансата σ^2 е константна во опсегот на независната променлива X_1. Ова својство се нарекува хомоскедастичност.
3. $\varepsilon_i \sim N(0, \sigma^2)$.
4. ε_i и x_{1i} се независни..
5. Од секое набљудување ε_i се меѓусебно независни.

Бидејќи во моделот (1) делот $\beta_0 + \beta_1 x_1$ е фиксен но непознат, може да се каже дека *Y* за дадено X_1 има иста варијанса σ^2 како ε. Ако на фиксниот дел се додаде случајната променлива ε, тој станува нова случајна променлива (*Y*) со различна аритметичка средина, но иста варијанса σ^2. Во моделот се претпоставува множество на вредности на X_1, од кои секоја има придружено нормално распоредена зависна променлива *Y*. Аритметичката средина на овие вредности на *Y* зависи од вредностите на X_1. Во случај вредностите на *Y* да зависеа само од вредностите на X_1 и да не вклучуваа случајна грешка, тогаш точките на *Y* ќе лежеа на правата линија. Кога претпоставуваме популациска регресиона права линија таа е дадена со

$$Y = \beta_0 + \beta_0 X_1$$

Коефициентите β_0 и β_1 се константи. Сепак, бидејќи се непознати истите се оценуваат од примерок на податоци, така што оценките $\hat{\beta}_0$ и $\hat{\beta}_1$ се функции од реализираните парови на податоци (x_i, y_i). Значи, користејќи ги оценките $\hat{\beta}_0$ и $\hat{\beta}_1$, се добива апроксимативната права линија

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i}$$

Набљудуваните вредности за y_i се еднакви на

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{e}_i$$

каде што $\hat{e}_i$ е набљудуваната разлика меѓу y_i и $\hat{Y}$. ($\hat{e}_i \neq \varepsilon_i$). Според методот на најмали квадрати, треба да се минимизира

$$\sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i_i})^2 = ESS(error\ sum\ of\ sqares)$$

Оценките по методот на најмали квадрати се добиваат така што се бараат парцијални изводи по непознатите оценки на коефициентите, кои потоа се изедначуваат со 0, за да се добијат двете „нормални“ равенки чие што решение ги дава оценките на

коефициентите по методот на најмали квадрати. Оваа постапка е добро познатата постапка за барање на минимум на функција од алгебра. Така, следува

$$\frac{\partial ESS}{\partial \hat{\beta}_0} = 0$$

$$2\sum_{i=1}^{n}(y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_{1i_i}\right)^2 (-1) = 0$$

$$\sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_{1i} = 0$$

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{1i} = \sum_{i=1}^{n} y_i \quad (*)$$

$$\frac{\partial ESS}{\partial \hat{\beta}_1} = 0$$

$$2\sum_{i=1}^{n}(y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_{1i_i}\right)^2 (-x_{1i}) = 0$$

$$\sum_{i=1}^{n} x_{1i} y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_{1i} - \hat{\beta}_1 \sum_{i=1}^{n} x^2{}_{1i} = 0$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_{1i} + \hat{\beta}_1 \sum_{i=1}^{n} x^2{}_{1i} = \sum_{i=1}^{n} x_{1i} y_i \quad (**)$$

Равенките (*) и (**) се двете симултани равенки (попознат како систем на нормални равенки) што кога ќе се решат се добиваат оценетите коефициенти $\hat{\beta}_0$ и $\hat{\beta}_1$.

Детерминантата на системот на нормални равенки е еднаква на

$$D = \begin{vmatrix} \sum_{i=1}^{n} x^2{}_{1i} & \sum_{i=1}^{n} x_{1i} \\ \sum_{i=1}^{n} x_{1i} & n \end{vmatrix} = n\sum_{i=1}^{n} x^2{}_{1i} - \left[\sum_{i=1}^{n} x^2{}_{1i}\right]^2 = n^2 \left\{ \frac{1}{n}\sum_{i=1}^{n} x^2{}_{1i} - \left[\sum_{i=1}^{n} x^2{}_{1i}\right]^2 \right\}$$

Ако ги означиме со $\bar{x}$ и s_x^2 аритметичката средина и варијансата на независната променлива во примерокот детерминантата на системот на нормални равенки е еднаква на

$$D = n^2 s_x^2$$

Бидејќи детерминантата на системот е поголема од 0, системот има едно и само едно решение. Тоа решение се т.н. оценки на најмали квадрати кои што можат да се напишат и како

$$\hat{\beta}_1 = \frac{1}{n s_x^2} \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})$$

$$\hat{\beta}_0 = \bar{y} - \frac{\bar{x}}{ns_x^2}\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) = \bar{y} - \hat{\beta}\,\bar{x}$$

Користени се познатите формули

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Оценките добиени по методот на најмали квадрати се најдобри, линеарни, непристрасни оценки.

Да го илустрираме претходното со еден едноставен пример. Нека се дадени следнве парови на податоци (x_{1i}, y_i) : (1,10), (2,40), (3,80), (4,60) и (5,90). Во тој случај, дијаграмот на растурање би изгледал како на графикот 1.

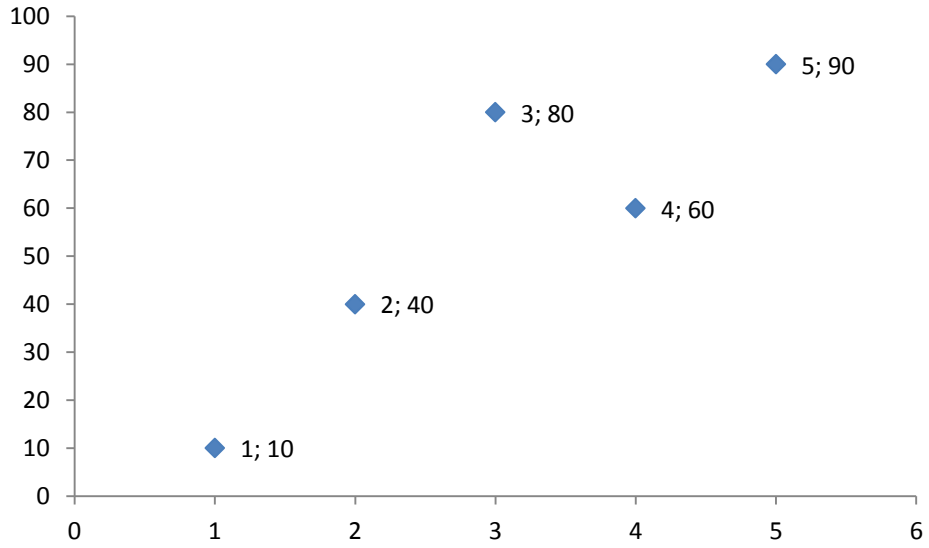

График 1. Дијаграм на растурање за примерот

Пресметаните оценки се

$$\hat{\beta}_1 = \frac{1}{ns_x^2}\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) = \frac{1}{5 \cdot 2}180 = 18$$

$$\hat{\beta}_0 = \bar{y} - \frac{\bar{x}}{ns_x^2}\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) = \bar{y} - \hat{\beta}\,\bar{x} = 56 - 3 \cdot 18 = 2$$

Според тоа, добиената права линија по методот на најмали квадрати е

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} = 2 + 18x_{1i}$$

Ако се нацрта, линијата и емпириските парови на податоци би изгледале како на графикот 2.

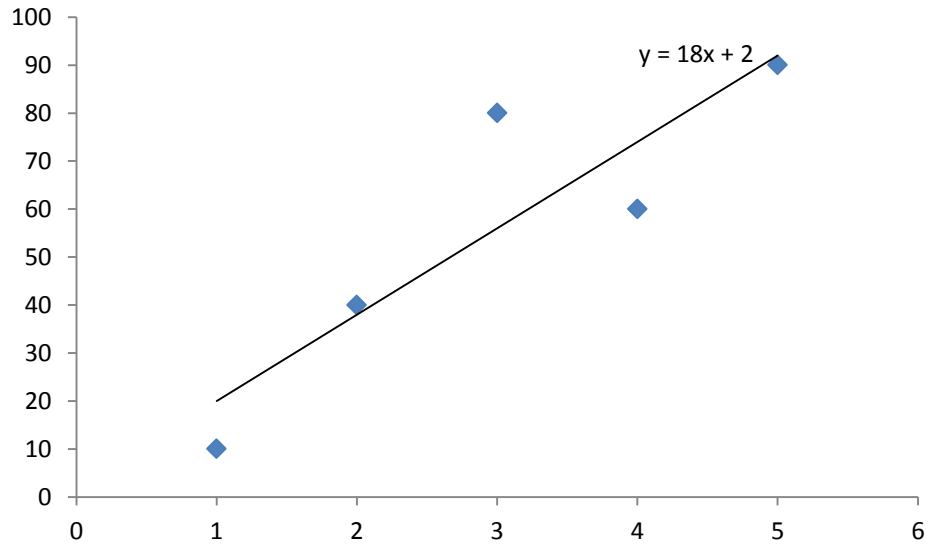

График 2. Линија на најмали квадрати

Методот на најмали квадрати се користи и при анализа на временски серии, кога треба да се оценатa коефициентите на избраната линија на тренд што најдобро ги апроксимира податоците дадени во временска рамка. Во овој случај, не постои зависност меѓу два белега како кај регересионата анализа, туку се врши соодветно кодирање на временските периоди, односно на она што во регресионата анализа се вредности на независната променлива x.

Наводи

1. Berenson, M.L., Levine, M.D, Krehbiel, T.C. *Basic Business Statistics – Concepts and Applications,* Pearson International Edition,2009, стр. 610
2. Carlson, W.L., Thorne, B. Applied *Statistical Methods for Business, Economics, and Social Sciences,* Prentice Hall, 1997, стр. 637
3. Ристески, С. Тевдовски, Д. *Статистика за бизнис и економија,* Економски факултет, 2010, стр.370
4. Sicich, T. *Business Statistics by Example,* Maxwell Macmillan Int. 1992, стр. 570
5. Vukovic N,Vukmirovic, D. *Statistika,* Fakultet organizacionih nauka, 2004,стр.289