ИНСТИТУТ
МАТЕМАТИКИ
МЕХАНИКИ
КОМПЬЮТЕРНЫХ
НАУК

ЮЖНЫЙ
ФЕДЕРАЛЬНЫЙ
УНИВЕРСИТЕТ

Regression model

Kurbatova Natalia Victorovna

nvkurbatova@sfedu.ru

ИНСТИТУТ МАТЕМАТИКИ МЕХАНИКИ КОМПЬЮТЕРНЫХ НАУК

Linear Regression

ЮЖНЫЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ

Definition: Let X - random variable depends on random variable or vector;

Z-values are given or observed.

Denote by $f(t)$ the function of dependence the average value of X on Z: $E(X|Z=t)=f(t)$,

$f(t)$ – is called regression line

$x=f(t)$ – regression equation

Experiment:

Under the condition of n values of Z: $t_1, t_2, \dots t_n$, observed values: $X_1, X_2, \dots X_n$; introduce $\varepsilon_i \equiv X_i - E(X|z=t_i) = X_i - f(t_i)$ – difference between the observed random variables in i-th experiment and expectation of X provided that $Z=t_i$.

About joint distribution ε_i it is assumed, that vector ε consists of independent, normally distributed random variables with zero mean:

$$E(\varepsilon_i)=E(X_i)-f(t_i)=E(X|Z=t_i)-E(X|X=t_i)=0$$

Goal – is to determine regression function! How?

Goal: To determine (estimate) $f\ (t)$;

Known: t_i – are not random; ε_i, X_i – are random

Family function: in general, we need to decide (not only theoretical) which class of functions $f\ (t)$ belongs to! This determines k - length of $\boldsymbol{\theta}$.

Strategy: use the most appropriate function class since the function is uniquely determined by the parameters $\boldsymbol{\theta}=(\theta_1,\theta_2,\ldots,\theta_k)$

Idea: based on maximization likelihood function depends on the sample $X=(X_1,X_2,\ldots X_n)$

How?

Likelihood method. The basic assumption

The basic assumption:

a) ε_i – independent, identically distributed

b) distribution $h(x)$ – symmetrical

c) family of distribution has zero mean and unknown variance (Normal, Students, …)

So X_i have density function $h(x - f(t_i))$ and likelihood function is:

$$f(X, \theta_1, \theta_2, \ldots, \theta_k) = \prod_{i=1,n} h(X_i - f(t_i)) = h(\varepsilon_1) \cdot h(\varepsilon_2) \cdot \ldots \cdot h(\varepsilon_k) \to \theta \; max$$

ИНСТИТУТ МАТЕМАТИКИ МЕХАНИКИ КОМПЬЮТЕРНЫХ НАУК

Least Square Methods, case of Normal distributed ε_i

ЮЖНЫЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ

Assumption: $\varepsilon_i \in N(0,\sigma^2)$, ε_i - independent

Likelihood method is connected with Least square method:

$$f(\vec{X},\boldsymbol{\theta}) = \prod \frac{1}{\sqrt{2\pi}\,\sigma}\, exp\left\{-\frac{(X_i - f(t_i))^2}{2\sigma^2}\right\} =$$

$$= \frac{1}{\sigma^n (2\pi)^{n/2}}\, exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - f(t_i))^2\right\} \to max$$

?

achieves at a minimum of the sum squares:

$$\sum_{i=1}^{n}(X_i - f(t_i))^2 = \sum \varepsilon_i^2, \qquad \sum_{i=1}^{n}(X_i - f(t_i))^2 \to min$$

Example: Linear Regression, expierimental data

$$X_i = \theta_1 + t_i\theta_2, \qquad i = 1, \dots, n$$

$$L \equiv \sum \varepsilon_i^2 = \sum (x_i - \theta_1 - t_i\theta_2)^2 \to min$$

Here, unknown parameter θ_1, θ_2, are determined by solving the linear system:

$$\left\{ \frac{\partial L}{\partial \theta_1} = 0, \qquad \frac{\partial L}{\partial \theta_2} = 0 \right\}$$

the points suspicious (suspicious [səsˈpɪʃəs] – подозрительный) of an extremum:

$$\left\{ \hat{\theta}_1 = \bar{X} - \bar{t}\hat{\theta}_2, \qquad \hat{\theta}_2 = \frac{\frac{1}{n}\sum X_i t_i - \bar{X}\,\bar{t}}{\frac{1}{n}\sum (t_i - \bar{t})^2} \right\}$$

var(t)= D(t) = E(t^2) - (E(t))^2

What about $\hat{\theta}_1, \hat{\theta}_2$ for centering data { X_i } and { t_i } ?

Sample correlation coefficient:

$$\rho = \left(\frac{1}{n}\sum (X_i - \bar{X})(t_i - \bar{t})\right) / \sqrt{\frac{1}{n}\sum (t_i - \bar{t})^2 \ \frac{1}{n}(X_i - \bar{X})^2}$$

is the *measure* of linear dependence between $X_1, X_2, \dots X_n$ and $t_1, t_2, \dots t_n$.

A practical approach to regression

Let (ξ, η) – two-dimensional random vector. η – dependent variable, ξ – independent

There are n trials of ξ; value of η is recorded in experiments.

$\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n$ – is the sample of trials (ξ, η); on it's basis is required construct the linear regression.

Linear model without free term ?

$$f = \theta_1 x + \theta_0.$$

Registration of experiment in matrix form

Let's consider experiments in matrix form $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{\theta}$:

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 - \bar{x} \\ 1 & x_2 - \bar{x} \\ \dots & \dots \\ 1 & x_n - \bar{x} \end{pmatrix}, \boldsymbol{Y} = \begin{pmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \dots \\ y_n - \bar{y} \end{pmatrix}, \boldsymbol{\theta} = \begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix},$$

here $\bar{x}$, $\bar{y}$ – empirical average (mean)

Models, for choosen class-function may be present as matrix equation:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\theta}$$

equivalent transformations

ИНСТИТУТ МАТЕМАТИКИ МЕХАНИКИ КОМПЬЮТЕРНЫХ НАУК

Least square solution

ЮЖНЫЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ

$$\boldsymbol{X}'\boldsymbol{Y} = (\boldsymbol{X}'\boldsymbol{X})\boldsymbol{\theta},$$

here

$$\boldsymbol{X}'\boldsymbol{X} = \begin{pmatrix} \sum_i (x_i - \bar{x})^2 & \sum_i (x_i - \bar{x}) \\ \sum_i (x_i - \bar{x}) & n \end{pmatrix}, \boldsymbol{X}'\boldsymbol{Y} = \begin{pmatrix} \sum_i (x_i - \bar{x})(y_i - \bar{y}) \\ \sum_i (y_i - \bar{y}) \end{pmatrix}$$

by using $\sum_i (y_i - \bar{y}) = \sum_i (x_i - \bar{x}) = 0,$

$$\boldsymbol{X}'\boldsymbol{X} = \begin{pmatrix} \sum_i (x_i - \bar{x})^2 & 0 \\ 0 & n \end{pmatrix}, \boldsymbol{X}'\boldsymbol{Y} = \begin{pmatrix} \sum_i (x_i - \bar{x})(y_i - \bar{y}) \\ 0 \end{pmatrix},$$

ИНСТИТУТ
МАТЕМАТИКИ
МЕХАНИКИ
КОМПЬЮТЕРНЫХ
НАУК

Least square estimates of θ

ЮЖНЫЙ
ФЕДЕРАЛЬНЫЙ
УНИВЕРСИТЕТ

Value of parameters follows from equality:

$$\begin{pmatrix} \theta_1 \sum_i (x_i - \bar{x})^2 \\ \theta_0 n \end{pmatrix} = \begin{pmatrix} \sum_i (x_i - \bar{x})(y_i - \bar{y}) \\ 0 \end{pmatrix}$$

if

$$\theta_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{\operatorname{cov}(X, Y)\ \sigma_Y}{\sigma_X^2\ \sigma_Y} = \frac{\rho_{XY}\sigma_Y}{\sigma_X}, \quad \theta_0 = 0.$$

Conclusion: *Regression for centering data has not free term!*

Case of standardized data

Let σ – empirical standard deviation, then

$$\boldsymbol{Y} = \boldsymbol{X\theta}, \ \boldsymbol{\theta} = \begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix},$$

equivalent transformations

$$\boldsymbol{X'Y} = (\boldsymbol{X'X})\boldsymbol{\theta},$$

$$\boldsymbol{X} = \begin{pmatrix} 1 & (x_1 - \bar{x})/\sigma_x \\ 1 & (x_2 - \bar{x})/\sigma_x \\ \dots & \dots \\ 1 & (x_n - \bar{x})/\sigma_x \end{pmatrix}, \boldsymbol{Y} = \begin{pmatrix} (y_1 - \bar{y})/\sigma_y \\ (y_2 - \bar{y})/\sigma_y \\ \dots \\ (y_n - \bar{y})/\sigma_y \end{pmatrix}.$$

Regression coefficients for standardized data are:

$$\theta_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})/(\sigma_x \sigma_y)}{\sum_i (x_i - \bar{x})^2/\sigma_x^2} = \rho_{X,Y}, \quad \theta_0 = 0,$$

here $\rho_{X,Y}$ – coefficients of correlation of X, Y.

Regression coefficients for standardized data are called β–coefficients.

Strength of links. Pairwise regression

Pairwise regression; the basic dispersion relation:

$$\sum_i (y_i - \bar{y})^2 = \sum_i (f(x_i) - \bar{y})^2 + \sum_i (f(x_i) - y_i)^2,$$

or

$$\sigma_y^2 = \sigma_f^2 + \sigma_{res}^2$$

σ_y^2 – $\{y_i\}$ variance ; σ_f^2 – is explained by regression and

σ_{res}^2 – residual between observed in trials and expected according to regression model

Measure (strength links) is coefficient of determination :

$$R^2 = 1 - \frac{\sigma_{res}^2}{\sigma_y^2}$$

The high value (is near one) of R^2 and small σ_{res}^2 is interpreted regression as almost functional dependence.

interpret $[in'tëprit]$

Model evaluation.

Statistical criterion H0 : ρ=0

When do coefficients of regression are equal to coefficients of correlation ?

Coefficients of correlation – function of random variables:

The evaluation of the model do by means statistical criteria:

- $\xi, \eta \in N(0,1)$ Normal distributed $E(\xi) = \mu = 0,\ D(\xi) = 1,$ $E(\eta) = \mu = 0,\ D(\eta) = 1,$
- Null hypotheses: factors is not correlated $(H_0 : \rho = 0)$,
- Criterion statistic : $t_{n-2} = \sqrt{(n-2)}\dfrac{r}{\sqrt{(1-r^2)}}$

  r – empirical coefficients of correlation
- t – has Student distribution with $n-2$ degree of freedom

Multidimensional linear regression. X – m-dimensional factor

Matrixes $\boldsymbol{X}$, $\boldsymbol{\theta}$ convert to next kind:

$$\boldsymbol{X} = \begin{pmatrix} 1 & (x_1^{(1)} - \bar{x}^{(1)})/\sigma_x^{(1)}, \cdots, (x_1^{(m)} - \bar{x}^{(m)})/\sigma_x^{(m)} \\ 1 & (x_2^{(1)} - \bar{x}^{(1)})/\sigma_x^{(1)}, \cdots, (x_2^{(m)} - \bar{x}^{(m)})/\sigma_x^{(m)} \\ \cdots & \cdots \\ 1 & (x_n^{(1)} - \bar{x}^{(1)})/\sigma_x^{(1)}, \cdots, (x_n^{(m)} - \bar{x}^{(m)})/\sigma_x^{(m)} \end{pmatrix},$$

$\boldsymbol{Y}$ – standardaized.

$\theta_0 = 0$, coefficients of regression: $\boldsymbol{\theta} = \dfrac{\boldsymbol{X}'\boldsymbol{Y}}{\boldsymbol{X}'\boldsymbol{X}}$,

For centering data expression of θ_k is equal to

$$\theta_k = \frac{\sum_i (x_i^{(k)} - \bar{x}^{(k)})(y_i - \bar{y})}{\|\boldsymbol{X}'\boldsymbol{X}\|} = \frac{\operatorname{cov}(X^{(k)}, Y)}{\sum_{X,X}},$$

here $\sum_{X,X}$ – matrix of multiple correlation

Model adequacy

Let y – response to $\boldsymbol{X}$ (m–dimensional) vector of factors

$$R^2 = \frac{\frac{1}{n-m-1}\sum_i (f(\boldsymbol{X}_i) - \bar{y})^2}{\frac{1}{n}\sum_i (y_i - \bar{y})^2}$$

Analysis of coefficient of determination imply model modification: **reduce number of factors**, if it needs

The way of factors adding is sequential:

- if value of R^2 at the current step becomes greater then current factor is included to model
- if value of R^2 is unchanged then at then the current variable (factor) is excluded from model

Adequacy of the model is proved by normal distributed residuals with zero mean - *need to check!*

ИНСТИТУТ МАТЕМАТИКИ МЕХАНИКИ КОМПЬЮТЕРНЫХ НАУК

Least square method and normal equations

ЮЖНЫЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ

Let $\vec{Z} = (Z_1, Z_2, \ \dots \ , Z_k)^T$ – random regression factors (object properties or variables),

$\beta = (\beta_1, \beta_2, \dots, \beta_k)^T$ – unknown linear regression parameters and

$E(X|\vec{Z}) = f(\vec{z}) = \beta_1 Z_1 + \dots + \beta_k Z_k$, here $Z^{(i)} = (Z_1^{(i)}, Z_2^{(i)}, \dots, Z_k^{(i)})$ – i-th experiment.

Results of n experiments, n≫ k, with responses $\vec{X} = (X_1, X_2, \dots X_n)^T$

are represented by this system:

$$\begin{cases} X_1 = \beta_1 Z_1^{(1)} + \dots + \beta_k Z_k^{(1)} + \varepsilon_1 \\ X_2 = \beta_1 Z_1^{(2)} + \dots + \beta_k Z_k^{(2)} + \varepsilon_2 \\ \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \\ X_n = \beta_1 Z_1^{(n)} + \dots + \beta_k Z_k^{(n)} + \varepsilon_n \end{cases},$$

ε_i - *they are implicitly represented in the system, are contained in the experimental data*

it is form of multivariate regression or by matrix notations :

$$\vec{X} = \vec{Z} * \vec{\beta} + \vec{\varepsilon} \qquad (*)$$

$$Z = \begin{pmatrix} Z_1^{(1)} & \dots & Z_k^{(1)} \\ & \dots & \\ Z_1^{(n)} & \dots & Z_k^{(n)} \end{pmatrix} \cdot \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \equiv \begin{pmatrix} \vec{Z}^{(1)} \\ \vdots \\ \vec{Z}^{(n)} \end{pmatrix}$$

Lemma 1. On the positivity of the matrix of the system

Lemma1

If Z has rank k, i.e. all column independent then A = $Z^T * Z$ – positive matrix (p.m.).

Proof

- According to definition of (p.m.)

$$\vec{t}^T At \geq 0 \text{ for } \forall \vec{t} = (t_1, t_2, \ldots, t_n) \in R^n \text{ and } \vec{t}^T At = 0 \quad \Leftrightarrow \vec{t} = \vec{0}$$

[as *for* $u = (u_1, \ldots, u_n)^T$, $||\vec{u}||^2 = \vec{u}^T \cdot \vec{u} = \sum u_i^2 \geq 0$, $\quad ||\vec{u}||^2 = 0 \Leftrightarrow \vec{u} = \vec{0}$].

- A – symmetrical,

$$A^T = A: \ \vec{t}^T A \vec{t} = \vec{t}^T Z^T Z \vec{t} = (Z\vec{t})^T Z\vec{t} = \|Z\vec{t}\|^2 \geq 0 \text{ and } \|Z\vec{t}\|^2 = \vec{0} \quad \text{if} \quad Z\vec{t} = \vec{0},$$

but rank (Z)=k $\Rightarrow \vec{t} = \vec{0}$, this is contradiction because of $\vec{t}$ – an arbitrary vector, $\vec{t} \in R^n$, so A – positive matrix.

ИНСТИТУТ МАТЕМАТИКИ МЕХАНИКИ КОМПЬЮТЕРНЫХ НАУК

Lemma 2. About the root factorization of the positive matrix

ЮЖНЫЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ

Positive definiteness and symmetry of matrix A imply the existence of $\sqrt{A}$ is a real symmetric matrix such that $\sqrt{A}\,\sqrt{A} = A$. **Lemma2**

Proof

For any positive and symmetrical matrix factorization A=$Q^T D Q$ exists, here D – diagonal matrix (with eigenvalues on diagonal, eigenvalues > 0) and Q – orthogonal (consists of eigenvectors) .

$$A = Q^T \sqrt{D}\sqrt{D}\ Q = \left(\sqrt{D}\,Q\right)^T \sqrt{D}\,Q \quad \Rightarrow \quad \sqrt{A} = \sqrt{D}\,Q\,,$$

proven.

Determination of the β -coefficient based on the optimization problem (matrix form)

Let's find least squares estimation $\hat{\beta} = \vec{\beta}^*$, satisfying the following optimization problem (minimal solution exists):

$$S(\vec{\beta}) \rightarrow \min, \qquad (1)$$

here

$$S(\vec{\beta}) = \sum_{i=1}^{n} \varepsilon_i^2 = ||\vec{\varepsilon}||^2 = \left|\left|\vec{X} - Z\,\vec{\beta}\right|\right|^2 = \left(\vec{X} - Z\,\vec{\beta}\right)^T \left(\vec{X} - Z\,\vec{\beta}\right)$$

There are two approaches to find $\hat{\beta}$:

1. solving system $\left\{\frac{\partial S(\vec{\beta})}{\partial \beta_1} = 0, \dots, \frac{\partial S(\vec{\beta})}{\partial \beta_k} = 0\right\}$ to find extremum points - *considered before*
2. $S(\vec{\beta})$ – square of distance between points $\vec{X} \in R^n$ and $Z\vec{\beta}$, $(\cdot)Z\vec{\beta} \in hyperplane$ where $\forall Z\vec{t}, (\vec{t} \in R^k)$ lies.

$S(\vec{\beta}^*)$ – minimal distance, because of vector $\vec{X} - z\hat{\beta}$ is orthogonal to all vectors of hyper plane $Z\vec{t}\,(\forall \vec{t} \in R^k)$,

so $\left(Z\vec{t}, \vec{X} - Z\hat{\beta}\right) = (Z\vec{t})^T\left(\vec{X} - Z\hat{\beta}\right) = \vec{t}^T\left(Z^T\vec{X} - Z^T Z\hat{\beta}\right) = 0$ and for any $\vec{t}^T \neq 0$, f.e. basis vector $\vec{t}^T = (0\ 0\ 0\ 1\ 0 \dots 0) \in R^k$

$\Rightarrow$ obtain *simple transformations*

$$Z^T\vec{X} - Z^T Z\hat{\beta} = 0 \quad \Rightarrow \quad Z^T Z\hat{\beta} = Z^T\vec{X}, \quad A\hat{\beta} = Z^T\vec{X} \qquad (2)$$

so least squares estimation – is solution of (2) According to Lemma 1 they have the same single solution

$$\hat{\beta} = \vec{\beta}^* = A^{-1} Z^T \vec{X} \qquad (3)$$

If rectangle matrix Z size (n,k) has rank = k, $k \leq n$, both equation (2) is called normal equations.

ИНСТИТУТ МАТЕМАТИКИ МЕХАНИКИ КОМПЬЮТЕРНЫХ НАУК

Least square estimation properties

ЮЖНЫЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ

Conclusion

If $\vec{\varepsilon}$ consists of independent r.v. $\forall \vec{\varepsilon}_i \in N(0,\sigma^2) \Rightarrow$ least squares estimation is the same as likelihood estimation for $\vec{\sigma}^2 = \frac{1}{n}\sum \hat{\varepsilon}_i^2 = \frac{1}{n}||\vec{X} - Z\hat{\beta}||^2 = \frac{1}{n}S(\hat{\beta})$

Properties of the least square estimation

1 $\hat{\beta} - \vec{\beta} = A^{-1}Z^T\vec{\varepsilon}$ (4)

proof:

Substitute (3) in (4) and use (*) and equality $A = Z^TZ$, can obtain $A^{-1}Z^T\vec{X} - \vec{\beta} =$

$= A^{-1}Z^T(Z * \vec{\beta} + \vec{\varepsilon}) - \vec{\beta} = A^{-1}A * \vec{\beta} + A^{-1}Z^T\vec{\varepsilon} - \vec{\beta} = A^{-1}Z^T\varepsilon,$ *proven*

2 If $E\vec{\varepsilon} = 0$ then $\hat{\beta}$ – unbiased estimation for $\vec{\beta}$.

proof:

$E\hat{\beta} =^{(p.1.)} \vec{\beta} + A^{-1}Z^TE\vec{\varepsilon} = \vec{\beta}.$ *proven*

ИНСТИТУТ МАТЕМАТИКИ МЕХАНИКИ КОМПЬЮТЕРНЫХ НАУК

Important assumptions

ЮЖНЫЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ

Matrix Z has rank=k and all columns of Z linear independent.

I. Vector $\vec{\varepsilon}$ consists of independent random values $\in N(0, \sigma^2)$.

Recall for any $\vec{x}$: $D\vec{x} = E(\vec{x} - E\vec{x})(\vec{x} - E\vec{x})^T$ – covariance matrix;

II. $\mathrm{cov}(x_i, x_j) = E(x_i - Ex_i)(x_j - Ex_j)$ and $D\vec{\varepsilon} = \sigma^2 E_n$;

$E_n \equiv eye(n)$ – identity matrix size n, ML notation

Multivariate Model Statistics

(3) Let I and II assumptions are true, then

$\sqrt{A} * \hat{\beta}$ has covariance matrix of diagonal type and equal to $\sigma^2 E_k$,

$(\sqrt{A} * \hat{\beta} = \sigma^2 E_k)$; it means that coordinates of $\sqrt{A}\hat{\beta}$ – uncorrelated

without proof

Theorem Let I and II assumptions are true, then

1) Vector $\frac{1}{\sigma}\sqrt{A}(\hat{\beta} - \vec{\beta})$ has k-dimensional normal standard distribution (consists of k independent random variables $\in N(0,1)$)

2) $n\hat{\sigma}^2/\sigma^2 = \|\vec{X} - Z\hat{\beta}\|^2/\sigma^2$ has χ^2_{n-k} distribution and doesn't depends on $\hat{\beta}$

(Interesting result!)

3) $(\sigma^2)^* = n\hat{\sigma}^2 = \frac{1}{n-k}||\vec{X} - Z\hat{\beta}||^2$ – unbiased estimation for σ^2

without proof

ИНСТИТУТ
МАТЕМАТИКИ
МЕХАНИКИ
КОМПЬЮТЕРНЫХ
НАУК

ЮЖНЫЙ
ФЕДЕРАЛЬНЫЙ
УНИВЕРСИТЕТ

Thank you for your attention!

DON'T FORGET TO DREAM - IT WILL INCREASE YOUR HEALTH!