Regresní analýza dat
01REAN - Cvičení 03

Jiri Franc

Czech Technical University
Faculty of Nuclear Sciences and Physical Engineering
Department of Mathematics

Outline of todays exercises:

- Summary of theory for simple regression.
- Ordinary least squares method.
- Linear regression in R.

Regression

Regression is a method for studying the relationship between a response variable Y and explanatory variables X.

General, we assume that there are the basic probability space (Ω, A, P) and $p, q, n \in N$ than the model

$$Y_i = g(X_i, \beta^0) + \varepsilon_i \qquad i = 1 \dots n$$

is called the general regression model and

- n is the sample size (number of observations).
- $g(X_i, \beta^0)$ is a smooth model function, $g : \mathbb{R}^q \times \mathbb{R}^p \to \mathbb{R}$.
- Y_i is called response variable, dependent variable.
- X_i's are called explanatory variables (independent variables, predictor variables, regressors, factor, carrier, features, etc.) and R^q is sometimes called factor space. ,
- β^0 is a vector of true regression parameters (in the linear regression model also called regression coefficients),
- ε_i is called disturbance (error term, fluctuation, etc.) and it represents unexplained variation in the dependent variable.

One way to summarize the relationship between X and Y is through the regression function $r(x) = \mathbb{E}(Y|X = x) = \int yf(y|x)dy$.

Linear Regression Model

Let $p, q \in \mathbb{N}$, $q = p$, and $g(X_i, \beta^0) = X_i^T \beta^0 \ \forall i = 1 \dots n$, then:

The linear regression model is the model

$$Y_i = X_{i1}\beta_1^0 + X_{i2}\beta_2^0 + \dots + X_{ip}\beta_p^0 + \varepsilon_i = X_i^T \beta^0 + \varepsilon_i, \qquad i = 1 \dots n$$

We can rewrite the previous definition with n equations in matrix notation.

$$\underbrace{\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}}_{\mathbf{Y}} = \underbrace{\begin{pmatrix} X_{1,1} & X_{1,2} & \dots & X_{1,p} \\ X_{2,1} & X_{2,2} & \dots & X_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n,1} & X_{n,2} & \dots & X_{n,n} \end{pmatrix}}_{\mathbf{X}} \cdot \underbrace{\begin{pmatrix} \beta_1^0 \\ \beta_2^0 \\ \vdots \\ \beta_p^0 \end{pmatrix}}_{\beta^0} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{\boldsymbol{\varepsilon}}.$$

Equivalently,

$$\mathbf{Y} = \mathbf{X}\beta^0 + \varepsilon, \tag{1}$$

where $\mathbf{Y}$ is an $n \times 1$ vector of response variables, $\mathbf{X}$ is an $n \times p$ matrix of predictors, β^0 is a $p \times 1$ vector of unknown coefficients and ε is an $n \times 1$ vector of unknown errors.

Simple Linear Regression

The simplest version of regression is when X_i is simple (a scalar not a vector) and $r(x)$ is assumed to be linear $r(x) = \beta_1 + \beta_2 x$ than:

The **simple linear regression model** (**Straight-line regression**) is defined as

$$Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i,$$

where $\mathbb{E}(\varepsilon_i | X_i) = 0$ and $\mathbb{V}(\varepsilon_i | X_i) = \sigma^2$.

The unknown parameters in the model are the intercept β_1, the slope β_2 and the variance σ^2.

Note:

$Y_i = \beta_1 + \beta_2 \ln(X_i) + \varepsilon_i$, or $Y_i = \beta_1 + \beta_2 (X_i)^2 + \varepsilon_i$ are still linear models.

Estimation

Let $\hat{\beta}_1$ and $\hat{\beta}_2$ denote estimates of β_1 and β_2, and $\hat{\sigma}$ estimate of σ.

The predicted values or fitted values are

$$\hat{Y}_i = \beta_1 + \beta_2 X_i.$$

The fitted line is defined to be

$$\hat{r}(x) = \beta_1 + \beta_2 x.$$

Residuals are defined to be

$$r_i = \hat{\varepsilon}_i = Y_i - \hat{Y}_i = Y_i - \beta_1 + \beta_2 X_i.$$

The residual sums of squares or RSS is defined by

$$RSS = \sum_{i=1}^{n} r_i^2.$$

Ordinary Least Squares estimation

The ordinary least squares estimator (OLS) denoted by $\hat{\beta}^{(OLS,n)}$ minimizes RSS and is given by

$$\hat{\beta}^{(OLS,n)} = \arg\min_{\beta \in R^p} \sum_{i=1}^{n} (Y_i - X_i^T \beta)^2 = \arg\min_{\beta \in R^p} (\mathbf{Y} - \mathbf{X}\beta)^T (\mathbf{Y} - \mathbf{X}\beta).$$

The solution of OLS always exists, we can differentiate previous relation with respect to β and put equal zero to obtain the system of equations which is called *normal equations*.

$$\frac{\partial \sum_{i=1}^{n} (Y_i - X_i^T \beta)^2}{\partial \beta_l} = -2 \sum_{i=1}^{n} (Y_i - X_i^T \beta) X_{il} = 0 \quad \text{for } l = 1 \ldots p.$$

Equivalently in matrix form

$$\mathbf{X}^T (\mathbf{Y} - \mathbf{X}\beta) = 0.$$

Ordinary Least Squares estimation

Assume that the design matrix $\mathbf{X}$ has full rank. Than the ordinary least squares estimator exists and $\hat{\beta}^{(OLS,n)}$ is given by formula

$$\hat{\beta}^{(OLS,n)} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

In our simple linear regression problem, we can obtain:

$$\begin{aligned} \hat{\beta}_2 &= \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}, \\ \hat{\beta}_1 &= \bar{Y} - \hat{\beta}_2\bar{X}. \end{aligned}$$

Estimation of variance

An unbiased estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n-1-p}\sum_{i=1}^{n} r_i^2.$$

Variance of estimated coefficients:

$$\begin{aligned}\mathbb{V}(\hat{\beta}_2) &= \sigma^2\frac{1}{\sum_{i=1}^{n}(X_i-\bar{X})^2},\\ \mathbb{V}(\hat{\beta}_1) &= \sigma^2\frac{\sum_{i=1}^{n}X_i^2}{n\sum_{i=1}^{n}(X_i-\bar{X})^2}\end{aligned}$$

lm - Formula Call in R

Syntax	Model	Comments
Y ~ A	$Y = \beta_o + \beta_1 A$	Straight-line with an implicit y-intercept
Y ~ -1 + A	$Y = \beta_1 A$	Straight-line with no y-intercept; that is, a fit forced through (0,0)
Y ~ A + I(A^2)	$Y = \beta_o + \beta_1 A + \beta_2 A^2$	Polynomial model; note that the identity function **I( )** allows terms in the model to include normal mathematical symbols.
Y ~ A + B	$Y = \beta_o + \beta_1 A + \beta_2 B$	A first-order model in A and B without interaction terms.
Y ~ A:B	$Y = \beta_o + \beta_1 AB$	A model containing only first-order interactions between A and B.
Y ~ A*B	$Y = \beta_o + \beta_1 A + \beta_2 B + \beta_3 AB$	A full first-order model with a term; an equivalent code is Y ~ A + B + A:B.
Y ~ (A + B + C)^2	$Y = \beta_o + \beta_1 A + \beta_2 B + \beta_3 C + \beta_4 AB + \beta_5 AC + \beta_6 AC$	A model including all first-order effects and interactions up to the n^{th} order, where n is given by **( )^n**. An equivalent code in this case is Y ~ A*B*C – A:B:C.

Úkoly:

- Study enclosed R code.
- Solve problems described at the end of the code.