DESCRIPTIVE STATISTICS

Dr Alina Gleska

Institute of Mathematics, PUT

April 22, 2018

1 Regression analysis

Two-dimensional data

Statistical observation – the pair (X, Y).

Types of analysis:

- correlation – we define the shape, the direction and the strength of relationships;
- regression – we define the mathematical function between correlated variables.

Regression analysis

Regression analysis – a set of statistical processes for estimating the relationships among variables.

Regression function – a mathematical function of the independent variable X.

We distinguish two types of dependence:

- functional - for each value of X we have only one value of Y;
- statistical (stochastic) - for each value of X we have many values of Y.

The statistical dependence

The statistical dependence we can write as:

$$Y = f(X) + \varepsilon,$$

where ε – a random error.

The regression equation (the regression model) – the equation describing the relationships among variables after adding a random error.

Regression

Types of regression:

$Y = f(X) + \varepsilon$ – a simple regression (a regression model with a single explanatory variable);

$Y = f(X_1, \ldots, X_n) + \varepsilon$ – a multiple regression;

X – a reason – the independent variable;

Y – a result, an effect – the dependent variable.

Types of simple regression:

- linear – the best fitted function is a linear function;
- nonlinear – the best fitted function is a nonlinear function (for example an exponential function or a logarithmic function).

The choice of the regression model

The regression model is chosen according to the dispersion of the empirical data on their scatterplot.

REMARK! In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. These model parameters can be easily interpreted – this is the high advantage of linear models!

A linear regression model

$$Y = a + bX + \varepsilon,$$

X – the independent variable (or the explanatory variable, or the predictor),

Y – the dependent variable (or the response variable, or the regressand),

a – the intercept,

b – the slope of the regression model,

ε – a random error.

The slope b – it gives the expected change of Y when the value of variable X increases by one unit.

REMARK! If $b = 0$, there is no relationship among variables.

A linear regression model

LP	Values of X	Values of Y
1	x_1	y_1
2	x_2	y_2
$\vdots$	$\vdots$	$\vdots$
n	x_n	y_n

$$y_i = a + bx_i + \varepsilon_i,$$

x_i, y_i - the empirical values of the variable X and Y
ε_i - the error terms (or the disturbance terms, or noise) (that is, if there were no error Y would be a deterministic linear function of X).

The error terms

The error terms – the deviations of empirical values y_i from the predicted values $\hat{y}_i$ appearing in the linear regression model:

$$e_i = y_i - \hat{y}_i$$

Property:

$$\sum_{i=1}^{n} e_i = 0.$$

Model Assumption: the random error component is independent of the X component.
The linear regression equation:

$$\hat{y}_i = a + bx_i$$

The linear regression equation

$$\hat{y}_i = a + bx_i$$

We can find a and b using the ordinary least squares (OLS) method. The linear least squares is a method for estimating the unknown parameters in a linear regression model. OLS chooses the parameters of a linear function of a set of explanatory variables by minimizing the sum of the squares of the differences between the observed dependent variable (values of the variable being predicted) in the given dataset and those predicted by the linear function:

$$S(a,b) = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \rightarrow \textit{minimize}$$

Geometrically, this is seen as the sum of the squared distances, parallel to the axis of the dependent variable, between each data point in the set and the corresponding point on the regression line – the smaller the differences, the better the model fits the data. The resulting estimator can be expressed by a simple formula, especially in the case of a single regressor on the right-hand side.

$$b = \frac{cov(X, Y)}{S_x^2}, \qquad a = \overline{y} - b \cdot \overline{x}$$

EXAMPLE

Regression validation

The goodness of fit of a statistical model describes how well it fits a set of observations. Measures of goodness of fit typically summarize the discrepancy between observed values and the values expected under the model in question.

Measures of regression validation:

- the variance of residuals: $s_e^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- the standard deviation of residuals: $s_e = \sqrt{s_e^2}$
- the coefficient of random variation: $V_e = \frac{s_e}{\overline{y}} \cdot 100\%$
- the coefficient of determination: $R^2 = r^2 \cdot 100\%$
  (where r - the Pearson's correlation coefficient)
- the coefficient of indetermination: $\varphi^2 = 100\% - R^2$

The coefficient of random variation V_e

The coefficient of random variation V_e defines the proportion of the mean prediction error to the mean of the variable Y.

Properties:

- belongs to the interval $[0\%, 100\%]$;
- the smaller value, the better fitting to the empirical data.

The coefficient of determination R^2

One useful aspect of regression is that it can divide the variation of Y into two parts: the variation of the predicted scores $\hat{y}_i - \overline{y}$ and the variation in the errors of prediction $y_i - \hat{y}_i$.

The coefficient of determination R^2

The deviations:

$$y_i - \overline{y} = (\hat{y}_i - \overline{y}) + (y_i - \hat{y}_i)$$

Sum of squares:

$$\begin{aligned} \sum_{i=1}^{n}(y_i - \overline{y})^2 &= \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ SST &= SSR + SSE \end{aligned}$$

The coefficient of determination R^2

SST – total sum of squares – the sum of the squared deviations of Y from the mean of Y;

SSR – sum of squares for regression (or the sum of squares predicted) – the sum of the squared deviations of the predicted scores from the mean;

SSE – sum of squares for errors – the sum of the squared errors of predictions.

The coefficient of determination defines the proportion of the explained variation of the Y to the total variation:

$$R^2 = \frac{SSR}{SST} = r^2.$$

The coefficient of determination R^2

Properties:

- belongs to the interval $[0\%, 100\%]$;
- the closer 100%, the better fitting to the empirical data:
  $R^2 = 100\%$ – the whole variation of Y is explained by the variation of X;
  $R^2 = 0\%$ – the whole variation of Y is random.

The coefficient of indetermination φ^2

The coefficient of indetermination φ^2 – defines the proportion of the unexplained variation of the Y to the total variation:

$$\varphi^2 = \frac{SSE}{SST} = 100\% - R^2.$$

Properties:

- belongs to the interval $[0\%, 100\%]$;
- the closer 0%, the better fitting to the empirical data:
  $\varphi^2 = 0\%$ – the perfect fitting;
  $\varphi^2 = 100\%$ – the whole variation of Y is random.