Probability and Statistics

Raluca Muresan

West University of Timisoara

raluca.muresan@e-uvt.ro

Overview

We will study **relations between variables**.

The type of their relation is called (mathematically) **regression**.

Establishing and testing such a relation enables us:

- to understand **interactions, causes, and effects** among variables;
- to **predict** unobserved variables based on the observed ones;
- to determine which variables **significantly affect** the variable of interest.

Regression models relate a **response variable** to one or several **predictors**.

Definition

Response or dependent variable Y is a variable of interest that we predict based on one or several predictors.
Predictors or independent variables $X_1, \ldots, X_k$ are used to predict the values and behavior of the response variable Y .
Regression of Y on $X_1, X_2 \ldots, X_k$ is the conditional expectation,

$$G(x_1, x_2, \ldots, x_k) = E(Y|X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k)$$

Example. House prices depend on house area, number of bedrooms and bathrooms, the backyard area, the average income of the neighborhood, etc.

Method of least squares I

How do we estimate the regression function G that connects response variable Y with predictors $X_1, X_2 \ldots, X_k$?

First we focus on **univariate regression** predicting response Y based on one predictor X. The method will be extended to k predictors.

We first observe the pairs of observed values

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

For accurate forecasting, we are looking for the function $\hat{G}(x)$ that passes **as close as possible** to the observed data points.

Method of least squares minimizes the sum of squared distances.

Method of least squares II

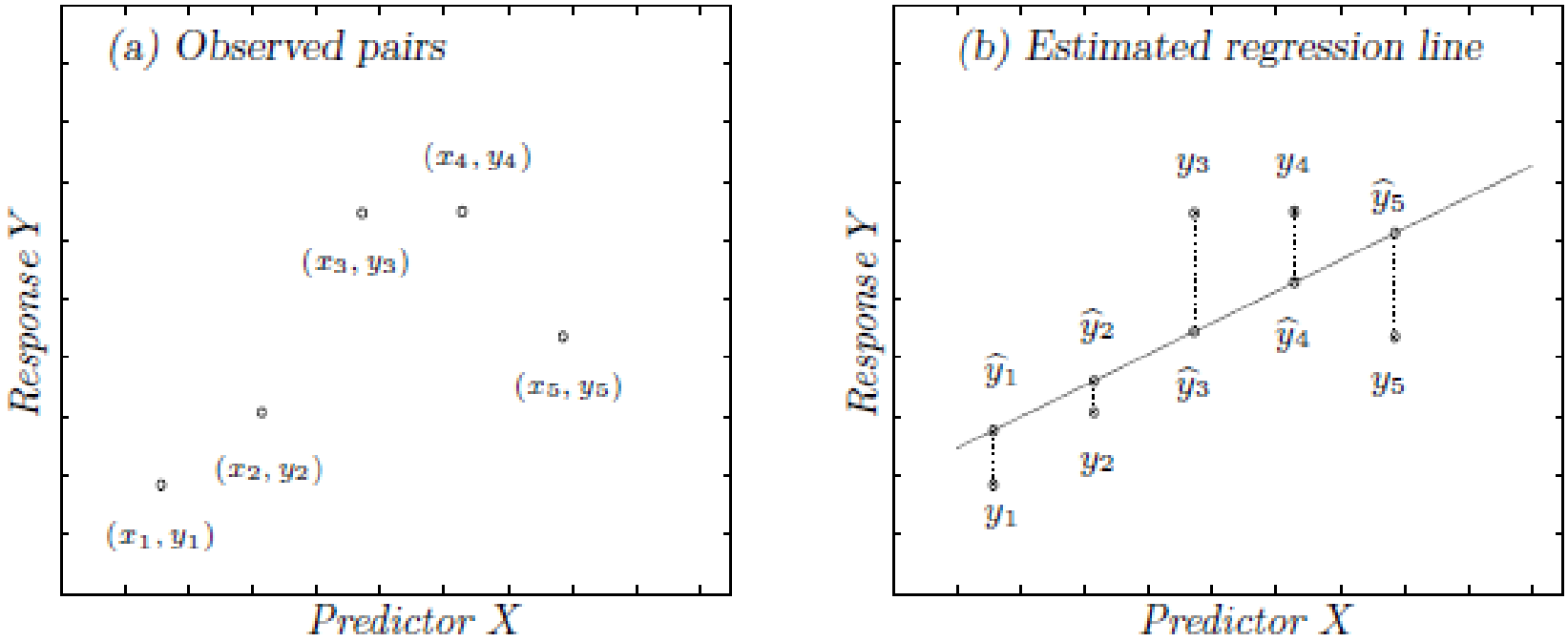

FIGURE 11.3: *Least squares estimation of the regression line.*

Method of least squares III

Definition

Residuals of the regression model

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

are differences between observed responses y_i and their fitted values $\hat{y}_i = \hat{G}(x_i)$.

Definition

Method of least squares finds a regression function $\hat{G}(x)$ that minimizes the sum of squared residuals

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Linear regression

Linear regression model assumes that the conditional expectation

$$G(x) = E(Y|X = x) = \beta_0 + \beta_1 x$$

is a linear function of x.

The slope

$$\beta_1 = G(x + 1) - G(x)$$

is the predicted change in the response variable when predictor changes by 1.

Estimation of parameters

Let us **estimate** the slope and intercept by method of least squares.

We have to minimize the sum of squared residuals

$$Q = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

We compute the partial derivatives and we equate them to 0:

$$\begin{cases} \frac{\partial Q}{\partial \beta_0} = 0 \\ \frac{\partial Q}{\partial \beta_1} = 0 \end{cases}$$

We have the system of normal equations:

$$\begin{cases} \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0 \\ \sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) = 0 \end{cases}$$

Regression estimates

$$b_0 = \widehat{\beta}_0 = \bar{y} - b_1\,\bar{x}$$

$$b_1 = \widehat{\beta}_1 = S_{xy}/S_{xx}$$

where

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

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Example 1. (World population). According to the International Data Base of the U.S. Census Bureau, population of the world grows according to Table 11.1. How can we use these data to predict the world population in years 2015 and 2020?

Year	Population mln. people	Year	Population mln. people	Year	Population mln. people
1950	2558	1975	4089	2000	6090
1955	2782	1980	4451	2005	6474
1960	3043	1985	4855	2010	6864
1965	3350	1990	5287	2015	?
1970	3712	1995	5700	2020	?

TABLE 11.1: *Population of the world, 1950–2020.*

Regression and correlation I

The correlation coefficient of the random variables X and Y is

$$\rho = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

where $Cov(X, Y) = E[(X - E(X))(Y - E(Y))]$.

It measures the **direction and strength** of a linear relationship between variables X and Y .

We can estimate ρ by the **sample correlation coefficient**:

$$r = \frac{s_{xy}}{s_x s_y}$$

Regression and correlation II

where s_{xy} is the sample covariance

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

and s_x, s_y are the sample standard deviations

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

The values of r are in $[-1, 1]$.

ANOVA and R^2 I

Analysis of variance (ANOVA) explores **variation** among the observed responses.

A portion of this variation can be explained by **predictors**. The rest is attributed to **"error"**.

The **total variation** among observed responses is measured by the **total sum of squares**

$$SS_{tot} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

ANOVA and R^2 II

A portion of this total variation is attributed to predictor X and the regression model, measured by the **regression sum of squares**

$$SS_{reg} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

The rest of total variation is attributed to "error". It is measured by the **error sum of squares**:

$$SS_{err} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

ANOVA and R^2 III

Proposition

We have

$$SS_{tot} = SS_{reg} + SS_{err}$$

Definition

The coefficient of determination R^2 is the proportion of the total variation explained by the model,

$$R^2 = \frac{SS_{reg}}{SS_{tot}}.$$

It is always between 0 and 1, with high values generally suggesting a good fit.

In univariate regression, R-square also equals the squared sample correlation coefficient

$$R^2 = r^2.$$

ANOVA and R^2 IV

Example 2. Compute SS_{tot}, SS_{reg}, SS_{err} and R^2 for the data in Example 1. Observing the value of R^2, what can we say about the model?

T tests

For further analysis, we introduce standard **regression assumptions**.

We will assume that observed responses y_i are independent Normal random variables with mean $E(Y_i) = \beta_0 + \beta_1 x_i$ and constant variance σ^2. Predictors x_i are considered non-random.

As a consequence, regression estimates b_0 and b_1 have Normal distribution. After we estimate the variance σ^2, they can be studied by T-tests and T-intervals.

Regression variance is

$$s^2 = \frac{SS_{err}}{n-2}$$

which estimates $\sigma^2 = Var(Y)$ unbiasedly.

The **ANOVA table**:

Univariate ANOVA

Source	Sum of squares	Degrees of freedom	Mean squares	F
Model	$SS_{\text{REG}} = \sum(\hat{y}_i - \bar{y})^2$	1	$MS_{\text{REG}} = SS_{\text{REG}}$	$\frac{MS_{\text{REG}}}{MS_{\text{ERR}}}$
Error	$SS_{\text{ERR}} = \sum(y_i - \hat{y}_i)^2$	$n-2$	$MS_{\text{ERR}} = \frac{SS_{\text{ERR}}}{n-2}$	
Total	$SS_{\text{TOT}} = \sum(y_i - \bar{y})^2$	$n-1$		

We see that the sample regression variance is the mean squared error,

$$s^2 = MS_{ERR}$$

.

Testing hypotheses about β_1:

$$H_0 : \beta_1 = 0$$

$$H_A : \beta_1 \neq 0$$

The test statistic:

$$T = \frac{b_1}{\frac{s}{\sqrt{S_{xx}}}}$$

has a Student t distribution with $n - 2$ degrees of freedom.

The F-ratio

$$F = \frac{MS_{REG}}{MS_{ERR}}$$

is used to test significance of the entire regression model and has a F-distribution with $df_{REG} = 1$ and $df_{ERR} = n - 2$ degrees of freedom.

ANOVA F-test is always one-sided and right-tail because only large values of the F-statistic show a large portion of explained variation and the overall significance of the model.

For univariate regression, the F test and t test are the same.

Example 3. (Efficiency of computer programs). A computer manager needs to know how efficiency of her new computer program depends on the size of incoming data. Efficiency will be measured by the number of processed requests per hour. Applying the program to data sets of different sizes, she gets the following results:

Data size (gigabytes), x	6	7	7	8	10	10	15
Processed requests, y	40	55	50	41	17	26	16

a) Estimate the regression line. b) Compute the ANOVA table and estimate the variance of the model. c) Perform the t test for the slope. d) Perform the F test for the model.

The End