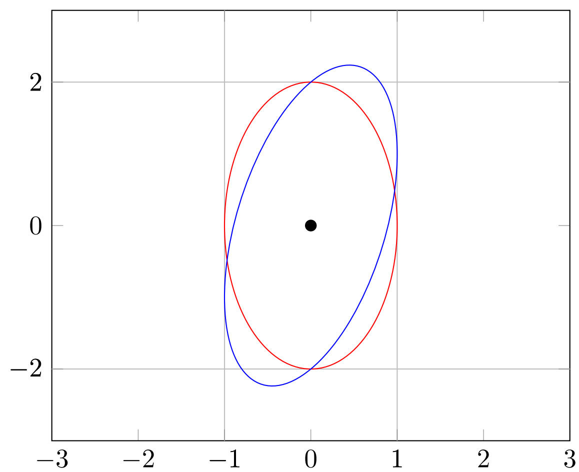
Contents

Linear Regression

TDung

1 Introduction



Linear regression is used to predict value of a variable based on the value of another one. The variable using to predict the other one is called *Independent variable* (or sometimes, it might be called *predictor variable*).The predicted variable is called *dependent variable* (or sometimes, it might be called *the outcome variable*).

For example, you can use linear regression to predict the population of a country based on the population of previous years or maybe you can predict a person's height depends on their weight...

In case we have more than one independent variables, we have to use multiple regression.

2 Ordinary Least Squares Estimation

The *simple linear regression* model consists of the *mean function* and the *variance function.*

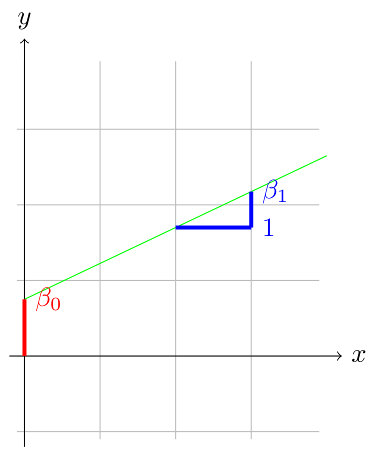
$$E = (Y|X = x) = \beta_0 + \beta_1 x \tag{1}$$

$$Var(Y|X = x) = \sigma^2 \tag{2}$$

This is a line where y is the output variable which we want to predict, X is the input variable we know and β_0 and β_1 are coefficients that we need to estimate that move the line around.

The parameters in the *mean function* (1) are the intercept β_0 and the slope β_1. The intercept is the value of $E(Y|X = x)$ when x equals to zero (*In machine learning, it is called the bias, because it is added to offset all predictions*) and the slope is define of change in $E(Y|X = x)$ for a unit change in X. The parameters are usually unknown and must be estimated by using data. By changing the parameters, we can get all possible straight lines.

The *variance function* in (2) is assumed to be a constant, with a positive value σ^2 that is usually unknown. When $\sigma^2 > 0$, the observed value of the i^{th} response y_i will typically not equal to its expected value $E(Y|X = x_i)$.



Equation of a straight line $E(Y|X = x) = \beta_0 + \beta_1 x$

We have many methods have been suggested for obtaining estimates of parameters in a model. And the method we discuss here is called *ordinary least squares* (or *OLS*), in which parameter estimates are chosen to minimize a quantity called the residual sum of squares.

Parameters are unknown quantities that characterize a model. Estimates of parameters are computable functions of data and are therefore statistics. Estimates of parameters are denoted by putting a "hat" over the corresponding Greek letter. For example with $\hat{\beta}_i$, read "beta i^{th} hat" is the estimator of β_i, and $\hat{\sigma}^2$ is the estimator of σ^2. The *fitted value* for case i is given by $\hat{E}(Y|X = x_i)$ and we use the shorthand notation y like this:

$$\hat{y}_i = \hat{E}(Y|X = x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{3}$$

All least squares computations for simple regression just depend only on averages, sums of squares and sums of cross-products. Definitions of the quantities used are given in **Table 2.1**. Sums of squares and cross-products have been centered by subtracting the average from each of the values before squaring or taking cross-products.

Alternative formulas for computing the corrected sums of squares and cross products from uncorrected sums of squares and cross products but they can be highly inaccurate when used on a computer and should be avoided.

Table 2.1 also lists definitions for the usual univariate and bivariate summary statistics, the sample averages ($\overline{x}$, $\overline{y}$), the sample variances (SD_x^2, SD_y^2) and estimated covariance and correlation (s_{xy}, r_{xy}). The "hat" rule described would suggest that different symbols should be used for these quantities; for example, p_{xy} might be more appropriate for the sample correlation if the population correlation is $\hat{p}_{xy}$.

Table 2.1: Definitions of Symbols

Quantity	Definition	Description
$\overline{x}$	$\sum\left(\frac{x_i}{n}\right)$	Average of x
$\overline{y}$	$\sum\left(\frac{y_i}{n}\right)$	Average of y
SXX	$\sum(x_i - \overline{x})^2$	Sum of squares for the x
SYY	$\sum(y_i - \overline{y})^2$	Sum of squares for the y
SXY	$\sum(x_i - \overline{x})(y_i - \overline{y})$	Sum of cross-products
SD_x	$\sqrt{\frac{SXX}{n-1}}$	Sample standard deviation of the x's
SD_y	$\sqrt{\frac{SYY}{n-1}}$	Sample standard deviation of the y's
s_{xy}	$\frac{SXY}{n-1}$	Sample covariance
r_{xy}	$\frac{s_{xy}}{SD_x SD_y}$	Sample correlation

The symbol $\sum$ means to add over all the values or pairs of values in the data.

This inconsistency is deliberate since in many regression situations, these statistics are not estimates of population parameters.

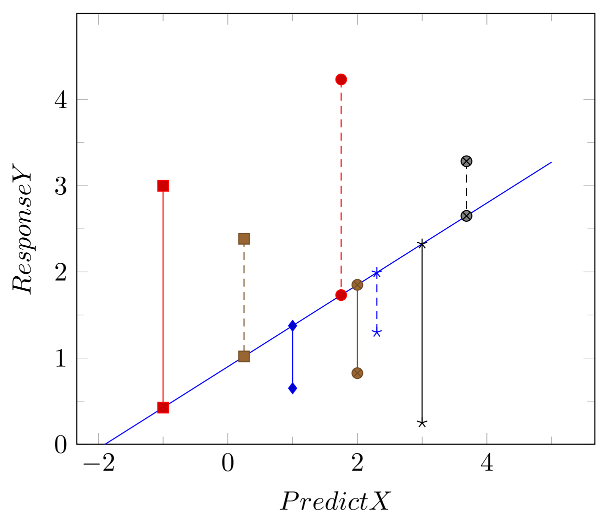
To illustrate computations, we will use Forbes' data, page 4, for which $n = 17$. The data are given in **Table 2.2**. In our analysis of these data, the response will be taken to be $Lpres = 100log10(Pressure)$, and the predictor is Temp.

Table 2.2: Forbes' 1857 Data on Boiling Point and Barometric Pressure for 17 Locations in the Alps and Scotland

Case	Temp(F)	Pressure (Inches Hg)	$Lpres = 100log(Pressure)$
1	194.5	20.79	131.79
2	194.3	20.79	131.79
3	197.9	22.40	135.02
4	198.4	22.67	135.55
5	199.4	23.15	136.46
6	199.9	23.35	136.83
7	200.9	23.89	137.82
8	201.1	23.99	138.00
9	201.4	24.02	138.06
10	201.3	24.01	138.04
11	203.6	25.14	140.04
12	204.6	26.57	142.44
13	209.5	28.49	145.47
14	208.6	27.76	144.34
15	210.7	29.04	146.30
16	211.9	29.88	147.54
17	212.2	30.06	147.80

Forbes' data were collected at 17 selected locations, so the sample variance of boiling points, $SD_x^2 = 33.17$, is not an estimate of any meaningful population variance. Similarly, r_{xy} depends as much on the method of sampling as it does on the population value p_{xy}, should such a population value make sense.

3 Least squares criterion



A schematic plot for ols fitting. Each data point is indicated by a small symbol.Points below the line have negative residuals, while points above the line have positive residuals.

The criterion function for obtaining estimators is based on the residuals. The residuals reflect the inherent asymmetry in the roles of the response and the predictor in regression problems.

The ols estimators are those values β_0 and β_1 that minimize the function.

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{N} |y_i - (\beta_0 + \beta_1 x_i)|^2 \tag{4}$$

When evaluated at $\left(\hat{\beta}_0, \hat{\beta}_1\right)$, we call the quantity $\mathrm{RSS}\left(\hat{\beta}_0, \hat{\beta}_1\right)$ the *residual sum of squares*, or just RSS.

The least squares estimates can be derived in many ways, one of which is outlined in Appendix A.3. They are given by the expressions

$$\hat{\beta}_1 = \frac{SXY}{SXX} = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} = r_{xy}\frac{SD_x}{SD_y} \tag{5}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \tag{6}$$

The several forms for $\hat{\beta}_1$ are all equivalent.

Using Forbes' data, we will write $\overline{x}$ to be the sample mean of *Temp* and $\overline{y}$ to be the sample mean of *Lpres*. The quantities needed for computing the least squares estimators are:

$$\begin{array}{lll} \overline{x} = 202.95294 & \mathrm{SXX} = 530.78235 & \mathrm{SXY} = 475.31224 \\ \overline{y} = 139.60529 & \mathrm{SYY} = 427.79402 & \end{array}$$

In case regression calculations are not done by using statistical software or a statistical calculator, intermediate calculations such as these should be done as accurately as possible, and rounding should be done only to final results. Using the results already mentioned, we can find

$$\hat{\beta}_1 = \frac{SXY}{SXX} = 0.895$$
$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = -42.138$$

The estimated line, given by either of the equations

$$\hat{E}(Lpres|Temp) = -42.138 + 0.895Temp$$

4 Estimating σ^2

Although the e_i are not parameters in the usual sense but we will use the hat notation to specify the residuals: the residual for the i^{th} case, denoted $\hat{e}_i$, is given by the equation

$$\hat{e_i} = \hat{y_i} - \hat{\beta}_0 + \hat{\beta}_1 \tag{7}$$

which compare with the equation for the statistical errors

$$e_i = y_i - \beta_0 + \beta_1 \tag{8}$$

5 About Forbes' Data

In an 1857 article, a Scottish physicist named James D. Forbes discussed a series of experiments that he had done concerning the relationship between atmospheric pressure and the boiling point of water. Forbes knew that altitude could be determined from atmospheric pressure, measured with a barometer, with lower pressures corresponding to higher altitudes.

In the middle of the nineteenth century, barometers were fragile instruments, and Forbes wondered if a simpler measurement of the boiling point of water could substitute for a direct reading of barometric pressure. Forbes collected data from n = 17 locations in the Alps and in Scotland. He measured at each location pressure in inches of mercury with a barometer and boiling point in degrees Fahrenheit. Let's take a look at the scatter plot. Here is the scatter plot. Of course we have to load the data first. After plotting the data, we add the best-fitting OLS line to the plot. This is the straight line that best fits the data according to the Ordinary Least Squares criterion, which we shall discuss in detail later.

Copied from James H. Steiger - Department of Psychology and Human Development Vanderbilt University

Read more on: $https://en.wikipedia.org/wiki/Simple_linear_regression$