Contents

1 Introduction **2**

2 Ordinary Least Squares Estimation **2**

3 Least squares criterion **5**

4 Estimating σ^2 **6**

5 Appendix A3 **6**

6 About Forbes' Data **7**

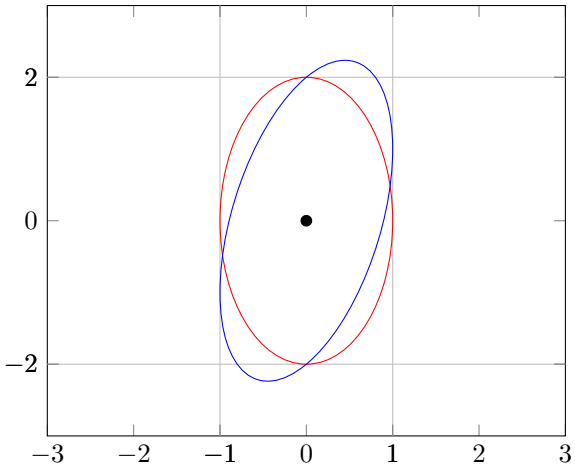
7 Linear Regression's Estimation Methods **7**

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 variable, we have to use multiple regression.

2 Ordinary Least Squares Estimation

Simple linear regression model includes the *mean function*(1) and the *variance function*(2).

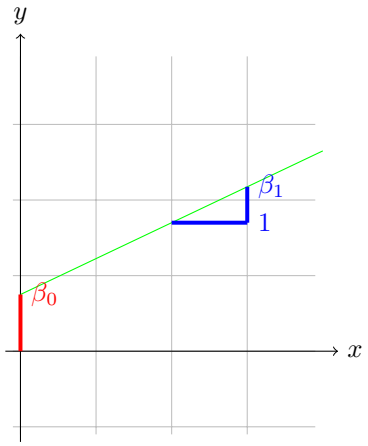
$$E = (Y|X = x) = \beta_0 + \beta_1 x \tag{1}$$

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

In this line, Y is the output variable we want to predict, X is the input variable we know, β_0 and β_1 are the coefficients we need to estimate that can 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 defined by 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* (2) is assumed to be 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 $E(Y|X = x) = \beta_0 + \beta_1 x$

There are many methods for obtaining estimates of parameters in a model. The method we are discussing 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, $\hat{\beta}_i$, read as "beta i^{th} hat", is the estimate of β_i, and $\hat{\sigma}^2$ is the estimate 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 used quantities are given in **Table 2.1**. Sums of squares and cross-products have been centered by subtracting the averages from each of the values before squaring or taking cross-products.

There are alternative formulas for computing the correct sums of squares and cross-products from the incorrect one. However, they can be highly inaccurate when being used on computers 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 the estimated covariance and correlation (s_{xy}, r_{xy}). The "hat" rule described above suggests 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	$\sum\left(x_i-\overline{x}\right)^2$	Sum of squares for x
SYY	$\sum\left(y_i-\overline{y}\right)^2$	Sum of squares for 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 x's
SD_y	$\sqrt{\frac{SYY}{n-1}}$	Sample standard deviation of y's
s_{xy}	$\frac{SXY}{n-1}$	Sample covariance
r_{xy}	$\frac{s_{xy}}{SD_xSD_y}$	Sample correlation

The symbol Σ means to add over all the values or pairs of values in the data.

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

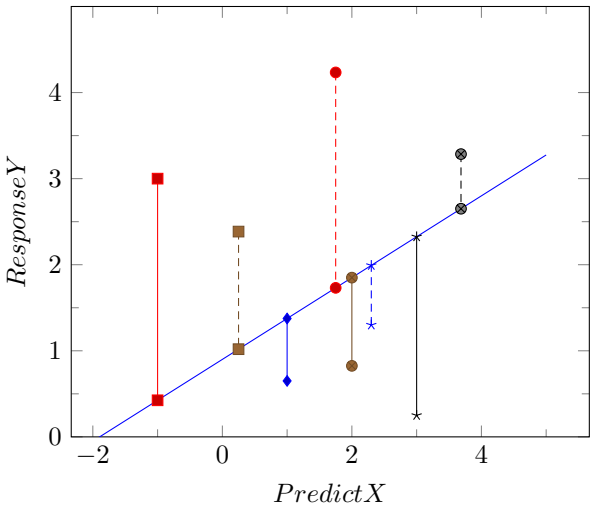
To illustrate computations, we will use the Forbes' data given in **Table 2.2**. In our analysis of these data, the response will be taken to be $Lpres = 100log10\ (Pressure)$ and the predictor is Temp.

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. Likewise, r_{xy} depends as much on the method of sampling as it does on the population value p_{xy}, should with 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 estimates 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 estimates are those values β_0 and β_1 that minimize the function.

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

We call the quantity $\text{RSS}\left(\hat{\beta}_0, \hat{\beta}_1\right)$ the *residual sum of squares*, or just RSS.

The least squares estimates can be derived in many ways, one of which is

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

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

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

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

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

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 only be used for final results. Using the above-mentioned results, we can find

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

The estimated line, given by both of the equations

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

4 Estimating σ^2

Although e_i are not usually parameters, 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 compares with the equation for the statistical errors

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

For simple regression, residual $df = n - 2$ (with $n = 17$, the number of data we are using; df = number of cases minus the number of parameters in the mean function and df is denoted from *degrees of freedom*), so the estimate of σ^2 is given by

$$\sigma^2 = \frac{RSS}{n-2} \tag{9}$$

Using summaries from the given results ($\overline{x}, \overline{y}, ...$), we find

$$RSS = SYY - \hat{\beta}_1^2 SXX = 2.15491$$
$$\sigma^2 = \frac{RSS}{n-2} = 0.14366$$

The square root of σ^2, $\sigma = \sqrt{0.14366} = 0.37903$, is often called the standard error of regression. It's in the same units as is the response variable.

5 Appendix A3

The ols estimates of values β_0 and β_1 that minimize the function.

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{N} |y_i - (\beta_0 + \beta_1 x_i)|^2 \tag{10}$$

The method of finding the minimize is differentiate with respect to β_0 and β_1, and make it equal 0.

$$\frac{\partial RSS(\beta_0, \beta_1)}{\beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \tag{11}$$

$$\frac{\partial RSS(\beta_0, \beta_1)}{\beta_1} = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0 \tag{12}$$

From (11), (12) we get (13) and (14) like this:

$$\beta_0 n + \beta_1 \sum x_i = \sum y_i \tag{13}$$

$$\beta_0 \sum x_i + \beta_1 \sum x_i^2 = \sum x_i y_i \tag{14}$$

Using fomulars:

$$SXX = \Sigma(x_i - \overline{x})^2 = \Sigma x_i^2 - n\overline{x}^2 \tag{15}$$

$$SXY = \Sigma(x_i - \overline{x})(y_i - \overline{y}) = \Sigma x_i y_i - n\overline{xy} \tag{16}$$

Solving (13), (14), we get:

$$\hat{\beta_0} = \overline{y} - \beta_1 \overline{x}, \tag{17}$$

$$\hat{\beta_1} = \frac{SXY}{SXX} \tag{18}$$

6 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 the 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, so Forbes wondered if a simpler measurement of the boiling point of water could work as a substitute for a direct reading of barometric pressure. He collected data from 17 locations in the Alps and Scotland. He measured 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.

7 Linear Regression's Estimation Methods

These are some of the most common estimations for linear regression:

Least-squares estimation.
Maximum-likelihood estimation.
Bayesian linear regression.
Quantile regression.
Mixed models.
Principal component regression.
Least-angle regression.
Theil–Sen estimator.

References: James H. Steiger - Department of Psychology and Human Development Vanderbilt University
Further reading on:
https : //en.wikipedia.org/wiki/Simple_linear_regression
https : //en.wikipedia.org/wiki/Linear_regression