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 used 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 The model behind linear regression

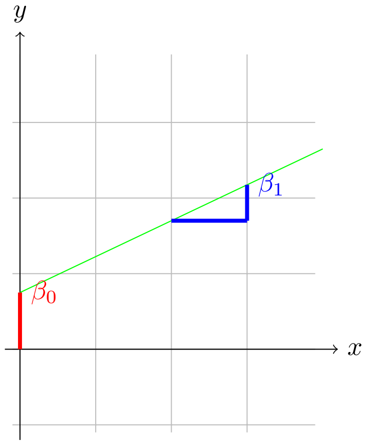
Simple linear regression model includes the *mean function*(1) and the *variance function*(2).

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

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

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.

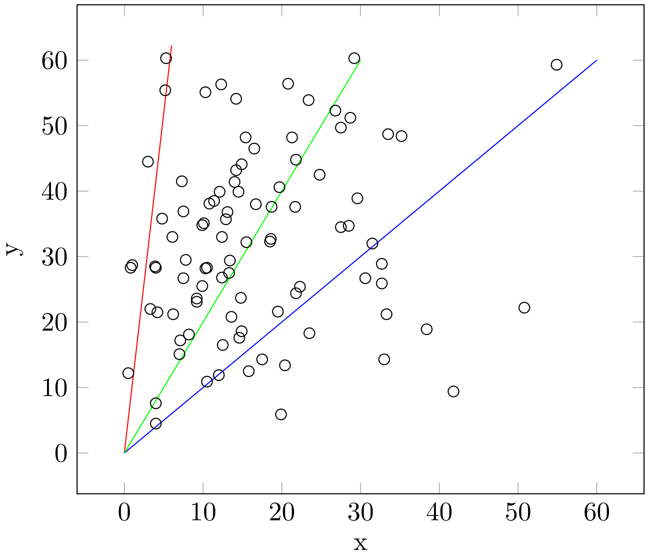
The parameters in the *mean function* (1) are the intercept β_0 and the slope β_1. The intercept is the value of $E(Y|X = x)$ when x equals to zero (*in machine learning, it is called the bias, because it is added to offset all predictions*) and the slope is defined by the change in $E(Y|X = x)$ for 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.



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

In fact, we can not accurately determine the parameter β_0 or β_1 because it is unknown but only estimated by using data. By changing the parameters, we can get all possible straight lines. That is why the general equation we have a certain error. Therefore, we will process and get results from data to the overall estimate.

The simplest way to explain is to suppose we have a data set with x and y values that make a lot of points on the graph, and through these points, we can plot many linear lines to show the relationship linearity between x and y, but we can only choose **one** line that best represents of this relationship, corresponding to the fact that we have only make a single linear regression equation.



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 not equal to its expected value $E(Y|X = x_i)$. To explain the difference between the observed data and the expected value, we have to make a quantitative called a statistical error or e_i= y_i - $E(Y|X = x_i)$.

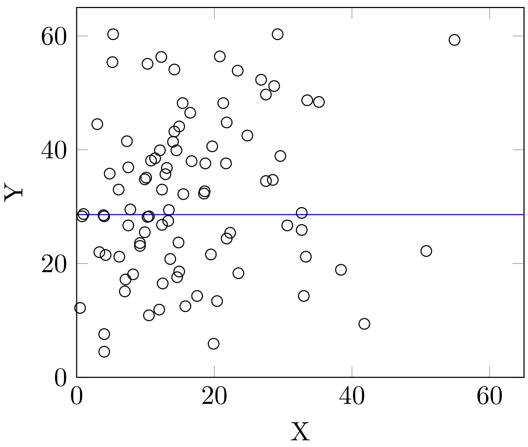
The errors e_i depend on unknown parameters in the mean function and so are not observable quantities.

3 Statistical Hypotheses

If we make assumptions when the chief null hypothesis is H_0: $B_1 = 0$ (i.e the equation $E(Y|X = x)$ is parallel to the horizontal axis) and the corresponding alternative hypothesis is H_1: $B_1 = 0$ (i.e the equation $E(Y|X = x)$ is not parallel to the horizontal axis or H_1 is negative. by H_0). If H_0 is true, we can conclude that for every variable x, it has no effect on Y.

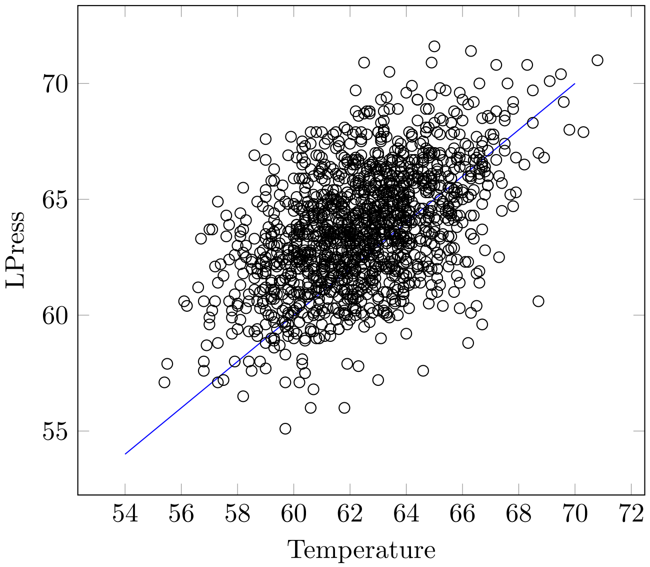
For **Figure 3.1**, we might expect the straight-line to be appropriate but with $B_1 = 0$. If the slope is zero, then the line is parallel to the horizontal axis. We will test for independence of x and y by testing the hypothesis that $\beta_1 = 0$ against the alternative hypothesis that $\beta_1 = 0$.

Figure 3.1



Sometimes it is reasonable to choose a different null hypothesis for B_1. For example, if x is some **gold standard** for a particular measurement, i.e., a best-quality measurement often involving great expense, and y is some cheaper substitute, then the obvious null hypothesis is $B_1 = 1$ with alternative $B_1! = 1$.

Figure 3.2



For Figure 3.2, which the blue line we assume the parameter $\beta_0 = 0$ and $\beta_1 = 1$

4 Simple linear regression

Table 4.1: 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.

5 Regression calculation

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 (RSS)***.

The table 5.1 will show all formulas that we use to calculate the two parameters: β_0 and β_1

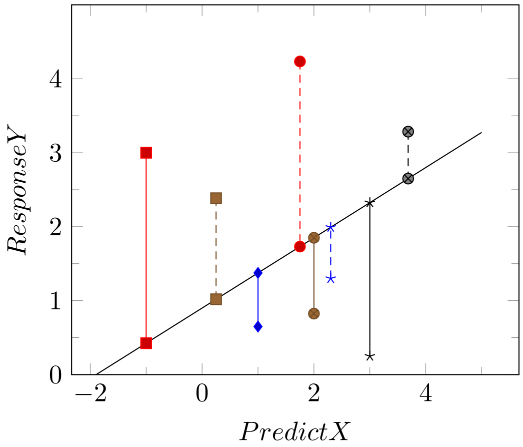
Table 5.1: Definitions of Symbols

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 x
SYY	$\sum(y_i - \overline{y})^2$	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}	$\frac{s_{xy}}{SD_x SD_y}$	Sample correlation

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

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.

Figure 5.2



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 OLS estimates are those values β_0 and β_1 that minimize the function. We call the quantity $\mathrm{RSS}\left(\hat{\beta}_0, \hat{\beta}_1\right)$ the *residual sum of squares* or just RSS.

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{N} |y_i - (\beta_0 + \beta_1 x_i)|^2 \tag{3}$$

One of method to find the minimum is differentiate with respect to β_0 and β_1, and make them equal 0.

$$\frac{\partial RSS(\beta_0, \beta_1)}{\beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0 \tag{4}$$

$$\frac{\partial RSS(\beta_0, \beta_1)}{\beta_1} = -2\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) = 0 \tag{5}$$

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

$$\beta_0 n + \beta_1 \sum x_i = \sum y_i \tag{6}$$

$$\beta_0 \sum x_i + \beta_1 \sum x_i^2 = \sum x_i y_i \tag{7}$$

Using fomulars:

$$
\begin{aligned}
SXX &= \Sigma(x_i - \overline{x})^2 \qquad (8)\\
&= \Sigma x_i^2 - 2\Sigma x_i \overline{x} + n\overline{x}^2\\
&= \Sigma x_i^2 - 2\Sigma x_i \frac{\Sigma x_i}{n} + n\frac{\Sigma x_i \Sigma x_i}{n^2}\\
&= \Sigma x_i^2 - \frac{\Sigma x_i \Sigma x_i}{n}\\
&= \Sigma x_i^2 - n\frac{\Sigma x_i \Sigma x_i}{nn}\\
&= \Sigma x_i^2 - n\overline{x}^2\\
SXY &= \Sigma(x_i - \overline{x})(y_i - \overline{y}) = \Sigma x_i y_i - n\overline{xy} \qquad (9)
\end{aligned}
$$

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

$$
\begin{aligned}
&\beta_0 n + \beta_1 \Sigma x_i = \Sigma y_i\\
&\beta_0 = \frac{\Sigma y}{n} - \beta_1 \frac{\Sigma x}{n}\\
&\beta_0 = \overline{y} - \beta_1 \overline{x} \qquad (10)
\end{aligned}
$$

$$
\begin{aligned}
&\beta_0 \Sigma x_i + \beta_1 \Sigma x_i^2 = \Sigma x_i y_i\\
&\beta_1 = \frac{\Sigma x_i y_i - \beta_0 \Sigma x_i}{\Sigma x_i^2}\\
&\beta_1 = \left(\frac{\Sigma x_i y_i}{\Sigma x_i^2} - \frac{\Sigma x_i \Sigma y_i}{n\Sigma x_i^2}\right)\frac{n\Sigma x_i^2}{n\Sigma x_i^2 - \Sigma x_i \Sigma x_i}\\
&\beta_1 = \left(\Sigma x_i y_i - \frac{\Sigma x_i \Sigma y_i}{n}\right)\frac{n}{n\Sigma x_i^2 - \Sigma x_i \Sigma x_i}\\
&\beta_1 = \frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{n\Sigma x_i^2 - \Sigma x_i \Sigma x_i}\\
&\beta_1 = \frac{\Sigma x_i y_i - n\overline{xy}}{n\Sigma x_i^2 - n\overline{x}^2}\\
&\beta_1 = \frac{SXY}{SXX} = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} = r_{xy}\frac{SD_x}{SD_y} \qquad (11)
\end{aligned}
$$

The several forms for $\hat{\beta_1}$ are all equivalent.

Using Forbes' data from **Table 5.1**, 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:

$$
\begin{array}{lll}
\overline{x} = 202.95294118 & SXX = 530.78235294 & SXY = 475.31223529\\
\overline{y} = 139.60529412 & SYY = 427.79402353 &
\end{array}
$$

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

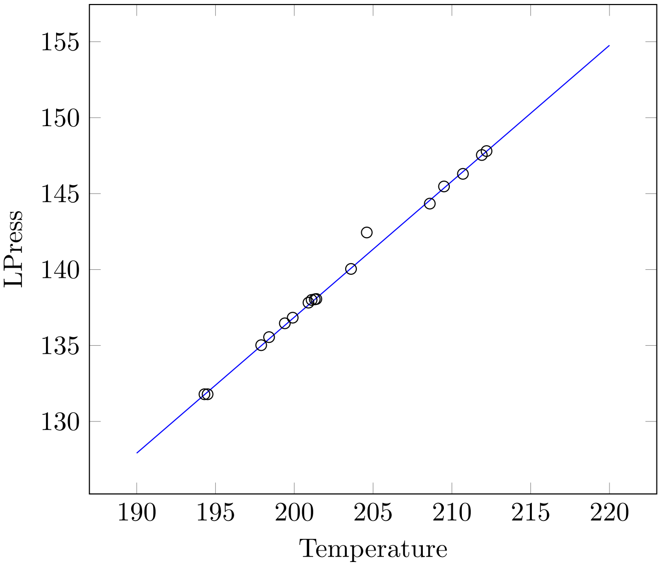
$$
\begin{aligned}
\hat{\beta_1} &= \frac{SXY}{SXX} = 0.895\\
\hat{\beta_0} &= \overline{y} - \hat{\beta_1}\overline{x} = -42.138
\end{aligned}
$$

The estimated line, given by both of the equations

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

Figure 5.3 below shows the straight-line of the mean function after calculating parameter B1 and B2 using the formulas in Table 5.1

Figure 5.3: The straight-line which has minimize residual



6 Interpreting Regression Coefficients

Firstly, we need to analysis the correlation between x and y by using the correlation analyze method with the main formula is the correlation coefficient. The result of the formula just evaluates whether or not there is a linear relationship, whether this relationship is positive or negative, stable or not.

The formula of the correlation coefficient is given from the Co-variance formula - which is also a method of evaluating the relationship between two variables x and y - but has a major limitation for which Co-variance formula is often rarely used and instead by the correlation coefficient. Specifically:

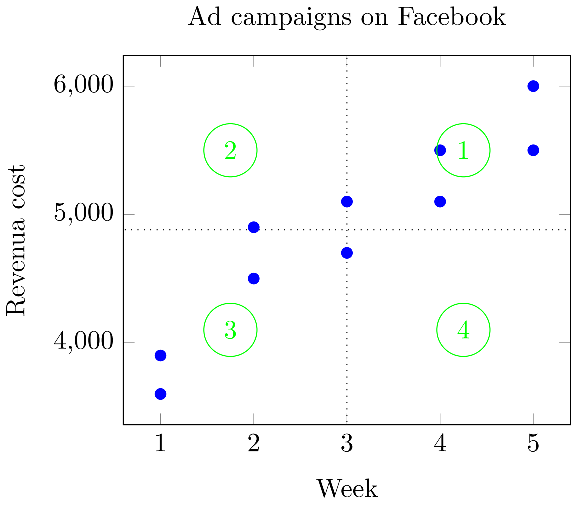
$$s_{xy} = \frac{\Sigma(x_i - \overline{x})(y_i - \overline{y})}{n-1}$$

For example, we have a sample data about the number of ad campaigns on Facebook per week for 10 weeks and how many million revenue cost (units of 1000 VND) are received, where x will be the number of campaigns per week, y is the sales

Table 6.1: the number of ad campaigns on Facebook per week for 10 weeks

Week	Ad campaign	revenue cost
1	2	4900
2	3	5100
3	5	5500
4	4	5100
5	1	3900
6	1	3600
7	3	4700
8	4	5500
9	2	4500
10	5	6000

Next, we will apply the above-mentioned formula to calculate the co-variance is 1022. From there, we will divide the graph into 4 parts.



Interpret the results:

The line in which value $x = 4$ is the average value over 10 weeks, the y line of 10-week average revenue. This line divides the graph into four parts. Part 1 will have x value greater than average of x, y greater than mean of y. Part 2 will have x value less than the mean of x and value of y greater than the mean of y. Similar to parts 3 and 4.

Therefore, we can see (x - $\overline{x}$)(y - $\overline{y}$) will be positive with parts 1 and 3 but negative with parts 2 and 4.

- If S_{xy} is positive, the points are mostly in parts 1 and 3, a positive linear relationship.
- If S_{xy} is negative, the points are mostly in parts 2 and 4, a negative linear relationship.
- If the points are completely divided equally into 4 parts then S_{xy} will be 0 and there is no linear relationship between x and y.

Looking on the graph we can see that 7/10 points are in parts 1 and 3 corresponding to positive S. From there, we can confirm that x and y have a linear relationship, which means that the advertising campaign increases, revenue also increases.

7 Residual Checking

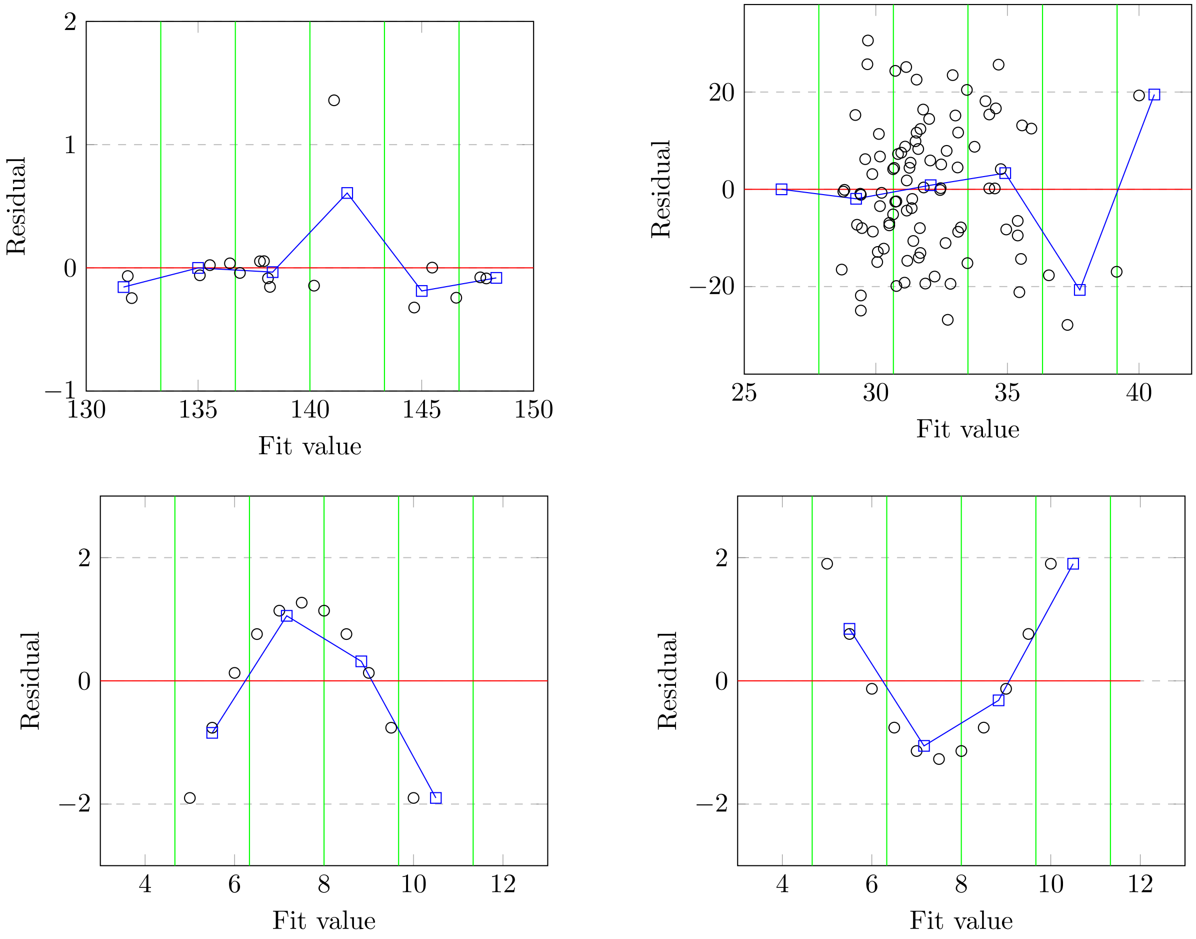


Figure 7.1 Sample residual versus fit plots for testing equal variance

A plot of all residuals on the y-axis versus. the predicted values on the x-axis, called a residual versus fit plot, is a good way to check the linearity and equal variance assumptions. The fixed-x assumption can not be checked with residual analysis (or any other data analysis).

To analyze a residual vs. fit plot, such as any of the examples shown in **Figure 7.1**, you should mentally divide it up into about 5 to 10 vertical stripes (in here we will divide it into 5 vertical stripes). Then each stripe represents all of the residuals for a number of subjects who have a similar predicted values.

To check the linearity assumption, consider that for each x value, if the mean of Y falls on a straight line, then the residuals have a mean of **zero**. If we incorrectly fit a straight line to a curve, then some or most of the predicted means are incorrect, and this causes the residuals for at least specific ranges of x (or the predicated Y) to be **non-zero** on average. Specifically if the data follow a simple curve, we will tend to have either a pattern of high then low then high residuals or the reverse. So the technique used to detect non-linearity in a residual vs. fit plot is to find the (vertical) mean of the residuals for each vertical stripe.

If the resultant connected segments or curve is close to a horizontal line at 0 on the y-axis, then we have no reason to doubt the linearity assumption. If there is a clear curve, most commonly a “smile” or “frown” shape, then we suspect non-linearity.

Four examples are shown in figure 9.4. In each band the mean residual is marked, and lines segments connect these. Plots A and B show no obvious pattern away from a horizontal line other that the small amount of expected “noise”. Plots C and D show clear deviations from normality, because the lines connecting the mean residuals of the vertical bands show a clear frown (C) and smile (D) pattern, rather than a flat line.