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

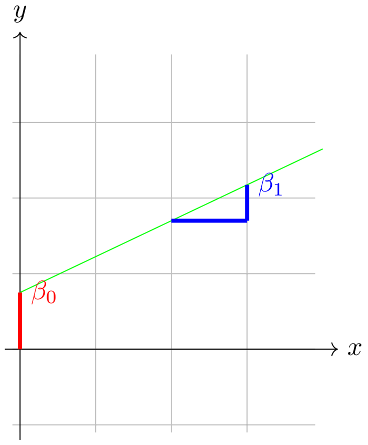
Simple linear regression model includes the *mean function*(1) and the *variance function*(2).

$$E = (Y|X = x) = \beta_0 + \beta_1 x \tag{2.1}$$

$$Var(Y|X = x) = \sigma^2 \tag{2.2}$$

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.

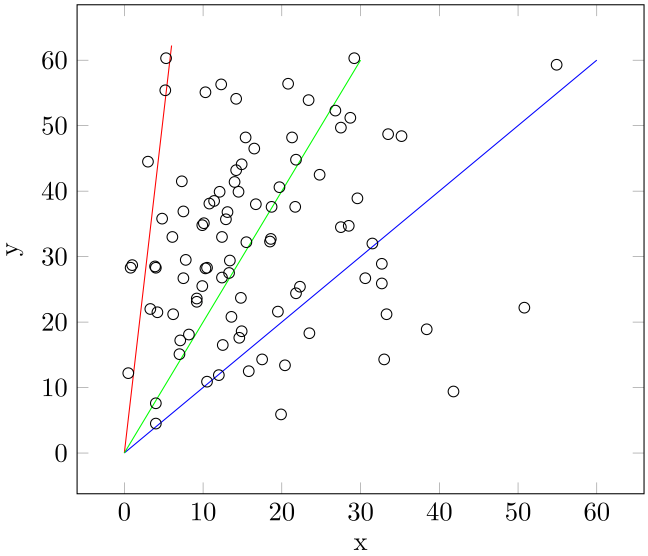
The parameters in the *mean function* (1) are the intercept β_0 and the slope β_1. The intercept is the value of $E(Y|X = x)$ when x equals to zero (*in machine learning, it is called the bias, because it is added to offset all predictions*) and the slope is defined by the change in $E(Y|X = x)$ for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

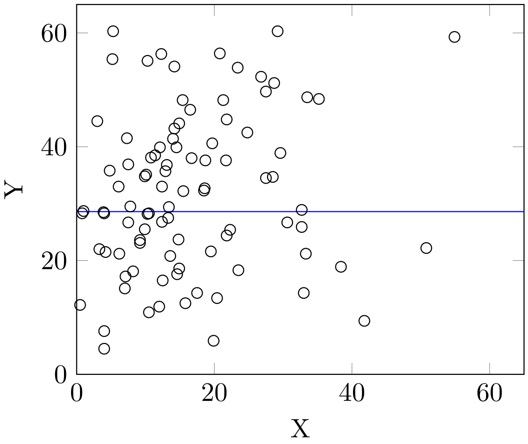
The errors e_i depend on unknown parameters in the mean function and so are not observable quantities.

3 Statistical Hypotheses

If we make assumptions when the chief null hypothesis is H_0: $\beta_1 = 0$ (i.e the equation $E(Y|X = x)$ is parallel to the horizontal axis) and the corresponding alternative hypothesis is H_1: $\beta_1 = 0$ (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.

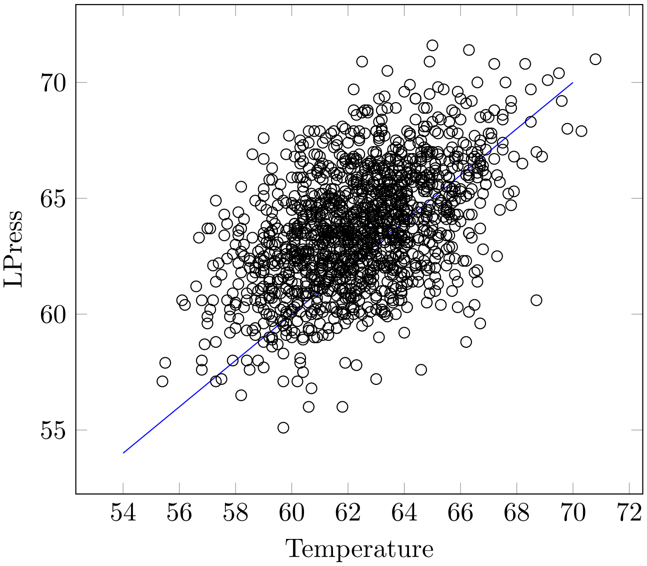
For **Figure 3.1**, we might expect the straight-line to be appropriate but with $\beta_1 = 0$. If the slope is zero, then the line is parallel to the horizontal axis. We will test for independence of x and y by testing the hypothesis that $\beta_1 = 0$ against the alternative hypothesis that $\beta_1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

Figure 3.1



Sometimes it is reasonable to choose a different null hypothesis for β_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 $\beta_1 = 1$ with alternative $\beta_1 \neq 1$.

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

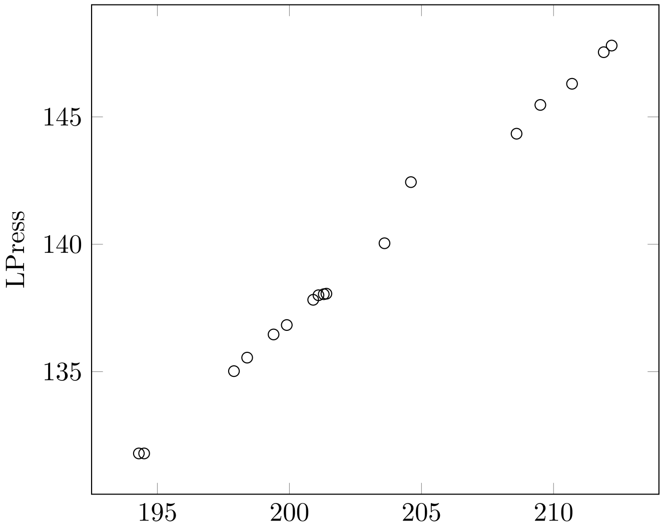
Table 4.1: Forbes' 1857 Data on Boiling Point and Barometric Pressure for 17 Locations in the Alps and Scotland

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

The data consists of 17 pairs of numbers corresponding to observed boiling point and corrected barometric pressure, at locations in the Alps.

Based on Table 4.1, we can make a scatterplot.

Figure 4.2 : Scatterplot of Forbes data



We want to use EDA (explanatory data analysis) to check that the assumptions are reasonable before trying a regression analysis. We can see that the assumptions of linearity seems plausible because we can imagine a straight line from bottom left to top right going through the center of the points. Also the assumption of equal spread is plausible because for any narrow range of temperature (horizontally), the spread of LPress value (vertically) is fairly similar. The assumption of Normality is something that human beings can not test by looking at a scatterplot.But there were only two possible outcomes in the whole experiment, we could reject the idea that the distribution of LPress is Normal at each Temperature level.

The assumption of fixed-x cannot be seen in the data.

The assumption of independent error is usually not visible in the data and must be judged by the way the experiment was run.

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

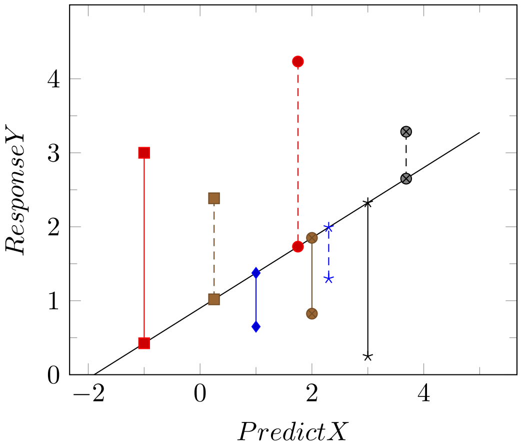
Table 5.1: Definitions of Symbols

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

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

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

$$\frac{\partial RSS(\beta_0, \beta_1)}{\beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0 \tag{5.1}$$

$$\frac{\partial RSS(\beta_0, \beta_1)}{\beta_1} = -2\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) = 0 \tag{5.2}$$

From (5.1), (5.2) we get:

$$\beta_0 n + \beta_1 \sum x_i = \sum y_i \tag{5.3}$$

$$\beta_0 \sum x_i + \beta_1 \sum x_i^2 = \sum x_i y_i \tag{5.4}$$

Using fomulars:

$$\begin{aligned} SXX &= \Sigma(x_i - \overline{x})^2 \qquad (5.5)\\ &= \Sigma x_i^2 - 2\Sigma x_i \overline{x} + n\overline{x}^2 \\ &= \Sigma x_i^2 - \frac{\Sigma x_i \Sigma x_i}{n} \\ &= \Sigma x_i^2 - n\overline{x}^2 \\ SXY &= \Sigma(x_i - \overline{x})(y_i - \overline{y}) = \Sigma x_i y_i - n\overline{xy} \qquad (5.6) \end{aligned}$$

Solving (5.3), (5.4), we get:

$$\begin{aligned} &\beta_0 n + \beta_1 \Sigma x_i = \Sigma y_i \\ &\beta_0 = \frac{\Sigma y}{n} - \beta_1 \frac{\Sigma x}{n} \\ &\beta_0 = \overline{y} - \beta_1 \overline{x} \qquad (5.7) \end{aligned}$$

$$\begin{aligned} &\beta_0 \Sigma x_i + \beta_1 \Sigma x_i^2 = \Sigma x_i y_i \\ &\beta_1 = \frac{\Sigma x_i y_i - \beta_0 \Sigma x_i}{\Sigma x_i^2} \\ &\beta_1 = \frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{n\Sigma x_i^2 - \Sigma x_i \Sigma x_i} \\ &\beta_1 = \frac{\Sigma x_i y_i - n\overline{xy}}{n\Sigma x_i^2 - n\overline{x}^2} \\ &\beta_1 = \frac{SXY}{SXX} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} = r_{xy}\frac{SD_x}{SD_y} \qquad (5.8) \end{aligned}$$

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:

$$\overline{x} = 202.95294118 \quad SXX = 530.78235294 \quad SXY = 475.31223529$$
$$\overline{y} = 139.60529412 \quad SYY = 427.79402353$$

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

$$\hat{\beta}_1 = \frac{SXY}{SXX} = 0.895 \tag{5.9}$$

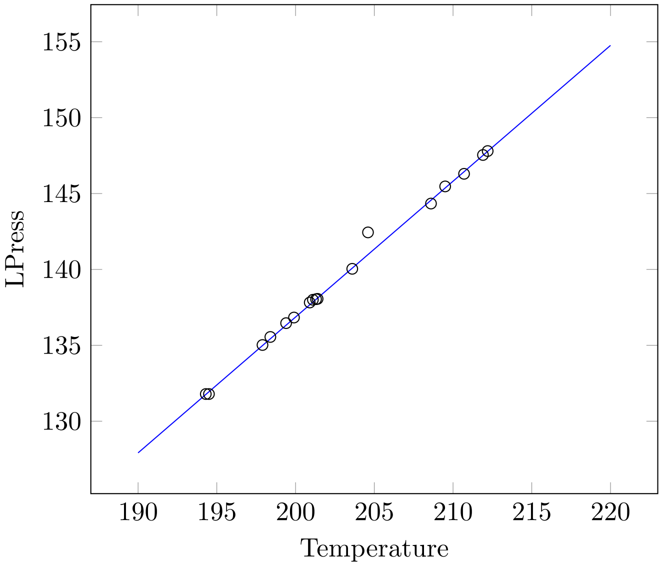
$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x} = -42.138 \tag{5.10}$$

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 β_0 and β_1 using the formulas in **Table 5.1**

Figure 5.3: The straight-line which has minimize residual



The variance σ^2 is the average squared size of the ${e_i}^2$, we should expect that its estimator $\hat{\sigma}^2$ is obtained by averaging the squared residuals. Under the assumption that the errors are uncorrelated random variables with zero means and common variance σ^2, an unbiased estimate of σ^2 is obtained by dividing RSS = $\Sigma{\hat{e}_i}^2$ by its degrees of freedom (df), where residual df = number of cases minus the number of parameters in the mean function. For simple regression, residual df = n - 2, so the estimate of σ^2 is given by:

$$\hat{\sigma}^2 = \frac{RSS}{n-2} \tag{5.11}$$

The variance $\sigma^2 = 0$ indicates that all values in the data set are the same. In another hand, there was no error. All nonzero variances will be positive numbers. Using the summaries for Forbes' data given, we find:

$$RSS = SYY - \frac{SXY^2}{SXX} = 2.15493 \tag{5.12}$$

$$\sigma^2 = \frac{2.15493}{17-2} = 0.14366 \tag{5.13}$$

The square root of $\hat{\sigma}^2$, $\hat{\sigma} = \sqrt{0.14366} = 0.37903$ is called the standard error of regression.

To analyze the variance, we have many methods to compare the fit of 2 or more mean functions for the collected data set. One of the ways to choose a regression line for the model is

$$E(Y|X = x) = \beta_0 \tag{5.14}$$

It is clearly that with all x_i values (X = x_i), we always get a constant Y value. It is a line parallel to the horizontal axis. If we use E (Y | X = x) = β_0 with OLS, we get the formula for RSS is:

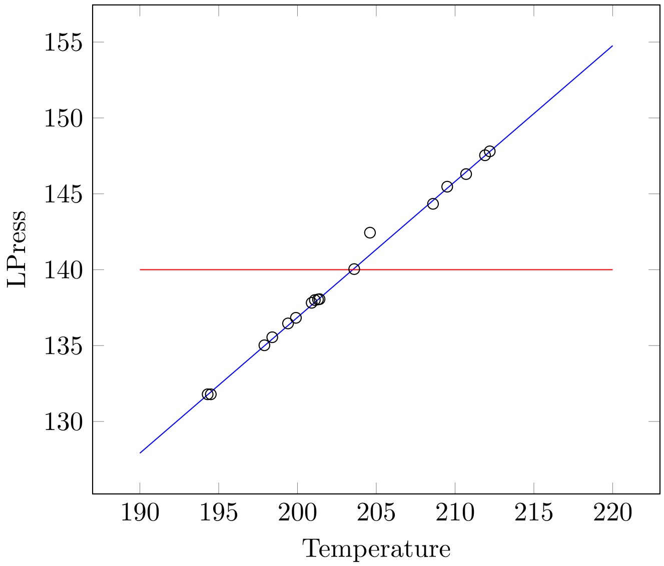
$$RSS = \Sigma(y_i - \hat{B}_0)^2 = \Sigma(y_i - \overline{y})^2 = SYY \tag{5.15}$$

The sum of the squares has df = n - 1.

Next, to increase the fit of the regression line, we will create a slope for the regression line by adding a variable so that Y changes depending on X.

$$E(Y|X = x) = \beta_0 + x\beta_1 \tag{5.16}$$

Figure 5.3: Two mean functions compared by the analysis of variance



Obviously, when we look at the straight line using E(Y | X = x) = $\beta_0 + \beta_1 x$ (the blue line) is better than E(Y | X = x) = β_0 (the red line). Of course, it is easy to see that the estimation of parameter β_0 by E(Y | X = x) = $\beta_0 + \beta_1 x$ and E(Y | X = x) = β_0 is completely different.

The difference between the sum of squares at (13) and that at (RSS) is the reduction in residual sum of squares due to enlarging the mean function from (2.13) to the simple regression mean function (2.16). This is the sum of squares due to regression, SSreg, defined by

$$SSR = SSY - RSS = \frac{SXY^2}{SXX} \tag{5.17}$$

The df associated with SSR is the difference in df for mean function (12), n - 1, and the df for mean function (14), n - 2, so the df for SSR is (n - 1) - (n - 2) = 1 for simple regression.

Table 5.4.1: The Analysis of variance

Source	df	SS	MS	F	p-value
Regression	1	SSreg	SSreg/1	MS/Sig	
Residual	n - 2	RSS	Sig		

Table 5.4.2: The Analysis of variance

Source	df	SS	MS	F	p-value
Regression	1	425.639	425.639	2962.79	0
Residual	15	2.155	0.144		

A smaller mean function obtained from the larger by setting some parameters to zero, or occasionally setting them to some other known value.

For example, equation (2.13) was obtained from (2.16) by setting $\beta_1 = 0$. The line in the ANOVA table for the total gives the residual sum of squares corresponding to the mean function with the fewest parameters

If the sum of squares for regression $SSreg$ is large, then the simple regression mean function E(Y | X = x) = $\beta_0 + \beta_1 x$ should be a significant improvement over the mean function given by (2.13), E(Y | X = x) = β_0. This is equivalent to saying that the additional parameter in the simple regression mean function $\beta_1 x$ is different from zero or that E(Y | X = x) is not constant as X varies.

We call this ratio F:

$$F = \frac{SSreg/1}{\sigma^2}$$

comparing the regression mean square, SSreg divided by its df, to the residual mean square σ^2. We need to judge how large is “large.”.

$$NH(\text{null hypothesis}) : E(Y|X = x) = \beta_0$$
$$AH(\text{alternative hypothesis}) : E(Y|X = x) = \beta_0 + \beta_0 x$$

If the error is follow normal distribution N(0, σ^2), under NH will follow an F-distribution with df associated with numerator and denominator of, 1 and n - 2 for simple regression (F $\sim$ F(1, n - 2)).

$$F = \frac{425.639}{0.144} = 2963$$

The p-value of F (or the significance of level of null hypothesis). The p-value in the table 5.2 is shown as “approximately zero”, meaning that, if the NH were true, the change of F exceeding its observed value is essentially zero. This is very strong evidence against NH and in favor of AH.

Table 5.5: The Analysis of variance

	Unstandardized Coefficients		t	Sig.	95% Confidence Interval for B	
	B	Std. Error			Lower Bound	Upper Bound
0	-42.138	3.340	12.616	0.000	-47.993	-36.282
1	0.895	0.0164	54.604	0.000	0.867	0.930

In this table, we see the number -42.138 to the right of the ”Constant” label and under the labels ”Unstandardized Coefficients” and ”B”. This is called the intercept estimate, estimated intercept coefficient, or estimated constant, and can be written as b_0, $\hat{\beta}0$ or rarely B_0 is incorrect, because the parameter value β_0 is a fixed, unknown ”secret of nature”. The number 0.0164 is the slope estimate, estimated slope coefficient, slope estimate.It can be written as b_1, $\hat{\beta}1$ or rarely β_1 is incorrect.

To the right of the intercept and slope coefficient, you will find their standard errors. Formula for the standard errors:

$$SE(\beta_0) = \sqrt{\sigma^2 \frac{1}{SXX}}$$
$$SE(\beta_1) = \sqrt{\sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{SXX} \right)}$$

Standard errors are estimated standard deviations of the corresponding sampling distributions.

The two statistic are calculated by using the default null hypotheses of $H_0 : \beta_j = 0$ according to the general t-statistic formula

$$t_j = \frac{b_j - \text{hypothesized value of } \beta_j}{SE(b_j)} \tag{5.18}$$

When the errors are NID(0, σ^2), parameter estimates, fitted values, and predictions will be normally distributed. Confidence intervals and tests can be based on the t-distribution. Let t(α/2, d) be the value that cuts off $\alpha/2$ x 100 % in the upper tail of the t-distribution with d df.

The standard error of the intercept is SE(β_0). Hence a (1 - α) x 100 % confidence interval for the intercept is the set of points β_0 in the interval

$$\hat{\beta}_0 - t(\alpha/2, n-2)SE(\hat{\beta}_0) \leq \beta_0 \leq \hat{\beta}_0 + t(\alpha/2, n-2)SE(\hat{\beta}_0)$$

Because

$$P\left(t(\alpha/2, n-2) \leq \frac{\beta_0 - \hat{\beta}0}{SE(\beta 0)} \leq t(1-\alpha/2, n-2) \right) = 1 - \alpha$$

For a 95% confidence interval, t(0.05, 15) = 2.131, and the interval is

$$-49.255 \leq \beta_0 \leq -35.021$$

A hypothesis test of

$$\begin{aligned} NH(\text{null hypothesis}) &: \beta_0 = 0 \\ AH(\text{alternative hypothesis}) &: \beta_0 \neq 0 \end{aligned}$$

is obtain by computing the t-statistic

$$t_0 = \frac{\hat{\beta}_0 - \beta_0}{SE(\hat{\beta}_0)}$$

In Forbes's data, consider testing the NH $\beta_0 = 0$ against the alternative that $\beta_0 \neq 0$. The statistic is

$$t_0 = \frac{-42.138 - 0}{3.340} = 12.616$$

which has a small p-value (approximately zero) provides evidence against the NH.
The standard error of $\hat{\beta}_1$ is $SE(\hat{\beta}_1) = 0.0164$. For a 95% confidence interval for slope is the set of β_1 is

$$0.876 \leq \beta_1 \leq 0.930$$

A hypothesis test of

$$NH(\text{null hypothesis}) : \beta_1 = 0$$
$$AH(\text{alternative hypothesis}) : \beta_1 \neq 0$$

For the Forbes's data, t = 54.604 which also has a small p-value(approximately zero) provides evidence against the NH.

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:

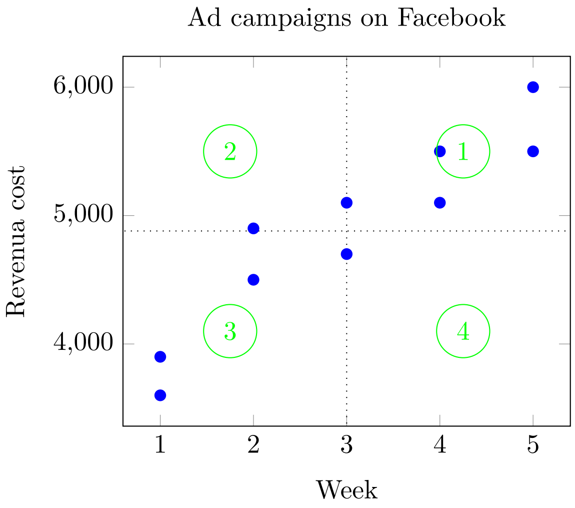
$$s_{xy} = \frac{\Sigma(x_i - \overline{x})(y_i - \overline{y})}{n-1}$$

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:

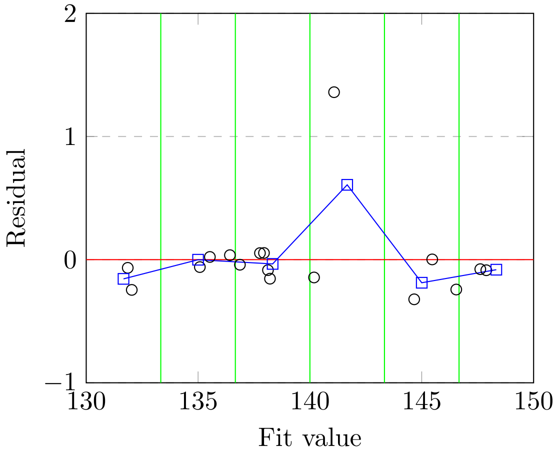
The line in which value $x = 4$ is the average value over 10 weeks, the y line of 10-week average revenue. This line divides the graph into four parts. Part 1 will have x value greater than average of x, y greater than mean of y. Part 2 will have x value less than the mean of x and value of y greater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

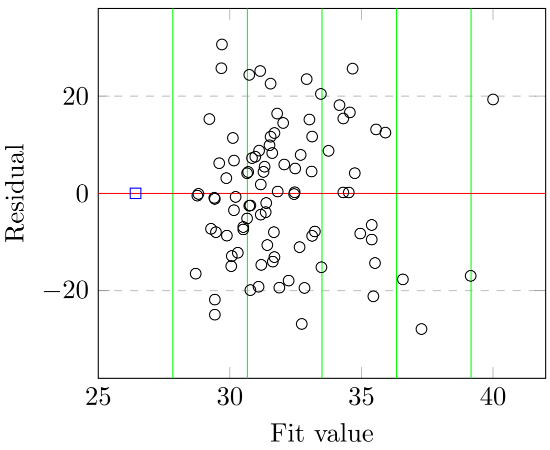
Therefore, we can see $(x - \overline{x})(y - \overline{y})$ will be positive with parts 1 and 3 but negative with parts 2 and 4.

- 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





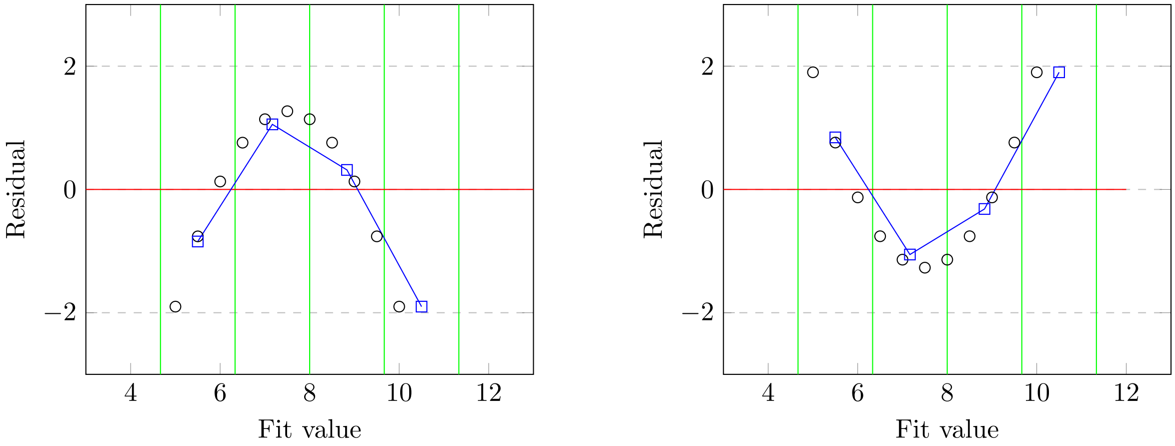


Figure 7.1 Sample residual versus fit plots for testing equal variance

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 7.1. 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.