

Simple Linear Regression

Let's consider two variables, one explanatory (**X**) and one response variable (**Y**), presented in **Table 1** and shown in **Fig 1**.

Nº	X	Y	Nº	X	Y
1	93.5	91.3	10	82	76
2	171	133	11	146	116
3	192.4	155.4	12	168	136
4	206.4	180.6	13	323	259
5	279	241.8	14	302	250
6	229.9	146.3	15	284	227
7	59	53	16	239	177
8	91	80	17	311	248
9	70	60			

Table 1. Observed data on **X** and **Y**.

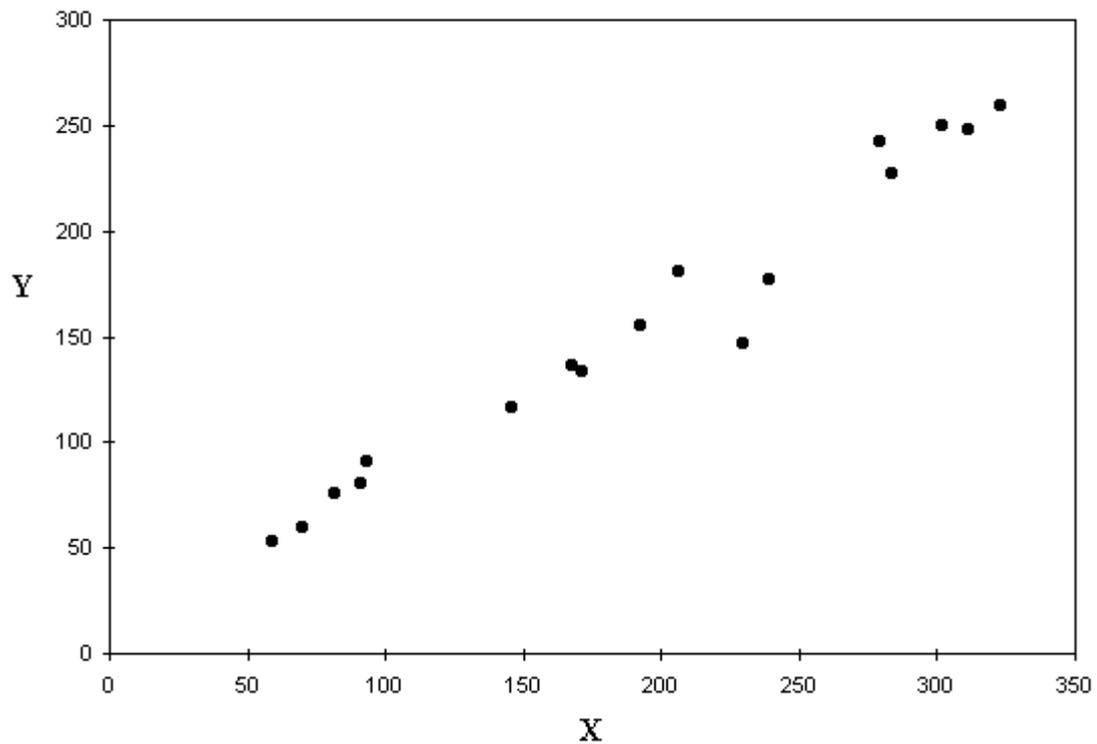


Figure 1 A plot of the data.

By looking at this scatter plot, it can be seen that variables **X** and **Y** have a close relationship that may be reasonably represented by a straight line. This would be represented mathematically as

$$\mathbf{Y} = \mathbf{a} + \mathbf{bX} + \boldsymbol{\varepsilon}$$

where **a** (the intercept) describes where the line crosses the y-axis, **b** describes the slope of the line, and $\boldsymbol{\varepsilon}$ is an error term that describes the variation of the real data above and below the line. Simple linear regression attempts to find a straight line that best 'fits' the data, where the variation of the real data above and below the line is minimized.

The Ordinary Least Squares estimates of the parameters are given by

$b = \frac{cov(x,y)}{\sigma^2(x)}$; $a = \bar{y} - b \cdot \bar{x}$, where $cov(x,y)$ is covariance between X and Y, \bar{y} and \bar{x} are the mean values of Y and X, respectively.

The fitted line has **a**=7.76 and **b**=0.769 (OLS estimates) and now that we know the equation,

$$\hat{\mathbf{Y}} = 7.76 + 0.769\mathbf{X},$$

we can plot the line onto the data ($\boldsymbol{\varepsilon}$ is not needed to plot the line); see **Fig 2.1**.

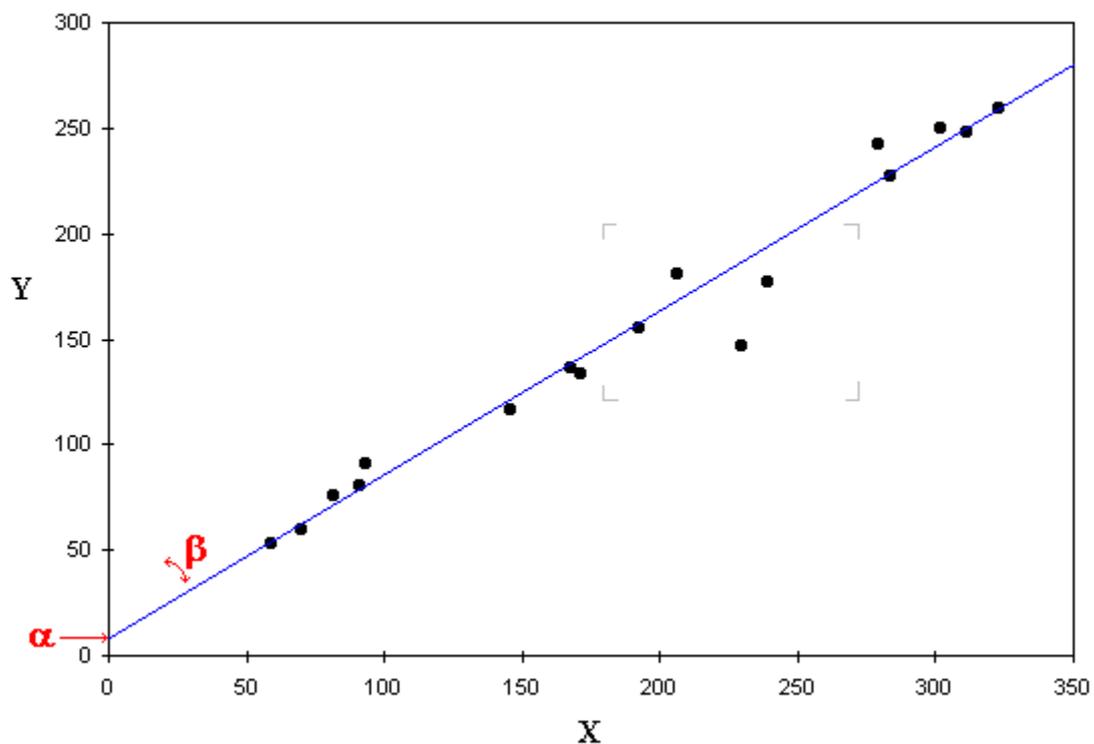


Figure 2.1 Plot showing the fitted regression line and data points.

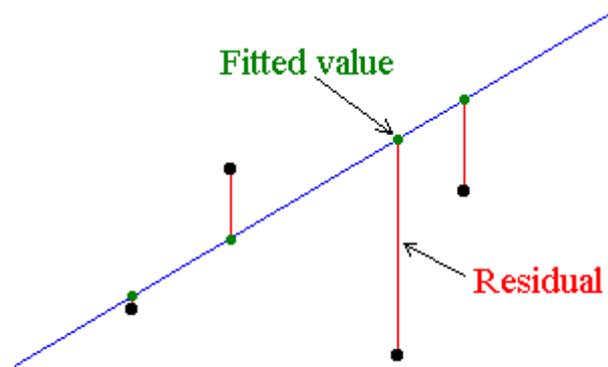


Figure 2.2 Detailed section of Fig. 2.1 with residuals and fitted values shown.

ANALYSIS OF VARIATION

Three sums of squares represent variation of **Y** from several sources (the squares are taken to 'remove' the sign (+ or -)).

1. **TSS** (total sum of squares) describes the variation within the values of **Y**, and is the sum of the squared difference between each value of **Y** and the mean of **Y**.
2. **ESS** (explained sum of squares) describes the variation within the fitted values of **Y**, and is the sum of the squared difference between each fitted value of **Y** and the mean of **Y** (the source of this variation is the variation of **X**).
3. **RSS** (residual sum of squares) describes the variation of observed **Y** from estimated (fitted) **Y**. It is derived from the cumulative addition of the square of each residual, where a residual is the distance of a data point above or below the fitted line (see Fig 2.2) (the source of this variation is the variation of ϵ).

Note: $TSS=ESS+RSS$ (normally)

TESTING THE MODEL

1. *What is the probability that the values for **a** and **b** are not derived by chance?*

The **p** (probability) values state the confidence that one can have in the estimated values being correct, given the constraints of the regression analysis (to be discussed later). Calculating **p-values** involves *t-test*.

The *t-statistic* is the *coefficient* divided by its *standard error* ($se_b = \sigma_b$). The standard error is an estimate of the *standard deviation* of the coefficient, the amount it varies across cases. It can be thought of as a measure of the precision with which the regression coefficient is measured. If a coefficient is large compared to its standard error, then it is probably different from 0.

How large is large? It is necessary to compare the *t-statistic* on your variable **X** with values in the *Student's t-distribution* to determine the *p-value*, which is the number that you really need to be looking at. If the *p-value* associated with the test is less than 0.05 we reject the hypothesis that $\beta=0$ (the true value of a coefficient is zero and therefore **X** does not influence **Y**). If the associated *p-value* is equal to or greater than 0.05, we do just the opposite, that is we accept the null hypothesis that $\beta=0$, which implies that you have to look for another explanatory variable for modeling your **Y**.

2. Goodness of fit

Statistical measure of how well a regression line approximates real data points is the *coefficient of determination* (**D**, R^2 , *R-squared*). It represents the % variation of the data explained by the fitted line; the closer the points to the line, the better the fit. *R-squared* of 1.0 (100%) indicates a perfect fit. *R-squared* is derived from

$$R^2 = 100 \times \frac{ESS}{TSS}.$$

Testing statistical significance of *R-squared* (and the overall statistical significance of the model) involves *p-value* again. Now it is associated with the *F-test*. The *F statistic* is a ratio of the explained and residual variation of **Y** taking account of their degrees of freedom¹ and is derived from

$$F = \frac{ESS/m}{RSS/(T-(m+1))},$$

¹ Degrees of freedom represent the number of independent values in a calculation, minus the number of estimated parameters. For example, the variance (unbiased) of *n* data points has *n-1* degrees of freedom, because the variance requires estimating another parameter (the mean) in its calculation.

Degrees of freedom can also be thought of as the number of values that are free to vary in a calculation. Examples:

- 1) If you have to take ten different courses to graduate, and only ten different courses are offered, then you have nine degrees of freedom. Nine semesters you will be able to choose which class to take; the tenth semester, there will only be one class left to take - there is no choice, if you want to graduate.
- 2) When ranking ten items, there are only nine degrees of freedom. That is, once the nine items are ranked, the tenth is already determined.

where m is the number of the explanatory variables (one in case of a simple regression), $m+1$ is the number of the coefficients (two for a and b), T is the number of observations in your data set. If the variance of Y accounted for the regression (X) is large compared to the variance accounted for the error (ε), then the model is good.

Again, how large is large? It is necessary to compare the F statistic on your variances with values in the F distribution to determine the p -value, which is the number that you really need to be looking at. If the p -value associated with the test is less than 0.05 we reject the hypothesis that $R^2=0$ (the true value of a coefficient is zero and therefore X does not determine Y or, alternatively, the true model is $\hat{Y}=\alpha$ rather than $\hat{Y}=\alpha+bX$).

For the data presented above:

standard error for β (σ_β) equals 0.03631, t -statistic equals 21.19, p -value is practically zero ($1.36 \cdot 10^{-12}$);

$R^2=0.968$, implying that 96.8% of the total variation within Y may be accounted for the variation within X or that the regression explains/determines 96.8% of the total variation within Y ;

F -statistic is 449.04, p -value is almost zero ($1.36 \cdot 10^{-12}$).

CONFIDENCE INTERVALS FOR THE PARAMETERS OF A LINEAR REGRESSION

A confidence interval is a measure of the reliability of an estimate. It is a type of interval estimate of a population parameter.

Confidence interval tells you the most likely range of the unknown population average.

It is an observed interval (i.e. it is calculated from the observations), in principle different from sample to sample, that frequently includes the parameter of interest if the experiment is repeated. How frequently the observed interval contains the parameter is determined by the confidence level (β).

More specifically, the meaning of the term "confidence level" is that, if confidence intervals are constructed across many separate data analyses of repeated (and possibly different) experiments, the proportion of such intervals that contain the true value of the parameter will match the confidence level; this is guaranteed by the reasoning underlying the construction of confidence intervals.

Whereas two-sided confidence limits form a confidence interval, their one-sided counterparts are referred to as lower or upper confidence bounds.

Confidence intervals consist of a range of values (interval) that act as good estimates of the unknown population parameter. However, in infrequent cases, none of these values may cover the value of the parameter. **The level of confidence of the confidence interval (β) would indicate the probability that**

the confidence range captures this true population parameter given a distribution of samples.

This value is represented by a percentage, so when we say, "**we are 95% confident that the true value of the parameter is in our confidence interval**", we express that 95% of the observed confidence intervals will hold the true value of the parameter. If a corresponding hypothesis test is performed, the confidence level is the complement of respective level of significance, i.e. a 95% confidence interval reflects a significance level of 0.05. The confidence interval contains the parameter values that, when tested, should not be rejected with the same sample. **Greater levels of variance yield larger confidence intervals, and hence less precise estimates of the parameter.**

In applied practice, confidence intervals are typically stated at the 95% confidence level.

A $(1 - \alpha)\%$ level confidence interval for β is given by

$$(b - t_{crit} \cdot \sigma_b; b + t_{crit} \cdot \sigma_b),$$

where t_{crit} is the upper $\alpha/2$ critical value of the t distribution with $(T-2)$ degrees of freedom (for a simple regression).