

Simple Linear Regression

W. Robert Stephenson
Department of Statistics
Iowa State University

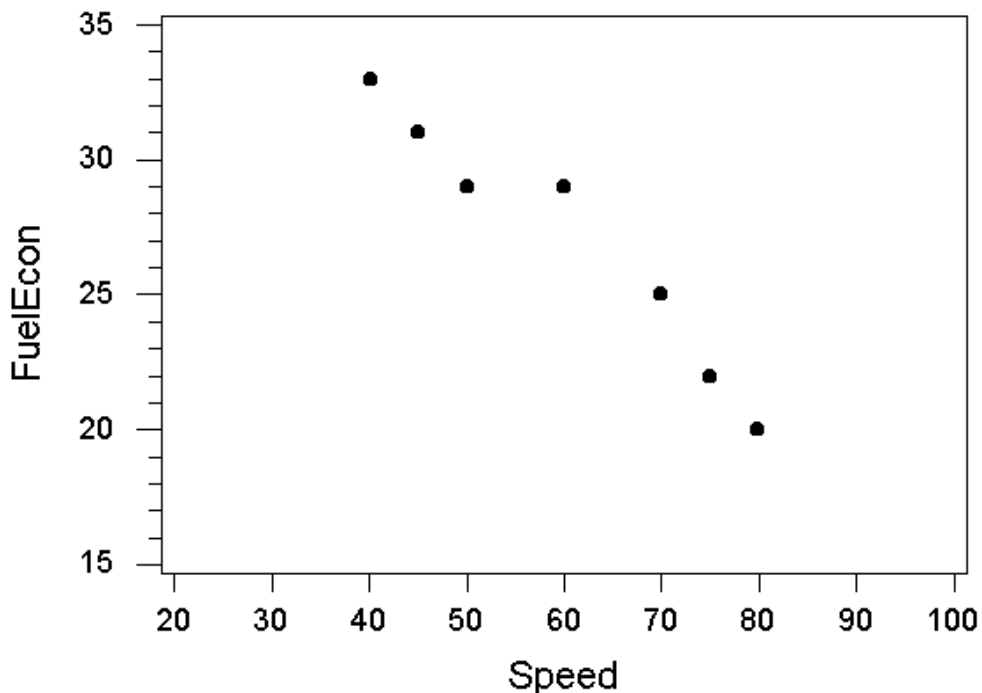
One of the most widely used statistical techniques is **simple linear regression**. This technique is used to relate a measured **response** variable, Y , to a single measured **predictor** (explanatory) variable, X , by means of a straight line. It uses the **principle of least squares** to come up with values of the “best” **slope** and **intercept** for a straight line that approximates the relationship. By means of the example below we motivate the technique, indicate the rationale underlying the calculations and come up with the formulas for the “best” slope and intercept based on the data.

Example: An experiment is run to investigate the relationship between speed and fuel economy for a car operating at highway speeds. The experiment is run in the laboratory rather than on the road in order to control for outside factors such as weather, road surface, tire wear and the like. This does limit the scope of the experiment since the fuel economies obtained in the laboratory may not match those obtained by the average driver on actual highways. There are 7 levels of speed in the experiment. At each level the car is maintained at that speed and the fuel economy in miles per gallon is measured. The speeds are run in random order. Below are the data.

Run Order	Speed, X (MPH)	Fuel Economy, Y (MPG)
4	40	33
7	45	31
1	50	29
2	60	29
6	70	25
3	75	22
5	80	20

In order to get an idea of the relationship between Speed and Fuel Economy a scatterplot is constructed. The predictor (explanatory) variable appears on the horizontal axis with the response variable on the vertical axis. One can see from the plot of Fuel Economy vs. Speed that as the speed is increased the fuel economy, in general, decreases. There is a general inverse, or negative, association.

Fuel Economy vs. Speed



The purpose of simple linear regression is to come up with a straight line that captures the relationship between the predictor and the response variable. One can see from the data that when the speed increases from 40 MPH to 80 MPH the fuel economy decreases from 33 MPG to 20 MPG. On average the fuel economy has decreased 13 MPG over an increase of 40 MPH. If one uses the change in Y over the change in X as an approximation for the slope of a line, one obtains a naive estimate of the slope, $\tilde{\beta}_1 = \frac{-13}{40} = -0.325$. Extending this back to the point where $X=0$, one gets a naive estimate of the intercept as $\tilde{\beta}_0 = 33 + 40 \cdot (0.325) = 46$. An equation of one line that might be used to approximate the relationship between speed, X, and fuel economy, Y, is: $\tilde{Y} = 46 - 0.325X$

Is this the “best” straight line for approximating the linear relationship? That depends on how we define “best.” Some might define best as the line that goes through the most data points. Others would use a criterion based on how closely the approximation comes to the actual responses. To measure “closeness” we need to look at the idea of a **residual**. The residual is the difference between the observed response and the predicted (approximated) response.

$$\text{residual} = \text{observed} - \text{predicted}$$

Example:

Speed,X (MPH)	Fuel Economy,Y (MPG)	Predicted $\tilde{Y}=46 - 0.325X$	Residual $(Y - \tilde{Y})$	Squared Residual $(Y - \tilde{Y})^2$
40	33	33.0	0.0	0.00
45	31	31.4	-0.4	0.16
50	29	29.8	-0.8	0.64
60	29	26.5	2.5	6.25
70	25	23.3	1.7	2.89
75	22	21.6	0.4	0.16
80	20	20.0	0.0	0.00
				10.1

The naive line does very well at the ends but not very well in the middle. We can summarize how well the line fits by using the **sum of squared residuals**. The smaller this value is the “closer” the predictions are to the observations in some overall sense. This leads us to the generally accepted idea of what the “best” straight line is. **The “best” straight line is the line that minimizes the sum of squared residuals.** This idea is often referred to as the **principle of least squares**.

Mathematically, we wish to minimize:

$$\sum(Y - (\beta_0 + \beta_1 X))^2$$

with respect to β_1 and β_0 . This can be done by taking derivatives and setting them equal to zero.

$$\frac{\partial}{\partial \beta_1} \sum(Y - (\beta_0 + \beta_1 X))^2 = \sum -2X(Y - (\beta_0 + \beta_1 X)) \equiv 0$$

$$\frac{\partial}{\partial \beta_0} \sum(Y - (\beta_0 + \beta_1 X))^2 = \sum -2(Y - (\beta_0 + \beta_1 X)) \equiv 0$$

This results in the so called **normal equations**:

$$\sum Y = \beta_0 n + \beta_1 \sum X$$

$$\sum XY = \beta_0 \sum X + \beta_1 \sum X^2$$

which can be solved simultaneously to obtain:

$$\hat{\beta}_1 = \frac{SS_{XY}}{SS_{XX}}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

where:

$$SS_{XY} = \left[\sum XY - \frac{(\sum X)(\sum Y)}{n} \right]$$

$$SS_{XX} = \left[\sum X^2 - \frac{(\sum X)^2}{n} \right]$$

Example:

X^2	Speed,X (MPH)	Fuel Economy,Y (MPG)	XY
1,600	40	33	1,320
2,025	45	31	1,395
2,500	50	29	1,450
3,600	60	29	1,740
4,900	70	25	1,750
5,625	75	22	1,650
6,400	80	20	1,600
26,650	420	189	10,905

$$SS_{XY} = \left[10,905 - \frac{(420)(189)}{7} \right] = -435$$

$$SS_{XX} = \left[26,650 - \frac{(420)^2}{7} \right] = 1,450$$

$$\hat{\beta}_1 = \frac{-435}{1,450} = -0.30$$

$$\hat{\beta}_0 = \frac{189}{7} - (-0.30)\left(\frac{420}{7}\right) = 27.0 + 18.0 = 45.0$$

$$\hat{Y} = 45.0 - 0.30X$$

Speed,X (MPH)	Fuel Economy,Y (MPG)	Predicted $\hat{Y} = 45.0 - 0.30X$	Residual $(Y - \hat{Y})$	Squared Residual $(Y - \hat{Y})^2$
40	33	33.0	0.0	0.00
45	31	31.5	-0.5	0.25
50	29	30.0	-1.0	1.00
60	29	27.0	2.0	4.00
70	25	24.0	1.0	1.00
75	22	22.5	-0.5	0.25
80	20	21.0	-1.0	1.00
			0.0	7.5

Note that the sum of squared residuals is smaller than that from the naive straight line. Indeed, there is no other straight line with a smaller sum of squared residuals.

Statistical analysis of regression goes beyond the calculation of the least squares line. Such further analysis looks at the **statistical significance** of the linear relationship and the **strength** of the linear relationship. This analysis depends on the separation of variability into **explained** and **unexplained** components. The sum of squared residuals summarizes the **unexplained** or **Error** variability. The **explained** variability, that due to regression, can be summarized in a “sum of squares” given by the following:

$$SS_{Regression} = (\hat{\beta}_1)^2 SS_{XX} = \frac{(SS_{XY})^2}{SS_{XX}}$$

Together the sum of squared residuals and sum of squares due to regression add up to the Total sum of squares. That is:

$$SS_{Residual} + SS_{Regression} = SS_{Total} = \left[\sum Y^2 - \frac{(\sum Y)^2}{n} \right]$$

Information about variability is summarized in an **Analysis of Variance** table, like the one below.

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
Regression	1	$SS_{Regression}$	$SS_{Regression}/1$
Residual (Error)	$n - 2$	$SS_{Residual}$	$SS_{Residual}/(n - 2)$
Total	$n - 1$	SS_{Total}	

The **Mean Square Residual** provides information about variability that is not explained by the straight line. This variability could be due to the **lack of fit**, the inability of a straight line to capture the relationship between predictor and response, and/or due to **random error**. The **Mean Square Residual** provides a basis for assessing **statistical significance**. **A relationship is said to be statistically significant if it cannot be explained by random error variability**. A formal test of this is based on the test statistic:

$$t = \frac{\hat{\beta}_1}{\sqrt{\frac{MS_{Error}}{SS_{XX}}}}$$

Extreme values of $|t|$, larger than 2 or 3 ¹, indicate that one gains significant information about the response by knowing the value of X and the regression equation.

Finally, the **strength** of the linear relationship can be assessed by the **coefficient of determination** given by the formula:

$$R^2 = \frac{SS_{Regression}}{SS_{Total}}$$

Which can be interpreted as the percentage of the total variability in the response variable that can be explained by the linear relationship with the predictor variable.

¹One can get more precise with the exact cutoffs using a table of the t distribution. However, for most practical purposes this rule of thumb is adequate.