Simple Linear Regression

Introduction

In a simple linear model, there's one independent variable and one dependent variable.

$$Y = \alpha + \beta X + \varepsilon \tag{1}$$

where

X and Y are the independent and dependent variables, respectively;

$\alpha, \beta \in \mathbb{R}$ (i.e. the coefficients are real numbers);

ε is the random error or residual, usually assumed to follow a normal distribution.

Least-Squares Regression

If we denote our model estimates for α and β by a and b, respectively, and the fitted line by

$$\hat{Y} = a + bX, \tag{2}$$

then the sum of the squared residuals (*sum of squared errors*, or SSE) for the model is given by

$$\text{SSE} = \sum (y - \hat{y})^2. \tag{3}$$

The values of a and b which minimize SSE are

$$\begin{aligned} a &= \frac{\sum x^2 \sum y - \sum x \sum xy}{n\sum x^2 - \left(\sum x\right)^2} \\ b &= \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2} \end{aligned} \tag{4}$$

Example

Using the x and y values shown, fill in the xy and x^2 columns, and the bottom row.

x	y	xy	x^2
0	-0.5		
1	0		
2	1.5		
3	2		
$\sum x =$	$\sum y =$	$\sum xy =$	$\sum x^2 =$

Table 1: Data and calculations for regression example

Use the values from the bottom row of Table 1, along with the number of data points (n), to calculate a and b.

$$a = \frac{\sum x^2 \sum y - \sum x \sum xy}{n\sum x^2 - \left(\sum x\right)^2}$$

$$= \qquad (5)$$

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2}$$

$$=$$

Use the a and b values from (5) to write the formula of the fitted line, following the form of (2).

$$\hat{Y} = \qquad (6)$$

To visualize the fit, plot the data points in Table 1, along with the fitted regression line in (4), on the graph in Figure 1.

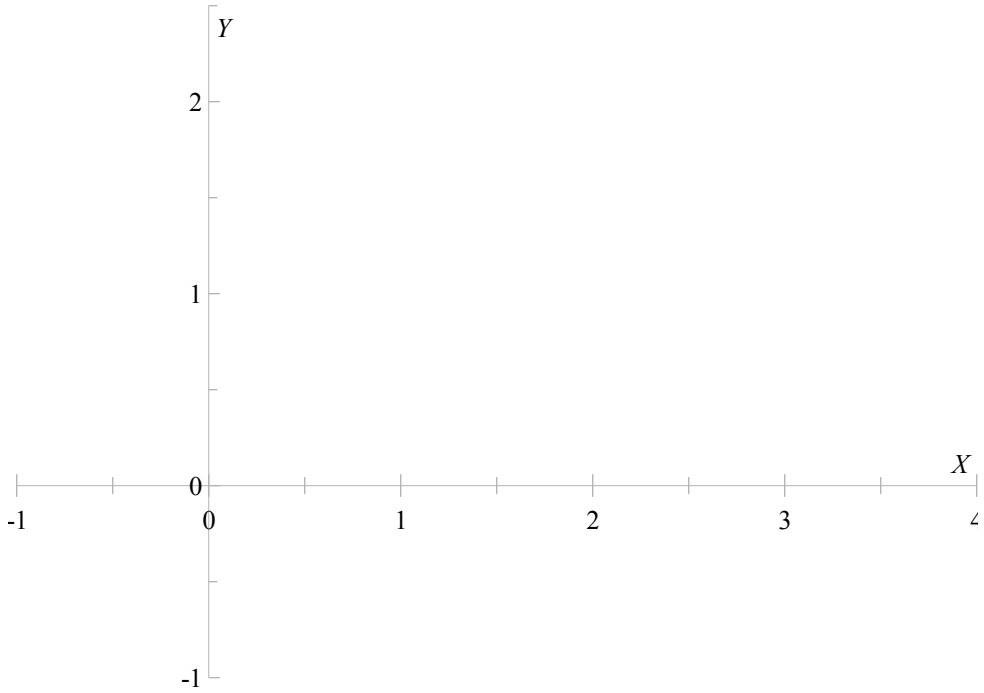

Figure 1: Example data points and fitted regression line

The Coefficient of Determination

To measure the goodness-of-fit of the regression model, we first quantify the total change in the dependent variable, by summing the squared deviations of its observed values from its sample mean. We call this sum the *total sum of squares*, or SST.

$$\begin{aligned} \text{SST} &= \sum \left(y - \bar{y} \right)^2 \\ &= \sum y^2 - \frac{\left(\sum y \right)^2}{n} \end{aligned} \tag{7}$$

SST can also be expressed as the sum of SSE and the *sum of squares of regression* (SSR), i.e.

$$\text{SST} = \text{SSE} + \text{SSR}, \tag{8}$$

where SSE is given by (3), and

$$SSR = \sum \left(\hat{y} - \bar{y} \right)^2. \tag{9}$$

From SST and SSR, we get a fundamental goodness-of-fit measure: the *coefficient of determination*, or R^2.

$$R^2 = \frac{\text{SSR}}{\text{SST}} \tag{10}$$

We can interpret R^2 as the fraction of the variation in the dependent variable that's determined or explained by the model.

Example

To calculate R^2 for the example data set, start by using the original data and the fitted line equation (6) to fill in the additional columns (including the bottom row) in the data table.

x	y	xy	x^2	y^2	$\hat{y}$	$\left(y - \hat{y} \right)^2$
0	-0.5	0	0			
1	0	0	1			
2	1.5	3	4			
3	2	6	9			
$\sum x = 6$	$\sum y = 3$	$\sum xy = 9$	$\sum x^2 = 14$	$\sum y^2 =$		$\sum \left(y - \hat{y} \right)^2 =$

Table 2: Calculations for coefficient of determination in regression example

Now, use formulas (3), (7), (8), and (10), along with the sums computed in Table 2, and the number of data points (n), to find R^2.

$$\text{SSE} = \sum \left(y - \hat{y} \right)^2$$

$$=$$

$$\text{SST} = \sum y^2 - \frac{\left(\sum y \right)^2}{n}$$

$$=$$

$$\text{SSR} = \text{SST} - \text{SSE}$$

$$=$$

$$R^2 = \frac{\text{SSR}}{\text{SST}}$$

$$= \qquad (11)$$

What does this value of R^2 allow us to say about the fitted model?