# Simple Linear Regression

**An Example**

## Example:

Suppose the following data were collected from a sample of 10 pizzerias of a certain firm:

| Restaurant | Student Population (1000s) | Quarterly Sales(1000s) |
|---|---|---|
| $i$ | $x_i$ | $y_i$ |
| 1 | 2 | 58 |
| 2 | 6 | 105 |
| 3 | 8 | 88 |
| 4 | 8 | 118 |
| 5 | 12 | 117 |
| 6 | 16 | 137 |
| 7 | 20 | 157 |
| 8 | 20 | 169 |
| 9 | 22 | 149 |
| 10 | 26 | 202 |

1.

Find a regression function describing the quarterly sales as a linear function of the student population.

*Solution:*

$$\begin{cases} n \cdot b_0 + b_1 \cdot \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ b_0 \cdot \sum_{i=1}^{n} x_i + b_1 \cdot \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i \cdot y_i \end{cases}$$

*Working Table*

| $x_i$ | $y_i$ | $x_i^2$ | $x_i \cdot y_i$ | $y_i^2$ |
|---|---|---|---|---|
| 2 | 58 | 4 | 116 | 3364 |
| 6 | 105 | 36 | 630 | 11025 |
| 8 | 88 | 64 | 704 | 7744 |
| 8 | 118 | 64 | 944 | 13924 |
| 12 | 117 | 144 | 1404 | 13689 |
| 16 | 137 | 256 | 2192 | 18769 |
| 20 | 157 | 400 | 3140 | 24649 |
| 20 | 169 | 400 | 3380 | 28561 |
| 22 | 149 | 484 | 3278 | 22201 |
| 26 | 202 | 676 | 5252 | 40804 |
| 140 | 1300 | 2528 | 21040 | 184730 |

$$\begin{cases} 10b_0 + 140b_1 = 1300 \\ 140b_0 + 2528b_1 = 21040 \end{cases} \qquad b_0 = 60, \quad b_1 = 5$$

$$y^* = 60 + 5x$$

2.

Find *SST*, *SSR* and *SSE*.

*Solution:*

$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$ : total sum of squares

$SSR = \sum_{i=1}^{n} (y_i^* - \bar{y})^2$ : sum of square due to regression

$SSE = \sum_{i=1}^{n} (y_i - y_i^*)^2$ : sum of squares due to error

$(SST = SSR + SSE)$

*Working Table*

| $i$ | $x_i$ | $y_i$ | $y_i^*$ | $y_i - y_i^*$ | $(y_i - y_i^*)^2$ | $(y_i - \bar{y})^2$ |
|---|---|---|---|---|---|---|
| 1 | 2 | 58 | 70 | -12 | 144 | 5184 |
| 2 | 6 | 105 | 90 | 15 | 225 | 625 |
| 3 | 8 | 88 | 100 | -12 | 144 | 1764 |
| 4 | 8 | 118 | 100 | 18 | 324 | 144 |
| 5 | 12 | 117 | 120 | -3 | 9 | 169 |
| 6 | 16 | 137 | 140 | -3 | 9 | 49 |
| 7 | 20 | 157 | 160 | -3 | 9 | 729 |
| 8 | 20 | 169 | 160 | 9 | 81 | 1521 |
| 9 | 22 | 149 | 170 | -21 | 441 | 361 |
| 10 | 26 | 202 | 190 | 12 | 144 | 5184 |
| | 140 | 1300 | 1300 | 0 | 1530 | 15730 |

$$SSR = SST - SSE = 15730 - 1530 = 14200$$

3.

Find and interpret the coefficients of determination and correlation.

*Coefficient of determination:*

$$r^2 := \frac{SSR}{SST}, \quad 0 \le r^2 \le 1$$

$$r^2 = \frac{14200}{15730} = 0.9027.$$

We can conclude that 90.27% of the total sum of squares can be explained by using the regression equation $y^* = 60 + 5x$ to predict quarterly sales. In other words, 90.27% of the variability in sales can be explained by the linear relationship between the size of the student population and sales. It is a good fit for the estimated regression equation.

*Coefficient of correlation:*

$$r := (\text{sign of } b_1)\sqrt{r^2}, \quad -1 \le \text{r} \le 1.$$

$$r = \sqrt{0.9027} = 0.9501.$$

As expected, there is a direct relationship between the quarterly sales and the student population

The coefficient of correlation (and of course also the coefficient of determination) for a simple linear regression function can alternatively be calculated according to the following formula:

$$r = \frac{n \cdot \sum_{i=1}^{n} x_i \cdot y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{\sqrt{\left( n \cdot \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} x_i \right) \left( n \cdot \sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} y_i \cdot \sum_{i=1}^{n} y_i \right)}}, \quad -1 \le r \le 1$$

$$r = \frac{10 \cdot 21040 - 140 \cdot 1300}{\sqrt{\left(10 \cdot 2528 - 140^2\right) \cdot \left(10 \cdot 184730 - 1300^2\right)}} \approx 0.9501.$$

4.

Carry out a T-test of significance for $\alpha = 0.01$.

*Solution:*

*Step 1:*

$$H_0: \beta_1 = 0, \qquad H_1: \beta_1 \neq 0$$

*Step 2:*

$$s = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{1530}{10-2}} = 13.829$$

$$s_{b_1} = \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \frac{13.829}{\sqrt{568}} = 0.580$$

$$t_{stat} = \frac{b_1}{s_{b_1}} = \frac{5}{0.5803} = 8.62$$

*Step 3:*

$$t_{stat} = 8.62 \geq 3.355 = t_{crit 8;0.01} \quad \Rightarrow \quad \text{reject } H_0$$

5.

Carry out an F-test of significance for $\alpha = 0.01$.

*Step 1:*

$$H_0 : \beta_1 = 0, \qquad H_1 : \beta_1 \neq 0$$

*Step 2:*

$$MSE = \frac{SSE}{n-2} = s^2 = 13.829^2 \approx 191.24$$

$$MSR = \frac{SSR}{\text{Number of independent variables}} = \frac{14200}{1} = 14200$$

$$F_{stat} = \frac{MSR}{MSE} = \frac{14200}{191.24} = 74.25$$

*Step 3:*

$$F_{stat} = 74.25 \geq 11.26 = F_{0.01} \quad \Rightarrow \quad \text{reject } H_0$$

6.

Construct a 1% confidence interval for $\beta_1$.

*Solution:*

$$\beta_1 \in \left[ b_1 - t_{\alpha/2} s_{b_1},\ b_1 + t_{\alpha/2} s_{b_1} \right]$$

$$\beta_1 \in \left[ 5 - 3.355 \cdot 0.5803,\ 5 + 3.355 \cdot 3.5803 \right] = \left[ 3.05,\ 6.95 \right].$$

7.

Predict the sales for a university with 10000 students.

*Solution:*

$$y^*(10) = 60 + 5 \cdot 10 = 110 \quad \text{(or \$110000)}$$

8.

Construct a confidence interval for the mean of $y$ with $x_p = 10$ and $\alpha = 0.05$.

$$E(y_p) \in \left[ y_p^* - t_{\alpha/2} \cdot s_{y_p^*},\ y_p^* + t_{\alpha/2} \cdot s_{y_p^*} \right]$$

where

$x_p$: the particular given value for $x$

$y_p$: the value of $y$ corresponding to the given

$E(y_p)$: the expected value of $y$

$y_p^* = b_0 + b_1 x_p$: the point estimate of $E(y_p)$, when $x = x_p$

$$s_{y_p^*} = s \cdot \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

$$s_{y_p^*} = 13.829 \cdot \sqrt{\frac{1}{10} + \frac{(10-14)^2}{568}} = 4.95$$

$$E(y_p) \in \left[110 - 2.306 \cdot 4.95,\ 110 + 2.306 \cdot 4.95\right] = \left[98.585,\ 121.415\right].$$