Linear regression

Part 3: Error and useful metrics

By: Noureddin Sadawi, PhD

University of London

Predicted and residual values

- Suppose we are given *n* pairs of observations: $(x_1,y_1), (x_2,y_2),\ldots,(x_n,y_n)$
- The estimated or predicted values $\hat{y_1}, \hat{y_2}, \cdots, \hat{y_n}$ are obtained by successfully substituting $x_1, x_2, \ldots, x_n$ into the equation of the estimated regression line. That is:

$$\hat{y_1} = a + bx_1, \hat{y_2} = a + bx_2, \ldots, \hat{y_n} = a + bx_n$$

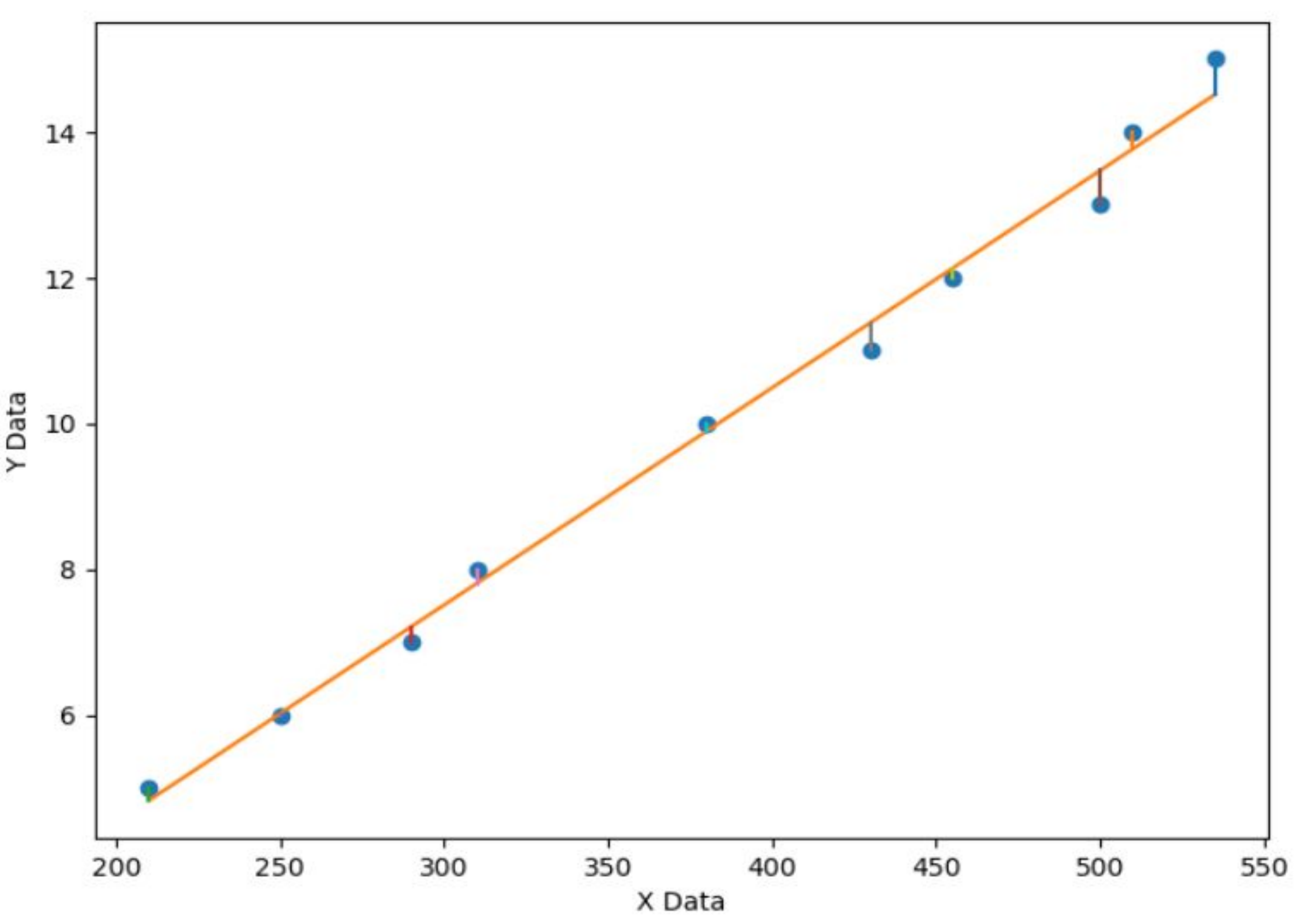
- The residuals $e_1, e_2, \ldots, e_n$ are the differences between the observed and predicted values. That is:

$$e_1 = y_1 - \hat{y_1}, e_2 = y_2 - \hat{y_2}, \ldots, e_n = y_n - \hat{y_n}$$

- In other words, residuals are the distances between each of the points and the line.

Residuals

- We project the points on the line and compute the vertical distance between each point and the line.
- To project means to plug in the values into the line equation.



Error sum of squares

- The residuals allow us to compute the *sum of squares due to error* (also known as the residual sum of squares).
- It is denoted SSE and is given by:

$$SSE = \sum (y_i - \hat{y}_i)^2$$

- The sum of squares due to error (SSE) measures the error in using the least squares regression equation to estimate the values of the dependent variable in the sample.

Total sum of squares

- A quantity that measures the total amount of variation in observed *y* values is known as the total sum of squares.
- Its denoted SST and given by:

$$SST = \sum (y_i - \bar{y})^2$$

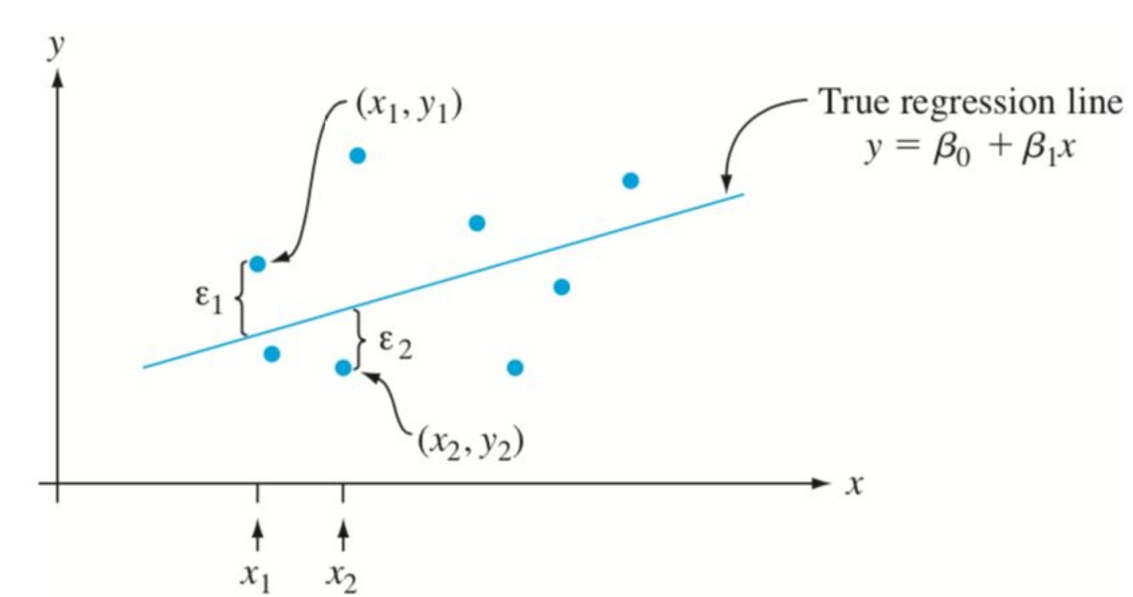
Regression sum of squares

- To measure how much the predicted values $\hat{y}$ deviate from the mean $\bar{y}$, another sum of squares known as the *sum of squares due to regression* is computed.
- Its denoted SSR and given by:

$$SSR = \sum (\hat{y}_i - \bar{y})^2$$

Relationship between SSE, SST and SSR

- SSE, SST and SSR are related as follows:
  SST = SSE + SSR
- If all the observed values of the dependent variable happen to lie on the estimated regression line, then that is a perfect fit and the residuals are 0, hence SSE would be 0.



- That means a perfect fit would mean that SST = SSR, or SSR/SST = 1.

Coefficient of determination

- The coefficient of determination, r^2 or R^2, is the ratio SSR/SST which takes value between 0 and 1.

$$r^2 = \frac{SSR}{SST}$$

- It is interpreted as the proportion of observed variation in *y* that can be explained by the simple linear regression model.

Coefficient of determination

- The higher the value of r^2 the more successful the simple linear regression is at explaining the variation of *y*.
- If r^2 is small, then an alternative model, either a non-linear model or a multiple regression model, maybe required which can more effectively explain the variation in *y*.
- When we express the coefficient of determination, r^2, as a percentage, then this can be viewed as a percentage of the total sum of squares that can be explained by using the regression line.

Useful metrics to measure error

- Remember: the residuals e_1, e_2, …, e_n:

$$e_1 = y_1 - \hat{y_1}, e_2 = y_2 - \hat{y_2}, \ldots, e_n = y_n - \hat{y_n}$$

- Mean squared error (MSE): $MSE = \frac{1}{n}\sum_{t=1}^{n} e_t^2$

- Root mean squared error (RMSE): $RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2}$

- Mean absolute error (MAE): $MAE = \frac{1}{n}\sum_{t=1}^{n} |e_t|$

Interpretation

- The table on the right shows the observations of transportation time and distance for a sample of ten rail shipments by a motor parts supplier.
- Let us determine the regression equation.
- It is easy to plug in the values in the table in the equations we explained before. The results should be:

Slope (i.e. b) = 0.029796

Intercept (i.e. a) = -1.431176

$$b = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} \quad \text{and} \quad a = \bar{y} - b\bar{x}$$

Distance (km)	Delivery time (days)
210	5
290	7
250	6
500	13
310	8
430	11
455	12
380	10
535	15
510	14

Interpretation

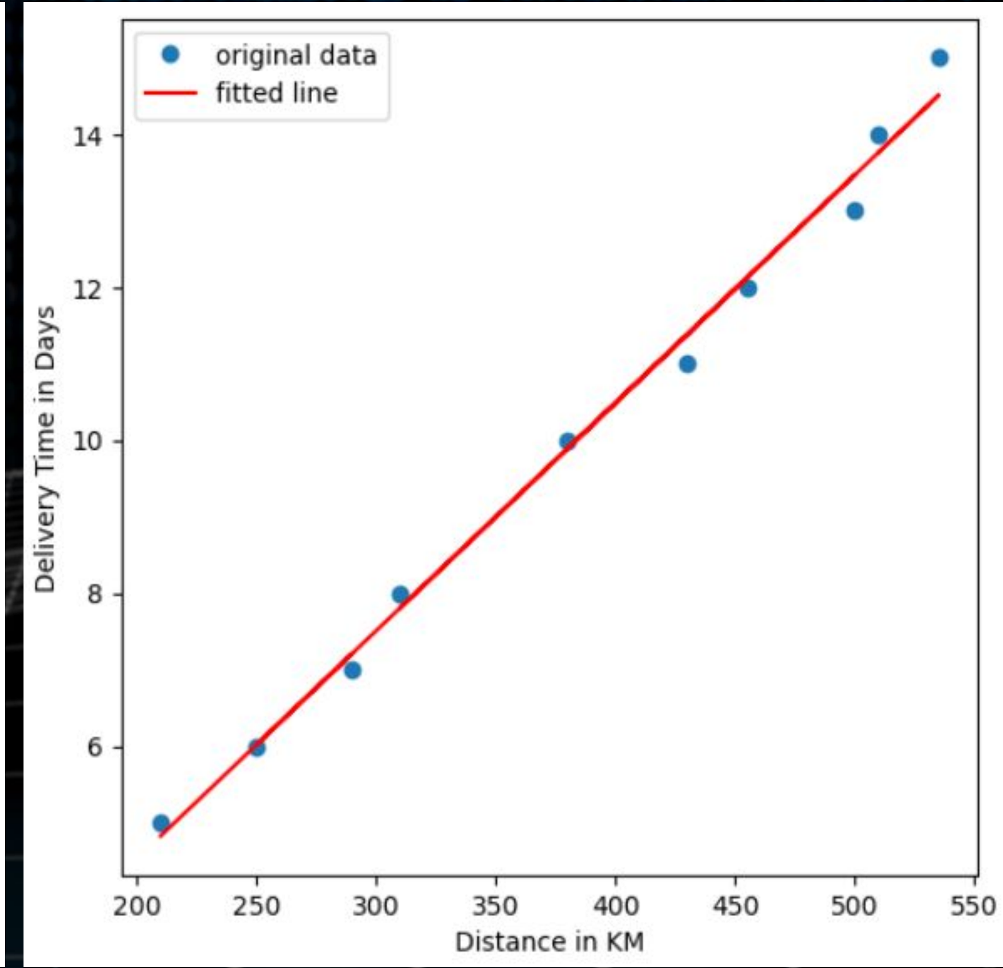
Remember: E(Y) = A + Bx

Estimated delivery time in days =

0.029796*Distance in km - 1.431176

Interpretation: for a unit increase in input (i.e. distance in km), there will be a 0.03 increase in the output (i.e. delivery time in days).

- Intercept in this example does not have much meaning.



Interpretation

Estimated delivery time = 0.03*Distance in km - 1.43

Examples:

1. For a distance of 350km, we expect delivery to take:

   0.029796*350 - 1.431176 = 8.997 days.

2. For a distance of 480km, we expect delivery to take:

   0.029796*480 - 1.431176 = 12.87 days.