THE MATH BEHIND LINEAR REGRESSION

By *Clutch Data Science*

The simple linear regression model:

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{1}$$

We use $\hat{\beta}_0, \hat{\beta}_1$ as the estimator of β_0, β_1. To find the best line that approximate the linear relartionship, we use the Least Squared method, finding the $\hat{\beta}_0, \hat{\beta}_1$ that minimize the sum of squared residuals (SSR).

$$SSR = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \tag{2}$$

1. THE CLOSED FORM SOLUTION FOR $\hat{\beta}_0$ AND $\hat{\beta}_1$

To derive the solution, we need to take the partial derivative w.r.t $\hat{\beta}_0$ and $\hat{\beta}_1$

$$\frac{\partial SSR}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \tag{3}$$

$$\frac{\partial SSR}{\partial \beta_1} = -2\sum_{i=1}^{n}x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \tag{4}$$

To get $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the SSR, we Set (3), (4) to 0. Then we can derive:

$$\bar{y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x} = 0 \tag{5}$$

$$\sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0 \tag{6}$$

From (5) we have:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{7}$$

Substitute (7) into (6):

$$
\begin{aligned}
\sum x_i y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) \sum x_i - \hat{\beta}_1 \sum x_i^2 &= 0 \\
\hat{\beta}_1 (\bar{x} \sum x_i - \sum x_i^2) &= - \sum x_i y_i + \bar{y} \sum x_i \\
\hat{\beta}_1 \sum x_i (\bar{x} - x_i) &= \sum x_i (\bar{y} - y_i) \\
\hat{\beta}_1 &= \frac{\sum x_i (\bar{y} - y_i)}{\sum x_i (\bar{x} - x_i)}
\end{aligned}
$$

Furthermore, we normally write $\hat{\beta}_1$ in the following format:

$$
\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})(x_i - \bar{x})} \tag{8}
$$

Because $\sum \bar{x}(y_i - \bar{y}) = \bar{x} \sum (y_i - \bar{y}) = 0$ and $\sum \bar{x}(x_i - \bar{x}) = \bar{x} \sum (x_i - \bar{x}) = 0$

Therefore, (7) and (8) gives the mathematical solution for $\hat{\beta}_0$ and $\hat{\beta}_1$.

2. THE VARIANCE OF $\hat{\beta}_0$ AND $\hat{\beta}_1$

From (7), (8) we have:

$$
\begin{aligned}
\mathrm{Var}\,(\hat{\beta}_0) &= \mathrm{Var}\,(\bar{y} - \hat{\beta}_1 \bar{x}) \\
&= \mathrm{Var}\,(\bar{y}) + \bar{x}^2 \mathrm{Var}\,(\hat{\beta}_1) - 2\bar{x} \mathrm{Cov}\,(\bar{y}, \hat{\beta}_1)
\end{aligned}
$$

$$
\mathrm{Var}\,(\bar{y}) = \mathrm{Var}\,(\frac{1}{n} \sum y) = \frac{1}{n^2} \sum \mathrm{Var}\,(y) = \frac{\sigma^2}{n}
$$

$$
\mathrm{Var}\,(\hat{\beta}_1) = \frac{1}{[\sum (x_i - \bar{x})^2]^2} \sum (x_i - \bar{x})^2 \mathrm{Var}\,(y_i) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}
$$

The covariance term is:

$$
\begin{aligned}
\mathrm{Cov}\,(\bar{y}, \hat{\beta}_1) &= \mathrm{Cov}\,(\frac{1}{n} \sum y_i, \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2}) \\
&= \frac{1}{n} \frac{1}{\sum (x_i - \bar{x})^2} \mathrm{Cov}\,(\sum_i y_i, \sum_j (x_j - \bar{x}) y_j) \\
&= \frac{1}{n \sum (x_i - \bar{x})^2} \sum_j (x_j - \bar{x}) \mathrm{Cov}\,(\sum_i y_i, y_j) \\
&= \frac{1}{n \sum (x_i - \bar{x})^2} \sum_j (x_j - \bar{x}) \sum_i \mathrm{Cov}\,(y_i, y_j) \\
&= \frac{1}{n \sum (x_i - \bar{x})^2} \sum_j (x_j - \bar{x}) n \sigma^2 \\
&= \frac{\sigma^2}{\sum (x_i - \bar{x})^2} \sum_j (x_j - \bar{x}) \\
&= 0
\end{aligned}
$$

Therefore, we have:

$$\mathrm{Var}\,(\beta_0) = \sigma^2 (\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}) \tag{9}$$

$$\mathrm{Var}\,(\beta_1) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2} \tag{10}$$