Before getting into Linear Regression- “True regression functions are never linear” but it is still a useful tool for predicting a quantitative response.
Topics: Linear Regresion Model concepts and Least Squares Approach.
Important questions that are important to address from a given data:

1. ***Is there a relationship between the variable and the response?*** Determine whether the data providence of an association between the variable and the response.

2. ***How strong is the relationship between the variable and the corresponding responses?*** Strength of this relationship.

3. ***Association of the response with the variable:*** The need to separate out the individual contribution of each variable.

4. ***How large is this association?***

5. ***Predicting the future and the accuracy of the same***

6. ***Is the relationship linear?*** If the relationship between the variable and the response is approximately a straight-line relationship then linear regression is an appropriate tool

7. Synergy effect (in marketing) or interaction effect (in statistics)

Simple Linear Regression

For a quantitative response Y on the basis of a single predictor variable X, we assume that there is approximately a linear relationship between X and Y. We can mathematically write this relationship as:

$$Y \approx \beta_0 + \beta_1 X$$

$\beta_0 \rightarrow$ **intercept** of the linear model
$\beta_1 \rightarrow$ **slope** of the linear model

They are known as the **model coefficients** or just **paraneters**. We will then use our training data and determine the *estimates* of $\hat{\beta}_0$ and $\hat{\beta}_1$ for the model coefficients.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \qquad \text{(estimates)}$$

where $\hat{y}$ indicates a prediction of Y on the basis of $X = x$

Estimating the Coefficients

β_0 and β_1 are unknown, we'll use our data to estimate the coefficients. Our data is:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

Our goal is to find the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ such that the linear model fits the available data well that is to say $y_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i$ for i = 1,2,. . . ,n
Our task is to measure the *closeness*. The most common approach involves **minimizing the least squares criterion**
We define e_i as the residual

$$e_i = y_i - \hat{y}_1$$

As evident, the residual is basically the difference between the observed i-th value and the i-th response value that is predicter by our linear model. We define the **residual sum of squares RSS** as

$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2$$

The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS so using some calculus, we obtain the following *least squares coefficient estimates* for the linear regression:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}\,\overline{x}$$

where $\overline{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\overline{x} \equiv \sum_{i=1}^{n} x_i$ are the sample means.

Assessing the Accuracy of the Coefficient Estimates

Assuming the *true* relationship between X and Y takes the form $Y = f(X) + \varepsilon$, here ε is a **mean-random error term**. Approximating f as a linear function we have:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

here, β_0 is the intercept term- expected value of Y when $X = 0$ and β_1 is the slope- the average increase in Y associated with a one-unit increase in X

The analogy between linear regression and estimation of the mean of a random variable is an apt based on the concept of **bias**. If we use the sample mean $\hat{\mu}$ to estimate μ, this estimate is *unbiased* since $\hat{\mu} = \mu$ is expected by us. $\hat{\mu}$ might *underestimate* or *overestimate* the value of μ but if we could average a huge number of sets of observations, then this average would *exactly* equal μ. Hence, an unbiased estimator does NOT systematically over- or under-estimate the true parameter.
The same could be said for estimating the values of β_0 and β_1, if we could average the estimates obtained over a huge number of data sets, then the average of these estimates would be spot on!

How far off will the estimate of $\hat{\mu}$ will be from μ:

$$\mathrm{Var}(\hat{\mu}) = \mathrm{SE}(\hat{\mu})^2 = \frac{\sigma^2}{n} \tag{Standard error}$$

The standard error tells us the average amount that this estimate of $\hat{\mu}$ differs from the actual value μ For computing the standard errors associated with $\hat{\beta}_0$ and $\hat{\beta}_1$,

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right]$$

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

σ^2 is the variance of the noise or $\sigma^2 = \mathrm{Var}(\varepsilon)$. Assuming that the errors ε_i for each observation have common variance σ^2 and are uncorrelated
In general σ^2 is not known and has to be estimated from the data. This estimate of σ is known as the *residual standard error* and is given by the formula

$$\mathrm{RSE} = \sqrt{\frac{\mathrm{RSS}}{(n-2)}}$$

Standard errors can be used to compute confidence intervals.

***