Machine Learning

Samatrix Consulting Pvt Ltd

Linear Regression

Why Regression?

- In this chapter, we will study a very simple approach in supervised learning, linear regression.
- Linear regression is one of the most useful and widely used tools for predicting the quantitative response.
- Many advanced machine learning models are generalizations or extensions of linear regression.
- However, it is very important to have a good understanding of linear regression before studying complex models.

Important Questions

- In the previous chapter, we discussed the case study based on Advertising data.
- Supposed the client wants us to make the recommendations about the marketing plan for next year that can lead to more sales.
- What information do we need to make our recommendation?
- We may ask a few important questions.

Important Questions

- Do we see a relationship between advertising budget and sales?
  - We should be able to explore whether there is evidence of a relationship between expenditure on advertisement and sales.
  - If we find that there is a weak relationship between advertisement expenditure and sales, we should reconsider our decision of spending money on advertisement.
- What is the strength of the relationship between the advertisement budget and sales?
  - If the relationship exists between the advertisement budget and sales, we would like to understand the strength of the relationship.
  - If an increase in spending increases the sales significantly, the relationship is strong else the relationship is weak.

Important Questions

- Which advertisement medium (out of TV, radio, and newspaper) contributes to the sales?
  - Whether all the three mediums contribute to the sales or there are one or two mediums that contribute to the sales.
  - We need to segregate the impact of the individual mediums on sales.
- Can we estimate the impact of each medium on sales accurately?
  - For each dollar spent on each medium what is the impact on the sales. Whether the sales increase or decreases and by what amount?
- Can we predict the future sale accurately?
- Whether the relationship is linear or more complex?

We can use linear regression to answer all the above questions.

Simple Linear Regression

- Simple linear regression is a simple approach to predict quantitative response Y on the basis of a single input variable X.
- It is based on the assumption that the relationship between X and Y is approximately linear.
- The mathematical notation of this linear relationship is

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

- The symbol $\approx$ means “is approximately modeled as”.
- We can say that we are regressing Y on X (or Y onto X).

Simple Linear Regression

- In our example, if X represents TV advertisement and Y represents sales, we can regress sales onto TV by fitting the model.

$$Sales \approx \beta_0 + \beta_1 \times TV$$

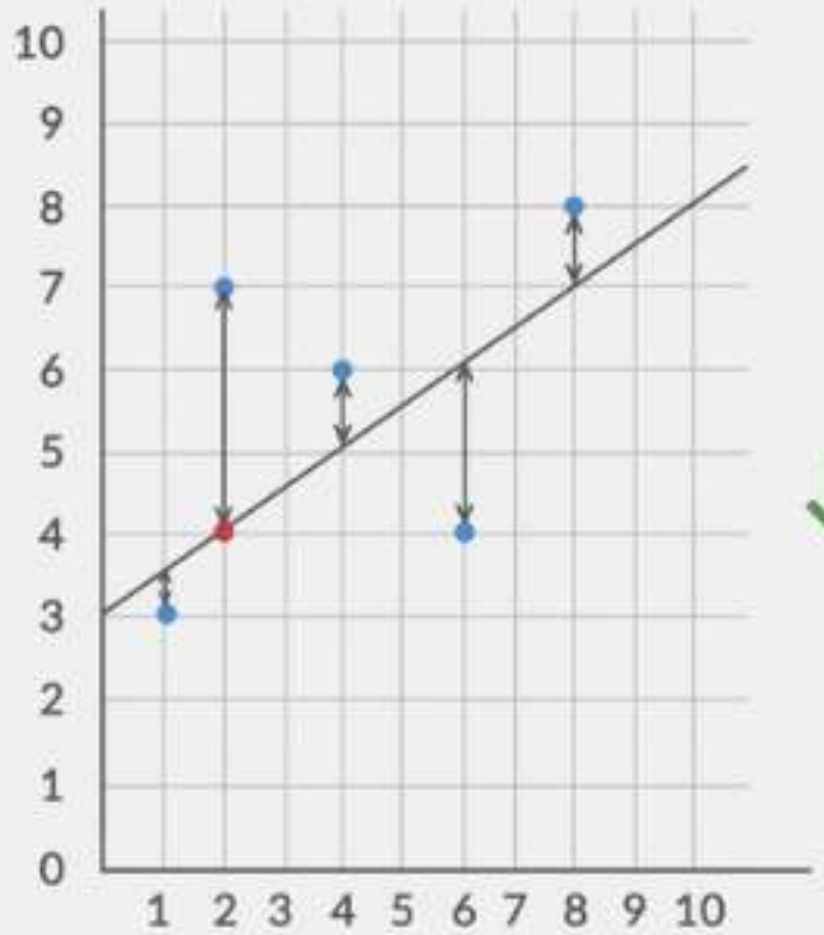
- In this equation, the β_0 and β_1 are unknown constants. β_0 represents the intercept and β_1 represents the slope terms in the lines model.
- Both the constants are known as model parameters or coefficients. We use our training data to estimate $\hat{\beta}_0$ and $\hat{\beta}_1$.

Simple Linear Regression

- We can predict the future sales given the value of expenditure on TV advertising from the following:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

- $\hat{y}$ indicates the prediction of Y given $X = x$.
- The estimated value or the predicted value of an unknown parameter or coefficient is denoted by a hat, ^, symbol.



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

Intercept Slope

$$e_i = y_i - y_{pred}$$

Ordinary Least Squares Method:

$$e_1^2 + e_2^2 + \_\_\_ + e_n^2 = \text{RSS (Residual Sum Of Squares)}$$

$$RSS = (Y_1-\beta_0-\beta_1 X_1)^2 + (Y_2-\beta_0-\beta_1 X_2)^2 + \_\_\_\_ + (Y_n-\beta_0-\beta_1 X_n)^2$$

$$RSS = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

Estimating the Coefficients

- The coefficients β_0 and β_1 are unknown.
- Before making the predictions, we should use the training data to estimate the coefficients. Let the data set

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

Represent n observation pairs.

- Each observation contains a measurement of X and a measurement of Y.

Estimating the Coefficients

- In the case of Advertising example, the data set consists of the TV advertising budget and product sales in $n = 200$ different markets.
- We would like to estimate the coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ so that the linear model can fit the data well. In other words, $y_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i$ for $i = 1, \dots, n$.
- We can also say that we want to find the intercept $\hat{\beta}_0$ and a slope $\hat{\beta}_1$ so that the resulting line is as close as possible to the $n = 200$ observations.
- We can measure the closeness by using the least square criterion.

Estimating the Coefficients

- If $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ is the prediction of Y based on the ith value of X. The ith residual is represented by $e_i = y_i - \hat{y}_i$.
- This is the difference between the ith response value that is predicted by our model and ith observed response value.
- We can define the **residual sum of square** as

$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2$$

$$RSS = \left(y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1\right)^2 + \left(y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2\right)^2 + \cdots + \left(y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n\right)^2$$

Estimating the Coefficients

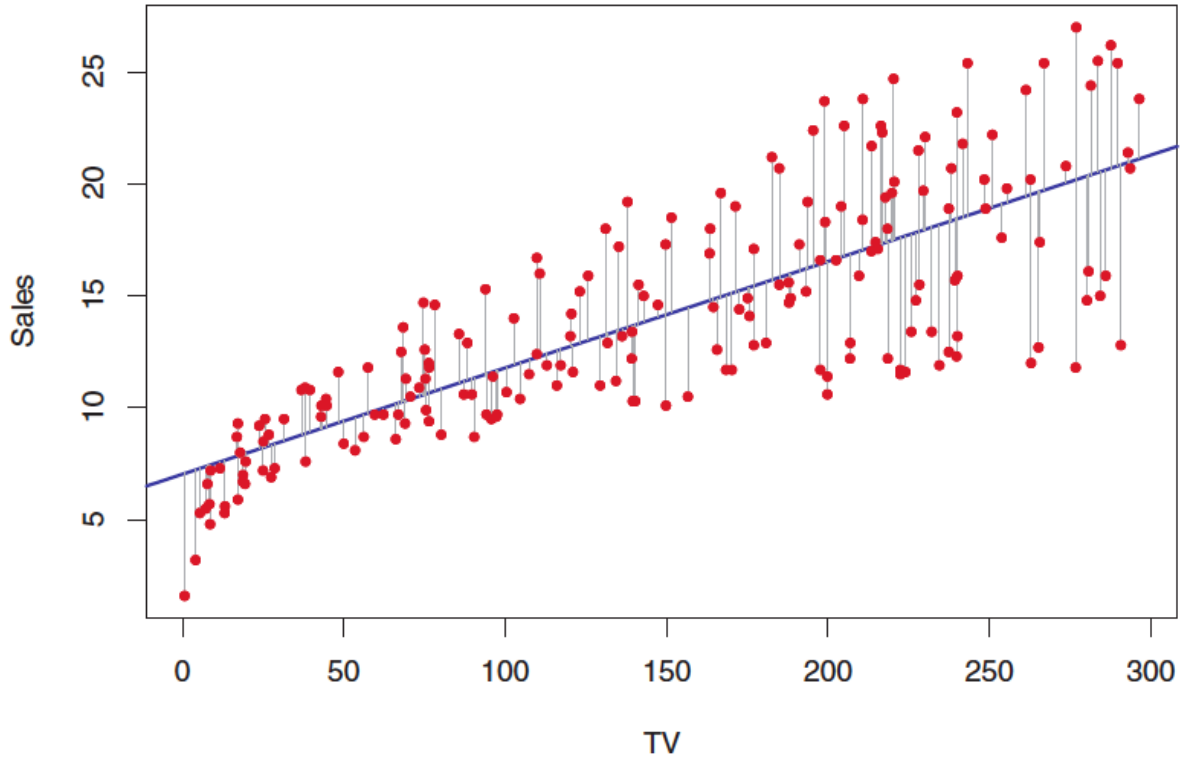
- For the least square method we choose the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ so that RSS is minimized. We can use calculus to show that these values are minimized at

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

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

- where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ are the sample mean.
- Hence, we get the **least squares coefficient estimates** for simple linear regression.

Estimating the Coefficients



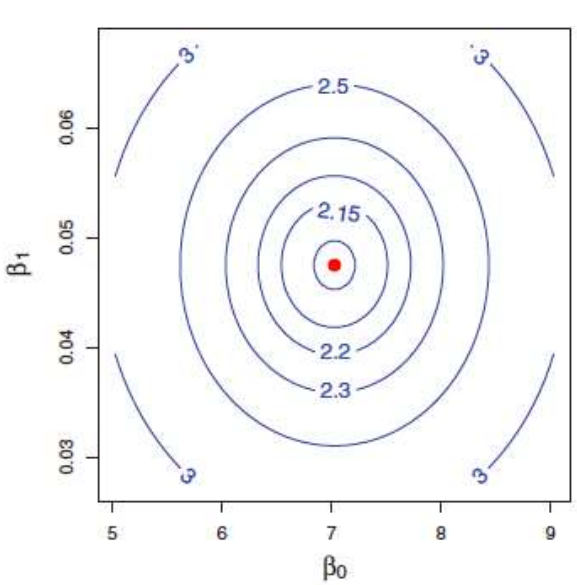
The least square fits the regression of sales to TV for the advertising data. The fit minimizes the sum of squared errors. The grey line represents an error. The fit makes a compromise by averaging their squares.

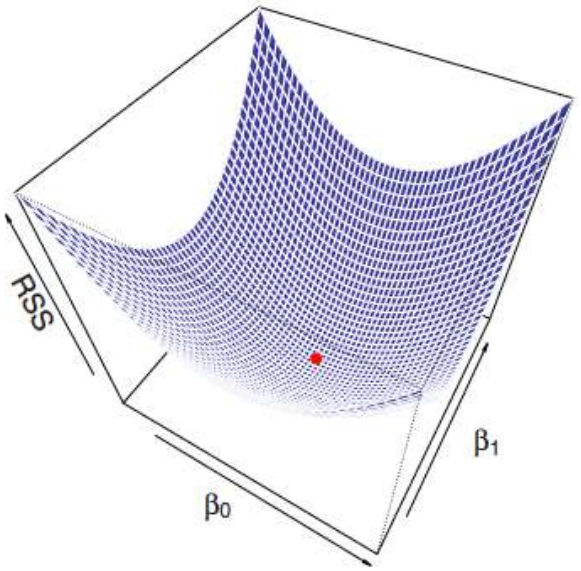
Estimating the Coefficients

- The least square fits the regression of sales to TV for the advertising data. The fit minimizes the sum of squared errors.
- The grey line represents an error.
- The fit makes a compromise by averaging their squares.
- The simple regression fit for advertising data is displayed in Figure 1. In this case, $\hat{\beta}_0 = 7.03$ and $\hat{\beta}_1 = 0.0475$.
- This means that if we spend an additional $\$1000$ on TV advertisements, the sales will increase by 47.5 unit of product.

Estimating the Coefficients

- The Figure 2 shows the RSS value for a number of values of β_0 and β_1 using the advertising data with the sales as the response and the TV as the predictor.
- The red dot represents the pair of the least square estimates $(\hat{\beta}_0, \hat{\beta}_1)$





Accuracy of Coefficient Estimates

Population Regression Line

- We have studied that the true relationship between X and Y is

$$Y = f(X) + \epsilon$$

- In this equation f is some unknown function and ϵ is a mean-zero random error term.
- If we approximate function f by a linear function, the relationship can be written as

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

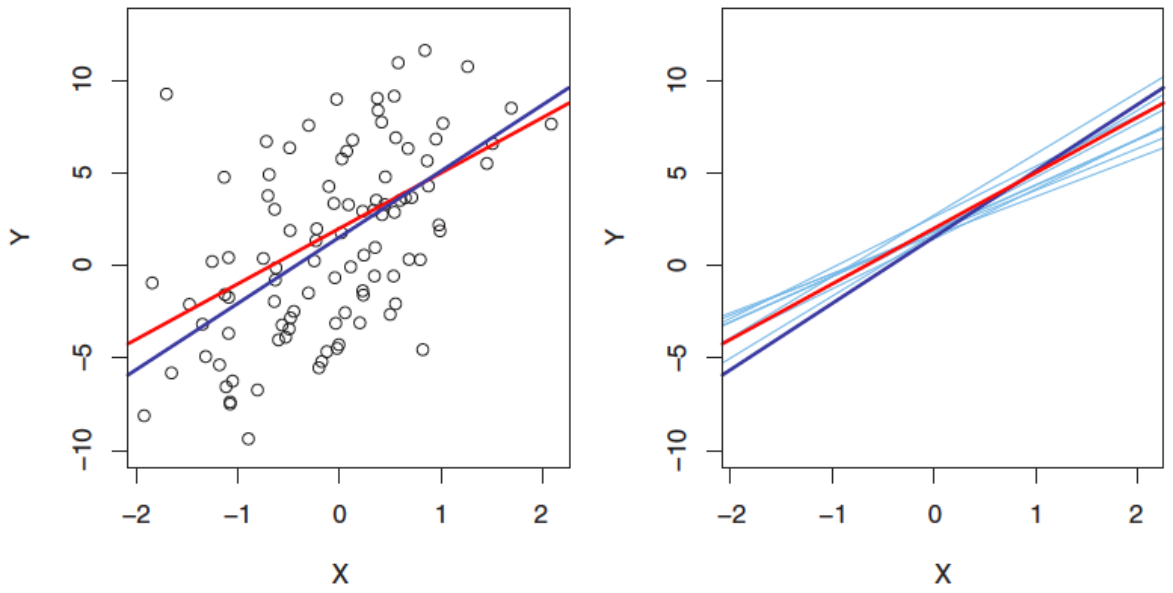
- Here β_0 is the intercept that is the expected value of Y when $X = 0$. β_1 is the slope that is expected to increase in Y with a one-unit increase in X.
- In this simple model, the error term is independent of X.

Population Regression Line

- The model given by $Y = \beta_0 + \beta_1 X + \epsilon$ defines the population regression line which is the best linear approximation to the true relationship between X and Y.
- The least square estimates given by $(\hat{\beta}_0, \hat{\beta}_1)$ characterize the least square line.

Population Regression Line

- LEFT - For a simulated data set, the red line represents the true relationship, $f(X) = 2 + 3X$. This is also known as the population regression line. The blue line is the least-squares estimate for $f(X)$ that is based on the observed data as shown in black.
- RIGHT – The population regression line is once again shown in red and the least square line in dark blue. The ten light blue lines show ten least-square lines that have been computed based on a separate set of observations. Each of the least-squares lines is different however the least square lines are very close to the population regression line.



Standard Error

- It means that if we estimate (β_0, β_1) based on a particular dataset, our estimates would not be exactly equal to (β_0, β_1).
- However, if we take the average of the estimates from a huge number of data set, then our average would be equal to (β_0, β_1).
- But a single estimate of the parameters may be an underestimate or overestimate of (β_0, β_1).
- But the question is how far off will the single estimate be?
- We can use **standard error** to answer these questions.
- For example, if we are estimating population mean μ using the estimator $\hat{\mu}$, the standard error of $\hat{\mu}$, $SE(\hat{\mu})$ would be

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

Standard Error

- The standard error tells us the average amount that the estimate $\hat{\mu}$ differs from the actual value of μ.
- The equation also tells us that with an increase in the number of observations n, the standard error reduces.
- Similarly, we need to know, how close $\hat{\beta}_0$ and $\hat{\beta}_1$ are to the true values of β_0 and β_1.
- The standard errors associated with $\hat{\beta}_0$ and $\hat{\beta}_1$ can be given by the following formulas:

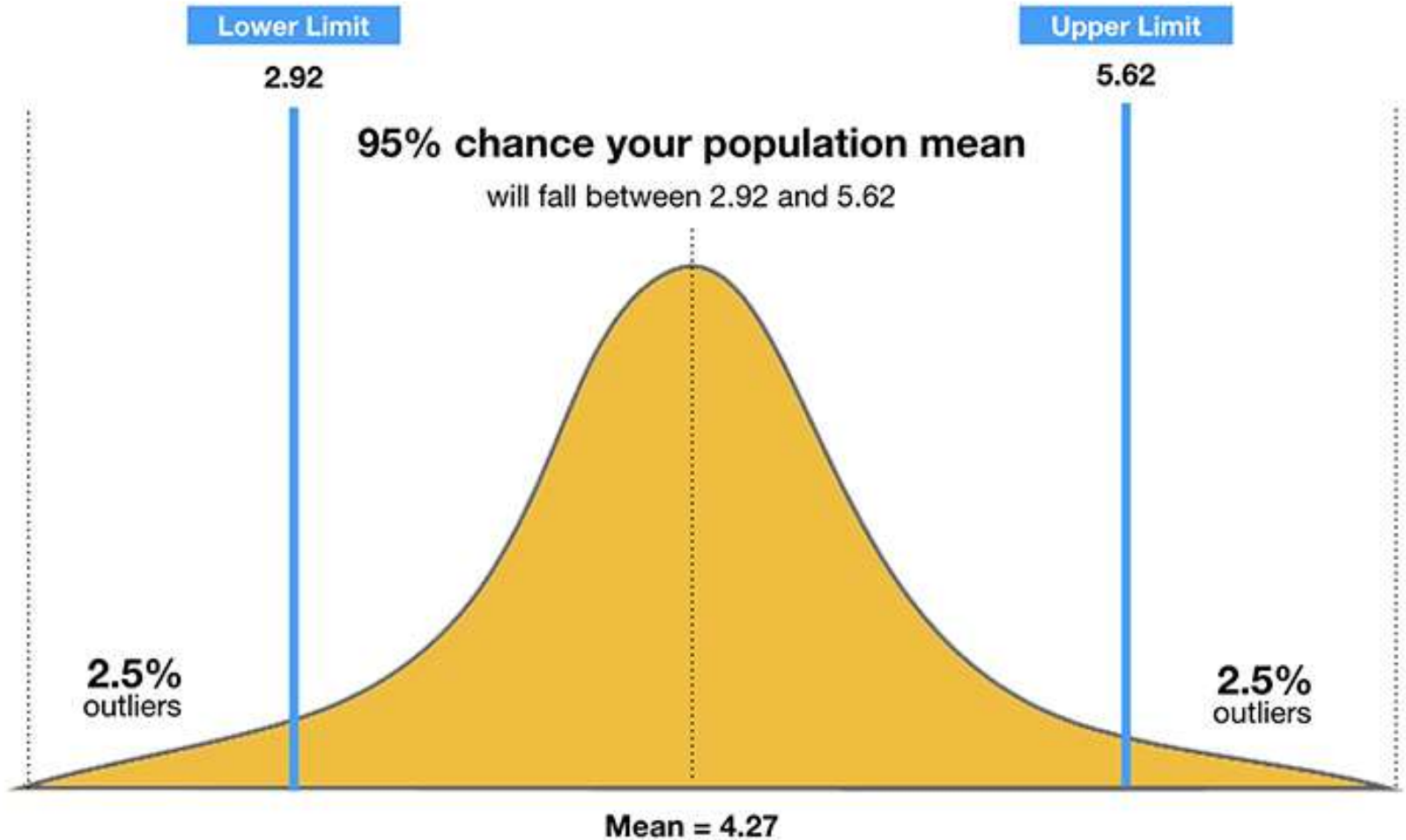
$$SE(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$

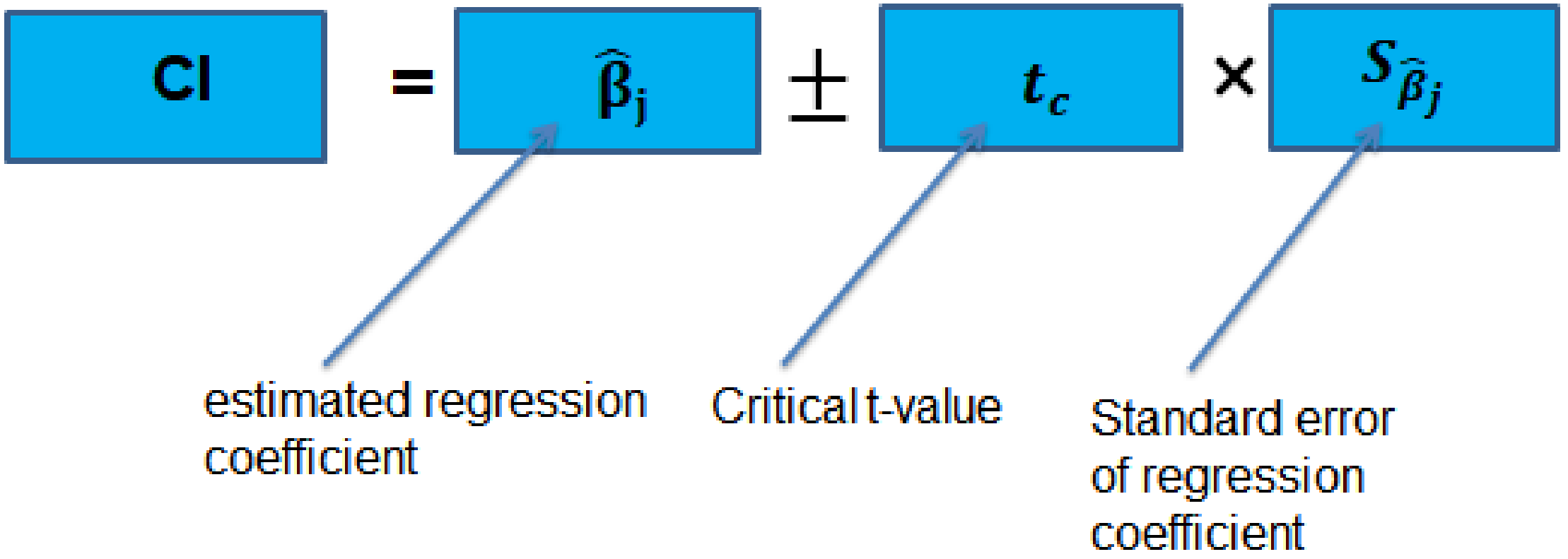
$$SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Residual Standard Error

- In general, σ^2 is not known, but can be estimated from the data. The estimate of σ is known as the **residual standard error**, and is given by the formula

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


Lower Limit
2.92
Upper Limit
5.62
95% chance your population mean
will fall between 2.92 and 5.62
2.5%
outliers
2.5%
outliers
Mean = 4.27


CI
=
$\hat{\beta}_j$
±
t_c
×
$S_{\hat{\beta}_j}$
estimated regression coefficient
Critical t-value
Standard error of regression coefficient

Confidence Interval

- We can use standard errors to compute confidence intervals. A 95% confidence interval can be defined as a range of values such that with 95% probability, the range will contain the true value of the parameter. This range is defined in the terms of the lower and upper limits computed from the sample of data.
- For the linear regression, the 95% confidence interval of β_1 takes the following form

$$\hat{\beta}_1 \pm 2 \times SE(\hat{\beta}_1)$$

- That is there is approximately a 95% chance that the interval

$$\left[\hat{\beta}_1 - 2 \times SE\left(\hat{\beta}_1\right), \hat{\beta}_1 + 2 \times SE\left(\hat{\beta}_1\right)\right]$$

- will contain the true value of β_1.
- Similarly, the confidence interval for β_0 takes the form

$$\hat{\beta}_0 \pm 2 \times SE(\hat{\beta}_0)$$

Hypothesis Testing

- We can also use the standard errors to perform hypothesis tests on the coefficients. The null hypothesis is

$$H_0 : There\ is\ no\ relationship\ between\ X\ and\ Y$$

- The alternative hypothesis

$$H_a : There\ is\ some\ relationship\ between\ X\ and\ Y$$

- This corresponds to testing

$$H_0 : \beta_1 = 0$$

- Versus

$$H_a : \beta_1 \neq 0$$

Hypothesis Testing

- If $\beta_1 = 0$, the model becomes $Y = \beta_0 + \epsilon$ and we can say that X is not associated with Y.
- For testing the null hypothesis, we explore whether $\hat{\beta}_1$ is sufficiently far off from zero so that we can be confident that β_1 is non-zero.
- But the question is how far it should be?
- This depends on the $SE(\hat{\beta}_1)$ that is the accuracy of $\hat{\beta}_1$.
- If the value of $SE(\hat{\beta}_1)$ is small, even the small values of $\hat{\beta}_1$ will suggest that $\beta_1 \neq 0$ and hence we can conclude there some relationship exists between X and Y.

Hypothesis Testing

- On the other hand, for large value of $SE(\hat{\beta}_1)$, the $\hat{\beta}_1$ should be large enough so that we can reject the null hypothesis.
- We compute t-statistic as

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

- That measure the number of standard deviations that $\hat{\beta}_1$ is away from 0.
- If there is no relationship between $X\ and\ Y$, then the equation will take t-distribution with $n - 2$ degree of freedom.

Hypothesis Testing

- We can compute the probability of observing the value equal to $|t|$ or larger by assuming $\beta_1 = 0$. We call the probability the $p - value$.
- A small p-value indicates that it is unlikely to observe such a substantial association between the predictor and the response due to chance, in the absence of any real association between the predictor and the response.
- Hence, if we see a small p-value then we can infer that there is an association between the predictor and the response.
- We reject the null hypothesis—that is, we declare a relationship to exist between X and Y —if the p-value is small enough.
- Typical p-value cutoffs for rejecting the null hypothesis are 5 or 1%

Accuracy of Model

Accessing the Accuracy of Model

- Once the null hypothesis is rejected in the favor of the alternative hypothesis, we would like to quantify the extent to which the model fits the data.
- We access the quality of the linear regression fit using residual standard error (RSE) and R^2 statistics

Residual Squared Error

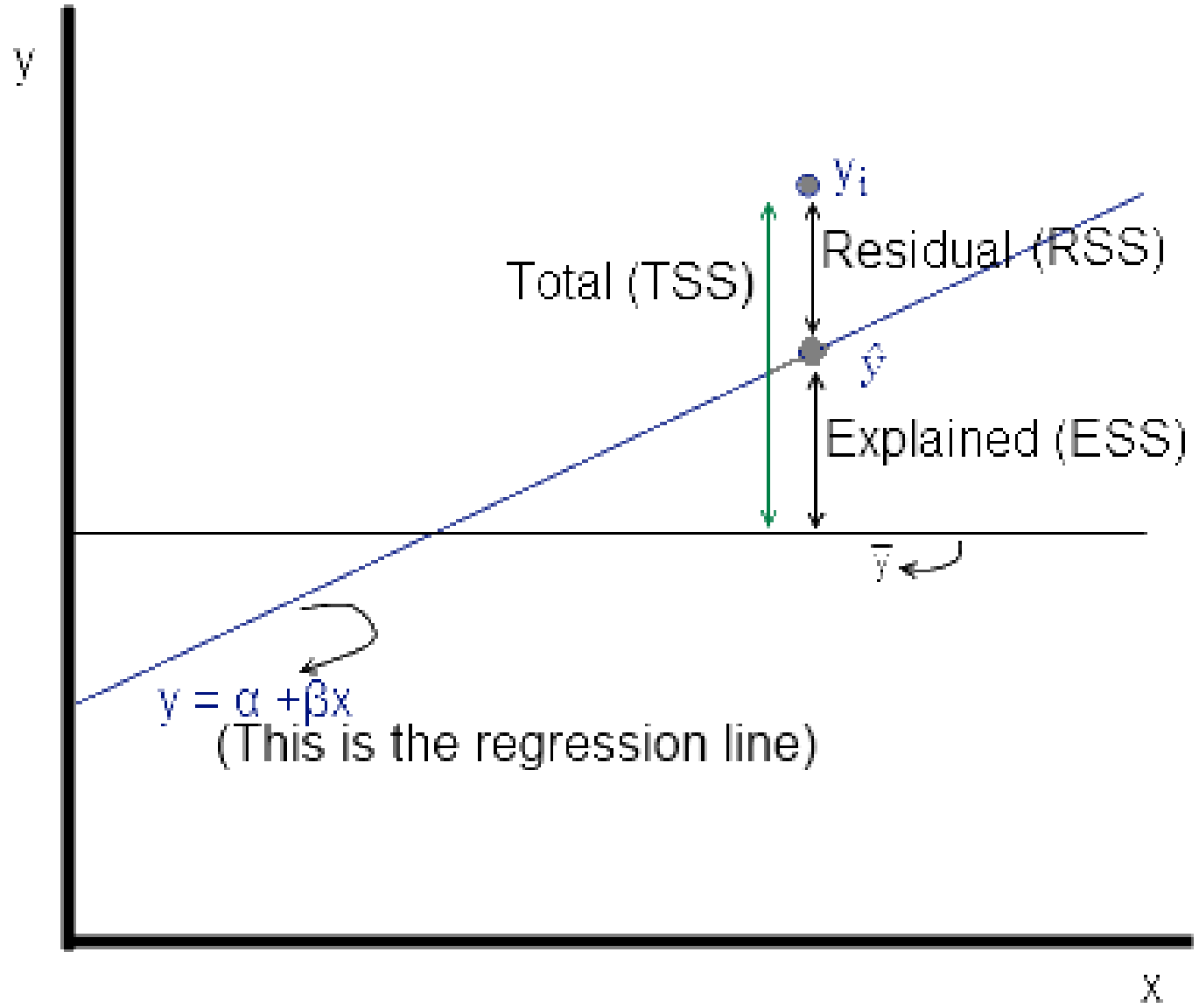
- We have seen that an error term ϵ is associated with each observation.
- Due to these errors, even if we know the true regression line, we will not be able to predict Y from X perfectly.
- We use residual standard error (RSE) to the estimate the standard deviation of ϵ.
- RSE gives an average amount that the response will deviate from the true regression line.

Residual Squared Error

- We compute RSE using the formula

$$RSE = \sqrt{\frac{1}{n-2}RSS} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- Where Residual Sum Square $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.
- From RSE we can measure the lack of the fit of the model to the data.
- If the predicted values are very close to the true outcome values, then the RSE will be small. In this case we can conclude that the model fits the data very well.
- Similarly, if the predicted values are far from the true outcome values, we may conclude that the model doesn't fit the data well.



$\hat{y}$ is the predicted value of y given x, using the equation $y = \alpha + \beta x$.

y_i is the actual observed value of y

$\bar{y}$ is the mean of y.

The distances that RSS, ESS and TSS represent are shown in the diagram to the left - but remember that the actual calculations are squares of these distances.

$$TSS = \Sigma (y_i - \bar{y})^2$$

$$RSS = \Sigma (y_i - \hat{y})^2$$

$$ESS = \Sigma (\hat{y} - \bar{y})^2$$

R Squared Statistics

- From RSE, we get an absolute measure of lack of fit of the model to the data.
- But it is measured in the units of Y. Hence, we cannot make out whether the RSE that we got is a good RSE.
- We can use R^2 statistics as an alternative measure of fit. R^2 provides the proportion of the variance explained hence it always returns the value between 0 and 1 and it is independent of the scale of Y.
- We can calculate R^2 as follows

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

R Squared Statistics

- The **Total sum of Square** $TSS = (y_i - \bar{y})^2$. We have already defined RSS.
- Using TSS, we can measure the total variance in the response Y.
- The R^2 statistics that is close to 1 indicates that a large proportion of the variability in the response has been explained by the regression.
- A number near 0 indicates that the regression did not explain much of the variability in the response.

Thanks

Samatrix Consulting Pvt Ltd

Samatrix.io