Simple Linear Regression

- Useful for Predicting quantitative Response
- Predicting Y on the basis of a single predictor variable X
- Assuming that there is a linear relationship between X and Y

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

- This can be read as regressing Y on X
- or Y onto X
  Example :
- TV ads $\rightarrow X$
- Sales $\rightarrow Y$
- Sales $\approx \beta_0 + \beta_1 TV$
  - β_0,β_1 two unknown **constants**
  - $\beta_0 \rightarrow$ Slop of X
  - $\beta_1 \rightarrow$ intercept of Y
  - They are called model **Coefficients** or **Parameters**

After Using Training data to estimate $\hat{\beta}_0, \hat{\beta}_1$

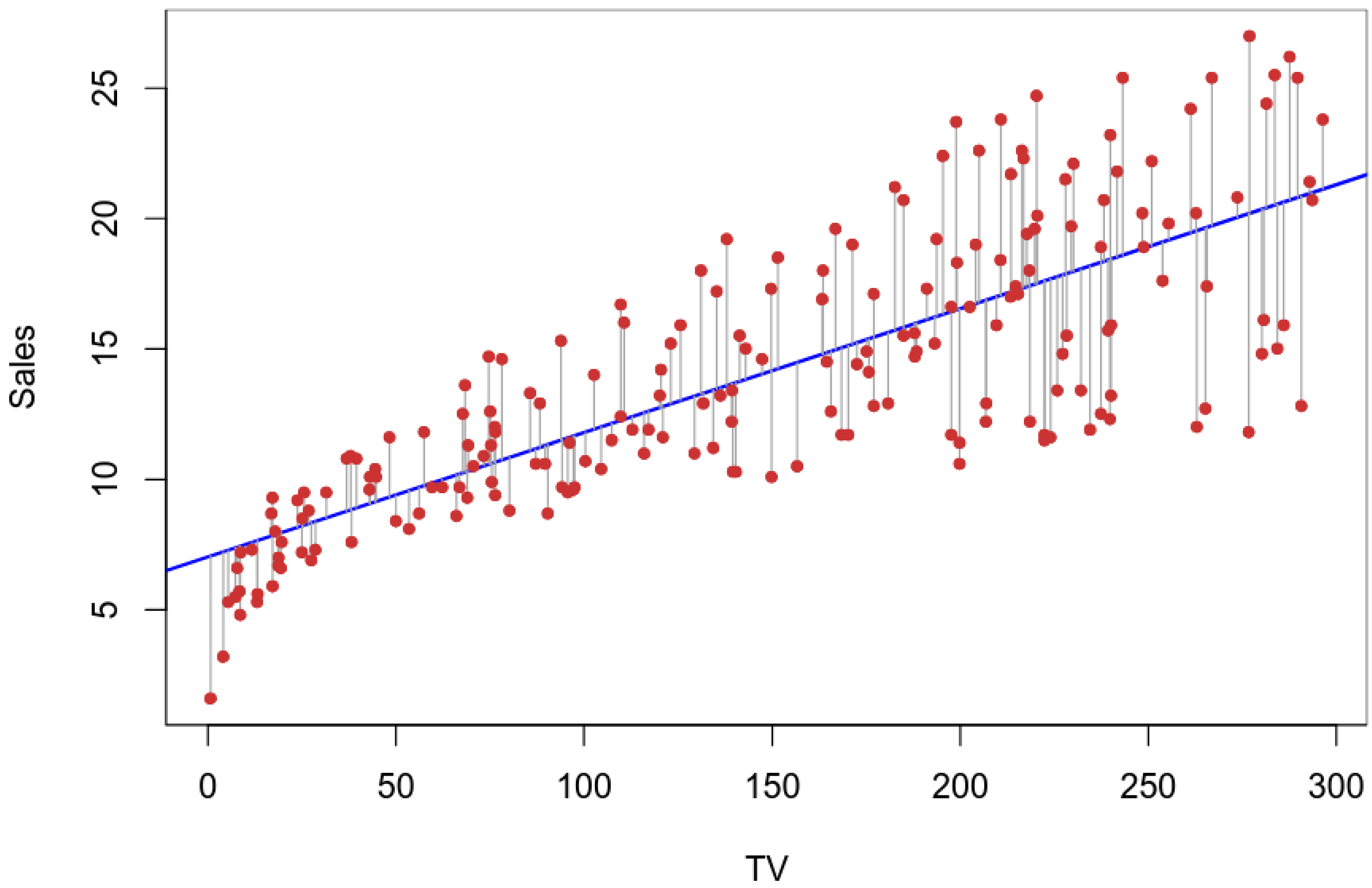
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Estimating the Coefficients

- In practice β_0, β_1 are unknown, we usually useTraining Data to estimate them :$(x_1, y_1), \ldots, (x_n, y_n)$
- we try to estimate β_0, β_1 as close as possible to the data points so:
  - $y_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i$
- To minimize as much as possible $\rightarrow$ we use Least squares criterion

Let $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i \rightarrow e_i = y_i - \hat{y}_i$

- y_i observed response $\rightarrow$ true values in Training Data
- $\hat{y}_i$ predicted response $\rightarrow$ from the regression line
  With that we have Residual Sum of Squares (RSS)



- The Blue line fit is found by the lease squares
- Minimizing the residual sum of squares
    - Minimizing both β_0, β_1 Ordinary Least Squares
- Every grey line is the residual of $y_i - \hat{y}_i$

Assessing the Accuracy of the Coefficients β_0, β_1

1. Standard Error $SE(\hat{\beta}_1)$
2. Confidence Interval
3. Hypothesis Testing

Standard Error $\hat{SE}(\hat{\beta}_1)$

- We assumed that the true relationship between X and Y is on form :

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

- If f is approximated to be linear then we can write the relationship as:

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

    - $\beta_1 \rightarrow$ The average increase in Y associated with one unite increase in X
- The population mean μ is usually unknown so the sample mean $\bar{\mu}$ will provide a good estimate to μ
- The sample mean $\bar{\mu}$ is unbiased $\rightarrow$ it averages out an estimation of huge biased estimations so that the sample mean $\bar{\mu}$ will be as close to the real population mean μ
  Using the same Logic we make multiple estimations for Coefficients $\hat{\beta}_0, \hat{\beta}_1$ and averaging it out will be spot on
- The question now is how a single estimation is far from the mean $\rightarrow \text{variance}$

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

- σ is the standard deviation of each realization y_i

In the same tone we can see how close $\hat{\beta}_0, \hat{\beta}_1$ are to true values β_0, β_1

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

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

Standard Error Derivation.

- When $\sigma^2 = \mathrm{Var}(\varepsilon)$ These formulas are valid (With uncorrelated error)
- $\hat{\beta}_0$ would equal the $\mathrm{Var}(\bar{\mu})$ if $\bar{x} = 0$ which implies $\hat{\beta}_0 = \bar{y}$
- σ^2 usually unknown so we estimate it from the data
  - Known as Residual Standard Error

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

  - $(n-2)$ Degrees of Freedom (Fixing the Slop,intercept)

Confidence Interval For Coefficient Estimates $\hat{\beta}_0, \hat{\beta}_1$

Standard Error can be used to compute Confidence interval, in 95% **Confidence Interval** the range of values such that with 95% probability the range will contain the true value of the Estimates $\hat{\beta}_0, \hat{\beta}_1$:

$$\hat{\beta}_0 \pm 2\hat{SE}(\hat{\beta}_0) \implies [\hat{\beta}_0 - 2\hat{SE}(\hat{\beta}_0), \hat{\beta}_0 + 2\hat{SE}(\hat{\beta}_0)]$$

$$\hat{\beta}_1 \pm 2\hat{SE}(\hat{\beta}_1) \implies [\hat{\beta}_1 - 2\hat{SE}(\hat{\beta}_1), \hat{\beta}_1 + 2\hat{SE}(\hat{\beta}_1)]$$

Example :

- If $\mathrm{Sales} \approx \beta_0 + \beta_1 \mathrm{TV}$
- If $\mathrm{TV} = 0$ no money spent on TV ads
- Then we are 95% Confidence that the $\mathrm{Sales} \approx \beta_0$ with $\beta_0 \in [\hat{\beta}_0 - 2\hat{SE}(\hat{\beta}_0), \hat{\beta}_0 + 2\hat{SE}(\hat{\beta}_0)$

Hypothesis Testing For Coefficients Estimates $\hat{\beta}_0, \hat{\beta}_1$

Standard Error can be used to perform **Hypothesis Testing** on the Coefficients $\hat{\beta}_0, \hat{\beta}_1$

- Let **Null Hypothesis** be
  $H_0 : \text{There is no relationship between X and Y } \rightarrow \beta_1 = 1$
- And the **Alternative Hypothesis** be $H_a : \text{There is some relationship between X and Y} \rightarrow \hat{\beta}_1 \neq 0$

If the **Null Hypothesis** is true $Y = \beta_0 + \varepsilon$ which means that X is not Associated with Y

- If the $\hat{SE}(\hat{\beta}_1)$ is small even small values of $\hat{\beta}_1$ may provide strong evidence to **reject the Null Hypothesis**
- if the $\hat{SE}(\hat{\beta}_1)$ is large, then $\hat{\beta}_1$ must be large enough to provide a strong evidence to reject H_0
- For that we perform a **T-test** or **T-statistic** as follows :

$$t_{n-2} = \frac{\hat{\beta}_1 - \beta_1}{\hat{SE}(\hat{\beta}_1)}$$

  - $n-2$ is the Degrees of Freedom (Since we Estimating β_0, β_1)
  - The estimated Regression Coefficients β_0, β_1 are random variables because they depend on **sample data**
  - We use the **T-statistic** cause if the sample size n is small the t-distribution have fatter tails (more uncertainty when it comes to smaller sample size)
  - As $n \rightarrow \infty$ the t-distribution converges to a normal distribution
    - **Why not Z-test:**
      The Z-test requires knowing the population standard deviation of the error σ^2 which is always unknown and can only be estimated

Now Testing the **Null Hypothesis** : $H_0 : \beta_1 = 0$

$$t_{n-2} = \frac{\hat{\beta}_1 - 0}{\hat{SE}(\hat{\beta}_1)}$$

- We calculate t_{n-2} and look the area corresponding to it $p-value$
- Given we decided The Significance Level before hand usually (5% or 1%)
- $p-value$ Probability that the Null hypothesis is true
- A small $p-value$ indicate its very unlikely it happened due to chance or statistical fluctuation

Assessing The Accuracy of the Model

Quantifying how much the model fit the data or the **Quality** of a Linear Regression fit and its typically assessed using :

1. Residual Standard Error **RSE**
2. R^2 Statistic

Residual Standard Error RSE

Its the Error term ε, even if we knew the true regression line we wont **predict** Y perfectly, RSE is the estimate of the **Standard Deviation of the residuals(errors)** ε

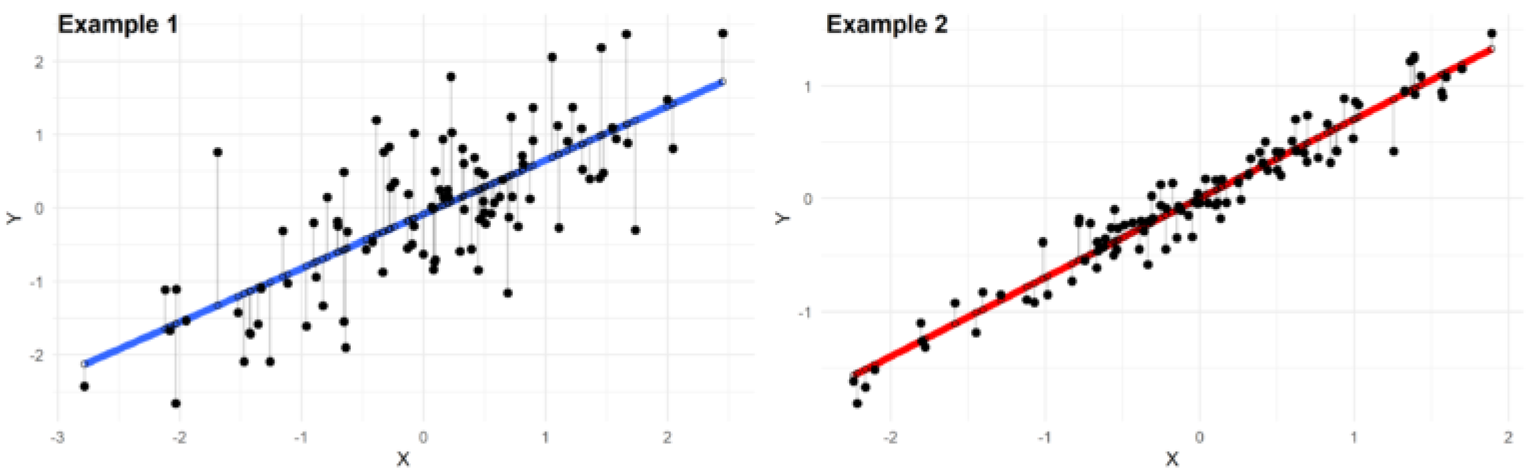
- The average amount the response Y will deviate from the true regression Line
- The RSE measures how well the regression line fits the data, The differences between the observed values y_i and the predicted values of $\hat{y}_i$

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

- RSS is the Residual Sum of Squares
- $n-2$ Degrees of Freedom

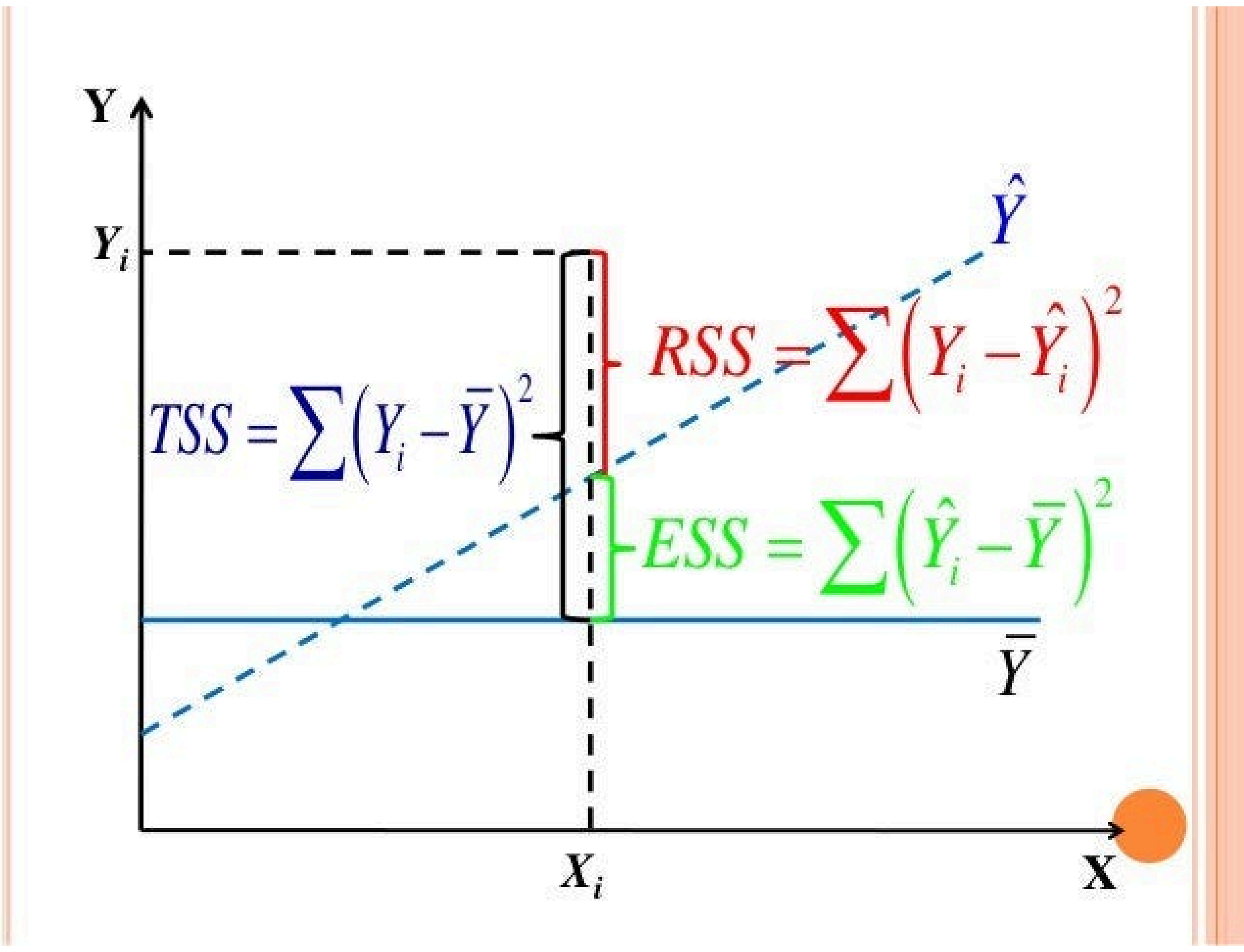
Interpretation of RSE :

- Lower values indicated a tighter fit and less unexplained variability, Which means that our regression line explained most of the variability in the TSS (The model fits the data well)
- Higher values indicates a poor fit of the model
- Can also be use to construct a Prediction interval



TSS **:(Total sum squared)**
Its the variance in the response Y before the regression line is fitted
unlike RSS which measures the amount of variability that is left after the regression line (Unexplained Variance).
TSS is simply the distance between the responses y_i and the mean response $\bar{y}$

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



R^2 Statistic :

R^2 Provides an alternative measure of fit, Unlike RSE which is measures in Y unites and its cant be clear which RSE value is good and also depends on the context of the problem and all

- R^2 provides a measure of proportion of variance/Variability in Y cause its include both TSS and RSS
- The before regression line variability TSS
- The after regression line variability (Unexplained variance)RSS
- It always takes value between 0 and 1

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

Interpretation of R^2:

- Closer to 1 value means the variability in the data points can be explained with regression
- Closer to 0 value means the regression don't explain much of the variability in the data or σ^2 is too high in the response Y
- R^2 is also a measure of the Linear relationship between Y, X , The higher R^2 means that the variation in Y is explained by X