# Multiple Linear Regression

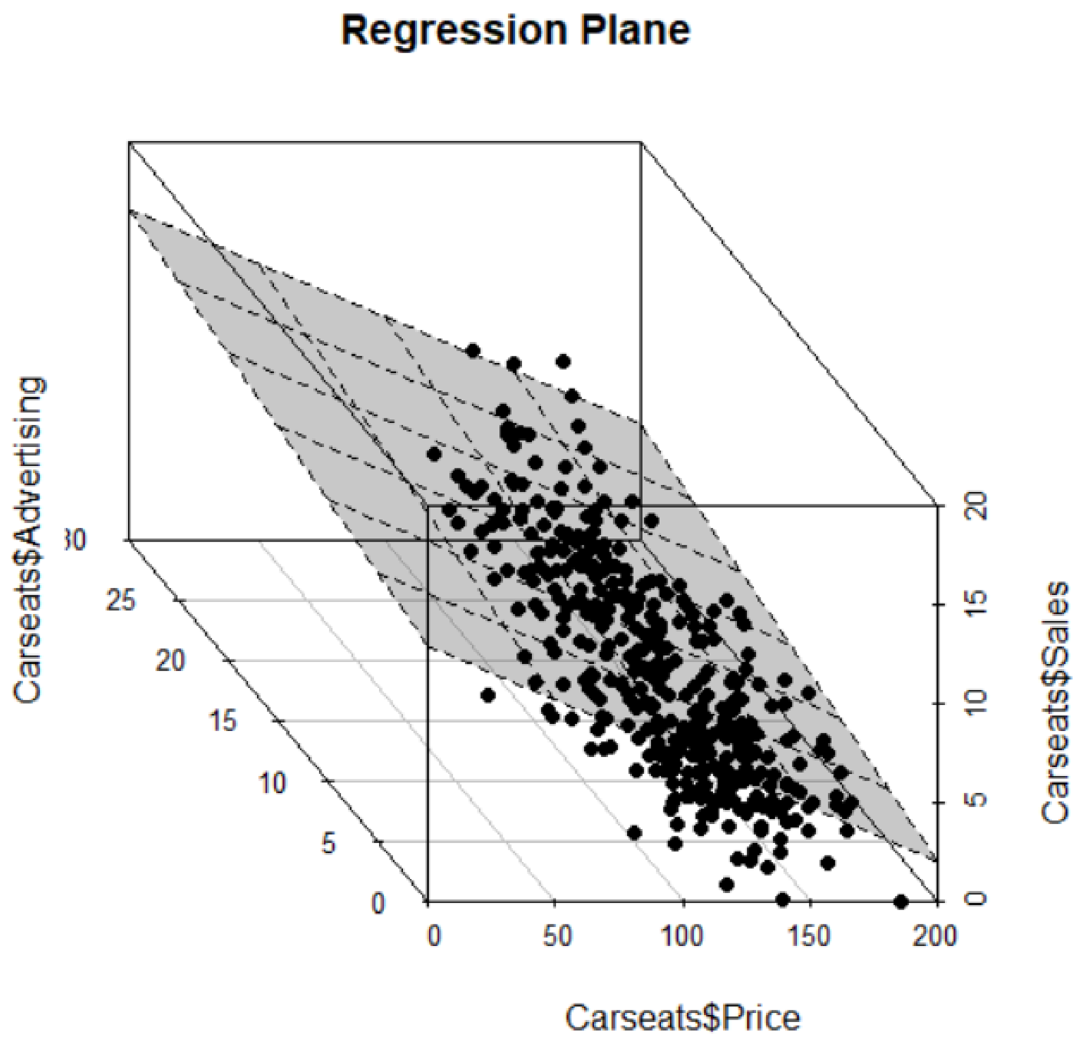
- Simple Linear Regression predict Response for a single predictor $X$, for example $\text{TV}$
- Multiple Linear Regression deals with multiple predictors, even fitting separate Simple regression to each predictor $X$ this will make us miss some key correlations and associations between predictors and Response

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

**For example :**

$$\text{sales} = \beta_0 + \beta_1 TV + \beta_2 Radio + \beta_3 Newspaper$$

## Geometric Interpretation (Regression Plane)




a

- Unlike Simple Linear Regression Multiple Linear regression have multiple predictors which is draw as a hyperplane

## Estimating The Regression Coefficients :

- The Coefficients in the multiple regression are unknown $\beta_0, \beta_1 \ldots, \beta_p$
- So we estimate them as as in the Simple Linear Regression

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

### Multiple Linear Regression Matrix Form :

- The multiple Regression formula can be written in a matrix form making it better to work with and derive the Coefficients
- $$Y = X\beta + \varepsilon$$

  With :
  - $Y \to$ Vector of dependent variables Response
  - $X \to$ Matrix of $n * p$ dimensions + **intercept** $\beta_0$
  - $\beta \to$ Vector of Coefficients (to be estimated)
  - $\varepsilon \to$ Vector of Error terms

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{21} & \dots & X_{p1} \\ 1 & X_{12} & X_{22} & \dots & X_{p2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & X_{2n} & \dots & X_{pn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

- $E(\varepsilon) = 0$
- $var(\varepsilon) = \sigma^2 I_{n*n}$
- $\beta's$ are called partial regression coefficients cause $\beta_1$ is the expected change in $Y$ per Unite change in $X_1$, While holding other $X's$ constant

## Least Squares Estimator :

Multiple Regression often reveal how much a predictor $X_i$ effect the prediction Response $Y$, that the Simple regression don't address

- Due to the slop in the Simple Linear Regression represent the average increase in $Y$ without association with the other predictors
- In Multiple Regression the average increase in $Y$ associated with increasing $X_1$ while holding the others $X$ fixed
- Multiple Regression can suggest a no relationship between $Y$ and a Predictor $X$

Deriving The coefficients estimates Using OLS method Ordinary Least Squares

# Assessing the Accuracy of the Coefficients $\hat{\beta}_p$

Same as in the Simple Linear Regression we use :

1. Standard Error / Variance
2. Confidence intervals
3. Hypothesis testing (F-test)

## Standard Error of $\hat{\beta}_p$

The Standard Error is the square root of its variance of $\hat{\beta}_j$ is how much $\hat{\beta}_j$ will vary from the mean or the expected value of $\hat{\beta}_j$ we found that its unbiased in Standard Error Derivation

$$E[\hat{\beta}] = \beta$$

We also Derived the the Standard Error of $\hat{\beta}$ in Standard Error Derivation and got:

$$\text{Var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

$$SE(\hat{\beta}) = \sigma \sqrt{(X^T X)^{-1}}$$

- $\sigma^2$ is almost unknown in all practical situations
- We use the Sample standard deviation $S^2$
  - $S^2 = \frac{\sum e_i^2}{n-p} = \frac{e^T e}{n-p} = MSE$

## Confidence Interval

- Constructing a how Confident we are on the estimated coefficients $\hat{\beta}_j$
  From Ordinary Least Squares and Standard Error Derivation we know

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

$$\text{Var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1} \ \rightarrow \text{SE}(\hat{\beta}) = \sigma \sqrt{(X^T X)^{-1}}$$

  Since $\sigma^2$ is most of the time :

$$\text{SE}(\hat{\beta}) = S \sqrt{(X^T X)^{-1}}$$

  We construct the following **confidence interval**:

$$t_{n-p} = \frac{\hat{\beta}_j - \beta_j}{SE(\hat{\beta})}$$

- $n - p$ Degrees of Freedom
- $\hat{\beta}_j - \beta_j$ How far our estimate to the real coefficient

$$P(\hat{\beta}_j - t \,.\, \text{SE}(\hat{\beta}_j) < \beta_j < \hat{\beta}_j + t \,.\, \text{SE}(\hat{\beta}_j)) = 1 - \alpha$$

and get :

$$\hat{\beta}_j \pm t_{\frac{\alpha}{2},n-p} \, . \, \mathrm{SE}(\hat{\beta}_j)$$

## Hypothesis Testing (F-Statistics)

- The question asked is : Is there a relationship between the Response and the Predictors $X$
- We check this using the hypothesis Testing

$$H_0 : \beta_1 = \beta_2 = \ldots \beta_p = 0 \text{ There is no realtionship between the predictors and the reponse}$$

$$H_a : \text{alteast one } \beta_j \neq 0$$

- To test this Hypothesis we use $F - statistics$ test

$$F = \frac{\frac{TSS-RSS}{p}}{\frac{RSS}{(n-p-1)}}$$

- $TSS = \sum(y_i - \bar{y})^2 \rightarrow$ total sum squared
- $RSS = \sum(y_i - \hat{y}_i)^2 \rightarrow$ Residual sum squared
- We divide by $RSS$ to have a proportion of difference
- $P$ number of predictors to explain $Y$
- $-1$ is the intercept $\beta_0$
- $TSS - RSS$ is the explained variance by the regression

If the linear model assumptions are correct:

$$E\left\{\frac{RSS}{(n-p-1)}\right\} = \sigma^2$$

- That the Expected value of the **unexplained variance** is due to irreducible error $\varepsilon$

if $H_0$ is true

$$E\left\{\frac{TSS-RSS}{p}\right\} = \sigma^2$$

- Which means that the predictors $X$ didn't effect the outcome response and have no relationship between each other
- Cause if there was a relationship between the predictors and response its gonna be :

$$E\left\{\frac{TSS-RSS}{p}\right\} > \sigma^2$$

The $F - test$ give us evidence to either reject or accept the **null hypothesis**, How big the $F - statistic$ should be to reject $H_0$

- Depends on the $n$ and $p$
- if $n$ number of Observation is large little larger than $1$ is enough
- if $n$ number is small $\rightarrow$ need larger $\text{F-Statistic}$

Sometimes we want to test a particular **subset** of predictors coefficients are zero

$$H_0 : \beta_{p-q+1} = \beta_{p-q+2} = \ldots \beta_p = 0$$

$$F = \frac{(RSS_0 - RSS)}{q} \times \frac{n-p-1}{RSS}$$

- $RSS_0 \rightarrow$ Residual sum of squares for the new model that only conclude $q$ coefficients we want to test
**WHY F-TEST**:
- when number of the variables is large $p = 100$ in the $H_0$ there is a $5\%$ chance of p-value being below $0.05$ by chance
- That's why individual t-test each predictor $X$ can lead to wrong assumptions
- F-Test avoids that by deciding and adjusting for the number of predictors $\frac{1}{p}$
- Nothing I learned till now will help if number of variables $p > n$

## Deciding On Important Variables:

Most of the time the Response is only associated with a subset of the predictors $X$, It would be better if we knew these predictors $X$ and fit them in a single model this can be done using **Variable Selection** which i will study later.

Here are some classical approaches:

- if $p$ is small we can text all four models and select the best
    - Model with no variables
    - Model containing only one predictor $X_1$

  - Model containing only the second predictor $X_2$
  - Model containing both of the predictors $X_1 X_2$
- For large numbers of $p$
  - **Forward Selection** : Its a greedy approach starting with a **Null model** (No variables), Adding variables with the lowest Residual Sum of Squares util we hit a threshold
    - **Disadvantages**:
      - May miss important predictors that are only significant when combined with others
  - **Backward Selection** : Start with all the variables and removing the ones with the highest $p - value$(Least statically Significant to the response $Y$), Stopping until all variables $p - value$ are below a certain value
    - **Disadvantage**
      - Expensive to compute if starting with all variables
  - **Mixed Selection** : Starting with no variables, Adding with **Forward Selection** and deleting with **Backward Selection** till we reach the desired outcome, Its still prone to Overfitting

# Assessing The Accuracy Of The Model

Assessing the accuracy of how well our model fits the given data using :

1. RSE
2. $R^2$
3. Confidence Interval for the mean Response
4. Prediction Interval
5. F-test

## RSE and $R^2$

The two most common ways to fit a model:

- **Residual Standard Error** $\rightarrow$ Its the sample standard deviation $S^2$, an estimate for the population standard deviation $\sigma^2$
  - The average amount the response $Y$ will deviate from the true regression hyperplane
  - Its measures how well the regression hyperplane fits the data

$$\text{RSE} = \sqrt{\frac{RSS}{n-p-1}}$$

  - $n - p - 1 \rightarrow p$ variables numbers, $1$ the intercept $\beta_0$ Degrees of Freedom
  - Adding any more predictors $X$ will in an increase in $\text{RSE}$ even if they have little association with the response, On the other hand it can result in a boost in the Response
  - Which can effect how well our multiple regression model fit
    **Interpretation of** $\text{RSE}$ :
  - Lower values indicate a tighter fit and less unexplained variability, which means our regression hyperplane explained a lot of the variance in the original data $\text{TSS}$
  - Higher values indicate a poor fit of the model
- $R^2 \rightarrow$ Its same mathematical concept in Simple Linear Regression
- $$R^2 = 1 - \frac{RSS}{TSS}$$
- It doesn't matter the amount of the predictors $X$
- With only risk of overfitting if the amount of variables $p$ is high

# Confidence Interval For The mean Response $Y_0$

Constructing a Confidence Interval for the expected value of $Y$ without taking into account:

- The irreducible error $\varepsilon$
- The Spread of the actual future outcome

Derived in Confidence And Prediction Intervals Derivations :

$$\hat{Y}_0 \pm t_{\frac{\alpha}{2},n-p} \hat{\text{SE}(\hat{Y}_0)}$$

- $Y_0 \rightarrow$ Predicted mean at $x_0$
- $t_{\frac{\alpha}{2},n-p} \rightarrow$ T$-value$
- $\hat{\text{SE}(\hat{Y}_0)} = S\sqrt{X_0^T (X^T X)^{-1} X_0} \rightarrow$ Standard Error
- This is the range where we believe the true mean response $Y_0$ falls in

The Confidence Interval is Narrower than prediction interval cause it doesn't take into account the irreducible error $\varepsilon$

## Prediction Interval For a New Response $Y_0$

Its the range where we expect an actual new future Response will fall into, given input $X_0$

- Here Consider the noise **irreducible error** $\varepsilon$

Derived in Confidence And Prediction Intervals Derivations

$$\hat{Y}_0 \pm t_{\frac{\alpha}{2},n-p}\hat{\text{SE}}(Y_0 - \hat{Y}_0)$$

- $\hat{\text{SE}}(Y_0 - \hat{Y}_0) = S\sqrt{1 + X_0^T(X^TX)^{-1}X_0}$
- We use the t-test cause of The sample standard deviation which is the estimate for the $\sigma^2$ the standard deviation of the population

## Confidence Interval vs Prediction Interval

| Key Points | Confidence Interval | Prediction Interval |
|---|---|---|
| **includes irreducible Error $\varepsilon$?** | No | Yes |
| **Target** | Mean response at a given Predictor value $x_0$ | Actual outcome $Y_0$ for a new observation m $X_0$ |
| **Tells you** | Where the mean is likely to fall | Where a single new data point is likely to fall |
| **Interested in** | The mean | The individual outcome value |
| **Used for** | Estimating trends, Model uncertainty | Capturing total uncertainty (model+ noise) |