Some Important Terms

Population: Group of subjects we want to draw conclusions
Sample: Subset of the entire population which we work with
Parameter: Is a number describing a whole population; it is generally unknown unless it is a **census** study
Statistics: A number describing the sample
Dependent Variable: Variables that measure the outcome of interest
Independent Variable: Variables whose relationship to the response is being studied
Univariate Data Set: Observations on a single variable.
Bivariate Data Set: Observations on two variables.
Multivariate Data Set: Observations on more than one variable.
Pearson's Correlation Coef: Numerical measure representing the strength and direction of the linear relationship. Correlation values between **-0.5** and **0.5** are considered to be weak. Correlation value between **-0.5** -- **-0.8** and **0.5 – 0.8** are considered to be moderate and values closer to **1/-1** is considered to be strong.
Correlation does not Imply Causation: Correlation does not imply causation and there's no way to determine or prove causation from a correlational study. Unless we make a census study, we can't confirm this. There can be a confounding variable in the study which also affects the final state of the population.

Simple Linear Regression

We have only a single predictor variable X, p=1.

$$Y = f(x) + \epsilon \; or \; Y = \beta_0 + \beta_1 X + \epsilon$$

B_0 is the intercept and B_1 is the slope. F(x) is the lin.func. eps. Res.er

Least Square Regressions Line

$$\hat{Y} = \widehat{B_0} + \widehat{B_1} X$$

$\widehat{B_0}$ is the mean response when the X is 0 -- it may or may not make sense to have X as 0. On average, we can expect $s.e.(\widehat{B_0})$ sample to sample variability in the intercept. When $\widehat{B_1}$ is 0.

$\widehat{B_1}$ is the change in the mean response produced by a unit change in X. On average, we can expect $s.e.(\widehat{B_1})$ sample to sample variability in the slope.

Goodness of Fit for the Fitted Model

Coeff. of Determination R^2: Proportion of variance in Y explained by X
Variability in Errors $\hat{\sigma}$: Average distance the observed values of Y around the regression line.
Res. Plot: Should have a random distribution without significant patterns.

Significance of Predictors

Set Hypothesis: $H_0: \beta_i = 0 \mid H_a: \beta_i \neq 0$
Level of Sign.: $\alpha = 0.05$
Calculate T-Stats: $t = \frac{\widehat{B_i} - B_{i,H_0}}{s.e(\widehat{B_i})}$, by default B_{i,H_0} is 0.
Find the P-Value and Conclusion: $P - value \leq \alpha \Leftrightarrow$ Reject H_0
If we reject H_0 → β_i is a significant predictor.
If we do not reject H_A → β_i is not a significant predictor.

LINE Assumptions for SLR

> L: There is a linear relation between X and Y.
Scatterplot Y vs. X & Scatterplot of $\hat{Y}$ vs. Y. → Show linear pattern
Residual plot → No clear pattern
Tukey's curve test →
Set Hypothesis: H_0: Linearity assumption holds.
H_a: Linearity assumption fails.
Level of Sign.: $\alpha = 0.05$
T-Statistic
Find the P-Value and Conclusion: $P - value \leq \alpha \Leftrightarrow$ Reject H_0
If we reject H_0 → Linearity assumption fails
If we do not reject H_A → Linearity assumptions hold

> I: Independence of Errors
Check index-plot of residuals → No clear cyclic pattern

> N: Normally Distributed Errors
Histogram of residuals → (normal distribution / bell curve)
QQ-Plot → (linear)

> E: Errors are Homoscedastic (they have a fixed variance)
Residual plot: No change in variability of errors
Breusch-Pagan Test:
Set Hypothesis: H_0: Constant variance. Homoscedastic
H_a: Non-Constant Variance. Heteroscedastic
Level of Sign.: $\alpha = 0.05$
Get T-Test
Find the P-Value and Conclusion: $P - value \leq \alpha \Leftrightarrow$ Reject H_0
If we reject H_0 → Heteroscedastic
If we do not reject H_0 → Homoscedastic

Transformations

When one or more of the standard regression assumptions fail, we can fix this by transforming the response/explanatory variables such that the assumptions will hold true after the transformation. If **linearity**, **normality** and **homoscedasticity** assumptions fail, we can use transformations.
Transformation of the predictor → *Can fix linearity*
Transformation of the response → *Can fix any of the 3 problems.*
Order of transformation is important as the order in which we deal with non-constant variance and non-linearity is important.
Transforming Y would influence 1, 2, and 3, while transforming X would only influence 1. Therefore, we should always deal with 2 and 3 before assessing the linearity.

Square-Root Transformation
Extent of Variability → Mild (Var(Y) increases slightly as Y increases)
Happens when Y counts are following a Poisson distribution.
If Y is not positive for all observations, use $\sqrt{Y + c}$

Logarithm
Extent of Variability → Medium (Var(Y) increases markedly as Y increases)
If range of Y is very broad
If Y is not positive use log(Y + c)

Reciprocal
Extent of Variability → Var(Y) drastically increases after some threshold.
If most values of Y are clustered and all are small numbers, whereas very few values are large.
When Y is waiting times for an event.

Transformation on the Predictor
We use the polynomial transformations when we need to apply a transformation to the predictor variable. This will give us a smoother model and a better fit overall.

Intercept when X is not 0
There can be cases in which X cannot be equal to 0. For example, Age, Height, Weight, etc. In this case, the Y-intercept would be 0 by default. To have a meaningful intercept, we can consider **centering** X, which will re-scale X to make another meaningful value (like mean) 0.

Unusual Observations

Generally, a small number of data points can have a large influence on a regression model, which can also violate the regression assumptions. There are three types of unusual observations:

Leverage Points
A data point with a high leverage can have a large influence on fitting the model.
hatvalues(fited_model) > 2 * mean(hatvalues(fitted_model))

Outliers
Outliers are points that do not fit the model well. Outliers may or may not have a large effect on the model. To identify outliers, we can check for observations with large residuals.
which(abs(rstandard(fitted_model))>2)

Influential Points
Influential points are values that have a large effect on the regression model. Observations in this category are also called high leverage, large residual points. To detect the influential points (points with high leverage and large residual), we use Cook's distance. A cook's distance for an observation is considered large when it is greater than **4/n** where **n** is the sample size, making that observation influential. This is only a heuristic to guess the observation and cannot be considered an exact rule.

Multiple Linear Regression (MLR)

When there is more than one predictor variable (p>1) in our model, we can classify it as a multiple linear regression model. Everything that is covered in SLR can be modified to MLR;

$$\boldsymbol{Y = f(X_1, X_2 \ldots X_p) + \epsilon}$$

where f() is a linear function and ϵ is the error term.

An MLR model can be written as follows;

$$Y = \beta_0 + \beta_1 X_1 + B_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

Similar to SLR, our goal is to find the least squares estimates of coefficients.

We can again use the coefficient of determination, variability in errors and residual plot ($\hat{Y}$ vs. Y) to measure the goodness of fit of our MLR model.

Similarly, we can again use the LINE assumptions but we reduce it to 6 plots (we do not have scatterplot of Y vs. X and Residuals vs. X).

In MLR, we also check for *multi-collinearity* to check if each predictor is independent of each other or not. At the same time, we check if they are non-random and not-fixed. And also, if they are measured without errors and no high-influence.

Detecting Multi-Collinearity

To check if the predictors are independent or not, we use,

> Scatter plot and Correlation plot matrices
Scatter plot matrix should not show a strong relationship.
Correlation plot should not have high values

> Variance Inflation Factor (ViF)
VIF value between 1 and 4 suggests that there is a *moderate correlation* but not enough.
VIF greater than or equal to 4 represents critical levels of multicollinearity. P-Values are questionable.

Comparing MLR Models

When a model is *nested* inside another model, this means that this model contains a subset of the predictors from a larger model. Comparing nested model, the coefficient of determination would not be a sufficient predictor.

The adjusted R^2 takes into account the number of degrees freedom. When we compare models with different number of predictors, we use adjusted R-Squared model.

Significance of Predictors

Set Hypothesis:

$\boldsymbol{H_0}: \boldsymbol{\beta_i = \beta_{i,H_0}} (\boldsymbol{for}$ **i=0, 1...,p**) There is no relationship between Y and the independent variable X_i (any predictor variable)

H_α: $\beta_i \neq \beta_{i,H_0}$ There is no relationship between Y and X_i.

Level of Sign.: $\boldsymbol{\alpha = 0.05}$

T-Statistic: $t = \frac{\widehat{B_i} - B_{i,H_0}}{s.e(\widehat{B_i})}$, by default B_{i,H_0} is 0.

Find the P-Value and Conclusion: $P - value \leq \alpha \Leftrightarrow$ Reject H_0

If we reject H_0 → The predictor is statistically significant.

If we do not reject H_0 → The predictor is not stats. significant.

Alternative Way to test Significance of Predictors

We can use the confidence intervals and check if it includes 0. If it includes 0, β_i=0 is a possibility which means that null hypothesis can also be true.

Overall Fit of Regression (F-Test)

Set Hypothesis

$\boldsymbol{H_0}: \boldsymbol{\beta_1 = 0, \beta_2 = 0, \dots, \beta_p = 0}$ None of the predictors can explain significant amount of variability in Y.

$H_\alpha: B_j \neq 0$ for at least one value of j = 1, 2, 3, … At least one predictor can explain significant amount of variability.

Level of Sign.: $\boldsymbol{\alpha = 0.05}$

T-Statistic: F = MSR/MSE ~ $F_{p,\ n-p-1}$

Find the P-Value and Conclusion: $P - value \leq \alpha \Leftrightarrow$ Reject H_0

If we reject H_0 → At least one of the predictors explain significant amount of variability.

If we do not reject H_0 → None of the predictors can explain significant amount of variability.

Likelihood Maximization

Mathematical method for evaluation how well a model fits the data it was generated from. There are three criteria AIC, AICc, BIC. We would try to select the model with the smallest criteria values to ensure that we have the best balance between the goodness of fit, data and complexity of the model.

$$AIC = ln\left(\frac{SSE_{(p+1)}}{n}\right) + \left(\frac{n + 2(p+1)}{n}\right)$$

$$AICc = ln\left(\frac{SSE_{(p+1)}}{n}\right) + \left(\frac{n + p + 1}{n - p - 1}\right)$$

$$BIC = ln\left(\frac{SSE_{(p+1)}}{n}\right) + \left(\frac{(p+1)ln(n)}{n}\right)$$

Where n is the sample size and (p+1) is the number of regression parameters.

When evaluating a model, we choose the lowest AIC, AICc, BIC and the highest R^2_{adj}.

Analysis of Variance (ANOVA)

Data = Model + Error

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \epsilon$$

Total variation in the response variable Y = Variation explained by regression model + Unexplained variation in errors.

Sum of Total Squares (SST)

$$[\sum_{i=1}^{n} (y_i - \bar{y})^2]$$

Sum of Regression Squares (SSR)

Partition of SST explained by the regression model.

$$[\sum_{i=1}^{n} (\hat{y_i} - \bar{y})^2]$$

Sum of Error Squares (SSE)

Part of SST that can not explained by the regression model.

$$[\sum_{i=1}^{n} (y_i - \hat{y_i})^2]$$

$$SST = SSR + SSE$$

Coefficient of Determination R^2

$$\boldsymbol{R^2 = \frac{SSR}{SST}}$$

Proportion of total variability in Y that is explained by the model.

Adjusted R-Squared ($R^2_{adjusted}$)

$$\boldsymbol{R^2_{adjusted} = 1 - \frac{(1 - R^2)(N-1)}{N - p - 1}}$$

Where R^2 is the sample R-square, p is the total number of predictors, and N is the sample size.

Degrees of Freedom (DF)

Degrees of Freedom Total = DfT = n-1

Degrees of Freedom due to regression = DfR = p

Degrees of Freedom due to Error = DfE = n - (p+1)

Mean Squares

Mean Squares Regression = MSR = SSR/DFR

Mean Squares Error = MSE = SSE/DFE

Mean Squares Total = MST = SST/DfT

Source of variation	***Regression***	***Error***	***Total***
Degrees of Freedom (Df)	p	n-(p+1)	n-1
Sum of Squares (SS)	SSR	SSE	SSR+SSE
Mean Squares (MS)	MSR=SSR/p	MSE=SSE/(n-p-1)	MST=SST/(n-1)
F-Statistic	F=MSR/MSE		
P-Value	2P(F>|F|)		

General Linear F-Test

Step 1: Define a larger **full (unrestricted) model** (with more parameters). The model thought to be the most appropriate for the data.

Step 2: Define a smaller **reduced (restricted) model**. Model described null hypothesis.

Step 3: Use an ***F-Statistic*** to decide whether or not to reject the smaller reduced model in favor of the larger full model.

Null hypothesis always favors the reduced model, while the alternative always favors the full model.

$$\boldsymbol{F = \frac{\frac{(SSE_{Red} - SSE_{Full})}{dfR - dfF}}{\frac{SSE_{Full}}{dfF}}}$$

dfR → Degrees of freedom for the reduced model

dfF → Degrees of freedom for full model.

We generally reject the null hypothesis if F-Value is large or its associated P-Value is small.

Set Hypothesis:

$H_0: B_n = 0$ → Favor reduced model

$H_\alpha: \beta_n \neq 0$ → Favor full model

Level of Sign.: $\boldsymbol{\alpha = 0.05}$

T-Statistic

Find the P-Value and Conclusion: $P - value \leq \alpha \Leftrightarrow$ Reject H_0

If we reject H_0 → Reject RM in favor of FM

If we do not reject H_0 → The predictor is not statistically significant. Do not reject the reduced model.

Three Types of F-Tests

When the research question asks if a specific predictor is significantly linearly related to the response.

When the research question asks if at least one predictor is useful in predicting the response.

When the research question asks about comparing two models, it can be used to test if one of the new predictors is statistically significant considering the other predictors, or it can be used to compare two models.

Categorical Predictor Variables

Categorical variables are variables that are used to classify observations. R identifies categorical variables as factor with levels.

Dummy Variable

We can incorporate categorical variables in our regression model by using dummy variables. The standard approach to adding categorical variables is to use a binary approach of 0/1. We check for the different dummy values in order to interpret the regression equation for different categories.

Hypothesis Test for Dummy Variables

For this test, we would use the same approach that we use in the significance of predictors.

Interactions

We use an interaction variable to avoid the "same slope" restriction. Similar to the dummy variables, we should be mindful of the category to have the different equations for different categories.