Assignment 4

1. **What is the difference between a univariate and a multivariate function?**

Ans:

## Univarate Analysis

Univariate analysis is the simplest form of data analysis where the data being analyzed contains only one variable. Since it's a single variable it doesn’t deal with causes or relationships.  The main purpose of univariate analysis is to describe the data and find patterns that exist within it.

You can think of the variable as a category that your data falls into. One example of a variable in univariate analysis might be "age". Another might be "height". Univariate analysis would not look at these two variables at the same time, nor would it look at the relationship between them.

Some ways you can describe patterns found in univariate data include looking at mean, mode, median, range, variance, maximum, minimum, quartiles, and standard deviation. Additionally, some ways you may display univariate data include frequency distribution tables, bar charts, histograms, frequency polygons, and pie charts.

**Bivarate Analysis**

Bivariate analysis is used to find out if there is a relationship between two different variables. Something as simple as creating a scatterplot by plotting one variable against another on a Cartesian plane (think X and Y axis) can sometimes give you a picture of what the data is trying to tell you. If the data seems to fit a line or curve then there is a relationship or correlation between the two variables.  For example, one might choose to plot caloric intake versus weight.

**Multivariate Analysis**

Multivariate analysis is the analysis of three or more variables.  There are many ways to perform multivariate analysis depending on your goals.  Some of these methods include:

* Additive Tree
* Canonical Correlation Analysis
* Cluster Analysis
* Correspondence Analysis / Multiple Correspondence Analysis
* Factor Analysis
* Generalized Procrustean Analysis
* MANOVA
* Multidimensional Scaling
* Multiple Regression Analysis
* Partial Least Square Regression
* Principal Component Analysis / Regression / PARAFAC
* Redundancy Analysis.

1. **What are the various ways to check the applicability of a particular regression model on a dataset?**

Ans:

The best way to take a look at a regression data is by **plotting the predicted values against the real values in the holdout set**. In a perfect condition, we expect that the points lie on the 45 degrees line passing through the origin (y = x is the equation). The nearer the points to this line, the better the regression.

Methods to determine the validity of regression models include **comparison of model predictions and coefficients with theory, collection of new data to check model predictions**. comparison of results with theoretical model calculations, and data splitting or cross-validation in which a portion of the data is used to *estimate* the model coefficients, and the remainder of the data is used to measure the *prediction* accuracy of the model. An expository review of these methods is presented. It is concluded that data splitting is an effective method of model validation when it is not practical to collect new data to test the model. The DUPLEX algorithm, developed by R. W. Kennard, is recommended for dividing the data into the estimation set and prediction set when there is no obvious variable such as time to use as a basis to split the data. Several examples are included to illustrate the various methods of model validation.

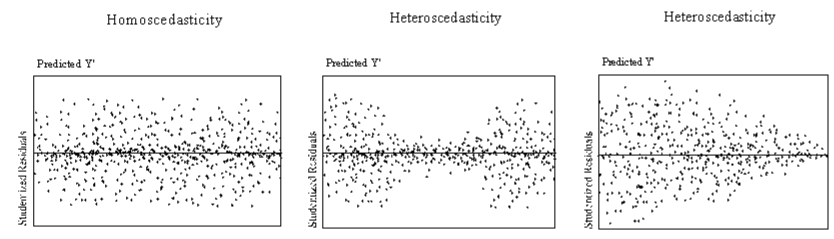
sssss

1. **What are the basic assumptions of the Linear Regression Algorithm?**

Ans:

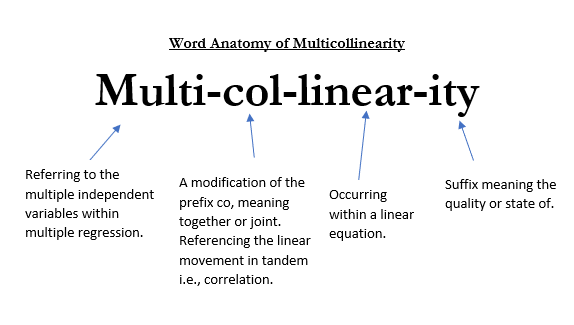
[**Linear regression**](https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/what-is-linear-regression/) is an analysis that assesses whether one or more predictor variables explain the dependent (criterion) variable.  The regression has five key assumptions:

1. **Linearity** (Linear features, relationship between input features and target value, It should be linear).
2. **Homoscedasticity**(Residual/error terms) (constant variance)



1. **Multicollinearity**: There should not be highly correlated features. \

**Detection**: VIF ,eigen value, correlation coefficient.



1. **Independence**: Features should be independent i.e. pairwise covariance is equal to zero.
2. The errors (residuals) should follow a normal/Gaussian Distribution. (µ=0,
3. Sum of residuals =0 (Your errors are normally distributed. )
4. **What are the different Evaluation metrics used in Linear Regression?**

Ans:

**Evaluation metrics for a linear regression model**

Evaluation metrics are a measure of how good a model performs and how well it approximates the relationship. Let us look at**MSE, MAE, R-squared, Adjusted R-squared, and RMSE.**

**Mean Squared Error (MSE)**

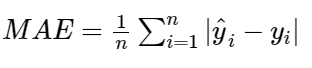
The most common metric for regression tasks is MSE. It has a convex shape. It is the average of the squared difference between the predicted and actual value. Since it is differentiable and has a convex shape, it is easier to optimize.

https://miro.medium.com/max/630/1*aLt5bWtuBr_V7TrBILe3qQ.png

MSE penalizes large errors.

**Mean Absolute Error (MAE)**

This is simply the average of the absolute difference between the target value and the value predicted by the model. Not preferred in cases where outliers are prominent.



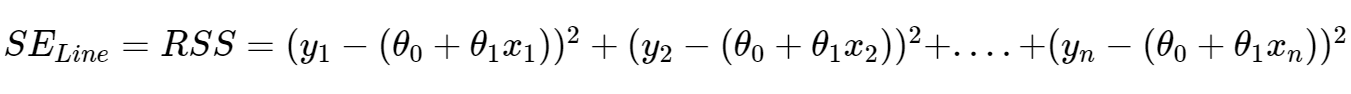
MAE does not penalize large errors.

**R-squared or Coefficient of Determination**

This metric represents the part of the variance of the dependent variable explained by the independent variables of the model. It measures the strength of the relationship between your model and the dependent variable.To understand what R-square really represents let us consider the following case where we measure the error of the model with and without the knowledge of the independent variables.

**Calculating regression error**When we know the values of the independent variables, we can calculate the regression error.

We know that residual is the difference between actual and predicted value. Thus, RSS (Residual sum of squares) can be calculated as follows.

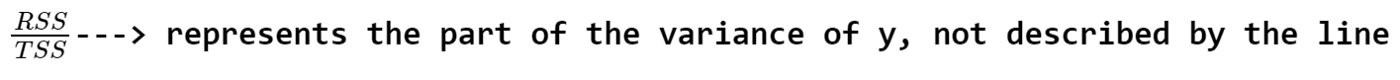


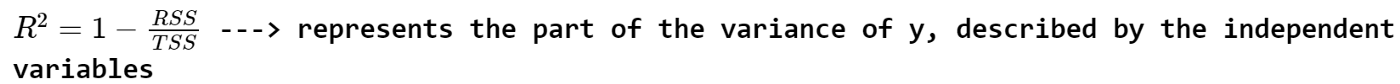
**Calculating squared residual error**Consider the case where we don't know the values of the independent variables. We only have the **y** values. With this, we calculate the **mean of the *y* values.** This point can be represented as a horizontal line. Now we calculate the sum of squared error between the **mean *y*value** and that of every other ***y*** value.The total variation in Y can be given as a sum of squared differences of the distance between every point and the arithmetic mean of Y values. This can be termed as **TSS**(Total sum of squares).

https://miro.medium.com/max/1268/1*nmhlSwtVwD3d244JcdBTvQ.png

Total variation in y or TSS

**Calculating the coefficient of determination with RSS & TSS**So we wanna find out the percentage of the total variation of Y, described by the independent variables X. If we know the percentage of the total variation of Y, that is **not**described by the regression line, we could just subtract the same from 1 to get the **coefficient of determination or R-squared.**





Coefficient of determination.

If the data points are very close to the regression line, then the model accounts for a good amount of variance, thus resulting in a high R² value.

*However do not let the R² value fool you. A good model can have low R² value and a biased model can have a high R² value as well. That is the reason you should make use of residual plots.*

To summarize, the ratio of the residual error (RSS) against the total error (TSS) tells you how much of the total error remains in your regression model. Subtracting that ratio from 1 gives how much error you removed using the regression analysis. That is the R-squared error.

***If R² is high (say 1), then the model represents the variance of the dependent variable.***

***If R² is very low, then the model does not represent the variance of the dependent variable and regression is no better than taking the mean value, i.e. you are not using any information from the other variables.***

***A Negative R² means you are doing worse than the mean value. It can have a negative value if the predictors do not explain the dependent variables at all such that RSS ~ TSS.***

Thus R² evaluates the scattered data points about the regression line.

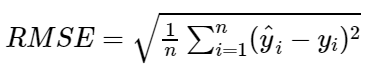
It is not possible to see a model with an R² of 1. In that case, all predicted values are the same as actual values and this essentially means that all values fall on the regression line.

**Root Mean Squared Error (RMSE)**

This is the square root of the average of the squared difference of the predicted and actual value.

R-squared error is better than RMSE. This is because R-squared is a relative measure while RMSE is an absolute measure of fit (highly dependent on the variables — not a normalized measure).

Basically, RMSE is just the root of the average of squared residuals. We know that residuals are a measure of how distant the points are from the regression line. Thus, RMSE measures the scatter of these residuals.



Root mean square error..

RMSE penalizes large errors.

**Model selection & Subset Regression**

Let me make it clear that, when you develop any model considering all of the predictors or regressor variables, it is termed as a **full model.**If you drop one or more regressor variables or predictors, then this model is a **subset model.**

***The general idea behind subset regression is to find which does better. The subset model or the full model.***

We select the subset of predictors that do the best of all the available candidate predictors, such that we have the largest *R²*value, largest adjusted R², or the smallest *MSE*.

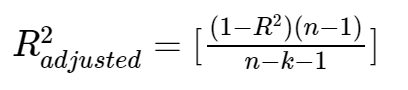
*However, R² is never used for comparing the models as the value of R² increases with the increase in the number of predictors (even if these predictors do not add any value to the model).*

***Reason for model selection****We set out to select the best subset of predictors that explain the data well.  
A simpler model that adequately explains the relationship is always a better option due to the reduced complexity. The addition of unnecessary regressor variables will add noise.*

**We will now look at the most common criteria and strategies for comparing and selecting the best models.**

**Adjusted R-squared — selection criterion**

The main difference between **adjusted R-squared**and R-square is that **R-squared** describes the amount of variance of the dependent variable represented by every single independent variable, while **adjusted R-squared** measures variation explained by only the independent variables that actually affect the dependent variable.



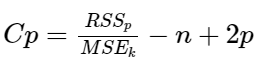
Adjusted R-squared.

*In the equation above, n is the number of data points while k is the number of variables in your model, excluding the constant.*

R² tends to increase with an increase in the number of independent variables. This could be misleading. Thus, the adjusted R-squared penalizes the model for adding furthermore independent variables (k in the equation) that do not fit the model.

**Mallow’s Cp — selection criterion**

Mallow’s Cp measures the usefulness of the model. It tries to compute the mean squared prediction error.



Mallow’s Cp statistic.

Here ***p***is the number of regressors, ***RSSₚ*** is the RSS of the model for the given ***p*** number of regressors, ***MSEₖ***is the total MSE for ***k***total number of predictors, and ***n*** is the sample size. ***This is useful when n>>k>p.***

Mallow’s Cp compares the full model with a subset model. If ***Cp*** is almost equal to ***p***(smaller the better), then the subset model is an appropriate choice.

One can plot***Cp vs p***for every subset model to find out the candidate model.

**Exhaustive and Best subset searching**

The **exhaustive search** looks at all the models. If there are ***k*** number of regressors, there ***2ᵏ*** possible models. This is a very slow process.

The **best subset strategy** simplifies the search by finding the model that minimizes RSS for every ***P-value***.

**Stepwise Regression**

Stepwise Regression is faster than Exhaustive and Best subset searching. It is an iterative procedure to choose the best model.  
*Stepwise regression is classified into backward and forward selection.*  
**Backward selection**starts with a full model, then step by step we reduce the regressor variables and find the model with the least RSS, largest R², or the least MSE. The variables to drop would be the ones with high p-values. It is however important to note that you cannot drop one of the levels of a categorical variable. Doing so would result in a biased model. You either drop all levels of the categorical variable or none.  
**Forward selection**starts with a null model, then step by step we increase the regressor variables until we can no longer improve the error performance of the model. We usually pick the model with the highest adjusted R²

1. Why do we square the residuals instead of the Modulus?

Ans:

By squaring the residual values, **we treat positive and negative discrepancies** in the same way. Why do we sum all the squared residuals? Because we cannot find a single straight line that minimizes all residuals simultaneously. Instead, we minimize the average (squared) residual value.

1. Which evaluation metric should you prefer if you have a lot of outliers present in it?

**Ans:**

**Mean Absolute Error(MAE)** is preferred when we have too many outliers present in the dataset because MAE is robust to outliers whereas MSE and RMSE are very susceptible to outliers and these start penalizing the outliers by squaring the error terms, commonly known as residuals.