Author: Sunil Yousef

Assignment Linear Regression Part2

Question-1:

List down at least three main assumptions of linear regression and explain them in your own words. To explain an assumption, take an example or a specific use case to show why the assumption makes sense.

Answer:

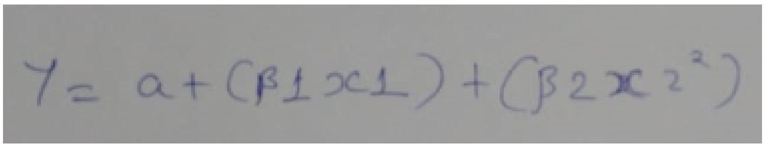
The 5 assumptions that are made for linear regression are Linearity, Outliers, Autocorrelation, Multicollinearity and Heteroskedasticity.

The 3 main assumptions are explained below.

1. Linearity

The assumption of linearity means we assume the association between variables and response variable are linear, that is the best fitting regression line is a straight line. A linear relationship means that a change in response Y due to one unit change in X(n) is constant, regardless of the value of X(n).

Example:

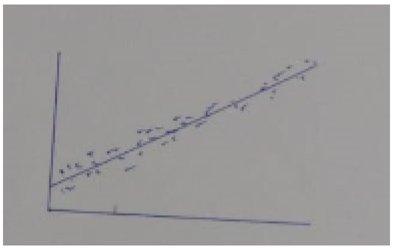


$$Y = a + (\beta 1 x1) + (\beta 2 x2^2)$$

Even though X2 is raised to power 2, the equation is still linear in beta parameters.

if a linear model is builded for a non linear dataset the model will not be able to capture the trend and will fail to analyse the unseen dataset. To check for linearity, a scatter plot of the data can be used and check whether the line fits the data points or not. If the variables are non linear then we need to identify new features that has linear behaviour.

***Plot showing linerarity*.**



***Plots showing no or little linearity*.**

Author: Sunil Yousef

2. Autocorrelation

Autocorrelation assumption says that, there should be no correlation between the residual (error) terms. Autocorrelation occurs when the residuals are not independent from each other, In other words value of y(x+1) is not independent from the value of y(x).

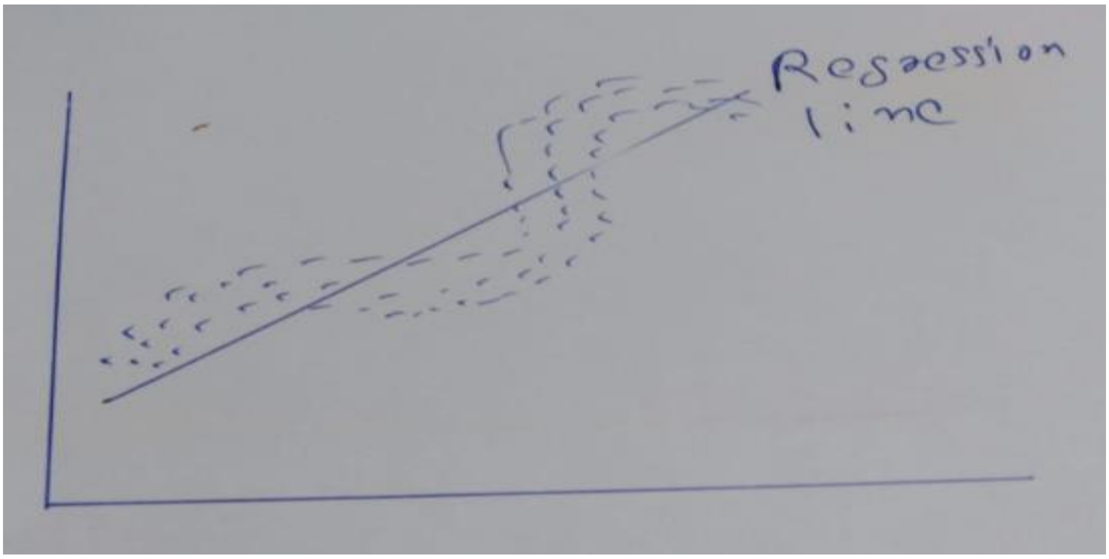
Autocorrelation mostly occurs in time series data or we can say autocorrelation applicable especially for time series data. If there is autocorrelation in data then one error will influence other errors and so on. Which means it can be the correlation of a time Series with lags of itself.

Linear regression analysis requires little or no autocorrelation in the data. If the error terms are correlated, the estimated standard errors tend to underestimate the true standard error.

It causes narrows confidence intervals and prediction intervals. Narrower confidence interval means that a 95% confidence interval would have lesser probability than 0.95 that it would contain the actual value of coefficients.

Examples of autocorrelation:

1. This typically occurs in stock prices, where the price is not independent from the previous price.
2. Consider the scenario where husband is the sole income owner of a house and both husband and wife is saving from the same income. in this case any change in income of husband will affect the savings of wife, then autocorrelation is present.



Methods to check for autocorrelation: The Durbin – Watson (DW) statistic, Plot a residual versus time plot and check for the correlated pattern in the residual values, Using acf plot, Using runs test.

Author: Sunil Yousef

3. Multicollinearity

Multicollinearity assumptions says the independent variables should not be moderately or highly correlated with each other, means there should not be any perfect linear relationship between explanatory variables.

The goal of regression analysis is to isolate the relationship between each independent variable and the dependent variable. If multi colinearity exists then the 1 unit of change in an independent variable will affect other corelated variables.

Problems with Multicollinearity

The stronger the correlation, the more difficult it is to change one variable without changing another. With presence of correlated predictors, the standard errors tend to increase. When predictors are correlated, the estimated regression coefficient of a correlated variable depends on which other predictors are available in the model.

Types of Multicollinearity

The two basic kinds of multicollinearities are Structural multicollinearity and Data multicollinearity.

1) Structural multicollinearity: Occurs when we create a model term using other terms. Example: if term X is squared to model curvature, clearly there is a correlation between X and X2.

2) Data multicollinearity: Multicollinearity present in the data itself. Example: Observational experiments are more likely to exhibit this.

Checking multicollinearity

1) Scatter plot - Use a scatter plot between the independent variables.

2) Variance Inflation Factor (VIF) – The simplest and preferred way is to check VIF values.

* VIF is a metric computed for every X variable that goes into a linear model.

* VIFs between 1 and 4 suggest that there is a moderate correlation, but not severe enough and can be left as it is.

* VIFs greater than 4 represent critical levels of multicollinearity where the coefficients are poorly estimated, and the p-values are questionable.

If the VIF of a variable is high, it means the information in that variable is already explained by other X variables present in the given model, which means, more redundant is that variable. So, lower the VIF (<2) the better.

VIF for a X var is calculated as: VIF = 1/(1-Rsquared)

Where, Rsq is the Rsq term for the model with given X as response against all other Xs that went into the model as predictors. If two of the X’s have high correlation, they will likely have high VIFs.

3) Correlation matrix - the correlation coefficients need to be smaller than 1.

Fixing Multicollinearity issues: The two ways to fix Multicollinearity issues are,

1) Iteratively remove the X var with the highest VIF.

2) See correlation between all variables and keep only one of all highly correlated pairs.

Question-2:

By now you have seen multiple **model evaluation metrics** used for regression models, such as r-squared, adjusted r-squared, RMSE, the residual plot etc.

In this question, you are required to **explain at least three regression model evaluation metrics** in your own words.

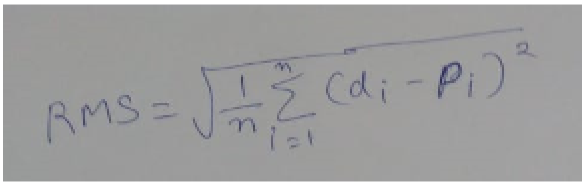
1. For the final model that you have built, explain each evaluation metric with its intuition (i.e. what and how it measures) and relate the intuition to its mathematical formula. You may use figures or examples to explain if needed. Limit your answer to 1000 words for this part.
2. Compare the advantages and disadvantages of any three evaluation metrics. If you do not think there's any advantage or disadvantage of a certain metric, mention that. Limit your answer to 1000 words for this part.

Answer:

1. RMSE:

    - The RMSE which is the square root of variance of the residuals indicates how close the observed data points are to the model’s predicted values, is a good measure of how accurately the model predicts the response.
    - RMSE has the same units as the response variable.
    - Lower values of RMSE indicates a better fit.
    - it is the most important criterion for prediction.

Equation for RMSE



$$RMS = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(d_i - p_i)^2}$$

For the final model builted as part of this assignment, RMSE is 0.05952925741364948, This indicates that the model 15 is a better fit as Lower values of RMSE(0.059) indicates a better fit and the Prediciton accuracy of model is good, as RMSE value is much lower

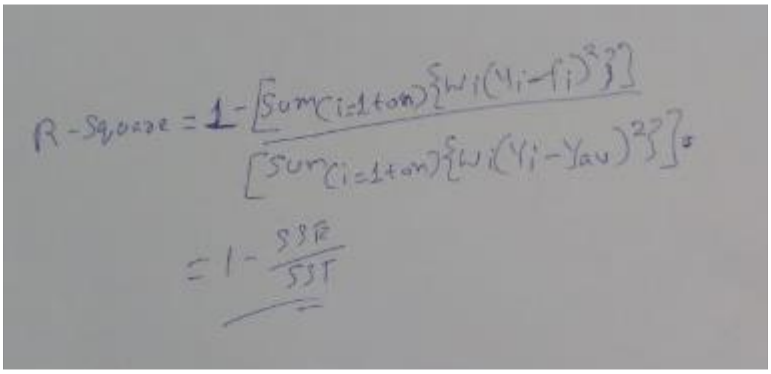
Disadvantages of RMSE:

- Root mean squared error measures the vertical distance between the point and the line, so if data is flat near the bottom and steep near the top. Then the RMSE will report greater distances to points high, but short distances to points low even when the distances are equivalent.

2. R-squared:

    - R-squared measures how close the data are to the fitted regression line and also known as the coefficient of determination.
    - R-squared is the percentage of the response variable variation that is explained by a linear model.
        - R-squared = Explained variation / Total variation
        - R-squared is always between 0 and 100%:
    - 0% indicates that the model explains none of the variability of the response data around its mean.
    - 100% indicates that the model explains all the variability of the response data around its mean.
    - This implies, the higher the R-squared, the better the model fits your data.

Equation of R-squared



$$R\text{-Square} = 1 - \frac{\left[\mathrm{sum}(i=1 \text{ to } n)\{w_i(y_i - f_i)^2\}\right]}{\left[\mathrm{sum}(i=1 \text{ to } n)\{w_i(y_i - y_{av})^2\}\right]} = 1 - \frac{SSE}{SST}$$

Where f_i is the predicted value from the fit, y_{av} is the mean of the observed data y_i is the observed data value. w_i is the weighting applied to each data point, usually w_i=1. SSE is the sum of squares due to error and SST is the total sum of squares.

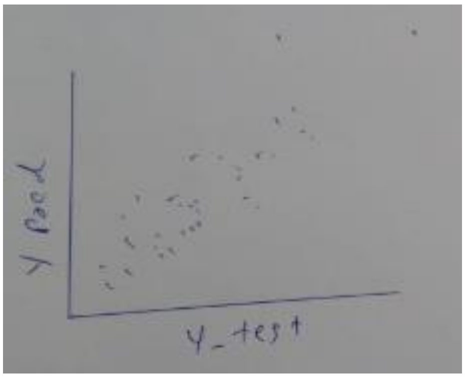
For the final model builted as part of this assignment, R-squared value is 0.8618851633496969, The 86% indicates that the model is fitting data much better.

3. Residual plot:

    - All errors are white error as they are randomly distributed
    - The errors is calculated by the differences between the actual and predicted values.
    - It must be randomly distributed and should not follow a pattern

- it should follow gausian distribution as per The Central Limit Theorem.
    - It states that the distribution of the sum of a large number of random variables will tend towards a normal distribution.
- Another assumption made is that each data point has errors that are independent from one another.
- Which helps us assume they occur randomly and not autocorrelated.
- Since the errors occur randomly, it is expected each data point has equal probability of appearing above or bellow the line of best fit.
- If errors followed a pattern then there is some variable that we might not have considered while building a model.
- Autocorrelation assumption says that, there should be no correlation between the residual (error) terms. if there is a correlation between error terms, then this phenomenon is known as Autocorrelation.
- Autocorrelation occurs when the residuals are not independent from each other, In other words when the value of y(x+1) is not independent from the value of y(x).

Example of good residual plot is given below



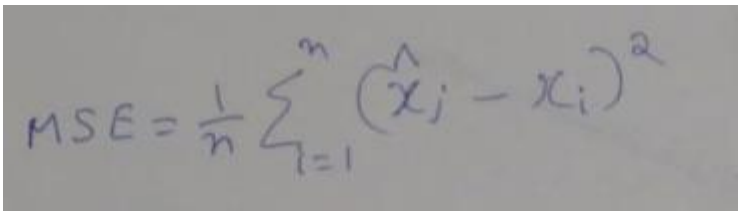
4. Mean squared error:

- The mean squared error (MSE) or mean squared deviation (MSD) is the average squared difference between the estimated values and what is estimated.
- MSE is a risk function, corresponding to the expected value of the squared error loss and is almost always strictly positive (and not zero).
- The MSE is a measure of the quality of an estimator, positive values closer to zero are better.

Equation of Mean squared error:

Let us suppose that **Xi bar** be the vector denoting values of n number of predictions. Also, Xi be a vector representing n number of true values.

Author: Sunil Yousef



$$MSE = \frac{1}{n} \sum_{i=1}^{n} (\hat{x}_j - x_i)^2$$

Comparison of methods

To test a model all the above mentioned four methods are really helpful,

- Comparison of R-squared and RMSE:
    - We know that R-squared is a relative measure of fit but RMSE is an absolute measure of fit, this is because RMSE is the square root of a variance.
- Advantages of Residual plot:
    - Residual plot helps us to quickly check the randomness of error and is an effective way to test model quickly.
- Comparison of MSE and R-squared and Residual plot:
    - The mean squared error is used to refer the unbiased estimate of error variance. This helps us to identify the risks involved in using the model.
    - After checking the residual plot it is a good idea to predict the risk of model using MSE.