---Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

A: We have season, weathersit, workingday, year, are few of the categorical variables and below are the observation for the plots I have drawn,

Bike Sharing count across different season : We can say that the bike's sharing is been approximately same during different seasons.

Bike sharing across different years: We can see that the number of bike shared in 2018 and 2019 are approximately same.

Bike sharing on working days vs holidays: we can clearly say that bike riding done on working day is drastically high when compared to weekend or holiday.

Bike sharing w.r.t to Weathersit below are the observations:[¶](http://localhost:8888/notebooks/ShreeKrishnaTechnologies/Linear%20Regression/Multiple%20Linear%20Regression/Poonam_Bike_Sharing_CaseStudy.ipynb#From-the-above-plot-on-Weathersit-below-are-the-observations:)

* When the weather is Clear, bike renting is done more than 450.
* When the weather is Mist+Cloudy, the bike rentin has reduced and it's falling beween 200 to 300.
* When the weather is with Light Snow, the bike rides has dropped less than 100.
* When there is a heavy rain, there is no bike ride at all.

Bike Sharing count across different months : we can say that bike renting is more than 50 irrespective of the month and from January

2. Why is it important to use drop\_first=True during dummy variable creation? (2 mark)

A: drop\_first=True used when we create the Dummy variables in Linear Regression model building. We use it when we have the categorical variables. This option help in removing the additional columns while dummy variable creation. If more dummy variables are created then collinearity may be too high. is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

A: The highest correlation is with temp and atemp variables.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

A: We assumed that the error terms are normally distributed. We have identified the error distribution by plotting the histogram w.r.t y-train data. We have got the decent bell curve upon plotting the histogram.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

A: The top 3 features temp, atemp and weathersit are the significantly contributing high towards the y-axis.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

A: In machine learning. Models are created to learn from the data and predict based on the variables provided.

Machine learning algorithms are classified as supervised learning and unsupervised learning.

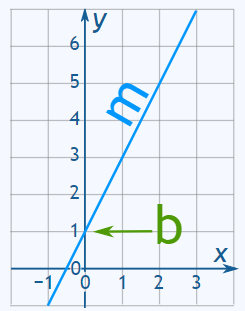
Again there is a classification in supervised learning and unsupervised learning. Under supervised learning,

1. Regression
2. Classification

Linear regression falls under regression algorithms. A linear regression model attempts to explain the relationship between a dependent and an independent variable using a straight line.The independent variable is also known as the predictor variable. And the dependent variables are also known as the output variables. In linear regression, we try to identify the best-fit line — the line which fits the given scatter-plot in the best way. The line is calculated using the straight line equation where m is the co-efficient of x and b is the constant and m is slope of the data.

**Equation of a Straight Line**

**Y=mx+b**



**Least Squares Regression Line**

The coefficients of the least squares regression line are determined by the Ordinary Least Squares method — which basically means minimising the sum of the squares of the: y-coordinates of actual data - y-coordinates of predicted data. The value of the correlation coefficient will always be: Between -1 and 1.

**Assumptions of Simple Linear Regression**

While building a linear model, you assume that the target variable and the input variables are linearly dependent. You are making inferences on the 'population' using a 'sample'. The assumption that variables are linearly dependent is not enough to generalise the results you obtain on a sample to the population, which is much larger in size than the sample. Thus, you need to have certain assumptions in place in order to make inferences.

**Hypothesis Testing in Linear Regression**

When you fit a straight line through the data, you'll obviously get the two parameters of the straight line, i.e. the intercept (b) and the slope (m). Now, while ‘b’ is not of much importance right now, but there are a few aspects surrounding ‘m’ which need to be checked and verified. We do this testing using Null Hypothesis testing.

Hence, every time you perform a linear regression, you need to test whether the fitted line is a significant one or not or to simply put it, you need to test whether ‘m’ is significant or not. And in comes the idea of Hypothesis Testing on ‘m’.

You start by saying that β1 is not significant, i.e. there is no relationship between X and y. So in order to perform the hypothesis test, we first propose the null hypothesis that m is 0. And the alternative hypothesis thus becomes m is not zero.

* Null Hypothesis (H0): m=0
* Alternate Hypothesis (HA): m≠0

Now, in order to perform the hypothesis test, you need to derive the p-value for the given beta.

calculate p-value anyway:

* Calculate the value of t-score for the mean point (in this case, zero, according to the Null hypothesis that we have stated) on the distribution
* Calculate the p-value from the cumulative probability for the given t-score using the t-table
* Make the decision on the basis of the p-value with respect to the given value of m(significance level)

 Now, if the p-value turns out to be less than 0.05, you can reject the null hypothesis and state that m is indeed significant.

Once we complete calculating the required fields, we next step is to build the model.

**Linear Regression Model Building**

The first important step before building a model is to perform the test-train split. To split the model, you use the train\_test\_split function.

**from sklearn.model\_selection import train\_test\_split**

After you import the statsmodel.api, you can create a simple linear regression model in just few steps.

**import statsmodels.api as sm**

**X\_train\_sm = sm.add\_constant(X\_train)**

**lr = sm.OLS(y\_train, X\_train\_sm)**

**lr\_model=lr.fit()**

By default, stats model does not involve the intercept i.e.,(y = mx+ c is y= mx) . 99% of the time we don’t ignore c unless we are sure that cis always) hence it is always recommended to add c while you use stats model for train test splitting.

X\_train\_sm = sm.add\_constant(X\_train)

Here, OLS stands for Ordinary Least Squares, which is the method that 'statsmodels' use to fit the line. You use the command 'add\_constant' so that statsmodels also fits an intercept. If you don't use this command, it will fit a line passing through the origin by default.

lr.summary() : This prints details of all the values like,

1. The coefficients and significance (p-values)
2. R-squared
3. F statistic and its significance

**Coefficients and p-values:**

The p-values of the coefficients tell you whether the coefficient is significant or not. If the coefficient of a particular variable came out to be 0.0545 with a standard error of about 0.002. Thus, you got a t-value of 24.722 which lead to a practically zero p-value. Hence, you can say that your coefficient is indeed significant.

**R-squared** value tells you exactly how much variance in the data has been explained by the model. If, the R-squared is about 0.816 which means that the model is able to explain 81.6% of the variance which is pretty good. Higher the R-Squared value, model is explaining well.

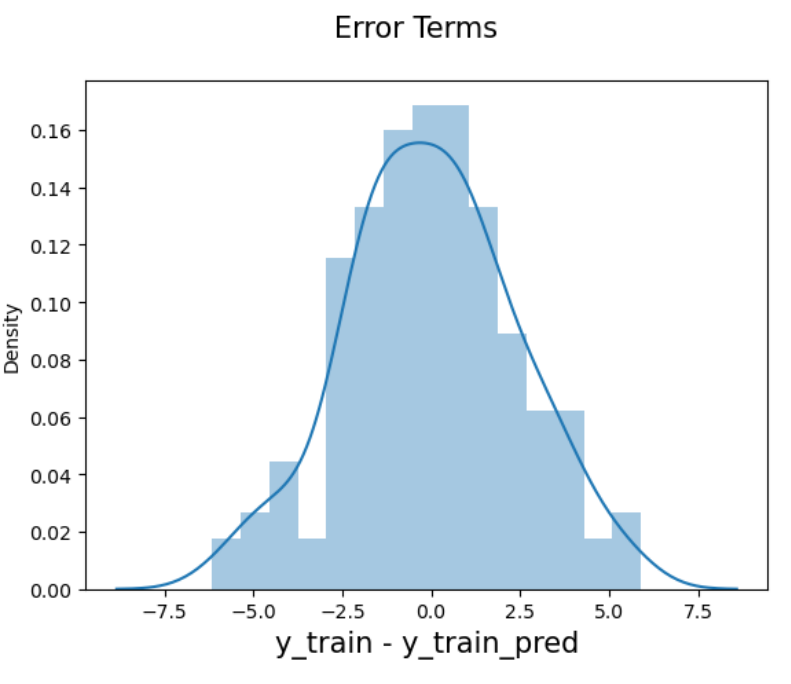
**F-statistic** is similar in the sense that now instead of testing the significance of each of the betas, it tells you whether the overall model fit is significant or not. This parameter is examined because many a time it happens that even though all of your betas are significant, but your overall model fit might happen just by chance. The heuristic is similar to what you learnt in the normal p-value calculation as well. If the 'Prob (F-statistic)' is less than 0.05, you can conclude that the overall model fit is significant. If it is greater than 0.05, you might need to review your model as the fit might be by chance, i.e. the line may have just luckily fit the data.

**Residual Analysis and Predictions**

Building the model on the train set has two parts: fitting a line and validating the assumptions of regression. Once linear regression model building is done, we need to perform residual analysis. You'd need to verify if your model is not violating this assumption. And doing this is fairly simple: you just plot a histogram of the error terms to check whether they are normally distributed. And another assumption was that the error terms should be independent of each other. Again, for this, you need to plot the error terms, this time with either of X or y to check for any patterns

The residuals are normally distributed, and there are no visible patterns in the error terms (except for the fact that the variance seems to be increasing a little for the higher values) which means, the model fit looks good.

Histogram plot Example:



By default, statsmodels fits a line passing through the origin, i.e. it doesn't fit an intercept. Hence, you need to use the command 'add\_constant' so that it also fits an intercept.

2.Explain the Anscombe’s quartet in detail. (3 marks)

A: It is a type a graph which is used to represent dataset in 4 different quartets. It was designed by the Statistician Francis Anscombe as Anscombe's quartet in 1973. It helps in analysing the data. When we look at the dataset, it seems to be very identical but when we plot, then we will get to know the real pictorial representation of data. It helps in analysing the data before we start working on it. It is very important to perform view the graph using Anscombe’s quartet as it shows the visually.

3.What is Pearson’s R? (3 marks)

A: It is a method of calculating the correlation between x and y variable using the formula, although in python is calculated through libraries. Basically, it calculates the correlation between variables which range between 1 and -1. The higher the value, there is strong correlation, if it’s 0 then there is no correlation and -1 means it has correlation in negative direction. We see this in python using head map. To plot the heatmap for observing correlation between variable we need to have the values to be numerical. We cannot plot it for categorical data.

4.What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

A: The dataset cannot always have numeric values. It may be combination of categorical and numerical. And also, the range of data cannot always be same Hence it is important to scale the feature variables to fit the correct values. Sometimes values can be more than 1, and all the categorical columns may have small integer values. So it is extremely important to rescale the variables so that they have a comparable scale. If we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients. This might become very annoying at the time of model evaluation. So we will use standardization or normalization so that the units of the coefficients obtained are all on the same scale. As you know, there are two common ways of rescaling:

1. Min-Max scaling : It have data scale to range between 0 and 1. It is helpful when we see the abnormal distribution of data.

2. Standardisation : It has mean-0 and sigma-1.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

A: VIF helps in calculating the variance. Large value indicates the correlation is higher and VIF value leads to infinite.  Variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables. The VIF values are considered as F-Statistics and for linear regression this values must be less than 0.5 otherwise we need to drop the features from the training model and it will also have impact on the model. This helps in determining the overall model fit in the liner regression.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

A: It is a scatter plot used in liner regression model. Scotter plot is very useful when we want to see the data distribution of type numeric. As liner regression is used to predict the numeric values, we calculate y-test and y-pred values. To see the data distribution we plot the scatter plot and view the results. We can visualize how well the data is scatter though the line from that we can predict if the model is a perfect fit or not.