Multiple Linear Regression – Bike sharing assignment

By Saunak MALLIK, ML-C64 Upgrad/ IIITB

Contents

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)

From Uni/Bi-variate analysis, below conclusions on categorical variables –

1. More rides are observed in summer and fall seasons.

2. The number of rides has increased in 2019 as compared to 2018... looks like bike-sharing is becoming more popular year-on-year.

3. Bike sharing is higher in the months of June-Sep months reflecting high holiday season.

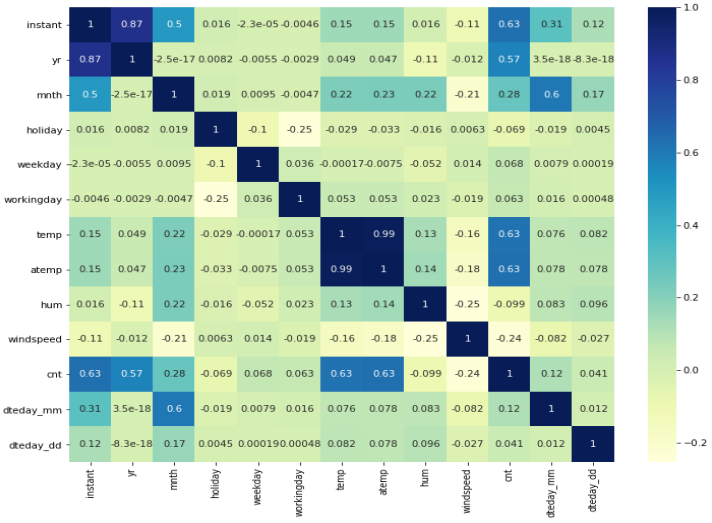
Additionally, later from the FINAL MODEL,

cnt = 0.2398 + 0.2631*yr - 0.1741*windspeed + 0.2926*spring + 0.1992*Mist + 0.2931*clear

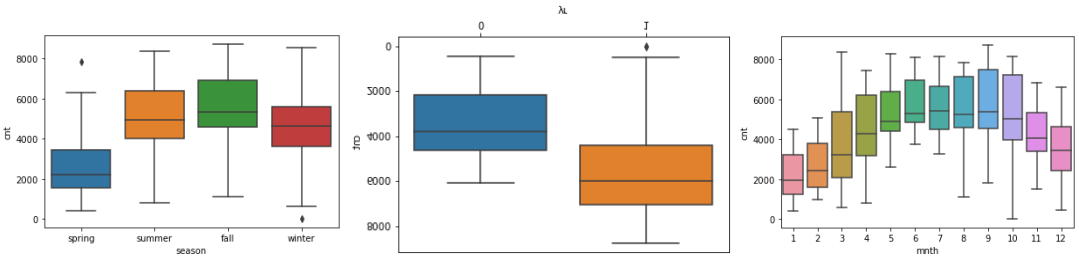
I could infer following about categorical variables –

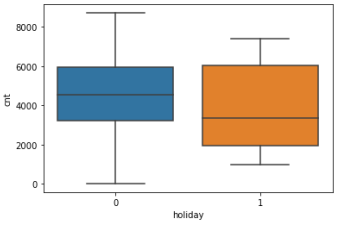
- mist weather conditions (Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist) has higher probability of hiring a bike
- clear weather conditions (Clear, Few clouds, Partly cloudy, Partly cloudy) has higher probability of hiring a bike
- In spring, there is a higher probability of hiring a bike
- The bike rental trend is increasing year-on-year. Hence, the popularity of the bike rentals seems on an increasing trend.

Multi-variate analysis (sns.heatmap(df.corr())), following about categorical variables –



Uni-variate analysis - sns.countplot() –





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

Answer: drop=True must be used during dummy variable creation while encoding categorical variables, to drop the reference group. This is to ensure that a categorical variable with n groups is encoded with (n-1) dummy variables. Each of the dummy coded variables uses one degree of freedom, so n groups have (n-1) degrees of freedom (~ analysis of variance).

One of the groups is encoded as 0, which is referred to as the "base group"/ "reference group". **This group is dropped with drop=True at the time of encoding**.

Why (n-1): That is to ensure one of the categorical variables serves as the base variable (all 0's).

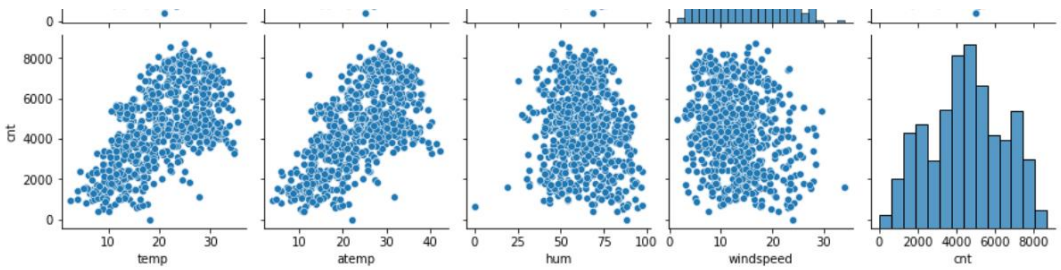
e.g., say a categorical variable has 4 categories – grp1, grp2, grp3 and grp4. We will need to create 3 (4-1) dummy coded variables.

base group = grp4

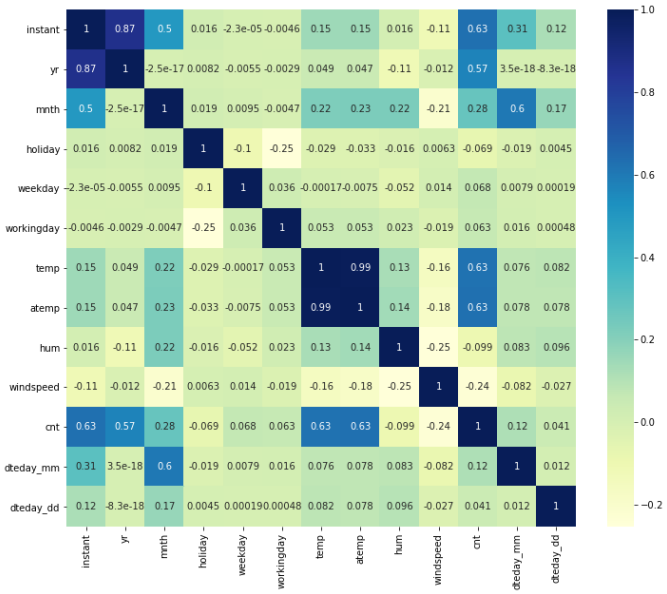
	Dummy_variable_1	Dummy_variable_2	Dummy_variable_3
grp1	1	0	0
grp2	0	1	0
grp3	0	0	1
grp4 = (base group)	0	0	0

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

Answer: Looking at the pairplot and Multivariate analysis (heatmap), it seems that temp and atemp variables exhibit highest the highest co-relation with the target variable, cnt –



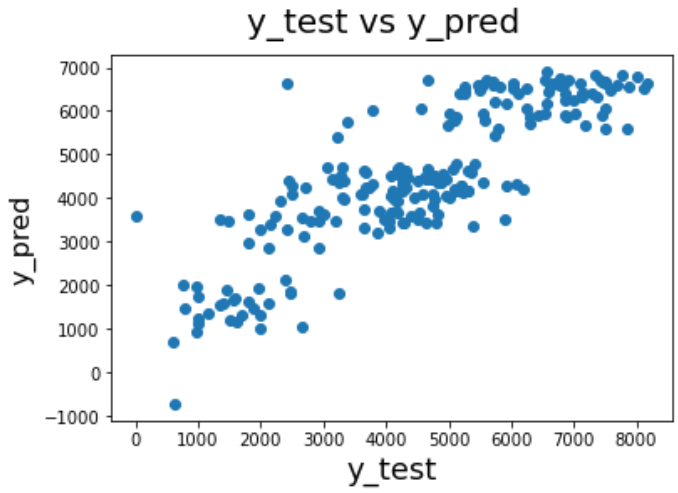
temp	0.63
atemp	0.63



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

Assumption 1: There is a *linear relationship* between X and Y

After building the model using the training dataset (df_train incl. X_train & y_train), I have tested the model on the testing_dataset (df_test incl. X_test & y_test). Below indicates a linear relationship.

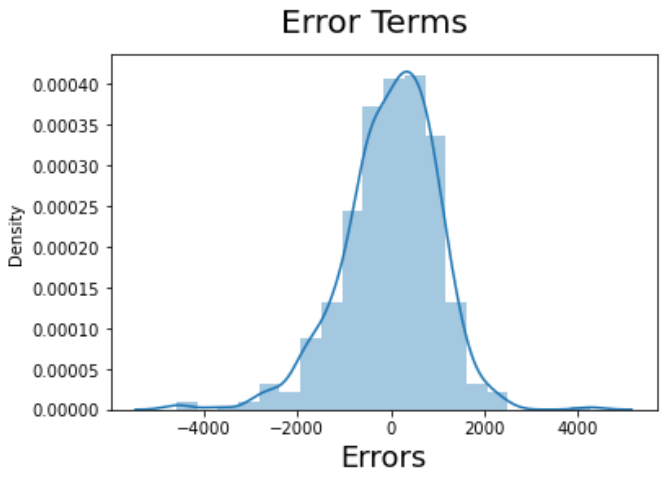


Assumption 2: Error terms are *normally distributed* with mean zero

After building the model using the training dataset (df_train incl. X_train & y_train), I have tested the model on the testing_dataset (df_test incl. X_test & y_test).

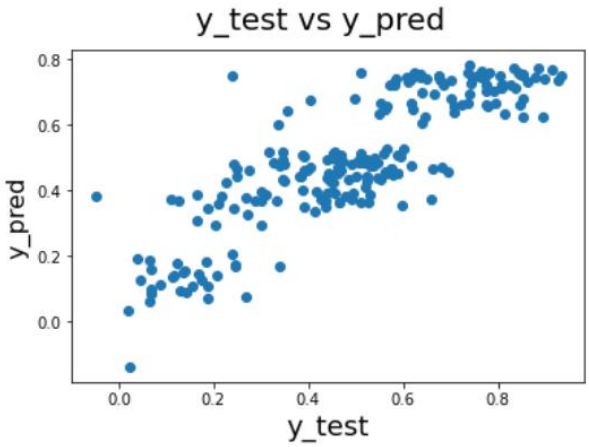
I have used the model to compute y_pred by using the training data, y_train.

And then computed and plotted error terms as: y_pred – y_train. I find that the error terms are Normally distributed with mean=0. (Below diagram)



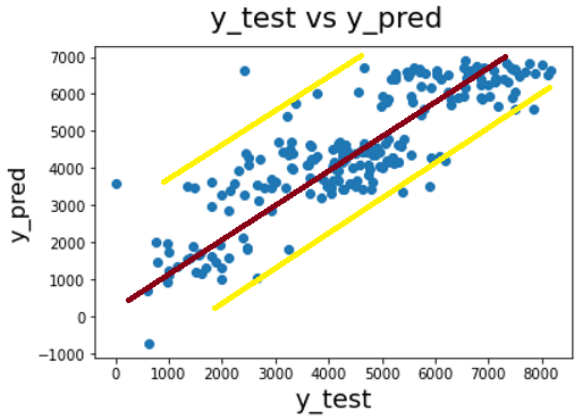
Assumption 3: Error terms are *independent* of each other.

There is no pattern exhibited by the Error terms. Hence, they're independent of each others. Additionally, this is proven from the Final model that where all features exhibits VIF<5.



Assumption 4: Error terms have *constant variance* (homoscedasticity).

The residuals/ errors does not increase or decrease as the error values changes (increases/ decreases).



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

Answer: Year, spring season and clear_sky are the most significantly contributing independent variables.

Final model:

cnt = 0.2398 + 0.2631*yr - 0.1741*windspeed + 0.2926*spring + 0.1992*Mist + 0.2931*clear

- p-values of all features < 0.05. Hence, they are all significant.
- VIF of all the features < 5. Hence, they do not exhibit significant relationship/ collinearity with other variables.

General Subjective Questions

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

- Linear Regression is a statistical method/ Machine Learning technique that allows us to summarize and study relationships between two continuous (quantitative) variables. Here, the labelled data is used to build the LR model and subsequently use the model to predict the future target variables.

Steps in building a Linear Regression model –

1. **Reading, Understanding and Visualising the dataset.** Perform Univariate, Bivariate and Multivariate analysis. Plot the numerical and categorical data.
2. **Data pre-processing/ Preparing the data for modelling (on-HOT encoding, train-test split, scaling etc.)**.
    - Data pre-processing step to prepare for building a linear model. This will include dealing with categorical variables – One HOT encoding, adding dummy variables etc.
    - Arrive at final dataset
    - Visualize the final dataset - Univariate, Bivariate and Multivariate analysis. Plot the numerical and categorical data.
    - **Split the dataset into training and test dataset**
    - Scaling the coefficients of the independent variables of the dataset
3. **Model building. Feature selection and Building/ Training the model recursively on the training dataset**
    1. Bottom-up approach OR
    2. Top-down approach OR
    3. Hybrid method (recommended)

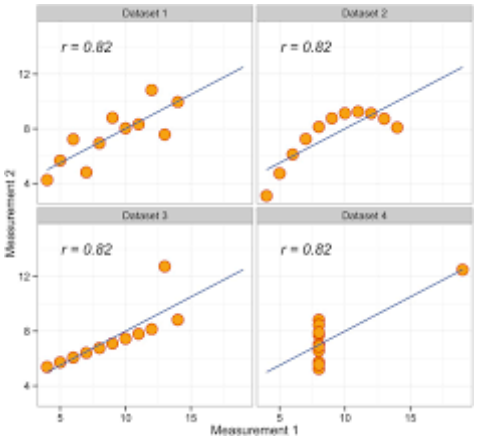
    It is good to use approach 3 (hybrid approach) whereby –

    - Use sklearn.RFE for Feature selection
    - Then build the model using OLS method from statsmodel.api. Check model
        - R^2

        - p-values of co-efficients
        - VIF of co-efficients
    - Recursively keep removing the dependent variables with p<0.05 & VIF<5.
    - Arrive at the final model
4. **Residual analysis**. Once all the variables have been added, you will perform a manual feature elimination and move on to the **residual analysis** and predictions, as usual.
5. **Predicting & evaluating the model on the Test Dataset**.
    - Ensure: The model R^2 on testing dataset **should not encounter** a very significant drop.
    - Else, this signifies over-fitting and we should go back and build the model over again.

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

- Anscombe's quartet is a group of qualitatively different datasets that have the –
    - same mean,
    - standard deviation,
    - and regression line,
- It is often used to illustrate the importance of looking at a set of data graphically and not **only** relying on basic statistical properties of the datasets.
- Anscombe's quartet emphasizes the importance of visualizing data through various graph types. From scatter plots to line graphs, we'll find how different datasets within the quartet can produce drastically different patterns, challenging our preconceived notions and highlighting the need for visual exploration.



3. What is Pearson's R? (3 marks)

- The Pearson correlation coefficient (r) measures the strength and direction of the relationship between two variables.
- It is the most common way of measuring a linear correlation.
- It is a number between –1 and 1.
    - r = 1 ➔ the two variables have a very high positive correlation i.e., When one variable changes, the other variable changes in the same direction.
    - r = 0 ➔ the two variables are not correlated.
    - r = -1 ➔ the two variables have a very high negative correlation i.e., When one variable changes, the other variable changes in the opposite direction.
- In EDA, Pearson correlation r is used to plot the heatmap as part of Multivariate analysis.

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

- Scaling is a technique performed to adjust the value of the coefficients of the numerical features in a Linear Regression model on to a similar scale, so as to avoid large variations of coefficients in the final model.

- Why is scaling performed: e.g., Scaling of variables is an important step because, there are some variables like 'windspeed, humidity, cnt which is on a different scale with respect to all other numerical variables, which take very small values. Also, the categorical variables that I have encoded earlier take either 0 or 1 as their values. Hence, it is important to have everything on the same scale for the model to be easily interpretable.

  Since I've performed scaling on my numerical features of my dataset, the co-efficients in my model is comparable (and in the same range (0-1):
  cnt = 0.2398 + 0.2631*yr - 0.1741*windspeed + 0.2926*spring + 0.1992*Mist + 0.2931*clear

- Scaling should be done after the test-train split.

- 2 scaling methods used in the industry –
  - 1. Min-max scaling: scaled values are in (0-1) range
  - 2. Standard scaling: No min-max range but it ensures mean = 0, std devn = 1

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

- Variance Inflation Factors (VIFs) is a measure of the correlation among independent variables in Ordinary least squares regression models.
- If there is a perfect correlation amongst the independent variables, then VIF = infinity i.e., some variables are able to create perfect multiple regressions on other variables.
  Here, $x_k = x_0 + x_1 + x_2 + \ldots\ldots\ldots\ldots x_{k-1}$
- Solution: remove such variables and build a new model.
- While building my model, had I not dropped the original categorical variables (e.g., Weathersit etc.) after doing a Dummy variables encoding, I'd have got VIF=infinite in my initial OLS model.

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

- QQ plot (or Quantile-Quantile) plot in statistics is a probability plot, which is a graphical method of comparing two probability distributions by plotting their quantiles against each other.
- QQ plot is a scatter plot by type.
- How to plot: The first quantile is that of the variable you are testing the hypothesis for and the second one is the actual distribution you are testing it against.
- To interpret a Q-Q plot, you need to look at the shape and pattern of the points. If the points lie on or close to a 45-degree line, it means that the data follow the reference distribution closely.