2022

Linear Regression Subjective Questions Solutions

SUBMITTED BY:
BALABHASKAR ASHOK KUMAR – MLC35

Assignment-based Subjective Questions

Q1) From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Soln.) Based on the Exploratory Data Analysis –

1. In the categorical variable of **year** – there is a clear increase in demand for bikes from 2018 to 2019. The average demand has increased from 2018 to 2019 (increased at least by 1.5 times). The increase in registered customer is more prominent than that of casual customers.
2. In the categorical variable months of **months** - The demand increases from start of the year and peaks at mid-year and starts dropping towards end of year (to be precise - during the period of May - October). Possible reasoning for this could include that weather tends to get worse during early year start and year end. It's typically Summer/fall (autumn) season during mid-year in US.
3. In the categorical variable months of **days of week** - We observe the pattern that most users tend to use the bikes less at start of the week (Mondays, Tuesdays ...) but the value increases over to weekend. However looking in depth tells us that Registered Users tend to decrease during Sundays whereas Casual users tend to use it more frequently during Sundays and Mondays. It decreases during the mid of week (Wednesdays, Thursdays ...).
4. In the categorical variable months of **seasons** - The highest demand is during the fall (autumn) season, followed by summer and winter and the least is during spring. The trend follows similarly for both casual and registered users.
5. In the categorical variable **workingday** - Majority of the demand is during the non-holiday day - this could possibly be that most of the user base use it could be working professional who use it for short commute

Q2) Why is it important to use drop_first=True during dummy variable creation?

Soln.) The importance of using drop_first = True or in general following the principle of having k-1 columns for k levels in categorical variable is to reduce **redundancy** and thereby creating a multi collinearity among the variables. You will be able to derive the kth case using only k-1 variables. If we do not drop a column we will be affected by this bottleneck which is sometimes referred to as **Dummy Variable Trap**

Eg: As we had seen in the home rental case study, a house could be furnished, unfurnished, semi-furnished. Creating dummy variables for this would be having any two columns. The dummy variable column would look like as below.

Furnished	Unfurnished
0	1
1	0
0	0

The row which has 0 in both furnished and unfurnished column simply indicates that it is a semi-furnished house.

Q3) Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Soln.) The highest correlation was observed for the variables temp and atemp (later while modelling we only retain temp variable) – assuming that we donot take into consideration casual and registered – as these are irrelevant to the analysis.

Q4) How did you validate the assumptions of Linear Regression after building the model on the training set?

Soln.) Assumption 1: Indications of Linearity between target variable and independent variable.

- This assumption was verified by plotting a pair plot. Looking at the target variable plots there was a clear linearity between the variable and some of the dependent variables such as temp.

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

- This assumption was verified by predicting the target variable of training set using the final linear model and plotting a distribution plot of the residual. The plot was a normal distribution curve at mean = 0.
- There was additionally no visible patterns on the plot of residuals against an index.

Assumption 3: Error terms have constant variance.

- To check whether the error terms are homoscedastic – the above residual plot versus indices were leveraged. The error terms were scattered around zero itself. However there was a slight larger variance towards the end of the dataset.

Q5) Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes

Soln.) Feature 1: temp (Positive Correlation)

- Temperature variable could be titled as the major driver for the “cnt” variable. For a unit increase in Temperature - the demand increases by 3339.8603 units
- Possible reasoning behind this could be as simple as that with drop in temperature, especially to very low condition people tend to stay indoors or prefer to use automobiles which has heating system.

Feature 2: yr (Positive Correlation)

- Although the data is only for two years – from EDA as well as regression model it is clear that the demand for the bikes has increased from 2018 to 2019. For a unit increase in the year the demand for the bikes increased by 2084 units.

Feature 3: windspeed (Negative Correlation)

- Wind speed had the largest negative correlation among the variables. For a unit increase in wind speed the demand drops by 1183 units.
- Possible reasoning behind this could be that the feeling temperature could be low when there is higher windspeed and it is practically difficult to ride bikes (in terms of resistance if you are moving against wind) when there are high wind speeds.

General Subjective Questions

Q1) Explain the linear regression algorithm in detail.

Soln.) Linear regression is a machine learning model that tries to fit a linear relationship between two variables.

$$y = \beta_0 + \sum \beta_i x_i$$

The overall goal is to minimize the error while predicting the y dataset. The most common algorithm for linear regression that we most commonly employ is the **method of Least Squares** – which calculates the value of coefficients by minimizing the sum of squares of residuals.

The residuals are actually the distance between the actual dependent variable and the predicted value of the dependent variable. The algorithm runs to minimize the sum of the square of this distance.

$$Minimize\ Residual\ Sum\ of\ Squares\ (RSS) = \sum (y_i - \left(\beta_0 + \sum \beta_i x_i\right))^2$$

A general metric use to explain this concept is the R^2 value which is given by

$$R^2 = 1 - \frac{RSS}{Total\ Sum\ of\ Squares\ (TSS)}$$

There are different types of Least Squares technique that we can employ, namely:

1. Ordinary Least Square (OLS) – used when the errors are homoscedastic
2. Weighted Least Square (WLS) – used when the errors are heteroscedastic
3. Generalized Least Square (GLS) – used when heteroscedasticity and correlations are present

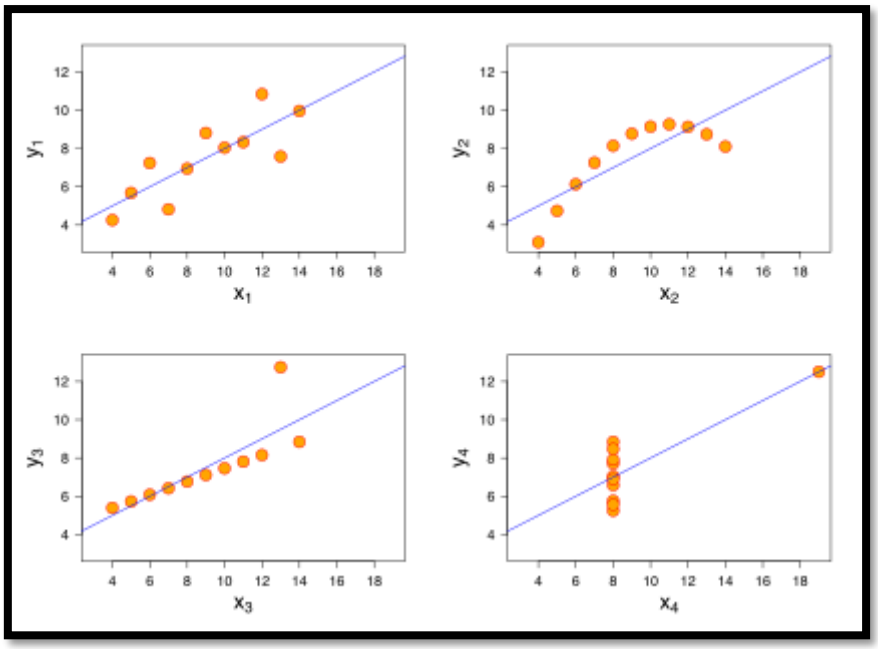
Assumptions for Ordinary Least Square Linear Regression Models:

1. There should be a linear relationship between X and y.
2. Residuals must be normally distributed.
3. Errors terms must be independent of each other and must have constant variance (homoscedasticity).

Q2) Explain the Anscombe's quartet in detail.

Soln.) Anscombe's quartet is four datasets that have the same statistical metrics but appear different when plotted on a graph. Below is the brief summary of the statistical metrics for all four dataset.

1. Mean of the x variables are : - 9
2. Mean of the y variables are : - 7.5
3. Variance of x : - 11
4. Variance of y : - 4.125
5. Correlation betwee n x and y : - 0.816
6. R2 Value : - 0.67



The purpose of using Anscombe's quartet is to

1. Explain that we cannot derive conclusion merely by looking into statistical metrics – there could be outliers that could be causing the effect. Hence graphing it is key to understanding the dataset.
2. We cannot solely rely on metrics to predict the linearity between two variables. As an example R2 value of 0.67 is often considered normal – however only the first graph in the quartet has actually a linear relationship.

Q3) What is Pearson's R?

Soln.) Pearson's R (Product Moment Correlation coefficient) can be defined as the measure of linear correlation between two variables. Mathematically it is defined as the ratio of product of covariance of the two variables to the product of their standard deviation. The value could range from -1 to +1 where +1 indicates a perfectly positive correlation, 0 indicates no correlation at all and -1 indicates a perfectly negative correlation.

$$r = \frac{N\sum xy - \sum x \sum y}{\sqrt{[N\sum x^2 - (\sum x)^2][N\sum y^2 - (\sum y)^2]}}$$

Where

N = Number of observations

Pearson's r also has assumes the same assumptions as that of linear regression, i.e. the residuals distributed normally, there is no variance in the error distribution and most importantly Y is linearly related to X.

Q4) What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Soln.) Scaling is the process of transforming the given dataset into a applicable range so that we can compare between variables and additionally it helps in faster computations.

Scaling is primarily done when you want to interpret/convey meanings through your coefficients of your model. If scaling is not done – there is a high chance that some of your variables might get ignored/ effect might not be translated prominently because of large magnitude of other variables. Scaling doesnot affect the accuracy/statistical metrics of the model.

Two of the most used scaling techniques are:

1. **Normalized Scaling (MinMax Scaling)** – This technique brings the dataset into the range of [0,1] by picking up the largest value in the dataset and smallest value in the dataset.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

2. **Standardized Scaling** -– This technique brings the dataset into a normal distribution format where the values are centred around with mean being zero and standard deviation being 1.

$$X' = \frac{X - \mu}{\sigma}$$

The preference of one over the other depends on multiple factors. If your distribution is Gaussian, you typically tend to stick with standardized scaling. When your model is distance based algorithm it would be good to follow normalized scaling. Standardized scaling does not affect outliers since they do not have a bounding range.

Q5) You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Soln.) $Variance\ Inflation\ Factor$, VIF $= \frac{1}{1-R_i^2}$

The Variance Inflation Factor will become infinite when the denominator turns to 0, i.e. When R2 value becomes 1 indicating that there is a perfect correlation between the ith variable (one that has turned infinite) and rest of the variables. This simply means that the ith variable can be explained completely by the rest of the variable.

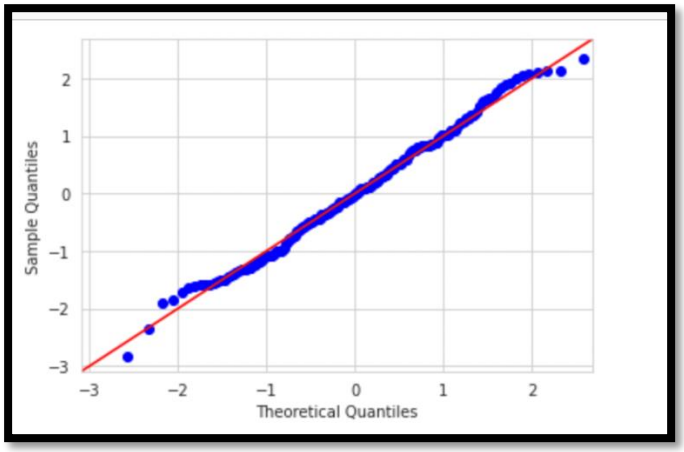
The solution of this problem is to drop one of the irrelevant variable that is causing this issue.

Q6) What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

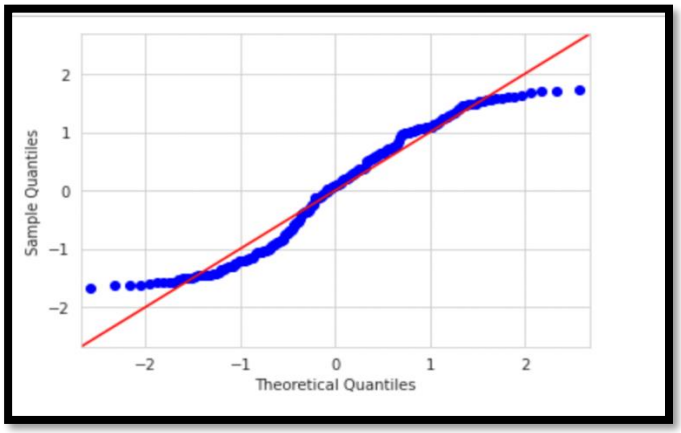
Soln.) Q-Q plot or Quantile – Quantile Plot is a graphical plot between theoretical sample and experimental sample and helps in determining whether both the dataset originate from population that have the same distribution, understand the skewness, tail behaviour etc. This is typically leveraged when your training dataset and test dataset are given separate.

Observables:

- A straight line of scatter plots where the points are really close to the error line makes it a normal distribution.



- An S Shape of the scatter plot should point out that the distribution is uniform



The importance of Q-Q plot in linear regression arises when you plot the residuals. You should be able to observe a straight lines for the residuals if it follows normal distribution. This can be achieved by the following code:

```
import statsmodels.api as sm
res = y_test – y_test_pred
sm.qqplot(res, fit = True,line = 45)
```