Linear Regression Subjective Questions

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?

I have done analysis on categorical and continuous variable using barplot and scatter plot respectively. Below are the mentioned points I can infer from the visualization:

1. High number of bike rents on fall and then summer season compare to other seasons
2. Year 2019 have higher number of bike rents than year 2018
3. Higher number of bike rents when on clear weather compare to other weather conditions
4. There are more bike rents on Saturdays, Thursdays, and Wednesdays compare to other weekdays.
5. More number of bike rents when humidity is high
6. More bike were rented on when the wind speed was slow
7. Most of the bookings has been done during the month of may, june, july, aug, sep and oct. Trend increased starting of the year till mid of the year and then it started decreasing as we approached the end of year.
8. More number of bike rents when it’s not a holiday

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

It is important to do drop_first=True as it helps in reducing the extra number of columns created during dummy variable creation. Thus it reduces the correlations created

among dummy variables.

The rule of thumb is to use always k-1 dummy variables out of k categories.

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

‘temp’ variable has the highest correlation with target variable

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

I done validation on the basis of these assumptions:

1. Multicollinearity check: There is should less or no significant multicollinearity among independent variables
2. Normalization of error/residuals terms: Error terms should be normally distributeda and mostly centered around 0
3. Linear relationship validation: There should be visible linearity among the variables
4. Independence of residuals: Residual should be dependent to each other
5. Homoscedasticity: There should be no visible pattern in residual values.

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

The top 3 features contributing significantly towards explaining the demand of the shared bikes were:

1. temp
2. year
3. Saturday

General Subjective Questions

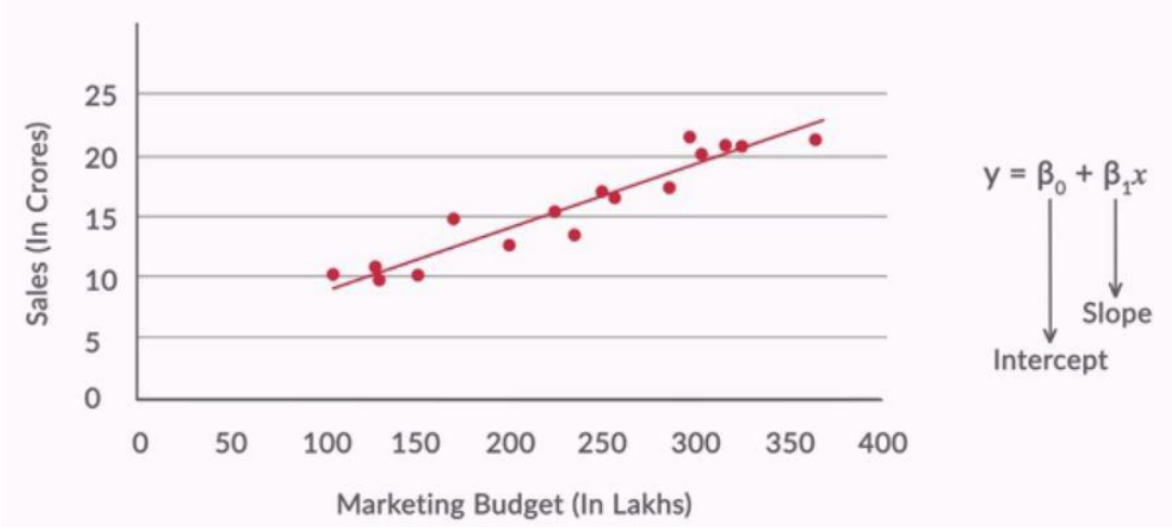
1. Explain the linear regression algorithm in detail.

The first and most elementanry type of model is regression model which describes the relationship between a dependent variable and set of independent variable using a staright line. The straight line is plotted on the scatter plot of these set of points.

The standard equation of the regression line is given by the following expression:

$$Y = mX + c$$

- Here Y is the dependent variable we are trying to predict
- X is the independent variable we are using to make predictions
- m is the slope of the regression line which represents the effect X has on Y
- c is a constant, known as the Y-intercept. If X=0, Y would be equal to c.



Now, there are can be Positive or Negative Linear Relationship

- Positive linear relationship: If both independent and dependent variable increases.
- Negative linear relationship: if independent increases and dependent variable decreases.

There are two types of Linear Regression:

1. Simple Linear Regression

a. In Simple Linear Regression we have only one independent and one target variable. The expression of this is:

$$Y = Beta0 + Beta1 * X$$

where **Beta0** denotes constant and **Beta1** denotes the slope of the regression line

2. Multiple Linear Regression:

a. In Simple Linear Regression we have only multiple independent and one target variable. The expression of this is:

$$Y = Beta0 + Beta1 * X1 + Beta2 * X2 + ... + Betap * Xp$$

where **Beta0** denotes constant and **Betap** denotes the slope of the regression line and p denotes the number of predictor of that variable.

The following are some assumptions about dataset that is made by Linear Regression model −

1. Multi-collinearity:

Linear regression model assumes that there is very little or no multi-collinearity in
the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.

2. Auto-correlation:

Another assumption Linear regression model assumes is that there is very little or
no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.

3. Relationship between variables:
Linear regression model assumes that the relationship between response and feature variables must be linear.

4. Normality of error terms: Error terms should be normally distributed

5. Homoscedasticity: There should be no visible pattern in residual values.

2. Explain the Anscombe’s quartet in detail.

Anscombe’s Quartet can be defined as a group of four data sets which are **nearly identical in simple descriptive statistics**, but there are some peculiarities in the dataset that **fools the regression model** if built. They have very different distributions and **appear differently** when plotted on scatter plots.

It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize COMPLETELY, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

Anscombe's Data										
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

Image from towardsdatascience.com

The summary statistics show that the means and the variances were identical for x and y
across the groups:

- Mean of x is 9 and mean of y is 7.50 for each dataset.
- Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset

- The correlation coefficient (how strong a relationship is between two variables) between
  x and y is 0.82 for each dataset

When we plot these four datasets on an x/y coordinate plane, we can observe that they show
the same regression lines as well but each dataset is telling a different story:

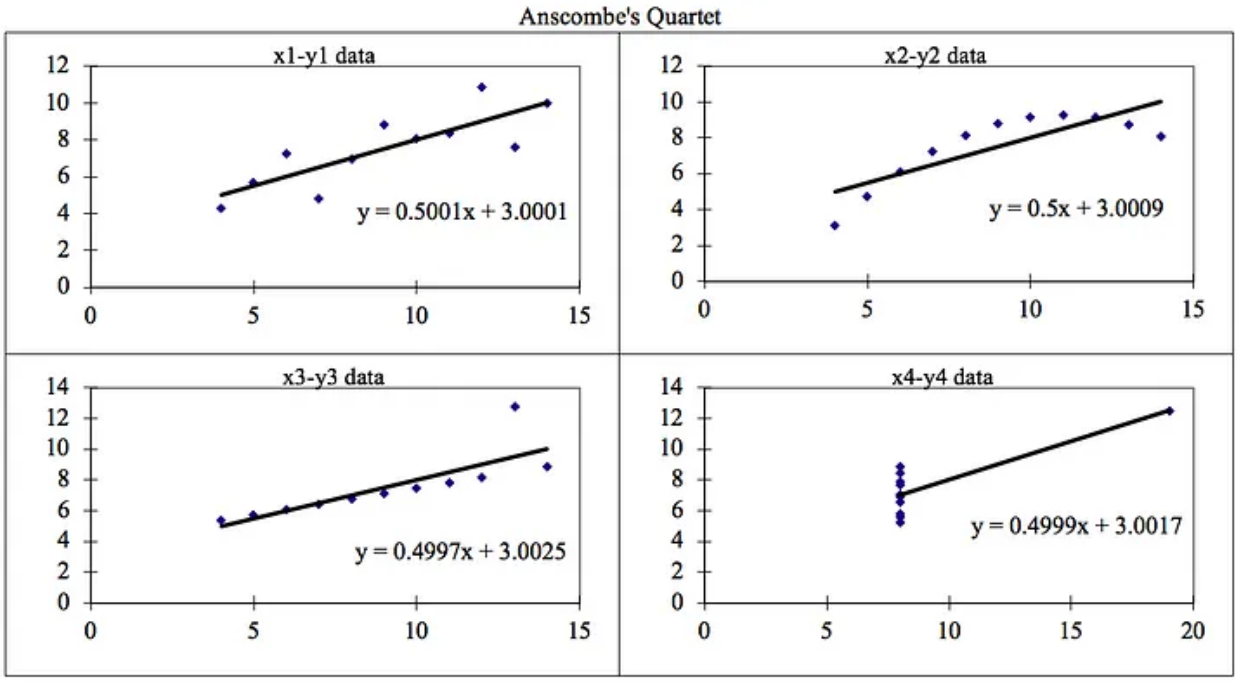


Image from towardsdatascience.com

The four datasets can be described as:

1. **Dataset 1:** this **fits** the linear regression model pretty well.
2. **Dataset 2:** this **could not fit** linear regression model on the data quite well as the data is non-linear.
3. **Dataset 3:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model
4. **Dataset 4:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model

3. What is Pearson's R?

The **Pearson correlation coefficient (*r*)** is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

Pearson correlation coefficient (*r*)	Correlation type	Interpretation	Example
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the **same direction**.	Baby length & weight: The longer the baby, the heavier their weight.
0	No correlation	There is **no relationship** between the variables.	Car price & width of windshield wipers: The price of a car is not related to the width of its windshield wipers.
Between0 and –1	Negative correlation	When one variable changes, the other variable changes in the **opposite direction**.	Elevation & air pressure: The higher the elevation, the lower the air pressure

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

It is one of the important aspect to consider the feature scaling. So what happens, when you have lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret.

So we need to scale features because of two reasons:

1. They can be then easy to interpret
2. Then there will faster convergence of gradient descent methods

There are two ways to scale the feature:

1. **Standardizing**: The variables are scaled in such a way that their mean is zero and standard deviation is one.
2. **MinMax Scaling:** The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

Normalized scaling	Standardized scaling
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
It is really affected by outliers.	It is much less affected by outliers.
Scikit-Learn provides transformer called MinMaxScaler for Normalization.	Scikit-Learn provides transformer called StandardScaler for standardization.

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

If there is absolute perfect correlation then VIF will be infinite. A large value of VIF indicates that there is correlatioin between the variables. If the VIF is >5, this means that the variance of the model coefficent is inflated by a factor which is worth examining and if the VIF is >10 then it is better to drop the variable because due to its presence there is high multicollinerity.

In the case where VIF is infinite then the perfect correlation between the independent variable will be visible and we get R-squared value as 1. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

The use of Q-Q plot can be described like this,

Suppose ,for example, if dataset 1, the age variable has 200 records and dataset 2, the age variable has 20 records, it is possible to compare the distributions of these datasets to see if they are indeed the same. This can be particularly helpful in machine learning, where we split data into train-validation-test to see if the distribution is indeed the same. It is also used in the post-deployment scenarios to identify covariate shift/dataset shift/concept shift visually.