

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)

Answer: The below are categorical variables present from dataset:

1. Season
Fall has the highest demand for rental bikes.
2. Year
The demand was increased from 2018 to 2019.
3. Month
Demand is continuously growing each month till June. September month has highest demand. After September, demand is decreasing.
4. Holiday
When there is holiday, demand has decreased.
5. Weekday
Weekday is not giving clear picture about demand.
6. Weather
The Clear weather has highest demand.
7. Working Day
The demand has increased on working day.

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

Answer: drop_first=True 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.

When you have a categorical variable with say 'n' levels, the idea of dummy variable creation is to build 'n-1' variables, indicating the levels. For a variable say, 'Relationship' with three levels namely, 'Single', 'In a relationship', and 'Married', you would create a dummy table like the following:

Relationship Status	Single	In a relationship	Married
Single	1	0	0
In a relationship	0	1	0
Married	0	0	1

But you can clearly see that there is no need of defining three different levels. If you drop a level, say 'Single', you would still be able to explain the three levels.

Let's drop the dummy variable 'Single' from the columns and see what the table looks like:

Relationship Status	In a relationship	Married
Single	0	0
In a relationship	1	0
Married	0	1

If both the dummy variables namely 'In a relationship' and 'Married' are equal to zero, that means that the person is single. If 'In a relationship' is one and 'Married' is zero, that means that the person is in a relationship and finally, if 'In a relationship' is zero and 'Married' is 1, that means that the person is married.

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

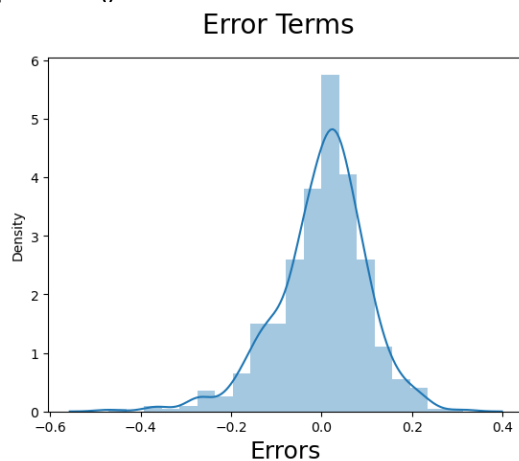
Answer: Temp and Atemp are having highest correlation with the target variable. Also Temp and Atemp are highly correlated to each other so that's the reason, I have dropped Atemp from model analysis and building.

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

Answer:

We checked if the error terms are also normally distributed which is in fact, one of the major assumptions of linear regression. We plotted histogram of the error terms and seen what it looks like.

```
y_train_pred = lm.predict(X_train_lm)
# Plot the histogram of the error terms
fig = plt.figure()
sns.distplot((y_train - y_train_pred), bins = 20)
fig.suptitle('Error Terms', fontsize = 20)      # Plot heading
plt.xlabel('Errors', fontsize = 18)            # X-label
plt.show()
```



As we can see that Errors are normally distributed here with mean 0. So, everything seems to be fine.

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:

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

1. Temperature
Demand of Shared bikes increases in pleasant temperature.
2. Holiday
The demand is very less in Holiday and hence, company should focus on working day like gives an offers or discounts on working day rather than holidays.
3. Season/Month
Company should focus on expanding business during Winter and Fall and Company should focus on expanding business during September.

General Subjective Questions

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

Answer: Linear regression may be defined as the statistical model that analyzes the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

Mathematically the relationship can be represented with the help of following equation –

$$Y = mX + b$$

Here, Y is the dependent variable we are trying to predict

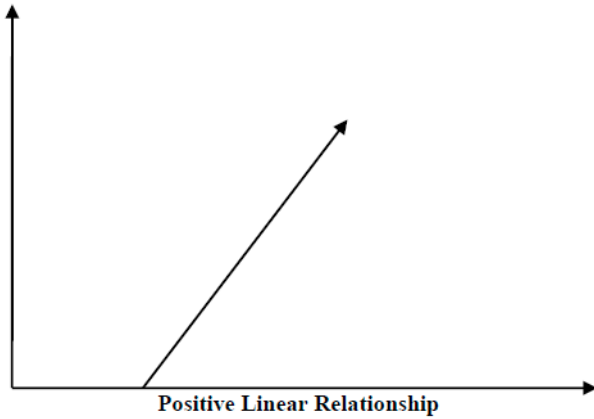
X is the dependent variable we are using to make predictions.

m is the slop of the regression line which represents the effect X has on Y

b is a constant, known as the Y-intercept. If $X = 0$, Y would be equal to b.

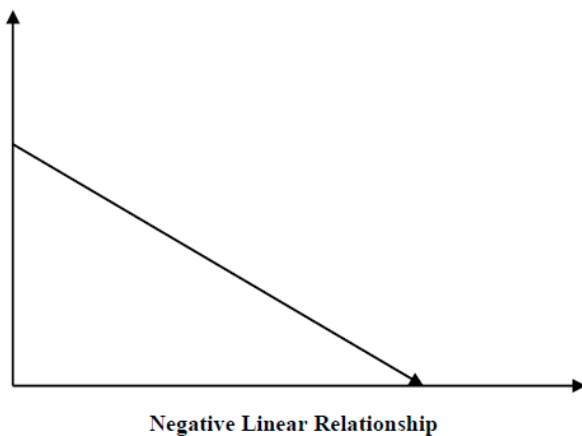
Positive Linear Relationship

A linear relationship will be called positive if both independent and dependent variable increases. It can be understood with the help of following graph –



Negative Linear relationship

A linear relationship will be called positive if independent increases and dependent variable decreases. It can be understood with the help of following graph



Types of Linear Regression

Linear regression is of the following two types –

Simple Linear Regression

Multiple Linear Regression

Simple Linear Regression (SLR)

It is the most basic version of linear regression which predicts a response using a single feature. The assumption in SLR is that the two variables are linearly related.

Multiple Linear Regression (MLR)

It is the extension of simple linear regression that predicts a response using two or more features. Mathematically we can explain it as follows –

Consider a dataset having n observations, p features i.e. independent variables and y as one response i.e. dependent variable the regression line for p features can be calculated as follows –

$$h(x_i) = b_0 + b_1x_{i1} + b_2x_{i2} + \dots + b_px_{ip}$$

Here, $h(x_i)$ is the predicted response value and $b_0, b_1, b_2, \dots, b_p$ are the regression coefficients.

Multiple Linear Regression models always includes the errors in the data known as residual error which changes the calculation as follows –

$$h(x_i) = b_0 + b_1x_{i1} + b_2x_{i2} + \dots + b_px_{ip} + e_i$$

However, before we conduct linear regression, we must first make sure that four assumptions are met:

1. Linear relationship: There exists a linear relationship between the independent variable, x, and the dependent variable, y.
2. Independence: The residuals are independent. In particular, there is no correlation between consecutive residuals in time series data.
3. Homoscedasticity: The residuals have constant variance at every level of x.
4. Normality: The residuals of the model are normally distributed.

If one or more of these assumptions are violated, then the results of our linear regression may be unreliable or even misleading.

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

Answer: Anscombe's quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting data before you analyze it and build your model. These four data sets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets. However, when you plot these data sets, they look very different from one another.

Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

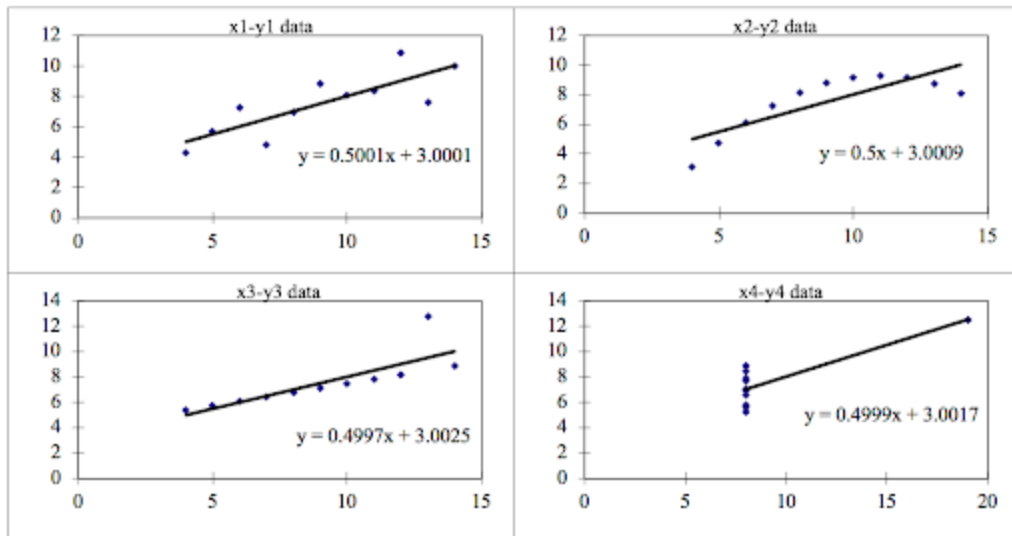
We can define these four plots as follows:

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	

The statistical information for these four data sets are approximately similar. We can compute them as follows:

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		

However, when these models are plotted on a scatter plot, each data set generates a different kind of plot that isn't interpretable by any regression algorithm, as you can see below:



We can describe the four data sets as:

ANSCOMBE'S QUARTET FOUR DATASETS

Data Set 1: fits the linear regression model pretty well.

Data Set 2: cannot fit the linear regression model because the data is non-linear.

Data Set 3: shows the outliers involved in the data set, which cannot be handled by the linear regression model.

Data Set 4: shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

As you can see, Anscombe's quartet helps us to understand the importance of data visualization and how easy it is to fool a regression algorithm. So, before attempting to interpret and model the data or implement any machine learning algorithm, we first need to visualize the data set in order to help build a well-fit model.

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

Answer: The Pearson correlation coefficient (r) is the most widely used correlation coefficient and is known by many names:

Pearson's r

Bivariate correlation

Pearson product-moment correlation coefficient (PPMCC)

The correlation coefficient

The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.

Although interpretations of the relationship strength (also known as effect size) vary between disciplines, the table below gives general rules of thumb:

Pearson correlation coefficient (r) value	Strength	Direction
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None
Between 0 and $-.3$	Weak	Negative
Between $-.3$ and $-.5$	Moderate	Negative
Less than $-.5$	Strong	Negative

The Pearson correlation coefficient is also an inferential statistic, meaning that it can be used to test statistical hypotheses. Specifically, we can test whether there is a significant relationship between two variables.

Another way to think of the Pearson correlation coefficient (r) is as a measure of how close the observations are to a line of best fit.

The Pearson correlation coefficient also tells you whether the slope of the line of best fit is negative or positive. When the slope is negative, r is negative. When the slope is positive, r is positive.

When r is 1 or -1 , all the points fall exactly on the line of best fit:

When r is greater than .5 or less than $-.5$, the points are close to the line of best fit:

When r is between 0 and .3 or between 0 and $-.3$, the points are far from the line of best fit:

When r is 0, a line of best fit is not helpful in describing the relationship between the variables:

The Pearson correlation coefficient (r) is one of several correlation coefficients that you need to choose between when you want to measure a correlation. The Pearson correlation coefficient is a good choice when all of the following are true:

Both variables are quantitative: You will need to use a different method if either of the variables is qualitative.

The variables are normally distributed: You can create a histogram of each variable to verify whether the distributions are approximately normal. It's not a problem if the variables are a little non-normal.

The data have no outliers: Outliers are observations that don't follow the same patterns as the rest of the data. A scatterplot is one way to check for outliers—look for points that are far away from the others.

The relationship is linear: "Linear" means that the relationship between the two variables can be described reasonably well by a straight line. You can use a scatterplot to check whether the relationship between two variables is linear.

Calculating the Pearson correlation coefficient

Below is a formula for calculating the Pearson correlation coefficient (r):

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

Step 1: Calculate the sums of x and y

Step 2: Calculate x^2 and y^2 and their sums

Step 3: Calculate the cross product and its sum

Step 4: Calculate r

Example: Calculating r $n = 10$

$$\sum x = 33.5$$

$$\sum y = 501.2$$

$$\sum x^2 = 113.05$$

$$\sum y^2 = 25\,264$$

$$\sum xy = 1\,684.2$$

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

$$r = \frac{10 \sum 1\,684.2 - (33.5)(501.2)}{\sqrt{[(10)(113.05) - (33.5)^2][(10)(25\,264) - (501.2)^2]}}$$

$$r = \frac{16\,842 - 16\,790.2}{\sqrt{[1\,130.5 - 1\,122.25][252\,640 - 251\,201.4]}}$$

$$r = \frac{51.8}{\sqrt{11\,868.45}}$$

$$r = 0.47$$

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

Answer: It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalization/Min-Max Scaling:

It brings all of the data in the range of 0 and 1. `sklearn.preprocessing.MinMaxScaler` helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Standardization Scaling:

Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

$$\text{Standardisation: } x = \frac{x - \text{mean}(x)}{\text{sd}(x)}$$

`sklearn.preprocessing.scale` helps to implement standardization in python.

One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

Answer: If there is perfect correlation, then $VIF = \infty$. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get $R^2 = 1$, which leads to $1/(1-R^2)$ infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

If VIF is large and multicollinearity affects your analysis results, then you need to take some corrective actions before you can use multiple regression. Here are the various options:

One approach is to review your independent variables and eliminate terms that are duplicates or not adding value to explain the variation in the model. For example, if your inputs are measuring the weight in kgs and lbs then just keep one of these variables in the model and drop the other one. Dropping the term with a large value of VIF will hopefully fix the VIF for the remaining terms and now all the VIF factors are within the threshold limits. If dropping one term is not enough, then you may need to drop more terms as required.

A second approach is to use principal component analysis and determine the optimal set of principal components that best describe your independent variables. Using this approach will get rid of your multicollinearity problem but it may be hard for you to interpret the meaning of these “new” independent variables.

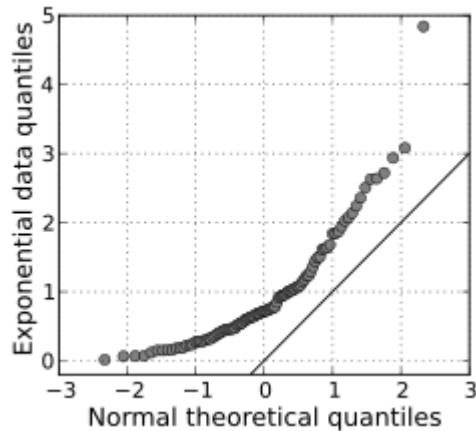
The third approach is to increase the sample size. By adding more data points to our model, hopefully, the confidence intervals for the model coefficients are narrower to overcome the problems associated with multicollinearity.

The fourth approach is to transform the data to a different space like using a log transformation so that the independent variables are no longer correlated as strongly with each other.

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

Answer: Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q-Q plots is to find out if two sets of data come from the same distribution. A 45-degree angle is plotted on the Q-Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

A Q-Q plot showing the 45-degree reference line:



If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line $y = x$. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line $y = x$. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.

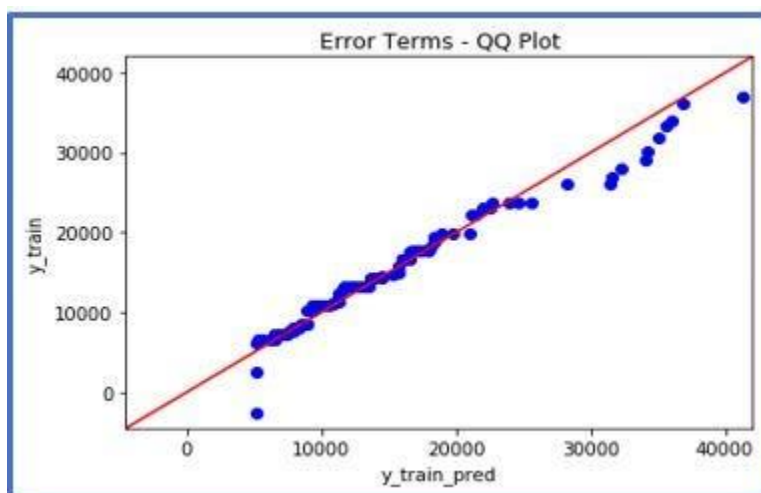
Interpretation:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

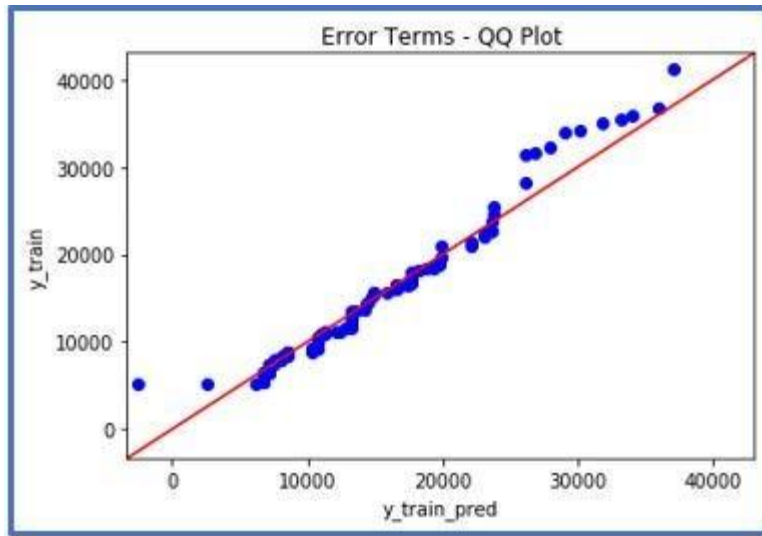
Below are the possible interpretations for two data sets.

a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.



c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.



d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis

Python:

statsmodels.api provide qqplot and qqplot_2samples to plot Q-Q graph for single and two different data sets respectively.