

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?

Year has a major impact on the bike rentals and we can see a steady growth in 2019 compare to 2018.

Weathersituation also has significant impact on the number of bike rentals. People tend to rent bikes more when weather is clear or there is no rain or snow. And we have seen rental goes down whenever there is any rain or snow.

Season has positive impact on bike rentals. Summer and winter has significant impact on the dependent variable.

Month also showing some pattern. Bike booking were high during the month of May to September.

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

`drop_first=True` is important to use, as it helps in reducing the extra column created during dummy variable creation because keeping k dummies for k levels of a categorical variable is not needed since one of the combination will be uniquely representing this redundant column. Hence, its better to drop one of the column and just have $k-1$ dummies(columns) to represent k levels. This Overall approach reduces Multi-collinearity in the dataset, which is one of the prime Assumption of Multiple Linear Regression.

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

Looking at the pair-plot among the numerical variables, "temp" has the highest correlation with the target variable closely followed by "atemp".

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

- The distribution plot of error term for both training and test set shows the **normal distribution with mean at 0**.
- The residual plot for the training set is **reasonably random**.
- Also the error terms satisfies to have reasonably **constant variance or standard deviation** across all values of the predictors (**homoscedasticity**).

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

1) Temperature (0.548): A coefficient value of 0.548 indicated that a temperature has significant impact on bike rentals.

2) Weathersit: light_rain_snow(-0.283): A coefficient value of -0.283 indicated that the light rain and snow deters people from renting out bikes.

3) year (0.233) : A coefficient value of -0.283 indicated that year wise bike rental numbers are increased.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a supervised machine learning method that is used to find a linear equation that helps us to determine the strength of the relationship between one dependent variable and several independent variables. This is achieved by fitting a line to the data.

In other words we can say, two variables which are on the x-axis and y-axis should be linearly correlated and mathematically, we can write a linear regression equation as:

$$y = a + bx$$

Where, Here, x and y are two variables on the regression line.

b = Slope of the line

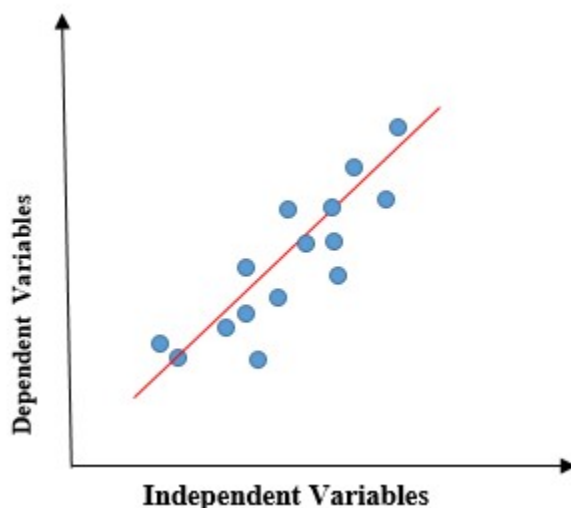
a = y-intercept of the line

x = Independent variable from dataset

y = Dependent variable from dataset

The process of regression is mostly used to forecast and determine relationships between two or more variables. The cost function is used to find out the actual learning rate by determining the error between the predicted value and the actual value. The concept of gradient descent is used to understand how big or small the learning rate has been in every step of the way. More so, it determines the number of steps that are needed for the algorithm to learn.

If there is a single input variable (x), such linear regression is called simple linear regression. And if there is more than one input variable, such linear regression is called multiple linear regression. The linear regression model gives a sloped straight line describing the relationship within the variables.



The above graph presents the linear relationship between the dependent variable and independent variables. When the value of x (independent variable) increases, the value of y (dependent variable) is likewise increasing. The red line is referred to as the best fit straight line. Based on the given data points, we try to plot a line that models the points the best.

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 was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties. There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.

This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.

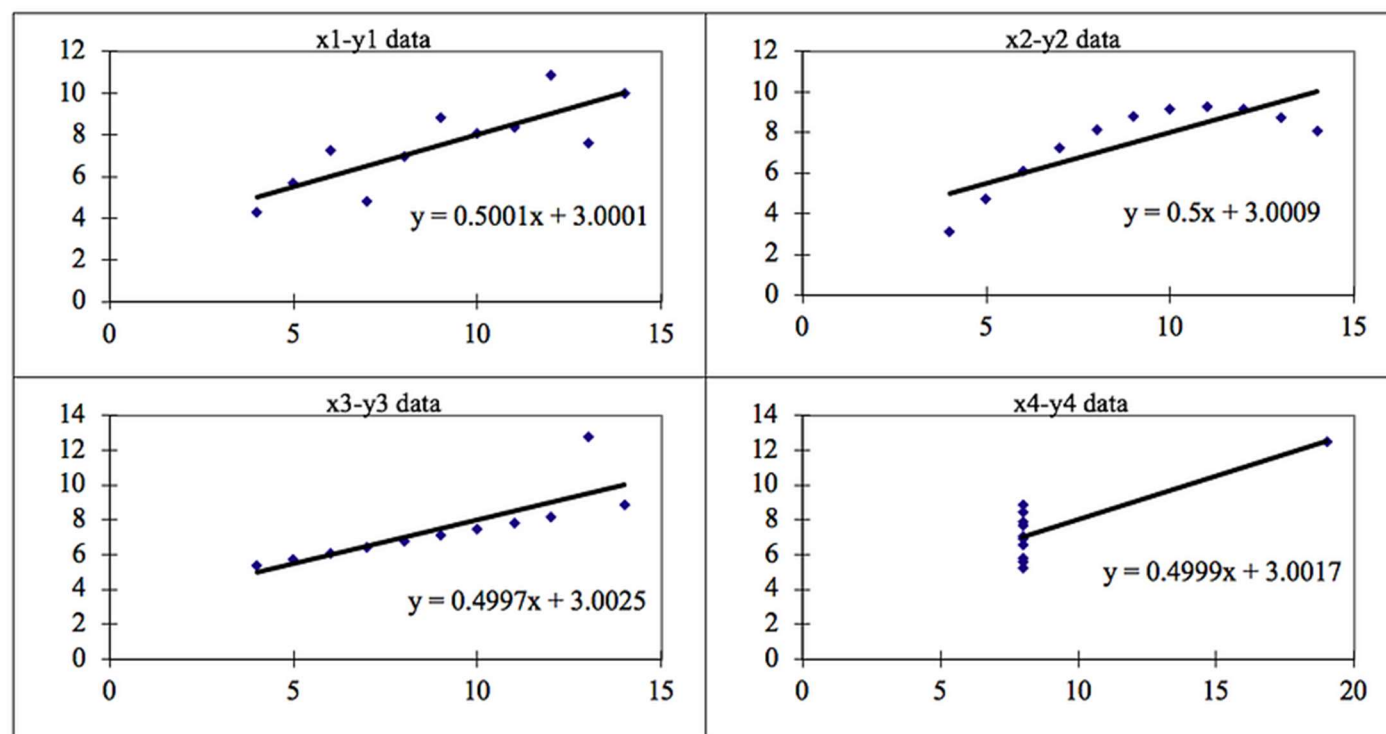
These four plots can be defined as follows:

| 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 all these four datasets are approximately similar and can be computed 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 | |

When these models are plotted on a scatter plot, all datasets generates a different kind of plot that is not interpretable by any regression algorithm which is fooled by these peculiarities and can be seen as follows:



The four datasets can be described as:

Dataset 1: this fits the linear regression model pretty well.

Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.

Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model

Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model

3. What is Pearson's R?

In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r , the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between -1 and 1 .

The Pearson's correlation coefficient varies between -1 and $+1$ where:

- $r = 1$ means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- $r = -1$ means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- $r = 0$ means there is no linear association
- $r > 0 < 5$ means there is a weak association
- $r > 5 < 8$ means there is a moderate association
- $r > 8$ means there is a strong association

Pearson r Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Here,

- r = correlation coefficient
- x_i = values of the x -variable in a sample
- \bar{x} = mean of the values of the x -variable
- y_i = values of the y -variable in a sample
- \bar{y} = mean of the values of the y -variable

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

What is scaling?

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.

Why is scaling performed?

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).

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

| NORMALIZATION | STANDARDIZATION |
|---|---|
| Minimum and maximum value of features are used for scaling. | Mean and standard deviation is used for 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. |
| This transformation squishes the n-dimensional data into an n-dimensional unit hypercube. | It translates the data to the mean vector of original data to the origin and squishes or expands. |
| It is useful when we don't know about the distribution | It is useful when the feature distribution is Normal or Gaussian. |

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

If there is perfect correlation between two independent variables, then $VIF = \infty$.

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).

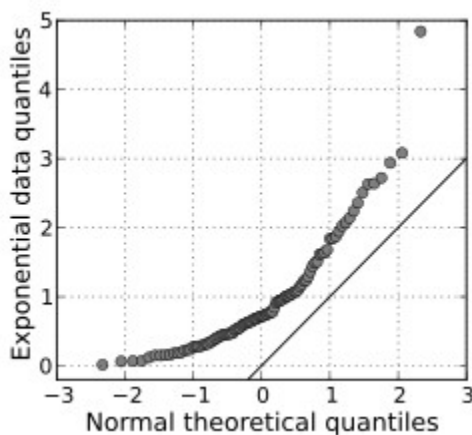
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Q-Q plots are also known as 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.

As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.

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.