Linear Regression Assignment 1

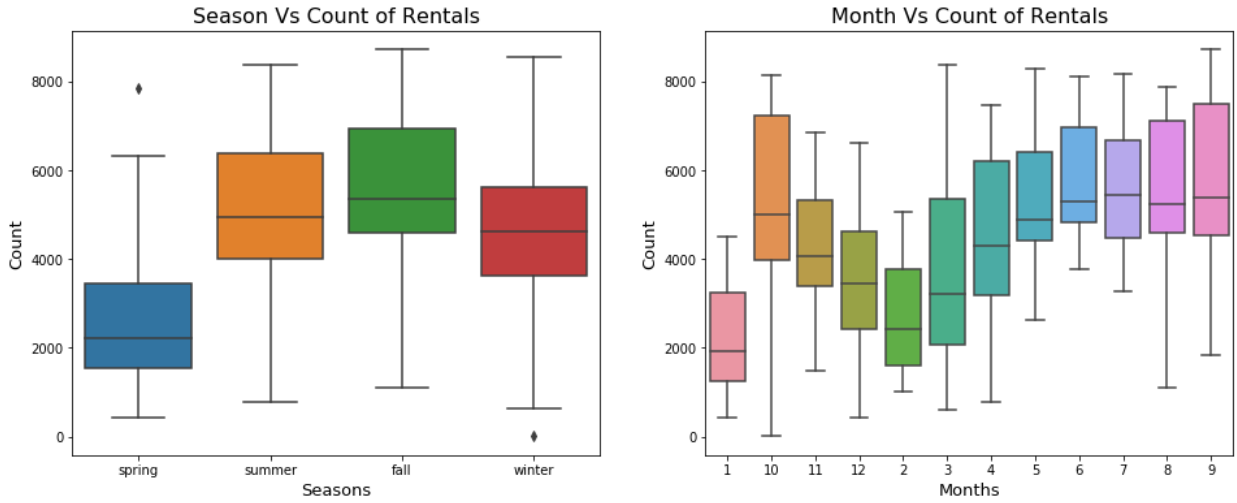
Abhijith Sharma (abhijiths16.instru@coep.ac.in)

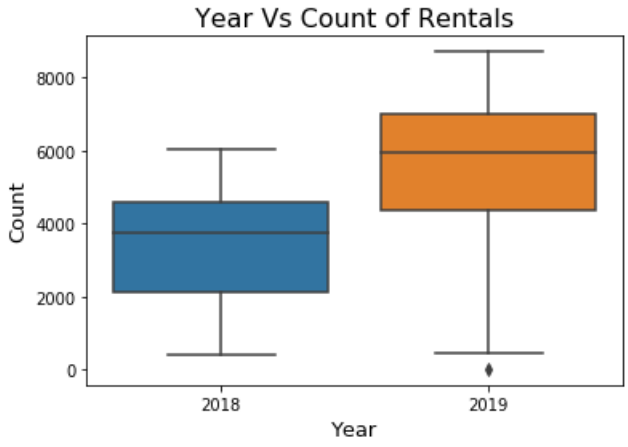
September 14, 2020

ASSIGNMENT BASED SUBJECTIVE QUESTIONS

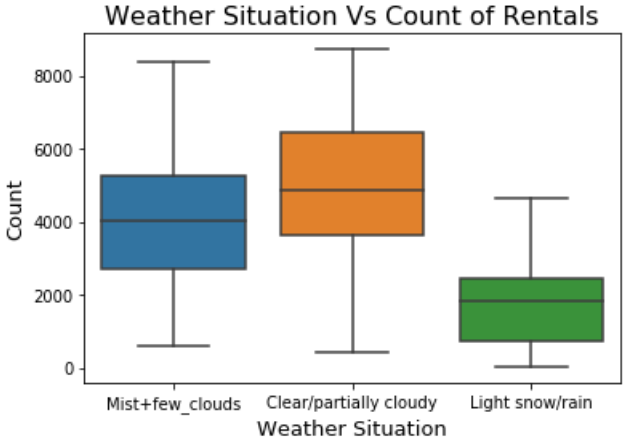
1. From your analysis of the categorical variables from the data set, what could you infer about their effect on the dependent variable?

Our target variable is continuous variable and to understand the influence of categorical variables on target variable, it is best to have box plots. We can clearly see that the season and more precisely month have an influence over the bikes rented. Fall (July-October) have seen the highest rentals and spring months like (february-april) has seen the lowest ones.

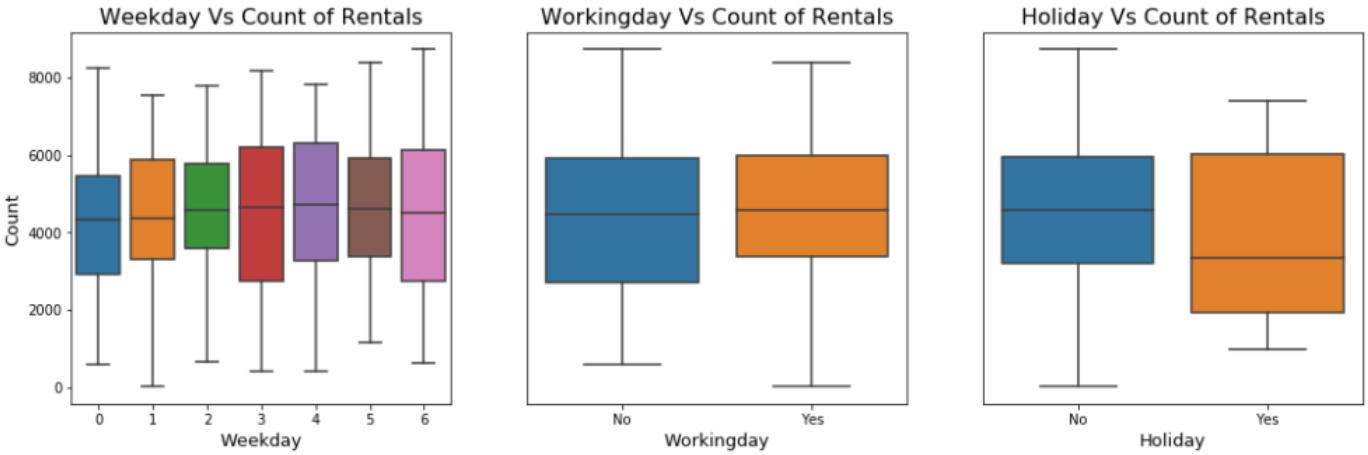




We can see that the business has expanded from 2018 to 2019 which states that people like the rental boom bikes



We can observe that during snow the count is less whereas as the weather is clear/partially cloudy the count is more.



We do not see significant difference in bike rentals during holidays and non holidays from above plots.

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

This is important because for n dummy variables nth variable would be redundant as it can occur when all other variables are zero which means that it occurs when all other variables doesn't occur.
Lets us consider an example for intuitive understanding. I invited three friends to my house, Ram, Suresh and Ramesh. My mother already knew their names as they were my friends but had never seen them. When they came, they introduced themselves to my mother. So the first one stepped forward and introduced himself as Ram and similarly second one as Suresh. Now my mother without even introduction knew that third one is Ramesh. This is the same what happens with dummy variables.
Like during dummification of say months I just need 11 variables and I would automatically know that if its not among these 11 variables it would definitely be the one which is dropped.

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

The variable having highest correlation with target variable is temp or atemp. They both have the same correlation of 0.63.

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

The assumptions can be validated by plotting suitable graphs:

- **Residual Plot**: This shows the predicted value (based on the regression equation) on the X axis, and the residuals on the Y axis. The residuals are essentially the difference between the predicted value and the actual value. Hence this plot is used to check homoscedasticity of residual.
- **Q-Q Normal plot**: This is used to assess if your residuals are normally distributed. This can also be evaluated by plotting histogram of residuals
- **Error Plot**: Plotting error against target variable in X Axis helps to understand the independence of error which is a assumption of linear regression. Ideally the distribution should be random with no emerging pattern.

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

The top three contributing variables are:

- **Light snow**: It has a negative influence on target variable that is the count reduces when it snows.(Coefficient -1.313)
- **Year**: It has a positive influence on the target variable.(Coefficient 1.0330)
- **Temperature**: It has a positive influence on target variables that is higher the temperature high is the probability of count being increased.(Coefficient 0.595)

GENERAL SUBJECTIVE QUESTIONS

6. Explain the linear regression algorithm in detail

Linear regression is a algorithm used for predictive analysis. The overall idea of regression is to examine two things:

- Does a set of predictor variables do a good job in predicting an outcome (dependent) variable?
- Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable?

In linear regression we try to find a line that best fits the data. One way to achieve this is to use the least squares criterion, which minimizes the sum of all the squared prediction errors. The equation in given by:

$$y_{Predicted} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_1 x_1 + \ldots\ldots + \beta_k x_k$$

This is then used to calculate cost function given by based on least squared criterion:

$$\frac{Min}{J\left(\beta_0, \beta_1, .. \beta_k\right)} \quad where,\ J\left(\beta_0, \beta_1, .. \beta_k\right) = \sum_{i=1}^{n} {\varepsilon_i}^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2$$

This cost function has to be optimised. It can be done in two ways
Matrix method in which we directly find β by matrix multiplication of inverse. However, the short coming of this method is that the inverse may not always exist.

$$\beta = (X^T X)^{-1} X^T y \quad where,\ \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ \beta_k \end{bmatrix} X = \begin{bmatrix} 1 & x_{11} & x_{12} & . & . & x_{1k} \\ 1 & x_{21} & x_{22} & . & . & x_{2k} \\ . & . & . & & & . \\ . & . & . & & & . \\ 1 & x_{n1} & x_{n2} & . & . & x_{nk} \end{bmatrix} y = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix}$$

Another is a numerical method that is gradient descent. In this we start with a guess of β and keep updating until convergence. The update is done as shown below:

$$\beta_0 := \beta_0 - \eta \frac{\partial}{\partial \beta_0} J(\beta_0, \beta_1 \ldots \beta_k) = \beta_0 - \eta \frac{2}{n} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right) . x_{i0}$$

$$\beta_1 := \beta_1 - \eta \frac{\partial}{\partial \beta_1} J(\beta_0, \beta_1 \ldots \beta_k) = \beta_1 - \eta \frac{2}{n} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right) . x_{i1}$$

.

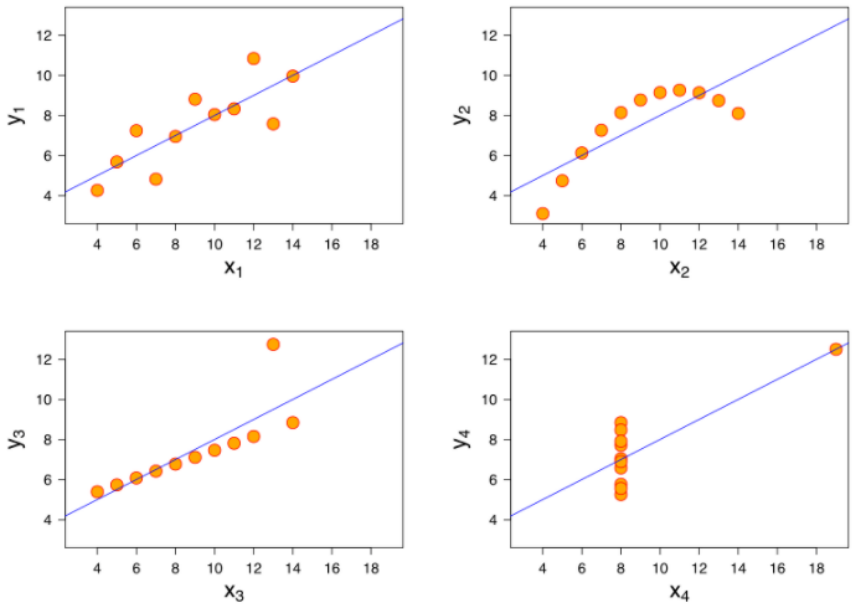
.

$$\beta_k := \beta_k - \eta \frac{\partial}{\partial \beta_k} J(\beta_0, \beta_1 \ldots \beta_k) = \beta_k - \eta \frac{2}{n} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right) . x_{ik}$$

The β formulated describes the significance of a particular attribute. That is higher the magnitude of β, more is the contribution of this variable in the prediction. The sign of β depicts whether the attribute has positive or negative influence

7. Explain the Anscombe's quartet in detail

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.



The plots can be described as follows.

- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables

correlated where y could be modelled as gaussian with mean linearly dependent on x.

- The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant.
- In the third graph (bottom left), the distribution is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- The fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

Yet the four graphs has exactly same mean, variance, correlation as well as same line as linear regression fit. Hence this quartet strongly demonstrate the idea that mere statistics can be misleading in some scenarios.Statistician Francis Anscombe made this quartet to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough.

8. What is Pearson's R?

Pearson's correlation coefficient also known as Pearson's R is the test statistics that measures the statistical relationship, or association, between two continuous variables. It basically gives the measure of linear relationship among two variables. It has a value between + 1 and 1. A value of + 1 means total positive linear correlation, 0 means no linear correlation, and 1 means total negative linear correlation.
Pearson's correlation coefficient is the co-variance of the two variables divided by the product of their standard deviations. Given a pair of random variables **(X,Y)**, the formula for ρ is given by: Assumptions while calculating ρ are:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

where
$\sigma_X$is the standard deviation of X
$\sigma_Y$is the standard deviation of Y

- **Independent of case**: Cases should be independent to each other.
- **Linear relationship**: Two variables should be linearly related to each other. This can be assessed with a scatterplot: plot the value of variables on a scatter diagram, and check if the plot yields a relatively straight line.
- **Homoscedasticity**: the residuals scatterplot should be roughly rectangular-shaped.

The plots below shows how the Pearson's coefficient varies in different scenarios.

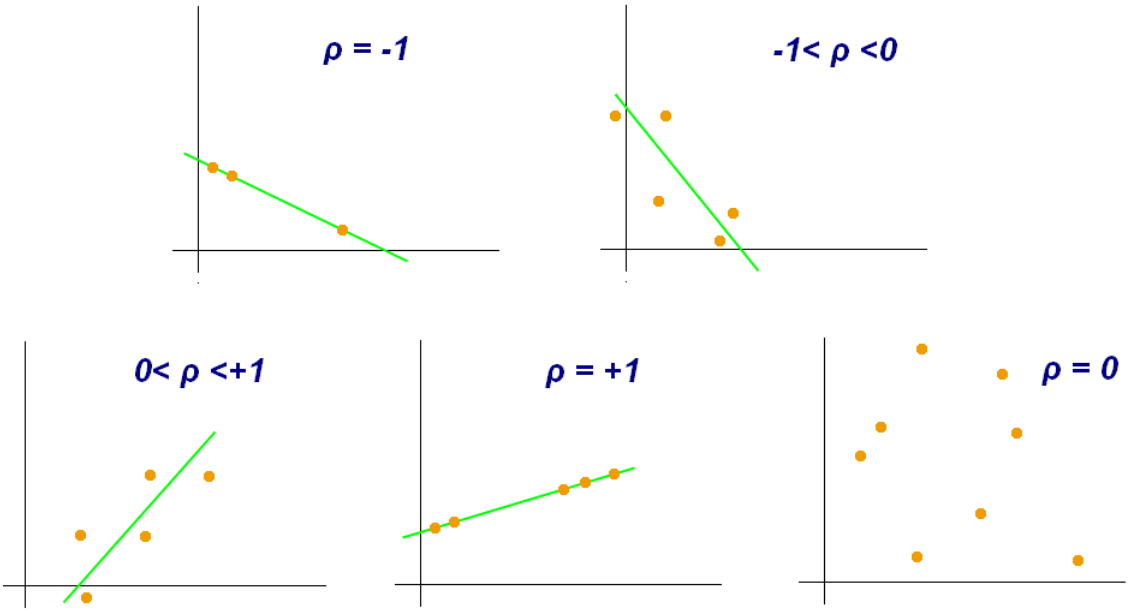


Figure 1: Source:Wikipedia

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

Feature scaling is a method used to normalize the range of independent variables or features of data. The scaling is performed for the following reasons:

- Having features on a similar scale can help the gradient descent converge more quickly towards the minima.
- When we scale our data before employing a distance based algorithm so that all the features contribute equally to the result and hence interpretation of coefficients becomes easy.

Normalization: Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling. The formula is:

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

Standardization Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation. The formula is:

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

μ is the mean of the feature values and σ is the standard deviation of the feature values.
When to use Normalisation and Standardisation?

- Normalization is good to use when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks.
- Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

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

VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables. For example, we would fit the following models to estimate the coefficient of determination R1 and use this value to estimate the VIF:

$$X_1 = \alpha_2 X_2 + \alpha_3 X_3 + ... \tag{1}$$

$$VIF_1 = \frac{1}{1 - {R_1}^2} \tag{2}$$

Next, we fit the model between X2 and the other independent variables to estimate the coefficient of determination R2:

$$X_2 = \alpha_1 X_1 + \alpha_3 X_3 + ... \tag{3}$$

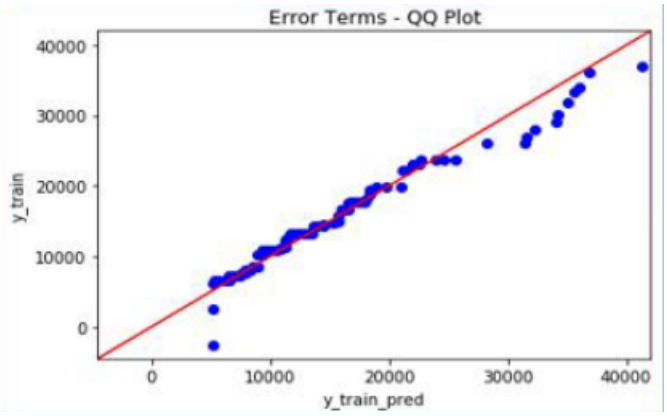
$$VIF_2 = \frac{1}{1 - {R_2}^2} \tag{4}$$

If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity. A general rule of thumb is that if VIF $>$ 5 then there is multi-collinearity. Hence to conclude if VIF is infinite this means that the attribute can be perfectly explained by the combinations of other variables in the data set i.e the attribute is perfectly co-linear with rest of the variables.

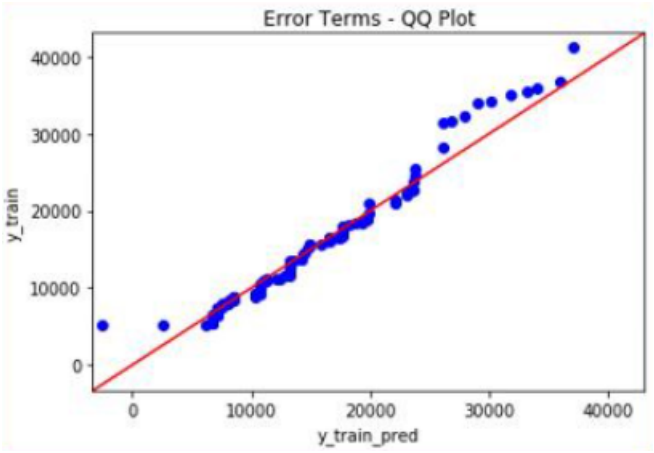
11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

a Q–Q (quantile-quantile) plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. The Q-Q plot can be interpreted as:

- Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis.
- Y-values $<$ X-values: If y-quantiles are lower than the x-quantiles.



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



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

In linear regression, normal Q-Q Plot is used to assess if residuals are normally distributed. Basically the data points must closely follow the straight line at a 45 percent angle upwards (left to right). Any patterns that deviate from this particularly anything that looks curvilinear (bending at either end) or s shaped is not desired and would fail the assumption of linear regression.