# BIKE SHARING ASSIGNMENT

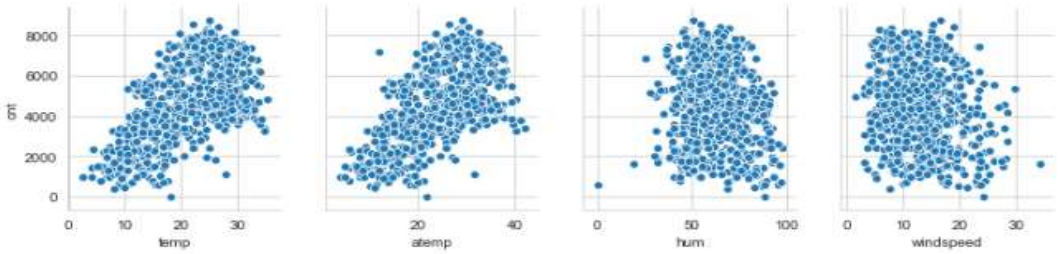
By: Hamza Tanveer

## Assignment-based Subjective Questions

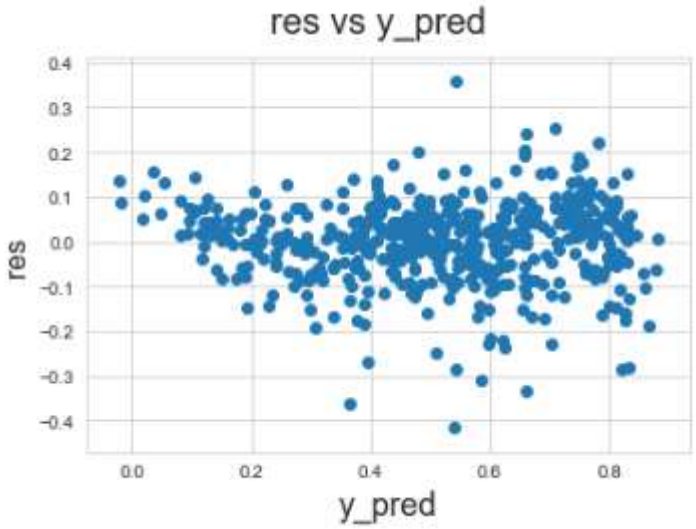
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**
   The effect of the categorical variables present in the dataset are as follows:
   a. SEASON: Spring has the least demand, whereas fall has the highest demand and extreme season (winter and summer) has intermediate demands.
   b. MONTH: The demand is highest in September ( in-line with season as fall is around September) and low during winter months (January and February)
   c. HOLIDAYS: Demand decreases during the holidays
   d. WEATHERSIT: The demands are highest for the 'Clear Day' and there is no demand during 'Heavy Rain'
   e. YR: The demand is higher in 2019 then of 2018 as business grew over a year.
   f. WEEKDAY: There is no as such impact of weekday on the demand.

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 columns created during the dummy variable creation. Hence it reduces the correlations created among the dummy variables as the same amount of information can be conveyed with one-less dummy variable.

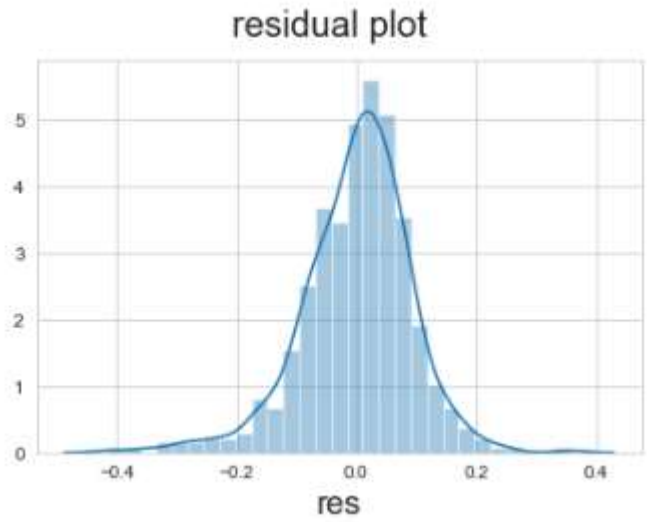
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
   Temp (atemp is highly co-linear with the temp) has the highest correlation with the target variable.



4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
   a. LINEAR RELATIONSHIP BETWEEN X AND Y: Residual vs Fitted values plot can be used to validate this assumption; no pattern is signifies there is linearity in data.



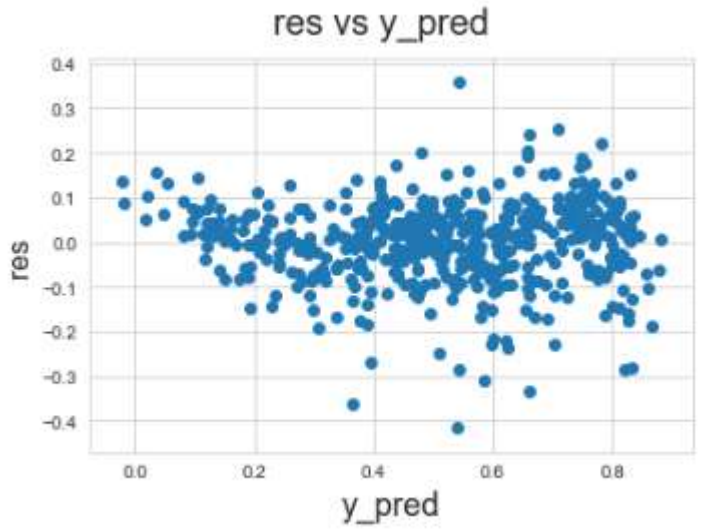

   b. ERROR TERMS ARE NORMALLY DISTRIBUTED WITH THE MEAN ZERO: Histogram of residuals shows that residuals are normally distributed with mean of zero.




   c. ERROR TERMS ARE INDEPENDENT OF EACH OTHER: Durbin – Watson(DW) Statistic should lie between 0-4 and for our model it is 2.051

| | | | |
|---|---|---|---|
| **Omnibus:** | 59.182 | **Durbin-Watson:** | 2.051 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 134.016 |
| **Skew:** | -0.629 | **Prob(JB):** | 7.92e-30 |
| **Kurtosis:** | 5.173 | **Cond. No.** | 17.3 |

d. ERROR TERMS HAVE CONSTANT VARIANCE (HOMOSCEDASTICITY): Residual vs Fitted values plot can be used to validate this assumption; no funnel shape pattern is present thus homoscedastic.




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 contributing features are:
   a. TEMP: Temperature is the major factor contributing to the demand of shared bikes. The demand is high for the moderate temperature (20-30) degree Celsius
   b. WEATHER: Snowy weather has a negative impact on the demand of shared bikes. The demand is high for the clear weather
   c. SEASON: During summers demand for the shared bike increases.

## General Subjective Questions

1. **Explain the linear regression algorithm in detail.**
   Linear regression is a type of supervised machine learning algorithm that is used for the prediction of numeric values. Linear regression is a linear model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x).
   There are two types of linear regression:
   SINGLE LINEAR REGRESSION: When there is a single input variable (x), it is referred to as simple linear regression.
   The general equation for simple linear regression is as follows:

   $$y(pred) = b0 + b1*X$$

   where b0 is constant and b1 is coefficient for x.
   MLTIPLE LINEAR REGRESSION: When there are multiple input variables, it is referred as multiple linear regression.
   The general equation for multiple linear regression is as follows:

y(pred)= b0 + b1*X1 + b2*X2 + b3*X3 + …..

where b0 is constant and b1,b2 ,b3 ,… are coefficients of X1, X2, X3 respectively

The most common technique to train linear regression equation is Ordinary Least Squares, it seeks to minimize the sum of squared residuals.

However there are certain assumptions that for building machine linear model:

a. There is a linear relationship between X and Y
b. Error terms are normally distributed with mean zero (not X and Y)
c. Error terms are independent o each other
d. Error terms have constant variance(homoscedasticity)

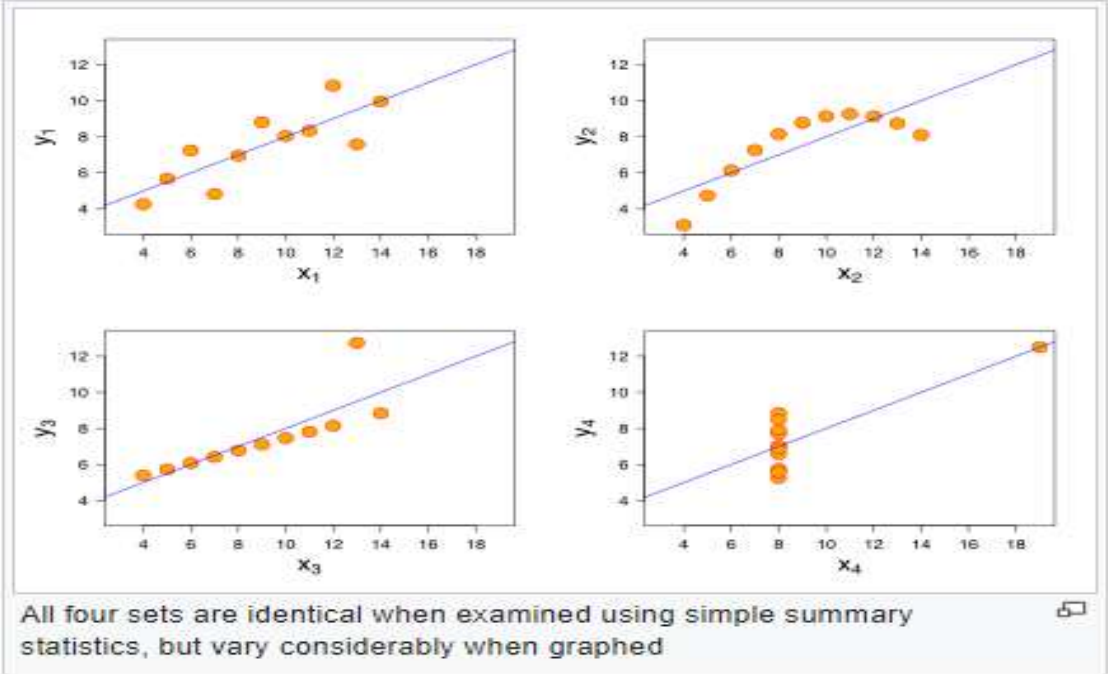
**2. Explain the Anscombe's quartet in detail.**

Anscombe's quartet was constructed by statistician Francis Anscombe. It includes four data sets that have almost identical statistical features, but they have a very different distribution and look totally different when plotted on a graph.

It was developed to emphasize

a. The importance of plotting data
b. Effect of outliers and other influential observations on statistical properties

The plot is as follows for 4 datasets:




All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

The summary for the dataset is

| Property | Value | Accuracy |
|---|---|---|
| Mean of $x$ | 9 | exact |
| Sample variance of $x$ : $s_x^2$ | 11 | exact |
| Mean of $y$ | 7.50 | to 2 decimal places |
| Sample variance of $y$ : $s_y^2$ | 4.125 | ±0.003 |
| Correlation between $x$ and $y$ | 0.816 | to 3 decimal places |
| Linear regression line | $y = 3.00 + 0.500x$ | to 2 and 3 decimal places, respectively |
| Coefficient of determination of the linear regression : $R^2$ | 0.67 | to 2 decimal places |

a. The first scatter plot (top left) appears to be a simple linear relationship
b. The second graph (top right) is not distributed normally; while a relationship between the two variables is not linear
c. In the third graph (bottom left), the distribution is linear, but have a different regression line. The calculated regression is offset by the one outlier
d. Finally, the fourth graph (bottom right) shows 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.

**3. What is Pearson's R?**

Pearson's R is a correlation coefficient. It is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and 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.In simple terms, it shows the linear relationship between tow datasets.

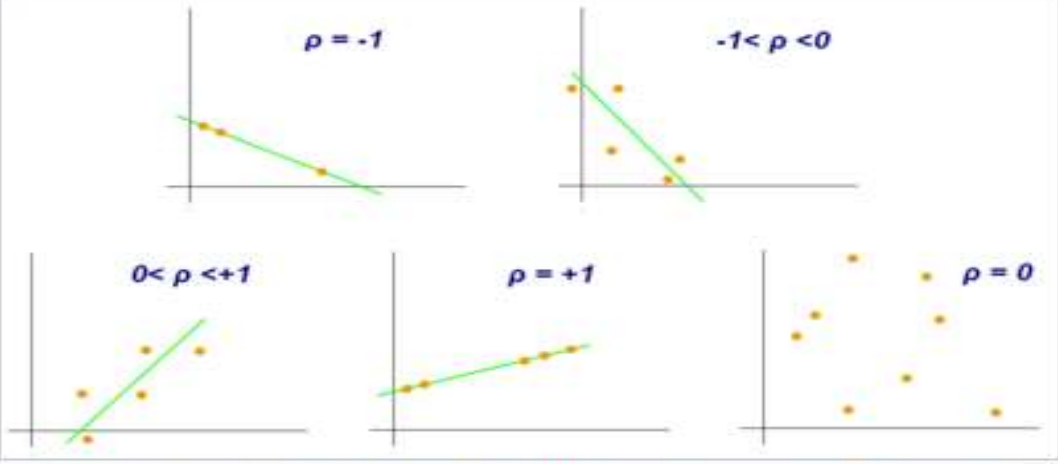
For a population it is given by,

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} \quad \text{(Eq.1)}$$

where,

- $\operatorname{cov}$ is the covariance
- $\sigma_X$ is the standard deviation of $X$
- $\sigma_Y$ is the standard deviation of $Y$

Different values of Pearson's R,




Examples of scatter diagrams with different values of correlation coefficient ($\rho$)

Pearson's R > 0 shows there is positive relation between the two datasets
Pearson's R = 0 shows there is no relation between the two datasets
Pearson's R < 0 shows there is negative relation between the two datasets

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

Feature scaling is a step of data Pre-Processing which is applied to independent variables to normalize the range of independent variables or features of data. Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without normalization.

Scaling is performed because of the following reasons:

a. If one of the features has a broad range of values, the target variable will be governed by this particular feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the target variable.
b. Gradient descent converges much faster with feature scaling than without it.

The difference between Normalized scaling and Standard scaling are:

| 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. |

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

Variance Inflation factor (VIF) dontes how much the variance of coefficient estimate is being inflated by collinearity. VIF is given by

$$\text{VIF} = \frac{1}{1-r2}$$

In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity, then VIF = infinity. This shows a perfect correlation between two independent variables. 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 (Quantile-Quantile plots) are plots of two quantiles against each other.

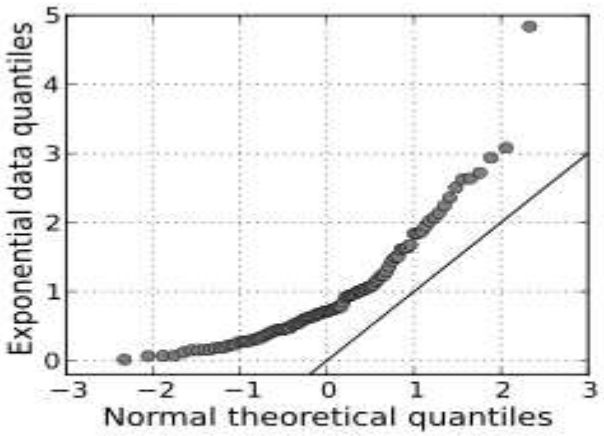
Advantages of Q-Q plot are :

a) It can be used with sample sizes also

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It can be used to check following scenarios: If two data sets
a. Come from populations with a common distribution
b. Have common location and scale
c. Have similar distributional shapes
d. Have similar tail behavior

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, If all point of quantiles lies away from the straight line at an angle of 45 degree from x –axis then it is from different distribution.




*A Q Q plot showing the 45 degree reference line. Image: skbkekas/Wikimedia Commons.*