# Bike Sharing Assignment

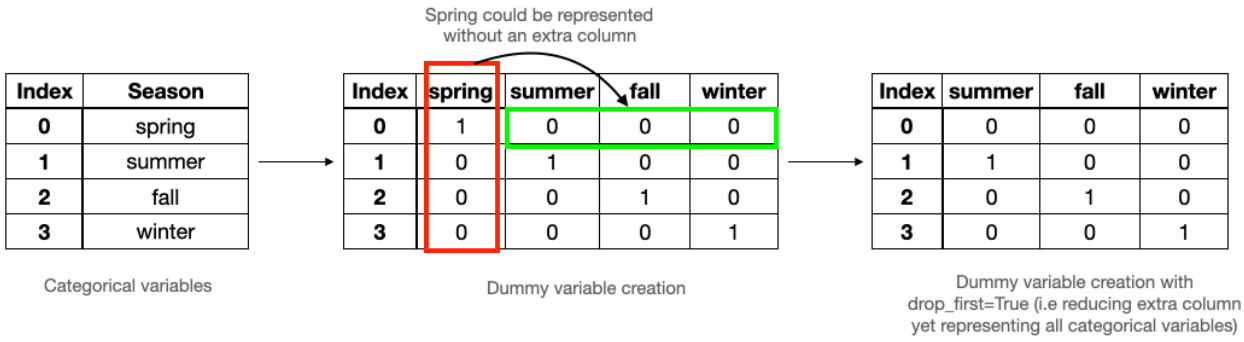
Student: Karthik Jayaraman

**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)**
   Based on the visual analysis (boxplot & bar plot) of categorical variables from the bike sharing dataset (Ref: Section 4b in Jupyter notebook), the following inferences could be made:
   *Note: The total demand of rental bikes includes both casual and registered.*

   a. The overall demand for Bike sharing rental bikes have increased from 2018 to 2019
   b. The average demand for Bike sharing rental bikes is highest in Fall season of 2019
   c. The bike demand trend is considerably high from April to October month.
   d. The bike demand trend decreases from November to March month.
   e. The booking for Rental bikes is high when it's not a holiday as people want to spend time in home and with family.
   f. Demand is higher when the weather situation is good (Clear, Few clouds, Partly cloudy, Partly cloudy)

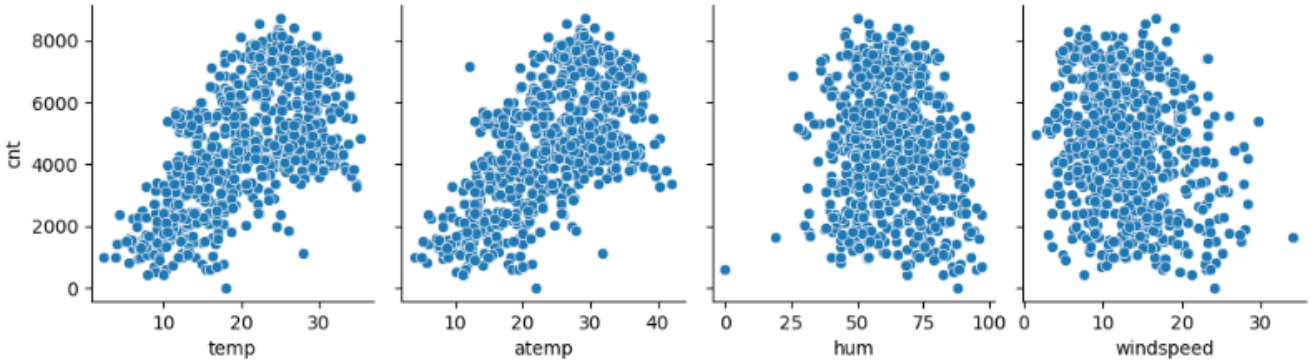
2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
   **(Answer)**
   Machine learning algorithms mostly works with numeric variables to understand relationships. Categorical variables, like season (spring, summer, fall, winter), may have important information for building a machine learning model but would have to be encoded to its numeric representations.
   Dummy variable creation is a process of converting or encoding categorical data to binary representation. (i.e., each variable is converted in as many 0/1 variables as there are different values.)
   Specifically, the Categorical variable with 'n' levels, dummy variable creation process **reduces extra levels (n-1)** each indicating whether that level exists or not using a zero or one.



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

**(Answer)**

Based on the visual analysis (pair plot) of numerical variables from the bike sharing dataset (Ref: Section 4a in Jupyter notebook), **'temp' (temperature in Celsius)** has the highest correlation with the target variable ('cnt' - count of total rental bikes including both casual and registered).

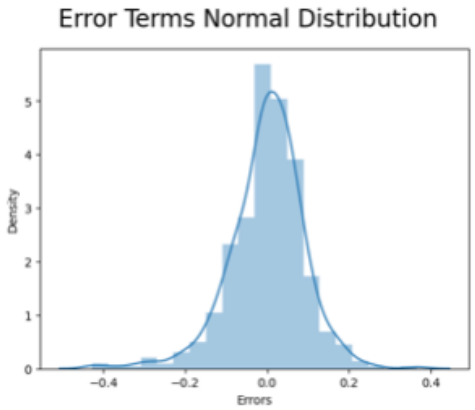


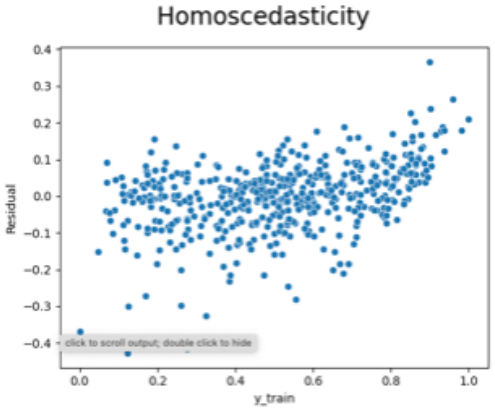

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

**(Answer)**

Based on the learning from course, the following assumptions were validated after building the model on training set:

a. Error terms to be normally distributed.
b. Multicollinearity among predictor variables to be insignificant.
c. Linear relationship among variables to be observed.
d. Homoscedasticity or the spread of errors (between observed & predicted values) in a regression model to be consistent without any visible pattern which will make prediction results more statistically accurate.



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 top 3 features are year, temp (temperature), and months (jul, sep) contributing significantly towards explaining the demand of the shared bikes.

## General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
   **(Answer)**
   In Machine learning broadly there are two variables to work with:
   a. Independent variable (X) –the predictors or features used to predict or explain changes in dependent variable.
   b. Dependent variable (Y) – or the target or prediction expected out of the model.

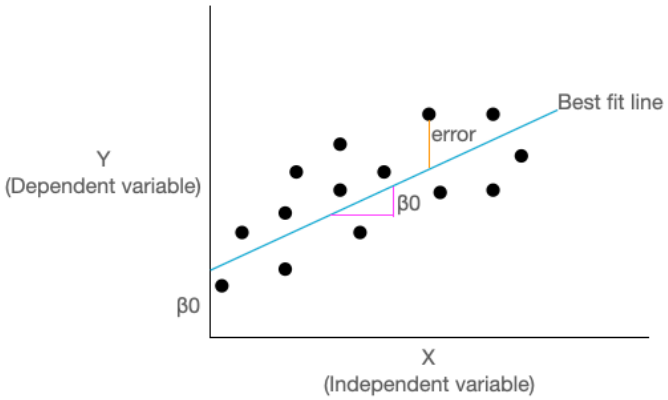
   Example: Marketing budget (X) used to predict Sales (Y)

   An equation with highest degree of 1 is called as a Linear equation.
   The graph of Linear equation forms a straight line.

   In algebraic terms, a Linear equation could be written as:

   $y = c + mx$
   (c = y-intercept & m = slope)

   Similarly, the equation of a Linear regression line is given by:
   $Y = \beta0 + \beta1 * X$
   (β0 = y-intercept & β1 = slope)



The Linear regression algorithms establishes a linear relationship between independent variables and dependent variable by finding the best fit line using observed data.

The best fit line is arrived by minimizing the sum of squared errors (RSS – Residual Sum of squares) by taking each data point in the plot.

The Ordinary Least Squares (OLS) method is employed by calculating the residuals at each data point in the plot.

Residual = Actual value - Predicted value

$e(i) = y(i) - y(pred)$

$RSS = e(1)^2 + e(2)^2 + \ldots + e(n)^2$

$= (y(1) - y(pred1))^2 + (y(2) - y(pred2))^2 + \ldots$

$= (y(1) - (\beta_o + \beta_1.X_1))^2 + \ldots$

$= \text{Sum of } \sum(y(i) - (\beta_o + \beta_1 * Xi))^2$

(where i = 1 to n)

- Goal is Minimize RSS for best value of $\beta_o$ & $\beta_1$

Once the model or best value of Linear regression question (best fit line) is arrived,
The strength of the Linear regression model is assessed using:

a. R2 or Coefficient of determination – Higher the R-squared value, the better fit model
b. Residual Standard Error (RSE) – Lower RSS means tighter model fit.

The following assumptions are also validated after building the model on training set:

a. Error terms to be normally distributed.
b. Multicollinearity among predictor variables to be insignificant.
c. Linear relationship among variables to be observed.
d. Homoscedasticity or the spread of errors (between observed & predicted values) in a regression model to be consistent without any visible pattern which will make prediction results more statistically accurate.

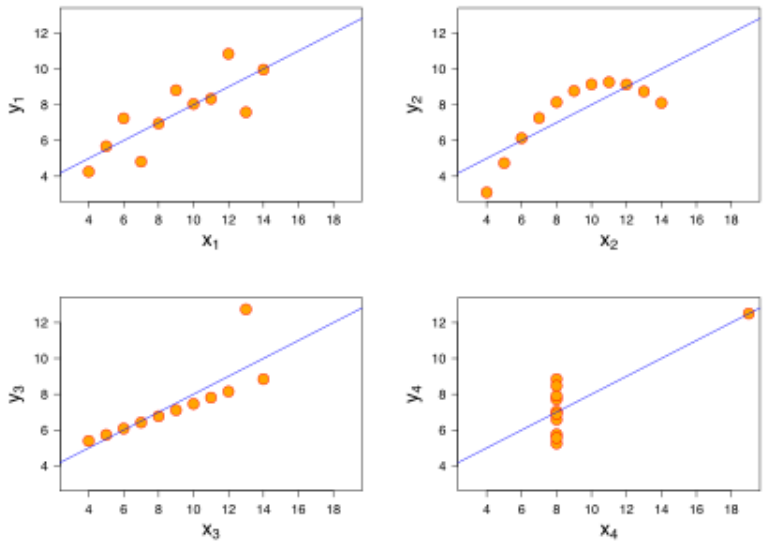
There are two types of Linear regression:

a. Simple Linear Regression (one independent variable)
b. Multiple Linear Regression (several independent variables)

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

In 1973, the statistician Francis Anscombe defined four data sets which are near identical in descriptive statistics (mean, variance, correlation) but appear different when visually observed using a graph (scatter plot).

This emphasizes the important of visualizing data before applying algorithms to build machine learning models.




*(Courtesy: Wikipedia)*

The first scatter plot (top left): Simple linear relationship, corresponding to two variables correlated where y with mean linearly dependent on x.

The second graph (top right): Non-linear relationship and Pearson correlation coefficient is not relevant.

The third graph (bottom left): Linear regression offset by one outlier.

The fourth graph (bottom right): High correlation coefficient, even though the other data points do not indicate any relationship between the variables.

The highlight of Anscombe's quartet, is the below descriptive statistics is nearly same for all four relationships mentioned above:

The mean of x, sample variance of x, mean of y, sample variance of y, correlation between x and y, linear regression line is same across all four data sets.

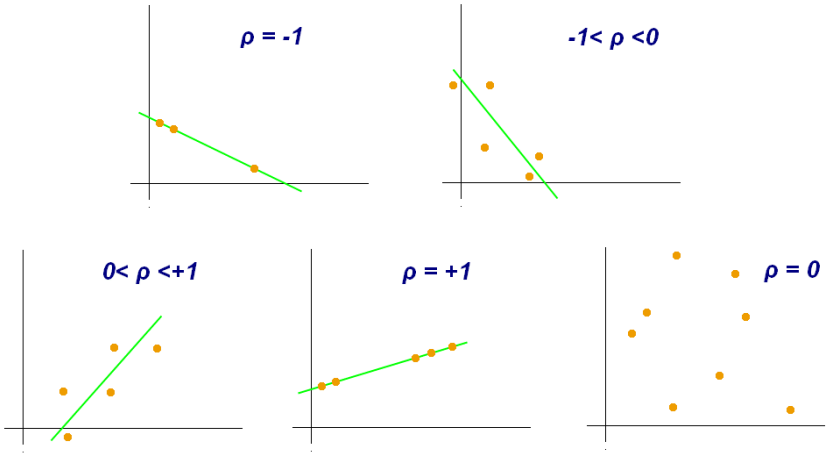
3. What is Pearson's R? (3 marks)
**(Answer)**

The Pearson correlation coefficient is a measure of linear correlation between two sets of data. It numerically signifies the strength and direction of linear relationship between two continuous variables.

It is the ratio between covariance between two variables and product of their standard deviations.

R = 1 (data is perfectly linear with positive slope)
R = -1 (data is perfectly linear with negative slope)
R = 0 (there is no linear association)
± 0.5 < R < ± 1 (Strong correlation)
± 0.3 < R < ± 0.49 (Medium correlation)
R < ± 0.29 (Weak correlation)

Scatter plots explaining the various correlation is given below.




*(Courtesy: Wikipedia)*

We could calculate Pearson (R) correlation coefficient using SciPy Library:

scipy.stats.pearsonr(…)

(https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pearsonr.html)

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

Lot of independent variables in a model would be on different scales. This will lead to a Machine learning model with unrealistic coefficients and difficult to interpret.

So, we employ feature scaling process for ease of interpretation and faster convergence (in gradient descent methods).

The Standardized scaling ensures variables are scaled in a way mean is zero and standard deviation is one.

The Normalized scaling (Min-Max scaling) ensures variables are scaled in a way where values lie between zero and one using max & min values of data.

The major differences are:

| Properties | Normalized scaling (Min-Max) | Standardized scaling |
|---|---|---|
| **Scaling** | Rescale to 0 to 1 | Not bounded |
| **Data used** | Min & Max values of features | Mean=0 and Std.Dev = 1 |
| **Outliers** | Sensitive to outliers | Less sensitive to outliers |
| **Modification** | Changes range of data | Changes shape of distribution of data |
| **Usage** | Image processing | Linear regression, K-means |
| **SkLearn library** | sklearn.preprocessing.**MinMaxScaler** | sklearn.preprocessing.**StandardScaler** |

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
**(Answer)**
Variance inflation factor (VIF) is a measure (estimate) of severity of multicollinearity in regression. VIF explains the relationship of an independent variable with all other independent variables.

Multicollinearity could have negative effects when building regression models of prediction and interpreting the results. This could also make it difficult to identify the contribution of significant variables.

The formula is:

VIF (i) = 1 / (1 – R-squared (i))

VIF > 10 is Very High and has to be investigated and corrected
VIF > 5 is High and has to be investigated and corrected

1 < VIF < 5 (moderately correlated)
VIF=1 (No multicollinearity)

When VIF is infinite (∞),
there is perfect multicollinearity in the regression model which suggests that
the independent variables are not suitable for prediction in the model. This could be corrected by removing variables and stabilize the model.

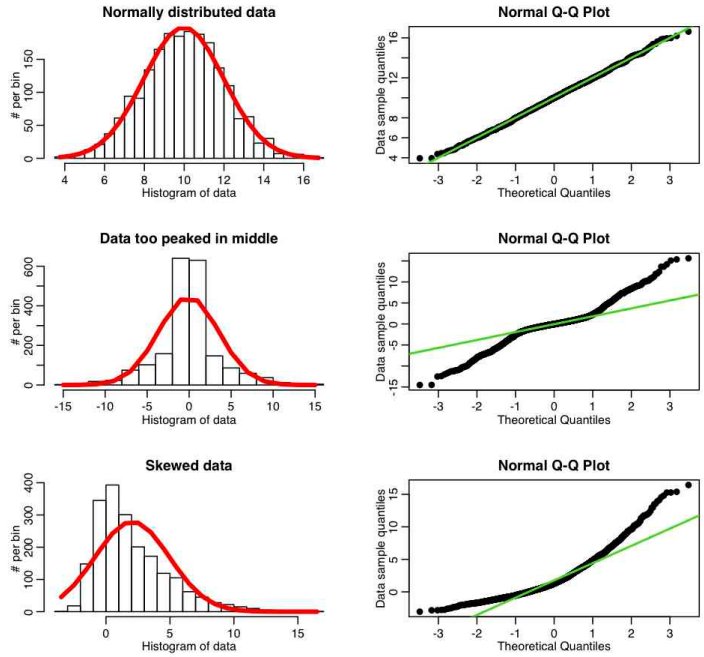
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
**(Answer)**

A Q-Q plot (quantile-quantile plot) is a graphical method for comparing two probability distributions by plotting their quantiles against each other.

Quantiles divide the range of a probability distribution into continuous intervals with equal probabilities.

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the straight-line y = x.
If the two distributions data has peaked in the middle or skewed, the Q-Q plot will have asymmetry.

In Linear regression, the assumption to validate is to check residuals follow a normal distribution.
The Q-Q plots are used to test one of these important Linear regression assumptions.
The Q-Q plots also identify skewness or tails and hints the data scientist to transform data or use other regression methods for abnormality.




*(Courtesy: sherrytowers.com)*