Linear Regression Assignment

Student: Jheser Guzman Illanes
Email: dicotips@gmail.com

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)**

- People prefers more Bike sharing on Working Days than in Holydays. This can conclude on commuting to the office riding a bike.
- People want to share the bike good weather conditions compared to cloudy, rainy or snowy.
- Regarding "**Season" vs "Count**" we conclude that the Bike Sharing is slightly higher in the Fall season than in Winter.
- Regarding **"Year" vs "Count"** we see that there are more Bike sharing from 2018 to 2019, which means that the business is rowing or it is getting more accepted by people.

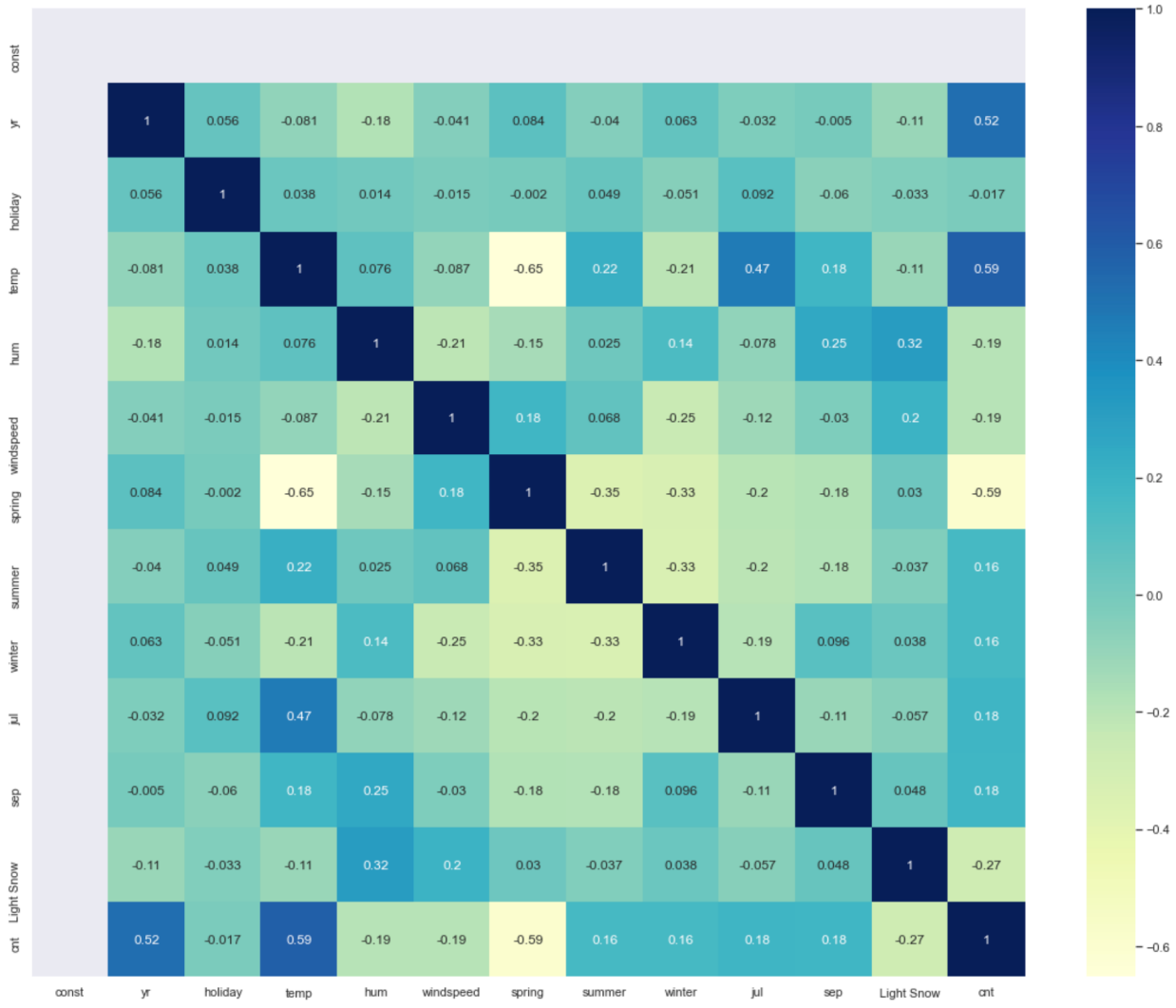
2. **Why is it important to use drop_first=True during dummy variable creation? (2 marks)**

The **get_dummies** function generates N dummy variables from N different values in a Categorical data. If we use drop_first=True, it will drop the first one due to the 00..0 tuple, which is not needed because it can be inferred by the other n-1 dummy variables.

Let's consider the following example with variable about months. The get_dummy function will generate 12 months, if we use drop_first=False. If we use, drop_first = True, the column of the first month, January, will be droped.

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

Based on the following HeatMap including the target variable:



We conclude that the **Temp**, and the **Year** are highly correlated with the target variable.

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

The validations for assumptions of Linear Regression that I took in the model building are:

- **Linear Relationship** between independent and dependent variables. I took only the variables which are highly correlated with the target variable using correlation values.

- **Homoscedasticity –** Which means that the residuals have constant variance. To verify homoscedasticity, I plotted the Residual plot which shows the variance of the error terms is constant across the values of the dependent variable.

- **Absence of Multicollinearity -** It refers to the fact that two or more independent variables are highly correlated, such as “Temp” and “aTemp”. To verify this, we took a look at the Variance Inflation Factors (VIF) and R-squared values to see if there is a high VIF and high correlation between variables.

- **Normality of Errors -** I plotted the QQ-plot to show normal distribution. This indicates unbiasedness of the residuals.

- **Independence of residuals (absence of auto-correlation):** I used *Durbin-Watson test* to verify this. We got 2.01 which indicating zero autocorrelation.

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

   Year, Temp, and a combination of **weather conditions** explain the demand of the shared bikes.

General Subjective Questions

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

Linear Regression is the Statistical technique to understanding the relationship between a dependent variable and its independent variables. We can categorize it in two types of models for Linear Regression:

1. *Simple Linear Regression (SLR):* If a single independent variable is used to predict the value of the dependent variable.
2. Multiple Linear Regression (MLR): If multiple independent variables are used to predict the value of the dependent variable.

Simple Linear Regression: This is the most basic form of regression model. Here we find the relationship between 1 dependent variable and 1 independent variable. This can be explained on the basis of simple linear equation:

$$Y = \beta_0 + \beta_1 X$$

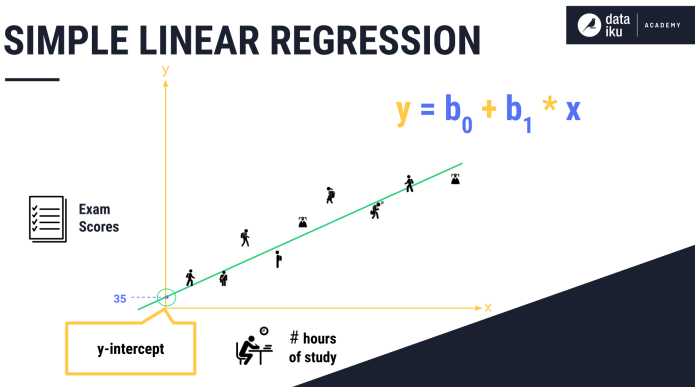
Here Y is the dependent variable, β_0 is the intercept, β_1 is the slope or correlation factor between Y and independent variable X.

The process to do SLR is:

- **Step 1:** Plot a scatter plot for the above variables.
- **Step 2:** Plot a best fit line shown by the above equation is drawn on the scatter plot between 2 variables. This line is known as Regression Line. There can be multiple regression lines in this plot. Then, we decide by calculating the **Residual Sum of Squares (RSS)**.

$$RSS = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$$

For any data point on the line, the difference between the actual data point of scatter plot and the line is the residual. RSS is determined by adding the squares of these residuals for every point in the scatter plot.



Source: https://knowledge.dataiku.com/latest/courses/intro-to-ml/regression/regression-summary.html

- **Step 3:** Determine the strength of the linear regression model by using R^2 or Coefficient of Determination:

$$R^2 = \frac{1 - RSS}{TSS}$$

TSS is Total Sum of Squares: It is the sum of errors of the data points from mean of response variables.

$$\text{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Multiple Linear Regression: Multiple Linear Regression is a statistical method to understand the relationship between 1 dependent and multiple independent variables. This can be explained on the basis of simple linear equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ldots . + \beta_p X_p + \epsilon$$

Where e is the error.

This model has new basic properties:

- Model now fits a “Hyperplane” instead of a line.
- Coefficients are still calculated using r-square.
- Assumptions of simple linear regression holds, that means that the **error term** have zero mean and are normally distributed.

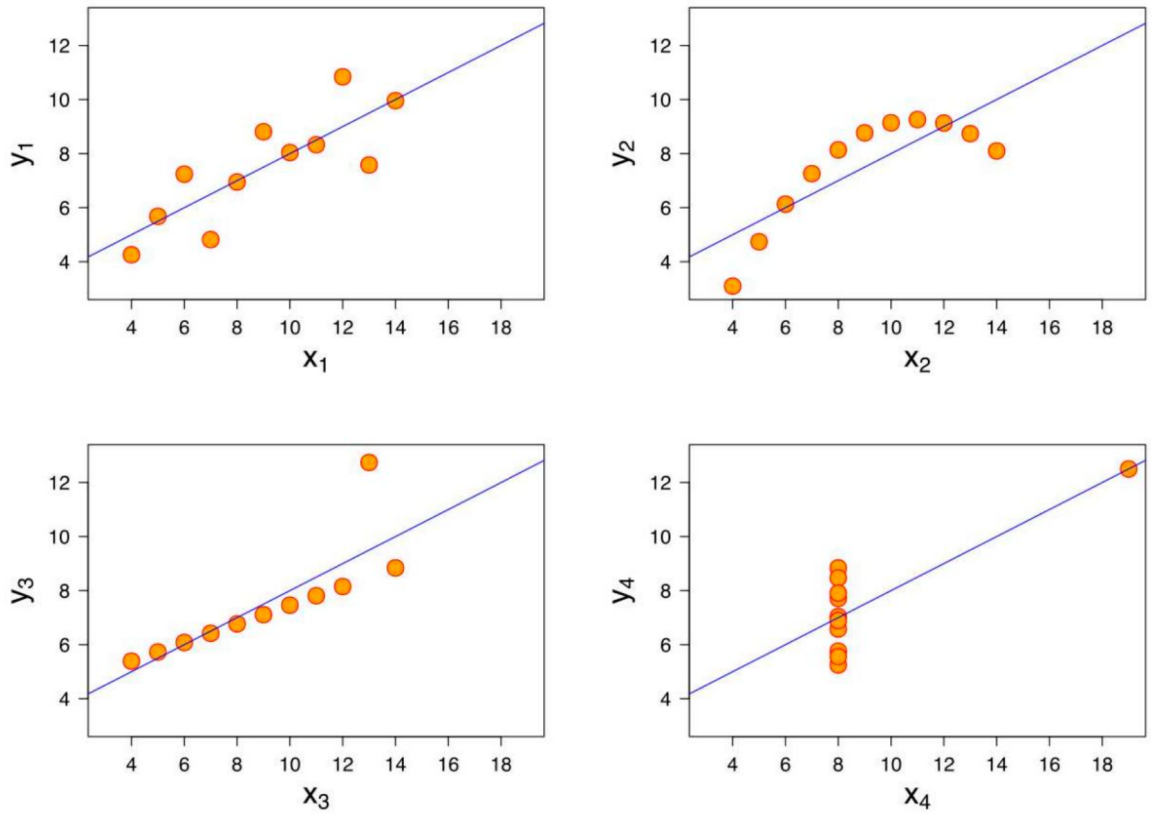
There are some new properties that we need to take care when using MLR:

- **Overfitting:** Adding more variables may cause overfitting. This will ed the model to memorize all the data points and will led to generalization. This will lead to decrease in accuracy.
- **Multicollinearity:** We must check that the independent variables should not be highly correlated. It affects the interpretation of data whether the change in Y when all other variables are constant apply or not. We can use heatmaps and pairplots to find the correlation between independent variables.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet can be defined by four datasets that have nearly identical statistical properties, but they look very different when graphed.

- Each dataset consists of eleven (x, y) data points.
- They were constructed in 1973 by the statistician Francis Anscombe to demonstrate the importance of graphing data before analysing it, and the effect of outliers on statistical properties.
- He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough." It has been rendered as an actual musical quartet.



The four datasets in above's graphs have the same mean in x and y. Have the same sum and variance. But they have a completely different distribution

1. Dataset 1 (top-left) is a linear model.
2. Data set 2 (top-right) is not normally distributed.
3. The distribution of the data set 3 is linear (Bottom-Left), but it has an outlier.
4. The dataset 4 (Bottom-Right) shows that one outlier is enough to change the mean and variance, and correlation and coefficient.

3. What is Pearson's R?

Pearson R coefficient is a measure of linear regression between two variables. It can take values between -1 to 1.

a. ***Positive Correlation:*** when Variable1 is directly proportional to Variable2. If V1 increases, V2 will increase too.
b. **Negative Correlation:** when Variable 1 is inversely proportional to Variable 2. It means if V1 increases, V2's value decreases.
c. **No Correlation:** When the value is close to 0. It means that the change in V1 will not have any effect on the change of V2.

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

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

In the case of a dataset with multiple *numeric variables*, they have different ranges and scales of magnitude. For quick and easy analysis and easy interpretation of the model, it is recommended to transform the values to a uniform scale range. We do scale for the following reasons:

1. Ease of interpretation
2. Faster convergence for Gradient Descent Methods (iterative).

Scaling does not affect the correlations, and it will only affect the coefficients of the model and none of the other parameters like t-statistic, F statistic, p-values, R-square, etc.

This can be done by 2 methods:

- **Standardizing:** The variables are scaled so that the mean is zero and standard deviation 1. It is useful when the variable has a central tendency distribution.

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

- **Max/Min Scaling:** The variable has constant distribution, and the transformation changes the values in range 0 to 1.

$$x = \frac{x - \min(x)}{max(x) - \min(x)}$$

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

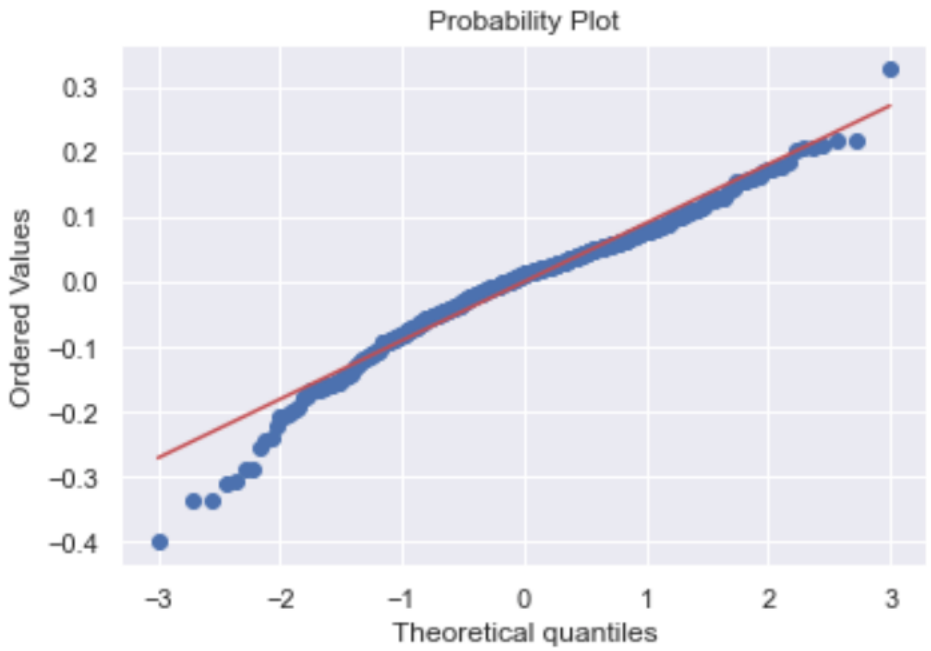
If VIF is infinite, it means that R^2 is close to 1. This means that there is close to perfect correlation (positive or negative).

$$VIF_i = \frac{1}{1 - R_i^2}$$

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

The Q-Q Plot or quantile-quantile plot is a scattered plot created by plotting 2 sets of quantiles. The chart determines if the data can be approximated by a statistical distribution such as Normal, exponential or a Uniform distribution.

If both the sets of quantiles which are plotted came from same distribution, then we will get a line which is roughly straight.



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 $y = x$.