# ML Assignment - Subjective Questions

Ramya D

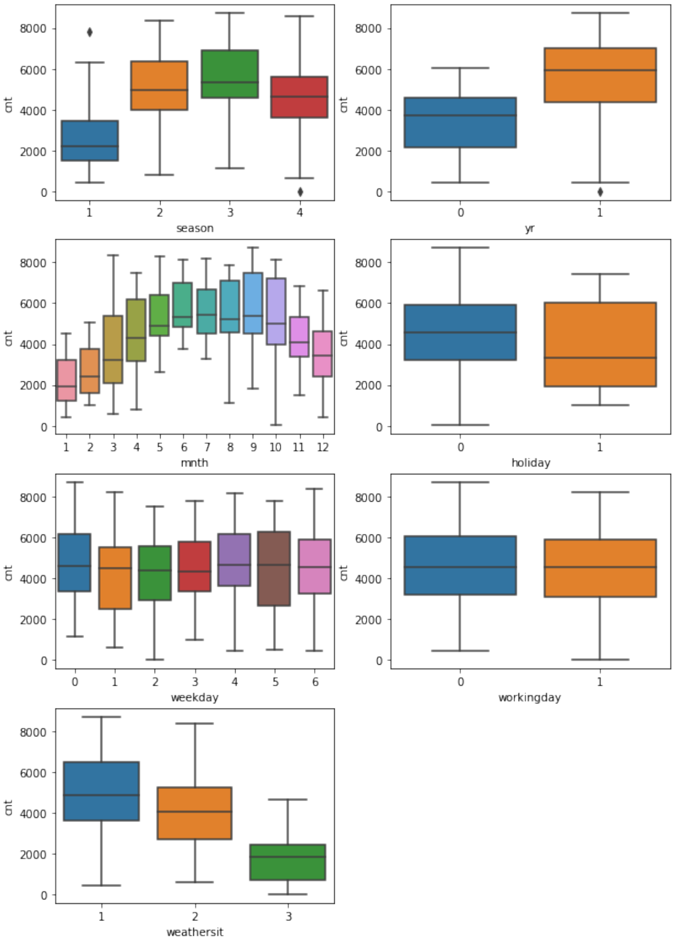
May 9, 2021

## 1 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)

   **A:**

Figure 1: Analysis of Categorical Variables



The categorical columns in the dataset are:

- `season` - Season (1:spring, 2:summer, 3:fall, 4:winter)
- `yr` - Year (0: 2018, 1:2019)
- `mnth` - Month ( 1 to 12)
- `holiday` - Indicator variable denoting whether a particular day is holiday or not
- `weekday` - Indicator variable denoting whether a particular day is weekday or not
- `workingday` - Indicator variable denoting whether a particular day is working day or not
- `weathersit` - Weather Situation for a particular day
  - 1: Clear, Few clouds, Partly cloudy, Partly cloudy
  - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
  - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
  - 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

We created a series of boxplots for these categorical columns with the dependent variable `cnt` to look at their dependencies. From the box plots above, we can infer that:

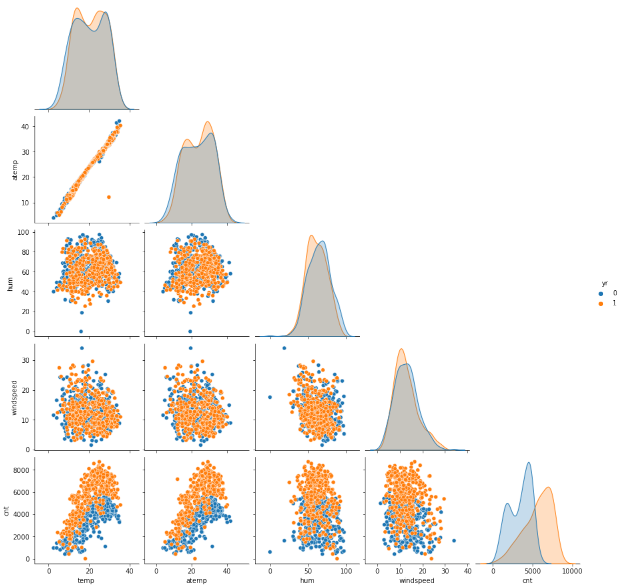
- In general we see a high demand for bikes in the `fall` season and this is the lowest for the season `spring`.The seasons `summer` and `winter` has almost has a similar distributions.
- We saw a higher demand in 2019 compared to the year 2018.Also, from the our final model and correlation analysis we observed that `yr` is the most significant variable in the data set which can explain the dependent variable `cnt`
- In any given calender year we have seen that the demand gradually increases from `January` to `August` and then decreases from there till `December`.
- As expected the demand is low on holidays as compared to nonholidays.
- We see a slight increase in demand on saturday and sunday.
- The categorical variable `Working day` has almost no influence on `cnt`.
- Usually when the weather is Clear or cloudy the demand is highest and when it is snowing or light raining the demand is lowest.

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

   **A:** The parameter `drop_first` determines whether to get $k-1$ dummies out of $k$ categorical levels by removing the first level.This needs to be set as `True` because it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.Also, we can explain all the $k$ categorical levels with just $k-1$ dummy variables without loosing any information about dropped level.

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

   **A:**

   From the last row in the above pair plot, we can see that the dependant variabloe `cnt` has a strong dependancy on the variables `atemp`,`temp`. This can also be confirmed from the table of correlations between the numerical columns and dependant variable below.

   | **Variable** | **Correlation with cnt** |
   |---|---|
   | `atemp` | 0.631 |
   | `temp` | 0.627 |
   | `hum` | -0.099 |
   | `windspeed` | -0.235 |

   Table 1: Correlations with the dependant variable cnt

   Thus from the above table 1, we can confirm that the numeric variable `atemp` has a highest correlation of 0.631 with the dependant variable `cnt`.

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

   **A:**

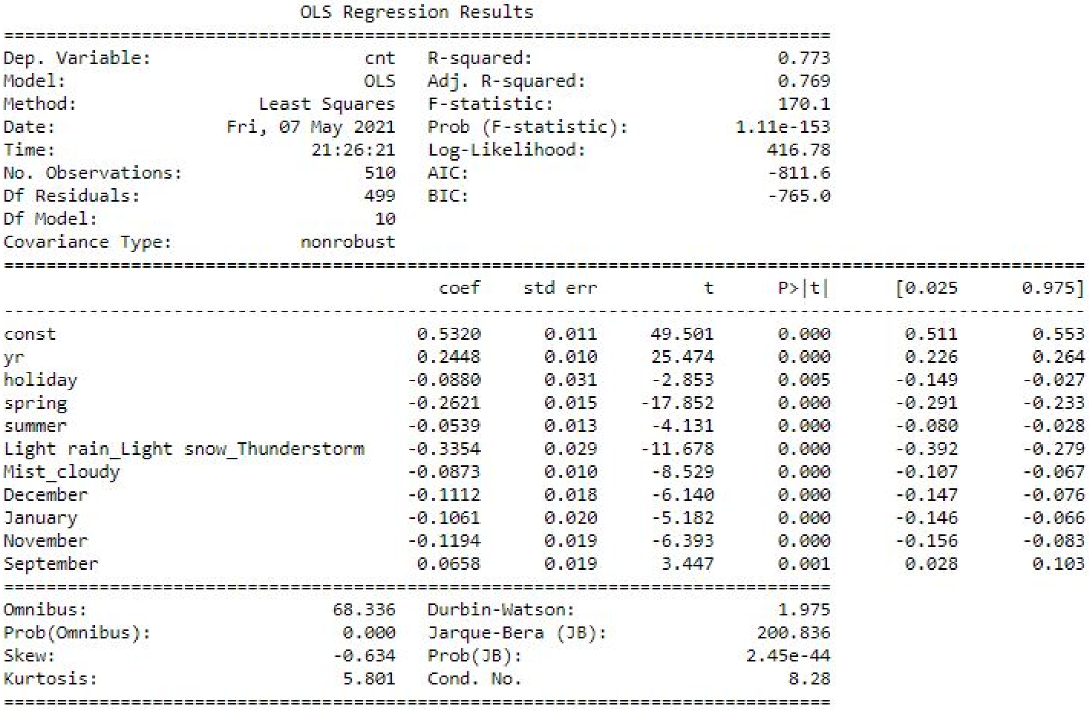
   - Residuals follow a normal distribution: In order to verify assumptions we looked at the normal QQ-plot of residuals and observed that it closely follows the normal distribution except for a slight light tail.
   - Residuals has a mean = 0: Though we did not perform any statistical z-test but from the histogram and kernel density function we plotted it is clearly evident that the mean of the residuals is a very small number which is close to 0.

- No multicollinearity: There are different metrics to measure amount of multicollinearity among the group of independent variables such as `VIF`,`correlations`,`tolerance` In our model we chose to restrict ourselves just to bi variate correlations and `VIF` . We followed a strict criteria of `VIF`$<2$ for all the independent variables.
- Homoscedasticity: We did a scatter plot of $y_{pred}$ and $y_{actual}$ to check if there are any patterns.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

   **A:**

Figure 2: Summary of our Final Model



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.773
Model:                            OLS   Adj. R-squared:                  0.769
Method:                 Least Squares   F-statistic:                     170.1
Date:                Fri, 07 May 2021   Prob (F-statistic):          1.11e-153
Time:                        21:26:21   Log-Likelihood:                 416.78
No. Observations:                 510   AIC:                            -811.6
Df Residuals:                     499   BIC:                            -765.0
Df Model:                          10
Covariance Type:            nonrobust
======================================================================================================
                                         coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------------
const                                  0.5320      0.011     49.501      0.000       0.511       0.553
yr                                     0.2448      0.010     25.474      0.000       0.226       0.264
holiday                               -0.0880      0.031     -2.853      0.005      -0.149      -0.027
spring                                -0.2621      0.015    -17.852      0.000      -0.291      -0.233
summer                                -0.0539      0.013     -4.131      0.000      -0.080      -0.028
Light rain_Light snow_Thunderstorm    -0.3354      0.029    -11.678      0.000      -0.392      -0.279
Mist_cloudy                           -0.0873      0.010     -8.529      0.000      -0.107      -0.067
December                              -0.1112      0.018     -6.140      0.000      -0.147      -0.076
January                               -0.1061      0.020     -5.182      0.000      -0.146      -0.066
November                              -0.1194      0.019     -6.393      0.000      -0.156      -0.083
September                              0.0658      0.019      3.447      0.001       0.028       0.103
==============================================================================
Omnibus:                       68.336   Durbin-Watson:                   1.975
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              200.836
Skew:                          -0.634   Prob(JB):                     2.45e-44
Kurtosis:                       5.801   Cond. No.                         8.28
==============================================================================
```

As we can observe from our final model summary above, apart from the variable `const`, the variables `yr`,`spring` and `Light rain_light_snow_Thunderstrom` explain the dependent variable `cnt` most with an absolute tstatastic values of 25.474, 17.852 and 11.678 respectively.

## 2 General Subjective Questions

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

   **A:**

   Often, in practice, one is called upon to solve problems involving sets of variables when it is known that there exists some inherent relationship among the variables. The concept of regression analysis deals with finding the best relationship between $Y$ and independent variables $x_i$, quantifying the strength of that relationship, and using methods that allow for prediction of the response values given values of the regressors.The complexity of most problems is such that in order to be able to predict an important response, a multiple regression model is needed. When this model is linear in the coefficients, it is called a linear regression model. For the case of $k$ independent variables $x_1, x_2, \cdots, x_k$ , the mean of $Y|(x_1, x_2, \cdots, x_k)$ is given by the multiple linear regression model

   $$\mu_Y|x_1, x_2, \cdots, x_k = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k,$$

   and the estimated response is obtained from the sample regression equation

   $$\hat{y} = b_0 + b_1 x_1 + \ + b_k x_k,$$

   where each regression coefficient $\beta_i$ is estimated by $b_i$ from the sample data using the method of least squares.

   We obtain the least squares estimators of the parameters $\beta_0, \beta_1, \cdots, \beta_k$ by fitting the above multiple linear regression model to the data points

   $$\left\{(x_{1i}, x_{2i}, \cdots, x_{ki}, y_i); i = 1, 2, \cdots, n; n > k\right\}$$

   ,

   where $y_i$ is the observed response to the values $x_{1i}, x_{2i}, \cdots, x_{ki}$ of the $k$ independent variables $x_1, x_2, \cdots, x_k$ . Each observation $(x_{1i}, x_{2i}, \cdots, x_{ki}, y_i)$ is assumed to satisfy the following equation:

   $$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}$$

   or

   $$y_i = \hat{y}_i + e_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki} + e_i$$

   , where $i$ and $e_i$ are the random error and residual, respectively, associated with the response $y_i$ and fitted value $\hat{y}_i$ .

   It is assumed that the $e_i$ are independent and identically distributed with mean 0 and common variance $\sigma^2$ . In using the concept of least squares to arrive at estimates $b_0, b_1, \cdots, b_k$, we minimize the expression

   $$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{1i} - b_2 x_{2i} - \cdots - b_k x_{ki})^2$$

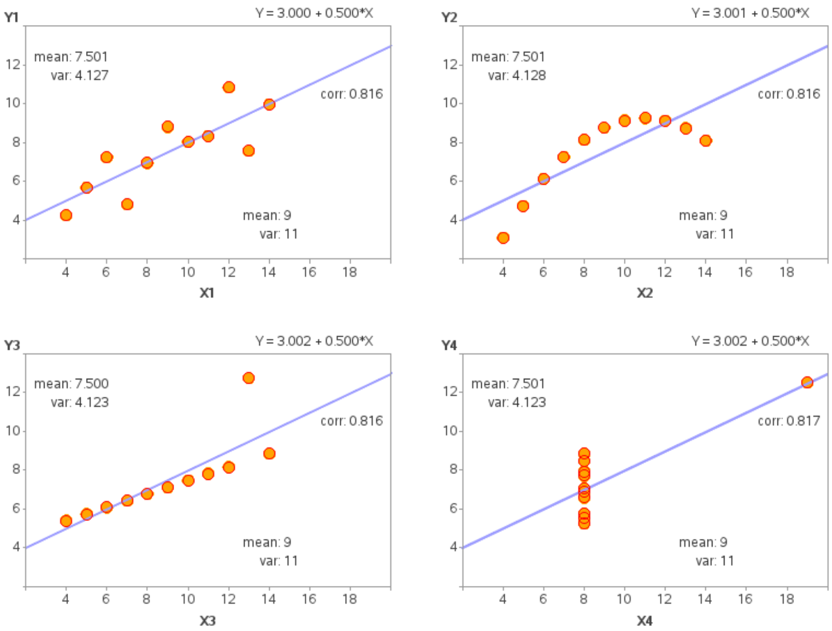
Differentiating $SSE$ in turn with respect to $b_0, b_1, \cdots, b_k$, and equating to zero, we generate the set of $k+1$ normal equations for multiple linear regression which needs to be solved to estimate the $b_i$'s.

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

   **A:**

   Anscombe's quartet is a collection of four $x$ - $y$ datasets which have the same traditional statistical properties (mean, variance, correlation, regression line, etc.), yet are quite different. Each dataset consists of eleven $(x, y)$ points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers and other influential observations 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.

   Figure 3: Anscombe's Quartet



   Based on the figure from Dataset 1, the linear regression model seems to fit the data quite closely. However, for the figure from Dataset 2, the data seems to be of a curvilinear nature, possibly quadratic and the linear model fitting is inappropriate. Similarly, the linear model on the figure based on Dataset 3 is also erroneous; only 1 data point passes through the fitted line and one point is far away from the regression fitted line. For the figure from Dataset 4, one point is a clear outlier, while all the other points are clustered at the same x value.

   Hence, one should check the validity of the data. Additionally, if the data is accurate, then the linear model fit should be reported as-is. However, one should mention that one of the data points have played a critical role in the linear regression model fitting of the data. Anscombe's quartet provides a quick way to the idea that sometimes the visual dimension can reveal a story that simple numerical analysis appears to deny.

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

   **A:**

   Pearson r correlation is the most widely used correlation statistic to measure the degree of the relationship between linearly related variables. 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.

   $$r_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

   As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationship. The absolute values of both the sample and population Pearson correlation coefficients are on or between 0 and 1. Correlations equal to $+1$ or $-1$ correspond to data points lying exactly on a line (in the case of the sample correlation), or to a bivariate distribution entirely supported on a line (in the case of the population correlation). A value of $-1$ implies that all data points lie on a line for which $Y$ decreases as $X$ increases. A value of 0 implies that there is no linear correlation between the variables.

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

   **A:**

   When you have a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret. So we need to scale features because of two reasons:

   (a) Ease of interpretation

   (b) Faster convergence for gradient descent methods

   You can scale the features using two very popular method:

   (a) Standardizing: The variables are scaled in such a way that their mean is zero and standard deviation is one.

   $$x_{scaled} = \frac{x - x_{mean}}{x_{std}}$$

   (b) MinMax Scaling: The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

   $$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

   .

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

   **A:** 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. The formulation of VIF is given below:

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

   Now,

   $$\begin{aligned} VIF = \infty \Rightarrow \frac{1}{1 - R_i^2} &= \infty \\ \Rightarrow 1 - R_i^2 &= 0 \\ \Rightarrow R_i &= \pm 1 \end{aligned}$$

   An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables ($R_i = \pm 1$) . To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

   **A:**

   The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

   A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value.

   A 45 degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

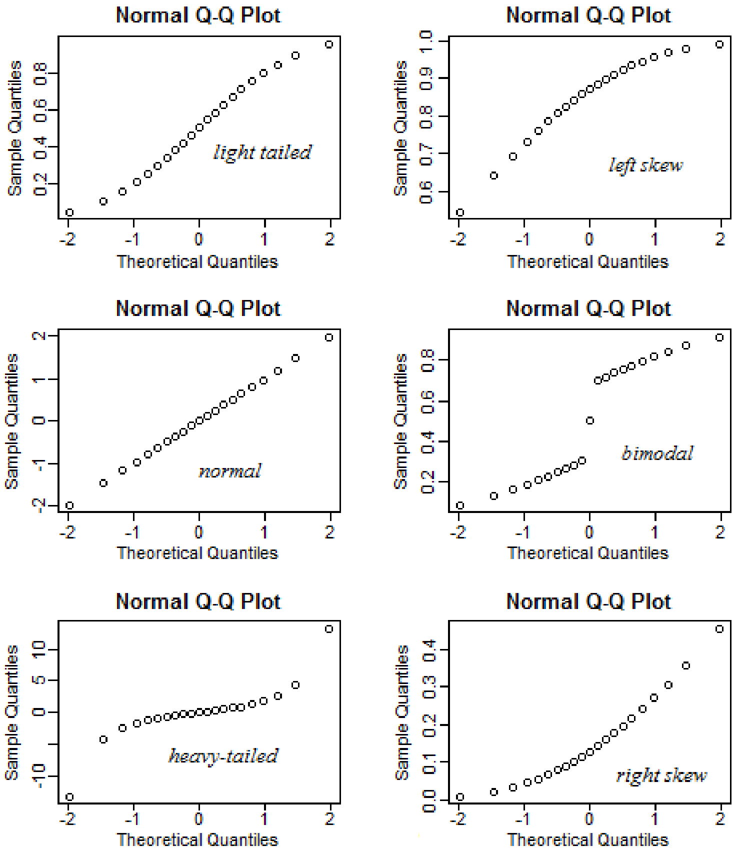
   The advantages of the q-q plot are:

   (a) The sample sizes do not need to be equal.

   (b) Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot. For example, if the two data sets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45 degree reference line.

The q-q plot is similar to a probability plot. For a probability plot, the quantiles for one of the data samples are replaced with the quantiles of a theoretical distribution.Please find the interpretation for a normal q-q plot below.

Figure 4: Normal q-q Plot Interpretation


Normal Q-Q Plot
Sample Quantiles
light tailed
Theoretical Quantiles
Normal Q-Q Plot
Sample Quantiles
left skew
Theoretical Quantiles
Normal Q-Q Plot
Sample Quantiles
normal
Theoretical Quantiles
Normal Q-Q Plot
Sample Quantiles
bimodal
Theoretical Quantiles
Normal Q-Q Plot
Sample Quantiles
heavy-tailed
Theoretical Quantiles
Normal Q-Q Plot
Sample Quantiles
right skew
Theoretical Quantiles