Assignment-based Subjective Questions

Submitted by Sreekumar R

(1) From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

A: There are five categorical variables worth considering:

- The weather(mild, strong, medium etc),
- The season(winter, summer, fall, spring)
- Whether the day is a holiday or not
- Whether the day is a working day
- What day of the week is it
- The real feel temperature

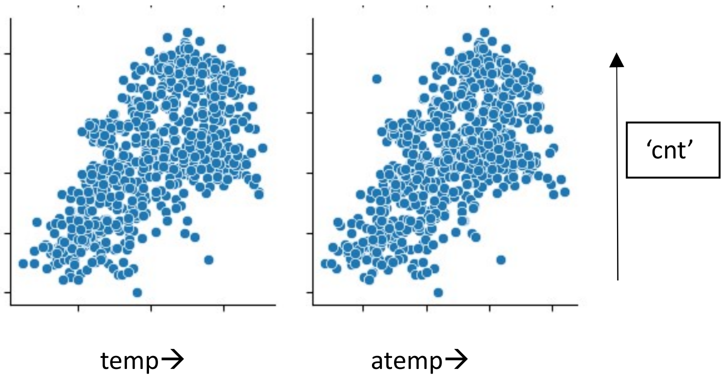
The dependent variable here 'cnt' i.e the total number bike sharing users per day prima facie seems to be having dependency on the above five categorical variables. Obviously if the weather is strong with rain, strong winds, snow etc, users will be abstaining from using the bike. In a similar manner season would also affect usage numbers. Likewise, it is logically expected that holiday and working day would have an influence in the manner that working day would see more people on the roads to offices, work locations, colleges etc so also the day of the week. However these are all guesses which can be confirmed only after predictive analysis using multiple linear regression.

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

A: Categorical variables cannot be used directly for modeling since the values are non-numeric. For converting categorical variables into numerical dummy variables, we use the Python function **pd.get_dummies().** The parameter **drop_first=True** is to be passed while calling the function. The reasons being, technically only (N-1) dummy variables are required to delineate a categorical variable with 'N' different values. It is always better to keep the model simple and with least number of predictor variables for model efficiency, easier to understand and interpret since the comparison would be with the base state where all values of dummy variables are 'zeroes'. Therefore parameter **drop_first=True** drops the first column of the dummy variables after converting the categorical variables into dummy variables.

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

A:



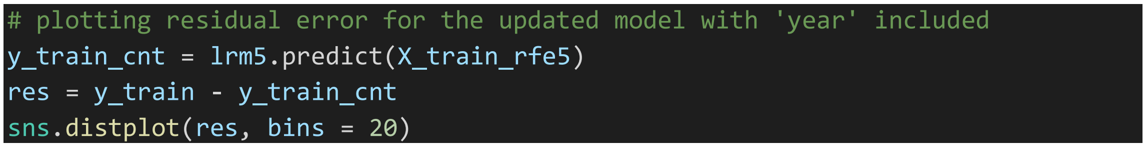
Looking at the pair plot, it is seen that two variables ‘temp’ the temperature of the day in Celsius and ‘atemp’ the feeling temperature are having the highest correlation with the target variable ‘cnt’ total number of users in a day.

Out of these two, based on visual observation both seem to have the same correlation. It is difficult to predict which has higher correlation visually since both seem to have the same pattern.

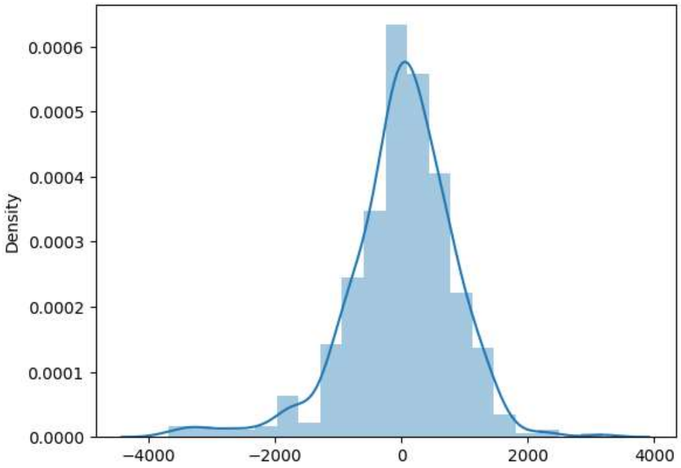
(4) How did you validate the assumptions of Linear Regression after building the model on the training set?

A: The following assumptions were validated.

a) Error terms are normally distributed with mean zero. Residuals on the training set data were calculated and plotted for the final chosen model. It can be seen that the distribution of residuals is normal and centred around mean zero. This is evident from the following plot. Residuals on the training data are calculated using the following code and plotted using sns.distplot() function.

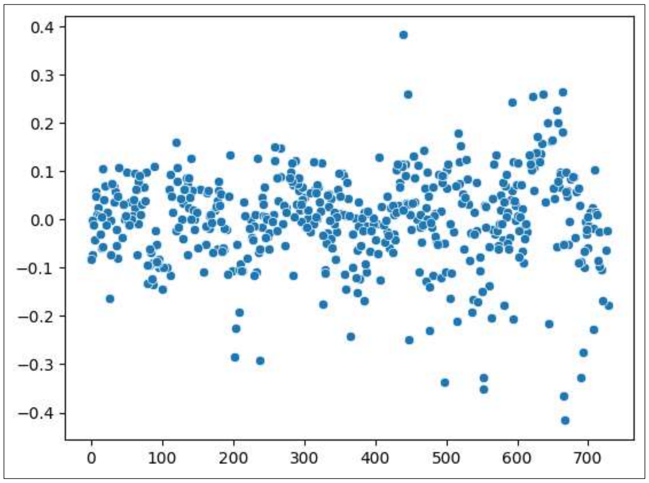


```
# plotting residual error for the updated model with 'year' included
y_train_cnt = lrm5.predict(X_train_rfe5)
res = y_train - y_train_cnt
sns.distplot(res, bins = 20)
```



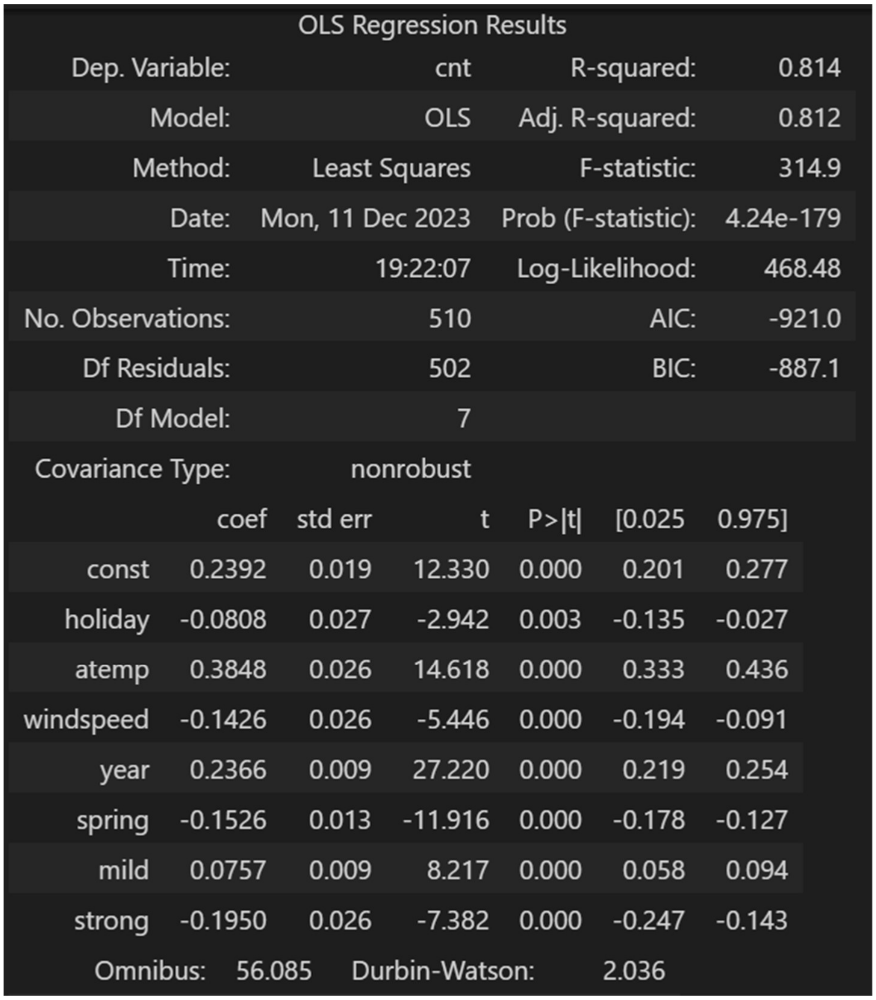
b) Error terms are independent of each other ie no visible patterns are observed. This is again visible from the plots.

c) Error terms on the test set have constant variance. This is evident from the plot given below.



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

A: According to the chosen model, atemp(real feel temperature), year, both these have a positive correlation, and strong weather(having negative correlation). Real feel temperature with a positive correlation and has the strongest influence. This is evident from the summary statistics of chosen model lrm5 given below.



```
                    OLS Regression Results
     Dep. Variable:              cnt        R-squared:      0.814
             Model:              OLS    Adj. R-squared:     0.812
            Method:    Least Squares       F-statistic:     314.9
              Date: Mon, 11 Dec 2023 Prob (F-statistic): 4.24e-179
              Time:         19:22:07    Log-Likelihood:    468.48
 No. Observations:              510               AIC:     -921.0
     Df Residuals:              502               BIC:     -887.1
          Df Model:                7
  Covariance Type:         nonrobust
              coef  std err        t  P>|t|  [0.025  0.975]
     const  0.2392    0.019   12.330  0.000   0.201   0.277
   holiday -0.0808    0.027   -2.942  0.003  -0.135  -0.027
     atemp  0.3848    0.026   14.618  0.000   0.333   0.436
 windspeed -0.1426    0.026   -5.446  0.000  -0.194  -0.091
      year  0.2366    0.009   27.220  0.000   0.219   0.254
    spring -0.1526    0.013  -11.916  0.000  -0.178  -0.127
      mild  0.0757    0.009    8.217  0.000   0.058   0.094
    strong -0.1950    0.026   -7.382  0.000  -0.247  -0.143
     Omnibus:  56.085   Durbin-Watson:    2.036
```

General Subjective Questions

(1) Explain the linear regression algorithm in detail.

A: Linear regression is a machine learning algorithm used to predict values in the case of a continuous variable. It is a supervised learning algorithm learning from past data, here termed as the training data. The linear regression(LR) model shows the linear relationship between the variables(between the predictor variables and the target variables).

A straight line(best fit line) is fitted between the independent(predictor) variables and the dependent(target) variables. The equation of the straight line y=mx+c shows the simple linear regression model, where m is the slope of the best fit line and c is the y intercept at x=0. If the value of m is high it shows the level of influence of the variable x. ie how much the value of y increases for a unit increase in the value of x.

The best fit line mentioned above is arrived at from the residuals. Residual is the difference between the actual value and the predicted value using the LR model. The algorithm used is minimizing the sum of squares of the residual errors (RSS: residual sum of squares). The best fit line is arrived at such that the RSS is minimal. RSS is called the cost function in the LR method. Minimizing the cost function is the prime aim of the LR models.

Cost function can be minimized using the differentiation method or the gradient descent method. Due to mathematical resources required to do a differentiation, generally gradient descent method is used for minimizing the cost function and arrive at an optimum model. Gradient descent function iterates over the m and c values to so that the cost function(RSS) is minimum.

RSS is an absolute quantity varying with units of the predictor variables. Therefor TSS(Total sum of squares) is used. Sum of the squares of the differences between actual y value and the mean of predicted y values.

Strength of the LR model is represented by R2 value which is (1- RSS/TSS). A high value generally indicates a good model. If the p-value of the variable is quite high, the model is not optimal. Generally p-value should be less than 5%. R2 value signifies the explainability like how well the variables explain the target variable.

The following are the assumptions in linear regression.

(1) Linear dependency between predictor and target variables
(2) Error terms are normally distributed with mean zero
(3) Error terms are independent of each other i.e no visible pattern of errors
(4) Error terms are having constant variance (Homoscedastic).

In the hypothesis testing, the null hypothesis is m = 0(m is not significant) and alternate hypothesis is m not equal to zero where m is the slope the regression line. Calculate the p-value based on the t-score and if p-value is less than 5%, we can conclude that null hypothesis is rejected.

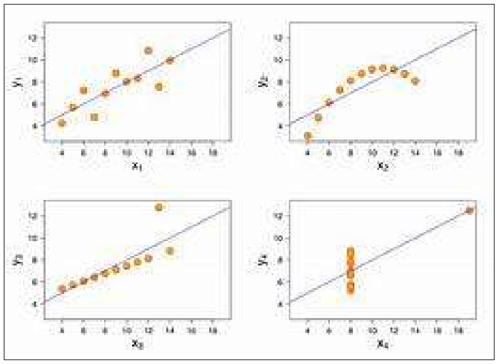
In multiple linear regression (MLR) models, there are number of independent variables and one predictor variable. In this case feature (predictor variables) selection is very important in this case, to identify the influential variables. MLR is required when one variable might not be sufficient to create a good model and make accurate predictions. Here the model fits a ‘hyperplane’ instead of a line. In MLR, overfitting, multicollinearity, feature selection are very important in deciding the best model.

(2) Explain the Anscombe’s quartet in detail.

A: Anscombe’s quartet is a popularly used example to show the significance of visual representation of data as charts, plots etc. Though the statistically derived values are useful for data analysis and understanding, these should be supplemented with analysis of visual representation of data to get the right inferences. Anscombe’s quartet has four sets of data points which all have the following statistical values as identical though the data points have four different patterns when observed visually in a plot.

1) Mean of X
2) Mean of Y
3) Variance of X

4) Variance of Y
5) Correlation between X and Y
6) Linear regression equation.



These data sets were arrived at by the statistician Francis Anscombe with an intention to signify the importance of visual representation of data sets. As it can be seen from the above picture, all four data sets have different pattens when viewed as plots, though the statistical values are all same which can be quite misleading. Therefore the essence of the quartet is that always supplement statistical values with visual representations for proper understanding of data and inferences.

(3) What is Pearson's R?

A: Pearson's R is also known as correlation coefficient. It is a measure of relationship between two variables linearly related. The value ranges between -1 and 1.

-1 means perfect negative relationship means when one variable increases, the other decreases the other too increases proportionately.

+1 means perfect positive relationship means when one variable increases, the other too increases proportionately.

Pearson's R value of zero means the two variables are not related and independent of each other.

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

A: Scaling is performed on data points of variables to bring the data point values to comparable level. Scaling is performed for ease of interpretation of the model. The coefficients will be comparable and influence of each can be determined easily looking at the LR model equation. Scaling also helps in faster convergence during gradient descent method since the numerical values are lower and in the same range. Scaling is normally done after the train-test split. Same transformation is applied to the test variables also during testing after building the model.
Scaling won't affect the metrics like r2, prob(F-statistic), p-values of variables and model accuracy. However coefficients will surely change in a scaled LR model compared to unscaled LR model. Scaling wont change the shape or distribution of the original variables.

There are two types of scaling- minmax scaling(called normalization) and standardization scaling.

Min Max scaling reduces the data point values to between 0 and 1(if there are only positive values). This takes care of outliers also since the range is compressed between zero and one. Equation is (X-Xmin)/(Xmax-Xmin)

Standardization scaling reduces the values of datapoints such that the mean is zero and the standard deviation is 1. This method doesn't compress the data values to a particular range. Equation is (X-Xmean)/Std dev of X

Dummy variables are generally not scaled since the values are already zeroes and ones.

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

A: Infinite VIF value means the variable is having a very strong correlation with the rest of the variables and is described fully by the other variables. This means the variable with infinite VIF value is redundant in the presence of other variables.
Equation of VIF is 1/(1-r^2). Here is the correlation coefficient. If the variables are strongly related, the value of r would be '1' and therefore the denominator would become zero. In such cases the VIF value would become infinite. It definitely corresponds to very high correlation with variables and certainly that variable with infinite VIF value can be dropped.

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

A: Q-Q plot or the Quantile-Quantile plot is a statistical tool to check if data points follow normal distribution. In linear distribution it is assumed that the residual errors follow a normal distribution. This can be checked using the Q-Q plot. If the points lie along the diagonal line, the data points follow the normal distribution. It is a tool to validate the assumption of normal distribution of residual errors. There is a provision in statsmodels library to plot the Q-Q plot.
In the X-axis theoretical quantiles of normal distribution are plotted. In the Y-axis, sample quantiles are plotted. If the plot follows a diagonal line, then the points follow normal distribution.

****