Linear Regression Assignment

Student name: Alhad Parashtekar

Assignment-based Subjective Questions

Q1) From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

The target variable is the total number of customers, i.e., 'cnt'. In the dataset, apart from 'cnt', 'casual' and 'registered' columns also describe number of casual (non-registered) and number of registered customers. Here, 'cnt' is sum of 'casual' and 'registered' variables (cnt = casual + registered). Hence, these three are not independent variables. Therefore, the effect of categorical variables on three of them is analysed individually.

The bar charts in the notebook show that 'seasons' and 'months' show a significant impact on all three customer count variables. The customer count is high in summer and fall, i.e., from May to September. However, it must be noted that both data are same but with different resolutions. The month is a finer division of time than seasons. This fact is considered while model building.

All customer count variables are higher in 2019 when compared with 2018. Though the year will be considered as a variable in the model, a detailed analysis is recommended to find out the reason behind it.

The impact of the weather situation on all three variables shows the same trend for all three customer count variables. Customer count is highest when the weather is misty while lowest in case of rains. The notebook shows the bar charts.

As expected, number of the customers is higher on holidays when compared with working days.

The 'day of the week' variable does not show a significant effect on 'cnt'. However, it shows a significant impact on 'casual' and 'registered'. Both these variables show opposing trends with respect to 'day of week'. The casual customers are highest on weekends, while registered customer takes the majority share on weekdays. The reason behind this should be investigated.

Q2) Why is it important to use drop_first=True during dummy variable creation?

The get_dummies method in python crates one dummy variable for every possible response in the categorical variable. 'drop_first=True' drops one of these variables from the data frame.

However, one dummy variable in this set would be redundant as it would not provide any new information. That is, when the rest of the variables are zero, this would automatically be one. Therefore, consideration of all the variables, would result into carry one extra variable per each categorical variable. Consequently, it is advisable to drop one dummy variable, which is accomplished by drop_first=True.

Q3) Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Among numerical variables, 'temp' and 'atemp' have highest correlation, i.e., 0.63 with 'cnt'.

Q4) How did you validate the assumptions of Linear Regression after building the model on the training set?

One of the assumptions of linear regression is that errors are normally distributed with the mean as zero. This assumption is verified by plotting the histogram of the errors (y_observed – y_predicted). One can observe the mean and approximate distribution from the histogram. The mean should be near zero and the distribution should be normal.

The assumption of homoscedasticity can be verified from the scatter plot of observed data and (preferable) line plot of predicted data. The enveloping lines, which envelop the data on either side, should be approximately parallel to the fitted line.

The randomness in the error can be verified from the same scatter plot or by plotting errors as a function of y_observed. The independence of the errors will not show any pattern in the errors.

Q5) Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Temperature, weather conditions, and year have the highest impact on the demand for shared bikes. As explained earlier, the 'year' is not really a variable, therefore my third impactful variable would be either seasons or months.

General Subjective Questions

Q1) Explain linear regression algorithm

The linear relationship between target variables and the regressors can be expressed in a general form as follows.

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_n X_n$$

Here y is a target variable, *X*s are regressors, and βs are coefficients. The linear regression algorithm finds the optimum values of βs for given a set of inputs and output. The linear regression algorithms find βs by minimizing a pre-determined cost function. Generally, root-mean-squared error or absolute error is utilized as a cost function for linear regression.

In most cases, the Gradient Descent Algorithm is used to minimize the cost function. Different variations of gradient descent can be utilized such as stochastic gradient descent, mini-batch gradient descent, and learning rate scheduling. In addition to gradient descent, Newton's methods can also be utilized for cost function minimization.

The parameters obtained by minimizing the cost function may not always be statistically significant. The statistical significance of the coefficients is checked generally by p-values. The statistical significance of the overall fit is checked using parameters such as F-statistic, AIC, or BIC.

Q2) Explain the Anscombe's quartet in detail.

F. J. Anscombe from Yale University published an article titled 'Graphs in Statistical Analysis' in The American Statistician in 1973. The article discusses the necessity of graphs, mainly scatter plots, for faultless statistical analysis.

In the article, Anscombe published four scatter plots (see following figure) showing different trends in input and output variables. These figures are known as Anscombe's quartet. Though the trends seen in the data are different, the data sets result in linear regression models with the same statistical metrics. These statistical metrics include mean and variance of input and output variables, correlation between input and output, coefficients of linear regression, and R-squared values.

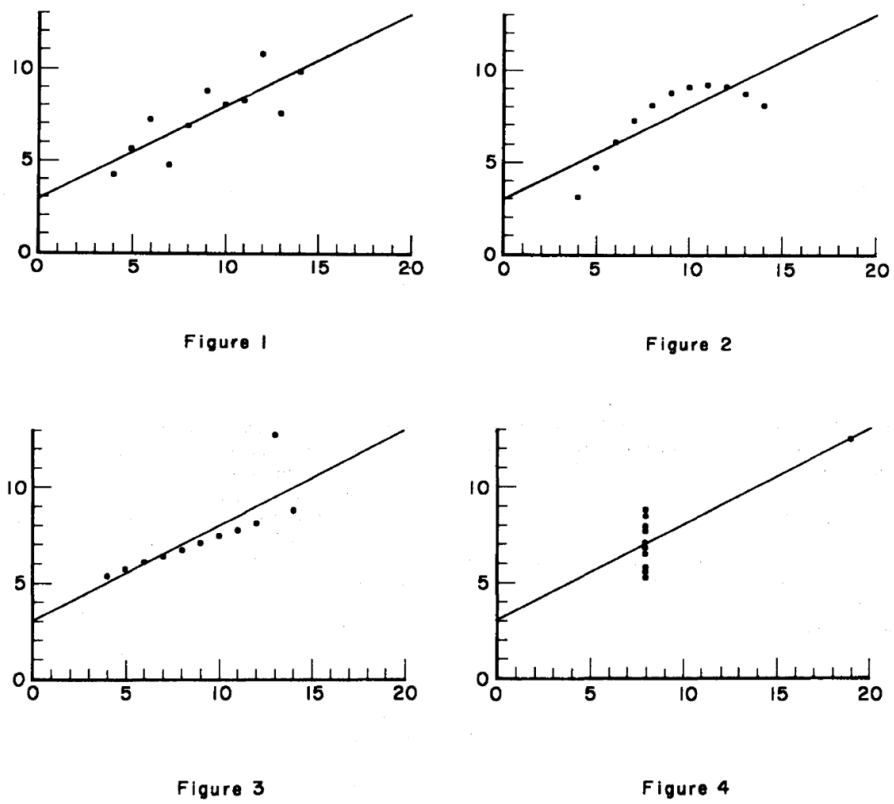

Fig. 1 The Anscombe's quartet. Figure 2 and 4 from the quartet clearly does not show linear relationship. Though, the data in Figure 1 and 3 can be modelled using linear relationship, the regression line should be different.

This analysis shows the importance of visualizing the data using simple techniques such as scatter plots. The article points to the fact that inferences drawn from a linear regression without prior visualization can lead to erroneous inferences.

Ref: Anscombe, F.J., 1973. Graphs in statistical analysis. The american statistician, 27(1), pp.17-21.

Q3) What is Pearson's R?

Pearson's R also known as Pearson's correlation coefficient quantifies the strength of the relationship between the two variables. The sign of the correlation, positive or negative, indicates the trend found in the relationship. The positive correlation indicates that the values of both variables would increase or decrease simultaneously, while the negative correlation indicates the value of one of the variables decreases while the other increases.

The magnitude of the correlation indicates the 'closeness' of the data points to the fitted lines. If all data-points perfectly coincides with the fitted line, the magnitude of correlation would be 1. The coincidence decreases as the magnitude reduces from 1 to 0.

The formula for Pearson's correlation can be given as

$$Correlation = \frac{Covariance(var1, var2)}{\sqrt{variance(var1)}\sqrt{variance(var2)}}$$

Here, var1 and var2 are variables 1 and 2, respectively. The covariance decides the sign of the correlation. The division by variances of both the variables makes the correlation independent of the scale of the variable.

Q4) What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

The different input variables, i.e., regressors, can have different units. This would result in larger variations in their absolute values. The optimization algorithms such as gradient descent would find it difficult to reach the optimum value when variables have large numeric differences.

The regression coefficients obtained from data having large differences in scale will also have similar scale differences. This would make it difficult to interpret which input variable has larger influence on the target variable. Therefore, scaling of the variables is carried out to map all the variables in a similar range of values.

Two scaling approaches are popular. The normalization strategy maps all the values on a scale of zero to one, with zero being minimum and one being maximum. The standardization strategy maps the data on a standard normal distribution.

Q5) You might have observed that sometimes the value of VIF is infinite. Why does this happen?

The formula for VIF is

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

When VIF is infinity, the R^2 is one. This suggests that one of the input variables for which VIF is calculated can be perfectly explained using all the remaining input variables. In other words, the inclusion of that variable in the model would result in the incorporation of redundant data. The inclusion of this data would result in an inflated value of overall regression giving a false impression about the linear regression.

Q6) What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression

The Q-Q plots are used to find whether a variable is distributed normally or not. The Y-axis, ordinate of the Q-Q plot contains quantile values (i.e., percentile) of all the data points for one variable in the dataset. The X-axis of the Q-Q plots contains values from a standard normal distribution corresponding to the same quantiles. The closer the Q-Q is to the straight line, the closer the variable's distribution is to the normal distribution. This scheme is useful to check whether the distribution of errors is normal and also whether the training and testing datasets are drawn from the same sample.