Linear Regression Subjective Questions

Student: Gert Agenbag

Assignment-based Subjective Questions

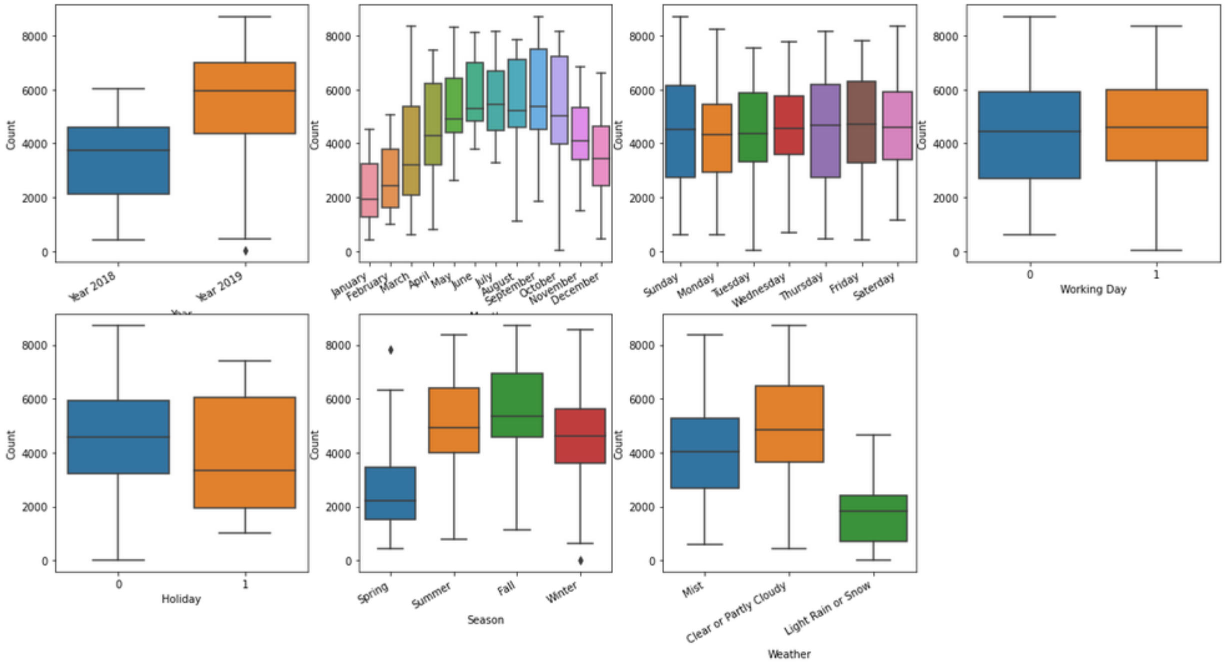
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Some categorical variables had little or no influence on the dependent variable. The categorical variable Working Day (‘workingday’) has a negligible effect on the dependent variable, Count (‘cnt’). In the boxplot, it appears that Holiday (‘holiday’) may have some small effect, but with only 21 records of Holidays in the entire dataset, it’s not possible to make an accurate estimate of its influence on the dependent variable. The Day of Week categorical variable (‘weekday’) also showed very little variability.

The Operational Years categorical variable (‘yr’) had a very strong influence on the dependent variable. The dataset only included two years of data, 2018 and 2019, so it is treated as a categorical variable. In essence, this variable captures the growth in bike rentals over time. If the dataset spanned over a longer period, it may made sense to treat is as a discrete quantitative variable, or even to substitute it with a continuous quantitative variable such as Instant (‘instant’). Another option may have been to derive the year as a continuous fractional number for each record. However, I found that using the categorical variable worked well in this model.

While the Season (‘season’) and Month (‘mnth’) categorical variables had a strong influence on the dependent variable, they are highly correlated with each other and with the Temperature variables. Using only a some of the seasons or some of the calendar months in the model would reduce the model’s interpretability. After trying to introduce the seasons into the model, I found that using Apparent Temperature (‘atemp’) instead gave almost identical results and opted to exclude the season variables from the model. Temperature is a good proxy for months and seasons as the temperature varies over the course of the year.

Finally, the Weather (‘weathersit’) categorical variables also had a strong influence on the dependent variable. Demand for bike rentals decrease when there is (1) Light Rain or Snow, (2) Wind, or (3) Mist. It may be reasonable to assume that demand will also decrease during Heavy Rain or Snow, but the supplied dataset did not include any records for such weather conditions.



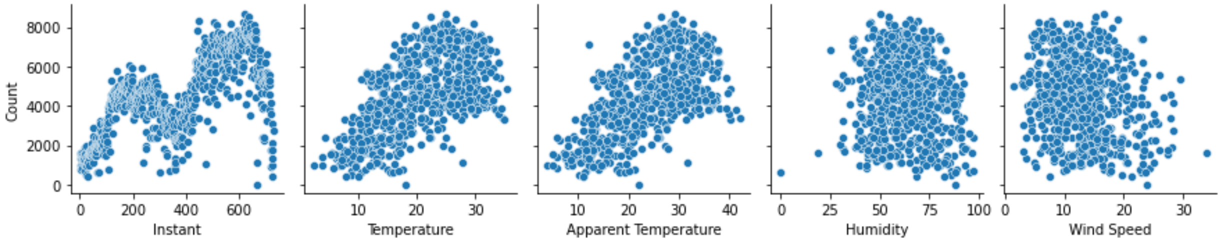
2. Why is it important to use drop_first=True during dummy variable creation?

When dummy variables are created, there is duplication of information. This results in multicollinearity between the dummy variables, which affects the performance of the model. By dropping one of the dummy variables, one can eliminate the multicollinearity that was introduced.

As a matter of convenience, the first dummy variable is usually dropped. However, this can affect the interpretability of the model. If interpretability is important, it is better to identify the baseline dummy variable from which other dummy variables deviate. If the baseline dummy variable is dropped, it's easier to understand how the deviating dummy variables relate to the target variable.

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

Apparent Temperature ('atemp') has the highest correlation with the target variable, Count ('cnt'), at 0.630. The closely related variable Temperature ('temp') has a correlation of 0.627 with the target variable, but a correlation of 0.991 with Apparent Temperature.

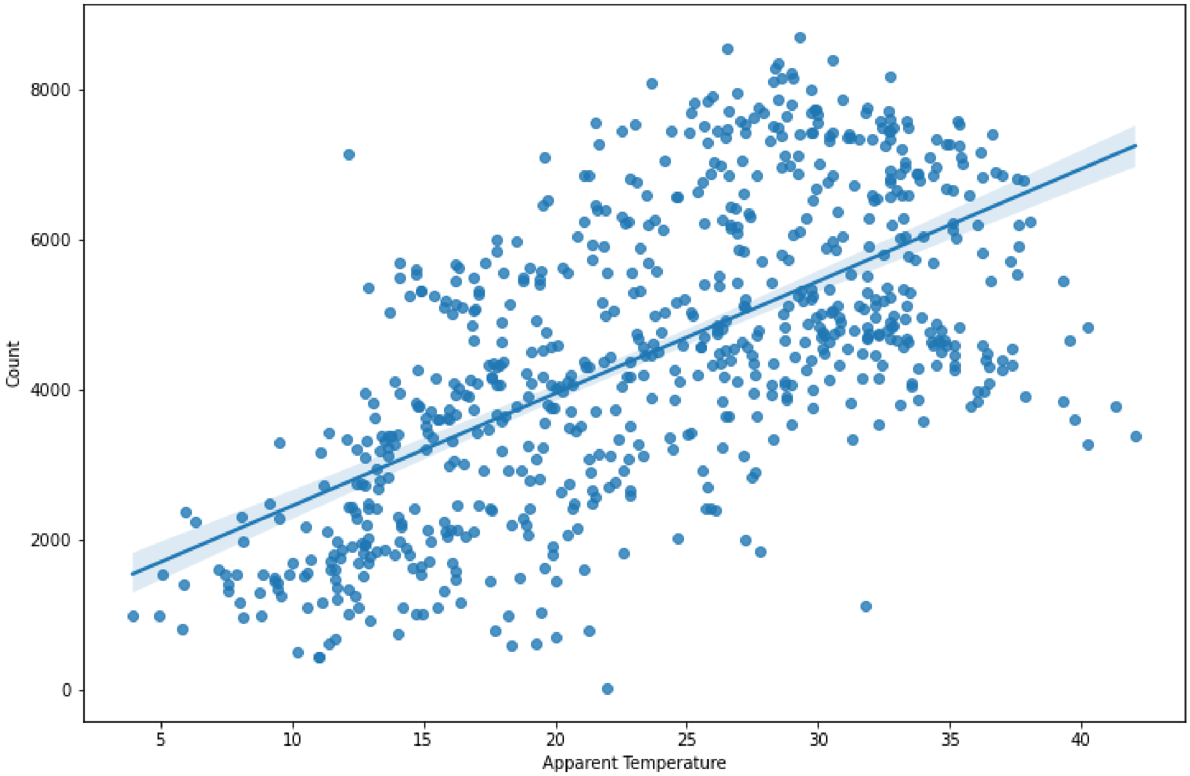


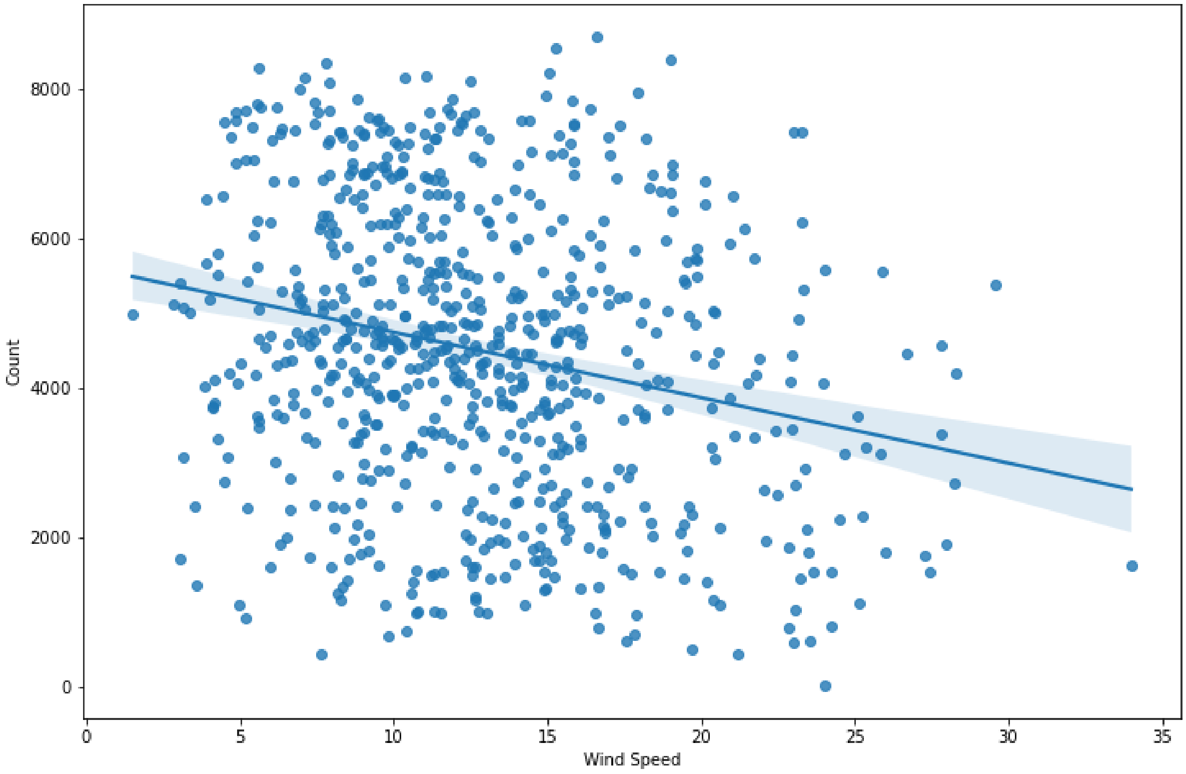
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

The four assumptions associated with a linear regression model are linearity, homoscedasticity, independence, and normality (Liu et al., 2016).

Linearity:

For categorical dummy variables, linearity is implied. The relationship between quantitative predictive variables and the target variable was visualized using pairplot. The pairplot showed a strong positive linear relationship between Temperature and the target variable, Count. The pair plot also showed a negative relationship between Wind Speed and Count, but with more variability. A regplot was created for each variable to verify these relationships.


Count
0
2000
4000
6000
8000
5
10
15
20
25
30
35
40
Apparent Temperature


Count
0
2000
4000
6000
8000
0
5
10
15
20
25
30
35
Wind Speed

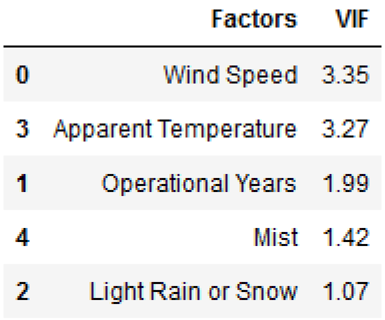
Homoscedasticity:

The p-value that the for the Breusch-Pagan test is 0.00005, and therefore the null hypothesis of homoscedasticity is not rejected. The f-statistic of the hypothesis that the error variance does not depend on x, is 5.50 and its corresponding p-value is 0.00006.

Independence:

The Durbin–Watson statistic for the for the model is 1.828. This is fairly close to the ideal value of 2, and within the acceptable range of 1.5 to 2.5.

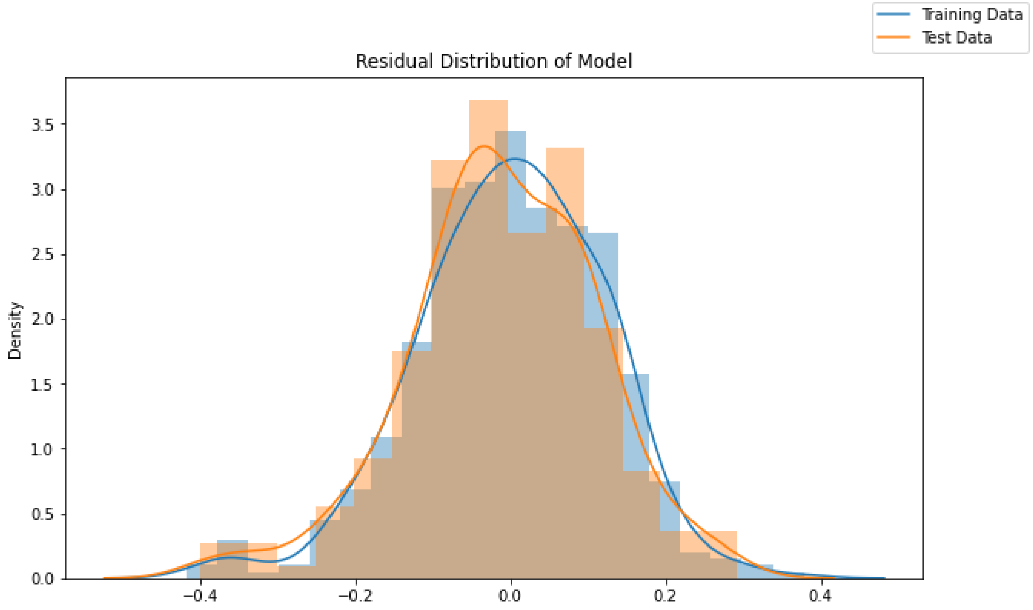
The variance inflation factors are below the generally acceptable threshold of 5.0.



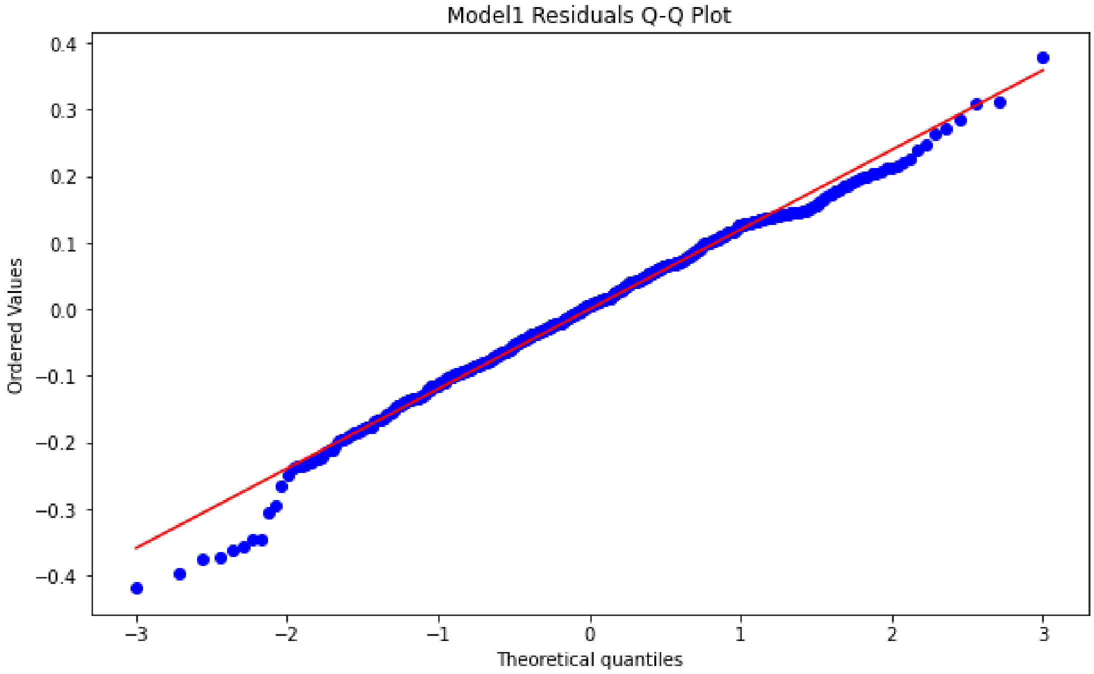
	Factors	VIF
0	Wind Speed	3.35
3	Apparent Temperature	3.27
1	Operational Years	1.99
4	Mist	1.42
2	Light Rain or Snow	1.07

Normality of Residuals:

The residual errors of this model follow a normal distribution. The distribution is visualized with a distplot. There is a long tail around -0.35 but with very few data points:



The points on the quantile-quantile plot mostly follow a straight line:



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

The top three factors, in order of importance, are:

Factor	**Field**	**Unscaled Values / Range**	**Type**	**Scaled Coefficient**
Apparent Temperature	atemp	3.95 – 42.04 deg C	Quantitative	0.630648
Operational Years	yr	2018 / 2019	Categorical	0.242467
Light Rain or Snow	weathersit	3	Categorical	-0.219629

The strongest predictor of demand is the Apparent Temperature (‘atemp’). Apparent temperature is highly correlated (0.99) with Temperature (‘temp’), so one would not use both these predictor variables in the same model.

The growth in demand over time is captured in the factor named Operational Years (‘yr’). It represents the increase in demand from the year 2018 to 2019. It is the second strongest predictor of demand.

Demand decreases substantially when there is Light Rain or Snow. It is the third strongest predictor or demand.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a statistical method used for predictive analysis. It is used to predict the value of the unknown predicted (dependent) variable from the values of one or more known predictor (independent) variables. Linear regression relies on the assumption that the relationship between the predicted and predictor variables are linear.

The equation for linear regression is:

$$\hat{Y} = \mathrm{b}_0 + \mathrm{b}_1 \mathrm{X}_1 + \mathrm{b}_2 \mathrm{X}_2 + \ldots + \mathrm{b}_\mathrm{n} X_n + \in$$

where:

$\hat{Y}$ = the predicted (dependent) variable
b_0 = the constant y-intercept value
$\mathrm{b}_\mathrm{n} X_n$ = the regression coefficient and value of the nth predictor (independent) variable
$\in$ = the error of the model

The coefficients of the predictor variables in a linear regression model are often estimated using ordinary least squares. Least squares is a approach in linear algebra to find an approximate solution to overdetermined systems, where there are more equations than unknowns. Least squares minimizes the sum of the squares of the residuals, where a residual is the difference between observed value and the value computed by the linear regression model for one equation.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is four datasets that have very different distributions, but almost identical summary statistics. Anscombe created the dataset in 1973 to demonstrate the effect of outliers on statistical properties, and to emphasize the importance of visualizing dataset when analyzing it.

Even through the four datasets look drastically different when visualized, their x and y means, x and y sample variances, x and y correlations, linear regression lines, and r-squared values are identical to 2 decimal places or better.

The quartet is often included in statistics handbooks and introductory statistics courses as an illustration to students about the pitfalls associated with evaluating basic summary statistics for realistic datasets.

3. What is Pearson's R?

Pearson's correlation coefficient is a normalized measurement of covariance between two variables with a range between -1 and 1. It only reflects linear correlations between datasets.

A value of 1 indicates a perfect positive linear correlation between the datasets, while a value of -1 indicates a perfect negative linear correlation. A value of 0 indicates that there is no linear correlation between the datasets.

Pearson's correlation coefficient is calculated as the covariance of the two datasets divided by the product of their standard deviations. In Python, it can computed using the Pandas Dataframe.corr() method.

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

Gradient descent-based algorithm and distance-based algorithms are sensitive to scale. Feature scaling is used to ensure that variables have equal and consistent scales so that these algorithms are not affected.

Scaling also ensures consistency between model coefficients – making it easier to distinguish between the influence of individual predictor variables when building a multiple linear regression model.

Normalization is a scaling technique that is used to scale variables so that they fall within a specific range, usually between zero and one. Normalization is well-suited to scale quantitative variables where they coexist with categorical variables or dummy variables that can take a discrete value of either zero or one. Normalizing such a dataset will not affect these categorical variables or dummy variables.

Standardization is a scaling technique that is used to scale variables so that they are centered around a mean with a unit of standard deviation. One advantage of standardization is that it not affected by outliers which can compress the range in the case of normalization.

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

An infinite variance inflation factor value indicates that there is a correlation of 1.0 between two variables. Thus, one variable completely explains the other variable. The variance inflation factor is infinite because R-squared is equal to 1.0, and variance inflation factor is calculated as 1 / (1 - R-squared), leading to a division by zero.

When there is high correlation between independent variables in a linear regression model, one should iteratively remove correlated variables until the variance inflation factor for all independent variables is at an acceptable level. A variance inflation factor above 5 is generally seen as problematic, and independent variables with values higher than 5 should be investigated and considered for removal from the model. A factor above 10 is unacceptable, as it indicates an R-squared value of 0.90.

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

Quantile-quantile plots are used to plot the quantiles of two probability distributions against each other (Varshney, 2020). It allows one to graphically analyze and compare the quantiles of two probability distributions.

Quantile-quantile plots are an effective mechanism to validate that a distribution is normal. If the distribution is normal, the points on the quantile-quantile plot will lie on the straight-line y=x. A deviation at the bottom end of the line indicates that the distribution is left-skewed (left tail). Likewise, a deviation at the top end of the line indicates that the distribution is right-skewed (right tail). A deviation on both ends of the line is an indication that the data is not normally distributed.

Ordinary least-squares estimators are maximum likelihood estimators for normally distributed data. If the dataset does not have a normal distribution, the ordinary least-squares will not be optimal, and the linear regression model will therefore not be accurate.

When building a linear regression model, quantile-quantile plots can be used to validate whether data is normally distributed. This is one of the mechanisms that can be used to ensure that the model will be accurate.

References

Liu, C.-T., Milton, J., & McIntosh, A. (2016, January 6). Simple Linear Regression. Retrieved June 15, 2022, from https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/R/R5_Correlation-Regression/R5_Correlation-Regression4.html

Varshney, P. (2020, April 14). Q-Q Plots Explained. Retrieved June 15, 2022, from https://towardsdatascience.com/q-q-plots-explained-5aa8495426c0