Linear Regression Assignment
-- Subhadip Ghosh

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?
Answer: We have created dummy variables out of the categorical variables. This helps to identify/infer the effect of the particular labels of the categorical variables. Not every labels/values of those categorical variable have impact on the dependent variable.
E.g. mnth(month) had 12 lables, out of only 9(september) have impact on the dependent variable.

2. Why is it important to use drop_first=True during dummy variable creation?
Answer: While creating dummy variable, the drop_first=true is important, because it reduces the number of dummy columns created. Also this in turn help reducing correlations among the dummy variables.

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

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
Answer:
Distribution of error term is Normal: is validated using Residual Analysis
Independent variables
There is a linear relationship between the X-independent variable(s) and the Y-dependent variable: Using Pairplot.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
Answer: Top 3 Features and their co-efficient:
Temperature(temp): + 0.5977
Humidity(hum): - 0.3425
WindSpeed(windspeed): - 0.2361

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
Answer: Linear regression is a supervised machine learning method that the finds a linear equation that best describes the correlation of the explanatory variables with the dependent variable.
This is achieved by fitting a line to the data using least squares. The line tries to minimize the sum of the squares of the residuals. The residual is the distance between the line and the actual value of the explanatory variable. Finding the line of best fit is an iterative process.
The following is an example of a resulting linear regression equation:

$$y = b_0 + b_1x_1 + b_2x_2 + \ldots$$

In the example above, y is the dependent variable, and x1, x2, and so on, are the explanatory variables. The coefficients (b1, b2, and so on) explain the correlation of the explanatory variables with the dependent variable. The sign of the coefficients (+/-) designates whether the variable is positively or negatively correlated. b0 is the intercept that indicates the value of the dependent variable assuming all explanatory variables are 0.

A linear regression model helps in predicting the value of a dependent variable, and it can also help explain how accurate the prediction is. This is denoted by the R-squared and p-value values. The

R-squared value indicates how much of the variation in the dependent variable can be explained by the explanatory variable and the p-value explains how reliable that explanation is. The R-squared values range between 0 and 1. A value of 0.8 means that the explanatory variable can explain 80 percent of the variation in the observed values of the dependent variable. A value of 1 means that a perfect prediction can be made, which is rare in practice. A value of 0 means the explanatory variable doesn't help at all in predicting the dependent variable. Using a p-value, you can test whether the explanatory variable's effect on the dependent variable is significantly different from 0.

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

Answer: Anscombe's quartet comprises a set of four dataset, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plot on graph. The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

The four datasets can be described as:

Dataset 1: this fits the linear regression model pretty well.

Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.

Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model

Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model

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

Answer: The Pearson correlation coefficient (r) is the most widely used correlation coefficient and is known by many names:

Pearson's r, Bivariate correlation, Pearson product-moment correlation coefficient (PPMCC), The correlation coefficient

The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.

Although interpretations of the relationship strength (also known as effect size) vary between disciplines, the table below gives general rules of thumb:

Pearson r

Pearson correlation coefficient (*r*) value	Strength	Direction
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None
Between 0 and –.3	Weak	Negative
Between –.3 and –.5	Moderate	Negative
Less than –.5	Strong	Negative

Formula :

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

Here,

r = correlation coefficient
xi = values of the x-variable in a sample
x̄ = mean of the values of the x-variable
Yi = values of the y-variable in a sample
ȳ =mean of the values of the y-variable

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

Answer:

What is scaling?

It is a step of data preparation which is applied to independent variables to normalize the data within a specific range. It also helps in speeding up the calculations in an algorithm.

Why is scaling performed?

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude. It is important to note that scaling just affects the coefficients and none of the other parameters like tstatistic, F-statistic, p-values, R-squared, etc.

Normalization/Min-Max Scaling:

It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

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

NORMALIZATION	STANDARDIZATION
Minimum and maximum value of features are used for scaling.	Mean and standard deviation is used for scaling.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
It is really affected by outliers.	It is much less affected by outliers.
This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.

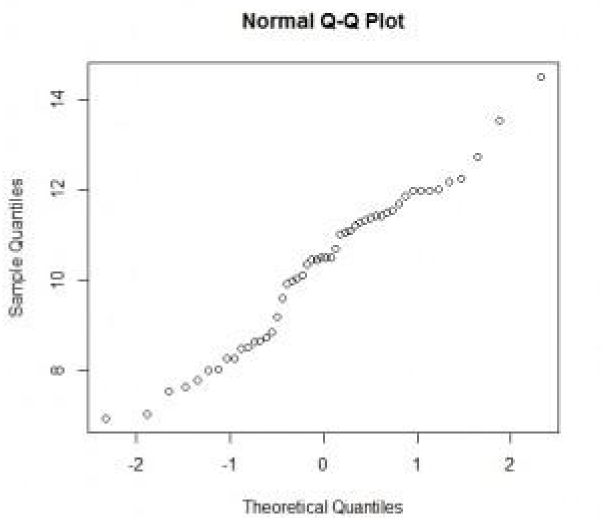
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer: If there is perfect correlation between two independent variables, then VIF = infinity.
In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/(1-R2) infinity.
To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.
An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Answer:

The Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential. For example, if we run a statistical analysis that assumes our residuals are normally distributed, we can use a Normal Q-Q plot to check that assumption. It's just a visual check, not an air-tight proof, so it is somewhat subjective. But it allows us to see at-a-glance if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation.
A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. Here's an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.



Now what are "quantiles"? These are often referred to as "percentiles". These are points in your data below which a certain proportion of your data fall. For example, imagine the classic bell-curve standard Normal distribution with a mean of 0. The 0.5 quantile, or 50th percentile, is 0. Half the data lie below 0. That's the peak of the hump in the curve. The 0.95 quantile, or 95th percentile, is about 1.64. 95 percent of the data lie below 1.64. The following R code generates the quantiles for a standard Normal distribution from 0.01 to 0.99 by increments of 0.01:
qnorm(seq(0.01,0.99,0.01))

We can also randomly generate data from a standard Normal distribution and then find the quantiles. Here we generate a sample of size 200 and find the quantiles for 0.01 to 0.99 using
the quantile function:
quantile(rnorm(200),probs = seq(0.01,0.99,0.01))

So we see that quantiles are basically just your data sorted in ascending order, with various data points labelled as being the point below which a certain proportion of the data fall. However it's worth noting there are many ways to calculate quantiles. In fact, the quantile function in R offers 9 different quantile algorithms! See help(quantile) for more information.

Q-Q plots take your sample data, sort it in ascending order, and then plot them versus quantiles calculated from a theoretical distribution. The number of quantiles is selected to match the size of your sample data. While Normal Q-Q Plots are the ones most often used in practice due to so many statistical methods assuming normality, Q-Q Plots can actually be created for any distribution.

As it is clear with name, the Q-Q plot is a graphical plotting of the quantiles of two distributions with respect to each other. In other words we can say plot quantiles against quantiles. Whenever we are interpreting a Q-Q plot, we shall concentrate on the ‘y = x’ line. We also call it the 45-degree line in statistics. It entails that each of our distributions has the same quantiles. In case if we witness a deviation from this line, one of the distributions could be skewed when compared to the other.