Linear Regression Assignment Solutions

By:- Aakash Sharma

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?**

Ans. Following are the effects for categorical variables on the dependent variable:

- ✓ Ride Count Seems to be in maximum in fall (autumn) followed by Summer, Spring & Winter respectively.
- ✓ Ride Count has increased drastically in 2019 as compared to 2018
- ✓ Ride Count seems to increase between May to October which are comparatively Fall(Autumn) & Summer Season in US
- ✓ Ride Count is lesser on Holidays as compared to other days.
- ✓ Working Day / Non-Working Day shows almost similar behaviour (after just visualising the data)
- ✓ Ride Count is more on Clear & Misty Days as compared to Light Snow / Rainfall
- ✓ Ride Count seems to be very much linearly dependent on Temperature
- ✓ Humidity & wind speed does not indicate any specific behaviour on just visualising the data.

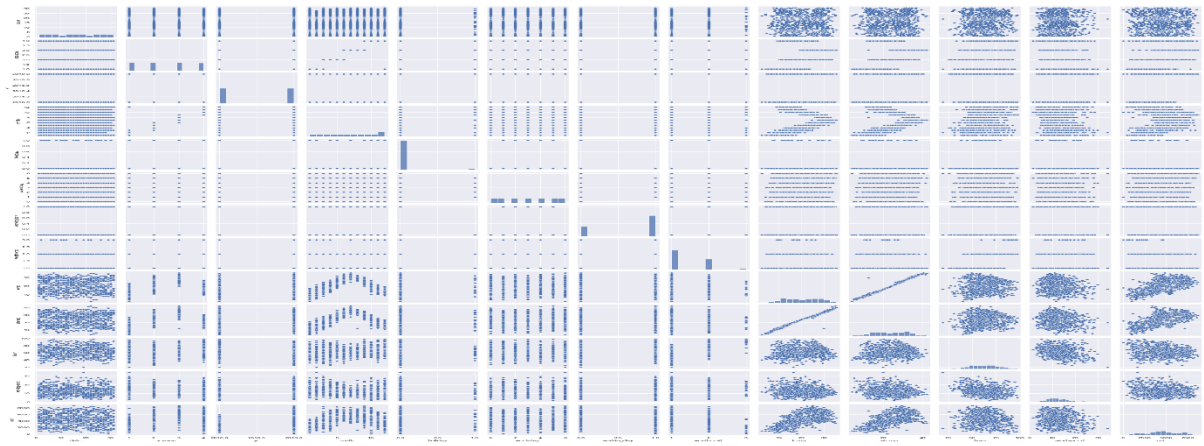
2. **Why is it important to use drop_first=True during dummy variable creation?**

Ans.

- ✓ While there are k levels in a categorical variable, it is tempting to create k dummy columns (or variables) to represent each level with a distinct column, it is important to consider the multicollinearity issues which arise with that outlook towards dummy variable creation.
- ✓ Multicollinearity is an issue that arises while building an ML model if two or more variables carry the same amount or very similar information. This essentially affects the interpretability of the model.
- ✓ If we create k dummy variables for k levels of categorical data, the kth variable contains no new information.
- ✓ This is why it is important to use drop_first = True when we create dummy variables, to avoid multicollinearity issues and maintain an interpretable model.

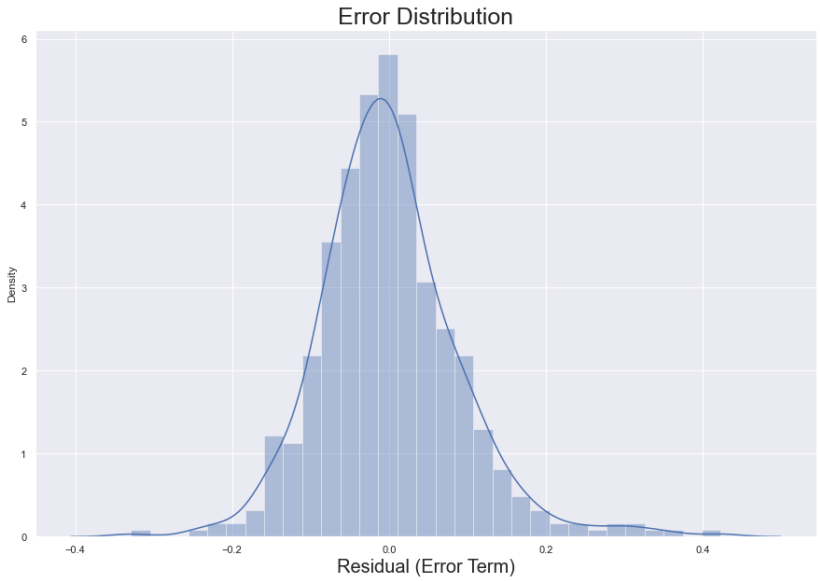
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

Ans. Temp and atemp are the numerical variables which are showing 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?

Ans. Residuals distribution should follow normal distribution and centred around 0. (mean = 0). We validate this assumption about residuals by plotting a distplot of residuals and see if residuals are following normal distribution or not. The above diagram shows that the residuals are distributed about mean = 0.



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

Ans. Based on the final model yr, temp and weather are the top 3 features contributing significantly towards explaining the demand of the shared bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans. Linear Regression Algorithm is a machine learning algorithm based on supervised learning where the predicted output is continuous and has a constant slope. It's used to predict values within a continuous range, rather than trying to classify them into categories. It is a part of regression analysis. Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), consequently called linear regression. If there is a single input variable (x), such linear regression is called simple linear regression. And if there is more than one input variable, such linear regression is called multiple linear regression.

Simple Linear Regression - Simple linear regression uses traditional slope-intercept form, where m and b are the variables our algorithm will try to "learn" to produce the most accurate predictions. x represents our input data and y represents our prediction.
y = mx + b
where,
y = dependent variable
x = independent variable
m = intercept of the line
b = linear regression coefficient

When the value of x (independent variable) increases, the value of y (dependent variable) is likewise increasing. Regression method tries to find the best fit line which shows the relationship between the dependent variable and predictors with least error.

Assumptions of Simple Linear Regression –

There are four assumptions associated with a linear regression model:

1. **Linearity**: The relationship between X and the mean of Y is linear.
2. **Homoscedasticity**: The variance of residual is the same for any value of X.
3. **Independence**: Observations are independent of each other.
4. **Normality**: For any fixed value of X, Y is normally distributed.

Multiple Linear Regression - A more complex, multi-variable linear equation might look like this, where w represents the coefficients, or weights, our model will try to learn.
f(x,y,z)=w1x + w2y + w3z

The variables x,y,z represent the attributes, or distinct pieces of information, we have about each observation.

Assumptions of Multiple Linear Regression –

Multiple linear regression analysis makes five key assumptions:

1**. Linear relationship:** There exists a linear relationship between each predictor variable and the response variable.
2**. No Multicollinearity:** None of the predictor variables are highly correlated with each other.
3. **Independence:** The observations are independent**.**
4**. Homoscedasticity:** The residuals have constant variance at every point in the linear model.
5**. Multivariate Normality:** The residuals of the model are normally distributed.

2. Explain the Anscombe's quartet in detail.

Ans. Anscombe's quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting data before you analyse it and build your model. These four data sets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets. However, when you plot these data sets, they look very different from one another. Anscombe's quartet intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough.

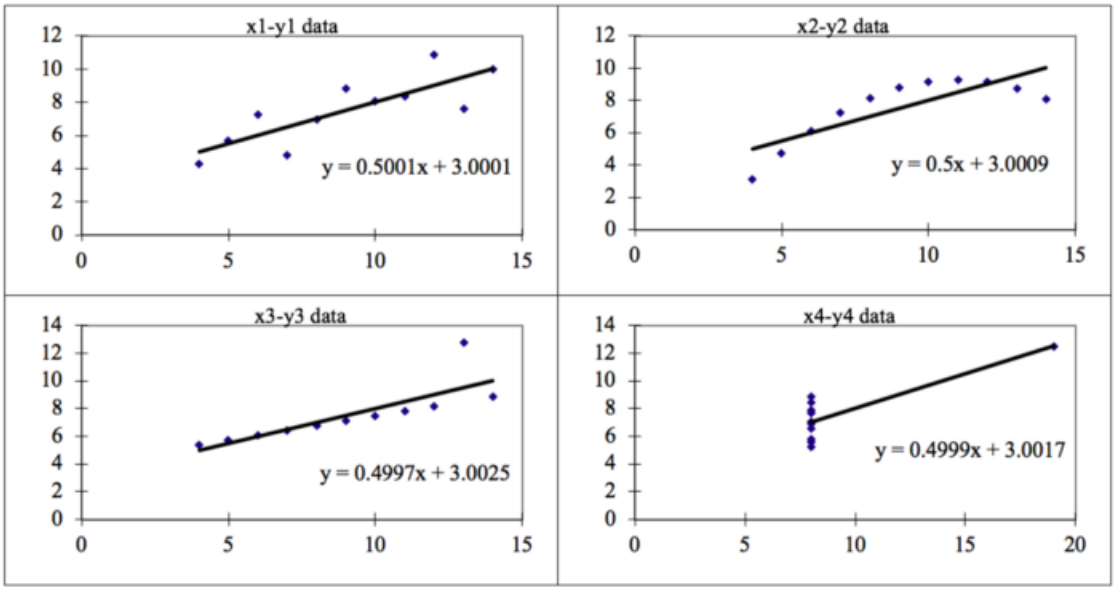
These four plots can be defined as follows:

Anscombe's Data										
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89

The statistical information for all these four datasets are approximately similar and can be computed as follows:

Anscombe's Data										
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

When these models are plotted on a scatter plot, all datasets generates a different kind of plot that is not interpretable by any regression algorithm which is fooled by these peculiarities and can be seen as follows:



The four datasets can be described as:

1. The first scatter plot (top left) appears to be a simple linear relationship,
2. The second graph (top right); cannot fit the linear regression model because the data is non-linear
3. In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line. It shows the outliers involved in the dataset which cannot be handled by linear regression model
4. Finally, the fourth graph (bottom right) shows the outliers involved in the dataset which cannot be handled by linear regression model. It shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables. It shows the outliers involved in the dataset which cannot be handled by linear regression model

3. What is Pearson's R?

Ans. The **Pearson correlation** method is the most common method used for numerical variables. It assigns a value between − 1 and 1, where 0 is no correlation, 1 is total positive correlation, and − 1 is total negative correlation. Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations.
Pearson's R Formula is as follows:

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

Here,

1. r = correlation coefficient
2. xi = values of the x-variable in a sample
3. $\bar{x}$ = mean of the values of the x-variable
4. yi = values of the y-variable in a sample
5. $\bar{y}$ = 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?

Ans. **Scaling** is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data pre-processing step. In simpler terms, in machine learning algorithms we need to bring all features in the same standing, so that one significant number doesn't impact the model just because of their large magnitude. This is called scaling or Feature scaling.
Feature scaling in machine learning is one of the most critical steps during the pre-processing of data before creating a machine learning model. Scaling can make a difference between a weak machine learning model and a better one. 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 which results in an incorrect model. 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 t-statistic, F-statistic, p-values, R-squared, etc.

Normalization is generally used when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks. It brings all of the data in the range of 0 and 1.

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

Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization. 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).

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

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

Ans. Variance Inflation Factor (VIF) is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. It is calculated by taking the ratio of the variance of all a given model's betas divide by the variance of a single beta if it were fit alone. The higher the VIF value, the greater the correlation of the variable with other variables. Values of more than 4 or 5 are sometimes regarded as being moderate to high, with values of 10 or more being regarded as very high. If there is perfect correlation, then **VIF** = **infinity**. An infinite VIF value means that the variable is exactly linear combination of other variable. If the independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and it's R-squared value will be equal to 1. So, VIF = 1/(1-1) which gives VIF = 1/0 which results in "infinity"

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

Ans. Quantile - Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, Exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution. A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions. The power of Q-Q plots lies in their ability to summarize any distribution visually.

The advantages of the Q-Q plot are:

1. The sample sizes do not need to be equal.
2. Many distributional aspects can be simultaneously tested.

Q-Q plot is very useful to determine:

1. If two populations are of the same distribution
2. If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it's met using this.
3. Skewness of distribution