Subjective Test Answers

Chaitanya Namburu

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?*

Answer:

Season: The count is higher in Summer and Fall, compared to Winter and Spring. It seems like hot seasons are preferred by the customers.

Holiday: The average count is higher on a non-holiday, compared to a holiday.

Weather Situation: The bikes were taken more when the weather was calm and clear, compared to misty conditions. It was followed by light snow and the bikes were literally not taken during heavy rains and thunderstorms.

Month: The count gradually increased from the start of the year to around Sep/Oct and decreased towards the end of the year. It coincided with our understanding of seasons. The busiest months were Jul-Oct.

Weekday: Thursday, Friday, Saturday & Sunday seemed to have higher count compared to the beginning of the week.

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

Answer:

"drop_first" option is important to avoid multicollinearity among the variables. This is used for categorical variables.

We need k-1 dummy variables to represent a categorical variable with k levels.

Example:

If there are 4 levels in a categorical variable (e.g., seasons = summer, winter, autumn, rainy) we need only 3 variables to represent the mentioned categorical variable. If we don't drop the first column, it leads to multicollinearity among the variables.

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

Answer:

It can be observed from the pair-plot that Count has the highest correlation with "temp" (Temperature) variable.

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

Answer:

I have validated the assumptions of the Linear Regression model by ensuring the following:

1. The residuals (error terms) should have a normal distribution. I have graphed out the residuals and observed that they are approximately in a normal distribution.
2. Homoscedasticity – I have ensured that there is no pattern among the residuals.
3. There was no autocorrelation among the residuals.
4. There was no multicollinearity among the variables. The VIF of the variables in the final model was less as per the acceptable standards.
5. Linear relationship – there was linear relationship between the variables and the dependent variable.

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

Answer:

The top features contributing significantly explaining the demand of bikes are:

- Temperature
- Weather Situation 3 (Light snow)
- Year
- Windspeed

General Subjective Questions

Q1. *Explain the linear regression algorithm in detail.*

Answer:

Linear regression is a machine learning technique to predict unknown values of a target variable by studying known values of the target variable with respect to a few other independent variables. The technique involves the development of a model by using a training set of data which establishes a linear relationship the independent variables and the dependent variable. Then we use the model to determine the values of the target variable for other values of independent variables.

It is classified as Simple Linear Regression and Multiple Linear Regression based on the number of independent variables.

The overall goal is to mathematically establish a linear relation using the formula:

$$Y = mx + c$$

Where Y = target variable and x = independent variable and c = constant

Multiple Linear Regression:

The formula in the case of multiple linear regression:

$Y = \underline{b0} + b1.x1 + b2.x2 + b3.x3 + \ldots b_n x_n$

Using the training data properly, we try to find the coefficients and constant value to best fit the available data. We use various techniques like gradient descent to find the optimal coefficients.

The generated model should satisfy some assumptions to be an efficient model:

- Linearity
- Absence of multicollinearity
- Absence of Autocorrelation
- Normality of residuals
- Homoscedasticity

Q2. Explain the Anscombe's quartet in detail.

Answer:

Developed by the French statistician Anscombe, the quartet consists of four datasets which have similar descriptive statistics but are different in the nature of the datasets. The 4 datasets appear very different if they are graphed out.

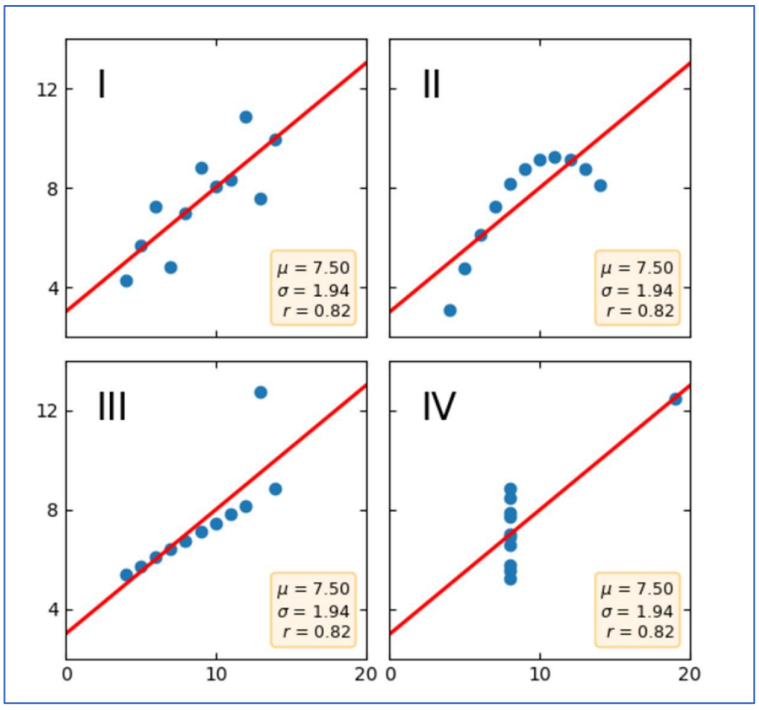
Example:

For the 4 data sets given below, the descriptive statistics are the same, but what happens after we graph them out can be seen below:

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 descriptive statistics are as follows for the 4 datasets:

- Mean of x = 9 and Mean of y = 7.5 (for all the datasets)
- Variance of x = 11 & Variance of y = 4.13 (for all the datasets)
- Correlation coefficient between x and y = .816 for each dataset



The datasets highlight the importance of visualization of the data instead of blindly following the data models or descriptive statistics.

Q3. *What is Pearson's R?*

Answer:

Pearson's r, also known as Coefficient of Correlation, is a numerical measure of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative. The Pearson correlation coefficient, r, can take a range of values from +1 to -1.

A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of

the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases.

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

Answer:

Feature scaling is the process of normalizing the range of independent variables or features in a dataset. It is done in order to train the model to treat all the independent variables with equal importance. If the features are not scaled, the coefficients will be quite variant and will be difficult to ascertain which of them are important for the generated model. Sometimes, the units of the variables also might cause undue importance given to them. For example, if one of the column is in kilograms and other is in pounds, the pounds variable will have higher number always and might seem unduly important in the generated model.

Scaling can be done in 2 ways:

- Normalised scaling
- Standardised scaling

Normalised scaling:

It uses the formula:

```
x_scaled = (x - min(x)) / (max(x) - min(x))
```

So, the values are scaled between 0 and 1. Generally, the MinMaxScaler from ScikitLearn is used for this. It is affected by outliers present in the data because everything is squeezed between 0 and 1.

Standardised scaling:

It uses the formula:

```
x_scaled = (x - mean(x)) / std(x)
```

The scaled data will have mean = 0 and it standard deviation = 1.

It is not affected by the presence of outliers. Generally, the StandardScaler from ScikitLearn is used for development.

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

Answer:

VIF = 1 / (1 – R2)

When R2 is 1, the value becomes infinity. This occurs when there is a perfect correlation between the data variables. The higher the VIF value, the higher the correlation between the variables. This leads to multicollinearity and renders the model inefficient.

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

Answer:

Quantile-Quantile plot or Q-Q plot is a scatter plot created by plotting 2 different quantiles against each other and is used to determine if they come from populations with a common distribution.

The first dataset is the test variable, and the second dataset is the actual distribution that we are testing it against.

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions. Importance of Q-Q plot: When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences.