Subjective Questions
and
Answers

Submitted by
Mohana

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?**

   I inferred the following about the effect of categorical variables on dependent variables by analyzing through visualization.

   - Highest number of bikes are rented during the 'fall' and 'summer' seasons.
   - More bikes are rented during the year 2019 than in 2018.
   - Bikes rental is increasing continuously from January till June.
   - The highest number of bikes are rented during 'September' month.
   - Number of bikes rented during holidays is comparatively more than the non-holidays.
   - There is not much difference in bike rentals during working days and weekends.
   - More bikes are rented when the sky is clear.

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

   drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

   Let's say we have 4 types of values in the Categorical column 'season' in our data set and we want to create a dummy variable for that column. If one variable is not Spring, not Summer, and not Winter, then it is obviously Fall. So, we do not need 4^{th} variable to identify the 'Fall' season.

Hence if we have a categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

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

'temp' and 'atemp' are the numerical variables that have the highest correlation with the target variable.

Correlation of temp with target variable = 0.63
Correlation of atemp with target variable = 0.63

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

- Linearity check

Validated the linearity relationship between dependent and independent variables.

- Normality check

Validated whether distribution of the error terms is normally distributed.

- Homoscedacity

Validated whether the error term have constant variance.

- Independence of error terms

Validated whether the residuals are independent of each other.

- Small or no multicollinearity between the features.

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

- atemp
- yr
- winter

General Subjective Questions

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

Linear regression is an algorithm that provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events. It is a statistical method used in data science and machine learning for predictive analysis.

The independent variable is also the predictor or explanatory variable. The regression model predicts the value of the dependent variable, which is the response or outcome variable being analyzed.

The equation of the simple linear regression is

$$Y = m*X + b$$

Where X = independent variable (target)

Y = dependent variable

m = slope of the line (slope is defined as the ‘rise’ over the ‘run’)

The equation for multiple linear regression is similar to the equation for a simple linear equation,

$$Y = b + m1*x1 + m2*x2 + \ldots\ldots + mn*xn$$

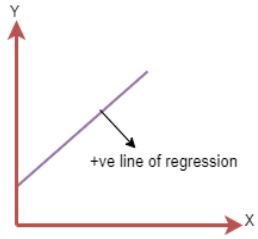
Where x1,x2,..xn are independent variables and m1,m2,..mn are the slope/coefficients.

Linear Regression Line

A linear line showing the relationship between the dependent and independent variables is called a regression line. A regression line can show two types of relationship:

Positive Linear Relationship:

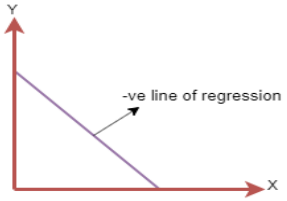
If the dependent variable increases n the Y-axis and the independent variable increases on X-axis, then such a relationship is termed as a Positive linear relationship.



The line equation will be: $Y= a_0+a_1x$

Negative Linear Relationship:

If the dependent variable decreases on the Y-axis and the independent variable increases on the X-axis, then such a relationship is called a negative linear relationship.



The line of equation will be: $Y= -a_0+a_1x$

Best Fit line

Our main goal is to find the best fit line which means the error between predicted values and actual values should be minimized. The best fit line will have the least error.

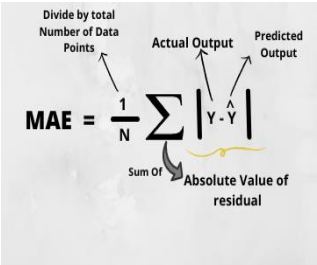
Metrics

We have the following metrics to measure evaluate the model.

1. Mean square error

$$MSE = \frac{1}{n}\Sigma\underbrace{\left(y - \hat{y}\right)^2}_{\text{The square of the difference between actual and predicted}}$$

2. Mean Absolute error



3. R2

$$\text{R2 Squared} = 1 - \frac{SSr}{SSm}$$

SSr = Squared sum error of regression line

SSm = Squared sum error of mean line

2. Explain the Anscombe's quartet in detail.

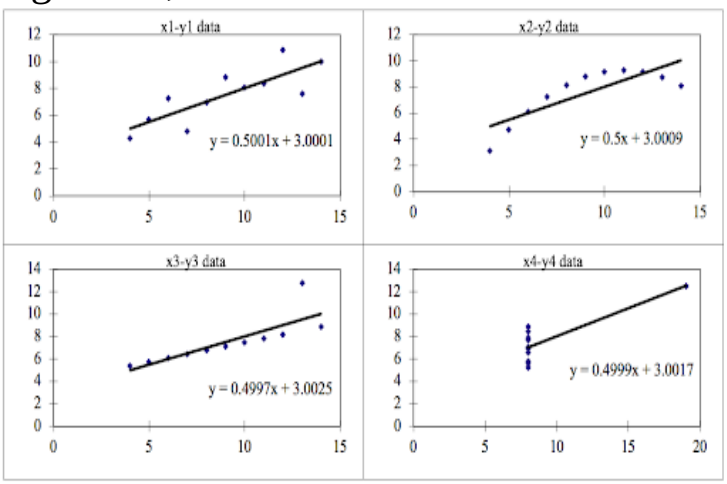
Anscombe's quartet is a group of four data sets that are nearly identical in simple descriptive statistics, but there are peculiarities that fool the regression model once you plot each data set. As you can see, the data sets have very different distributions so they look completely different from one another when you visualize the data on scatter plots.

The statistical information for these four data sets are approximately similar. We can compute them 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	

However, when these models are plotted on a scatter plot, each data set generates a different kind of plot that isn't interpretable by any regression

algorithm, as we can see below:



We can describe the four data sets as:

ANSCOMBE'S QUARTET FOUR DATASETS

- Data Set 1: fits the linear regression model pretty well.
- Data Set 2: cannot fit the linear regression model because the data is non-linear.
- Data Set 3: shows the outliers involved in the data set, which cannot be handled by the linear regression model.
- Data Set 4: shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

As we can see, Anscombe's quartet helps us to understand the importance of data visualization and how easy it is to fool a regression algorithm. So, before attempting to interpret and model the data or implement any machine learning algorithm, we first need to visualize the data set in order to help build a well-fit model.

3. **What is Pearson's R?**

The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables. When one variable changes, the other variable changes in the same direction.

Formula

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

r = correlation coefficient
x_i = values of the x-variable in a sample
$\bar{x}$ = mean of the values of the x-variable
y_i = values of the y-variable in a sample
$\bar{y}$ = 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?**

Scaling is a process of normalizing the independent variables within a particular range.

In real-time, we get data with a mix of magnitudes or units. That is each variable may have different units. If we are using the data as such without scaling, we may result in an incorrect model. The main purpose of scaling is to bring all the variables to the same magnitude/unit. Scaling also speeds up the calculation in the algorithm.

There are two types of scaling methods in data preprocessing. They are

1. Normalization
2. Standardization

Normalization

Normalization brings all the data within the range of 0 to 1. The formula is as follows:

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

Standardization

Standardization uses the z-score method. It brings all the data to standard normal distribution with mean as 0 and standard deviation as 1

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

In our dataset, we have applied MinMaxScaler function that is we applied Normalization for the below numeric variables.
'temp','atemp','hum','windspeed','cnt'.

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

When there is very high collinearity (multicollinearity) between the independent variables, then the value of VIF results in infinite.

For perfect collinearity, R2 results in 1. Hence VIF which is 1/(1-R2) becomes (1/0) infinite.

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

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other.The purpose of Q Q plots is to find out if two sets of data come from the same distribution.

A 45-degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.
If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

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 advantages of the q-q plot are: The sample sizes do not need to be equal. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.