Linear Regression Subjective Questions

Assignment-based Subjective Questions

Question1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer: Below are the observations:

- Bikes are less in demand in the spring than other seasons
- Demand increased in 2019 from 2018.
- Jun to Sep - demand is high and Jan - lowest demand
- demand is less in holidays.
- weekdays- demand is almost similar.
- No specific change in demand with workign day and non working day.
- weathersit :
    - 1: Clear, Few clouds, Partly cloudy, Partly cloudy ---demand is high
    - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds ---comparatively less demand
    - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist -very less demand
    - 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog -not much data probably no demand

Question2: Why is it important to use drop_first=True during dummy variable creation?

Answer: By using parameter drop_first = True, this will drop the first dummy variable, thus it will give n-1 dummies out of n discrete categorical levels by removing the first level. If we don't use drop_first=True then instead of n-1, n discrete categorical levels will be used and there will be one extra column get created. E.g.

```
sales_data = pd.DataFrame({"name":["William","Emma","Sofia","Markus","Edward"]

                           ,"sales":[50000,52000,90000,34000,42000]

                           ,"region":["East","North","East","South","West"]

                           }

                          )


print(sales_data)
```

Out

	Name	Sales	Region
0	William	50000	East
1	Emma	52000	North
3	Sofia	90000	East
4	Markus	34000	South
5	Edward	42000	West

The region variable is a categorical variable that we'll be able to transform into 0/1 dummy variables.

pd.get_dummies(sales_data.region)

OUT:

	East	North	South	West
0	1	0	0	0
1	0	1	0	0
3	1	0	0	0
4	0	0	1	0
5	0	0	0	1

EXPLANATION

The output of pd.get_dummies is a group of 4 new variables:

East, North, South, West

There's one new variable for every level of the original categorical variable.

Where the value was 'East' in the original Series, the new East variable has a value of 1 (and the values for the other variables are 0).

Where the value was 'North' in the original Series, the new North variable has a value of 1 (and the values for the other variables are 0).

So the get_dummies function has recorded a single variable with 4 values, into 4 variables with 0 or 1 values. The new structure effectively contains the same information, but it's represented in a different way.

by setting drop_first = True causes get_dummies to exclude the dummy variable for the first category of the variable you're operating on.

When you have a categorical variable with K mutually exclusive categories, you actually only need K – 1 new dummy variables to encode the same information.

This is because if all of the existing dummy variables equal 0, then we know that the value should be 1 for the remaining dummy variable.

So for example, if region_North == 0, and region_South == 0, and region_West == 0, then region_East must equal 1. This is implied by the existing 3 dummy variables, so we don't need the 4th. The extra dummy variable literally contains redundant information.

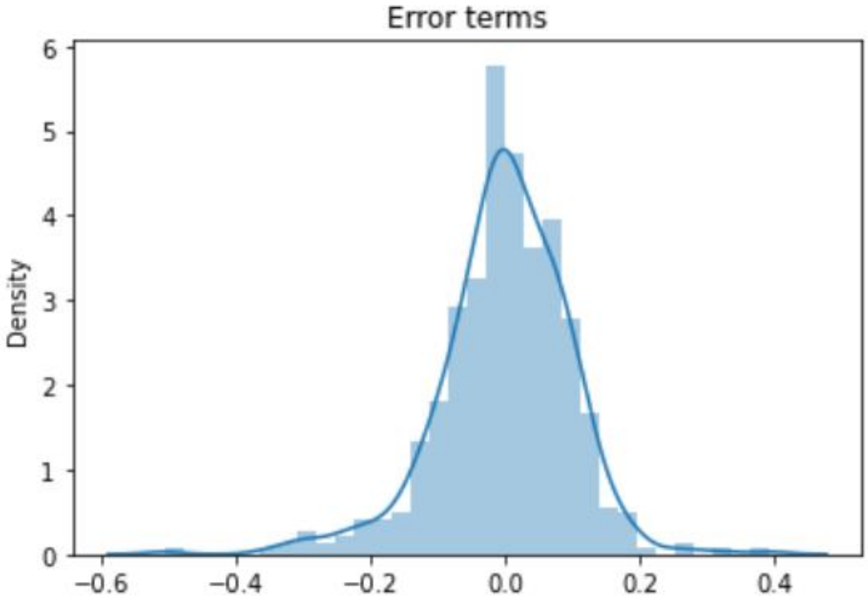
So, it's a common convention to drop the dummy variable for the first level of the categorical variable that you're encoding.

Question3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Answer: atemp and temp both have the same correlation with target variable which is the highest among all numerical variables.

Question4: How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer: To validate the assumptions of Linear Regression after building the model on the training set we plotted distplot of the residuals and the distribution plot of error term shows the normal distribution with mean at Zero.



Question5: Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Answer: Based on final model top three features contributing significantly towards explaining the demand are:

- Temperature (0.414)
- weathersit : weathersit_LightSnow_LightRain (-0.266)
- yr (0.256)

General Subjective Questions

Question1: Explain the linear regression algorithm in detail?

Answer: Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used. There are many names for a regression's dependent variable. It may be called an outcome variable, criterion variable, endogenous variable, or regress and. The independent variables can be called exogenous variables, predictor variables, or regressors.

Linear regression is used in many different fields, including finance, economics, and psychology, to understand and predict the behavior of a particular variable. For example, in finance, linear regression might be used to understand the relationship between a company's stock price and its earnings, or to predict the future value of a currency based on its past performance.

One of the most important supervised learning tanks is regression. In regression set of records are present with X and Y values and this values are used to learn a function, so that if you want to predict Y from an unknown X this learn function can be used. In regression we have to find value of Y, So, a function is required which predicts Y given XY is continuous in case of regression.

Here Y is called as criterion variable and X is called as predictor variable. There are many types of functions or modules which can be used for regression. Linear function is the simplest type of function. Here, X may be a single feature or multiple features representing the problem.

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x)). Hence, the name is Linear Regression. In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

Mathematically, we can write a linear regression equation as:

$y = a + bx$

Where a and b given by the formulas:

$$b(slobe) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a(intercept) = \frac{n\sum y - b(\sum x)}{n}$$

Here, x and y are two variables on the regression line.

b = Slope of the line

a = y-intercept of the line

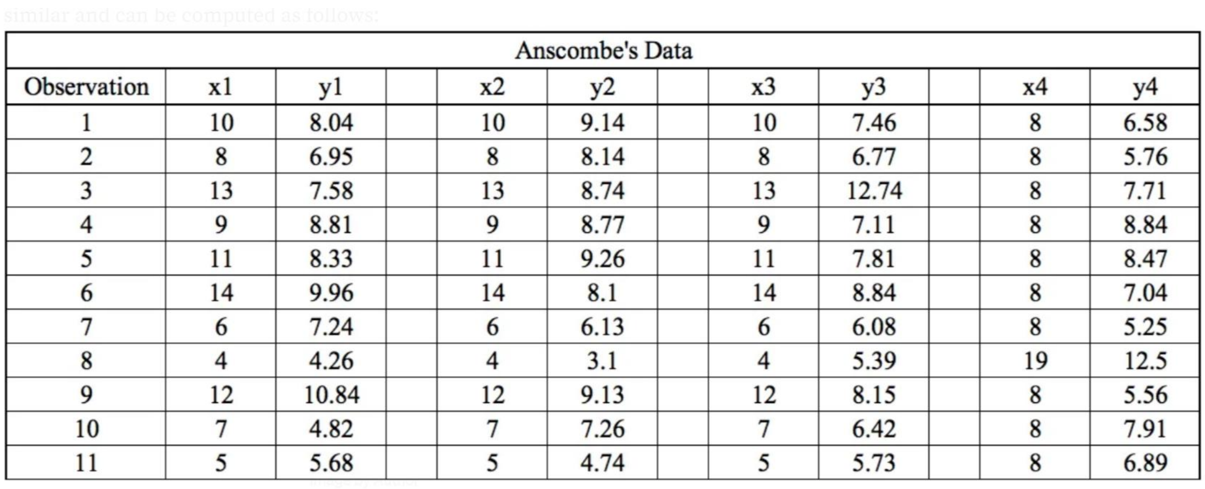
x = Independent variable from dataset

y = Dependent variable from dataset

Question2: Explain the Anscombe’s quartet in detail.

Answer: Anscombe’s Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

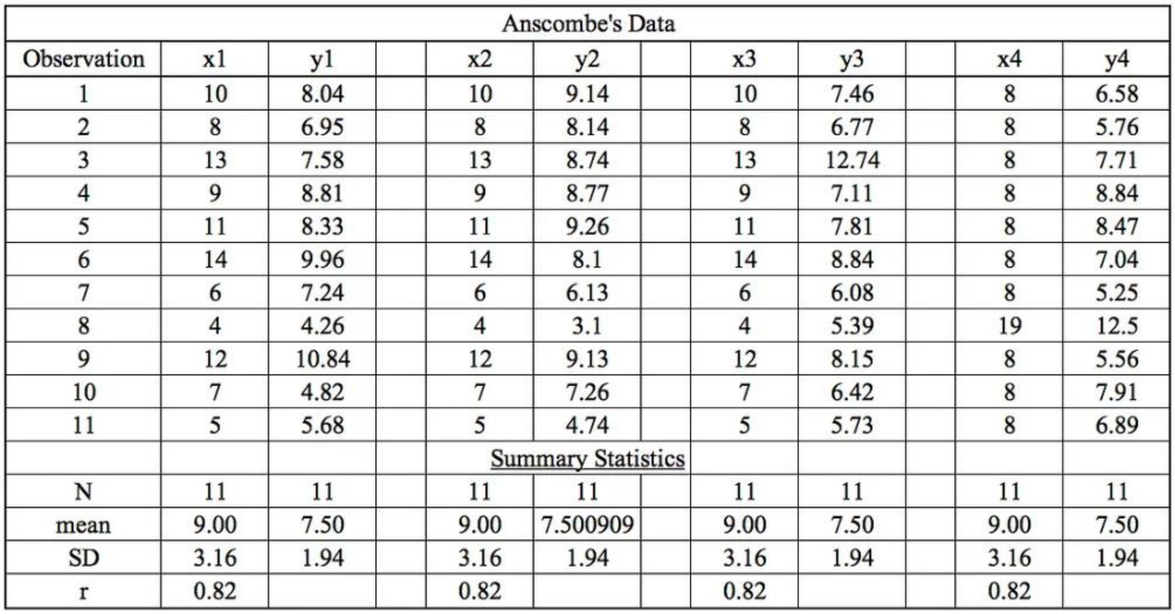
Anscombe’s Quartet tells us about the importance of visualizing the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets. These four plots can be defined as follows:



similar and can be computed as follows:

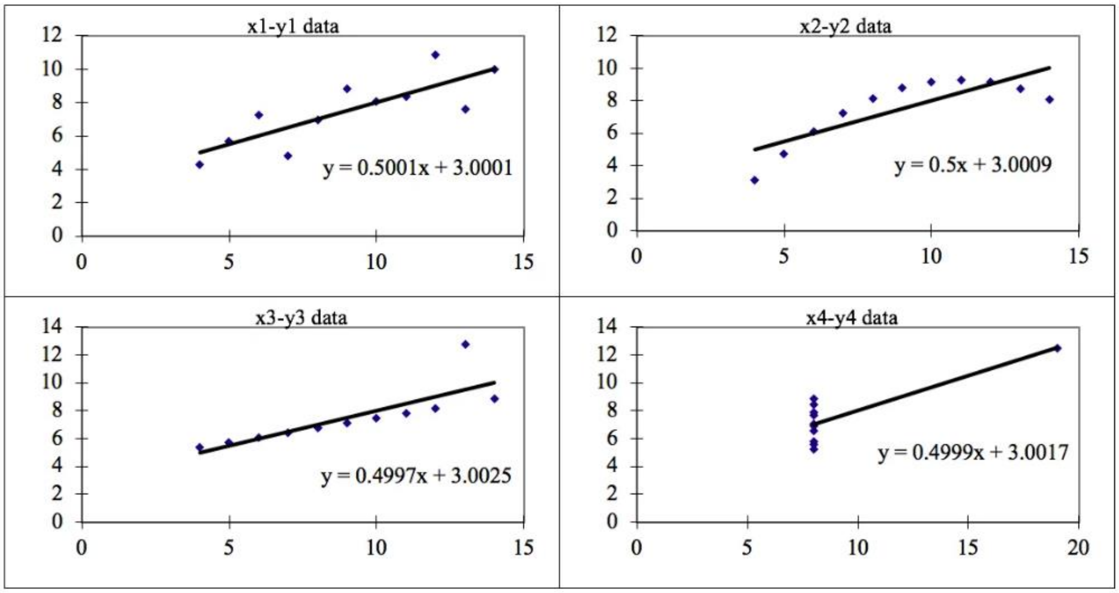
Anscombe's Data											
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		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

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:

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

Question3: What is Pearson’s R?

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 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

The Pearson correlation coefficient is also an inferential statistic, meaning that it can be used to test statistical hypotheses. Specifically, we can test whether there is a significant relationship between two variables.

When to use the Pearson correlation coefficient

The Pearson correlation coefficient (r) is one of several correlation coefficients that you need to choose between when you want to measure a correlation. The Pearson correlation coefficient is a good choice when all of the following are true:

Both variables are quantitative: You will need to use a different method if either of the variables is qualitative.

The variables are normally distributed: You can create a histogram of each variable to verify whether the distributions are approximately normal. It's not a problem if the variables are a little non-normal.

The data have no outliers: Outliers are observations that don't follow the same patterns as the rest of the data. A scatterplot is one way to check for outliers—look for points that are far away from the others.

The relationship is linear: "Linear" means that the relationship between the two variables can be described reasonably well by a straight line. You can use a scatterplot to check whether the relationship between two variables is linear.

Question4: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer:

Feature Scaling: In general Data set contains different types of variables having different magnitude and units (kilograms, grams, Age in years, salary in thousands etc).The significant issue with variables is that they might differ in terms of range of values. So the feature with large range of values will start dominating against other variables. Models could be biased towards those high ranged features. So to overcome this problem, we do feature scaling. The goal of applying Feature Scaling is to make sure features are on almost the same scale so that each feature is equally important and make it easier to process by most ML algorithms.

Why: 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 t-statistic, F-statistic, p-values, R-squared, etc.

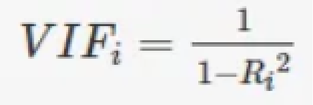
Normalization vs Standardization

- If you have outliers in your feature (column), normalizing your data will scale most of the data to a small interval, which means all features will have the same scale and hence it will not handle outliers well.
- Standardization is more robust to outliers, and in many cases, it is preferable over Max-Min Normalization.
- Normalization is good to use when your data does not follow a Normal distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks.
- Standardization, on the other hand, can be helpful in cases where the data follows a Normal 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.

<u>Question5:</u> You might have observed that sometimes the value of VIF is infinite. Why does this happen?

<u>Answer:</u> A variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis. Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model. This can adversely affect the regression results. Thus, the variance inflation factor can estimate how much the variance of a regression coefficient is inflated due to multicollinearity.

The value of VIF is calculated by the below formula:



$$VIF_i = \frac{1}{1-R_i^2}$$

Where, ‘i’ refers to the ith variable.

If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables.

Question6: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Answer: 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. This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Advantages:

- It can be used with sample sizes also
- Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios:

If two data sets —

- come from populations with a common distribution
- have common location and scale
- have similar distributional shapes
- have similar tail behavior

Interpretation:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.