

# Linear Regression Assignment

BIKE-SHARING PROVIDER BOOMBIKES

CHETAN DESAI

## Assignment-based Subjective Questions

**Question 1.** From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

**Answer:** Following are the categorical variables present in the given dataset

Season, Year, Month, Weathersit, Weekday

According the boxplot analyses, we have the following conclusion:

Plot1: **SeasonVsCount** - The count is least during spring

Plot2: **YearVsCount** - The count has increased in the next year (2019)

Plot3: **MonthVsCount** - The count has decrease in the first 2 and last 2 months of the year

Plot4: **HolidayVsCount** - We can say that the count is more on a no holiday day

Plot5: **WeekdayVsCount** - Can't say much on this trend, but we can say that count is relatively more on Wed and Sat

Plot6: **WheatherVsCount** - The count is more during condition '1'

Plot7: **WorkingdayVsCount** - not giving much clear picture

**Question 2.** Why is it important to use drop\_first=True during dummy variable creation?


**Answer:** Main reason behind dropping the column is the problem arising due to the **correlation** between the independent variables (features).

If there is collinearity between the dummy variables it leads to **multicollinearity** which violates the postulates/assumptions of **linear regression model**.

It just eliminates one extra column (redundant column) created while dummy variable creation.

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

**Answer:** By looking at the **pair-plot** among numerical variables, **temp** and **atemp** are the two variables which are highly correlated.



PS: Please refer the pair-plot from the python notebook

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

**Answer:** **Residual distribution** should follow normal distribution with the central mean = 0

**Displot** plays important role in analysing the residual distribution.

With the graph present in the python notebook we can conclude that the residual errors are normally distributed (bell shaped curve)

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

**Answer:** According to my prediction of the model following are the top 3 features significantly contributing towards the demand of the shared bikes.

a) **Temp**

b) **Weathersit (Good&Clear)**

c) **Year**

## General Subjective Questions

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


**Answer:** Linear regression is a statistical regression method used for predictive analysis and observe the relationship between the continuous variables. It shows the relationship between the **independent variable** (X) also known as predictor variable and the **dependent variable** (Y-axis) also called target variable.

Depending on the number of input **variables** or predictor variables, there are two types:

**Simple linear regression:** When the number of independent variables is 1

**Multiple linear regression:** When the number of independent variables is more than 1

Linear regression estimates the relationship between a dependent variable and an independent variable using the **line equation**:



$y = mx + c$  where  $Y$  is the **target variable**,  $x$  is the **input variable**,  $m$  and  $c$  are **slope** and **y intercept** respectively

The above equation could be also written as  $y = b_1x + b_0$ , which implies that for every rise/fall in the value of  $x$ , the  $y$  is rising or falling  $b_1$  times provided  $b_0$  is kept constant.

A regression line can be a **Positive** Linear Relationship or a **Negative** Linear Relationship based on the coefficient values

The main aim of the linear regression algorithm is to get the best values for  $b_0$  and  $b_1$  to find the best fit line for the model. The best fit line should have the least error means the error between predicted values and actual values should be minimized. It can be done using Cost function.

**Cost function** adjusts the regression coefficients and measures how a linear regression model is acting. The cost function is used to find the correctness of the mapping function that maps the predictor variable to the target variable. This mapping function is also known as the Hypothesis function.

Following are the steps to be followed in linear regression algorithm

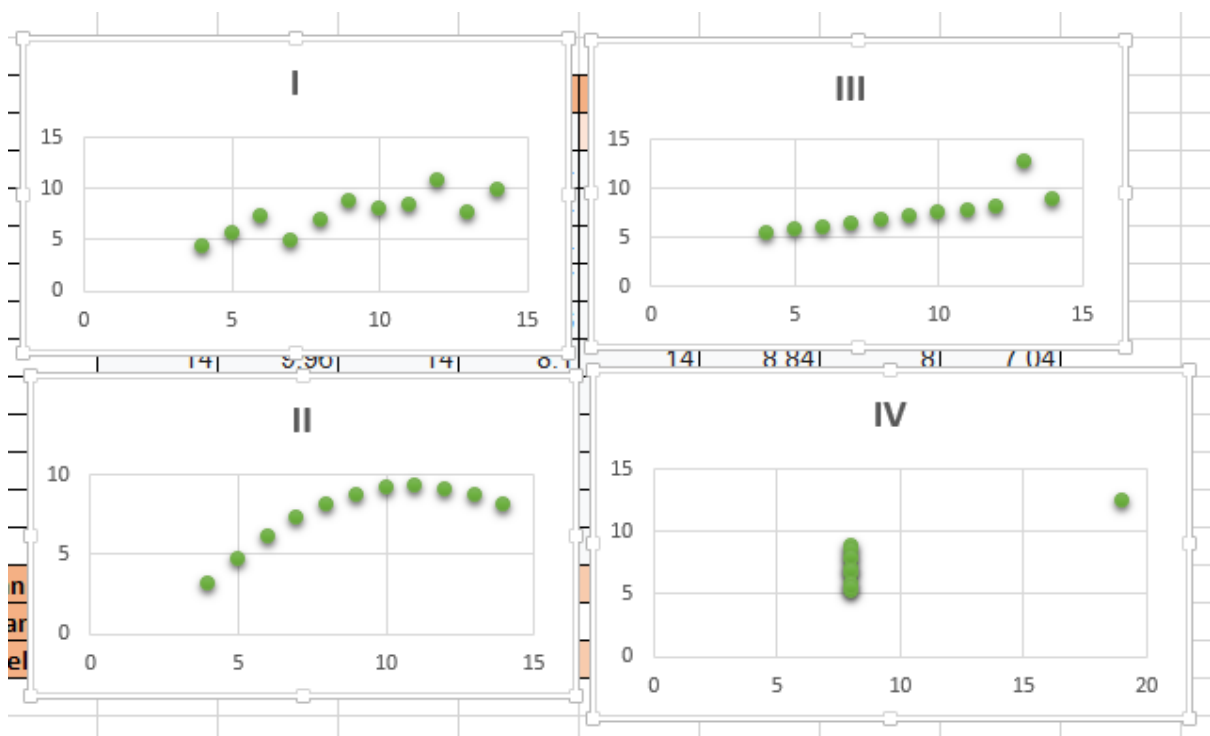
- 1. Reading and understanding the data:** Importing required libraries like pandas & numpy for data analysis and manipulation and seaborn & matplotlib for data visualization
- 2. Performing EDA on the data:** Visualization of data - Visualizing numerical variables using scatter or pair plots and for categorical variables using bar plots or boxplots in order to interpret the inferences.
- 3. Data preparation:** modification - Converting categorical variables with varying degrees of levels into dummy variables which are numerical in nature
- 4. Splitting the data into training and test sets either 70-30 or 80-20 :** Splitting the data into 2 sections in order to train a subset of given dataset to generate a trained model that will very well generalize how test data will be evaluated,
- 5. Build a linear model with help of library functions:** We add all the variables at once and then eliminate variables based on high multicollinearity ( $VIF > 5$ ) or insignificance (high p-values).
- 6. Residual analysis of the train data, error terms are analysed:** It tells us how much the errors ( $y_{\text{actual}} - y_{\text{pred}}$ ) are distributed across the model. A good residual analysis will signify that the mean is centred on 0.


**7. Making predictions using the final model and evaluation:** We will predict the test dataset by transforming it onto the trained dataset

**Question 2.** Explain the Anscombe's quartet in detail.

**Answer:** Anscombe's quartet contains 4 datasets that have almost identical simple statistical properties, but appear different when they are plotted.

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8.04	10	9.14	10	7.46	8	6.58
	8	6.95	8	8.14	8	6.77	8	5.76
	13	7.58	13	8.74	13	12.74	8	7.71
	9	8.81	9	8.77	9	7.11	8	8.84
	11	8.33	11	9.26	11	7.81	8	8.47
	14	9.96	14	8.1	14	8.84	8	7.04
	6	7.24	6	6.13	6	6.08	8	5.25
	4	4.26	4	3.1	4	5.39	19	12.5
	12	10.84	12	9.13	12	8.15	8	5.56
	7	4.82	7	7.26	7	6.42	8	7.91
	5	5.68	5	4.74	5	5.73	8	6.89
Mean	9	7.50090	9	7.50090	9	7.5	9	7.50090
Variance	11	4.12726	11	4.12762	11	4.1226	11	4.12324
Correlation	0.81642		0.81623		0.81628		0.81652	
n	1		7		7		1	





Datasets which are indistinguishable over a number of statistical properties, still produce different graphs, are often used to illustrate the importance of graphical representations when exploring data.

Sometimes statistics summary of the data are misleading on their own. So it's important to use graphical or visualization of data for larger data analysis process. Visualizing our data allows us to revisit our summary statistics and re-contextualize them as needed.

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

**Answer:** The Pearson's R referred to as Pearson's Correlation Coefficient in statistics. It is a statistic that measures the **linear correlation** between two variables. Like all correlations, it also has a numerical value that lies between - **1.0 and +1.0**.

However, it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations

**Question 4.** What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

**Answer:** scaling is a step of data processing in data analyse using models, it is applied to independent variables to normalize the data within a particular range. It also helps in faster calculations in an algorithm.

The data collected often contains features/variables which are highly varying in magnitudes or units or 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**.

Normalized scaling	Standardize scaling
It brings all of the data in the range of 0 and 1.	It brings all of the data into a standard normal distribution which has mean zero and standard deviation one
Uses <b>MinMaxScaler</b> from sklearn sklearn.preprocessing.MinMaxScaler	Uses <b>scale</b> from sklearn sklearn.preprocessing.scale
$x = \frac{x - \min(x)}{\max(x) - \min(x)}$	$x = \frac{x - \text{mean}(x)}{sd(x)}$
it loses some information in the data, if there are outliers in the dataset	it retains the information in the data, if there are outliers in the dataset

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

**Answer:** VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. A large value of VIF indicates that there is a correlation between the variables.

If the VIF is infinite which implies that there is a perfect correlation between two predictor variables. It is the case of perfect correlation.

In this case we get **Rsquared** value equal to 1.


Due to which the term **1/ (1-Rsquared)** reached infinity.

For this we need to identify the feature which is causing this perfect correlation and should be dropped to get a best model

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables

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

**Answer:** Q-Q Plots are Quantile-Quantile plots. These are the plots of two quantiles against each other. They plot the quantiles of a sample distribution against quantiles of a theoretical distribution.



It helps us to decide if a dataset follows any specific type of probability distribution such as normal, uniform or exponential.

If two populations are of the same distribution we can use Q-Q plot to summarise the distribution.

We can check if the residuals follow normal distribution in linear regression model.