Bike Sharing (Linear Regression) Assignment (DS-C52 - Dhara Khamar)

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? (3 marks)

Ans:

The final Multiple Linear Regression model contains many predictor variables that are categorical in nature and some of them have been encoded to dummy variables

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2317	0.027	8.607	0.000	0.179	0.285
yr	0.2286	0.008	28.154	0.000	0.213	0.245
holiday	-0.0958	0.026	-3.741	0.000	-0.146	-0.045
temp	0.5395	0.022	24.503	0.000	0.496	0.583
hum	-0.1759	0.037	-4.694	0.000	-0.250	-0.102
windspeed	-0.1835	0.026	-7.151	0.000	-0.234	-0.133
summer	0.1030	0.011	9.441	0.000	0.082	0.124
winter	0.1482	0.011	13.942	0.000	0.127	0.169
weathersit_2	-0.0544	0.011	-5.167	0.000	-0.075	-0.034
weathersit_3	-0.2351	0.026	-8.932	0.000	-0.287	-0.183
month_8	0.0553	0.016	3.393	0.001	0.023	0.087
month_9	0.1222	0.016	7.542	0.000	0.090	0.154

- Summer and winter are dummy variables for season category.
- weathersit_2 and weathersit_3 fare dummy variables for weathersit category.
- month_* are dummy variables for month category.

We can infer from above image that these variables are statistically significant and explain the variance in model very well.

---

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

Ans:

To encode categorical data, one hot encoding is done, where a dummy variable is to be created for each discrete categorical variable for a feature. This can be done by using pandas.get_dummies() which will return dummy-coded data.

Here we use 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 do not use drop_first = True, then n dummy variables will be created, and these predictors(n dummy variables) are themselves correlated which is known as multicollinearity and it, in turn, leads to Dummy Variable Trap.

---

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans:

From the pair plot we can say that **atemp** is having highest correlation with target variable **cnt** which is followed by **temp**. As per the correlation heatmap, correlation coefficient between **atemp** and **cnt** is **0.65**. And correlation coefficient between **temp** and **cnt** is **0.64.**

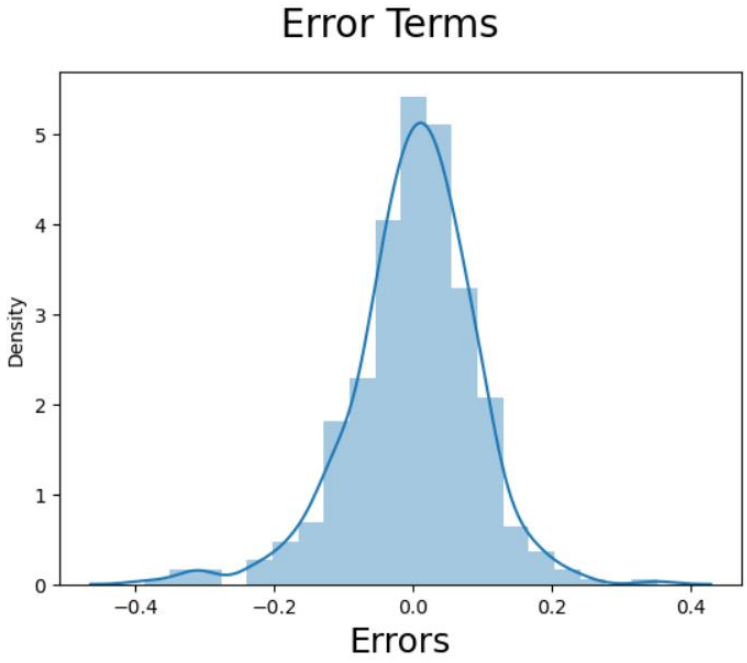
---

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans:

To validate assumptions of the model, and hence the reliability for inference, we go with the following procedures:

• Residual Analysis: We need to check if the error terms are also normally distributed (which is in fact, one of the major assumptions of linear regression). I have plotted the histogram of the error terms and this is what it looks like: The residuals are following the normally distribution with a mean 0.



• Linear relationship between predictor variables and target variable: This is happening because all the predictor variables are statistically significant (p-values are less than 0.05). Also, **R-Squared** value on training set is **0.841** and **adjusted R-Squared** value on training set is **0.838**. This means that variance in data is being explained by all these predictor variable

• Error terms are independent of each other: Handled properly in the model. The predictor variables are independent of each other. Multicollinearity issue is not there because VIF (Variance Inflation Factor) for all predictor variables are below 5.

---

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Ans:

Top 3 features significantly contributing towards demand of shared bikes are:

1) temp (coef: 0.5308)

2) yr (coef: 0.2288)

3) Weather Situation 3 (weathersit_3) (coef: -0.2351)

General Subjective Questions

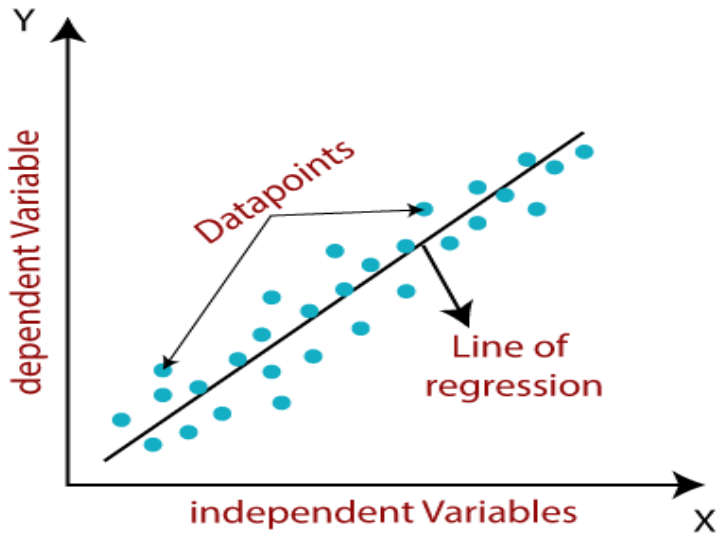
1. Explain the linear regression algorithm in detail. (4 marks)

Ans:

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as **sales, salary, age, product price,** etc.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

The linear regression model provides a sloped straight line representing the relationship between the variables. Consider the below image:



Mathematically, we can represent a linear regression as:

$$y = a_0 + a_1x + \varepsilon$$

Y = Dependent Variable (Target Variable)
X = Independent Variable (predictor Variable)
a0 = intercept of the line (Gives an additional degree of freedom)

a1 = Linear regression coefficient (scale factor to each input value).
ε = random error

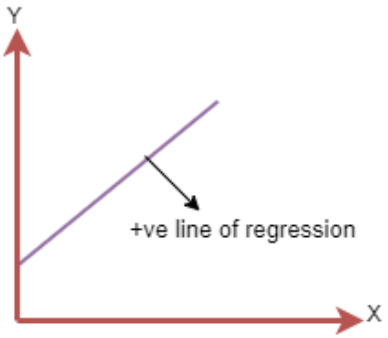
Types of Linear Regression

- **Simple Linear Regression:**
  If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.
- **Multiple Linear regression:**
  If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

Linear Regression Line

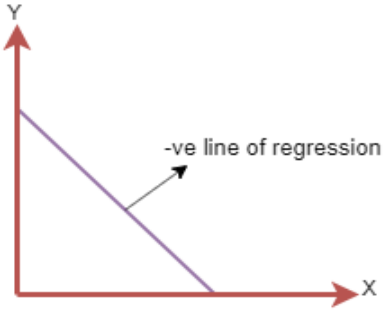
A linear line showing the relationship between the dependent and independent variables is called a **regression line**. A regression line can show two types of relationship:

- **Positive Linear Relationship:**
  If the dependent variable increases on the Y-axis and 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 independent variable increases on the X-axis, then such a relationship is called a negative linear relationship.



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

2. Explain the Anscombe's quartet in detail. (3 marks)

Ans:

Anscombe's Quartet is the modal example to demonstrate the importance of data visualization which was developed by the statistician Francis Anscombe in 1973 to signify both the importance of plotting data before analysing it with statistical properties.
It comprises of four data-set and each data-set consists of eleven (x,y) points. The basic thing to analyse about these data-sets is that they all share the same descriptive statistics(mean, variance, standard deviation etc) but different graphical representation.

Each graph plot shows the different behaviour irrespective of statistical analysis.

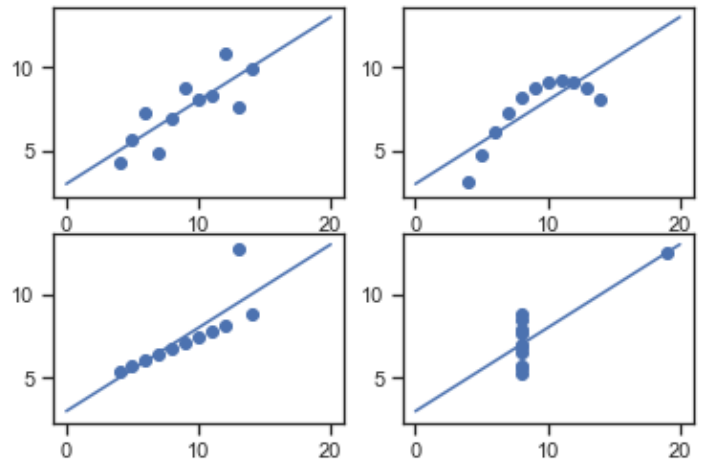
x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7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

Four Data-sets

Apply the statistical formula on the above data-set,

Average Value of x = 9
Average Value of y = 7.50
Variance of x = 11
Variance of y =4.12
Correlation Coefficient = 0.816
Linear Regression Equation : y = 0.5 x + 3

However, the statistical analysis of these four data-sets are pretty much similar. But when we plot these four data-sets across the x & y coordinate plane, we get the following results & each pictorial view represent the different behaviour.



Graphical Representation of Anscombe's Quartet

Data-set I — consists of a set of (x,y) points that represent a linear relationship with some variance.

Data-set II — shows a curve shape but doesn't show a linear relationship (might be quadratic?).

Data-set III — looks like a tight linear relationship between x and y, except for one large outlier.

Data-set IV — looks like the value of x remains constant, except for one outlier as well

---

3. What is Pearson's R? (3 marks)

Ans:

The Pearson correlation coefficient represents the relationship between the two variables, measured on the same interval or ratio scale. It measures the strength of the relationship between the two continuous variables.

The coefficient not only states the presence or absence of the correlation between the two variables but also determines the exact extent to which those variables are correlated.

It is independent of the unit of measurement of the variables where the values of the correlation coefficient can range from the value +1 to the value -1. However, it is insufficient to tell the difference between the dependent and independent variables.

It is independent of the unit of measurement of the variables. For example, suppose the unit of measurement of one variable is in years while the unit of measurement of the second variable is in kilograms. In that case, even then, the value of this coefficient does not change.

The correlation coefficient between the variables is symmetric, which means that the value of the correlation coefficient between Y and X or X and Y will remain the same.

Formula

The Pearson Correlation Coefficient formula is as follows:

$$\mathbf{r} = \frac{\mathbf{n}(\sum \mathbf{xy}) - (\sum \mathbf{x})(\sum \mathbf{y})}{\sqrt{[\mathbf{n}\sum \mathbf{x^2} - (\sum \mathbf{x})^2][\mathbf{n}\sum \mathbf{y^2} - (\sum \mathbf{y})^2]}}$$

Where,

r = Pearson Coefficient

n= number of pairs of the stock

∑xy = sum of products of the paired stocks

∑x = sum of the x scores

∑y= sum of the y scores

∑x2 = sum of the squared x scores

∑y2 = sum of the squared y scores

The steps for Pearson correlation coefficient calculation are as follows:

- Find out the number of pairs of variables denoted by n. Suppose x consists of 3 variables – 6, 8, 10. Suppose y consists of corresponding three variables: 12, 10, and 20.
- List down the variables in two columns.

x	y
6	12
8	10
10	20

- Find out the product of x and y in the 3rd column.

x	y	x*y
6	12	72
8	10	80
10	20	200

- Find the sum of values of all x and y variables. Write the results at the bottom of the 1st and 2nd columns. Then, write the sum of x*y in the 3rd column.

x	y	x*y
6	12	72
8	10	80
10	20	200
24	**42**	**352**

- Find out x2 and y2 in the 4th and 5th columns and their sum at the bottom of the columns.

x	y	x*y	x^2	y^2
6	12	72	36	144
8	10	80	64	100
10	20	200	100	400
24	**42**	**352**	**200**	**644**

- Insert the values found above in the formula and solve it.

  r = 3*352-24*42 / √(3*200-24^2)*(3*644-42^2) = 0.7559

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans:

Scaling: It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Why scaling ? 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/Min-Max Scaling:

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 Scaling:

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).

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

Normalization	Standardization
Rescales values to a range between 0 and 1	Centres data around the mean and scales to a standard deviation of 1
Useful when the distribution of the data is unknown or not Gaussian	Useful when the distribution of the data is Gaussian or unknown
Sensitive to outliers	Less sensitive to outliers
Retains the shape of the original distribution	Changes the shape of the original distribution
May not preserve the relationships between the data points	Preserves the relationships between the data points
Equation: (x – min)/(max – min)	Equation: (x – mean)/standard deviation

However, at the end of the day, the choice of using normalization or standardization will depend on your problem and the machine learning algorithm you are using. There is no hard and fast rule to tell you when to normalize or standardize your data. You can always start by fitting your model to raw, normalized, and standardized data and comparing the performance for the best results.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Ans:

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.

The formula for VIF is:

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

Where, Ri2 = Unadjusted coefficient of determination forregressing the ith independent variable on the remaining ones

- VIF equal to 1 = variables are not correlated
- VIF between 1 and 5 = variables are moderately correlated
- VIF greater than 5 = variables are highly correlated

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity.

To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well)

---

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

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.

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.

Few advantages:

a) It can be used with sample sizes also

b) 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 behaviour

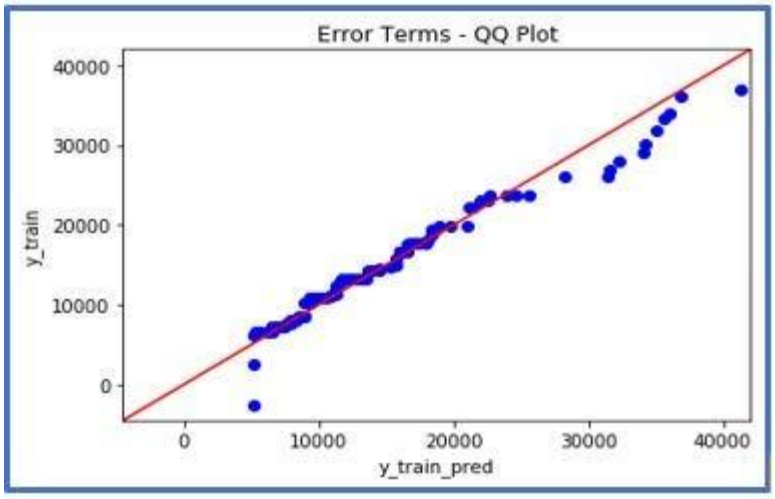
Interpretation:

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

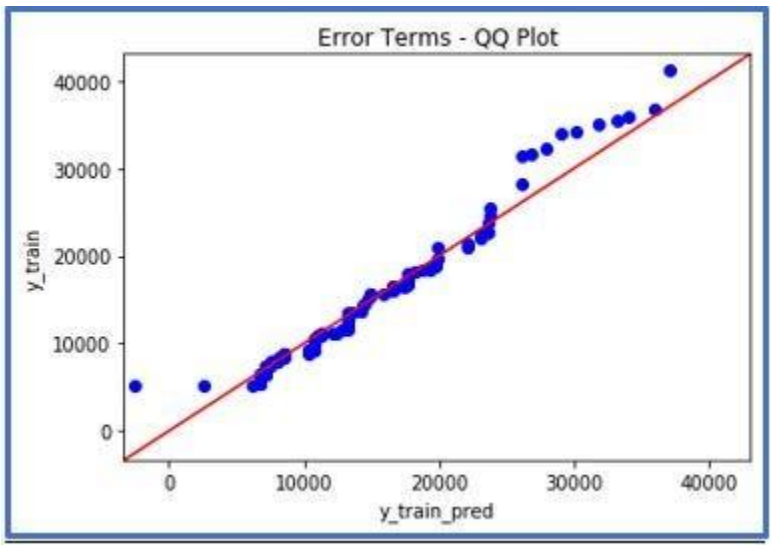
Below are the possible interpretations for two data sets.

- Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

- Y-values < X-values: If y-quantiles are lower than the x-quantiles.



- X-values < Y-values: If x-quantiles are lower than the y-quantiles.



- Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis