Linear Regression

Subjective Questions & Answers

By

Renuka Mokka

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

Inferences from charts of categorical variables vs target carriable:

1. Season -3: fall has highest demand for rental bikes

2. I see that demand for year 2019 has grown

3. Demand is continuously growing each month till June. September month has highest demand. After September, demand is decreasing

4. When there is a holiday, demand has decreased.

5. Weekday is not giving clear picture about demand.

6. The clear weather_situation has highest demand

7. During September, bike sharing is more. During the year end and beginning, it is less, could be due to extreme weather conditions

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

- Dummy variable creation is converting categorical column values into binary (0 and 1)
- Ex: Season has 4 values: Autumn, Spring, Summer, Winter
- Dummy variable splits it into 4 columns having labels Season_Autumn, Season_Spring, Season_Summer, Season_Winter
- The values for those columns will be 1 0 0 0,0 1 0 0,0 0 1 0,0 0 0 1
- We can explain it with the help of just 3 columns(p-1) as if all 3 0's means it's Autumn, otherwise one of those 3 seasons.
- Hence, we can drop first column using drop_first=True , while creating dummy variables to reduce size of dataset without loosing any insights.

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

- Looks like temp and atemp has highest correlation with the target variable.
- Temp and atemp are highly correlated with each other.

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

By doing Residual anyalysis:

1. Linearity check by plotting y_train vs residual
2. By plotting distplot for error term(y_train – y_train_pred), observed that error curve is normally distributed with mean 0.
3. By plotting error term vs index from 1 to length (X_train), it’s clear that error term does not show any clear trend, i.e. error terms are independent of each other.
4. By plotting y_test and y_test_pred to understand the spread

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

1. Temp -> directly related to demand
2. Year(2019) ->has seen steady growth in demand
3. Weather situation bad ->Inversely related to demand
4. Windspeed also inversely related to demand

General Subjective Questions

1.Explain the linear regression algorithm in detail.

- Linear regression is a quiet and simple statistical regression method used for predictive analysis and shows the relationship between the continuous variables.
- The linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), consequently called linear regression.
- *If there is a single input variable (x), such linear regression is called **simple linear regression**. And if there is more than one input variable, such linear regression is called **multiple linear regression.*** The linear regression model gives a sloped straight line describing the relationship within the variables.
- Simple linear regression can be defined as y = a0+a1X+E

  y=Target/dependent variable

  X=independent/predictor variable

  a0=intercept of the line

  a1=coefficient

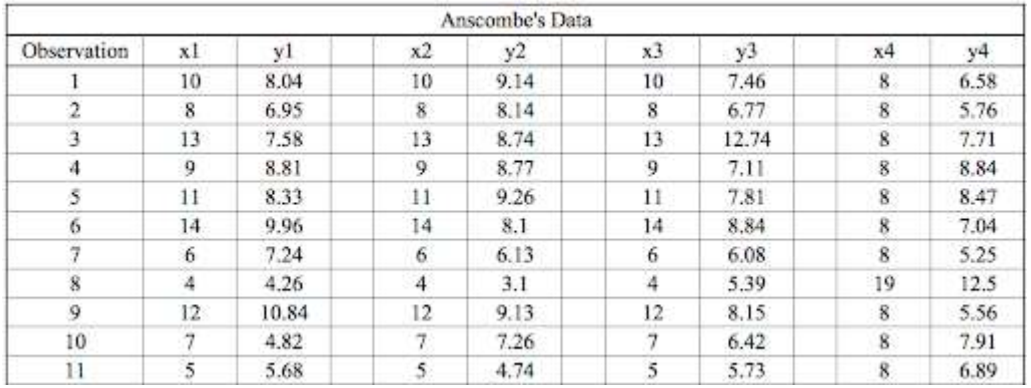
  E=random error

1.Explain the linear regression algorithm in detail (contd.)

- When working with linear regression, our main goal is to find the best fit line that means the error between predicted values and actual values should be less.
- **Residuals:** The distance between the actual value and predicted values is called residual. If the observed points are far from the regression line, then the residual will be high, and so cost function will high. If the scatter points are close to the regression line, then the residual will be small and hence the cost function. This should be minimized. The best fit line will have the least error.

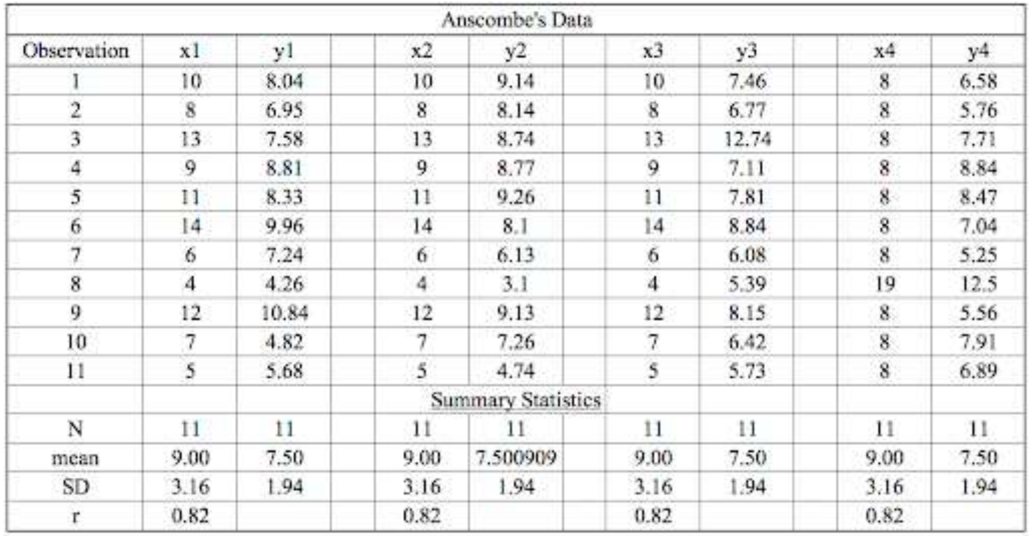
2.Explain the Anscombe's Quartet in detail:

- Anscombe's quartet is a group of datasets (x, y) that have the same mean, standard deviation, and regression line, but which are qualitatively different.
- It is often used to illustrate the importance of looking at a set of data graphically and not only relying on basic statistic properties.
- **Anscombe's quartet** is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.



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



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	

Fig:3 Statistical data of 4 datasets which looked very similar

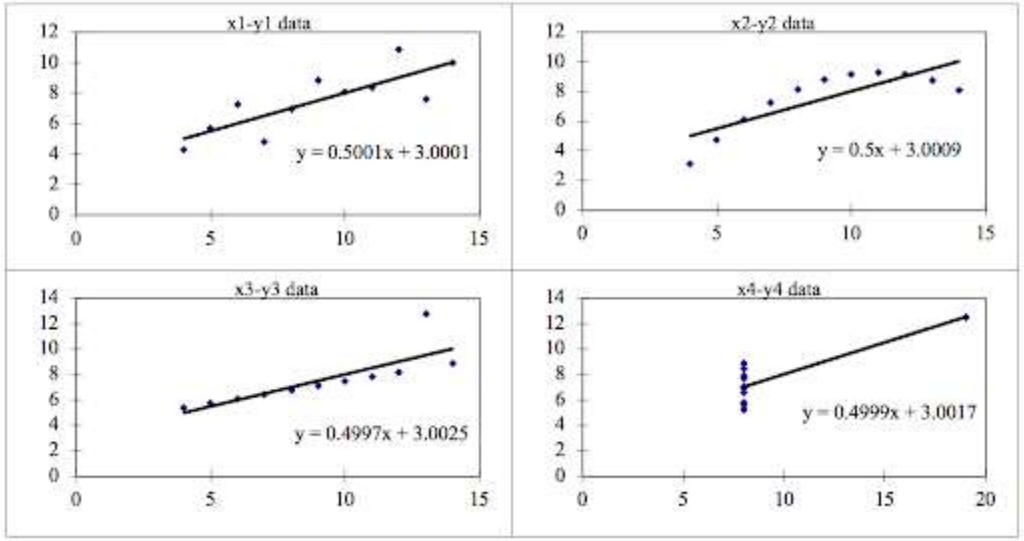


Fig:4 Scatter plots of 4 datasets

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 you 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?

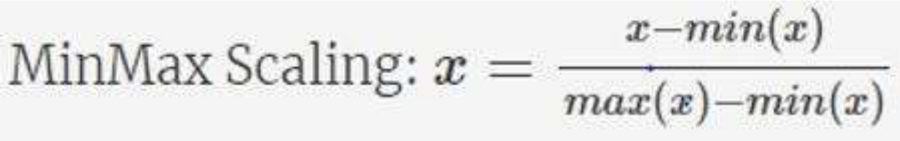
The **Pearson correlation coefficient (*r*)** is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

Pearson correlation coefficient (r)	Correlation type	Interpretation	Example
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the **same direction**.	Baby length & weight: The longer the baby, the heavier their weight.
0	No correlation	There is **no relationship** between the variables.	Car price & width of windshield wipers: The price of a car is not related to the width of its windshield wipers.
Between 0 and –1	Negative correlation	When one variable changes, the other variable changes in the **opposite direction**.	Elevation & air pressure: The higher the elevation, the lower the air pressure.

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

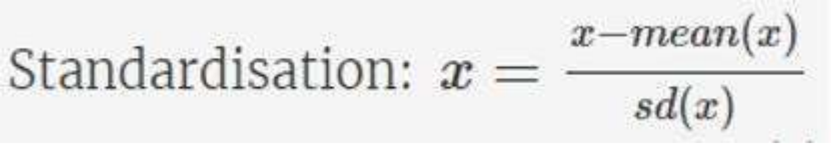
- Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.
- Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done, then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we must 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. **sklearn.preprocessing.MinMaxScaler** helps to implement normalization in python.



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

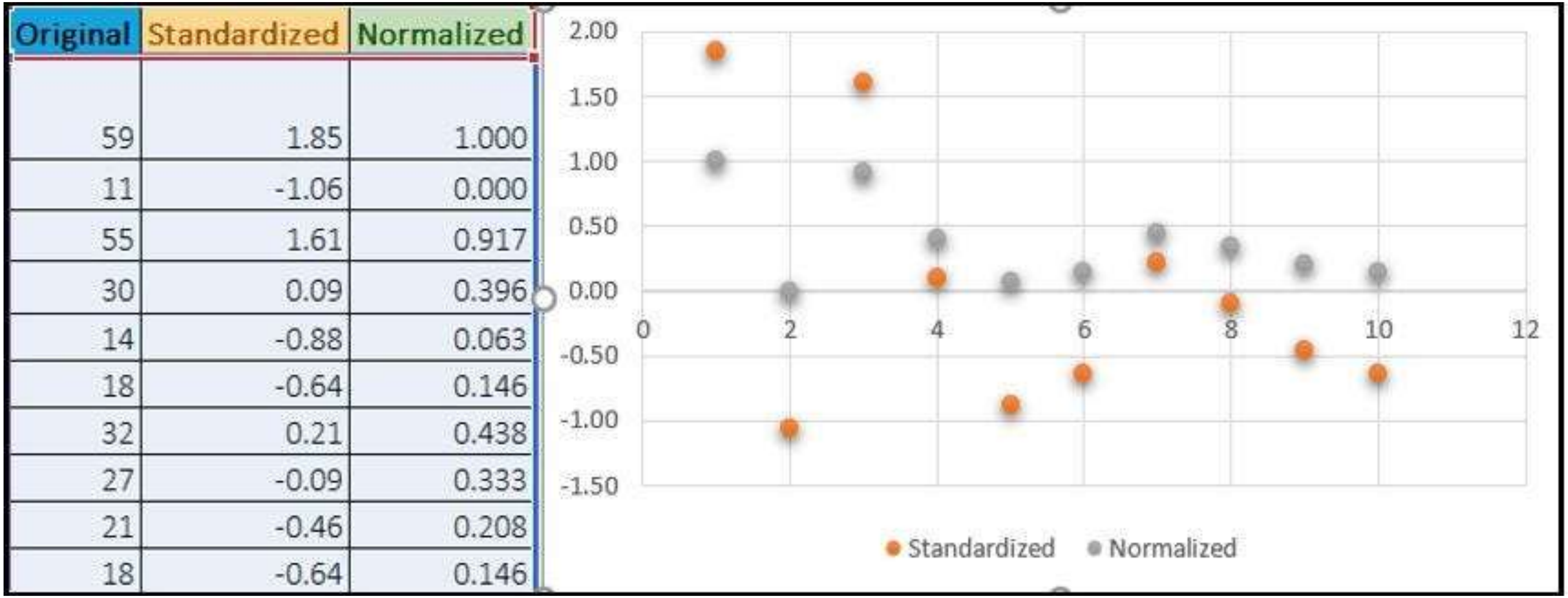
- **Standardization Scaling:**
- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).



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

- sklearn.preprocessing.scaler helps to implement standardization in python.
- One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

Normalized vs Standardized Scaling



Original	Standardized	Normalized
59	1.85	1.000
11	-1.06	0.000
55	1.61	0.917
30	0.09	0.396
14	-0.88	0.063
18	-0.64	0.146
32	0.21	0.438
27	-0.09	0.333
21	-0.46	0.208
18	-0.64	0.146

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

- VIF infinity 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.
- It happens when we have some columns which are highly correlated with each other as it causes multi-collinearity.
- To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

- Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.
- The power of Q-Q plots lies in their ability to summarize any distribution visually.
- QQ plots is very useful to determine

  If two populations are of the same distribution

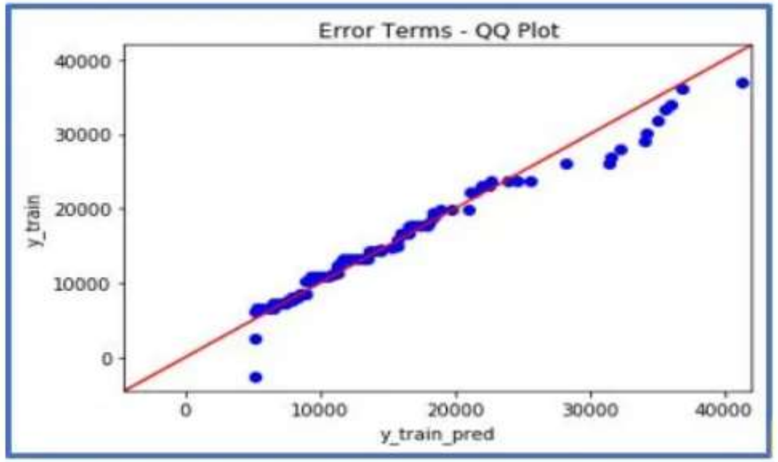
  If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it’s met using this.

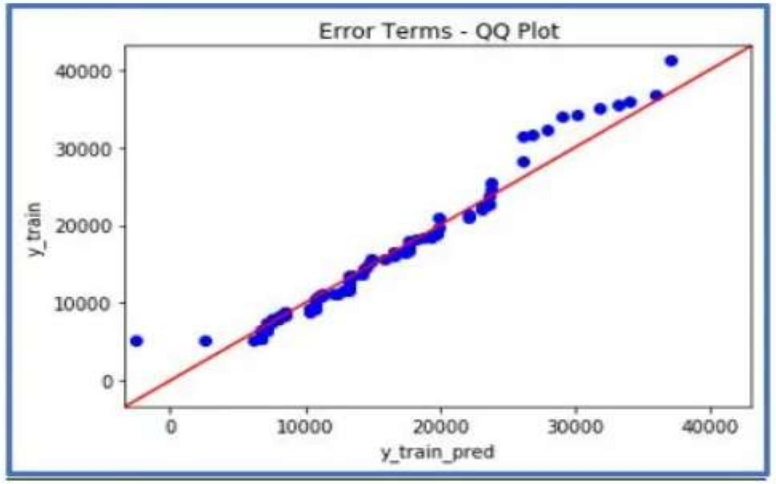
  Skewness of distribution

  In Q-Q plots, we plot the theoretical Quantile values with the sample Quantile values. Quantiles are obtained by sorting the data. It determines how many values in a distribution are above or below a certain limit.

Interpreting Q-Q plots:

- Below are the possible interpretations for two data sets.
  a) **Similar distribution**: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x –axis





b) **Y-values < X-values:** If y-quantiles are lower than X-quanitles

c) **X-values < Y-values:** If x-quantiles are lower than the Y-quantiles.

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