Name: Abhijeet Srivastava
Email: abhijeet.s92@outlook.com

# Linear Regression Case Study

## 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)

Bike Rentals are:

- more during the Fall season and followed by summer, then winter and lastly spring
- more in the 2019 compared to 2018
- more in May through Oct and lower in the remaining months
- the practically the same when compared with workingday
- more in Clear weather, followed by mist+cloudy
- almost same during the weekdays: highest on Monday, Thursday and Friday
- less on holidays

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

Drop_first = True is important during dummy variable creation as it firstly reduces the correlation among the dummy variables and also reduces the number of columns in the data frame.

In the assignment for example, season variable has Fall, Spring, Summer, and Winter has 4 categories. Since there are four variables, if 3 of them are already defined and are false then it implies that the remaining one has to be true. So, we use drop_first = True and get rid of the first column, reducing correlation and number of columns in the df.

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

The variable ‘temp’ has the highest correlation with target variable ‘cnt’

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

We validate the assumptions of Linear Regression after building the model on the training set by:

a. checking the p-value and VIF, low p-value and low VIF tells us that the variables are independent of each other (no multicollinearity as well)

b. running a residual analysis on the error terms to check that they are normally distributed around the mean

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

For the final model, Equation for the best fit line:

$$count=0.2325\times yr-0.1008\times holiday+0.5034\times temp-0.0764\times Spring+0.0355\times Summer+0.0842\times Winter-0.0527\times July+0.0810\times Sep-0.2999\times LightSnow-0.0798\times MistCloudy$$

The top 3 features that explain the demand are:

Name: Abhijeet Srivastava
Email: abhijeet.s92@outlook.com

i. ‘temp’ for every 1-unit increase in year, there is 0.5034 increase in demand

ii. ‘yr’ every 1-unit increase causes a 0.2325 increase in demand

iii. ‘Light Snow every 1-unit increase causes a 0.2999 decrease in demand

## General Subjective Questions

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

Linear Regression is an algorithm in machine learning based supervised learning. It predicts values for a target variable based on independent variable(s). In this form of ML, the independent variables

- need to have a linear relationship with the target
- need to have the same variance (homoscedasticity)
- need to be normally distributed with target variable

Mathematically, y = b0 + b1x1 + b2x2 + …. + bixi
Where b0 is the y intercept and all other b’s are the corresponding slope values for x’s
This algorithm shows how the independent variables change the target variable value

LR is mainly used for:

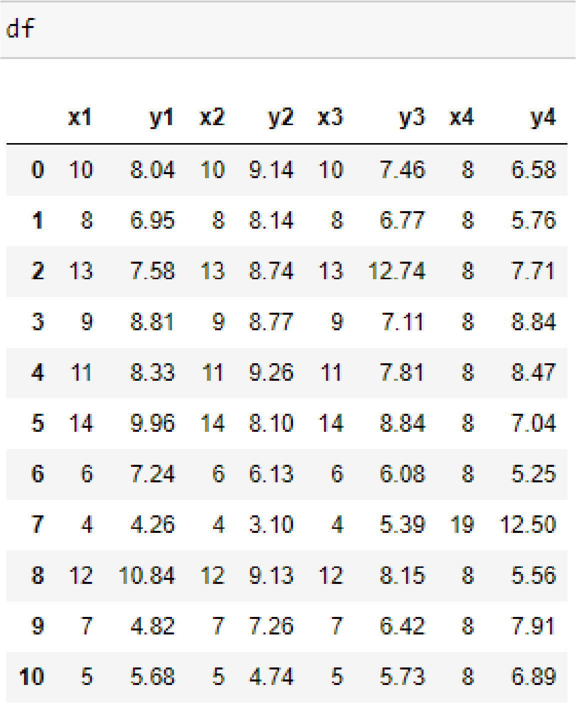
- prediction of trends and sales targets
- price prediction
- risk management

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

Francis Anscombe used 4 data sets that have nearly identical descriptive statistics and very different distributions to show the importance of graphing data before analysing it and the effect of outliers on statistical properties. On the right there are the 4 datasets stored in a pandas data frame

For all the 4 sets, (see below)

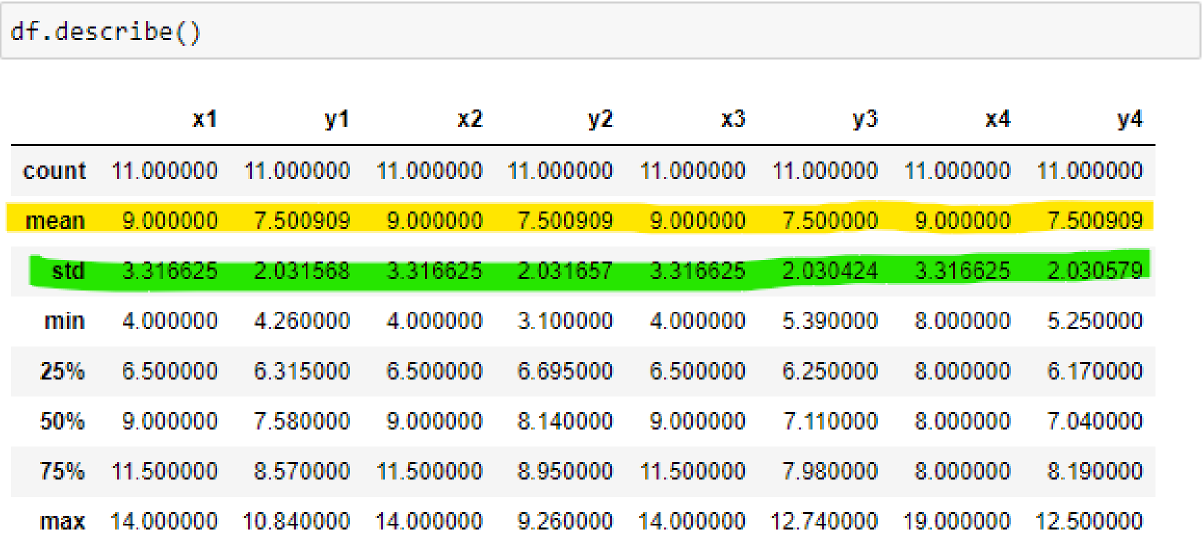
- x
    - ○ mean = 9
    - ○ variance = 11
- y
    - ○ mean = 7.5
    - ○ variance = 4.125
- correlation = 8.16
- Regression line y = 3+ 0.5x
- R-squared = 0.67



```
df
```

| | x1 | y1 | x2 | y2 | x3 | y3 | x4 | y4 |
|---|---|---|---|---|---|---|---|---|
| **0** | 10 | 8.04 | 10 | 9.14 | 10 | 7.46 | 8 | 6.58 |
| **1** | 8 | 6.95 | 8 | 8.14 | 8 | 6.77 | 8 | 5.76 |
| **2** | 13 | 7.58 | 13 | 8.74 | 13 | 12.74 | 8 | 7.71 |
| **3** | 9 | 8.81 | 9 | 8.77 | 9 | 7.11 | 8 | 8.84 |
| **4** | 11 | 8.33 | 11 | 9.26 | 11 | 7.81 | 8 | 8.47 |
| **5** | 14 | 9.96 | 14 | 8.10 | 14 | 8.84 | 8 | 7.04 |
| **6** | 6 | 7.24 | 6 | 6.13 | 6 | 6.08 | 8 | 5.25 |
| **7** | 4 | 4.26 | 4 | 3.10 | 4 | 5.39 | 19 | 12.50 |
| **8** | 12 | 10.84 | 12 | 9.13 | 12 | 8.15 | 8 | 5.56 |
| **9** | 7 | 4.82 | 7 | 7.26 | 7 | 6.42 | 8 | 7.91 |
| **10** | 5 | 5.68 | 5 | 4.74 | 5 | 5.73 | 8 | 6.89 |

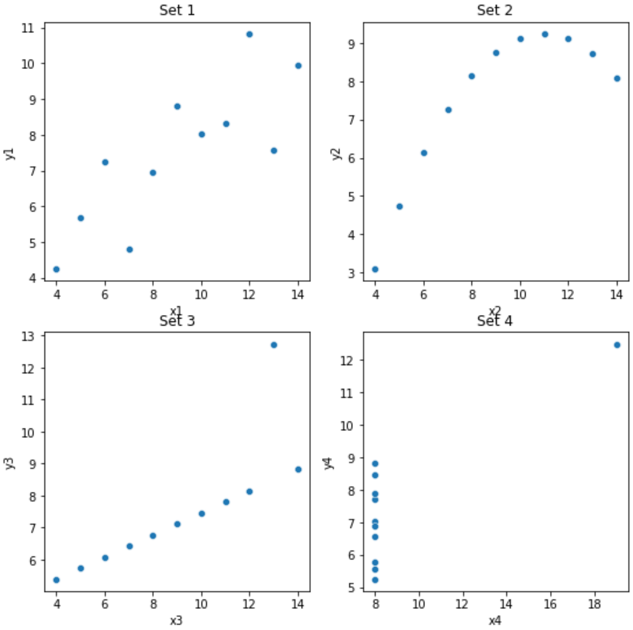
Name: Abhijeet Srivastava
Email: abhijeet.s92@outlook.com

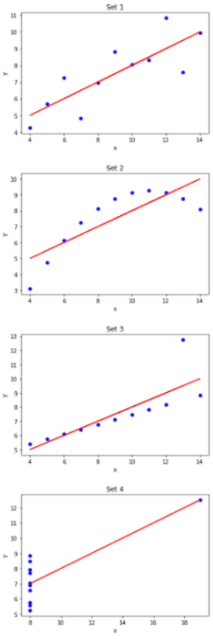


```
df.describe()
```

| | x1 | y1 | x2 | y2 | x3 | y3 | x4 | y4 |
|---|---|---|---|---|---|---|---|---|
| count | 11.000000 | 11.000000 | 11.000000 | 11.000000 | 11.000000 | 11.000000 | 11.000000 | 11.000000 |
| mean | 9.000000 | 7.500909 | 9.000000 | 7.500909 | 9.000000 | 7.500000 | 9.000000 | 7.500909 |
| std | 3.316625 | 2.031568 | 3.316625 | 2.031657 | 3.316625 | 2.030424 | 3.316625 | 2.030579 |
| min | 4.000000 | 4.260000 | 4.000000 | 3.100000 | 4.000000 | 5.390000 | 8.000000 | 5.250000 |
| 25% | 6.500000 | 6.315000 | 6.500000 | 6.695000 | 6.500000 | 6.250000 | 8.000000 | 6.170000 |
| 50% | 9.000000 | 7.580000 | 9.000000 | 8.140000 | 9.000000 | 7.110000 | 8.000000 | 7.040000 |
| 75% | 11.500000 | 8.570000 | 11.500000 | 8.950000 | 11.500000 | 7.980000 | 8.000000 | 8.190000 |
| max | 14.000000 | 10.840000 | 14.000000 | 9.260000 | 14.000000 | 12.740000 | 19.000000 | 12.500000 |

But have different distributions:





- Set1 has a simple linear relationship,
- Set 2 has a non-linear relationship,
- Set 3 linear relationship but with outliers
- Set 4 outlier changing the correlation coefficient where data has no obvious relationship

3. What is Pearson’s R? (3 marks)

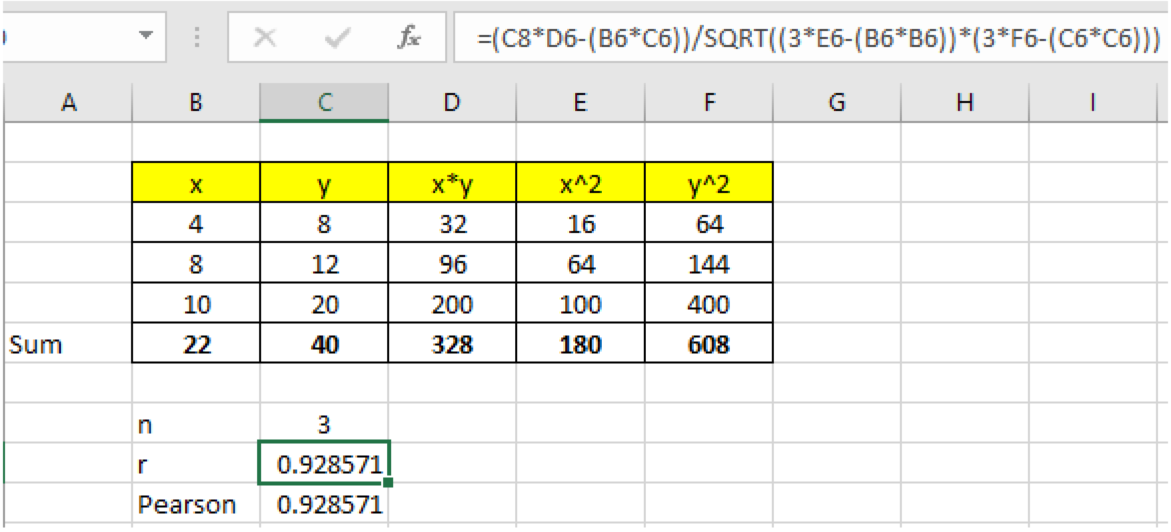
It measures the linear correlation between two variables and ranges from -1 to 1
It is the covariance of the 2 variables divided by the product of their SDs

$$r = \frac{N\sum xy - (\sum x)(\sum y)}{\sqrt{(N\sum x^2 - (\sum x)^2)(N\sum y^2 - (\sum y)^2}}$$

Where N is the number of pairs of scores

Name: Abhijeet Srivastava
Email: abhijeet.s92@outlook.com

∑xy is the sum of products of paired scores
∑x is sum of x scores
∑y is sum of y scores



=(C8*D6-(B6*C6))/SQRT((3*E6-(B6*B6))*(3*F6-(C6*C6)))

| A | B | C | D | E | F |
|---|---|---|---|---|---|
| | | | | | |
| | x | y | x*y | x^2 | y^2 |
| | 4 | 8 | 32 | 16 | 64 |
| | 8 | 12 | 96 | 64 | 144 |
| | 10 | 20 | 200 | 100 | 400 |
| Sum | **22** | **40** | **328** | **180** | **608** |
| | | | | | |
| | n | 3 | | | |
| | r | 0.928571 | | | |
| | Pearson | 0.928571 | | | |

Using the numbers and making the formula from scratch vs using the excel function

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

It is a step of the data processing part, applied to independent variables to normalize them within a particular range. Most of the time, the independent variables have different scales for example in the Bike sharing case study, the data has column variable in Celsius, humidity variable in %, windspeed in kmph. Modelling without scaling these variables will give us very different coefficients and might skew the best fit line.

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

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

In MinMax Scaling, the variable is scaled via taking the difference of it with the min value in column over the difference in the max and min value of the column

Standardisation mean and SD are used for scaling.

Differences:

| | |
|---|---|
| MinMax | Standardisation |
| Scale values between -1 to 1 or 0 to 1 | Not bounded |
| Heavily Affected by outliers | Not affected by outliers that much |
| Useful when distribution isn't known | Useful when distribution is known to be Normal/Gaussian |

Name: Abhijeet Srivastava
Email: abhijeet.s92@outlook.com

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

VIF = infinity indicates that there is correlation = 1 (I.e. perfect correlation) between the 2 independent variables. When there is perfect correlation, R-squared = 1, plugging in VIF formula VIF = 1/(1-R-sqaured) results in VIF = infinity.

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

A Q-Q plot (Quantile-Quantile plot) is a graphical tool where 2 quantiles are plotted against each other. Its purpose is to check if the two sets of data are from the same distribution or not and if the sets are from some theoretical distribution or not.
A Q-Q plot compares the shapes of the sets by showing the

- Location
- Scale
- Skewness

For Linear Regression, we can use Q-Q plots to confirm if both the train and test data sets (received separately) are from populations with the same distributions.