



International
Institute of Information
Technology Bangalore



Bike Sharing Assignment

Submitted By

ABHISHEK DAS

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

Ans: Categorical variables require special attention in regression analysis because, unlike dichotomous or continuous variables, they cannot be entered into the regression equation just as they are. Instead, they need to be recoded into a series of variables which can then be entered into the regression model.

If the factor has 2 classes then we can make dummy variable with 1 and 0 since it's a binary case.

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

Ans : drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables

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

Ans: atemp and temp

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

Ans: We can check below points to on model to validate assumptions:

The regression model is linear in parameters

The mean of the residuals is zero - Check the mean of the residuals. If it zero (or very close), then this assumption is held true for that model.

Homoscedasticity of residuals or equal variance

No autocorrelation of residuals.

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

Ans: 1. Holidays 2. Season 3. Weathersit

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression may be defined as the statistical model that analyzes the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

Mathematically the relationship can be represented with the help of following equation –

$$Y=mX+b$$

Here,

Y = dependent variable we are trying to predict.

X = independent variable we are using to make predictions.

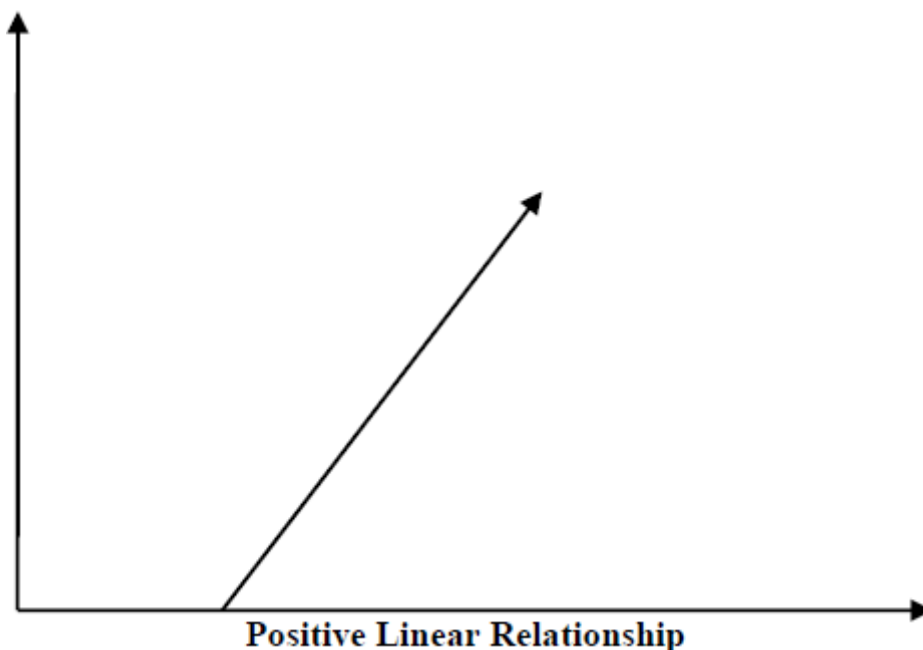
m = Slope of the regression line which represents the effect X has on Y

b = known as the Y -intercept. If $X = 0$, Y would be equal to b .

Furthermore, the linear relationship can be positive or negative in nature as explained below –

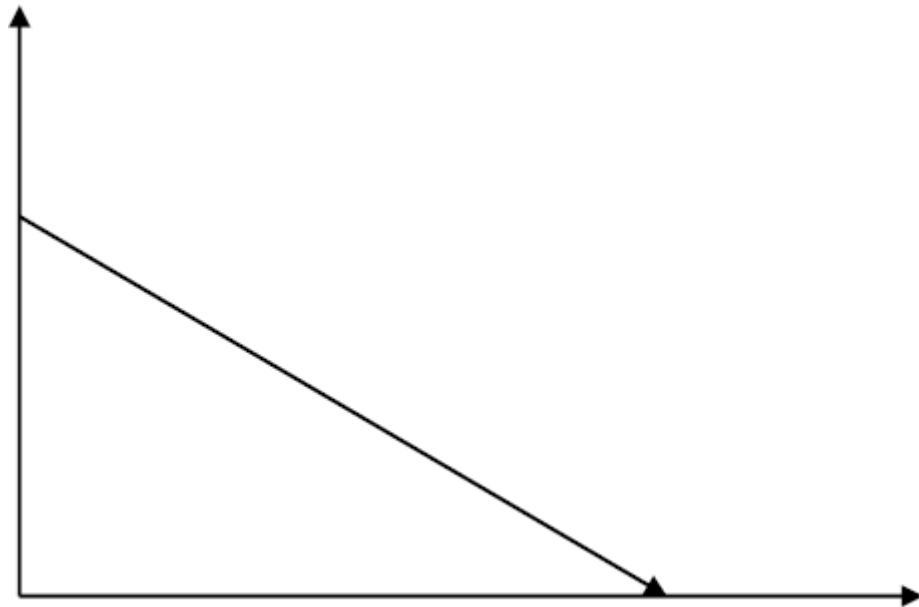
Positive Linear Relationship

A linear relationship will be called positive if both independent and dependent variable increases. It can be understood with the help of following graph.



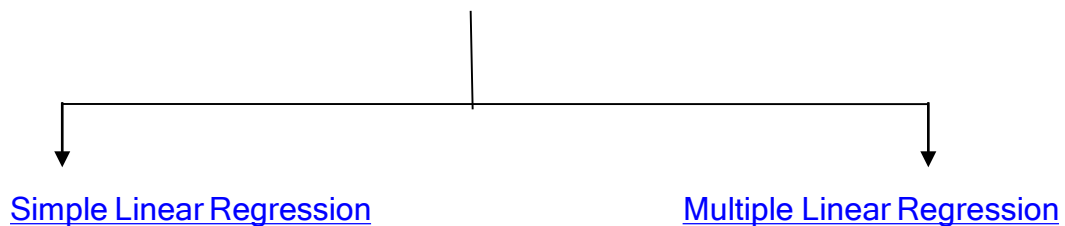
Negative Linear relationship

A linear relationship will be called positive if independent increases and dependent variable increases. It can be understood with the help of following graph -



Negative Linear Relationship

Types of Linear Regression



Simple linear regression is used to estimate the relationship between two quantitative variables:

We can use simple linear regression when you want to know:

1. How strong the relationship is between two variables (e.g. the relationship between rainfall and soil erosion).
2. The value of the dependent variable at a certain value of the independent variable (e.g. the amount of soil erosion at a certain level of rainfall).

Multiple linear regression is used to estimate the relationship between two or more independent variables and one dependent variable.

We can use multiple linear regression when you want to know:

1. How strong the relationship is between two or more independent variables and one dependent variable (e.g. how rainfall, temperature, and amount of fertilizer added affect crop growth).
2. The value of the dependent variable at a certain value of the independent variables (e.g. the expected yield of a crop at certain levels of rainfall, temperature, and fertilizer addition).

Assumptions

The following are some assumptions about dataset that is made by Linear Regression model –

Multi-co linearity – Linear regression model assumes that there is very little or no multi-co linearity in the data. Basically, multi-co linearity occurs when the independent variables or features have dependency in them.

Auto-correlation – Another assumption Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.

Relationship between variables – Linear regression model assumes that the relationship between response and feature variables must be linear.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet was constructed in 1973 by the [statistician Francis Anscombe](#) to demonstrate both the importance of graphing data before analyzing it and the effect of [outliers](#) and other [influential observations](#) on statistical properties.

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points.

Simple understanding:

Once Francis John "Frank" Anscombe who was a statistician of great repute found 4 sets of 11 data-points in his dream and requested the council as his last wish to plot those points. Those 4 sets of 11 data-points are given below.

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

After that, the council analyzed them using only descriptive statistics and found the mean, standard deviation, and correlation between x and y.

Code: Python program to find mean, standard deviation, and the correlation between x and y

```
# Import the required libraries
import pandas as pd
import statistics
from scipy.stats import pearsonr

# Import the csv file
df = pd.read_csv("anscombe.csv")

# Convert pandas dataframe into pandas series
list1 = df['x1']
list2 = df['y1']

# Calculating mean for x1
print('%0.1f' % statistics.mean(list1))

# Calculating standard deviation for x1
print('%0.2f' % statistics.stdev(list1))

# Calculating mean for y1
print('%0.1f' % statistics.mean(list2))

# Calculating standard deviation for y1
print('%0.2f' % statistics.stdev(list2))

# Calculating pearson correlation
corr, _ = pearsonr(list1, list2)
print('%0.3f' % corr)
```

Output :

```
9.0
3.32
7.5
2.03
0.816
```

Summary						
Set	mean (X)	sd (X)	mean (Y)	sd (Y)	cor (X,Y)	
1	9	3.32	7.5	2.03	0.816	
2	9	3.32	7.5	2.03	0.816	
3	9	3.32	7.5	2.03	0.816	
4	9	3.32	7.5	2.03	0.817	

Code: Python program to plot scatter plot :

```
# Import the required libraries
from matplotlib import pyplot as plt
import pandas as pd

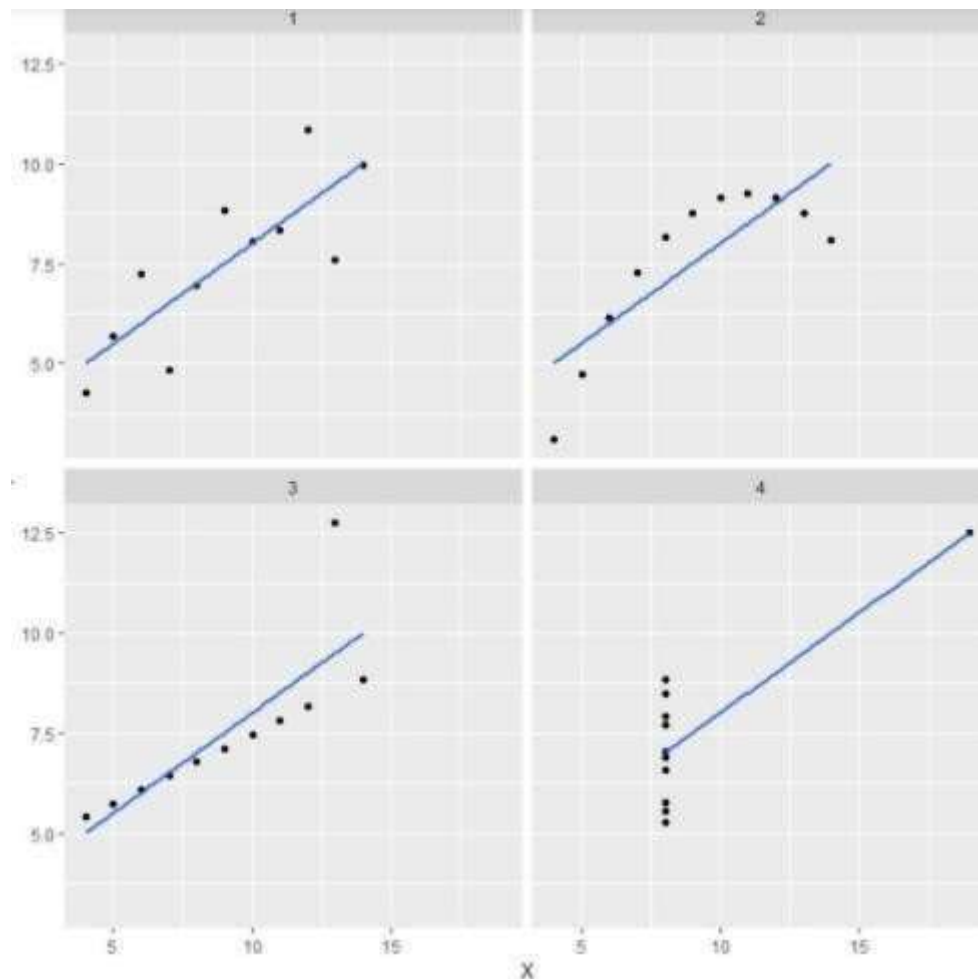
# Import the csv file
df = pd.read_csv("anscombe.csv")

# Convert pandas dataframe into pandas series
list1 = df['x1']
list2 = df['y1']

# Function to plot scatter
plt.scatter(list1, list2)

# Function to show the plot
plt.show()
```

Output:



3. What is Pearson's R?

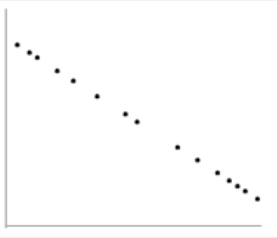

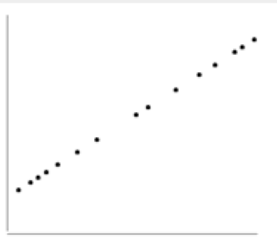
Pearson r correlation: Pearson r correlation is the most widely used correlation statistic to measure the degree of the relationship between linearly related variables. For example, in the stock market, if we want to measure how two stocks are related to each other, Pearson r correlation is used to measure the degree of relationship between the two. The point-biserial correlation is conducted with the Pearson correlation formula except that one of the variables is dichotomous. The following formula is used to calculate the Pearson r correlation:

$$r_{xy} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}}$$

r_{xy} = Pearson r correlation coefficient between x and y
 n = number of observations
 x_i = value of x (for i th observation)
 y_i = value of y (for i th observation)

Values of Pearson's correlation coefficient

Pearson's correlation coefficient (r) for continuous (interval level) data ranges from -1 to +1

$r = -1$		data lie on a perfect straight line with a negative slope
$r = 0$		no linear relationship between the variables
$r = +1$		data lie on a perfect straight line with a positive slope

Positive correlation indicates that both variables increase or decrease together, whereas negative correlation indicates that as one variable increases, so the other decreases, and vice versa.

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

Feature Scalling: Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step

Example: If an algorithm is not using feature scaling method then it can consider the value 3000 meter to be greater than 5 km but that's actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

Techniques to perform Feature Scaling

Min-Max Normalization: This technique re-scales a feature or observation value with distribution value between 0 and 1.

$$X_{\text{new}} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

Standardization: It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

$$X_{\text{new}} = \frac{X_i - X_{\text{mean}}}{\text{Standard Deviation}}$$

Why Scaling is used ?

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.

Difference between Normalisation and Standardisation

Normalisation	Standardisation
1. Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2. It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3. Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4. It is really affected by outliers.	It is much less affected by outliers.
5. Scikit-Learn provides a transformer called <code>MinMaxScaler</code> for Normalization.	Scikit-Learn provides a transformer called <code>StandardScaler</code> for standardization.
6. This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7. It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8. It is often called as Scaling Normalization	It is often called as Z-Score Normalization.

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

VIF = variance inflation factor

Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables. Mathematically, the VIF for a regression model variable is equal to the ratio of the overall model variance to the variance of a model that includes only that single independent variable. This ratio is calculated for each independent variable.

The higher the value, the greater the correlation of the variable with other variables. Values of more than 4 or 5 are sometimes regarded as being moderate to high, with values of 10 or more being regarded as very high. These numbers are just rules of thumb; in some contexts a VIF of 2 could be a great problem.

The Variance Inflation Factor (VIF) is a measure of co linearity among predictor variables within a multiple regression. It is calculated by taking the the ratio of the variance of all a given model's betas divide by the variance of a single beta if it were fit alone.

In this way, why is Vif infinite?

If there is perfect correlation, then $VIF = \text{infinity}$. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

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

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 –

- i. come from populations with a common distribution
- ii. Have common location and scale
- iii. Have similar distributional shapes
- iv. Have similar tail behaviour