Subjective Questions

Assignment-based Subjective Questions

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

The linear model equation after training:

y = 526.45
+2001.47*yr
-1045.42*holiday
+4180.74*temp
-508.86*season_spring
+603.29*season_summer
+833.91*season_winter
+387.93*day_11
+452.76*day_16
+388.27*day_17
+433.68*mnth_10
+340.26*mnth_8
+912.70*mnth_9
-460.89*weekday_Tues
-2076.14*weathersit_bad_weather
+707.02*weathersit_best_weather

As per this model,

- Variables with negative effect: holidays, spring season, Tuesdays, and bad weather
- Variables with positive effect: year (2019 has more business than 2018), temparature, summer and winter season, 11th 16th and 17th of months, August September and October months, best weather
- Business is drastically reduced if it's a holiday
- Business increases drastically with increase in temperature
- Summer season has the best positive effect among seasons
- Business is a little better on the 11th, 16th, and 17th of a month
- Business is good on the September month compared to other months
- Business has a negative effect on Tuesdays
- Bad weather is the biggest negative factor to business

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

Drop_first=True creates only n-1 column, where n -> no of categories.

We can drop one category because

- The value of the dropped category can be inferred from the rest of the n-1 categories.
- There will be high correlation between the dropped category and the rest of the category. So, to make sure our assumption of independent X variables, we drop one category.

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

Variable “temp” has the highes correlation with the target variable.

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

Validated the assumptions of Linear Regression by plotting a distribution plot of the residuals.

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

Top 3 features:

- Weather (business very bad if bad weather)
- Holidays (business bad during holidays)
- Temperature (business increases a good amount with increase in temperature)

General Subjective Questions

Explain the linear regression algorithm in detail.

Introduction

Linear Regression algorithm is an ML algorithm used to build models that predict a target variable in the population (y), give values to a set of independent variables(X). This algorithm is the simplest of the ML algorithm, which generates a simple linear equation as the model, of the form

$$y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \ldots. + \beta_i \cdot X_i + \ldots + \beta_n \cdot X_n$$

Where

y -> target variable / dependent variable

X_i -> independent variable

n -> number of independent variables

How it works

The Linear Regression algorithm helps in finding out the coefficient(β_i) values of the above model/linear equation. Once the coefficients are found, we can substitute the X_i values to get our needed target variable y.

Model does so by using a sample from the population. Let's call this sample a dataset.

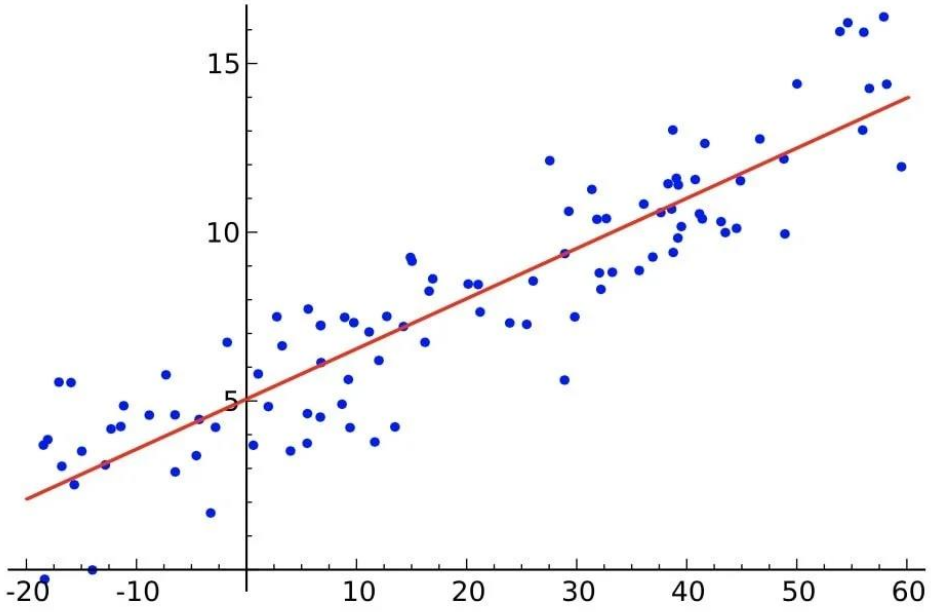
Simple Linear Regression

Let's take a single X value for the model for simplicity. The dataset now involves a lot of points that can be plotted on a graph. The model will look like

$$y = \beta_0 + \beta_1 \cdot X$$

Which is the equation of a line.

The Linear Regression model will plot the best fit line that explains all the data points. It will look something like this



How Linear Regression algorithm finds the Best Fit LIne

To explain this, we introduce the concept of **Residual Sum of Errors (RSS)**. It is the sum of square of the y distance between the model/line and each data point.

Or in other words

$$RSS = \sum (y_{datapoint} - y_{line})^2$$

The best-fit line is the one with the lowest RSS. The Linear Regression algorithm finds the coefficients(β) for the best-fit line using the same.

How to find Model/Line with Lowest RSS

We use gradient descent to find the lowest RSS.

How Gradient Descent Works

A lot of theory behind it. In short, conceptually, they find the gradient of the coefficients(β) which gives the relative magnitude and direction (+/-) of how much each coefficient should be changed to lower RSS.

Multiple Linear Regression

Expand Simple Linear Regression to Multiple dependent variables (X), and we have Multiple Linear Regression algorithm.

With this, we have gone through all major concepts in Linear Regression algorithm.

Explain the Anscombe's quartet in detail.

Anscombe's quartet is a set of 4 datasets, each with 11 (x,y) points. This is the dataset:

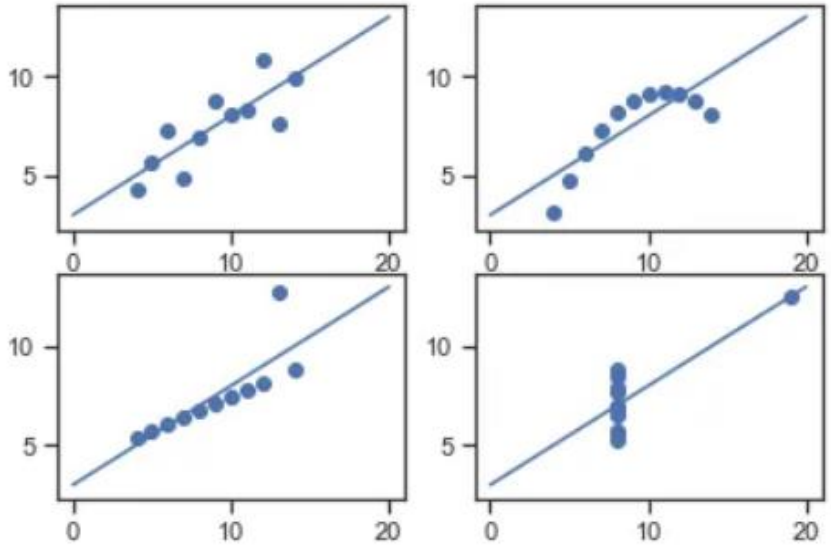
x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7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

Four Data-sets

They all contains the same descriptive statistics, which are:

- 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

But on plotting the dataset with the Linear Regression Equation, we get:



Graphical Representation of Anscombe's Quartet

As can be seen, all of them differ greatly even if they all have the same descriptive statistics.

This is a demonstration by Francis Anscombe as to why graphical representation is important even with all the descriptive statistics.

What is Pearson's R?

Pearson's R, also called Pearson's Correlation Coefficient, is a measure of linear relationship between two variables.

Before explaining Pearson's R, we need to talk about Covariance between two variables

Covariance

Covariance is a statistical measure that quantifies how two variables change together.

$$Cov(X,Y) = \frac{\sum (X_i - X_{mean}) \cdot (Y_i - Y_{mean})}{n-1}$$

This does give a measure of relationship between two variables, but two Covariance values of different variables can't be compared together since they are scaled different.

Pearson's Correlation Coefficient

Pearson's Correlation Coefficient is the normalized version of Covariance, by dividing the Covariance with the standard deviation of X and Y.

$$Pearson's\ Correlation,\ r = \frac{Cov(X,Y)}{\sigma(X) \cdot \sigma(Y)}$$

Since its normalized, this value can be used to compare the relationship between multiple pairs of variables.

What R value means

- r = 1 : perfect positive linear relationship
- r = -1 : perfect negative linear relationship
- r = 0 : no linear relationship

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

Scaling

Scaling is the process of transforming different variable values to the same order of magnitude/scale.

Why are they performed?

They are performed for two reasons:

- Interpretation of Coefficients

  If the variable values are not in the same scale/range/order of magnitude, nor will the coefficent value (β) after training. So, it will be difficult to compare the coefficients, and hence, difficult to interpret the trained model

- Faster Gradient Descent

Difference between Normalized Scaling and Standardized Scaling

- Normalized Scaling
  - Scales the variables between 0 and 1.
  - $X_{normalized} = \frac{X - \min(X)}{\max(X) - \min(X)}$
- Standardized Scaling
  - Performs transformation on the variables, making their mean 0 and standard deviation 1.
  - $X_{standardized} = \frac{X - mean(X)}{std(X)}$

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

Formula for VIF is

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

Where

R^2 -> R2-score

The R2-score is a measure of how well a model explains the variance in the dataset.

If a model completely explains the variance of a dataset, then the r2-score will be 1, and inturn, the VIF value becomes infinite.

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

What is Q-Q plot

Q-Q plot, or Quantile-Quantile plot, is a plot of (X_i, Y_i), where

X -> Value of the Theoretical distribution laying in some quantile i

Y -> Value of the Provided distribution laying in the same quantile i

If the plot follows

X=Y

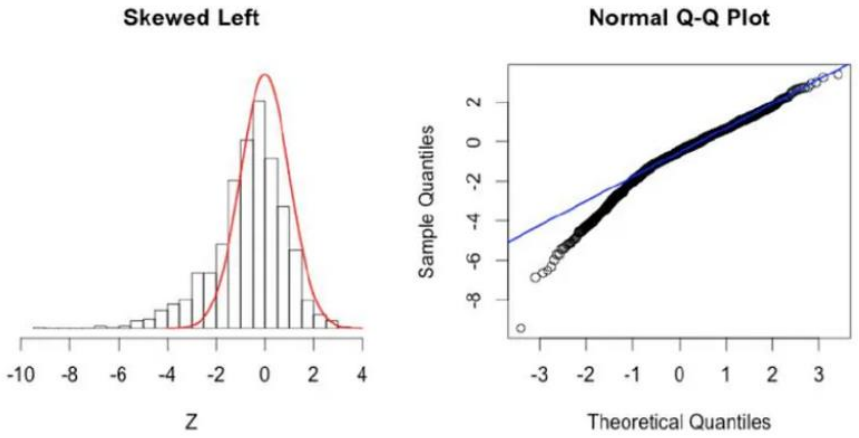
then, the provided distribution matches the expected theoretical distribution

This plotting is used:

- to validate if a provided distribution matches a certain theoretical distribution
- find in which all quantiles the given distribution deviates from the expected theoretical distribution

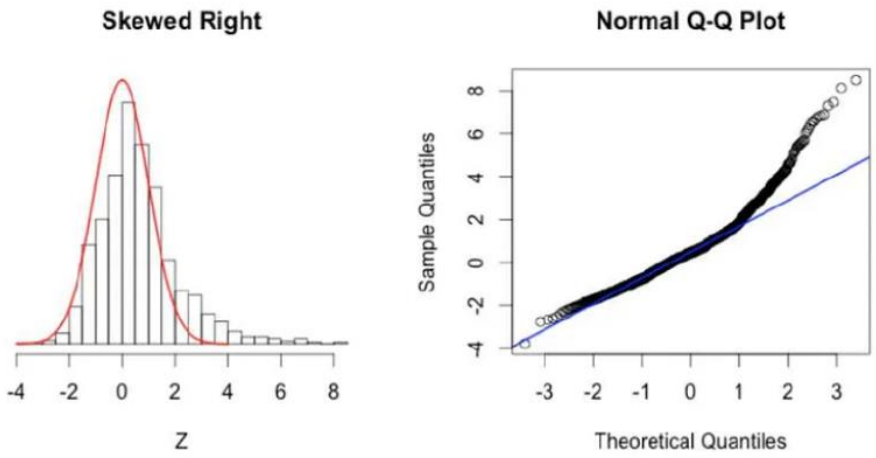
For example,

- If the given distribution is skewed left wrt the expected theoretical distribution



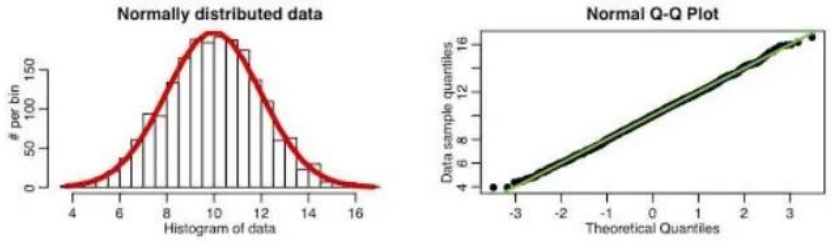
Left Skewed Q-Q plot for Normal DIstribution

- Similarly, if its skewed right



Right Skewed Q-Q plot for Normal Distribution

- But if the given distribution and expected theoretical distribution matches



<u>Importance of Q-Q Plot in Linear Regression</u>

- Residual Analysis: Can be used to check if the residuals follow a normal distribution