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)

Answer:

Some of the categorical variables do seem to have an impact on the target variable. For example –

- We have the highest bike-bookings in fall followed by summer and winter. Lowest in spring.
- Year has more bookings in year 1 (2019) vs year 0 (2018).
- Holiday, working day, and weekday have weak to no relationship with target variable.
- Weather situation has a significant impact on target variable. Highest bike rentals happen in clear weather followed by misty weather and then light rain. No data for heavy rain available.
- Months also have a significant impact which is correlated with season.

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

Answer:

To avoid multicollinearity, since, for a variable with k categories, automatically k dummy variables get created. Together these dummy variables have perfect linear relationship among themselves since any of these can be fully described by the others. Hence, we drop one of them using drop_first=True.

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

Answer:

Just looking at the pair-plot, both temp and atemp have the highest correlation with the target variable.

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

Answer:

Part of the validation was already done during EDA by ascertaining relationship between predictor and target variables and determining that at least some have a linear relationship.

After model building, I plotted a histogram of the error terms and found that they were approximately normally distributed with mean close to zero.

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

Answer:

We will use magnitude of coefficients obtained from our model to determine importance since all the features were either binary or scaled to 0-1 range. Based on that, the top 3 features in order are temperature, year, and windspeed.

General Subjective Questions

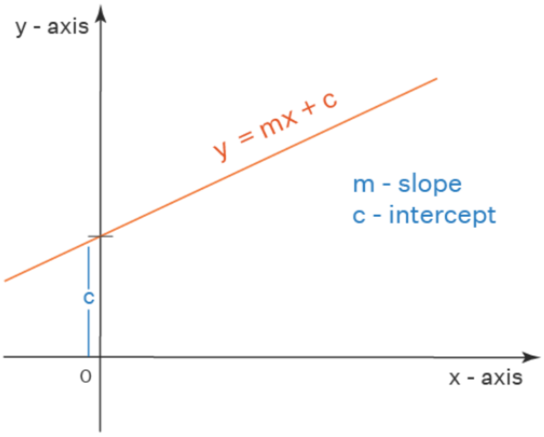
1. Explain the linear regression algorithm in detail. (4 marks)

Answer:

In linear regression, we try to figure out a straight-line relationship between the independent and dependent variables. This relationship follows the straight-line equation –

$$y = mx + c$$

where m is the slope and c is the intercept



Types of Regression

1. Simple Linear Regression
   - Model with only 1 independent variable
2. Multiple Linear regression
   - Model with more than 1 independent variables

The linear regression model attempts to fit a straight line through the data. The equation of the straight line would be as follows for simple linear regression –

$$y = \beta_0 + \beta_1 x$$

The model tries to find the best-fit line for the data. We define error or residual as the difference between the observed value of y and the predicted value of y (by our fitted line) for any datapoint.

$$e_i = y_i - y_{pred}$$

We generally use the Ordinary Least Squares method to best fit this line. First, we get a sum of the squares of all the residuals (RSS) and then minimize this sum to get the line with the smallest error.

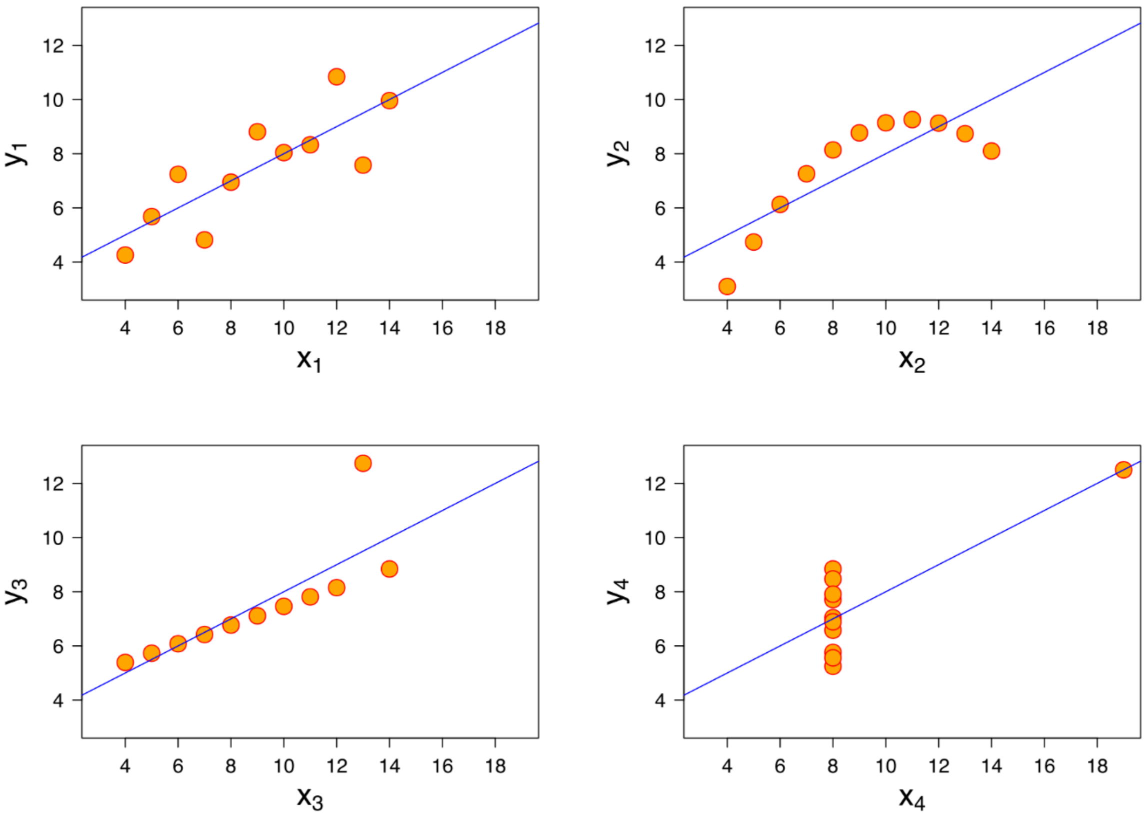
$$\text{RSS (Residual Sum of Squares)} = e_1^2 + e_2^2 + \ldots + e_n^2$$

$$= \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)$$

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

Answer:

The Anscombe's quartet was developed by statistician Francis Anscombe in 1973 to demonstrate how important EDA and detailed data visualization can be since relying only on summary stats can be very misleading. As an example, look at the below set of charts (copied from Wikipedia) which shows fours data distributions with the same mean and standard deviations of x & y variables but vastly different distributions.



https://en.wikipedia.org/wiki/Anscombe%27s_quartet#:~:text=Anscombe's%20quartet%20comprises%20four%20data,(x%2Cy)%20points.

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

Answer:

It is the commonly used correlation coefficient sued to denote the strength of linear relationship between two variables. It takes values between -1 to 1 indicating a perfect negative relationship to a perfect positive relationship. The R^2 obtained in linear regression is same as the square of this correlation coefficient.

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

Answer:

Scaling refers to condensing the range of values that a variable can take. There are two key benefits of scaling –

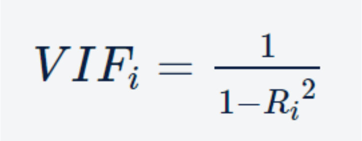
1. Ease of interpretation – if different X variables have different scales, then it is difficult to compare them and interpret the relative value of their coefficients. Scaling helps a lot here, especially in MLR.
2. Faster convergence for gradient descent methods – scaling the features and getting them to similar ranges of values helps the model to converge faster.

Normalized scaling, or min-max scaling, bring all the values in the range of 0 to 1 using the minimum and maximum values of the variable. Whereas, standardized scaling converts all the values to Z-scores using the mean and standard deviation of the variable.

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

Answer:

VIF increases with increase in R^2. Hence, more the correlation, higher the VIF and when the correlation becomes perfect (very close to 1), VIF tends to Infinity.



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

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

Answer:

A Q-Q plot, which means a quantile-quantile plot is used to compare two probability distributions by plotting their quantiles against each other. If you want to know whether a given distribution is Gaussian, or Uniform, or exponential, or Normal, etc., you can plot the quantiles of the data against calculated theoretical quantiles of that distribution.

This would be very useful in linear regression so ascertain the normality of the error terms and validate the model assumptions.