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)

A. Albeit any of the categorical variables' dummies did not have biggest influenced, if we combine the individual absolute weights of each of these dummies of "weathersit", "season", "extended_weekend" etc., they would have a very big influence overall on the model.

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

A. If we do not drop one of the categories in the form of drop_first=True, that dummy variable could be predicted (or correlated high) with other dummy variables of the same categorical variable. This would in turn result in high VIF for that dummy variable.

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

A. temp (or even atemp) has the highest correlation with the dependent variable.

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

A.

1. Overall distribution of the train residuals is following normal distribution centered at 0.
2. There is no significant pattern for residuals vs y train.

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

A. Top 3 factors in **absolute** contribution of bike demand are:

1. atemp
2. "Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds" category of weathersit. (negatively)
3. yr (year).

General Subjective Questions

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

A.

1. Linear Regression (Ordinary Least Squares) aims to predict an unique real value for each sample of independent variables.

2. This is done in a linear way where each independent variable is assigned a weightage.

3. The algorithm learns weights such that the average squared distance between the predicted line (equation) and the actuals are minimized

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

A. Anscombe's quartet consists of 4 sets of dependent and dependent variables. They all have same descriptive statistics:

1. Mean and variance of dependent and independent variables are the same.

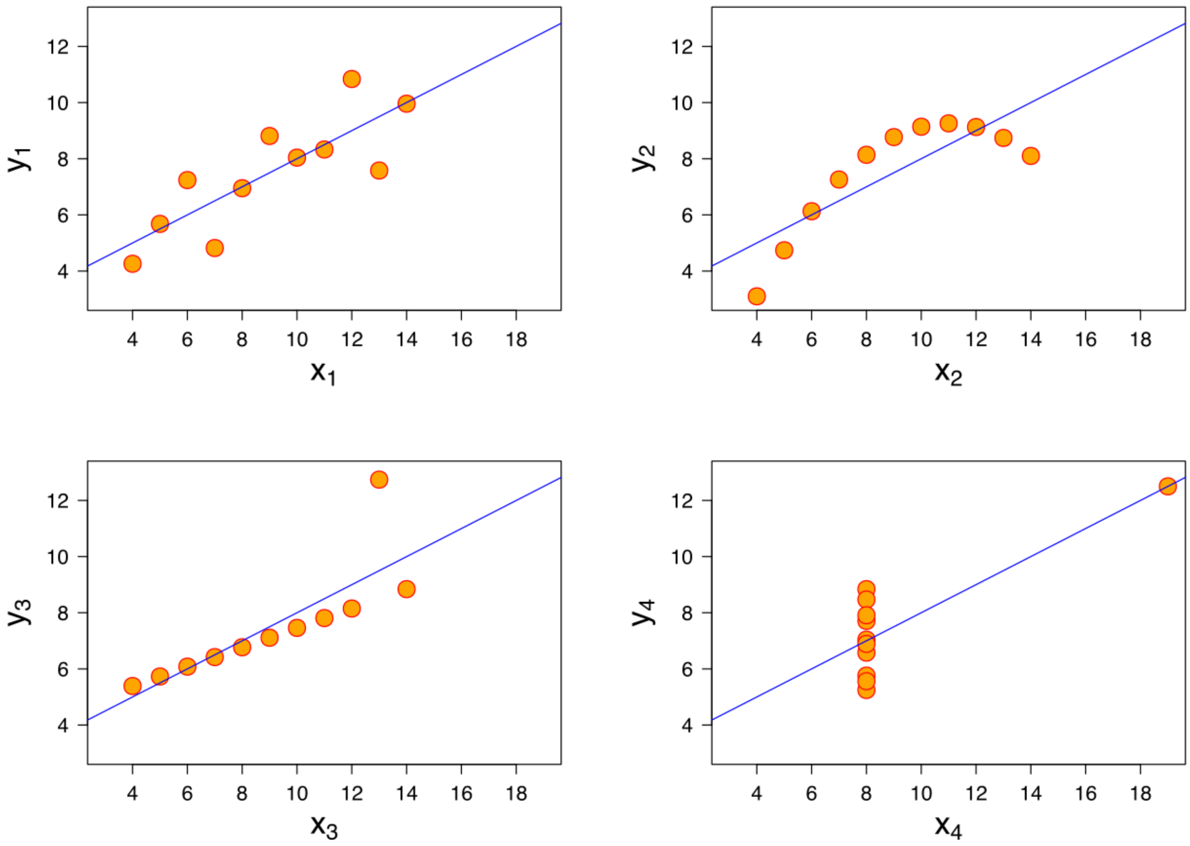
2. Correlation between the independent and dependent variables are the same.

3. Linear regression line equation is the same.

4. Coefficient of determination (r-squared) is the same.

Yet they all have different relationships and distributions.

This is generally demonstrated to show the importance of plotting and exploratory analysis.



Source: Wikipedia

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

A.

1. This is a correlation coefficient which measures the level of association (linearly) between 2 real variables.

2. Its upper and lower limits are +1 and -1 respectively.

3. High positive values mean that the association is positive and if there is an increase in one variable, the other tends to increase as well. Similarly, negative correlation results in general decrease of one variable with increase in other.

4. Correlation near 0 on either side means there is no linear association between the variables.

5. This is calculated using covariance of the variables which are standardized first by subtracting mean and then divided by standard deviation.

Correlation does not mean causation!

A lower absolute value of Pearson’s R does not mean absence of association, it just means that there is no linear association!

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

and standardized scaling? (3 marks)

A.

1. Scaling is done to ensure all the independent variables are on the similar scale. This is generally done by either subtracting or dividing by or combination of both for the independent variable.

2. This is done to make sure that the influence of independent variables is more accurate than the unscaled version. Also, this speeds up the gradient descent.

3. a. In normalization, mean of the variable is subtracted and divided by standard deviation.

b. In standardization, minimum value of the variable is subtracted and divided by the range (maximum value – minimum value). This ensures that all the values are in the range 0-1.

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

(3 marks)

A. Since VIF = 1/(1-r2). VIF of infinity means that r2 is 1, which in turn means that whole variance of dependent variable is explained by the model alone. This could happen in the cases of data leakage or overfitting.

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 is a quantile-quantile plot which compares probability distributions of 2 variables by taking respective quantiles (orders) and then plotting respective quantile pairs on a scatter plot. If the distributions are similar, there tends to be a linear relationship on the Q-Q plot.

For linear regression, since one of the assumptions is that residuals follow a normal distribution, this is validated by plotting residuals on the Q-Q plot against normal distribution samples.