Assignment-based Subjective Questions

--Submission by Nitin Balaji Srinivasan

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: From a boxplot analysis for the following categorical variables on the target dependent variable 'cnt', I can infer the following.

a. Season – Fall season has more bookings 'cnt' than any other season
b. Year – 2019 has more bookings 'cnt' than 2018 (business improvement)
c. Month – July and Sep have the highest medians
d. Holiday – Most bookings happen on non-holidays
e. Weekday – Similar medians for each of the week days
f. Workingday – Almost the same median of bookings between working day and non-working day, though the number of bookings are high on working day
g. Weathersit – More bookings on a clear day vs Light snow + rain

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

Answer: Drop_first=True ensures that an additional dummy variable is not created, as (n-1) dummy variables are sufficient to cover 'n' categorical values.

Syntax - drop_first: bool, default False

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

Answer: temp variable has the highest correlation with the target variable 'cnt'

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

Answer: I validated the assumptions of Linear Regression based on the following checks:

a. Normal distribution of error terms
b. Multicollinearity and no auto-correlation
c. Homoscedasticity – No visible pattern in residual values – equal distance from the mean error
d. linear relationship validation

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

Answer: The following are the top 3 features contributing significantly towards explaining the demand of the shared bikes –

- temp (Coefficient: 0.4910)
- weathersit – “Light snow rain” (Coefficient: -0.2842)
- Year (Coefficient: 0.2336)

General Subjective Questions

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

Answer: Linear regression may be defined as the statistical model that analyses 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 −

$Y = mX + c$

Here, Y is the dependent variable we are trying to predict.

X is the independent variable we are using to make predictions.

m is the slope of the regression line which represents the effect X has on Y

c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.

Linear regression is of the following two types −

- Simple Linear Regression
- Multiple Linear Regression

Line of best fit refers to a line through a scatter plot of data points that best expresses the relationship between those points. Statisticians typically use the least squares method (sometimes known as ordinary least squares, or OLS) to arrive at the geometric equation for the line, either through manual calculations or by using software.

A straight line will result from a simple linear regression analysis of two or more independent variables. A **multiple regression** involving several related variables can produce a curved line in some cases.

Ordinary Least Squares Method:

Ordinary least squares (OLS) is a type of linear least squares method for choosing the unknown parameters in a linear regression model by the principle of least squares: minimizing the sum of the squares of the differences between the observed

dependent variable (values of the variable being observed) in the input dataset and the output of the (linear) function of the independent variable.

In mathematical terms, the OLS formula can be written as the following:

Minimize $\sum(y_i - \hat{y}_i)^2$

where y_i is the actual value, $\hat{y}_i$ is the predicted value. A linear regression model used for determining the value of the response variable, $\hat{y}$, can be represented as the following equation.

$$y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n + e$$

where:

- y is the dependent variable
- b0 is the intercept
- b1, b2, ..., bn are the coefficients of the independent variables x1, x2, ..., xn
- e is the error term

Assumptions for linear regression -

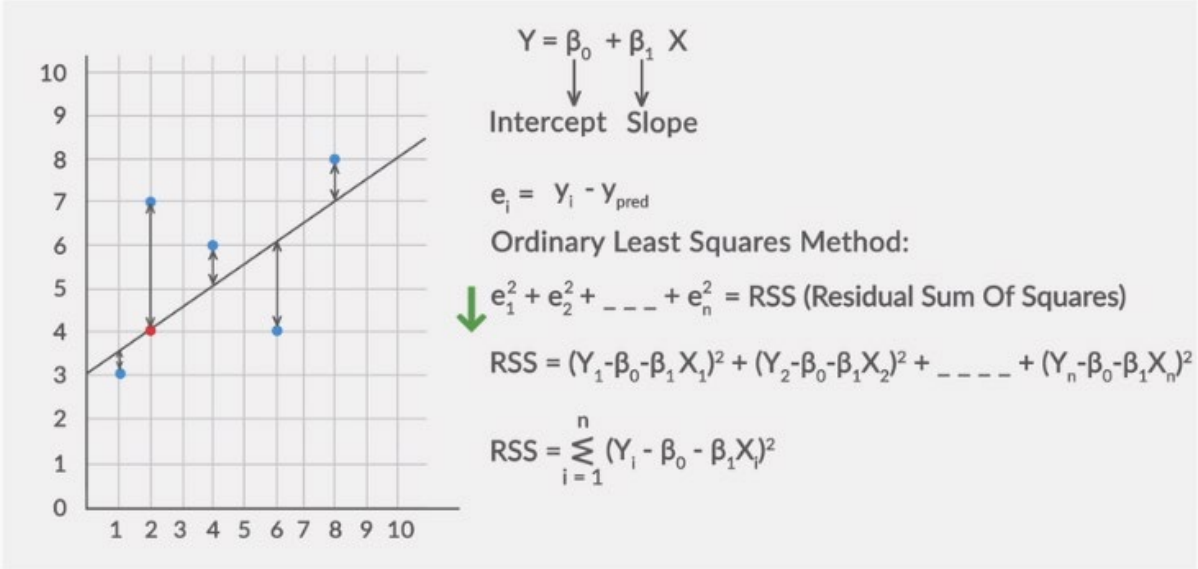
- Multi-collinearity –
  Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.
- Auto-correlation – Another assumption Linear regression model assumes is 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.
- Normality of error terms – Error terms should be normally distributed
- Homoscedasticity –There should be no visible pattern in residual values.

Loss function:

A loss function or cost function (sometimes also called an error function) is a function that maps an event or values of one or more variables onto a real number intuitively representing some "cost" associated with the event. An optimization problem seeks to minimize a loss function.

RSS, TSS and R2:

1. Ordinary least square or Residual Sum of squares (RSS) — Here the cost function is the (y(i) — y(pred))² which is minimized to find that value of β0 and β1, to find that best fit of the predicted line.



2. This is an absolute difference between the actual y and the predicted y. Now, if the units of the actual y and predicted y changes the RSS will change. So, we use the relative term R² which is 1-RSS/TSS

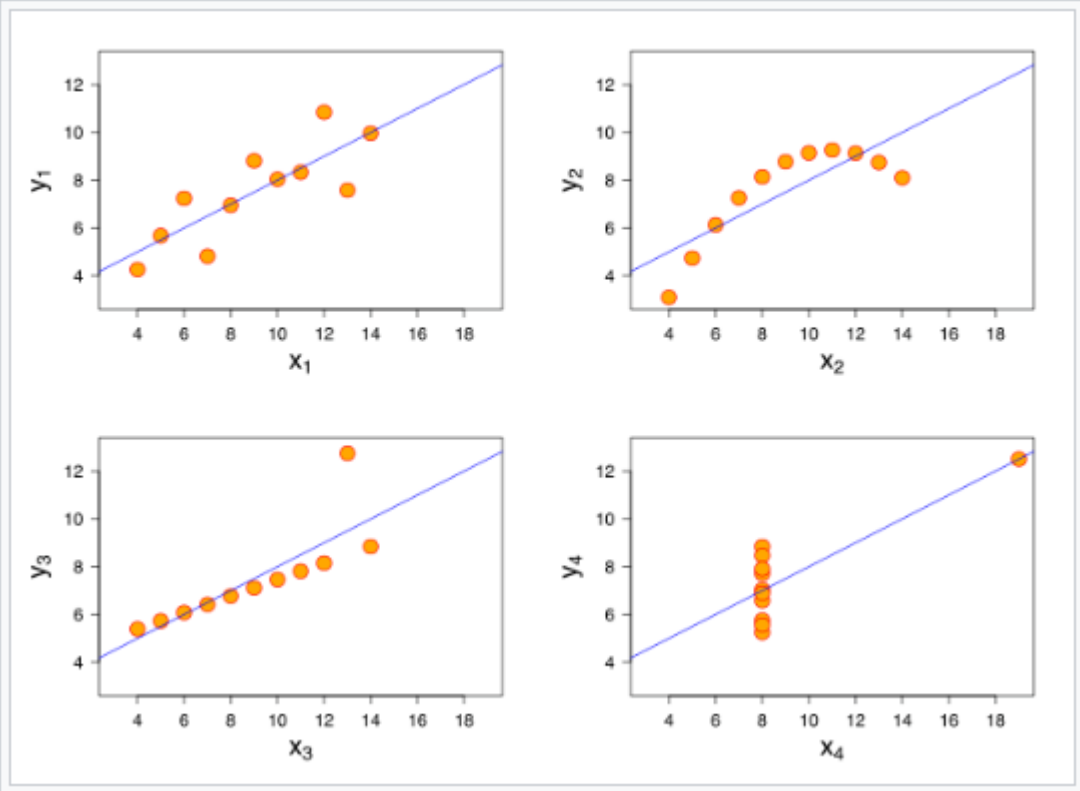
3. TSS — total sum of squares - Instead of adding the actual value's difference from the predicted value, in the TSS, we find the difference from the mean y the actual value.

4. TSS works as a cost function for a model which does not have an independent variable, but only y intercept (mean $\bar{y}$). This gives how good is the model without any independent variable. When independent variable is added the model performance is given by RSS. The ration of RSS/TSS gives how good is the model as compared to the mean value without variance.

Lesser is this ratio lesser is the residual error with actual values, and greater is the residual error with the mean. This implies that the model is more robust. So, 1-RSS/TSS is considered as the measure of robustness of the model and is known as R^2

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

Answer: 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. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties.



- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two correlated variables, where y could be modelled as gaussian with mean linearly dependent on x.
- For the second graph (top right), while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.

- In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier, which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.
- The datasets are as follows. The *x* values are the same for the first three datasets.

Anscombe's quartet

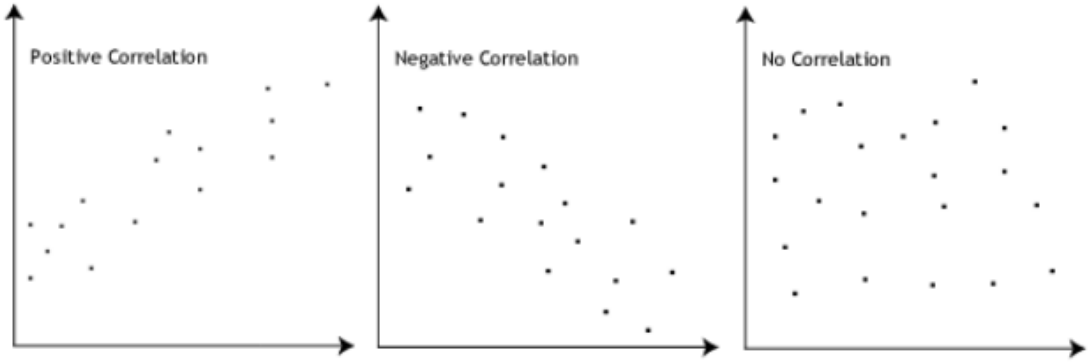
I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	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

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

Answer: Pearson's correlation coefficient, a measurement quantifying the strength of the association between two variables. Pearson's correlation coefficient r takes on the values of −1 through +1.

Values of −1 or +1 indicate a perfect linear relationship between the two variables, whereas a value of 0 indicates no linear relationship. (Negative values simply indicate the direction of the association, whereby as one variable increases, the other decreases.) Correlation coefficients that differ from 0 but are not −1 or +1 indicate a linear relationship, although not a perfect linear relationship.



The Pearson's correlation coefficient formula is

$$r = [n(\Sigma xy) - \Sigma x \Sigma y] / \sqrt{[n(\Sigma x^2) - (\Sigma x)^2][n(\Sigma y^2) - (\Sigma y)^2]}$$

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

Answer: Feature scaling is a data preprocessing technique used to transform the values of features or variables in a dataset to a similar scale. The purpose is to ensure that all features contribute equally to the model and to avoid the domination of features with larger values. Feature scaling becomes necessary when dealing with datasets containing features that have different ranges, units of measurement, or orders of magnitude. There are several common techniques for feature scaling, including standardization, normalization, and min-max scaling. These methods adjust the feature values while preserving their relative relationships and distributions. By applying feature scaling, the dataset's features can be transformed to a more consistent scale, making it easier to build accurate and effective machine learning models.

Difference between normalised scaling and standardised scaling

Min-max scaling or min-max normalization is the simplest method and consists in rescaling the range of features to scale the range in [0, 1] or [−1, 1]. Selecting the target range depends on the nature of the data. The general formula for a min-max of [0, 1] is given as:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

where x is an original value, x' is the normalized value.

Standardisation scaling: Feature standardization makes the values of each feature in the data have zero-mean (when subtracting the mean in the numerator) and unit-variance. This method is widely used for normalization in many machine learning algorithms (e.g., support vector machines, logistic regression, and artificial neural networks).[3][4] The general method of calculation is to determine the distribution mean and standard deviation for each feature. Next we subtract the mean from each feature. Then we divide the values (mean is already subtracted) of each feature by its standard deviation.

$$x' = \frac{x - \bar{x}}{\sigma}$$

Where x is the original feature vector, $\bar{x} = \text{average}(x)$ is the mean of that feature vector, and σ is its standard deviation.

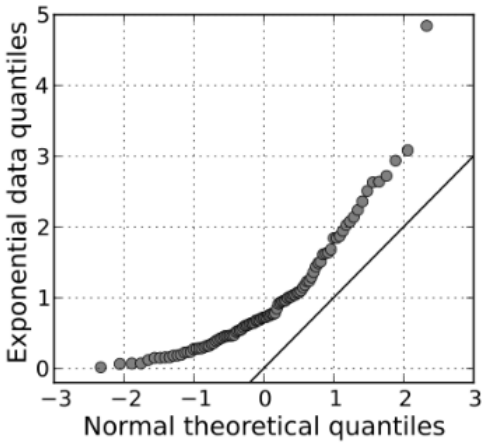
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer: Perfect correlation results in VIF = Infinity. High correlation results in high VIF. In case of perfect correlation R2=1, therefore VIF = 1/(1-R2) will result in infinity.

This shows that there is a perfect correlation between two independent variables suggesting that one of the variables will need to be dropped.

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

Answer: In statistics, a Q–Q plot (quantile–quantile plot) is a probability plot, a graphical method for comparing two probability distributions by plotting their quantiles against each other. A point (x, y) on the plot corresponds to one of the quantiles of the second distribution (y-coordinate) plotted against the same quantile of the first distribution (x-coordinate). This defines a parametric curve where the parameter is the index of the quantile interval.



If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the identity line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

The Q-Q plot is a graph that tests the conformity between the empirical distribution and the given theoretical distribution.

In **linear regression**, a Q-Q plot is used to check if the residuals are normally distributed. If the residuals are normally distributed, the points on the Q-Q plot will fall approximately on a straight line. The Q-Q plot can be used to identify departures from normality. If there are departures from normality, we can transform the data or use other methods to address the issue.