Assignment based subjective Questions: Answers

1) On analysis of categorical variables versus dependent variable the following can be inferred

- For weekday variable, the medians are similar for all categories
- For month variable, the distribution shows greater density in the middle months of the year. Example: 4th month to 10th month
- For Season variable, there is a good variance in distribution

2) Dummy variable creation is usually done on a categorical variable. If we have a variable that has 3 categories, for Ex:

Season
Summer
Winter
Monsoon

This can not directly be used during analysis or model building, hence we break them down to binary format as follows

Summer	Winter	Monsoon
1	0	0
0	1	0
0	0	1

From this we can see that in any case one variable will be 1, so it would be more efficient and clean to remove one of the dummy variables. In this case on using drop_first=True, we remove the Summer variable. Hence when Winter and Monsoon are 0, it would automatically mean that Summer is 1.

3) On looking at the pair-plot we can see that Registered variable shows highest correlation with the target variable

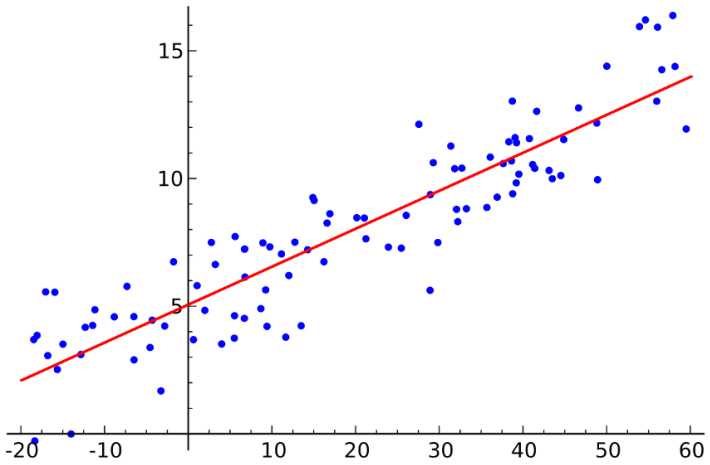
4) After building the model and training it, it is important to validate the model and this can be done the following way:

- The model has a high adjusted R squared value and the dependable variables show a p value less than 0.05. This helps in understanding the linear relationship between X and Y variables
- Plotting the Error terms by calculating the difference between ground truth Y value and predicted Y value on train dataset. The distribution plotted shows a normal distribution which in turn favors the assumptions on linear regression
- The mean of error terms are also checked to be 0

5) On checking the correlation matrix for the decided variables we can conclude that the temp, yr and summer season are the top 3 features contributing towards the final model

General Subjective Questions : Answers

1) Regression is a method of modelling a target value based on independent predictors. This method is mostly used for forecasting and finding out the cause and effect relationship between variables. Regression techniques mostly differ based on the number of independent variables and the type of relationship between the independent and dependent variables.



Simple linear regression is a type of regression analysis where the number of independent variables is one and there is a linear relationship between the independent x and dependent y variable. The red line in the above graph is referred to as the best fit straight line. Based on the given data points, we plot a line that models the points the best. The line can be modelled based on the linear equation shown below.

y = a_0 + a_1 * x ## Linear Equation

The motive of the linear regression algorithm is to find the best values for a0 and a1. The same concept is followed for multiple regressions when you have multiple variables. The other two important concepts here are the cost function and gradient descent.

The cost function helps us to figure out best possible values for a0 and a1 (all coefficients in case of multiple regression)which would provide the best fit line for the data points. Since we want the best values for a0 and a1, we convert this search problem into a minimization problem where we would like to minimize the error between the predicted value and the

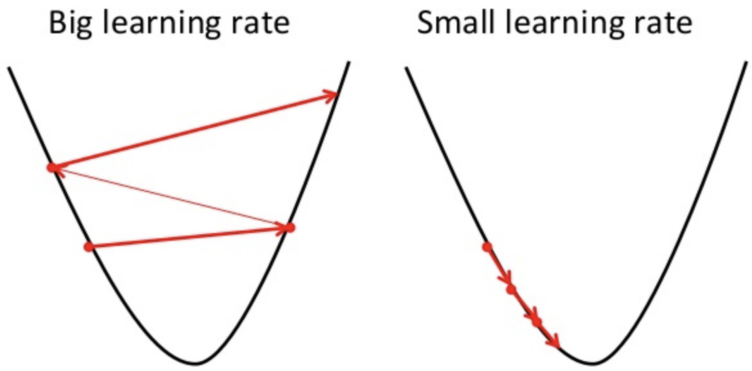
$$minimize \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

$$J = \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

actual value.

We choose the above function to minimize. The difference between the predicted values and ground truth measures the error difference. We square the error difference and sum over all data points and divide that value by the total number of data points. This provides the average squared error over all the data points. This cost function is also known as the Mean Squared Error(MSE) function. Now, using this MSE function we are going to change the values of a_0 and a_1 such that the MSE value settles at the minima.

The next important concept needed to understand linear regression is gradient descent. Gradient descent is a method of updating a0 and a1 to reduce the cost function(MSE). The idea is that we start with some values for a0 and a1 and then we change these values iteratively to reduce the cost. Gradient descent helps the model to move towards local minima.



(Image sourced from internet)

To draw an analogy, imagine a pit in the shape of U and you are standing at the topmost point in the pit and your objective is to reach the bottom of the pit. There is a catch, you can only take a discrete number of steps to reach the bottom. If you decide to take one step at a time you would eventually reach the bottom of the pit but this would take a longer time. If you choose to take longer steps each time, you would reach sooner but, there is a chance that you could overshoot the bottom of the pit and not exactly at the bottom. In the gradient descent algorithm, the number of steps you take is the learning rate. This decides on how fast the algorithm converges to the minima.

In gradient descent, to update a_0 and a_1, we take gradients from the cost function. To find these gradients, we take partial derivatives with respect to a0 and a1 is.

$$a_0 = a_0 - \alpha \cdot \frac{2}{n} \sum_{i=1}^{n} (pred_i - y_i)$$

$$a_1 = a_1 - \alpha \cdot \frac{2}{n} \sum_{i=1}^{n} (pred_i - y_i) \cdot x_i$$

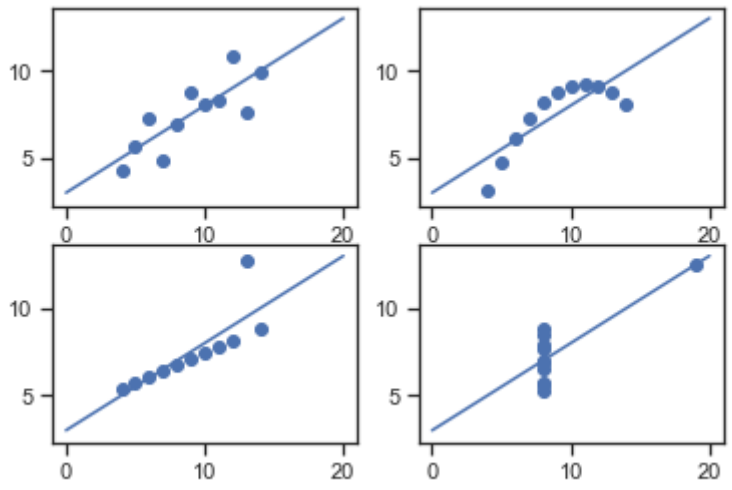
The partial-derivates are the gradients and they are used to update the values of a0 and a1. Alpha is the learning rate which is a hyperparameter that you must specify. A smaller learning rate could get you closer to the minima but takes more time to reach the minima, a larger learning rate converges sooner but there is a chance that you could overshoot the minima.

Once we converge on the minima, we can plot the line for the coefficients and see that it fits the data.

2) *Anscombe's Quartet* is the modal example to demonstrate the importance of data visualization that signify both the importance of plotting data before analyzing it with statistical properties. It comprises of four data-set and each data-set consists of eleven (x,y) points. The common thing to analyze in these data-sets is that they all share the same descriptive statistics(mean, variance, standard deviation etc) but different graphical representation. Each graph plot shows the different behavior irrespective of statistical analysis.
Following are statistical values of an assumed dataset

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.

We can see that the statistical analysis of the data-sets are pretty much similar. But on plotting these datasets across the x & y coordinate plane, we can assume a representation as follows:



- Data-set I — consists of a set of (x,y) points that represent a linear relationship with some variance.
- Data-set II — shows a curve shape but doesn't show a linear relationship (might be quadratic?).

- Data-set III — looks like a tight linear relationship between *x* and *y*, except for one large outlier.
- Data-set IV — looks like the value of *x* remains constant, except for one outlier as well.

Data-sets which are identical over a number of statistical properties, yet produce dissimilar graphs, are frequently used to illustrate the importance of graphical representations when exploring data. This isn't to say that summary statistics are useless. They're just misleading on their own. This shows the importance of such a tool in a larger data analysis process. Visualizing our data allows us to revisit our summary statistics and re-contextualize them as needed.

3) Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

4) Scaling is a technique applied to bring all the independent variables within the same range. This technique is performed so that during model building the values of independent variables are brought to the same range and the coefficients of the model would be more comparable and not extreme due to the different sized values of the variables. Keeping the data scaled also helps in cleaner visualization and analysis.

Normalized Scaling	Standardized scaling
X_new = (X - X_min)/(X_max - X_min)	X_new = (X - mean)/Std
Useful when features are of different scales	Useful when we want to ensure 0 mean and unit standard deviation
Scaled values will be bound between [-1,+1]	No bounds to the scaled values

5) If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.Hence when VIF is infinity it is most likely that the same variable is present twice in the dataset or there are two variables which are analogous of each other

6) Quantile-Quantile or Q-Q plot, is a graphical tool which tells us if a set of data came from some theoretical distribution such as a Normal, uniform or Exponential distribution. Also, it helps to determine if two data sets come from populations with a common distribution. It is

interpreted as a plot of the quantiles of the first data set against the quantiles of the second data set.

This is especially helpful in 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 the same distributions. The possible results are as follows:

1. Similar distribution: If all quantile points lie on or close to straight line at45 degree angle from x -axis
1. Y-values < X-values: If y-quantiles are lesser than the x-quantiles.
2. X-values < Y-values: If x-quantiles are lesser than the y-quantiles.
3. Different distribution: If all quantile points lie away from the straight line at 45 degree angle from x -axis