SUBJECTIVE QUESTIONS ASSIGNMENT

By- SATADHRITI CHAKRABARTY, DS C17 FEB 2020, GROUP 2

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?

- Analyzing the categorical variables from the dataset, we can infer:

- **Season**: For all the seasons, the spread is slightly more in summer and fall and the median value for fall is more than the rest. So, renting in autumn is generally more.
- **Holiday**: Spread on holidays is more but the median value of the variable not being a holiday is higher.
- **Workingday**: This is in tandem with the 'holiday' variable as we can clearly see that the spread on working days is less but the median value on a working day is higher.
- **Weathersit**: Spread as well as the median is the highest when the weather is clear with little or no cloud.
- **Weekday**: There is no visible trend but the spread on Wednesday and Saturday is a bit higher.
- **Yr**: Year 2019 has certainly seen a steep rise in rents from 2018.

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

- After creating dummies for a categorical variable, if we do not use the command drop_first=True, the first column won't be dropped whereas it is extremely important to drop it to avoid a **dummy variable trap**. For example, if we have **gender** as a categorical independent variable in our model, having **dummies for both males and females will lead to multicollinearity** problem as male = 1 would strictly imply that female = 0 and vice versa. Dropping one would negate this problem as male = 1 would mean that the person is a male and male = 0 would mean that the person is a female.

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

- The variable '**atemp**' 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?

- After building the model on the training set, to validate the assumptions of Linear Regression, we check the **normality of error terms** and find out that they follow a **Normal Distribution**. Checking for the **homoskedasticity** assumption, we plot a scatterplot of the Predicted Y with the residuals and find out **they do not have a cone or fan-shaped pattern** (which we find in case of heteroskedasticity). We have also checked for **multicollinearity** and removed the variables which displayed a strong linear relationship amongst them.

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

- The top 3 Features are '**yr**', '**weathersit**' (Light Rain/Snow), '**season**'(Spring) have the highest contributing factors towards explaining the demand for shared bikes. While the 'yr' has a positive effect on the dependent variable 'cnt', the other two variables have negative coefficients.

GENERAL SUBJECTIVE QUESTIONS

1) Explain the linear regression algorithm in detail

- Linear regression is the **simplest form of a machine learning algorithm** that depicts the relationship between the dependent (target) variable and independent (predictor) variables.

  There are broadly two types of linear regression:

  i) **Simple Linear Regression**, the most basic one dealing with only one predictor variable

  ii) **Multiple Linear Regression**, showing the relationship between the target and a set of more than one predictor variables

How is it done?

A straight line is first fitted on the scatter plot between the dependent and the independent variables.

The standard equation of the regression line of a **simple linear regression** is given by the expression:

$Y = \beta_0 + \beta_1 X$, where β_0 is the intercept and β_1 is the slope of the line. Intuitively, β_1 is the change in Y that occurs due to one unit change in X.

Similarly, for a **multiple linear regression**, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots + \beta_p X_p$, where X_1, X_2,…, X_p are the regressor variables.

The objective of linear regression is to find the best fit line on the scatter plot (in case of multiple linear regression, the line becomes a hyperplane, but the basic intuition is the same) which is obtained by ***minimizing the Residual Sum of Squares (RSS)***. Residual for any point is found by subtracting predicted value of the dependent variable from its actual value.

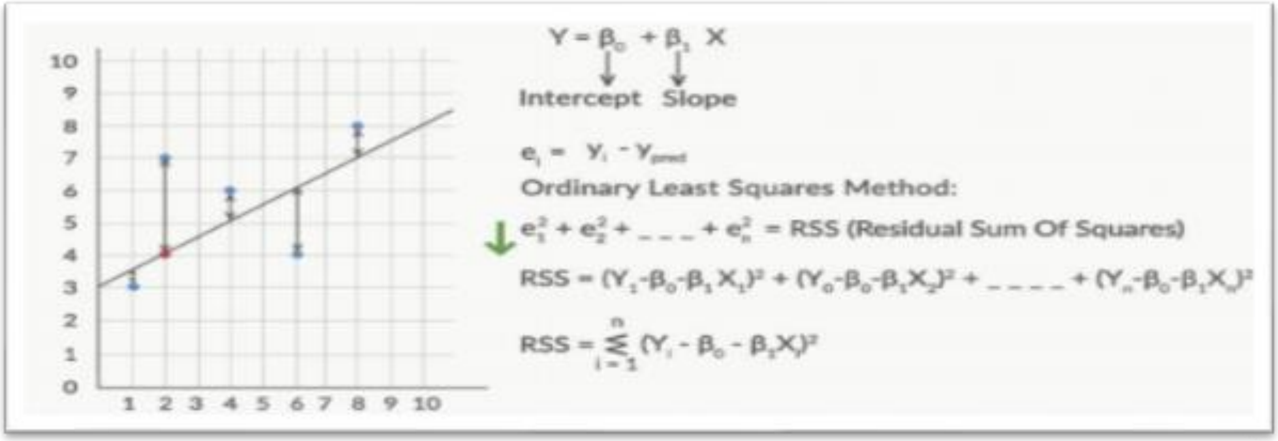

10
9
8
7
6
5
4
3
2
1
0
1 2 3 4 5 6 7 8 9 10
$Y = \beta_0 + \beta_1 X$
Intercept
Slope
$e_i = Y_i - Y_{pred}$
Ordinary Least Squares Method:
$e_1^2 + e_2^2 + \_\_\_ + e_n^2$ = RSS (Residual Sum Of Squares)
$RSS = (Y_1-\beta_0-\beta_1X_1)^2 + (Y_0-\beta_0-\beta_1X_2)^2 + \_\_\_\_ + (Y_n-\beta_0-\beta_1X_n)^2$
$RSS = \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1X_i)^2$

Fig: Estimating the RSS in a simple linear regression

The most commonly used method to determine the strength of fit of the model is the $\boldsymbol{R^2}$ or the **coefficient of determination**. R^2 tells us the percentage of variance of the dependent variable that can be explained by the model's predictor variables. The formula for $\boldsymbol{R^2}$ can be written as:

$\boldsymbol{R^2}$ **= 1-RSS/TSS** where TSS is the Total Sum of Squares. Generally, higher the R^2, better the model fit. But it is not always true. High R^2 value can also result from **overfitting** which happens when the model performs very well on the training set by remembering all the datapoints from the training set and fails to perform well with the unknown test set. Now what do we mean by training and test sets?

Training and Test Sets

We divide the data in hand into two parts- **Training and Test Set**. Training set is built using 70-75 % of the data and we try to fit a model with it. The Test Set is the data that our model, based on the training set is being tested on. We say, our model performance is good if we get an R^2 value of around 80% for both the training and the test sets using the same model.

But before going to test our model for the test set, we need to satisfy a series of questions at hand.

Let's discuss from the point of view of a multiple linear regression as simple linear regressions are rarely applicable in real life.

Preprocessing before Splitting the Data

First, before we split our data into the training and test sets, we need to do a few steps of preprocessing like dummy encoding for the categorical feature variables and scaling of the continuous feature variables. Not going into too much detail about the dummy encoding, we just need to be sure that our **dummy encoding** is done by dropping one value of that variable for which we are using dummy. For example, if we are doing a dummy encoding for a categorical variable 'season' which has four distinct levels- 'summer', 'monsoon', 'autumn', 'winter', we need to make sure that we have 3 dummy variables to avoid a dummy variable trap. Now, for the continuous variables, we need to do scaling to bring all of them to a common **scale** i.e. if we have 2 continuous variables, 'number of bathrooms in the flat' and 'area of the flat', normally their scales would be different and running a regression without making any change will result in a biased improper regression. After we are done with these steps, we divide the data into training and test sets.

Adjusted R-squared or R-squared?

Now, for a multiple linear regression, we check the adjusted R-squared value instead of the simple R-squared as adjusted R-squared **penalizes** the model if it keeps on adding redundant variables whereas R-squared only increases as we add more variables, not a realistic scenario.

Model Building

In the training set, we keep on adding variables and see for the adjusted R-squared values. We then build a model with all the regressor variables and check for their p-values and the F-value in the summary statistics after running the regression. The F-statistic gives us an idea of the overall fit of the model. If the value is negligibly small, we can assert that our model is a good fit. Now, even if the adjusted R-squared value is high and the F-statistic tells us that our model is a good fit, it may happen that the p-values of the model for some of the variables is high. But before looking into the p-values, we need to check for multicollinearity amongst the predictor variables. We use the concept of Variance Inflation Factor, VIF for that purpose. When we run a VIF for a particular variable, that variable is taken as the dependent variable and all the other variables are regressed against it. If there is multicollinearity, the R-squared value in this regression will be high as that would mean that the predictor variables can potentially explain the variance in the dependent variable in this regression. So, by the formula,

$$\textbf{VIF} = {}^{\boldsymbol{1}}\!/_{\boldsymbol{1 - R_i^2}},$$

we can see that increasing the R^2 value when we take the VIF for the i^{th} variable will only increase the VIF and confirm multicollinearity. There may be other variable also with high VIF's, but it may be because of the existence of the variable with the highest VIF.

Now, while dropping the variables one by one, we need to check the combination of its p-value and VIF score. Generally, the **rule** is:

i) If the p-values and the VIF both are high, we will surely drop that variable.
ii) If the VIF is low but still the p-value is high, that variable will be dropped next.
iii) Next comes the turn for the variables with low p-values but high VIF's (>5 in most cases. Variables with VIF score more than 10 must be dropped).
iv) We can retain the variables with both low p-values and low VIF.

We need to make sure that we drop the variables one by one, in tandem with the above logic and get to our final model where we will have variables with low p-values, a considerably good adjusted R-squared value (around 80%) and VIF's <=5.

After performing these steps, we need to do a residual analysis of the train data where by the assumptions of linear regression, we need to make sure that the error terms are normally distributed and homoscedastic.

We make predictions on the test set after this step using the model that we have finalized with the train set. If we get a R-squared value between the actual Y and the predicted Y around the R-squared value we got in our final regression using the training set, we can say that our model is a fine predictor of our unknown test set and we can go ahead with the mode. The coefficients we use for our model equation are the coefficients we obtain from the final regression on our training set.

2) Explain the Anscombe's quartet in detail

- Anscombe's quartet was developed by statistician Francis Anscombe consisting of four datasets, each containing eleven (x, y) pairs. All these datasets have the same descriptive statistics. The following table shows us the four sets of data and computation of some of the basic statistics and how they are exactly the same in all the cases.

	I		II		III		IV	
	X	Y	X	Y	X	Y	X	Y
	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
SUM	99.0	82.5	99.0	82.5	99.0	82.5	99.0	82.5
AVERAGE	9.0	7.5	9.0	7.5	9.0	7.5	9.0	7.5
STD DEV	3.3	2.0	3.3	2.0	3.3	2.0	3.3	2.0
CORR COEFF	0.82		0.82		0.82		0.82	

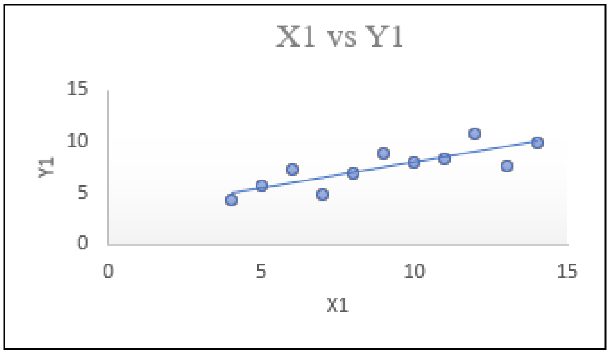
Observations

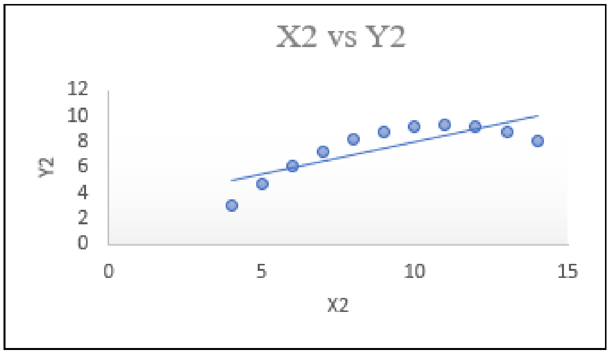
i) Mean of x = 9, Mean of y =7.5 for all the four datasets

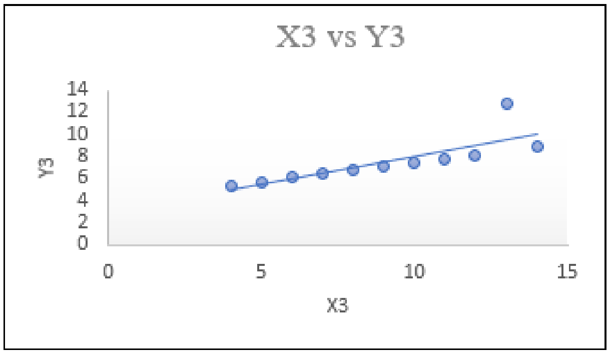
ii) Standard deviation of x = 3.3, Standard deviation of y = 2 for all the four datasets

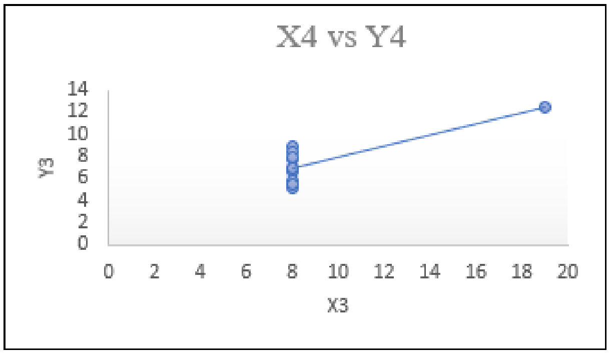
iii) The correlation coefficient for each pair of x and y is 0.82.

With the same descriptive statistics, these four datasets follow the same regression line on the scatter plot but each one of them tells us a different story.









i) The first scatter plot depicts a simple linear relationship.

ii) The second plot is not normally distributed. Linear regression is not applicable and the Pearson's correlation coefficient is not relevant.

iii) The third graph has a linear distribution, but the presence of an outlier lowers the correlation coefficient to a great extent.

iv) Just one outlier in the fourth visual pulls up the correlation coefficient to such an extent that it becomes comparable with the other three graphs, despite the other data points not showing any sort of relationship.

Anscombe's quartet depicts the importance of visualization in analyzing a dataset as a descriptive statistics summary can be misleading many a times.

3) What is Pearson's R?

- Pearson's correlation coefficient, R is a measure of the strength of association between two quantitative, continuous variables, for example, age and height.

To calculate Pearson's R, we first need to ensure that the two continuous variable follow a linear relationship using a scatter plot. A straight line is then fitted on the scatter plot so that the distances between the datapoints and the straight line is minimum. The closer the datapoints are to the straight line, higher the value of R.

The value of R ranges from -1 to 1, with -1 denoting a perfect straight line connecting all the points with a negative slope and a +1 denoting a perfect straight line connecting all the datapoints with a positive slope. R= 0 would mean that there is no linear relationship between the variables. Pearson's R for a pair of random variables (x, y) can be calculated as cov (x, y)/$\sigma_x \sigma_Y$ where cov is the covariance and σ_x, σ_y are the standard deviations of x and y respectively.

Assumptions while calculating Pearson's R would be:

i) Both the variables should be normally distributed.

ii) There should not be any significant outliers. Presence of an outlier may pull the Pearson's R value to a great extent as it is very sensitive to outliers.

iii) Each variable should be continuous and must have a linear relationship.

iv) The observations are paired i.e. same number of observations for both the variables we are calculation the correlation on.

However, correlation can be misleading sometimes as just because two variables have a high strength of association, we cannot conclude in some cases that one directly causes the other. To assert logically, correlation does not always imply causation.

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

- Scaling is a technique that is used to bring the independent variables within a comparable range. It is a part of data preprocessing before performing a regression.

The values for the independent variables in a regression can vary to a great extent, for example, if we are predicting price of a new flat, factors such as area and number of rooms will come into question. Now that area is measured in square feet and number of rooms is generally a smaller number, running a regression with the actual values will lead to misinterpretation of the model as a whole. We need to perform scaling to bring these variables within a common smaller range.

There are basically two types of scaling broadly used:

i) Normalization: Rescaling the values in a range of [0,1]. Also known as min-max scaling.

$$x' = \frac{x-x_{\min}}{x_{max}-x_{min}}$$

ii) Standardization: Rescales the data so that the values have a mean (μ) of 0 and a standard deviation (σ) of 1. It is not affected by outliers.

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

Now, when should we use normalization and standardization? We can use normalization when the distribution of the data doesn't follow a normal distribution like in the case of K-Nearest Neighbors and standardization when the distribution is Gaussian or normal.

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

- Variance Inflation Factor is used to check multicollinearity in a linear regression. If the VIF for a variable is high (above 5), we normally drop that variable and re-run the regression. However, in some cases, we might notice that value of VIF is exceptionally large. Now, the formula for VIF:

$$\textbf{VIF} = {}^{\mathbf{1}}/_{\mathbf{1}-\boldsymbol{R_i^2}},$$

we can assert that higher the value of $\boldsymbol{R_i^2}$**,** higher the value of VIF. Now, when we check for VIF of a variable, we regress all the other explanatory variables in the model against it. i.e. we take the variable in question to be a dependent variable and run a regression. The model will provide us a value of R^2.A higher R^2 would mean that the variance in the dependent variable can be explained to a great extent by the regressor variables, i.e. the variables have a linear relationship amongst them and any one or more of the regressor variables can successfully explain the dependent variable in this regression. So, there can be cases where the value of R^2 is so high (nearly 100%) that the VIF tends to infinity. There is nothing to worry about this as this would mean that the **variable in question can be exactly replicated by a linear combination of the other explanatory variables.** We surely need to drop one or some independent variables from the model as one or more of them are redundant and must be removed to avoid this high multicollinearity.

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

- Q-Q plot or quantile-quantile plot is a scatter plot that is used to determine whether two data sets belong to a common distribution. Using a 45-degree reference line, the statement is tested. If the two data sets are representatives of populations with similar distributions, the points should fall approximately along the reference line. The greater the deviation from this reference line, the greater the chance of the data sets coming from differently-distributed populations. If the distributions have a linear relationship and are not exactly same, the points in the Q-Q plot will lie on a straight line but not on the 45-degree line.

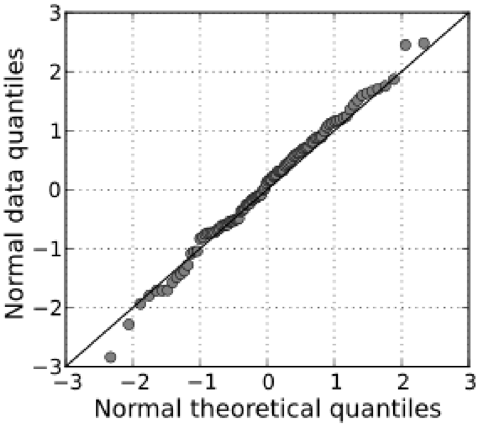


Fig: Plotting quantiles from sample data against a Normal distribution

Q-Q plot sorts our data in ascending order and plots them versus quantiles obtained from a theoretical distribution. The number of quantiles is selected to match the size of our data. Q-Q plots can be used for any distribution though it is mostly used with Normal distribution because of its widespread use. If the quantiles of our sample data do not lie on the 45-degree line, we can infer that the data follows a skewed distribution.