

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?

Ans: The Linear regression equation came out to be:

$$\text{cnt} = 0.35 + 0.25 \cdot \text{yr} + 0.48 \cdot \text{atemp} - 0.14 \cdot \text{hum} - 0.16 \cdot \text{windspeed} - 0.10 \cdot \text{spring} + 0.04 \cdot \text{winter} - 0.24 \cdot \text{Light rain} - 0.05 \cdot \text{misty weather} - 0.04 \cdot \text{Sun} - 0.04 \cdot \text{Jan} - 0.07 \cdot \text{Jul} + 0.06 \cdot \text{Sep}$$

In the above equation apart from variables yr, atemp, hum, windspeed the rest are categorical variables.

The variable Light rain, Spring has significant impact on y(target variable). For 1 unit change in Light rain there is 0.24 unit change in the target variable. For 1 unit change in spring there is 0.1 unit change in cnt.

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

Ans: For a Linear regression to happen we need data points for dependent variable as well as predictor variables. So we need to convert categorical variables to numerals.

If there are 'n' values for a categorical variable, then for it to be represented in numerals we need to convert it to n columns. And to identify 'n' values we effectively need (n-1) values as the absence of all (n-1) variables can be treated as the nth value. So we can drop any one of n values. Here we are dropping the first one to make it into (n-1) columns.

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

Ans: atemp followed by temp has high correlation with target variable(cnt).

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

Ans:

1) The error variable is normally distributed and by plotting it.

When the (y_predicted from the final model - y value from train) is plotted, we could see that they are normally distributed around mean 0. Our assumption is true and also validated.

2) Constant variance of error terms.

This is validated by plotting the y_test and y_predicted from the final model and they are almost in a straight line. So the variance is not changing.

3) The predictor variables are not interdependent.

By removing the highly correlated variables and the variables with high VIF we make sure this assumption is validated and here it's done before arriving at the final model.

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

Ans: $Y = 0.35 + 0.25 \cdot \text{yr} + 0.48 \cdot \text{atemp} - 0.14 \cdot \text{hum} - 0.16 \cdot \text{windspeed} - 0.10 \cdot \text{spring} + 0.04 \cdot \text{winter} - 0.24 \cdot \text{Light rain} - 0.05 \cdot \text{misty weather} - 0.04 \cdot \text{Sun} - 0.04 \cdot \text{Jan} - 0.07 \cdot \text{Jul} + 0.06 \cdot \text{Sep}$

'atemp', 'humidity' and 'Light rain ' (whether it's raining or not).

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: Linear regression is one of the supervised learning methods where we depend on past data. It is an analysis mechanism where we try to predict the outcome of a variable (target variable) based on the behavior of other variables (predictor variables). In machine learning we try to fit the best possible linear equation which can explain the relation between the target variable and predictor variables.

A general Linear Regression equation looks like this,

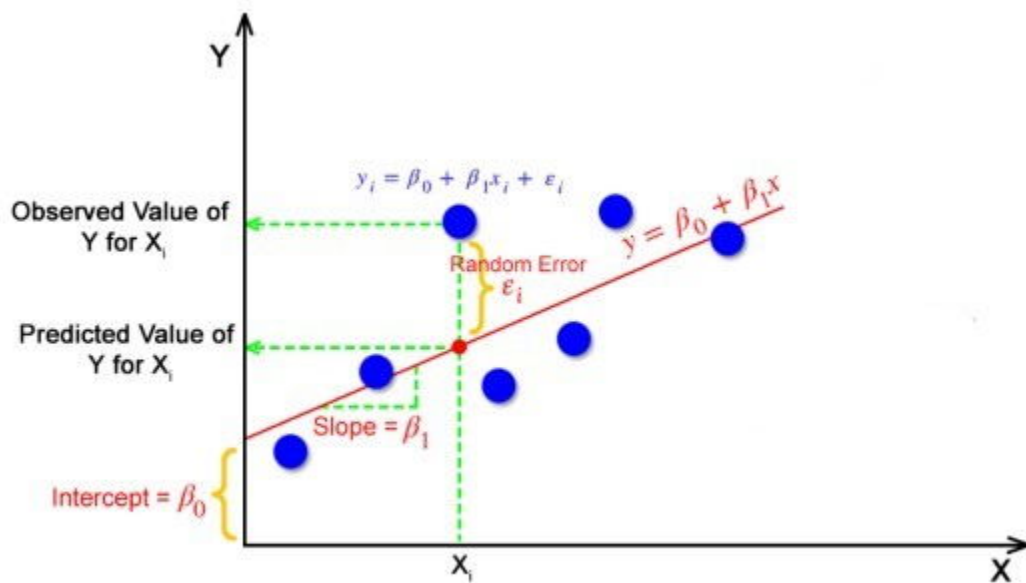
$$y = c_0 + c_1x_1 + c_2x_2 + c_3x_3 + \dots + c_nx_n.$$

Here Y is the item which we want to predict based on the behavior of $x_1, x_2, x_3, \dots, x_n$.

C_0 is the intercept/constant and $c_1, c_2, c_3, \dots, c_n$ are the coefficients of the variables.

Our attempt is to find the best possible values of the intercept and coefficients based on the behavior of the predictor variables. The best fit is the line with the least possible error values.

Below image is an example for one predictor variable.



The blue dots represent the data points and the distance between the blue point and the line along the Y-axis is the error. The line with the least sum of errors is the best possible line.

$$\sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2$$

Sum(Y actual - Y predicted from the Linear regression model).

There are certain assumptions while performing linear regression.

1) Linearity: The predictor variables are linearly related to the target variable.

This can be validated by plotting the variables against the target variable one by one.

2) Homoscedasticity: The variance of the error terms should be constant.

This can be validated by plotting the residual error terms, and there should not be any particular pattern/skew in the distribution.

3) Distribution of error terms: When plotted, the error terms should be normally distributed around 0.

4) No multicollinearity: There should be any collinearity between the predictor variables, this can be identified by using Pearson's R or by plotting VIF.

Executing the linear Regression: As mentioned earlier, the target variable is predicted by observing the behavior of the predictors variables from past data.

The past data is taken and is split into train data set and test data set in the ratio 70:30 or 80:20 or 90:10 respectively.




By using the train dataset we build the model, meaning we try to find the best fit line. Then we use this line/model to predict on the test dataset.

Factors which helps us evaluate if a model is worthy or no:

1) R squared: The most commonly used metric for model evaluation in regression analysis is R squared. It can be defined as a Ratio of variation to the Total Variation. The value of R squared lies between 0 to 1, the value closer to 1 the better the model.

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$
$$1 - (RSS/TSS)$$

2) Adjusted R squared: As the number of predictors increase the r2 value always increases and this gives an inflated value of r2 and might give false confidence, so this Adjusted r2 takes into account the n (number of points in dataset) and k (number of independent predictors). The nearer the value to 1, the better is the model.


$$\text{Adjusted R Squared Formula} = 1 - \left[\frac{(1 - R^2) \times (n - 1)}{(n - k - 1)} \right]$$


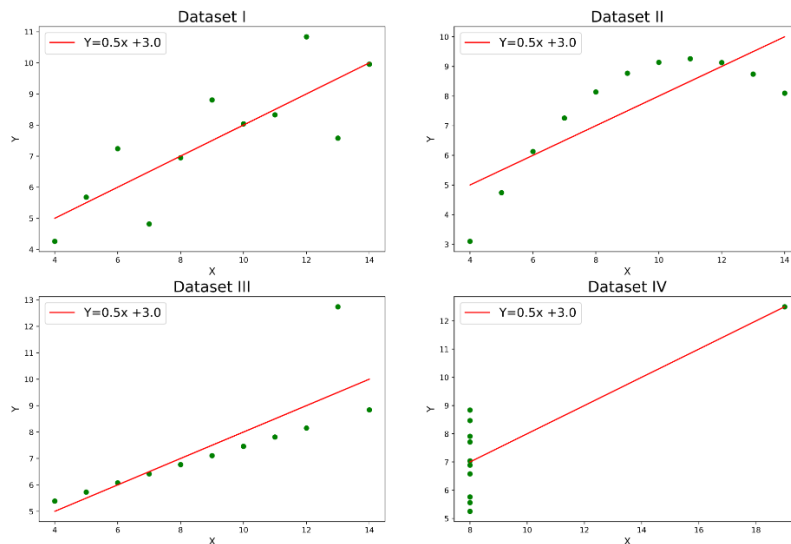
2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet is a set of four datasets with identical summary statistics; the mean, variance, r^2 , correlation coefficient between x and y , Linear regression equation but different data points when we plot them on graph.

The four datasets comprise of 11 x-y pairs.

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

When we find the best fit line for all these 4 data sets the Linear regression line comes out to be almost the same, but when these data points are plotted on graphs we see a different story.



Dataset 1: LR can be used to explain the data points.

Dataset 2: X-Y are not linearly related.

Dataset 3: Outliers are not properly fit.

Dataset 4: One single point is affecting the coefficient and deviating entirely.

This experiment clearly explains that it's always best to do EDA and see the distribution of data, rather than solely relying on summary statistics.

3. What is Pearson's R?

Ans: Pearson's R is a way of measuring the linear correlation between variables.

It is the ratio between the covariance of two variables and the product of their standard deviations. It gives a number between 1 and -1. Correlation coefficient of 1 and -1 means both variables are very strongly correlated.

Eg: $a = b$, $a = 0.2b$. Here 'a' and 'b' are positively correlated. $a = -b$, $a = -0.5b$. Here 'a' and 'b' are negatively correlated.

And a correlation coefficient of 0 means the variables are not in any way related at all. Change in one variable, there is no change in the other variable.

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

Ans: Scaling is performed to improve model performance and for faster convergence. We almost always run multiple Linear regression models before finalizing a model. If the values are not scaled and if they have large differences, then it will affect the speed of convergence (time to arrive at predictor variables).

Also if the predictors are not scaled, then the respective coefficients might vary largely in numeric value.

Eg: for a scaled model the LR can look like $y = 10*a + 9*b + c_1$ and the same equation for an unscaled model looks like $y = 1000*a + 9*b + c_2$. Plotting and analyzing the first equation is much easier than the second equation, though both are for the same dataset.

Normalized scaling converts every value to less than 1 and greater than 0. It is obtained by the formula, $\text{value} = (x - \min) / (\max - \min)$.

Standardization scaling uses 'mean' to express every other value. It is obtained by the formula, $\text{value} = (x - \text{mean}) / \text{standard deviation}$.

Normalization is less sensitive to outliers whereas Standardization is more sensitive to outliers. The shape of the distribution is preserved in Normalization, the shape of distribution changes in Standardization.

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

Ans: A very high correlation among variables gives a high VIF. If a variable has a high Variance Inflation Factor then it means that variable can easily be explained by all/some of the other variables.

That variable with high VIF has a strong linear relation with other variables. VIF is more like a cross product between the predictors, a VIF of 1.0 implies there is no relation between the predictor variables.

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

Ans: Q-Q plot stands for Quantile-Quantile plot. As the name suggests its a graph which maps the quantile of one data distribution to the respective quantile in another. It can be used to find the type of distribution of a dataset; whether it's a uniform distribution, normal distribution etc...

Eg: If we are trying to find if the data set is normally distributed or not.

If there are 'n' data points we divide the distribution into 'n' quantiles and the area under the normal distribution is divided into (n-1) regions and identify each quantile with its z-value.

Map 'n' points from the dataset with normal distribution 'n' points. If all the points lie on the $y=x$ line then the dataset is normally distributed.

In this way we can find the type of distribution of any data set.

In Linear regression if we are given two datasets, it's always best to check if they are from the same population or not. By using Q-Q plot we can plot the quantiles of the train set with the respective quantiles of the test set and if they are appearing on the $y=x$ line.

One of the assumptions of Linear regression is that the residuals or the error terms follow normal distribution. We can validate the assumption by using the Q-Q plot.