Assignment Linear Regression

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?

    1. The categorical variables available in the assignment are “season”, “workingday”, “weathersit”, “weekday”, “yr”, “holiday”, and “mnth”. Based on EDA and Linear Regression we can give few information as below:
    2. “season” –
        a. Based on the data available, the most favourable seasons for biking are summer and fall. We can run good campaign for season of summer to grow the demand.
        b. Spring has significant low consumption ratio.
    3. “workingday” –
        a. Working day represents weekday and weekend/holiday information.
        b. Registered and casual users’ identity and relevant strategy for working and not working days shall help to increase the numbers.
        c. Working Day shows that whether that day is not having holiday or weekend.
        d. Registered users are tend to rent bikes on working day. In contrast, casual demand is high on weekends. Company should come up with good strategy to convert that casual demand to registered one as well by offering them sign up discount.
    4. “weathersit” –
        a. Most favourable weather condition is the clean/few clouds days.
        b. Registered users count is comparatively high even on the light rainy days, so the assumption can be drawn that the bikes are being used for daily commute to the workplace.
        c. There is no data available for heavy rain/snow days. One assumption could be there is no demand when that is the case or may be data is not complete!
    5. “weekday” –
        a. If we consider “cnt” column we do not find any significant pattern with the weekday.
        b. However if the relation is plotted with “registered” users, we observe that bike usage is higher on working days. And with “casual” users it opposite.
    6. “yr” –
        a. 2 years data is available and the increase in the bikes has increased from 2018 to 2019. We see demand keep on increasing YoY(Year on Year).
    7. “holiday” –
        a. On Holiday does not provide good prediction with Total Demand.
    8. “mnth” –
        a. The bike rental ratio is higher for June, July, August, September and October months.

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

When we create dummy variables for categorical variable each dummy variable have 1 and 0 values. 1 is used to depict the presence and 0 for absence of the respective category. This means if the category variable has 3 categories, there will be 3 dummy variables. This process also called as One Hot Encoding.

The drop_first = True is used while creating dummy variables to drop the base/reference category. The reason for this is to avoid the multi-collinearity getting added into the model if all dummy variables are included. The reference category can be easily identified where 0 is present in a single row for all the other dummy variables of a particular category.

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

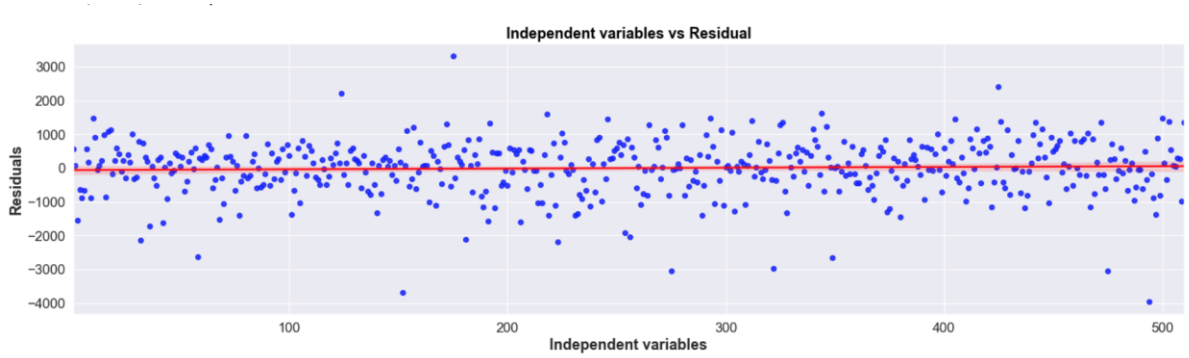
- "temp" is the variable which has the highest correlation with target variable i.e. 0.63.
- The casual and registered variables are actually part of the target variable as values of these columns sum up to get the target variable, hence ignoring the correlation of these 2 variables.
- "atemp" is the derived parameter from temp, humidity and windspeed, hence not considering it as it is eliminated in the model preparation.

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

- X and y should have linear relationship: Linear relationship between independent and dependent variables – The linearity is validated by looking at the points distributed symmetrically around the diagonal line of the actual vs predicted plot as shown in the below figure. We do that by finding Pearson correlation.

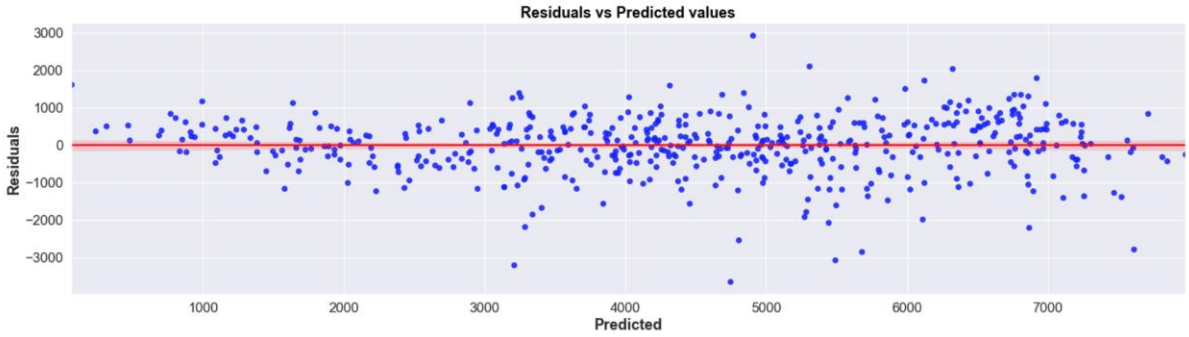
	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
season	1.00	-0.00	0.83	-0.01	-0.00	0.01	0.02	0.33	0.34	0.20	-0.21	0.21	0.41	0.40
yr	-0.00	1.00	-0.00	0.01	-0.01	-0.00	-0.05	0.05	0.05	-0.13	-0.03	0.25	0.60	0.57
mnth	0.83	-0.00	1.00	0.02	0.01	-0.00	0.05	0.22	0.23	0.22	-0.19	0.12	0.29	0.28
holiday	-0.01	0.01	0.02	1.00	-0.10	-0.25	-0.03	-0.03	-0.03	-0.02	0.02	0.05	-0.11	-0.07
weekday	-0.00	-0.01	0.01	-0.10	1.00	0.04	0.03	-0.00	-0.01	-0.04	0.00	0.06	0.06	0.07
workingday	0.01	-0.00	-0.00	-0.25	0.04	1.00	0.06	0.05	0.05	0.02	-0.01	-0.52	0.31	0.06
weathersit	0.02	-0.05	0.05	-0.03	0.03	0.06	1.00	-0.12	-0.12	0.62	0.07	-0.25	-0.26	-0.30
temp	0.33	0.05	0.22	-0.03	-0.00	0.05	-0.12	1.00	0.99	0.13	-0.13	0.54	0.54	0.63
atemp	0.34	0.05	0.23	-0.03	-0.01	0.05	-0.12	0.99	1.00	0.14	-0.16	0.54	0.54	0.63
hum	0.20	-0.13	0.22	-0.02	-0.04	0.02	0.62	0.13	0.14	1.00	-0.19	-0.09	-0.11	-0.12
windspeed	-0.21	-0.03	-0.19	0.02	0.00	-0.01	0.07	-0.13	-0.16	-0.19	1.00	-0.14	-0.20	-0.21
casual	0.21	0.25	0.12	0.05	0.06	-0.52	-0.25	0.54	0.54	-0.09	-0.14	1.00	0.39	0.67
registered	0.41	0.60	0.29	-0.11	0.06	0.31	-0.26	0.54	0.54	-0.11	-0.20	0.39	1.00	0.95
cnt	0.40	0.57	0.28	-0.07	0.07	0.06	-0.30	0.63	0.63	-0.12	-0.21	0.67	0.95	1.00

- Error terms are independent of each other i.e. No Autocorrelation – The dependence of the error terms means that the model is not able to predict the errors and hence there is possibility of missing predictor and we should find one and re-build our model.



- Error terms are normally distributed: Histogram and distribution plot helps to understand the normal distribution of error terms along with the mean of 0. The figure below clearly depicts the same.

- Error terms have constant variance (homoscedasticity): The errors plotted against the fitted variable shows if there is a constant variance in the error or not. The constant variance helps to confirm the homoscedasticity of the model. The below figure clearly depicts the model is homoscedastic.



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 variables are:

- "yr" (2060) - Year – The growth year on year seems organic given the geological attributes.
- "atemp" (819)- as temperature increases by 1 standard deviation then Total demand increases by 819 units.
- "season" (587) – Winter season is playing the crucial role in the demand of shared bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

- Linear regression is the method of finding the best linear relationship within the independent variables and dependent variables.
- The algorithm uses the best fitting line to map the association between independent variables with dependent variable.
- There are 2 types of linear regression algorithms
    - Simple Linear Regression – Single independent variable is used.
        - $Y = \beta_0 + \beta_1 X$ is the line equation used for SLR.
    - Multiple Linear Regression – Multiple independent variables are used.
        - $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \in$ is the line equation for MLR.
    - $\beta_0 = value\ of\ the\ Y\ when\ X = 0\ (Y\ intercept)$
    - $\beta_1, \beta_2, \ldots, \beta_p = Slope\ or\ the\ gradient.$
- Cost functions – The cost functions helps to identify the best possible values for the $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ which helps to predict the probability of the target variable. The minimization approach is used to reduce the cost functions to get the best fitting line to predict the dependent variable. There are 2 types of cost function minimization approaches – **Unconstrained and constrained**.
    - Sum of squared function is used as a cost function to identify the best fit line. The cost functions are usually represented as
        - The straight-line equation is $Y = \beta_0 + \beta_1 X$
        - The prediction line equation would be $Y_{pred} = \beta_0 + \beta_1 x_i$ and the actual Y is as Y_i.
        - $Now\ the\ cost\ function\ will\ be\ J(\beta_1, \beta_0) = \sum(y_i - \beta_1 x_i - \beta_0)^2$
    - The unconstrained minimization are solved using 2 methods
        - Closed form
        - Gradient descent
- While finding the best fit line we encounter that there are errors while mapping the actual values to the line. These errors are nothing but the residuals. To minimize the error squares OLS (Ordinary least square) is used.
    - $e_i = y_i - y_{pred}$ is provides the error for each of the data point.
    - OLS is used to minimize the total e^2 which is called as Residual sum of squares.
    - $\text{RSS} = = \sum_{i=1}^{n} (y_i - y_{pred})^2$

- Ordinary Lease Squares method is used to minimize Residual Sum of Squares and estimate beta coefficients.

---

2. Explain the Anscombe's quartet in detail.

Statistics like variance and standard deviation are usually considered good enough parameters to understand the variation of some data without actually looking at every data point. The statistics are great to for describing the general trends and aspects of the data.

Francis Anscombe realized in 1973 that only statistical measures are not good enough to depict the data sets. He created several data sets all with several identical statistical properties to illustrate the fact.

1. Illustrations
    a. One of the data set is as follows:

Anscombe's quartet

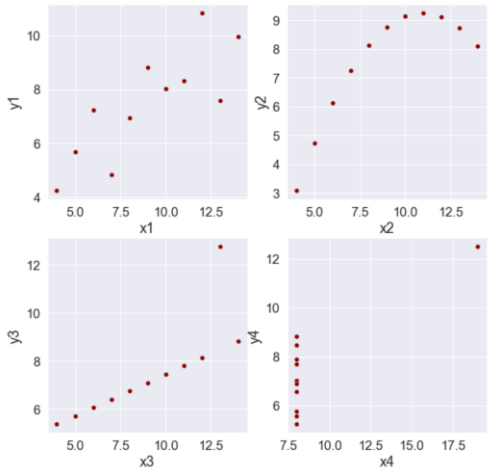
I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08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

(reference: https://en.wikipedia.org/wiki/Anscombe%27s_quartet)

    b. If the descriptive statistics are checked for above data set then all looks same:

Property	Value	Accuracy
Mean of *x*	9	exact
Sample variance of *x* : s_x^2	11	exact
Mean of *y*	7.50	to 2 decimal places
Sample variance of *y* : s_y^2	4.125	±0.003
Correlation between *x* and *y*	0.816	to 3 decimal places
Linear regression line	*y* = 3.00 + 0.500*x*	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression : R^2	0.67	to 2 decimal places

c. However, when plotted these points, the relation looks completely different as depicted below.



2. Anscombe's Quartet signifies that multiple data sets with many similar statistical properties could still be different from one another when plotted.
3. The dangers of outliers in data sets are warned by the quartet. Check the bottom 2 graphs. If those outliers would have not been there the descriptive stats would have been completely different in that case.
4. Important points
    a. Plotting the data is very important and a good practice before analysing the data.
    b. Outliers should be removed while analysing the data.
    c. Descriptive statistics do not fully depict the data set in its entirety.

---

3. What is Pearson's R?

The Pearson's R (also known as Pearson's correlation coefficients) measures the strength between the different variables and the Linear relation with each other. The Pearson's R returns values between -1 and 1. The interpretation of the coefficients are:

Pearson's R is very useful metric to identify how two variable and at what strength they are linearly related.

1. *-1 coefficient indicates strong inversely proportional relationship.*
2. *0 coefficient indicates no relationship.*
3. *1 coefficient indicates strong proportional relationship.*

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

Where:

N = the number of pairs of scores

Σxy = the sum of the products of paired scores

Σx = the sum of x scores

Σy = the sum of y scores

$Σx^2$ = the sum of squared x scores

Σy^2 = the sum of squared y scores

---

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

- The scaling is the data preparation step for regression model. The scaling normalizes these varied datatypes to a particular data range.
- Most of the times the feature data is collected at public domains where the interpretation of variables and units of those variables are kept open collect as much as possible. This results into the high variance in units and ranges of data. If scaling is not done on these data sets, then the chances of processing the data without the appropriate unit conversion are high. Also, the higher the range then higher the possibility that the coefficients are impaired to compare the dependent variable variance.
  The scaling only affects the coefficients. The prediction and precision of prediction is unaffected after scaling.
- Usually Scaling is done to find out ranking of X variables beta coefficients w.r.t y variable.
- Normalization/Min-Max scaling – The Min max scaling normalizes the data within the range of 0 and 1. The Min max scaling helps to normalize the outliers as well.

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

- Standardization converges all the data points into a standard normal distribution where mean is 0 and standard deviation is 1.

$$Standardization: x = \frac{x - mean(x)}{sd(x)}$$

---

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

$$VIF = \frac{1}{1 - R^2}$$

If the R^2 is 1 then the VIF is infinite. The reason for R^2 to be 1 is that there is a perfect correlation between one or more than one independent variables with Dependent Variable.

---

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

Q-Q plots are the quantile-quantile plots also called as probability plot. It is a graphical method comparing two probability distributions by plotting their quantiles against each other. The theoretical distributions could be of type normal, exponential or uniform. The Q-Q plots are useful in the linear regression to identify the train data set and test data set are from the populations with same distributions. This is another method to check the normal distribution of the data sets in a straight line with patterns explained below

- Interpretations
  - Similar distribution: If all the data points of quantile are lying around the straight line at an angle of 45 degree from x-axis.
  - Y values < X values: If y-values quantiles are lower than x-values quantiles.
  - X values < Y values: If x-values quantiles are lower than y-values quantiles.
  - Different distributions – If all the data points are lying away from the straight line.

- The plot has a provision to mention the sample size as well.