Bike Sharing Assignment

Assignment-based/General Subjective Questions

Arjun Khanna

Question #1

From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

- W.r.t seasons, during ’fall’, the demand was the highest, followed by summer and winter. The least demand was in spring.
- W.r.t. weathersit, demand is high during ‘clear_few clouds’, followed by ‘mist_cloudy’. The demand is very poor during ‘Light rain_Light snow_Thunderstorm’
- Demand does not fluctuate greatly on any day of the week.
- *Demands increases in the month of 3, 5, 6, 8 ,9, 7 , 10 and yr*
- *Demand decreases if it is holiday , Spring, Light rain_Light snow_Thunderstorm, Mist_cloudy, Sunday*

Question #2

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

- drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables. More importantly, it is always a good practice if you can achieve higher efficiency with lesser number of dimensions.
- Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

Question #3

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

- Temp, atemp has 0.63 correlation with target variable cnt.

Question #4

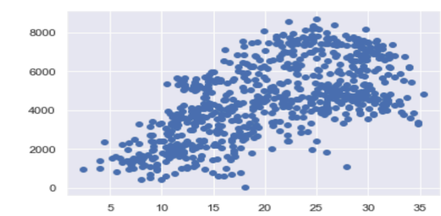
How did you validate the assumptions of Linear Regression after building the model on the training set?

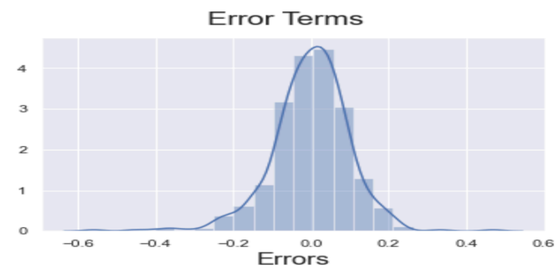
- **Linear Relationship –** use Pair-wise scatterplots in validating the linearity assumption as it is easy to visualize a linear relationship on a plot.
- **Normality of errors -** draw the distribution of residuals against levels of the dependent variable to identify if Residuals are distributed normally. If the resulting curve is not normal (i.e. is skewed), it may highlight a problem.
- **Homoscedasticity –** use residual plot and verify that the variance of the error terms is constant across the values of the dependent variable.
- Absence of **Multicollinearity** – use heatmaps & VIF (Variance Inflation Factors < 5) to identify the presence of multicollinearity.

```
#Performing EDA
# 1)PAIRPLOTS TO UNDERSTAND NUMERICAL VARIABLES
```

```
plt.scatter('temp','cnt',data=df)
```

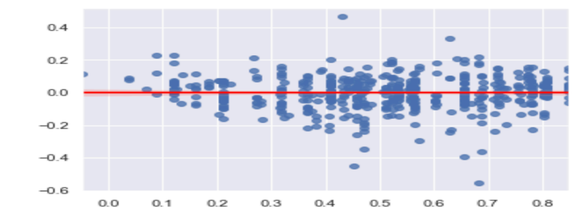
```
<matplotlib.collections.PathCollection at 0x7f93e42c35e0>
```

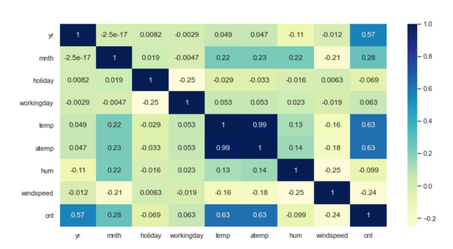





```
sns.regplot(x=y_train_cnt, y=res,line_kws={'color': 'red'})
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f93e7726070>
```





	Features	VIF
0	yr	1.68
2	spring	1.45
4	Mist_cloudy	1.41
5	3	1.23
12	10	1.17
8	8	1.14
10	Sunday	1.14
9	9	1.13
6	5	1.12
11	7	1.09
7	6	1.08
3	Light rain_Light snow_Thunderstorm	1.06
1	holiday	1.03

Question #5

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

- cnt= 0.246 X yr - 0.083 X holiday - 0.198 X Spring - 0.321 X Light rain_Light snow_Thunderstorm - 0.090 X Mist_Cloudy +0.063 X 3 + 0.123 X 5 +0.148 X 6 +0.153 X 8 + 0.193 X 9 -0.049 X Sunday + 0.126 X 7 + 0.116 X 10

- **Demands increases in the month of 3, 5, 6, 8 ,9, 7 , 10 and yr**
- **Demand decreases if it is holiday , Spring, Light rain_Light snow_Thunderstorm, Mist_cloudy, Sunday.**

General Subjective Question #1

Explain the linear regression algorithm in detail.

- Linear regression is used for finding linear relationship between target and one or more predictors. There are two types of linear regression- Simple and Multiple.
- Simple linear regression is useful for finding STATISTICAL relationship between two continuous variables. One is a predictor or independent variable and other is dependent variable.
- The core idea is to obtain a line that best fits the data. The best fit line is the one for which total prediction error (all data points) are as small as possible. Error is the distance between the point to the regression line. The standard equation of the regression line is given by the following expression: $Y = \beta_0 + \beta_1 X$
- Residual Analysis - Consider, we have a dataset which predicts sales of juice when given a temperature of place. Value predicted from regression equation will always have some difference with the actual value. Sales will not match exactly with the true output value. This difference is called as residue.
- Residual plot helps in analyzing the model using the values of residues. It is plotted between predicted values and residue. Their values are standardized. The distance of the point from 0 specifies how bad the prediction was for that value. If the value is positive, then the prediction is low. If the value is negative, then the prediction is high. 0 value indicates prefect prediction. Detecting residual pattern can improve the model.
- Non-random pattern of the residual plot indicates that the model is,
  - Missing a variable which has significant contribution to the model target
  - Missing to capture non-linearity (using polynomial term)
  - No interaction between terms in model
- Characteristics of a residue: Residuals do not exhibit any pattern; Adjacent residuals should not be same as they indicate that there is some information missed by system.

General Subjective Question #1 (contd.)

Null-Hypothesis and P-value

Null hypothesis is the initial claim that researcher specify using previous research or knowledge.
Low P-value: Rejects null hypothesis indicating that the predictor value is related to the response
High P-value: Changes in predictor are not associated with change in target

Metrics for model evaluation: ***R-Squared value***

R2 is a number which explains what portion of the given data variation is explained by the developed model. This value ranges from 0 to 1. Value '1' indicates predictor perfectly accounts for all the variation in Y. Value '0' indicates that predictor 'x' accounts for no variation in 'y'. Overall, the higher the R-squared, the better the model fits your data. Mathematically, it is represented as: $R^2 = 1 - (RSS / TSS)$

1. Regression sum of squares (SSR) - This gives information about how far estimated regression line is from the horizontal 'no relationship' line (average of actual output).
2. Sum of Squared error (SSE) - How much the target value varies around the regression line (predicted value).
3. Total sum of squares (SSTO) - This tells how much the data point move around the mean.

Let's take a look at what the assumptions of simple linear regression were:
1. Linear relationship between X and Y
2. Error terms are normally distributed (not X, Y)
3. Error terms are independent of each other
4. Error terms have constant variance (homoscedasticity

General Subjective Question #1 (contd.)

- Multiple linear regression (MLR) is a statistical technique to understand the relationship between one dependent variable and several independent variables. The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X.
- The formulation for multiple linear regression is also similar to simple linear linear regression with the small change that instead of having beta for just one variable, you will now have betas for all the variables used. The formula now can be simply given as: *Y = B0 + B1X1 + B2X2 + …. + BpXp + E.* *Specific* Considerations for MLR:

a) Adding more isn't always helpful, as Model may 'overfit' by becoming too complex. Model fits the train set 'too well', doesn't generalize. Symptoms: high train accuracy, low test accuracy. Multicollinearity - Associations between predictor variables. Use pairwise plots, Heatmaps & VIF (Variance Inflation Factor) to detect Multicollinearity.

b) Feature Scaling: Another important aspect to consider is feature scaling. When you have a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret.

c) Handling Categorical Variables - One way to deal with them is creating dummy variables. The key idea behind creating dummy variables is that for a categorical variable with 'n' levels, you create 'n-1' new columns each indicating whether that level exists or not using a zero or one.

General Subjective Question #1 (contd.)

- Feature selection becomes an important aspect
- It is generally recommended that you follow a balanced approach, i.e., use a combination of automated (coarse tuning using Recursive Feature Elimination {RFE}) + manual (fine tuning) selection in order to get an optimal model.

- Since, a multiple linear regression can be built with different combinations of the variables present, model comparison and hence, selection of the best model becomes extremely essential. The key aspect while selecting the best model is the trade-off between selecting the model explaining the variance best and the model which is fairly simple. So to implement this idea, you need a few parameters apart from the original ones (like R-squared) that would test the goodness of the model as well as penalise the model for using more number of predictor variables.

General Subjective Question #2

UpGrad

Explain the Anscombe's quartet in detail.

- It was constructed in 1973 by statistician **Francis Anscombe** to illustrate the **importance** of **plotting the graphs** before analyzing and model building, and the effect of other **observations on statistical properties**.

Anscombe's Data									
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

Image by Author

- There are these four data set plots which have nearly **same statistical observations (refer table)**, which provides same statistical information that involves **variance**, and **mean** of all x,y points in all four datasets.
- However, when these data is graphed, each data set tells different stories (first one fits pretty well, while the second could not fit the linear regression well, while 3rd and 4th dataset shows outliers that cannot be handled by regressions.
- Data visualization is extremely important as modelling based on simple statistics can fool the model.

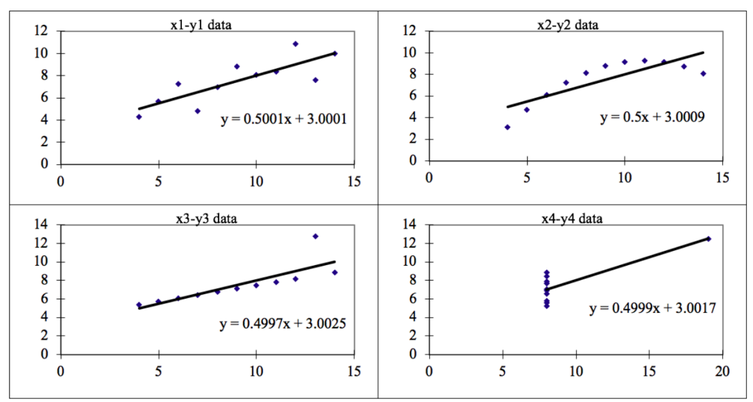
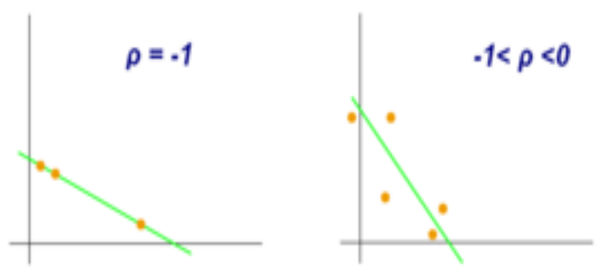


Image by Author

General Subjective Question #3

What is Pearson's R?

- Pearson's R measures the strength of relationship between two variables.



- It is always measured between -1 and 1, while 1 represents perfect positive corelation, -1 represents perfect inverse correlation and 0 represents no correlation, >0 represents positive correlation while <0 represents negative correlation.

General Subjective Question #4

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

- When you have a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret. So it is important to scale features because of two reasons:
    - Ease of interpretation & Faster convergence for gradient descent methods
- Normalization & standardization are two techniques used for feature scaling.
- Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.
- Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.
- It is recommended to use standardization when there are outliers, as standardization does not affect bounding range in your data.
- It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F statistic, p-values, R-square, etc.

General Subjective Question #5

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

- 'INFINITE' VIF represents perfect correlation between other variables. In other words, this indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an **infinite VIF** as well).
- Variables with infinite/high VIF must be removed from the model.

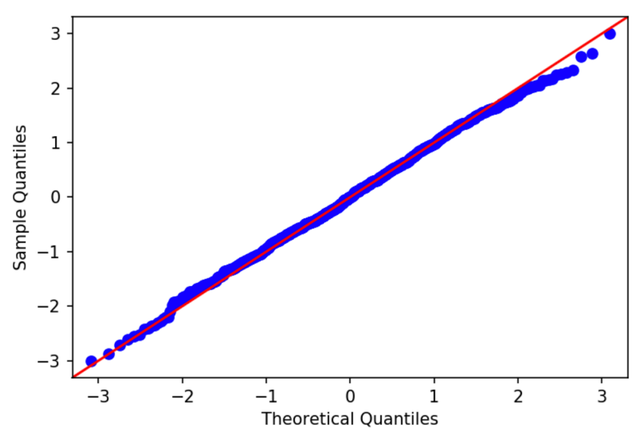
General Subjective Question #6

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

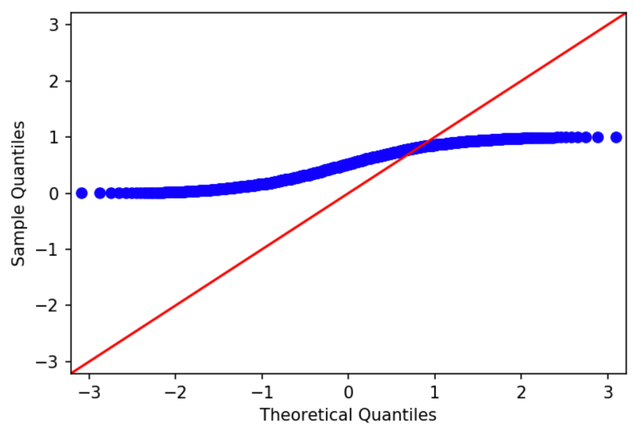
- Things that are normally distributed are great. Knowing that something conforms to the normal distribution (and knowing its mean and standard deviation) allows us to make all kinds of useful inferences about it. For example, we can be reasonably sure where its value will fall say 95% of the time (between -1.96 and +1.96 standard deviations of the mean). But if our variable is actually not normally distributed, then our inferences will be wrong, sometimes very wrong. And depending on the application, the consequences of our inaccurate inferences can range from being merely inconvenient to even dangerous.

- That's where QQ plots come in. They're a quick and visual way to assess whether a variable is normal or not. we can use QQ plots to check our data against any distribution, not just the normal distribution.

- If our data adheres to the red 45 degree line or close to it, it's normal and if it does not, then it's not a normal distribution.



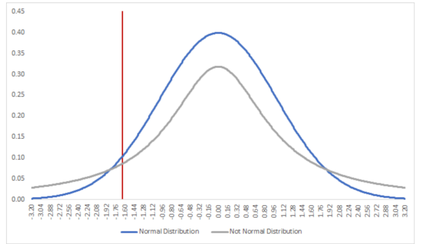
QQ plot of a normally distributed random variable



QQ plot of a random variable that is not normally distributed

General Subjective Question #6

- How QQ Plots Work - The “QQ” in QQ plot means quantile-quantile — that is, **the QQ plot compares the quantiles of our data against the quantiles of the desired distribution** (defaults to the normal distribution, but it can be other distributions too as long as we supply the proper quantiles).
- **Quantiles are breakpoints that divide our numerically ordered data into equally proportioned buckets.** For example, you’ve probably heard of percentiles before — percentiles are quantiles that divide our data into 100 buckets (that are ordered by value), with each bucket containing 1% of observations. Quartiles are quantiles that divide our data into 4 buckets (0–25%, 25–50%, 50–75%, 75–100%).
- **So if the distribution of our portfolio is actually the grey line, but we model it with the blue line, we will be significantly understating the frequency of a terrible outcome** (terrible outcomes are ones to the left of our threshold, the red line). We would be assuming that there is only a 5% chance of a terrible outcome, when in reality 17% of the area under the gray line (its cumulative density function) lies to the left of our terrible outcomes threshold.
- That’s why it’s important to check that something is normal. And that’s where QQ plots really shine. In essence, QQ plots do what we just did with our overlaid histograms (and threshold), but it does it for every observation in our data.



Normal and Not Normal Distribution visual comparison

 # We can use the QQ plot function from the statsmodels library:

```
import statsmodels.api as sm
from matplotlib import pyplot as plt
# Create QQ plot
sm.qqplot(np.array(random_normals), line='45')
plt.show()
```

General Subjective Question #6 (contd.,)

- **What The QQ Plot Is Telling Us** - We already know that if the dots of our plot fall on a 45 degree line, then our data is normally distributed (assuming we are using the normal distribution's theoretical quantiles). But when they do not fall on the line, we can still learn a lot about the distribution of our data. Here are some general tips for reading a QQ plot:

- **The slope tells us whether the steps in our data are too big or too small** (or just right). **Remember, each step (where a step is going from one quantile to the next) in the data traverses a fixed and constant percentage** — *for example, if we have N observations, then each step traverses 1/(N-1) of the data. So we are seeing how the step sizes (a.k.a. quantiles) compare between our data and the normal distribution.*

- A steeply sloping section of the QQ plot means that in this part of our data, the observations are more spread out than we would expect them to be if they were normally distributed. One example cause of this would be an unusually large number of outliers (like in the QQ plot we drew with our code previously).

- A flat QQ plot means that our data is more bunched together than we would expect from a normal distribution. For example, in a uniform distribution, our data is bounded between 0 and 1. And within that range, each value is equally likely. So the extremes of the range (like 0.01 and 0.99) are just as likely as something in the middle like 0.50. This is very different from a normal distribution (with mean of 0 and standard deviation of 1) where something like a -3 or a 4 would be much less likely to be observed than a 0. So the QQ plot of a uniformly distributed variable (where the observations are equally spaced and therefore more bunched up relative to a normal distribution) would have a very shallow slope.