BIKE SHARING ASSIGNMENT

Name : Hareshkumar.K.M

Date : 29-Nov-2020

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?***

   The inference on the effect of the categorical variable(predictor) on a dependent variable(target) can be identified using their co-efficients. Below are the list of co-efficients of categorical variables which impacts the target variable (cnt)

S.no	Feature	Description	Weight/coefficient
1	*winter*	A category of season	*612.77*
2	*spring*	A category of season	*-1051.17*
3	*workingday*	A category of day which is represented as 1	*182.28*
4	*weather_cls2*	A category of weathersit : Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist	*-416.59*
5	weather_cls3	A category of weathersit :(Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds)	*-2302.99*

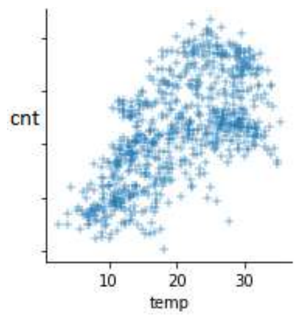
2. ***Why is it important to use drop_first=True during dummy variable creation?***

   If the dummy variable is created is a feature consisting of N categories , adding the drop_first=True parameter will get N-1 dummies by deleting the first level. If we don't drop the first column then the dummy variables will be correlated and may affect some models adversely and the effect is stronger when the cardinality is smaller.

   Example : In the assignment, all the categorical features are handled using dummy variables with drop_first=True.

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

   *The 'temp' feature has the highest correlation with 'cnt' (target). The correlation coefficient is **0.63***



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

Assumptions of Linear Regressions are as follows,

i. The target variable is always has linear relationship with independent variables. This is proved in the assignment model as we can see the linear coefficients of the model summary for every independent variable

OLS Regression Results

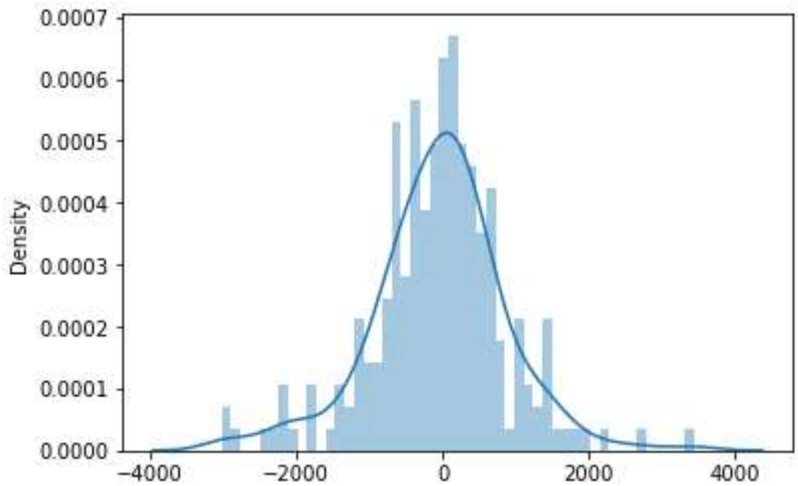
Dep. Variable:	cnt	**R-squared:**	0.828
Model:	OLS	**Adj. R-squared:**	0.825
Method:	Least Squares	**F-statistic:**	301.5
Date:	Sun, 29 Nov 2020	**Prob (F-statistic):**	2.93e-186
Time:	21:51:40	**Log-Likelihood:**	-4151.1
No. Observations:	511	**AIC:**	8320.
Df Residuals:	502	**BIC:**	8358.
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3707.5200	96.720	38.332	0.000	3517.493	3897.547
yr	2054.1000	75.337	27.265	0.000	1906.085	2202.115
workingday	182.2786	79.661	2.288	0.023	25.768	338.789
atemp	940.4500	59.102	15.912	0.000	824.333	1056.567
hum	-140.7650	52.663	-2.673	0.008	-244.232	-37.298
spring	-1051.1707	134.590	-7.810	0.000	-1315.599	-786.742
winter	612.7689	110.980	5.521	0.000	394.727	830.811
weather_cls2	-416.5869	101.438	-4.107	0.000	-615.883	-217.291
weather_cls3	-2302.9936	240.808	-9.564	0.000	-2776.110	-1829.877

Omnibus:	96.932	**Durbin-Watson:**	1.988
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	236.638

cnt = 3707.519988*const + 2054.099990*yr + 182.278640* workingday + 940.449952*atemp - 140.765000*hum - 1051.170651*spring + 612.768884*winter - 416.586930*weather_cls2 - 2302.993645*weather_cls3

ii. The residuals (errors) are normally distributed with mean 0. This was proved in the assignment by a distribution plot of all the residual plots of test dataset.



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

Coefficient of the final model :

```
const            3707.519988
yr               2054.099990
atemp             940.449952
winter            612.768884
workingday        182.278640
hum              -140.765000
weather_cls2     -416.586930
spring          -1051.170651
weather_cls3    -2302.993645
```

Top 3 Features are :

S.no	Feature	Weight/coefficient
1	weather_cls3 (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds)	2302
2	yr	2054
3	spring (season :1)	1051

Since the const (intercept) is not a feature, it is not listed above even it has high coefficient value.

General Subjective Questions:

1. ***Explain the linear regression algorithm in detail***

Linear Regression is a supervised machine learning algorithm where the target variable is always continuous.

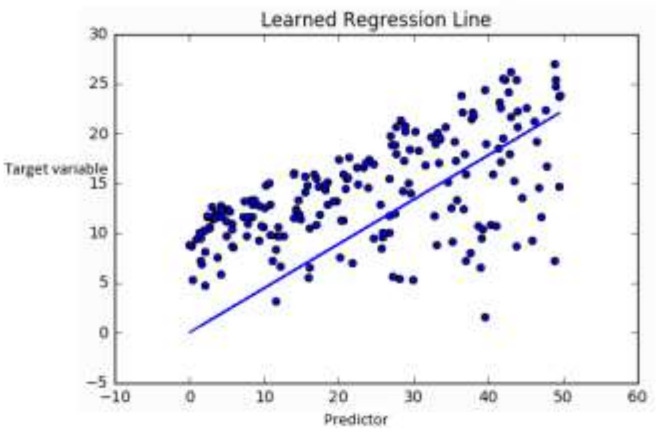
*Y =m1*X1 + m2*X2 +.... ..+ mn*Xn + C +error*

Y - target variable
X1,X2,...,Xn - predictors
C - constant /intercept

A regression algorithm is said to be linear if it follows the below assumptions,

i. The target variable always has a linear relation with one/more independent variables.
ii. The errors are always normally distributed with mean as 0
iii. The variance of the independent variables are always constant (*homoscedasticity*)

The algorithm creates a best fit line which represents the correlation between target variable and predictors as below,



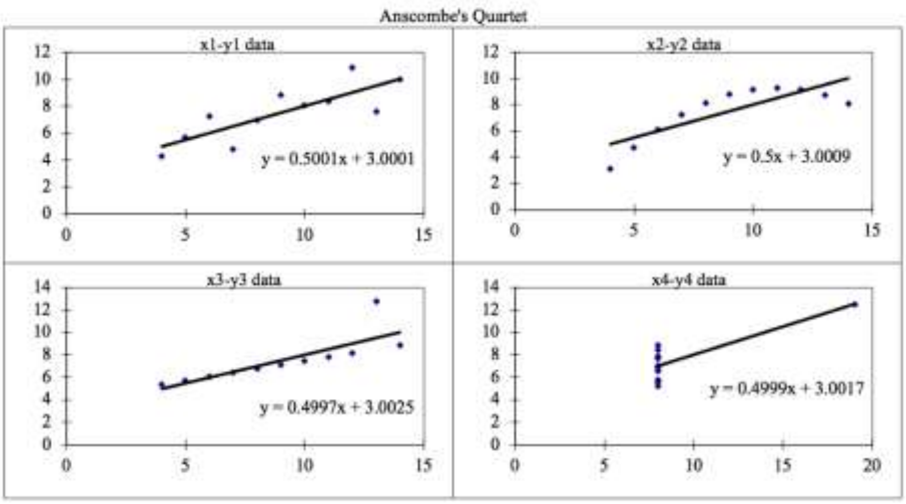
2. ***Explain the Anscombe's quartet in detail.***

Anscombe's Quartet is a group of **four data sets** which looks identical in statistics perspective, but there are some peculiarities in the dataset that fools the regression model . They have very different distributions and appear differently when plotted on scatter plots. It empahsize on the importance of visualization before creating the model.

The 4 datasets mentioned below have same statistical observations,

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	

However, when they are plotted they take a different pattern, which emphasis on the importance of visualization to identify, if the regression line is really a best fit.



,

3.What is Pearson's R?

Pearson's R or the Person coefficient explains the correlation between the dependent and the target variables. It ranges between **-1 to 1.**

i. If the value is **+1** then the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
ii. If the value is **-1** then the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
iii. If the value is **0** then there is no linear association between target and the dependent variable

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

Scaling/feature scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

- **Normalized Scaling**: This technique re-scales a feature or predictors with distribution value between 0 and 1. They are highly sensitive to outliners .They are used predominantly in deep learning

$$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

- **Standardized scaling:** It is a very effective technique which re-scales a feature value so that it has distribution value between -1 and +1 . It has 0 mean value and variance equals to 1. They are not sensitive to outliners in the data.

$$X_{new} = \frac{X_i - X_{mean}}{\text{Standard Deviation}}$$

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

Variable Inflation factor explains the correlation of one predictor with another predictor.
Below if the formula for VIF

VIF = 1 /(1- R-squared)

VIF becomes infinite when the **R-squared between the predictors is 1**. Which means, one predictor variable has very high correlation with other predictor variable. It is always advisable to drop the predictors which has high value.

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

Q–Q (quantile-quantile) plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. It is a technique to determine if two data sets come from populations with a common distribution.

The q-q plot is formed by:

Vertical axis: Estimated quantiles from data set 1

Horizontal axis: Estimated quantiles from data set 2

Both axes are in units of their respective data sets. That is, the actual quantile level is not plotted. For a given point on the q-q plot, we know that the quantile level is the same for both points, but not what that quantile level actually is.

For linear regression , QQ plt can be used to n identify if the residuals are normally distributed. If we plot the quantiles of the residuals against the quantiles of the normal distribution, and the quantiles of the residuals are near enough to the quantiles of the corresponding values computed from the normal distribution, then the residuals are normally distributed.