Bike Sharing Assignment

Sanjay Singh

Bike Sharing Assignment –

Assignment-based Subjective Questions

- Q1 : From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable
- Ans 1 : There are total 11 vars (1 const & 10 independent var) in this model, except const, temp and windspeed, rest all 8 vars are categorical. Which indicate a significant effect on dependent variable
- Q2 : Why is it important to use **drop_first=True** during dummy variable creation?
- Ans 2: When we use ML we work with algorithms and these algorithms cannot process categorical variables. In such situations, we need to make sure that columns values are transformed into individual separate columns with values as 0s and 1s. This is called getting dummies pandas columns. ex. columns labels of Food Type are Veg, NonVeg and others and these labels are transformed into Veg and NonVeg and we don't need others as once both Veg and nonveg is 0 that means automatically it is other type. Highlighted Other we don't require.

Other	NonVeg	Veg
1	0	0
0	1	0
0	0	1

Bike Sharing Assignment – Continue..

Assignment-based Subjective Questions

- Q3 : Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
- Ans 3 : temp, atemp
- Q4 : How did you validate the assumptions of Linear Regression after building the model on the training set?
- Ans 4:
  - Residual Analysis
  - Validating VIF values
  - P-value, Adjusted R-Sq from OLS summary

Bike Sharing Assignment – Continue..

Assignment-based Subjective Questions

- Q5 : Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
- Ans 5 : Temp, workingday, windspeed

Bike Sharing Assignment – Continue..

General Subjective Questions

- Q1 : Explain the linear regression algorithm in detail.

➢ Ans 1 : Linear regression is to find a linear relation between dependent and one or multiple independent variables. It is of 2 types.

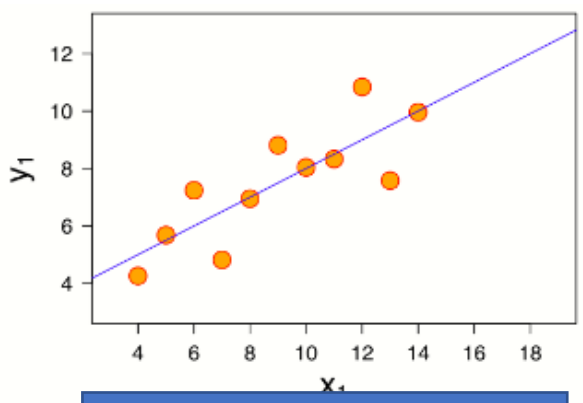
- ➢ **Simple linear regression** [one independent and one dependent variable]
  - ➢ Y pred = b0 + b1*x [linear equation]
  - ➢ b0 & b1 is chosen in such that to minimize the error ex. Error = sum of Squares actual output – predicted output ➔ called Sum of Squared error.
  - ➢ b0 is called intercept & b1 is called coefficient
  - ➢ If b1 >0 means x & y has positive relation in case x increases y also increase and b1 <0 its negative relation means x increase and y decreases.
  - ➢ Ignoring b0 will ensure that model passes through origin and it will not predict correct result [bias]
  - ➢ Residual analysis is very important after prediction to validate Error is normally distributed
  - ➢ R-square value
  - ➢ Low p-value
- ➢ **Multiple linear regression** [more than one independent and one dependent variable]
  - ➢ Y pred = b0 + b1*x + b2*y……b3*z
  - ➢ Same analysis is performed and only difference is we have more than one independent variables.
  - ➢ Multicollinearity is to be taken care while modelling as 2 Independent variable might have co-relation between each other.

Bike Sharing Assignment – Continue..

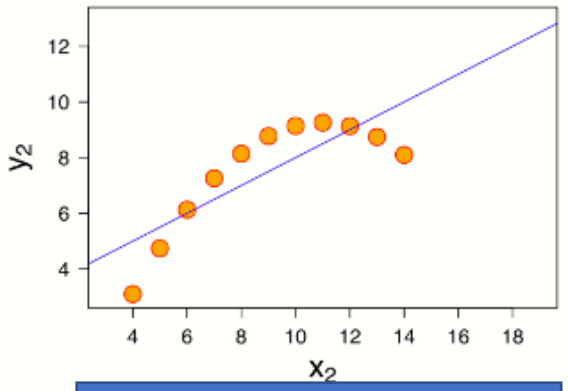
General Subjective Questions

- Q2 : Explain the Anscombe's quartet in detail

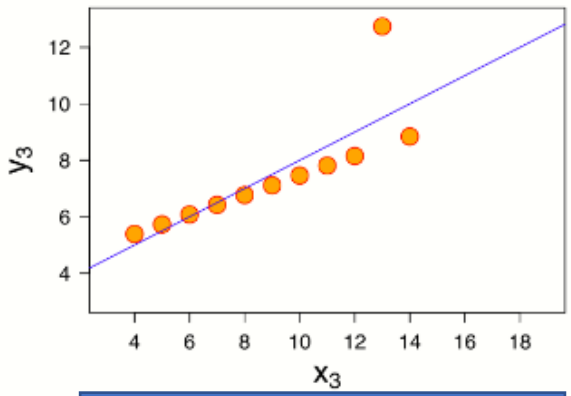
- Ans 2 : Anscombe's quartet tells that visualization is very important before doing any analysis and/or building models. It highlights that even though different data set have same mean, variance, correlation coefficient and best fit line but when you plot them it could be very different result which may not be linear at all. So it means don't just rely only on statistics but plot the data set.
- Example of data set below has same statistics but very different when you plot them.



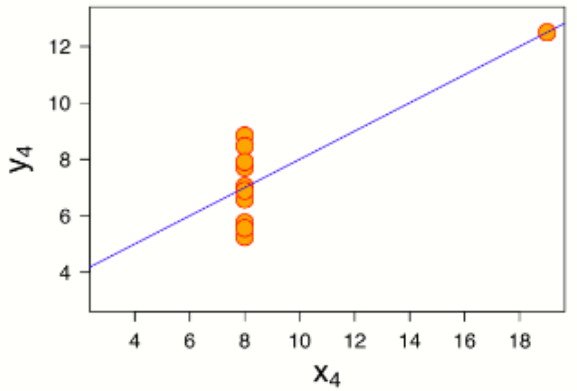
Linear Model



Not a linear model



Linear but outliers



Not a Linear model and outliers

Bike Sharing Assignment – Continue..

General Subjective Questions

	X1	Y1	X2	Y2	X3	Y3	X4	Y4
	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	5.74	5	5.73	8	6.89
Sum	**99**	**82.51**	**99**	**82.51**	**99**	**82.51**	**99**	**82.51**
Avg	**9**	**7.5**	**9**	**7.5**	**9**	**7.5**	**9**	**7.5**
STDEV	**3.32**	**2.03**	**3.32**	**2.03**	**3.32**	**2.03**	**3.32**	**2.03**

Bike Sharing Assignment – Continue..

General Subjective Questions

- Q3 : What is Pearson's R?

- Ans 3 : Pearson's R is used to measure linear trends between 2 variables. Its possible values range is -1 to 1.
  - When r = 1 means it a perfect +tve relation between 2 variable and all samples are only on best fit line with upwards slope
  - When r >0 and r < 1 means samples are scattered around the best fit line and 2 variables are having +tve linear relation. Increase in x will result increase in y.
  - When r = -1 means it a perfect -tve relation between 2 variable and all samples are only on best fit line with downwards slope
  - When r >-1 and r <0 means samples are scattered around the best fit line and 2 variables are having -tve linear relation. Increase in x will result decrease in y.
  - When r = 0 means there is no linear relation between 1 variables.

Bike Sharing Assignment – Continue..

General Subjective Questions

- Q4 : What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
- Ans 4 : Scaling is a technique to bring down the independent variables values of a data set to the same scale. It is like normalizing the values to same scale. There are two types of scaling method we perform. If we donn;t scale then difference between variable will be more, which will result into incorrect model as higher values will have more impact on model.
  - Normalization : This is know as Min-Max Scaling. This will scale the range between 0 and 1.
  - X-new = (X – X min) / (X-max - X-min)
  - Standardization : will transform of data by subtracting mean and divided by standard deviation. This is also called Z-score. In this scaling all data points are centered around mean.
  - X-new = X – mean / (St D)

Data before scaling

```
In [37]: bikes_train.head()
```

Out[37]:

	holiday	workingday	temp	atemp	hum	windspeed	cnt	spring	summer	winter	...	Sep	Li
576	0	1	29.246653	33.1448	70.4167	11.083475	7216	0	0	0	...	0	
426	0	0	16.980847	20.6746	62.1250	10.792293	4066	1	0	0	...	0	
728	0	0	10.489153	11.5850	48.3333	23.500518	1796	1	0	0	...	0	
482	0	0	15.443347	18.8752	48.9583	8.708325	4220	0	1	0	...	0	
111	0	1	13.803347	16.0977	72.9583	14.707907	1683	0	1	0	...	0	

5 rows × 30 columns

Data after Scaling

```
In [91]: bikes_train.head()
```

Out[91]:

	holiday	workingday	temp	atemp	hum	windspeed	spring	summer	winter	Aug	...	Sep
576	0	1	0.815169	0.766351	0.725633	0.264686	0	0	0	0	...	0
426	0	0	0.442393	0.438975	0.640189	0.255342	1	0	0	0	...	0
728	0	0	0.245101	0.200348	0.498067	0.663106	1	0	0	0	...	0
482	0	0	0.395666	0.391735	0.504508	0.188475	0	1	0	0	...	0
111	0	1	0.345824	0.318819	0.751824	0.380981	0	1	0	0	...	0

5 rows × 29 columns

Bike Sharing Assignment – Continue..

General Subjective Questions

- Q5 : You might have observed that sometimes the value of VIF is infinite. Why does this happen?
- Ans 5 : VIF infinite mean very high correlation (Perfect correlation) between two independent variables. In modeling VIF more than 5 variables are dropped. We need to look after other factor as well while dropping. In this scenario we need to drop 1 variable.

Bike Sharing Assignment – Continue..

General Subjective Questions

- Q6 : What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
- Ans 6 : Q-Q plot is a graphical tool to help find if 2 data set are from same populations with common distribution. You need to plot the 2 quantiles (ex. Median 50 % data below this value). A 45 % degree angle is plotted on Q-Q plot if the two data sets come from a common distributions. Q-Q model help checking model, we need to ensure and check the distribution of error terms. Graph should not have significant deviation from the mean.