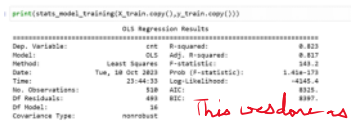
Linear Regression Subjective Questions

Tuesday, October 10, 2023 11:49 PM

Assignment Based

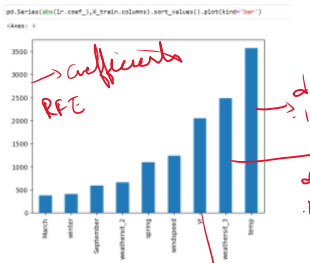
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

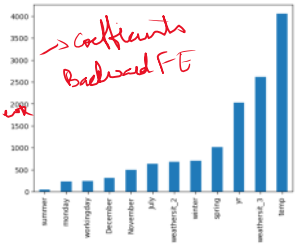
(A) Yr has the highest positive correlation(.57) with the target variable among categorical variables. It is also the 2nd the highest coefficient (slope) among the categorical variables. While weathersit_3 (a component generated from the weather sit column) has the highest coefficient weight among categorical variables. Spring a component of dummy encoding the seasons column is important too , also spring got a correlation score (r– Pearson's Correlation coefficient) of - 0.56 , a negative correlation really close to Yr , but in the opposite direction. It was observed that the yr column had considerable impact on the r2-score . Removing yr column led to drop of approx. .30 in the r2-score, this was observed while experimenting with different feature sets. Another notable but less important feature would be be components of mnth(month).



removed yr

→ As seen removing year led to such a large drop in r2-Score





As seen temp, weather sit 3 & yr came out as top features in both RFE & Backward Feature Selection models

Although going by the coeffs temp is most imp, going by drop in r2-score based on feature absence yr is most important.

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

(A) It is important to use drop_first = True , because not doing so can lead to multicollinearity. The variance of the nth dummy variable created explained by the n-1 dummy variables that come are extracted . Multicollinearity cause problems in model interpretation and it also makes the model more complex by adding a additional features . The interpretation that the coefficient of x1 , m1 represent the change in y with wrt x1 ,when x1 increases by 1 unit and all other independent variables are constant doesn't h old.

Eg: Lets say you have a categorical variable that represents the final result of a students exam

Result → Pass, Fail, Absent

dummy encoding

Pass	Fail	Absent
1	0	0
0	1	0
0	0	1

Eg from assignment

```
# Dummy encoding categorical variables
df = pd.concat([df,pd.get_dummies(df['season'],drop_first=True)],axis=1)
df = pd.concat([df,pd.get_dummies(df['weathersit'],drop_first=True)],axis=1)
df = pd.concat([df,pd.get_dummies(df['weekday'],drop_first=True)],axis=1)
df = pd.concat([df,pd.get_dummies(df['mnth'],drop_first=True)],axis=1)
```

Eg of what happens when there is multicollinearity

Pass & Fail can completely explain Absent. This also gives a high VIF (Variance Inflation Factor) Score

∴ It can be concluded that drop_first = True, is very important

→ it can be inferred from here the variance/results of the Absent column can be inferred from Pass & Fail

	Features	VIF
1	holiday	inf
2	workingday	inf
7	monday	inf
8	tuesday	inf
9	wednesday	inf
10	thrusday	inf
11	friday	inf
14	temp	4.71
15	windspeed	4.02
0	yr	2.06
3	spring	1.84
5	weathersit_2	1.53
4	winter	1.41
12	March	1.26
13	September	1.16
6	weathersit_3	1.09

→ The results of VIF = Inf is because holiday, working day & weekdays can completely explain variance of the target when other 2 are present

Variance of holiday can be completely explained by weekday + working day & vice versa

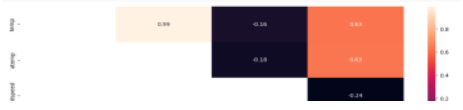
```
print(vif_method(X_train.cop
```

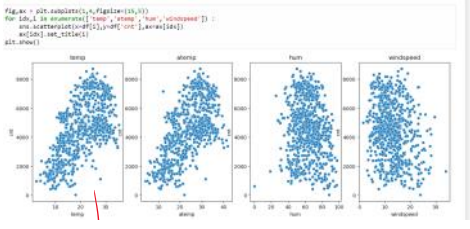
	Features	VIF
1	workingday	30.31
8	wednesday	8.58
7	tuesday	7.27
9	thrusday	7.03
10	friday	6.92
6	monday	6.57
13	temp	4.71
14	windspeed	4.02
0	yr	2.06
2	spring	1.84
4	weathersit_2	1.53
3	winter	1.41
11	March	1.26
12	September	1.16
5	weathersit_3	1.09

→ Effect of removing holiday & working day still has high VIF because its variance can explained by weekdays

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

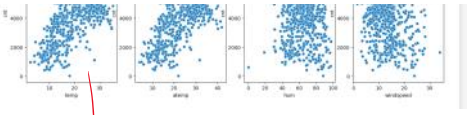
(A) Disregarding registered and casual , which are a part of cnt (cnt = registered + casual) . The features with highest correlation to the target variable are temp and atemp





Also temp & atemp have a 0.99 correlation with each other. Since they are temp & what the temp feels like

↳ As seen temp & atemp have high correlation



↳ As seen temp & atemp have a comparatively higher positive correlation

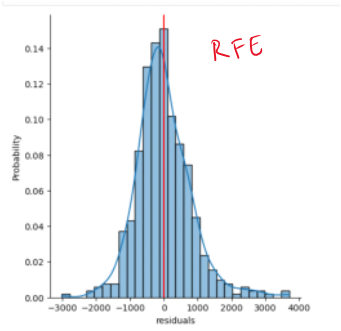
↳ Vice versa

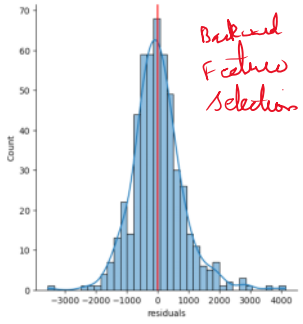
can explained by weekdays

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

(A) The model was validated by checking the mean of the residuals . The residuals followed a normal distribution and meanof the residuals was infinitesimally close to zero . Also the there weren't any large notable patterns overall . There was slight decrease in residuals as cnt(target variable) increased. Also a qq-plot(quantile-quantile plot) was plotted to check if the residuals were normally distributed in addition to this .

Residual distribution





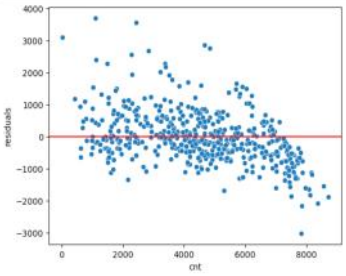

```
(y_pred_train - y_train).mean()
2.7142175217615624e-12
```

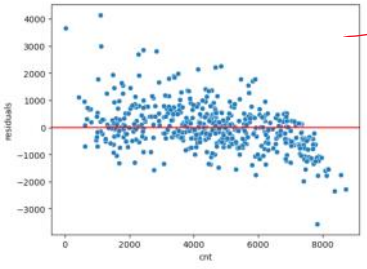
↳ RFE

```
(y_pred_train - y_train).mean()
1.2137259961306096e-11
```

↳ Backward Feature Selection

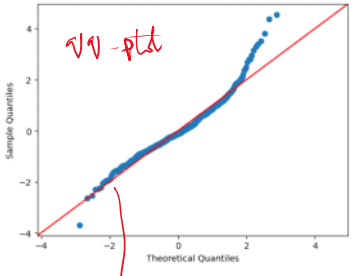
Left: RFE based model
Right: Backward Feature selection model

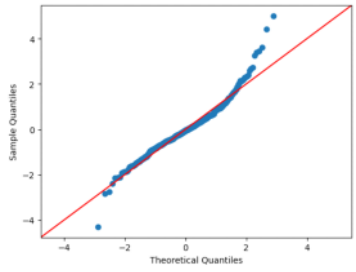




→ Residuals variance

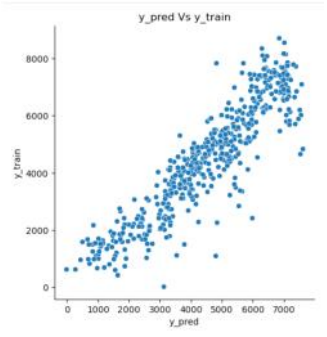
* It can be observed that there seems to be pattern to variance towards higher counts.

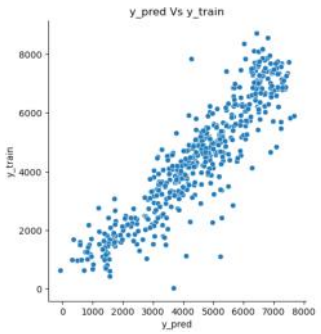




* Although the residuals seem to follow a normal distribution they are tapering off towards the right tail

↳ check if residuals follow a normal distribution

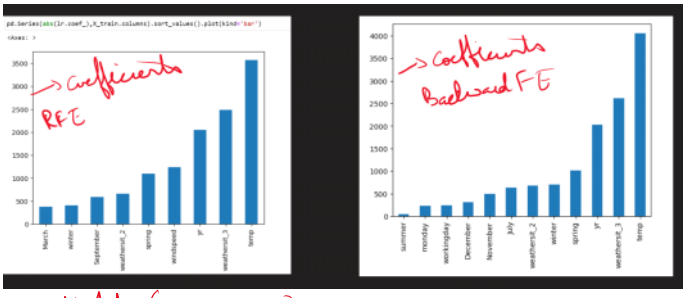




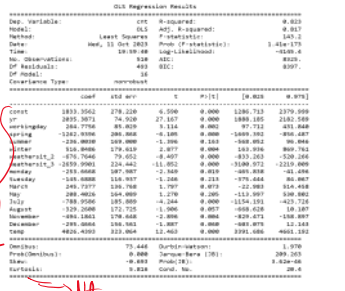
→ ypred vs y-train
There is a +ve linear relationship here which is a good thing

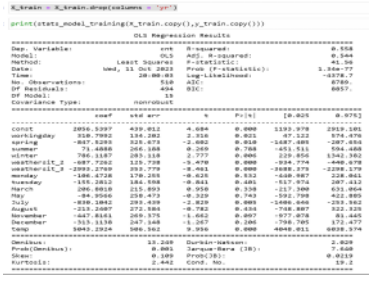
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

(A) It can be observed from the below plots that represent coefficients that temp , weather_sit3 and Yr are the most important features to the model. This was verified across 2 models that used RFE and the other backward feature elimination.



Base Model (FE)





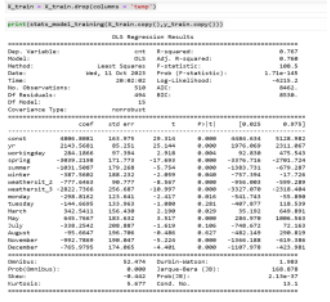
R-squared 0.823 to 0.558

→ yr removed

After feature Elimination

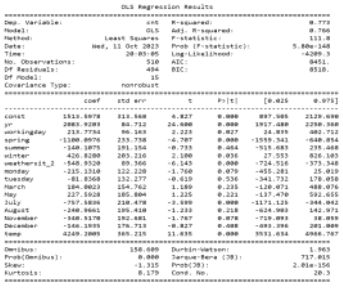
* Note this is not final model. This is the output of feature Elimination

→ being displayed for example purpose



→ temp removed

R-Squared 0.823 to 0.767



→ Weathersit_3 removed

R-squared 0.823 to 0.773

Generic Questions

1. Explain the linear regression algorithm in detail ?

(A) Is a supervised machine learning algorithm , meaning it required target labels for the model to be trained . It is part of the regression family of ml algorithms . Linear regression aims to fit the target variable using a straight line as its name suggests. It is used to predict the behavior of a continuous variable .

Equation of Linear Regression

$y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i$, where β_0 is your intercept & β_i is the coefficient / slope of the i^{th} feature (x_i)

$n \rightarrow$ no of features

Assumptions of Linear Regression :

* There should be linear relationship between X (Independent/Predictor) & Y (dependent/target) variables .

* The error terms that is the residuals should be normally distributed , with their mean centered on 0.

* The variance of residuals should be constant (homostadacity)

* Error terms should be independent of one another. If the error at $y = 10$ is high it is not that the error at $y = 20$, should be higher or lower. The errors shouldn't dependent on the previous error

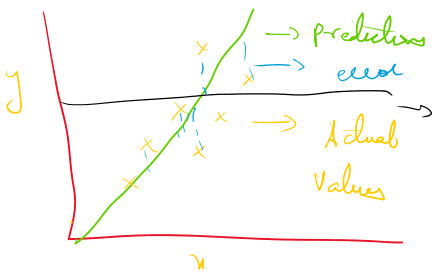
The equation of linear regression can be solved in 2 ways

OLS & Gradient Descent .

↳ closure function ↳ Iterative approach

Error terms :

Mean Squared Error $= \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|^2 = \frac{1}{n} \sum_{i=1}^{n} |y_i - (\beta_0 + \beta_i x_i)|^2$



$y - \hat{y}$

The errors here are called residuals .

* You want the error to be as low

* You want the error to be as low as possible

* Although Mean Squared error gives you error of the model, it cannot be determined if the model you have is good or bad, as there is no values that defines it.

This is where R2-score / R-Squared comes in

$$R2\text{-Score} = 1 - \frac{MSE \text{ of Regression Model}}{MSE \text{ of Mean Line}}$$

- Mean of all y values is taken & a line is drawn
  - ↳ calculate MSE for this
- Values lie b/w 0 & 1,
  - closer you are to 1 better your model predicts
  - closer you are to 0, the worse your model predicts

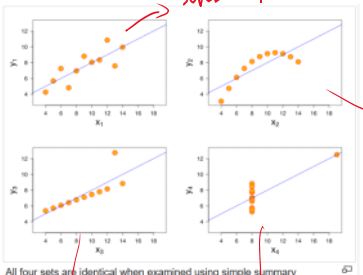
2. Explain the Anscombe's quartet in detail ?

(A) Anscombe's quartet contains 4 datasets which extremely similar descriptive stats i.e the same mean , variance and correlations. However as seen in the same they have very different distribution . This emphasizes the need to always plot the data , as these stats can be misleading . Anscombe's quartet provides important of plotting graphs especially in linear regression . These 4 datasets give extremely similar coefficients , intercepts and r2-scores for . If not plotted it could lead to the misconception that the datasets are identical .

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

→ this is the dataset



this represents linear relationship

There is relationship but not linear

linear relationship with outliers

No linear relationship at all. But these produce high correlation

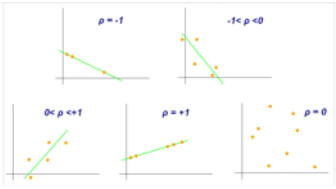
This proves that plotting data is important. As it is not a good idea to use linear model in II & IV

III shows that a single outlier can lower the score of a model

I seems to be simple linear relationship model

3. What is Pearson's R ?

(A) Is pearson's correlation coefficient . It is used to measure the correlation between linear data . This is the most widely used correlation coefficient , though it doesn't to well on non-linear data . Usually spearman's correlation is used to measure correlations in case data is monotonic . The value of r ranges from -1 to 1 . Closer to 1 indicates a strong positive correlation . Closer to zero indicates a very weak correlation



↳ Covariance btw x & y

$r = P(x,y) = \frac{Cov(x,y)}{\sigma_x \sigma_y}$ → Variance of y

↳ Variance of x

Pd.corr() → gives you pearson correlation coefficient (r) by default for all combinations in the dataframe

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

(A) Scaling is the process of bringing continuous variables in a dataset to a similar range . Scaling is performed to prevent a single feature from overly influencing the outcome , especially in models like liner regression , logistic regression and knn , as values with a higher scale usually tend to dominate the predictions .

Standardized Scaling or Standardization : is a scaling technique which scales features into a standard normal distribution with mean = 0 and std =1 . Also 99.7% of the data is present within 3 std from the mean.

Normalization or Normalized Scaling or MinMax Scaling : is a scaling technique which brings all the features into a range between 0 and 1 . The max value would be 1 and the minimum value would be 0. It also handles outliers . It compresses the outliers , so that need they are no longer considered outliers after scaling

Standardization

Z-Score Scaling = $\frac{x-\mu}{\sigma}$, $\mu = 0, \sigma = 1$

Min Max Scaling = $\frac{X - X_{min}}{X_{max} - X_{min}} \in [0, 1]$

↳ handles outliers

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

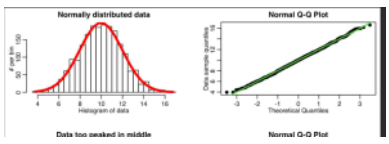
(A) VIF also known as Variance Inflation Factor estimates how much variance of the a dependent variables is being explained by the other dependent variables .
The VIF would be infinite if the variance of the dependent variable is being completely explained by the other dependent variables in the dataset . It is sign of very high multi collinearity .

$VIF = \frac{1}{1 - R^2}$, $VIF = \infty$, when $R^2 = 1$ → Variance completely explained

↳ perfect linear relation

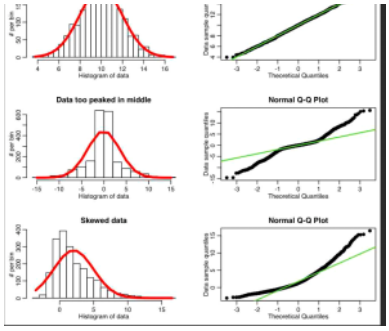
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

(A) Q-Q plots that expand to Quantile-Quantile Plots are used to measure if 2 datasets come from the same distribution . In linear regression q-q plots can be used to determine if the residuals follow normal distribution . q-q plots have a theoretical quantile and a sample quantile . The theoretical quantile in this case of linear regression would be a normal distribution and the sample quantiles would be the percentile values of the residuals.



Interpretation of Q-Q plot

If the sample distribution is same as



If the sample distribution is same as theoretical distribution the pts would line on the straight line.

∴ The more the pts lie on the straight line, more similar the distributions

Inversely, as more & more pts tapper away from the line, more distinct the distributions