




LINEAR REGRESSION

(DESCRIPTIVE STATISTIC ANALYSIS)

Maryam Najimigoshtasb



A national chain of women's clothing stores with locations in the large shopping malls thinks that it can do a better job of planning more renovations and expansions if it understands what variables impact sales. It plans a small pilot study on stores in 25 different mall locations. The data it collects consist of monthly sales, store size (sq. ft), number of linear feet of window display, number of competitors located in mall, size of the mall (sq. ft), and distance to nearest competitor (ft).

Find a multiple regression model for the data.

Sales = $\beta_0 + \beta_1$ (store size)
+ β_2 (number of linear feet of window display)
+ β_3 (number of competitors located in mall)
+ β_4 (size of the mall)
+ β_5 (distance to nearest competitor)

```
libname mylib '/home/u58699890/My practice/mylib';  
filename Mall '/home/u58699890/My practice/File_MALL.xlsx';  
proc import datafile=Mall out=mylib.Mall replace DBMS=xlsx ;getnames=yes;  
run;  
  
proc reg data=mylib.mall;  
model sales= Size Windows Competitors MallSize NearestCompetitor;  
run;
```

Interpret the values of the coefficients in the model

Intercept: 1506.80179
this value indicates the least square value of mean sale and is calculated when other factors are zero, as having zero value for other factors is not practical so interpretation of y intercept is not logical.

LEAST SQUARE SLOPES:
Size: 0.91937, it shows that mean sale increase by 0.91937 for each sq.ft increase in size of store.
WindowsB: 9.07598 it shows that mean sale increase by 9.07598 for each number of linear feet increase of window display
Competitors: - 67.68553 it shows that mean sale decrease by 67.68553 for each number of competitors located in mall increase
Mallsize: -0.00090285 , it shows that mean sale decrease by 0.00090285 for each sq.ft increase in size of Mall
Nearst competitor: 2.09589 it shows that mean sale increase by 2.09589 for each ft increase distance to nearest competitor.

Test whether the model as a whole is significant. At the 0.05 level of significance, what is your conclusion?

The most important indicator of having a good model is f-test for this level of significant, it is the good enough to consider it as a good model how ever the f value in not desirable.

The REG Procedure
Model: MODEL1
Dependent Variable: Sales Sales

Number of Observations Read	25
Number of Observations Used	25

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	5761406	1152281	19.21	<.0001
Error	19	1139390	59968		
Corrected Total	24	6900796			

Root MSE	244.88345	R-Square	0.8349
Dependent Mean	4535.48000	Adj R-Sq	0.7914
Coeff Var	5.39928		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	Intercept	1	1506.80179	672.18680	2.24	0.0371
Size	Size	1	0.91937	0.30063	3.06	0.0065
Windows	Windows	1	9.07598	28.82343	0.31	0.7563
Competitors	Competitors	1	-67.68553	21.95288	-3.08	0.0061
MallSize	MallSize	1	-0.00090285	0.00028062	-3.22	0.0045
NearestCompetitor	NearestCompetitor	1	2.09589	1.59443	1.31	0.2043



Use the model to predict monthly sales for each of the stores in the study

(store size) x1

(number of linear feet of window display) x2

(number of competitors located in mall) x3

(size of the mall) x4

(distance to nearest competitor) x4

$$\text{Sales} = 1506.80179 + 0.91937 X1 + 9.07598X2 + -67.68553 X3 + -0.00090285 X4 + 2.09589X5$$



Plot the residuals versus the actual values. Do you think that the model does a good job of predicting monthly sales? Why or why not?

There 4 assumptions to check the linear regression model :

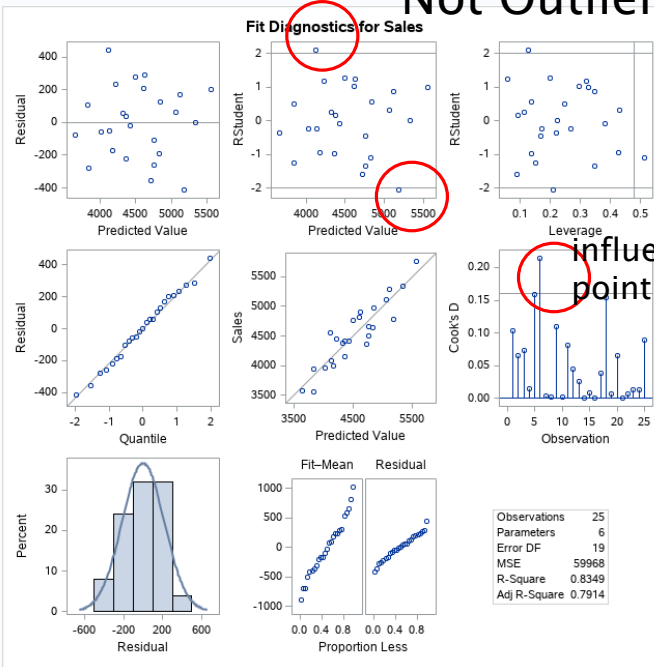
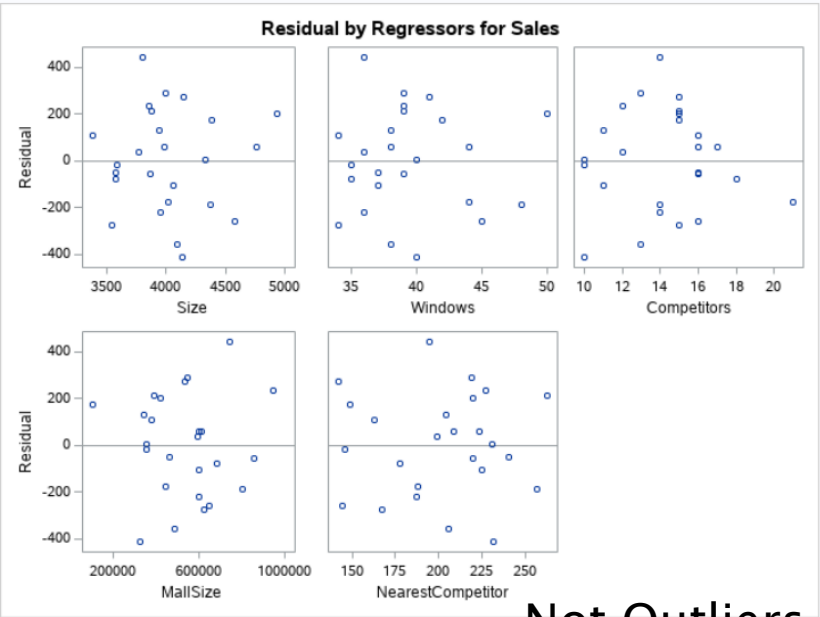
Linearity :there is no any pattern in the residual plot and error point is randomly spreaded equally around the line , so the linearity fulfilled

Normality : the afore mentioned reason is confirm it as well. Moreover, we can use percent – residual histogram on the graph.which Is following the normal distribution.

Homoskatsticity: variance spread within normal range out around the line and there is no pattern we can not see any increase or decrease of error by increasing x.

Independency : we can not see any increase or decrease of error by increasing x.

All the Assumptions are fulfilled we can conclude that model is doing great



Not Outliers

influential point

Observations	25
Parameters	6
Error DF	19
MSE	59968
R-Square	0.8349
Adj R-Square	0.7914

Find and interpret the value of R^2 for this model.

R square is 83 percent which shows our variable 83 percent give us correct value of sales, R adjusted is less because it penalize us for carp variables but still above 75 percent. We might can improve it .

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	5761406	1152281	19.21	<.0001
Error	19	1139390	59968		
Corrected Total	24	6900796			

Root MSE	244.88345	R-Square	0.8349
Dependent Mean	4535.48000	Adj R-Sq	0.7914
Coeff Var	5.39928		

Do you think that this model will be useful in helping the planners? Why or why not?

At this stage , F-test and R or better adjusted R square confirm that this model can be a good model for the purpose, but we need to investigate why R square is 4 percent lower it might have carp variables. Next we go to the t-test of variables .

Test the individual regression coefficients. At the 0.05 level of significance, what are your conclusions?

All of them except **Windows** and **Nearstcompetitors** are less than level of significant, which is good it means our variables explain our model but **Windows** and **Nearstcompetitors** are not good, are above 0.05.

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	Intercept	1	1506.80179	672.18680	2.24	0.0371
Size	Size	1	0.91937	0.30063	3.06	0.0065
Windows	Windows	1	9.07598	28.82343	0.31	0.7563
Competitors	Competitors	1	-67.68553	21.95288	-3.08	0.0061
Mall Size	MallSize	1	-0.00090285	0.00028062	-3.22	0.0045
NearestCompetitor	NearestCompetitor	1	2.09589	1.59443	1.31	0.2043

If you were going to drop just one variable from the model, which one would you choose? Why?

Windows and **Nearstcompetitors** are not explaining the model and the t-test is above of the level of significant so we need to first delete **Windows** to see if our model explained by variable will be better or not , looking at the **F value** , then if still **Nearstcompetitors** are above of 0.05 then do the same for that.

Use stepwise regression to find the best model for the data.

```
title " Stepwise Selection Methods";
title2 "Using Default Values for SLENTRY and SLSTAY";
proc reg data=mylib.mall;

Stepwise: model sales= Size Windows Competitors MallSize NearestCompetitor /
selection = stepwise;
run;
```

Stepwise Selection Methods
Using Default Values for SLENTRY and SLSTAY

The REG Procedure
Model: Stepwise
Dependent Variable: Sales Sales

Number of Observations Read	25
Number of Observations Used	25

Stepwise Selection: Step 1

Variable Size Entered: R-Square = 0.5814 and C(p) = 27.1707

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	4012100	4012100	31.94	<.0001
Error	23	2888696	125595		
Corrected Total	24	6900796			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	222.40809	766.39576	10577	0.08	0.7743
Size	1.07258	0.18977	4012100	31.94	<.0001

Bounds on condition number: 1, 1

Stepwise Selection: Step 2

Variable Competitors Entered: R-Square = 0.7409 and C(p) = 10.8132

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	5112961	2556481	31.46	<.0001
Error	22	1787835	81265		
Corrected Total	24	6900796			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	1287.75994	681.05078	290546	3.58	0.0719
Size	1.08865	0.15271	4129796	50.82	<.0001
Competitors	-79.57360	21.61997	1100861	13.55	0.0013

Bounds on condition number: 1.0008, 4.0033

Stepwise Selection: Step 3

Variable MallSize Entered: R-Square = 0.8155 and C(p) = 4.2301

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	5627674	1875891	30.94	<.0001
Error	21	1273122	60625		
Corrected Total	24	6900796			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	1769.60574	611.03962	508470	8.39	0.0086
Size	1.04482	0.13276	3755185	61.94	<.0001
Competitors	-71.03060	18.90237	856069	14.12	0.0012
MallSize	-0.00079216	0.00027187	514713	8.49	0.0083

Bounds on condition number: 1.0367, 9.2279

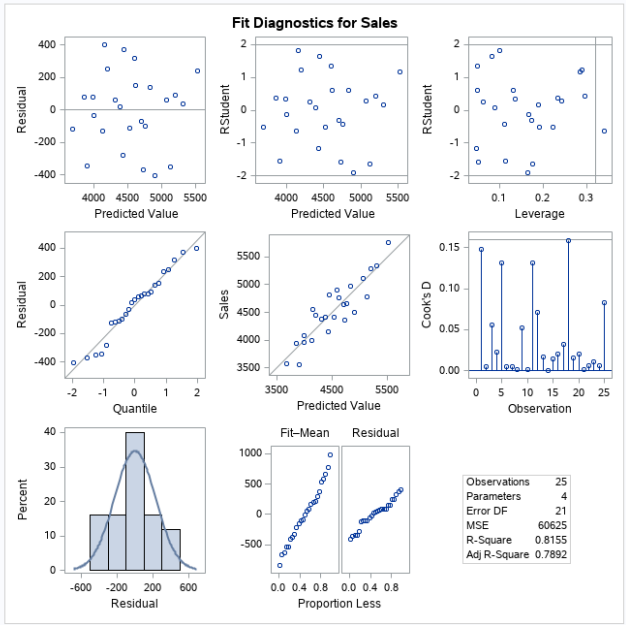
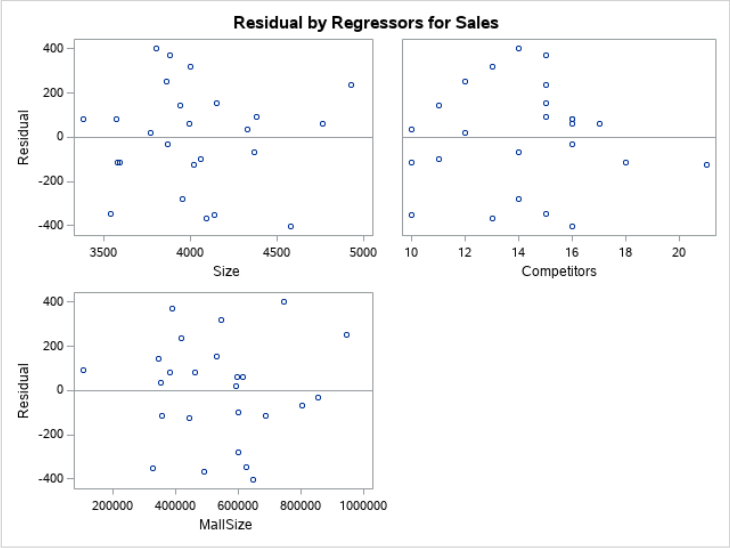
All variables left in the model are significant at the 0.1500 level.

No other variable met the 0.1500 significance level for entry into the model.

Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	Pr > F
1	Size		Size	1	0.5814	0.5814	27.1707	<.0001
2	Competitors		Competitors	2	0.1595	0.7409	10.8132	0.0013
3	MallSize		MallSize	3	0.0746	0.8155	4.2301	0.0083

Analyze the model you have identified to determine whether it has any problems.

this model starts by adding size in the model in the first step and then select the competitors as the next significant factor respect to the size.it tries to increase the f-value , it counties to the step 3 and added Mallsize considering the significance of existing variables and then stop adding as the windows and nearestcompetitors as they will increase the probability of failing the model. All the four assumptions are met using the residual plot there is not any pattern and normality linearity and variance equality and independency of residual are met . There is no outlier using studentize plot there is not outliers. Comparing these cooks D's of stepwise and multiple regression it seems the high influential points are deleted.





Write a memo reporting your findings to your boss. Identify the strengths and weaknesses of the model you have chosen.

The principal drawbacks of stepwise multiple regression include bias in parameter estimation, inconsistencies among model selection algorithms, an inherent (but often overlooked) problem of multiple hypothesis testing, and an inappropriate focus or reliance on a single best model. The model has the better F-value compare to the multiple regression has been done in last slides.



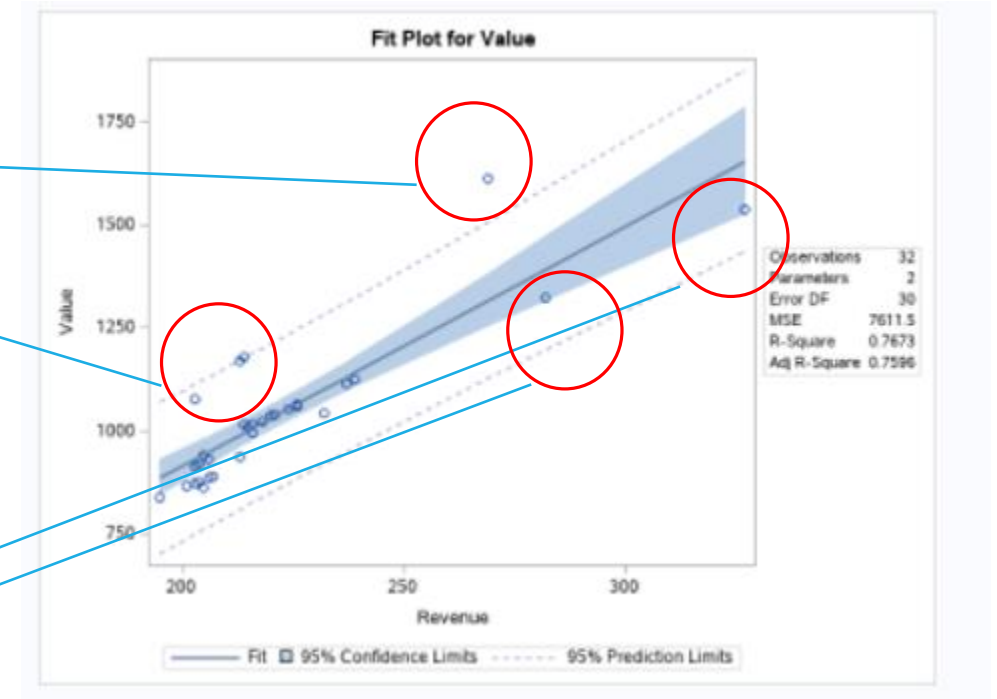
The File NFLValues.xlsx show the annual revenue (\$ millions) and the estimated team value (\$ millions) for the 32 teams in the National Football League.

```
.....  
proc reg data=mylib.NFL;  
model Value= Revenue/r;  
output out = demo cookd= cook student=  
studresids;  
run;  
.....  
proc sort data= demo; by studresids;  
run;  
.....  
proc print data=demo;  
run;
```

Develop a scatter diagram with Revenue on the horizontal axis and Value on the vertical axis. Does it appear that there are any outliers and/or influential observations in the data?

Influential points and outliers

Influential points but not outliers



Develop the estimated regression equation that can be used to predict team value given the value of annual revenue.

$$\text{Value} = -252.07830 + 5.83167 \text{ Revenue}$$

The REG Procedure

Model: MODEL1

Dependent Variable: Value Value

Number of Observations Read	32
Number of Observations Used	32

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	753008	753008	98.93	<.0001
Error	30	228346	7611.53579		
Corrected Total	31	981354			

Root MSE	87.24412	R-Square	0.7673
Dependent Mean	1040.00000	Adj R-Sq	0.7596
Coeff Var	8.38886		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	Intercept	1	-252.07830	130.81712	-1.93	0.0635
Revenue	Revenue	1	5.83167	0.58631	9.95	<.0001

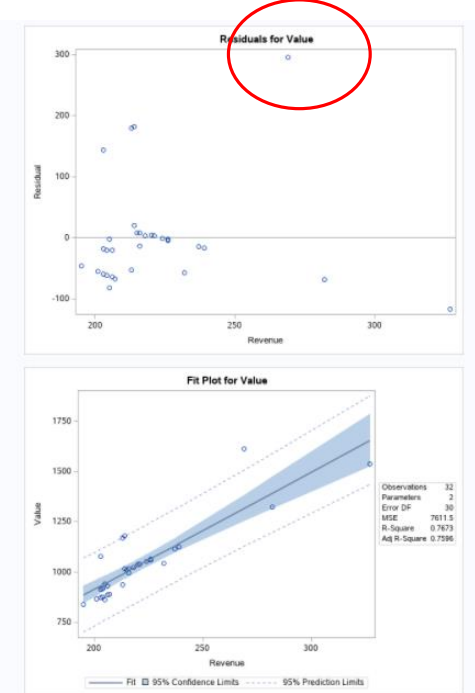
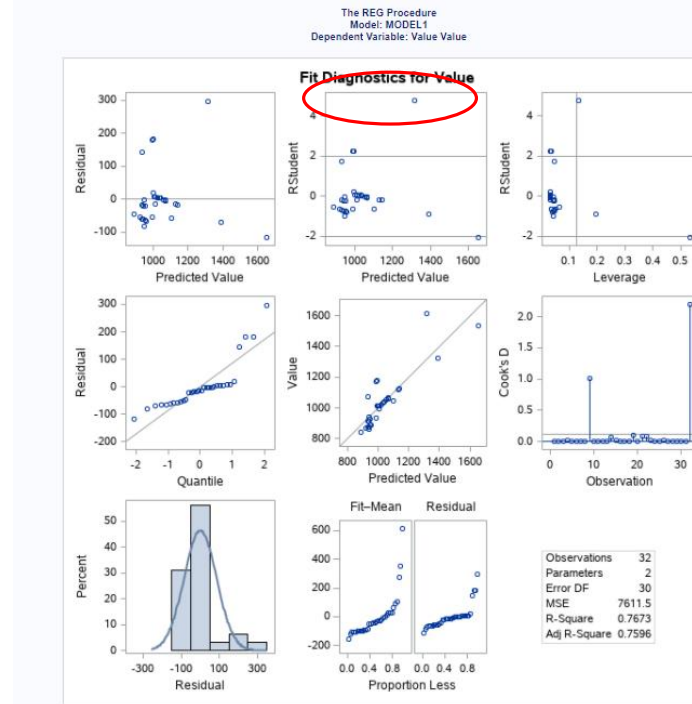
c. Use residual analysis to determine whether any outliers and/or influential observations are present. Briefly summarize your findings and conclusions.

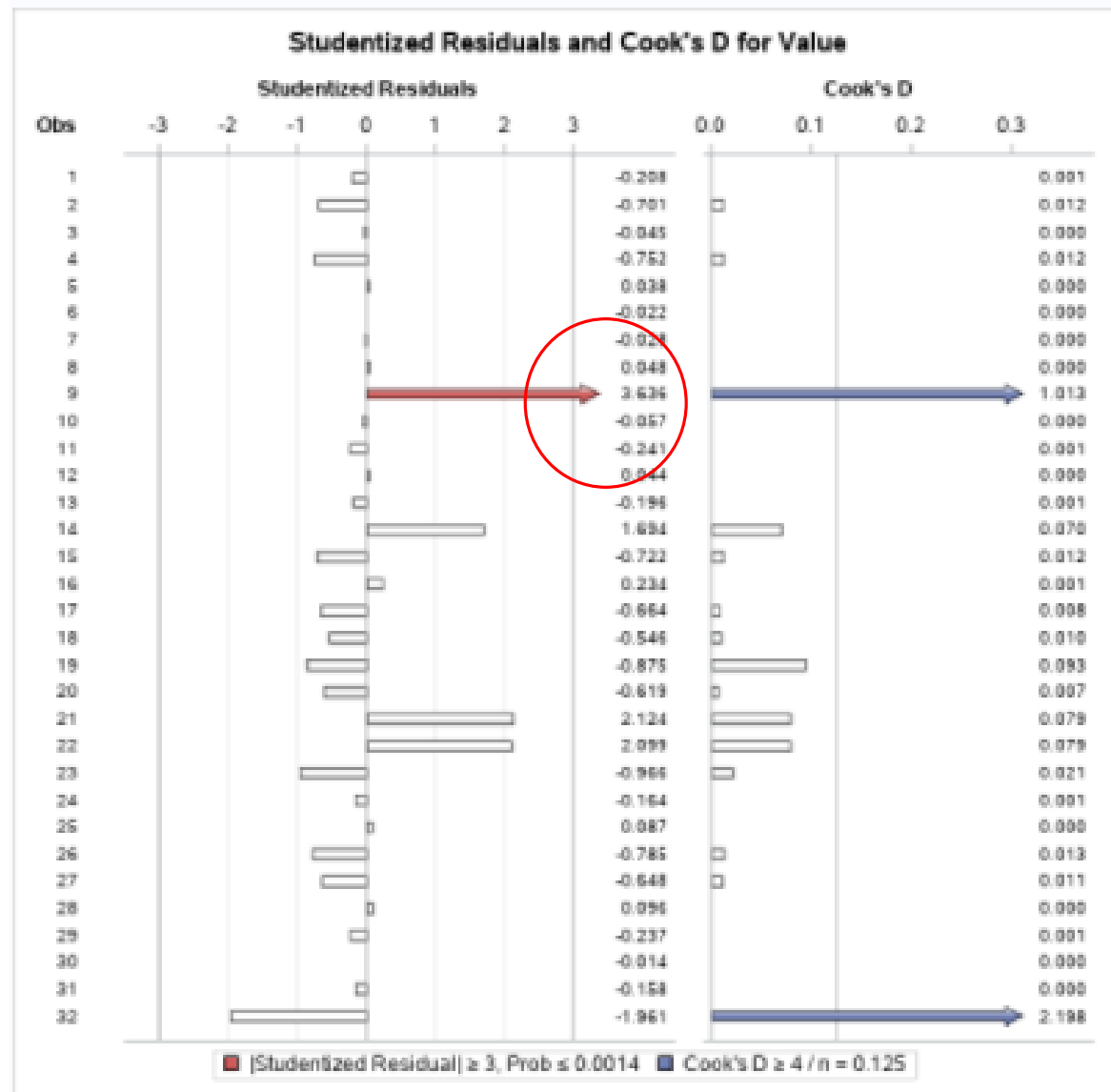
Yes we can use two plots of residuals to show the outliers is data :

One using studentized plot : one point is out side of the range, above 4

Two using error point out side of std of errors. one point is out side of the range, close or out of 3std of error

Following slides confirm this.





Obs	Team	Revenue	Value	studentresid	cook
1	Washington Redskins	327	1538	-1.98104	2.19753
2	Oakland Raiders	205	881	-0.98594	0.02129
3	New England Patriots	282	1324	-0.87514	0.09348
4	San Diego Chargers	207	888	-0.78503	0.01312
5	Buffalo Bills	208	885	-0.75242	0.01247
6	Jacksonville Jaguars	204	878	-0.72238	0.01235
7	Atlanta Falcons	203	872	-0.70147	0.01208
8	Miami Dolphins	232	1044	-0.66394	0.00827
9	San Francisco 49ers	201	885	-0.64793	0.01113
10	New Orleans Saints	213	937	-0.61904	0.00888
11	Minnesota Vikings	195	839	-0.54587	0.01004
12	Detroit Lions	204	917	-0.24142	0.00138
13	St Louis Rams	208	929	-0.23710	0.00124
14	Arizona Cardinals	203	914	-0.20838	0.00107
15	Houston Texans	239	1125	-0.19575	0.00090
16	Philadelphia Eagles	237	1118	-0.18428	0.00059
17	Tennessee Titans	218	994	-0.15804	0.00042
18	Denver Broncos	228	1081	-0.05683	0.00005
19	Baltimore Ravens	228	1082	-0.04518	0.00003
20	Cincinnati Bengals	205	941	-0.02828	0.00002
21	Chicago Bears	228	1084	-0.02188	0.00001
22	Tampa Bay Buccaneers	224	1053	-0.01415	0.00000
23	Carolina Panthers	221	1040	0.03820	0.00002
24	Green Bay Packers	218	1023	0.04398	0.00003
25	Cleveland Browns	220	1035	0.04789	0.00004
26	Pittsburgh Steelers	218	1015	0.08869	0.00013
27	Seattle Seahawks	215	1010	0.09841	0.00018
28	Kansas City Chiefs	214	1018	0.23441	0.00098
29	Indianapolis Colts	203	1078	1.89352	0.07043
30	New York Jets	213	1170	2.09901	0.07888
31	New York Giants	214	1178	2.12350	0.07895
32	Dallas Cowboys	289	1812	3.63581	1.01277



END