Assignment 5: Correlation

Sukanya Mukherjee

128041217

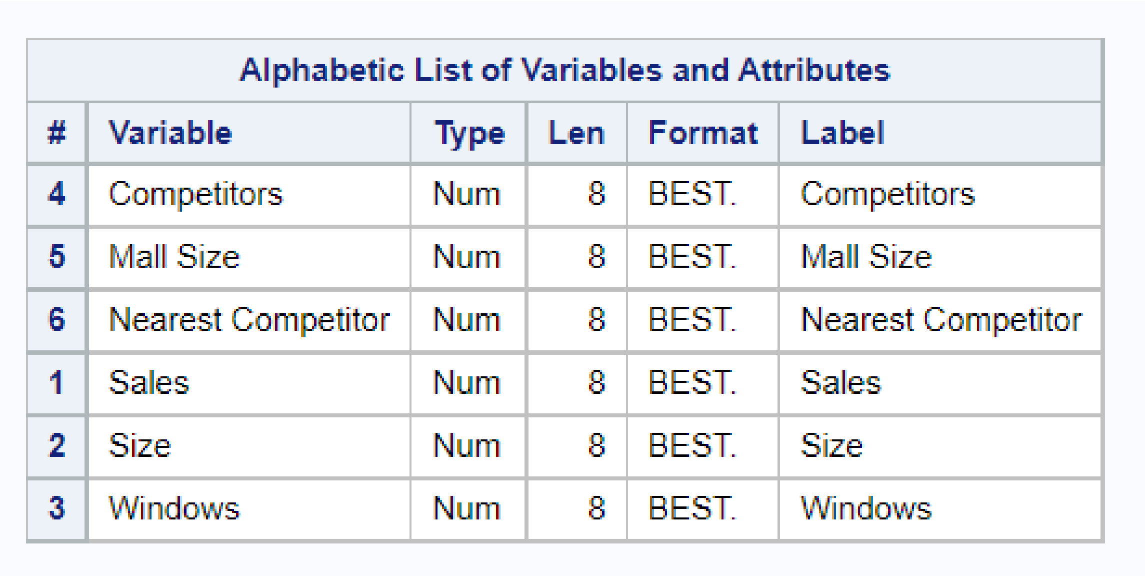
Statistics for Analytics

BAN100ZBB

Samaneh Gholami

Metadata

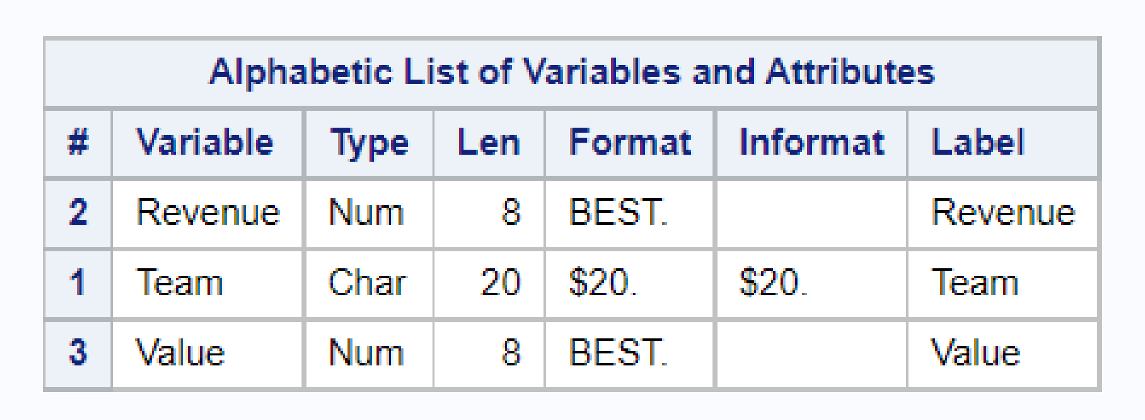
MALL DATASET



Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Label
4	Competitors	Num	8	BEST.	Competitors
5	Mall Size	Num	8	BEST.	Mall Size
6	Nearest Competitor	Num	8	BEST.	Nearest Competitor
1	Sales	Num	8	BEST.	Sales
2	Size	Num	8	BEST.	Size
3	Windows	Num	8	BEST.	Windows

NFL VALUES DATASET



Alphabetic List of Variables and Attributes

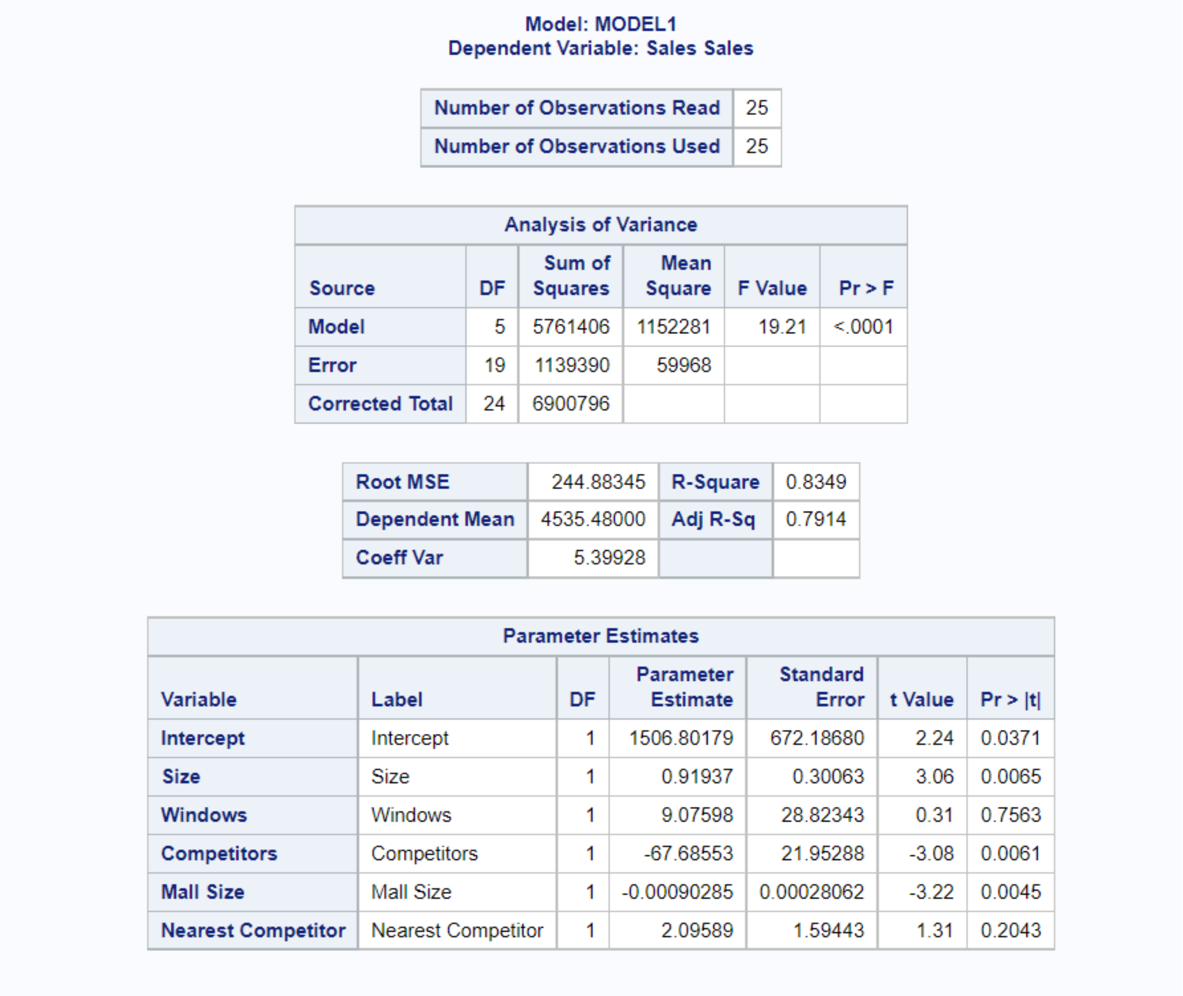
#	Variable	Type	Len	Format	Informat	Label
2	Revenue	Num	8	BEST.		Revenue
1	Team	Char	20	$20.	$20.	Team
3	Value	Num	8	BEST.		Value

Given Data

- **Problem 1 (10 marks) File: MALL. XLS**
- 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).
- a. Find a multiple regression model for the data.
- b. Interpret the values of the coefficients in the model.
- c. Test whether the model as a whole is significant. At the 0.05 level of significance, what is your conclusion?
- d. Use the model to predict monthly sales for each of the stores in the study.
- e. 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?
- f. Find and interpret the value of R! for this model.
- g. Do you think that this model will be useful in helping the planners? Why or why not?
- h. Test the individual regression coefficients. At the 0.05 level of significance, what are your conclusions?
- i. If you were going to drop just one variable from the model, which one would you choose? Why?
- **The store planners for the women's clothing chain want to find the best model that they can for understanding what store characteristics impact monthly sales.**
- j. Use stepwise regression to find the best model for the data.
- k. Analyze the model you have identified to determine whether it has any problems.
- l. Write a memo reporting your findings to your boss. Identify the strengths and weaknesses of the model you have chosen.

A. Multiple Regression Model

- First table (ANOVA Table):
- SSR = 5761406
- SSE = 1139390
- SST = 6900796
- Second Table: R square value. (explained further)
- Third Table: Value of Coefficients. (explained further)



Model: MODEL1
Dependent Variable: Sales Sales

Number of Observations Read	25
Number of Observations Used	25

Analysis of Variance

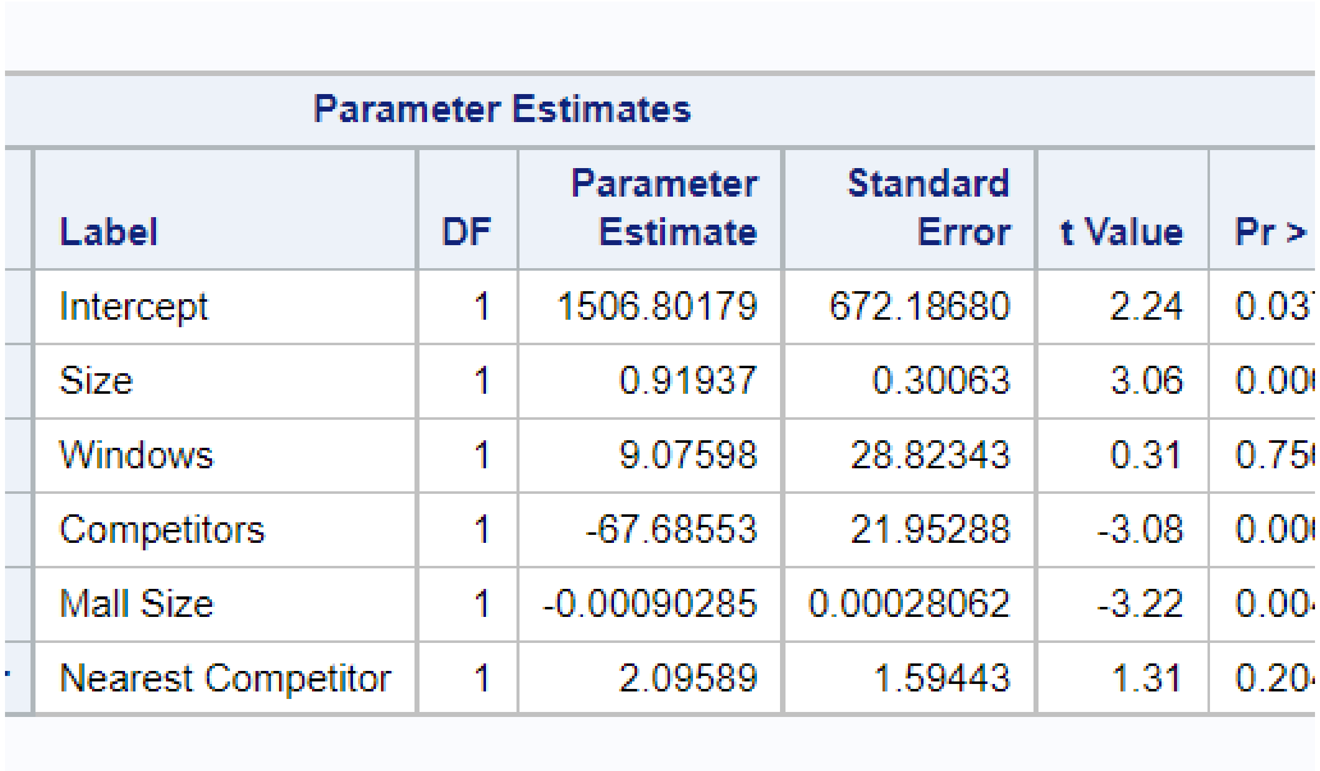
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	5761406	1152281	19.21	<.0001
Error	19	1139390	59968		
Corrected Total	24	6900796			

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
Mall Size	Mall Size	1	-0.00090285	0.00028062	-3.22	0.0045
Nearest Competitor	Nearest Competitor	1	2.09589	1.59443	1.31	0.2043

B. Value of Coefficients



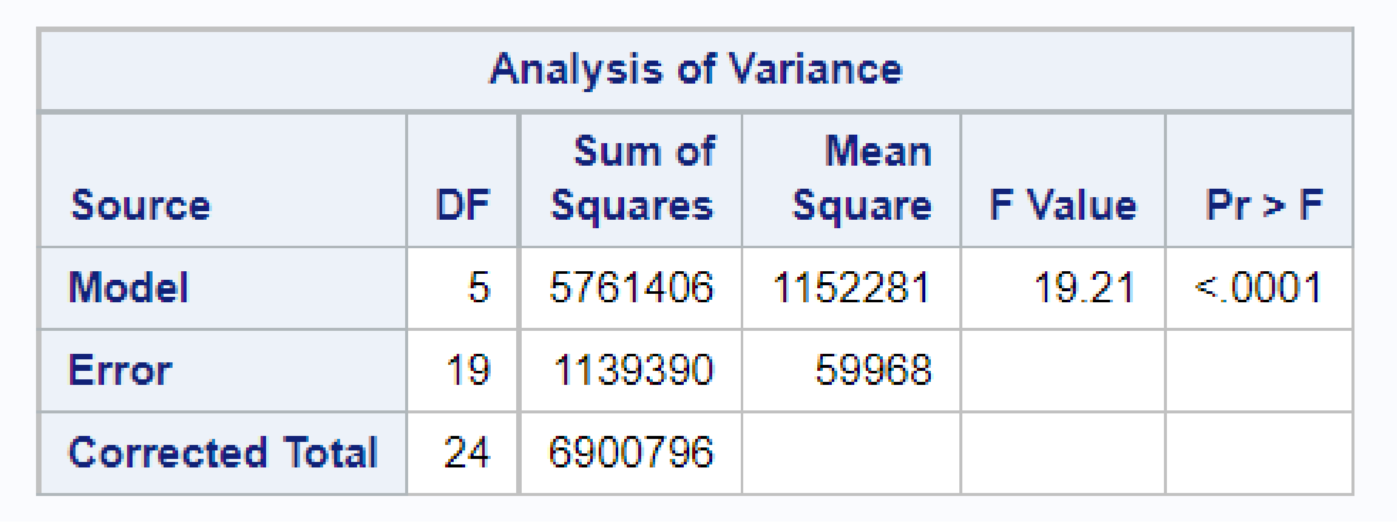
Parameter Estimates

Label	DF	Parameter Estimate	Standard Error	t Value	Pr >
Intercept	1	1506.80179	672.18680	2.24	0.03
Size	1	0.91937	0.30063	3.06	0.00
Windows	1	9.07598	28.82343	0.31	0.75
Competitors	1	-67.68553	21.95288	-3.08	0.00
Mall Size	1	-0.00090285	0.00028062	-3.22	0.00
Nearest Competitor	1	2.09589	1.59443	1.31	0.20

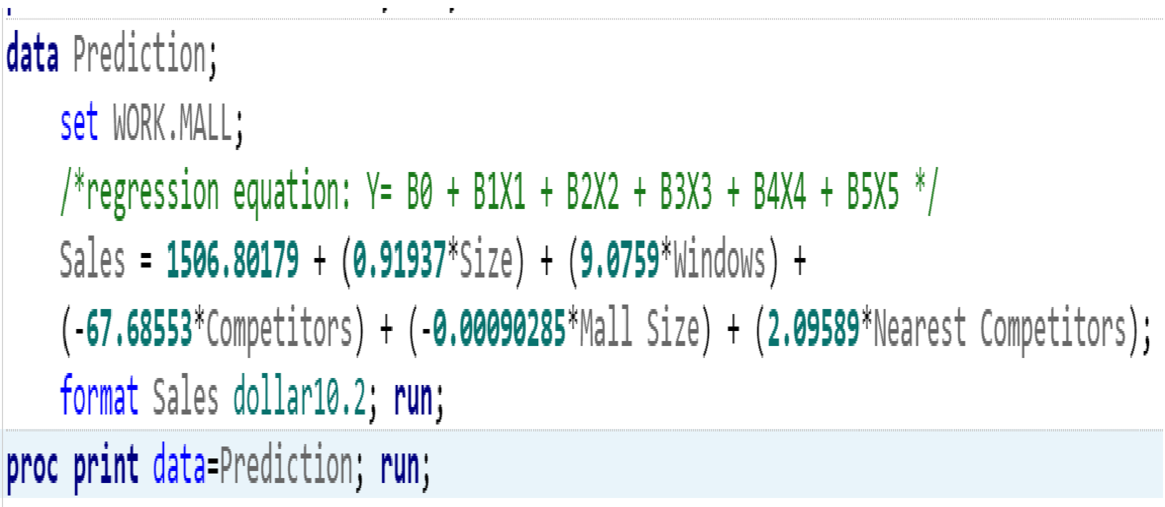
- Regression equation:
- Y= B0 + B1X1 + B2X2 + B3X3 + B4X4 + B5X5 .
- Where:
- B0 (Intercept) = 1506.80179
- B1(Size) = 0.91937
- B2 (Windows) = 9.0759
- B3 (Competitors) = - 67.68553
- B4 (Mall Size) = - 0.00090285
- B5 (Nearest Competitors) = 2.09589
- Size, Windows and Nearest Competitors are directly proportional to Sales.
- Competitors and Mall Size are inversely proportional to Sales.

C. Model as a whole

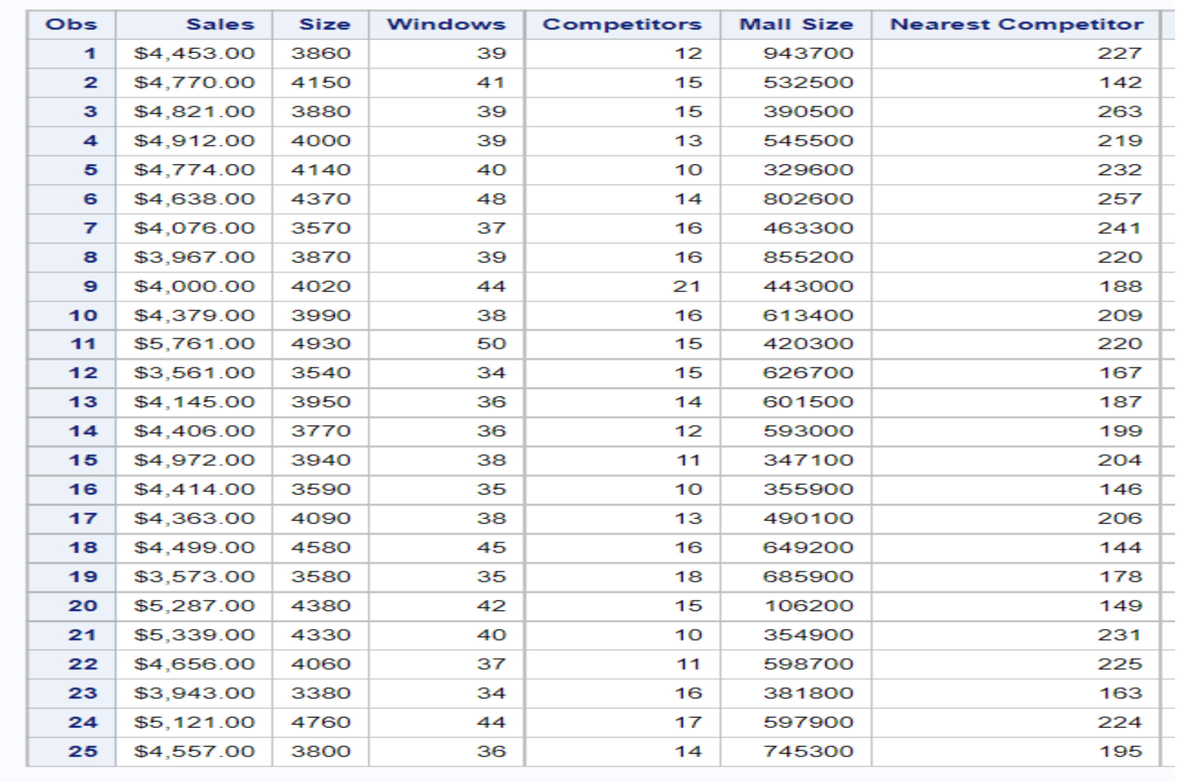
- As seen in the table, p-value (<0.0001) is less than alpha (0.05) , which means it is significant.



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			



```
data Prediction;
    set WORK.MALL;
    /*regression equation: Y= B0 + B1X1 + B2X2 + B3X3 + B4X4 + B5X5 */
    Sales = 1506.80179 + (0.91937*Size) + (9.0759*Windows) +
    (-67.68553*Competitors) + (-0.00090285*Mall Size) + (2.09589*Nearest Competitors);
    format Sales dollar10.2; run;
proc print data=Prediction; run;
```

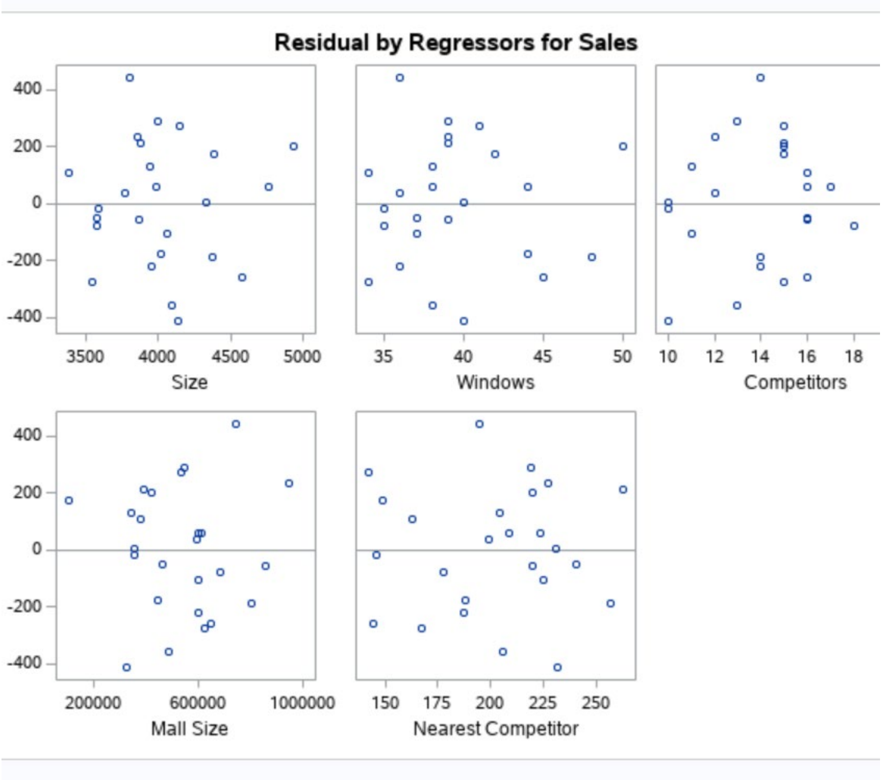


Obs	Sales	Size	Windows	Competitors	Mall Size	Nearest Competitor
1	$4,453.00	3860	39	12	943700	227
2	$4,770.00	4150	41	15	532500	142
3	$4,821.00	3880	39	15	390500	263
4	$4,912.00	4000	39	13	545500	219
5	$4,774.00	4140	40	10	329600	232
6	$4,638.00	4370	48	14	802600	257
7	$4,076.00	3570	37	16	463300	241
8	$3,967.00	3870	39	16	855200	220
9	$4,000.00	4020	44	21	443000	188
10	$4,379.00	3990	38	16	613400	209
11	$5,761.00	4930	50	15	420300	220
12	$3,561.00	3540	34	15	626700	167
13	$4,145.00	3950	36	14	601500	187
14	$4,406.00	3770	36	12	593000	199
15	$4,972.00	3940	38	11	347100	204
16	$4,414.00	3590	35	10	355900	146
17	$4,363.00	4090	38	13	490100	206
18	$4,499.00	4580	45	16	649200	144
19	$3,573.00	3580	35	18	685900	178
20	$5,287.00	4380	42	15	106200	149
21	$5,339.00	4330	40	10	354900	231
22	$4,656.00	4060	37	11	598700	225
23	$3,943.00	3380	34	16	381800	163
24	$5,121.00	4760	44	17	597900	224
25	$4,557.00	3800	36	14	745300	195

D. Model to predict monthly sales for each of the stores

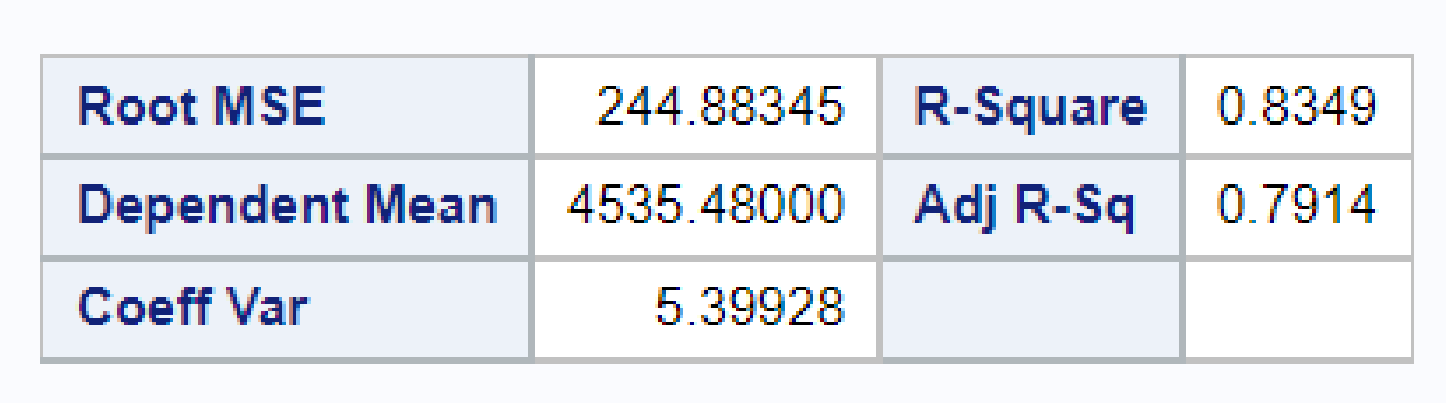
E. Residuals versus the Actual values

- No pattern identified
- The points are scattered all over.
- But least square method is not violated, so this proves to be a trustable data.



F. Value of R square

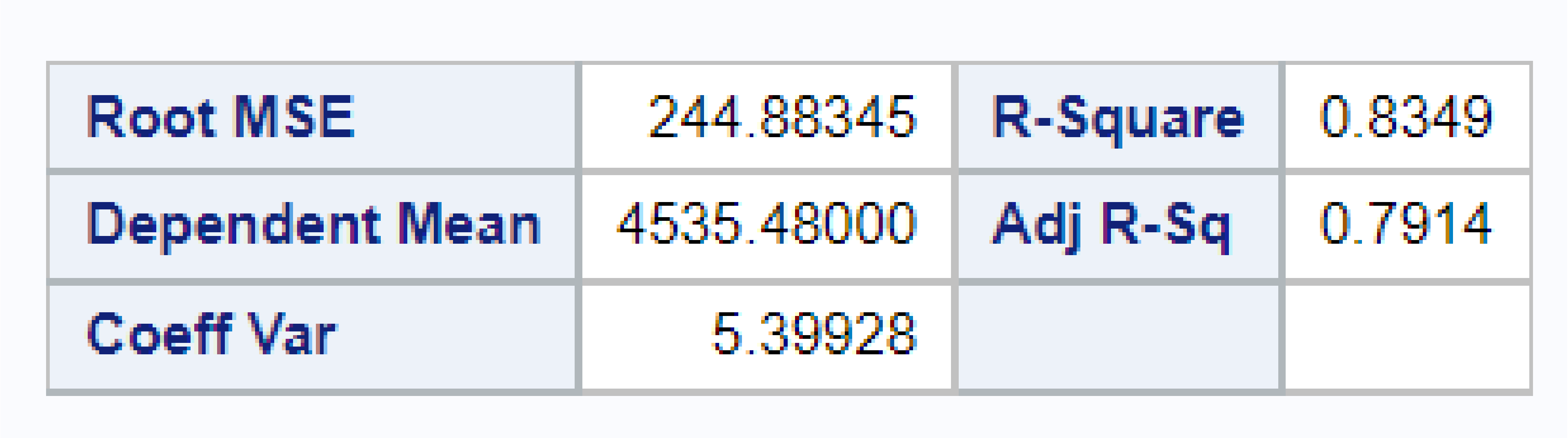
- R-square should be between 0 to 1 which is true.
- R-square is very close to 1 (0.8349) which means data is about 83% correct.



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

G. Model will be useful?

As the accuracy is 83%, this model is useful.



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

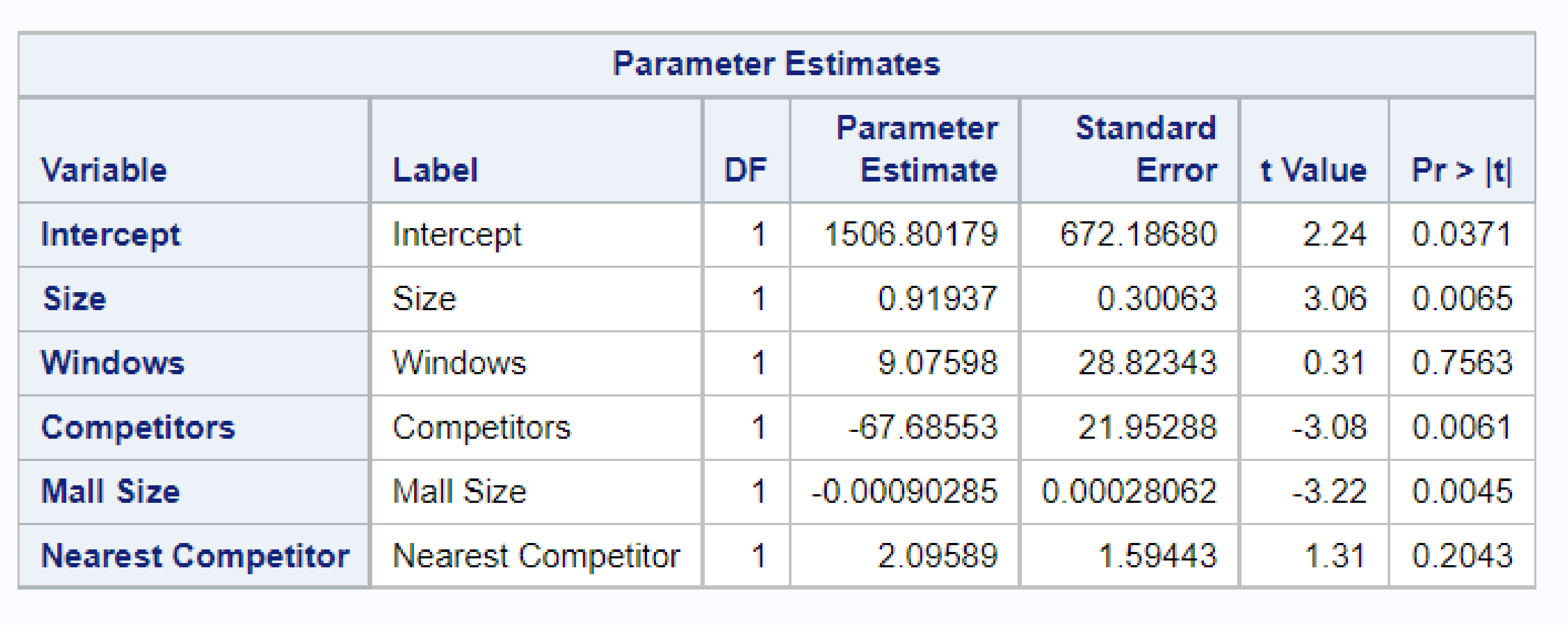
H. Test the individual regression coefficients

Null hypothesis: Variable 1 is not dependent on Variable2

Alternate hypothesis: Variable 1 is dependent on Variable2

P-values of **Size**, **Competitors** and **Mall Size** are less than alpha (0.05); hence we **reject the null hypothesis**, they are **dependent on Sales**

P-values of **Windows** and **Nearest Competitor** are more than alpha; hence we **fail to reject null hypothesis**, they are **independent of Sales**.



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	Mall Size	1	-0.00090285	0.00028062	-3.22	0.0045
Nearest Competitor	Nearest Competitor	1	2.09589	1.59443	1.31	0.2043

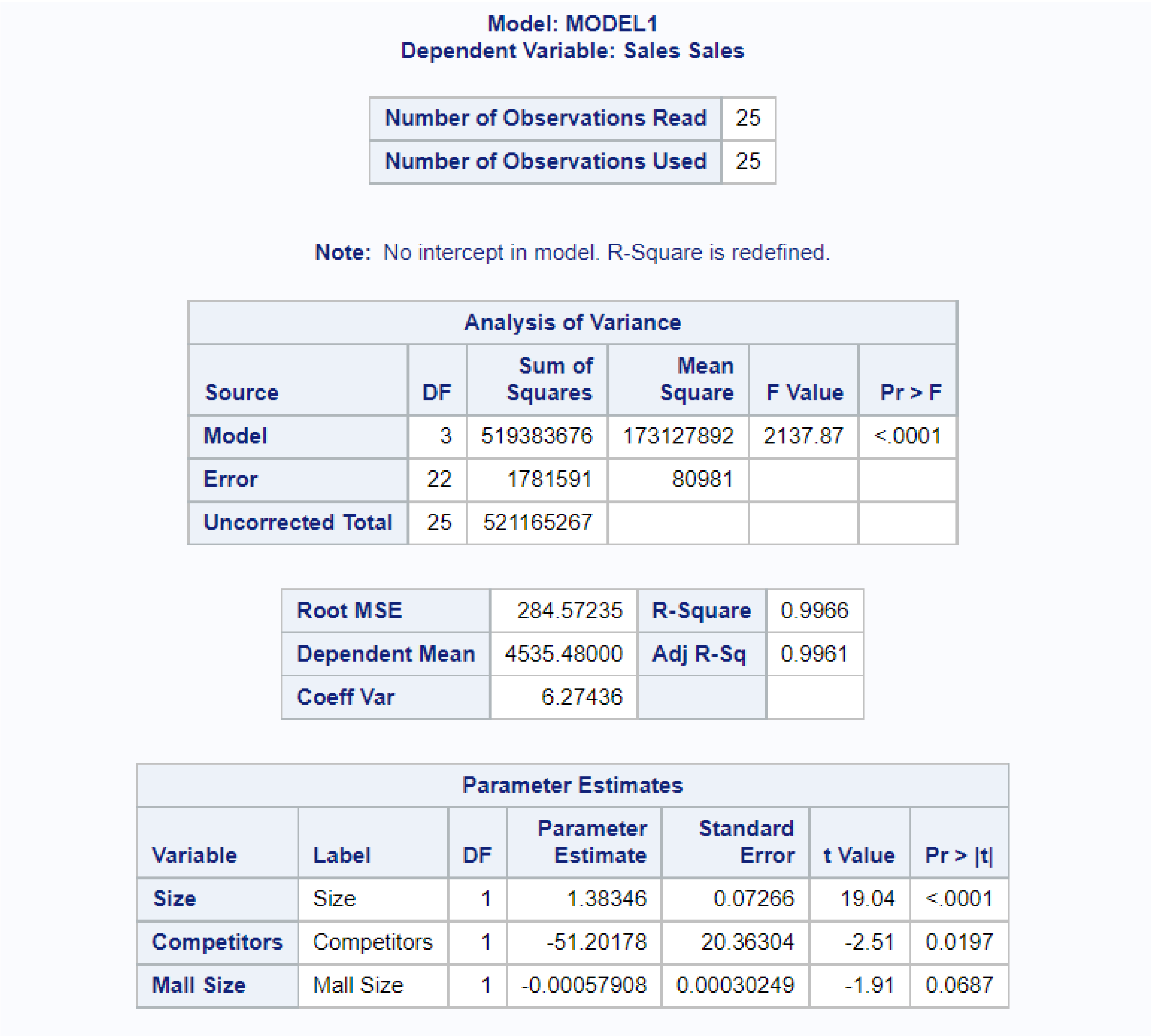
I. Drop just one variable from the model

Windows should be dropped as its value is 0.7563.

Dropping it can make the model even more robust and accurate.

J. Stepwise regression to find the best model

- Variables Size, Competitors and Mall Size is chosen for the model as overall R-square is 0.9966 which is almost 1 and perfect for the model and overall p-value is less than alpha.



Model: MODEL1
Dependent Variable: Sales Sales

Number of Observations Read	25
Number of Observations Used	25

Note: No intercept in model. R-Square is redefined.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	519383676	173127892	2137.87	<.0001
Error	22	1781591	80981		
Uncorrected Total	25	521165267			

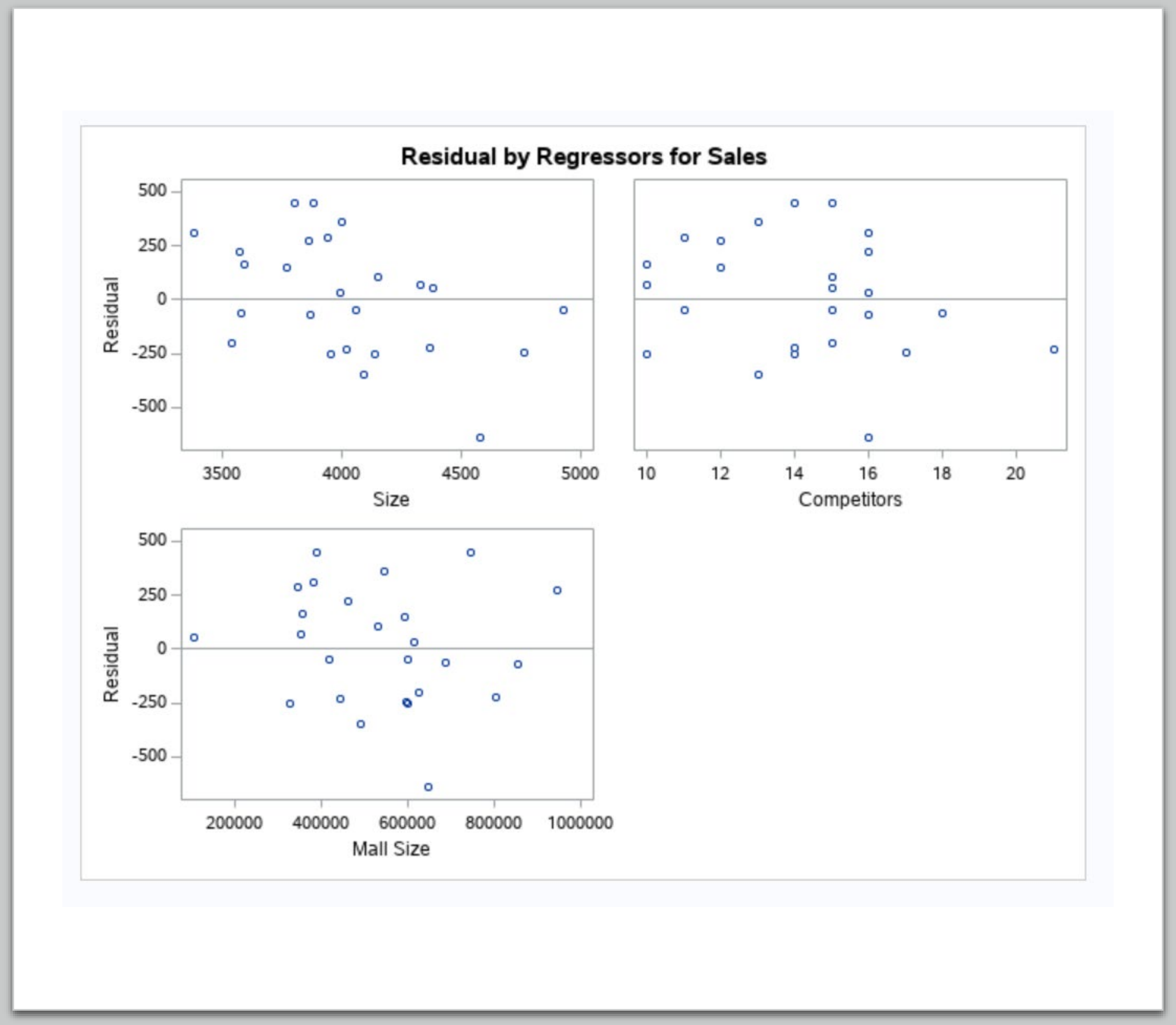
Root MSE	284.57235	R-Square	0.9966
Dependent Mean	4535.48000	Adj R-Sq	0.9961
Coeff Var	6.27436		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Size	Size	1	1.38346	0.07266	19.04	<.0001
Competitors	Competitors	1	-51.20178	20.36304	-2.51	0.0197
Mall Size	Mall Size	1	-0.00057908	0.00030249	-1.91	0.0687

K. Determine whether it has any problems

- No Specific pattern or trend is found, the model can be trusted.



L. Identify the strengths and weaknesses

Regression equation:

Y= B0 + B1X1 + B2X2 + B3X3 + B4X4 + B5X5

Sales = 1506.80179 + (0.91937*Size) + (9.0759*Windows) + (-67.68553*Competitors) + (-0.0090285*Mall Size) + (2.09589*Nearest Competitors)

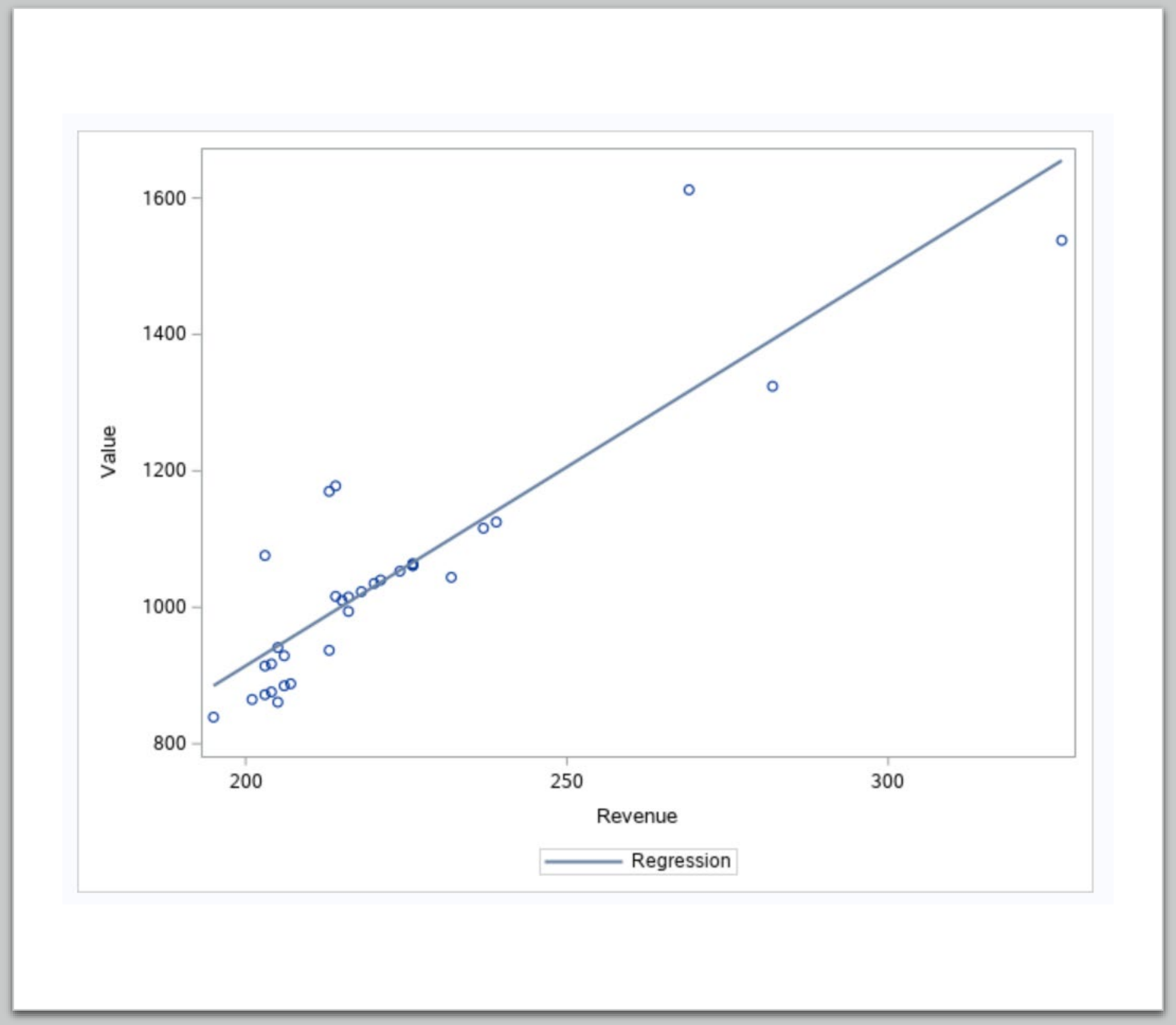
1. Weaknesses: Windows and Nearest Competitors should be removed.
2. Sales = 1506.80179 + (0.91937*Size) + (-67.68553*Competitors) + (-0.0090285*Mall Size)
3. Strengths: Size, Competitors and Mall size
4. Removing Windows and Nearest Competitors will make the R-square 0.9966 which is almost perfect and closest to 1.

Given Data

- **Problem 2 (10 marks)**
- **The File NFLValues.xlsx** show the annual revenue ($ millions) and the estimated team value ($ millions) for the 32 teams in the National Football League.
- a. 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?
- b. Develop the estimated regression equation that can be used to predict team value given the value of annual revenue.
- c. Use residual analysis to determine whether any outliers and/or influential observations are present. Briefly summarize your findings and conclusions.

A. Scatter diagram

- Most of the points are surrounding the slope line.
- Few points are quite far from the line and can be the outliers influencing the model.



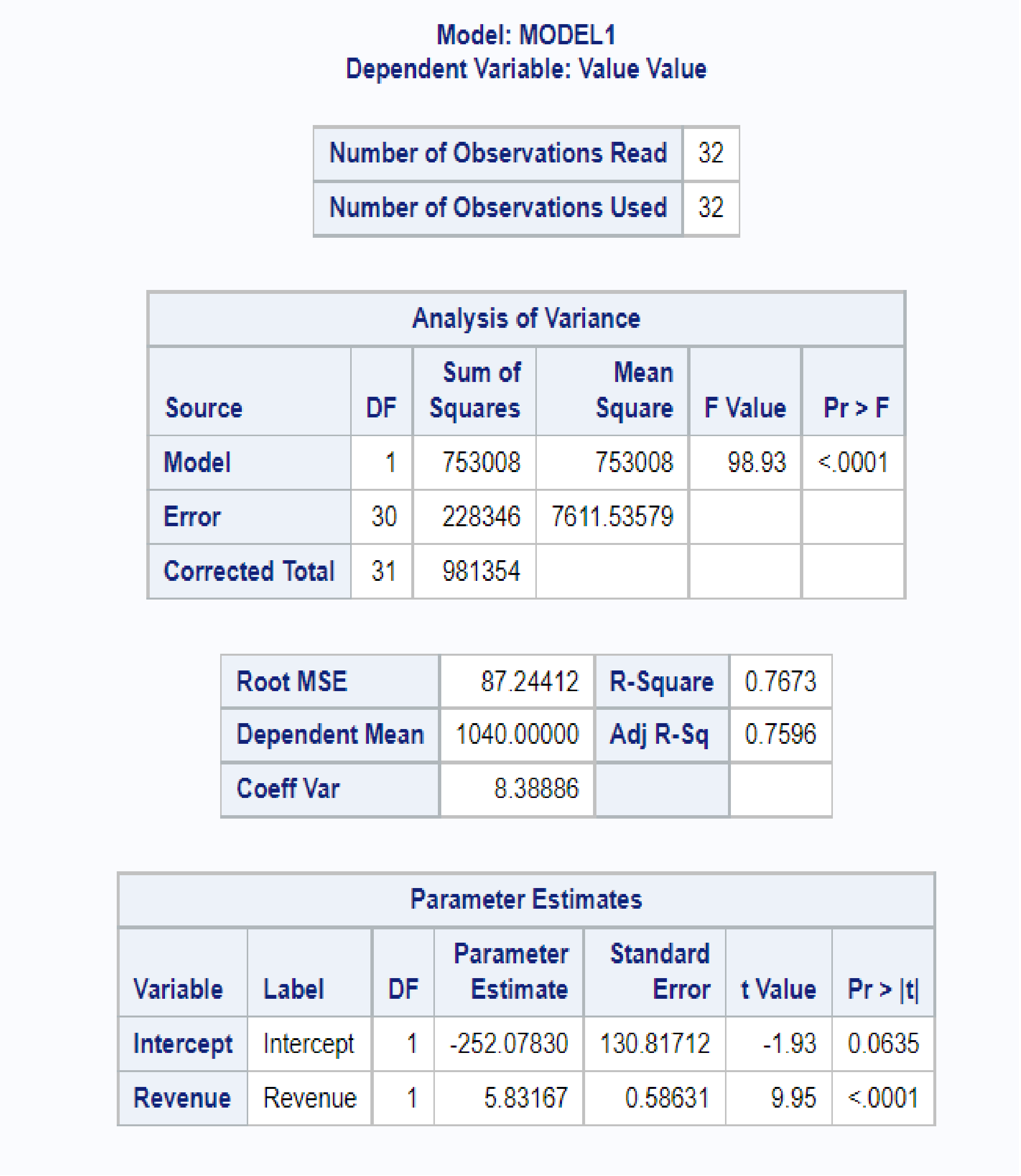
B. Estimated regression equation

- **First table (ANOVA):**
- SSR=753008
- SSE= 228346
- SST=981354
- Overall p-value (0.0001) is less than alpha.
- **Second table**: Overall r-square is 0.7673, which is betwee and 1 and close to 1, also tells us that the data is around 7 accurate.
- **Third table**: Revenue is directly proportional to Value.

Regression equation:

Y= B0 + B1X

Value = -252.07830 + 5.83167*Revenue



Model: MODEL1
Dependent Variable: Value Value

Number of Observations Read	32
Number of Observations Used	32

Analysis of Variance

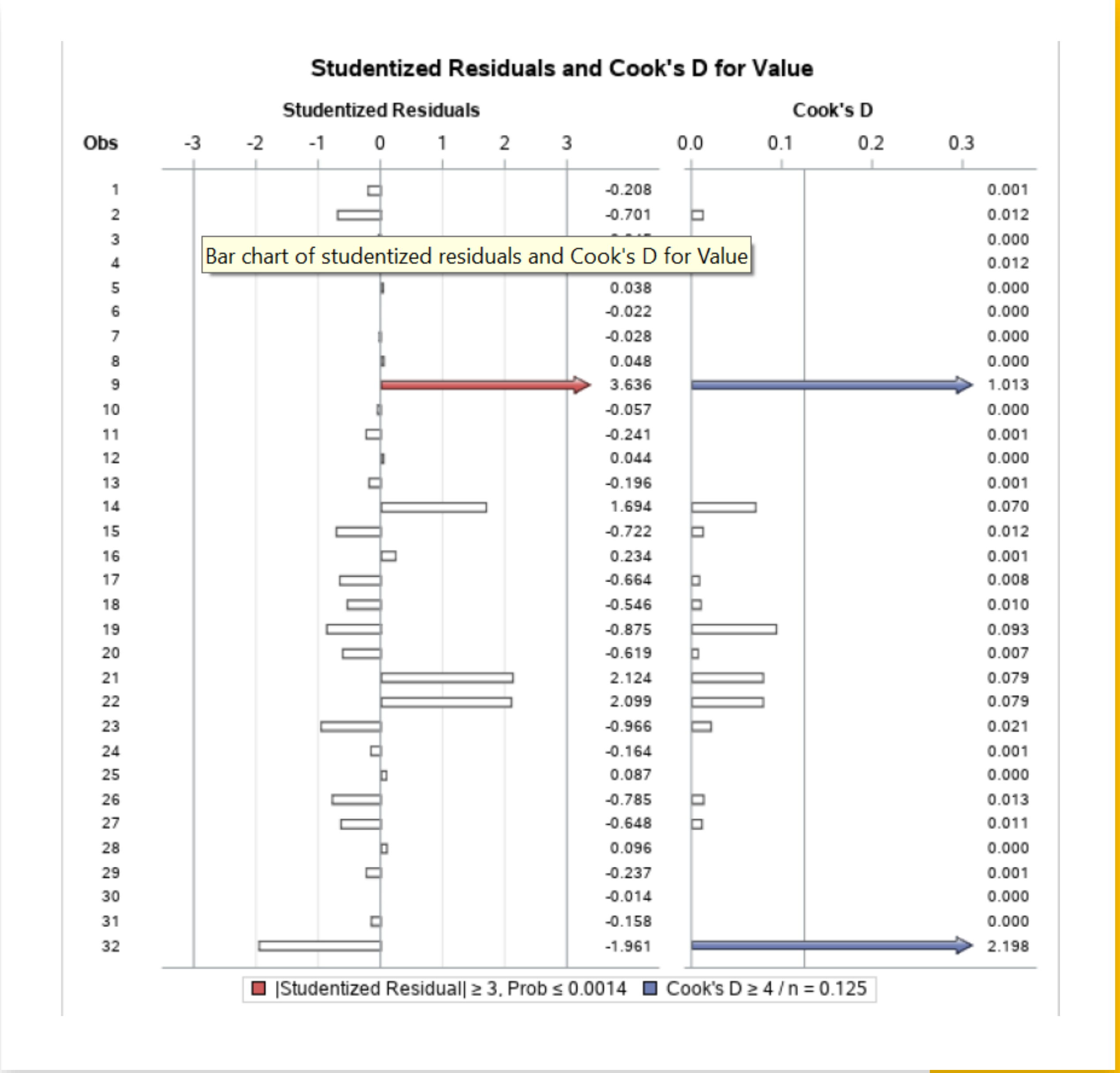
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	753008	753008	98.93	<.0001
Error	30	228346	7611.53579		
Corrected Total	31	981354			

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. Residual analysis to determine any outliers and/or influential observations



As seen in the graph, observation number 9 is an outlier and observation numbers 9 and 32 are influential for the dataset.