**Chapter 15**

**Multiple Regression**

**Case Problem 1: Consumer Research, Inc.**

Descriptive statistics for these data are shown below:

N MEAN MEDIAN TRMEAN STDEV SEMEAN

INCOME 50 43.48 42.00 43.41 14.55 2.06

SIZE 50 3.420 3.000 3.341 1.739 0.246

AMOUNT 50 3964 4090 3973 933 132

MIN MAX Q1 Q3

INCOME 21.00 67.00 30.00 55.00

SIZE 1.000 7.000 2.000 5.000

AMOUNT 1864 5678 3109 4747

The following scatter diagrams suggest a linear relationship.

Minitab was used to obtain the following regression analysis output:

The regression equation is

AMOUNT = 2204 + 40.5 INCOME

Predictor Coef SE Coef T p

Constant 2204.0 329.0 6.70 0.000

INCOME 40.480 7.184 5.63 0.000

S = 731.7 R-sq = 39.8% R-sq(adj) = 38.6%

Analysis of Variance

SOURCE DF SS MS F p

Regression 1 16999744 16999744 31.75 0.000

Residual Error 48 25699404 535404

Total 49 42699148

Unusual Observations

Obs. INCOME AMOUNT Fit Stdev.Fit Residual St.Resid

3 32.0 5100 3499 132 1601 2.22R

5 31.0 1864 3459 137 -1595 -2.22R

R denotes an observation with a large standardized residual

The regression equation is

AMOUNT = 2582 + 404 SIZE

Predictor Coef SE Coef T p

Constant 2581.9 195.3 13.22 0.000

SIZE 404.13 51.00 7.92 0.000

S = 620.8 R-sq = 56.7% R-sq(adj) = 55.8%

Analysis of Variance

SOURCE DF SS MS F p

Regression 1 24200718 24200718 62.80 0.000

Residual Error 48 18498432 385384

Total 49 42699152

Unusual Observations

Obs SIZE AMOUNT Fit SE Fit Residual St Resid

5 2.00 1864.0 3390.2 113.8 -1526.2 -2.50R

R denotes an observation with a large standardized residual

The regression equation is

AMOUNT = 1305 + 33.1 INCOME + 356 SIZE

Predictor Coef SE Coef T p

Constant 1304.9 197.7 6.60 0.000

INCOME 33.133 3.968 8.35 0.000

SIZE 356.30 33.20 10.73 0.000

S = 398.1 R-sq = 82.6% R-sq(adj) = 81.8%

Analysis of Variance

SOURCE DF SS MS F p

Regression 2 35250756 17625378 111.22 0.000

Residual Error 47 7448393 158476

Total 49 42699148

SOURCE DF SEQ SS

INCOME 1 16999744

SIZE 1 18251010

Unusual Observations

Obs INCOME AMOUNT Fit SE Fit Residual St Resid

3 32.0 5100.0 3790.3 76.9 1309.7 3.35R

5 31.0 1864.0 3044.6 83.9 -1180.6 -3.03R

11 25.0 4208.0 3202.1 91.6 1005.9 2.60R

R denotes an observation with a large standardized residual

The standardized residual plot for the model involving both independent variables is shown below:



Although the multiple regression model explains a high percentage of the variability in the dependent variable, the output identifies three observations as having a large standardized residual; thus, these 3 observations should be treated as possible outliers.

# Case Problem 2: Predicting Winnings for NASCAR Drivers

1. The Minitab output showing the sample correlation coefficients follows.

Winnings ($) Poles Wins Top 5

Poles 0.406

0.015

Wins 0.662 0.133

0.000 0.446

Top 5 0.861 0.437 0.725

0.000 0.009 0.000

Top 10 0.898 0.458 0.697 0.902

0.000 0.006 0.000 0.000

Cell Contents: Pearson correlation

The variable most highly correlated with Winnings ($) is the number of top ten finishes. A portion of the Minitab output that uses the Top 10 independent variable to predict Winnings ($) follows.

The regression equation is

Winnings ($) = 3049157 + 161934 Top 10

Predictor Coef SE Coef T P

Constant 3049157 171769 17.75 0.000

Top 10 161934 13831 11.71 0.000

S = 576313 R-Sq = 80.6% R-Sq(adj) = 80.0%

Analysis of Variance

Source DF SS MS F P

Regression 1 4.55270E+13 4.55270E+13 137.07 0.000

Residual Error 33 1.09605E+13 3.32137E+11

Total 34 5.64875E+13

2. A portion of the Minitab output follows.

The regression equation is

Winnings ($) = 3140367 - 12939 Poles + 13545 Wins + 71629 Top 5 + 117071 Top 10

Predictor Coef SE Coef T P

Constant 3140367 184229 17.05 0.000

Poles -12939 107205 -0.12 0.905

Wins 13545 111226 0.12 0.904

Top 5 71629 50667 1.41 0.168

Top 10 117071 33433 3.50 0.001

S = 581382 R-Sq = 82.0% R-Sq(adj) = 79.7%

Analysis of Variance

Source DF SS MS F P

Regression 4 4.63473E+13 1.15868E+13 34.28 0.000

Residual Error 30 1.01402E+13 3.38005E+11

Total 34 5.64875E+13

Source DF Seq SS

Poles 1 9.31517E+12

Wins 1 2.12230E+13

Top 5 1 1.16646E+13

Top 10 1 4.14449E+12

Looking at the *p*-values corresponding to the *t* values for each of the independent variables, the only significant variable is Top 10, with a *p*-value of .001. Also note that this model has an *R*2 of 0.82, while the model that included only Top 10 as an independent variable had an *R*2 of .806. Adding Poles, Wins, and Top 5 to the model as independent variables added little to the model’s ability to explain variation in Winnings.

3. A portion of the Minitab output follows.

The regression equation is

Winnings ($) = 3140367 - 12939 Poles + 202245 Wins + 188700 Top 2-5

+ 117071 Top 6-10

Predictor Coef SE Coef T P

Constant 3140367 184229 17.05 0.000

Poles -12939 107205 -0.12 0.905

Wins 202245 90226 2.24 0.033

Top 2-5 188700 34586 5.46 0.000

Top 6-10 117071 33433 3.50 0.001

S = 581382 R-Sq = 82.0% R-Sq(adj) = 79.7%

Analysis of Variance

Source DF SS MS F P

Regression 4 4.63473E+13 1.15868E+13 34.28 0.000

Residual Error 30 1.01402E+13 3.38005E+11

Total 34 5.64875E+13

Looking at the *p*-values corresponding to the *t* values for each of the independent variables, the only independent variable that is not significant is Poles, with a *p*-value of .905.

This occurred because we greatly reduced the multicolinearity in the model by replacing Top % with Top 2-5 and replacing Top 10 with Top 6-10. The Minitab output below provides evidence of this.

Winnings ($) Poles Wins Top 2-5

Poles 0.406

0.015

Wins 0.662 0.133

0.000 0.446

Top 2-5 0.823 0.489 0.537

0.000 0.003 0.001

Top 6-10 0.642 0.330 0.420 0.411

0.000 0.053 0.012 0.014

Cell Contents: Pearson correlation

4. A portion of the Minitab output follows.

The regression equation is

Winnings ($) = 3138094 + 204735 Wins + 186778 Top 2-5 + 116189 Top 6-10

Predictor Coef SE Coef T P

Constant 3138094 180327 17.40 0.000

Wins 204735 86427 2.37 0.024

Top 2-5 186778 30210 6.18 0.000

Top 6-10 116189 32102 3.62 0.001

S = 572067 R-Sq = 82.0% R-Sq(adj) = 80.3%

Analysis of Variance

Source DF SS MS F P

Regression 3 4.63424E+13 1.54475E+13 47.20 0.000

Residual Error 31 1.01451E+13 3.27261E+11

Total 34 5.64875E+13

The regression equation is

Winnings ($) = 3138094 + 204735 Wins + 186778 Top 2-5 + 116189 Top 6-10

Wins: An estimate of the expected increase in total winnings corresponding to having one more win, when all the other independent variables are held constant, is $204,735.

Top 2-5: An estimate of the expected increase in total winnings corresponding to having one more Top 2-5 finish, when all the other independent variables are held constant, is $186,778.

Top 6-10: An estimate of the expected increase in total winnings corresponding to having one more Top 6-10 finish, when all the other independent variables are held constant, is $116,189.

# Case Problem 3: Finding the Best Car Value

1. A portion of the Minitab output follows.

The regression equation is

Cost/Mile = 0.523 + 0.119 Family-Sedan + 0.230 Upscale-Sedan

Predictor Coef SE Coef T P

Constant 0.52308 0.01443 36.25 0.000

Family-Sedan 0.11892 0.01854 6.42 0.000

Upscale-Sedan 0.23026 0.01836 12.54 0.000

S = 0.0520301 R-Sq = 75.8% R-Sq(adj) = 74.9%

Analysis of Variance

Source DF SS MS F P

Regression 2 0.43254 0.21627 79.89 0.000

Residual Error 51 0.13806 0.00271

Total 53 0.57060

There appears to be a significant relationship between the type of car and the cost/mile. Approximately 75% of the variability in Cost/Mile is explained using the Family-Sedan and Upscale-Sedan variables. Note that for a small sedan, Family-Sedan = 0 and Upscale-Sedan = 1. Thus the estimate of the Cost/Mile for a small sedan is .523. Note that the Cost/Mile increases by .119 for a family sedan and .230 for an upscale sedan. Conclusion: smaller cars have lower five-year owner costs.

2. A portion of the Minitab output follows.

The regression equation is

Value Score = 1.37 - 2.27 Cost/Mile + 0.0111 Road-Test Score

+ 0.166 Predicted Reliability + 0.0228 Family-Sedan

+ 0.0681 Upscale-Sedan

Predictor Coef SE Coef T P

Constant 1.3710 0.1397 9.82 0.000

Cost/Mile -2.2659 0.1938 -11.69 0.000

Road-Test Score 0.011133 0.001313 8.48 0.000

Predicted Reliability 0.16621 0.01043 15.93 0.000

Family-Sedan 0.02278 0.03799 0.60 0.552

Upscale-Sedan 0.06811 0.05371 1.27 0.211

S = 0.0719876 R-Sq = 93.5% R-Sq(adj) = 92.9%

Analysis of Variance

Source DF SS MS F P

Regression 5 3.59669 0.71934 138.81 0.000

Residual Error 48 0.24875 0.00518

Total 53 3.84543

3. Family-Sedan and Upscale-Sedan are both not significant. The Minitab regression output corresponding to a model using only Cost/Mile, Road-Test Score, and Predicted Reliability follows.

The regression equation is

Value Score = 1.24 - 2.04 Cost/Mile + 0.0114 Road-Test Score

+ 0.165 Predicted Reliability

Predictor Coef SE Coef T P

Constant 1.24443 0.09273 13.42 0.000

Cost/Mile -2.0433 0.1047 -19.51 0.000

Road-Test Score 0.011377 0.001230 9.25 0.000

Predicted Reliability 0.16510 0.01016 16.26 0.000

S = 0.0721260 R-Sq = 93.2% R-Sq(adj) = 92.8%

Analysis of Variance

Source DF SS MS F P

Regression 3 3.5853 1.1951 229.73 0.000

Residual Error 50 0.2601 0.0052

Total 53 3.8454

4. The estimated regression equation developed in part (3) shows that the three best predictors of Value Score are Cost/Mile, Road-Test Score, and Predicted Reliability. So, it does appear that the analysis supports this claim. Moreover, these three variables are the variables that Consumer Reports used to calculate the values for Value Score. But, consider the following regression output in which the Cost/Mile independent variable is replaced by the two dummy variables, Family-Sedan and Upscale-Sedan.

The regression equation is

Value Score = 0.172 - 0.250 Family-Sedan - 0.455 Upscale-Sedan

+ 0.0112 Road-Test Score + 0.170 Predicted Reliability

Predictor Coef SE Coef T P

Constant 0.1719 0.1840 0.93 0.355

Family-Sedan -0.24995 0.05820 -4.29 0.000

Upscale-Sedan -0.45529 0.05757 -7.91 0.000

Road-Test Score 0.011166 0.002549 4.38 0.000

Predicted Reliability 0.17008 0.02024 8.40 0.000

S = 0.139741 R-Sq = 75.1% R-Sq(adj) = 73.1%

Analysis of Variance

Source DF SS MS F P

Regression 4 2.88858 0.72214 36.98 0.000

Residual Error 49 0.95685 0.01953

Total 53 3.84543

This regression output shows that the size of the car, as represented by the two dummy variables is also a significant factor in predicting Value Score. But, note that in part (1) the estimated regression equation shows that there is a significant relationship between Cost/Mile and the two dummy variables representing size. So, once the effect of Cost/Mile has been accounted for, any effects that might be due to size have already been incorporated into the model.