**Chapter 16**

**Regression Analysis: Model Building**

**Case Problem 1: Analysis of PGA Tour Statistics**

Descriptive statistics and the sample correlation coefficients for the data follow:

Variable N N\* Mean SE Mean StDev Minimum Q1 Median

Money ($) 125 0 1791113 92688 1036283 800694 1060815 1488214

Scoring Average 125 0 71.026 0.0377 0.422 70.110 70.700 71.020

DrDist 125 0 288.01 0.773 8.64 261.40 281.95 287.50

DrAccu 125 0 63.385 0.462 5.163 51.110 59.820 62.940

GIR 125 0 64.892 0.238 2.658 57.950 63.230 64.900

Sand Saves 125 0 50.029 0.501 5.606 36.800 45.685 50.640

PPR 125 0 29.227 0.0476 0.533 27.920 28.845 29.240

Scrambling 125 0 57.744 0.245 2.741 48.410 56.100 57.760

Bounce Back 125 0 18.599 0.216 2.411 13.750 16.895 18.720

Variable Q3 Maximum

Money ($) 2156281 6601094

Scoring Average 71.300 72.080

DrDist 293.90 315.10

DrAccu 66.625 73.950

GIR 66.895 71.100

Sand Saves 53.450 63.710

PPR 29.540 30.860

Scrambling 59.660 64.820

Bounce Back 20.205 26.210

We see that for the top 125 players the average earnings is $1,791,113, the average score is 70.03, the average distance per drive is 289.0, and so on.

Money ($) Scoring Average DrDist

Scoring Average -0.439

0.000

DrDist 0.185 0.070

0.039 0.437

DrAccu -0.240 -0.214 -0.618

0.007 0.017 0.000

GIR 0.122 -0.492 0.244

0.176 0.000 0.006

Sand Saves 0.285 -0.352 -0.266

0.001 0.000 0.003

PPR -0.111 0.134 0.382

0.217 0.136 0.000

Scrambling 0.111 -0.481 -0.531

0.216 0.000 0.000

Bounce Back 0.171 -0.225 0.189

0.056 0.012 0.035

DrAccu GIR Sand Saves

GIR 0.276

0.002

Sand Saves -0.042 -0.205

0.644 0.022

PPR 0.115 0.734 -0.485

0.201 0.000 0.000

Scrambling 0.274 -0.255 0.531

0.002 0.004 0.000

Bounce Back -0.059 0.094 0.039

0.514 0.295 0.667

PPR Scrambling

Scrambling -0.650

0.000

Bounce Back -0.051 0.029

0.571 0.745

Cell Contents: Pearson correlation

P-Value

The sample correlation coefficient between earnings and the average score is -.439; thus, lower scores are associated with higher earnings. In analyzing the data in an attempt to predict the average score, earnings would not be considered an independent variable; it is simply another output measure that has been used to rank the data.

The sample correlation coefficients show that the independent variable most highly correlated with the scoring average is the percentage of time a player is able to hit the green in regulation (GIR). Thus, the best single independent variable model uses GIR to predict Scoring Average. The corresponding Minitab regression output is shown below:

The regression equation is

Scoring Average = 76.1 - 0.0781 GIR

Predictor Coef SE Coef T P

Constant 76.0932 0.8096 93.99 0.000

GIR -0.07808 0.01247 -6.26 0.000

S = 0.368987 R-Sq = 24.2% R-Sq(adj) = 23.6%

Analysis of Variance

Source DF SS MS F P

Regression 1 5.3417 5.3417 39.23 0.000

Residual Error 123 16.7466 0.1362

Total 124 22.0883

The best single independent variable equation is able to explain 24% of the variation in Scoring Average. To investigate what other independent variables might be useful in predicting the average score we used Minitab’s best-subsets procedure.

Response is Scoring Average

B

S S o

a c u

n r n

d a c

D D m e

r r S b

D A a l B

i c G v P i a

Mallows s c I e P n c

Vars R-Sq R-Sq(adj) Cp S t u R s R g k

1 24.2 23.6 363.6 0.36899 X

1 23.1 22.5 370.4 0.37160 X

2 77.2 76.8 26.7 0.20317 X X

2 63.5 62.9 114.2 0.25702 X X

3 79.8 79.3 12.0 0.19196 X X X

3 78.4 77.9 20.7 0.19834 X X X

4 81.0 80.3 6.6 0.18716 X X X X

4 80.6 79.9 9.2 0.18915 X X X X

5 81.5 80.7 5.3 0.18536 X X X X X

5 81.2 80.4 7.4 0.18700 X X X X X

6 81.7 80.8 6.0 0.18513 X X X X X X

6 81.5 80.5 7.3 0.18614 X X X X X X

7 81.7 80.6 8.0 0.18590 X X X X X X X

This output indicates that three independent variables (GIR., PPR, and Scrambling) can be used to develop an estimated regression equation with R-Sq (adj) = 79.3. The Minitab regression output for this estimated regression equation follows:

The regression equation is

Scoring Average = 65.8 - 0.185 GIR + 0.662 PPR - 0.0362 Scrambling

Predictor Coef SE Coef T P

Constant 65.762 1.913 34.38 0.000

GIR -0.18502 0.01056 -17.52 0.000

PPR 0.66245 0.06702 9.88 0.000

Scrambling -0.036221 0.009152 -3.96 0.000

S = 0.191961 R-Sq = 79.8% R-Sq(adj) = 79.3%

Analysis of Variance

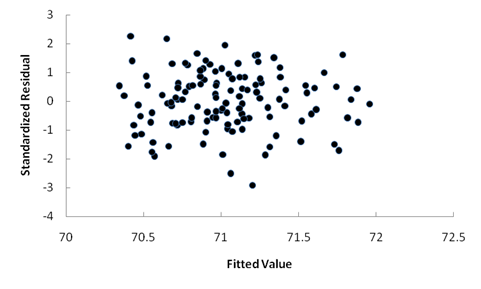
Source DF SS MS F P

Regression 3 17.6296 5.8765 159.48 0.000

Residual Error 121 4.4587 0.0368

Total 124 22.0883

At the .05 level of significance, the *p*-values (overall and individual) indicate that the estimated regression equation is highly significant. The corresponding standardized residual plot is shown below:



There does not appear to be anything in the residual plot that would lead us to question the assumptions regarding the error term.

Our analysis indicates that the three most important variables in terms of predicting the average score are GIR, PPR, and Scrambling. In the Minitab output that follows we show the results of adding DrDist as a fourth independent variable.

The regression equation is

Scoring Average = 68.3 - 0.181 GIR + 0.650 PPR - 0.0473 Scrambling

- 0.00626 DrDist

Predictor Coef SE Coef T P

Constant 68.316 2.091 32.67 0.000

GIR -0.18115 0.01040 -17.43 0.000

PPR 0.65008 0.06550 9.92 0.000

Scrambling -0.047304 0.009823 -4.82 0.000

DrDist -0.006262 0.002320 -2.70 0.008

S = 0.187161 R-Sq = 81.0% R-Sq(adj) = 80.3%

Analysis of Variance

Source DF SS MS F P

Regression 4 17.8848 4.4712 127.64 0.000

Residual Error 120 4.2035 0.0350

Total 124 22.0883

At the .05 level of significance, the *p*-values (overall and individual) indicate that the estimated regression equation is highly significant. But, the addition of DrDist to the estimated regression equation only increases the value of R-Sq (adj) by 1%.

The following Minitab output shows the estimated regression equation obtained by adding a fifth independent variables, SandSaves.

The regression equation is

Scoring Average = 69.1 - 0.179 GIR + 0.621 PPR - 0.0430 Scrambling

- 0.00619 DrDist - 0.00663 Sand Saves

Predictor Coef SE Coef T P

Constant 69.071 2.112 32.70 0.000

GIR -0.17867 0.01038 -17.21 0.000

PPR 0.62097 0.06680 9.30 0.000

Scrambling -0.04304 0.01000 -4.30 0.000

DrDist -0.006189 0.002298 -2.69 0.008

Sand Saves -0.006626 0.003624 -1.83 0.070

S = 0.185360 R-Sq = 81.5% R-Sq(adj) = 80.7%

Analysis of Variance

Source DF SS MS F P

Regression 5 17.9997 3.5999 104.78 0.000

Residual Error 119 4.0886 0.0344

Total 124 22.0883

At the .05 level of significance, the *p*-values (overall and individual) indicate that the estimated regression equation is highly significant. But, the addition of SandSaves is not significant because the p-value = .070 > *α* = .05. However, at the .10 level of significance, SandSaves would have been significant. But, the increase in explanatory power as compared to the previous four independent variable equation is negligible.

In terms of predictive ability it appears that GIR, PPR, and Scrambling are the most important factors. It may be surprising that DrDist and DrAccu were not important factors in predicting Average Score, or that SandSaves was not a key factor. This does not mean that these variables are not important factors in how well a player scores, but that at the level the top 125 PGA Tour pros players have reached, these variables are not as important in terms of achieving lower scores than GIR, PPR, and Scrambling.

**Case Problem 2: Rating Wines from the Piedmont Region of Italy**

1. A table showing the nunber of wines and average price for each rating follows.

|  |  |  |
| --- | --- | --- |
| **Rating** | **Number** | **Average Price ($)** |
| Classic | 7 | 269.57 |
| Outstanding | 40 | 73.00 |
| Very Good | 45 | 30.36 |
| Good | 7 | 16.71 |
| Mediocre | 1 | 21.00 |

None of the wines reviewed received a Not Recommended rating and ony one one wine was rated Mediocre. Overall, 85% of the wines received a Very Good or Outstanading rating. But, with the exception of the one wine that was rated Mediocre, the average price paid is greater for wines that are rated higher.

2. The scatter diagram follows.

The relationship between price and score does not appear to be linear. The scatter diagram indicates a curvinliear relationship between the two variables.

3. A portion of the Minitab output follows.

The regression equation is

Score = 87.8 + 0.0280 Price

Predictor Coef SE Coef T P

Constant 87.7632 0.3425 256.27 0.000

Price 0.027995 0.003419 8.19 0.000

S = 2.65911 R-Sq = 40.6% R-Sq(adj) = 40.0%

Analysis of Variance

Source DF SS MS F P

Regression 1 473.96 473.96 67.03 0.000

Residual Error 98 692.95 7.07

Total 99 1166.91

4. A portion of the Minitab output follows.

The regression equation is

Score = 86.2 + 0.0713 Price - 0.000113 PriceSq

Predictor Coef SE Coef T P

Constant 86.1660 0.4496 191.66 0.000

Price 0.071307 0.009390 7.59 0.000

PriceSq -0.00011327 0.00002320 -4.88 0.000

S = 2.39466 R-Sq = 52.3% R-Sq(adj) = 51.3%

Analysis of Variance

Source DF SS MS F P

Regression 2 610.67 305.34 53.25 0.000

Residual Error 97 556.24 5.73

Total 99 1166.91

5. Standardized residual plots for each type of model follow.

**Standardized Residual Plot for Simple Linear Regression**



**Standardized Residual Plot for the Second-Order Model**



The second-order model removes the curvilinear effect present in the data. In additition the second-order model provides a better fit; R-Sq (adj) = 51.3% for the second-order model compared to R-Sq = 40.6% for simple linear regression.

6. A portion of the Minitab output follows.

The regression equation is

Score = 77.7 + 3.16 lnPrice

Predictor Coef SE Coef T P

Constant 77.731 1.047 74.22 0.000

lnPrice 3.1559 0.2736 11.53 0.000

S = 2.24734 R-Sq = 57.6% R-Sq(adj) = 57.2%

Analysis of Variance

Source DF SS MS F P

Regression 1 671.96 671.96 133.05 0.000

Residual Error 98 494.95 5.05

Total 99 1166.91

The model using the natural logarithm of Price provides a slightly better fit than the second-order model, with R-Sq (adj) = 57.2%.

Looking at a scatter plot showing the equation of the second-order estimated regression line as compared to a scatter plot showing the line fr the logaritmic estimated regression line helps illustrate how the two models differ in providing a fit for these data.

7. Based upon the judgment of *Wine Spectator* reviewers, spending more for a bottle of wine will, in general, provide a better wine. But, because each individual usually has a different opion regarding the taste of a particular wine, this may not be true for you.

8. A scatter diagram showing only the wines with a price of $30 or less follows.

There does not appear to be much of a relationship between price and score for this group of wines. In fact, there is only one bottle of wine in this group that received a rating of less than Good. So, you could probably pick a wine at random from this group and end up with a wine that *Wine Spectator* would rate as Good or Very Good.