**Chapter 14**

**Simple Linear Regression**

**Case Problem 1: Measuring Stock Market Risk**

a. Selected descriptive statistics follow:

**Variable** **N Mean StDev Minimum Median Maximum**

Microsoft 36 0.00503 0.04537 -0.08201 0.00400 0.08883

Exxon Mobil 36 0.01664 0.05534 -0.11646 0.01279 0.23217

Caterpillar 36 0.03010 0.06860 -0.10060 0.04080 0.21850

Johnson & Johnson 36 0.00530 0.03487 -0.05917 -0.00148 0.10334

McDonald’s 36 0.02450 0.06810 -0.11440 0.03700 0.18260

Sandisk 36 0.06930 0.19540 -0.28330 0.07410 0.50170

Qualcomm 36 0.02840 0.08620 -0.12170 0.03870 0.21060

Procter & Gamble 36 0.01059 0.03707 -0.05365 0.01333 0.08783

S&P 500 36 0.01010 0.02633 -0.03429 0.01034 0.08104

From the descriptive statistics we see that six of the companies had a higher mean monthly return than the market (as measured by the S&P 500): Exxon Mobil, Caterpillar, McDonald’s, Sandisk, Qualcomm, and Procter & Gamble. Microsoft and Johnson & Johnson had lower mean monthly returns.

Using the standard deviation as a measure of volatility, Sandisk was the most volatile stock with a standard deviation of .1954. The stocks of Johnson & Johnson and P & G exhibit less volatility than the other individual stocks. But, all of the individual stocks are more volatile than the market as a whole. The diversification embodied in the S&P 500 reduces its volatility.

b. The estimated regression equation relating each of the individual stocks to the S&P 500 is shown below. The value of for each equation is also shown.

Microsoft = 0.00040 + 0.458 S&P 500 R-Sq = 7.1%

Exxon Mobil = 0.00926 + 0.731 S&P 500 R-Sq = 12.1%

Caterpillar = 0.015000 + 1.49 S&P 500 R-Sq = 32.9%

Johnson & Johnson = 0.00521 + 0.009 S&P 500 R-Sq = 0.0%

McDonald’s = 0.00930 + 1.500 S&P 500 R-Sq = 33.8%

Sandisk = 0.04300 + 2.600 S&P 500 R-Sq = 12.3%

Qualcomm = 0.01410 + 1.410 S&P 500 R-Sq = 18.7%

Procter & Gamble = 0.00548 + 0.507 S&P 500 R-Sq = 12.9%

The betas (slope of estimated regression equation) for the individual stocks can be obtained from the regression output.

**Company Beta**

Microsoft .458

Exxon Mobil .731

Caterpillar 1.490

Johnson & Johnson .009

McDonald’s 1.500

Sandisk 2.600

Qualcomm 1.410

The beta for the market as a whole is 1. So, any stock with a beta greater than 1 will move up faster than the market when the market goes up. Any stock with a beta less than 1 will not go down as fast as the market in periods where the market declines.

We would expect Sandisk, with a beta of 2.6, to benefit most from an up market. Johnson & Johnson, with a beta of .009 is least affected by the market. The effect of the market going down cannot be expected to exert much downward pressure on shares of Johnson & Johnson.

c. The values seem to indicate that from 0% to 33.8% of the variability of the returns in these individual stocks is explained by the return for the market.

**Case Problem 2: U.S. Department of Transportation**

Descriptive statistics for these data are shown below:

N MEAN MEDIAN TRMEAN STDEV SEMEAN

PERCENT 42 12.262 12.000 12.184 3.132 0.483

FATAL 42 1.922 1.881 1.906 1.071 0.165

MIN MAX Q1 Q3

PERCENT 8.000 18.000 9.000 15.000

FATAL 0.039 4.100 0.992 2.824

The following scatter diagram suggests a linear relationship between these two variables:

Minitab was used to develop the following regression analysis output:

The regression equation is

FATAL = - 1.60 + 0.287 PERCENT

Predictor Coef SE Coef t-ratio p

Constant -1.5974 0.3717 -4.30 0.000

PERCENT 0.28705 0.02939 9.77 0.000

S = 0.5894 R-sq = 70.5% R-sq(adj) = 69.7%

Analysis of Variance

SOURCE DF SS MS F p

Regression 1 33.134 33.134 95.40 0.000

Residual Error 40 13.893 0.347

Total 41 47.028

Unusual Observations

Obs. PERCENT FATAL Fit Stdev.Fit Residual St.Resid

15 10.0 0.0390 1.2731 0.1126 -1.2341 -2.13R

23 8.0 2.1900 0.6990 0.1548 1.4910 2.62R

R denotes an obs. with a large st. resid.

There is a significant relationship between the two variables. Two observations are identified as having a large standardized residual and should be treated as possible outliers; the following standardized residual plot does not indicate any other problems with the residuals.



Conclusion: It appears that the number of fatal accidents per 1000 licenses is linearly related to the percentage of licensed drivers under the age of 21; that is, the higher the percentage of drivers under 21, the larger the number of total accidents.

# Case Problem 3: Selecting a Point and Shoot Digital Camera

1. Descriptive statistics for the data set follow.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Price ($)** | **Megapixels** | **Weight (oz.)** | **Score** |
| Mean | 175.36 | 12.86 | 5.82 | 56.36 |
| Standard Error | 15.65 | 0.35 | 0.19 | 1.27 |
| Median | 160 | 12 | 6 | 56.5 |
| Mode | 200 | 12 | 5 | 66 |
| Standard Deviation | 82.80 | 1.84 | 0.98 | 6.70 |
| Sample Variance | 6855.42 | 3.39 | 0.97 | 44.83 |
| Kurtosis | 0.66 | -0.63 | -1.19 | -0.62 |
| Skewness | 1.06 | 0.23 | -0.12 | -0.43 |
| Range | 320 | 6 | 3 | 24 |
| Minimum | 80 | 10 | 4 | 42 |
| Maximum | 400 | 16 | 7 | 66 |
| Sum | 4910 | 360 | 163 | 1578 |
| Count | 28 | 28 | 28 | 28 |

The sample correlation coefficients for this data set follow.

Score Price ($) Megapixels

Price ($) 0.683

0.000

Megapixels -0.008 0.139

0.969 0.481

Weight (oz.) 0.286 0.349 -0.199

0.141 0.069 0.310

With a sample correlation coefficient of .683, price appears to be the best predictor of the overall score.

2. Scatter diagrams for the data are shown below.

There appears to be a positive relationship between the price of the camera and the overall score. But, observation 17, a Nikon camera with a price of $400, appears to be an observation that will have a significant impact when we fit a linear model to these data. It may be worth considering restricting the analysis to cameras that have a price of less than $400. Another possible explanation for what we observe here is that the underlying relationship may not be linear. In other words, the somewhat curvilinear trend in the data may be due to diminishing returns.

The number of megapixels does not appear to have much effect on the overall score. But, note that as the number of megapixels increase from 10 to 14, the overall score appears to have a downward trend; that is the overall score is decreasing. This seems to be counterintuitive in that generally speaking, higher megapixel cameras are usually considered to have better picture quality. But, the overall score for the 16 megapixel cameras does increase somewhat.

There may be a modest increase in overall score for cameras that weigh more. Also note the large variability in the score for cameras with a weight of 5 ounces and cameras with a weight of 7 ounces. The pattern in the data may also be an indication that the effect of weight may also involve some curvilinear effect.

Conclusion: The variable that appears to be the best predictor of overall score is the price of the camera.

3. A portion of the Minitab output follows:

The regression equation is

Score = 46.7 + 0.0552 Price ($)

Predictor Coef SE Coef T P

Constant 46.669 2.238 20.85 0.000

Price ($) 0.05525 0.01158 4.77 0.000

S = 4.98238 R-Sq = 46.7% R-Sq(adj) = 44.6%

Analysis of Variance

Source DF SS MS F P

Regression 1 565.00 565.00 22.76 0.000

Residual Error 26 645.43 24.82

Total 27 1210.43

Unusual Observations

Price

Obs ($) Score Fit SE Fit Residual St Resid

17 400 59.000 68.768 2.767 -9.768 -2.36RX

28 130 42.000 53.851 1.078 -11.851 -2.44R

R denotes an observation with a large standardized residual.

X denotes an observation whose X value gives it large leverage.

With a p-value = .000, price is a significant factor in predicting the overall score. The estimated regression equation explained 46.7% of the variability in the overall score. Note that there are two observations that are unusual: observation 17 and observation 28. But, observation 17 is listed as being an observation with a large leverage and thus it is considered to be an influential observation. To confirm this conclusion, the following regression output show the results after removing observation 17 from the data.

The regression equation is

Score = 44.2 + 0.0724 Price ($)

Predictor Coef SE Coef T P

Constant 44.167 2.240 19.72 0.000

Price ($) 0.07239 0.01236 5.85 0.000

S = 4.50538 R-Sq = 57.8% R-Sq(adj) = 56.1%

Analysis of Variance

Source DF SS MS F P

Regression 1 695.72 695.72 34.27 0.000

Residual Error 25 507.46 20.30

Total 26 1203.19

Note that the slope of the estimated line without observation 17 is .0724 as compared to the slope of .0552 with observation 17. And, the fit has also improved.

Are we justified in simply discarding observation 17 just because it is influential and provides a better fit? No. But, if we are interested in only exploring the relationship between price and overall score for cameras that cost less than $400, then removing observation 17 from the data set is acceptable.

4. Using only the data for the Canon cameras, the scatter diagram using the price of the camera as the independent variable follows.

There does appear to be a relationship between the price of the camera and the overall score. But, the relationship appears to be curvilinear. However, using simple linear regression for these data we obtain the following output.

The regression equation is

Score = 47.3 + 0.0665 Price ($)

Predictor Coef SE Coef T P

Constant 47.288 2.573 18.38 0.000

Price ($) 0.06648 0.01363 4.88 0.000

S = 3.61854 R-Sq = 68.4% R-Sq(adj) = 65.5%

Analysis of Variance

Source DF SS MS F P

Regression 1 311.66 311.66 23.80 0.000

Residual Error 11 144.03 13.09

Total 12 455.69

Unusual Observations

Price

Obs ($) Score Fit SE Fit Residual St Resid

13 90 46.00 53.27 1.52 -7.27 -2.21R

R denotes an observation with a large standardized residual.

The estimated regression equation is significant and explains 68.4% of the variability in the overall score using the price of the camera. But, the curvilinear relationship we observed in the scatter diagram is still a concern. If we are willing to only consider cameras with a price of $200 or less, a linear relationship may be able to be used as an approximation. For instance, the following regression output show the results using only Canon cameras with a price of $200 or less.

The regression equation is

Score = 41.8 + 0.107 Price ($)

Predictor Coef SE Coef T P

Constant 41.809 3.210 13.02 0.000

Price ($) 0.10681 0.02070 5.16 0.001

S = 3.13708 R-Sq = 74.7% R-Sq(adj) = 71.9%

Analysis of Variance

Source DF SS MS F P

Regression 1 261.97 261.97 26.62 0.001

Residual Error 9 88.57 9.84

Total 10 350.55

The fit has improved slightly, but the issue whether the underlying relationship may be better described by curvilinear model cannot be resolved using the methods introduced in this chapter.

**Case Problem 4: Finding the Best Car Value**

1. Descriptive statistics follow.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Price ($)** | **Cost/Mile** | **Road-Test Score** | **Predicted Reliability** | **Value Score** |
| Mean | 26886.20 | 0.642 | 80.45 | 3.75 | 1.46 |
| Standard Error | 754.51 | 0.01 | 2.21 | 0.14 | 0.04 |
| Median | 28067.50 | 0.665 | 82 | 4 | 1.43 |
| Mode | #N/A | 0.67 | 81 | 4 | 1.73 |
| Standard Deviation | 3374.284152 | 0.06 | 9.90 | 0.64 | 0.20 |
| Sample Variance | 11385793.54 | 0.00 | 98.05 | 0.41 | 0.04 |
| Kurtosis | -1.41 | -1.58 | 2.58 | -0.44 | -0.64 |
| Skewness | -0.23 | -0.04 | -1.41 | 0.25 | -0.18 |
| Range | 10560 | 0.18 | 41 | 2 | 0.7 |
| Minimum | 21800 | 0.56 | 52 | 3 | 1.05 |
| Maximum | 32360 | 0.74 | 93 | 5 | 1.75 |
| Sum | 537724 | 12.84 | 1609 | 75 | 29.16 |
| Count | 20 | 20 | 20 | 20 | 20 |

2. A portion of the Minitab output follows.

The regression equation is

Value Score = 2.36 - 0.000033 Price ($)

Predictor Coef SE Coef T P

Constant 2.3587 0.3063 7.70 0.000

Price ($) -0.00003350 0.00001131 -2.96 0.008

S = 0.166326 R-Sq = 32.8% R-Sq(adj) = 29.0%

Analysis of Variance

Source DF SS MS F P

Regression 1 0.24276 0.24276 8.78 0.008

Residual Error 18 0.49796 0.02766

Total 19 0.74072

3. A portion of the Minitab output follows.

The regression equation is

Value Score = 2.94 - 2.31 Cost/Mile

Predictor Coef SE Coef T P

Constant 2.9422 0.3422 8.60 0.000

Cost/Mile -2.3119 0.5307 -4.36 0.000

S = 0.141540 R-Sq = 51.3% R-Sq(adj) = 48.6%

Analysis of Variance

Source DF SS MS F P

Regression 1 0.38012 0.38012 18.97 0.000

Residual Error 18 0.36060 0.02003

Total 19 0.74072

4. A portion of the Minitab output follows.

The regression equation is

Value Score = 0.798 + 0.00821 Road-Test Score

Predictor Coef SE Coef T P

Constant 0.7978 0.3471 2.30 0.034

Road-Test Score 0.008206 0.004283 1.92 0.071

S = 0.184882 R-Sq = 16.9% R-Sq(adj) = 12.3%

Analysis of Variance

Source DF SS MS F P

Regression 1 0.12546 0.12546 3.67 0.071

Residual Error 18 0.61526 0.03418

Total 19 0.74072

5. A portion of the Minitab output follows.

The regression equation is

Value Score = 1.05 + 0.108 Predicted Reliability

Predictor Coef SE Coef T P

Constant 1.0515 0.2594 4.05 0.001

Predicted Reliability 0.10839 0.06824 1.59 0.130

S = 0.189982 R-Sq = 12.3% R-Sq(adj) = 7.4%

Analysis of Variance

Source DF SS MS F P

Regression 1 0.09105 0.09105 2.52 0.130

Residual Error 18 0.64967 0.03609

Total 19 0.74072

6. Although Consumer Reports did not include price as one of the components of value score, the regression output shown in part (2) shows that there is a significant statistical relationship between price ($) and value score (*p*-value = .008). Reviewing the regression output in parts (3) – (5) indicates that cost/mile is the best single predictor of value score (R-Sq = 51.3%). To further investigate the relationship among these variables we really need to use multiple regression analysis.