Chapter 3 Project: Investigating Relationships

What is a better predictor of battery life in netbooks, weight or cost? What is a better predictor of the cost of a used car, age or number of miles? What is a better predictor of winning percentage, points scored or points allowed? What is a better predictor of success in AP Statistics, SAT score or GPA?

In this project you will investigate which of two possible explanatory variables is a better predictor of a response variable by doing a thorough analysis and comparison of the relationships between each pair of variables.

Your report/poster should include the following components:

1. Introduction: In this section you will introduce the context of your study, define the variables you will be investigating, and discuss any preliminary hypotheses you might have about the relationships between the variables.
2. Data Collection: In this section you will describe how you obtained your data. If it is from the Internet, make sure to cite the specific page. Include the data in a table and make sure you have at least 10 observations.
3. Graphs: Display the relationships in well-labeled scatterplots. Make sure to display the response variable on the same scale in each plot. Describe the relationships in each scatterplot and compare the relationships.
4. Numerical Summaries and Interpretations: Calculate and interpret the correlation, equation of the least-squares regression line, the standard deviation of the residuals s and r^2 for each relationship. Also, make and describe the residual plots for each relationship.
5. Conclusion and Discussion: Decide which of your explanatory variables does a better job of predicting the response variable, citing specific evidence from the graphs and numerical summaries. Discuss when it would be appropriate to make predictions using the least-squares regression line and any potential limitations of your model.

Chapter 3 Project Rubric

Note: If a project doesn't meet the minimum requirements for a 1 in a category, a score of 0 is possible.

Introduction and Data Collection

4 = Complete

- Describes the context of the research
- Clearly defines the variables and any preliminary hypotheses
- Specifically describes how the data were collected (including source, if appropriate)
- Includes appropriate amount of data and displays in a table

3 = Substantial

- Clearly introduces the context of the research and the variables being used
- Describes how the data were collected or includes the data in a table

2 = Developing

- Introduces the context of the research, but doesn't specifically define variables.
- Describes how data were collected, but doesn't include the data (or vice-versa)

1 = Minimal

- Briefly describes the context of the research or the method of data collection

Graphs

4 = Complete

- Scatterplots are correctly drawn, clearly labeled and easy to compare
- Important characteristics of the graphs are described and compared
- Residual plots are correctly displayed and interpreted

3 = Substantial

- Includes all three characteristics above, but makes one of the following errors
  - Scatterplots are correctly drawn, but some labels are missing
  - Scatterplots are compared, but the descriptions are weak or some comparisons are missing
  - Residual plot is included, but not interpreted correctly

2 = Developing

- Includes scatterplots with appropriate descriptions and comparisons, but no residual plots OR includes both scatterplots and residual plots with weak descriptions or no descriptions

1 = Minimal

- Only scatterplots are included with little or no descriptions or interpretations

Numerical Summaries

4 = Complete

- Includes all of the numerical summaries (r, slope, y intercept, s, r^2)
- All numerical summaries are interpreted correctly in context

3 = Substantial

- Includes all of the numerical summaries, but the interpretations are weak and/or lack context

2 = Developing

- Includes most or all of the numerical summaries but several interpretations are missing or incorrect or not written in context

1 = Minimal

- Some numerical summaries are included

Conclusions

4 = Complete

- Makes a reasonable conclusion about which explanatory variable is a better predictor
- Decision is based on specific evidence from the graphs and numerical summaries
- Discusses when making predictions is appropriate (i.e. discusses extrapolation)
- Shows evidence of critical reflection (discusses possible errors, shortcomings, limitations, etc.)

3 = Substantial

- Makes a reasonable conclusion citing evidence from graphs and numerical summaries
- Discusses when to make predictions or shows some other evidence of critical reflection

2 = Developing

- Makes a reasonable conclusion based on evidence from graphs and numerical summaries

1 = Minimal

- Makes a reasonable conclusion with little or no reference to specific evidence

Overall Presentation/Communication

4 = Complete

- Clear, holistic picture of the project
- Project is well organized, neat and easy to read
- Ideas are well communicated, including appropriate transitions between sections.

3 = Substantial

- Project is organized and easy to read, but lacks clear communication or a holistic picture of the project

2 = Developing

- Project is not well organized or communication is poor

1 = Minimal

- Communication and organization are very poor

Examples of Computer Output

Example: Body Weight and Pack Weight

Minitab Output:

```
Predictor       Coef  SE Coef     T      P
Constant      16.265    3.937  4.13  0.006
Body Weight  0.09080  0.02831  3.21  0.018


S = 2.26954   R-Sq = 63.2%   R-Sq(adj) = 57.0%
```

JMP Output:

Summary of Fit

RSquare	0.631536
RSquare Adj	0.570126
Root Mean Square Error	2.26954
Mean of Response	28.625
Observations (or Sum Wgts)	8

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	16.264927	3.93692	4.13	0.0061
Body Weight	0.0907994	0.028314	3.21	0.0184

Example: Age and Gesell Scores

Minitab Output:

```
Predictor        Coef  SE Coef      T      P
Constant      109.874    5.068  21.68  0.000
Age (months)  -1.1270   0.3102  -3.63  0.002


S = 11.0229   R-Sq = 41.0%   R-Sq(adj) = 37.9%
```

JMP Output:

Summary of Fit

RSquare	0.409971
RSquare Adj	0.378917
Root Mean Square Error	11.02291
Mean of Response	93.66667
Observations (or Sum Wgts)	21

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	109.87384	5.067802	21.68	<.0001
Age	-1.126989	0.310172	-3.63	0.0018

Alternate Example: Used Hondas

Minitab Output:

```
Predictor      Coef  SE Coef      T      P
Constant    18773.3    856.2  21.93  0.000
Miles        -86.18    15.95  -5.40  0.000


S = 971.647   R-Sq = 76.4%   R-Sq(adj) = 73.8%
```

JMP Output:

Summary of Fit

RSquare	0.764451
RSquare Adj	0.738279
Root Mean Square Error	971.6474
Mean of Response	14425
Observations (or Sum Wgts)	11

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	18773.284	856.2452	21.93	<.0001
Miles	-86.1822	15.94638	-5.40	0.0004

Alternate Example: Track and Field

Minitab Output:

```
Predictor      Coef  SE Coef      T      P
Constant     304.56    50.73   6.00  0.000
sprint      -27.629    7.413  -3.73  0.003


S = 31.9772   R-Sq = 55.8%   R-Sq(adj) = 51.8%
```

JMP Output:

Summary of Fit

RSquare	0.55809
RSquare Adj	0.517916
Root Mean Square Error	31.97723
Mean of Response	118.3846
Observations (or Sum Wgts)	13

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	304.55979	50.73181	6.00	<.0001
Sprint	-27.62874	7.412755	-3.73	0.0033

Alternate Example: Netbooks (using weight to predict battery life, outlier included)

Minitab Output:

```
Predictor      Coef  SE Coef      T      P
Constant   -16.046    4.622  -3.47  0.002
weight       7.944    1.698   4.68  0.000


S = 1.61655   R-Sq = 52.2%   R-Sq(adj) = 49.9%
```

JMP Output:

Summary of Fit

RSquare	0.522413
RSquare Adj	0.498534
Root Mean Square Error	1.616548
Mean of Response	5.511364
Observations (or Sum Wgts)	22

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-16.04591	4.621775	-3.47	0.0024
weight	7.9440542	1.698425	4.68	0.0001

Alternate Example: Netbooks (using price to predict battery life, outliers included)

Minitab Output:

```
Predictor       Coef   SE Coef      T      P
Constant       6.553     3.326   1.97  0.063
price      -0.002933  0.009258  -0.32  0.755


S = 2.33333   R-Sq = 0.5%   R-Sq(adj) = 0.0%
```

JMP Output:

Summary of Fit

RSquare	0.004993
RSquare Adj	-0.04476
Root Mean Square Error	2.333326
Mean of Response	5.511364
Observations (or Sum Wgts)	22

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	6.5531967	3.32603	1.97	0.0628
price	-0.002933	0.009258	-0.32	0.7547

Data Exploration: The SAT Essay

Minitab Output:

```
Predictor       Coef    SE Coef       T       P
Constant      1.1728     0.3193    3.67   0.001
Words       0.010370   0.001037   10.00   0.000


S = 0.792095   R-Sq = 77.5%   R-Sq(adj) = 76.8%
```

JMP Output:

Summary of Fit

RSquare	0.77528
RSquare Adj	0.767532
Root Mean Square Error	0.792095
Mean of Response	4.032258
Observations (or Sum Wgts)	31

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	1.1727949	0.319318	3.67	0.0010
Words	0.0103701	0.001037	10.00	<.0001

