Oregon Hill Wireless Survey
Regression Model and Statistical Evaluation

Sky Huvard

Business Statistics
Dr. George Canavos
4 May 2003

Overview

I am interested in starting a wireless broadband project in my neighborhood, Oregon Hill. I decided to survey (Appendix 1) the community to determine factors that influence their Internet access budget. I designed a 10 question online poll and distributed flyers. The first week of the poll, I got less than 10 responses. I then decided to take a proactive approach and set up a kiosk on Cary Street and asked everyone that walked by to take the survey. After 12 hours of polling over two days, I had 31 usable responses.

Prediction

I assumed factors in a person's Internet access budget include (1) if they are a home owner, (2) if they have a home office, (3) if they are a student, (4) if they are a gamer ["avid video game player"], (5) if they want a static IP [a fixed "Internet Protocol" address], (6) desired upload speed, and (7) desired download speed. I defined the following intial operational variables: Y = Internet access budget; X_1 = 1 if they own their home, 0 if not; X_2 = 1 if they have a home office; 0 if not; X_3 = 1 if they are a student, 0 if not; X_4 = 1 if they are a gamer, 0 if not; X_5 = 1 if they want a static IP, 0 if not; X_6 = desired upload speed in kbps ["kilobits per second"]; X_7 = desired download speed in kbps.

Sample Data

The sample data (Appendix 2) was obtained online via the survey. Responses were deemed representative of the community if their zip code was 23220. A total of 31 responses were recorded.

Initial Observation

I assumed the model:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \varepsilon$$

correctly represents the relationship between the response variable Y and the potential predictor variables X_1-X_6. I fit the model to the sample data using Minitab which output the least squares equation:

$$Y = 7.0 - 17.3X_1 + 19.8X_2 - 18.6X_3 + 33.0X_4 - 7.8X_5 + 0.0189X_6 + 0.0155X_7$$

The signs of the least squares coefficients for the predictor variables X_2, X_3, X_4, X_6, and X_7, are consistent with my expections. I expected positive relationships between Y and X_1, X_2, X_4, X_5, X_6, and X_7, and a negative relationship between Y and X_3. The signs of the least squares coefficient for the predictor variable X_1 is unexpected, but not unrealistic. The negative sign of the least squares coefficient for X_5 is not expected. This implies that people would pay less for a premium service such as a static IP, with all other variables being held constant.

Initial Evaluation

The analysis of variance for the least squares equation shows the *P*-value for the null hypothesis H_0: $\beta_1 = \beta_2 = \ldots = \beta_7 = 0$ is .028, which implies that the evidence against the null hypothesis *is* convincing. However, examining individual *P*-values show the only predictor variable that helps to explain the variation in the sample *Y*-values is X_4. Graphs of the residuals versus each of the predictor variables indicate a few outlier values, but they are not omitted because they are an acceptable value. The adjusted R^2 value for this least squares equation shows that only 30% of the total variation in the sample *Y*-values are attributed to the predictor variables. A stepwise regression shows that X_2, X_4, and X_7 are the only predictor variables that help explain the variation.

Revised Observation

I assumed the model:

$$Y = \beta_2 X_2 + \beta_4 X_4 + \beta_7 X_7 + \varepsilon$$

correctly represents the relationship between the response variable *Y* and the potential predictor variables X_2, X_4, and X_7. I fit the model to the sample data using Minitab which output the least squares equation:

$$Y = -7.9 + 29.0 X_2 + 30.7 X_4 + 0.0209 X_7$$

The signs of the least squares coefficients for the predictor variables X_2, X_3, X_4, X_6,

and X_7, are consistent with my expections. I expected positive relationships between Y and X_2, X_4, and X_7. The signs of the least squares coefficient for the constant unexpected. This implies that people would not pay for service with download speeds lower than 378 kbps, with all other variables being held constant.

Revised Evaluation

The null hypothesis H_0: $\beta_2 = \beta_4 = \beta_7 = 0$ is clearly contradicted (P-value ≈ .004). The presence of X_2, X_4, and X_7 are helpful in explaining the variation in the sample Y-values (P-values for T statistics of .026, .024, and .014, respectively). The residual variance was slightly reduced, from $S_e^2 = 29.4077$ to $S_e^2 = 29.1761$. The adjusted R^2 value for this least squares equation shows that 31% of the total variation in the sample Y-values are attributed to the predictor variables. Graphs of the residuals against each of the three predictor variables provide evidence that the error variance is not constant, and could be improved using the *weighted least squares* method. A graph of the residuals versus the response variable indicate a linear association with an unknown predictor.

Conclusion

The revised regression equation is the least deficient. An alternate regression equation is provided in the appendix.

Appendix

1. Oregon Hill Wireless Survey

2. Survey Results

3. Initial Regression Analysis

4. Revised Regression Analysis

5. Alternate Regression Analysis

Oregon Hill Wireless Survey

1. Are you a home owner?

2. Do you have a home office?

3. Are you a student?

4. Are you a gamer?

5. Would you like a static IP?

6. What is your current budget for Internet connectivity? (dollars per month)

7. What upload speed do you desire? (kbps)

8. What download speed do you desire? (kbps)

9. Would you be willing to host a small antenna to accommodate delivery of service to your location?

Survey Results

y	x1	x2	x3	x4	x5	x6	x7
50	0	0	0	1	1	512	512
10	0	1	0	0	1	512	1280
12	0	0	1	0	0	512	512
25	1	1	0	0	1	768	768
49	1	1	0	0	1	1024	2048
35	0	1	1	0	1	768	2048
25	1	0	1	1	0	256	512
25	0	0	1	1	1	1024	2048
20	1	0	0	1	0	128	256
100	0	1	0	0	0	1024	2048
20	1	0	0	0	1	1024	1024
20	0	0	1	1	0	1024	2048
20	1	1	0	0	0	128	128
40	0	1	0	0	1	1024	1024
40	0	1	0	0	1	1024	1024
40	0	1	0	0	1	1024	1024
80	0	1	0	0	1	1024	2048
50	0	0	0	0	0	128	2048
40	1	1	0	0	1	1024	2048
50	0	0	1	0	0	1024	1280
30	0	0	1	0	0	512	1024
60	0	1	0	0	1	1024	2048
20	0	0	1	0	1	256	512
45	1	0	1	1	1	768	1536
30	0	0	1	1	1	256	512
75	0	0	1	1	1	1024	2048
30	0	0	0	1	0	128	1024
40	1	0	0	0	1	1024	2048
40	1	0	0	0	1	1024	2048
40	1	1	0	0	1	1024	2048
200	0	1	0	1	1	1024	2048

Initial Regression Analysis

Minitab output:

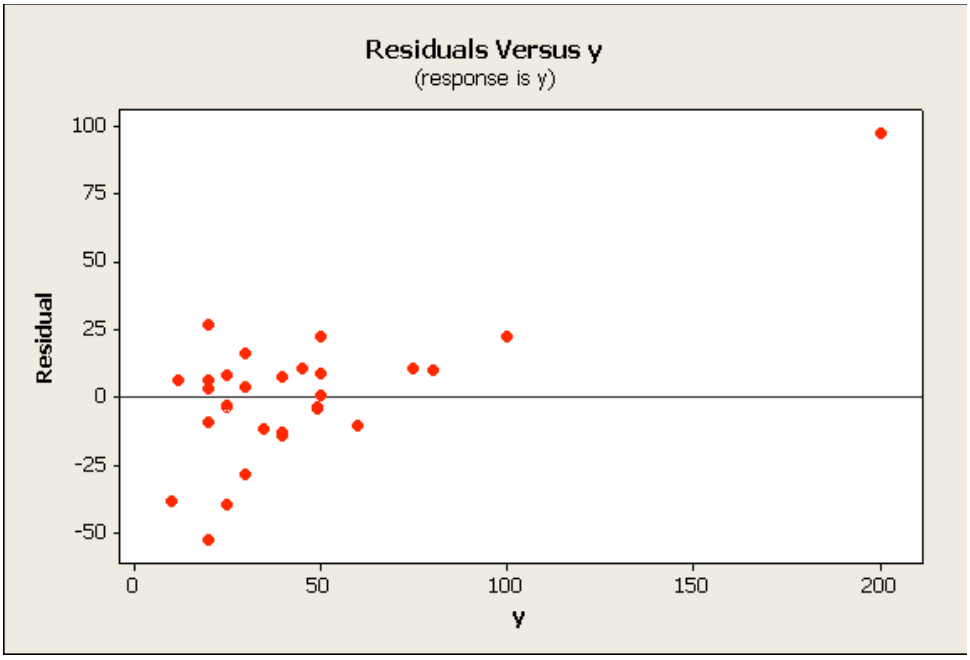
```
The regression equation is
y = 7.0 - 17.3 x1 + 19.8 x2 - 18.6 x3 + 33.0 x4 - 7.8 x5 + 0.0189 x6 + 0.0155 x7


Predictor      Coef  SE Coef      T      P
Constant       6.99    18.01   0.39  0.701
x1           -17.30    11.75  -1.47  0.154
x2            19.76    14.73   1.34  0.193
x3           -18.64    13.97  -1.33  0.195
x4            33.01    13.27   2.49  0.021
x5            -7.77    13.60  -0.57  0.573
x6          0.01885  0.02351   0.80  0.431
x7          0.01551  0.01067   1.45  0.160


S = 29.4077   R-Sq = 46.3%   R-Sq(adj) = 30.0%


Analysis of Variance

Source           DF       SS      MS     F      P
Regression        7  17151.8  2450.3  2.83  0.028
Residual Error   23  19890.7   864.8
Total            30  37042.6
```



Revised Regression Analysis

Minitab output:

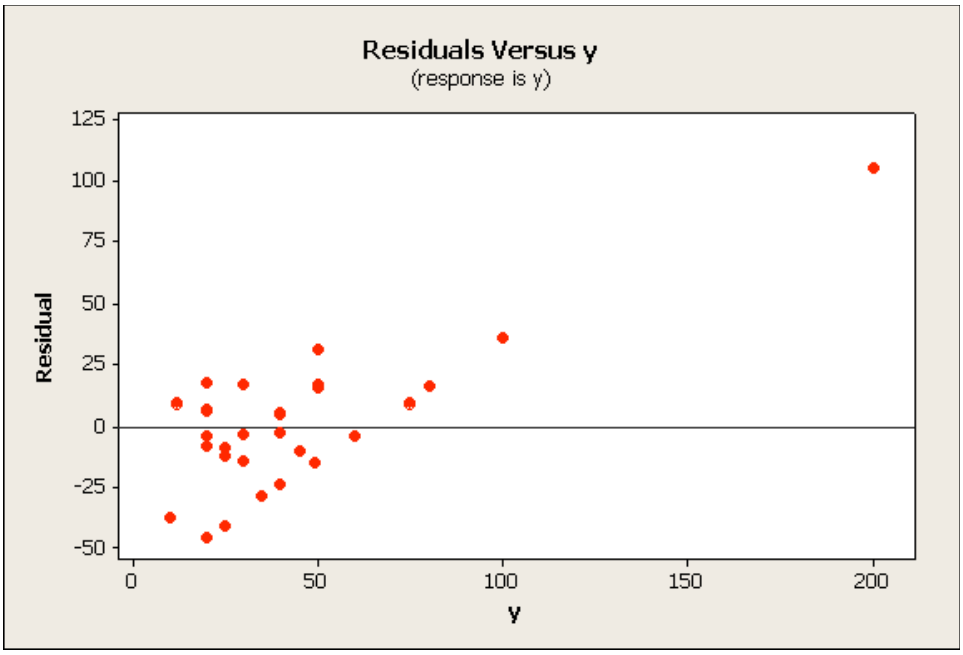
```
The regression equation is
y = - 7.9 + 29.0 x2 + 30.7 x4 + 0.0209 x7


Predictor       Coef    SE Coef       T       P
Constant       -7.86      14.00   -0.56   0.579
x2             28.98      12.30    2.36   0.026
x4             30.72      12.84    2.39   0.024
x7          0.020921   0.008006    2.61   0.014


S = 29.1761   R-Sq = 38.0%   R-Sq(adj) = 31.1%


Analysis of Variance

Source           DF       SS      MS     F      P
Regression        3  14058.9  4686.3  5.51  0.004
Residual Error   27  22983.7   851.3
Total            30  37042.6
```



Alternate Observation

My subject-matter knowledge told me predictor variables that should weigh heavily on budget are: bandwidth [upload + download speed], home office use, and preference for a static IP. I set up a pair of qualitative variables to classify these and sort out unreasonably low budgets. $X_8 = 1$ for advanced functionality, and are willing to pay for it, 0 if not; $X_9 = 1$ for mid-level functionality, 0 if not; both X_8 and $X_9 = 0$ for basic functionality. I then included these new predictor variables in another stepwise regression, which presented X_2, X_4, X_8, and X_9 for the new least squares equation.

The alternate model is:

$$Y = \beta_2 X_2 + \beta_4 X_4 + \beta_8 X_8 + \beta_9 X_9 + \varepsilon$$

and is assumed to correctly represent the relationship between the response variable Y and the potential predictor variables X_1-X_9. I fit the model to the sample data using Minitab which output the least squares equation:

$$Y = 11.9 + 11.2X_2 + 15.6X_4 + 47.8X_8 + 19.1X_9$$

The signs of the least squares coefficients for the predictor variables X_2, X_4, X_8, and X_9, are consistent with my expectations.

Alternate Evaluation

The null hypothesis H_0: $\beta_2 = \beta_4 = \beta_8 = \beta_9 = 0$ is clearly contradicted (P-value ≈ .000). The presence of X_2, X_4, X_8, and X_9 are helpful in explaining the variation in the sample Y-values (P-values for T statistics of .000, .000, .019, .003, respectively). The residual variance was reduced by almost half, from $S_e^2 = 17.365$ to $S_e^2 = 9.981$. The adjusted R^2 value for this least squares equation shows that 80% of the total variation in the sample Y-values are attributed to the predictor variables. Graphs of the residuals against each of the four predictor variables provide evidence that the error variance is not constant, and could be improved using the *weighted least squares* method. A graph of the residuals versus the response variable indicate a few reasonable outliers.

Alternate Regression Analysis

Minitab output:

```
The regression equation is
y = 11.9 + 47.8 x8 + 19.1 x9 + 11.2 x2 + 15.6 x4


Predictor    Coef  SE Coef     T      P
Constant   11.920    4.573  2.61  0.015
x8         47.816    5.099  9.38  0.000
x9         19.143    4.757  4.02  0.000
x2         11.195    4.448  2.52  0.019
x4         15.569    4.707  3.31  0.003


S = 9.98084   R-Sq = 82.5%   R-Sq(adj) = 79.5%
```

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	4	11238.4	2809.6	28.20	0.000
Residual Error	24	2390.8	99.6		
Total	28	13629.2