TRASHBALL: A LOGISTIC REGRESSION CLASSROOM ACTIVITY

Christopher H. Morrell and Richard E Auer
Mathematical Sciences Department, Loyola College in Maryland,
4501 North Charles Street, Baltimore, MD 21210-2699 chm@loyola.edu

Abstract

A classroom activity is described that may be used to motivate and illustrate the use of logistic regression for a binary response variable. The activity involves students attempting to toss a ball into a trashcan from various distances. The outcome is whether or not the student is successful in tossing the ball into the trashcan. The resulting data allows the instructor to discuss the binary nature of the response variable, the need for a logistic regression model, fitting and interpreting the model, multiple logistic regression, and variable selection in this context.

1. Introduction

In many statistical methods or linear models courses instructors initially concentrate on continuous numerical response variables. Recently it has become easier to also consider response variables that are either binary or categorical. This has been made possible as more introductory linear models and statistical methods books now include chapters or sections devoted to logistic regression (see Kleinbaum et al. (1998), Kutner et. al. (2003), Ott, and Longnecker (2001), Ryan (1996)).

In the next section, we describe a classroom activity that can be used to motivate the need for the logistic regression model and provides an entertaining way for students to become familiar with the model. This activity may also be used to discuss experimental design issues, if that topic fits into the course objectives/description.

The activity was conducted in subsequent offerings of Experimental Research Methods, a junior/senior level course taken by Mathematical Science majors and minors at Loyola College in Maryland. Section 3 considers the actual data collected from the activity in the fall of 2003.

Section 4 provides suggestions and conclusions based on our experience.

2. Trashball: The Activity

The activity involves students attempting to toss a ball into a trashcan. Consequently, the outcome or response variable is whether or not the ball ends up in the trashcan. This binary variable could be termed "ShotMade." Whether or not the student is successful in making the shot likely depends on the distance from the trashcan that the student tosses the ball. A number of additional explanatory variables may be included in the design of the experiment. For example, four factors may be included in the design of the experiment: the distance from the trashcan (from 5 to 12 feet), the orientation of the trashcan (a rectangular trashcan was used and the long side can be aimed at or the trashcan may be rotated through 90^o to provide a narrower target), the gender of the student, and the type of ball used (tennis ball or racquetball). It is not surprising that this trashcan activity was nicknamed "Trashball" by one of the students in the fall 2003 class.

This activity may be most beneficial to conduct just after finishing the topic of multiple linear regression on a continuous numerical dependent variable. As the set-up for Trashball is described to the class, the students realize that the response variable has only two outcomes. A discussion of the assumptions behind linear regression leads to the realization that linear regression is not appropriate for this data. In addition, it may be pointed out that linear regression could lead to predictions that are negative or greater then one. By now, the students will have discovered that they are actually trying to model the probability of a success and that the results must be values between 0 and 1. At this point, the logistic regression function should be introduced and its properties explained.

When the activity actually begins, the data may be entered into a computer using the Minitab software package (Ryan and Joiner, 2001). Optimally, the results are immediately displayed using a classroom projection system. The data set includes all the values of the explanatory variables for each shot taken. The settings of these explanatory variables make up the design of the experiment. Students may be told how the various factors should be balanced across the experiment so that interaction terms can be estimated. But such plans may likely prove to be difficult to achieve as was the case in our class; some of the students were absent on the day of the experiment. This merely necessitates some adjustments to the design.

To increase the sample size, consider having each student make three attempts from varying combinations of distance, orientation, and type of ball. The repeated observations may induce some non-independence and this should be discussed during the execution of the experiment. Tables 1(a)-(f) display cross tabulations that illustrate the resulting balance in the explanatory variables in the fall of 2003. Note that Tables 1(a) and 1(b) demonstrate that gender is well balanced across ball-type and orientation. Tables 1(c)-(f) similarly demonstrate a reasonable balance among the other explanatory variables.

Table 1(a). Cross tabulation of gender by type of ball.

	Gender		
	Male	Female	Total
Racquet Ball	9	12	21
Tennis Ball	9	12	21
Total	18	24	42

Table 1(b). Cross tabulation of gender by orientation of target.

	Gender		
	Male	Female	Total
Narrow Target	9	12	21
Wide Target	9	12	21
Total	18	24	

Table 1(c). Cross tabulation of orientation by type of ball.

	Orientation		
	Narrow Target	Wide Target	Total
Racquet Ball	10	11	21
Tennis Ball	11	10	21
Total	21	21	42

Table 1(d). Cross tabulation of type of ball by distance from target.

	Shot Distance (in feet)								
	5	6	7	8	9	10	11	12	Total
Racquet Ball	3	2	3	3	2	3	2	3	21
Tennis Ball	3	2	3	2	3	3	2	3	21
Total	6	4	6	5	5	6	4	6	42

Table 1(e). Cross tabulation of orientation by distance from target.

	Shot Distance (in feet)								
	5	6	7	8	9	10	11	12	Total
Narrow Target	5	0	5	0	5	2	4	0	21
Wide Target	1	4	1	5	0	4	0	6	21
Total	6	4	6	5	5	6	4	6	42

Table 1(f). Cross tabulation of gender by distance from target.

	Shot Distance (in feet)								
	5	6	7	8	9	10	11	12	Total
Male	2	2	3	2	2	3	2	2	18
Female	4	2	3	3	3	3	2	4	24
Total	6	4	6	5	5	6	4	6	42

3. Results of Conducting the Classroom Activity

After the data is finally entered into the computer, applying the logistic model in Minitab in class provides much drama for the students. They likely will be anticipating which explanatory variables, if any, prove to be statistically significant.

Figure 1 is a plot of ShotMade versus distance (with jitter added to the points) and the Lowess curve is overlaid on the plot to illustrate the trend in the data. It is clear that as the distance increases there are more misses and, consequently, one can expect the probability of making the shot to decline with distance.

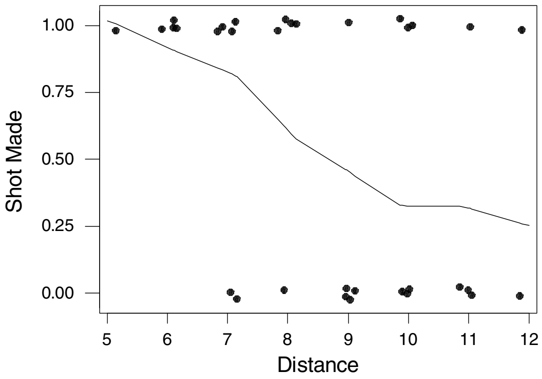

Figure 1. ShotMade versus distance between the thrower and the trashcan. Jitter is added to the points to show the repeated observations. The Lowess curve is overlaid.

Minitab Output 1. Logistic regression fit for ShotMade with distance between the thrower and the trashcan.

Binary Logistic Regression: ShotMade versus Distance

```
Link Function:  Logit
Response Information

Variable  Value       Count
ShotMade  1              25  (Event)
          0              17
          Total          42

Logistic Regression Table
                                                     Odds       95% CI
Predictor       Coef    SE Coef         Z      P    Ratio    Lower    Upper
Constant       5.204      1.695      3.07 0.002
Distance     -0.5499     0.1842     -2.98 0.003     0.58     0.40     0.83

Log-Likelihood = -22.294
Test that all slopes are zero: G = 12.102, DF = 1, P-Value = 0.001

Goodness-of-Fit Tests
Method                 Chi-Square    DF      P
Pearson                     5.542     6  0.476
Deviance                    6.488     6  0.371
Hosmer-Lemeshow             5.542     6  0.476

Table of Observed and Expected Frequencies:
(See Hosmer-Lemeshow Test for the Pearson Chi-Square Statistic)

                              Group
Value      1     2     3     4     5     6     7     8   Total
1
  Obs      2     1     3     1     4     4     4     6      25
  Exp      1.2   1.2   2.6   2.8   3.5   4.8   3.5   5.5
0
  Obs      4     3     3     4     1     2     0     0      17
  Exp      4.8   2.8   3.4   2.2   1.5   1.2   0.5   0.5

  Total    6     4     6     5     5     6     4     6      42
```

The goodness of fit tests indicate that this model provides an adequate description of this data. The predicted probability of making a shot as a function of distance x is given by:
P(ShotMade from distance x) = exp(5.204-0.5499*x) / (1+exp(5.204-0.5499*x)).

Figure 2 illustrates the fitted linear and logistic regression models. The fitted linear model clearly shows that predictions can fall outside the allowable range for probabilities. But the two models do agree quite well in the 0.2 to 0.8 range of probabilities. The odds ratio of 0.58 ($e^{-0.54999}$) indicates that the odds of making a shot changes by a factor of 0.58 for each additional foot from the trash can.

It is interesting to note that fitting a linear regression model to the data provides a prediction of greater than 1 when the distance is 4 feet (Figure 2) and the residuals show a decidedly non-random pattern (not shown). Hence, we now describe the fit of the appropriate logistic model (see Minitab Output 1).

Once simple logistic regression has been discussed and the various ideas and concepts of logistic regression have been introduced, the class can move onto multiple logistic regression by incorporating the additional explanatory variables measured during the experiment. The fitted logistic model containing all explanatory variables (but no interactions) is given in Minitab Output 2.

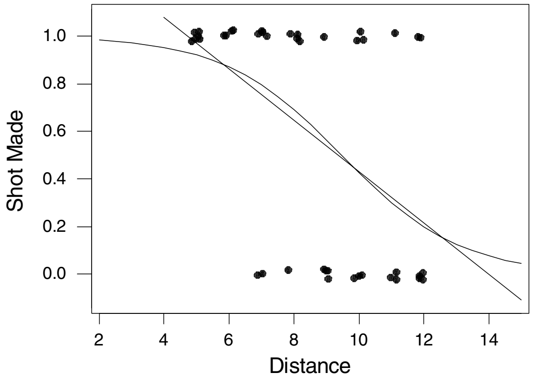

Figure 2. The fitted linear and logistic regression models. Jitter is included in the observed data points.

Minitab Output 2. Multiple Logistic regression fit for ShotMade with distance, type of ball, gender, and orientation of trash can.

Binary Logistic Regression: ShotMade versus Distance, Ball, Orientation, Gender

```
Logistic Regression Table
                                                        Odds        95% CI
Predictor          Coef    SE Coef        Z      P     Ratio    Lower    Upper
Constant          5.782      1.990     2.91  0.004
Distance        -0.7649     0.2349    -3.26  0.001      0.47     0.29     0.74
Ball             0.6156     0.8437     0.73  0.466      1.85     0.35     9.67
Orientation       2.420      1.015     2.38  0.017     11.24     1.54    82.16
Gender          -0.1532     0.8330    -0.18  0.854      0.86     0.17     4.39

Log-Likelihood = -18.394
Test that all slopes are zero: G = 19.904, DF = 4, P-Value = 0.001
```

Minitab Output 3. Parameter estimates of the final multiple logistic regression fit after backward elimination.

Binary Logistic Regression: ShotMade versus Distance, Orientation

```
Logistic Regression Table
                                                        Odds        95% CI
Predictor          Coef    SE Coef        Z      P     Ratio    Lower    Upper
Constant          5.857      1.913     3.06  0.002
Distance        -0.7425     0.2282    -3.25  0.001      0.48     0.30     0.74
Orientation      2.3096     0.9827     2.35  0.019     10.07     1.47    69.11

Log-Likelihood = -18.684
Test that all slopes are zero: G = 19.323, DF = 2, P-Value = 0.000

Goodness-of-Fit Tests

Method               Chi-Square    DF       P
Pearson                   3.441     8   0.904
Deviance                  3.994     8   0.858
Hosmer-Lemeshow           3.316     7   0.854
```

```
Table of Observed and Expected Frequencies:
(See Hosmer-Lemeshow Test for the Pearson Chi-Square Statistic)

                                  Group
Value       1     2     3     4     5     6     7     8     9   Total
1
  Obs       1     1     2     3     3     5     4     5     1      25
  Exp       0.4   1.9   1.9   3.3   2.7   4.5   4.5   4.9   1.0
0
  Obs       3     6     4     2     1     0     1     0     0      17
  Exp       3.6   5.1   4.1   1.7   1.3   0.5   0.5   0.1   0.0

  Total     4     7     6     5     4     5     5     5     1      42
```

Table 2. Observed proportion and modeled probabilities by orientation and distance.

Observed Proportion	Shot Distance (in feet)							
Logistic Probability	5	6	7	8	9	10	11	12
Wide Target	1.000 0.895	-	0.600 0.659	-	0.200 0.305	0.000 0.173	0.250 0.090	-
Narrow Target	1.000 0.989	1.000 0.976	1.000 0.951	0.800 0.903	-	0.750 0.677	-	0.333 0.322

Since gender is the least significant variable, it is dropped from the model. Type of ball also remains statistically non-significant leading to the final model (see Minitab Output 3).

The goodness of fit tests again indicate that this model provides an excellent description of this data. The estimated parameters indicate that the probability of making the shot decreases with distance and that one has a higher probability of making the shot if the orientation of the trashcan has the longer target facing the thrower.

The odds ratio for orientation tells us that the odds are 10 times higher to be successful in throwing the ball into the can if one is throwing at the long target versus the narrow target. In addition, the odds ratio of 0.48 states that the odds are reduced by roughly a factor of two for each foot the student moves away from the trash can.

Table 2 compares the observed proportion of shots made with the logistic probabilities for orientation and distance. The modeled probabilities generally conform to the observed proportions in the cells containing data.

4. Conclusions

Chapters or sections on logistic regression are appearing more frequently in texts on statistical methods/linear models. Trashball may be used to motivate the use of logistic regression to model a binary response variable. Our students enjoyed this activity. When conducting a mid-semester evaluation of the course, one student responded “More Trashball.”

Some suggested future modifications to this reported experiment are:

1. Use balls that are more different (for example, tennis and table tennis, or a larger ball that would require more precision to make it stay in the trashcan).
2. Hand student uses to toss the ball (writing hand, other hand – would not want to use left/right).

References

Kleinbaum, D.G., Lawrence L. Kupper, L.L., Muller, K.E.. Azhar Nizam, A. (1998), *Applied Regression Analysis and Multivariable Methods*, 3rd Edition, Duxbury Press.

Kutner, M.H., Nachtsheim, C.J., Neter, J., and Wasserman, W . (2003) *Applied Linear Regression Models*, McGraw-Hill/Irwin.

Ott, R.L. and Longnecker. M.T. (2001) *An Introduction to Statistical Methods and Data Analysi*s, 5th Edition, Duxbury Press.

Ryan, B.F. and Joiner, B.L. (2001) *Minitab Handbook*, 4th Edition, Duxbury Press.

Ryan, T.P. (1996) *Modern Regression Methods,* John Wiley and Sons.