Lecture 7

January 22, 2019 11:11 AM

STATISTICS 475: Applied Discrete Data Analysis

Logistic Regression Models for Binary Responses

(B&L Section 2.1–2.2.1)

1 Problem to be solved

- So far we have learned about the Bernoulli distribution for binary RVs and the binomial distribution for multiple independent Bernoullis.
- We have learned inference for one probability and for comparisons of two probabilities
  - This involved special techniques that are similar to, but also different from, those used for means of normal distributions
- Very often, we observe binary or binomial responses *along with one or more explanatory variables*
  - Now we need to generalize the notion of regression to apply to binary/binomial data

2 Example for this chapter

I will start with the example that we will use in this chapter:

Example: Placekicking (Adapted from the book) (Lecture 7 scripts.R, Placekick.csv)

In American and Canadian football, points can be scores by a "placekicker" kicking a ball through a target area at an end of the field. A success occurs when the football is kicked over the crossbar and between the two uprights of the goal posts. The placekicker's team receives either 1 or 3 points for a successful kick, where a "point after touchdown" (PAT) receives 1 point and a "field goal" receives 3 points. A placekick that is not successful receives 0 points. See the videos of "famous" placekicks on our Canvas site.

1

My coauthor, Chris Bilder, collected data on every placekick attempted in the 1995 National Football League (NFL—the main professional football league in the US) Season for a research project on the factors that affect the chance of success of a placekick. We examined a number of explanatory variables, including:

- `week`: Week of the season
- `distance`: Distance of the placekick from the goalposts, in yards (1 yard = 0.91 meters)
- `change`: Binary (or dummy or indicator) variable denoting lead-change (1) vs. non-lead-change (0) placekicks; lead-changing placekicks are those that have the potential to change which team is winning the game (for example, if a field goal is attempted by a team that is losing by 3 points or less, they will no longer be losing if the kick is successful)
- `elap30`: Number of minutes remaining before the end of the half, with overtime placekicks receiving a value of 0
- `PAT`: Binary variable denoting the type of placekick, where a PAT attempt is a 1 and a field goal attempt is a 0. Until this year, PATs were always attempted from 20 yards, except in rare circumstances.
- `type`: Binary variable denoting outdoor (1) vs. dome (0) placekicks
- `field`: Binary variable denoting grass (1) vs. artificial turf (0) placekicks
- `wind`: Binary variable for placekicks attempted in windy conditions (1) vs. non-windy conditions (0); we define windy as a wind stronger than 15 miles per hour at kickoff in an outdoor stadium

The response variable is `good`: a binary variable denoting successful (1) vs. failed (0) placekicks. There are 1,425 placekick observations (individual placekick attempts) in the data set. Below is how the data are read into R:

```
> placekick <- read.csv(file="C:\\Users\\tloughin\\Dropbox\\
475 Applied Discrete Data Analysis\\R\\Placekick.csv", header=TRUE)
> head(placekick)
  week distance change  elap30 PAT type field wind good
1    1       21      1 24.7167   0    1     1    0    1
2    1       21      0 15.8500   0    1     1    0    1
3    1       20      0  0.4500   1    1     1    0    1
4    1       28      0 13.5500   0    1     1    0    1
5    1       20      0 21.8667   1    0     0    0    1
6    1       25      0 17.6833   0    0     0    0    1
```

Our goal will be to develop a model that explains how the probability of successful placekick—$P(\texttt{good} = 1)$—relates to the explanatory variables that were collected. In particular, it is well known that `distance` should affect the probability of success. How much of a role do the other variables play?

---

3 Review of Linear Regression

- Linear regression model:

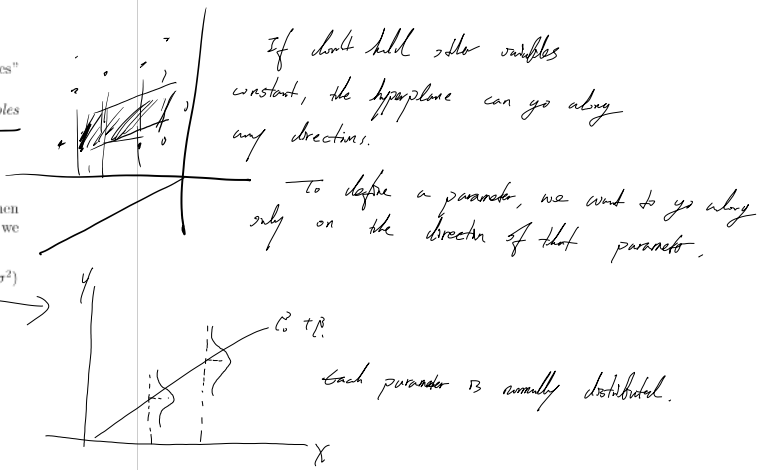
$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i$$

for $i = 1, \ldots, n$

  - Y_i is the RV representing the response for observation i
  - $x_{i1}, \ldots, x_{ip}$ are observed values of the corresponding explanatory variables $x_1, \ldots, x_p$
    - Get used to p as the number of explanatory variables
    - We will index explanatory variables by j: x_j, $j = 1, \ldots, p$
  - $\beta_0, \ldots, \beta_p$ are regression parameters
    - β_0 is the intercept (mean value of Y when all $x_j = 0$)
    - β_j, $j = 1, \ldots, p$ are "(partial) regression coefficients" (not really "slopes" anymore)
      - Change in mean Y for 1-unit change in x_j *holding all other variables constant.*
  - ϵ_i, $i = 1, \ldots, n$ are independent errors
    - Assume $\epsi_i \sim N(0, \sigma^2)$
  - We may write the models with or without the "i" subscripts. It is expected when we are referring to a model for a specific to a set of data, but is not needed if we are referring more generally to a model structure

- As a consequence of this model, we can write that $Y \sim N(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p, \sigma^2)$
  - In other words, we have a distributional MODEL for Y
    - See Appendix B.1 for a discussion of what it means to be a model
  - Mean value of Y is $E(Y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ Expected Y
  - Equal variance at all combinations of x_j's

- Goal is to estimate parameters of the model
  - Use least squares



    - *Requires equal variances*
  - Use the estimates for inference (e.g., $H_0 : \beta_j = 0$)
    - *Requires normality of Y*
  - Construct additional quantities (e.g., $\hat{Y}$, predicted value or mean for a particular combination of x_j's).

3.1 Linear regression for binary responses?

- Suppose Y_i are independent Bernoulli(π_i) RVs, where $E(Y_i) = \pi_i$ might depend on $x_{i1}, \ldots, x_{ip}$
- What if we do linear regression here?
  - Y_i does not have a normal distribution
  - $\text{Var}(Y_i) = \pi_i(1 - \pi_i)$ leads to potentially different variances
  - $\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ is not constrained to be within 0 and 1

Assumptions are too badly violated.

- As a result, we often have a very poor fitting line, and inferences based on the line are unreliable.
  - Practically never do this
- Instead we develop a regression model for the mean π that uses the correct distributional model
  - Use the fact that $Y_i \sim$Bernoulli(π)
  - Use a model for π that respects the 0–1 boundaries

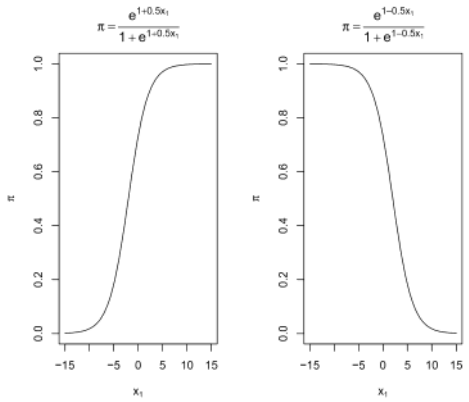
4 Logistic Regression Model

- The fundamental assumption of the logistic regression model is that

$$\pi = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)} \tag{1}$$

  - As $\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ goes toward $-\infty$, $\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)$ approaches 0, so π approaches 0
  - As $\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ goes toward ∞, $\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)$ approaches ∞, so π approaches 1
  - Thus, this form respects the boundaries of a probability
  - Also, the variance is $\pi(1 - \pi)$ from Binomial model

Figure 1: Shape of the logistic curve for 1-variable model with $\beta_1 = +0.5$ (left) and $\beta_1 = -0.5$ (right)



  - If we untangle this relationship, we can show that the log of the odds has a linear model:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \quad (2)$$

    * The log of the odds is also called the LOGIT of π:

$$\text{logit}(\pi) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \quad (3)$$

  - Formulas (1), (2), and (3) are equivalent representations of the logistic regression model and we will use them interchangeably.

- See Figure 1 for a picture of what the relationship looks like between π and any particular x_j (**PiPlot.R** from the book's website)
  - $0 < \pi < 1$
  - When $\beta_1 > 0$, there is a positive relationship between x_1 and π. When $\beta_1 < 0$, there is a negative relationship between x_1 and π.
  - The shape of the curve is somewhat similar to the letter s (this shape is called SIGMOIDAL).

  - The slope of the curve is dependent on the value of x_1.
    - We can show this mathematically by taking the derivative of π with respect to x_1: $\partial\pi/\partial x_1 = \beta_1\pi(1-\pi)$.
- Above $\pi = 0.5$ is a reversed mirror image of below $\pi = 0.5$.
  - If you rotate the graph 180 degrees, the picture is the same
- Resembles a linear relation between about $0.3 < \pi < 0.7$, but otherwise not very linear

4.1 Parameter estimation

- Gather data y_i, $i = 1, \ldots, n$ with corresponding explanatory variables
- The parameters we need to estimate are the regression coefficients $\beta_1, \ldots, \beta_p$
- Maximum likelihood estimation:

$$L(\boldsymbol{\beta}|\mathbf{y}) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

where $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)'$, $\mathbf{y} = (y_1, \ldots, y_n)'$, and

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})}$$

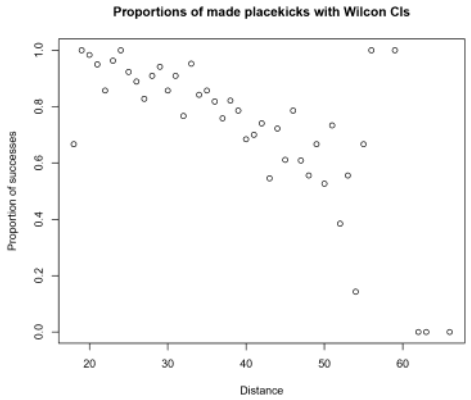
  - Plugging this definition for π_i into the likelihood and doing some algebra gives

$$\log\left[L(\boldsymbol{\beta}|\mathbf{y})\right] = \sum_{i=1}^{n} y_i\left(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}\right) - \log\left(1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})\right),$$

    which is reasonably easy to work with mathematically and computationally (see p. 65 in the book)

- Iterative numerical methods are used to find the $\boldsymbol{\beta}$ that maximize the likelihood function
- Regression parameter estimates (MLEs): $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)'$
- Estimated variance-covariance matrix for regression parameter estimates is found from the curvature (second derivative) of the log likelihood
  - Doesn't lead to a *simple* formula, but is not hard to compute.
  - Results in a *matrix* of values, corresponding to the estimated variances of the parameter estimates, and their estimated covariances.

Figure 2: Plot of proportions of successful placekicks against distance. Note: Not all distances have the same number of trials (over half of the kicks occur at 20 yards!).



Example: Placekicking (Lecture 7 scripts.R, Placekick.csv)

We will start by looking at a simple model of the probability of a success ($\pi = P(\texttt{good} = 1)$) against distance:

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1.$$

So $p = 1$ here, and x_1 is `distance`. See Figure 2 for a plot of the proportions of success at each distance. Note the distinct trend toward decreasing proportion of successes as distance increases. (*Therefore will expect β_1 has what sign?*)

- Logistic regression is fit using `glm()` in R
  - Stands for "generalized linear models", which is a group of models that includes logistic regression
    - This function fits many types of models
    - *Get used to using it!*
  - Model for logit is entered using `formula=<binary y> ~ <x1>+<x2>+...+<xp>`

    * "<>" is my code for something that gets replaced by a name you have chosen
  - Logistic regression is specified with `family=binomial(link = logit)`
  - `summary(<glm object>)` gives back some of the essential information

- Here are the results for the placekick data (not all numbers have been explained yet...)

```
> mod.fit <- glm(formula=good ~ distance,
                 family=binomial(link = logit), data=placekick)
> summary(mod.fit)  # summary(object = mod.fit) more completely

Call: glm(formula = good ~ distance, family = binomial(link = logit),
data = placekick)

Deviance Residuals:
     Min       1Q   Median       3Q      Max
 -2.7441   0.2425   0.2425   0.3801   1.6092

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.812080   0.326277   17.81   <2e-16 ***
distance    -0.115027   0.008339  -13.79   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1013.43  on 1424  degrees of freedom
Residual deviance:  775.75  on 1423  degrees of freedom
AIC: 779.75

Number of Fisher Scoring iterations: 6
```

- The estimated logistic regression model is

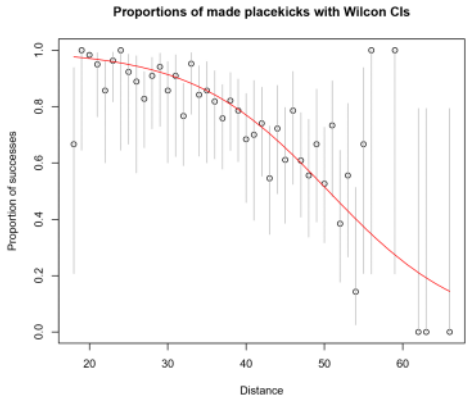
$$\text{logit}(\hat{\pi}) = 5.8 - 0.115\texttt{distance}$$

- The standard errors are

$$\sqrt{\widehat{Var}(\hat{\beta}_0)} = 0.33,\ \sqrt{\widehat{Var}(\hat{\beta}_1)} = 0.0083$$

- There are times when you will need to use the parameter estimates and/or standard errors in further computations.
  - Parameter estimates can be accessed directly using `coef(<glm object>)`

Figure 3: Plot of observed proportions of successes at each distance, with corresponding Wilson score confidence intervals (gray) and fitted logistic regression curve (red)



Not all CI. go across the curve:

Because we expect (only) $(1-\alpha)\%$ of all similarly constructed intervals to contain the true parameter.

i.e. $(1-\alpha)\%$ coverage probability

  - Estimated variances and covariances (like correlations between estimates) can be accessed directly using `vcov(<glm object>)`
  - See program for examples

- The computation took 6 iterations to achieve convergence
  - This number isn't important, but in rare cases where it does *not* achieve convergence, you will not want to report these results.
    * If it fails to converge, will see number of iterations = 25, the default maximum
    * *May* also see a warning message that it did not converge.

- The model fit is depicted in Figure 3

---

5 What to learn from this

1. Logistic regression is a *hugely* popular, go-to statistical method for binary/binomial responses. Learn to love it.
2. Get used to all of the notation, terminology, and formulas here, especially equations (1), (2), (3). We will keep using them.
3. Understand the shape and properties of the model.
4. You will have a chance to practice R code!

6 Exercises (due when announced)

Compete the following exercises from B&L, Chapter 2:

1. Exercise 3
2. Exercise 4(a)
3. Exercise 5(a).
   (a) Report the model in equation form similar to

$$\text{logit}(\hat{\pi}) = 5.8121 - 0.1150\texttt{distance}$$

   (b) Plot the resulting points, confidence intervals, and curve, similar to Figure 3.
   (c) Program the formula to compute probabilities of o-ring failure at various temperatures. Compute the probabilities at 81, 61, and 31 degrees Fahrenheit.
   (d) Note that 31 degrees is an extrapolation. Comment both on the numerical estimate of the probability of o-ring failure (would *you* want to ride in the shuttle if this is true?), and on your faith in the reliability of that estimate.

Here are two more exercises for practice, because I'd hate for you not to have something to do involving placekicking...

1. Exercise 7
2. Exercise 8(a,b). Report the equation for the logit in equation form as in the previous problem.

Notice that I want you to become *very* comfortable with estimating logistic regression models. These models are among the most important models in this course, and in all of statistics.