Chapter 6 Part 1

05/06/2021

Chapter Packages

```r
# Packages required for Chapter 6
library(gridExtra)
library(mnormt)
library(lme4)
library(knitr)
library(pander)
library(tidyverse)
library(kableExtra)
```

Logistic Regression

Learning Objectives

- Identify a binomial random variable and assess the validity of the binomial assumptions.
- Write a generalized linear model for binomial responses in two forms, one as a function of the logit and one as a function of p.
- Explain how fitting a logistic regression differs from fitting a linear least squares regression (LLSR) model.
- Interpret estimated coefficients in logistic regression.
- Differentiate between logistic regression models with binary and binomial responses.
- Use the residual deviance to compare models, to test for lack-of-fit when appropriate, and to check for unusual observations or needed transformations.

Re-Introduction to Logistic Regression

- Logistic regression is characterized by research questions with binary (yes/no or success/failure) or binomial (number of yesses or successes in *n* trials) responses:

a. Are students with poor grades more likely to binge drink?

b. Is exposure to a particular chemical associated with a cancer diagnosis?

c. Are the number of votes for a congressional candidate associated with the amount of campaign contributions?

Re-Introduction to Logistic Regression

- **Binary Responses:** Recall that binary responses take on only two values: success (Y=1) or failure (Y=0), Yes (Y=1) or No (Y=0), etc.
- Binary responses are ubiquitous; they are one of the most common types of data that statisticians encounter.
- We are often interested in modeling the probability of success p based on a set of covariates, although sometimes we wish to use those covariates to classify a future observation as a success or a failure.
- Examples (a) and (b) above would be considered to have binary responses (Does a student binge drink? Was a patient diagnosed with cancer?), assuming that we have a unique set of covariates for each individual student or patient.

Re-Introduction to Logistic Regression

- **Binomial Responses:** Recall binomial responses are the number of successes in n identical, independent trials with constant probability p of success.
- A sequence of independent trials like this with the same probability of success is called a **Bernoulli process**.
- As with binary responses, our objective in modeling binomial responses is to quantify how the probability of success, p, is associated with relevant covariates.

Example (c) above would be considered to have a binomial response, assuming we have vote totals at the congressional district level rather than information on individual voters.

Logistic Regression Assumptions

- Much like ordinary least squares (OLS), using **logistic regression** to make inferences requires model assumptions.

1. **Binary Response** The response variable is dichotomous (two possible responses) or the sum of dichotomous responses.
2. **Independence** The observations must be independent of one another.
3. **Variance Structure** By definition, the variance of a binomial random variable is $np(1-p)$, so that variability is highest when $p = .5$.
4. **Linearity** The log of the odds ratio, $\log(\frac{p}{1-p})$, must be a linear function of x. This will be explained further in the context of the first case study.

A Graphical Look at Logistic Regression

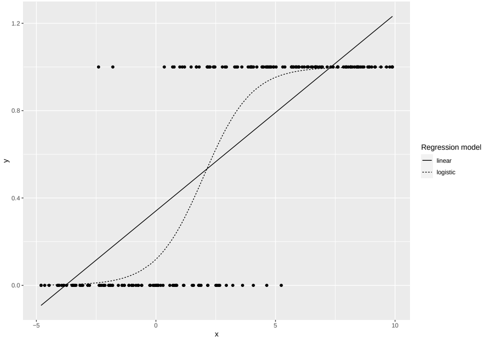

Figure 1: Linear vs. logistic regression models for binary response data.

Logistic Curve

- This graph illustrates a data set with a binary (0 or 1) response (Y) and a single continuous predictor (X).
- The solid line is a linear regression fit with least squares to model the probability of a success (Y=1) for a given value of X.
- With a binary response, the line doesn't fit the data well, and it produces predicted probabilities below 0 and above 1.
- On the other hand, the logistic regression fit (dashed curve) with its typical "S" shape follows the data closely and always produces predicted probabilities between 0 and 1.

Logistic Regression

For these and several other reasons detailed in this chapter, we will focus on the following model for logistic regression with binary or binomial responses:

$$log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 x_i$$

where the observed values $Y_i \sim$ binomial with $p = p_i$ for a given x_i and $n = 1$ for binary responses.

Case Studies Overview

- The first two involve binomial responses (Soccer Goalkeepers and Reconstructing Alabama), while the last case uses a binary response (Trying to Lose Weight).
- Even though binary responses are much more common, their models have a very similar form to binomial responses, so the first two case studies will illustrate important principles that also apply to the binary case.

Case Study Soccer

- The soccer goalkeeper data can be written in the form of a 2 $\times$ 2 table.
- This example is used to describe some of the underlying theory for logistic regression.
- Demonstrate is how binomial probability mass functions (pmfs) can be written in one-parameter exponential family form, from which we can identify the canonical link.
- Using the canonical link, we write a Generalized Linear Model for binomial counts and determine corresponding MLEs for model coefficients.
- Interpretation of the estimated parameters involves a fundamental concept, the odds ratio.

Case Study Reconstructing Alabama

- The Reconstructing Alabama case study is another binomial example which introduces the notion of deviances, which are used to compare and assess models.
- Thus, we will investigate hypothesis tests and confidence intervals, including issues of interaction terms, overdispersion, and lack-of-fit.
- We will also check the assumptions of logistic regression using empirical logit plots and deviance residuals.

Case Studies Soccer

- he last case study addresses why teens try to lose weight.
- Here the response is a binary variable which allows us to analyze individual level data.
- The analysis builds on concepts from the previous sections in the context of a random sample from CDC's Youth Risk Behavior Survey (YRBS).

Case Study: Soccer Goalkeepers

- Does the probability of a save in a soccer match depend upon whether the goalkeeper's team is behind or not?
- @Roskes2011 looked at penalty kicks in the men's World Cup soccer championships from 1982 to 2010, and they assembled data on 204 penalty kicks during shootouts. The data for this study is summarized in Table on the next slide.
- We need to type this into R

Modeling Odds

- Odds are one way to quantify a goalkeeper's performance.
- Here the odds that a goalkeeper makes a save when his team is behind is 2 to 22 or 0.09 to 1.
- Or equivalently, the odds that a goal is scored on a penalty kick is 22 to 2 or 11 to 1. An odds of 11 to 1 tells you that a shooter whose team is ahead will score 11 times for every 1 shot that the goalkeeper saves.

Modeling Odds

- When the goalkeeper's team is not behind the odds a goal is scored is 141 to 39 or 3.61 to 1.
- We see that the odds of a goal scored on a penalty kick are better when the goalkeeper's team is behind than when it is not behind (i.e., better odds of scoring for the shooter when the shooter's team is ahead).
- We can compare these odds by calculating the **odds ratio** (OR), 11/3.61 or 3.05, which tells us that the *odds* of a successful penalty kick are 3.05 times higher when the shooter's team is leading.

Modeling Odds

- In our example, it is also possible to estimate the probability of a goal, p, for either circumstance.
- When the goalkeeper's team is behind, the probability of a successful penalty kick is $p = 22/24$ or 0.833.
- We can see that the ratio of the probability of a goal scored divided by the probability of no goal is $(22/24)/(2/24) = 22/2$ or 11, the odds we had calculated above.
- The same calculation can be made when the goalkeeper's team is not behind. In general, we now have several ways of finding the odds of success under certain circumstances:

$$\text{Odds} = \frac{\#\text{successes}}{\#\text{failures}} = \frac{\#\text{successes}/n}{\#\text{failures}/n} = \frac{p}{1-p}.$$

Logistic Regression Models for Binomial Responses

- We would like to model the odds of success; however, odds are strictly positive.
- Similar to modeling $\log(\lambda)$ in Poisson regression, which allowed the response to take on values from $-\infty$ to ∞, we will model the log(odds), the **logit**, in logistic regression.
- Logits will be suitable for modeling with a linear function of the predictors:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$$

Logistic Regression Models for Binomial Responses

- Models of this form are referred to as **binomial regression models**, or more generally as **logistic regression models**.
- Here we provide intuition for using and interpreting logistic regression models, and then in the short optional section that follows, we present rationale for these models using GLM theory.

Soccer

- In our example we could define $X = 0$ for not behind and $X = 1$ for behind and fit the model:

$$\log\left(\frac{p_X}{1 - p_X}\right) = \beta_0 + \beta_1 X \tag{1}$$

where p_X is the probability of a successful penalty kick given X.

- Based on this model, the log odds of a successful penalty kick when the goalkeeper's team is not behind is:

$$\log\left(\frac{p_0}{1 - p_0}\right) = \beta_0,$$

and the log odds when the team is behind is:

$$\log\left(\frac{p_1}{1 - p_1}\right) = \beta_0 + \beta_1.$$

Soccer

- We can see that β_1 is the difference between the log odds of a successful penalty kick between games when the goalkeeper's team is behind and games when the team is not behind.
  Using rules of logs:

$$\beta_1 = (\beta_0+\beta_1)-\beta_0 = \log\left(\frac{p_1}{1-p_1}\right)-\log\left(\frac{p_0}{1-p_0}\right) = \log\left(\frac{p_1/(1-p_1)}{p_0/(1-p_0)}\right).$$

Soccer

- e^{β_1} is the ratio of the odds of scoring when the goalkeeper's team is not behind compared to scoring when the team is behind.
- In general, *exponentiated coefficients in logistic regression are odds ratios (OR)*.
- A general interpretation of an OR is the odds of success for group A compared to the odds of success for group B—how many times greater the odds of success are in group A compared to group B.

Soccer

- The logit model (Equation @ref(eq:logitXform)) can also be re-written in a **probability form**:

$$p_X = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

which can be re-written for games when the goalkeeper's team is behind as:

$$p_1 = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}} \tag{2}$$

and for games when the goalkeeper's team is not behind as:

$$p_0 = \frac{e^{\beta_0}}{1+e^{\beta_0}} \tag{3}$$

Likelihood

- We use likelihood methods to estimate β_0 and β_1.
- We can write the likelihood for this example in the following form:

$$\mathrm{Lik}(p_1, p_0) = \binom{24}{22} p_1^{22}(1-p_1)^2 \binom{180}{141} p_0^{141}(1-p_0)^{39}$$

- Our interest centers on estimating $\hat{\beta}_0$ and $\hat{\beta}_1$, not p_1 or p_0.
- We replace p_1 in the likelihood with an expression for p_1 in terms of β_0 and β_1
- Similarly, p_0 involves only β_0.

Likelihood

- After removing constants, the new likelihood looks like:

$$\mathrm{Lik}(\beta_0, \beta_1) \propto$$

$$\left(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right)^{22} \left(1 - \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right)^{2} \left(\frac{e^{\beta_0}}{1+e^{\beta_0}}\right)^{141} \left(1 - \frac{e^{\beta_0}}{1+e^{\beta_0}}\right)^{39}$$

- Now what? Fitting the model means finding estimates of β_0 and β_1, but familiar methods from calculus for maximizing the likelihood don't work here.
- Instead, we consider all possible combinations of β_0 and β_1. That is, we will pick that pair of values for β_0 and β_1 that yield the largest likelihood for our data.

Finding the MLE

- Trial and error to find the best pair is tedious at best, but more efficient numerical methods are available.
- The MLEs for the coefficients in the soccer goalkeeper study are $\hat{\beta}_0 = 1.2852$ and $\hat{\beta}_1 = 1.1127$.
- Lets fit the model in R.
- Exponentiating $\hat{\beta}_1$ provides an estimate of the odds ratio (the odds of scoring when the goalkeeper's team is behind, compared to the odds of scoring when the team is not behind) of 3.04, which is consistent with our calculations using the 2 × 2 table.
- We estimate that the odds of scoring when the goalkeeper's team is behind is over 3 times that of when the team is not behind or, in other words, the odds a shooter is successful in a penalty kick shootout are 3.04 times higher when his team is leading.