Chapter 6 Logistic Regression

6.1 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.

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

6.2 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?

Binary Responses: Recall from Section 3.3.1 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.

Binomial Responses: Also recall from Section 3.3.2 that 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.

6.2.1 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.

6.2.2 A Graphical Look at Logistic Regression

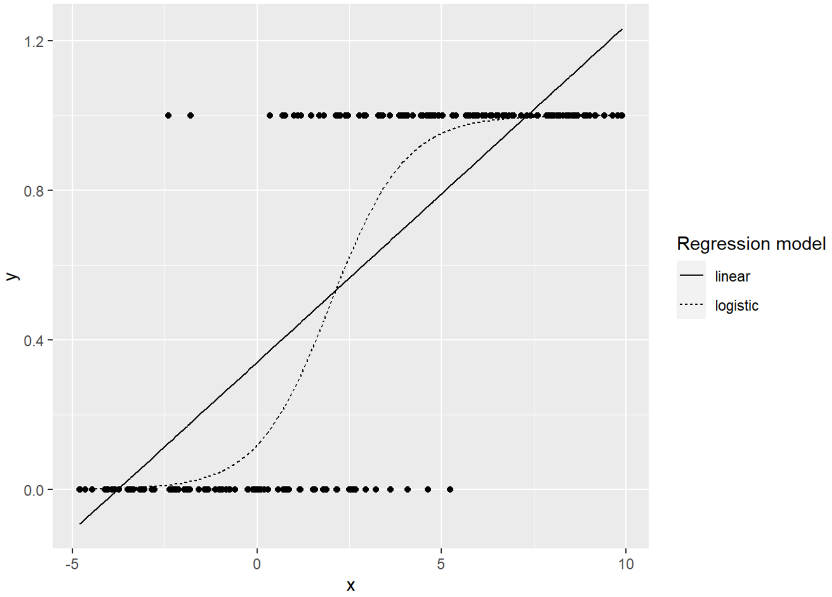

Figure 6.1: Linear vs. logistic regression models for binary response data.

Figure 6.1 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. 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.

6.3 Case Studies Overview

We consider three case studies in this chapter. 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. Here are the statistical concepts you will encounter for each case study.

The soccer goalkeeper data can be written in the form of a 2 × 2 table. This example is used to describe some of the underlying theory for logistic regression. We demonstrate how binomial probability mass functions (pmfs) can be written in one-parameter exponential family form, from which we can identify the canonical link as in Chapter 5. 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.

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.

The 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).

6.4 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? Roskes et al. (2011) 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 6.1.

Table 6.1: Soccer goalkeepers' penalty kick saves when their team is and is not behind.

	Saves	Scores	Total
Behind	2	22	24
Not Behind	39	141	180
Total	41	163	204

(Source: Roskes et al. 2011.)

6.4.1 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. 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.

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:

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

6.4.2 Logistic Regression Models for Binomial Responses

We would like to model the odds of success; however, odds are strictly positive. Therefore, 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:

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

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.

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

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

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

So, 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:

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

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:

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

Thus 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.

The logit model (Equation (6.1)) 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:

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

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

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

We use likelihood methods to estimate β_0 and β_1. As we had done in Chapter 2, we can write the likelihood for this example in the following form:

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

Our interest centers on estimating $\hat{\beta}_0$ and $\hat{\beta}_1$, not p_1 or p_0. So we replace p_1 in the likelihood with an expression for p_1 in terms of β_0 and β_1 as in Equation (6.2). Similarly, p_0 in Equation (6.3) involves only β_0. After removing constants, the new likelihood looks like:

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

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

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. 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 $\hat{\beta}_1 = 1.1127$.

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.

Time out for study discussion (optional).

- Discuss the following quote from the study abstract: "Because penalty takers shot at the two sides of the goal equally often, the goalkeepers' right-oriented bias was dysfunctional, allowing more goals to be scored."
- Construct an argument for why the greater success observed when the goalkeeper's team was behind might be better explained from the shooter's perspective.

Before we go on, you may be curious as to why there is *no error term* in our model statements for logistic or Poisson regression. One way to look at it is to consider that all models describe how observed values are generated. With the logistic model we assume that the observations are generated as binomial random variables. Each observation or realization of Y = number of successes in n independent and identical trials with a probability of success on any one trial of p is produced by $Y \sim \textrm{Binomial}(n,p)$. So the randomness in this model is not introduced by an added error term, but rather by appealing to a binomial probability distribution, where variability depends only on n and p through $\textrm{Var}(Y) = np(1-p)$, and where n is usually considered fixed and p the parameter of interest.

6.4.3 Theoretical Rationale (optional)

Recall from Chapter 5 that generalized linear models (GLMs) are a way in which to model a variety of different types of responses. In this chapter, we apply the general results of GLMs to the specific application of binomial responses. Let Y = the number scored out of n penalty kicks.

The parameter, p, is the probability of a score on a single penalty kick. Recall that the theory of GLMs is based on the unifying notion of the one-parameter exponential family form:

$$f(y;\theta) = e^{[a(y)b(\theta)+c(\theta)+d(y)]}$$

To see that we can apply the general approach of GLMs to binomial responses, we first write an expression for the probability of a binomial response and then use a little algebra to rewrite it until we can demonstrate that it, too, can be written in one-parameter exponential family form with $\theta = p$. This will provide a way in which to specify the canonical link and the form for the model. Additional theory allows us to deduce the mean, standard deviation, and more from this form.

If Y follows a binomial distribution with n trials and probability of success p, we can write:

$$\begin{aligned} P(Y=y) &= \binom{n}{y} p^y (1-p)^{(n-y)} \\ &= e^{y\log(p) + (n-y)\log(1-p) + \log\binom{n}{y}} \end{aligned}$$

However, this probability mass function is not quite in one-parameter exponential family form. Note that there are two terms in the exponent which consist of a product of functions of y and p. So more simplification is in order:

$$P(Y=y) = e^{y\log\left(\frac{p}{1-p}\right) + n\log(1-p) + \log\binom{n}{y}}$$

Don’t forget to consider the support; we must make sure that the set of possible values for this response is not dependent upon p. For fixed n and any value of p, $0 < p < 1$, all integer values from 0 to n are possible, so the support is indeed independent of p.

The one-parameter exponential family form for binomial responses shows that the canonical link is $\log\left(\frac{p}{1-p}\right)$. Thus, GLM theory suggests that constructing a model using the logit, the log odds of a score, as a linear function of covariates is a reasonable approach.

6.5 Case Study: Reconstructing Alabama

This case study demonstrates how wide-ranging applications of statistics can be. Many would not associate statistics with historical research, but this case study shows that it can be done. U.S. Census data from 1870 helped historian Michael Fitzgerald of St. Olaf College gain insight into important questions about how railroads were supported during the Reconstruction Era.

In a paper entitled “Reconstructing Alabama: Reconstruction Era Demographic and Statistical Research,” Ben Bayer performs an analysis of data from 1870 to explain factors that influence voting on referendums related to railroad subsidies (Bayer and Fitzgerald 2011). Positive votes are hypothesized to be inversely proportional to the distance a voter is from the proposed railroad, but the racial composition of a community (as measured by the percentage of Black residents) is hypothesized to be associated with voting behavior as well. Separate analyses of three counties in Alabama—Hale, Clarke, and Dallas—were performed; we discuss Hale County here. This example differs from the soccer example in that it includes continuous covariates. Was voting on railroad referenda related to distance from the proposed railroad line and the racial composition of a community?

6.5.1 Data Organization

The unit of observation for this data is a community in Hale County. We will focus on the following variables from `RR_Data_Hale.csv` collected for each community (see Table 6.2):

- `pctBlack` = the percentage of Black residents in the community
- `distance` = the distance, in miles, the proposed railroad is from the community
- `YesVotes` = the number of “Yes” votes in favor of the proposed railroad line (our primary response variable)
- `NumVotes` = total number of votes cast in the election

Table 6.2: Sample of the data for the Hale County, Alabama, railroad subsidy vote.

community	pctBlack	distance	YesVotes	NumVotes
Carthage	58.4	17	61	110
Cederville	92.4	7	0	15
Greensboro	59.4	0	1790	1804
Havana	58.4	12	16	68

6.5.2 Exploratory Analyses

We first look at a coded scatterplot to see our data. Figure 6.2 portrays the relationship between `distance` and `pctBlack` coded by the `InFavor` status (whether a community supported the referendum with over 50% Yes votes). From this scatterplot, we can see that all of the communities in favor of the railroad referendum had over 55% Black residents, and all of those opposed are 7 miles or farther from the proposed line. The overall percentage of voters in Hale County in favor of the railroad is 87.9%.

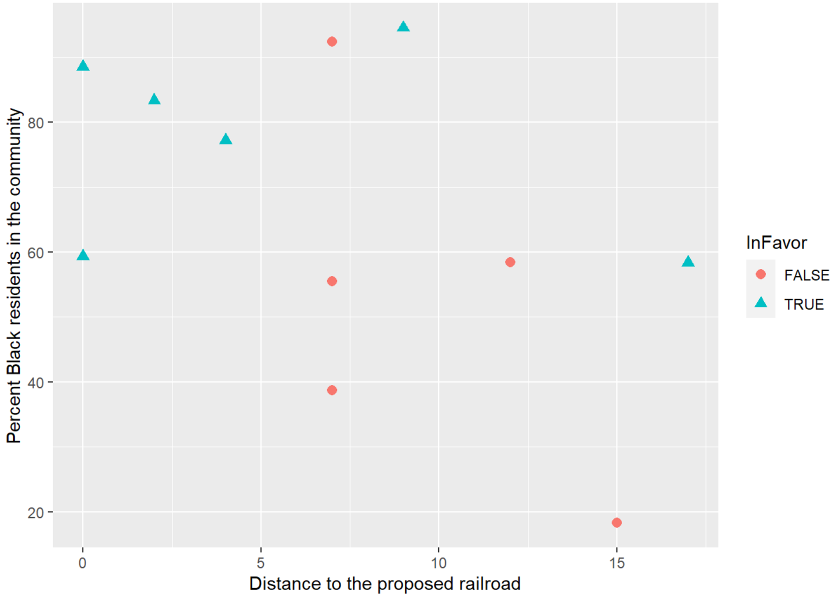

Figure 6.2: Scatterplot of distance from a proposed rail line and percent Black residents in the community coded by whether the community was in favor of the referendum or not.

Recall that a model with two covariates has the form:

$$\log(\text{odds}) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2.$$

where p is the proportion of Yes votes in a community. In logistic regression, we expect the logits to be a linear function of X, the predictors. To assess the linearity assumption, we construct **empirical logit plots**, where “empirical” means “based on sample data.” Empirical logits are computed for each community by taking $\log\left(\frac{\text{number of successes}}{\text{number of failures}}\right)$. In Figure 6.3, we see that the plot of empirical logits versus distance produces a plot that looks linear, as needed for the logistic regression assumption. In contrast, the empirical logits by percent Black residents reveal that Greensboro deviates quite a bit from the otherwise linear pattern; this suggests that Greensboro is an outlier and possibly an influential point. Greensboro has 99.2% voting yes, with only 59.4% Black residents.

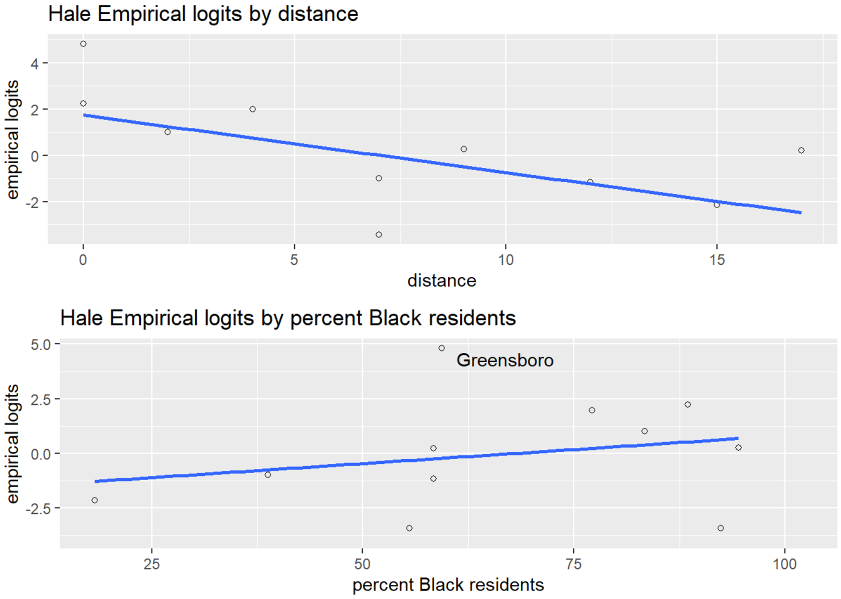

Figure 6.3: Empirical logit plots for the Railroad Referendum data.

In addition to examining how the response correlates with the predictors, it is a good idea to determine whether the predictors correlate with one another. Here, the correlation between distance and percent Black residents is negative and moderately strong with $r = -0.49$. We'll watch to see if the correlation affects the stability of our odds ratio estimates.

6.5.3 Initial Models

The first model includes only one covariate, distance.

```
# Model with just distance
model.HaleD <- glm(cbind(YesVotes, NumVotes - YesVotes) ~
    distance, family = binomial, data = rrHale.df)
# alternative expression
model.HaleD.alt <- glm(YesVotes / NumVotes ~ distance,
    weights = NumVotes, family = binomial, data = rrHale.df)
```

```
##             Estimate Std. Error z value   Pr(>|z|)
## (Intercept)   3.3093    0.11313   29.25 4.268e-188
## distance     -0.2876    0.01302  -22.08 4.447e-108
##  Residual deviance =  318.4  on  9 df
##  Dispersion parameter =  1
```

Our estimated binomial regression model is:

$$\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = 3.309 - 0.288\text{distance}_i$$

where $\hat{p}_i$ is the estimated proportion of Yes votes in community i. The estimated odds ratio for distance, that is the exponentiated coefficient for distance, in this model is $e^{-0.288} = 0.750$. It can be interpreted as follows: for each additional mile from the proposed railroad, the support (odds of a Yes vote) declines by 25.0%.

The covariate `pctBlack` is then added to the first model.

```
model.HaleBD <- glm(cbind(YesVotes, NumVotes - YesVotes) ~
  distance + pctBlack, family = binomial, data = rrHale.df)
```

```
##               Estimate Std. Error z value   Pr(>|z|)
## (Intercept)  4.22202   0.296963  14.217  7.155e-46
## distance    -0.29173   0.013100 -22.270 7.236e-110
## pctBlack    -0.01323   0.003897  -3.394  6.881e-04
```

```
##  Residual deviance =  307.2  on  8 df
##  Dispersion parameter =  1
```

Despite the somewhat strong negative correlation between percent Black residents and distance, the estimated odds ratio for distance remains approximately the same in this new model (OR $= e^{-0.29} = 0.747$); controlling for percent Black residents does little to change our estimate of the effect of distance. For each additional mile from the proposed railroad, odds of a Yes vote declines by 25.3% after adjusting for the racial composition of a community. We also see that, for a fixed distance from the proposed railroad, the odds of a Yes vote declines by 1.3% (OR $= e^{-.0132} = .987$) for each additional percent of Black residents in the community.

6.5.4 Tests for Significance of Model Coefficients

Do we have statistically significant evidence that support for the railroad referendum decreases with higher proportions of Black residents in a community, after accounting for the distance a community is from the railroad line? As discussed in Section 4.4 with Poisson regression, there are two primary approaches to testing significance of model coefficients: **Drop-in-deviance test to compare models** and **Wald test for a single coefficient**.

With our larger model given by $\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1\text{distance}_i + \beta_2\text{pctBlack}_i$, the Wald test produces a highly significant p-value ($Z = \frac{-0.0132}{0.0039} = -3.394$, $p = .00069$) indicating significant evidence that support for the railroad referendum decreases with higher proportions of Black residents in a community, after adjusting for the distance a community is from the railroad line.

The drop-in-deviance test would compare the larger model above to the reduced model $\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1\text{distance}_i$ by comparing residual deviances from the two models.

```
drop_in_dev <- anova(model.HaleD, model.HaleBD, test = "Chisq")
```

```
  ResidDF ResidDev Deviance Df      pval
1       9    318.4       NA NA        NA
2       8    307.2    11.22  1 0.0008083
```

The drop-in-deviance test statistic is $318.44 - 307.22 = 11.22$ on $9 - 8 = 1$ df, producing a p-value of .00081, in close agreement with the Wald test.

A third approach to determining significance of β_2 would be to generate a 95% confidence interval and then checking if 0 falls within the interval or, equivalently, if 1 falls within a 95% confidence interval for e^{β_2}. The next section describes two approaches to producing a confidence interval for coefficients in logistic regression models.

6.5.5 Confidence Intervals for Model Coefficients

Since the Wald statistic follows a normal distribution with n large, we could generate a Wald-type (normal-based) confidence interval for β_2 using:

$$\hat{\beta}_2 \pm 1.96 \cdot \mathrm{SE}(\hat{\beta}_2)$$

and then exponentiating endpoints if we prefer a confidence interval for the odds ratio e^{β_2}. In this case,

$$\begin{aligned}
95\% \text{ CI for } \beta_2 &= \hat{\beta}_2 \pm 1.96 \cdot \mathrm{SE}(\hat{\beta}_2) \\
&= -0.0132 \pm 1.96 \cdot 0.0039 \\
&= -0.0132 \pm 0.00764 \\
&= (-0.0208, -0.0056) \\
95\% \text{ CI for } e^{\beta_2} &= (e^{-0.0208}, e^{-0.0056}) \\
&= (.979, .994) \\
95\% \text{ CI for } e^{10\beta_2} &= (e^{-0.208}, e^{-0.056}) \\
&= (.812, .946)
\end{aligned}$$

Thus, we can be 95% confident that a 10% increase in the proportion of Black residents is associated with a 5.4% to 18.8% decrease in the odds of a Yes vote for the railroad referendum after controlling for distance. This same relationship could be expressed as (a) between a 0.6% and a 2.1% decrease in odds for each 1% increase in the Black population, or (b) between a 5.7% ($1/e^{-.056}$) and a 23.1% ($1/e^{-.208}$) increase in odds for each 10% decrease in the Black population, after adjusting for distance. Of course, with $n = 11$, we should be cautious about relying on a Wald-type interval in this example.

Another approach available in R is the **profile likelihood method**, similar to Section 4.4.

```
exp(confint(model.HaleBD))
```

```
              2.5 %   97.5 %
(Intercept) 38.2285 122.6116
distance     0.7276   0.7660
pctBlack     0.9794   0.9945
```

In the model with `distance` and `pctBlack`, the profile likelihood 95% confidence interval for e^{β_2} is (.979, .994), which is approximately equal to the Wald-based interval despite the small sample size. We can also confirm the statistically significant association between percent Black residents and odds of voting Yes (after controlling for distance), because 1 is not a plausible value of e^{β_2} (where an odds ratio of 1 would imply that the odds of voting Yes do not change with percent Black residents).

6.5.6 Testing for Goodness-of-Fit

As in Section 4.4.9, we can evaluate the goodness-of-fit for our model by comparing the residual deviance (307.22) to a χ^2 distribution with $n - p$ (8) degrees of freedom.

```
1-pchisq(307.2173, 8)  # Goodness-of-fit test
```

```
[1] 0
```

The model with `pctBlack` and `distance` has statistically significant evidence of lack-of-fit ($p < .001$).

Similar to the Poisson regression models, this lack-of-fit could result from (a) missing covariates, (b) outliers, or (c) overdispersion. We will first attempt to address (a) by fitting a model with an interaction between distance and percent Black residents, to determine whether the effect of racial composition differs based on how far a community is from the proposed railroad.

```
model.HaleBxD <- glm(cbind(YesVotes, NumVotes - YesVotes) ~
  distance + pctBlack + distance:pctBlack,
  family = binomial, data = rrHale.df)
```

```
##                    Estimate Std. Error z value
## (Intercept)        7.550902  0.6383697  11.828
## distance          -0.614005  0.0573808 -10.701
## pctBlack          -0.064731  0.0091723  -7.057
## distance:pctBlack  0.005367  0.0008984   5.974
##                    Pr(>|z|)
## (Intercept)       2.783e-32
## distance          1.012e-26
## pctBlack          1.698e-12
## distance:pctBlack 2.321e-09
```

```
##  Residual deviance =  274.2  on  7 df
##  Dispersion parameter =  1
```

```
drop_in_dev <- anova(model.HaleBD, model.HaleBxD,
                     test = "Chisq")
```

```
  ResidDF ResidDev Deviance Df      pval
1       8    307.2       NA NA        NA
2       7    274.2    32.98  1 9.294e-09
```

We have statistically significant evidence (Wald test: $Z = 5.974, p < .001$; Drop-in-deviance test: $\chi^2 = 32.984, p < .001$) that the effect of the proportion of community residents who are Black on the odds of voting Yes depends on the distance of the community from the proposed railroad.

To interpret the interaction coefficient in context, we will compare two cases: one where a community is right on the proposed railroad (`distance` = 0), and the other where the community is 15 miles away (`distance` = 15). The significant interaction implies that the effect of `pctBlack` should differ in these two cases. In the first case, the coefficient for `pctBlack` is -0.0647, while in the second case, the relevant coefficient is $-0.0647 + 15(.00537) = 0.0158$. Thus, for a community right on the proposed railroad, a 1% increase in percent Black residents is associated with a 6.3% ($e^{-.0647} = .937$) decrease in the odds of voting Yes, while for a community 15 miles away, a

1% increase in percent Black residents is associated with a ($e^{.0158} = 1.016$) 1.6% *increase* in the odds of voting Yes. A significant interaction term doesn't always imply a change in the direction of the association, but it does here.

Because our interaction model still exhibits lack-of-fit (residual deviance of 274.23 on just 7 df), and because we have used the covariates at our disposal, we will assess this model for potential outliers and overdispersion by examining the model's residuals.

6.5.7 Residuals for Binomial Regression

With LLSR, residuals were used to assess model assumptions and identify outliers. For binomial regression, two different types of residuals are typically used. One residual, the **Pearson residual**, has a form similar to that used with LLSR. Specifically, the Pearson residual is calculated using:

$$\text{Pearson residual}_i = \frac{\text{actual count} - \text{predicted count}}{\text{SD of count}} = \frac{Y_i - m_i\hat{p}_i}{\sqrt{m_i\hat{p}_i(1 - \hat{p}_i)}}$$

where m_i is the number of trials for the i^{th} observation and $\hat{p}_i$ is the estimated probability of success for that same observation.

A **deviance residual** is an alternative residual for binomial regression based on the discrepancy between the observed values and those estimated using the likelihood. A deviance residual can be calculated for each observation using:

$$\text{d}_i = \text{sign}(Y_i - m_i\hat{p}_i)\sqrt{2[Y_i\log\left(\frac{Y_i}{m_i\hat{p}_i}\right) + (m_i - Y_i)\log\left(\frac{m_i - Y_i}{m_i - m_i\hat{p}_i}\right)]}$$

When the number of trials is large for all of the observations and the models are appropriate, both sets of residuals should follow a standard normal distribution.

The sum of the individual deviance residuals is referred to as the **deviance** or **residual deviance**. The residual deviance is used to assess the model. As the name suggests, a model with a small deviance is preferred. In the case of binomial regression, when the denominators, m_i, are large and a model fits, the residual deviance follows a χ^2 distribution with $n-p$ degrees of freedom (the residual degrees of freedom). Thus for a good fitting model the residual deviance should be approximately equal to its corresponding degrees of freedom. When binomial data meets these conditions, the deviance can be used for a goodness-of-fit test. The p-value for lack-of-fit is the proportion of values from a χ^2_{n-p} distribution that are greater than the observed residual deviance.

We begin a residual analysis of our interaction model by plotting the residuals against the fitted values in Figure 6.4. This kind of plot for binomial regression would produce two linear trends with similar negative slopes if there were equal sample sizes m_i for each observation.

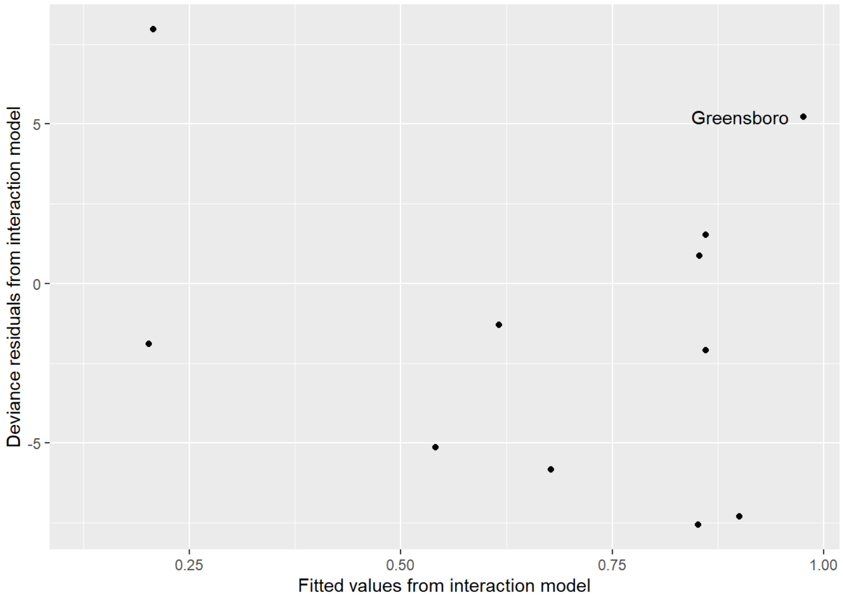

Figure 6.4: Fitted values by residuals for the interaction model for the Railroad Referendum data.

From this residual plot, Greensboro does not stand out as an outlier. If it did, we could remove Greensboro and refit our interaction model, checking to see if model coefficients changed in a noticeable way. Instead, we will continue to include Greensboro in our modeling efforts. Because the large residual deviance cannot be explained by outliers, and given we have included all of the covariates at hand as well as an interaction term, the observed binomial counts are likely overdispersed. This means that they exhibit more variation than the model would suggest, and we must consider ways to handle this overdispersion.

6.5.8 Overdispersion

Similar to Poisson regression, we can adjust for overdispersion in binomial regression. With overdispersion there is **extra-binomial variation**, so the actual variance will be greater than the variance of a binomial variable, $np(1-p)$. One way to adjust for overdispersion is to estimate a multiplier (dispersion parameter), $\hat{\phi}$, for the variance that will inflate it and reflect the reduction in the amount of information we would otherwise have with independent observations. We used a similar approach to adjust for overdispersion in a Poisson regression model in Section 4.9, and we will use the same estimate here: $\hat{\phi} = \frac{\sum(\text{Pearson residuals})^2}{n-p}$.

When overdispersion is adjusted for in this way, we can no longer use maximum likelihood to fit our regression model; instead we use a quasilikelihood approach. Quasilikelihood is similar to likelihood-based inference, but because the model uses the dispersion parameter, it is not a binomial model with a true likelihood (we call it **quasibinomial**). R offers quasilikelihood as an option when model fitting. The quasilikelihood approach will yield the same coefficient point estimates as maximum likelihood; however, the variances will be larger in the presence of overdispersion (assuming $\phi > 1$). We will see other ways in which to deal with overdispersion and clusters in the remaining chapters in the book, but the following describes how overdispersion is accounted for using $\hat{\phi}$:

Summary: accounting for overdispersion

- Use the dispersion parameter $\hat{\phi} = \frac{\sum(\text{Pearson residuals})^2}{n-p}$ to inflate standard errors of model coefficients.
- Wald test statistics: multiply the standard errors by $\sqrt{\hat{\phi}}$ so that $\text{SE}_\text{Q}(\hat{\beta}) = \sqrt{\hat{\phi}} \cdot \text{SE}(\hat{\beta})$ and conduct tests using the t-distribution.
- Confidence intervals use the adjusted standard errors and multiplier based on t, so they are thereby wider: $\hat{\beta} \pm t_{n-p} \cdot \text{SE}_\text{Q}(\hat{\beta})$.
- Drop-in-deviance test statistic comparing Model 1 (larger model with p parameters) to Model 2 (smaller model with $q < p$ parameters) is $F = \frac{1}{\hat{\phi}} \cdot \frac{D_2 - D_1}{p-q}$ where D_1 and D_2 are the residual deviances for models 1 and 2, respectively, and $p-q$ is the difference in the number of parameters for the two models. Note that both $D_2 - D_1$ and $p-q$ are positive. This test statistic is compared to an F-distribution with $p-q$ and $n-p$ degrees of freedom.

Output for a model which adjusts our interaction model for overdispersion appears below, where $\hat{\phi} = 51.6$ is used to adjust the standard errors for the coefficients and the drop-in-deviance tests during model building. Standard errors will be inflated by a factor of $\sqrt{51.6} = 7.2$. As a result, there are no significant terms in the adjusted interaction model below.

```
model.HaleBxDq <- glm(cbind(YesVotes, NumVotes - YesVotes) ~
  distance + pctBlack + distance:pctBlack,
  family = quasibinomial, data = rrHale.df)
```

```
##                     Estimate Std. Error t value
## (Intercept)         7.550902   4.585464  1.6467
## distance           -0.614005   0.412171 -1.4897
## pctBlack           -0.064731   0.065885 -0.9825
## distance:pctBlack   0.005367   0.006453  0.8316
##                    Pr(>|t|)
## (Intercept)          0.1436
## distance             0.1799
## pctBlack             0.3586
## distance:pctBlack    0.4331
```

```
##  Residual deviance =  274.2  on  7 df
##  Dispersion parameter =  51.6
```

We therefore remove the interaction term and refit the model, adjusting for the extra-binomial variation that still exists.

```
model.HaleBDq <- glm(cbind(YesVotes, NumVotes - YesVotes) ~
  distance + pctBlack,
  family = quasibinomial, data = rrHale.df)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.22202    1.99031  2.1213  0.06669
## distance    -0.29173    0.08780 -3.3228  0.01050
## pctBlack    -0.01323    0.02612 -0.5064  0.62620
##  Residual deviance =  307.2  on  8 df
##  Dispersion parameter =  44.92
```

By removing the interaction term and using the overdispersion parameter, we see that distance is significantly associated with support, but percent Black residents is no longer significant after adjusting for distance.

Because quasilikelihood methods do not change estimated coefficients, we still estimate a 25% decline $(1-e^{-0.292})$ in support for each additional mile from the proposed railroad (odds ratio of .75).

```
exp(confint(model.HaleBDq))
```

```
             2.5 %   97.5 %
(Intercept) 1.3609 5006.722
distance    0.6091    0.871
pctBlack    0.9366    1.044
```

While we previously found a 95% confidence interval for the odds ratio associated with distance of (.728, .766), our confidence interval is now much wider: (.609, .871). Appropriately accounting for overdispersion has changed both the significance of certain terms and the precision of our coefficient estimates.

6.5.9 Summary

We began by fitting a logistic regression model with `distance` alone. Then we added the covariate `pctBlack` , and the Wald-type test and the drop-in-deviance test both provided strong support for the addition of `pctBlack` to the model. The model with `distance` and `pctBlack` had a large residual deviance suggesting an ill-fitted model. When we looked at the residuals, we saw that Greensboro is an extreme observation. Models without Greensboro were fitted and compared to our initial models. Seeing no appreciable improvement or differences with Greensboro removed, we left it in the model. There remained a large residual deviance so we attempted to account for it by using an estimated dispersion parameter similar to Section 4.9 with Poisson regression. The final model included distance and percent Black residents, although percent Black residents was no longer significant after adjusting for overdispersion.

6.6 Linear Least Squares vs. Binomial Regression

$$
\begin{gathered}
\text{Response} \\
\_ \\
\textbf{LLSR:}\ \text{normal} \\
\textbf{Binomial Regression:}\ \text{number of successes in n trials}
\end{gathered}
$$

$$
\begin{gathered}
\text{Variance} \\
\_ \\
\textbf{LLSR:}\ \text{equal for each level of } X \\
\textbf{Binomial Regression:}\ np(1-p) \text{ for each level of } X
\end{gathered}
$$

$$
\begin{gathered}
\text{Model Fitting} \\
\_ \\
\textbf{LLSR:}\ \mu = \beta_0 + \beta_1 x \text{ using Least Squares} \\
\textbf{Binomial Regression:}\ \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x \text{ using Maximum Likelihood}
\end{gathered}
$$

$$
\begin{gathered}
\text{EDA} \\
\_ \\
\textbf{LLSR:}\ \text{plot } X \text{ vs. } Y\text{; add line} \\
\textbf{Binomial Regression:}\ \text{find log(odds) for several subgroups; plot vs. } X
\end{gathered}
$$

$$
\begin{gathered}
\text{Comparing Models} \\
\_ \\
\textbf{LLSR:}\ \text{extra sum of squares F-tests; AIC/BIC} \\
\textbf{Binomial Regression:}\ \text{drop-in-deviance tests; AIC/BIC}
\end{gathered}
$$

$$
\begin{gathered}
\text{Interpreting Coefficients} \\
\_ \\
\textbf{LLSR:}\ \beta_1 = \text{ change in mean response for unit change in } X \\
\textbf{Binomial Regression:}\ e^{\beta_1} = \text{ percent change in odds for unit change in } X
\end{gathered}
$$

6.7 Case Study: Trying to Lose Weight

The final case study uses individual-specific information so that our response, rather than the number of successes out of some number of trials, is simply a binary variable taking on values of 0 or 1 (for failure/success, no/yes, etc.). This type of problem—**binary logistic regression**—is exceedingly common in practice. Here we examine characteristics of young people who are trying to lose weight. The prevalence of obesity among U.S. youth suggests that wanting to lose weight is sensible and desirable for some young people such as those with a high body mass index (BMI). On the flip side, there are young people who do not need to lose weight but make ill-advised attempts to do so nonetheless. A multitude of studies on weight loss focus specifically on youth and propose a variety of motivations for the young wanting to lose weight; athletics and the media are two commonly cited sources of motivation for losing weight for young people.

Sports have been implicated as a reason for young people wanting to shed pounds, but not all studies are consistent with this idea. For example, a study by Martinsen et al. (2009) reported that, despite preconceptions to the contrary, there was a higher rate of self-reported eating disorders among controls (non-elite athletes) as opposed to elite athletes. Interestingly, the kind of sport was not found to be a factor, as participants in leanness sports (for example, distance running, swimming, gymnastics, dance, and diving) did not differ in the proportion with eating disorders when compared to those in non-leanness sports. So, in our analysis, we will not make a distinction between different sports.

Other studies suggest that mass media is the culprit. They argue that students' exposure to unrealistically thin celebrities may provide unhealthy motivation for some, particularly young women, to try to slim down. An examination and analysis of a large number of related studies (referred to as a **meta-analysis**) (Grabe, Hyde, and Ward 2008) found a strong relationship between exposure to mass media and the amount of time that adolescents spend talking about what they see in the media, deciphering what it means, and figuring out how they can be more like the celebrities.

We are interested in the following questions: Are the odds that young females report trying to lose weight greater than the odds that males do? Is increasing BMI associated with an interest in losing weight, regardless of sex? Does sports participation increase the desire to lose weight? Is media exposure associated with more interest in losing weight?

We have a sample of 500 teens from data collected in 2009 through the U.S. Youth Risk Behavior Surveillance System (YRBSS) (Centers for Disease Control and Prevention 2009). The YRBSS is an annual national school-based survey conducted by the Centers for Disease Control and Prevention (CDC) and state, territorial, and local education and health agencies and tribal governments. More information on this survey can be found here.

6.7.1 Data Organization

Here are the three questions from the YRBSS we use for our investigation:

Q66. Which of the following are you trying to do about your weight?

- A. Lose weight
- B. Gain weight
- C. Stay the same weight
- D. I am not trying to do anything about my weight

Q81. On an average school day, how many hours do you watch TV?

- A. I do not watch TV on an average school day
- B. Less than 1 hour per day
- C. 1 hour per day
- D. 2 hours per day
- E. 3 hours per day
- F. 4 hours per day
- G. 5 or more hours per day

Q84. During the past 12 months, on how many sports teams did you play? (Include any teams run by your school or community groups.)

- A. 0 teams
- B. 1 team
- C. 2 teams
- D. 3 or more teams

Answers to Q66 are used to define our response variable: Y = 1 corresponds to “(A) trying to lose weight”, while Y = 0 corresponds to the other non-missing values. Q84 provides information on students’ sports participation and is treated as numerical, 0 through 3, with 3 representing 3 or more. As a proxy for media exposure, we use answers to Q81 as numerical values 0, 0.5, 1, 2, 3, 4, and 5, with 5 representing 5 or more. Media exposure and sports participation are also considered as categorical factors, that is, as variables with distinct levels which can be denoted by indicator variables as opposed to their numerical values.

BMI is included in this study as the percentile for a given BMI for members of the same sex. This facilitates comparisons when modeling with males and females. We will use the terms *BMI* and *BMI percentile* interchangeably with the understanding that we are always referring to the percentile.

With our sample, we use only the cases that include all of the data for these four questions. This is referred to as a **complete case analysis**. That brings our sample of 500 to 445. There are limitations of complete case analyses that we address in the Discussion.

6.7.2 Exploratory Data Analysis

Nearly half (44.7%) of our sample of 445 youths report that they are trying to lose weight, 48.1% of the sample are females, and 59.3% play on one or more sports teams. Also, 8.8% report that they do not watch any TV on school days, whereas another 13.0% watched 5 or more hours each day. Interestingly, the median BMI percentile for our 445 youths is 68. The most dramatic difference in the proportions of those who are trying to lose weight is by sex; 58% of the females want to lose weight in contrast to only 32% of the males (see Figure 6.5). This provides strong support for the inclusion of a sex term in every model considered.

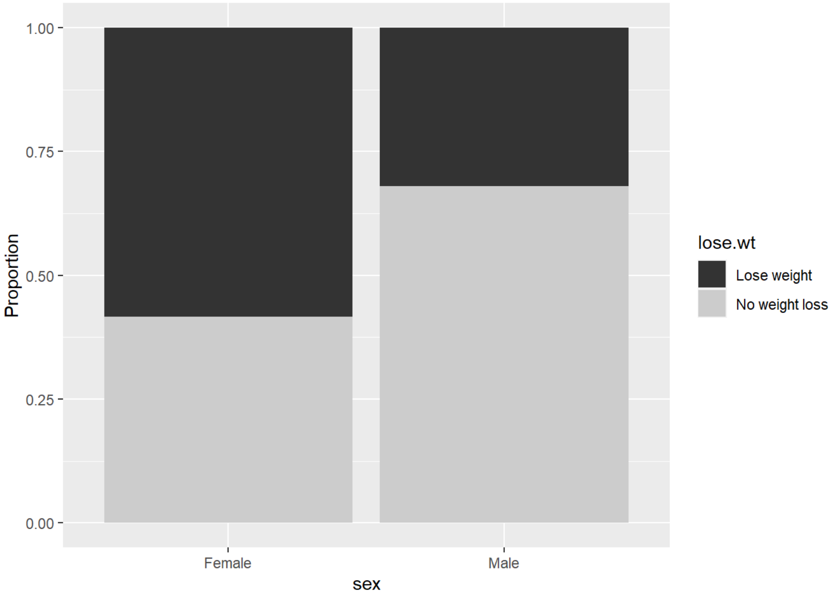

Figure 6.5: Weight loss plans vs. sex.

Table 6.3: Mean BMI percentile by sex and desire to lose weight.

Sex	Weight loss status	mean BMI percentile	SD	n
Female	No weight loss	43.2	25.8	89
	Lose weight	72.4	23.0	125
Male	No weight loss	58.8	28.2	157
	Lose weight	85.7	18.0	74

Table 6.3 displays the mean BMI of those wanting and not wanting to lose weight for males and females. The mean BMI is greater for those trying to lose weight compared to those not trying to lose weight, regardless of sex. The size of the difference is remarkably similar for the two sexes.

If we consider including a BMI term in our model(s), the logit should be linearly related to BMI. We can investigate this assumption by constructing an empirical logit plot. In order to calculate empirical logits, we first divide our data by sex. Within each sex, we generate 10 groups of equal sizes, the first holding the bottom 10% in BMI percentile for that sex, the second holding the next lowest 10%, etc. Within each group, we calculate the proportion, $\hat{p}$ that reported wanting to lose weight, and then the empirical log odds, $log(\frac{\hat{p}}{1-\hat{p}})$, that a young person in that group wants to lose weight.

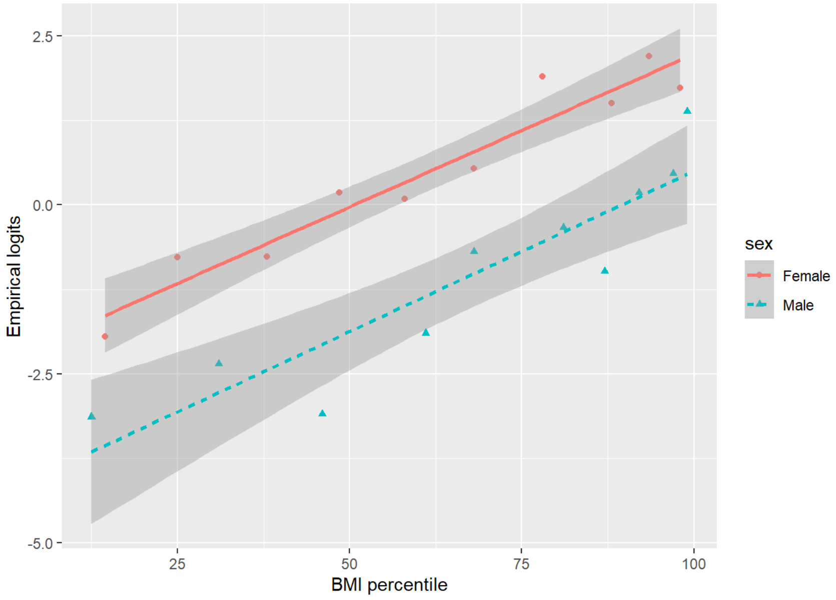

Figure 6.6: Empirical logits of trying to lose weight by BMI and sex.

Figure 6.6 presents the empirical logits for the BMI intervals by sex. Both males and females exhibit an increasing linear trend on the logit scale indicating that increasing BMI is associated with a greater desire to lose weight and that modeling log odds as a linear function of BMI is reasonable. The slope for the females appears to be similar to the slope for males, so we do not need to consider an interaction term between BMI and sex in the model.

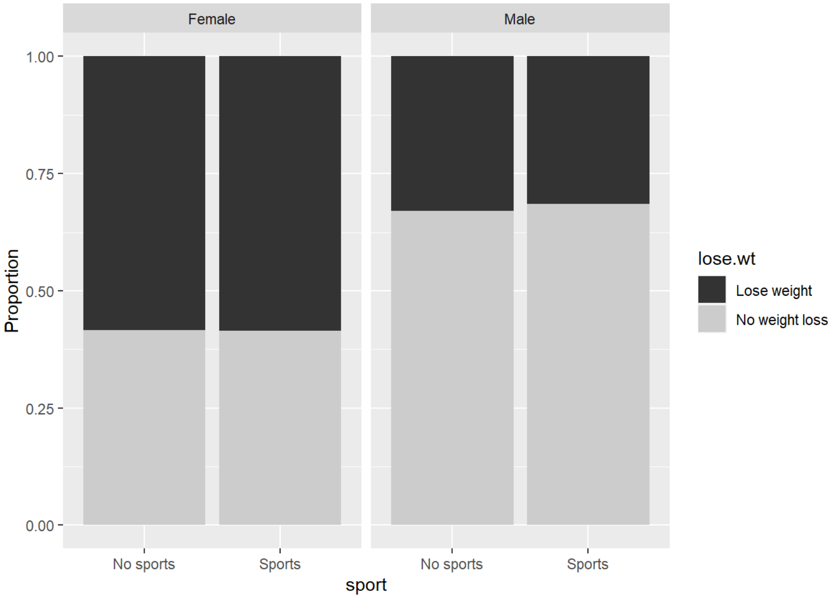

Figure 6.7: Weight loss plans vs. sex and sports participation.

Out of those who play sports, 44% want to lose weight, whereas 46% want to lose weight among those who do not play sports. Figure 6.7 compares the proportion of respondents who want to lose weight by their sex and sport participation. The data suggest that sports participation is associated with the same or even a slightly lower desire to lose weight, contrary to what had originally been hypothesized. While the overall levels of those wanting to lose weight differ considerably between the sexes, the differences between those in and out of sports within sex appear to be very small. A term for sports participation or number of teams will be considered, but there is not compelling evidence that an interaction term will be needed.

It was posited that increased exposure to media, here measured as hours of TV daily, is associated with increased desire to lose weight, particularly for females. Overall, the percentage who want to lose weight ranges from 38% of those watching 5 hours of TV per day to 55% among those watching 2 hours daily. There is minimal variation in the proportion wanting to lose weight with both sexes combined. However, we are more interested in differences between the sexes (see Figure 6.8). We create empirical logits using the proportion of students trying to lose weight for each level of hours spent watching TV daily and look at the trends in the logits separately for males and females. From Figure 6.9, there does not appear to be a linear relationship for males or females.

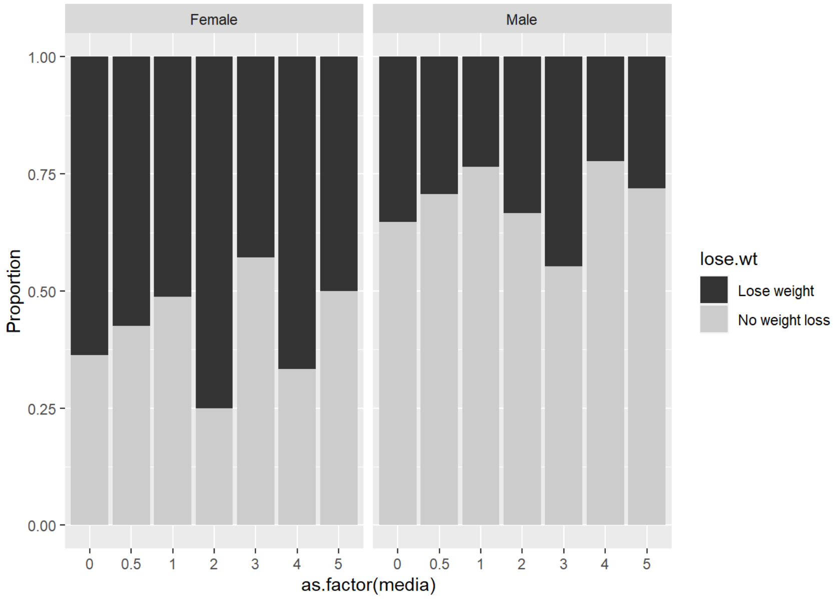

Figure 6.8: Weight loss plans vs. daily hours of TV and sex.

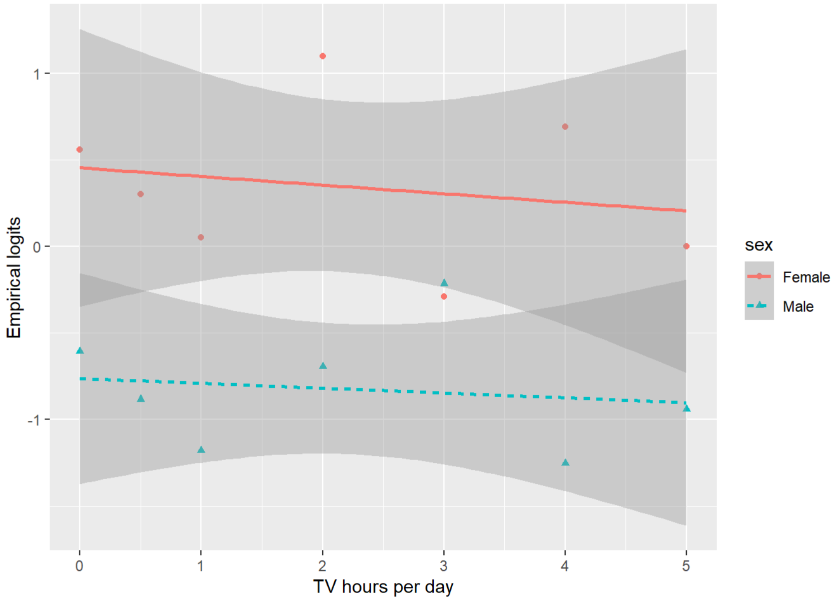

Figure 6.9: Empirical logits for the odds of trying to lose weight by TV watching and sex.

6.7.3 Initial Models

Our strategy for modeling is to use our questions of interest and what we have learned in the exploratory data analysis. For each model we interpret the coefficient of interest, look at the corresponding Wald test and, as a final step, compare the deviances for the different models we considered.

We first use a model where sex is our only predictor.

```
model1 <- glm(lose.wt.01 ~ female, family = binomial,
              data = risk2009)
```

```
##               Estimate Std. Error z value  Pr(>|z|)
## (Intercept)  -0.7522     0.1410  -5.334 9.588e-08
## female        1.0919     0.1978   5.520 3.382e-08
```

```
##  Residual deviance =  580.3  on  443 df
```

Our estimated binomial regression model is:

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -0.75 + 1.09\text{female}$$

where $\hat{p}$ is the estimated proportion of youth wanting to lose weight. We can interpret the coefficient on `female` by exponentiating $e^{1.0919} = 2.98$ (95% CI = $(2.03, 4.41)$) indicating that the odds of a female trying to lose weight is nearly three times the odds of a male trying to lose weight ($Z = 5.520$, $p = 3.38e-08$). We retain sex in the model and consider adding the BMI percentile:

```
model2 <- glm(lose.wt.01 ~ female + bmipct,
              family = binomial, data = risk2009)
```

```
##               Estimate Std. Error z value  Pr(>|z|)
## (Intercept) -4.25914    0.44927  -9.480 2.541e-21
## female       1.86067    0.25896   7.185 6.714e-13
## bmipct       0.04715    0.00524   8.997 2.313e-19
```

```
##  Residual deviance =  463  on  442 df
```

We see that there is statistically significant evidence ($Z = 8.997, p < .001$) that BMI is positively associated with the odds of trying to lose weight, after controlling for sex. Clearly BMI percentile belongs in the model with sex.

Our estimated binomial regression model is:

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -4.26 + 1.86\text{female} + 0.047\text{bmipct}$$

To interpret the coefficient on `bmipct`, we will consider a 10-unit increase in `bmipct`. Because $e^{10*0.047} = 1.602$, then there is an estimated 60.2% increase in the odds of wanting to lose weight for each additional 10 percentile points of BMI for members of the same sex. Just as we had done in other multiple regression models, we need to interpret our coefficient *given that the other variables remain constant*. An interaction term for BMI by sex was tested (not shown) and it was not significant ($Z = -0.70$, $p = 0.485$), so the effect of BMI does not differ by sex.

We next add `sport` to our model. Sports participation was considered for inclusion in the model in three ways: an indicator of sports participation (0 = no teams, 1 = one or more teams), treating the number of teams (0, 1, 2, or 3) as numeric, and treating the number of teams as a factor. The models below treat sports participation using an indicator variable, but all three models produced similar results.

```
model3 <- glm(lose.wt.01 ~ female + bmipct + sport,
              family = binomial, data = risk2009)
```

```
##             Estimate Std. Error z value  Pr(>|z|)
## (Intercept) -4.17138   0.468463 -8.9044 5.367e-19
## female       1.84951   0.259514  7.1268 1.027e-12
## bmipct       0.04728   0.005251  9.0032 2.193e-19
## sportSports -0.14767   0.235101 -0.6281 5.299e-01
```

```
##  Residual deviance =  462.6  on  441 df
```

```
model3int <- glm(lose.wt.01 ~ female + bmipct + sport +
              female:sport + bmipct:sport,
              family = binomial, data = risk2009)
```

```
##                      Estimate Std. Error z value
## (Intercept)        -3.643635   0.604821  -6.024
## female              1.451017   0.378547   3.833
## bmipct              0.042530   0.007211   5.898
## sportSports        -1.187199   0.893057  -1.329
## female:sportSports  0.731516   0.523566   1.397
## bmipct:sportSports  0.009908   0.010463   0.947
##                      Pr(>|z|)
## (Intercept)        1.698e-09
## female             1.265e-04
## bmipct             3.684e-09
## sportSports        1.837e-01
## female:sportSports 1.624e-01
## bmipct:sportSports 3.436e-01
```

```
##  Residual deviance =  460.5  on  439 df
```

Sports teams were not significant in any of these models, nor were interaction terms (sex by sports and bmipct by sports). As a result, sports participation was no longer considered for inclusion in the model.

We last look at adding `media` to our model.

```
model4 <- glm(lose.wt.01 ~ female + bmipct + media,
              family = binomial, data = risk2009)
```

```
##             Estimate Std. Error z value  Pr(>|z|)
## (Intercept) -4.08892   0.462947  -8.832 1.025e-18
## female       1.84776   0.259636   7.117 1.105e-12
## bmipct       0.04783   0.005287   9.046 1.485e-19
## media       -0.09938   0.072464  -1.371 1.702e-01
```

```
##  Residual deviance =  461.1  on  441 df
```

Media is not a statistically significant term ($Z = -1.371, p = 0.170$). However, because our interest centers on how media may affect attempts to lose weight and how its effect might be different for females and males, we fit a model with a media term and a sex by media interaction term (not shown). Neither term was statistically significant, so we have no support in our data that media exposure as measured by hours spent watching TV is associated with the odds a teen is trying to lose weight after accounting for sex and BMI.

6.7.4 Drop-in-Deviance Tests

```
drop_in_dev <- anova(model1, model2, model3, model4,
                     test="Chisq")
```

```
  ResidDF ResidDev Deviance Df      pval
1     443    580.3       NA NA        NA
2     442    463.0 117.3301  1 2.431e-27
3     441    462.6   0.3947  1 5.298e-01
4     441    461.1   1.5007  0        NA
```

```
       df   AIC
model1  2 584.3
model2  3 469.0
model3  4 470.6
model4  4 469.1
```

Comparing models using differences in deviances requires that the models be **nested**, meaning each smaller model is a simplified version of the larger model. In our case, Models 1, 2, and 4 are nested, as are Models 1, 2, and 3, but Models 3 and 4 cannot be compared using a drop-in-deviance test.

There is a large drop-in-deviance adding BMI to the model with sex (Model 1 to Model 2, 117.3), which is clearly statistically significant when compared to a χ^2 distribution with 1 df. The drop-in-deviance for adding an indicator variable for sports to the model with sex and BMI is only 462.99 - 462.59 = 0.40. There is a difference of a single parameter, so the drop-in-deviance would be compared to a χ^2 distribution with 1 df. The resulting p-value is very large (.53) suggesting that adding an indicator for sports is not helpful once we've already accounted for BMI and sex. For comparing Models 3 and 4, one approach is to look at the AIC. In this case, the AIC is (barely) smaller for the model with media, providing evidence that the latter model is slightly preferable.

6.7.5 Model Discussion and Summary

We found that the odds of wanting to lose weight are considerably greater for females compared to males. In addition, respondents with greater BMI percentiles express a greater desire to lose weight for members of the same sex. Regardless of sex or BMI percentile, sports participation and TV watching are not associated with different odds for wanting to lose weight.

A limitation of this analysis is that we used complete cases in place of a method of imputing responses or modeling missingness. This reduced our sample from 500 to 445, and it may have introduced bias. For example, if respondents who watch a lot of TV were unwilling to reveal as much, and if they differed with respect to their desire to lose weight from those respondents who reported watching little TV, our inferences regarding the relationship between lots of TV and desire to lose weight may be biased.

Other limitations may result from definitions. Trying to lose weight is self-reported and may not correlate with any action undertaken to do so. The number of sports teams may not accurately reflect sports-related pressures to lose weight. For example, elite athletes may focus on a single sport and be subject to greater pressures, whereas athletes who casually participate in three sports may not feel any pressure to lose weight. Hours spent watching TV are not likely to encompass the totality of media exposure, particularly because exposure to celebrities occurs often online. Furthermore, this analysis does not explore in any detail maladaptions—inappropriate motivations for wanting to lose weight. For example, we did not focus our study on subsets of respondents with low BMI who are attempting to lose weight.

It would be instructive to use data science methodologies to explore the entire data set of 16,000 instead of sampling 500. However, the types of exploration and models used here could translate to the larger sample size.

Finally a limitation may be introduced as a result of the acknowledged variation in the administration of the YRBSS. States and local authorities are allowed to administer the survey as they see fit, which at times results in significant variation in sample selection and response.

6.8 Exercises

6.8.1 Conceptual Exercises

1. List the explanatory and response variable(s) for each research question.

    a. Are students with poor grades more likely to binge drink?
    b. What is the chance you are accepted into medical school given your GPA and MCAT scores?
    c. Is a single mom more likely to marry the baby’s father if she has a boy?
    d. Are students participating in sports in college more or less likely to graduate?
    e. Is exposure to a particular chemical associated with a cancer diagnosis?

2. Interpret the odds ratios in the following abstract.

    Daycare Centers and Respiratory Health (Nafstad et al. 1999)

    - **Objective**. To estimate the effects of the type of daycare on respiratory health in preschool children.

    - **Methods**. A population-based, cross-sectional study of Oslo children born in 1992 was conducted at the end of 1996. A self-administered questionnaire inquired about daycare arrangements, environmental conditions, and family characteristics (n = 3853; response rate, 79%).

    - **Results**. In a logistic regression controlling for confounding, children in daycare centers had more often nightly cough (adjusted odds ratio, 1.89; 95% confidence interval 1.34-2.67), and blocked or runny nose without common cold (1.55; 1.07-1.61) during the past 12 months compared with children in home care.

3. Construct a table and calculate the corresponding odds and odds ratios. Comment on the reported and calculated results in this *New York Times* article from Kolata (2009).

    - In November, the Centers for Disease Control and Prevention published a paper reporting that babies conceived with IVF, or with a technique in which sperm are injected directly into eggs, have a slightly increased risk of several birth defects, including a hole between the two chambers of the heart, a cleft lip or palate, an improperly developed esophagus and a malformed rectum. The study involved 9,584 babies with birth defects and 4,792 babies without. Among the mothers of babies without birth defects, 1.1% had used IVF or related methods, compared with 2.4% of mothers of babies with birth defects.

    - The findings are considered preliminary, and researchers say they believe IVF does not carry excessive risks. There is a 3% chance that any given baby will have a birth defect.

4. In a small pilot study, researchers compared two groups of 3 turbine wheels each under low humidity and two groups of 3 turbine wheels each under high-humidity conditions to determine if humidity is related to the number of fissures that occur. If Y = number of turbine wheels that develop fissures, then assume that $Y \sim \text{Binomial}(n = 3, p = p_L)$ under low humidity, and $Y \sim \text{Binomial}(n = 3, p = p_H)$ under high humidity, where $f(y; p) = \binom{n}{y} p^y (1 - p)^{n-y}$. Write out the log-likelihood function $\log\text{L}(p_L, p_H)$, using the data in Table 6.4 and simplifying where possible.

Table 6.4: Data for Conceptual Exercise 4.

Turbine group	**1**	**2**	**3**	**4**
Humidity	Low	Low	High	High
n = number of turbine wheels	3	3	3	3
y = number of fissures	1	2	1	0

6.8.2 Guided Exercises

1. **Soccer goals on target.** Data comes from an article in *Psychological Science* (Roskes et al. 2011). The authors report on the success rate of penalty kicks that were on-target, so that either the keeper saved the shot or the shot scored, for FIFA World Cup shootouts between 1982 and 2010. They found that 18 out of 20 shots were scored when the goalkeeper's team was behind, 71 out of 90 shots were scored when the game was tied, and 55 out of 75 shots were scored with the goalkeeper's team ahead.

    a. Calculate the odds of a successful penalty kick for games in which the goalkeeper's team was behind, tied, or ahead. Then, construct empirical odds ratios for successful penalty kicks for (a) behind versus tied, and (b) tied versus ahead.

    b. Fit a model with the categorical predictor c("behind","tied","ahead") and interpret the exponentiated coefficients. How do they compare to the empirical odds ratios you calculated?

2. **Medical school admissions.** The data for Medical School Admissions is in `MedGPA.csv` , taken from undergraduates from a small liberal arts school over several years. We are interested in student attributes that are associated with higher acceptance rates.

    - `Accept` = accepted (A) into medical school or denied (D)
    - `Acceptance` = accepted (1) into medical school or denied (0)
    - `Sex` = male (M) or female (F)
    - `BCPM` = GPA in natural sciences and mathematics
    - `GPA` = overall GPA
    - `VR` = verbal reasoning subscale score of the MCAT
    - `PS` = physical sciences subscale score of the MCAT
    - `WS` = writing samples subscale score of the MCAT
    - `BS` = biological sciences subscale score of the MCAT
    - `MCAT` = MCAT total score
    - `Apps` = number of schools applied to

    Be sure to interpret model coefficients and associated tests of significance or confidence intervals when answering the following questions:

    a. Compare the relative effects of improving your MCAT score versus improving your GPA on your odds of being accepted to medical school.
    b. After controlling for MCAT and GPA, is the number of applications related to odds of getting into medical school?
    c. Is one MCAT subscale more important than the others?

d. Is there any evidence that the effect of MCAT score or GPA differs for males and females?

3. **Moths.** An article in the *Journal of Animal Ecology* by Bishop (1972) investigated whether moths provide evidence of "survival of the fittest" with their camouflage traits. Researchers glued equal numbers of light and dark morph moths in lifelike positions on tree trunks at 7 locations from 0 to 51.2 km from Liverpool. They then recorded the number of moths removed after 24 hours, presumably by predators. The hypothesis was that, since tree trunks near Liverpool were blackened by pollution, light morph moths would be more likely to be removed near Liverpool.

   Data (Ramsey and Schafer 2002) can be found in `moth.csv` and contains the variables below. In addition, R code at the end of the problem can be used to input the data and create additional useful variables.

   - `MORPH` = light or dark
   - `DISTANCE` = kilometers from Liverpool
   - `PLACED` = number of moths of a specific morph glued to trees at that location
   - `REMOVED` = number of moths of a specific morph removed after 24 hours

   a. What are logits in this study?
   b. Create an empirical logit plot of logits vs. distance by morph. What can we conclude from this plot?
   c. Create a model with `DISTANCE` and `dark`. Interpret all the coefficients.
   d. Create a model with `DISTANCE`, `dark`, and the interaction between both variables. Interpret all the coefficients.
   e. Interpret a drop-in-deviance test and a Wald test to test the significance of the interaction term in (d).
   f. Test the goodness-of-fit for the interaction model. What can we conclude about this model?
   g. Is there evidence of overdispersion in the interaction model? What factors might lead to overdispersion in this case? Regardless of your answer, repeat (d) adjusting for overdispersion.
   h. Compare confidence intervals for coefficients in your models from (g) and (d).
   i. What happens if we expand the data set to contain one row per moth (968 rows)? Now we can run a logistic binary regression model. How does the logistic binary regression model compare to the binomial regression model? What are similarities and differences? Would there be any reason to run a logistic binomial regression rather than a logistic

binary regression in a case like this? Some starter code can be found below the input code.

```
moth <- read_csv("data/moth.csv")
moth <- mutate(moth,
               notremoved = PLACED - REMOVED,
               logit1 = log(REMOVED / notremoved),
               prop1 = REMOVED / PLACED,
               dark = ifelse(MORPH=="dark",1,0) )


mtemp1 = rep(moth$dark[1],moth$REMOVED[1])
dtemp1 = rep(moth$DISTANCE[1],moth$REMOVED[1])
rtemp1 = rep(1,moth$REMOVED[1])
mtemp1 = c(mtemp1,rep(moth$dark[1],
                      moth$PLACED[1]-moth$REMOVED[1]))
dtemp1 = c(dtemp1,rep(moth$DISTANCE[1],
                      moth$PLACED[1]-moth$REMOVED[1]))
rtemp1 = c(rtemp1,rep(0,moth$PLACED[1]-moth$REMOVED[1]))
for(i in 2:14)  {
  mtemp1 = c(mtemp1,rep(moth$dark[i],moth$REMOVED[i]))
  dtemp1 = c(dtemp1,rep(moth$DISTANCE[i],moth$REMOVED[i]))
  rtemp1 = c(rtemp1,rep(1,moth$REMOVED[i]))
  mtemp1 = c(mtemp1,rep(moth$dark[i],
                        moth$PLACED[i]-moth$REMOVED[i]))
  dtemp1 = c(dtemp1,rep(moth$DISTANCE[i],
                        moth$PLACED[i]-moth$REMOVED[i]))
  rtemp1 = c(rtemp1,rep(0,moth$PLACED[i]-moth$REMOVED[i]))  }
newdata = data.frame(removed=rtemp1,dark=mtemp1,dist=dtemp1)
newdata[1:25,]
cdplot(as.factor(rtemp1)~dtemp1)
```

4. **Birdkeeping and lung cancer: a retrospective observational study.** A 1972-1981 health survey in The Hague, Netherlands, discovered an association between keeping pet birds and increased risk of lung cancer. To investigate birdkeeping as a risk factor, researchers conducted a case-control study of patients in 1985 at four hospitals in The Hague. They

identified 49 cases of lung cancer among patients who were registered with a general practice, who were age 65 or younger, and who had resided in the city since 1965. Each patient (case) with cancer was matched with two control subjects (without cancer) by age and sex. Further details can be found in Holst, Kromhout, and Brand (1988).

Age, sex, and smoking history are all known to be associated with lung cancer incidence. Thus, researchers wished to determine after age, sex, socioeconomic status, and smoking have been controlled for, is an additional risk associated with birdkeeping? The data (Ramsey and Schafer 2002) is found in `birdkeeping.csv` , and the variables are listed below. In addition, R code at the end of the problem can be used to input the data and create additional useful variables.

- `female` = sex (1 = Female, 0 = Male)
- `age` = age, in years
- `highstatus` = socioeconomic status (1 = High, 0 = Low), determined by the occupation of the household’s primary wage earner
- `yrsmoke` = years of smoking prior to diagnosis or examination
- `cigsday` = average rate of smoking, in cigarettes per day
- `bird` = indicator of birdkeeping (1 = Yes, 0 = No), determined by whether or not there were caged birds in the home for more than 6 consecutive months from 5 to 14 years before diagnosis (cases) or examination (controls)
- `cancer` = indicator of lung cancer diagnosis (1 = Cancer, 0 = No Cancer)

a. Perform an exploratory data analysis to see how each explanatory variable is related to the response (`cancer`). Summarize each relationship in one sentence.
   - For quantitative explanatory variables (`age` , `yrsmoke` , `cigsday`), produce a cdplot, a boxplot, and summary statistics by cancer diagnosis.
   - For categorical explanatory variables (`female` or `sex` , `highstatus` or `socioecon_status` , `bird` or `keep_bird`), produce a segmented bar chart and an appropriate table of proportions showing the relationship with cancer diagnosis.

b. In (a), you should have found no relationship between whether or not a patient develops lung cancer and either their age or sex. Why might this be? What implications will this have on your modeling?

c. Based on a two-way table with keeping birds and developing lung cancer from (a), find an unadjusted odds ratio comparing birdkeepers to non-birdkeepers and interpret this odds ratio in context. (Note: an *unadjusted* odds ratio is found by *not* controlling for any other variables.) Also, find an analogous relative risk and interpret it in context as well.

d. Are the elogits reasonably linear relating number of years smoked to the estimated log odds of developing lung cancer? Demonstrate with an appropriate plot.

e. Does there appear to be an interaction between number of years smoked and whether the subject keeps a bird? Demonstrate with an interaction plot and a coded scatterplot with empirical logits on the y-axis.

Before answering the next questions, fit logistic regression models in R with `cancer` as the response and the following sets of explanatory variables:

- `model1` = `age` , `yrsmoke` , `cigsday` , `female` , `highstatus` , `bird`
- `model2` = `yrsmoke` , `cigsday` , `highstatus` , `bird`
- `model4` = `yrsmoke` , `bird`
- `model5` = the complete second order version of `model4` (add squared terms and an interaction)
- `model6` = `yrsmoke` , `bird` , `yrsmoke:bird`

f. Is there evidence that we can remove `age` and `female` from our model? Perform an appropriate test comparing `model1` to `model2` ; give a test statistic and p-value, and state a conclusion in context.

g. Is there evidence that the complete second order version of `model4` improves its performance? Perform an appropriate test comparing `model4` to `model5` ; give a test statistic and p-value, and state a conclusion in context.

h. Carefully interpret each of the four model coefficients in `model6` in context.

i. If you replaced `yrsmoke` everywhere it appears in `model6` with a mean-centered version of `yrsmoke` , tell what would change among these elements: the 4 coefficients, the 4 p-values for coefficients, and the residual deviance.

j. `model4` is a potential final model based on this set of explanatory variables. Find and carefully interpret 95% confidence intervals based on profile likelihoods for the coefficients of `yrsmoke` and `bird` .

k. How does the adjusted odds ratio for birdkeeping from `model4` compare with the unadjusted odds ratio you found in (c)? Is birdkeeping associated with a significant increase in the odds of developing lung cancer, even after adjusting for other factors?

l. Use the categorical variable `years_factor` based on `yrsmoke` and replace `yrsmoke` in `model4` with your new variable to create `model4a` . First, interpret the coefficient for `years_factorOver 25 years` in context. Then tell if you prefer `model4` with years smoked as a numeric predictor or `model4a` with years smoked as a categorical predictor, and explain your reasoning.

m. Discuss the scope of inference in this study. Can we generalize our findings beyond the subjects in this study? Can we conclude that birdkeeping causes increased odds of developing lung cancer? Do you have other concerns with this study design or the analysis you carried out?

n. Read the article that appeared in the *British Medical Journal*. What similarities and differences do you see between their analyses and yours? What are a couple of things you learned from the article that weren't apparent in the short summary at the beginning of the assignment.

```
birds <- read_csv("data/birdkeeping.csv") %>%
  mutate(sex = ifelse(female == 1, "Female", "Male"),
         socioecon_status = ifelse(highstatus == 1,
                                   "High", "Low"),
         keep_bird = ifelse(bird == 1, "Keep Bird", "No Bird"),
         lung_cancer = ifelse(cancer == 1, "Cancer",
                              "No Cancer")) %>%
  mutate(years_factor = cut(yrsmoke,
                            breaks = c(-Inf, 0, 25, Inf),
            labels = c("No smoking", "1-25 years",
                       "Over 25 years")))
```

5. **2016 Election.** An award-winning student project (Blakeman, Renier, and Shandaq 2018) examined driving forces behind Donald Trump's surprising victory in the 2016 Presidential Election, using data from nearly 40,000 voters collected as part of the 2016 Cooperative Congressional Election Survey (CCES). The student researchers investigated two theories: (1) Trump was seen as the candidate of change for voters experiencing economic hardship, and (2) Trump exploited voter fears about immigrants and minorities.

The data set `electiondata.csv` has individual level data on voters in the 2016 Presidential Election, collected from the CCES and subsequently tidied. We will focus on the following variables:

- `Vote` = 1 if Trump; 0 if another candidate
- `zfaminc` = family income expressed as a z-score (number of standard deviations above or below the mean)
- `zmedinc` = state median income expressed as a z-score
- `EconWorse` = 1 if the voter believed the economy had gotten worse in the past 4 years; 0 otherwise
- `EducStatus` = 1 if the voter had at least a bachelor's degree; 0 otherwise
- `republican` = 1 if the voter identified as Republican; 0 otherwise
- `Noimmigrants` = 1 if the voter supported at least 1 of 2 anti-immigrant policy statements; 0 if neither
- `propforeign` = proportion foreign born in the state
- `evangelical` = 1 if `pew_bornagain` is 2; otherwise 0

The questions below address Theory 1 (Economic Model). We want to see if there is significant evidence that voting for Trump was associated with family income level and/or with a belief that the economy became worse during the Obama Administration.

a. Create a plot showing the relationship between whether voters voted for Trump and their opinion about the status of the economy. What do you find?
b. Repeat (a) separately for Republicans and non-Republicans. Again describe what you find.
c. Create a plot with one observation per state showing the relationship between a state's median income and the log odds of a resident of that state voting for Trump. What can you conclude from this plot?

Answer (d)-(f) based on `model1a` below:

d. Interpret the coefficient for `zmedinc` in context.
e. Interpret the coefficient for `republican` in context.
f. Interpret the coefficient for `EconWorse:republican` in context. What does this allow us to conclude about Theory 1?
g. Repeat the above process for Theory 2 (Immigration Model). That is, produce meaningful exploratory plots, fit a model to the data with special emphasis on `Noimmigrants`, interpret meaningful coefficients, and state what can be concluded about Theory 2.

h. Is there any concern about the independence assumption in these models? We will return to this question in later chapters.

```
model1a <- glm(Vote01 ~ zfaminc + zmedinc + EconWorse +
    EducStatus + republican + EducStatus:republican +
    EconWorse:zfaminc + EconWorse:republican,
    family = binomial, data = electiondata)
summary(model1a)
```

6.8.3 Open-Ended Exercises

1. **2008 Presidential voting in Minnesota counties.** Data in `mn08.csv` contains results from the 2008 U.S. Presidential Election by county in Minnesota, focusing on the two primary candidates (Democrat Barack Obama and Republican John McCain). You can consider the response to be either the percent of Obama votes in a county (binomial) or whether or not Obama had more votes than McCain (binary). Then build a model for your response using county-level predictors listed below. Interpret the results of your model.

    - `County` = county name
    - `Obama` = total votes for Obama
    - `McCain` = total votes for McCain
    - `pct_Obama` = percent of votes for Obama
    - `pct_rural` = percent of county who live in a rural setting
    - `medHHinc` = median household income
    - `unemp_rate` = unemployment rate
    - `pct_poverty` = percent living below the poverty line
    - `medAge2007` = median age in 2007
    - `medAge2000` = median age in 2000
    - `Gini_Index` = measure of income disparity in a county
    - `pct_native` = percent of native born residents

2. **Crime on campus.** The data set `c_data2.csv` contains statistics on violent crimes and property crimes for a sample of 81 U.S. colleges and universities. Characterize rates of violent crimes as a proportion of total crimes reported (i.e., `num_viol` / `total_crime`). Do they differ based on type of institution, size of institution, or region of the country?

    - `Enrollment` = number of students enrolled

   - `type` = university (U) or college (C)
   - `num_viol` = number of violent crimes reported
   - `num_prop` = number of property crimes reported
   - `viol_rate_10000` = violent crime rate per 10,000 students enrolled
   - `prop_rate_10000` = property crime rate per 10,000 students enrolled
   - `total_crime` = total crimes reported (property and violent)
   - `region` = region of the country

3. **NBA data.** Data in `NBA1718team.csv` (Kaggle 2018b) looks at factors that are associated with a professional basketball team's winning percentage in the 2017-18 season. After thorough exploratory data analyses, create the best model you can to predict a team's winning percentage; be careful of collinearity between the covariates. Based on your EDA and modeling, describe the factors that seem to explain a team's success.

   - `win_pct` = Percentage of Wins,
   - `FT_pct` = Average Free Throw Percentage per game,
   - `TOV` = Average Turnovers per game,
   - `FGA` = Average Field Goal Attempts per game,
   - `FG` = Average Field Goals Made per game,
   - `attempts_3P` = Average 3 Point Attempts per game,
   - `avg_3P_pct` = Average 3 Point Percentage per game,
   - `PTS` = Average Points per game,
   - `OREB` = Average Offensive Rebounds per game,
   - `DREB` = Average Defensive Rebounds per game,
   - `REB` = Average Total Rebounds per game,
   - `AST` = Average Assists per game,
   - `STL` = Average Steals per game,
   - `BLK` = Average Blocks per game,
   - `PF` = Average Fouls per game,
   - `attempts_2P` = Average 2 Point Attempts per game

4. **Trashball.** Great for a rainy day! A fun way to generate overdispersed binomial data. Each student crumbles an 8.5 by 11 inch sheet and tosses it from three prescribed distances ten times each. The response is the number of made baskets out of 10 tosses, keeping track of the distance. Have the class generate and collect potential covariates, and include them in your data set (e.g., years of basketball experience, using a tennis ball instead of a sheet of paper, height). Some sample analysis steps:

a. Create scatterplots of logits vs. continuous predictors (distance, height, shot number, etc.) and boxplots of logit vs. categorical variables (sex, type of ball, etc.). Summarize important trends in one or two sentences.
b. Create a graph with empirical logits vs. distance plotted separately by type of ball. What might you conclude from this plot?
c. Find a binomial model using the variables that you collected. Give a brief discussion on your findings.

References

Bayer, Ben, and Michael Fitzgerald. 2011. “Reconstructing Alabama: Reconstruction Era Demographic and Statistical Research.”

Bishop, J. A. 1972. “An Experimental Study of the Cline of Industrial Melanism in Biston Betularia (L.) (Lepidoptera) Between Urban Liverpool and Rural North Wales.” *Journal of Animal Ecology* 41 (1): 209–43. https://doi.org/10.2307/3513.

Blakeman, Margaret, Tim Renier, and Rami Shandaq. 2018. “Modeling Donald Trump’s Voters in the 2016 Election.”

Centers for Disease Control and Prevention. 2009. “Youth Risk Behavior Survey Data.” http://www.cdc.gov/HealthyYouth/yrbs/index.htm.

Grabe, Shelly, Janet Shibley Hyde, and L. Moniquee Ward. 2008. “The Role of the Media in Body Image Concerns Among Women: A Meta-Analysis of Experimental and Correlational Studies.” *Pyschological Bulletin* 134 (3): 460–76. https://doi.org/10.1037/0033-2909.134.3.460.

Holst, P. A., D. Kromhout, and R. Brand. 1988. “For Debate: Pet Birds as an Independent Risk Factor for Lung Cancer.” *British Medical Journal* 297 (6659): 1319–21. https://doi.org/10.1136/bmj.297.6659.1319.

Kaggle. 2018b. “NBA Enhanced Box Scores and Standings.” https://www.kaggle.com/pablote/nba-enhanced-stats.

Kolata, Gina. 2009. “Picture Emerging on Genetic Risks of Ivf.” *The New York Times*.

Martinsen, M, S Bratland-Sanda, A K Eriksson, and J Sundgot-Borgen. 2009. “Dieting to Win or to Be Thin? A Study of Dieting and Disordered Eating Among Adolescent Elite Athletes and Non-Athlete Controls.” *British Journal of Sports Medicine* 44 (1): 70–76.

http://bjsm.bmj.com/content/44/1/70.

Nafstad, Per, Jorgen A. Hagen, Leif Oie, Per Magnus, and Jouni J. K. Jaakkola. 1999. “Day Care Centers and Respiratory Health.” *Pediatrics* 103 (4): 753–58. http://pediatrics.aappublications.org/content/103/4/753.

Ramsey, Fred, and Daniel Schafer. 2002. *The Statistical Sleuth: A Course in Methods of Data Analysis*. 2nd ed. Boston, Massachusetts: Brooks/Cole Cengage.

Roskes, Marieke, Daniel Sligte, Shaul Shalvi, and Carsten K. W. De Dreu. 2011. “The Right Side? Under Time Pressure, Approach Motivation Leads to Right-Oriented Bias.” *Psychology Science* 22 (11): 1403–7. https://doi.org/10.1177/0956797611418677.