Logistic Regression (Part II)

Rebecca C. Steorts some material from Chapter 6 of Roback and Legler text.

Computing set up

```r
library(tidyverse)
library(tidymodels)
library(knitr)
library(patchwork)
library(viridis)
library(kableExtra)
library(magrittr)
library(gridExtra)

knitr::opts_chunk$set(fig.width = 8,
                      fig.asp = 0.618,
                      fig.retina = 3,
                      dpt = 300,
                      out.width = "90%",
                      fig.align = "center")

ggplot2::theme_set(ggplot2::theme_bw(base_size = 16))
```

Topics

- Case Study of Reconstruction of Alabama

Notes based on Chapter 6 Roback and Legler (2021) unless noted otherwise.

Basics of logistic regression

Bernoulli + Binomial random variables

Logistic regression is used to analyze data with two types of responses:

- **Binary**: These responses take on two values success ($Y = 1$) or failure ($Y = 0$), yes ($Y = 1$) or no ($Y = 0$), etc.

$$P(Y = y) = p^y(1-p)^{1-y} \qquad y = 0, 1$$

- **Binomial**: Number of successes in a Bernoulli process, n independent trials with a constant probability of success p.

$$P(Y = y) = \binom{n}{y} p^y(1-p)^{n-y} \qquad y = 0, 1, \ldots, n$$

In both instances, the goal is to model p the probability of success.

Logistic regression model

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

- The response variable, $\log\left(\frac{p}{1-p}\right)$, is the log(odds) of success, i.e. the logit
- Use the model to calculate the probability of success

$$\hat{p} = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}$$

- When the response is a Bernoulli random variable, the probabilities can be used to classify each observation as a success or failure

Logistic vs linear regression model

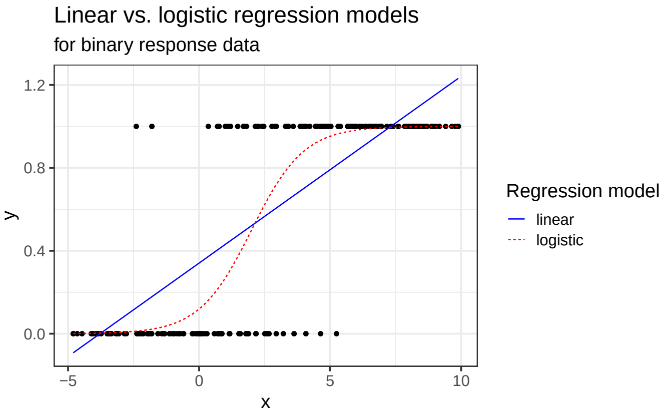

Figure 1: Graph from BMLR Chapter 6

Logit link

Bernoulli and Binomial random variables can be written in one-parameter exponential family form, $f(y;\theta) = e^{[a(y)b(\theta)+c(\theta)+d(y)]}$

Bernoulli

$$f(y;p) = e^{y\log(\frac{p}{1-p})+\log(1-p)}$$

Binomial

$$f(y;n,p) = e^{y\log(\frac{p}{1-p})+n\log(1-p)+\log\binom{n}{y}}$$

Logit link

Bernoulli and Binomial random variables can be written in one-parameter exponential family form, $f(y;\theta) = e^{[a(y)b(\theta)+c(\theta)+d(y)]}$

Bernoulli

$$f(y;p) = e^{y\log(\frac{p}{1-p})+\log(1-p)}$$

Binomial

$$f(y;n,p) = e^{y\log(\frac{p}{1-p})+n\log(1-p)+\log\binom{n}{y}}$$

They have the same canonical link $b(p) = \log\left(\frac{p}{1-p}\right)$

Assumptions for logistic regression

The following assumptions need to be satisfied to use logistic regression to make inferences

1. **Binary response**: The response is dichotomous (has two possible outcomes) or is the sum of dichotomous responses
2. **Independence**: The observations must be independent of one another
3. **Variance structure**: Variance of a binomial random variable is $np(1-p)$ ($n = 1$ for Bernoulli) , so the variability is highest when $p = 0.5$
4. **Linearity**: The log of the odds ratio, $\log\left(\frac{p}{1-p}\right)$, must be a linear function of the predictors $x_1, \ldots, x_p$

Case Study: Reconstructing Alabama

In a paper entitled “Reconstructing Alabama: Reconstruction Era Demographic and Statistical Research,” Ben Bayer performed an analysis of data from 1870 to explain factors that influence voting on referendums related to railroad subsidies (**Bayer2011?**).

Case Study: Reconstructing Alabama

Positive votes are thought to be inversely proportional to the distance a voter is from the proposed railroad.

The racial composition of a community (as measured by the percentage of Black residents) is thought to be associated with voting behavior too.

Goal: Was voting on railroad referenda related to distance from the proposed railroad line and the racial composition of a community (in Hale County)?

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.

- `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

Data Summary

Table 1: 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

EDA

```
## [1] -0.4915505

## [1] 0.8792619

## [1] 0.5834543
```

Scatterplot

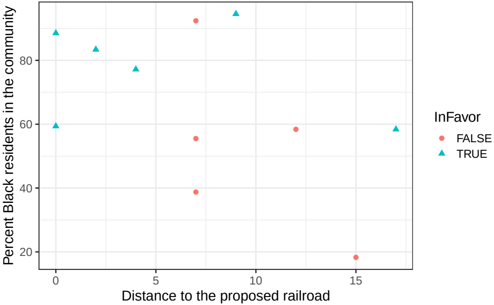

Figure 2: Scatterplot of distance from a proposed rail line and percent Black residents in the community coded by whether the community was in favor of the referendum or not.

Empirical Logit Plots

To assess the linearity assumption, we construct empirical logit plots, where this is based upon the sample.

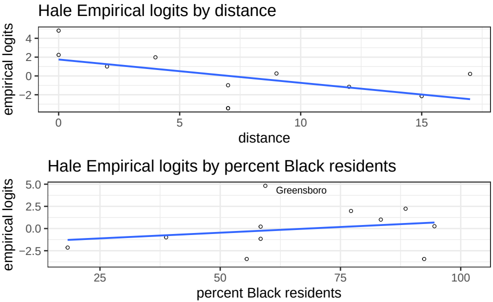

Figure 3: Empirical logit plots for the Railroad Referendum data. The top plot is linear; the bottom plot reveals Greensboro deviates from the linear pattern.

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)  3.3092686 0.11313068  29.25173 4.267877e-188
## distance    -0.2875828 0.01302188 -22.08458 4.446572e-108
```

```
##  Residual deviance =  318.4394  on  9 df
##  Dispersion parameter =  1
```

Initial Model

Our estimated binomial regression model is:

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

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%.

Second Model

The covariate pctBlack is then added to the first model.

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

```
##                Estimate  Std. Error    z value      Pr(>|z|)
## (Intercept)  4.22202114 0.296963480  14.217308  7.155332e-46
## distance    -0.29173451 0.013099945 -22.269903 7.235697e-110
## pctBlack    -0.01322713 0.003896876  -3.394291  6.880665e-04
```

```
##  Residual deviance =  307.2173  on  8 df
##  Dispersion parameter =  1
```

Second Model

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 $= e^{-0.29} = 0.747$).

That is, 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.

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?

We will investigate this using a drop in deviance test.

Recall: the null is the reduced model and the alternative is the full model.

Tests for Significance of Model Coefficients

Our larger model is given by
$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \mathrm{distance}_i + \beta_2 \mathrm{pctBlack}_i.$

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

Tests for Significance of Model Coefficients

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

```
  ResidDF ResidDev Deviance Df         pval
1       9 318.4394       NA NA           NA
2       8 307.2173 11.22207  1 0.0008083041
```

The drop-in-deviance test statistic is $318.44 - 307.22 = 11.22$ on $9 - 8 = 1$ df, producing a p-value of .00081. Thus, we reject the null in favor of the full model.

We find that there is significant evidence that supports for the railroad referendum decreases with higher black residents in the community, after adjusting for the distance a community is from the railroad.

Confidence Intervals for Model Coefficients

We can use the **profile likelihood method**, to find confidence intervals for our model coefficients. (Similar to Section 4.4).

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

```
                2.5 %      97.5 %
(Intercept) 38.2284603 122.6115988
distance     0.7276167   0.7659900
pctBlack     0.9793819   0.9944779
```

In the model with `distance` and `pctBlack`, the profile likelihood 95% confidence interval for e^{β_2} is (0.979, 0.994).

Thus, we can be 95 percent confident that a 1 percent increase in the proportion of black residents is associated with a 0.6 percent to 2.1 percent decrease in the odds of a yes vote for the railroad, after controlling for distance.

Confidence Intervals for Model Coefficients

We can alternatively calculate the confidence interval using the Wald statistic.

The 95% CI for β_2 is

$$\begin{aligned} \hat{\beta}_2 &\pm 1.96 \times \mathrm{SE}(\hat{\beta}_2) && (1) \\ &= -0.0132 \pm 1.96 \times 0.0039 && (2) \\ &= (-0.0208, -0.0056) && (3) \end{aligned}$$

The 95% CI for e^{β_2} is

$$\begin{aligned} &= (e^{-0.0208}, e^{-0.0056}) && (4) \\ &= (0.979, 0.994), && (5) \end{aligned}$$

matching the results for the profile likelihood.

Testing Goodness of Fit

We can conduct a goodness of fit test similar to Poisson models by comparing the residual deviance (307.22) to a χ^2 distribution with $n - p$ degree of freedom (8).

```r
1 - pchisq(307.2173, 8)
```

```
## [1] 0
```

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

Testing Goodness of Fit

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.

Addressing Missing Covariates

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

```
##                      Estimate    Std. Error    z value     Pr(>|z|)
## (Intercept)        7.550901738 0.6383697118  11.828415 2.783488e-32
## distance          -0.614005206 0.0573808237 -10.700530 1.011981e-26
## pctBlack          -0.064730817 0.0091722561  -7.057240 1.698416e-12
## distance:pctBlack  0.005366531 0.0008983743   5.973603 2.320705e-09

##  Residual deviance =  274.2337  on  7 df
##  Dispersion parameter =  1
```

Drop in Deviance Test

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

```
  ResidDF ResidDev Deviance Df         pval
1       8 307.2173       NA NA           NA
2       7 274.2337 32.98364  1 9.293761e-09
```

Drop in Deviance Test

We have statistically significant evidence (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.

Interaction Model

Our interaction model still exhibits lack-of-fit (residual deviance of 274.23 on just 7 df).

A residual deviance of 274.23 is extremely large for a chi-squared distribution with 7 df. This corresponds to a p-value that is essentially zero — meaning the probability of observing such a large deviance by chance, if the model were correct, is virtually 0.

We will now assess the model potential outliers and overdispersion by examining the model's residuals.

Residuals for Binomial Regression

There are two types of residuals used for Binomial regression:

1. the Pearson residual and
2. the deviance residual

Pearson residual

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 observation i and $\hat{p}_i$ is the estimated probability of success for that same observation.

Deviance residual

A deviance residual is an alternative residual for binomial regression based on the discrepency between the observed values and those estimated using the likelihood.

A deviance residual can be calculated for each observation using

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

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

Residual Deviance

The sum of the individual deviance residual is called the **deviance** or the **residual deviance**.

A smaller deviance is preferred.

For a good model fit, the deviance should follow χ^2_{n-p}

Residual Analysis

We consider a residual analysis of our interaction model by plotting the residuals against the fitted values.

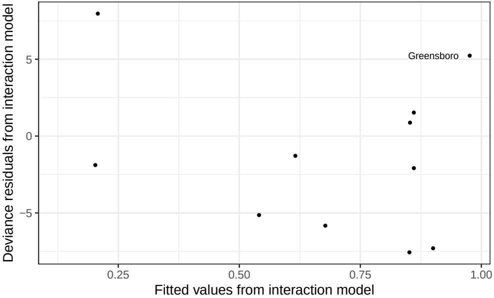

Figure 4: Fitted values by residuals for the interaction model for the Railroad Referendum data.

Residual Analysis

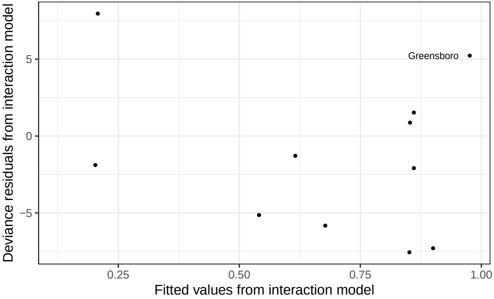

Figure 5: Fitted values by residuals for the interaction model for the Railroad Referendum data.

- Greensboro does not appear to be an outlier.
- It is possible that the binomial counts are overdispersed.

Overdispersion

Similar to Poisson regression, we can adjust for overdispersion in binomial regression.

There is extra-binomial variation, meaning the actual variance will be greater than the variance of the binomial variable $np(1-p)$.

Adjustment for Overdispersion

To adjust for overdispersion, we can estimate a dispersion parameter $\hat{\phi}$ for the variance that will inflate it.

Specifically,

$$\hat{\phi} = \frac{\sum(\text{Pearson residuals})^2}{n - p}$$

and this was the same approach taken in Section 4.9 for Poisson regression.

Adjustment for Overdispersion

When overdispersion is adjusted for in this manner, we cannot use MLE to fit our regression model.

We must use a quasi-likelihood approach, which is similar to likelihood based inference, but because the model uses a dispersion parameter it is no longer a binomial model with a binomial likelihood.

Thus, we call it quasi-binomial; R provides this for fitting the model.

Analysis for Overdispersion

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

Analysis for Overdispersion

```
##                       Estimate  Std. Error    t value  Pr(>|t|)
## (Intercept)        7.550901738 4.585463565  1.6467041 0.1436126
## distance          -0.614005206 0.412171304 -1.4896845 0.1799209
## pctBlack          -0.064730817 0.065885091 -0.9824805 0.3585934
## distance:pctBlack  0.005366531 0.006453098  0.8316208 0.4330721
```

```
##  Residual deviance =  274.2337  on  7 df
##  Dispersion parameter =  51.5967
```

Output adjusting the interaction model for overdispersion, where $\hat{\phi} = 51.6$ is used to adjust the standard errors for the coefficients and drop in deviance test.

Standard errors are inflated by a factor of $\sqrt{(51.6)} = 7.2$.

Observe, there are no significant terms in the model below.

Thus, we remove the interaction terms and refit the model.

Removal of Interaction Term

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

Removal of Interaction Term

```
##                Estimate Std. Error    t value   Pr(>|t|)
## (Intercept)  4.22202114 1.99030675  2.1212917 0.06669098
## distance    -0.29173451 0.08779837 -3.3227783 0.01049658
## pctBlack    -0.01322713 0.02611762 -0.5064447 0.62620393

##  Residual deviance =  307.2173  on  8 df
##  Dispersion parameter =  44.9194
```

By removing the interaction term and using the dispersion parameter, we see:

- distance is significantly associated with referendum support
- percent black is not significantly associated with referendum support (after adjusting for distance)

Estimated coefficients

Quasi-likelihood methods do not change the estimated coefficients.

We still estimate a 25 percent decline $(1 - e^{-0.292})$ in referdendum support for each additional mile from the proposed railroad.

Confidence intervals

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

```
                2.5 %       97.5 %
(Intercept) 1.3608623 5006.7224182
distance    0.6091007    0.8710322
pctBlack    0.9365625    1.0437861
```

Our previous 95% confidence interval for the odds ratio associated with distance was $(0.728, 0.766)$.

Our new 95% confidence interval for the odds ratio associated with a distance is $(0.609, 0.891)$, which is **wider.**

Summary

- We began fitting a logistic regression model with distance, solely.
- We added pctBlack.
- We performed a drop in deviance test, providing strong support for the addition of pctBlack to the model.
- The model with both distance and pctBlack had a large residual deviance suggesting an ill model fit.

Summary (continued)

- We investigated issues with lack of model fit.
- Greensboro was perhaps an outlier, however, models with and without Greensboro were effectively the same.
- To account for the large deviance, we attempted to adjust for overdispersion.
- The final model included distance and pctBlack, although pctBlack was no longer significant after adjust for dispersion.

References

Roback, Paul, and Julie Legler. 2021. *Beyond multiple linear regression: applied generalized linear models and multilevel models in R*. CRC Press.