Chapter 6 Part 2

05/07/2021

Case Study: Reconstructing Alabama

- 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 [@Bayer2011].
- 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.

Case Study: Reconstructing Alabama

- 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?

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 Sample

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

Getting Started

- Lets read the data into R
- Rename some variables
- Create some new percents
- Create a scatterplot of distnace and pctBlack by InFavor

Exploratory Analyses

- We first look at a coded scatterplot to see our data. This 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%.

Two Covariate Model

- 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.

Logit Plots

- 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 $\log\left(\frac{\text{number of successes}}{\text{number of failures}}\right)$.
- What do we want these plots to look like?
- Lets create these in R (two this time)
- Do they look good? Outliers or influential points?

Logit Plots

- 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.

Relationship Between Variables

- 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.
- Lets make the scatterplot in R, and find the correlation between distance and pctBlack
- 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.

Initial Models

- The first model includes only one covariate, distance.
- Let us fit it in R

Initial Models

- 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%.

Second Model

- Add covariate `pctBlack` to the first model.
- Do this in R

Second Model with Distance and PctBlack

- 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.

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 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**.

Tests for Significance of Model Coefficients

- 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.
- Let us do a drop in deviance test in R

Drop in Deviance Result

- 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.

Third Apporach

- 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.
- What two approaches come to mind?

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}.

Confidence Intervals for Model Coefficients

- In this case,

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

- Let's do this in R

Ways to Interpret

- 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.

Profile Likelihood Method

- Another approach available in R is the **profile likelihood method**, similar to Section @ref(cs-philippines).
- This is much easier to do in R...
- 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).

Testing for Goodness-of-Fit

- 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.
- Let us do this i R

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

Testing for 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.
- Let's do this in R

Model Check and Interpretation

- 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.

Interpretation

- Case 1 - `distance` $= 0$; case 2 - `distance` $= 1$

  $$\mathit{logit}(\hat{p}_i) = 7.551 - .6140 \cdot \mathit{distance} - 0.0647 \cdot \mathit{pctBlack} + .0054 \cdot \mathit{distance} \cdot$$

- In the first case, the coefficient for `pctBlack` is -0.0647, while in the second case, the relevant coefficient is $-0.0647 + 15(.00537) = 0.0158$.
- 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
- 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.

How is the LOF on the Interaction Model?

- Lets check our GOF in R
- 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.

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.

Residuals for Binomial Regression

- 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:

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

- 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*.

Deviance

- 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).

Deviance

- *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.
- Let's create a deviance rsidual vs fitted values
- 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.

Residual Comments

- 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.

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, and we will use the same estimate here: $\hat{\phi} = \frac{\sum(\text{Pearson residuals})^2}{n-p}$.

Overdispersion

- 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 $\phi > 1$).

Summary: accounting for overdispersion

- Use the dispersion parameter $\hat{\phi} = \frac{\sum(Pearson residuals)^2}{n-p}$ to inflate standard errors of model coefficients.
- Wald test statistics: multiply the standard errors by $\sqrt{\hat{\phi}}$ so that $SE_Q(\hat{\beta}) = \sqrt{\hat{\phi}} \cdot SE(\hat{\beta})$ and conduct tests using the t-distribution.
- Confidence intervals use the adjusted standard errors and multiplier based on t, so they are thereby wider: $\hat{\beta} \pm t_{n-p} \cdot SE_Q(\hat{\beta})$.

Summary: accounting for overdispersion

- 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.

Summary: accounting for overdispersion

- Lets do this in our example with interaction
- 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.
- We therefore remove the interaction term and refit the model, adjusting for the extra-binomial variation that still exists
- Let us do this in R

New Quasibinomial Model

- 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).
- 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.

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.

Summary

- 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.
- The final model included distance and percent Black residents, although percent Black residents was no longer significant after adjusting for overdispersion.

Linear Least Squares vs. Binomial Regression

Response

LLSR : normal

Binomial Regression : number of successes in n trials

Variance

LLSR : equal for each level of X

Binomial Regression : $np(1-p)$ for each level of X

Model Fitting

LLSR : $\mu = \beta_0 + \beta_1 x$ using Least Squares

Binomial Regression : $\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$ using Maximum Likeli