Logistic regression

Binomial responses + Ordinal logistic models

02.21.22

Click here for PDF of slides

Announcements

- HW 03 **due Mon, Feb 28 at 11:59pm**
    - Released after class
- Mini-Project 01 grades posted in Sakai
    - See Github issues for feedback on written report
- Mid-semester grades by Wednesday (apologies for the delay!)

Learning goals

- Fit and interpret logistic regression model for binomial response variable
- Interpret coefficients and results from an ordinal logistic regression model
- Summarize GLMs for independent observations

Logistic regression for binomial response variable

Data: Supporting railroads in the 1870s

The data set **`RR_Data_Hale.csv`** contains information on support for referendums related to railroad subsidies for 11 communities in Alabama in the 1870s. The data were originally analyzed as part of a thesis project by a student at St. Olaf College. The variables in the data are

- **`pctBlack`**: percentage of Black residents in the county
- **`distance`**: distance the proposed railroad is from the community (in miles)
- **`YesVotes`**: number of "yes" votes in favor of the proposed railroad line
- **`NumVotes`**: number of votes cast in the election

Primary question: Was voting on the railroad referendum related to the distance from the proposed railroad line, after adjusting for the racial composition of a community?

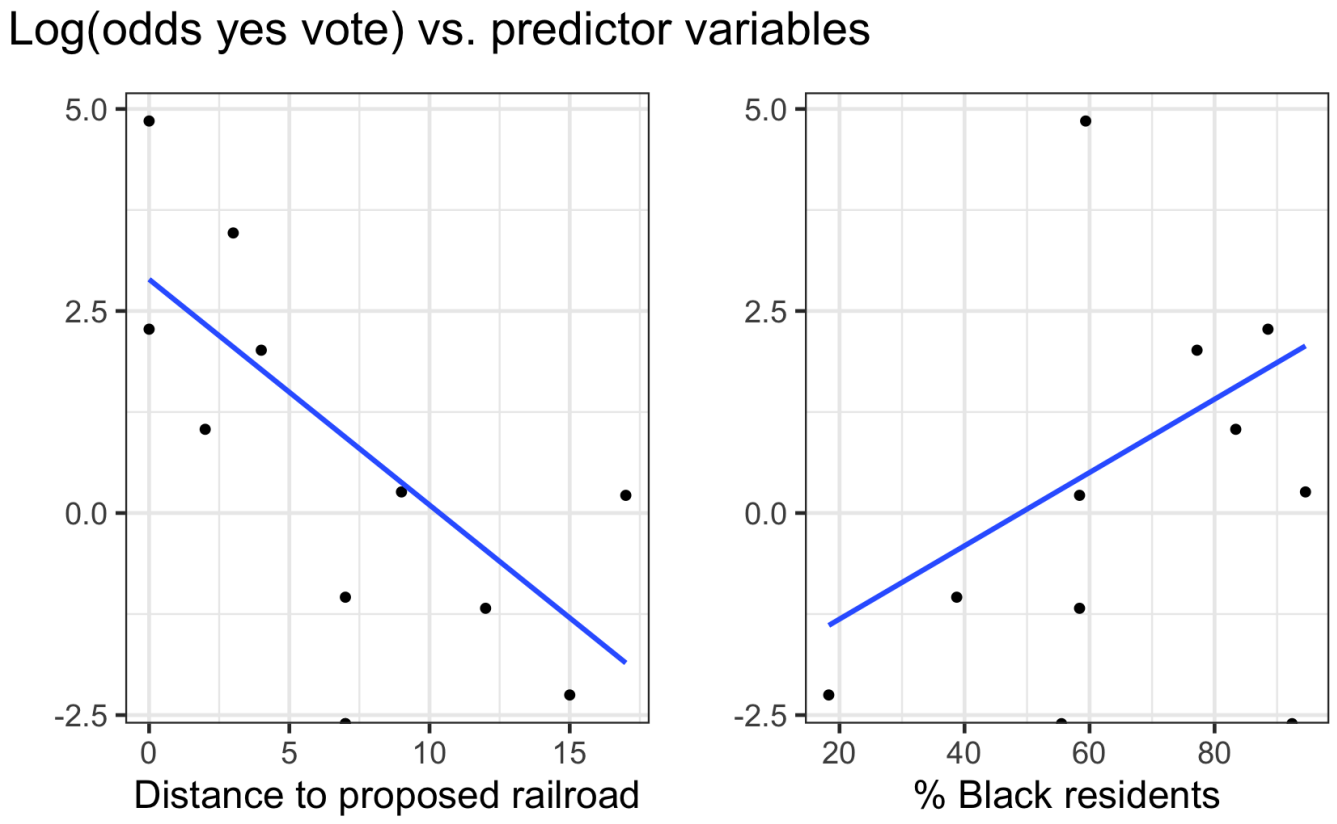
The data

```
rr <- read_csv("data/RR_Data_Hale.csv")
rr %>% slice(1:5) %>% kable()
```

County	popBlack	popWhite	popTotal	pctBlack	distance	YesVotes	NumVotes
Carthage	841	599	1440	58.40	17	61	110
Cederville	1774	146	1920	92.40	7	0	15
Five Mile Creek	140	626	766	18.28	15	4	42
Greensboro	1425	975	2400	59.38	0	1790	1804
Harrison	443	355	798	55.51	7	0	15

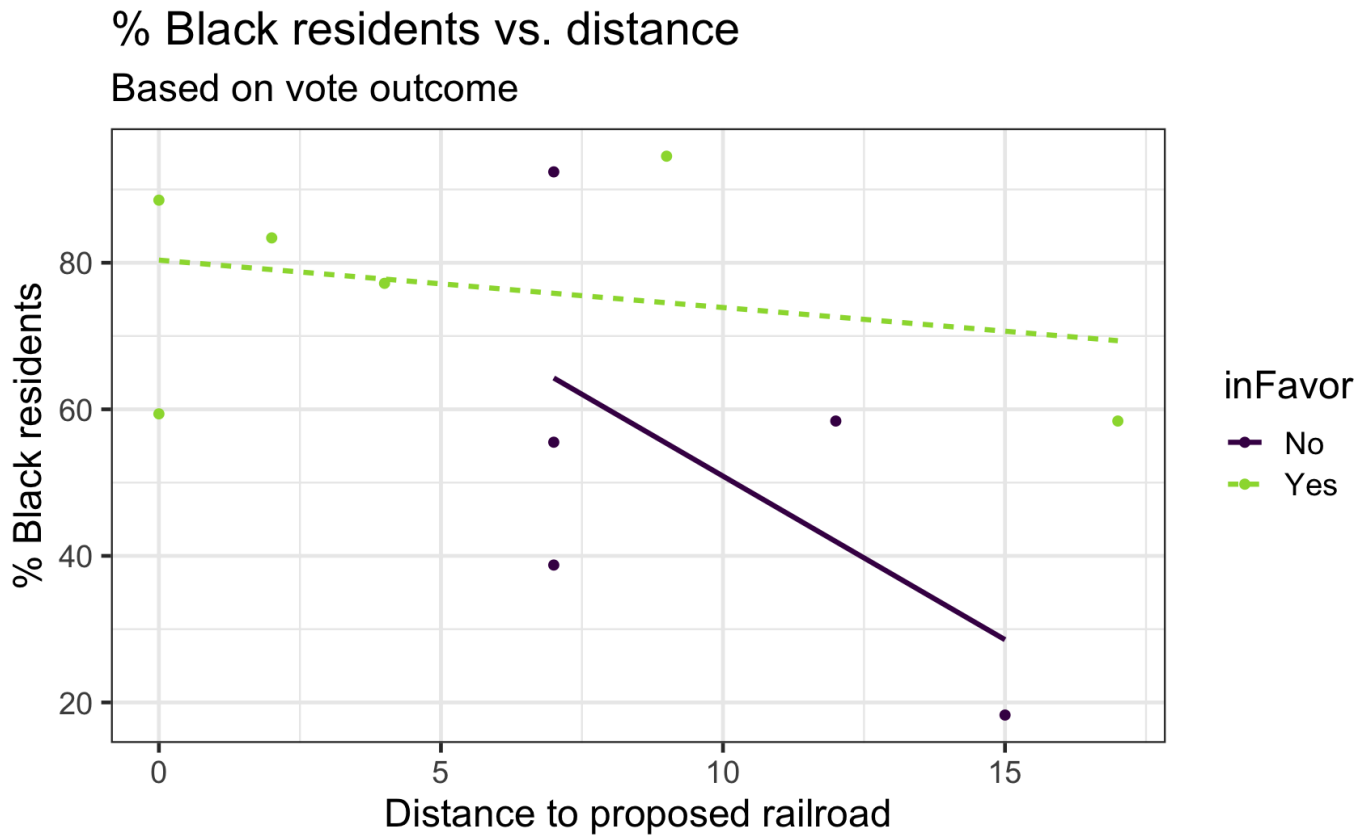
Exploratory data analysis

```
rr <- rr %>%
  mutate(pctYes = YesVotes/NumVotes,
         emp_logit = log(pctYes / (1 - pctYes)))
```



Exploratory data analysis

```
rr <- rr %>%
  mutate(inFavor = if_else(pctYes > 0.5, "Yes", "No"))
```



Model

Let p be the percent of yes votes in a county. We'll start by fitting the following model:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1\, dist + \beta_2\, pctBlack$$

Likelihood

$$\begin{aligned} L(p) &= \prod_{i=1}^{n} \binom{m_i}{y_i} p_i^{y_i} (1-p_i)^{m_i - y_i} \\ &= \prod_{i=1}^{n} \binom{m_i}{y_i} \left[\frac{e^{\beta_0 + \beta_1\, dist_i + \beta_2\, pctBlack_i}}{1 + e^{\beta_0 + \beta_1\, dist_i + \beta_2\, pctBlack_i}}\right]^{y_i} \left[\frac{1}{e^{\beta_0 + \beta_1\, dist_i + \beta_2\, pctBlack_i}}\right]^{m_i - y_i} \end{aligned}$$

Use IWLS to find $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$.

Model in R

```
rr_model <- glm(cbind(YesVotes, NumVotes - YesVotes) ~ distance + pctBlack,
                data = rr, family = binomial)
tidy(rr_model, conf.int = TRUE) %>%
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	4.222	0.297	14.217	0.000	3.644	4.809
distance	-0.292	0.013	-22.270	0.000	-0.318	-0.267
pctBlack	-0.013	0.004	-3.394	0.001	-0.021	-0.006

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = 4.22 - 0.292\, dist - 0.013\, pctBlack$$

See Section 6.5 of *Generalized Linear Models with Examples in R* by Dunn and Smyth (available through Duke library) for details on estimating the standard errors.

Part 1 of `lecture-13.Rmd`

10:00

Residuals

Similar to Poisson regression, there are two types of residuals: Pearson and deviance residuals

Pearson residuals

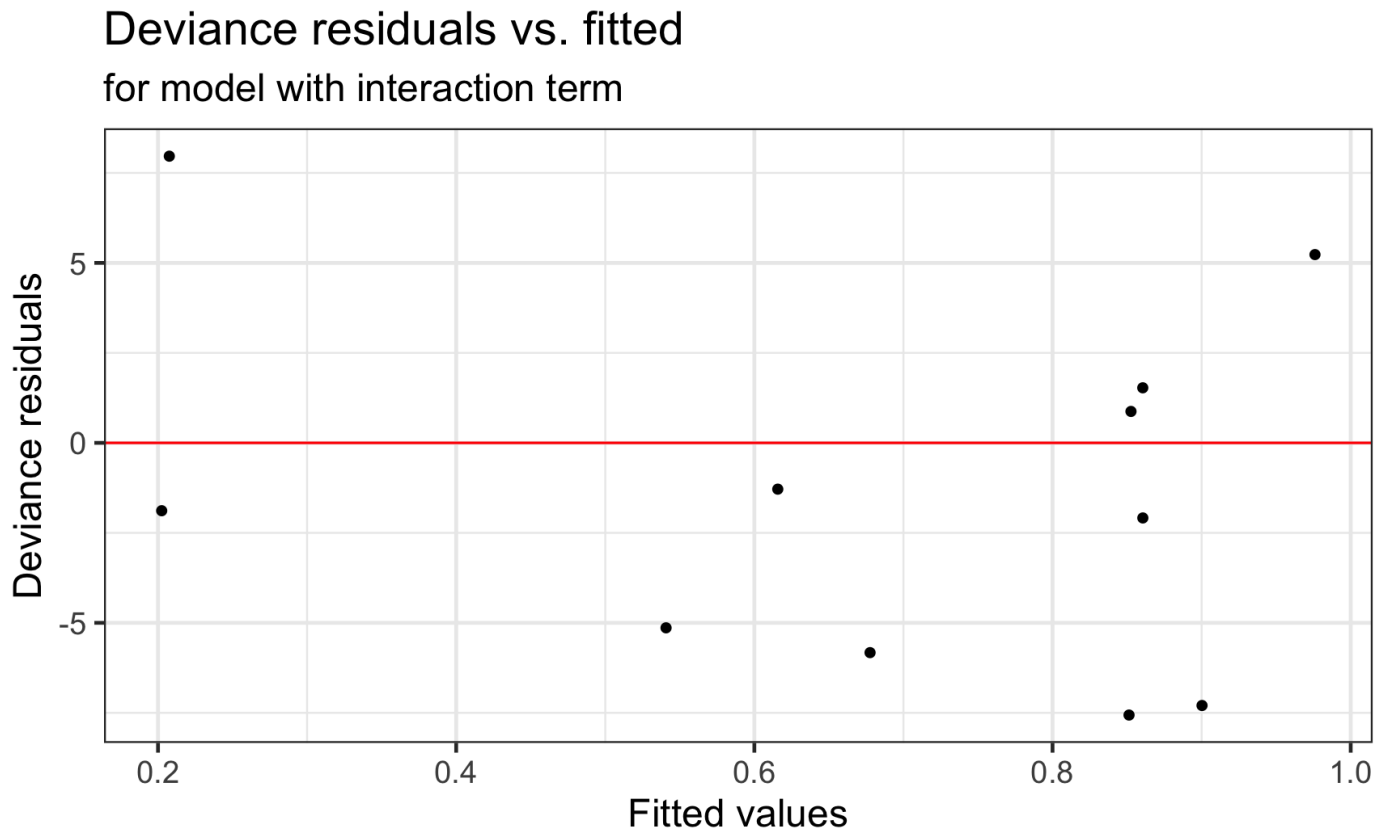
$$\text{Pearson residual}_i = \frac{\text{actual count} - \text{predicted count}}{\text{SD count}} = \frac{Y_i - m_i\hat{p}_i}{\sqrt{m_i\hat{p}_i(1-\hat{p}_i)}}$$

Deviance residuals

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

Plot of deviance residuals

```
rr_int_aug <- augment(rr_int_model, type.predict = "response",
                      type.residuals = "deviance")
```



Goodness of fit

Similar to Poisson regression, the sum of the squared deviance residuals is used to assess goodness of fit.

$$H_0: \text{ Model is a good fit}$$
$$H_a: \text{ Model is not a good fit}$$

- When m_i is large and the model is a good fit $(H_0 \text{ true})$ the residual deviance follows a χ^2 distribution with $n - p$ degrees of freedom.
    - Recall $n - p$ is the residual degrees of freedom.
- If the model fits, we expect the residual deviance to be approximately what value?

Overdispersion

Adjusting for overdispersion

- Overdispersion occurs when there is **extra-binomial variation**, i.e. the variance is greater than what we would expect, $np(1-p)$.

- Similar to Poisson regression, we can adjust for overdispersion in the binomial regression model by using a dispersion parameter

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

  - By multiplying by $\hat{\phi}$, we are accounting for the reduction in information we would expect from independent observations.

Adjusting for overdispersion

- We adjust for overdispersion using a **quasibinomial** model.
    - "Quasi" reflects the fact we are no longer using a binomial model with true likelihood.
- The standard errors of the coefficients are $SE_Q(\hat{\beta}_j) = \sqrt{\hat{\phi}} SE(\hat{\beta})$
    - Inference is done using the t distribution to account for extra variability

Part 2 of `lecture-13.Rmd`

Predicting emergency department wait and treatment times

Predicting ED wait and treatment times

Ataman and Sariyer (2021) use ordinal logistic regression to predict patient wait and treatment times in an emergency department (ED). The goal is to identify relevant factors that can be used to inform recommendations for reducing wait and treatment times, thus improving the quality of care in the ED.

Data: Daily records for ED arrivals in August 2018 at a public hospital in Izmir, Turkey.

Response variable: Wait time, a categorical variable with three levels:

- Patients who wait less than 10 minutes
- Patients whose waiting time is in the range of 10-60 minutes
- Patients who wait more than 60 minutes

Ataman, M. G., & Sarıyer, G. (2021). Predicting waiting and treatment times in emergency departments using ordinal logistic regression models. The American Journal of Emergency Medicine, 46, 45-50.

Ordinal logistic regression

Let Y be an ordinal response variable that takes levels $1, 2, \ldots, J$ with associated probabilities $p_1, p_2, \ldots, p_J$.

The **proportional odds model** can be written as the following:

$$\log\left(\frac{P(Y \leq 1)}{P(Y > 1)}\right) = \beta_{01} + \beta_1 x_1 + \cdots + \beta_p x_p$$

$$\log\left(\frac{P(Y \leq 2)}{P(Y > 2)}\right) = \beta_{02} + \beta_1 x_1 + \cdots + \beta_p x_p$$

$$\ldots$$

$$\log\left(\frac{P(Y \leq J)}{P(Y > J)}\right) = \beta_{0J} + \beta_1 x_1 + \cdots + \beta_p x_p$$

- How is the proportional odds model similar to the multinomial logistic model?
- How is it different?

03:00

Effect of arrival mode

Table 5
OLR models results

Input variable	OLR model for waiting time			
	Parameter estimate	*p*-value	95% confidence interval	
			Lower bound	Upper bound
Gender	−0.022	0.261	−0.061	0.016
Age	−0.116	0.000	−0.154	−0.079
Arrival mode	−3.398	0.000	−3.616	−3.180
Triage level	0.016	0.153	−0.006	0.037
ICD-10 diagnosis	−0.067	0.000	−0.071	−0.063
	Model fitting information: Chi-square = 3740.277; *p*-value: 0.000			
	Model summary: Cox & Snell R square = 0.194; Nagelkerke R square = 0.207			

The variable **`arrival mode`** takes two categories: ambulance and walk-in. Describe the effect of arrival mode in this model. Note that the baseline level is "walk-in".

Effect of triage level

Consider the full output with the ordinal logistic models for wait and treatment times.

Table 5
OLR models results

Input variable	OLR model for waiting time				OLR model for treatment time			
	Parameter estimate	*p*-value	95% confidence interval		Parameter estimate	*p*-value	95% confidence interval	
			Lower bound	Upper bound			Lower bound	Upper bound
Gender	−0.022	0.261	−0.061	0.016	0.041	0.056	−0.001	0.084
Age	−0.116	0.000	−0.154	−0.079	0.151	0.000	0.111	0.190
Arrival mode	−3.398	0.000	−3.616	−3.180	1.215	0.000	1.095	1.335
Triage level	0.016	0.153	−0.006	0.037	−0.950	0.000	−0.973	−0.926
ICD-10 diagnosis	−0.067	0.000	−0.071	−0.063	0.054	0.000	0.049	0.058
	Model fitting information: Chi-square = 3740.277; *p*-value: 0.000 Model summary: Cox & Snell R square = 0.194; Nagelkerke R square = 0.207				Model fitting information: Chi-square = 10,504.755; *p*-value: 0.000 Model summary: Cox & Snell R square = 0.343; Nagelkerke R square = 0.382			

Use the results from both models to describe the effect of triage level (red = urgent, green = non-urgent) on the wait and treatment times in the ED. Note that "green" is the baseline level. Is this what you expected?

02:00

Wrap up GLM for independent observations

Wrap up

- Covered fitting, interpreting, and drawing conclusions from GLMs
    - Looked at Poisson, Negative Binomial, and Logistic (binary, binomial, ordinal) in detail
- Used Pearson and deviance residuals to assess model fit and determine if new variables should be added to the model
- Addressed issues of overdispersion and zero-inflation
- Used the properties of the one-parameter exponential family to identify the best link function for any GLM

Everything we've done thus far as been under the assumption that the observations are *independent*. Looking ahead we will consider models for data with **dependent (correlated) observations**.

Acknowledgements

These slides are based on content in

- BMLR: Chapter 6 - Logistic Regression