Stat 539 Homework 4

Kenny Flagg

February 21, 2017

1. *Agresti Exercise 5.17 (p. 196-7) Use the following toy data to illustrate comments in Section 5.5 about grouped versus ungrouped binary data in the effect on the deviance:*

x	Number of Trials	Number of Successes
0	4	1
1	4	2
2	4	4

Denote by M_0 the null model $\text{logit}(\pi_i) = \beta_0$ *and by M_1 the model* $\text{logit}(\pi_i) = \beta_0 + \beta_1 x_i$.

(a) *Create a data file in two ways, entering the data as (i) ungrouped data: $n_i = 1$, $i = 1, \ldots, 12$, (ii) grouped data: $n_i = 4$, $i = 1, 2, 3$. Fit M_0 and M_1 for each data file. Show that the deviances for M_0 and M_1 differ for the two forms of data entry. Why is this?*

Please see page 10 for my R code to create the datasets and fit the models. The deviances differ because these are different models fit to dfferent datasets wth different numbers of observations.

	M_0 Deviance	M_1 Deviance	Difference
Grouped	6.2568	0.9844	5.2724
Ungrouped	16.3006	11.0283	5.2724

(b) *Show that the difference between the deviances for M_0 and M_1 is the same for each form of data entry. Why is this? (Thus, the data file format does not matter for inference, but it does matter for goodness-of-fit testing.)*

The differences in deviance are the same because, even though the datasets differ, they are equivalent ways to present the same information. Thus the estimated proportion of successes $\hat{\pi}_i$ for each value of x_i will be equal for both models, and thus the estimated model coefficients will also be equal so M_1 should have the same amount of improvement over M_0 for both data formats. Let $\hat{\beta}_0^*$ denote the coefficient estimate for M_0 and let $\hat{\beta}_0$, $\hat{\beta}_1$ be the estimates for M_1.

For the ungrouped data, note that $w_i = n_i = 1$ for $i = 1, \ldots, 12$, so the deviances are

$$D_0 = 2\sum_{i=1}^{12} y_i \left[\log(y_i) - \hat{\beta}_0^* + \log\left(1+\exp\left(\hat{\beta}_0^*\right)\right)\right] + 2\sum_{i=1}^{12}(1-y_i)\left[\log(1-y_i) + \log\left(1+\exp\left(\hat{\beta}_0^*\right)\right)\right]$$

and

$$D_1 = 2\sum_{i=1}^{12} y_i \left[\log(y_i) - \hat{\beta}_0 - \hat{\beta}_1 x_i + \log\left(1+\exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right)\right] + 2\sum_{i=1}^{12}(1-y_i)\left[\log(1-y_i) + \log\left(1+\exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right)\right]$$

so then

$$D_0 - D_1 = 2\sum_{i=1}^{12} y_i \left[\hat{\beta}_0 + \hat{\beta}_1 x_i - \hat{\beta}_0^* + \log\left(\frac{1+\exp\left(\hat{\beta}_0^*\right)}{1+\exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)}\right)\right] + 2\sum_{i=1}^{12}(1-y_i)\log\left(\frac{1+\exp\left(\hat{\beta}_0^*\right)}{1+\exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)}\right).$$

Denote the grouped response as $y_j^{(g)}$ and note that $w_j^{(g)} = n_j^{(g)} = 4$ for $j = 1, 2, 3$. The deviances are

$$D_0^{(g)} = 2\sum_{j=1}^{3} n_j^{(g)} y_j^{(g)} \left[\log\left(y_j^{(g)}\right) - \hat{\beta}_0^* + \log\left(1+\exp\left(\hat{\beta}_0^*\right)\right)\right] + 2\sum_{j=1}^{3} n_j^{(g)}\left(1-y_j^{(g)}\right)\left[\log\left(1-y_j^{(g)}\right) + \log\left(1+\exp\left(\hat{\beta}_0^*\right)\right)\right]$$

and

$$D_1^{(g)} = 2\sum_{j=1}^{3} n_j^{(g)} y_j^{(g)} \left[\log\left(y_j^{(g)}\right) - \hat{\beta}_0 - \hat{\beta}_1 x_j + \log\left(1+\exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_j\right)\right)\right] + 2\sum_{j=1}^{3} n_j^{(g)}\left(1-y_j^{(g)}\right)\left[\log\left(1-y_j^{(g)}\right) + \log\left(1+\exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_j\right)\right)\right]$$

so then

$$D_0^{(g)} - D_1^{(g)} = 2\sum_{j=1}^{3} n_j^{(g)} y_j \left[\hat{\beta}_0 + \hat{\beta}_1 x_j - \hat{\beta}_0^* + \log\left(\frac{1+\exp\left(\hat{\beta}_0^*\right)}{1+\exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_j\right)}\right)\right] + 2\sum_{j=1}^{3} n_j^{(g)}\left(1-y_j^{(g)}\right)\log\left(\frac{1+\exp\left(\hat{\beta}_0^*\right)}{1+\exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_j\right)}\right).$$

Therefore, $D_0 - D_1 = D_0^{(g)} - D_1^{(g)}$ because $n_j^{(j)} y_j^{(g)} = \sum_{i:x_i = x_j} y_i$.

2. *Agresti Exercise 5.30 (p. 199) In one of the first studies of the link between lung cancer and smoking, Richard Doll and Austin Bradford Hill collected data from 20 hostpitals in London, England. Each patient admitted with lung cancer in the preceeding year was queried about their smoking behavior. For each of the 709 patients admitted, they recorded the smoking behavior of a noncancer patient at the same hospital of the same gender and within the same 5-year grouping on age. A smoker was defined as a person who had smoked at least one cigarette a day for at least a year. Of the 709 cases having lung cancer, 688 reported being smokers. Of the 709 controls, 650 reported being smokers. Specify a relevant logistic regression model, explain what can be estimated and what cannot, and conduct a statistical analysis.*

   This is a case-control study with observed counts as shown in the table below.

	Lung Cancer	
Smoker	Yes	No
Yes	688	650
No	21	59

   A logistic regression model with cancer as the response and smoking as the predictor is

$$\begin{aligned} n_i y_i &\sim \text{Binomial}(n_i, \pi_i), \\ \text{logit}(\pi_i) &= \beta_0 + \beta_1 x_i \end{aligned}$$

   where y_i is the sample proportion of people in the ith group who have lung cancer, $i = 1$ for nonsmokers and $i = 2$ for smokers, n_i is the size of the ith group, π_i is the probability that an individual in the ith group gets lung cancer, and x_i is a smoking indicator variable with $x_1 = 0$ for the nonsmoker group and $x_2 = 1$ for the smoker group.

   Because the numbers of people with and without cancer are fixed, we cannot estimate the population probability of having lung cancer. (In the estimated model, $\exp(\hat{\beta}_0)$ is the proportion of nonsmokers in the *sample* who have lung cancer.) However, we can estimate the probability of being a smoker conditional on cancer status, and we can estimate the odds ratio for smoking status between people with and without cancer, which is equal to the odds ratio for cancer between smokers and nonsmokers ($\exp(\hat{\beta}_1)$).

```
prob2 <- data.frame(Cancer = c(688, 21), No_Cancer = c(650, 59), Smoker = c(1, 0))
prob2_fit <- glm(cbind(Cancer, No_Cancer) ~ Smoker, data = prob2, family = binomial)
xtable(summary(prob2_fit)$coefficients, digits = 5)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.03302	0.25410	-4.06534	0.00005
Smoker	1.08983	0.25992	4.19290	0.00003

   With a Wald p-value of 0.00003, we have very strong evidence that the odds of lung cancer differ between smokers and nonsmokers.

3. *Using the data given in Agresti Exercise 5.35 (p. 201), answer the following questions:*

   (a) *Enter the data into R in grouped format. Show the R code used to create the data frame and the first few rows.*

```
prob3 <- data.frame(AZT = c(1, 1, 0, 0), White = c(0, 1, 0, 1),
                    AIDS = c(11, 14, 12, 32), No_AIDS = c(52, 93, 43, 81))
print(prob3)
```

```
  AZT White AIDS No_AIDS
1   1     0   11      52
2   1     1   14      93
3   0     0   12      43
4   0     1   32      81
```

   (b) *Calculate and interpret the sample marginal odds ratio of developing AIDS for those taking AZT immediately to those who wait until their T cells showed severe immune weakness.*

   The table below shows the marginal counts.

	AIDS	No AIDS
AZT Immediately	25	145
AZT Later	44	124

   Then the estimated odds ratio is

$$\widehat{OR} = \frac{(25/145)}{(44/124)} = 0.486.$$

   For people in this sample, the odds of getting AIDS for those who recieved AZT immediately are 51.4% lower than the odds of getting AIDS for those who waited to get AZT.

   (c) *Calculate and interpret the sample conditional odds ratio of developing AIDS for those taking AZT immediately to those who wait until their T cells showed severe immune weakness, for black subjects.*

   The contingency table for blacks is

	AIDS	No AIDS
AZT Immediately	11	52
AZT Later	12	43

   so the estimated odds ratio is

$$\widehat{OR}_{\text{black}} = \frac{(11/52)}{(12/43)} = 0.758.$$

   For black people in this sample, the odds of getting AIDS for those who recieved AZT immediately are 24.2% lower than the odds of getting AIDS for those who waited to get AZT.

(d) *Do these data exhibit Simpson's Paradox? Why or why not?*

No, these data do not exhibit Simpson's Paradox because, both marginally and conditionally, recieving AZT immediately is associated with lower odds of AIDS.

(e) *Fit a logistic regression model to these data using AZT treatment and race as predictors, with no interaction.*

```
prob3_fit <- glm(cbind(AIDS, No_AIDS) ~ White + AZT, data = prob3, family = binomial)
xtable(summary(prob3_fit)$coefficients, digits = 5)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.07357	0.26294	-4.08295	0.00004
White	0.05548	0.28861	0.19225	0.84755
AZT	-0.71946	0.27898	-2.57890	0.00991

i. *Write the equation of the fitted model, clearly defining any variables and symbols used. Write a sentence interpreting each of the estimated coefficients.*

The estimated model is

$$\begin{aligned} n_i y_i &\sim \text{Binomial}(n_i, \pi_i) \\ \text{logit}(\pi_i) &= -1.07357 + 0.05548 \times \text{White}_i - 0.71946 \times \text{AZT}_i \end{aligned}$$

where

- n_i is the number of people in group i,
- y_i is the proportion of people in group i who developed AIDS,
- π_i is the probability that an individual in group i develops AIDS,
- $\text{White}_2 = \text{White}_4 = 1$ indicate groups of white people and $\text{White}_1 = \text{White}_3 = 0$ indicate groups of black people, and
- $\text{AZT}_1 = \text{AZT}_2 = 1$ indicate groups of people who recieved AZT immediately and $\text{AZT}_3 = \text{AZT}_4 = 0$ indicate groups of people who watied to recieve AZT.

$\hat{\beta}_0 = -1.07357$: For black people who wait to recieve AZT, we estimate that the odds of developing AIDS are $\exp(-1.07357) = 0.342$ to 1.

$\hat{\beta}_1 = 0.05548$: For white people with a given AZT treatment, we estimate that the odds of developing AIDS are $\exp(0.05548 - 1) = 5.71\%$ lower than odds of developing AIDS for black people with the same AZT treatment.

$\hat{\beta}_2 = -0.71946$: For people of a given race who recieve AZT immediately, we estimate that the odds of developing AIDS are $1 - \exp(-0.71946) = 51.3\%$ lower than odds of developing AIDS for people of the same race who wait to recieve AZT.

ii. *Express the conditional odds ratio of developing AIDS for those taking AZT immediately to those who wait until their T cells showed severe immune weakness, for black subjects, in terms of the estimated model coefficients. Calculate and interpret a 95% confidence interval for this quantity. How does your estimate compare to the sample conditional odds ratio found in part (c)?*

The conditional odds ratio estimated by the model is

$$\begin{aligned}\widehat{OR}_{\text{black}} &= \frac{\exp(\hat{\beta}_0 + \hat{\beta}_2)}{\exp(\hat{\beta}_0)} \\ &= \frac{\exp(-1.07357 - 0.71946)}{\exp(-1.07357)} \\ &= \exp(-0.71946) \\ &= 0.48702\end{aligned}$$

with an approximate 95% confidence interval of

$$\exp(-0.71946 \pm 1.96 \times 0.27898) = (0.28189, 0.84141).$$

We are 95% confidence that, for people of a given race who recieved AZT immediately, the odds of developing AIDS are between 15.9% and 71.8% lower than the odds of developing AIDS for people of the same race who waited to recieve AZT.

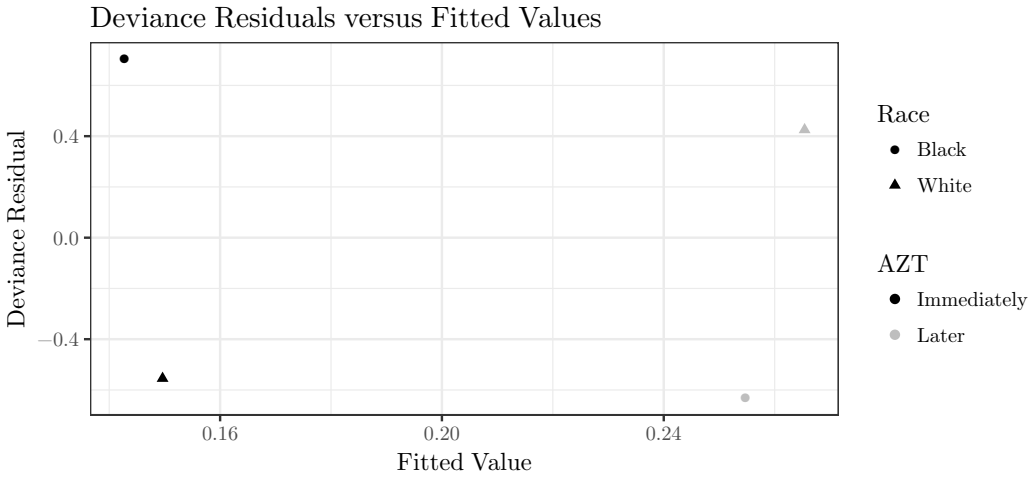
This odds ratio estimate is a bit smaller than the one from part (c), but the estimate from (c) is inside this confidence interval and therefore not unreasonable. Computing the odds ratio from the conditional table is equivalent to estimating the odds ratio from an interaction model, so we should not expect them to be equal. The no-interaction model forces the conditional odds ratio to be the same for both races, and in fact the odds ratio computed from the model is very close to the marginal odds ratio from part (b).

iii. *Use the deviance of this model to perform a goodness of fit test.*

We are testing H_0: the model has an adequate fit against H_a: a more complicated model is needed. Under H_0, the residual deviance follows a χ^2_1 distribution. The residual deviance is 1.384 with a p-value of 0.2395. There is no evidence of an inadequate fit.

iv. *Plot the deviance residuals versus the fitted values. Which covariate groups contribute most to lack of fit? how?*

```
ggplot(prob3 %>% mutate(AZT = ifelse(AZT == 1, 'Immediately', 'Later'),
                        Race = ifelse(White == 1, 'White', 'Black')),
       aes(x = fitted(prob3_fit),
           y = residuals(prob3_fit, type = 'deviance'),
           col = AZT, shape = Race)) +
  geom_point() +
  scale_color_manual(values = c('black', 'grey')) +
  xlab('Fitted Value') +
  ylab('Deviance Residual') +
  ggtitle('Deviance Residuals versus Fitted Values')
```



The largest magnitude deviance residual is for blacks who recieved AZT immediately. This model underestimates the odds of developing AIDS for that group.

4. *Agresti Exercise 5.38 (p. 201). Your analysis should include a short description of how you chose your model with relevant tests, model checking methods including Hosmer-Lemeshow goodness of fit test and residual analysis, interpretation of the coefficients in the final model, and a short paragraph summarizing the results. Your write-up should be no more than two pages. Upload a well-commented .R file of the relevant R code you used for the data analysis process to the "Homework 4 R Code" D2L Assignment folder.*

   The data are from a study of the associants between sore throat outcomes after surgery and the type of device used to secure the airway. A total of 35 patients were included in the sample and had either a tracheal tube or a laryngeal mask airway used during surgery. One patient who recieved a tracheal tube had an unusually long surgery of 135 minutes and will be omitted from the analysis. The remaining 16 patients who recieved tracheal tubes had surgeries lasting from 0 to 90 minutes, with a mean of 41.12 minutes, and 7 of these patients had sore throats upon waking. The 18 patients who had laryngeal mask airways used had surgeries 0 to 95 minutes long, with a mean of 45.33 minutes, and 14 of these patients had sore throats afterwards.

   The observed surgery durations and sore throat outcomes appear in Figure 1 along with smoothed curves to approximate the proportion of patients with a given surgery duration and device who got sore throats. The proportion with a sore throat increases with time and tends to be higher for patients who recieved the laryngeal mask airways. Both of the curves dip downward at some point but this is likely because few patients in the sample had surgeries of those durations rather than being an actual feature of the relationship between the device and the sore throat outcome.

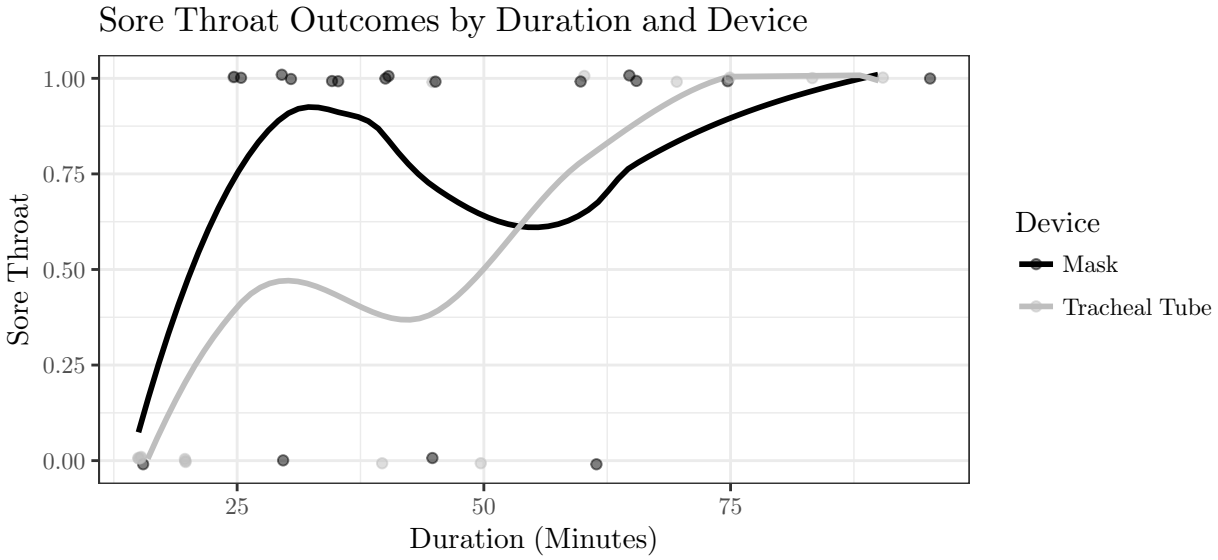


Figure 1: Plot showing the observed sore throat outcomes (points, semitransparent and jittered slightly to reduce overlap) versus the duration of the surgery, colored by device. The curves are local polynomial smoothers approximating the proportion of patients using each device who wake up with a sore throat after a surgery of a given duration.

I begin by fitting the main effects model

$$y_i \sim \text{Binomial}(1, \pi_i)$$
$$\text{logit}(\pi_i) = \beta_0 + \beta_1 d_i + \beta_2 t_i$$

where $y_i = 1$ if patient i had a sore throat and 0 otherwise, π_i is the ith patient's true probability of getting a sore throat, d_i is the duration of patient i's surgery in minutes, and $t_i = 1$ if patient i had a tracheal tube used and 0 otherwise, for $i = 1, 2, \ldots, 34$. The coefficient estimates appear in Table 1. I used the Hosmer-Lemeshow goodness of fit test with four groups to asses model adequacy. Due to the small sample size, the group with the smallest estimated probabilities only had 1.5 expected sore throats, but the other three groups each had expected sore throat counts of at least 4.6. The goodness of fit test statistic is $\chi^2_2 = 1.297$ with a p-value of 0.52293 giving no evidence of a lack of fit.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.41265	1.09850	-1.28599	0.19845
D	0.06854	0.02654	2.58217	0.00982
T	-1.65913	0.92240	-1.79870	0.07207

Table 1: Estimated logistic regression coefficients.

There is very strong evidence of an association between surgery duration and the odds of a sore throat after controlling for the device type (p-value of 0.00982); for patients with given device, the odds of waking with a sore throat are an estimated 7.09% higher the odds of a sore throat for patients with the same type of device and a surgery one minute shorter. There is weak evidence of an association between the device type and the odds of a sore throat after controlling for the sugery duration (p-value of 0.07207); for patients with tracheal tubes, the odds of waking with a sore throat are an estimated 81.0% lower the odds of a sore throat for patients with a laryngeal mask airway and the same surgery duration.

R Code for Problem 1

```
prob1_ungrouped <- data.frame(
  x = rep(0:2, each = 4),
  Success = c(1, 0, 0, 0,  1, 1, 0, 0,  1, 1, 1, 1)
)

m0_ungrouped <- glm(Success ~ 1, data = prob1_ungrouped, family = binomial)
m1_ungrouped <- glm(Success ~ x, data = prob1_ungrouped, family = binomial)

prob1_grouped <- data.frame(
  x = 0:2,
  Trials = rep(4, 3),
  Successes = c(1, 2, 4)
)

m0_grouped <- glm(cbind(Successes, Trials - Successes) ~ 1,
                  data = prob1_grouped, family = binomial)
m1_grouped <- glm(cbind(Successes, Trials - Successes) ~ x,
                  data = prob1_grouped, family = binomial)

prob1_deviances <- data.frame(
  `\\(M_{0}\\) Deviance` = c(
    Grouped = deviance(m0_grouped),
    Ungrouped = deviance(m0_ungrouped)
  ),
  `\\(M_{1}\\) Deviance` = c(
    Grouped = deviance(m1_grouped),
    Ungrouped = deviance(m1_ungrouped)
  ),
  Difference = c(
    Grouped = deviance(m0_grouped) - deviance(m1_grouped),
    Ungrouped = deviance(m0_ungrouped) - deviance(m1_ungrouped)
  ),
  check.names = FALSE
)

xtable(prob1_deviances, digits = 4)
```