Stat 525 Homework 6

Kenny Flagg

October 24, 2016

1. *The data are described in Problems 9.1 on pages 144-145 and the problem continues as Problem 10.3 on page 163. Refer to the specific questions below. R code to load these data into R is attached. Use logistic regression to answer the questions below.*

   (a) *Calculate the ratio of the odds of TB given exposure to biomass fuel over the odds of TB given no exposure for the pooled data. Also give an approximate 95% confidence interval and interpret in terms of the problem. The data you need are in `TB.pool`. Fit a logistic regression model with just biomass as the explanatory variable. Compare the results to those you got on the midterm exam.*

```
TB <- cbind(c(38, 102, 12, 136), c(12, 141, 9, 383))
biomass <- rep(c('Yes', 'No'), 2)
income <- rep(c('<1000', '>=1000'), each = 2)
TB.data <- data.frame(biomass, income, cases = TB[,1], controls = TB[,2])
fit1 <- glm(cbind(cases,controls)˜biomass, family=binomial, data=TB.data)
summary(fit1)$coefficients
```

```
             Estimate Std. Error    z value     Pr(>|z|)
(Intercept) -0.789221 0.07816699 -10.096603 5.718904e-24
biomassYes   1.656722 0.27153108   6.101407 1.051389e-09
```

```
confint.default(fit1)
```

```
                 2.5 %     97.5 %
(Intercept) -0.9424255 -0.6360165
biomassYes   1.1245304  2.1889127
```

```
exp(summary(fit1)$coefficients['biomassYes', 'Estimate'])
```

```
[1] 5.242097
```

```
exp(confint.default(fit1)['biomassYes',])
```

```
   2.5 %   97.5 %
3.078771 8.925503
```

The odds of TB for individuals exposed to biomass are estimated to be 5.242 times larger than the odds of TB for unexposed individuals. We are 95% confident that the true odds ratio is between 3.079 and 8.926. The point estimate and confidence interval are identical to the cross-product odds ratio and its Wald interval.

(b) *Provide me with an estimate of the common odds ratio adjusted for income along with an approximate 95% interval. Fit a logistic regression model with biomass and income as explanatory variables. Compare these results to the those you got using the CMH method on the exam.*

```
fit2 <- glm(cbind(cases,controls)~biomass+income, family=binomial, data=TB.data)
summary(fit2)$coefficients

                Estimate Std. Error   z value     Pr(>|z|)
(Intercept)   -0.3159599  0.1265509 -2.496703 1.253539e-02
biomassYes     1.4184651  0.2783881  5.095279 3.482276e-07
income>=1000 -0.7240439  0.1571051 -4.608660 4.052716e-06

confint.default(fit2)

                   2.5 %      97.5 %
(Intercept)   -0.5639951 -0.06792478
biomassYes     0.8728344  1.96409570
income>=1000 -1.0319642 -0.41612365

exp(summary(fit2)$coefficients['biomassYes', 'Estimate'])

[1] 4.130775

exp(confint.default(fit2)['biomassYes',])

   2.5 %   97.5 %
2.393686 7.128463
```

After adjusting for income level, the odds of TB for individuals exposed to biomass are estimated to be 4.131 times larger than the odds of TB for unexposed individuals. We are 95% confident that the true odds ratio is between 2.394 and 7.128. The point estimate is slightly smaller than the CMH adjusted odds ratio (which was 4.158) and the confidence interval is shifted down accordingly.

(c) *Fit a model assuming an interaction between biomass and income. Use a drop-in-deviance test to assess the evidence for the existence of an interaction. Compare the results to the results you got from the Breslow-Day method on the exam.*

```
fit3 <- glm(cbind(cases,controls)~biomass*income,family=binomial,data=TB.data)
anova(fit2, fit3, test = 'Chisq')

Analysis of Deviance Table

Model 1: cbind(cases, controls) ~ biomass + income
Model 2: cbind(cases, controls) ~ biomass * income
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         1   0.071115
2         0   0.000000  1 0.071115   0.7897
```

With $\chi^2_1 = 0.0711$ and p-value = 0.7897 there is no evidence of an interaction between biomass exposure and income level. This agrees with the Breslow-Day result, which had

$\chi^2_1 = 0.0689$ and p-value $= 0.7930$.

(d) *Summarize the results as you did on the exam.*

There is little to no evidence that the effect of biomass fuel exposure on tuberculosis occurrence differs between individuals with a monthly family income below 1,000 pesos and individuals with monthly family income of at least 1,000 pesos ($\chi^2_1 = 0.0711$, p-value $= 0.7897$). There is evidence that biomass fuel exposure is associated with higher tuberculosis rates; adjusting for income level, the odds of tuberculosis for an exposed individual are estimated to be 4.131 times larger than the odds of TB for unexposed individuals, with a 95% confidence interval of 2.424 to 7.255.

2. *Extra Credit for STAT 425 and required for STAT 525: Above you were able to get results when you fit the model with an interaction in it. Note however that the model fit perfectly (residual deviance of 0 on 0 df). We were even able to compare this model to the simpler additive model with a drop-in-deviance test. I computed the sample proportions, transformed them to logits, and attempted to fit an ordinary least squares (OLS) model involving the interaction. I got the following.*

```
TB.data$p <- with(TB.data, cases/(cases+controls))
TB.data$logit <- with(TB.data, log(p/(1-p)))
summary(lm(logit~biomass*income, data=TB.data))
```

```

Call:
lm(formula = logit ~ biomass * income, data = TB.data)

Residuals:
ALL 4 residuals are 0: no residual degrees of freedom!

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                -0.3238         NA      NA       NA
biomassYes                  1.4765         NA      NA       NA
income>=1000               -0.7116         NA      NA       NA
biomassYes:income>=1000    -0.1534         NA      NA       NA

Residual standard error: NaN on 0 degrees of freedom
Multiple R-squared:      1,Adjusted R-squared:    NaN
F-statistic:   NaN on 3 and 0 DF,  p-value: NA
```

This model did not fit. Why can we get valid results fitting a logistic regression model but not the OLS model? (The answer has nothing to do with the OLS fit being invalid because it is applied to proportions. For example, compare the coefficient estimates from the OLS fit to the logistic regression model fit.).

This OLS model includes as many coefficients as there are observed logits, so the model perfectly fits the observed data. Thus this model has a residual variance of zero, so standard errors cannot be computed. The logistic regression assumes that the variance is a function of the estimated proportion of individuals with TB, so it does not matter that the model leaves no unexplained variation.

3. *We looked at using* `prop.trend.test` *to test for a linear trend in risk. Recall the data were the presence or absence of coronary heart disease with body weight being the risk factor. Weight was binned into 5 ordinal categories. The data were.*

```
chd<-array(c(32,31,50,66,78,558,505,594,501,739),dim=c(5,2),
dimnames=list(c("<150","151-160","161-170","171-180",">180"),
c("CHD","no CHD")) )
# get the number of trials by row
n<-rowSums(chd)
# get the counts
d<-chd[,1]
prop.trend.test(d,n)
```

```
Chi-squared Test for Trend in Proportions

data:  d out of n ,
 using scores: 1 2 3 4 5
X-squared = 15.361, df = 1, p-value = 0.0000888
```

Recall that the test is based on fitting a weighted least squares model with the proportions as responses and the weight categories treated as a quantitative variable with values 1 through 5. We will fit a logistic regression model to these data to test for a linear trend.

```
score<-1:5
fit<-glm(cbind(chd[,1],chd[,2])~score,family=binomial)
summary(fit)
```

```
Call:
glm(formula = cbind(chd[, 1], chd[, 2]) ~ score, family = binomial)

Deviance Residuals:
    <150   151-160   161-170   171-180      >180
-0.10609  -0.72539   0.02465   1.99968  -1.12004

Coefficients:
            Estimate Std. Error z value  Pr(>|z|)
(Intercept) -3.01985    0.17339 -17.417   < 2e-16
score        0.18045    0.04636   3.893 0.0000991

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 21.3981  on 4  degrees of freedom
Residual deviance:  5.7913  on 3  degrees of freedom
AIC: 37.937

Number of Fisher Scoring iterations: 4
```

We talked about the `Residual Deviance` *but not about the* `Null Deviance` *This is the deviance associated with the reduced model that arises if all the regression coefficients are 0. In this case it is the deviance associated with the null hypothesis $H_0 : \beta_1 = 0$. To test for a trend we compare the full model*

$$logit(\pi(x)) = \beta_0 + \beta_1 x$$

to the reduced model

$$logit(\pi(x)) = \beta_0$$

(a) *Using the residual and null deviances carry out a drop in deviance test to compare these two models. Compare the results (value of test statistic and p-value) to those from* ***prop.trend.test****.*

```
21.3981-5.7913
```

```
[1] 15.6068
```

```
pchisq(15.6068, 1, lower.tail = FALSE)
```

```
[1] 0.00007797367
```

The drop-in-deviance test statistic is $\chi^2_1 = 15.607$ with p-value $= 0.0000780$, giving very strong evidence of a linear association between the log-odds of coronary heart disease and body weight. This test statistic is similar (but larger by 0.3) to the `prop.test` statistic and so the p-values are very close as well.

(b) *What advantages do the logistic regression model have over the results from* ***prop.trend.test****?*

The biggest advantage would be the ability to adjust for other variables by including them in the model as predictors.

4. *Suppose you have a single categorical risk factor measured at 3 levels (L = low, M = medium, and H = high). Assuming that H is chosen to be the reference category define two dummy variables X_1 and X_2 for the L and M groups. A logistic regression model*

$$logit(\pi(x)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

will be fit to the data.

(a) *Interpret the parameters in terms of relevant odds ratios.*

$\boldsymbol{\beta_0}$: For a high-risk individual, the odds of disease are e^{β_0}.

$\boldsymbol{\beta_1}$: The odds of disease for a low-risk individual are e^{β_1} times the odds of disease for a high-risk individual.

$\boldsymbol{\beta_2}$: The odds of disease for a medium-risk individual are e^{β_2} times the odds of disease for a high-risk individual.

(b) *Are there any structural relationships imposed on the odds ratios by this model? (Hint: This question may make more sense after you read the next problem).*

No, this model accommodates any type of relationship among the odds ratios.

5. *The labeling of the categorical variable in the previous question implies a natural ordering. Suppose you code the risk factor as a single quantitative variable X with values 0 for L, 1 for M , and 2 for H.*

   (a) *Describe an appropriate logistic regression model for this situation, providing an interpretation of the parameters.*

   An appropriate model would be

   $$logit(\pi(x)) = \beta_0 + \beta_1 X.$$

   The odds of disease for a low-risk individual are e^{β_0}. Moving up by one level of risk is associated with the odds of disease increasing by a multiplicative factor of e^{β_1}.

   (b) *Are there any structural relationships imposed on the odds ratios by this model?*

   Yes, this model forces the odds ratio for medium-risk individuals against low-risk individuals to be the same as the odds ratio for high-risk individuals versus medium-risk individuals.