Stat 525 Homework 8

Kenny Flagg

November 7, 2016

1. *Diabetes melitus is a major public health problem. Patients who have had the disease longer than 20 years tend to develop diabetic retinopathy (damage to blood vessels in the retina). Early detection is crucial to minimizing damage. A study was conducted to determine if general practitioners could be trained to detect diabetic retinopathy after a short training session. The eye skills of 85 general practioners were evaluated before and after training workshops. Each doctor was asked to evaluate 4 patients who did and 4 patients who did not have retinopathy before and after the workshop. We will look at the results for those who did not have retinopathy, i.e. we will look at the evidence for an improvement in specificity. Let X denote post-workshop with $X = 1$ denoting not satisfactory and $X = 2$ denoting satisfactory. Let Y denote pre-workshop with $Y = 1$ denoting not satisfactory and $Y = 2$ denoting satisfactory.*

	$X = 1$	$X = 2$	
$Y = 1$	15	50	65
$Y = 2$	5	15	20
Total	20	65	85

(a) *Let p_1 be the proportion of non-satisfactory results pre-workshop and p_2 be the proportion of non-satisfactory results post-workshop. Estimate the difference in proportions $p_1 - p_2$ and give a standard error for that estimate.*

$$\widehat{p}_1 - \widehat{p}_2 = \frac{50 - 5}{85} = 0.529$$

$$SE\left(\widehat{p}_1 - \widehat{p}_2\right) = \frac{1}{85}\sqrt{(50 + 5) - \frac{(50 - 5)^2}{85}} = 0.0657$$

(b) *Construct an approximate 95% confidence interval for $p_1 - p_2$ and interpret it.*

$$0.529 \pm 1.96 \times 0.0657 = (0.401, 0.658)$$

We are 95% confident that the true proportion of doctors whose retinopathy evaluations were unsatisfactory before the workshop is between 0.401 and 0.658 higher than the proportion doctors whose retinopathy evaluations were unsatisfactory after the workshop.

(c) *Test the hypothesis that the proportions of non-satisfactory results are equal pre and post-workshop.*

```
diabetes <- cbind(c(15, 5), c(50, 15))
mcnemar.test(diabetes)


	McNemar's Chi-squared test with continuity correction

data:  diabetes
McNemar's chi-squared = 35.2, df = 1, p-value = 2.975e-09
```

With $\chi^2_1 = 35.2$ and p-value < 0.0001 we have very convincing evidence that the true proportion of doctors whose retinopathy evaluations were unsatisfactory before the workshop and the proportion of doctors whose retinopathy evaluations were unsatisfactory after the workshop are not equal.

2. *To study an association between the risk of a low birth weight newborn and maternal smoking, an infant who weighed less than 2500 grams at birth (case) was matched to an infant whose birth weight was greater than 2500 grams at birth (control) so that the mother of each infant had the same prepregnancy weight. The risk factor is the mother's smoking exposure (E = smoker, $\overline{E}$ = nonsmoker). A total of $n = 167$ matched pairs (matched on prepregnancy weight) was included in the study. The table below contains the results.*

	$< 2500\ E$	$< 2500\ \overline{E}$	
$\geq 2500\ E$	15	22	37
$\geq 2500\ \overline{E}$	40	90	130
Total	55	112	167

Fit a logistic regression model to these data. Give an estimate of the ratio of the odds that an infant exposed to smoking is low birthweigh to the odds that an infant not exposed is low birthweight. Give an approximate 95% confidence interval and interpret the interval. R-code for this analysis will be provided in a separate script file.

```
id <- rep(1:167, each = 2)
status <- rep(c(1, 0), 167)
smk <- c(rep(c(1, 1), 15), rep(c(1, 0), 40), rep(c(0, 1), 22), rep(c(0, 0), 90))
fun <- function(x){x[1] - x[2]}
z1 <- by(smk, id, FUN = fun)
resp <- status[seq(1, 333, 2)]
fit <- glm(resp ˜ z1 - 1, family = binomial)

summary(fit)


Call:
glm(formula = resp ˜ z1 - 1, family = binomial)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
```

```
0.9362  1.1774  1.1774  1.1774  1.4395

Coefficients:
   Estimate Std. Error z value Pr(>|z|)
z1   0.5978     0.2654   2.252   0.0243

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 231.51  on 167  degrees of freedom
Residual deviance: 226.21  on 166  degrees of freedom
AIC: 228.21

Number of Fisher Scoring iterations: 4

exp(coef(fit))

      z1
1.818182

exp(confint(fit))

   2.5 %   97.5 %
1.091734 3.109002
```

We estimate the odds of low birthweight given smoking exposure to be 1.82 times the odds of low birthweight given no smoking exposure. We are 95% confident that the true odds ratio is between 1.09 and 3.11.

3. *A data set* `lipcancer.txt` *is attached. The data set contains the observed numbers of lip cancer cases (*`obs`*) in 56 Scottish districts between 1975-1980, the expected number of cases (*`exp`*), the percentage of the district population employed in agriculture, fishing, and forestry (*`aff`*), and the latitude (*`lat`*) and longitude (*`long`*) coordinate of the center of each district. The observed counts by themselves can be misleading. For example, district 33 had 7 observed cases but the expected number of cases was also 7. But district 8 had 7 observed cases with only 2.3 cases expected. The ratio of observed to expected counts is a Standardized Morbidity Ratio (SMR). We will model the SMR using a Poisson rate model with the number of expected cases included as an offset term in the model.*

   (a) *Fit a Poisson regression model with observed number of cases as the response and* `aff` *as the explanatory variable. Summarize the results. Show me the output. Interpret the estimated coefficient associated with* `aff`*. Give an approximate 95% confidence interval for that parameter.*

```
lip.cancer <- read.table('Scotland.txt', header = TRUE)
names(lip.cancer) <- c('id', 'obs', 'exp', 'aff', 'lat', 'long')
fit.a <- glm(obs ˜ aff, family = poisson, offset = log(exp), data = lip.cancer)
summary(fit.a)


Call:
glm(formula = obs ˜ aff, family = poisson, data = lip.cancer,
    offset = log(exp))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.7632  -1.2156   0.0967   1.3362   4.7130

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.542268   0.069525   -7.80 6.21e-15
aff          0.073732   0.005956   12.38  < 2e-16

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 380.73  on 55  degrees of freedom
Residual deviance: 238.62  on 54  degrees of freedom
AIC: 450.6

Number of Fisher Scoring iterations: 5
```

```
exp(coef(fit.a)['aff'])
```

```
     aff
1.076518
```

```
exp(confint(fit.a)['aff',])
```

```
   2.5 %   97.5 %
1.063968 1.089107
```

For each additional 1% of the district employed in agriculture, fishing, or farming, we estimate the SMR for lip cancer to be 7.65% larger. We are 95% confident that the true increase is between 6.40% and 8.91%.

(b) *One possible confounding issue is how far north a district lies. More northern districts are more rural and can be expected to have a higher percentage of people who work outside. But the farther north one lives the less the sun exposure. Fit a Poisson regression model with observed number of cases as the response and* `aff` *and* `lat` *as explanatory variables.* `lat` *accounts for the northinghess of the district. Summarize the results. Show me the output. Interpret the estimated coefficient associated with* `aff` *after accounting for the possible confounding effects of* `lat`*. Give an approximate 95% confidence interval for that parameter.*

```
fit.b <- glm(obs ~ aff + lat, family = poisson, offset = log(exp), data = lip.cancer)
summary(fit.b)
```

```
Call:
glm(formula = obs ~ aff + lat, family = poisson, data = lip.cancer,
    offset = log(exp))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3979  -1.0943   0.2294   1.3201   3.0078
```

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -24.586752   2.537728  -9.688  < 2e-16
aff           0.054241   0.006707   8.087  6.1e-16
lat           0.429060   0.045241   9.484  < 2e-16

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 380.73  on 55  degrees of freedom
Residual deviance: 159.71  on 53  degrees of freedom
AIC: 373.69

Number of Fisher Scoring iterations: 5

exp(coef(fit.b)['aff'])

     aff
1.055739

exp(confint(fit.b)['aff',])

   2.5 %   97.5 %
1.041867 1.069625
```

After accounting for lattitude, for each additional 1% of the population employed in agriculture, fishing, or farming, we estimate the SMR for lip cancer to be 5.57% larger, with a 95% confidence interval from 4.19% to 6.96%.

4. *Refer to the (AC, AM, CM) model from the Alcohol-Cigarette-Marijuana use example. On page 211 in the notes I gave the estimated conditional odds ratio of alcohol use among cigarette smokers to alcohol use among non-smokers $\widehat{\theta}_{AC|M} = 7.8$. I also gave approximate 95% (Wald) confidence intervals for $\theta_{AM|C}$ and $\theta_{CM|A}$. Give the points estimate for these quantities for these latter 2 odds ratios and confirm the intervals.*

   Using the model summary on page 209,

$$\widehat{\theta}_{AM|C} = \exp\left(2.98601\right) = 19.8$$

   with 95% confidence interval

$$\exp\left(2.98601 \pm 1.96 \times 0.46468\right) = (7.97, 49.2)$$

   and

$$\widehat{\theta}_{CM|A} = \exp\left(2.84789\right) = 17.3$$

   with 95% confidence interval

$$\exp\left(2.84789 \pm 1.96 \times 0.16384\right) = (12.5, 23.8).$$