Biometry 711 - Categorical Data Analysis
Summer 2015
Homework 4
Due Wednesday 5 August 2015, 6pm

This homework assignment addresses the problem of obtaining adjusted relative risk estimates.

1. In homework 2, we reviewed how to obtain adjusted odds ratios using the Cochran-Mantel-Haenszel approach. A similar approach can be used to obtain relative risk estimates adjusted for the confounding effects of a categorical variable. (In fact, the code we used in HW2 produces adjusted RR estimates in addition to adjusted OR estimates.) Using `SAS' PROC FREQ` (or otherwise) and the `IMPACT` data set, estimate the relative risk and corresponding 95% CI for being drug free at 12 months for subjects randomized to the long treatment arm compared to those randomized to the short treatment arm, controlling for treatment site. Interpret the relative risk estimate using a complete sentence.

2. When the confounding variable is continuous or the number of confounders is large, stratified techniques for adjustment fail. Instead, we turn to regression models to tackle this problem. We begin within the simple framework of a binary response, y_i, and a single binary predictor, x_i which takes on values of 1 or 0 for exposed and unexposed conditions, respectively. Let $\pi(x_i) = P(Y_i = 1|x_i)$ and consider the model $\log(\pi(x_i)) = \alpha + \beta x_i$. Demonstrate that β is the log relative risk comparing subjects with $x_i = 1$ to those with $x_i = 0$.

3. The model in question 2 can easily be extended to accommodate multiple covariates as $\log(\pi(\mathbf{x}_i)) = \mathbf{x}_i'\beta$. If we assume the $\{y_i\}$ are realizations from a binomial distribution, we call this a *log-binomial model*. Although you can't fit this model in `PROC LOGISTIC` (it won't allow a log link), you can easily fit this model in `PROC GENMOD` by specifying a binomial distribution and a log link function. One thing to notice about the log-binomial model is the requirement that the linear predictor be negative (why?). This constraint can result in convergence issues. Therefore, although the log-binomial model is an attractive and simple approach to constructing adjusted RR estimates, it can be problematic in practice. Use a log-binomial model to construct the same adjusted RR and corresponding 95% CI as estimated in question 1.

4. An alternative regression model assumes the $\{y_i\}$ are realizations from a Poisson distribution. The Poisson distribution is a good approximation to the binomial distribution when the sample size is large and the event is rare. This too is easily fit using `PROC GENMOD` and it doesn't suffer from the convergence issues of the log-binomial model, although in practice the veracity of the assumptions may be questioned. Use a Poisson regression model to construct the same adjusted RR and corresponding 95% CI as estimated in question 1.

5. The last approach we will explore is one proposed by Zou. Read the AJE article by Guangyong Zou. Note that there are a number of errors in the manuscript.

   (a) The log-likelihood on page 703 should read

$$\ell(\alpha, \beta) = C + \sum_{i=1}^{n} [y_i(\alpha + \beta x_i) - \exp(\alpha + \beta x_i)].$$

(b) The equation on the bottom of page 703 in the first column should read

$$\widehat{\text{var}}(\hat{\beta}) = \widehat{\text{var}}(\log \widehat{\text{RR}}) = 1/a + 1/c.$$

(c) The first equation at the top of page 703 in the second column should read

$$\widehat{\text{var}}(\log \widehat{\text{RR}}) = \frac{1}{a^2}\sum_{i=1}^{n_1}[y_i - \exp(\hat{\alpha} + \hat{\beta})]^2 + \frac{1}{c^2}\sum_{i=1}^{n_0}[y_i - \exp(\hat{\alpha})]^2.$$

(d) The second equation at the top of page 703 in the second column is an alternative formula for $\widehat{\text{var}}(\log \widehat{\text{RR}})$ where $\hat{\alpha}$ and $\hat{\beta}$ are replaced by their estimates.

Show the MLEs for $\exp(\alpha)$ and $\exp(\beta)$ are $\exp(\hat{\alpha}) = c/n_0$ and $\exp(\hat{\beta}) = (an_0)/(cn_1)$, as stated in the paper.

6. Although Zou derives the results on page 703 using a Poisson likelihood-based approach, we can also tackle this problem using Quasi-likelihood where we specify:

   (a) The relationship between the marginal mean, μ_i and the covariate x_i. Specifically, $g(\mu_i) = \log(\pi(x_i)) = \alpha + \beta x_i$ so that $\mu_i = \pi(x_i) = \exp(\alpha + \beta x_i)$.

   (b) The variance of Y_i as a function of the mean. Specifically, $\text{Var}(Y_i) = \phi v(\mu_i) = \mu_i$ (the usual Poisson assumptions with scale parameter = 1 and variance = mean).

   Based on these assumptions, find an explicit expression for

$$\mathbf{D} = \begin{pmatrix} \partial\mu_1/\partial\alpha & \partial\mu_1/\partial\beta \\ \partial\mu_2/\partial\alpha & \partial\mu_2/\partial\beta \\ \vdots & \vdots \\ \partial\mu_n/\partial\alpha & \partial\mu_n/\partial\beta \end{pmatrix}.$$

   Then use the definition $\mathbf{V} = \phi\ \text{diag}(v(\mu_i))$, to show that the model-based estimate of the variance-covariance matrix for the estimated parameter vector $(\hat{\alpha}, \hat{\beta})$ defined in class as $[\mathbf{D}'\mathbf{V}^{-1}\mathbf{D}]^{-1}$ is equivalent to

$$\begin{pmatrix} \frac{1}{n_0 e^{\alpha}} & -\frac{1}{n_0 e^{\alpha}} \\ -\frac{1}{n_0 e^{\alpha}} & \frac{1}{n_1 e^{\alpha} e^{\beta}} + \frac{1}{n_0 e^{\alpha}} \end{pmatrix}.$$

7. Using the expressions for $\hat{\alpha}$ and $\hat{\beta}$ shown in the paper, show that the estimate of $[\mathbf{D}'\mathbf{V}^{-1}\mathbf{D}]^{-1}$ is

$$\begin{pmatrix} \frac{1}{c} & -\frac{1}{c} \\ -\frac{1}{c} & \frac{1}{a} + \frac{1}{c} \end{pmatrix}$$

   and hence $\widehat{\text{var}}(\hat{\beta}) = \widehat{\text{var}}(\log \widehat{\text{RR}}) = 1/a + 1/c$ as stated on the bottom of page 703.

8. As stated on page 703 of the paper, "... since the error term is misspecified when the underlying data are binomially distributed, the sandwich estimator is used to make the appropriate correction." Recall that the sandwich estimator is given by

$$[\mathbf{D}'\mathbf{V}^{-1}\mathbf{D}]^{-1}[\mathbf{D}'\mathbf{V}^{-1}\mathbf{V}^{*}\mathbf{V}^{-1}\mathbf{D}][\mathbf{D}'\mathbf{V}^{-1}\mathbf{D}]^{-1}$$

where $\mathbf{V}^*$ is the true variance-covariance matrix of the data, and is estimated by

$$\hat{\mathbf{V}}^* = \text{diag}[(y_i - \hat{\mu}_i)^2].$$

Show that an estimate of the "meat" of the sandwich estimator can be written as

$$\begin{pmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{pmatrix}$$

where

$$v_{11} = \sum_{i=1}^{n_0}(y_i - e^{\hat{\alpha}})^2 + \sum_{i=1}^{n_1}(y_i - e^{\hat{\alpha}+\hat{\beta}})^2,$$

and

$$v_{12} = v_{21} = v_{22} = \sum_{i=1}^{n_1}(y_i - e^{\hat{\alpha}+\hat{\beta}})^2.$$

9. Using the estimate of the model-based variance derived in question 7, show that the element in the second row, second column of the sandwich estimator is given by

$$\frac{v_{11}}{c^2} - \frac{v_{21}}{c}\left(\frac{1}{a} + \frac{1}{c}\right) - \frac{v_{12}}{c}\left(\frac{1}{a} + \frac{1}{c}\right) + v_{22}\left(\frac{1}{a} + \frac{1}{c}\right)^2$$

and show that this yields the expression for $\widehat{\text{var}}(\log \widehat{\text{RR}})$ given on the top of page 703, second column.

10. Finally, substitute the expressions for $\hat{\alpha}$ and $\hat{\beta}$ shown in the paper into the variance estimate derived in question 9 and demonstrate its equivalence to the expression for $\widehat{\text{var}}(\log \widehat{\text{RR}})$, the second equation on the top of page 703, second column. Note that this expression is equivalent to Equation 3.5 given in Agresti.

11. Now, using the procedure outlined in the paper, use Zou's modified Poisson regression approach to construct the same adjusted RR and corresponding 95% CI as estimated in question 1.