Coursework 5: STAT 570

Philip Pham

October 31, 2018

1. Consider the data given in Table 1, which are a simplified version of those reported in Breslow and Day (1980). These data arose from a case-control study that was carried out to investigate the relationship between esophageal cancer and various risk factors. Disease status is denoted Y with $Y = 0$ and $Y = 1$ corresponding to without/with disease and alcohol consumption is represented by X with $X = 0$ and $X = 1$ denoting less than 80g and greater than or equal to 80g on average per day. Let the probabilities of high alcohol consumption in the cases and controls be denoted

$$p_1 = \mathbb{P}\left(X = 1 \mid Y = 1\right) \text{ and } p_2 = \mathbb{P}\left(X = 1 \mid Y = 0\right), \tag{1}$$

respectively. Further, let X_1 be the number exposed from n_1 cases and X_2 be the number exposed from n_2 controls. Suppose $X_i \mid p_i \sim \text{Binomial}(n_i, p_i)$ in the case $(i = 1)$ and control $(i = 2)$ groups.

   (a) Of particular interest in studies such as this is the odds ratio defined by

$$\theta = \frac{\mathbb{P}\left(Y = 1 \mid X = 1\right)/\mathbb{P}\left(Y = 0 \mid X = 1\right)}{\mathbb{P}\left(Y = 1 \mid X = 0\right)/\mathbb{P}\left(Y = 0 \mid X = 0\right)}. \tag{2}$$

   Show that the odds ratio is equal to

$$\theta = \frac{\mathbb{P}\left(X = 1 \mid Y = 1\right)/\mathbb{P}\left(X = 0 \mid Y = 1\right)}{\mathbb{P}\left(X = 1 \mid Y = 0\right)/\mathbb{P}\left(X = 0 \mid Y = 0\right)} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}. \tag{3}$$

   Solution: We have that

$$\mathbb{P}\left(Y = y \mid X = x\right) = \frac{\mathbb{P}\left(X = x \mid Y = y\right)\mathbb{P}\left(Y = y\right)}{\mathbb{P}\left(X = x\right)} \tag{4}$$

   by Bayes' rule. Applying Equation 4 to Equation 2, we get

$$\theta = \frac{\left[\mathbb{P}\left(X = 1 \mid Y = 1\right)\mathbb{P}\left(Y = 1\right)\right]/\left[\mathbb{P}\left(X = 0 \mid Y = 1\right)\mathbb{P}\left(Y = 0\right)\right]}{\left[\mathbb{P}\left(X = 0 \mid Y = 1\right)\mathbb{P}\left(Y = 1\right)\right]/\left[\mathbb{P}\left(X = 0 \mid Y = 0\right)\mathbb{P}\left(Y = 0\right)\right]}. \tag{5}$$

   The $\mathbb{P}\left(Y = y\right)$ factors cancel and we obtain the first part of Equation 3. Using Equation 1, we substitute to obtain the second part of Equation 3.

	$X = 0$	$X = 1$	
$Y = 1$	104	96	200
$Y = 0$	666	109	775

Table 1: Case-control data: $Y = 1$ corresponds to the event of esophageal cancer, and $X = 1$ exposure to greater than 80g of alcohol per day. There are 200 cases and 775 controls.

(b) Obtain the MLE and a 90% confidence interval for θ, for the data of Table 1.

Solution: The likelihood and log-likelihood functions are

$$\begin{aligned} L(p_1, p_2) &= \binom{n_1}{x_1} p_1^{x_1} (1-p_1)^{n_1-x_1} + \binom{n_2}{x_2} p_2^{x_2} (1-p_2)^{n_2-x_2} && (6) \\ l(p_1, p_2) &= \log L(p_1, p_2) \\ &= \sum_{i=1}^{2} \left[\log \binom{n_i}{x_i} + x_i \log p_i + (n_i - x_i) \log(1-p_i) \right], \end{aligned}$$

so the score function is

$$S(p_1, p_2) = \nabla \log L(p_1, p_2) = \begin{pmatrix} \frac{x_1 - n_1 p_1}{p_1(1-p_1)} \\ \frac{x_2 - n_2 p_2}{p_2(1-p_2)}. \end{pmatrix} \quad (7)$$

Thus, the Fisher information is

$$I(p_1, p_2) = mathbbE\left[S(p_1, p_2) S(p_1, p_2)^\intercal\right] = \begin{pmatrix} \frac{n_1}{p_1(1-p_1)} & 0 \\ 0 & \frac{n_2}{p_2(1-p_2)} \end{pmatrix}. \quad (8)$$

From Equation 7, we can solve $S(\hat{p}_1, \hat{p}_2) = \mathbf{0}$ to get the MLEs $\hat{p}_1 = x_1/n_1$ and $\hat{p}_2 = x_2/n_2$. Since the MLE is invariant to reparameterization, we have the MLE for θ:

$$\boxed{\hat{\theta} = \frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_2/(1-\hat{p}_2)} = \frac{1992}{1417} \approx 5.640.} \quad (9)$$

We estimate the confidence interval for $\log \hat{\theta}$ which works since log is a monotonic transform. Using the delta method and Equation 8, we have that

$$\begin{aligned} \operatorname{Var}\left(\log \hat{\theta}\right) &\approx \left(\nabla \log \hat{\theta}\right)^\intercal \left(I(\hat{p}_1, \hat{p}_2)\right)^{-1} \left(\nabla \log \hat{\theta}\right) \\ &= \begin{pmatrix} \frac{1}{\hat{p}_1(1-\hat{p}_1)} & \frac{1}{\hat{p}_2(1-\hat{p}_2)} \end{pmatrix} \begin{pmatrix} \frac{\hat{p}_1(1-\hat{p}_1)}{n_1} & 0 \\ 0 & \frac{\hat{p}_2(1-\hat{p}_2)}{n_2} \end{pmatrix} \begin{pmatrix} \frac{1}{\hat{p}_1(1-\hat{p}_1)} \\ \frac{1}{\hat{p}_2(1-\hat{p}_2)} \end{pmatrix} \\ &= \frac{1}{n_1 \hat{p}_1 (1-\hat{p}_1)} + \frac{1}{n_2 \hat{p}_2 (1-\hat{p}_2)} \\ &= \frac{1}{n_1 \hat{p}_1} + \frac{1}{n_1(1-\hat{p}_1)} + \frac{1}{n_2 \hat{p}_2} + \frac{1}{n_2(1-\hat{p}_2)}. && (10) \end{aligned}$$

Numerically, this is $\operatorname{Var}\left(\log \hat{\theta}\right) \approx 0.0307$.

The 90% confidence interval for $\log \hat{\theta}$ is approximately

$$\left(\log \hat{\theta} - \Phi^{-1}(0.95) \sqrt{\operatorname{Var}\left(\log \hat{\theta}\right)}, \log \hat{\theta} + \Phi^{-1}(0.95) \sqrt{\operatorname{Var}\left(\log \hat{\theta}\right)}\right), \quad (11)$$

which is about $(1.441, 2.018)$. Taking the exponent of both sides, we have a 90% confidence interval for $\hat{\theta}$ of $\boxed{(4.228, 7.524).}$

(c) We now consider a Bayesian analysis. Assume that the prior distribution for p_i is the beta distribution $\operatorname{Beta}(a, b)$ for $i = 1, 2$. Show that the posterior distribution $p_i \mid x_i$ is given by the beta distribution $\operatorname{Beta}(a + x_i, b + n_i - x_i)$, $i = 1, 2$.

Solution: