Coursework 7: STAT 570

Philip Pham

November 24, 2018

1. Create a binary variable Z_i, with $Z_i = 0$ corresponding to $Y_i \in \{0, 1\}$ and $Z_i = 1$ corresponding to $Y_i \in \{2, 3\}$. Let $q(x_i) = \mathbb{P}(Z_i = 1 \mid x_i)$, with $\mathbf{x}_i = \begin{pmatrix}1 & x_{1i} & x_{2i}\end{pmatrix}^\intercal$, represent the probability of mental impairment being *Moderate* or *Impaired*, given covariates $\mathbf{x}_i$, $i = 1, \ldots, n = 40$. Provide a single plot that shows the association between $q(x_i)$ and x_{1i} and x_{2i}, on a response scale you feel is appropriate. Comment on the plot.

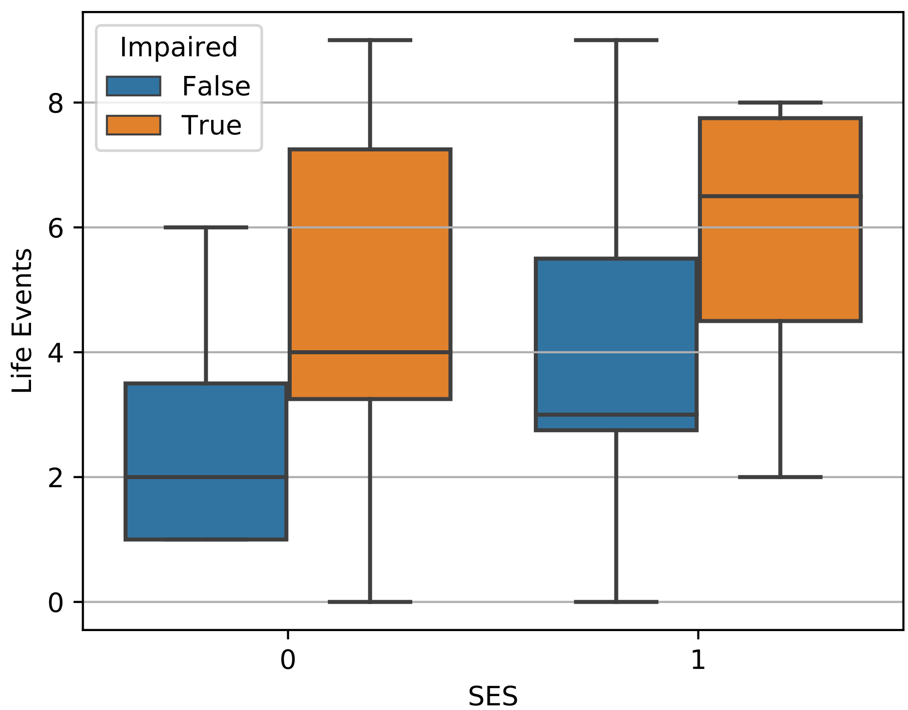

Figure 1: Orange denotes $Z_i = 1$ and blue denotes $Z_i = 0$.

Solution: See Figure 1. Conditioned on SES, those that are impaired ($Z_i = 1$) have a greater number of life events on average.

2. Suppose $Z_i \mid q_i \sim \text{Binomial}(1, q_i)$ independently for $i = 1, \ldots, n = 40$, where $q_i = q(x_i)$. Consider the logistic regression model,

$$q(x_i) = \log\left(\frac{q(\mathbf{x}_i)}{1 - q(\mathbf{x}_i)}\right) = \mathbf{x}_i^\intercal \boldsymbol{\gamma} = \gamma_0 + \gamma_1 x_{1i} + \gamma_2 x_{2i}, \tag{1}$$

where $\boldsymbol{\gamma} = \begin{pmatrix} \gamma_0 & \gamma_1 & \gamma_2 \end{pmatrix}^\intercal$. Write down the log-likelihood $l(\boldsymbol{\gamma})$ for the sample z_i, $i = 1, \ldots, n$.

Solution: Solving for $q(\mathbf{x}_i)$ in Equation 1, we find

$$q(\mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^\intercal \boldsymbol{\gamma})}{1 + \exp(\mathbf{x}_i^\intercal \boldsymbol{\gamma})} = \frac{1}{1 + \exp(-\mathbf{x}_i^\intercal \boldsymbol{\gamma})}. \tag{2}$$

The likelihood function is $L(\boldsymbol{\gamma}) = \prod_{i=1}^n (q(\mathbf{x}_i))^{z_i} (1 - q(\mathbf{x}_i))^{1-z_i}$, so the log-likelihood function becomes

$$\begin{aligned}
l(\boldsymbol{\gamma}) = \log L(\boldsymbol{\gamma}) &= \sum_{i=1}^n (z_i \log q(\mathbf{x}_i) + (1 - z_i) \log(1 - q(\mathbf{x}_i))) \qquad (3)\\
&= \sum_{i=1}^n \left(z_i \log \frac{q(\mathbf{x}_i)}{1 - q(\mathbf{x}_i)} + \log(1 - q(\mathbf{x}_i)) \right) \\
&= \sum_{i=1}^n \left(z_i \mathbf{x}_i^\intercal \boldsymbol{\gamma} + \log \frac{1}{1 + \exp(\mathbf{x}_i^\intercal \boldsymbol{\gamma})} \right) = \sum_{i=1}^n - \log(1 + \exp((1 - 2z_i) \mathbf{x}_i^\intercal \boldsymbol{\gamma})).
\end{aligned}$$

3. Fit the model described in the previous part, and give confidence intervals for the odds ratios.

   Carefully interpret these odds ratios.

	Estimate	Standard error	95% CI lower bound	95% CI upper bound
γ_0	-0.925065	0.723346	-2.342797	0.492666
γ_1	-1.629731	0.780849	-3.160167	-0.099296
γ_2	0.309899	0.147920	0.019980	0.599818

Table 1: Estimates and confidence intervals for $\hat{\boldsymbol{\gamma}}$ using maximum likelihood estimation.

Solution: Taking the derivative of Equation 3, we have the score function:

$$\begin{aligned}
S(\boldsymbol{\gamma}) = \nabla^\intercal l(\boldsymbol{\gamma}) &= \sum_{i=1}^n \frac{2z_i - 1}{1 + \exp((1 - 2z_i) \mathbf{x}_i^\intercal \boldsymbol{\gamma})} \exp((1 - 2z_i) \mathbf{x}_i^\intercal \boldsymbol{\gamma}) \mathbf{x}_i. \\
&= \sum_{i=1}^n \frac{2z_i - 1}{1 + \exp((2z_i - 1) \mathbf{x}_i^\intercal \boldsymbol{\gamma})} \mathbf{x}_i \\
&= X^\intercal (\mathbf{z} - \mathbf{q}(X)), \qquad (4)
\end{aligned}$$

where $\mathbf{z} = \begin{pmatrix} z_1 & z_2 & \cdots & z_n \end{pmatrix}^\intercal$ and $\mathbf{q}(X) = \begin{pmatrix} q_1 & q_2 & \cdots & q_n \end{pmatrix}^\intercal$.
From Equation 4, we have the Fisher information matrix:

$$\begin{aligned}
I_n(\boldsymbol{\gamma}) &= \operatorname{var}(S(\boldsymbol{\gamma}) \mid \boldsymbol{\gamma}) = \mathbb{E}[S(\boldsymbol{\gamma}) S(\boldsymbol{\gamma})^\intercal \mid \boldsymbol{\gamma}] \\
&= \mathbb{E}[X^\intercal (\mathbf{z} - \mathbf{q}(X))(\mathbf{z} - \mathbf{q}(X))^\intercal X \mid \boldsymbol{\gamma}] \\
&= X^\intercal \mathbb{E}[(\mathbf{z} - \mathbf{q}(X))(\mathbf{z} - \mathbf{q}(X))^\intercal \mid \boldsymbol{\gamma}] X \\
&= \sum_{i=1}^n q(\mathbf{x}_i)(1 - q(\mathbf{x}_i)) \mathbf{x}_i \mathbf{x}_i^\intercal = \sum_{i=1}^n \frac{1}{2 + \exp(-\mathbf{x}_i^\intercal \boldsymbol{\gamma}) + \exp(\mathbf{x}_i^\intercal \boldsymbol{\gamma})} \mathbf{x}_i \mathbf{x}_i^\intercal, \qquad (5)
\end{aligned}$$

where we have used independence of the observations and variance of the binomial distribution to get the last line.

We solve Equation 4, $S(\hat{\boldsymbol{\gamma}}) = \mathbf{0}$, to get an estimate for $\boldsymbol{\gamma}$. Using Equation 5, we have that

$$\hat{\boldsymbol{\gamma}} \xrightarrow{\mathcal{D}} \mathcal{N}\left(\gamma, I_n^{-1}(\hat{\gamma})\right), \tag{6}$$

that is, $\hat{\boldsymbol{\gamma}}$ is asymptotically normal.

Using Equation 6, we obtain the estimates and intervals in Table 1.

The predicted log odds ratio given some $\mathbf{x}_i$ is

$$\hat{\theta}_i = \mathbf{x}_i^\intercal \hat{\boldsymbol{\gamma}}, \tag{7}$$

which will have variance

$$\operatorname{var}\left(\hat{\theta}_i\right) = \mathbf{x}_i^\intercal \operatorname{var}(\hat{\gamma})\, \mathbf{x}_i \approx \mathbf{x}_i^\intercal I_n^{-1}(\hat{\gamma})\, \mathbf{x}_i, \tag{8}$$

using Equation 6.