Coursework 8: STAT 570

Philip Pham

November 27, 2018

1. Consider n experiments with Z_{ij} , $j = 1, 2, \ldots, N_i$, the binary outcomes within cluster (experiment) i with $Y_i = \sum_{j=1}^{N} Z_{ij}$, $i = 1, \ldots, n$.

   (a) By writing
$$\operatorname{var}(Y_i) = \sum_{j=1}^{N_i} \operatorname{var}(Z_{ij}) + \sum_{j=1}^{N_i} \sum_{j \neq k} \operatorname{cov}(Z_{ij}, Z_{ik}), \tag{1}$$
   show that
$$\operatorname{var}(Y_i) = N_i p_i (1 - p_i) \times \left[1 + (N_i - 1)\tau_i^2\right], \tag{2}$$
   where $p_i = \mathbb{E}[Z_{ij}]$ and τ_i^2 is the correlation of outcomes within cluster i.

   Solution: Using the variance for a Bernoulli random variable and the definition of the correlation coefficient, we have that
$$\begin{aligned} \operatorname{var}(Z_{ij}) &= p_i (1 - p_i) \\ \operatorname{cov}(Z_{ij}, Z_{ik}) &= \tau_i^2 p_i (1 - p_i) \text{ for } j \neq k. \end{aligned} \tag{3}$$
   Since Z_{ij} are identically distributed for different j, we can rewrite Equation 1 as
$$\operatorname{var}(Y_i) = N_i \operatorname{var}(Z_{i1}) + N_i (N_i - 1) \operatorname{cov}(Z_{i1}, Z_{i2}). \tag{4}$$
   Applying Equation 3 to Equation 4, we have the result
$$\begin{aligned} \operatorname{var}(Y_i) &= N_i p_i (1 - p_i) + N_i (N_i - 1) \tau_i^2 p_i (1 - p_i) \\ &= N_i p_i (1 - p_i) \times \left[1 + (N_i - 1)\tau_i^2\right] \end{aligned}$$
   as desired.

   (b) Consider the model
$$Y_i \mid q_i \sim \operatorname{Binomial}(N_i, q_i) \tag{5}$$
$$q_i \sim \operatorname{Beta}(a_i, b_i), \tag{6}$$
   where we can parameterize as $a_i = dp_i$, $b_i = d(1 - p_i)$, so that
$$\mathbb{E}[q_i] = p_i = \frac{a_i}{d} \tag{7}$$
$$\operatorname{var}(q_i) = \frac{p_1 (1 - p_i)}{d + 1}. \tag{8}$$
   Obtain the marginal moments and show that the variance is of the form in Equation 2, and identify τ_i^2.

Solution: We have that

$$\begin{aligned}
\mathbb{P}(Y_i = y) &= \int_0^1 \mathbb{P}(Y_i \mid q_i)\, p(q_i)\, \mathrm{d}q_i \\
&= \int_0^1 \left(\binom{N_i}{y} q_i^y (1-q_i)^{N_i - y} \right) \left(\frac{\Gamma(a_i + b_i)}{\Gamma(a_i)\Gamma(b_i)} q_i^{a_i - 1} (1-q_i)^{b_i - 1} \right) \mathrm{d}q_i \\
&= \binom{N_i}{y} \frac{\Gamma(a_i + b_i)}{\Gamma(a_i)\Gamma(b_i)} \int_0^1 q_i^{a_i + y - 1} (1-q_i)^{b_i + N_i - y - 1}\, \mathrm{d}q_i \\
&= \binom{N_i}{y} \left(\frac{\Gamma(a_i + b_i)}{\Gamma(a_i)\Gamma(b_i)} \right) \left(\frac{\Gamma(y + a_i)\Gamma(N_i - y + b_i)}{\Gamma(N_i + a_i + b_i)} \right),
\end{aligned} \tag{9}$$

so $Y_i \sim \text{BetaBinomial}(N_i, a_i, b_i)$.
Using Equation 9, the expectation of Y_i is

$$\mathbb{E}[Y_i] = \sum_{y=0}^{N_i} y \mathbb{P}(Y_i = y) = \sum_{y=1}^{N_i} y \mathbb{P}(Y_i = y). \tag{10}$$

Note that when $N_i = 1$, Equation 10 trivially becomes $a_i / (a_i + b_i)$. In general, we can show that $\mathbb{E}[Y_i] = N_i \dfrac{a_i}{a_i + b_i}$. With the $N_i = 1$ base case established, we now have

$$\begin{aligned}
\mathbb{E}[Y_i] &= \sum_{y=1}^{N_i} y \mathbb{P}(Y_i = y) \\
&= \sum_{y=1}^{N_i} y \binom{N_i}{y} \left(\frac{\Gamma(a_i + b_i)}{\Gamma(a_i)\Gamma(b_i)} \right) \left(\frac{\Gamma(y + a_i)\Gamma(N_i - y + b_i)}{\Gamma(N_i + a_i + b_i)} \right) \\
&= \frac{N_i}{N_i - 1 + a_i + b_i} \sum_{y=1}^{N_i} (y - 1 + a_i)\, \text{BetaBinomial}_{N_i - 1, a_i, b_i}(y - 1) \\
&= \frac{N_i}{N_i - 1 + a_i + b_i} \left(\frac{(N_i - 1) a_i}{a_i + b_i} + a_i \right) = N_i \frac{a_i}{a_i + b_i}
\end{aligned} \tag{11}$$

as expected. Substituting $a_i = dp_i$ and $b_i = d(1 - p_i)$, we have that

$$\mathbb{E}[Y_i] = N_i \frac{dp_i}{dp_i + d(1 - p_i)} = N_i p_i. \tag{12}$$

For the variance, we can use the law of total variance to obtain

$$\begin{aligned}
\operatorname{var}(Y_i) &= \mathbb{E}[\operatorname{var}(Y_i \mid q_i)] + \operatorname{var}(\mathbb{E}[Y_i \mid q_i]) \\
&= \mathbb{E}[N_i q_i (1 - q_i)] + \operatorname{var}(N_i q_i) \\
&= N_i \left(\frac{a_i}{a_i + b_i} - \left(\frac{a_i b_i}{(a_i + b_i)^2 (a_i + b_i + 1)} + \left(\frac{a_i}{a_i + b_i} \right)^2 \right) \right) \\
&\quad + N_i^2 \frac{a_i b_i}{(a_i + b_i)^2 (a_i + b_i + 1)} \\
&= N_i \frac{a_i b_i (a_i + b_i + N_i)}{(a_i + b_i)^2 (a_i + b_i + 1)}.
\end{aligned} \tag{13}$$

From Equations 11 and 13, we obtain the second moment

$$\begin{aligned}\mathbb{E}\left[Y_i^2\right] &= \mathrm{var}\left(Y_i\right) + \left(\mathbb{E}\left[Y_i\right]\right)^2 \\ &= N_i \frac{a_i b_i \left(a_i + b_i + N_i\right)}{\left(a_i + b_i\right)^2 \left(a_i + b_i + 1\right)} + \left(N_i \frac{a_i}{a_i + b_i}\right)^2 \\ &= N_i \frac{a_i \left(N_i \left(a_i + 1\right) + b_i\right)}{\left(a_i + b_i\right)\left(a_i + b_i + 1\right)}.\end{aligned}$$

Substituting $a_i = dp_i$ and $b_i = d\left(1 - p_i\right)$ into Equation 13, we have

$$\begin{aligned}\mathrm{var}\left(Y_i\right) &= N_i p_i \left(1 - p_i\right) \frac{a_i + b_i + N_i}{a_i + b_i + 1} \\ &= N_i p_i \left(1 - p_i\right) \frac{a_i + b_i + 1 + \left(N_i - 1\right)}{a_i + b_i + 1} \\ &= N_i p_i \left(1 - p_i\right) \times \left[1 + \left(N_i - 1\right) \frac{1}{d + 1}\right]. \qquad (14)\end{aligned}$$

Thus, we have that $\tau^2 = 1/\left(d + 1\right)$, so small values of d mean that the Z_{ij} are highly correlated. This is consistent with the behavior of the beta distribution since for small d, q_i is likely to be close to 0 and 1.

2. In this question a simulation study to investigate the impact on inference of omitting covariates in logistic regression will be performed, in the situation in which the covariates are independent of the exposure of interest. Let x be the covariate of interest and z another covariate. Suppose the true (adjusted) model is independently distributed $Y_i \mid x_i, z_i \sim \mathrm{Bernoulli}\left(p_i\right)$, where

$$\log \frac{p_i}{1 - p_i} = \beta_0 + \beta_1 x_i + \beta_2 z_i.. \qquad (15)$$

A comparison with the unadjusted model $Y_i \mid x_i \sim \mathrm{Bernoulli}\left(p_i^\star\right)$ independently distributed, where

$$\log \frac{p_i^\star}{1 - p_i^\star} = \beta_0^\star + \beta_1^\star x_i, \qquad (16)$$

for $i = 1, \ldots, n = 1000$ will be made. Suppose x is binary with $\mathbb{P}\left(X = 1\right) = 0.5$ and $Z \sim \mathcal{N}\left(0, 1\right)$ independent and identically distributed with x and z independent. Combinations of the parameters $\beta_1 \in \{0.5, 1\}$ and $\beta_2 \in \{0.5, 1.0, 2.0, 3.0\}$, with $\beta_0 = -2$ in all cases will be considered.

For each combination of parameters compare the results from the two models, Equation 15 and Equation 16, with respect to:

(a) $\mathbb{E}\left[\hat{\beta}_1\right]$ and $\mathbb{E}\left[\hat{\beta}_1^\star\right]$, as compared to β_1.

(b) The standard errors of $\hat{\beta}_1$ and $\hat{\beta}_1^\star$.

(c) The coverage of 95% confidence intervals for β_1 and $\beta_1^\star$.

(d) The probability of rejecting $H_0 : \beta_1 = 0$ (the power) under both models using a Wald test.

Based on the results, summarize the effect of omitting a covariate that is independent of the exposure of interest, in particular in comparison with the linear model.

Solution: