# Coursework 8: STAT 570

Philip Pham

December 2, 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 $\tau_i^2$ 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 $\tau_i^2$.

**Solution:** We have that

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

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

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

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

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

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

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

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

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

From Equations 11 and 13, we obtain the second moment

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

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

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

Thus, we have that $\boxed{\tau^2 = 1/(d + 1)}$, 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 $Y_i \mid x_i, z_i \sim \text{Bernoulli}(p_i)$, 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 $Y_i \mid x_i \sim \text{Bernoulli}(p_i^\star)$ 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 $\mathbb{P}(X = 1) = 0.5$ and $Z \sim \mathcal{N}(0, 1)$ 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 $\beta_1$.

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

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

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

|  |  |  |  |  | $\mathbb{E}\left[\hat{\beta}_1\right]$ | $\hat{\sigma}_{\hat{\beta}_1}$ | Coverage of 95% CI | Wald test power |
|---|---|---|---|---|---|---|---|---|
| $n$ | $\beta_0$ | $\beta_1$ | $\beta_2$ | Adjusted |  |  |  |  |
| 1000 | -2.0 | 0.5 | 0.5 | False | 0.486574 | 0.176321 | 0.949120 | 0.787659 |
|  |  |  |  | True | 0.503676 | 0.179792 | 0.949654 | 0.800568 |
|  |  |  | 1.0 | False | 0.437434 | 0.165271 | 0.933632 | 0.753868 |
|  |  |  |  | True | 0.503888 | 0.177877 | 0.950001 | 0.808620 |
|  |  |  | 2.0 | False | 0.319081 | 0.146253 | 0.763893 | 0.586224 |
|  |  |  |  | True | 0.503771 | 0.185850 | 0.949600 | 0.773979 |
|  |  |  | 3.0 | False | 0.237999 | 0.137772 | 0.522350 | 0.407654 |
|  |  |  |  | True | 0.503356 | 0.203177 | 0.950344 | 0.695671 |
|  |  | 1.0 | 0.5 | False | 0.966688 | 0.168091 | 0.946236 | 0.999924 |
|  |  |  |  | True | 1.005611 | 0.171580 | 0.949684 | 0.999981 |
|  |  |  | 1.0 | False | 0.864523 | 0.158277 | 0.862934 | 0.999844 |
|  |  |  |  | True | 1.005120 | 0.172235 | 0.949677 | 0.999981 |
|  |  |  | 2.0 | False | 0.631023 | 0.142659 | 0.263969 | 0.993729 |
|  |  |  |  | True | 1.006883 | 0.185441 | 0.949352 | 0.999783 |
|  |  |  | 3.0 | False | 0.472034 | 0.135913 | 0.027699 | 0.935383 |
|  |  |  |  | True | 1.007629 | 0.204881 | 0.949635 | 0.998844 |

Table 1: Result of doing $2^{20}$ simulations for each set of parameters.

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:** See Table 1. For each set of parameters $2^{20}$ simulations were done. The expectation was estimated with Monte Carlo integration. The standard error was estimated by taking the square root of the unbiased variance estimate.

Using the standard error estimate, 95% confidence intervals were calculated for each simulation to see if they contained the true parameter value.

The estimated standard error was also used to compute the Wald test statistic. To estimate the probability of rejecting the null hypothesis ($\beta_1 = 0$), a 95% test was performed for each simulation.

Estimates for $\beta_1$ are biased in both the adjusted and unadjusted model. In the adjusted model, the bias is very small and does not vary much with $\beta_2$. In the unadjusted model, $\hat{\beta}_1^\star$ has bias depedent on $\beta_2$.

When $\beta_2$ is small, say 0.5, the bias is not too bad, and the 95% confidence interval contains the true parameter value close to 95% of the time as we would expect. As $\beta_2$ grows larger, the estimate becomes more and more biased, coverage of the confidence interval worsens, and the power of the Wald test declines. When $\beta_1 = 1.0$, the power is still high because the effect size is large.

In the adjusted model, regardless of $\beta_2$, we obtain the expected 95% coverage of the confidence interval, and the power stays consistently high.

One way to understand this behavior is by looking at the score function:

$$S\left(\beta\right) = X^\intercal \left(Y - \frac{1}{1 + \exp\left(-X\beta\right)}\right). \tag{17}$$

Note that

$$0 = \mathbb{E}\left[S\left(\beta\right) \mid \beta\right] = X^\intercal \left(\mathbf{p} - \frac{1}{1 + \exp\left(-X\beta\right)}\right),$$

where $\mathbf{p} = \begin{pmatrix} p_1 & p_2 & \cdots & p_n \end{pmatrix}^\intercal$. Thus, we are choosing $\beta$ to fit $\operatorname{expit}\left(X\beta\right)$ to $\mathbf{p}$.

In the adjusted model, this amounts to choosing

$$\hat{\beta}_0^\star = \operatorname{logit}\left(\frac{\sum_{i=1}^n (1 - x_i)\, y_i}{\sum_{i=1}^n (1 - x_i)}\right) = \operatorname{logit}\hat{\mathbb{P}}\left(Y_i = 1 \mid X_i = 0\right) \tag{18}$$

$$\hat{\beta}_0^\star + \hat{\beta}_1^\star = \operatorname{logit}\left(\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n 1 - x_i}\right) = \operatorname{logit}\hat{\mathbb{P}}\left(Y_i = 1 \mid X_i = 1\right). \tag{19}$$

Let $\phi$ be the probability density function for the standard normal distribution. Due to nonlinearity, we have that

$$\begin{aligned}\mathbb{P}\left(Y_i = 1 \mid X_i = 0\right) &= \int_{-\infty}^{\infty} \operatorname{expit}\left(\beta_0 + \beta_2 z\right) \phi\left(z\right)\, \mathrm{d}z \neq \operatorname{expit}\left(\beta_0\right) \\ \mathbb{P}\left(Y_i = 1 \mid X_i = 1\right) &= \int_{-\infty}^{\infty} \operatorname{expit}\left(\beta_0 + \beta_1 + \beta_2 z\right) \phi\left(z\right)\, \mathrm{d}z \neq \operatorname{expit}\left(\beta_0 + \beta_1\right).\end{aligned}$$

Substituting the true values for the observed values in Equations 18 and 19, it's clear that we're fitting to the incorrect log odds ratio.

Code for the simulations can be found in `logistic_regression.ipynb`.