Chase Joyner

885 Homework 2

October 13, 2015

## Problem 1:

(a) First note that

$$P(Z_j = 0) = P(Y_{ij} = 0)^c = (1-p)^c$$

and so

$$P(Z_j = 1) = 1 - (1-p)^c.$$

Define $p^\star = 1-(1-p)^c$. Then, we have $Z_j \overset{iid}{\sim} \text{Bernoulli}(p^\star)$. Recall that the MLE for a Bernoulli random variable is the mean, and so

$$\widehat{p^\star} = \overline{Z} = \frac{1}{J}\sum_{j=1}^{J} Z_j.$$

By the invariance property of MLEs, the MLE of $p$ satisfies

$$1-(1-\widehat{p})^c = \widehat{p^\star}.$$

Solving for $\widehat{p}$ and plugging in $\bar{Z}$, we have

$$\widehat{p} = 1-(1-\overline{Z})^{1/c}.$$

Then, we have by CLT,

$$\sqrt{n}\left(\widehat{p^\star} - p^\star\right) \xrightarrow{d} N\big(0, p^\star(1-p^\star)\big).$$

Define a function $g(x) = 1-(1-x)^{1/c}$. Then, $g'(x) = \frac{1}{c}(1-x)^{\frac{1}{c}-1}$, and so by the Delta method,

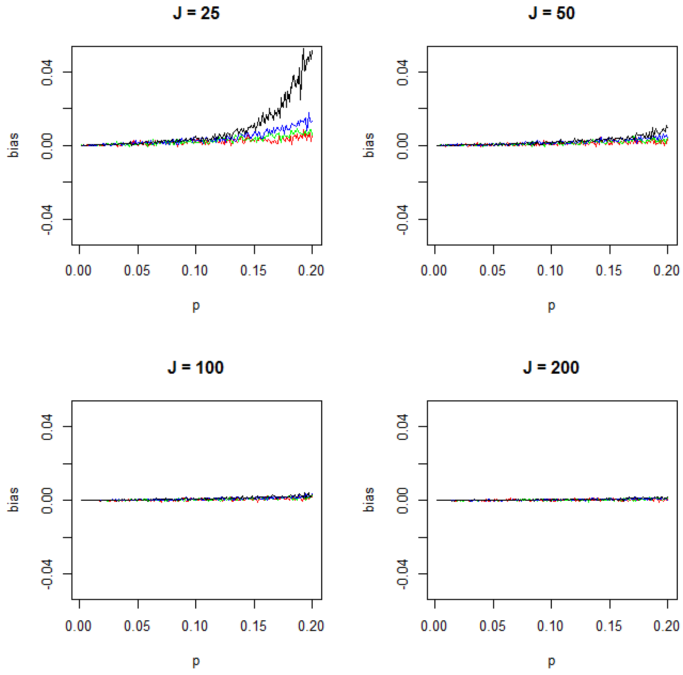
$$\sqrt{n}\left(\widehat{p} - p\right) \xrightarrow{d} N\left(0, \frac{1}{c^2}p^\star(1-p^\star)^{\frac{2}{c}-1}\right).$$

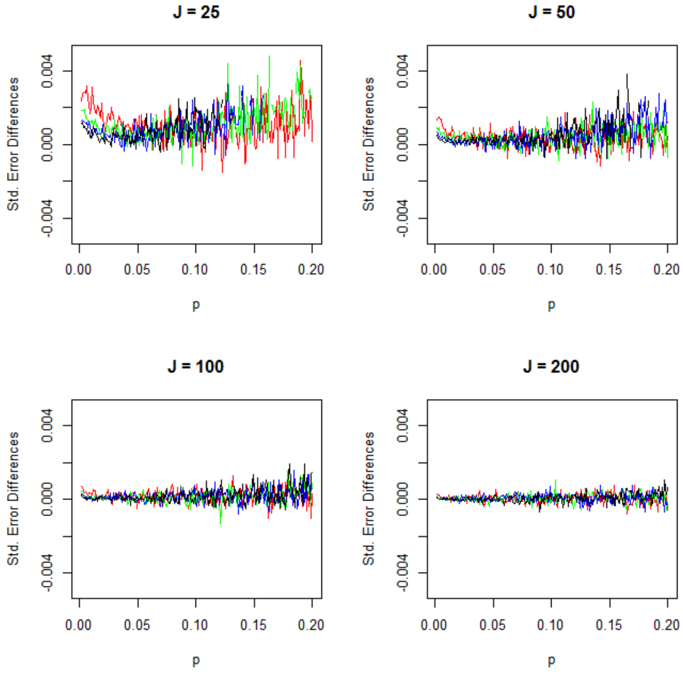
Therefore, a $(1-\alpha)\times 100\%$ asymptotic confidence interval for $p$ is

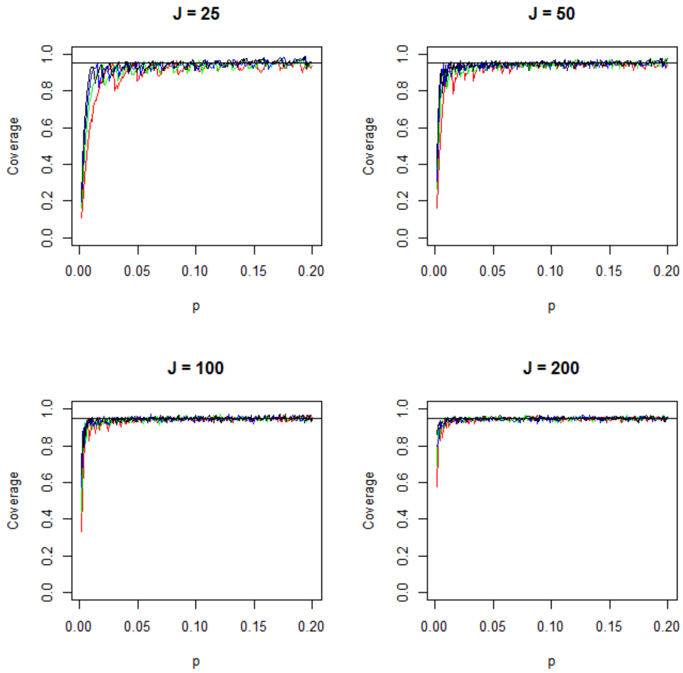
$$\left[\widehat{p} \pm z_{1-\alpha/2}\frac{1}{c}\sqrt{\frac{p^\star(1-p^\star)^{\frac{2}{c}-1}}{n}}\,\right].$$

The R function for this problem can be found in the appendix. We have no simulation for this problem since the function is used for other parts of Problem 1.

(b) Now we look at some results for these simulations. Below we have Figure 1, Figure 2, and Figure 3, which show the bias, difference in asymptotic standard errors and sample standard errors, and the coverage probabilities, respectively.







It appears that as the sample size $J$ increases, the bias, errors, and coverage probabilities begin to act as they should. More specifically, the asymptotics seem to kick in pretty well around $J = 100$. This is because the bias is almost 0 for $J = 100$, and the variance bounces around 0 for this sample size, and also the coverage probabilities seem to be closer to .95 and stay there for $J = 100$. Also, we see that the standard errors increases in $p$, which is to be expected because we have more 1s and 0s as $p$ increases, causing more variability. Lastly, though difficult to see in the plots without color, as the pool sizes increase, the bias, variance, and coverage probabilities become a bit worse. Actually, I expected perhaps the opposite since larger pools would allow for more tests, which in turn we would gain more information. However, maybe there is some sort of diminishing effect for larger pools, to the point where we don't gain more information.

(c) Now we allow for different pool sizes $c_j$ and account for possible measurement error. Under these assumptions, we have

$$\begin{aligned} P\left(\widetilde{Z}_j = 1\right) &= P\left(\widetilde{Z}_j = 1 \mid Z_j = 1\right) P(Z_j = 1) + P\left(\widetilde{Z}_j = 1 \mid Z_j = 0\right) P(Z_j = 0) \\ &= Se \cdot p_j^\star + (1 - Sp)(1 - p_j^\star) =: p_j^{\star\star}, \end{aligned}$$

where $p_j^\star = 1 - (1-p)^{c_j}$. Therefore, the likelihood function is

$$L(p_j^{\star\star}) = \prod_{j=1}^{J} {p_j^{\star\star}}^{\widetilde{Z}_j} (1 - p_j^{\star\star})^{1-\widetilde{Z}_j}.$$

At this point, we use R to optimize the log likelihood function to obtain the MLE $\widehat{p}_j^{\star\star}$, and therefore obtain $\widehat{p}$. Once obtaining $\widehat{p}$, we calculate the asymptotic variance to be

$$\begin{aligned}
\widehat{I}_J(\widehat{p}_J) &= -\nabla_p^2 \frac{1}{J}\ell(\widehat{p}_J) \\
&= -\nabla_p^2 \frac{1}{J}\sum_{j=1}^{J}\left\{ \widetilde{Z}_j \log\left(p_j^{\star\star}\right) + (1-\widetilde{Z}_j)\log\left(1-p_j^{\star\star}\right)\right\} \\
&= -\frac{1}{J}\sum_{j=1}^{J}\Bigg\{ \widetilde{Z}_j \left[ -\frac{1}{\left(p_j^{\star\star}\right)^2}\cdot\frac{dp_j^{\star\star}}{dp} + \frac{1}{p_j^{\star\star}}\cdot\frac{d^2p_j^{\star\star}}{dp^2}\right] - \\
&\qquad (1-\widetilde{Z}_j)\left[\frac{1}{\left(1-p_j^{\star\star}\right)^2}\left(\frac{dp_j^{\star\star}}{dp}\right)^2 + \frac{1}{1-p_j^{\star\star}}\cdot\frac{d^2p_j^{\star\star}}{dp^2}\right]\Bigg\},
\end{aligned}$$

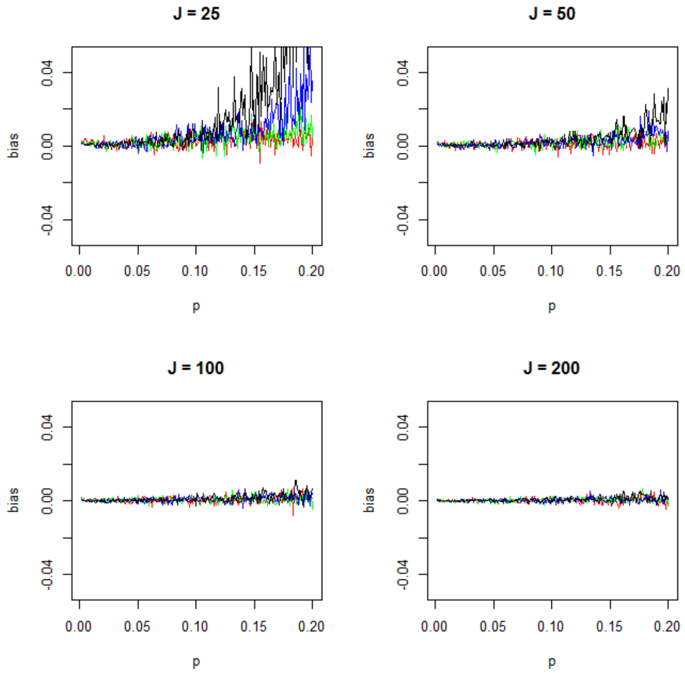
where $p_j^{\star} = 1-(1-p)^{c_j}, p_j^{\star\star} = Se + (1-p)^{c_j}(1-Se-Sp)$, and

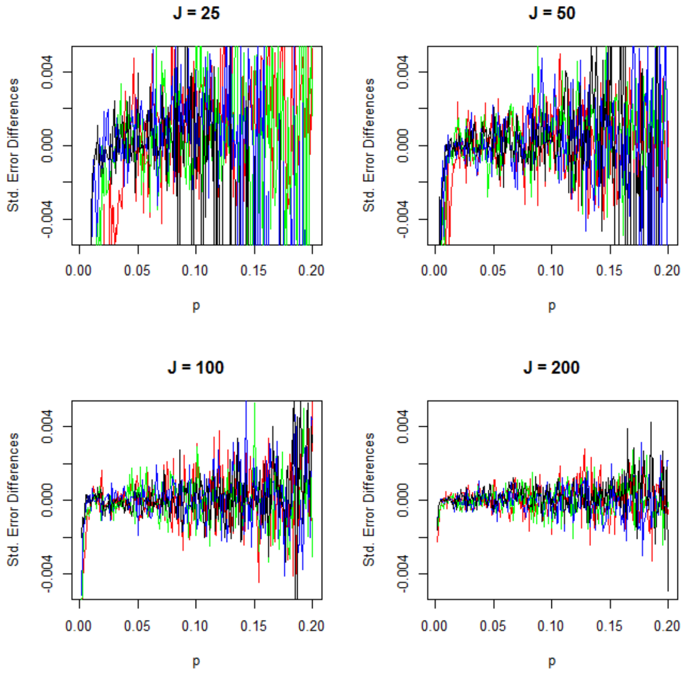
$$\begin{aligned}
\frac{dp_j^{\star\star}}{dp} &= -c_j(1-Se-Sp)(1-p)^{c_j-1} \\
\frac{d^2p_j^{\star\star}}{dp^2} &= c_j(c_j-1)(1-Se-Sp)(1-p)^{c_j-2}.
\end{aligned}$$

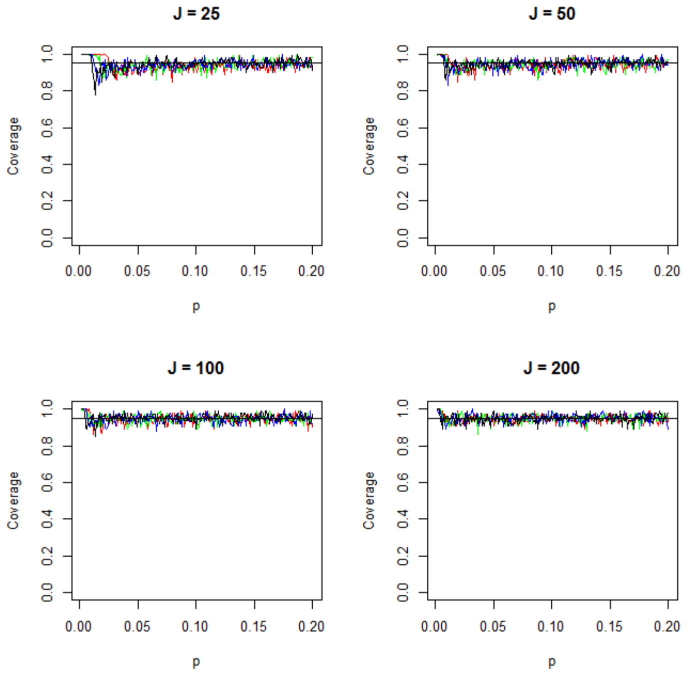
Therefore, a $(1-\alpha)\times 100\%$ asymptotic confidence interval for $p$ is

$$\left[\widehat{p} \pm z_{1-\alpha/2}\sqrt{n^{-1}\widehat{I}_J\left(\widehat{p}_J\right)^{-1}}\right].$$

(d) Now we look at some results for these simulations. This is similar to part (b), however now we look at how random pool sizes affects things. Again, below we have Figures 1, 2, and 3, which show the bias, difference in asymptotic standard errors and sample standard errors, and the coverage probabilities, respectively.


J = 25
J = 50
J = 100
J = 200
bias
0.04
0.00
-0.04
0.00
0.05
0.10
0.15
0.20
p


J = 25
J = 50
J = 100
J = 200
Std. Error Differences
0.004
0.000
-0.004
0.00
0.05
0.10
0.15
0.20
p



Here, we again see that the bias diminishes as the sample size $J$ increases and seems to do pretty well around $J = 100$. Similarly, the difference in the asymptotic standard errors and sample standard error diminishes as $J$ increases. However, perhaps here one would prefer a sample size of $J = 200$ or larger here based on Figure 2. One note to make here is that as $p$ heads towards 0, the standard errors blow up but as $p$ increases, everything seems to work properly. Lastly, the coverage probabilities seem to act similarly. As $p$ heads towards zero, since the errors blow up, the coverage is so large and so we have a coverage probability of 1. As $p$ increases however, the coverage probabilities bounce around 0.95 as expected.

## Problem 2:

(a) As before, we have $Z_j \sim \text{Bernoulli}(p_j^\star)$, but now $p_j^\star = 1 - \prod_{i=1}^{c_j}(1 - p_{ij})$ and

$$p_{ij} = p(\mathbf{x}_{ij}) = \frac{\exp\{\mathbf{x}_{ij}'\boldsymbol{\beta}\}}{1 + \exp\{\mathbf{x}_{ij}'\boldsymbol{\beta}\}}.$$

Therefore, we have

$$\begin{aligned} P\left(\widetilde{Z}_j = 1\right) &= P\left(\widetilde{Z}_j = 1 \mid Z_j = 1\right) P(Z_j = 1) + P\left(\widetilde{Z}_j = 1 \mid Z_j = 0\right) P(Z_j = 0) \\ &= Se \cdot p_j^\star + (1 - Sp)(1 - p_j^\star) = p_j^{\star\star}. \end{aligned}$$

From this, we are able to write the likelihood for the measurement error prone observations $\widetilde{Z}_j$ as

$$L(\boldsymbol{\beta} \mid \widetilde{\mathbf{Z}}) = \prod_{j=1}^{J} {p_j^{\star\star}}^{\widetilde{Z}_j}(1 - p_j^{\star\star})^{1-\widetilde{Z}_j},$$

and so we obtain the log likelihood function

$$\ell(\boldsymbol{\beta}) = \sum_{j=1}^{J} \widetilde{Z}_j \log p_j^{\star\star} + \big(1 - \widetilde{Z}_j\big)\log(1 - p_j^{\star\star}).$$

We now use R to maximize this.

(b) By MLE theory, we know that

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} N\Big(0, I(\boldsymbol{\beta})^{-1}\Big).$$

Define the function $g(\mathbf{y}) = \frac{\exp\{\mathbf{x}_{ij}'\mathbf{y}\}}{1+\exp\{\mathbf{x}_{ij}'\mathbf{y}\}}$. Then,

$$g'(\mathbf{y}) = \frac{\exp\{\mathbf{x}_{ij}'\mathbf{y}\}\mathbf{x}_{ij}'}{(1+\exp\{\mathbf{x}_{ij}'\mathbf{y}\})^2}.$$

Therefore, by the Delta method,

$$\sqrt{n}\big(\widehat{p}_{ij} - p_{ij}\big) \xrightarrow{d} N\left(0, \left(\frac{\exp\{\mathbf{x}_{ij}'\boldsymbol{\beta}\}\mathbf{x}_{ij}'}{(1+\exp\{\mathbf{x}_{ij}'\boldsymbol{\beta}\})^2}\right)^2 I(\boldsymbol{\beta})^{-1}\right).$$

Then, we have the asymptotic confidence interval

$$\left[\widehat{p}_{ij} \pm z_{\alpha/2}\left(\frac{\exp\{\mathbf{x}_{ij}'\boldsymbol{\beta}\}\mathbf{x}_{ij}'}{(1+\exp\{\mathbf{x}_{ij}'\boldsymbol{\beta}\})^2}\right)^2 I(\boldsymbol{\beta})^{-1}\right].$$

(c) I have written a simulation for this part, however my code was giving an error at times about singular matrices or not being able to square root the hessian matrix that optim returned. I can provide my code if necessary.

(d) In the table below, we have the marginal tests, $H_0\colon \beta_k = 0$ vs $H_1\colon \beta_k \neq 0$ for $k = 0, 1, 2, 3$. Also, we have the MLEs for the regression coefficients. First, note that the LRT returned negative numbers, indicating that the ratio was greater than 1. This cannot happen and so there is a coding error somewhere, but I was unable to find it.

| Chlamydia Data Analysis | | | | |
|---|---|---|---|---|
| | $H_0\colon \beta_0 = 0$ | $H_0\colon \beta_1 = 0$ | $H_0\colon \beta_2 = 0$ | $H_0\colon \beta_3 = 0$ |
| Estimate | $\widehat{\beta}_0 = -1.1318$ | $\widehat{\beta}_1 = -0.0597$ | $\widehat{\beta}_2 = 0.7868$ | $\widehat{\beta}_3 = 0.0153$ |
| Wald Test | 27.6691 | 55.91654 | 161.33315 | 0.4564274 |
| P-value | 0 | 0 | 0 | 0.648 |
| LRT | -2.006694 | -2.020476 | -2.040717 | -2.00012 |

We see that the Wald test is indicating that $\beta_3 = 0$ and so the number of hours playing video games is not statistically significant. The LRT should return a similar conclusion if the code were working properly.