STAT 570 Final (Take Home)

Xinyu Gao 1828056

Dec 2019

1 Problem 1

By strong Law of Large Number (L.L.N), we have

$$l_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} logp_\theta(Y_i|X_i) \overset{a.s}{\rightarrow} E_F[logp_\theta(Y|\theta)] \tag{1}$$

where F is the true model and

$$E_F[logp_\theta(Y|X)] = E_{X,Y}[logp_\theta(Y|X)]$$

Then we will prove the following equation (2) that might be used in the following problems:

$$E_{X,Y}(logp_\theta(Y|X)) = E_X E_{Y|X}(logp_\theta(Y|X)) \tag{2}$$

Since

$$E_{X,Y}(logp_\theta(Y|X)) = \int\int p_*(X,Y)logp_\theta(Y|X)dydx = \int p_*(X)\int p_*(Y|X)logp_\theta(Y|X)dydx$$

which is equal to

$$\int p_*(X)(E_{Y|X}(logp_\theta(Y|X)))dx = E_X E_{Y|X}(logp_\theta(Y|X))$$

hence (1) is proved.
Then we apply (2) to (1), we have

$$E_{X,Y}[logp_\theta(Y|X)] = E_X E_{Y|X}(logp_\theta(Y|X)) = E_X[\int p_*(Y|X)logp_\theta(Y|X)dy]$$

where

$$\int p_*(Y|X)logp_\theta(Y|X)dy = \int log\frac{p_\theta(Y|X)}{p_*(Y|X)}p_*(Y|X)dy + \int logp_*(Y|X)p_*(Y|X)dy$$

and the second term on the right hand side is $E_{Y|X}(logp_*(Y|X))$ is not related to θ, so we can ignore it when maximizing to find the optimal θ. Therefore

$$\arg\max_\theta E_F(logp_\theta(Y|X)) = \arg\max_\theta E_X(-KL(p_*,p_\theta|X)) = \arg\min_\theta E_X(KL(p_*,p_\theta|X))$$

Therefore MLE asymptotically minimizes the mean conditional K-L divergence of p_θ from p_*.

2 Problem 2

In this problem, we assume data are generated $\{(Y_i, X_i, Z_i, S_i)\} \overset{i.i.d}{\sim} G$. In this problem, we should note that we cannot directly apply L.L.N to $l_n^1(\theta)$ because the log-likelihood is not divided by n, so we need to transform it before using L.L.N,

$$l_n^1(\theta) = \frac{1/n \sum_{i=1}^n S_i logp_\theta(Y_i|X_i)}{n_1/n}$$

where

$$\frac{n_1}{n} \xrightarrow{p} E_G(S)$$

and

$$\frac{1}{n}\sum_{i=1}^n S_i logp_\theta(Y_i|X_i) \overset{a.s}{\to} E_G(Slogp_\theta(Y|X))$$

Hence by Slutsky's Theorem, we have

$$l_n^1(\theta) \overset{a.s}{\to} \frac{E_G(Slogp_\theta(Y|X))}{E_G(S)}$$

we only need to focus on $E_G(Slogp_\theta(Y|X))$, it can bve expressed as $E_{X,Y,Z,S}(Slogp_\theta(Y|X))$, by Eq.(2), we have

$$E_X E_{Y|X} E_{Z|X,Y} E_{S|X,Y,X}(Slogp_\theta(Y|X)) \tag{3}$$

where

$$E_{S|X,Y,X}(Slogp_\theta(Y|X)) = logp_\theta(Y|X)E_{S|X,Y,Z}(S) = logp_\theta(Y|X)(1 * P(S=1|X,Y))$$

plugging it into Eq.(3), we have

$$E_X E_{Y|X} E_{Z|X,Y}(logp_\theta(Y|X)P(S=1|X,Y))$$

since the interior is not related to Z, and the integral of Z is equal to 1, we have

$$E_X E_{Y|X}(logp_\theta(Y|X)P(S=1|X,Y))$$

Similarly to the procedure in problem 1, maximizing $l_n^1\theta$ is asymptotically maximizing

$$E_X(-\int log\frac{p_*(Y|X)}{p_\theta(Y|X)}p_*(Y|X)p(S=1|X,Y)dy)$$

i.e minimizing

$$E_X(\int log\frac{p_*(Y|X)}{p_\theta(Y|X)}p(S=1|X,Y)dy)$$

The result is different from that in Problem 1.

3 Problem 3

Similarly to Problem 2, we first transform $l_n^{\pi}(\theta)$

$$l_n^{\pi}(\theta) = \frac{1/n \sum_{i=1}^{n} S_i logp_{\theta}(Y_i|X_i)/\pi(Y_i, X_i, Z_i)}{n_1/n}$$

and it is asymptotically approaches to

$$\frac{E_G(S\frac{logp_{\theta}(Y|X)}{\pi(Y,X,Z)})}{E_G(S)}$$

where

$$E_G(S\frac{log_{\theta}(Y|X)}{\pi(Y,X,Z)}) = E_{X,Y,Z,S}(S\frac{log_{\theta}(Y|X)}{\pi(Y,X,Z)})$$

We apply Eq.(2) here again, the above can be thus expressed as

$$E_X E_{Y|X} E_{Z|X,Y} E_{S|X,Y,Z}(S\frac{log_{\theta}(Y|X)}{\pi(Y,X,Z)}) \tag{4}$$

Since the last conditional expectation is only related to S, we can calculated it as

$$E_{S|X,Y,Z}(S\frac{log_{\theta}(Y|X)}{\pi(Y,X,Z)}) = \frac{log_{\theta}(Y|X)}{\pi(Y,X,Z)}(1{*}p(S=1|X,Y,Z)+0{*}p(S=0|X,Y,Z)) = \frac{log_{\theta}(Y|X)}{\pi(Y,X,Z)}\pi(Y,X,Z)$$

plugging in into Eq.(4), we have

$$E_X E_{Y_X} E_{Z|X,Y}(log_{\theta}(Y|X)) = E_X E_{Y|X}(log_{\theta}(Y|X))$$

which is exactly the same as in Problem 1, so the following steps can be seen in Problem 1, and we can derive the conclusion here: maximizing $l_n^{\pi}(\theta)$ asymptotically minimizes the mean conditional K-L divergence $E_X(KL(p_*, p_{\theta}|X))$.

4 Problem 4

We first assume n_1 is not related to θ, and if we want to maximize $l_n^{\pi}(\theta)$, we can take derivative to it

$$G_n(\theta) = \frac{\partial}{\partial\theta} l_n^{\pi}(\theta) = \frac{1}{n_1}\sum_{i=1}^{n} \frac{S_i}{\pi(Y_i, X_i, Z_i)} \frac{\partial}{\partial\theta} logp_{\theta}(Y_i|X_i)$$

hence

$$G_n(\theta) = \frac{1}{n_1}\sum_{i=1}^{n} \frac{S_i}{\pi(Y_i, X_i, Z_i)} \frac{\partial}{\partial\theta} logp_{\theta}(Y_i|X_i)$$

Suppose $\widehat{\theta}_{\pi}$ can maximize the log-likelihood, so $G_n(\widehat{\theta}_{\pi}) = 0$. If the assumed model is the true model, then the estimating function is unbiased. Then we find this $\widehat{\theta}_{\pi}$.
Case1: If the equation system $G_n(\theta) = 0$ is a closed form, then we can find the solution by solving the equation of system directly.

Case2: If the equation system is not a closed form, we can apply numerical approach to find the approximate solution, we take Newton-Raphson method as an example:
We first randomly choose a initial value for $\theta^{(0)}$, then find the approximation by the iterated formula

$$\theta^{(t+1)} = \theta^{(t)} - G_n^{'}(\theta^{(t)})^{-1}G_n(\theta^{(t)})$$

Next we find asymptotic distribution of $\hat{\theta}_\pi$. Based on Result 2.1 (i.i.d case), we have

$$\sqrt{n_1}(\hat{\theta}_\pi - \theta) \overset{a.s}{\to} N(0, A^{-1}B(A^{-1})^T)$$

where A and B can be calculated in an empirically way

$$\hat{A}_n = \frac{1}{n_1}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}G(\theta) = \frac{1}{n_1}\sum_{i=1}^{n}\frac{S_i}{\pi(Y_i, X_i, Z_i)}\frac{\partial^2}{\partial\theta^2}logp_\theta(Y_i|X_i)$$

and

$$\hat{B}_n = \frac{1}{n_1}\sum_{i=1}^{n}\frac{S_i^2}{\pi^2(Y_i, X_i, Z_i)}\frac{\partial}{\partial\theta}logp_\theta(Y_i|X_i)(\frac{\partial}{\partial\theta}logp_\theta(Y_i|X_i))^T$$

Then

$$\sqrt{n_1}(\hat{\theta}_\pi - \theta) \overset{a.s}{\to} N(0, \hat{A}_n^{-1}\hat{B}_n(\hat{A}_n^{-1})^T)$$

We denote $\hat{\Sigma} = \hat{A}_n^{-1}\hat{B}_n(\hat{A}_n^{-1})^T$ and suppose $\theta_* = (\theta_1, \theta_2, ..., \theta_p)^T$ is a p-dimensional vector, then for each element, we have

$$\sqrt{n_1}(\hat{\theta}_j - \theta_j) \overset{a.s}{\to} N(0, \hat{\Sigma}_{jj})$$

where $\hat{\Sigma}_{jj}$ is the j-th element in the diagonal of $\hat{\Sigma}$. Hence the confidence interval at α level is given as

$$(\hat{\theta}_j - \frac{Z_{\frac{\alpha}{2}}}{\sqrt{n_1}}\sqrt{\hat{\Sigma}_{jj}},\ \ \hat{\theta}_j + \frac{Z_{\frac{\alpha}{2}}}{\sqrt{n_1}}\sqrt{\hat{\Sigma}_{jj}})$$

5 Problem 5

(a)
We first find the MLE, i.e. minimizing the log-likelihood

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\max_{\beta_0,\beta_1} l_n^1(\beta_0, \beta_1)$$

where $l_n^1 = \sum_{i=1}^{n} S_i logp_\beta(y_i|x_i)$ and $y_i|x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ hence we just need to minimize

$$l = \sum_{i=1}^{n} S_i(y_i - \beta_0 - \beta_1 x_i)^2$$

Taking the derivative to it in terms of β_0 and β_1 and then set be 0, we have

$$\frac{\partial}{\partial\beta_0}l = 2\sum_{i=1}^{n} S_i(y_i - \beta_0 - \beta_1 x_i)(-1) = 0$$

$$\frac{\partial}{\partial \beta_1} l = 2\sum_{i=1}^{n} S_i(y_i - \beta_0 - \beta_1 x_i)(-x_i) = 0$$

This equation of system can be expressed as

$$\sum_{i=1}^{n} S_i y_i = \beta_0 \sum_{i=1}^{n} S_i + \beta_1 \sum_{i=1}^{n} x_i S_i \tag{5}$$

$$\sum_{i=1}^{n} S_i x_i y_i = \beta_0 \sum_{i=1}^{n} S_i x_i + \beta_1 \sum_{i=1}^{n} S_i x_i^2 \tag{6}$$

then MLE can be computed as

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} S_i \sum_{i=1}^{n} S_i x_i y_i - \sum_{i=1}^{n} S_i x_i \sum_{i=1}^{n} S_i y_i}{\sum_{i=1}^{n} S_i x_i^2 \sum_{i=1}^{n} S_i - \sum_{i=1}^{n} S_i x_i \sum_{i=1}^{n} S_i x_i}$$

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} S_i y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i S_i}{\sum_{i=1}^{n} S_i}$$

Then from the equation system Eq.(5) and Eq.(6) above, we divide n on both sides

$$\frac{1}{n}\sum_{i=1}^{n} S_i y_i = \beta_0 \frac{1}{n}\sum_{i=1}^{n} S_i + \beta_1 \frac{1}{n}\sum_{i=1}^{n} S_i x_i$$

$$\frac{1}{n}\sum_{i=1}^{n} S_i x_i y_i = \beta_0 \frac{1}{n}\sum_{i=1}^{n} S_i x_i + \beta_1 \frac{1}{n}\sum_{i=1}^{n} S_i x_i^2$$

and by L.L.N, we have

$$\frac{1}{n}\sum_{i=1}^{n} S_i y_i \xrightarrow{p} E(Sy)$$

$$\frac{1}{n}\sum_{i=1}^{n} S_i \xrightarrow{p} E(S)$$

$$\frac{1}{n}\sum_{i=1}^{n} S_i x_i \xrightarrow{p} E(Sx)$$

$$\frac{1}{n}\sum_{i=1}^{n} S_i x_i y_i \xrightarrow{p} E(Sxy)$$

$$\frac{1}{n}\sum_{i=1}^{n} S_i x_i^2 \xrightarrow{p} E(Sx^2)$$

Hence now the equation of system is

$$E(Sy) = \tilde{\beta}_0 E(S) + \tilde{\beta}_1 E(Sx) \tag{7}$$

$$E(Sxy) = \tilde{\beta}_0 E(Sx) + \tilde{\beta}_1 E(Sx^2) \tag{8}$$

Then we will find each element in Eq.(7) and Eq.(8).

$$E_{X,Y,Z,S}(S) = E_Z E_{S|X,Y,Z}(S) = E_Z(\tau I(z_1 = z_2) + \rho I(z_1 \neq z_2))$$

which is equal to $0.9 * 0.01 + 0.1 * 0.09 = 0.018$

$$E_{X,Y,Z,S}(Sy) = E_Z E_{X|Z} E_{Y|X,Z} E_{S|X,Y,Z}(Sy) = E_Z E_{X|Z} E_{Y|X,Z}[y(\tau I(z_1 = z_2) + \rho I(z_1 \neq z_2))]$$

where $E_{Y|X,Z}(y) = \mu_i = x_i(I(Z_{1i} = Z_{2i}) - I(Z_{1i} \neq Z_{2i}))$ and therefore $E_{X|Z}(x_i) = -1 + 2z_{1i}$, hence we have

$$E_{X,Y,Z,S}(Sy) = E_Z((-1 + 2z_{1i})(\tau I(Z_{1i} = Z_{2i}) - \rho I(Z_{1i} \neq Z_{2i})))$$

The above expectation can be calculated under four conditions:(1) $Z_1 = Z_2 = 1$, (2) $Z_1 = Z_2 = 0$, (3) $Z_1 = 1, Z_2 = 0$, (4) $Z_1 = 0, Z_2 = 1$, and

$$p(Z_1, Z_1) = p(Z_2|Z_1)p(Z_1) = (0.1 + 0.8Z_1)^{Z_2}(0.9 - 0.8Z_1)^{1-Z_2} 0.5^{Z_1} 0.5^{1-Z_1}$$

so it is easy to find $p(Z_1 = 1, Z_2 = 1) = p(Z_1 = 0, Z_2 = 0) = 0.45$ and $p(Z_1 = 1, Z_2 = 0) = p(Z_1 = 1, Z_2 = 0) = 0.05$, then we can find

$$E_{X,Y,Z,S}(Sy) = 0.45\tau - 0.45\tau + 0.05\rho - 0.05\rho = 0$$

Similarly, we have $E(Sx) = 0$ and $E(Sxy) = 0$ and $E(Sx^2) = 0.036$. Finally, the equation of system (7) and (8) can be

$$0 = \tilde{\beta}_0 0.018 + \tilde{\beta}_1 0$$
$$0 = \tilde{\beta}_0 0 + \tilde{\beta}_1 0.036$$

Apparently, $(\tilde{\beta}_0, \tilde{\beta}_1)^T = (0, 0)^T$.

(b)

Since we want to find the true distribution of $Y_i|X_i$ (at the super population level), we need to find

$$f(y_i|x_i) = \frac{f(y_i, x_i)}{f(x_i)}$$

which is equivalent to

$$\frac{\sum_{Z_{1i}} \sum_{Z_{2i}} f(x_i, y_i, z_{1i}, z_{2i})}{\sum_{Z_{1i}} \sum_{Z_{21}} f(x_i, z_{1i}, z_{2i})} \tag{9}$$

Since

$$f(x, y, z_1, z_2) = f(y|x, z_1, z_2) f(x, z_1, z_2)$$

and

$$f(x, z_1, z_2) = f(x|z_1, z_2) p(z_2) | p(z_1) p(z_1)$$

Eq.(9) can be written as

$$\frac{\sum_{Z_{1i}} \sum_{Z_{2i}} f(y_i|x_i, z_{1i}, z_{2i}) f(x_i|z_{1i}, z_{2i}) p(z_{2i})|p(z_{1i}) p(z_{1i})}{\sum_{Z_{1i}} \sum_{Z_{2i}} f(x_i|z_{1i}, z_{2i}) p(z_{2i})|p(z_{1i}) p(z_{1i})} \tag{10}$$

where $f(x_i|z_{1i}, z_{2i}) = f(x_i|z_{1i})$ due to $x_i \perp\!\!\!\perp z_{2i}|z_{1i}$ and

$$y_i|x_i, z_{1i}, z_{2i} \sim Normal(\mu_i, \sigma_i^2)$$

$$x_i|z_{1i} \sim Normal(-1 + 2z_{1i}, 1)$$

$$z_{2i}|z_{1i} \sim Bernoulli(0.1 + 0.8z_{1i})$$

$$z_{1i} \sim Bernoulli(0.5)$$

Based on these distribution and conditional distribution, we will calculated the numerator and denominator in Eq.(6) separately.
We first find its numerator *numerator*,

$$numerator = a + b + c + d$$

where a is under the condition that $z_{1i} = 1$ and $z_{2i} = 1$ in terms of the numerator.
b is under the condition that $z_{1i} = 0$ and $z_{2i} = 0$ in terms of the numerator.
c is under the condition that $z_{1i} = 1$ and $z_{2i} = 0$ in terms of the numerator.
d is under the condition that $z_{1i} = 0$ and $z_{2i} = 1$ in terms of the numerator.
By simplifying the equation, we have

$$numerator = numerator_1 numerator_2$$

where

$$numerator_1 = [exp(-1/2(x_i - 1)^2) + exp(-1/2(x_i + 1)^2)]$$

$$numerator_2 = [\frac{0.45}{2\pi exp(0.2|x_i|)} exp(-1/2\frac{(y_i - x_i)^2}{exp(0.4|x_i|)}) + \frac{0.05}{2\pi exp(0.2|x_i|)} exp(-1/2\frac{(y_i + x_i)^2}{exp(0.4|x_i|)})]$$

By a similar way as we find *numerator*, the denominator *denominator* can be computed as follows

$$denominator = \frac{1}{2\sqrt{2\pi}}[exp(-1/2(x_i - 1)^2) + exp(-1/2(x_i + 1)^2)]$$

Finally we have

$$f(y_i|x_i) = \frac{numerator}{denominator}$$

which is equal to

$$\frac{0.9}{\sqrt{2\pi} exp(0.2|x_i|)} exp(-\frac{1}{2}\frac{(y_i - x_i)^2}{exp(0.4|x_i|)}) + \frac{0.1}{\sqrt{2\pi} exp(0.2|x_i|)} exp(-\frac{1}{2}\frac{(y_i + x_i)^2}{exp(0.4|x_i|)})$$

Here we find the true distribution of $y_i|x_i$.

(c)

The only difference of (a) and (c) is that (a) is balanced sampling, while (c) is imbalanced sampling, i.e. in (c), we have $\rho = \tau$, specifically, we find the expectations in (a) again with new parameters.

$$E(S) = 0.9 * \tau + 0.1 * \rho = \rho$$
$$E(Sy) = 0$$
$$E(Sxy) = 2(0.9 * \tau - 0.1 * \rho) = 1.6\rho$$
$$E(Sx) = 0$$
$$E(Sx^2) = 2(0.1\rho + 0.9\rho) = 2\rho$$

Then the equation of system can be

$$0 = \rho\beta_0^* + 0\beta_1^*$$
$$1.6\rho = 0\beta_0^* + 2\rho\beta_1^*$$

Solving this equation of system, we find the $\beta^* = (\beta_0^*, \beta_1^*)^T = (0, 0.8)^T$.

(d)

The estimation strategy devised in Problem 4 is as follows:

$$G_n(\theta) = \frac{1}{n_1}\sum_{i=1}^{n} \frac{S_i}{\pi(Y_i, X_i, Z_i)} \frac{\partial}{\partial \theta} logp_\theta(y_i|x_i)$$

where $y_i|x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$, and θ here is $\theta = (\beta_0, \beta_1, \sigma^2)^T$ and

$$p_\theta(y_i|x_i) = \frac{1}{\sqrt{2\pi}\sigma} exp[-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2]$$

hence

$$G_n(\theta) = \frac{1}{n_1}\sum_{i=1}^{n} \frac{S_i}{\pi_i} \frac{\partial}{\partial \theta}[-\frac{1}{2}log\sigma^2 + \frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2]$$

We first find the derivatives inside of $G_n(\theta)$, and we denote them as $\frac{\partial}{\partial \beta_0} l(\beta_0, \beta_1, \sigma^2)$, $\frac{\partial}{\partial \beta_1} l(\beta_0, \beta_1, \sigma^2)$ and $\frac{\partial}{\partial \sigma^2} l(\beta_0, \beta_1, \sigma^2)$

$$\frac{\partial}{\partial \beta_0} l(\beta_0, \beta_1, \sigma^2) = \frac{1}{\sigma^2}(y_i - \beta_0 - \beta_1 x_i) \tag{11}$$

$$\frac{\partial}{\partial \beta_1} l(\beta_0, \beta_1, \sigma^2) = \frac{1}{\sigma^2}(y_i x_i - \beta_0 x_i - \beta_1 x_i^2) \tag{12}$$

$$\frac{\partial}{\partial \sigma^2} l(\beta_0, \beta_1, \sigma^2) = -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(y_i - \beta_0 - \beta_1 x_i)^2 \tag{13}$$

Then $G_n(\theta)$ can be expressed as

$$G_n(\theta) = \frac{1}{n_1} \sum_{i=1}^{n} \frac{S_i}{\pi_i} \begin{pmatrix} \frac{1}{\sigma^2}(y_i - \beta_0 - \beta_1 x_i) \\ \frac{1}{\sigma^2}(y_i x_i - \beta_0 x_i - \beta_1 x_i^2) \\ -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(y_i - \beta_0 - \beta_1 x_i)^2 \end{pmatrix}$$

Based on Eq.(11), (12) and (13), and set $G_n(\theta) = 0$ we have

$$\sum_{i=1}^{n} \frac{S_i}{\pi_i} y_i = \beta_0 \sum_{i=1}^{n} \frac{S_i}{\pi_i} + \beta_1 \sum_{i=1}^{n} \frac{S_i}{\pi_i} x_i \tag{14}$$

$$\sum_{i=1}^{n} \frac{S_i}{\pi_i} x_i y_i = \beta_0 \sum_{i=1}^{n} \frac{S_i}{\pi_i} x_i + \beta_1 \sum_{i=1}^{n} \frac{S_i}{\pi_i} x_i^2 \tag{15}$$

$$\sum_{i=1}^{n} \frac{S_i}{\pi_i} \frac{1}{\sigma^2} (y_i - \beta_0 - \beta_1 x_i)^2 = \sum_{i=1}^{n} \frac{S_i}{\pi_i} \tag{16}$$

By solving the above equation, we have

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} \frac{S_i}{\pi_i} y_i - \hat{\beta}_1 \sum_{i=1}^{n} \frac{S_i}{\pi_i} x_i}{\sum_{i=1}^{n} \frac{S_i}{\pi_i}} \tag{17}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \frac{S_i}{\pi_i} x_i y_i \sum_{i=1}^{n} \frac{S_i}{\pi_i} - \sum_{i=1}^{n} \frac{S_i}{\pi_i} x_i \sum_{i=1}^{n} \frac{S_i}{\pi_i} y_i}{\sum_{i=1}^{n} \frac{S_i}{\pi_i} x_i^2 \sum_{i=1}^{n} \frac{S_i}{\pi_i} - \sum_{i=1}^{n} \frac{S_i}{\pi_i} x_i \sum_{i=1}^{n} \frac{S_i}{\pi_i} x_i} \tag{18}$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \frac{S_i}{\pi_i} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\sum_{i=1}^{n} \frac{S_i}{\pi_i}} \tag{19}$$

In order to find the confidence interval, we need to calculate the sandwich estimation,

$$\hat{A}_n = \frac{1}{n_1} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} G(\theta)$$

it is equal to

$$\frac{1}{n_1} \sum_{i=1}^{n} \frac{S_i}{\pi_i} \begin{pmatrix} -\frac{1}{\sigma^2} & -\frac{1}{\sigma^2} x_i & -\frac{1}{\sigma^4}(y_i - \beta_0 - \beta_1 x_i) \\ -\frac{1}{\sigma^2} x_i & -\frac{1}{\sigma^2} x_i^2 & -\frac{1}{\sigma^4}(y_i x_i - \beta_0 x_i - \beta_1 x_i^2) \\ -\frac{1}{\sigma^4}(y_i - \beta_0 - \beta_1 x_i) & -\frac{1}{\sigma^4}(y_i x_i - \beta_0 x_i - \beta_1 x_i^2) & \frac{1}{2\sigma^4} - \frac{1}{\sigma^6}(y_i - \beta_0 - \beta_1 x_i)^2 \end{pmatrix}$$

we then plug in the MLE to get

$$\hat{A}_n = \frac{1}{n_1} \sum_{i=1}^{n} \frac{S_i}{\pi_i} \begin{pmatrix} -\frac{1}{\hat{\sigma}^2} & -\frac{1}{\hat{\sigma}^2} x_i & 0 \\ -\frac{1}{\hat{\sigma}^2} x_i & -\frac{1}{\hat{\sigma}^2} x_i^2 & 0 \\ 0 & 0 & \frac{1}{2\hat{\sigma}^4} - \frac{1}{\hat{\sigma}^6}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \end{pmatrix}$$

Then we find $\hat{B}_n$

$$\hat{B}_n = \frac{1}{n_1}\sum_{i=1}^{n}\frac{S_i^2}{\pi_i^2}G(\theta)G(\theta)^T$$

plugging the MLE, we have

$$\hat{B}_n = \frac{1}{n_1}\sum_{i=1}^{n}\frac{S_i^2}{\pi_i^2}\begin{pmatrix} \tfrac{1}{\hat{\sigma}^4}(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)^2 & \tfrac{1}{\hat{\sigma}^4}(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)(y_ix_i-\hat{\beta}_0x_i-\hat{\beta}_1x_i^2) & *_1 \\ . & \tfrac{1}{\hat{\sigma}^4}(y_ix_i-\hat{\beta}_0x_i-\hat{\beta}_1x_i^2)^2 & *_2 \\ . & . & *_3 \end{pmatrix}$$

(note:we use . because B is a symmetric matrix)
where

$$*_1 = -\frac{1}{2\hat{\sigma}^4}(y_i-\hat{\beta}_0-\hat{\beta}_1x_i) + \frac{1}{2\hat{\sigma}^6}(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)^3$$

$$*_2 = (y_ix_i-\hat{\beta}_0x_i-\hat{\beta}_1x_i^2)(-\frac{1}{2\hat{\sigma}^4} + \frac{1}{2\hat{\sigma}^6}(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)^2)$$

$$*_3 = (-\frac{1}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4}(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)^2)^2$$

Then the asymptotically variance for $\hat{\theta}$ is

$$\hat{\Sigma} = \hat{A}_n^{-1}\hat{B}_n(\hat{A}_n^{-1})^T$$

Then we can find the confidence interval of β_1^* as

$$(\hat{\beta}_1 - \frac{Z_{\frac{\alpha}{2}}}{\sqrt{n_1}}\sqrt{\hat{\Sigma}_{22}},\ \ \hat{\beta}_1 + \frac{Z_{\frac{\alpha}{2}}}{\sqrt{n_1}}\sqrt{\hat{\Sigma}_{22}})$$

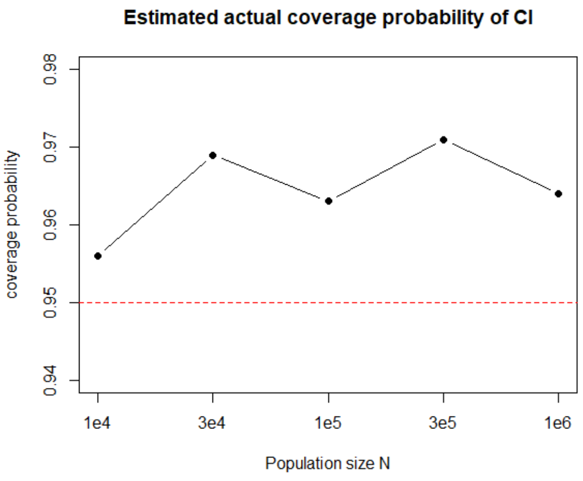

Figure 1: The coverage probability

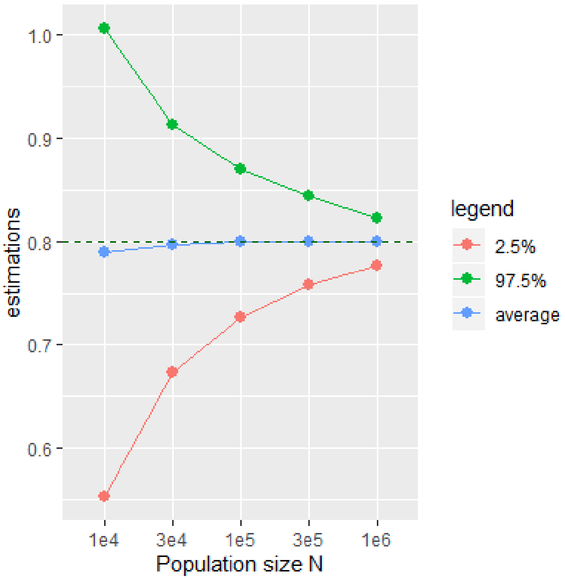

Figure 2: the average and 2.5 and 97.5 percentiles of estimator

We hope this confidence interval can hold the true value i.e. 0.8. The complete implementation in R can be seen in Appendix. Here we display two plots: Fig.1: the estimated actual coverage probability of confidence interval with different N. The horizontal line at 0.95 is for reference (red); Fig.2: the average and 2.5 and 97.5 percentiles of estimates as a function of N and a horizontal line at the true value of β_1^* is in darkgreen.
From Fig.1, all of the coverage probabilities are over 0.95, although they fluctuate around 0.965.
Fro Fig.2, we find that the average, 2.5 percent and 97.5 percent quantiles of estimates are gradually close to the true value (0.8) with the growth of population size N.

6 Extra

1. Lecture Notes Chapter 2 p102 line 7 instead of $l_n(\lambda)$ it should be "$I(\theta)$".
2. Lecture Notes Chapter 5B p15 line -1 instead of $B(\beta) = (X^TX)^{-1}\sigma^2$, it should be $B(\beta) = (X^TX)\sigma^2$.
3. Book p54 line 13, 18, instead of "(2.36) and (3.36)", it should be "(2.36) and (2.37)".

7 Appendix

Simulations:
step1: we simulate data by "musample" in R:

```
mysample <- function(N) {
  # z1
  Z1 <- rbern(N, 0.5)
  # z2
  Z2 <- sapply(Z1, function(x) rbern(1, 0.1 + 0.8 *x )     )
  # x
  X <- rnorm(N, -1+2*Z1, 1)
  # Y
  mu <- X * ifelse(Z1==Z2 ,1,0) - X*ifelse(Z1!=Z2, 1,0)
  sd <- exp(0.2*abs(X))
  Y <- rnorm(N, mu, sd)
  # Pi
  PI <- 0.01 * ifelse(Z1==Z2, 1, 0) + 0.09 * ifelse(Z1!=Z2, 1,0)
  # S
  S <- rbern(N, PI)
  sampling <- data.frame(X=X, Y=Y,Z1=Z1,Z2=Z2, PI=PI, S=S)
  # deleting the data when S=1
  sampling <- sampling[which(sampling$S == 1), ]
  return (sampling)
}
```

step2: we find the estimator of β and σ^2 and $\hat{y}$ by $findparameters$ in R

```
find_parameters <- function(data){
  # beta_0 and beta_1
  m_left <- matrix(NA,2,2)
  m_left[1,1] <- sum(1/data$PI)
  m_left[1,2] <- sum(data$X/data$PI)
  m_left[2,1] <-  sum(data$X/data$PI)
  m_left[2,2] <- sum((data$X)^2/data$PI)
  m_right <- matrix(NA, 2,1)
  m_right[1,1] <-  sum(data$Y/data$PI)
  m_right[2,1] <-  sum(data$X*data$Y/data$PI)
  # hat beta
  beta <- solve(m_left) %*% m_right
  # yhat
  X  <- cbind(rep(1, dim(data)[1])  , data$X)
  yhat <- X %*% beta
  # estimate sigma^2
  numerator <- sum(1/data$PI * (data$Y - yhat)^2)
  denominator <- sum(1/data$PI)
  sigma2 <- numerator / denominator
  para <- data.frame(beta0 = beta[1], beta1 = beta[2], sigma2 = sigma2)
  return (list(para=para, yhat=yhat))
}
```

step3: we find the sandwich estimator and confidence interval

```
# sandwich variance
```

```
sandiwch_estimation <- function(data, parameters) {
  para <- parameters$para
  yhat <- parameters$yhat
  # A
  A <- matrix(NA, 3,3)
  A[1,1] <- 1/dim(data)[1] * sum(1/ data$PI * (- 1 / para$sigma2) )
  A[1,2] <- 1/dim(data)[1] * sum(1/ data$PI * (- data$X / para$sigma2) )
  A[2,1] <- A[1,2]
  A[1,3] <- 0
  A[3,1] <- 0
  A[2,2] <- 1/dim(data)[1] * sum(1/ data$PI * (- (data$X)^2 / para$sigma2)
)
  A[2,3] <- 0
  A[3,2] <- A[2,3]
  A[3,3] <- 1/dim(data)[1] * sum(1/ data$PI * ( 1/ (2* (para$sigma2)^2 ) - 1 /
(para$sigma2)^3 * (data$Y - yhat)^2 ))

  # B
  B <- matrix(NA, 3,3)
  B[1,1] <- 1/dim(data)[1] * sum(1/ (data$PI)^2 *  (1/ (para$sigma2)^2) * (data$Y - yha
  B[1,2] <- 1/dim(data)[1] * sum(1/ (data$PI)^2 *  (1/ (para$sigma2)^2) * (data$Y - yha
  B[2,1] <- B[1,2]
  B[1,3] <- 1/dim(data)[1] * sum(1/ (data$PI)^2 * (- 1 / (2*(para$sigma2)^2) * (data$Y -
+  1 / (2*(para$sigma2)^3) * ( data$Y - yhat)^3  ) )
  B[3,1] <- B[1,3]
  B[2,2] <- 1/dim(data)[1] * sum(1/ (data$PI)^2 * 1 / (para$sigma2)^2 * (data$Y - yhat)
)
  B[2,3] <- 1/dim(data)[1] * sum(1/ (data$PI)^2 * ( - 1 / (2*(para$sigma2)^2) * (data$Y
+  1 / (2*(para$sigma2)^3) * data$X * (data$Y -yhat)^3 )     )
  B[3,2] <- B[2,3]
  B[3,3] <- 1/dim(data)[1] * sum(1/ (data$PI)^2 *  (-1 / (2*para$sigma2)
+ 1 / (2*(para$sigma2)^2) * (data$Y - yhat)^2    )^2  )
  # sandwich
  A_inv <- solve(A)
  var_estimate <- A_inv %*% B %*% t(A_inv)
  # estimated variance of beta_1
  var_beta_1_estimate <- var_estimate[2,2]
  return (var_beta_1_estimate)
}
var_b1 <- sandiwch_estimation(data, parameters)
CI_lower <- parameters$para$beta1 - 1.96 / sqrt(dim(data)[1]) * sqrt(var_b1)
CI_upper <- parameters$para$beta1 + 1.96 / sqrt(dim(data)[1]) * sqrt(var_b1)
c(CI_lower, CI_upper)
```

step4: corresponding to different N, we find the coverage probability, the average and quantiles of estimators.

```
# Simulations
count <- 0
N = 1000000
iters <- 1000
betas <- matrix(NA, iters,1)
for (i in 1:iters) {

  if (i %% 10 == 0) {
    print(i)
  }
  data <- mysample(N)
  parameters <- find_parameters(data)
  hat_beta1 <- parameters$para$beta1
  betas[i] <- hat_beta1
  var_b1 <- sandiwch_estimation(data, parameters)
  CI_lower <- hat_beta1 - 1.96 / sqrt(dim(data)[1]) * sqrt(var_b1)
  CI_upper <- hat_beta1 + 1.96 / sqrt(dim(data)[1]) * sqrt(var_b1)
  if (CI_lower <= 0.8 & CI_upper >= 0.8) {
    count = count + 1
  }
}
print(N)
print(count)
print(mean(betas))
print(quantile(betas, c(0.025,0.975)))
```

step5: plot

```
# plot 1: coverage probability
library(ggplot2)
Names <- c("1e4", "3e4", "1e5", "3e5", "1e6")
prob <- c(0.972, 0.969, 0.980, 0.981, 0.978)

plot(prob, xlab = "Population␣size␣N", ylab="coverage␣probability", ylim = c(0.94,1) ,c
     main = "Estimated␣actual␣coverage␣probability␣of␣CI", xaxt="n")
abline(h=0.95, lty=2, col="red")
axis(1,at=seq(1,5,1),label=Names)
# plot 2: the average and 2.5 and 97.5 percentiles of my point estimates
legend <- rep(c("average", "2.5%", "97.5%"),each = 5)
num <- rep(1:5, times=3)
beta_val <- c(c(0.7895, 0.7961, 0.7995, 0.7999, 0.8001), c(0.5528, 0.6735, 0.7268, 0.75
                    c(1.0063, 0.912, 0.8693,0.8436, 0.8229)    )
d <- data.frame(order=num, type = type, beta = beta_val)
ggplot(data = d, mapping = aes(x = order, y = beta, colour = legend)) + geom_line() + g
  labs(x = "Population␣size␣N", y = "estimations") +
  scale_x_discrete(limits=c("1e4", "3e4", "1e5", "3e5", "1e6")) +
geom_hline(aes(yintercept=0.8), colour=”darkgreen”, linetype=”dashed”)
```