Mathematical Foundations of Biostatistics

Assignment 2

The assignment is due Wednesday October 16 by 11:59pm. Questions Q1, Q2 and Q3 have the weights of 25%, 25% and 50%, respectively.

1. (a) Let $\mu_n = 1/n$ for $n = 1, 2, \ldots$ and let $X_n \sim \text{Poisson}(\mu_n)$.
   i. Show that $X_n \xrightarrow{p} 0$.
   ii. Let $Y_n = nX_n$. Show that $Y_n \xrightarrow{p} 0$.

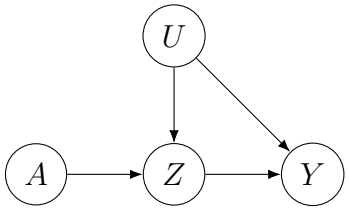
   (b) Suppose we have a computer program consisting of $n = 100$ pages of code. Let X_i be the number of errors on the ith page of code. Suppose that the X_i's are Poisson with mean 1 and that they are independent. Let $Y = \sum_{i=1}^{n} X_i$ be the total number of errors. Use the central limit theorem to approximate $P(Y < 90)$.

   (c) Let $X_1, \ldots, X_n$ be IID Poisson random variables with mean μ.
   i. Find the limiting distribution of $\sqrt{n}(\log(\overline{X}_n) - \log(\mu))$.
   ii. Find a function g such that $\sqrt{n}(g(\overline{X}_n) - g(\mu)) \xrightarrow{d} N(0, 1)$ (known as a variance stabilizing transformation).

2. Consider the data-generating mechanism

$$Y_i = \theta Z_i + U_i + \varepsilon_i$$
$$Z_i = 0.5A_i + U_i + \tau_i,$$

   where ε_i, τ_i, A_i and U_i are independent and standard normal distributed. We are interested in estimating the parameter θ, and consider the situation where the observed data are (a_i, y_i, z_i), $i = 1, \ldots, n$, but where U_i is unobserved, so we cannot adjust for it in the analysis. The situation can be illustrated through the below DAG.



Olli Saarela, Dalla Lana School of Public Health, 155 College Street, 6th floor, Toronto, Ontario M5T 3M7, Canada. E-mail: olli.saarela@utoronto.ca

Suppose that, because U_i is unobserved, we fit the model

$$Y_i = \theta_0^* + \theta_1^* Z_i + \varepsilon_i^*,$$

to use the ordinary least squares (OLS) estimator

$$\widehat{\theta}_1^* = \frac{\sum_{i=1}^n (Y_i - \overline{Y})(Z_i - \overline{Z})}{\sum_{i=1}^n (Z_i - \overline{Z})}$$

as an estimator of θ.

(a) Find $\text{plim}_{n\to\infty} \widehat{\theta}_1^*$. Under which circumstances $\widehat{\theta}_1^*$ is a consistent estimator of θ? Is this true in the present setting? How do you interpret the result? What is this phenomenon called? (Hint: you can assume that sample variance and covariance are consistent estimators of their theoretical counterparts, and make use the properties of theoretical covariance.)

(b) An alternative estimator is obtained by taking $\widetilde{\theta} = \widehat{\phi}_1 / \widehat{\psi}_1$, where $\hat{\phi}_1$ and $\hat{\psi}_1$ are the OLS estimators of coefficients ϕ_1 and ψ_1 in the models

$$Y_i = \phi_0 + \phi_1 A_i + \kappa_i$$
$$Z_i = \psi_0 + \psi_1 A_i + \xi_i.$$

Find $\text{plim}_{n\to\infty} \widetilde{\theta}$. Under which circumstances $\widetilde{\theta}$ is a consistent estimator of θ? Is this true in the present setting? What is this estimator/method called?

3. Carry out and report a simulation study of the properties of (the sampling distributions of) $\widehat{\theta}_1^*$ and $\widetilde{\theta}$. The data generating mechanism is specified as above. Choose $\theta = 0.5$, $m = 1000$ and sample size scenarios $n = 250, 500$. Report mean, bias and standard deviation of the estimator, Monte Carlo error of the mean, mean standard error and 95% normal approximation confidence interval coverage probability. The standard error of $\widehat{\theta}_1^*$ may be obtained from linear regression. For $\widetilde{\theta}$ you can use the variance estimator

$$\widehat{V}[\widetilde{\theta}] = \frac{\widehat{V}[Y_i - \widetilde{\theta} Z_i]}{n \widehat{V}[Z_i] \widehat{\rho}_{A,Z}^2},$$

where $\widehat{\rho}$ is the sample correlation coefficient. Report also the bootstrap standard error, and coverage probability of 95% bootstrap normal approximation and bootstrap percentile confidence intervals. Which one of the two point estimators is better and why? How do the results compare to the theoretical ones in Q2?

1. (a) Let $\mu_n = 1/n$ for $n = 1, 2, \ldots$ and let $X_n \sim \text{Poisson}(\mu_n)$.
   i. Show that $X_n \xrightarrow{p} 0$.
   ii. Let $Y_n = nX_n$. Show that $Y_n \xrightarrow{p} 0$.

   (b) Suppose we have a computer program consisting of $n = 100$ pages of code. Let X_i be the number of errors on the ith page of code. Suppose that the X_i's are Poisson with mean 1 and that they are independent. Let $Y = \sum_{i=1}^{n} X_i$ be the total number of errors. Use the central limit theorem to approximate $P(Y < 90)$.

   (c) Let $X_1, \ldots, X_n$ be IID Poisson random variables with mean μ.
   i. Find the limiting distribution of $\sqrt{n}(\log(\overline{X}_n) - \log(\mu))$.
   ii. Find a function g such that $\sqrt{n}(g(\overline{X}_n) - g(\mu)) \xrightarrow{d} N(0, 1)$ (known as a variance stabilizing transformation).

a) i. Need to show $\lim_{n\to\infty} P(|X_n - 0| > \varepsilon) = 0$

Since $X_n \sim \text{Poisson}(\mu_n)$, then $E(X_n) = \frac{1}{n} = V(X_n)$

From Chebyshev's Inequality,

$$P(|X_n - 0| \geq \varepsilon) \leq \frac{V(X_n)}{\varepsilon^2}$$

$$= \frac{1/n}{\varepsilon^2} \longrightarrow 0 \quad \text{as} \quad n \to \infty$$

Thus, $X_n \xrightarrow{P} 0$ as needed. ∎

ii. We need to show $\lim_{n\to\infty} P(|Y_n - 0| > \varepsilon) = 0$.

Note $P(|Y_n| > \varepsilon) = P(nX_n > \varepsilon)$

$= P(X_n > \varepsilon/n)$

Since $X_i \sim \text{Poisson}(1/n)$, then $P(X_n = k) = \frac{(\frac{1}{n})^k e^{-\frac{1}{n}}}{k!}$

① $P(X_n = 1) = \frac{1}{n} e^{-\frac{1}{n}} < \frac{1}{n}$ as $n \to \infty$

Note $1/n \to 0$ as $n \to \infty$

② $P(X_n \geq 1) = 1 - P(X_n = 0) = 1 - e^{-\frac{1}{n}}$

From Taylor expansion

$$e^x = 1 - \frac{1}{n} + \frac{(-\frac{1}{n})^2}{2!} - \frac{(-\frac{1}{n})^3}{3!} + \cdots$$

$$\Rightarrow e^{-\frac{1}{n}} = 1 - \frac{1}{n} + O_n \quad \text{where } O_n \text{ is negligible}$$

$$\Rightarrow 1 - e^{-\frac{1}{n}} \approx \frac{1}{n}$$

So $P(X_n \geq 1) = 1 - e^{-\frac{1}{n}} \approx \frac{1}{n} \to 0$ as $n \to \infty$

③ $P(X_n = 0) = e^{-\frac{1}{n}} \approx 1 - \frac{1}{n} \to 1$ as $n \to \infty$

Then, $P(Y_n = 0) = P(nX_n = 0) = P(X_n = 0) \to 1$ as $n \to \infty$.

Hence, combining (1), (2), and (3) we can say

$$P(|Y_n - 0| > \varepsilon) = 1 - P(Y_n = 0) = 1 - 1 = 0$$

$$\Rightarrow Y_n \xrightarrow{P} 0$$

b) Let $Y = \sum X_i$. Since $X_i \sim \text{Poisson}(1)$, then

$Y \sim \text{Poisson}(100) \approx N(100, 100)$ by CLT. Then

$$P(Y < 90) = P\left(Z < \frac{90 - 100}{10}\right)$$

$$= P(Z < -1)$$

$$= 0.1587.$$

(c) Let $X_1, \ldots, X_n$ be IID Poisson random variables with mean μ.

i. Find the limiting distribution of $\sqrt{n}(\log(\overline{X}_n) - \log(\mu))$.

ii. Find a function g such that $\sqrt{n}(g(\overline{X}_n) - g(\mu)) \xrightarrow{d} N(0, 1)$ (known as a variance stabilizing transformation).

c) i. Recall $\sqrt{n}(\bar{x}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$ by CLT

↑ $V(X_i) = \mu$

Then by Delta Method,

$$\sqrt{n}\left[g(\bar{x}_n) - g(\mu)\right] \xrightarrow{d} N(0, (g'(\mu))^2\sigma^2)$$

$$\Rightarrow \sqrt{n}\left(\log(\bar{x}_n) - \log(\mu)\right) \xrightarrow{d} N\left(0, ((\log\mu)')^2\sigma^2\right)$$

$$= N(0, 1/\mu)$$

ii. Since $\sqrt{n}(\bar{x}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$ by CLT.

Then by Delta Method,

$$\sqrt{n}\left[g(\bar{x}_n) - g(\mu)\right] \xrightarrow{d} N(0, (g'(\mu))^2\sigma^2)$$

We want $(g'(x))^2\sigma^2 = 1$

Since $X_1, \ldots, X_n \overset{iid}{\sim}$ Poisson, $E(X_i) = V(X_i) = \mu$

So then $(g'(x))^2\mu = 1$

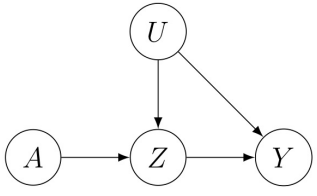
$$\Rightarrow g(x) = \int \frac{1}{\sqrt{x}}\, dx = 2\sqrt{x}$$

Hence $g(x) = 2\sqrt{x}$ is the variance stabilizing transformation

2. Consider the data-generating mechanism

$$Y_i = \theta Z_i + U_i + \varepsilon_i$$
$$Z_i = 0.5A_i + U_i + \tau_i,$$

where ε_i, τ_i, A_i and U_i are independent and standard normal distributed. We are interested in estimating the parameter θ, and consider the situation where the observed data are (a_i, y_i, z_i), $i = 1, \ldots, n$, but where U_i is unobserved, so we cannot adjust for it in the analysis. The situation can be illustrated through the below DAG.



Suppose that, because U_i is unobserved, we fit the model

$$Y_i = \theta_0^* + \theta_1^* Z_i + \varepsilon_i^*,$$

to use the ordinary least squares (OLS) estimator

$$\widehat{\theta}_1^* = \frac{\sum_{i=1}^n (Y_i - \overline{Y})(Z_i - \overline{Z})}{\sum_{i=1}^n (Z_i - \overline{Z})}$$

as an estimator of θ.

(a) Find $\text{plim}_{n\to\infty} \widehat{\theta}_1^*$. Under which circumstances $\widehat{\theta}_1^*$ is a consistent estimator of θ? Is this true in the present setting? How do you interpret the result? What is this phenomenon called? (Hint: you can assume that sample variance and covariance are consistent estimators of their theoretical counterparts, and make use the properties of theoretical covariance.)

a) As $n \to \infty$, from WLLN

$\sum (Y_i - \bar{Y})(z_i - \bar{z}) \xrightarrow{P} \text{cov}(Y, z)$ and
$\sum (z_i - \bar{z})^2 \xrightarrow{P} V(z)$

So, we have

$$\lim_{n\to\infty} \hat{\theta}_1^{*} = \lim_{n\to\infty} \frac{\sum (Y_i-\bar{Y})(z_i-\bar{z})}{\sum (z_i-\bar{z})^2} = \frac{cov(Y,z)}{V(z)}$$

Since $Y_i = \theta z_i + U_i + \varepsilon_i$ by the mechanism,
$z_i = 0.5A_i + U_i + \tau_i$

① Then

$$Cov(Y,z) = Cov(\theta z_i + U_i + \varepsilon_i, z_i)$$

$$= Cov(\theta z_i, z_i) + Cov(U_i, z_i) + Cov(\varepsilon_i, z_i)$$

$$= \theta V(z) + \underline{Cov(U_i, z_i)}$$

$$Cov(U_i, z_i) = Cov(U_i, 0.5A_i + U_i + \tau_i)$$

$$= \cancel{Cov(U_i, 0.5A_i)}^{\,0} + V(U_i) + cov(U_i, \tau_i)$$

$$= V(U_i)$$

$$= 1 \qquad \text{as } U_i \sim N(0,1)$$

So $Cov(Y,z) = \theta V(z) + 1$

② Now, $V(z_i) = V(0.5A_i + U_i + \tau_i)$

$$\overset{ind}{=} (0.5)^2 V(A_i) + V(V_i) + V(I_i)$$

$$= 0.25 + 1 + 1 \quad \text{as } A_i, V_i, I_i \sim N(0,1)$$

$$= 2.25$$

Hence, $\hat{\theta}_1^* = \dfrac{\sum (y_i - \bar{y})(z_i - \bar{z})}{\sum (z_i - \bar{z})^2}$

$$\approx \frac{cov(Y,z)}{V(z)} = \frac{\theta V(z)}{V(z)} + \frac{V(U)}{V(z)}$$

$$= \frac{2.25\theta + 1}{2.25}$$

$$= \theta + \frac{4}{9}$$

By def^n, $\hat{\theta}_n$ is consistent if $\hat{\theta}_n \xrightarrow{P} \theta$, which is not the case for $\hat{\theta}_1^*$. This occurs as both z and Y (regressor and outcome) depends on the unobserved V_i, but V_i is not omitted from the model. If is consistent if $Cov(z, V) = 0$,

V_i is a confounding variable. It violates the independence assumption in OLS models. This pheromenon is called omitted-variable bias (wikipedia).

(b) An alternative estimator is obtained by taking $\widetilde{\theta} = \widehat{\phi}_1 / \widehat{\psi}_1$, where $\hat{\phi}_1$ and $\hat{\psi}_1$ are the OLS estimators of coefficients ϕ_1 and ψ_1 in the models

$$Y_i = \phi_0 + \phi_1 A_i + \kappa_i$$
$$Z_i = \psi_0 + \psi_1 A_i + \xi_i.$$

Find $\text{plim}_{n\to\infty} \widetilde{\theta}$. Under which circumstances $\widetilde{\theta}$ is a consistent estimator of θ? Is this true in the present setting? What is this estimator/method called?

To use our new estimator $\tilde{\theta} = \hat{\phi}_1 / \hat{\psi}_1$, recall our original data generating mechanism

$$\begin{cases} Y_i = \theta Z_i + U_i + \varepsilon_i \\ Z_i = 0.5 A_i + U_i + \tau_i \end{cases}$$

(1) Now for our model $Y_i = \phi_0 + \phi_1 A_i + \kappa_i$

Substitute the expression Y_i from the original data-generating mechanism,

$$\begin{aligned} Y_i &= \theta Z_i + U_i + \varepsilon_i \\ &= \theta (0.5 A_i + U_i + \tau_i) + U_i + \varepsilon_i \\ &= 0.5\theta A_i + \text{other terms} \end{aligned}$$

As $A_i \perp$ other terms, we see $\hat{\phi}_1 = 0.5\theta$.
As $n \to \infty$, since OLS estimators are consistent
ie. $\underset{n\to\infty}{\text{plim}}\ \hat{\phi}_1 = \phi_1 = 0.5\theta$

(2) Now for $z_i = \psi_0 + \psi_1 A_i + \xi_i$, from the original data generating mechanism,

$$z_i = 0.5A_i + v_i + \varepsilon_i$$

We see $\hat{\psi}_1 = 0.5$ as $A_i \perp$ other terms.

Now

$$\underset{n\to\infty}{\text{plim}}\ \hat{\psi}_1 = \psi_1 = 0.5$$

Then, $\tilde{\theta} = \hat{\phi}_1 / \hat{\psi}_1 = \frac{0.5\theta}{0.5} = \theta$. Both $\hat{\phi}_1$ and $\hat{\psi}_1$ converge to their true values due to OLS estimator consistency. Hence, $\underset{n\to\infty}{\text{plim}}\ \tilde{\theta} = \theta$ is true.

This is called 2 stage least squares (2SLS Wikipedia). Another name that came up was instrumental variable analysis, which is the seneria for something like this.

3. Carry out and report a simulation study of the properties of (the sampling distributions of) $\widehat{\theta}_1^*$ and $\widetilde{\theta}$. The data generating mechanism is specified as above. Choose $\theta = 0.5$, $m = 1000$ and sample size scenarios $n = 250, 500$. Report mean, bias and standard deviation of the estimator, Monte Carlo error of the mean, mean standard error and 95% normal approximation confidence interval coverage probability. The standard error of $\widehat{\theta}_1^*$ may be obtained from linear regression. For $\widetilde{\theta}$ you can use the variance estimator

$$\widehat{V}[\widetilde{\theta}] = \frac{\widehat{V}[Y_i - \widetilde{\theta} Z_i]}{n \widehat{V}[Z_i] \widehat{\rho}_{A,Z}^2},$$

where $\widehat{\rho}$ is the sample correlation coefficient. Report also the bootstrap standard error, and coverage probability of 95% bootstrap normal approximation and bootstrap percentile confidence intervals. Which one of the two point estimators is better and why? How do the results compare to the theoretical ones in Q2?

Output:

1) $\hat{\theta}^*$ with $n=250$

```
[1] "theta_star"
     No. runs nobs  Mean  Bias    SD   MCE Mean SE Normal CI coverage Mean BS SE
[1,]     1000  250 0.943 0.443 0.052 0.002   0.053                  0      0.052
     Normal BS CI coverage BS Percent lower BS Percent upper
[1,]                     0            0.845            1.043
```

2) $\tilde{\theta}$ with $n=250$

```
[1] "theta_tilde"
     #runs nobs  Mean   Bias    SD   MCE Mean SE Normal CI coverage Mean BS SE Normal BS CI coverage BS Percent lower
[1,]  1000  250 0.481 -0.019 0.197 0.006   0.193               0.958      0.247                  0.968           -0.032
     BS Percent upper
[1,]           -0.032
```

3) $\hat{\theta}^*$ with $n = 500$

```
[1] "theta_star"
     No. runs nobs  Mean  Bias    SD   MCE Mean SE Normal CI coverage Mean BS SE Normal BS CI coverage BS Percent lower
[1,]     1000  500 0.943 0.443 0.038 0.001   0.037                  0      0.037                     0            0.873
     BS Percent upper
[1,]            1.014
```

4) $\tilde{\theta}$ with $n = 500$

```
[1] "theta_tilde"
     #runs nobs  Mean  Bias    SD   MCE Mean SE Normal CI coverage Mean BS SE Normal BS CI coverage BS Percent lower
[1,]  1000  500 0.491 -0.009 0.132 0.004    0.13              0.954      0.136                 0.962            0.201
     BS Percent upper
[1,]            0.201
```

We can see $\tilde{\theta}$ is a better estimator than $\hat{\theta}^*$ from the Mean (around 0.5), Bias (Close to 0). Also the 95% CI coverage probability is close to 1 for $\tilde{\theta}$ compared to 0 for $\hat{\theta}^*$ for both $n=250$ and $n=500$ which is good.

It matches with our theoretical analysis from Q2.

Note: Code included as QMD.