

Final Exam (January 2022)

Problem 1. (10 points) Suppose that $\hat{\theta}_n \rightarrow_p 0$ and $\sqrt{n}\hat{\theta}_n \rightarrow_d N(0, \sigma^2)$, where $\sigma^2 > 0$ is a finite constant (scalar). Let $\hat{\sigma}_n^2$ be a consistent estimator of σ^2 . Use the delta method or the continuous mapping theorem (CMT) to find the asymptotic distributions of the following quantities:

- (i) $\Phi\left(\sqrt{n}\hat{\theta}_n/\hat{\sigma}_n\right)$, where $\Phi(\cdot)$ is the standard normal CDF.
- (ii) $\sqrt{n}\left(\Phi\left(\hat{\theta}_n\right) - 0.5\right)/\hat{\sigma}_n$.

Solution.

- (i) First,

$$\sqrt{n}\hat{\theta}_n/\hat{\sigma}_n \rightarrow_d \frac{1}{\sigma}N(0, \sigma^2) =^d N(0, 1).$$

By the CMT (since $\Phi(\cdot)$ is continuous everywhere),

$$\Phi\left(\sqrt{n}\hat{\theta}_n/\hat{\sigma}_n\right) \rightarrow_d \Phi(Z),$$

where $Z \sim N(0, 1)$. However,

$$\begin{aligned} P(\Phi(Z) \leq u) &= P(Z \leq \Phi^{-1}(u)) \\ &= \Phi(\Phi^{-1}(u)) \\ &= u \end{aligned}$$

for all $u \in (0, 1)$. This is a CDF of a *Uniform*(0, 1) distribution. Therefore, we conclude that

$$\Phi\left(\sqrt{n}\hat{\theta}_n/\hat{\sigma}_n\right) \rightarrow_d \text{Uniform}(0, 1).$$

- (ii) Since $\Phi(0) = 0.5$, and $\hat{\theta}_n \rightarrow_p 0$, we can write,

$$\sqrt{n}\left(\Phi\left(\hat{\theta}_n\right) - 0.5\right)/\hat{\sigma}_n = \sqrt{n}\left(\Phi\left(\hat{\theta}_n\right) - \Phi(0)\right)/\hat{\sigma}_n$$

By the delta method,

$$\begin{aligned} \sqrt{n}\left(\Phi\left(\hat{\theta}_n\right) - \Phi(0)\right)/\hat{\sigma}_n &\rightarrow_d \frac{d\Phi(0)}{d\theta} \frac{1}{\sigma}N(0, \sigma^2) \\ &=^d \phi(0)N(0, 1) \\ &=^d N(0, \phi^2(0)), \end{aligned}$$

where $\phi(0)$ is the standard normal PDF at zero.

Problem 2. (10 points) Suppose we observe the i.i.d. random sample $\{(Y_i, X_i)\}_{i=1}^n$ with X_i being a scalar. Take the linear model

$$\begin{aligned} Y_i &= X_i\beta + e_i \\ \mathbb{E}(e_i | X_i) &= 0 \\ \Omega &= \mathbb{E}(X_i^2 e_i^2). \end{aligned}$$

Let $\hat{\beta}$ be the LS estimate of β with residuals $\hat{e}_i = Y_i - X_i\hat{\beta}$. Consider the estimates of Ω :

$$\begin{aligned} \tilde{\Omega}_n &= \frac{1}{n} \sum_{i=1}^n X_i^2 e_i^2 \\ \hat{\Omega}_n &= \frac{1}{n} \sum_{i=1}^n X_i^2 \hat{e}_i^2. \end{aligned}$$

- (i) Find the asymptotic distribution of $\sqrt{n}\left(\tilde{\Omega}_n - \Omega\right)$.

(ii) Find the asymptotic distribution of $\sqrt{n}(\hat{\Omega}_n - \Omega)$.

Solution. By CLT,

$$\sqrt{n}(\tilde{\Omega} - \Omega) \rightarrow_d N(0, \text{Var}(X_i^2 e_i^2)).$$

$$\text{Var}(X_i^2 e_i^2) = \mathbb{E}(X_i^4 e_i^4) - \mathbb{E}(X_i^2 e_i^2)^2.$$

Use the expansion

$$\begin{aligned} \hat{e}_i^2 &= (Y_i - X_i \hat{\beta})^2 \\ &= (e_i - X_i(\hat{\beta} - \beta))^2 \\ &= e_i^2 + X_i^2(\hat{\beta} - \beta)^2 - 2e_i X_i(\hat{\beta} - \beta). \end{aligned}$$

Then,

$$\begin{aligned} \hat{\Omega} &= \frac{1}{n} \sum_{i=1}^n X_i^2 \hat{e}_i^2 \\ &= \tilde{\Omega} + \frac{1}{n} \sum_{i=1}^n X_i^4 (\hat{\beta} - \beta)^2 - 2 \frac{1}{n} \sum_{i=1}^n X_i^3 e_i (\hat{\beta} - \beta) \end{aligned}$$

and

$$\sqrt{n}(\hat{\Omega} - \Omega) = \sqrt{n}(\tilde{\Omega} - \Omega) + \frac{1}{n} \sum_{i=1}^n X_i^4 \sqrt{n}(\hat{\beta} - \beta)^2 - 2 \frac{1}{n} \sum_{i=1}^n X_i^3 e_i \sqrt{n}(\hat{\beta} - \beta).$$

Note that we have

$$\begin{aligned} \sqrt{n}(\hat{\beta} - \beta) &\rightarrow_d N(0, V_\beta) \\ \frac{1}{n} \sum_{i=1}^n X_i^4 &\rightarrow_p \mathbb{E}(X_i^4) \\ \frac{1}{n} \sum_{i=1}^n X_i^3 e_i &\rightarrow_p \mathbb{E}(X_i^3 e_i) = 0 \\ \hat{\beta} - \beta &\rightarrow_p 0. \end{aligned}$$

By Slutsky's theorem,

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n X_i^4 \sqrt{n}(\hat{\beta} - \beta)^2 &\rightarrow_p 0 \\ \frac{1}{n} \sum_{i=1}^n X_i^3 e_i \sqrt{n}(\hat{\beta} - \beta) &\rightarrow_p 0. \end{aligned}$$

So

$$\frac{1}{n} \sum_{i=1}^n X_i^4 \sqrt{n}(\hat{\beta} - \beta)^2 - 2 \frac{1}{n} \sum_{i=1}^n X_i^3 e_i \sqrt{n}(\hat{\beta} - \beta) \rightarrow_p 0$$

and by Slutsky's theorem,

$$\sqrt{n}(\hat{\Omega} - \Omega) \rightarrow_d N(0, \text{Var}(X_i^2 e_i^2)).$$

Problem 3. (10 Points) Consider the linear regression model with endogenous regressors:

$$\begin{aligned} Y_i &= X_i' \beta + U_i, \\ \mathbb{E}(X_i U_i) &\neq 0, \end{aligned}$$

where β is the unknown k -vector of parameters ($k > 4$). Let Z_i be an $l > k$ vector of instruments. Describe how to test

$$H_0 : \beta_1 \beta_3 = \beta_2 \beta_4 \text{ against } H_1 : \beta_1 \beta_3 \neq \beta_2 \beta_4$$

using the efficient GMM estimator.

- (i) Explain step by step how to construct the Wald test statistic allowing for heteroskedastic errors.
- (ii) Describe the asymptotic distribution of the proposed statistic under H_0 . Assume that the data are i.i.d. and that the usual assumptions on the moments of X 's, Z 's, and U 's are satisfied. Describe the decision rule.

Solution.

- (i) The efficient GMM estimator is

$$\begin{aligned}\hat{\beta}_{GMM} &= \left(\left(\sum_{i=1}^n X_i Z_i' \right) \hat{\Omega}_n \left(\sum_{i=1}^n Z_i X_i' \right) \right)^{-1} \left(\left(\sum_{i=1}^n X_i Z_i' \right) \hat{\Omega}_n \left(\sum_{i=1}^n Z_i Y_i \right) \right) \\ &= \left(\left(\sum_{i=1}^n X_i Z_i' \right) \left(\sum_{i=1}^n \hat{U}_i^2 Z_i Z_i' \right)^{-1} \left(\sum_{i=1}^n Z_i X_i' \right) \right)^{-1} \left(\left(\sum_{i=1}^n X_i Z_i' \right) \left(\sum_{i=1}^n \hat{U}_i^2 Z_i Z_i' \right)^{-1} \left(\sum_{i=1}^n Z_i Y_i \right) \right)\end{aligned}$$

where $\hat{U}_i = Y_i - X_i' \hat{\beta}^{2SLS}$ ($\hat{\beta}^{2SLS}$ is the two stage least square estimator), therefore, $\sqrt{n}(\hat{\beta}_{GMM} - \beta) \rightarrow_d N(0, V)$, where

$$V = \left(E[X_i Z_i'] \cdot (E[U_i^2 Z_i Z_i'])^{-1} E[Z_i X_i'] \right)^{-1}.$$

Define $h(\beta) = \beta_1 \beta_3 - \beta_2 \beta_4$. By delta method,

$$\sqrt{n} \left(h(\hat{\beta}_{GMM}) - h(\beta) \right) \rightarrow_d N \left(0, \frac{\partial h(\beta)}{\partial \beta'} V \frac{\partial h(\beta)}{\partial \beta} \right).$$

$\frac{\partial h(\beta)}{\partial \beta'} = [\beta_3 \quad -\beta_4 \quad \beta_1 \quad -\beta_2 \quad 0 \quad \dots \quad 0]$. Under H_0 , $h(\beta) = 0$. The Wald statistic:

$$W_n = n \cdot h(\hat{\beta}_{GMM})' \left(\frac{\partial h(\hat{\beta}_{GMM})}{\partial \beta'} \hat{V}_n \frac{\partial h(\hat{\beta}_{GMM})}{\partial \beta} \right)^{-1} h(\hat{\beta}_{GMM}),$$

where $\frac{\partial h(\hat{\beta}_{GMM})}{\partial \beta'} = [\hat{\beta}_{3,GMM} \quad -\hat{\beta}_{4,GMM} \quad \hat{\beta}_{1,GMM} \quad -\hat{\beta}_{2,GMM} \quad 0 \quad \dots \quad 0]$ and

$$\hat{V}_n = \left(\left(\frac{1}{n} \sum_{i=1}^n X_i Z_i' \right) \left(\frac{1}{n} \sum_{i=1}^n (Y_i - X_i' \hat{\beta}_{GMM})^2 Z_i Z_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n Z_i X_i' \right) \right)^{-1}.$$

W_n is the Wald test statistic allowing for heteroskedastic errors.

- (ii) Under H_0 ($h(\beta) = 0$), $W_n \rightarrow_d \chi_1^2$. Decision rule: reject H_0 if $W_n > \chi_{1,1-\alpha}^2$, where $\Pr[W > \chi_{1,1-\alpha}^2] = \alpha$ with $W \sim \chi_1^2$.

Problem 4. (15 points) Consider the model

$$\begin{aligned}Y_i &= \beta_0 + \beta_1 X_i + e_i \\ \mathbb{E}(e_i) &= 0 \\ \mathbb{E}(X_i e_i) &= 0\end{aligned}$$

with both Y_i and X_i scalar. Assume $\beta_0 > 0$ and $\beta_1 < 0$. Suppose the parameter of interest is the area under the regression curve (e.g., consumer surplus), which is $A = -\frac{\beta_0^2}{2\beta_1}$. Let $\hat{\theta} = (\hat{\beta}_0, \hat{\beta}_1)'$ be the LS estimator of $\theta = (\beta_0, \beta_1)'$ so that $\sqrt{n}(\hat{\theta} - \theta) \rightarrow_d N(0, V_\theta)$. Let \hat{V}_θ be a standard consistent estimator for V_θ . You do not need to write out these estimators.

- (i) Given the above, describe an estimator of A .
- (ii) Construct an asymptotic $1 - \alpha$ coverage probability confidence interval for A .
- (iii) Construct an asymptotic $1 - \alpha$ coverage probability bootstrap percentile confidence interval for A .

Solution.

(i) The plug-in estimator is $\hat{A} = -\hat{\beta}_0^2/2\hat{\beta}_1$.

(ii) Define

$$\mathbf{a} = \begin{pmatrix} \frac{\partial A}{\partial \beta_0} \\ \frac{\partial A}{\partial \beta_1} \end{pmatrix} = \begin{pmatrix} -\beta_0/\beta_1 \\ \beta_0^2/2\beta_1^2 \end{pmatrix}.$$

By delta method,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \rightarrow_d N(\mathbf{0}, \mathbf{a}' \mathbf{V}_{\boldsymbol{\theta}} \mathbf{a}),$$

where $\mathbf{V}_{\boldsymbol{\theta}} = \mathbf{Q}^{-1} \boldsymbol{\Omega} \mathbf{Q}^{-1}$, $\mathbf{Q} = \mathbb{E}(\mathbf{X}_i \mathbf{X}_i')$, $\mathbf{X}_i = (1, X_i)'$ and $\boldsymbol{\Omega} = \mathbb{E}(e_i^2 \mathbf{X}_i \mathbf{X}_i')$. An estimate of the asymptotic variance is $\hat{\mathbf{a}}' \hat{\mathbf{V}}_{\boldsymbol{\theta}} \hat{\mathbf{a}}$, where

$$\hat{\mathbf{a}} = \begin{pmatrix} -\hat{\beta}_0/\hat{\beta}_1 \\ \hat{\beta}_0^2/2\hat{\beta}_1^2 \end{pmatrix},$$

$\hat{\mathbf{V}}_{\boldsymbol{\theta}} = \hat{\mathbf{Q}}^{-1} \hat{\boldsymbol{\Omega}} \hat{\mathbf{Q}}^{-1}$, $\hat{\mathbf{Q}} = n^{-1} \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i'$, $\hat{\boldsymbol{\Omega}} = n^{-1} \sum_{i=1}^n \hat{e}_i^2 \mathbf{X}_i \mathbf{X}_i'$ and $\hat{e}_i = Y_i - \mathbf{X}_i' \hat{\boldsymbol{\theta}}$. An asymptotic $1 - \alpha$ coverage probability confidence interval for A is

$$\left[\hat{A} - 1.96 \times \sqrt{n^{-1} \hat{\mathbf{a}}' \hat{\mathbf{V}}_{\boldsymbol{\theta}} \hat{\mathbf{a}}}, \hat{A} + 1.96 \times \sqrt{n^{-1} \hat{\mathbf{a}}' \hat{\mathbf{V}}_{\boldsymbol{\theta}} \hat{\mathbf{a}}} \right].$$

(iii) percentile bootstrap

- Step 1: Generate an i.i.d. sample $\{(Y_i^*, X_i^*)\}_{i=1}^n$ by drawing with replacement from the sample $\{(Y_i, X_i) : i = 1, \dots, n\}$.
- Step 2: Use the bootstrap sample to calculate $\hat{\boldsymbol{\theta}}^* = (\hat{\beta}_0^*, \hat{\beta}_1^*)' = (\sum_{i=1}^n \mathbf{X}_i^* \mathbf{X}_i^{*'})^{-1} (\sum_{i=1}^n \mathbf{X}_i^* Y_i^*)$ ($\mathbf{X}_i^* = (1, X_i^*)'$) and $\hat{A}^* = -\hat{\beta}_0^{*2}/2\hat{\beta}_1^*$.
- Step 3: Repeat B times. Collect B replications: $\hat{A}^{*1}, \dots, \hat{A}^{*B}$.
- Step 4: Order the B values of $\hat{A}^{*1}, \dots, \hat{A}^{*B}$. Denote the ordered values by $\hat{A}_{(1)}^*, \dots, \hat{A}_{(B)}^*$. Suppose $B = 1000$. The bootstrap percentile confidence interval is $[\hat{A}_{(25)}^*, \hat{A}_{(975)}^*]$.

Problem 5. (10 points) Consider the simple regression model (with independently and identically distributed (i.i.d.) observations):

$$Y_i = \beta_0 + \beta_1 X_i^* + U_i.$$

Assume that $\mathbb{E}(U_i) = \mathbb{E}(X_i^* U_i) = 0$. However, instead of observing X_i^* , we only observed $X_i = X_i^* + e_i$. We think of X_i as some measurement of X_i^* that is subject to error. Assume

$$\mathbb{E}(e_i) = \mathbb{E}(e_i U_i) = \mathbb{E}(X_i^* e_i) = 0.$$

(i) Suppose we estimate the model using LS with the observed X_i in place of X_i^* . Let $\hat{\beta}_{1,n}^{LS}$ denote the LS estimator. Show that

$$\hat{\beta}_{1,n}^{LS} \rightarrow_p \beta_1 \frac{\text{Var}(X_i^*)}{\text{Var}(X_i^*) + \text{Var}(e_i)}.$$

This means when there is measurement error, the LS estimate is closer to zero than β_1 .

(ii) Suppose we have a second (subject-to-error) measurement of X_i^* , $Z_i = X_i^* + \eta_i$, where

$$\mathbb{E}(\eta_i) = \mathbb{E}(\eta_i U_i) = \mathbb{E}(X_i^* \eta_i) = \mathbb{E}(\eta_i e_i) = 0.$$

Show that

$$\tilde{\beta}_{1,n} = \frac{\sum_{i=1}^n (Z_i - \bar{Z}_n) Y_i}{\sum_{i=1}^n (Z_i - \bar{Z}_n) X_i}$$

is a consistent estimator for β_1 , where $\bar{Z}_n = n^{-1} \sum_{i=1}^n Z_i$.

Solution.

(i) The linear model:

$$Y_i = \beta_0 + \beta_1 X_i^* + U_i = Y_i = \beta_0 + \beta_1 (X_i - e_i) + U_i = \beta_0 + \beta_1 X_i + U_i - \beta_1 e_i.$$

The LS estimator:

$$\hat{\beta}_1^{LS} = \beta_1 + \frac{n^{-1} \sum_{i=1}^n (X_i - \bar{X}) (U_i - \beta_1 e_i)}{n^{-1} \sum_{i=1}^n (X_i - \bar{X})^2}.$$

$\mathbb{E}(X_i U_i) = \mathbb{E}(X_i^* + e_i) U_i = 0$ and $\mathbb{E}(X_i e_i) = \mathbb{E}(X_i^* + e_i) e_i = \text{Var}(e_i)$. $\text{Var}(X_i) = \text{Var}(X_i^*) + \text{Var}(e_i) + 2 \cdot \text{Cov}(X_i^*, e_i) = \text{Var}(X_i^*) + \text{Var}(e_i)$. So, by WLLN and Continuous Mapping Theorem,

$$\begin{aligned} \frac{n^{-1} \sum_{i=1}^n (X_i - \bar{X}) (U_i - \beta_1 e_i)}{n^{-1} \sum_{i=1}^n (X_i - \bar{X})^2} &= \frac{n^{-1} \sum_{i=1}^n (X_i - \bar{X}) U_i - \beta_1 n^{-1} \sum_{i=1}^n (X_i - \bar{X}) e_i}{n^{-1} \sum_{i=1}^n (X_i - \bar{X})^2} \\ &\xrightarrow{p} \frac{\mathbb{E}(X_i U_i) - \mathbb{E}(X_i) \mathbb{E}(U_i) - \beta_1 (\mathbb{E}(X_i e_i) - \mathbb{E}(X_i) \mathbb{E}(e_i))}{\text{Var}(X_i)} \\ &= -\beta_1 \frac{\text{Var}(e_i)}{\text{Var}(X_i^*) + \text{Var}(e_i)}. \end{aligned}$$

(ii) And

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^n (Z_i - \bar{Z}) Y_i}{\sum_{i=1}^n (Z_i - \bar{Z}) X_i} = \frac{\sum_{i=1}^n (Z_i - \bar{Z}) (\beta_0 + \beta_1 X_i + U_i)}{\sum_{i=1}^n (Z_i - \bar{Z}) X_i} = \beta_1 + \frac{n^{-1} \sum_{i=1}^n (Z_i - \bar{Z}) U_i}{n^{-1} \sum_{i=1}^n (Z_i - \bar{Z}) (X_i - \bar{X})}.$$

Similarly,

$$\frac{n^{-1} \sum_{i=1}^n (Z_i - \bar{Z}) U_i}{n^{-1} \sum_{i=1}^n (Z_i - \bar{Z}) (X_i - \bar{X})} \xrightarrow{p} \frac{\mathbb{E}(Z_i U_i) - \mathbb{E}(Z_i) \mathbb{E}(U_i)}{\text{Cov}(X_i, Z_i)} = 0.$$

Problem 6. (15 points) Consider the regression model

$$\begin{aligned} Y_i &= \beta X_i + U_i, \\ \mathbb{E}(U_i | X_i) &= 0, \\ \mathbb{E}(U_i^2 | X_i) &= \sigma^2, \end{aligned}$$

where $\beta \in \mathbb{R}$ is an unknown scalar parameter. Assume that $\{(Y_i, X_i) : i = 1, \dots, n\}$ are iid. Consider the following estimator of β :

$$\tilde{\beta}_n = \frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n X_i}.$$

(i) Show that $\tilde{\beta}_n \xrightarrow{p} \beta$ as $n \rightarrow \infty$.

(ii) Show that $\sqrt{n}(\tilde{\beta}_n - \beta)$ is asymptotically normal, and find the asymptotic variance.

(iii) Compare the asymptotic variance of $\tilde{\beta}_n$ with the asymptotic variance of the OLS estimator. Which estimator is asymptotically more efficient?

Solution. First, write

$$\tilde{\beta}_n = \beta + \frac{n^{-1} \sum_{i=1}^n U_i}{n^{-1} \sum_{i=1}^n X_i}.$$

(i) Suppose that $E|X_i| < \infty$ and $EX_i \neq 0$, then

$$\begin{aligned} \tilde{\beta}_n &\xrightarrow{p} \beta + \frac{EU_i}{EX_i} \\ &= \beta. \end{aligned}$$

(ii) Write

$$\begin{aligned} \sqrt{n}(\tilde{\beta}_n - \beta) &= \frac{n^{-1/2} \sum_{i=1}^n U_i}{n^{-1} \sum_{i=1}^n X_i} \\ &\xrightarrow{d} \frac{N(0, \sigma^2)}{EX_i} \\ &=^d N\left(0, \frac{\sigma^2}{(EX_i)^2}\right). \end{aligned}$$

For the CLT to hold, we need an additional assumption that $\sigma^2 > 0$.

(iii) The asymptotic variance of the OLS estimator under the assumed conditions is

$$\frac{\sigma^2}{EX_i^2}.$$

Since

$$EX_i^2 \geq (EX_i)^2,$$

the OLS estimator has a smaller asymptotic variance and, therefore, is asymptotically more efficient.

Problem 7. (8 points) In April 1992, New Jersey (NJ) increased the state minimum wage from \$4.25 to \$5.05. The neighboring state, Pennsylvania, (PA) had minimum wage stay at \$4.25. Suppose you can collect random samples of firms in both states in February and November in 1992. Let emp be a variable that is equal to the number of employees in the firm. Without controlling for any other factors, write down a linear model that allows you to test whether the minimum wage policy reduces employment. Which coefficient in your model measures the effect of the minimum wage policy? Why might you want to control for other factors (explanatory variables) in the model?

Solution. Denote

$$G_i = \begin{cases} 1 & \text{if the } i\text{th firm in the sample is in NJ} \\ 0 & \text{if the } i\text{th firm in the sample is in PA} \end{cases}$$

and

$$T_i = \begin{cases} 1 & \text{if the } i\text{th firm is surveyed in November} \\ 0 & \text{if the } i\text{th firm is surveyed in February.} \end{cases}$$

Let emp_i denote the number of employees in the i th firm. We estimate the linear model

$$emp_i = \alpha + \gamma T_i + \lambda G_i + \delta T_i \times G_i + U_i.$$

The coefficient δ measures the effect of the policy. We may want to control for other explanatory variables in the model to avoid omitted variable problem.

Problem 8. (15 points) Consider the following regression model estimated using individual-level data on workers:

$$wage = \beta_0 + \beta_1 education + \beta_2 ability + u.$$

Assume that ability is unobserved, however, for each worker it is known if the town where he lives has a college or does not have a college. We make the following assumptions:

$$\mathbb{E}(education^C) > \mathbb{E}(education^{NC}) \quad (1)$$

and

$$\mathbb{E}(ability^C) = \mathbb{E}(ability^{NC}), \quad (2)$$

where $\mathbb{E}(education^C)$ and $\mathbb{E}(education^{NC})$ denote the expected years of education for workers living in a town with a college and no college respectively. Similarly, $\mathbb{E}(ability^C)$ and $\mathbb{E}(ability^{NC})$ denote the expected ability of workers living in a town with a college and no college respectively.

(i) (3 points) Is assumption (1) likely to be true? Explain why or why not.

(ii) (3 points) Is assumption (2) likely to be true? Explain why or why not.

(iii) (9 points) The econometrician used the following estimator of β_1 :

$$\hat{\beta}_1 = (\overline{wage}^C - \overline{wage}^{NC}) / (\overline{education}^C - \overline{education}^{NC}),$$

where \overline{wage}^C denotes the average wage in the group of workers who live in a town with a college (assume that there are n_C workers in that group), and \overline{wage}^{NC} denotes the average wage in the group of workers who live in a town with no college (assume that there are n_{NC} workers in that group); $\overline{education}^C$ denotes the average years of education in the group of workers who live in a town with a college, and $\overline{education}^{NC}$ denotes the average years of education in the group of workers who live in a town with no college. Assume that assumptions (1) and (2) hold, and that the expected value of the error term u is zero in both groups. Use the law of large numbers to show that $\hat{\beta}_1$ is a consistent estimator of β_1 , i.e. show that $\hat{\beta}_1 \rightarrow_p \beta_1$ as $n_C \rightarrow \infty$ and $n_{NC} \rightarrow \infty$.

Solution. We have the model

$$wage = \beta_0 + \beta_1 education + \beta_2 ability + u,$$

which implies:

$$\overline{wage}^C = \beta_0 + \beta_1 \overline{education}^C + \beta_2 \overline{ability}^C + \bar{u}^C,$$

and similarly:

$$\overline{wage}^{NC} = \beta_0 + \beta_1 \overline{education}^{NC} + \beta_2 \overline{ability}^{NC} + \bar{u}^{NC},$$

We have then for $\hat{\beta}_1$ that

$$\begin{aligned} \hat{\beta}_1 &= \frac{\overline{wage}^C - \overline{wage}^{NC}}{\overline{education}^C - \overline{education}^{NC}} \\ &= \frac{\left(\beta_0 + \beta_1 \overline{education}^C + \beta_2 \overline{ability}^C + \bar{u}^C \right) - \left(\beta_0 + \beta_1 \overline{education}^{NC} + \beta_2 \overline{ability}^{NC} + \bar{u}^{NC} \right)}{\overline{education}^C - \overline{education}^{NC}} \\ &= \beta_1 + \beta_2 \frac{\overline{ability}^C - \overline{ability}^{NC}}{\overline{education}^C - \overline{education}^{NC}} + \frac{\bar{u}^C - \bar{u}^{NC}}{\overline{education}^C - \overline{education}^{NC}} \\ &\rightarrow_p \beta_1 + \beta_2 \frac{E(ability^C) - E(ability^{NC})}{E(education^C) - E(education^{NC})} + \frac{E(u^C) - E(u^{NC})}{E(education^C) - E(education^{NC})}, \end{aligned}$$

where convergence in the last line is by the LLN and Slutsky's Lemma. Slutsky's lemma can be applied here because $E(education^C) > E(education^{NC})$ so that $E(education^C) - E(education^{NC}) \neq 0$. It follows that $\hat{\beta}_1 \rightarrow_p \beta_1$ because

$$\begin{aligned} E(u^C) &= 0, \\ E(u^{NC}) &= 0, \text{ and} \\ E(ability^C) - E(ability^{NC}) &= 0. \end{aligned}$$

Problem 9. (9 Points) Suppose you collect data from a survey ($n = 1000$) on wages, years of schooling, years of experience and gender. In addition, you ask for information about marijuana usage. The original question is: "On how many times last month did you smoke marijuana?"

- Write an equation that would allow you to estimate the effect of marijuana usage on wage, while controlling for other factors as education, experience and gender.
- Write a single equation that would allow you to test whether marijuana usage has different effects on wages for men and women.
- Based on your model in (ii), explain in detail how to test the null hypothesis that marijuana usage has the same effect on wages for men and women. Hint: Describe how to construct the test statistic and the decision rule.
- Suppose you think it is better to measure marijuana usage by putting people into one of 4 categories: (a) nonuser, (b) light user (1 to 5 times per month), (c) moderate user (6 to 10 times per month), (d) heavy user (more than 10 times per month). Write a model that allows you to estimate the effect of marijuana usage on wage with the help of these categories, using dummy variables. Define your dummy variables.
- Based on your model in (iv), explain in detail how to test the null hypothesis that smoking marijuana at least 10 times per month has at least as large effect on wage as smoking marijuana 6 to 10 times per month. Hint: Describe how to construct the test statistic and the decision rule.

Solution.

- The model is $\ln wage_i = \alpha_0 + \alpha_1 marijuana_i + \alpha_2 educ_i + \alpha_3 exper_i + \alpha_4 female_i + u_i$, where $\ln wage_i$ is the natural logarithm of wages, $marijuana_i$ is the frequency of marijuana smoking, $educ_i$ is years of schooling, $exper_i$ is years of experience for individual i , $female_i$ is 1 if individual i is female, and u_i is the error term.
- $\ln wage_i = \beta_0 + \beta_1 marijuana_i + \beta_2 educ_i + \beta_3 exper_i + \beta_4 female_i + \beta_5 marijuana_i \times female_i + \varepsilon_i$.
- Test $H_0 : \beta_5 = 0$ against $H_1 : \beta_5 \neq 0$, using the following test-statistic: $T = \frac{\hat{\beta}_5}{se(\hat{\beta}_5)}$, comparing to the critical value of $z_{1-\alpha/2}$. The decision rule is to reject H_0 if $|T| > z_{1-\alpha/2}$.

- (iv) The model is $\ln wage_i = \delta_0 + \delta_1 light_i + \delta_2 moderate_i + \delta_3 heavy_i + \delta_4 educ_i + \delta_5 exper_i + \delta_6 female_i + \eta_i$, where $light_i$, $moderate_i$, $heavy_i$ are 1 if individual i is a light, moderate and heavy marijuana-user, respectively, and 0 otherwise. The omitted category is non-user.
- (v) Test $H_0 : \delta_3 \geq \delta_2$ against $H_1 : \delta_3 < \delta_2$, using the following test-statistic: $T = \frac{\hat{\delta}_3 - \hat{\delta}_2}{se(\hat{\delta}_3 - \hat{\delta}_2)}$, comparing to the critical value of $z_{1-\alpha}$. The decision rule is to reject H_0 if $T < z_{1-\alpha}$.