# Final Exam (January 2022)

**Problem 1.** (8 points) Consider the linear model

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + U_i \\ \mathbb{E}\left(U_i \mid X_i\right) &= 0. \end{aligned}$$

Suppose that $Y_i$ is binary: $Y_i \in \{0,1\}$ so that $\Pr\left(Y_i = 1 \mid X_i\right) = \mathbb{E}\left(Y_i \mid X_i\right) = \beta_0 + \beta_1 X_i$. Show that $U_i$ is heteroskedastic (i.e., $\mathbb{E}\left(U_i^2 \mid X_i\right)$ depends on $X_i$).

**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

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

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

$$\widetilde{\Omega}_n = \frac{1}{n}\sum_{i=1}^n X_i^2 e_i^2$$

$$\widehat{\Omega}_n = \frac{1}{n}\sum_{i=1}^n X_i^2 \widehat{e}_i^2.$$

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

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

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

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

where $\beta$ 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.

**Problem 4.** (15 points) Consider the model

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

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 $\widehat{\boldsymbol{\theta}} = \left(\widehat{\beta}_0, \widehat{\beta}_1\right)'$ be the LS estimator of $\boldsymbol{\theta} = (\beta_0, \beta_1)'$ so that $\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \to_d N\left(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{\theta}}\right)$. Let $\widehat{\boldsymbol{V}}_{\boldsymbol{\theta}}$ be a standard consistent estimator for $\boldsymbol{V}_{\boldsymbol{\theta}}$. 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 $A$.

**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 $\widehat{\beta}_{1,n}^{LS}$ denote the LS estimator. Show that

$$\widehat{\beta}_{1,n}^{LS} \to_p \beta_1 \frac{\mathrm{Var}(X_i^*)}{\mathrm{Var}(X_i^*) + \mathrm{Var}(e_i)}.$$

This means when there is measurement error, the LS estimate is closer to zero than $\beta_1$.

(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

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

is a consistent estimator for $\beta_1$, where $\overline{Z}_n = n^{-1} \sum_{i=1}^n Z_i$.

**Problem 6.** (15 points) Consider the regression model

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

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

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

(i) Show that $\widetilde{\beta}_n \to_p \beta$ as $n \to \infty$.

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

(iii) Compare the asymptotic variance of $\widetilde{\beta}_n$ with the asymptotic variance of the OLS estimator. Which estimator 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?

**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:

$$\mathbb{E}\left(education^C\right) > \mathbb{E}\left(education^{NC}\right) \tag{1}$$

and

$$\mathbb{E}\left(ability^C\right) = \mathbb{E}\left(ability^{NC}\right), \tag{2}$$

where $\mathbb{E}\left(education^C\right)$ and $\mathbb{E}\left(education^{NC}\right)$ denote the expected years of education for workers living in a town with a college and no college respectively. Similarly, $\mathrm{E}\left[ability^C\right]$ and $\mathrm{E}\left[ability^{NC}\right]$ 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 $\beta_1$:

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

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 $\beta_1$, i.e. show that $\hat{\beta}_1 \to_p \beta_1$ as $n_C \to \infty$ and $n_{NC} \to \infty$.

**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?"

(i) 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.

(ii) Write a single equation that would allow you to test whether marijuana usage has different effects on wages for men and women.

(iii) 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.

(iv) 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.

(v) 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.

(vi) Based on your model in (iv), explain in detail how to test the null hypothesis that smoking marijuana has no effect on wage. Hint: Again, describe the test statistic and the decision rule.