Final Exam (January 2022)

Problem 1. (5 points) Consider the following simple regression model:

$$Y_i = \alpha + \beta X_i + U_i.$$

Suppose the observations (Y_i, X_i), $i = 1, 2, ..., n$ are iid. Assume $\mathrm{E}[|U_i|] < \infty$, $\mathrm{E}[|X_i|] < \infty$ and $\mathrm{E}[U_i] = 0$. Let $\widetilde{\beta}_n$ be any consistent estimator of β (not necessarily the OLS estimator). Define the following estimator for α: $\widetilde{\alpha}_n = \overline{Y}_n - \widetilde{\beta}_n \overline{X}_n$, where $\overline{Y}_n = n^{-1}\sum_{i=1}^n Y_i$ and $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$. Prove that $\widetilde{\alpha}_n$ is a consistent estimator of α.

Problem 2. (10 points) Consider the linear model

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + U_i \\ \mathrm{E}[U_i \mid X_i] &= 0. \end{aligned}$$

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

Problem 3. (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 $\mathrm{E}[U_i] = \mathrm{E}[X_i^* U_i] = 0$. 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

$$\mathrm{E}[e_i] = \mathrm{E}[e_i U_i] = \mathrm{E}[X_i^* e_i] = 0.$$

(i) (5 points) Suppose we estimate the model using OLS with the observed X_i in place of X_i^*. Let $\widehat{\beta}_{1,n}^{LS}$ denote the OLS estimator. Show that $\widehat{\beta}_{1,n}^{LS} \to_p \bar{\beta}_1$ and find the expression of $\bar{\beta}_1$.

(ii) (5 points) Suppose we have a second (subject-to-error) measurement of X_i^*, $Z_i = X_i^* + \eta_i$, where

$$\mathrm{E}[\eta_i] = \mathrm{E}[\eta_i U_i] = \mathrm{E}[X_i^* \eta_i] = \mathrm{E}[\eta_i e_i] = 0.$$

Show that

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

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

Problem 4. (12 points) Consider the regression model

$$\begin{aligned} Y_i &= \beta X_i + U_i, \\ \mathrm{E}[U_i \mid X_i] &= 0, \\ \mathrm{E}\left[U_i^2 \mid X_i\right] &= \sigma^2, \end{aligned}$$

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 β:

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

(i) (4 points) Show that $\widetilde{\beta}_n \to_p \beta$ as $n \to \infty$.

(ii) (4 points) Show that $\sqrt{n}\left(\widetilde{\beta}_n - \beta\right)$ is asymptotically normal, and find the asymptotic variance.

(iii) (4 points) Compare the asymptotic variance of $\widetilde{\beta}_n$ with the asymptotic variance of the OLS estimator. Which estimator is asymptotically more efficient?

Problem 5. (10 points) Suppose that the econometrician has data on random variables X_i and Y_i generated from the following model:

$$\begin{aligned} Y_i &= X_i^3 + \varepsilon_i, \\ \mathrm{E}\left[\varepsilon_i \mid X_i\right] &= 0. \end{aligned}$$

The true model is unknown to the econometrician, and he estimates a linear regression of Y_i against a constant and X_i:

$$Y_i = \hat{\beta}_{0,n} + \hat{\beta}_{1,n} X_i + \hat{U}_i,$$

where $\hat{\beta}_{0,n}$ and $\hat{\beta}_{1,n}$ are the OLS estimators, and $\hat{U}_i$ denotes the OLS residuals. Suppose that data are iid and $X_i \sim \mathrm{N}(0,1)$. Find the probability limits of $\hat{\beta}_{0,n}$ and $\hat{\beta}_{1,n}$. Does the linear regression model correctly estimates the sign of the true marginal effect of X_i on Y_i? Hints: Since $X_i \sim \mathrm{N}(0,1)$, $\mathrm{E}\left[X_i^3\right] = 0$ and $\mathrm{E}\left[X_i^4\right] = 3$.

Problem 6. (10 points) Recall that the probability mass function for the Poisson distribution with parameter λ is

$$f\left(x \mid \lambda\right) = \frac{e^{-\lambda}\lambda^x}{x!},$$

where $x = 0, 1, 2, \ldots.$ Suppose that $X_1, ..., X_n$ is an i.i.d. sample with probability mass function $f\left(\cdot \mid \lambda_*\right)$. It can be shown that $\mathrm{E}\left[X_1\right] = \mathrm{Var}\left[X_1\right] = \lambda_*$ (you are not required to show this result.).

(i) (5 Points) Find the maximum likelihood estimator $\hat{\lambda}^{MLE}$.

(ii) (5 Points) What is the asymptotic distribution of $\sqrt{n}\left(\hat{\lambda}^{MLE} - \lambda_*\right)$? How do you estimate the asymptotic variance?

Problem 7. (10 points) Let $Y_i\left(1\right)$ and $Y_i\left(0\right)$ be the potential outcomes for individual i for the treated status and the untreated status respectively. There exists a random vector X_i of covariates for which we make the unconfoundedness assumption $Y_i\left(0\right) \perp\!\!\!\perp D_i \mid X_i$. We observe D_i, X_i and $Y_i = D_i Y_i\left(1\right) + \left(1 - D_i\right) Y_i\left(0\right)$ for each individual i. Here X_i is a discrete covariate random vector. We are interested in the conditional average treatment effect (at $X_i = x$) on the treated defined by $\mathrm{CATT}\left(x\right) = \mathrm{E}\left[Y_i\left(1\right) - Y_i\left(0\right) \mid D_i = 1, X_i = x\right]$, where x is some value chosen by the econometrician.

(i) (5 points) Show how the conditional average treatment effect is identified, which means: can you write $\mathrm{CATT}\left(x\right)$ as a quantity that depends on the joint distribution of the observed elements Y_i, D_i, X_i?

(ii) (5 points) Suggest a reasonable estimator of the conditional average treatment effect using the sample analogue principle we taught in class.

Problem 8. (12 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:

$$\mathrm{E}\left[education^C\right] > \mathrm{E}\left[education^{NC}\right] \tag{1}$$

and

$$\mathrm{E}\left[ability^C\right] = \mathrm{E}\left[ability^{NC}\right], \tag{2}$$

where $\mathrm{E}\left[education^C\right]$ and $\mathrm{E}\left[education^{NC}\right]$ 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) (6 points) The econometrician used the following estimator of β_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 β_1, i.e. show that $\hat{\beta}_1 \to_p \beta_1$ as $n_C \to \infty$ and $n_{NC} \to \infty$.

Problem 9. (12 Points) A researcher interested in the relationship between parenting, age and schooling has data for the year 2000 for a sample of 1,167 married males and 870 married females aged 35 to 42 in the National Longitudinal Survey of Youth which was carried out in the United States. In particular, she is interested in how the presence of young children in the household is related to the age and education of the respondent. She defines $CHILDL6$ to be 1 if there is a child less than 6 years old in the household and 0 otherwise and regresses it on AGE, age, and S, years of schooling, for males and females separately using the Probit model. She obtains the results shown in the table below (standard errors in parentheses). Φ is the standard normal distribution function and ϕ is the the standard normal probability density function. For males and females separately, she calculates

$$\overline{Z} = \hat{\beta}_1 + \hat{\beta}_2 \overline{AGE} + \hat{\beta}_3 \overline{S},$$

where $\overline{AGE}$ and $\overline{S}$ are the sample averages of AGE and S and $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ are the Probit estimates. The values of $\overline{Z}$ and $\phi\left(\overline{Z}\right)$ are shown in the following table:

	Males	Females
AGE	-0.137 (0.018)	-0.154 (0.023)
S	0.132 (0.015)	0.094 (0.020)
$Constant$	0.194 (0.358)	0.547 (0.492)
$\overline{Z}$	-0.399	-0.874
$\phi\left(\overline{Z}\right)$	0.368	0.272

(i) (3 Points) Explain how one may derive the marginal effects of the explanatory variables on the probability of having a child less than 6 in the household, and calculate for both males and females the average marginal effects calculated at the sample averages of AGE and S.

(ii) (3 Points) Explain whether the signs of the marginal effects are reasonable. Explain whether you would expect the marginal effect of schooling to be higher for males or for females.

(iii) (3 Points) Someone asks the researcher whether the marginal effect of S is significantly different for males and females. The researcher does not know how to test whether the difference is significant and asks you for advice. What would you say?

(iv) (3 Points) Describe the procedure (for males only) to obtain a bootstrap percentile confidence interval for β_2, the coefficient of AGE.

Problem 10. (9 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?