Econometrics 1

Final Exam, January 2025.

2 hours ; books, notes, slides, and calculators are forbidden.

The scale of points assigned to each exercise is only indicative and might change. We assume that all random variables are bounded, so all the moment conditions needed for applying the Law of Large Numbers or the Central Limit Theorem hold.

Exercise 1 (5 points)

1 point for a correct answer, -0,25 points for the wrong one, 0 points if no answer is provided.

Please, mark the correct answer only. You do not need to justify your answer.

In all the questions, the sample considered consists of i.i.d. variables with the same distribution as the same variables without the subscript i (for example, (Y_i, D_i) in question 1 has the same distribution as (Y, D)).

1. Let $(Y_i, D_i)_{i=1,\dots,n}$ be a sample and assume $P(D=0) = P(D=1) = P(D=2) = 1/3$. Let $(\widehat{\alpha}, \widehat{\beta}_D)$ be, respectively, the intercept and slope OLS coefficients of the regression of Y on D. Let $\widehat{Y} = \widehat{\alpha} + \widehat{\beta}_D D$ and $\widehat{\varepsilon} = Y - \widehat{Y}$. Then,
   (a) The slope coefficient β_D of the theoretical regression of Y on D satisfies $\beta_D = E(Y|D=1) - E(Y|D=0)$.
   (b) If the $(D_i)_{i=1,\dots,n}$ are not all equal, $\widehat{\text{Cov}}(\widehat{Y}, \widehat{\varepsilon}) = 0$.
   (c) If the $(D_i)_{i=1,\dots,n}$ are not all equal, it is possible to have $\sum_{i=1}^n \widehat{\varepsilon}_i \neq \sum_{i=1}^n \widehat{\varepsilon}_i D_i$.
   (d) The OLS estimator $\widehat{\beta}_D$ may not converge in probability because the invertibility condition on $E[(1, D)(1, D)^T]$ may not hold under the stated assumptions.
2. Let $(Y_i, X_i^T)_{i=1,\dots,n}$ be the observed sample, with $X_i \in \mathbb{R}^k$ and $E(XX^T)$ invertible. Let $\beta_0 \in \mathbb{R}^k$ be the coefficients of the theoretical regression of Y on X, with $\beta_0 \neq (0, \dots, 0)^T \in \mathbb{R}^k$. Let $\varepsilon = Y - \beta_0^T X$ be the residual of the theoretical regression of Y on X, and let $E(\varepsilon^2 XX^T) \neq E(\varepsilon^2)E(XX^T)$. We denote with $\widehat{\beta}$ the OLS coefficients of the regression of Y on X. Then,
   (a) $\sqrt{n}(\widehat{\beta} - \beta_0) \xrightarrow{d} \mathcal{N}(0, E[\varepsilon^2][E(XX^T)]^{-1})$.
   (b) $\|\sqrt{n}(\widehat{\beta} - \beta_0)\|^2 \xrightarrow{d} \chi^2_k$.
   (c) $\|\sqrt{n}\widehat{\beta}\| \xrightarrow{P} \infty$.
   (d) $\widehat{\beta}$ cannot be computed if $\widehat{V}(Y) = 0$.
3. Let $(Y_i, X_i^T)_{i=1,\dots,n}$ be the observed sample and consider the minimization problem

$$\min_{\beta \in \mathbb{R}^k} \sum_{i=1}^n |Y_i - X_i^T \beta|^2 + \lambda \|\beta\|^2$$

   where $\lambda > 0$, $\|\beta\|$ is the Euclidean norm of β, $X \in \mathbb{R}^k$, and $k \geq n$. Then,

(a) if $k > n$, the invertibility condition on $E(XX^T)$ does not hold, so the solution is not unique.

(b) the solution may or may not be unique depending on the value of $\lambda > 0$.

(c) the solution is always unique and is sparse, i.e., the solution $\widehat{\beta}$ will have many of its components equal to zero.

(d) the solution is always unique and is not sparse.

4. Let $Y_i(d)$ be the potential outcome of individual i under treatment status $d \in \{0, 1\}$, $Y_i = Y_i(D_i)$, and let D_i be the observed treatment of individual i. We observe a sample $(Y_i, D_i)_{i=1,\dots,n}$ with $P(D = 1) > 0$ and let β_D be the coefficient of the theoretical regression of Y on D. Then,

(a) $\mathrm{Cov}(D, Y(0)) = 0$ implies $\beta_D = E[Y(1) - Y(0)]$.

(b) for any $i \in \{1, \dots, n\}$, the probability of observing $Y_i(1)$ is zero.

(c) to estimate consistently $E[Y(1) - Y(0)|D = 1]$ we need to assume that the treatment effects are constant : $Y_i(1) - Y_i(0) = \delta$ for some non-random δ.

(d) $D \perp (Y(0), Y(1))$ implies $\beta_D = E[Y(1) - Y(0)|D = 1]$.

5. Let $Y_i(d)$ be the potential outcome of individual i under treatment status $d \in \{0, 1\}$; $Z_i = 1$ if individual i is allocated to the treatment arm and $Z_i = 0$ otherwise ; let $D_i(z) = 1$ if individual i takes the treatment under allocation $z \in \{0, 1\}$ and $D_i(z) = 0$ otherwise ; let $Y_i = Y_i(D_i)$ and $D_i = D_i(Z_i)$ be the observed outcome and treatment for individual i. We assume that $D(1) \geq D(0)$ and $E(D|Z = 1) - E(D|Z = 0) > 0$. Also, $Z \perp (Y(1), Y(0), D(1), D(0))$. We also observe a sample $(Y_i, D_i, Z_i)_{i=1,\dots,n}$, and assume that $V(D) > 0$ and $V(Z) > 0$. Then,

(a) $\widehat{\beta}_D \xrightarrow{P} E(Y|Z = 1) - E(Y|Z = 0)$, where $\widehat{\beta}_D$ is the slope coefficient of the OLS regression of Y on D.

(b) with D^* being the fitted value of the theoretical regression of D on Z, we have

$$\frac{E(Y|Z = 1) - E(Y|Z = 0)}{E(D|Z = 1) - E(D|Z = 0)} = \frac{\mathrm{Cov}(Y, D^*)}{\mathrm{V}(D^*)} .$$

(c) the causal effect $E[Y(1) - Y(0)|D = 1]$ satisfies

$$E[Y(1) - Y(0)|D = 1] = \frac{E(Y|Z = 1) - E(Y|Z = 0)}{E(D|Z = 1) - E(D|Z = 0)} .$$

(d) we cannot identify any causal effect if $\mathrm{Cov}(D, Y(0)) \neq 0$.

Exercice 2 (10 points)

We aim to measure the returns to education for men in the United States. For this purpose, we have a dataset where we measure the logarithm of wages (`lwage`), the number of years of education starting from the equivalent of elementary school (`educ`), the number of years of experience (`exper`) and its square (`expersq`), whether the individual is Black (`black`), lives in the southern United States (`south`), resides in a city with more than 50,000 inhabitants (`city`), or has lived there 10 years earlier (`city10`)[1].

1. The data used here is from the « National Longitudinal Survey of Young Men »(NLSYM) of 1966.

We present the following descriptive statistics :

Variable	Mean	Standard deviation
`lwage`	6,26	0,44
`educ`	13,26	2,68
`exper`	8,86	4,14
`expersq`	95,58	84,62

Table 1 – Mean and standard deviation of some variables

1. *(2 points)* We consider the regression below. What do the columns "Estimate", "Std. Error", and "Pr(>|t|)" correspond to ? How can we interpret the number 0.093171 (line `educ`) ? Comment on the statement "there is probably a selection bias because the R^2 of the regression is low".

```
Call:
lm_robust(formula = lwage ~ educ + exper + expersq, data = card)

Standard error type:  HC2

Coefficients:
             Estimate Std. Error t value   Pr(>|t|)  CI Lower  CI Upper   DF
(Intercept)  4.468541  0.0703272  63.539  0.000e+00  4.330646  4.606435 3006
educ         0.093171  0.0036805  25.315 2.503e-128  0.085954  0.100387 3006
exper        0.089783  0.0070933  12.657  8.217e-36  0.075875  0.103691 3006
expersq     -0.002486  0.0003414  -7.281  4.210e-13 -0.003155 -0.001816 3006

Multiple R-squared:  0.1958 , Adjusted R-squared:  0.195
F-statistic: 231.4 on 3 and 3006 DF,  p-value: < 2.2e-16
```

2. *(2,5 points)* We seek to measure the average marginal effect of experience. Provide the theoretical formula as well as an estimator for this parameter. Can we compute this estimator here ? If yes, replace the theoretical values with the numerical values (without doing the final calculation), otherwise indicate the missing information needed to do so. Same questions for the average marginal effect of the experiment for individuals having 12 years or less of education.

3. *(1 points)* We now consider the regression below. Comment on the statistical significance of `city` and `city10`. Also, comment on the evolution of the coefficient of `educ`. Was such an evolution expected ?

```
Call:
lm_robust(formula = lwage ~ educ + exper + expersq + south +
    black + city + city10, data = card)

Standard error type:  HC2

Coefficients:
             Estimate Std. Error t value  Pr(>|t|)  CI Lower  CI Upper   DF
(Intercept)  4.731360  0.0701260  67.469 0.000e+00  4.593860  4.868860 3002
educ         0.073850  0.0036446  20.263 1.135e-85  0.066704  0.080996 3002
exper        0.083516  0.0067402  12.391 2.007e-34  0.070300  0.096732 3002
expersq     -0.002246  0.0003185  -7.053 2.172e-12 -0.002871 -0.001622 3002
south       -0.123210  0.0153980  -8.002 1.736e-15 -0.153402 -0.093018 3002
black       -0.187532  0.0175203 -10.704 2.894e-26 -0.221885 -0.153179 3002
city         0.144433  0.0191052   7.560 5.329e-14  0.106972  0.181893 3002
city10       0.025528  0.0181646   1.405 1.600e-01 -0.010089  0.061144 3002

Multiple R-squared:  0.2909 , Adjusted R-squared:  0.2893
F-statistic:   187 on 7 and 3002 DF,  p-value: < 2.2e-16
```

4. *(1,5 points)* We now seek to use an instrumental variable for `educ`. Explain why and give the sign of the bias that can be expected on the estimator of the coefficient of `educ` in the previous question. We observe the variables `IQ` (measure of cognitive skills), `nearc` (proximity to a university), `married` (being married), `whitecol` (indicator of being a white-collar worker). Which variable seems the most appropriate as an instrument ? Justify your answer.

5. *(1,5 points)* Let z be the chosen instrument, and consider the regression below. Why is it important to study this regression ? Indicate which assumption(s) necessary for the validity of the instrumental approach can be tested using this regression, and the result(s) of these test(s).

```
Call:
lm_robust(formula = educ ~ z + exper + expersq + south + black +
    city + city10, data = card)

Standard error type:  HC2

Coefficients:
              Estimate Std. Error  t value  Pr(>|t|)  CI Lower  CI Upper   DF
(Intercept) 16.6553648   0.147054 113.2602 0.000e+00 16.367028 16.943702 3002
z            0.3211497   0.084520   3.7997 1.477e-04  0.155427  0.486872 3002
exper       -0.4100488   0.032089 -12.7783 1.897e-36 -0.472968 -0.347129 3002
expersq      0.0007195   0.001711   0.4205 6.742e-01 -0.002636  0.004075 3002
south       -0.2897870   0.078555  -3.6890 2.292e-04 -0.443814 -0.135760 3002
black       -1.0008330   0.088216 -11.3452 3.059e-29 -1.173803 -0.827863 3002
city         0.3679973   0.111622   3.2968 9.893e-04  0.149134  0.586861 3002
city10       0.0617206   0.109529   0.5635 5.731e-01 -0.153039  0.276480 3002

Multiple R-squared:  0.4745 , Adjusted R-squared:  0.4733
F-statistic: 521.2 on 7 and 3002 DF,  p-value: < 2.2e-16
```

6. *(2 points)* Comment on the output below and indicate which estimator it allows us to obtain. Comment on the evolution of the `educ` coefficient compared to question 3. Explain why this evolution may seem surprising and provide an explanation.

```
Call:
ivreg(formula = lwage ~ exper + expersq + south + black + city +
    city10 | educ | z, data = card)

Residuals:
     Min       1Q   Median       3Q      Max
-1.76404 -0.22980  0.02407  0.24842  1.40655

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.9768470  0.8988598   4.424 1.00e-05 ***
educ         0.1187744  0.0534800   2.221   0.0264 *
exper        0.1019455  0.0229290   4.446 9.06e-06 ***
expersq     -0.0022778  0.0003285  -6.933 5.02e-12 ***
south       -0.1083668  0.0235244  -4.607 4.26e-06 ***
black       -0.1429038  0.0560350  -2.550   0.0108 *
city         0.1266673  0.0294666   4.299 1.77e-05 ***
city10       0.0173992  0.0217012   0.802   0.4228
```

Exercise 3 (5 points)

Let $Y_i(d)$ the potential outcome of individual i under treatment status d. We assume the following model for the potential outcome

$$Y_i(d) = c_0 + \Delta_i d + \lambda_0 Z_i + \eta_i \text{ with } E(\eta_i) = 0 \text{ and } E(\Delta_i | Z_i, D_i) = \delta_0\,, \tag{0.1}$$

where c_0 and λ_0 are fixed coefficients. We also assume that

$$D_i(z) = \alpha_0 + \gamma_0 z + U_i + \zeta_0 \eta_i \text{ where } E(U_i) = 0 \text{ and } U_i \perp \eta_i\,. \tag{0.2}$$

We also assume that

$$Z_i \perp (U_i, \eta_i) \text{ and } \mathrm{V}(Z_i) > 0\,.$$

Let $D_i := D_i(Z_i)$ and $Y_i := Y_i(D_i)$. We observe an i.i.d. sample $(Y_i, D_i, Z_i)_{i=1,\ldots,n}$.

1. *(1 point)* From Equation (0.1) obtain a causal regression where Y is the dependent (response) variable, $(1, D, Z)$ are the regressors (each multiplied by a fixed coefficient), and the residual has zero mean.
2. *(1 point)* Show that if $\zeta_0 \neq 0$, the OLS estimator of δ_0 corresponding to the regression above is not consistent.
3. *(1 point)* Assume that $\zeta_0 \neq 0$. Under which conditions on λ_0 and γ_0 can δ_0 be consistently estimated ? Under such conditions, propose an estimator $\widehat{\delta}$ of δ_0 and, by using the results seen in the course, argue that it is consistent.
4. We want to test the null hypothesis of no treatment effect versus the alternative of positive treatment effects, i.e.,

$$H_0 : \delta_0 = 0 \text{ Versus } H_1 : \delta_0 > 0\,.$$

   (a) *(0.75 points)* Recall that

$$\sqrt{n}\left(\widehat{\delta} - \delta_0\right) \xrightarrow{d} \mathcal{N}(0, V_a)$$

   for some V_a. Give the expression of V_a and propose a consistent estimator of it.

   (b) *(0.5 points)* Provide a test statistic for testing H_0 versus H_1.

   (c) *(0.75 points)* Write (i) the critical value and (ii) the decision rule for testing H_0 versus H_1.