GENEVA
GRADUATE
INSTITUTE

INSTITUT DE HAUTES
ÉTUDES INTERNATIONALES
ET DU DÉVELOPPEMENT
GRADUATE INSTITUTE
OF INTERNATIONAL AND
DEVELOPMENT STUDIES

ECONOMETRICS I - MIE

Take-Home Exam 1

Prof. Julia Cajal-Grossi - TA Viktoria Vidahazy

Deadline: Final submissions must be sent over the email to the TA and the professor no later than Thursday 27 October at 5 PM.

Teams: This take-home exam can be solved in groups of no more than three students. You are welcome to choose your team partners freely. You are also welcome to work on your own if you wish to do so. The list of team members needs to the clearly stated on the front cover of the submission. Note that collaboration across work teams is not permitted in this take home exam.

Formatting: Please submit a single .pdf file with the answers to all four questions. If you are not familiar with LaTex, feel free to type in Word and convert to .pdf. The document (excluding a cover page) cannot exceed 10 pages with standard margins, single spacing and a font size of 12.

Grading: This take-home exam accounts for 20% of the overall grade of the course. The grading criteria do not depend on the size of the team – 1, 2 or 3 people.

Plagiarism: Please, observe the rules against plagiarism at the Institute.

Contain all variables. Without the one with collinearity problem. => co variance

Test significance.

Question 1 (40%)

Collinarity => whats the problem of it

You are an applied economist working for the research division of an international organization. Your team leader asks you to propose an econometric approach for answering the following research questions: *are factory workers more productive when their work environment meets good social compliance standards? if so, how much more? is this different for female and male workers?* In order to do so you are given access to a dataset that was obtained through a survey of 10,000 line operators in garment plants in Dhaka, Bangladesh (i.e. N=10,000). Workers may be employed at the same factory but no two workers work on the same division or production line. All survey data were collected during September 2022, after an audit of the performance of all factories in Dhaka along several dimensions of social compliance. The dataset you are given contains the following variables:

• All the variables

• control

Discussion: in this question we have this data

=> Cause problem

=> This model first: results

This another model

=> Do this test

- worker_id: a numeric identifier for the worker; this variable uniquely identifies an observation in the data
- factory_id: a numeric identifier for the factory of the worker (there are several thousand different factories in the data)
- productivity: the number of pieces of garment the worker produced per hour of work the day before the survey (which was a normal day and should be taken as representative of the worker's productivity)
- product: a categorical variable that contains the specific product the worker was producing the previous day (t-shirts, jeans, sweaters, etc.)
- female: a dummy that takes value one if the worker is a female (zero if male)
- age: the age of the worker in years (between 18 and 65)
- marital status: categorical variable indicating the marital status of the worker (single, married, widowed, etc.)
- education: categorical variable containing the maximum level of education attained by the individual (none, incomplete primary, complete primary, etc.)
- experience: years of experience in their current role in the factory
- household_size: number of people leaving in the same household as the worker
- children: number of children (zero if none)
- factory_size: the number of employees in the factory
- division_size: the number of employees in the division of the worker
- happiness_score: a discrete variable ranging from 1 (extremely unhappy) to 5 (extremely happy) collecting answers from workers to the question *how happy are you with the working conditions in your plant?*
- x_1 harassment_pass: dummy that takes value one if the audit showed that the factory complies with good standards around the harassment of workers (zero if the factory is below a good standard)
- x_2 voice_pass: dummy that takes value one if the audit showed that the factory complies with good standards around workers' voice and dialogue in the factory (zero if the factory is below a good standard)
- x_3 maternity_pass: dummy that takes value one if the audit shows that the factory complies with good standards around maternity and childcare provisions (zero if the factory is below a good standard)
- x_4 investment_sc: dollars invested by the factory in 2021 on social compliance activities (note that only 10% of the workers are in factories with positive investments; the rest of the sample shows investments equal to zero)

• Standardize variables.

=> insignificant

$t = 0$

• Correlations

• Isolate effects

$y = x_1 \boxed{\beta} + \varepsilon$

$y = x_1 \boxed{\beta} + \dots x_4 \beta_4 + \varepsilon$

Write a model (or set of models) and an estimation approach for answering the research question you were tasked with. Note that you can only use the dataset at hand and the estimators and tests studied in this course. Please be specific in indicating the variables that you include / exclude in your model(s), how you would transform the variables (if at all), the expected signs and units of measurement of the relationships you are targeting, the construction of any tests you may want to perform, the hypothesis, rejection rules, and degrees of

(5) => t => F =>

(b(a.) b(b)) => chow differ man.

(b) ⇒ education
age — experience
factory
$u_{1t}+u_{2t}+\dots+u_{nt}$
$=\sum_{i=1}^{N} u_{it}$

$E\left(\frac{\sum u_{it}}{n}\right)=E(u_{1t})+\dots+E(u_{nt})$

$Var\left(\frac{\sum u_{it}}{n}\right)=Var(u_{1t}+u_{2t}+\dots+u_{nt})=$

$\frac{1}{n^2}(Var(u_{1t})+\dots+Var(u_{nt}))$

$=\frac{\sigma^2}{n}$?

freedom, etc.. Justify both the economics and the econometrics of your proposed approach. Describe any shortcomings of your exercise and be sure to explain your team leader what caveats they need to have in mind when interpreting the results that your approach would yield.

Question 2 (20%)

What happens for a model ⇒ To a model aggregation over i to t specific

Suppose that for the population of firms in Switzerland, the relationship over time between dividends and some observable regressors (such as the firm's size) follows the assumptions of the Classical Linear Regression Model, conditional on the realized values of the regressors. Let the model be:

$i, t \to$ time
$\downarrow$
individual

$$y_{it} = \boldsymbol{x}'_{it}\boldsymbol{\beta} + u_{it}$$

with u_{it} i.i.d. (over i and t), $\sim \mathcal{N}(0, \sigma^2)$, and independent of $\boldsymbol{x}_{it}$

$$i = 1 \dots N \text{ and } t = 1, \dots T$$

$y_{it} = \begin{bmatrix} y_1 \\ \vdots \\ y_T \end{bmatrix}$

where $y_1 = \sum_{i=1}^{N} y_{i1}$

Across firm in each time period ⇒ Only indexed by t

Rather than a random sample of firms over time, assume you have a sample of T averages (across all firms) of dividends and regressors – i.e., you have T observations. You estimate an alternative model adapted to the data you have at hand.

Aggregation over i for T specific

1. What would this model look like? Please spell out the model in as much detail as above.

2. Would the OLS estimator of the parameters in the adapted model be unbiased for $\boldsymbol{\beta}$ in the model presented in the introduction? Discuss your reasoning in detail (you do not need to formalize a proof if not needed for your argument).

3. Would it be efficient? X Discuss your reasoning in detail (you do not need to formalize a proof if not needed for your argument).

Question 3 (20%)

For some observations, observe both x & y
For some observations, only observe x but not y

Let a dataset containing N observations be formed by N_c *complete* observations (this is, we have information both for $\boldsymbol{X}$ and $\boldsymbol{y}$) and N_m observations for which we have data on $\boldsymbol{X}$ but $\boldsymbol{y}$ is *missing*, $N = N_c + N_m$. We are interested in the parameters of the model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$. A colleague suggests that you proceed as follows: (i) regress $\boldsymbol{y}_c$ on $\boldsymbol{X}_c$ to obtain $\hat{\boldsymbol{\beta}}_c$; (ii) predict $\hat{\boldsymbol{y}}_m$ using $\boldsymbol{X}_m\hat{\boldsymbol{\beta}}_c$; (iii) regress vector $(\boldsymbol{y}_c, \hat{\boldsymbol{y}}_m)$ onto $(\boldsymbol{X}_c, \boldsymbol{X}_m)$ to obtain $\hat{\boldsymbol{\beta}}_{c,m}$.

1. Is $\hat{\boldsymbol{\beta}}_{c,m}$ unbiased for $\boldsymbol{\beta}$? If biased, ✓ in what direction? Discuss your reasoning in detail (you do not need to formalize a proof if not needed for your argument).

$$\begin{pmatrix} y_c \\ \hat{y}_m \end{pmatrix} = \begin{pmatrix} x_c \\ x_m \end{pmatrix}\beta + \begin{pmatrix} u_c \\ u_m \end{pmatrix}$$

(i) $y_c = x_c\beta + u_c$

$\hat{\beta}_c = (x_c' x_c)^{-1} x_c' y_c \quad \hat{y}_m = x_m \hat{\beta}_c$

$$\hat{\beta}_{cm} = \left[(X_c \; X_m)' \begin{pmatrix} X_c \\ X_m \end{pmatrix} \right]^{-1} \begin{bmatrix} X_c \\ X_m \end{bmatrix}' \begin{bmatrix} y_c \\ y_m \end{bmatrix}$$

$$= \cdot - - \cdot \cdot$$

$$= \hat{\beta}_c$$

X biased

2. Is $\hat{\sigma}^2_{c,m}$, the estimator of the variance of the error in the 'imputed' regression unbiased? And, if biased, in what direction? Discuss your reasoning in detail (you do not need to formalize a proof if not needed for your argument).

3. Are there any implications of the approach proposed by your colleague, if you are considering testing the significance of an individual coefficient in your regression?

✓ **Question 4** (20%)

Consider a partitioned regression model, whose normal equations are:

$$\boldsymbol{X_1'X_1}\hat{\boldsymbol{\beta}}_1 + \boldsymbol{X_1'X_2}\hat{\boldsymbol{\beta}}_2 = \boldsymbol{X_1'y}$$

$$\boldsymbol{X_2'X_1}\hat{\boldsymbol{\beta}}_1 + \boldsymbol{X_2'X_2}\hat{\boldsymbol{\beta}}_2 = \boldsymbol{X_2'y}$$

Let $\hat{\boldsymbol{\epsilon}}$ denote the residuals of the full regression of $\boldsymbol{y}$ on $\boldsymbol{X}$. Let $\tilde{\boldsymbol{X}}_2 = \boldsymbol{M}_1\boldsymbol{X}_2$ and $\tilde{\boldsymbol{y}} = \boldsymbol{M}_1\boldsymbol{y}$.

1. Is it true that the residuals in the regression of $\tilde{\boldsymbol{y}}$ on $\tilde{\boldsymbol{X}}_2$ numerically equal $\hat{\boldsymbol{\epsilon}}$? Provide a formal proof.

2. Provide an intuitive explanation for why the statement in (1) is true / is false.

   Need to be true or not because partition regression means . . . ,

3. Is it true that $\tilde{\boldsymbol{y}}'\tilde{\boldsymbol{y}} - \hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}} = \tilde{\boldsymbol{y}}'\boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\tilde{\boldsymbol{y}}$? Provide a formal proof.

   after residualize this over that, there is no variation . . .

4. Provide an intuitive explanation for why the statement in (3) is true / is false.

$$\bar{P} = \tilde{X}_2'(\tilde{X}_2'\tilde{X}_2)^{-1}\tilde{X}_2^{t} \quad \tilde{X}_2' \; (\tilde{X}_2\tilde{X}_2)^{-1} \tilde{X}_2'$$