MOEC0021
Empirical Methods
University of Zurich HS 2020
PROF: Emanuele Dicarlo / TA: A. Jenni, M.R. Greco, S. Bagagli, L. Schubert

---

Problem Set 4 - Endogeneity and Panel Data

This problem set is due on the **14th of December** at **23:59**. Solutions should be turned in on OLAT in a single PDF file. Include in your file any code you wrote to answer the questions. Please name your file as GroupName_PS4.pdf.
One of your goals is to communicate efficiently. Please keep your answers succinct; lengthy answers will be marked down.

---

1. Theory - Endogeneity

Unemployment benefits provide workers with insurance against unemployment risk at the cost of incentive distortions in the labor market. According to economic theory, a crucial parameter in determining the optimal level of unemployment benefits is the elasticity of the unemployment duration with respect to the benefit level.
Suppose that you have cross-sectional data on individual unemployment spell duration and unemployment benefit. You wish to estimate the unemployment duration elasticity.

(a) You start by estimating the following model

$$\ln u_i = \beta_0 + \beta_1 \ln b_i + \nu_i, \tag{1}$$

where u_i is individual i's unemployment duration, b_i denotes the level of unemployment benefits for which individual i was eligible upon loosing her job and ν_i is the error term. Why does it make sense to estimate this model in logs? What problem arises when taking logs in this setting?

(b) Most economists would think that your model (1) suffers from an endogeneity problem? What is a mechanism that would create such an endogeneity problem? Why is endogeneity a concern?

(c) Having noticed that unemployment benefit levels are partly determined by state regulations, you decide to run the following regression

$$\ln u_{i,s} = \beta_0 + \beta_1 \ln \overline{b}_s + \beta_2 \mathsf{X}_{i,s} + \nu_{i,s}. \tag{2}$$

In this model, s indexes the state in which individual i lives, $\overline{b}_s$ is the average level of unemployment benefit in state s and $\mathsf{X}_{i,s}$ are individual controls (such as age, past employment history or education). How does using the average state benefit $\overline{b}_s$ instead of the individual benefit level $b_{i,s}$ help solve the endogeneity problem? Why is it important to control for individual characteristics in the regression?

(d) Give one reason why some economists would insist that your model (2) also suffers from an endogeneity problem?

(e) Abstracting from endogeneity concerns, are your results from regression (2) sufficient to obtain an estimate of the unemployment duration elasticity? Explain.

Let us now consider explicitly the case in which the government adjusts the level of unemployment benefits with the predicted average unemployment duration $\hat{u}_s$. Suppose that the government decision can be approximated using the following formula

$$\ln \bar{b}_s = \gamma_0 + \gamma_1 \ln \hat{u}_s + \nu_s, \tag{3}$$

where γ_0 and γ_1 are unknown parameters and ν_s is an approximation error satisfying $\mathbb{E}(\nu_s) = 0$. Suppose further that the prediction error in the average unemployment duration, η_s, is multiplicative

$$\hat{u}_s = \overline{u}_s \eta_s, \tag{4}$$

where $\overline{u}_s$ is the true average unemployment duration (which is only observed after the government sets $\bar{b}_s$), $\eta_s > 0$ and $\mathbb{E}(\ln \eta_s) = 0$.
We are interested in the following structural equation for the log average unemployment duration $\overline{u}_s$

$$\ln \overline{u}_s = \beta_0 + \beta_1 \ln \bar{b}_s + \mathsf{X}_s \Gamma + \varepsilon_s, \qquad \mathbb{E}(\varepsilon_s | \mathsf{X}_s) = 0, \tag{5}$$

where X_s contains other economic variables that determine the unemployment rate duration. Assume that the approximation error ν_s and the prediction error η_s are independent from each other and from X_s and ε_s.

(f) Use equations (3), (4) and (5) to obtain a reduced-form expression for $\ln \overline{u}_s$. That is, express $\ln \overline{u}_s$ as a function of the parameters and the exogenous variables of the model.

(g) Show that estimating β_1 directly from (5) yields an inconsistent estimate. Can you sign the asymptotic bias?

(h) Describe a strategy to obtain a consistent estimate of the unemployment duration elasticity in this context. What could be a limitation of your approach?

(i) (Bonus) What would motivate the government to set the benefit level as a function of the predicted unemployment duration?

2. Empirical Application - Panel Data

Given the rise in temperature predicted by climate scientists, the health consequences of high temperatures have recently received more attention.
In this problem, you will use panel data to investigate the temperature-mortality relationship over the second half of the twentieth century in the United States. This problem is based on a paper by Barreca, Clay, Deschenes, Greenstone and Shapiro ("Adapting to Climate Change: The Remarkable Decline in the US Temperature-Mortality Relationship over the Twentieth Century", *Journal of Political Economy*, 2016). You can download a simplified version of their data set on OLAT (`mortality_temp.csv`).
The data set consists of a monthly panel for each state for the years 1960-2004. It contains the following variables: `year`, `month`, `stfips` (state identifier), `lndrate` (log mortality rate), `lndrate_cvd` (log mortality rate due to cardiovascular diseases), `lndrate_mva` (log mortality rate due to motor vehicle accidents), `devp25` and `devp75` which are indicators for precipitations below the 25th percentile (respectively above the 75th percentile) of the precipitation distribution in a given month-state cell, and `bin_j` (number of days with daily average temperature in temperature bin j[1]). The bins correspond to the following temperatures:

- $< 10°\text{F}$ for $j = 1$
- $[(j-1) * 10°\text{F}, j * 10°\text{F})$ for $j = 2, \ldots, 9$
- $> 90°\text{F}$ for $j = 10$.

Throughout this exercise you will run several regressions of log mortality rate on the temperature bins. When reporting your results, please use the 60°F-69°F bin as the excluded group, i.e. the coefficient on a given temperature bin should correspond to an additional day in this bin instead of a day in the 60°F-69°F bin.

(a) Run a pooled OLS regression of log mortality rate on the temperature bins. What is the advantage of using temperature bins as independent variables instead of the average monthly temperature? Report the coefficient on the hottest temperature bin. Does it have the sign you expected?

(b) What is the crucial assumption for the pooled OLS estimator to be consistent in this context? Describe a mechanism that would lead this assumption to be violated. Could y use the random effect estimator to solve this problem?

(c) Estimate a model with state FE (fixed effects) and month FE. Report and interpret the coefficient on the hottest temperature bin with this specification. Why might this coefficient fail to recover the causal impact of temperature on log mortality?

[1] These variables are based on averages over all the weather stations in each state, which explains why they sometimes take rational values.

(d) Estimate a model with state-by-month FE. Report the coefficient on the hottest temperature bin. Does the coefficient have the sign you expected? What is the variation in the data that allows you to estimate the coefficients on the temperature bin variables?

(e) Write down the estimated equation and the required assumption for your estimator in (d) to be consistent. Do you find this assumption credible? Why?

(f) Plot the evolution over time of i) the average number of days in the hottest temperature bin per year ii) the average log mortality rate per year.

(g) Add the two dummies for unusual precipitations and a quadratic time trend (using the year variable) to the previous model. How does the coefficient on the hottest temperature bin compare with the one obtained in (d)? Use your plots from the previous question to interpret your results.

(h) Estimate again the model in (g) but using the log mortality rate due to motor-vehicle accidents and the log mortality rate due to cardiovascular diseases as dependent variables. Report the coefficient on the hottest temperature bin for each regression. Do these results add credibility to a causal interpretation of your estimate in (g)?

(i) (Bonus) What do you think of the magnitude of the effect found in (g)?