Problem Set 3
Due Monday, April 20 at the beginning of lecture
ECON 381-2, Northwestern University, Spring 2015

You may work in groups of 2 or 3 as long as each group member turns in their own handwritten copy and all group members are clearly indicated at the top of every copy. No credit is given if no work is shown. For Stata/R problems include both your code and the output.

1. (a) Consider the linear regression model

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

where (Y_i, X_i) are i.i.d. with finite fourth moments and $\mathbb{E}(U_i \mid X_i) = 0$. Assume that $\alpha = 0$. Explain why

$$\widehat{\beta}_{lrnc} \equiv \frac{\sum_{i=1}^N Y_i X_i}{\sum_{i=1}^N X_i^2}$$

is a consistent estimator of β.

(b) Suppose that we do not assume that $\alpha = 0$, but we know that $\overline{X} \equiv \frac{1}{N}\sum_{i=1}^N X_i = 0$ with probability 1 and $\overline{Y} \equiv \frac{1}{N}\sum_{i=1}^N Y_i = 0$ with probability 1. Explain why $\widehat{\beta}_{lrnc}$ is a consistent estimator of β.

(c) If we do not assume $\alpha = 0$ and we know that $\overline{X} = 0$ with probability 1, but we do not know the same for $\overline{Y}$, is $\widehat{\beta}_{lrnc}$ consistent for β? If not, find the probability limit of $\widehat{\beta}_{lrnc}$.

(d) If we do not assume $\alpha = 0$ and we know that $\overline{Y} = 0$ with probability 1, but we do not know the same for $\overline{X}$, is $\widehat{\beta}_{lrnc}$ consistent for β? If not, find the probability limit of $\widehat{\beta}_{lrnc}$.

(e) Suppose that $X_{i1}, \ldots, X_{iT}$ are T random variables and let $\widetilde{X}_{it} \equiv X_{it} - \frac{1}{T}\sum_{s=1}^T X_{is}$. Show that

$$\overline{\widetilde{X}} \equiv \frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \widetilde{X}_{it},$$

is equal to 0.

(f) Now consider the two period panel data model

$$Y_{it} = \alpha + \beta X_{it} + A_i + V_{it},$$

where $(X_{i1}, Y_{i1}, X_{i2}, Y_{i2})$ is an i.i.d. sample. Let $\widehat{\beta}_{fd}$ be the first-difference estimator of β that is formed by regressing $\Delta Y_i \equiv Y_{i2} - Y_{i1}$ on $\Delta X_i \equiv X_{i2} - X_{i1}$ without a constant (i.e. using the analog of $\widehat{\beta}_{lrnc}$). Let $\widehat{\beta}_{fe}$ be the fixed effects estimator. Show that $\widehat{\beta}_{fe} = \widehat{\beta}_{fd}$.

2. Let $\boldsymbol{X}_{it}$ be a K–dimensional column vector for each of $t = 1, \ldots, T$, and define $\dot{\boldsymbol{X}}_{it} = \boldsymbol{X}_{it} - \frac{1}{T}\sum_{s=1}^{T} \boldsymbol{X}_{is}$.

   (a) Show that $\sum_{t=1}^{T} \mathbb{E}[\dot{\boldsymbol{X}}_{it}\dot{\boldsymbol{X}}_{it}']$ is not invertible if any of the components of $\boldsymbol{X}_{it}$ do not vary over time.

   (b) Show that $\sum_{t=1}^{T} \mathbb{E}[\dot{\boldsymbol{X}}_{it}\dot{\boldsymbol{X}}_{it}']$ is not invertible if for every t one of the components of $\boldsymbol{X}_{it}$ can be written as a perfect linear combination (that does *not* vary across t) of the other components of $\boldsymbol{X}_{it}$.

   (c) Explain the significance of your findings in relation to Assumptions MLR2 and FE2.

3. Consider the panel data model

$$Y_{it} = \alpha + \beta X_{it} + \gamma W_{it} + A_i + V_{it}.$$

   Let $\boldsymbol{Y}_i \equiv [Y_{i1}, \ldots, Y_{iT}]$, $\boldsymbol{X}_i \equiv [X_{i1}, \ldots, X_{iT}]'$ and $\boldsymbol{W}_i \equiv [W_{i1}, \ldots, W_{iT}]'$. Assume that we have an i.i.d. sample $(\boldsymbol{Y}_i, \boldsymbol{W}_i, \boldsymbol{X}_i)$ for $i = 1, \ldots, N$ with finite fourth moments.

   (a) Define $\dot{X}_{it} \equiv X_{it} - \frac{1}{T}\sum_{s=1}^{T} X_{is}$ and $\dot{V}_{it} \equiv V_{it} - \frac{1}{T}\sum_{s=1}^{T} V_{is}$. Show that if $\mathbb{E}(V_{it} \,|\, \boldsymbol{X}_i) = 0$ for every t, then $\mathbb{E}(\dot{V}_{it} \,|\, \dot{X}_{it}) = 0$ for every t. Explain the significance in relation to Assumptions FE.

   (b) Suppose that $\mathbb{E}[V_{it}|\boldsymbol{X}_i, \boldsymbol{W}_i] = \mathbb{E}[V_{it}|\boldsymbol{W}_i] = \lambda + \delta W_{it}$ for every t. Construct a consistent estimator of β and justify its consistency. You may assume that FE2 holds when necessary.

   (c) Suppose instead that

$$\mathbb{E}[V_{it}|\boldsymbol{X}_i, \boldsymbol{W}_i] = \mathbb{E}[V_{it}|\boldsymbol{W}_i] = \lambda + \sum_{s=1}^{t} \delta W_{is},$$

   which depends on W_{is} for all $s \leq t$. Construct consistent estimators of, β, γ and δ, and justify their consistency. Explain why it is possible to consistently estimate both γ and δ separately in this case. You may assume that FE2 holds when necessary. *Hint: The first-differenced estimator may be easier to analyze than the fixed effects estimator in this problem.*

Stata/R

4. This question is based on "Estimating the Peace Dividend: The Impact of Violence on House Prices in Northern Ireland" by Besley and Mueller, which was published in *The American Economic Review* in 2012. The paper is available on Canvas, as is the dataset the authors used (`ireland2.dta`). Read enough of the paper to be able to answer the following questions.[1]

   (a) Is the panel balanced? What is the unit of observation i? What is the length of measurement for time t?

[1] You shouldn't have to read past pg. 818.

(b) Does the dataset fit into the asymptotic framework that we discussed in Lecture Note 3? Why or why not?

(c) Regardless of your answers to parts a) and b), replicate columns (1), (2), (3), (5) and (6) from Table 1.
Hints: (i) If `X` *is a variable in Stata, then* `LN.X` *is the Nth lag of that variable, where* `N` *is an integer. (ii) If you have a variable* `A` *that takes on discrete values then including* `i.A` *as a variable in a regression tells Stata to put a dummy for every value that* `A` *takes. (iii) Your answers to parts a) and b) do not affect the way you would run regressions in Stata. They are definitely important for interpreting these regressions, but that is not the point of this problem.*

5. This problem uses the dataset contained in `wagepan.dta`, which is available on Blackboard. The dataset is a panel of working age men for the years 1980-1987, taken from the National Longitudinal Survey.

   (a) Is the panel balanced? What is N and what is T?

   (b) Run a pooled regression of log wage on years of education. Do you think that it is a reliable estimate of the causal effect of an extra year of schooling? Why or why not? Is we were to interpret these estimates causally, what would be the percent increase in wages caused by an extra year of schooling?

   (c) Run a pooled regression of log wage on years of education with time (but not unit) fixed effects. Compare with your answer in (b).

   (d) Run a fixed effects regression of log wage on years of education with unit (but not time) fixed effects. What happened and why?

   (e) Run a pooled regression of log wage on union status. Interpret your results. Is your estimate a reliable indication of the causal returns to union status? Why or why not?

   (f) Of the variables `manuf, nrthcen, exper, expersq`, which should be controlled for and why? Report a pooled regression of your preferred specification and compare with the results from (e).

   (g) Run a fixed effects regression of log wage on union status with both unit and time time effects and the variables you thought should be controlled for. Compare to your results to (e) and your preferred specification in (f) and provide an explanation. If you didn't think that `exper` needed to be controlled for, run a separate fixed effects regression with `exper` added. What happened, and why?

6. This question is about the linear regression model

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

Assume that Y_i is generated according to the above and that the rest of the variables are determined as follows:

$$X_i = \delta + \lambda W_i + V_i$$

with $W_i \sim N(0, \sigma_W^2)$, $V_i \sim N(0, \sigma_V^2)$, $U_i \sim N(0, \sigma_U^2)$, all mutually independent.

(a) Write $\text{Var}(X_i)$ and $\text{Corr}(X_i, W_i)$ in terms of λ, σ_W and σ_V.

(b) Conduct a Monte Carlo simulation to approximate the finite-sample variance of $\widehat{\beta}_{mlr}$ from a multiple linear regression of Y_i on X_i, W_i and a constant. Let the sample size be $N = 500$, the number of simulations be $M = 5000$ and use the following values when generating the data:

$$\alpha = 0, \beta = 1, \gamma = 0, \delta = 0, \lambda = .3, \sigma_W = 1, \sigma_V = 2, \sigma_U = 2.$$

Hint: Study `regression.do` *from Lecture Note 2 and modify it accordingly.*

(c) Re-run the Monte Carlo simulation with the same parameter values as in (b) except change σ_U to 1. What happened to the variance of $\widehat{\beta}_{mlr}$ and why?

(d) Re-run the Monte Carlo simulation with the same parameter values as in (b) except change σ_V to 1.5. What happened to the variance of $\widehat{\beta}_{mlr}$ and why?

(e) Re-run the Monte Carlo simulation with the same parameter values as in (b) except change β to 0. What happened to the variance of $\widehat{\beta}_{mlr}$ and why?

(f) Re-run the Monte Carlo simulation with the same parameter values as in (b) except change λ to 0. What happened to the variance of $\widehat{\beta}_{mlr}$ and why?