Assignment 3: Panel, Nonparametric and Bootstrap (b11)

A. Colin Cameron U.C.-Davis

Data are at http://cameron.econ.ucdavis.edu/bgpe2011

1. Use data in file `mus08psidextract.dta`
We will do analysis similar to that in the slides, but with the only regressors `wks` and `ed`.
The model is $\texttt{lwage}_{it} = \alpha_i + \beta \texttt{wks}_{it} + \gamma \texttt{ed}_i + \varepsilon_{it}$, $i = 1, ..., N$, $t = 1, ..., T$.
We are interested in effect of `wks` (after controlling for `ed`).

(a) Run the following code

```
use mus08psidextract.dta, clear
regress lwage wks ed, noheader vce(cluster id)
estimates store POLS
xtreg lwage wks ed, re vce(robust) theta
estimates store RE
xtreg lwage wks ed, fe vce(robust)
estimates store FE
regress D.lwage D.wks D.ed, vce(cluster id)
estimates store FD
estimates table POLS RE FE FD, b(%9.5f) se stats(N)
```

(b) Are the estimates of β similar?

(c) Why are some of the estimates of γ equal to zero?

(d) If the model is correctly specified and the error α_i is correlated with the regressors while the error ε_{it} is uncorrelated with the regressors, which of any of the four estimates of β given above would you expect to be similar? Was this the case here?

(e) Re-estimate the pooled OLS model with OLS default standard errors. What is the lesson here?

(f) Re-estimate the fixed effects model with default standard errors. What is the lesson here?

2. Continue with the preceding example.
We use command `xtdata` which converts data x_{it} to $\bar{x}_i$ (option `be`), or to $x_{it} - \bar{x}_i + \bar{\bar{x}}$ (option `fe`) or to $x_{it} - \widehat{\theta}_i \bar{x}_i$ (option `re`). Here $\bar{\bar{x}} = \frac{1}{T} \sum_t \bar{x}_i = \frac{1}{NT} \sum_t x_{it}$ is the grand mean of x_{it}.

(a) Give command `scatter lwage wks || lfit lwage wks`
Does there appear to be a relationship between `lwage` and `wks` using total variation in the data.

(b) Give command `xtdata lwage wks ed, be`
and then `scatter lwage wks || lfit lwage wks`
Does there appear to be a relationship between `lwage` and `wks` using between variation in the data?

(c) Give command `use mus08psidextract.dta, clear`
and then `xtdata lwage wks ed, fe`
and then `scatter lwage wks || lfit lwage wks`
Does there appear to be a relationship between `lwage` and `wks` using within variation in the data?

(d) Now give command `scatter lwage wks || lfit lwage wks`
Explain what is happening here.

3. Consider panel data regression with a balanced panel and single regressor x_i.
The fixed effect estimator is OLS of $(y_{it} - \bar{y}_i)$ on $(x_{it} - x_i)$.
The first difference estimator is OLS of $(y_{it} - y_{i,t-1})$ on $(x_{it} - x_{i,t-1})$.

(a) When $T = 2$ (so two periods of data) express $(y_{it} - \bar{y}_i)$ as a function of y_{i1} and y_{i2}. What do you conclude about the relationship between the within and first difference estimator when $T = 2$?

(b) When $T = 3$ (so three periods of data) express $(y_{it} - \bar{y}_i)$ as a function of y_{i1}, y_{i2}, and y_{i3}. What do you conclude about the relationship between the within and first difference estimator when $T = 3$?

4. Nonparametric density estimation for log hourly wage.

(a) Read in data in file `nonparametric.dta`. Drop observations with low hourly wages (< 0.5) due to possible measurement error.

(b) Use command `histogram` to get a histogram density estimate for `lnhwage`.

(c) Use command `kdensity` to get a kernel density estimate for `lnhwage`.
How does this compare to the histogram?

(d) The kernel density estimate is smoother because it is essentially a rolling histogram, with averaging at many points, that is then graphed with estimates connected directly rather than by using a step function. To see this, obtain the kernel density estimates with evaluation at just 35 points, use the rectangle kernel, save the estimates and the evaluation points, and then . The commands are

```
kdensity lnhwage, lwidth(thick) n(35) gen(evalpoint density)kernel(rectangle)
scatter density evalpoint, connect(stairstep) lwidth(medthick)
```

How do the two estimates compare?

5. Nonparametric regression of hourly wage on years of schooling. [**Note: wage not ln(wage)**]

(a) Read in data in file `nonparametric.dta`. Drop observations with low hourly wages (< 0.5) due to possible measurement error. Also drop those with education $<= 8$ years as relatively few observations.

(b) Nonparametric regression is local averaging of y around a given value of x. A simple way to do this is to calculate the average of `hwage` at each level of education (in this example there are multiple y for most distinct x values). To do this give commands

```
bysort educ:  egen localaverage = mean(hwage)
graph twoway scatter localaverage educatn, c(l)
```

(c) One method for nonparametric regression is LOWESS. Use command `lowess` to perform nonparametric regression of `hwage` on `educ`.
How does this compare to the local average in part (b)?

(d) The LOWESS estimate is smoother because it is essentially a rolling local average of y, over a wide window around each value of x. (For example, at $x = 12$ we also include in the average values of y when $10 \leq x \leq 14$). To see this we do LOWESS with a narrow bandwidth of 0.2 (so $0.2 \times N$ observations are used for calculating smoothed values for each point of evaluation). The commands are

```
lowess hwage educatn, bwidth(0.2) generate(lowessfitted)
graph twoway scatter lowessfitted educatn, c(l) || scatter localaverage educatn, c(l)
```

How do the two estimates compare?

(e) Alternative nonparametric regression methods are kernel regression (command `lpoly`) and local linear linear (command `lpoly, degree(1)`). Do these and compare to Lowess in part (c).

6. Bootstrap estimate of standard error.
Give the following commands

```
use bootdata.dta, replace
poisson docvis chronic, vce(boot, reps(50) seed(10101))
bootstrap, reps(50) seed(10101) saving(bootoutput, replace): poisson docvis chronic
use bootoutput.dta
list
summarize
```

(a) Give the estimate of the standard error of $\widehat{\beta}_{\text{chronic}}$ from option `vce(boot)`

(b) Is this the same as that from the command `bootstrap?`
What if we had done a different number of bootstraps or a different seed?

(c) What were the estimated slope coefficients of `chronic` from the first three bootstraps.

(d) How does the standard deviation of estimated slope coefficients of `chronic` from the 50 bootstraps compare to the standard error estimate from part (a)?