Comments on chapters 13 and 14 in Wooldridge – Introductory Econometrics 2e

Alois Geyer
Institute for Financial Research
WU Vienna University of Economics and Business
Vienna Graduate School of Finance (VGSF)
alois.geyer@wu.ac.at
http://www.wu.ac.at/~geyer

December 2, 2020

p.426: *independently pooled cross sections*:

corresponds to *independent* samples

increases the precision of estimates

cross sections are not identically distributed

p.426: *panel data set*:

corresponds to *paired* samples

cross sections are not identically distributed

accommodate unobserved or omitted regressors

p.428: F-test for m restrictions and K regressors (incl. a constant):

$$F = \frac{(n-K)(R_u^2 - R_r^2)}{m(1-R_u^2)} \sim F(m, n-K)$$

R_r^2 ... R^2 of the restricted model, R_u^2 ... R^2 of the unrestricted model

$$F = \frac{(n-K)(\text{SSE}_r - \text{SSE}_u)}{m\text{SSE}_u} \sim F(m, n-K)$$

SSE_r ... sum of squared errors from the restricted model,
SSE_u ... sum of squared errors from the unrestricted model

p.428: $0.128 \cdot 4 = 0.512$ (4 ... high school takes *four* years longer than college)

p.428: *turning point of the quadratic*:

$$y = 0.532x - 0.0058x^2$$

$$\frac{\partial y}{\partial x} = 0.532 - 0.0058 \cdot 2 \cdot x = 0 \quad \Longrightarrow \quad x = \frac{0.532}{0.0058 \cdot 2} \approx 46$$

p.431: 27.2% *is the more accurate estimate*:

$$\Delta \ln \text{wage} = -0.317 \Delta x \quad \Longrightarrow \quad \ln \text{wage}_1 - \ln \text{wage}_0 = -0.317 \cdot 1$$

$$\text{wage}_1 = \text{wage}_0 \exp\{-0.317\} = \text{wage}_0 \cdot 0.728 \qquad 1 - 0.728 = 0.272$$

p.434: δ_0 *captures changes in* all *housing values*: in fact, δ_0=18790.3 only measures the change in prices *not near* the incinerator.

S.434: *provided we assume that houses both near and far from the site did not appreciate at different rates for other reasons*: this statement refers to the *parallel trends assumption*. If prices evolve *differently* over time because of other reasons, the estimated effect would be partially and falsely attributed to the reason represented by the dummy variable. A time-dummy cannot be used to account for this problem, because the regression already contains an interaction with time. The problem can only be overcome if other reasons can be quantified and the corresponding regressors are interacted with a time-dummy.

p.438: *unobserved factors affecting the dependent variable*:

basic model: $y_{it} = \beta_0 + \beta_1 x_{it} + a_i + u_{it}$
a_i ... *unobserved* or *fixed effect* (is responsible for *unobserved heterogeneity*)

Which problem is associated with the existence of a_i? If a_i and x_{it} are correlated, all estimates are biased and inconsistent (*omitted variable bias*)

p.439: $y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 x_{it} + v_{it} \qquad v_{it} = a_i + u_{it}$
estimates are unbiased and consistent only if v_{it} and x_{it} are uncorrelated

p.439: *Question 13.3: Show that* $\text{Cov}(v_{i1}, v_{i2})$=$\text{Var}(a_i)$:

$$v_{i1} = a_i + u_{i1} \qquad v_{i2} = a_i + u_{i2}$$

$$\text{assumptions: } [a_i] = 0,\ [u_{i1}] = 0,\ [u_{i2}] = 0,\ \text{Cov}(u_{i1}, u_{i2}) = 0$$

$$\text{Cov}(a_i, u_{i1}) = 0,\ \text{Cov}(a_i, u_{i2}) = 0$$

$$\text{Cov}(v_{i1}, v_{i2}) = [(a_i + u_{i1})(a_i + u_{i2})] =$$

$$= [a_i^2 + a_i u_{i1} + a_i u_{i2} + u_{i1} u_{i2}] = [a_i^2] = \text{Var}[a_i]$$

This fact will become relevant in the context of the *random effects* model (see comment related to p.470).

p.440: *standard errors in this equation are incorrect*: standard errors are computed under the assumption of no serial correlation

p.440: *main reason to collect panel data*: using a single *cross section* creates an omitted variable problem

p.440: *first-differenced equation* $\Delta y_{it} = \beta_1 \Delta x_{it} + \Delta u_{it}$: β_1 does not change by taking differences! β_0 in the original equation gets lost.

p.441: Δx_i *must have some variation across* i: if the variance of Δx_i is low the standard error of its coefficient will be high. This is true for the regressor *educ* in *Example* 13.5. If Δx_i has zero variance it must be eliminated. This problem occurs when estimating equation (13.19) since only two years are considered.

p.441: *The only other assumption*: it should be added that Δu_i in (13.17) must not be autocorrelated. In case of only two periods (as here) this condition is fulfilled by construction. In case of more than two periods this property must be fulfilled and checked (see second paragraph on p.449: *When using more than two time periods, we must assume that* Δu_{it} *is uncorrelated over time for the usual standard errors and test statistics to be valid*).

p.441: the coefficient 15.40 in equation (13.18) *corresponds to* the coefficient of the dummy $d87$ in equation (13.16) (see comment on p.467).

p.445: the results in section 13.4 are based on using the years 1987 and 1988 only!

S.448: *We can also appeal to asymptotic results*: this refers to the consistency property if u_t and regressors are uncorrelated (which is weaker than the assumption of *strong exogeneity*).

p.448: *Therefore (13.30) does not contain an intercept*: taking differences of the dummies $d2$ and $d3$ in equation (13.28) results in

t	c	d_2	d_3	Δd_2	Δd_3
1	1	0	0	.	.
2	1	1	0	1	0
3	1	0	1	–1	1
1	1	0	0	.	.
2	1	1	0	1	0
3	1	0	1	–1	1

There exist linear combinations of Δd_2 and Δd_3 which are identical to c; e.g. $(1 + \Delta d_2)/2 + \Delta d_3$; thus, only two of the three variables c, Δd_2 and Δd_3 can be used.

Estimating equation (13.30) – without intercept – results in (p-values in parenthesis)

$$\widehat{\mathsf{DLOG(SCRAP)}} = -0.139750.1044\mathsf{D(D88)} - 0.426880.0003\mathsf{D(D89)} - 0.08310.368\mathsf{D(GRANT)}.$$

Adding an intercept results in the EViews error message Near singular matrix, which indicates the identity of the linear combination of Δd_2 and Δd_3, and c. Removing the dummy D88 from (13.30) and using an intercept instead results in (p-values in parenthesis)

$$\widehat{\mathsf{DLOG(SCRAP)}} = -0.139750.1044 - 0.147370.2275\mathsf{D(D89)} - 0.08310.368\mathsf{D(GRANT)}.$$

When comparing the coefficients of the dummy variables it must be taken into account that D(D88) equals +1 in 1988 and −1 in 1989! Thus, the constant on the right hand side of the first equation in 1989 is: +0.13975−0.42688=−0.2871. This corresponds exactly to the constant in 1989 from the second equation: −0.13975−0.14737= −0.2871.

p.449: *The correlation between* Δu_{it} *and* $\Delta u_{i,t+1}$ *can be shown to be –.5*:

$y_t = u_t - u_{t-1} \qquad [u_t] = 0 \quad [u_t u_{t-1}] = 0$ (u_t is not autocorrelated)

autocovariance of y_t: $\gamma_1 = E[y_t y_{t-1}] = [(u_t - u_{t-1})(u_{t-1} - u_{t-2})]$

$\gamma_1 = [(u_t u_{t-1} - u_t u_{t-1} - u_{t-1}^2 + u_{t-1} u_{t-2})]$

$= [u_t u_{t-1}] - [u_t u_{t-1}] - [u_{t-1}^2] + [u_{t-1} u_{t-2}] = -[u_{t-1}^2] = -[u_t]$

autocorrelation of y_t: $\rho_1 = \dfrac{\gamma_1}{[y_t]} = \dfrac{-[u_t]}{[u_t] + [u_{t-1}]} = -0.5$

p.449: *random walk*: y_t is a *random walk* if $y_t - y_{t-1}$ is not autocorrelated.

p.449: *feasible GLS* or *Prais-Winsten* vs. *Cochrane-Orcutt*: these approaches correct for the autocorrelation of errors. The observed variables (y and x) are transformed on the basis of ρ_1 as follows: $y_t^* = y_t - \rho_1 y_{t-1}$ (similarly for x_t). Using *Cochrane-Orcutt* the first observation gets lost. *Prais-Winsten* overcomes this problem.

p.452: *the police variable might be endogenous* and *this additional form of endogeneity*: In this example it is argued that the regressor *polpc* depends on the *expected* but unobservable *crime rate*. Thus, the expected *crime rate* is part of the error term. If the regressor *polpc* depends on this variable, the error term and the regressor are correlated (which violates the exogeneity assumption).

p.463: *Table 14.1*: note that the results in this table are based on the fact that grant_1 is *assumed 0 in 1987*. As a matter of fact grant_1 should be coded as 'NA' in the first year of each *cross section*.

p.464: *The R-squared given in Table 14.1 is based on the within transformation*: Note that there is major difference between the R^2 from using dummy variables to account for fixed effects or using the demeaned variables (*within transformation*). Using a dummy variable for each cross section usually produces a rather high R^2 (see p.466). If different orders of magnitude of y_{it} in each cross section are the main source of variance in the dependent variable, this will be captured by the cross section dummies (many degrees of freedom). This source of variance is eliminated when using demeaned variables, and R^2 measures "the amount of time variation ... that is explained by the time variation in the explanatory variables".

p.464: *time-constant variables cannot be included*: a disadvantage of FD and FE is that variables which are constant over time or change deterministically over time (e.g. a linear time trend) cannot be used. Interactions with dummies are possible, however.

Differencing interaction terms: suppose the original equation is

$$y_t = c + b_1 D + b_2 x_t + b_3 D \cdot x_t + e$$

(e.g. D is a dummy distinguishing two groups). This implies

$$\begin{array}{ll} D = 0: & y_t = c + b_2 x_t \\ D = 1: & y_t = (c + b_1) + (b_2 + b_3) x_t \end{array}$$

Taking differences correctly means that the variable z_t=$D \cdot x_t$ must be differenced:

$$\Delta y_t = b_2 \Delta x_t + b_3 \Delta z_t + \Delta e.$$

In other words, D and x_t *must not* be differenced separately and multiplied afterwards (D would be the same in each *cross section* and ΔD would be zero). The first differences are given by

$$\begin{array}{ll} D = 0: & \Delta y_t = b_2 \Delta x_t \\ D = 1: & \Delta y_t = (b_2 + b_3) \Delta x_t. \end{array}$$

Thus, in the FD model b_3 is the additional slope with respect to x_t in the second group.

p.466 *The R-squared from the dummy variable regression is usually rather high*: see comment on p.464.

p.466 *Some econometrics packages ... report an "intercept"*: In fact, the constant 75.4 should not (cannot) be part of a FE model. However, "*EViews automatically includes a constant term so that the fixed effects estimates sum to zero and should be interpreted as deviations from an overall mean* (EViews Users Guide p.837)."

p.467: *when $T = 2$, the FE and FD estimates and all test statistics are identical*: This statement must be made more precise. For example, estimating equation (13.18) as a FE model results in (p-values in parenthesis)

$$\hat{crmrte} = 100.9350.000 - 0.0180.976 unem.$$

Adding the dummy *d87* results in

$$\hat{crmrte} = 75.40.000 + 15.40.002 d87 + 2.220.015 unem,$$

which agrees with (13.18). The coefficient 15.4 corresponds to the constant in equation (13.18).

distinguishing between FD and FE:

if u_{it} is (strongly) autocorrelated: FD

if u_{it} is not autocorrelated: FE

p.467: *It is difficult to test whether the u_{it} are serially uncorrelated after FE estimation*: Note that the estimated residuals from the FE regression $\ddot{u}_{it}$ are not equal to the "true" residuals u_{it} which are the object of this assumption.
indirect test for autocorrelation:

estimated FD model

check Δu_{it} for autocorrelation

if Δu_{it} is not autocorrelated: FD

If Δu_{it} is *negatively* autocorrelated: FE
p.467: *If there is substantial negative serial correlation in the Δu_{it}*: if u_{it} is not autocorrelated, taking differences makes Δu_{it} negatively autocorrelated (see comment on p.449).

Alternative: FE and AR(1) correction

p.469 (line 3): *sample section problem* *sample* selection *problem*

p.469 *a_i is* uncorrelated *with each explanatory variable*: the *random effects* (RE) model should be chosen if a_i and x_{it} are uncorrelated, since FD and FE are inefficient in this case.

p.470: *there is no need for panel data at all*: this is a *preliminary* implication of $[a_i, x_{it}]$=0.

p.470: *But it ignores a key feature of the model*: if the RE assumption holds (a_i and x_{it} are uncorrelated) the errors v_{it} *must be* positively autocorrelated! For a proof of the positive correlation between v_{it} and v_{is} see the comment on p.439. This positive autocorrelation requires GLS estimation.

p.471: *the whole reason for using panel data*: if panel data is used to account for *unobserved heterogeneity* (i.e. it is assumed that there are fixed effects – see comment on p.438), a RE model is most likely not the first choice.

p.472: exper *is dropped in the FE analysis (but* exper^2 *remains)*: the reason why *exper* has to be dropped is given in the next sentence: because the *regression also contains a full set of year dummies*. Since *exper* increases every year by a constant amount (one year), this variable is perfectly correlated with the year dummies.

p.473: *The estimate of λ for the random effects estimation*: $\hat{\lambda}$=0.643 can be obtained in EViews from the estimated RE model using equation (14.10), and substituting σ_u by the value 0.350990 from the line Idiosyncratic random and σ_a by the value 0.324603 from the line Cross-section random. Using the results from *Example* 14.4 one obtains

$$1 - [0.35099^2/(0.35099^2 + 8 \cdot 0.324603^2)]^{1/2} = 0.643,$$

where T=8 (the number of years in each *cross section*).

p.473: RE or FE?

relevant assumption/fact: a_i and x_{it} are uncorrelated

if a_i and x_{it} are correlated RE is inconsistent

Summary:

$$y_{it} = \beta_0 + \beta_1 x_{it} + a_i + u_{it}$$

$$y_{it} = \beta_0 + \beta_1 x_{it} + v_{it}$$

FD: $\Delta y_{it} = \beta_1 \Delta x_{it} + \Delta u_{it}$

FE: $(y_{it} - \bar{y}_i) = \beta_1 (x_{it} - \bar{x}_i) + (v_{it} - \bar{v}_i)$

RE: $(y_{it} - \lambda \bar{y}_i) = \beta_0 (1 - \lambda) + \beta_1 (x_{it} - \lambda \bar{x}_i) + (v_{it} - \lambda \bar{v}_i)$

If *unobserved effects* a_i are

correlated with regressors x_{it}: FD or FE

uncorrelated with regressors x_{it}: RE

If residuals u_{it} are

(strongly) autocorrelated: FD or FE with AR-correction

not autocorrelated: FE

Hausman-test to distinguish between FE and RE: H_0: $[a_i, x_{i,t}]$=0