Applied Methods for PhD, Week 07

Basic Linear Panel Data

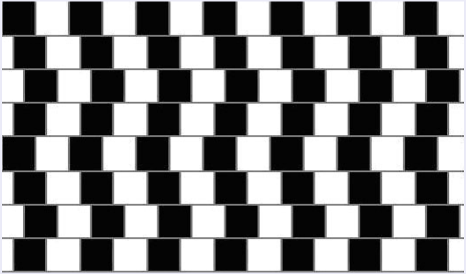
Michael E. Kummer
Theoretical Slide Set, inspired byStephen Kastoryano

NovaSBE, OTIM

PANEL DATA

- Panel data follow a cross-section of individuals over several time-periods.
- Outcomes and characteristics are observed at multiple points in time.
- Advantages of panel data (compared to cross-section data):
  - Control for unobserved time-invariant heterogeneity (without strong functional form restrictions or instrumental variables).
  - Can provide information on dynamics.

Figure: The fuzzy lines of parallel universes.



PSEUDO-PANEL DATA

- Pseudo-panel data: Repeated cross-section, where cohorts of individuals with the same year-of-birth are treated as panels.
- Usually higher response rate in pseudo-panel data than in true panel data.
- Disadvantage of pseudo-panel data is that dynamics cannot be studied, as it is unobserved how individuals move within the distribution of outcomes.
- A panel data set is *balanced* if all individuals are observed at each moment in time. The panel data set is *unbalanced* if some of the observations are missing.
- Unbalanced panel data can be caused by *attrition* or *sample-selection*. This can be less problematic than in cross-sections (depending on the source of attrition).
- For now, assume that we observe a balanced panel data set, which contains N individuals, which are observed over T periods.
- Panel data sets are often 'short', which implies that T is small while N is large.

LINEAR PANEL DATA MODEL

- Basic linear panel data model:

$$Y_{it} = \alpha + X_{it}\beta + \eta_i + \lambda_t + U_{it} \qquad i = 1, \dots, N \quad t = 1, \dots, T$$

- η_i is an individual specific effect, which captures components that are unobserved by the econometrician.
- λ_t is some function (polynomial) in time (henceforth, for simplicity of notation, we include λ_t in X_{it}).
- Often, the individual specific effect contains omitted variables which are correlated with the regressors, $\mathrm{E}[\eta_i | X_{i1}, \dots, X_{it}] \neq 0$.

LINEAR PANEL DATA MODEL

- Classical Example: Agricultural production (Mundlak 1961, Chamberlain 1984)
  - $Y_{it} = log(\text{output})$, $X_{it} = log(\text{input})(\text{labour})$, $\eta_i =$an input that remains constant over time (soil quality), $U_{it} =$a stochastic input outside the farmer's control (rainfall)
  - Questionable to assume labour is uncorrelated to soil quality.
- Suppose that η_i is ignored and the (pooled) model is estimated by OLS

$$Y_{it} = \alpha + X_{it}\beta + U^*_{it} \qquad \text{with} \quad U^*_{it} = \eta_i + U_{it}$$

- This assumes $\mathrm{E}[\eta_i | X_{i1}, \ldots, X_{it}] = 0$, and violation of this assumption causes biased and inconsistent estimators for β.

EXOGENEITY

- The regressors are strictly exogenous if (also conditional on the unobserved individuals specific effect)

$$\mathrm{E}[U_{it}|X_{i1},\ldots,X_{iT},\eta_i]=0 \qquad i=1,\ldots,N \quad t=1,\ldots,T$$

- Intuitively, you should think about whether your regressors contain
  - *lagged-endogenous* variables ($\mathrm{E}[U_{it-1}|X_{it}]\neq 0$). X_{it} should not contain 'intermediate outcomes'.
  - *feedback*, i.e. X_{it} depends on Y_{it-1} ($\mathrm{E}[U_{it}|X_{it+1}]\neq 0$). Ex: treatment and unobserved selection into treatment. Next lecture: what type of feedback can be modeled?
- In the statistical model, lagged-endogenous variables and feedback are the same.
- The alternative to strict exogeneity is sequential exogeneity (or weak exogeneity, which implies

$$\mathrm{E}[U_{it}|X_{i1},\ldots,X_{it},\eta_i]=0 \qquad i=1,\ldots,N \quad t=1,\ldots,t$$

- This allows for lagged-endogenous variables as regressors and feedback, i.e. the model can be dynamic.
- Weak exogeneity implies only $\mathrm{E}[U_{it}|X_{it},\eta_i]=0 \qquad i=1,\ldots,N \quad t=1,\ldots,T$.

STATIC FIXED EFFECT MODEL

- Consider the static fixed-effect model:

$$Y_{it} = \alpha + X_{it}\beta + \eta_i + \lambda_t + U_{it}$$

- X_{it} is a $(1 \times K)$-vector of strictly exogenous regressors and U_{it} is independent over time and across individuals.
- We do not rule out correlation between η_i and X_{it}.
- Either α should be normalized or a restriction on the η_i-parameters is required. Common to impose $\alpha = 0$.
- Under these assumptions β can be estimated using within estimation.

FIRST-DIFFERENCE ESTIMATION

- Instead of within estimation, take first-differences (also a fixed effect estimator)

$$\begin{aligned} Y_{it} - Y_{it-1} &= X_{it}\beta + \eta_i + U_{it} - X_{it-1}\beta - \eta_i - U_{it-1} \\ &= (X_{it} - X_{it-1})\beta + (U_{it} - U_{it-1}) \qquad t = 2, \dots, T \end{aligned}$$

- or $\Delta Y_{it} = \Delta X_{it}\beta + \Delta U_{it}$.
- Taking first-differences eliminates η_i from the model.
- Estimating by OLS we obtain the first-difference estimator $\widehat{\beta}_{fd}$.
- The first-difference estimator requires $\mathrm{E}[\Delta X_{it}\Delta U_{it}] = 0$ for consistency.
- If the regressors are only weakly exogenous, the first-difference estimator is not necessarily consistent ($Cov(U_{it-1}, X_{it})$ will not be 0).

WITHIN ESTIMATION

- Within estimation is a fixed-effect methods
- It does not impose any stochastic structure on η_i (as opposed to random effects).
- In the first step, averages are taken over time for all individuals

$$\overline{Y}_i = \overline{X}_i \beta + \eta_i \overline{U}_i$$

$$\overline{Y}_i = \frac{1}{T}\sum_{t=1}^{T} Y_{it} \qquad \overline{X}_i = \frac{1}{T}\sum_{t=1}^{T} X_{it} \qquad \overline{U}_i = \frac{1}{T}\sum_{t=1}^{T} U_{it}$$

- Next subtract $\overline{Y}_i$ from Y_{it},

$$Y_{it} - \overline{Y}_i = X_{it}\beta + \eta_i + U_{it} - \overline{X}_i\beta - \eta_i - \overline{U}_i = (X_{it} - \overline{X}_i)\beta + (U_{it} - \overline{U}_i)$$

- Subtracting the mean eliminates η_i from the estimation.

WITHIN ESTIMATION

- This implies that we get the estimating equation

$$\tilde{Y}_{it} = \tilde{X}_{it}\beta + \tilde{U}_{it}$$

with

$$\tilde{Y}_{it} = Y_{it} - \overline{Y}_i \quad \tilde{X}_{it} = X_{it} - \overline{X}_i \quad \tilde{U}_{it} = U_{it} - \overline{U}_i$$

- Estimating by OLS we obtain within estimator $\widehat{\beta}_{within}$.
- Within estimator requires $\mathrm{E}[\tilde{X}_{it}\tilde{U}_{it}] = 0$ for consistency.
- Under homoskedasticity, $Var(U_{it}|X_{i1}, \dots, X_{it}, \eta_i) = \sigma^2$

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\widehat{U}_{it}^2}{N(T-1)-K} \quad \text{with} \quad \widehat{U}_{it} = \tilde{Y}_{it} - \tilde{X}_{it}\widehat{\beta}_{within}$$

- Note that the denominator is $N(T-1)-K$ instead of $NT-K$ (as would be used by standard OLS packages).

ESTIMATING INDIVIDUAL SPECIFIC EFFECTS

- Finally, the estimators for the individual specific effects are

$$\widehat{\eta}_i = \overline{Y}_i - \overline{X}_i \widehat{\beta}_{within} \qquad i = 1, \dots, N$$

- If N is not too large, one could simply include dummy variables for each individual and estimate the original model by OLS. This provides the within estimators and $\widehat{\eta}_i$ in a single step.
- Questions:
  - Which variation in the data identifies parameters β? What does this imply for time-invariant covariates?
  - For estimators η_i and β to be consistent, must asymptotics imply $T \to \infty$? $N \to \infty$? or both?

WITHIN ESTIMATION QUESTIONS

- Some remarks:
  - The parameters β are identified due to (within) variation in X_{it} over time.
  - Influence of time-invariant covariates can not be estimated.
  - Estimators for η_i and β are consistent if the asymptotics imply that T becomes large.
  - If instead T is fixed and N goes to infinity, only $\widehat{\beta}_{within}$ within is consistent, but $\widehat{\eta_i}$ is not (so called *incidental parameters*).
  - Problem for Non-linear panel: *incidental parameters problem*: What happens when N grows large and within estimation is not possible?

WITHIN VS. FIRST-DIFFERENCE?

- If $T = 2$, the within estimator and first-difference estimator are the same.
- For $T > 2$, if within estimates differ much from first-difference estimates, then either the assumption of strict exogeneity is violated or the model is incorrectly specified (important time-varying regressors are missing).
- If U_{it} is uncorrelated over time, the within estimator is more efficient than the first-difference estimator.
- If the U_{it} follow random walk $U_{it} = U_{it-1} + error_{it}$, the first-difference estimator is more efficient.
- In many cases, the serial correlation is probably going to lie somewhere between these two extremes.
- If strict exogeneity is violated, the first-difference estimator and the within estimator become both inconsistent and have different probability limits.

SERIAL CORRELATION IN FIXED EFFECTS

- In case of serial correlation within estimation is consistent, but the standard errors of the estimators should be corrected (problem increases with large T).
- The most-often used method for computing correct standard errors is Newey-West.
- Newey-West uses that

$$\widehat{Var}\left(\widehat{\beta}_{within}\right) = \widehat{A}^{-1}\widehat{B}\widehat{A}^{-1}$$

  with

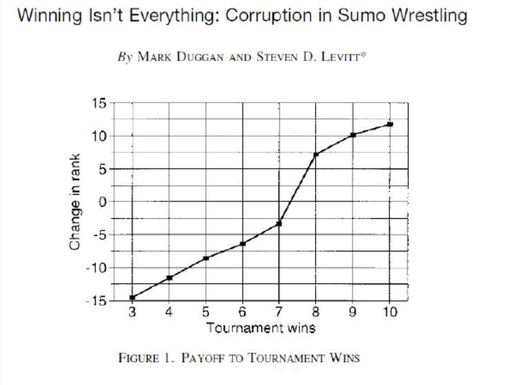
$$\widehat{A} = \sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{X}'_{it}\widetilde{X}_{it} \qquad \widehat{B} = \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}\widehat{U}_{it}\widehat{U}_{is}\widetilde{X}'_{is}\widetilde{X}_{it}$$

- where $\widehat{U}_{it} = \widetilde{Y}_{it} - \widetilde{X}_{it}\widehat{\beta}_{within}$.
- STATA within estimation: **xtset idvar timevar** followed by **xtreg depvar indvar, fe robust**.
- STATA first-difference estimation: **reg (or areg) d.depvar d.indvar, cluster(idvar)**

DISADVANTAGES OF THE FIXED-EFFECT MODEL

- Time-invariant regressors cannot be included in fixed-effect estimation.
- These variables drop out when taking all variables in derivation of their sample means (time-demeaning) so their coefficients are unidentified from the individual specific effect.
- Out of sample prediction is impossible. For individuals not included in the panel, one cannot 'observe' η_i. Therefore, even if the values of all regressors for this individual are observed, it is still impossible to predict an outcome.
- If most variation in time-varying regressors is *between* individuals, parameter estimates might not be very precise.

DUGGAN AND LEVITT (2002)



Winning Isn't Everything: Corruption in Sumo Wrestling

By Mark Duggan and Steven D. Levitt*

Figure 1. Payoff to Tournament Wins

- Each win equivalent to ~ 3 jumps in ranking with nonlinearity at 7 to 8 wins
- A wrestler who achieves a winning record (eight wins or more, known as kachi-koshi) is guaranteed to rise up the official ranking.

DUGGAN AND LEVITT (2002)

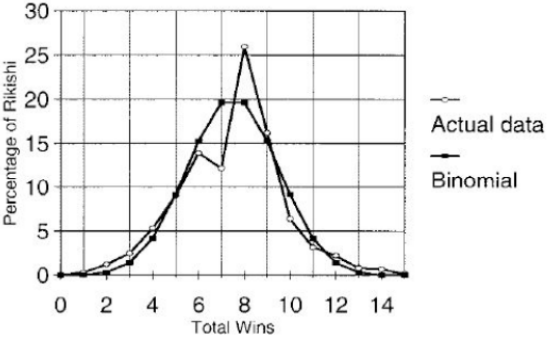
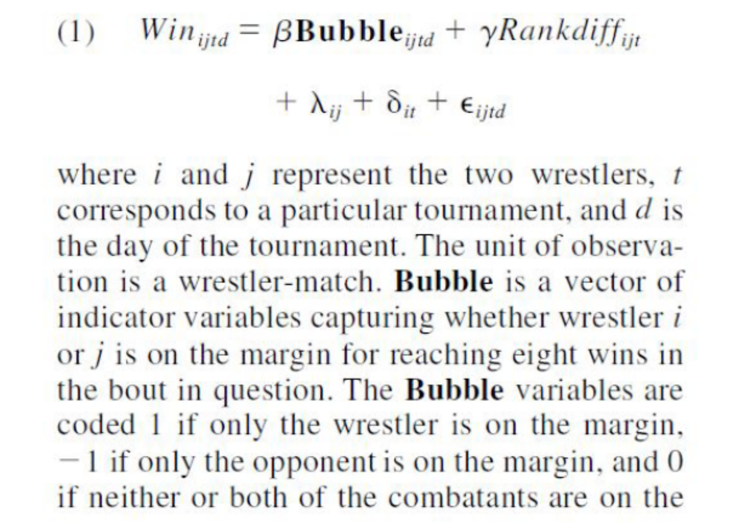


FIGURE 2. WINS IN A SUMO TOURNAMENT
(ACTUAL VS. BINOMIAL)

DUGGAN AND LEVITT (2002)



$$(1) \quad Win_{ijtd} = \beta \mathbf{Bubble}_{ijtd} + \gamma Rankdiff_{ijt} + \lambda_{ij} + \delta_{it} + \epsilon_{ijtd}$$

where i and j represent the two wrestlers, t corresponds to a particular tournament, and d is the day of the tournament. The unit of observation is a wrestler-match. **Bubble** is a vector of indicator variables capturing whether wrestler i or j is on the margin for reaching eight wins in the bout in question. The **Bubble** variables are coded 1 if only the wrestler is on the margin, -1 if only the opponent is on the margin, and 0 if neither or both of the combatants are on the

DUGGAN AND LEVITT (2002)

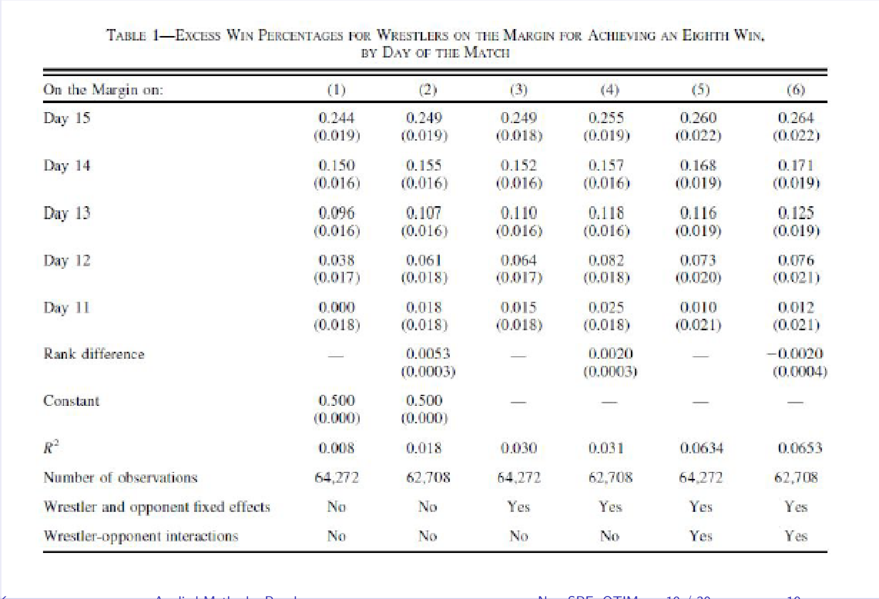


TABLE 1—EXCESS WIN PERCENTAGES FOR WRESTLERS ON THE MARGIN FOR ACHIEVING AN EIGHTH WIN, BY DAY OF THE MATCH

On the Margin on:	(1)	(2)	(3)	(4)	(5)	(6)
Day 15	0.244 (0.019)	0.249 (0.019)	0.249 (0.018)	0.255 (0.019)	0.260 (0.022)	0.264 (0.022)
Day 14	0.150 (0.016)	0.155 (0.016)	0.152 (0.016)	0.157 (0.016)	0.168 (0.019)	0.171 (0.019)
Day 13	0.096 (0.016)	0.107 (0.016)	0.110 (0.016)	0.118 (0.016)	0.116 (0.019)	0.125 (0.019)
Day 12	0.038 (0.017)	0.061 (0.018)	0.064 (0.017)	0.082 (0.018)	0.073 (0.020)	0.076 (0.021)
Day 11	0.000 (0.018)	0.018 (0.018)	0.015 (0.018)	0.025 (0.018)	0.010 (0.021)	0.012 (0.021)
Rank difference	—	0.0053 (0.0003)	—	0.0020 (0.0003)	—	−0.0020 (0.0004)
Constant	0.500 (0.000)	0.500 (0.000)	—	—	—	—
R^2	0.008	0.018	0.030	0.031	0.0634	0.0653
Number of observations	64,272	62,708	64,272	62,708	64,272	62,708
Wrestler and opponent fixed effects	No	No	Yes	Yes	Yes	Yes
Wrestler-opponent interactions	No	No	No	No	Yes	Yes

STATIC RANDOM-EFFECT MODEL

- The specification of the static random-effect model is the same as the static fixed-effect model
$$Y_{it} = \alpha + X_{it}\beta + \eta_i + U_{it}$$
- But η_i is assumed to have a stochastic structure with $\mathrm{E}[\eta_i | X_{i1}, \ldots, X_{iT}] = 0$ and $Var(\eta_i) = \sigma^2_\eta$.
- And η_i are uncorrelated across individuals, $Cov(\eta_i, \eta_j) = 0$ for $i \neq j$.
- Furthermore, U_{it} are usual error terms, i.e. $\mathrm{E}[U_{it} | X_{i1}, \ldots, X_{iT}, \eta_i] = 0$ and $Var(U_{it}) = \sigma^2_u$. And U_{it} independent across individuals and across time.
- Finally, η_i and U_{it} independent of each other and $V_i = \eta_i + U_i$ is *composite error term*.
- In brief: $U_{it} | (X_{i1}, \ldots, X_{iT}, \eta_i) \sim IID(0, \sigma^2_u)$, $\eta_i \sim IID(0, \sigma^2_\eta)$ and U_{it}, η_i independent.

RANDOM-EFFECTS SPECIFICATION

- For ease of exposition, we stack the observation of each individual

$$Y_i = X_i\beta + e_T\eta_i + U_i$$

- where

$$Y_i = \begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{iT} \end{pmatrix} \qquad X_i = \begin{pmatrix} X_{i1} \\ \vdots \\ X_{iT} \end{pmatrix} \qquad U_i = \begin{pmatrix} U_{i1} \\ \vdots \\ U_{iT} \end{pmatrix}$$

- and e_T is a (T x 1)-vector with all elements equal to 1. The random-effect specification implies

$$Var(e_T\eta_i + U_i) = \sigma_u^2 I_T + \sigma_\eta^2 e_T e_T' = \sigma_u^2\left(I_T + \frac{\sigma_\eta^2}{\sigma_u^2} e_T e_T'\right) = \sigma_u^2\Omega$$

- with I_T the identity matrix of size T.

RANDOM-EFFECTS ESTIMATION

- If σ_u^2 and σ_η^2 are known, the *Generalized Least Squares* (GLS) estimator for β is the *Best Linear Unbiased Estimator* (BLUE).

$$\widehat{\beta}_{GLS} = \sum_{i=1}^{N} (X_i' \Omega^{-1} X_i)^{-1} \sum_{i=1}^{N} X_i' \Omega^{-1} Y_i$$

- However, in most cases σ_u^2 and σ_η^2 are unknown, so one has to apply feasible GLS, which proceeds in the following way.

1. In the first-step within estimation is used to obtain an estimate for σ_u^2.
2. Next perform *between estimation* by using OLS on

$$\overline{Y}_i = \overline{X}_i \beta + \varepsilon_i$$

Where $\sigma_\varepsilon^2 = \sigma_\eta^2 + \frac{\sigma_u^2}{T}$ as $\varepsilon_i = \eta_i + \overline{U}_i$. Therefore, $\widehat{\sigma_\varepsilon^2} = \widehat{\sigma_\eta^2} + \frac{\widehat{\sigma_u^2}}{T}$.

3. Finally, substitute σ_u^2 and σ_η^2 in the expression for Ω to obtain $\widehat{\Omega}$ and perform GLS

$$\widehat{\beta}_{FGLS} = \sum_{i=1}^{N} (X_i' \widehat{\Omega}^{-1} X_i)^{-1} \sum_{i=1}^{N} X_i' \widehat{\Omega}^{-1} Y_i$$

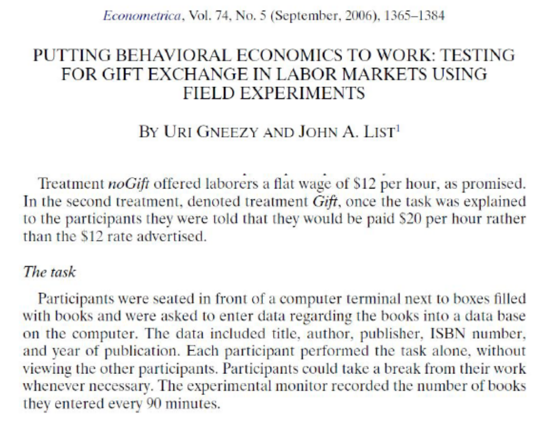
RANDOM-EFFECTS ESTIMATION

- In STATA: **xtset idvar timevar** followed by **xtreg depvar indvar, re robust**.
- Remark: The within estimator in the first step does not exploit $\mathrm{E}[\eta_i|X_{i1},\ldots,X_{iT}]=0$ and therefore this feasible GLS estimator is not efficient.
- β could be estimated consistently by Maximum Likelihood under the additional assumptions $\eta_i \sim N(0,\sigma_\eta^2)$ and $U_{it} \sim N(0,\sigma_u^2)$. Consistency of the parameters only requires either N or T to go to infinity.
- As in FE, one can add additional *error component* U_t, which varies over time but is the same for individuals.

$$Y_{it} = \alpha + X_{it}\beta + \eta_i + \lambda_t + U_{it}$$

- Can you think of a way to change this model which allow correlation between individual random effects η_i and covariates?

GNEEZY AND LIST 2006



Econometrica, Vol. 74, No. 5 (September, 2006), 1365–1384

PUTTING BEHAVIORAL ECONOMICS TO WORK: TESTING FOR GIFT EXCHANGE IN LABOR MARKETS USING FIELD EXPERIMENTS

BY URI GNEEZY AND JOHN A. LIST[1]

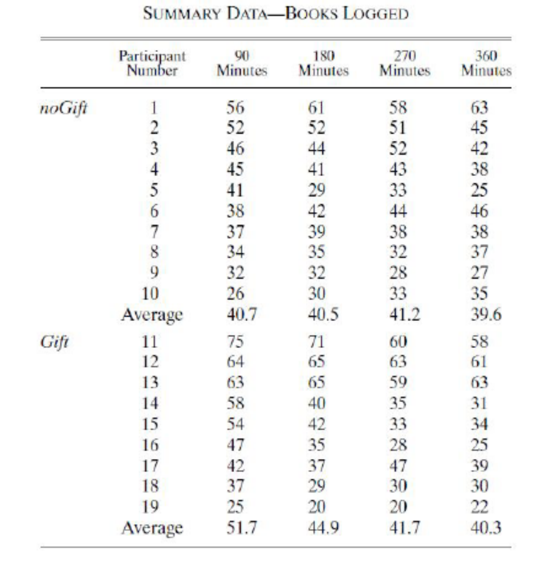
Treatment *noGift* offered laborers a flat wage of $12 per hour, as promised. In the second treatment, denoted treatment *Gift*, once the task was explained to the participants they were told that they would be paid $20 per hour rather than the $12 rate advertised.

The task

Participants were seated in front of a computer terminal next to boxes filled with books and were asked to enter data regarding the books into a data base on the computer. The data included title, author, publisher, ISBN number, and year of publication. Each participant performed the task alone, without viewing the other participants. Participants could take a break from their work whenever necessary. The experimental monitor recorded the number of books they entered every 90 minutes.

- Use random-effects since treatment dummy variable is static so using fixed effects would violate rank condition (no time variation in treatment).

GNEEZY AND LIST 2006



SUMMARY DATA—BOOKS LOGGED

	Participant Number	90 Minutes	180 Minutes	270 Minutes	360 Minutes
noGift	1	56	61	58	63
	2	52	52	51	45
	3	46	44	52	42
	4	45	41	43	38
	5	41	29	33	25
	6	38	42	44	46
	7	37	39	38	38
	8	34	35	32	37
	9	32	32	28	27
	10	26	30	33	35
	Average	40.7	40.5	41.2	39.6
Gift	11	75	71	60	58
	12	64	65	63	61
	13	63	65	59	63
	14	58	40	35	31
	15	54	42	33	34
	16	47	35	28	25
	17	42	37	47	39
	18	37	29	30	30
	19	25	20	20	22
	Average	51.7	44.9	41.7	40.3

FIXED EFFECTS VS. RANDOM EFFECTS

- Random effects can estimate the coefficients of time-invariant regressors.
- Random effects can be used to make predictions outside the sample for which time-invariant regressors are informative.
- Random effects assumes a stochastic structure on the individual specific effects, so it makes stronger assumptions than fixed effects.
- Fixed effects estimation is robust against departures from the imposed stochastic structure on the individual specific effects, but less efficient than random effects estimation if the stochastic structure is correct.
- The stochastic structure on the individual specific effect implies that the parameters β could be estimated using a single cross-section. Random effects only needs panel data to disentangle η_i from U_{it}.

HAUSMAN TEST

- An alternative approach to testing between random effects and fixed effects is proposed by Hausman (Ectra, 1978).
- Under $H_0 : \mathrm{E}[\eta_i | X_{i1}, \ldots, X_{iT}] = 0$, both random-effects and fixed-effects estimators are consistent, but the random-effect estimator is more efficient.
- Under $H_1 : \mathrm{E}[\eta_i | X_{i1}, \ldots, X_{iT}] \neq 0$, only the fixed-effects estimator is consistent.
- Therefore, investigate the test statistic

$$T = \left(\widehat{\beta}_{FE} - \widehat{\beta}_{RE}\right)\left(Var(\widehat{\beta}_{FE}) - Var(\widehat{\beta}_{RE})\right)^{-1}\left(\widehat{\beta}_{FE} - \widehat{\beta}_{RE}\right)$$

- which converges to a χ^2-distribution with the degrees of freedom equal to the number of time-varying regressors.
- Hausman-test is asymptotically equivalent to Mundlak-test (STATA: **hausmann fe re, sigmamore**).

FIXED EFFECTS VS. RANDOM EFFECTS

- If the stochastic structure of the static random-effects model is correct, this model should be preferred.
- Mundlak (Ectra, 1978) proposed a model on pooling time-series and cross-section data.

$$Y_{it} = X_{it}\beta + \overline{X}_{2,i}\gamma + \omega_i + U_{it}$$

- X_{it} can included time-invariant variables and intercept. $\overline{X}_{2,i}$ is time average of time-variant regressors in X_{it} with $X_{it} = [X_{1,i} \quad X_{2,it}]$.
- where ω_i is a random effect which is uncorrelated with X_{it}. Therefore, this model should be estimated using (feasible) GLS.
- Mundlak showed that the random-effects estimator for β in this specification is identical to the within estimator.

MUNDLAK SPECIFICATION

- Given Mundlak's specification

$$Y_{it} = X_{it}\beta + \overline{X}_{2,i}\gamma + \omega_i + U_{it}$$

- the individual specific effect equals $\eta_i = \overline{X}_{2,i}\gamma + \omega_i$.
- No economic interpretation should be attached to γ.
- If $\gamma = 0$, then the individual specific effects η_i are uncorrelated with the regressors.
- This proposes an easy test for random effects $H_0 : \gamma = 0$ against fixed effects $H_1 : \gamma \neq 0$.
- After having estimated Mundlak's model a Wald-test can be applied (ie. compare to χ^2-distribution).

OTHER CONCERNS

- So far, we only considered balanced panels. There can be many reasons why a panel is not balanced.
- The source of attrition is important in deciding how to treat unbalanced panels.
- In case of a *rotating panel*, i.e. individuals only participate for a fixed number of time period and are then replaced by new individuals, the usual fixed-effects and random-effects estimators can be used (as long as individuals are observed at least twice). When using random-effects estimation the covariance matrix should be weighted correctly for the number of times each individual is observed.
- Having an unbalanced panel does not cause any problems to fixed-effects estimators as long as each individual is observed twice, and the reason for attrition is not related to U_{it}.
- If attrition is related to U_{it}, sample-selection models for panel data should be used.