Structure and Regression

Prf. José Fajardo
FGV/EBAPE

A Statistical Relationship

- A relationship of interest:
  - Number of hospital visits: H = 0,1,2,…
  - Covariates: x_1=Age, x_2=Sex, x_3=Income, x_4=Health
- Causality and covariation
  - Theoretical implications of 'causation'
  - Comovement and association
  - Intervention of omitted variables
  - Temporal relationship – movement of the "causal variable" precedes the effect.

Structure vs. Regression

- Reduced Form vs. Stuctural Model
- Simultaneous equations origin
  - $Q(d) = a_0 + a_1P + a_2I + e(d)$ (demand)
    $Q(s) = b_0 + b_1P + b_2C + e(s)$ (supply)
    $Q(.) = Q(d) = Q(s)$
    What is the effect of a change in I on Q(.)?
    (Not a regression)
  - Reduced form: $Q = c_0 + c_1I + c_2C + v$.
    (Regression)
- Modern concepts of structure vs. regression: The search for causal effects.

Models

- Conditional mean function: E[y | **x**] (Regression)
- Other conditional characteristics of interest:
  - Conditional variance function: Var[y | **x**]
  - Conditional quantiles, e.g., median [y | **x**]
  - Other conditional moments
- Conditional probabilities: P(y|**x**) (Discrete choice)
- **What is the sense in which "y varies with x?"**

Partial Effects

- What did the model tell us?
- Covariation and partial effects: How does the y "vary" with the x?
- Marginal Effects: Effect on what?????
  - For continuous variables
    $\delta(x)=\partial E[y|x]/\partial x$, usually not coefficients
  - For dummy variables
    $E[y|x,d=1] - E[y|x,d=0]$
- Elasticities: $\varepsilon(x)=\delta(x) * x / E[y|x]$

Average Partial Effects

- When $\delta(x) \neq \beta$, APE = $E_x[\delta(x)] = \int_x \delta(x)f(x)dx$
- Approximation: Is $\delta(E[x]) = E_x[\delta(x)]$? **(no)**
- Empirically: Estimated APE = $(1/N)\Sigma_{i=1}^{N}\hat{\delta}(x_i)$
- Empirical approximation: Est.APE = $\hat{\delta}(\bar{x})$

APE and PE at the Mean

$\delta(x)=\partial E[y|x]/\partial x$, $\mu=E[x]$

$\delta(x) \approx \delta(\mu)+\delta'(\mu)(x-\mu)+(1/2)\delta''(\mu)(x-\mu)^2+ \varepsilon$

$E[\delta(x)]=APE \approx \delta(\mu) + (1/2)\delta''(\mu)\sigma_x^2$

The Linear Model

- $\mathbf{y} = \mathbf{X\beta}+\boldsymbol{\varepsilon}$, N observations, K columns in $\mathbf{X}$, including a column of ones.
  - Standard assumptions about $\mathbf{X}$
  - Standard assumptions about $\boldsymbol{\varepsilon}|\mathbf{X}$
  - **$E[\boldsymbol{\varepsilon}|\mathbf{X}]=0$, $E[\boldsymbol{\varepsilon}]=0$ and $Cov[\boldsymbol{\varepsilon},\mathbf{x}]=0$**
- Regression?
  - If $E[\mathbf{y}|\mathbf{X}] = \mathbf{X\beta}$

Endogeneity

- Definition: $E[\boldsymbol{\varepsilon}|\mathbf{x}]\neq 0$
- Why not?
  - Omitted variables
  - Unobserved heterogeneity (equivalent to omitted variables)
  - Measurement error on the RHS (equivalent to omitted variables)
  - Endogenous sampling and attrition

Structure and Regression

- Simultaneity? What if $E[\varepsilon|\mathbf{x}]\neq 0$
- $y=x\beta+\varepsilon$, $x=\delta y+u$. $Cov[x, \varepsilon]\neq 0$
  - $x\beta$ is not the regression?
  - What is the regression?
    - Reduced form: Assume ε and u are uncorrelated.
    - $y = [\beta/(1- \beta\delta)]u + [1/(1- \beta\delta)]\varepsilon$
    - $x= [1/(1- \beta\delta)]u + [\delta /(1- \beta\delta)]\varepsilon$
    - $Cov[x,y]/Var[x] = \lambda$

$$= [\beta\sigma_u^2 + \delta\sigma_\varepsilon^2]/[\sigma_u^2 + \delta^2\sigma_\varepsilon^2]$$

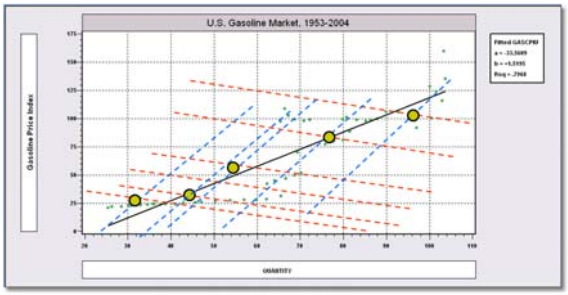
$$= w\beta + (1-w)(1/\delta) \text{ where } w=\sigma_u^2/[\sigma_u^2 + \delta^2\sigma_\varepsilon^2]$$

    - The regression is $y = \lambda x + v$, where $E[v|x]=0$

Structure vs. Regression

Supply = a + b*Price + c*Capacity

Demand = A + B*Price + C*Income



Implications

- The structure is the theory
- The regression is the conditional mean
- There is always a conditional mean
  - It may not equal the structure
  - It may be linear in the same variables
  - What is the implication for least squares estimation?
    - LS estimates regressions
    - LS does not necessarily estimate structures
    - Structures may not be estimable – they may not be identified.

Model Selection

- Regression models: Fit measure = R^2
- Nested models: log likelihood, GMM criterion function (distance function)
- Nonnested models, nonlinear models:
  - Classical
    - Akaike information criterion= – (logL – 2K)/N
    - Bayes (Schwartz) information criterion = –(logL-K(logN))/N
  - Bayesian: Bayes factor = Posterior odds/Prior odds
    (For noninformative priors, BF=ratio of posteriors)

Remaining to Consider for the Linear Regression Model

- Failures of standard assumptions
  - Heteroscedasticity
  - Autocorrelation and Spatial Correlation
  - Robust estimation
- Omitted variables
- Measurement error

How do panel data fit into this?

- We can use the usual models.
  - We can use far more elaborate models
  - We can study effects through time
- Observations are surely correlated.
  - The same individual is observed more than once
  - Unobserved heterogeneity that appears in the disturbance in a cross section remains persistent across observations (on the same 'unit').
  - Procedures must be adjusted.
- Dynamic effects are likely to be present.

Benefits of Panel Data

- Time and individual variation in behavior unobservable in cross sections or aggregate time series
- Observable and unobservable individual heterogeneity
- Rich hierarchical structures
- More complicated models
- Features that cannot be modeled with only cross section or aggregate time series data alone
- Dynamics in economic behavior

CR (1988) example cont'...

Cornwell and Rupert Data

Cornwell and Rupert Returns to Schooling Data, 595 Individuals, 7 Years
Variables in the file are

EXP	= work experience
WKS	= weeks worked
OCC	= occupation, 1 if blue collar,
IND	= 1 if manufacturing industry
SOUTH	= 1 if resides in south
SMSA	= 1 if resides in a city (SMSA)
MS	= 1 if married
FEM	= 1 if female
UNION	= 1 if wage set by union contract
ED	= years of education
LWAGE	= log of wage = dependent variable in regressions

These data were analyzed in Cornwell, C. and Rupert, P., "Efficient Estimation with Panel Data: An Empirical Comparison of Instrumental Variable Estimators," Journal of Applied Econometrics, 3, 1988, pp. 149-155. See Baltagi, page 122 for further analysis. The data were downloaded from the website for Baltagi's text.

Exercicio1.do

```
Insheet using "e:\paneldata\paneldata\ebape\cornwell&rupert.csv"
generate person=group(595)
bysort person: generate period=group(7)
* panel data definition
xtset person period
xtdes
xtsum
* plots of panel data
xtline lwage if person<=10, overlay
generate exp2=exp^2
local x1 exp exp2 wks occ ind south smsa ms union
local x2 ed blk fem
* panel data regression: y=lwage
* x1=[1 exp exp2 wks occ ind south smsa ms union],
* x2=[ed blk fem] (time-invariant regressors)
regress lwage `x1' `x2'
estimates store OLS
*autocorrelations?
```

Correlation.do

```
regress lwage `x1' `x2'
predict uhat, residuals
*analize autocorrelation of residuals
regress lwage `x1' `x2', vce(cluster person)
estimates store OLSvce
```

FE estimation: Exercicio2.do

```
local x1 exp exp2 wks occ ind south smsa ms union
local x2 ed blk fem
xtreg lwage `x1' `x2', fe
estimates store FE
*findit xtserial
xtserial lwage `x1' `x2'
xtreg lwage `x1' `x2', fe vce(cluster person)
estimates store FEvce
estimates table OLS OLSvce FE FEvce, b se t stats(r2_o rho)
```

Panel Data

Prf. José Fajardo
FGV/EBAPE

Panel Data –Basic Approach

- Counterfactual approach to causality (Rubin's model)

$$Y^T_{i,t_0} - Y^C_{i,t_0}$$ (T: Treatment, C: Control)

- With cross-sectional data (between estimation)

$$Y^T_{i,t_0} - Y^C_{j,t_0}$$

  - Assumption of unit homogeneity (no unobserved heterogeneity)

- With panel data I (within estimation)

$$Y^T_{i,t_1} - Y^C_{i,t_0}$$

  - Assumption of temporal homogeneity (no period effects, no maturation)

- With panel data II (within estimation with comparison group)

$$(Y^T_{i,t_1} - Y^C_{i,t_0}) - (Y^C_{j,t_1} - Y^C_{j,t_0})$$

  - Assumption of parallel trends

Panel data and Causal Inference

The two major problems in Social Research	Solution with experimental design	Solution with panel design
Self-selection (leading to unobserved heterogeneity)	Randomization	Within estimation (before-after comparison)
Reverse Causality (treatment depends on Y)	Controlled treatment	No simple solution (e.g. no time-varying unobserved heterog.)

- With panel data we can tackle one of the two major problems of Social Research

Fixed and Random Effects

- Unobserved individual effects in regression: $E[y_{it} \mid \mathbf{x}_{it}, c_i]$
  Notation:
  $$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + c_i + \varepsilon_{it}$$
  $$\mathbf{X}_i = \begin{bmatrix} \mathbf{x}_{i1}' \\ \mathbf{x}_{i2}' \\ \vdots \\ \mathbf{x}_{iT_i}' \end{bmatrix}$$ T_i rows, K columns
- Linear specification:
  Fixed Effects: $E[c_i \mid \mathbf{X}_i] = g(\mathbf{X}_i)$. $Cov[\mathbf{x}_{it}, c_i] \neq \mathbf{0}$
  effects are correlated with included variables.
  - **Random Effects:** $E[c_i \mid \mathbf{X}_i] = \mu$; effects are uncorrelated with included variables. If $\mathbf{X}_i$ contains a constant term, $\mu=0$ WLOG. Common: $Cov[\mathbf{x}_{it}, c_i] = \mathbf{0}$, but $E[c_i \mid \mathbf{X}_i] = \mu$ is needed for the full model

Convenient Notation

- Fixed Effects – the 'dummy variable model'
  $$y_{it} = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}$$
  Individual specific constant terms.
  $$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \Sigma_{j=1}^{N} \alpha_j d_{ijt} + \varepsilon_{it},\ d_{ijt} = \mathbf{1}(i=j)$$
- Random Effects – the 'error components model'
  $$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it} + u_i$$
  Compound ("composed") disturbance

Exogeneity

- Contemporaneous exogeneity
  - $E[\varepsilon_{it} \mid \mathbf{x}_{it}, c_i] = 0$ → Not sufficient for regression
  - Doesn't imply how to estimate $\boldsymbol{\beta}$
- Strict exogeneity – the most common assumption
  - $E[\varepsilon_{it} \mid \mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT}, c_i] = 0$
  - Can use first difference or fixed effects
  - Cannot hold if $\mathbf{x}_{it}$ contains lagged values of y_{it}
- Sequential exogeneity?
  - $E[\varepsilon_{it} \mid \mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{it}, c_i] = 0$

Assumptions for Asymptotics

- Convergence of moments involving cross section $\mathbf{X}_i$.
- N increasing, T or T_i assumed fixed.
  - "Fixed T asymptotics" (see Greene, p. 175)
  - Time series characteristics are not relevant (may be nonstationary)
  - If T is also growing, need to treat as multivariate time series.
- Ranks of matrices. $\mathbf{X}$ must have full column rank. ($\mathbf{X}_i$ may not, if $T_i < K$.)
- Strict exogeneity and dynamics. If $\mathbf{x}_{it}$ contains $y_{i,t-1}$ then $\mathbf{x}_{it}$ cannot be strictly exogenous. $\mathbf{X}_{it}$ will be correlated with the unobservables in period t-1.

Estimating β

- $\boldsymbol{\beta}$ is the partial effect of interest
- Can it be estimated (consistently) in the presence of (unmeasured) c_i?
  - Does pooled least squares "work?"
  - Strategies for "controlling for c_i" using the sample data
  - Using a proxy variable.

The Regression

- Presence of omitted effects

  $y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + c_i + \varepsilon_{it}$, observation for person i at time t

  $\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + c_i\mathbf{i} + \boldsymbol{\varepsilon}_i$, T_i observations in group i

  $= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{c}_i + \boldsymbol{\varepsilon}_i$, note $\mathbf{c}_i = (c_i, c_i, \ldots, c_i)'$

  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{c} + \boldsymbol{\varepsilon}$, $\sum_{i=1}^{N} T_i$ observations in the sample

- Potential bias/inconsistency of OLS – depends on 'fixed' or 'random'

OLS with Individual Effects

$\mathbf{b}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta}+\mathbf{c}+\boldsymbol{\varepsilon})$

$=\boldsymbol{\beta} + \left[(1/N)\Sigma_{i=1}^{N}\mathbf{X}_i'\mathbf{X}_i\right]^{-1}\left[(1/N)\Sigma_{i=1}^{N}\mathbf{X}_i'\mathbf{c}_i\right]$ (part due to the omitted c_j)

$+\left[(1/N)\Sigma_{i=1}^{N}\mathbf{X}_i'\mathbf{X}_i\right]^{-1}\left[(1/N)\Sigma_{i=1}^{N}\mathbf{X}_i'\boldsymbol{\varepsilon}_i\right]$ (covariance of $\mathbf{X}$ and $\boldsymbol{\varepsilon}$ will = 0)

The third term vanishes asymptotically by assumption

$\text{plim}\ \mathbf{b} = \boldsymbol{\beta} + \text{plim}\left[\frac{1}{N}\Sigma_{i=1}^{N}\mathbf{X}_i'\mathbf{X}_i\right]^{-1}\left[\Sigma_{i=1}^{N}\frac{T_i}{N}\bar{\mathbf{x}}_i c_j\right]$ (left out variable formula)

So, what becomes of $\left[\Sigma_{i=1}^{N} w_i \bar{\mathbf{x}}_i c_j\right]$?

plim $\mathbf{b}$ = $\boldsymbol{\beta}$ if the covariance of $\bar{\mathbf{x}}_i$ and c_j converges to zero.

Proxy Variables

- Proxies for unobserved effects: e.g., Test score for unobserved ability
- Interest is in $\delta(\mathbf{x}_{it},c_i)=\partial E[y_{it}|\mathbf{x}_{it},c_i]/\partial \mathbf{x}_{it}$
- Since c_i is unobserved, we seek APE = $E_c[\delta(\mathbf{x}_{it},c_i)]$
- Proxy has two characteristics
  - Ignorable in the model: $E[y_{it}|\mathbf{x}_{it},z_i,c_i] = E[y_{it}|\mathbf{x}_{it},c_i]$
  - 'Explains' c_i in that $E[c_i|z_i,\mathbf{x}_{it}] = E[c_i|z_i]$. In the presence of z_i, $\mathbf{x}_{it}$ does not further 'explain c_i.'
- Then, $E_c[\delta(\mathbf{x}_{it},c_i)] = E_z\{\partial E[y_{it}|\mathbf{x}_{it},z_i]/\partial \mathbf{x}_{it}\}$
  - Proof: See Wooldridge, pp. 23-24.
  - Loose ends:
    - Where do you get the proxy?
    - What is $E[y_{it}|\mathbf{x}_{it},z_i]$? Use the linear projection and hope for the best.

Estimating the Sampling Variance of *b*

- $s^2(\mathbf{X}'\mathbf{X})^{-1}$?
  - Correlation across observations
  - Heteroscedasticity
- A "robust" covariance matrix
  - Robust estimation (in general)
  - The White estimator
  - A Robust estimator for OLS.

A 'Cluster' Estimator

$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + (c_i + \varepsilon_{it})$
$= \mathbf{x}_{it}'\boldsymbol{\beta} + v_{it}$, $\text{Cov}[v_{it}, v_{is}] \neq 0$

Pseudo-log likelihood that produces OLS as the estimator

$\log L^* = \sum_{i=1}^{N}\left[(-1/2)\sum_{t=1}^{T_i}(\log\sigma^2 + \log 2\pi + v_{it}^2/\sigma^2\right]$

The solution for σ^2 will always be $[\sum_{i=1}^{N}\sum_{t=1}^{T_i}\hat{v}_{it}^2] / \sum_{i=1}^{N} T_i$,

so concentrate on $\boldsymbol{\beta}$. The solution will be $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$

$\partial \log L^*/\partial\boldsymbol{\beta} = \sum_{i=1}^{N}\left[\sum_{t=1}^{T_i}\mathbf{x}_{it}v_{it}/\sigma^2\right] = \sum_{i=1}^{N}\mathbf{g}_i = \mathbf{g}$.

$\partial^2 \log L^*/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}' = -\sum_{i=1}^{N}\sum_{t=1}^{T_i}\mathbf{x}_{it}\mathbf{x}_{it}'/\sigma^2 = -(1/\sigma^2)\mathbf{X}'\mathbf{X} = \mathbf{H}$ and $= \mathbf{E}[\mathbf{H}]$

$\text{Var}[\mathbf{b}] = (-\mathbf{H}^{-1})\text{Var}[\mathbf{g}](-\mathbf{H}^{-1})$

Var[**g**] is usually **H**, but not here because of correlation across observations. Approximate Var[**g**] with $\sum_{i=1}^{N}\mathbf{g}_i\mathbf{g}_i'$.

Cluster Estimator (cont.)

Robust variance estimator for Var[**b**]

Est.Var[**b**]

$$= \left[(\mathbf{X}'\mathbf{X})^{-1}\left[\sum_{i=1}^{N}\left(\sum_{t=1}^{T_i}\mathbf{x}_{it}\hat{v}_{it}\right)\left(\sum_{t=1}^{T_i}\mathbf{x}_{it}'\hat{v}_{it}\right)\right](\mathbf{X}'\mathbf{X})^{-1}\right]$$

$$= (\mathbf{X}'\mathbf{X})^{-1}\left[\sum_{i=1}^{N}\left(\sum_{t=1}^{T_i}\sum_{s=1}^{T_i}\hat{v}_{it}\hat{v}_{is}\mathbf{x}_{it}\mathbf{x}_{is}'\right)\right](\mathbf{X}'\mathbf{X})^{-1}$$

$\hat{v}_{it}$ = a least squares residual = $\widehat{\varepsilon_{it} + c_i}$

(If $T_i = 1$, this is the White estimator.)

Using First Differences

$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + c_i + \varepsilon_{it}$, observation for person i at time t

Eliminating the heterogeneity

$\Delta y_{it} = y_{it} - y_{i,t-1} = (\Delta\mathbf{x}_{it}')\boldsymbol{\beta} + \Delta c_i + \Delta\varepsilon_{it}$
$= (\Delta\mathbf{x}_{it}')\boldsymbol{\beta} + u_{it}$

Note: Time invariant variables become zero
Time trend becomes the constant term
Time dummy variables become (0,...,1,-1,0,0...)

OLS with First Differences

With strict exogeneity of $(\mathbf{X}_i, c_i)$, OLS regression of Δy_{it} on $\Delta \mathbf{x}_{it}$ is unbiased and consistent but inefficient.

$$\text{Var}\begin{pmatrix} \varepsilon_{i,2} - \varepsilon_{i,1} \\ \varepsilon_{i,3} - \varepsilon_{i,2} \\ \vdots \\ \varepsilon_{i,T_i} - \varepsilon_{i,T_i-1} \end{pmatrix} = \begin{bmatrix} 2\sigma_\varepsilon^2 & -\sigma_\varepsilon^2 & 0 & 0 \\ -\sigma_\varepsilon^2 & 2\sigma_\varepsilon^2 & -\sigma_\varepsilon^2 & \vdots \\ 0 & -\sigma_\varepsilon^2 & \ddots & -\sigma_\varepsilon^2 \\ 0 & \cdots & -\sigma_\varepsilon^2 & 2\sigma_\varepsilon^2 \end{bmatrix} \text{ (Toeplitz form)}$$

GLS is unpleasantly complicated. In order to compute a first step estimator of σ_ε^2 we would use fixed effects. We should just stop there. Or, use OLS in first differences and use Newey-West with one lag.

Two Periods

With two periods and strict exogeneity,

$\Delta y_{it} = y_{i2} - y_{i,1} = \delta_0 + (\mathbf{x}'_{i2} - \mathbf{x}'_{i1})\boldsymbol{\beta} + u_i$

Consider a "treatment, D_i," that takes place between time 1 and time 2 for some of the individuals

$\Delta y_i = \delta_0 + (\Delta \mathbf{x}_i)'\boldsymbol{\beta} + \delta_1 D_i + u_i$

D_i = the "treatment dummy"

$\hat{\delta}_1 = \overline{\Delta y} \mid \text{treatment} - \overline{\Delta y} \mid \text{control}$

= "difference in differences" estimator.

$\hat{\delta}_0 =$ Average change in y_i for the "treated"

This is a classical regression model. If there are no regressors,

D-in-D Model

With two periods and strict exogeneity,

$y_{it} = \beta_0 + \beta_1 D_{i2} + \beta_2 T_t + \beta_3 T_t D_{it} + \varepsilon_{it}$

D_{i2} = dummy variable for a treatment that takes place between time 1 and time 2 for some of the individuals,

T_t = a time period dummy variable, 0 in period 1, 1 in period 2.

Using least squares,

$$b_3 = (\bar{y}_2 - \bar{y}_1)_{D=1} - (\bar{y}_2 - \bar{y}_1)_{D=0}$$

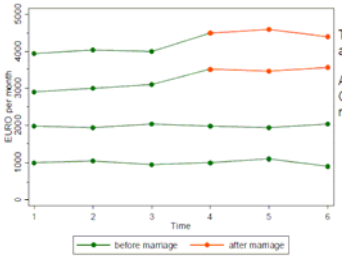
This is a classical regression model.

Example: Marriage Premium

- Fabricated data: long-format

```
. list id time wage marr, separator(6)

     +-------------------------+     +-------------------------+
     | id   time   wage   marr |     | id   time   wage   marr |
     |-------------------------|     |-------------------------|
  1. |  1      1   1000      0 |  13. |  3      1   2900      0 |
  2. |  1      2   1050      0 |  14. |  3      2   3000      0 |
  3. |  1      3    950      0 |  15. |  3      3   3100      0 |
  4. |  1      4   1000      0 |  16. |  3      4   3500      1 |
  5. |  1      5   1100      0 |  17. |  3      5   3450      1 |
  6. |  1      6    900      0 |  18. |  3      6   3550      1 |
     |-------------------------|     |-------------------------|
  7. |  2      1   2000      0 |  19. |  4      1   3950      0 |
  8. |  2      2   1950      0 |  20. |  4      2   4050      0 |
  9. |  2      3   2050      0 |  21. |  4      3   4000      0 |
 10. |  2      4   2000      0 |  22. |  4      4   4500      1 |
 11. |  2      5   1950      0 |  23. |  4      5   4600      1 |
 12. |  2      6   2050      0 |  24. |  4      6   4400      1 |
     +-------------------------+     +-------------------------+
```

Example: Marriage Premium I



There is a causal effect: a marriage-premium

And there is selectivity: Only high wage men marry

Computing Marriage Premium

- These data are like experimental data
  - Treatment and control group
  - Before (t = 1, 2, 3) and after (t = 4, 5, 6) measurements
  - However, treatment assignment is not under control of the researcher (no randomization)
    - This is the reason for our problem with self-selection
- A within approach
  - However, the fact that we have before-after measurements, allows a within approach. This compensates for the missing randomization.
    - Thus, we can compute the causal effect despite self-selection
  - For each man we compute a individual causal effect, by subtracting the before-mean from the after-mean
  - To get the average treatment effect (ATE) we average over the married (treatment) and the unmarried (control) men

Computing Marriage Premium

This is a kind of Difference-in-Differences (DID) estimator that uses all before and after information

- To get the individual causal effects (Δ_i) we make within-person comparisons (the after-before difference)
  - $\Delta_1 = 4500 - 4000 = 500$
  - $\Delta_2 = 3500 - 3000 = 500$
  - $\Delta_3 = 2000 - 2000 = 0$
  - $\Delta_4 = 1000 - 1000 = 0$

$$\Delta_i = \frac{1}{3}\sum_{t=4}^{6} y_{it} - \frac{1}{3}\sum_{t=1}^{3} y_{it}$$

- To get the average treatment effect (ATE) we average over the married (treatment) and the unmarried (control) men

$$ATE = \frac{500 + 500}{2} - \frac{0 + 0}{2} = 500$$

- The marriage-premium in our data is **+500 €**

Problems

- Result of a cross-sectional regression at T=4:

$$y_{i4} = \beta_0 + \beta_1 x_{i4} + u_{i4}$$

  - Between-comparison: compare wages of married and unmarried men

$$\hat{\beta}_1 = \frac{4500 + 3500}{2} - \frac{2000 + 1000}{2} = 2500$$

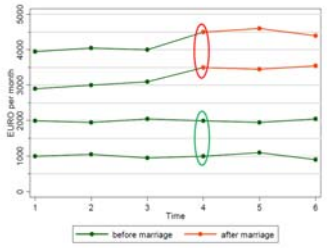
  - We get a very large marital wage premium!
- Obviously, a cross-sectional regression is highly misleading!
  - The bias is due to unobserved heterogeneity
    - High-wage men self-select into marriage
  - Technically: endogeneity (x_{i4} and u_{i4} are correlated)
- Self-selection is ubiquitous in non-experimental research

→ Many (most?) cross-sectional regression results are biased!

→ Be always critical when using cross-sectional data: Might unobserved heterogeneity distort my results?

Cross-sectional Regression

- This is the information used by a cross-sectional regression



No-solution: Pooled OLS

- Pool the data and estimate an OLS regression (POLS)

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

- The result is $\hat{\beta}_1 = 1833$
  - This is the mean of the red points minus the mean of the green points
  - The bias is still heavy
  - POLS also relies on a between comparison. It is thus biased due to unobserved heterogeneity: x_{it} and u_{it} are correlated
- Panel data per se do not remedy the problem of unobserved heterogeneity!
  - One has to use appropriate methods of analysis

Panel Data and Within Estimation

(2) Between- and within-group variation

Define the individual-specific or group mean for any variable, *e.g.* y_{it} as:

$$\bar{y}_i = \frac{1}{T_i}\sum_{t=1}^{T_i} y_{it}$$

y_{it} can be decomposed into 2 components:

$$y_{it} - \bar{y} = (y_{it} - \bar{y}_i) + (\bar{y}_i - \bar{y})$$
$$= \text{within} \quad + \quad \text{between}$$

where $\bar{y} = \sum_{i=1}^{n}\sum_{t=1}^{T_i} y_{it} \Big/ n\bar{T}$ and $\bar{T}$ is average no. of periods per case

Corresponding decomposition of sum of squares:

$$\sum_{i=1}^{n}\sum_{t=1}^{T_i}(y_{it} - \bar{y})^2 = \sum_{i=1}^{n}\sum_{t=1}^{T_i}(y_{it} - \bar{y}_i)^2 + \sum_{i=1}^{n}\sum_{t=1}^{T_i}(\bar{y}_i - \bar{y})^2$$

or: $T_{yy} = W_{yy} + B_{yy}$

Beatles example

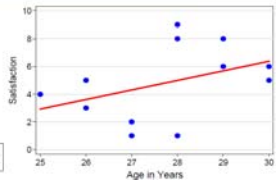
- Life satisfaction of four Englishmen

	1968	1969	1970
John	8	6	5
Paul	5	2	1
George	4	3	1
Ringo	9	8	6

- How depends satisfaction on age?
  - It seems, as if the four grew happier
  - $\hat{\beta} = 0.69$

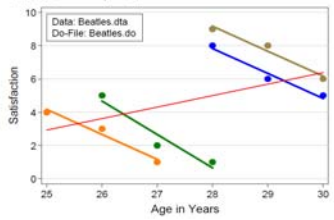
Data: Beatles.dta
Do-File: Beatles.do

Source: Kohler/Kreuter (2009) Data Analysis Using Stata. Stata Press. Pp. 242 ff.



Within estimation

- Between estimation is completely misleading
  - Because there is much heterogeneity between the men
- Within estimation shows a negative age effect
  - Separate regression line for each man
  - All four men grew less happier
    - Green: $\hat{\beta} = -2.0$
    - The rest: $\hat{\beta} = -1.5$



Within estimation

- Within estimators implement a "before-after comparison"
  - Difference-in-differences, first-difference, fixed-effects
- Starting point: error decomposition $u_{it} = \nu_i + \varepsilon_{it}$
- This yields the error component model

$$y_{it} = \beta_1 x_{it} + \nu_i + \varepsilon_{it}$$

  - ν_i: person-specific time-constant unobserved heterogeneity
    - Assumption: person-specific constants (fixed-effects)
    - Therefore the constant β_0 is dropped
  - ε_{it}: idiosyncratic error term
    - Assumption: zero mean, homoscedasticity, no autocorrelation
- POLS is consistent only, if x_{it} is uncorrelated with **both** error-components
  - $Cov(x_{it}, \nu_i) = 0$: no person-specific unobserved heterogeneity
  - $Cov(x_{it}, \varepsilon_{it}) = 0$: contemporaneous exogeneity assumption

First-Difference Estimator

- How can we get rid of the fixed-effects?
- With panel data we can „difference them out"

$$y_{it-1} = \beta_1 x_{it-1} + v_i + \varepsilon_{it-1}$$
$$y_{it} = \beta_1 x_{it} + v_i + \varepsilon_{it}$$

Subtracting the first equation from the second gives:

$$\Delta y_i = \beta_1 \Delta x_i + \Delta \varepsilon_i$$

where "Δ" denotes the change from $t-1$ to t.

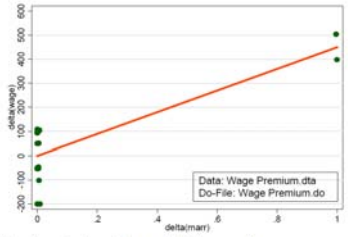
- Fixed-effects have been "differenced out"
- Pooled OLS applied to these transformed data provides the first-difference (FD) estimator. The FD-estimator is consistent if $Cov(x_{it}, \varepsilon_{is}) = 0 \quad \text{for } t \leq s$ (sequential exogeneity assumption)
  - Intuition: otherwise Δx_i and $\Delta \varepsilon_i$ would be correlated
- However, $Cov(x_{it}, v_i) = 0$ is no longer required
  - FD identifies the causal effect under weaker assumptions
  - ➢ Time-constant unobserved heterogeneity is no longer a problem

Example: FD Estimation

```
* We generate the first-differenced variables by hand
generate  dwage = wage - L.wage              // L. is the lag-operator
generate  dmarr = marr - L.marr

* We run an OLS regression (without constant)
regress   dwage dmarr, noconstant

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  1,    19) =   39.97
       Model |      405000     1      405000           Prob > F      =  0.0000
    Residual |      192500    19  10131.5789           R-squared     =  0.6778
-------------+------------------------------           Adj R-squared =  0.6609
       Total |      597500    20       29875           Root MSE      =  100.66

------------------------------------------------------------------------------
       dwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       dmarr |         450   71.17436     6.32   0.000     301.0304    598.9696
------------------------------------------------------------------------------
```

- The FD-estimator is 450, which is very close to the true causal effect

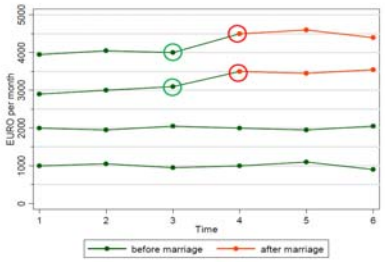
``Mechanics´´ of FD regression



- The slope is based on only 2 observations (the immediate wage increase after marriage)
- With T>2 FD-estimation is inefficient

FD Regression

- This is the information used by a FD-regression



Fixed Effect Estimator

- Fixed-effects estimation
  - "Time-demeaning" the data (within transformation)

$$y_{it} = \beta_1 x_{it} + v_i + \varepsilon_{it} \quad (1)$$

Average over t for each i $\quad \bar{y}_i = \beta_1 \bar{x}_i + v_i + \bar{\varepsilon}_i \quad (2)$

Subtract (2) from (1) $\quad y_{it} - \bar{y}_i = \beta_1 (x_{it} - \bar{x}_i) + \varepsilon_{it} - \bar{\varepsilon}_i \quad (3)$

  - Fixed-effects have been "differenced out". Only within variation is left
  - Pooled OLS applied to time-demeaned data provides the fixed-effects (FE) estimator. The FE-estimator is consistent if

    $Cov(x_{it}, \varepsilon_{is}) = 0$ for all t and s (strict exogeneity assumption)
    - Intuition: otherwise $x_{it} - \bar{x}_i$ and $\varepsilon_{it} - \bar{\varepsilon}_i$ would be correlated
  - However, $Cov(x_{it}, v_i) = 0$ is no longer required
    - FE identifies the causal effect under weaker assumptions
    - Time-constant unobserved heterogeneity is no longer a problem

Example: Fixed Effect Estimator

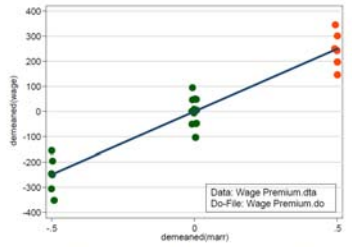
Data: Wage Premium.dta
Do-File: Wage Premium.do

```
. xtset id time
       panel variable:  id (strongly balanced)
        time variable:  time, 1 to 6
                delta:  1 unit

. xtreg wage marr, fe

Fixed-effects (within) regression               Number of obs      =        24
Group variable: id                              Number of groups   =         4

R-sq:  within  = 0.8982                         Obs per group: min =         6
       between = 0.8351                                        avg =       6.0
       overall = 0.4065                                        max =         6

                                                F(1,19)            =    167.65
corr(u_i, Xb)  = 0.5164                         Prob > F           =    0.0000

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        marr |         500   38.61642    12.95   0.000     419.1749    580.8251
       _cons |        2500    16.7214   149.51   0.000     2465.002    2534.998
-------------+----------------------------------------------------------------
     sigma_u |  1290.9944
     sigma_e |  66.885605
         rho |  .99732298   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Interpreting FE Output

- The FE model succeeds in identifying the true causal effect!
  - Marriage increases the wage by 500 €
  - The effect is significant (judged by the t-value or the p-value)
  - A constant is reported, since Stata adds back the overall wage mean, which is 2500 here
  - Model fit can be judged by the within R^2 as usual (referring to (3))
    - 90% of the within wage variation is explained by marital status change
    - The between and overall R^2 refer to different models and are not useful here
  - Variance of the error-components
    - `sigma_u` is the estimated standard deviation of $\hat{v}_i$
    - `sigma_e` is the estimated standard deviation of $\hat{\ell}_{it}$
- Interpretation of within estimation results
  - Not: "X causes Y"
  - But: "A change in X causes a change in Y"

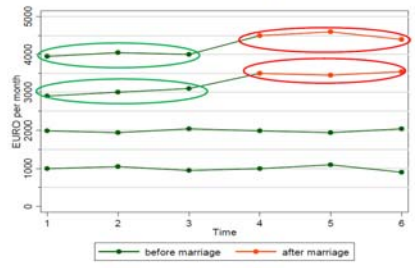
``Mechanics´´ of FE regression



- Those, never marrying are at X=0. They contribute nothing to the regression.
- The slope is only determined by the wages of those marrying:
  It is the difference in the mean wage before and after marriage.

FE Regression

This is the information used by a FE-regression



Fixed Effect Estimator

- $\tilde{\beta}$ will be unbiased and consistent as $N \to \infty$ and for T fixed. It will also be consistent if $T \to \infty$ regardless of N.
- Define $\hat{\mu}_i = T^{-1} \sum_{t=1}^{T} (y_{it} - X_{it}' \hat{\beta}_{WG}) = \bar{y}_i - \bar{X}_i' \hat{\beta}_{WG}$ for $i = 1, 2, ..., N$ as the fixed-effects estimates.
- Note that $\hat{\mu}_i$ will be unbiased but will only be consistent for $T \to \infty$
- In general the FE estimator is more efficient than the FD estimator when $v_{it} \sim IID(0, \sigma_v^2)$.
- The FD estimator is more efficient thatn the FE estimator when the v remainder component is a random walk.

Equivalent FE Estimator I: LSDV

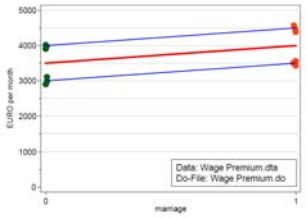
```
tabulate id, gen(pers)
```

```
. regress wage marr pers1-pers4, noconstant

      Source |       SS       df       MS              Number of obs =      24
-------------+------------------------------           F(  5,    19) = 9052.94
       Model |  202500000        5    40500000          Prob > F      =  0.0000
    Residual |      85000       19  4473.68421          R-squared     =  0.9996
-------------+------------------------------           Adj R-squared =  0.9995
       Total |  202585000       24  8441041.67          Root MSE      =  66.886

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        marr |         500   38.61642    12.95   0.000     419.1749    580.8251
       pers1 |        1000   27.30593    36.62   0.000      942.848    1057.152
       pers2 |        2000   27.30593    73.24   0.000     1942.848    2057.152
       pers3 |        3000    33.4428    89.71   0.000     2930.003    3069.997
       pers4 |        4000    33.4428   119.61   0.000     3930.003    4069.997
------------------------------------------------------------------------------
```

- Least-squares-dummy-variables-estimator (LSDV)
- Practical only when N is small
- We get estimates for the v_i

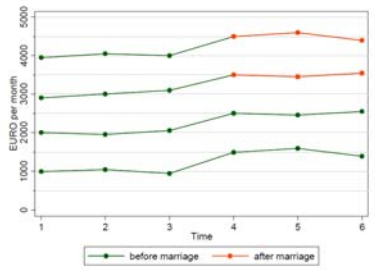
Data: Wage Premium.dta
Do-File: Wage Premium.do

Equivalent FE Estimator II: Individual Slope Reg.



- Estimate a regression for every man marrying (blue)
- The FE-estimator is the (weighted) mean of the individual slopes (red)

Problem: No Control Group



Modified Data:

"wage3" is a new wage variable

There is a period effect: general wage increase at t=4.

Now, there is no causal effect!

But there is still selectivity.

Problem: No Control Group

```
. xtreg wage3 marr, fe

Fixed-effects (within) regression               Number of obs      =        24
Group variable: id                              Number of groups   =         4

R-sq:  within  = 0.4732                         Obs per group: min =         6
       between = 0.8000                                        avg =       6.0
       overall = 0.3958                                        max =         6

                                                F(1,19)            =     17.07
                                                Prob > F           =    0.0006

------------------------------------------------------------------------------
      wage3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------+----------------------------------------------------------------
       marr |         500   121.0336     4.13   0.001     246.6738    753.3262
      _cons |        2625   52.40907    50.09   0.000     2515.307    2734.693
------------------------------------------------------------------------------
```

Data: Wage Premium.dta
Do-File: Wage Premium.do

- FE-regression yields the wrong answer
  - Reason is that FE does not use the control group information
  - This is generally true: groups where X does not change contribute nothing to the FE-estimator
    - Note that Stata reports the N in the data, not the N used for FE-estimation!

Solution: Two-way FE Regression

- Including time fixed-effects (period dummies)

$$y_{it} = \beta_1 x_{it} + \mu_t + v_i + \varepsilon_{it}$$

  - Now also the control group information is used

```
. tab time, gen(t)
. xtreg wage3 marr t2-t6, fe

------------------------------------------------------------------------------
      wage3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------+----------------------------------------------------------------
       marr |   -4.59e-13   58.24824    -0.00   1.000     -124.93      124.93
         t2 |          50   50.44445     0.99   0.338    -58.19259    158.1926
         t3 |        62.5   50.44445     1.24   0.236    -45.69259    170.6926
         t4 |       537.5   58.24824     9.23   0.000      412.57      662.43
         t5 |       562.5   58.24824     9.66   0.000      437.57      687.43
         t6 |       512.5   58.24824     8.80   0.000      387.57      637.43
      _cons |      2462.5   35.66961    69.04   0.000     2385.996    2539.004
------------------------------------------------------------------------------
```

One should always include time in a FE-regression! (period or age effects)

Data: Wage Premium.dta
Do-File: Wage Premium.do

Two basic identification problems

(1) Unobservable variables
- Can we identify the impact of unobservables?
- Can we distinguish the impact of unobservables from the impact of time-invariant observables?

(2) Age, cohort and time effects – can they be distinguished?
- Behaviour may change with age
- Current behaviour may be affected by experience in "formative years" ⇒ cohort or year-of-birth effect
- Time may affect behaviour through changing social environment

A second identification problem: Age, cohort & time effects

Fundamental identity relating age (A_{it}), time of interview (t) and birth cohort (B_i):

$$A_{it} \equiv t - B_i$$

These three cannot be distinguished in principle. To do so would require an ability to move a cohort forward or back in time (!) to measure the effect of time holding age and cohort constant.

- In a cross-section, t doesn't vary, so time effects can't be estimated and age or cohort are collinear – only their joint effect can be estimated
- In a panel, t varies but A_{it}, t and B_i are collinear - only two of the three effects can be estimated.
- So we can use (t ,B_i) , (A_{it} ,B_i) or (A_{it} ,t) as covariates, but not all three.

Age, cohort and time effects

A possible solution is to think more deeply about the effects of time and cohort and introduce further information.

E.g. we may think it is the social environment at the time of birth that generates differences between cohorts and the present social environment that generates time effects.

Let $\mathbf{w}(t)$ be variables describing the social environment at historical time t (*e.g.* unemployment rate, income inequality, crime rate).

Then our model would use A_{it} , $\mathbf{w}(t)$ and $\mathbf{w}(B_i)$) as covariates

This breaks the exact relationship between age, time and cohort effects and permits identification.

Summary of FE Estimation

- Panel data and within estimation (FD-, FE-regression) can identify a causal effect under weaker assumptions
  - Time-constant person-specific unobserved heterogeneity no longer biases the estimates
- However, with FE-regressions we cannot estimate the effects of time-constant covariates. These are all cancelled out by the within transformation.
  - This reflects the fact that panel data do not help to identify the causal effect of a time-constant covariate!
- The "within logic" applies only with time-varying covariates
  - Something has to "happen"
    → Analyzing the effects of events
  - Only then a before-after comparison is possible

Summary of FE Estimation

- FE-estimator has to assume (strict exogeneity assumption)

$$Cov(x_{it}, \varepsilon_{is}) = 0 \quad \text{for all } t \text{ and } s$$

  - If this assumption is violated (endogeneity) FE-estimates are biased
- Sources of endogeneity
  - Time-varying unobserved heterogeneity (captured by ε_{it})
  - Y shocks trigger the change in X (reverse causality)
  - Errors in reporting X (measurement errors)
- Supposed remedy
  - IV-estimation (`xtivreg`)
  - Structural equation modeling (LISREL)
  - These methods rest on untestable assumptions. Therefore, it is unclear whether they improve estimates.
  - These methods have produced a big mess in social research.