

# Longitudinal/Panel Data Analysis: Lecture 1

Raymond Duch

University of Oxford  
Nuffield College  
raymond.duch@nuffield.ox.ac.uk  
[raymondduch.com/class/2011/paneldata](http://raymondduch.com/class/2011/paneldata)

May 3, 2011

- Week 1: Introduction to panel data
- Week 2: Longitudinal Panels, Fixed, Random Effects, Dynamic Models
- Week 3: Hierarchical nested data sets
- Week 4: Cross section time series

- Stata 10.0 Manual Longitudinal/Panel Data, xtabond, xtabond postestimation, xtdpdsys, xtivreg
- Halaby, Charles N. 2004. "Panel Models in Sociological Research: Theory into Practice." *Annual Review of Sociology*. 30:507-44.
- Hsiao, Cheng. 2003. "Analysis of Panel Data: Second Edition." *Political Analysis*. 15: chapters 1 and 3.

- Leveraging panel data to solve problems of causal inference
- Problems of unobservables in non-randomized studies
  - time-invariant unit-specific unobservables
  - time-varying unit-specific unobservables that represent transitory and idiosyncratic forces acting upon units

# Why Panel Data?

$$y_{it} = \alpha^* + \beta' x_{it} + \rho' z_{it} + u_{it},$$

$$i = 1, \dots, N,$$

$$t = 1, \dots, T,$$

$x_{it}$  and  $z_{it}$  are  $k_1 \times 1$  and  $k_2 \times 1$  vectors of exogenous variables

(1)

# Unobserved Unit Heterogeneity

- when error term  $u_{it}$  is independently, identically distributed over  $i$  and  $t$ , with mean zero and variance  $\sigma^2$  OLS give unbiased estimates of  $\alpha^*$ ,  $\beta x_{it}$ , and  $\rho z_{it}$
- Now suppose that  $z_{it}$  values are unobservable, and the covariances between  $x_{it}$  and  $z_{it}$  are nonzero?
- Then the least-squares regression coefficients of  $y_{it}$  on  $x_{it}$  are biased.
- Repeated observations of individuals allows us to get rid of this effect of  $z$

# Unobserved Unit Heterogeneity

- when error term  $u_{it}$  is independently, identically distributed over  $i$  and  $t$ , with mean zero and variance  $\sigma^2$  OLS give unbiased estimates of  $\alpha^*$ ,  $\beta x_{it}$ , and  $\rho z_{it}$
- Now suppose that  $z_{it}$  values are unobservable, and the covariances between  $x_{it}$  and  $z_{it}$  are nonzero?
- Then the least-squares regression coefficients of  $y_{it}$  on  $x_{it}$  are biased.
- Repeated observations of individuals allows us to get rid of this effect of  $z$

## First Differencing to Eliminate $z$

If  $z_{it} = z_i$  for all  $t$  (i.e.,  $z$  values stay constant through time for a given individual but vary across individuals):

$$y_{it} - y_{i,t-1} = \beta' (x_{it} - x_{i,t-1}) + (u_{it} - u_{i,t-1}),$$

$$i = 1, \dots, N,$$

$$t = 1, \dots, T, \quad (2)$$

## Mean Deviation to Eliminate $z$

If  $z_{it} = z_i$  for all  $t$  (i.e.,  $z$  values stay constant through time for a given individual but vary across individuals):

$$y_{it} - \bar{y}_i = \beta' (x_{it} - \bar{x}_i) + (u_{it} - \bar{u}_i),$$

$$i = 1, \dots, N,$$

$$t = 1, \dots, T,$$

This is the Fixed Effect estimation employed by the Stata `xtreg` command (3)

Ignoring such heterogeneity could lead to inconsistent or meaningless estimates of interesting parameters.

$$y_{it} = \alpha_i^* + \beta_i x_{it} + u_{it},$$

$$i = 1, \dots, N,$$

$$t = 1, \dots, T,$$

In this model the intercept and slope coefficients may differ across the  $i$  units. (4)

$$y_{it} = \alpha^* + \beta x_{it} + u_{it},$$

$$i = 1, \dots, N,$$

$$t = 1, \dots, T,$$

(5)

$$y_{it} = \alpha_{it}^* + \sum_{k=1}^K \beta_k x_{kit} + u_{it},$$

$$i = 1, \dots, N,$$

$$t = 1, \dots, T,$$

(6)

## All coefficients vary over individuals

$$y_{it} = \alpha_i^* + \sum_{k=1}^K \beta_{ki} x_{kit} + u_{it},$$

$$i = 1, \dots, N,$$

$$t = 1, \dots, T,$$

(7)

# All coefficients vary over time and individuals

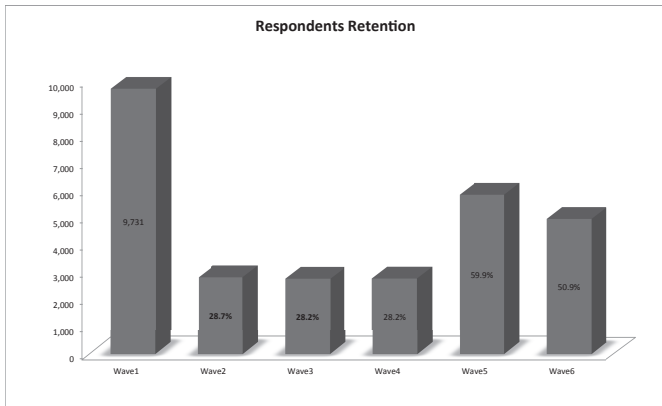
$$y_{it} = \alpha_{it}^* + \sum_{k=1}^K \beta_{kit} x_{kit} + u_{it},$$

$$i = 1, \dots, N,$$

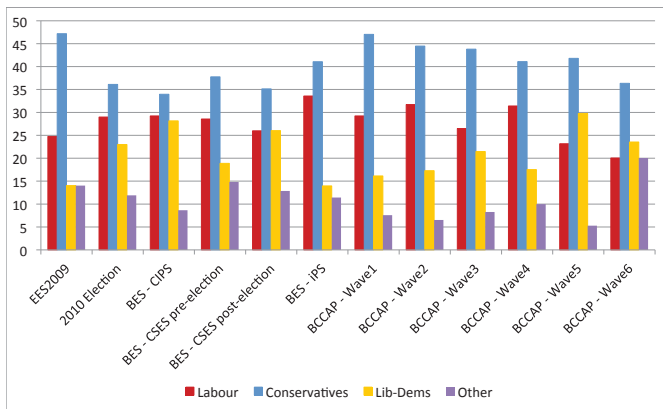
$$t = 1, \dots, T,$$

(8)

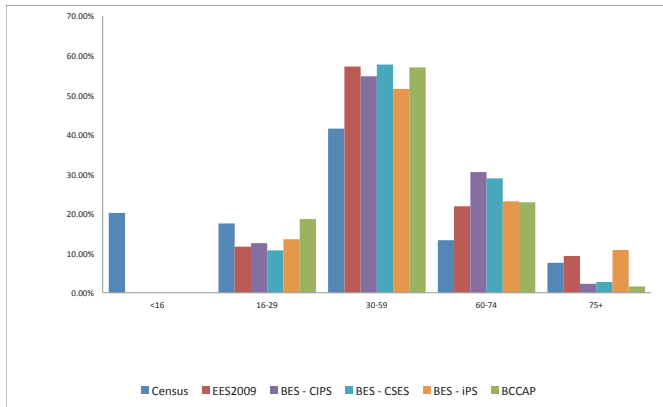
# Panel Study: British CCAP Study



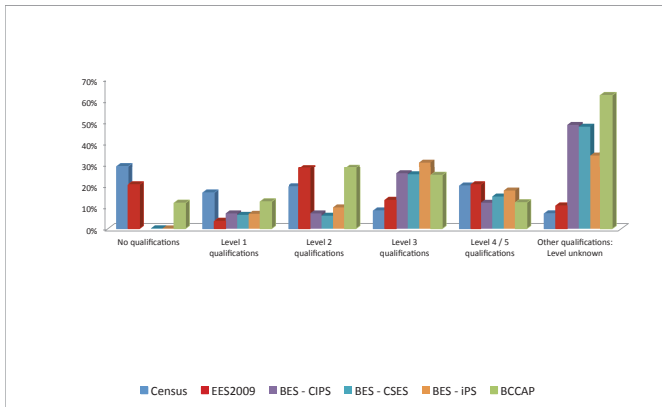
# Panel Study: British CCAP Study



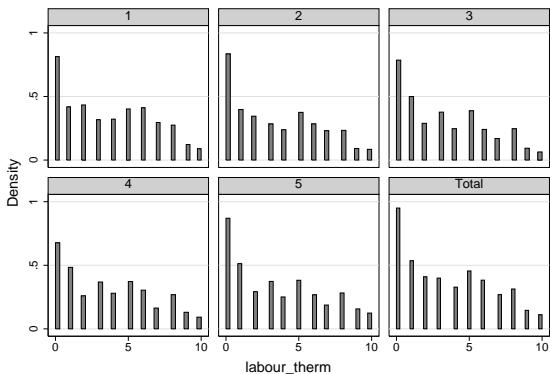
# Panel Study: British CCAP Study



# Panel Study: British CCAP Study

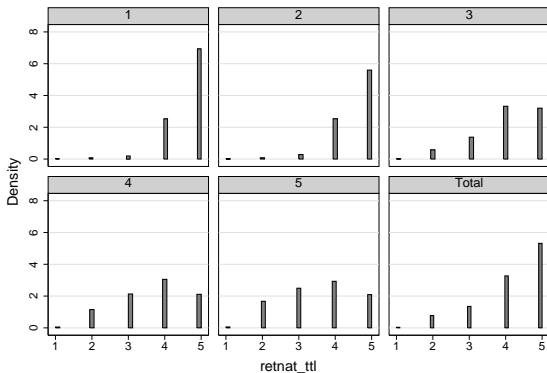


# Labour evaluations over five panel waves



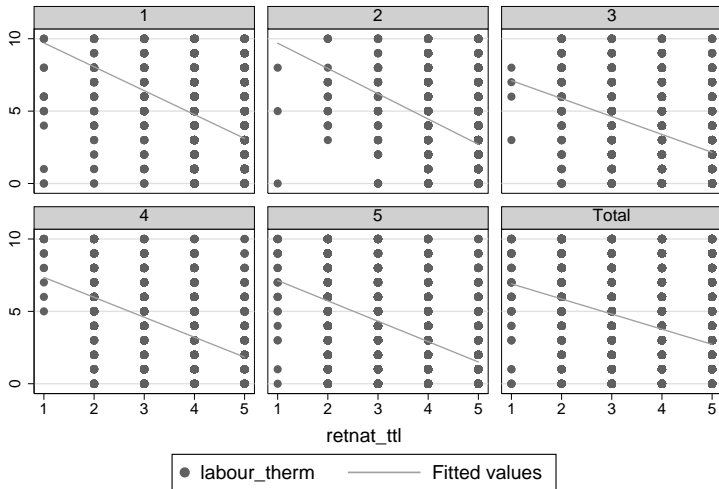
Graphs by wave\_id

# Economic evaluations over five panel waves



Graphs by wave\_id

# Scatter plot of labour evaluations against economy



Graphs by wave\_id

- If  $\theta$  is a vector of parameters, then  $\hat{\theta}_U$  is the ML estimator of  $\theta$  without restrictions, and  $\hat{\theta}_R$  is the estimator with the constraints (restrictions).
- We have an “unrestricted” log of likelihood,  $\hat{L}_U$ . That will be the maximum (by construction); the “restricted” log of likelihood,  $\hat{L}_R$ , by definition, will be smaller.
- If a restriction is valid, then the likelihood of the restriction won't cause a large reduction in the log of likelihood.

The likelihood ratio is:

$$\lambda = \frac{\hat{L}_R}{\hat{L}_U}$$

- Both of these likelihoods are positive
- since  $\hat{L}_U > \hat{L}_R$ , these two conditions mean that  $\lambda$  must be between zero and one.
- The large-sample distribution of  $-2\ln\lambda$  is chi-squared with  $df =$  number of restrictions imposed.

# Correlations

```
. corr income_1 income_2 income_3 income_4 income_5  
(obs=800)
```

```
      | income_1 income_2 income_3 income_4 income_5  
-----+-----  
income_1 | 1.0000  
income_2 | 0.9416 1.0000  
income_3 | 0.9142 0.9347 1.0000  
income_4 | 0.9031 0.9228 0.9570 1.0000  
income_5 | 0.8932 0.9085 0.9339 0.9488 1.0000
```

```
. corr retnat_1 retnat_2 retnat_3 retnat_4 retnat_5  
(obs=1105)
```

```
      | retnat_1 retnat_2 retnat_3 retnat_4 retnat_5  
-----+-----  
retnat_1 | 1.0000  
retnat_2 | 0.5293 1.0000  
retnat_3 | 0.3518 0.3867 1.0000  
retnat_4 | 0.3164 0.3217 0.5670 1.0000  
retnat_5 | 0.2949 0.3212 0.4843 0.5779 1.0000
```

# Pooled Regression

```
. regress labour_therm retnat_ttl income_ttl
```

Source	SS	df	MS
Model	17549.7573	2	8774.87865
Residual	147573.07	18650	7.91276515
Total	165122.827	18652	8.85282154

Number of obs	=	18653
F( 2, 18650)	=	1108.95
Prob > F	=	0.0000
R-squared	=	0.1063
Adj R-squared	=	0.1062
Root MSE	=	2.813

labour_therm	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
retnat_ttl	-1.015335	.0215739	-47.06	0.000	-1.057622 - .973048
income_ttl	.0043702	.0058558	0.75	0.455	-.0071077 .0158482
_cons	7.861831	.1025247	76.68	0.000	7.660873 8.062789

# LR Test of Unit Effect Specifications

```
. regr labour_therm retnat_ttl income_ttl wave_2 wave_3 wave_4 wave_5
```

Source	SS	df	MS	Number of obs =	18653
Model	25173.4644	6	4195.5774	F( 6, 18646) =	558.99
Residual	139949.363	18646	7.50559707	Prob > F =	0.0000
				R-squared =	0.1525
				Adj R-squared =	0.1522
Total	165122.827	18652	8.85282154	Root MSE =	2.7396

labour_therm	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
retnat_ttl	-1.420642	.0246644	-57.60	0.000	-1.468987	-1.372298
income_ttl	.0001437	.0057065	0.03	0.980	-.0110416	.011329
wave_2	-.3066489	.0665855	-4.61	0.000	-.4371627	-.1761352
wave_3	-1.202404	.0707193	-17.00	0.000	-1.34102	-1.063788
wave_4	-1.375404	.0711866	-19.32	0.000	-1.514936	-1.235871
wave_5	-1.723747	.0579514	-29.74	0.000	-1.837337	-1.610157
_cons	10.34627	.1271365	81.38	0.000	10.09707	10.59547

```
. estimates store intercept
```

# LR Test of Unit Effect Specifications

```
.   regr labour_therm retnat_ttl income_ttl
```

Source	SS	df	MS	Number of obs =	18653
Model	17549.7573	2	8774.87865	F( 2, 18650) =	1108.95
Residual	147573.07	18650	7.91276515	Prob > F	= 0.0000
-----+-----				R-squared	= 0.1063
Total	165122.827	18652	8.85282154	Adj R-squared	= 0.1062
-----+-----				Root MSE	= 2.813

labour_therm	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
retnat_ttl	-1.015335	.0215739	-47.06	0.000	-1.057622	-.973048
income_ttl	.0043702	.0058558	0.75	0.455	-.0071077	.0158482
_cons	7.861831	.1025247	76.68	0.000	7.660873	8.062789

```
.   lrtest intercept
```

Likelihood-ratio test	LR chi2(4) =	989.41
(Assumption: . nested in intercept)	Prob > chi2 =	0.0000

# LR Test of Unit Effect Specifications

```
.      regr labour_therm retnat_ttl income_ttl wave_2_int wave_3_int wave_4_int wave_5_int
note: wave_4_int omitted because of collinearity
```

Source	SS	df	MS	Number of obs =	18653
Model	22432.0709	5	4486.41418	F( 5, 18647) =	586.29
Residual	142690.756	18647	7.65220981	Prob > F =	0.0000
-----				R-squared =	0.1359
-----				Adj R-squared =	0.1356
Total	165122.827	18652	8.85282154	Root MSE =	2.7663

labour_therm	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
retnat_ttl	-1.110824	.0218887	-50.75	0.000	-1.153728	-1.06792
income_ttl	.000946	.0057612	0.16	0.870	-.0103466	.0122386
wave_2_int	-.0200303	.0141997	-1.41	0.158	-.0478631	.0078024
wave_3_int	-.1893241	.0162476	-11.65	0.000	-.2211708	-.1574774
wave_4_int	(omitted)					
wave_5_int	-.3167173	.0132225	-23.95	0.000	-.3426346	-.2908
_cons	8.659463	.1057869	81.86	0.000	8.452111	8.866815

```
.      estimates store interact
```

# LR Test of Unit Effect Specifications

```
.      regr labour_therm retnat_ttl income_ttl
```

Source	SS	df	MS	Number of obs =	18653
Model	17549.7573	2	8774.87865	F( 2, 18650) =	1108.95
Residual	147573.07	18650	7.91276515	Prob > F =	0.0000
				R-squared =	0.1063
				Adj R-squared =	0.1062
Total	165122.827	18652	8.85282154	Root MSE =	2.813

labour_therm	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
retnat_ttl	-1.015335	.0215739	-47.06	0.000	-1.057622	-.973048
income_ttl	.0043702	.0058558	0.75	0.455	-.0071077	.0158482
_cons	7.861831	.1025247	76.68	0.000	7.660873	8.062789

```
.      lrtest interact
```

Likelihood-ratio test	LR chi2(3) =	627.56
(Assumption: . nested in interact)	Prob > chi2 =	0.0000

Consider the following data and treatment condition:

- cross-sectional data ( $t = 1$ )
- causal variable (or treatment)  $d_i$  scored 1 for the treatment group and 0 for the control group
- the treatment occurs a period,  $\tau$ , prior to  $t = 1$
- $y$  and  $d$  are the only observed variables
- $\gamma$  is a parameter for the causal effect
- $\delta_1$  is a period effect common to all units
- $\theta_{i|d}$  is a time-invariant unit-specific effect that captures unobserved unit heterogeneity and is conditional on  $d$
- $\epsilon_{i1|d}$  is white noise unique to the  $i$ th unit at  $t = 1$  conditional on  $d$  and  $\theta_i$

The control group:

$$y_{i1|d=0} = \theta_{i|d=0} + \delta_1 + \epsilon_{i1|d=0} \quad (9)$$

The treatment group:

$$y_{i1|d=1} = \gamma + \theta_{i|d=1} + \delta_1 + \epsilon_{i1|d=1} \quad (10)$$

The treatment effect?

$$E(y_{i1|d=1} - y_{i1|d=0}) = \gamma + E(\theta_{i|d=1} - \theta_{i|d=0}) + E(\epsilon_{i1|d=1} - \epsilon_{i1|d=0}) \quad (11)$$

- $\delta_1$  drops out
- mean of disturbances is independent of  $d$ :  
$$E(\epsilon_{1t|d=1} - \epsilon_{i1|d=0}) = 0$$
- unobserved heterogeneity is mean independent of the causal variable:  $[E(\theta_{i|d=1}) = E(\theta_{i|d=0})]$

which implies the OLS regression:

$$y_{i1} = \alpha + \gamma d_i + \mu_{i1} \quad (12)$$

where  $\mu_{i1} = \theta_{i|d} + \epsilon_{i1|d}$

The control group:

$$y_{i0|d=0} = \theta_{i|d=0} + \delta_0 + \epsilon_{i0|d=0} \quad (13)$$

$$y_{i1|d=0} = \theta_{i|d=0} + \delta_1 + \epsilon_{i1|d=0} \quad (14)$$

The treatment group:

$$y_{i0|d=1} = \gamma + \theta_{i|d=1} + \delta_0 + \epsilon_{i0|d=1} \quad (15)$$

$$y_{i1|d=1} = \gamma + \theta_{i|d=1} + \delta_1 + \epsilon_{i1|d=1} \quad (16)$$

The treatment effect?

$$E(y_{i1|d=1} - y_{i1|d=0}) - E(y_{i0|d=1} - y_{i0|d=0}) = \gamma + E(\epsilon_{i1|d=1} - \epsilon_{i0|d=1}) - E(\epsilon_{i1|d=1} - \epsilon_{i0|d=0})$$

(17)

- unit effects that were source of heterogeneity bias are eliminated
- identification does not require assumption that period effects are temporally stable
- exogeneity identification restriction holds:
- $E(\epsilon_{i1|d=1} - \epsilon_{i0|d=1}) - E(\epsilon_{i1|d=1} - \epsilon_{i0|d=0}) = 0$

The control and treatment groups:

$$y_{i0} = \delta_0 + \theta_{i|d} + \epsilon_{i0} \quad (18)$$

$$y_{i1} = \delta_1 + \gamma d_i + \theta_{i|d} + \epsilon_{i1} \quad (19)$$

Differencing then gives you:

$$(y_{i1} - y_{i0}) = (\delta_1 - \delta_0) + \gamma d_i + (\epsilon_{i1} - \epsilon_{i0}) \quad (20)$$

Least squares gives an estimate of treatment effect:

$$\hat{\gamma}_{dd} = (\bar{y}_{.1|d=1} - \bar{y}_{.1|d=0}) - (\bar{y}_{.0|d=1} - \bar{y}_{.0|d=0}) \quad (21)$$

Deviations from within-unit means:

$$\bar{y}_i = (\delta_1 + \delta_0)/2 + \gamma \bar{d}_i + \theta_i \hat{\epsilon}_i \quad (22)$$

Consistent fixed effects estimator:

$$(y_{it} - \bar{y}_i) = (\delta_0 - \delta_1)/2 + (\delta_1 - \delta_0)p_1 + \gamma(d_{it} - \bar{d}_i) + (\epsilon_{it} - \bar{\epsilon}_i) \quad (23)$$

# Regression Labour Evaluations, Economy and Income

```
.      regr labour_therm retnat_ttl income_ttl
```

Source	SS	df	MS	Number of obs =	18653
Model	17549.7573	2	8774.87865	F( 2, 18650) =	1108.95
Residual	147573.07	18650	7.91276515	Prob > F	= 0.0000
-----				R-squared	= 0.1063
Total	165122.827	18652	8.85282154	Adj R-squared	= 0.1062
-----				Root MSE	= 2.813

labour_therm	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
retnat_ttl	-1.015335	.0215739	-47.06	0.000	-1.057622	-.973048
income_ttl	.0043702	.0058558	0.75	0.455	-.0071077	.0158482
_cons	7.861831	.1025247	76.68	0.000	7.660873	8.062789

# Fixed Effects with Income

```
. xtreg labour_therm retnat_ttl income_ttl, fe
Fixed-effects (within) regression      Number of obs   =   18653
Group variable: id                    Number of groups =    8215

R-sq:  within = 0.0045                Obs per group:  min =     1
      between = 0.0568                  avg   =     2.3
      overall = 0.0557                  max   =     5

corr(u_i, Xb) = 0.2075                F(2,10436)      =   23.76
                                          Prob > F        =   0.0000
```

```
-----+-----
labour_therm |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----+-----
retnat_ttl |  -.0903765   .013828    -6.54  0.000   -0.1174819   -0.063271
income_ttl |  .0246439   .011645     2.12  0.034   .0018175    .0474702
   _cons |   3.831079  .1016761   37.68  0.000   3.631775    4.030384
-----+-----
sigma_u |  2.7985428
sigma_e |  1.282659
rho |   .8264003   (fraction of variance due to u_i)
-----+-----
F test that all u_i=0:      F(8214, 10436) =     9.65      Prob > F = 0.0000
```

# Fixed Effects with Gender

```
.      xtreg  labour_therm retnat_ttl female, fe
note: female omitted because of collinearity

Fixed-effects (within) regression              Number of obs   =   22890
Group variable: id                            Number of groups =   9538

R-sq:  within = 0.0035                          Obs per group:  min =    1
        between = 0.1729                          avg   =    2.4
        overall = 0.1101                          max   =    5

corr(u_i, Xb) = 0.3258                          F(1,13351)      =   46.90
                                                Prob > F        =   0.0000

-----+-----
labour_therm |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----+-----
retnat_ttl |   -.084394   .0123232    -6.85  0.000    - .1085491   -.0602388
female | (omitted)
_cons |   3.897776   .0526257    74.07  0.000     3.794622    4.00093
-----+-----
sigma_u |  2.7990027
sigma_e |  1.2753294
rho |   .82808525   (fraction of variance due to u_i)
-----+-----
F test that all u_i=0:      F(9537, 13351) =   10.28      Prob > F = 0.0000
```