Introduction to Panel-Data Analysis using Stata

Chris Cheng, Ph.D.

Staff Econometrician, StataCorp LLC
College Station, TX, USA

November 9, 2021

Outline

1. Basic concepts
   - Incentives
   - Stata tools
   - The data and the data generating process
   - The model
2. Linear models for panel data
   - Data generating process
   - Random effects
   - Fixed effects
   - Fixed or random effects
   - Marginal analysis
3. Dynamic panel-data linear models
   - Arellano–Bond
   - Arellano–Bover/Blundell–Bond
4. Extended regression models for panel data

What is panel?

- (mathstats) repeated measures (wide vs long format)
- (biostats) longitudinal data
- (economics) panel data

How it looks

– a typical panel vs. cross-sectional data structure

```
. list country year consumption gdp irate ///
        if CountryName=="Mexico" |        ///
        CountryName=="United States",     ///
        sepby(country) abbreviate(12) noobs
```

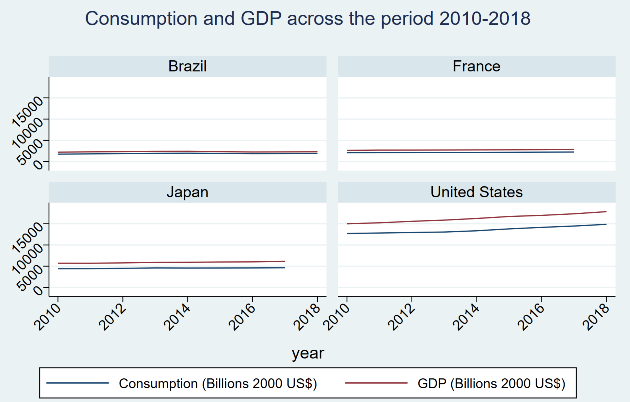
country	year	consumption	gdp	irate
Mexico	2010	815.78416	1057.8013	1.2125
Mexico	2011	842.78459	1096.5486	.95583333
Mexico	2012	863.83937	1136.4885	1.0816667
United States	2010	12695.979	14992.053	2.4000001
United States	2011	12812.144	15224.555	6.5
United States	2012	12932.334	15567.038	7

```
. list country consumption gdp irate           ///
        if country == 128 | country == 207, ///
              abbreviate(12) noobs
```

country	consumption	gdp	irate
Mexico	905.63994	1189.6593	1.4738889
United States	13561.267	16307.542	5.0000001

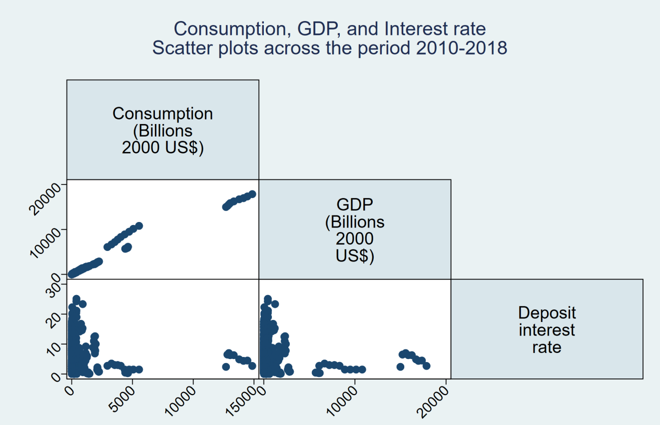
Source: http://databank.worldbank.org/data/Home.aspx

How it looks



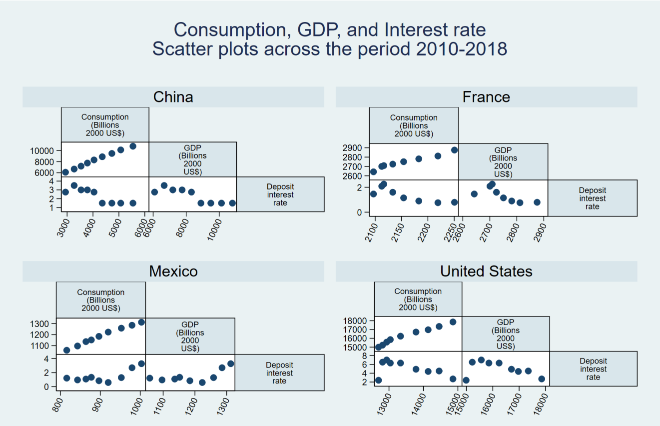
Source: http://databank.worldbank.org/data/Home.aspx
Stata command: xtline

How it works - Pooled vs. Panel



Source: http://databank.worldbank.org/data/Home.aspx
Stata command: graph matrix

How it works - Pooled vs. Panel



Source: http://databank.worldbank.org/data/Home.aspx
Stata command: graph matrix

Stata tools

- Data management
- Linear regression estimators
- Dynamic panel-data estimators
- Nonlinear regression estimators
- Postestimation tools
- Extended regression models

Tools related to panel

- **reshape** converts data from wide to long form and vice versa
- **xtsum** summarizes xt (panel) data
- **xttab** tabulates xt (panel) data, one-way table for categorical variables
- **xttrans** tabulates xt (panel) data and reports transition probabilities
- **duplicates** reports, tags, or drops duplicate observations
- **panelstat** a *community-contributed* (user-written) command, computes statistics for panel data
  https://www.stata.com/meeting/portugal17/slides/portugal17_Silva.pdf

LINEAR PANEL-DATA MODELS

Data-generating process (DGP)

- The data-generating process is given by

$$
\begin{aligned}
y_{it} &= \beta_0 + \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + \eta_{it} \\
\eta_{it} &\equiv \alpha_i + \varepsilon_{it} \\
i &= 1, \ldots n \\
t &= 1, \ldots T
\end{aligned}
$$

- The random disturbance (η_{it}) includes two parts:
  - α_i: the ***unobservable*** component is particular to each panel and is ***time-invariant*** (e.g. for individuals: ability, intelligence, work ethic). As in the regression case, the assumptions made on η_{it}, with particular emphasis on α_i, define the models we work with.
  - ε_{it}: the idiosyncratic error term

Model for aggregate consumption

$$\text{consumption}_{it} = \beta_0 + \beta_1 * \text{gdp}_{it} + \beta_2 * \text{irate}_{it} + \alpha_i + \varepsilon_{it}$$

- World Bank public online data on
  consumption: Final consumption expenditure (2010 US$)
  gdp: Gross domestic product (2010 US$)
  irate: deposit interest rate
- Example : 2010-2018 for 131 countries
- Source: http://databank.worldbank.org/data/Home.aspx

Specifying the panel structure in Stata

Users can tell Stata that data have a special structure for various types of datasets

- Repeated measures/Panel data/Longitudinal data datasets – see **help xtset**
- Time-series datasets – see **help tsset**
- Survival time datasets – see **help stset**
- Datasets arising from complex survey designs (called survey datasets) – see **help svyset**

Specifying the panel structure in Stata

Assuming that the second dimension corresponds to time series, we use the **xtset** command to specify the panel structure with

- Panel identifier variable (e.g. country)
- Time identifier variable (e.g. year)

Specifying the panel structure in Stata

Assuming that the second dimension corresponds to time series, we use the **xtset** command to specify the panel structure with

- Panel identifier variable (e.g. country)
- Time identifier variable (e.g. year)

```
. xtset country year

       Panel variable:  country (unbalanced)
        Time variable:  year, 2010 to 2018, but with a gap
                Delta:  1 unit
```

P.S. You can specify the panel structure using **xtset panelvar** if you want to ignore the time structure.

Theoretical framework (DGP revisit)

$$\begin{aligned} y_{it} &= \beta_0 + \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + \eta_{it} \\ \eta_{it} &= \alpha_i + \varepsilon_{it} \end{aligned}$$

- As in the classical linear regression, all models are defined by two components:
  1. The data-generating process (DGP)
  2. The relationship between the random disturbance or idiosyncratic shock and the explanatory variables
  3. How the relationship is structured defines your model: random- vs. fixed- effects

Random-effects model

- The regressors are unrelated to the ***unobserved time-invariant component*** α_i

$$E\left(\alpha_i | x_{it1}, \ldots x_{itk}\right) = E\left(\alpha_i\right)$$

Random-effects model

- The regressors are unrelated to the ***unobserved time-invariant component*** α_i

$$E\left(\alpha_i | x_{it1}, \ldots x_{itk}\right) = E\left(\alpha_i\right)$$

- strict exogeneity, no lagged dependent variables

$$E\left(\varepsilon_{it} | x_{it1}, \ldots, x_{itk}, \alpha_i\right) = 0$$

Random-effects model

- The regressors are unrelated to the ***unobserved time-invariant component*** α_i

$$E(\alpha_i | x_{it1}, \ldots x_{itk}) = E(\alpha_i)$$

- strict exogeneity, no lagged dependent variables

$$E(\varepsilon_{it} | x_{it1}, \ldots, x_{itk}, \alpha_i) = 0$$

- The previous two assumptions allow us to think about using a regression. But:

$$\begin{aligned} E\left(\varepsilon_i \varepsilon_i' | x_i, \alpha_i\right) &= \sigma_\varepsilon^2 I_T \\ E\left(\varepsilon_{it}^2\right) &= \sigma_\varepsilon^2 \\ E\left(\varepsilon_{it}\varepsilon_{is}\right) &= 0 \\ V\left(\alpha_i^2\right) &= E\left(\alpha_i^2 | x_i\right) = \sigma_\alpha^2 \end{aligned}$$

Random-effects covariance-variance matrix

- For each individual, we have that

$$\Omega = E\left(\eta_i \eta_i'\right) = \begin{pmatrix} \sigma_\varepsilon^2 + \sigma_\alpha^2 & \sigma_\alpha^2 & \dots & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_\varepsilon^2 + \sigma_\alpha^2 & \dots & \vdots \\ \vdots & & \ddots & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_\alpha^2 & \dots & \sigma_\varepsilon^2 + \sigma_\alpha^2 \end{pmatrix}$$

Random-effects covariance-variance matrix

- For each individual, we have that

$$\Omega = E\left(\eta_i \eta_i'\right) = \begin{pmatrix} \sigma_\varepsilon^2 + \sigma_\alpha^2 & \sigma_\alpha^2 & \dots & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_\varepsilon^2 + \sigma_\alpha^2 & \dots & \vdots \\ \vdots & & \ddots & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_\alpha^2 & \dots & \sigma_\varepsilon^2 + \sigma_\alpha^2 \end{pmatrix}$$

- This gives rise to an efficient estimator:

$$\begin{aligned} \Omega^{-1/2} y_i &= \Omega^{-1/2} x_i \beta + \Omega^{-1/2} \eta_i \\ \Omega^{-1/2} z_i &\equiv z_i^* \end{aligned}$$

- This implies that we have the following model:

$$\begin{aligned} y_i^* &= x_i^* \beta + \eta_i^* \\ E\left(\eta_i^* \eta_i^{*\prime}\right) &= I_T \end{aligned}$$

Random-effects estimation with Stata

```
. use panel_webinar_stata
. xtset country year
       Panel variable:  country (unbalanced)
        Time variable:  year, 2010 to 2018, but with a gap
                Delta:  1 unit
```

```
. describe
Contains data from panel_webinar_stata.dta
 Observations:         1,016
    Variables:            10                  13 Aug 2020 16:37
--------------------------------------------------------------------------------
Variable      Storage   Display    Value
    name         type    format    label      Variable label
--------------------------------------------------------------------------------
CountryName     str30   %30s                  Country Name
year            float   %9.0g                 Year
irate           double  %10.0g                Deposit interest rate
consumption     double  %10.0g                Consumption (Billions 2000 US$)
gdp             double  %10.0g                GDP (Billions 2000 US$)
country         long    %30.0g     country    Country Name
ln_cons         float   %9.0g                 Log of consumption
ln_gdp          float   %9.0g                 Log of gdp
ln_irate        float   %9.0g                 Log of irate
region          long    %12.0g     region     Regional groups
--------------------------------------------------------------------------------
Sorted by: country  year
```

Random-Effects Estimation with Stata

```
. xtreg ln_cons ln_gdp ln_irate, re

Random-effects GLS regression                   Number of obs     =      1,016
Group variable: country                         Number of groups  =        131
R-sq:                                           Obs per group:
     within  = 0.8033                                         min =          1
     between = 0.9859                                         avg =        7.8
     overall = 0.9847                                         max =          9
                                                Wald chi2(2)      =   13277.81
corr(u_i, X) = 0 (assumed)                      Prob > chi2       =     0.0000

------------------------------------------------------------------------------
     ln_cons | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
      ln_gdp |    .958856   .0084128   113.98   0.000     .9423672    .9753449
    ln_irate |  -.0039294   .0041147    -0.95   0.340     -.011994    .0041352
       _cons |    .760708   .2065915     3.68   0.000     .3557961     1.16562
-------------+----------------------------------------------------------------
     sigma_u |   .2339765
     sigma_e |  .05205235
         rho |  .95284182   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Interpreting results

- The `Wald chi2(df)` statistic is the equivalent of the `F` and regards the overall relevance of the model
- The three different `R-sq` statistics represent the variability of y explained by its predicted values. But there are three possible measures of y:
  1. y_{it} OVERALL
  2. $\bar{y}_i$ BETWEEN
  3. $y_{it} - \bar{y}_i$ WITHIN
- `corr(u_i, X)` refers to the correlation between the time-invariant component α_i, in this case called `u_i`, and the regressors. For the random effects, we assume it is zero.
- `sigma_u`$= \sigma_\alpha$,
  `sigma_e`$= \sigma_\varepsilon$,
  `rho`$= \sigma_\alpha^2 \left(\sigma_\varepsilon^2 + \sigma_\alpha^2\right)^{-1}$

Random-effects estimation with Stata

Random effects vs. Pooled OLS

```
. xttest0

Breusch and Pagan Lagrangian multiplier test for random effects

    ln_cons[country,t] = Xb + u[country] + e[country,t]

       Estimated results:
                        |       Var     SD = sqrt(Var)
                ---------+-----------------------------
                 ln_cons |   4.027035       2.006747
                       e |   .0027094       .0520524
                       u |    .054745       .2339765

       Test: Var(u) = 0

                         chibar2(01) =   3108.85

                      Prob > chibar2 =    0.0000
```

Fixed effects models

- The regressors are correlated with the ***unobserved time-invariant*** component α_i

$$Cov\left(\alpha_i, x_i\right) \neq 0$$

Fixed effects models

- The regressors are correlated with the ***unobserved time-invariant*** component α_i

$$Cov\left(\alpha_i, x_i\right) \neq 0$$

- strict exogeneity, no lagged dependent variables:

$$E\left(\varepsilon_{it} | x_{it1}, \ldots, x_{itk}, \alpha_i\right) = 0$$

Interpretation

- In the model we have been discussing:

  $$ln\,(\text{consumption}_{it}) = \beta_0 + \beta_1 ln\,(\text{GDP}_{it}) + \beta_2 ln\,(\text{irate}_{it}) + \alpha_i + \varepsilon_{it}$$

- It is difficult to maintain, for a particular model, that the unobserved individual component is independent of all regressors

Within transform

$$y_{it} = \beta_0 + \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + \alpha_i + \varepsilon_{it} \quad (1)$$

- If we take the average over the T observations of each panel, we obtain

$$\bar{y}_i = \beta_0 + \beta_1 \bar{x}_{i1} + \ldots + \beta_k \bar{x}_{ik} + \alpha_i + \bar{\varepsilon}_i$$

Within transform

$$y_{it} = \beta_0 + \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + \alpha_i + \varepsilon_{it} \qquad (1)$$

- If we take the average over the T observations of each panel, we obtain

$$\bar{y}_i = \beta_0 + \beta_1 \bar{x}_{i1} + \ldots + \beta_k \bar{x}_{ik} + \alpha_i + \bar{\varepsilon}_i$$

Where

$$\bar{y}_i = T^{-1} \sum_{t=1}^{T} y_{it},$$

$$\bar{x}_{ij} = T^{-1} \sum_{t=1}^{T} x_{itj}$$

Within transform

$$y_{it} = \beta_0 + \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + \alpha_i + \varepsilon_{it} \qquad (1)$$

- If we take the average over the T observations of each panel, we obtain

$$\bar{y}_i = \beta_0 + \beta_1 \bar{x}_{i1} + \ldots + \beta_k \bar{x}_{ik} + \alpha_i + \bar{\varepsilon}_i$$

Where

$$\bar{y}_i = T^{-1} \sum_{t=1}^{T} y_{it},$$

$$\bar{x}_{ij} = T^{-1} \sum_{t=1}^{T} x_{itj}$$

- We now can construct the following object:

$$y_{it} - \bar{y}_i = (\beta_0 - \beta_0) + \beta_1 (x_{it1} - \bar{x}_{i1}) + \ldots + \beta_k (x_{itk} - \bar{x}_{ik}) + (\alpha_i - \alpha_i) + (\varepsilon_{it} - \bar{\varepsilon}_i)$$

- And we can then estimate the parameters of interest from equation (1):

$$\tilde{y}_i = \beta_1 \tilde{x}_{i1} + \ldots + \beta_k \tilde{x}_{ik} + \tilde{\varepsilon}_i$$

Within estimation

```
. xtreg ln_cons ln_gdp ln_irate, fe

Fixed-effects (within) regression               Number of obs     =      1,016
Group variable: country                         Number of groups  =        131
R-sq:                                           Obs per group:
     within  = 0.8034                                         min =          1
     between = 0.9858                                         avg =        7.8
     overall = 0.9845                                         max =          9
                                                F(2,883)          =    1804.60
corr(u_i, Xb)  = -0.0175                        Prob > F          =     0.0000

------------------------------------------------------------------------------
     ln_cons | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      ln_gdp |    .958713    .016523    58.02   0.000      .926284     .991142
    ln_irate |  -.0074047   .0042761    -1.73   0.084    -.0157972    .0009878
       _cons |   .7750608   .4063998     1.91   0.057    -.0225615    1.572683
-------------+----------------------------------------------------------------
     sigma_u |  .24585324
     sigma_e |  .05205235
         rho |  .95709727   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0: F(130, 883) = 152.63                  Prob > F = 0.0000
```

Fixed effects vs. Random effects

- Theory should be one of the main factors guiding your modeling decision
- However, you should present a statistical test to back up your claims
  1. Hausman test for fixed effects vs random effects
  2. Mundlak test for fixed effects vs random effects

Hausman test

- The following object has a Chi-Squared distribution with degrees of freedom equal to the number of regressors:

$$H = \left(\hat{\beta}_{fe} - \hat{\beta}_{re}\right)' \left[\widehat{VCE}_{fe} - \widehat{VCE}_{re}\right]^{-1} \left(\hat{\beta}_{fe} - \hat{\beta}_{re}\right)$$

- The test implicitly assumes that the random effects model is efficient, which in turn makes $\left[\widehat{VCE}_{fe} - \widehat{VCE}_{re}\right]$ positive definite.
- The test rules out heteroskedasticity and serial correlation

Hausman test


```
. quietly xtreg ln_cons ln_gdp ln_irate, fe
. estimates store eq_fe
. quietly xtreg ln_cons ln_gdp ln_irate, re
. estimates store eq_re

. hausman eq_fe eq_re

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |     eq_fe        eq_re        Difference       Std.err.
-------------+----------------------------------------------------------------
      ln_gdp |    .958713      .958856        -.000143        .0142209
    ln_irate |  -.0074047    -.0039294       -.0034753        .0011638
------------------------------------------------------------------------------
                 b = Consistent under Ho and Ha; obtained from xtreg.
  B = Inconsistent under Ha, efficient under Ho; obtained from xtreg.

Test of Ho: Difference in coefficients not systematic

    chi2(2) = (b-B)'[(V_b-V_B)^(-1)](b-B)
            =  17.25

Prob > chi2 = 0.0002
```

Mundlak test

Alternatively, the Mundlak test can be used for comparing fixed effects and random effects.

The Stata Blog

"Fixed effects or random effects: The Mundlak approach"

Enrique Pinzon, Associate Director of Econometrics
https://blog.stata.com/2015/10/29/fixed-effects-or-random-effects-the-mundlak-approach/

MARGINAL ANALYSIS

Interpreting the coefficients

- Notice that all the variables are in natural logs. Therefore:

$$E\left(ln\left(y_{it}\right)|lnx_{it},\alpha_i\right) = \beta_0 + \beta_1 lnx_{it1} + \ldots + \beta_k lnx_{itk} + \alpha_i$$

- If you want the impact of a continuous regressor on y_{it}:

$$\frac{\partial E\left(y_{it}|x_{it},\alpha_i\right)}{\partial x_{itj}} \frac{x_{itj}}{E\left(y_{it}|x_{it},\alpha_i\right)} = \beta_j$$

Interpreting the coefficients

- Notice that all the variables are in natural logs. Therefore:

$$E\left(ln\left(y_{it}\right)|lnx_{it}, \alpha_i\right) = \beta_0 + \beta_1 lnx_{it1} + \ldots + \beta_k lnx_{itk} + \alpha_i$$

- If you want the impact of a continuous regressor on y_{it}:

$$\frac{\partial E\left(y_{it}|x_{it}, \alpha_i\right)}{\partial x_{itj}} \frac{x_{itj}}{E\left(y_{it}|x_{it}, \alpha_i\right)} = \beta_j$$

- Use **margins** to get the elasticities (**dydx()** in this particular case):

```
. quietly xtreg ln_cons ln_gdp ln_irate, fe
. margins, dydx(*)
```

```
Average marginal effects                           Number of obs = 1,016
Model VCE     : Conventional

Expression    : Linear prediction, predict()
dy/dx w.r.t.  : ln_gdp ln_irate
```

	dy/dx	Delta-method Std. err.	z	P>|z|	[95% conf.	interval]
ln_gdp	.958713	.016523	58.02	0.000	.9263284	.9910976
ln_irate	-.0074047	.0042761	-1.73	0.083	-.0157857	.0009763

Marginal effects with interactions

- Regional interactions with **ln_gdp**:

```
. quietly xtreg ln_cons ln_gdp ///
        i.region#c.ln_gdp ln_irate, fe

. margins, dydx(ln_gdp) over(region)

Average marginal effects                        Number of obs = 986
Model VCE    : Conventional

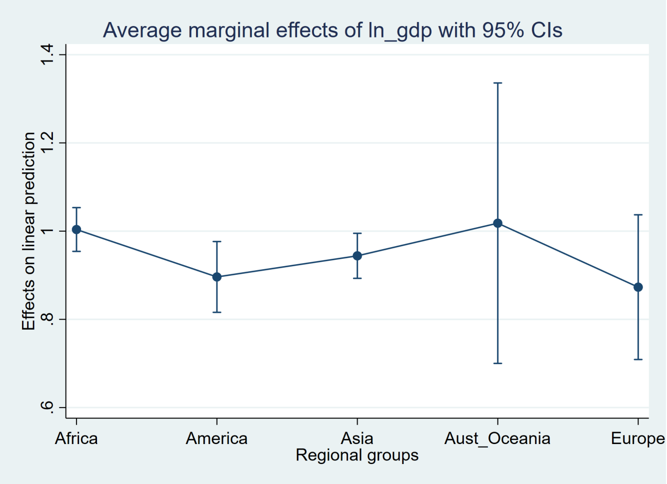
Expression   : Linear prediction, predict()
dy/dx w.r.t. : ln_gdp
over         : region
```

		Delta-method				
	dy/dx	Std. err.	z	P>|z|	[95% conf.	interval]
ln_gdp						
region						
Africa	1.003669	.0253091	39.66	0.000	.9540644	1.053274
America	.8961536	.0409304	21.89	0.000	.8159314	.9763758
Asia	.9440403	.0260334	36.26	0.000	.8930157	.9950649
Aust_Oceania	1.017993	.1622033	6.28	0.000	.70008	1.335905
Europe	.8729883	.0837015	10.43	0.000	.7089363	1.03704

Marginal effects by region

```
. marginsplot

Variables that uniquely identify margins: region
```



Predictive margins with interactions

- Regional interactions with **ln_irate**:

```
. quietly xtreg ln_cons ln_gdp i.region##c.ln_irate, re

. margins region, at(ln_irate=(-4(2)0))
```

```
Predictive margins                              Number of obs = 986
Model VCE     : Conventional

Expression    : Linear prediction, predict()

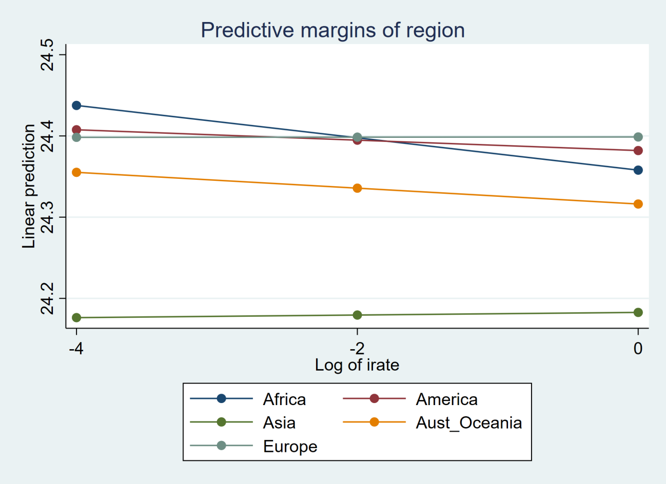
1._at: ln_irate = -4
2._at: ln_irate = -2
3._at: ln_irate =  0

------------------------------------------------------------------------------
              |            Delta-method
              |     Margin   Std. err.      z    P>|z|     [95% conf. interval]
------------------------------------------------------------------------------
output omitted
------------------------------------------------------------------------------
```

Predictive margins by region

```
. marginsplot, noci
```

```
Variables that uniquely identify margins: ln_irate region
```



DYNAMIC PANEL-DATA MODELS

Dynamic models

$$y_{it} = \beta_0 + \beta_1 y_{i(t-1)} + x_{it}'\beta_2 + \alpha_i + \varepsilon_{it}$$

- In the model above, x_{it} could also include lagged variables.

Dynamic models

$$y_{it} = \beta_0 + \beta_1 y_{i(t-1)} + x_{it}'\beta_2 + \alpha_i + \varepsilon_{it}$$

- In the model above, x_{it} could also include lagged variables.
- Taking first differences:

$$\Delta y_{it} = \beta_1 \Delta y_{i(t-1)} + \Delta x_{it}'\beta_2 + (\alpha_i - \alpha_i) + \Delta\varepsilon_{it}$$

Dynamic models

$$y_{it} = \beta_0 + \beta_1 y_{i(t-1)} + x_{it}'\beta_2 + \alpha_i + \varepsilon_{it}$$

- In the model above, x_{it} could also include lagged variables.
- Taking first differences:

$$\Delta y_{it} = \beta_1 \Delta y_{i(t-1)} + \Delta x_{it}'\beta_2 + (\alpha_i - \alpha_i) + \Delta\varepsilon_{it}$$

- We have eliminated the fixed effect but notice that

$$E\left(\Delta y_{i(t-1)} \Delta\varepsilon_{it}\right) \neq 0$$

Instrumental-variable (GMM) estimation

- The key to estimation is to find a set of instruments that satisfy
$$E\left(z_{it}\Delta\varepsilon_{it}\right) = 0$$
- This gives rise to the following models:
  - Anderson–Hsiao $y_{i(t-2)}$ and $\Delta y_{i(t-2)}$ (**xtivreg, fd**).

Instrumental-variable (GMM) estimation

- The key to estimation is to find a set of instruments that satisfy
$$E\left(z_{it}\Delta\varepsilon_{it}\right) = 0$$
- This gives rise to the following models:
  - Anderson–Hsiao $y_{i(t-2)}$ and $\Delta y_{i(t-2)}$ (**xtivreg, fd**).
  - Arellano and Bond suggest using all available lag levels (not only the second lag) for the first difference equation (**xtabond**).

Arellano-Bond

```
. xtabond ln_cons ln_gdp ln_irate, twostep
Arellano-Bond dynamic panel-data estimation     Number of obs     =        761
Group variable: country                         Number of groups  =        121
Time variable: year
                                                Obs per group:
                                                              min =          1
                                                              avg =   6.289256
                                                              max =          7

Number of instruments =     31                  Wald chi2(3)      =    9345.33
                                                Prob > chi2       =     0.0000
Two-step results
------------------------------------------------------------------------------
     ln_cons | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
     ln_cons |
         L1. |   .3616734   .0072234    50.07   0.000     .3475158     .375831
             |
      ln_gdp |    .602238   .0073699    81.72   0.000     .5877932    .6166828
    ln_irate |  -.0085773   .0024087    -3.56   0.000    -.0132982   -.0038564
       _cons |   .7702696   .2566304     3.00   0.003     .2672833    1.273256
------------------------------------------------------------------------------
Warning: gmm two-step standard errors are biased; robust standard
         errors are recommended.
Instruments for differenced equation
        GMM-type: L(2/.).ln_cons
        Standard: D.ln_gdp D.ln_irate

Instruments for level equation
        Standard: _cons
```

Model specification

```
. estat sargan

Sargan test of overidentifying restrictions
H0: Overidentifying restrictions are valid

      chi2(27)      =  32.56842
      Prob > chi2   =    0.2117
```

- The overidentification restriction is a test of the validity of the instruments under correct specification.

Model specification

```
. estat abond
Arellano-Bond test for zero autocorrelation in first-differenced errors
H0: No autocorrelation
```

Order	z	Prob > z
1	-2.5248	0.0116
2	1.5938	0.1110

- The Arellano–Bond test is testing that H0: $E\left[\Delta\varepsilon_{it}\Delta\varepsilon_{i(t-1)}\right] \neq 0$:

$$
\begin{aligned}
E\left[\Delta\varepsilon_{it}\Delta\varepsilon_{i(t-1)}\right] &= E\left[\left(\varepsilon_{it}-\varepsilon_{i(t-1)}\right)\left(\varepsilon_{i(t-1)}-\varepsilon_{i(t-2)}\right)\right] \\
&= E\left[\varepsilon^2_{i(t-1)}\right] + 0
\end{aligned}
$$

- According to our asssumptions, we should reject this hypothesis. Also, according to our hypothesis,

$$
\begin{aligned}
E\left[\Delta\varepsilon_{it}\Delta\varepsilon_{i(t-2)}\right] &= E\left[\left(\varepsilon_{it}-\varepsilon_{i(t-1)}\right)\left(\varepsilon_{i(t-2)}-\varepsilon_{i(t-3)}\right)\right] \\
&= E\left(\varepsilon_{it}\varepsilon_{i(t-2)}\right) - E\left(\varepsilon_{it}\varepsilon_{i(t-3)}\right) + E\left(\varepsilon_{i(t-1)}\varepsilon_{i(t-2)}\right) \\
&\quad - E\left(\varepsilon_{i(t-1)}\varepsilon_{i(t-3)}\right) \\
&= 0
\end{aligned}
$$

A new set of moment conditions

- The lagged-level instruments in **xtabond** become weak as the AR process becomes too persistent or σ_u^2/σ_e^2 becomes too large, so a new set of moments conditions are proposed:

A new set of moment conditions

- The lagged-level instruments in **xtabond** become weak as the AR process becomes too persistent or σ_u^2/σ_e^2 becomes too large, so a new set of moments conditions are proposed:

$$
\begin{aligned}
E\left(z_{it}\Delta\varepsilon_{it}\right) &= 0 \\
E\left(\Delta z_{it}\varepsilon_{it}\right) &= 0
\end{aligned}
$$

- These are defined by Arellano–Bover/Blundell–Bond
- Notice that you have moments for the equation in levels and for the equation in first difference
- Fit this model with **xtdpdsys**

Arellano–Bover/Blundell–Bond


```
. xtdpdsys ln_cons ln_gdp ln_irate, twostep
System dynamic panel-data estimation            Number of obs      =       884
Group variable: country                         Number of groups   =       122
Time variable: year
                                                Obs per group:
                                                              min =          1
                                                              avg =   7.245902
                                                              max =          8

Number of instruments =     38                  Wald chi2(3)       =  36908.02
                                                Prob > chi2        =    0.0000
Two-step results
------------------------------------------------------------------------------
     ln_cons | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
     ln_cons |
         L1. |    .464653   .0063034    73.71   0.000     .4522985    .4770074
             |
      ln_gdp |   .4918536   .0051095    96.26   0.000     .4818391     .501868
    ln_irate |  -.0092232   .0029176    -3.16   0.002    -.0149415   -.0035049
       _cons |   .9754017   .1538629     6.34   0.000     .6738359    1.276967
------------------------------------------------------------------------------
Warning: gmm two-step standard errors are biased; robust standard
         errors are recommended.
Instruments for differenced equation
        GMM-type: L(2/.).ln_cons
        Standard: D.ln_gdp D.ln_irate

Instruments for level equation
        GMM-type: LD.ln_cons
        Standard: _cons
```

Arellano–Bover/Blundell–Bond

Overidentification and Autocorrelation Tests

```
. estat sargan

 Sargan test of overidentifying restrictions
H0: Overidentifying restrictions are valid

        chi2(34)      =  46.01339
        Prob > chi2   =    0.0819

. estat abond

 Arellano-Bond test for zero autocorrelation in first-differenced errors

H0: no autocorrelation
```

Order	z	Prob > z
1	-2.6633	0.0077
2	1.6218	0.1048

Your Own Dynamic Model

- This model relies heavily on the idea that the dynamics are correctly specified

Your Own Dynamic Model

- This model relies heavily on the idea that the dynamics are correctly specified
- For instance, you could have

$$\begin{aligned} y_{it} &= \beta_0 + \beta_1 y_{i(t-1)} + x_{it}'\beta_2 + \alpha_i + \varepsilon_{it} + \gamma\varepsilon_{i(t-1)} \\ \Delta y_{it} &= \Delta\beta_1 y_{i(t-1)} + \Delta x_{it}'\beta_2 + \Delta\varepsilon_{it} + \gamma\Delta\varepsilon_{i(t-1)} \end{aligned}$$

- You now need to construct a new set of instruments that satisfy the moment conditions.
- Stata allows you to do this with **xtdpd**. You need to specify the instruments for the level and difference equations.

Extended Regression Models

Extended regression models (ERMs)

Problems: Endogeneity, selection, and nonrandom treatment assignment

- ***Endogeneity*** Unobserved variable affects causal relation
- ***Selection*** Part of sample (outcome) is missing not at random
- ***Nonrandom treatment*** You want something that looks like an experiment

ERMs account for all of these problems simultaneously, whether you have a continuous, binary, interval, or ordered outcome variable.

ERMs can also be used to fit panel-data random effects and two-level multilevel models.

Extended regression models (ERMs) for panel data

- Random-effects linear regression with endogenous covariates
  -
    ```
    xteregress y x1 x2, ///
            endogenous(w = x1 z1 z2)
    ```
- Random-effects linear regression with sample selection
  -
    ```
    xteregress y x1 x2, ///
            select(selected = x2 w2)
    ```
- Random-effects linear regression with endogenous treatment
  -
    ```
    xteregress y x1 x2, ///
            entreat(treatment = w z2 z3)
    ```

Extended regression models for panel data

- Random-effects probit regression
  - ```
    xteprobit y x1 x2,                      ///
            endogenous(w = x1 z1 z2) ///
            select(selected = x2 w2) ///
            entreat(treatment = w z2 z3)
    ```
- Random-effects ordered probit regression
  - `xteoprobit`
- Random-effects interval regression
  - `xteintreg`
- Random-effects Heckman model
  - `xtheckman`
    ```

Random-effects probit regression with sample selection

```
. webuse womenhlthre,clear
(Women´s health status panel)
. xtset personid year
       panel variable:  personid (strongly balanced)
        time variable:  year, 2010 to 2013
                delta:  1 unit
. generate goodhlth = health>3 if !missing(health)
. label var goodhlth "Good-Excellent Health condition"


. describe
Contains data from https://www.stata-press.com/data/r17/womenhlthre.dta
 Observations:         7,200                 Women´s health status panel
    Variables:            10                 6 Sep 2020 16:14
-----------------------------------------------------------------------------
Variable      Storage   Display    Value
    name         type    format    label      Variable label
-----------------------------------------------------------------------------
grade           byte    %8.0g                 Years of education
personid        int     %9.0g                 Person ID
year            int     %9.0g                 Year
workschool      byte    %8.0g      yesno      Employed or in school
insured         byte    %8.0g      yesno      Has health insurance
regcheck        byte    %8.0g      yesno      Has regular checkups
select          byte    %8.0g                 In sample
exercise        byte    %8.0g      yesno      Exercises regularly
health          byte    %9.0g      status     Health status
goodhlth        float   %9.0g                 Good-Excellent Health condition
-----------------------------------------------------------------------------
Sorted by: personid  year
```

Random-effects probit regression with sample selection

```
. xteprobit goodhlth i.exercise grade, select(select = grade i.regcheck)

(setting technique to bhhh)
Iteration 0:   log likelihood =  -6840.671
Iteration 1:   log likelihood = -6808.6475
Iteration 2:   log likelihood = -6808.1535
Iteration 3:   log likelihood = -6808.1515
Iteration 4:   log likelihood = -6808.1515
Extended probit regression                  Number of obs     =      7,200
                                                  Selected    =      5,421
                                                  Nonselected =      1,779
Group variable: personid                    Number of groups  =      1,800
                                            Obs. per group:
                                                          min =          4
                                                          avg =        4.0
                                                          max =          4
Integration method: mvaghermite             Integration pts.  =          7
                                            Wald chi2(2)      =     348.34
Log likelihood = -6808.1515                 Prob > chi2       =     0.0000
```

Random-effects probit regression with sample selection


```
. xteprobit goodhlth i.exercise grade, select(select = grade i.regcheck)
```

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
goodhlth						
exercise						
yes	.3554439	.0400762	8.87	0.000	.276896	.4339919
grade	.1743015	.0095533	18.25	0.000	.1555774	.1930256
_cons	-2.252753	.1154867	-19.51	0.000	-2.479102	-2.026403
select						
grade	.0832256	.007392	11.26	0.000	.0687376	.0977137
regcheck						
yes	.4800144	.036039	13.32	0.000	.4093793	.5506495
_cons	-.5420435	.0964841	-5.62	0.000	-.731149	-.3529381
corr(e.select,e.goodhlth)	.8060986	.0855705	9.42	0.000	.5627727	.9208657
var(goodhlth[personid])	.2640095	.0364768			.2013787	.346119
var(select[personid])	.1538155	.0271043			.1088948	.2172667
corr(select[personid], goodhlth[personid])	.6224091	.0808206	7.70	0.000	.4384837	.7562961

Random-effects probit regression with endogenous treatment

```
. xteprobit goodhlth i.exercise grade,      ///
>     entreat(insured = i.workschool, nore) nolog

Extended probit regression                     Number of obs     =      7,200
Group variable: personid                       Number of groups  =      1,800

                                               Obs. per group:
                                                             min =          4
                                                             avg =        4.0
                                                             max =          4

Integration method: mvaghermite                Integration pts.  =          7

                                               Wald chi2(6)      =     265.10
Log likelihood =  -7572.592                    Prob > chi2       =     0.0000
```

Random-effects probit regression with endogenous treatment


```
. xteprobit goodhlth i.exercise grade,      ///
>    entreat(insured = i.workschool, nore) nolog
```

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
goodhlth						
exercise#insured						
yes#no	.5563098	.0916258	6.07	0.000	.3767266	.735893
yes#yes	.486376	.0454754	10.70	0.000	.3972458	.5755062
insured#c.grade						
no	.0125397	.0207005	0.61	0.545	-.0280325	.053112
yes	.0788714	.0100576	7.84	0.000	.0591589	.098584
insured						
no	-1.398234	.3668983	-3.81	0.000	-2.117342	-.679127
yes	-.6820556	.1458962	-4.67	0.000	-.9680069	-.3961043
insured						
workschool						
yes	.6620277	.058127	11.39	0.000	.5481008	.7759545
_cons	-.0088057	.0557336	-0.16	0.874	-.1180415	.1004301
corr(e.insured,e.goodhlth)	.3433395	.1522733	2.25	0.024	.0195374	.6019547
var(goodhlth[personid])	.3394691	.0451158			.2616222	.4404797

Summary

1. Basic Concepts
2. Linear panel-data models
    - Random effects
    - Fixed effects
    - Marginal analysis
3. Dynamic panel-data models
    - Arellano–Bond
    - Arellano–Bover/Blundell–Bond
4. Extended regression models for panel data

New in Stata 17

- Fixed-effects and random-effects multinomial logit models
  - `xtmlogit y x1 x2, re|fe`
- Bayesian longitudinal/panel-data models
  - https://www.stata.com/new-in-stata/bayesian-longitudinal-panel-data-models/
- Difference-in-differences (DID) and DDD models
  - `xtdidregress (ovar omvarlist) (tvar[, continuous]), group(groupvars) [time(timevar)]`

Online documentation
-xt- commands

https://www.stata.com/bookstore/longitudinal-panel-data-reference-manual/

Send questions to Tech Support

tech-support@stata.com

Upcoming webinars

https://www.stata.com/training/webinar/