Panel GLMs

Department of Political Science and Government
Aarhus University

May 12, 2015

Segue From Event-History

- Event history analysis involves the analysis of durations and probabilities of state changes over time across many units
- Each unit's trajectory or history can begin at an arbitrary point in time
  - Ex. 1: Colony's time to independence after 1900
  - Ex. 2: Durability of democratic government after independence
- In problems (like Ex. 1), we are interested in studying units over the *same period of time*

Panel Analysis

- In event history analysis, time is our key variable
- In panel analysis:
  - unit characteristics are our key variables
  - observations exist simultaneously
- We are interested in effects of X on Y

Terminology

Terminology

- *Panel*

Terminology

- *Panel*
- *Wide* versus *Long* data

Terminology

- *Panel*
- *Wide* versus *Long* data
- *Time-varying* versus *time-invariant*

Terminology

- *Panel*
- *Wide* versus *Long* data
- *Time-varying* versus *time-invariant*
- *Balanced* versus *Unbalanced* panel

Terminology

- *Panel*
- *Wide* versus *Long* data
- *Time-varying* versus *time-invariant*
- *Balanced* versus *Unbalanced* panel
- *Fixed effects*

Terminology

- *Panel*
- *Wide* versus *Long* data
- *Time-varying* versus *time-invariant*
- *Balanced* versus *Unbalanced* panel
- *Fixed effects*
- *Random effects*

Panel versus Time-Series

- Cross-sectional data involve many units observed at one time
- Panel data involve many units over at multiple points in time
- Time-series data involve one (or more) units observed at multiple points time
- Time-Series, Cross-Sectional (TSCS) data are panel data
  - Sometimes the units are aggregations
- *Within-subjects* analysis is panel analysis

Causal Inference

- What is the goal of causal inference?
- How do we define a causal effect (in terms of counterfactuals)?

Causal Inference

- What is the goal of causal inference?
- How do we define a causal effect (in terms of counterfactuals)?
- If X_i is time-varying, we observe Y_i for the same unit i when X_i takes on different values

Causal Inference

- What is the goal of causal inference?
- How do we define a causal effect (in terms of counterfactuals)?
- If X_i is time-varying, we observe Y_i for the same unit i when X_i takes on different values
- Is this the same as observing both $Y_0 it$ and $Y_1 it$?

Causal Inference

- What is the goal of causal inference?
- How do we define a causal effect (in terms of counterfactuals)?
- If X_i is time-varying, we observe Y_i for the same unit i when X_i takes on different values
- Is this the same as observing both $Y_0 it$ and $Y_1 it$?
- Then why are panel data useful?

Nonlinear Panel Models
Examples

Nonlinear Panel Models
Examples

- Binary outcome

Nonlinear Panel Models
Examples

- Binary outcome
- Ordered outcome

Nonlinear Panel Models
Examples

- Binary outcome
- Ordered outcome
- Count outcome

Nonlinear Panel Models
Examples

- Binary outcome
- Ordered outcome
- Count outcome
- Multinomial outcome

Nonlinear Panel Models
Examples

- Binary outcome
- Ordered outcome
- Count outcome
- Multinomial outcome
- Censored

Research Questions

- Form groups of 4
- Generate a research question involving:
  - Binary outcome
  - Ordered outcome
  - Count outcome
- For each type, generate an institutional- and an individual-level question
- So 6 research questions total

Review: Basic Panel Approaches

- Pooled estimator
- Fixed effects estimator
- Random effects estimator

Review: Basic Panel Approaches

- Pooled estimator
- Fixed effects estimator
- Random effects estimator
- We'll focus on binary models first

Estimation Issues

- Cross-sectional OLS models are easy to estimate
- Linear panel models are fairly easy to estimate

Estimation Issues

- Cross-sectional OLS models are easy to estimate
- Linear panel models are fairly easy to estimate
- Cross-sectional GLMs are modestly hard to estimate
    - No closed-form solution
    - Often rely on maximization algorithms

Estimation Issues

- Cross-sectional OLS models are easy to estimate
- Linear panel models are fairly easy to estimate
- Cross-sectional GLMs are modestly hard to estimate
  - No closed-form solution
  - Often rely on maximization algorithms
- Nonlinear panel models are harder to estimate

Who cares?

- If Stata can give us numbers, who cares what's happening?
- More difficult problem means greater diversity of solutions
  - No obvious best solution
  - Terminology overload
  - Assumptions!

Who cares?

- If Stata can give us numbers, who cares what's happening?
- More difficult problem means greater diversity of solutions
  - No obvious best solution
  - Terminology overload
  - Assumptions!
- Be cautious when treading into unfamiliar waters!

Terms You Might See

- Quadrature
- Conditional Likelihood
- Simulated Likelihood
- Generalized Estimating Equation (GEE)
- Generalized Method of Moments (GMM)

Pooled Estimator

- $y_{it} = \beta_0 + \beta_1 x_{it} + \cdots + \epsilon_{it}$
- Ignores panel structure (interdependence)
- Ignores heterogeneity between units
- But, we can actually easily estimate and interpret this model!
- Estimation uses “generalized estimating equations” (GEE)
- Note: Also called *population-averaged* model

Pooled Estimator

- Continuous outcomes:
  $y_{it} = \beta_0 + \beta_1 x_{it} + \cdots + \epsilon_{it}$
- Binary outcomes:
  $y_{it}* = \beta_0 + \beta_1 x_{it} + \cdots + \epsilon_{it}$
  $y_{it} = 1$ if $y_{it}* > 0$, and 0 otherwise
- Link functions are the same in panel as in cross-sectional
  - Logit
  - Probit
- Use clustered standard errors

Respecting the Panel Structure

- With a panel structure, ϵ_{it} can be decomposed into two parts:
  - υ_{it}
  - u_i
- If we assume u_i is unrelated to X: fixed effects
- If we allow a correlation: random effects

Fixed Effects Estimator

- This gives us:

$$\begin{aligned} y_{it} &= \beta_0 + \beta_1 x_{it} + \cdots + \upsilon_{it} + u_i \\ y_{it} &= \beta_{0i} d_{it} + \beta_1 x_{it} + \cdots + \upsilon_{it} \end{aligned} \tag{1}$$

- Varying intercepts (one for each unit)
- Can generalize to other specifications (e.g., fixed period effects)

Fixed Effects Estimator

- Fixed effects terms absorb all time-invariant between-unit heterogeneity
- Effects of time-invariant variables cannot be estimated
- Each unit is its own control ("within" estimation)
- Two ways to estimate this:
  - Unconditional maximum likelihood
  - Conditional maximum likelihood
- Both are problematic

Fixed Effects Estimator

- Unconditional maximum likelihood
  - From OLS: dummy variables for each unit
  - Number of parameters to estimate increases with sample size
  - For logit/probit: *incidental parameters problem*
  - Estimate become inconsistent

- Conditional maximum likelihood
  - From OLS: "De-meaned" data to avoid estimating unit-specific intercepts
  - For logit: condition on $Pr(Y_i = 1)$ across all t periods
  - Does not work for probit!

Conditional MLE

- Estimates only based on units that change in Y
- Effects of time-invariant variables are not estimable
- Observations with time-invariant outcome are dropped

Conditional MLE

- Estimates only based on units that change in Y
- Effects of time-invariant variables are not estimable
- Observations with time-invariant outcome are dropped
- Estimation of two-wave panel using fixed-effects logistic regression is same as a pooled logistic regression where the outcome is direction of change regressed on time-differenced explanatory variables

Fixed Effects Estimator

- Interpretation is difficult
- Use `predict` to get fitted values on the latent scale
- `margins, dydx()` is also problematic
  - Use `, predict(xb)` to obtain log-odds marginal effects
  - Use `, predict(pu0)` to assume fixed effect is zero
  - Neither of those is the default

Questions?

Random Effects Estimator

- If we are willing to assume that unit-specific error term is uncorrelated with other variables
- Why might this not be the case?

Random Effects Estimator

- If we are willing to assume that unit-specific error term is uncorrelated with other variables
- Why might this not be the case?
- Pooled estimator also makes this assumption
- But that estimator ignores panel structure (non-independence)

Estimation hell!

Estimation hell!

- Due to *incidental parameters problem* we cannot consistently estimate both the regression coefficients and the unit-specific effects

Estimation hell!

- Due to *incidental parameters problem* we cannot consistently estimate both the regression coefficients and the unit-specific effects
- We have to make some assumptions about the unit-specific error terms
- But assumptions get us to a likelihood function that can only be maximized via *integration* of a complicated function
- Quadrature (a form of numerical approximation of an integral) is therefore used (costly!)

Random Effects Estimator

- Can be used with logit or probit
- Interpretation is messy because unit-specific error terms are unobserved
- Thus marginal effects calculation must make an assumption of about the random effects:
    - Predict log-odds: `margins, dydx(*)`
    - Assume they are 0: `, predict(pu0)`

Random versus Fixed Effects

- Different assumptions
- Very different estimation strategies
  - These are consequential for interpretation
- Use Hausman test to decide between estimators:

```
xtlogit ..., fe
estimates store fixed
xtlogit ..., re
estimates store random
hausman fixed random
```

- Use FE if H_0 rejected

Reminder!

- Some outcomes are binary but are constant before and after an "event"
  - Individual graduates from university
  - Country transitions to democracy
- We can analyze these using binary outcome panel models *or* using event-history methods from last week
- Either might be appropriate, depending on the research question, hypothesis, and data

Questions about Binary Models?

Example: Wawro

- Form groups of three
- Discuss:
  - What is the research question?
  - What is the method used?
  - What are the results?

Ordered Outcome Models

- Estimators exist, but only random effects is implemented in Stata
  - Logit and probit available
- Other possible analysis strategies:
  - Use a linear panel specification (`xtreg`)
  - Estimate a pooled model (`ologit`/`oprobit`) with clustered SEs
  - Recode categories to binary and use `xtlogit`
  - Use a mixed effects specification (`meologit`/`meoprobit`)

Count Outcome Models

- Count outcome models are somewhat easier to estimate than binary outcome models
- Still have pooled, fixed effects, and random effects strategies
- As in cross-sectional data, prefer negative binomial regression over Poisson regression when there is *overdispersion*
- Methods using unconditional maximum likelihood (fixed effects) are computationally expensive

Interpreting Count Models

- Predict the linear/latent scale:
  `margins, predict(xb)`
- Predict outcomes, assuming fixed/random effect is zero:
  `margins, predict(nu0)`
- With RE, assuming random effect is zero:
  `margins, predict(pr0(`n`))`, where n is number of events

Interpreting Count Models

- Coefficients can be translated into *incidence rate ratios* using `, irr` option in Stata

- This is sort of like the odds-ratio interpretation for binary outcome models

- Meaning: a unit change in x produces a change in the incidence rate for the outcome
  - If $IRR > 1$: unit change in x increases rate of y
  - If $IRR < 1$: unit change in x decreases rate of y

- May be helpful, may not. You can choose for yourself.

Example: Seeberg

- Form groups of three
- Discuss:
  - What is the research question?
  - What is the method used?
  - What are the results?

Questions about Count Models?

Standard Errors

- Standard errors can be complicated
- For pooled model, use standard errors clustered by unit
  - `vce(robust)`
  - `vce(cluster id)`
- For random effects, you may want bootstrapped standard errors
- Always check for robustness

Interpretation: Quick Review

- Usual rules don't apply
- Estimation via an MLE variant usually means marginal effects are undefined
- Depending on model specification, predicted values may also be *conditional*
- We have to make further assumptions to create an interpretable quantity of interest from the model

Intepretation: Trade-offs

- Analytic trade-off between model choice and interpretability
- Pooled estimates are interpretable in conventional ways, but use assumptions
  - Ignores panel structure
  - No unobserved confounding/heterogeneity
- Other models are harder to estimate and interpret, but may be more "correct," though:
  - RE assumes heterogeneity is not confounding
  - FE disallows effects of time-invariant variables

Mixed Effects

- We can also estimate mixed effects models for non-linear outcomes
- This works more or less as with linear outcomes
  - Binary: `melogit`, `meprobit`
  - Ordered: `meologit`, `meoprobit`
  - Count: `mepoisson`, `menbreg`
  - Linear: `mixed`
- Estimation and interpretation is similar to hierarchical linear models

Questions about anything?

Where have we been?

- What have we learned in this course?
- What haven't we learned in this course?

What have we learned?

What have we learned?

- Thinking about causality as counterfactuals

What have we learned?

- Thinking about causality as counterfactuals
- How to obtain causal inference from observational data

What have we learned?

- Thinking about causality as counterfactuals
- How to obtain causal inference from observational data
- Analyzing continuous outcome data

What have we learned?

- Thinking about causality as counterfactuals
- How to obtain causal inference from observational data
- Analyzing continuous outcome data
- Analyzing binary, ordered, and count outcome data

What have we learned?

- Thinking about causality as counterfactuals
- How to obtain causal inference from observational data
- Analyzing continuous outcome data
- Analyzing binary, ordered, and count outcome data
- Analyzing event histories

What have we learned?

- Thinking about causality as counterfactuals
- How to obtain causal inference from observational data
- Analyzing continuous outcome data
- Analyzing binary, ordered, and count outcome data
- Analyzing event histories
- Analyzing data over time

What have we learned?

- Thinking about causality as counterfactuals
- How to obtain causal inference from observational data
- Analyzing continuous outcome data
- Analyzing binary, ordered, and count outcome data
- Analyzing event histories
- Analyzing data over time
- Managing complex data structures

What have we learned?

- Thinking about causality as counterfactuals
- How to obtain causal inference from observational data
- Analyzing continuous outcome data
- Analyzing binary, ordered, and count outcome data
- Analyzing event histories
- Analyzing data over time
- Managing complex data structures
- Data interpretation!

What should I learn next?

What should I learn next?

- Measurement: factor analysis, principal components, IRT

What should I learn next?

- Measurement: factor analysis, principal components, IRT
- Design: surveys, experiments, data gathering

What should I learn next?

- Measurement: factor analysis, principal components, IRT
- Design: surveys, experiments, data gathering
- Classification: regression trees, classifiers, SVM

What should I learn next?

- Measurement: factor analysis, principal components, IRT
- Design: surveys, experiments, data gathering
- Classification: regression trees, classifiers, SVM
- Clustering: K-means, hierarchical clustering

What should I learn next?

- Measurement: factor analysis, principal components, IRT
- Design: surveys, experiments, data gathering
- Classification: regression trees, classifiers, SVM
- Clustering: K-means, hierarchical clustering
- Nonparametric statistics

What should I learn next?

- Measurement: factor analysis, principal components, IRT
- Design: surveys, experiments, data gathering
- Classification: regression trees, classifiers, SVM
- Clustering: K-means, hierarchical clustering
- Nonparametric statistics
- Bayesian statistics

What should I learn next?

- Measurement: factor analysis, principal components, IRT
- Design: surveys, experiments, data gathering
- Classification: regression trees, classifiers, SVM
- Clustering: K-means, hierarchical clustering
- Nonparametric statistics
- Bayesian statistics
- Time series analysis

What should I learn next?

- Measurement: factor analysis, principal components, IRT
- Design: surveys, experiments, data gathering
- Classification: regression trees, classifiers, SVM
- Clustering: K-means, hierarchical clustering
- Nonparametric statistics
- Bayesian statistics
- Time series analysis
- Data visualization

What should I learn next?

- Measurement: factor analysis, principal components, IRT
- Design: surveys, experiments, data gathering
- Classification: regression trees, classifiers, SVM
- Clustering: K-means, hierarchical clustering
- Nonparametric statistics
- Bayesian statistics
- Time series analysis
- Data visualization
- "Big data"

Goals for this course

Goals for this course

- Describe politically relevant research questions and hypotheses

Goals for this course

- Describe politically relevant research questions and hypotheses
- Evaluate and deduce observable implications from political science theories

Goals for this course

- Describe politically relevant research questions and hypotheses
- Evaluate and deduce observable implications from political science theories
- Explain statistical procedures and their appropriate usages

Goals for this course

- Describe politically relevant research questions and hypotheses
- Evaluate and deduce observable implications from political science theories
- Explain statistical procedures and their appropriate usages
- Apply statistical procedures to relevant research problems

Goals for this course

- Describe politically relevant research questions and hypotheses
- Evaluate and deduce observable implications from political science theories
- Explain statistical procedures and their appropriate usages
- Apply statistical procedures to relevant research problems
- Synthesize results from statistical analyses into well-written and well-structured essays

Goals for this course

- Describe politically relevant research questions and hypotheses
- Evaluate and deduce observable implications from political science theories
- Explain statistical procedures and their appropriate usages
- Apply statistical procedures to relevant research problems
- Synthesize results from statistical analyses into well-written and well-structured essays
- Demonstrate how to use Stata for statistical analysis

Exam

- Standard 7-day home assignment
- We will give you a question and data
- You write an essay that answers that question
- To do well:
  - Understand your analysis
  - Justify your analysis
  - Interpret your analysis
- Exam allows for considerable flexibility

Questions?

Course Evaluations

- What went well in this course?
- What would you like to have gone differently?

Course Evaluations

- What went well in this course?
- What would you like to have gone differently?
- http://www.survey-xact.dk/LinkCollector?key=YAV25A9Q359N

Preview

- Tomorrow: More panel GLMs in Stata
- Next week:
  - Optional Q/A Session (14:15–15:00)
  - In this room
  - Readings test your knowledge on complex articles
- PhD Students: meet here next week at 15:00