Analysis of Economics Data
Chapter 17: Special Topics

© A. Colin Cameron
Univ. of Calif. Davis

November 2022

CHAPTER 17: Special Topics

- Consider special issues that arise with
  - cross-section data
  - panel data
  - time series data.
- Consider the most commonly used nonlinear regression models.
- Provide a brief discussion of various methods to allow causal inference given observational data.
- Time series presentation is very dense - just provide a list of topics here.

Outline

Datasets: EARNINGS_COMPLETE, NBA, INTERESTRATES

17.1 Cross-section Data

- If **independent errors** use heteroskedastic-robust standard errors.
- If **clustered errors** with individual i in cluster g we have
  - $y_{ig} = \beta_1 + \beta_2 x_{2ig} + \cdots + \beta_k x_{kig} + u_{ig}$, $g = 1, .., G$.
- For OLS use cluster-robust standard errors where cluster on g
  - or use the following alternative estimation methods.
- Cluster-specific **random effects estimator** models the error as
  - $u_{ig} = \alpha_g + \varepsilon_{ig}$ where $\alpha_g \sim (0, \sigma^2_\alpha)$ and $\varepsilon_{ig} \sim (0, \sigma^2_\varepsilon)$
  - advantage: FGLS in this model could be more efficient than OLS.
- Cluster-specific **fixed effects estimator** again models the error as
  - $u_{ig} = \alpha_g + u_{ig}$ but treat α_g as an individuals-specific fixed effect
  - can eliminate α_g and consistently estimate β's by OLS in model $y_{ig} - \bar{y}_g = \beta_2(x_{2ig} - \bar{x}_{2g}) + \cdots + \beta_k(x_{kig} - \bar{x}_{kg}) + (\varepsilon_{ig} - \bar{\varepsilon}_g)$
  - advantage: allows regressors to be correlated with α_g so inconsistency only arises if regressors correlated with the ε_{ig} component of the error.

17.2 Panel Data

- Now have data for individual i in years $t = 1, \ldots T$
  - $y_{it} = \beta_1 + \beta_2 x_{2it} + \cdots + \beta_k x_{kit} + u_{it},\ i = 1, ..., n$
- Use OLS with cluster-robust standard errors where cluster on i
  - or use the following alternative estimators.
- **Random effects** and related models
  - estimate by FGLS after specifying a model for the correlation over time for a given individual in the error u_{it}.
- Individual-specific **fixed effects** for individual i in cluster g
  - now specify $y_{it} = \alpha_i + \beta_1 + \beta_2 x_{2it} + \cdots + \beta_k x_{kit} + u_{it}$
  - can eliminate α_i and consistently estimate β's by OLS in model $y_{it} - \bar{y}_i = \beta_2 (x_{2it} - \bar{x}_{2i}) + \cdots + \beta_k (x_{kit} - \bar{x}_{ki}) + (\varepsilon_{it} - \bar{\varepsilon}_t)$
  - advantage: allows regressors to be correlated with α_t so inconsistency only arises if regressors correlated with ε_{it}.
- **Dynamic models** allow lagged y_{it}'s as regressors
  - more complicated.

Panel Data (continued)

- With panel data, variables potentially vary over both time and individuals.
- This variation for a variable z_{it} can be decomposed as follows.
- **Total variation** is the variation of z_{it} around the overall mean $\bar{z} = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} z_{it}$.
- **Within variation** is variation over time for a given individual, the variation of z_{it} around the individual mean $\bar{z}_i = \frac{1}{T}\sum_{t=1}^{T} z_{it}$.
- **Between variation** is variation across individuals, the variation of the individual mean $\bar{z}_i$ around the overall mean $\bar{z}$.
- The corresponding **decomposition** for the overall variance is

$$\begin{aligned}
\text{Within variance:} \quad & s^2_{\text{W}} = \frac{1}{nT-1}\sum_i\sum_t (z_{it}-\bar{z}_i)^2 \\
\text{Between variance:} \quad & s^2_{\text{B}} = \frac{1}{n-1}\sum_i (\bar{z}_i-\bar{z})^2 \\
\text{Overall variance:} \quad & s^2_{\text{O}} = \frac{1}{nT-1}\sum_i\sum_t (z_{it}-\bar{z})^2.
\end{aligned}$$

- OLS (and random effects) use both within and between variation.
- The fixed effects estimator uses only within variation.

17.3 Panel Data Example: NBA Team Revenue

- Dataset NBA has data on 29 teams for the 10 seasons 2001-02 to 2010-11
  - view as short panel dataset (T fixed and n large).

			Standard deviation		
Variable	Definition	Mean	Overall	Between	Within
Revenue	Team revenue in 1999 $millions	95.714	24.442	22.467	10.319
Lnrevenue	Natural logarithm of team revenue	4.532	0.236	0.213	0.108
Wins	Number of wins including playoff	41.04	12.438	7.044	10.356
Playoff	=1 if made playoffs in prev.season	0.545	0.499	0.243	0.439
Champ	=1 if champion in previous season	0.035	0.184	0.094	0.159
Allstars	Number of players voted Allstars	0.860	0.871	0.524	0.704
Lncitypop	Log of city population in millions	1.301	0.801	0.807	0.097
Teamid	Team identifier	14.86	8.355	8.517	0.000
Season	Season identifier	5.54	2.872	0.371	2.858

Panel Data Example: NBA Team Revenue (cont.)

- Log-linear model: dependent variable is natural logarithm of team revenue.

Variable	Estimator, coefficients and standard errors					
	Pooled OLS		Random Effects		Fixed Effects	
	Het-robust	Robust	Default	Robust	Default	Robust
Wins	.0049***	.0049***	.0024***	.0024***	.0027***	.0027***
	(.0014)	(.0015)	(.0008)	(.0008)	(.0007)	(.0007)
Season	.0180***	.0180***	.0188***	.0188***	.0200***	.0200***
	(.0035)	(.0033)	(.0017)	(.0033)	(.0017)	(.0029)
Playoff	.0306	.0306	.0385**	.0385*	.0362**	.0362*
	(.0359)	(.0447)	(.0176)	(.0200)	(.0167)	(.0209)
Champion	.1089***	.1089***	.0118	.0118	.0052	.0052
	(.0331)	(.0473)	(.0316)	(.0163)	(.0300)	(.0167)
Allstars	.0353***	.0353*	.0372***	.0372***	.0356***	.0356***
	(.0127)	(.0178)	(.0075)	(.0066)	(.0071)	(.0068)
Lncitypop	.1440***	.1440***	.0196	.0196	-.2021***	-.2021***
	(.0196)	(.0598)	(.0315)	(.0872)	(.0491)	(.0632)
Intercept	3.9945	3.9945	4.2477	4.2477	4.5222	4.5222
	(.0491)	(.0596)	(.0560)	(.1076)	(.0649)	(.0957)
Observations	286	286	286	286	286	286

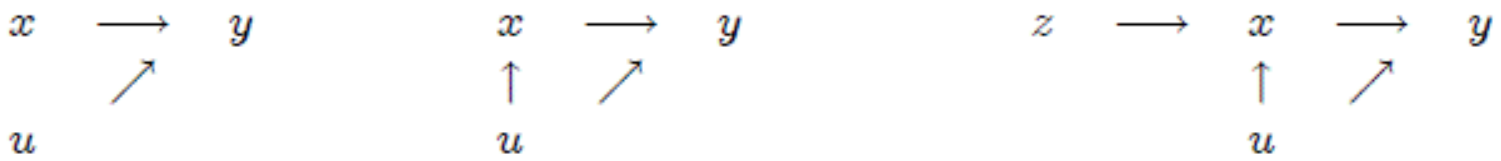
Panel Data Example: NBA Team Revenue (cont.)

- Pooled OLS, random effects and fixed effects estimators
  - use cluster-robust standard errors, not the default
  - asterisks: single for 10%, double for 5%, triple for 1%.
- The fixed effects slope estimate of 0.0027 means that one more win per season is associated with a 0.27% increase in team revenue, after controlling for city characteristics, some immediate past performance measures, and unobserved team characteristics $(\alpha_i m)$ that are time invariant.

17.4 Instrumental Variables

- Problem: in model $y = \beta_1 + \beta_2 x + u$ we have $\mathrm{E}[u|x] = 0$
  - then x is endogenous and OLS is inconsistent.
- Solution: assume there exists an instrument z that
  - does not belong in the model for y (exclusion restriction)
  - is correlated with x.

1. OLS consistent **2. OLS inconsistent** **3. IV consistent**



- Note: This is only possible if one can find a valid instrument.

Instrumental Variables (continued)

- The **instrumental variables (IV) estimator** of β_2 is

$$b_{IV} = \frac{\sum_i (z_i - \bar{z})(y_i - \bar{y})}{\sum_i (z_i - \bar{z})(x_i - \bar{x})}$$

- Intuitively it estimates $\frac{\Delta y}{\Delta x}$ as the ratio $\frac{\Delta y}{\Delta z} / \frac{\Delta z}{\Delta x}$
  - if one-unit change in z is associated with a one unit increase in x of 2 and increase of y of 3 then $b_{IV} = 3/2 = 1.5$.
  - and this can be given a causal interpretation of $\frac{\Delta y}{\Delta x} = 1.5$.
- Can extend to multiple regression
  - exogenous regressors (uncorrelated with u) are instruments for themselves
  - if more instruments (z) than endogenous regressors (x) then use two-stage least squares.
- Example: in log-wage model treat schooling as endogenous
  - use distance to closest college as an instrument.

17.5 Causal Inference: An Overview

- **Causal inference**
  - goal is to get causal estimate of effect of x on z using observational data.
- Stereotypical problem is returns to training where self-select into training
  - $y_i = \beta_1 + \gamma d_i + u_i$ where d_i is a binary indicator for training
  - people choose to get training and we expect that those with higher (unobserved) expected benefits to training will select training
  - then $\mathrm{E}[u_i|d_i = 1] > \mathrm{E}[u_i|d_i = 0]$ so $\mathrm{E}[u_i|d_i] \neq 0$.
- There are many different methods to nonetheless obtain a causal estimate
  - each method has its own distinct assumptions and data requirements.

Causal Inference: Potential Outcomes Model

- **Potential outcomes model** or Rubin causal model
  - standard framework that is used.
- Consider a binary treatment D
  - $D_i = 1$ for individual i if treated
  - $D_i = 0$ if individual i is not treated (a control).
- There are two potential outcomes for Y_i
  - Y_{1i} if $D_i = 1$ and Y_{0i} if $D_i = 0$.
- Interest lies in estimating the treatment affect $\gamma_i \equiv Y_{1i} - Y_{0i}$
  - we cannot estimate γ_i as we only observe one of Y_{1i} and Y_{0i}
  - so restrict attention to more aggregated measures.
- The **average treatment effect** (ATE) in the population is

$$\text{ATE} = E[\gamma_i] = E[Y_{1i} - Y_{0i}].$$

- The **average treatment effect on the treated** (ATET) is

$$\text{ATET} = E[\gamma_i | D_i = 1] = E[(Y_{1i} - Y_{0i}) | D_i = 1].$$

Causal Inference: Differences-in-Differences

- A **random controlled trial** (RCT) is an **experiment** where randomly assign people to treatment and control.
  - then estimate ATE by the difference in means $\bar{y}_{1i} - \bar{y}_{0i}$
  - done more often in economics but still not a lot.
- A **difference-in-difference** estimate uses the following
  - simplest case two periods of time
  - no individuals are treated in the first period
  - some are treated in the second period and some are not
  - $\widehat{\text{ATET}}$ = average change in y over time for those treated in second period
    minus average change in y over time for those not treated in second period.
  - example of a natural experiment.

Causal Inference: Regression Adjustment

- **Control function approach** adds controls
  - $\widehat{\text{ATE}} = \widehat{\gamma}$ from OLS of $y_i = \beta_1 + \gamma d_i + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + u_i$
  - need to ensure d_i and u_i uncorrelated once the controls are added.
- Richer **regression adjustment** estimator runs separate regressions
  - regress y_i on intercept and $x_{2i}, ..., x_{ki}$ for $d_i = 1$ only compute $\frac{1}{n}\sum_{i=1}^{n} \widehat{y}_{1i}$ where $\widehat{y}_{1i}$ is resulting prediction for $d_i = 0$ and $d_i = 1$.
  - regress y_i on intercept and $x_{2i}, ..., x_{ki}$ for $d_i = 0$ only compute $\frac{1}{n}\sum_{i=1}^{n} \widehat{y}_{0i}$ where $\widehat{y}_{1i}$ is resulting prediction for $d_i = 0$ and $d_i = 1$.
  - $\widehat{\text{ATE}} = \frac{1}{N}\sum_{i=1}^{N} \widehat{y}_{1i} - \frac{1}{N}\sum_{i=1}^{N} \widehat{y}_{0i}$.
- **Fixed effects** estimators
  - $y_{it} = \beta_1 + \gamma d_{it} + \beta_2 x_{2it} + \cdots + \beta_k x_{kit} + \alpha_i + \varepsilon_{it}$
  - need to assume d_{it} and ε_{it} uncorrelated once the controls and α_i are added.

Causal Inference: Regression Discontinuity Design

- **Regression discontinuity design (RDD)**
  - a threshold variable determines treatment status
  - e.g. admission into treatment is based on a score denoted s, with scores above 100, say, leading to treatment $(d = 1)$.
- A simple RDD estimate compares the average value of y for individuals on either side of the threshold.
- Complication: usually the outcome variable y itself varies with s
  - suppose that $y = \beta_1 + \beta_2 s + u$ without treatment
  - then a simple RDD estimate of ATET is $\widehat{\gamma}$ from OLS of
    - ★ $y_i = \beta_1 + \gamma d_i + \beta_2 s_i + u_i$.
- In practice more flexible models are used
  - e.g. different linear or quadratic trends on either side of the threshold
  - estimates are focused on observations close to the threshold
  - or nonparametric methods are used either side of the threshold.

Causal Inference: Local Average Treatment Effects

- Instrumental variables (IV) estimator from chapter 17.4
  - IV estimator in model $y_i = \beta_1 + \gamma d_i$ where z_i is instrument for x_i.
- This restricts constant treatment effect γ for all individuals.
- Instead allow different (**heterogeneous**) treatment effects γ_i.
- Specialize to a **binary treatment** D and suppose for simplicity that higher value of Z makes selection into treatment $(D = 1)$ more likely.
- Distinguish between four types of people:
  - (1) **Always-takers** chose treatment $(D = 1)$ regardless of the value of Z
  - (2) **Never-takers** never chose treatment $(D = 0)$ regardless of the value of Z;
  - (3) **Compliers** are induced into treatment so $D = 1$ when $Z = 1$ and $D = 0$ when $Z = 0$
  - (4) **Defiers** are induced away from treatment so $D = 0$ when $Z = 1$ and $D = 1$ when $Z = 0$.
- Then under the crucial and nontestable assumption that there are no defiers, also called the monotonicity assumption, the IV estimator

Causal Inference: IPW and Matching

- **Inverse probability weighting** uses weighted averages of the outcome
  - binary treatment $d_i = 1$ or $d_i = 0$.
  - we observe $d_i y_i$ if the individual is treated and $(1 - d_i)y_i$ if untreated
  - ATE is the weighted average $\frac{1}{N}\sum_{i=1}^{n} w_i\{d_i y_i - (1 - d_i)y_i\}$
    - ★ where the weights $w_i = 1/\widehat{p}_i$ if treated and $w_i = 1/(1 - \widehat{p}_i)$ if untreated
    - ★ and $\widehat{p}_i = \widehat{\Pr}(d_i = 1|(x_{2i}, ..., x_{ki})$ is the **propensity score**, the predicted probability of treatment.
  - key: assume that the weights control for selection into treatment.
- **Matching** compares treated person to a similar (on x's) untreated
  - **nearest neighbor matching** compare outcome for each treated observation to the average outcome of the k observations whose values of $x_2, ..., x_k$ are closest to those for the treated observation.
  - **propensity score matching** instead compares outcomes with similar probability of treatment..

17.6 Nonlinear Regression Models

- Binary outcome $y_i = 0$ or 1
  - model $\Pr[y_i = 1|\text{regressors}]$ using a logit model or probit model
  - maximum likelihood estimation by computer is straightforward, interpretation of estimates is more difficult.
- The **logit model** specifies that

$$\begin{aligned}\Pr[y = 1|x_2, ..., x_k] &= \frac{\exp(\beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}{1 + \exp(\beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k)} \\ \Pr[y = 0|x_2, ..., x_k] &= 1 - \Pr[y = 1|x_2, ..., x_k].\end{aligned}$$

- For the j^{th} regressor $\text{ME}_j = \frac{\Delta \widehat{p}}{\Delta x_j} = \widehat{p}(1 - \widehat{p})b_j$
  - where $\widehat{p}$ is the predicted probability
  - $\text{ME}_j| \leq 0.25 \times |b_j|$ and sign of b_j gives sign of ME.

Nonlinear Regression Models (continued)

- The **probit model** specifies that

$$\begin{aligned} \Pr[y &= 1|x_2, ..., x_k] = \Phi(\beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k) \\ \Pr[y &= 0|x_2, ..., x_k] = 1 - \Pr[y = 1|x_2, ..., x_k], \end{aligned}$$

  where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

- For the probit model $\text{ME}_j = \phi(\widehat{p})b_j$
  - where $\phi(\cdot)$ is the standard normal density function
  - $\text{ME}_j| \leq 0.4 \times |b_j|$ and sign of b_j gives sign of ME.

Nonlinear Regression Models (continued)

- Suppose the conditional mean is exponential, so that

$$\mathsf{E}[y|x_2, ..., x_k] = \exp(\beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k).$$

- This model is applicable to nonnegative data as $\mathsf{E}[y|x_2, ..., x_k] > 0$.
- Estimation is by a method called **quasi-maximum likelihood**
  - rather than by least squares regression
  - b_1, b_2,...,b_k maximize $\sum_{i=1}^{n}\{y_i \ln \mu_i - \mu_i\}$ where $\mu_i = \exp(b_1 + b_2 x_{2i} + \cdots + b_k x_{ki})$.
- Now $\text{ME}_j = \frac{\Delta \widehat{y}}{\Delta x_j} = \widehat{y} b_j$
  - where $\widehat{y}$ is the predicted value of y
  - and AME$= \bar{y} b_j$.
- Coefficients b_j can be directly interpreted as semi-elasticities.

17.7 Time Series Data

- Topics covered in the text
  - HAC Standard errors
  - stationary process and data transformation
  - sample autocorrelations
  - tests for autocorrelation
  - autoregressive models
  - finite distributed lag models
  - autoregressive distributed lag models
  - autoregressive error models
  - nonstationary time series and unit roots
  - spurious regression
  - regression with nonstationary data
  - forecasting.

17.8 Time Series Data: U.S. Treasury Security Interest Rates

- Dataset INTERESTRATES has monthly data from January 1982 to January 2015 on 1-year and 10-year treasury note constant maturity rates.
- Application regresses 10-year rate on 1-year rate.