what is Econometrics

Example Questions:

1. How does an additional year of school effect earnings
2. Does sunscreen really prevent skin cancer

- Econometrics uses data to learn about Economic relationships (models)

Econometric Process

1. Question
2. Model
3. Data
4. Estimation
5. Inference

- Causal Questions, if I change one thing how does other change

The Model

Y → Dependent Variable

X → Independent Variable

↳ for each question we must determine which is which

Intro to Linear Bivariate Model

say that,

$wage(schooling) = A\, e^{(r \times schooling)}$

$\ln(wage) = \ln[A] + r * schooling$

$\underbrace{\ln(wage)}_{y_i(x)} = E[\ln A_i] + r * schooling$

$$y_i(x) = \beta_0 + \beta_1 x + \varepsilon_i$$

↳ Linear Bivariate Model

Important: If you can transform relation into linear form you can use OLS

- y_i → Dependent/outcome Variable
- β_0 → Intercept
- β_1 → Slope or Marginal effect
- ε_i → Error term or Disturbance
- x_i → regressor, variable of interest

14.32 Mid term

Empirical Models

- we make the assumption $E[\varepsilon_i | X_i = x] = 0$
- Assumption justified if
  1. X_i is randomly assigned
  2. Economic Theory implies conditional Moment Restriction
  3. * control variables related to X unrelated to y

why Matters

* On average, People w/ different X dont have different ε

* Proof ①, Assumption 2 *

Data

1. Cross-Section → 1 time, many individuals
2. Time series → one thing, many times
3. Panel Data → Many time Periods, Many individuals

↳ we assume each observation is an i.i.d Draw from some joint Distribution F(x,y)

Simple (Bivariate) Regression

- If we knew (x_i, y_i) of all individuals we'd know exactly β_1

Sample

- we use a random sample as a proxy for Population

Formula for Sample Mean (2)

* Ordinary Least Squares *

- Estimate line that minimizes Average Squared Error

$$\min_{b_0, b_1} \hat{E}\left[(y_i - (b_0 + b_1 x_i))^2\right] \quad (3)$$

$$\hat{b}_1 = \frac{\widehat{Cov}(x,y)}{\widehat{Var}(x)}$$

$$\hat{b}_0 = \bar{Y} - \hat{b}_1 \bar{X}$$

Sample Regression Line

Properties of OLS coeffs

$(\hat{b}_0, \hat{b}_1)$ are Random Variables

↳ it'll take on different values when we consider different samples

Proofs that b's = β's (4)

- An estimate is an RV who's centered about Pop. value

Predicted or fitted value

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x$$

Residual

$$\hat{u}_i = y_i - \hat{y}_i$$

Regression breaks up Y into two parts

$$y_i = \hat{y}_i + \hat{u}_i$$

$\hat{y}_i$ = explained part of x

$\hat{u}_i$ = unexplained part of x

~~Multivariate Regression Analysis~~

Omitted Variable Bias

Derivation of OVB (5)

↳ Factors that influence X & correlate w/ Y cause Bias

control strategies

1. Randomized Controltrials (unethical)
2. Hold other factors fixed (Matching) → decimates sample
3. Multiple Regression

Multiple Regression Analysis

- Putting omitted variables into regression solves problem of bias

$$y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki} + \varepsilon_i$$

$$Y = X\beta + \varepsilon$$

Derivation of $\hat{b}$ Matrix (6)

$$\hat{b} = (X'X)^{-1} X'Y$$

Predicted values:

$$\hat{y} = X\hat{b}$$

Residuals:

$$\hat{u} = y - \hat{y}$$

Properties of Residuals

- Average is zero
- Residual uncorrelated w/ every x

R^2

$$R^2 = 1 - (SSR/SST)$$

↳ Note: will always increase if we add another variable

Alternate Form of OLS COEF

$$\hat{b}_1 = \frac{\widehat{Cov}(y_i, \hat{r}_{1i})}{\widehat{Var}(\hat{r}_{1i})}$$

↳ $\hat{r}_{1i}$ is residual from regressing x_1 on other included variables

Projections

- Regression is a projection of the outcome variable on the column space of X-variables

$$\hat{Y} = X\hat{B}$$

$$= X(X'X)^{-1}X'Y$$

$$= P_x Y$$

$$\hat{U} = Y - P_x Y$$

$$= (I - P_x) Y$$

$$= M_x Y$$

Multiple Regression Ols Form

$$\hat{b}_1 = \frac{\sum_{i=1}^{n} y_i \hat{r}_{1i}}{\sum_{i=1}^{n} \hat{r}_{1i}^2}$$

↳ $\hat{r}_{1i}$ is residual from regressing x_1 on other included variables

$$\sum_{i=1}^{n} \hat{u}_i = 0 \qquad \sum_{i=1}^{n} \hat{u}_i x_i = 0$$

Properties of OLS Estimators

(Gauss-Markov Assumptions)

A0 No perfect collinearity

↳ w/ this assumption $E[\hat{b}|X] \to \beta$

(7) $(X'X)^{-1}$ exists

A1: Conditional expectation = 0

$$E[\varepsilon | x_1, \dots x_k] = 0$$

A2: Homoskedasticity

A3: No serial correlation (8)

$$V(\varepsilon|X) = \sigma^2 I$$

$$V(\hat{b}|X) = \sigma^2 (X'X)^{-1}$$

$$V(\hat{b}_j) = \frac{\sigma^2}{SST_{x_j}(1 - R_j^2)}$$

Gauss-Markov theorem

if A0 → A3 hold true

OLS is BLUE

↳ "Best Linear Unbiased Operator"

Proof of Gauss Markov theorem

"Refer to sheet"

Inference

- we have $\hat{b}_j$, we want to know more about β_j

Hypothesis tests

1. A null Hypothesis H_0
2. Alternative Hypothesis H_1
3. a test statistic, $\hat{T}$
4. A distribution of ~~test~~ statistic if H_0 true

Specific Mechanics of t-test

- we run regression & obtain

$$\hat{b} = (X'X)^{-1} X'Y$$

↳ $(K+1) \times 1$ column vector of estimates

- Say we want to test if $\beta_j = 0$ or 0.25 or -1

Step 1: Null Hypothesis

$$H_0: \beta_j = h_0$$

$$H_1: \beta_j \neq h_0$$

Step 2: test statistic

Given, $\hat{\sigma}^2 = \frac{SSR}{n-(K+1)}$

$$s.e(\hat{b}_j) = \sqrt{\frac{\hat{\sigma}^2}{SST_{x_j}(1 - R_j^2)}}$$

Step 3: t-stat

$$t_{\hat{b}_j} = \frac{\hat{b}_j - h_0}{s.e(\hat{b}_j)}$$

Step 4: Distribution

$$\hat{t}_{\hat{b}_j} \sim t(n-(k+1))$$

↳ Distribution

$$p = F_{t(n-(k+1))}\left(-|t_{\hat{b}_j}|\right) + 1 - F_{t(n-(k+1))}\left(|t|\right)$$

↳ Finally you test

if $p \leq a$ "significance level"

↳ Dont reject if true

↳ reject if False

Note: as $n-(k+1) \geq 0$, .05 critical value is around 2

Confidance intervals

- $\frac{\hat{b}_j - h_0}{s.e(\hat{b}_j)} = \pm c_\alpha$ (critical value of significance α)

$$h_0 = \hat{b}_j \pm c_\alpha (s.e(\hat{b}_j))$$

- Means out of 100 draws, correct value will be contained 95 times out of 100

Test of Hypotheses on linear combo

- what if we want to test the relationship between more than just one type of parameter

Precedure:

step 1

$H_0: \beta_j - \beta_n = 0$

$H_1: \beta_j - \beta_n \neq 0$

Step 2: t-stat

$$\hat{T} = \frac{\hat{\beta}_j - \hat{\beta}_n - 0}{s.e(\hat{b}_j - \hat{b}_n)}$$

Step 3: Distribution

$$\hat{T} \sim t(n-(k+1))$$

Standard error (different)

$$s.e(\hat{b}_j - \hat{b}_n) = \sqrt{\hat{V}(\hat{b}_j - \hat{b}_n)}$$

$$= \sqrt{\hat{v}(\hat{b}_j) + \hat{v}(\hat{b}_n) - 2\,\hat{Cov}(\hat{b}_j, \hat{b}_n)}$$

↳ All pieces can be found in Var-Cov matrix

$$\hat{V}(\hat{b}) = \frac{SSR}{n-(k+1)}(X'X)^{-1}$$

Hypotheses on linear combo: General

$H_0: a_0\beta_0 + a_1\beta_1 + \ldots a_k\beta_k = 0$

↳ $a'\beta = 0$

$$\hat{T} = a'\hat{b} \Big/ \sqrt{a'\hat{V}(\hat{b})a}$$

Testing Multiple Linear Restriction

- t-tests are good for a test of one relationship

F-Test

- To test Multiple Linear relationships we perform an F-Test

Procedure

1. Step 1

$H_0: \beta_1 = 0, \beta_m = 0, \beta_n = 0$

H_1: one or more of parameters is not zero

Step 2 (F-stat)

$$\hat{T} = \frac{(SSR_{restrict} - SSR_{unrestrict})/q}{SSR_{unrestrict}/(n-(k+1))}$$

$$\hat{T} \sim F(q, n-(k+1))$$

Violation of A4: Errors Not Normal

Assumption: Errors Distributed Normally

↳ If normality doesn't hold we still get this condition in large samples

Large Sample Asymptotic OLS Properties

- $\hat{b}_j$ is consistent as sample size gets larger

$$\lim_{n\to\infty} Pr(|\hat{b}_j - \beta_j| > \varepsilon) = 0$$

- $\hat{b}$ is asymptotically normally Distributed (9)

Violation of A2: Heteroskedasticity

Assumption: $Var(\varepsilon_i|X) = \sigma^2$

when it fails: when Error fans out depending on X

Test: Breusch-Pagan Test

steel: run a regression to achieve SSR

step2: Regress SSR on X

step3: F-test that all $\beta_{\varepsilon_i^2|X_i} = 0$ &

call Heteroskedastic if F test fails

Test: white test

↳ also add X_i^2 & X_1X_2

$$\hat{V}_{white}(\hat{b}) = (X'X)^{-1}\left(\sum_{i=1}^{n} x_i x_i' \varepsilon_i^2\right)(X'X)^{-1}$$

Fix: to make OLS Blue use GLS

Violation of A1: Functional form

Proposed Model

$$\ln(w_i) = \beta_0 + \beta_1 educ_i + \beta_2 xp_i + \eta_i$$

true model:

$$\ln(w_i) = \beta_0 + \beta_1 educ_i + \beta_2 xp_i + \beta_3 xp_i^2 + \varepsilon_i$$

- OLS estimate no longer is consistent
  (return is $\beta_2 + 2\beta_3 xp_i$ for experience)
- To test: Throw in Polynomial terms & conduct F-test

Violation of A_3: Serial Correlation

- when outcome of one observation influenced by outcome of another

when can go wrong:

- Time series data
- Panel Data
- Clustering

Consequences:

- OLS still unbiased & consistent
- Standard errors are wrong
- Not efficient

Suppose: $u_t = \rho u_{t-1} + e_t$

Test: Durbin-Watson Test

1. Run original Regression of Y on X
2. calculate $\hat{u}_i$ & regress on $\hat{u}_{i-1}$
3. T-test $\rho = 0$

$Cov(\varepsilon, X) = 0$

causes

1. omitted variables

$$y_i = \alpha + \beta x_i + \gamma w_i + \eta_i$$

↳ if $\gamma \neq 0$, $C(x_i, w_i) \neq 0$

14.32 Final

Simultaneous Equations Model

Problem: $\text{Cov}(X_i, \varepsilon_i) = 0$

↳ possible cause: simultaneity

Demand Function

$$q_{dt} = \alpha + \beta p_t + Z_t^{d\prime} \lambda^d + u_{1t}$$

q_{dt} = quantity of good demanded

p_t = price of good

→ w/ LOG specification: β = demand elasticity

Z_t^d = Demand shifters (something that effects demand but not supply)

Supply Function

$$p_t = \gamma + \delta q_t + Z_t^{s\prime} \lambda^s + u_{2t}$$

→ $1/\delta$ is supply elasticity

→ Z_t^s = Supply shifters

- Equations are called structural Equations
- Coefficients called structural Parameters

Solution: use exogenous shifters

* See equation sheet: Simultaneity Problem

consequence: OLS is biased (upward prob)

Step 1:

→ Solve for endogenous variables in terms of exogenous variables

Reduced Form Equations

$$p_t = \frac{\alpha\delta}{1-\beta\delta} + \frac{\delta}{1-\beta\delta} Z_t^{d\prime}\lambda^d + \frac{1}{1-\beta\delta} Z_t^{s\prime}\lambda^s + \frac{\delta u_{1t} + u_{2t}}{1-\beta\delta}$$

$$q_t = \frac{\alpha + \beta\gamma}{1-\beta\delta} + \frac{1}{1-\beta\delta} Z_t^{d\prime}\lambda^d + \frac{\beta}{1-\beta\delta} Z_t^{s\prime}\lambda^s + \frac{u_{1t} + \beta u_{2t}}{1-\beta\delta}$$

$$p_t = \pi_{10} + Z_t^{d\prime}\pi_{11} + Z_t^{s\prime}\pi_{12} + v_{1t}$$

$$q_t = \pi_{20} + Z_t^{d\prime}\pi_{21} + Z_t^{s\prime}\pi_{22} + v_{2t}$$

- Coefficients → Reduced Form Parameters

$\hat{\pi}_{12} \to$	$\lambda^s \frac{\beta}{1-\beta\delta}$
$\hat{\pi}_{22} \to$	$\lambda^s \frac{1}{1-\beta\delta}$
$\hat{\beta} = \hat{\pi}_{12}/\hat{\pi}_{22} \to$	β

→ Known as method of indirect least squares

Summary

- Supply shifter Z_t^s serves as an instrument to endogenous variable p_t

conditions on Z_t^s

1. Exclusion from demand Equation
2. Has to be relevant to supply equation

$$\hat{\beta} = \frac{\hat{C}(Z_t^s, q_t)}{\hat{C}(Z_t^s, p_t)}$$

Instrumental Variables

Problem: Endogenous explanatory variables

Causes:
- measurement error
- Omitted Variables
- Simultaneity

* Endogeneity of X_i makes the OLS estimators biased & inconsistent

Simple Bi-variate IV setup

Causal Model: $y_i = \beta_0 + \beta_1 X_i + u_i$

Problem: $\text{cov}(X_i, u_i) \neq 0$

what makes a good instrument?

1. Exclusion: $\text{cov}(Z, u) = 0$
2. Relevance: $\text{cov}(Z, X) \neq 0$

- can test "Relevance" on t-test of regression

$$X_i = \pi_{20} + \pi_{22} Z_i + v_{2i}$$

$H_0: \pi_{22} = 0$

- Exclusion just determined by common sense

Identification and Estimation

→ start w/ $\text{cov}(Z, y)$ and solve for $\beta_1 = \text{cov}(Z, y)/\text{cov}(Z, X)$

$$\hat{b}_1^{IV} = \frac{\sum_{i=1}^n (Z_i - \bar{Z})(y_i - \bar{y})}{\sum_{i=1}^n (Z_i - \bar{Z})(X_i - \bar{X})}$$

$$s.e(\hat{b}_1^{IV}) = \frac{\hat{\sigma}^2}{SST_X \cdot R^2_{X,Z}}$$

↳ construct t-tests and confidance intervals as usual

2 Comments

- If X_i not endogenous, and used as instrument for self

$\hat{b}^{IV} \to \hat{b}^{OLS}$

- If X_i not endogenous, but use Z any ways

Still consistent but larger s.e

IV w/ Multiple Regressors

Regression { $X \to$ endogenous; $Z_i \to$ exogenous regressor

$$y_i = \beta_0 + \beta_1 X_i + Z_{1i}'\lambda_1 + \varepsilon_i$$

- Propose Z_2 as indicator

$-\text{cov}[Z_{2i}, Z_{1i}] = 0$ (exclusion)

$-\text{cov}[Z_{2i}, X] \neq 0$ (relevance)

Matrices

$$X = \begin{bmatrix} 1 & X_1 & Z_{11} \\ \vdots & \vdots & \vdots \\ 1 & X_n & Z_{1n} \end{bmatrix}, Z = \begin{bmatrix} 1 & Z_{21} & Z_{11} \\ \vdots & \vdots & \vdots \\ 1 & Z_{2n} & Z_{1n} \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \lambda_1 \end{bmatrix}$$

Notation of Model

Model: $Y = X\beta + \varepsilon$

$X \rightarrow$ matrix including all RHS variables in Model

$Z \rightarrow$ matrix w/ all exogenous variables (included in model or not)

$\hat{b}_{IV} = (Z'X)^{-1}Z'Y$ → see IV estimator Derivation

Identification

- Need: # of RHS Variables = # of Exogenous Variables

# of excluded instruments = # of endogenous variables

(order condition)

Two Stage Least Squares (2SLS)

Reasons to use more instruments than needed

1. More Precise estimation
2. Allows for statistical testing of Assumptions

# of intruments = # of variables $\Rightarrow$ just identified

# of instruments > # of variables $\Rightarrow$ over identified

SETUP

$$X = \begin{bmatrix} x_{1,1} & \cdots & x_{1,p} & x_{1,p+1} & \cdots & x_{1,K} \\ \vdots & & \vdots & \vdots & & \vdots \\ x_{1,n} & \cdots & x_{p,n} & x_{p+1,n} & \cdots & x_{K,n} \end{bmatrix}$$

$$Z = \begin{bmatrix} z_{1,1} & \cdots & z_{r,1} & x_{p+1,1} & \cdots & x_{K,n} \\ \vdots & & \vdots & \vdots & & \vdots \\ z_{1,n} & \cdots & z_{r,n} & x_{p+1,n} & \cdots & x_{K,n} \end{bmatrix}$$

- p endogenous regressors
- r excluded exogenous instruments
- $K-p$ exogenous regressors

Order condition: $r = p$

Computing 2SLS Estimator

- computing simple IV estimator won't work ($Z'X$ not square)
- We must take our $K+r-p$ instrumental variables and combine them to form one instrument for each of the regressors

Exogenous regressors

↳ can act as its own instrument

Endogenous Regressors*

- what combination of exogenous instruments would work as best instrument for any variable?

$$w_i = \pi_0 + \pi_1 z_1 + \cdots + \pi_r z_r + \pi_{p+1} x_{p+1} + \cdots + \pi_K x_K$$

↳ w_i = instrument for x_i

Instrument coefficients are OLS

Conditions to make w_i good

1. $\text{cov}(w_i, \varepsilon) = 0$
2. $\text{cov}(w_i, x) = 0$

} All z's uncorrelated

What set of coefficients results in highest R^2 value between x and z's?

↓

OLS coefficients!

$$w_1 = Z(Z'Z)^{-1}Z'x_1$$

$$w_2 = Z(Z'Z)^{-1}Z'x_2$$

Putting it together

$$W = Z(Z'Z)^{-1}Z'X = \hat{X}$$

- Now that we have an instrument for each x

$$\hat{\beta}_{2SLS} = (\hat{X}'\hat{X})^{-1}\hat{X}'Y$$

or $= (X'P_Z X)^{-1}X'P_Z Y$

↳ $P_Z = Z(Z'Z)^{-1}Z'$

Reasons why its called 2SLS

1. Regress endogenous variables on ALL exogenous instruments
2. Regress y on predicted values from first stage

Proof: simple IV = Just identified

Testing Endogeneity

- Only want to use IV if we have to, will be less accurate

test of endogeneity

step1: Regress Potentially Endogenous variables on all exogenous variables

calculate residuals $\hat{v}_2$

step2: include residual $\hat{v}_2$ to OLS regression of structural Equation

step3: test if its coefficient is zero

reject → endogenous

Don't Reject → exogenous

Testing Validity of Instruments

1. Hausman Test

↳ requires over-identification

Idea: [a] 2SLS (overidentified)

[b] Just-identified (subset of all possible instruments)

↳ These two should be consistent

$$\hat{T} = \frac{\hat{b}_{\text{Just-identified}} - \hat{b}_{\text{2SLS}}}{S.e(\hat{b}_{\text{JI}} - \hat{b}_{\text{2SLS}})}$$

$H_0: \hat{T} = 0$ — instruments valid

$H_1: \hat{T} \neq 0$ — one or more not valid

2. Omnibus test

[a] estimate via 2SLS to form residuals

[b] Regress $\hat{u}_t$ on all instruments and get R^2

[c] nR_u will follow χ^2 with $r-q-p$ degrees of freedom

Examples of IV Estimation

1. Demand Elasticity

Endogenous: Simultaneity

Instrument: harvest-yield

2. military service on later earnings

Endogenous: high self selection

Instrument: draft-eligibility

3. effect of many districts on student performance

endogenous: wealthier areas have more

exogenous: # of major streams in area

Time Series

- we now look at how outcome variable evolves over time (index by time t)

Model: $\beta_0 + \beta_1 X_{1t} + \dots + \beta_K X_{Kt} + u_t$

$X_t'\beta + u_t, \; t = 1, \dots, n$

Important difference between time series and cross-section

- we allow one observation to depend on past observations

y_t = house price, X_{1t} = int$_t$, X_{2t} = int$_{t-1}$

Dynamic Models — Dependent Data — create problems

Types of time series Models

1. Static Model

$$y_t = \beta_0 + \beta_1 Z_t + u_t$$

2. Finite Distributed Lag (FDL)

$$y_t = \alpha_0 + \delta_0 Z_t + \delta_1 Z_{t-1} + u_t$$

Assumptions for Time Series

A0

same as before, No perfect collinearity

$(X'X)^{-1}$ exists

A1

$E[u_t|X] = 0$, u_t must be uncorrelated across all time periods

↳ strict exogeneity

A2

same as before: Homoskedasticity

A3

same as before: No serial correlation

↳ much more likely to fail

consequences

A0-A2 unbiased

Trends and Seasonality

- Time series can grow over time

↳ model this by including Time in Regression

$$y_t = \alpha_0 + \alpha_1 t + \dots + u_t$$

- Trends also exhibit seasonality

↳ Model this by including dummy variables for months of the year

$$y_t = \alpha_0 + \alpha_1 feb_t + \alpha_2 mar_t + \dots + \alpha_{11} dec_t + u_t$$

Dependent Data & Dynamic Models

- observations from one period often depend on observations from another

- Moving Average Process of order one

$$x_t = e_t + \alpha_1 e_{t-1}$$

- Autoregressive Process AR(1)

$$y_t = \rho_1 y_{t-1} + e_t$$

↳ to be stable, $|\rho_1| < 1$

weak/strong dependence processes

- weakly dependent processes

↳ correlation between two observations dies down as time between them increases

- Strongly dependent (AR(1) where $\rho_1 = 1$)

$$y_t = y_{t-1} + e_t$$

↳ effectively a random walk

$$Var(y_t) = \sigma_e^2 \times t \Rightarrow \infty$$

$$corr(y_t, y_{t+h}) = \sqrt{t/(t+h)} \Rightarrow 1$$

Solution for high persistence processes

- Random Walk w/ Drift

$$y_t = \alpha_0 + y_{t-1} + e_t$$

- Solution: First-Difference (transformed)

$$\Delta y_t = y_t - y_{t-1} = e_t$$

↳ nicely behaved

Panel Data

- Many Observations over Many time Periods
- We label individuals using index i; and time periods using index t
- # of individuals = N
  # of time Periods = T

What Panel is used for?

- Learn about dynamic relationships
- control for time-invariant variables

Panel Data: workhorse Model

$$y_{i,t} = \beta_0 + W_i'\beta_1 + S_{i,t}'\beta_2 + Z_t'\beta_3 + \alpha_i + v_{i,t}$$

W_i → characteristics of individuals that don't change over time

$S_{i,t}$ → variables that change over time across individuals

Z_t → variables that change over time but are the same for all individuals

α_i ** individual effect

Estimation: Panel Models

- What would happen if we ran regression with all variables

↳ if α_i: uncorrelated (consistent & unbiased)

↳ if α_i: correlated (inconsistent & biased)

First Differencing & Fixed effects

- Imagine forming a new variable (year-to-year change in y)

$$\Delta y_{i,t} = y_{i,t} - y_{i,t-1} = \Delta S_{i,t}'\beta_2 + \Delta Z_t'\beta_3 + \Delta v_{i,t}$$

↳ using this we can at least get β_2 and β_3

- Option 2: Run OLS with indicator variable for each individual

$$y_{i,t} = \beta_0 + W_i'\beta_1 + S_{i,t}'\beta_2 + Z_t'\beta_3 + \sum_{i=1}^{N}\alpha_i D_i + v_{i,t}$$

identical to making this transform

$$\tilde{y}_{i,t} = y_{i,t} - \bar{y}_i$$

(Running fixed effects) $= \tilde{S}_{i,t}'\beta_2 + \tilde{Z}_t'\beta_3 + \tilde{v}_{i,t}$

Random Effects

- Assume α_i are uncorrelated with the regressors
- Random Effects

$$\tilde{y}_{i,t} = y_{i,t} - \hat{\lambda}\bar{y}_i$$

$$\tilde{x}_{i,t} = x_{i,t} - \hat{\lambda}\tilde{x}_i$$

$$\hat{\lambda} = 1 - \sqrt{\frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + T\hat{\sigma}_\alpha^2}}$$

- Run Pooled OLS on transformed variables

Important Consequences

1. will be consistent
2. does allow estimation of coefficients on time-invarying variables
3. Proof: Prove $\hat{\beta}_{RE}$ is Blue by proving A3 Holds

Difference in Differences

example: effect of NYC soft drink ban on obesity rates

- People of NYC: Treated$_i$ = 1
- People of Westchester: = 0
- Cross-section & time series alone will fail

Model:

$$\text{Obesity}_{i,t} = \beta_0 + \beta_1 \text{Treated}_i + \beta_2 \text{Post}_t + \delta\,\text{Treated}_i * \text{Post}_t + v_{i,t}$$

$\hat{\delta}$ → Difference in Post/Pre-Difference between NYC & Westchester

Non-linear Models

- suppose we explore age of marriage vs. Probability of Divorce
- y_i = Divorce$_i$ (0 or 1)
  x_i = Age of Marriage

Problem: Probability has to fall between 0 and 1 but linear function of x doesn't have to

- We model effect of X on a Binary Dependent variable as

$$Pr(y_i = 1 | x_i) = G(x_i'\beta)$$

↳ $G(\cdot)$ { link function that lies between zero and one }

Maximum Likelihood Estimation

- Data

obs	Y_i	X_i
1	1	x_1
	0	x_2
	1	
n	1	x_n

- Probability that $y_i = 1$

$$G(x_i'\beta)$$

- Probability that $y_i = 0$

$$1 - G(x_i'\beta)$$

- In general $Pr(y_i = y)$

$$f(y | x_i;\beta) = G(x_i'\beta)^y (1 - G(x_i'\beta))^{1-y}$$

likelihood Function

$$L(Y | X;\beta) = \prod_{i=1}^{n} f(y_i | x_i;\beta)$$

$$Log(L) = \sum_{i=1}^{n} y_i \ln(G(x_i'\beta)) + (1-y_i)\ln(1 - G(x_i'\beta))$$

$$\hat{b}_{MLE} = \arg\max_{\beta} \left(\ln[\mathcal{L}(\beta)] \right)$$

Link Functions:

Probit → $G(\cdot) \rightarrow \Phi(\cdot)$

Logit → $G(\cdot) \rightarrow \frac{\exp(\cdot)}{(1+\exp(\cdot))}$

Quantile Regression

- The Models we've looked at have explored how explanatory variables effect the mean of the outcome variable

Setup

Quantile Function: $Q_Y(\tau) = a : Pr(Y \le a) = \tau$

also, $Q_Y(\tau) = F_y^{-1}(\tau)$

Model

$$Y_i = x_i'\beta(\tau) + \varepsilon_i(\tau)$$

$$Q_{\varepsilon | x}(\tau | x_i) = 0$$

Quantile Regression Estimation

$$\hat{b}_{QR} = \underset{b}{\arg\min} \frac{1}{n} \sum^{n} \rho_\tau(Y_i - x_i' b)$$

↳ ρ_τ is a check function

$$\rho(\varepsilon) = (\tau - 1(\varepsilon < 0))\varepsilon$$