

Linear Regression Model

$$y = X\beta + \varepsilon$$

- $\varepsilon_i \sim \text{iid}(0, \sigma^2) \rightarrow E(\varepsilon) = 0, \text{Var}(\varepsilon) = \sigma^2 \cdot I_{N \times N}$
- $E(X_{ik} \cdot \varepsilon_i) = 0$ for all k
- X has full column rank K

$$\hat{\beta}_{ols} = (X'X)^{-1} X'y, \quad E\left(\hat{\beta}_{ols}\right) = \beta$$

$$\hat{\text{Var}}\left(\hat{\beta}_{ols}\right) = \hat{\sigma}^2 (X'X)^{-1}, \quad \hat{\sigma}^2 = \frac{\hat{\varepsilon}' \hat{\varepsilon}}{n - K} = \frac{RSS}{n - K}$$

$\hat{\beta}_{ols}$ is BLUE

Statistical problems with ε

1. $\text{Var}(\varepsilon) = \text{diag}[\sigma_1^2 \dots \sigma_n^2]$, heteroskedasticity
2. $E(\varepsilon_i \cdot \varepsilon_j) \neq 0$ for $i \neq j$
 - \rightarrow pairwise correlation: Ex. random group effects
 - $E(\varepsilon_i \cdot \varepsilon_{i-s}) \neq 0$, serial correlation
 - Ex. Twins share common unobservables (family effects)
 - Moulton(1986)

Identification problems with X

1. measurement error in X_i
2. omitted variables bias, simultaneity/self-selection
3. incorrect functional form of conditional expectation $X\beta$
 - Ex. Interactions and polynomials
4. $\text{rank}(X'X) < k \rightarrow$ horrible design

Identification problems with β

β constant for all $i \rightarrow$ external validity

β_i varies \rightarrow heterogeneous responses \equiv estimates may be context-specific particularly difficult if individuals select levels of treatment based on β_i (expected gain)

Heteroskedasticity

$$Var(\varepsilon_i) = E(\varepsilon_i^2) = \sigma_i^2, \quad Var(\varepsilon) = \Sigma$$

- $\hat{\beta}_{ols}$ unbiased and inefficient

- $Var(\hat{\beta}_{ols}) = \hat{\sigma}^2 (X'X)^{-1}$ inconsistent

Consistent estimator for $Var(\hat{\beta}_{ols})$: Eicker-White formula

$$\begin{aligned} \hat{Var}(\hat{\beta}_{ols}) &= (X'X)^{-1} X' \hat{\Sigma} X (X'X)^{-1}, \quad \hat{\Sigma} = \text{diag} \left[\hat{\varepsilon}_1^2, \hat{\varepsilon}_2^2, \dots, \hat{\varepsilon}_n^2 \right] \\ &= \left(\sum x_i x_i' \right)^{-1} \left(\sum \hat{\varepsilon}_i^2 x_i x_i' \right) \left(\sum x_i x_i' \right)^{-1} \end{aligned}$$

Eicker-White corrected s.e.'s \equiv square root of diagonal

- t, F tests valid asymptotically
- OLS inefficient

Test for hetero: Does $Var(\varepsilon_i)$ vary over $i \approx$ vary with observables?

i) White, ii) Breusch-Pagan, iii) Goldfeld-Quandt

- essentially gauge correlation of $\hat{\varepsilon}_i^2$ with X_i (and Z_i , etc.)

White test:

- regress $\hat{\varepsilon}_i^2$ on X 's squares and cross products $\rightarrow R^2$
- Under H_0 : No heteroskedasticity
$$nR^2 \xrightarrow{d} \chi^2(q), \quad q = \# \text{ of regressors} - 1 \text{ (constant)}$$
- if $nR^2 > 5\%$ critical value then heteroskedasticity
- calculate White s.e. ($\hat{\beta}_{ols}$)

Breusch-Pagan \equiv Lagrange Multiplier version

- many specification tests have this form
- test depends on knowing variables causing hetero.

Efficient estimation: GLS \equiv WLS

- presumes knowledge of structural form of heteroskedasticity
- benefit: efficiency gain
- cost: inaccurate approximation \rightarrow inconsistent s.e.'s

Makes sense when know “exact” form of hetero.

Ex. Regression based on cell/group means

$$\sigma_j^2 \propto \frac{1}{n_j} \rightarrow \text{reg y x [w = n}_j\text{]}$$

Conclusion: (in STATA) reg y x, robust

Pairwise Correlation (autocorrelation)

$E(\varepsilon_i \cdot \varepsilon_j) \neq 0 \rightarrow$ clustering, random group effects

$E(\varepsilon_i \cdot \varepsilon_s) \neq 0 \rightarrow$ serial correlations

$Var(\varepsilon) = \Sigma$, off-diagonal elements $\neq 0$

- $\hat{\beta}_{ols}$ unbiased and inefficient

- $Var(\hat{\beta}_{ols}) = \hat{\sigma}^2 (X'X)^{-1}$ inconsistent

Generally, $E(\varepsilon_i \cdot \varepsilon_j) > 0 \rightarrow Var(\hat{\beta}_{ols})$ biased down (can be large)

$E(\varepsilon_i \cdot \varepsilon_j) < 0 \rightarrow$ biased up

Approaches

1. Estimate $\hat{\beta}_{ols}$ and correct s.e. ($\hat{\beta}_{ols}$) for clustering

2. Based on estimated form of correlation, estimate efficient $\hat{\beta}_{fgls}$

Case 1: $E(\varepsilon_i \cdot \varepsilon_j) \neq 0$

Ex. y_{is} = wage for i in state s

X_s = worker compensation law in state s

$$y_{is} = \beta X_s + \varepsilon_{is}, \quad \varepsilon_{is} = \alpha_s + u_{is}, \quad u_{is} \sim iid(0, \sigma_u^2)$$

α_s = random state effect, $\alpha_s \sim (0, \sigma_s^2)$, $E(\alpha_s \cdot u_{is}) = 0$

- individuals in same state have “similar unobservables/shocks”

$$E(\varepsilon_{is} \cdot \varepsilon_{js}) = \sigma_s^2 > 0$$

$$E(\varepsilon \varepsilon') = \Sigma = \begin{bmatrix} \ddots & \sigma_s^2 & & & & \\ \sigma_s^2 & \ddots & & & & \\ & & \ddots & \sigma_s^2 & & \\ & & \sigma_s^2 & \ddots & & \\ & & & & \ddots & \sigma_s^2 \\ & & & & \sigma_s^2 & \ddots \end{bmatrix} \equiv \text{Block diagonal with S blocks}$$

$$Var\left(\hat{\beta}_{ols}\right) = (X'X)^{-1} X' \Sigma X (X'X)^{-1}$$

$\hat{\sigma}^2 (X'X)^{-1}$ can drastically understate true $Var\left(\hat{\beta}_{ols}\right)$ – Moulton (1986)

1. run OLS and calculate consistent s.e.’s corrected for “clustering

- estimate $\hat{\sigma}_s^2$ for $s = 1, \dots, 50$ and plug into $\hat{\Sigma}$
- **STATA:** reg y x, cluster(state) does something similar (actually cluster command is more robust).
- consistent and inefficient
- difference between conventional and consistent s.e.’s provide evidence on clustering

2. run OLS using group sample means at the level of the treatment

$$\bar{y}_s = \beta X_s + \bar{\varepsilon}_s, \quad \bar{\varepsilon}_s = \bar{u}_s \sim iid(0, \sigma_u^2)$$

use N_s as weights

consistent and inefficient (more inefficient than OLS with s.e.'s corrected for clustering)

3. Efficient estimation \equiv FGLS

- presumes knowledge on form of clustering

Case 2: Serial correlation – ARMA(p,q)

Durbin-Watson test, Cochrane-Orcutt FGLS

Case 3: Panel Data

$$y_{it} = X_{it}' \beta + \varepsilon_{it}, \quad \varepsilon_{it} = \alpha_i + u_{it}, \quad u_{it} \sim iid(0, \sigma_u^2), \quad \alpha_i \sim (0, \sigma_\alpha^2)$$

$\alpha_i \equiv$ “random effects”

reg y x, cluster(id)

areg y x, absorb(id) cluster(id)

Twins Ex.

$$y_{ij} = X_{ij}' \beta + \varepsilon_{ij}, \quad \varepsilon_{ij} = \alpha_i + u_{ij}, \quad u_{ij} \sim iid(0, \sigma_u^2), \quad \alpha_i \sim (0, \sigma_\alpha^2)$$

$$E(\alpha_i \cdot u_{ij}) = 0$$

i = family, j = 1 (twin 1), 2 (twin 2)

$E(\alpha_i \cdot X_{ij}) = 0 \rightarrow \alpha_i =$ family “random” effect

(may think $E(\alpha_i \cdot X_{ij}) \neq 0$)

- unobservables of twins in same family may be correlated

- cluster on family id when running regression on pooled sample of twins

Power Calculations

Sample size required for precise estimate of treatment effect

Ex. Random assignment of training

$y_i = \text{log-income}$, $\sigma^2 = \text{Var}(y_i)$

Treatment: $\bar{y}_T \xrightarrow{p} \mu_T$, Control: $\bar{y}_C \xrightarrow{p} \mu_C$

$$\hat{ATE} = \bar{y}_T - \bar{y}_C = \Delta y \leq 0.1$$

$$\sigma_{\Delta y}^2 = \frac{\sigma^2}{N_T} + \frac{\sigma^2}{N_C} = \frac{\sigma^2}{Np(1-p)}, \quad p = \frac{N_T}{N_C}, \quad \text{optimal } p^* = 0.5$$

$$\sigma_{\Delta y} = \frac{\sigma}{[Np(1-p)]^{0.5}}, \quad \sigma \approx 0.5 \text{ for log-income}$$

For $p = 0.5$ with random assignment (optimal case)

$N = 400$ for s.e. $(\Delta y) = 0.05$

$N = 625$ for s.e. $(\Delta y) = 0.04$

In general, need $N > 1,000$

Violations such that OLS inconsistent:

$E(X_i \cdot \varepsilon_i) \neq 0$

$$E\left(\hat{\beta}_{ols}\right) = \beta + E\left[(X'X)^{-1}\right]E(X'\varepsilon) \neq \beta \text{ if } E(X'\varepsilon) \neq 0$$

Measurement error/Errors-in-variables

Case 1: M.E. in y

$$y_i^* = X_i' \beta + \varepsilon_i, \text{ observe } y_i = y_i^* + u_i$$

$$y_i = X_i' \beta + (\varepsilon_i + u_i)$$

if $E(u_i \cdot X_i) = 0$, $\hat{\beta}_{ols}$ unbiased but error variance increases (larger s.e.'s)

if $E(u_i \cdot X_i) \neq 0$, $\hat{\beta}_{ols}$ biased

Case 2: M.E. in X

Bivariate EIV:

$$y_i = \gamma \cdot s_i^* + \varepsilon_i, \quad s_i^* = \text{true schooling}$$

observe $s_i = s_i^* + u_i$, $u_i \sim iid(0, \sigma_u^2)$, (u_i, s_i^*) independent

→ Classical M.E.

$$y_i = \gamma \cdot s_i + (\varepsilon_i - \gamma u_i) = \gamma s_i + \tilde{\varepsilon}_i$$

$\hat{\gamma}_{ols}$ biased down \equiv attenuation bias

$$- Cov\left(\tilde{\varepsilon}_i, s_i\right) = Cov(-\gamma u_i, u_i) = -\gamma \sigma_u^2 < 0$$

- find lower correlation between observed schooling and earnings partially due to misreporting in schooling (some variation not due to true variation in treatment).

$$E\left(\hat{\gamma}_{ols}\right) = \gamma - \frac{\gamma \sigma_u^2}{Var(s_i)} = (1 - \lambda)\gamma, \quad \lambda = \frac{\sigma_u^2}{\sigma_s^2} = \frac{\text{Noise}}{\text{Total Variance}} \text{ ratio}$$

$$\lambda = \frac{\sigma_u^2}{\sigma_{s^*}^2 + \sigma_u^2} = \frac{\text{Noise}}{\text{Signal} + \text{Noise}}$$

Ex. $\lambda \approx 0.1 \rightarrow 10\%$ attenuation bias in bivariate regression

Multivariate EIV:

Add X_i 's to regression to control for confounding (O.V.B)

$$y_i = \gamma \cdot s_i^* + X_i' \beta + \varepsilon_i, \quad s_i = s_i^* + u_i, \quad u_i = \text{classical M.E. (also uncorrelated with } X)$$

regress y_i on s_i, X_i

$$E\left(\hat{\gamma}_{ols}\right) = \left(1 - \frac{\lambda}{1 - R_{s,X}^2}\right) \gamma, \quad R_{s,X}^2 = \text{R-squared from } s_i \text{ on } X_i \text{ regression}$$

$R_{s,X}^2 \uparrow \Rightarrow$ attenuation bias \uparrow for fixed λ

- X_i 's correlated with $s_i^* \Rightarrow$ soak up signal in s_i
- If (u_i, X_i) independent, then X_i soaks up no noise variance \Rightarrow exacerbates attenuation bias problem

Ex. X_i = family background (parents' education, income), age, race, etc.

$$R_{s,X}^2 \approx \frac{1}{3} \Rightarrow \text{Bias} = 15\%$$

Ex. Identical twins

$$y_{ij} = \gamma \cdot s_{ij} + \varepsilon_{ij}, \quad \varepsilon_{ij} = \alpha_i + u_{ij}, \quad (\alpha_i, u_{ij}) \text{ independent, etc.}$$

i = family, j = 1, 2 twin

if $E(\alpha_i \cdot s_{ij}) = 0$ then $\alpha_i =$ random effect, not a source of bias

More likely: $E(\alpha_i \cdot X_{ij}) \neq 0 \Rightarrow$ omitted vars. bias in OLS due to unobservable family-level variables

If α_i is "equal" for twins 1 and 2, can difference it out

$$(y_{i1} - y_{i2}) = \gamma(s_{i1} - s_{i2}) + (\varepsilon_{i1} - \varepsilon_{i2}), \quad (\varepsilon_{i1} - \varepsilon_{i2}) = (u_{i1} - u_{i2})$$

- Regression based on 1st differences

Problem: $s_{i1} = s_{i1}^* + v_{i1}$, $s_{i2} = s_{i2}^* + v_{i2}$

- differencing data can exacerbate M.E. problem
- if $v_{ij} \sim$ classical and (v_{i1}, v_{i2}) independent, then difference out a lot of signal but no noise.
- 1st-differences \approx adding J family-specific dummy variables to regression \rightarrow high correlation between true schooling and family dummies.

$r_s =$ within-family correlation in s

$$E\left(\hat{\gamma}_{fd}\right) = \gamma \left(1 - \frac{Var(v)}{[Var(s^*) + Var(v)](1 - r_s)} \right)$$

Ex. $r_s \approx 0.7$, $\lambda = 0.1 \rightarrow$ Bias $\approx 30\%$

Why would we see a lot of variation in s between twins in same family?

- Ashenfelter and Krueger figure
- If identical twins have, in truth, identical education levels, then all measured differences are due to M.E. \rightarrow drastic downward bias

Generally,

- Above result comes from assuming (v_{i1}, v_{i2}) independent
- If $Cov(v_{i1}, v_{i2}) > 0$, first-differencing still exacerbates M.E. bias as long as $Cov(v_{i1}, v_{i2}) < Cov(s_{i1}^*, s_{i2}^*)$.
- If $Cov(v_{i1}, v_{i2}) < 0$, will exacerbate M.E. bias even more than above.

Above highlights trade-off between addressing omitted variables bias versus measurement error bias.

Approaches:

1. Obtain 3rd party estimate of data reliability (noise-total variance ratio) based on validation studies.

2. Fit group averages (so-called WALD estimator)

$$y_{ij} = \gamma \cdot s_{ij}^* + \varepsilon_{ij} \quad s_{ij} = s_{ij}^* + u_{ij}, \quad j = 1, \dots, J \text{ groups}$$

regress \bar{y}_j on \bar{s}_j

- if a lot of observations in each group $j \approx$ "Average out M.E."

3. Instrumental Variables

- Have z correlated with s but uncorrelated with ε_i and u_i

$z \equiv$ instrumental variable (2 birds with 1 stone)

$$y_i = \gamma \cdot s_i^* + \varepsilon_i, \quad s_i = s_i^* + u_i$$

$$y_i = \gamma \cdot s_i + \tilde{\varepsilon}_i$$

$$\hat{\gamma}_{iv} = (z's)^{-1} z'y = \gamma + (z's)^{-1} z' \tilde{\varepsilon} \xrightarrow{p} \gamma$$

$$\hat{\gamma}_{2sls} = \hat{\gamma}_{iv} \text{ when } \# \text{ of } z = \# \text{ of } X$$

$$s_i = \hat{s}_i + \hat{u}_i = \text{signal} + \text{noise} \rightarrow \text{variation in } s \text{ due to } z \text{ is signal}$$

Ex. Multiple measures

Get another report on s_i^* from *independent* source

- ask twin i about twin j 's education level
- ask both husband and wife about each other's weight and height (may question independence of reporting errors)

s_j^k , j = reportee, k = reporter

s_1^1, s_2^2 = self - reports, s_1^2 = sibling 2's report on sibling 1

$$s_1^1 = s_1^* + e_1^1, \quad s_1^2 = s_1^* + e_1^2$$

Assume “Classical” M.E.

i) $(e_1^1, e_1^2) \sim iid(0, \sigma_e^2)$

ii) $Cov(e_1^1, e_1^2) = 0$

- in pooled cross-section regression, can use s_1^2 as IV for s_1^1

- $corr(s_1^1, s_1^2) = \frac{Var(s_1^*)}{[Var(s_1^1) \cdot Var(s_1^2)]^{0.5}} = 1 - \frac{\sigma_e^2}{\sigma_s^2} = \text{Reliability Ratio}$

Ashenfelter and Krueger find 0.88-0.92 reliability

- 1st differences regression

- $(s_1^2 - s_2^1)$ valid IV for $(s_1^1 - s_2^2) \rightarrow$ over-identified

- Invalid if $Cov(e_1^1, e_1^2) \neq 0 \rightarrow$ correlated M.E.’s

- Under certain assumptions on form of correlation of self- and cross-report errors, $(s_1^2 - s_2^2)$ valid IV for $(s_1^1 - s_2^1)$

- Can also use method-of-moments estimation framework since overidentified in some cases

Omitted Variables Bias:

“True” model:

$$y_i = \alpha + \gamma s_i + \theta A_i + \varepsilon_i$$

A_i = ability, unobserved and correlated with s_i and y_i

Estimated model:

$$y_i = \tilde{\alpha} + \tilde{\gamma} s_i + \tilde{\varepsilon}_i$$

$$E\left(s_i \cdot \tilde{\varepsilon}_i\right) \neq 0 \rightarrow \text{bias due to omitted } A_i$$

If $Cov(A_i, s_i) > 0$ and $Cov(A_i, y_i) > 0$ then positive bias

$$E\left(\tilde{\gamma}\right) = \gamma + \theta \cdot \frac{Cov(s, A)}{Var(s)}$$

Auxilliary model: $A_i = \lambda s_i + u_i$

$$\tilde{\gamma} = \gamma + \lambda\theta, \quad \tilde{\varepsilon}_i = \lambda u_i + \varepsilon_i,$$

$$\text{Bias}(\tilde{\gamma}) = \lambda\theta$$

Twins Ex.: Causality issues

Credibility of comparing twins – does first-differencing necessarily reduce/remove OVB?

$$\frac{\text{Cov}(s_{i1} - s_{i2}, \varepsilon_{i1} - \varepsilon_{i2})}{\text{Var}(s_{i1} - s_{i2})} < \frac{\text{Cov}(s_{ij}, \varepsilon_{ij})}{\text{Var}(s_{ij})}$$

Issue: $\text{Var}(s_{i1} - s_{i2}) \ll \text{Var}(s_{ij})$

Is education the only thing that differs between identical twins? If so, why? If not, then there may be bias in the twins first-differences estimate.

Regression analysis \equiv Tool:

For “causal” interpretation need

- i) linearity and additivity (correct functional form)**
- ii) accurate measurement**
- iii) No omitted variables bias**

Solutions? Good research design eliminates competing hypotheses ex-ante (e.g., random assignment).

Freedman \rightarrow Snow cholera quasi-experiment

DiNardo and Pischke \rightarrow “misuse” of OLS in economics

Next Lecture: Selection on observables (program evaluation)

- concerned mostly with violation of i)**