ExpEcon Methods:
Robust SEs, Clustering, Fixed & Random Effects

ECON 8877

P.J. Healy

First version thanks to Han Wang

Updated 2023-11-17

Introduction

Outline

- Homoskedasticity/Heteroskedasticity

Outline

- Homoskedasticity/Heteroskedasticity
- Clustering

Outline

- Homoskedasticity/Heteroskedasticity
- Clustering
- Fixed Effects and Random Effects

Let's start by reviewing the asymptotic results for OLS.

Model
For $i = 1, 2, \ldots, n$,

$$y_i = x_i'\beta + \epsilon_i$$

where y_i and ϵ_i are scalar, and x_i and β are $k \times 1$ column vectors.

Assumptions

(OLS 0) $(y_i, x_i')_{i=1,\ldots,n}$ is an i.i.d. sequence.

(OLS 1) $\mathbb{E}[x_i x_i']$ is finite and nonsingular.

(OLS 2) $\mathbb{E}[x_i \epsilon_i] = 0$.

- $\hat{\beta} - \beta = (\frac{1}{n}\sum_i x_i x_i')^{-1}(\frac{1}{n}\sum_i x_i \epsilon_i) \xrightarrow{p} (\mathbb{E}[x_i x_i'])^{-1}\mathbb{E}[x_i \epsilon_i'] = 0$

By CLT and Slutsky's theorem,
$\sqrt{n}(\hat{\beta} - \beta) \sim N(0, (\mathbb{E}[x_i x_i'])^{-1}\mathbb{E}[x_i x_i' \epsilon_i^2](\mathbb{E}[x_i x_i'])^{-1})$.

- The variance of the estimator $\hat{\beta}$ is $\mathbb{V}(\hat{\beta}) = n^{-1}(\mathbb{E}[x_i x_i'])^{-1}\mathbb{E}[x_i x_i' \epsilon_i^2](\mathbb{E}[x_i x_i'])^{-1}$

Homoskedasticity: $Cov(\epsilon_i, \epsilon_j) = 0$, and $Var(\epsilon_i | x_i) = \sigma^2$.

- Under homoskedasticity, the middle term $\mathbb{E}[x_i x_i' \epsilon_i^2] = \sigma^2 \mathbb{E}[x_i x_i']$. This simplifies our variance:

$$\mathbb{V}(\hat{\beta})_{homoskedasticity} = n^{-1}\sigma^2(\mathbb{E}[x_i x_i'])^{-1}$$

- Feasible estimator:

$$\hat{\mathbb{V}}(\hat{\beta})_{homoskedasticity} = \hat{\sigma}^2 (X'X)^{-1}$$

where $\hat{\sigma}^2 = (n - k - 1)^{-1}\hat{\epsilon}'\hat{\epsilon}$.

Heteroskedasticity: $Cov(\epsilon_i, \epsilon_j) = 0$, and $Var(\epsilon_i|x_i) = \sigma^2(x_i)$.

- Under heteroskedasticity, the (Eicker(1967)-)Huber (1967)- White (1980) robust estimator is

$$\hat{\mathbb{V}}(\hat{\beta})_{HW} = (X'X)^{-1} \sum_i x_i x_i' \hat{\epsilon}_i^2 (X'X)^{-1}$$

Doing this in practice

- What are the options for estimating the variance? (Long & Ervin, 2000)

$$\hat{\mathbb{V}}(\hat{\beta})_{HW} = (X'X)^{-1}\sum_i x_i x_i' \hat{\epsilon}_i^2 (X'X)^{-1} \tag{HC0}$$

$$\hat{\mathbb{V}}(\hat{\beta})_{robust} = (X'X)^{-1}\sum_i \frac{n}{n-k} x_i x_i' \hat{\epsilon}_i^2 (X'X)^{-1} \tag{HC1}$$

$$\hat{\mathbb{V}}(\hat{\beta})_{HC2} = (X'X)^{-1}\sum_i (1-h_{ii})^{-1} x_i x_i' \hat{\epsilon}_i^2 (X'X)^{-1} \tag{HC2}$$

$$\hat{\mathbb{V}}(\hat{\beta})_{HC3} = (X'X)^{-1}\sum_i (1-h_{ii})^{-2} x_i x_i' \hat{\epsilon}_i^2 (X'X)^{-1} \tag{HC3}$$

where h_{ii} is the diagonal element of the "hat matrix" ($X(X'X)^{-1}X'$).
Note: HC1 is just a d.o.f. adjustment to HC0 ($n/(n-k)$)

$\hat{\mathbb{V}}$ determines our confidence intervals. Thus, our size & power

Doing this in practice

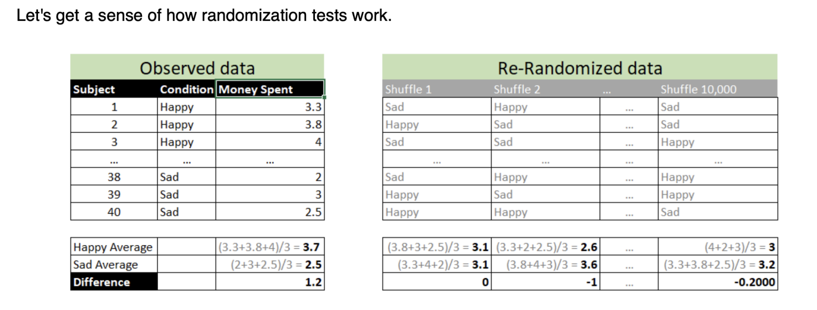
- Let's say $A <= B$ if $B - A$ is PSD (B is more conservative)

  $$HC0 <= HC1 <= HC2 <= HC3$$

  Woodridge ran simulations to show HC2-HC1 is PSD (n=200, k=3, 1,000,000 replications, always true).
- Simulation studies show that HC2 and HC3 lead to better—with small n, possibly much better—confidence intervals than HC1.
- The Stata default with vce(robust) uses HC1.
- The R default with sandwich uses HC3. For R, see estimateR, clubSandwich and Kolesar's github repo.

Related: Young (2019)

- A QJE paper (595 Google cites):
  we need "randomization tests", instead of regressions, to get correct p-values.
- look at 53 experimental papers from the journals of the AEA
- compare randomization tests to conventional tests
  - individual significance results: 13-22 percent fewer
  - joint significance results: 33-49 percent fewer



Let's get a sense of how randomization tests work.

Observed data

Subject	Condition	Money Spent
1	Happy	3.3
2	Happy	3.8
3	Happy	4
...	...	...
38	Sad	2
39	Sad	3
40	Sad	2.5

Happy Average		(3.3+3.8+4)/3 = **3.7**
Sad Average		(2+3+2.5)/3 = **2.5**
Difference		**1.2**

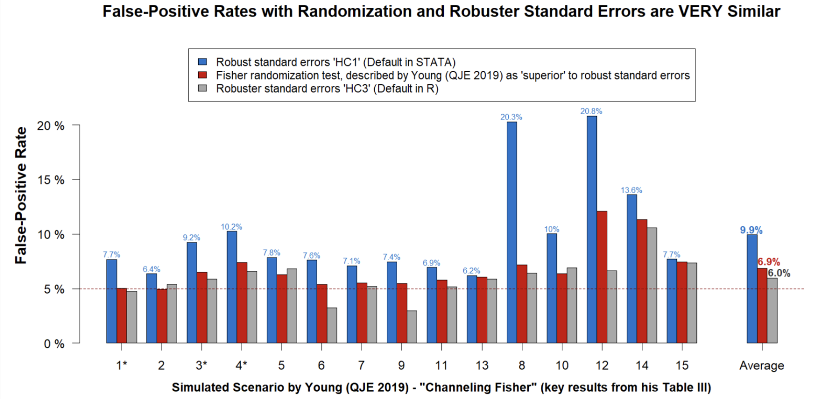
Re-Randomized data

Shuffle 1	Shuffle 2	...	Shuffle 10,000
Sad	Happy	...	Sad
Happy	Sad	...	Sad
Sad	Sad	...	Happy
...	...	...	...
Sad	Happy	...	Happy
Happy	Sad	...	Happy
Happy	Happy	...	Sad

(3.8+3+2.5)/3 = **3.1**	(3.3+2+2.5)/3 = **2.6**	...	(4+2+3)/3 = **3**
(3.3+4+2)/3 = **3.1**	(3.8+4+3)/3 = **3.6**	...	(3.3+3.8+2.5)/3 = **3.2**
0	**-1**	...	**-0.2000**

Related: A post on Data Colada...

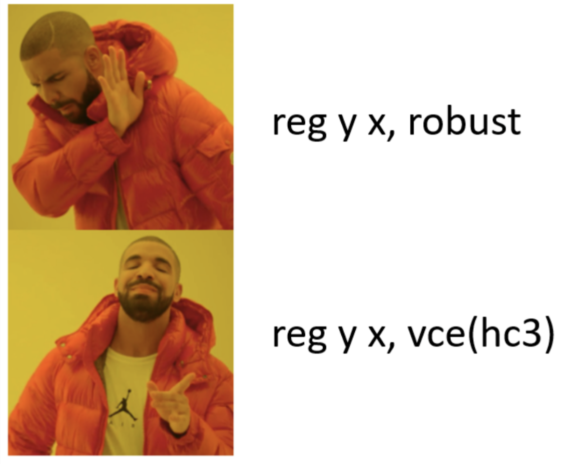
- The QJE study cited used HC1 when comparing with randomization inference in experiments.
- The datacolada post shows that using HC1 and HC3 can be very different when sample sizes are not large.
- But HC3 turns out to work quite well even with pretty small n.



False-Positive Rates with Randomization and Robuster Standard Errors are VERY Similar

Main Takeaway

For Stata users:



```
reg y x, robust
```

```
reg y x, vce(hc3)
```

Clustering and generalizing $\mathbb{E}[\epsilon\epsilon'|X]$

$$\Omega_{homoskedasticity} = \begin{bmatrix} \sigma^2 & & & \\ & \ddots & & \\ & & \ddots & \\ & & & \sigma^2 \end{bmatrix}$$

$$\Omega_{heteroskedasticity} = \begin{bmatrix} \sigma_1^2 & & & \\ & \ddots & & \\ & & \ddots & \\ & & & \sigma_n^2 \end{bmatrix}$$

- We've ignored any correlation structure in Ω.
- In many cases, we don't have that. Instead, Ω has clusters.
  - units are people, and clusters are cities, states or countries
  - units are choices, and clusters are subjects, groups or sessions

Clustering and generalizing $\mathbb{E}[\epsilon\epsilon'|X]$

Let C_i denote unit *i* 's cluster assignment.

- A simple example:

$$\Omega_{ij} = \begin{cases} \sigma^2 & \text{if} \quad i = j \\ \rho\sigma^2 & \text{if} \quad C_i = C_j \;\&\; i \neq j \\ 0 & \text{if} \quad C_i \neq C_j \;\&\; i \neq j \end{cases}$$

$$\Omega_{cluster} = \begin{bmatrix} \sigma^2 & \rho\sigma^2 & & & \\ \rho\sigma^2 & \sigma^2 & & & \\ & & \ddots & & \\ & & & \sigma^2 & \rho\sigma^2 \\ & & & \rho\sigma^2 & \sigma^2 \end{bmatrix}$$

  e.g., if we study individual choices, it might be ok to assume away the correlation between different subjects.

- A more unstructured example: $\Omega_{ij} = \sigma_{ij}$ if $C_i = C_j$.

Let the number of clusters be G, indexed by g

$$\hat{\mathbb{V}}(\hat{\beta})_{LZ} = (X'X)^{-1}\left(\sum_{g} X'_{g,n}\hat{\epsilon}_{g,n}\hat{\epsilon}'_{g,n}X_{g,n}\right)(X'X)^{-1} \text{ (Liang \& Zeger, 1986)}$$

- This makes us think more generally, it's about getting the structure of Ω right. (So better to err on the conservative side)
- However, A recent QJE paper (Abadie, Athey, Imbens & Wooldridge, 2023) argues that this intuition is not correct.

Related: Abadie, Athey, Imbens & Wooldridge (2023)

Misconceptions:

- “The presence of within-cluster correlation implies the need for clustering.”
- “Being as conservative as necessary.”
    - Suppose we want to use the sample average to estimate the population mean. Suppose the population can be partitioned into clusters, e.g., in geographical units. If outcomes are positively correlated in clusters, the cluster variance will be larger than the robust variance.
      But there’s no need to cluster...
- “Researchers have only two choices: to cluster or not to cluster.”

Related: Abadie, Athey, Imbens & Wooldridge (2023)

Main Takeaways:

- “The decision on when and how to cluster standard errors depends on the nature of the sampling and the assignment processes only, not on the presence of within-cluster error components in the outcome variable.”
- The traditional advice of being as conservative as necessary is likely misguided.
- They suggest new ways to estimate variance: causal cluster variance (CCV) and two-stage cluster bootstrap (TSCB).
  - These are designed for applications with large number of observations and substantial variation in treatment assignment within clusters.
- Fixed effects do NOT remove need for clustering.

Doing this in practice

- There are ongoing debates on clustering...
- If we know the appropriate cluster level, we can implement this using the cluster command in Stata:

$$reg\ y\ x, cluster(g)$$

For experimentalists: Cluster by

- Subject?
- Session?
- Other??

Fixed Effects vs Random Effects

$$y_{it} = x_{it}'\beta + u_i + e_{it}$$

where y_{it}, u_i and e_{it} are scalar, and x_{it} and β are $k \times 1$ column vectors.

Key Difference:

- Random effects: u_i is part of the error.
  Need to assume no correlation between u_i and x_{it}.
  $Cov(u_i, x_{it}) = 0$ for $t = 1, \ldots, T$
  (or $\mathbb{E}[c_i | x_{i1}, \ldots, x_{iT}] = \mathbb{E}[c_i]$)
- Fixed effects: u_i is part of the intercept.
  u_i can be arbitrarily correlated with x_{it}.

Random effects

RE approach exploits the implied correlation structure of errors.

Let $v_{it} = u_i + e_{it}$. Stacking for T periods, we have $y_i = x_i\beta + v_i$. Define $\Omega = \mathbb{E}[v_i v_i' | x_i]$.

Assumptions

(RE 1) $\mathbb{E}[e_{it} | x_{i1}, \ldots, x_{iT}] = 0$ and $\mathbb{E}[u_i | x_{i1}, \ldots, x_{iT}] = 0$.

(RE 2) $\mathbb{E}[e_i e_i' | x_i, u_i] = \sigma_e^2 I_T$ and $\mathbb{E}[u_i^2 | x_i] = \sigma_u^2$

$$\Omega = \begin{bmatrix} \sigma_u^2 + \sigma_e^2 & \sigma_u^2 & \cdots & \sigma_u^2 \\ \sigma_u^2 & \sigma_u^2 + \sigma_e^2 & & \sigma_u^2 \\ \sigma_u^2 & & \ddots & \sigma_u^2 \\ \sigma_u^2 & \cdots & \sigma_u^2 & \sigma_u^2 + \sigma_e^2 \end{bmatrix}$$

- $\hat{\beta}_{RE} = (\sum_i x_i' \hat{\Omega}^{-1} x_i)^{-1} (\sum_i x_i' \hat{\Omega}^{-1} y_i)$.

Feasible GLS estimation of RE model

Step 1. Run a pooled OLS of y_{it} on x_{it} and get the residuals $\hat{v}_{it}$.

Step 2. Estimate $\sigma_v^2 = \sigma_u^2 + \sigma_e^2$ by $\hat{\sigma}_v^2 = \frac{1}{nT-k} \sum_i \sum_t \hat{v}_{it}^2$.

Step 3. Estimate σ_u^2 using cross terms only:
$\hat{\sigma}_u^2 = \frac{1}{nT(T-1)/2-k} \sum_{i=1}^{n} \sum_{t=1}^{T-1} \sum_{s=t+1}^{T} \hat{v}_{it}\hat{v}_{is}$.

Step 4. Form $\hat{\Omega}$ using $\hat{\sigma}_v^2$ and $\hat{\sigma}_u^2$.

Step 5. Estimate β by GLS: $\hat{\beta}_{RE} = (\sum_i x_i'\hat{\Omega}^{-1}x_i)^{-1}(\sum_i x_i'\hat{\Omega}^{-1}y_i)$

Fixed effects

Assumptions

(FE 1) $\mathbb{E}[e_{it}|x_{i1},\ldots,x_{iT}]=0$.

(FE 2) $\mathbb{E}[e_i e_i'|x_i,u_i]=\sigma_e^2 I_T$.

There are several derivations of the estimator.

- Add individual specific dummies: $y=X\beta+Du+e$. Then OLS estimation of β proceeds by the Frisch–Waugh–Lovell theorem. Define $y^*=y-D\left(D'D\right)^{-1}D'y$ and $X^*=X-D\left(D'D\right)^{-1}D'X$.

$$\hat{\beta}_{FE}=(X^{*\prime}X^*)^{-1}X^{*\prime}y^*$$

- De-mean/differencing: $\hat{\beta}_{FE}=\hat{\beta}_{within}$

Doing this in practice

- We can use the Hausman test to choose RE vs FE. (HO is in favor of "random effects")
- In stata, RE or FE estimation:

  $$xtset$$

  $$xtreg\ y\ x, re$$

  $$xtreg\ y\ x, fe$$

  Note that the default panel structure in Stata has two dimensions (individual *i* and time *t*). There are packages for higher dimensions, e.g. in Changkuk's MPL paper, he has "individual", "product" and "round".
- Estimating FE using dummies is very flexible when we want to control different levels of fixed effects. But the # of regressors can be very large.

Final thoughts

- Many ways of estimating variance: analytical/ bootstrap
- With iid data, if we worry about heteroskedasticity, there are HC0, HC1 (HW), HC2, HC3... When sample size is small, we'd better use HC2 or HC3.
- With data that is not iid, clustering can adjust the variance. We need to motivate why and how to cluster.
- Random effects or fixed effects are on the model level. It's helpful, e.g., when we want to control some individual-specific effects.
- Individual-specific effects are treated as part of the error in RE models, while as part of the intercept in FE models.

References

- *Econometrics* textbook by Bruce Hansen
- *“Yale Applied Empirical Methods PhD Courses”* by Paul Goldsmith-Pinkham [link]
- *“[99] Hyping Fisher: The Most Cited 2019 QJE Paper Relied on an Outdated Stata Default to Conclude Regression p-values Are Inadequate”* by Uri Simonsohn, on Data Colada [link]
- Jeffery Wooldridge’s comments on the Data Colada post [link]
- Alwyn Young, Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results, The Quarterly Journal of Economics, Volume 134, Issue 2, May 2019, Pages 557–598, https://doi.org/10.1093/qje/qjy029
- Alberto Abadie and others, When Should You Adjust Standard Errors for Clustering?, The Quarterly Journal of Economics, Volume 138, Issue 1, February 2023, Pages 1–35, https://doi.org/10.1093/qje/qjac038