Section 4: Inference

Valentine Gilbert

February 24, 2022

Overview

Derivation

Analytic standard errors rely on an **asymptotic approximation** (what does this mean?)

- Recall that $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$
- With a little algebra, we can rewrite this as: details

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \left(\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x_i}\boldsymbol{x_i}'\right)^{-1} \frac{\sqrt{N}}{N}\sum_{i=1}^{N} \boldsymbol{x_i}u_i$$

- What happens to the purple term as $N \to \infty$ and why?
  - It converges in probability to $\alpha = E[\boldsymbol{x_i}\boldsymbol{x_i}']^{-1}$ by the LLN and Slutsky
- And what happens to the red term as $N \to \infty$ and why?
  - It converges in distribution to $\mathcal{N}(0, \Sigma)$ by the CLT, where $\Sigma = E[(\boldsymbol{x_i}u_i)(\boldsymbol{x_i}u_i)']$
- So what happens to the left hand side as $N \to \infty$ and why?
  - It converges in distribution to $\mathcal{N}(0, \alpha\Sigma\alpha')$ by Slutsky

Homoskedasticity-Only Standard Errors

Homoskedasticity-only standard errors are wrong but useful for intution

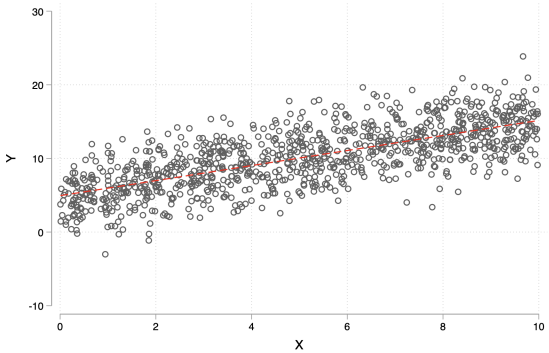
- Homoskedasticity holds if $E[u_i^2|\boldsymbol{x_i}] = \sigma^2$
- This greatly simplifies the variance-covariance formula:

$$var(\hat{\boldsymbol{\beta}}) = \sigma^2 E[\boldsymbol{x_i x_i'}]^{-1}/N$$

- The standard error for $\hat{\beta}_k$ is:

$$SE(\hat{\beta}_k) = \sqrt{\frac{\sigma^2}{var(\tilde{x}_{ki})N}}$$

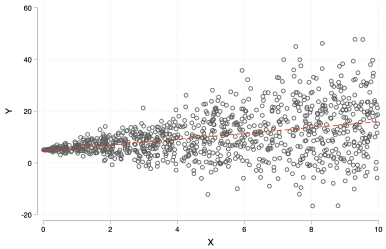
- So how can I get more precise estimates?

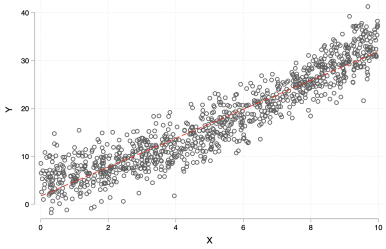


Heteroskedasticity-Consistent Standard Errors

Residuals are heteroskedastic if $E[u_i^2|\boldsymbol{x_i}]$ isn't constant

- This can happen if $var(Y_i|\boldsymbol{x_i})$ isn't constant...
- ...or if the CEF is non-linear
- Examples?
- There are several varieties of heteroskedasticity-consistent standard error estimators, all of which try to get better estimates in finite samples

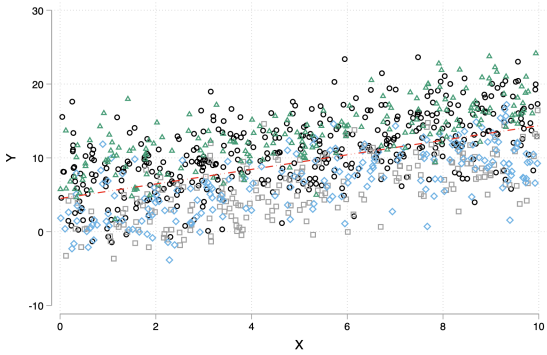




Clustered Standard Errors

Residuals are clustered if they're correlated within groups: $E[u_i u_j] \neq 0$ for some i, j

- Examples?
- Clustered standard errors allow for arbitrary correlation of residuals within groups but assume residuals are uncorrelated across groups
- Rule of thumb: Cluster at the level of treatment assignment
- Should also cluster by unit in panel data (Bertrand, Duflo, Mullainathan 2004)
- You can cluster on multiple dimensions!



A Basic Bootstrap

Suppose we have N observations of data (Y_i, X_i) and estimate $Y_i = \beta_0 + \beta_1 X_i + u_i$. The simplest bootstrap procedure estimates sampling uncertainty by repeatedly resampling from the observed data:

1. Use the original data to estimate β_1. Call this estimate $\hat{\beta}_1^*$
2. Do B iterations, indexed by b, of the following:
   1. Form a new sample by drawing N observations from the original data with replacement
   2. Estimate the regression function using the new sample and call the estimate $\hat{\beta}_1^b$
3. Use the sample standard deviation of $\hat{\beta}_1^b$ as an estimate of the standard error of $\hat{\beta}_1^*$

Why does this work? The idea is that drawing from the observed data approximates drawing samples from the population

Asymptotics

The bootstrap still relies on asymptotics, but some varieties can provide **asymptotic refinements** to analytic standard errors

- This means that our finite sample confidence intervals have closer to nominal "coverage"
- On the previous slide, our test statistic was just the coefficient of interest $\hat{\beta}_1$, which is asymptotically $\mathcal{N}(0, \Sigma)$
- If we choose an asymptotically pivotal test statistic (like the t-statistic), we get better finite sample performance

The Nonparametric Bootstrap-t

Here's a bootstrap procedure that uses the t-statistic as the test statistic:

1. Use the original data to estimate $\hat{\beta}_1$ and $se(\hat{\beta}_1)$. Calculate $t^* \equiv (\hat{\beta}_1 - \beta_{H_0}) / se(\hat{\beta}_1)$
2. Do B iterations, indexed by b, of the following:
   1. Form a new sample by drawing N observations from the original data with replacement
   2. Estimate the regression function using the new sample and calculate $t^b \equiv (\hat{\beta}^b - \hat{\beta}^*) / se(\hat{\beta}^b)$
3. Use the empirical CDF of t^b to test $H_0 : \beta_1 = \beta_{H_0}$
   - If $\widehat{Pr}(|t^b| \geq t^*) \leq .05$, reject H_0
   - Otherwise fail to reject H_0

How could we form a confidence interval?

Residual and Wild Bootstraps-t

The **residual bootstrap-t** does the following:

1. Calculate t^* as well as $\hat{Y}_i^*$ and $\hat{u}_i^*$ for all i
2. Resample with replacement from $\{\hat{u}_i^*\}$ to form $\{\hat{u}_i^b\}$, then form new outcomes $Y_i^b = \hat{Y}_i^* + \hat{u}_i^b$
3. Calculate t^b using estimates from regression of Y_i^b on X_i

The **wild bootstrap-t** does the following

1. Calculate t^* as well as $\hat{Y}_i^*$ and $\hat{u}_i^*$ for all i
2. Form Y_i^b by taking $\hat{Y}_i^*$ and adding $\hat{u}_i^*$ w/ probability .5 and subtracting $\hat{u}_i^*$ w/ probability .5
3. Calculate t^b using estimates from regression of Y_i^b on X_i

Possible to do all three varieties with clustering

Monte Carlo Results - Cameron, Gelbach, and Miller (2008 ReStat)

TABLE 2.—1,000 SIMULATIONS FROM DGP WITH GROUP-LEVEL RANDOM ERRORS
(Rejection rates for tests of nominal size 0.05 with simulation standard errors in parentheses)

Estimator #	Method	Number of Groups (G)					
		5	10	15	20	25	30
1	Assume i.i.d.	0.426	0.479	0.489	0.490	0.504	0.472
		(0.016)	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)
2	Moulton-type estimator	0.130	0.084	0.086	0.074	0.080	0.052
		(0.011)	(0.009)	(0.009)	(0.008)	(0.009)	(0.007)
3	Cluster-robust	0.195	0.132	0.096	0.093	0.095	0.069
		(0.013)	(0.011)	(0.009)	(0.009)	(0.009)	(0.008)
4	CR3 residual correction	0.088	0.084	0.065	0.072	0.067	0.057
		(0.009)	(0.009)	(0.008)	(0.008)	(0.008)	(0.007)
5	Pairs cluster bootstrap-se	0.152	0.122	0.095	0.096	0.100	0.072
		(0.011)	(0.010)	(0.009)	(0.009)	(0.009)	(0.008)
6	Residual cluster bootstrap-se	0.047	0.049	0.063	0.062	0.066	0.043
		(0.007)	(0.007)	(0.008)	(0.008)	(0.008)	(0.006)
7	Wild cluster bootstrap-se	0.012	0.031	0.039	0.041	0.056	0.040
		(0.003)	(0.005)	(0.006)	(0.006)	(0.007)	(0.006)
8	Pairs cluster bootstrap-BCA	0.161	0.106	0.101	0.087	0.094	0.068
		(0.012)	(0.010)	(0.010)	(0.009)	(0.009)	(0.008)
9	BDM bootstrap-t	0.117	0.109	0.094	0.094	0.095	0.068
		(0.010)	(0.010)	(0.009)	(0.009)	(0.009)	(0.008)
10	Pairs cluster bootstrap-t	0.081	0.082	0.075	0.073	0.070	0.054
		(0.009)	(0.009)	(0.008)	(0.008)	(0.008)	(0.007)
11	Pairs CR3 bootstrap-t	0.081	0.085	0.070	0.072	0.069	0.051
		(0.009)	(0.009)	(0.008)	(0.008)	(0.008)	(0.007)
12	Residual cluster bootstrap-t	0.034	0.052	0.049	0.044	0.056	0.050
		(0.006)	(0.007)	(0.007)	(0.006)	(0.007)	(0.007)
13	Wild cluster bootstrap-t	0.054	0.062	0.056	0.045	0.060	0.045
		(0.007)	(0.008)	(0.007)	(0.007)	(0.008)	(0.007)
	T_distribution(G-2)	0.145	0.086	0.072	0.066	0.062	0.060

Design-Based Uncertainty

Design-Based Uncertainty

- In randomized experiments, variability in our causal estimates comes from randomization, not sampling
  - We observe outcomes associated with only one of many potential randomization vectors
- Unlike with sampling uncertainty, we know the underlying probability distribution of randomization vectors
- For certain kinds of null hypotheses, we can re-randomize to generate the **exact distribution** of the test statistic under the null hypothesis
- No approximations (asymptotic or otherwise) involved!

$\Longrightarrow$ Can be especially useful when N is small and asymptotic approximations may be poor

Sharp and Dull Nulls

- Randomization inference allows us to test **sharp null hypotheses**
- A sharp null hypothesis is any hypothesis that lets us fill in counterfactual outcomes (i.e. the question marks)
- For example, the null hypothesis that the treatment effect is 0 *for everyone* is sharp
- The null hypothesis that the treatment effect is 5 for everyone is also sharp
- The null hypothesis that the treatment effect is 0 for men and 5 for women is sharp
- **Question**: Why isn't the null hypothesis of 0 average treatment effect sharp?

id	D_i	$Y_i(0)$	$Y_i(1)$
1	1	?	12
⋮	⋮	⋮	⋮

Example with N = 4

Suppose you observe the following data:

id	D_i	$Y_i(0)$	$Y_i(1)$
1	1	?	12
2	1	?	15
3	0	3	?
4	0	7	?

- Under a sharp null of a constant treatment effect of 0, what are the unobserved potential outcomes?

Example with N = 4

Suppose you observe the following data:

id	D_i	$Y_i(0)$	$Y_i(1)$
1	1	12	12
2	1	15	15
3	0	3	3
4	0	7	7

- Under a sharp null of a constant treatment effect of 0, what are the unobserved potential outcomes?

Example with N = 4

Suppose you observe the following data:

id	D_i	$Y_i(0)$	$Y_i(1)$
1	1	?	12
2	1	?	15
3	0	3	?
4	0	7	?

- Under a sharp null of a constant treatment effect of 5, what are the unobserved potential outcomes?

Example with N = 4

Suppose you observe the following data:

id	D_i	$Y_i(0)$	$Y_i(1)$
1	1	7	12
2	1	10	15
3	0	3	8
4	0	7	12

- Under a sharp null of a constant treatment effect of 5, what are the unobserved potential outcomes?

Hypothesis Testing

- So a sharp null hypothesis is any hypothesis that lets us fill in unobserved potential outcomes (i.e. counterfactual outcomes)
- How does this help us test if an observed difference in means is due to a causal effect or due to chance?
- Calculate the test statistic you would have calculated under different possible randomization draws
  - With $N = 4$, we can do this for all possible randomization vectors
  - When N is large, we can do this for a random sample of randomization vectors
- Compare the test statistic you calculated with the observed data to the distribution of test statistics under the null
  - If the observed test statistic is very extreme, it's unlikely to be due to chance

All Potential Randomizations and Differences in Means Under $H_0 : \beta_i = 0$ for all i

D_1	D_2	D_3	D_4	$\bar{Y}(1)$ - $\bar{Y}(0)$
0	0	1	1	-8.5
0	1	0	1	3.5
0	1	1	0	-.5
1	0	0	1	.5
1	0	1	0	-3.5
1	1	0	0	8.5

Question: What's the implied one-sided p-value of the observed difference in means? What about the two-sided p-value?

Monte Carlo Results - Young (2019 QJE)

SIZE AT THE 0.05 LEVEL IN 10,000 SIMULATIONS
(REJECTION RATES IN TESTS OF THE TRUE MEAN OF THE DATA-GENERATING PROCESS)

	Robust (1)	Rand-t (2)	Rand-c (3)	Boot-t (4)	Boot-c (5)	J-knife (6)	Robust (1)	Rand-t (2)	Rand-c (3)	Boot-t (4)	Boot-c (5)	J-knife (6)
Panel A: Tests of effects of binary treatment (t_i) given the data-generating process $y_i = \alpha + t_i{}^*\beta_i + \varepsilon_i$												
	Balanced regression design						Unbalanced regression design					
Fixed treatment effects: $\beta_i = \beta$, $\varepsilon_i \sim$ standard normal												
20	0.048	0.048	0.048	0.039	0.068	0.044	0.241	0.046	0.048	0.000	0.108	0.140
200	0.048	0.048	0.048	0.050	0.051	0.048	0.067	0.051	0.050	0.049	0.063	0.057
2,000	0.049	0.049	0.049	0.050	0.050	0.049	0.053	0.052	0.052	0.051	0.053	0.052
Heterogeneous treatment effects: $\beta_i \sim$ standard normal, $\varepsilon_i \sim$ standard normal												
20	0.052	0.052	0.052	0.040	0.073	0.046	0.283	0.089	0.129	0.000	0.129	0.172
200	0.053	0.052	0.052	0.053	0.055	0.052	0.064	0.051	0.131	0.045	0.060	0.055
2,000	0.049	0.048	0.048	0.048	0.048	0.049	0.052	0.052	0.137	0.051	0.052	0.051
Heterogeneous treatment effects: $\beta_i \sim \text{chi}^2$, $\varepsilon_i \sim$ standard normal												
20	0.060	0.062	0.062	0.046	0.082	0.055	0.290	0.091	0.144	0.000	0.131	0.174
200	0.054	0.055	0.055	0.051	0.055	0.053	0.083	0.065	0.189	0.056	0.079	0.071
2,000	0.045	0.045	0.045	0.045	0.045	0.045	0.054	0.052	0.195	0.051	0.054	0.053

A Little Algebra 1/2

The regression of interest is $y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + u_i$

- Let's start with our estimate of $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \left(\sum_{i=1}^{N} \boldsymbol{x_i}\boldsymbol{x_i}'\right)^{-1} \sum_{i=1}^{N} \boldsymbol{x_i} y_i$$

$$= \left(\sum_{i=1}^{N} \boldsymbol{x_i}\boldsymbol{x_i}'\right)^{-1} \sum_{i=1}^{N} \boldsymbol{x_i}(\boldsymbol{x_i}'\boldsymbol{\beta} + u_i)$$

$$= \boldsymbol{\beta} + \left(\sum_{i=1}^{N} \boldsymbol{x_i}\boldsymbol{x_i}'\right)^{-1} \sum_{i=1}^{N} \boldsymbol{x_i} u_i$$

$$\implies \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \left(\sum_{i=1}^{N} \boldsymbol{x_i}\boldsymbol{x_i}'\right)^{-1} \sum_{i=1}^{N} \boldsymbol{x_i} u_i$$

A Little Algebra 2/2

back

Now rewrite $\hat{\beta} - \beta$ in terms of sample averages so we can apply the LLN and CLT

$$\begin{aligned}\hat{\beta} - \beta &= \left(\sum_{i=1}^{N} \boldsymbol{x_i} \boldsymbol{x_i}'\right)^{-1} \sum_{i=1}^{N} \boldsymbol{x_i} u_i \\ &= \left(\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x_i} \boldsymbol{x_i}'\right)^{-1} \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x_i} u_i\end{aligned}$$

Remember that the standard deviation of a sample average shrinks at a rate of $\sqrt{N}$

- We want to understand what happens to the sampling distribution of $\hat{\beta}$ as $N \to \infty$
- So we multiply by $\sqrt{N}$ to keep the standard deviation of a sample average constant

$$\sqrt{N}(\hat{\beta} - \beta) = \left(\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x_i} \boldsymbol{x_i}'\right)^{-1} \frac{\sqrt{N}}{N} \sum_{i=1}^{N} \boldsymbol{x_i} u_i$$