

Inference: Clustering

EC 607, Set 10

Edward Rubin
Spring 2020

Prologue

Schedule

Last time

Regression discontinuities

Today

Inference and clustering

Inference

Inference

Motivation

So far, we've focused on carefully **obtaining causal estimates** of the effect of some treatment \mathbf{D}_i on our outcome \mathbf{Y}_i .

Inference

Motivation

So far, we've focused on carefully **obtaining causal estimates** of the effect of some treatment D_i on our outcome Y_i .

Our discussion of research designs and their requirements/assumptions has centered on **avoiding selection and securing unbiased and/or consistent estimates** for τ .

Inference

Motivation

So far, we've focused on carefully **obtaining causal estimates** of the effect of some treatment D_i on our outcome Y_i .

Our discussion of research designs and their requirements/assumptions has centered on **avoiding selection and securing unbiased and/or consistent estimates** for τ .

In other words, we've concentrated on **point estimates**.

What about **inference**?

Inference

Shminference †

Q Why care about inference?

† What is *shminference*?

Inference

Shminference †

Q Why care about inference?

A I'll give you two reasons.

† What is *shminference*?

Inference

Shminference †

Q Why care about inference?

A I'll give you two reasons.

1. We often want to **test theories/hypotheses**. Point estimates (*i.e.*, $\hat{\beta}$) can't do this alone. Inference finishes the job.

† What is *shminference*?

Inference

Shminference †

Q Why care about inference?

A I'll give you two reasons.

1. We often want to **test theories/hypotheses**. Point estimates (*i.e.*, $\hat{\beta}$) can't do this alone. Inference finishes the job.
2. Other times, we want to **measure the effect** of a treatment. Inference helps us think about the **precision** of our estimates.

† What is *shminference*?

Inference

Shminference †

Q Why care about inference?

A I'll give you two reasons.

1. We often want to **test theories/hypotheses**. Point estimates (*i.e.*, $\hat{\beta}$) can't do this alone. Inference finishes the job.
2. Other times, we want to **measure the effect** of a treatment. Inference helps us think about the **precision** of our estimates.

Note: Similar reasoning can apply to bounding forecasting/predictions.

† What is *shminference*?

Inference

Shminference †

Q Why care about inference?

A I'll give you two reasons.

1. We often want to **test theories/hypotheses**. Point estimates (*i.e.*, $\hat{\beta}$) can't do this alone. Inference finishes the job.
2. Other times, we want to **measure the effect** of a treatment. Inference helps us think about the **precision** of our estimates.

Note: Similar reasoning can apply to bounding forecasting/predictions.

If you want answers, then you need to do inference

† What is *shminference*?

Inference

Shminference †

Q Why care about inference?

A I'll give you two reasons.

1. We often want to **test theories/hypotheses**. Point estimates (*i.e.*, $\hat{\beta}$) can't do this alone. Inference finishes the job.
2. Other times, we want to **measure the effect** of a treatment. Inference helps us think about the **precision** of our estimates.

Note: Similar reasoning can apply to bounding forecasting/predictions.

If you want answers, then you need to do inference correctly.

† What is *shminference*?

Inference

What's so complicated?

Angrist and Pischke told us that "correcting" our standard errors for heteroskedasticity may increase the standard errors up to 25%.

What else are we worried about?

Inference

What we're worried about

- **Transformations of estimators**, *i.e.*, $\text{Var} \left[f \left(\hat{\beta} \right) \right] \neq f \left(\text{Var} \left[\hat{\beta} \right] \right)$

Inference

What we're worried about

- **Transformations of estimators**, *i.e.*, $\text{Var} \left[f \left(\hat{\beta} \right) \right] \neq f \left(\text{Var} \left[\hat{\beta} \right] \right)$
- **Dependence/correlation in our disturbance**, *i.e.*, $\text{Cov}(\varepsilon_i, \varepsilon_j) \neq 0$
 - Autocorrelation $\varepsilon_t = \rho\varepsilon_{t-1} + \varepsilon_t$
 - Correlated shocks within groups $\varepsilon_i = \varepsilon_{g(i)} + \varepsilon_i$

Inference

What we're worried about

- **Transformations of estimators**, *i.e.*, $\text{Var} \left[f \left(\hat{\beta} \right) \right] \neq f \left(\text{Var} \left[\hat{\beta} \right] \right)$
- **Dependence/correlation in our disturbance**, *i.e.*, $\text{Cov}(\varepsilon_i, \varepsilon_j) \neq 0$
 - Autocorrelation $\varepsilon_t = \rho\varepsilon_{t-1} + \varepsilon_t$
 - Correlated shocks within groups $\varepsilon_i = \varepsilon_{g(i)} + \varepsilon_i$
- **Finite-sample properties** vs. asymptotic properties

Inference

What we're worried about

- **Transformations of estimators**, *i.e.*, $\text{Var} \left[f \left(\hat{\beta} \right) \right] \neq f \left(\text{Var} \left[\hat{\beta} \right] \right)$
- **Dependence/correlation in our disturbance**, *i.e.*, $\text{Cov}(\varepsilon_i, \varepsilon_j) \neq 0$
 - Autocorrelation $\varepsilon_t = \rho\varepsilon_{t-1} + \varepsilon_t$
 - Correlated shocks within groups $\varepsilon_i = \varepsilon_{g(i)} + \varepsilon_i$
- **Finite-sample properties** vs. asymptotic properties
- **Power** and **minimal detectable effects**

Inference

What we're worried about

- **Transformations of estimators**, *i.e.*, $\text{Var} \left[f \left(\hat{\beta} \right) \right] \neq f \left(\text{Var} \left[\hat{\beta} \right] \right)$
- **Dependence/correlation in our disturbance**, *i.e.*, $\text{Cov}(\varepsilon_i, \varepsilon_j) \neq 0$
 - Autocorrelation $\varepsilon_t = \rho\varepsilon_{t-1} + \varepsilon_t$
 - Correlated shocks within groups $\varepsilon_i = \varepsilon_{g(i)} + \varepsilon_i$
- **Finite-sample properties** vs. asymptotic properties
- **Power** and **minimal detectable effects**
- **Multiple-hypothesis testing** and ***p-hacking***

Inference

What we're worried about

- **Transformations of estimators**, *i.e.*, $\text{Var} \left[f \left(\hat{\beta} \right) \right] \neq f \left(\text{Var} \left[\hat{\beta} \right] \right)$
- **Dependence/correlation in our disturbance**, *i.e.*, $\text{Cov}(\varepsilon_i, \varepsilon_j) \neq 0$
 - Autocorrelation $\varepsilon_t = \rho\varepsilon_{t-1} + \varepsilon_t$
 - Correlated shocks within groups $\varepsilon_i = \varepsilon_{g(i)} + \varepsilon_i$
- **Finite-sample properties** vs. asymptotic properties
- **Power** and **minimal detectable effects**
- **Multiple-hypothesis testing** and ***p-hacking***

In other words: We've got a lot to worry/think about.

Clustering

Clustering

Setup

Many studies—observational and experimental—have a treatment that is assigned to all/most individuals within a group.

- Classrooms/schools
- Households
- Villages/counties/states

Clustering

Setup

Many studies—observational and experimental—have a treatment that is assigned to all/most individuals within a group.

- Classrooms/schools
- Households
- Villages/counties/states

Furthermore, we might imagine individuals within the same group may have correlated disturbances. For i and j in group g

$$\text{Cov}(\varepsilon_i, \varepsilon_j) = E[\varepsilon_i \varepsilon_j] = \rho_\varepsilon \sigma_\varepsilon^2$$

where ρ_ε gives the within-group correlation of disturbances—what MHE calls the **intraclass correlation coefficient**.

Clustering

Setup

In other words, we have a regression

$$y_i = \beta_0 + \beta_1 x_{g(i)} + \varepsilon_i$$

where individual i is in group g , and $\mathbf{X}_{g(i)}$ only varies across groups.

Clustering

Setup

In other words, we have a regression

$$y_i = \beta_0 + \beta_1 x_{g(i)} + \varepsilon_i$$

where individual i is in group g , and $\mathbf{X}_{g(i)}$ only varies across groups.

For within-group correlation, we can use an additive random-effects model

$$\varepsilon_i = \nu_{g(i)} + \eta_i$$

meaning group members all receive a common shock $\nu_{g(i)}$, and individuals receive independent shocks η_i .

Clustering

Setup

In other words, we have a regression

$$y_i = \beta_0 + \beta_1 x_{g(i)} + \varepsilon_i$$

where individual i is in group g , and $\mathbf{X}_{g(i)}$ only varies across groups.

For within-group correlation, we can use an additive random-effects model

$$\varepsilon_i = \nu_{g(i)} + \eta_i$$

meaning group members all receive a common shock $\nu_{g(i)}$, and individuals receive independent shocks η_i .

Note We assume η_i is independent of η_j ($i \neq j$) and ν_g ($\forall g$).

Clustering

Additive random effects

Based upon this model we've set up

$$\varepsilon_i = \nu_{g(i)} + \eta_i$$

the covariance between individuals i and j in group g is

$$\begin{aligned}\text{Cov}(\varepsilon_i, \varepsilon_j) &= E[\varepsilon_i \varepsilon_j] = E[(\nu_g + \eta_i)(\nu_g + \eta_j)] = E[\nu_g^2] = \sigma_\nu^2 \\ &= \rho_\varepsilon \sigma_\varepsilon^2 \\ &= \rho_\varepsilon (\sigma_\nu^2 + \sigma_\eta^2)\end{aligned}$$

Thus, we can write the intraclass correlation coefficient as

$$\rho_\varepsilon = \frac{\sigma_\nu^2}{\sigma_\varepsilon^2} = \frac{\sigma_\nu^2}{\sigma_\nu^2 + \sigma_\eta^2}$$

Clustering

What is ρ_ε ?

Let's review what we know.

$$\varepsilon_i = \nu_{g(i)} + \eta_i \quad \text{and} \quad \rho_\varepsilon = \frac{\sigma_\nu^2}{\sigma_\varepsilon^2} = \frac{\sigma_\nu^2}{\sigma_\nu^2 + \sigma_\eta^2}$$

Clustering

What is ρ_ε ?

Let's review what we know.

$$\varepsilon_i = \nu_{g(i)} + \eta_i \quad \text{and} \quad \rho_\varepsilon = \frac{\sigma_\nu^2}{\sigma_\varepsilon^2} = \frac{\sigma_\nu^2}{\sigma_\nu^2 + \sigma_\eta^2}$$

One way to think about ρ_ε is as the **share of the variance of the disturbance ε_i accounted for by the shared disturbance $\nu_{g(i)}$.**

Clustering

What is ρ_ε ?

Let's review what we know.

$$\varepsilon_i = \nu_{g(i)} + \eta_i \quad \text{and} \quad \rho_\varepsilon = \frac{\sigma_\nu^2}{\sigma_\varepsilon^2} = \frac{\sigma_\nu^2}{\sigma_\nu^2 + \sigma_\eta^2}$$

One way to think about ρ_ε is as the **share of the variance of the disturbance ε_i accounted for by the shared disturbance $\nu_{g(i)}$** .

As $\nu_{g(i)}$ accounts for more and more of the variation in ε_i , $\rho_\varepsilon \rightarrow 1$.

Clustering

So...

Q Why do we care about ρ_ϵ ?

Clustering

So...

Q Why do we care about ρ_ε ?

A It tells us by how wrong our standard errors can be if we treat all observations as independent.

Clustering

So...

Q Why do we care about ρ_ε ?

A It tells us by how wrong our standard errors can be if we treat all observations as independent.

Let $\text{Var}_o(\hat{\beta}_1)$ denote the conventional variance formula for OLS estimator.[†]

[†] which treats all disturbances as independent (and identically distributed).

Clustering

So...

Q Why do we care about ρ_ε ?

A It tells us by how wrong our standard errors can be if we treat all observations as independent.

Let $\text{Var}_o(\hat{\beta}_1)$ denote the conventional variance formula for OLS estimator.[†]

Let $\text{Var}(\hat{\beta}_1)$ denote the actual variance of $\hat{\beta}_1$.

[†] which treats all disturbances as independent (and identically distributed).

Clustering

So....

With **(1)** nonstochastic regressors fixed by group *and* **(2)** groups of size n

$$\frac{\text{Var}(\hat{\beta}_1)}{\text{Var}_o(\hat{\beta}_1)} = 1 + (n - 1)\rho_\varepsilon$$

Clustering

So....

With **(1)** nonstochastic regressors fixed by group *and* **(2)** groups of size n

$$\frac{\text{Var}(\hat{\beta}_1)}{\text{Var}_o(\hat{\beta}_1)} = 1 + (n - 1)\rho_\varepsilon \quad \implies \quad \frac{\text{S.E.}(\hat{\beta}_1)}{\text{S.E.}_o(\hat{\beta}_1)} = \sqrt{1 + (n - 1)\rho_\varepsilon}$$

Clustering

So....

With **(1)** nonstochastic regressors fixed by group and **(2)** groups of size n

$$\frac{\text{Var}(\hat{\beta}_1)}{\text{Var}_o(\hat{\beta}_1)} = 1 + (n-1)\rho_\varepsilon \quad \implies \quad \frac{\text{S.E.}(\hat{\beta}_1)}{\text{S.E.}_o(\hat{\beta}_1)} = \sqrt{1 + (n-1)\rho_\varepsilon}$$

The term $\sqrt{1 + (n-1)\rho_\varepsilon}$ is called the **Moulton factor**[†].

The **Moulton factor** tells us by what factor standard errors will be wrong if we ignore within-group correlation (conditional on assumptions **1** and **2**).

[†] After Moulton (1986). Derivation: MHE 323–325.

Clustering

So....

With **(1)** nonstochastic regressors fixed by group and **(2)** groups of size n

$$\frac{\text{Var}(\hat{\beta}_1)}{\text{Var}_o(\hat{\beta}_1)} = 1 + (n-1)\rho_\varepsilon \quad \implies \quad \frac{\text{S.E.}(\hat{\beta}_1)}{\text{S.E.}_o(\hat{\beta}_1)} = \sqrt{1 + (n-1)\rho_\varepsilon}$$

The term $\sqrt{1 + (n-1)\rho_\varepsilon}$ is called the **Moulton factor**[†].

The **Moulton factor** tells us by what factor standard errors will be wrong if we ignore within-group correlation (conditional on assumptions **1** and **2**).

Q What happens if $\rho = 1$?

[†] After [Moulton \(1986\)](#). Derivation: *MHE* 323–325.

Clustering

So....

With **(1)** nonstochastic regressors fixed by group and **(2)** groups of size n

$$\frac{\text{Var}(\hat{\beta}_1)}{\text{Var}_o(\hat{\beta}_1)} = 1 + (n-1)\rho_\varepsilon \quad \implies \quad \frac{\text{S.E.}(\hat{\beta}_1)}{\text{S.E.}_o(\hat{\beta}_1)} = \sqrt{1 + (n-1)\rho_\varepsilon}$$

The term $\sqrt{1 + (n-1)\rho_\varepsilon}$ is called the **Moulton factor**[†].

The **Moulton factor** tells us by what factor standard errors will be wrong if we ignore within-group correlation (conditional on assumptions **1** and **2**).

Q What happens if $\rho = 1$? What if you duplicated your dataset?

[†] After [Moulton \(1986\)](#). Derivation: *MHE* 323–325.

Clustering

So....

With **(1)** nonstochastic regressors fixed by group and **(2)** groups of size n

$$\frac{\text{Var}(\hat{\beta}_1)}{\text{Var}_o(\hat{\beta}_1)} = 1 + (n-1)\rho_\varepsilon \quad \implies \quad \frac{\text{S.E.}(\hat{\beta}_1)}{\text{S.E.}_o(\hat{\beta}_1)} = \sqrt{1 + (n-1)\rho_\varepsilon}$$

The term $\sqrt{1 + (n-1)\rho_\varepsilon}$ is called the **Moulton factor**[†].

The **Moulton factor** tells us by what factor standard errors will be wrong if we ignore within-group correlation (conditional on assumptions **1** and **2**).

- Q What happens if $\rho = 1$? What if you duplicated your dataset?
- Q What happens as n increases?

[†] After Moulton (1986). Derivation: MHE 323–325.

Clustering

The Moulton factor

The Moulton factor

$$\frac{\text{S.E.}(\hat{\beta}_1)}{\text{S.E.}_o(\hat{\beta}_1)} = \sqrt{1 + (n - 1)\rho_\varepsilon}$$

shows even when ρ_ε is small, we can have vary large standard error issues.

Clustering

The Moulton factor

The Moulton factor

$$\frac{\text{S.E.}(\hat{\beta}_1)}{\text{S.E.}_o(\hat{\beta}_1)} = \sqrt{1 + (n - 1)\rho_\varepsilon}$$

shows even when ρ_ε is small, we can have vary large standard error issues.

Ex An experiment on 400 schools, each with 1,000 students.

Clustering

The Moulton factor

The Moulton factor

$$\frac{\text{S.E.}(\hat{\beta}_1)}{\text{S.E.}_o(\hat{\beta}_1)} = \sqrt{1 + (n - 1)\rho_\varepsilon}$$

shows even when ρ_ε is small, we can have vary large standard error issues.

Ex An experiment on 400 schools, each with 1,000 students.

If $\rho_\varepsilon = 0.01$, the Moulton factor is $\sqrt{1 + (1,000 - 1) \times 0.01} \approx 3.32$.

Clustering

Test statistics

Recall $t_{\text{stat}} = \frac{\hat{\beta}_1}{\text{S.E.}(\hat{\beta}_1)}$.

Clustering

Test statistics

Recall $t_{\text{stat}} = \frac{\hat{\beta}_1}{\text{S.E.}(\hat{\beta}_1)}$.

$\therefore \frac{t_o}{t}$

Clustering

Test statistics

$$\text{Recall } t_{\text{stat}} = \frac{\hat{\beta}_1}{\text{S.E.}(\hat{\beta}_1)}.$$

$$\therefore \frac{t_o}{t} = \frac{\hat{\beta}_1 / \text{S.E.}_o(\hat{\beta}_1)}{\hat{\beta}_1 / \text{S.E.}(\hat{\beta}_1)}$$

Clustering

Test statistics

Recall $t_{\text{stat}} = \frac{\hat{\beta}_1}{\text{S.E.}(\hat{\beta}_1)}$.

$$\therefore \frac{t_o}{t} = \frac{\hat{\beta}_1 / \text{S.E.}_o(\hat{\beta}_1)}{\hat{\beta}_1 / \text{S.E.}(\hat{\beta}_1)} = \frac{\text{S.E.}(\hat{\beta}_1)}{\text{S.E.}_o(\hat{\beta}_1)}$$

Clustering

Test statistics

Recall $t_{\text{stat}} = \frac{\hat{\beta}_1}{\text{S.E.}(\hat{\beta}_1)}$.

$$\therefore \frac{t_o}{t} = \frac{\hat{\beta}_1 / \text{S.E.}_o(\hat{\beta}_1)}{\hat{\beta}_1 / \text{S.E.}(\hat{\beta}_1)} = \frac{\text{S.E.}(\hat{\beta}_1)}{\text{S.E.}_o(\hat{\beta}_1)} = \text{the Moulton factor.}$$

Clustering

Test statistics

$$\text{Recall } t_{\text{stat}} = \frac{\hat{\beta}_1}{\text{S.E.}(\hat{\beta}_1)}.$$

$$\therefore \frac{t_o}{t} = \frac{\hat{\beta}_1 / \text{S.E.}_o(\hat{\beta}_1)}{\hat{\beta}_1 / \text{S.E.}(\hat{\beta}_1)} = \frac{\text{S.E.}(\hat{\beta}_1)}{\text{S.E.}_o(\hat{\beta}_1)} = \text{the Moulton factor.}$$

Ex Thus, in our example of 400 schools with 1,000 students, ignoring within-school correlation of $\rho_\varepsilon = 0.01$ would lead us test statistics that are more than 3 times as large as they should be.

Clustering

Test statistics

$$\text{Recall } t_{\text{stat}} = \frac{\hat{\beta}_1}{\text{S.E.}(\hat{\beta}_1)}.$$

$$\therefore \frac{t_o}{t} = \frac{\hat{\beta}_1 / \text{S.E.}_o(\hat{\beta}_1)}{\hat{\beta}_1 / \text{S.E.}(\hat{\beta}_1)} = \frac{\text{S.E.}(\hat{\beta}_1)}{\text{S.E.}_o(\hat{\beta}_1)} = \text{the Moulton factor.}$$

Ex Thus, in our example of 400 schools with 1,000 students, ignoring within-school correlation of $\rho_\varepsilon = 0.01$ would lead us test statistics that are more than 3 times as large as they should be.

This is why economics seminars have standard-error police.

Clustering

Relaxing assumptions

If we allow regressors to vary by individual and groups to differ in size (n_g),

$$\frac{\text{Var}(\hat{\beta}_1)}{\text{Var}_o(\hat{\beta}_1)} = 1 + \left[\frac{\text{Var}(n_g)}{\bar{n}} + \bar{n} - 1 \right] \rho_x \rho_\varepsilon$$

where ρ_x denotes the intraclass (within-group) correlation of x_i .[†]

[†] See *MHE* for mathematical definitions and the derivation.

Clustering

Relaxing assumptions

If we allow regressors to vary by individual and groups to differ in size (n_g),

$$\frac{\text{Var}(\hat{\beta}_1)}{\text{Var}_o(\hat{\beta}_1)} = 1 + \left[\frac{\text{Var}(n_g)}{\bar{n}} + \bar{n} - 1 \right] \rho_x \rho_\varepsilon$$

where ρ_x denotes the intraclass (within-group) correlation of x_i .[†]

Important The Moulton factor for this general model depends upon the amount of within-group correlation in x_i and ε_i .

[†] See *MHE* for mathematical definitions and the derivation.

Clustering

Relaxing assumptions

If we allow regressors to vary by individual and groups to differ in size (n_g),

$$\frac{\text{Var}(\hat{\beta}_1)}{\text{Var}_o(\hat{\beta}_1)} = 1 + \left[\frac{\text{Var}(n_g)}{\bar{n}} + \bar{n} - 1 \right] \rho_x \rho_\varepsilon$$

where ρ_x denotes the intraclass (within-group) correlation of x_i .[†]

Important The Moulton factor for this general model depends upon the amount of within-group correlation in x_i and ε_i .

The special case is also important, as treatment is often fixed at some level.

[†] See *MHE* for mathematical definitions and the derivation.

Clustering

The answer

Q So what do we do now?

Clustering

The answer

Q So what do we do now?

A We've got options (as usual)

1. Parametrically model the random effects
2. Cluster-robust standard error (estimator)
3. Aggregate up to the group (or a similar method)
4. Block (group-based) bootstrap
5. GLS/MLE modeling y_i and ε_i

Clustering

The answer

Q So what do we do now?

A We've got options (as usual)

1. Parametrically model the random effects
2. Cluster-robust standard error (estimator)
3. Aggregate up to the group (or a similar method)
4. Block (group-based) bootstrap
5. GLS/MLE modeling y_i and ε_i

Most common: Cluster-robust standard errors

Runner up: Block bootstrap

Second runner up: Group-level analysis

Clustering

Cluster-robust standard errors

Liang and Zeger (1986) extend White's heteroskedasticity-robust covariance matrix to allow for both clustering and heteroskedasticity.[†]

[†] When people say *clustering*, they typically mean *correlated disturbances within a group*.

Clustering

Cluster-robust standard errors

Liang and Zeger (1986) extend White's heteroskedasticity-robust covariance matrix to allow for both clustering and heteroskedasticity.[†]

$$\hat{\Omega}_{\text{cl}} = (\mathbf{X}'\mathbf{X})^{-1} \left(\sum_g \mathbf{X}'_g \hat{\Psi}_g \mathbf{X}_g \right) (\mathbf{X}'\mathbf{X})^{-1}$$

[†] When people say *clustering*, they typically mean *correlated disturbances within a group*.

Clustering

Cluster-robust standard errors

Liang and Zeger (1986) extend White's heteroskedasticity-robust covariance matrix to allow for both clustering and heteroskedasticity.[†]

$$\hat{\Omega}_{\text{cl}} = (\mathbf{X}'\mathbf{X})^{-1} \left(\sum_g \mathbf{X}'_g \hat{\Psi}_g \mathbf{X}_g \right) (\mathbf{X}'\mathbf{X})^{-1}$$
$$\hat{\Psi}_g = a e_g e'_g = a \begin{bmatrix} e_{1g}^2 & e_{1g}e_{2g} & \cdots & e_{1g}e_{n_gg} \\ e_{1g}e_{2g} & e_{2g}^2 & e_{2g} \cdots & e_{2g}e_{n_gg} \\ \vdots & \vdots & \ddots & \vdots \\ e_{1g}e_{n_gg} & e_{2g}e_{n_gg} & \cdots & e_{n_gg}^2 \end{bmatrix}$$

where e_g are the OLS residuals for group g , e_{ig} is the residual for individual i in group g , and a is a degrees-of-freedom adjustment.

[†] When people say *clustering*, they typically mean *correlated disturbances within a group*.

Clustering

Cluster-robust standard errors

Derivation Let \mathbf{x}_i denote observation i (row) from \mathbf{X} .

Clustering

Cluster-robust standard errors

Derivation Let \mathbf{x}_i denote observation i (row) from \mathbf{X} .

$$\text{Var}\left(\hat{\beta}|\mathbf{X}\right) = E\left[\left(\hat{\beta} - \beta\right)\left(\hat{\beta} - \beta\right)'|\mathbf{X}\right] = E\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\varepsilon\varepsilon'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}|\mathbf{X}\right]$$

Clustering

Cluster-robust standard errors

Derivation Let \mathbf{x}_i denote observation i (row) from \mathbf{X} .

$$\begin{aligned}\text{Var}\left(\hat{\beta}|\mathbf{X}\right) &= E\left[\left(\hat{\beta} - \beta\right)\left(\hat{\beta} - \beta\right)'|\mathbf{X}\right] = E\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\varepsilon\varepsilon'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}|\mathbf{X}\right] \\ &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}' E\left[\varepsilon\varepsilon'|\mathbf{X}\right]\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\end{aligned}$$

Clustering

Cluster-robust standard errors

Derivation Let \mathbf{x}_i denote observation i (row) from \mathbf{X} .

$$\begin{aligned}\text{Var}\left(\hat{\beta}|\mathbf{X}\right) &= E\left[\left(\hat{\beta} - \beta\right)\left(\hat{\beta} - \beta\right)'|\mathbf{X}\right] = E\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\varepsilon\varepsilon'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}|\mathbf{X}\right] \\ &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}' E\left[\varepsilon\varepsilon'|\mathbf{X}\right]\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \\ &= \left(\sum_{i=1}^N \mathbf{x}'_i \mathbf{x}_i\right)^{-1} \left(\sum_{i=1}^N \sum_{j=1}^N \mathbf{x}'_i \mathbf{x}_j E\left[\varepsilon_j \varepsilon_i|\mathbf{X}\right]\right) \left(\sum_{i=1}^N \mathbf{x}'_i \mathbf{x}_i\right)^{-1}\end{aligned}$$

Clustering

Cluster-robust standard errors

Derivation Let \mathbf{x}_i denote observation i (row) from \mathbf{X} .

$$\begin{aligned}\text{Var}\left(\hat{\beta}|\mathbf{X}\right) &= E\left[\left(\hat{\beta} - \beta\right)\left(\hat{\beta} - \beta\right)'|\mathbf{X}\right] = E\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\varepsilon\varepsilon'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}|\mathbf{X}\right] \\ &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}' E\left[\varepsilon\varepsilon'|\mathbf{X}\right]\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \\ &= \left(\sum_{i=1}^N \mathbf{x}'_i \mathbf{x}_i\right)^{-1} \left(\sum_{i=1}^N \sum_{j=1}^N \mathbf{x}'_i \mathbf{x}_j E\left[\varepsilon_j \varepsilon_i|\mathbf{X}\right]\right) \left(\sum_{i=1}^N \mathbf{x}'_i \mathbf{x}_i\right)^{-1}\end{aligned}$$

Q Can we estimate $\left(\sum_i \sum_j \mathbf{x}'_i \mathbf{x}_j E\left[\varepsilon_j \varepsilon_i|\mathbf{X}\right]\right)$ with $\sum_i \sum_j \mathbf{x}'_i \mathbf{x}_j e_j e_i = \mathbf{X}' e e' \mathbf{X}$?

Clustering

Cluster-robust standard errors

Derivation Let \mathbf{x}_i denote observation i (row) from \mathbf{X} .

$$\begin{aligned}\text{Var}\left(\hat{\beta}|\mathbf{X}\right) &= E\left[\left(\hat{\beta} - \beta\right)\left(\hat{\beta} - \beta\right)'|\mathbf{X}\right] = E\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\varepsilon\varepsilon'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}|\mathbf{X}\right] \\ &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}' E\left[\varepsilon\varepsilon'|\mathbf{X}\right]\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \\ &= \left(\sum_{i=1}^N \mathbf{x}'_i \mathbf{x}_i\right)^{-1} \left(\sum_{i=1}^N \sum_{j=1}^N \mathbf{x}'_i \mathbf{x}_j E\left[\varepsilon_j \varepsilon_i|\mathbf{X}\right]\right) \left(\sum_{i=1}^N \mathbf{x}'_i \mathbf{x}_i\right)^{-1}\end{aligned}$$

Q Can we estimate $\left(\sum_i \sum_j \mathbf{x}'_i \mathbf{x}_j E\left[\varepsilon_j \varepsilon_i|\mathbf{X}\right]\right)$ with $\sum_i \sum_j \mathbf{x}'_i \mathbf{x}_j e_j e_i = \mathbf{X}' e e' \mathbf{X}$?

A No. Recall with OLS, $\mathbf{X}' e = 0$.

Clustering

Cluster-robust standard errors

Derivation Let \mathbf{x}_i denote observation i (row) from \mathbf{X} .

$$\begin{aligned}\text{Var}\left(\hat{\beta}|\mathbf{X}\right) &= E\left[\left(\hat{\beta} - \beta\right)\left(\hat{\beta} - \beta\right)'|\mathbf{X}\right] = E\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\varepsilon\varepsilon'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}|\mathbf{X}\right] \\ &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}' E\left[\varepsilon\varepsilon'|\mathbf{X}\right]\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \\ &= \left(\sum_{i=1}^N \mathbf{x}'_i \mathbf{x}_i\right)^{-1} \left(\sum_{i=1}^N \sum_{j=1}^N \mathbf{x}'_i \mathbf{x}_j E\left[\varepsilon_j \varepsilon_i|\mathbf{X}\right]\right) \left(\sum_{i=1}^N \mathbf{x}'_i \mathbf{x}_i\right)^{-1}\end{aligned}$$

Q Can we estimate $\left(\sum_i \sum_j \mathbf{x}'_i \mathbf{x}_j E\left[\varepsilon_j \varepsilon_i|\mathbf{X}\right]\right)$ with $\sum_i \sum_j \mathbf{x}'_i \mathbf{x}_j e_j e_i = \mathbf{X}' e e' \mathbf{X}$?

A No. Recall with OLS, $\mathbf{X}' e = \mathbf{0}$. But we will do something similar.

Clustering

Cluster-robust standard errors

Imagine we have G clusters with some unknown **dependence between observations within a cluster** and **independence between clusters**.

Clustering

Cluster-robust standard errors

Imagine we have G clusters with some unknown dependence between observations within a cluster and independence between clusters.

Then we can ignore $\mathbf{x}'_i \mathbf{x}_j E[\varepsilon_j \varepsilon_i | \mathbf{X}]$ if i and j are in different clusters.

Clustering

Cluster-robust standard errors

Imagine we have G clusters with some unknown **dependence between observations within a cluster** and **independence between clusters**.

Then we can ignore $\mathbf{x}'_i \mathbf{x}_j E[\varepsilon_j \varepsilon_i | \mathbf{X}]$ if i and j are in **different clusters**.

We can estimate $\sum_i \sum_j \mathbf{x}'_i \mathbf{x}_j E[\varepsilon_j \varepsilon_i | \mathbf{X}]$ with

$$\sum_{g=1}^G \left(\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \mathbf{x}'_i \mathbf{x}_j e_j e_i \right) = \sum_{g=1}^G \mathbf{X}'_g e_g e'_g \mathbf{X}_g$$

I.e., to learn about **within-group** covariance, we calculate these **within-group** cross products and then sum over groups.[†]

[†] Group sizes can vary.

Clustering

Guidelines for group number/size

Large G , Small N_g

Clustered standard errors work well. $G > N_g$ and $G > 20$.

Large G , Large N_g

We might be concerned about the number of within-group cross terms here. However, for moderately large G (50?), cluster-robust standard errors appear to perform well with large N_g .

Small G , Large N_g

Cluster-robust standard errors do not work well (definitely $G < 10$).

Options Collapse groups? Wild clustered bootstrap?

Small G , Small N_g

Essentially the same issues and solutions as small G with large N_g .

Clustering

Further extensions

We've discussed the standard cluster-robust variance-covariance estimator.

Multi-way clustering allows multiple levels/dimensions in which individuals are *clustered*.

- For *nested clusters* (e.g., state and county), people commonly cluster at the highest (largest) unit.
- For *non-nested clusters* (e.g., state and year), **Cameron, Gelbach, and Miller (2011)** provide a covariance estimator

$$\mathbf{Var}(\hat{\beta}) = \mathbf{Var}_{\text{State}}(\hat{\beta}) + \mathbf{Var}_{\text{Year}}(\hat{\beta}) - \mathbf{Var}_{\text{State-Year}}(\hat{\beta})$$

where $\mathbf{Var}_{\text{State}}(\hat{\beta})$ denotes the covariance of $\hat{\beta}$ clustered by state.

Clustering

Further extensions

We've discussed the standard cluster-robust variance-covariance estimator.

The term **Conley standard errors** is often used to describe situations in which you have spatial clustering/correlation that you can describe via a function like spatial distance.[†]

See [Conley \(1999\)](#) for the paper and [this blog by Dan Christensen and Thiemo Fetzer](#) for practical implementation in R and Stata.

[†] They also are robust to heteroskedasticity and autocorrelation within units.

Clustering

Cluster-robust standard errors

So now you know what `lm_robust()`, `iv_robust()`, *etc.* are doing when you specify a variable for clustering (*e.g.*, `clusters = var`).

Clustering

Cluster-robust standard errors

So now you know what `lm_robust()`, `iv_robust()`, etc. are doing when you specify a variable for clustering (e.g., `clusters = var`).

`lm_robust()` **without clustering**

```
# Estimate without clusters
vote_no ← lm_robust(
  voteA ~ expendA + expendB,
  fixed_effects = state,
  data = wooldridge::vote1
)
```

`lm_robust()` **with clustering**

```
# Estimate with clusters
vote_cl ← lm_robust(
  voteA ~ expendA + expendB,
  fixed_effects = state,
  clusters = state,
  data = wooldridge::vote1
)
```

Clustering

Cluster-robust standard errors

Alternatives for clustering: `fe1m()` from `lfe` and `feols()` from `fixest`.

Clustering

Cluster-robust standard errors

Alternatives for clustering: `feelm()` from `lfe` and `feols()` from `fixest`.

`feelm()` clustering by state

```
# Estimate with clusters
est_felm = feelm(
  voteA ~ expendA + expendB |
  state |
  0 |
  state,
  data = wooldridge::vote1
)
```

`feols()` clustering by state

```
# Estimate with clusters
est_feols = feols(
  voteA ~ expendA + expendB |
  state,
  data = wooldridge::vote1
)
# Force cluster-rob. SEs
summary(
  est_feols,
  se = "cluster",
  cluster = "state"
)
```

Time for a simulation.

Cluster simulation

Cluster simulation

The DGP

Let's opt for a simple-ish example.[†]

$$y_{ig} = (\beta_0 = 1) + (\beta_1 = 2) x_{1,g} + (\beta_2 = 0) x_{2,g} + \varepsilon_{ig}$$

$$\varepsilon_{ig} = \nu_g + \eta_i$$

where the $\eta_i \perp \eta_j$, $\eta_i \perp \nu_g$, and $\nu_g \perp \nu_h$.

[†] So we have more room for problem sets/exams.

Cluster simulation

The DGP

Let's opt for a simple-ish example.[†]

$$y_{ig} = (\beta_0 = 1) + (\beta_1 = 2) x_{1,g} + (\beta_2 = 0) x_{2,g} + \varepsilon_{ig}$$
$$\varepsilon_{ig} = \nu_g + \eta_i$$

where the $\eta_i \perp \eta_j$, $\eta_i \perp \nu_g$, and $\nu_g \perp \nu_h$.

Let's assume $\eta_i \sim N(0, 1)$ and $\nu_g \sim N(0, 1)$. And $x_g \sim N(0, 1)$.

Plus $N_g = 100$ with 10 groups.

[†] So we have more room for problem sets/exams.

Cluster simulation

The DGP

Let's opt for a simple-ish example.[†]

$$y_{ig} = (\beta_0 = 1) + (\beta_1 = 2) x_{1,g} + (\beta_2 = 0) x_{2,g} + \varepsilon_{ig}$$
$$\varepsilon_{ig} = \nu_g + \eta_i$$

where the $\eta_i \perp \eta_j$, $\eta_i \perp \nu_g$, and $\nu_g \perp \nu_h$.

Let's assume $\eta_i \sim N(0, 1)$ and $\nu_g \sim N(0, 1)$. And $x_g \sim N(0, 1)$.

Plus $N_g = 100$ with 10 groups.

Note Small G with large-ish N_g .

[†] So we have more room for problem sets/exams.

First we need to write the **data generating process for one iteration**.

```
# The DGP
sim_dgp ← function(n = 100, n_grps = 10, σv = 1, ση = 1) {
  # Create the right number of observations
  sample_df ← expand.grid(i = 1:n, g = 1:n_grps) %>% as_tibble()
  # Create a unique ID (from 1 to number of observations)
  sample_df %<>% mutate(id = 1:(n * n_grps))
  # Sample v at the group level (NOTE: DON'T FORGET TO UNGROUP)
  sample_df %<>% group_by(g) %>%
    mutate(v = rnorm(1, sd = σv)) %>% ungroup()
  # Sample η at the individual level
  sample_df %<>% mutate(η = rnorm(n * n_grps, sd = ση))
  # Sample x_g from N(0,1)
  sample_df %<>% group_by(g) %>%
    mutate(x1 = rnorm(1), x2 = rnorm(1)) %>% ungroup()
  # Calculate y
  sample_df %<>% mutate(y = 1 + 2 * x1 + 0 * x2 + v + η)
  # Return
  return(sample_df)
}
```

Now we **analyze** the data within one iteration.

```
# Analyze 'data'
sim_analyze ← function(data) {
  # Conventional SEs
  result_ols ← lm_robust(
    y ~ x1 + x2, data = data, se_type = "classical"
  ) %>% tidy() %>% filter(term %in% c("x1", "x2")) %>% select(1:5) %>%
  mutate(type = "conventional")
  # Cluster-robust SEs
  result_cl ← lm_robust(
    y ~ x1 + x2, data = data, clusters = g
  ) %>% tidy() %>% filter(term %in% c("x1", "x2")) %>% select(1:5) %>%
  mutate(type = "clustered")
  # Bind results together and add column for standard errors
  results_df ← bind_rows(result_ols, result_cl)
  # Return results
  return(results_df)
}
```

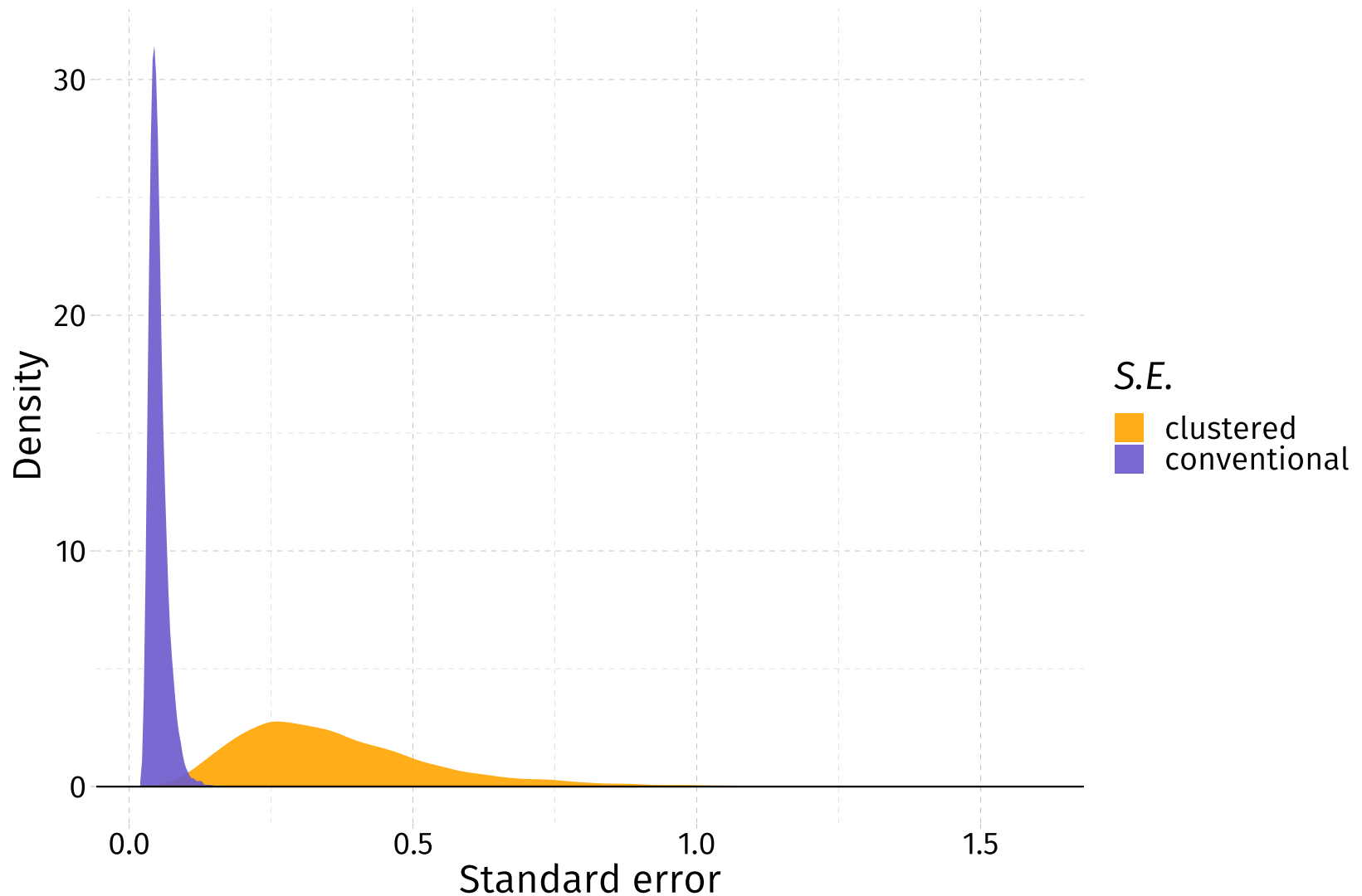
Now put the pieces together.

```
# Join sim_dgp and sim_analyze
sim_iter ← function(n = 100, n_grps = 10,  $\sigma_v$  = 1,  $\sigma_\eta$  = 1) {
  # Run the analysis in sim_analyze on the output of sim_dgp
  sim_dgp(n = 100, n_grps = 10,  $\sigma_v$  = 1,  $\sigma_\eta$  = 1) %>% sim_analyze()
}
```

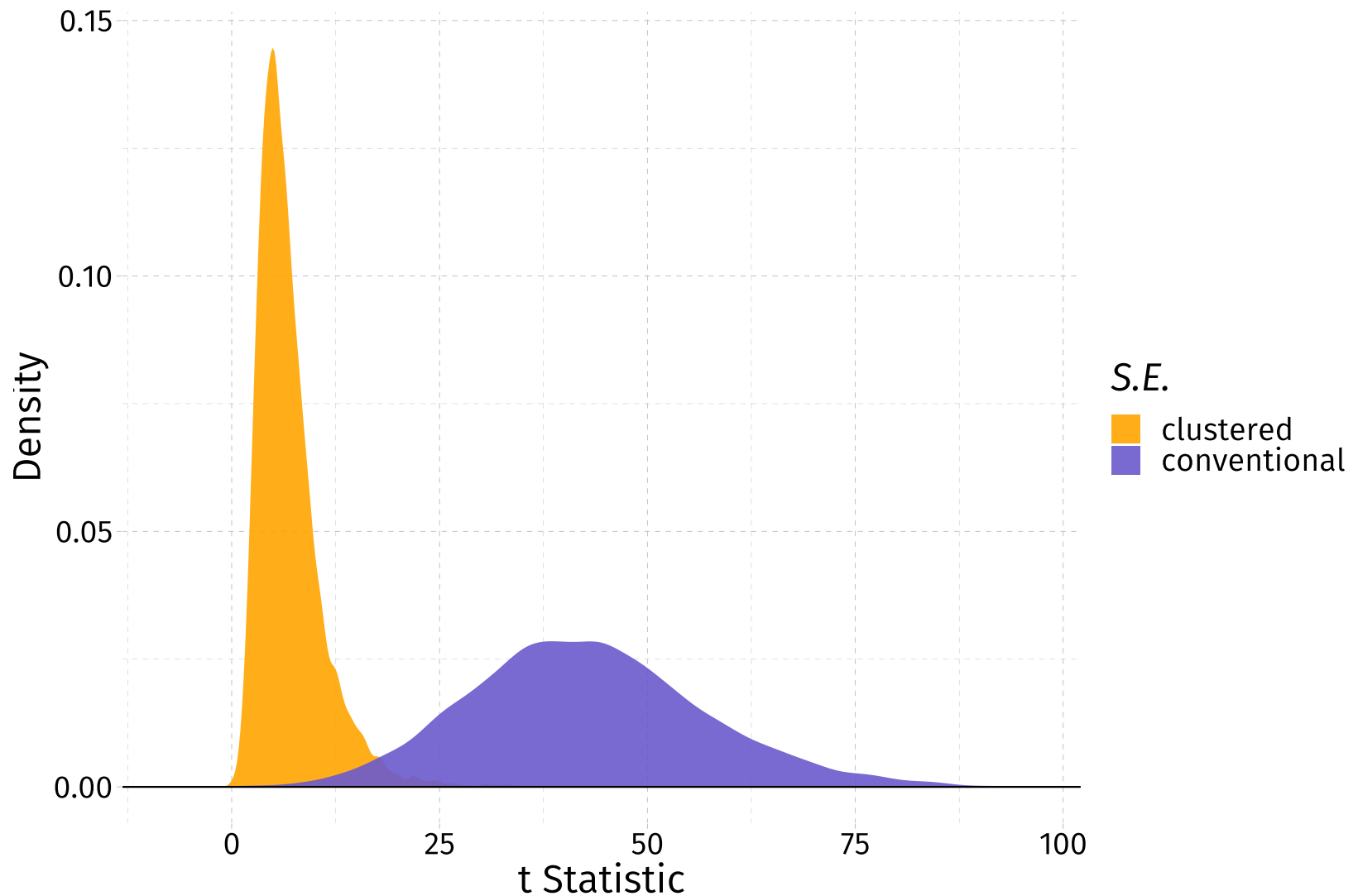
And we **run the simulation** (10,000 times).

```
# Load and set up furrr
p_load(furrr)
plan(multiprocess, workers = 10)
# Set a seed
set.seed(1234)
# Run the simulation 1e4 times
sim_df ← future_map_dfr(
  # Repeat sample size 100 for 1e4 times
  rep(100, 1e4),
  # Our function
  sim_iter,
  # Let furrr know we want to set a seed
  .options = future_options(seed = T)
)
```

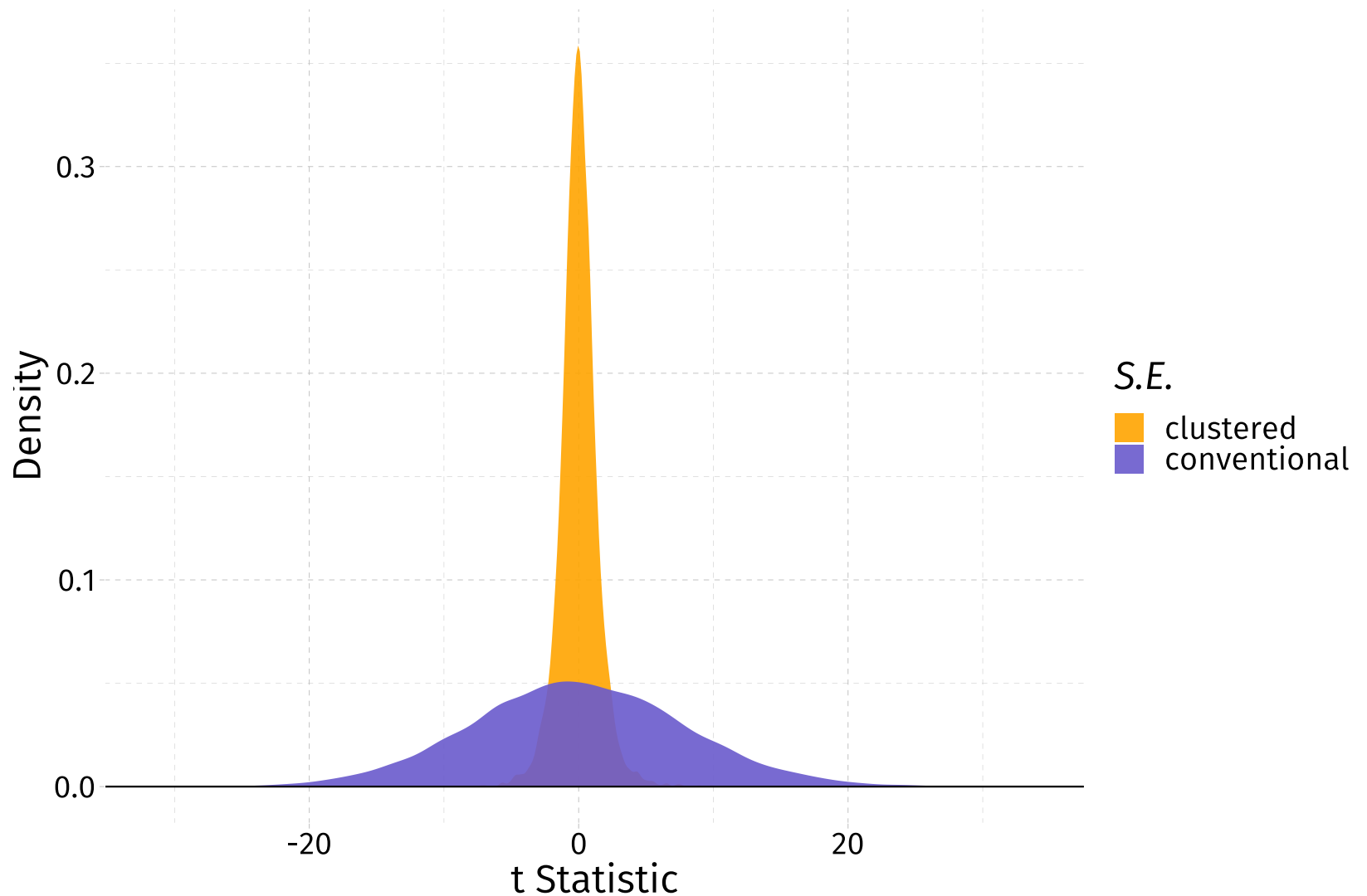
Comparing standard errors for $\hat{\beta}_1$ (coefficient on x_1)



Comparing t statistics for $\hat{\beta}_1$ (coefficient on x_1)



Comparing t statistics for $\hat{\beta}_2$ (coefficient on x_2)



Rejection rates

x1	clustered	0.878
x1	conventional	0.999
x2	clustered	0.0371
x2	conventional	0.801

1. We definitely can see the **need for clustering**.

Conventional standard errors are rejecting a **true** H_0 80% of the time.

2. **Cluster-robust standard errors are struggling** a bit in this situation.

Small G ; large N_g . Rejecting **false** H_0 88% and **true** H_0 3.7% of the time.

Resources from the literature

When Should You Adjust Standard Errors for Clustering?

Abadie, Athey, Imbens, and Wooldridge

A Practitioner's Guide to Cluster-Robust Inference

Cameron and Miller (2015)

Robust Inference With Multiway Clustering

Cameron, Gelbach, and Miller (2011)

Bootstrap-Based Improvements for Inference with Clustered Errors

Cameron, Gelbach, and Miller (2008)

How Much Should We Trust Differences-In-Differences Estimates?

Bertrand, Duflo, and Mullainathan (2004)

Table of contents

Inference

1. Motivation
2. Clustering
3. Moulton factors
 - Example
 - Test statistics
4. Answers
5. Cluster-robust S.E.s
6. Clustering extensions
7. Simulation
8. Resources