13. Properties of Least Squares

Fall 2023

Matthew Blackwell

Gov 2002 (Harvard)

Where are we? Where are we going?

- Before: learned about CEFs and linear projections in the population.

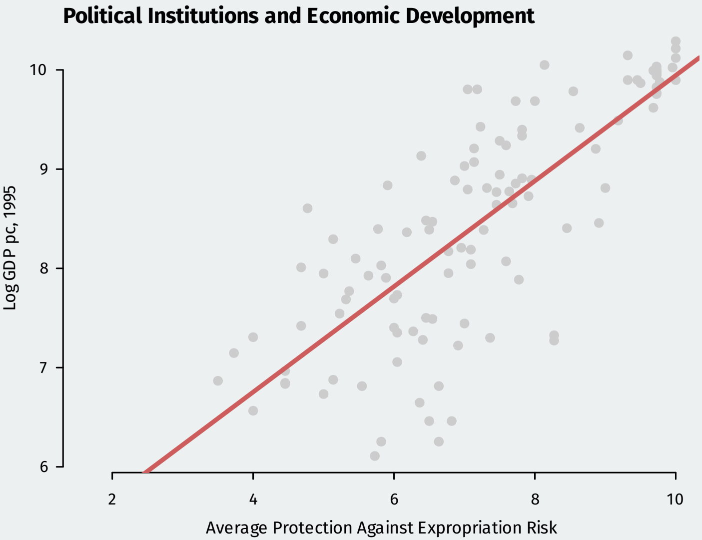
Where are we? Where are we going?

- Before: learned about CEFs and linear projections in the population.
- Last time: OLS estimator, its algebraic properties.

Where are we? Where are we going?

- Before: learned about CEFs and linear projections in the population.
- Last time: OLS estimator, its algebraic properties.
- Now: its statistical properties, both finite-sample and asymptotic.

Acemoglu, Johnson, and Robinson (2001)



Sampling distribution of the OLS estimator

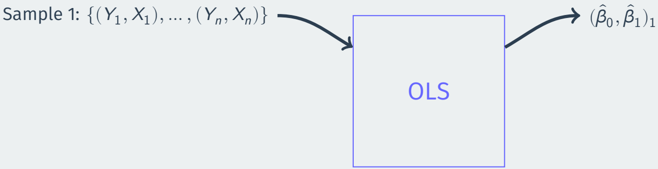
- OLS is an estimator—we plug data into and we get out estimates.

Sampling distribution of the OLS estimator

- OLS is an estimator—we plug data into and we get out estimates.

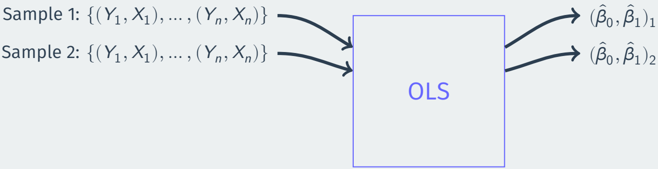
Sampling distribution of the OLS estimator

- OLS is an estimator—we plug data into and we get out estimates.



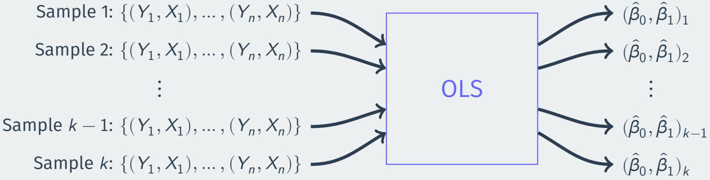
Sampling distribution of the OLS estimator

- OLS is an estimator—we plug data into and we get out estimates.



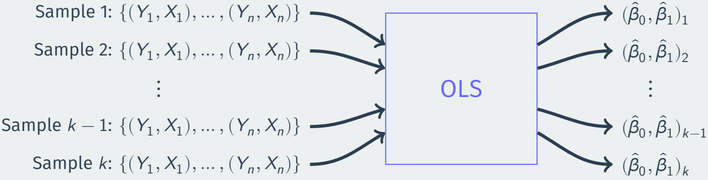
Sampling distribution of the OLS estimator

- OLS is an estimator—we plug data into and we get out estimates.



Sampling distribution of the OLS estimator

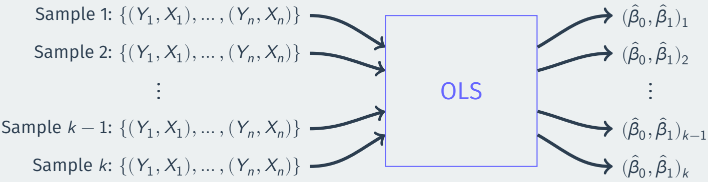
- OLS is an estimator—we plug data into and we get out estimates.



- Just like the sample mean or sample difference in means

Sampling distribution of the OLS estimator

- OLS is an estimator—we plug data into and we get out estimates.



- Just like the sample mean or sample difference in means
- Has a sampling distribution, with a sampling variance/standard error.

Simulation procedure

- Let's take a simulation approach to demonstrate:

Simulation procedure

- Let's take a simulation approach to demonstrate:
    - Pretend that the AJR data represents the population of interest

Simulation procedure

- Let's take a simulation approach to demonstrate:
    - Pretend that the AJR data represents the population of interest
    - See how the line varies from sample to sample

Simulation procedure

- Let's take a simulation approach to demonstrate:
  - Pretend that the AJR data represents the population of interest
  - See how the line varies from sample to sample

1. Draw a random sample of size $n = 30$ with replacement using `sample()`

Simulation procedure

- Let's take a simulation approach to demonstrate:
  - Pretend that the AJR data represents the population of interest
  - See how the line varies from sample to sample

1. Draw a random sample of size $n = 30$ with replacement using `sample()`
2. Use `lm()` to calculate the OLS estimates of the slope and intercept

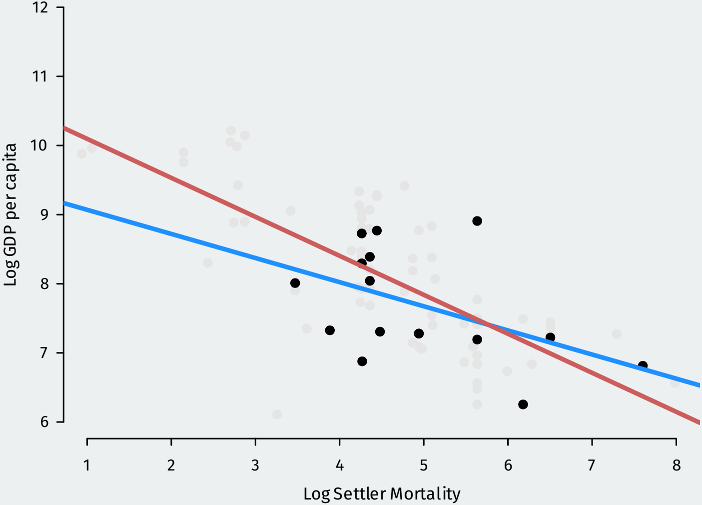
Simulation procedure

- Let's take a simulation approach to demonstrate:
  - Pretend that the AJR data represents the population of interest
  - See how the line varies from sample to sample

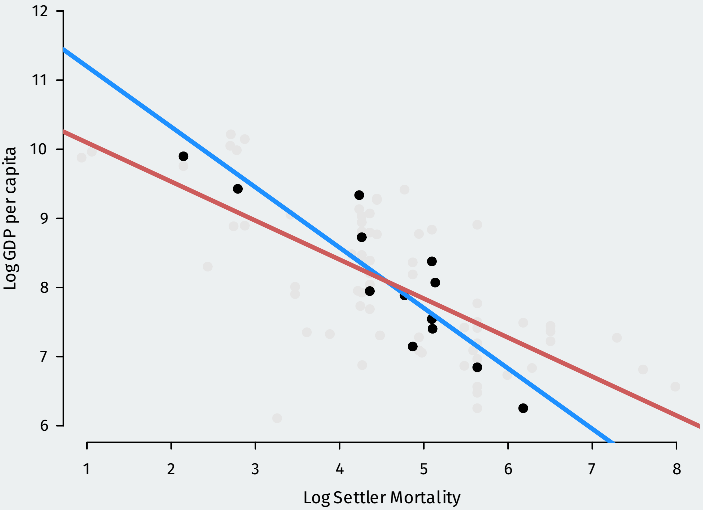
1. Draw a random sample of size $n = 30$ with replacement using `sample()`
2. Use `lm()` to calculate the OLS estimates of the slope and intercept
3. Plot the estimated regression line

Population Regression

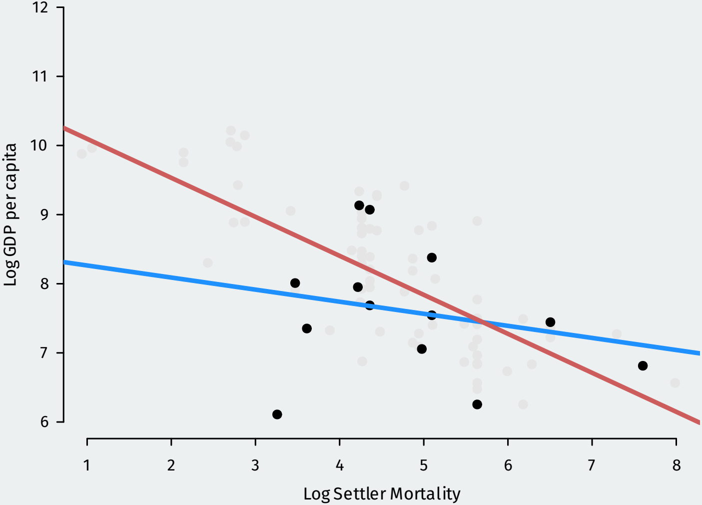
Randomly sample from AJR



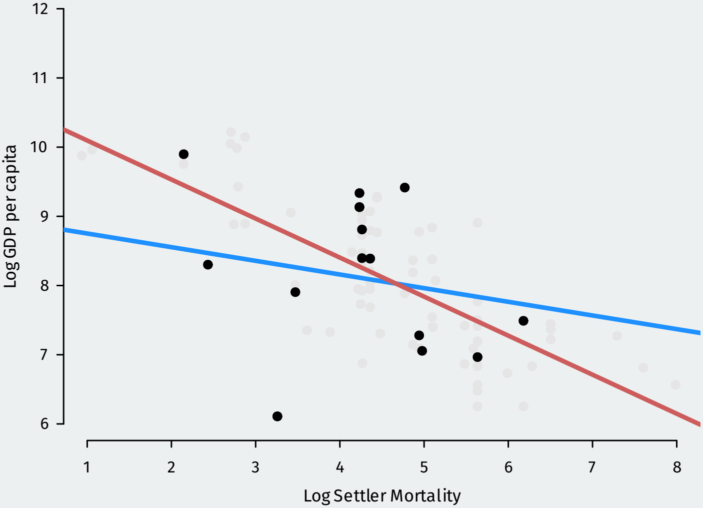
Randomly sample from AJR



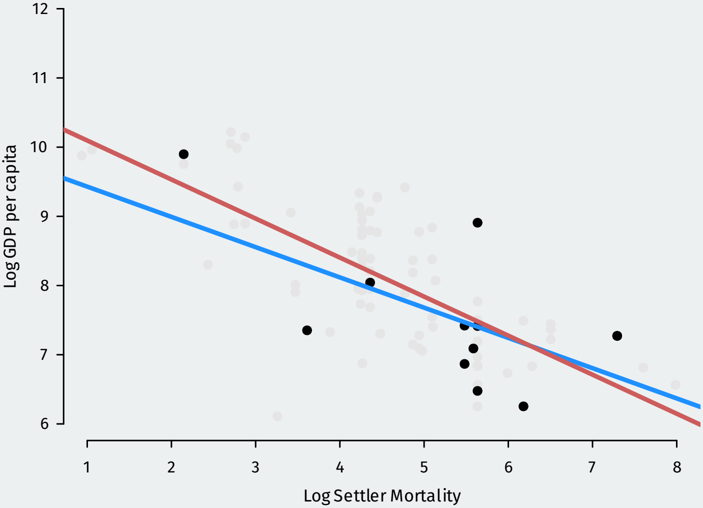
Randomly sample from AJR



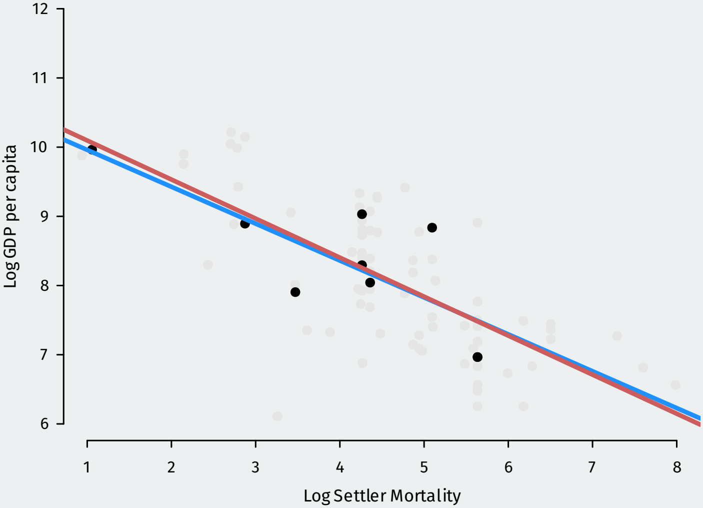
Randomly sample from AJR



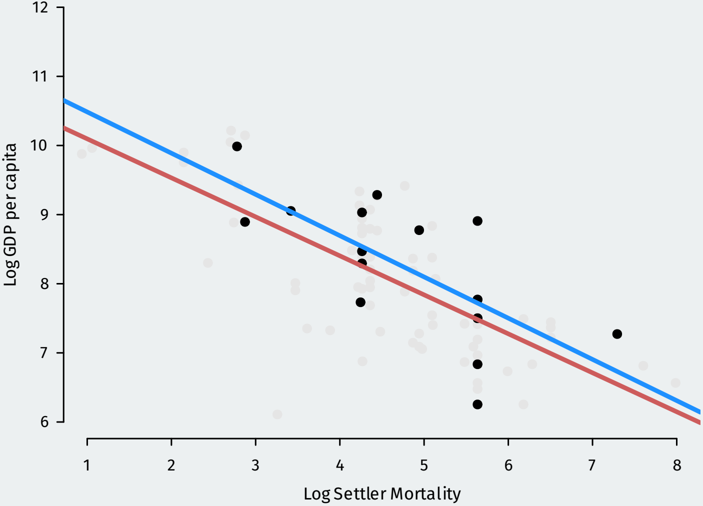
Randomly sample from AJR



Randomly sample from AJR



Randomly sample from AJR



Big picture

- We want finite-sample guarantees about our estimates.

Big picture

- We want finite-sample guarantees about our estimates.
    - Unbiasedness, exact sampling distribution, etc.

Big picture

- We want finite-sample guarantees about our estimates.
    - Unbiasedness, exact sampling distribution, etc.
- But finite-sample results come at a price in terms of assumptions.

Big picture

- We want finite-sample guarantees about our estimates.
    - Unbiasedness, exact sampling distribution, etc.
- But finite-sample results come at a price in terms of assumptions.
    - Unbiasedness: CEF is linear.

Big picture

- We want finite-sample guarantees about our estimates.
    - Unbiasedness, exact sampling distribution, etc.
- But finite-sample results come at a price in terms of assumptions.
    - Unbiasedness: CEF is linear.
    - Exact sampling distribution: normal errors.

Big picture

- We want finite-sample guarantees about our estimates.
    - Unbiasedness, exact sampling distribution, etc.
- But finite-sample results come at a price in terms of assumptions.
    - Unbiasedness: CEF is linear.
    - Exact sampling distribution: normal errors.
- Asymptotic results hold under much weaker assumptions, but require more data.

Big picture

- We want finite-sample guarantees about our estimates.
    - Unbiasedness, exact sampling distribution, etc.
- But finite-sample results come at a price in terms of assumptions.
    - Unbiasedness: CEF is linear.
    - Exact sampling distribution: normal errors.
- Asymptotic results hold under much weaker assumptions, but require more data.
    - OLS consistent for the linear projection even with nonlinear CEF.

Big picture

- We want finite-sample guarantees about our estimates.
    - Unbiasedness, exact sampling distribution, etc.
- But finite-sample results come at a price in terms of assumptions.
    - Unbiasedness: CEF is linear.
    - Exact sampling distribution: normal errors.
- Asymptotic results hold under much weaker assumptions, but require more data.
    - OLS consistent for the linear projection even with nonlinear CEF.
    - Asymptotic normality for sampling distribution under mild assumptions.

Big picture

- We want finite-sample guarantees about our estimates.
    - Unbiasedness, exact sampling distribution, etc.
- But finite-sample results come at a price in terms of assumptions.
    - Unbiasedness: CEF is linear.
    - Exact sampling distribution: normal errors.
- Asymptotic results hold under much weaker assumptions, but require more data.
    - OLS consistent for the linear projection even with nonlinear CEF.
    - Asymptotic normality for sampling distribution under mild assumptions.
- Focus on two models:

Big picture

- We want finite-sample guarantees about our estimates.
    - Unbiasedness, exact sampling distribution, etc.
- But finite-sample results come at a price in terms of assumptions.
    - Unbiasedness: CEF is linear.
    - Exact sampling distribution: normal errors.
- Asymptotic results hold under much weaker assumptions, but require more data.
    - OLS consistent for the linear projection even with nonlinear CEF.
    - Asymptotic normality for sampling distribution under mild assumptions.
- Focus on two models:
    - **Linear projection model** for asymptotic results.

Big picture

- We want finite-sample guarantees about our estimates.
    - Unbiasedness, exact sampling distribution, etc.
- But finite-sample results come at a price in terms of assumptions.
    - Unbiasedness: CEF is linear.
    - Exact sampling distribution: normal errors.
- Asymptotic results hold under much weaker assumptions, but require more data.
    - OLS consistent for the linear projection even with nonlinear CEF.
    - Asymptotic normality for sampling distribution under mild assumptions.
- Focus on two models:
    - **Linear projection model** for asymptotic results.
    - **Linear regression/CEF model** for finite samples.

1/ Linear projection model and Large-sample Properties

Linear projection model

- We'll start at the most broad, fewest assumptions

Linear projection model

- We'll start at the most broad, fewest assumptions

Linear projection model

1. For the variables $(Y, \mathbf{X})$, we assume the linear projection of Y on $\mathbf{X}$ is defined as:

$$Y = \mathbf{X}'\boldsymbol{\beta} + e$$
$$\mathbb{E}[\mathbf{X}e] = 0.$$

Linear projection model

- We'll start at the most broad, fewest assumptions

Linear projection model

1. For the variables $(Y, \mathbf{X})$, we assume the linear projection of Y on $\mathbf{X}$ is defined as:
$$Y = \mathbf{X}'\boldsymbol{\beta} + e$$
$$\mathbb{E}[\mathbf{X}e] = 0.$$
2. The design matrix is invertible, so $\mathbb{E}[\mathbf{X}_i\mathbf{X}_i'] > 0$ (positive definite).

Linear projection model

- We'll start at the most broad, fewest assumptions

Linear projection model

1. For the variables $(Y, \mathbf{X})$, we assume the linear projection of Y on $\mathbf{X}$ is defined as:

$$Y = \mathbf{X}'\boldsymbol{\beta} + e$$

$$\mathbb{E}[\mathbf{X}e] = 0.$$

2. The design matrix is invertible, so $\mathbb{E}[\mathbf{X}_i\mathbf{X}_i'] > 0$ (positive definite).

- Linear projection model holds under **very** mild assumptions.

Linear projection model

- We'll start at the most broad, fewest assumptions

Linear projection model

1. For the variables $(Y, \mathbf{X})$, we assume the linear projection of Y on $\mathbf{X}$ is defined as:
$$Y = \mathbf{X}'\boldsymbol{\beta} + e$$
$$\mathbb{E}[\mathbf{X}e] = 0.$$
2. The design matrix is invertible, so $\mathbb{E}[\mathbf{X}_i\mathbf{X}_i'] > 0$ (positive definite).

- Linear projection model holds under **very** mild assumptions.
  - Remember: not even assuming linear CEF!

Linear projection model

- We'll start at the most broad, fewest assumptions

Linear projection model

1. For the variables $(Y, \mathbf{X})$, we assume the linear projection of Y on $\mathbf{X}$ is defined as:

$$Y = \mathbf{X}'\boldsymbol{\beta} + e$$

$$\mathbb{E}[\mathbf{X}e] = 0.$$

2. The design matrix is invertible, so $\mathbb{E}[\mathbf{X}_i\mathbf{X}_i'] > 0$ (positive definite).

- Linear projection model holds under **very** mild assumptions.
  - Remember: not even assuming linear CEF!
  - Implies coefficients are $\boldsymbol{\beta} = (\mathbb{E}[\mathbf{X}\mathbf{X}'])^{-1}\mathbb{E}[\mathbf{X}Y]$

Linear projection model

- We'll start at the most broad, fewest assumptions

Linear projection model

1. For the variables $(Y, \mathbf{X})$, we assume the linear projection of Y on $\mathbf{X}$ is defined as:
$$Y = \mathbf{X}'\boldsymbol{\beta} + e$$
$$\mathbb{E}[\mathbf{X}e] = 0.$$
2. The design matrix is invertible, so $\mathbb{E}[\mathbf{X}_i\mathbf{X}_i'] > 0$ (positive definite).

- Linear projection model holds under **very** mild assumptions.
  - Remember: not even assuming linear CEF!
  - Implies coefficients are $\boldsymbol{\beta} = (\mathbb{E}[\mathbf{X}\mathbf{X}'])^{-1}\mathbb{E}[\mathbf{X}Y]$
- What properties can we derive under such weak assumptions?

A very useful decomposition

$$\hat{\boldsymbol{\beta}} = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i Y_i\right) = \boldsymbol{\beta} + \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i e_i\right)}_{\text{estimation error}}$$

- OLS estimates are the truth plus some estimation error.

A very useful decomposition

$$\hat{\boldsymbol{\beta}} = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_iY_i\right) = \boldsymbol{\beta} + \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_ie_i\right)}_{\text{estimation error}}$$

- OLS estimates are the truth plus some estimation error.
- Most of what we derive about OLS comes from this view.

A very useful decomposition

$$\hat{\boldsymbol{\beta}} = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_iY_i\right) = \boldsymbol{\beta} + \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_ie_i\right)}_{\text{estimation error}}$$

- OLS estimates are the truth plus some estimation error.
- Most of what we derive about OLS comes from this view.
- Sample means in the estimation error follow the law of large numbers:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i' \xrightarrow{p} \mathbb{E}[\mathbf{X}_i\mathbf{X}_i'] \equiv \mathbf{Q}_{\mathbf{XX}} \qquad \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_ie_i \xrightarrow{p} \mathbb{E}[\mathbf{X}e] = \mathbf{0}$$

A very useful decomposition

$$\hat{\boldsymbol{\beta}} = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i Y_i\right) = \boldsymbol{\beta} + \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i e_i\right)}_{\text{estimation error}}$$

- OLS estimates are the truth plus some estimation error.
- Most of what we derive about OLS comes from this view.
- Sample means in the estimation error follow the law of large numbers:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i' \xrightarrow{p} \mathbb{E}[\mathbf{X}_i\mathbf{X}_i'] \equiv \mathbf{Q}_{\mathbf{XX}} \qquad \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i e_i \xrightarrow{p} \mathbb{E}[\mathbf{X}e] = \mathbf{0}$$

- $\mathbf{Q}_{\mathbf{XX}}$ is invertible by assumption, so by the continuous mapping theorem:

$$\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'\right)^{-1} \xrightarrow{p} \mathbf{Q}_{\mathbf{XX}}^{-1} \quad \Longrightarrow \quad \hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta} + \mathbf{Q}_{\mathbf{XX}}^{-1}\cdot\mathbf{0} = \boldsymbol{\beta},$$

Consistency of OLS

Theorem (Consistency of OLS)

Under the linear projection model and i.i.d. data, $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}$.

- Simple proof, but powerful result.

Consistency of OLS

Theorem (Consistency of OLS)

Under the linear projection model and i.i.d. data, $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}$.

- Simple proof, but powerful result.
- OLS consistently estimates the linear projection coefficients, $\boldsymbol{\beta}$.

Consistency of OLS

Theorem (Consistency of OLS)

Under the linear projection model and i.i.d. data, $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}$.

- Simple proof, but powerful result.
- OLS consistently estimates the linear projection coefficients, $\boldsymbol{\beta}$.
    - No guarantees about what the β_j represent!

Consistency of OLS

Theorem (Consistency of OLS)

Under the linear projection model and i.i.d. data, $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}$.

- Simple proof, but powerful result.
- OLS consistently estimates the linear projection coefficients, $\boldsymbol{\beta}$.
  - No guarantees about what the β_j represent!
  - Best linear approximation to $\mathbb{E}[Y \mid \mathbf{X}]$.

Consistency of OLS

Theorem (Consistency of OLS)

Under the linear projection model and i.i.d. data, $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}$.

- Simple proof, but powerful result.
- OLS consistently estimates the linear projection coefficients, $\boldsymbol{\beta}$.
    - No guarantees about what the β_j represent!
    - Best linear approximation to $\mathbb{E}[Y \mid \mathbf{X}]$.
    - If we have a linear CEF, then it's consistent for the CEF coefficients.

Consistency of OLS

Theorem (Consistency of OLS)

Under the linear projection model and i.i.d. data, $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}$.

- Simple proof, but powerful result.
- OLS consistently estimates the linear projection coefficients, $\boldsymbol{\beta}$.
  - No guarantees about what the β_j represent!
  - Best linear approximation to $\mathbb{E}[Y \mid \mathbf{X}]$.
  - If we have a linear CEF, then it's consistent for the CEF coefficients.
- Valid with no restrictions on Y: could be binary, discrete, etc.

Consistency of OLS

Theorem (Consistency of OLS)

Under the linear projection model and i.i.d. data, $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}$.

- Simple proof, but powerful result.
- OLS consistently estimates the linear projection coefficients, $\boldsymbol{\beta}$.
  - No guarantees about what the β_j represent!
  - Best linear approximation to $\mathbb{E}[Y \mid \mathbf{X}]$.
  - If we have a linear CEF, then it's consistent for the CEF coefficients.
- Valid with no restrictions on Y: could be binary, discrete, etc.
- Not guaranteed to be unbiased (unless CEF is linear, as we'll see...)

Central limit theorem, reminders

- We'll want to approximate the sampling distribution of $\hat{\boldsymbol{\beta}}$. CLT!

Central limit theorem, reminders

- We'll want to approximate the sampling distribution of $\hat{\boldsymbol{\beta}}$. CLT!
- Consider some sample mean of i.i.d. data: $n^{-1}\sum_{i=1}^{n} g(\mathbf{X}_i)$. We have:

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{X}_i)\right] = \mathbb{E}[g(\mathbf{X}_i)] \qquad \text{var}\left[\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{X}_i)\right] = \frac{\text{var}[g(\mathbf{X}_i)]}{n}$$

Central limit theorem, reminders

- We'll want to approximate the sampling distribution of $\hat{\boldsymbol{\beta}}$. CLT!
- Consider some sample mean of i.i.d. data: $n^{-1}\sum_{i=1}^{n} g(\mathbf{X}_i)$. We have:

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{X}_i)\right] = \mathbb{E}[g(\mathbf{X}_i)] \qquad \text{var}\left[\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{X}_i)\right] = \frac{\text{var}[g(\mathbf{X}_i)]}{n}$$

- CLT implies:

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{X}_i) - \mathbb{E}[g(\mathbf{X}_i)]\right) \xrightarrow{d} \mathcal{N}(0, \text{var}[g(\mathbf{X}_i)])$$

Central limit theorem, reminders

- We'll want to approximate the sampling distribution of $\hat{\boldsymbol{\beta}}$. CLT!
- Consider some sample mean of i.i.d. data: $n^{-1}\sum_{i=1}^{n} g(\mathbf{X}_i)$. We have:

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{X}_i)\right] = \mathbb{E}[g(\mathbf{X}_i)] \qquad \text{var}\left[\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{X}_i)\right] = \frac{\text{var}[g(\mathbf{X}_i)]}{n}$$

- CLT implies:

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{X}_i) - \mathbb{E}[g(\mathbf{X}_i)]\right) \xrightarrow{d} \mathcal{N}(0, \text{var}[g(\mathbf{X}_i)])$$

- If $\mathbb{E}[g(\mathbf{X}_i)] = 0$, then we have

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{X}_i)\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(\mathbf{X}_i) \xrightarrow{d} \mathcal{N}(0, \mathbb{E}[g(\mathbf{X}_i)g(\mathbf{X}_i)'])$$

Standardized estimator

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{X}_i e_i\right)$$

Standardized estimator

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{X}_i e_i\right)$$

- Remember that $\left(n^{-1}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'\right)^{-1} \xrightarrow{p} \mathbf{Q}_{\mathbf{XX}}^{-1}$ so we have

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \approx \mathbf{Q}_{\mathbf{XX}}^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{X}_i e_i\right)$$

Standardized estimator

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{X}_i e_i\right)$$

- Remember that $(n^{-1}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i')^{-1} \xrightarrow{p} \mathbf{Q}_{\mathbf{XX}}^{-1}$ so we have

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \approx \mathbf{Q}_{\mathbf{XX}}^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{X}_i e_i\right)$$

- What about $n^{-1/2}\sum_{i=1}^{n}\mathbf{X}_i e_i$? Notice that:

Standardized estimator

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{X}_i e_i\right)$$

- Remember that $(n^{-1}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i')^{-1} \overset{p}{\to} \mathbf{Q}_{\mathbf{XX}}^{-1}$ so we have

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \approx \mathbf{Q}_{\mathbf{XX}}^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{X}_i e_i\right)$$

- What about $n^{-1/2}\sum_{i=1}^{n}\mathbf{X}_i e_i$? Notice that:
    - $n^{-1}\sum_{i=1}^{n}\mathbf{X}_i e_i$ is a sample average with $\mathbb{E}[\mathbf{X}_i e_i] = 0$.

Standardized estimator

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{X}_i e_i\right)$$

- Remember that $(n^{-1}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i')^{-1} \xrightarrow{p} \mathbf{Q}_{\mathbf{XX}}^{-1}$ so we have

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \approx \mathbf{Q}_{\mathbf{XX}}^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{X}_i e_i\right)$$

- What about $n^{-1/2}\sum_{i=1}^{n}\mathbf{X}_i e_i$? Notice that:
    - $n^{-1}\sum_{i=1}^{n}\mathbf{X}_i e_i$ is a sample average with $\mathbb{E}[\mathbf{X}_i e_i] = 0$.
    - Rewrite as $\sqrt{n}$ times an average of i.i.d. mean-zero random vectors.

Standardized estimator

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{X}_i e_i\right)$$

- Remember that $(n^{-1}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i')^{-1} \xrightarrow{p} \mathbf{Q}_{\mathbf{XX}}^{-1}$ so we have

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \approx \mathbf{Q}_{\mathbf{XX}}^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{X}_i e_i\right)$$

- What about $n^{-1/2}\sum_{i=1}^{n}\mathbf{X}_i e_i$? Notice that:
    - $n^{-1}\sum_{i=1}^{n}\mathbf{X}_i e_i$ is a sample average with $\mathbb{E}[\mathbf{X}_i e_i] = 0$.
    - Rewrite as $\sqrt{n}$ times an average of i.i.d. mean-zero random vectors.
- Let $\boldsymbol{\Omega} = \mathbb{E}[e_i^2\mathbf{X}_i\mathbf{X}_i']$ and apply the CLT:

$$\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{X}_i e_i\right) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{\Omega})$$

Asymptotic normality

Theorem (Asymptotic Normality of OLS)

Under the linear projection model,

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathcal{N}(0, \mathbf{V}_{\boldsymbol{\beta}}),$$

where,

$$\mathbf{V}_{\boldsymbol{\beta}} = \mathbf{Q}_{\mathbf{XX}}^{-1}\boldsymbol{\Omega}\mathbf{Q}_{\mathbf{XX}}^{-1} = \left(\mathbb{E}[\mathbf{X}_i\mathbf{X}_i']\right)^{-1}\mathbb{E}[e_i^2\mathbf{X}_i\mathbf{X}_i']\left(\mathbb{E}[\mathbf{X}_i\mathbf{X}_i']\right)^{-1}$$

- $\hat{\boldsymbol{\beta}}$ is approximately normal with mean $\boldsymbol{\beta}$ and variance $\mathbf{V}_{\hat{\boldsymbol{\beta}}} = \mathbf{Q}_{\mathbf{XX}}^{-1}\boldsymbol{\Omega}\mathbf{Q}_{\mathbf{XX}}^{-1}/n$

Asymptotic normality

Theorem (Asymptotic Normality of OLS)

Under the linear projection model,

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \overset{d}{\rightarrow} \mathcal{N}(0, \mathbf{V}_{\boldsymbol{\beta}}),$$

where,

$$\mathbf{V}_{\boldsymbol{\beta}} = \mathbf{Q}_{\mathbf{XX}}^{-1}\boldsymbol{\Omega}\mathbf{Q}_{\mathbf{XX}}^{-1} = \left(\mathbb{E}[\mathbf{X}_i\mathbf{X}_i']\right)^{-1}\mathbb{E}[e_i^2\mathbf{X}_i\mathbf{X}_i']\left(\mathbb{E}[\mathbf{X}_i\mathbf{X}_i']\right)^{-1}$$

- $\hat{\boldsymbol{\beta}}$ is approximately normal with mean $\boldsymbol{\beta}$ and variance $\mathbf{V}_{\hat{\boldsymbol{\beta}}} = \mathbf{Q}_{\mathbf{XX}}^{-1}\boldsymbol{\Omega}\mathbf{Q}_{\mathbf{XX}}^{-1}/n$
- $\mathbf{V}_{\hat{\boldsymbol{\beta}}} = \mathbf{V}_{\boldsymbol{\beta}}/n$ is the **asymptotic covariance matrix** of $\hat{\boldsymbol{\beta}}$

Asymptotic normality

Theorem (Asymptotic Normality of OLS)

Under the linear projection model,

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathcal{N}(0, \mathbf{V}_{\boldsymbol{\beta}}),$$

where,

$$\mathbf{V}_{\boldsymbol{\beta}} = \mathbf{Q}_{\mathbf{XX}}^{-1}\boldsymbol{\Omega}\mathbf{Q}_{\mathbf{XX}}^{-1} = \left(\mathbb{E}[\mathbf{X}_i\mathbf{X}_i']\right)^{-1}\mathbb{E}[e_i^2\mathbf{X}_i\mathbf{X}_i']\left(\mathbb{E}[\mathbf{X}_i\mathbf{X}_i']\right)^{-1}$$

- $\hat{\boldsymbol{\beta}}$ is approximately normal with mean $\boldsymbol{\beta}$ and variance $\mathbf{V}_{\hat{\boldsymbol{\beta}}} = \mathbf{Q}_{\mathbf{XX}}^{-1}\boldsymbol{\Omega}\mathbf{Q}_{\mathbf{XX}}^{-1}/n$
- $\mathbf{V}_{\hat{\boldsymbol{\beta}}} = \mathbf{V}_{\boldsymbol{\beta}}/n$ is the **asymptotic covariance matrix** of $\hat{\boldsymbol{\beta}}$
  - Square root of the diagonal of $\mathbf{V}_{\hat{\boldsymbol{\beta}}}$ = standard errors for $\hat{\beta}_j$

Asymptotic normality

Theorem (Asymptotic Normality of OLS)

Under the linear projection model,

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathcal{N}(0, \mathbf{V}_{\boldsymbol{\beta}}),$$

where,

$$\mathbf{V}_{\boldsymbol{\beta}} = \mathbf{Q}_{\mathbf{XX}}^{-1}\boldsymbol{\Omega}\mathbf{Q}_{\mathbf{XX}}^{-1} = \left(\mathbb{E}[\mathbf{X}_i\mathbf{X}_i']\right)^{-1}\mathbb{E}[e_i^2\mathbf{X}_i\mathbf{X}_i']\left(\mathbb{E}[\mathbf{X}_i\mathbf{X}_i']\right)^{-1}$$

- $\hat{\boldsymbol{\beta}}$ is approximately normal with mean $\boldsymbol{\beta}$ and variance $\mathbf{V}_{\hat{\boldsymbol{\beta}}} = \mathbf{Q}_{\mathbf{XX}}^{-1}\boldsymbol{\Omega}\mathbf{Q}_{\mathbf{XX}}^{-1}/n$
- $\mathbf{V}_{\hat{\boldsymbol{\beta}}} = \mathbf{V}_{\boldsymbol{\beta}}/n$ is the **asymptotic covariance matrix** of $\hat{\boldsymbol{\beta}}$
    - Square root of the diagonal of $\mathbf{V}_{\hat{\boldsymbol{\beta}}}$ = standard errors for $\hat{\beta}_j$
- Allows us to formulate (approximate) confidence intervals, tests.

2/ OLS variance estimation

Estimating OLS variance

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathcal{N}(0, \mathbf{V}_{\boldsymbol{\beta}}), \qquad \mathbf{V}_{\boldsymbol{\beta}} = \mathbf{Q}_{\mathbf{XX}}^{-1} \boldsymbol{\Omega} \mathbf{Q}_{\mathbf{XX}}^{-1}$$

- Estimation of $\mathbf{V}_{\boldsymbol{\beta}}$ uses plug-in estimators.

Estimating OLS variance

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathcal{N}(0, \mathbf{V}_{\boldsymbol{\beta}}), \qquad \mathbf{V}_{\boldsymbol{\beta}} = \mathbf{Q}_{\mathbf{XX}}^{-1} \boldsymbol{\Omega} \mathbf{Q}_{\mathbf{XX}}^{-1}$$

- Estimation of $\mathbf{V}_{\boldsymbol{\beta}}$ uses plug-in estimators.
  - Replace $\mathbf{Q}_{\mathbf{XX}} = \mathbb{E}[\mathbf{X}_i \mathbf{X}_i']$ with $\widehat{\mathbf{Q}}_{\mathbf{XX}} = n^{-1} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i' = \mathbb{X}'\mathbb{X}/n$.

Estimating OLS variance

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathcal{N}(0, \mathbf{V}_{\boldsymbol{\beta}}), \qquad \mathbf{V}_{\boldsymbol{\beta}} = \mathbf{Q}_{\mathbf{XX}}^{-1}\boldsymbol{\Omega}\mathbf{Q}_{\mathbf{XX}}^{-1}$$

- Estimation of $\mathbf{V}_{\boldsymbol{\beta}}$ uses plug-in estimators.
  - Replace $\mathbf{Q}_{\mathbf{XX}} = \mathbb{E}[\mathbf{X}_i\mathbf{X}_i']$ with $\widehat{\mathbf{Q}}_{\mathbf{XX}} = n^{-1}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i' = \mathbb{X}'\mathbb{X}/n$.
  - Replace $\boldsymbol{\Omega} = \mathbb{E}[e_i^2\mathbf{X}_i\mathbf{X}_i']$ with $\widehat{\boldsymbol{\Omega}} = n^{-1}\sum_{i=1}^{n}\hat{e}_i^2\mathbf{X}_i\mathbf{X}_i'$

Estimating OLS variance

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathcal{N}(0, \mathbf{V}_{\boldsymbol{\beta}}), \qquad \mathbf{V}_{\boldsymbol{\beta}} = \mathbf{Q}_{\mathbf{XX}}^{-1}\boldsymbol{\Omega}\mathbf{Q}_{\mathbf{XX}}^{-1}$$

- Estimation of $\mathbf{V}_{\boldsymbol{\beta}}$ uses plug-in estimators.
  - Replace $\mathbf{Q}_{\mathbf{XX}} = \mathbb{E}[\mathbf{X}_i\mathbf{X}_i']$ with $\widehat{\mathbf{Q}}_{\mathbf{XX}} = n^{-1}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i' = \mathbb{X}'\mathbb{X}/n$.
  - Replace $\boldsymbol{\Omega} = \mathbb{E}[e_i^2\mathbf{X}_i\mathbf{X}_i']$ with $\widehat{\boldsymbol{\Omega}} = n^{-1}\sum_{i=1}^n \hat{e}_i^2\mathbf{X}_i\mathbf{X}_i'$
- Putting these together to get a **consistent** estimator:

$$\widehat{\mathbf{V}}_{\boldsymbol{\beta}} = \left(\frac{1}{n}\mathbb{X}'\mathbb{X}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n \hat{e}_i^2\mathbf{X}_i\mathbf{X}_i'\right)\left(\frac{1}{n}\mathbb{X}'\mathbb{X}\right)^{-1} \xrightarrow{p} \mathbf{V}_{\boldsymbol{\beta}}$$

Estimating OLS variance

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathcal{N}(0, \mathbf{V}_{\boldsymbol{\beta}}), \qquad \mathbf{V}_{\boldsymbol{\beta}} = \mathbf{Q}_{\mathbf{XX}}^{-1}\boldsymbol{\Omega}\mathbf{Q}_{\mathbf{XX}}^{-1}$$

- Estimation of $\mathbf{V}_{\boldsymbol{\beta}}$ uses plug-in estimators.
  - Replace $\mathbf{Q}_{\mathbf{XX}} = \mathbb{E}[\mathbf{X}_i\mathbf{X}_i']$ with $\widehat{\mathbf{Q}}_{\mathbf{XX}} = n^{-1}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i' = \mathbb{X}'\mathbb{X}/n$.
  - Replace $\boldsymbol{\Omega} = \mathbb{E}[e_i^2\mathbf{X}_i\mathbf{X}_i']$ with $\widehat{\boldsymbol{\Omega}} = n^{-1}\sum_{i=1}^{n}\hat{e}_i^2\mathbf{X}_i\mathbf{X}_i'$
- Putting these together to get a **consistent** estimator:

$$\widehat{\mathbf{V}}_{\boldsymbol{\beta}} = \left(\frac{1}{n}\mathbb{X}'\mathbb{X}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{e}_i^2\mathbf{X}_i\mathbf{X}_i'\right)\left(\frac{1}{n}\mathbb{X}'\mathbb{X}\right)^{-1} \xrightarrow{p} \mathbf{V}_{\boldsymbol{\beta}}$$

- Approximate variance of the coefficients:

$$\widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}} = \frac{1}{n}\widehat{\mathbf{V}}_{\boldsymbol{\beta}} = (\mathbb{X}'\mathbb{X})^{-1}\left(\sum_{i=1}^{n}\hat{e}_i^2\mathbf{X}_i\mathbf{X}_i'\right)(\mathbb{X}'\mathbb{X})^{-1}$$

Estimating OLS variance

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathcal{N}(0, \mathbf{V}_{\boldsymbol{\beta}}), \qquad \mathbf{V}_{\boldsymbol{\beta}} = \mathbf{Q}_{\mathbf{XX}}^{-1}\boldsymbol{\Omega}\mathbf{Q}_{\mathbf{XX}}^{-1}$$

- Estimation of $\mathbf{V}_{\boldsymbol{\beta}}$ uses plug-in estimators.
  - Replace $\mathbf{Q}_{\mathbf{XX}} = \mathbb{E}[\mathbf{X}_i\mathbf{X}_i']$ with $\widehat{\mathbf{Q}}_{\mathbf{XX}} = n^{-1}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i' = \mathbb{X}'\mathbb{X}/n$.
  - Replace $\boldsymbol{\Omega} = \mathbb{E}[e_i^2\mathbf{X}_i\mathbf{X}_i']$ with $\widehat{\boldsymbol{\Omega}} = n^{-1}\sum_{i=1}^{n}\hat{e}_i^2\mathbf{X}_i\mathbf{X}_i'$
- Putting these together to get a **consistent** estimator:

$$\widehat{\mathbf{V}}_{\boldsymbol{\beta}} = \left(\frac{1}{n}\mathbb{X}'\mathbb{X}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{e}_i^2\mathbf{X}_i\mathbf{X}_i'\right)\left(\frac{1}{n}\mathbb{X}'\mathbb{X}\right)^{-1} \xrightarrow{p} \mathbf{V}_{\boldsymbol{\beta}}$$

- Approximate variance of the coefficients:

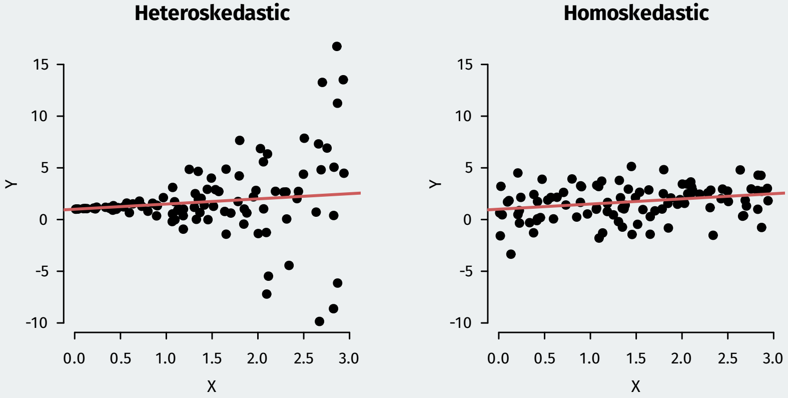
$$\widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}} = \frac{1}{n}\widehat{\mathbf{V}}_{\boldsymbol{\beta}} = (\mathbb{X}'\mathbb{X})^{-1}\left(\sum_{i=1}^{n}\hat{e}_i^2\mathbf{X}_i\mathbf{X}_i'\right)(\mathbb{X}'\mathbb{X})^{-1}$$

- Square root of the diagonal of $\widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}}$: **heteroskedasticity-consistent (HC) SEs** (aka “robust SEs”)

Homoskedasticity

Assumption: Homoskedasticity

The variance of the error terms is constant in $\mathbf{X}$, $\mathbb{E}[e^2 \mid \mathbf{X}] = \sigma^2(\mathbf{X}) = \sigma^2$.



Consequences of homoskedasticity

- Homoskedasticity implies $\mathbb{E}[e_i^2 \mathbf{X}_i \mathbf{X}_i'] = \mathbb{E}[e_i^2]\mathbb{E}[\mathbf{X}_i \mathbf{X}_i'] = \sigma^2 \mathbf{Q}_{\mathbf{XX}}$

Consequences of homoskedasticity

- Homoskedasticity implies $\mathbb{E}[e_i^2 \mathbf{X}_i \mathbf{X}_i'] = \mathbb{E}[e_i^2]\mathbb{E}[\mathbf{X}_i \mathbf{X}_i'] = \sigma^2 \mathbf{Q}_{\mathbf{XX}}$
- Simplifies the expression for the variance of $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$:

$$\mathbf{V}_{\boldsymbol{\beta}}^{\texttt{lm}} = \mathbf{Q}_{\mathbf{XX}}^{-1} \mathbb{E}[e_i^2] \mathbf{Q}_{\mathbf{XX}} \mathbf{Q}_{\mathbf{XX}}^{-1} = \sigma^2 \mathbf{Q}_{\mathbf{XX}}^{-1}$$

Consequences of homoskedasticity

- Homoskedasticity implies $\mathbb{E}[e_i^2 \mathbf{X}_i \mathbf{X}_i'] = \mathbb{E}[e_i^2]\mathbb{E}[\mathbf{X}_i \mathbf{X}_i'] = \sigma^2 \mathbf{Q}_{\mathbf{XX}}$
- Simplifies the expression for the variance of $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$:

$$\mathbf{V}_{\boldsymbol{\beta}}^{\texttt{lm}} = \mathbf{Q}_{\mathbf{XX}}^{-1} \mathbb{E}[e_i^2] \mathbf{Q}_{\mathbf{XX}} \mathbf{Q}_{\mathbf{XX}}^{-1} = \sigma^2 \mathbf{Q}_{\mathbf{XX}}^{-1}$$

- Estimated variance of $\hat{\boldsymbol{\beta}}$ under homoskedasticity

$$s^2 = \frac{1}{n-k} \sum_{i=1}^{n} \hat{e}_i^2 \qquad \widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}}^{\texttt{lm}} = \frac{1}{n} s^2 \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i' \right)^{-1} = s^2 \left(\mathbb{X}' \mathbb{X} \right)^{-1}$$

Consequences of homoskedasticity

- Homoskedasticity implies $\mathbb{E}[e_i^2 \mathbf{X}_i \mathbf{X}_i'] = \mathbb{E}[e_i^2]\mathbb{E}[\mathbf{X}_i \mathbf{X}_i'] = \sigma^2 \mathbf{Q}_{\mathbf{XX}}$
- Simplifies the expression for the variance of $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$:

$$\mathbf{V}_{\boldsymbol{\beta}}^{\texttt{lm}} = \mathbf{Q}_{\mathbf{XX}}^{-1} \mathbb{E}[e_i^2] \mathbf{Q}_{\mathbf{XX}} \mathbf{Q}_{\mathbf{XX}}^{-1} = \sigma^2 \mathbf{Q}_{\mathbf{XX}}^{-1}$$

- Estimated variance of $\hat{\boldsymbol{\beta}}$ under homoskedasticity

$$s^2 = \frac{1}{n-k} \sum_{i=1}^{n} \hat{e}_i^2 \qquad \widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}}^{\texttt{lm}} = \frac{1}{n} s^2 \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i' \right)^{-1} = s^2 \left(\mathbb{X}' \mathbb{X} \right)^{-1}$$

- LLN implies $s^2 \overset{p}{\to} \sigma^2$ and so $n \widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}}^{\texttt{lm}}$ is consistent for $\mathbf{V}_{\boldsymbol{\beta}}^{\texttt{lm}}$

Notes on skedasticity

- Homoskedasticity: strong assumption that isn't needed for consistency.

Notes on skedasticity

- Homoskedasticity: strong assumption that isn't needed for consistency.
- Software: almost always reports $\widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}}^{\texttt{lm}}$ by default.

Notes on skedasticity

- Homoskedasticity: strong assumption that isn't needed for consistency.
- Software: almost always reports $\widehat{\mathbf{V}}^{\texttt{lm}}_{\hat{\boldsymbol{\beta}}}$ by default.
    - e.g. `lm()` in R or `reg` in Stata.

Notes on skedasticity

- Homoskedasticity: strong assumption that isn't needed for consistency.
- Software: almost always reports $\widehat{\mathbf{V}}^{\texttt{lm}}_{\hat{\beta}}$ by default.
  - e.g. `lm()` in R or `reg` in Stata.
- Separate commands for HC SEs $\widehat{\mathbf{V}}_{\hat{\beta}}$

Notes on skedasticity

- Homoskedasticity: strong assumption that isn't needed for consistency.
- Software: almost always reports $\widehat{\mathbf{V}}^{\texttt{lm}}_{\hat{\boldsymbol{\beta}}}$ by default.
  - e.g. `lm()` in R or `reg` in Stata.
- Separate commands for HC SEs $\widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}}$
  - Use **`{sandwich}`** package in R or **`,robust`** in Stata.

Notes on skedasticity

- Homoskedasticity: strong assumption that isn't needed for consistency.
- Software: almost always reports $\widehat{\mathbf{V}}_{\hat{\beta}}^{\texttt{lm}}$ by default.
    - e.g. `lm()` in R or `reg` in Stata.
- Separate commands for HC SEs $\widehat{\mathbf{V}}_{\hat{\beta}}$
    - Use **`{sandwich}`** package in R or **`,robust`** in Stata.
- If $\widehat{\mathbf{V}}_{\hat{\beta}}^{\texttt{lm}}$ and $\widehat{\mathbf{V}}_{\hat{\beta}}$ differ a lot, maybe check modeling assumptions (King and Roberts, PA 2015)

Notes on skedasticity

- Homoskedasticity: strong assumption that isn't needed for consistency.
- Software: almost always reports $\widehat{\mathbf{V}}^{\texttt{lm}}_{\hat{\boldsymbol{\beta}}}$ by default.
  - e.g. `lm()` in R or `reg` in Stata.
- Separate commands for HC SEs $\widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}}$
  - Use **`{sandwich}`** package in R or **`,robust`** in Stata.
- If $\widehat{\mathbf{V}}^{\texttt{lm}}_{\hat{\boldsymbol{\beta}}}$ and $\widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}}$ differ a lot, maybe check modeling assumptions (King and Roberts, PA 2015)
- Lots of "flavors" of HC variance estimators (HC0, HC1, HC2, etc).

Notes on skedasticity

- Homoskedasticity: strong assumption that isn't needed for consistency.
- Software: almost always reports $\widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}}^{\texttt{lm}}$ by default.
    - e.g. `lm()` in R or `reg` in Stata.
- Separate commands for HC SEs $\widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}}$
    - Use **`{sandwich}`** package in R or **`,robust`** in Stata.
- If $\widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}}^{\texttt{lm}}$ and $\widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}}$ differ a lot, maybe check modeling assumptions (King and Roberts, PA 2015)
- Lots of "flavors" of HC variance estimators (HC0, HC1, HC2, etc).
    - Mostly small, ad hoc changes to improve finite-sample performance.

AJR data

```
library(sandwich)
mod <- lm(logpgp95 ~ avexpr + lat_abst + meantemp, data = ajr)
vcov(mod) ## homoskdastic V_\hat{beta}
```

```
##             (Intercept)    avexpr  lat_abst  meantemp
## (Intercept)      0.9079 -0.040952 -0.537463 -0.023246
## avexpr          -0.0410  0.004162 -0.000778  0.000605
## lat_abst        -0.5375 -0.000778  0.867588  0.016717
## meantemp        -0.0232  0.000605  0.016717  0.000705
```

```
sandwich::vcovHC(mod, type = "HC2") ## HC2
```

```
##             (Intercept)   avexpr lat_abst  meantemp
## (Intercept)      0.9764 -0.05735 -0.29548 -0.024639
## avexpr          -0.0573  0.00538 -0.00358  0.001107
## lat_abst        -0.2955 -0.00358  0.60821  0.008792
## meantemp        -0.0246  0.00111  0.00879  0.000706
```

Inference with OLS

- Inference is basically the same as any asymptotically normal estimator.

Inference with OLS

- Inference is basically the same as any asymptotically normal estimator.
- Let $\widehat{\text{se}}(\hat{\beta}_j)$ be the estimated SE for $\hat{\beta}_j$.

Inference with OLS

- Inference is basically the same as any asymptotically normal estimator.
- Let $\widehat{\text{se}}(\hat{\beta}_j)$ be the estimated SE for $\hat{\beta}_j$.
  - Square root of jth diagonal entry: $\sqrt{[\widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}}]_{jj}}$

Inference with OLS

- Inference is basically the same as any asymptotically normal estimator.
- Let $\widehat{\text{se}}(\hat{\beta}_j)$ be the estimated SE for $\hat{\beta}_j$.
    - Square root of jth diagonal entry: $\sqrt{[\widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}}]_{jj}}$
- Hypothesis test of $\beta_j = b_0$:

$$\text{general t-statistic} = \frac{\hat{\beta}_j - b_0}{\widehat{\text{se}}(\hat{\beta}_j)} \qquad \text{"usual" t-statistic} = \frac{\hat{\beta}_j}{\widehat{\text{se}}(\hat{\beta}_j)}$$

Inference with OLS

- Inference is basically the same as any asymptotically normal estimator.
- Let $\widehat{\text{se}}(\hat{\beta}_j)$ be the estimated SE for $\hat{\beta}_j$.
  - Square root of jth diagonal entry: $\sqrt{[\widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}}]_{jj}}$
- Hypothesis test of $\beta_j = b_0$:

$$\text{general t-statistic} = \frac{\hat{\beta}_j - b_0}{\widehat{\text{se}}(\hat{\beta}_j)} \qquad \text{“usual” t-statistic} = \frac{\hat{\beta}_j}{\widehat{\text{se}}(\hat{\beta}_j)}$$

  - Use same critical values from the normal as usual $z_{\alpha/2} = 1.96$.

Inference with OLS

- Inference is basically the same as any asymptotically normal estimator.
- Let $\widehat{\text{se}}(\hat{\beta}_j)$ be the estimated SE for $\hat{\beta}_j$.
    - Square root of jth diagonal entry: $\sqrt{[\widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}}]_{jj}}$
- Hypothesis test of $\beta_j = b_0$:

$$\text{general t-statistic} = \frac{\hat{\beta}_j - b_0}{\widehat{\text{se}}(\hat{\beta}_j)} \qquad \text{“usual” t-statistic} = \frac{\hat{\beta}_j}{\widehat{\text{se}}(\hat{\beta}_j)}$$

    - Use same critical values from the normal as usual $z_{\alpha/2} = 1.96$.
- 95% (asymptotic) confidence interval for $\hat{\beta}_j$:

$$\left[\hat{\beta}_j - 1.96\, \widehat{\text{se}}(\hat{\beta}_j),\ \hat{\beta}_j + 1.96\, \widehat{\text{se}}(\hat{\beta}_j)\right]$$

Inference with OLS

- Inference is basically the same as any asymptotically normal estimator.
- Let $\widehat{\text{se}}(\hat{\beta}_j)$ be the estimated SE for $\hat{\beta}_j$.
    - Square root of jth diagonal entry: $\sqrt{[\widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}}]_{jj}}$
- Hypothesis test of $\beta_j = b_0$:

$$\text{general t-statistic} = \frac{\hat{\beta}_j - b_0}{\widehat{\text{se}}(\hat{\beta}_j)} \qquad \text{“usual” t-statistic} = \frac{\hat{\beta}_j}{\widehat{\text{se}}(\hat{\beta}_j)}$$

    - Use same critical values from the normal as usual $z_{\alpha/2} = 1.96$.
- 95% (asymptotic) confidence interval for $\hat{\beta}_j$:

$$\left[\hat{\beta}_j - 1.96\,\widehat{\text{se}}(\hat{\beta}_j),\ \hat{\beta}_j + 1.96\,\widehat{\text{se}}(\hat{\beta}_j)\right]$$

- Software often uses t critical values instead of normal (we'll see why).

Inference with `lmtest::coeftest()`

```r
library(lmtest)
## homoskedastic error
lmtest::coeftest(mod)
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.9289     0.9528    7.27  1.2e-09 ***
## avexpr        0.4059     0.0645    6.29  5.1e-08 ***
## lat_abst     -0.1980     0.9314   -0.21    0.832
## meantemp     -0.0641     0.0266   -2.41    0.019 *
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
## HC2 variance estimator
lmtest::coeftest(mod, vcov = vcovHC(mod, type = "HC2"))
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.9289     0.9881    7.01  3.3e-09 ***
## avexpr        0.4059     0.0733    5.53  8.6e-07 ***
## lat_abst     -0.1980     0.7799   -0.25    0.801
## meantemp     -0.0641     0.0266   -2.41    0.019 *
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3/ Inference for Multiple Parameters

Inference for interactions

$$m(x, z) = \beta_0 + X\beta_1 + Z\beta_2 + XZ\beta_3$$

Inference for interactions

$$m(x,z) = \beta_0 + X\beta_1 + Z\beta_2 + XZ\beta_3$$

- **Partial** or **marginal** effect of X at Z: $\frac{\partial m(x,z)}{\partial x} = \beta_1 + z\beta_3$

Inference for interactions

$$m(x,z) = \beta_0 + X\beta_1 + Z\beta_2 + XZ\beta_3$$

- **Partial** or **marginal** effect of X at Z: $\frac{\partial m(x,z)}{\partial x} = \beta_1 + z\beta_3$
- Estimate it by plugging in the estimated coefficients: $\frac{\partial \widehat{m}(x,z)}{\partial x} = \hat{\beta}_1 + z\hat{\beta}_3$

Inference for interactions

$$m(x, z) = \beta_0 + X\beta_1 + Z\beta_2 + XZ\beta_3$$

- **Partial** or **marginal** effect of X at Z: $\frac{\partial m(x,z)}{\partial x} = \beta_1 + z\beta_3$
- Estimate it by plugging in the estimated coefficients: $\frac{\partial \widehat{m}(x,z)}{\partial x} = \hat{\beta}_1 + z\hat{\beta}_3$
- What if we want the variance of this effect for any value of Z?

$$\mathbb{V}\left(\frac{\partial \widehat{m}(x,z)}{\partial x}\right) = \mathbb{V}\left[\hat{\beta}_1 + z\hat{\beta}_3\right] = \mathbb{V}[\hat{\beta}_1] + z^2\mathbb{V}[\hat{\beta}_3] + 2z\text{cov}[\hat{\beta}_1, \hat{\beta}_3]$$

Inference for interactions

$$m(x, z) = \beta_0 + X\beta_1 + Z\beta_2 + XZ\beta_3$$

- **Partial** or **marginal** effect of X at Z: $\frac{\partial m(x,z)}{\partial x} = \beta_1 + z\beta_3$
- Estimate it by plugging in the estimated coefficients: $\frac{\partial \widehat{m}(x,z)}{\partial x} = \hat{\beta}_1 + z\hat{\beta}_3$
- What if we want the variance of this effect for any value of Z?

$$\mathbb{V}\left(\frac{\partial \widehat{m}(x,z)}{\partial x}\right) = \mathbb{V}\left[\hat{\beta}_1 + z\hat{\beta}_3\right] = \mathbb{V}[\hat{\beta}_1] + z^2\mathbb{V}[\hat{\beta}_3] + 2z\text{cov}[\hat{\beta}_1, \hat{\beta}_3]$$

- Use the estimated covariance matrix:

$$\hat{\mathbb{V}}\left(\frac{\partial \widehat{m}(x,z)}{\partial x}\right) = \widehat{V}_{\hat{\beta}_1} + z^2\widehat{V}_{\hat{\beta}_3} + 2z\widehat{V}_{\hat{\beta}_1\hat{\beta}_3}$$

Inference for interactions

$$m(x, z) = \beta_0 + X\beta_1 + Z\beta_2 + XZ\beta_3$$

- **Partial** or **marginal** effect of X at Z: $\frac{\partial m(x,z)}{\partial x} = \beta_1 + z\beta_3$
- Estimate it by plugging in the estimated coefficients: $\frac{\partial \widehat{m}(x,z)}{\partial x} = \widehat{\beta}_1 + z\widehat{\beta}_3$
- What if we want the variance of this effect for any value of Z?

$$\mathbb{V}\left(\frac{\partial \widehat{m}(x, z)}{\partial x}\right) = \mathbb{V}\left[\widehat{\beta}_1 + z\widehat{\beta}_3\right] = \mathbb{V}[\widehat{\beta}_1] + z^2\mathbb{V}[\widehat{\beta}_3] + 2z\text{cov}[\widehat{\beta}_1, \widehat{\beta}_3]$$

- Use the estimated covariance matrix:

$$\widehat{\mathbb{V}}\left(\frac{\partial \widehat{m}(x, z)}{\partial x}\right) = \widehat{V}_{\widehat{\beta}_1} + z^2\widehat{V}_{\widehat{\beta}_3} + 2z\widehat{V}_{\widehat{\beta}_1\widehat{\beta}_3}$$

  - $\widehat{V}_{\widehat{\beta}_1}$ is the diagonal entry of $\widehat{\mathbf{V}}_{\widehat{\boldsymbol{\beta}}}$ for $\widehat{\beta}_1$

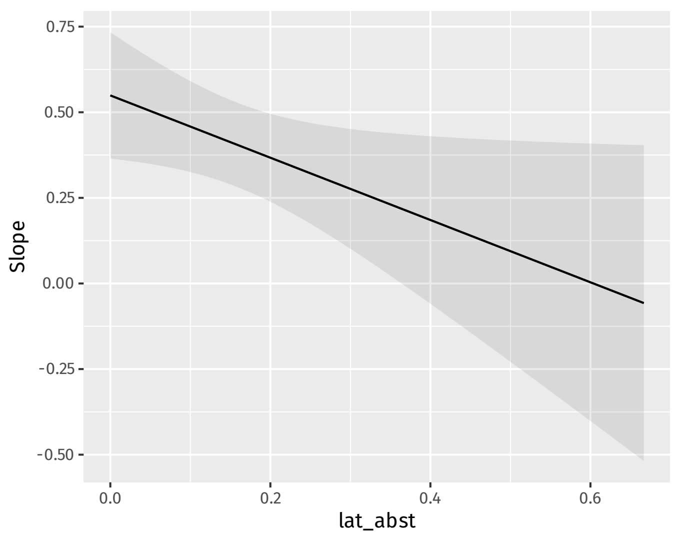
Visualizing via `marginaleffects`

```r
int_mod <- lm(logpgp95 ~ avexpr * lat_abst + meantemp, data = ajr)
coeftest(int_mod)
```

```
##
## t test of coefficients:
##
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       6.9864     0.9273    7.53    5e-10
## avexpr            0.5491     0.0941    5.84    3e-07
## lat_abst          5.8152     3.0791    1.89   0.0642
## meantemp         -0.1048     0.0326   -3.21   0.0022
## avexpr:lat_abst  -0.9095     0.4451   -2.04   0.0458
##
## (Intercept)     ***
## avexpr          ***
## lat_abst        .
## meantemp        **
## avexpr:lat_abst *
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Visualizing marginal effects

```
library(marginaleffects)
plot_slopes(int_mod, variables = "avexpr", condition = "lat_abst")
```



Tests of multiple coefficients

$$m(X, Z) = \beta_0 + X\beta_1 + Z\beta_2 + XZ\beta_3$$

- What about a test of no effect of X ever? Involves 2 coeffcients:

$$H_0 : \beta_1 = \beta_3 = 0$$

Tests of multiple coefficients

$$m(X, Z) = \beta_0 + X\beta_1 + Z\beta_2 + XZ\beta_3$$

- What about a test of no effect of X ever? Involves 2 coeffcients:

$$H_0 : \beta_1 = \beta_3 = 0$$

- Alternative: $H_1 : \beta_1 \neq 0$ or $\beta_3 \neq 0$

Tests of multiple coefficients

$$m(X,Z) = \beta_0 + X\beta_1 + Z\beta_2 + XZ\beta_3$$

- What about a test of no effect of X ever? Involves 2 coeffcients:

$$H_0 : \beta_1 = \beta_3 = 0$$

- Alternative: $H_1 : \beta_1 \neq 0$ or $\beta_3 \neq 0$
- We would like a test statistic that is large when the null is implausible.

Tests of multiple coefficients

$$m(X, Z) = \beta_0 + X\beta_1 + Z\beta_2 + XZ\beta_3$$

- What about a test of no effect of X ever? Involves 2 coeffcients:

$$H_0 : \beta_1 = \beta_3 = 0$$

- Alternative: $H_1 : \beta_1 \neq 0$ or $\beta_3 \neq 0$
- We would like a test statistic that is large when the null is implausible.
    - What about $\hat{\beta}_1^2 + \hat{\beta}_3^2$?

Tests of multiple coefficients

$$m(X, Z) = \beta_0 + X\beta_1 + Z\beta_2 + XZ\beta_3$$

- What about a test of no effect of X ever? Involves 2 coeffcients:

$$H_0 : \beta_1 = \beta_3 = 0$$

- Alternative: $H_1 : \beta_1 \neq 0$ or $\beta_3 \neq 0$
- We would like a test statistic that is large when the null is implausible.
    - What about $\widehat{\beta}_1^2 + \widehat{\beta}_3^2$?
    - Distribution depends on the variance/covariance of the coefficients.

Tests of multiple coefficients

$$m(X, Z) = \beta_0 + X\beta_1 + Z\beta_2 + XZ\beta_3$$

- What about a test of no effect of X ever? Involves 2 coeffcients:

$$H_0 : \beta_1 = \beta_3 = 0$$

- Alternative: $H_1 : \beta_1 \neq 0$ or $\beta_3 \neq 0$
- We would like a test statistic that is large when the null is implausible.
    - What about $\hat{\beta}_1^2 + \hat{\beta}_3^2$?
    - Distribution depends on the variance/covariance of the coefficients.
    - Need to normalize like the t-statistic.

Alternative test for one coefficient

- Usually t-test of $H_0 : \beta_j = b_0$ based on the t-statistic:

$$t = \frac{\hat{\beta}_j - b_0}{\widehat{\text{se}}(\hat{\beta}_j)},$$

Alternative test for one coefficient

- Usually t-test of $H_0 : \beta_j = b_0$ based on the t-statistic:

$$t = \frac{\hat{\beta}_j - b_0}{\widehat{\text{se}}(\hat{\beta}_j)},$$

  - Reject when $|t| > c$ for some critical value c from the standard normal.

Alternative test for one coefficient

- Usually t-test of $H_0 : \beta_j = b_0$ based on the t-statistic:

$$t = \frac{\hat{\beta}_j - b_0}{\widehat{\text{se}}(\hat{\beta}_j)},$$

  - Reject when $|t| > c$ for some critical value c from the standard normal.

- Equivalent test based rejects when $t^2 > c^2$

$$t^2 = \frac{\left(\hat{\beta}_j - b_0\right)^2}{\mathbb{V}[\hat{\beta}_j]} = \frac{n\left(\hat{\beta}_j - b_0\right)^2}{[\mathbf{V}_{\hat{\boldsymbol{\beta}}}]_{jj}}$$

Alternative test for one coefficient

- Usually t-test of $H_0 : \beta_j = b_0$ based on the t-statistic:

$$t = \frac{\hat{\beta}_j - b_0}{\widehat{\text{se}}(\hat{\beta}_j)},$$

  - Reject when $|t| > c$ for some critical value c from the standard normal.

- Equivalent test based rejects when $t^2 > c^2$

$$t^2 = \frac{\left(\hat{\beta}_j - b_0\right)^2}{\mathbb{V}[\hat{\beta}_j]} = \frac{n\left(\hat{\beta}_j - b_0\right)^2}{[\mathbf{V}_{\hat{\boldsymbol{\beta}}}]_{jj}}$$

- Because $t \overset{d}{\to} \mathcal{N}(0,1)$, we'll have t^2 converging to a χ^2_1 distribution

Alternative test for one coefficient

- Usually t-test of $H_0 : \beta_j = b_0$ based on the t-statistic:

$$t = \frac{\hat{\beta}_j - b_0}{\widehat{\text{se}}(\hat{\beta}_j)},$$

  - Reject when $|t| > c$ for some critical value c from the standard normal.
- Equivalent test based rejects when $t^2 > c^2$

$$t^2 = \frac{\left(\hat{\beta}_j - b_0\right)^2}{\mathbb{V}[\hat{\beta}_j]} = \frac{n\left(\hat{\beta}_j - b_0\right)^2}{[\mathbf{V}_{\hat{\boldsymbol{\beta}}}]_{jj}}$$

- Because $t \overset{d}{\to} \mathcal{N}(0,1)$, we'll have t^2 converging to a χ^2_1 distribution
  - Reminder: χ^2_k is the sum of k squared standard normals.

Alternative test for one coefficient

- Usually t-test of $H_0 : \beta_j = b_0$ based on the t-statistic:

$$t = \frac{\hat{\beta}_j - b_0}{\widehat{\text{se}}(\hat{\beta}_j)},$$

  - Reject when $|t| > c$ for some critical value c from the standard normal.
- Equivalent test based rejects when $t^2 > c^2$

$$t^2 = \frac{\left(\hat{\beta}_j - b_0\right)^2}{\mathbb{V}[\hat{\beta}_j]} = \frac{n\left(\hat{\beta}_j - b_0\right)^2}{[\mathbf{V}_{\hat{\boldsymbol{\beta}}}]_{jj}}$$

- Because $t \overset{d}{\to} \mathcal{N}(0,1)$, we'll have t^2 converging to a χ^2_1 distribution
  - Reminder: χ^2_k is the sum of k squared standard normals.
  - Could get the critical value for t^2 directly from χ^2_1.

Rewriting hypotheses with matrices

- We can rewrite the null hypothesis as $H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{c}$ where,

$$\mathbf{L} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \qquad \mathbf{c} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Rewriting hypotheses with matrices

- We can rewrite the null hypothesis as $H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{c}$ where,

$$\mathbf{L} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \qquad \mathbf{c} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

  - $\mathbf{L}$ has q rows or restriction and $k + 1$ columns (one for each coefficient)

Rewriting hypotheses with matrices

- We can rewrite the null hypothesis as $H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{c}$ where,

$$\mathbf{L} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \qquad \mathbf{c} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

  - $\mathbf{L}$ has q rows or restriction and $k + 1$ columns (one for each coefficient)

- Estimated version of the constraint: $\mathbf{L}\widehat{\boldsymbol{\beta}}$

Rewriting hypotheses with matrices

- We can rewrite the null hypothesis as $H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{c}$ where,

$$\mathbf{L} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \qquad \mathbf{c} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

  - $\mathbf{L}$ has q rows or restriction and $k + 1$ columns (one for each coefficient)

- Estimated version of the constraint: $\mathbf{L}\hat{\boldsymbol{\beta}}$

- By the Delta method, under the null hypothesis we have

$$\sqrt{n}\left(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{L}\boldsymbol{\beta}\right) \overset{d}{\to} \mathcal{N}(0, \mathbf{L}'\mathbf{V}_{\boldsymbol{\beta}}\mathbf{L}).$$

Rewriting hypotheses with matrices

- We can rewrite the null hypothesis as $H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{c}$ where,

$$\mathbf{L} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \qquad \mathbf{c} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

  - $\mathbf{L}$ has q rows or restriction and $k + 1$ columns (one for each coefficient)

- Estimated version of the constraint: $\mathbf{L}\widehat{\boldsymbol{\beta}}$

- By the Delta method, under the null hypothesis we have

$$\sqrt{n}\left(\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{L}\boldsymbol{\beta}\right) \overset{d}{\to} \mathcal{N}(0, \mathbf{L}'\mathbf{V}_{\boldsymbol{\beta}}\mathbf{L}).$$

- In this case:

$$\sqrt{n}\left(\begin{bmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_3 \end{bmatrix}\right) \overset{d}{\to} \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} [\mathbf{V}_{\boldsymbol{\beta}}]_{[11]} & [\mathbf{V}_{\boldsymbol{\beta}}]_{[13]} \\ [\mathbf{V}_{\boldsymbol{\beta}}]_{[31]} & [\mathbf{V}_{\boldsymbol{\beta}}]_{[33]} \end{bmatrix}\right)$$

Rewriting hypotheses with matrices

- We can rewrite the null hypothesis as $H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{c}$ where,

$$\mathbf{L} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \qquad \mathbf{c} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

  - $\mathbf{L}$ has q rows or restriction and $k+1$ columns (one for each coefficient)
- Estimated version of the constraint: $\mathbf{L}\widehat{\boldsymbol{\beta}}$
- By the Delta method, under the null hypothesis we have

$$\sqrt{n}\left(\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{L}\boldsymbol{\beta}\right) \xrightarrow{d} \mathcal{N}(0, \mathbf{L}'\mathbf{V}_{\boldsymbol{\beta}}\mathbf{L}).$$

- In this case:

$$\sqrt{n}\left(\begin{bmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_3 \end{bmatrix}\right) \xrightarrow{d} \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} [\mathbf{V}_{\boldsymbol{\beta}}]_{[11]} & [\mathbf{V}_{\boldsymbol{\beta}}]_{[13]} \\ [\mathbf{V}_{\boldsymbol{\beta}}]_{[31]} & [\mathbf{V}_{\boldsymbol{\beta}}]_{[33]} \end{bmatrix}\right)$$

  - If this covariance matrix where identity, then these would be standard normal and $\widehat{\beta}_1^2 + \widehat{\beta}_3^2$ would be χ^2_2 under the null

Wald statistic

- Under the null, $\sqrt{n}\left(\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{c}\right) \xrightarrow{d} \mathcal{N}(0, \mathbf{L}'\mathbf{V}_{\boldsymbol{\beta}}\mathbf{L})$

Wald statistic

- Under the null, $\sqrt{n}\left(\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{c}\right) \overset{d}{\to} \mathcal{N}(0, \mathbf{L}'\mathbf{V}_{\boldsymbol{\beta}}\mathbf{L})$
- $(\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{c})'(\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{c})$ is the squared deviations from the null.

Wald statistic

- Under the null, $\sqrt{n}\left(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{c}\right) \overset{d}{\to} \mathcal{N}(0, \mathbf{L}'\mathbf{V}_{\boldsymbol{\beta}}\mathbf{L})$
- $(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{c})'(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{c})$ is the squared deviations from the null.
    - Problem: doesn't account for variance/covariance of the estimated coefficients.

Wald statistic

- Under the null, $\sqrt{n}\left(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{c}\right) \overset{d}{\to} \mathcal{N}(0, \mathbf{L}'\mathbf{V}_{\boldsymbol{\beta}}\mathbf{L})$
- $(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{c})'(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{c})$ is the squared deviations from the null.
  - Problem: doesn't account for variance/covariance of the estimated coefficients.
- **Wald statistic** normalize by the covariance matrix:

$$W = n\left(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{c}\right)'\left(\mathbf{L}'\widehat{\mathbf{V}}_{\boldsymbol{\beta}}\mathbf{L}\right)^{-1}\left(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{c}\right)$$

Wald statistic

- Under the null, $\sqrt{n}\left(\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{c}\right) \overset{d}{\to} \mathcal{N}(0, \mathbf{L}'\mathbf{V}_{\boldsymbol{\beta}}\mathbf{L})$
- $(\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{c})'(\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{c})$ is the squared deviations from the null.
  - Problem: doesn't account for variance/covariance of the estimated coefficients.
- **Wald statistic** normalize by the covariance matrix:

$$W = n\left(\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{c}\right)'\left(\mathbf{L}'\widehat{\mathbf{V}}_{\boldsymbol{\beta}}\mathbf{L}\right)^{-1}\left(\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{c}\right)$$

  - Similar to dividing by the SE for the t-test

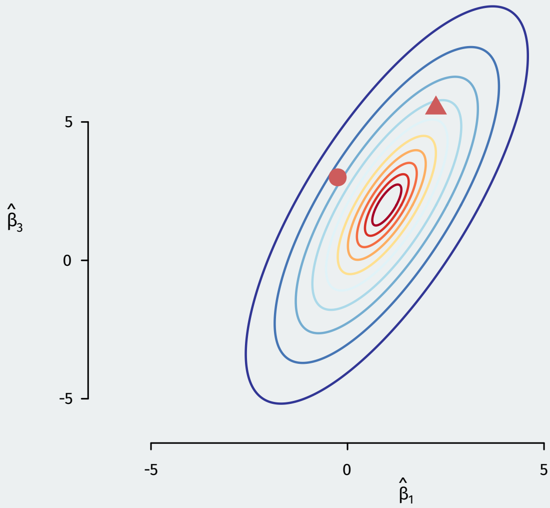
Wald statistic

- Under the null, $\sqrt{n}\left(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{c}\right) \xrightarrow{d} \mathcal{N}(0, \mathbf{L}'\mathbf{V}_{\boldsymbol{\beta}}\mathbf{L})$
- $(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{c})'(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{c})$ is the squared deviations from the null.
  - Problem: doesn't account for variance/covariance of the estimated coefficients.
- **Wald statistic** normalize by the covariance matrix:

$$W = n\left(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{c}\right)'\left(\mathbf{L}'\widehat{\mathbf{V}}_{\boldsymbol{\beta}}\mathbf{L}\right)^{-1}\left(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{c}\right)$$

  - Similar to dividing by the SE for the t-test
  - Squared distance of observed values from the null, weighted by the distribution of the parameters under the null

Weighting by the distribution



Wald test

$$W = n\left(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{c}\right)'\left(\mathbf{L}'\widehat{\mathbf{V}}_{\boldsymbol{\beta}}\mathbf{L}\right)^{-1}\left(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{c}\right)$$

- Asymptotically under the null $W \overset{d}{\to} \chi^2_q$ where q is rows of $\mathbf{L}$

Wald test

$$W = n\left(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{c}\right)'\left(\mathbf{L}'\widehat{\mathbf{V}}_{\boldsymbol{\beta}}\mathbf{L}\right)^{-1}\left(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{c}\right)$$

- Asymptotically under the null $W \xrightarrow{d} \chi^2_q$ where q is rows of $\mathbf{L}$
  - q is the number of linear restrictions in the null

Wald test

$$W = n\left(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{c}\right)'\left(\mathbf{L}'\widehat{\mathbf{V}}_{\boldsymbol{\beta}}\mathbf{L}\right)^{-1}\left(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{c}\right)$$

- Asymptotically under the null $W \overset{d}{\to} \chi^2_q$ where q is rows of $\mathbf{L}$
  - q is the number of linear restrictions in the null
- **Wald test**: reject when $W > w_\alpha$, where $\mathbb{P}(W > w_\alpha) = \alpha$ under the null.

Wald test

$$W = n\left(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{c}\right)'\left(\mathbf{L}'\widehat{\mathbf{V}}_{\boldsymbol{\beta}}\mathbf{L}\right)^{-1}\left(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{c}\right)$$

- Asymptotically under the null $W \overset{d}{\to} \chi^2_q$ where q is rows of $\mathbf{L}$
  - q is the number of linear restrictions in the null
- **Wald test**: reject when $W > w_\alpha$, where $\mathbb{P}(W > w_\alpha) = \alpha$ under the null.
  - Use χ^2_q distribution for critical values, p-values

Wald test

$$W = n\left(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{c}\right)'\left(\mathbf{L}'\widehat{\mathbf{V}}_{\boldsymbol{\beta}}\mathbf{L}\right)^{-1}\left(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{c}\right)$$

- Asymptotically under the null $W \overset{d}{\to} \chi^2_q$ where q is rows of $\mathbf{L}$
    - q is the number of linear restrictions in the null
- **Wald test**: reject when $W > w_\alpha$, where $\mathbb{P}(W > w_\alpha) = \alpha$ under the null.
    - Use χ^2_q distribution for critical values, p-values
- Typical software output: **F-statistic** $F = W/q$

Wald test

$$W = n\left(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{c}\right)'\left(\mathbf{L}'\widehat{\mathbf{V}}_{\boldsymbol{\beta}}\mathbf{L}\right)^{-1}\left(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{c}\right)$$

- Asymptotically under the null $W \overset{d}{\to} \chi^2_q$ where q is rows of $\mathbf{L}$
  - q is the number of linear restrictions in the null
- **Wald test**: reject when $W > w_\alpha$, where $\mathbb{P}(W > w_\alpha) = \alpha$ under the null.
  - Use χ^2_q distribution for critical values, p-values
- Typical software output: **F-statistic** $F = W/q$
  - p-values and critical values come from F distribution with q and $n - k - 1$ dfs.

Wald test

$$W = n\left(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{c}\right)'\left(\mathbf{L}'\widehat{\mathbf{V}}_{\boldsymbol{\beta}}\mathbf{L}\right)^{-1}\left(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{c}\right)$$

- Asymptotically under the null $W \overset{d}{\to} \chi^2_q$ where q is rows of $\mathbf{L}$
  - q is the number of linear restrictions in the null
- **Wald test**: reject when $W > w_\alpha$, where $\mathbb{P}(W > w_\alpha) = \alpha$ under the null.
  - Use χ^2_q distribution for critical values, p-values
- Typical software output: **F-statistic** $F = W/q$
  - p-values and critical values come from F distribution with q and $n - k - 1$ dfs.
  - As $n \to \infty$, $F_{q,n-k-1} \overset{d}{\to} \chi^2_q$ so asymptotically similar to Wald under homoskedascity (slightly more conservative).

Wald test

$$W = n\left(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{c}\right)'\left(\mathbf{L}'\widehat{\mathbf{V}}_{\boldsymbol{\beta}}\mathbf{L}\right)^{-1}\left(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{c}\right)$$

- Asymptotically under the null $W \overset{d}{\to} \chi^2_q$ where q is rows of $\mathbf{L}$
    - q is the number of linear restrictions in the null
- **Wald test**: reject when $W > w_\alpha$, where $\mathbb{P}(W > w_\alpha) = \alpha$ under the null.
    - Use χ^2_q distribution for critical values, p-values
- Typical software output: **F-statistic** $F = W/q$
    - p-values and critical values come from F distribution with q and $n-k-1$ dfs.
    - As $n \to \infty$, $F_{q,n-k-1} \overset{d}{\to} \chi^2_q$ so asymptotically similar to Wald under homoskedascity (slightly more conservative).
    - No justification for F test under heteroskedasticity.

Wald test

$$W = n\left(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{c}\right)'\left(\mathbf{L}'\widehat{\mathbf{V}}_{\boldsymbol{\beta}}\mathbf{L}\right)^{-1}\left(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{c}\right)$$

- Asymptotically under the null $W \overset{d}{\to} \chi^2_q$ where q is rows of $\mathbf{L}$
  - q is the number of linear restrictions in the null
- **Wald test**: reject when $W > w_\alpha$, where $\mathbb{P}(W > w_\alpha) = \alpha$ under the null.
  - Use χ^2_q distribution for critical values, p-values
- Typical software output: **F-statistic** $F = W/q$
  - p-values and critical values come from F distribution with q and $n - k - 1$ dfs.
  - As $n \to \infty$, $F_{q,n-k-1} \overset{d}{\to} \chi^2_q$ so asymptotically similar to Wald under homoskedascity (slightly more conservative).
  - No justification for F test under heteroskedasticity.
  - "Usual" F-test reports test of all coef = 0 except intercept (pointless?)

Wald test steps

1. Choose a Type I error rate, α.

Wald test steps

1. Choose a Type I error rate, α.
    - Same interpretation: rate of false positives you are willing to accept

Wald test steps

1. Choose a Type I error rate, α.
    - Same interpretation: rate of false positives you are willing to accept
2. Calculate the rejection region for the test (one-sided)

Wald test steps

1. Choose a Type I error rate, α.
    - Same interpretation: rate of false positives you are willing to accept
2. Calculate the rejection region for the test (one-sided)
    - Rejection region is the region $W > w_\alpha$ such that $\mathbb{P}(W > w_\alpha) = \alpha$

Wald test steps

1. Choose a Type I error rate, α.
   - Same interpretation: rate of false positives you are willing to accept
2. Calculate the rejection region for the test (one-sided)
   - Rejection region is the region $W > w_\alpha$ such that $\mathbb{P}(W > w_\alpha) = \alpha$
   - We can get this from R using the `qchisq()` function

Wald test steps

1. Choose a Type I error rate, α.
    - Same interpretation: rate of false positives you are willing to accept
2. Calculate the rejection region for the test (one-sided)
    - Rejection region is the region $W > w_\alpha$ such that $\mathbb{P}(W > w_\alpha) = \alpha$
    - We can get this from R using the `qchisq()` function
3. Reject if observed statistic is bigger than critical value

Wald test steps

1. Choose a Type I error rate, α.
    - Same interpretation: rate of false positives you are willing to accept
2. Calculate the rejection region for the test (one-sided)
    - Rejection region is the region $W > w_\alpha$ such that $\mathbb{P}(W > w_\alpha) = \alpha$
    - We can get this from R using the `qchisq()` function
3. Reject if observed statistic is bigger than critical value
    - Use `pchisq()` to get p-values if needed.

Wald test steps

1. Choose a Type I error rate, α.
    - Same interpretation: rate of false positives you are willing to accept
2. Calculate the rejection region for the test (one-sided)
    - Rejection region is the region $W > w_\alpha$ such that $\mathbb{P}(W > w_\alpha) = \alpha$
    - We can get this from R using the `qchisq()` function
3. Reject if observed statistic is bigger than critical value
    - Use `pchisq()` to get p-values if needed.

- When applied to a single coefficient, equivalent to a t-test.

Wald test steps

1. Choose a Type I error rate, α.
    - Same interpretation: rate of false positives you are willing to accept
2. Calculate the rejection region for the test (one-sided)
    - Rejection region is the region $W > w_\alpha$ such that $\mathbb{P}(W > w_\alpha) = \alpha$
    - We can get this from R using the `qchisq()` function
3. Reject if observed statistic is bigger than critical value
    - Use `pchisq()` to get p-values if needed.

- When applied to a single coefficient, equivalent to a t-test.
- Use packages like `{lmtest}` or `{clubSandwich}` in R.

Wald test in lmtest

```r
## run OLS with the restrictions imposed (avexpr removed)
restricted <- lm(logpgp95 ~ lat_abst + meantemp, data = ajr)

## pass estimated model and estimated null model to
## wald test with HC variance estimator
lmtest::waldtest(restricted, int_mod, test = "Chisq",
                 vcov = vcovHC)
```

```
## Wald test
##
## Model 1: logpgp95 ~ lat_abst + meantemp
## Model 2: logpgp95 ~ avexpr * lat_abst + meantemp
##   Res.Df Df Chisq Pr(>Chisq)
## 1     57
## 2     55  2  34.2    3.7e-08 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Multiple testing

- Separate t-tests for each β_j: α of them will be significant by chance.

Multiple testing

- Separate t-tests for each β_j: α of them will be significant by chance.
- Illustration:

Multiple testing

- Separate t-tests for each β_j: α of them will be significant by chance.
- Illustration:
  - Randomly draw 21 variables independently.

Multiple testing

- Separate t-tests for each β_j: α of them will be significant by chance.
- Illustration:
  - Randomly draw 21 variables independently.
  - Run a regression of the first variable on the rest.

Multiple testing

- Separate t-tests for each β_j: α of them will be significant by chance.
- Illustration:
    - Randomly draw 21 variables independently.
    - Run a regression of the first variable on the rest.
- By design, no effect of any variable on any other.

Multiple test example

```
noise <- data.frame(matrix(rnorm(2100), nrow = 100, ncol = 21))
summary(lm(noise))
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.028039   0.113820   -0.25   0.8061
## X2          -0.150390   0.112181   -1.34   0.1839
## X3           0.079158   0.095028    0.83   0.4074
## X4          -0.071742   0.104579   -0.69   0.4947
## X5           0.172078   0.114002    1.51   0.1352
## X6           0.080852   0.108341    0.75   0.4577
## X7           0.102913   0.114156    0.90   0.3701
## X8          -0.321053   0.120673   -2.66   0.0094 **
## X9          -0.053122   0.107983   -0.49   0.6241
## X10          0.180105   0.126443    1.42   0.1583
## X11          0.166386   0.110947    1.50   0.1377
## X12          0.008011   0.103766    0.08   0.9387
## X13          0.000212   0.103785    0.00   0.9984
## X14         -0.065969   0.112214   -0.59   0.5583
## X15         -0.129654   0.111575   -1.16   0.2487
## X16         -0.054446   0.125140   -0.44   0.6647
## X17          0.004335   0.112012    0.04   0.9692
## X18         -0.080796   0.109853   -0.74   0.4642
## X19         -0.085806   0.118553   -0.72   0.4713
## X20         -0.186006   0.104560   -1.78   0.0791 .
## X21          0.002111   0.108118    0.02   0.9845
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.999 on 79 degrees of freedom
## Multiple R-squared:  0.201,  Adjusted R-squared:  -0.00142
## F-statistic: 0.993 on 20 and 79 DF,  p-value: 0.48
```

Multiple testing gives false positives

- 1 out of 20 variables significant at $\alpha = 0.05$

Multiple testing gives false positives

- 1 out of 20 variables significant at $\alpha = 0.05$
- 2 out of 20 variables significant at $\alpha = 0.1$

Multiple testing gives false positives

- 1 out of 20 variables significant at $\alpha = 0.05$
- 2 out of 20 variables significant at $\alpha = 0.1$
- Exactly the number of false positives we would expect.

Multiple testing gives false positives

- 1 out of 20 variables significant at $\alpha = 0.05$
- 2 out of 20 variables significant at $\alpha = 0.1$
- Exactly the number of false positives we would expect.
- But notice the F-statistic: the variables are not **jointly** significant

Multiple testing gives false positives

- 1 out of 20 variables significant at $\alpha = 0.05$
- 2 out of 20 variables significant at $\alpha = 0.1$
- Exactly the number of false positives we would expect.
- But notice the F-statistic: the variables are not **jointly** significant
- **Bonferroni correction**: use p-value cutoff α/m where m is the number of hypotheses.

Multiple testing gives false positives

- 1 out of 20 variables significant at $\alpha = 0.05$
- 2 out of 20 variables significant at $\alpha = 0.1$
- Exactly the number of false positives we would expect.
- But notice the F-statistic: the variables are not **jointly** significant
- **Bonferroni correction**: use p-value cutoff α/m where m is the number of hypotheses.
  - Example: $0.05/20 = 0.0025$

Multiple testing gives false positives

- 1 out of 20 variables significant at $\alpha = 0.05$
- 2 out of 20 variables significant at $\alpha = 0.1$
- Exactly the number of false positives we would expect.
- But notice the F-statistic: the variables are not **jointly** significant
- **Bonferroni correction**: use p-value cutoff α/m where m is the number of hypotheses.
  - Example: $0.05/20 = 0.0025$
  - Ensures that the family-wise error rate (probability of making at least 1 Type I error) is less than α.

4/ Linear Regression Model and Finite-sample Properties

Standard linear regression model

- Standard textbook model: **correctly specified linear CEF**

Standard linear regression model

- Standard textbook model: **correctly specified linear CEF**
  - Designed for finite-sample results.

Standard linear regression model

- Standard textbook model: **correctly specified linear CEF**
  - Designed for finite-sample results.

Assumption: Linear Regression Model

1. The variables $(Y, \mathbf{X})$ satisfy the the linear CEF assumption.

$$Y = \mathbf{X}'\boldsymbol{\beta} + e$$
$$\mathbb{E}[e \mid \mathbf{X}] = 0.$$

Standard linear regression model

- Standard textbook model: **correctly specified linear CEF**
  - Designed for finite-sample results.

Assumption: Linear Regression Model

1. The variables $(Y, \mathbf{X})$ satisfy the the linear CEF assumption.

$$Y = \mathbf{X}'\boldsymbol{\beta} + e$$
$$\mathbb{E}[e \mid \mathbf{X}] = 0.$$

2. The design matrix is invertible $\mathbb{E}[\mathbf{X}\mathbf{X}'] > 0$ (positive definite).

Standard linear regression model

- Standard textbook model: **correctly specified linear CEF**
  - Designed for finite-sample results.

Assumption: Linear Regression Model

1. The variables $(Y, \mathbf{X})$ satisfy the the linear CEF assumption.

$$Y = \mathbf{X}'\boldsymbol{\beta} + e$$
$$\mathbb{E}[e \mid \mathbf{X}] = 0.$$

2. The design matrix is invertible $\mathbb{E}[\mathbf{X}\mathbf{X}'] > 0$ (positive definite).

- Basically this assumes the CEF of Y given $\mathbf{X}$ is linear.

Standard linear regression model

- Standard textbook model: **correctly specified linear CEF**
  - Designed for finite-sample results.

Assumption: Linear Regression Model

1. The variables $(Y, \mathbf{X})$ satisfy the the linear CEF assumption.

$$Y = \mathbf{X}'\boldsymbol{\beta} + e$$
$$\mathbb{E}[e \mid \mathbf{X}] = 0.$$

2. The design matrix is invertible $\mathbb{E}[\mathbf{X}\mathbf{X}'] > 0$ (positive definite).

- Basically this assumes the CEF of Y given $\mathbf{X}$ is linear.
- We continue to maintain $\{(Y_i, \mathbf{X}_i)\}$ are i.i.d.

Properties of OLS under linear CEF

- Linear CEFs imply stronger finite-sample guarantees:

Properties of OLS under linear CEF

- Linear CEFs imply stronger finite-sample guarantees:

1. **Unbiasedness:** $\mathbb{E}\left[\hat{\boldsymbol{\beta}} \mid \mathbb{X}\right] = \boldsymbol{\beta}$

Properties of OLS under linear CEF

- Linear CEFs imply stronger finite-sample guarantees:

1. **Unbiasedness:** $\mathbb{E}\left[\hat{\boldsymbol{\beta}} \mid \mathbb{X}\right] = \boldsymbol{\beta}$
2. **Conditional sampling variance**: let $\sigma_i^2 = \mathbb{E}[e_i^2 \mid \mathbf{X}_i]$

$$\mathbb{V}[\hat{\boldsymbol{\beta}} \mid \mathbb{X}] = (\mathbb{X}'\mathbb{X})^{-1}\left(\sum_{i=1}^{n} \sigma_i^2 \mathbf{X}_i \mathbf{X}_i'\right)(\mathbb{X}'\mathbb{X})^{-1}$$

Properties of OLS under linear CEF

- Linear CEFs imply stronger finite-sample guarantees:

1. **Unbiasedness:** $\mathbb{E}\left[\hat{\boldsymbol{\beta}} \mid \mathbb{X}\right] = \boldsymbol{\beta}$
2. **Conditional sampling variance**: let $\sigma_i^2 = \mathbb{E}[e_i^2 \mid \mathbf{X}_i]$

$$\mathbb{V}[\hat{\boldsymbol{\beta}} \mid \mathbb{X}] = (\mathbb{X}'\mathbb{X})^{-1}\left(\sum_{i=1}^{n} \sigma_i^2 \mathbf{X}_i \mathbf{X}_i'\right)(\mathbb{X}'\mathbb{X})^{-1}$$

- Useful when linearity holds by default (discrete X in experiments, etc)

Linear CEF under homoskedasticity

- Under homoskedasticity, we have a few other finite-sample results:

Linear CEF under homoskedasticity

- Under homoskedasticity, we have a few other finite-sample results:

3. **Conditional sampling variance**: $\mathbb{V}[\widehat{\boldsymbol{\beta}} \mid \mathbb{X}] = \sigma^2 \left(\mathbb{X}'\mathbb{X}\right)^{-1}$

Linear CEF under homoskedasticity

- Under homoskedasticity, we have a few other finite-sample results:

3. **Conditional sampling variance**: $\mathbb{V}[\hat{\boldsymbol{\beta}} \mid \mathbb{X}] = \sigma^2 \left(\mathbb{X}'\mathbb{X}\right)^{-1}$
4. **Unbiased variance estimator**: $\mathbb{E}\left[\widehat{\mathbb{V}}^0[\hat{\boldsymbol{\beta}}] \mid \mathbf{X}\right] = \sigma^2(\mathbb{X}'\mathbb{X})^{-1}$

Linear CEF under homoskedasticity

- Under homoskedasticity, we have a few other finite-sample results:

3. **Conditional sampling variance**: $\mathbb{V}[\hat{\boldsymbol{\beta}} \mid \mathbb{X}] = \sigma^2 \left(\mathbb{X}'\mathbb{X}\right)^{-1}$
4. **Unbiased variance estimator**: $\mathbb{E}\left[\widehat{\mathbb{V}}^0[\hat{\boldsymbol{\beta}}] \mid \mathbf{X}\right] = \sigma^2(\mathbb{X}'\mathbb{X})^{-1}$
5. **Gauss-Markov**: OLS is the best linear unbiased estimator of $\boldsymbol{\beta}$ (BLUE). If $\tilde{\boldsymbol{\beta}}$ is a linear estimator,

$$\mathbb{V}[\tilde{\boldsymbol{\beta}} \mid \mathbb{X}] \geq \mathbb{V}[\hat{\boldsymbol{\beta}} \mid \mathbb{X}] = \sigma^2 \left(\mathbb{X}'\mathbb{X}\right)^{-1}$$

Linear CEF under homoskedasticity

- Under homoskedasticity, we have a few other finite-sample results:

3. **Conditional sampling variance**: $\mathbb{V}[\hat{\boldsymbol{\beta}} \mid \mathbb{X}] = \sigma^2 \left(\mathbb{X}'\mathbb{X}\right)^{-1}$
4. **Unbiased variance estimator**: $\mathbb{E}\left[\widehat{\mathbb{V}}^0[\hat{\boldsymbol{\beta}}] \mid \mathbf{X}\right] = \sigma^2(\mathbb{X}'\mathbb{X})^{-1}$
5. **Gauss-Markov**: OLS is the best linear unbiased estimator of $\boldsymbol{\beta}$ (BLUE). If $\tilde{\boldsymbol{\beta}}$ is a linear estimator,

$$\mathbb{V}[\tilde{\boldsymbol{\beta}} \mid \mathbb{X}] \geq \mathbb{V}[\hat{\boldsymbol{\beta}} \mid \mathbb{X}] = \sigma^2 \left(\mathbb{X}'\mathbb{X}\right)^{-1}$$

  - For matrices, $\mathbf{A} \geq \mathbf{B}$ means that $\mathbf{A} - \mathbf{B}$ is positive semidefinite.

Linear CEF under homoskedasticity

- Under homoskedasticity, we have a few other finite-sample results:

3. **Conditional sampling variance**: $\mathbb{V}[\hat{\boldsymbol{\beta}} \mid \mathbb{X}] = \sigma^2 \left(\mathbb{X}'\mathbb{X}\right)^{-1}$
4. **Unbiased variance estimator**: $\mathbb{E}\left[\widehat{\mathbb{V}}^0[\hat{\boldsymbol{\beta}}] \mid \mathbf{X}\right] = \sigma^2(\mathbb{X}'\mathbb{X})^{-1}$
5. **Gauss-Markov**: OLS is the best linear unbiased estimator of $\boldsymbol{\beta}$ (BLUE). If $\tilde{\boldsymbol{\beta}}$ is a linear estimator,

$$\mathbb{V}[\tilde{\boldsymbol{\beta}} \mid \mathbb{X}] \geq \mathbb{V}[\hat{\boldsymbol{\beta}} \mid \mathbb{X}] = \sigma^2 \left(\mathbb{X}'\mathbb{X}\right)^{-1}$$

  - For matrices, $\mathbf{A} \geq \mathbf{B}$ means that $\mathbf{A} - \mathbf{B}$ is positive semidefinite.
  - A matrix $\mathbf{C}$ is p.s.d. if $\mathbf{x}'\mathbf{C}\mathbf{x} \geq 0$.

Linear CEF under homoskedasticity

- Under homoskedasticity, we have a few other finite-sample results:

3. **Conditional sampling variance**: $\mathbb{V}[\hat{\boldsymbol{\beta}} \mid \mathbb{X}] = \sigma^2 \left(\mathbb{X}'\mathbb{X}\right)^{-1}$
4. **Unbiased variance estimator**: $\mathbb{E}\left[\widehat{\mathbb{V}}^0[\hat{\boldsymbol{\beta}}] \mid \mathbf{X}\right] = \sigma^2(\mathbb{X}'\mathbb{X})^{-1}$
5. **Gauss-Markov**: OLS is the best linear unbiased estimator of $\boldsymbol{\beta}$ (BLUE). If $\tilde{\boldsymbol{\beta}}$ is a linear estimator,

$$\mathbb{V}[\tilde{\boldsymbol{\beta}} \mid \mathbb{X}] \geq \mathbb{V}[\hat{\boldsymbol{\beta}} \mid \mathbb{X}] = \sigma^2 \left(\mathbb{X}'\mathbb{X}\right)^{-1}$$

  - For matrices, $\mathbf{A} \geq \mathbf{B}$ means that $\mathbf{A} - \mathbf{B}$ is positive semidefinite.
  - A matrix $\mathbf{C}$ is p.s.d. if $\mathbf{x}'\mathbf{C}\mathbf{x} \geq 0$.
  - Upshot: OLS will have the smaller SEs than any other linear estimator.

Normal regression model

- Most parametric: $Y \sim \mathcal{N}(\mathbf{X}'\boldsymbol{\beta}, \sigma^2)$.

Normal regression model

- Most parametric: $Y \sim \mathcal{N}(\mathbf{X}'\boldsymbol{\beta}, \sigma^2)$.
    - Normal error model since $e = Y - \mathbf{X}'\boldsymbol{\beta} \sim \mathcal{N}(0, \sigma^2)$.

Normal regression model

- Most parametric: $Y \sim \mathcal{N}(\mathbf{X}'\boldsymbol{\beta}, \sigma^2)$.
  - Normal error model since $e = Y - \mathbf{X}'\boldsymbol{\beta} \sim \mathcal{N}(0, \sigma^2)$.
- Rarely believed, but allows for exact inference for all n.

Normal regression model

- Most parametric: $Y \sim \mathcal{N}(\mathbf{X}'\boldsymbol{\beta}, \sigma^2)$.
    - Normal error model since $e = Y - \mathbf{X}'\boldsymbol{\beta} \sim \mathcal{N}(0, \sigma^2)$.
- Rarely believed, but allows for exact inference for all n.
    - $(\hat{\beta}_j - \beta_j)/\widehat{\text{se}}(\hat{\beta}_j)$ follows a t distribution with $n - k$ degrees of freedom.

Normal regression model

- Most parametric: $Y \sim \mathcal{N}(\mathbf{X}'\boldsymbol{\beta}, \sigma^2)$.
    - Normal error model since $e = Y - \mathbf{X}'\boldsymbol{\beta} \sim \mathcal{N}(0, \sigma^2)$.
- Rarely believed, but allows for exact inference for all n.
    - $(\hat{\beta}_j - \beta_j)/\widehat{\text{se}}(\hat{\beta}_j)$ follows a t distribution with $n - k$ degrees of freedom.
    - F statistics follows F distribution exactly rather than approximately.

Normal regression model

- Most parametric: $Y \sim \mathcal{N}(\mathbf{X}'\boldsymbol{\beta}, \sigma^2)$.
    - Normal error model since $e = Y - \mathbf{X}'\boldsymbol{\beta} \sim \mathcal{N}(0, \sigma^2)$.
- Rarely believed, but allows for exact inference for all n.
    - $(\hat{\beta}_j - \beta_j)/\widehat{\text{se}}(\hat{\beta}_j)$ follows a t distribution with $n - k$ degrees of freedom.
    - F statistics follows F distribution exactly rather than approximately.
- Software often implicitly assumes this for p-values.

Normal regression model

- Most parametric: $Y \sim \mathcal{N}(\mathbf{X}'\boldsymbol{\beta}, \sigma^2)$.
    - Normal error model since $e = Y - \mathbf{X}'\boldsymbol{\beta} \sim \mathcal{N}(0, \sigma^2)$.
- Rarely believed, but allows for exact inference for all n.
    - $(\hat{\beta}_j - \beta_j)/\widehat{\text{se}}(\hat{\beta}_j)$ follows a t distribution with $n - k$ degrees of freedom.
    - F statistics follows F distribution exactly rather than approximately.
- Software often implicitly assumes this for p-values.
- With reasonable n, asymptotic normality has the same effect.

5/ Clustering

Clustered dependence: intuition

- Think back to the Gerber, Green, and Larimer (2008) social pressure mailer example.

Clustered dependence: intuition

- Think back to the Gerber, Green, and Larimer (2008) social pressure mailer example.
    - Randomly assign households to different treatment conditions.

Clustered dependence: intuition

- Think back to the Gerber, Green, and Larimer (2008) social pressure mailer example.
    - Randomly assign households to different treatment conditions.
    - But the measurement of turnout is at the individual level.

Clustered dependence: intuition

- Think back to the Gerber, Green, and Larimer (2008) social pressure mailer example.
    - Randomly assign households to different treatment conditions.
    - But the measurement of turnout is at the individual level.
- Zero conditional mean error holds here (random assignment)

Clustered dependence: intuition

- Think back to the Gerber, Green, and Larimer (2008) social pressure mailer example.
    - Randomly assign households to different treatment conditions.
    - But the measurement of turnout is at the individual level.
- Zero conditional mean error holds here (random assignment)
- Violation of **iid/random sampling**:

Clustered dependence: intuition

- Think back to the Gerber, Green, and Larimer (2008) social pressure mailer example.
    - Randomly assign households to different treatment conditions.
    - But the measurement of turnout is at the individual level.
- Zero conditional mean error holds here (random assignment)
- Violation of **iid/random sampling**:
    - errors of individuals within the same household are correlated.

Clustered dependence: intuition

- Think back to the Gerber, Green, and Larimer (2008) social pressure mailer example.
    - Randomly assign households to different treatment conditions.
    - But the measurement of turnout is at the individual level.
- Zero conditional mean error holds here (random assignment)
- Violation of **iid/random sampling**:
    - errors of individuals within the same household are correlated.
    - SEs are going to be wrong.

Clustered dependence: intuition

- Think back to the Gerber, Green, and Larimer (2008) social pressure mailer example.
    - Randomly assign households to different treatment conditions.
    - But the measurement of turnout is at the individual level.
- Zero conditional mean error holds here (random assignment)
- Violation of **iid/random sampling**:
    - errors of individuals within the same household are correlated.
    - SEs are going to be wrong.
- Called **clustering** or **clustered dependence**

Clustered dependence: notation

- Clusters (groups): $g = 1, \dots, m$

Clustered dependence: notation

- Clusters (groups): $g = 1, \ldots, m$
- Units: $i = 1, \ldots, n_g$

Clustered dependence: notation

- Clusters (groups): $g = 1, \dots, m$
- Units: $i = 1, \dots, n_g$
- n_g is the number of units in cluster g

Clustered dependence: notation

- Clusters (groups): $g = 1, \dots, m$
- Units: $i = 1, \dots, n_g$
- n_g is the number of units in cluster g
- $n = \sum_{g=1}^{m} n_g$ is the total number of units

Clustered dependence: notation

- Clusters (groups): $g = 1, \ldots, m$
- Units: $i = 1, \ldots, n_g$
- n_g is the number of units in cluster g
- $n = \sum_{g=1}^{m} n_g$ is the total number of units
- Units are (usually) belong to a single cluster:

Clustered dependence: notation

- Clusters (groups): $g = 1, \dots, m$
- Units: $i = 1, \dots, n_g$
- n_g is the number of units in cluster g
- $n = \sum_{g=1}^{m} n_g$ is the total number of units
- Units are (usually) belong to a single cluster:
    - voters in households

Clustered dependence: notation

- Clusters (groups): $g = 1, \dots, m$
- Units: $i = 1, \dots, n_g$
- n_g is the number of units in cluster g
- $n = \sum_{g=1}^{m} n_g$ is the total number of units
- Units are (usually) belong to a single cluster:
    - voters in households
    - individuals in states

Clustered dependence: notation

- Clusters (groups): $g = 1, \ldots, m$
- Units: $i = 1, \ldots, n_g$
- n_g is the number of units in cluster g
- $n = \sum_{g=1}^{m} n_g$ is the total number of units
- Units are (usually) belong to a single cluster:
    - voters in households
    - individuals in states
    - students in classes

Clustered dependence: notation

- Clusters (groups): $g = 1, \ldots, m$
- Units: $i = 1, \ldots, n_g$
- n_g is the number of units in cluster g
- $n = \sum_{g=1}^{m} n_g$ is the total number of units
- Units are (usually) belong to a single cluster:
    - voters in households
    - individuals in states
    - students in classes
    - rulings in judges

Clustered dependence: notation

- Clusters (groups): $g = 1, \dots, m$
- Units: $i = 1, \dots, n_g$
- n_g is the number of units in cluster g
- $n = \sum_{g=1}^{m} n_g$ is the total number of units
- Units are (usually) belong to a single cluster:
    - voters in households
    - individuals in states
    - students in classes
    - rulings in judges
- Outcome varies at the unit-level, Y_{ig} and the main independent variable varies at the cluster level, X_g.

Clustered dependence: example model

$$
\begin{aligned}
Y_{ig} &= \beta_0 + X_g\beta_1 + v_{ig} \\
&= \beta_0 + X_g\beta_1 + c_g + u_{ig}
\end{aligned}
$$

- u_{ig} unit error component with $\mathbb{V}[u_{ig}|X_g] = \sigma_u^2$

Clustered dependence: example model

$$
\begin{aligned}
Y_{ig} &= \beta_0 + X_g\beta_1 + v_{ig} \\
&= \beta_0 + X_g\beta_1 + c_g + u_{ig}
\end{aligned}
$$

- u_{ig} unit error component with $\mathbb{V}[u_{ig}|X_g] = \sigma_u^2$
- c_g cluster error component with $\mathbb{V}[c_g|X_g] = \sigma_c^2$

Clustered dependence: example model

$$
\begin{aligned}
Y_{ig} &= \beta_0 + X_g\beta_1 + v_{ig} \\
&= \beta_0 + X_g\beta_1 + c_g + u_{ig}
\end{aligned}
$$

- u_{ig} unit error component with $\mathbb{V}[u_{ig}|X_g] = \sigma_u^2$
- c_g cluster error component with $\mathbb{V}[c_g|X_g] = \sigma_c^2$
- c_g and u_{ig} are assumed to be independent of each other.

Clustered dependence: example model

$$\begin{aligned} Y_{ig} &= \beta_0 + X_g\beta_1 + v_{ig} \\ &= \beta_0 + X_g\beta_1 + c_g + u_{ig} \end{aligned}$$

- u_{ig} unit error component with $\mathbb{V}[u_{ig}|X_g] = \sigma_u^2$
- c_g cluster error component with $\mathbb{V}[c_g|X_g] = \sigma_c^2$
- c_g and u_{ig} are assumed to be independent of each other.
    - $\rightsquigarrow \mathbb{V}[v_{ig}|X_g] = \sigma_c^2 + \sigma_u^2$

Clustered dependence: example model

$$\begin{aligned} Y_{ig} &= \beta_0 + X_g\beta_1 + v_{ig} \\ &= \beta_0 + X_g\beta_1 + c_g + u_{ig} \end{aligned}$$

- u_{ig} unit error component with $\mathbb{V}[u_{ig}|X_g] = \sigma_u^2$
- c_g cluster error component with $\mathbb{V}[c_g|X_g] = \sigma_c^2$
- c_g and u_{ig} are assumed to be independent of each other.
    - $\rightsquigarrow \mathbb{V}[v_{ig}|X_g] = \sigma_c^2 + \sigma_u^2$
- What if we ignore this structure and just use v_{ig} as the error?

Lack of independence

- Covariance between two units i and s in the same cluster:

$$\text{Cov}[v_{ig}, v_{sg}] = \sigma_c^2$$

Lack of independence

- Covariance between two units i and s in the same cluster:

$$\text{Cov}[v_{ig}, v_{sg}] = \sigma_c^2$$

- Correlation between units in the same group is called the **intra-class correlation coefficient**, or ρ_c:

$$\text{Cor}[v_{ig}, v_{sg}] = \frac{\sigma_c^2}{\sigma_c^2 + \sigma_u^2} = \rho_c$$

Lack of independence

- Covariance between two units i and s in the same cluster:

$$\text{Cov}[v_{ig}, v_{sg}] = \sigma_c^2$$

- Correlation between units in the same group is called the **intra-class correlation coefficient**, or ρ_c:

$$\text{Cor}[v_{ig}, v_{sg}] = \frac{\sigma_c^2}{\sigma_c^2 + \sigma_u^2} = \rho_c$$

- Zero covariance of two units i and s in different clusters g and k:

$$\text{Cov}[v_{ig}, v_{sk}] = 0$$

Example covariance matrix

- $\mathbf{v}' = \begin{bmatrix} v_{1,1} & v_{2,1} & v_{3,1} & v_{4,2} & v_{5,2} & v_{6,2} \end{bmatrix}$

Example covariance matrix

- $\mathbf{v}' = \begin{bmatrix} v_{1,1} & v_{2,1} & v_{3,1} & v_{4,2} & v_{5,2} & v_{6,2} \end{bmatrix}$
- Variance matrix under clustering:

Example covariance matrix

- $\mathbf{v}' = \begin{bmatrix} v_{1,1} & v_{2,1} & v_{3,1} & v_{4,2} & v_{5,2} & v_{6,2} \end{bmatrix}$
- Variance matrix under clustering:

$$\mathbb{V}[\mathbf{v}|\mathbf{X}] = \begin{bmatrix} \sigma_c^2 + \sigma_u^2 & \sigma_c^2 & \sigma_c^2 & 0 & 0 & 0 \\ \sigma_c^2 & \sigma_c^2 + \sigma_u^2 & \sigma_c^2 & 0 & 0 & 0 \\ \sigma_c^2 & \sigma_c^2 & \sigma_c^2 + \sigma_u^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_c^2 + \sigma_u^2 & \sigma_c^2 & \sigma_c^2 \\ 0 & 0 & 0 & \sigma_c^2 & \sigma_c^2 + \sigma_u^2 & \sigma_c^2 \\ 0 & 0 & 0 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 + \sigma_u^2 \end{bmatrix}$$

Example covariance matrix

- $\mathbf{v}' = \begin{bmatrix} v_{1,1} & v_{2,1} & v_{3,1} & v_{4,2} & v_{5,2} & v_{6,2} \end{bmatrix}$
- Variance matrix under clustering:

$$\mathbb{V}[\mathbf{v}|\mathbf{X}] = \begin{bmatrix} \sigma_c^2 + \sigma_u^2 & \sigma_c^2 & \sigma_c^2 & 0 & 0 & 0 \\ \sigma_c^2 & \sigma_c^2 + \sigma_u^2 & \sigma_c^2 & 0 & 0 & 0 \\ \sigma_c^2 & \sigma_c^2 & \sigma_c^2 + \sigma_u^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_c^2 + \sigma_u^2 & \sigma_c^2 & \sigma_c^2 \\ 0 & 0 & 0 & \sigma_c^2 & \sigma_c^2 + \sigma_u^2 & \sigma_c^2 \\ 0 & 0 & 0 & \sigma_c^2 & \sigma_c^2 & \sigma_c^2 + \sigma_u^2 \end{bmatrix}$$

- Variance matrix under i.i.d.:

$$\mathbb{V}[\mathbf{v}|\mathbf{X}] = \begin{bmatrix} \sigma_u^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma_u^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_u^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_u^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_u^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma_u^2 \end{bmatrix}$$

Effects of clustering

$$Y_{ig} = \beta_0 + X_g\beta_1 + c_g + u_{ig}$$

- $\mathbb{V}^0[\hat{\beta}_1]$ = **conventional** OLS variance assuming i.i.d./homoskedasticity.

Effects of clustering

$$Y_{ig} = \beta_0 + X_g\beta_1 + c_g + u_{ig}$$

- $\mathbb{V}^0[\hat{\beta}_1]$ = **conventional** OLS variance assuming i.i.d./homoskedasticity.
- Let $\mathbb{V}[\hat{\beta}_1]$ be the true sampling variance under clustering.

Effects of clustering

$$Y_{ig} = \beta_0 + X_g\beta_1 + c_g + u_{ig}$$

- $\mathbb{V}^0[\hat{\beta}_1]$ = **conventional** OLS variance assuming i.i.d./homoskedasticity.
- Let $\mathbb{V}[\hat{\beta}_1]$ be the true sampling variance under clustering.
- When clusters are balanced, $n^* = n_g$, comparison of clustered to conventional:

$$\mathbb{V}[\hat{\beta}_1] \approx \mathbb{V}^0[\hat{\beta}_1]\left(1 + (n^* - 1)\rho_c\right)$$

Effects of clustering

$$Y_{ig} = \beta_0 + X_g\beta_1 + c_g + u_{ig}$$

- $\mathbb{V}^0[\hat{\beta}_1] =$ **conventional** OLS variance assuming i.i.d./homoskedasticity.
- Let $\mathbb{V}[\hat{\beta}_1]$ be the true sampling variance under clustering.
- When clusters are balanced, $n^* = n_g$, comparison of clustered to conventional:
$$\mathbb{V}[\hat{\beta}_1] \approx \mathbb{V}^0[\hat{\beta}_1]\left(1 + (n^* - 1)\rho_c\right)$$
- True variance will be higher than conventional when within-cluster correlation is positive, $\rho_c > 0$.

Linear model with clustering

$$Y_{ig} = \mathbf{X}_{ig}'\boldsymbol{\beta} + v_{ig}$$

- Assumptions:

Linear model with clustering

$$Y_{ig} = \mathbf{X}'_{ig}\boldsymbol{\beta} + v_{ig}$$

- Assumptions:
  - $\mathbb{E}[v_{ig} \mid \mathbf{X}_{ig}] = 0$ so we have the correct CEF.

Linear model with clustering

$$Y_{ig} = \mathbf{X}_{ig}'\boldsymbol{\beta} + v_{ig}$$

- Assumptions:
  - $\mathbb{E}[v_{ig} \mid \mathbf{X}_{ig}] = 0$ so we have the correct CEF.
  - $\mathbb{E}[v_{ig} v_{jg'} \mid \mathbf{X}_{ig}, \mathbf{X}_{jg'}] = 0$ unless $g = g'$.

Linear model with clustering

$$Y_{ig} = \mathbf{X}_{ig}'\boldsymbol{\beta} + v_{ig}$$

- Assumptions:
  - $\mathbb{E}[v_{ig} \mid \mathbf{X}_{ig}] = 0$ so we have the correct CEF.
  - $\mathbb{E}[v_{ig}v_{jg'} \mid \mathbf{X}_{ig}, \mathbf{X}_{jg'}] = 0$ unless $g = g'$.
  - Correlated errors allowed within groups, uncorrelated across. Allows heteroskedasticity.

Linear model with clustering

$$Y_{ig} = \mathbf{X}_{ig}'\boldsymbol{\beta} + v_{ig}$$

- Assumptions:
  - $\mathbb{E}[v_{ig} \mid \mathbf{X}_{ig}] = 0$ so we have the correct CEF.
  - $\mathbb{E}[v_{ig}v_{jg'} \mid \mathbf{X}_{ig}, \mathbf{X}_{jg'}] = 0$ unless $g = g'$.
  - Correlated errors allowed within groups, uncorrelated across. Allows heteroskedasticity.
- Pooled OLS under clustered dependence:

$$\mathbf{Y}_g = \mathbb{X}_g\boldsymbol{\beta} + \mathbf{v}_g$$

Linear model with clustering

$$Y_{ig} = \mathbf{X}_{ig}'\boldsymbol{\beta} + v_{ig}$$

- Assumptions:
  - $\mathbb{E}[v_{ig} \mid \mathbf{X}_{ig}] = 0$ so we have the correct CEF.
  - $\mathbb{E}[v_{ig} v_{jg'} \mid \mathbf{X}_{ig}, \mathbf{X}_{jg'}] = 0$ unless $g = g'$.
  - Correlated errors allowed within groups, uncorrelated across. Allows heteroskedasticity.
- Pooled OLS under clustered dependence:

$$\mathbf{Y}_g = \mathbb{X}_g\boldsymbol{\beta} + \mathbf{v}_g$$

  - $\mathbf{Y}_g$ is the $n_g \times 1$ vector of responses for cluster g

Linear model with clustering

$$Y_{ig} = \mathbf{X}'_{ig}\boldsymbol{\beta} + v_{ig}$$

- Assumptions:
  - $\mathbb{E}[v_{ig} \mid \mathbf{X}_{ig}] = 0$ so we have the correct CEF.
  - $\mathbb{E}[v_{ig}v_{jg'} \mid \mathbf{X}_{ig}, \mathbf{X}_{jg'}] = 0$ unless $g = g'$.
  - Correlated errors allowed within groups, uncorrelated across. Allows heteroskedasticity.
- Pooled OLS under clustered dependence:

$$\mathbf{Y}_g = \mathbb{X}_g\boldsymbol{\beta} + \mathbf{v}_g$$

  - $\mathbf{Y}_g$ is the $n_g \times 1$ vector of responses for cluster g
  - $\mathbb{X}_g$ is the $n_g \times k$ matrix of data for the gth cluster.

Linear model with clustering

$$Y_{ig} = \mathbf{X}_{ig}'\boldsymbol{\beta} + v_{ig}$$

- Assumptions:
  - $\mathbb{E}[v_{ig} \mid \mathbf{X}_{ig}] = 0$ so we have the correct CEF.
  - $\mathbb{E}[v_{ig}v_{jg'} \mid \mathbf{X}_{ig}, \mathbf{X}_{jg'}] = 0$ unless $g = g'$.
  - Correlated errors allowed within groups, uncorrelated across. Allows heteroskedasticity.
- Pooled OLS under clustered dependence:

$$\mathbf{Y}_g = \mathbb{X}_g\boldsymbol{\beta} + \mathbf{v}_g$$

  - $\mathbf{Y}_g$ is the $n_g \times 1$ vector of responses for cluster g
  - $\mathbb{X}_g$ is the $n_g \times k$ matrix of data for the gth cluster.
- We can write the OLS estimator as:

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{g=1}^{m} \mathbb{X}_g'\mathbb{X}_g\right)\left(\sum_{g=1}^{m} \mathbb{X}_g'\mathbf{Y}_g\right)$$

Cluster-robust variance estimator

- Independence is across clusters so the CLT holds as m gets big.

Cluster-robust variance estimator

- Independence is across clusters so the CLT holds as m gets big.
    - Key intuition: we're sampling clusters, not individual units.

Cluster-robust variance estimator

- Independence is across clusters so the CLT holds as m gets big.
    - Key intuition: we're sampling clusters, not individual units.
- CLT implies $\sqrt{m}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ will be asymp. normal with mean 0 and variance:

$$\left(\mathbb{E}[\mathbb{X}_g'\mathbb{X}_g]\right)^{-1} \mathbb{E}[\mathbb{X}_g'\mathbf{v}_g\mathbf{v}_g'\mathbb{X}_g] \left(\mathbb{E}[\mathbb{X}_g'\mathbb{X}_g]\right)^{-1}$$

Cluster-robust variance estimator

- Independence is across clusters so the CLT holds as m gets big.
  - Key intuition: we're sampling clusters, not individual units.
- CLT implies $\sqrt{m}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ will be asymp. normal with mean 0 and variance:

$$\left(\mathbb{E}[\mathbb{X}_g'\mathbb{X}_g]\right)^{-1} \mathbb{E}[\mathbb{X}_g'\mathbf{v}_g\mathbf{v}_g'\mathbb{X}_g] \left(\mathbb{E}[\mathbb{X}_g'\mathbb{X}_g]\right)^{-1}$$

- Similar to the iid case, replace population quantities with sample versions (and divide by m):

$$\widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}}^{\text{CL0}} = (\mathbb{X}'\mathbb{X})^{-1} \left(\sum_{g=1}^{m} \mathbb{X}_g'\hat{\mathbf{v}}_g\hat{\mathbf{v}}_g'\mathbb{X}_g\right) (\mathbb{X}'\mathbb{X})^{-1}$$

Cluster-robust variance estimator

- Independence is across clusters so the CLT holds as m gets big.
    - Key intuition: we're sampling clusters, not individual units.
- CLT implies $\sqrt{m}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ will be asymp. normal with mean 0 and variance:

$$\left(\mathbb{E}[\mathbb{X}_g'\mathbb{X}_g]\right)^{-1}\mathbb{E}[\mathbb{X}_g'\mathbf{v}_g\mathbf{v}_g'\mathbb{X}_g]\left(\mathbb{E}[\mathbb{X}_g'\mathbb{X}_g]\right)^{-1}$$

- Similar to the iid case, replace population quantities with sample versions (and divide by m):

$$\widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}}^{\text{CL0}} = (\mathbb{X}'\mathbb{X})^{-1}\left(\sum_{g=1}^{m}\mathbb{X}_g'\hat{\mathbf{v}}_g\hat{\mathbf{v}}_g'\mathbb{X}_g\right)(\mathbb{X}'\mathbb{X})^{-1}$$

    - Noting: $\mathbb{X}'\mathbb{X}/m = m^{-1}\sum_{g=1}^{m}\mathbb{X}_g'\mathbb{X}_g$

Cluster-robust variance estimator

- Independence is across clusters so the CLT holds as m gets big.
    - Key intuition: we're sampling clusters, not individual units.
- CLT implies $\sqrt{m}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ will be asymp. normal with mean 0 and variance:

$$\left(\mathbb{E}[\mathbb{X}_g'\mathbb{X}_g]\right)^{-1}\mathbb{E}[\mathbb{X}_g'\mathbf{v}_g\mathbf{v}_g'\mathbb{X}_g]\left(\mathbb{E}[\mathbb{X}_g'\mathbb{X}_g]\right)^{-1}$$

- Similar to the iid case, replace population quantities with sample versions (and divide by m):

$$\widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}}^{\text{CL0}} = (\mathbb{X}'\mathbb{X})^{-1}\left(\sum_{g=1}^{m}\mathbb{X}_g'\hat{\mathbf{v}}_g\hat{\mathbf{v}}_g'\mathbb{X}_g\right)(\mathbb{X}'\mathbb{X})^{-1}$$

    - Noting: $\mathbb{X}'\mathbb{X}/m = m^{-1}\sum_{g=1}^{m}\mathbb{X}_g'\mathbb{X}_g$
- With small-sample adjustment (reported by most software):

$$\widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}}^{\text{CL1}} = \frac{m}{m-1}\frac{n-1}{n-k}(\mathbb{X}'\mathbb{X})^{-1}\left(\sum_{g=1}^{m}\mathbb{X}_g'\hat{\mathbf{v}}_g\hat{\mathbf{v}}_g'\mathbb{X}_g\right)(\mathbb{X}'\mathbb{X})^{-1}$$

Example: Gerber, Green, Larimer

Dear Registered Voter:

WHAT IF YOUR NEIGHBORS KNEW WHETHER YOU VOTED?

Why do so many people fail to vote? We've been talking about the problem for years, but it only seems to get worse. This year, we're taking a new approach. We're sending this mailing to you and your neighbors to publicize who does and does not vote.

The chart shows the names of some of your neighbors, showing which have voted in the past. After the August 8 election, we intend to mail an updated chart. You and your neighbors will all know who voted and who did not.

DO YOUR CIVIC DUTY — VOTE!

MAPLE DR	Aug 04	Nov 04	Aug 06
9995 JOSEPH JAMES SMITH	Voted	Voted	______
9995 JENNIFER KAY SMITH		Voted	______
9997 RICHARD B JACKSON		Voted	______
9999 KATHY MARIE JACKSON		Voted	______

Social pressure model

```
load("../assets/gerber_green_larimer.RData")
library(lmtest)
social$voted <- 1 * (social$voted == "Yes")
social$treatment <- factor(
  social$treatment,
  levels = c("Control", "Hawthorne", "Civic Duty", "Neighbors", "Self")
)
mod1 <- lm(voted ~ treatment, data = social)
coeftest(mod1)
```

```
##
## t test of coefficients:
##
##                       Estimate Std. Error t value
## (Intercept)            0.29664    0.00106  279.53
## treatmentHawthorne     0.02574    0.00260    9.90
## treatmentCivic Duty    0.01790    0.00260    6.88
## treatmentNeighbors     0.08131    0.00260   31.26
## treatmentSelf          0.04851    0.00260   18.66
##                       Pr(>|t|)
## (Intercept)            < 2e-16 ***
## treatmentHawthorne     < 2e-16 ***
## treatmentCivic Duty    5.8e-12 ***
## treatmentNeighbors     < 2e-16 ***
## treatmentSelf          < 2e-16 ***
## ---
```

Social pressure model, CRSEs

```r
library(sandwich)
coeftest(mod1, vcov = sandwich::vcovCL(mod1, cluster = social$hh_id))
```

```
##
## t test of coefficients:
##
##                      Estimate Std. Error t value
## (Intercept)           0.29664    0.00131  226.52
## treatmentHawthorne    0.02574    0.00326    7.90
## treatmentCivic Duty   0.01790    0.00324    5.53
## treatmentNeighbors    0.08131    0.00337   24.13
## treatmentSelf         0.04851    0.00330   14.70
##                      Pr(>|t|)
## (Intercept)           < 2e-16 ***
## treatmentHawthorne    2.8e-15 ***
## treatmentCivic Duty   3.2e-08 ***
## treatmentNeighbors    < 2e-16 ***
## treatmentSelf         < 2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Cluster-robust standard errors

- CRSE do not change our estimates $\hat{\beta}$, cannot fix bias

Cluster-robust standard errors

- CRSE do not change our estimates $\hat{\beta}$, cannot fix bias
- Valid under **clustered dependence** when main variable is constant within cluster

Cluster-robust standard errors

- CRSE do not change our estimates $\hat{\boldsymbol{\beta}}$, cannot fix bias
- Valid under **clustered dependence** when main variable is constant within cluster
    - Relies on independence between clusters

Cluster-robust standard errors

- CRSE do not change our estimates $\hat{\beta}$, cannot fix bias
- Valid under **clustered dependence** when main variable is constant within cluster
    - Relies on independence between clusters
    - Allows for arbitrary dependence within clusters

Cluster-robust standard errors

- CRSE do not change our estimates $\hat{\boldsymbol{\beta}}$, cannot fix bias
- Valid under **clustered dependence** when main variable is constant within cluster
  - Relies on independence between clusters
  - Allows for arbitrary dependence within clusters
  - CRSEs usually $>$ conventional SEs—use when you suspect clustering

Cluster-robust standard errors

- CRSE do not change our estimates $\hat{\beta}$, cannot fix bias
- Valid under **clustered dependence** when main variable is constant within cluster
  - Relies on independence between clusters
  - Allows for arbitrary dependence within clusters
  - CRSEs usually $>$ conventional SEs—use when you suspect clustering
- When X_{ig} not constant within cluster, but just correlated $\rightsquigarrow$ more complicated.

Cluster-robust standard errors

- CRSE do not change our estimates $\hat{\beta}$, cannot fix bias
- Valid under **clustered dependence** when main variable is constant within cluster
    - Relies on independence between clusters
    - Allows for arbitrary dependence within clusters
    - CRSEs usually $>$ conventional SEs—use when you suspect clustering
- When X_{ig} not constant within cluster, but just correlated $\rightsquigarrow$ more complicated.
    - See Abadie, Athey, Imbens, and Wooldridge (2021).

Cluster-robust standard errors

- CRSE do not change our estimates $\hat{\boldsymbol{\beta}}$, cannot fix bias
- Valid under **clustered dependence** when main variable is constant within cluster
  - Relies on independence between clusters
  - Allows for arbitrary dependence within clusters
  - CRSEs usually $>$ conventional SEs—use when you suspect clustering
- When X_{ig} not constant within cluster, but just correlated $\rightsquigarrow$ more complicated.
  - See Abadie, Athey, Imbens, and Wooldridge (2021).
- Consistency of the CRSE are in the number of groups, not the number of individuals

Cluster-robust standard errors

- CRSE do not change our estimates $\hat{\beta}$, cannot fix bias
- Valid under **clustered dependence** when main variable is constant within cluster
  - Relies on independence between clusters
  - Allows for arbitrary dependence within clusters
  - CRSEs usually $>$ conventional SEs—use when you suspect clustering
- When X_{ig} not constant within cluster, but just correlated $\rightsquigarrow$ more complicated.
  - See Abadie, Athey, Imbens, and Wooldridge (2021).
- Consistency of the CRSE are in the number of groups, not the number of individuals
  - CRSEs can be incorrect with a small (< 50 maybe) number of clusters