Linear Models Lecture 6: Heteroskedasticity

Robert Gulotty

University of Chicago

February 20, 2026

Theory based vs "Agnostic" strategies

- Two approaches to handling heteroskedasticity:

1) Use theoretical knowledge to specify a corrected model.
2) Use statistical estimators that perform well even when assumptions are violated.

- Which of these two approaches works for you will depend on the complexity of your problem and the development of theory in your field.

Why homoskedastic SEs can be dangerously wrong (Hansen 4.13)

Under homoskedasticity, we estimate $\hat{\boldsymbol{V}}^0_{\hat{\beta}} = (\boldsymbol{X'X})^{-1}s^2$. But the true variance under heteroskedasticity is $\boldsymbol{V}_{\hat{\beta}} = (\boldsymbol{X'X})^{-1}(\boldsymbol{X'DX})(\boldsymbol{X'X})^{-1}$.

Example (Hansen): Suppose $k = 1$, $\sigma_i^2 = X_i^2$, and $\mathbb{E}[X] = 0$. Then:

$$\frac{V_{\hat{\beta}}}{\mathbb{E}\left[\hat{V}^0_{\hat{\beta}}\right]} \approx \frac{\mathbb{E}[X^4]}{(\mathbb{E}[X^2])^2} \stackrel{\text{def}}{=} \kappa$$

- If $X \sim N(0, \sigma^2)$: $\kappa = 3$. True variance is $3\times$ the homoskedastic estimate.
- For *wage* in the CPS: $\kappa = 30$. True variance is $30\times$ the homoskedastic estimate.
- The homoskedastic SE understates uncertainty by a factor of $\sqrt{\kappa}$.

Takeaway: The classical covariance estimator can be wildly misleading. Always use a heteroskedasticity-robust estimator.

Why homoskedastic SEs can be dangerously wrong (Hansen 4.13)

Under homoskedasticity, we estimate $\hat{\boldsymbol{V}}^0_{\hat{\beta}} = (\boldsymbol{X'X})^{-1}s^2$. But the true variance under heteroskedasticity is $\boldsymbol{V}_{\hat{\beta}} = (\boldsymbol{X'X})^{-1}(\boldsymbol{X'DX})(\boldsymbol{X'X})^{-1}$.

Example (Hansen): Suppose $k = 1$, $\sigma_i^2 = X_i^2$, and $\mathbb{E}[X] = 0$. Then:

$$\frac{V_{\hat{\beta}}}{\mathbb{E}\left[\hat{V}^0_{\hat{\beta}}\right]} \approx \frac{\mathbb{E}[X^4]}{(\mathbb{E}[X^2])^2} \stackrel{\text{def}}{=} \kappa$$

- If $X \sim N(0, \sigma^2)$: $\kappa = 3$. True variance is $3\times$ the homoskedastic estimate.
- For *wage* in the CPS: $\kappa = 30$. True variance is $30\times$ the homoskedastic estimate.
- The homoskedastic SE understates uncertainty by a factor of $\sqrt{\kappa}$.

Takeaway: The classical covariance estimator can be wildly misleading. Always use a heteroskedasticity-robust estimator.

Why homoskedastic SEs can be dangerously wrong (Hansen 4.13)

Under homoskedasticity, we estimate $\hat{\boldsymbol{V}}^0_{\hat{\beta}} = (\boldsymbol{X'X})^{-1}s^2$. But the true variance under heteroskedasticity is $\boldsymbol{V}_{\hat{\beta}} = (\boldsymbol{X'X})^{-1}(\boldsymbol{X'DX})(\boldsymbol{X'X})^{-1}$.

Example (Hansen): Suppose $k = 1$, $\sigma_i^2 = X_i^2$, and $\mathbb{E}[X] = 0$. Then:

$$\frac{V_{\hat{\beta}}}{\mathbb{E}\left[\hat{V}^0_{\hat{\beta}}\right]} \approx \frac{\mathbb{E}[X^4]}{(\mathbb{E}[X^2])^2} \stackrel{\text{def}}{=} \kappa$$

- If $X \sim N(0, \sigma^2)$: $\kappa = 3$. True variance is $3\times$ the homoskedastic estimate.
- For *wage* in the CPS: $\kappa = 30$. True variance is $30\times$ the homoskedastic estimate.
- The homoskedastic SE understates uncertainty by a factor of $\sqrt{\kappa}$.

Takeaway: The classical covariance estimator can be wildly misleading. Always use a heteroskedasticity-robust estimator.

Why homoskedastic SEs can be dangerously wrong (Hansen 4.13)

Under homoskedasticity, we estimate $\hat{\boldsymbol{V}}^0_{\hat{\beta}} = (\boldsymbol{X'X})^{-1}s^2$. But the true variance under heteroskedasticity is $\boldsymbol{V}_{\hat{\beta}} = (\boldsymbol{X'X})^{-1}(\boldsymbol{X'DX})(\boldsymbol{X'X})^{-1}$.

Example (Hansen): Suppose $k = 1$, $\sigma_i^2 = X_i^2$, and $\mathbb{E}[X] = 0$. Then:

$$\frac{V_{\hat{\beta}}}{\mathbb{E}\left[\hat{V}^0_{\hat{\beta}}\right]} \approx \frac{\mathbb{E}[X^4]}{(\mathbb{E}[X^2])^2} \stackrel{\text{def}}{=} \kappa$$

- If $X \sim N(0, \sigma^2)$: $\kappa = 3$. True variance is $3\times$ the homoskedastic estimate.
- For *wage* in the CPS: $\kappa = 30$. True variance is $30\times$ the homoskedastic estimate.
- The homoskedastic SE understates uncertainty by a factor of $\sqrt{\kappa}$.

Takeaway: The classical covariance estimator can be wildly misleading. Always use a heteroskedasticity-robust estimator.

"Agnostic" Covariance Matrix Estimation under heteroskedasticity

Suppose we have heteroskedasticity:

$$var(\hat{\beta}) = (\boldsymbol{X'X})^{-1}\boldsymbol{X'DX}(\boldsymbol{X'X})^{-1}$$
$$\boldsymbol{D} = \text{diag}(\sigma_1^2, \ldots \sigma_n^2)$$

If we knew $e_1^2, \ldots e_n^2$, then we could just plug them in:

$$\tilde{var}(\hat{\beta}) = (\boldsymbol{X'X})^{-1}(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' e_i^2)(\boldsymbol{X'X})^{-1}$$

If we use observed $\hat{e}^2$, then this is called the HC0 (heteroskedasticity consistent) estimator, or White covariance matrix estimator.

"Agnostic" Covariance Matrix Estimation under heteroskedasticity

Suppose we have heteroskedasticity:

$$var(\hat{\beta}) = (\boldsymbol{X'X})^{-1}\boldsymbol{X'DX}(\boldsymbol{X'X})^{-1}$$
$$\boldsymbol{D} = \text{diag}(\sigma_1^2, \ldots \sigma_n^2)$$

If we knew $e_1^2, \ldots e_n^2$, then we could just plug them in:

$$\tilde{var}(\hat{\beta}) = (\boldsymbol{X'X})^{-1}(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' e_i^2)(\boldsymbol{X'X})^{-1}$$

If we use observed $\hat{e}^2$, then this is called the HC0 (heteroskedasticity consistent) estimator, or White covariance matrix estimator.

Problems with White covariance estimators

- HC0 is downward biased in finite samples.
- It is missing the $\frac{1}{n-k}$ term for degrees of freedom.

$\rightarrow$ Fixing this gets us HC1.

- It fails to account for leverage:

$$\mathbb{E}[\hat{e}_i^2] = \sigma_i^2(1 - h_{ii}) < \sigma_i^2$$

$\rightarrow$ Fixing this gets us HC2

HC2 and HC3: Leverage-corrected estimators

HC2 (unbiased) and HC3 (conservative) correct for the leverage of each observation:

$$\hat{\boldsymbol{V}}_{\hat{\beta}}^{HC2} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\left(\sum_{i=1}^{n}(1-h_{ii})^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'\hat{e}_i^2\right)(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

$$\hat{\boldsymbol{V}}_{\hat{\beta}}^{HC3} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\left(\sum_{i=1}^{n}(1-h_{ii})^{-2}\boldsymbol{x}_i\boldsymbol{x}_i'\hat{e}_i^2\right)(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

Since $(1-h_{ii})^{-2} > (1-h_{ii})^{-1} > 1$, we always have:

$$\hat{\boldsymbol{V}}_{\hat{\beta}}^{HC0} < \hat{\boldsymbol{V}}_{\hat{\beta}}^{HC2} < \hat{\boldsymbol{V}}_{\hat{\beta}}^{HC3}$$

Implementation in R

Option 1: `estimatr` **package** (recommended for applied work)

```
library(estimatr)
# HC2 is the default -- unbiased under homoskedasticity
lm_robust(y ~ x1 + x2, data = dta, se_type = "HC2")
# HC3 is conservative (biased away from zero)
lm_robust(y ~ x1 + x2, data = dta, se_type = "HC3")
```

Option 2: `sandwich` + `lmtest` (flexible, works with any `lm` object)

```
library(sandwich); library(lmtest)
model <- lm(y ~ x1 + x2, data = dta)
coeftest(model, vcov = vcovHC(model, type = "HC2"))
```

Implementation in R

Option 1: `estimatr` **package** (recommended for applied work)

```
library(estimatr)
# HC2 is the default -- unbiased under homoskedasticity
lm_robust(y ~ x1 + x2, data = dta, se_type = "HC2")
# HC3 is conservative (biased away from zero)
lm_robust(y ~ x1 + x2, data = dta, se_type = "HC3")
```

Option 2: `sandwich` + `lmtest` (flexible, works with any `lm` object)

```
library(sandwich); library(lmtest)
model <- lm(y ~ x1 + x2, data = dta)
coeftest(model, vcov = vcovHC(model, type = "HC2"))
```

R Example: Comparing Standard Errors (Hansen 4.15)

```r
library(estimatr); library(sandwich); library(lmtest)
data(mtcars)
m <- lm(mpg ~ wt + hp, data = mtcars)

# Homoskedastic (classical) SEs
se_classical <- summary(m)$coefficients[, "Std. Error"]

# Robust SEs using sandwich
se_hc0 <- sqrt(diag(vcovHC(m, type = "HC0")))
se_hc1 <- sqrt(diag(vcovHC(m, type = "HC1")))
se_hc2 <- sqrt(diag(vcovHC(m, type = "HC2")))
se_hc3 <- sqrt(diag(vcovHC(m, type = "HC3")))

cbind(Classical = se_classical, HC0 = se_hc0,
      HC1 = se_hc1, HC2 = se_hc2, HC3 = se_hc3)
```

Note: HC0 $<$ HC2 $<$ HC3 always holds. The differences grow when observations have high leverage (h_{ii} close to 1).

Notes on Interpretation

- Robust SEs do **not** affect coefficient estimates or R^2/RMSE.
- They only change standard errors, t-statistics, and confidence intervals.
- Use Wald tests (not classical F-tests) for joint hypotheses:

```
library(car)
linearHypothesis(model, c("x1 = 0", "x2 = 0"),
                 white.adjust = "hc2")
```

- Hansen recommends HC2 (unbiased under homoskedasticity) or HC3 (conservative for any $\boldsymbol{X}$) over HC1.
- In most applications HC1, HC2, HC3 are similar. They diverge when some h_{ii} is large (high-leverage observations).

Notes on Interpretation

- Robust SEs do **not** affect coefficient estimates or R^2/RMSE.
- They only change standard errors, t-statistics, and confidence intervals.
- Use Wald tests (not classical F-tests) for joint hypotheses:

```
library(car)
linearHypothesis(model, c("x1 = 0", "x2 = 0"),
                 white.adjust = "hc2")
```

- Hansen recommends HC2 (unbiased under homoskedasticity) or HC3 (conservative for any $\boldsymbol{X}$) over HC1.
- In most applications HC1, HC2, HC3 are similar. They diverge when some h_{ii} is large (high-leverage observations).

Notes on Interpretation

- Robust SEs do **not** affect coefficient estimates or R^2/RMSE.
- They only change standard errors, t-statistics, and confidence intervals.
- Use Wald tests (not classical F-tests) for joint hypotheses:

```
library(car)
linearHypothesis(model, c("x1 = 0", "x2 = 0"),
                 white.adjust = "hc2")
```

- Hansen recommends HC2 (unbiased under homoskedasticity) or HC3 (conservative for any $\boldsymbol{X}$) over HC1.
- In most applications HC1, HC2, HC3 are similar. They diverge when some h_{ii} is large (high-leverage observations).

HC1 does not work with Sparse Dummy Variabes

- Suppose $Y = \beta_1 D + \beta_2 + e$, where $D_i = 1$ for n_1 cases.
- In the extreme case, $n_1 = 1$:

$$V_{\hat{\beta}} = \sigma^2(\boldsymbol{X'X})^{-1} = \sigma^2 \begin{pmatrix} 1 & 1 \\ 1 & n \end{pmatrix}^{-1} = \sigma^2 \frac{1}{n-1} \begin{pmatrix} n & -1 \\ -1 & 1 \end{pmatrix}$$

$$V_{\hat{\beta}_1} = \sigma^2 \frac{n}{n-1}$$

- Consider the estimator $\hat{\theta} = \hat{\beta}_1 + \hat{\beta}_2$, with variance $\sigma^2 \frac{n}{n-1} + \sigma^2 \frac{1}{n-1} - \sigma^2 \frac{2}{n-1} = \sigma^2$

$$\hat{V}_{\beta}^{HCI} = s^2 \frac{n}{(n-1)^2} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

$$\hat{V}_{\theta}^{HCI} = s^2 \frac{n}{(n-1)^2} + s^2 \frac{n}{(n-1)^2} - s^2 \frac{n}{(n-1)^2} - s^2 \frac{n}{(n-1)^2} + s^2 \frac{n}{(n-1)^2} = 0$$

Measures of Fit (Hansen 4.18)

- $R^2 = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}_Y^2}$ *always increases* when regressors are added $\rightarrow$ cannot be used for model selection.
- $\bar{R}^2$ (adjusted) corrects with $(n-1)/(n-k)$, but Hansen argues it still tends to select models with too many parameters.
- **Recommended:** the leave-one-out cross-validation R^2:

$$\tilde{R}^2 = 1 - \frac{\sum_{i=1}^{n} \tilde{e}_i^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} = 1 - \frac{\bar{\sigma}^2}{\hat{\sigma}_Y^2}$$

  where $\tilde{e}_i = \hat{e}_i/(1 - h_{ii})$ are the prediction errors.
- $\tilde{R}^2$ estimates the percentage of forecast variance explained; it can be *negative* if the model predicts worse than the mean.
- Hansen: "It is recommended to omit R^2 and $\bar{R}^2$. If a measure of fit is desired, report $\tilde{R}^2$ or $\bar{\sigma}^2$."

Measures of Fit (Hansen 4.18)

- $R^2 = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}_Y^2}$ *always increases* when regressors are added $\rightarrow$ cannot be used for model selection.
- $\bar{R}^2$ (adjusted) corrects with $(n-1)/(n-k)$, but Hansen argues it still tends to select models with too many parameters.
- **Recommended:** the leave-one-out cross-validation R^2:

$$\tilde{R}^2 = 1 - \frac{\sum_{i=1}^{n} \tilde{e}_i^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} = 1 - \frac{\bar{\sigma}^2}{\hat{\sigma}_Y^2}$$

  where $\tilde{e}_i = \hat{e}_i/(1 - h_{ii})$ are the prediction errors.
- $\tilde{R}^2$ estimates the percentage of forecast variance explained; it can be *negative* if the model predicts worse than the mean.
- Hansen: "It is recommended to omit R^2 and $\bar{R}^2$. If a measure of fit is desired, report $\tilde{R}^2$ or $\bar{\sigma}^2$."

Measures of Fit (Hansen 4.18)

- $R^2 = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}_Y^2}$ *always increases* when regressors are added $\rightarrow$ cannot be used for model selection.
- $\bar{R}^2$ (adjusted) corrects with $(n-1)/(n-k)$, but Hansen argues it still tends to select models with too many parameters.
- **Recommended:** the leave-one-out cross-validation R^2:

$$\tilde{R}^2 = 1 - \frac{\sum_{i=1}^{n} \tilde{e}_i^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} = 1 - \frac{\bar{\sigma}^2}{\hat{\sigma}_Y^2}$$

where $\tilde{e}_i = \hat{e}_i / (1 - h_{ii})$ are the prediction errors.

- $\tilde{R}^2$ estimates the percentage of forecast variance explained; it can be *negative* if the model predicts worse than the mean.
- Hansen: "It is recommended to omit R^2 and $\bar{R}^2$. If a measure of fit is desired, report $\tilde{R}^2$ or $\bar{\sigma}^2$."

Measures of Fit (Hansen 4.18)

- $R^2 = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}_Y^2}$ *always increases* when regressors are added $\rightarrow$ cannot be used for model selection.
- $\bar{R}^2$ (adjusted) corrects with $(n-1)/(n-k)$, but Hansen argues it still tends to select models with too many parameters.
- **Recommended:** the leave-one-out cross-validation R^2:

$$\tilde{R}^2 = 1 - \frac{\sum_{i=1}^{n} \tilde{e}_i^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} = 1 - \frac{\bar{\sigma}^2}{\hat{\sigma}_Y^2}$$

  where $\tilde{e}_i = \hat{e}_i/(1 - h_{ii})$ are the prediction errors.
- $\tilde{R}^2$ estimates the percentage of forecast variance explained; it can be *negative* if the model predicts worse than the mean.
- Hansen: "It is recommended to omit R^2 and $\bar{R}^2$. If a measure of fit is desired, report $\tilde{R}^2$ or $\bar{\sigma}^2$."

Measures of Fit (Hansen 4.18)

- $R^2 = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}^2_Y}$ *always increases* when regressors are added $\rightarrow$ cannot be used for model selection.
- $\bar{R}^2$ (adjusted) corrects with $(n-1)/(n-k)$, but Hansen argues it still tends to select models with too many parameters.
- **Recommended:** the leave-one-out cross-validation R^2:

$$\tilde{R}^2 = 1 - \frac{\sum_{i=1}^{n} \tilde{e}_i^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = 1 - \frac{\bar{\sigma}^2}{\hat{\sigma}^2_Y}$$

  where $\tilde{e}_i = \hat{e}_i/(1 - h_{ii})$ are the prediction errors.
- $\tilde{R}^2$ estimates the percentage of forecast variance explained; it can be *negative* if the model predicts worse than the mean.
- Hansen: "It is recommended to omit R^2 and $\bar{R}^2$. If a measure of fit is desired, report $\tilde{R}^2$ or $\bar{\sigma}^2$."

Clustered Standard Errors

- Suppose that our data are drawn from groups.

  E.g. students in a school, districts in a state, apartments in a building.
- These observations are subject to common shocks (even if observations don't affect one another).
- These data are called *clustered*.
- We will assume that clusters are known to the researcher and observations are independent across clusters.

Motivating Example: Duflo, Dupas, and Kremer (2011)

In 2005, 140 primary schools in Kenya received funding to hire an extra teacher. Half assigned students to classrooms by prior test score ("tracking").

$$\widehat{TestScore}_{ig} = -0.071 + 0.138\ Tracking_g + \hat{e}_{ig}$$

- With **conventional robust SEs**: $s(\hat{\gamma}) = 0.026$
- With **cluster-robust SEs** (at school level): $s(\hat{\gamma}) = 0.078$

The cluster-robust SEs are $3\times$ larger than the conventional ones!

Ignoring clustering would vastly overstate the precision of the estimated treatment effect. The cluster-robust standard error is the appropriate one because student achievement within a school is correlated.

Motivating Example: Duflo, Dupas, and Kremer (2011)

In 2005, 140 primary schools in Kenya received funding to hire an extra teacher. Half assigned students to classrooms by prior test score ("tracking").

$$\widehat{TestScore}_{ig} = -0.071 + 0.138\ Tracking_g + \hat{e}_{ig}$$

- With **conventional robust SEs**: $s(\hat{\gamma}) = 0.026$
- With **cluster-robust SEs** (at school level): $s(\hat{\gamma}) = 0.078$

The cluster-robust SEs are $3\times$ larger than the conventional ones!

Ignoring clustering would vastly overstate the precision of the estimated treatment effect. The cluster-robust standard error is the appropriate one because student achievement within a school is correlated.

Motivating Example: Duflo, Dupas, and Kremer (2011)

In 2005, 140 primary schools in Kenya received funding to hire an extra teacher. Half assigned students to classrooms by prior test score ("tracking").

$$\widehat{TestScore}_{ig} = -0.071 + 0.138\ Tracking_g + \hat{e}_{ig}$$

- With **conventional robust SEs**: $s(\hat{\gamma}) = 0.026$
- With **cluster-robust SEs** (at school level): $s(\hat{\gamma}) = 0.078$

The cluster-robust SEs are $3\times$ larger than the conventional ones!

Ignoring clustering would vastly overstate the precision of the estimated treatment effect. The cluster-robust standard error is the appropriate one because student achievement within a school is correlated.

Formalism for Clustered Regression

- $(Y_{ig},\ \boldsymbol{x}_{ig})$ where $g = 1, \ldots, G$ indexes a cluster and $i = 1, \ldots, n_g$ indexes individuals in cluster g.
- $\boldsymbol{Y}_g = (Y_{1g}, \ldots Y_{ng})'$, $\boldsymbol{X}_g = (\boldsymbol{x}_{1g}, \ldots \boldsymbol{x}_{ng})'$ at the group level.

$$
\begin{aligned}
Y_{ig} &= \boldsymbol{x}_{ig}'\beta + e_{ig} \\
\hat{\boldsymbol{\beta}} &= \left(\sum_{g=1}^{G} \sum_{i=1}^{n_g} \boldsymbol{x}_{ig} \boldsymbol{x}_{ig}' \right)^{-1} \left(\sum_{g=1}^{G} \sum_{i=1}^{n_g} \boldsymbol{x}_{ig} Y_{ig} \right) \\
&= \left(\sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{X}_g \right)^{-1} \left(\sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{y}_g \right) \\
&= \left(\boldsymbol{X}'\boldsymbol{X} \right)^{-1} \left(\boldsymbol{X}'\boldsymbol{y} \right)
\end{aligned}
$$

Formalism for Clustered Regression

- $(Y_{ig},\ \boldsymbol{x}_{ig})$ where $g = 1, \ldots, G$ indexes a cluster and $i = 1, \ldots, n_g$ indexes individuals in cluster g.
- $\boldsymbol{Y}_g = (Y_{1g}, \ldots Y_{ng})'$, $\boldsymbol{X}_g = (\boldsymbol{x}_{1g}, \ldots \boldsymbol{x}_{ng})'$ at the group level.

$$
\begin{aligned}
Y_{ig} &= \boldsymbol{x}_{ig}'\beta + e_{ig} \\
\hat{\beta} &= \left(\sum_{g=1}^{G} \sum_{i=1}^{n_g} \boldsymbol{x}_{ig} \boldsymbol{x}_{ig}' \right)^{-1} \left(\sum_{g=1}^{G} \sum_{i=1}^{n_g} \boldsymbol{x}_{ig} Y_{ig} \right) \\
&= \left(\sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{X}_g \right)^{-1} \left(\sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{y}_g \right) \\
&= \left(\boldsymbol{X}'\boldsymbol{X} \right)^{-1} \left(\boldsymbol{X}'\boldsymbol{y} \right)
\end{aligned}
$$

Formalism for Clustered Regression

- $(Y_{ig},\ \boldsymbol{x}_{ig})$ where $g = 1, \ldots, G$ indexes a cluster and $i = 1, \ldots, n_g$ indexes individuals in cluster g.
- $\boldsymbol{Y}_g = (Y_{1g}, \ldots Y_{ng})'$, $\boldsymbol{X}_g = (\boldsymbol{x}_{1g}, \ldots \boldsymbol{x}_{ng})'$ at the group level.

$$Y_{ig} = \boldsymbol{x}_{ig}'\beta + e_{ig}$$

$$\hat{\beta} = \left(\sum_{g=1}^{G} \sum_{i=1}^{n_g} \boldsymbol{x}_{ig} \boldsymbol{x}_{ig}' \right)^{-1} \left(\sum_{g=1}^{G} \sum_{i=1}^{n_g} \boldsymbol{x}_{ig} Y_{ig} \right)$$

$$= \left(\sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{X}_g \right)^{-1} \left(\sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{y}_g \right)$$

$$= \left(\boldsymbol{X}'\boldsymbol{X} \right)^{-1} \left(\boldsymbol{X}'\boldsymbol{y} \right)$$

Formalism for Clustered Regression

- $(Y_{ig},\ \boldsymbol{x}_{ig})$ where $g = 1, \ldots, G$ indexes a cluster and $i = 1, \ldots, n_g$ indexes individuals in cluster g.
- $\boldsymbol{Y}_g = (Y_{1g}, \ldots Y_{ng})'$, $\boldsymbol{X}_g = (\boldsymbol{x}_{1g}, \ldots \boldsymbol{x}_{ng})'$ at the group level.

$$Y_{ig} = \boldsymbol{x}_{ig}'\beta + e_{ig}$$

$$\hat{\boldsymbol{\beta}} = \left(\sum_{g=1}^{G} \sum_{i=1}^{n_g} \boldsymbol{x}_{ig} \boldsymbol{x}_{ig}' \right)^{-1} \left(\sum_{g=1}^{G} \sum_{i=1}^{n_g} \boldsymbol{x}_{ig} Y_{ig} \right)$$

$$= \left(\sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{X}_g \right)^{-1} \left(\sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{y}_g \right)$$

$$= \left(\boldsymbol{X}'\boldsymbol{X} \right)^{-1} \left(\boldsymbol{X}'\boldsymbol{y} \right)$$

Formalism for Clustered Regression

- $(Y_{ig},\ \boldsymbol{x}_{ig})$ where $g = 1, \ldots, G$ indexes a cluster and $i = 1, \ldots, n_g$ indexes individuals in cluster g.
- $\boldsymbol{Y}_g = (Y_{1g}, \ldots Y_{ng})'$, $\boldsymbol{X}_g = (\boldsymbol{x}_{1g}, \ldots \boldsymbol{x}_{ng})'$ at the group level.

$$
\begin{aligned}
Y_{ig} &= \boldsymbol{x}_{ig}'\beta + e_{ig} \\
\hat{\boldsymbol{\beta}} &= \left(\sum_{g=1}^{G} \sum_{i=1}^{n_g} \boldsymbol{x}_{ig} \boldsymbol{x}_{ig}' \right)^{-1} \left(\sum_{g=1}^{G} \sum_{i=1}^{n_g} \boldsymbol{x}_{ig} Y_{ig} \right) \\
&= \left(\sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{X}_g \right)^{-1} \left(\sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{y}_g \right) \\
&= \left(\boldsymbol{X}'\boldsymbol{X} \right)^{-1} \left(\boldsymbol{X}'\boldsymbol{y} \right)
\end{aligned}
$$

Variance of Clustered Regression

Call $\mathbf{\Sigma}_g = E\left[\boldsymbol{e}_g \boldsymbol{e}_g'\right]$ the $n_g \times n_g$ covariance in the g cluster.

$$var\left[\left(\sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{e}_g\right)\right] = \sum_{g=1}^{G} var\left[\boldsymbol{X}_g' \boldsymbol{e}_g\right] \quad \text{(By independence across cluster)}$$

$$= \sum_{g=1}^{G} \boldsymbol{X}_g' var\left[\boldsymbol{e}_g\right] \boldsymbol{X}_g$$

$$= \sum_{g=1}^{G} \boldsymbol{X}_g' \Sigma_g \boldsymbol{X}_g$$

$$\equiv \Omega_n$$

$$\boldsymbol{V}_{\hat{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \Omega_n \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$$

Variance of Clustered Regression

Call $\mathbf{\Sigma}_g = E\left[\boldsymbol{e}_g \boldsymbol{e}_g'\right]$ the $n_g \times n_g$ covariance in the g cluster.

$$var\left[\left(\sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{e}_g\right)\right] = \sum_{g=1}^{G} var\left[\boldsymbol{X}_g' \boldsymbol{e}_g\right] \qquad \text{(By independence across cluster)}$$

$$= \sum_{g=1}^{G} \boldsymbol{X}_g' var\left[\boldsymbol{e}_g\right] \boldsymbol{X}_g$$

$$= \sum_{g=1}^{G} \boldsymbol{X}_g' \Sigma_g \boldsymbol{X}_g$$

$$\equiv \Omega_n$$

$$\boldsymbol{V}_{\hat{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \Omega_n \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$$

Variance of Clustered Regression

Call $\mathbf{\Sigma}_g = E\left[\boldsymbol{e}_g \boldsymbol{e}_g'\right]$ the $n_g \times n_g$ covariance in the g cluster.

$$var\left[\left(\sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{e}_g\right)\right] = \sum_{g=1}^{G} var\left[\boldsymbol{X}_g' \boldsymbol{e}_g\right] \quad \text{(By independence across cluster)}$$

$$= \sum_{g=1}^{G} \boldsymbol{X}_g' var\left[\boldsymbol{e}_g\right] \boldsymbol{X}_g$$

$$= \sum_{g=1}^{G} \boldsymbol{X}_g' \Sigma_g \boldsymbol{X}_g$$

$$\equiv \Omega_n$$

$$\boldsymbol{V}_{\hat{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \Omega_n \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$$

Variance of Clustered Regression

Call $\boldsymbol{\Sigma}_g = E\left[\boldsymbol{e}_g \boldsymbol{e}_g'\right]$ the $n_g \times n_g$ covariance in the g cluster.

$$var\left[\left(\sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{e}_g\right)\right] = \sum_{g=1}^{G} var\left[\boldsymbol{X}_g' \boldsymbol{e}_g\right] \qquad \text{(By independence across cluster)}$$

$$= \sum_{g=1}^{G} \boldsymbol{X}_g' var\left[\boldsymbol{e}_g\right] \boldsymbol{X}_g$$

$$= \sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{\Sigma}_g \boldsymbol{X}_g$$

$$\equiv \Omega_n$$

$$\boldsymbol{V}_{\hat{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \Omega_n \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$$

Variance of Clustered Regression

Call $\boldsymbol{\Sigma}_g = E\left[\boldsymbol{e}_g \boldsymbol{e}_g'\right]$ the $n_g \times n_g$ covariance in the g cluster.

$$
\begin{aligned}
var\left[\left(\sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{e}_g\right)\right] &= \sum_{g=1}^{G} var\left[\boldsymbol{X}_g' \boldsymbol{e}_g\right] && \text{(By independence across cluster)} \\
&= \sum_{g=1}^{G} \boldsymbol{X}_g' var\left[\boldsymbol{e}_g\right] \boldsymbol{X}_g \\
&= \sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{\Sigma}_g \boldsymbol{X}_g \\
&\equiv \Omega_n
\end{aligned}
$$

$$
\boldsymbol{V}_{\hat{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \Omega_n \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}
$$

Variance of Clustered Regression

Call $\boldsymbol{\Sigma}_g = E\left[\boldsymbol{e}_g \boldsymbol{e}_g'\right]$ the $n_g \times n_g$ covariance in the g cluster.

$$
\begin{aligned}
var\left[\left(\sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{e}_g\right)\right] &= \sum_{g=1}^{G} var\left[\boldsymbol{X}_g' \boldsymbol{e}_g\right] && \text{(By independence across cluster)} \\
&= \sum_{g=1}^{G} \boldsymbol{X}_g' var\left[\boldsymbol{e}_g\right] \boldsymbol{X}_g \\
&= \sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{\Sigma}_g \boldsymbol{X}_g \\
&\equiv \Omega_n \\
\boldsymbol{V}_{\hat{\beta}} &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \Omega_n \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}
\end{aligned}
$$

Moulton (1990) Formula

Suppose all clusters are equal size N, homoskedastic within cluster $\mathbb{E}[e_{ig}^2] = \sigma^2$, with intra-cluster correlation $\mathbb{E}[e_{ig}e_{\ell g}] = \sigma^2\rho$ for $i \neq \ell$, and X_{ig} does not vary within a cluster:

$$\boldsymbol{V}_{\hat{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2(1+\rho(N-1))$$

The inflation factor $1+\rho(N-1)$ can be enormous:

ρ	Cluster size N	Inflation factor
0.05	20	1.95
0.10	48	5.7
0.25	48	12.75
0.25	100	25.75

Even modest intra-cluster correlation with moderate cluster sizes produces large distortions in conventional SEs.

Moulton (1990) Formula

Suppose all clusters are equal size N, homoskedastic within cluster $\mathbb{E}[e_{ig}^2] = \sigma^2$, with intra-cluster correlation $\mathbb{E}[e_{ig}e_{\ell g}] = \sigma^2\rho$ for $i \neq \ell$, and X_{ig} does not vary within a cluster:

$$\boldsymbol{V}_{\hat{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \sigma^2(1 + \rho(N-1))$$

The inflation factor $1 + \rho(N-1)$ can be enormous:

ρ	Cluster size N	Inflation factor
0.05	20	1.95
0.10	48	5.7
0.25	48	12.75
0.25	100	25.75

Even modest intra-cluster correlation with moderate cluster sizes produces large distortions in conventional SEs.

Strategy 1: Modeling intra-cluster dependence

- A random-effects model assumes:

$$u_{gi} = \lambda e_g + e_{gi}$$

- Where $e_{gi} \sim f(0,\ \omega^2)$ is individual specific, $e_g \sim f(0, 1)$ is cluster wide and the two are independent.

$$\Omega_g = \begin{bmatrix} \lambda^2 + \omega^2 & \lambda^2 & \dots & \lambda^2 \\ \lambda^2 & \lambda^2 + \omega^2 & \dots & \lambda^2 \\ \lambda^2 & \lambda^2 & \dots & \lambda^2 + \omega^2 \end{bmatrix}$$

FGLS

- Estimate pooled model of $\boldsymbol{y}$ on $\boldsymbol{X}$ using OLS. Record $\hat{\sigma}^2$
- Estimate model with fixed effects: regress $\boldsymbol{y}$ on $[\boldsymbol{X}\ \boldsymbol{D}]$. This is our estimate of ω^2, $\hat{\omega}^2$.
- $\hat{\lambda}^2 = \hat{\sigma}^2 - \hat{\omega}^2$.
- Plug into GLS for Σ.
- The Random Effects Estimator is consistent, biased, and asymptotically efficient (if the errors are correct).
- However, if we needed group fixed effects in $\boldsymbol{X}$, we already estimate λe_g, and this model will not work.

FGLS

- Estimate pooled model of $\boldsymbol{y}$ on $\boldsymbol{X}$ using OLS. Record $\hat{\sigma}^2$
- Estimate model with fixed effects: regress $\boldsymbol{y}$ on $[\boldsymbol{X}\ \boldsymbol{D}]$. This is our estimate of ω^2, $\hat{\omega}^2$.
- $\hat{\lambda}^2 = \hat{\sigma}^2 - \hat{\omega}^2$.
- Plug into GLS for Σ.
- The Random Effects Estimator is consistent, biased, and asymptotically efficient (if the errors are correct).
- However, if we needed group fixed effects in $\boldsymbol{X}$, we already estimate λe_g, and this model will not work.

FGLS

- Estimate pooled model of $\boldsymbol{y}$ on $\boldsymbol{X}$ using OLS. Record $\hat{\sigma}^2$
- Estimate model with fixed effects: regress $\boldsymbol{y}$ on $[\boldsymbol{X}\ \boldsymbol{D}]$. This is our estimate of ω^2, $\hat{\omega}^2$.
- $\hat{\lambda}^2 = \hat{\sigma}^2 - \hat{\omega}^2$.
- Plug into GLS for Σ.
- The Random Effects Estimator is consistent, biased, and asymptotically efficient (if the errors are correct).
- However, if we needed group fixed effects in $\boldsymbol{X}$, we already estimate λe_g, and this model will not work.

FGLS

- Estimate pooled model of $\boldsymbol{y}$ on $\boldsymbol{X}$ using OLS. Record $\hat{\sigma}^2$
- Estimate model with fixed effects: regress $\boldsymbol{y}$ on $[\boldsymbol{X}\ \boldsymbol{D}]$. This is our estimate of ω^2, $\hat{\omega}^2$.
- $\hat{\lambda}^2 = \hat{\sigma}^2 - \hat{\omega}^2$.
- Plug into GLS for Σ.
- The Random Effects Estimator is consistent, biased, and asymptotically efficient (if the errors are correct).
- However, if we needed group fixed effects in $\boldsymbol{X}$, we already estimate λe_g, and this model will not work.

FGLS

- Estimate pooled model of $\boldsymbol{y}$ on $\boldsymbol{X}$ using OLS. Record $\hat{\sigma}^2$
- Estimate model with fixed effects: regress $\boldsymbol{y}$ on $[\boldsymbol{X}\ \boldsymbol{D}]$. This is our estimate of ω^2, $\hat{\omega}^2$.
- $\hat{\lambda}^2 = \hat{\sigma}^2 - \hat{\omega}^2$.
- Plug into GLS for Σ.
- The Random Effects Estimator is consistent, biased, and asymptotically efficient (if the errors are correct).
- However, if we needed group fixed effects in $\boldsymbol{X}$, we already estimate λe_g, and this model will not work.

FGLS

- Estimate pooled model of $\boldsymbol{y}$ on $\boldsymbol{X}$ using OLS. Record $\hat{\sigma}^2$
- Estimate model with fixed effects: regress $\boldsymbol{y}$ on $[\boldsymbol{X}\ \boldsymbol{D}]$. This is our estimate of ω^2, $\hat{\omega}^2$.
- $\hat{\lambda}^2 = \hat{\sigma}^2 - \hat{\omega}^2$.
- Plug into GLS for Σ.
- The Random Effects Estimator is consistent, biased, and asymptotically efficient (if the errors are correct).
- However, if we needed group fixed effects in $\boldsymbol{X}$, we already estimate λe_g, and this model will not work.

Intra-cluster dependence that survives fixed effects

- Contrast the random effects model with a factor model

$$u_{gi} = \lambda_{gi} e_g + e_{gi}$$

- Where $e_{gi} \sim f(0,\ \omega^2)$ is individual specific, $e_g \sim f(0, 1)$ is cluster wide and the two are independent, but now λ_{gi} depends on the individual.
- For example, some students may be affected more by teacher quality than others.
- Now if we use cluster fixed effects:

$$u_{gi} - \bar{u}_g = (\lambda_{gi} - \bar{\lambda}_g) e_g + e_{gi} - \bar{e}_g$$

$$cov(u_{gi} - \bar{u}_g, u_{gl} - \bar{u}_g) = (\lambda_{gi} - \bar{\lambda}_g)(\lambda_{gl} - \bar{\lambda}_g)$$

- Which is zero only if λ_{gi} is the same for all i.
- Unfortunately, it is not obvious how to implement this model.

Intra-cluster dependence that survives fixed effects

- Contrast the random effects model with a factor model

$$u_{gi} = \lambda_{gi} e_g + e_{gi}$$

- Where $e_{gi} \sim f(0,\ \omega^2)$ is individual specific, $e_g \sim f(0, 1)$ is cluster wide and the two are independent, but now λ_{gi} depends on the individual.
- For example, some students may be affected more by teacher quality than others.
- Now if we use cluster fixed effects:

$$u_{gi} - \bar{u}_g = (\lambda_{gi} - \bar{\lambda}_g) e_g + e_{gi} - \bar{e}_g$$

$$cov(u_{gi} - \bar{u}_g, u_{gl} - \bar{u}_g) = (\lambda_{gi} - \bar{\lambda}_g)(\lambda_{gl} - \bar{\lambda}_g)$$

- Which is zero only if λ_{gi} is the same for all i.
- Unfortunately, it is not obvious how to implement this model.

Intra-cluster dependence that survives fixed effects

- Contrast the random effects model with a factor model

$$u_{gi} = \lambda_{gi} e_g + e_{gi}$$

- Where $e_{gi} \sim f(0,\ \omega^2)$ is individual specific, $e_g \sim f(0, 1)$ is cluster wide and the two are independent, but now λ_{gi} depends on the individual.
- For example, some students may be affected more by teacher quality than others.
- Now if we use cluster fixed effects:

$$u_{gi} - \bar{u}_g = (\lambda_{gi} - \bar{\lambda}_g) e_g + e_{gi} - \bar{e}_g$$

$$cov(u_{gi} - \bar{u}_g, u_{gl} - \bar{u}_g) = (\lambda_{gi} - \bar{\lambda}_g)(\lambda_{gl} - \bar{\lambda}_g)$$

- Which is zero only if λ_{gi} is the same for all i.
- Unfortunately, it is not obvious how to implement this model.

Intra-cluster dependence that survives fixed effects

- Contrast the random effects model with a factor model

$$u_{gi} = \lambda_{gi} e_g + e_{gi}$$

- Where $e_{gi} \sim f(0,\ \omega^2)$ is individual specific, $e_g \sim f(0, 1)$ is cluster wide and the two are independent, but now λ_{gi} depends on the individual.
- For example, some students may be affected more by teacher quality than others.
- Now if we use cluster fixed effects:

$$u_{gi} - \bar{u}_g = (\lambda_{gi} - \bar{\lambda}_g) e_g + e_{gi} - \bar{e}_g$$

$$cov(u_{gi} - \bar{u}_g, u_{gl} - \bar{u}_g) = (\lambda_{gi} - \bar{\lambda}_g)(\lambda_{gl} - \bar{\lambda}_g)$$

- Which is zero only if λ_{gi} is the same for all i.
- Unfortunately, it is not obvious how to implement this model.

Intra-cluster dependence that survives fixed effects

- Contrast the random effects model with a factor model

$$u_{gi} = \lambda_{gi} e_g + e_{gi}$$

- Where $e_{gi} \sim f(0,\ \omega^2)$ is individual specific, $e_g \sim f(0, 1)$ is cluster wide and the two are independent, but now λ_{gi} depends on the individual.
- For example, some students may be affected more by teacher quality than others.
- Now if we use cluster fixed effects:

$$u_{gi} - \bar{u}_g = (\lambda_{gi} - \bar{\lambda}_g) e_g + e_{gi} - \bar{e}_g$$

$$cov(u_{gi} - \bar{u}_g, u_{gl} - \bar{u}_g) = (\lambda_{gi} - \bar{\lambda}_g)(\lambda_{gl} - \bar{\lambda}_g)$$

- Which is zero only if λ_{gi} is the same for all i.
- Unfortunately, it is not obvious how to implement this model.

Intra-cluster dependence that survives fixed effects

- Contrast the random effects model with a factor model

$$u_{gi} = \lambda_{gi} e_g + e_{gi}$$

- Where $e_{gi} \sim f(0,\ \omega^2)$ is individual specific, $e_g \sim f(0, 1)$ is cluster wide and the two are independent, but now λ_{gi} depends on the individual.
- For example, some students may be affected more by teacher quality than others.
- Now if we use cluster fixed effects:

$$u_{gi} - \bar{u}_g = (\lambda_{gi} - \bar{\lambda}_g) e_g + e_{gi} - \bar{e}_g$$

$$cov(u_{gi} - \bar{u}_g, u_{gl} - \bar{u}_g) = (\lambda_{gi} - \bar{\lambda}_g)(\lambda_{gl} - \bar{\lambda}_g)$$

- Which is zero only if λ_{gi} is the same for all i.
- Unfortunately, it is not obvious how to implement this model.

Strategy 2: Cluster Robust Standard Errors Arellano (1987), Hansen (2007)

The squared error e_i^2 is an unbiased estimate for $E[e_i^2]$, $\boldsymbol{e}_g \boldsymbol{e}_g'$ is unbiased for $E[\boldsymbol{e}_g \boldsymbol{e}_g']$. As with White, we can estimate e_i^2 with $\hat{e}_i^2$:

$$\hat{\Omega}_n = \sum_{g=1}^{G} \boldsymbol{X}_g' \hat{\boldsymbol{e}}_g \hat{\boldsymbol{e}}_g' \boldsymbol{X}_g$$

$$\hat{\boldsymbol{V}}_{\hat{\beta}} = a_n \left(\boldsymbol{X}\boldsymbol{X}'\right)^{-1} \hat{\Omega}_n \left(\boldsymbol{X}\boldsymbol{X}'\right)^{-1}$$

$$a_n = \left(\frac{n-1}{n-k}\right)\left(\frac{G}{G-1}\right)$$

a_n is the small sample correction used by STATA, recommended by Hansen (2007).

Clustered SEs in R

Option 1: `estimatr` (simplest)

```
library(estimatr)
lm_robust(y ~ x1 + x2, data=dta, clusters=cluster_id, se_type="CR2")
```

Option 2: `sandwich` + `lmtest`

```
library(sandwich); library(lmtest)
model <- lm(y ~ x1 + x2, data = dta)
coeftest(model, vcov = vcovCL(model, cluster = dta$cluster_id))
```

Option 3: Manual construction (Hansen Ch. 4 R code)

```
xe <- x * rep(e, times = k)        # X_i * e_i
xe_sum <- rowsum(xe, cluster_id)  # sum within clusters
G <- nrow(xe_sum); omega <- t(xe_sum) %*% xe_sum
scale <- G/(G-1) * (n-1)/(n-k)
V_cl <- scale * invx %*% omega %*% invx
```

Clustered SEs in R

Option 1: `estimatr` (simplest)

```
library(estimatr)
lm_robust(y ~ x1 + x2, data=dta, clusters=cluster_id, se_type="CR2")
```

Option 2: `sandwich` + `lmtest`

```
library(sandwich); library(lmtest)
model <- lm(y ~ x1 + x2, data = dta)
coeftest(model, vcov = vcovCL(model, cluster = dta$cluster_id))
```

Option 3: Manual construction (Hansen Ch. 4 R code)

```
xe <- x * rep(e, times = k)        # X_i * e_i
xe_sum <- rowsum(xe, cluster_id)   # sum within clusters
G <- nrow(xe_sum); omega <- t(xe_sum) %*% xe_sum
scale <- G/(G-1) * (n-1)/(n-k)
V_cl <- scale * invx %*% omega %*% invx
```

Clustered SEs in R

Option 1: `estimatr` (simplest)

```
library(estimatr)
lm_robust(y ~ x1 + x2, data=dta, clusters=cluster_id, se_type="CR2")
```

Option 2: `sandwich` + `lmtest`

```
library(sandwich); library(lmtest)
model <- lm(y ~ x1 + x2, data = dta)
coeftest(model, vcov = vcovCL(model, cluster = dta$cluster_id))
```

Option 3: Manual construction (Hansen Ch. 4 R code)

```
xe <- x * rep(e, times = k)        # X_i * e_i
xe_sum <- rowsum(xe, cluster_id)  # sum within clusters
G <- nrow(xe_sum); omega <- t(xe_sum) %*% xe_sum
scale <- G/(G-1) * (n-1)/(n-k)
V_cl <- scale * invx %*% omega %*% invx
```

Inference with Clustered Samples (Hansen 4.22)

- The effective sample size for cluster-robust inference is G (number of clusters), **not** n (number of observations).
- The cluster-robust estimator treats each cluster as a single observation and estimates the covariance from the variation across cluster means.
- If $G = 50$, inference quality is comparable to heteroskedasticity-robust inference with $n = 50$.
- Most cluster-robust theory assumes **homogeneous** cluster sizes. When cluster sizes are highly unequal, cluster sums have heterogeneous variances $\rightarrow$ compounding problem.
- When the number of *treated* clusters is small (e.g. only a few schools received treatment), the cluster-robust SE on the treatment coefficient can be severely downward biased—analogous to the sparse dummy variable problem (Section 4.16).

Inference with Clustered Samples (Hansen 4.22)

- The effective sample size for cluster-robust inference is G (number of clusters), **not** n (number of observations).
- The cluster-robust estimator treats each cluster as a single observation and estimates the covariance from the variation across cluster means.
- If $G = 50$, inference quality is comparable to heteroskedasticity-robust inference with $n = 50$.
- Most cluster-robust theory assumes **homogeneous** cluster sizes. When cluster sizes are highly unequal, cluster sums have heterogeneous variances → compounding problem.
- When the number of *treated* clusters is small (e.g. only a few schools received treatment), the cluster-robust SE on the treatment coefficient can be severely downward biased—analogous to the sparse dummy variable problem (Section 4.16).

Inference with Clustered Samples (Hansen 4.22)

- The effective sample size for cluster-robust inference is G (number of clusters), **not** n (number of observations).
- The cluster-robust estimator treats each cluster as a single observation and estimates the covariance from the variation across cluster means.
- If $G = 50$, inference quality is comparable to heteroskedasticity-robust inference with $n = 50$.
- Most cluster-robust theory assumes **homogeneous** cluster sizes. When cluster sizes are highly unequal, cluster sums have heterogeneous variances → compounding problem.
- When the number of *treated* clusters is small (e.g. only a few schools received treatment), the cluster-robust SE on the treatment coefficient can be severely downward biased—analogous to the sparse dummy variable problem (Section 4.16).

Inference with Clustered Samples (Hansen 4.22)

- The effective sample size for cluster-robust inference is G (number of clusters), **not** n (number of observations).
- The cluster-robust estimator treats each cluster as a single observation and estimates the covariance from the variation across cluster means.
- If $G = 50$, inference quality is comparable to heteroskedasticity-robust inference with $n = 50$.
- Most cluster-robust theory assumes **homogeneous** cluster sizes. When cluster sizes are highly unequal, cluster sums have heterogeneous variances $\rightarrow$ compounding problem.
- When the number of *treated* clusters is small (e.g. only a few schools received treatment), the cluster-robust SE on the treatment coefficient can be severely downward biased—analogous to the sparse dummy variable problem (Section 4.16).

Inference with Clustered Samples (Hansen 4.22)

- The effective sample size for cluster-robust inference is G (number of clusters), **not** n (number of observations).
- The cluster-robust estimator treats each cluster as a single observation and estimates the covariance from the variation across cluster means.
- If $G = 50$, inference quality is comparable to heteroskedasticity-robust inference with $n = 50$.
- Most cluster-robust theory assumes **homogeneous** cluster sizes. When cluster sizes are highly unequal, cluster sums have heterogeneous variances $\rightarrow$ compounding problem.
- When the number of *treated* clusters is small (e.g. only a few schools received treatment), the cluster-robust SE on the treatment coefficient can be severely downward biased—analogous to the sparse dummy variable problem (Section 4.16).

At what level ought one cluster? (Hansen 4.23)

- Should we cluster by individual, county, state, or region?
- There is a **bias–variance tradeoff**:
  - **Too fine** (e.g. household instead of village): omits covariance terms $\rightarrow$ SEs biased *downward*, spurious significance.
  - **Too coarse** (e.g. state instead of county): adds noise $\rightarrow$ SEs imprecise, less power.
- Rules of thumb:
  - Cluster at the level where treatment is assigned.
  - Cluster at the coarsest level defensible by theory, provided G is not too small ($G \geq 50$ preferred).
  - Your effective sample size is G, not n.
- Honest assessment (Hansen): "We really do not know what is the 'correct' level at which to do cluster-robust inference."

At what level ought one cluster? (Hansen 4.23)

- Should we cluster by individual, county, state, or region?
- There is a **bias–variance tradeoff**:
  - **Too fine** (e.g. household instead of village): omits covariance terms $\rightarrow$ SEs biased *downward*, spurious significance.
  - **Too coarse** (e.g. state instead of county): adds noise $\rightarrow$ SEs imprecise, less power.
- Rules of thumb:
  - Cluster at the level where treatment is assigned.
  - Cluster at the coarsest level defensible by theory, provided G is not too small ($G \geq 50$ preferred).
  - Your effective sample size is G, not n.
- Honest assessment (Hansen): "We really do not know what is the 'correct' level at which to do cluster-robust inference."

At what level ought one cluster? (Hansen 4.23)

- Should we cluster by individual, county, state, or region?
- There is a **bias–variance tradeoff**:
  - **Too fine** (e.g. household instead of village): omits covariance terms $\rightarrow$ SEs biased *downward*, spurious significance.
  - **Too coarse** (e.g. state instead of county): adds noise $\rightarrow$ SEs imprecise, less power.
- Rules of thumb:
  - Cluster at the level where treatment is assigned.
  - Cluster at the coarsest level defensible by theory, provided G is not too small ($G \geq 50$ preferred).
  - Your effective sample size is G, not n.
- Honest assessment (Hansen): "We really do not know what is the 'correct' level at which to do cluster-robust inference."

At what level ought one cluster? (Hansen 4.23)

- Should we cluster by individual, county, state, or region?
- There is a **bias–variance tradeoff**:
  - **Too fine** (e.g. household instead of village): omits covariance terms $\rightarrow$ SEs biased *downward*, spurious significance.
  - **Too coarse** (e.g. state instead of county): adds noise $\rightarrow$ SEs imprecise, less power.
- Rules of thumb:
  - Cluster at the level where treatment is assigned.
  - Cluster at the coarsest level defensible by theory, provided G is not too small ($G \geq 50$ preferred).
  - Your effective sample size is G, not n.
- Honest assessment (Hansen): "We really do not know what is the 'correct' level at which to do cluster-robust inference."

At what level ought one cluster? (Hansen 4.23)

- Should we cluster by individual, county, state, or region?
- There is a **bias–variance tradeoff**:
  - **Too fine** (e.g. household instead of village): omits covariance terms $\rightarrow$ SEs biased *downward*, spurious significance.
  - **Too coarse** (e.g. state instead of county): adds noise $\rightarrow$ SEs imprecise, less power.
- Rules of thumb:
  - Cluster at the level where treatment is assigned.
  - Cluster at the coarsest level defensible by theory, provided G is not too small ($G \geq 50$ preferred).
  - Your effective sample size is G, not n.
- Honest assessment (Hansen): "We really do not know what is the 'correct' level at which to do cluster-robust inference."

Practical Recommendations (Hansen Ch. 4 Summary)

1. **Always report robust standard errors.** The classical homoskedastic formula $s^2(\boldsymbol{X'X})^{-1}$ is only valid under the strong (and rarely true) assumption of conditional homoskedasticity.
2. **Use HC2 by default** for cross-sectional data. It is unbiased under homoskedasticity and performs well under heteroskedasticity. Use HC3 if you want a conservative alternative.
3. **If data are clustered**, use cluster-robust SEs. Your effective sample size is G (clusters), not n (observations). Cluster at the coarsest defensible level.
4. **For measures of fit**, prefer $\tilde{R}^2$ (leave-one-out cross-validation) over R^2 or $\bar{R}^2$. Hansen: "It is recommended to omit R^2 and $\bar{R}^2$."
5. **Watch for sparse dummies and high leverage.** These can cause robust SE estimators to break down or be severely biased.

Practical Recommendations (Hansen Ch. 4 Summary)

1. **Always report robust standard errors.** The classical homoskedastic formula $s^2(\boldsymbol{X'X})^{-1}$ is only valid under the strong (and rarely true) assumption of conditional homoskedasticity.
2. **Use HC2 by default** for cross-sectional data. It is unbiased under homoskedasticity and performs well under heteroskedasticity. Use HC3 if you want a conservative alternative.
3. **If data are clustered**, use cluster-robust SEs. Your effective sample size is G (clusters), not n (observations). Cluster at the coarsest defensible level.
4. **For measures of fit**, prefer $\tilde{R}^2$ (leave-one-out cross-validation) over R^2 or $\bar{R}^2$. Hansen: "It is recommended to omit R^2 and $\bar{R}^2$."
5. **Watch for sparse dummies and high leverage.** These can cause robust SE estimators to break down or be severely biased.

Practical Recommendations (Hansen Ch. 4 Summary)

1. **Always report robust standard errors.** The classical homoskedastic formula $s^2(\boldsymbol{X'X})^{-1}$ is only valid under the strong (and rarely true) assumption of conditional homoskedasticity.
2. **Use HC2 by default** for cross-sectional data. It is unbiased under homoskedasticity and performs well under heteroskedasticity. Use HC3 if you want a conservative alternative.
3. **If data are clustered**, use cluster-robust SEs. Your effective sample size is G (clusters), not n (observations). Cluster at the coarsest defensible level.
4. **For measures of fit**, prefer $\tilde{R}^2$ (leave-one-out cross-validation) over R^2 or $\bar{R}^2$. Hansen: "It is recommended to omit R^2 and $\bar{R}^2$."
5. **Watch for sparse dummies and high leverage.** These can cause robust SE estimators to break down or be severely biased.

Practical Recommendations (Hansen Ch. 4 Summary)

1. **Always report robust standard errors.** The classical homoskedastic formula $s^2(\boldsymbol{X'X})^{-1}$ is only valid under the strong (and rarely true) assumption of conditional homoskedasticity.
2. **Use HC2 by default** for cross-sectional data. It is unbiased under homoskedasticity and performs well under heteroskedasticity. Use HC3 if you want a conservative alternative.
3. **If data are clustered**, use cluster-robust SEs. Your effective sample size is G (clusters), not n (observations). Cluster at the coarsest defensible level.
4. **For measures of fit**, prefer $\tilde{R}^2$ (leave-one-out cross-validation) over R^2 or $\bar{R}^2$. Hansen: "It is recommended to omit R^2 and $\bar{R}^2$."
5. **Watch for sparse dummies and high leverage.** These can cause robust SE estimators to break down or be severely biased.

Practical Recommendations (Hansen Ch. 4 Summary)

1. **Always report robust standard errors.** The classical homoskedastic formula $s^2(\boldsymbol{X'X})^{-1}$ is only valid under the strong (and rarely true) assumption of conditional homoskedasticity.
2. **Use HC2 by default** for cross-sectional data. It is unbiased under homoskedasticity and performs well under heteroskedasticity. Use HC3 if you want a conservative alternative.
3. **If data are clustered**, use cluster-robust SEs. Your effective sample size is G (clusters), not n (observations). Cluster at the coarsest defensible level.
4. **For measures of fit**, prefer $\tilde{R}^2$ (leave-one-out cross-validation) over R^2 or $\bar{R}^2$. Hansen: "It is recommended to omit R^2 and $\bar{R}^2$."
5. **Watch for sparse dummies and high leverage.** These can cause robust SE estimators to break down or be severely biased.

Multicollinearity (Hansen 4.20)

- **Strict multicollinearity**: $\boldsymbol{X'X}$ is singular $\rightarrow \hat{\beta}$ is undefined. Typically caused by redundant fixed effects or a dummy variable trap.
- **Near multicollinearity**: regressors are highly correlated. With two regressors and correlation ρ:

$$\text{var}[\hat{\beta}_j|\boldsymbol{X}] = \frac{\sigma^2}{n(1-\rho^2)}$$

  As $\rho \rightarrow 1$, the variance $\rightarrow \infty$.
- This is *not* a bias problem—it is purely a precision problem, equivalent to having a small sample (Goldberger's "micronumerosity" critique).
- Robust standard errors can be sensitive to high leverage under near multicollinearity, producing misleadingly small SEs even when coefficient estimates are imprecise.

Multicollinearity (Hansen 4.20)

- **Strict multicollinearity**: $\boldsymbol{X'X}$ is singular $\rightarrow \hat{\beta}$ is undefined. Typically caused by redundant fixed effects or a dummy variable trap.
- **Near multicollinearity**: regressors are highly correlated. With two regressors and correlation ρ:

$$\text{var}[\hat{\beta}_j|\boldsymbol{X}] = \frac{\sigma^2}{n(1-\rho^2)}$$

  As $\rho \rightarrow 1$, the variance $\rightarrow \infty$.
- This is *not* a bias problem—it is purely a precision problem, equivalent to having a small sample (Goldberger's "micronumerosity" critique).
- Robust standard errors can be sensitive to high leverage under near multicollinearity, producing misleadingly small SEs even when coefficient estimates are imprecise.

Multicollinearity (Hansen 4.20)

- **Strict multicollinearity**: $\boldsymbol{X'X}$ is singular $\rightarrow \hat{\beta}$ is undefined. Typically caused by redundant fixed effects or a dummy variable trap.
- **Near multicollinearity**: regressors are highly correlated. With two regressors and correlation ρ:

$$\text{var}[\hat{\beta}_j|\boldsymbol{X}] = \frac{\sigma^2}{n(1-\rho^2)}$$

  As $\rho \rightarrow 1$, the variance $\rightarrow \infty$.
- This is *not* a bias problem—it is purely a precision problem, equivalent to having a small sample (Goldberger's "micronumerosity" critique).
- Robust standard errors can be sensitive to high leverage under near multicollinearity, producing misleadingly small SEs even when coefficient estimates are imprecise.

Multicollinearity (Hansen 4.20)

- **Strict multicollinearity**: $\boldsymbol{X'X}$ is singular $\rightarrow \hat{\beta}$ is undefined. Typically caused by redundant fixed effects or a dummy variable trap.
- **Near multicollinearity**: regressors are highly correlated. With two regressors and correlation ρ:

$$\text{var}[\hat{\beta}_j|\boldsymbol{X}] = \frac{\sigma^2}{n(1-\rho^2)}$$

  As $\rho \rightarrow 1$, the variance $\rightarrow \infty$.
- This is *not* a bias problem—it is purely a precision problem, equivalent to having a small sample (Goldberger's "micronumerosity" critique).
- Robust standard errors can be sensitive to high leverage under near multicollinearity, producing misleadingly small SEs even when coefficient estimates are imprecise.

Ridge Regression

- Suppose $\mathbb{E}[\boldsymbol{y}|\boldsymbol{X}] = \boldsymbol{X}\beta$.
- The Ridge regression estimator, given a constant $\lambda > 0$ is

$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{I}_k\lambda)^{-1}\boldsymbol{X}'\boldsymbol{y}$$

$$\begin{aligned}\mathbb{E}[\hat{\beta}|\boldsymbol{X}] &= \mathbb{E}[(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{I}_k\lambda)^{-1}\boldsymbol{X}'\boldsymbol{y}|\boldsymbol{X}] \\ &= (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{I}_k\lambda)^{-1}\boldsymbol{X}'\mathbb{E}[\boldsymbol{y}|\boldsymbol{X}] \\ &= (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{I}_k\lambda)^{-1}\boldsymbol{X}'\boldsymbol{X}\beta\end{aligned}$$

- Ridge purposefully introduces bias to reduce variance (MASS::lm.ridge)
- Note $\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{I}_k\lambda$ is always full rank, so you can use ridge even if $k > n$.

Ridge Regression

- Suppose $\mathbb{E}[\boldsymbol{y}|\boldsymbol{X}] = \boldsymbol{X}\beta$.
- The Ridge regression estimator, given a constant $\lambda > 0$ is

$$\begin{aligned}\hat{\beta} &= (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{I}_k\lambda)^{-1}\boldsymbol{X}'\boldsymbol{y} \\ \mathbb{E}[\hat{\beta}|\boldsymbol{X}] &= \mathbb{E}[(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{I}_k\lambda)^{-1}\boldsymbol{X}'\boldsymbol{y}|\boldsymbol{X}] \\ &= (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{I}_k\lambda)^{-1}\boldsymbol{X}'\mathbb{E}[\boldsymbol{y}|\boldsymbol{X}] \\ &= (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{I}_k\lambda)^{-1}\boldsymbol{X}'\boldsymbol{X}\beta\end{aligned}$$

- Ridge purposefully introduces bias to reduce variance (MASS::lm.ridge)
- Note $\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{I}_k\lambda$ is always full rank, so you can use ridge even if $k > n$.

Ridge Regression

- Suppose $\mathbb{E}[\boldsymbol{y}|\boldsymbol{X}] = \boldsymbol{X}\beta$.
- The Ridge regression estimator, given a constant $\lambda > 0$ is

$$\begin{aligned}\hat{\beta} &= (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{I}_k\lambda)^{-1}\boldsymbol{X}'\boldsymbol{y} \\ \mathbb{E}[\hat{\beta}|\boldsymbol{X}] &= \mathbb{E}[(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{I}_k\lambda)^{-1}\boldsymbol{X}'\boldsymbol{y}|\boldsymbol{X}] \\ &= (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{I}_k\lambda)^{-1}\boldsymbol{X}'\mathbb{E}[\boldsymbol{y}|\boldsymbol{X}] \\ &= (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{I}_k\lambda)^{-1}\boldsymbol{X}'\boldsymbol{X}\beta\end{aligned}$$

- Ridge purposefully introduces bias to reduce variance (MASS::lm.ridge)
- Note $\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{I}_k\lambda$ is always full rank, so you can use ridge even if $k > n$.

Ridge Regression

- Suppose $\mathbb{E}[\boldsymbol{y}|\boldsymbol{X}] = \boldsymbol{X}\beta$.
- The Ridge regression estimator, given a constant $\lambda > 0$ is

$$\begin{aligned}\hat{\beta} &= (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{I}_k\lambda)^{-1}\boldsymbol{X}'\boldsymbol{y} \\ \mathbb{E}[\hat{\beta}|\boldsymbol{X}] &= \mathbb{E}[(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{I}_k\lambda)^{-1}\boldsymbol{X}'\boldsymbol{y}|\boldsymbol{X}] \\ &= (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{I}_k\lambda)^{-1}\boldsymbol{X}'\mathbb{E}[\boldsymbol{y}|\boldsymbol{X}] \\ &= (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{I}_k\lambda)^{-1}\boldsymbol{X}'\boldsymbol{X}\beta\end{aligned}$$

- Ridge purposefully introduces bias to reduce variance (MASS::lm.ridge)
- Note $\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{I}_k\lambda$ is always full rank, so you can use ridge even if $k > n$.