1 OLS Properties Under Gauss-Markov Assumptions

1. Linearity: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$
2. Random/iid sample: $(y_i, \mathbf{x}_i')$ are a iid sample from the population.
3. No perfect collinearity: $\mathbf{X}$ is an $n \times (k+1)$ matrix with rank $k+1$
4. Zero conditional mean: $E[\mathbf{u}|\mathbf{X}] = \mathbf{0}$
5. Homoskedasticity: $\text{var}(\mathbf{u}|\mathbf{X}) = \sigma_u^2 \mathbf{I}_n$
6. Normality: $\mathbf{u}|\mathbf{X} \sim N(\mathbf{0}, \sigma_u^2 \mathbf{I}_n)$

The rank of a matrix is the maximum number of linearly independent columns.

- If $\mathbf{X}$ has rank $k+1$, then all of its columns are linearly independent
- ...If all of the columns are linearly independent, then the assumption of no perfect collinearity hold.
- If $\mathbf{X}$ has rank $k+1$, then $(\mathbf{X}'\mathbf{X})$ is invertible
  (see essentially any linear algebra book for proof)
- Just like variation in X led us to be able to divide by the variance in simple OLS

2 Unbiasedness of $\hat{\boldsymbol{\beta}}$

Is $\hat{\boldsymbol{\beta}}$ still unbiased under assumptions 1-4? Does $E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$?

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \text{ (linearity and no collinearity)} \\
\hat{\boldsymbol{\beta}} &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) \\
\hat{\boldsymbol{\beta}} &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u} \\
\hat{\boldsymbol{\beta}} &= \mathbf{I}\boldsymbol{\beta} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u} \\
\hat{\boldsymbol{\beta}} &= \boldsymbol{\beta} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u} \\
E[\hat{\boldsymbol{\beta}}|\mathbf{X}] &= E[\boldsymbol{\beta}|\mathbf{X}] + E[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] \\
E[\hat{\boldsymbol{\beta}}|\mathbf{X}] &= \boldsymbol{\beta} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'E[\mathbf{u}|\mathbf{X}] \\
E[\hat{\boldsymbol{\beta}}|\mathbf{X}] &= \boldsymbol{\beta} \text{ (zero conditional mean)} \\
E[E[\hat{\boldsymbol{\beta}}|\mathbf{X}]] &= \boldsymbol{\beta} \text{ (law of iterated expectations)}
\end{aligned}$$

2.1 Classical Standard Errors

2.2 Assumption 5: Homoskedasticity

- The stated homoskedasticity assumption is: $\text{var}(\mathbf{u}|\mathbf{X}) = \sigma_u^2\mathbf{I}_n$
- To really understand this we need to know what $\text{var}(\mathbf{u}|\mathbf{X})$ is in full generality.
- The variance of a vector is actually a matrix:

$$\text{var}[\mathbf{u}] = \Sigma_u = \begin{bmatrix} \text{var}(u_1) & \text{cov}(u_1, u_2) & \dots & \text{cov}(u_1, u_n) \\ \text{cov}(u_2, u_1) & \text{var}(u_2) & \dots & \text{cov}(u_2, u_n) \\ \vdots & & \ddots & \\ \text{cov}(u_n, u_1) & \text{cov}(u_n, u_2) & \dots & \text{var}(u_n) \end{bmatrix}$$

- This matrix is always symmetric since $\text{cov}(u_i, u_j) = \text{cov}(u_j, u_i)$ by definition.
- What does $\text{var}(\mathbf{u}|\mathbf{X}) = \sigma_u^2\mathbf{I}_n$ mean?
- $\mathbf{I}_n$ is the $n \times n$ identity matrix, σ_u^2 is a scalar.

- Visually:

$$\text{var}[\mathbf{u}] = \sigma_u^2 \mathbf{I}_n = \begin{bmatrix} \sigma_u^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma_u^2 & 0 & \dots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \dots & \sigma_u^2 \end{bmatrix}$$

- In less matrix notation:
  - $\text{var}(u_i) = \sigma_u^2$ for all i (constant variance)
  - $\text{cov}(u_i, u_j) = 0$ for all $i \neq j$ (implied by iid)

2.3 Variance of Linear Function of Random Vector

Recall that for a linear transformation of a random variable X we have $V[aX + b] = a^2 V[X]$ with constants a and b.

We will need an analogous rule for linear functions of random vectors.

Variance of Linear Transformation of Random Vector Let $f(\mathbf{u}) = \mathbf{Au} + \mathbf{B}$ be a linear transformation of a random vector $\mathbf{u}$ with non-random vectors or matrices $\mathbf{A}$ and $\mathbf{B}$. Then the variance of the transformation is given by:

$$V[f(\mathbf{u})] = V[\mathbf{Au} + \mathbf{B}] = \mathbf{A}V[\mathbf{u}]\mathbf{A}'$$

2.4 Variance of $\hat{\boldsymbol{\beta}}$

$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$ and $E[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \boldsymbol{\beta} + E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] = \boldsymbol{\beta}$ so the OLS estimator is a linear function of the errors. Thus:

$$\begin{aligned}
V[\hat{\boldsymbol{\beta}}|\mathbf{X}] &= V[\boldsymbol{\beta}|\mathbf{X}] + V[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] \\
&= V[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V[\mathbf{u}|\mathbf{X}]((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')' \quad (\mathbf{X} \text{ is nonrandom given } \mathbf{X}) \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V[\mathbf{u}|\mathbf{X}]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \quad (\text{by homoskedasticity}) \\
&= \sigma^2\mathbf{I}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}
\end{aligned}$$

This gives the $(k+1) \times (k+1)$ variance-covariance matrix of $\hat{\boldsymbol{\beta}}$.

To estimate $V[\hat{\boldsymbol{\beta}}|\mathbf{X}]$, we replace σ^2 with its unbiased estimator $\hat{\sigma}^2$, which is now written using matrix notation as:

$$\hat{\sigma}^2 = \frac{\sum_i \hat{u}_i^2}{n-(k+1)} = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n-(k+1)}$$

2.5 Back to Slide 22

3 Agnostic Inference

3.1 Consistency and Robust SEs

We know the population value of β is:

$$\beta = E\left[\mathbf{X}'\mathbf{X}\right]^{-1} E\left[\mathbf{X}'\mathbf{y}\right]$$

Plug-in principle $\rightsquigarrow$ replace population expectation with sample versions:

$$\hat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{y}$$

- With this representation, we can write the OLS estimator as follows:

$$\hat{\beta} = \beta + \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{u}$$

- Core idea: $\mathbf{X}'\mathbf{u}$ is the sum of random variables so the CLT applies.
- That, plus some asymptotic theory allows us to say:

$$\sqrt{N}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \Omega)$$

- The covariance matrix, Ω is given as:

$$\Omega = E[\mathbf{X}'\mathbf{X}]^{-1} E[\mathbf{X}'\text{Diag}(\mathbf{u^2})\mathbf{X}] E[\mathbf{X}'\mathbf{X}]^{-1}$$

- We will again be able to replace $\mathbf{u}$ with its empirical counterpart (the residuals) $\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\hat{\beta}$, and $\mathbf{X}$ with its sample counterpart.
- Note that we have only used IID sampling and no perfect collinearity so far.

Asymptotic Normality of OLS Under Random Sampling Under the Best Linear Projection model,

$$\sqrt{N}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \mathbf{\Omega})$$

where

$$\mathbf{\Omega} = E[\mathbf{X}'\mathbf{X}]^{-1} E[\mathbf{X}'\text{Diag}(\mathbf{u^2})\mathbf{X}] E[\mathbf{X}'\mathbf{X}]^{-1}$$

and

- $\hat{\beta}$ is approximately normal asymptotically with mean β and variance $\mathbf{\Omega}/n$.
- $\mathbf{\Omega}/n$ is the asymptotic covariance matrix of $\hat{\beta}$ and the square root of the diagonal are the standard errors.
- Asymptotically this provides us with all the necessary infrastructure for confidence intervals and tests.

3.1.1 Comparing to the Classical Approach (Homoskedasticity)

- Remember what we did before:

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

- Let $\text{Var}[\mathbf{u}|\mathbf{X}] = \mathbf{\Sigma}$
- Recall before we used Assumptions 1-4 to show:

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

- With homoskedasticity, $\mathbf{\Sigma} = \sigma^2\mathbf{I}$, we simplified

$$\begin{aligned}
\text{Var}[\hat{\boldsymbol{\beta}}|\mathbf{X}] &= \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{\Sigma}\mathbf{X} \left(\mathbf{X}'\mathbf{X}\right)^{-1} \\
&= \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\sigma^2\mathbf{I}\mathbf{X} \left(\mathbf{X}'\mathbf{X}\right)^{-1} \text{ (by homoskedasticity)} \\
&= \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{X} \left(\mathbf{X}'\mathbf{X}\right)^{-1} \\
&= \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}
\end{aligned}$$

- Replace σ^2 with estimate $\widehat{\sigma}^2$ will give us our estimate of the covariance matrix

When $\mathbf{\Sigma} \neq \sigma^2\mathbf{I}$, we are stuck with this expression:

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

Idea: If we can consistently estimate the components of $\mathbf{\Sigma}$, we could directly use this expression by replacing $\mathbf{\Sigma}$ with its estimate, $\hat{\mathbf{\Sigma}}$.

3.1.2 White's Heteroskedasticity Consistent Estimator

Suppose we have heteroskedasticity of unknown form (but zero covariance):

$$V[\mathbf{u}] = \mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & 0 & \dots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \dots & \sigma_n^2 \end{bmatrix}$$

then $V[\hat{\boldsymbol{\beta}}|\mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Sigma}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$ and White (1980) shows that

$$\widehat{V[\hat{\boldsymbol{\beta}}|\mathbf{X}]} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \begin{bmatrix} \hat{u}_1^2 & 0 & 0 & \dots & 0 \\ 0 & \hat{u}_2^2 & 0 & \dots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \dots & \hat{u}_n^2 \end{bmatrix} \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

is a consistent estimator of $V[\hat{\boldsymbol{\beta}}|\mathbf{X}]$ under any form of heteroskedasticity consistent with $V[\mathbf{u}]$ above.

The estimate based on the above is called the heteroskedasticity consistent (HC) or robust standard errors. This also coincides with the agnostic standard errors!

Core intuition: while $\widehat{\mathbf{\Sigma}}$ is an $n \times n$ matrix, $\mathbf{X}'\widehat{\mathbf{\Sigma}}\mathbf{X}$ is a $(k+1) \times (k+1)$ matrix. So there is hope of estimating it consistently as sample size grows *even when every true error variance is unique.*

$$\widehat{\mathbf{\Sigma}} = \begin{bmatrix} \hat{u}_1^2 & 0 & 0 & \dots & 0 \\ 0 & \hat{u}_2^2 & 0 & \dots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \dots & \hat{u}_n^2 \end{bmatrix}$$

$$\mathbf{X}'\mathbf{\Sigma}\mathbf{X} = \begin{bmatrix} \sum_i x_{i,1}x_{i,1}\hat{u}_i^2 & \sum_i x_{i,1}x_{i,2}\hat{u}_i^2 & \dots & \sum_i x_{i,1}x_{i,k+1}\hat{u}_i^2 \\ \sum_i x_{i,2}x_{i,1}\hat{u}_i^2 & \sum_i x_{i,2}x_{i,2}\hat{u}_i^2 & \dots & \sum_i x_{i,2}x_{i,k+1}\hat{u}_i^2 \\ & & \vdots & \\ \sum_i x_{i,k+1}x_{i,1}\hat{u}_i^2 & \sum_i x_{i,k+1}x_{i,2}\hat{u}_i^2 & \dots & \sum_i x_{i,k+1}x_{i,k+1}\hat{u}_i^2 \end{bmatrix}$$

3.2 White's Heteroskedasticity Consistent Estimator

Robust standard errors are easily computed with the "sandwich" formula:

1. Fit the regression and obtain the residuals $\hat{\mathbf{u}}$
2. Construct the "meat" matrix $\widehat{\mathbf{\Sigma}}$ with squared residuals in diagonal:

$$\widehat{\mathbf{\Sigma}} = \begin{bmatrix} \hat{u}_1^2 & 0 & 0 & \dots & 0 \\ 0 & \hat{u}_2^2 & 0 & \dots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \dots & \hat{u}_n^2 \end{bmatrix}$$

3. Plug $\widehat{\mathbf{\Sigma}}$ into the sandwich formula to obtain the robust estimator of the variance-covariance matrix

$$V[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\widehat{\mathbf{\Sigma}}\mathbf{X} \left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

- There are various small sample corrections to improve performance when sample size is small. The most common variant (sometimes labeled HC1) is:

$$V[\hat{\boldsymbol{\beta}}|\mathbf{X}] \;=\; \frac{n}{n-k-1} \cdot \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\widehat{\mathbf{\Sigma}}\mathbf{X} \left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

4 Clustering

Clustered Dependence: Intuition

- Think back to the Gerber, Green, and Larimer (2008) social pressure mailer example.
- Their design: randomly sample households and randomly assign them to different treatment conditions.
- But the measurement of turnout is at the individual level and we have one unit per person.
- Violation of iid/random sampling called clustering or clustered dependence.
- Loosely speaking, you are tricking the regression into thinking it has more information than it does.
- Recall our general sandwich expression:

$$V[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{\Sigma}\mathbf{X} \left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

- We need to adjust to address our new sampling strategy.
- We assume that respondents are independent across clusters, but dependent within clusters. Thus, we have

$$\text{Var}[\boldsymbol{u}_g|\mathbf{X}_g] = \boldsymbol{\Sigma}_g$$

- This leads to a block diagonal expression for $\boldsymbol{\Sigma}$ overall

$$V[\boldsymbol{u}] = \boldsymbol{\Sigma} = \left[\begin{array}{c|c|c|c} \boldsymbol{\Sigma}_1 & 0 & \dots & 0 \\ \hline \mathbf{0} & \boldsymbol{\Sigma}_2 & \dots & \mathbf{0} \\ \hline & & \ddots & \\ \hline \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{\Sigma}_M \end{array}\right]$$

- Under this clustered dependence, we can write this as:

$$V[\hat{\boldsymbol{\beta}}|\mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{g=1}^{m}\mathbf{X}_g'\boldsymbol{\Sigma}_g\mathbf{X}_g\right)(\mathbf{X}'\mathbf{X})^{-1}$$

- The independence is now across clusters, so we will need asymptotics in the number of clusters for CLT.
- We can use a plugin estimator for the sampling variance on the previous slide with a small sample correction (many variants available, here we give the CR1 variant Stata uses):

$$\hat{V}_{CR1}[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \frac{m}{(m-1)}\frac{n-1}{n-k}(\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{g=1}^{m}\mathbf{X}_g'\hat{\boldsymbol{u}}_g\hat{\boldsymbol{u}}_g'\mathbf{X}_g\right)(\mathbf{X}'\mathbf{X})^{-1}$$

- Another excellent strategy is to use the block bootstrap (resample clusters instead of units).
- CRSE do not change our estimates $\widehat{\boldsymbol{\beta}}$, cannot fix bias.
- Consistency of the CRSE are in the number of groups, not the number of individuals.
- Works well in the following settings:
  - sampling complete blocks of units where sampled clusters are a small fraction of the population.

  - sampling a small fraction of units from all or nearly all clusters in the population.
  - assignment of key variable is constant within the cluster.

- Just because something is a group, doesn't mean you need to cluster standard errors, key is to think about the sampling strategy.

- Abadie, Athey, Imbens and Wooldridge (2022) "When Should You Adjust Standard Errors for Clustering" *QJE* provides guidance and approaches for other cases.

- NB: this broad strategy of sandwich estimators works for other kinds of settings (time-series etc.). See `sandwich` package in `R`.