Lecture 4: Standard Errors

September 24, 2025

Today's Plan

- Recap OLS and various forms of standard errors
- Standard errors are tedious but I guess you are supposed to know this stuff
- Hopefully first and last time we talk about this

Recap: Asymptotics for OLS and the Linear Model

OLS

$$y_i = \beta_0 + \beta x_i + u_i$$

Recall the three basic OLS assumptions

1. $\mathbb{E}(u_i|X_i) = 0$
2. (X_i, Y_i), $i = 1, \ldots, n$, are i.i.d.
3. Large outliers are rare $\mathbb{E}[Y^4] < \infty$ and $\mathbb{E}[X^4] < \infty$.

Unbiasedness and Consistency

- Unbiasedness means on average we don't over or under estimate $\widehat{\beta}$

$$\mathbb{E}[\widehat{\beta}] - \beta_0 = 0$$

- Consistency tells us that we approach the true β_0 as $n \longrightarrow \infty$.

$$\widehat{\beta} \xrightarrow{p} \beta_0$$

- Example: $X_{(1)}$ is unbiased but not consistent for the mean.
- Example $\frac{n}{n-5}\overline{X}$ is consistent but biased for the mean.

Bias Variance Decomposition

We can decompose any estimator into two components

$$\underbrace{\mathbb{E}[(y-\hat{f}(x))^2]}_{MSE} = \underbrace{\left(\mathbb{E}[\hat{f}(x)-f(x)]\right)^2}_{Bias^2} + \underbrace{\mathbb{E}\left[\left(\hat{f}(x)-\mathbb{E}[\hat{f}(x)]\right)^2\right]}_{Variance}$$

- What minimizes MSE?

$$f(x_i) = \mathbb{E}[Y_i \mid X_i]$$

- In general we face a tradeoff between bias and variance.
- In OLS we minimize the variance among unbiased estimators assuming that the true $f(x_i) = X_i\beta$ is linear. (But is it?)

Outliers and Leverage

One way to find outliers is to calculate the leverage of each observation i. We begin with the hat matrix:

$$P = X(X'X)^{-1}X'$$

and consider the diagonal elements which for some reason are labeled h_{ii}

$$h_{ii} = x_i(X'X)^{-1}x_i'$$

This tells us how influential an observation is in our estimate of $\widehat{\beta}$.
Particularly important for $\{0, 1\}$ dummy variables with uneven groups.

Leave One Out Regression

- This is sometimes called the Jackknife
- Sometimes it is helpful to know what would happen if we omitted a single observation i
- Turns out we don't need to run N regressions

$$
\begin{aligned}
\widehat{\beta}_{-i} &= (X'_{-i}X_{-i})^{-1}X'_{-i}Y_{-i} \\
&= \widehat{\beta} - (X'X)^{-1}x_i\tilde{u}_i \quad \text{where } \tilde{u}_i = (1-h_{ii})^{-1}\hat{u}_i
\end{aligned}
$$

- $\tilde{u}_i$ has the interpretation of the LOO prediction error.
- high leverage observations move $\widehat{\beta}$ a lot.

You can read more about this in Ch3 of Hansen. [Skip derivation]

Gauss Markov Theorem

Gauss Markov Adds two assumptions:

1. $E(u_i|X_i) = 0$
2. (X_i, Y_i), $i = 1, \ldots, n$, are i.i.d.
3. Large outliers are rare $\mathbb{E}[Y^4] < \infty$ and $\mathbb{E}[X^4] < \infty$.
4. $Var(u_i) = \sigma^2$ (homoskedasticity)
5. $u_i \sim N(0, \sigma^2)$ (normal errors)

Under these assumptions you learned that OLS is BLUE

Variance of $\widehat{\beta}$

Start with the variance of the residuals to form a diagonal matrix D:

$$\mathrm{Var}(\mathbf{u}|\mathbf{X}) = \mathbb{E}\left(\mathbf{u}\mathbf{u}' \mid \mathbf{X}\right) = \mathbf{D}$$

$$\mathbf{D} = \mathrm{diag}\left(\sigma_1^2, \ldots, \sigma_n^2\right) = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix}$$

- $\mathbf{D}$ is diagonal because $\mathbb{E}[u_i u_j \mid X] = \mathbb{E}[u_i \mid x_i]\mathbb{E}[u_j \mid x_j] = 0$ (independence)
- The elements of D_i are given by $\mathbb{E}[u_i^2 \mid X] = \mathbb{E}[u_i^2 \mid x_i] = \sigma_i^2$.
- In the homoskedastic case $\mathbf{D} = \sigma^2 \mathbf{I}_n$.

Variance of $\widehat{\beta}$

A useful identity for linear algebra:

$$\mathrm{Var}(a\mathbf{Z}) = a^2\,\mathrm{Var}(\mathbf{Z})$$

$$\mathrm{Var}(A\mathbf{Z}) = A\,\mathrm{Var}(\mathbf{Z})A'$$

Recall that $\mathrm{Var}(\mathbf{Y}|\mathbf{X}) = \mathrm{Var}(\mathbf{u}|\mathbf{X})$ and also recall the formula for $\widehat{\beta}$:

$$\widehat{\beta} = \underbrace{(X'X)^{-1}X'}_{A}\,Y = A'Y$$

$$\begin{aligned}\mathbf{V}_{\widehat{\beta}} = \mathrm{Var}(\widehat{\beta}|X) &= (X'X)^{-1}X'\,\mathrm{Var}(Y|X)X(X'X)^{-1} \\ &= (X'X)^{-1}(X'\mathbf{D}X)(X'X)^{-1}\end{aligned}$$

We have that $(X'\mathbf{D}X) = \sum_{i=1}^{N} x_i x_i' \sigma_i^2$. Under homoskedasticity $\mathbf{D} = \sigma^2\mathbf{I}_n$ and $\mathbf{V}_{\widehat{\beta}} = \sigma^2(X'X)^{-1}$.

$$\mathsf{D} = \operatorname{diag}\left(\sigma_1^2, \ldots, \sigma_n^2\right) = \mathbb{E}\left(u_i u_i' \mid \mathsf{X}\right) = \mathbb{E}\left(\widetilde{\mathsf{D}} \mid \mathsf{X}\right)$$

We can estimate $\widehat{\mathsf{V}}_{\widehat{\beta}}$ by plugging in $\mathsf{D} \longrightarrow \widetilde{\mathsf{D}}$:

$$\begin{aligned} \mathsf{V}_{\widehat{\beta}} &= (X'X)^{-1}(X'\widetilde{\mathsf{D}}X)(X'X)^{-1} \\ &= (X'X)^{-1}\left(\sum_{i=1}^{N} x_i x_i' u_i^2\right)(X'X)^{-1} \end{aligned}$$

The expectation shows us this estimator is unbiased:

$$\mathbb{E}[\mathsf{V}_{\widehat{\beta}} \mid X] = (X'X)^{-1}\left(\sum_{i=1}^{N} x_i x_i' \, \mathbb{E}[u_i^2|X]\right)(X'X)^{-1}$$

Heteroskedasticity Consistent (HC) Variance Estimates

What we need is a consistent estimator for $\hat{u}_i^2$.

$$V_{\widehat{\beta}}^{HC0} = (X'X)^{-1}\left(\sum_{i=1}^{N} x_i x_i' \hat{u}_i^2\right)(X'X)^{-1}$$

$$V_{\widehat{\beta}}^{HC1} = (X'X)^{-1}\left(\sum_{i=1}^{N} x_i x_i' \hat{u}_i^2\right)(X'X)^{-1} \cdot \left(\frac{n}{n-k}\right)$$

Could use leave one out variance estimate:

$$V_{\widehat{\beta}}^{HC2} = (X'X)^{-1}\left(\sum_{i=1}^{N} (1-h_{ii})^{-1} x_i x_i' \hat{u}_i^2\right)(X'X)^{-1}$$

$$V_{\widehat{\beta}}^{HC3} = (X'X)^{-1}\left(\sum_{i=1}^{N} (1-h_{ii})^{-2} x_i x_i' \hat{u}_i^2\right)(X'X)^{-1}$$

- We know that $\mathbb{V}_{\hat{\beta}}^{HC3} > \mathbb{V}_{\hat{\beta}}^{HC2} > \mathbb{V}_{\hat{\beta}}^{HC0}$ because $(1 - h_{ii}) < 1$.
- $HC3$ are the most conservative and also place the most weight on potential outliers.
- `Stata` uses $HC1$ as the default and it is what most people refer to when they say robust standard errors.
- These are often called White (1980) SE's or Eicher-Huber-White SE's.
- In small sample some evidence that $HC2$ has better coverage, (what is that?)

What is Clustering?

Suppose we want to relax our i.i.d. assumption:

- Each observation i is a villager and each group g is a village
- Each observation i is a student and each group g is a class.
- Each observation t is a year and each entity i is a state.
- Each observation t is a week and each entity i is a shopper.

We might expect that $\text{Cov}(u_{g1}, u_{g2}, \ldots, u_{gN}) \neq 0 \longrightarrow$ independence is a bad assumption.

Clustering: Intuition

The groups (villages, classrooms, states) are independent of one another, but within each group we can allow for arbitrary correlation.

- If correlation is within an individual over time we call it serial correlation or autocorrelation
- Just like in time-series$\longrightarrow$ we have fewer effective independent observations in our sample.
- Asymptotics now about the number of groups $G \longrightarrow \infty$ not observations $N \longrightarrow \infty$

Begin by stacking up observations in each group $\mathbf{y}_g = [y_{g1}, \ldots, y_{gn_g}]$, we can write OLS three ways:

$$y_{ig} = x_{ig}'\beta + u_{ig}$$
$$\mathbf{y}_g = \mathbf{X}_g\beta + \mathbf{u}_g$$
$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{u}$$

All of these are equivalent:

$$\widehat{\beta} = \left(\sum_{g=1}^{G}\sum_{i=1}^{n_g} x_{ig}'x_{ig}\right)^{-1}\left(\sum_{g=1}^{G}\sum_{i=1}^{n_g} x_{ig}'y_{ig}\right)$$
$$\widehat{\beta} = \left(\sum_{g=1}^{G}\mathbf{X}_g'\mathbf{X}_g\right)^{-1}\left(\sum_{g=1}^{G}\mathbf{X}_g'\mathbf{y}_g\right)$$

Clustering (Continued)

The error terms have covariance within each cluster g as:

$$\boldsymbol{\Sigma}_g = \mathbb{E}\left(\mathbf{u}_g \mathbf{u}_g' \mid X_g\right)$$

In order to calculate $\widehat{V_{\hat{\beta}}}$ we replace the covariance matrix **D** with Ω and consider an estimator $\widehat{\Omega}_n$. We exploit independence across clusters:

$$\operatorname{var}\left(\left(\sum_{g=1}^{G} X_g' \mathbf{u}_g\right) \mid X\right) = \sum_{g=1}^{G} \operatorname{var}\left(X_g' u_g | X_g\right) = \sum_{g=1}^{G} X_g' \mathbb{E}\left(u_g u_g' | X_g\right) X_g = \sum_{g=1}^{G} X_g' \boldsymbol{\Sigma}_g \mathbf{X}_g \equiv \Omega_N$$

And an estimate of the variance:

$$V_{\hat{\beta}} = \operatorname{var}(\widehat{\beta} \mid X) = (\mathbf{X}'\mathbf{X})^{-1} \boldsymbol{\Omega}_n (\mathbf{X}'\mathbf{X})^{-1}$$

Clustered SE's

$$
\begin{aligned}
\widehat{\Omega}_n &= \sum_{g=1}^{G} X_g' \widehat{\mathbf{u}}_g \widehat{\mathbf{u}}_g' X_g \\
&= \sum_{g=1}^{G} \sum_{i=1}^{n_g} \sum_{\ell=1}^{n_g} x_{ig} x_{\ell g}' \widehat{u}_{ig} \widehat{u}_{\ell g} \\
&= \sum_{g=1}^{G} \left(\sum_{i=1}^{n_g} x_{ig} \widehat{u}_{ig} \right) \left(\sum_{\ell=1}^{n_g} x_{\ell g} \widehat{u}_{\ell g} \right)'
\end{aligned}
$$

- First line makes explicit: independence over each of G clusters
- Last line easiest for computer

$$\widehat{V}_{\hat{\beta}}^{\mathrm{CR1}} = (X'X)^{-1} \left(\sum_{g=1}^{G} X_g' \widehat{u}_g \widehat{u}_g' X_g \right) (X'X)^{-1}$$

$$\widehat{V}_{\hat{\beta}}^{\mathrm{CR3}} = (X'X)^{-1} \left(\sum_{g=1}^{G} X_g' \widetilde{u}_g \widetilde{u}_g' X_g \right) (X'X)^{-1}$$

- Can replace $\hat{\mathrm{u}}_g \rightarrow \tilde{\mathrm{u}}_g$ for leave-one out like *HC*3 (these are called *CR*3).

Most Asked PhD Student Econometric Question

How should I cluster my standard errors?

- Heck if I know.
- This is very problem specific
- It matters a lot $\longrightarrow$ standard errors can get orders of magnitude larger.
- Do you believe across group independence or not? [this is the only thing that matters]
- If you include fixed effects probably you need at least clustering at that level.

Newey West Standard Errors (HAC)

- In serially correlated data we need to account for $\text{Cov}(u_t, u_{t-1}, \ldots) \neq 0$.
- Clustering is one solution, but we may end up throwing away all of our data.
- Instead we could estimate the serial correlation.
- May also want standard errors that are heteroskedasticity AND autocorrelation consistent (HAC).
- Have to select a number of lags L

$$
\begin{aligned}
\widehat{\Omega}_{n,L}^{HAC} &= \sum_{t=1}^{T} u_t^2 x_t x_t' + \sum_{l=1}^{L} \sum_{t=l+1}^{T} w_l u_t u_{t-l} \left(x_t x_{t-l}' + x_{t-l} x_t' \right) \\
w_l &= 1 - \frac{l}{L+1}
\end{aligned}
$$

What about β?

- All of the estimates above should produce identical point estimates
- We have just been talking about adjusting standard errors
- Should the presence of heteroskedasticity change our estimates of $\widehat{\beta}$ as well?

A simple extension is Weighted Least Squares (WLS)

- Different motivations
- Suppose we have sampling weights that are not $\frac{1}{n}$ from survey data, etc:
  - If my population is supposed to represent all US residents and my sample is 75% Women...
  - Relax LSA (2) (X_i, Y_i), $i = 1, \ldots, n$, are i.i.d.
- In this case, OLS is still unbiased and consistent, just inefficient

WLS

Can weight each observation as w_i so that $\sum_{i=1}^{N} w_i = 1$ instead of $w_i = \frac{1}{N}$.

Can define a diagonal matrix W with entries w_i.

$$\arg\min_{\beta} \sum_{i=1}^{N} w_i (y_i - X_i\beta)^2 = \arg\min_{\beta} \left\| W^{1/2} \left| Y - X\beta \right| \right\|$$

Can also consider a transformation of the data

$$\tilde{y}_i = \sqrt{w_i} y_i, \quad \tilde{x}_i = \sqrt{w_i} x_i$$
$$\tilde{Y} = W^{1/2} Y, \quad \tilde{X} = W^{1/2} X$$

A regression of $\tilde{Y}$ on $\tilde{X}$:

$$\widehat{\beta}_{WLS} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{Y} = (X'WX)^{-1}X'WY$$

WLS

Also used as a solution to heteroskedasticity

- Relax LSA (2) (X_i, Y_i), $i = 1, \ldots, n$, are i.i.d.
- Relax LSA (4) $Var(u_i) = \sigma^2$ (homoskedasticity)

Why? We are minimizing weighted sum of squared residuals:

$$\sum_{i=1}^{N} w_i(y_i - \hat{y}_i)^2 = \sum_{i=1}^{N} w_i u_i^2$$

Suppose we have heteroskedasticity so that $Var(\varepsilon_i) = \sigma_i^2$ and $w_i \propto \frac{1}{\sigma_i^2}$.

In this setting WLS is BLUE.

WLS

Why does anyone ever run OLS instead of WLS?

- Problem is that σ_i^2 is unknown before we run our regression.
- We can estimate $\widehat{\sigma}_i^2$.

This procedure is known as Iteratively Re-weighted Least Squares IRLS

1. Intialize weights to identity matrix: $W = I$
2. Regress Y on X with weights W
3. Obtain $\widehat{u}_i$.
4. Update W with $w_{ii} = \frac{1}{\widehat{u}_i^2}$
5. Repeat until parameter estimates don't change

GLS and FGLS

There is no reason to require that W be diagonal. This gives us Generalized Least Squares

$$\widehat{\beta}_{GLS} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{Y} = (X'\Omega X)^{-1}\Omega' WY$$

The idea is to use the inverse covariance matrix of residuals. But this is high dimensional ($N \times N$) and estimating it is harder than our original problem!

Feasible Generalized Least Squares FGLS:

1. Intialize weights to identity matrix: $\widehat{\Omega} = I$
2. Regress Y on X with weighting matrix $\widehat{\Omega}$
3. Obtain $\widehat{u}_i$.
4. Construct $\mathbb{E}[u_i^2 \mid X, Z]$ via (nonlinear) regression: $\exp[\gamma_0 + \gamma_1 x_i + \gamma_2 z_i]$.
5. Update $\widehat{\Omega}$ with $\mathbb{E}[u_i^2 \mid X, Z]$
6. Repeat until parameter estimates don't change

Thanks!