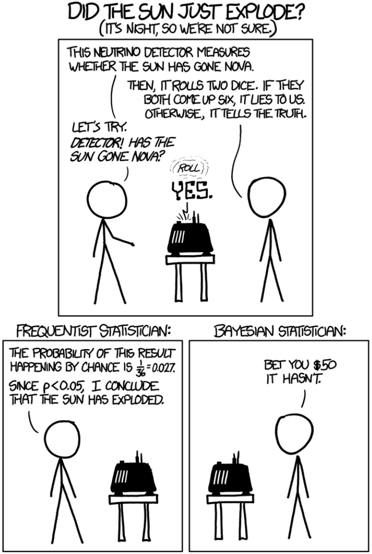
Econometrics I

Lecture 4: Inference and Standard Errors

Chris Conlon

Fall 2025



Source: xkcd.com

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.

Outliers

- **Outliers** refer to observations that are "far away" from the rest of the data. They can be due to errors in the data. There is no standard formal definition.
- What to do? Greene:"*It is difficult to draw firm general conclusions... It remains likely that in very small samples, some caution and close scrutiny of the data are called for.*" I'd say that's true even in large samples, but there isn't a generally accepted way of quantifying what counts as appropriate "caution and close scrutiny."

Removing Outliers?

- Removing extreme outliers (in x) from datasets is often considered good practice. But we should be mindful about why as dropping observations creates the potential for manipulation.
- Sometimes extreme outliers are just errors, in which case they should almost certainly be dropped.
- Even if they aren't errors, they may reflect a different mode in the data generating process. They may require a different or more general model to account for them properly. Consider the justificaiton of a linear model based on Taylor's theorem (local linear approximation). With such a justification for your modeling strategy, it would not make sense to include an outlier in x.
- It's important to be transparent about how dropped outliers affect results.

Outliers and Leverage

- One way to find influential observations is to calculate the **leverage** of each observation i. We begin with the hat matrix:

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

  and consider the diagonal elements, which are labeled h_{ii}

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

- This tells us how influential an observation is in our estimate of $\mathbf{b}_{OLS}$. 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

$$
\begin{aligned}
\mathbf{b}_{-i} &= (\mathbf{X}'_{-i}\mathbf{X}_{-i})^{-1}\mathbf{X}'_{-i}\mathbf{y}_{-i} \\
&= \mathbf{b}_{OLS} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\tilde{e}_i \quad \text{where } \tilde{e}_i = (1-h_{ii})^{-1}e_i
\end{aligned}
$$

- $\tilde{e}_i$ has the interpretation of the LOO prediction error.
- high leverage observations move $\mathbf{b}_{OLS}$ a lot.

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?)

Variance of $\widehat{\beta}_{OLS}$

- A useful identity for linear algebra:

$$\text{Var}(a\mathbf{Z}) = a^2 \,\text{Var}(\mathbf{Z})$$
$$\text{Var}(A\mathbf{Z}) = A\,\text{Var}(\mathbf{Z})A'$$

- Since $\mathbf{b}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$,

$$\text{Var}(\mathbf{b}_{OLS}|\mathbf{X}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\text{Var}(\mathbf{y}|\mathbf{X})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

- Recalling that $\text{Var}(\mathbf{y}|\mathbf{X}) = \text{Var}(\varepsilon|\mathbf{X})$ (because $\text{Var}(\mathbf{X}|\mathbf{X}) = 0$)

$$\text{Var}(\mathbf{b}_{OLS}|\mathbf{X}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\text{Var}(\varepsilon|\mathbf{X})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

Variance of $\widehat{\beta}_{OLS}$

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

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

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

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

Variance of $\widehat{\beta}$

$$\mathbf{D} = \text{diag}\left(\sigma_1^2, \ldots, \sigma_n^2\right) = \mathbb{E}\left(\varepsilon_i \varepsilon_i' \mid \mathbf{X}\right) = \mathbb{E}\left(\widetilde{\mathbf{D}} \mid \mathbf{X}\right)$$

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

$$\mathbf{V}_{\widehat{\beta}} = (X'X)^{-1}(X'\widetilde{\mathbf{D}}X)(X'X)^{-1} = (X'X)^{-1}\left(\sum_{i=1}^{N} x_i x_i' \varepsilon_i^2\right)(X'X)^{-1}$$

The expectation shows us this estimator is unbiased:

$$\begin{aligned}
\mathbb{E}[\mathbf{V}_{\widehat{\beta}} \mid X] &= (X'X)^{-1}\left(\sum_{i=1}^{N} x_i x_i' \,\mathbb{E}[\varepsilon_i^2 | X]\right)(X'X)^{-1} \\
&= (X'X)^{-1}\left(\sum_{i=1}^{N} x_i x_i' \,\sigma_i^2\right)(X'X)^{-1} = (X'X)^{-1}(X'\mathbf{D}X)(X'X)^{-1}
\end{aligned}$$

Heteroskedasticity Consistent (HC) Variance Estimates

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

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

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

Could use leave one out variance estimate:

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

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

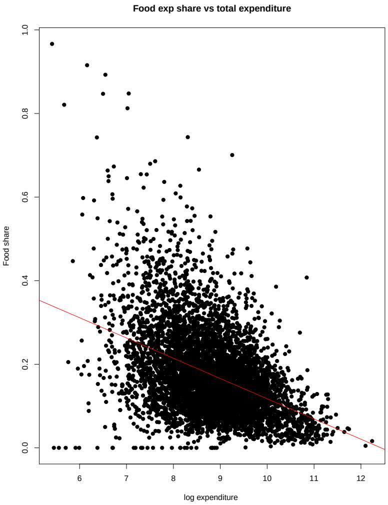
Heteroskedasticity Consistent (HC) Variance Estimates

- We know that $\mathbf{V}_{\widehat{\beta}}^{HC3} > \mathbf{V}_{\widehat{\beta}}^{HC2} > \mathbf{V}_{\widehat{\beta}}^{HC0}$ 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?)

Example: Engel Curves

- Engel curves refer to the relationship between a household's expenditure share on a good and income (or total expenditure).
- Engel curves for food are typically downward sloping – as total expenditure of a household increases, the proportion of its expenditure dedicated to food falls.
    - Expenditure on food still rises as total expenditure rises, but less than proportionally, so that food's expenditure *share* falls.

Food Engel Curves



- If we plotted total food expenditure (rather than the expenditure share), the heteroscedasticity would go in the other direction.

Data source: BLS Consumer Expenditure Survey data

Heteroscedasticity vs. Correlation

- Recall that we defined the homoscedasticity assumption as:

$$Var(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}$$

  this assumption has two aspects:

  1. The disturbance for each observation has the same variance
  2. Imposing zero correlation between disturbances for different observations

- The terminology can be misleading here, because what people refer to as "heteroscedasticity-robust" standard errors (the variance estimators on the previous slide) are robust to violations of 1 but not 2.
- We need to do a bit more to estimate standard errors in a way that is robust to correlated data.

Correlation I

- The baseline assumptions of the linear regression framework imply that the disturbances are uncorrelated across observations. There are many ways for this to be violated.
    - Example 1: we might have county-level data for a regression and be concerned that different counties within a given state have correlated disturbances because all counties are subject to the same (unobserved) state-level policies.
    - Example 2: time series data (asset prices), and we are worried that some unobserved factors within the disturbances are serially correlated
    - Example 3: county level data again, and we are worried about geographically correlated factors such as weather.

Correlation II

Different correlation patterns call for different estimators of $\boldsymbol{\Sigma}$, the variance of $\mathbf{b}_{OLS}$ Some common alternatives to the no-correlation baseline:

1. Clustered standard errors, when there is correlation between observations within well-defined groups, but no correlation between observations in different groups.
2. Newey-West standard errors (and extensions) to deal with serial correlation in time series data.
3. Conley-Newey-West standard errors that allow for correlation in multiple dimensions (especially popular in the context of spatially explicit models).

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 $N \longrightarrow \infty$

Clustering I

- Suppose data are organized into distinct groups $g = 1, 2, \ldots, G$. Let $g(i)$ be the group identity of observation i.
  - e.g., with county-level data, we have $g(\mathit{Manhattan}) = \mathit{NY}$.
- We assume $\mathbb{E}\left[\varepsilon_i \varepsilon_j\right] = 0$ as long as $g(i) \neq g(j)$, and we do not restrict the correlation $\mathbb{E}\left[\varepsilon_i \varepsilon_j\right]$ for observations within the same group.
- Intuition: the linear regression framework with no correlation in observations will overstate the precision of our estimates. If we add another observation within a cluster, and that observation is highly correlated with the other observations, it's not actually as good as adding another independent observation.

Clustering II

- Recall the sandwich formula for standard errors:

$$n^{-1}\mathbb{E}\left[\mathbf{x}_i\mathbf{x}_i'\right]^{-1}\mathbb{V}\left[\mathbf{x}_i\varepsilon_i\right]\mathbb{E}\left[\mathbf{x}_i\mathbf{x}_i'\right]^{-1}.$$

- The estimator for the middle part without clustering was

$$\mathbb{V}\left[\mathbf{x}_i\varepsilon_i\right] = n^{-1}\sum_i \mathbf{x}_i\mathbf{x}_i' e_i^2$$

- With clustering, it will be

$$\mathbf{V}_{clu} = n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{x}_i\,\mathbf{x}_j'\,e_i\,e_j\cdot\mathbb{I}\left[(i,j)\in \text{Group}_g\right]$$

where the **I** function is 1 when i, j come from the same group and zero otherwise.

Clustering III

- The cluster-robust estimate of standard errors will be consistent as the number of groups gets large.
- Note that this estimator adds extra terms (covariance terms) to the estimate of variance, so this is going to make standard errors larger as long as covariances $\mathbb{E}\left[\varepsilon_i \varepsilon_j\right]$ are positive.
- Thus, if standard formulas are used in the presence of cluster-correlated disturbances, standard errors will be too small.
- Statistical software packages typically make it easy to compute cluster-robust errors.
- Clustering often makes a **huge** difference in standard errors.

Clustering Derviation

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

$$
\begin{aligned}
y_{ig} &= x_{ig}'\beta + \varepsilon_{ig} \\
\mathbf{y}_g &= \mathbf{X}_g\beta + \boldsymbol{\varepsilon}_g \\
\mathbf{Y} &= \mathbf{X}\beta + \boldsymbol{\varepsilon}
\end{aligned}
$$

All of these are equivalent:

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

Clustering Derivation (Continued)

The error terms have covariance within each cluster g as:

$$\boldsymbol{\Sigma}_g = \mathbb{E}\left(\varepsilon_g \varepsilon_g' \mid \boldsymbol{X}_g\right)$$

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

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

And an estimate of the variance:

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

Clustered SE's

$$\widehat{\boldsymbol{V}}^{\mathrm{CR1}}_{OLS} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\left(\sum_{g=1}^{G}\boldsymbol{X}'_g\boldsymbol{e}_g\boldsymbol{e}'_g\boldsymbol{X}_g\right)(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

$$\widehat{\boldsymbol{V}}^{\mathrm{CR3}}_{OLS} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\left(\sum_{g=1}^{G}\boldsymbol{X}'_g\widetilde{\boldsymbol{e}}_g\widetilde{\boldsymbol{e}}'_g\boldsymbol{X}_g\right)(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

- Can replace $\mathbf{e}_g \longrightarrow \tilde{\mathbf{e}}_g$ for leave-one out like $HC3$ (these are called $CR3$).

Clustering in R

```
feols(y~ x1 + x2, data=df, vcov=~group_id )
feols(y~ x1 + x2, data=df, vcov=~group_id+time_id)
```

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.

Bootstrap I

- Another approach to estimating the standard errors of $\mathbf{b}_{OLS}$ is the **bootstrap**
- The basic idea:
  1. Simulate a new data set (same number of observations) by sampling (with replacement) from the original data set
  2. Estimate $\mathbf{b}_{OLS,s}$ for the new data set.
  3. Repeat lots of times, resulting in a bunch of different estimates of $\mathbf{b}_{OLS,s}$, say $s = 1, \ldots, 10000$
  4. Look at the variance of the $\mathbf{b}_{OLS,s}$ estimates across the various simulated data sets. This is your estimate of $\mathbf{\Sigma}$, or $Var(\mathbf{b}_{OLS})$

Bootstrap II

- The bootstrap's main appeal is that it can provide a better finite-sample approximation of the distribution of the parameter estimates.
  - Note that the Eicker-Huber-White standard errors estimates are *consistent*, but not generally *unbiased* in finite samples
  - The bootstrap is probably worth trying if you're ever working with non-linear estimators (which can be consistent but are typically not unbiased in finite samples).

- Also, it can potentially deliver good estimates of standard errors even with correlated errors, but this depends on the version of the bootstrap (see **block bootstrap**). Exploring formally the conditions under which the bootstrap works well is beyond our scope.