

# Lecture 1: Review (Mostly)

---

Chris Conlon

January 31, 2023

NYU Stern

# Packages for Today

Let's load some packages so that I can make some better looking plots:

```
#always  
library(tidyverse)  
# for SE's  
library(estimatr)  
library(broom)  
# for regressions  
library(fixest)  
library(sandwich)
```

# 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

---

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

Recall the three basic OLS assumptions

1.  $\mathbb{E}(u_i|X_i) = 0$
2.  $(X_i, Y_i)$ ,  $i = 1, \dots, n$ , 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  $\beta_0$  as  $n \rightarrow \infty$ .

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

- 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

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

- What minimizes MSE?

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

- 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  $\hat{\beta}$ .  
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]

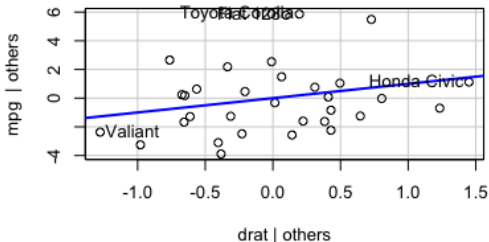
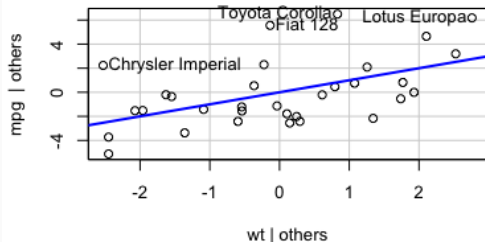
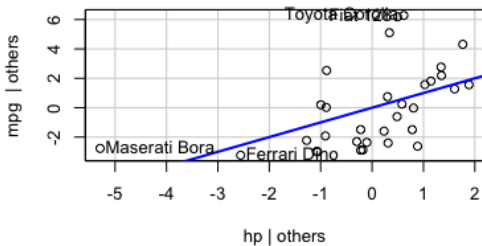
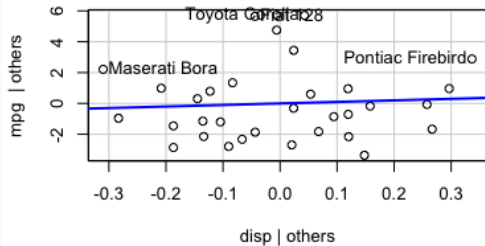
## Leverage and QQ plots

```
library(car)
fit <- lm(mpg~disp+hp+wt+drat, data=mtcars)

# Assessing Outliers
outlierTest(fit) # Bonferonni p-value for most extreme obs
qqPlot(fit, main="QQ Plot") #qq plot for studentized resid
leveragePlots(fit) # leverage plots
```

# Leverage Plot

Leverage Plots



# Gauss Markov Theorem

Gauss Markov Adds two assumptions:

1.  $E(u_i|X_i) = 0$
2.  $(X_i, Y_i), i = 1, \dots, n$ , 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 $\hat{\beta}$

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

$$\text{Var}(\mathbf{u}|\mathbf{X}) = \mathbb{E}(\mathbf{u}\mathbf{u}' | \mathbf{X}) = \mathbf{D}$$

$$\mathbf{D} = \text{diag}(\sigma_1^2, \dots, \sigma_n^2) = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n^2 \end{pmatrix}$$

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

## Variance of $\widehat{\beta}$

A useful identity for linear algebra:

$$\begin{aligned}\text{Var}(a\mathbf{Z}) &= a^2 \text{Var}(\mathbf{Z}) \\ \text{Var}(A\mathbf{Z}) &= A \text{Var}(\mathbf{Z}) A'\end{aligned}$$

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

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

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  $\mathbf{V}_{\widehat{\beta}} = \sigma^2 (X'X)^{-1}$ .

## Variance of $\widehat{\beta}$

$$\mathbf{D} = \text{diag}(\sigma_1^2, \dots, \sigma_n^2) = \mathbb{E}(u_i u_i' | \mathbf{X}) = \mathbb{E}(\widetilde{\mathbf{D}} | \mathbf{X})$$

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

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

The expectation shows us this estimator is unbiased:

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

## Heteroskedasticity Consistent (HC) Variance Estimates

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

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

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

Could use leave one out variance estimate:

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

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



## Heteroskedasticity Consistent (HC) Variance Estimates

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

# Heteroskedasticity Consistent (HC) Variance Estimates

To read about SE's in estimatr:

<https://declaredesign.org/r/estimatr/articles/mathematical-notes.html> or fixest

[https://lrberge.github.io/fixest/articles/fixest\\_walkthrough.html#the-vcov-argument-1](https://lrberge.github.io/fixest/articles/fixest_walkthrough.html#the-vcov-argument-1)

```
dat <- data.frame(X = matrix(rnorm(2000*5), 2000), y = rnorm(2000))
```

```
res<-feols(y ~ X.1+X.2+X.3+X.4+X.5, data = dat)
```

```
hc0<-summary(res, vcov = sandwich::vcovHC, type = "HC0")$se
```

```
hc1<-summary(res, vcov = sandwich::vcovHC, type = "HC1")$se
```

```
hc2<-summary(res, vcov = sandwich::vcovHC, type = "HC2")$se
```

```
hc3<-summary(res, vcov = sandwich::vcovHC, type = "HC3")$se
```

```
all(hc2 > hc0 )
```

```
[1] TRUE
```

```
all(hc3> hc2 )
```

```
[1] TRUE
```

# 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  $\text{Cov}(u_{g1}, u_{g2}, \dots, u_{gN}) \neq 0 \rightarrow$  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→ we have fewer effective independent observations in our sample.
- Asymptotics now about the number of groups  $G \rightarrow \infty$  not observations  $N \rightarrow \infty$

# Clustering

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

$$\widehat{\beta} = \left( \sum_{g=1}^G \sum_{i=1}^{n_g} x'_{ig} x_{ig} \right)^{-1} \left( \sum_{g=1}^G \sum_{i=1}^{n_g} x'_{ig} y_{ig} \right)$$

$$\widehat{\beta} = \left( \sum_{g=1}^G \mathbf{x}'_g \mathbf{x}_g \right)^{-1} \left( \sum_{g=1}^G \mathbf{x}'_g \mathbf{y}_g \right)$$

$$\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1} (\mathbf{X}'\mathbf{Y})$$

## Clustering (Continued)

The error terms have covariance within each cluster  $g$  as:

$$\Sigma_g = \mathbb{E}(\mathbf{u}_g \mathbf{u}_g' | \mathbf{X}_g)$$

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

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

And an estimate of the variance:

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

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

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

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

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

- Can replace  $\widehat{\mathbf{u}}_g \rightarrow \widetilde{\mathbf{u}}_g$  for leave-one out like *HC3* (these are called *CR3*).



```
lm_robust(y~ x1 + x2, data=df, se_type="CR0", cluster=group_id )  
lm_robust(y~ x1 + x2, data=df, se_type="CR2", cluster=group_id )  
lm_robust(y~ x1 + x2, data=df, se_type="CR1", cluster=group_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 → 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  $\text{Cov}(u_t, u_{t-1}, \dots) \neq 0$ .
- 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  $L$

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

## What about $\beta$ ?

- 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  $\hat{\beta}$  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)  $(X_i, Y_i)$ ,  $i = 1, \dots, n$ , are i.i.d.
- In this case, OLS is still unbiased and consistent, just **inefficient**

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  $w_i$ .

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

Can also consider a transformation of the data

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

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

$$\hat{\beta}_{WLS} = (\tilde{X}' \tilde{X})^{-1} \tilde{X}' \tilde{Y} = (X' W X)^{-1} X' W Y$$

Also used as a solution to heteroskedasticity

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

Why? We are minimizing weighted sum of squared residuals:

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

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**.

Why does anyone ever run OLS instead of WLS?

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



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

**Thanks!**

---