Lecture 5 – Linear Regression 1 – Inference

Paul Goldsmith-Pinkham

January 31, 2024

This lecture note will discuss focus on the simple linear model, and studying various cases for understanding inference. So far we've focused on *identification* – e.g. what estimands can we know from the data generating process? Now, given estimators for these estimands, we want to discuss uncertainty and inference.

In this lecture note, we'll be dancing around a fundamental tension. Much of the existing literature and approach to inference is model-based, and as a result, you may feel a bit of a bait-and-switch: I promised a lot of design-based discussion, and all of a sudden we're back in the old world.[1]

The reason for this is that the model-based world is a useful starting point for understanding the basic ideas of inference. A lot of important statistical ideas will show up in this setting, and it's also the world that most of the literature is written in, so it's important to understand the basic ideas. Overall, just think of these different approaches as different tools in your toolbox.

[1] Recall that when I talk about design-based inference, I am thinking about treating the potential outcomes as fixed $(Y(1), Y(0))$ and the treatment assignment D as random. Model-based inference, in the context of lecture note, is considering the model $Y = X\beta + D\tau + \epsilon$, and then thinking about the random sampling of the (Y, X, D) creating uncertainty in the estimates. In other words, it's about the variation in $\epsilon|X, D$.

Review: random sampling and linear regression

To fix notation, I want to do a refresher on notation and concepts about random sampling. This section is heavily based on review of materials from Gary Chamberlain, to which I am very grateful.[2] Conceptually, we have considered random variables (Y, X, D) from a joint distribution F:

$$(Y, X, D) \sim F. \tag{1}$$

[2] Gary Chamberlain was an extraordinary econometrician and former teacher of mine who had an amazing set of lecture notes that I got permissiont to post online.

Now, we will formalize the concept of the data we observe that approximates the full population. We'll consider a sample of size n, where the ith draw of the data gives us the variables (Y_i, X_i, D_i).[3] We will initially consider random sampling where each draw is independent and identically distributed (i.i.d.):

$$(Y_i, X_i, D_i) \overset{i.i.d.}{\sim} F. \tag{2}$$

[3] In panel settings, with T observations, it is easy to consider Y_i being a vector of $Y_{i1}, Y_{i2}, \ldots, Y_{iT}$, and similarly for the other variables. Then, these vectors will be treated as a unit.

Notationally, I'll often group $W_i = (X_i, D_i)$[4] in order to focus on a single set of non-outcome variables. The random sampling happens jointly – we make no distinctions between Y and W in the sampling process. We make that distinction later when we consider the conditional expectation of Y_i conditional on W_i. We will then stack the data $Y_n = (Y_1, \ldots, Y_n)$, $W_n = (W_1, \ldots, W_n)$.

[4] When I need to make a distinction, X will usually denote controls and D will be the causal variable(s) of interest.

We consider the linear predictor

$$E^*(Y_i|W_i) = W_i'\beta$$

where E^* denotes the linear predictor such that β are the minimizer of the expected squared error loss $E((Y - E^*(Y_i|W_i))^2)$.[5] I will assume W_i includes a constant for purposes of linear regression.

[5] Note that this is the traditional OLS estimator, and if we want to allow for more flexible functional forms of W_i, we'll need to include higher-order interactions and functions, etc.

Then, recall that

$$\beta = E(W_i W_i')^{-1} E(W_i Y_i).$$

Note that β is a population object, defined based on two moments.

Next recall that the least-squares estimator of β is

$$b(Y_n, W_n) = \left(n^{-1} \sum_{i=1}^{n} W_i W_i' \right)^{-1} \left(n^{-1} \sum_{i=1}^{n} W_i Y_i \right) \tag{3}$$

$$= (W_n' W_n)^{-1} W_n' Y_n. \tag{4}$$

Recall that b is a *function* of random variables Y_n and W_n (an estimator) and as such also a random variable. Since we can't directly study $E(b)$, we study instead $E(b|W_n)$, which focuses on the conditional distribution of $Y_i|W_i$.[6]

[6] Why can't we study $E(b)$? The nonlinearity makes it hard to study expectations – the expectation of a ratio is not equal to the ratio of expectations.

If we consider $E(b(Y_n, W_n|W_n = w))$, we see

$$E(b(Y_n, W_n|W_n = w)) = (w'w)^{-1} w' E(Y_n|W_n = w).$$

When we are correctly specified, and the conditional expectation $E(Y_n|W_n) = W_n\beta$, then we have:

$$E(b(Y_n, W_n|W_n = w)) = \beta.$$

Since this is true for any w, we can use the law of iterated expectations and $E(E(b(Y_n, W_n|W_n = w))) = \beta$.

We will now turn to inference in this setting.

Model-based inference

Given our linear project, $E^*(Y_i|W_i) = W_i'\beta$, we write

$$Y_i = W_i'\beta + \epsilon_i,$$

where ϵ_i denotes the error term that is mechanically orthogonal to W_i. As we saw above, we'll often consider the uncertainty in the sampling *conditional* on W_i, and hence the uncertainty that drives our estimate is from ϵ_i (the unexplained part of Y_i). Restating our estimator from before,

$$\hat{\beta} = \beta + (\mathbf{W}_n'\mathbf{W}_n)^{-1} \mathbf{W}_n' \boldsymbol{\epsilon}_n.$$

Typically we take $\mathbf{W}_n$ as given, and so the uncertainty (in the model based world) is driven by $\boldsymbol{\epsilon}_n$.

Now we consider the variance of $\hat{\beta}$. Formally, we see that this revolves around the structure of $\mathbb{E}(\boldsymbol{\epsilon}_n\boldsymbol{\epsilon}_n'|\mathbf{W}_n) = \Omega_n$:

$$\mathbb{V}(\hat{\beta}|\mathbf{W}_n) = (\mathbf{W}_n'\mathbf{W}_n)^{-1}\mathbf{W}_n'\mathbb{E}(\boldsymbol{\epsilon}_n\boldsymbol{\epsilon}_n'|\mathbf{W}_n)\mathbf{W}_n(\mathbf{W}_n'\mathbf{W}_n)^{-1} \tag{5}$$

$$= (\mathbf{W}_n'\mathbf{W}_n)^{-1}\mathbf{W}_n'\Omega_n\mathbf{W}_n(\mathbf{W}_n'\mathbf{W}_n)^{-1} \tag{6}$$

Everything pivots around the structure of $\mathbb{E}(\boldsymbol{\epsilon}_n\boldsymbol{\epsilon}_n'|\mathbf{W}_n) = \Omega_n$.

So far, we have assumed that the draws are independent, so we already know that Ω_n is a diagonal matrix.[7] To simplify further, we can consider is homoskedasticity, where $\Omega = \sigma^2 I_n$. This simplifies our variance:

$$\mathbb{V}(\hat{\beta})_{homoskedastic} = \sigma^2(\mathbf{W}_n'\mathbf{W}_n)^{-1} \tag{7}$$

What is the content of this assumption? Beyond $Cov(\epsilon_i, \epsilon_j) = 0$, it assumes $Var(\epsilon_i|W_i) = Var(\epsilon_i)$.[8]

[7] Why? Verify this for yourself if this is not clear.

[8] This means that conditional on $W = w$, the variance of Y is the same, regardless of the value of w. That's pretty strong.

Comment 1

Consider the linear regression model when posed as potential outcomes with unobserved heterogeneity and strong ignorability:

$$\begin{aligned}
Y_i &= \alpha + D_i\beta + \epsilon_i \\
\alpha &= E(Y_i(0)) \\
\beta &= E(Y_i(1)) - E(Y_i(0)) \\
\epsilon_i &= D_i\Big(\underbrace{(Y_i(1) - Y_i(0))}_{\beta_i} - \beta\Big) + \Big(Y_i(0) - E(Y_i(0))\Big).
\end{aligned}$$

Hence, the assumptions about $Var(\epsilon_i|D_i)$ relate directly the extent of heterogeneity in the treatment effect. Namely,

$$Var(\epsilon_i|D_i = 1) = Var(\beta_i|D_i = 1) + Var(Y_i(0)|D_i = 1) \tag{8}$$

$$Var(\epsilon_i|D_i = 0) = Var(Y_i(0)|D_i = 0). \tag{9}$$

These can only satisfy homoskedasticity and be equal if $Var(\beta_i) = 0$, and hence the treatment effect is constant. Under random assignment, the latter terms are equal by construction.

A feasible estimator for the homoskedastic variance estimand, where k is the number of regressors in β (excluding the constant), follows using empirical analogs:

$$\hat{\mathbb{V}}(\hat{\beta})_{homoskedastic} = \hat{\sigma}^2(\mathbf{W}_n'\mathbf{W}_n)^{-1} \qquad \hat{\sigma}^2 = (n-k-1)^{-1}\hat{\boldsymbol{\epsilon}}_n'\hat{\boldsymbol{\epsilon}}_n \tag{10}$$

If we instead want to allow $Var(\epsilon_i|W_i) = \sigma^2(W_i)$ to vary with W, this implies a potentially complicated function $\sigma^2(W_i)$. But, as

it turns out, the estimator for $\mathbb{V}(\hat{\beta})$ is quite straightforward. This is often referred to as the "robust" or EHW estimator [Eicker, 1963, Huber et al., 1967, White, 1980]:

$$\hat{\mathbb{V}}(\hat{\beta})_{EHW} = (\mathbf{W}_n'\mathbf{W}_n)^{-1}\sum_i \hat{\epsilon}_i^2 W_i W_i' (\mathbf{W}_n'\mathbf{W}_n)^{-1}.$$

Example 1
Consider the case where $W_i = (1, D_i)$, and D_i is a dummy treatment variable. Then, the variance of the coefficient on D_i reduces to

$$\hat{\mathbb{V}}(\hat{\beta})_{homoskedastic} = \frac{\hat{\sigma}^2}{n_0} + \frac{\hat{\sigma}^2}{n_1} = \frac{\hat{\sigma}^2}{n}, \qquad \hat{\mathbb{V}}(\hat{\beta})_{EHW} = \frac{\hat{\sigma}^2(0)}{n_0} + \frac{\hat{\sigma}^2(1)}{n_1},$$

where $n_0 = \sum_i (1 - D_i)$, $n_1 = \sum_i D_i$ and $\hat{\sigma}^2(x)$ is the estimated variance of ϵ_i for observations with $D_i = x$.

We then consider confidence intervals based around distributional assumptions. Recall that our distributional assumptions come from considering the following statistics:

$$T = \frac{\hat{\beta} - \beta}{\sqrt{\hat{\mathbb{V}}/\nu}}$$

When we assume that the distribution of ϵ is homoskedastic and Normal, we know the exact distribution of T. That's because β is consequently the sum of independent normals, and the variance of $\hat{\beta}$ is a scaled chi-squared (since a squared normal is chi-squared). This is the basis for the t-distribution.[9]

Without homoskedastic Normality, the distribution for T is only Student-t asymptotically. This approximation works pretty well in many settings, but there are edge cases where issues can arise.[10] One straightforward edge case we'll consider now is when n_1 and n_0 are not both simultaneously growing large.

[9] E.g. $T = Z/\sqrt{V/\nu}$, where $Z \sim \mathcal{N}(0,1)$, $V \sim \chi^2(\nu)$

[10] Edge cases that you likely know well include weak IV and unit roots. I call them edge cases because they usually involve sending a parameter to a boundary case, such as the first-stage coefficient to zero, or AR parameter to one.

Confidence intervals, finite sample performance, and the Behrens-Fisher problem

Note that in our discussion above, the feasible estimator for $\hat{\sigma}^2(x)$ was not made explicit. For the homoskedastic case, we have a simple estimator in Equation (10). We have adjusted for k in this estimator to account for the bias that arises from our estimation of β. A similar adjustment needs to occur for the heteroskedastic case to account for the estimation of the parameters as well. Note that this is a *finite sample* adjustment, since it will not matter as n gets large.

It is worth highlighting the different relevant adjustments that get made in the heteroskedastic case, which will matter for practical

inference.[11] The original EHW estimator is biased, and does not adjust for the k parameters. MacKinnon and White [1985] propose a second estimator, referred to as the HC2 estimator, that adjusts for the bias in the variance estimator.[12] It's exactly unbiased in the binary treatment case, but this is not generally true. In the binary case, it adjusts simply by subtracting one:

$$\hat{\sigma}^2(d) = \frac{1}{n_d - 1} \sum_i D_i (Y_i - \overline{Y}_d)^2.$$

So far, we have just discussed different estimators for $\mathbb{V}$. Recall that we want to use these to make confidence intervals, based on the distribution of T. That distribution requires an assumption about the degrees of freedom for T. Then, we can construct 95% confidence intervals based on these asymptotic results:

$$\text{CI} = \left(\beta - t^{n-2}_{0.975} \times \sqrt{\hat{\mathbb{V}}}, \beta + t^{n-2}_{0.975} \times \sqrt{\hat{\mathbb{V}}} \right)$$

where t^n_q is the qth quantile the t distribution with degrees of freedom n. The issue is that the degrees of freedom are not clear in the heteroskedastic case. Why? Because the variance is the weighted sum of *different* Chi-squared distributions. This is very concrete in Example 1, where the denominator of the scaling for the variance is driven by the relative size of the treatment and control groups.

We now consider the Behrens-Fisher problem. Imagine that $n_0 >> n_1$, that is, there are few treated units relative to the control.[13] Then, the distribution is really driven by $\sigma^2(1)/n_1$, and n_1 is the correct degrees of freedom. This makes a big difference! Contrast the degrees of freedom between the two cases: $t^3_{0.975} = 3.182$ vs. $t^{28}_{0.975} = 2.048$.[14] This naturally creates much wider confidence intervals for a given dataset, which implies that the coverage under the n_0 degrees of freedom would have much lower coverage than the n_1 degrees of freedom. We can see this coverage difference in simulations done by Imbens and Kolesar [2016] in Figure 1.

The estimator with the best performance in this setting is the Bell-Maccaffrey adjustment, which is a generalization of the HC2 estimator. This estimator adjusts for the degrees of freedom issue by finding the parameter K which creates a t-distribution that most closely matches the first two moments of the dispersion of the estimated $\hat{V}_{HC2}$ around the estimand.[15] In the binary case, this reduces to

$$K_{BM} = \frac{(n_0 + n_1)^2 (n_0 - 1)(n_1 - 1)}{n_1^2 (n_1 - 1) + n_0^2 (n_0 - 1)}.$$

Some intuition arises in the case when n_0 and n_1 are similar: we get $K_{BM} = n - 2$, which is the degrees of freedom in the homoskedastic case. If $n_0 >> n_1$, then we get $K_{BM} \approx n_1$. This is a very intuitive

[11] See the discussion in Comment 2 for implementations, and the implications of using biased estimators in Figure 1.

[12] There is also a third estimator, referred to as HC3, from MacKinnon [2012], that also adjusts for the bias in the variance estimator. It is quite conservative (it's biased upwards in the case of binary treatment). For the binary case, it is given by

$$\hat{\mathbb{V}}(\hat{\beta})_{EHW} = \hat{\sigma}^2(0) \frac{n_0}{(n_0 - 1)^2} + \hat{\sigma}^2(1) \frac{n_1}{(n_1 - 1)^2}.$$

I provide it for completeness, but it is not widely used.

[13] These results and discussion come from Imbens and Kolesar [2016].

[14] Notably, this issue starts to disappear as the minimum size gets large.

[15] This is done under the assumption of homoskedasticity, for *just* the purposes of estimating the degrees of freedom.

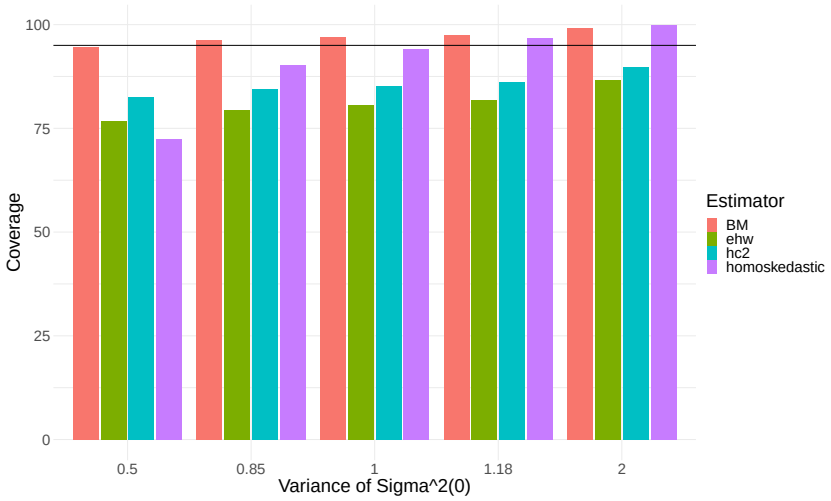


Figure 1: These simulation results come from Imbens and Kolesar [2016] and study the impact of unbalanced treatment and control. In the simulation, $n = 30, n_0 = 27, n_1 = 3$. The coverage of the different estimators discussed in **??** are shown. The Bell-Maccaffrey adjustments are the most accurate, and EHW does poorly in all cases. The data is simulated to be Normally distributed conditional on treatment, and the Variance of the error term is allowed to vary.

result, and lines up with what we'd want. Imbens and Kolesar [2016] recommend this estimator to researchers in practice.

The Behrens-Fisher issue (and Bell-Maccaffrey solution) generalizes to general regression setting, even when the treatment is not binary. The key idea is that the variance we're scaling by is not a Chi-squared with the full degrees of freedom n. The approximation that we use matches the degrees of freedom to get the first and second moment as close as possible to the "right" chi-squared.[16]

Comment 2

What are the statistical packages that we use in these cases, and what's doable?

- *Stata uses the EHW standard errors by default (with the* `robust` *command). But, I believe as of Stata 17, it does the correct finite sample adjustment for the degrees of freedom to make it unbiased in the simple case. You can use* `vce(hc2)` *to get the HC2 estimator, and* `vce(hc3)` *to get the HC3 estimator. For the Bell-Maccaffrey adjustments, see* `reg_sandwich`*.*
- *In R, there are many pacakges, but the ones I would look see* `estimatR`*,* `clubSandwich`*, and Kolesar's github repo.*

[16] These issues can rear their head when the the regressor of interest is highly skewed (e.g. a log-normal right hand side variable). The intuition comes from the fact that the distribution of the regression will affect the distribution of $(W_n' W_n)^{-1} W_n' \epsilon_n$, warping the finite sample behavior of the sum of the ϵ_n sum.

Example 2

You might be asking yourself, why do I have to care about these things? For a lovely discussion of finite sample issues, you should peruse Data Colada's discussion of Alwyn Young's QJE paper on randomization inference. The paper is a great example of how finite sample issues can matter in practice.

To quote the post: "In this post I show that this conclusion only holds when relying on an unfortunate default setting in Stata. In contrast, when regression results are computed using the default setting in R [1], a setting that's also available in Stata, a setting shown over 20 years ago to be more appropriate than the one used by Stata... the supposed superiority of the randomization test goes away."

Bootstrap One nice alternative approach to distributional assumptions on T is to use the bootstrap. The bootstrap is a general purpose tool for constructing confidence intervals, and it can be used in the context of linear regression. The idea is to resample the data with replacement, and then re-estimate the model. This comes in many forms, but the most intuitive and straightforward form is called the non-parametric bootstrap, which would involve resampling n observations of (Y_i, W_i) from the data with replacement, and then re-estimating the model. After B samples are re-estimated, this will give us a distribution of the parameter and we can describe the statistical properties of this distribution (e.g. the 95% interval of this distribution). It is also plausible to construct t-statistics $t_b = (\hat{\beta}_b - \beta^0)/\sqrt{V_b}$ for each bootstrap sample b and null hypothesis β^0, and use this instead of $\hat{\beta}$. Since the t-statistic is asymptotically pivotal, it can have nice properties.

However, the non-parametric bootstrap can have issues if the sample is small or the regressors are skewed [Imbens and Kolesar, 2016], as the additional noise introduced by resampling creates worse distributional approximations. One very successful bootstrap alternative is the wild bootstrap. Concretely, this works as follows (see Davidson and Flachaire [2008] for details):

1. Estimate the linear model $\hat{\beta}$ and obtain residuals $\hat{\epsilon}_i$.
2. In each bootstrap step b:
   (a) For each observation i, the X_i is fixed, along with $\hat{\beta}$ and $\hat{\epsilon}_i$. We then draw a binary variable $U_{i,b}$ that is either 1 or -1 with equal probability. We set $Y_{i,b} = \hat{\beta} X_i + U_{i,b}\hat{\epsilon}_i$.
   (b) With the new dataset, we re-estimate the model and construct a t-statistic $t_b^1 = (\hat{\beta}_b - \hat{\beta})/\sqrt{\hat{V}_b}$, where $\sqrt{\hat{V}_b}$ is the standard

error.

3. With the full set of t_b, we can construct a confidence interval by calculating the $q_{0.95}(|t_b|)$, the 0.95 quantile of $|t_b|$, and using it in the place of our usual critical value:

$$\text{CI}_{WILD} = \left(\hat{\beta} - q_{0.95}(|t_b|)\sqrt{\hat{V}}, \hat{\beta} + q_{0.95}(|t_b|)\sqrt{\hat{V}}\right)$$

Combining Sampling- and Design-based uncertainty

How should we be thinking about inference anyway? What's the error in ϵ *mean*? The thought experiment typically comes from a sampling perspective – we consider that this is a small sample from a broader population, and uncertainty comes from whether the estimates reflect the true underlying population. Note that this contrasts with our design-based thought experiment!

This starts to get very confusing when thinking about some settings. For example, how do we think about sampling "new states" when we have all 50 states? Worse yet, what if we have access to all the census data? We observe the whole population! What's the uncertainty in our estimates then? In estimation of causal effects, we still uncertainty because there's uncertainty driven by the fundamental problem of causal inference!

Using work from Abadie et al. [2020], we will now consider two sources of uncertainty: sampling and design. There exists a population of size N and a sample of size $n \leq N$.[17] Let $R_i = \{0, 1\}$ denotes whether or not an observation is in the sample. There are also potential outcomes $Y^*(D_i)$. Now we have both sampling uncertainty (e.g. does our sample reflect the population) and design uncertainty (e.g. does the causal comparison reflect the true causal effect). We can now combine the two sources of uncertainty to get a better understanding of the variance of our estimator.

[17] My n, N are reversed from the paper.

I will focus on just binary case of a single treatment, but the paper considers full regression setting. We consider three estimands:

1. $\theta^{descr} = N_1^{-1} \sum_{i=1}^{N} D_i Y_i - N_0^{-1} \sum_{i=1}^{N} (1 - D_i) Y_i$
2. $\theta^{causal,sample} = n^{-1} \sum_{i=1}^{N} R_i (Y_i^*(1) - Y_i^*(0))$
3. $\theta^{causal} = N^{-1} \sum_{i=1}^{N} (Y_i^*(1) - Y_i^*(0))$

We have a single estimator we can consider:

$$\hat{\theta} = n_1^{-1} \sum_{i=1}^{N} R_i D_i Y_i - n_0^{-1} \sum_{i=1}^{N} R_i (1 - D_i) Y_i$$

The key point of paper – the variance of this estimator depends on what we condition on. If we condition on D, we focus on sampling

uncertainty. If we condition on R we focus on causal uncertainty within sample. If we condition on neither, we focus on causal uncertainty as well as sampling uncertainty.

What are these variance estimands? Let S_x denote the population variance for each potential outcome, and S_θ denote the population variance of the treatment effect outcomes (which recall, we cannot directly estimate).[18] Then, we have the following variance estimands:

1. Sampling: $E(Var(\hat{\theta}|\mathbf{D}, n_1, n_0)|n_1, n_0) = \frac{S_1^2}{n_1}\left(1 - \frac{n_1}{N_1}\right) + \frac{S_0^2}{n_0}\left(1 - \frac{n_0}{N_0}\right)$,
2. Design: $E(Var(\hat{\theta}|\mathbf{R}, n_1, n_0)|n_1, n_0) = \frac{S_1^2}{n_1} + \frac{S_0^2}{n_0} - \frac{S_\theta^2}{n_1+n_0}$
3. Both: $Var(\hat{\theta}|n_1, n_0) = \frac{S_1^2}{n_1} + \frac{S_0^2}{n_0} - \frac{S_\theta^2}{N_1+N_0}$

[18] This variance is $S_\theta = \frac{1}{N-1}\sum_i \left(Y_i(1) - Y_i(0) - N^{-1}\sum_i (Y_i(1) - Y_i(0))\right)^2$.

The S_θ^2 term is what we usually ignore (because not feasibly estimable). Consider the following thought experiments: let n get close to N, or let n get small relative to N. This will move around the sampling and total variance estimands, but not the sample causal estimand. Next, notice that the difference between Sampling and Design is not obvious. Sampling can be small if n is close to N, but when S_θ^2 is large, that can make the design variance small.

Clustering and generalizing Ω

This ignored any sort of unusual correlation structure in Ω, and assumed random assignment. In many cases, we don't have that. Instead, Ω has a clustering structure. This can get quite complex. Let's start with the simple case of known clusters. E.g. units are people, and clusters are cities, counties or states. For today, we're ignoring the very important question of panel data. We'll come back to that later.

Let C_i denote unit i's cluster assignment. A very simple version of Ω is now

$$\Omega_{ij} = \begin{cases} \sigma^2 & \text{if} \quad i = j \\ \rho\sigma^2 & \text{if} \quad C_i = C_j \;\&\; i \neq j \\ 0 & \text{if} \quad C_i \neq C_j \;\&\; i \neq j \end{cases} \tag{11}$$

This matrix can also be more unstructured. For exmaple, one could have a flexible block diagonal with $\Omega_{ij} = \sigma_{ij}$ if $C_i = C_j$. A key issue that arises here is in the relative size of the block versus the number of blocks. See Hansen [2007] for a discussion of one example.

Let's start with a model-based approach to this. Denote the number of clusters by K. Following Liang and Zeger [1986], the estimator

for the variance of $\hat{\beta}$ is

$$\hat{\mathbb{V}}(\hat{\beta}|\mathbf{W}_n, \mathbf{C}_n)_{LZ} = (\mathbf{W}_n'\mathbf{W}_n)^{-1} \left(\sum_{k=1}^{K} \mathbf{W}_{k,n}' \hat{\epsilon}_{k,n} \hat{\epsilon}_{k,n}' \mathbf{W}_{k,n} \right) (\mathbf{W}_n'\mathbf{W}_n)^{-1}. \tag{12}$$

This estimator allows for flexible covariance within the block, and assumes that the size of the block is fixed, and that K is large.[19] Historically, clustering in this setting has focused the structure of Ω. Why? Well, take the simple case in Equation (11). In this case,

$$\mathbb{V}(\hat{\beta}) = \mathbb{V}_{homoskedastic} \times \left(1 + \rho_\epsilon \rho_W \frac{n}{K_n}\right), \tag{13}$$

where ρ_ϵ and ρ_W are the within-cluster correlations of each r.v. This makes you think that these are the main terms that matter, and more generally it's about getting the structure of Ω right. E.g., better to err on the conservative side and let the blocks be large.

[19] In fact, the degrees of freedom are defined by the size of K.

However this intuition is *not* correct in contexts with any meaningful heteroskedasticity. Abadie et al. [2023] can generate an example with tiny within-cluster correlation, and large clusters (with many clusters) where the Liang-Zeger estimator $\hat{\mathbb{V}}(\hat{\beta})_{LZ}$ is large and $\hat{\mathbb{V}}(\hat{\beta})_{EHW}$ is small. How come? Recall from Comment 1 that the variance of the error term depends on two pieces: the variance of the poetntial outcome, and the variance of the tratement effect, as it correlates with treatment. Namely, it's all about the correlation *between* W and ϵ, and heterogeneity in our effects across clusters.

In Abadie et al. [2023], they construct an example where $N = 10,000,000$ with 100 equal sized clusters. There is a binary W with *equal* probability. THere are significant heterogeneous effects of W across clusters – some clusters have positive effect, some have negative. Overall ATE is 0. What does that mean intuitively? If there is heterogeneity in effects, it causes correlation between treatment and residual.

So why do the two standard error estimators vary so much? To quote the Abadie et al. [2023] working paper:

> The reason for the difference between the EHW and LZ standard errors is simple, but reflects the fundamental source of confusion in this literature. Given the random assignment both standard errors are correct, but for different estimands. The LZ standard errors are based on the presumption that there are clusters in the population of interest beyond the 100 clusters that are seen in the sample. The EHW standard errors assume the sample is drawn randomly from the population of interest. It is this presumption underlying the LZ standard errors of existence of clusters that are not observed in the sample, but that are part of the population of interest, that is critical, and often implicit, in the model-based motivation for clustering the standard errors. It is of course explicit in the sampling design literature (e.g., Kish [1965]). If we changed the set up to one where the population of 10,000,000 consisted of say 1,000 clusters, with 100 clusters drawn at random, and then sampling units randomly from those sampled clusters, the LZ standard errors would be correct, and the EHW standard errors would be incorrect. Obviously one cannot tell from the sample itself whether there exist such clusters that are part of the population of interest that are not in the sample, and therefore one needs to choose between the two standard errors on the basis of substantive knowledge of the study design.

What are the key takeaways from this paper? First, cluster your regression at the unit of randomization. Being conservative can be quite bad! It depends on what you are trying to do. The traditional advice of being as conservative as necessary is likely misguided. Fixed effects do NOT remove the need for clustering. We'll revisit this in panel settings.

Discussion Questions 1
What is the "unit of randomization" in a case like CARD and KRUEGER [1994]?

Comment 3 (Spatial and Network Error)
Things get more complicated with more general error structures. Consider two additional cases:

- *Spatial correlation = $\rho_{ij} = f(d_{ij})$, where d_{ij} is a function of some economic distance.*
- *Social network correlation = $\rho_{ij} = f(d_{ij})$, where d_{ij} is a function of path length in a network*

This can matter especially when SUTVA is violated However, Barrios et al. (2012) show that, under SUTVA, if treatments are randomly assigned at a given cluster level, we can ignore the broader spatial correlations
Conley (1999) provides a flexible way to consider clustering on spatial distances. Consider our matrix Ω again. Now, Ω_{ij} is a function of the distance, d_{ij}, between each person. Unfortunately, this means that every person can be correlated. Key assumption – the correlation declines with distance. Hence, far away distances matter less in practice. Hence, when we estimate this, we "window" our estimator (this is exactly the same as Newey-West estimators). Then we allow correlation as in the Liang-Zeger estimator, as a function of distances. This estimator is consistent for general forms of spatial correlation

- *Estiators available in both Stata and R*

Example 3
Consequences of ignoring spatial correlation
Spatial correlation can be a big deal. Consider the analogy to time series.

- *A big rule: worry about highly autocorrelated data! Can inflate your t-statistics substantially*
- *Why? Because if we treat observations as independent, we will infer more information than actually exists*

Kelly (2019) claims that spatial correlation in outcomes can cause this same issue. Consider a regression of some modern outcome, e.g. city income, on a historical characteristic, such as colonial boundaries

- *Claim in Kelly (2019) is that t-statistics in these types of regressions are grossly amplified by spatial correlation*
- *Fixable with Conley standard errors?*
- *This is a huge deal for a lot of literatures (economic history especially) – matters for corporate governance literature too (LLSV)*

Concluding thoughts

This stuff is *hard*. We are doing the simplest case (linear regression) and still have lots of questions. As always, asking what the knowable estimand is can be very helpful. Next, if you are unsure, it is very useful to consider simulating data. [20] In many cases, there is not an obvious "best" answer, and simulating your data is the best solution. This is because many results are asymptotic in nature, and hence approximations.

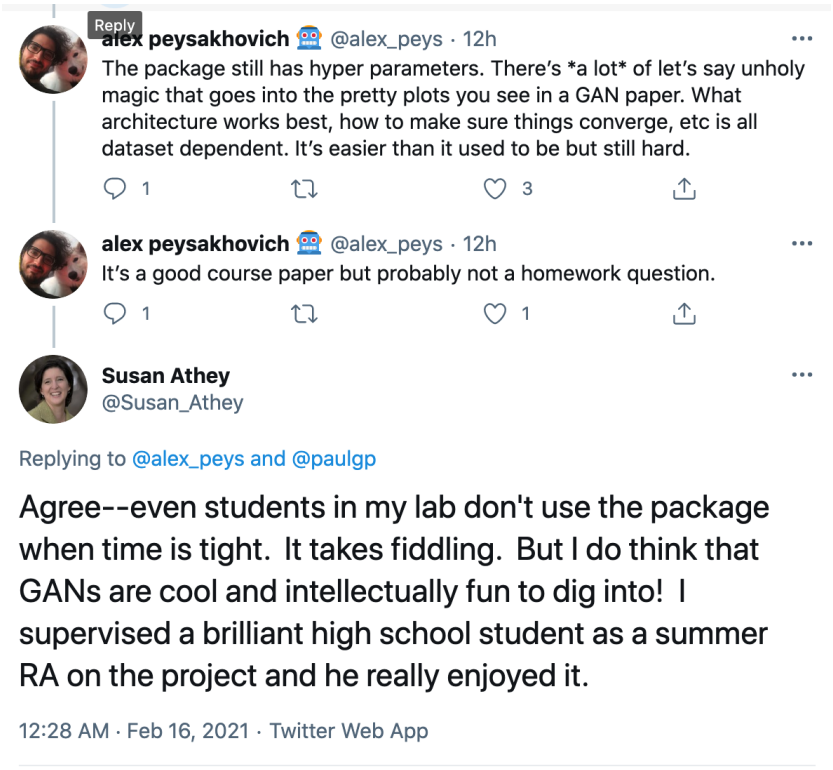
[20] This is the approach advocated in Blair et al. [2023].

So how does one implement a simulation? Goal is to generate data that matches your dataset's distributions. However, for very simple simulations, you'll have to make parametric assumptions that may not match your actual data. Athey et al. (2020) propose a method for matching the data as closely as possible, using a Generative Adversarial Network. In other words, construct distributions that match the "true" data as closely as possible

- Computationally expensive, but great way to evaluate performance
- Code is available here: `https://github.com/gsbDBI/ds-wgan`
- Docs are here: `https://github.com/gsbDBI/ds-wgan`

However this stuff is really hard to implement. If you intuitively

know the issue, try doing something simple with normals Or try bootstrapping!



References

Alberto Abadie, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge. Sampling-based versus design-based uncertainty in regression analysis. *Econometrica*, 88(1):265–296, 2020.

Alberto Abadie, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge. When should you adjust standard errors for clustering? *The Quarterly Journal of Economics*, 138(1):1–35, 2023.

Graeme Blair, Alexander Coppock, and Macartan Humphreys. *Research Design in the Social Sciences: Declaration, Diagnosis, and Redesign*. Princeton University Press, 2023.

DAVID CARD and ALAN B KRUEGER. Minimum wages and employment: A case study of the fast-food industry in new jersey and pennsylvania. *The American Economic Review*, 84(4):772–793, 1994.

Russell Davidson and Emmanuel Flachaire. The wild bootstrap, tamed at last. *Journal of Econometrics*, 146(1):162–169, 2008. ISSN

0304-4076. DOI: https://doi.org/10.1016/j.jeconom.2008.08.003. URL https://www.sciencedirect.com/science/article/pii/S0304407608000833.

Friedhelm Eicker. Asymptotic normality and consistency of the least squares estimators for families of linear regressions. *The annals of mathematical statistics*, 34(2):447–456, 1963.

Christian B Hansen. Asymptotic properties of a robust variance matrix estimator for panel data when t is large. *Journal of Econometrics*, 141(2):597–620, 2007.

Peter J Huber et al. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 221–233. Berkeley, CA: University of California Press, 1967.

Guido W Imbens and Michal Kolesar. Robust standard errors in small samples: Some practical advice. *Review of Economics and Statistics*, 98(4):701–712, 2016.

Kung-Yee Liang and Scott L Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

James G MacKinnon. Thirty years of heteroskedasticity-robust inference. In *Recent advances and future directions in causality, prediction, and specification analysis: Essays in honor of Halbert L. White Jr*, pages 437–461. Springer, 2012.

James G MacKinnon and Halbert White. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of econometrics*, 29(3):305–325, 1985.

Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: journal of the Econometric Society*, pages 817–838, 1980.