Linear Regression I: Inference

Paul Goldsmith-Pinkham

January 31, 2023

Linear Regression and Inference

- Today, we'll be focusing 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

- Let's define some notation to start. We'll consider a sample of size n:
  - causal variable D_i (can be continuous valued) $\rightarrow$ $\mathbf{D}_n$;
  - outcome $Y_i \rightarrow$ $\mathbf{Y}_n$;
  - controls X_i (relevant for exogeneity, and can be vector valued) $\rightarrow$ $\mathbf{X}_n$.

- We'll assume SUTVA holds, and that D_i is exogeneous conditional on X_i

Start with the traditional model-based approach

- Regression is typically written as

$$Y_i = X_i\gamma + D_i\tau + \epsilon_i,$$

where ϵ_i denotes the error term, and is driving the randomness to estimate τ (and γ). To see this, let $W_i = (X_i, D_i)$ such that

$$Y_i = W_i\beta + \epsilon_i.$$

- Note that $\hat{\beta} = \beta + (\mathbf{W}_n'\mathbf{W}_n)^{-1}\mathbf{W}_n'\boldsymbol{\epsilon}_n$
  - Typically we take $\mathbf{W}_n$ as given, and so the uncertainty (in the model based world) is driven by ϵ_i

Start with the traditional model-based approach

- The variance of the estimator of $\hat{\beta}$ is

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

  Everything pivots around the structure of $\mathbb{E}(\boldsymbol{\epsilon}_n\boldsymbol{\epsilon}_n'|\mathbf{W}_n) = \Omega_n$. The simplest case we can consider is homoskedasticity, where $\Omega = \sigma^2 I_n$.

- This beautifully simplifies our variance:

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

- What is the content of this assumption? That $Cov(\epsilon_i, \epsilon_j) = 0$ and $Var(\epsilon_i|W_i) = Var(\epsilon_i)$.

- Feasible estimator (k = number of regressors):

$$\hat{\mathbb{V}}(\hat{\beta})_{homoskedastic} = \hat{\sigma}^2(\mathbf{W}_n'\mathbf{W}_n)^{-1} \qquad \hat{\sigma}^2 = (n-k-1)^{-1}\hat{\boldsymbol{\epsilon}}_n'\hat{\boldsymbol{\epsilon}}_n \quad (3)$$

Start with the traditional model-based approach

- Under heteroskedasticity (which implies that $Var(\epsilon_i|W_i) = \sigma^2(W_i)$, the robust estimator comes from Eicker (1967), Huber (1967) and White (1980) [EHW]:

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

- Consider the simple dummy treatment case. Here, we have

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

- Key question for later: how do we estimate $\hat{\sigma}^2(x)$?

Confidence intervals

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

- We then consider confidence intervals based around distributional assumptions.
- Recall that our distributional assumptions come from considering the following statistics: $T = \frac{\hat{\beta}-\beta}{\hat{V}}$
  - When we assume that the distribution of ϵ is Normal, we know the exact distribution of T (under homoskedasticity).
  - That's because β is normally distributed, and a Normal divided by square root of the variance (which is chi-squared in distribution).
  - E.g. $T = Z/\sqrt{V/\nu}$, where $Z \sim \mathcal{N}(0,1)$, $V \sim \chi^2(\nu)$
- Without Normality, only holds asymptotically (asymptotically pivotal). Without homoskedasticity, holds asymptotically, but becomes more complicated if n_1 and n_0 are not both growing large.

Confidence intervals and the Behrens-Fisher problem

- We construct 95% confidence intervals based on these asymptotic results:

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

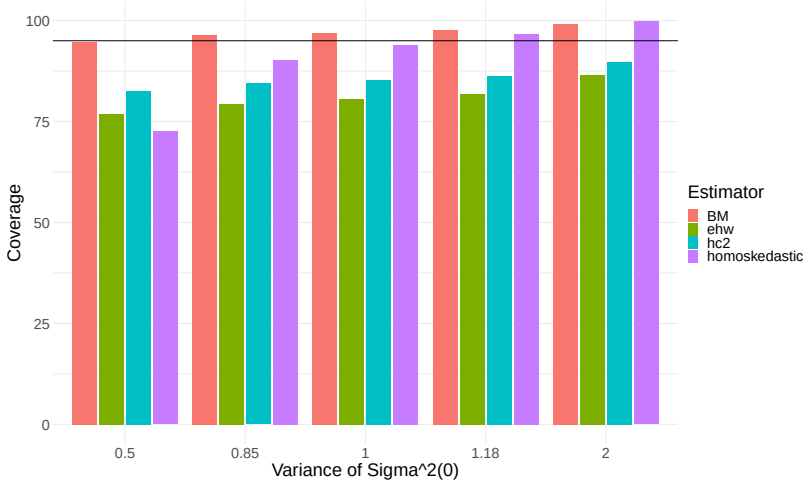
  where t^n_q is the qth quantile the t distribution with degrees of freedom n

- The trick – when we construct this statistic in the heteroskedastic case, we're not dividing by the right finite sample variance.
  - Why? Because the variance is the weighted sum of *different* Chi-squared distributions

- The Behrens-Fisher problem – imagine that $n_0 >> n_1$. (E.g. many untreated, a few treated).
  - 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 $t^3_{0.975} = 3.182$ vs. $t^{28}_{0.975} = 2.048$

Simulation Example (Imbens and Kolesar)

- Simulation: $n = 30, n_0 = 27, n_1 = 3$
- Normally distributed conditional on treatment
- Relative variance varies from 0.5 to 2



Confidence intervals and the Behrens-Fisher problem

- This generalizes to general regression setting (even when it's not binary)
- The key idea – the variance we're scaling by is not a Chi-squared with the full degrees of freedom. We want to match the distribution as best we can.
- 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.
- This accounts for issues like a highly skewed regressor (log-normal right hand side variable)

Doing this in practice

- These are solveable and packages exist. Stata uses the EHW standard errors by default. For the Bell-Maccaffrey adjustments, see `reg_sandwich`. For R, see `estimatR`, `clubSandwich`, and Kolesar's github repo: `https://github.com/kolesarm/Robust-Small-Sample-Standard-Errors`

- Key point is that these are all approximations in finite sample

- What are the options for estimating $\hat{\sigma}^2(x)$?

$$u_j = Y_j - W_j\beta, h = \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'$$

$$\hat{\mathbb{V}}(\hat{\beta})_{robust} = (\mathbf{W}'\mathbf{W})^{-1}\sum_j (n/n-k)u_j^2 W_i' W_i (\mathbf{W}'\mathbf{W})^{-1} \quad \text{(robust)}$$

$$\hat{\mathbb{V}}(\hat{\beta})_{HC2} = (\mathbf{W}'\mathbf{W})^{-1}\sum_j (1/(1-h_{jj}))u_j^2 W_i' W_i (\mathbf{W}'\mathbf{W})^{-1} \quad \text{(HC2)}$$

$$\hat{\mathbb{V}}(\hat{\beta})_{HC3} = (\mathbf{W}'\mathbf{W})^{-1}\sum_j (1/(1-h_{jj})^2)u_j^2 W_i' W_i (\mathbf{W}'\mathbf{W})^{-1} \quad \text{(HC3)}$$

Uncertainty – what is it?

- How should we be thinking about inference anyway? What's the error in ϵ mean anyway?

- 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.
  - E.g., How do we think about sampling "new states" when we have all 50 states?
  - What if we have access to all the census data?

- We still think we have uncertainty here when we do estimates. That's because there's uncertainty driven by the fundamental problem of causal inference!

Combining Sampling and Design Based uncertainty (Abadie et al. (2020))

- Consider now two sources of uncertainty. There exists a population of size N and a sample of size $n \leq N$.
- $R_i = \{0, 1\}$ denotes whether or not an observation is in the sample
- There is also potential outcomes $Y^*(D_i)$, driven by the causal variable.
- 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)
- But what is amazing is that we can combine the two – e.g. design uncertainty within a sample, versus within the population (internal validity vs. external validity)

Combining Sampling and Design Based uncertainty

- Will focus on just binary case, but paper considers full regression setting

- Three estimands (apologies in advance, my n, N are reversed from the paper):
  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:

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

- Key point of paper – the variance of this estimator depends on:
  1. We condition on D – e.g. focus on sampling uncertainty
  2. We condition on R – e.g. focus on causal uncertainty within sample
  3. We allow for both variances

Combining Sampling and Design Based uncertainty

- What are these variances?
  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: $Var(\hat{\theta}|n_1, n_0) = \frac{S_1^2}{n_1} + \frac{S_0^2}{n_0} - \frac{S_\theta^2}{N_1+N_0}$

  The S_θ^2 term is what we usually ignore (because not feasibly estimable)

- Thought experiments:
  1. Let n get small relative to N.
  2. Not obvious which of Sampling and Design is bigger
  3. Ignoring finite sample features will lead us to overstate variance!

Combining Sampling and Design Based uncertainty

Key takeaways from paper:

1. Design-based uncertainty can be smaller than traditional estimates, especially when the sample is large relative to the population
2. Defining the relevant estimand is important for your variance estimates (e.g. causal estimates within sample, or causal estimates across populations)
    - This is particularly true when your sample is a convenience sample
    - Or, when the given sample makes sense – e.g. 50 states
3. If you have a finite sample that is a non-trivial share of the population, you can improve your standard errors (see the paper)
4. Don't get confused by the idea of having uncertainty in your estimates when using the population. This comes from design-based uncertainty, not sampling uncertainty

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 quesiton of panel data
  - We're going to discuss this! Just not today
- Let C_i denote unit i's cluster assignment. A very simple version of Ω is now

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

- Can also be more unstructured – e.g. flexible block diagonal with $\Omega_{ij} = \sigma_{ij}$ if $C_i = C_j$. Key issue – asymptotics (see Hansen (2007))

Start with the traditional model-based approach

- Let the number of clusters be K. In this case, the estimator for the variance of $\hat{\beta}$ (this comes from Liang and Zeger (1986)) is

$$\hat{\mathbb{V}}(\hat{\beta}|\mathbf{W}_n, \mathbf{C}_n)_{LZ} = (\mathbf{W}_n'\mathbf{W}_n)^{-1} \left(\sum_{k=1}^{K} \mathbf{W}_{k,n}' \hat{\boldsymbol{\epsilon}}_{k,n} \hat{\boldsymbol{\epsilon}}_{k,n}' \boldsymbol{W}_{k,n} \right) (\mathbf{W}_n'\mathbf{W}_n)^{-1} \tag{4}$$

- Historically, clustering in this setting has focused the structure of Ω. Why? Well, take the simple case from last slide. In this case,

$$\mathbb{V}(\hat{\beta}) = \mathbb{V}_{homoskedastic} \times \left(1 + \rho_\epsilon \rho_W \frac{n}{K_n} \right), \tag{5}$$

  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

Clustering is about correlation between treatment and error

- However this intuition is *not* correct. (Abadie et al. (2017)) Can generate an example with tiny within-cluster correlation, and large clusters (with many clusters) where:
  - $\hat{\mathbb{V}}(\hat{\beta})_{LZ}$ is large
  - $\hat{\mathbb{V}}(\hat{\beta})_{EHW}$ is small

- Why? It's all about the correlation between W and ϵ, and heterogeneity in our effects across clusters.

- In Abadie et al example:
  - Pop = 10M, 100 equal sized clusters. Binary W with *equal* prob. = 0.5
  - However, heterogeneous effects – 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
  - Why do the two standard errors vary so much?

Abadie et al

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.

Takeaways from paper

- Key takeaways:
  1. Cluster your regression at the unit of randomization
  2. Being conservative can be quite bad! It depends on what you are trying to do.
  3. The traditional advice of being as conservative as necessary is likely misguided
  4. Fixed effects do NOT remove need for clstering

- We'll revisit this in panel settings. However, a question: what is the "unit of randomization" in a case like Card and Krueger (1993)?
  - Not totally obvious. Likely impossible to say something truly causal without strong assumptions about simultaneous shocks

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) distances

- 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

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)

Final 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
  - 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
- Ok, so how to do simulations?

Final thoughts

- 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 Adverserial 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`

Final thoughts

- However this stuff is really hard to implement.
- If you intuitively know the issue, try doing something simple with normals
- Or try bootstrapping!