1 Consistency

Consider a model

$$Y = X\beta + U \quad X \not\perp U$$

with instrument Z such that $\mathbb{E}\left[Z'X\right]$ has full rank and $\mathbb{E}\left[Z'U\right] = 0$. Compare the consistency of 2 estimators:

1. IV (or 2SLS) applied to sample of size N where $N \to \infty$
2. OLS for $N \to N^* \leq 10^{12}$ and TSLS for $N^* > 10^{12}$.

Problem 1.1. IV (or 2 SLS) applied to sample of size N where $N \to \infty$.

Solution. We have the following model (redefining notation so that X and Z are column vectors)

$$Y = X'\beta + U.$$

The 2SLS estimator assumes the following:

$$X = \Pi' Z + V, \mathbb{E}[Z'U] = \mathbb{E}[Z'V] = 0.$$

Using that $\mathbb{E}[Z'V] = 0$, we have

$$\Pi = \mathbb{E}[ZZ']^{-1}\mathbb{E}[ZX'].$$

Using that $\mathbb{E}[Z'U] = 0$, we have

$$\beta = \mathbb{E}[\Pi' Z Z' \Pi]^{-1}\mathbb{E}[\Pi' Z Y].$$

Thus the 2SLS estimator is:

$$\hat{\Pi} = \left(\frac{1}{N}\sum_{i=1}^{N} Z_i Z_i'\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} Z_i' X_i\right)$$

$$\hat{\beta}_{2SLS} = \left(\frac{1}{N}\sum_{i=1}^{N} \hat{\Pi}' Z_i Z_i' \hat{\Pi}\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} \hat{\Pi} Z_i' Y_i\right).$$

By the law of large numbers (LLN)

$$\frac{1}{N}\sum_{i=1}^{N} Z_i Z_i' \xrightarrow{P} \mathbb{E}[ZZ']$$

$$\frac{1}{N}\sum_{i=1}^{N} Z_i' X_i \xrightarrow{P} \mathbb{E}[ZX'].$$

So by the continuous mapping theorem (CMT)

$$\hat{\Pi} \xrightarrow{P} \Pi.$$

Then by CMT

$$\left(\frac{1}{N}\sum_{i=1}^{N} \hat{\Pi}' Z_i Z_i' \hat{\Pi}\right)^{-1} \xrightarrow{P} \mathbb{E}[\Pi' Z Z' \Pi]^{-1}.$$

By LLN:

$$\frac{1}{N}\sum_{i=1}^{N} Z_i' Y_i \xrightarrow{P} \mathbb{E}[ZY],$$

and by CMT

$$\frac{1}{N}\sum_{i=1}^{N} \hat{\Pi} Z_i' Y_i \xrightarrow{P} \mathbb{E}[\Pi' ZY],$$

Thus, by CMT

$$\hat{\beta}_{2SLS} \xrightarrow{P} \beta.$$

Problem 1.2. OLS for $N \to N^* \leq 10^{12}$ and TSLS for $N^* > 10^{12}$.

Solution. The OLS estimator is:

$$\begin{aligned}\hat{\beta}_{OLS} &= \left(\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} X_i Y_i\right)\\ &\xrightarrow{P} \mathbb{E}[XX']^{-1}\mathbb{E}[X_i Y_i],\end{aligned}$$

by LLN (on each sum) and CMT (on the inverse in the first factor and on the product of the two factors). But since $\mathbb{E}[XU] \neq 0$,

$$\beta \neq \mathbb{E}[XX']^{-1}\mathbb{E}[X_i Y_i].$$

Instead:

$$\begin{aligned} Y &= X'\beta + U\\ \leftrightarrow XY &= XX'\beta + XU\\ \leftrightarrow \mathbb{E}[XY] &= \mathbb{E}[XX']\beta + \mathbb{E}[XU]\\ \leftrightarrow \beta &= \mathbb{E}[XX']^{-1}(\mathbb{E}[XY] - \mathbb{E}[XU])\\ &\neq \mathbb{E}[XX']^{-1}\mathbb{E}[X_i Y_i],\end{aligned}$$

where the last expression is the asymptotic limit of $\hat{\beta}_{OLS}$ and the non-equality follows since $\mathbb{E}[XU]$ needn't equal 0 since $X \not\perp U$. Thus, even in infinite samples $\hat{\beta}_{OLS}$ is not consistent, so it definitely won't be consistent in finite samples.

Thus, this compound estimator will converge (get close to) to the wrong value ($\mathbb{E}[XX']^{-1}\mathbb{E}[X_i Y_i]$) as $N \to N^*$, and then once N switches to to be greater than N^*, the estimator will snap to be close to β (since $N^* = 10^12$ is large), and will converge monotonically to β after that.

2 Properties of Estimators

Discuss the properties of estimators obtained from the following specification-testing regression procedure for model

$$Y = X_1\beta_1 + X_2\beta_2 + U$$

where (X_1, X_2) are scalars; $(X_1, X_2) \perp U$; $|\text{Cov}\,(X_1, X_2)| > 0$; $(X_1, X_2, U) \sim N\left([\mu_1, \mu_2, 0]\,; \Sigma_{XU}\right)$.

Problem 2.1. Analyze the OLS estimators of β_1 from the full model

$$(*) : (X_1, X_2, U) \sim N\left(\mu_1, \mu_2, 0; \Sigma_{XU}\right)$$

Solution. Let $\hat{X}_1$ be the residuals of regressing X_1 on X_2, respectively. Then, regressing Y on $\hat{X}_1$ gives us

$$\begin{aligned}
\hat{\beta}_1 &= [\sum_{i=1}^{N}(\hat{X}_{1i}^2)]^{-1}[\sum_{i=1}^{N}(\hat{X}_{1i}Y_i)] \\
&= \beta_1 + [\sum_{i=1}^{N}(\hat{X}_{1i}^2)]^{-1}[\sum_{i=1}^{N}(\hat{X}_{1i}X_{2i})]\beta_2 + [\sum_{i=1}^{N}(\hat{X}_{1i}^2)]^{-1}[\sum_{i=1}^{N}(\hat{X}_{1i}U_i)] \\
&= \beta_1 + [\sum_{i=1}^{N}(\hat{X}_{1i}^2)]^{-1}[\sum_{i=1}^{N}(\hat{X}_{1i}U_i)]
\end{aligned}$$

where the middle term goes away since $\hat{X}_1$ is uncorrelated to X_2 by construction. Since $(X_1, X_2)U$, we have

$$E[\hat{\beta}_1] = \beta_1 + E[[\sum_{i=1}^{N}(\hat{X}_{1i}^2)]^{-1}[\sum_{i=1}^{N}(\hat{X}_{1i}U_i)]] = \beta_1$$

and our estimator is unbiased. By the Central Limit Theorem, noting U is independent of (X_1, X_2),

$$\sqrt{N}(\hat{\beta}_1 - \beta_1) \sim N(0, Var[U_i]Var[\hat{X}_{1i}]^{-1}).$$

Problem 2.2. Derive OLS estimators of β_1 derived from the following procedure:

1. Estimate (β_1, β_2) by OLS if $t_2 > 1.964$ where t_2 is the t-statistic fro OLS estimate of β_2 when X_1 and X_2 are both entered.
2. Estimate β_1 as the value of π from the regression $Y = X_1\pi_1 + V; \mathbb{E}\left[V\right] = 0$ if $t_2 < 1.964$.

This is called a *pretest estimator.* (Hint: See Bancroft, 1944).

Solution. For (a), the OLS estimator is the same as above. For (b) we have

$$\begin{aligned}\tilde{\beta}_1 &= (X_1'X_1)^{-1}(X_1'Y) \\ &= \beta_1 + (X_1'X_1)^{-1}(X_1'X_2)\beta_2 + (X_1'X_1)^{-1}(X_1'U)\end{aligned}$$

and

$$E[\tilde{\beta}_1] = \beta_1 + E[(X_1'X_1)^{-1}(X_1'X_2)]\beta_2$$

and so our estimate is biased unless X_2 is uncorrelated with either Y or X_1.

Our pretest estimator is

$$\hat{\beta}_{1,pre} = \hat{\beta}_1 1\{t_2 > 1.964\} + \tilde{\beta}_1 1\{t_2 \leq 1.964\}$$

Problem 2.3. What does this example say about the common practice of testing models using "specification tests"? How should we interpret the sampling distributions of estimates derived from pretest estimators?

Solution. This tells us that specification tests distorts our estimator of target parameters. Thus, if we were strict frequentists, we should adjust our hypothesis testing to test on $\hat{\beta}_{1,pre}$, when normally in research we choose $\hat{\beta}_1$ or $\tilde{\beta}_1$

The sampling distribution of estimates derived from pretest estimators is now a normal mixture distribution.

3 Methodology of Positive Economics (1953)

Read Friedman's "Methodology of Positive Economics" (1953) and especially footnote 11.

Problem 3.1. Define identification in general terms. Give an example.

Solution. Generally speaking, identification refers to determining models from data. It asks whether theoretical constructs have any empirical content in a hypothetical population or in real samples.

More formally, identification requires that given the observed distribution g of W (vector of observables), we have that (1) $\theta \in \Theta$ and (2) $g_\theta = g$ where $\theta \in \Theta$ is a model and $\pi(\theta)$ is our parameter of interest. In this case, the set of identified parameters of interest is

$$\Pi^* = \{\pi(\theta) : \theta \in \Theta^*\}$$

where Θ^* denotes the set of θ that satisfies the above two conditions.

One notable example is that of separating supply and demand curves based on on observations of price data. Philip Wright (1915), for example, pointed out that what appeared to be an upward sloping demand curve was actually a supply curve, traced out by a moving demand curve. If we were able to determine whether the observations are caused by shifts in supply curve vs. demand curve, this would be an example of a successful identification.

Problem 3.2. Compare the Cowles "structural approach" to the approach based on Friedman's methodology.

Solution. The Cowles approach trivializes the process of selecting one of many evidence-consistent hypotheses by restricting the choice set. Specifically, it endorses the division of step of consistent hypothesis selection into two steps:

1. Choose a model: select a class of admissible hypotheses;
2. Impose structure: choose one hypothesis from this class.

Friedman argues that this subdivision is inherently an arbitrary process, and in doing so the Cowles "structural approach" does not address all possible alternative models that may be consistent with the evidence. Instead, he argues that these two stages must be proceeded jointly, involving continued construction of new hypotheses and revision of old ones. Successful economics should present an inventory of findings and develop new hypotheses (models) in light of empirical rejections of old models or findings that suggest new models.

Problem 3.3. If you use Friedman's approach, how do you learn from data? Compare it with the Cowles approach. Can you use the same data to build a model and test it? Can this be done rigorously? Does data reuse not risk pretest bias? Compare Bayesian, classical, and abductive approaches to learning from data. How does each approach deal with surprise? (Something unanticipated.) Distinguish between numerical surprises from model surprises, i.e. fundamentally new phenomena.

Solution. See below:

1. Using Friedman's approach, we would use economic theory as an engine of economic analysis i.e. a device to generate predictions. Cosnequently, theories should be judged according to their ability to predict phenomena of interest. Naturally, this calls for repeated interactions between hypotheses and empirical results, augmenting the models to enhance our understanding of the data.

2. The Cowles approach, on the other hand, would use such economic theory to generate a set of hypotheses. It takes its classes of admissible models as determined before an empirical investigation begins.

   ▷ Friedman argues that this is subject to the identification problem since if one hypothesis is consistent with available evidence, an infinite number of hypotheses are, at which point the selection among alternatives must happen through some arbitrary principle such as Occam's razor.

3. Friedman encourages the researcher to interact with all of the available data and theory to learn and to augment it. Unlike Cowles, he views testing hypotheses as only a stage of an investigation, not its end. And to him, generating and testing new hypotheses in response to rejection of initial candidate hypotheses is a central feature of the research process.

4. To guard against the problem of using the same data, we can test provisional models on fresh data, possibly of a different character than the data used to formulate initial hypotheses. We can then draw new testable implications from hypotheses that survive an initial stage of scrutiny.

5. It helps to juxtapose classical, Bayesian and abductive approaches.

   ▷ Bayesians agree that learning from data is an integral part of Bayesian reasoning, but they are correct only to the extent that they describe learning about events that are a priori thought to be possible as formalized in some prior. This prior could be arrived at for whatever reason. Therefore, they do not have a way to cope with the totally unexpected surprises, since the priors rules out such cases.

   ▷ Abductive approaches allow the agent to learn from total surprises. One example may be to rely on their proximity to recombinations and extensions of past case histories and received theory. Heckman and Singer also suggest that agents revisit past events in light of empirical surprises to assemble interpretations of novel findings based on proximity to a reconfigured past.

   ▷ Note that classical approaches misses the part about learning from data. It is inherently slave to the model it starts out with, which obviates the need for learning from data. It ignores the recognition of benefit for knowledge of going back and forth with data.

Problem 3.4. Evaluate the following statement: *"Proper statistical inference requires that we specify hypotheses in advance of the dat." No data snooping allowed.*

Solution. The statement is contentious for several reasons:

1. The separation of hypotheses generation and testing, albeit analytically convenient, is artificial.

2. Successful empirical economics presents an inventory of findings and develops new hypotheses in light of empirical rejections of old ones.

   ▷ For example, literature on the consumption function offers a rich source of examples in which the abductive approach has been successful.

In short, the statement does not recognize the benefit of knowledge of going back and forth with data â learning from it, revising hypotheses in light of it, augmenting with fresh data and new theoretical insights.

4 Estimating ATE

Suppose you have a finite fixed size sample of a population of size N with some treated and some not. Treatment is randomly assigned. Each person has a (Y_0, Y_1) which is fixed. Only the sample of treatment assignments differ across individuals. They are determined by the toss of a coin. Let $X_i = 1$ is unit i assigned treatment. Unit i is assigned to control otherwise. You seek to estimate ATE by

$$\frac{1}{N}\sum_{i=1}^{N}(Y_{1i} - Y_{0i})$$

Problem 4.1. What is the standard error of the estimator

$$\frac{1}{N_1}\sum_{i=1}^{N} X_i Y_i - \frac{1}{N_0}\sum_{i=1}^{N}(1 - X_i) Y_i$$

where $N_1 = \sum_{i=1}^{N} X_0 N_0 = \sum_{i=1}^{N}(1 - X_i)$?

Solution. Rewrite the estimator as follows:

$$\frac{1}{N_1}\sum_{i=1}^{N} X_i Y_i - \frac{1}{N_0}\sum_{i=1}^{N}(1 - X_i) Y_i = \frac{1}{N}\left(\sum_{i=1}^{N}\frac{N}{N_1}X_i Y_{1i} - \frac{N}{N_0}(1 - X_i) Y_{0i}\right).$$

Define a new centered treatment indicator:

$$D_i = X_i - \frac{N_1}{N} = \begin{cases} \frac{N_0}{N} & \text{if } X_i = 1 \\ -\frac{N_1}{N} & \text{if } X_i = 0 \end{cases}$$

By construction the sample and population averages of D_i are both zero. Moreover, note that

$$V[D_i] = \mathbb{E}[D_i^2] - \mathbb{E}[D_i]^2 = \frac{N_1}{N}\frac{N_0^2}{N^2} + \frac{N_0}{N}\frac{N_1^2}{N^2} = \frac{N_1 N_0}{N^2}$$

Furthermore for $i \neq j$, the co variance between different D's is

$$P(D_i \cdot D_j = d) = \begin{cases} \frac{N_1(N_1-1)}{N(N-1)} & \text{if } d = \frac{N_0^2}{N^2} \\ 2\frac{N_1 N_0}{N(N-1)} & \text{if } d = \frac{N_0 N_1}{N^2} \\ \frac{N_0(N_0-1)}{N(N-1)} & \text{if } d = \frac{N_1^2}{N^2} \\ 0 & \text{else} \end{cases},$$

and so

$$\mathbb{E}[D_i D_j] = \begin{cases} \frac{N_1 N_0}{N^2} & \text{if } i = j \\ -\frac{N_1 N_0}{N^2(N-1)} & \text{if } i \neq j \end{cases}$$

Now rewriting the ATE estimator in terms of D_i:

$$\begin{aligned} \frac{1}{N}\left(\sum_{i=1}^{N}\frac{N}{N_1}X_i Y_{1i} - \frac{N}{N_0}(1 - X_i) Y_{0i}\right) &= \frac{1}{N}\left(\sum_{i=1}^{N}\frac{N}{N_1}\left(D_i + \frac{N_1}{N}\right) Y_{1i} - \frac{N}{N_0}\left(-D_i + \frac{N_0}{N}\right) Y_{0i}\right) \\ &= \frac{1}{N}\left(\sum_{i=1}^{N} Y_{1i} - Y_{0i}\right) + \frac{1}{N}\sum_{i=1}^{N} D_i\left(\frac{N}{N_1}Y_{1i} - \frac{N}{N_0}Y_{0i}\right) \end{aligned}$$

where the first difference is the ATE we seek to measure. Now note that the only variance in this previous equation is from D_i since the potential outcomes are fixed for each individual. Thus,

$$\begin{aligned}
V\left[\frac{1}{N}\left(\sum_{i=1}^{N}\frac{N}{N_1}X_iY_{1i}-\frac{N}{N_0}\left(1-X_i\right)Y_{0i}\right)\right] &= \frac{1}{N^2}V\left[\sum_{i=1}^{N}D_i\left(\frac{N}{N_1}Y_{1i}-\frac{N}{N_0}Y_{0i}\right)\right] \\
&= \frac{1}{N^2}\mathbb{E}\left[\left(\sum_{i=1}^{N}D_i\left(\frac{N}{N_1}Y_{1i}-\frac{N}{N_0}Y_{0i}\right)\right)^2\right], \text{ since } \mathbb{E}[D_i]=0 \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\left(D_i\left(\frac{N}{N_1}Y_{1i}-\frac{N}{N_0}Y_{0i}\right)\right)^2\right] \\
&\quad+\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j\neq i}\mathbb{E}\left[D_i\left(\frac{N}{N_1}Y_{1i}-\frac{N}{N_0}Y_{0i}\right)\cdot D_j\left(\frac{N}{N_1}Y_{1j}-\frac{N}{N_0}Y_{0j}\right)\right] \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}[D_i^2]\left(\left(\frac{N}{N_1}Y_{1i}-\frac{N}{N_0}Y_{0i}\right)\right)^2 \\
&\quad+\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j\neq i}\mathbb{E}[D_iD_j]\left(\frac{N}{N_1}Y_{1i}-\frac{N}{N_0}Y_{0i}\right)\cdot\left(\frac{N}{N_1}Y_{1j}-\frac{N}{N_0}Y_{0j}\right) \\
&= \frac{N_1N_0}{N^2}\frac{1}{N^2}\sum_{i=1}^{N}\left(\left(\frac{N}{N_1}Y_{1i}-\frac{N}{N_0}Y_{0i}\right)\right)^2 \\
&\quad-\frac{N_1N_0}{N^2(N-1)}\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j\neq i}\left(\frac{N}{N_1}Y_{1i}-\frac{N}{N_0}Y_{0i}\right)\cdot\left(\frac{N}{N_1}Y_{1j}-\frac{N}{N_0}Y_{0j}\right) \\
&= \frac{N_1N_0}{N^3(N-1)}\sum_{i=1}^{N}\left(\left(\frac{N}{N_1}Y_{1i}-\frac{N}{N_0}Y_{0i}\right)\right)^2 \\
&\quad-\frac{N_1N_0}{N^2(N-1)}\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{N}{N_1}Y_{1i}-\frac{N}{N_0}Y_{0i}\right)\cdot\left(\frac{N}{N_1}Y_{1j}-\frac{N}{N_0}Y_{0j}\right) \\
&= \frac{N_1N_0}{N^3(N-1)}\sum_{i=1}^{N}\left(\left(\frac{N}{N_1}Y_{1i}-\frac{N}{N_0}Y_{0i}\right)\right)^2 \\
&\quad-\frac{N_1N_0}{N^2(N-1)}\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{N}{N_1}Y_{1i}-\frac{N}{N_0}Y_{0i}\right)\cdot\left(\frac{N}{N_1}Y_{1j}-\frac{N}{N_0}Y_{0j}\right) \\
&= \frac{N_1N_0}{N^3(N-1)}\sum_{i=1}^{N}\left(\left(\frac{N}{N_1}Y_{1i}-\frac{N}{N_0}Y_{0i}\right)\right)^2 \\
&\quad-\frac{N_1N_0}{N^2(N-1)}\left(\frac{1}{N}\sum_{i=1}^{N}\left(\frac{N}{N_1}Y_{1i}-\frac{N}{N_0}Y_{0i}\right)\right)^2.
\end{aligned}$$

Now let

$$\overline{\Delta Y}=\frac{1}{N}\sum_{i=1}^{N}\left(\frac{N}{N_1}Y_{1i}-\frac{N}{N_0}Y_{0i}\right)=\frac{N}{N_1}\frac{1}{N_1}\sum_{i=1}^{N}Y_{1i}-\frac{N}{N_0}\frac{1}{N_0}\sum_{i=1}^{N}Y_{0i}=\frac{N}{N_1}\bar{Y}_1-\frac{N}{N_0}\bar{Y}_0.$$

Following the above set of inequalities, we then have

$$\begin{aligned}
V\left[\frac{1}{N}\left(\sum_{i=1}^{N}\frac{N}{N_1}X_iY_{1i}-\frac{N}{N_0}\left(1-X_i\right)Y_{0i}\right)\right] &= \frac{N_1N_0}{N^3(N-1)}\sum_{i=1}^{N}\left(\left(\frac{N}{N_1}Y_{1i}-\frac{N}{N_0}Y_{0i}\right)\right)^2-\frac{N_1N_0}{N^2(N-1)}\overline{\Delta Y}^2 \\
&= \frac{N_1N_0}{N^3(N-1)}\sum_{i=1}^{N}\left(\left(\frac{N}{N_1}Y_{1i}-\frac{N}{N_0}Y_{0i}\right)-\overline{\Delta Y}\right)^2 \\
&= \frac{N_1N_0}{N^3(N-1)}\sum_{i=1}^{N}\Bigg(\frac{N}{N_1}\Bigg((Y_{1i}-\bar{Y_1}\Bigg)-\frac{N}{N_0}\Bigg(Y_{0i}-\bar{Y_0}\Bigg)\Bigg)^2 \\
&= \frac{N_1N_0}{N^3(N-1)}\frac{N^2}{N_1^2}\sum_{i=1}^{N}\left(Y_{1i}-\bar{Y_1}\right)^2 \\
&\quad+\frac{N_1N_0}{N^3(N-1)}\frac{N^2}{N_0^2}\sum_{i=1}^{N}\left(Y_{0i}-\bar{Y_0}\right)^2 \\
&\quad+\frac{2N_1N_0}{N^3(N-1)}\frac{N^2}{N_1N_0}\sum_{i=1}^{N}\left(Y_{1i}-\bar{Y_1}\right)\left(Y_{0i}-\bar{Y_0}\right).
\end{aligned}$$

Now let

$$\begin{aligned}
S_1^2 &= \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_{1i}-\bar{Y_1}\right)^2 \\
S_0^2 &= \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_{0i}-\bar{Y_0}\right)^2 \\
S_\Delta^2 &= \frac{1}{N-1}\sum_{i=1}^{N}\left((Y_{1i}-\bar{Y_1})-(Y_{0i}-\bar{Y_0})\right)^2 \\
&= \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_{1i}-\bar{Y_1}\right)^2+\frac{1}{N-1}\sum_{i=1}^{N}\left(Y_{0i}-\bar{Y_0}\right)^2-\frac{2}{N-1}\sum_{i=1}^{N}\left(Y_{1i}-\bar{Y_1}\right)\left(Y_{0i}-\bar{Y_0}\right) \\
&= S_1^2+S_0^2-\frac{2}{N-1}\sum_{i=1}^{N}\left(Y_{1i}-\bar{Y_1}\right)\left(Y_{0i}-\bar{Y_0}\right).
\end{aligned}$$

Thus,

$$\begin{aligned}
V\left[\frac{1}{N}\left(\sum_{i=1}^{N}\frac{N}{N_1}X_iY_{1i}-\frac{N}{N_0}\left(1-X_i\right)Y_{0i}\right)\right] &= \frac{N_0}{NN_1}S_1^2+\frac{N_1}{NN_0}S_1^2+\frac{1}{N}(-S_\Delta^2+S_1^2+S_0^2) \\
&= \frac{1}{N_1}S_1^2+\frac{1}{N_0}S_0^2-\frac{1}{N}S_\Delta^2.
\end{aligned}$$

Hence the standard error is

$$\sqrt{\frac{1}{N_1}S_1^2+\frac{1}{N_0}S_0^2-\frac{1}{N}S_\Delta^2}.$$

Problem 4.2. Compare this sampling error with that from a model-based (super populative) approach. Compare the sources of variation in each approach.

Solution. Note that in population case, we have 2 sources of uncertainty: the random realization of the sample and the random assignment of treatment to individuals in the sample. In the finite sample case, we only have one source of uncertainty: the random assignment of treatment to individuals in the sample.

Deonte the true population (super-populative = sp) ATE:

$$\delta_{sp} = \frac{1}{N_{\text{sp}}} \sum_{i=1}^{N_{\text{sp}}} (Y_{1i} - Y_{0i}),$$

where N_{sp} is the total number of individuals in the population. The finite-sample (fs) ATE is

$$\delta_{\text{fs}} = \frac{1}{N} \sum_{i=1}^{N_{\text{sp}}} R_i \cdot (Y_{1i} - Y_{0i}),$$

where N is the finite sample size and R_i represents the dummy for if individual i is sampled. Clearly $\mathbb{E}[R_i] = N/N_{sp}$ since we have a random sample. Note that

$$\mathbb{E}[\delta_{\text{fs}}] = \mathbb{E}[\frac{1}{N} \sum_{i=1}^{N_{\text{sp}}} R_i \cdot (Y_{1i} - Y_{0i})] = \frac{1}{N} \sum_{i=1}^{N_{\text{sp}}} \mathbb{E}[R_i] \cdot (Y_{1i} - Y_{0i}) = \frac{1}{N} \sum_{i=1}^{N_{\text{sp}}} \frac{N}{N_{sp}} \cdot (Y_{1i} - Y_{0i}) = \delta_{sp}.$$

Thus, we have:

$$\begin{aligned}
V[\delta_{\text{fs}}] &= \frac{1}{N^2} \mathbb{E}\left[\left(\sum_{i=1}^{N_{\text{sp}}} R_i \cdot (Y_{1i} - Y_{0i}) - \delta_{sp} \right)^2 \right] \\
&= \frac{1}{N^2} \mathbb{E}\left[\left(\sum_{i=1}^{N_{\text{sp}}} \left(R_i - \frac{N}{N_{sp}} \right) \cdot (Y_{1i} - Y_{0i} - \delta_{sp}) \right)^2 \right], \text{ where we center } R_i \text{ as we did } D_i \text{ in part a (WLOG)} \\
&= \frac{1}{N^2} \sum_{i=1}^{N_{\text{sp}}} \mathbb{E}\left[\left(R_i - \frac{N}{N_{sp}} \right)^2 \right] \cdot (Y_{1i} - Y_{0i} - \delta_{sp})^2 \\
&\quad + \sum_{i=1}^{N_{\text{sp}}} \sum_{j \neq i} \mathbb{E}\left[\left(R_i - \frac{N}{N_{sp}} \right) \left(R_j - \frac{N}{N_{sp}} \right) \right] \cdot (Y_{1i} - Y_{0i} - \delta_{sp}) \cdot (Y_{1j} - Y_{0j} - \delta_{sp}).
\end{aligned}$$

Note that

$$\mathbb{E}\left[\left(R_i - \frac{N}{N_{sp}} \right)^2 \right] = \mathbb{E}[R_i^2] - \frac{N^2}{N_{sp}^2} = \frac{N}{N_{sp}} - \frac{N^2}{N_{sp}^2},$$

and

$$\begin{aligned}
\mathbb{E}\left[\left(R_i - \frac{N}{N_{sp}} \right) \left(R_j - \frac{N}{N_{sp}} \right) \right] &= \mathbb{E}\left[R_i - \frac{N}{N_{sp}} \right] \mathbb{E}\left[R_j - \frac{N}{N_{sp}} \right], \text{ by } R_i \perp R_j \text{, from random assignment} \\
&= \mathbb{E}\left[R_i - \frac{N}{N_{sp}} \right]^2 \\
&= 0.
\end{aligned}$$

Note that here, unlike in part a, since we assume N_{sp} is large relative to N we ignore the finite sample correction when calculating the cross-moment (i.e. we don't worry about the fact that given $R_i = 1$, technically the probability $Pr(R_j = 1) = (N-1)/(N_{sp}-1)$). Therefore,

$$V[\delta_{\text{fs}}] = \left(\frac{1}{NN_{sp}} - \frac{1}{N_{sp}^2}\right)\sum_{i=1}^{N_{\text{sp}}} \cdot (Y_{1i} - Y_{0i} - \delta_{sp})^2 = \frac{\sigma_\Delta^2}{N} - \frac{\sigma_\Delta^2}{N_{sp}},$$

where

$$\sigma_\Delta^2 = \mathbb{E}[((Y_{1i} - \bar{Y_1}) - (Y_{0i} - \bar{Y_0}))^2] = \mathbb{E}[S_\Delta^2],$$

where the last equality follows since the estimator we used in part a is unbiased. Analogously define,

$$\sigma_1^2 = \mathbb{E}[(Y_{1i} - \bar{Y_1})^2] = \mathbb{E}[S_1^2]$$
$$\sigma_0^2 = \mathbb{E}[(Y_{0i} - \bar{Y_0})^2] = \mathbb{E}[S_0^2].$$

Since we assume $N_{sp} >> N$, we have

$$V[\delta_{\text{fs}}] \approx \frac{\sigma_\Delta^2}{N}.$$

Now denote the estimator of interest:

$$\hat{\delta} = \frac{1}{N_1}\sum_{i=1}^{N_{\text{sp}}} R_i \cdot X_i \cdot Y_i - \frac{1}{N_0}\sum_{i=1}^{N_{\text{sp}}} R_i \cdot (1 - X_i) \cdot Y_i.$$

Note that the expectation of $\hat{\delta}$ **conditional on the realized sample is:**

$$\begin{aligned}
\mathbb{E}[\hat{\delta}|R] &= \frac{1}{N_1}\sum_{i=1}^{N_{sp}} R_i \mathbb{E}[X_i] Y_{1i} - \frac{1}{N_0}\sum_{i=1}^{N_{sp}} R_i \mathbb{E}[1 - X_i] Y_{0i} \\
&= \frac{1}{N_1}\sum_{i=1}^{N_{sp}} R_i \frac{N_1}{NY_{1i}} - \frac{1}{N_0}\sum_{i=1}^{N_{sp}} R_i \frac{N_0}{N} Y_{0i} \\
&= \delta_{fs}.
\end{aligned}$$

Thus, by LIE

$$\mathbb{E}[\hat{\delta}] = \mathbb{E}[\mathbb{E}[\hat{\delta}|R]] = \mathbb{E}[\delta_{fs}] = \delta_{sp}.$$

We now calculate the standard error

$$\begin{aligned}
V[\hat{\delta}] &= \mathbb{E}[V[\hat{\delta}|R]] + V[\mathbb{E}[\hat{\delta}|R]], \text{ by LIE} \\
&= \mathbb{E}\left[\frac{1}{N_1}S_1^2 + \frac{1}{N_0}S_0^2 - \frac{1}{N}S_\Delta^2\right] + V[\delta_{fs}], \text{ where the 1st term is from part a} \\
&\approx \frac{1}{N_1}\sigma_1^2 + \frac{1}{N_0}\sigma_0^2 - \frac{1}{N}\sigma_\Delta^2 + \frac{1}{N}\sigma_\Delta^2, \text{ using the approximation result derived above} \\
&= \frac{1}{N_1}\sigma_1^2 + \frac{1}{N_0}\sigma_0^2.
\end{aligned}$$

Thus the standard error is

$$\sqrt{\frac{1}{N_1}S_1^2 + \frac{1}{N_0}S_0^2}.$$

Note that this value is larger than that in part a, since we've added in the varaiance due to sample realization.

5 Defining Concepts

Define the following concepts:

- Random sample
- Truncated sample
- Stratified random sample
- Censored sample
- Censored random variable
- Choice based sample
- Design-based inference vs. model-based inference

Solution. Let $\omega\left(y,x\right)$ be a weighting function that alters the population density. Then define the density of sampled data $g\left(y^{\star},x^{\star}\right)$ as

$$g\left(y^{\star},x^{\star}\right)=\frac{\omega\left(y^{\star},x^{\star}\right)f\left(y^{\star},x^{\star}\right)}{\int\omega\left(y^{\star},x^{\star}\right)f\left(y^{\star},x^{\star}\right)dy^{\star}dx^{\star}}$$

Next define the indicator function

$$i\left(x,y\right)=\begin{cases}0 & \text{potential observation at } \left(y,x\right) \text{ cannot be sampled}\\ 1 & \text{otherwise}\end{cases}$$

Further, let $\Delta=1$ record the occurence of the event "a potential observation is sampled, i.e. the value of $\mathbf{y},\mathbf{x}$ is observed" and let $\Delta=0$ if it is not. We then define the proportion that is sampled as

$$\mathbb{P}\left(\Delta=1\right)=\int i\left(y,x\right)f\left(y,x\right)dydx$$

we can then use these definitions to define the terms above.

- Random sample: A randomly chosen subset of the population where the weighting is $\omega(y,x)=1$ and so the likelihood of choosing an observation for the sample is equal to its likelihood in the population.
- Truncated sample: A sample where the proportion of the population that can be sample is unknown and cannot be identified. That is, $P(\Delta=0)>0$ and $P(\Delta=1)$ is not known and cannot be identified.
- Stratified random sample: A stratified random sample is a random sample where $\omega(y,x)=\omega(x)$ and so its likelihood of being observed in sample only depends on x.
- Censored sample: A random sample where the proportion of the population that can be sample is known or can be identified. That is, $P(\Delta=0)>0$ but $P(\Delta=1)$ is known or can be identified.
- Censored random variable: Let Y_1 and Y_2 be two random variables. The random variable Y_1 is a censored random variable if whether we observe it or not depends on the realized value of Y_2.
- Choice based sample: A random sample of random variables (D,X) where selection is based only on D.

▷ Design-based inference vs model-based inference: These are two opposite views of causal inference. Design-based inference is causal inference through experimental designs and data. Model-based inference is causal inference through building a model that predicts outcomes.

6 Causal Parameter

Define "causal parameter." Also define "structural parameter." Is a structural parameter a causal parameter? Is a causal parameter a structural parameter?

Solution. Structural parameters are parameters of interest in a set of equations that stem from an underlying economic model which imposes structure on our measurement. These parameters are usually taken to be invariant, thereby enabling a rigorous counterfactual analysis. Causal parameters, on the other hand, are parameters of interest in a reduced-form setting, and we do not need a full model to identify them. MTE and ATE are examples of such causal parameters.

To illustrate with an example, consider a data of prices p_t and quantities q_t, both of which are chosen by the monopolist. Let c_t be the cost of the monopolist. Using a reduced-form approach, recognizing that both price and quantity seem to be associated with changes in cost, we can try to estimate the following system of equations:

$$\begin{aligned} q_t &= \alpha + \beta c_t + \epsilon_t \\ p_t &= \gamma + \delta c_t + \nu_t \end{aligned}$$

in which case $(\alpha, \beta, \gamma, \delta)$ are the causal parameters to be estimated. A structural model, on the other hand, would start with individual's utility maximization problem to derive the demand curve

$$p_t = \alpha + \beta q_t$$

and solve the monopolist's revenue-maximization problem

$$\max_{p_t, q_t} \sum_{t=0}^{\infty} \delta^t \left(p_t - c_t\right) q_t \left(p_t\right)$$

which would yield an equation of the following nature:

$$\begin{aligned} q_t &= \alpha^* + \beta^* c_t + \epsilon_t \\ p_t &= \gamma^* + \delta^* c_t + \nu_t \end{aligned}$$

in which case $(\alpha^*, \beta^*, \gamma^*, \delta^*)$ are the causal parameters to be estimated.

As we can see here, causal parameters would coincide with structural parameters only if the reduced form equations have a meaningful structural interpretation. Structural parameters are causal parameters, but causal parameters are not necessarily structural parameters.

7 Credibility Revolution

Discuss the credibility revolution in Econometrics (see Angrist and Pischke, 2010, JEP). What constitutes credibility? (See also companion paper by Keane.)

Solution. The credibility revolution is the movement in economics since the 1980s towards increasingly robust and valid empirical work. Edward Leaner pointed out in 1983 that “hardly anyone takes anyone else’s data seriously anymore" due to lack of robustness in empirical work to to changes in key assumptions. Thus, “credibility” here means that your empirical results are legitimate and convincing to others. Specifically, empirical work has become more crdible over the last 30 years due to:

1. Better and more data,

2. Less focus on esoteric econometric issues and more on justifying a causal interpretation of one’s results, and

3. Most importantly, better research designs. In particular, researchers now

   ▷ Focus more on identifying the specific sources of variation that yields a causal interpretation by using experimental and/or quasi-experimental methods.

8 Meta-analysis

Suppose you have a set of 10 studies on the effect of repeal of drug laws on addiction.

Problem 8.1. Define the properties of the vote counting method (i.e., what % of studies show a positive impact). What are its properties as the number of studies gets large?

Solution. Vote counting rule for M studies. If proportion of studies significant, i.e. the test statistic is greater than some cutoff ν, we pronounce "significant" and then proceed, e.g. $\nu = 0.6$.

$$\begin{aligned}
&Pr(\text{Proportion of studies significant} > \nu) \\
&= Pr(\frac{\#\text{studies sig.}}{M} > v) \\
&= 1 - \sum_{i=0}^{[\nu M]} \binom{M}{i} P^i (1-P)^{M-i}
\end{aligned}$$

where $[\nu M]$ is the greatest integer less than or equal to νM.

To analyze large samples, we use the binomial Central Limit Theorem. If Q studies yield significant results then

$$\frac{Q}{M} \sim N(P, \frac{P(1-P)}{M})$$

as $M \to \infty$ for fixed P, which is the power under the alternative hypothesis.

Vote counting probability of an effect in studies is

$$\mathbb{P}\left(\frac{Q}{M} > v\right) = \mathbb{P}\left(\underbrace{\frac{\frac{Q}{M} - P}{\left[\frac{P(1-P)}{M}\right]}}_{N(0,1)} > \frac{v-P}{\left[\frac{P(1-P)}{M}\right]^{\frac{1}{2}}}\right)$$

Case 1: $v > P > 0$

If $v > P > 0$ then

$$\lim_{M\to\infty} \frac{v-P}{\left[\frac{P(1-P)}{M}\right]^{\frac{1}{2}}} = \lim_{M\to\infty} \frac{(v-P)\, M^{\frac{1}{2}}}{(P(1-P))^{\frac{1}{2}}} \to \infty$$

and

$$\lim_{M\to\infty} \mathbb{P}\left(\frac{Q}{M} > v\right) = \mathbb{P}\left(\underbrace{\frac{\frac{Q}{M} - P}{\left[\frac{P(1-P)}{M}\right]^{\frac{1}{2}}}}_{N(0,1)} > \frac{v-P}{\left[\frac{P(1-P)}{M}\right]^{\frac{1}{2}}}\right) \to 0$$

If the power of the test for each study is low, we encounter serious problems. Power of the vote counting procedure goes to zero.

Case 2: $P > v$

Then we have that

$$\lim_{M\to\infty} \frac{v-P}{\left[\frac{P(1-P)}{M}\right]^{\frac{1}{2}}} \to -\infty$$

and then

$$\lim_{M\to\infty} \mathbb{P}\left(\frac{Q}{M} > v\right) \to 1$$

Problem 8.2. What would a meta-analyst report?

Solution. Let P_i be the p-value from study i. A meta-analyst would report

$$-2\log\left(\prod_{i=1}^{10} P_i\right) \sim \chi^2\left(20\right)$$

and reject the null if

$$P = -2\sum_{i=1}^{10} \log P_i \geq C$$

for some critical value C of the χ^2 distribution. Likewise, the analyst might report a weighted multiplicative sum of the P_i values, written aas

$$P_N = \prod_{i=1}^{10} P_i^{\nu_i}$$

where $-2\log P_N$ is a weighted average of χ^2 random variables.

Problem 8.3. Under what assumptions is the method of meta-analysis valid?

Solution. Assume equal competence of all investigators:

- ▷ M studies, same protocol for each study
- ▷ Independent studies (samples different and independent of each other)
- ▷ Summarizing the same types of data and studies