# Applied Microeconometrics Problem Set 2

Yixin Sun

November 10, 2020

## Problem 1

A model is falsifiable if there is a known function $\tau : \mathcal{G} \to \{0, 1\}$ s.t.

$$\tau(G) = 1 \quad \Rightarrow \quad \Theta^\star(G) = \emptyset \quad \text{(misspecified)}$$

and $\tau(G) = 1$ for at least one $G \in \mathcal{G}$

(a)

$$\Pi^*(G) = \{E_F[U] : F \in \mathcal{H}, G_F(y) = G(y)\}$$

This model is not falsifiable. Let $g = E_G[Y]$. If $G(0) > 0$, then $\Pi^*(G)$ is in the set $(-\infty, g]$. If $G(0) = 0$, then $\Pi^*(G) = g$. Neither of these sets are empty, so the model is not falsifiable.

(b)

$$\Pi^*(G) = \{E_F[U] : F \in \mathcal{H}, F(-1) = 0, F(2) = 1, G_F(y) = G(y)\}$$

This model is falsifiable because any distribution where $G(2) < 1$ would give $\Pi^*(G) = \emptyset$. So we have

$$\tau(G) = \mathbb{1}\{G(2) < 1\}$$

(c)

$$\Pi^*(G) = \left\{E_F[U] : F \in \mathcal{H}, F(0) = \frac{1}{2}, G_F(y) = G(y)\right\}$$

This model is falsifiable using the test

$$\tau(G) = \mathbb{1}\left\{G(0) \neq \frac{1}{2}\right\}$$

(d)

$$\Pi^*(G) = \{E_F[U] : F \in \mathcal{H}, E_F[U] = 0, G_F(y) = G(y)\}$$

For $E_F[U]$ to equal zero, it must be that either $U = 0$ for all $U$, or $F$ has support over both positive and negative numbers. Since it must be the case that $F(0) > 0$, the model is then falsifiable using the test

$$\tau(G) = \mathbb{1}\{G(0) = 0\}$$

(e)

$$\Pi^*(G) = \{\text{med}_F(U) : F \in \mathcal{H}, G_F(y) = G(y)\}$$

Similar to (a), let $g' = med(Y)$. If $G(0) = 0$ then $\Pi^*(G) = med(Y)$. If $G(0) > 0$, then $\Pi^*(G)$ is in the set $(-\infty, g']$. Neither of these sets are empty, so the model is not falsifiable.

## Problem 2

From the supplement, we have the following

$$\begin{aligned}
\hat{F} &\equiv n\frac{\hat{\pi}^2}{\hat{\sigma}_\pi^2} = \left(\frac{\sqrt{n}\hat{\pi}}{\hat{\sigma}_\pi}\right)^2 \\
\sqrt{n}\hat{\pi} &\equiv \frac{\sqrt{n}\frac{1}{n}\sum_{i=1}^n Z_i\left(Z_i\pi_n + V_i\right)}{\left(\frac{1}{n}\sum_{i=1}^n Z_i^2\right)} = \sqrt{n}\pi_n + \frac{\sqrt{n}\frac{1}{n}\sum_{i=1}^n Z_iV_i}{\left(\frac{1}{n}\sum_{i=1}^n Z_i^2\right)} \xrightarrow{d} \pi + \frac{R_{\text{fs}}}{E\left[Z^2\right]}
\end{aligned}$$

From the second line, we can see that an asymptotically unbiased alternative estimator needs to get rid of the $\frac{\sqrt{n}\frac{1}{n}\sum_{i=1}^n Z_iV_i}{\left(\frac{1}{n}\sum_{i=1}^n Z_i^2\right)}$ term. So we can create the new estimator:

$$\sqrt{n}\hat{\pi} - \frac{\sqrt{n}\frac{1}{n}\sum_{i=1}^n Z_iV_i}{\left(\frac{1}{n}\sum_{i=1}^n Z_i^2\right)}$$

Plugging this into our F-statistics:

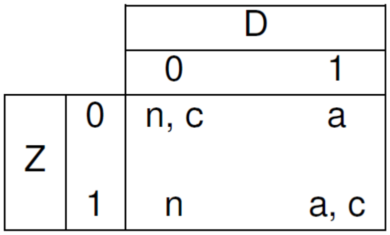
$$\begin{aligned}
\hat{F}^{unbiased} &\equiv \left(\frac{\sqrt{n}\hat{\pi} - \frac{\sqrt{n}\frac{1}{n}\sum_{i=1}^n Z_iV_i}{\left(\frac{1}{n}\sum_{i=1}^n Z_i^2\right)}}{\hat{\sigma}_\pi}\right)^2 \\
&= \left(\frac{\sqrt{n}\pi_n}{\hat{\sigma}_\pi}\right)^2 \\
&\xrightarrow{d} \left(\frac{\pi E\left[Z^2\right]^{1/2}}{\sigma_V}\right) \equiv \mu^2
\end{aligned}$$

## Problem 3

From the data, we observe the following:

$$\begin{aligned}
F_{zd}(y) &= P(Y \leq y \mid D = d, Z = 1) \\
p_a &= P(D = 1 \mid Z = 0) \\
p_n &= P(D = 0 \mid Z = 1) \\
p_c &= 1 - P(D = 1 \mid Z = 0) - P(D = 0 \mid Z = z)
\end{aligned}$$

By ruling out defiers through monotonicity, we know the following for different combinations of $Z$ and $D$ in the observed data:[1]



| | | D | |
|---|---|---|---|
| | | 0 | 1 |
| Z | 0 | n, c | a |
| | 1 | n | a, c |

[1] Table copied from Magne Mogstad's Empirical Analysis III slides

(a) From the information above, we can point identify the distribution of always-takers using the observed data,

$$F_1(y \mid a) = F_{01}(y) = P(Y \leq y \mid D = 1, Z = 0)$$

(b) Similarly, $F_0(y \mid n) = F_{01}(y) = P(Y \leq y \mid D = 0, Z = 1)$

(c) We know that the distribution of people where $Z = 1$ and $D = 1$ is a weighted average of always-takers and compliers,

$$P(Y \leq y \mid D = 1, Z = 1) = F_1(y \mid c)\frac{p_c}{p_c + p_a} + F_1(y \mid a)\frac{p_a}{p_c + p_a}$$

Rearranging the above equation and plugging in for observed data, we have

$$F_1(y \mid c) = \left[P(Y \leq y \mid D = 1, Z = 1) - P(Y \leq y \mid D = 1, Z = 0)\frac{p_a}{p_c + p_a}\right]\frac{p_c + p_a}{p_c}$$

Similarly, the distribution of people where $Z = 0$ and $D = 0$ is a weighted average of never-takers and compliers,

$$P(Y \leq y \mid D = 0, Z = 0) = F_0(y \mid c)\frac{p_c}{p_c + p_n} + F_0(y \mid n)\frac{p_n}{p_c + p_n}$$

So we can also point-identify

$$F_0(y \mid c) = \left[P(Y \leq y \mid D = 0, Z = 0) - P(Y \leq y \mid D = 0, Z = 1)\frac{p_n}{p_c + p_n}\right]\frac{p_c + p_n}{p_c}$$

(d) We first assume assume that $F_0(y \mid c)$ and $F_1(y \mid c)$ are both continuously and strictly between 0 and 1. We can write

$$\begin{aligned} F_1(y) &= P(Y(1) \leq y) \\ &= P(Y(1) \leq y|a)p_a + P(Y(1) \leq y|n)p_n + P(Y(1) \leq y|c)p_c \end{aligned}$$

The first and the third term we observe in the data, as shown from above (a) and (c). While we do not have $P(Y(1) \leq y|n)$, we can choose $y'$ such that

$$F_0(y|n) = F_0(y'|c)$$

where the assumption given in the problem gives us

$$F_1(y|n) = F_1(y'|c)$$

which is point identified, which gives us that $F_1(y)$ is point identified.

Similarly, we can solve for $F_0(y)$,

$$\begin{aligned} F_0(y) &= P(Y(0) \leq y) \\ &= P(Y(0) \leq y|a)p_a + P(Y(0) \leq y|n)p_n + P(Y(0) \leq y|c)p_c \\ &= P(Y(0) \leq y'|c)p_a + P(Y(0) \leq y|n)p_n + P(Y(0) \leq y|c)p_c \end{aligned}$$

(e) We can point identify $G$ by differentiating to get

$$\begin{aligned} G(y) &= P(Y(1) - Y(0) < y) \\ &= \int_{-\infty}^{\infty} f(x)\mathbb{1}\left\{F_1^{-1}(x) - y \leq F_0^{-1}(x)\right\} dx \end{aligned}$$

which is point identified from (d)

## Problem 4

(a) In this context, the monotonicity condition states that having two children of the same sex in a family pushes the family towards having a third child, and never makes the family more likely to stop having children. This monotonicity condition does not seem credible, since it is plausible that families may be biased towards wanting only boys or only girls, which means those families are more likely to stop having children after two kids.

(b) With the twins instrument, the monotonicity condition states that if your second child turns out to be a set of twins, then the family is pushed towards have a third child. Never-takers are not possible in this case, since it is not possible to choose to only have one child if you are pregnant with twins. The monotonicity condition here seems valid.

(c) With multiple instruments, the monotonicity assumption needs to rule out defiers for both instruments. Since we cannot rule out defiers in (a), the monotonicity condition does not hold when using both instruments.

## Problem 5

(a) The monotonicity condition on slide 10 states

$$P(D(1) \geq D(0) \mid X) = 1$$

So the monotonicity can flip between different X values. The condition given in the problem set is

$$P(D(1) \geq D(0) \mid X = k) = 1, \forall k = 1, ..., K$$

so $D(1) \geq D(0)$ must hold for all $X$ groups.

(b) From the regression setup for the potential outcomes model, we can write

$$\beta_{tsls} = \frac{Cov(Y, Z)}{Cov(D, Z)}$$

Applying the usual LATE interpretation, we can write

$$E[Y(1) - Y(0) \mid D(1) = 1, D(0) = 0, X = k] = \frac{\text{Cov}(Y, Z \mid X = k)}{\text{Cov}(D, Z \mid X = k)}$$

Using these two facts, we work backwards

$$\begin{aligned}
&E\left[\frac{\text{Cov}(D, Z \mid X)}{E[\text{Cov}(D, Z \mid X)]} E[Y(1) - Y(0) \mid D(1) = 1, D(0) = 0, X]\right] \\
&= E\left[\frac{\text{Cov}(D, Z \mid X)\,\text{Cov}(Y, Z \mid X)}{E[\text{Cov}(D, Z \mid X)]\,\text{Cov}(D, Z \mid X)}\right] \\
&= E\left[\frac{\text{Cov}(Y, Z \mid X)}{E[\text{Cov}(D, Z \mid X)]}\right] \\
&= \frac{E[\text{Cov}(Y, Z \mid X)]}{E[\text{Cov}(D, Z \mid X)]} \\
&= \frac{\text{Cov}(Y, Z)}{\text{Cov}(D, Z)} \\
&= \beta_{tsls}
\end{aligned}$$

(c) $\beta_{tsls}$ is a weighted average of $LATE(X)$. The weights here are the covariance between the treatment and the instrument, or how strong the instrument is, conditional on $X$. The intuition here is that we give more weight to people who are more likely to be pushed into treatment by the instrument, or the complier group.

## Problem 6

From the set up, we know that OLS will be biased, and we only need $Z_1$ to properly instrument for $X$. The results below are consistent with what we would expect. OLS is very biased, and its 95% confidence interval never covers the true value of the coefficient, no matter the sample size. Both the TSLS and Jackknife with one instrument does very well, producing estimates with little bias and a robust confidence interval coverage, even in small samples.

Comparing the TSLS to the Jackknife estimates in the many instrument case is where we see the Jackknife procedure realy shine. As discussed in class, TSLS with a many instrument problem does poorly in small samples. We see this at work here, where the TSLS with many instruments does poorly, although improving with the sample size. On the otherhand, Jackknife with many instruments does considerably better, and has a coverage level that is close to the one instrument approaches as the sample gets larger.

The intuition here is that the first stage fitted values, $\hat{\dot{X}}_i \equiv \hat{\Pi} Z_i \equiv \left(X'Z\right)\left(Z'Z\right)^{-1} Z$, are still correlated with the error, $U$. While this correlation goes away in large samples, it increases with the number of instruments for a fixed sample, and biases the TSLS coefficient towards OLS (which we observe to be true in our results). The jackknife approach breaks this correlation by essentially running the first stage regression $N$ times, estimating $\hat{\Pi}_i$ without observation $i$.

Table 1: Monte Carlo Results

| Value | N = 100 | N = 200 | N = 400 | N = 800 |
|---|---|---|---|---|
| **OLS** | | | | |
| Median | 1.589 | 1.591 | 1.586 | 1.587 |
| Bias | 0.587 | 0.589 | 0.586 | 0.587 |
| SD | 0.066 | 0.044 | 0.031 | 0.023 |
| Coverage | 0.000 | 0.000 | 0.000 | 0.000 |
| **TSLS - 1 Instrument** | | | | |
| Median | 0.998 | 0.994 | 0.999 | 1.006 |
| Bias | -0.025 | -0.016 | -0.005 | 0.002 |
| SD | 0.183 | 0.131 | 0.082 | 0.057 |
| Coverage | 0.950 | 0.928 | 0.952 | 0.960 |
| **TSLS - Many Instruments** | | | | |
| Median | 1.276 | 1.172 | 1.094 | 1.055 |
| Bias | 0.272 | 0.167 | 0.094 | 0.053 |
| SD | 0.113 | 0.093 | 0.071 | 0.053 |
| Coverage | 0.334 | 0.526 | 0.684 | 0.804 |
| **Jackknife - 1 Instrument** | | | | |
| Median | 0.952 | 0.969 | 0.988 | 1.001 |
| Bias | -0.086 | -0.042 | -0.016 | -0.003 |
| SD | 0.224 | 0.142 | 0.085 | 0.058 |
| Coverage | 0.960 | 0.930 | 0.954 | 0.950 |
| **Jackknife - Many Instruments** | | | | |
| Median | 0.960 | 0.976 | 0.992 | 1.001 |
| Bias | 0.280 | -0.045 | -0.014 | -0.003 |
| SD | 19.219 | 0.172 | 0.097 | 0.063 |
| Coverage | 0.660 | 0.822 | 0.906 | 0.940 |

This table reports the medians, SD, and average bias of Monte Carlo simulations. Coverage is the fraction of time the 95% confidence interval contains our true coefficient value, 1.

# Problem 7

## 1. Reproduce Table 2

The first three columns below match almost exactly the table in Abadie (2003). The last column does not match as well, although all of the coefficients are quite close and within the implied 95% confidence interval of the Abadie (2003) estimates.

Table 2: Replication

| | | Two stage least squares | | |
|---|---|---|---|---|
| | | Endogenous Treatment | | |
| Var | OLS | First Stage | Second Stage | Least squares treated |
| Participation in 401(k) | 13,527.05 | | 9,418.83 | 11,036.28 |
| | (1,809.59) | | (2,094.31) | (2,162.54) |
| Constant | -23,549.00 | -0.0306 | -23,298.74 | -24,823.01 |
| | (2,177.26) | (0.0087) | (2,163.10) | (2,450.73) |
| Family Income (thousand $) | 976.93 | 0.0013 | 997.19 | 991.58 |
| | (83.34) | (0.0001) | (82.98) | (86.36) |
| Age (minus 25) | -376.17 | -0.0022 | -345.95 | -7.74 |
| | (236.89) | (0.0010) | (236.86) | (272.10) |
| Age (minus 25) sq. | 38.70 | 0.0001 | 37.85 | 29.68 |
| | (7.66) | (0.0000) | (7.67) | (8.71) |
| Married | -8,369.47 | -0.0005 | -8,355.87 | -7,851.63 |
| | (1,829.24) | (0.0079) | (1,829.17) | (2,127.42) |
| Family Size | -785.65 | 0.0001 | -818.96 | -823.66 |
| | (410.62) | (0.0024) | (409.93) | (492.91) |
| Eligibility for 401(k) | | 0.6883 | | |
| | | (0.0080) | | |

This table replicates Table 2 from Abadie (2003). (1) presents OLS estimates. (2) and (3) present the first and second stage estimates from 2SLS. (4) presents the results of LARF estimator using logit to estimate $\tau(x_i)$. Standard errors are heteroskedasticity robust.

## 2. Anderson-Rubin confidence interval

The first row shows the 95% Anderson-Rubin confidence interval, estimated by calculating the AR statistic for a grid of 1,000 $\bar{\beta}$ values. The second row is a confidence interval obtained using standard asymptotic approximations for TSLS. The confidence intervals are similar, but the AR interval is a bit wider, which is consistent with the fact that we are not in a weak instrument setting here, and that the AR test sacrifices power in its construction.

Table 3: Confidence Intervals

| Name | Lower Bound | Upper Bound |
|---|---:|---:|
| Anderson-Rubin | 6,340.75 | 12,496.91 |
| TSLS | 5,313.98 | 13,523.68 |

## 3. Jackknife Estimates

The Jackknife estimates are roughly the same as the TSLS estimates, which makes sense since we do not have a many instruments problem here, so the two methods are first-order equivalent.

Table 4: Jackknife IV

| Var | Second Stage | Jackknife |
|---|---|---|
| Participation in 401(k) | 9,418.83 | 9,418.53 |
| | (2,094.31) | (1,858.56) |
| Constant | -23,298.74 | -23,298.72 |
| | (2,163.10) | (1,995.94) |
| Family Income (thousand $) | 997.19 | 997.19 |
| | (82.98) | (28.91) |
| Age (minus 25) | -345.95 | -345.95 |
| | (236.86) | (217.91) |
| Age (minus 25) sq. | 37.85 | 37.85 |
| | (7.67) | (5.78) |
| Married | -8,355.87 | -8,355.87 |
| | (1,829.17) | (1,639.99) |
| Family Size | -818.96 | -818.96 |
| | (409.93) | (497.41) |

(1) reports the estimates from the second stage of 2SLS, same as (3) from the previous table. (2) reports estimates from a jackknife IV estimator with bootstrapped standard errors.