E31720 Problem Set 1

Franco Calle
(incidental discussions with Hugo Lopez, Phillip Monagan and Ed Jee)

University of Chicago

October 19, 2021

Problem 1

First we state the two models and our condition:

$$\begin{aligned}\text{Eq1: } & Y = \alpha_1 + \beta D + \epsilon_1 \\ \text{Eq2: } & Y = \alpha_1 + \gamma D + \omega X + \epsilon_2\end{aligned}$$

Potential outcomes:

$$\begin{aligned}Y &= DY(1) + (1-D)Y(0) \\ &= Y(0) + D(Y(1) - Y(0)) \\ &= Y(0) + \alpha D \end{aligned} \tag{1}$$

We can use FWL to partial out the both equations and leave them in terms of D. Our equations will be:

$$\begin{aligned}\text{Eq1: } & \hat{Y}_1 = \beta \hat{D}_1 + \hat{\epsilon}_1 \\ \text{Eq2: } & \hat{Y}_2 = \gamma \hat{D}_1 + \hat{\epsilon}_2\end{aligned}$$

Where:

$$\begin{aligned}\hat{Y}_1 &= Y - \text{BLP}(Y|1) \\ \hat{Y}_2 &= Y - \text{BLP}(Y|1, X) \\ \hat{D}_1 &= D - \text{BLP}(D|1) \\ \hat{D}_2 &= D - \text{BLP}(D|1, X)\end{aligned}$$

And:

$$\begin{aligned}\text{BLP}(Y|1) &= E[Y] \\ \text{BLP}(Y|1, X) &= E[Y] + \frac{Cov(Y, X)}{Var(X)} X \\ \text{BLP}(D|1) &= E[D] \\ \text{BLP}(D|1, X) &= E[D] + \frac{Cov(D, X)}{Var(X)} X\end{aligned}$$

Then, we can express both β and γ as covariances. Let's start first with β

$$
\begin{aligned}
\beta &= \frac{Cov(\hat{Y}_1, \hat{D}_1)}{Var(\hat{D}_1)} \\
&= \frac{Cov(Y, \hat{D}_1)}{Var(\hat{D}_1)} && \text{Remove cov } \hat{D}_1 \text{ on BLP(Y|1,X)} \\
&= \frac{Cov(Y(0) + \alpha D, \hat{D}_1)}{Var(\hat{D}_1)} && \text{Def. of potential outcomes} \\
&= \frac{Cov(Y(0) + \alpha D, \hat{D}_1)}{Var(D)} && \text{Since } Var(\hat{D}_1) = Var(D) \\
&= \frac{Cov(Y(0) + \alpha D, D - E[D])}{Var(D)} \\
&= \frac{Cov(Y(0), D - E[D])}{Var(D)} + \alpha \frac{Var(D)}{Var(D)} && \text{Since } Cov(D, \hat{D}_1) = Var(D) \\
&= \alpha + \frac{Cov(Y(0), D)}{Var(D)} && \text{Rearranging and simplifying}
\end{aligned}
$$

Now Let's check the case of γ.

$$
\begin{aligned}
\gamma &= \frac{Cov(\hat{Y}_2, \hat{D}_2)}{Var(\hat{D}_2)} \\
&= \frac{Cov(Y, \hat{D}_2)}{Var(\hat{D}_2)} \\
&= \frac{Cov(Y(0) + \alpha D, D - E[D] - \frac{Cov(D,X)}{Var(X)}X)}{Var(\hat{D}_2)} \\
&= \frac{Cov(Y(0) + \alpha D, D - \frac{Cov(D,X)}{Var(X)}X)}{Var(\hat{D}_2)} \\
&= \left[Cov(Y(0), D - \frac{Cov(D,X)}{Var(X)}X) + \alpha Cov(D, D - \frac{Cov(D,X)}{Var(X)}X) \right] \frac{1}{Var(\hat{D}_2)} \\
&= \left[Cov(Y(0), D - \frac{Cov(D,X)}{Var(X)}X) + \alpha \left(Var(D) - \frac{Cov(D,X)^2}{Var(X)} \right) \right] \frac{1}{Var(D) - \frac{Cov(D,X)^2}{Var(X)}} \\
&= \alpha + \left[Cov(Y(0), D - \frac{Cov(D,X)}{Var(X)}X) \right] \left[Var(D) - \frac{Cov(D,X)^2}{Var(X)} \right]^{-1} \\
&= \alpha + \left[Cov(Y(0), D) - \frac{Cov(D,X)}{Var(X)} Cov(Y(0), X) \right] \left[Var(D) - \frac{Cov(D,X)^2}{Var(X)} \right]^{-1}
\end{aligned}
$$

Now compute $|\alpha - \beta|$ and $|\alpha - \gamma|$:

$$|\alpha - \beta| = |\frac{Cov(Y(0), D)}{Var(D)}|$$

$$|\alpha - \gamma| = |\left[Cov(Y(0), D) - \frac{Cov(D, X)}{Var(X)} Cov(Y(0), X)\right] \left[Var(D) - \frac{Cov(D, X)^2}{Var(X)}\right]^{-1}|$$

Part A

Is it true that $|\alpha - \gamma| \leq |\alpha - \beta|$? If not, find counterexample.

This is not necessarily true, for instance, assume Cov(X, D) < 0, Cov(Y(0), X) > 0. Then We get that the numerator of $|\alpha - \gamma|$ become greater than $Cov(Y(0), D)$ and denominator becomes smaller than $Var(D)$ which in turn increases $|\alpha - \gamma|$ even further. In this case we have that $|\alpha - \gamma| > |\alpha - \beta|$

Part B

Suppose D and X are uncorrelated. Does this change answer to (a)?

Yes, if $D \perp X$ then Cov(D,X) $= 0$ which basically gives us that $|\alpha - \gamma| = |\alpha - \beta|$

Part C

Suppose that X is uncorrelated with Y(0) and Y(1). Does this change the answer to a?

It would be modified slightly. In this case we get that both numerators are equal, the only difference would be in the denominators. Then, we will have the following condition:

$$|\alpha - \beta| \leq |\alpha - \gamma| \iff |\frac{Cov(Y(0), D)}{Var(D)}| \leq |\frac{Cov(Y(0), D)}{Var(D) - \frac{Cov(D,X)^2}{Var(X)}}|$$

First case: equality can happen if Cov(D, X) $= 0$, then both expressions are equivalent.
Second case: $|\alpha - \beta| < |\alpha - \gamma|$ when $Cov(D, X) \neq 0$. This is because denominator on the second expression can only be smaller than the denominator in the first one. Recall that Var($\hat{D}_2$) is bounded below by zero, and $Cov(D, X)^2$ is squared so the denominator can only be lower in absolute value.

Part D

Suppose that $E[Y(0)|D = d, X = x] = E[Y(0)|X = x]$ does this change answer in (a)?

Yes, now the inequality will depend on the difference in weights conditional on D. First, looking at the condition, we know that $E[Y(0)|D = 1, X = x] = E[Y(0)|D = 0, X = x] = E[Y(0)|X = x]$. The only place where this would affect is Cov(Y(0), D).

Recall that $Cov(X, D) = p(1-p)\,[E[X|D = 1] - E[X|D = 0]]$, where p = E[D] = Pr(D = 1). Then we can use this identity for Y(0) and D:

$$
\begin{aligned}
Cov(Y(0), D) &= p(1-p)\left[E[Y(0)|D=1] - E[Y(0)|D=0]\right] \\
&= p(1-p)\left\{E\left[E[Y(0)|D=1,X]|D=1\right] - E\left[E[Y(0)|D=0,X]|D=0\right]\right\} \\
&= p(1-p)\left\{E\left[E[Y(0)|X]|D=1\right] - E\left[E[Y(0)|X]|D=0\right]\right\} \\
&= p(1-p)\sum_{x\in X} E[Y(0)|X=x](Pr(X=x|D=1) - Pr(X=x|D=0))
\end{aligned}
$$

As can be seen, now we will have to consider the density of X conditional on D = 1 and D = 0. There are two cases:

- The weighted average can be negative or positive, if it is either negative or positive, then we go again to the case in (a) where any inequality can happen.
- If Pr(X = x | D = 1) = Pr(X = x | D = 0) for all x, then Cov(Y(0), D) = 0 which implies $|\alpha - \beta| = 0$ and $|\alpha - \gamma|$ will be greater in almost all cases with the exception of the case where $D \perp X$ where both absolute values are equal to 0.

Part E

Suppose E[Y(0) | X = x] is a linear function of x. Does this change answer to a)?

No it does not, the only thing it does is that $\frac{Cov(Y(0),X|X=x)}{Var(X|X=x)}$ is a constant for all values of X. We can replace this into the the second component of the numerator of $|\alpha - \gamma|$ and we will get that the bias will be equal for all bins of X, or in other words, conclusion in (a) holds for any $x \in X$.

Problem 2

Consider Binary Treatment Model:

$$Y = DY(1) + (1 - D)Y(0)$$

With $D \in \{0, 1\}$. and suppose X is vector of observable variables such that $(Y(0), Y(1)) \perp D|X$. Let $p(x) = P[D = 1|X = x]$ and $P = P[D = 1|X]$ denote the propensity score.

Part A

Show that $D \perp X|P$:

Proof. The strategy here is to find that the distribution of D conditional on P is equal to the distribution of D conditional on X. This implies that D is independent of X conditional on P. First let's define P(D = 1|X) as P since we will use it here a lot.

$$\begin{aligned} P(D = 1|P) &= E[D|P] \\ &= E[E[D|P, X]|P] \\ &= E[E[D|X]|P] \\ &= E[P(D = 1|X)|P] \\ &= E[P|P] \\ &= P(D = 1|X) \\ &= P \end{aligned}$$

So the first line used the expectation counterpart of the probability, conditional on P. Second line used LIE. Third line eliminated P from the conditioning since P is function of X therefore it is redundant. Fourth we replaced the expectation counterpart to the probability one. Fifth replaced the expresion by P and the sixth eliminates the expectation since conditioning on P gives us exact values of P, which is the expression we wanted.

□

Part B

Let b be any function of X for which $D \perp X|b(X)$ show that $D \perp (Y(0), Y(1))|b(X)$. Discuss the intuition behind the result.

Proof. We need to show:

$$
\begin{aligned}
P[D=1|Y(0),Y(1),b(X)] &= E[P[D=1|Y(1),Y(0),X,b(X)]|Y(0),Y(1),b(X)] \\
&= E[P[D=1|Y(1),Y(0),X]|Y(0),Y(1),b(X)] \\
&= E[P[D=1|X]|Y(0),Y(1),b(X)] \\
&= E[P|Y(0),Y(1),b(X)] \\
&= P[D=1|X] \\
&= P[D=1|P]
\end{aligned}
$$

This implies that we can condition on b(X) instead of X. The first line uses LIE; the second eliminates b(X) since it is redundant; the third uses the fact that $D \perp (Y(0), Y(1))|X$ which is given by the problem; the fourth just uses de definition of P; the fifth uses conclusion from part A to get the desired equality.

The intuition for this result is that as long as we have a function b(X) which comprises information of X so that $D \perp D|b(X)$, then we can use that function and it might not be the propensity score. The potential relevance for implementation is that it gives us more flexibility in terms of the functional form to measure the score as long as the orthogonality property holds. There could be some functions that make the dimension reduction faster than common probabilistic methods (probit/logit).

□

Part C

Continue assuming that b is function of X for which $D \perp X|b(X)$. Show that propensity score, p, can always be written as function of b. That is, show that there exists a well-defined function f such that p(x) = f(b(x)) for every x (almost every x, if you want to be precise) in the support of X.

Proof. Consider p(x) = f(b(x)) for almost every x, then p(x) maps x to a number of elements equal or lower than b(x). This as a consequence implies that b(x) has more granular information than p(x), or as Rosenbaum and Rubin (1983) say, P(X) is the 'coarsest' function of X such that the conditional independence assumption holds. So if we condition on b(x) that informs us more than p(x).

$$
\begin{aligned}
P[D=1|Y(0),Y(1),p(x)] &= E[P[D=1|Y(1),Y(0),X,b(X),p(X)]|Y(0),Y(1),p(X)] \\
&= E[P[D=1|Y(1),Y(0),X,b(X)]|Y(0),Y(1),p(X)] \\
&= E[P[D=1|Y(1),Y(0),X]|Y(0),Y(1),p(X)] \\
&= E[P[D=1|X]|Y(0),Y(1),p(X)] \\
&= E[P|Y(0),Y(1),p(X)] \\
&= P[D=1|X]
\end{aligned}
$$

□

In the previous equations, the first line uses LIE. The second removes p(X) since b(X) has more information than p(X), the third removes b(X) because the very X has more information, finally we remove Y(0) and Y(1) because conditional on X they do not provide any information as shown in B.

Problem 3

Consider the binary treatment potential outcomes model:

$$Y = DY(1) + (1 - D)Y(0)$$

Maintain selection on observables assumption that $(Y(0), Y(1)) \perp D|X$. Let $\tilde{p}(x)$ and $\tilde{\mu}_1(x)$ be two functions of x and consider the quantity.

$$\beta_1 = E\left[\frac{DY}{\tilde{p}(X)} - \frac{(D - \tilde{p}(X))}{\tilde{p}(X)}\tilde{\mu}_1(x)\right]$$

Part A

Show that if $\tilde{p}(x) = \mathbb{P}[D = 1|X = x]$ for all x, then $\beta_1 = E[Y(1)]$.

Proof. Recall from 3.c we already proved that Pr[D = 1 | Y(0), Y(1), p(X)] = P[D=1 | X], we will use it in our LIE:

$$\begin{aligned}
\beta_1 &= \mathrm{E}\left[\frac{DY}{\tilde{p}(X)} - \left(\frac{D - \tilde{p}(X)}{\tilde{p}(X)}\right)\tilde{\mu}_1(x)\right] \\
&= \mathrm{E}\left[\mathrm{E}\left[\frac{DY}{\tilde{p}(X)} - \left(\frac{D - \tilde{p}(X)}{\tilde{p}(X)}\right)\tilde{\mu}_1(X) \;\middle|\; \tilde{p}(X), X\right]\right] \\
&= \mathrm{E}\left[\frac{\mathrm{E}[DY \mid \tilde{p}(X), X]}{\tilde{p}(X)} - \left(\frac{\mathrm{E}[D \mid \tilde{p}(X), X] - \tilde{p}(X)}{\tilde{p}(X)}\right)\tilde{\mu}_1(X)\right] \\
&= \mathrm{E}\left[\frac{\mathrm{E}[DY \mid X]}{\tilde{p}(X)} - \left(\frac{\mathrm{E}[D \mid X] - \tilde{p}(X)}{\tilde{p}(X)}\right)\tilde{\mu}_1(X)\right] \\
&= \mathrm{E}\left[\frac{\mathrm{E}[DY \mid X]}{\tilde{p}(X)} - \left(\frac{\tilde{p}(X) - \tilde{p}(X)}{\tilde{p}(X)}\right)\tilde{\mu}_1(X)\right] \\
&= \mathrm{E}\left[\frac{\mathrm{E}[DY \mid X]}{\tilde{p}(X)}\right] \\
&= \mathrm{E}\left[\frac{\tilde{p}(X)\,\mathrm{E}[Y \times 1 \mid D = 1, X] + (1 - \tilde{p}(X))\,\mathrm{E}[Y \times 0 \mid D = 0, X]}{\tilde{p}(X)}\right] \\
&= \mathrm{E}\left[\frac{\tilde{p}(X)\,\mathrm{E}[Y \mid D = 1, X]}{\tilde{p}(X)}\right] \\
&= \mathrm{E}[\mathrm{E}[Y(1) \mid X]] \\
&= \mathrm{E}[Y(1)]
\end{aligned}$$

So the first line is just the definition of β_1. The second line applies LIE over the expectation conditional on X and p(X). The third line inserts the expectation over DY and D taking advantage that p(X) and $\mu(X)$ information is already contained over the condition. The fourth line removes $\tilde{p}(x)$ from the conditioning set since it is redundant, we are already conditioning on X. The fifth

just changes E[D | X] to its probability counterpart. The sixth eliminates the second term in the RHS. The seventh line applies law of total probability over events D and takes advantage of the selection on observables assumption since we are already conditioning on X. The eight eliminates the component that conditions on D=0. The ninth simplifies $\tilde{p}(X)$. And the tenth integrates over X's (or uses LIE). □

Part B

Show that if $\tilde{\mu}_1(x) = \mathrm{E}[Y(1) \mid X = x]$ for all x, then $\beta_1 = E[Y(1)]$.

Proof. Now we are assuming $\tilde{\mu}_1(x) = \mathrm{E}[Y(1) \mid X = x]$ and nothing about $\tilde{p}(X)$. So let's start from the definition of β_1.

$$
\begin{aligned}
\beta_1 &= \mathrm{E}\left[\frac{DY}{\tilde{p}(X)} - \left(\frac{D - \tilde{p}(X)}{\tilde{p}(X)}\right)\tilde{\mu}_1(x)\right] \\
&= \mathrm{E}\left[\mathrm{E}\left[\frac{DY}{\tilde{p}(X)} - \left(\frac{D - \tilde{p}(X)}{\tilde{p}(X)}\right)\mathrm{E}[Y(1) \mid X] \;\middle|\; X\right]\right] \\
&= \mathrm{E}\left[\frac{\mathrm{E}[DY \mid X]}{\tilde{p}(X)} - \mathrm{E}[Y(1) \mid X]\left(\frac{\mathrm{E}[D \mid X] - \tilde{p}(X)}{\tilde{p}(X)}\right)\right] \\
&= \mathrm{E}\left[\frac{P[D=1|X]\,\mathrm{E}[Y(1) \mid X]}{\tilde{p}(X)} - \left(\frac{\mathrm{E}[Y(1) \mid X]P[D=1|X] - \mathrm{E}[Y(1) \mid X]\tilde{p}(X)}{\tilde{p}(X)}\right)\right] \\
&= \mathrm{E}\left[\frac{\mathrm{E}[Y(1) \mid X]\tilde{p}(X)}{\tilde{p}(X)}\right] \\
&= E[Y(1)]
\end{aligned}
$$

Here we start with the definition of β, then apply LIE to get an expression conditional on X. The third equality moves out E[Y(1)|X] which is a function of X, and introduces the conditional expectation over the variables which are not function of X. The fourth line applies law of total probability on the first component and eliminates the part that conditions on D=0, while in the second component we change the expectation of D for its probability counterpart. The fifth equality just subtracts the first component within the expectation with the first component of the second fraction. And finally the sixth equality applies LIE to back out the unconditional expectation which is the result we were expecting.

□

Part C

Derive an expression β_0 that is analogous to β_1 and has the property that $\beta_0 = E[Y(0)]$ if either $\tilde{p}(x) = P[D = 1|X = x]$ for all x, or $\tilde{\mu}_0(x) = E[Y(0)|X = x]$ for all x.

Proof. An analogous counterpart for β_0 is as follows:

$$\beta_0 = \mathrm{E}\left[\frac{(1-D)Y}{1-\tilde{p}(X)} + \frac{D-\tilde{p}(X)}{1-\tilde{p}(X)}\tilde{\mu}_0(X)\right]$$

Now consider the condition $\tilde{p}(X) = P[D=1|X]$.

$$\begin{aligned}
\beta_0 &= \mathrm{E}\left[\frac{(1-D)Y}{1-\tilde{p}(X)} + \frac{D-\tilde{p}(X)}{1-\tilde{p}(X)}\tilde{\mu}_0(X)\right] \\
&= \mathrm{E}\left[\mathrm{E}\left[\frac{(1-D)Y}{1-\tilde{p}(X)} + \frac{D-\tilde{p}(X)}{1-\tilde{p}(X)}\tilde{\mu}_0(X) \;\middle|\; X\right]\right] \\
&= \mathrm{E}\left[\frac{\mathrm{E}[(1-D)Y \mid X]}{1-\tilde{p}(X)} + \frac{\mathrm{E}[D \mid X]-\tilde{p}(X)}{1-\tilde{p}(X)}\tilde{\mu}_0(X)\right] \\
&= \mathrm{E}\left[\frac{\mathrm{E}[(1-D)Y \mid X]}{1-\tilde{p}(X)}\right] \\
&= \mathrm{E}\left[\frac{(1-P[D=1|X])\,\mathrm{E}[(1-D)Y(0) \mid D=0,X] + P[D=1|X]\,\mathrm{E}[(1-D)Y(1) \mid D=1,X]}{1-\tilde{p}(X)}\right] \\
&= \mathrm{E}\left[\frac{(1-P[D=1|X])\,\mathrm{E}[(1-0)Y(0) \mid X]}{1-P[D=1|X]}\right] \\
&= \mathrm{E}[\mathrm{E}[Y(0) \mid X]] \\
&= \mathrm{E}[Y(0)]
\end{aligned}$$

Here we followed the same steps as in question (A). Second line LIE. Third introduced the expectation to the unconditioned random variables. Fourth, eliminated second component inside unconditional expectation. Fifth, law of total probability over D. Sixth eliminated part where (1-D) equals zero when conditioned on D=1. Seventh canceled $1-\tilde{p}(X)$ from num and denom. Eighth applied LIE again.

Now consider the condition $\tilde{\mu}(X) = E[Y(0)|X]$.

$$\begin{aligned}
\beta_0 &= \mathrm{E}\left[\frac{(1-D)Y}{1-\tilde{p}(X)} + \frac{D-\tilde{p}(X)}{1-\tilde{p}(X)}\,\mathrm{E}[Y(0) \mid X]\right] \\
&= \mathrm{E}\left[\mathrm{E}\left[\frac{(1-D)Y}{1-\tilde{p}(X)} + \frac{D-\tilde{p}(X)}{1-\tilde{p}(X)}\,\mathrm{E}[Y(0) \mid X] \;\middle|\; X\right]\right] \\
&= \mathrm{E}\left[\frac{\mathrm{E}[(1-D)Y \mid X]}{1-\tilde{p}(X)} + \frac{\mathrm{E}[D \mid X]-\tilde{p}(X)}{1-\tilde{p}(X)}\,\mathrm{E}[Y(0) \mid X]\right] \\
&= \mathrm{E}\left[\frac{(1-P[D=1|X])\,\mathrm{E}[Y(0) \mid X]}{1-\tilde{p}(X)} + \frac{P[D=1|X]\,\mathrm{E}[Y(0) \mid X]-\tilde{p}(X)\,\mathrm{E}[Y(0) \mid X]}{1-\tilde{p}(X)}\right] \\
&= \mathrm{E}\left[\frac{\mathrm{E}[Y(0) \mid X]-\tilde{p}(X)\,\mathrm{E}[Y(0) \mid X]}{1-\tilde{p}(X)}\right] \\
&= \mathrm{E}[\mathrm{E}[Y(0) \mid X]] \\
&= E[Y(0)]
\end{aligned}$$

In this part I first replace the problem condition in the definition of β_0. The second line applies LIE. The third applies the conditional expectation over random variables that do not depend on X. The fourth line expands the first component within the expectation using the Law of total probability, then cancels the term (1-D) when D=1, and in the second term changes E[D|X] to its prob counterpart. The fifth line operates the numerator. The sixth line factorizes (1-$\tilde{p}(X)$) and cancels out the term in the num and denom. The seventh equality applies LIE to get the unconditional expectation.

□

Part D

Solution:

A to C gives us insights to think about doubly robust estimation under selection on observable. Doubly robust in the sense that we can generate an ATE estimator on the basis of either the propensity score or the conditional mean. However, this may not be too reliable under a finite sample context.

Problem 4

Suppose assumptions IV.E and IV.X are satisfied. Recall:

- **(IV.E)** $Y_i(d,z) = Y_i(d,z')$ always (with prob. 1), for all z, z' and every d.
- **(IV.X)** $Y_i(d,z)$ and Z_i are independent, conditional on W_i for every d and z.

We will derive 4.5 without introducing new unobservable U_i. Consider the following two assumptions.

- **(IV.CTE)** There exists a known function h, an unknown parameter β_h and a known value d_h such that:

$$Y_i(d) - Y_i(d_h) = h(d, W_i)'\beta_h \text{ for all } d$$

- **(IV.LIN)** There exists a known function g, an unknown parameter vector β_g and a known value d_g such that:

$$\mathrm{E}[Y_i(d_g) \mid W_i = w] = g(w)'\beta_g \text{ for all w}$$

Part A

Show that if assumption IV.L is satisfied, then IV.CTE and IV.LIN are satisfied:

Proof. Recall assumption **(IV.L)**: $Y_i(d) = x(d, W_i)'\beta + U_i$ for all d, where x is known, and β is unknown. Also U_i is unobserved r.v. Now let's fix d_h and substract $Y_i(d_h)$ from the outcome for i evaluated at any treatment d.

$$\begin{aligned} Y_i(d) - Y_i(d_h) &= h(d, W_i)'\beta - h(d_h, W_i)'\beta_h + (U_i - U_i) \quad \forall \text{ d} \\ &= x(d, W_i)'\beta_h - x(d_h, W_i)'\beta_h \\ &= [x(d, W_i) - x(d_h, W_i)]'\beta_h \\ &= h(d, d_h, W_i)'\beta_h \\ &= h(d, W_i)'\beta_h \end{aligned}$$

The first equation just substracts $Y(d)$ minus $Y(d_h)$ for some potential treatment d_h fixed for all treatments d. The second line eliminates the unobservables which are the same regardless of treatment status and uses the fact that $\beta = \beta_h$ for any d since IV.L states that treatment effects are constant. The third line factorizes β_h. The fourth line defines the term in brackets by function $h(d, d_h, W_i) = x(d, W_i) - x(d_h, W_i)$. And the fifth line removes d_h since it is fixed. Then we obtain **IV.CTE** for all d, where β_h is unknown, and, if d_h is known we know function h

Now the case for **IV.LIN**. Let's take conditional expectation of the outcome $Y_i(d_g)$ on some treatment d_g conditional on W = w.

$$
\begin{aligned}
\mathrm{E}[Y_i(d_g) \mid W_i = w] &= \mathrm{E}[x(d_g, W_i)'\beta + U_i \mid W_i = w] \\
&= \mathrm{E}[x(d_g, W_i)'\beta \mid W_i = w] + \mathrm{E}[U_i \mid W_i = w] \\
&= \mathrm{E}[x(d_g, W_i)'\beta \mid W_i = w] \\
&= \mathrm{E}[x(d_g, w)'\beta] \\
&= \mathrm{E}[x(d_g, w)']\beta \\
&= \mathrm{E}[x(d_g, w)']\beta_g \\
&= g(d_g, w)'\beta_g \\
&= g(w)'\beta_g
\end{aligned}
$$

The second line takes advantage of linearity of the model. The third line uses the fact that by IV.L $\mathrm{E}[U_i \mid W_i = w] = 0$ for all w and this condition applies for any d, therefore also applies to d_g. The fourth replaces w in function x. The fifth line moves unknown parameter β out of the expectation. The sixth line uses the fact that β is always the same regardless of d_g, so we just change the subscript. The last line redefines $g(d_g, w) = E[x(d_g, w)']$, which is the expression we wanted. Since we know function x, then we know the expectation of that function, so g associated to d_g is known, and β_g is unknown. □

Part B

Show that if IV.CTE and IV.LIN are satisfied, then Assumption IV.L is satisfied.

Proof. We will start by fixing h that satisfies IV.CTE and use β_h and d_h.

$$
\begin{aligned}
&Y_i(d) - Y_i(d_h) = h(d, W_i)'\beta_h \quad \forall \text{ d} \\
&\quad \implies Y_i(d) = Y_i(d_h) + h(d, W_i)'\beta_h && \forall \text{ d} \\
&\quad \implies Y_i(d_g) = Y_i(d_h) + h(d_g, W_i)'\beta_h && \text{for } d_g \text{ that satisfies LIN} \\
&\quad \implies Y_i(d_g) = [Y(d) - h(d, W_i)'\beta_h] + h(d_g, W_i)'\beta_h \\
&\quad \implies Y_i(d_g) = Y(d) + [h(d_g, W_i)' - h(d, W_i)']\beta_h \\
&\quad \implies Y(d) = Y_i(d_g) + [h(d, W_i) - h(d_g, W_i)]'\beta_h \\
&\quad \implies Y(d) - g(W)'\beta_g = \underbrace{Y_i(d_g) - g(W_i)'\beta_g}_{U_i} + [h(d, W_i) - h(d_g, W_i)]'\beta_h \\
&\quad \implies Y(d) = g(W_i)'\beta_g + [h(d, W_i) - h(d_g, W_i)]'\beta_h + U_i \\
&\quad \implies Y(d) = \underbrace{[g(W_i) + h(d, W_i)' - h(d_g, W_i)]}_{x(d, W_i)}\beta_h + U_i \\
&\quad \implies Y(d) = x(d, W_i)'\beta_h + U_i
\end{aligned}
$$

By IV.LIN we have that:

$$\begin{aligned} g(W)'\beta_g &= E[Y(d_g)|W] \\ &= E[Y_i(d_h) + h(d_g, W_i)'\beta_h|W] \\ &= E[Y_i(d_h)|W] + E[h(d_g, W_i)'|W]\beta_h \qquad (2) \\ & \qquad (3) \end{aligned}$$

In the second line we just pass $Y_i(d_h)$ to RHS. The third selects d_g that satisfies IV.LIN since the equality holds for all d. The fourth line replaces $Y_i(d_h)$ using the expression from first line, this holds for every d. The fifth line factorizes β_h. The sixth substract both LHS and RHS by $g(W_i)'\beta_g$ which by IV.LIN is the expected value of $Y_i(d_g)$ conditional on W, we define the difference of the value vs its prediction as U_i. The seventh line passes $g(W_i)'\beta_g$ to the RHS. The eighth line factorizes β_h and takes advantage that $\beta = \beta_h = \beta_g$ from IV.CTE. Finally, we define $x(d, W_i)'\beta_h$ which represents the constant, linear treatment effects expression.

□

Part C

Show that if IV.CTE and IV.LIN hold for some d_h and d_g respectively, then they hold for all d_h and d_g.

Proof. From B we have shown that if IV.CTE holds for some d_h, and if IV.LIN holds for some d_g, then we can recover IV.L. Since IV.L holds for every d, then we can do the same procedure we did in Part A to proof IV.CTE and IV.LIN but for any d_h and d_g respectively (instead of a unique case d_h and d_g), obtaining the desired result.

□

Problem 5

Suppose that $d_f = d_x$ and $E[F_i X_i']$ has full rank d_x. Show that $E[F_i F_i']$ is invertible. Is the opposite implication true? If so, prove it. If not, provide a counterexample. Do these conclusions change if $d_f > d_x$? If so, how?

First argument: $E[F_i X_i']$ has full rank d_x, then $E[F_i F_i']$ is invertible.

Proof. By contradiction, assume $E[F_i F_i']$ not invertible.

$$
\begin{aligned}
&\implies \exists c \in R^{d_f} \;\; s.t. \;\; \mathrm{E}[F_i F_i']c = 0_{d_f} \\
&\implies c' \exists F_i F_i' c = 0 \\
&\implies \exists c' F_i F_i' c = 0 \\
&\implies P[c'F = 0] = 1 \\
&\implies cE[F_i X_i'] = 0
\end{aligned}
$$

Then, we get that $\mathrm{E}[F_i X_i']$ not full rank. Which contradicts the initial condition, then the statement is true.

□

Second argument: If $E[F_i F_i']$ is invertible, then $E[F_i X_i']$ has full rank d_x?

Proof. False. We can think of a simple matrix $X_i = 0_{d_x}$. When multiplied $F_i X_i'$ the resulting matrix does not have full rank, which contradicts the argument. □

First argument but considering $d_f > d_x$.

Results can be generalized using $d_f > d_x$.

Second argument but considering $d_f > d_x$.

False as well. We can use same argument we used in $d_f = d_x$.

Problem 6

Suppose that $d_f = d_x$. Let:

$$\beta_{iv} = E[F_i X_i']^{-1} E[F_i Y_i].$$

Part A

Decompose $\beta = [\beta_1' \beta_2']$ comformably with $X_i = [H_i', W_i']'$ and $F_i = [Z_i', W_i']'$ show that:

$$\beta_1 = \left(E[\tilde{Z}_i \tilde{Z}_i{}']^{-1} E[\tilde{Z}_i H_i']\right)^{-1} \left(E[\tilde{Z}_i \tilde{Z}_i{}']^{-1} E[\tilde{Z}_i Y_i']\right)$$

Where $\tilde{Z}_i = Z_i - L[Z_i|W_i]$. Provide an intuitive interpretation of this formula.

Proof. ...

Some preliminaries first:

A.1 First, we begin by defining the Linear projection of W as $\mathbb{P}_w = W(W^T W)^{-1} W^T$, and the residual generator as $\mathbb{M}_w = I - \mathbb{P}_w$. This matrix has the properties of being invertible and idempotent. Then, we can represent Z partialed out as follows: $\tilde{Z}_i = Z_i - L[Z_i|W_i] = \mathbb{M}_w Z$.

A.2 Second, the problem requires us to consider β_{iv}, I suppose that the problem is assuming that it is unique and exists, otherwise the problem would not make sense (I guess). Since β_{iv} is a function of $E[F_i X_i']^{-1}$, then it follows that $E[F_i X_i']$, which implies $E[F_i F_i']$ is also invertible as shown in Problem 5.

A.3 Since $E[F_i F_i']$ invertible, F must have full rank, which implies Z and W are full rank.

A.4 Third, since F and X both contain matrix W, then condition $d_f = d_x$ plus full rank condition in [A.2] implies $d_H = d_Z$, which means we are in the exactly identified case.

Now, lets conjecture the following equality and show it's true:

$$\begin{aligned}
\beta_{iv} &= E[(F\mathbb{M}_w X_i^T)]^{-1} E[(F\mathbb{M}_w Y)] \\
&= E[(F X_i^T)^{-1} \mathbb{M}_w^{-1}] E[\mathbb{M}_w (Y^T F^T)^T] \\
&= E[\mathrm{E}[(F X_i^T)^{-1} \mathbb{M}_w^{-1} \mid W]\, \mathrm{E}[\mathbb{M}_w (Y^T F^T)^T \mid W]] \\
&= E[\mathrm{E}[(F X_i^T)^{-1} \mid W] \mathbb{M}_w^{-1} \mathbb{M}_w\, \mathrm{E}[FY \mid W]] \\
&= E[\mathrm{E}[(F X_i^T)^{-1} \mid W] I\, \mathrm{E}[FY \mid W]] \\
&= E[F_i X_i^T]^{-1} E[F_i Y_i]
\end{aligned}$$

We have shown it is true. Now, when we apply the residual generator $\mathbb{M}_w$ over F we get the following $F\mathbb{M}_w = [Z_i'\mathbb{M}_w; W_i'\mathbb{M}_w] = [\tilde{Z}_i'; 0]$. Now, replace this expression into β_{iv} and retrieve the first set of parameters β_1 .

$$
\begin{aligned}
\implies \beta_1 &= \mathrm{E}[\tilde{Z}_i H_i^T]^{-1}\,\mathrm{E}[\tilde{Z}_i Y_i] \\
&= \mathrm{E}[\tilde{Z}_i H_i^T]^{-1}\left(E[\tilde{Z}_i\tilde{Z}_i^T]E[\tilde{Z}_i\tilde{Z}_i^T]^{-1}\right)\mathrm{E}[\tilde{Z}_i Y_i] \\
&= \left(\mathrm{E}[\tilde{Z}_i H_i^T]^{-1}E[\tilde{Z}_i\tilde{Z}_i^T]\right)\left(E[\tilde{Z}_i\tilde{Z}_i^T]^{-1}\,\mathrm{E}[\tilde{Z}_i Y_i]\right) \\
&= \left(E[\tilde{Z}_i\tilde{Z}_i^T]^{-1}\,\mathrm{E}[\tilde{Z}_i H_i^T]\right)^{-1}\left(E[\tilde{Z}_i\tilde{Z}_i^T]^{-1}\,\mathrm{E}[\tilde{Z}_i Y_i]\right)
\end{aligned}
$$

The first line just gets the first set of parameters. The second line uses the fact that $E[ZZ']$ invertible, and $\mathbb{M}_w$ is invertible which implies then $E[(\mathbb{M}_w Z)(\mathbb{M}_w Z)'] = E[\tilde{Z}\tilde{Z}']$ is also invertible, so we just multiply and divide in between. third we just rearrange matrices, and fourth we take the inverse out of the parenthesis in the first component, obtaining the desired result.

The intuitive interpretation is that after partialling out covariates W, β_1 (which is the wald estimate of our endogenous variables over outcome Y), is just the ratio between the first stage and the reduced form estimates.

□

Part B

Suppose that $d_f \geq d_x$. Let:

$$\beta_{TSLS} = E[\hat{X}_i\hat{X}_i']^{-1}E[\hat{X}_i Y_i']$$

Where $\hat{X}_i = c_{tsls}F_i$. Decompose $\beta_{TSLS} = [\beta'_{TSLS,1}, \beta'_{TSLS,2}]'$ conformably with $X_i = [H_i', W_i']'$. Partition c_{TSLS} as:

$$c_{TSLS} = \begin{bmatrix} c_{tsls}^{h;z} & c_{tsls}^{h;w} \\ c_{tsls}^{w;z} & c_{tsls}^{w;w} \end{bmatrix}$$

where $c_{tsls}^{h;z}$ is the matrix of first-stage coefficients on Z_i for regressions with components of H_i as the regressands, and similarly for the other submatrices. Show that:

$$\beta_{tsls,1} = E[\dot{Z}\dot{Z}']^{-1}E[\dot{Z}Y]$$

Where $\dot{Z}_i = c_{tsls}^{h,z}\tilde{Z}_i$

Proof. □

Part C

Problem 7[1]

Part A

This just reproduces the OLS estimates. I did clustered errors at the Tribe and State. They do not match the paper because they use a twoway culstering, which is slightly different. However, my t-statistics are still closer to the ones reported in the paper than just controlling for heteroskedasticity.

Table 1: OLS and Tribe Fixed - Effects Results

	log(per capita income)				
Dependent	(1)	(2)	(3)	(4)	(5)
			Panel A: OLS		
Forced coexistence	-0.358***	-0.334***	-0.364***	-0.302***	-0.291***
	(-5.699)	(-5.792)	(-6.094)	(-4.777)	(-4.503)
Historical centralization	0.278***	0.304***	0.351***	0.313***	0.282***
	(4.288)	(5.540)	(5.808)	(4.989)	(2.775)
R^2	0.212	0.360	0.393	0.457	0.600
			Panel B: Tribe Fixed Effects		
Forced coexistence	-0.401***	-0.318***	-	-0.274***	-0.276***
	(-2.442)	(-2.486)	-	(-0.013)	(0.000)
R^2	0.596	0.652	-	-	-
Reservation controls		Y	Y	Y	Y
Tribe controls			Y	Y	Y
Additional reservation-controls				Y	Y
State fixed effecs					Y

Part B

In this section I implement a propensity score estimation for both ATE and ATT for each of the two panels reported in Tables 3 and 5 in the paper. The way I do it is the following:

- First estimate parameters from a Probit model that characterizes selection into treatment conditional on covariates that are described in the bottom of each panel of table 2, and including the exogenous variables which are value of gold and silver separate (for table 2) and added (for table 3).
- Once I have estimated the probit model, I predict the score for all observations in the sample.
- Then I use k nearest neighbors to match people in each group to the k most similar observations in the other group according to the score. I use the euclidean distance as a metric of similarity and when averaging the neighbors I assume equal weights (did not compute a kernel due to time).

[1]The code for this question uses low level commands for all functions needed. My code can be found in my github repository

- Finally, I compute the mean differences for all the sample to find ATE, and on the treated to find ATT.

Results: Results from the exercise look similar in the order of magnitude to IV results, which are also close to estimates OLS in the paper. We do the matching for specifications using 3, 4, 5 and 6 neighbors and random tie breaking.

The trade-offs of using propensity score matching instead of an IV specification is that results can be very sensible to the number of neighbors used to build the counterfactual because the sample size is small (182 observations). We could add more observations and add a smoothing function that gives lower weights to more distant neighbors, however I do not think results would change that much.

However, when using this approach we can build average treatment on the treated. Our results show that ATT seem to be higher than ATE in most specifications. In other words, the effect over income of removing Forced Coexistence for the bands that were treated is greater than the average effect. This is rather intuitive because those bands according to the author were less likely to negotiate to get favorable treaty terms, which is a sign of weak governance -unobservable to the econometrician. So, if one were to eliminate forced coexistence out of those bands, they would benefit the most compared to other type of bands which had stronger forms governance and more capacity to negotiate.

Finally, when comparing estimates from Table 2 and 3 it looks like using a single exogenous indicator seems to give more stable estimates. Perhaps this is linked to the authors' argument that single variables are weakly correlated with the treatment.

Table 2: Propensity Score Matching Estimates - Gold and Silver separate

	log(per capita income)					
	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A: Propensity Score Matching - KNN = 3					
Avg. Treatment Effect	-0.324	-0.303	-0.37	-0.342	-0.291	-0.287
Avg. Treatment on Treated	-0.325	-0.28	-0.401	-0.371	-0.321	-0.322
	Panel B: Propensity Score Matching - KNN = 4					
Avg. Treatment Effect	-0.362	-0.33	-0.375	-0.397	-0.297	-0.296
Avg. Treatment on Treated	-0.408	-0.333	-0.401	-0.433	-0.298	-0.3
	Panel C: Propensity Score Matching - KNN = 5					
Avg. Treatment Effect	-0.328	-0.326	-0.366	-0.354	-0.292	-0.286
Avg. Treatment on Treated	-0.351	-0.328	-0.392	-0.39	-0.271	-0.265
	Panel D: Propensity Score Matching - KNN = 6					
Average Treatment Effect	-0.309	-0.315	-0.363	-0.33	-0.281	-0.277
Average Treatment on Treated	-0.328	-0.309	-0.389	-0.361	-0.263	-0.257
Historical Centralization	Y	Y	Y	Y	Y	Y
Res-controls		Y	Y	Y	Y	Y
Add. tribe-controls			Y	Y	Y	Y
Endog. res-controls				Y	Y	Y
State fixed effects					Y	Y
Add exclusion controls						Y

Table 3: Propensity Score Matching Estimates - Gold and Silver index

	log(per capita income)					
	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A: Propensity Score Matching - KNN = 3					
Avg. Treatment Effect	-0.335	-0.346	-0.448	-0.351	-0.319	-0.332
Avg. Treatment on Treated	-0.329	-0.353	-0.513	-0.366	-0.347	-0.361
	Panel B: Propensity Score Matching - KNN = 4					
Avg. Treatment Effect	-0.378	-0.337	-0.41	-0.385	-0.293	-0.302
Avg. Treatment on Treated	-0.41	-0.342	-0.463	-0.415	-0.306	-0.312
	Panel C: Propensity Score Matching - KNN = 5					
Avg. Treatment Effect	-0.334	-0.335	-0.391	-0.348	-0.292	-0.291
Avg. Treatment on Treated	-0.343	-0.338	-0.452	-0.379	-0.281	-0.279
	Panel D: Propensity Score Matching - KNN = 6					
Average Treatment Effect	-0.316	-0.324	-0.36	-0.329	-0.277	-0.285
Average Treatment on Treated	-0.321	-0.322	-0.408	-0.356	-0.266	-0.274
Historical Centralization	Y	Y	Y	Y	Y	Y
Res-controls		Y	Y	Y	Y	Y
Add. tribe-controls			Y	Y	Y	Y
Endog. res-controls				Y	Y	Y
State fixed effects					Y	Y
Add exclusion controls						Y

Part C

Recall the author's baseline specification:

$$y_{ie} = \alpha_e + \alpha_s + \beta FC_{ie} + \gamma_1 \text{res-controls}_{ie} + \gamma_2 \text{tribe-controls}_{ie} + \epsilon ie$$

Where α_e is tribe fixed effects and α_s are state fixed effects and $\epsilon_{ie} = \eta_e + \nu_{ie}$ is the error term that has a component that is common to all reservations of a tribe, and β_1 is the parameter of interest which is the effect of Forced Coexistence on reservation i.

The author argues that if we assume that only tribe level unobservable affected selection into FC_{ie}, the specification with tribe fixed effects α_e would estimate the causal effect β. However, since it is likely that selection into FC_i is based also on unobservable band characteristics, then he follows an IV approach.

The IV approach uses historical mining rushes, measured as the dollar value of silver and gold mines located outside the border of reservations.

Exogeneity: According to the author, where the U.S. government deemed Native American lands more valuable, it formed fewer, more concentrated reservations in order to free up land and to better monitor tribes to prevent them from migrating back to their ancestral homelands.

He further argues that the instrument is exogenous because allocation was essentially random, since the Native North Americans, unlike the Incas did not have mining, it was unlikely that they selected their location on the basis of the existence of mining reservations.

However, even when tribes did not select their location on the basis of the existence of mining value potential, mining areas surely have geographical aspects that may have influenced their initial settlement, for instance, land altitude, agricultural land infertility.

He controls for land ruggedness to address unobservables affecting the outcome. He should also mention that it could help to address issues of settlement as well. However, I consider it is insufficient since there could be many other unobservables that make assignment not exogenous.

Exclusion The author argues that since the Native Americans did not have mining, then historical silver and gold value should be uncorrelated to tribe's income. However, even when they did not have mining as their main economic activity, the very presence of gold and silver areas could be correlated with unobservables that affect income.

For instance, tribes in more valuable lands could have more bargaining power against the government to decide whether they wanted to coexist with other tribes or not. Also, it could incentive the government to invest in infrastructure necessary to explore mine reservations those areas. Additionally mines could be negatively correlated with fertile land in the area which would affect tribe economic activities.

Relevance

The author argues that instruments are relevant since results from the first stage show a positive and significant correlation between FC_{ie} and historical silver and gold land value, even though coefficients look small.

He also considers adding up the two variables into a single value of precious metals. He uses this strategy because it increases the F-stat which makes his first stage assessment of relevance more credible. However, this could happen just mechanically since it increases the degrees of freedom in his very small sample of 182 observations.

Part D

Here I just reproduce the point estimates from table V. I get the same values.

Table 4: IV Results

	log(per capita income)					
Dependent	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A: Two Instruments					
Forced Coexistence	-0.329	-0.304	-0.36	-0.316	-0.302	-0.403
	Panel B: One Instrument					
Forced Coexistence	-0.406	-0.371	-0.397	-0.35	-0.339	-0.443
Historical Centralization	Y	Y	Y	Y	Y	Y
Res-controls		Y	Y	Y	Y	Y
Add. tribe-controls			Y	Y	Y	Y
Endog. res-controls				Y	Y	Y
State fixed effects					Y	Y
Add. exclusion controls						Y

Part E

Panel B is the specification that considers one single index instrument of gold and silver historical mining value instead of the two separately. The index is the addition of the two variables.

The rationale for including Panel B is because a single index reports stronger F-statistic for weak instruments than in Panel A. In the paper, table V panel A has almost all columns with F values below 10, while panel B reports F values above 10 for all columns except for column 2.

I think including panel B does not make sense since adding the two instruments is just imposing the restriction that each of the instruments has the same contribution per dollar value in the first stage. Why does that have to be the case? Could not gold rush induce governments to enforce coexistence more than silver, or the other way around? Panel A is more flexible in that regard and allows the data speak by estimating the corresponding weights in a linear way.

The reason why the F statistic increases in panel B is most likely to be mechanic. Since the number of observations is very small, reducing the instrument to one variable increases the degrees of freedom relatively more than if we were in a large sample context.