Math Review Notes—Causal Inference and Econometrics

Gregory Faletto

Contents

Last updated October 19, 2020

Chapter 1

Causal Inference and Econometrics

1.1 Generalized Method of Moments (Chapter 13 of Hansen [2020])

1.1.1 Overidentified Moment Equations (Section 13.4 of Hansen [2020])

Consider the instrumental variables model (see Section 1.2.2). The estimator $\hat{\beta}$ is the solution of the moment condition

$$\overline{g}_n(\beta) = \frac{1}{n}\sum_{i=1}^n g_i(\beta) = \frac{1}{n}\sum_{i=1}^n Z_i(Y_i - X_i^\top \beta) = \frac{1}{n}\left(\boldsymbol{Z}^\top \boldsymbol{Y} - \boldsymbol{Z}^\top \boldsymbol{X}\boldsymbol{\beta}\right).$$

If this model is overidentified (that is, the number of instruments ℓ—and therefore moment conditions to satisfy—exceeds the number of variables p in $\boldsymbol{X}$—and therefore the number of parameters to estimate in $\boldsymbol{\beta}$), in general this estimator does not exist, so the method of moments estimator is not defined.

The idea of the generalized method of moments estimator is to make $\overline{g}_n(\beta)$ as close to zero as possible. Define the vector $\boldsymbol{\mu} := \boldsymbol{Z}^\top \boldsymbol{Y} \in \mathbb{R}^\ell$, the matrix $\boldsymbol{G} := \boldsymbol{Z}^\top \boldsymbol{X} \in \mathbb{R}^{\ell \times p}$, and the "error" $\boldsymbol{\eta} := \boldsymbol{\mu} - \boldsymbol{G}\boldsymbol{\beta}$. Then we can write the finite-sample analogue of the above equation as

$$\begin{aligned} & \boldsymbol{Z}^\top \boldsymbol{Y} = \boldsymbol{Z}^\top \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\eta} \\ \iff \quad & \boldsymbol{\mu} = \boldsymbol{G}\boldsymbol{\beta} + \boldsymbol{\eta}. \end{aligned}$$

Therefore the least squares estimator (if we take all moment conditions to be equally important) is $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{G}^\top \boldsymbol{G}\right)^{-1} \boldsymbol{G}^\top \boldsymbol{\mu}$. In general, we may want to weigh some moment conditions as more important than others (possibly because errors are non-homogeneous, in which case this increases efficiency). Then by analogy to weighted least squares (see Section **??**), for some positive definite weight matrix $\boldsymbol{W}$ we have the **generalized method of moments estimator**

$$\hat{\boldsymbol{\beta}} := \left(\boldsymbol{G}^\top \boldsymbol{W} \boldsymbol{G}\right)^{-1} \boldsymbol{G}^\top \boldsymbol{W} \boldsymbol{\mu} = \left(\boldsymbol{X}^\top \boldsymbol{Z} \boldsymbol{W} \boldsymbol{Z}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{Z} \boldsymbol{W} \boldsymbol{Z}^\top \boldsymbol{Y}. \tag{1.1}$$

This minimizes the weighted sum of squares $\boldsymbol{\eta}^\top \boldsymbol{W} \boldsymbol{\eta}$.

Definition 1.1.1 (Generalized Method of Moments estimator; Definition 13.1 in Hansen [2020]). For a positive definite square weight matrix $\boldsymbol{W}$, define the GMM criterion function

$$J(\boldsymbol{\beta}) := n\overline{g}_n(\boldsymbol{\beta})^\top \boldsymbol{W} \overline{g}_n(\boldsymbol{\beta}). \tag{1.2}$$

Then the **generalized method of moments estimator** is

$$\hat{\boldsymbol{\beta}}_{\text{gmm}} := \arg\min_{\beta} \left\{ J_n(\boldsymbol{\beta}) \right\}.$$

Note that GMM includes the method of moments estimator as a special case. This implies that all results for GMM apply to any method of moments estimators. In this case $\boldsymbol{W}$ does not matter. In the overidentified case, the choice of $\boldsymbol{W}$ is important.

1.2 Instrumental Variables (Section 4.8 of Cameron and Trivedi [2005])

1.2.1 Inconsistency of OLS and Examples of Endogeneity (Section 4.8.1 of Cameron and Trivedi [2005], Section 12.3 in Hansen [2020])

- **Measurement error in the regressor.** Suppose $\mathbb{E}[Y \mid Z] = Z^\top \beta$, but Z is not observed; instead, $X = Z + u$ is observed, where u is measurement error with $\mathbb{E}(u) = 0$ and u is independent of e and Z. We have

  $$Y = Z^\top \beta + e = (X - u)^\top \beta + e = X^\top \beta + \nu$$

  where $\nu = e - u^\top \beta$. Therefore

  $$Y = X^\top \beta + \nu,$$

  but

  $$\mathbb{E}[X\nu] = \mathbb{E}[(Z + u)(e - u^\top \beta)] = -\mathbb{E}[uu^\top]\beta \neq 0.$$

  Therefore least squares estimation is inconsistent, and X is endogenous. The projection coefficient (the quantity least squares is consistent for) is (in the case $p = 1$)

  $$\beta^* = \beta + \frac{\mathbb{E}[X\nu]}{\mathbb{E}[X^2]} = \beta \left(1 - \frac{\mathbb{E}[u^2]}{\mathbb{E}[X^2]} \right).$$

  Since $\mathbb{E}[u^2]/\mathbb{E}[X^2] < 1$, the projection coefficient shrinks the structural parameter β towards zero. This is called **measurement error bias** or **attentuation bias**.

- **Simultaneous equations bias.** Suppose that quantity Q and price P are determined jointly by demand

$$Q = -\beta_1 P e_1$$

and supply

$$Q = \beta_2 P + e_2,$$

with (for simplicity) $e = (e_1, e_2)$ satisfying $\mathbb{E}[e] = 0$ and $\mathbb{E}[ee'] = I_2$. In matrix notation, we have

$$\begin{aligned} \begin{pmatrix} 1 & \beta_1 \\ 1 & -\beta_2 \end{pmatrix} \begin{pmatrix} Q \\ P \end{pmatrix} &= \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} \\ \iff \qquad \begin{pmatrix} Q \\ P \end{pmatrix} &= \begin{pmatrix} 1 & \beta_1 \\ 1 & -\beta_2 \end{pmatrix}^{-1} \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} \\ &= \frac{1}{\beta_1 + \beta_2} \begin{pmatrix} \beta_2 & \beta_1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} \\ &= \begin{pmatrix} (\beta_2 e_1 + \beta_1 e_2)/(\beta_1 + \beta_2) \\ (e_1 - e_2)/(\beta_1 + \beta_2) \end{pmatrix}. \end{aligned}$$

The projection of Q on P yields $Q = \beta^* P + e^*$ with $\mathbb{E}[Pe^*] = 0$ and the coefficient defined by projection as

$$\beta^* = \mathbb{E}[P^2]^{-1} \mathbb{E}[PQ] = \frac{\beta_2 - \beta_1}{2}.$$

The projection coefficient β^* equals neither the demand slope β_1 nor the supply slope β_2, but equals an average of the two. (The fact that it is a simple average is an artifact of the covariance structure.) Hence the OLS estimate satisfies $\hat{\beta} \xrightarrow{p} \beta^*$, and the limit does not equal β_1 or β_2. The fact that the limit is neither the supply nor demand slope is called **simultaneous equations bias**. This occurs generally when Y and X are jointly determined, as in market equilibrium. Generally, when both the dependent variable and a regressor are simultaneously determined, the variables should be treated as endogenous.

- **Choice variables as regressors.** Suppose we are interested in outcome y, log-earnings, and we have predictor x, years of schooling. We are interested in the causal effect on y of an **exogenous** change in x—a change in amount of schooling that is not the choice of the individual; for example, an increase in the minimum age at which students leave school. The OLS regression model specifies

$$y = \beta x + u$$

where u is an error term. Regression of y on x yields OLS estimate $\hat{\beta}$ of β. If we assume that x is uncorrelated with u, OLS yields a consistent estimator for the true causal effect. However, u (which contains the effects of all variables besides schooling on earnings) could be correlated with x. For example, unobserved *ability* may be correlated with both earnings and increased levels of schooling. In that case, OLS will be consistent for

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \beta + \frac{\mathrm{d}u}{\mathrm{d}x} > \beta.$$

That is, the positive correlation between x and u means that the linear projection coefficient β^* is upwardly biased relative to the structural coefficient β. The OLS estimator is therefore biased and inconsistent for β, over-estimating the causal effect of education on wages.

This type of endogeneity occurs generally when Y and X are both choices made by an economic agent, even if they are made at different points in time. Generally, when both the dependent variable and a regressor are choice variables made by the same agent, the variables should be treated as endogenous.

A more formal treatment of the linear regression model with K regressors leads to the same conclusion. Under standard assumptions, a necessary condition for consistency of OLS is that $\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{u} \xrightarrow{p} \boldsymbol{0}$; we can see this because

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y} \\
&= \left(\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \frac{1}{n}\boldsymbol{X}^\top \left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}\right) \\
&= \left(\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \frac{1}{n}\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} + \left(\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \frac{1}{n}\boldsymbol{X}^\top \boldsymbol{u} \\
&= \boldsymbol{\beta} + \left(\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \frac{1}{n}\boldsymbol{X}^\top \boldsymbol{u};
\end{aligned}$$

we see this converges to $\boldsymbol{\beta}$ in probability if $\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{u} \xrightarrow{p} \boldsymbol{0}$ (see also Section 4.7.1 of Cameron and Trivedi [2005]).

1.2.2 Instrumental Variable

The inconsistency of OLS is due to the endogeneity of x, meaning that changes in x are associated not only with changes in y bu also changes in the error u. What is needed is a method to generate only exogenous variation in x. An obvious way is through a randomized experiment, but for many economic applications such experiments are too expensive, infeasible, or unethical. One alternative approach is using an instrument.

An **instrument** z is a variable that is correlated with x but not with u or directly with y (that is, z is associated with y only through its effect on x).

Definition 1.2.1 (Instrumental variable; Definition 12.1 in Hansen [2020]). The random vector $Z \in \mathbb{R}^\ell$ is an **instrumental variable** if the following are true:

$$\begin{aligned}
\mathbb{E}[Z^\top e] &= 0, \\
\mathbb{E}[ZZ^\top] &= 0, \qquad \text{and} \\
\operatorname{rank}(\mathbb{E}[ZX^\top]) &= p.
\end{aligned}$$

The first component of this definition is that the instruments are uncorrelated with the regression error. Second, we must exclude linearly dependent instruments. The third condition is often called the **relevance condition** and is essential for the identification of the model. A necessary condition for the relevance condition is $\ell \geq p$.

1.2.3 Instrumental Variables Estimator

For regression with scalar regressor x and scalar instrument z, the **instrumental variables (IV) estimator** is defined as

$$\hat{\beta}_{IV} := \left(\boldsymbol{z}^\top \boldsymbol{x}\right)^{-1} \boldsymbol{z}^\top \boldsymbol{y}.$$

This estimator is consistent for the slope coefficient β in the linear model if z is correlated with x and uncorrelated with u.

We will derive this estimator. Note that under our assumptions,

$$\mathbb{E}\left[\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta} \mid \boldsymbol{z}\right] = \boldsymbol{0}.$$

Using this, we have

$$\boldsymbol{0} = \mathbb{E}\left[\boldsymbol{z}^\top \boldsymbol{0}\right] = \mathbb{E}\left[\boldsymbol{z}^\top \mathbb{E}\left[\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta} \mid \boldsymbol{z}\right]\right] = \mathbb{E}\left[\mathbb{E}\left[\boldsymbol{z}^\top \left(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta}\right) \mid \boldsymbol{z}\right]\right] = \mathbb{E}\left[\boldsymbol{z}^\top \left(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta}\right)\right].$$

If the number of instruments equals the number of regressors ($\dim(\boldsymbol{z}) = p$), the method of moments estimator is then the solution to the corresponding sample moment condition

$$\begin{aligned} & \frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i(y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}) = \boldsymbol{0} \\ \iff \quad & \boldsymbol{z}^\top \left(\boldsymbol{y} - \boldsymbol{x}\hat{\boldsymbol{\beta}}\right) = \boldsymbol{0} \\ \iff \quad & \boldsymbol{z}^\top \boldsymbol{y} = \boldsymbol{z}^\top \boldsymbol{x}\hat{\boldsymbol{\beta}} \\ \iff \quad & \hat{\boldsymbol{\beta}} = \left(\boldsymbol{z}^\top \boldsymbol{x}\right)^{-1} \boldsymbol{z}^\top \boldsymbol{y}, \end{aligned}$$

as shown in (1.4).

1.2.4 Two-Stage Least Squares (Section 8.3.4 of Greene [2003])

Suppose there may be more instruments than endogenous variables. Then $Z^\top X$ is not invertible (it is rank p but has ℓ rows), and a new analysis is required. Since Z is uncorrelated with e, we can express an approximation $\hat{X}$ of X in the column space of Z by projection:

$$\hat{X} = Z(Z^\top Z)^{-1} Z^\top X.$$

Then we can regress y against $\hat{X}$ to get a consistent estimator for the endogenous (structural) coefficient:

$$
\begin{aligned}
\beta_{\mathrm{IV}} &= \left(\hat{X}^\top \hat{X}\right)^{-1} \hat{X}^\top y \\
&= \left(\left[Z(Z^\top Z)^{-1} Z^\top X\right]^\top Z(Z^\top Z)^{-1} Z^\top X\right)^{-1} \left[Z(Z^\top Z)^{-1} Z^\top X\right]^\top y \\
&= \left(X^\top Z(Z^\top Z)^{-1} Z^\top Z(Z^\top Z)^{-1} Z^\top X\right)^{-1} X^\top Z(Z^\top Z)^{-1} Z^\top y \\
&= \left(X^\top Z(Z^\top Z)^{-1} Z^\top X\right)^{-1} X^\top Z(Z^\top Z)^{-1} Z^\top y. \qquad (1.3)
\end{aligned}
$$

Similarly, when p endogenous regressors are in X and p (an equal number) of instruments are available, we have

$$
\hat{\beta}_{IV} := \left(\boldsymbol{Z}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{Z}^\top \boldsymbol{y}. \qquad (1.4)
$$

1.2.5 LATE/CATE Theorem

Theorem 1.2.5.1 (LATE Theorem (Special case of Theorem 2 in Imbens and Angrist [1994])).

$$
\frac{\mathbb{E}[Y_i \mid Z_i = 1] - \mathbb{E}[Y_i \mid Z_i = 0]}{\mathbb{E}[A_i \mid Z_i = 1] - \mathbb{E}[A_i \mid Z_i = 0]} = \mathbb{E}[Y_i(1) - Y_i(0) \mid A_i(1) > A_i(0)].
$$

Proof.

$$
\begin{aligned}
\mathbb{E}[Y_i \mid Z_i = 1] &= \mathbb{E}\left[\underbrace{Y_i(0) + A_i(Y_i(1) - Y_i(0))}_{Y_i} \mid Z_i = 1\right] \\
&= \mathbb{E}\left[Y_i(0) \mid Z_i = 1\right] + \mathbb{E}\left[A_i(Y_i(1) - Y_i(0)) \mid Z_i = 1\right] \\
&\overset{(*)}{=} \mathbb{E}\left[Y_i(0)\right] + \mathbb{E}\left[A_i(Y_i(1) - Y_i(0))\right]
\end{aligned}
$$

where $(*)$ follows from the randomization assumption. Similarly,

$$
\begin{aligned}
\mathbb{E}[Y_i \mid Z_i = 0] &= \mathbb{E}\left[\underbrace{Y_i(0) + A_i(Y_i(1) - Y_i(0))}_{Y_i} \mid Z_i = 0\right] \\
&= \mathbb{E}\left[Y_i(0) \mid Z_i = 0\right] + \mathbb{E}\left[A_i(Y_i(1) - Y_i(0)) \mid Z_i = 0\right] \\
&\overset{(*)}{=} \mathbb{E}\left[Y_i(0)\right] + \mathbb{E}\left[A_i(Y_i(1) - Y_i(0))\right],
\end{aligned}
$$

so

$$
\begin{aligned}
\mathbb{E}[Y_i \mid Z_i = 1] - \mathbb{E}[Y_i \mid Z_i = 0] = &\ \mathbb{E}\left[(A_i(1) - A_i(0))(Y_i(1) - Y_i(0))\right] \\
= &\ \mathbb{E}\Bigg[(A_i(1) - A_i(0))(Y_i(1) - Y_i(0)) \mid \underbrace{A_i(1) > A_i(0)}_{\text{(compliers)}}\Bigg] \mathbb{P}(A_i(1) > A_i(0)) \\
&+ \mathbb{E}\Bigg[(A_i(1) - A_i(0))(Y_i(1) - Y_i(0)) \mid \underbrace{A_i(1) < A_i(0)}_{\text{(defiers)}}\Bigg] \mathbb{P}(A_i(1) < A_i(0)) \\
\overset{(**)}{=} &\ \mathbb{E}\left[(A_i(1) - A_i(0))(Y_i(1) - Y_i(0)) \mid A_i(1) > A_i(0)\right] \mathbb{P}(A_i(1) > A_i(0))
\end{aligned}
$$

where $(**)$ follows from the monotonicity assumption ($\mathbb{P}(A_i(1) < A_i(0)) = 0$). Can get denominator (easier)

$$
\mathbb{E}[A_i \mid Z_i = 1] - \mathbb{E}[A_i \mid Z_i = 0] = \mathbb{E}[A_i(1) - A_i(0)] = \mathbb{P}(A_i(1) > A_i(0))
$$

□

Lemma 1.2.5.2.

$$
\frac{\mathbb{E}[Y_i \mid Z_i = 1] - \mathbb{E}[Y_i \mid Z_i = 0]}{\mathbb{E}[A_i \mid Z_i = 1] - \mathbb{E}[A_i \mid Z_i = 0]} = \mathbb{E}[Y_i(1) - Y_i(0) \mid A_i = 1].
$$

Proof. We have $Y_i = Y_i(0) + A_i(Y_i(1) - Y_i(0))$. Also, $\mathbb{P}(A_i = 1 \mid Z_i = 0) = 0$ by the assumption of no defiers. That is, $\mathbb{E}[A_i(Y_i(1) - Y_i(0)) \mid Z_i = 0] = 0$. Then

$$
\begin{aligned}
\mathbb{E}[Y_i \mid Z_i = 1] - \mathbb{E}[Y_i \mid Z_i = 0] &= \mathbb{E}[Y_i(0) + A_i(Y_i(1) - Y_i(0)) \mid Z_i = 1] - \mathbb{E}[Y_i(0) + A_i(Y_i(1) - Y_i(0)) \mid Z_i = 0] \\
&= \mathbb{E}[A_i(Y_i(1) - Y_i(0)) \mid Z_i = 1] - \mathbb{E}[A_i(Y_i(1) - Y_i(0)) \mid Z_i = 0] \\
&= \mathbb{E}[A_i(Y_i(1) - Y_i(0)) \mid Z_i = 1] \\
&= \mathbb{E}[A_i(Y_i(1) - Y_i(0)) \mid Z_i = 1, A_i = 1] \cdot \mathbb{P}(A_i = 1 \mid Z_i = 1) \\
&= \mathbb{E}[A_i(Y_i(1) - Y_i(0)) \mid A_i = 1] \cdot \mathbb{P}(A_i = 1 \mid Z_i = 1)
\end{aligned}
$$

so this is the numerator.

□

1.2.6 GMM Estimator (Section 13.6 of Hansen [2020])

As discussed in Section 1.1.1, the moment equations for instrumental variables are

$$
\boldsymbol{Z}^\top \boldsymbol{Y} - \boldsymbol{Z}^\top \boldsymbol{X} \boldsymbol{\beta} = 0,
$$

so the GMM criterion (1.2) can be written as

$$J(\beta) = n\left(\boldsymbol{Z}^\top \boldsymbol{Y} - \boldsymbol{Z}^\top \boldsymbol{X}\boldsymbol{\beta}\right)^\top \boldsymbol{W}\left(\boldsymbol{Z}^\top \boldsymbol{Y} - \boldsymbol{Z}^\top \boldsymbol{X}\boldsymbol{\beta}\right).$$

The GMM estimator minimizes $J(\beta)$. The first order conditions are

$$\begin{aligned}
0 &= \frac{\partial}{\partial \beta} J(\hat{\beta}) \\
&= 2\frac{\partial}{\partial \beta}\overline{g}_n(\hat{\beta})^\top \boldsymbol{W} \overline{g}_n(\hat{\beta}) \\
&= -2\left(\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{Z}\right)\boldsymbol{W}\left(\frac{1}{n}\boldsymbol{Z}^\top(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})\right).
\end{aligned}$$

The solution is the GMM estimator for the overidentified IV model,

$$\hat{\boldsymbol{\beta}}_{\text{gmm}} = \left(\boldsymbol{X}^\top \boldsymbol{Z}\boldsymbol{W}\boldsymbol{Z}^\top \boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top \boldsymbol{Z}\boldsymbol{W}\boldsymbol{Z}^\top \boldsymbol{Y},$$

the same estimator as in (1.1). The dependence on the estimator $\boldsymbol{W}$ is only up to scale; that is, if $\boldsymbol{W}$ is replaced by $c\boldsymbol{W}$ for some $c > 0$, $\hat{\boldsymbol{\beta}}_{\text{gmm}}$ does not change. When $\boldsymbol{W}$ is fixed by the user, we call $\hat{\boldsymbol{\beta}}_{\text{gmm}}$ a **one-step GMM** estimator. Note that by comparison to (1.3), we see that if $\boldsymbol{W} = \left(\boldsymbol{Z}^\top \boldsymbol{Z}\right)^{-1}$ then we have the two stage least squares estimator. Also note that if $\ell = p$ then $\boldsymbol{X}^\top \boldsymbol{Z}$ is invertible (as is $\boldsymbol{W}$ since it is positive definite by assumption) and we have

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{\text{gmm}} &= \left(\boldsymbol{Z}^\top \boldsymbol{X}\right)^{-1}\boldsymbol{W}^{-1}\left(\boldsymbol{X}^\top \boldsymbol{Z}\right)^{-1}\boldsymbol{X}^\top \boldsymbol{Z}\boldsymbol{W}\boldsymbol{Z}^\top \boldsymbol{Y} \\
&= \left(\boldsymbol{Z}^\top \boldsymbol{X}\right)^{-1}\boldsymbol{W}^{-1}\boldsymbol{W}\boldsymbol{Z}^\top \boldsymbol{Y} \\
&= \left(\boldsymbol{Z}^\top \boldsymbol{X}\right)^{-1}\boldsymbol{Z}^\top \boldsymbol{Y},
\end{aligned}$$

which matches the estimator in (1.4).

1.3 DSO 699 [Imbens and Rubin, 2015]

1.3.1 Causal Estimands (Section 1.20 of Imbens and Rubin [2015], p. 18)

Different treatment effects:

1.
$$\text{Ave}(Y_i(\text{Asp}) - Y_i(\text{No})) = \frac{1}{N}\sum_{i=1}^{N}(Y_i(\text{Asp}) - Y_i(\text{No}))$$

$$\text{Ave}(Y_i(\text{Asp}) - Y_i(\text{No})) = \mathbb{E}(Y_i(\text{Asp}) - Y_i(\text{No}))$$

2.
$$\text{Median}(Y_i(\text{Asp}) - Y_i(\text{No}))$$

3.
$$\text{Median}(Y_i(\text{Asp})) - \text{Median}(Y_i(\text{No}))$$

4. (Subpopulations, or heterogeneous treatment effects [HTE])

$$\tau_{fs}(f) = \text{Ave}_{X_i=\text{female}}(Y_i(\text{Asp}) - Y_i(\text{No})) = \frac{1}{N(f)} \sum_{i:X_i=\text{female}} (Y_i(\text{Asp}) - Y_i(\text{No}))$$

Third one is easier to study than second.

1.3.2 Assignment Mechanisms (Chapter 4 of Imbens and Rubin [2015])

- Completely randomized design
- Bernoulli assignment
- Bernoulli assignment within blocks
- Probability of treatment depending on covariates
- Randomized within matched pairs

1.4 Regression Methods (Chapter 7 of Imbens and Rubin [2015])

Selection bias problem: we have

$$\begin{aligned} Y_i^{\text{obs}} &= \begin{cases} Y_i(1), & W_i = 1, \\ Y_i(0), & W_i = 0 \end{cases} \\ &= Y_i(1)W_i + Y_i(0)(1 - W_i) \\ &= Y_i(0) + (Y_i(1) - Y_i(0))W_i. \end{aligned}$$

Note that

$$\begin{aligned} \mathbb{E}[Y_i^{\text{obs}} \mid W_i = 1] - \mathbb{E}[Y_i^{\text{obs}} \mid W_i = 0] &= \mathbb{E}[Y_i(1) \mid W_i = 1] - \mathbb{E}[Y_i(0) \mid W_i = 0] \\ &= \mathbb{E}[Y_i(1) \mid W_i = 1] - \mathbb{E}[Y_i(0) \mid W_i = 1] + \mathbb{E}[Y_i(0) \mid W_i = 1] - \mathbb{E}[Y_i(0) \mid W_i = 0] \\ &= \underbrace{\mathbb{E}[Y_i(1) - Y_i(0) \mid W_i = 1]}_{\text{average treatment effect on treated}} + \underbrace{\mathbb{E}[Y_i(0) \mid W_i = 1] - \mathbb{E}[Y_i(0) \mid W_i = 0]}_{\text{selection bias}} \end{aligned}$$

Random assignment solves the selection bias issue:

$$\mathbb{P}(W_i = 1 \mid Y_i(0), Y_i(1)) = \mathbb{P}(W_i = 1)$$

$$W_i \perp\!\!\!\perp (Y_i(0), Y_i(1))$$

1.4.1 Linear Regression with No Covariates (Section 7.4 of Imbens and Rubin [2015])

Regression approach: Let $\alpha = \mathbb{E}[Y_i(0)]$. Define

$$\epsilon_i = \begin{cases} Y_i^{\text{obs}} - \alpha, & W_i = 0 \\ Y_i^{\text{obs}} - \alpha - \tau, & W_i = 1. \end{cases}$$

(Note that Y_i^{obs} is random due to (1) random sampling of which observational units are included in the sample and (2) (possibly) randomized treatment assignment. We have

$$Y_i^{\text{obs}} = \alpha + \tau W_i + \epsilon_i.$$

In order for least squares to be consistent, we need to verify whether $\mathbb{E}[\epsilon_i \mid W_i = 1] = \mathbb{E}[\epsilon_i \mid W_i = 0]$. Under the assumption that $W_i \perp\!\!\!\perp (Y_i(0), Y_i(1))$,

$$\begin{aligned} \mathbb{E}[\epsilon_i \mid W_i = 1] &= \mathbb{E}\left[Y_i^{\text{obs}} - \alpha - \tau \mid W_i = 1\right] \\ &= \mathbb{E}\left[Y_i(1) - \alpha - \tau \mid W_i = 1\right] \\ &= \mathbb{E}\left[Y_i(1)\right] - \alpha - \tau \\ &= \mathbb{E}\left[Y_i(1)\right] - \mathbb{E}[Y_i(0)] - (\mathbb{E}[Y_i(1)] - \mathbb{E}[y_i(0)]) \\ &= 0. \end{aligned}$$

Similarly, under this assumption $\mathbb{E}[\epsilon_i \mid W_i = 0] = 0$. We can estimate τ usingOLS:

$$\hat{\tau}^{\text{obs}} = \frac{1}{\sum_{i=1}^N (W_i - \overline{W})^2} \sum_{i=1}^N (Y_i^{\text{obs}} - \overline{Y}^{\text{obs}})(W_i - \overline{W}) = \frac{1}{\sum_{i=1}^N (W_i - \overline{W})^2} \sum_{i=1}^N (Y_i^{\text{obs}} - \overline{Y}^{\text{obs}}) W_i$$

We have

$$\sum_{i=1}^N (W_i - \overline{W})^2 = \sum_{i=1}^N W_i^2 - N\overline{W}^2 = N_t - N\left(\frac{N_t}{N}\right)^2 = \frac{NN_t - N_t^2}{N} = \frac{N_t N_c}{N}.$$

In the numerator,

$$
\begin{aligned}
\sum_{i=1}^{N}(Y_i^{\text{obs}} - \overline{Y}^{\text{obs}})(W_i - \overline{W}) &= \sum_{i=1}^{N}(Y_i^{\text{obs}} - \overline{Y}^{\text{obs}})W_i \\
&= \sum_{i=1}^{N} Y_i^{\text{obs}} W_i - \sum_{i=1}^{N} \overline{Y}^{\text{obs}} W_i \\
&= \sum_{W_i=1} Y_i(1) - \overline{Y}^{\text{obs}} N_t \\
&= N_t \overline{Y}^{\text{obs}}(1) - \overline{Y}^{\text{obs}} N_t \\
&= N_t \left(\overline{Y}^{\text{obs}}(1) - \overline{Y}^{\text{obs}} \right) \\
&= N_t \left(\overline{Y}^{\text{obs}}(1) - \frac{1}{N} \sum_{W_i=1} Y_i(1) - \frac{1}{N} \sum_{W_i=0} Y_i(0) \right) \\
&= N_t \left(\overline{Y}^{\text{obs}}(1) - \frac{N_t}{N} \overline{Y}_i^{\text{obs}}(1) - \frac{N_c}{N} \overline{Y}_i^{\text{obs}}(0) \right) \\
&= \frac{N_t N_c}{N} \left(\overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}} \right) \\
&\vdots \\
&= \sum_{W_i=1} (Y_i(1) - \overline{Y}^{\text{obs}}) W_i + \sum_{W_0=1} (Y_i(0) - \overline{Y}^{\text{obs}}) W_i \\
&= \sum_{W_i=1} (Y_i(1) - \overline{Y}^{\text{obs}})
\end{aligned}
$$

Therefore

$$
\hat{\tau}^{\text{obs}} = \frac{N_t N_c}{N} \left(\overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}} \right) \Big/ \frac{N_t N_c}{N} = \overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}},
$$

which is the same as Neyman's estimator. Now consider the variance estimator.

$$
\text{Var}(\hat{\tau}^{\text{obs}}) = \frac{\hat{\sigma}^2_{Y|W}}{\sum_{i=1}^{N}(W_i - \overline{W}})^2.
$$

$$
\begin{aligned}
\hat{\sigma}^2_{Y|X} &= \frac{1}{N-2} \sum_{i=1}^{N} (Y_i^{\text{obs}} - \hat{Y}_I^{\text{obs}})^2 \\
&= \frac{1}{N-2} \left[\left(\sum_{W_i=1} Y_I^{\text{obs}} - \overline{Y}_t^{\text{obs}} \right)^2 + \left(\sum_{W_i=0} Y_I^{\text{obs}} - \overline{Y}_c^{\text{obs}} \right)^2 \right]
\end{aligned}
$$

$$
\hat{V} = \frac{1}{N-2} \left[\left(\sum_{W_i=1} Y_I^{\text{obs}} - \overline{Y}_t^{\text{obs}} \right)^2 + \left(\sum_{W_i=0} Y_I^{\text{obs}} - \overline{Y}_c^{\text{obs}} \right)^2 \right] \Big/ \frac{N_t N_c}{N}
$$

$$N\hat{V} \to \frac{S_t^2}{\rho} + \frac{S_c^2}{1-\rho}$$

Variance estimator (p. 121):

$$\hat{V}^{\text{homosk}} = \left(\frac{1}{N_c} + \frac{1}{N_t} \right) \hat{\sigma}^2_{Y|w}$$

Calculations for heteroskedastic robust variance estimator (p. 121):

$$\left(\sum_{i=1}^{N} \left(W_i - \overline{W} \right)^2 \right) = \frac{N_t N_c}{N}.$$

$$\begin{aligned}
\hat{\epsilon}_i &= Y_i^{\text{obs}} - \hat{Y}_i^{\text{obs}} = Y_i^{\text{obs}} - \hat{\alpha} - \hat{\tau} W_i \\
&= \begin{cases} Y_i^{\text{obs}} - \hat{\alpha}, & W_i = 0, \\ Y_i^{\text{obs}} - \hat{\alpha} - \hat{\tau}, & W_i = 1 \end{cases} \\
&= \begin{cases} Y_i(0) - \left(\overline{Y} - \hat{\tau}\overline{W} \right), & W_i = 0, \\ Y_i(1) - \left(\overline{Y} - \hat{\tau}\overline{W} \right) - \hat{\tau}, & W_i = 1 \end{cases}
\end{aligned}$$

where $\hat{\alpha} = \overline{Y} - \hat{\tau}\overline{W}$. Consider $W_i = 0$.

$$\begin{aligned}
Y_i(0) - \left(\overline{Y} - \hat{\tau}\overline{W} \right) &= Y_i(0) - \overline{Y} - \left(\overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}} \right) \frac{N_t}{N} \\
&= Y_i(0) - \frac{1}{N} \sum_{i=1}^{N} Y_i^{\text{obs}} - \frac{1}{N} \sum_{W_i=1} Y_i^{\text{obs}} + \frac{N_t}{N_c N} \sum_{W_i=0} Y_i^{\text{obs}} \\
&= Y_i(0) - \frac{N_t}{N_c N} \left(\frac{N_c}{N_t} \sum_{i=1}^{N} Y_i^{\text{obs}} - \sum_{W_i=0} Y_i^{\text{obs}} \right) - \frac{1}{N} \sum_{W_i=1} Y_i^{\text{obs}} \\
&= Y_i(0) - \frac{N_t}{N_c N} \left(\frac{N_c}{N_t} \sum_{i=1}^{N} Y_i^{\text{obs}} - \sum_{W_i=0} Y_i^{\text{obs}} \right) - \frac{1}{N} \sum_{W_i=1} Y_i^{\text{obs}} \\
&= Y_i(0) - \frac{N_t}{N_c N} \left(\frac{N_c}{N_t} \sum_{W_i=1} Y_i^{\text{obs}} + \frac{N_c}{N_t} \sum_{W_i=0} Y_i^{\text{obs}} - \sum_{W_i=0} Y_i^{\text{obs}} \right) - \frac{1}{N} \sum_{W_i=1} Y_i^{\text{obs}} \\
&\vdots \\
&= Y_i(0) - \overline{Y}_c^{\text{obs}}
\end{aligned}$$

Similarly, for $W_i = 1$

$$Y_i(1) - (\overline{Y} - \hat{\tau}^{\text{obs}}\overline{W}) - \hat{\tau}^{\text{ols}} = Y_i(1) - \overline{Y}_t^{\text{obs}}.$$

We have

$$\hat{\epsilon}_i = \begin{cases} Y_i(0) - \overline{Y}_c^{\text{obs}}, & W_i = 0 \\ Y_i(1) - \overline{Y}_t^{\text{obs}}, & W_i = 1 \end{cases}$$

$$\begin{aligned} \sum_{i=1}^N \hat{\epsilon}_i^2 (W_i - \overline{W})^2 &= \sum_{W_i=1} \hat{\epsilon}_i^2 (1 - \frac{N_t}{N})^2 + \sum_{W_i=0} \hat{\epsilon}_i^2 (0 - \frac{N_t}{N})^2 \\ &\vdots \\ &= \frac{N_c^2}{N^2} \sum_{W_i=1} (Y_i(1) - \overline{Y}_t^{\text{obs}})^2 + \frac{N_t^2}{N^2} \sum_{W_i=0} (Y_i(0) - \overline{Y}_c^{\text{obs}})^2 \end{aligned}$$

Then

$$\begin{aligned} \hat{V}^{\text{hetero}} &= \frac{1}{N_t^2} \sum_{W_i=1} (Y_i(10 - \overline{Y}_t^{\text{obs}})^2 + \frac{1}{N_c^2} \sum_{W_i=0} (Y_i(0) - \overline{Y}_c^{\text{obs}})^2 \\ &\approx \frac{1}{N_t} S_t^2 + \frac{1}{N_c} S_c^2 \end{aligned}$$

where

$$S_t^2 = \frac{1}{N_t - 1} \sum_{W_i=1} (Y_i(1) - \overline{Y}_t^{\text{obs}})^2$$

We can also use weighted least squares if we think the errors are heteroskedastic. The estimator is

$$\hat{\tau}_{wls} = \sum_{i=1}^N \frac{1}{\sigma_i^2} (y_i - \alpha - \beta^\top w_i)^2$$

with $\sigma_i^2 = \text{Var}(\epsilon_i)$. Then

$$N\hat{V}^{\text{hetero}} \approx \frac{1}{N_t/N} S_t^2 + \frac{1}{N_c/N} S_c^2 \xrightarrow{p} \frac{1}{\rho}\sigma_t^2 + \frac{1}{1-\rho}\sigma_c^2$$

where σ_t^2 is the population variance of $Y_i(1)$ and σ_c^2 is the population variance of $Y_i(0)$.

1.4.2 Linear Regression with Additional Covariates (Section 7.5 of Imbens and Rubin [2015])

Notes on Theorem 7.1(i):

$$
\begin{aligned}
\tau^* &= \frac{\mathrm{Cov}(Y_i^{\mathrm{obs}}, W_i)}{\mathrm{Var}(Y_i^{\mathrm{obs}})} \\
&= \frac{\mathbb{E}[Y_i^{\mathrm{obs}} W_i] - \mathbb{E}[Y_i^{\mathrm{obs}}]\mathbb{E}[W_i]}{p(1-p)} \\
&= \frac{\mathbb{E}[Y_i(1)W_i] - \mathbb{E}[W_i Y_i(1) + (1-W_i)Y_i(0)]p}{p(1-p)} \\
&= \frac{p\mu_t - p[p\mu_t + (1-p)\mu_c]}{p(1-p)} \\
&= \mu_t - \mu_c \\
&= \tau.
\end{aligned}
$$

where p is the probability of treatment.

1.4.3 Testing for the Presence of Treatment Effects (Section 7.9 of Imbens and Rubin [2015])

1.5 Model-Based Inference for Completely Randomized Experiments (Chapter 8 of Imbens and Rubin [2015])

1.5.1 A Simple Example: Naive and More Sophisticated Approaches to Estimation (Section 8.3 of Imbens and Rubin [2015])

Theorem 1.5.1.1. For the mean imputation method,

$$
\hat{\tau}_{\text{impute}} = \hat{\tau}^{\text{dif}} = \overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}}
$$

For the sampling imputation method,

$$
\mathbb{E}\left[\hat{\tau}_{\text{impute}} \mid \boldsymbol{Y}^{\text{obs}}, \boldsymbol{W}\right] = \hat{\tau}^{\text{dif}} = \overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}}.
$$

Proof. (a) For mean imputation, we impute the missing observation for observation i by taking the mean among all observations j with $W_j = 1 - W_i$. That is,

$$
\hat{Y}_i^{\text{mis}} = (1 - W_i)\,\overline{Y}_t^{\text{obs}} + W_i \overline{Y}_c^{\text{obs}}.
$$

Then (using $W_i(1 - W_i) = 0$, $W_i^2 = W_i$, and $(1 - W_i)^2 = (1 - W_i)$ for all i)

$$\begin{aligned}
\hat{\tau}^{\text{impute}} &= \frac{1}{N}\sum_{i=1}^{N}(2W_i-1)(Y_i^{\text{obs}}-\hat{Y}_i^{\text{mis}}) \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(W_iY_i^{\text{obs}}+(1-W_i)\hat{Y}_i^{\text{mis}}-\left[(1-W_i)Y_i^{\text{obs}}+W_i\hat{Y}_i^{\text{mis}}\right]\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(W_iY_i^{\text{obs}}+(1-W_i)\left[(1-W_i)\,\overline{Y}_t^{\text{obs}}+W_i\overline{Y}_c^{\text{obs}}\right]\right. \\
&\quad \left.-\left[(1-W_i)Y_i^{\text{obs}}+W_i\left[(1-W_i)\,\overline{Y}_t^{\text{obs}}+W_i\overline{Y}_c^{\text{obs}}\right]\right]\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(W_iY_i^{\text{obs}}+(1-W_i)^2\overline{Y}_t^{\text{obs}}-\left[(1-W_i)Y_i^{\text{obs}}+W_i^2\overline{Y}_c^{\text{obs}}\right]\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}W_i\left[Y_i^{\text{obs}}-\overline{Y}_c^{\text{obs}}\right]+\frac{1}{N}\sum_{i=1}^{N}(1-W_i)\left[\overline{Y}_t^{\text{obs}}-Y_i^{\text{obs}}\right] \\
&= \frac{1}{N}\cdot N_t\left(\overline{Y}_t^{\text{obs}}-\overline{Y}_c^{\text{obs}}\right)+\frac{1}{N}\cdot N_c\left(\overline{Y}_t^{\text{obs}}-\overline{Y}_c^{\text{obs}}\right) \\
&= \frac{N_t+N_c}{N}\left(\overline{Y}_t^{\text{obs}}-\overline{Y}_c^{\text{obs}}\right) \\
&= \overline{Y}_t^{\text{obs}}-\overline{Y}_c^{\text{obs}} \\
&= \hat{\tau}^{\text{dif}}.
\end{aligned}$$

(b) For the second imputation method, observe that $\hat{Y}_i^{\text{mis}}$ is a random variable, with

$$\mathbb{E}\left[\hat{Y}_i^{\text{mis}} \mid \{Y_i^{\text{obs}},W_i\}_{i=1}^N\right]=(1-W_i)\,\overline{Y}_t^{\text{obs}}+W_i\overline{Y}_c^{\text{obs}}.$$

$$\begin{aligned}
\mathbb{E}\left[\hat{\tau}^{\text{impute}} \mid \{Y_i^{\text{obs}},W_i\}_{i=1}^N\right] &= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}(2W_i-1)(Y_i^{\text{obs}}-\hat{Y}_i^{\text{mis}}) \mid \{Y_i^{\text{obs}},W_i\}_{i=1}^N\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}(2W_i-1)\left(Y_i^{\text{obs}}-\mathbb{E}\left[\hat{Y}_i^{\text{mis}} \mid \{Y_i^{\text{obs}},W_i\}_{i=1}^N\right]\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}(2W_i-1)\left(Y_i^{\text{obs}}-\left[(1-W_i)\,\overline{Y}_t^{\text{obs}}+W_i\overline{Y}_c^{\text{obs}}\right]\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(W_iY_i^{\text{obs}}+(1-W_i)\hat{Y}_i^{\text{mis}}-\left[(1-W_i)Y_i^{\text{obs}}+W_i\hat{Y}_i^{\text{mis}}\right]\right).
\end{aligned}$$

Then the rest follows from the proof of part (a).

□

1.5.2 Bayesian Model-Based Imputation in the Absence of Covariates (Section 8.4 of Imbens and Rubin [2015])

Bibliography

A. C. Cameron and P. K. Trivedi. *Microeconometrics: Methods and Applications*. Cambridge University Press, 2005. doi: 10.1017/CBO9780511811241.

W. H. Greene. *Econometric Analysis*. Pearson Education, fifth edition, 2003. ISBN 0-13-066189-9. URL http://pages.stern.nyu.edu/~wgreene/Text/econometricanalysis.htm.

B. E. Hansen. *Econometrics*. August 2020.

G. Imbens and D. B. Rubin. *Causal inference for statistics, social, and biomedical sciences: an introduction*. Cambridge University Press, 2015.

G. W. Imbens and J. D. Angrist. Identification and Estimation of Local Average Treatment Effects. *Econometrica*, 62(2):467, 1994. ISSN 00129682. doi: 10.2307/2951620.