Model Based Instrumental Variables

Bryan S. Graham, UC - Berkeley & NBER

February 13, 2021

Angrist (1990) studies the relationship between military service in Vietnam and earnings for American men born in the early 1950s. For concreteness let the population of interest be American men born in 1950. Let Y denote earnings in 1981 for a random draw from this population; let D equal one if this individual served in Vietnam and zero otherwise; finally, let X equal one if the individual was "draft eligible" (i.e., had a low draft lottery number) and zero otherwise. Our goal is to use the random variation in military service induced by the draft lottery to study its effect on earnings.

Draft eligibility ($X = 1$) encourages military service. In what follows we will refer to X as the *encouragement* or *instrumental variable.* It is helpful to categorize people into types based on their military service behavior under encouragement as well as its absence. Does an individual always serve? Does he serve only when encouraged? This involves some counterfactual or hypothetical thinking; we only observe an individual's behavior under encouragement or not, never both at the same time.

Let $D(1) \in \{0,1\}$ denote an individual's veteran status when, *possibly contrary to fact*, they are draft eligible. Let $D(0) \in \{0,1\}$ denote their veteran status, again possibly contrary to fact, when they are draft ineligible. We can partition individuals into four strata based upon their military service behavior with and without encouragement.

Table 1 depicts the four possible *compliance strata.* First, there are *never-takers* ($D(0) = 0, D(1) = 0$), who never serve in the military irrespective of draft eligibility. Second, there are *always-takers* ($D(0) = 1, D(1) = 1$), who serve in the military regardless of draft eligibility. Third, and important for our analysis, are *compliers* ($D(0) = 0, D(1) = 1$), who

Table 1: Compliance Strata

	$D(1) = 0$	$D(1) = 1$
$D(0) = 0$	Never-Taker	Complier
$D(0) = 1$	Defier	Always-Taker

serve when encouraged (i.e., when draft eligible), but who do not serve when not encouraged. Finally, there are *defiers* $(D(0) = 1, D(1) = 0)$, who avoid service when encouraged, but – inexplicably – serve when not. Will be rule out the existence of this last compliance stratum by making the **monotonicity assumption** that $D(1) \geq D(0)$: no unit is less likely to serve when encouraged to do so.

Let $Y(d, x)$ denote an individual's earnings in 1981 when, possibly contrary to fact, their military service and draft eligibility are, respectively, $D = d$ and $X = x$. We will additionally make the **exclusion restriction** that $Y(d, 1) = Y(d, 0)$ for $d = 0, 1$. Conditional on serving in the military, whether one was draft eligible or not has no effect on earnings. Likewise conditional on not serving in the military, draft eligibility has no effect on earning. An implication of this assumption is that any effect of draft eligibility on earnings must operate indirectly by inducing individuals to serve or not to serve. There is not direct effect of an individuals draft eligibility on earnings.

Our final assumption is that the encouragement/instrument is **randomly assigned**.

$$(D(0), D(1), Y(0), Y(1)) \perp X. \tag{1}$$

This X is a function of an randomly assigned lottery number condition (1) is reasonable.

The causal effect of military service on earnings for individual i equals

$$Y_i(1) - Y_i(0).$$

While this effect is well defined, it is never observable since we either observe earnings given military service or not, but never both simultaneously. Indeed observed earnings equal

$$Y = (1 - D) Y(0) + DY(1). \tag{2}$$

with military service being generated according to

$$D = (1 - X) D(0) + XD(1). \tag{3}$$

While individual-level treatment effects are never identified, our hope is that average effects are. Below we will show that the average effect of service among compliers is identified:

$$\beta^{\text{LATE}} = \mathbb{E}\left[Y(1) - Y(0) \middle| \text{Unit is a complier}\right].$$

This result was first show by Imbens & Angrist (1994), who called the identified average the

local average treatment effect or LATE.

Identification

Let A be a 3×1 vector of compliance strata indicator variables. If an individual is a never-taker the first element of this vector is a one, with the other elements being zero. If he is a complier, the second element is a one, with the other elements being zero. Finally, if he is an always-taker the last element is a one with the balance being zero. Recall that there are no defiers when we maintain the monotonicity assumption.

Let a_j be a vector of zeros with the exception of the j^{th} element, which is equal to one. Further let $\pi_j = \Pr(A = a_j)$ equal the population frequency of compliance strata j.

By the law of total probability

$$\Pr(D = 1|\, X = 0) = 0 \cdot \pi_1 + 0 \cdot \pi_2 + 1 \cdot \pi_3 \quad (4)$$

$$\Pr(D = 1|\, X = 0) = 0 \cdot \pi_1 + 1 \cdot \pi_2 + 1 \cdot \pi_3. \quad (5)$$

To derive (4) and (5) it is helpful to observe that an individual's compliance strata A is fully determined by their configuration of $(D(0), D(1))$. By random assignment we therefore have that A and X are independent (i.e., $\Pr(A = a_j|\, X) = \Pr(A = a_j)$). Note also that D is a degenerate random variable given a unit's compliance strata and encouragement.

Equation (4) indicates that the frequency of service among the *unencouraged* coincides with the frequency of always-takers in the population. This makes sense as compliers don't serve unless encouraged, while never-takers never serve, hence any unit serving when *not* encouraged must be an always-taker. By random assignment the frequency of never-takers, compliers and always-takers is the same in the $X = 0$ and $X = 1$ populations. Result (4) follows.

Together (4) and (5) give:

$$\begin{aligned} \Pr(D = 1|\, X = 0) &= \pi_3 \\ \Pr(D = 1|\, X = 1) - \Pr(D = 1|\, X = 0) &= \pi_2 \\ 1 - \Pr(D = 1|\, X = 1) &= \pi_1. \end{aligned}$$

The the population shares of the three compliance strata are identified.

Next consider the distribution of earnings among individuals who serve, but were no encouraged to do so. From our discussion above we know that this group consists solely of always-takers:

$$\begin{aligned}\Pr\left(Y \leq y\right| D=1, X=0) &= \Pr\left(Y \leq y\right| D=1, X=0, A=a_3) \\ &= \Pr\left(Y\left(1\right) \leq y\right| D\left(0\right)=1, X=0, A=a_3) \\ &= \Pr\left(Y\left(1\right) \leq y\right| A=a_3). \end{aligned} \tag{6}$$

The first equality follows because the event $D = 1$ and $X = 0$ coincides with the event "the individual is an always-takers". The second equality follows from (2) and (3), while the third is an implication of random assignment. The distribution of potential earnings under military service for always-takers is identified by the observed distribution of earnings among unencouraged veterans.

An entirely analogous argument gives

$$\Pr\left(Y \leq y\right| D=0, X=1) = \Pr\left(Y\left(0\right) \leq y\right| A=a_1). \tag{7}$$

The distribution of potential earnings under military service for never-takers is identified by the observed distribution of earnings among encouraged non-veterans.

To derive the distribution of potential earnings without service among compliers we again use the law of total probability.

$$\begin{aligned}\Pr\left(Y \leq y\right| X=0) =& \Pr\left(A=a_1\right| X=0) \Pr\left(Y \leq y\right| X=0, A=a_1) \\ &+ \Pr\left(A=a_2\right| X=0) \Pr\left(Y \leq y\right| X=0, A=a_2) \\ &+ \Pr\left(A=a_3\right| X=0) \Pr\left(Y \leq y\right| X=0, A=a_3) \\ =& \pi_1 \Pr\left(Y\left(0\right) \leq y\right| A=a_1) + \pi_2 \Pr\left(Y\left(0\right) \leq y\right| A=a_2) \\ &+ \pi_3 \Pr\left(Y\left(1\right) \leq y\right| A=a_1). \end{aligned} \tag{8}$$

From (4),(5), and (8) we get

$$\begin{aligned}\Pr\left(Y\left(0\right) \leq y\right| A=a_2) =& \frac{\Pr\left(Y \leq y\right| X=0)}{\Pr\left(D=1\right| X=1) - \Pr\left(D=1\right| X=0)} \\ &- \frac{\left[1-\Pr\left(D=1\right| X=1)\right]}{\Pr\left(D=1\right| X=1) - \Pr\left(D=1\right| X=0)} \Pr\left(Y \leq y\right| D=0, X=1) \\ &- \frac{\left[\Pr\left(D=1\right| X=0)\right]}{\Pr\left(D=1\right| X=1) - \Pr\left(D=1\right| X=0)} \Pr\left(Y \leq y\right| D=1, X=0). \end{aligned} \tag{9}$$

Analogous arguments give

$$\begin{aligned}\Pr\left(Y\left(1\right)\leq y\right|A=a_{2}) =&\frac{\Pr\left(Y\leq y\right|X=1)}{\Pr\left(D=1\right|X=1)-\Pr\left(D=1\right|X=0)} \\ &-\frac{\left[1-\Pr\left(D=1\right|X=1)\right]}{\Pr\left(D=1\right|X=1)-\Pr\left(D=1\right|X=0)}\Pr\left(Y\leq y\right|D=0,X=1) \\ &-\frac{\left[\Pr\left(D=1\right|X=0)\right]}{\Pr\left(D=1\right|X=1)-\Pr\left(D=1\right|X=0)}\Pr\left(Y\leq y\right|D=1,X=0).\end{aligned} \tag{10}$$

Hence the distribution of both $Y(0)$ and $Y(1)$ are identified for the sub-population of compliers; a result due to Imbens & Rubin (1997).

The difference in the marginal distribution of the potential outcome under treatment, $Y(1)$, versus that under control, $Y(0)$, is also identified for the subpopulation of compliers. A consequence is that the **local average treatment effect** (LATE)

$$\beta^{\mathrm{LATE}}=\mathbb{E}\left[Y\left(1\right)-Y\left(0\right)\right|A=a_{2}]=\frac{\mathbb{E}\left[Y\right|X=1]-\mathbb{E}\left[Y\right|X=0]}{\Pr\left(D=1\right|X=1)-\Pr\left(D=1\right|X=0)} \tag{11}$$

is also identified. Since the expression to the right of the equality equals $\mathbb{C}\left(Y,X\right)/\mathbb{C}\left(D,X\right)$ the LATE also coincides with the probability limit of the coefficient on D in the instrumental variables fit of Y onto a constant and D using X as an instrument for D (Imbens & Angrist, 1994; Angrist et al., 1996).

In the empirical economics literature, following Imbens & Angrist (1994), the LATE is typically estimated by the sample analog of the right-hand-side of (11), or the so-called WALD-IV estimator. This is convenient and simple, however it also has a number of problems. In applications researchers typically need to condition on additional pre-treatment covariates in order to justify the random assignment assumption. Let W be a vector of such covariates where we are willing to assume **conditional random assignment**:

$$\left(D\left(0\right),D\left(1\right),Y\left(0\right),Y\left(1\right)\right)\perp X|\,W=w,\ w\in\mathbb{W}. \tag{12}$$

A direct extension of the argument given above yields

$$\begin{aligned}\beta^{\mathrm{LATE}}\left(w\right)&=\mathbb{E}\left[Y\left(1\right)-Y\left(0\right)\right|A=a_{2},W=w] \\ &=\frac{\mathbb{E}\left[Y\right|X=1,W=w]-\mathbb{E}\left[Y\right|X=0,W=w]}{\Pr\left(D=1\right|X=1,W=w)-\Pr\left(D=1\right|X=0,W=w)}.\end{aligned}$$

A conditional version of the WALD-IV estimator identifies a conditional LATE. Unfortunately, when W is high-dimensional, analog estimation based upon the right-hand-side of the above expression is not straightforward (cf., Frölich, 2007; Ogburn et al., 2015).

In practice researchers often proceed by computing the linear IV fit of Y onto a constant, D, and W with X serving as an excluded instrument for D. This IV fit does not identify the LATE except under rather special and implausible functional form assumptions; nevertheless it is common for researchers to proceed (incorrectly) as if it identifies the LATE. See Imbens & Rubin (2015) for additional discussion.

Although we will not emphasize the role of additional covariates here, one advantage of the likelihood based estimator outline below is that it can easily be adapted to include pre-treatment covariates. While the likelihood based approach does require additional modeling assumptions, it further leads to more precise inference when these assumptions are true. Likelihood based IV methods are due to Imbens & Rubin (1997). While rarely used by economists, they are used with some regularity in applied statistics.

Likelihood

Our goal is to write down a likelihood for (D, Y) given X. To do this we will first write down a likelihood for (A, D, Y) given X. Will will then treat a unit's compliance strata as missing data. We will call the latter likelihood the **complete data likelihood**, the former the observed or **integrated likelihood**.

We begin by observing that

$$\begin{aligned} f\left(A, D, Y \mid X\right) &= f\left(D, Y \mid X, A\right) f\left(A \mid X\right) \\ &= f\left(D, Y \mid X, A\right) f\left(A\right). \end{aligned}$$

The second equality follows by random assignment of X.

We will assume that the distributions of potential (log) earnings given compliance strata are Gaussian, with strata-specific location and scale parameters (other parametric assumptions could be made here depending on the outcome of interest). We have that the distribution of potential earnings given non-service among never-takers is

$$Y(0) \mid A = a_1 \sim \mathcal{N}\left(\mu_{N0}, \sigma_{N0}^2\right).$$

Since never-takers never serve the distribution of $Y(1) \mid A = a_1$ is undefined. For compliers we do need to define both potential earning distributions:

$$Y(0)|\, A = a_2 \sim \mathcal{N}\left(\mu_{C0}, \sigma^2_{C0}\right).$$
$$Y(1)|\, A = a_2 \sim \mathcal{N}\left(\mu_{C1}, \sigma^2_{C1}\right).$$

Finally, for always-takers we need only the $Y(1)$ distribution:

$$Y(1)|\, A = a_3 \sim \mathcal{N}\left(\mu_{A1}, \sigma^2_{A1}\right).$$

Let $\phi(y; \mu, \sigma^2)$ be the density of a $\mathcal{N}(\mu, \sigma^2)$ random variable at $Y = y$. With these assumptions we compute conditional densities for service and earnings across the three compliance strata of

$$\begin{aligned}
f(D, Y|\, X, A = a_1; \delta) &= \phi\left(Y; \mu_{N0}, \sigma^2_{N0}\right)^{(1-D)(1-X)} \phi\left(Y; \mu_{N0}, \sigma^2_{N0}\right)^{(1-D)X} 0^{D(1-X)} 0^{DX} \\
f(D, Y|\, X, A = a_2; \delta) &= \phi\left(Y; \mu_{C0}, \sigma^2_{C0}\right)^{(1-D)(1-X)} 0^{(1-D)X} 0^{D(1-X)} \phi\left(Y; \mu_{C1}, \sigma^2_{C1}\right)^{DX} \\
f(D, Y|\, X, A = a_3; \delta) &= 0^{(1-D)(1-X)} 0^{(1-D)X} \phi\left(Y; \mu_{A1}, \sigma^2_{A1}\right)^{D(1-X)} \phi\left(Y; \mu_{A1}, \sigma^2_{A1}\right)^{DX}.
\end{aligned}$$

You should check that the above densities equal zero for unsupported events.

With the above notation we get a complete data log-likelihood contribution for unit i of

$$\ln L_i^C(\theta, A_i) = \sum_{j=1}^{3} A_{ji} \left\{\ln f(D_i, Y_i|\, X_i, A_i = a_j; \delta) + \ln \pi_j\right\}. \tag{13}$$

where we let $\delta = (\mu_{N0}, \sigma^2_{N0}, \mu_{C0}, \sigma^2_{C0}, \mu_{C1}, \sigma^2_{C1}, \mu_{N0}, \sigma^2_{N0})'$ and $\pi = (\pi_1, \pi_2, \pi_3)'$ with $\theta = (\delta', \pi')'$.

Complete data analysis

It is instructive to first consider an analysis which presumes compliance strata are observed. Assume we have access to the random sample $\{A_i, D_i, Y_i, X_i\}_{i=1}^{N}$. Since the compliance strata population shares sum to one we maximize the Lagrangian

$$\mathcal{L} = \sum_{i=1}^{N} \sum_{j=1}^{3} A_{ji} \left\{\ln f(D_i, Y_i|\, X_i, A_i = a_j; \delta) + \ln \pi_j\right\} + \lambda(1 - \pi_1 - \pi_2 - \pi_3)$$

yielding MLEs of

$$\hat{\pi}_1 = \frac{1}{N}\sum A_{1i},\ \hat{\pi}_2 = \frac{1}{N}\sum A_{2i},\ \hat{\pi}_3 = \frac{1}{N}\sum A_{3i}$$

and

$$\begin{aligned}
\hat{\mu}_{N0} &= \frac{\sum_{i=1}^{N} A_{1i}\left\{(1-D_i)(1-X_i)+(1-D_i)X_i\right\}Y_i}{\sum_{i=1}^{N} A_{1i}\left\{(1-D_i)(1-X_i)+(1-D_i)X_i\right\}}, \\
\hat{\sigma}^2_{N0} &= \frac{\sum_{i=1}^{N} A_{1i}\left\{(1-D_i)(1-X_i)+(1-D_i)X_i\right\}(Y_i-\hat{\mu}_{N0})^2}{\sum_{i=1}^{N} A_{1i}\left\{(1-D_i)(1-X_i)+(1-D_i)X_i\right\}}, \\
\hat{\mu}_{C0} &= \frac{\sum_{i=1}^{N} A_{2i}(1-D_i)(1-X_i)Y_i}{\sum_{i=1}^{N} A_{2i}(1-D_i)(1-X_i)},\ \hat{\sigma}^2_{C0} = \frac{\sum_{i=1}^{N} A_{2i}(1-D_i)(1-X_i)(Y_i-\hat{\mu}_{C0})^2}{\sum_{i=1}^{N} A_{2i}(1-D_i)(1-X_i)}, \\
\hat{\mu}_{C1} &= \frac{\sum_{i=1}^{N} A_{2i}D_iX_iY_i}{\sum_{i=1}^{N} A_{2i}D_iX_i},\ \hat{\sigma}^2_{C1} = \frac{\sum_{i=1}^{N} A_{2i}D_iX_i(Y_i-\hat{\mu}_{C1})^2}{\sum_{i=1}^{N} A_{2i}D_iX_i}, \\
\hat{\mu}_{A1} &= \frac{\sum_{i=1}^{N} A_{3i}\left\{D_i(1-X_i)+D_iX_i\right\}Y_i}{\sum_{i=1}^{N} A_{3i}\left\{D_i(1-X_i)+D_iX_i\right\}}, \\
\hat{\sigma}^2_{A1} &= \frac{\sum_{i=1}^{N} A_{3i}\left\{D_i(1-X_i)+D_iX_i\right\}(Y_i-\hat{\mu}_{A1})^2}{\sum_{i=1}^{N} A_{3i}\left\{D_i(1-X_i)+D_iX_i\right\}}.
\end{aligned} \tag{14}$$

Consider the mean outcome for compliers under military service, μ_{C1}. If knowledge of the compliance strata were available we would estimate this by the average outcome among compliers who served in the military. Why is this okay? Within a compliance strata the only reason why one unit serves and another doesn't is because those units received different encouragements (in this case a different draft lottery number). Since encouragement is randomly assigned, then so is actual treatment *within a compliance strata*. Encouraged compliers (who serve) and non-encouraged compliers (who do not serve) are fully comparable. Therefore

$$\hat{\beta}^{\text{LATE}} = \hat{\mu}_{C1} - \hat{\mu}_{C1}$$

provides a consistent estimate of the LATE, or the average effect of military service on earnings among compliers. Note we cannot estimate an average effect for never-takers or always-takers, since in these two strata only one of the two potential outcome distributions is identified. The encouragement/instrument does not generate exogenous variation in treatment status in these two strata.

Incomplete data analysis

In practice compliance strata are unobserved. Unit i's contribution to the **integrated or observed log-likelihood** is computed by "averaging over" the compliance strata distribution:

$$l_i^I(\theta) = \ln\left(\sum_{l=1}^{K} L_i^C(\theta, a_l)\right).$$

The integrated log-likelihood for the entire sample is then

$$l_N^I(\theta) = \sum_{i=1}^{N} l_i^I(\theta). \tag{15}$$

We will use the EM-Algorithm to maximize (15) with respect to θ.

Let $\tilde{A}_{1i} \in [0,1]$ be the subjective probability the econometrician attaches to the statement "unit i is a never-taker" being true. For the time being assume that the parameter $\theta = (\delta', \pi')'$ is known. Our likelihood, indexed by θ, and the data can be used – via Bayes' rule – to compute the probability that unit i belongs to a certain compliance strata.

The probability that "unit i is a never-taker" given that he did not serve *and* was encouraged to do so is 1:

$$\Pr\left(A_i = a_1 \middle| D_i = 0, Y_i, X_i = 1; \theta\right) = 1. \tag{16}$$

This follows because always-takers always serve, and compliers served if encouraged, hence any encouraged unit *not* serving is definitely a never-taker. The situation is a bit more complicated for non-serving units who were not encouraged. An unencouraged unit who did not serve could be either be a complier or a never-taker. To compute the probability of the latter we use Bayes' rule:

$$\Pr\left(A_i = a_1 \middle| D_i = 0, Y_i, X_i = 0; \theta\right) = \frac{f\left(D_i = 0, Y_i \middle| X_i = 0, A_i = a_1; \theta\right) \Pr\left(A_i = a_1 \middle| X_i = 0; \theta\right)}{f\left(D_i = 0, Y_i \middle| X_i = 0; \theta\right)}.$$

We have that $\Pr\left(A_i = a_1 \middle| X_i = 0; \theta\right) = \pi_1$ and $f\left(D_i = 0, Y_i \middle| X_i = 0, A_i = a_1; \theta\right) = \phi\left(Y_i; \mu_{N0}, \sigma_{N0}^2\right)$. This gives the numerator in the expression above.

Since non-serving units who were not encouraged are a mixture of never-takers and compliers, the marginal density $f\left(D_i = 0, Y_i \middle| X_i = 0; \theta\right)$ equals

$$f\left(D_i = 0, Y_i \middle| X_i = 0; \theta\right) = \pi_1 \phi\left(Y_i; \mu_{N0}, \sigma_{N0}^2\right) + \pi_2 \phi\left(Y_i; \mu_{C0}, \sigma_{C0}^2\right).$$

To verify this claim use the identity

$$f\left(D_i=0,Y_i\middle|\,X_i=0;\theta\right)=\sum_{j=1}^{3}f\left(D_i=0,Y_i\middle|\,X_i=0,A_i=a_j;\theta\right)\Pr\left(A_i=a_j\middle|\,X_i=0;\theta\right)$$

and also note that since always-takers always serve $f\left(D_i=0,Y_i\middle|\,X_i=0,A_i=a_3;\theta\right)=0$. Putting these results together yields

$$\Pr\left(A_i=a_1\middle|\,D_i=0,Y_i,X_i=0;\theta\right)=\frac{\pi_1\phi\left(Y_i;\mu_{N0},\sigma_{N0}^2\right)}{\pi_1\phi\left(Y_i;\mu_{N0},\sigma_{N0}^2\right)+\pi_2\phi\left(Y_i;\mu_{C0},\sigma_{C0}^2\right)}. \tag{17}$$

From (16) and (17) we get a posterior probability of unit i being a never taker of

$$\tilde{A}_{1i}=\left(1-X_i\right)\left(1-D_i\right)\frac{\pi_1\phi\left(Y_i;\mu_{N0},\sigma_{N0}^2\right)}{\pi_1\phi\left(Y_i;\mu_{N0},\sigma_{N0}^2\right)+\pi_2\phi\left(Y_i;\mu_{C0},\sigma_{C0}^2\right)}+X_i\left(1-D_i\right). \tag{18}$$

Note that (18) evaluates to zero whenever $D_i=1$ (since never-takers never serve in the military). Similarly if $X_i=1$ and $D_i=0$ it evaluates to one, since an encouraged unit who does not serve is a never-taker with probability one. The interesting case is when both X_i and D_i equal zero; units in this group constitute a mixture of never takers and compliers.

Similar arguments yield

$$\begin{aligned}\tilde{A}_{2i}=&\left(1-X_i\right)\left(1-D_i\right)\frac{\pi_2\phi\left(Y_i;\mu_{C0},\sigma_{C0}^2\right)}{\pi_1\phi\left(Y_i;\mu_{N0},\sigma_{N0}^2\right)+\pi_2\phi\left(Y_i;\mu_{C0},\sigma_{C0}^2\right)}\\&+X_iD_i\frac{\pi_2\phi\left(Y_i;\mu_{C1},\sigma_{C1}^2\right)}{\pi_2\phi\left(Y_i;\mu_{C1},\sigma_{C1}^2\right)+\pi_3\phi\left(Y_i;\mu_{A1},\sigma_{A1}^2\right)}.\end{aligned} \tag{19}$$

and

$$\tilde{A}_{3i}=\left(1-X_i\right)D_i+X_iD_i\frac{\pi_3\phi\left(Y_i;\mu_{A1},\sigma_{A1}^2\right)}{\pi_2\phi\left(Y_i;\mu_{C1},\sigma_{C1}^2\right)+\pi_3\phi\left(Y_i;\mu_{A1},\sigma_{A1}^2\right)}. \tag{20}$$

It is a good exercise to verify that expressions (19) and (20) are correct.

Since A_i is unobserved we can not evaluate unit i's contribution to the complete data log-likelihood (13). However, using (18), (19) and (20), we can evaluate this unit's expected log-likelihood contribution:

$$\mathbb{E}\left[\ln L_i^C\left(\theta,A_i\right)\middle|\,X_i,D_i,Y_i;\hat{\theta}^{(s)}\right]=\sum_{j=1}^{3}\tilde{A}_{ji}\left\{\ln f\left(D_i,Y_i\middle|\,X_i,A_i=a_j;\delta\right)+\ln\pi_j\right\} \tag{21}$$

Note that the value of θ used to compute the above expectation, which involves an average over the posterior distribution of A_i, may differ from the value of θ at which the expected log-likelihood is evaluated. We will use $\hat{\theta}^{(s)}$ to denote the former and θ to denote the latter. Summing over all $i = 1, \ldots, N$ units we get a criterion function which is proportional to something called the Q-function (defined carefully below):

$$Q\left(\theta, \hat{\theta}^{(s)}\right) \propto \sum_{i=1}^{N} \sum_{j=1}^{3} \tilde{A}_{ji} \left\{ \ln f\left(\left. D_i, Y_i \right| X_i, A_i = a_j; \delta\right) + \ln \pi_j \right\} \tag{22}$$

Maximizing (22) with respect to θ yields

$$\hat{\pi}_1^{(s+1)} = \frac{1}{N} \sum \tilde{A}_{1i},\ \hat{\pi}_2^{(s+1)} = \frac{1}{N} \sum \tilde{A}_{2i},\ \hat{\pi}_3^{(s+1)} = \frac{1}{N} \sum \tilde{A}_{3i}$$

and

$$\begin{aligned}
\hat{\mu}_{N0}^{(s+1)} &= \frac{\sum_{i=1}^{N} \tilde{A}_{1i} \left\{ (1-D_i)(1-X_i) + (1-D_i) X_i \right\} Y_i}{\sum_{i=1}^{N} \tilde{A}_{1i} \left\{ (1-D_i)(1-X_i) + (1-D_i) X_i \right\}}, \\
\left(\hat{\sigma}_{N0}^2\right)^{(s+1)} &= \frac{\sum_{i=1}^{N} \tilde{A}_{1i} \left\{ (1-D_i)(1-X_i) + (1-D_i) X_i \right\} \left(Y_i - \hat{\mu}_{N0}^{(s+1)}\right)^2}{\sum_{i=1}^{N} \tilde{A}_{1i} \left\{ (1-D_i)(1-X_i) + (1-D_i) X_i \right\}}, \\
\hat{\mu}_{C0}^{(s+1)} &= \frac{\sum_{i=1}^{N} \tilde{A}_{2i} (1-D_i)(1-X_i) Y_i}{\sum_{i=1}^{N} \tilde{A}_{2i} (1-D_i)(1-X_i)},\ \left(\hat{\sigma}_{C0}^2\right)^{(s+1)} = \frac{\sum_{i=1}^{N} \tilde{A}_{2i} (1-D_i)(1-X_i) \left(Y_i - \hat{\mu}_{C0}^{(s+1)}\right)^2}{\sum_{i=1}^{N} \tilde{A}_{2i} (1-D_i)(1-X_i)}, \\
\hat{\mu}_{C1}^{(s+1)} &= \frac{\sum_{i=1}^{N} \tilde{A}_{2i} D_i X_i Y_i}{\sum_{i=1}^{N} \tilde{A}_{2i} D_i X_i},\ \left(\hat{\sigma}_{C1}^2\right)^{(s+1)} = \frac{\sum_{i=1}^{N} \tilde{A}_{2i} D_i X_i \left(Y_i - \hat{\mu}_{C1}^{(s+1)}\right)^2}{\sum_{i=1}^{N} \tilde{A}_{2i} D_i X_i}, \\
\hat{\mu}_{A1}^{(s+1)} &= \frac{\sum_{i=1}^{N} \tilde{A}_{3i} \left\{ D_i (1-X_i) + D_i X_i \right\} Y_i}{\sum_{i=1}^{N} \tilde{A}_{3i} \left\{ D_i (1-X_i) + D_i X_i \right\}}, \\
\left(\hat{\sigma}_{A1}^2\right)^{(s+1)} &= \frac{\sum_{i=1}^{N} \tilde{A}_{3i} \left\{ D_i (1-X_i) + D_i X_i \right\} \left(Y_i - \hat{\mu}_{A1}^{(s+1)}\right)^2}{\sum_{i=1}^{N} \tilde{A}_{3i} \left\{ D_i (1-X_i) + D_i X_i \right\}}.
\end{aligned}$$

These expressions are identical to those derived in the complete data case after replacing stratum indicator A_{1i} with the posterior probability $\tilde{A}_{1i}$ and so on.

The above discussion suggests the following estimation algorithm.

1. Let $\hat{\theta}^{(s)}$ for $s = 0$ be an initial value for the parameter θ.
2. **E-Step:** Compute the expected log-likelihood given the data and current parameter value $\hat{\theta}^{(s)}$ according to equation (22).
3. **M-Step:** Maximize $Q\left(\theta, \hat{\theta}^{(s)}\right)$ with respect to θ. Denote the solution to this maximization problem by $\hat{\theta}^{(s+1)}$.

4. Repeat steps 2 and 3 until $Q\left(\hat{\theta}^{(s+1)},\hat{\theta}^{(s)}\right) \approx Q\left(\hat{\theta}^{(s)},\hat{\theta}^{(s-1)}\right)$ is small and/or until $\hat{\theta}^{(s+1)} \approx \hat{\theta}^{(s)}$.

Both the E- and M-steps have closed-form solutions in the current setting.

Theory of the EM Algorithm

Let $q(a)$ be some assignment of probability mass to the K possible types such that $q(a_k) > 0$ for all $k = 1, \ldots, K$ and $\sum_{k=1}^{K} q(a_k) = 1$. Here "types" correspond to the three compliance strata so that $K = 3$.

We can show that the i^{th} unit's contribution to the observed log likelihood is bounded below by

$$\begin{aligned}\ln\left(\sum_{l=1}^{K} L_i^C(\theta, a_l)\right) &= \ln\left(\sum_{l=1}^{K} q(a_l)\frac{L_i^C(\theta, a_l)}{q(a_l)}\right) \\ &\geq \sum_{l=1}^{K} q(a_l)\ln\left(\frac{L_i^C(\theta, a_l)}{q(a_l)}\right) \\ &= Q_i^*(\theta, q) \qquad (23)\end{aligned}$$

where the middle line follows from Jensen's inequality: $g(\mathbb{E}[Y]) \geq \mathbb{E}[g(Y)]$ for $g(\cdot)$ concave. Here $\ln(\cdot)$ is concave and expectations are with respect to $q(a)$. The last line defines $Q_i^*(\theta, q)$ Equation (23) gives

$$l_N^I(\theta) \geq \sum_{i=1}^{N} Q_i^*(\theta, q_i)$$

for any set of valid distribution functions $\{q_i\}_{i=1}^{N}$ that assign positive probability to each $\{a_k\}_{k=1}^{K}$.

Bayes' Theorem, and the form of the complete data likelihood (13), yields the conditional type probabilities

$$\Pr(A = a_k|\, X_i, D_i, Y_i; \theta) \stackrel{def}{\equiv} \tilde{A}_{ki}(\theta) = \frac{L_i^C(\theta, a_k)}{\sum_{l=1}^{K} L_i^C(\theta, a_l)}, \qquad (24)$$

for $k = 1, \ldots, K$. In machine learning literature on classification (24) is called the "responsibility" of cluster k for unit i. We can use (24) to factor the i^{th} unit's contribution to the complete data likelihood as

$$L_i^C(\theta; a_k) = \tilde{A}_{ki}(\theta)\left[\sum_{l=1}^{K} L_i^C(\theta, a_l)\right].$$

This gives a re-arrangement of the lower bound (23) equal to

$$\begin{aligned} Q_i^*(\theta, q_i) &= \sum_{l=1}^{K} q_i(a_l) \ln\left(\frac{L_i^C(\theta, a_l)}{q_i(a_l)}\right) \\ &= \sum_{l=1}^{K} q_i(a_l) \ln\left(\frac{\tilde{A}_{li}(\theta)\left[\sum_{m=1}^{K} L_i^C(\theta, a_m)\right]}{q_i(a_l)}\right) \\ &= -D_{\mathrm{KL}}\left(q_i \| \tilde{A}_i\right) + \left[\sum_{l=1}^{K} q_i(a_l)\right] \ln\left(\sum_{m=1}^{K} L_i^C(\theta, a_m)\right) \\ &= -D_{\mathrm{KL}}\left(q_i \| \tilde{A}_i\right) + \ln\left(\sum_{m=1}^{K} L_i^C(\theta, a_m)\right), \end{aligned} \tag{25}$$

where $D_{\mathrm{KL}}\left(q_i \| \tilde{A}_i\right) = \sum_{l=1}^{K} q_i(a_l) \ln\left(\frac{q_i(a_l)}{\tilde{A}_{li}(\theta)}\right)$ is the Kullback-Leibler divergence of $\tilde{A}_i$ from q_i.

Now consider, once again, the our maximization procedure:

1. Let $\hat{\theta}^{(s)}$ for $s = 0$ be an initial value for θ.

2. **E-Step:** Set $q(a_k) = \tilde{A}_{ki}\left(\hat{\theta}^{(s)}\right)$ for $k = 1, \ldots, K$ and form the observed log-likelihood lower bound

$$\begin{aligned} Q\left(\theta, \hat{\theta}^{(s)}\right) &= \sum_{i=1}^{N} Q_i^*\left(\theta, \tilde{A}_i\left(\hat{\theta}^{(s)}\right)\right) \\ &= \sum_{i=1}^{N} \left\{\sum_{l=1}^{K} \tilde{A}_{li}\left(\hat{\theta}^{(s)}\right) \ln\left(L_i^C(\theta, a_l)\right) + \mathbf{S}\left(\tilde{A}_i\right)\right\} \\ &= \sum_{i=1}^{N} \left\{\mathbb{E}\left[\ln\left(L_i^C(\theta, A)\right) \middle| X_i, D_i, Y_i; \hat{\theta}^{(s)}\right] + \mathbf{S}\left(\tilde{A}_i\right)\right\} \end{aligned} \tag{26}$$

where $\mathbb{E}\left[\ln\left(L_i^C(\theta, A)\right) \middle| X_i, D_i, Y_i; \hat{\theta}^{(s)}\right]$ is the expected value of the i^{th} unit's contribution to the complete data log-likelihood (given her observed encouragement, service choice and earnings, and the current parameter value $\hat{\theta}^{(s)}$) and $\mathbf{S}(q) = -\sum_l q_l \ln q_l$ is the entropy of q.

3. **M-Step**: Choose $\hat{\theta}^{(s+1)}$ to maximize $Q\left(\theta, \hat{\theta}^{(s)}\right)$ with respect to θ. Note that since $\mathbf{S}\left(\tilde{A}_i\right)$ is constant in θ this term is often omitted from the "Q-function" in practice (as was done above).

4. Repeat steps 2 and 3 until $Q\left(\hat{\theta}^{(s+1)}, \hat{\theta}^{(s)}\right) \approx Q\left(\hat{\theta}^{(s)}, \hat{\theta}^{(s-1)}\right)$ is small and/or until $\hat{\theta}^{(s+1)} \approx \hat{\theta}^{(s)}$.

Note that $Q(\theta, \theta) = l_N^I(\theta)$: after the E-Step the "Q-function" coincides with the observed log-likelihood (the Kullback-Leibler term is zero at $q_i = \tilde{A}_i$). We also have that the M-Step weakly increases the "Q-function". Putting things together we have

$$l_N^I\left(\hat{\theta}^{(s+1)}\right) \geq Q\left(\hat{\theta}^{(s+1)}, \hat{\theta}^{(s)}\right) \geq Q\left(\hat{\theta}^{(s)}, \hat{\theta}^{(s)}\right) = l_N^I\left(\hat{\theta}^{(s)}\right). \tag{27}$$

From left-to-right the first inequality follows from (23), the second from the definition of maximization, and the third from (25) evaluated at $q_i = \tilde{A}_i(\theta)$. By virtue of (27) the observed log-likelihood $Q\left(\hat{\theta}^{(s)}, \hat{\theta}^{(s)}\right) = l_N^I\left(\hat{\theta}^{(s)}\right)$ is monotonically increasing in s. The EM algorithm will therefore find a local maximum (or saddle point) of the *observed log-likelihood* (15). Running the algorithm from a variety of starting points is advised as there may be multiple local maxima.

Further reading

The estimator introduced here is due to Imbens & Rubin (1997). While used frequently in applied statistics, its use in economics is less common, where method-of-moments estimators predominate (Angrist et al., 1996). One advantage of the model-based approach is the ease with which pre-treatment covariates are introduced into the analysis in a coherent way; something that is more difficult with the method-of-moments approach (cf., Imbens & Rubin, 1997). Non-parametric and semi-parametric LATE estimators, allowing for pre-treatment covariates, are discussed by Frölich (2007) and Ogburn et al. (2015).

Murphy (2012, Chapter 12) provides a elementary introduction to the EM algorithm from a machine learning perspective. Gupta & Chen (2010) provide a survey with signal processing applications. Ruud (1991) provides a nice theoretical discussion with applications to discrete choice models common in econometrics. The EM algorithm has numerous applications in econometrics, particularly for discrete heterogeneity modeling in duration analysis and panel data analysis.

References

Angrist, J. D. (1990). Lifetime earnings and the vietnam era draft lottery: evidence from social security administrative records. the. *American Economic Review*, 80(3), 313 – 336.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434), 444 – 455.

Frölich, M. (2007). Nonparametric iv estimation of local average treatment effects with covariates. *Journal of Econometrics*, 139(1), 35 – 75.

Gupta, M. R. & Chen, Y. (2010). Theory and use of the em algorithm. *Foundations and Trends in Signal Processing*, 4(3), 223 – 296.

Imbens, G. W. & Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62(2), 467 – 475.

Imbens, G. W. & Rubin, D. B. (1997). Estimating outcome distributions for compliers in instrumental variable models. *Review of Economic Studies*, 64(4), 555 – 574.

Imbens, G. W. & Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge: Cambridge University Press.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: The MIT Press.

Ogburn, E. L., Rotnitzky, A., & Robins, J. M. (2015). Doubly robust estimation of the local average treatment effect curve. *Journal of the Royal Statistical Society: Series B*, 77(2), 373 – 396.

Ruud, P. A. (1991). Extensions of estimation methods using the em algorithm. *Journal of Econometrics*, 49(3), 305 – 341.

