

**Economics 696F, Causal Inference and Program Evaluation**  
**Lecture Note 11: Imbens and Newey (2004)**

Consider the following model:

$$T = h(Z, \eta), \tag{1}$$

$$Y = g(T, \epsilon). \tag{2}$$

We interpret  $Y$  as the outcome, and  $T$  as the (observed) treatment.

The functions  $h$  and  $g$  are unknown, so the model is “nonparametric.” We do not assume that  $\eta$  and  $\epsilon$  are independent, so that in general  $T$  is related to  $\epsilon$ . That is,  $T$  is “endogenous.” Unlike some of the previous models we have seen, the disturbance term is not additively separable. As a consequence, individual heterogeneity can interact with the “treatment” in determining the outcome. Provided that (a) the distribution of  $\epsilon$  is sufficiently rich, and (b) the function  $g$  is sufficiently flexible, we can get a very wide range of potential outcome functions.

We can connect this to our potential outcomes notation:

$$Y_i(t) = g(t, \epsilon_i), \quad t \in \mathcal{T}.$$

Since  $t$  can interact with  $\epsilon_i$ , treatment effects like

$$Y_i(t_1) - Y_i(t_0) = g(t_1, \epsilon_i) - g(t_0, \epsilon_i)$$

are generally functions of  $\epsilon_i$ , hence individual-specific.

We make some further assumptions:<sup>1</sup>

1.  $\epsilon$  and  $\eta$  are scalar and continuously distributed.
2.  $h(z, \eta)$  is strictly increasing in  $\eta$ .
3.  $g(t, \epsilon)$  is strictly increasing in  $\epsilon$ .

The strict monotonicity of  $h$  and  $g$  in the disturbance terms does imply some restrictions on the potential outcomes functions, although for continuous outcomes they are not testable in the sense of implying observable restrictions on the data.

Instrumental Variable

$$Z \perp (\eta, \epsilon).$$

Unlike the linear IV case, we need full independence rather than zero covariance to be able to estimate interesting quantities.

---

<sup>1</sup>Some of these can be relaxed slightly; see Imbens and Newey (2004). We are making slightly stronger assumptions than necessary to simplify the exposition.

Next, we need to define what quantities we are interested in estimating:

Average Conditional Response (ACR):

$$\begin{aligned}\beta(t, \eta) &:= E[g(t, \epsilon) | \eta] \\ &= \int g(t, \epsilon) dF_{\epsilon | \eta}(\epsilon | \eta).\end{aligned}\tag{3}$$

Note: here I am using  $F_{\cdot | \cdot}(\cdot | \cdot)$  to denote the CDF of the relevant conditional distribution. The notation  $\int k(w) dF(w)$  means the integral with respect to the distribution  $F(w)$ . It may help to think of  $dF(w)$  as standing for  $f(w)dw$ .

Although the ACR might not be of direct interest, it is useful in obtaining a number of meaningful quantities:

Average Structural Function (ASF):

$$\mu(t) := E[g(t, \epsilon)] = \int g(t, \epsilon) dF_{\epsilon}(\epsilon).$$

Useful fact:

$$\begin{aligned}\mu(t) &= \int \int g(t, \epsilon) dF_{\epsilon | \eta}(\epsilon | \eta) dF_{\eta}(\eta) \\ &= \int \beta(t, \eta) dF_{\eta}(\eta).\end{aligned}\tag{4}$$

Average Marginal Productivity:

$$\begin{aligned}E \left[ \frac{\partial g(T, \epsilon)}{\partial t} \right] &= E \left[ E \left[ \frac{\partial g(T, \epsilon)}{\partial t} \middle| \eta \right] \right] \\ &= E \left[ \int \frac{\partial g(T, \epsilon)}{\partial t} dF_{\epsilon | \eta}(\epsilon | \eta) \right] \\ &= E \left[ \frac{\partial \beta(T, \eta)}{\partial t} \right],\end{aligned}\tag{5}$$

where the last equality follows by interchanging differentiation and integration.

### Identification

We want to show that the various quantities defined above, along with the functions  $g$  and  $h$ , are identified.

First, it is useful to note that since  $h(z, \eta)$  is strictly increasing in  $\eta$ , we can normalize  $\eta$  to have a marginal distribution that is uniform on  $[0, 1]$ . Likewise, we can normalize  $\epsilon$  to be uniform on  $[0, 1]$ .<sup>2</sup>

---

<sup>2</sup>To see why, note that for any random variable  $\eta$ ,  $v = F_{\eta}(\eta)$  is *Unif* $[0, 1]$ . So we can write  $h(z, \eta) = h(z, F_{\eta}^{-1}(v))$ .

**Theorem 1**  $\beta(t, \eta)$  is identified on the joint support of  $(T, \eta)$  from the joint distribution of  $(Y, T, Z)$ .

**Proof:** By the normalization on  $\eta$ ,  $Pr(\eta \leq c) = c$ . Consider the conditional CDF of  $T$  given  $Z$ :

$$\begin{aligned}
F_{T|Z}(t_0|z_0) &= Pr(T \leq t_0|Z = z_0) \\
&= Pr(h(Z, \eta) \leq t_0|Z = z_0) \\
&= Pr(\eta \leq h^{-1}(Z, t_0)|Z = z_0) \\
&= Pr(\eta \leq h^{-1}(z_0, t_0)|Z) \\
&= F_\eta(h^{-1}(z_0, t_0)) \\
&= h^{-1}(z_0, t_0).
\end{aligned}$$

Since  $F_{T|Y}$  is identified, so is  $h^{-1}$ , hence  $h$  is identified. If we know  $h^{-1}$ , then we can calculate

$$\eta = h^{-1}(Z, T) = F_{T|Z}(T|Z).$$

Next, since  $(\eta, \epsilon) \perp Z$ ,

$$\epsilon \perp Z|\eta \Rightarrow \epsilon \perp h(Z, \eta)|\eta \Rightarrow \epsilon \perp T|\eta.$$

Therefore

$$\begin{aligned}
\beta(t, \eta) &= E[g(t, \epsilon)|\eta] = E[g(T, \epsilon)|T = t, \eta] \\
&= E[Y|T = t, \eta] \\
&= E[Y|T = t, F_{T|Z}(T|Z) = \eta].
\end{aligned}$$

Thus, the joint distribution of  $(Y, T, Z)$  identifies  $\beta(t, \eta)$ . □

In order to identify the ASF  $\mu(t)$ , we need an additional assumption on the support of  $T$ :

Support Assumption: The support of  $T$  given  $\eta$  does not depend on the value of  $\eta$ .

This is a strong assumption.

**Theorem 2** Suppose the support assumption holds. Then the ASF  $\mu(t)$  is identified from the joint distribution of  $(Y, T, Z)$ .

**Proof:** Let  $\mathcal{T}$  denote the support of  $T$ . By the support assumption, the joint support of  $T, \eta$  is  $\mathcal{T} \times [0, 1]$ . The function  $\beta(t, \eta)$  is identified on this support by Theorem 1. Then

$$\begin{aligned}
\int_0^1 \beta(t, \eta) d\eta &= \int_0^1 \int g(t, \epsilon) dF_{\epsilon|\eta}(\epsilon|\eta) d\eta \\
&= \mu(t).
\end{aligned}$$

□

The argument for the next theorem is similar to the argument in Theorem 1:

**Theorem 3** *The joint distribution of  $(T, \eta, \epsilon)$  is identified, up to normalizations on  $\eta, \epsilon$ , and  $g(t, \epsilon)$  is identified on the support of  $(T, \epsilon)$ .*

### Estimation

The proofs of Theorems 1 and 2 are important, because they suggest how to estimate the functions  $\beta(t, \eta)$  and  $\mu(t)$ .

We saw in the proof of Theorem 1 that  $\eta = F_{T|Z}(T|Z)$ . So if we have an estimate of this conditional CDF, we can estimate the  $\eta_i$  by:

$$\hat{\eta}_i = \hat{F}_{T|Z}(T_i|Z_i).$$

Next, since  $\beta(t, \eta) = E[Y|T = t, \eta]$ , we can run a regression of  $Y_i$  on  $T_i$  and  $\hat{\eta}_i$  to estimate the  $\beta$  function:

$$\hat{\beta}(t, \eta) = \hat{E}[Y|T = t, \eta].$$

Note: the fact that we are using an estimated version of  $\eta$  in the regression will affect the asymptotic distribution.

To get an estimate of the average structural function  $\mu(t)$  we can follow the proof of Theorem 2:

$$\hat{\mu}(t) = \int_0^1 \hat{\beta}(t, \eta) d\eta.$$

For the average marginal productivity, we can use

$$E \left[ \widehat{\frac{\partial g(T, \epsilon)}{\partial t}} \right] = \frac{1}{n} \sum_{i=1}^n \frac{\partial \hat{\beta}(T_i, \hat{\eta}_i)}{\partial t}.$$

To fill in the details, we need to describe how to estimate  $F_{T|Z}$  and  $E[Y|T, \eta]$ . Note that  $F_{T|Z}$  can be written as

$$F_{T|Z}(t|z) = \Pr(T \leq t|Z = z) = E[1(T \leq t)|Z = z].$$

Thus, both  $F_{T|Z}$  and  $E[Y|T, \eta]$  can be viewed as conditional expectations. This suggests that we can use a nonparametric regression method to estimate these two functions.

Imbens and Newey (2004) develop results for series estimators. (It should also be possible to use kernel estimators and other types of nonparametric estimators.)

First, consider  $F_{T|Z}$ . Let

$$q_{l,L}(z), \quad l = 1, \dots, L; L = 1, 2, \dots$$

denote approximating functions, and let

$$q^L(z) = \begin{pmatrix} q_{1L}(z) \\ \vdots \\ q_{LL}(z) \end{pmatrix}.$$

For example, if  $Z$  is scalar we could use a power series:

$$q^L(z) = \begin{pmatrix} 1 \\ z \\ z^2 \\ \vdots \\ z^{L-1} \end{pmatrix}.$$

Basically, we want to regress  $1(T_i \leq t)$  on  $q^L(Z_i)$ . The least squares coefficients are

$$\left( \sum_{i=1}^n q^L(Z_i)q^L(Z_i)' \right)^{-1} \sum_{i=1}^n q^L(Z_i)1(T_i \leq t).$$

So the fitted values for a particular  $(t, z)$  are:

$$\tilde{\eta} = \tilde{F}(t|z) = q^L(z)' \left( \sum_{i=1}^n q^L(Z_i)q^L(Z_i)' \right)^{-1} \sum_{i=1}^n q^L(Z_i)1(T_i \leq t).$$

Note: we can replace the inverse with a generalized inverse if the matrix is singular. Also, it makes sense to “trim” the  $\tilde{\eta}$  to ensure that they are between 0 and 1:

$$\hat{\eta}_i = 1(\tilde{\eta}_i > 0) \cdot \min\{\tilde{\eta}_i, 1\}.$$

Once we have the  $\hat{\eta}_i$ , we can use them in a regression of  $Y$  on  $T$  and  $\hat{\eta}$  to get an estimate of  $\beta$ .

Again, we will use a series estimators. Let  $w = (t, \eta)$  be the vector of regressors. Let

$$p_{k,K}(w), \quad l = 1, \dots, K; K = 1, 2, \dots$$

denote approximating functions, and let

$$p^K(w) = \begin{pmatrix} p_{1K}(w) \\ \vdots \\ p_{KK}(w) \end{pmatrix}.$$

Then, using  $\hat{w}_i = (T_i, \hat{\eta}_i)$ , we can estimate  $\beta(w)$  by

$$\hat{\beta}(w) := p^K(w)' \hat{\gamma},$$

where

$$\hat{\gamma} := \left( \sum_{i=1}^n p^K(\hat{w}_i)p^K(\hat{w}_i)' \right)^{-1} \sum_{i=1}^n p^K(\hat{w}_i)Y_i.$$