TREATMENT HETEROGENEITY

PAUL SCHRIMPF
NOVEMBER 2, 2011
UNIVERSITY OF BRITISH COLUMBIA
ECONOMICS 628: TOPICS IN ECONOMETRICS

Consider some experimental treatment, such as taking a drug or attending a job training program. It is very likely that different people respond differently to the treatment. For example, with the training program, some people may earn the same whether or not they receive the training, while other people's earnings may be much greater with the training than without. Recognizing this seemingly simple fact greatly affects how we can interpret instrumental variable estimates of the effect of the treatment.

1. CONTEXT

In addition to being empirically relevant, treatment heterogeneity has been important in the development of econometric thought. The thing that distinguishes econometrics from statistics more than anything else is that econometrics focuses far more on estimating causal relationships from observational data. Traditional econometrics focuses on combining economic theory with observational data to infer causal effects. Simultaneous equation methods to estimate e.g. demand and supply, and the Heckman selection model to estimate e.g. the effect of education on earnings are canonical examples of this approach. Roughly in the 1980s, some researchers grew increasingly skeptical of this approach. Their view was that many traditional econometric models made too strong of assumptions. There was a recognition that some of the basic assumption of idealized economic theory may not hold. Moreover, many traditional econometric models invoked functional form and distributional assumptions for tractibility. These assumptions are difficult to defend. Additionally, people became aware that these assumptions can lead to erroneous estimates. An influential paper by LaLonde (1986) compared the estimated effect of a job training program obtained from a randomized experiment to various non-experimental estimates. He found that the non-experimental estimates were sensitive to auxillary assumptions, and often did not agree with the experimental estimate. Results such as this led some economists to reject the traditional approach to econometrics and instead think of causality as only what could be estimated in an idealized experiment. This approach to econometrics is sometimes called the reduced form approach. The traditional approach to econometrics is called the structural approach.

Naturally, there has been some tension between adherents to each of these two approaches. This tension has helped spur progress in both approaches. Since the 1980s, reduced form advocates have greatly clarified exactly what they are estimating. Meanwhile, structural advocates have greatly relaxed functional form and distributional assumptions, and clarified to what extent identification comes from data and to what extent identification comes from other assumptions. Many of the advances on both fronts came from thinking about models with heterogeneous treatment effects.

2. SETUP

I am going to shamelessly use Imbens's slides on IV with treatment heterogeneity in lecture, so I will follow that notation here. We have a cross section of observations indexed by i. There is a treatment $W_i \in \mathcal{W}$. To begin with we will focus on binary treatments, so $W_i \in \{0, 1\}$. Later, we will look at multi-valued treatments. Associated with each treatment is a potential outcome, $Y_i(W_i)$, where $Y_i : \mathcal{W} \to \mathcal{Y}$. Y_i is a function from treatments to potential outcomes. We only observe one of these outcomes, $Y_i(Wi)$, but we are interested in the effect treatment, which is just the difference in potential outcomes,

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

Of course, we cannot estimate $Y_i(1) - Y_i(0)$ for each individual without some unrealistically strong assumptions. However, we can come up with reasonable assumptions to estimate e.g.

$$\mathrm{E}[Y_i(1) - Y_i(0)] = \mathrm{ATE}$$

This quantity is called the average treatment effect, and is often abbreviated ATE. A related quantity of interest is the average effect of treatment for those that receive treatment.

$$\mathrm{E}[Y_i(1) - Y_i(0)|W_i = 1] = \mathrm{ATT}$$

This is called the average effect of treatment on the treated.

When could we estimate the ATE and ATT? Well, the simplest case is if we have a randomized experiment. That is, suppose W_i is randomly assigned, independent of $Y_i(1)$ and $Y_i(0)$. Then

$$\mathrm{E}[Y_i(1)|W_i = 1] = \mathrm{E}[Y_i(1)]$$

and

$$\mathrm{E}[Y_i(0)|W_i = 0] = \mathrm{E}[Y_i(0)].$$

So we can estimate the average treatment effect by[1]

$$\mathbb{E}_n[Y_i(1)|W_i = 1] - \mathbb{E}_n[Y_i(0)|W_i = 0].$$

Also, it is easy to see that the average treatment effect is the same as the average treatment effect for the treated.

If there we do not have a randomized experiment, but we do have an instrument Z_i such that Z affects W but not Y, then with some assumptions, we can estimate the average treatment effect using IV. In particular, suppose

$$Y_i = \beta_0 + \beta_1 W_i + \epsilon_i.$$

Also assume that Z_i is independent of the potential outcomes and potential treatments.

A1 (Independence)**.** *Let $Z_i \in \mathcal{Z}$ and $W_i : \mathcal{Z} \to \{0, 1\}$ then Z_i is independent of $(Y_i(0), Y_i(1), W_i)$, which we denote*

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

It is important to emphasize the fact that W_i is a function of Z_i. Z_i affects the observed treatment through this function, but the distribution of the function is independent of Z_i. In particular, things such as $W_i(z_1) - W_i(z_2)$ for two particular values of the instrument, z_1, z_2 are independent of Z_i. Note that this is a slight change in notation compared to earlier. Earlier, W_i was just the observed treatment, Y_i was a function of W_i, and $Y_i(W_i)$ was the observed outcome. Now, W_i is also a function and $W_i(Z_i)$ is the observed outcome. Henceforth, we will let lower case letters, $y_i = Y_i(W_i(Z_I))$ and $w_i = W_i(Z_i)$ denote the observed outcome and treatment.

[1] I use $\mathbb{E}_n[x_i] = \frac{1}{n}\sum_{i=1}^n x_i$ to denote the empirical expectation of x_i.

Throughout we have also been assuming the following exclusion.

A2 (Exclusion). *Y_i is a function of only $W_i(Z_i)$, and not Z_i directly.*

This assumption is built into our notation, but it is good to state it explicitly, so that we do not forget that we are making it.

The third assumption that we need is the instrument is relevant.

A3 (Instrument relevance). $\mathrm{E}[W_i(z)] \neq 0$ *(as a function of z)*

Then

$$\hat{\beta}_1^{IV} = \frac{\mathbb{E}_n\left[Y_i(Z_i - \mathbb{E}_n[Z_i])\right]}{\mathbb{E}_n\left[W_i(Z_i - \mathbb{E}_n[Z_i])\right]}$$

is a consistent estimate of the average treatment effect and the treatment effect on the treated.

3. Local average treatment effects

From the previous paragraph, we see that IV consistently estimates the average treatment effect when the treatment effect is homogeneous. What happens when the treatment effect is heterogeneous? For ease of exposition let's assume that Z_i is also binary. Then plim of the IV estimate can be written as[2]

$$\begin{aligned}
\operatorname{plim} \hat{\beta}_1^{IV} =& \frac{\mathrm{E}[Y_i(w_i)|Z_i = 1] - \mathrm{E}[Y_i(w_i)|Z_i = 0]}{\mathrm{E}[w_i|Z_i = 1] - \mathrm{E}[w_i|Z_i = 0]} \\
=& \frac{\mathrm{E}[Y_i(1)w_i + Y_i(0)(1 - w_i)|Z_i = 1] - \mathrm{E}[Y_i(1)w_i + Y_i(0)(1 - w_i)|Z_i = 0]}{\mathrm{E}[w_i|Z_i = 1] - \mathrm{E}[w_i|Z_i = 0]} \\
=& \frac{\mathrm{E}[Y_i(1)W_i(1) - Y_i(0)W_i(1)] - \mathrm{E}[Y_i(1)W_i(0) - Y_i(0)W_i(0)]}{\mathrm{E}[w_i|Z_i = 1] - \mathrm{E}[w_i|Z_i = 0]} \\
=& \frac{\mathrm{E}\left[(Y_i(1) - Y_i(0))\,(W_i(1) - W_i(0))\right]}{\mathrm{E}[w_i|Z_i = 1] - \mathrm{E}[w_i|Z_i = 0]} \\
=& \frac{\mathrm{P}(\Delta W_i = 1)\mathrm{E}\left[Y_i(1) - Y_i(0)|\Delta W_i = 1\right] - \mathrm{P}(\Delta W_i = -1)\mathrm{E}\left[Y_i(1) - Y_i(0)|\Delta W_i = -1\right]}{\mathrm{P}(\Delta W_i = 1) - \mathrm{P}(\Delta W_i = -1)} \qquad (1)
\end{aligned}$$

where $\Delta W_i = W_i(1) - W_i(0)$ is the change in treatment when the instrument changes from 0 to 1. The expressions $\mathrm{E}[Y_i(1) - Y_i(0)|\Delta W_i = 1]$ and $\mathrm{E}[Y_i(1) - Y_i(0)|\Delta W_i = -1]$ are average treatment effects conditional on W changing when the instrument changes. This is useful because although these conditional expectation are not the average treatment effect or the average treatment effect on the treated, they are average treatment effects for certain subgroups. However, β_1^{IV} does not estimate these conditional expectations separately. It only estimates the weighted sum in (1). Also notice that even if $\mathrm{E}[Y_i(1) - Y_i(0)|\Delta W_i = 1]$ and $\mathrm{E}[Y_i(1) - Y_i(0)|\Delta W_i = -1]$ are both positive, β_1^{IV} can be positive, negative, or zero. Without a further restriction, the IV estimate might not even have a meaningful sign.

Fortunately, there is a reasonable restriction that can be made. In many cases, we think of instruments that have a monotonic effect on the probability of receiving treatment. For example, in the military service application of Angrist (1990) that we talked about in class, it is sensible to assume that lower draft numbers only increase the probability of serving in the military. In other words, it can be reasonable to assume that $\Delta W_i \geq 0$.

A1 (Monotone instrument). $W_i(1) \geq W_i(0)$

[2]This uses the Wald estimate formula for IV with a binary instrument. We went through this in class, but I'm not going to write it here. See e.g. Angrist and Pischke (2009) for a derivation.

This means that there are no people who receive treatment when the instrument is 0, but do not receive treatment when the instrument is 1. In general, we can divide the population into four groups:

(1) Always takers always receive treatment, $W_i(1) = W_i(0) = 1$
(2) Never takers never receive treatment, $W_i(1) = W_i(0) = 0$
(3) Compliers receive treatment only when the instrument is 1, $W_i(1) = 1$, $W_i(0) = 0$.
(4) Deniers receive treatment only when the instrument is 0, $W_i(1) = 0$, $W_i(0) = 1$.

If we assume that there no deniers, then

$$\begin{aligned} \operatorname{plim} \hat{\beta}_1^{IV} =& \frac{\mathrm{P}(\Delta W_i = 1)\mathrm{E}\left[Y_i(1) - Y_i(0)|\Delta W_i = 1\right]}{\mathrm{P}(\Delta W_i = 1)} && (2) \\ =& \mathrm{E}[Y_i(1) - Y_i(0)|\Delta W_i = 1]. && (3) \end{aligned}$$

This expression is what Imbens and Angrist (1994) call the local average treatment effect, abbreviated LATE. It is the average treatment effect for compliers.

3.1. **Representativeness of compliers.** One natural question is how similar the compliers are to the rest of the population. There is no definitive way to answer this question, but you can get some idea by comparing the compliers to the always takers and the never takers. We can estimate the portion of always takers, compliers, and never takers as follows. Let a denote always takers, n never takers, and c compliers.

$$\begin{aligned} \mathrm{E}[w_i|Z_i = 0] =& \mathrm{E}[w_i|Z_i = 0, a]\mathrm{P}(a|Z_i = 0) + \mathrm{E}[w_i|Z_i = 0, n]\mathrm{P}(n|Z_i = 0) + \mathrm{E}[w_i|Z_i = 0, c]\mathrm{P}(c|Z_i = 0) \\ =& \mathrm{P}(a|Z_i = 0) \end{aligned}$$

The second line follows from the fact that by definition compliers and never takers have $W_i = 0$ when $Z_i = 0$, and always takers have $w_i = 1$. Now an always taker is just someone with $W_i(1) = W_i(0) = 1$. Assumption A1 says that the function W_i is independent of Z_i. Always takers are defined by W_i, therefore being an always taker (or never taker or complier) is independent of Z_i, and $\mathrm{P}(a) = \mathrm{P}(a|Z_i)$. Thus,

$$\mathrm{P}(a) = \mathrm{E}[w_i|Z_i = 0].$$

Identical reasoning[3] shows that

$$\mathrm{P}(n) = 1 - \mathrm{E}[w_i|Z_i = 1],$$

and

$$\mathrm{P}(c) = 1 - \mathrm{P}(n) - \mathrm{P}(a) = \mathrm{E}[w_i|Z_i = 1] - \mathrm{E}[w_i|Z_i = 0].$$

This is useful, but our interpretation of any given local average treatment effect likely depends on $\mathrm{P}(c)$. If we know that the compliers are most of the population ($\mathrm{P}(c)$ is near 1), then we should expect that the LATE is near the ATE (although the difference can still be arbitrarily large if $\mathcal{Y}$ is unbounded).

We can get an even better idea of how the compliers compare the rest of the population by comparing $\mathrm{E}[Y_i(0)|c]$ with $\mathrm{E}[Y_i(0)|n]$, and $\mathrm{E}[Y_i(1)|c]$ with $\mathrm{E}[Y_i(1)|a]$. We have already shown that $\mathrm{E}[Y_i(1) - Y_i(0)|c]$ is identified. Now we will show that $\mathrm{E}[Y_i(0)|c]$, $\mathrm{E}[Y_i(0)|n]$, $\mathrm{E}[Y_i(1)|c]$, and $\mathrm{E}[Y_i(1)|a]$ can be identified.

First note that by the independence assumption (A1),

$$\begin{aligned} \mathrm{E}[Y_i(1)|a] &= \mathrm{E}[Y_i(1)|W_i(1) = W_i(0) = 1] = \\ &= \mathrm{E}[Y_i(1)|W_i(1) = W_i(0) = 1, Z_i = 1] = \mathrm{E}[Y_i(1)|W_i(1) = W_i(0) = 1, Z_i = 0] \end{aligned}$$

[3]It might be a useful exercise to write out the argument.

Also, since anyone with $w_i = 1$ and $Z_i = 0$ is an always taker,

$$\mathrm{E}[Y_i(1)|W_i(1) = W_i(0) = 1, Z_i = 0] = \mathrm{E}[Y_i(w_i)|w_i = 1, Z_i = 0].$$

Thus,

$$\mathrm{E}[Y_i(1)|a] = \mathrm{E}[y_i|w_i = 1, Z_i = 0]$$

is identified. Similarly,[4]

$$\mathrm{E}[Y_i(0)|n] = \mathrm{E}[y_i|w_i = 0, Z_i = 1]$$

is identified. Now observe that

$$\begin{aligned}\mathrm{E}[y_i|w_i = 1, Z_i = 1] =& \mathrm{E}[y_i|w_i = 1, Z_i = 1, c]\mathrm{P}(c|w_i = 1, Z_i = 1) + \mathrm{E}[y_i|w_i = 1, Z_i = 1, a]\mathrm{P}(a|w_i = 1, Z_i = 1)\\ =& \mathrm{E}[Y_i(1)|c]\mathrm{P}(c|w_i = 1, Z_i = 1) + \mathrm{E}[Y_i(1)|a]\mathrm{P}(a|w_i = 1, Z_i = 1)\end{aligned}$$

$w_i = 1$ when $Z_i = 1$ only for compliers and always takers, so

$$\begin{aligned}\mathrm{P}(c|w_i = 1, Z_i = 1) =& \frac{\mathrm{P}(c|Z_i = 1)}{\mathrm{P}(c|Z_i = 1) + \mathrm{P}(a|Z_i = 1)}\\ =& \frac{\mathrm{P}(c)}{\mathrm{P}(c) + \mathrm{P}(a)}\end{aligned}$$

Thus,

$$\mathrm{E}[y_i|w_i = 1, Z_i = 1] = \mathrm{E}[y_i|w_i = 1, c]\frac{\mathrm{P}(c)}{\mathrm{P}(c) + \mathrm{P}(a)} + \mathrm{E}[Y_i(1)|a]\frac{\mathrm{P}(a)}{\mathrm{P}(c) + \mathrm{P}(a)}$$

and

$$\begin{aligned}\mathrm{E}[Y_i(1)|c] =& \mathrm{E}[y_i|w_i = 1, Z_i = 1]\frac{\mathrm{P}(c) + \mathrm{P}(a)}{\mathrm{P}(c)} - \mathrm{E}[Y_i(1)|a]\frac{\mathrm{P}(a)}{\mathrm{P}(c)}\\ =& \mathrm{E}[y_i|w_i = 1, Z_i = 1]\frac{\mathrm{P}(c) + \mathrm{P}(a)}{\mathrm{P}(c)} - \mathrm{E}[y_i|w_i = 1, Z_i = 0]\frac{\mathrm{P}(a)}{\mathrm{P}(c)}.\end{aligned}$$

Similarly,

$$\mathrm{E}[Y_i(0)|c] = \mathrm{E}[y_i|w_i = 0, Z_i = 0]\frac{\mathrm{P}(c) + \mathrm{P}(n)}{\mathrm{P}(c)} - \mathrm{E}[y_i|w_i = 0, Z_i = 1]\frac{\mathrm{P}(n)}{\mathrm{P}(c)}.$$

So you can estimate and compare $\mathrm{E}[Y_i(0)|c]$ with $\mathrm{E}[Y_i(0)|n]$ and $\mathrm{E}[Y_i(1)|c]$ with $\mathrm{E}[Y_i(1)|a]$. For an example of this see Imbens and Wooldridge (2007), which we talked about in lecture.

3.2. **Multi-valued instruments.** In our analysis of LATE above we assumed that the instrument in binary. If the instrument takes on multiple values, say $Z_i \in \mathcal{Z}$ then for any pair $z_0, z_1 \in \mathcal{Z}$, we could repeat the analysis above to show that

$$LATE(z_0, z_1) = \mathrm{E}[Y_i(1) - Y_i(0)|W_i(z_1) = 1, W_i(z_0) = 0]$$

is identified. Also, as above we could define populations of compliers, always takers, and never takers for each z_0, z_1. Of course, do to this we need assumption A1 to hold for each z_0, z_1 i.e. $W_i(z_0) \leq W_i(z_1)$.

What does $\hat{\beta}_1^{IV}$ estimate when Z is multi-valued? Well, in general you can't get a nice interpretable expression for it. However, with some further assumptions you can show that the IV estimate is a weighted average of $LATE(z_0, z_1)$ across different values of z_0 and z_1. Imbens and Angrist (1994) state this result for when Z has discrete support. In the next section, we will give an analogous result for continuously distributed Z.

[4]It might be a useful exercise to write out the argument.

4. CONTINUOUS INSTRUMENTS AND MARGINAL TREATMENT EFFECTS

This section is largely based on Heckman and Vytlacil (1999) and Heckman and Vytlacil (2007). The more structural approach to treatment effects typically treat treatment assignment as a selection problem. That is, they assume that treatment is determined by a latent index,

$$W_i(Z_i) = 1\{\nu(Z_i) - U_i \geq 0\},$$

where $\nu : \mathcal{Z} \to \mathbb{R}$, U_i is some real valued random variable, and $U_i \perp\!\!\!\perp Z_i$. It is easy to see that a latent index model implies the monotonicity assumption (A1) of LATE. For any z_0, z_1, either $\nu(z_0) \leq \nu(z_1)$ or $\nu(z_0) \geq \nu(z_1)$, and then either $W_i(z_0) \leq W_i(z_1)$ or $W_i(z_0) \geq W_i(z_1)$ for all i. On the other hand, it is not clear that the assumptions of LATE imply the existence of such an index model. In fact, early papers on LATE emphasized that the LATE framework does not include a potentially restrictive latent index assumption. You might think that the latent index model is completely unrestrictive since you can always let $\nu(z) = P(W_i(Z_i) = 1|Z_i = z)$ and make U uniform. However, such a U need not be independent of Z. Nonetheless, it turns out that the four LATE assumptions imply the existence of a latent index model with $U_i \perp\!\!\!\perp Z_i$. This result was shown by Vytlacil (2002). This is a useful observation because there are some results that are easier to show directly from the LATE assumptions, and other results that are easier to show from the latent index selection assumption.

Let $\pi(z) = \mathrm{P}(w_i = 1|Z_i = z)$. As in the previous section we can define

$$LATE(p_0, p_1) = \frac{\mathrm{E}[y_i|\pi(Z_i) = p_1] - \mathrm{E}[y_i|\pi(Z_i) = p_0]}{p_1 - p_0}$$

and we should expect that this is the average treatment group for a certain group of compliers. However, this group is a bit complicated because it involves all z_1, z_0 such that $\pi(z_1) = p_1$, and $\pi(z_0) = p_0$. We can get a more tractable expression by using the latent index assumption. Notice that

$$\begin{aligned}\mathrm{E}[y_i|\pi(Z_i) = p] =& p\mathrm{E}[Y_i(1)|\pi(Z_i) = p, D = 1] + (1-p)\mathrm{E}[Y_i(0)|\pi(Z_i) = p, D = 0] \\ =& p\int_0^p \mathrm{E}[Y_i(1)|\tilde{U}_i = u]du + (1-p)\int_0^p \mathrm{E}[Y_i(0)|\tilde{U}_i = u]du\end{aligned}$$

where $\tilde{U}_i = F_U(U_i)$ is uniformly distributed (if we assume U_i is absolutely continuous with respect to Lebesgue measure, something that Vytlacil (2002), shows we can do without loss of generality) and $U_i \leq \nu(Z_i)$ iff $\tilde{U}_i \leq \pi(Z_i)$. Then,

$$\begin{aligned}LATE(p_0, p_1) =& \frac{\int_{p_0}^{p_1} \mathrm{E}[Y_i(1) - Y_i(0)|\tilde{U}_i = p]dp}{p_1 - p_0} \\ =& \mathrm{E}[\Delta Y_i|P(z_0) \leq \tilde{U}_i \leq P(z_1)]. \end{aligned} \tag{4}$$

So $LATE(p_0, p_1)$ is the average treatment effect for people with $\tilde{U}_i$ between p_0 and p_1. We can estimate $LATE(p_0, p_1)$ only when we observe z_0 and z_1 such that $\pi(z_0) = p_0$ and $\pi(z_1) = p_1$. Also, $LATE(0, 1)$ is the average treatment effect. This is the well known result that in selection models, we can identify the average treatment effect only if we have an exclusion with "large support" i.e. $\exists z_0, z_1 \in \mathcal{Z}$ such that $\pi(z_0) = 0$ and $\pi(z_1) = 1$.

The expression in the integrand of (4),

$$MTE(p) = \mathrm{E}[\Delta Y_i|\tilde{U}_i = p]$$

is called the marginal treatment effect. It is the effect of treatment for people with $\tilde{U} = p_0$, i.e. those with $U_i = \nu(z_0)$ where $\pi(z_0) = p_0$. These people are indifferent between receiving treatment or

not. We can write pretty much any other possible treatment effect of interest as an integral of the marginal treatment effect. For example,

$$ATE = \int_0^1 MTE(p)dp.$$

We can identify the marginal treatment as follows. If we take the limit as p_1 approaches p_0 of $LATE(p_1, p_0)$, we get

$$LIV(p_0) = \lim_{p_1 \to \pi 0} \frac{\mathrm{E}[y_i|\pi(Z_i) = p_1] - \mathrm{E}[y_i|\pi(Z_i) = p_0]}{p_1 - p_0}$$

Heckman and Vytlacil (1999) call this the local instrumental variables estimate. It is clear that

$$LIV(p) = \mathrm{E}[\Delta Y_i|\tilde{U}_i = p] = MTE(p),$$

so LIV is an estimate of MTE.

4.1. β^{IV} **as a weighted average of MTE.** We can show that β_1^{IV} estimates a weighted average of marginal treatment effects. Suppose we use some function of Z_i, $g(Z_i)$ as an instrument. Then,

$$\beta_1^{IV}(g) = \frac{\mathrm{E}[y_i(g(z_i) - \mathrm{E}[g(z_i)])]}{\mathrm{E}[y_i(g(z_i) - \mathrm{E}[g(z_i)])]}.$$

Following Heckman and Vytlacil (2007), we will deal with the numerator and denominator separately. Let $\tilde{g}(Z_i) = g(Z_i) - \mathrm{E}[g(Z_i)]$. Note that

$$\begin{aligned}
\mathrm{E}[y_i(g(z_i) - \mathrm{E}[g(z_i)])] =& \mathrm{E}\left[(Y_i(0) + w_i(Y_i(1) - Y_i(0)))\,\tilde{g}_i(Z)\right] \\
=& \mathrm{E}\left[w_i\,(Y_i(1) - Y_i(0))\,\tilde{g}_i(Z)\right] \text{ (independence of } z_i \text{ and } Y_i(1)) \\
=& \mathrm{E}\left[1\{\tilde{U}_i \leq \pi(z_i)\}\,(\Delta Y_i)\,\tilde{g}_i(Z)\right] \\
=& \mathrm{E}\left[1\{\tilde{U}_i \leq \pi(z_i)\}\,(\Delta Y_i)\,\tilde{g}(z_i)\right] \\
=& \mathrm{E}\left[1\{\tilde{U}_i \leq \pi(z_i)\}\mathrm{E}_Y[\Delta Y_i|\tilde{U} = u]\tilde{g}(z_i)\right] \\
=& \mathrm{E}_U\left[\mathrm{E}_Y[\Delta Y_i|\tilde{U} = u]\mathrm{E}_Z[\tilde{g}(z_i)|\pi(z_i) \geq \tilde{U}_i]\mathrm{P}_Z(\tilde{U}_i \leq \pi(z_i))\right] \\
=& \int_0^1 MTE(u)\mathrm{E}_Z[\tilde{g}(z_i)|\pi(z_i) \geq u]\mathrm{P}_Z(u \leq \pi(z_i))du
\end{aligned}$$

where the subscripts on expectations and probabilities are simply to emphasize what the expectation is being taken over. Finally, observe that $Cov(g(z), W) = Cov(g(z), \pi(z))$, so

$$\beta^{IV}(g) = \int_0^1 MTE(u)\omega_g(u)du \tag{5}$$

where

$$\omega_g(u) = \frac{\mathrm{E}_Z[\tilde{g}(z_i)|\pi(z_i) \geq u]\mathrm{P}_Z(u \leq \pi(z_i))}{Cov(g(z), \pi(z))}.$$

It can be shown that these weights integrate to one. Also, if $g(z) = \pi(z)$, it is easy to see that weights are positive. Also, since these weights depend on z and w, they are estimable. We could estimate these weights to get some idea of which weighted average of marginal treatment effects IV is estimating. A final interesting observation is that $\beta^{IV}(g)$ depends on g. In the traditional IV setup, the choice of g affects efficiency, but it does not affect what is being estimated.

5. POLICY RELEVANT TREATMENT EFFECTS

This section is based largely on Carneiro, Heckman, and Vytlacil (2010). In the previous sections we have focused on identifying the effect of administering some treatment. If we think about evaluating some potential policy, the average treatment effect is the effect of the policy that forces everyone to receive treatment. This is often not the most realistic or relevant policy. The majority of policies do not force people to receive undergo some treatment. Instead, policies typically provide some incentive to receive treatment. For example, attending college is a treatment that has been widely studied. However, no one thinks that any government would or should force everyone to attend college. In light of that, although it may be an interesting thing to think about, the average treatment effect of college does not have much practical relevance. The policy interventions with respect to college that we see, such as direct subsidies and subsidized loans, change the incentives to go to college. Our current setup gives us a nice way to think about such policies.

Suppose we observe some baseline policy and want to evaluate an alternative policy. Define the policy relevant treatment effect as

$$PRTE = \frac{\mathrm{E}[y_i|alt] - \mathrm{E}[y_i|base]}{\mathrm{E}[w_i|alt] - \mathrm{E}[w_i|base]}.$$

More generally, we might be interested in the four conditional expectations in this expression separately. We define the PRTE as this ratio so that it has the same form as other treatment effects. It is the effect of the policy per person induced to receive treatment.

If we assume that the policy only affects $\pi(z)$ and not the distribution of Y_i, W_i then we can use our observation of the baseline policy to extrapolate what will happen in the alternate policy. Let $\pi_b(z)$ denote the baseline probability of treatment and $\pi_a(z)$ the alternative probability of treatment. Also, let

$$F_{\pi(z)}(p) = \mathrm{P}(\pi(z) \leq p)$$

be the cdf of $\pi(z)$. Then

$$\begin{aligned}
\mathrm{E}[y_i|base] &= \int_0^1 \mathrm{E}[y_i|base, \pi(z) = p] dF_{\pi_b(z)}(p) \\
&= \int_0^1 \mathrm{E}[w_i Y_i(1) + (1 - w_i) Y_i(0) | \pi(z) = u] dF_{\pi_b(z)}(p) \\
&= \int_0^1 \left(\int_0^1 1\{p \geq u\} \mathrm{E}[Y_i(1)|\tilde{U}_i = u] du \right) dF_{\pi_b(z)}(p) + \\
&\quad + \int_0^1 \left(\int_0^1 1\{p < u\} \mathrm{E}[Y_i(0)|\tilde{U}_i = u] du \right) dF_{\pi_b(z)}(p) \\
&= \int_0^1 F_{\pi(z)}(u) \mathrm{E}[Y_i(1)|\tilde{U}_i = u] + \left(1 - F_{\pi_b(z)}(u)\right) \mathrm{E}[Y_i(0)|\tilde{U}_i = u]
\end{aligned}$$

Similarly,

$$\mathrm{E}[y_i|alt] = \int_0^1 F_{\pi_a(z)}(u) \mathrm{E}[Y_i(1)|\tilde{U}_i = u] + \left(1 - F_{\pi_a(z)}(u)\right) \mathrm{E}[Y_i(0)|\tilde{U}_i = u]$$

Thus,

$$\begin{aligned}
\mathrm{E}[y_i|alt] - \mathrm{E}[y_i|base] &= \int_0^1 \mathrm{E}[Y_i(1)|\tilde{U}_i = u](F_{\pi_a(z)}(u) - F_{\pi_b(z)}(u)) - \mathrm{E}[Y_i(0)|\tilde{U}_i = u](F_{\pi_a(z)}(u) - F_{\pi_b(z)}(u)) du \\
&= \int_0^1 MTE(u)(F_{\pi_a(z)}(u) - F_{\pi_b(z)}(u)) du
\end{aligned}$$

and

$$PRTE = \int_0^1 MTE(u)\omega_{a,b}(u)du \tag{6}$$

where

$$\omega_{a,b}(u) = \frac{F_{\pi_a(z)}(u) - F_{\pi_b(z)}(u)}{\mathrm{E}_z[\pi_a(z)|a] - \mathrm{E}[\pi_b(z)|b]}.$$

As above, we can identify $MTE(u)$ for $u \in \pi_b(\mathcal{Z})$ where $\mathcal{Z}$ is the observed support of z. If we then specify an alternate policy such that $\pi_a(\mathcal{Z}) \subseteq \pi_b(\mathcal{Z})$ (i.e. the alternate policy does not push $\pi_a(z)$ outside of the range observed in the baseline, $\pi_a(z)$ can differ from $\pi_b(z)$ for individual z), then we can identify the policy relevant treatment effect. If we are thinking about an abstract policy, we can just set $\pi_a(z)$ to whatever we think is interesting. If we have a real policy in mind, we might want to estimate $\pi_a(z)$. The appropriate estimation method and assumptions will depend on the particular policy being considered.

If we have a sequence of alternative policies indexed by α such that we can look the limit as $\alpha \to 0$. Also suppose that $\alpha = 0$ is the baseline policy, then we can consider

$$\lim_{\alpha \to 0} PRTE(\alpha) = MPRTE(\alpha)$$

which we call the marginal policy relevant treatment effect. This sequence of alternate policies corresponds to a sequence of $F_{\pi_\alpha(z)}$ which we will just denote as F_α to reduce notation. From the above,

$$PRTE(\alpha) = \int_0^1 MTE(u)\frac{F_\alpha(u) - F_0(u)}{\mathrm{E}[\pi_a(z)] - \mathrm{E}[\pi_b(z)]}$$

Assuming we can interchange limits and integrals, then

$$MPRTE(u) = \int_0^1 MTE(u)\frac{\frac{\partial F_\alpha}{\partial \alpha}(u)|_{\alpha=0}}{\int_0^1 \frac{\partial F_\alpha}{\partial \alpha}(u)|_{\alpha=0}du}du$$

So the MPRTE is again a weighted average of the MTE.

6. Estimation and inference

The key ingredient for estimating policy relevant treatment effects and marginal policy relevant treatment effects is the marginal treatment. As above, $MTE(u)$ is equal to $LIV(u)$, which is

$$LIV(u) = \frac{\partial \mathrm{E}[y_i|\pi(Z_i) = p]}{\partial p}(u)$$

To estimate the marginal treatment effect, we need to estimate the derivative of a conditional expectation function. In the spirit of the section on context, we would like to do so with little or no assumptions about functional form or the distribution of unobservables. There are two primary ways of doing this: kernel regression and series regression. Since $MTE(u)$ is an unknown function that cannot be described by a finite dimensional parameter, estimating $MTE(u)$ is a non-parametric problem. We will see that $MTE(u)$ converges more slowly that the usual parametric ($\sqrt{n}$) rate.

As in the previous section, many other treatment effects of interest can be written as weighted integrals of the marginal treatment effect. There is an econometrics literature about estimating weighted average derivative estimators. The estimate of the marginal treatment effect is a derivative, so various treatment effects can be analyzed as weighted average derivative estimators.

References: Our treatment of series estimators is based on Chernozhukov (2009). Newey (1997) is the standard reference for series estimation. Our presentation of kernel estimation is based largely on Hansen (2009). Hansen (2008) has useful result on uniform convergence rates of kernel estimators.

6.1. **Estimators.** To estimate the marginal treatment effect,

$$MTE(u) = \mathrm{E}[\Delta Y_i|\pi(Z_i) = u] = \frac{\partial \mathrm{E}[y_i|\pi(Z_i) = p]}{\partial p}(u),$$

we must first estimate

$$\pi(Z_i) = \mathrm{E}[w_i|Z_i].$$

$\pi(Z_i)$ is often called the propensity score.[5] To estimate the propensity score, we must estimate the conditional expectation of treatment given the instruments. To estimate the marginal treatment effect, we must estimate the derivative of the conditional expectation of the outcome. The problem of estimating conditional expectations and their derivatives appears quite often, so it has been widely studied. There are two basic approaches: series and kernel based.

6.1.1. *Series regression.*

Suppose we are interested in estimating $\mathrm{E}[y_i|x_i]$ with $x_i \in \mathcal{X} \subseteq \mathbb{R}^d$ and $y_i \in \mathbb{R}$, where $\mathcal{X}$ is compact. We will let $g : \mathcal{X} \to \mathbb{R}$ be $g(x) = \mathrm{E}[y_i|x_i = x]$, and assume that $g \in \mathcal{G}_n$. $\mathcal{G}_n$ is some space of functions from $\mathcal{X}$ to $\mathbb{R}$. We index it by n to allow for the possibility that $\mathcal{G}_n$ changes with the sample size, but in typical applications, it will not. We will denote the probability measure of x by F.

We will approximate g by a linear combination of k functions, $p_j : \mathcal{X} \to \mathbb{R}$, for $j = 1, .., k$. Let

$$p(x) = \begin{pmatrix} p_1(x) \\ \vdots \\ p_k(x) \end{pmatrix}.$$

[5] This name is often associated with propensity score matching, which is another method of estimating treatment effects, but makes very different assumptions.

We approximate $g(x)$ by $p(x)'b$ for some $b \in \mathbb{R}^k$. We will allow k to increase with n so that $p(x)'b$ can get closer and closer to $g(x)$. Series regression gets its name from the sequence of approximating functions p_j. The specification of $\mathcal{G}_n$ affects the rate at which $p(x)'b$ can approach $g(x)$. The most common specification of $\mathcal{G}_n$ is some H older space. We will discuss the $\mathcal{G}_n$ in more detail in the next section.

Common choices of approximating functions include:

(1) Polynomials:
$$p(x) = \begin{pmatrix} 1 & x & cdots & x^{k-1} \end{pmatrix}^T$$
for $d = 1$

(2) Fourier series:
$$p(x) = \begin{pmatrix} 1 & \cos(2\pi x) & \sin(2\pi x) & cdots & \cos(2(k/2-1)\pi x) & \sin(2(k/2-1)\pi x) \end{pmatrix}^T$$
or equivalently,
$$p(x) = \begin{pmatrix} 1 & \cos(\pi x) & \cos(2\pi x) & cdots & \cos(k\pi x) \end{pmatrix}^T$$
for $\mathcal{X} = [0, 1]$.

(3) Splines: a spline of order l with r knots, $t_1, ..., t_r$ is a piecewise polynomial function that is $l-1$ times continuously differentiable. The associated series can be written
$$p(x) = \begin{pmatrix} 1 & x & \cdots & x^l & (x-t_1)_+^3 & \cdots & (x-t_l)_+^3 \end{pmatrix}$$
where $(x)_+ = \max\{0, x\}$. Note that $k = l + 1 + r$.

The above series are often transformed to reduce collinearity of the approximating functions. Polynomials are often orthonormalized, sometimes with respect to the uniform measure on $[0, 1]$, other times with respect to an estimate or guess at the probability measure of x_i. Fourier series are already orthonormal with respect to the uniform measure on $[0, 1]$, but could be orthonormalized with respect to another measure. Splines are often transformed into B-splines instead of the form given above.

Given a choice of series functions and k, we estimate g by regressing y on $p(x)$. We will let
$$\hat{\beta} = \arg\min_{b \in \mathbb{R}^k} \mathbb{E}_n[(y_i - p(x_i)'b)^2]$$
and $\hat{g}(x) = p(x)'\hat{\beta}$. We estimate the derivatives of g by the derivatives of $\hat{g}$,
$$\frac{\partial \hat{g}}{\partial x_j}(x) = \frac{\partial p'}{\partial x_j}(x)\beta.$$
In the next section we will analyze the asymptotics of these estimators.

6.1.2. *Kernel regression.* The usual way of estimating expectations is to just use the sample average. If x has discrete support, then there is no problem with estimating $\mathrm{E}[y_i|x_i = x]$ by $\mathbb{E}_n[y_i|x_i = x]$. However, when x is continuously distributed, the probability of observing any $x_i = x$ is zero for each value of x. Instead of conditioning on $x_i = x$, we could condition on x_i being close to x. For example,
$$\mathbb{E}_n[y_i| \, \|x_i - x\| < h].$$
If h is not too small, we will have some observations with $\|x_i - x\|$. Also, if we let $h \to 0$ then,
$$\mathrm{E}[y_i| \, \|x_i - x\| < h] \to \mathrm{E}[y_i|x_i = x].$$
Therefore we should expect that if we make $h \to 0$ as $n \to \infty$, then $\mathbb{E}_n[y_i| \, \|x_i - x\| < h]$ should converge to $\mathrm{E}[y_i|x_i = x]$.

Rather than just looking at the average y_i for all x_i near x, we could look at a weighted average,

$$\hat{g}_k(x) = \frac{\mathbb{E}_n[y_i K(H^{-1}(x_i - x))]}{\mathbb{E}_n[y_i K(H^{-1}(x_i - x))]}$$

where $K : \mathbb{R}^d \to \mathbb{R}$ is a kernel function such that $\int_{\mathbb{R}^d} K(u)du = 1$ and $K(-u) = K(u)$. H is called the bandwidth. It will depend on n and $\|H\| \to 0$ as $n \to \infty$. Common one-dimensional kernels include:

(1) Uniform: $K(u) = \frac{1}{2}1\{|u| < 1\}$
(2) Gaussian: $K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$
(3) Epanechnikov: $K(u) = \frac{3}{4}(u^2 - 1)1\{|u| < 1\}$

Multivariate kernels are often constructed as products of univariate kernels. There are also more natural multivariate version of each of these kernels.

(1) Uniform: $K(u) = \frac{1}{2}1\{\|u\| < 1\}$
(2) Gaussian: $K(u) = \frac{1}{\sqrt{2\pi}} e^{-u'u/2}$
(3) Epanechnikov: $K(u) = \frac{3}{4}(u'u - 1)1\{\|u\| < 1\}$

The multivariate bandwidth is often $H = hS$ for some fixed nonsingular matrix S and $h \in \mathbb{R}$.

The jth moment of a univariate kernel is

$$\int_{\mathbb{R}} u^j K(u)du.$$

There is a natural definition for multivariate kernels as well, but the notation is more cumbersome, so we do not write it. Symmetric kernels have all of their odd moments equal to zero. A jth order kernel has its $1, 2, ..., j-1$ moments equal to zero and its jth moment non-zero. The three examples above are second order kernels. Higher order kernels can have lower bias, and are necessary for some results. A perceived disadvantage of higher order kernels is that since $\int u^2 K(u)du = 0$, it must be that $K(u) < 0$ for some values of u. Higher order kernels can be constructed from any lower order kernel, see Hansen (2005). The fourth order versions of the Epanechnikov and Gaussian kernels are:

(1) 4th order Epanechnikov: $K(u) = \frac{15}{8}(1 - \frac{7}{3}u^2)\frac{3}{4}(u^2 - 1)1\{|u| < 1\}$
(2) 4th order Gaussian: $K(u) = \frac{1}{2}(3 - u^2)\frac{1}{\sqrt{2\pi}} e^{-u'u/2}$.

As above, to get an estimate of $\frac{\partial \mathrm{E}[y_i|x_i=x]}{\partial x}$ instead of $\mathrm{E}[y_i|x_i = x]$, we simply differentiate our estimate of $g(x) = \mathrm{E}[y_i|x_i = x]$. That is,

$$\widehat{\frac{\partial g}{\partial x_j}}(x) = \frac{\partial \hat{g}_k}{\partial x_j}(x) = \frac{\partial}{\partial x_j} \frac{\mathbb{E}_n[y_i K(H^{-1}(x_i - x))]}{\mathbb{E}_n[y_i K(H^{-1}(x_i - x))]}.$$

We will see that for this estimate to be any good, we will need to use a higher order kernel.

Local polynomial regression. One way of viewing kernel regression is that it solves

$$\hat{g}_k(x) = \min_{b_0} \mathbb{E}_n \left[(y_i - b_0)^2 K\left(H^{-1}(x_i - x)\right) \right]$$

In other words $\hat{g}(x)$ is the weighted least squares regression of y_i on a constant with weights given by $K\left(H^{-1}(x_i - x)\right)$. We could also think of a usual weighted least squares regression,

$$\hat{g}_{ll}(x) \min_{b_0, b} \mathbb{E}_n \left[(y_i - b_0 - x_i' b)^2 K\left(H^{-1}(x_i - x)\right) \right]$$

This estimator is called local linear regression. One advantage of local linear regression is that if we let

$$\hat{b}_0(x), \hat{b}(x) = \arg\min_{b_0, b} \mathbb{E}_n \left[(y_i - b_0 - x_i' b)^2 K \left(H^{-1}(x_i - x) \right) \right]$$

Then the jth component of $\hat{b}(x)$ will be a valid estimation of $\frac{\partial g}{\partial x_j}$. The natural generalization of local linear regression local polynomial regression. Let $p(x_i)$ denote a polynomial series of fixed degree evaluated at x_i. Local polynomial regression is

$$\hat{g}_{lp}(x) \min_b \mathbb{E}_n \left[(y_i - p(x_i)' b)^2 K \left(H^{-1}(x_i - x) \right) \right].$$

Note that unlike with series regression the degree of this polynomial is fixed. It does not change with the sample size. As with linear regression the coefficients of a local polynomial regression of degree m are valid estimates of the first m derivatives of g.

6.2. **Asymptotic theory.** In this section, we will establish conditions under which series and kernel regression are consistent. We will also find the rate of convergence and the limit distribution of these estimators. We will see that the results depend on the rate at which k grows with n for series or the rate at which h shrinks with n for kernels. Since g is a function, we can look at these results both pointwise for some particular value of x, and for the entire function g. That is, we can examine the pointwise convergence of

$$\hat{g}(x) - g(x)$$

and the convergence of the entire function by analyzing

$$\|\hat{g} - g\|.$$

Note that there are a variety of potential norms we could consider. We will focus on $\mathcal{L}^2$ and $\mathcal{L}^\infty$ norms.

6.2.1. *Series regression.* Let

$$\hat{\beta} = \arg\min_b \mathbb{E}_n \left[(y_i - p(x_i)' b)^2 \right]$$

so that

$$\hat{g}(x) = p(x)' \hat{\beta}.$$

Also define the population version of $\hat{\beta}$ as

$$\beta = \arg\min_b \mathrm{E} \left[(y_i - p(x_i)' b)^2 \right].$$

Then we can write

$$\begin{aligned} \hat{g}(x) - g(x) =& p(x)' \hat{\beta} - p(x)' \beta + p(x)' \beta - g(x) \\ =& \underbrace{p(x)' \left(\hat{\beta} - \beta \right)}_{\text{estimation error}} + \underbrace{\left(p(x)' \beta - g(x) \right)}_{\text{approximation error}} \end{aligned}$$

We begin by assuming the following.

A1. *Let $\epsilon_i = y_i - g(x_i)$. (ϵ_i, x_i) are i.i.d. and $\sigma_i^2 \equiv \mathrm{E}[\epsilon_i^2 | x_i]$ is bounded.*

Approximation by series is a topic that has been widely studied, so there are many results available about the approximation error,

$$r(x) \equiv p(x)' \beta - g(x).$$

We will make the following assumption about $r(x)$.

A2 (approximation error). *For each $g \in \mathcal{G}_n$ there are finite constants c_k and ℓ_k such that*

$$\|r\|_{F,2} \equiv \left(\int r(x)^2 dF(x) \right)^{1/2} \leq c_k$$

and

$$\|r\|_\infty \equiv \sup_{x \in \mathcal{X}} |r(x)| \leq \ell_k c_k.$$

Recall that F is the probability measure of x. Feasible values of c_k and ℓ_k depend on $\mathcal{G}_n$, the series being used, and F. Suppose that $\mathcal{G}_n$ is a H older (α, s) class, i.e.

$$\mathcal{G}_n = \{g \in C^s(\mathcal{X}) : \|D^s g(x_1) - D^s g(x_2)\| \leq M \|x_1 - x_2\|^\alpha \, \forall x_1, x_2 \in \mathcal{X}\}$$

where $C^s(\mathcal{X})$ is the set of s times continuously differentiable functions from $\mathcal{X}$ to $\mathbb{R}$, Then for polynomials and Fourier series,

$$c_k \lesssim k^{-s/d}$$

where the notation $f(k) \lesssim g(k)$ means that there exists $0 \leq M < \infty$ such that $f(k) \leq M g(k)$ for all k. For splines of order s_0,

$$c_k \lesssim k^{-\max\{s, s_0\}/d}.$$

See the references in Chernozhukov (2009) or Newey (1997) for the source of these results. The bound on ℓ_k is even more dependent on the series and other assumptions. The following results are useful.

(1) **Polynomials**: for Chebyshev polynomials on $\mathcal{X} = [-1, 1]$ with $dF(x)/dx = \frac{1}{\sqrt{1-x^2}}$ (so the Chebyshev polynomials are orthonormal with the inner product induced by dF), and $\mathcal{G}_n \subseteq C(\mathcal{X})$, then

$$\ell_k \leq c_0 \log k + c_1$$

for some fixed c_0 and c_1.

(2) **Fourier series**: if $\mathcal{X} = [0, 1]$, F is uniform (so the Fourier series is orthonormal with the inner product induced by dF), and $\mathcal{G}_n \subseteq C(\mathcal{X})$, then

$$\ell_k \leq c_0 \log k + c_1$$

for some fixed c_0 and c_1.

(3) **Splines**: if $\mathcal{X} = [0, 1]$, F is uniform, and $\mathcal{G}_n \subseteq C(\mathcal{X})$, then

$$\ell_k \leq c_0$$

for some fixed c_0.

These results are stated in Chernozhukov (2009). Presumably, he got them from one of the references he listed. I'd guess DeVore and Lorentz (1993), but I really do not know.

We will go into more detail about how to control the sampling error,

$$p(x_i)(\hat{\beta} - \beta)$$

The usual formula for $\hat{\beta}$ gives that

$$\hat{\beta} = \mathbb{E}_n[p(x_i)p(x_i)']^{-1}\mathbb{E}_n[p(x_i)y]$$

and

$$\beta == \mathrm{E}[p(x_i)p(x_i)']^{-1}\mathrm{E}[p(x_i)y].$$

The matrix $\mathrm{E}[p(x_i)p(x_i)']$ will appear often. We will also need to control how large $\|p(x)\|$ can be. To do so, we assume the following.

A3. *The eigenvalues of* $Q \equiv= \mathrm{E}[p(x_i)p(x_i)']$ *are bounded above and away from zero. Let*

$$\xi_k \equiv \sup_{x\in\mathcal{X}} \|p(x)\|.$$

Assume that k is such that

$$\xi_k^2 n^{-1} \log n \to 0.$$

As stated in Newey (1997), for polynomials, $\xi_k \lesssim k$. For Fourier series and splines $\xi_k \lesssim \sqrt{k}$.

We can now state some results. We begin by establishing the $\mathcal{L}^2$ rate of convergence.

Theorem 6.1 ($\mathcal{L}^2$ rate for series)**.** *Under A1, A2, and A3, and if* $c_k \to 0$*, then*

$$\|\hat{g} - h\|_{F,2} \lesssim_p \sqrt{\frac{k}{n}} + c_k$$

Proof. See Chernozhukov (2009). □

It is common to choose the number of series terms to achieve the fastest $\mathcal{L}^2$ convergence rate. The first part of the rat,e $\sqrt{k/n}$ represents sampling error and increases with k. The second term, c_k, is approximation error. It decreases with k. The fastest convergence rate is achieved when the rate from sampling error and rate approximation error are made equal.

$$\sqrt{k/n} \propto c_k$$

As stated above, for polynomials and Fourier series, $c_k \lesssim k^{-s/d}$ where g is assumed to be s times differentiable. In that case, the optimal k is proportional to $n^{\frac{d}{s+d}}$ and the optimal rate is $\sqrt{n^{\frac{-s}{s+d}}}$. Thus, the $\mathcal{L}^2$ nonparametric rate is slower than $n^{-1/2}$, but the nonparametric $\mathcal{L}^2$ rate gets close to $n^{-1/2}$ if we assume the function we are estimating is very smooth.

Lemma 6.1 (Pointwise linearization)**.** *Under A1-A3, for* $\alpha \in \mathbb{R}^k$ *with* $\|\alpha\| = 1$ *we have*

$$\sqrt{n}\alpha'(\hat{\beta} - \beta) = \alpha' \mathbb{G}_n \left[p(x_i)(\epsilon_i + r(x_i))\right] + R_{1n}$$

where

$$R_{1n} \lesssim_p \sqrt{\frac{\xi_k^2 \log n}{n}} \left(1 + \ell_k c_k \sqrt{k \log n}\right)$$

and

$$\alpha' \mathbb{G}_n [p(x_i)' r(x_i)] \lesssim_p \ell_k c_k.$$

Proof. See Chernozhukov (2009). □

Notice that if $(\xi_k^2 \log n)/n \to 0$ and $\ell_k c_k \to 0$, then all that remains is $\alpha' \mathbb{G}_n [p_i(x_i)\epsilon_i]$. Given an appropriate assumption on ϵ_i, we can show that this term is asymptotically normal.

A4. ϵ_i *is such that for each* $M \to \infty$*,*

$$\sup_{x\in\mathcal{X}} \mathrm{E}\left[\epsilon_i^2 1\{|\epsilon_i| > M\} | x_i = x\right] \to 0,$$

and approximation error obeys $|r(x_i)| \leq \ell_k c_k = o(\sqrt{n/\xi_k})$*.*

Theorem 6.2 (Pointwise asymptotic normality)**.** *Suppose A1-A4 hold. If* $R_{1n} \xrightarrow{p} 0$ *and* $\ell_k c_k \to 0$*, then*

$$\sqrt{n} \frac{\alpha'(\hat{\beta} - \beta)}{\|\alpha' \Omega^{1/2}\|} \xrightarrow{d} N(0, 1),$$

where $\Omega = Q^{-1} \mathrm{E}[\epsilon_i^2 p(x_i) p(x_i)'] Q^{-1}$*.*

Proof. See Chernozhukov (2009). □

Note that we can take $\alpha = p(x)$ to obtain

$$\sqrt{n}\frac{p(x)'(\hat{\beta}-\beta)}{\|p(x)'\Omega^{1/2}\|} \xrightarrow{d} N(0,1).$$

If additionally $\frac{\sqrt{n}r(x)}{\|p(x)'\Omega^{1/2}\|} \to 0$, then we have

$$\sqrt{n}\frac{p(x)'\hat{\beta}-g(x))}{\|p(x)'\Omega^{1/2}\|} \xrightarrow{d} N(0,1).$$

This is why the theorem is label pointwise asymptotic normality.

Another thing to notice about theorem 6.3 is that it is always true that a $N(0,1)$ has the same distribution as $\frac{\alpha'\Omega^{1/2}}{\|\alpha'\Omega^{1/2}\|}N(0,I_k)$. If k were fixed we would have

$$\sqrt{n}(\hat{\beta}-\beta)\Omega^{-1/2} \xrightarrow{d} N(0,I_k).$$

We cannot get this sort of result here because k is increasing with n. However, to emphasize this parallel, we could have stated the result of theorem 6.3 as

$$\sqrt{n}\frac{p(x)'\hat{\beta}-g(x))}{\|p(x)'\Omega^{1/2}\|} \xrightarrow{d} \frac{p(x)'\Omega^{1/2}}{\|p(x)'\Omega^{1/2}\|}N(0,I_k).$$

We could also state this result as

$$\left|\sqrt{n}\frac{p(x)'\hat{\beta}-g(x))}{\|p(x)'\Omega^{1/2}\|} - \frac{p(x)'\Omega^{1/2}}{\|p(x)'\Omega^{1/2}\|}\mathcal{N}_k\right| = o_p(1),$$

for some $\mathcal{N}_k \sim N(0,I_k)$. When we look at the uniform limit distribution, we will get a result with this form, so it is useful to draw attention to the similarity. We did not originally state the theorem in this form to emphasize that theorem 6.3 is really a result of applying a standard central limit theorem.

To obtain a uniform linearization and asymptotic distribution, we need a stronger assumption on the errors.

A5. *The errors are conditionally sub-Gaussian, which means*

$$\sup_{x\in\mathcal{X}} \mathrm{E}\left[e^{\epsilon_i^2/2}|x_i = x\right] < \infty.$$

Additionally for $\alpha(x) \equiv p(x)/\|p(x)\|$*, we have*

$$\|\alpha(x_1)-\alpha(x_2)\| \leq \xi_{1k}\|x_1-x_2\|$$

with $\xi_{1k} \lesssim k^a$ *for some* $a < \infty$*.*

Lemma 6.2 (uniform linearization)**.** *Suppose that A1-A5 hold. Then uniformly in* $x \in \mathcal{X}$*,*

$$\sqrt{n}\alpha(x)'(\hat{\beta}-\beta) = \alpha(x)'\mathbb{G}_n\left[p(x_i)(\epsilon_i + r(x_i))\right] + R_{1n}$$

where

$$R_{1n} \lesssim_p \sqrt{\frac{\xi_k^2(\log n)^2}{n}}\left(1+\ell_k c_k\sqrt{k\log n}\right)$$

and

$$\alpha(x)'\mathbb{G}_n[p(x_i)r(x_i)] = R_{2n} \lesssim_p \ell_k c_k \log n$$

Proof. See Chernozhukov (2009). □

Theorem 6.3 (uniform rate). *Under A1-A5 we have*

$$\sup_{x\in\mathcal{X}} \left|\alpha(x)'\mathbb{G}_n\left[p(x_i)(\epsilon_i + r(x_i))\right]\right| \lesssim_p (\log n)^{3/2}$$

so

$$\sup_{x\in\mathcal{X}} |\hat{g}(x) - g(x)| \lesssim_p \frac{\xi_k}{\sqrt{n}}\left((\log n)^{3/2} + R_{1n} + R_{2n}\right) + \ell_k c_k$$

Proof. See Chernozhukov (2009). □

Finally, we state a uniform convergence in distribution result.

Theorem 6.4 (strong approximation). *Suppose A1-A5 hold and $R_{1n} = o_p(a_n^{-1})$, and that*

$$a_n^3 k^4 \xi_k^2 (1 + \ell_k^3 c_k^3)(\log n)^2/n \to 0$$

Then for some $\mathcal{N}_k \sim N(0, I_k)$ we have

$$\sup_{\alpha\in S^{k-1}} \left|\sqrt{n}\frac{\alpha'(\hat{\beta}-\beta)}{\|\alpha'\Omega^{1/2}\|} - \frac{\alpha'\Omega^{1/2}}{\|\alpha'\Omega^{1/2}\|}\mathcal{N}_k\right| = o_p(a_n^{-1})$$

As with the pointwise limit theorem 6.3, if we replace α with $p(x)$ and we have $\sup_{x\in\mathcal{X}} \sqrt{n}\frac{|r(x)|}{\|p(x)'\Omega^{1/2}\|} = o_P(a_n^-1)$, then 6.4 implies that

$$\sup_{x\in\mathcal{X}} \left|\sqrt{n}\frac{\hat{g}(x) - g(x)}{\|\alpha'\Omega^{1/2}\|} - \frac{p(x)'\Omega^{1/2}}{\|p(x)'\Omega^{1/2}\|}\mathcal{N}_k\right| = o_p(a_n^{-1}).$$

Note that unlike the pointwise asymptotic distribution (6.3), the uniform limiting theory is not a traditional weak convergence result. For a given x, regardless of k, $\frac{p(x)'\Omega^{1/2}}{\|p(x)'\Omega^{1/2}\|}\mathcal{N}_k$ has a standard normal distribution. However, as a function of x, the Gaussian process, $\frac{p(x)'\Omega^{1/2}}{\|p(x)'\Omega^{1/2}\|}\mathcal{N}_k$ changes with k. Theorem 6.4 says nothing about whether $\frac{p(x)'\Omega^{1/2}}{\|p(x)'\Omega^{1/2}\|}\mathcal{N}_k$ ever converges to a fixed Gaussian process, so in particular, the theorem does not show weak convergence. Nonetheless, for any k, $\frac{p(x)'\Omega^{1/2}}{\|p(x)'\Omega^{1/2}\|}\mathcal{N}_k$ is a tractable process and we can find its distribution either analytically or through simulation. This is enough to perform inference.

To get some idea of how this approximating process behaves, figure 1 shows the covariance function of $\frac{p(x)'\Omega^{1/2}}{\|p(x)'\Omega^{1/2}\|}\mathcal{N}_k$ for $d = 1$ and polynomials for various k. That is, it plots

$$\text{Cov}\left(\frac{p(x_1)'\Omega^{1/2}}{\|p(x_1)'\Omega^{1/2}\|}\mathcal{N}_k, fracp(x_2)'\Omega^{1/2}\left\|p(x_2)'\Omega^{1/2}\right\|\mathcal{N}_k\right)$$

as a function of x_1 and x_2. When $x_1 = x_2$ the variance is always one. When $x_1 \neq x_2$, the covariance approaches 0 as k increases. The approximating processes eventually converge to white noise. However, we cannot perform inference based on white noise as the limiting distribution because we have not shown how quickly the approximating processes approach white noise.

Figure 2 shows the same thing for Fourier series.

A uniform confidence band of $g(x)$ of level $1-\alpha$ is a pair of functions, $l_k(x), u_k(x)$ such that

$$\text{P}(l_k(x) \leq g(x) \leq u_k(x) \forall x \in \mathcal{X}) = 1 - \alpha$$

There are a number of ways to construct such bands, but it is standard to focus on bands of the form

$$(l_k(x), u_k(x)) = \hat{g}(x) \pm \kappa(1-\alpha)\left\|p(x)'\Omega^{1/2}\right\|$$

FIGURE 1. Covariance function for polynomials

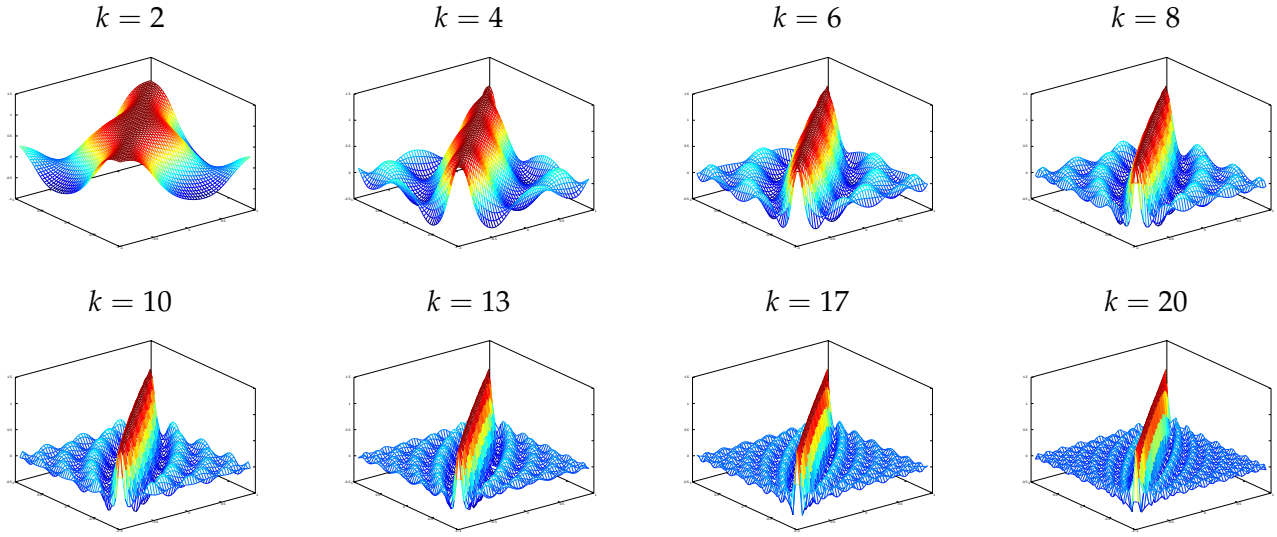
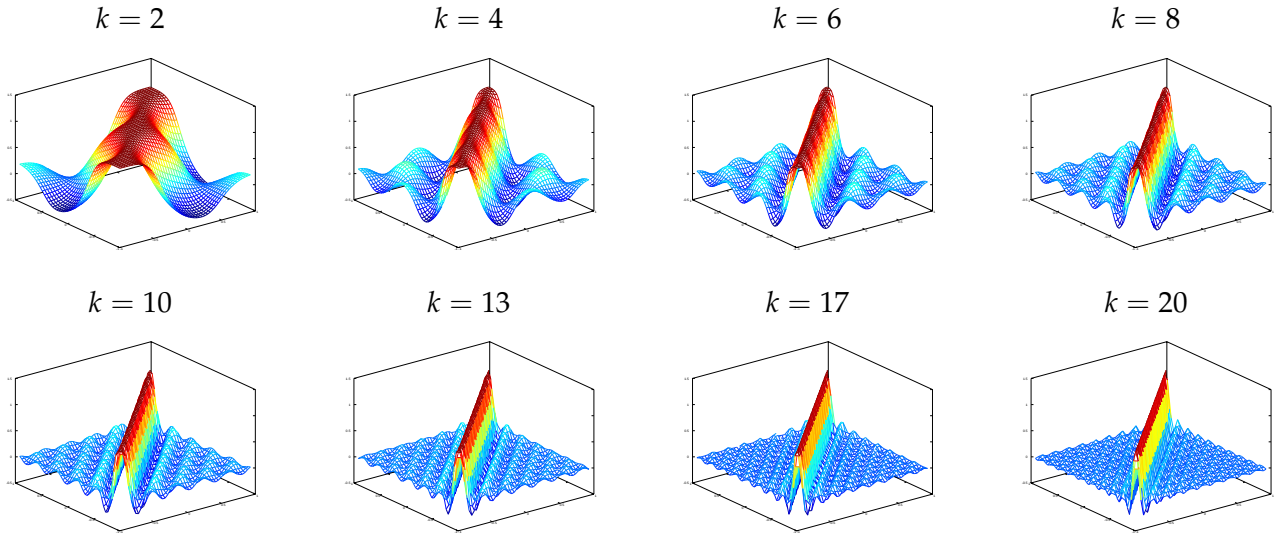


FIGURE 2. Covariance function for Fourier series



where $\kappa(1-\alpha)$ is chosen so as to get the correct coverage probability. From theorem 6.4, we know that $\sqrt{n}\frac{\hat{g}(x)-g(x)}{\left\|p(x)'\Omega^{1/2}\right\|}$ is uniformly close to $\frac{p(x)'\Omega^{1/2}}{\left\|p(x)'\Omega^{1/2}\right\|}\mathcal{N}_k$. Let

$$Z_n(x) \equiv \frac{p(x)'\Omega^{1/2}}{\left\|p(x)'\Omega^{1/2}\right\|}\mathcal{N}_k.$$

We can set $\kappa(1-\alpha)$ to be the $(1-\alpha)$ quantile of

$$\sup_{x\in\mathcal{X}} |Z_n(x)|.$$

There are analytic results for this, which are useful for comparing the widths of these confidence bands to confidence bands from other methods or of other estimators. However, in practice, it is easier to use simulation. Thus, you could compute confidence bands by:

(1) Estimate $\hat{g}(x)$.
(2) Estimate $\hat{\Omega} = \mathbb{E}_n[p(x_i)p(x_i)']^{-1}\mathbb{E}_n[\hat{\epsilon}_i^2 p(x_i)p(x_i)']\mathbb{E}_n[p(x_i)p(x_i)']^{-1}$
(3) Simulate a large number of draws, say $z_1, ..., z_R$ from $N(0, I_k)$, set

$$Z_{n,r}(x) = \frac{p(x)'\hat{\Omega}^{1/2}}{\left\|p(x)'\hat{\Omega}^{1/2}\right\|} z_r$$

and find $\sup_{x\in\mathcal{X}} |Z_{n,r}(x)|$
(4) Set $\hat{\kappa}(1-\alpha) = 1 - \alpha/2$ quantile of $\sup_{x\in\mathcal{X}} |Z_{n,r}(x)|$
(5) The confidence bands are

$$(\hat{l}_k(x), \hat{u}_k(x)) = \hat{g}(x) \pm \hat{\kappa}(1-\alpha) \left\|p(x)'\hat{\Omega}^{1/2}\right\|$$

Note that all of our results above treated Ω as known. We could show that the results go through when using $\hat{\Omega}$ instead, see Chernozhukov, Lee, and Rosen (2009).

Throughout, we have had these constants c_k, ℓ_k, etc that depend on various details of the problem. Wang and Yang (2009) obtain similar results for spline regression, but they make explicit assumptions about what c_k, ℓ_k, etc will be.

All of our results have been for $\hat{g}(x)$ and not $\frac{\partial \hat{g}}{\partial x_j}(x)$. However, the result in theorem 6.4 also applies to

$$\sup_{x\in\mathcal{X}} \left| \sqrt{n} \frac{p^j(x)'(\hat{\beta} - \beta)}{\left\|p^j(x)'\Omega^{1/2}\right\|} - \frac{p^j(x)'\Omega^{1/2}}{\|\alpha'\Omega^{1/2}\|} \mathcal{N}_k \right| = o_p(a_n^{-1})$$

where $p^j(x) = \frac{\partial p}{\partial x_j}(x)$. If we redefine the approximation error as

$$r(x) = p^j(x)'\beta - \frac{\partial g}{\partial x_j}$$

then we just need to control this approximation error instead. If I recall correctly, we will generally get $c_k = k^{-\frac{s-m}{d}}$ when we approximate the mth derivative. I believe ℓ_k will not change, but I am not at all certain. Finally ξ_k must be redefined as $\sup \left\|p^j(x)\right\|$, and it increases to $k^{1/2+m}$ for splines, and k^{1+2m} for polynomials. I am not sure about Fourier series, but I suspect $k^{1/2+m}$ as well. It is easy to show that $k^{1/2+m}$ works, but it may be possible to get a sharper bound. In any case, both ξ_k and c_k are worse when estimating derivatives instead of functions themselves. Because of this, we will get a slower rate of convergence when estimating derivatives.

6.2.2. *Kernel regression.* I am running out of time, so just refer to Hansen (2009) for kernel regression. Hansen's notes on nonoparametrics have 16 parts. The most relevant is the second part, `http://www.ssc.wisc.edu/~bhansen/718/NonParametrics2.pdf`. Hansen's notes show the same sort of pointwise asymptotic normality and uniform convergence rate results as above for series esitmators. Hansen's notes do not cover a uniform limiting distribution. However, something like theorem 6.4 can be shown for kernel regression as well. See e.g. Chernozhukov, Lee, and Rosen (2009), although the result was first shown much earlier.

6.2.3. *Bootstrap.* Someone asked whether you can construct uniform confidence bands using the bootstrap. Yes, you can, but only if you bootstrap in the correct way. It has not been proven that the standard nonparametric bootstrap works (i.e. resampling observations with replacement). However, certain variants of the bootstrap do work. For kernel regression, Hardle and Marron (1991) propose using a wild bootstrap procedure. Claeskens and van Keilegom (2003) propose a smoothed bootstrap procedure is consistent for local polynomial regression. I do not know of any analogous result for series regression. However, I am fairly certain that a combination of the

arguments in Chernozhukov (2009) and Chandrasekhar, Chernozhukov, Molinari, and Schrimpf (2011) would show consistency of another smoothed bootstrap procedure.

REFERENCES

ANGRIST, J. (1990): "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," *American Economic Review*, 80(3), 313–36.

ANGRIST, J., AND J.-S. PISCHKE (2009): *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

CARNEIRO, P., J. J. HECKMAN, AND E. VYTLACIL (2010): "Evaluating Marginal Policy Changes and the Average Effect of Treatment for Individuals at the Margin," *Econometrica*, 78(1), 377–394.

CHANDRASEKHAR, A., V. CHERNOZHUKOV, F. MOLINARI, AND P. SCHRIMPF (2011): "Inference for best linear approximations to set identified functions," .

CHERNOZHUKOV, V. (2009): "(Some) new asymptotic theory for series esitmators," Lecture notes.

CHERNOZHUKOV, V., S. LEE, AND A. ROSEN (2009): "Intersection Bounds: Estimation and Inference," CeMMAP Working Paper CWP19/09.

CLAESKENS, G., AND I. VAN KEILEGOM (2003): "Bootstrap confidence bands for regression curves and their derivatives," *Annals of Statistics*, 31(6), 1852–1884.

DEVORE, R. A., AND G. G. LORENTZ (1993): *Constructive Approximation*. Springer.

HANSEN, B. (2009): "Lecture notes on nonparametrics," Lecture notes.

HANSEN, B. E. (2005): "Exact Mean Integrated Squared Error of Higher Order Kernel Estimators," *Econometric Theory*, 21(6), pp. 1031–1057.

——— (2008): "Uniform Convergence Rates for Kernel Estimation with Dependent Data," *Econometric Theory*, 24(03), 726–748.

HARDLE, W., AND J. S. MARRON (1991): "Bootstrap Simultaneous Error Bars for Nonparametric Regression," *Annals of Statistics*, 31(6), 1852–1884.

HECKMAN, J., AND E. VYTLACIL (1999): "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects," *Proceedings of the National Academy of Sciences*, 96.

HECKMAN, J. J., AND E. J. VYTLACIL (2007): "Chapter 71 Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast their Effects in New Environments," in *Handbook of Econometrics*, ed. by J. J. Heckman, and E. E. Leamer, vol. 6, Part B of *Handbook of Econometrics*, pp. 4875 – 5143. Elsevier.

IMBENS, G., AND J. WOOLDRIDGE (2007): "Discrete Choice Models," Discussion paper, NBER Lecture Notes 11.

IMBENS, G. W., AND J. D. ANGRIST (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62(2), pp. 467–475.

LALONDE, R. J. (1986): "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review*, 76(4), 604–620.

NEWEY, W. K. (1997): "Convergence rates and asymptotic normality for series estimators," *Journal of Econometrics*, 79(1), 147–168.

VYTLACIL, E. (2002): "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*, 70(1), 331–341.

WANG, J., AND L. YANG (2009): "Polynomial spline confidence bands for regression curves," *Statistica Sinica*, 19, 325–342.