1 A Framework For Counterfactuals And Causal Inference

(A-1) Two potential states, “untreated” and “treated,” with corresponding potential outcomes given by the variables \((y_0, y_1)\), \(y_0, y_1 \in \mathbb{R}\). Extension to more than two potential outcomes is straightforward.

(A-2) No market or social interactions among agents in the hypothetical world.

(A-3) A static model.
A causal function for each state,

\[ y_0 = g_0(x) \]
\[ y_1 = g_1(x). \]

Letting \( D_i \) be the domain of function \( g_i \),

\[ g_0: D_0 \rightarrow \mathbb{R}, \ g_1: D_1 \rightarrow \mathbb{R} \text{ and } D_i \subseteq \mathbb{R}^J. \]

Let \((\Omega, \mathcal{A}, P)\) denote a probability space. The random variables for the potential outcomes are \((Y_0(\omega), Y_1(\omega))\). \(X(\omega)\) produces the random variable \(Y(\omega)\),

\[ Y_0(\omega) = g_0(X(\omega)) \]

\[ Y_1(\omega) = g_1(X(\omega)) \]
$g_0, g_1$ functions are playing two roles in our analysis.

They are describing not only how the random vector $X(\omega)$ is functionally related to the random variables $Y_0(\omega), Y_1(\omega)$, but also are specifying what values the outcomes would have taken had the causes $X(\omega)$ taken alternative values.

A causal relationship is only well defined if a theory relating causes to outcomes is articulated and a mechanism generating
variation in the causes is clearly specified.
The causal effect of $x_j$ on $y_i (i = 0, 1)$ is obtained from varying $x_j$. Thus consider two values of $x_j : x'_j$ and $x''_j$. For this variation, we define the causal effect of $x_j$ on $y_i$, fixing the remaining coordinates of $x$, as

\begin{align*}
(2-2) \ [\Delta y_i \mid \Delta x_j = x'_j - x''_j] &= g_i(x_1, \ldots, x'_j, \ldots, x_J) \\
&\quad - g_i(x_1, \ldots, x''_j, \ldots, x_J).
\end{align*}

$g_0, g_1$, it is thus required that these variables be variation-free: $(x_1, \ldots, x_J) \in X_1 \times \ldots \times X_J$.

*Person-specific* causal effect of $x_j$
\[ g_i(X_1(\omega), \ldots, X_{j-1}(\omega), x'_j, X_{j+1}(\omega), \ldots, X_J(\omega)) \]

\[ -g_i(X_1(\omega), \ldots, X_{j-1}(\omega), x_j, X_{j+1}(\omega), \ldots, X_J(\omega)) \]
If the $g_i$ are differentiable with respect to $x$, then the *Limit Causal Effect* of $x_j$ on $y_i$ is

\[(2-3) \text{ Limit Causal Effect of } x_j \text{ on } y_i = \frac{\partial y_i}{\partial x_j} = \frac{\partial g_i(x)}{\partial x_j}.\]

Treatment effect causal effects are versions of Marshallian causal effects. Define $D(\omega)$ as an indicator denoting whether in a hypothetical population person $\omega$ is in regime “1” or not.

Thus $D(\omega) = 1$ if $Y_1(\omega)$ is observed and $D(\omega) = 0$ if $Y_0(\omega)$ is observed.

$D(\omega)$ is assumed to be a random variable
determined by the random vector $Z(\omega)$. Let $\mathcal{Z}$ denote the support of $Z(\omega)$. The assumption that $D(\omega)$ is measurable $\sigma(Z(\omega))$ is equivalent to the following representation:

\begin{equation}
D(\omega) = 1 \text{ if } Z(\omega) \in \tilde{\mathcal{Z}}, \quad D(\omega) = 0 \text{ if } Z(\omega) \in \mathcal{Z} \setminus \tilde{\mathcal{Z}}
\end{equation}

where $\tilde{\mathcal{Z}}$ is a measurable subset of the support of $Z(\omega)$.

\begin{equation}
d = 1 \text{ if } z \in \tilde{\mathcal{Z}}, d = 0 \text{ if } z \in \mathcal{D}_d \setminus \tilde{\mathcal{Z}}
\end{equation}

where $\mathcal{D}_d$ is the domain on which the function is defined and $\tilde{\mathcal{Z}} = \tilde{\mathcal{Z}} \cap \mathcal{Z}$. Measured outcome, random variable $Y(\omega)$ in a hypothetical...
population, as:

\[(2-5)Y(\omega) = D(\omega)Y_1(\omega) + (1 - D(\omega))Y_0(\omega).\]
(2-6) *Treatment Effect Causal Effect:*

\[ \Delta(x) = y_1(x) - y_0(x) = g_1(x) - g_0(x). \]

This expression evaluated at the characteristics of individual \( \omega \) is

\[ \Delta(X(\omega)) = Y_1(\omega) - Y_0(\omega) = g_1(X(\omega)) - g_0(X(\omega)). \]
Define the causal effect of changing $D$:

$$ h(d(z), y_1(x), y_0(x)) = d(z)y_1(x) + (1 - d(z))y_0(x), $$

and

$$ \tilde{h}(z, x) = d(z)y_1(x) + (1 - d(z))y_0(x). $$
Then provided that there is some mechanism for changing $D$ while holding $X$ fixed, it is meaningful in the hypothetical world to define the treatment effect causal effect as

$$
\Delta(x) = h(1, y_1(x), y_0(x)) - h(0, y_1(x), y_0(x))
$$

or in terms of $\tilde{h}$ as

$$
\Delta(x) = \tilde{h}(z, x) - \tilde{h}(z', x)
$$

for any $z \in \bar{Z}$ and $z' \in \bar{Z}^c$. 
By hypothetical external manipulations of the causal variables in hypothetical worlds. Notice that if there is a functional relationship connecting $Z$ and $X$ then this expression cannot be evaluated for all $x$ given $z$. $X$ and $Z$ have to be variation free to generate a treatment effect causal effect.

Any definition of a causal parameter is intrinsically incomplete unless a mechanism is defined specifying how the change in the causal variable is implemented.
Suppose that $X$ and $Z$ are not variation free. If the only way to vary $D$ is to vary $X$, then no definition of a \textit{ceteris paribus} causal effect of $D$ on $Y$ is possible.

As a consequence of (2-1a), (2-1b), and (2-4), the random variables $Y_0(\omega), Y_1(\omega)$ and $D(\omega)$ in a hypothetical population are degenerate given the causes, $X(\omega)$ and $Z(\omega)$, respectively. That is what is meant by a full accounting of causes. Most statisticians implicitly assume an unexplained source of
randomness without justifying its origin.

To account for random variables that are nondegenerate conditional on the observed covariates, a common convention in econometrics breaks $X(\omega)$ and $Z(\omega)$ into observed

$(X_o(\omega), Z_o(\omega))$ and unobserved components $(X_u(\omega), Z_u(\omega))$.

\begin{align}
(2-7a) \quad Y_0(\omega) &= g_{0o}(X_o(\omega)) + \\
& \quad g_{0u}(X_u(\omega)) \\
(2-7b) \quad Y_1(\omega) &= g_{1o}(X_o(\omega)) + \\
& \quad g_{1u}(X_u(\omega)).
\end{align}
\[ U_0(\omega) = g_{0u}(X_u(\omega)) \text{ and } U_1(\omega) = g_{1u}(X_u(\omega)) \]

\( (U_0(\omega), U_1(\omega)) \)
(2-8a) \[ E(U_0(\omega) \mid X_o(\omega)) = 0 \]

(2-8b) \[ E(U_1(\omega) \mid X_o(\omega)) = 0, \]

we thus obtain the mean outcome equations

(2-9a) \[ E(Y_0(\omega) \mid X_o(\omega)) = g_{0o}(X_o(\omega)) \]

(2-9b) \[ E(Y_1(\omega) \mid X_o(\omega)) = g_{1o}(X_o(\omega)). \]

\( g_{0u}(X_o(\omega)) \) and \( g_{1u}(X_o(\omega)) \) are conditional expectations of \( Y_0(\omega) \) and \( Y_1(\omega) \) given \( X_o(\omega) \).

Define the causal effect of a change in \( x_o \) for person \( \omega \) as
\[ \Delta Y_i(\omega) = g_i(x_o, X_u(\omega)) - g_i(x'_o, X_u(\omega)). \]

Then as a consequence of assumptions (2-7) and (2-8),
\[ E(Y_i(\omega) \mid X_o(\omega) = x_o) - E(Y_i(\omega) \mid X_o(\omega) = x'_o) = \Delta Y_i(\omega) = g_{io}(x_o) - g_{io}(x'_o). \]

Failure of (2-8a) and (2-8b) gives rise to *simultaneous equations bias*: \( E(Y_i(\omega) \mid X_o(\omega)) \neq g_{io}(X_o(\omega)) \). Even if \( X_o(\omega) \) and \( X_u(\omega) \) are independent but conditions (2-7a) and (2-7b) do not hold. Assuming that the support of \( X(\omega), \mathcal{X} \), is open and connected, and the conditional means of \( Y_i \) given \( X(\omega) \) are continuous and bounded functions of
$X(\omega)$, then for each value of $X_0(\omega)$ there is a point in the support of $X(\omega)$ say $(X_0(\omega), X^*_u(\omega))$ where, by the mean value function for integrals,
\begin{equation}
\int_{X_u} E(Y_i(\omega) \mid X_o(\omega), X_u(\omega)) g(X_u(\omega) \mid X_o(\omega)) dX_u(\omega) = E(Y_i(\omega) \mid X_o(\omega), X^*_u(\omega))
\end{equation}
1.1 The Problem of Causal Inference

It is not necessary to observe the same person in both the treated and untreated states to form the Treatment Effect Causal Effect (2-6) for that person. It is sufficient to observe two individuals with the same values of $X$ but different values of $D$. In other words, if $\omega, \omega'$ are such that $X(\omega) = X(\omega')$, and $D(\omega) = 1$ but $D(\omega') = 0$, then $Y_0(\omega)$ can be determined by $Y_0(\omega')$ and $Y_1(\omega')$ can be determined.
by $Y_1(\omega)$. Such $(\omega, \omega')$ pairs are possible if there are pairs such that $X(\omega) = X(\omega')$ but

$$Z(\omega) \in \tilde{\mathcal{Z}}, Z(\omega') \in \tilde{\mathcal{Z}}^c.$$
1.2 Reformulating the Evaluation Problem to the Population Level

\[ 0 < Pr(D(\omega) = 1 \mid X_o(\omega)) < 1 \]

(A-4) \[ 0 < Pr(D(\omega) = 1 \mid X_o(\omega)) < 1. \]

(2-11) \[ ATE(X_o(\omega) = x) \]
\[ = E(Y_1(\omega) - Y_0(\omega) \mid X_o(\omega) = x). \]

(2-12) \[ TT(X_o(\omega) = x) \]
\[ = E(Y_1(\omega) - Y_0(\omega) \mid X_o(\omega) = x, D(\omega) = 1). \]

(2-13) \[ TUT(X_o(\omega) = x) \]
\[ = E(Y_1(\omega) - Y_0(\omega) \mid X_o(\omega) = x, D(\omega) = 0). \]
These definitions do not require that $X_o(\omega) \perp X_u(\omega)$. However, in order to identify the full effect of treatment, the $X_o(\omega)$ must satisfy a no-feedback condition.

A sufficient no-feedback condition works with a counterfactual $X_{o,d}(\omega)$ process, defined as the realization of the $X_o(\omega)$ process when $D$ is fixed externally at $d$ ($D = d$, $d \in \{0, 1\}$). The no-feedback condition is that

\[(A-5) \ X_{o,1}(\omega) = X_{o,0}(\omega) \quad a.e.\]
i.e. that the realized values of the $X_o(\omega)$ process are essentially the same irrespective of the values assumed by $D$. $X_{o,1}(\omega)$ and $X_{o,0}(\omega)$ need not be stochastically independent of $(Y_0(\omega), Y_1(\omega))$: $X_{o,d} \perp \neg \perp (Y_0(\omega), Y_1(\omega))$. If (A-5) is violated, conditioning on $X_o(\omega)$ masks the full effect of $D(\omega)$ on outcomes. Then the estimated effect of treatment is marginal of its effect as it operates through the conditioning variables.

The mean finite change in $Y(w) =$
\[ D(Z(\omega))Y_1(X(\omega)) + (1 - D(Z(\omega)))Y_0(X(\omega)) \]

with respect to the finite change in the jth coordinate of \( Z_o(\omega), Z_j(\omega) \) is

\[(2-14) \quad E \left[ \frac{\Delta Y(\omega)}{\Delta Z_j(\omega)} \mid X_o(\omega) = x, Z_o(\omega) = z, \Delta Z_j(\omega) \neq 0 \right].\]

**Marginal Treatment Effect (MTE)** parameter of Heckman and Vytlacil (1999, 2000a,b,c). It is formally defined as

\[(2-15) \quad MTE(X_o(\omega) = x, Z_o(\omega) = z) = \lim_{\Delta Z_j(\omega) \to 0} E \left[ \frac{\Delta Y(\omega)}{\Delta Z_j(\omega)} \mid X_o(\omega) = x, Z_o(\omega) = z \right].\]
1.3 Relationships Among Population Treatment Effects Causal Parameters, Marshallian Causal Functions and Structural Equation Models

Structural equations

\[(2-16a) \ Y_0 = g_0 (X (\omega)) = f_0 (X_o (\omega), X_u (\omega); \theta_0)\]

\[(2-16b) \ Y_1 = g_1 (X (\omega)) = f_1 (X_o (\omega), X_u (\omega); \theta_1)\]

In the standard linear structural equations case

\[(2-17a) \ f_0 = X_o (\omega)' \theta_o + \nu_0 (\omega)\]

\[(2-17b) \ f_1 = X_o (\omega)' \theta_1 + \nu_1 (\omega)\]
Consider $ATE$:

$$E(Y_1(\omega) - Y_0(\omega) \mid X_o(\omega) = x) - \int [g_1(X_o(\omega) = x, X_u(\omega)) - g_0(X_o(\omega) = x)] \cdot dF(X_u(\omega) \mid X_o(\omega) = x)$$

$ATE$ generalizes to different populations as long as

$$E_F(U_0(\omega) \mid X_0(\omega) = x) = E_{F^*}(U_0(\omega) \mid X_0(\omega) = x).$$

$ATE$ determined in one population will apply.
1.4 Relationships with the Time Series Causality Literature

The notions of causality presented in the previous subsection all have counterparts in dynamic settings, where the variables are time dated. We present these definitions later on. Time-series notions of causality as developed by Granger (1969) and Sims (1972), are conceptually distinct and sometimes at odds with the notion of causality based on controlled variation that is used in this chapter.
The time-series literature on causality uses time dating of variables (temporal precedence relationships) to determine empirical causes and does not define or establish Marshallian *ceteris paribus* relationships. Thus letting $t$ denote time, past outcome, $Y_t$, is said to cause future conditioning variable $X_t$ if past $Y_t$ helps to predict future $X_t$, given past $X_t$ determines (“causes” in the sense of this chapter) current and past $Y_t$ as often arises in dynamic economic models. The “causality”
determined from such testing procedures does not correspond to causality as defined in this paper, and in this instance is in direct conflict with it.
1.5 The Value of Structural Equations in Making Policy Forecasts

The problem of forecasting the effects of a policy evaluated on one population but applied to another population can be formulated in the following way. Let \( Y = \varphi(X_o(\omega), X_u(\omega)) \), where \( \varphi : D \rightarrow \mathcal{Y} \), \( D \subseteq \mathbb{R}^J \), \( \mathcal{Y} \subseteq \mathbb{R} \). \( \varphi \) is a Marshallian causal function determining outcome \( Y \), and we assume that it is known only over

\[
\text{Supp}(X_o(\omega), X_u(\omega)) = \mathcal{X}_o \times \mathcal{X}_u. \quad X_o \text{ and }
\]

33
$X_u$ are the random variables in the source population. The mean outcome conditional on $X_o(\omega) = x$ is

$$E_S(Y(\omega) \mid X_o(\omega) = x) = \int_{\chi_u} \varphi(X_o(\omega) = x, X_u(\omega))dF_S(X_u(\omega) \mid X_o(\omega) = x)$$

The average outcome in the target population $(T)$ is

$$E_T(Y(\omega) \mid X_o(\omega) = x) = \int_{\chi_u} \varphi(X_0(\omega) = x, X_u(\omega))dF_T(X_u(\omega) \mid X_o(\omega) = x)$$

where $\chi_u$ is the support of $U$ in the target population. Provided the support of
\((X_o(w), X_u(w))\) is the same in the source and the target populations, from knowledge of \(F_T\) it is possible to produce a correct value of \(E_T(Y(\omega) \mid X_o(\omega) = x)\) on the target population.

Additive separability in \(\varphi\) simplifies the extrapolation problem. If \(\varphi\) is additively separable

\[
\varphi = \varphi_o(X_o(\omega)) + \varphi_u(X_u(\omega)),
\]

(1) Structural or Marshallian causal functions are determined (e.g. $\varphi(X(\omega))$ in the previous discussion).

(2) The new policy is characterized by an invertible mapping from observed random variables to the characteristics associated with the policy: $c(\omega) = q(X(\omega))$, where $c(\omega)$ is the set of characteristics associated with the policy and $q, q : \mathbb{R}^J \rightarrow \mathbb{R}^J$, is a known invertible mapping.

(3) $X(\omega) = q^{-1}(c(\omega))$ is solved to associate characteristics that in principle can be observed with the policy. This places the characteristics of the new policy on the same footing as those of the old.

(4) It is assumed that $\text{Supp}(q^{-1}(c(\omega))) \subseteq \text{Supp}(X(\omega))$. This ensures that the support of the new characteristics mapped into $X(\omega)$ space is contained in the support of $X(\omega)$. If this condition is not met, structural versions of the nonparametric Marshallian causal functions must be used to forecast the effects of the new policy, to
extend it beyond the support of the source population.
(5) The forecast effect of the policy on $Y$ is $Y_c(\omega) = \varphi(q^{-1}(c(\omega)))$.

$Supp_{post\ tax}(P(\omega)(1+t)) \subseteq Supp_{pretax}(P(\omega))$. 