Economics 6003
Quantitative Economics

1. Causation and Identification
2. Important Concepts and Tools in Time Series Analysis

John Bluedorn

Lecturer, Economics Division

Monday/Thursday, 9/12 March 2009

What is a causal effect?

- The causal effect of an intervention $x \in \{0, 1\}$ (e.g., an extra year of education for a person, a shift from a fixed exchange rate regime to a floating exchange rate regime, etc.) for a unit i (e.g., a person, a country, etc.) of a population upon an outcome y (e.g., a person's wage rate, a country's economic growth rate, etc.) is defined to be the difference between their outcome if the intervention occurs and their outcome if the intervention does not occur.
  - Note that the intervention here is binary. In general, it need not be. Let $x = 0$ denote the *control* intervention and $x = 1$ denote the *treatment* intervention.
  - For simplicity, we assume that there are *no interactions across units*. In other words, unit i's response does not depend upon unit j's intervention status. The assumption is referred to as the stable unit treatment value assumption (SUTVA).

Potential outcomes and causal effects

- Notationally, it is:

$$c_i = y_i\,(x = 1) - y_i\,(x = 0)\,,$$

where c is the causal effect for individual i of intervention x.

- The **fundamental problem of causal inference** is that we only observe one of the quantities which form the causal effect. Why? If an individual/unit receives the treatment, then the possibility that the individual does *not* receive the treatment at that point in time is closed off (and *vice versa*).
- We refer to this framework for thinking about causation as the *potential outcomes* (sometimes known as the *counterfactuals*) framework.

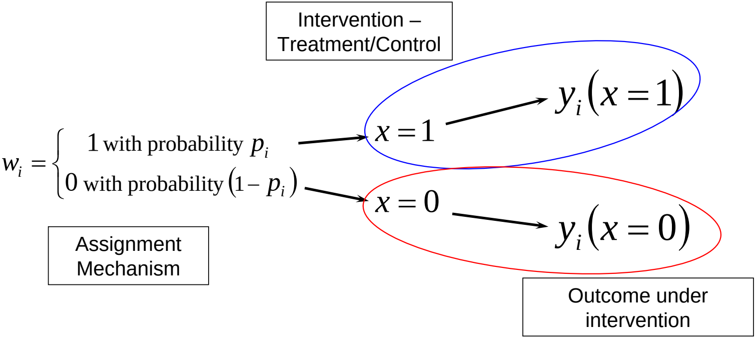
Potential outcomes and causal effects

- The intuition that such a way of thinking about causation is useful dates back centuries (albeit informally).
- The modern formulation is originally due to Neyman (1923) in the context of experimental studies (random assignment of the intervention). Its modern form, in particular as applied to observational studies (nonrandom assignment of the intervention), is due to Rubin (1974).
- Taking a historical perspective, we compare the outcome under the intervention with what *would have happened* if the intervention had not been applied – the *counterfactual*, which is inherently *un*observable.

How to proceed?

- We need to figure out how to construct an estimate of the *un*observed (or missing) outcome for unit i.
- The general strategy is to compare the outcomes of units who are treated with similar units who are not.
  - In the absence of an exact match of units across treatment/control, we consider the *average* causal effect for the population $\Rightarrow c = \bar{y}(x = 1) - \bar{y}(x = 0)$.
- In addition to SUTVA, we need assumptions on *how* the intervention is assigned to units. Denote the assignment mechanism by $w \in \{0, 1\}$. Note that w is either 0 or 1 for an individual, while $x = 0$ and $x = 1$, as *potential* outcomes both exist for each individual.

An individual's potential outcomes tree . . .



Potential Outcomes Framework with Assignment of Binary Intervention to individual *i* (SUTV assumed)

How to proceed?

- We can distinguish 2 types of assignment mechanism w.
  - If assignment is *random*, then individual characteristics and the intervention are *independent* (viz., unrelated) $\Rightarrow p_i$ does not depend upon individual i or $y_i(1), y_i(0)$. There is *no* confounding of differences in individuals with differences in outcomes.
  - If assignment is *nonrandom*, then individual characteristics and the intervention are *dependent* (viz., related) $\Rightarrow p_i$ depends upon individual i or $y_i(1), y_i(0)$. Differences in individuals may confound differences in outcomes, *even* conditional upon the intervention.
- With nonrandom assignment, we must correct for the probability that a unit i receives the treatment. To accomplish this, we must assume that p_i is *only* dependent upon *observables*. If unobservables drive p_i, then we are out of luck; it is not feasible to correct for the treatment probability.

In observational studies?

- Whether or not SUTVA is a reasonable assumption depends upon the particular research question.
    - If the intervention is an additional year of schooling, then SUTVA seems reasonable.
    - If the intervention is a shift from a fixed to a floating exchange rate regime, then SUTVA seems less reasonable (the effects of the exchange rate regime choice likely depend upon what other countries' exchange rate regimes are).
- Economists usually deal with observational data, which means that random assignment is not guaranteed to hold. For economists, an intervention is an *explanatory variable* of interest.
    - The acquisition of an additional year of schooling is an individual choice. It is *not* randomly assigned.
    - The shift from a fixed to a floating exchange rate regime is a policy decision made by a country. It is *not* randomly assigned.

Response to such threats?

- How should an empirical researcher address the possible failure of SUTVA and random assignment?
  - If SUTVA is thought to fail, then the appropriate response is to incorporate the postulated interactions directly into the model. In this manner, we account for the possible spillovers.
    - For example, we include the exchange rate regime of neighbors as an explanatory variable for a country's exchange rate regime choice. Of course, this leads to a possible simultaneity problem which must be considered.
  - If random assignment is thought to fail, then the appropriate responses are:
    - Find a plausible *natural* experiment, where random assignment does apply. This is usually tough, but extremely useful if successful.
    - Model the assignment mechanism explicitly, thereby correcting for the nonrandom assignment. This can be straightforward to accomplish, but difficult to defend convincingly.

Identification in econometric models

- Pearl (2000) demonstrated that structural equation modeling (SEM), which is typical in economics (think about instrumental variables), is equivalent to making a set of assumptions on the assignment mechanism and interactions in the potential outcomes framework.
  - The structural restrictions in an SEM allow for us to interpret the estimated effects as causal effects.
  - The validity of the causal inference depends upon the structural restrictions being true.
- In practice, the structural restrictions in econometrics often take the form of *orthogonality* assumptions.
  - For example, in OLS, we require that: $E(X'\varepsilon) = 0$, where $E(\cdot)$ denotes the expectations operator, X denotes the set of linear explanatory variables, and ε denotes the latent, mean-zero error term.
  - In TSLS, we require that: $E(Z'\varepsilon) = 0$ and $E(Z'X) \neq 0$, where Z denotes the set of instruments.
  - Orthogonality and independence are closely related concepts (but not identical).

Identification in econometric models

- As we proceed, we will return to the potential outcomes/counterfactuals framework when we evaluate the *economic meaning* of econometric identification assumptions for causal inference.
  - In OLS research design, we require that conditional upon any other explanatory variables, the explanatory variable of interest is *unrelated* to latent unit characteristics $\Rightarrow$ similar to random assignment.
  - A similar requirement is used in a TSLS research design, where the instrument takes the place of the explanatory variable of interest.
- Identification is the set of assumptions required to recover a unique representation for a model from the data.
- Whether or not the identification assumptions used in a study are plausible is a matter of judgment (the *jurisprudence* of economics). Thinking about the relevant potential outcome/counterfactual may help us make this judgment.

Important Concepts and Tools in Time Series Analysis

- There are a few conceptual ingredients and tools that we need before we dive into the practice of time series analysis. They are:
  1. Stochastic processes and random sequences
  2. Ergodicity – of a distribution and of a moment
  3. Stationarity – strong (strict) and weak (covariance)
  4. Linear difference equations and their solution

Stochastic processes and random sequences

- Most simply, a *stochastic process* is any process whose evolution we can follow and predict (it is ordered) in terms of probability (it is random). When ordered by a discrete index, we have a *random sequence*. When the discrete index is time t, the random sequence is a *time series*.
    - A dataset on a random variable Y_t of sample size T:

      $$\{y_1, y_2, \ldots, y_T\}$$

      $\{Y_t\}$ is the stochastic process. The sample/dataset drawn from the stochastic process are the *realizations*.
    - If $E(Y_t) = 0 \forall t$, $E\left(Y_t^2\right) = \sigma^2$ where σ is some constant, and $E(Y_t Y_s) = 0 \forall t \neq s$, then we call Y_t a *white noise* process.
- More formally, a stochastic process is an indexed/ordered collection of random variables defined over some probability space.

Ergodicity

- The concept of ergodicity is a tricky one, requiring a bit of preparation via a thought experiment.
  - The *many-worlds* interpretation of quantum mechanics states that reality is structured like a giant decision tree, where each branch represents a particular *worldline* or sequence of realizations. Each branch is weighted with a probability (derived from the quantum theory).
  - Suppose that we draw a set of K trajectories/worldlines from our stochastic process $\{Y_t\}$ and index them by k:

$$\{Y_t^{(k)}\}_{t=-\infty,\infty}^{k=1,K}$$

    Each k when realized represents a single history for the stochastic process $\{Y_t\}$. The collection of K worldlines is known as an *ensemble*.

- The process $\{Y_t\}$ is *ergodic* if the distributional properties of any worldline k are representative of the properties of all the other possible worldlines for $\{Y_t\}$.

Ergodicity

- This implies that the distributional properties of $\{Y_t\}$ over large enough ranges of time do not depend upon the *initial conditions*.
- Ergodicity is thus sometimes referred to as *asymptotic independence*.
- Usually, a weaker set of conditions where the process $\{Y_t\}$ is ergodic for some set of its moments is sufficient. For example, practically speaking, we always need for $\{Y_t\}$ to be *ergodic for the mean* $\Rightarrow$ the ensemble average of Y and the time average of Y converge:

$$\lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K} Y_{t_0}^{(k)} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=-T}^{T} Y_t^{(k_0)} = \mu$$

for any particular worldline k_0 and instant t_0, where $\mu = E\left(Y_t^{(k)}\right)$ is some constant.

Ergodicity

- A similar logic can be readily applied to other moments, such as second moments or variance/covariances.
- What would a failure of ergodicity mean?
  - there is strong *path dependence* $\Rightarrow$ where we have been or started from, has an irrevocable influence on where we can go. Note how nonergodicity is closely related to chaos.
  - the data we have, which is drawn from a single worldline, is not revelatory about the true properties of $\{Y_t\}$.
    - Example: Let $Y_t^{(k)} = \mu^{(k)} + \varepsilon_t$, where μ and ε are mean-zero and independent for all t, k. This means that $E\left(Y_t^{(k)}\right) = 0$. But, for a given worldline k_0, the time average does *not* converge to zero $\Rightarrow$:

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=-T}^{T} Y_t^{(k_0)} = \mu^{(k_0)} \neq 0$$

- Note the conceptual similarity between the many worlds interpretation and the potential outcomes framework.

Stationarity

- A stochastic process $\{Y_t\}$ is *strictly (strongly) stationary* if:

$$\left(Y_{t+h_1}, Y_{t+h_2}, \ldots, Y_{t+h_H}\right) \stackrel{\text{dist}}{=} \left(Y_{s+h_1}, Y_{s+h_2}, \ldots, Y_{s+h_H}\right)$$

$\forall t, s$ and any given $h_1, \ldots h_H$.

In other words, the finite dimensional marginal distributions of the process $\{Y_t\}$ are *calendar-invariant* or translation-invariant; they do not depend upon the date t. The distribution is constructed with respect to the set of possible worldlines.

- Define the *autocovariance function* of a stochastic process $\{Y_t\}$ as:

$$\gamma_{YY}(t, h) = E\left\{\left[Y_t - E(Y_t)\right]\left[Y_{t-h} - E(Y_{t-h})\right]\right\}$$

Note how if $h = 0$, this is just the variance of the random variable Y_t.

Stationarity

- A stochastic process $\{Y_t\}$ is *covariance (weakly) stationary* if:

$$\begin{aligned} E(Y_t) &= \mu \forall t \text{ and } \mu \text{ finite} \\ \gamma_{YY}(t,h) &= \gamma_h \forall t \text{ and } \gamma_h \text{ finite} \end{aligned}$$

  In other words, the first and second moments of the process $\{Y_t\}$ do *not* depend upon the calendar date/time t. They are *calendar-invariant*. The second moments (autocovariances) only depend upon the time shift h. Again, the underlying distribution is constructed with respect to the set of possible worldlines.
- For most applied work, we only require that covariance or weak stationarity hold, since we will usually only work with the first and second moments of a process. Thus, we will use *stationary* to loosely mean weakly stationary.

Ergodicity + Stationarity

- Unless we explicitly note otherwise, henceforth we will always assume that the stochastic process $\{Y_t\}$ is:
  1. ergodic for first and second moments $\Rightarrow$ time averages converge to the ensemble averages, meaning that they accurately reflect the first and second moments of the stochastic process
  2. weakly stationary (calendar-invariant first and second moments)
- With these assumptions, we can take the empirical analogs of the first and second moments of the stochastic process $\{Y_t\}$, (the time averages), confident that they *converge* to the *true* moments *regardless* of the point in time at which we start the averaging.

Ergodicity + Stationarity

- The three concepts of ergodicity, strong stationarity, and weak stationarity are *distinct*.
  - strong stationarity $\nRightarrow$ weak stationarity, *unless* first and second moments are finite. Strong is about distributions, while weak is about first two moments. Clearly, weak $\nRightarrow$ strong.
  - weak stationarity $\nLeftrightarrow$ ergodic for first two moments *unless* all worldlines have identical first two moments. Weak is about calendar-invariance of first two moments, while ergodic for first two moments is about convergence of the time/ensemble averages. Clearly, ergodic $\Rightarrow$ weak.
- However, with more assumptions, we can sometimes link the concepts:
  - If we have a weakly stationary process $\{Y_t\}$ and $\sum_{h=0}^{\infty} \|\gamma_h\| < \infty$, then $\{Y_t\}$ is ergodic for the mean. The autocovariances shrink quickly enough that the memory of the process is limited enough that time averages equal ensemble averages.

Linear difference equations

- A *difference equation* is the discrete-time analog of a continuous time differential equation (both are equations of motion).
- It describes how the process's realization in the next instant is some function of past realizations (and possibly some forcing variables).
- A *linear* difference equation states that the process's next realization is a linear function of the past realizations. Examples:

$$\begin{aligned} Y_t &= a + bY_{t-1}, \text{ AR(1) process} \\ Y_t &= a + bY_{t-1} + cY_{t-2}, \text{ AR(2) process} \end{aligned}$$

  where a, b, c are constants.
- If they are stochastic from the perspective of the past, then there is an forcing variable which is added; it is known as the *innovation*.

How to solve?

- We'll use lag operators to work through solving linear difference equations. When we use them, we are implicitly invoking the Riesz-Fischer Theorem and facts about Fourier and z-transforms (see Sargent, 1987, chapter 11 for the full details).
- Denote the lag operator by L. When applied to a variable, it has the effect of shifting its time index one-step backwards: $LY_t = Y_{t-1}$.
- It is a linear operator, so that $LaY_t = aLY_t = aY_{t-1}$ for a constant a and $L(Y_t + Y_{t-1}) = LY_t + LY_{t-1} = Y_{t-1} + Y_{t-2}$.
- Define $L^n = \prod_{i=1}^{n} L$, an application of the lag operator n times. If we take a negative power, it shifts the time index forward $\Rightarrow L^{-1}Y_t = Y_{t+1}$.
- With these rules, we can construct polynomials in the lag operator.

How to solve?

- Our general strategy is to take a difference equation, factor it in terms of lag operators, and then divide out the lag operator terms somehow, eliminating the dependence of Y_t on its own past.
- We'll use the fundamental theorem of algebra to help us factor the lag polynomials.
- For an AR(1) difference equation:

$$\begin{aligned} Y_t &= a + bY_{t-1} \\ Y_t &= a + bLY_t \\ (1 - bL)\, Y_t &= a \end{aligned}$$

- Now, we are stuck unless we can determine what $(1 - bL)^{-1}$ means (this is the divide out step).

How to solve?

- By inspection, it sort of looks like the solution to a geometric series. We postulate:

$$\begin{aligned}(1-bL)^{-1} &= \sum_{i}^{\infty}(bL)^i \\ &= 1 + bL + b^2L^2 + \cdots\end{aligned}$$

assuming that $|b| < 1$ for the series to converge.

- We can verify that this is correct by noting that:

$$\begin{aligned}\frac{1-bL}{1-bL} &= (1-bL)\left(\sum_{i}^{\infty}(bL)^i\right) \\ &= \left(1 + bL + b^2L^2 + \cdots\right) \\ &\quad -bL\left(1 + bL + b^2L^2 + \cdots\right) \\ &= 1.\end{aligned}$$

How to solve?

- So, the AR(1) difference equation solution is:

$$Y_t = \frac{a}{1-b} + c_0 b^t$$

where we have used the fact that a is a constant (so it doesn't change with the lag operator). c_0 is some unknown constant.

- What is this additional $c_0 b^t$ term? Note that:

$$\begin{aligned}(1-bL)\, c_0 b^t &= c_0 \left(b^t - bLb^t\right) \\ &= c_0 \left(b^t - bb^{t-1}\right) = 0,\end{aligned}$$

which is consistent with the difference equation.

- In general, we need an additional restriction to get rid of this "bubble" term (e.g., a particular value for Y_t at some time or boundedness of the path of Y_t). We'll usually assume that some such conditions hold.

How to solve?

- What do we do if $|b| > 1$? We can use a clever rewriting of divide-out (inversion) step:

$$
\begin{aligned}
\frac{1}{1-bL} &= \frac{-(bL)^{-1}}{1-(bL)^{-1}} \\
&= \left(\frac{-1}{bL}\right)\left(\frac{1}{1-\left(\frac{1}{b}\right)L^{-1}}\right) \\
&= -\left(\frac{1}{b}\right)L^{-1} - \left(\frac{1}{b}\right)^2 L^{-2} - \left(\frac{1}{b}\right)^3 L^{-3} - \cdots
\end{aligned}
$$

which will converge since $\left|\frac{1}{b}\right| < 1$.

- Note how this means that the present t depends upon the future potentially! Such expressions naturally arise in forward-looking models.

Factoring higher order polynomials ...

- For the AR(2) case, it gets more complicated:

$$\begin{aligned} Y_t &= a + bY_{t-1} + cY_{t-2} \\ Y_t &= a + bLY_t + cL^2 Y_t \\ \left(1 - bL - cL^2\right) Y_t &= a \end{aligned}$$

- Now what? We don't know how to invert a polynomial higher than order 1 (which we just did). The fundamental theorem of algebra helps us out, since it states that we can always factor a polynomial out like:

$$\begin{aligned} &\left(a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n\right) \\ = \quad & a_0 \left(1 - \lambda_1 x\right)\left(1 - \lambda_2 x\right) \cdots \left(1 - \lambda_n x\right) \end{aligned}$$

where the as are the coefficients, x is the variable, and the λs represent the roots of the equation. The domain of the coefficients, variables, and roots are assumed to be the complex numbers. Roots may repeat.

Factoring higher order polynomials ...

- Note that this means that:

$$\begin{aligned}\left(1 - bL - cL^2\right) &= \left(1 - \lambda_1 L\right)\left(1 - \lambda_2 L\right) \\ &= \left(1 - \lambda_1 L - \lambda_2 L + \lambda_1 \lambda_2 L^2\right) \\ &= 1 - \left(\lambda_1 - \lambda_2\right) L + \lambda_1 \lambda_2 L^2\end{aligned}$$

which implies that:

$$\begin{aligned} b &= \lambda_1 + \lambda_2 \\ c &= -\lambda_1 \lambda_2 \end{aligned}$$

- So AR(2) models are not too bad to solve via substitution and the quadratic formula.
- Then, we can apply our usual divide-out strategy, if the roots are less than one in modulus (they might be complex and the modulus is like the absolute value). Otherwise, we have to do the trick we used for $|b| > 1$ to rewrite and solve forwards.

Factoring higher order polynomials . . .

- Note that this will get pretty tricky if we get above a fourth-order difference equation.
- Why? For the AR(2), we had to use the quadratic formula. Unfortunately, no explicit formula to solve polynomials that are of an order higher than 4 exists (it has been proven). So, it has to be done numerically.

Summary

- Definition of the causal effect of an intervention – a comparison of potential outcomes/counterfactuals to the intervention.
  - fundamental problem of causal inference – we only observe one outcome for an individual.
  - with additional assumptions, we can recover an estimate of the causal effects (e.g., SUTVA, assignment mechanism).
- Identification is the set of assumptions (e.g., SUTVA, assignment mechanism) that are required to recover a unique representation for a model from the data.
- We always need to ask what the economic meaning of the identification assumptions we make in empirical work is. This helps us gauge their reasonableness.

Summary

- Stochastic process – an indexed/ordered collection of random variables defined over some probability space.
- White noise
- Ergodicity – a kind of asymptotic independence, where the distributional properties of a stochastic process do not depend upon the initial conditions. Recall the many-worlds interpretation.
- Ergodic for the mean – the ensemble and time averages converge.
- Strong stationarity – marginal distributions are calendar-invariant.
- Weak stationarity – first and second moments are calendar-invariant.
- Linear difference equations and lag operators – recall the restrictions on the modulus of roots and on a particular value of the process that we make in the solution.