Motivation

Thomas Wiemann
University of Chicago

Econometrics
Econ 21020

Updated: March 28, 2022

Outline

1. Learning Goals of the Course

2. On the Analysis of Causal Questions

   ▷ Descriptive versus causal questions

   ▷ Example: Returns to higher education

Outline

1. **Learning Goals of the Course**
2. On the Analysis of Causal Questions
   - ▷ Descriptive versus causal questions
   - ▷ Example: Returns to higher education

What is Econometrics?

Econometrics is a field at the intersection of economics, statistics and mathematics, created out of the desire to quantify economic theory using data.

Jan Tinbergen, one of the founders of the field, suggests:
"*Econometrics could be defined as 'statistical observation of theoretically founded concepts,' or, alternatively, 'mathematical economics working with measured data.'* (Tinbergen, 1954, p. 10)"

Note that he does not simply define econometrics as "the application of statistical methods to economic data." As Qin Duo puts it in her book:
"[...] *econometrics is not merely a tool-kit for economics but a subject rich in ideas and theories which have the potential to revolutionize economics.* (Duo, 1993, p. 1)"

One such topic that we will spend lots of time on is a formal theory for the analysis of causal questions.

Learning Goals of the Course

This is a course in econometrics: We will draw from economics, mathematics (in particular, probability theory), *and* statistics.

After the course, you are (hopefully) able to

- understand and apply basic concepts in probability theory and statistics;
- differentiate between descriptive and causal questions;
- combine economics and probability theory to define causal parameters;
- develop logically consistent causal identification arguments;
- construct and evaluate regression-based estimators of causal parameters;
- draw insights from real data using R.

Learning Goals of the Course (Contd.)

To achieve the learning goals, the course is multifaceted:

- Lectures present and discuss key concepts.
- Problem sets challenge you to apply the learned concepts. This is meant to deepen your understanding of lecture material as well as to let you develop new insights on your own.
- TA sessions cover programming in R, work through example exercises, and review the problem sets and the midterm.

The goal of this course is not to be another letter on your transcript!

- I hope you will *learn* something useful, regardless of whether you continue in academia, the private sector, or the public sector.
- If you have questions, please don't hesitate to reach out. Send a message via Slack, or come to the weekly office hours.

Outline

1. Goals of the Course

2. **On the Analysis of Causal Questions**

   ▷ **Descriptive versus causal questions**

   ▷ Example: Returns to higher education

Descriptive versus causal questions

We will draw a strong contrast between *descriptive* and *causal* questions.

Descriptive questions are concerned with the *realized* state of the world:

- ▷ *What is* the median income in the United States?
- ▷ *What is* the number of COVID-19 cases in Chicago?
- ▷ *What is* the size of the gray/black market in Illinois?

"Descriptive" does not mean easy. Indeed, large resources are required to answer any of the three questions above reliably.

Descriptive questions are also encountered in everyday-life:

- ▷ *What is* the amount of weight I gained over the Winter break?
- ▷ *What is* the amount of money I spent on coffee each week?
- ▷ *What is* the number of hours needed to solve the problem sets?

Descriptive versus causal questions (Contd.)

Causal questions are concerned with the *potential* states of the world:

- ▷ *What* would the median income in the United States be *if* there were no minimum wage laws?
- ▷ *What* would the number of COVID-19 in Chicago be today *if* there had been no the Winter mask mandates?
- ▷ *What* would the size of the gray/black market in the Illinois be *if* recreational drugs (e.g., alcohol) were to be de-legalized?

Causal questions are also encountered in everyday-life:

- ▷ *What* would the amount of weight I gained over the Winter break be *if* had I signed up for the gym?
- ▷ *What* would the amount of money I spent on coffee each week be *if* I buy an espresso machine?
- ▷ *What* would the number of hours needed to solve the problem sets be *if* students study with peers?

Descriptive versus causal questions (Contd.)

A key insight we will develop is that data alone *cannot* suffice to draw causal conclusions.

- Data w/o assumptions allows for descriptive statements.
- Data w/ assumptions *can* allow for causal conclusions.

Causal conclusions are often more interesting because they can help evaluate and guide decisions, descriptive statements alone cannot.

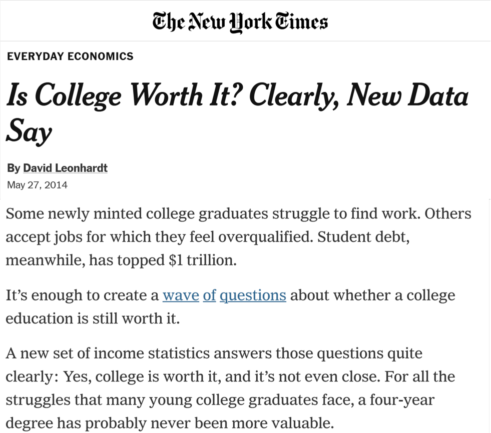
- Were the Chicago Winter mask mandates effective?
- Should you form study groups for this course or work in solitary?

But careful causal reasoning is challenging.

- Descriptive statements are often confused for causal conclusions in public discourse and private discussions.
- Errors may misguide actions in important decision problems.

Descriptive and causal questions in the New York Times

Consider this New York Times article on the returns to education:



The New York Times

EVERYDAY ECONOMICS

Is College Worth It? Clearly, New Data Say

By David Leonhardt
May 27, 2014

Some newly minted college graduates struggle to find work. Others accept jobs for which they feel overqualified. Student debt, meanwhile, has topped $1 trillion.

It's enough to create a wave of questions about whether a college education is still worth it.

A new set of income statistics answers those questions quite clearly: Yes, college is worth it, and it's not even close. For all the struggles that many young college graduates face, a four-year degree has probably never been more valuable.

Notes. Screenshot from a NYT article available here: link. This example is inspired by lectures of Prof. Alexander Torgovitsky.

Descriptive and causal questions in the New York Times (Contd.)

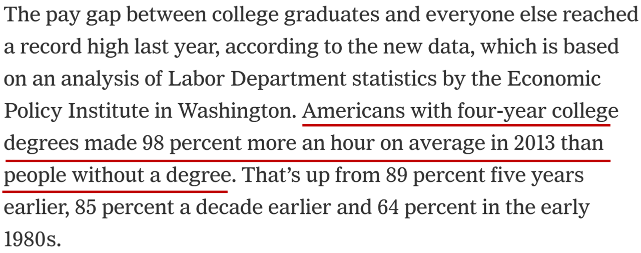
The author is concerned with the decision problem on whether or not to pursue a college education in the United States.

A causal question relevant for this decision is whether or not a college degree is (financially) worth it.

The headline says college is “clearly” worth it.

- ▷ What is this conclusion based on?

Descriptive and causal questions in the New York Times (Contd.)



The pay gap between college graduates and everyone else reached a record high last year, according to the new data, which is based on an analysis of Labor Department statistics by the Economic Policy Institute in Washington. Americans with four-year college degrees made 98 percent more an hour on average in 2013 than people without a degree. That's up from 89 percent five years earlier, 85 percent a decade earlier and 64 percent in the early 1980s.

Notes. Screenshot from a NYT article available here: link.

That is a descriptive statement!

- Would the college graduates have earned less if they had not pursued a higher education?
- Would the non-college graduates have earned more if they had a college degree?

From the data alone, it's impossible to tell: let's see why.

Outline

1. What is Econometrics?

2. Goals of the Course

3. **On the Analysis of Causal Questions**

    ▷ Descriptive versus causal questions

    ▷ **Example: Returns to higher education**

Example: Returns to higher education

The final part of today's lecture is a brief introduction to the analysis of causal questions using the New York Times-example.

- ▷ Use the returns of education example to illustrate a formal theory for causal analysis.
- ▷ Highlight why causal conclusions can't be drawn from data alone, and which assumption could be combined with the data to allow for a causal conclusion.
- ▷ Relies on economic intuition, basic probability theory, and statistics that should have been covered in the prerequisites.

Don't worry if this introduction seems very challenging today:

- ▷ If it does, that's an excellent motivation for our review of probability and statistics in the next few lectures!
- ▷ We will revisit today's example in lectures 6 & 7 after the review.

Three distinct tasks arising in the analysis of causal questions

A careful causal analysis requires three distinct tasks:

Table 1: Three distinct tasks in the analysis of causal questions

Task	Description	Requirements
1	Definition of counterfactuals (or hypotheticals)	A scientific theory
2	Parameter identification from a hypothetical population	Mathematical analysis
3	Parameter estimation and inference from real data	Statistics

Notes. Paraphrased Table 1 from Heckman and Vytlacil (2007).

- ▷ Definition of counterfactuals develops quantified versions of the "what if" questions (i.e., expressed using mathematics).
- ▷ Identification concerns the task of linking counterfactuals to known functions of observables in a logically consistent manner.
- ▷ Estimation and inference is concerned with calculation of the counterfactuals from a finite number of data points, as well as assessing the associated sampling uncertainty.

***Note**: The tasks may occasionally be referred to simply as (1) definition, (2) identification, and (3) estimation.*

Example: Returns to higher education (Contd.)

To illustrate the three tasks, consider the question on the returns of education discussed in the New York Times article.

A key question of interest (for both you and I) in this context is the returns of a college degree for college graduates:

▷ *What is the change in hourly wages for college graduates if they had not pursued higher education?*

Knowing the answer would allow us to to evaluate our life decisions. It may also guide decisions of those who are considering college.

We are now in need of a formal language that allows us to be logically consistent when thinking about causal questions. This formal language is mathematics (and in particular probability theory).

Task 1: Definition

To quantify this “what if” question, we develop an economic model of hourly wages expressed in mathematical terms.

- ▷ Denote hourly wages by Y.
- ▷ Denote having a college degree by $W = 1$, and $W = 0$ otherwise.

Now formulate an economic model g that relates W to Y. A very general version of such a model is

$$Y = g(W, U), \tag{1}$$

where U are all determinants of Y other than W.

Economic theory is crucial for formulating and interpreting (1).

- ▷ Helps construct a set of variables that are in U. For example, talent or intellect (or connections, or luck...).
- ▷ Helps assess which restrictions on the model g are sensible or not sensible. For example, suppose you are considering to assume $g(W, U) = W\beta + U$. Do you think that is a sensible restriction?

Task 1: Definition (Contd.)

Since at least Alfred Marshall's Principles in Economics (1890), economists are keen in studying problems in isolation, holding all other determinants constant. This known as the *ceteris paribus*-principle. In the context of (1), we apply the principle by holding the other determinants of hourly wages fixed.

Consider a particular individual with wage $Y = y$, who obtained a college degree $W = 1$, and whose other determinants of wages are $U = u$.

- ▷ u denotes, for example, a specific level of talent and intellect.
- ▷ $g(1, u)$ denotes her wages if she had obtained a college degree.
- ▷ $g(0, u)$ denotes her wages if she had not obtained a college degree.

Since the individual did obtain a college degree, we have the fact that $g(1, u) = y$. In contrast, $g(0, u)$ is a counterfactual whose value is unkown to us.

Task 1: Definition (Contd.)

The return to a college degree for the individual can now be defined as

$$g(1, u) - g(0, u), \tag{2}$$

whose value is unknown because $g(0, u)$ is unkown.

More generally, we can define the return to higher education as a function of the other determinants of hourly wages:

$$\tau(U) = g(1, U) - g(0, U). \tag{3}$$

Note that for a particular individual with $U = u$, we only ever observe either $g(1, u)$ or $g(0, u)$ but never both. Hence it is impossible to observe $\tau(u)$ for any individual. This is known as *the fundamental problem of causal inference* (Holland, 1986).

Task 1: Definition (Contd.)

The fundamental problem of causal inference is a key challenge in the social sciences:

- If all other determinants U were known, then we could compare two individuals – one college graduate and one non-graduate – who have the same value $U = u$ to calculate $\tau(u)$.
- A key characteristic of social sciences is that U is beyond our full understanding. What are *all* other determinants of hourly wages?

We consider (Y, W) to be *observables* and U to be *unobservables*.

- You can think of observables as variables about which you could make descriptive statements. For example, you could assess the hourly wages of college graduates and non-graduates.
- You can think of unobservables as variables about which you could not make descriptive statements. For example, you could not assess *all* other determinants of hourly wages.

Task 1: Definition (Contd.)

We leverage probability theory to make progress despite this difficulty.

- ▷ Model (Y, W, U) as random variables to capture the idea that we are working with both observables and unobservables.
- ▷ U is an unobservable and hence not known with certainty.
- ▷ (Y, W) are observables but modeled as functions of determinants that are not fully observable themselves.

This approach using probability theory has proven to be invaluable in the analysis of causal questions. It allows for the formulation of counterfactuals that are agnostic towards the unobservables U (in some appropriate manner).

For example,

$$E_U[\tau(U)] = E_U[g(1, U) - g(0, U)], \tag{4}$$

gives the *expected* returns to higher education.

Task 1: Definition (Contd.)

Our causal question of interest is concerned with the returns to higher education for college graduates. We can adapt (4) appropriately using basic notation for expectation operators:

$$E_U[\tau(U)|W = 1] = E_U[g(1, U) - g(0, U)|W = 1], \tag{5}$$

which is the expected returns to higher education for college graduates.

We will proceed with (5) as our object of interest.

- ▷ In social sciences, statements about (3) are nearly impossible.
- ▷ Choose the slightly easier object (5) instead.

This concludes the first task: *Definition* of the counterfactual of interest.

Task 2: Identification

Note that modeling (Y, W, U) as random variables allowed us to formulate a useful counterfactual, but did not address the fundamental problem of causal inference!

Using basic probability calculus, we have

$$\begin{aligned} E_U[\tau(U)|W = 1] &= E_U[g(1, U) - g(0, U)|W = 1] \\ &= E_U[g(1, U)|W = 1] - E_U[g(0, U)|W = 1] \\ &= E_Y[Y|W = 1] - E_U[g(0, U)|W = 1]. \end{aligned} \tag{6}$$

- ▷ $E_Y[Y|W = 1]$ is an expression involving only the observables (Y, W). Here, it is the expected hourly wage of college graduates.
- ▷ $E_U[g(0, U)|W = 1]$ is an expression involving the unobservable U. Here, it is the expected hourly wage of college graduates if they had not pursued higher education. There's no hope of knowing this since there are no college graduates without a college degree.

Task 2: Identification (Contd.)

The task of identification establishes a bridge between the counterfactual of interest and the observables.

Here, Equation (6) showed that $E_U[\tau(U)|W=1]$ is a function of $E_Y[Y|W=1]$ and $E_U[g(0,U)|W=1]$, where only the latter involves the unobservables U.

The goal of the identification analysis is now to express $E_U[g(0,U)|W=1]$ as a function of only the observables (Y, W).

- ▷ Because of the fundamental problem of causal inference, this requires assumptions!

In this course, we will consider three key identifying assumptions:

- ▷ Random assignment;
- ▷ Selection on Observables;
- ▷ Instrumental Variables.

Which assumption is plausible crucially depends on the economic setting.

Task 2: Identification (Contd.)

Today, we only consider the strongest one: Random assignment.

Assumption 1 (Random assignment; RA)

Let (Y, W, U) be random variables with joint distribution characterized by

$$\begin{aligned} & Y = g(W, U), \\ \text{and} \quad & W \perp\!\!\!\perp U, \end{aligned} \tag{7}$$

where $g : \operatorname{supp} W \times \operatorname{supp} U \to \operatorname{supp} Y$.

In the returns to education context, Assumption RA states that obtaining a college degree is independent of all other determinants of hourly wages.

- ▷ Not particularly plausible... but will proceed with it today.
- ▷ Don't worry, we'll make things *more* complicated after the midterm!

Task 2: Identification (Contd.)

Under Assumption RA, identification relies on the following result:

Corollary 1

Let (W, U) be random variables such that $W \perp\!\!\!\perp U$. Then

$$E_U[h(U)|W] = E_U[h(U)], \tag{8}$$

for all functions h such that $E_U[|h(U)|] < \infty$.

Proof.

Left as a self-study exercise. (Hint: Assume that (W, U) have a joint probability density $f_{w,u}(w, u)$ with marginals $f_w(w)$ and $f_u(u)$, and apply the law of the unconscious statistician.) □

Task 2: Identification (Contd.)

Under Assumption RA, the identification proof then is

$$\begin{aligned} E_U[\tau(U)|W=1] &\overset{(1)}{=} E_Y[Y|W=1] - E_U[g(0,U)|W=1] \\ &\overset{(2)}{=} E_Y[Y|W=1] - E_U[g(0,U)|W=0] \\ &\overset{(3)}{=} E_Y[Y|W=1] - E_Y[Y|W=0]. \end{aligned} \tag{9}$$

Here,

- ▷ (1) follows from Equation (6);
- ▷ (2) follows from Assumption RA and Corollary 1;
- ▷ and (3) follows from the model in Equation (1).

This shows that knowing $(E_Y[Y|W=1], E_Y[Y|W=0])$, suffices for knowing $E_U[\tau(U)|W=1]$.

Because $E_Y[Y|W=1]$ and $E_Y[Y|W=0]$) are known functions of only the observables, we say that under Assumption RA, $E_U[\tau(U)|W=1]$ *is identified* (in the sense of Hurwicz, 1950).

Task 3: Estimation

The identification analysis showed that it suffices to know

- ▷ the expected hourly wage of college graduates ($E_Y[Y|W=1]$), and
- ▷ the expected hourly wage of non-college graduates ($E_Y[Y|W=0]$).

The final task in the analysis of causal questions is concerned with estimating these expressions of observables from real data and quantifying the associated sampling uncertainty.

To make progress, we now introduce the final pillar of econometrics into our toolbox: Statistics

Task 3: Estimation

Suppose that we observe a sample of size n: $\{(y_i, w_i)\}_{i=1}^n$.

- ▷ This is a collection n individuals, where y_i denotes the ith individual's hourly wage and w_i denotes whether or not they have a college degree.

For the data to be useful, we assume that the data is a sample from the random variables (Y, W) discussed previously.

- ▷ $(y_i, w_i) \sim (Y, W)$, for $i = 1, \ldots, n$.
- ▷ Note that we don't have a sample of the unobservables U.

We can now construct estimates of $E_Y[Y|W = 1]$ and $E_Y[Y|W = 0]$:

$$\widehat{E}_Y^{(n)}[Y|W = 1] = \frac{1}{\sum_{i=1}^n w_i} \sum_{i=1}^n y_i w_i \tag{10}$$

$$\widehat{E}_Y^{(n)}[Y|W = 0] = \frac{1}{\sum_{i=1}^n (1 - w_i)} \sum_{i=1}^n y_i (1 - w_i) \tag{11}$$

Task 3: Estimation

$\widehat{E}_Y^{(n)}[Y|W=1]$ and $\widehat{E}_Y^{(n)}[Y|W=0]$ are descriptive statements.

- ▷ These are the objects that the NYT article referred to!
- ▷ According to the article:

$$\widehat{E}_Y^{(n)}[Y|W=1] \approx \$32.60, \quad \text{and} \quad \widehat{E}_Y^{(n)}[Y|W=0] \approx \$16.50.$$

Given our estimates $\widehat{E}_Y^{(n)}[Y|W=0]$ and $\widehat{E}_Y^{(n)}[Y|W=1]$, we can construct an estimate of the counterfactual of interest:

$$\widehat{E}_U^{(n)}[\tau(U)|W=1] = \widehat{E}_Y^{(n)}[Y|W=1] - \widehat{E}_Y^{(n)}[Y|W=0]. \tag{12}$$

Using the estimates from the NYT article, $\widehat{E}_U^{(n)}[\tau(U)|W=1] \approx \$16.10$.

Task 3: Estimation

Is $\widehat{E}_U^{(n)}[\tau(U)|W=1]$ a "good" estimate of $E_U[\tau(U)|W=1]$?

Statistics provides tools that allow for evaluation of our estimates.

Two concrete questions that we will be concerned with are:

(a) Does $\widehat{E}_U^{(n)}[\tau(U)|W=1] \xrightarrow{p} E_U[\tau(U)|W=1]$ as $n \to \infty$?

(b) Does $\delta_n \widehat{E}_U^{(n)}[\tau(U)|W=1] \xrightarrow{d} P(\theta)$ as $n \to \infty$ for some scaling sequence $(\delta_i)_{i=1}^{\infty}$, where $P(\theta)$ is a known distribution parameterized by a known θ.

These are examples on convergence in probability and convergence in distribution.

- (a) is concerned with whether our estimate eventually converges to the true value as the sample size increases. (This is often considered the minimal requirement of an estimator.)
- (b) is concerned with quantifying the uncertainty of our estimate. (Important to understand how much we can "rely" on it.)

Example: Returns to higher education (Contd.)

We will develop answers to (a) and (b) in lectures 6 & 7. For now, assume that we have found a satisfactory answer to both questions so that our causal analysis is complete.

The initial causal question was:

- ▷ *What is the change in hourly wages for college graduates if they had not pursued higher education?*

Our analysis + the NYT data allows us to reply:

- ▷ "Assuming that having college degree is independent of other determinants of hourly wages (Assumption RA), hourly wages of college graduates would be approximately $16.10 lower on average if they had not pursued higher education."

That's a logically consistent causal conclusion, as our analysis showed.

- ▷ But is it useful? The causal analysis we conducted crucially relied on Assumption RA, which may not be plausible here. Without the assumption, our conclusion is not guaranteed to hold.
- ▷ This was not clear from the NYT article!

Example: Returns to higher education (Contd.)

Summary:

- Task 1: *Definition*. We used economic theory and probability theory to formulate and interpret a counterfactual that is informative for the causal question of interest.
- Task 2: *Identification*. We used probability theory to show that the counterfactual can be expressed as a known function of only observables *under* an economically interpretable assumption.
- Task 3: *Estimation*. We developed an estimate of the counterfactual of interest, assessed its statistical properties, and calculated it using real data.

Note: We needed to combined economics, mathematics (in particular, probability theory), *and* statistics to arrive at our causal conclusion!

Course Schedule

Today's discussion was a motivation for lectures 2-5.

- ▷ We will revisit the example in more depth in lectures 6 & 7.

	Date	Topic
1	Mar 28	Logistics & Motivation
2	Mar 30	Review of Probability Theory
3	Apr 4	Review of Probability Theory
4	Apr 6	Review of Statistics
5	Apr 11	Review of Statistics
6	Apr 13	Introduction to Causal Inference
7	Apr 18	Random Assignment
8	Apr 20	Simple Linear Regression
9	Apr 25	Simple Linear Regression
10	Apr 27	Simple Linear Regression
–	May 2	Midterm
11	May 4	Selection on Observables
12	May 9	Multivariate Linear Regression
13	May 11	Multivariate Linear Regression
14	May 16	Multivariate Linear Regression
15	May 18	Instrumental Variables
16	May 23	Instrumental Variables
17	May 25	Instrumental Variables
–	TBD	Final Exam

References

Duo, Q. (1993). *The formation of econometrics: A historical perspective*. Clarendon Press.

Heckman, J. J. and Vytlacil, E. J. (2007). Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation. In Heckman, J. J. and Leamer, E., editors, *Handbook of Econometrics*, volume 6, chapter 70, pages 4779–4874. Elsevier, Amsterdam.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960.

Hurwicz, L. (1950). Generalization of the concept of identification. In Koopmans, T. C., editor, *Statistical inference in dynamic economic models*, volume 6, chapter 4, pages 245–257. Wiley, New York.

Tinbergen, J. (1954). *Econometrics*. George Allen & Unwin Ltd. 2005 Reprint published by Routledge.