2 Evidence, Models and Decision Making

"Make your theories elaborate"

—R. A. Fisher (1945) commenting on how to clarify the step from association to causation.

2.1 Introduction

A ubiquitous feature of modern training in economics is the division of duties between theory and evidence, particularly in graduate programs where micro-economic theory and econometrics are taught as separate subjects with often little overlap or connection between the two. A common view among theorists, as nicely articulated by Ariel Rubinstein (1991), is that economic theory (or more specifically game theory) is a form of story telling about the world, but does not present a realistic representation of the world. In many ways this is an update on the views of Friedman (1953) who stated:

> The ultimate goal of a positive science is the development of a "theory" or, "hypothesis" that yields valid and meaningful (i.e., not truistic) predictions about phenomena not yet observed. Such a theory is, in general, a complex intermixture of two elements. In part, it is a "language" designed to promote "systematic and organized methods of reasoning." In part, it is a body of substantive hypotheses designed to abstract essential features of complex reality.[1]

The standard first year graduate textbook in micro-economics, Mas-Colell et al. (1995), does not have a single citation to empirical evidence. For the empirically minded student, graduate micro-economics is often viewed as a course one must survive, before moving on with more interesting empirical questions. Conversely, for the theorist the reverse is true.

The purpose of this chapter is to provide some basic language and concepts for the empirical study of socio-economic phenomena. This will allow me to discuss how one can move from the theory to evidence at various points in the book (though this is a book on economic theory).

1 Page 7.

One of the reasons this issue is often sidestepped in graduate training is because it is very hard! Economics, like physics, has a wide variety of models, and for a particular problem there may be a great multiplicity of choice regarding the most appropriate model. In the case of physics this is beautifully illustrated in the classic text by Feynman et al. (1963) where every page is full of both theory and evidence, with the theory chosen as a function of the phenomena to be studied.

In chapter 2 of Volume III, Feynman explicitly discusses the problem of the wave and particle views of matter, and the impossibility of ever having a completely deterministic model of the universe. The point is that even in physics there is not, nor likely to be, a single model or world view that is useful in all contexts. Even though physics is viewed as a well developed science it is important to keep in mind that modern physics does not consist of a single model, but is a collection of models, whose choice depends upon the phenomena at hand, as well as the time and space scales. One uses different models for the study of stars from the ones used for studying sub-atomic particles.

This is even more true for complex socio-economic phenomena and naturally leads to some controversy regarding the best way forward. For example, the June 2010 issue of the *Journal of Economic Literature* is devoted to the question of how best to do applied research in development economics.[2] Deaton (2010) suggests the recent emphasis upon the use of experiments has resulted in less integration of the theory with the evidence. More generally, theory is sometimes viewed with suspicion by applied economists, as we can see from Angrist and Pischke (2009)'s conclusion to their entertaining and insightful book:

> If applied econometrics were easy, theorists would do it. But it's not as hard as the dense pages of *Econometrica* might lead you to believe. Carefully applied to coherent causal questions, regression and 2SLS almost always make sense. Your standard errors won't be quite right, but they rarely are. Avoid embarrassment by being your own best skeptic, and especially, DON'T PANIC!

The advice "DON'T PANIC!" is good advice that Douglas Adams also embraces, as it appears inscribed in large friendly letters on the *Hitchhikers Guide to the Galaxy.* A serious lesson one can learn from Adams' brilliant book is the impossibility of predicting the future. The author begins his book with the imagined destruction of the earth by the

2 See the articles by Deaton (2010), Heckman (2010) and Imbens (2010).

Vogons completing an inter-galactic construction project. This example illustrates a number of points.

First, this might be fiction, but there are many real life examples of the difficulty in predicting events that lead to monumental upheavals in individual lives, such as the 2011 Tsunami in Japan, the 2008 crash in financial markets, or maybe Donald Trump becoming the President of the United States in 2017. These events were difficult to predict, and resulted in thousands of individuals and firms having to make unexpected, life changing decisions. We can certainly expect similar events in the future, though maybe not in the form of a Vogon super-highway, but in some other, equally unexpected form.

For example, there might be the discovery of cheap energy from nuclear fusion. What would happen to all the long term supply contracts between fuel supplies (coal, oil, natural gas) and utilities that have explicit incentive terms that make it costly for one or the other party to leave the relationships?[3] At the time parties entered into these contracts, they had a model of how the future would unfold, and then designed their relationship in the context of this model. A dramatic fall in energy prices would necessarily lead to both contract renegotiation and contract breach, that in turn would require courts to decide how contracts should be enforced to deal with these unforeseen contingencies.[4]

In this Chapter I briefly outline the role of models in the assessment of empirical evidence, and how this evidence can be used to modify our beliefs and expectations regarding an uncertain future. For our purposes a model has three roles. The first of these is to provide a *representation* of observed phenomena. By this, one means that a model provides a parsimonious way to encode and represent information numerically. The second is to provide information about a *population* of units - individuals, firms or countries. The goal of such a model is to capture statistical relationships between observed characteristics that are common to the population. This allows us to make inferences about the characteristics of any member of the population. Finally, models can provide a representation of causal mechanisms that explicitly address the question of how changing a choice affects observed or experienced outcomes.

3 See Joskow (1987) on the value of long term contracts.

4 The problem of foreseability is an important part of modern contract law. See Farnsworth (2004), Section 3(B).

Thus, I explicitly assume that building better models of the world helps us make better decisions. This is an unprovable assumption that is controversial. For example, in his book, *The Poverty of Historicism,* Popper (1957) explicitly argues that one cannot hope to have a scientific theory of human affairs. However, there is evidence that model building is a fundamental ingredient for decision making for any successful individual. This point was nicely illustrated by BF Skinner in the case of pigeons!

2.1.1 "Superstition" in the Pigeon

The ability to see patterns in data, even when such patterns do not really exist, is a skill that is not restricted to humans. Skinner (1948) reports an ingenious experiment where he provides food pellets to a number of pigeons. Before describing the experiment, consider first the observed outcomes. After one treatment, the birds exhibit a variety of behaviors. One repeatedly turns counter clockwise in the cage, another repeatedly thrusts its head into one of the corners of the cage, while a third develops "a 'tossing' response, as if placing its head beneath an invisible bar and lifting it repeatedly"[5]

If presented with this data, one can imagine building an area of research for understanding why different birds have different behaviors. These observed rituals might be due to the way they were brought up, the types of parents they had, or maybe their genes. Skinner convincingly shows that it is none of these. Rather, food was randomly supplied to the birds, and each bird learned to associate the behavior with what it was doing at the time the food arrived.

For example, if a bird was turning at the time food arrived, then it is natural to have the hypothesis that further turning would lead to more food. If food supply was sufficiently frequent, then further turning would be rewarded with more food, and the bird would believe that by turning in its cage it *caused* food to arrive! Skinner carefully modified the arrival rates of food, and showed that he could extinguish the behavior, or increase its frequency by manipulating the arrival rates of the food.

This experiment has a number of useful lessons. First, it illustrates that animals, humans included, are designed to build models of the

5 page 168, Skinner (1948).

world, including false models. The difference between humans and birds is that we have developed a statistics toolbox that can help distinguish between good and bad models, and particularly the difference between causal relationships and spurious correlations.

Second, and most importantly, it is very easy to build a false model. For example, there might be underlying characteristics of the pigeons that lead them to turn or toss their head. In the absence of an experiment, one might easily collect data that is correlated with many of the observed behaviors, without actually understanding that the root cause is a spurious correlation between behavior and food supply.

In human affairs there are many many examples of such behaviors and "superstitions", including astrology, palm reading, lucky tokens and so on. A particularly important example is predicting future prices in financial markets. Low and Hasanhodzic (2010) provide an interesting history of "technical analysis" in financial markets. By a technical analysis, one means identifying patterns in past financial data in order to predict future trends, and thereby identify profitable trades. Given the potential gain from such predictions, this encourages a large industry in financial advising in which some individuals claim to be able to do this better than others.[6] It is a sad fact that even though we have some excellent statistical tools, individuals continue to be attracted to "experts" who claim to be able to beat the market.[7] A good example of this is Bernard Madoff, who was convicted of a Ponzi scheme using funds from many respectable investors.[8]

One goal of economics is to help improve the rules that regulate economies, given the actual behavior of individuals. This begins by building good, evidence based models, which in turn provide a guide to decision making. These models will not be perfect. One reason that "unscientific" models and world views survive is for exactly the same reason that pigeons end up building false models - in a large complex economy there is a strictly positive probability that some trader using unscientific technical analysis beats the market several periods in a row,

6 In a large market, there will always be some individuals who seem to do well, even if they are simply lucky. See exercises 1 and 2.

7 See Hamilton (1994) for an excellent review of the literature on time series.

8 See the files on his case at http://www.justice.gov/usao/nys/madoff.html.

even though he or she has no special skill. The challenge is to build models of the world that can distinguish the lucky trader from one that has true insights into how markets work. The next three sections outlines three distinct reasons for building useful models.

2.2 Models for Representation

One of the most ubiquitous applications of a formal model is as a *parsimonious representation* of physical phenomena. For example modern photography is heavily reliant upon representing visual data in computer memory. Most casual photographers use the jpeg standard, which is a mathematical model of the photograph. When a photograph is taken, the image is divided into a grid of pixels that record the intensity of each of the three primary colors at the pixel location (or have repeated groups pixels, each one specialized to a color). This information, called a raster image, is a large vector $\alpha = \{\alpha_i\}_{i=1}^{10^7}$ where for each pixel i, $\alpha_i = (R_i, G_i, B_i)$ provides the intensity of each of the three primary colors.[9] The intensity is recorded as a binary number, usually with 16 or 24 bits (and hence 2^{16} or 2^{24} intensity levels). The original file typically contains millions of bits of information, which require large amounts of storage space. The jpeg format is a *mathematical model* that is used to reduce the image to a smaller set of parameters, while retaining quality as measured by the perception of a human viewer of the image. Essentially, the jpeg algorithm projects this information into a lower dimensional vector space, in the same way one does a linear regression. [10]

Formally, the jpeg standard produces two functions. The first takes the data and produces a lower dimensional representation: $\beta = f^{jpeg}(\alpha, r) \in I^{n_{raw}/r}$, where r denotes the degree of compression and I is the set of possible intensity levels $(0 - (2^{24} - 1))$. It relies upon how the brain interprets visual data to interpolate between pixels. In most cameras the function f^{jpeg} is programmed into the camera, so that only the data β

9 This corresponds to the case where the image is split into three separate images, a process used in more expensive cameras. Most cameras use a single sensor, which the pixels on the sensor are divided by the three colors. Conceptually the systems are very similar.

10 See www.jpeg.org for more information about ISO/IEC IS 10918-1 | ITU-T Recommendation T.81, or more simply the jpeg standard.

is recorded and sent to the computer. Given this data, a second function recreates the raster image that is displayed on the screen that is the "inverse" of the function f^{jpeg}:

$$\alpha' = g^{jpeg}(\beta) \in I^{n_{raw}/r}.$$

In general, since β is in a lower dimensional space than α, it is normally the case that $\alpha \neq \alpha'$.

This example illustrates the point that even though we could in principle store the high quality raster file, the jpeg model results in a smaller file size representing the data, though with some information loss. What makes it a good model is that one can achieve a much more parsimonious representation of the data, even though, as Pauli observes, the model is wrong in the sense it does not perfectly represent the original data.[11] Rather, the best way to evaluate the jpeg algorithm is subjective - does it do a good job representing an image.

The representation problem is complex because is can be hierarchical, with models building upon models. Figure 2.1 is based upon a jpeg image of New York City (NYC) circa 1964.[12] This image can be viewed as a very useful model. First, it is not "true" in the sense that it provides a perfect representation of NYC. Rather it provides information about the streets in the city at the time. More over, it can be used for a what if analysis. Using the map, and given a particular location, one can determine the route that will get one to the coastline as quickly as possible.

Second, it provides a natural way to think about the "external validity" of the models. That is to what extent can I use the model for new questions. We would expect that in 1965 the map would still be accurate, and useful to determine how to go to a particular location, or the name of a street in a particular location. Today, more than 50 years later, the map is still useful, but we would expect changes - some street names change and there may be new construction that changes the optimal route we might use. In other words, this map is a "false" model in that

11 See Peierls (1960), page 186, where he states: "Quite recently, a friend showed him the paper of a young physicist which he suspected was not of great value but on which he wanted Pauli's views. Pauli remarked sadly, 'It is not even wrong."'

12 Diversified Map Corporation. Street map, New York City. St. Louis: Diversified Map Corp., 1964. Retrieved from the Library of Congress, https://www.loc.gov/item/2007630435. (Accessed April 12, 2016.)

it is not perfectly accurate, and is less accurate today than at the time it was produced. In the absence of a better model/map it remains useful even today.

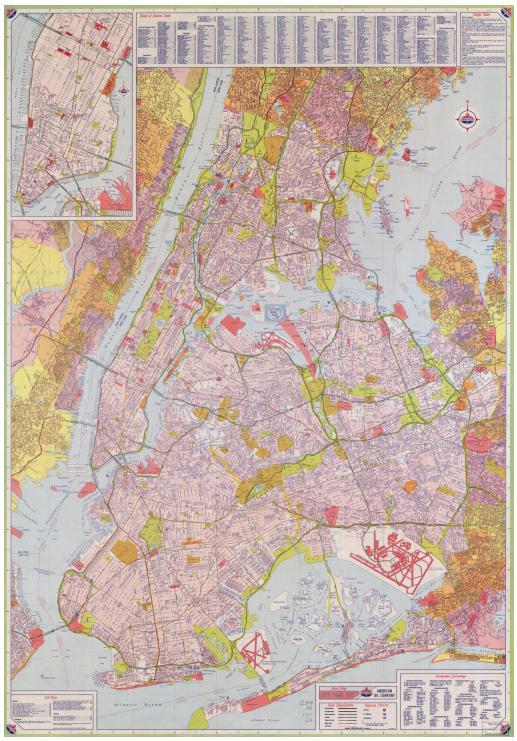

Figure 2.1
Map of New York City

We can push the map analogy further and observe that one may choose a different map/model for different questions. The map in Figure 2.2 is a 1954 subway map of New York City.[13] This is a model of the same thing, New York City, but at a different time and with a different goal - providing information on the subway system. In particular, the subway map is intended to answer a different set of questions than the street

13 Voorhies, Stephen J, and Union Dime Savings Bank. Map of the New York city subway system. New York: Union Dime Savings Bank, c, 1954. Map. Retrieved from the Library of Congress, `https://www.loc.gov/item/82690521`. (Accessed April 12, 2016.)

map. Street maps can contain subway information, but it can be hard to see the subway markers, and adding line overlays would cover some streets. Thus, subway maps can be viewed as models with a different goal from a street map - they allow individuals to make decisions on how to use the subway quickly and efficiently.

Like a street map, the external validity decreases with time. Moreover, notice that the external validity of this model is very *non-linear!* In the year a new subway line is opened, there is a large and discrete change to the model. For the parts of the city that are not served by the new line, the new model is unchanged, but it can have a large effect upon decision makers who live close to the new line.

Figure 2.2
Subway Map for New York City

In summary, the map perspective captures the main features of a good model. The most important feature is that a map/model is a decision

aid - the model's representation allows one to make better choices more quickly than in the absence of a map/model. Second, maps/models of the same phenomena are not unique. The map one uses to represent a city depends upon the question one is going to ask (where is the closest park or which subway one should take). Maps/models have external validity because the world they are describing changes slowly and smoothly enough that the representation from a period is useful for many periods into the future. Finally, the failure of a map/model can be very rapid and occur in non-linear, unpredictable ways.

All these observations are true for economic models as well. For example, we can study the US economy from the perspective of micro-economists, where the concern might be about the level of inequality, and appropriate policy responses. In contrast, macro-economists are more concerned with issues such as inflation and monetary policy. These questions are clearly linked, yet in practice one does not use a single model to study them, but builds a model for the specific problem. Hence, the models in this book should not be viewed as the end point, but as a set of potential starting points for building models to address specific issues and problems.

2.3 Population Models

The map example illustrates why models are useful for decision making. The example also illustrates a feature that is common to all models - they are not "true", but rather very useful, but imperfect decision aids. A second feature of a map, say the map of New York City, is that we do not expect it to be useful for other cities. However, there are features of New York City that we might expect to apply to other places. For example, the relationship between population density and rents. The purpose of a population model is to be able to make statements about situations we have not seen before based upon information from other, similar situations.

More precisely, a population model begins with a large number of units, denoted by $i \in I$. These units might be cities, but more often in economics they are individuals (workers or consumers), firms or countries. Suppose that at date t there are potentially three sources of information. The first is a vector providing individual specific characteristics

Table 2.1
Population Features

Unit $i \in N$	Characteristics: $x_{it} \in X$	Actions: $a_{it} \in A$	Outcomes: $y_{it} \in Y$
Person	age, sex, race, wealth, health	education, labor supply	income, health
Firm	age, size, location	location, investment	growth, profits, size
Country	health, climate, geography	taxes,services	health, climate, wealth

or properties, denoted by $x_{it} \in X \subset \Re^n$. This might be the gender of a person or the number of employees in a firm.

The purpose of a model is to understand and evaluate choices. Hence, the second ingredient is the set of possible choices that the unit might make, given by a finite set A, where the action is $a_{it} \in A$. The combination of choices and characteristics can be associated with a set of *potential outcomes,* Y, where $y_{it} \in Y$, is the realized outcome at date t for unit i. For the current discussion we remain agnostic regarding the relationship between these variables. Specifically, we discuss the statistical properties of the population that in turn allows us to describe the expected properties of a unit randomly selected from the population.

Table (2.1) provides some examples of population models. Notice that the same information can appear in several columns. For example, the current health of an individual can be an X variable because it affects labor supply. However, it can also be an outcome variable because choices, such as the level of education, can affect health.[14]

As an example, consider the 1979 National Longitudinal Survey of Youth, better known as the NLSY79. This is a nationally (United States) representative survey of 12,686 men and women born in the years 1957-64. These individuals were interviewed annually from 1979 till 1994, and they are now interviewed on a biennial basis since 1994. [15] The data includes information on a variety of characteristics, including gender, age, education, wage and hours worked in a year.

14 SeeCurrie (2009)

15 See `http://www.bls.gov/nls` for more information.

Suppose we let education, $e_i \in \{0, 1, 2, ..., 20\}$ be a choice variable, and let the outcome $y_{it} \in \Re$ denote the log wage in year t for a male worker i in year 2000. Let N be the set of male workers in the United States in the year 2000, and let $Z_i = \{e_i, y_{i,2000}\}$, $i \in N$ be a potential observation. The population information is denoted by:

$$Population = \{Z_{i,2000} | i \in N\}.$$

A number of population models are possible. For example, this set could include several million observations if one had data from every individuals in the economy. This is not practical, hence the actual data is from a small survey, that is used to make general statements regarding the whole population. The survey results can be denoted by a set $N^s = I^s \times T^s$, where $I^s \subset I$ are the individuals observed, and $T^s \subset T$ is the set of dates at which observations are taken, where T is the set of possible dates. Let $N = I \times T$ be the index set for the whole population, and let $n = \#N$ and $n^s = \#N^s$ be the size of each index set.

In this case, letting $z_{it} = \{y_{it}, x_{it}\}$denote the observations, the data set is given by:

$$Data = \{z_{it} | it \in N^s\}.$$

While the *population* of units is given by:

$$Pop = \{z_{it} | it \in N\}.$$

A central concern of statistics is to ask what we can learn about the population *Pop* using the observations in *Data*. For example, we may wish to know some of the basic features of the Population as given by *descriptive statistics,* such as the mean and variance:

$$mean(z_{it}, N) = \sum_{it \in N} z_{it}/n,$$
$$var(z_{it}, N) = \sum_{it \in N} (z_{it} - mean(Z_{it}))^2 /(n-1).$$

Notice that z_{it} is *not* being viewed as a random variable. The mean and variance are in principle well defined numbers that we can measure if we had data on all individuals. In general, access to such data is very difficult. One can estimate these numbers by drawing a random sample

from N of size $n^s < n$. Let $N^s(m)$ represent this sample, and then create the statistic:

$$mean(n^s) = \sum_{it \in N^s(m)} Z_{it}/n^s.$$

If the sample is randomly selected it follows that (see exercises)

$$E\{mean(n^s)\} = mean(n). \tag{2.1}$$

This statement makes no assumptions regarding the distribution of the underlying observations. Rather, 2.1 is the consequence of supposing that the sample $N^s(m)$ is randomly selected from the whole population.

In practice it can be difficult to obtain a random selection. For example, suppose that the data is collected by a phone interview, and that lower wage individuals are less likely to have a phone. In that case one over-samples from the high wage individuals, leading to a sample for which the mean wage is higher than mean wage in the population. When one uses the statement that the sample is "representative", one is supposing that equation 2.1 holds. There are many applications for such models. For example, the Gallup poll tries to work out how a population of individuals will vote based upon a sample for a small subset of the population.

To move beyond simple means, the statistician must add some additional structure, such as supposing that the z_{it} are realizations of a random variable with specific properties. For example, if it were the case that $z_{it} = z_{it'}$ for all $it, it' \in N$ then a single draw from the population would reveal all the other values. The more common case is to suppose that $\{z_{it}\}_{it \in N}$ come from a set of independently and identically distributed random variables, whose distribution is from a parametric class of distribution functions:

$$Z_{it} \sim f(z|\theta), \tag{2.2}$$

where $\theta \in \Re^k$ is an unknown parameter. For example, one might suppose that $Z_{it} \in \Re$ is normally distributed with unknown mean m and variance σ^2, in which case $\theta = \left(m, \sigma^2\right)$.

This assumption provides one with another way to *represent* the data as discussed above. Given the data, one can estimate the value of these parameters using maximum likelihood estimation (MLE)

$$max_\theta \Sigma_{it \in N^S} log f(z_{it}|\theta). \tag{2.3}$$

In the case of normally distributed data, if $\theta^0 = \left(m^0, \sigma^2\right)$ is the MLE solving (2.3), then the best guess of any value of z_{it} would be m^0, with standard deviation of σ. A good example is the Mincer wage equation.

The Mincer Wage Equation

The work of Jacob Mincer (1958), Theodore Schultz (1961)and Gary Becker (1962) introduced the idea that education is an investment activity that increases a person's future earnings. By education they mean the number of years of formal schooling. These ideas were very influential because they provided a way to explicitly measure the value of education via what become known as the Mincer wage equation. The basic form supposes that:

$$log(wage_i) = \beta \times e_i + \alpha + \epsilon_i,$$

where $wage_i$ and e_i are respectively the wage and education level of worker i. If ϵ_i is normally distributed then we can computed the maximum likelihood estimates of the parameters $\theta = \left\{\beta, \alpha.\sigma^2\right\}$ using ordinary least squares (OLS). The year 2000 sample of the NSLY used in Lemieux et al. (2009) has about 2200 men, with OLS estimates given by:

$$w_i = .087 \times Education_i - .43 + noise_i \tag{2.4}$$

The coefficients represent the correlation between a person's log wage and education, and a constant term. These coefficients are chosen to minimize the mean squared error:

$$MSE = \Sigma_{i \in I} noise_i^2 / 2200 = 0.43.$$

where I is the set of individuals in the sample. The relationship between wages and education is illustrated in figure 2.3. Wages clearly rise on average with education, but there is also a great deal of variation. This equation has a couple of interpretations. One interpretation is as parsimonious representation of the data, where the vector $\{.087, -0.43, 0.43\}$ provides estimates of the return to education, the intercept and the variance of the error term. Thus, one reduces a data set with 4400 parameters to one with 3.

This is of practical importance for an employer who is trying to decide on what wage to offer a worker. If the only information the employer has

is the years of education, then a reasonable offer might be the expected population wage conditional upon years of education:

$$\hat{w_i} = E\{w_i|e_i\} = \frac{\int_w wf\left(w, e_i|\theta^0\right)dw}{\int_w f\left(w, e_i|\theta^0\right)dw} = .087 \times e_i - .43.$$

It is worth emphasizing that the kinds of predictions one makes using population models are relatively "atheoretical" in the sense that they do not depend upon a specific mechanism linking education to wages. The main key ingredient in population models are the characteristics of individual unit i, e_i, which in turn allows one to make an inference regarding their log wage, w_i.

Like the jpeg algorithm, this regression can be viewed as creating a lower dimensional representation of the data in which 4400 data points are replaced by three parameters. However, there is still a great deal of unexplained variation. In this case the R^2 is .18, and hence *noise* accounts for about 82% of variation in wages, with the rest "explained" by a person's education.[16]

One can dramatically improve the quality of the fit by conditioning upon more variables, such as the age of the worker, occupation and so on. This would result in a fit with a smaller R^2, though it would not explain why workers in different occupations earn different amounts. What we have done is simply describe a pattern in the data that can be used to provide good predictions regarding the characteristics of a randomly chosen member of the population.

2.4 Causality

The Mincer wage equation, and its extensions are very useful for employers who would like to know how much they should pay an individual worker that is randomly chosen from the population. However, without additional assumptions, it cannot be used to reliably guide individuals as they decide how much education they should accumulate. For an individual worker, the decision they face is choosing between an additional year of education, versus the *counterfactual* - the wage they would get if they enter the labor market immediately. To make a rational choice they

16 Recall $R^2 = 1 - \frac{SE}{TSS}$, where $TSS = \Sigma_{i \in I} w_i^2$ is the total sum of squares.

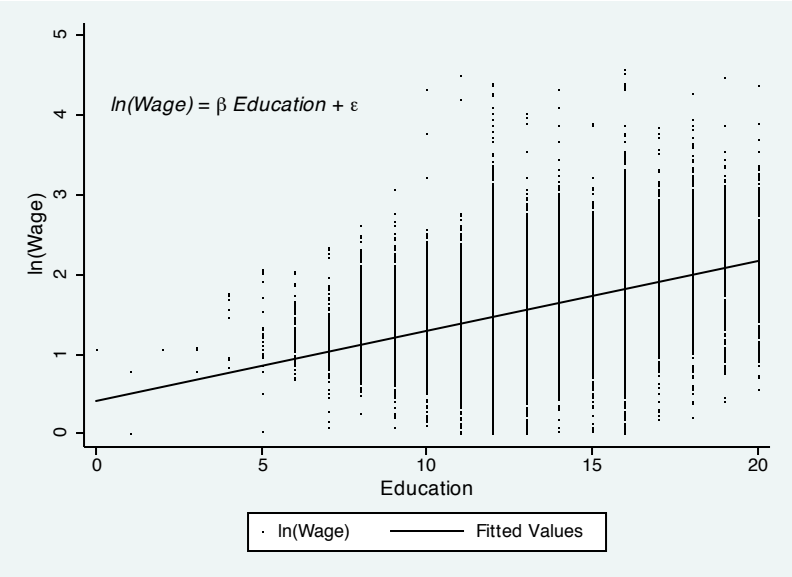

Figure 2.3
Wages versus Education

need to know the *potential outcomes* $\{u^0, u^1\}$, where u^0 is the lifetime utility from going to the labor market immediate, while u^1 is the lifetime utility from an additional year of education.

We can always frame the choice as if we are doing an experimental trial. In this case, an additional year of education can be viewed as the treatment, while going to the labor market immediately is the control. The *treatment effect* is defined by $\tau = u^1 - u^0$. The difficulty is that one makes this choice only once, and so there is no way for the individual herself to experiment with both choices (there is no way to undo the treatment with an additional year of schooling). Holland (1986) calls this the *fundamental problem of causal inference.* The only reliable way to do this is via time travel. For example, in the film *Groundhog Day* our hapless hero, Phil Connors played by Bill Murray, would keep reliving groundhog day in Punxsutawney, Pennsylvania, until he worked out how to deal with his personality and connect with the heroine of the movie, Katie, played by Andie MacDowell. In that case, each day was exactly

the same, so Phil could experiment with different strategies, until finally he was able to win Katie's heart.

As Holland (1986) points out, in order to make progress one necessarily has to make some untestable assumptions. In labor economics the most common approach is to suppose that one can group individuals with similar characteristics, some of whom get the additional year, and some do not.[17] The difficulty with using population level data to make this choice is due to *self-selection.* One reason college educated individuals earn more is because they are more able - many would earn more even in the absence of a college education (e.g. Bill Gates, the founder of Microsoft, a college dropout).

It is easy to see such an effect in the data. The NLSY data set is used by the US military to standardize the Armed Forces Qualifying Exam (AFQT) given to all new recruits. This is essentially an IQ test whose score is recorded in the NLSY. If we run the Mincer regression with the addition of the AFQT score we get:

$$w_i = .067 \times Education + .006 \times AFQT - 0.552 + \epsilon_i. \tag{2.5}$$

The addition of this single variable reduces the coefficient on education almost by half . Moreover, the standard error on AFQT is small, so that the statistical significance of AFQT is the same as Education. In this case the R^2 is now .36, about 20% larger than in the case without the AFQT score. This result is consistent with the hypothesis that more able individuals get more education, and that wages are determined in part by ability. This leaves open the question of how to measure the return to education - that is for a given person, how will his future outcome vary with the education he receives?

The solution outlined in Holland (1986) proceeds as follows.[18] We begin with a universe of individuals whose characteristics are described by a compact set $X \subset \Re^n$. For example, this might be all persons in a country who had a fever last year. Individuals may also be firms or countries, though for the current discussion we can think of them as a

17 See Angrist and Krueger (1999) for a nice discussion.

18 See Rosenbaum (2010) and Imbens and Rubin (2015) for recent reviews. The discussion here follows Deaton (2010) and MacLeod (2016).

collection of persons denoted by:

$$U = \{i \in P | x_i \in X\},$$

where x_i is the characteristic of individual i, and P denotes the universe of all possible individuals. Here I deviate slightly from Holland where the primitive is typically the set P. The reason is that the external validity of any experiment is defined by the set of persons for whom the results are valid. These individuals are typically not listed, but described by features such as race or where they live. Notice that this formulation includes the special case in which each person is a unique point in X.

For each person i, we would like to know for each choice $d_i \in \{1, 0\}$, the set of *potential outcomes:*

$$\left\{\left(x_i, u_i^1, u_i^0\right) | i \in U\right\},$$

where u_i^1, u_i^0 are the outcomes for choices 1 and 0 respectively. These are potential outcomes because the choice is made at a given date, with payoffs realized in the future, and hence for each unit we can at best observe u_i^1 or u_i^0, but not both. I maintain throughout the *stable unit treatment value assumption (STUVA)* - the decision for unit $j \neq i$ does not affect the potential outcomes for unit i. The *average treatment effect (ATE)* of choice 1 is given by:

$$\tau^{ATE} = E\left\{u_i^1 - u_i^0 | i \in U\right\}.$$

This is the parameter estimated with a randomized control trial (RCT). One procedure to measure ATE is as follows. Randomly select from U - the set of individuals that match the criteria in set X - $2n$ individuals, who are randomly assigned to group 1 - U_1 and group 0 - U_0. This generates data, $Data\,(n) = \left\{x_i, u_i^{d_i} | i \in U_A \cup U_B\right\}$, where $d_i = 1$ if $i \in U_1$ and $d_i = 0$ if $i \in U_0$. The point here is that $Data(n)$ cannot contain both potential outcomes for the same unit, but it can be used to compute an estimate of average treatment effect:

$$\hat{\tau}^{ATE}\left(Data(n)\right) = \frac{1}{n}\left\{\sum_{i \in U_1} u_i^1 - \sum_{i \in U_0} u_i^0\right\}.$$

When the assignment is random ($x_i \perp\!\!\!\perp d_i$), then we have the well known result:

Proposition 2.1 . *If units are randomly assigned to choices* 1 *and* 0*, and the stable unit treatment value assumption is satisfied, then the average treatment effect satisfies:*

$$\tau^{ATE} = E\left\{\hat{\tau}^{ATE}\left(Data(n)\right)\right\} = lim_{n\to\infty}\hat{\tau}^{ATE}\left(Data\left(n\right)\right).$$

Proof. We follow Deaton (2010). First:

$$\begin{aligned} E\left\{\hat{\tau}^{ATE}\left(Data(n)\right)\right\} &= \frac{1}{n}\left\{\sum_{i\in U_1} E\{u_i^1|d_i=1\} - \sum_{i\in U_0} E\{u_i^0|d_i=1\}\right\}. \\ &= E\left\{u_i^1|d_i=1\right\} - E\left\{u_i^0|d_i=0\right\} \\ &= lim_{n\to\infty}\hat{\tau}^{ATE}\left(Data\left(n\right)\right) \end{aligned}$$

Next observe that:

$$\begin{aligned} E\left\{\hat{\tau}^{ATE}\left(Data(n)\right)\right\} &= E\left\{u_i^1|d_i=1\right\} - E\left\{u_i^0|d_i=0\right\}, \\ &= E\left\{u_i^1|d_i=1\right\} - E\left\{u_i^0|d_i=1\right\}, \\ &= E\left\{u_i^0|d_i=1\right\} - E\left\{u_i^0|d_i=0\right\}. \end{aligned}$$

Observe that by SUTVA and random assignment, we have that the final line is zero. Random assignment also implies that the expected value of a potential outcome (observed or not) is not affected by the assignment. Hence we have:

$$\begin{aligned} lim_{n\to\infty}\hat{\tau}^{ATE}\left(Data\left(n\right)\right) &= E\left\{u_i^1|d_i=1\right\} - E\left\{u_i^0|d_i=1\right\}, \\ &= E\left\{u_i^1 - u_i^0|d_i=1\right\}, \\ &= E\left\{u_i^1 - u_i^0|i\in U\right\}, \\ &= \tau^{ATE}. \end{aligned}$$

□

Though quite simple, this result nicely illustrates the power of RCTs: under the appropriate assumptions they allow for the measurement of the average treatment effect for a *population*. There is a large literature on constructing bounds to τ^{ATE} given finite data from an RCT. Our concern here is not with the implementation details for an RCT, but with the problem of making *decisions* using observational data.

The first condition, $\tau^{ATE} = E\left\{\hat{\tau}^{ATE}\left(Data(n)\right)\right\}$, is called the *ignorability condition.* It means that regardless of the sample size, the mean

is an unbiased estimate of the treatment effect. This result has helped stimulate a literature in experimental economics (see Roth (1995), Duflo et al. (2008) and Charness and Kuhn (2011) for reviews). A typical laboratory experiment in economics draws from a pool of students, and randomly allocates them to different treatments. The results of these experiments provide information upon the average effect of the treatment for the population. Experiments are often viewed as the gold standard for the estimation of a causal effect. This is because the individuals from the target population are randomly assigned to different treatments.

These ideas have also influenced empirical work using observational data. A famous example is Angrist and Krueger (1991), who measure the effect of compulsory school laws upon outcomes. The idea is that all compulsory schooling laws are based upon requirements for students to be in school up to a certain age, such as 16 years of age. They observe that the law binds differently depending upon the quarter in which an individual is born. Individuals born in the first quarter of the year are relatively older compared to their peers in the same class. This means that they can drop out earlier, and hence on average would have less education than individuals born in other quarters. The idea is that the change in years of schooling depends only upon the quarter of birth and the law. As long as the quarter of birth is not related to labor market earnings in some unobservable way, then we have a *natural experiment.*[19] This is a feature of the environment that leads to a random assignment of individuals, when combined with the unit homogeneity assumption, allows one to estimate the casual effect of a treatment. (see Angrist and Krueger (1999) for a nice discussion of how this and related techniques have been used in labor economics).

The external validity of these studies - the extent to which the treatment effects are the same in other situations - explicitly depends upon some version of *unit homogeneity.* By this, one means that the effect is the same for all units (individuals, firms or countries) with similar characteristics. This is a claim that cannot always be tested in practice. A more serious issue is that in many cases, as both Deaton (2010) and

19 Labor market experience does increase with age. Angrist and Krueger control for this possibility by looking at workers from 40-49 years of age where one quarter of experience would have a small effect upon earnings.

Heckman (2010) observe, one may also be interested in the treatment effect for sub-populations of X.

For example, suppose that we wish to know the effect of treatment of an infection with a new antibiotic. For some people who have natural immunity, an antibiotic may have no effect, for others it may save their life, and for still others who are allergic it may cause death! In such case one may be interested in the conditional average treatment effect:

$$\tau(x) = E\left\{u_i^1 - u_i^0 | i \in U, x_i = x\right\}.$$

Given the large variability in human populations it is too expensive/impossible to run RCTs for all values of x (see MacLeod (2016) for a discussion). Hence, one needs some way to generalize from a small number of trials to the full population. The role of theory is precisely to answer such questions. A theoretical model is intended to capture the structure of existing data, and to help us predict counterfactual consequences of choice in new situations. Hence, as Deaton (2010) observes, one needs to complement RCTs with an investigation of the mechanisms leading to treatment effects. Rosenbaum (1984) made this point much earlier, and recalls one of the quotes from distinguished statistician, RA Fisher: "Make your theories elaborate". The point is that theory helps structure the experiment, and allows one to learn what are the valid inferences one can make from an experiment.

2.5 Summary

This chapter discusses the role that formal models play a central role in organizing and summarizing large amounts of information. In the case of economic models, the first step in building a useful model is to identify the units to which the model will apply. This could be an individual, a family, a firm or even a country. The extent to which the model has empirical content depends upon the extent to which observations of a number of units provide information regarding the characteristics or behavior of similar units in the future. These predictions take two possible forms.

The first are population models. These are models for which it is assumed that a group or population of units have similar properties. For example, observing the wage for a randomly selected group of workers

provides information on the expected wage for the population at large. These models can be used to address a number of useful questions; however, they are not causal in the sense of providing substantial information on how a decision can affect an outcome.

This is the domain of the second class of empirical models that rely upon counterfactual analysis. Specifically, when choosing between different courses of action, one would like to know how the outcome varies with choice. As Holland (1986) has observed, the *fundamental problem of causal analysis* is the impossibility of simultaneously observing the outcome for each choice. Hence, every empirical study has an explicit or implicit identification strategy that allows one to infer the consequences of different treatments. Typical solutions include supposing that one has *time homogeneity* - the same unit can be used to compare the outcomes of decisions at two points in time, or *unit homogeneity* - the effect of the decision is the same for different units with similar observable characteristics.

The goal of this book is to understand incentive contracts between entities. The analysis very much depends upon the hypothesis that parties can anticipate the consequence of a contract for behavior, which in turn depends upon some form of unit invariance that we need to make precise. Since the environment within which entities act varies with time, then even an experiment with a credible identification strategy may not necessarily apply to a new environment. The purpose of a good theory is to allow one to extend our ability to make decisions to situations for which we have little previous experience. As we shall see, incentive contracts are very complex, and we hope to show that the theory can be, with caution, fruitfully applied to a number of important issues.

2.6 Exercises.

1. Consider the market for financial advice, and the problem of picking stocks that will go up in the next month. Suppose that there are N stocks that go up with probability p each month. Now suppose that there are M advisors who pick stocks at random. What is the probability that at least one advisor is correct in a month? What is the probability that at least one advisor has picked correctly more than t times, in 12 months? Have a look at actual markets and think

about the likelihood that there are advisors who consistently perform better than the market.

2. Now suppose there is an advisor who is better than average, that is all the stocks she picks have a probability $q > 1/2$ of going up in price. What is the probability that she is the top performer over 12 months compared to individuals who pick randomly?
3. Prove equation 2.1 when there is sampling with no replacement and sampling with replacement. In the later case after choosing a worker, that same worker is eligible for being chosen again. In each case compute the variance of the statistic $m\left(Z_i, I^S\right)$ and compute the variance as the sample size m^S approaches m. Given this, which procedure would you choose for sampling from the population?

References

Angrist, J. and A. Krueger (1991, NOV). Does compulsory school attendance affect schooling and earnings. *Quarterly Journal Of Economics 106*(4), 979–1014.

Angrist, J. and J.-S. Pischke (2009). *Mainly Harmless Econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.

Angrist, J. D. and A. B. Krueger (1999). Empirical strategies in labor economics. In O. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics*, pp. 1278–1357. Elsevier Science B.V.

Becker, G. (1962, October). Investment in human capital: A theoretical analysis. *Journal of Political Economy 70*, 9–49.

Charness, G. and P. Kuhn (2011). Lab labor: What can labor economists learn from the lab? In O. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics, Volume 4*, Volume 4. Elsevier.

Currie, J. (2009, March). Healthy, wealthy, and wise: Socioeconomic status, poor health in childhood, and human capital development. *Journal of Economic Literature 47*(1), 87–122.

Deaton, A. (2010, June). Instruments, randomization, and learning about development. *Journal of Economic Literature 48*(2), 424–455.

Duflo, E., R. Glennerster, and M. Kremer (2008). Using randomization in development economics research: A toolkit. In T. P. Schultz and J. A. Strauss (Eds.), *Handbook of Development Economics*, Volume 4, Chapter 61, pp. 3895–3962. Amsterdam, The Netherlands: Elsevier B.V.

Farnsworth, E. A. (2004). *Contracts, 4th edition*. New York: Aspen Publishers.

Feynman, R. P., R. B. Leighton, and M. Sands (1963). *The Fynman Lectures on Physics*. Reading, MA: Addison-Wesley Publishing Company.

Friedman, M. (1953). The methodology of positive economics. In *Essays in Positive Economics*. Chicago, IL: University of Chicago Press.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press.

Heckman, J. J. (2010). Building bridges between structural and program evaluation approaches to evaluating policy. *Journal of Economic Literature 48*(2), 356–98.

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

Imbens, G. (2010, June). Better late than nothing: Some comments on deaton (2009) and heckman and urzua (2009). *Journal of Economic Literature 48*(2), 399–423.

Imbens, G. W. and D. B. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Pr.

Joskow, P. (1987, March). Contract duration and relation-specific investments: Empirical evidence from coal markets. *American Economic Review 77*, 168–185.

Lemieux, T., W. B. MacLeod, and D. Parent (2009, February). Performance pay and wage inequality. *Quarterly Journal of Economics 124*(1), 1–49.

Low, A. W. and J. Hasanhodzic (2010). *The Evolution of Technical Analysis*. Hoboken, NJ: John Wiley & Sons.

MacLeod, W. B. (2016, March). Viewpoint: The human capital approach to inference. Working Paper 22123, National Bureau of Economic Research.

Mas-Colell, A., M. D. Whinston, and J. R. Green (1995). *Microeconomic Theory*. New York, NY: Oxford University Press.

Mincer, J. (1958). Investment in human-capital and personal income-distribution. *Journal of Political Economy 66*(4), 281–302.

Peierls, R. E. (1960). Wolfgang ernst pauli. 1900-1958. *Biographical Memoirs of Fellows of the Royal Society 5*, 175–192.

Popper, K. R. (2002 (first published 1957), September). *The poverty of historicism.* Routledge.

Rosenbaum, P. R. (1984). From association to causation in observational studies: The role of tests of strongly ignorable treatment assignment. *Journal of the American Statistical Association 79*(385), pp.41–48.

Rosenbaum, P. R. (2010). *Design of Observational Studies.* Springer Series in Statistics.

Roth, A. E. (1995). Introduction to experimental economics. In J. H. Kagel and A. E. Roth (Eds.), *The Handbook of Experimental Economics*, pp. 3–109. Princeton, NJ, U.S.A.: Princeton University Press.

Rubinstein, A. (1991, July). Comments on the interpretation of game theory. *Econometrica 59*(4), 909–924. 00129682 Econometric Society.

Schultz, T. W. (1961, March). Investment in human capital. *American Economic Review*, 1–17.

Skinner, B. F. (1948). 'superstition' in the pigeon. *Journal of Experimental Psychology 38*, 168–172.

.