Economics 361
Introduction

Jun Ishii *
Department of Economics
Amherst College

Fall 2023

Overview

> "There are known knowns. There are things we know we know. We also know there are **known unknowns**. That is to say, we know there are some things we do not know. But there are also **unknown unknowns**, the ones we don't know we don't know."

Donald Rumsfeld, February 12th 2002

What is **uncertainty**? Are there events that are intrinsically uncertain? Or does uncertainty simply reflect our ignorance, our inability to observe all of the relevant state variables underlying the event? In short, as Einstein famously grappled, "does God play dice?"

Perhaps more appropriately for a **social** science like economics, is there a meaningful distinction between these two views?

Consider the following **blackbox** which we will dub an **experiment**: the **experiment** takes inputs which we will call **conditioning variables** and spits out output which we will call **outcomes**. A collection of these **outcomes** from repeated application of the **experiment** – whether with the same or differing values of the conditioning variables – we will dub **data** (and later relate to the concept of a **sample**).

If knowledge of the conditioning variables serving as inputs for the experiment were sufficient to predict accurately the outcomes of the experiment, the experiment would be considered **deterministic**. But our analysis of the experiment may still be uncertain as, most notably, we may have incomplete knowledge of the key inputs. In such a scenario, the best our analysis could do would be to offer predictions that are conditional on the values of the unknown inputs ("known unknowns"). Thus, to the analyst, this deterministic experiment is indistinguishable from an intrinsically uncertain experiment involving only the known inputs. Uncertainty surrounding the unknown inputs

*Office: converse Hall 315 Phone: (413) 542-2901 E-mail: jishii@amherst.edu

introduces uncertainty into the experiment.

Consider the following deterministic experiment with three conditioning variables $\{X_1, X_2, X_3\}$, each of which can take some value from the real line. The experiment simply spits out the sum of the three conditioning variables as its outcome. Thus, the outcome of this experiment, which we will dub Y, can be expressed deterministically by the following equation:

$$Y = \underbrace{X_1 + X_2 + X_3}_{Y(X_1,X_2,X_3)} \tag{1}$$

Therefore, knowledge of $\{X_1, X_2, X_3\}$ along with the process by which the experiment generates its outcome (i.e. the above equation) allows the analyst to "predict" the value of the outcome associated with the given set of $\{X_1, X_2, X_3\}$ values.

Notationally, if we dub a specific set of $\{X_1, X_2, X_3\}$ values as $\{X_1 = x_1, X_2 = x_2, X_3 = x_3\}$ or, even more succinctly, $\{x_1, x_2, x_3\}$, we can express the associated "predicted" outcome as

$$y = x_1 + x_2 + x_3 \tag{2}$$

But suppose the analyst knew the process and the values of two but not all three conditioning variables associated with the above experiment. Without loss of generality, let us say the analyst knows $\{X_1 = x_1, X_2 = x_3\}$ but not the value of X_3. In which case, we could arrive at some expression that lies in-between the above two

$$Y(X_1, X_2, X_3)|_{X_1=x_1,X_2=x_2} = x_1 + x_2 + X_3$$

where the lefthand side term " $|_{X_1=x_1,X_2=x_2}$ " serves as shorthand for "conditional on $\{X_1 = x_1, X_2 = x_2\}$. Our uncertainty concerning the value of X_3 makes what was a deterministic relationship an uncertain ("stochastic") one.

What this hybrid expression provides is a statement of what the outcome prediction would be for a given X_3 value, an algebraic expression in terms of the unknown variable X_3. The values of X_1 and X_2 are known and, thus, assumed fixed. You are now conditioning your outcome prediction on the known $\{X_1, X_2\}$ values but not the unknown X_3 value.

To proceed further, one must impose/assume/know more about this unknown X_3. Suppose the analyst were able to surmise the range of values that X_3 could take. This information, combined with the information already provided above, would enable the analyst to surmise the range of values that the outcome Y could take given $\{X_1 = x_1, X_2 = x_2\}$. Suppose further that the analyst knew that some allowable values of X_3 were more "likely" than other values. This would enable the analyst to conclude some allowable values of Y being more "likely" than others (again, conditioning on $\{X_1 = x_1, X_2 = x_2\}$).

For the above scenario, one squarely within the "known unknowns" category, our "statistical" statements about outcomes (Y) derive from our "statistical" statements about the unknown conditioning variable(s) (X_3). This line of thought lies at the heart of much of modern statistics.

But what if the analyst knew what the set of relevant conditioning variables were for some experiment but not the process by which those conditioning variables effected the outcome. In terms of the above example, the analyst knew that the relevant conditioning variables were $\{X_1, X_2, X_3\}$ but not how those variables effected Y. If the analyst believed the process to be deterministic, by which we mean that, for any given allowable set of values for the conditioning variables $\{X_1 = x_1, X_2 = x_2, X_3 = x_3\}$, the experiment results in one and only possible outcome y, the analyst might be able to infer what this process is through data analysis.

Suppose the analyst observed N repetitions of the experiment, not all of which involved the same set of $\{x_1, x_2, x_3\}$ values. If the analyst believes the deterministic process can be expressed by some proper function $f(\cdot)$

$$
\begin{aligned}
Y &= f(X_1, X_2, X_3) \\
y &= f(x_1, x_2, x_3)
\end{aligned}
$$

Without further information about $f(\cdot)$, the analyst is stuck. But consider two further scenarios:

- The analyst knows $f(\cdot)$ but only up to some set of (unknown) parameters
- The analyst believes $f(\cdot)$ could be approximated by some known function, again up to some set of (unknown) parameters

Aside: what is the difference between a **variable** and a **parameter**? For our purposes here, a variable can (but not necessarily) *vary* in value across repetitions of the experiment but a parameter is *fixed* in value across all repetitions of the experiment.

Let us consider a concrete version of the first scenario involving our earlier example. Suppose the analyst knows Y is derived from the following linear relationship with $\{X_1, X_2, X_3\}$ involving parameters $\{\beta_0, \beta_1, \beta_2, \beta_3\}$

$$
Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3
$$

Leveraging the earlier referenced data involving N repetitions, the analyst has N equations by which they can try to infer the values of the unknown parameters $\{\beta_0, \beta_1, \beta_2, \beta_3\}$. If the N equations include at least 4 linearly independent equations, the analyst should be able to derive precisely the previously unknown 4 parameter values. If not, this suggests that the above is an imperfect expression of the process by which $\{X_1, X_2, X_3\}$ jointly realize the value of Y.

But consider the scenario where the above linear relationship (involving the unknown $\{\beta_0, \beta_1, \beta_2, \beta_3\}$) holds but that the analyst only observes $\{X_1, X_2\}$; the value of X_3 is unknown. In this scenario, not only are the four parameter values unknown but so is the value of X_3 in any given repetition. As the value of X_3 may vary across repetitions, it is now impossible to infer the values of the unknown parameters precisely as the unknown X_3 effectively makes the number of unknowns $N + 4$ associated with a data involving N repetitions; there will always be more unknowns than linearly independent equations. Again, uncertainty surrounding X_3 effectively makes the process determining Y uncertain, even with data. Statistical statements about Y again necessarily derive from statistical statements about X_3.

Now consider the scenario where the analyst does not know the "true" function $f(\cdot)$ underlying the process by which Y is realized from $\{X_1, X_2, X_3\}$. In this scenario, we could still leverage some of what we had discussed earlier if we are able (or simply willing) to arrive at some sufficient approximation for $f(\cdot)$: $\hat{f}(\cdot) \approx f(\cdot)$.

If we thought $f(\cdot)$ a sufficiently differentiable function in $\{X_1, X_2, X_3\}$, we might – in the spirit of Taylor's Theorem – use some polynomial of $\{X_1, X_2, X_3\}$ as our approximation. For a first order approximation

$$Y(X_1, X_2, X_3) \quad = \quad \underbrace{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3}_{\hat{f}(X_1, X_2, X_3)} \; + \; \epsilon$$

where $\epsilon \equiv Y(X_1, X_2, X_3) - \hat{f}(X_1, X_2, X_3)$ represents the approximation error associated with $\hat{f}(\cdot)$

The analyst might then use data to arrive at values of $\{\beta_0, \beta_1, \beta_2, \beta_3\}$ that, by some metric, "minimizes" the error between Y and $\hat{f}(\cdot)$.

Were some conditioning variable, say X_3, unobserved, the unobserved variable would get lumped with the approximation error

$$Y(X_1, X_2, X_3) \quad = \quad \beta_0 + \beta_1 X_1 + \beta_2 X_2 \; + \; (\; \beta_3 X_3 + \epsilon \;)$$

This further complicates the challenge faced by the analyst as the interpretation of $\{\beta_0, \beta_1, \beta_2\}$ becomes less clear without further statements concerning X_3 and/or ϵ.

When appropriately informative statements about the unknowns, whether X_3 or ϵ, cannot be made, the analyst encroaches into the world of "unknown unknowns." Much of modern classical statistics, developed during the latter half of the 20th century, deals with "known unknowns." When faced with "unknown unknowns," the traditional short-run response has been to act **as if** the scenario were one of "known unknowns" and the more long-run response to seek to reduce the set of truly "unknown unknowns."

Statistical Inference

> An astronomer, a physicist, and a mathematician were on vacation in Scotland. From a train window, they saw a black sheep in the middle of a field. "How interesting," observed the astronomer, "all Scottish sheep are black." To which the physicist replied "No, no! Some Scottish sheep are black!" The mathematician gazed heavenward, then intoned, "In Scotland, there exists at least one field, containing at least one sheep, at least one side of which is black."
>
> A version of a joke that has been told and re-told over many decades

Inference is the act of deriving conclusions based on known particulars. When the known particulars support but do not necessarily entail the conclusions, inference follows **inductive reasoning**.

Statistical inference is inference on the **population** based on the **sample**. Heuristically-speaking, sample refers to the observed data and **population** the underlying data generating process (**DGP**). As such, it is a form of inductive reasoning.

There are three main types of conclusions about the DGP one derives from the data:

- Measurement of the factors/parameters underlying the DGP (**Estimation**)
- "Confirmation" or "rejection" of some conjecture on the DGP (**Hypothesis Testing**)
- Statement about some future realization of the DGP (**Prediction**)

Econometrics is the practice of statistical inference on economic data.

Econometric Theory, like Statistical Theory, is a science based on deductive reasoning. But the application of econometric (statistical) theory – econometrics (statistics) – is based on inductive reasoning and is more a formalized practice/art than a "pure science."

An Econ 111 Example

One of the most famous example of deductive reasoning is Aristotle's

> All men are mortal
> Socrates is a man
> Socrates is mortal

There are also examples of deductive reasoning in your introductory economics course. One of the most famous ends with the punch-line, "P = MC"

- Premise: competitive market, identical price-taking profit maximizing firms
- Conclusion: market equilibrium price is equal to marginal cost

The conclusion, "P = MC," is correct if the premise is correct. Let us consider some related pieces of inductive reasoning:

- **Point Estimation**: We observe a market price of P and conclude that each firm's marginal cost must be equal to P
- **Interval Estimation**: We observe a market price of P and conclude that each firm's marginal cost lies in some neighborhood around P, e.g. $P \in \{0.8P\ ,\ 1.2P\}$
- **Hypothesis-Testing**: We observe a market price of P and a marginal cost of $C \neq P$ and conclude that the market is not competitive
- **Prediction**: We know that marginal cost is rising in the near future and conclude that market price will also be rising in the near future

In each of the above, the given known particular is not sufficient to ensure the validity of the conclusion; we also need the premise given in the above deductive reasoning example. e.g. Given a competitive market, identical price-taking profit-maximizing firms, and a market price of P, we can conclude that each firm's marginal cost equals P.

A Messier Econ 111 Example

Suppose you are given yesterday's market price and quantity from 50 different widget markets: $\{P_i, Q_i\}_{i=1}^{50}$. From this **sample**, what can you conclude about the underlying market forces?

- **Estimation**: estimate the price elasticity of demand for widgets?
- **Hypothesis Testing**: test whether demand for widgets are price inelastic?
- **Prediction**: predict the impact of a proposed sales tax on the widget market?

First Take: Jump from descriptive statistics; calculate the sample correlation between P_i and Q_i. Make inference from the calculated sample correlation.

But what does the sample correlation tell us, exactly? It summarizes aspects of the data but there is no clear link between this summary and the inferences we want to make. A sample correlation of -0.1 tells us what about the price elasticity of demand?

Second Take: Fit a line (curve) through the (P_i, Q_i) scatter plot. Make inference based on the fitted line (curve).

How do we fit this line/curve? Carl Friedrich Gauss proposed the method of "least squares" – choose the line/curve that minimizes the sum of the squared difference between observed and fitted values. This approach has distinct advantages, explored later in the course. But this approach works to the extent that the fitted line (curve) appropriately represents the desired underlying market force (e.g. supply/demand curve). Additionally, a naïve curve fitting through $\{P_i, Q_i\}_{i=1}^{50}$ ignores possible differences (heterogeneity) in demand/supply forces across the 50 markets.

Third Take: Collect further data that characterizes how the sampled markets differ. "Condition" on these characteristics while fitting the line (curve).

This is essentially what multivariate regression analysis does, using Gauss' least squares concept.

Why econometricians earn more than statisticians ... and it's not just because they know how to bargain ...

What does economics tell us about the relationship between market price and quantity? If the market is in equilibrium, the market price and quantity pair lies at the intersection of the supply *and* demand curves. So a fitted line through market price and quantity pairs represents both market supply and demand forces. Economics provides information about the data generating process.

Fourth Take: Simultaneous equations model (which we visit later in the course)