1
Introduction

We believe progress in the study of accounting (perhaps any scientific endeavor) is characterized by attention to theory, data, and model specification. Understanding the role of accounting in the world typically revolves around questions of causal effects. That is, holding other things equal what is the impact on outcome (welfare) of some accounting choice. The ceteris paribus conditions are often awkward because of simultaneity or endogeneity. In these pages we attempt to survey some strategies for addressing these challenging questions and share our experiences. These shared experiences typically take the form of identifying the theory through stylized accounting examples, and exploring the implications of varieties of available data (to the analyst or social scientist). Theory development is crucial for careful identification of the focal problem. Progress can be seriously compromised when the problem is not carefully defined. Once the problem is carefully defined, identifying the appropriate data is more straightforward but, of course, data collection often remains elusive.[1] Recognizing information available to the economic agents as well as limitations of data available to the analyst is of paramount importance. While our econometric tool kit continues to grow richer, frequently there is no substitute for finding data better suited to the problem at hand. The combination of theory (problem identification) and data leads to model specification. Model specification and testing frequently lead us to revisit theory development and data collection. This three-legged, iterative strategy for "creating order from chaos" proceeds without end.

[1] We define empiricists. as individuals who have special talents in identification and collection of task-appropriate data. A skill we regard as frequently undervalued and, alas, one which we do not possess (or at least, have not developed).

1.1 Problematic illustration

The following composite illustration discusses some of our concerns when we fail to faithfully apply these principles.[2] It is common for analysts (social scientists) to deal with two (or more) alternative first order considerations (theories or framings) of the setting at hand. One theory is seemingly more readily manageable as it proceeds with a more partial equilibrium view and accordingly suppresses considerations that may otherwise enter as first order influences. The second view is more sweeping, more of a general equilibrium perspective of the setting at hand.

Different perspectives may call for different data (regressands and/or regressors in a conditional analysis). Yet, frequently in the extant literature some of the data employed reflect one perspective, some a second perspective, some both perspectives, and perhaps, some data reflect an alternate, unspoken theory. Is this cause for concern?

Consider asset valuation in public information versus private information settings. A *CAPM* (public information) equilibrium (Sharpe [1964], Lintner [1965], and Mossin [1966]; also see Lambert, Leuz, and Verrecchia [2007]) calls for the aggregation of risky assets into an efficient market portfolio and the market portfolio is a fundamental right-hand side variable. However, in a world where private information is a first order consideration, there exists no such simple aggregation of assets to form an efficient (market) portfolio (Admati [1985]). Hence, while diversification remains a concern for the agents in the economy it is less clear what role any market index plays in the analysis.[3]

Empirical model building (specification and diagnostic checking) seems vastly different in the two worlds. In the simpler *CAPM* world it is perhaps sensible to consider the market index as exogenous. However, its measurement is of critical importance (Roll [1977]).[4] Measures of the market index are almost surely inadequate and produce an errors-in-variables (in other words, correlated omitted

---

[2] The example is a composite critique of the current literature. Some will take offense at these criticisms even though no individual studies are referenced. The intent is not to place blame or dwell on the negative but rather to move forward (hopefully, by inventing new mistakes rather than repeating the old ones). Our work (research) is forever incomplete.

[3] Another simple example involving care in data selection comes from cost of capital analysis where, say, the focus is on cost of debt capital. Many economic analyses involve the influence of various (often endogenous) factors on the marginal cost of debt capital. Nevertheless, the analysts employ a historical weighted average of a firm's debt cost (some variant of reported interest scaled by reported outstanding debt). What does this tell us about influences on the firm's cost of raising debt capital?

[4] Arbitrage pricing (Ross [1976]) is a promising complete information alternative that potentially avoids this trap. However, identification of risk factors remains elusive.

variable) problem.[5] When experimental variables are added,[6] care is required as they may pick up measurement error in the market index rather than the effect being studied. In addition, it may be unwise to treat the factors of interest exogenously. Whether endogeneity arises as a first order consideration or not in this seemingly simpler setting has become much more challenging than perhaps was initially suspected.

In our alternate private information world, inclusion of a market index may serve as a (weak) proxy for some other fundamental factor or factors. Further, these missing fundamental variables may be inherently endogenous. Of course, the diagnostic checks we choose to employ depend on our perception of the setting (theory or framing) and appropriate data.

Our point is econometric analysis associated with either perspective calls for a careful matching of theory, data, and model specification. Diagnostic checking follows first order considerations outlined by our theoretical perspective and data choices. We hope evidence from such analyses provides a foundation for discriminating between theories or perspectives as a better first order approximation. In any case, we cannot investigate every possible source of misspecification but rather we must focus our attention on problematic issues to which our perspective (theory) guides us. While challenging, the iterative, three-legged model building strategy is a cornerstone of scientific inquiry.

In writing these pages (including the above discussion), we found ourselves to be significantly influenced by Jaynes' [2003] discussion of probability theory as the logic of science. Next, we briefly outline some of the principles he describes.

---

[5] Is the lack of a significant relation between individual stocks, or even portfolios of stocks, with the market index a result of greater information asymmetry (has there been a marked jump in the exploitation of privately informed-opportunism? – Enron, Worldcom, etc.), or the exclusion of more assets in the index (think of the financial engineering explosion) over the past twenty years?

[6] The quality of accounting information and how it affects some response variable (say, firm value) is often the subject of inquiry. Data is an enormous challenge here. We know from Blackwell [1953] (see also Blackwell and Girshick [1954], Marschak and Miyasawa [1968], and Demski [1973]), information systems, in general, are not comparable as fineness is the only generally consistent ranking metric and it is incomplete. This means that we have to pay attention to the context and are only able to make contextual comparisons of information systems. As accounting is not a myopic supplier of information, informational complementarities abound. What is meant by accounting quality surely cannot be effectively captured by vague proxies for relevance, reliability, precision, etc. that ignore other information and belie Blackwell comparability. Further, suppose we are able to surmount these issues, what is learned in say the valuation context may be of no consequence in a stewardship context (surely a concern in accounting). Demski [1994,2008] and Christensen and Demski [2003] provide numerous examples illustrating this point. Are we forgetting the idea of statistical sufficiency? A statistic is not designed to be sufficient for the data in the address of all questions but for a specific question (often a particular moment). Moving these discussions forward demands more creativity in identifying and measuring the data.

1.2 Jaynes' desiderata for scientific reasoning

Jaynes' discussion of probability as logic (the logic of science) suggests the following desiderata regarding the assessment of plausible propositions:

1. Degrees of plausibility are represented by real numbers;
2. Reasoning conveys a qualitative correspondence with common sense;
3. Reasoning is logically consistent.

Jaynes' [2003, p. 86] goes on to argue the fundamental principle of probabilistic inference is

> To form a judgment about the likely truth or falsity of any proposition A, the correct procedure is to calculate the probability that A is true
>
> $$\Pr\left(A \mid E_1, E_2, \ldots\right)$$
>
> conditional on all the evidence at hand.

Again, care in problem or proposition definition is fundamental to scientific inquiry.

In our survey of econometric challenges associated with analysis of accounting choice, we attempt to follow these guiding principles. However, the preponderance of extant econometric work on endogeneity is classical, our synthesis reflects this, and, as Jaynes points out, classical methods sometimes fail to consider all evidence. Therefore, where classical approaches may be problematic, we revisit the issue with a "more complete" Bayesian analysis. The final chapter synthesizes (albeit incompletely) Jaynes' thesis on probability as logic and especially informed, maximum entropy priors. Meanwhile, we offer a simple but provocative example of probability as logic.

1.2.1 Probability as logic illustration[7]

Suppose we only know a variable, call it X_1, has support from $(-1, 1)$ and a second variable, X_2, has support from $(-2, 2)$. Then, we receive an aggregate report — their sum, $Y = X_1 + X_2$, equals $\frac{1}{2}$. What do we know about X_1 and X_2? Jayne's maximum entropy principle (*MEP*) suggests we assign probabilities based on what we know but only what we know. Consider X_1 alone. Since we only know support, consistent probability assignment leads to the uniform density

$$f\left(X_1 : \{-1 < X_1 < 1\}\right) = \frac{1}{2}$$

Similarly, for X_2 we have

$$f\left(X_2 : -2 < X_2 < 2\right) = \frac{1}{4}$$

[7] This example was developed from conversations with Anil Arya and Brian Mittendorf.

Now, considered jointly we have[8]

$$f\left(X_1, X_2 : \{-1 < X_1 < 1, -2 < X_2 < 2\}\right) = \frac{1}{8}$$

What is learned from the aggregate report $y = \frac{1}{2}$? Bayesian updating based on the evidence suggests

$$f\left(X_1 \mid y = \frac{1}{2}\right) = \frac{f\left(X_1, y = \frac{1}{2}\right)}{f\left(y = \frac{1}{2}\right)}$$

and

$$f\left(X_2 \mid y = \frac{1}{2}\right) = \frac{f\left(X_2, y = \frac{1}{2}\right)}{f\left(y = \frac{1}{2}\right)}$$

Hence, updating follows from probability assignment of $f\left(X_1, Y\right)$, $f\left(X_2, Y\right)$, and $f\left(Y\right)$. Since we have $f\left(X_1, X_2\right)$ and $Y = X_1 + X_2$ plus knowledge of any two of (Y, X_1, X_2) supplies the third, we know

$$f\left(X_1, Y : \left\{ \begin{array}{c} \{-3 < Y < -1, -1 < X_1 < Y + 2\} \\ \{-1 < Y < 1, -1 < X_1 < 1\} \\ \{1 < Y < 3, Y - 2 < X_1 < 1\} \end{array} \right\}\right) = \frac{1}{8}$$

and

$$f\left(X_2, Y : \left\{ \begin{array}{c} \{-3 < Y < -1, -2 < X_2 < Y + 1\} \\ \{-1 < Y < 1, -1 < X_2 < 1\} \\ \{1 < Y < 3, Y - 1 < X_2 < 2\} \end{array} \right\}\right) = \frac{1}{8}$$

Further,

$$\begin{aligned} f\left(Y\right) &= \int f\left(X_1, Y\right) dX_1 \\ &= \int f\left(X_2, Y\right) dX_2 \end{aligned}$$

Hence, integrating out X_1 or X_2 yields

$$\int_{-1}^{Y+2} f\left(X_1, Y\right) dX_1 = \int_{-2}^{Y+1} f\left(X_2, Y\right) dX_2 \quad \text{for } -3 < Y < -1$$

$$\int_{-1}^{1} f\left(X_1, Y\right) dX_1 = \int_{-1}^{1} f\left(X_2, Y\right) dX_2 \quad \text{for } -1 < Y < 1$$

and

$$\int_{Y-2}^{1} f\left(X_1, Y\right) dX_1 = \int_{Y-1}^{2} f\left(X_2, Y\right) dX_2 \quad \text{for } 1 < Y < 3$$

[8] *MEP* treats X_1 and X_2 as independent random variables as we have no knowledge regarding their relationship.

Collectively, we have[9]

$$\begin{aligned} f\left(Y : \{-3 < Y < -1\}\right) &= \frac{3+Y}{8} \\ f\left(Y : \{-1 < Y < 1\}\right) &= \frac{1}{4} \\ f\left(Y : \{1 < Y < 3\}\right) &= \frac{3-Y}{8} \end{aligned}$$

Now, conditional probability assignment given $y = \frac{1}{2}$ is

$$f\left(X_1 : \{-1 < X_1 < 1\} \mid y = \frac{1}{2}\right) = \frac{\frac{1}{8}}{\frac{1}{4}} = \frac{1}{2}$$

and

$$f\left(X_2 : \{Y - 1 < X_2 < Y + 1\} \mid y = \frac{1}{2}\right) = \frac{\frac{1}{8}}{\frac{1}{4}}$$

or

$$f\left(X_2 : \left\{-\frac{1}{2} < X_2 < \frac{3}{2}\right\} \mid y\right) = \frac{1}{2}$$

Hence, the aggregate report tells us nothing about X_1 (our unconditional beliefs are unaltered) but a good deal about X_2 (support is cut in half). For instance, updated beliefs conditional on the aggregate report imply $E\left[X_1 \mid y = \frac{1}{2}\right] = 0$ and $E\left[X_2 \mid y = \frac{1}{2}\right] = \frac{1}{2}$. This is logically consistent as $E\left[X_1 + X_2 \mid y = \frac{1}{2}\right] = E\left[Y \mid y = \frac{1}{2}\right]$ must be equal to $\frac{1}{2}$.

On the other hand, if the aggregate report is $y = 2$, then revised beliefs are

$$f\left(X_1 : \{Y - 2 < X_1 < 1\} \mid y = 2\right) = \frac{\frac{1}{8}}{\frac{3-Y}{8}} = \frac{1}{3-2}$$

or

$$f\left(X_1 : \{0 < X_1 < 1\} \mid y = 2\right) = 1$$

---

[9] Likewise, the marginal densities for X_1 and X_2 are identified by integrating out the other variable from their joint density. That is

$$\begin{aligned} &\int_{-2}^{2} f\left(X_1, X_2\right) dX_2 \\ = \quad & f\left(X_1 : \{-1 < X_1 < 1\}\right) = \frac{1}{2} \end{aligned}$$

and

$$\begin{aligned} &\int_{-1}^{1} f\left(X_1, X_2\right) dX_1 \\ = \quad & f\left(X_2 : \{-2 < X_2 < 2\}\right) = \frac{1}{4} \end{aligned}$$

This consistency check brings us back to our starting point.

and

$$f\left(X_2 : \{Y - 1 < X_2 < Y + 1\} \mid y = 2\right) = \frac{1}{3 - 2}$$

or

$$f\left(X_2 : \{1 < X_2 < 2\} \mid y = 2\right) = 1$$

The aggregate report is informative for both variables, X_1 and X_2. For example, updated beliefs imply

$$E\left[X_1 \mid y = 2\right] = \frac{1}{2}$$

and

$$E\left[X_2 \mid y = 2\right] = \frac{3}{2}$$

and

$$E\left[X_1 + X_2 \mid y = 2\right] = 2$$

Following a brief overview of chapter organization, we explore probability as logic in other accounting settings.

1.3 Overview

The second chapter introduces several recurring accounting examples and their underlying theory including any equilibrium strategies. We make repeated reference to these examples throughout later chapters as well as develop other sparser examples. Chapter three reviews linear models including double residual regression (*FWL*) and linear instrumental variable estimation. Prominent examples survey some econometric issues which arise in the study of earnings management as equilibrium reporting behavior and econometric challenges associated with documenting information content in the presence of multiple sources of information.

Chapter four continues where we left off with linear models by surveying loss functions and estimation. The discussion includes maximum likelihood estimation, nonlinear regression, and James-Stein shrinkage estimators. Chapter five utilizes estimation results surveyed in chapter four to discuss discrete choice models — our point of entry for limited dependent variable models. Discrete choice models and other limited dependent variable models play a key role in many identification and estimation strategies associated with causal effects.

Distributional and structural conditions can sometimes be relaxed via nonparametric and semiparametric approaches. A brief survey is presented in chapter six. Nonparametric regression is referenced in the treatment effect discussions in chapters 8 through 12. In addition, nonparametric regression can be utilized to evaluate information content in the presence of multiple sources of information as introduced in chapter three. Chapter seven surveys repeated-sampling inference methods with special attention to bootstrapping and Bayesian simulation. Analytic demands of Bayesian inference are substantially reduced via Markov chain

Monte Carlo (*McMC*) methods which are briefly discussed in chapter seven and applied to the treatment effect problem in chapter 12.

Causal effects are emphasized in the latter chapters — chapters 8 through 13. A survey of econometric challenges associated with endogeneity is included in chapter eight. This is not intended to be comprehensive but a wide range of issues are reviewed to emphasize the breadth of extant work on endogeneity including simultaneous probit, strategic choice models, duration models, and selection analysis. Again, the Tuebingen-style treatment effect examples are introduced at the end of chapter eight.

Chapter nine surveys identification of treatment effects via ignorable treatment conditions, or selection on observables, including the popular and intuitively appealing propensity score matching. Tuebingen-style examples are extended to incorporate potential regressors and ask whether, conditional on these regressors, average treatment effects are identified. In addition, treatment effects associated with the asset revaluation regulation example introduced in chapter two are extensively analyzed.

Chapter ten reviews some instrumental variable (*IV*) approaches. *IV* approaches are a natural response when available data do not satisfy ignorable treatment conditions. Again, Tuebingen-style examples incorporating instruments are explored. Further, treatment effects associated with the report precision regulation setting introduced in chapter two are analyzed.

Chapter 11 surveys marginal treatment effects and their connection to other (average) treatment effects. The chapter also briefly mentions newer developments such as dynamics and distributions of treatment effects as well as general equilibrium considerations though in-depth exploration of these issues are beyond the scope of this book. Bayesian (*McMC*) analysis of treatment effects are surveyed in chapter 12. Analyses of marginal and average treatment effects in prototypical selection setting are illustrated and the regulated report precision setting is revisited.

Chapter 13 brings the discussion full circle. Informed priors are fundamental to probability as logic. Jayne's [2003] widget problem is a clever illustration of the principles of consistent reasoning in an uncertain setting. Earnings management as equilibrium reporting behavior is revisited with informed priors explicitly recognized. We only scratch the surface of potential issues to be addressed but hope that others are inspired to continue the quest for a richer and deeper understanding of causal effects associated with accounting choices.

1.4 Additional reading

Jaynes [2003] describes a deep and lucid account of probability theory as the logic of science. Probabilities are assigned based on the maximum entropy principle (*MEP*).