Towards a Formalization of Policy Analytics

Dustin Chambers

MERCATUS WORKING PAPER

All studies in the Mercatus Working Paper series have followed a rigorous process of academic evaluation, including (except where otherwise noted) at least one double-blind peer review. Working Papers present an author's provisional findings, which, upon further consideration and revision, are likely to be republished in an academic journal. The opinions expressed in Mercatus Working Papers are the authors' and do not represent official positions of the Mercatus Center or George Mason University.

MERCATUS CENTER
George Mason University
3434 Washington Blvd., 4th Floor, Arlington, Virginia 22201
www.mercatus.org

Dustin Chambers, "Toward a Formalization of Policy Analytics," Mercatus Working Paper, Mercatus Center at George Mason University, Arlington, VA, March 2021.

Abstract

A relatively new field within economics, policy analytics has experienced rapid growth and yielded insights in many disciplines within the profession. Nonetheless, a theoretical framework from which to conceptualize this nascent field has yet to emerge in the literature. This paper introduces a few concepts that should be useful in establishing a formal policy analytics framework.

JEL codes: C18, C49, D80, K20

Keywords: policy analytics; framework; latent variable estimation; natural language processing

Author Affiliation and Contact Information

Dustin Chambers
Department of Economics and Finance
Salisbury University
DLChambers@salisbury.edu

This paper can be accessed at https://www.mercatus.org/publications/regulation/formalization-policy-analytics

Toward a Formalization of Policy Analytics

Dustin Chambers

Introduction

In the span of just a few years, researchers in the field of policy analytics have examined the impact of public policy on a host of economic outcomes. These studies have yielded novel empirical insights in a diverse and growing set of economic fields, including agricultural economics (see, for example, Chambers and Malone [2017]), development economics (see, for example, Chambers, McLaughlin, and Stanley [2019]), finance (see, for example, Rotthoff and Richards [2020]), industrial organization (see, for example, Bailey and Thomas [2017]), labor economics (see, for example, Bailey, Thomas, and Anderson [2019]), macroeconomics (see, for example, Coffee, McLaughlin, and Peretto [2020]), and public choice (see, for example, Mulholland [2019]). Many of these discoveries, which would have been impossible 20 years ago, are a direct consequence of machine-learning algorithms, which leverage large datasets and modern computers to detect complex patterns in data, thereby yielding improved predictions over time. Despite these rapid gains, a theoretical framework from which to conceptualize the field of policy analytics has yet to emerge. This paper, therefore, seeks to fill this gap in the literature by sketching a formal policy analytics framework.[1]

Section 1 will introduce a few instrumental concepts of a basic framework in which latent information is embedded in natural language expressions. Section 2 elaborates on these concepts

[1] Within information theory, Shannon (1948) developed a concept known as Shannon entropy, which when applied to text, measures the level of uncertainty regarding missing letters within a body of text. Alternatively, Shannon entropy can be interpreted as a measure of the average information content of each character in a text. Regardless, Shannon entropy is unrelated to the interpretation of text (e.g., the policy intent of a text), which is the principle focus of this paper.

to describe policy texts and policy variables. Section 3 applies these concepts to the estimation of policy variables, and section 4 concludes the paper.

1. Framework

Consider the individual alphanumeric characters and symbols (c) belonging to a character set (C)—for example, the ASCII character set. Concatenating said characters results in a string ($s \in C^N$), and the set of all possible strings is denoted $S = \{s|s \in C^N\}$. Most strings contained in S are nonsensical to a human and not considered a valid expression of any language. The subset of strings that are valid expressions of any written language is denoted $L \subset S$. Each element of L is a natural language expression and denoted $\ell \in L$. For the purposes of policy analytics, these natural language expressions (ℓ) consist of any government text (both sections of whole documents and entire documents), including the Constitution, legal statutes, civil and criminal laws, case law, government regulations, administrative law, guidance documents, and so on. For simplicity, these various government texts will be generically labeled as "policy text." Further, assume that the policy text contains information of interest, denoted by y. This information, henceforth the "policy variable," may (or may not) be plainly stated within the text. If clearly stated, the extraction of y is trivial. However, if the policy variable is not clearly stated (e.g., can be inferred by someone with background knowledge, training, and experience), then additional information (X) may be needed to extract (or estimate) the policy variable. In the latter case, y is essentially a latent policy variable. Finally, consider an indicator function, $I(y, \ell)$, which equals 1 if the policy variable (y) is informationally contained within the policy text (ℓ) and which equals 0 otherwise.

Given the above assumptions, this can be formally modeled as follows. Consider a probability space $(\Omega, \mathcal{F}, P)$ in which Ω is a sample space, $\mathcal{F}$ is an event space, and P is a probability function over the event space. The sample space consists of all pairs of policy variables and the policy texts in which they are contained, that is, $\Omega \equiv \{(\mathrm{y}, \ell) | \ell \in L, I(y, \ell) = 1\}$. The event space ($\mathcal{F}$) is a σ-algebra over Ω, and $P: \mathcal{F} \to [0,1]$ is a measure satisfying $P(\Omega) = 1$.

2. Policy Texts and Variables

The process of encoding policy variables with text yields natural language expressions of varying quality and clarity. This section builds a simple taxonomy of these policy texts and discusses the implications for policy analytics.

2.1 Unambiguous and Informationally Complete Statements

Given a policy text, the task of policy analytics is to extract specific, targeted information (i.e., the policy variable) contained within that text. If the desired information is stated plainly within the body of the text, it can be easily extracted by a simple reading of the text or by use of automated "data scraping" algorithms. Stated more formally, in such cases there exists a function that maps policy text onto the policy variable (i.e., $W: \ell \to y$), and the policy text is said to be an *unambiguous and informationally complete natural language expression with respect to* y. In other words, if the policy text is clearly written and contains enough information about the policy variable such that someone who understands the language it is written in can unambiguously determine its value (i.e., with probability 1), then W produces perfect estimates of y. For example, if someone is interested in the maximum contaminant level of arsenic allowed by the Environmental Protection Agency in community water systems (as reported in the US Code of Federal Regulations [CFR]), then 40 C.F.R. § 141.62(b) contains a table providing this value (0.010 mg/l), along with maximum contaminant levels for 15 other inorganic contaminants.

2.2 Unambiguous and Informationally Incomplete Statements

If the desired information is not plainly stated within the body of the text, the task of extracting it is much harder and may require a subject matter expert (e.g., a lawyer) or an appropriately trained machine-learning algorithm to deduce the latent policy variable encoded within the text. This can be modeled by assuming that exogenous information (X) is required to estimate or determine the value of the latent policy variable. Stated more formally, consider the information set (Ψ) which includes all factors influencing the language in which the latent policy variable is encoded, including authorizing legislation, the agency or body drafting the policy, the identity of the author penning the rule, existing case law and precedent, politics, and so on. If there exists a function (W) and exogenous information set ($X \subseteq \Psi$) enabling the mapping of policy text and exogenous information onto the policy variable of interest (i.e., $W: \{\ell, X\} \to y$), the information set X is said to be *perfectly informative*, and the policy text is said to be an *unambiguous and informationally incomplete natural language expression with respect to* y. However, if one lacks enough exogenous information, the second-best option is to form an estimate of the latent policy variable (this will be covered in section 3).

In other words, the additional context provided by the exogenous information enables someone who understands the language the policy text is written in to determine the value of the policy variable (possibly with probability 1 if sufficient information is available). For example, 14 C.F.R. § 91 provides clear and detailed federal regulations regarding aircraft flight rules. Although the language contained in these regulations are generally indecipherable to a layperson, they are clearly understood by someone with sufficient background information and knowledge of the subject matter (e.g., a commercial pilot).

2.3 Ambiguous Statements

Now consider a policy text that is so poorly written that the policy variable encoded within it cannot be extracted by a subject matter expert or an appropriately trained machine-learning algorithm with certainty. Such a policy text is said to be an *ambiguous natural language expression with respect to* y.[2] This can be modeled by assuming that there does not exist an exogenous information set and estimator capable of determining the value of the latent policy variable contained within a policy text with probability 1. Instead, it is assumed that if appropriate exogenous information is available, then one can form an estimate of the latent policy variable (this will be covered in section 3).

Many academic researchers have documented laws, regulations, and other policy texts that they deem to be ambiguous, but IRS publications and the Internal Revenue Code appear to be especially problematic (see Givati [2009] and Blank and Osofsky [2017], among others). For example, in response to the Tax Cuts and Jobs Act of 2017 (Public Law 115–97), the US Chamber of Commerce sent a 15-page document to the Department of the Treasury seeking clarification regarding how specific provisions would impact member firms (see Harris [2018]). In such cases, policy experts can narrow the possible interpretation of a statute or regulation and predict its likely intent but cannot determine the true meaning with certainty.

[2] This implies that possessing complete knowledge (Ψ) does not permit the determination of y with certainty since by assumption no subset $X \subseteq \Psi$ will permit the determination of y with certainty.

2.4 Implications for Policy Analytics

If the regulation is clearly stated and sufficiently informative, then the value of the latent policy variable can be ascertained by the mere reading of the regulation—that is, ℓ is an *unambiguous and informationally complete natural language expression with respect to* y. In the real world, this clarity is rare as some level of context (i.e., conditional information) is required to understand a given regulation and estimate the value of the latent policy variable. Therefore, a policy analyst would be fortunate to find an *unambiguous and informationally incomplete natural language expression with respect to* y. In most cases, policy analysts must be prepared to deal with *ambiguous natural language expressions with respect to* y. The salient issue then becomes one of sufficient conditional information (X) to produce a reasonable forecast of y and, insofar as y is measurable (in the space of real numbers), a statistically unbiased estimate of y.

3. Estimation of Policy Variables

Given that most policy text is complex legalese, policy variables are rarely ascertainable with certainty via a simple reading of policy text. In such cases, additional information must be brought to bear in order to produce an estimate of the target policy variable. The following sections describe the unbiased estimation of policy variables and apply these concepts to the real-world example of the RegData database and discuss the efficiency of such estimates.

3.1 Conditions for Unbiased Estimation of y

Suppose a given policy text is *unambiguous and informationally incomplete*. By definition, if one possesses sufficient information, there exists a function ($W: \{\ell, X\} \rightarrow y$) that will determine the value of the policy variable with probability 1. However, if the available information (X) is *not* a perfectly informative subset of Ψ, then said information is insufficient to determine the value of y with certainty (i.e., with probability 1). Alternatively, if a policy text is ambiguous, then the latent policy variable contained within a policy text cannot be determined with probability 1. In either case, an estimate of the policy variable must be obtained using the available information.

Assuming that y is a real number ($y \in \mathbb{R}$), consider a candidate estimator of y, $H\colon \{\ell, X\} \rightarrow \hat{y}$, which maps policy text and exogenous information onto estimated values of y (denoted $\hat{y}$):

$$\hat{y} = H(\ell, X).$$

If the available information is sufficient to generate an unbiased estimate of y from a given policy text (ℓ_0), then by definition

$$\mathrm{E}[H(\ell_0, X)|X \in \tilde{X}(\ell_0, H)] = \mathrm{E}[\hat{y}|X \in \tilde{X}(\ell_0, H)] = y,$$

where $\tilde{X}(\ell_0, H)$ is the set of all information sets that produce unbiased estimates of y for a given estimator (H) and policy text (ℓ_0):

$$\tilde{X}(\ell_0, H) \equiv \{X \subset \Psi | \ \mathrm{E}[H(\ell_0, X)] = y\}.$$

If true, the relationship between y and H can be rewritten in the following familiar form:

$$y = H(\ell, X) + u,$$

where u is a mean zero error term that is orthogonal to X—that is, $\mathrm{E}[u|X] = \mathrm{E}[u] = 0$.

However, if the set $\tilde{X}(\ell_0, H)$ is empty, then unbiased estimation of a given policy text is not possible by way of the estimator (H). If the policy text is *unambiguous and informationally incomplete*, then by construction there exists an estimator (say, G) for which the resulting set $\tilde{X}(\ell_0, G)$ is non-empty. That said, if the policy text is *ambiguous*, then there is no guarantee that there exists an estimator (G) for which the resulting set $\tilde{X}(\ell_0, G)$ is non-empty, rendering an unbiased estimation of y impossible.

3.2 Application: RegData

Natural language processing (NLP) machine-learning algorithms train classifiers using training datasets—that is, a large pool of natural language expressions ($\ell' \in L_0$) and corresponding known values of the latent policy variable ($y' \in Y_0$). This training data constitutes the conditioning data ($X_0 = \{L_0, Y_0\}$) used to classify regulatory text:

$$H(\ell, X_0) = \hat{y}.$$

In the example of the Mercatus Center's RegData dataset,[3] training data (X_0) are used to examine regulatory text (ℓ) and determine the probability that said text belongs to any of a large number (*N*) of North American Industrial Classification System–coded industries ($\gamma_i, i = 1, \dots, N$):

$$H(\ell, X_0) = \widehat{\Pr}(y = \gamma_i | \ell, X_0)\ \forall i.$$

Clearly, RegData's industry probability estimates are valid if and only if the conditional data (X_0) used to train the NLP classifiers are *informationally sufficient* ($X_0 \in \tilde{X}(\ell, H)$). Although a direct test of this assumption is hard to devise, this framework provides a way of testing the validity of alternative estimates of a common, latent policy variable.

Prior to RegData, the most common way to measure the extent of regulations by industry (or for the entire economy) was to count the pages of applicable regulations in the *Code of Federal Regulations* or the *Federal Register* (see, for example, Friedman [1962] and Dawson

[3] For detailed information about the methodology used to construct RegData, see Al-Ubaydli and McLaughlin (2017) and McLaughlin and Sherouse (2019). For the latest version of RegData, see McLaughlin (2020).

and Seater [2013]). Even if we make the false assumption that all federal regulations are unambiguous and informationally incomplete natural language expressions with respect to industry (or aggregate) regulation, this estimator still cannot produce unbiased estimates of regulation. To see why this is true, consider the page-count estimator, $H(\ell, X_C)$, where ℓ is the combined corpus of all regulations pertaining to an industry (or nation) and X_C is the set of conditioning information used to estimate the level of said regulation. Since the page-count estimator simply measures the length of ℓ (expressed in page counts rather than words or characters), the corresponding conditioning set X_C must be empty. However, by assumption federal regulations are unambiguous and informationally incomplete natural language expressions, thus they do not contain enough information to estimate the level of regulation at the industry level or in the aggregate.

3.3 Efficient Estimation of y

Suppose a given policy text (ℓ_0) is either *unambiguous and informationally incomplete or ambiguous* and there exist two estimators (H and G) for which unbiased estimates of y can be obtained—that is, $\tilde{X}(\ell_0, H)$ and $\tilde{X}(\ell_0, G)$ are both non-empty sets. Then, if the following condition is satisfied, H is a *relatively more efficient estimator* than G with respect to ℓ_0:

$$Var\big(H(\ell_0, X)\big) < Var\big(G(\ell_0, X)\big) \; \forall \, X \in \{\tilde{X}(\ell_0, H) \cap \tilde{X}(\ell_0, G)\}.$$

It is important to note that this concept of relative efficiency pertains to common information sets that produce unbiased estimates using both estimators and does not rule out the possibility that G is more efficient than H with respect to ℓ_0, as defined below:

$$\min_{X_1 \in \tilde{X}(\ell_0, G)} Var\big(G(\ell_0, X_1)\big) < \min_{X_2 \in \tilde{X}(\ell_0, H)} Var\big(H(\ell_0, X_2)\big).$$

Finally, an unbiased estimator (H) is said to be *inefficient* if it satisfies the following condition:

$$Var\big(H(\ell_0, X_1)\big) < Var\big(H(\ell_0, X_2)\big)\ s.t. X_1 \subset X_2 \in \tilde{X}(\ell_0, H).$$

If a policy text (ℓ_0) is *unambiguous and informationally incomplete*, then by definition there exist information $X^* \subseteq \Psi$ and a mapping Wenabling the determination of y with complete certainty—that is, $Var\big(W(\ell_0, X^*)\big) = 0$. Nonetheless, in the real world, knowledge of both X^* and W are likely infeasible. If a policy text (ℓ_0) is *ambiguous*, then by definition there does *not* exist an information set $X^* \subseteq \Psi$ and a corresponding mapping $W: \{\ell_0, X^*\} \rightarrow y$. As such, any estimator (H) capable of producing unbiased estimates of y must satisfy the following condition:

$$\min_{X \in \tilde{X}(\ell_0, H)} Var\big(H(\ell_0, X)\big) > 0.$$

4. Conclusion

In its relatively short existence, the rapidly evolving field of policy analytics has impacted many economic disciplines. Yet, a formal framework from which to conceptualize the field of policy analytics has yet to emerge. This paper is the first attempt (to our knowledge) to fill this gap in the literature. Using this framework, the paper describes the conditions in which unbiased estimation of the latent policy variable is possible and applies these concepts to the RegData database.

References

Al-Ubaydli, O., and P. A. McLaughlin. 2017. "RegData: A Numerical Database on Industry-Specific Regulations for All United States Industries and Federal Regulations, 1997–2012." *Regulation and Governance* 11: 109–23.

Bailey, J., and D. Thomas. 2017. "Regulating Away Competition: The Effect of Regulation on Entrepreneurship and Employment." *Journal of Regulatory Economics* 52: 237–54.

Bailey, J., D. Thomas, and J. R. Anderson. 2019. "Regressive Effects of Regulation on Wages." *Public Choice* 180(1): 91–103.

Blank, J. D., and L. Osofsky. 2017. "Simplexity: Plain Language and the Tax Law." *Emory Law Journal* 66: 189–264.

Chambers, D., and T. Malone. 2017. "Quantifying Federal Regulatory Burdens in the Beer Value Chain." *Agribusiness: An International Journal* 33(3): 466–71.

Chambers, D., P. A. McLaughlin, and L. Stanley. 2019. "Regulation and Poverty: An Empirical Examination of the Relationship between the Incidence of Federal Regulation and the Occurrence of Poverty across the US States." *Public Choice* 180(1): 131–44.

Coffee, B., P. A. McLaughlin, and P. Peretto. 2020. "The Cumulative Cost of Regulations." *Review of Economic Dynamics* 38: 1–21.

Dawson, J., and J. Seater. 2013. "Federal Regulation and Aggregate Economic Growth." *Journal of Economic Growth* 18(2): 137–77.

Friedman, M. 1962. *Capitalism and Freedom*. Chicago: University of Chicago Press.

Givati, Y. 2009. "Resolving Legal Uncertainty: The Unfulfilled Promise of Advance Tax Rulings." *Virginia Tax Review* 29: 137–75.

Harris, C. L. 2018. "U.S. Chamber Comments regarding Regulatory Guidance under Public Law No. 115–97." Comments to David J. Kautter and William M. Paul, US Chamber of Commerce, Washington, DC, March 7. https://www.uschamber.com/comment/us-chamber-comments-regarding-regulatory-guidance-under-public-law-no-115-97.

McLaughlin, P. A. 2020. "RegData US 3.2 Annual (dataset)." Mercatus Center at George Mason University, Arlington, VA.

McLaughlin, P., and O. Sherouse. 2019. "RegData 2.2: A Panel Dataset on US Federal Regulations." *Public Choice* 180(1–2): 43–55.

Mulholland, S. E. 2019. "Stratification by Regulation: Are Bootleggers and Baptists Biased?" *Public Choice* 180(1): 105–30.

Rotthoff, K. W., and T. Richards. 2020. “Regulatory Shocks and Market Valuation.” Mercatus Working Paper, Mercatus Center at George Mason University, Arlington, VA.

Shannon, C. E. 1948. “A Mathematical Theory of Communication.” *Bell System Technical Journal* 27(3): 379–423.