Should Legal Empiricists Go Bayesian?

Jeff Strnad*

Abstract

Bayesian empirical approaches appear frequently in fields such as engineering, computer science, political science and medicine, but almost never in law. This article illustrates how such approaches might be very useful in empirical legal studies. In particular, Bayesian approaches enable a much more natural connection between the normative or positive issues that typically motivate such studies and the empirical results.

Preliminary – do not quote or circulate without author's permission.

September 4, 2006 Version

*Charles A. Beardsley Professor of Law, Stanford University. Marc Fernandes, Jonathan Hennessy and Ethan Siller provided excellent research assistance. I am thankful for generous financial support from the John M. Olin Program in Law and Economics at Stanford Law School.

Contents

1 Introduction

Legal academics doing empirical work have used frequentist rather than Bayesian approaches almost exclusively.[1] In contrast, Bayesian methods are common in fields such as political science, engineering and medicine. This article argues that both a Bayesian perspective and particular Bayesian methods have much to offer legal empiricists. In addition to theoretical discussion, the article applies both the perspective and some of the methods to the heated dispute over the impact of right-to-carry laws on various types of crime.[2]

At the core of the frequentist approach is the idea that the available data is a sample from a larger, real or imagined population. The statistical challenge is to develop "estimators" for various parameters of interest such as regression coefficients. A good estimator will have desirable properties in repeated samples. In law and other areas, the repeated samples often are hypothetical since the data is observational rather than experimental. In addition, a frequentist will be interested in the asymptotic properties of various estimators, the behavior of the estimator as the sample size from the population tends to infinity. For example, one trait that makes an estimator "good" is "consistency," the asymptotic convergence of the estimator to some "true" population value.

A strict version of the frequentist approach requires that the researcher

[1] There are a number of examples, some noted in what follows, where researchers apply Bayesian approaches in empirical examinations of legal issues, but almost none of this work is by legal academics.

[2] In order to make the article accessible to a broad audience, the discussion in the text is at a very basic level, deliberately avoiding mathematics beyond simple algebra and developing some concepts in detail that already will be familiar to many readers engaged in empirical work. In contrast, in the interest of brevity, some of the footnotes freely presume technical expertise and use statistical or mathematical terminology that may be unfamiliar to some readers. With such readers in mind, these more technical footnotes cover only topics that are not essential to understanding the main arguments and results in the article. In the few places where the text does include mathematics beyond simple algebra, explanations are given that should obviate the need to understand the mathematics.

Since there are many excellent texts available, I do not attempt to describe the basics of the Bayesian methods that underlie the applications presented here. Readers familiar with frequentist econometrics might find [Koop 2003], [Lancaster 2004], [Gelman, et.al. 2004], [Geweke 2005], [Jackman 2006] or [Poirier 1996] to be good starting points among others. The author intends to make the programs (primarily MATLAB-based) and data that underlie the reported results generally available upon publication of the paper.

specify a single model and then use frequentist (rather than Bayesian) methods to test that model or to estimate some parameter of interest. Statistical significance often is the measure used to assess the model or particular parameter estimates. E.g., for a regression with more than twenty observations, econometrics students learn the rule of thumb that a t-statistic with absolute value greater than two indicates that the null hypothesis that the associated coefficient is zero may be rejected at the 5% level.

In the frequentist framework, returning to the same data with a different model, perhaps based on what the researcher has learned from the results under the first model, is not legitimate unless certain major adjustments are made. As discussed in [Theil 1971], if one simply runs a second regression on the same data, one cannot interpret the resulting t-statistics and other measures of statistical significance in the usual way. The intuition is simple and may be illustrated by considering the rule of thumb for 5% significance. A t-statistic of greater than two in absolute value indicates that, under the null hypothesis that the associated coefficient is zero, there is less than a 5% probability that the associated coefficient would deviate that much from zero through chance fluctuation. However, if the researcher runs two separate models, the probability that one will see a t-statistic greater than two in absolute value for a particular coefficient in at least one of them is higher than 5%.

Estimating regressions with different specifications in a frequentist framework is sometimes referred to as "pretesting." A final reported regression is not the only one that the researcher ran. The researcher "pretested" by estimating several models before settling on the reported variant. As demonstrated by [Danilov and Magnus 2004], the error (in terms of overstating statistical significance) resulting from pretesting can be very large.

Many frequentist techniques have their roots in experimental sciences. In that setting, if the researcher wishes to test an alternative model, there is the option of gathering more data. Testing the second model on a second, independent set of data avoids any pretesting problem. In law, however, it is typically the case that the researcher is dealing with observational data that cannot be extended by additional experimentation.

The pretesting problem raises issues of conscious or unconscious researcher bias. A researcher is likely to favor a specification that is "reasonable," but it is hard to imagine that researchers are not influenced by their prior beliefs about the outcomes. The potential for trouble becomes more pronounced when one considers aggregate effects. Suppose that one researcher publishes

results linking a particular class of crimes to economic rather than "punishment" variables. If this publication motivates researchers with strong beliefs that punishment deters to try alternative specifications, a few of these specifications might overturn the original result simply by chance, i.e., even if the original result is "true." It is likely that some of the deterrence-oriented researchers will find these specifications to be "reasonable" and will report them. Note that this result can occur even if each researcher only estimates a single alternative specification. At the individual level, there may be no pretesting or conscious bias. At the aggregate level, we have an outcome that brings to mind the famous joke attributed by [Leamer 1983] to Coase, "If you torture the data long enough, Nature will confess."

In short, there is both an "internal" and an "external" problem. The "internal" problem is that a researcher might be unsure of the true model and therefore be motivated, quite legitimately, to try more than one specification. In the frequentist framework, this experimentation degrades the final reported results. In addition, conscious or unconscious bias may motivate specification choices. The "external" problem is that there often are multiple specifications from various researchers, all with very human biases. Each researcher may have a cogent set of arguments about why his or her specification is "reasonable," but the specifications may result in very different positive conclusions or policy prescriptions. A paper that reports a single specification leads to questions about why the author chose that particular one and whether conscious or unconscious bias had anything to do with it. Would you, the reader, have come up with a different specification and different results if you had done the work?

Bayesian approaches to estimation and testing differ substantially from frequentist ones. The target of a Bayesian analysis typically is the "posterior" probability distribution of some parameter or hypothesis, θ, of interest to the researcher, conditional on observing some data, D. The analysis begins with a specification of prior beliefs in the form of a probability distribution for θ. These beliefs represent (conceptually) the views of the research (or his or her audience) before seeing the data. Bayesian reasoning typically comes down to the application of "Bayes' rule." Consider the case where θ represents a finite number of discrete elements, such as alternative models or hypotheses. In this case, we can write down Bayes rule in terms of probabilities for each

discrete element.[3] In particular, we have:

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} \tag{1}$$

The posterior probability, $P(\theta|D)$ is on the left hand side. The prior probability, $P(\theta)$, occurs as a factor in the numerator of the right hand side. The other factor in that numerator is $P(D|\theta)$, usually termed the "likelihood." It is the probability that the data D occurs if in fact the value of the parameter is θ. The denominator of the right hand side is typically called the "marginal likelihood." It is the sum or integral of the numerator over all the values of θ, resulting in the probability that the data occurs summed across all values of θ.

Several general features are worth noting. In contrast to the frequentist approaches, Bayesian analysis requires the researcher to specify a likelihood function and a prior. These extra requirements are not necessarily disadvantages. Being able to describe the dependence of the results on prior beliefs (often in ways not possible using frequentist approaches) can be a big plus, especially when the context involves hotly contested issues, a common situation in law. For example, with respect to model selection, Bayesian approaches allow some headway both on the "external" and "internal" problems described above. One may begin explicitly with the a set of prior beliefs about the cogency of various models. As a result, the researcher is not forced into the straightjacket of adopting a single model for estimation. The Bayesian analysis produces posterior probabilities for the models. A researcher may report the results for a variety of priors or for very weak priors (expressing high uncertainty about the proper model) and thereby make the article more

[3]The rule sometimes is called "Bayes' Theorem," but it is a tautological relationship based on logical consistency for probability relationships rather than a theorem that requires some difficult proof. Consider, for example, the probability, $P(A \wedge B)$, that events A and B both occur. It follows from the definition of conditional probability that:

$$P(A \wedge B) = P(A|B)P(B)$$

and also that:

$$P(A \wedge B) = P(B|A)P(A).$$

As a result:

$$P(A|B)P(B) = P(B|A)P(A).$$

Bayes' rule follows from dividing through both sides of this equation by $P(A)$ or by $P(B)$.

relevant to readers who may have very different beliefs than the researcher about the models. Similarly, on the "external" front, Bayesian analysis allows one to assign probabilities to various alternative specifications advocated by different researchers and then run the models against each other.

Second, a strict Bayesian applies Bayes' rule to all aspects of a regression or other estimation. Thus, for a regression, one begins with a prior probability distribution for the coefficients and the error parameters. In some situations, it is possible to choose these prior distributions to be "noninformative," i.e., reflecting a situation where the researcher wishes to assert no prior knowledge about the coefficients or other aspects of the regression. The output of a Bayesian regression is a set of posterior distributions for the coefficients and the error parameters. The researcher can interpret these posterior distributions in any way that makes sense. For example, the researcher might report the mean and standard error of a coefficient based on the posterior distribution for that coefficient. This approach would parallel the information usually reported for frequentist regressions.[4] But the availability of posterior distributions means that the researcher may go further. In particular, it is possible to specify posterior probabilities for hypotheses rather than being confined to reporting test statistics and p values. A p value type of approach may be very misleading, and often fails to address the most cogent questions implicit in the research.

I use data and models collected in [Donohue 2004] to illustrate various points. That article examines the issue of whether adoption of right-to-carry ("RTC") laws, allowing citizens to carry concealed handguns, increases or decreases the incidence of nine categories of crime: violent crime, murder, rape, aggravated assault, robbery, property crime, burglary, larceny, and auto theft.[5] The data are panel data at the state level from the United States

[4]Instead of a confidence intervals, however, Bayesians typically report "highest probability density intervals." Unlike confidence intervals, these "HPDIs" have a direct probability interpretation regarding the coefficient. Under the researcher's posterior distribution, there is a 95% probability that the coefficient is in the 95% HPDI. For the Bayesian, there are only probability statements based on the available data. The Bayesian framework does no more than assure the consistency of probability judgments. In contrast, frequentist analysis presumes the coefficient has some true value that would be apparent if we had more data than we actually do have. Any probability statement in the frequentist framework about the value of the coefficient being in an interval is specious. It is either in the interval or it is not. The interval itself is random and what one can say is that there is a 95% chance that the interval includes the true value of the coefficient.

[5]The violent crime category is a composite that includes murder, rape, aggravated

covering the years 1977-1999. The article begins by noting the heated nature of the controversy, leading to accusations that various parties manufactured data or engaged in other questionable practices. The article concludes that the models used to support the view, championed most prominently by John Lott, that permitting concealed handguns reduces crime are very sensitive to changes in specification. To reach that conclusion, the article uses a two layered approach. First, it considers various "standard panel data" model specifications. Second, it develops its own specification approach, one that avoids giving heavier weight to states that adopt RTC laws early in the sample period by limiting the number of years pre- and post-adoption that can affect estimates of the coefficients of the RTC dummy variables.

The article implements the first layer by considering various specifications developed for purposes other than studying RTC laws in addition to considering a modified version of John Lott's specifications. The rationale is that examining these other models along with Lott's model will result in a greater likelihood of arriving at an objective answer:

> Because of the powerful ideological motivations of many gun researchers, a legitimate fear is that an analyst trying to prove a certain point might choose among a vast array of possible statistical models simply to generate a desired result. To address this concern, I report not only a modified version of Lott's original model (called the "Modified Lott" set of explanatory variables), but also the results of three other models that were developed by researchers to answer questions having nothing to do with RTC laws – one by Wentong Zheng (developed to look at the impact of lotteries on crime), one by William Spelman (developed to look at the impact of incarceration on crime), and one by John Donohue and Steve Levitt (developed to look at the impact of abortion legalization on crime).... Whatever infirmities these last three models have, we know that they were created by serious academics without any intention of skewing the estimates of the impact of the RTC laws. When we add a variable identifying the date of adoption of the RTC laws to these pre-existing statistical models, we an see if the results support – or refute – the more guns, less crime hypothesis. [Donohue 2004, pp. 631–32 (footnotes omitted)]

In essence, the article selects a group of models designed to have strong

assault and robbery. The property crime category is a composite that includes burglary, larceny and auto theft. Thus, seven of the nine categories are mutually exclusive while two are composites.

predictive power with respect to crime rates and then assesses the signs and statistical significance of added dummy variables that indicate the presence or absence of RTC laws.

I have chosen to use the material from [Donohue 2004] in the examples that follow for two reasons. First, it is an extremely thoughtful and well-executed piece of empirical work based on frequentist methods. To the extent that the Bayesian approaches highlighted here add value, it will not be on account of technical or analytical defects in the original work. Second, the material involves considerable complexity. Facing this complexity will reveal both strengths and various difficulties with Bayesian methods.

The rest of the article consists of three sections. Section 2 focuses on evaluating hypotheses. As is the case with much of the legal empirical literature, [Donohue 2004] casts many of its results in terms of whether or not certain coefficients or relationships are "statistically significant." This approach is equivalent to using p values and has the associated weaknesses alluded to earlier. Section 2 begins by discussing these weaknesses and then applies some Bayesian approximations and computations as an alternative approach.

Section 3 picks up on the idea in [Donohue 2004] of looking at a variety of models based on their likely predictive value with respect to crime rates. Bayesian approaches allow ready comparison of models and the capability of "averaging" across models in the face of uncertainty about which model (if any!) is correct. Section 3 reports the results of several comparison and averaging exercises in addition to discussing theoretical considerations.

Section 4 presents concluding thoughts. Appendix A lists the variables and models presented in [Donohue 2004], permitting a simplified discussion of the variables and models in the text. Appendix B contains some results arising out of the approaches in section 3 that are interesting but not central to the argument in that section.

The primary goal of the article is to illustrate how some Bayesian perspectives and methods can add value to legal empirical analysis. Although some of the results contribute to the empirical assessment of right-to-carry laws, I refrain from an all-out effort on that front, preferring to emphasize conceptual aspects and maintain a uniformly accessible technical level throughout. I leave some important aspects to a more technical sequel.

The range of Bayesian methods is very large and growing quickly. I make no attempt here to be comprehensive, preferring to focus on a few methods that are broadly illustrative. Many Bayesian methods, including some used in this article, require computational methods that go beyond what

is readily available in popular frequentist-based statistical packages such as STATA. With this fact in mind, I deliberately include several methods and approximations that are easy to implement within such packages.

2 Evaluating Hypotheses

Bayes's rule is essentially a consistency requirement for probabilistic reasoning. If you have a certain set of prior beliefs, the rule tells you what your posterior beliefs should be based on a certain likelihood relationship involving some evidence or data. For hypotheses, Bayesian approaches result in posterior probability information about whether or not a particular hypothesis is true. Consider, for example, the common approach in regression analysis of testing the null hypothesis that a certain regression coefficient, β, is equal to zero. The alternative hypothesis is that the coefficient does not equal zero. Typical terminology would be to call the null and alternative hypotheses something like H_0 and H_1 respectively:

$$H_0 : \beta = 0 \text{ (null hypothesis)}$$

$$H_1 : \beta \neq 0 \text{ (alternative hypothesis)}$$

A common assumption in regression analysis is that the dependent variable, "y," is stochastic while the independent variables collected in a matrix, "X," are not. If we follow that assumption and refer to the stochastic portion of the regression data as "y," then the posterior quantities of interest will be denoted $P(H_0|y)$ and $P(H_1|y)$, the respective posterior probabilities that the null and alternative are true.[6] As a starting point, the analysis requires the prior probabilities for each hypothesis, denoted $P(H_0)$ and $P(H_1)$ respectively. The prior transforms into the posterior via the likelihoods, $P(y|H_0)$ and $P(y|H_1)$.

For the null hypothesis, Bayes' rule is:

$$P(H_0|y) = \frac{P(y|H_0)P(H_0)}{P(y)}. \tag{2}$$

[6]In some models or data sets, X can be or must be taken as stochastic also. In that case, we would use terminology like $P(H_0|y, X)$ instead of $P(H_0|y)$. It is simpler to use the latter in the text, and the fuller former expression would add nothing to the discussion.

It is often convenient to consider the "posterior odds," $P(H_1|y)/P(H_0|y)$. It follows readily from Bayes rule that:

$$\frac{P(H_1|y)}{P(H_0|y)} = \frac{P(y|H_1)}{P(y|H_0)} \times \frac{P(H_1)}{P(H_0)} \tag{3}$$

where the three fractions in the equation are commonly referred to using the following names:

$$\text{posterior odds} \;=\; \text{Bayes factor} \;\times\; \text{prior odds.}$$

The "Bayes factor" looks like a frequentist likelihood ratio, and in some cases it is precisely such a likelihood or can be interpreted as such.[7] The name "Bayes factor" simply indicates that, in a Bayesian framework, this ratio transforms the prior odds into posterior odds based on what the observer has learned from the data.

There are two cases where the evidence does not change the odds that a hypothesis is true. The first is data driven. Under the data, the Bayes factor may turn out to be exactly one. In this situation, the posterior odds will be equal to the prior odds. The second case arises because the observer has "dogmatic" beliefs, putting a prior probability of zero or one on the hypothesis being true. Observers with dogmatic beliefs will not be influenced by the data.[8]

Under Bayesian approaches, the posterior odds ratio or the posterior probability for each of the hypotheses is the goal of the analysis. On the other hand, a common frequentist approach is to generate and examine "p values." Often results are stated in terms of "statistical significance" defined as p values less than some critical value such as $.10, .05$, or $.01$.

[7]Both the numerator of the Bayes factor and the denominator may incorporate prior probability elements. In these situations, the factor has a Bayesian taint, and, consequently, it is unclear whether interpreting it as a purely frequentist quantity makes sense. Developments later in the text will illustrate this point.

[8]The mathematics is simple. If the prior probability of the null hypothesis is $P(H_0) = 1$, then the prior probability of the alternative hypothesis, $P(H_1)$ as well as the ratio $P(H_1)/P(H_0)$ must be zero. As a result, no matter what value for the Bayes factor in equation (3) arises from the data, the posterior odds ratio will be zero. I.e., $P(H_0|y) = 1$, and the observer still is certain that the null hypothesis is true. If the observer has the opposite dogmatic belief, $P(H_0) = 0$, then both the prior and posterior odds ratios will be ∞. As a consequence, it will be true that $P(H_0|y) = 0$, and the observer will remain certain that the null hypothesis is false regardless of the weight of the evidence captured by the Bayes factor.

More formally, suppose that a random variable X has density $f(x, \beta)$ where the parameter, β, is unknown. To test the null hypothesis $H_0 : \beta = \beta_0$ against the alternative hypothesis $H_1 : \beta \neq \beta_0$, a typical frequentist approach would be to use a test statistic $T(x)$ with a known sampling distribution under the null to derive a "p value." If the sampling distribution is symmetric around β_0 with a density that is decreasing in $|x - \beta_0|$, then the p value is:

$$p = P(|T(x)| \geq |T(x_{observed})| \mid H_0 : \beta = \beta_0).$$

p is the probability of observing an outcome for X at least $|x_{observed} - \beta_0|$ away from β_0 under the sampling distribution for the test statistic T. For example, in an ordinary least squares regression where one assumes that error disturbances are normally and identically distributed, the sampling distribution for the coefficient estimator is a Student t distribution, and $T(x)$ is the "t statistic" commonly reported in regression results. Researchers commonly state that the associated coefficient is "statistically significant" if the associated p-value is less than some value such as $.10, .05$, or $.01$, and they often report regression results with stars or bold emphasis to indicate coefficients that are significant. [Donohue 2004] uses the .05 level as a benchmark as is quite common.[9]

The next subsection describes the applications of the p value approach in [Donohue 2004]. Three subsequent subsections discuss the weaknesses of this approach from a Bayesian perspective and how the results in [Donohue 2004] appear from that perspective.

2.1 P Values in John Donohue's Right-to-Carry Regressions

[Donohue 2004] contains a staggering number of regressions involving various specifications. The focus is on one or more dummy variables indicating the presence, absence or duration of state right-to-carry laws in regressions where the logs of various crime rates are the dependent variables. The coefficients of the dummy variables represent the percentage change in crime induced by the presence of right-to-carry laws. Using the standard frequentist approach, the centerpiece of the analysis is an examination of the signs and statistical

[9] Habit dating from the salient presence of the .05 level in the earliest tables for p values may be part of the reason. At the time, tables were critical since electronic computation devices were not available. [Freedman, Pisani & Purves 1998, p. 548]

significance of these coefficients. In some cases, the regressions contain one dummy variable (or duration variable) that aims at estimating the average treatment effect (across all states) on each of nine crime categories of right-to-carry laws. In contrast to this "aggregate" approach, other regressions include 26 separate dummies, one for each state that adopted a right-to-carry law during the 1977-1999 period studied. These "state specific" dummies yield a disaggregated view of the situation, using the states that did not institute a right-to-carry law during the period as controls.

In the "standard panel data" part of the paper, Donohue considers three specifications for each of the four models mentioned in the introduction: Modified Lott, Donohue-Levitt, Spelman and Zheng. One specification contains a simple dummy variable (or in the state specific case, 26 simple dummy variables, one for each of the states adopting right-to-carry laws during the study period) indicating the presence or absence of a right-to-carry law in the applicable state and year. A second "spline" specification replaces the dummy or dummies for the presence of such a law with a variable indicating the number of years since adoption. A third specification uses the simple dummy or dummies but adds a time trend dummy aimed at picking up background crime trends in states that adopted right-to-carry laws during the period of the study.

These twelve variations (four models each with three specifications) crossed with nine crime categories result in $108 = 12 \times 9$ coefficient estimates for the aggregate case where there is one RTC ("right-to-carry") dummy in each regression and $2808 = 12 \times 9 \times 26$ coefficient estimates for RTC dummies in state specific regressions. As a result, the "standard panel data" part of the paper involves $2916 = 108 + 2808$ significance tests based on p values. For the aggregate RTC dummies, Donohue simply reports the results and their statistical significance. For the state specific RTC dummies, Donohue takes three separate approaches to reporting and interpreting the results for each of the 108 variations of specifications crossed with crime categories:

(1) the number of positive estimated coefficients minus number of negative estimated coefficients;

(2) the number of statistically significant positive estimated coefficients minus the number of statistically significant negative estimated coefficients;

(3) the state population weighted mean of the state specific estimates and a significance indiction for the mean.

In addition to the "standard panel data" models, for each of nine crime categories [Donohue 2004] also reports the results of eight specifications that

implement an approach developed by [Autor, Donohue & Schwab 2002]. This "ADS" approach limits the operation of the RTC dummy to a window around the adoption year which Donohue calls the "treatment window." Without this limitation, there is a danger that crime patterns in early-adopting states that occur long after adoption will heavily influence the RTC dummy coefficients even though they do not plausibly flow from the RTC laws.[10] Because of the short span of the data, there are only a handful of early-adopters. Part of the motivation for the applying the ADS approach is that Donohue found (in earlier research with Ian Ayres) that the model predicts large swings in crime with a delay of 10 years or more. As discussed in [Donohue 2004, pp. 635-636], Ayres and Donohue speculate that this result arises from the coincident timing of the crack epidemic, making the RTC laws appear to be much more effective than they are. In any event, the reported ADS results involve $72 = 8 \times 9$ additional p-value based evaluations of the significance of an RTC dummy, raising the total to $2988 = 108 + 2808 + 72$.

It is important to consider which questions these various approaches and the huge set of p value based tests address. It is clear from the discussion in the article that multiple issues are of interest, including at a minimum: assessing the "more guns, less crime" claim generally, examining outcomes for various crime categories, determining how sensitive the results are to alterative specifications, and weighing the possibility that RTC laws generally have negligible effects on crime. The analysis below will address many of these issues, but the bulk of the discussion will be more general. The main goal is not to critique [Donohue 2004] in particular or to make new empirical claims about RTC laws, but, rather, to look at some of the ways in which Bayesian perspectives and methods might add value.

One nice feature of Bayesian approaches is that they are capable of speaking to researchers or readers who have different prior beliefs about particular issues. For example, there is the issue of whether the RTC laws have a negligible effect. Different readers might assign different prior probabilities to the negligible effect hypothesis for which the null hypothesis that the coefficients on the RTC dummies is zero is an approximation. By computing Bayes factors, we can tell any of these readers how their beliefs should change in the

[10] In effect, the ADS approach takes the dogmatic position that the bulk of the differences induced by RTC law adoptions will show up within some span of years after adoption. Of course, it is possible that such changes take longer to percolate into society. It may take quite a while for the gun carrying population to increase and for criminals to learn that there is a much higher likelihood that potential victims will respond with gunfire.

face of the evidence. An alternative way of proceeding is to aim at what some would call an "objective" Bayesian analysis by picking a prior which is "neutral." An appealing choice in facing the "negligible effect" hypothesis might be to consider a prior assigning equal, 50%, probabilities to the null hypothesis that the effects are zero (or negligible) and to the alternative that they are not. We will use this choice below in studying p values. However, skepticism about whether there can be any generally accepted "objective" Bayesian analysis is warranted. A central feature of that analysis is to connect potentially diverse prior beliefs with posterior beliefs. In some cases, it may turn out that the posterior results are not very sensitive to the prior beliefs, but that is a serendipitous empirical outcome. An intellectually safe course is to consider "objective" Bayesian analysis in the form of so-called "neutral" priors as a method for generating interesting examples rather than as anything more meaningful. With this point in mind, I will refer to the situation of equal prior probabilities on the null and alternative as my particular choice as a "reference prior," avoiding the normative connotations of words like "neutral" or "objective."

In a frequentist setting, the null hypothesis that the RTC effects are negligible or zero naturally suggests the two-sided p-value-based significance test for the RTC dummy regression coefficients that is the pre-eminent approach in most legal regression-based studies and that [Donohue 2004] employs 2988 times.[11] Unfortunately, there is a strong argument that this traditional two-sided significance test tends to be a very misleading indicator with respect to the questions of interest. The next section details that argument, and the section following that one discusses the RTC results in light of the argument.

2.2 Bayesian Perspectives on p Values

How does the p value for the null hypothesis that a regression coefficient is zero relate to the Bayesian analysis inherent in equation (3) above, and, in particular to the Bayes factor which indicates how one should shift one's beliefs after observing the data? The p value is a likelihood, related to or equal to $P(y|H_0)$, the probability of observing the data conditional on the null hypothesis being true. The p value is not the posterior probability of

[11] Other issues might call for different approaches. For instance, a frequentist might use a one-sided test if the question is whether or not the crime reduction or increase due to RTC laws is greater than a certain amount. I discuss one-sided as well as two-sided tests in the next subsection.

any hypothesis being true and it is, at most, only one of two components in the Bayes factor. Econometrics and statistics textbooks explicating frequentist techniques consistently and emphatically issue warnings with respect to this point.[12] Nonetheless, many applied papers, including the bulk of the empirical legal literature, take low p values (e.g. $< .05$) as strong evidence that regression coefficients or other parameters are non-zero.

A key question is whether low p values tend to coincide with Bayes factors that indicate a shift in the researcher's probability assessment sharply away from the null hypothesis, or, equivalently, with low posterior probabilities in the face of a 50/50 reference prior. The answer is that "it depends." In the case of two-sided tests, as documented in [Berger & Delampady 1987], it has been known since the middle of the last century that p values are often extremely misleading, tending to greatly underestimate the posterior probability that a point null hypothesis (such as $H_0 = \beta_0$) is true.[13] In contrast, as demonstrated by [Casella & Berger 1987], p values for one-sided tests may be reasonably close to posterior probabilities in some circumstances. The distinction between the efficacy of p values with respect to two-sided and one-sided tests will play a significant role in the assessment of the RTC evidence below.

Before turning to that discussion, it is worth explaining why there often is a large discrepancy between p values and posterior probabilities in the case of two-sided tests.[14] Since tests on regression coefficients are of interest, we will consider an example where the sampling distribution of a standardized coefficient estimator is a t-distribution under the null hypothesis that the coefficient, β, is zero. In particular, let $t = \hat{\beta}/se(\beta)$, where $\hat{\beta}$ is the OLS estimator for a coefficient with true value β, standardized using the

[12]For example, [Freedman, Pisani & Purves 1998], a leading elementary statistics textbook, includes the following warning set off in a box to emphasize its importance:

> The P-value of a test is the chance of getting a big test statistic – assuming the null hypothesis to be right. P is not the chance of the null hypothesis being right. [Freedman, Pisani & Purves 1998, p. 482]

[13]The literature discussing the issues covered in this subsection is vast. For the interested reader, [Berger & Delampady 1987] and [Berger & Sellke 1987] provide an excellent starting point. Both articles are accompanied by the comments of several top statisticians with a wide variety of viewpoints on the subject. In addition, the articles collect the previous literature.

[14][Sellke, Bayarri & Berger 2001] provide a somewhat different but very intuitive and cogent explanation.

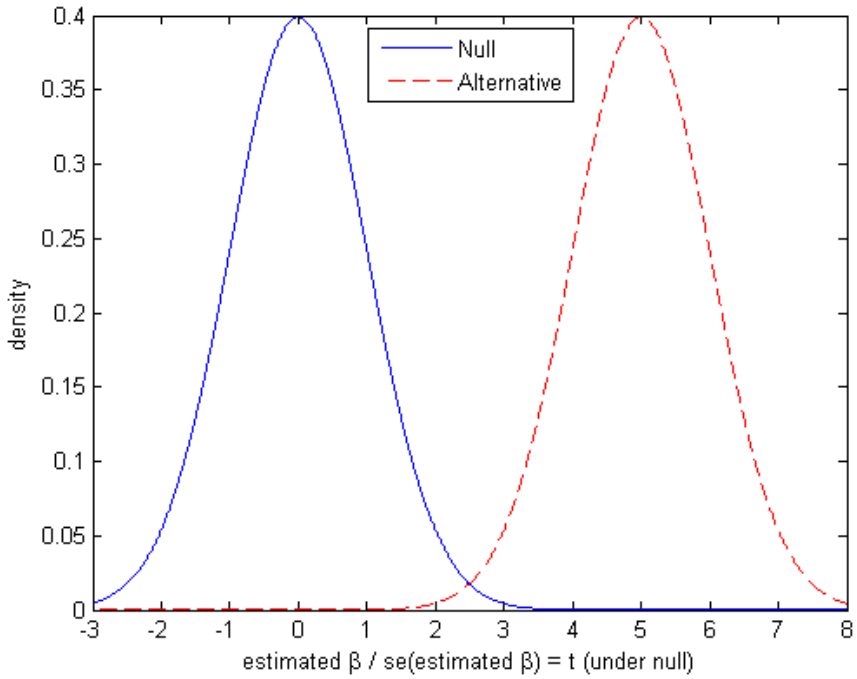

Figure 1: Null and Alternative Sampling Distributions

estimated standard error for $\hat{\beta}$. Under the null hypothesis that $\beta = 0$ and the usual assumption that the disturbances are normally distributed, t will be t-distributed with mean β and $n - k$ degrees of freedom where n is the number of observations and k is the number of regressors. As $n - k$ becomes large, the distribution of t will tend toward the standard normal distribution.

It is useful to begin by considering a very artificial special case: a test of the point null hypothesis that $\beta = 0$ versus the point alternative hypothesis that $\beta = 5$ where the outcome for the estimator, t, is 1.96. The 1.96 value is the upper boundary for a two-sided test at the 5% level in the standard normal limiting case $[(n - k) \rightarrow \infty]$ – i.e., rejection of the null at the 5% level occurs for $|t| \geq 1.96$ when there are a large number of observations compared to the number of regressors. If the alternative is true, then the sampling distribution will be the same, but shifted to center on a mean of $\hat{\beta} = 5$ instead of $\hat{\beta} = 0$. Figure 1 plots the probability density functions of the sampling distributions under the two hypotheses and the standard normal

limiting case. A value as high as 1.96 in absolute value is rare under the null hypothesis that the mean of the estimator is zero. The two-side p value for 1.96 is .05. Nonetheless, under the alternative, the outcome, $t = 1.96$, is even rarer as is evident from a glance at the levels in Figure 1 of the probability density functions for that outcome. That outcome is more than three standard deviations below the mean, $\hat{\beta} = 5$ of the sampling distribution under the alternative hypothesis. As a result, the researcher should conclude that the evidence *increases* the odds that the null hypothesis is true versus the alternative. It is easy to quantify this intuition. The Bayes factor, B_{10}, for the alternative versus the null is simply the ratio of the density functions, $f(t|\beta)$, at the observed value of the estimator:

$$B_{10} = \frac{f(t = 1.96|\beta = 5)}{f(t = 1.96|\beta = 0)} = \frac{1}{12.18}$$

where the subscripts in the notation "B_{10}" indicate that we are comparing the likelihood under the alternative, H_1, in the numerator to the likelihood under the null, H_0, in the denominator. Whatever prior odds the researcher had for the alternative versus the null, these odds would drop by more than a factor of twelve, a very strong shift in favor of the null.

The case of testing a point null hypothesis $H_0 : \beta = 0$ against the general alternative hypothesis $H_1 : \beta \neq 0$ is more complicated. This general alternative hypothesis presumes some weighted combination of point alternatives with different true values, β, rather than a single point alternative such as $H_1 : \beta = 5$ which we just looked at in the example. In effect, instead of the single dashed line sampling distribution curve centered at 5 in Figure 1, we will take a weighted average over cases where the same curve is centered at all possible values other than 0. The weighting function, $g(\beta|H_1)$, is a probability density function for β conditional on the alternative hypothesis being true. In a Bayesian framework, $g(\beta|H_1)$ represents the prior beliefs about the location of β under the alterative. For an observed value $\hat{\beta} = x$ the Bayes factor will be:

$$B_{10}(x) = \frac{\int f(x|\beta)g(\beta|H_1)d\beta}{f(x|\beta = 0)}.$$

The value of B_{10} depends on the choice of the weighting function, $g(\beta|H_1)$, and, as a result, there is no one value of the Bayes factor corresponding to values of x that are just barely significant at the 5% level under the null

hypothesis. This situation is an instance of a general feature of Bayesian analysis that can create complexity and difficulty: The results may be sensitive to the prior. Although this feature has the virtue of allowing us to present results for different prior beliefs, it sometimes leads to a situation where general assertions are not possible.

In the case of p values for two-sided tests, one way to generalize in the face of prior sensitivity is to investigate the question of the maximum value of B_{10} for various classes G of distributions, $g(\beta|H_1)$. Since p values underestimate the strength of the null hypothesis, these maximum values give a general indication of how serious the underestimation is *by using a prior that casts p values in the best possible light within each class G.* The maximum value will never decrease and will tend to increase as broader and broader nested classes, G, are considered. As a result, it is most favorable to the p-value approach to choose the class G to be as broad as possible: the set of all possible distributions. In this case, the maximum value of B_{10} will be attained by putting all of the weight on the alternative sampling distribution with its mode directly above the observed value, $\hat{\beta} = 1.96$ in the example here as pictured in Figure 2.[15] For the standard normal limiting case, the result will be a Bayes factor of 6.83. How does this Bayes factor relate to posterior probabilities of the hypotheses? It depends on the prior. Under our reference prior of 50% probability for the null hypothesis, the researcher's posterior probability of the null hypothesis being true would be 12.78%, much larger than the 5% suggested by the p value. This 12.78% value follows from equation (3), which appears as follows as a function of the Bayes factor:

$$\frac{P(H_1|y)}{P(H_0|y)} = B_{10} \times \frac{P(H_1)}{P(H_0)}. \tag{4}$$

The reference prior assigns equal prior probability to H_1 and H_0 so that the final fraction on the right hand side is equal to 1. Since H_1 and H_0 are mutually exclusive and exhaustive events, $P(H_1|y) = 1 - P(H_0|y)$, and it follows that:

$$P(H_0|y) = \frac{1}{1 + B_{10}}. \tag{5}$$

For $B_{10} = 6.83$ this equation yields $P(H_0|y) = 0.1278$, or 12.78% expressed in percentage terms. The discussion that follows is cast primarily in terms

[15]The probability density function, $g(\beta|H_1)$ is degenerate – a single point mass at $\theta = 1.96$.

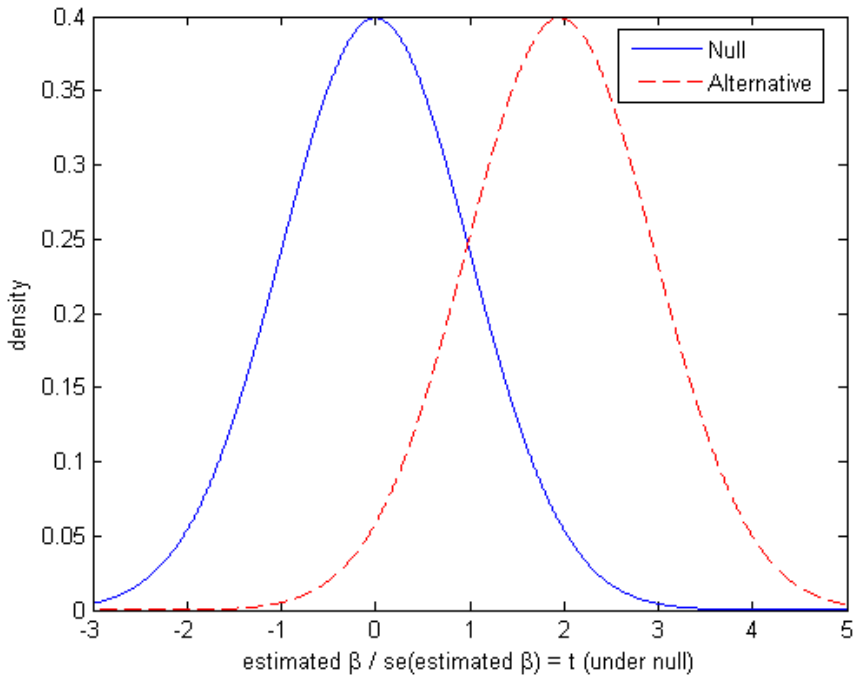

Figure 2: Null and Alternative Sampling Distributions

of posterior probabilities under the 50% reference prior instead of the Bayes factor since these probabilities are directly comparable to p values. Note from the equation above that choosing prior beliefs about the location of β under the alterative hypothesis that *maximizes* B_{10} will *minimize* the posterior probability of the null hypothesis under the 50% reference prior for the hypotheses themselves. The following table presents the Bayes factor and the posterior probability (under the 50% reference prior) for various different degrees of freedom for the t-distribution of the standardized estimator. In each case we set the outcome for t at the value δ such that $Pr(|t| \geq \delta) = .05$, i.e., the positive value for which $p = .05$ in a two-sided significance test.

Bayes Factor and Posterior Probability of Null Hypothesis under 50% Reference Prior $G = \{\text{all distributions}\}$ observed t level corresponds to $p = .05$, two-sided			
degrees of freedom	observed t	Bayes factor	posterior probability
5	2.5127	12.5127	0.0740
20	2.0860	7.9006	0.1124
30	2.0423	7.5213	0.1174
50	2.0086	7.2334	0.1215
100	1.9840	7.0261	0.1246
200	1.9719	6.9251	0.1262
1000	1.9623	6.8456	0.1275
∞ (normal)	1.9600	6.8259	0.1278

It is apparent that, except for the case of very few degrees of freedom, the outcomes are quite close to the limiting standard normal result.[16] In [Donohue 2004] the degrees of freedom typically are around 1000.

The assumption that $G = \{\text{all distributions}\}$ implies a very extreme choice for the weighting function, $g(\beta|H_1)$: The researcher masses all of

[16]The outcomes in the table are easy to calculate. Since both the t distribution and the normal distribution are symmetric and unimodal, the maximum value of the probability density functions occurs at the mode. Each distribution is conditional on some "true" value of β, and the mode is at this true value. The Bayes factor is simply

$$B_{10} = \frac{f(t_{.975}|\beta = t_{.975})}{f(t_{.975}|\beta = 0)}$$

where $f(x|\beta)$ is the probability density function at x for the distribution centered at β, and $t_{.975}$ is the value of t for which 2.5% of the distribution conditional on $\beta = 0$ is in the right tail. The numerator of the fraction is the value of the probability density function

the weight on the value of β that happens to equal the estimate of β that will emerge from the regression. As mentioned above, in Bayesian terms, the weighting function g represents the researcher's prior belief about the distribution of β conditional on the alternative hypothesis, $\beta \neq 0$, being true. Three classes of distributions considered in the literature as interesting candidates for "reasonable" priors in the face of significant uncertainty about β are: all symmetric distributions, all unimodal symmetric distributions and all normal distributions, where in each case the symmetry is around the null point which is $\beta = 0$ in our case. Symmetry about the null point reflects the idea that the actual result is equally likely to be greater or less than that point, a prior belief consistent with using a two-sided test approach in the first place. For the limiting normal case of the regression coefficient example we have been developing and for a reference prior with 50% probability for the null,[17] the minimum probabilities developed in the literature for the null being true are as follows:

---

at its mode. For a t distribution with ν degrees of freedom,

$$f(x|\beta) = \frac{\Gamma\left[(\nu+1)/2\right]}{(\pi\nu)^{1/2}\Gamma(\nu/2)} \times \frac{1}{\left[1 + ((x-\beta)^2/\nu)\right]^{(\nu+1)/2}}.$$

The first fraction on the right hand side is a constant term that will drop out of B_{10}, and when $x = \beta$, the second fraction on the right hand side equals 1. As a result,

$$B_{10} = \left[1 + (t^2_{.975}/\nu)\right]^{(\nu+1)/2}.$$

In the limiting case where $\nu \to \infty$, $f(x|\beta)$ becomes standard normal and,

$$B_{10} = exp(t^2_{.975}/2).$$

As documented in [Berger & Sellke 1987] this result for the standard normal case is derived and discussed in the literature as early as 1963. The result for the more general t distribution probably also is developed and discussed early on somewhere in the literature, but I did not make any attempt to search for appropriate references.

[17] As was true in the example above where $G = \{\text{all distributions}\}$, the results would not be very different if we considered a t distribution with high degrees of freedom, e.g., > 100, instead of the limiting normal case where the degrees of freedom tends to ∞.

Minimum Posterior Probability of Null Hypothesis under 50% Reference Prior and Various Classes of Distributions, G where observed t level corresponds to $p = .05$, two-sided source: [Berger & Sellke 1987]	
G	posterior probability
all distributions	0.128
all symmetric distributions	0.227
all unimodal symmetric distributions	0.290
all normal distributions	0.321

Clearly all of these probabilities are greater than the .05 p value might suggest. These probabilities are *minima* for various classes of distributions, G. Computed probabilities in actual applications may be much higher. For regression coefficients, one should not be surprised to find posterior probabilities in the $30 - 70\%$ range for the null hypothesis being true when the associated two-sided t statistic sits at the .05 threshold of significance. Much more extreme values of the t statistic will be required to bring the posterior probabilities down to around 5%. As shown in [Berger & Sellke 1987] and [Delampady 1989], the situation is similar when instead of a point null hypothesis, e.g., $\beta = 0$, one considers a null hypothesis that the true value of the parameter is close to some point value. The p values for the point null hypothesis are good approximations for the p values that would result from testing a null hypothesis that the true value is in a small interval (e.g., of total width equal to $.2 - .5$ of a sample standard deviation) centered at the point value, and one obtains lower bounds for the posterior probability of the null hypothesis that remain far above the p value.

The situation is quite different for one-sided tests. In the regression coefficient context, interesting one-sided tests might include testing $H_0 : \beta \leq 0$ versus $H_1 : \beta > 0$ or vice versa. [Casella & Berger 1987] derive minima for the posterior probability (based on a reference prior giving 50% probability to the null hypothesis) for one-sided tests and various classes G of weighting distributions, including some classes set forth in the table above for two-sided tests. When G is all symmetric unimodal distributions or all normal distributions, then the p value is the minimum possible value of the posterior probability. The minimum is attained in the interesting case of the "improper prior" that puts equal weight on all values of the parameter.[18] In

[18]This weighting function is "improper" because it is not a probability density function.

other cases, such as $G = \{\text{all symmetric distributions}\}$, the minimum possible posterior probability may be less than the p value. Thus, in the case of one-sided tests, it is *possible* that the appropriate posterior probabilities are close to the p value.[19] Nonetheless, even in the case of one-sided tests, the safest course is to calculate the posterior probabilities rather than rely on the possible salience of the p value.

In section 3 we will calculate posterior probabilities for various models and coefficient values including some of the models and RTC coefficients estimated in [Donohue 2004]. In the rest of this subsection and the next subsection, we will focus on a particular approximation to the Bayes factor appropriate to the two-sided test of the point null hypothesis that a regression coefficient is zero. This approximation is valid for large samples and needs only the t statistic for the coefficient and the sample size as inputs. Using the approximation will allow us to assess the 2988 two-sided tests developed in [Donohue 2004], and this assessment will show how misleading the p value approach is for those tests. The approximation also is useful in other contexts.[20]

The basis for the approximation is the insight that in a Bayesian framework, assessing relative probabilities for a null hypothesis (H_0) that a coefficient is zero in a linear regression versus the alternative hypothesis (H_1) that it is not is equivalent to comparing two models: M_0, a linear regression without the corresponding variable; and M_1, the "full model" consisting of a linear regression that includes the corresponding variable. The Bayes factor

---

The constant function for any constant greater than zero does not have a finite integral over the real line. "Improper priors" are important in Bayesian analysis because they are one way to represent "complete ignorance" about the value of a parameter. In many cases, proper posterior densities emerge from analysis that begins with improper priors because the infinite integral problem washes out after the improper prior is multiplied by the likelihood before integrating. Improper priors, however, tend to cause other problems. In particular, as discussed in later parts of the article, model comparison is sometimes impossible under improper priors.

[19]There is a connection between this one-sided test result and the results for point nulls. [Delampady 1989] shows that when $G = \{\text{unimodal symmetric distributions}\}$, the lower bounds on the posterior probability for an interval null centered at a point value decline gradually from the high values indicated in the table above to the p value as the interval expands. Delampady notes that this phenomenon connects the point null result to the result for one-sided test. As the interval expands, the two-sided test with a value outside of the interval becomes closer and closer to a suitably framed one-sided test.

[20]Since it is easy to calculate, it often provides an immediate cogent workshop comment when the presenting authors are making claims based on p values.

for this model comparison is the same as the Bayes factor for comparing the hypotheses:

$$B_{10} = \frac{P(y|H_1)}{P(y|H_0)} = \frac{P(y|M_1)}{P(y|M_0)}.$$

The expression $P(y|M_i)$ is the marginal likelihood for model i.

Bayes factors may be approximated using the Bayesian Information Criterion, "BIC" for short:[21]

$$2log(B_{10}) = 2log\left[\frac{P(y|\hat{\theta}_1, M_1)}{P(y|\hat{\theta}_0, M_0)}\right] - (k_1 - k_0)log(n)$$

where: $P(y|\theta_i, M_i)$ is the marginal likelihood for model M_i; $\hat{\theta}_i$ is the value of the vector of parameters, θ_i, that maximizes $P(y|\theta_i, M_i)$; k_i is the number of variables in model M_i, and n is the sample size. In our case, $k_1 - k_0 = 1$ since M_0 is simply M_1 with one variable omitted. Furthermore, as pointed out by [Kass & Raftery (1995)] and others, for large n, twice the log of the ratio of (maximum) marginal likelihoods in this situation is approximately equal to the square of the t-statistic for the omitted variable in the regression where it is included. Thus, the large sample approximation boils down to:[22]

$$2log(B_{10}) = t^2 - log(n). \tag{6}$$

[21] [Schwarz (1978)] originated BIC, and indeed, an equivalent quantity, minus one-half of BIC, is called "the Schwarz criterion."

[22] This approximation appears to be "prior free" since it does not depend explicitly on a particular prior or on an assumed distribution of the parameter (e.g., β in the example above) under the alternative hypothesis. This appearance is at least in part an illusion. For large n, where n is the number of observations, approximation of Bayes Factors by BIC will be appropriate only under certain types of priors used in Bayesian estimation of the applicable linear regressions. So-called "unit information" priors fall into this category. A well known example of a unit information prior in a regression context is to specify that the regression coefficients are distributed $N(\underline{\beta}, \frac{1}{n}\sigma^2(X'X)^{-1})$ where $\underline{\beta}$ is a vector of prior means, X is the matrix of independent variable data, and the error terms are assumed to be distributed $N(0, \sigma^2 I)$ where σ^2 is unknown, and I is the identity matrix. If we also assume that the prior distribution for $h = (\sigma^2)^{-1}$ is a gamma distribution, we will be using a "natural conjugate prior" that also is a "g-prior," both of which are discussed in section 3 below. The posterior mean and precision under this prior will be weighted averages of the prior and ordinary least squares results with weights $\frac{1}{n+1}$ and $\frac{n}{n+1}$ respectively. For details, see [Koop 2003, chapter 3]. In this formulation, the prior effectively will have weight equal to one data point in the posterior and, as a result, will not have very much impact. Unit information priors are very popular since they are a way of building in weak prior information without using improper priors. We will use these priors in section 3 below but also will consider issues of prior sensitivity. [Fernandez, Ley & Steel 2001] provide a good dis-

The sample sizes in the estimated models in [Donohue 2004] are around 1000. The following table indicates the Bayes factor, posterior probability, and p values that follow for various values of the t statistic using the approximation just developed:

---

cussion of various priors and their asymptotic consequences. [Kass & Wasserman (1995)] discuss the connection between unit information priors and BIC more generally, i.e., not restricted to natural conjugate priors in a regression context.

Despite the fact that BIC is not the final asymptotic form of the Bayes factor for all reasonable priors, there is a sense in which BIC gives the correct asymptotic answer quite broadly in the hypothesis testing framework. In section 3 we discuss the "consistency" properties of various processes and approximations in a model comparison setting. A consistent process or approximation will assign probability 1 to the "correct model" or the "best model" asymptotically. BIC is one such approximation, and that result is no accident. As shown in [O'Hagan & Forster 2004] and elsewhere, the general form of equation (6) is equal to the BIC version in the text plus a term that is $O(1)$. The first two terms in equation (6) are $O(n)$ and $O(log(n))$ respectively if the null hypothesis is false. If the null hypothesis is true then the first term is $O(1)$. As a consequence, in the regression coefficient case, the Bayes factor will tend to 1 or 0 asymptotically depending on the relative strength of the hypotheses. The same result will apply for any "consistent" approximation. Some of these differ from BIC by an $O(1)$ term. Others have a second term that is not $O(log(n))$ but asymptotically dominates any $O(1)$ term while being asymptotically dominated by the first term. The $n \approx 1000$ observations in the [Donohue 2004] regressions are not enough for the BIC approximation to be the whole story. $O(1)$ terms will still matter. Under a unit information prior, [Kass & Wasserman (1995)] show that the third term is $O(n^{-1/2})$ instead of $O(1)$.

We defer a discussion of methods for dealing with prior sensitivity until section 3. BIC provides an asymptotically valid method of approximating the Bayes factor in the frequentist hypothesis testing framework. It is a very nice approximation because one can compute it based on two simple numbers: t values and the sample size. Typically, these numbers are included in published reports. However, BIC is not a perfect substitute for computing actual Bayes factors. A big part of the problem is prior sensitivity.

BIC Approximation for $n = 1000$ posterior probability of null under 50% Reference Prior			
t statistic (abs. value)	Bayes factor	p value	posterior probability
0.9572	0.05	0.33870	0.9524
1.5174	0.1	0.12950	0.9091
1.9206	0.2	0.055060	0.8333
2.6283	1	0.00871	0.5000
3.1822	5	0.00151	0.1667
3.3619	9	0.00080	0.1000
3.5772	19	0.00036	0.0500
4.0122	99	0.00006	0.0100

The numbers in the table should be striking if not shocking for readers who are accustomed to p value analysis. A Bayes factor of 1 indicates that, based on the evidence, there should be no change in one's prior beliefs about the truth of the null hypothesis. The posterior odds should be equal to whatever the prior odds were. Under the approximation, a Bayes factor of 1 coincides with a t statistic equal to 2.6283, a value at which a frequentist would reject the null hypothesis at the .008 level! Starting with equal prior odds on the null and the alternative, a Bayes factor of 19 is required to drive the posterior probability of the null down to .05. Under the approximation, this level of posterior probability is reached only when t is 3.5772 in absolute value. If one takes this "equal priors" view as appropriate and requires the evidence to be strong enough that the null has only 5% posterior probability, then this level of 3.5772 is the critical value for $|t|$ rather than the usual benchmark of 1.96 in large samples and "around 2" more generally.

The table suggests that use of two-sided p values as indicia is likely to be very misleading, especially in situations where researchers rely on "statistical significance" based on t statistics that range from 2-3 in absolute value. In the next subsection, we will use the approximation to assess the p-value based conclusions in [Donohue 2004]. Before doing so, it is worth making some general points.

If one is interested in making actual probability assessments of the truth of the null hypothesis in a two-sided test situation, the usual benchmarks for critical levels of $|t|$ are not trustworthy. Instead of the usual benchmarks of around $1.65, 2$, and 2.3 for the $10\%, 5\%$, and 1% levels respectively, appropriate benchmarks might be in the $3 - 4.5$ range. Consider the following

table which uses the BIC approximation to compute "critical" levels of $|t|$ for various Bayes factors and sample sizes:

"Critical" t Values under BIC Approximation various sample sizes, n, and Bayes factors, "BF"				
n	BF = 1	BF = 9	BF = 19	BF = 99
200	2.3018	3.1133	3.3447	3.8064
500	2.4929	3.2572	3.4790	3.9249
1000	2.6283	3.3619	3.5772	4.0122
10000	3.0349	3.6885	3.8858	4.2896
100000	3.3931	3.9884	4.1715	4.5501

Assuming equal prior probabilities for the hypotheses, the Bayes factors of $9, 19$, and 99 correspond to posterior probabilities of $10\%, 5\%$, and 1% for the null being true, paralleling the corresponding p values. The idea of looking for t-statistics greater than three in absolute value corresponds to the observation by [Jeffreys 1980, p. 452] about a rule of thumb used by astronomers:[23]

> ...the rough rule long known to astronomers, i.e. that differences up to twice the standard error usually disappear when more or better observations become available, and that those of three or more times usually persist.

Of course, in a Bayesian context it always is best simply to compute any Bayes factors of interest rather than relying on an approximation or a rule of thumb. Furthermore, as discussed in [Sellke, Bayarri & Berger 2001], there are more general methods, both Bayesian and frequentist, of "calibrating" p values to avoid misleading interpretations than revised rules of thumb.

2.3 The RTC Results Revisited

We use the approximation developed in the previous subsection to assess the results in [Donohue 2004]. Thankfully(!), we will not revisit all 2988 instances of the use of two-sided p values in the paper individually. Instead, we will consider the overall impact on three separate groups of results: (i) the results under the "ADS" approach favored by Donohue; (ii) the aggregate

[23] [Berger & Sellke 1987] begin with this quotation and develop a nice qualitative example involving the "astronomers."

results for the "standard panel data" models; (iii) the state specific results for those models.

In assessing results, it is important to consider different viewpoints. A great strength of Bayesian approaches is that one can speak from the same data to readers with different prior beliefs, indicating what changes in belief would be warranted given the evidence. As mentioned above, typical Bayesian output would be posterior distributions for the parameters of interest based on various prior distributions for the same parameters. In later sections of the paper we will report results in terms of various features of the relevant posterior distributions. This subsection will focus on reinterpreting the frequentist results in [Donohue 2004] based on the theoretical discussion in the previous subsection. We will leave Bayesian estimation to later parts of the paper. The arguments and calculations in the reinterpretation should be of general interest. They are applicable to scrutinizing the results from a wide variety of papers that employ frequentist linear regression analysis.

[Donohue 2004] takes the typical tack of relying on two-sided rather than one-sided p-value analysis to evaluate linear regression output.[24] The previous subsection indicates that two-sided p values may be very misleading: the p value typically will overstate the evidence against the null hypothesis that a regression coefficient is zero or negligible. Implicit in the two-sided analysis is a desire to test this null hypothesis. We will use the approximation developed in the previous subsection to estimate Bayes factors indicating how one's views about this null hypothesis should shift based on the evidence. The picture will be very different from the one that emerges from p value analysis.

Some hypotheses may be of interest that are associated with one-sided rather than two-sided tests, such as whether the coefficients on the critical RTC dummies are less than zero (indicating a crime reduction resulting from RTC laws). Based on some of the theoretical results for one-sided tests dis-

[24]The use of two-sided rather than one-sided tests has a strong frequentist rationale. Conditional on an estimator ending up being positive or negative, a one-sided test will be statistically significant at lower values of the estimator. For example, in the normal limiting case, $t \geq 1.96$ is required for statistical significance at the .05 level for a two-sided test, but a one-sided test only requires that $t \geq 1.645$. Using a two-sided test removes the temptation to peek at the sign of the results and then use a one-sided test with ensuing greater likelihood that the variable is "significant." In a Bayesian framework, this problem disappears. The output is a posterior probability distribution rather than some test result or p value, and researchers may employ alternative priors in the face of divergent prior beliefs among "consumers" of the research.

cussed in the previous subsections, we can interpret some of the frequentist output in a Bayesian manner. In particular, one-sided p values will be equivalent to Bayesian posterior probabilities if (i) the prior puts 50% probability on the coefficient being above or below the point that is being tested and, (ii) the prior under the alternative hypothesis is improper, placing equal weight on all possible values of the parameter. We will reinterpret the evidence in [Donohue 2004] based on these priors.[25]

We begin with the results from ADS-type models, the most favored approach in [Donohue 2004]. These models have some very interesting and cogent features. They are explicit treatment models, evaluating the impact of adoption of RTC laws in a given state by examining effects in that state within a "treatment period" window of time surrounding adoption. One or two years of data from adopting states are dropped during an "adjustment period" centered on the effective date of the RTC law. There are eight alternative specifications: the treatment period may be five or seven years, the adjustment period may be one or two years, and two different weighting schemes are used to adjust for state population – raw population and population share. Since not much will turn on differences between specifications, we simply refer to the specifications as " models 1-8."[26] The crucial dummy variable is called "post" and is 1 during the portion of the treatment period following adoption and 0 otherwise.[27] The model also includes all of the covariates in the Modified Lott panel regression specification, and is estimated for all nine crime categories.

The t statistics for the 72 instances of "post" are as follows:

[25]It would be easy to back out Bayes factors from the posterior probabilities. Then we would not need to rely on assumption (i). However, the direct use of posterior probabilities based on (i) makes the implications very tangible.

[26]Models 1-4 use the raw population weights, and models 5-8 use population share. The odd numbered models drop only the year of adoption while the even numbered models drop that year and the following year. Models 1-2 and 5-6 use five-year treatment periods while the other models use seven-year treatment periods.

[27]A control variable called "treatment" is 1 during the entire treatment period and 0 otherwise. Another dummy variable "postpost" is 1 following the end of the treatment period and 0 otherwise.

ADS Specifications: t Statistics for "Post" eight specifications; nine crime categories									
model	violent	murder	rape	robbery	assault	property	burglary	larceny	auto
1	-0.19	1.06	-1.49	1.59	-1.16	1.89	0.61	2.18	2.49
2	-0.51	0.68	-1.58	1.47	-1.42	1.70	0.39	2.09	2.06
3	-0.53	0.81	-1.28	1.15	-1.27	1.82	0.10	2.39	2.55
4	-0.78	0.54	-1.35	0.86	-1.39	1.55	-0.17	2.25	2.01
5	-0.14	1.11	-1.53	1.60	-1.15	1.94	0.64	2.24	2.52
6	-0.38	0.80	-1.60	1.49	-1.34	1.77	0.42	2.21	2.07
7	-0.49	0.85	-1.30	1.16	-1.28	1.86	0.13	2.44	2.56
8	-0.70	0.62	-1.38	0.89	-1.36	1.60	-0.13	2.33	2.00

Based on these t-statistics [Donohue 2004, p. 637] states that "if one preferred this [ADS model] approach ... one would essentially discard the previous suggestion [from the panel data models] that the RTC laws reduce rape, and conclude that RTC laws appear to have no effect on violent crime and increase property crime (except for burglary) during the first two or three years following adoption." In a sophisticated discussion, the article goes on to discuss the possible origin or meaning of these results, for instance, what the causal mechanism might be behind the pattern. Much of this discussion, however, may be superfluous. Using the large sample approximation discussed in the previous subsection, the (approximated) Bayes factors for the alternative hypothesis (coefficient $\neq 0$) versus the null (coefficient $= 0$) are as follows:

ADS Specifications: Approximate Bayes Factors alternative hypothesis ("Post" coefficient is non-zero) versus null (coefficient is zero) eight specifications; nine crime categories									
model	violent	murder	rape	robbery	assault	property	burglary	larceny	auto
1	0.03	0.05	0.09	0.11	0.06	0.18	0.04	0.32	0.66
2	0.03	0.04	0.10	0.09	0.08	0.13	0.03	0.27	0.25
3	0.03	0.04	0.07	0.06	0.07	0.16	0.03	0.51	0.75
4	0.04	0.03	0.07	0.04	0.08	0.10	0.03	0.37	0.23
5	0.03	0.05	0.10	0.11	0.06	0.19	0.04	0.36	0.71
6	0.03	0.04	0.11	0.09	0.07	0.14	0.03	0.34	0.26
7	0.03	0.04	0.07	0.06	0.07	0.17	0.03	0.57	0.77
8	0.04	0.04	0.08	0.04	0.08	0.11	0.03	0.45	0.22

Strikingly, all 72 of the Bayes factors are less than one, with largest being 0.77. Thus, *whatever prior beliefs one started with, one would shift one views toward the null that the RTC laws have zero or negligible effect for all nine crime categories and all eight specifications of the model.* Given the amount of data, over 1000 observations, none of the t statistics is large enough in absolute value to shift one's beliefs against the null of zero or negligible effects. For many of the outcomes the small Bayes factors dictate a sharp shift toward the null, but even in the case of larceny and auto theft, the crime categories with the highest Bayes factors, the shift is decidedly toward the null.

It is easy to extract the outcomes for one-sided tests of coefficient values from frequentist linear regression estimates, and we have seen that the associated p values have a probability interpretation under particular prior assumptions. For the ADS specifications in [Donohue 2004], the one-sided p values against the null hypothesis that the key coefficient (on "Post") is greater than or equal to zero are as follows:

ADS Specifications: One-sided p Values null hypothesis ("Post" coefficient ≥ 0) versus alternative (coefficient < 0) eight specifications; nine crime categories									
model	violent	murder	rape	robbery	assault	property	burglary	larceny	auto
1	0.42	0.86	0.07	0.94	0.12	0.97	0.73	0.99	0.99
2	0.31	0.75	0.06	0.93	0.08	0.95	0.65	0.98	0.98
3	0.30	0.79	0.10	0.87	0.10	0.97	0.54	0.99	0.99
4	0.22	0.70	0.09	0.81	0.08	0.94	0.43	0.99	0.98
5	0.44	0.87	0.06	0.94	0.12	0.97	0.74	0.99	0.99
6	0.35	0.79	0.05	0.93	0.09	0.96	0.66	0.99	0.98
7	0.31	0.80	0.10	0.88	0.10	0.97	0.55	0.99	0.99
8	0.24	0.73	0.08	0.81	0.09	0.95	0.45	0.99	0.98

The symmetry of the sampling distribution for the coefficients means that one minus the numbers in the table will be the p values for testing in the other direction, i.e., versus a null hypothesis that the coefficients are less than zero. Interpreting these values as probabilities brings back into the picture the possibility that rape is deterred by RTC laws. The results also are suggestive on that score for assault, while for the general category of property crime and (more weakly) the specific category of robbery, the possibility of increases in crime induced by the RTC laws looms large.

We can draw several implications from this attempt to extract hypothesis probability information from the frequentist ADS regressions. First, the standard two-sided p value analysis employed in [Donohue 2004] is quite misleading. The goal that seems implicit from choosing the two-sided point null approach is to evaluate the hypothesis that the impact of crime on the RTC laws is zero or negligible. Under the approximation, the data clearly would strengthen one's degree of belief in that hypothesis *for all the crime categories and all eight specifications*. This result is very much in consonance with one of the themes in [Donohue 2004]: that the evidence for an impact (positive or negative) of the RTC laws is shaky. However, the article understates the degree of consistency of that theme with the data because the p value approach hides the strength of the hypothesis that the RTC laws have little or no effect. Second, suppose the set of goals is different: merely gauging the sign of the effects, however small the magnitude – or more generally, deciding on the likelihood that the effects exceed or are below a certain level. Then one is in the domain of one-sided tests, and we have seen that the p-value based frequentist version of these tests has a posterior hypothesis probability interpretation under at least some sets of prior beliefs. Viewed in that light, the evidence from the ADS results in [Donohue 2004] suggests a shift in view toward distinct positive or negative effects for five of the nine crime categories versus the two categories that had statistically significant effects under p value analysis. With respect to one of the three added categories, rape, the paper had argued that the lack of statistical significance under the ADS specification belied the apparent significance under some of the other models.

The picture is similar for the 108 aggregate estimates from the twelve panel data models covering nine crime categories each. [Donohue 2004] reports 33 t-statistics significant at the 5% level for the regression coefficient on the dummy variable indicating the presence or absence of a RTC law. Using the BIC approximation, only four of the coefficients end up with a Bayes factor of 19 or higher, and 14 of the 33 "significant" coefficients are associated with a Bayes factor less than one, indicating that one should shift one's beliefs in favor of the null hypothesis that the RTC laws have zero effect. The t-statistics and Bayes factors are as follows:

Panel Data Specifications: t Statistics for RTC Dummy twelve specifications; nine crime categories									
model	violent	murder	rape	robbery	assault	property	burglary	larceny	auto
DL d	1.19	-3.16	-1.95	1.74	0.84	3.42	2.13	2.91	3.00
DL t	0.27	-1.21	-1.84	0.38	0.40	1.78	0.80	1.33	2.18
DL s	1.57	1.10	0.44	0.20	1.53	1.46	0.36	1.25	1.45
ML d	-1.96	-0.84	-3.71	-2.27	0.11	1.98	3.44	2.37	3.31
ML t	0.15	1.34	-2.55	-1.23	2.29	2.90	2.45	1.15	3.41
ML s	-1.53	-2.51	-2.79	-1.45	-2.69	-0.37	0.39	-1.88	-0.59
SP d	1.17	-0.65	0.08	1.71	0.28	3.30	1.66	3.30	3.11
SP t	0.36	0.09	-1.30	0.91	-0.27	1.07	-0.06	1.34	1.04
SP s	0.24	-0.02	0.59	-0.39	0.05	0.54	-0.13	0.41	0.84
ZH d	1.15	-1.12	-0.53	-0.50	2.64	6.48	2.19	7.01	6.62
ZH t	0.47	0.54	-2.19	-0.76	1.86	1.77	0.50	1.99	2.32
ZH s	-0.76	-0.48	0.98	-0.50	-2.93	-0.42	-2.74	0.01	-0.58
models: ML = Modified Lott; DL = Donohue/Levitt; SP = Spelman; ZH = Zheng RTC specifications: d = dummy alone; t = with state trends; s = spline;									

Panel Data Specifications: Approximate Bayes Factors alternative hypothesis ("Post" coefficient is non-zero) versus null (coefficient is zero) twelve specifications; nine crime categories									
model	violent	murder	rape	robbery	assault	property	burglary	larceny	auto
DL d	0.06	4.31	0.20	0.13	0.04	10.06	0.28	2.04	2.65
DL t	0.03	0.06	0.16	0.03	0.03	0.14	0.04	0.07	0.31
DL s	0.10	0.05	0.03	0.03	0.09	0.08	0.03	0.06	0.08
ML d	0.20	0.04	28.90	0.39	0.03	0.21	10.89	0.48	7.09
ML t	0.03	0.07	0.76	0.06	0.40	1.95	0.59	0.06	9.64
ML s	0.09	0.68	1.45	0.08	1.09	0.03	0.03	0.17	0.03
SP d	0.06	0.04	0.03	0.13	0.03	6.77	0.12	6.81	3.64
SP t	0.03	0.03	0.07	0.04	0.03	0.05	0.03	0.07	0.05
SP s	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.04
ZH d	0.06	0.05	0.03	0.03	0.95	3.89E+7	0.32	1.40E+9	9.71E+8
ZH t	0.03	0.03	0.33	0.04	0.17	0.14	0.03	0.21	0.43
ZH s	0.04	0.03	0.05	0.03	2.17	0.03	1.25	0.03	0.03
models: ML = Modified Lott; DL = Donohue/Levitt; SP = Spelman; ZH = Zheng RTC specifications: d = dummy alone; t = with state trends; s = spline;									

Out of 108 Bayes factors, 89 are less than one, indicating that one should

shift one's beliefs toward the null hypothesis of no effect. Only four are greater than 19, and three of these (all very large!) are concentrated under one model. Only seven are greater than 9. Again, we have a general picture of not very much strength against the null hypothesis.

For one-sided tests, the results are similar to the ADS specification outcomes in two respects. First, many of the one-sided tests suggest evidence in favor of a particular sign if one is willing to adopt the prior discussed above. Out of 108 tests, 43 are significant at the 5% level and 24 at the 1% level. Second, the pattern across crimes bears some similarity to the pattern for the ADS specification: Most of the tests that suggest a positive coefficient (crime enhanced by RTC laws) are concentrated among the property crimes while the tests that suggest a negative coefficient are concentrated among the violent crimes, especially rape. However, the results are more mixed for each individual crime with some specifications resulting in quite different results, even for larceny and auto theft. Part 3 will report posterior distribution characteristics while at the same time attempting to unpack some of the specification issues. As a result, there is no reason to go into more detail here or to report the one-sided test results in a table.

There are 2808 state specific p value tests of RTC law coefficients in [Donohue 2004]. The purpose of the state specific estimates is not to argue that RTC laws may be appropriate in some states but not others but as an alternative way to gauge overall effects. Thus, for each crime category and panel data specification, [Donohue 2004] reports two comparisons: positive coefficients versus negative coefficients and positive statistically significant (5% level) coefficients versus negative statistically significant coefficients. Applying our approximation creates the same kind of culling as in the case of the ADS and aggregate panel data specifications. [Donohue 2004] finds 1116 out of the 2808 coefficients statistically significant. Out of the 1116, 385 end up having a Bayes factor less than one, and 407 have Bayes factors greater than 19. That 407 is about a seventh of the total, a much larger proportion than in the case of the ADS specifications (zero) or the aggregate panel data specifications (about four percent). Rather than go into detail about how this might shift the picture or engage in a discussion about one-sided results, the next subsection presents an alternative Bayesian approach for getting at the same issues.

The specific results across all of the specifications very much track the conceptual discussion in the previous subsection. In particular, p values do not have an interpretation as a probability that any hypothesis of interest

is true or false, and letting the p value play a probability role, even unconsciously, can be a serious mistake. On this basis, some believe that p values should be avoided in two-sided tests of a point null hypothesis such as the typical test for the "statistical significance" of regression coefficients. For example, in a section of their paper entitled "What Should be Done?," [Berger & Delampady 1987] assert that:

> First and foremost, when testing precise hypothesis, formal use of P-values should be abandoned. Almost anything will give a better indication of the evidence provided by the data against H_0.

In response to this position, some prominent statisticians see a role for p values, but only as a general indication or warning concerning a parameter that suggests a need for further analysis, not as precise or even approximate evidentiary weight pro or con with respect to a hypothesis.[28] The problem is that many empiricists, including legal scholars, have developed the bad habit of attributing more weight to p values than they deserve. One commonly sees "conclusions" about evidentiary salience based on the statistical significance of regression coefficients at the 5% or even 10% level in two-sided tests against a null that they are zero. Berger & Sellke's 1987 assessment seems apt even today:

> ... there might be cries of outrage to the effect that $p = .05$ was never meant to provide an absolute measure of evidence against H_0 and any

[28] Responding to [Berger & Sellke 1987], D.R. Cox asserts that:

> ...it [a conventional significance test] is to serve as a general warning that something is wrong (or not), not as explicit support for a particular alternative explanation. Thus, such tests have a very limited aim and often one should be doing something more strongly focused, but that does not make the P-value misleading or useless. [Cox 1987]

Commenting on the same article, Arnold Zellner notes that:

> ... I have the impression from my own experience, from Jeffrey's report of what astronomers do and from talking with others that many tend not to reject a null hypothesis when $t = 1.96$, but view the matter as a situation in which more information is needed.[Zellner 1987]

And, I.J. Good, responding to similar pessimism about p values in [Berger & Delampady 1987] states that

> One result, I hope, will be that the conventional P value of approximately .05, when testing a simple statistical hypothesis H_0 will be correctly interpreted: *not as a good reason for rejecting H_0 but as a reason for obtaining more evidence provided that the original experiment was worth doing in the first place.* [Good 1987, emphasis in the original]

> such interpretation is erroneous. The trouble with this view is that, like it or not, people do hypothesis testing to obtain evidence as to whether or not the hypotheses are true, and it is hard to fault the vast majority of nonspecialists for assuming that, if $p = .05$, then H_0 is very likely wrong. This is especially so since we know of no elementary textbooks that teach that $p = .05$ (for a point null) really means that there is at best weak evidence against H_0. Indeed, most nonspecialists interpret p precisely as $Pr(H_0|x)$ ... which only compounds the problem. [Berger & Sellke 1987, p. 114, citation omitted]

Even amongst "specialists," it is easy to find examples where legal scholars draw conclusions based on p value outcomes when they should not. This observation leads to a question: Why aren't legal scholars more like Jeffreys' astronomers, holding p value results lightly and looking for t statistics north of three rather than two in absolute value as a guidepost? This question is not a trivial one and applies to empirical work in many fields other than law. It is likely that there are some potentially complex psychological and/or sociological factors at work. I limit consideration here to one simple observation. As noted in Jeffreys' comments, it is the frailty of t values around two in the face of "more or better observations" that banishes any tendency to attach strong significance to t values of that magnitude. This frailty will be quite evident in a field where it is easy to generate more data through additional experiments or measurements. In law, however, researchers typically are limited to observational evidence that cannot be expanded. E.g., we cannot rerun history repeatedly with states making various different decisions about whether or not to enact RTC laws. As a result, in many empirical legal inquiries, it is unlikely that drawing excessively strong conclusions from p values around .05 will come back to bite the researcher. The bad habit of doing so can persist more easily than in fields where experimentation is more prominent or where evidence accumulates more rapidly.[29]

[29]Of course, it is possible that some endogenous elements are involved. Scholars may sort into fields according to the ease with which theories and ideas may be falsified based on evidence. Falsification tends to be particularly difficult in many areas of legal academic inquiry due to the complexity of the phenomena studied and the inability to generate trenchant data when needed. Perhaps the freedom to theorize with only very mild discipline from empirical evidence is appealing to many of the people who become legal academics.

2.4 A Hierarchical Approach

As mentioned in the previous subsection, [Donohue 2004] uses the state specific versions of the panel data models to gauge whether there was an overall effect of the RTC laws rather than to test for different effects in different states. These versions focus on the RTC dummies for the 26 states that adopted RTC laws during the sample period. The idea is to look at the distribution of the 26 coefficients for each specification and crime combination (108 combinations in all – 12 specifications and 9 crimes): There should be more negative coefficients than positive coefficients if the right-to-carry laws deter crime. [Donohue 2004] uses two distinct methods for counting negatives and positives: one comparing all positives to all negatives for each model and the other comparing only statistically significant positives to statistically significant negatives.

There is a major problem with this approach. Suppose that out of 26 states we have 8 with statistically significant negative coefficients on the RTC dummy and 12 with statistically significant positive coefficients. In the spirit of [Donohue 2004] we would interpret this result as evidence that RTC laws do not deter crime in general. An alternative interpretation would be to conclude (leaving aside issues concerning the validity of p-values) that it is likely that the right-to-carry laws have a deterrent effect in the eight states with statistically significant negative coefficients, but probably not in the others. This alternative interpretation is consistent with the model structure since estimating the RTC effects via 26 separate independent variables presumes independent effects in the 26 states – leaving no scope for the possibility that there is some common element in the effects across states. The other extreme, having a single RTC dummy for all states, presumes that there is only a common effect. Neither of these models is consistent with the idea, implicit in [Donohue 2004], that the individual state RTC effects reflect draws from some common distribution illuminated by comparing the number of negative coefficients to the number of positive ones.

A Bayesian hierarchical model is an ideal vehicle for modeling individual coefficients or other parameters as draws from a common distribution, and I construct a simple one here. As a first step, assume that the 26 RTC dummies are drawn from a common normal distribution with mean μ_β and variance V_β^2. These two numbers are "hyperparameters" since they characterize the distribution of other parameters in the models, and they are interesting in their own right. We will assign a prior distribution to each of

them, bring them to the data and emerge with posterior distributions. The V_β^2 hyperparameter indicates where we are between the extremes of a single common effect ($V_\beta^2 = 0$) and 26 independent effects ($V_\beta^2 \to \infty$). As is discussed cogently in [Gelman, et.al. 2004, pp. 131-133] a great advantage of the Bayesian hierarchical approach is that it allows us to consider a continuous range of cases with stronger and weaker degrees of commonality among a set of parameters. We do not have to make an either-or choice between assuming complete commonality or total independence, and we can gauge how the data should affect our beliefs about the degree of commonality that is present.

Consider the 108 panel data regressions in [Donohue 2004]. In each case, we can write the regression in the following form:

$$y_i = \tilde{Z}_i\tilde{\beta}_i + \tilde{X}_i\gamma + \epsilon_i \tag{7}$$

for $i = 1, 2, ..., N$ blocks of observations covering N different states each over a common set of T time periods. The variable $\tilde{Z}_i$ is the RTC dummy for state i.[30] All of the other independent variables including state fixed effects dummies are included in $\tilde{X}$ with a vector of coefficients γ that are common across states.[31] Initially, we will assume dogmatically that the errors are homoscedastic with zero mean and a shared precision, h.

In the Bayesian approach, we need to specify priors and compute or simulate posterior distributions for all of the parameters: $\beta_i, \gamma, h, \mu_\beta,$ and V_β^2. Under the priors we will use, the joint posterior distribution for the parameters is not a known distribution.[32] Instead, we simulate the relevant posterior

[30]This variable is a mixture of zeros and ones for states that changed their RTC laws during the T periods and is all zeros for the other states. The latter all-zero specification avoids multicollinearity with the state fixed effects dummies. Effectively the regression includes RTC dummies only for the 26 states that changed their RTC laws during the sample period.

[31]Readers may recognize that this model is similar to the more general "random coefficients" model where all the coefficients vary across the cross sectional units in the panel. Another related variant is the "random effects" model which specifies the fixed effects as arising from a distribution while restricting all other coefficients to be the same across all entities. Some of these variants have non-Bayesian implementations, but the Bayesian approaches are particularly transparent, flexible and easy to implement.

[32]In particular, we employ an independent Normal-Gamma prior for γ and h which precludes analytic solutions for these parameters or the other ones. This approach is described in many books. For a reader who has a background in frequentist econometrics, a good starting point is [Koop 2003, ch. 4]. I do not develop the particular hierarchical

distributions using iterative Markov Chain Monte Carlo techniques that are conventional in Bayesian applications. These techniques involve drawing a large sample from the posterior distribution for the parameters and then using this sample to compute desired quantities such as the posterior mean of regression coefficients. Here, as in many cases, although the joint posterior distribution for the parameters is not a known distribution, the distributions of individual parameters conditional on holding the other parameters fixed are known distributions. We sample the posterior by sequentially drawing from these conditional distributions. We choose "non-informative" priors with respect to μ_β and V_β^2. This choice means that our prior beliefs have no weight in forming the posterior for these hyperparameters. The result is a "data driven" posterior characterized by quantities similar or identical to frequentist estimators.[33]

The conditional posterior distributions for the RTC dummies, β_i, are normal with separate means, $\bar{\beta}_i$, and variances, V_i. The formula for the conditional posterior means is particularly illuminating:

$$\bar{\beta}_i = \frac{V_\beta(y_i - \tilde{X}_i\tilde{\gamma})\tilde{Z}_i + h^{-1}\mu_\beta}{V_\beta\tilde{Z}_i'\tilde{Z}_i + h^{-1}}$$

where it is important to keep in mind that all of the other parameters are fixed based on draws from their conditional posterior distributions.[34] The mean formula is governed by the two parameters, V_β and h^{-1}. If V_β is large relative to h^{-1}, then the "independence" of the RTC dummies from each

model used for the estimates fully here. It closely parallels similar models in [Koop 2003, §§7.3-7.4].

[33]In particular, μ_β has a posterior normal distribution whose mean at any given draw in the posterior simulation is simply

$$\frac{1}{N}\sum_{i=1}^{N}\beta_i$$

which is the sample mean of the hierarchical coefficients – the RTC dummies. Similarly, posterior draws for V_β come from a gamma distribution with a mean constructed from the sample variance of the hierarchical coefficients.

[34]These conditional posterior means are not final summary statistics but are used to draw values of β_i to populate the simulated posterior distribution. Conditional on draws for the other variables such as h and $\tilde{\gamma}$, the conditional posterior means and variances indicate the appropriate normal distribution for making the β_i draws.

other predominates. In the limit as this dominance grows, we would have:

$$\bar{\beta}_i = \frac{(y_i - \tilde{X}_i\tilde{\gamma})\tilde{Z}_i}{\tilde{Z}_i'\tilde{Z}_i}$$

which is in form an OLS regression coefficient for the RTC dummy variable, $\tilde{Z}_i$.[35] The RTC dummy coefficients only will be similar to each other if these pseudo OLS coefficients are similar across different states, i. On the other hand, if the variance (the inverse of the precision, h^{-1}) of the disturbances dominates, then we should not pay much attention to the variation in the pseudo OLS estimates. As this dominance grows, in the limit we would simply take $\beta_i = \mu_\beta$ for all i. I.e., either V_β is so small that we essentially face a common value for the RTC dummies or the disturbance variation is so large that we would respect apparent differences in these dummies only at our peril.

The results for the 108 panel data regressions in [Donohue 2004] indicate that neither extreme specification (complete independence or complete uniformity) is appropriate. The posterior means of $\sqrt{V_\beta}$, the hyperparameter value characterizing the "standard deviation" of the hypothesized common normal distribution for the RTC dummy coefficients ranges from 0.0624 to 0.3296 for the 108 models with a mean of 0.1744. These levels for the "standard deviation" are substantial since they are in percentage point crime rate units, suggesting that there may well be considerable "inherent" differences in the responses of crime in the different states to RTC laws. On the other hand, some scale comparisons indicate that there is significantly less variation than the amount present in estimates for the complete independence specification. In particular, standard deviations in the 108 cases for the coefficient estimates under the hierarchical model range from 0.6453 to 0.8463 of the magnitude of the standard deviations for same cases under the complete independence specification. The mean ratio was 0.7791. Thus, the hierarchical model in a rough sense tends to lie about 7/9 of the way toward the complete independence specification from the other extreme of complete uniformity.

The estimate posterior mean values of $\sqrt{V_\beta}$ are sensitive to h^{-1}, the variance of the disturbance terms. The same may be true for the observed dispersion in the estimated RTC dummy coefficients. The estimates and observed dispersions reported above for the 108 models rest on a hierarchical

[35]Since $\tilde{\gamma}$ is taken as fixed, $(y_i - \tilde{X}_i\tilde{\gamma})$ is effectively a vector of dependent variable values for state i.

structure that dogmatically restricts the disturbances to be homoscedastic. Since the results may hinge on the disturbance structure, I also ran a "robust regression" version. This version follows the approach in [Geweke 1993], specifying a Student-t distribution for the errors. The degrees of freedom for the distribution are a hyperparameter, creating another layer of hierarchy. Following the literature, the prior distribution for the degrees of freedom is taken to be χ^2, and in the runs performed, I set the prior mean at 25. As shown in [Geweke 1993], this structure is equivalent to disturbances that are scale mixture of normals, drawing the disturbance precisions from a gamma distribution with a mean of one and the same number of degrees of freedom as the t-distribution. This structure is extremely flexible, allowing the disturbance terms to be normally distributed but with a very dispersed set of precisions. It also has the advantage of greatly reducing the parameters that need characterization from a number that is about the size of the square of the (possibly very large) number of observations, "n," to a single hyperparameter.[36] The results under the Student-t approach did not differ substantially from those reported above for the case of homoscedastic errors. The ratios of standard deviations for the coefficient estimates in the hierarchical Student-t errors model versus the independent OLS specification averaged 0.8269, not far from the 0.7791 average reported above when the hierarchical model was constrained by presuming homoscedastic errors.[37]

[36] A rapidly growing literature documents a diverse and very rich set of Bayesian approaches for dealing with covariance structures such as the covariance matrix for disturbances in a linear regression model. At one extreme, the researcher can specify a prior distribution for each of the $n(n+1)/2$ distinct covariance elements when there are n observations and then update using the data to generate a posterior, subject to the covariance matrix being positive definite and symmetric. When n is large, this approach may require a great deal of work to specify priors, and limits on computational power or speed may make the approach infeasible. At the other extreme are hierarchical approaches that effectively reduce the parameters to a small number.

[37] Adding autocorrelation in AR(1) form as in the Donohue-Levitt specifications of [Donohue 2004] to the Student-t error structure would be another step to take. Including the AR(1) structure would involve a third level of hierarchy, with an additional hyperparameter, "ρ," the autocorrelation at the first lag. A more general structure, e.g., allowing AR(p) errors for arbitrary p also would be possible. But Student-t errors and AR(1), would, in a rough sense, parallel to the most complex structure used in the frequentist estimates in [Donohue 2004]. The Student-t approach when layered on top of the independent Normal-Gamma prior is computationally intensive, each regression taking over an hour using a Windows-based MATLAB version on a PC with a 3 GHz processor. Adding the AR(1) feature would increase the computation time further. As a result, I did not try

In sum, treating the RTC dummy coefficients in the 108 panel data regressions strictly as indicative of a common response across states that is masked by estimation error appears to be quite inappropriate based on the results from hierarchical modeling. Of course, the seeming distinctness of the responses also may be an illusion due to omitted variables or other problems. The main point is methodological: the Bayesian hierarchical approach allows one to probe for the degree of uniformity rather than being bound to choose between the extremes of complete independence or complete uniformity.

This ability has two aspects that foreshadow the discussion of model comparison and model averaging in the next section. First, as noted by [Gelman, et.al. 2004, pp. 131-133], we can conceptualize the hierarchical model as a weighted average of the two extreme models. Second, making the hyperparameters characterizing the hypothesized distribution of coefficients explicit with prior and posterior distributions, enables us to assess what the weights are likely to be, the end result being a preferred model intermediate between the extremes.

3 Comparing Models and Model Averaging

3.1 The Setting

When one is comparing or choosing among models, the same framework described in section 2 applies. For any particular model "M_i," the researcher begins with a prior probability $P(M_i)$ that the model is true. The researcher specifies a likelihood function $P(y|M_i)$ based on data "y." The marginal likelihood $P(y)$ is the total probability that we will see the data "y" across all possible models. The researcher's posterior probability $P(M_i|y)$ that model M_i is true after observing data "y" is given by Bayes' rule:

$$P(M_i|y) = \frac{P(y|M_i)P(M_i)}{P(y)} \tag{8}$$

It is obvious that the researcher's choice of prior, $P(M_i)$, may have a big impact on the posterior probability. The frequentist regression approach of reporting or relying on the results for a single model corresponds to having a noninformative prior about the coefficients and other aspects of the chosen

it.

model but being dogmatic about the model being true to the exclusion of all others, i.e., $P(M_i) = 1$. The other models typically have a different set of variables. In the case where the chosen model rules out some variables, the decision not to include them corresponds to a dogmatic belief that they have coefficients of zero.

The contrast between the treatment of the coefficients within the chosen model and any excluded variables is striking. The researcher is being very open by asserting no beliefs about the magnitudes of the included coefficients but is being utterly dogmatic about the excluded variables, insisting that their coefficients must be zero.[38]

Under a Bayesian approach, these inconsistencies are either absent or made very explicit. The researcher states prior probabilities for each model under consideration and for all the coefficients and the error structure parameters in each model and then derives posterior probability distributions for the models, coefficients and error parameters.

If more than one model emerges with nonzero probability and the researcher is interested in the value of a particular coefficient, a common approach is to compute the posterior distribution for the coefficient via "Bayesian model averaging." Under this approach the posterior distributions for the coefficient are combined into one distribution via a weighted average of the posterior distributions for the coefficient emerging from each model. The weights are the posterior probabilities of the models. Thus, posterior information about the coefficient from high probability models has more influence on the model averaged posterior for the coefficient than posterior information from low probability models.

Using model averaged results typically makes more sense than choosing one particular model, even the highest probability model. The averaged results reflect the researcher's uncertainty about which model is true. Examining only one model disregards this uncertainty and often involves very different coefficient "estimates" and "standard errors." Suppose, for example, that the coefficient standard errors are smaller in the highest probability

[38]This inconsistency has not escaped notice. One prominent statistician observed:

> It is my impression that rather generally, not just in econometrics, it is considered decent to use judgment in choosing a functional form but indecent to use judgment in choosing a coefficient. If judgment about important things is quite all right, why should it not be used for less important ones as well? [Tukey 1978]

model than in the averaged model. Reporting only the highest probability model means that the researcher is falsely representing his or her true beliefs about the degree of accuracy of the coefficient estimates.

The heart of the Bayesian method is to make consistent probability judgments. One aspect of consistency is to be clear about dogmatic elements. In a given empirical situation, there are various sets of models that a researcher might consider. It is important to specify which models the researcher is (dogmatically) ignoring in the analysis. Two general situations, termed M-open and M-closed by [Bernardo & Smith 1994] are relevant here.

In the M-closed situation, the researcher believes that the true model is within the set being considered.[39] This belief has a dogmatic element since the researcher's prior probability for all models outside of the set is zero. In the M-closed situation, the researcher can use the approach suggested above: assign prior probabilities to all of the models, compute the resulting posterior probabilities, and model average to obtain coefficient estimates. The approach is not limited to the case of a small number of candidate models. For instance, suppose that a set of models in the literature includes (cumulatively) 100 independent variables. If the researcher believes that the true model consists of some subset of these variables but is very unsure what the subset is, the researcher may wish to consider the set of all possible models that combine the variables, assigning equal prior probability to each model. That set is large, 2^{100} models in all. Modern Bayesian computation techniques combined with the high degree of existing computer power makes the outlined approach potentially feasible because search algorithms exist that will locate most or all of the high probability models without having to compute posterior probabilities for all 2^{100} possibilities.

In many situations, the M-closed perspective is inappropriate. Suppose, for example, that the researcher believes there are soft variables that may have a significant influence on the results but that these soft variables are not fully captured in the available numerical data. In this case, the researcher does not believe that the true model is in any set that the researcher might construct with presently available quantitative data. The rationale for model averaging, assigning prior probabilities adding up to one for the models in a given set and then computing posterior probabilities that add up to one, becomes less clear.

[39]The term "M-closed" expresses the idea that the set of models M available to the researcher is "closed" in the sense that it is known to include the true model.

Some major empirical legal disputes consist of the parties contending for various alternative model specifications that suggest very different positive or normative conclusions. Typically, each party will focus on one favorite specification or a small set of favored specifications. In some disputes, the specifications evolve as the disputants clash, often with little or no change in policy positions on the part of individual disputants. A cynical observer wonders to what extent the specification choices reflect researcher prior beliefs rather than anything revealed by the data. The fact that we cannot observe all of the specifications that a particular researcher considered or estimated makes the situation worse.[40] In addition, there are many instances where researchers try various specifications but then never report their results on the topic by publication or otherwise. These cases may involve results that are not "interesting" or that conflict with the researchers' prior beliefs. In sum,

[40]One is reminded of the grim assessment by [Leamer 1983] of econometrics as practiced circa the early 1980s in an article entitled "Let's Take the Con out of Econometrics:"

> The false idol of objectivity has done great damage to economic science. Theoretical econometricians have interpreted scientific objectivity to mean that an economist must identify exactly the variables in the model, the functional form, and the distribution of the errors. Given these assumptions, and given a data set, the econometric method produces an objective inference from a data set, unencumbered by the subjective opinions of the researcher.
>
> This advice could be treated as ludicrous, except that it fills all the econometric textbooks. Fortunately, it is ignored by applied econometricians. The econometric art as it is practiced at the computer terminal involves fitting many, perhaps thousands, of statistical models. One or several that the researcher finds pleasing are selected for reporting purposes. This searching for a model is often well intentioned, but there can be no doubt that such a specification search invalidates the traditional theories of inference. The concepts of unbiasedness, consistency, efficiency, maximum-likelihood estimation, in fact, all the concepts of traditional theory, utterly lose their meaning by the time an applied researcher pulls from the bramble of the computer output the one thorn of a model he likes best, the one he chooses to portray as a rose. The consuming public is hardly fooled by this chicanery. The econometrician's shabby art is humorously and disparagingly labeled "data mining," "fishing," "grubbing," "number crunching." ... Or how about "There are two things you are better off not watching in the making: sausages and econometric estimates."
>
> This is a sad and decidedly unscientific state of affairs we find ourselves in. Hardly anyone takes data analyses seriously. Or perhaps more accurately, hardly anyone takes anyone else's data analyses seriously. ...

the published literature may be subject to various "reporting bias," "publication bias" and "data mining" influences that render it untrustworthy. Formal studies such as [DeLong & Lang 1992] and [Donohue & Wolfers 2005] have found that these phenomena appear to be present if not pervasive.

It is exactly these concerns about the empirical analysis of right-to-carry laws that motivates [Donohue 2004] to consider the results under alternative models. As mentioned above, in addition to six different specifications of a modified version of John Lott's model, [Donohue 2004] considers the same six specifications of three other models designed to predict crime but for purposes other than analyzing the right-to-carry laws.[41] The idea is that the researchers who promulgated these three other models cannot have had any conscious or unconscious motivation to lend support to their favored hypotheses about right-to-carry laws.

For the skeptic, however, two major problems remain. First, there are a very large number of actual and possible predictive models for crime. Why choose these three? The results under these three might be aberrant compared to other models that might have been chosen. This first problem acquires added salience because of the possibility that the results from three unrepresentative models have been reported and published because they accord with the conscious or unconscious bias of researchers or journal editors.

Second, [Donohue 2004] simply presents the results from the different models and notes that the results are quite sensitive to which model we choose. Having shown that specification sensitivity exists, it becomes imperative to know whether it matters. If some of the models are much weaker in terms of "fit" or predictive power than the others, why should we give them equal weight in our thinking? Why not just ignore them? More generally, as is well known, adding variables that are irrelevant but that happen to be correlated with subsets of the independent variables may have a dramatic effect on the significance, sign or magnitude of the critical RTC variables. It is not clear that, absent taking further steps, we can trust an indication of specification sensitivity that comes from adding some models with new variables and observing that the results are model dependent.

Bayesian analysis offers some important tools for addressing this situation. Researchers can avoid the inflexibility inherent in dogmatically focusing on one particular model by examining multiple models and admitting vari-

[41] The six specifications involve the different setups for the RTC dummies mentioned above, including the state specific and aggregate variants.

ous priors that might reflect the different stances of a diversity of readers. It becomes possible to put competing models advanced by rival academics into a single framework where one specifies prior model probabilities and then sees how these probabilities are affected by the data. Before discussing specific results for the models in [Donohue 2004], the next two sections discuss Bayesian model comparison and Bayesian model averaging in a more theoretical setting, making clear particular strengths and weaknesses that will play a role in examining model choice in the right-to-carry context.

3.2 Some Theory

Resolving model uncertainty by using a weighted average based on posterior model probabilities has a formal basis as well as intuitive appeal. Suppose that the marginal posterior distribution of some parameter θ is $f_{M_i}(\theta|x)$ under a particular model, M_i, and data, x. $f(\theta|x)$, the marginal posterior distribution of θ across all models, $M_1, M_2, ..., M_m$, is simply the weighted average of the marginal posterior distributions over the individual models where the weights are the posterior probabilities of each model, e.g., $f(M_i|x)$ for the i^{th} model:

$$f(\theta|x) = \sum_i f_{M_i}(\theta|x) f(M_i|x) \tag{9}$$

If one wanted a frequentist minimum variance "estimate" of the value of some regression coefficient in this context, one would choose the mean of the marginal posterior distribution for the coefficient across models which will be equal to the corresponding means in the underlying models weighted by the posterior probability of each model. Similarly, the standard error for this estimate would be a weighted average of the standard errors under each of the models.

It is important to point out that choosing the coefficient estimate from a single model, even the one with highest posterior probability, is not an optimal estimate under the minimum variance criterion. Failing to average together the marginal posteriors from the various models violates the logical consistency requirements that follow from Bayes' theorem. One is using the wrong posterior distribution to make estimates. A common move among legal empiricists is to add as many control variables as possible to a regression in order to test the effect of some variable, as represented by the coefficient estimate for that variable. This approach is clearly not optimal if there is any doubt about the appropriateness of any of the variables. The "full model"

consisting of all of the explanatory variables at the researcher's disposal may be a low probability model, and even if it is the highest probability model, one should use the marginal posterior that arises from model averaging instead. That posterior fully reflects the researcher's uncertainty about the appropriate model.

There are some objective functions that dictate choosing the highest probability model rather than an average across models. For example, if there is a positive payoff for choosing the true model from a discrete set but a zero payoff otherwise, then an average across models is a sure loser and the model with the highest posterior probability is the best choice. Note, however, that model comparison is front and center: identifying the model with highest posterior probability is critical.

Bayesian model averaging and model comparison have some attractive "asymptotic" properties under the frequentist paradigm of trying to identify some "true" underlying data generating process. These properties are somewhat technical but easy to summarize. Each regression model with a particular set of coefficients and particular assumptions about the distribution of the disturbance terms represents a specific data generating process. Define the "true model" as one with the precise set of variables that generate the data. The true model with the correct set of coefficients and the correct error structure is the true data generating process. Assume that the regression approach is sufficiently flexible that the correct set of coefficients and the correct error structure is among the possibilities. Under this assumption, in an M-closed setting, the true model is among the candidate models, and as the number of independent observations increases, the probability assigned to the true model converges to one. The M-open setting is somewhat more complicated. If the true model is not itself one of the alternatives but is contained in one or more of the alternative models, then the model containing the true model that is most parsimonious will be the asymptotic choice. If none of the available models contains the true model, then the model that contains a data generating process (i.e. a particular set of coefficients and error structure admissible under the regression procedure) closest to the true data generating process will prevail asymptotically.[42] Thus, although

[42] [Dawid 1992] and [Dawid 1999] derive most of these results in more general form. [O'Hagan & Forster 2004, pp. 180-183] is a good intuitive survey-like treatment of the subject.

Comparing data generating processes involves comparing probability distributions. The measure of "closeness" under which the text results are true is the Kullback-Liebler diver-

Bayesian model averaging does not lead to the true model in an M-open setting, it is in an important sense the best one can do given that the true model is inaccessible.

These asymptotic properties are appealing and reassuring, but the added value from Bayesian model comparison or model averaging typically arises in the context where there is limited data. In an asymptotic setting, frequentist methods often will have the same property of being able to identify the true model if that model is available. Suppose, for example, that consistent estimators are available in a regression context and the full model includes all of the variables in the true model. The coefficient estimates of the extraneous variables will converge to zero as the available independent observations tend toward infinity. What will remain is the true model with correct coefficients. But without the benefit of a potentially unlimited amount of data, model uncertainty generally will be present. As a consequence, using the full model by itself will tend to be a very deficient approach compared to optimally allowing for model uncertainty using Bayesian methods.

It is important to make clear the sense in which the Bayesian model comparison or Bayesian model averaging approaches in later sections "optimally allow for model uncertainty." The model probabilities follow from the marginal likelihood of the dependent variable outcome under different models.[43] Thus, *the approaches are choosing among models based on their differential abilities to predict the dependent variable.* This predictive criterion matches up superbly with the idea in [Donohue 2004] to address specification issues by looking at a variety of models that have been chosen for their ability to predict various crime rates and then assessing how RTC related variables play out in those models. Bayesian model comparison and averaging implement this idea in a more general and systematic way, permitting more meaningful conclusions.[44]

Although Bayesian model comparison and averaging are appealing ways

gence, the difference between expected loss and its minimum under a loss function equal to $-ln(d(\theta))$ where $d(\theta)$ is a density function. [O'Hagan & Forster 2004, pp. 57-59] contains a nice, concise description of this measure, the associated loss function and the rationale for using that particular loss function.

[43]Similarly, the Kullback-Liebler divergence is with respect to the posterior distribution of the dependent variable.

[44]The predictive criterion is not the only relevant one, nor the only one advocated in [Donohue 2004]. In later sections, I discuss some others and their relationship to some of the Bayesian approaches presented in the article.

to address model uncertainty in some contexts, they are subject to the same two potential weaknesses that afflict Bayesian approaches more generally: the possible sensitivity of the results to the choice of priors and potential computational difficulties. First, and most important, the marginal likelihoods that lead to the model probabilities are sensitive to the priors chosen both for estimating each model and for choosing among models. I will be very specific about these "within model" and "across models" prior choices when applying a Bayesian model averaging approach to the models in [Donohue 2004]. Prior sensitivity does not matter if the researcher actually has a well-defined prior or can specify a set of priors that might be held by readers. In the later case, the researcher simply can present the model averaging results under alternative priors. The exercise in [Donohue 2004], however, is of a different character, trying to make a more general statement about the impact of RTC laws. In this vein, it is appropriate to aim at results that are "neutral" in the sense of being independent of any particular prior. We will see in the following discussion that pursuing this goal raises significant difficulties along several fronts. The underlying reason is simple and already has been mentioned. Bayesian analysis involves logical consistency – what priors and likelihoods imply about posteriors – and nothing more. In choosing priors, it often is the case that "neutrality" is not an objective property but is in the eye of the beholder. As a general matter, it would seem best to choose priors that are "interesting" or "relevant" in the sense of being meaningful to researchers or their audience. Some of these priors may be attractive because they seem "neutral" to part or all of that audience. As stated above, one nice feature of Bayesian analysis is the possibility of using alternative priors that will appeal to individuals with different beliefs or who are concerned about different issues. Second, computation can be an issue, especially when the number of models is large. We address both of these issues in the next subsection as part of the process of introducing the methods we will apply to parse the right-to-carry models.

3.3 Implementation Issues

On the computational side, there is a divide between prior/likelihood combinations that result in known posterior distributions and those that do not. In the former case, one can calculate the key parameters of the posterior distribution in closed form and generally quickly. In the later case, it often is possible to simulate the posterior distribution, but simulation typically

involves computation times that are orders of magnitude longer.

For a classical linear regression context model under the assumption of homoscedastic normally distributed errors, a workhorse closed form approach is to use a "natural conjugate prior" in normal-gamma form.[45] We explicate enough of this approach at this point to make the model comparison issues clear.[46]

Suppose that the normal linear regression model is:

$$y = X\beta + \epsilon$$

where y is $n \times 1$, X is an $n \times k$ data matrix possibly containing an "intercept" column of all ones, β is a $k \times 1$ vector of coefficients, and ϵ is an $n \times 1$ vector of uncorrelated homoscedastic mean-zero normally distributed errors with unknown variance, σ^2. Bayesian treatment of this model requires prior distributions for the regression coefficients, β, and the error precision, $h = \frac{1}{\sigma^2}$.[47] The likelihood in this model is normal but can be decomposed into the product of a normal distribution for $\beta - \hat{\beta}$ where $\hat{\beta}$ is the vector of OLS estimates for β and a gamma distribution for h. Choosing a gamma prior distribution for h and a distribution for β that is normal conditional on h results in a natural conjugate prior – the posterior distributions for β and h also will be conditionally normal and gamma respectively.

Using notation from [Koop 2003], assume that the prior distribution of β is $N(\underline{\beta}, h^{-1}\underline{V})$ while the prior of h is gamma with mean $\underline{s}^{-2}$ and $\underline{v}$ degrees of freedom. Rather than write down the general formula for the parameters of the posterior distributions, we first consider a specific form for $\underline{V}$, namely:

$$\underline{V}^{-1} = gX'X$$

where g is a positive constant.[48] This form is known as the "g-prior," an approach developed originally by [Zellner 1986]. It a very popular choice

[45] A "conjugate prior" is one that results in a posterior distribution from the same family as the prior distribution. A "natural conjugate prior" is one where the likelihood also is from the same family.

[46] The are many full treatments in the literature. The text discussion follows closely a very good one in [Koop 2003, ch. 3].

[47] In many frequentist treatments, the focus is on the error variance, σ^2. It often is much simpler to work with the precision, $h = \frac{1}{\sigma^2}$, in a Bayesian context.

[48] X'X is a $k \times k$ matrix, very familiar to anyone who has studied regression in a matrix context, consisting of sums of squares and summed cross products for the variables.

because it leads to greatly simplified computations and clear intuition while remaining very flexible. An important special case is the particular choice $g = \frac{1}{n}$. We develop that special case here. When $g = \frac{1}{n}$, the posterior distribution of β conditional on h is $N(\overline{\beta}, h^{-1}\overline{V})$ where

$$\overline{\beta} = \frac{1}{n+1}\underline{\beta} + \frac{n}{n+1}\hat{\beta} \tag{10}$$

$$\overline{V} = \left(\frac{n+1}{n}X'X\right)^{-1}.$$

The posterior mean for β is a weighted average of the prior mean, $\underline{\beta}$ and the OLS estimate, $\hat{\beta}$, where the weights are 1 and n respectively.[49] This case is an example of a "unit information prior," a prior that has a weight equal to one observation versus the n observations in the data.

It is evident that for large values of n, the posterior mean for the coefficients will be very close to the corresponding OLS estimates. The unit information prior effectively gives very little weight to the prior in the large n situation and can be characterized as "relatively noninformative" in the sense that the prior has almost no impact on the posterior. It is possible to go further and use a prior that is "noninformative," with the resulting posterior mean for the coefficients being exactly equal to the OLS estimates. This prior would assign equal probability to all possible values of the coefficients and would be "improper" in the sense that there is no probability density function representing the prior.[50] Estimates are possible because multiplying by the likelihood may result in a proper density function. This prior would be attractive to use in model comparison since it is apparently "neutral" in the sense that we are not giving any weight to prior beliefs versus the data. However, there is a problem. Model comparison will not work if noninformative priors are assigned to parameters that are present in one model but absent in others. Specifically, if two models have an unequal numbers of parameters, then the more parsimonious model will end up with 100%

[49] If we take $\underline{\beta} = 0$, a common choice, then it is evident from equation (10) that the posterior mean "shrinks" the OLS estimates by the factor $n/(n+1)$. As a result, the literature sometimes refers to this phenomenon as "shrinkage." A wide range of factors ≤ 1 other than $n/(n+1)$ can arise in various applications.

[50] The density would have to be a positive constant, c, over the full range $(-\infty, \infty)$ of possible values for β, but the constant function does not have a finite integral over that range.

posterior probability regardless of how salient that model is. In the case of equal numbers of parameters, the outcome will be completely dependent on the scaling of the variables.[51] As a result, using "relatively noninformative" priors is a move that attempts to come close to "neutrality" in the sense of giving little weight to prior beliefs while at the same time preserving the ability to compare models. It is worth noting that when model comparison focuses on the independent variables and h is common to all models, it does no harm to use a noninformative prior (e.g., $\underline{\nu} = 0$) for h. This approach is common.

Even with a relatively noninformative prior for the coefficients, two aspects of the prior can plague model comparison if "neutrality" is the goal. One is the difference between $\underline{\beta}$, the vector of prior means for the coefficients and $\hat{\beta}$, the vector of OLS estimates. Assuming a noninformative prior for h and a g-prior with $g = \frac{1}{n}$, the marginal likelihood for model M_j is proportional to:

$$\left\{ RSS_j + \frac{1}{n+1}(\widehat{\beta}_j - \underline{\beta}_j)' X_j' X_j (\widehat{\beta}_j - \underline{\beta}_j) \right\}^{-\frac{n}{2}} \tag{11}$$

where $RSS_j = (y - X\widehat{\beta})'(y - X\widehat{\beta})$ is the residual sum of squares under OLS for the model. The second term in the curly brackets measures the coherence of the OLS estimates with the prior mean for the coefficients. If the prior mean deviates sharply from the OLS estimates, this term will be large, the marginal likelihood will be small, and the model will fare poorly in comparisons with others. Although this term becomes insignificant compared to the RSS term as n grows large, we will see that it still has a big effect in the [Donohue 2004] model comparisons even though $n \approx 1000$.

A similar "coherence with the prior" term affects the posterior standard error of the coefficients under a natural conjugate prior. Larger differences between the OLS estimates and the prior mean result in larger posterior standard errors for the coefficients, an outcome that makes sense since one should be less confident in estimates that differ more from one's prior. This coherence with the prior problem that affects the posterior standard errors for the coefficients may be alleviated by using an independent normal-gamma prior instead of the natural conjugate prior: the prior distribution for β will be $N(\underline{\beta}, \underline{V})$ instead of $N(\underline{\beta}, h^{-1}\underline{V})$. The posterior under this prior is

[51] See [Koop 2003, pp. 40-43] for details.

not a known distribution, and the need to simulate the posterior instead of computing it directly makes computation more difficult. Estimating the marginal likelihood for each model is particularly difficult, and, as detailed in [Koop 2003, pp. 165-168], use of relatively noninformative priors will strongly and unduly favor parsimonious models. Model comparison will be meaningful only when the models have roughly an equal number of parameters. The driving force is a variant of the "coherence with the prior" phenomenon. A relatively noninformative prior allows very little reward for estimates coming close to the prior – most of the probability mass in the prior will be far away from any particular point estimate. A model with more parameters will look even worse in this regard – even more "misses." In the limit where the prior becomes noninformative, the phenomenon becomes ironclad: the most parsimonious model wins regardless of the relative fit that it offers. In sum, the independent normal-gamma approach alleviates the coherence with the prior effect on the posterior standard errors for the coefficients but not the effect on model comparison.

The second aspect involves the choice of g. Although $g = \frac{1}{n}$ has some very attractive properties such as "consistency" in the sense that the true model emerges with probability one as the number of independent observations increases, the results can be sensitive to the precise value of g used in the model comparison exercise. As detailed in [Fernandez, Ley & Steel 2001, Theorem 1, p. 421], a range of alternatives to $g = \frac{1}{n}$ also are consistent.

The focus of the discussion so far has been on the sensitivity of the model comparison results to the "within model" prior choices. The results also may be sensitive to "among model" prior choices. Recall that the Bayesian approach requires that we choose a prior probability, $P(M_i)$ for each model M_i. One useful formulation is to postulate a prior probability, "w" for any one independent variable being in the model. If model M_i has q_i independent variables out of a possible p, then

$$P(M_i|w) = w^{q_i}(1-w)^{p-q_i}.$$

A common simplification is to assume $w = 1/2$, a necessary and sufficient condition for all models to have equal prior probability, 2^{-p}. In contrast, choosing a low or high value of w will favor sparse and nearly saturated models respectively. Equal prior probabilities for all models sounds like an attractive ("neutral" perhaps?) starting point. However, context is critical, and equal probabilities will be a poor or inappropriate prior in many

instances. For example, if the researcher and the audience believe that the model is exploratory and that many of the variables probably are extraneous, choosing a prior with a value of w much less than 1/2 would be reasonable.

The difficulty here arises because there is no one particular prior appropriate to the problem when the goal is to be agnostic about the choice of priors. As pointed out by [George & Foster 2000], a purely Bayesian response to this dilemma is to express agnosticism directly by creating a hierarchical model with hyperparameters representing the possible range of priors and then choosing a prior for these hyperparameters that is noninformative or relatively noninformative. This prior for the hyperparameters effectively expresses agnosticism, "neutrality" or "ignorance," allowing a posterior distribution of particular priors for the model parameters to emerge from the data itself. In the model comparison case we are examining, "g" and "w" would be the hyperparameters.[52] [George & Foster 2000] consider this very case, although they use slightly different notation: $c = 1/g$ in place of g. The downside of using a purely Bayesian hierarchical approach in this context where there are many models is that adding the hyperparameter calculations to the mix can make computation prohibitively slow. [George & Foster 2000] therefore consider two "empirical Bayes" approaches. We will use the simpler of these two, the "conditional maximum likelihood criterion" or "CML" for short, in what follows.

An "empirical Bayes" approach substitutes hyperparameter "point estimates" for the posterior distribution that would emerge from a fully Bayesian treatment.[53] Under CML the point estimates are the values of c and w that

[52]The approach is very similar to the use of hyperparameters in other contexts such as the hierarchical treatment of the RTC dummy discussed in subsection 2.4 above. That treatment admits a continuum of models between complete uniformity (same impact of RTC laws in all states) and complete independence (no commonality in the responses) and then emerges with a posterior distribution for the location along the continuum. Here the task is explicit model comparison, and the goal is to express the researcher's degree of uncertainty about appropriate priors for the model parameters.

[53]Although the term "empirical Bayes" is well established and perhaps definitive at this point, some very astute commentators object to it. [O'Hagan & Forster 2004, p. 126] notes that "[e]mpirical Bayes is not Bayesian because it does not admit a distribution [for the parameter of interest]." [Gelman, et.al. 2004, p. 112] "... prefer to avoid the term ... because it misleadingly suggests that the fully Bayesian method ... is not empirical." Terminology aside, it is clear is that the empirical Bayes approach includes a frequentist element (use of point estimates) which probably motivated including the word "empirical" in the description.

maximize the marginal likelihood *conditional* on a particular model being true. As a result, there will be distinct estimates, $\hat{c}$ and $\hat{w}$, for each model. The other empirical Bayes approach considered by [George & Foster 2000] is the "maximum marginal likelihood criterion," "MML" for short. MML involves maximizing the marginal likelihood across all models, the result being a single $\hat{c}$ and single $\hat{w}$ applicable for all models. The downside of empirical Bayes versus fully Bayesian approaches is that the uncertainty inherent in the posterior distributions for c and w is ignored. This uncertainty would be fully captured if we integrated out these hyperparameters using the joint posterior distribution of all the parameters and hyperparameters in the model. Instead, we are simply evaluating all relevant parameters at the point estimate value of the hyperparameters, with CML taking the point estimate approach a step further than MML.[54] Both methods perform very well in [George & Foster 2000]'s simulations when compared to other model comparison criteria such as BIC. Although MML does better than CML, MML is computationally prohibitive for large models. In contrast, CML requires only OLS output – a few lines of code in STATA or similar packages will do the job. This trait makes CML potentially useful for researchers who shy away from Bayesian approaches because of the computational requirements.

As long as the true model is not the null model (all coefficients zero), CML has the same "consistency" properties as BIC. As the available number of independent observations grows large, it will assign a probability approaching one to the true model if that model is one of the alternatives. In addition, some popular model comparison criteria such as BIC or AIC (the Aikake information criterion) are special cases where c and w take on particular values.[55] The [George & Foster 2000] approaches (full Bayes and MML as well as CML) allow these criteria to emerge as applicable based on the data.[56]

[54]While MML integrates over the model space to arrive at global estimates, $\hat{c}$ and $\hat{w}$, CML uses maximum likelihood estimates of these values under the conditional likelihood for each model. CML is therefore an additional step removed from a fully Bayesian treatment.

[55]In the case of BIC, these are $c = n$, where n is the number of observations, and $w = 1/2$.

[56][George & Foster 2000, p. 739] identify a weakness of CML, not shared by MML or a full Bayes treatment, that can be important in certain applications: "... unless the true coefficients are large, [CML] tends to be bimodal over the model space, with one mode closer to the true model and the other at the completely saturated model." When the coefficients are small, the data will favor a small value of c, putting more weight on the prior which includes a mean of zero for the coefficients. As a result, in that situation,

Finally, it is worth noting that the [George & Foster 2000] approaches implicitly address the "within model" coherence with the prior problem that comes from choosing a prior mean of zero for the coefficients. If the data conflict with this prior mean, then the $\hat{c}$ values under CML or MML and the posterior distribution of c under the general Bayesian approach will be located near very large values, indicating very low weight on the prior. I.e., the flexibility with respect to c will allow the data to wash out the zero prior mean if indicated even if the number of observations is "small." The problem with fixing c is that it forces a particular weight for the prior regardless of the divergence of the prior mean of the coefficients from the values suggested by the likelihood function that represents the data.

3.4 Comparing the Panel Data Models

Both "within model" prior sensitivities discussed in the previous subsection emerge clearly from the data. Consider first the choice of prior mean for the

---

the "[bimodal] behavior stems from the fact that the likelihood does not distinguish well between models with small c and small w [low coefficient values, not many variables in the model] and models with even smaller c but large w [e.g., the saturated model with even lower coefficient values]." Id.

CML plays an important role in some of the applications in the following sections. As a result, it is important to assess whether the bimodal problem is significant for these applications. The set up in [George & Foster 2000] envisions a model space that includes all possible combinations of variables. The applications which follow do not employ this setup. They either involve comparing the small set of models ($\sim 20 - 30$) studied in [Donohue 2004] or consider a much larger set of models ($\sim 221,000$) that consist of mutually exclusive choices among alternatives within each of 12 variable groups. As a result, there is no coherent model space, and it is not clear in either context how we would define or identify a bimodal outcome. Nonetheless, the problem remains. In response, I check in each instance on the values of $\hat{c}$ associated with the model. $\hat{c}$ ends up being very large for the models that have significant probability or that comprise averages – typically between 2 and several hundred times n, the sample size. This result is not surprising since the model posteriors typically include "big" coefficients (on a standardized basis) for many of the variables. The small values of $\hat{c}$ that would create the bimodal problem are absent. The prior is receiving very little weight, substantially less than the $1/(n+1)$ weight that would arise from the "unit information prior" (inherent in BIC) for which $c = n$. Consistent with these $\hat{c}$ values, in the CML applications with the large set of models ($\sim 221,000$), saturated models are far from predominant. In fact, as detailed in Appendix B, for one of the twelve variable groups, the high probability models almost always omit the group entirely; two others are omitted more than 85% of time; six more are omitted between 44% and 77% of the time. Only two of the twelve groups are almost always present.

coefficients. The "coherence with the prior" term in the marginal likelihood (the second term in the sum, equation (11)) is quadratic in the difference between the prior mean and the OLS estimate. A larger difference increases a term raised to a large negative power, therefore reducing the marginal likelihood. Under a "prophetic prior" vector that is equal to the vector of OLS estimates, the coherence with the prior term will be zero. A traditional choice for the prior mean vector is zero. Compared to the prophetic prior, this choice will favor models that have fewer variables with estimated OLS coefficients substantially different from zero. As discussed above, this effect disappears asymptotically.

The effect, however, is quite present and salient in the [Donohue 2004] model comparison. The following table presents model probabilities based on the reference prior for violent crime and the 24 panel data models under three separate approaches:[57]

1. a natural conjugate g-prior with $g = 1/n$ and a zero prior mean for the coefficients;

2. a natural conjugate g-prior with $g = 1/n$ and a prophetic prior mean for the coefficients;

3. BIC – the asymptotic approximation for items 1 and 2.

[57]The reference prior endows each model with equal prior probability. With this prior, the posterior model probabilities for the 24 models will equal the marginal likelihood for each model divided by the sum of the marginal likelihoods for all 24 models.

Panel Data Specifications for Violent Crime Model Probabilities under Three Approaches (column maximums in bold)			
model & specification	zero prior	prophetic prior	BIC
ML: dum–agg	5.77 e-042	3.56 e-054	3.49 e-054
DL: dum–agg	**9.71 e-001**	7.30 e-079	7.03 e-079
SP: dum–agg	7.81 e-013	7.69 e-114	7.41 e-114
ZH: dum–agg	1.24 e-015	2.96 e-098	2.86 e-098
ML: spl–agg	2.42 e-043	1.09 e-054	1.06 e-054
DL: spl–agg	2.94 e-002	2.46 e-080	2.37 e-080
SP: spl–agg	5.06 e-014	8.68 e-114	8.38 e-114
ZH: spl–agg	5.28 e-017	5.51 e-099	5.33 e-099
ML: st_tr–agg	8.02 e-072	9.10 e-020	9.00 e-020
DL: st_tr–agg	2.71 e-029	9.73 e-056	9.47 e-056
SP: st_tr–agg	8.36 e-042	3.36 e-099	3.28 e-099
ZH: st_tr–agg	1.58 e-041	9.18 e-060	8.98 e-060
ML: dum–st_sp	2.21 e-074	4.01 e-051	3.97 e-051
DL: dum–st_sp	6.94 e-032	8.06 e-079	7.85 e-079
SP: dum–st_sp	1.42 e-043	2.65 e-115	2.58 e-115
ZH: dum–st_sp	3.26 e-045	5.10 e-090	4.98 e-090
ML: spl–st_sp	5.27 e-104	**1.00 e+000**	**1.00 e+000**
DL: spl–st_sp	3.70 e-062	2.51 e-058	2.47 e-058
SP: spl–st_sp	9.24 e-074	1.25 e-100	1.24 e-100
ZH: spl–st_sp	6.79 e-073	6.37 e-050	6.29 e-050
ML: st_tr–st_sp	1.51 e-108	1.32 e-044	1.32 e-044
DL: st_tr–st_sp	1.64 e-065	5.51 e-081	5.42 e-081
SP: st_tr–st_sp	2.51 e-078	9.87 e-128	9.74 e-128
ZH: st_tr–st_sp	6.81 e-078	1.48 e-085	1.47 e-085
models: ML = Modified Lott; DL = Donohue/Levitt; SP = Spelman; ZH = Zheng			
RTC specifications: dum = dummy alone; spl = spline; st_tr = with state trends; agg = aggregate (1 RTC variable); st_sp = state specific (26 RTC variables)			

It is clear from the table that choice of the prior mean for the coefficients makes a huge difference. Under a natural conjugate approach with $g = 1/n$ and a zero prior, the two model/specification combinations with the smallest

number of variables, the Donohue-Levitt models with a single RTC variable, have the highest probabilities.[58] With a prophetic prior, the Modified Lott model with state-specific spline RTC dummies dominates. The probability is virtually one, with the next best model/specification combination having a probability of $\sim 10^{-19}$. This model/specification has a relatively large number of variables although not the largest number. Furthermore, the model probabilities for the prophetic prior are nearly identical to the probabilities computed using BIC. After removing the "coherence with the prior effect," the results under the natural conjugate prior and $g = 1/n$ are close to the result under the asymptotic approximation. Furthermore, the fact that the models with the smallest number of variables dominate in the zero prior mean case is no big surprise. As discussed above, the coherence with the prior term will favor models with fewer variables with estimated OLS coefficients substantially different from zero.

This pattern for violent crime typifies the results for the other eight crime categories. The results under a prophetic prior closely match BIC, and a Modified Lott variant dominates. With a zero prior, one of the two (Donohue-Levitt) model/specifications with the smallest number of variables dominates with one exception: In the case of auto theft, the dominant model/specification is Modified Lott with an aggregate RTC dummy, the Modified Lott variant with the smallest number of variables.

The same coefficient prior mean effects arise for CML and CML-related regressions but the relationships are not quite as tight. Using CML-generated $\hat{c}$ and $\hat{w}$ values under a natural conjugate approach with a zero prior tends to favor small models. The two models with the smallest number of variables dominate for two of the nine crime categories, "small" Modified Lott variants dominate in another four, and "moderate" sized variants of Modified Lott and Donohue-Levitt dominate in the remaining two. With a prophetic prior the results are close to outcomes under CML. The dominant model matches for seven of the nine crime categories. The somewhat looser "coherence with

[58] Not counting the RTC, state trends, year dummy or state dummy variables or variables that must be removed due to exact multicollinearity, the Donohue-Levitt models have 6 independent variables, the Spelman model has 10, the Zheng model has 15, and the Modified Lott model has 42. The "state specific" specifications replace a single RTC dummy or spline variable with 26 such variables and, where present, replace a single state trends variable with 26 such variables. All specifications have 50 state dummies and 22 year dummies – one of the 23 total year dummies having been removed to prevent exact multicollinearity.

the prior effect" is not a surprise since CML implicitly adjusts for model size to some extent through the $\hat{w}$ parameter.

The other "within model" prior sensitivity discussed in the previous section involved the choice of g. This choice has a major impact. For example, using $g = 1/n^2$ instead of $g = 1/n$ in the natural conjugate approaches above pulls the prophetic prior results sharply away from the BIC results. The dominant models for each crime category under the prophetic prior become equivalent to those under the zero prior, and the pattern of dominant models under the zero prior differ from the $g = 1/n$ case: The "small" variants of the Modified Lott model dominate for five of the nine crime categories instead of just for auto theft. This sensitivity motivates use of CML or CML-related approaches where g (or equivalently, c) emerges from the data instead of simply being fixed at $1/n$ or some other value.

Because of the ability of the CML approach to ferret out some of the prior sensitivity, the rest of the article emphasizes results under this approach. In most instances, however, I also report results under BIC. One reason is that the p-value discussion in subsection 2.3 above used BIC to translate p-values under a null hypothesis that the RTC indicator regression coefficients are zero into pseudo-Bayesian quantities. This section will explore a more general hypothesis of RTC-irrelevance that nests this null hypothesis in an explicit model comparison framework, and it is interesting to see how the results play out under BIC for purposes of comparison with the previous section. BIC also has some desirable properties, being a "conservative choice" that arguably tilts toward the null in a way that is reminiscent of the often-employed 5% significance frequentist set up.[59]

As discussed in the previous section, computation is greatly facilitated by using a natural conjugate prior or by relying on approximations such as BIC or CML because the marginal likelihood or posterior probability for each model is simply a formula. However, these approaches presume a homoscedastic normal error structure. Adding in a more complex error

[59]The literature concerning BIC is immense. [Raftery 1999, pp. 414-417] makes the case for BIC as a "conservative" choice. BIC results in a fairly diffuse prior since it imposes influence equivalent to only one data point. A totally diffuse (non-informative) prior would result in the null being favored in all instances. BIC tilts the results toward the null but is not so excessively spread out that the null is impervious to rejection given moderately strong indications in the data. [Raftery 1999] notes that in some cases a less diffuse prior than BIC will be a superior choice. CML-based estimates allow flexibility concerning how diffuse the prior is.

structure generally means that the posterior distribution is not a known distribution. As a result, marginal likelihoods and the corresponding model probabilities are not available as formulas. Instead, they must be simulated by draws from the posterior distribution. The required computational time rises sharply. At the same time, there is no reason to use the restrictive natural conjugate prior structure. An independent normal-gamma prior is a popular alternative in the regression context, but, as mentioned above, under a relatively noninformative prior for the coefficients, the coherence with the prior problem remains under this alternative. In particular, under the common techniques for estimating marginal likelihoods, models with small numbers of parameters will be strongly favored regardless of quality. This feature makes relying on the independent normal-gamma prior a poor choice for model comparison in the [Donohue 2004] context since the number of parameters differs greatly across alternative models and specifications. Nonetheless, using this prior is useful for gaining insight into the question of whether using a more general error structure might matter because we will be able to compare paired models with similar numbers of parameters.

Along these lines, I generated results for the 24 combinations of model and RTC specifications in [Donohue 2004] using an relatively noninformative independent normal-gamma prior with and without Student-t errors. The use of Student-t errors resulted in much larger marginal likelihoods than the normal homoscedastic error structure.[60] However, the relative positions of the models did not shift very much. This outcome is reflected in the following table which reports the highest posterior probability model for each crime category under the two error structures – they are the same for eight out of nine categories.[61]

[60]The models with Student t errors contain only one additional parameter (the degrees of freedom for the t distribution) versus the models with homoscedastic errors. The small difference in number of parameters means that it is unlikely that the number of parameters will dominate the marginal likelihood differences between paired models. The fact that the model in each pair with one more parameter has a higher marginal likelihoods in the comparison here exemplifies that fact.

[61]The reference prior applies – we assign equal prior probability to each model. Not all of the highest probability models were "dominant" in the sense of having virtually 100% probability, but all had probabilities greater than 50%.

Highest Probability Model/Specification Panel Data Specification Regressions with Independent Normal Gamma Prior		
crime category	normal errors	Student-t errors
violent	ML: dum–agg	ML: spl–agg
murder	DL: spl–agg	DL: spl–agg
rape	DL: dum–agg	DL: dum–agg
robbery	ML: dum–agg	ML: dum–agg
agg. assault	DL: dum–agg	DL: dum–agg
property	ML: dum–agg	ML: dum–agg
burglary	ML: spl–agg	ML: spl–agg
larceny	ML: dum–agg	ML: dum–agg
auto theft	ML: dum–agg	ML: dum–agg
models: ML = Modified Lott; DL = Donohue/Levitt; SP = Spelman; ZH = Zheng		
RTC specifications: dum = dummy alone; spl = spline; st_tr = with state trends; agg = aggregate (1 RTC variable); st_sp = state specific (26 RTC variables)		

The tendency for model/specification combinations with few variables to have high probability also is evident from the table. The four highest probability combinations include the two with smallest number of variables (DL: dum–agg; DL: spl–agg) or the two sparsest specifications of the Modified Lott model.

With some assurance that a different error structure most likely will not upset model rankings, we turn to the results of comparing the 24 panel data models. For each crime category and comparison approach, there is a "dominant" model/specification combination. "Dominant" means that the posterior probability for the combination (under a reference prior assigning equal initial probability to each combination) is close to 100%.[62] The following table lists the dominant combinations:

[62]In 13 out 18 instances, the probability is within 10^{-7} of 1. In four others, it is at least 0.997. The lowest probability is 0.989.

Dominant Model/Specification under BIC and CML Panel Data Regressions (under equal probability reference prior)		
crime category	BIC	CML
violent	ML: spl–st_sp	ML: spl–st_sp
murder	ML: st_tr–agg	ML: st_tr–agg
rape	ML: st_tr–agg	ML: st_tr–agg
robbery	ML: spl–st_sp	ML: spl–st_sp
agg. assault	ML: spl–st_sp	ML: st_tr–agg
property	ML: st_tr–agg	ML: dum–agg
burglary	ML: st_tr–agg	ML: st_tr–agg
larceny	ML: st_tr–agg	ML: st_tr–agg
auto theft	ML: spl–st_sp	ML: st_tr–agg
models: ML = Modified Lott; DL = Donohue/Levitt; SP = Spelman; ZH = Zheng		
RTC specifications: dum = dummy alone; spl = spline; st_tr = with state trends; agg = aggregate (1 RTC variable); st_sp = state specific (26 RTC variables)		

Two features are evident. First, two RTC specifications of the Modified Lott model predominate: the aggregate specification that includes a state trends variable along with a single RTC dummy; the state specific specification that includes 26 state specific RTC "spline" variables each of which equals years since RTC adoption in a particular state and zero otherwise. Second, the BIC and CML lists correspond closely. This correspondence is not a big surprise given that: (i) model choices under CML and BIC converge asymptotically; and (ii) we have lots of data here – 1150 observations.

The outcomes in the table are reflected more broadly in patterns that are not easily reported in tabular form given the large number of crimes, specifications and models. Consider first the exercise of holding the model fixed and considering how the six RTC specifications rank with respect to posterior probability. The results are similar regardless of model. The three state specific RTC specifications tend to dominate the three aggregate ones except for the aggregate specification with state trends. That aggregate specification tends to do very well, often ranking first or second among the six. The fact that state specific specifications do well in general is not surprising given that the results under the hierarchical model in subsection 2.4 above ended up 80% of the way toward independent coefficients (distinct state ef-

fects) versus a single aggregate coefficient. Second, consider holding each RTC specification fixed and asking which model has highest posterior probability. The answer is very clear: holding the RTC specification constant, the Modified Lott model *always* has the highest probability. It is important to note, however, that this result does not preclude the other models having useful features that are absent from Lott's approach. For one thing, the dominance of the Modified Lott variants does not hold up when we allow the RTC specification to vary. The Donohue-Levitt and Zheng models have higher probabilities under some RTC specifications than the Modified Lott models under different specifications. In fact, the results in the next section will indicate that once we relax the restriction of considering only these 24 models but are allowed to mix and match, the Modified Lott versions will be totally superseded. The posterior probabilities concerning the impact of RTC laws on crime also will shift.

Before moving to a more general framework, it is useful to ask: What do these results comparing the 24 model/specification combinations portend for the impact of RTC laws on crime? The answer to this question indicates how the analysis in [Donohue 2004] might have changed by engaging in model comparison rather than just tabulating outcomes under different model/specification combinations, an approach that effectively gives each model/specification combination equal weight. Taking (temporarily!) an M-closed view, the dogmatic belief that the true model is among the 24, the general approach would be to examine the results after averaging across the model/specification combinations. Because a single combination dominates for each crime, averaging is equivalent to looking at the results under the dominant model.

The posterior mean and standard deviation of the RTC coefficients turn out to be very similar to the frequentist estimates in [Donohue 2004].[63] The

[63]The reason is that the Bayesian approaches here result in a posterior mean and standard deviation that is similar to the corresponding OLS estimates except for a shrinkage factor, as is evident from equation (10) and the surrounding discussion above. Under CML, that factor is $\hat{c}/(1+\hat{c})$ where $\hat{c}$ is the empirical Bayes estimate that emerges from the data. In all of the regressions, $\hat{c}$ turns out to be at least equal to the number of observations, $n = 1150$, the value that would obtain if we fixed $g = 1/n$ in a natural conjugate g-prior setting, and in many cases is much larger. There is a separate $\hat{c}$ estimate for each model/specification combination. The actual range over the 216 combinations of crime category, model and RTC specification is from 2673 to $753,245$ with a mean of $167,761$. As a result, the shrinkage factor is very close to 1 for all of the regressions, and the posterior mean and standard deviation are close to the corresponding OLS quantities.

difference after adding a model comparison step is that we can focus on the results under the dominant model/specification combination for each crime rather than just facing different outcomes under alternative combinations.

For the general category of violent crime, the state specific spline specification of the Modified Lott model dominates. There are around five coefficients that are strongly negative in the sense that the posterior is concentrated in negative regions. On the other hand, there are ten that are strongly positive, suggesting an increase in violent crime in those states as a result of RTC laws. For the four components that make up the violent crime category, the outcomes differ:

1. **Murder**: The dummy with state trends specification of the Modified Lott model dominates. The coefficient on the RTC dummy has a posterior mean of about .026, but the posterior probability that the coefficient is negative is not trivial, around 12%.

2. **Rape**: The dummy with state trends specification of the Modified Lott model dominates. The posterior mean for the RTC dummy coefficient is $-.0339$, and the posterior probability that the coefficient is negative is 98.6%.

3. **Robbery**: The state specific spline specification of the Modified Lott model dominates. There are 6 RTC coefficients that are strongly negative and 7 that are strongly positive.

4. **Aggravated Assault**: Here there was a split between BIC and CML and the assessments about the impact of the RTC laws differ for the two approaches. Under BIC, the state specific spline specification of the Modified Lott model dominates. There are 2 RTC coefficients that are strongly negative and 8 that are strongly positive. Under CML, the dummy with state trends specification of the Modified Lott model dominates. The posterior mean for the RTC dummy coefficient is $-.0302$, and the posterior probability that the coefficient is negative is 95.2%.

For the general category of property crimes, the dummy with state trends specification of the Modified Lott model dominates. The posterior mean for the RTC dummy coefficient is .0157, and the posterior probability that the

This outcome, however convenient, is not inevitable.

coefficient is positive is 98.1%. The posteriors for the three components that make up the property crime category suggest positive responses (crime increase) to RTC laws in various degrees:

1. **Burglary**: The dummy with state trends specification of the Modified Lott model dominates. The posterior mean for the RTC dummy coefficient is .0122, and the posterior probability that the coefficient is positive is 84.7%.

2. **Larceny**: The dummy with state trends specification of the Modified Lott model dominates. The posterior mean for the RTC dummy coefficient is .0326, and the posterior probability that the coefficient is negative is less than 0.003%.

3. **Auto Theft**: Here there was a split between BIC and CML and the assessments about the impact of the RTC laws differ for the two approaches. Under BIC, the state specific spline specification of the Modified Lott model dominates. There are 8 RTC coefficients that are strongly negative and 8 that are strongly positive. Under CML, The dummy with state trends specification of the Modified Lott model dominates. The posterior mean for the RTC dummy coefficient is .0586, and the posterior probability that the coefficient is positive is 99.9%.

A final task remains with respect to the 24 model/specification combinations: assessing the strength of the hypothesis that the RTC dummy or set of dummies is zero. I address this task by applying the model comparison to an expanded set of model/specification combinations, adding 12 "RTC-null" model/specifications consisting of three different RTC specifications for each of the four models:

1. **"null"**: no RTC dummies or state trend variables.

2. **"null_sttr_ag"**: no RTC dummies but inclusion of the aggregate state trends variable;

3. **"null_sttr_sp"**: no RTC dummies but inclusion of the 26 state specific state trends variables.

Two aspects of this exercise deserve mention. First, in the state specific specifications, it treats the 26 RTC dummies or spline variables as a set

rather than moving them in and out of the specification one-by-one. Comparing models with the 26 variables in or out as a group is analogous to the frequentist approach of using an F test with the null hypothesis that all 26 coefficients are zero. Second, the model comparison exercise is more general and more meaningful than frequentist t-test or F test approaches. In particular, models with RTC dummies or spline variables present can be "defeated" by models other than the model that is identical except for omission of those dummies or variables.[64]

The results are less straightforward than for the model comparison without "RTC-null" specifications because in some cases there is no dominant model. The following table reports results including all models that achieve at least 0.01% (rounded) probability in each instance:

[64]For example, an "RTC-null" model that includes the 26 state specific state trends variables but no RTC variables of any kind might dominate a specification with the aggregate RTC dummy and *aggregate* state trends variable present. This comparison is not analogous to a t-test on the RTC dummy variable in the second model. To be analogous, the first model would have to include the *aggregate* state trends variables instead of the 26 state specific state trend variables.

Model/Specification Probabilities under BIC and CML Panel Data Regressions including RTC-null Specifications (under equal probability reference prior) (RTC-null outcomes in **bold**)		
crime category	BIC	CML
violent	ML: spl–st_sp 100%	ML: spl–st_sp 100%
murder	**ML: null_sttr_sp 94.37%** ML: st_tr–agg 5.63%	**ML: null_sttr_sp 99.63%** ML: st_tr–agg 0.37%
rape	**ML: null_sttr_sp 77.17%** ML: st_tr–agg 22.83%	**ML: null_sttr_sp 97.87%** ML: st_tr–agg 2.13%
robbery	ML: spl–st_sp 100%	ML: spl–st_sp 100%
agg. assault	ML: spl–st_sp 100%	**ML: null_sttr_sp 99.84%** ML: st_tr–agg 0.16%
property	ML: st_tr–agg 93.63% **ML: null_sttr_sp 6.37%**	**ML: null 99.85%** ML: dum–agg 0.15%
burglary	**ML: null_sttr_sp 95.19%** ML: st_tr–agg 4.79% ML: spl–st_sp 0.02%	**ML: null_sttr_sp 100%**
larceny	ML: st_tr–agg 99.12% **ML: null_sttr_sp 0.88%**	ML: st_tr–agg 99.95% **ML: null_sttr_sp 0.05%**
auto theft	ML: spl–st_sp 99.95% ML: st_tr–agg 0.04% **ML: null_sttr_sp 0.01%**	ML: st_tr–agg 87.21% **ML: null_sttr_sp 12.79%**
models: ML = Modified Lott; DL = Donohue/Levitt; SP = Spelman; ZH = Zheng		

RTC specifications: dum = dummy alone; spl = spline;
st_tr = with state trends; agg = aggregate (1 RTC variable);
st_sp = state specific (26 RTC variables)
null = null, no state trends;
null_sttr_ag = null + aggregate state trends variable
null_sttr_sp = null + 26 individual state trends variables		

It is apparent from the table that under CML, the RTC-null specifications of the Modified Lott model are dominant or near dominant for five of the nine crimes and are a contender for a sixth. RTC-null specifications are the leading model for three of the nine crimes under BIC. Although CML is

an empirical Bayes measure and BIC an asymptotic approximation, corresponding fully Bayesian treatments are similar or support RTC-null models even more comprehensively across crime categories.[65] A tentative conclusion emerges: the models that appear to have the strongest predictive power with respect to several crime rates exclude RTC variables entirely. This conclusion becomes even stronger when we allow a wider spectrum of models as discussed in the next section.

3.5 Expanding the Class of Potential Models

The four panel data models considered in [Donohue 2004] are highly specific. Each involves a distinct and very detailed specification. The models include different explanatory variables representing demographics, poverty, unemployment, police, prisons/prisoners, and population density. Some models include variables or variable groups that are absent from the others: for instance various abortion rate variables in Donohue-Levitt, and alcohol consumption and a group of political variables in Zheng. It is highly implausible that these exact four complex specifications represent the universe of plausible ones.

One posture in the face of this situation has an M-closed flavor: We believe that the researchers have identified the correct set of variables, but we are uncertain of the exact subset that represents the "true model." This subset may cut across the four specific models, e.g., combining the demographic variables from the Modified Lott model with the Donohue-Levitt poverty variable. If we are totally agnostic about the right subset, our prior might be that each possible subset is equally likely. This position would dictate comparing all possible subsets under the equal probability "reference prior" employed in some of the examples above. The total number of variables is large, 230 to be precise, or 221 after eliminating variables that are identical or nearly so. The number of possible subsets, each one representing a distinct candidate specification, is huge, $\sim 2^{221}$.

Model comparison or model averaging in the face of such a large num-

[65] The unit information natural conjugate prophetic prior case is almost identical to BIC. The natural conjugate prophetic prior case using $\hat{c}$ and $\hat{w}$ from CML is very similar to CML. Using zero priors in each instance results in stronger support for RTC-null models. In the unit information case RTC-null variants are the leading model for eight out of nine crime categories. In the CML case, RTC-null variants are the leading model for all nine categories.

ber of possibilities is not necessarily infeasible. Although the time required to compute marginal likelihoods or model probabilities for all 2^{221} possibilities almost certainly would be prohibitive, there often are simulation approaches capable of locating most or all of the high probability models in a reasonably short time, leaving negligible probability across the unexplored alternatives.[66] Even if such a simulation is feasible, however, there is a potential problem identified by [George 1999] and discussed extensively in [Chipman, George & McCulloch 2001, pp. 78-79] that is highly relevant to the analysis of the RTC models here. Expanding a group of variables (e.g., variants representing "income") by adding many new highly correlated alternatives will generate a large number of very similar models involving the various combinations of variables from the group. These models all will have similar probabilities relative to alternatives, but the greatly increased number of the combinations will mean that the models as a group will have much higher combined probability at the expense of other possibilities than previously. This phenomenon can result in biasing model averages away from the alternatives which may include the "good models." Although this bias would wash out asymptotically, it is not wise to neglect it in actual applications. The danger is particularly acute for variables from the models considered in [Donohue 2004] since many of them are highly correlated. For instance, there are 36 highly correlated demographic variables in the Modified Lott specification. If these 36 variables are treated separately, there are 2^{36} within-group combinations!

A solution is to use so-called "dilution priors" to prevent highly correlated groups from receiving undue influence. A natural way to implement that approach in the current context is to divide the 221 variables into groups. E.g., if there are several alternatives representing "income," we can restrict the possible models to those that contain one of the alternative variables or none, ruling out instances where more than one income variable enters into a model by adopting a (dogmatic) dilution prior assigning zero probability to such models. The results that follow employ 12 such groups covering all 221 variables:[67]

[66] One popular approach is to use the "MC3" algorithm developed by [Madigan & York 1995]. This algorithm uses a Metropolis-Hastings step to sample from the model space in a way that the frequency of draws converges to the posterior probabilities of the associated models. [Fernandez, Ley & Steel 2001] and [Koop 2003, ch. 11] provide clear explanations and some nice examples of this approach.

[67] As mentioned above, the total number of variables was 230, but I removed 9 that are

1. demographics: three groups of variables (the groups from the Modified Lott, Spelman, and Zheng models);

2. poverty: three variables (alternatives from the Modified Lott, Zheng and Donohue-Levitt models);

3. unemployment: three variables (alternatives from the Modified Lott, Spelman and Donohue-Levitt models);

4. police: three variables (alternatives from the Zheng, Spelman and Donohue-Levitt models);

5. prison: two variables (alternatives from the Modified Lott and Donohue-Levitt models);

6. population density: two variables (alternatives from the Modified Lott and Zheng models);

7. income: two variables (alternatives from the Spelman and Zheng models);

8. state population: one variable (from the Modified Lott model);

9. abortion: one variable, three variants (depending on crime category) (from the Donohue-Levitt model);

10. alcohol consumption: one variable (from the Zheng model);

11. political variables: one group (from the Zheng model);

12. RTC and RTC-null specifications: eight groups that include at least one variable.

These groupings (plus the alternative with respect to each group of having no variables from the group in the model) result in a little less than a million possible specifications, a large number, but much smaller than 2^{221}. It turns

identical or nearly identical to others. The removed variables along with their highest correlation with a remaining variable include: Zheng prison (.9962); Spelman prison (.9915); Spelman population density (1.0000); Modified Lott income (.9932); Donohue-Levitt income (.9925); Zheng unemployment (.9997); Zheng, Spelman and Donohue-Levitt state population (1.0000). Appendix A details the variables included in each of the four models, broken down into the twelve groups.

out to be feasible to compute marginal likelihoods and model probabilities for all models – more elaborate simulation methods are not necessary.[68] This grouping method should minimize problems arising from treating too many highly correlated variables separately. It also is a conservative approach to exploring alternative specifications since selecting within each of the twelve groups to construct the alternatives creates models that are similar in spirit and structure to the original four.

The interpretative approach for the results necessarily differs from the approach for the model comparison exercise in the previous section. Hundreds or thousands of models may contribute significant probability to the average rather than the 36 models in the comparison exercise. Highest probability models often involve only around 10% of the total probability, and sometimes much less. A meaningful way to assess how various RTC specifications and other independent variables fare is to compute how much total probability each specification or variable receives across all models.[69] For the RTC specifications, the results are as follows, reporting only specifications that receive at least 1% of the total probability:

[68]For each crime category and eight variants including CML, BIC and six different natural conjugate approaches, computations require about a day on a PC with a $3Ghz$ CPU.

[69]To speed up the computations in the actual exercise, I retained only the $24,000$ highest probability models, using them to calculate reported values. I also computed the total probability of the remaining models to ensure that it was negligible. The highest total probability for the remaining models was $< 10^{-8}$ for the BIC runs and $< 10^{-26}$ for the CML runs, virtually ensuring no effect on any of the results up the number of significant digits reported.

RTC Specification Probabilities under BIC and CML Bayesian Model Averaging (under equal probability reference prior) (RTC-null outcomes in bold)		
crime category	BIC	CML
violent	spl–st_sp 100%	**null_sttr_sp 100%**
murder	**null_sttr_sp 87.20%** spl–agg 12.15%	spl–agg 97.87% **null_sttr_sp 1.77%**
rape	**null_sttr_sp 100%**	**null_sttr_sp 100%**
robbery	spl–st_sp 100%	spl–st_sp 100%
agg. assault	spl–st_sp 99.92%	**null_sttr_sp 100%**
property	**null_sttr_sp 100%**	**null 100%**
burglary	**null_sttr_sp 98.99%**	**null_sttr_sp 98.94%** st_tr–agg 1.06%
larceny	**null_sttr_sp 100%**	**null_sttr_sp 100%**
auto theft	spl–st_sp 100%	**null_sttr_sp 100%**

RTC specifications: dum = dummy alone; spl = spline;
st_tr = with state trends; agg = aggregate (1 RTC variable);
st_sp = state specific (26 RTC variables)
null = null, no state trends;
null_sttr_ag = null + aggregate state trends variable
null_sttr_sp = null + 26 individual state trends variables		

Many of the winning specifications are similar to the winners in the model comparison exercise in the previous section, but the predominance of RTC-null variants is much stronger.[70] Consider the picture based on taking CML as definitive. RTC-null variants dominate or nearly dominate in seven out

[70]This phenomenon extends to the purely Bayesian natural conjugate approaches not reported here in favor of CML and BIC. As usual, with prophetic priors, the results are similar to the corresponding CML and BIC results. The zero prior results bear much more resemblance to the corresponding CML and BIC outcomes than in the model comparison

of nine crime categories, including some such as rape, larceny and auto theft that tended to have highly significant coefficients on RTC dummies in the [Donohue 2004] regressions. The exceptions are murder and robbery. For murder, the model averaged coefficient of the RTC spline dummy is $-.0146$. Only about .009% of the posterior distribution for the coefficient is above zero – the standard deviation of the distribution is about 0.0039. What appears is a modest but fairly certain negative effect of the RTC laws on murder. For robbery, since a state specific variant predominates, there are 26 relevant model averaged RTC coefficients. They scatter widely on both sides of zero, averaging 0.0049. For seven states, 95% or more of the posterior probability distribution is in positive territory and for six states 95% or more is in negative territory. This suggests that the RTC laws may have a deterrent effect in some states that enacted them but by no means in all or even a predominance.

I would approach these results with considerable trepidation because I would not put much prior weight on the possibility that we are in an M-closed environment. It seems likely to me that various important variables and factors are missing, with the unevenly geographically-distributed crack cocaine epidemic mentioned in [Donohue 2004] being just the tip of the iceberg. On the other hand, the results here provide much stronger evidence of a weak role for the RTC variables in a predictive environment than is possible from simply examining the outcomes under the four panel data models in [Donohue 2004]. First, the model comparison aspect reduces the risk that we will be influenced by the outcomes of models, however numerous, that are weak in a predictive sense. Second, the ability to consider a wide range of possible specifications in a systematic way obviates the need to rely on a few highly specific and complex models where the rationale for many of the detailed aspects is not clear. Allowing a range of treatments for these aspects increases confidence in the results.

These conclusions are strengthened when we examine how the four panel data models would fare (each with 9 total combinations involving nine different RTC specifications) if thrown into the mix. As discussed in a previous footnote, I retained only the $24,000$ models with highest probability out of roughly a million for each crime category in the more general exercise. This move eases computation and does not impact the results given that the total

exercise, but they still display even more tendency toward RTC-null predominance than the CML and BIC outcomes.

probability of the excluded models is very small. One way of addressing the strength of the panel data models is to ask where the highest probability panel data model/specification combination for each crime category would rank among the top $24,000$ from the general exercise. The answer is stunning and is the same for all nine crime categories: *the highest probability panel data model would not rank among the top 24,000.* In fact, none of the highest probability panel data model/specification combinations even comes close. Each one would have positive but negligible probability if added to the more general exercise.[71]

Why do the highest probability variants of the four panel data models do so poorly? The answer is evident from examining the nature of the high probability models in the more general exercise. As detailed in Appendix B, these models almost always include the applicable Donohue-Levitt abortion rate variable but also typically contain many more variables than the rather sparse Donohue-Levitt panel data model. By simply dividing the available variables into twelve coherent groups and considering "menu" selections from the twelve groups, a multitude of combinations much stronger in a predictive sense than the four panel data models emerge. It is important to emphasize how conservative the approach is. Within each of the twelve groups, the choice among the variables seems quite contestable as a matter of modeling. We are not considering models containing random collections of possibly highly correlated and redundant variables or models that lack theoretical coherence compared to the four panel data models.

The extreme fragility of a frequentist approach that focuses on one or a handful of complex models that differ across a multitude of contestable dimensions is evident here. *It is highly unlikely that the handful of models considered contains the most salient ones in a predictive sense. As a result, much of the ensuing analysis or dispute arguably will be a waste of time – an argument about how outcomes differ under a set of very weak specifications.* In light of the results here, this characterization seems quite apt with respect to the whole dispute in the literature about the impact of the RTC laws on crime. Using Bayesian methods greatly broadens the group of models considered, creating flexibility across all of the contestable choices.

It is clear that Bayesian model comparison and model averaging add

[71] Here "negligible" means less than 2.2251×10^{-308}, the smallest real number cognizable in MATLAB. In contrast, the probabilities of the $24,000^{th}$ model for the nine crime categories range from 3.62×10^{-30} to 1.05×10^{-76}.

considerable value when the game is assessing alternative specifications based on predictive power. It is a great tool in the hands of skeptics faced with very complex competing models promulgated by true believers of various stripes. It is much harder to manipulate results via the details when a process is available to gauge the impact of systematically considering variations in the details that are contestable. However, it is important to keep in mind that these approaches are not panaceas. As noted by [Cremers 2002], deciding on the global set of variables to be considered in the comparison or averaging exercise is a new dimension where manipulation and unconscious bias can take hold. This situation exists because adding or subtracting variables from the global mix from which the candidate models are generated may affect the results. We also have seen that there is some leeway with respect to how the variables are parsed. Adding one or more groups of very similar, highly correlated variables can draw away probability from specifications that include variables orthogonal to the added variables. Addressing this situation via "dilution priors" or similar strategies has elements of art in it – there is no one right way to do it that would leave no scope for manipulation or unconscious bias. In sum, Bayesian model averaging and comparison approaches are very helpful, but they will not cure all ills by themselves.

The limitations of the model averaging and model comparison approaches underscore more general themes. A Bayesian framework adds only one element to empirical analysis: It ensures logical consistency about probability assessments. This element is quite valuable since it promotes clear thinking. But it cannot create magical "neutral" machinery that overcomes all possible manifestations of conscious or unconscious bias. In a sense, the lesson that emerges from the logical discipline inherent in Bayesian analysis is the opposite one: It reminds us of the key role of prior beliefs, and the consequent difficulty or even impossibility of any real "neutrality," in the absence of infinite pool of data and variables. Frequentist approaches are even more limited on this score since there is a single prior, often unexamined, lurking beneath each frequentist exercise. At its best, Bayesian analysis spurs the researcher and the audience to face the prior choices inherent in the given exercise.[72]

[72]There is an exchange about Bayesian model selection in the literature that, I believe, read as a whole reflects this position. Adrian Raftery has been one of the leading figures in developing Bayesian model averaging and related techniques. Much of discussion here concerning the rationales for Bayesian model averaging and also the critique of p-values in earlier sections is present in or derives from Raftery's work. Some parts of this article

3.6 Alternative Approaches to Specification Sensitivity

The Bayesian (and empirical Bayesian) approaches discussed above are certainly not the only techniques available to address specification sensitivity. There are many others in both the frequentist and Bayesian literatures. Although a comprehensive discussion is far beyond the scope of this article, it is worth making a few points that add some important perspective.

As a preliminary matter, consider why the choice of specification matters. The central concern underlying the work is the impact of the RTC laws on crime. A multiple regression approach can reveal a (not necessarily causal) relationship between indicators of the presence or duration of these laws and crime rates conditional on a set of independent variables. If the RTC indicators are related to the other independent variables, empirical estimates of this relationship will depend on the specification chosen. To be more precise, consider the picture that emerges under OLS. If the RTC indicators are orthogonal to all of the other independent variables, then the estimated regression coefficients for the indicators will be the same regardless of which of the other variables are included or excluded. The OLS formula for the

---

are very similar to [Raftery 1995a] which emphasizes the inadequacy of p values and the utility of Bayesian model averaging. One of the numerical examples in [Raftery 1995a] is law-related, examining Isaac Ehrlich's work on the deterrent effect of punishment. [Raftery 1995a] led to an interesting exchange with Andrew Gelman and Donald Rubin, prominent statisticians who, among other accomplishments, have played a very significant role in developing various aspects of Bayesian statistics. [Gelman & Rubin 1995] stress the importance of the context of the problem being examined and argue that, there may be cases where fitting a single complicated model ("probably using Bayesian methods") is a better approach than model selection based on Bayesian model averaging or related methods. They also suggest that in the Ehrlich application "hierarchical modeling might be more compelling." As discussed above and in some of Gelman and Rubin's other work, hierarchical modeling has a model comparison element. [Raftery 1995b] in replying to Gelman and Rubin stresses that social research typically involves some form of model selection, and that Bayesian model averaging often can be useful in that regard. A takeaway that none of the parties would disagree with is that there is no universally applicable Bayesian approach. Model averaging, hierarchical models or even a single complex model may be the best approach. A big part of the choice depends on prior information and beliefs. For instance, with well-developed prior information that points strongly toward a single complex model, it would be foolish (and, in fact, wrong in a Bayesian sense) to use a "neutral" approach that allocates equal prior probability to a wide range of variants based on the variables in the complex model.

sampling variance of a particular coefficient estimate, $\hat{\beta}_j$, is:

$$\text{var}\left(\hat{\beta}_j\right) = \frac{\sigma^2}{\sum x_j^2}\left(\frac{1}{1-R_j^2}\right) \tag{12}$$

where σ^2 is the true variance of the error term, $\sum x_j^2$ is the sum of the squared deviations of the jth independent variable from its sample mean, and R_j^2 is "R^2" from the auxiliary regression of X_j, the jth independent variable, on all of the other independent variables. If $R_j^2 = 0$, then X_j is orthogonal to the space defined by the other independent variables. As a consequence, the estimated coefficient for X_j will not depend on what collection of other independent variables are present in the regression. On the other hand, if $R_j^2 > 0$, then the estimated coefficient for X_j will depend on the specification and the variance of that estimate will be "inflated" by "VIF", the so-called "variance inflation factor":

$$VIF_j = \left(\frac{1}{1-R_j^2}\right). \tag{13}$$

The higher the value for VIF_j, the greater the multicollinearity problem arising from the association of X_j with the other independent variables. A high VIF_j means a high variance for the estimated coefficient indicating uncertainty about whether X_j influences the dependent variable or whether the influence is through the other independent variables. The reported standard error for X_j will be inflated by the factor $\sqrt{VIF_j}$.[73] One common rule of thumb is that a VIF of 10 or more for any independent variable in the regression indicates a significant degree of multicollinearity. High values for the VIFs of the RTC indicators would be particularly serious since these high values would indicate enhanced uncertainty about the value of coefficients of the indicators due to substantial entanglement with the other independent variables. The standard errors for the estimated coefficients of the RTC indicators will be inflated due to this entanglement. As a result, these coefficients (as well as the coefficients of the independent variables entangled with the RTC indicators) may be statistically insignificant even though the

[73]The estimated standard error is not given directly by equation (12) unless σ^2, the true variance of the error term, is known. Under OLS, typically only an estimate of σ^2 is available, replacing the actual σ^2 in equation (12). This estimated σ^2 may be sensitive to the specification even if $R_j = 0$. The variance inflation factor, however, remains the same. Thus, the estimated standard error is inflated by the square root of this factor.

RTC variables in fact have a big impact. I.e., there is a danger that the multicollinearity surrounding the RTC indicators will hide a real relationship. In addition, the coefficient estimates that emerge will be unreliable, as is correctly indicated by the elevated estimated standard errors.

Reporting VIFs for key variables is all too rare in the legal empirical literature. The reader is left in the dark as to whether the coefficient estimates for these variables are plagued by multicollinearity. Of the panel data specifications in [Donohue 2004] that emerged as winners under CML and BIC in the model comparison exercises were three that included RTC indicators. All three were variants of the Modified Lott model. Two of them were aggregate models with only one RTC indicator: the general RTC dummy variable. One was the aggregate model without state trends while the other included state trends. The VIFs for the RTC dummy are 4.55 and 6.92 in these models respectively. These VIFs are not trivial, implying inflation of the standard errors of the coefficient estimates by factors of 2.13 and 2.63 respectively. The third variant was the spline model with 26 state specific RTC dummies. VIFs for these 26 dummies range from 1.91 to 153.87 with a mean VIF of 13.58. In the reduced data set used in the Bayesian model averaging exercise, the VIF for the RTC dummy is 7.38 (square root = 2.71) after excluding all the other RTC variants. It is clear that substantial multicollinearity surrounds the RTC indicators. As a result, the coefficient estimates for these indicators will be sensitive to the specification and the estimated standard errors of the estimates will be inflated.

The Bayesian model comparison and averaging approaches applied here address this problem by choosing the best predictive combination of specifications. The touchstone is to come as close to the true random process that generates each dependent variable as possible. This goal is consistent with one major strand in [Donohue 2004]: choosing good predictive crime models, inserting RTC indicators, and seeing what coefficient estimates emerge for those indicators.

Two sets of comments are germane to this predictive approach. First, there are tools that readers may be familiar with and that populate some widely used statistics and econometrics packages designed to accomplish the same end. Three such tools are ridge regression, principal components analysis ("PCA") and the lasso. PCA reduces the independent variables to an orthogonal set of principal components and then (possibly) drops one or more principal components with low eigenvalues. The principal component with the lowest eigenvalue represents the direction (in the space generated by the

independent variables) with the least variation. Ridge regression shrinks the independent variables along the principal components with the most shrinkage along the principal component directions with the smallest eigenvalues. It is equivalent to minimizing the sum of squared residuals subject to an upper bound on the sum of the squares of the coefficient estimates (typically after standardizing the independent variables). The lasso bears similarities to ridge regression, being equivalent to minimizing the sum of squared residuals subject to an upper bound on the sum of the absolute values of the coefficients. Unlike ridge regression, the lasso can result in variables being dropped once the shrinkage proceeds far enough.

As discussed in [Hastie, Tibshirani & Friedman 2001, pp. 70-72], ridge regression, the lasso, and best subset selection (choose the best model with a fixed number of independent variables or fewer) can be conceptualized as Bayes estimates with different prior densities for the coefficients. (Principal component analysis may be viewed as crude form of ridge regression and as related to best subset selection.) More precisely, the conceptualization takes the log prior for coefficient β_j to be proportional to $|\beta_j|^q$ where $q = 0, 1, 2$ respectively for best subset selection, the lasso and ridge regression. The frequentist coefficient estimates that emerge from these techniques are modes of the posterior distributions for the coefficients. Once again we have an example where each frequentist method is equivalent to a Bayesian method with a particular prior.

The second set of comments strike directly at the predictive approach. Arguably, the goal is not to predict crime rates (the dependent variables) as well as possible, but to assess the impact of RTC laws on the crime rates. Taking this view, using some kind of "treatment model" is indicated. The idea is to compare outcomes when a treatment is applied (the RTC law) to outcomes where the treatment is not applied. A second major strand in [Donohue 2004] takes exactly this approach. The ADS specifications explicitly incorporate treatment modeling.

The ideal treatment-based approach would be to run an experiment in which outcomes for each subject (in the RTC context, states or state/year combinations involving particular values for the independent variables) are observed in both the treated and untreated condition. Unfortunately, for most legal applications we are stuck with observational data. E.g., we cannot rerun history with some states taking the opposite position on the RTC laws than they actually did.

There are many approaches that address this situation in a treatment

model context. One of them, "preprocessing matching" is of particular interest here because it is aimed directly at specification problems. This approach is described cogently in [Ho, et.al. 2005]. In the absence of having experimental data, one can cull the observational data so that treatment and control groups closely resemble each other. Exact matching would result in paired observations that are identical across all independent variables except for the treatment variable. The remaining data would then resemble an idealized experiment.[74] Exact matching is often unattainable, especially in the face of continuous rather than discrete independent variables or where there are a very large number of variables. There may be no matching pairs at all.[75] In these situations, as described in [Ho, et.al. 2005], there are a number of approaches available. The ultimate goal is to achieve "balance," which obtains completely when the sample distribution of each independent variable conditional on treatment is the same as the distribution conditional on no treatment. "Balance" has an obvious connection with the multicollinearity discussion above. If balance obtains for all independent variables, then each of the variables will be orthogonal to the treatment variable.[76] In a regression context, the coefficient estimates for the treatment variable will be unaffected by the specification. One may as well leave out the other independent variables.

[74]One important difference would be that the paired observations in the paired matching exercise involve different agents or cases. Although these agents or cases are identical up to all of the independent variables that are observed, they may differ in ways that are not observed or observable and these differences might affect the outcome (dependent variable) in a way that will confound the treatment effect. In an idealized experiment, this problem does not exist because the same agent or case would be subjected to both the treatment and control regimes. Any unobserved differences would be constant across regimes and would not be correlated with the treatment variable. In an actual experiment, random assignment to the treatment group along with a large enough number of observations serves as a statistical approximation of the idealized experiment. With observational data, it is not possible to use random assignment to achieve this statistical approximation. One is stuck with the possibility of confounding by unobserved factors.

[75]As is the case in many panel data situations, the RTC data involves a unique set of independent variables for each observation. Observations range over fifty states plus the District of Columbia and twenty-three years. At the same time, the independent variables include state and year dummy variables. These dummy variables will be unique for each observation. Exact matching will not be possible, and the presence of these dummy variables also will make it more difficult to achieve "balance" in the alternative (non-exact) matching schemes discussed below.

[76]Balance is more comprehensive since it requires the higher order moments as well as the mean to be identical across the two conditional distributions.

Typically, perfect balance is unobtainable. As a result, [Ho, et.al. 2005] suggest a two step process: (1) preprocessing matching to reduce imbalance; (2) parametric estimation (e.g., regression) to attempt to address any remaining bias. They note that their two-step approach is "doubly robust." If either step works perfectly, then the specification problem is solved. If the steps are individually imperfect, they may bolster one another.

Where would an approach such as Bayesian model averaging fit in this scheme? [Ho, et.al. 2005, p. 16, n. 8] view it as an alternative to the whole scheme. However, if imbalance is left after step (1), the argument for using Bayesian model averaging in the second step is the same one as usual. Specification uncertainty remains, and the forthright approach is to reflect this uncertainty in any estimates for treatment effects.

It is clear that a treatment model type of approach to the question of the impact of RTC laws would be desirable. Treatment models may include Bayesian aspects or be set in a Bayesian framework. Approaches of this type are interesting but typically require extensive discussion and technical development. As a result, I leave them (as well as several other aspects) to a sequel.

4 Concluding Thoughts

Bayesian analysis has much to offer legal academics both conceptually and operationally. At the conceptual level, some common frequentist ways of proceeding have serious flaws, especially when used improperly. A striking example is p-value based hypothesis testing. Although it is textbook wisdom that p values are not probabilities of any hypothesis being true, it is quite common for legal academics to derive conclusions as if they were. This error would not be serious if p values approximated posterior probabilities, but in some important instances there are systematic and substantial deviations. In particular, the traditional two-sided test against the null hypothesis that a regression coefficient is zero greatly overstates the evidence against the null if the p value is treated as if it were a posterior probability. In many cases where that null hypothesis is "rejected" at the 5% level, the evidence actually suggests that one should increase one's prior probability that the null is true. The associated rule of thumb that t statistics of 2 or greater indicate "statistically significant" results suffers from the same problem since this rule of thumb embodies rejecting the null hypothesis of zero coefficients

at the 5% level.

Perhaps the best alternative approach would be to drop p-value based analysis and focus explicitly on posterior probabilities computed using Bayesian methods. It is clear, however, that readers who do not wish to go the full Bayesian route have readily available and helpful tools at their disposal. The BIC-based approximation discussed in the article, a simple formula, gives a sense of the probability implications of t statistic outcomes for regression coefficients. Even short of applying (or remembering!) this formula, it is clear that the probability effects sometimes mistakenly attributed to 5% level results and t statistics of around 2 for two-sided tests do not emerge until t statistics are greater than 3 or maybe even 4. At a minimum, legal academics, whether producers or consumers of empirical work, should no longer passively accept conclusions or arguments mistakenly built around p value analysis.

Gauging specification sensitivity by considering alternative predictive crime models is a great strength of [Donohue 2004]. Nonetheless, simply comparing results under a few leading models is not very satisfying. There clearly are thousands if not millions of plausible specifications even limiting consideration to the sets of variables in the models examined. Any analysis that examines only a few models leads to the fear that the results are idiosyncratic even in the absence of unconscious or conscious bias that motivates the choice. Furthermore, simply listing results under various specifications is inadequate given that some specifications may have much more explanatory power than others. The Bayesian palette of methods includes some very powerful tools for model comparison and specification analysis. Using some of these tools, we considered a very broad range of models similar to the four that are the focus in [Donohue 2004]. In place of p value analysis, this consideration included various "RTC-null" models, specifications without RTC indicators present. The results are dramatic. None of the models in [Donohue 2004] rank highly against the alternatives. In addition, the results for specific crime categories are substantially different.[77]

[77] A Bayesian perspective underscores a general conceptual point: there is no bright line between theory and empirical analysis. Uncertainty among various models representing alternative theories may be parameterized in a Bayesian framework in a way that is parallel to considering a regression coefficient as an unknown parameter. The bright line is elsewhere: between items that the researcher is (dogmatically) taking as known, including "the data," and items treated as unknown, i.e., endowed with a non-dogmatic prior distribution. [Lancaster 2004, p. 9] stresses that in a Bayesian framework, "[t]he only

The model comparison and averaging exercises here also illustrate another point: Bayesian analysis does not deliver an "objective" approach that circumvents the need to take into account prior beliefs. Instead, the whole point is to make prior beliefs very explicit and ensure that the ensuing probability assessments are logically coherent. The Bayesian model comparison methods discussed and applied in section 3 are a case in point. These methods clearly seem to have greater robustness and arguably are much less prone to hidden researcher bias than reliance on a single selected model, but they depend on multiple choices with respect to prior distributions about which researchers may easily disagree. Perhaps most salient is the assignment of prior probabilities to particular models. In subsection 3.5, the prior assigned equal probabilities to models combining a single variable or single group of variables from each of twelve sets of such variables or groups.[78] This choice seems reasonable (to the author) since it generates a set of models most of which are quite similar to the four that are the focus in [Donohue 2004] and therefore hard to exclude as equally weighted possibilities a priori. Another feature of the choice is more disconcerting. It implicitly puts zero prior weight on models that might be constructed using variables not in the collection generated by the four models. Taking these results as conclusive implies an M-closed stance, a dogmatic prior belief that all of the relevant variables are present. My view is that such a stance is unwarranted: The results are interesting, but it is likely that important variables are missing.

Bayes' rule ensures logical consistency but puts no restriction on the choice of priors. It is easy to imagine plausible priors other than the one adopted in section 3.5 for the model averaging exercise. Going even further, it is possible to hold prior beliefs that require rejecting the model averaging exercise entirely. One could assert that the four models considered by [Donohue 2004] are the only ones that should have non-zero prior probability or could limit examination to a single complex model, assigning zero prior probability to all other possibilities. Researchers adopting such priors may have very good reasons or strong intuitions that support their beliefs. More generally, for any particular problem, there are a multitude of approaches consistent with Bayes' rule. Fervent Bayesians may disagree strongly about which approaches are best for addressing a particular empirical task. Typi-

distinction between objects is whether they will become known for sure when the data are in, in which case they are data (!); or whether they will not become known for sure, in which case they are parameters."

[78] For each set, omitting the set from the specification also was an option.

cally, these disagreements hinge on choosing different priors for some aspect or aspects of the task. The advantage of Bayesian analysis is that it makes these choices explicit in the face of the fact that no one choice can be said to be "objective" or "right."

This article has limited goals. It focuses on some applications that illustrate the value and nature of Bayesian analysis and has emphasized certain Bayesian and pseudo-Bayesian tools that are particularly accessible.[79] No attempt was made to go the whole route by applying the author's most favored Bayesian approach to analyze the right-to-carry laws. Nonetheless, some of the results on the RTC front move the ball forward. The strength of the RTC-null models in the model averaging exercise for seven of the nine crime categories is impressive and strengthens the viewpoint that RTC laws may make no appreciable difference one way or the other. These negative results occur for some crime categories, such as rape, larceny and auto theft, where a narrower analysis (focusing on a few models) indicated non-trivial effects. The hierarchical model results suggest considerable unexplained state variation in the impact of the laws, favoring a position closer to independent effects across states than to a single common effect.

To sum up, there are two overarching advantages of Bayesian thinking and methodology for legal researchers. First, the final product of the methods are posterior probability distributions about quantities of interest. With these distributions in hand, researchers can address the normative and positive issues of concern directly and naturally. Second, making prior beliefs and the dependence of the results on those beliefs explicit is important in a sphere characterized by sharp disputes and strongly held judgments.

It is important to emphasize that we have only scratched the surface here. The range and power of Bayesian approaches to empirical problems are impressive and are growing rapidly. The reader interested in these approaches should consider this article the barest introduction. At the same time, it should be clear from what we have considered that Bayesian thinking and methods have much to offer. Both producers and consumers of legal empirical work stand to benefit substantially. Yes, legal empiricists should go Bayesian.

[79]BIC and CML, for example, are pseudo-Bayesian approaches that are easy to implement with a few lines of code in frequentist statistical packages. Although a full Bayesian approach is typically preferred, these shortcuts offer some tools to those researchers who do not want to devote the time or resources required for more pure approaches.

A Appendix A: Variable and Model Description

[Donohue 2004, p. 641] contains a table summarizing the variables in the four models considered as alternatives. This appendix describes the variables in somewhat more detail (not possible in a table subject to reasonable aesthetics) and breaks them down across the twelve groups used in the Bayesian model averaging exercise as described in subsection 3.5. The appendix begins by describing the variables comprising each of the four models and ends with a description of the alternative RTC specifications. For each model, potentially there are variables in eleven of the twelve groups. The twelfth group contains the RTC specifications.

A.1 Modified Lott

The Modified Lott model contains the following variables:

1. demographics: percentage of state population for 36 race-age-gender categories = 6 age ranges $\times$ 3 race categories $\times$ 2 genders; race includes white, black and neither white nor black; the age ranges are 10-19, 20-29, 30-39, 40-49, 50-64, 65 and over;
2. poverty: real per capita income maintenance;
3. unemployment: real per capita unemployment insurance payments;
4. police: none;
5. prison: lagged incarceration rate per 100,000 state residents;
6. population density: population per square mile;
7. income: real per capita personal income;
8. state population: US Census state population;
9. abortion: none;
10. alcohol consumption: none;
11. political variables: none.

A.2 Donohue/Levitt

The Donohue/Levitt models contain the following variables:

1. demographics: none;
2. poverty: percent of population below poverty line;
3. unemployment: percent unemployed;
4. police: lagged log number of police per capita;
5. prison: lagged log number of prisoners per capita;
6. population density: none;
7. income: log of income per capita;
8. state population: none;
9. abortion: abortion rate, three variants (depending on crime category);
10. alcohol consumption: none;
11. political variables: none;

A.3 Spelman

The Spelman model contains the following variables:

1. demographics: percentage black plus percentages in four age ranges: 0-14, 15-17, 18-24, 25-34;
2. poverty: none;
3. unemployment: log of unemployment rate;
4. police: log of police per capita;
5. prison: log of lagged rate of sentenced prisoners per 100,000 residents;
6. population density: percentage urban;
7. income: log of real per capita income;

8. state population: none;
9. abortion: none;
10. alcohol consumption: none;
11. political variables: none;

A.4 Zheng

The Zheng model contains the following variables:

1. demographics: percentage black plus percentages for three age ranges: 15-17, 18-24, 25-34;
2. poverty: percentage of persons below poverty line;
3. unemployment: unemployment rate;
4. police: lagged police per capita;
5. prison: lagged prisoners per capita;
6. population density: percentage urban;
7. income: real per capita income;
8. state population: US Census state population;
9. abortion: none;
10. alcohol consumption: per capita alcohol consumption;
11. political variables: four dummy variables indicating governor's party (democrat, republican, independent, other);

A.5 RTC Specifications

There are three basic RTC specifications, each of which has an aggregate version and a state specific version:

1. dummy alone: in the aggregate version, a dummy variable indicating the presence of a RTC law on the books for states that adopted such laws during the sample period – 0 for the year of adoption and previous years, 1 for future years; 0 for states not adopting RTC laws during the sample period; in the state specific version, 26 separate dummy variables for the 26 states that adopted RTC laws during the sample period;

2. spline: same as dummy alone, except that the variable indicates number of years since adoption of the RTC laws rather than being 1 in post-adoption years;

3. dummy with state trends: includes dummy or dummies from dummy alone specification plus aggregate or state specific "state trends" variables; the state trends variables indicate the number of years since 1976 for the 26 states adopting RTC laws during the sample period; zero for the other states; in the aggregate version, there is only one state trends variable; in the state specific version, there are 26 that sum to the aggregate one.

In addition to these specifications, some parts of the paper consider three separate null specifications:

1. no RTC, spline or state trends variables;

2. no RTC or spline variables, but include the aggregate state trends variable;

3. no RTC or spline variables, but include the 26 state specific state trends variables.

B Appendix B: Variable Inclusion Probabilities

This Appendix discusses the variable inclusion probabilities for the model comparison exercise with an expanded class of models in section 3.5. There are twelve categories of variables: demographics, poverty, unemployment, police, prison, population density, income, state population, abortion, alcohol, politics, and RTC variant. For each category, one option is to include no variables from the category.

Six tables follow, two sets of three. The first set focuses on the eleven groups of variables other than the RTC group. The first table in this set states the minimum, average, and maximum inclusion probability across all nine crime categories for each of the eleven groups of non-RTC variables. Each group is separated by double lines. This table presents an overview, serving as an introduction to the later tables that state results for each crime category.

Variables from the abortion and prison group are included with probability close to one for all crime categories.[80] In contrast, the probability for inclusion of the political variables is almost always close to zero, and, except for one crime category, the probability of inclusion for the alcohol variable also is close to zero. For the other seven groups, the average probability that no variable in the group appears ranges from around 0.17 to around 0.86.

The second and third tables in the first set indicate the results for each crime category. It is evident that groups are typically either in or out with high probability. It also is clear that the favored groups and variables vary substantially across crime categories. I leave exploration of the differences and alternatives such as using a multivariate regression model to future work.

The second set of tables parallels the first except that the focus is on RTC variants. Variants differ with respect to: use of state specific ("st_sp") or aggregate ("agg") RTC dummies; whether or not state trend variables ("st_tr") are included (either aggregate ("agg") or state specific ("st_sp")); whether a dummy ("dum") or spline ("spl") type of RTC variable is involved. The final line in each table gives the probability that some form of RTC variable is included. The body of the article discusses the results set forth in these tables in great detail. The tables are included here for completeness.

[80]The "0.0000" values in the table typically mask some very small probability rather than a probability equal to precisely 0.

Non-RTC Variable Probabilities under CML (across all nine crime categories)			
variable	minimum	average	maximum
no demographics	0.0000	0.1734	0.9963
Lott demographics	0.0000	0.7148	1.0000
Spelman demographics	0.0000	0.1117	1.0000
Zheng demographics	0.0000	0.0000	0.0002
no poverty	0.0022	0.8626	1.0000
Lott poverty	0.0000	0.1264	0.9978
Zheng poverty	0.0000	0.0086	0.0736
Don-Lev poverty	0.0000	0.0024	0.0198
no unemployment	0.0000	0.4570	0.9999
Lott unemployment	0.0000	0.1118	0.9985
Spelman unemployment	0.0000	0.0010	0.0049
Don-Lev unemployment	0.0000	0.4302	1.0000
no police	0.0000	0.4795	0.9998
Zheng police	0.0000	0.2876	1.0000
Spelman police	0.0000	0.0055	0.0478
Don-Lev police	0.0000	0.2274	0.9997
no prison	0.0000	0.0000	0.0000
Lott prison	0.0000	0.7778	1.0000
Don-Lev prison	0.0000	0.2222	1.0000
no density	0.0000	0.5751	0.9998
Lott density	0.0000	0.1805	1.0000
Zheng density	0.0000	0.2444	0.9990
no income	0.0000	0.4430	1.0000
Spelman income	0.0000	0.1542	0.9802
Zheng income	0.0000	0.4028	1.0000
no state pop	0.0000	0.5900	1.0000
Lott state pop	0.0000	0.4100	1.0000
no abortion	0.0000	0.0000	0.0000
Don-Lev abortion	1.0000	1.0000	1.0000
no alcohol	0.3758	0.9292	1.0000
Zheng alcohol	0.0000	0.0708	0.6242
no political	1.0000	1.0000	1.0000
Zheng political	0.0000	0.0000	0.0000

Non-RTC Variable Probabilities under CML Violent Crimes					
variable	violent crime	murder	rape	robbery	aggravated assault
no demographics	0.0000	0.0000	0.5646	0.0000	0.9963
Lott demographics	0.0000	1.0000	0.4336	1.0000	0.0000
Spelman demographics	1.0000	0.0000	0.0018	0.0000	0.0034
Zheng demographics	0.0000	0.0000	0.0000	0.0000	0.0002
no poverty	0.9999	0.9771	0.8725	0.9954	0.0022
Lott poverty	0.0001	0.0011	0.1267	0.0026	0.9978
Zheng poverty	0.0001	0.0020	0.0002	0.0011	0.0000
Don-Lev poverty	0.0000	0.0198	0.0006	0.0008	0.0000
no unemployment	0.9493	0.9963	0.0015	0.9930	0.9999
Lott unemployment	0.0000	0.0020	0.9985	0.0060	0.0000
Spelman unemployment	0.0008	0.0008	0.0000	0.0004	0.0000
Don-Lev unemployment	0.0499	0.0010	0.0000	0.0006	0.0000
no police	0.9969	0.0002	0.3920	0.9275	0.9988
Zheng police	0.0000	0.0000	0.5660	0.0706	0.0000
Spelman police	0.0000	0.0001	0.0004	0.0003	0.0000
Don-Lev police	0.0030	0.9997	0.0416	0.0016	0.0012
no prison	0.0000	0.0000	0.0000	0.0000	0.0000
Lott prison	1.0000	1.0000	1.0000	1.0000	1.0000
Don-Lev prison	0.0000	0.0000	0.0000	0.0000	0.0000
no density	0.0004	0.3051	0.9539	0.9989	0.9998
Lott density	0.0062	0.6092	0.0009	0.0007	0.0000
Zheng density	0.9934	0.0857	0.0453	0.0004	0.0002
no income	1.0000	0.0172	0.9994	0.0000	0.0030
Spelman income	0.0000	0.9802	0.0003	0.3875	0.0001
Zheng income	0.0000	0.0026	0.0003	0.6125	0.9969
no state pop	0.0000	0.8817	0.4347	0.9958	0.0000
Lott state pop	1.0000	0.1183	0.5653	0.0042	1.0000
no abortion	0.0000	0.0000	0.0000	0.0000	0.0000
Don-Lev abortion	1.0000	1.0000	1.0000	1.0000	1.0000
no alcohol	1.0000	0.9992	0.3758	0.9990	0.9984
Zheng alcohol	0.0000	0.0008	0.6242	0.0010	0.0016
no political	1.0000	1.0000	1.0000	1.0000	1.0000
Zheng political	0.0000	0.0000	0.0000	0.0000	0.0000

Non-RTC Variable Probabilities under CML Property Crimes				
variable	property crime	burglary	larceny	auto theft
no demographics	0.0000	0.0000	0.0000	0.0000
Lott demographics	1.0000	1.0000	1.0000	1.0000
Spelman demographics	0.0000	0.0000	0.0000	0.0000
Zheng demographics	0.0000	0.0000	0.0000	0.0000
no poverty	1.0000	0.9987	0.9999	0.9178
Lott poverty	0.0000	0.0012	0.0000	0.0084
Zheng poverty	0.0000	0.0000	0.0000	0.0736
Don-Lev poverty	0.0000	0.0001	0.0000	0.0002
no unemployment	0.0000	0.0000	0.0000	0.1726
Lott unemployment	0.0000	0.0000	0.0000	0.0000
Spelman unemployment	0.0001	0.0000	0.0049	0.0021
Don-Lev unemployment	0.9999	1.0000	0.9951	0.8252
no police	0.0000	0.0000	0.9998	0.0008
Zheng police	1.0000	0.0000	0.0001	0.9512
Spelman police	0.0000	0.0004	0.0000	0.0478
Don-Lev police	0.0000	0.9996	0.0001	0.0002
no prison	0.0000	0.0000	0.0000	0.0000
Lott prison	1.0000	0.0000	1.0000	0.0001
Don-Lev prison	0.0000	1.0000	0.0000	0.9999
no density	0.0010	0.0000	0.9250	0.9918
Lott density	0.0000	1.0000	0.0001	0.0078
Zheng density	0.9990	0.0000	0.0750	0.0003
no income	0.0000	0.0000	0.9880	0.9791
Spelman income	0.0017	0.0000	0.0120	0.0064
Zheng income	0.9983	1.0000	0.0001	0.0145
no state pop	0.9999	0.9979	1.0000	0.0000
Lott state pop	0.0001	0.0021	0.0000	1.0000
no abortion	0.0000	0.0000	0.0000	0.0000
Don-Lev abortion	1.0000	1.0000	1.0000	1.0000
no alcohol	1.0000	1.0000	0.9931	0.9971
Zheng alcohol	0.0000	0.0000	0.0069	0.0029
no political	1.0000	1.0000	1.0000	1.0000
Zheng political	0.0000	0.0000	0.0000	0.0000

RTC Variant Probabilities under CML (across all nine crime categories)			
variable	minimum	average	maximum
no RTC or st_tr	0.0000	0.1111	1.0000
no RTC, st_tr–agg	0.0000	0.1103	0.9893
no RTC, st_tr–st_sp	0.0000	0.5575	1.0000
dum–agg	0.0000	0.0000	0.0000
spl–agg	0.0000	0.1087	0.9787
RTC st_tr–agg	0.0000	0.0012	0.0106
dum–st_sp	0.0000	0.0000	0.0000
spl–st_sp	0.0000	0.1111	1.0000
RTC st_tr–st_sp	0.0000	0.0000	0.0000
any RTC	0.0000	0.2210	1.0000

RTC Variant Probabilities under CML Violent Crimes					
variable	violent crime	murder	rape	robbery	aggravated assault
no RTC or st_tr	0.0000	0.0000	0.0000	0.0000	0.0000
no RTC, st_tr–agg	0.0000	0.0036	0.0000	0.0000	0.0000
no RTC, st_tr–st_sp	1.0000	0.0177	1.0000	0.0000	1.0000
dum–agg	0.0000	0.0000	0.0000	0.0000	0.0000
spl–agg	0.0000	0.9787	0.0000	0.0000	0.0000
RTC st_tr–agg	0.0000	0.0000	0.0000	0.0000	0.0000
dum–st_sp	0.0000	0.0000	0.0000	0.0000	0.0000
spl–st_sp	0.0000	0.0000	0.0000	1.0000	0.0000
RTC st_tr–st_sp	0.0000	0.0000	0.0000	0.0000	0.0000
any RTC	0.0000	0.9787	0.0000	1.0000	0.0000

RTC Variable Probabilities under CML Property Crimes				
variable	property crime	burglary	larceny	auto theft
no RTC or st_tr	1.0000	0.0000	0.0000	0.0000
no RTC, st_tr–agg	0.0000	0.9893	0.0000	0.0000
no RTC, st_tr–st_sp	0.0000	0.0000	1.0000	1.0000
dum–agg	0.0000	0.0000	0.0000	0.0000
spl–agg	0.0000	0.0000	0.0000	0.0000
RTC st_tr–agg	0.0000	0.0106	0.0000	0.0000
dum–st_sp	0.0000	0.0000	0.0000	0.0000
spl–st_sp	0.0000	0.0000	0.0000	0.0000
RTC st_tr–st_sp	0.0000	0.0000	0.0000	0.0000
any RTC	0.0000	0.0107	0.0000	0.0000

References

[Autor, Donohue & Schwab 2002] Autor, D.H., Donohue, J.J. & Schwab, S.J. 2002. "The Costs of Wrongful-Discharge Laws." NBER Working Paper #9425, *available at* http://www.nber.org/papers/w9425.

[Bernardo & Smith 1994] Bernardo, J.M. & Smith A.F.M. 1994. "Bayesian Theory." Wiley, Chichester.

[Berger & Sellke 1987] Berger, J.O. & Sellke, T. 1987. "Testing a Point Null Hypothesis: The Irreconicilabiltiy of P Values and Evidence." Journal of the American Statistical Association, 82, 112-122.

[Berger & Delampady 1987] Berger, J.O. & Delampady, M. 1987. "Testing Precise Hypotheses." Statistical Science, 2, 317-335.

[Casella & Berger 1987] Casella, G. & Berger, R.L. 1987. "Reconciling Bayesian and Frequentist Evidence in the One-Sided Testing Problem." Journal of the American Statistical Association, 82, 106-111.

[Chipman, George & McCulloch 2001] Chipman, H., George, E.I. & McCulloch, R.E. 2001. "The Practical Implementation of Bayesian Model Selection." IMS Lecture Notes – Monograph Series, 38, 67-134.

[Cox 1987] Cox, D.R. 1987. "Comment." Statistical Science, 2, 335-336.

[Cremers 2002] Cremers, K.J.M. 2002. "Stock Return Predictability: A Bayesian Model Selection Perspective." The Review of Financial Studies, 15, 1223-1249.

[Danilov and Magnus 2004] Danilov, D., Magnus, J.R. 2004. "On the harm that ignoring pretesting can cause." Journal of Econometrics, 122, 27-46.

[Dawid 1992] Dawid, A.P. 1992. "Prequential analysis, stochastic complexity and Bayesian inference (with Discussion)." In "Bayesian Statistics 4" (ed. J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith), 109-125. Clarendon Press, London.

[Dawid 1999] Dawid, A.P. 1999. "The Trouble with Bayes Factors." Research Report No. 202, Department of Statistical Science, University College, London.

[Delampady 1989] Delampady, M. 1989. "Lower Bounds on Bayes Factors for Interval Null Hypotheses." Jounral of the American Statistical Association, 84, 120-124.

[DeLong & Lang 1992] DeLong, J.B., Lang, K. 1992. "Are All Economic Hypotheses False?" Journal of Political Economy, 100, 1257-1272.

[Donohue 2004] Donohue, J.J. 2004. "Guns, Crime, and the Impact of State Right-to-Carry Laws." Fordham Law Review, 73, 623-652.

[Donohue & Wolfers 2005] Donohue, J.J., Wolfers, J. 2005. "Uses and Abuses of Empirical Evidence in the Death Penalty Debate." Stanford Law Review, 58, 789-843.

[Fernandez, Ley & Steel 2001] Fernandez, C., Ley, E. & Steel, M.F.J. 2001. "Benchmark priors for Bayesian model averaging." Journal of Econometrics, 100, 381-427.

[Freedman, Pisani & Purves 1998] Freedman, D., Pisani, R. & Purves, R. "Statistics." 3rd edition. Norton.

[Gelman, et.al. 2004] Gelman, A., Carlin, J.B., Stern, H.S. & Rubin, D.B. 2004. "Bayesian Data Analysis." 2nd edition. Chapman & Hall.

[Gelman & Rubin 1995] Gelman, A. & Rubin, D.B. 1995. "Avoiding Model Selection in Bayesian Social Research." In "Sociological Methodology" (Peter V. Marsden, ed.). Blackwell.

[George 1999] George, E.I. 1999. Discussion of "Bayesian model averaging and model search strategies" by M.A. Clyde. In "Bayesian Statistics 6" (J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, eds.) 175-177, Oxford University Press.

[George & Foster 2000] George, E.I. & Foster, D.P. 2000. "Calibration and empirical Bayes variable selection." Biometrika, 87, 731-747.

[Geweke 1993] Geweke, J. 1993. "Bayesian Treatment of the Independent Student-t Linear Model." Journal of Applied Econometrics, 8, S19-S40.

[Geweke 2005] Geweke, J. 2005. "Contemporary Bayesian Econometrics and Statistics." Wiley.

[Good 1987] Good, I.J. 1987. "Comment." Journal of the American Statistical Association, 82, 125-128.

[Hastie, Tibshirani & Friedman 2001] Hastie, T., Tibshirani, R. & Friedman, J. 2001. "The Elements of Statistical Learning: Data Mining, Inference, and Prediction." Springer.

[Ho, et.al. 2005] Ho, D., Imai, K., King, G. & Stuart, E. 2005. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." Working Paper.

[Jackman 2006] Jackman, S. 2006. "Bayesian Analysis for the Social Sciences." Forthcoming. Wiley.

[Jeffreys 1980] Jeffreys, H. 1980. "Some General Points in Probability Theory," in *Bayesian Analysis in Econometrics and Statistics*, ed. A. Zellner, Amsterdam: North-Holland, pp. 451-454.

[Kass & Raftery (1995)] Kass, R.E. & Raftery, A.E. 1995. "Bayes Factors." Journal of the American Statistical Association, 90, 773-795.

[Kass & Wasserman (1995)] Kass, R.E. & Wasserman, L. 1995. "A Reference Bayesian Test for Nested Hypotheses and Its Relationship to the Schwarz Criterion." Journal of the American Statistical Association, 90, 928-934.

[Koop 2003] Koop, G. 2003. "Bayesian Econometrics." Wiley.

[Lancaster 2004] Lancaster, T. 2004. "An Introduction to Modern Bayesian Econometrics." Blackwell.

[Leamer 1983] Leamer, Edward E. 1983. "Let's Take the Con out of Econometrics." American Economic Review, 73, 31-43.

[Madigan & York 1995] Madigan, D. & York J. 1995. "Bayesian Graphical Methods for Discrete Data." International Statistical Review, 63, 215-232.

[O'Hagan & Forster 2004] O'Hagan, A., Forster, J. 2004. "Kendall's Advanced Theory of Statistics, Volume 2B: Bayesian Inference (2nd Edition)." Arnold, London.

[Poirier 1996] Poirier, D.J. 1996. "Intermediate Statistics and Econometrics." MIT Press.

[Raftery 1995a] Raftery, A.E. 1995. "Bayesian Model Selection in Social Research." In "Sociological Methodology" (Peter V. Marsden, ed.). Blackwell.

[Raftery 1995b] Raftery, A.E. 1995. "Rejoinder: Model Selection is Unaviodable in Social Research." In "Sociological Methodology" (Peter V. Marsden, ed.). Blackwell.

[Raftery 1999] Raftery, A.E. 1999. "Bayes Factors and BIC." Sociological Methods & Research, 27, 411-427.

[Schwarz (1978)] Schwarz, G. 1978. "Estimating the Dimension of a Model." The Annals of Statistics, 6, 461-464.

[Sellke, Bayarri & Berger 2001] Sellke, T., Bayarri, M.J. & Berger, J.O. 2001. "Calibration of p Values for Testing Precise Null Hypotheses." The American Statistician, 55, 62-71.

[Theil 1971] Theil, Henry 1971. "Principles of Econometrics." Wiley, New York.

[Tukey 1978] Tukey, J.W. 1978. Discussion of Granger on seasonality. In "Seasonal Analysis of Economic Time Series," ed. A. Zellner. Washington, DC: U.S. Government Printing Office, pp. 50-53.

[Zellner 1986] Zellner, A. 1986. "On assessing prior distributions and Bayesian regression analysis with g-prior distributions." In "Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finietti," Ed. P.K. Goel & A. Zellner, pp. 233-243. Amsterdam: North-Holland.

[Zellner 1987] Zellner, A. 1987. "Comment." Statistical Science, 2, 339-341.