Conditional Frailty Model." It is clear that our tools and our contributions are appreciated by many outside of the discipline with methodologists making contributions other fields besides health/medicine, including: sociology, epidemiology, pure statistics, criminology, finance, demographics, economics, data processing, law, linguistics, and so on. This demonstrates that the subfield has reached a new level of scientific maturity wherein other fields are interested in importing our tools. Notice that this is the opposite effect than that described by Neal Beck only ten years ago in Political Methodology–A Welcoming Discipline (*Journal of the American Statistical Association*, Volume 95 (2000), pp. 651-654). We appear to have gone from net importers to net exporters, which is an exciting development.

Making Regression and Related Output More Helpful to Users[1]

Nathaniel Beck
New York University
nathaniel.beck@nyu.edu

All of us interact every day with computer screens showing tables of regression and related output.[1] We do this so often that our eyes have perhaps gotten used to ignoring the useless output that appears on those screens, though all those extraneous numbers we constantly see surely cannot improve analysis. Too many statistics make it harder for our eyes (and brains) to focus on the important ones. Worse, for students newly encountering regression, this extraneous output can often be misleading. Do we really want to teach our students that every regression coefficient should be tested against the null hypothesis that it is zero? We know enough to do much better, and it is easy to do better. This diatribe is an effort to push us in this direction.

Before beginning let me be clear that I am discussing output seen by the user on a computer screen. Clearly we can write an article or paper picking output as we choose, but, as we shall see, we often have little control over the standard output we see on a screen.[2] It is screen output that concerns me since it is screen output with which we normally interact. The issue is also of concern for students newly coming to regression: if Stata or R[3] produces some output surely it must be important. (Even worse, I as the instructor have forced them to spend effort learning these things, and now I say that much of it is useless!) It should be stressed that fault does not lie with the programmers of Stata or R; they both produce excellent software that does what users want. The problem is with *our* not demanding more useful output, and *our* continuing to perpetuate the mistakes of the past, both in our practice and teaching. And we know that what we do now is not quite right. Bastante![4]

Figure 1 shows what a user sees on the screen after typing a regression command in Stata; the output is for a generic model of votes for House candidates in the US (Jacobson et al., 1994). This output is not customizable by the normal user. While Stata is very commonly used in our discipline, maybe "higher end" packages like R do better? Figure 2, which shows standard regression output using the *summary()* method, disabuses us of that notion. Here, the only possible customization is to allow for "magic stars" indicating significance, hardly a useful customization.[5] Table 1 shows what I think standard output should look like. The alert reader will note that the change is non-trivial. The rest of the article discusses these differences in more detail. There are many good discussions of these issues from a statistical perspective; here I simply reference a few of those discussions. If you have not been convinced by previous articles that not every regression coefficient need come with a test of a null hypothesis that nothing is going on, or that R^2 is comforting but not very useful, this piece is not going to convince you and you can stop reading now. Know,

[1] Thanks to too many friends who do political methodology to name and two friends who realized how awful I am at coding. I also want to acknowledge the kind hospitality of CEACS at the Fundación Juan March in Madrid for allowing me to present an informal seminar which started me down this path.

[1] All said here generalizes to more complicated regression-like output produced by standard maximum likelihood routines.

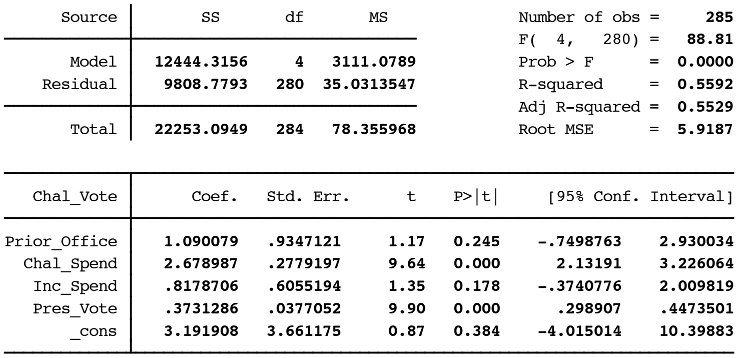
[2] Obviously there have been numerous previous attempts to make points similar to those made here, but directed towards how results are presented in journals. Gelman, Pasarica, and Dodhia (2002) and Kastellec and Leoni (2007) make an excellent argument for replacing all regression tables with graphs in journal articles. Here I am concerned with the computer screens we look at well before writing the journal article or senior thesis. I would be most happy if journals adopted the perspective of Gelman and Kastellec and Leoni; I would be quite happy if journals simply adopted some of the points below. But this diatribe is only indirectly aimed a journals.

[3] I discuss these two packages because they are most commonly used in political science, and they are also the most sophisticated of the general packages.

[4] Gigerenzer (2004, 604) concludes his article on the foolishness of null hypothesis testing equally strongly. "To stop the ritual [of null hypothesis testing], we also need more guts and nerves. We need some pounds of courage to cease playing along in this embarrassing game. This may cause friction with editors and colleagues, but it will in the end help them to enter the dawn of statistical thinking."

[5] Gelman and Hill's (2007) R package, *arm* provides the *display()* method, which is superior to *summary()* in that it eliminates significance tests and some unnecessary output and limits the number of decimal place (not significant digits!) to two. Moreover, it does not provide confidence (or highest posterior density) intervals. But *display()* is far superior to *summary* though not as good as Jeff Gill's *graph.summary()* mentioned below.

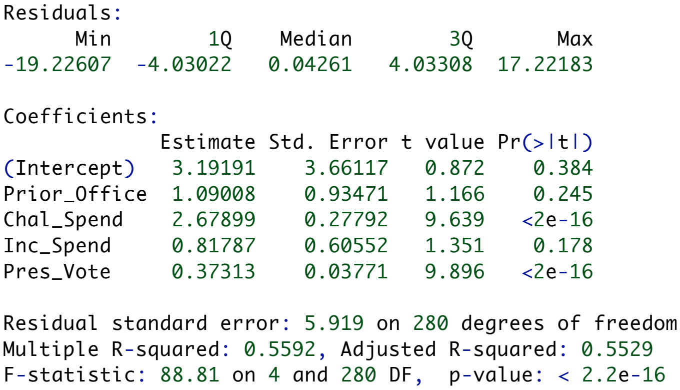
however, that you have decided to play along in Gigerenzer's embarrassing game.[6] If one does not wish to continue playing the game, Stata and R output routines (kindly written by programmers more competent than me) are available. The R program is available on Jeff Gill's website (http://artsci.wustl.edu/~jgill/Models/graph.summary.s; the Stata program can be installed by typing *ssc install leanout*.



```
      Source |       SS       df       MS              Number of obs =     285
-------------+------------------------------           F(  4,   280) =   88.81
       Model |  12444.3156     4  3111.0789            Prob > F      =  0.0000
    Residual |   9808.7793   280  35.0313547           R-squared     =  0.5592
-------------+------------------------------           Adj R-squared =  0.5529
       Total |  22253.0949   284   78.355968           Root MSE      =  5.9187

------------------------------------------------------------------------------
   Chal_Vote |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Prior_Office |   1.090079   .9347121     1.17   0.245    -.7498763    2.930034
  Chal_Spend |   2.678987   .2779197     9.64   0.000      2.13191    3.226064
   Inc_Spend |   .8178706   .6055194     1.35   0.178    -.3740776    2.009819
   Pres_Vote |   .3731286   .0377052     9.90   0.000      .298907    .4473501
       _cons |   3.191908   3.661175     0.87   0.384    -4.015014    10.39883
------------------------------------------------------------------------------
```

Figure 1: Standard Stata regression output: regression of vote for House challenger, 1992



```
Residuals:
      Min        1Q    Median        3Q       Max
-19.22607  -4.03022   0.04261   4.03308  17.22183

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.19191    3.66117   0.872    0.384
Prior_Office  1.09008    0.93471   1.166    0.245
Chal_Spend    2.67899    0.27792   9.639   <2e-16
Inc_Spend     0.81787    0.60552   1.351    0.178
Pres_Vote     0.37313    0.03771   9.896   <2e-16

Residual standard error: 5.919 on 280 degrees of freedom
Multiple R-squared: 0.5592, Adjusted R-squared: 0.5529
F-statistic: 88.81 on 4 and 280 DF,  p-value: < 2.2e-16
```

Figure 2: Standard R output: regression of vote for House challenger

Variable	$\hat{\beta}$	SE	95% CI
Prior Office	1.1	.9	(-.8 & 2.9)
Chal Spend	2.7	.3	(2.1 & 3.2)
Inc Spend	.8	.6	(-.4 & 2.0)
Pres Vote	.4	.04	(.3 & 0.4)
Constant	3.2	3.7	(-4.0 & 10.)
$\hat{\sigma} = 5.9$			
Number of observations: 285			

Table 1: What output should look like: regression of vote for House challenger

[6] One might decide to play that game for final publication given the power of editors and referees, but surely this game does not need to be played on every screen of regression output. Or one might decide that the only way to change things is to change editors and referees.

Significant and decimal digits

Let me start with something simple. I hope that none of us[7] believe that our regression estimates are accurate to seven significant digits. But this is how many digits Stata outputs, and there is no option to change this.[8] Ah the high priests say, Stata is just for the untutored, let them use R and all problems will be solved. But while R has a *digits()* option, which is a great idea, the output in R, as produced by the standard *summary.lm()* command, is, as written, not capable of producing fewer than 3 decimal places (and will produce as many integer digits as the regression indicates). So we all look at perhaps seven digits for every number we see, we all know this is ridiculous, and we all continue to allow this to happen. And this is for only the simplest issue, where there can be no cogent argument for what is, by default, current practice.

Computer programmers strive for enormous numerical accuracy but, alas, social science data are not quite so accurate. When I feel optimistic I might believe our results are accurate to two digits. We see fewer and fewer articles and presentations which simply take the digits reported by whatever computer package. But even if some journals can enforce a more reasonable number of digits, the number of digits seen on a screen simply confuses the eye, and does so for no good purpose. And surely there is more danger of the student being misled about the accuracy of regression results.

Now, alas, the number of significant digits and the number of digits reported is not the same thing; does 0.00034 have two, five or six significant digits? The analyst should know, but a computer program cannot. Journals (including the one I edited) often limit regression coefficients to two decimal places, but allow coefficients like 9462273.24. There is no easy way around this, and it is not obvious that users would understand that 9462273.24 should probably be reported as 9500000 (or perhaps 9000000).[9]

The issue is problematic since we (in general) are not really good at comprehending really small or really big numbers. Physicists and astronomers, after all, have decided they need both Ångstrom units and parsecs. Thus analysts should strive to have coefficients that are easy to understand, that is, a relatively small integer followed by perhaps a single decimal place. This is particularly helpful in regression, where the coefficient tells us the "impact" (whether causal or not) of a unit change in a variable. If that unit is too small (say measuring household income in dollars), the coefficient will be minuscule and hard to interpret; a simple rescaling to income in thousands of dollars solves many problems at once. Similarly, we would not want household income measured in millions of dollars; the counterfactual of a one million dollar increase in my income is truly a counterfactual, and the regression coefficient is going to be misleadingly large. Researchers should also try for meaningful scalings, and attempting to get reasonably sized (small) regression coefficients is one good way to try to enforce this. There is no perfect solution; we do not want income measured in hundreds of dollars, or area measured in hectohectares.[10]

Thus, for all these reasons, regression program should (at least by default) output numbers as n.d or nn. If coefficients do not fit into this scheme, it should normally be easy enough (and good) to rescale that variable (or, if relevant) the dependent variable. This should be the easy part of the argument, though practice shows this might not be as easy as I would hope. I looked at the 10 most recent quantitative articles in what should be our most sophisticated methodological journals, *The American Journal of Political Science* and *Political Analysis*. In both journals, the typical number of decimal places reported was three, with an additional units place, leading readers to believe that authors felt their results were accurate to four significant digits. While this piece deals only with output on computer screens, clearly we also need to worry about how that output appears in published articles. But that is a different task.

No irrelevant ancillary statistics

Stata and R (and all other statistical packages that I know of) show a number of ancillary statistics and tests that are of little or no interest, and certainly not of such interest that they should appear on every screen of output. (Where they are are of interest it is easy produce them after estimation.) These statistics and tests mislead many students into being overly excited when they should not be (and vice versa) and they may lead even experienced analysts into not focussing on what is really important.

The most egregious output is the F-test of the null

[7] In this diatribe "us" is some combination of readers of TPM and the larger number of users of quantitative methods in our discipline. Apologies to (the small number of) the innocent.

[8] On June 4, 2010, while this piece was in production, Stata 11.1 was released. This version allows the user to format how statistics appear. It is now trivial to limit output to two decimal places (with no attention to significant digits), or to force, as I would prefer, all output to appear as n.d or nn.d. I would urge all Stata users to upgrade to 11.1 and use the command "set cformat %3.1f, perm" in their profile. Clarity would also be improved by setting pformat and sformat to %1.0.

[9] Scientific notation can solve this problem. In some sense it solves the problem too well, by providing sensible mantissas, though at the cost of somewhat hard to comprehend characteristics. By solving the problem of significant digits too well, scientific notation provides no incentive for meaningfully rescaling the data.

[10] Thus, sometimes it might make most sense to report a coefficient of 0.032. No hard and fast rule is going to work here. But any system similar to the metric system allows us to get good scaling within a range of at most 3 digits (decimal plus integer).

hypothesis that all coefficients are zero, and the associated ANOVA table. This must be the least interesting null hypothesis in the world; why this is standard is beyond me. Even more puzzling, why do I always want to see the ANOVA table which is the basis for this test? I do not think it controversial to advocate dropping these items from standard screen output.

It is more controversial, but correct, to drop R^2 from the output (King, 1986, 675-8).This is a meaningless, unitless number that is supposed to give us comfort if it is close to one (how close?). There is no reason for a student to believe that a high R^2 is good or that a low R^2 is bad; we surely do not want students just adding variables to build up that R^2. Do we think that a regression which includes an independent variable that is almost identical to the dependent variable is superior to a regression without such a variable? And if R^2 is useless, so is the adjusted R^2; for any decent size data set the adjustment is trivial. (If one wanted some arbitrary number, the BIC would be better, but I am not advocating any numbers that are not directly of interest here.) So R^2-related statistics (and their maximum likelihood wannabes) should also disappear from standard screen output.

This leaves the number of observations and the estimate of σ, the standard deviation of the distribution which theoretically has generated the errors. N is useful because analysts so often compare regressions with different numbers of observations (whether due to missing data or something else). Such a comparison is difficult, at best, and researchers should always know how large their "sample" is. For time-series and time-series–cross-sectional data, programs should report the "sample period" (in meaningful dates) and, in the latter case, the number of units as well as the overall N.

Why $\hat{\sigma}$? This is a very nice interpretable number, a number which has the same units as the dependent variable. It tells the analyst how far a typical observation is from the regression line. This, unlike R^2, is an intuitive and meaningful number; if a dependent variable is GDP per capita in thousands of 2005 US dollars, and if the standard error or estimate is 10 (thousand US dollars), we know that a typical country is within about \$10,000 of the regression line (which may be good or bad).[11]

The regression table

Turning to the regression coefficients, clearly we need the coefficient (with fewer than 7 digits) and its standard error and its 95% confidence interval.[12] But why does every coefficient estimate we ever see come with an associated t-test of one specific null hypothesis ($H_0{:}\beta_k = 0$) and the associated p-value. As Gigerenzer (2004) and Gill (1999) and many others, have persuasively argued, hypothesis testing is a deeply flawed activity. But we need not even go this far to note that there is no reason *always* to look at the t-statistic and p-value for one specific null hypothesis test that the true value of a parameter might be zero. If we care about that hypothesis then we can simply check whether zero is contained in the reported confidence interval.

Alas, students (and others) misunderstand the meaning of significance tests. They often think that a failure to reject the null hypothesis means that they have shown that a parameter value is zero (or even small); they think that lower p-values indicate that a parameter is more important. And how many scholars, thumbing through a huge list of coefficient estimates, ignore those that have $p > .05$ and focus on the others, sorting estimates into significant and insignificant. We all know this is not the right practice, and it should not be aided and abetted by our computer programs. We should not be interested in simply whether a coefficient is "significant." We go to great trouble to estimate coefficients in units that give a huge amount of information; simply looking at the unit-less t or p just discards that information.

Focussing on the one simple test of the null that $\beta = 0$ also misleads students into not thinking about the hypothesis of interest. Sometimes we are interested in a series of coefficients, sometimes we are interested in the equality of coefficients, sometimes we care if they are near one, and so on. Current regression output makes it appear that the thing we naturally care about is one specific null. So whatever one thinks about hypothesis testing logic, current regression output is highly misleading.[13]

[11] The same examination of recent regression tables in the two journals indicates that everyone seems to believe that R^2 (or its pseudo-friends) is important, as is the likelihood. No one seems to believe that $\hat{\sigma}$ is worth reporting.

[12] 95% is about as good as any other choice. In some recent articles using Bayesian methods, authors have reported 80% "highest posterior density" (i.e. confidence) intervals. Why those who use Bayesian computations are happy with 4:1 odds while others are used to 19:1 odds is, at best, unclear, though of course the 80% intervals are comfortingly smaller.

[13] Alas, as previous, the output in our top journals indicates that no one finds confidence intervals of enough value to report and lots of results get labeled with magic stars, with the magic star always related to a test of whether some true parameter value might be zero. As Gerber and Malhotra (2008) clearly show, our journals seem to lack results which correspond to p-values just above .05. Gerber and Malhotra focus on publication bias; my interpretation of their results is that anyone clever enough to do a regression, and employed in a profession that values publication, upon seeing a key result with a p slightly above .05, will have no trouble finding a new specification with a p happily just below .05. So all this focus on stars and p-values simply leads to p-values which in fact are not p-values at all. Looking at the Gerber and Malhotra results, we can be quite sure that, even if we believe in the null hypothesis testing paradigm, our own work must be violating that paradigm.

Interpretative Bayesianism/Subjectivism

So now we see only useful output. If the coefficients themselves are of interest (as in regression), it is hoped that analysts will focus on those, looking at the numbers in terms of units, not simply asking if the estimate is "significant." For more complicated models, clearly other quantities of interest (and the uncertainty associated with those) must be estimated (King, Tomz and Wittenberg, 2000). But what is critical is that analysts and students not undo all their good work by basically running a hypothesis test in their head, that is, simply seeing which confidence intervals contain zero (in which case magic stars will do the same thing more efficiently). So how can better use be made of the uncertainty estimates?

Confidence intervals are difficult for classical (frequentist) statisticians to interpret. Few if any students remember the correct frequentist interpretation of a confidence interval five minutes after the final exam in their first course (if they ever knew it). Most people I know interpret a confidence interval as "it is likely that the parameter value lies in the interval." Such a statement makes sense only to someone who believes in subjective probability (Savage, 1954), where the probability of a statement being true is given by the odds you would be willing to give on a bet that the statement is true. So there are no frequentists in fox-holes. But can we use classical frequentist methods and then interpret results like a Bayesian?

Subjective probability developed independently of Bayesian inference. As Fienberg (2006, 16) notes, Savage's book mentions Bayes only once. But clearly subjective probability and Bayesian inference are now joined more closely than that.

Fortunately (or not!) most (not all, but most) Bayesian analyses done in political science are not really Bayesian, in that they use a highly uninformative prior (and I have yet to see a second study use the first to update said prior). Thus most Bayesians in political science are what I would call computational Bayesians, that is, they take advantage of the great power of Bayesian computational methods to produce results for very complicated models where standard classical methods fail. But, for simple things like regression (and simple maximum likelihood like logit and probit), for a reasonable sized sample (say at least 50) and a highly uninformative prior, the numerical results from a Bayesian and classical analysis are essentially the same (remembering how many significant digits we really have).[14] Thus one can take the 95% confidence interval computed classically and say that one would offer a bet at 19:1 odds that the parameter value lies in this range. This is simply a formalization of how almost all of us interpret confidence intervals in practice. Thus we can use the nice output to say that we are pretty sure that the true parameter is at least so big and and no bigger than something else. This seems like the most useful way to summarize what the data is saying about the parameters and their associated uncertainty.

Ten Commandments

1. Produce screen (and journal) output that is as meaningful to the analyst (and reader) as possible.

2. Make your output as easy to read as possible. In particular, variables should have meaningful names that relate to the underlying concepts.

3. Produce no more digits than are significant. If unsure, two is a generous guess.

4. Produce numbers that the human brain can easily process (typically between .1 and 9.9).

5. Choose units for your variables that make interpretation simpler.

6. Report all interesting numbers in meaningful units.

7. Do not provide uninteresting summary statistics; if they are really needed, they can be produced later. Provide interesting summary measures (such as $\hat{\sigma}$) that have units.

8. Provide only parameter estimates and indications of uncertainty of those estimates. This will usually be done via standard errors and confidence intervals.

9. Do not routinely produce tests of standard null hypotheses that a parameter is zero. Do not use stars or other markers to denote levels of significance.

10. Break any rules that conflict with the first.

For those who bristle at commandments, all of the items above can be rephrased as promises, with first-person pronouns. My goal is to get the data to speak as clearly as possible, particularly to students. So I conclude with some vows. I will no longer teach what I know to be nonsense, and no longer participate in nonsensical statistical rituals to please reviewers and editors. I will implement best practices, and endeavor to have my tools enhance those practices.

References:

Fienberg, Stephen E. 2006. "When Did Bayesian Inference Become 'Bayesian'?" *Bayesian Analysis* 1 (1):1–40.

Gelman, Andrew, Cristian Pasarica and Rahul Dodhia. 2002. "Let's Practice What We Preach: Turning Ta-

[14] This is not to deny that Bayesian computations are superior for hard problems, even with uninformative priors. Nor would I deny that Bayesian methods provide superior results for more complicated models. But even in those cases, one can still interpret the classical confidence intervals as a subjectivist would, conditional on the somewhat inferior classical model.

bles into Graphs." *American Statistician* 56 (2):121–30.

Gelman, Andrew and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York, NY: Cambridge University Press.

Gerber, Alan and Neil Malhotra. 2008. "Do Statistical Reporting Standards Affect What Is Published? Publication Bias in Two Leading Political Science Journals." *Quarterly Journal of Political Science* 3 (3):313–26.

Gigerenzer, Gerd. 2004. "Mindless Statistics." *The Journal of Socio-Economics* 33 (5):587–606.

Gill, Jeff. 1999."The Insignificance of Null Hypothesis Statistical Testing." *Political Research Quarterly* 52 (3):647–74.

Jacobson, Gary C. and Michael A. Dimock. 1994. "Checking Out: The Effects of Bank Overdrafts on the 1992 House Elections." *American Journal of Political Science* 38 (3):601–24.

Kastellec, Jonathan P. and Eduardo L. Leoni. 2007. "Using Graphs Instead of Tables in Political Science." *Perspectives on Politics* 5 (4):755–71.

King, Gary. 1986."How Not to Lie with Statistics: Avoiding Common Mistakes in Quantitative Political Science." *American Journal of Political Science* 30 (3):666–87.

King, Gary, Michael Tomz and Jason Wittenberg. 2000."Making the Most of Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science* 44 (2):341–55.

Savage, Leonard J. 1954. *The Foundations of Statistics.* New York, NY: Wiley.

SLAMM Abstracts

The 2010 St. Louis Area Methods Meeting (SLAMM) was held April 16 and 17 at Washington University, co-hosted by the Center for Applied Statistics and the Department of Political Science. The second day of the conference was reserved for graduate student presentations. The editors invited students who made presentations at the meeting to submit long abstracts of their work.

Are We Testing What We Think We're Testing? A Theoretical Evaluation of Methods for Testing Hypotheses about Temporal Changepoints

Michael P. Fix
University of South Carolina
fixm@email.sc.edu

Applied time series analysis is frequently used to study questions of great importance in political science. For example, scholars may be interested in patterns of democratization, changes in Supreme Court voting behavior, or the determinants of civil conflict. In this research we often make the implicit assumption that the relationship of interest is static across all subsets of the time series. While it is possible that this assumption holds in some instances, without sound theory it is important that this assumption be explicitly addressed and tested. Yet, determining how to properly conduct these tests is a more complex question.

In this paper, I advocate the use of Bayesian multiple changepoint models to test for potential structural breaks in time series data. In doing so, I focus primarily on the theoretical congruence (or the lack thereof) between the nature of the question tested and the methodological approach used. Certain approaches commonly used for testing for structural breaks (e.g. Chow tests) require the researcher to specify the potential changepoints *a priori.* Further, many of these tests are limited to the detection of a single structural break. Bayesian multiple changepoint models provide a theoretically more appropriate alternative to the commonly used techniques for dealing with changepoint problems by allowing the changepoints to be estimated as a parameter simultaneously with the other parameters. Moreover, Bayesian changepoint models allow for the estimation of the number and location of this structural breaks without having to specify their values *a priori.*

To illustrate the application of this approach, I present a substantive example from an analysis of the determinants of judicial decision making when reviewing administrative agency decision making. *Chevron USA, Inc. v. Natural Resources Defense Council* is one of the most widely cited decisions in the history of the U.S. Supreme Court, and many scholars claim that it completely reshaped administrative law. In essence, the *Chevron* decision held that courts were to defer to agency interpretation of statutes