Explaining Something from Nothing: Understanding the Influences of Model Specification and the Implications for Criminological Research

David W. McClure

July 22nd, 2011

Introduction

Regression analysis is an important research tool in the discipline of criminology. In criminal justice education, it is among the most prominent and common components of the curriculum. In criminological research, the many variations are perhaps the discipline's most commonly used statistical techniques. When properly applied, regression analysis can provide criminologists with detailed descriptions of genuine relationships between the many social phenomena associated with crime and justice. When regression is not used properly, however, it can provide erroneous descriptions of such relationships, which could ultimately lead to incorrect conclusions and false understandings.

Fortunately, criminologists can reduce the possibility of drawing false conclusions from the results of regression analysis through use of statistical and methodological techniques, as well as adherence to disciplinary conventions. In order for these solutions to be effective, however, it is first necessary to understand the problem. This paper is meant to provide such an understanding of several problems with regression analysis, and to demonstrate the extent to which the known solutions are able to address these problems.

More specifically, this paper is meant to accomplish two objectives: first, present a clear demonstration of the effects of over-specification and post-hoc specification in regression modeling; and second, discuss the disciplinary developments necessary to address these issues with model specification. To achieve this, the paper will describe the results of a series of Monte Carlo simulations that analyze random-value datasets in order to demonstrate the extent of the effects specification issues can have on the results. Because there should be no meaningful relationship among the variables in these random-value dataset, any reported effect can be attributed to the combination of chance co-variation and statistical practices used, rather than a genuine relationship in the data. Where they exist, these simulations also demonstrate the solutions to these problems. Where solutions do not exist, these simulations are used to demonstrate the importance of considering how these effects might influence the discipline and what can be done to reduce the threat. But first, it is necessary to more completely understand the role of regression analysis in criminology.

Regression in Criminology

Regression in Criminal Justice Education

Beginning their assessment at the final stages of graduate level social science programs, Wiles, Durrant, De Broe, & Powell, (2009) surveyed prospective employers of social science graduates to identify their most desired skills in new hires. The authors found that various data analysis and research abilities were most desired by employers. Examining the satisfaction of such a demand, Lytle & Travis (2008) reviewed the types of courses required in MA level criminal justice programs. The authors concluded that research methods and data analysis courses are virtually the only courses consistently required in criminal justice programs. "As a social science, criminal justice programs require students to understand research methods and statistics. Beyond those classes, only a course on criminological theory is required by as many as three-quarters of degree programs" (348). After reviewing and comparing all the required courses of the 186 identified programs, these authors found 93% required research methods courses and 61% required statistics courses. With very little consistency across programs, the existence of a rare common focus on research methods and analysis suggests these skill-sets are very important within the discipline.

Exploring the characteristics of the common focus on research methods and analysis across criminal justice programs, Sullivan & Maxfield (2003) reviewed the syllabi of the required research methods and data analysis courses in 27 university criminology PhD programs. The authors concluded,

> The vast majority" of research on the issue of paradigms in criminal justice/criminology has focused on theoretical and ideological issues. While acknowledging that these are important issues for study, it is just as important that the self-analysis be diversified. As Blankenship and Brown (1993) point out, in setting forth the term "paradigm," Kuhn was most concerned with research methods. We should be equally concerned with research methods in our discussions of paradigms within the field of criminology and criminal justice. This should lead to continued focus on the study of what is being consumed by doctoral students as an indicator of where the study of crime and criminal justice is going. (284)

Recognizing the demand for research methods and analysis (Wiles et al., 2009), that it is nearly the only common bond across criminology graduate programs (Lytle & Travis, 2008), and that discussions of regression analysis represent the single largest portion of these courses (Sullivan & Maxfield, 2003), regression analysis should is clearly a very prominent component of criminology's curriculum.

Concept of Regression

The popularity and prominence of regression modeling in criminological education is certainly understandable; the ability to examine relationships between a variety of social phenomena is highly applicable to researching crime. On a conceptual level, it allows for a more complete understanding of a given variable by describing it in terms of its relationships with other related variables, which allows regression analysis to more accurately describe variation in a dependent variable than is possible through simple descriptive statistics.

To illustrate, consider how one might answer the question, "how long were the prison sentences assigned in 2010?" As with many inquiries in the social sciences, even such a simple question does not have a simple answer. One possibility would be to provide the average length of prison sentences in 2010.

While this single value would provide a simple description of the many different prison sentences assigned in 2010, it would not be very precise. Many 2010 prison sentences would be much longer and much shorter than the average length.

Bivariate regression provides a more precise alternative to the average as a description of the 2010 sentences. Instead of using one value to describe all 2010 sentences, bivariate regression ties the response to one other variable (e.g., prior criminal history), which allows for a more precise description of the 2010 prison sentences (i.e., prison sentences increased X number of months for every prior criminal conviction). Though such a description is more complex than the average sentence, it is also more a precise description of the variation in the length of prison sentences (i.e., the actual sentences do not fall as far above and below their described values as compared to the average sentence). The trade-off between simplicity and precision continues as multivariate regression uses even more variables to explain the length of prison sentences (e.g., convicted crime, age, gender, etc.).

One method of describing the increased precision of regression over the average is the amount of variance explained, which is represented by the R-squared statistic.

$$R^2 = \frac{Improved\ Precision\ of\ the\ Regression\ Model\ beyond\ the\ Mean}{Imprecision\ of\ the\ Mean}$$

Essentially, the R-squared statistic is calculated as the total distance between the actual values of the data and the model's description of that data divided by the total distance between the actual values of the data and the average's description of that data. As a regression model will always be at least as precise as the mean, the results of the R-squared calculations will always fall between zero and one, and can be interpreted as the percent of variation in the data the model explains more precisely than the average.

Variance Explained in Criminological Research

The R-squared coefficient has been treated by many in the social sciences as an indication of a regression model's ability to accurately describe data. As J.S. Cramer (1987) explains,

> These measures of goodness of fit have a fatal attraction. Although it is generally conceded among insiders that they do not mean a thing, high values are still a source of pride and satisfaction to their authors, however hard they may try to conceal these feelings. (253)

Criminology is no exception. Looking at the use of regression analysis in criminology, Weisburd & Piquero (2008) examined the R-squared coefficients reported from 1968-2005 in *Criminology*, the discipline's flagship journal. In their description of the resulting data, the authors explain,

> The median R-squared value was .365. A histogram of the average R-squared values displayed in [figure 1] shows a clustering on the lower third, with rapidly decreasing numbers as average R-squared increases. In a large number of articles, R-squared values are extremely small. Some 25 percent of the 169 articles exhibit R-squared values below .20, and over 70 percent have an R-squared under .50. About one-tenth report R-squared values below 10. This means that most crime studies explain less than 50 percent of the variance beyond the threshold criterion (the mean of Y for linear models), and many models leave 80-90 percent variance unaccounted for. This occurs despite our having biased our results upward by selecting the highest R-squared value reported in each article. (472-473)

FIGURE 1

The results lead the authors to further conclude that "a sizable knowledge gap remains with respect to a deeper understanding of what processes cause people to commit crime and why certain areas experience more crime that others" (494). Additionally,

> While the low variance explained in articles in Criminology is troubling, it is more troubling that there has been little improvement in explanatory models over time. Our data suggest a negative correlation between R-squared and time, implying that criminologists reporting their work in Criminology are doing somewhat worse as the discipline matures. (488)

These concerns over the small R-squared values in criminology become even more disconcerting in light of the non-substantive influences that can significantly inflate the reported variance explained.

Non-Substantive Influences on Variance Explained

Conceptually, variation in the values of data can be thought of as a mixture of actual relationships between variables (signal) and the stochastic or random variation that is unexplainable (noise). Regression modeling used to describe the variation in a dependent variable through its relationship with the independent variables in the model, but certain features of the data (e.g., sample size) can have major impacts on its ability to distinguish the signal from the noise.

Examining "sample size *n* as a variable," Cramer (1987) documented the effects small sample sizes have on the ability of regression analysis to accurately describe data variation (257). By calculating the expected values of R-squared as a function of sample size (*n*) model variables, and the amount of variation in the data that a model should explain (signal, or ϕ), Cramer was able to demonstrate the R-squared coefficient can misrepresent the precision of the model, and that this effect can vary across model and dataset characteristics.

TABLE 1

Based on his assessment, Cramer (1987) concluded the accuracy of the R-squared coefficient varies with the characteristics of a dataset, and that it should not be relied upon as an assessment of model variance explained for sample sizes smaller than 50 observations. He also discussed the adjusted R-squared as a better alternative measure for regression models with a small sample sizes, but ultimately concluded that even this improved assessment does not sufficiently represent the accuracy of such models.

Cramer (1987) was able to demonstrate the weakness of regression modeling, specifically the R-squared as an indicator of variance explained, when sample sizes are small. Given the apparent importance of regression in Criminology, as well as the generally low amount of variance explained and the vulnerabilities of such analyses to dataset characteristics, it seems all the more important to understand the other factors that can affect the accuracy and precision of this approach. The following sections of this paper contribute to such an understanding by demonstrating the extent of the influence of errors in model specification, how these misleading influences vary over model/data characteristics, and how solutions to these problems mitigate these influences with varying degrees of success.

The next section of this paper builds upon Cramer’s (1987) examination of non-substantive features of regression analysis that can influence the goodness-of-fit statistic, but shifts focus from the biasing influence of small sample sizes to a focus on biasing influences of issues in model specification.

Simulating the Influence of Model Specification on Variance Explained

Despite the apparent objectivity associated with statistics, there is actually a lot of subjectivity and room for error in the application of regression modeling. Among the many influences on the accuracy of the analysis, model specification (or model construction) is a common source of difficulty. Ideally, variables are identified for inclusion in the model prior to any examination of the data and according to a theoretical rationale. When this isn’t done, and variables are included in the model without any expectation of a relationship (over-specification) or based on relationships merely identified through examination of the data sample (post-hoc specification), the resulting model can significantly over-estimate any relationships present in the data.

Similar to Cramer’s (1987) discussion of influences on the model variance explained statistics, this paper is concerned with the effects of sample size and model variables on the validity of these measures. This paper, however, builds upon Cohen’s discussion by extending focus to larger sample sizes and a much wider range of model variables. Additionally, and unlike Cramer, this paper isolates the effects of over-specification and post-hoc model specification by demonstrating their effects on the analysis of data where there should be no relationship to explain.

To clearly demonstrate the extent of the misrepresentations caused by those problems in model specification, datasets of normally distributed and random values were analyzed under conditions simulating these problems. In such datasets, the amount of variance that can be explained by genuine relationships among the variables is zero, since all the values are completely random and independent of one another. However, one cannot go so far as to refer to this as representing a null-effect. It is critical to understand the values in these datasets are not perfectly uncorrelated (Kluger & Tikochinsky, 2001), but merely correlated on the basis of random chance variation or error instead of any genuine or meaningful relationship in the data (Wilson & Shadish, 2006). As a result, these random-value datasets create a condition under which any amount of variation explained by an analysis *should* be zero, and any result greater than zero can be entirely attributed to the combined influence of randomly occurring co-variance and the statistical method used to describe the data. The results described in the following sections are the product of such simulations of model specification problems on random data, which were run using STATA 11.

Problem: Over-Specification

As described earlier, the R-squared statistic is a measure of the variance in the data explained by the model. In other words, it is a measure of how well the model is able to explain the variation that occurs in the dependent variable. More specifically, the measure compares all the variation in the values of the dependent variable (which are not captured by the mean) with the improved explanation of that variation achieved by the regression model. As mentioned earlier, though, there is always some common variation (or co-variation) that occurs between variables merely by chance, even if it is only a very small amount. So, every variable that is added to a model adds at least some explanation of the variation in the dependent variable, even if it is very small and entirely by chance.

It is possible to see this effect by running a regression analysis on a random-value dataset, as previously described. For instance, running a regression model to explain the variation in a dataset of 200 individual observations over 1 dependent variable and 10 independent variables produced an R-squared statistic of 0.0731. This value indicates the regression model explains 7.31% of the variation in the values of the dependent variable, which seems quite high given the description is based entirely on the random chance co-variation in the data.

Each iteration produces another R-squared value (e.g. 2nd time: 0.0720; 3rd time: 0.0713; 4th time: 0.0274; etc.), and while each additional value is almost always different than the previous, it is often not zero. The more the process is repeated (i.e., the greater the number of iterations of the simulation), the more apparent the tendency of these values to fall around a non-zero value. Using a Monte Carlo simulation to repeat the process 1000 times produces a fairly clear distribution of these values.

FIGURE 2

The majority of the R-squared statistics fall around a central tendency of approximately 0.05, or 5% of variance explained. Some values fall below this central tendency (as low as 0.00), while other values fall far above it (as high as 0.17). While the occasional R-squared value reached as high as 0.17, it is more important to recognize the values only very rarely approach the 0.00 level, as they should given the nature of the datasets. Similar to Cramer's examination, this misrepresentation of variance explained can vary by the number of variables in the model and the number of observations in the dataset. Adding variables to the model increases the variance explained, while the increase in the sample size lessens the effect of additional model variables. The bias recognized in Cramer's (1987) assessment is even more prominent when the theoretical R-squared coefficient (ϕ) is set at zero. Unlike Cramer (1987), the results offer the mean R-squared values for a larger range of selected values for both number of observations (n) and for the number of model variables (v).

TABLE 2

FIGURE 3

The biasing effect of smaller sample sizes continues to diminish as the sample size increases. The addition of further variables to the model, however, becomes even more pronounced as the number of independent variables increases from ten to one-hundred.

TABLE 3

FIGURE 4

The cumulative results of these simulations demonstrate the need to be skeptical of the R-squared statistic as a valid measure of variance explained, as it can become quite large independent of any genuine relationship among the variables in the model. Examined in terms of criminology, we can see that this upper extreme corner contains several values that are greater than the median .365 R-squared value Weisburd & Piquero (2008) identified in the study of reported regression results in *Criminology* (those values to the upper-right of the bold line). Beyond median values, several of the values on this table cross the .365 median within two standard deviations of the mean (those values to the upper-right of the thin line).

As remarkable as these results are, they are not especially disconcerting given that the calculation of the R-squared coefficient makes no consideration of the number of variables included in the model. As it is calculated, adding further variables to the model can only ever increase the R-squared, even if only by a small amount. To correct for the bias created by this effect, the adjusted R-squared coefficient is calculated to account for the addition of further variables to the model.

Solution: Adjusted R-squared

The root of the problem with the R-squared statistic is that the amount of variance explained by a model can only ever increase. Due to the way it is calculated (sum of the squared variation in the dependent variable explained by the model/ sum of the squared total variation in the dependent variable), there is no consequence for adding more variables to a model. Over-specifying a model to include variables with even very small amounts of chance co-variation can only increase the variance explained because the R-squared has no penalty for including additional variables in the model. The adjusted R-squared corrects this issue.

$$adjusted\ R^2 = 1 - \frac{\frac{Imprecision\ of\ the\ Regression\ Model}{sample\ size - number\ of\ variables\ in\ the\ model}}{\frac{Imprecision\ of\ the\ Mean}{sample\ size - 1}}$$

Conceptually, the adjusted R-squared accounts for the amount of chance co-variation each variable contributes to the model by effectively "raising the bar" of how much variation in the dependent variable the model must explain, given the number of model variables and sample size. By raising the bar, the adjusted R-squared *adjusts* the distribution of the measure of variance explained to reduce the influence of any chance co-variation, thus leaving only the genuine co-variation between that model variables and the dependent variable to improve the model.

FIGURE 5

Comparing the distribution of R-squared values created earlier with the distribution of adjusted R-squared values reported for the same 1000 random-value datasets, it is apparent that the adjusted R-squared reduced the misrepresentation of the R-squared for over-specified models. However, the effect is not perfect. Not all values become zero; instead, the distribution of R-squared values that had been centered around 5% has been shifted (or adjusted) to now fall around zero. In fact, because the shift has almost no effect on the shape of the distribution, but only its location, many of the adjusted R-squared values are negative. The distance below zero these values fall is not as important as the fact that they are less-than-or-equal-to zero, which ultimately means that the model does not actually explain any of the variation in the dependent variable. This same effect operates across different sample sizes and number of model variables.

FIGURE 6

The corrective effect also works in the extreme cases where there were very large numbers of variables included in the model. Likewise, central tendencies in table 3 are adjusted to 0.000, with occasional ± 0.005.

FIGURE 7

Importantly, the adjusted R-squared statistic is not a complete solution to the problem of over-specification. The distributions demonstrate that there are still many values that fall far from the zero-level expected of a random-value dataset. Additionally, the distribution of these values stretch further as the number of model variables increases and the sample size decreases. Nevertheless, the adjusted R-squared provides an important and largely effective protection against the problem of model over-specification causing inflated measures of variance explained by the model.

Solution: Model Significance

Concerning the influence of dataset characteristics on regression analyses' description of data, it is clear there is a significant misrepresentation of a model's accuracy conveyed by the R-squared coefficient, which exerts more influence as sample sizes decrease and the number of variables included in the model increases. However, this bias in the description of a model's accuracy is almost entirely mitigated by the adjusted R-squared coefficient's ability to account for the incorporation of additional variables into the regression model. In addition to correctly recognizing the absences of a relationship to describe in the data, the model significance also guarded against accepting these descriptions of the data as a description of reality.

As can be seen in figures 6 & 7, the adjusted-R-squared merely shifts the central tendency to zero, where we know it should be; there is still a distribution of values that fall above and below zero. The distribution of values that fall below zero are not much of a problem because their interpretation allows them to be erroneous: such models would explain X percent *less* variance than the mean of the dependent variable alone. The distribution of values above zero, however, still appear valid, even though they are lower than those provided by the R-squared.

To address this, the amount/values of model significance depicts the likelihood of the observed results are due to chance. This reduces the chance of falsely accepting the model as a description of a relationship that doesn't exist (i.e., mistakenly interpreting noise as signal, also known as type I error). Figure 8 depicts how the distribution of variance explained is shifted by the adjusted-R-squared from the R-squared. As can be seen, there is still a portion of the results that are considered significant, but it is less than half the distribution; it is only the 5% of the values that fall far above the central tendency. Figure 9 shows how this influence is affected as a function of sample size and model variables, and how increased sample size increases the precision of the identified relationship (statistical power/type II error).

FIGURE 8

FIGURE 9

Additionally, the mere presence of a large number of variables in a model with only a modest amount of variation explained presents a prima facie indication the model may be over-specified. For instance, one should be skeptical of any model with as many variables as those simulated in the latter examples. Even where the adjusted R-squared is reported, there is still reason to be suspicious of a model with so many variables without a compelling theoretical rationale for including so many variables in the model. Such solutions, however, provide little protection from the consequences of post-analysis (or post-hoc) model specification.

Problem: Post-Analysis Specification

Different than the problem of over-specifying a model by including too many variables that only coincidentally capture the co-variance, post-hoc model specification includes only those few variables that explain a high level of covariance in the data.

Ideally, models should be based on a theoretical expectation of a relationship between variables, rather than "cherry-picking" the variables in the dataset that happen to have large amounts of co-variation with the dependent variable. This can result in models that merely account for chance co-variation, rather than the genuine relationships in the data. For instance, say we are interested in a phenomenon (either a relationship or occurrence) that is present very rarely, perhaps only in 5% of cases. If we observe 100 cases, we should expect to find about 5 instances of this phenomenon. If we were to then ignore the other 95 cases and treat the data from those 5 cases as if they appeared from an observation of only 5 cases, then any analysis of that data would produce a biased description.

The bias created by such a practice has not been completely ignored in the social science literature (Leeb & Potscher, 2005), or even by some in criminology (Berk, Brown, & Zhao, 2010). In an excellent technical explanation of the problem of post-hoc specification, Berk, Brown, & Zhao (2010) review the mathematical and statistical errors caused by post-hoc estimation, and explain how they can affect the statistical significance of the model. Though these and other authors explain the *severe* consequences of engaging in post-hoc model specification, their cautions have not raised much attention within criminology. This section is intended to provide an extremely clear demonstration of the extensive influence post-hoc model specification can have on regression analysis and statistical inference through the same approach used in the previous section, but with slightly modified simulations.

To provide a more conceptual demonstration of the problem, this section uses stepwise regression to simulate the effects of post-hoc. Instead of simply creating a random-value dataset and adding all the variables to the model, the "cherry-picking" aspect of post-hoc model specification is simulated by stepwise regression, which essentially evaluates the significance of the available variables to build a model of only those variables that are significant at the .05 level.

Since it would not be realistic to compare models based on the number of variables *available* in the dataset, as this is not a conventionally reported element of the analysis, it is fairer to compare based on the appearance of the actual models themselves. So, after the stepwise regression was run for 190 possible combinations of dataset characteristics, they were reclassified by the number of variables included in the model.[1]

In the first simulation, the adjusted R-squared coefficient almost completely negated the R-squared coefficient's misleading assessment of the model's accuracy by accounting for the number of variables in the model, relative to their contribution to the increased accuracy of the model. That corrective effect does not provide the same level of benefit in the case of post-hoc model specification.

FIGURE 10

TABLE 4

FIGURE 11

In the first simulation, the adjusted R-squared was able to effectively reduce the biased description of the data provided by the R-squared across dataset characteristics. By accounting for the number of variables (*v*) included in the model, the adjusted R-squared was able to recognize there was essentially no variance being explained by the model. In the case of post-hoc specification, however, many fewer variables are actually included in each model, so there is less opportunity for the corrective effects of the adjusted R-squared coefficient to neutralize an erroneous description of the data where no explanation exists. The adjusted R-squared is only accounting for the small amount of over-specification that is occurring in the models. Model significance also doesn't help because the model is specifically defined by its significance. This creates major cause for concern, as these reported values appear valid according to conventional measures of goodness-of-fit and statistical significance, even where we know there is no real effect in the data on which to base the reported relationships.

[1] Interestingly, one would expect that only 5% of the available variables would be included in the model when significance is set at .05. However, it appears that there is an exponential increase in the number of model variables that are significant as the available number of variables increases.

The inflated R-squared values resulting from over-specification, which approached and exceeded the median values reported in *Criminology*, were not a major cause for alarm among criminologists for several reasons: 1) the adjusted R-squared greatly moderated the biased estimate; 2) the model significance indicated that most of the models were little more than a product of chance co-variation; and, 3) even a prima facie inspection of model size raises suspicion as the larger R-squared values resulted from unreasonably large models with up to 100 independent variables. This is not true of the high adjusted R-squared values resulting from post-hoc specification. While these values also approach the median values reported in *Criminology*, the same guards provide no protection from interpreting their results as valid: 1) the adjusted R-squared only slightly adjusts the distribution of R-squared values; 2) all these models are necessarily significant as each included model variable was specifically selected *because* of its significance; and, 3) a prima facie inspection does not raise alarm as these models are all reasonably sized with 10 or fewer model variables.

Barriers to a Statistical Solution

Unfortunately, there is not an easy solution to the problem of post-hoc specification. The adjusted R-squared was able to correct for the effects of over-specification by "raising the bar" of variation to be explained for each variable added to the model. A similar solution could be applied to the problem of post-hoc model specification. By raising the bar for each variable added or removed from the model. However, the number of attempted specifications is not as self-evident as the number of variables in the model. While it would be very difficult to credibly report the results of a model without also providing mention of the number of variables in the model, it is very easy (and common) to omit any discussion of the attempts to specify a model.

Model Significance

FIGURE 12

FIGURE 13

This number of attempts would be critical for calculating a re-adjusted R-squared value to address the issue of post-hoc specification. However, there is little incentive to report the number of attempts since each attempt quickly reduces the amount of variation explained by the model. The incentive to report is even further reduced by the lack of any professional expectation to do so. Given the sizeable effects on the variance explained caused by post-hoc specification, the lack of an incentive structure raises several important considerations for the discipline of criminology.

Considerations for the Discipline of Criminology

Up to this point, much of the discussion in this paper has focused on the analytic difficulties associated with model specification in regression analysis. This concluding section, however, moves the conversation of these challenges from a statistical context into a context of the larger discipline of criminology and criminal justice education.

In the conclusion of their examination of the R-squared values reported in "Criminology," Weisburd & Piquero (2008) pose the following questions.

> Perhaps most important, criminologists need to pay much greater attention to what they do not explain. Are low R-squared values due to inadequate theory or poor data, or is there some more general principle operating that limits ability to explain crime and criminals? To what extent can we assume that unexplained variance is stochastic, and to what extent must we accept that systematic biases enter into the quantitative models we develop. (493)

These questions begin to bridge the gap between an individual application of a statistical analyses in a single study with a larger discussion of disciplinary development. Concerning the influence of post-hoc model specification as an issue within criminology, consideration of education, research replication, and theory testing/development are the most important aspects of the discipline for offering protection against overstated or erroneous analyses entering the discipline's knowledge-base as valid descriptions of criminological phenomena.

The first place to begin protecting criminology from the effects of post-hoc specification begins with education. As mentioned earlier in the paper, regression is a commonly taught statistical method in graduate criminology programs, making it something many criminologists are at least familiar with. That familiarity should include an understanding of the problems and consequences of post-hoc model specification. Equipped with such an appreciation, it would be possible to reduce at least the danger of criminologists accidentally or unknowingly engaging in the practice. Relatedly, members of the profession should more explicitly demonstrate their commitment to more transparent statistical practices by reporting and expecting some mention of the number of attempts to specify a model.

Replication is another avenue for protection against the effects of post-hoc model specification. While actually replicating a study would be one excellent means of validating the results of a prior study as more than a random flash of co-variation, it is not the only option. Combining like studies through meta-analysis provides a basis for recognizing how far the findings of any one study might depart from the findings of related studies, possibly as the result of issues in model specification (for a discussion of the practice of meta-analysis, see Lipsey & Wislon (2001); for a discussion of meta-analysis in criminology, see Pratt (2010)). Similarly, a Bayesian approach to statistical inference examines an individual study's results relative to the distribution of effects reported in prior and related studies (for a discussion of the practice of Bayesian analysis, see Jackman (2009); for a discussion of Bayesian analysis in criminology, see Sullivan & Mieczkowski (2008)). It is even possible to promote replication within a single study and without reference to prior research, which is especially useful where no similar research exists. By simply dividing a dataset into two random sub-samples prior to any examination of the data, a final model can be specified through multiple attempts in one of the samples, and then be applied to the other sub-sample (Fox, 2008).

A final area of disciplinary consideration is the connection between data and theory. The importance of recognizing the process of research as an ongoing cycle between deduction and induction cannot be understated. On the deductive side of the cycle, theory guides expectations about the relationships between variables. On the inductive side of the cycle, empirical observations of relationships among variables guide what theory must explain. Viable and valid understandings of criminological phenomena must be able to continually circulate through this research cycle. Focusing on only one side without also considering the other can lead to erroneous conclusions about the social phenomena under study.

In terms of post-hoc model specification, focusing merely on induction from relationships observed in a dataset, without also considering any previously conceived theoretical explanation of such a relationship, could lead one to conclude a genuine relationship exists between variables where it is actually nothing more than chance co-variation that explains the apparent relationship. Importantly, the deductive side of the research cycle provides an opportunity to make an important point and final clarification about the issue of model specification.

Throughout this paper, it has been argued and demonstrated that issues of model specification, particularly post-hoc model specification, can have incredible effects on the ability of regression analysis to distinguish genuine relationships among variables in a dataset from the mere chance co-variation that inevitably occurs in all data. These effects can be quite large, and appear to provide better explanations of criminological phenomena than even some of those analyses published in the discipline's flagship journal. Despite this apparent threat to the validity of statistical research in the discipline, the process of exploratory model specification still holds an important role in the larger cycle of criminological research. By describing observed relationships in a given dataset, even if done so erroneously through over-specification or post-hoc specification, the results of regression modeling provide an important basis for theory testing and development, which then subsequently informs future expectations for observable relationships among data. It is not until a study is considered outside this cycle between data and theory, or induction and deduction, that issues of model specification truly begin to threaten the validity of criminological research. By addressing the threats of model specification in criminal justice education, including consideration of it in publications, and appreciating the impact of a single study's analyses as only part of a larger process of knowledge development in criminology, the discipline can reduce the threat of these analysis effects.

References

Anderson, D. R., Burnham, K. P., & Thompson, W. L. (2000). Null hypothesis testing: Problems, prevalence, and an alternative. *The Journal of Wildlife Management*, *64*(4), 912-923.

Berk, R., Brown, L., & Zhao, L. (2010). Statistical inference after model selection. *Journal of Quantitative Criminology, 26*, 217-236.

Blankenship, M. B., & Brown, S. E. (1993). Paradigm or perspective? A note to the discourse community." *Journal of Crime and Justice, 16*, 167-175.

Boruch, R. (2007). The null hypothesis is not called that for nothing: Statistical tests in randomized trials. *Journal of Experimental Criminology*, *3*(1), 1-20.

Bösch, H., Steinkamp, F., & Boller, E. (2006). In the eye of the beholder: Reply to Wilson and Shadish (2006) and Radin, Nelson, Dobyns, and Houtkooper (2006). *Psychological Bulletin, 132*(4), 533-537.

Bushway, S., Sweeten, G., & Wilson, D. (2006). Size matters: Standard errors in the application of null hypothesis significance testing in criminology and criminal justice. *Journal of Experimental Criminology*, *2*(1), 1-22.

Clear, T. R. (2001). Has academic criminal justice come of age - ACJS presidential address Washington, DC, April 2001. *Justice Quarterly*, *18*, 709-726.

Cramer, J. S. (1987). Mean and variance of R^2 in small and moderate samples. *Journal of Econometrics*, *35*(2-3), 253-266.

Fox, J. (2008). *Applied regression analysis and generalized linear models* (2nd ed.). Los Angeles, CA: Sage Publications.

Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods*, *1*(4), 379-390.

Heckman, J. J., & Smith, J. A. (1995). Assessing the case for social experiments. *The Journal of Economic Perspectives*, *9*(2), 85-110.

Hotelling, H. (1940). The selection of variates for use in prediction with some comments on the general problem of nuisance parameters. *The Annals of Mathematical Statistics*, *11*(3), 271-283.

Jackman, S. (2009). *Bayesian analysis for the social sciences*. West Sussex, UK: John Wiley and Sons, Ltd.

Kluger, A. N., & Tikochinsky, J. (2001). The error of accepting the "theoretical" null hypothesis: The rise, fall, and resurrection of commonsense hypotheses in psychology. *Psychological Bulletin, 127*(3), 408-423.

Leeb, H., & Pötscher, B. M. (2005). Model selection and inference: facts and fiction. *Econometric Theory*, *21*(1), 21-59.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Applied Social Research Methods Series (Vol. 49). Thousand Oaks, CA: Sage Publications.

Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review*, *1*(4), 476-490.

Lytle, D. J., & Travis, L. F. (2008). Graduate education in criminology or criminal justice: Assessing course requirements. *Journal of Criminal Justice Education*, *19*(3), 339-350.

Maltz, M. D. (1994). Deviating from the mean: The declining significance of significance. *Journal of Research in Crime and Delinquency*, *31*(4), 434-463.

Pratt, T. C. (2010). Meta-analysis in criminal justice and criminology: What it is, when it's useful, and what to watch out for. *Journal of Criminal Justice Education*, *21*(2), 152-168.

Sarat, A., & Silbey, S. (1988). The pull of the policy audience. *Law & Policy*, *10*(2-3), 97-166.

Southerland, M. D. (2002). Criminal justice curricula in the United States: A decade of change. *Justice Quarterly*, *19*(4), 589-601.

Sullivan, C., & Mieczkowski, T. (2008). Bayesian analysis and the accumulation of evidence in crime and justice intervention studies. *Journal of Experimental Criminology*, *4*(4), 381-402.

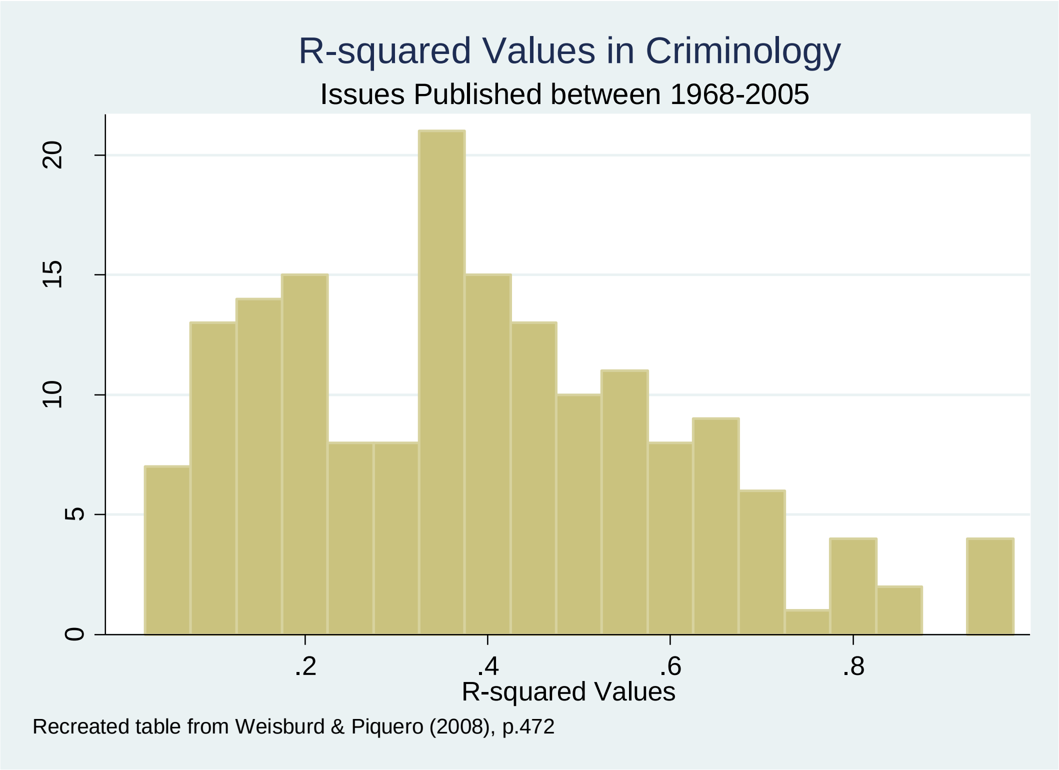
Sullivan, C. J., & Maxfield, M. G. (2003). Examining paradigmatic development in criminology and criminal justice: A content analysis of research methods syllabi in doctoral programs. *Journal of Criminal Justice Education*, *14*(2), 269-285.

Weisburd, D., & Piquero, A. (2008). How well do criminologists explain crime? Statistical modeling in published studies. In M. Tonry (Ed.), *Crime and Justice: A Review of Research* (Vol. 37, pp. 453-502). Chicago, IL: University of Chicago Press.

Wiles, R., Durrant, G., De Broe, S., & Powell, J. (2009). Methodological approaches at PhD and skills sought for research posts in academia: A mismatch? *International Journal of Social Research Methodology*, *12*(3), 257-269.

Wilson, D. B., & Shadish, W. R. (2006). On blowing trumpets to the tulips: To prove or not to prove the null hypothesis--comment on Bösch, Steinkamp, and Boller (2006). *Psychological Bulletin*, *132*(4), 524-528.

Figure 1



Recreated table from Weisburd & Piquero (2008), p.472

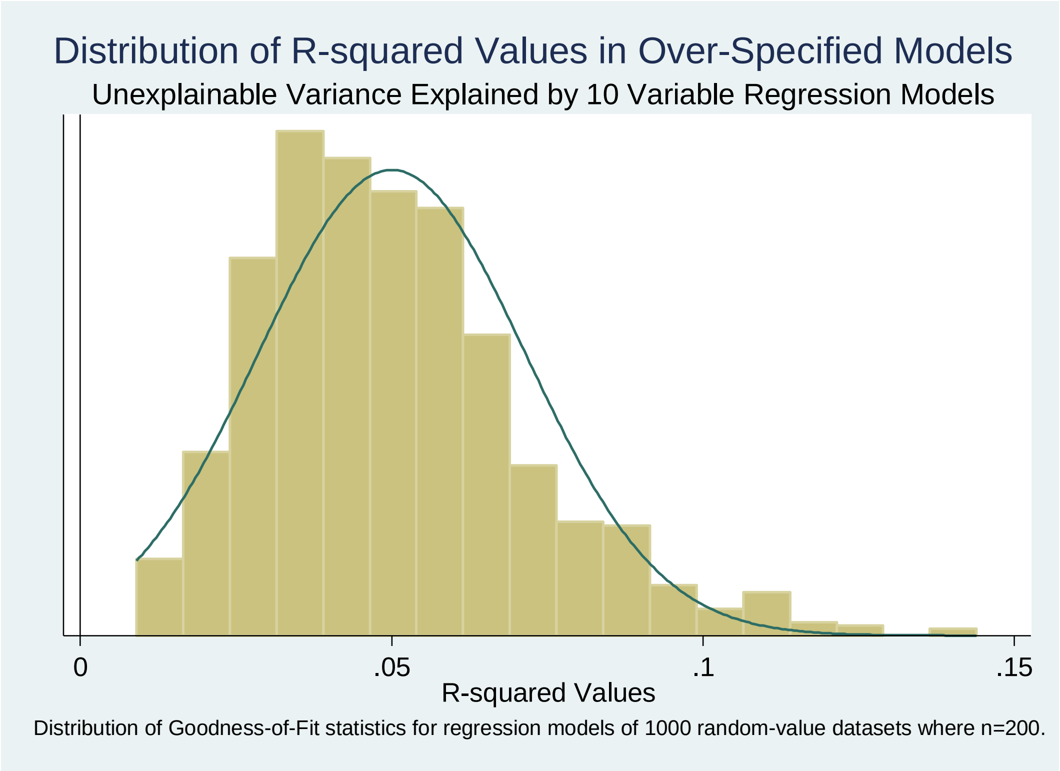
Table 1

Mean of R^2 for selected values of v, of ϕ, and of n.

	$\phi = 0.9$		$\phi = 0.667$		$\phi = 0.5$		$\phi = 0.333$	
	$v = 1$	$v = 2$	$v = 1$	$v = 2$	$v = 1$	$v = 2$	$v = 1$	$v = 2$
$n = 5$	0.936	0.957	0.760	0.840	0.620	0.747	0.485	0.657
6	0.930	0.947	0.742	0.807	0.597	0.697	0.454	0.590
7	0.925	0.940	0.730	0.784	0.581	0.665	0.433	0.547
8	0.922	0.935	0.722	0.768	0.569	0.641	0.419	0.516
9	0.920	0.931	0.715	0.756	0.561	0.624	0.408	0.493
10	0.918	0.928	0.710	0.746	0.554	0.610	0.400	0.475
20	0.908	0.914	0.688	0.705	0.526	0.552	0.365	0.400
30	0.906	0.909	0.681	0.692	0.517	0.534	0.354	0.377
40	0.904	0.907	0.677	0.686	0.513	0.526	0.349	0.366
50	0.903	0.905	0.675	0.682	0.510	0.520	0.345	0.359
100	0.902	0.903	0.671	0.674	0.505	0.510	0.339	0.346
150	0.901	0.902	0.670	0.672	0.503	0.507	0.337	0.342
200	0.901	0.901	0.668	0.670	0.502	0.504	0.336	0.338

Each cell provides the mean R-squared coefficient value (with a Poisson distribution) for regression models at each of the indicated number of independent variables (v), the indicated number of observations (n), the indicated level of explainable variance (ϕ), and calculated to a 10^{-8} level of accuracy.

Figure 2


Distribution of R-squared Values in Over-Specified Models

Unexplainable Variance Explained by 10 Variable Regression Models

Distribution of Goodness-of-Fit statistics for regression models of 1000 random-value datasets where n=200.

Table 2

Mean of R-squared for selected values of *v* and of *n*, at $\phi = 0.0$

	v = 1	*v* = 2	*v* = 3	*v* = 4	*v* = 5	*v* = 6	*v* = 7	*v* = 8	*v* = 9	*v* = 10
n = 200	**0.005**	0.010	0.015	**0.021**	0.025	0.030	**0.034**	0.039	0.046	**0.051**
250	0.004	0.008	0.012	0.015	0.020	0.024	0.028	0.033	0.036	0.040
300	0.003	0.007	0.010	0.014	0.017	0.020	0.023	0.026	0.030	0.033
350	**0.003**	0.006	0.009	**0.011**	0.015	0.017	**0.020**	0.023	0.026	**0.028**
400	0.002	0.005	0.008	0.010	0.013	0.015	0.018	0.020	0.022	0.025
450	0.002	0.004	0.007	0.009	0.011	0.014	0.016	0.018	0.020	0.023
500	**0.002**	0.004	0.006	**0.008**	0.010	0.012	**0.014**	0.016	0.018	**0.020**
550	0.002	0.004	0.005	0.007	0.009	0.011	0.013	0.014	0.016	0.018
600	0.002	0.003	0.005	0.007	0.008	0.010	0.012	0.013	0.015	0.017
650	**0.002**	0.003	0.005	**0.006**	0.008	0.009	**0.011**	0.012	0.014	**0.016**

Each cell provides the mean R-squared coefficient value for 1,000 regression models, each including the indicated number of independent variables (*v*), describing a computer generated random and uncorrelated dataset ($\phi = 0.0$) with the indicated number of observations (*n*).

Figure 3

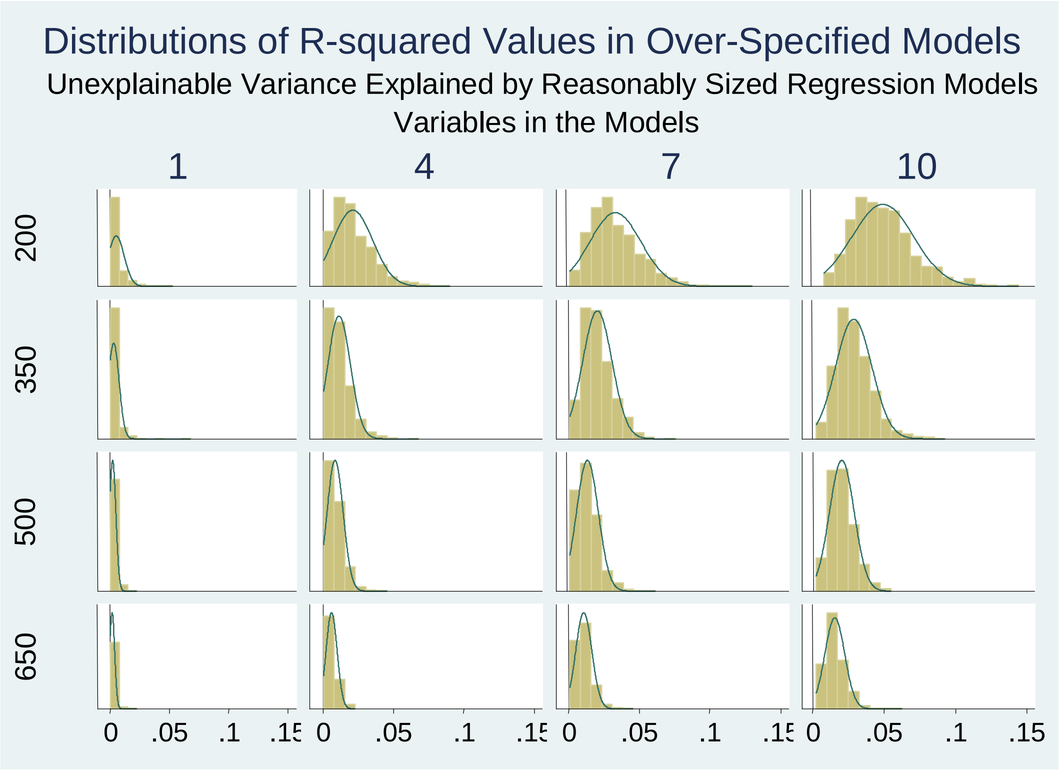


Table 3

Mean of R-squared for selected values of *v* and of *n*, at ϕ = 0.0

	v = 10	*v* = 20	*v* = 30	*v* = 40	*v* = 50	*v* = 60	*v* = 70	*v* = 80	*v* = 90	*v* = 100
n = 200	**0.051**	0.100	0.152	**0.200**	0.251	0.301	**0.353**	0.402	0.451	**0.503**
250	0.040	0.080	0.120	0.160	0.200	0.239	0.281	0.320	0.362	0.404
300	0.033	0.067	0.100	0.135	0.168	0.201	0.234	0.267	0.301	0.334
350	**0.028**	0.058	0.086	**0.115**	0.143	0.172	**0.201**	0.228	0.257	**0.286**
400	0.025	0.050	0.075	0.101	0.124	0.150	0.176	0.201	0.226	0.251
450	0.023	0.045	0.067	0.090	0.112	0.134	0.156	0.179	0.203	0.224
500	**0.020**	0.039	0.061	**0.081**	0.101	0.119	**0.141**	0.160	0.180	**0.201**
550	0.018	0.037	0.054	0.073	0.092	0.108	0.128	0.146	0.164	0.183
600	0.017	0.034	0.051	0.066	0.083	0.100	0.117	0.133	0.150	0.169
650	**0.016**	0.030	0.046	**0.062**	0.077	0.091	**0.108**	0.123	0.139	**0.154**

Each cell provides the mean R-squared coefficient value for 1,000 regression models, each including the indicated number of independent variables (*v*), describing a computer generated random and uncorrelated dataset (ϕ = 0.0) with the indicated number of observations (*n*).

Figure 4

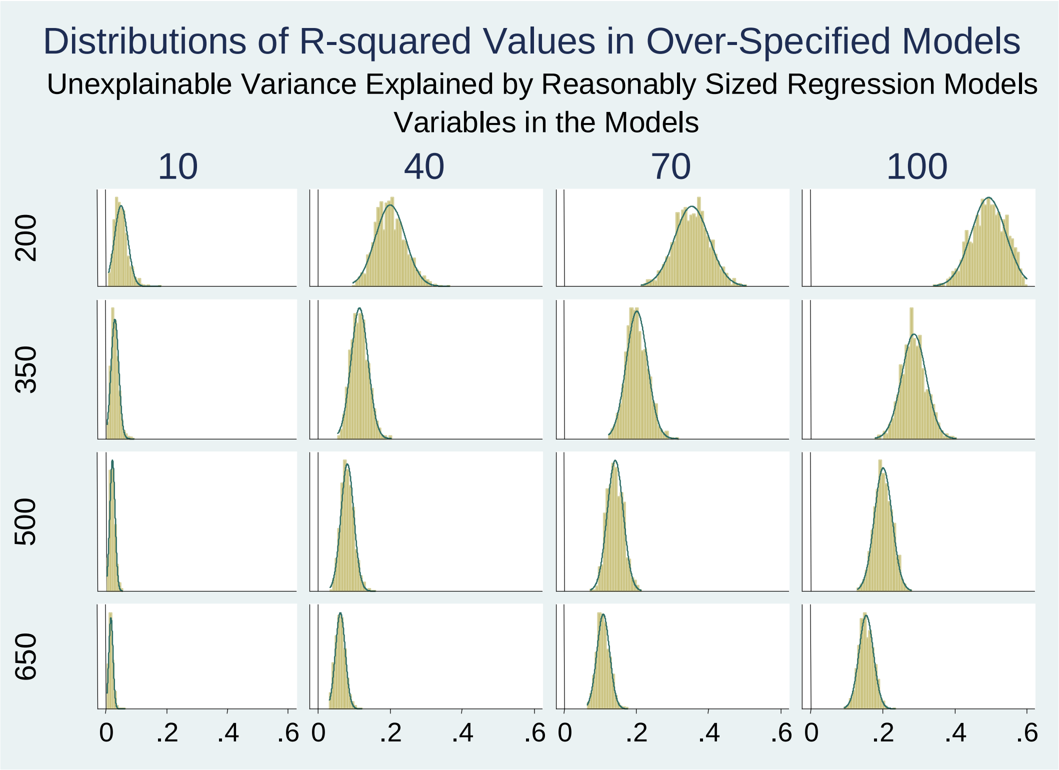


Figure 5

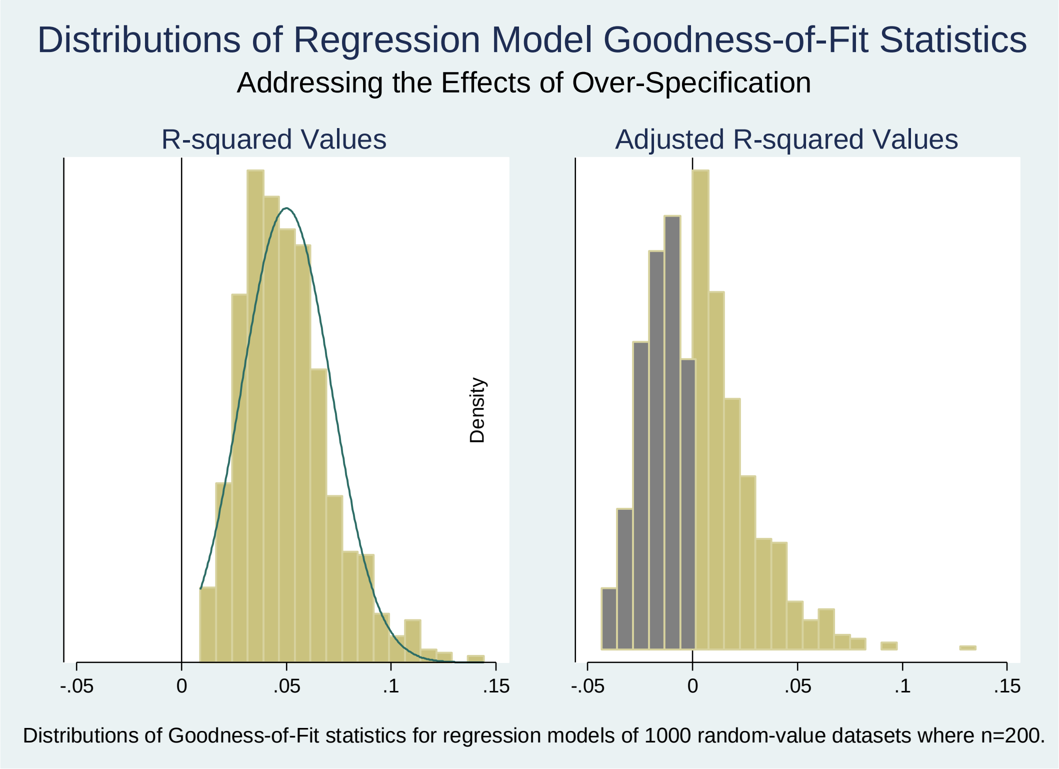


Figure 6

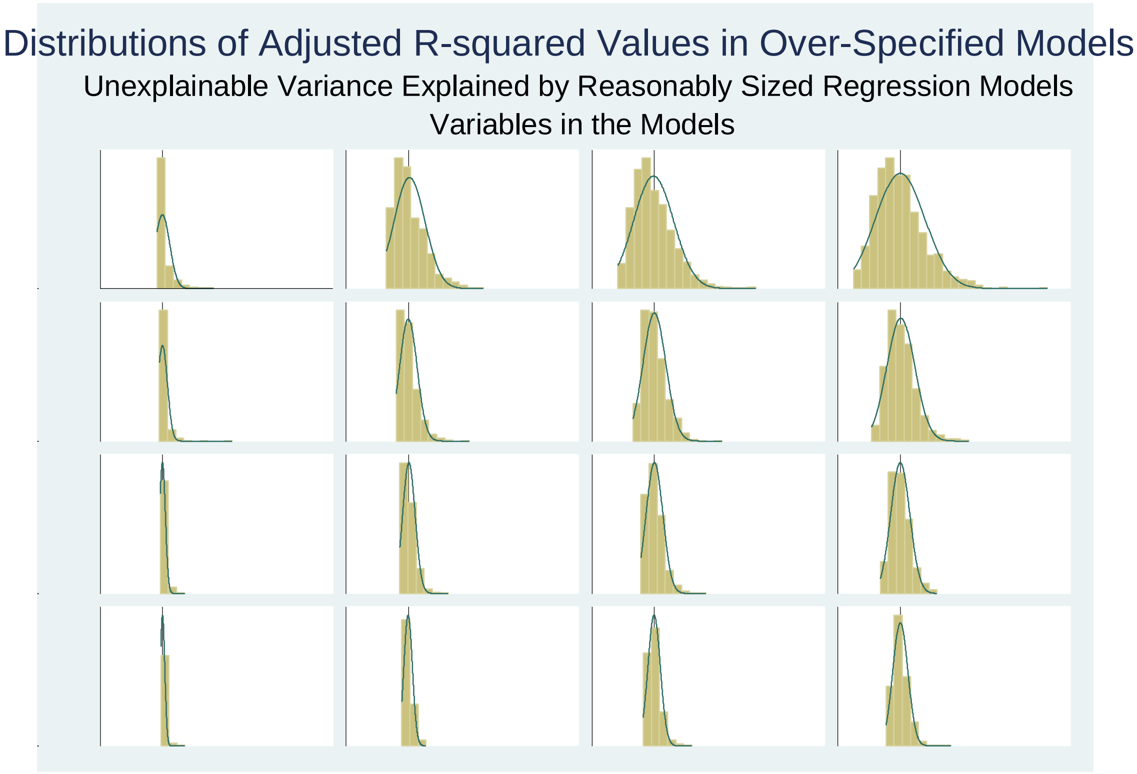


Figure 7

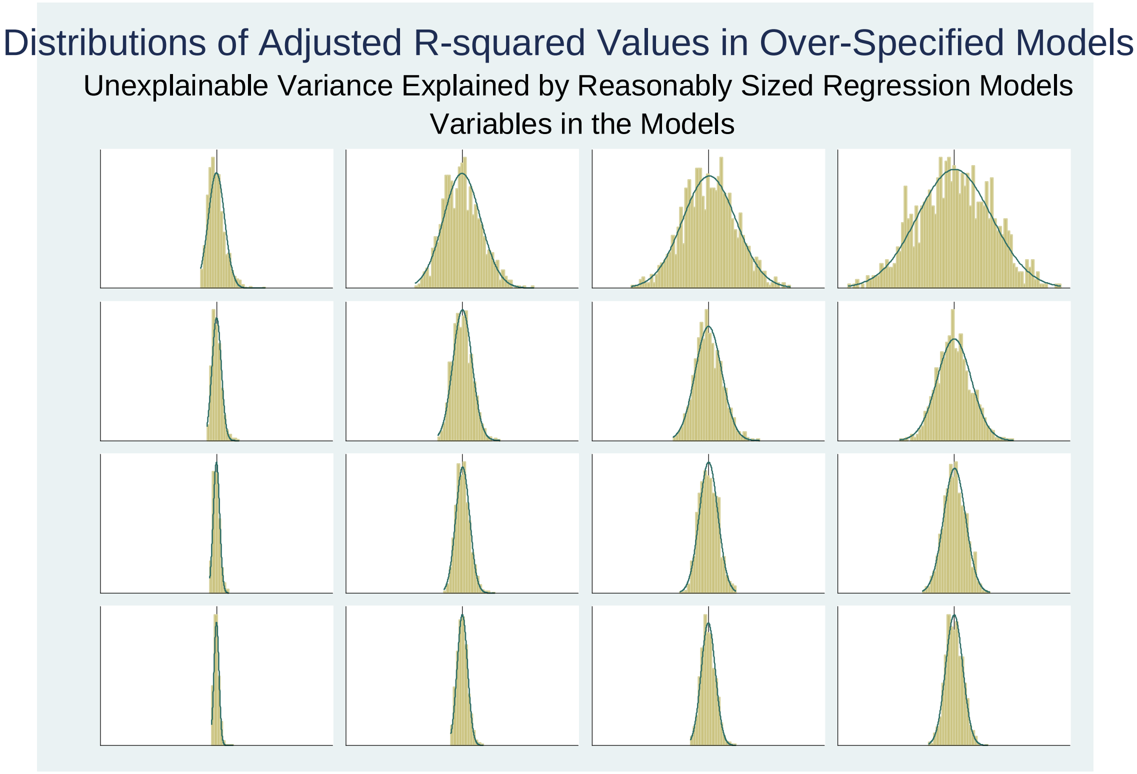
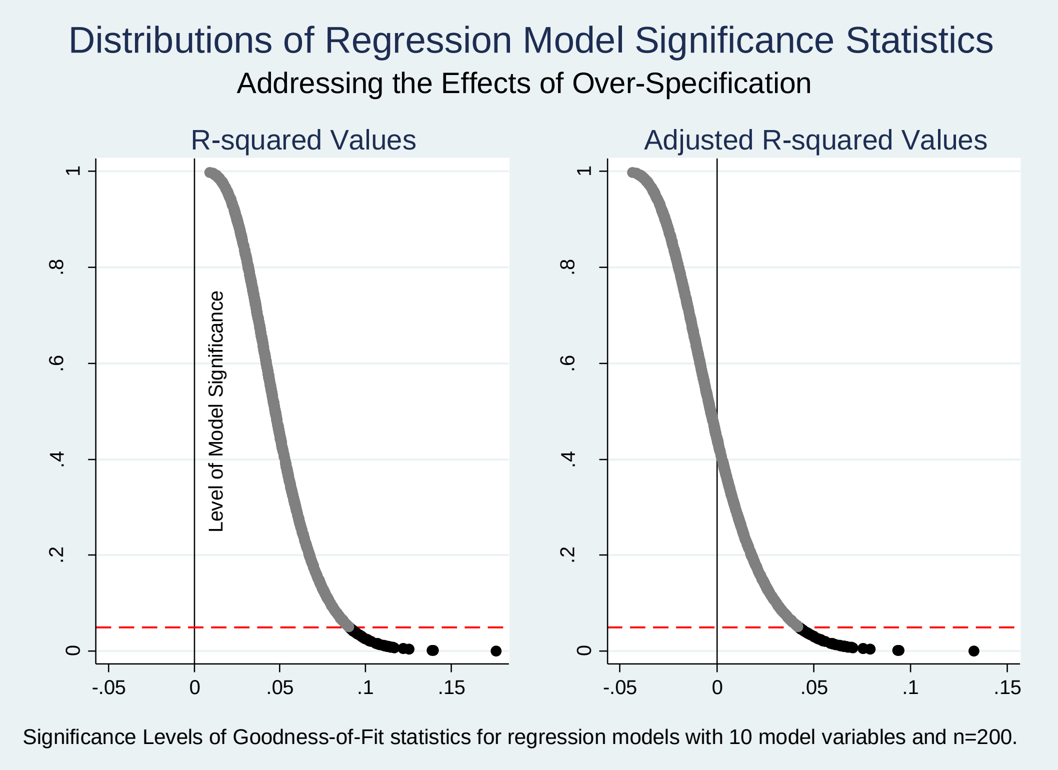


Figure 8



Distributions of Regression Model Significance Statistics

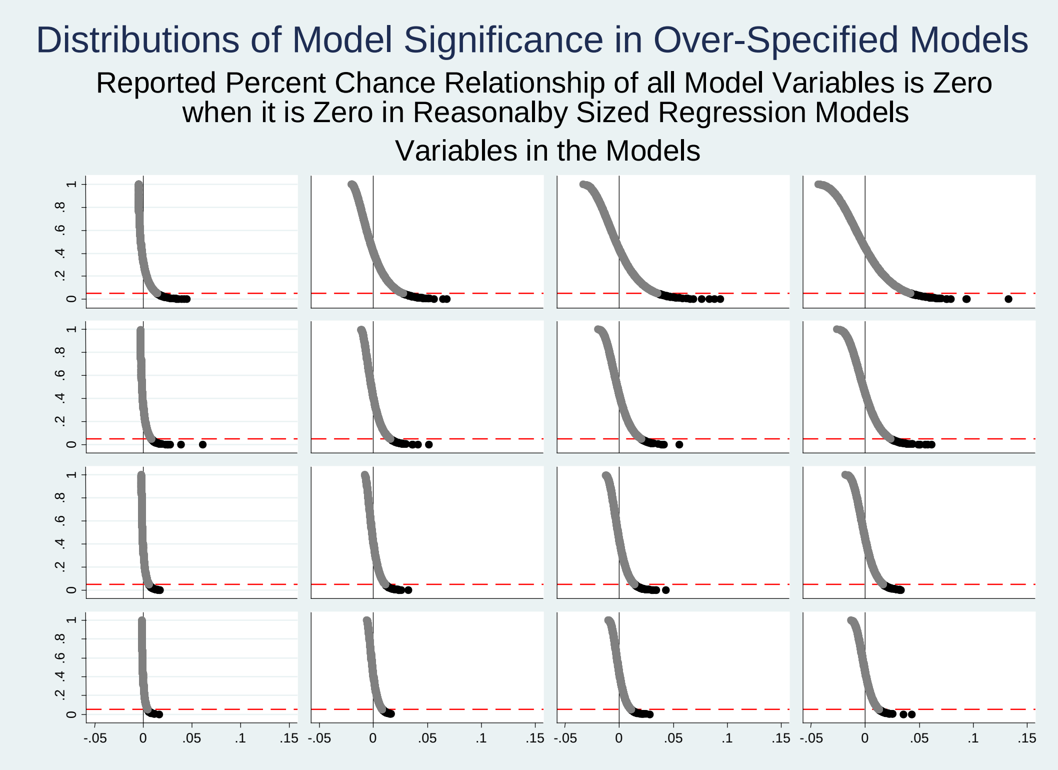
Addressing the Effects of Over-Specification

R-squared Values

Adjusted R-squared Values

Significance Levels of Goodness-of-Fit statistics for regression models with 10 model variables and n=200.

Figure 9

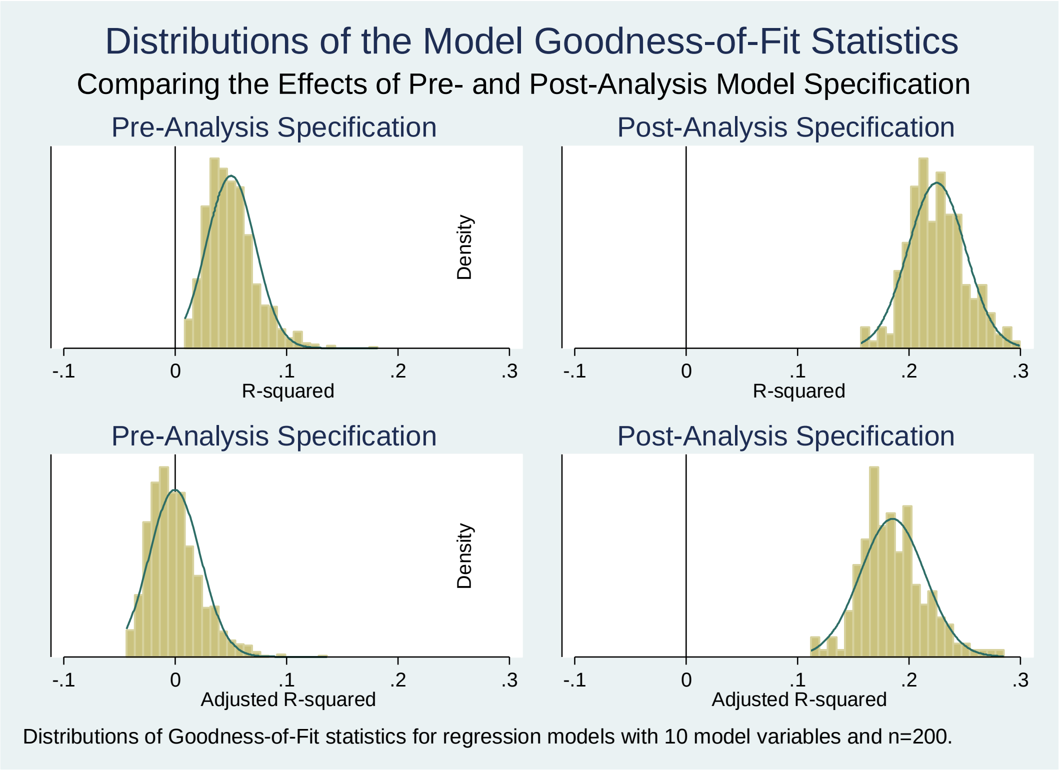


Distributions of Model Significance in Over-Specified Models

Reported Percent Chance Relationship of all Model Variables is Zero
when it is Zero in Reasonalby Sized Regression Models

Variables in the Models

Figure 10



Distributions of the Model Goodness-of-Fit Statistics

Comparing the Effects of Pre- and Post-Analysis Model Specification

Pre-Analysis Specification

Post-Analysis Specification

Pre-Analysis Specification

Post-Analysis Specification

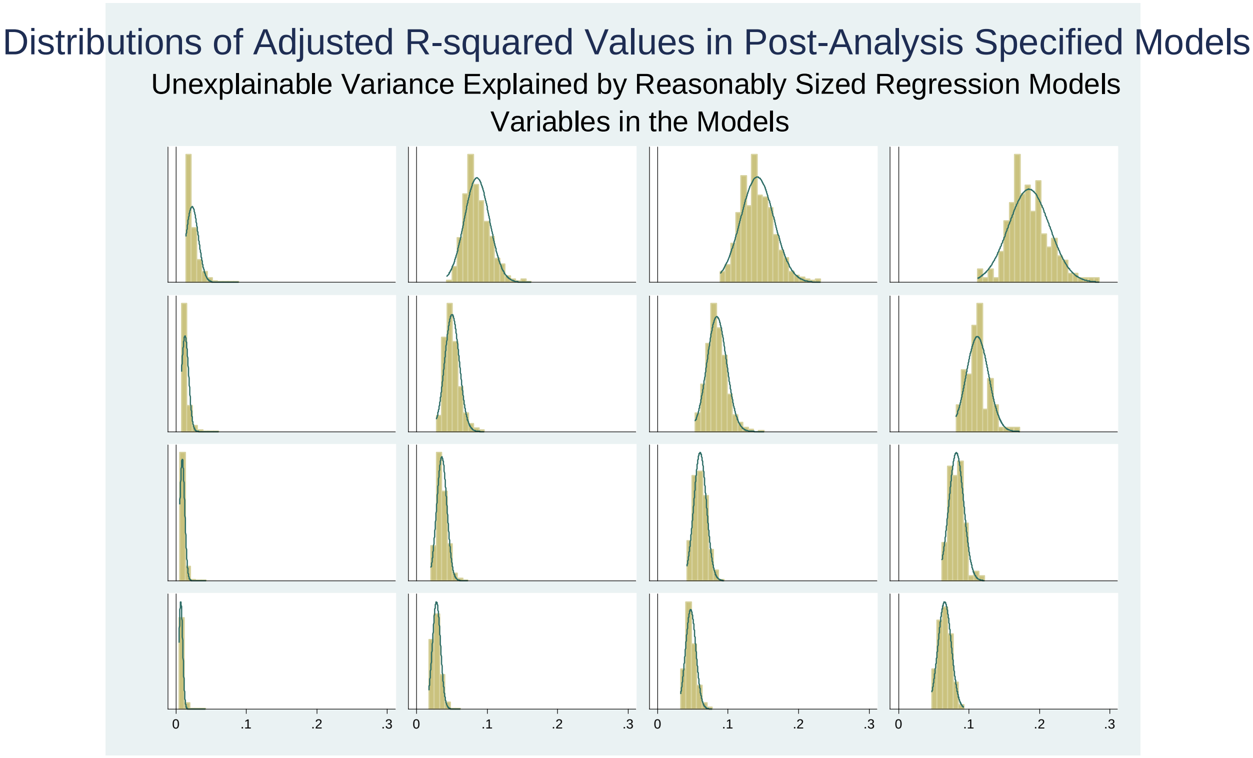
Distributions of Goodness-of-Fit statistics for regression models with 10 model variables and n=200.

Table 4

Mean of *adjusted R-squared* for post-hoc specified models at selected values of *v* and of *n*, at $\phi = 0.0$

	v = 1	*v* = 2	*v* = 3	*v* = 4	*v* = 5	*v* = 6	*v* = 7	*v* = 8	*v* = 9	*v* = 10
n = 200	**0.023**	0.045	0.066	**0.085**	0.104	0.122	**0.141**	0.156	0.173	**0.185**
250	0.018	0.037	0.054	0.069	0.086	0.100	0.114	0.129	0.141	0.153
300	0.015	0.030	0.044	0.058	0.072	0.085	0.096	0.110	0.121	0.134
350	**0.013**	0.026	0.039	**0.050**	0.061	0.074	**0.084**	0.092	0.106	**0.112**
400	0.012	0.023	0.034	0.044	0.055	0.064	0.075	0.084	0.094	0.101
450	0.010	0.020	0.030	0.040	0.049	0.057	0.066	0.075	0.085	0.092
500	**0.009**	0.018	0.027	**0.036**	0.044	0.052	**0.060**	0.067	0.075	**0.082**
550	0.008	0.016	0.025	0.033	0.040	0.048	0.055	0.062	0.068	0.076
600	0.008	0.015	0.023	0.030	0.036	0.045	0.050	0.057	0.063	0.071
650	**0.007**	0.014	0.021	**0.028**	0.034	0.041	**0.047**	0.053	0.060	**0.066**

Figure 11

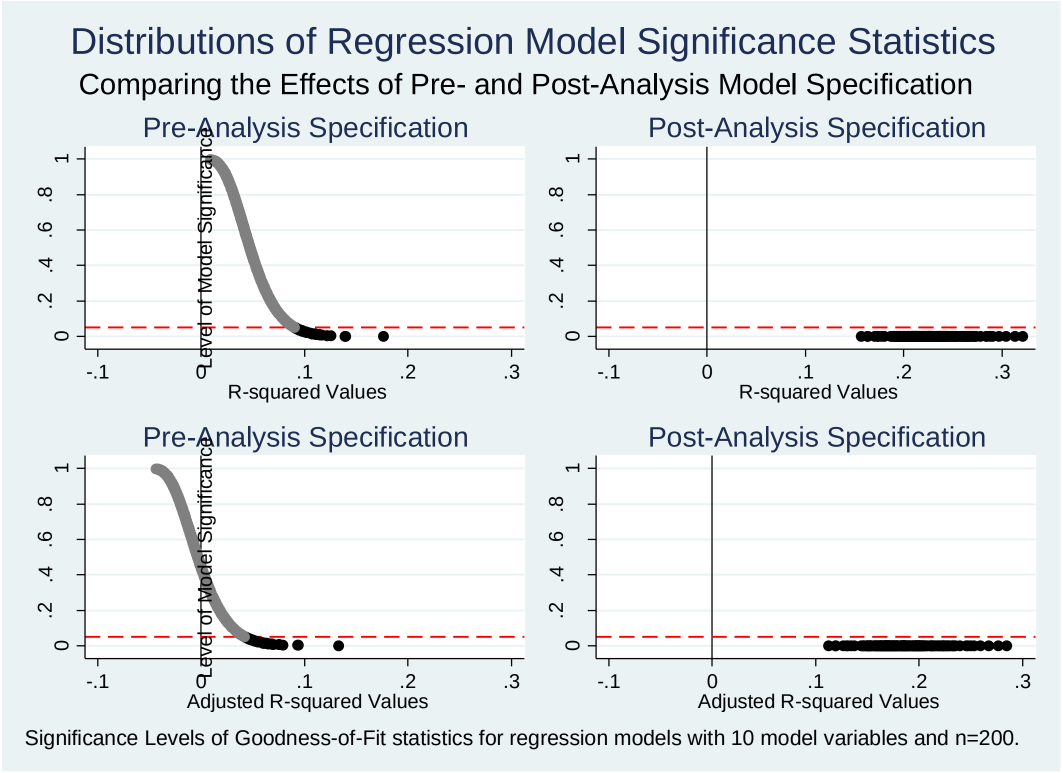


Distributions of Adjusted R-squared Values in Post-Analysis Specified Models

Unexplainable Variance Explained by Reasonably Sized Regression Models

Variables in the Models

Figure 12



Distributions of Regression Model Significance Statistics

Comparing the Effects of Pre- and Post-Analysis Model Specification

Pre-Analysis Specification

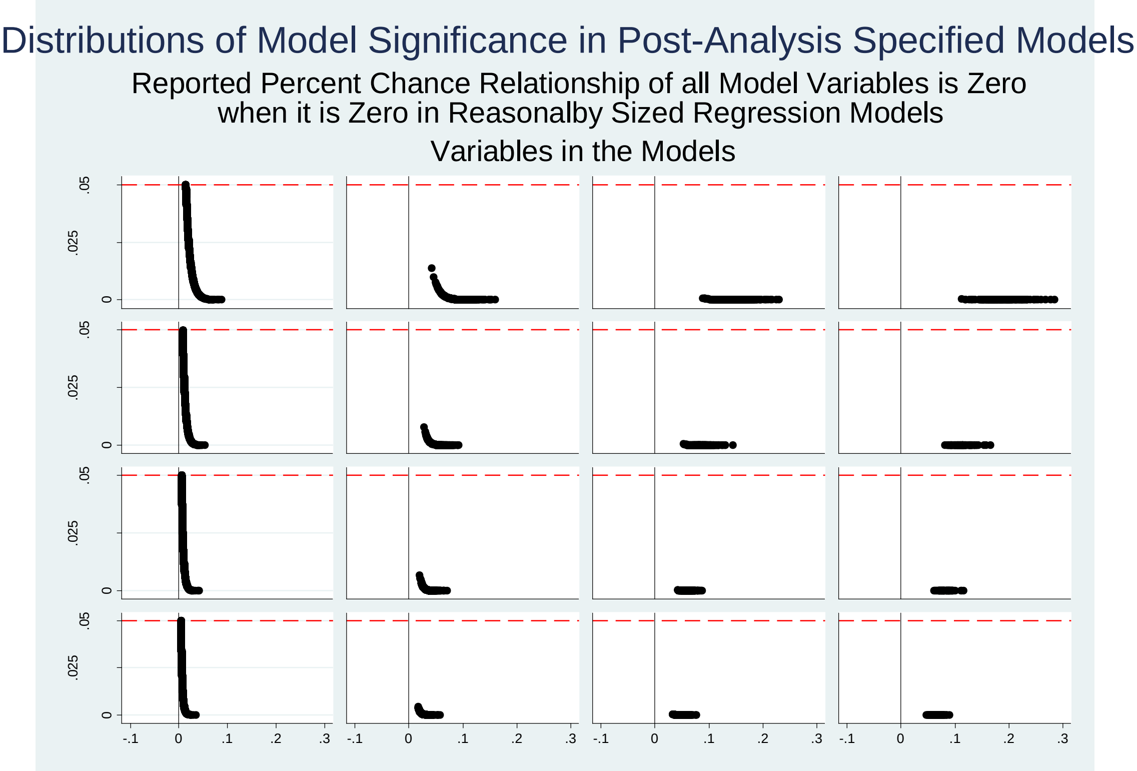
Post-Analysis Specification

Pre-Analysis Specification

Post-Analysis Specification

Significance Levels of Goodness-of-Fit statistics for regression models with 10 model variables and n=200.

Figure 13



Distributions of Model Significance in Post-Analysis Specified Models

Reported Percent Chance Relationship of all Model Variables is Zero
when it is Zero in Reasonalby Sized Regression Models

Variables in the Models