J Quant Criminol (2010) 26:217–236
DOI 10.1007/s10940-009-9077-7

ORIGINAL PAPER

Statistical Inference After Model Selection

Richard Berk · Lawrence Brown · Linda Zhao

Published online: 20 October 2009

Abstract Conventional statistical inference requires that a model of how the data were generated be known before the data are analyzed. Yet in criminology, and in the social sciences more broadly, a variety of model selection procedures are routinely undertaken followed by statistical tests and confidence intervals computed for a "final" model. In this paper, we examine such practices and show how they are typically misguided. The parameters being estimated are no longer well defined, and post-model-selection sampling distributions are mixtures with properties that are very different from what is conventionally assumed. Confidence intervals and statistical tests do not perform as they should. We examine in some detail the specific mechanisms responsible. We also offer some suggestions for better practice and show though a criminal justice example using real data how proper statistical inference in principle may be obtained.

Keywords Model selection · Statistical inference · Mixtures of distributions

Introduction

In textbook treatments of regression analysis, a model is a theory of how the data on hand were generated. Regressors are canonically treated as fixed, and the model specifies how the realized distribution of the response variable came to be, given the values of the regressors (Freedman 2005: 42). Causal interpretations can be introduced from information external to the model (Berk 2003: Chap. 5). Statistical tests and confidence intervals can be constructed.

This basic framework subsumes a variety of special cases. Popular instances are included under the generalized linear model and its extensions (McCullagh and Nelder

R. Berk (✉) · L. Brown · L. Zhao
Department of Statistics, University of Pennsylvania, Philadelphia, PA, USA
e-mail: berkr@sas.upenn.edu

R. Berk
Department of Criminology, University of Pennsylvania, Philadelphia, PA, USA

1989). Logistic regression is common example. Models with more than one regression equation (Greene 2003: Chaps. 14, 15) are for purposes of this paper also special cases.

The ubiquitous application of regression models in criminology, and in the social sciences more generally, has been criticized from a variety of perspectives for well over a generation (Box 1976; Leamer 1978; Rubin 1986; Freedman 1987; 2004; Manski 1990; Breiman 2001; Berk 2003; Morgan and Winship 2007). Despite the real merits of this literature, we assume in this paper that the idea of a "correct model" makes sense and consider statistical inference when the correct model is not known before the data are analyzed. We proceed, therefore, consistent with much common practice.

Statistical inference for regression assumes that there is a correct model that accurately characterizes the data generation process. This model is known, except for the values of certain parameters, before the data are examined (Freedman 2005: 64–65). The same holds for statistical inference more generally (Barnett 1983: Sect. 5.1).[1] Consequently, arriving at one or more regression models through data analysis would seem to make subsequent statistical inference problematic. Yet, model selection is a routine activity and is taught in any number of respected textbooks (Cook and Weisberg 1999; Greene 2003). This practice and pedagogy, therefore, would seem to warrant some scrutiny. Is there a problem? If so, is it important? And if so, what can be done about it? What are the consequences, for instance, of deleting "insignificant" predictors from a regression equation, re-estimating the model's parameter values, and applying statistical tests to the "final" model?

In the pages ahead, we show that when data used to arrive at one or more regression models are also used to estimate the values of model parameters and to conduct statistical tests or construct confidence intervals, the sampling distributions on which proper estimates, statistical tests and confidence intervals depend can be badly compromised. It follows that the parameters estimates, statistical tests and confidence intervals can be badly compromised as well. Moreover, because the compromised sampling distributions depend on complex interactions between a suite of possible models and the data to be analyzed, inferential errors are typically very difficult to identify and correct. It is far better, therefore, to avoid the problems to begin with. We suggest several ways by which this can be done.

In section "Framing the problem of post-model-selection statistical inference", the difficulties with "post-model-selection" statistical inference are introduced. Section "A more formal treatment" considers the particular mechanisms by which model selection can undermine statistical inference. To our knowledge, this discussion is novel. Section "Simulations of model-selection" illustrates through simulations the kinds of distortions that can result. Section "Potential solutions" discusses some potential remedies and shows with real data one example of appropriate practice. Section "Conclusions" draws some overall conclusions.

Framing the Problem of Post-Model-Selection Statistical Inference

When in a regression analysis the correct model is unknown before the data are introduced, researchers will often proceed in four steps.

[1] "Thus **sample data**, x, are assumed to arise from observing a random variable X defined on a **sample space**, $\Re$. The random variable X has a probability distribution $p_\theta(x)$ which is assumed known except for the value of the parameter Θ. The parameter Θ is some member of a specified **parameter space** Ω; x (and the random variable X) and Θ may have one or many components" (Barnett 1983: 121). Emphasis in the original. Some minor changes in notation have been made.

1. A set of models is constructed.
2. The data are examined, and a "final" model is selected.
3. The parameters of that model are estimated.
4. Statistical inference is applied to the parameter estimates.

Criminologists are certainly no exception. Davies and Dedel (2006), for instance, develop a violence risk screening instrument to be used in community corrections settings by reducing a logistic regression model with nine regressors to a logistic regression model with three regressors. Wald tests are applied to regression coefficients of the three-regressor model.

In many crime and justice analyses, some of the steps can be combined and are typically more complicated when there is more than one "final" model. For example, Ousey et al. (2008), consider whether a criminal victimization changes the likelihood of subsequent victimizations. Several competing models are developed. Some are discarded because of unsatisfactory statistical and interpretative properties. A variety of statistical tests are applied to each model, including the preferred ones. Lalond and Cho (2008), undertake a similar exercise for the impact of incarcerations on female inmates' employment prospects. There are both statistical and substantive considerations that lead the authors to favor some of their models over others, and there is a liberal use of statistical tests. Schroeder et al. (2007) consider the relative impact of drug and alcohol use on crime desistance by constructing a large number of competing models, some of which are deemed more instructive than others. Again, there is a liberal use of statistical tests, including for the models taken to be most instructive. Sampson and Raudenbush (2004) examine the causes of perceived neighborhood disorder through several different models, some of which are discarded for spurious associations. Conventional *t*-tests are used for all of the models including the subset of preferable models on which the substantive conclusions rest. In short, post-model-selection statistical inference is a routine activity in crime and justice research.

Each of the four steps can individually be legitimate. The problems addressed in this paper occur when all four steps are undertaken with the *same* data set. Perhaps the most apparent difficulty is that the model selection process in the first step is a form of data snooping. Standard errors conventionally estimated under such circumstances are well know to be incorrect; they are likely to too small (Freedman et al. 1988). False statistical power can result. In effect, there is an opportunity to look at all the face-down cards before a bet is placed.[2] For most thoughtful crime and justice researchers, this is old news.

The problems addressed here are more fundamental. It has long been recognized by some that when any parameter estimates are discarded, the sampling distribution of the remaining parameter estimates can be distorted. (Brown 1967; Olshen 1973). The rules by which some parameters are discarded do not even have to exploit information from the data being analyzed. The rules may be "ancillary" in the sense that they are constructed independently of the parameters responsible for the realized data (Brown 1990: 489).[3]

For example, suppose the model a researcher selects depends on the day of the week. On Mondays it's model A, on Tuesdays it's model B, and so on up to seven different models

[2] There are also the well-known difficulties that can follow from undertaking a large number of statistical tests. With every 20 statistical tests a researcher undertakes, for example, one null hypothesis will on the average be rejected at the .05 level even if all the null hypotheses are true. Remedies for the multiplicity problem are currently a lively research area in statistics (e.g., Benjamini and Yekutieli 2001; Efron 2007).

[3] The context in which Brown is working is rather different from the present context. But the same basic principles apply.

on seven different days. Each model, therefore, is the "final" model with a probability of 1/7th that has nothing to do with the values of the regression parameters. Then, if the data analysis happens to be done on a Thursday, say, it is the results from model D that are reported. All of the other model results that could have been reported are not. Those parameter estimates are summarily discarded.

Distorted sampling distributions of the sort discussed in the pages ahead can materialize under ancillary selection rules. It is the exclusion of certain estimates by itself that does the damage. In practice, however, the selection rule will not likely be ancillary; it will be dependent on regression parameters. For example, the selection rule may favor models with smaller residual variances.

Recent work has rediscovered and extended these insights while focussing particularly on the regression case. Leeb and Pötscher (2006, 2008) show that the sampling distributions of post-model-selection parameter estimates are likely to be unknown, and probably unknowable, even asymptotically. Moreover, it does not seem to matter what kind of model selection approach is used. Informal data exploration and tinkering produces the same kinds of difficulties as automated procedures in which the researcher does not participate except at the very beginning and the very end. Judgment-based model selection is in this sense no different from nested testing, stepwise regression, all subsets regression, shrinkage estimators (Efron et al. 2007), the Dantzig Selector (Candes and Tao 2007), and all other model selection methods considered to date (Leeb and Pötscher 2005). Likewise, the particular screening statistic applied is immaterial: an adjusted R^2, AIC, BIC, Mallows' Cp, cross-validiation, *p*-values or others (Leeb and Pötscher 2005).

A More Formal Treatment

The conceptual apparatus in which frequentist statistical inference is placed imagines a limitless number of independent probability samples drawn from a well-defined population. From each sample, one or more sample statistics are computed. The distribution of each sample statistic, over samples, is the sampling distribution for that sample statistic. Statistical inference is undertaken from these sampling distributions. To take a common textbook illustration, the sample statistic is a mean. A sampling distribution for the mean is the basis for any confidence intervals or a statistical tests that follow. The same framework applies to the parameters of a regression equation when the structure of the regression equation is known before the data are analyzed.[4]

Definitional Problems

Past discussions of post-model-selection statistical inference have apparently not recognized that when the regression model is not fully specified before the data are examined, important definitional ambiguities are introduced. Any definition of regression parameters depends on an assumed model. With different models come different definitions of what

[4] There is an alternative formulation that mathematically amounts to the same thing. In the real world, nature generates the data through a stochastic process characterized by a regression model. Inferences are made to the parameters of this model, not to the parameters of a regression in a well-defined population. Data on hand are not a random sample from that population, but are a random realization of a stochastic process, and there can be a limitless number of independent realizations. These two formulations and others are discussed in far more detail elsewhere. The material to follow is effectively the same under either account, but the random sampling approach is less abstract and easier build upon.

one is trying to estimate. In the absence of a model, the estimation enterprise is unclear. This follows even if attention is centered in the role of a given regressor.

Suppose, for example, that for a response variable Y, there are two potential regressors, X and Z. There is interest in the relationship between Y and X holding Z constant. A linear regression model is imposed with the corresponding population regression coefficient of $\beta_{yx\cdot z}$. Then,

$$\beta_{yx\cdot z} = \frac{\rho_{yx} - \rho_{xz}\rho_{yz}}{\left(1 - \rho_{xz}^2\right)} \times \frac{\sigma_y}{\sigma_x}, \tag{1}$$

where in the population ρ is a correlation coefficient, σ is a standard deviation, and the subscripts denote the variables involved. Unless the two regressors X and Z happen to be uncorrelated (i.e., $\rho_{xz} = 0$), an extremely unlikely occurrence in most observational studies, the value of the population parameter $\beta_{yx\cdot z}$ will depend on whether Z is included in the regression model. If Z is excluded, all of the correlations involving Z are equivalent to zero, and one is left with $\beta_{yx} = \rho_{yx}(\sigma_y/\sigma_x)$. β_{yx} is not the same as $\beta_{yx\cdot z}$, so the definition of regression parameter for X depends on the model in which X is placed. Similar issues can arise with any modeling enterprise, not just linear regression.

Thus, when a single model is not specified before the analysis begins, it is not clear what population parameter is the subject of study. And without this clarity, the reasoning behind statistical inference becomes obscure. For example, unbiased estimates are desirable, but unbiased estimates of what? In practice, there will typically be a suite of possible models and a large number of population regression coefficients, even for a single regressor. In section "Simulations of model-selection", simulations are presented to illustrate and help fix these ideas.

The response may be that a model selection procedure will be applied to find an appropriate model. Once that model is determined, it will be apparent what features of the population are being estimated. This argument can have merit if for a given population, model selection is undertaken on one random sample, while parameter estimation and statistical inference is undertaken on another random sample. When the entire process is undertaken on a single random sample, the argument stumbles badly, as we will now see.

Estimation Problems

Estimates of the regression parameters always depend on the composition of a realized random sample, but now also on the regression model selected. There is, therefore, a new source of uncertainty. How regressors perform can vary with the model in which they are placed, and the regression model selected can vary from sample to sample.

Figure 1 illustrates the implications of model selection for sampling distributions. Estimates are shown with the conventional hats, and the different models are denoted by subscripts. For example, $\hat{M}_1$, denotes Model 1 selected in a given sample, and $\hat{\Omega}_1$ represents that model's estimated parameters. There are nine random samples standing in for the usual limitless number.

One begins with a well-defined population. In that population, there is a process that generates values of the response variable conditional on a set of regressors. This process can be represented by a particular model that some may characterize as the "correct" model. But the form the model takes is unknown, and in the population there are many candidate models that in principle could be called correct. For example, the model consistent with how the data were generated might contain eight regressors, and a competitor might contain five regressors, some of which are not in the larger model at all.

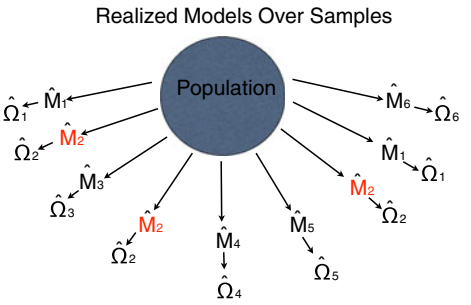

Fig. 1 The model selection thought experiment

A random sample is drawn. For that sample, the preferred model selection procedure is applied, and a winning model determined. The model chosen is sample dependent, and in that sense is an estimate. It is an informed guess of the correct model, given the random sample and the model selection procedure. The parameters of the selected model are estimated.

A critical assumption is that all of the regressors in the correct model are present in the sample. There can be, and usually are, other regressors as well. Some model selection procedures seek the correct model only. Other model selection procedures seek a model that includes the correct regressors and perhaps includes some regressors that actually do not belong in the model. For purposes of this paper, the same difficulties arise under either approach.

One imagines repeating the entire process—drawing a random sample, undertaking model selection, parameter estimation, and statistical inference—a limitless number of times. In this cartoon version with nine samples, Model 2 (M_2) is the unknown correct model, and from sample to sample other models are chosen as well. For these nine samples, the correct model happens to be chosen most frequently among the candidate models, but not the majority of the time. The same basic reasoning applies when from a suite of possible models several winners are chosen, although the exposition becomes rather more clumsy. There are now several final models, each with its own regression estimates, associated with each random sample.

The fundamental point is this: model selection intervenes between the realized sample and estimates of the regression parameters. Therefore, a sampling distribution consistent with how the regression estimates were generated must take the model selection step into account. Moreover, because there is only one correct model, the sampling distribution of the estimated regression parameters can include estimates made from incorrect models as well as the correct one. The result is a sampling distribution that is a mixture of distributions.

Drawing on the exposition of Leeb and Pötscher (2005: 24–26), and for our purposes with no important loss in generality, consider a regression analysis in which there are two candidate models, one of which some would call the correct model. The researcher does not know which model is the correct one. Call the two models M_1 and M_2. To take a very simple example, M_1 could be

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i, \tag{2}$$

and M_2 could be

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i. \tag{3}$$

These are just conventional regression equations consistent with the discussion surrounding Eq. 1. They imply different conditional distributions for y_i. M_1 differs from M_2 by

whether z_i is in the model, or equivalently, whether $\beta_2 \neq 0$.[5] A model selection procedure is employed to make that determination.

Suppose interest centers on a least squares estimate of β_1, the regression coefficient associated with the regressor x_i present in both models.[6] There is also interest in a t-test undertaken for the null hypothesis that $\beta_1 = 0$. There are two constants: C_1 associated with M_1 and C_2 associated with M_2. Likewise, there are two estimates of the standard error of $\hat{\beta}_1$. $\widehat{\text{SE}}_1$ denotes the estimated standard error under M_1, and $\widehat{\text{SE}}_2$ denotes the estimated standard error under M_2. When $\hat{\beta}_1/\widehat{\text{SE}}_1 \geq C_1$, the p-value for the test is equal to or less than .05, and the null hypothesis is rejected. C_2 serves the same function for M_2.[7] Then,

$$\begin{aligned} &P\left[\left(\hat{M}_1 \text{ and } \frac{\hat{\beta}_1}{\widehat{\text{SE}}_1} \geq C_1\right) \text{ or } \left(\hat{M}_2 \text{ and } \frac{\hat{\beta}_1}{\widehat{\text{SE}}_2} \geq C_2\right)\right] \\ &= P\left(\frac{\hat{\beta}_1}{\widehat{\text{SE}}_1} \geq C_1 | \hat{M}_1\right) P(\hat{M}_1) + P\left(\frac{\hat{\beta}_1}{\widehat{\text{SE}}_2} \geq C_2 | \hat{M}_2\right) P(\hat{M}_2) \end{aligned} \tag{4}$$

where $\hat{M}_1$ denotes that model 1 is selected, and $\hat{M}_2$ denotes that model 2 is selected. Equation 4 means that the probability that the null hypothesis for β_1 is rejected is a linear combination of two weighted conditional probabilities: (1) the probability of rejecting the null hypothesis given that model 1 is selected multiplied by the probability that model 1 is selected and (2) the probability of rejecting the null hypothesis given that model 2 is selected, multiplied by the probability that model 2 is selected. Thus, the sampling distribution for β_1 is a mixture to two distributions, and such mixtures can depart dramatically from the distributions that conventional statistical inference assumes.

To help fix these ideas, we turn to a demonstration of how a simple mixture of normal distributions can behave in a non-normal manner. Normality is the focus because it can play so central a role in statistical inference. The demonstration sets the stage for when we later consider more realistic and complex situations like those represented in Eqs. 2 and 3. We will see then that the results from our demonstration have important parallels in a linear regression setting but that simulations of actual model selection procedures produce a number of additional and unexpected results.

Figure 2 draws on Eq. 4. As before, there is a correct but unknown model. To simplify matters, and with no important loss of generality for this demonstration, the mean squared errors for both regression models are assumed to be approximately 1.0. Then, the sampling distribution for β_1, conditional on M_1 being selected, is taken to be normal with a mean of

[5] Setting some regression coefficients to 0 is perhaps the most common kind of restriction imposed on regression coefficients, but others can also lead to interesting models (e.g., $\beta_1 = \beta_2$, which would mean that $y_i = \beta_0 + \beta_1[x_i + z_i] + \varepsilon_i$). Also, in the interest of simplicity, we are being a little sloppy with notation. We should be using different symbols for the regression coefficients in the two equations because they are in different models and are, therefore, defined differently. But the added notional complexity is probably not worth it.

[6] The problems that follow can materialize regardless of the estimation procedure applied: least squares, maximum likelihood, generalized method of moments, and so on. Likewise, the problems can result from any of the usual model selection procedures.

[7] There are two constants because the constants are model dependent. For example, they can depend on the number of regressors in the model and which ones they are. There are two values for the mean squared error, because the fit of the two models will likely differ. To make this more concrete, $\hat{\beta}_1/\widehat{\text{SE}}_1 \geq C_1$ may be nothing more than a conventional t-test. Written this way, the random variation is isolated on the left hand side and the fixed variation is isolated in the right hand side. It is how the random variables behave that is the focus of this paper.

12.0 and a standard deviation of 2.0. The sampling distribution for β_1, conditional on M_2 being selected, is taken to be normal with a mean of 4.0 and a standard deviation of 1. To minimize the complications, an ancilliary selection procedure is applied such that $P(\hat{M}_1) = .2$ and $P(\hat{M}_2) = .8$; the model selection procedure chooses M_2 four times more often than M_1. Figure 2 is constructed by making 10,000 draws from the first normal distribution and 40,000 draws from the second normal distribution. The line overlaid is the true combined distribution, which the simulation is meant to approximate. The combined distribution has a mean of approximately 5.6 and a standard deviation of approximately 3.4.

The two sampling distributions, conditional on M_1 or M_2, are by design normal, but the simulated sampling distribution after model selection is decidedly non-normal. The distribution is bimodal and skewed to the right. Moreover, if M_1 is the correct model, $\hat{\beta}_1$ from the combined distribution is biased downward from 12.0 to 5.6. If M_2 is the correct model, $\hat{\beta}_1$ is biased upward from 4.0 to 5.6. In either case, the estimate will be systematically in error. The standard deviation of the combined distribution is substantially larger than the standard deviations of its constituent distributions: 3.4 compared to 2.0 or 1.0. Again, the error is systematic.

Biased estimates of both regression coefficients and their standard errors are to be anticipated in post-model-selection mixture distributions. Moreover, the biases are unaffected by sample size; larger samples cannot be expected to reduce the biases. Before one even gets to confidence intervals and statistical tests, the estimation process can be badly compromised.[8]

In summary, model selection is a procedure by which some models are chosen over others. But model selection is subject to uncertainty. Because regression parameter estimates depend on the model in which they are embedded, there is in post-model-selection estimates additional uncertainty not present when a model is specified in advance. The uncertainty translates into sampling distributions that are a mixture of distributions, whose properties can differ dramatically from those required for convention statistical inference.

Underlying Mechanisms

Although mixed sampling distributions are the core problem, the particular forms a mixed distribution can take determine its practical consequences. It is necessary, therefore, to look under the hood. We need to consider the factors that shape the mixed sampling distributions that can result from model selection. The specific mechanisms underlying post-model-selection statistical inference has apparently not been addressed in past work.

First, model selection implies that some regression coefficients values are censored. If sufficiently close to zero, they are treated as equal to zero, and their associated regressors are dropped from the model. Consequently, a post-selection sampling distribution can be truncated or contain gaps in regions that are near 0.0. These effects are a *direct* result of the model selection. For example, Fig. 2, might have censored at values $\hat{\beta}_1$ less than 2.0. There would have been no simulated values below that threshold.[9]

Second, there can be *indirect* effects. When for a given model a regressor is excluded, the performance of other regressors can be affected. For ease of exposition, as before suppose there are only two regressors, X and Z, as candidate explanatory variables. Then, the sample regression coefficient for X holding Z constant can be written as,

[8] Simulations much like those in section "Simulations of model-selection" can be used to produce virtually identical results starting with appropriate raw data and Eqs. 2 and 3.

[9] The selection mechanism would then not have been ancillary.

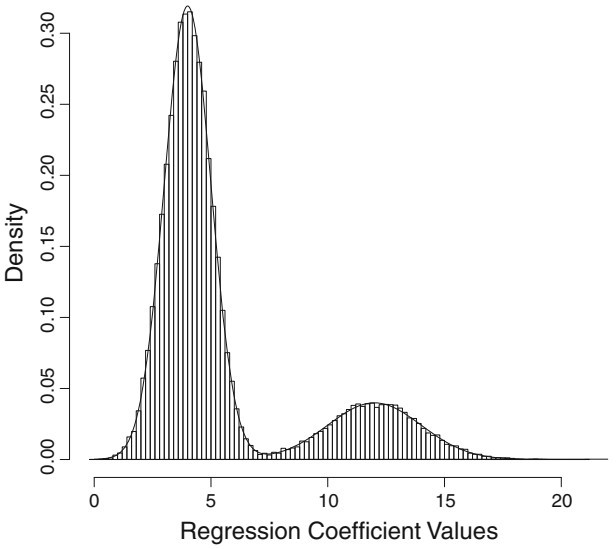

Fig. 2 An illustrative combined sampling distribution

$$\hat{\beta}_{yx.z} = \frac{r_{yx} - r_{xz}r_{yz}}{\left(1 - r_{xz}^2\right)} \times \frac{s_y}{s_x}, \tag{5}$$

where r is a sample correlation coefficient, s is a sample standard deviation, and for both, the subscripts denote the regressors involved. Equation 5 is the sample analogue of Eq. 1. If Z is dropped from the equation, all of the correlations involving Z are equivalent to 0.0. One is then left with the bivariate correlation between Y and X and the ratio of their two standard deviations. Insofar as r_{yz} and r_{xz} are large in absolute value, the value of $\hat{\beta}_{yx\cdot z}$ can change dramatically when Z is dropped from the model. Conversely, if the regressors are orthogonal (here, $r_{xz} = 0$), regression coefficient estimates will not be model dependent, and their sampling distributions will be unaffected by these indirect processes. These results generalize to models with more than two candidate regressors.[10]

Third, sampling distributions can vary in their dispersion alone, although this will typically be far less important than the direct or indirect effects of model selection. One can see how the dispersion can be affected through an equation for the standard error of a conventional regression coefficient estimate, not subject to model selection, say, $\hat{\beta}_{yx\cdot z}$. For the two regressors X and Z,

$$SE\left(\hat{\beta}_{yx\cdot z}\right) = \frac{\hat{\sigma}_\varepsilon}{s_x\sqrt{n-1}}\sqrt{\frac{1}{1-r_{xz}^2}} \tag{6}$$

where $\hat{\sigma}_\varepsilon$ is an estimate of the residual standard deviation, s_x is the sample standard deviation of x, and r_{xz}^2 is the square of the sample correlation between x and z.[11] From this equation, one learns that the sampling distribution will be more dispersed if there is more residual variance, less variance in regressor x, more linear dependence between regressors,

[10] The correlations in the numerator are replaced by partial correlations controlling for all other predictors, and the correlation in the denominator is replaced by the multiple correlation of the predictor in question with all other predictors.

[11] Because in regression x and z are usually treated as fixed, s_x and r_{xz}^2 are not random variables and not treated as estimates.

and a smaller sample size. All except the sample size can be affected by the particular model selected. In addition, $\hat{\sigma}_\varepsilon$ is affected directly by random sampling error.[12]

The three underlying selection mechanisms—direct censoring, indirect censoring, alterations in the dispersions of regression parameter estimates—can interact in complex ways that depend on the models considered and the data being analyzed. There are, therefore, no general expressions through which the impact of model selection may be summarized. However, simulations can provide some insight into what can happen. Because in simulations one knows the model responsible for the data, there is a benchmark to which the post-model-selection results can be compared. When working with real data, there is no known standard of absolute truth to which the empirical results can be held.[13]

Simulations of Model-Selection

We now turn to simulations of model selection effects. The form the selection takes is not a primary concern because in broad brush strokes at least, the implications are the same regardless of how the selection is accomplished. For an initial simulation, selection is implemented through forward stepwise regression using the AIC as a fit criterion. At each step, the term is added that leads to the model with the smallest AIC. The procedure stops when no remaining regressor improves the AIC.

For this simulation, the full regression model takes the form of

$$y_i = \beta_0 + \beta_1 w_i + \beta_2 x_i + \beta_3 z_i + \varepsilon_i, \tag{7}$$

where $\beta_0 = 3.0$, $\beta_1 = 0.0$, $\beta_2 = 1.0$, and $\beta_3 = 2.0$. Because the parameter $\beta_1 = 0$ in Eq. 7, many would refer to a submodel that excluded W as the "correct" model. But Eq. 7 is also correct as long at $\beta_1 = 0$ is allowed. Therefore, we will use the adjective "preferred" for the model with w excluded. The full model and the preferred model will generate the same conditional expectations for the response. The smaller model is preferred because it is simpler and uses up one less degree of freedom.

All three predictors are drawn at random from a multivariate normal distribution, although with fixed regressors, this is of no importance. The variances and covariance are set as follows: $\sigma_\varepsilon^2 = 10.0$, $\sigma_w^2 = 5.0$, $\sigma_x^2 = 6.0$, $\sigma_z^2 = 7.0$, $\sigma_{w,x} = 4.0$, $\sigma_{w,z} = 5.0$, and $\sigma_{x,z} = 5.0$. The sample size is 200. The intent is to construct a data set broadly like data sets used in practice. The data were not constructed to illustrate a best case or worst case scenario.

10,000 samples were drawn, stepwise regression applied to each, and sampling distributions were constructed from the final models selected. Rather than plotting estimated regression coefficients, conventional t-values are plotted. The usual null hypothesis was assumed; each regression coefficient has a population value 0.0. A distribution of t-values is more informative than a distribution of regression coefficients because it takes the regression coefficients and their standard errors into account. The R^2s varied over the simulations between about .3 and .4.

[12] When there are more than two regressors, the only change in Eq. 6 is that r_{xz}^2 is replaced by the square of the multiple correlation coefficient between the given regressor and all other regressors.

[13] If there were, there would be no need to do the research.

Simulation Results

With three regressors, there are eight possible models, including a model in which none of the regressors is chosen. The distribution shown in Table 1 indicates that the preferred model (i.e., with regressors *X* and *Z*) is selected about 66% of the time. The next most common model, chosen about 17% of the time, includes only regressor *Z*. The full model with regressors *W*, *X*, and *Z* is chosen 11% of the time. In short, for this simulation a practitioner has a little less than a two-thirds chance of selecting the preferred model.

Figure 3 shows two simulated *t*-value distributions for regressor *X*. The solid black line represents the distribution that would result if the preferred model were known and its parameter values estimated; there is no model selection. The broken line represents the post-model-selection sampling distribution. Both lines are the product of a kernel density smoother applied to the histograms from the simulation. The distribution assuming that the preferred model was known was constructed from all the 10,000 samples. For the post-model-selection *t*-values were available only if *x* was in a final model. That happened 7355 times out of 10,000 over four of the eight models (i.e., *X* alone, *X* with either *W* or *Z*, and *X* with *W* and *Z*). So, the broken line in Fig. 3 is the result conditional on *X* being in the model.

It is apparent that the two distributions in Fig. 3 are quite different. The post-model-selection distribution has a greater mean (2.6–2.2), a smaller standard deviation (.79–1.0) and is skewed to the right. Figure 4 is based on the same procedures as Fig. 3, but now *z* is the regressor of interest. The results are even more dramatic. The post-model-selection distribution is bimodal and strongly skewed to the right. Both the mean and the standard deviation are biased substantially upward: from 4.9 to 5.5 for the mean and from 1.0 to 2.3 for the standard deviation. Statistical inference assuming the solid black line would be very misleading insofar as the broken line captured what was really going on. This illustrates a fundamental point made in the Leeb and Pötscher papers cited.

Figures 3 and 4 demonstrate well the consequences for statistical inference that can follow from model selection. They show post-model-selection sampling distributions over the full set models in which the regressors are included. However, in practice researchers commonly settle on a final model, or a small number of final model candidates. Therefore, it is instructive to consider the distributions of *t*-values conditional on a given model; a model is chosen and tests are applied to that model only.

It is especially telling to condition on the preferred model even though in practice the preferred model would not be known a priori. One might hope that at least when the preferred model is selected, conventional statistical inference would be on sound footing. However, selecting the preferred model only guarantees that the proper regressors are included. It does not guarantee any of the desirable properties of the regression coefficient estimates.

Figure 5 shows that statistical inference remains problematic *even when the statistical inference is conditional on arriving at the preferred model*. In fact, Figs. 3 and 5 are very similar because for this simulation, *x* is usually selected as part of the chosen model. In effect, therefore, conditioning on the preferred model is already taking place much of the

Table 1 Distribution of models selected in 10,000 draws

None	*W*	*X*	*Z*	WX	WZ	XZ	WXZ
0%	0%	.0001%	17.4%	1.0%	4.9%	65.7%	10.8%

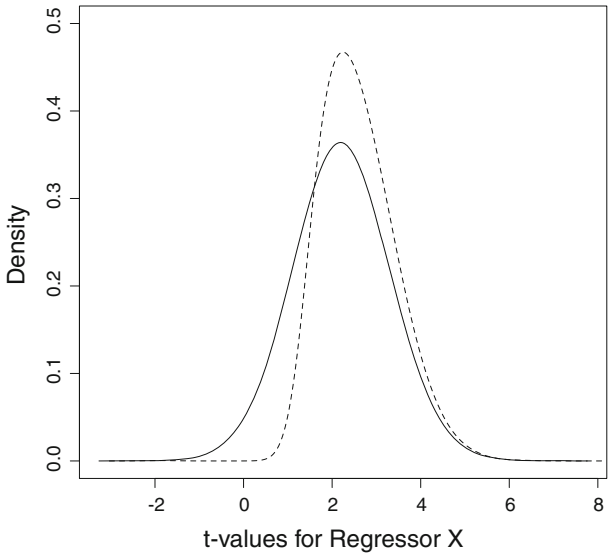

Fig. 3 Stepwise regression sampling distributions of the regression coefficient *t*-values for regressor *X* (The *solid line* is conditional on the preferred model being known. The *broken line* is conditional on *X* being included in a model)

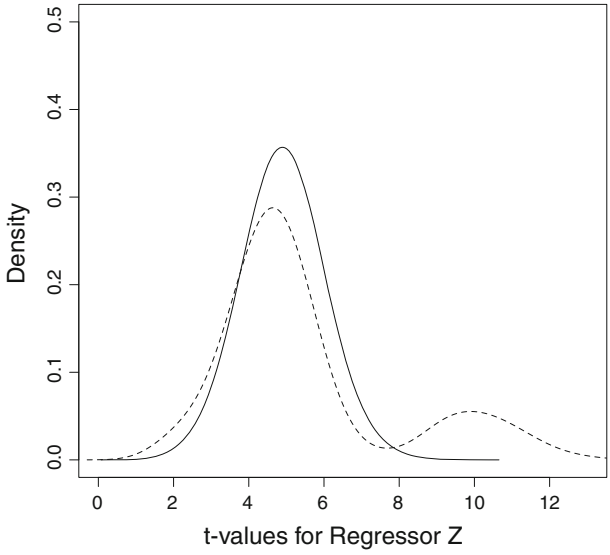

Fig. 4 Stepwise regression sampling distributions of the regression coefficient *t*-values for regressor *Z* (The *solid line* is conditional on the preferred model being known. The *broken line* is conditional on *Z* being included in a model)

time. However, this is not a general result and is peripheral to our discussion in any case. The point is that the biases noted for Fig. 3 remain. Thus, for the preferred model the null hypothesis that $\beta_2 = 0$ should be rejected at the .05 level with a probability of approximately .60. But after model selection, that probability is about .76. This represents an increase of about 27% driven substantially by bias in the estimated regression coefficient. It is not a legitimate increase in power. False power seems to be a common occurrence for post-model-selection sampling distributions.

Figure 6 is constructed from the same simulation as Figure 5 except that the true value for β_2 is now .5 not 1.0. The post-selection *t*-distribution, even when conditioning on the preferred model, is now strongly bimodal and nothing like the assumed normal

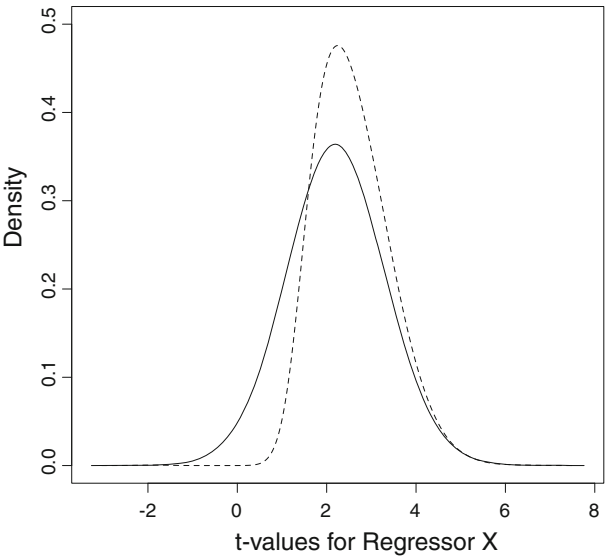

Fig. 5 Stepwise regression sampling distributions of the regression coefficient *t*-values for regressor *X*. (The *solid line* is conditional on the preferred model being known. The *broken line* is conditional on the preferred model being selected)

distribution. The general point is that the post-model-selection sampling distribution can take on a wide variety of shapes, none much like the normal, even when the researcher happens to have selected the model accurately representing how the data were generated.

Figure 7 shows two distributions for *Z*, one based on the preferred model being known (i.e., the solid line) and one conditioning on the selected preferred model (i.e., the broken line). Figure 7 is very different from Fig. 4. In this instance, conditioning on the selected preferred model brings the proper and post-model-selection distributions of *t*-values into a rough correspondence, and for both, the probability of rejecting at the .05 level the null hypothesis that $\beta_3 = 0$ is about .99. This underscores that post-model-selection sampling distributions are not automatically misleading. Sometimes the correct sampling distribution and the post-model-selection sampling distribution will be very similar.

Figures 8 and 9 are a recapitulation but when selection is by all subsets regression using Mallow's *Cp* as the selection criterion. Much like the earlier stepwise regression, the correct model is chosen 64% of the time. Compared to Figs. 5 and 7, very little changes. As noted earlier, for purposes of this paper the particular selection procedure used does not materially matter. The kinds of distortions introduced can vary with the selection procedure, but the overall message is unchanged. In this case, the distortions were about the same whether selection was by stepwise regression or all subsets regression.

Potential Solutions

Post-model-selection sampling distributions can be highly non-normal, very complex, and with unknown finite sample properties even when the model responsible for the data happens to be selected. There can be substantial bias in the regression estimates, and conventional tests and confidence intervals are undertaken at some peril. At this point,

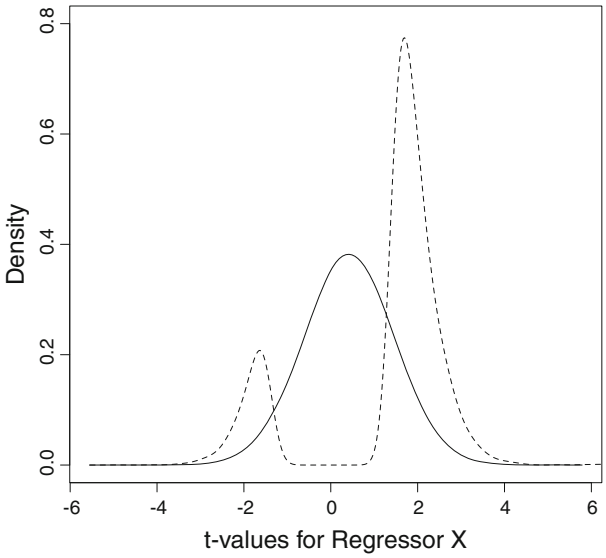

Fig. 6 Stepwise regression sampling distributions of the regression coefficient *t*-values for regressor *X*, $\beta_2 = .5$. (The *solid line* is conditional on the preferred model being known. The *broken line* is conditional on the preferred model being selected)

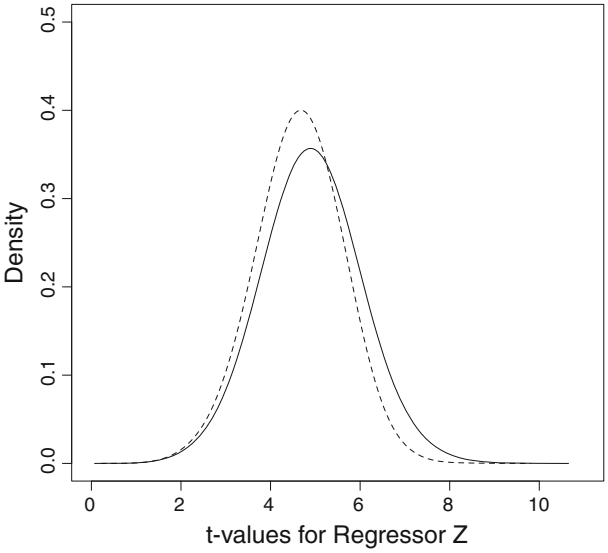

Fig. 7 Stepwise Regression sampling distributions of the regression coefficient *t*-values for regressor *Z*. (The *solid line* is conditional on the preferred model being known. The *broken line* is conditional on the preferred model being selected)

there seems to be no way to anticipate the nature of the problems or their magnitude except in a few very special cases. The three mechanisms described in section "Underlying mechanisms" by which the difficulties are introduced interact in complicated ways that are highly data and model dependent.

As already noted, however, there can be situations in which the consequences of model selection are not necessarily problematic. When a sample is very large relative to the number of regression parameters being estimated, and there are regression coefficients with true values sufficiently different from zero, many procedures will select the very same model over and over. In effect, one model has a probability of selection near 1.0, and all other models have probabilities of selection near 0.0. The sampling distributions are not

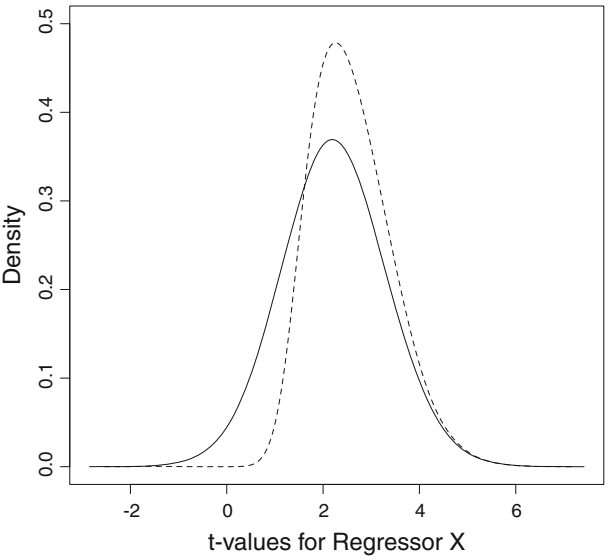

Fig. 8 All subsets regression sampling distributions of the regression coefficient *t*-values for regressor *X*. (The *solid line* is conditional on the preferred model being known. The *broken line* is conditional on the preferred model being selected)

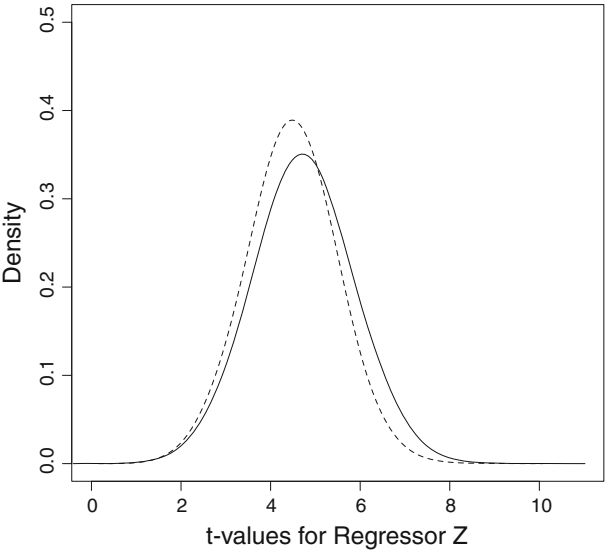

Fig. 9 All subsets regression sampling distributions of the regression coefficient *t*-values for regressor *Z*. (The *solid line* is conditional on the preferred model being known. The *broken line* is conditional on the preferred model being selected)

combinations of sampling distributions. In practice, one could not know for certain whether such a situation exists, but power analyses could provide information from which such a case might be made.

If the post-model-selection sampling distributions may be problematic, probably the most effective solution is to have two random samples from the population of interest: a training sample and a test sample. The training sample is used to arrive at a preferred model. The test sample is used to estimate the parameters of the chosen model and to apply statistical inference. For the test sample, the model is known in advance. The requisite structure for proper statistical inference is in place, and problems resulting from post-

model-selection statistical inference are prevented. The dual-sample approach is easy to implement once there are two samples.

When there is one sample, an option is to randomly partition that sample into two subsets. We call this the split-sample approach. One can then proceed as if there were two samples to begin with. Whether this will work in practice depends on the size each sample needs to be. Sample size determinations can be addressed by appropriate power analyses for each partition.

For example, suppose a researcher is prepared to assume, as required, that all of the necessary regressors in their appropriate functional forms are included in the data.[14] The researcher also assumes that the regression coefficients associated with at least some of the regressors are actually 0.0. This is known as the assumption of "sparsity." Then, the researcher assumes that in the data the regressors belonging in the preferred model will have large regression coefficients relative to their standard errors and that regressors that do not belong in the preferred model will have small regression coefficients relative to their standard errors. It follows that a relatively small sample will be able to find reliably the preferred model. The remaining data can serve as the test sample.

If the number of observations in the available data is too small to implement a split-sample approach, one can fall back on a traditional textbook strategy. Before the data are examined, a best guess is made about what the appropriate model should be. Parameter values can be estimated and statistical inference applied just as the textbooks describe. If this confirmatory step works out well, one can report the results with no concerns about post-model-selection inference.[15] One can then follow up with an exploratory data analysis. A range of other models can be constructed and evaluated as long as any statistical inference for selected models is not taken seriously. The exploratory results may well be very helpful for future research with new data. In short, there can be a confirmatory data analysis followed by an exploratory data analysis, each undertaken by its own inferential rules.

In the longer term, there are prospects for developing useful post-model-selection inference. A key will be to adjust properly for the additional uncertainty resulting from model selection. We are working on that problem and have some initial promising results. Unfortunately, the implications may be disappointing. It will probably turn out that when model uncertainty is properly taken into account, confidence intervals will be far larger and statistical tests will have substantially reduced power.

A Split-Sample Example

To help make the discussion of potential remedies more concrete, we turn briefly to an empirical illustration using data on sentencing. Determinants of post-conviction sentences have long been of interest to criminologists and to researchers from other disciplines who study sanctions (Blumstein et al. 1983; Wooldredge 2005; Johnson 2006). Probation decisions have received considerable attention (Morris and Tonry 1980; Petersilia 1997). When a decision is made to place an individual on probation, one might be interested in the factors that could affect the length of the suspended incarceration sentence. Suspended sentence length can be important. It can be an ongoing threat with which law-abiding

[14] This would include all necessary interaction effects.

[15] All of the usual caveats would still apply. For example, if the model specified does not properly represent how the data were generated, the regression estimates will be biased, and statistical tests will be not have their assumed properties.

behavior is shaped. It can also help to determine the length of the probation period and the probation conditions imposed. Therefore, the factors that might help to explain the length of suspended sentences are important too.

We use real data and a split-sample approach. The data are a random sample of 500 individuals sentenced to probation in a large American city. The length of the suspended sentence in months is the outcome of interest. For these data, mean sentence length is a little less than 28 months. The distribution is skewed to the right so that the median is only 18 months. 75% of the probationers have a suspended sentence of about 38 months or less. Because of the long right tail, it can make good sense to work with the log of sentence length as the response variable. Using the log of sentence length is also consistent with a theory that judges think in proportional terms when they determine sentence length. For instance, a sentence could be made 25% longer if an offender has a prior felony conviction.

The intent, therefore, is to consider how various features of the convicted individual and the crimes for which the individual was convicted may be related to log of sentence length.[16] The follow regressors were available.

1. Assault as the conviction offense
2. Drug possession as the conviction offense
3. Burglary as the conviction offense
4. Gun-related crime as the conviction offense
5. Number of juvenile arrests
6. Number of prior arrests
7. Age at first contact with the adult courts
8. Age at conviction
9. Race black
10. Not married
11. High school degree
12. Referred for drug treatment

There were other conviction crimes in the data. But given the nature of this population, these crimes were never reported, or were reported so rarely that they could not be used in the analysis. For example, if individuals were convicted of armed robbery, murder or rape, they were not placed on probation. A substantial majority of the conviction offenses were for drug-related offenses, burglaries and assaults.

The regression model was not known before the data were analyzed. Consequently, the sample of 500 was partitioned at random into 250 training observations and 250 test observations. All subsets regression was applied to the training data with the BIC as the screening statistic.[17] The parameters of the model selected were then estimated separately for the training and the test data. Statistical tests were undertaken in both cases. The results from the test data do not suffer from the problems we have been considering because the model was determined with another data set. But do the results for the test data differ materially from the results for the training data? Was there a problem to be fixed?

[16] The logarithm of zero is undefined. So, a value of .5 (i.e., about two weeks) was used instead. Other reasonable strategies led to results that for purposes of this paper were effectively the same. A suspended sentence of zero months can occur, for instance, if a sentencing judge gives sufficient credit for time served awaiting trial.

[17] The model selection was done with the procedure *regsubsets* in *R*.

Table 2 Results from training data

	Estimate	Multiplier	Standard error	*p*-value
Intercept	1.686	–	0.19	0.0000
Assault conviction	1.089	2.97	0.29	0.0002
Drug conviction	0.729	2.07	0.19	0.0001
Gun conviction	1.147	3.15	0.49	0.0196
Number of priors	0.024	1.02	0.005	0.00001

There were three models that had very similar BIC values and very similar structures. Table 2 shows the results for the model selected. Output from the other two models was much the same. The overall conclusions were as well.

Four regressors were included: whether the conviction was for an assault, whether the conviction was for a drug offense, whether the conviction was for a gun-related offense, and the number of prior arrests. For each estimated regression coefficient, we separately tested the null hypothesis that the population regression coefficient was zero. A two-tailed tests was applied using .05 as the critical value.

For the training data used to do the model selection, the null hypothesis was easily rejected for each regressor, and the associations were all strong. Consistent with a theory that judges determine sentences proportionally, all of the regression coefficients were exponentiated so that they became multiplicative constants. The baseline conviction offense is essentially burglary.[18] Then in Table 2, the average burglary sentence is multiplied by 2.97 if the conviction offense is assault, by 2.07 if the conviction offense is for drugs, and by 3.15 if for a gun related offense. For each additional prior arrest, the sentence length is multiplied by 1.02. Thus, an offender with 10 prior arrests would have a sentence that is about 1.22 times longer than an offender with no prior arrests. In short, what matters is the conviction offense and prior record, with the multipliers that are substantial in practical terms. Few criminologists would find this surprising.

Table 3 shows the result when the same model is used with the test data. The results are rather different. Conviction for an assault or for a gun-related offense are no longer statistically significant. Very small *p*-values in Table 2 are greater the .05 in Table 3. The standard errors for the two regression coefficients are essentially unchanged, but the sizes of the regression coefficients are substantially reduced. The other two regressors also show smaller associations with the response in Table 3, but still have *p*-values less than .05. Thus, given the nature of the crimes for which one can receive probation, what matters for the length of the suspended sentence is only whether the conviction is for a drug offense and the offender's prior record.

Because of the large number of possible models and the three mechanisms by which model selection effects are produced, it is effectively impossible to know exactly why the two tables differ. Matters are further complicated by random sampling variation in the training sample and the test sample. But, insofar as the usual requirements for credible models are met, the results in Table 3 should be the results reported. With no test data, one would be left with Table 2.

In summary, dual-sample or split-sample procedures are easy to implement. Model selection is undertaken with the training data. Estimation and statistical inference is undertaken with the test data. The model selection procedure does not matter, and can

[18] There is a scattering of a few other minor crimes are in the baseline.

Table 3 Results from test data

	Estimate	Multiplier	Standard error	*p*-value
Intercept	2.010	–	0.20	0.0000
Assault conviction	0.425	1.52	0.28	0.1322
Drug conviction	0.584	1.77	0.19	0.0022
Gun conviction	0.763	2.14	0.47	0.1059
Number of priors	0.019	1.02	0.006	0.0011

range from exhaustive searches of the sort just illustrated to informal approaches that drop predictors from a "full model" for any reason whatsoever. When dual-sample or split-sample procedures are not practical, one is best off making a very clear distinction between analyses that are confirmatory and analyses that are exploratory. Statistical inference can be justified only for confirmatory analyses. Finally, model selection by itself implies little about the ultimate credibility of the model chosen. Conventional assumption still have to be reasonable well met.

Conclusions

There is no doubt that post-model-selection statistical inference can lead to biased regression parameter estimates and seriously misleading statistical tests and confidence intervals. The approaches by which the selection is done are for the issues raised in this paper unimportant. Informal data snooping is as dangerous as state-of-the-art model selection procedures.

Currently, there are five possible responses. If a case can be made for an appropriate model before the data are analyzed, one can proceed as the textbooks describe. Alternatively, the problem can be ignored if one can credibly argue that the model selection procedures will select a single model with a probability of near 1.0. If there are two random samples from the same population, or if it is possible to construct the equivalent from the data on hand, appropriate statistical inference may be undertaken even if there are substantial post-model-selection difficulties. When neither approach is practical, one can make a clear distinction between data analyses that are confirmatory and analyses that are exploratory. Statistical inference is appropriate only for the former. Finally, should all else fail, one can simply forego formal statistical inference altogether.

If after model selection, there remains more than one candidate model, new complications are introduced. The underlying inferential logic is flawed. If there are several candidate models, at best only one can correctly represent how the data were generated. The confidence intervals and statistical tests for all models but one will not perform as required, and the one model for which statistical inference can be appropriate is unknown. There is also the real possibility that all of the models are suspect in which case, all of the tests and confidence intervals can be compromised. In short, post-model-selection statistical inference can be further jeopardized when more than one model is selected.

Acknowledgments Richard Berk's work on this paper was funded by a grant from the National Science Foundation: SES-0437169, "Ensemble methods for Data Analysis in the Behavioral, Social and Economic Sciences." The work by Lawrence Brown and Linda Zhao was supported in part by NSF grant DMS-07-07033. Thanks also go to Andreas Buja, Sam Preston, Jasjeet Sekhon, Herb Smith, Phillip Stark, and three reviewers for helpful suggestions about the material discussed in this paper.

References

Barnett V (1983) Comparative statistical inference, 2nd edn. Wiley, New York

Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. Ann Stat 29(4):1165–188

Berk RA (2003) Regression analysis: a constructive critique. Sage Publications, Newbury Park

Blumstein A, Cohen J, Martin SE, Tonrey MH (eds) (1983) Research on sentencing: the search for reform, vols 1 and 2. National Academy Press, Washington, DC

Box GEP (1976) Science and statistics. J Am Stat Assoc 71:791–799

Breiman L (2001) Statistical modeling: two cultures (with discussion). Stat Sci 16:199–231

Brown LD (1967) The conditional level of student's *t* test. Ann Math Stat 38(4):1068–1071

Brown LD (1990) An ancillarity paradox which appears in multiple linear regression. Ann Stat 18(2):471–493

Candes E, Tao T (2007) The Dantzig selector: statistical estimation when *p* is much larger than *n*. Ann Stat 35(6):2313–2331

Cook DR, Weisberg S (1999) Applied regression including computing and graphics. Wiley, New York

Davies G, Dedel K (2006) Violence screening in community corrections. Criminol Public Policy 5(4):743–770

Efron B, Hastie T, Tibshinani R (2007) Discussion: the Dantzig selector: statistical estimation with *p* much larger than *n*. Ann Stat 35(6):2358–2364

Efron B (2007) Correlation and large-scale simultaneous significance testing. J Am Stat Assoc 102(477): 93–103

Freedman DA (1987) As others see us: a case study in path analysis (with discussion). J Educ Stat 12:101–223

Freedman DA (2004) Graphical models for causation and the identification problem. Eval Rev 28:267–293

Freedman DA (2005) Statistical models: theory and practice. Cambridge University Press, Cambridge

Freedman DA, Navidi W, Peters SC (1988) On the impact of variable selection in fitting regression equations. In: Dijkstra TK (eds) On model uncertainty and its statistical implications. Springer, Berlin, pp 1–16

Greene WH (2003) Econometric methods, 5th edn. Prentice Hall, New Yortk

Johnson BD (2006) The multilevel context of criminal sentencing: integrating judge- and county-level influences. Criminology 44(2):235–258

Lalonde RJ, Cho RM (2008) The impact of incarceration in state prison on the employment prospects of women. J Quant Criminol 24:243–265

Leeb H, Pötscher BM (2005) Model selection and inference: facts and fiction. Econ Theory 21:21–59

Leeb H, Pötscher BM (2006) Can one estimate the conditional distribution of post-model-selection estimators? Ann Stat 34(5):2554–2591

Leeb H, Pötscher BM (2008) Model selection. In: Anderson TG, Davis RA, Kreib J-P, Mikosch T (eds) The handbook of financial time series. Springer, New York, pp 785–821

Leamer EE (1978) Specification searches: ad hoc inference with non-experimental data. Wiley, New York

Manski CF (1990) Nonparametric bounds on treatment effects. Am Econ Rev Pap Proc 80:319–323

McCullagh P, Nelder JA (1989) Generalized linear models. 2nd edn. Chapman & Hall, New York

Morgan SL, Winship C (2007) Counterfactuals and causal inference: methods and principles for social research. Cambridge University Press, Cambridge

Morris N, Tonry M (1990) Prison and probation: intermediate punishment in a rational sentencing system. Oxford, University Press, New York

Olshen RA (1973) The conditional level of the *F*-test. J Am Stat Assoc 68(343):692–698

Ousey GC, Wilcox P, Brummel S (2008) Déjà vu all over again: investigating temporal continuity of adolescent victimization. J Quant Criminol 24:307–335

Petersilia J (1997) Probation in the United States. Crime Justice 22:149–200

Rubin DB (1986) Which ifs have causal answers. J Am Stat Assoc 81:961–962

Sampson RJ, Raudenbush SW (2004) Seeing disorder: neighborhood stigma and the social construction of broken windows. Soc Psychol Q 67(4):319–342

Schroeder RD, Giordano PC, Cernkovich SA (2007) Drug use and desistance processes. Criminology 45(1):191–222

Wooldredge J, Griffin T, Rauschenberg F (2005) (Un)anticipated effects of sentencing reform on disparate treatment of defendants. Law Soc Rev 39(4):835–874