SPECIFICATION SEARCH AND LEVELS OF SIGNIFICANCE IN ECONOMETRIC MODELS

Steven B. Caudill
Auburn University

and

Randall G. Holcombe
Florida State University

INTRODUCTION

One of the many underlying assumptions of the classical regression model is that the model is correctly specified. The estimates of a regression equation may be biased or inefficient when specification errors are present, so the researcher will want to guard against specification errors whenever possible. Before the use of the computer in econometrics, specification tests and the comparison of different specifications were time consuming processes. Today, however, alternative specifications of a model can be easily examined, and frequently they are.[1] Often, empirical articles in economic journals display tables of regression results showing the effects of adding or deleting variables.[2] Also common is the reporting of results in linear and log linear form, and the examination of trend stationary versus difference stationary models. Sometimes only a footnote mentions that alternate specifications were examined even though no mention of the fact is made in the final réporting of the results. There are good reasons for examining alternate specifications of a model and reporting the results;[3] however, the primary theme of this paper is that when a specification search is undertaken, levels of significance cannot be interpreted in the same way as when a single specification is examined. The computer printout might say that a coefficient of a variable is significant at the 5 percent level, meaning that the probability of rejecting the null hypothesis of no effect when the null hypothesis is true is .05, but if the estimate is one from many specifications examined, the actual level of confidence, or probability of rejecting a true null hypothesis, may considerably exceed .05.

This basic point has long been recognized by econometricians, in econometric theory at least,[4] but in practice, levels of significance are almost always reported as if only one specification of the model were examined, even when many specifications are openly reported.[5] One reason is that adjusting significance levels for the results of specification search is only recently possible. Some adjustment is possible using the

Steven B. Caudill: Economics Department, College of Business, Auburn University, 415 W. Magnolia, Room 203, Auburn, Alabama 36849-5242. E-mail: scaudill@business.auburn.edu
Randall G. Holcombe: Department of Economics, 246 Bellamy Building, Florida State University, Tallahassee, FL 32306-2045. E-mail: holcombe@coss.fsu.edu

bootstrap procedure described by Veall [1992]. The bootstrap provides an alternative method of evaluating significance levels, which is based on viewing the entire specification search as the estimator.

This paper describes the problem specification searches pose for inference, presents the results of some simulations for purposes of illustration, and uses the bootstrapping procedure to give a better estimate of statistical significance than a standard t-test. The value of the illustrations of specification searches is that they help demonstrate the severity of the problem. The examples presented below illustrate that in most cases, a researcher can undertake a specification search and report a statistically significant result regardless of whether the variables in a regression equation are actually related. The bootstrap procedure used to analyze the specification searches does provide another way to examine the true statistical significance of empirical results. Two different specification searches are examined: a "drop insignificant coefficients" search and a "biggest t-ratio" search. Both are shown to lead to larger than reported standard errors.

AN EXAMPLE WITH RANDOM NUMBERS

To illustrate the problem, a specification search was undertaken to test the hypothesis that $Y = f(X1)$, using data sets generated by a random number generator.[6] Because other variables might influence Y, the complete model was $Y = f(X1,..., X7)$. All of the observations were generated by a random number generator, and 60 observations were generated for each variable. One hundred data sets were generated so that the specification search procedure could be replicated 100 times.

Specification search was attempted in two dimensions. First, the intercept and all possible combinations of independent variables were included in regression models in an attempt to find a specification that would show the coefficient of $X1$ to be significant at a reported 5 percent level of confidence. A two-tailed test was used, since either sign on $X1$ is considered an acceptable (publishable) result. This specification search follows the common practice for empirical researchers to report that variables have been added or deleted from their models, due to multicollinearity or other types of problems. The second type of specification search was undertaken by looking at only half of the sample; that is, examining the model with only 30 observations rather than 60. Only one half of the sample was examined this way, although both halves could have been. This type of specification search is done for various reasons with various types of data. In time series, the years since 1950 might have been examined, or all years since 1900. Alternatively, the war years might be omitted from a sample. With cross-sectional data, some countries might be dropped from the data because they are LDCs, or smaller firms might be dropped from the sample because they behave differently from large firms.

The types of specification search done here were quite limited. Frequently, logs and first differences are examined, and the imaginative researcher can often think of more than just seven independent variables. Indeed, the test variable itself can be changed, if for example, M2 fits better than M1, or if the corporate AAA bond rate fits better than the 90-day treasury bill rate. In some cases, the entire data set can be changed, for example, by examining SMSA rather than state data, or by using quarterly rather than monthly data. The point is that the two types of specification search done here — choosing the best combination of *ceteris paribus* independent variables, and eliminating some observations — are a small subset of the possible specification searches that could be undertaken, but these are two of the commonly used types of specification search.

When the regression $Y = a + bX1$ was estimated using the one hundred data sets of random numbers, $X1$ was significant at the reported .05 level 6 times, which is close to what would have been expected with random numbers. When the sample was divided in half and run with only 30 observations, $X1$ was also significant at the reported .05 level 6 times, but only three of the occurrences overlapped. Thus, in 3 of the 100 data sets of random numbers, $X1$ was significant only with the larger sample of 60 observations, in 3 it was significant only in the sample of 30 observations, and in 3 cases it was significant both in the large and small samples.

Next, every different combination of the *ceteris paribus* variables $X2$ through $X6$ were tried along with $X1$ in both the small and the large samples. The result was that in 17 out of the 100 data sets of random numbers, $X1$ was "significant at a reported .05 level" in at least one specification. With random numbers one would expect results significant at the .05 level in only 5 out of 100 cases, but even a simple specification search greatly increases the probability of finding "significant" results.[7] In this case our specification searches produced "significant" results more than three times as often. The implication is that even a simple specification search like the one done here can triple the researcher's chance of finding significant results, even if no relationship exists, and may give the researcher as much as a 17 percent chance of finding results reported as "significant at the 5 percent level." If a more complex search is done, the odds obviously improve, and as noted above, one need not just search over one data set as was done here.

AN EXAMPLE WITH REAL DATA

Gujarati [1988, 243] presents a small data set of 16 annual observations for the years 1968-1983 on variables related to the demand for telephone cable in the United States. The dependent variable in the demand regression is Total Plastic Purchases, which represents sales in millions of paired feet, and the independent variables are Gross National Product (*GNP*), housing starts (*STARTS*), the unemployment rate (*UNEMP*), the prime rate lagged 6 months (*PRIME*), and customer line gains (*LINE*). The full model contains five independent variables which means that if one wants to search for the best specification by omitting variables from the full model, there are 2^5 or 32 possible combinations of variables. If one eliminates the specification that contains only an intercept, that leaves 31 specifications that can be searched. To examine the impact of various specification searches on statistical significance, this section focuses on two independent variables, *GNP* and *PRIME*. The analysis first assumes that the researcher is searching for evidence that *GNP* has a statistically significant effect on Total Plastic Purchases, and then that the researcher is searching for evidence that *PRIME* has a statistically significant effect.

The first step in the investigation is to estimate the full model which contains all five of the independent variables. The results are given in Table 1, and are shown as model number 1. This specification has an R^2 of 0.82, and shows that *GNP* is significant at better than the .10 level, and that the coefficients of *STARTS*, *UNEMP*, and *LINE* are significant at better than the .05 level. Only the coefficient of PRIME is insignificant in this specification. The variables *GNP* and *PRIME* were chosen for further investigation because from the specification including all variables, it appears that the coefficient of *GNP* is statistically significant and the coefficient of *PRIME* is not.

This section will consider only the simplest kind of specification search: trying out different combinations of independent variables. The remainder of Table 1 shows the results from every possible combination of independent variables. In the full model the coefficient on *GNP* is 4.88, but it varies quite a bit in other specifications, ranging from a high of 9.69 in model 3 to a low of −0.64 in model 15.[8] This variation is certainly due to collinearity between *GNP* and the other regressors.[9] *GNP* has simple correlations of 0.73, 0.81, and −0.67 with *UNEMP*, *PRIME*, and *LINE*, respectively. The large range of coefficient estimates on *GNP* occurs even though in the full model the coefficient is statistically significant. The highest *t*-ratio on *GNP* also occurs in model 3, where it is an impressive 4.15. In a theoretical paper on the subject, Caudill and Holcombe [1987] show that when a specification search is undertaken to look for "statistically significant" results associated with a specific variable, the coefficient will be biased away from zero, so it is not surprising that the specification with the highest *t*-ratio on *GNP* is also the specification with the largest coefficient. Of the 16 specifications in which *GNP* appears, the associated *t*-ratio is at least 2 in 6 specifications, and exceeds 1.9 in another 2 specifications, suggesting that it should be relatively easy to uncover a specification in which the coefficient of *GNP* is statistically significant.

A specification search on *PRIME* uncovers even more interesting possibilities. The coefficient estimates on *PRIME* range from 346.96 in model 10 to -315.67 in model 3, and both are statistically significant at well above the .05 level, despite the fact that in the model with all of the independent variables, the *t*-ratio was only 0.08. Again, multicollinearity is probably responsible for the large range of estimates. *PRIME* has simple correlations of 0.81, −0.75, and 0.70 with *GNP*, *STARTS*, and *UNEMP*, respectively. If one is searching for a specification to support a particular hypothesis with regard to *PRIME*, one could pick a specification to show that its effect is positive and strong, negative and strong, or anything in between. While one might find a single regression result convincing if it showed a coefficient with a *t*-ratio of 3.75, as the coefficient of *PRIME* does in model 10, or −2.78, as the coefficient of *PRIME* does in model 3, this particular example shows that if the regression equation was chosen as the result of a specification search, inferences may be misleading. In this example we can actually see all of the other possible regressions so we can easily

Table 1
All Possible Regressions

Model Number	CONST	GNP	STARTS	UMEMP	PRIME	LINE	R²
1	5962.66	4.88	2.36	-819.13	12.01	-851.39	0.82
	(2.38)*	(1.94)	(2.80)	(4.46)	(0.08)	(2.91)	
2	6292.37		3.48	-635.61	247.04	-657.97	0.76
	(2.25)		(5.05)	(3.50)	(2.64)	(2.14)	
3	8196.68	9.69		-957.12	-315.67	-1020.25	0.68
	(2.71)	(4.15)		(4.15)	(2.78)	(2.80)	0.68
4	1037.31	-0.63	3.33		231.71	-42.05	
	(0.29)	(0.18)	(2.52)		(1.03)	(0.12)	0.49
5	6031.92	5.05	2.31	-824.38		-864.44	0.82
	(2.68)	(3.71)	(4.73)	(4.90)		(3.71)	
6	581.72	2.39	2.87	-471.85	246.34		0.67
	(0.26)	(0.78)	(2.68)	(2.51)	(1.54)		
7	13360.36			-517.62	-18.81	-598.29	0.19
	(3.16)			(1.65)	(0.14)	(1.11)	
8	789.32		3.17		199.56	-44.14	0.48
	(0.25)		(3.33)		(1.55)	(0.13)	
9	11415.14		2.46	-5.66		-1062	0.60
	(4.63)		(3.53)	(2.58)		(3.26)	
10	1440.87		3.43	-413.94	346.96		0.65
	(0.77)		(4.37)	(2.44)	(3.75)		
11	3161.43	5.24			-207.95	-91.23	0.19
	(0.74)	(1.65)			(1.23)	(0.21)	
12	6946.61	5.60		-810.53		-367.39	0.46
	(1.86)	(2.47)		(2.89)		(1.06)	
13	2061.76	7.86		-559.52	-110.36		0.46
	(0.79)	(2.80)		(2.45)	(1.01)		
14	1861.27	2.28	2.27			-213.85	0.44
	(0.52)	(1.08)	(2.72)			(0.65)	
15	764.90	-0.64	3.34		243.39		0.48
	(0.29)	(0.19)	(2.64)		(1.27)		
16	195.60	6.21	1.50	-469.72			0.60
	(0.09)	(3.26)	(2.39)	(2.37)			
17	382.56	3.14	1.98				0.42
	(0.14)	(1.93)	(2.87)				
18	3664.04	6.11		-624.10			0.41
	(1.75)	(2.75)		(2.85)			
19	2579.94	5.24			-184.86		0.19
	(0.84)	(1.71)			(1.50)		
20	2828.56	2.87				263.67	0.09
	(0.65)	(1.11)				(0.78)	
21	5444.79		1.45	-29.84			0.25
	(2.50)		(1.76)	(0.16)			
22	497.72		3.18		211.15		0.48
	(0.23)		(3.48)		(2.44)		
23	5511.65		2.36			-432.92	0.38
	(5.02)		(2.83)			(1.67)	
24	8837.60			-317.51	76.12		0.49
	(7.16)			(1.22)	(0.72)		
25	13020.99			-521.49		-551.36	0.19
	(3.92)			(1.73)		(1.38)	

Table 1 (Cont.)
All Possible Regressions

Model Number	CONST	GNP	STARTS	UMEMP	PRIME	LINE	R²
26	8281.81 (2.68)				-38.51 (0.27)	-94.40 (0.20)	0.01
27	5599.39 (2.30)	1.51 (0.81)					0.04
28	5160.58 (4.51)		1.51 (2.14)				0.25
29	8734.34 (7.25)			-186.49 (1.02)			0.07
30	7683.00 (9.54)				-14.51 (0.19)		0.00
31	7503.50 (7.29)					10.18 (0.04)	0.00

assess the alternative regression models. In general, the effects of specification searches on inferences depend on the type of search and the data set searched.

In the preceding analysis, among other things, we reported the highest and lowest parameter estimates resulting from the specification searches. Presenting this range is advocated by Leamer [1983] and is called extreme bounds analysis; it does provide some evidence on the consequences of specification search. Veall [1992] extends Leamer's extreme bounds analysis to obtain the entire distribution of the parameter being investigated, and not simply the range, providing more evidence on the consequences of specification searches. The next section examines the procedure suggested by Veall.

SYSTEMATIC SEARCH PROCEDURES

Assume that the researcher begins with an initial model, but has undertaken some specification search, presumably due to some dissatisfaction with the results of the initial model.[10] Two types of specification search rules are examined in this section. The first is the "drop insignificant coefficients" rule.[11] In this search, the full model is first estimated and then variables associated with statistically insignificant coefficients (except for the variable of interest) are deleted. Then, this reduced model is estimated and the parameter of interest is noted. The second search procedure follows the "biggest t-ratio" rule, which involves estimating all regressions which include the variable of interest and then choosing the specification that produces the most statistically significant coefficient estimate for that variable. Veall [1992] investigates the effects of a third type of specification search which involves the stepwise elimination of variables associated with insignificant coefficients.

If one is interested in the coefficient of *GNP*, the only insignificant variable in the full model is *PRIME*, and when it is dropped from the regression, the coefficient on *GNP* rises from 4.88 to 5.05 (in model 5) and its *t*-ratio rises from 1.94 to 3.71, again due to their highly collinear relationship. *GNP* appears to be much more significant after this apparently reasonable specification search, and also has a slightly larger estimated effect. If the search is undertaken to find the specification with the highest *t*-ratio on *GNP*, then model 3 is the result, with a *t*-ratio of 4.15 and a coefficient of 9.69, more than doubling the estimated effect of *GNP*. Clearly, the estimated magnitude of the *GNP* effect is changed by the specification search. In a theoretical paper on the subject, Caudill and Holcombe [1987] show that specification search biases the absolute value of estimated coefficients away from zero, and that effect is clearly visible here.

PRIME presents an interesting case with these same types of specification searches. If one uses the .10 level to identify statistically significant coefficients, then the coefficients of all of the variables except for *PRIME* are statistically significant in the full model, so no independent variables would be dropped and *PRIME* would be found insignificant. However, if the .05 level of significance is chosen, *GNP* is not significant in the full model, so is dropped to produce model 2, in which *PRIME* has a coefficient of 247.04 and a *t*-ratio of 2.64, making it very significant. Under the "drop the insignificant coefficients" rule, ironically, if one chooses the more inclusive .10 level of significance, then *PRIME* is not statistically significant, but if the more stringent .05 level is chosen, *PRIME* becomes significant. Again, this result is due to multicollinearity which exists, to some degree, in all data sets. If the specification with the highest *t*-ratio on *PRIME* is chosen, the coefficient is 346.96. Again, note that the coefficients are biased away from zero. The example shows that even what appears to be a reasonable specification search procedure, like dropping insignificant coefficients, may not produce results much different from choosing the specification with the highest *t*-ratio.

AN ASSESSMENT

In an effort to examine the statistical consequences of these common types of specification searches in a more rigorous manner, this section employs the bootstrap method of Efron [1982], which is used in a randomization test which is discussed by Kennedy [1995]. A similar application of bootstrapping was used in conjunction with model selection and a different type of specification search by Veall [1992]. Once the initial model has been estimated, the residuals, e_i, and the predicted values of the dependent variable are obtained, where

$$\begin{aligned} e_i &= y_i - \hat{y}_i \\ \hat{y}_i &= X_i\hat{\beta} \end{aligned} \tag{1}$$

and y refers to the dependent variable, Total Plastic Purchases.

Each of the residuals is assigned probability $1/n$ (in our case n=16) and a discrete random variable is created with probability function given by

$$p(e_i) = \frac{1}{n} \; for \; i = 1, ..., n. \tag{2}$$

These residuals are randomly drawn, resulting in a new set of residuals called e_i^*. For example, e_1^* might happen to be e_3. These new residuals are next added to the predicted values of y from the original regression to obtain new values of y called, y_i^*, where

$$y_i* = \hat{y}_i = e_i* \; for \; i = 1, ..., n. \tag{3}$$

These new ys are used with the same independent variables to produce new parameter estimates. These parameter estimates are noted and then the process is repeated some suitably large number of times, in this case 1000, and the maxima, minima, means, and variances of the estimates of the parameters of interest are tabulated. Veall [1992] refers to this procedure as residual bootstrapping.

First consider the variable *GNP*. Recall that in the full model the estimated coefficient is 4.88 with a t-ratio of 1.94. Using the "drop insignificant coefficients" decision rule, 1000 bootstrap replications found the lowest coefficient estimate to be −2.88 and the highest estimate to be 11.04. The mean coefficient estimate was 5.08 and the standard deviation from these 1000 coefficients was 1.57, producing a t-ratio of 3.24. Note that this is not the mean t-ratio from the 1000 data sets produced by the bootstrap, but rather is the t-ratio produced by taking the standard deviation of the 1000 coefficients from all of the data sets produced by bootstrapping. The result of this procedure is close to model 5 in Table 1, which was the model selected originally, following the "drop insignificant coefficients" specification search. The coefficient is very close, and the t-ratio is slightly lower. When the specification with the highest t-ratio is chosen, the results from the 1000 bootstraps produced a mean value of 7.59 with a t-ratio of 3.04. When compared to model 3 in Table 1, the coefficient and t-ratio are both lower, but are still considerably higher than in the original full model. Furthermore, a t-test on the coefficients produced from the bootstrap shows *GNP* to be statistically significant at better than the .05 level.

When the same exercise is undertaken with *PRIME*, the "drop insignificant coefficients" rule yields a mean coefficient value of 35.84 with a standard deviation of 179.83, so the bootstrap procedure indicates that PRIME is not actually statistically significant. Similarly, when the model with the biggest t-ratio is selected, the mean value of the coefficient is 135.95 with a standard deviation of 342.85, again suggesting that *PRIME* is not actually statistically significant. The bootstrap procedure confirms what the entire set of regressions in Table 1 suggested: *GNP* is a statistically significant independent variable, and *PRIME* is not. This bootstrapping procedure thus can be a way of identifying whether the result of a systematic specification search is really statistically significant.

Although *PRIME* is not statistically significant when the mean of 1000 coefficients in the "biggest t-ratio" models is compared with its standard deviation, in every case the individual regression equation generated a statistically significant coefficient on *PRIME*. The reason the average did not produce a statistically significant result when the individual regressions did is that the large range of values on *PRIME* was distributed almost symmetrically around zero. The coefficients from the specification searches in the 1000 data sets ranged in value from −433.14 to 461.97, with a mean of 35.84. In contrast, the range for the GNP coefficients under the "biggest t-ratio" specification search was from −2.88 to 11.04, with a mean of 7.89. Because the distribution of estimated coefficients was shifted toward the positive for GNP, the average coefficient was more than two times the average standard deviation, indicating statistical significance for GNP.

TABLE 2
Number of Specifications Yielding "Significant" Results

Number of Regressions (Out of 16 Possible) With "Significant" Results	Significant at 0.05 level Frequency	Cumulative Percent	Significant at the 0.10 level Frequency	Cumulative Percent
0	0	100.0	0	100.0
1	8	100.0	0	100.0
2	101	99.1	17	100.0
3	167	89.0	49	98.3
4	350	72.3	204	93.4
5	200	37.3	209	73.0
6	120	17.3	231	52.1
7	31	5.3	166	29.0
8	20	2.3	76	12.4
9	3	0.3	44	4.8
10	0	0.0	4	0.4

An alternative analysis of bootstrapped results is given by Efron and Gong [1983] and Freedman and Navidi [1986]. These authors suggest examining the percentage of bootstrap simulations containing a "significant" result for a variable. A different view of the bootstrapped results is possible in this framework. Although the bootstrap procedure suggests that *GNP* is statistically significant but *PRIME* is not, a specification search on this data set using the "biggest t-ratio" criterion will always turn up a specification with a "significant" coefficient on *PRIME*. Table 2 examines results from the regressions run on the 1000 bootstrapped data sets in a search for a "significant" coefficient on *PRIME*, and shows that for every one of the 1000 data sets, there is at least one model specification in which *PRIME* is found to be statistically significant at the reported level. Using all different combinations of independent variables, 16 specifications contain *PRIME*. The first column of Table 2 counts the number of specifications out of 16 that produced a statistically significant coefficient on *PRIME*. The next two columns show the results of testing for significance at the .05 level. Thus, reading the table, out of the 1000 bootstrapped data sets, in 8 of them only one model specification produced a statistically significant coefficient on *PRIME*. In an additional 101 data sets, two specifications produced a coefficient on *PRIME* significant at the reported .05 level. An additional 167 data sets produced three significant specifications, and so forth down the table. The cumulative percent column shows what percentage of the 1000 data sets produced at least the number of specifications in the left column with statistically significant *PRIME* coefficients. Thus, 100 percent of the time one could find at least one specification of the model with a significant *PRIME* coefficient, 99.1 percent of the time one could find at least two significant specifications, and so forth. One could find as many as 9 out of 16 specifications with *PRIME* coefficient reported significant at the .05 level in .3 percent of the data sets, but in no instances did 10 of the specifications have significant coefficients.

One is 10 percent more likely to find results reported to be significant at the .10 level even if no relationship exists, and the same type of results are shown in the right-most two columns of Table 2. In that case, 100 percent of the data sets had at least two specifications in which *PRIME* could be reported significant at the .10 level, 98.3 percent had at least three regressions with reported significant coefficients, and at the bottom of the table, 0.4 percent of the time, there were ten out of 16 specifications with *PRIME* coefficients reported to be significant at the .10 level. The bootstrapping results suggested that *PRIME* is not truly a statistically significant variable, yet in an examination of these 1000 data sets created by bootstrapping, at least one specification in which *PRIME* was reported to be significant could always be found, and in most cases there were many specifications to choose from. At the reported .05 level of significance, at least four specifications reporting "statistically significant" results could be found 72.3 percent of the time. Thus, with some specification search, a researcher could report many different models supporting the hypothesis that *PRIME* is a statistically significant determinant of Total Plastic Purchases.

CONCLUSION

More than a decade ago, Mayer [1980] and Leamer [1983] warned that one cannot take regression results and tests of statistical significance at face value. When many specifications of a model have been examined, actual levels of significance are not as high as they would appear from looking at reported significance levels. This paper has shown using a data set of random numbers that relatively minor and generally accepted techniques of specification search can greatly enhance the probability of finding "statistically significant results," when no actual relationship is present. Then, using a real-world data set, it demonstrated that even when a coefficient is not actually statistically significant, it is likely that many different specifications can be produced reporting it to be "statistically significant," using common specification search techniques and measuring significance with a standard *t*-test. Results like this are reported in academic journals all the time. Leamer [1983] advocated reporting all different plausible specifications to demonstrate the robustness of the results, but as the demonstration above has shown, even then it may be easy to report many plausible "significant" results for variables that really are not statistically significant.

Unfortunately, there is no simple test like a *t*-test that can indicate the significance level of a regression coefficient if it is being reported after a specification search.[12] However, building on Veall [1992], this paper has further examined a methodology that may be able to shed some light on the true statistical significance of a coefficient by using the results from many bootstrapped data sets. The bootstrap method, introduced in this context by Veall, is examined for two types of specification search. A "drop insignificant coefficients" rule and a "biggest *t*-ratio" are examined. The specification search is applied the same way to all of the bootstrapped data sets, producing coefficients on the variable of interest. The resulting set of coefficients is then examined to see if its mean is truly statistically significant. This method can provide an indication of the true significance of a coefficient, even after a complicated specification search. The technique is very general and can be used in to investigate the effects of any specification search, but the impact of the specification search depends on the type of search and the relationships in the data set. In general, standard errors get larger if a specification search has taken place, but exactly how much larger must be determined on a case-by-case basis.

Reported significance levels must be interpreted with caution when more than one specification of a model is examined. There is a greater than 5 percent probability of finding results "significant at the 5 percent level" in one of many specifications, but less than a 5 percent probability of finding significant results in all specifications. Beyond that, the type of specification search undertaken and the frequency of "significant" results will influence the actual significance of the findings.[13] The bootstrapping technique described above can provide some additional insight, but because few researchers are applying Veall's bootstrap method, one must be skeptical of any empirical results reported in economics journals. Articles normally do not provide enough information about the entire process that produced the results for readers to accurately evaluate their statistical significance.

NOTES

1. One of the earliest methods of specification search initiated by the power of the computer was stepwise regression. While the technique has fallen out of favor, it is probably more benign than the typical specification search leading up to published results. Leamer [1978] discusses several types of specification searches.
2. For example, in the March 1997 issue of the *American Economic Review* (the most recent issue at the time of the draft), five of eight empirical articles contain tables with multiple reported specifications.
3. For a discussion of this point, see Mayer [1980] and Leamer [1983].
4. See Lovell [1983] for a discussion of this point. In a thoughtful article, Keuzenkamp and Magnus [1995] question the entire undertaking of testing for statistical significance, and challenge readers to offer concrete examples of when such tests have affected economists' beliefs about economic propositions.
5. See Rizzo [1978] for a discussion of econometric tests, considering the problem of selecting *ceteris paribus* variables. Holcombe [1989] also considers problems of interpreting results after specification searches.
6. All empirical work in this paper, including the generation of the random numbers, was done in SAS. All random data is generated from a uniform distribution. The seed value was set to 0 so the first variate was determined by the time on the university clock.
7. This finding is in large part due to multicollinearity in the random data. Every regression model contains the variable *X1*, and *X1* is known to have no effect on *Y*. Adding additional explanatory variables should not change that situation. In fact, if the variables *X2* through *X7* were truly uncorrelated with *X1* in every sample, the coefficient on *X1* would not change from regression to regression (it should be zero). In this situation the only way additional regressors could lead to "statistically significant" results for *X1* would be if the variance were reduced sufficiently so that the *t*-ratios for *X1* could become larger. However, the variables *X2* through *X7* are evidently not completely uncorrelated with *X1* in the random samples we generated as evidenced by the increase in the numbers of "significant" results when regressors are added and specifications are searched.

8. Reporting the highest and lowest estimate from a specification search is called extreme bounds analysis and is advocated by Leamer [1983].

9. Some of the simple correlations are quite high, even exceeding 0.7 in absolute value, as the following correlation matrix indicates

Variable	TPP	GNP	STARTS	UNEMP	PRIME	LINE
TPP	1.00	0.21	0.50	-0.26	-0.05	0.01
GNP		1.00	-0.35	0.73	0.81	-0.67
STARTS			1.00	-0.46	-0.75	0.61
UNEMP				1.00	0.70	-0.81
PRIME					1.00	-0.83
LINE						1.00

10. As Spanos [1995] explains, the assumption that the correct model is known can never be verified with nonexperimental data because one can never know what other real-world factors might have influenced the data in a data set. Spanos notes, "The *ceteris paribus* clause cannot be made operational without carefully designed experiments" [1995, 224].

11. Dropping insignificant coefficients, in theory, leads to one type of pretest estimator. Interested readers are directed to Judge et al. [1985, 75]. This search is examined here because dropping insignificant coefficients is a common practice in our profession.

12. However, this problem was discussed, and a test proposed, by Malinvaud [1966, 205-207].

13. As Keuzenkamp and Magnus [1995, 11] note, "Any level of significance can be obtained by making the sample size large enough, unless the null hypothesis is exactly true."

REFERENCES

Caudill, S. B., and Holcombe, R. G. Coefficient Bias due to Specification Search in Econometric Models. *Atlantic Economic Journal*, September 1987, 30-34.

Efron, B. *The Jackknife, the Bootstrap, and Other Resampling Plans.* SIAM 1982.

Efron, B. and Gong, G. A Leisurely Look at the Bootstrap, the Jackknife, and Cross-validation, *American Statistician*, 1983, 36-48.

Freedman, D. A. and Navidi, W.C. Models for Adjusting the Census. *Statistical Science*, 1986, 3-11.

Gujarati, D. N. *Basic Econometrics*. New York: McGraw-Hill, 1988.

Holcombe, R. G. *Economic Models and Methodology*. New York: Greenwood, 1989.

Judge, G. J., Griffiths, W. E., Hill, R. C., Luktpohl, H., and Lee, T. *The Theory and Practice of Econometrics*. New York: John Wiley and Sons, 1985.

Kennedy, P. Randomization Tests in Econometrics. *Journal of Business and Economic Statistics*, 1995, 85-94.

Keuzenkamp, H. A. and Magnus, J. R. On Tests and Significance in Econometrics. *Journal of Econometrics*, 1995, 5-24.

Leamer, E. E. *Specification Searches: Ad Hoc Inference with Non-Experimental Data.* New York: Wiley, 1978.

__________. Let's Take the Con out of Econometrics. *American Economic Review*, 1983, 31-43.

Lovell, M. C. Data Mining. *Review of Economics and Statistics,* February 1983, 1-12.

Malinvaud, E. *Statistical Methods in Econometrics.* Chicago: Rand McNally, 1966.

Mayer, T. Economics as a Hard Science: Realistic Goal or Wishful Thinking. *Economic Inquiry*, April 1980, 165-78.

Rizzo, M. J. Praxeology and Econometrics: A Critique of Positivist Economics, in *New Directions in Austrian Economics,* edited by L. M. Spadaro. Kansas City: Sheed, Andrews, and McMeel, 1978.

Spanos, A. On Theory Testing in Econometrics Modeling with Nonexperimental Data. *Journal of Econometrics*, 1995, 189-226.

Veall, M. Bootstrapping the Process of Model Selection: An Econometric Approach. *Journal of Applied Econometrics*, 1992, 93-9.