Department of Economics
McAnulty College of Liberal Arts
Duquesne University
Pittsburgh, Pennsylvania

ESTIMATED EXHAUSTIVE REGRESSION: REDEFINING THE CROSS-MODEL TEST STATISTIC VIA SIMULATION

Jeremy Phillips

Submitted to the Economics Faculty
In partial fulfillment of the requirements for the degree of
Bachelor of Arts in Economics
December 2008

Faculty Advisor Signature Page

Amy Phelps, Ph.D. Date
Assistant Professor of Quantitative Sciences

Antony Davies, Ph.D. Date
Associate Professor of Economics

ESTIMATED EXHAUSTIVE REGRESSION: REDEFINING THE CROSS-MODEL TEST STATISTIC VIA SIMULATION

Jeremy Phillips, BA

Duquesne University, 2008

Regression-based data mining techniques, such as stepwise regression, are often used by econometricians in situations where specifying a single hypothesis is not desirable, often because the number of possible explanatory variables is too large or because not enough is known about the subject to construct a theory driven hypothesis. While such methods can be extremely helpful, they are highly susceptible to generating spurious results. Davies (2008) proposes a data mining technique called Estimated Exhaustive Regression (EER) which, he shows, can effectively differentiate between deterministic and spurious factors within data sets where varying degrees of multicollinearity are present. A critical step of EER is the calculation of the cross-model chi-square test statistic; however, it can be reasonably argued that this test statistic is not actually from a chi-square distribution. The purpose of this research is to (1) redefine the cross-model test statistic, (2) use simulation to show that the redefined test statistic is standard normally distributed, and (3) show that EER is less susceptible to generating spurious results than stepwise regression.

The results of the analysis show that EER has the potential to be a very useful tool for econometricians. It is shown that EER selects approximately half as many extraneous variables as stepwise regression.

Keywords: regression based data mining, data mining, all subsets regression, best subsets regression

Table of Contents

I. Introduction……..... 5

II. Literature Review……. 7

III. Methodology……..... 13

IV. Results and Analysis…… 20

V. Economic Implications……. 25

VI. Suggestions for Future Research…….. 26

VII. Conclusion……..... 27

VIII. References……..... 29

I. Introduction

Multiple regression analysis is a powerful tool that allows researchers, such as econometricians, to test a hypothesized relationship between a dependent variable and a set of independent variables. A researcher properly implements a regression analysis by testing a hypothesis that was carefully developed using a theory driven technique; however, in practice, it is often not desirable to specify a single hypothesis, usually because the set of possible explanatory variables is too large or simply because not enough is known about the subject to construct a theoretically sound hypothesis. In either case, the researcher might resort to a regression-based data mining technique, such as stepwise regression, that attempts to sort through the data and arrive at the optimal hypothesis.

But what exactly is data mining and how does it work? Data mining refers to a wide array of algorithms, methods, and techniques that can be applied to data sets in order to extract useful information. The fundamental strategy of most data mining techniques is to search and/or sort through data sets and detect trends and relationships within them. Multiple regression analysis is the most widely used statistical technique in the data mining community. Typically, a researcher will use some sort of regression-based data mining algorithm, such as stepwise regression, to find the set of explanatory variables that best fits the data, i.e. the optimal hypothesis. These techniques are not all necessarily as systematic as stepwise regression; for example, many researchers practice "data-peeking" which is done by running various models prior to running the final model. Data mining is useful because it often reveals previously unknown or unexplored relationships that are not immediately obvious from theory. These findings often open up

new avenues for research or offer potential for improvements to previous research. Despite its effectiveness, data mining is somewhat of a controversial practice because many, if not most, data mining techniques are highly susceptible to generating spurious results. One reason for this is because most data mining algorithms search for the model that best fits the data, while in reality the true model may not be the best fitting one.

Although data mining has proven to be an extremely useful and practical tool in many disciplines and sciences, this paper focuses on data mining and its application to economics. Economics data has a tendency to contain varying degrees of multicollinearity and, as a result, many regression-based data mining techniques are highly susceptible to spurious results. The multicollinearity makes it difficult for many of these techniques to differentiate between two variables that are highly correlated with one another and many times the variable with the most explanatory power is kept, despite the possibility that this extra explanatory power may simply be the result of random chance. Also, variables sometimes show up as significant because of a direct relationship they have with a true explanatory variable, creating an indirect, yet seemingly significant, relationship with the dependent variable. Variables such as these are regularly and wrongfully selected by many regression-based data mining algorithms.

The regression-based data mining technique in question in this paper is one first proposed by Davies (2008) called *Estimated Exhaustive Regression* (*EER*). *EER* is a technique that attempts to alleviate the spurious results problem by targeting and testing variables individually rather than comparing entire models or model fits. This is essentially done by examining the behavior of the parameter estimates of a particular variable across multiple models. Although Davies has shown empirically that *EER*

works, a critical step of *EER* is the calculation of the *cross-model chi-square test statistic*, but, it can be reasonably argued that this test statistic is not actually from a chi-square distribution. The purpose of this research is to (1) redefine the cross-model test statistic, (2) use simulation to show that the redefined test statistic is standard normally distributed, and (3) show that EER is less susceptible to generating spurious results than stepwise regression.

II. Literature Review

Data mining is certainly not a new practice; however, its popularity has increased over the last decade due to an explosion in computational technology and advancements in the construction and use of databases. Despite the decreasing cost of computational resources, many data mining procedures are still not feasible to implement simply because the problems are so complex that it would take years to solve them on a personal computer. For this paper we focus on regression-based data mining techniques that use an *exhaustive* approach to model selection. Such techniques are often classified as *all subsets regression* (Davies, 2008) or *best subset regression* (Hocking, 1967) because they consider all possible models that can be specified using a single data set. The following framework and notation will be used to better explain previous research:

Suppose a researcher is interested in a dependent variable, Y, and has a set, X, of k possible explanatory variables:

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \tag{2.1}$$

$$X = \{x_1, x_2, \dots, x_k\} \text{ where } x_i = \begin{bmatrix} x_{i,1} \\ \vdots \\ x_{i,n} \end{bmatrix} \tag{2.2}$$

Note that $\mathcal{P}(X)$, the powerset of X, consists of $2^k - 1$ unique subsets of X, excluding the empty set:

$$\mathcal{P}(X) = \big\{\{x_1\}, \{x_1, x_2\}, \dots, \{x_1, \dots, x_k\}\big\} \tag{2.3}$$

Let each of element of $\mathcal{P}(X)$ correspond to a unique regression model where only the variables in that element are included. For example:

$$\{x_i, x_j, x_k\} \longrightarrow \hat{Y} = \hat{\alpha} + \hat{\beta}_i x_i + + \hat{\beta}_j x_j + + \hat{\beta}_k x_k \tag{2.4}$$

There are then $2^k - 1$ possible regression models the researcher can construct using X.

An *exhaustive* regression-based data mining technique considers all of these $2^k - 1$ possible regression models in an attempt to find the best possible hypothesis. There are two primary schools of thought on how to make this selection. The first is to consider the entire model and select the best one based on some sort of logical comparison of competing models. Stepwise regression is an example of this because as it adds and removes variables it uses Mallows' C_p to compare the fits between competing models and then takes a *step* in the direction of the superior one. While stepwise regression is commonly thought to be an *all subsets* approach to model selection, in reality only a small number of models are actually considered. The fact is that stepwise regression uses an iterative approach that will eventually settle at a local optimum and it is not guaranteed that this optimum will be global. Kuk (1984) attempts to get around this problem by applying the same principle of stepwise regression, using Mallows' C_p to compare models, to a proportional hazards model while also using an *exhaustive* or *all subsets* approach. He applies his methodology to multiple myeloma data and finds that his results are remarkably different from the results obtained by previous studies which

had utilized stepwise regression. He concludes, "new insights are gained and the superiority of all subsets regression over stepwise regression is clearly demonstrated."

Kuk's application of all subsets regression did not require an extreme amount of computational resources because of the nature and small size of the problem. In most cases; however, an all subsets approach may be very time consuming, expensive, and sometimes not even feasible. Hocking and Leslie (1967) developed an algorithm that attempts to return the same result as all subsets regression without using the computational resources required to actually examine all subsets. They found that their methodology consistently returned 'good' models while using minimal computing resources. To achieve this, they used a procedure similar to stepwise regression. One key difference; however, is the method they used for selecting competing models; they used the *standardized total squared error* as the criterion. One interesting and relevant result of their research is they found that if a set consists of 10 possible explanatory variables (1023 possible models) they needed to only compute, on average, 299 subsets in order to find the *best* one. Although they claim their methodology is one which should be used for discarding variables in multiple regression analysis, I do not find that their research truly addresses the problem of spurious results as it does not target individual variables, rather, it only compares models which can easily be influenced by spurious relationships.

The second school of thought is to consider all $2^k - 1$ models and examine an individual variable's behavior across them, individually choosing which variables should be in the model and which ones should not. An interesting take on cross-model comparisons is presented by Leamer (1985). He applies an *extreme-bounds* test to a

variable's parameter estimates across all possible models as a method for measuring that variables *robustness*. By calculating the l*ower extreme bound* and *upper extreme bound* for each variable and comparing them, Leamer argues one can draw conclusions about that variable's *robust* nature. Consider the variable x_i in the framework presented above. If you were to run all $2^k - 1$ regression models you would have a collection of 2^{k-1} parameter estimates and standard errors for x_i[1]. The *lower extreme bound* is defined as $min(\hat{\beta}_i - 2s_{\beta_i})$ across all sampled models where s_{β_i} is the standard error of β_i. The *upper extreme bound* is defined as greatest value of $max(\hat{\beta}_i + 2s_{\beta_i})$. Leamer argues that if the *lower extreme bound* is negative and the *upper extreme bound* is positive then x_i is not a *robust* variable. The basic premise here is that if a variable's parameter estimate tends to change signs between models then it is less likely to be a true explanatory variable; you would expect it to be more stable if it were. Critics generally argue that this is an extremely intolerant test that results in a high probability of discounting important variables simply because it takes only one model, and some random chance, to throw a variable out.

Sala-I-Martin (1997) is also interested in *extreme-bounds* tests; however, he is not as quick as Leamer to discard variables. He argues, for example, that if 90% of a variable's parameter estimates are greater than zero then this variable is more *robust* than a variable with only 50% of its parameter estimates greater than zero. While Leamer would reject both of these variables, Sala-I-Martin believes that the degree of *robustness* is worth considering. Sala-I-Martin applies his methodology to the selection of variables

[1] Note that while there are $2^k - 1$ possible models, only half of them will include a given variable, thus each variable will have $\frac{2^k}{2} = 2^{k-1}$ parameter estimates.

used to model economic growth and challenges the popular view found in empirical growth literature, that no variables are robust. He concludes, "a substantial number of variables can be found to be strongly related to growth."

This paper focuses on a cross-model variable selection procedure that was originally proposed by Davies (2008) called *Estimated Exhaustive Regression* (*EER*). The principle argument behind *EER* is that the parameter estimates of a true explanatory variables across a sufficiently large sample of all $2^k - 1$ possible regression models will be more stable than the parameter estimates of variables with no underlying relationship. As a measurement of a variable's stability, Davies proposes a *cross-model test chi-square statistic* and defines it as follows: Considering the framework mentioned above, suppose we randomly select *N* elements from $\mathcal{P}(X)$ and run the corresponding *N* regression models defined by (2.4). Let *j*[2] be equal to the number of *N* models that include the variable X_i. We then have j parameter estimates, $\{\hat{\beta}_{i1} \dots \hat{\beta}_{ij}\}$, and j corresponding standard errors, $\{s_{\beta_{i1}} \dots s_{\beta_{ij}}\}$. The *cross-model test chi-square statistic* for variable X_i is calculated as follows:

$$c_i = \sum_{h=1}^{j} \left(\frac{\hat{\beta}_{ih}}{s_{\beta_{ih}}} \right)^2 \sim \chi_j^2 \qquad (2.5)$$

Davies asserts that under reasonable assumptions the *cross-model test statistic* comes from a *chi-square* distribution with *j* degrees of freedom. To correct for type II errors, he divides the entire test statistic by j and then assumes it to be chi-square distributed with one degree of freedom. This test statistic is then used to test the null hypothesis that $\beta_i = 0$. Davies experiments with a variety of critical values; however, in his final results

[2] Note that $E(j) = \frac{N}{2}$

he uses a critical value of 2.0. If the test statistic is greater than or equal to 2.0 then the null hypothesis is rejected and the variable is considered to be a true explanatory variable; otherwise, it is considered to be insignificant. This procedure is then repeated for all *k* variables and only the true explanatory variables are included in the final model.

By running Monte-Carlo tests on c_i, Davies (2008) is able to measure the effectiveness of *EER*. Generic data was generated, ensuring that the data did not violate any of the assumptions of linear regression and building into it varying degrees of multicollinearity in an effort to mimic real, economic data. Simulations run on these data sets show that *EER* is able to correctly select deterministic factors upwards of 85% of the time while selecting each spurious variable less than 20% of the time. Stepwise regression was slightly more effective at selecting true deterministic factors; however, it was more susceptible to falsely identifying a variable as deterministic when it was not. The results also show that as the number of possible explanatory variables, k, increases, *EER* becomes more effective. These results serve as strong evidence that *EER* is a very practical and effective regression-based data mining technique that can differentiate between deterministic and spurious factors.

It is important to note that a key assumption of *EER* is that the parameter estimate of a given variable from a given model is independent from the parameter estimate of the same variable from a different model, i.e. $\hat{\beta}_{ih}$ is independent of $\hat{\beta}_{ij}$. For various reasons, this assumption is highly unlikely to be true and as a result of this dependence, the *cross-model test statistic* is not *chi-square* distributed. In fact, its distribution does not appear to be a well known one such as a gamma or Weibull. The central purpose of this paper is to redefine the *cross-model test statistic* in order to make it more consistent and

theoretically sound. Once we have redefined the test statistic and established its properties, we can run Monte Carlo tests similar to those run by Davies.

III. Methodology

Suppose you are using linear regression to test the following competing hypotheses:

$$y = \alpha + \beta_{1.1}X_1 + \beta_{2.1}X_2 + u \tag{3.1}$$

$$y = \alpha + \beta_{1.2}X_1 + u \tag{3.2}$$

For both hypotheses you run the following linear regressions:

$$\hat{y} = \hat{\alpha} + \hat{\beta}_{1.1}X_1 + \hat{\beta}_{2.1}X_2 \tag{3.3}$$

$$\hat{y} = \hat{\alpha} + \hat{\beta}_{1.2}X_1 \tag{3.4}$$

You then have two different parameter estimates for β_1, namely $\hat{\beta}_{1.1}$ and $\hat{\beta}_{1.2}$. Assuming that $\beta_1 = 0$, it is well known that the standardized form of $\hat{\beta}_{1.1}$ and $\hat{\beta}_{1.2}$ will be *t-distributed* and, assuming a sufficiently large number of degrees of freedom, will be essentially *standard* normally distributed.

It is well known that the sum of two normally distributed, independent random variables will also be normally distributed; thus, assuming independence, it follows that $\hat{\beta}_{1.1} + \hat{\beta}_{1.2}$ will also be normally distributed, specifically:

$$\hat{\beta}_{1.1} + \hat{\beta}_{1.2} \sim N\left(0, s_{\beta_{1.1}} + s_{\beta_{1.2}}\right) \tag{3.5}$$

It then follows that:

$$\frac{\hat{\beta}_{1.1} + \hat{\beta}_{1.2}}{2\left(s_{\beta_{1.1}} + s_{\beta_{1.2}}\right)} \sim z \tag{3.6}$$

For various reasons, it is clear that $\hat{\beta}_{1.1}$ and $\hat{\beta}_{1.2}$ will not be independent, in fact they will generally be highly dependent and correlated. We can use the following argument to suggest that regardless of this dependence, their sum will still be normally distributed: We have no reason to believe that the dependence between $\hat{\beta}_{1.1}$ and $\hat{\beta}_{1.2}$ is any stronger than what can be explained by their correlation; therefore, we consider the following linear relationship:

$$\hat{\beta}_{1.1} = \alpha + \mathrm{B}\hat{\beta}_{1.2} + \upsilon \tag{3.7}$$

Where α is a constant and $\upsilon \sim N(\bar{\upsilon}, s_{\upsilon})$. By substituting (3.7) into (3.5) we have:

$$\hat{\beta}_{1.1} + \hat{\beta}_{1.2} \longrightarrow \quad \alpha + \mathrm{B}\hat{\beta}_{1.2} + \upsilon + \hat{\beta}_{1.2} \tag{3.8}$$

$$\longrightarrow \quad \alpha + (\mathrm{B} + 1)\hat{\beta}_{1.2} + \upsilon \tag{3.9}$$

We can then break this sum down into 3 components:

$$\alpha \tag{3.10}$$

$$(\mathrm{B} + 1)\hat{\beta}_{1.2} \tag{3.11}$$

$$\upsilon \tag{3.12}$$

We know that $\hat{\beta}_{1.2}$ is normally distributed, thus a scalar multiple of it, (3.11), will also be normally distributed, specifically $(\mathrm{B} + 1)\hat{\beta}_{1.2} \sim N\left(0, \mathrm{B} + 1 + s_{\beta_{1.2}}\right)$. Since (3.10) is a constant, adding it to (3.11) will only change its distribution's mean,

We can then say under reasonable certainty that:

$$\alpha + (\mathrm{B} + 1)\hat{\beta}_{1.2} \sim N\left(\alpha, \mathrm{B} + 1 + s_{\beta_{1.2}}\right) \tag{3.13}$$

Recall that υ is essentially the residual we get when we fit (3.7), and as an assumption of *GLS*, it must be uncorrelated with $\hat{\beta}_{1.2}$. We make the reasonable assumption that zero correlation between $\hat{\beta}_{1.2}$ and υ establishes their independence or implies that any

dependence between them is *weak* at best. We can then deduce that (3.13) is independent of υ.

Recalling that $\upsilon \sim N(\bar{\upsilon}, s_{\upsilon})$, we can conclude the following:

$$\alpha + (\mathrm{B}+1)\hat{\beta}_{1.2} + \upsilon \sim N\left(\alpha + \bar{\upsilon}, \mathrm{B}+1+s_{\beta_{1.2}} + s_{\upsilon}\right) \tag{3.14}$$

Therefore:

$$\hat{\beta}_{1.1} + \hat{\beta}_{1.2} \sim N\left(\alpha + \bar{\upsilon}, \mathrm{B}+1+s_{\beta_{1.2}} + s_{\upsilon}\right) \tag{3.15}$$

$$\frac{\hat{\beta}_{1.1} + \hat{\beta}_{1.2}}{2} \sim N\left(\frac{\alpha + \bar{\upsilon}}{2}, \frac{\mathrm{B}+1++s_{\beta_{1.2}} s_{\upsilon}}{2}\right) \tag{3.16}$$

Result (3.15) tells us that the sum of a variable's parameter estimates across models is normally distributed and result (3.16) essentially extends (3.15) to a variable's *average* parameter estimate across models. What we are really after is how a variable's *average* standardized parameter estimate across models is distributed. We can extend result (3.16) to the standardized case and reason that $\frac{\mathrm{B}+1+s_{\beta_{1.2}}+s_{\upsilon}}{2} = 1$.[3]

Based on this argument, the mean of this distribution, $\frac{\alpha+\bar{\upsilon}}{2}$, will be impossible to calculate exactly using only one sample because it is a function of the correlation of parameter estimates across *multiple* samples; however, we will show empirically that the expected value of this average across variables is zero.

The argument above serves as motivation for the construction of a *cross-model test statistic* similar to that proposed by Davies (2008). Keeping consistent with the earlier notation, we define c_i as follows:

[3] Note that the standard deviation of $\frac{\widehat{\beta}_{1.1}+\widehat{\beta}_{1.2}}{2\left(s_{\beta_{1.1}}+s_{\beta_{1.2}}\right)}$ will equal one regardless of dependence.

- $$c_i = \frac{1}{j}\sum_{h=1}^{j}\left(\frac{\hat{\beta}_{ih}}{s_{\beta_{ih}}}\right) \sim z \quad (3.16)$$

We can use properly constructed simulations to get empirical estimates of our new *cross-model test statistic's* sampling distribution nd test our hypothesis that it is standard normally distributed. Because what we are looking for is *a sampling distribution*, we must define populations from which we can sample data from. From each sample we can calculate test statistics and by re-sampling the same populations a sufficiently large number of times we can construct *empirical cumulative distribution functions* for these test statistics. We define our data populations using the following procedure:

Data Populations[4]

1. Let X_1 be defined as a standard normal random variable. Thus the population for X_1 can be defined by the standard normal distribution.
2. Let the populations of $X_2 - X_{15}$ be defined such that $X_i = \gamma_i X_1 + \upsilon_j$, where γ_i are randomly selected from the standard normal distribution and are held constant across all observations. υ_j are randomly selected from a normal distribution with mean 0 and variance 0.1 and are randomly selected for each observation.
3. Let the population of y be defined such that $y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_2 X_2 + u_j$ where $\alpha = \beta_1 = \beta_2 = \beta_3 = 1$ and u_j is standard normally distributed and is randomly selected for each observation .

[4] This is a slightly modified version of the procedure used by Davies (2008)

When we sample data from these populations we expect the relationships to vary slightly across samples. F*igure 3.1* and *Figure 3.2* illustrate an example of two samples from the same population. Notice that although the relationships are very similar, they do vary slightly across the samples. It is this variation that allows us to use this data generation methodology to simulate how the *cross-model test statistic* is distributed.

Figure 3.1 - Sample 1

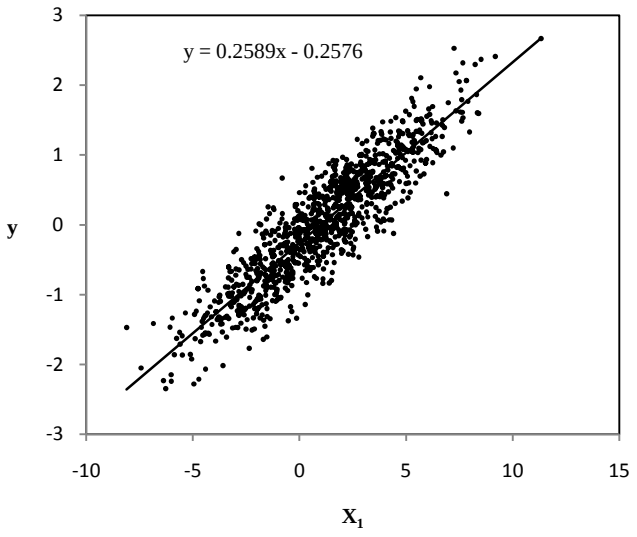
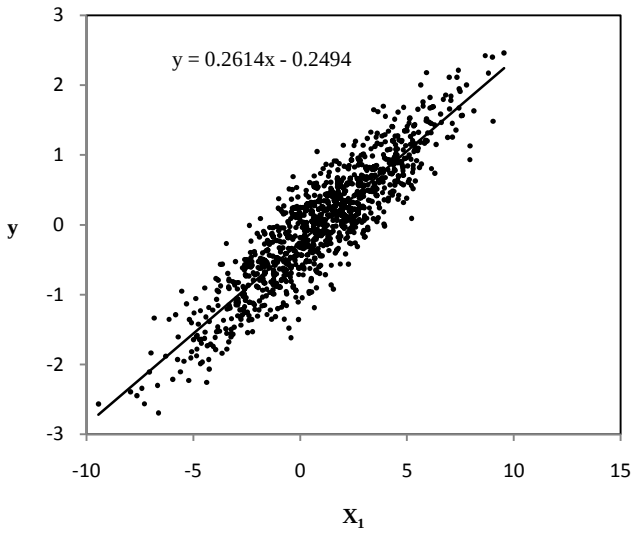


Figure 3.2 - Sample 2



The data is designed to represent economic data, which tends to contain varying degrees of multicollinearity. The correlation matrix in *Figure 3.3* clearly illustrates the presence of multicollinearity in a typical data set generated by this methodology.
It is easy to see in this example that all the variables are highly correlated with the dependent variable even though the dependent is truly only a function of X_1, X_2, and X_3. It is also easy to see the strong relationships the independent variables have with one another. Both of these conditions compound the model selection process and cause confusion for the researcher when trying to develop a hypothesis. This type of data

structure offers us a great opportunity to test how effectively *EER* can see through spurious correlations and the multicollinearity.

Figure 3.3 -Sample Correlation Matrix

	y	**X_1**	**X_2**	**X_3**	**X_4**	**X_5**	**X_6**	**X_7**	**X_8**	**X_9**	**X_{10}**	**X_{11}**	**X_{12}**	**X_{13}**	**X_{14}**	**X_{15}**
y	1.00	0.94	0.94	-0.50	-0.93	0.75	-0.76	0.85	0.85	-0.88	0.93	-0.88	0.91	0.28	0.83	-0.17
X_1	0.94	1.00	0.99	-0.62	-0.99	0.79	-0.82	0.91	0.90	-0.94	0.99	-0.95	0.97	0.30	0.88	-0.18
X_2	0.94	0.99	1.00	-0.61	-0.98	0.78	-0.81	0.91	0.90	-0.93	0.98	-0.93	0.96	0.30	0.87	-0.18
X_3	-0.50	-0.62	-0.61	1.00	0.62	-0.51	0.52	-0.57	-0.56	0.59	-0.62	0.57	-0.61	-0.21	-0.56	0.08
X_4	-0.93	-0.99	-0.98	0.62	1.00	-0.78	0.81	-0.90	-0.89	0.94	-0.98	0.94	-0.96	-0.30	-0.87	0.18
X_5	0.75	0.79	0.78	-0.51	-0.78	1.00	-0.64	0.74	0.71	-0.76	0.78	-0.74	0.77	0.23	0.69	-0.13
X_6	-0.76	-0.82	-0.81	0.52	0.81	-0.64	1.00	-0.74	-0.75	0.78	-0.81	0.77	-0.79	-0.25	-0.71	0.13
X_7	0.85	0.91	0.91	-0.57	-0.90	0.74	-0.74	1.00	0.83	-0.86	0.90	-0.86	0.88	0.29	0.82	-0.16
X_8	0.85	0.90	0.90	-0.56	-0.89	0.71	-0.75	0.83	1.00	-0.85	0.89	-0.86	0.87	0.27	0.80	-0.17
X_9	-0.88	-0.94	-0.93	0.59	0.94	-0.76	0.78	-0.86	-0.85	1.00	-0.93	0.89	-0.92	-0.29	-0.83	0.17
X_{10}	0.93	0.99	0.98	-0.62	-0.98	0.78	-0.81	0.90	0.89	-0.93	1.00	-0.93	0.96	0.30	0.88	-0.18
X_{11}	-0.88	-0.95	-0.93	0.57	0.94	-0.74	0.77	-0.86	-0.86	0.89	-0.93	1.00	-0.91	-0.30	-0.84	0.17
X_{12}	0.91	0.97	0.96	-0.61	-0.96	0.77	-0.79	0.88	0.87	-0.92	0.96	-0.91	1.00	0.30	0.86	-0.15
X_{13}	0.28	0.30	0.30	-0.21	-0.30	0.23	-0.25	0.29	0.27	-0.29	0.30	-0.30	0.30	1.00	0.27	-0.08
X_{14}	0.83	0.88	0.87	-0.56	-0.87	0.69	-0.71	0.82	0.80	-0.83	0.88	-0.84	0.86	0.27	1.00	-0.16
X_{15}	-0.17	-0.18	-0.18	0.08	0.18	-0.13	0.13	-0.16	-0.17	0.17	-0.18	0.17	-0.15	-0.08	-0.16	1.00

We use the following procedure to generate empirical examples of how the *cross-model test statistic* is distributed:

Sampling Distribution Simulation

1. Generate data populations using the procedure outlined above.
2. Take a 1,000 observation sample of all variables from the data populations defined in Step 1.
3. Randomly select 1,000 regression models from the possible 32,767 that can be constructed using variables $X_1 - X_{15}$.

4. Run the 1,000 regression models selected in Step 3 using the sampled data from Step 2.
5. Calculate the *cross-model test statistic*, as defined by (3.16), for variables $X_1 - X_{15}$ using the results from Step 4.
6. Repeat steps 2-5 a total of 1,000 times, being sure that the all the data is being sampled from the same populations that were defined in Step 1.

As a result of this simulation, we will have a set of 1,000 *cross-model test statistics* for each variable, $X_1 - X_{15}$. Because these test statistics come from multiple samples of the same population, we can use them to construct an empirical estimation of our target sampling distribution.

If these simulations confirm our hypothesis, that our *cross-model test statistic* is standard normally distributed, we can use the test statistic as a tool for model selection. By running Monte Carlo tests on *EER* similar to those performed by Davies, we can compare *EER's* effectiveness to the effectiveness of backward-stepwise regression. The following procedure is used to perform said Monte Carlo tests:

Monte Carlo Testing *EER*

1. Generate data populations using the procedure outlined above.
2. Take a 1,000 observation sample of all variables from the data populations defined in Step 1.
3. Randomly select 500 regression models from the possible 32,767 that can be constructed using variables $X_1 - X_{15}$.

4. Run the 500 regression models selected in Step 3 using the data set generated in Step 2.
5. Calculate the *cross-model test statistic*, as defined by (3.16), for variables $X_1 - X_{15}$ using the results from Step 4.
6. Perform a two-tailed hypothesis test on each test statistic calculated in Step 5 using a .05 level of significance and assuming that the test statistic is standard normally distributed under the null hypothesis that $\beta_i = 0$. The final model will consist of only those variables that do not pass this hypothesis test.
7. Run backwards-stepwise regression on the data set generated in Step 2 using .05 as the level of significance.
8. Repeat steps 1-7 a total of 2,000 times, generating new data populations and data sets for each iteration.

IV. Results and Analysis

The results from the sampling distribution simulation show us considerable empirical evidence that our *cross-model test statistic* comes from a normal distribution. *Figure 4.1* shows us the empirical cumulative distribution function obtained from our simulation for the *cross-model test statistics* of variables $X_4 - X_{15}$. These appear to all be normally distributed; however, we do not see convincing visual evidence that they standard normally distributed. We know that none of variables should be statistically significant in the correctly specified model; thus, we can argue that there should be nothing inherently different between the distributions of their *cross-model test statistics*. Following that reasoning, it may make sense to blend or average all of these distributions

together.[5] The result of blending these distributions is shown in *Figure 4.2*. We see overwhelming visual evidence that this blended distribution is a standard normal; in fact, the comparison passes a Kolmogorov-Smirnov test at a .05 level of significance.

Figure 4.1 - Empirical CDF of the cross-model test statistic of spurious factors

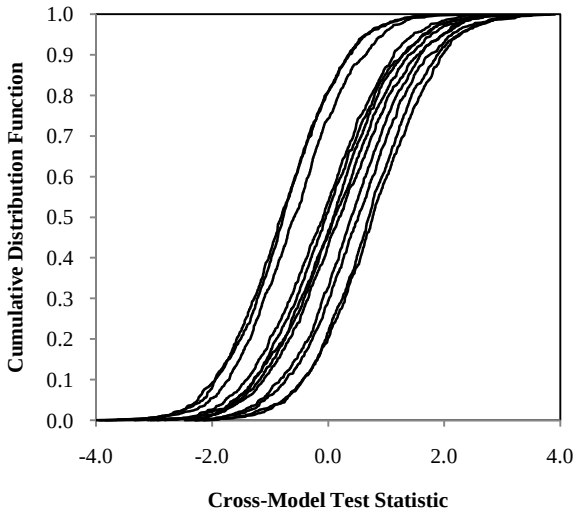
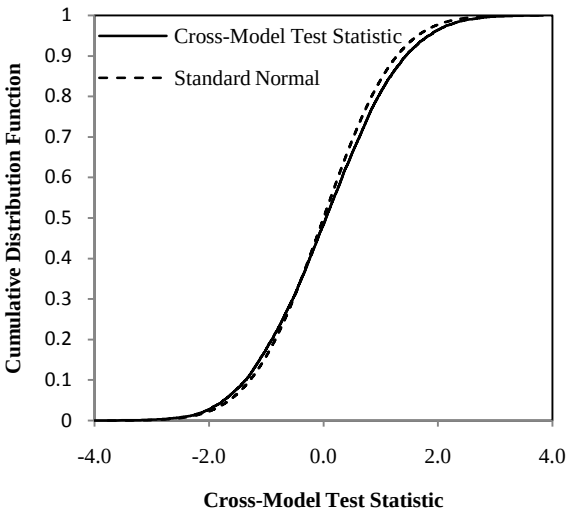


Figure 4.2 - Blended Empirical CDF of the cross-model test statistic for a spurious factor



The distributions in *Figure 4.1* are from a simulation that consisted of the random selection of 1,000 regression models over 1,000 population samples. *Figure 4.3* illustrates what happens as we increase the number of regression models to 2,000 and the number of population samples to 2,000. *Figure 4.4* illustrates the results of a similar simulation that used only 50 regression models and 5,000 samples. Notice that as the number of samples increases, the distributions converge toward the standard normal. The empirical evidence supports our hypothesis that the cross-model test statistic is approximately standard normally distributed. Lastly, *Figure 4.5* illustrates the results of a simulation using 25 regression models and 10,000 samples. As the sample size becomes larger, these distributions begin to slightly converge. Blending these

[5] Blending the distributions is done by merging all of the sets of test statistics into one set.

distributions in the same manner as performed above results in a distribution that is almost exactly equal to the standard normal distribution; this is shown in *Figure 4.6*. Again, the comparison passes a Kolmogorov-Smirnov test at a .05 level of significance. This result supports our claim that the expected mean of the *cross-model test statistic* for a given variable will be zero. From these results we can conclude that our hypothesis is correct, the distribution of a given variable's *cross-model test statistic*, when its parameter estimate in the correctly specified model is zero, is expected to be approximately standard normal.

Figure 4.3 - 2,000 Models / 2,000 Samples

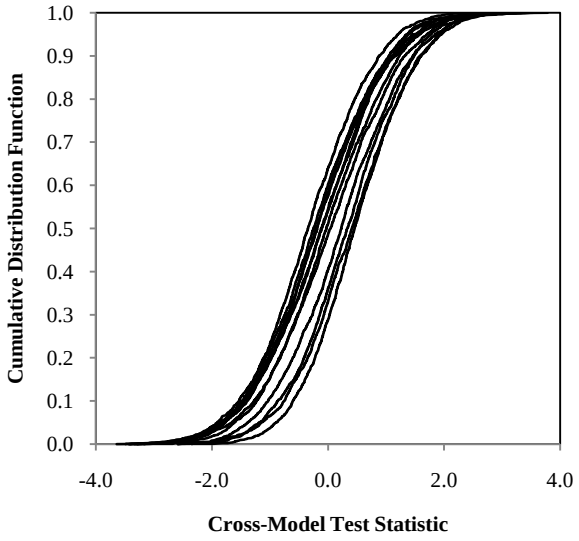


Figure 4.4 - 50 Models / 5,000 Samples

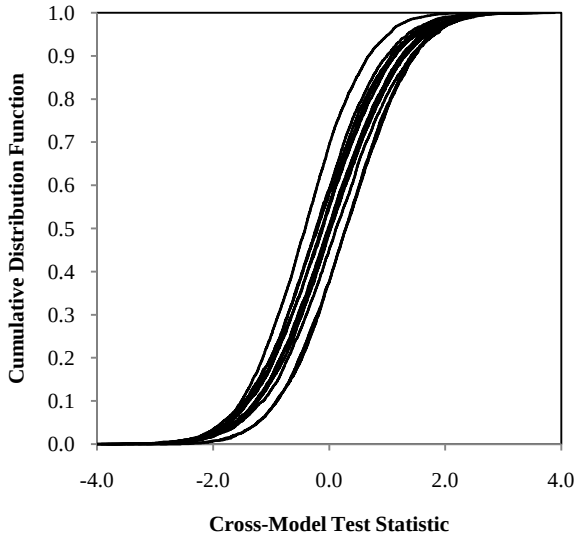


Figure 4.5 - 25 Models / 10,000 Samples

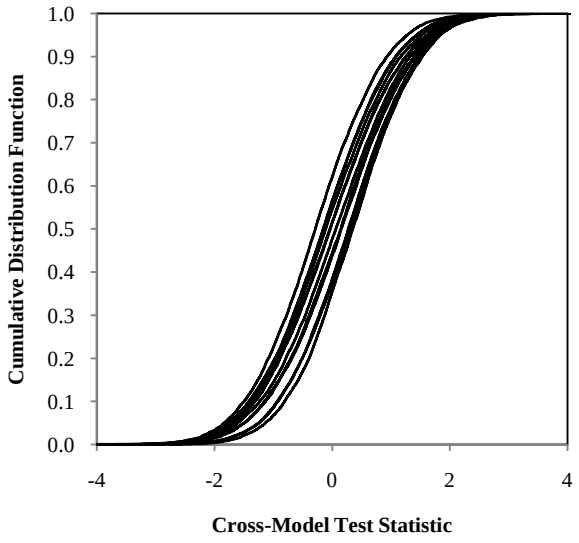
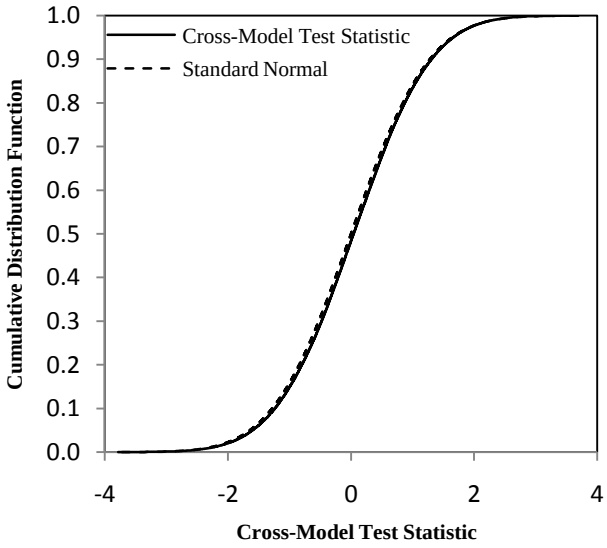
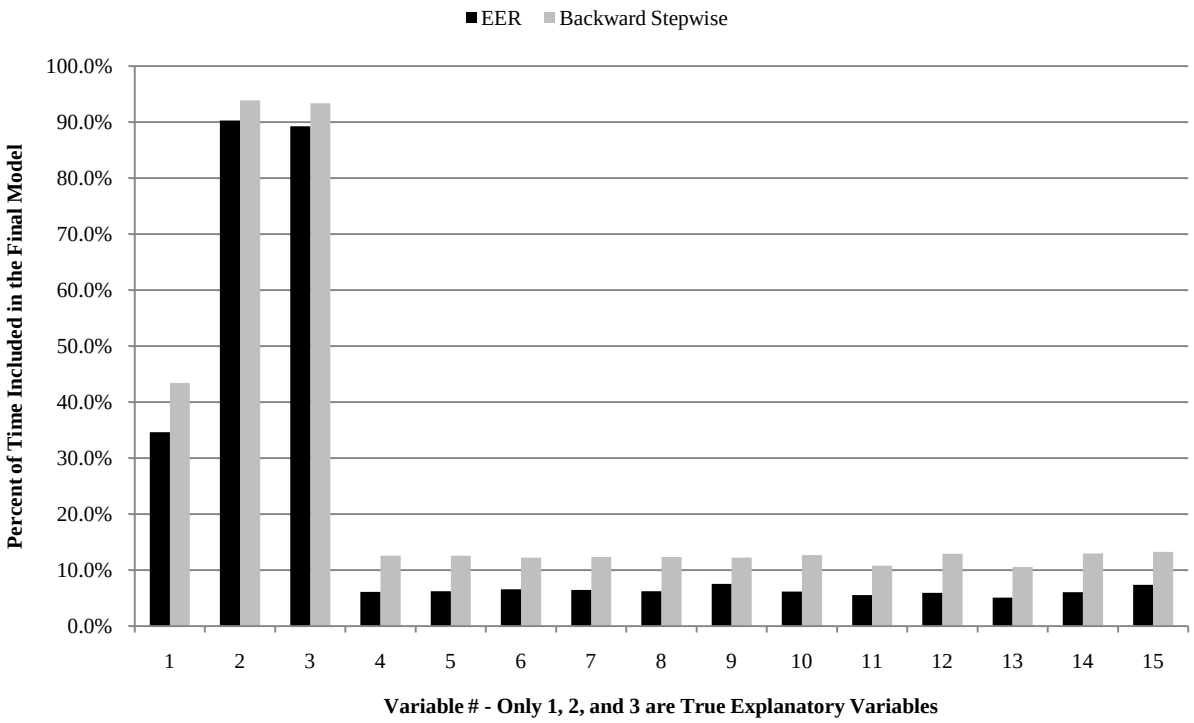


Figure 4.6 - 25 Models / 10,000 Samples - Blended



We can conclude that when performing the hypothesis test in the *EER* procedure we should use a two-tailed test with critical values of approximately -1.96 and 1.96 and only include a variable in the final model if its *cross-model test statistic* is either less than -1.96 or greater than 1.96. The results of performing the simulation outlined above using these critical values is summarize in *Figure 4.4.*

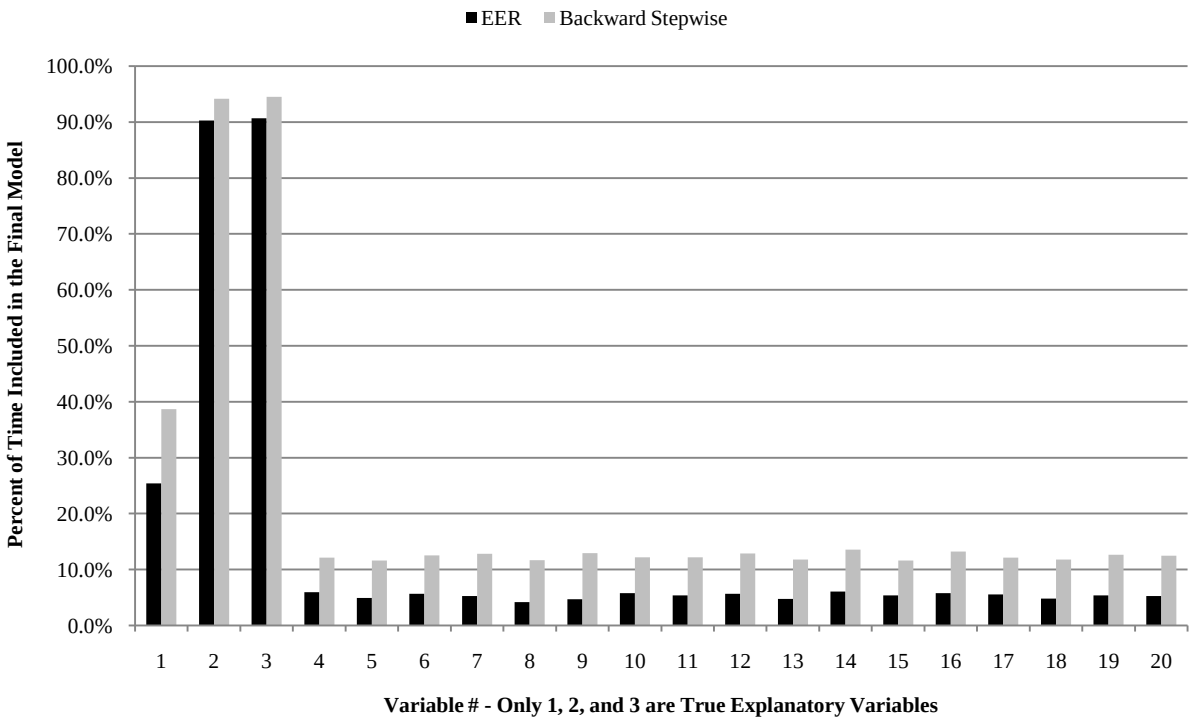
Figure 4.7 - Monte Carlo Test on EER - 2,000 Trials - 15 Variables



These results are interesting for a couple reasons. The first reason is that both *EER* and stepwise regression have a difficult time including X_1 in the final model. This is likely because the multicollinearity between the variables is inherently stronger across X_1 since all other variables are generated as a function of it. While this step in the data generation process creates the multicollinearity between all variables, it will naturally results in stronger correlations for X_1. The second interesting result is that the rate at which *EER* selected an unrelated variable to be in the model was 6.3% while for stepwise regression this rate was 12.3%. This evidence supports our claim that *EER* is less susceptible to generating spurious results than stepwise regression.

Figure 4.8 illustrates the results of the same Monte Carlo test outlined above; however, this time the number of spurious variables is increased from 12 to 17 ($X_4 - X_{20}$).

Figure 4.8 - Monte Carlo Test on EER - 2,000 Trials - 20 Variables



The results are similar to the case where there were only 15 total variables. Again *EER* does a better job than stepwise regression at leaving extraneous variables out of the model. The rate at which *EER* selected an unrelated variable to be in the model was 5.3% while for stepwise regression this rate was 12.4%. *EER's* performance improved slightly in this respect.

V. Economic Implications

Although the results of this analysis can be applied to several different sciences and disciplines, their application to economics stands out. For example, there is a large collection of literature written concerning variable selection in economic growth models and there are many contradicting conclusions. Some researchers conclude that there exist very few or no variables that are robust enough to correlate systematically with

economic growth (Lavine & Renelt, 1992), while others find strong evidence that certain sets of variables are strongly correlated with growth (Sala-I-Martin, 1997). One of the problems with modeling economic growth is the vary degrees of multicollinearity that are generally found across variables. The multicollinearity makes the variable selection process confusing for the researcher and it is common practice for them to resort to a regression-based data mining technique. These techniques do not necessarily have to be as systematic as stepwise regression. For example, any time a researcher begins adding and removing variables or testing multiple hypotheses they are essentially data mining. The problem with these techniques is that of multiple comparisons because the more models that are tested, the higher the probability that a variable randomly becomes significant.

We have shown that in the case of multicollinearity, stepwise regression is highly susceptible to generating spurious results. We have also shown that *EER* performs significantly better than stepwise regression in terms of being able to eliminate variables that do not belong in the model. Because it is natural for economic data to contain varying degrees of multicollinearity, and because the simulations used in this analysis target variable selection in the presence of multicollinearity, the results suggest that *EER* has the potential to be a very effective tool for economists.

VI. Suggestions for Future Research

The results of the sampling distribution simulations run in this analysis offer no evidence that the *cross-model test statistic* we have defined are *not* normally distributed.

They seem to be very well behaved; however, the argument I offer for why this is the case may not be accepted by some as a sound *mathematical proof*. It is clear that the standardized parameter estimates of a given variable across multiple models are dependent with one another. The average of these estimates will only be normally distributed if it is the case that their joint probability distribution is also normal. One avenue for possible future research would be to prove that this is always the case.

The theory and effectiveness of *EER* has only been tested with data sets specific to those used in this analysis. It is unclear whether or not these results can be extended to all variable selection problems. Further testing and development of the theory is required in order to extend these results to the general case. One area that requires particular attention is our assumption that the expected mean of the *cross-model test statistic* is zero for all variables. In reality, each variable is likely to have a different mean. As a whole, all these means will likely be distributed about zero; however, it may be possible to estimate this mean more exactly. By assuming it is always zero, we are essentially guaranteeing that over a large enough sample our false negative rate will be only 5%; however, this does not guarantee that the false negative rate for a particular variable will be 5%. This is a problem that should be further investigated, though we may not be able to avoid it.

VII. Conclusion

The purpose of this research was to improve upon the *EER* procedure proposed by Davies (2008) by redefining the *cross-model test statistic* and showing that this test statistic is from a standard normal distribution and to run simulations testing *EER's*

effectiveness compared to stepwise regression. We are able to show both theoretically and empirically that the *cross-model test statistic* we defined in (3.16) is from a standard normal distribution. By running a two-tailed hypothesis test using this test statistic we are able to show that *EER* is an effective tool for variable selection. What is encouraging about our results is that our false negative rate (that is the rate at which we include a variable in the final model when it should not be included) is only around 5-6%, which is essentially right where it should be considering we are working at a .05 level of significance. These results support our hypothesis that (3.16) is from a standard normal distribution. The stepwise procedure typically has a false negative rate of around 12%, even though it will only include a variable in a model if it is significant at the .05 level. The problem is that many variables make it into the final models only because they are significantly influential in that model simply due to random chance.

We have demonstrated that *EER* is a more superior data-mining technique than stepwise regression in the sense that it is less susceptible to generating spurious results. Specifically, this analysis targeted data sets that contained varying degrees of multicollinearity in an attempt to mimic typical economic data. The results show that EER is more superior in cases of multicollinearity than stepwise regression; thus, this suggests that EER has the potential to be a very useful variable selection tool for economists.

VIII. References

Davies, A. (2008, August). Exhaustive Regression: An Exploration of Regression-Based Data Mining Techniques Using Super Computation. *Research Program on Forecasting* , Working Paper No. 2008-008.

Farrar, D. E., & Glaber, R. R. (1967). Multicollinearity in regression Anlysis: The Problem Revisted. *The Review of Economics and Statistics , 49* (1), 92-107.

Hocking, R. R., & Leslie, R. N. (1967). Selection of the Best Subset in Regression Analysis. *Technometrics , 9* (4), 531-54.

Kuk, A. Y. (1984). All Subsets Regression in a Proportional Hazards Model. *Biometika , 71* (3), 587-592.

Leamer, E. E. (1985). Sensitivity Analses Would Help. *American Economic Review , 57* (3), 309-313.

Sala-I-Martin, X. X. (1997). I Just Ran Two Million Regressions. *American Economic Association , 87* (2), 178-183.