Research Program on Forecasting

Exhaustive Regression

An Exploration of Regression-Based Data Mining Techniques Using Super Computation

Antony Davies, Ph.D.
Associate Professor of Economics
Duquesne University
Pittsburgh, PA 15282

Research Fellow
The Mercatus Center
George Mason University
Arlington, VA 22201

antony@antolin-davies.com

RPF Working Paper No. 2008-008
http://www.gwu.edu/~forcpgm/2008-008.pdf

August 8, 2008

RESEARCH PROGRAM ON FORECASTING
Center of Economic Research
Department of Economics
The George Washington University
Washington, DC 20052
http://www.gwu.edu/~forcpgm

Research Program on Forecasting (RPF) Working Papers represent preliminary work circulated for comment and discussion. Please contact the author(s) before citing this paper in any publications. The views expressed in RPF Working Papers are solely those of the author(s) and do not necessarily represent the views of RPF or George Washington University.

Exhaustive Regression

An Exploration of Regression-Based Data Mining Techniques Using Super Computation

Antony Davies, Ph.D.
Associate Professor of Economics
Duquesne University
Pittsburgh, PA 15282

Research Fellow
The Mercatus Center
George Mason University
Arlington, VA 22201

antony@antolin-davies.com

Regression analysis is intended to be used when the researcher seeks to test a given hypothesis against a data set. Unfortunately, in many applications it is either not possible to specify a hypothesis, typically because the research is in a very early stage, or it is not desirable to form a hypothesis, typically because the number of potential explanatory variables is very large. In these cases, researchers have resorted either to overt data mining techniques such as stepwise regression, or covert data mining techniques such as running variations on regression models prior to running the final model (also known as “data peeking”). While data mining side-steps the need to form a hypothesis, it is highly susceptible to generating spurious results. This paper draws on the known properties of OLS estimators in the presence of omitted and extraneous variable models to propose a procedure for data mining that attempts to distinguish between parameter estimates that are significant due to an underlying structural relationship and those that are significant due to random chance.

1. Background

Regression analysis is designed to estimate the probability of observing a given data set given that a pre-determined hypothesis about the relationship between an outcome variable and a set of factors is assumed to be true. Unfortunately, in many applications it is not possible to specify a hypothesis. Two possible reasons why a researcher would want to perform analysis absent a hypothesis are:

Large Data Scope: Data size refers to the number of observations in a data set. Data scope refers to the number of potential explanatory factors ("candidate factors") in the data set. In the case of large data scope (e.g., economic and financial data sets), the number of candidate factors is so large that the cost of forming a tractable hypothesis is prohibitive.

Early Stage Analysis: In the case of early stage analysis (e.g., clinical data), there has not been enough observation to as yet for a hypothesis.

To date, stepwise regression (SR) has been one of the more widely used techniques for performing these "hypothesis-less" analyses. SR methods perform a "smart" sampling of regression models in an attempt to find a regression model that best fits the data. The procedure for "smartly" sampling is ad-hoc. Two typically used procedures are backward (wherein the first model includes all factors and factors are removed one at a time) and forward (wherein the first model includes no factors and factors are added one at a time). SR procedures contain three major flaws:

Sampling size: The number of regression models that can be constructed from a given data set can be incredibly large. For example, with only 30 candidate factors one could construct more than 1 billion regression models ($2^{30} - 1$). Stepwise procedures sample only a small number (typically less than 100) of the set of possible regression models. While stepwise methods can find models that fit the data reasonably well, as the number of factors rises, the probability of stepwise methods finding the best-fit model is virtually zero.

Fit criterion: In evaluating competing models, stepwise methods typically employ an F-statistic criterion. This criterion causes stepwise to methods to seek out the model that comes closest to explaining the data set. However, as the number of candidate factors increases, what also increases is the probability of a given factor adding explanatory power *simply by random chance*. Thus, the stepwise fit criterion cannot distinguish between factors that contribute explanatory power to the outcome variable because of an underlying relationship ("deterministic factors") and those that contribute by random chance only ("spurious factors").

Initial condition: Because SR only examines a small subset of the space of possible models and because the space of "fits" of the models (potentially) contains many local optima, the solution SR returns varies based on the starting point of the search. For example, for the same data set, SR backward and SR forward can yield different results. As the starting points for SR backward and SR forward are arbitrary, there are many other potential starting points each of which potentially yields a different result. For example, Figure 1 depicts the space of possible regression models that can be formed using *K* factors. Each block represents one regression model. The shade of the block indicates the "quality" of the model (e.g., goodness of fit). A stepwise procedure that starts at model A evaluates models in the vicinity of A, moves to the best model, then re-evaluates in the new vicinity. This continues until the procedure cannot find a better model in the vicinity. In this example, stepwise would move along the indicated path starting at point A. Were stepwise to start at point B, however, it would end up at a different "optimal" model. Out of the four starting points shown, only starting point D finds the best model.

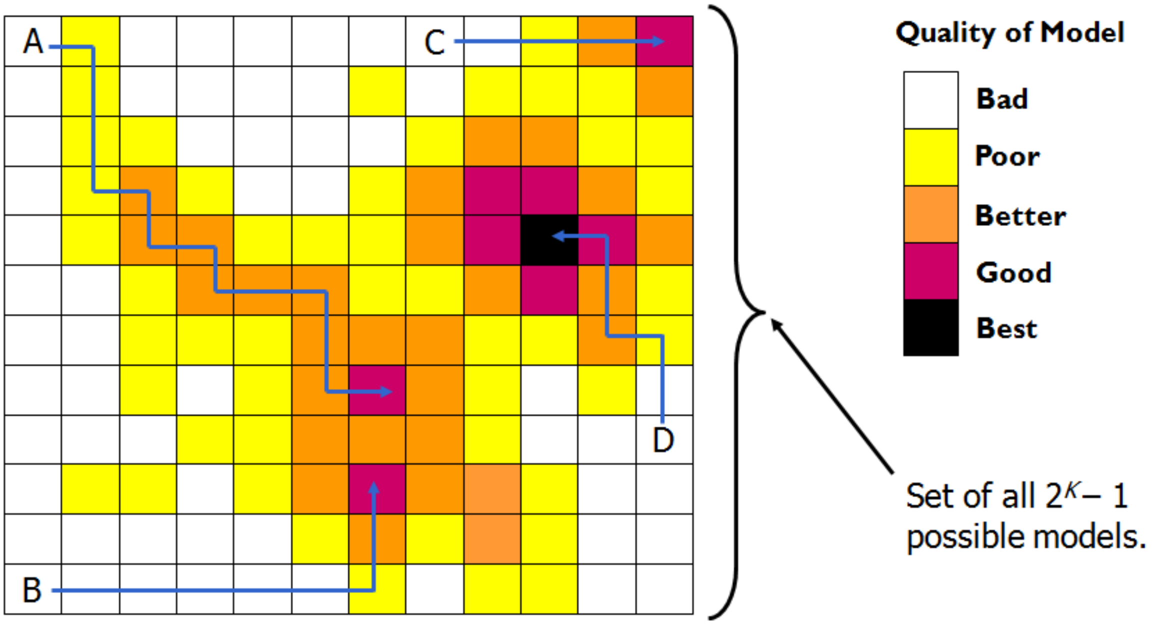

Figure 1. Results from stepwise procedures are dependent on the initial condition from which the search commences.

2. All Subsets Regression

All Subsets Regression (ASR) is a procedure intended to be used when a researcher wants to perform analysis in the absence of a hypothesis and wants to avoid the sampling size problem inherent in stepwise procedures. For K potential explanatory factors, ASR examines all $2^K - 1$ linear models that can be constructed. Until recently, ASR has been infeasible due to the massive computing power required. By employing grid-enabled super computation, it is now feasible to conduct ASR for moderately sized data sets. For example, as of today an average computer would require nearly 100 years of continuous work to perform ASR on a data set containing 40 candidate factors (see Figure 2), while a 10,000 node grid computer could complete the same analysis in less than a week. While ASR solves the sampling size and initial condition problems inherent in SR, ASR remains subject to the fit criterion problem. If anything, the fit criterion is more of a problem for ASR as the procedure examines a much larger space of models than does SR and therefore is more likely to find spurious factors.

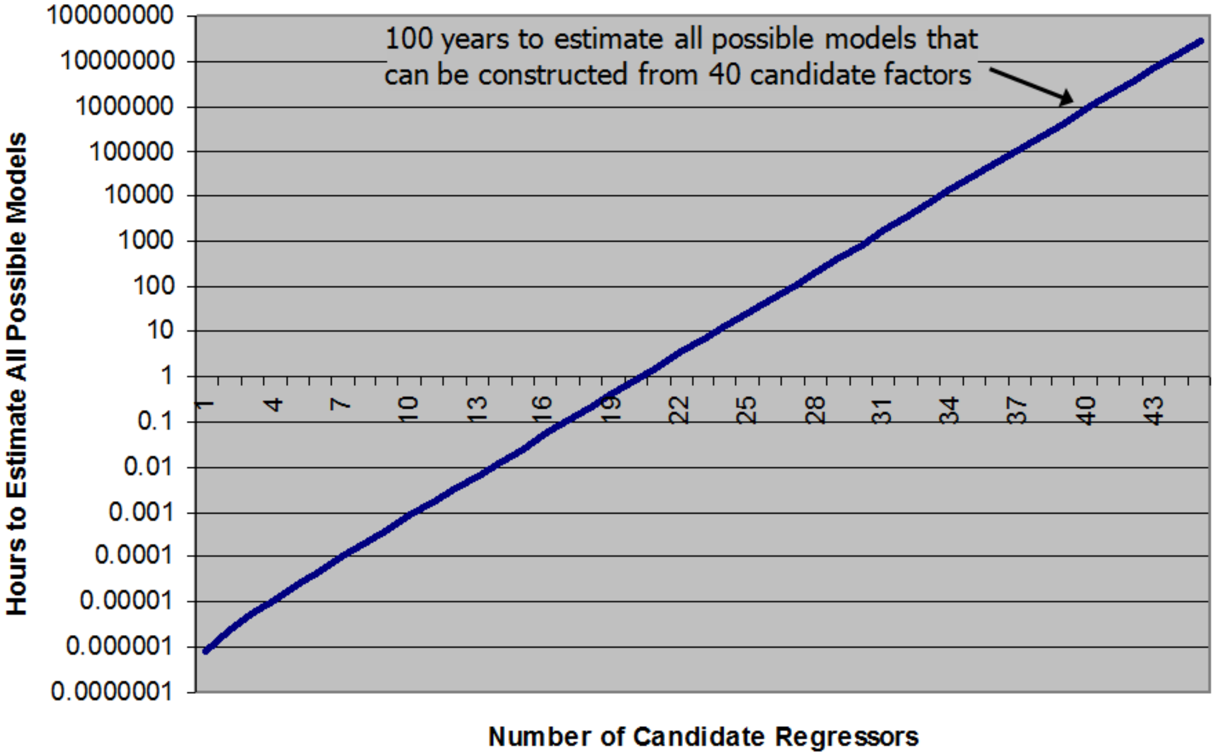

Figure 2. The time a single computer requires to perform ASR rises exponentially with the number of candidate factors.

3. Exhaustive Regression

Exhaustive Regression (ER) utilizes the ASR procedure, but attempts to identify spurious factors via a *cross-model chi-square statistic* that tests for stability in parameter estimates across models. The cross-model stability test compares parameter estimates for each factor across all 2^{K-1} models in which the factor appears. Factors whose parameter estimates yield significantly different results across models are identified as spurious. A given factor can exist in one of three types of models: omitted factor model, correctly specified model, and extraneous factor model. A correctly specified model contains all of the explanatory factors (i.e., the factors contribute to explaining the outcome variable because of some underlying relationship between the factor and the outcome) and no other factors. An omitted variable model includes at least one, but not all explanatory factors and (possibly) other factors. An extraneous variable model includes all explanatory factors and at least one other factor.

In the cases of the correctly specified model and the extraneous variable model, estimates of parameters associated with the factors ("slope coefficients") are unbiased.[1] In the case of the omitted variable model, however, estimates of the slope coefficients are biased and the direction of bias is a function of the covariances of the omitted factor with the outcome variable and the omitted factor with the factors included in the model. For example, consider a data set containing $k_1+k_2+k_3$ factors, for each of which there are N observations. Let the factors be arranged into three Nxk_j, j={1,2,3} matrices $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$, and let the sets of slope coefficients associated with each set of factors be the k_jx1 vectors $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, and $\boldsymbol{\beta}_3$, respectively. Let $\mathbf{Y}$ be an Nx1 vector of observations on the outcome variable. Suppose that, unknown to the researcher, the process that determines the outcome variable is

[1] Assuming, of course, that the remaining classical linear model assumptions hold.

$$\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u} \tag{1}$$

where **u** is an *N*x1 vector of independently and identically normally distributed errors. The three cases are attained when we apply the ordinary least squares (OLS) procedure to estimate the following models:

Omitted Variable Case: $\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{u}_\mathbf{O}$

Correctly Specified Case: $\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u}$

Extraneous Variable Case: $\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{X}_3\boldsymbol{\beta}_3 + \mathbf{u}_\mathbf{E}$

In the correctly specified case, the ordinary least squares regression procedure yields the slope estimate:

$$\begin{bmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{bmatrix} = \left([\mathbf{X}_1 \quad \mathbf{X}_2]' [\mathbf{X}_1 \quad \mathbf{X}_2] \right)^{-1} [\mathbf{X}_1 \quad \mathbf{X}_2]' \mathbf{Y} \tag{2}$$

where the square brackets indicate a partitioned matrix. Substituting the definition for **Y** from (1) into (2), it can be shown that the expected values of the slope estimates equal the true slope values:

$$\mathrm{E}\left(\begin{bmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{bmatrix} \right) = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix}$$

This is the unbiasedness condition. Similarly, in the extraneous variable case, we have:

$$\begin{bmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \\ \hat{\boldsymbol{\beta}}_3 \end{bmatrix} = \left([\mathbf{X}_1 \quad \mathbf{X}_2 \quad \mathbf{X}_3]' [\mathbf{X}_1 \quad \mathbf{X}_2 \quad \mathbf{X}_3] \right)^{-1} [\mathbf{X}_1 \quad \mathbf{X}_2 \quad \mathbf{X}_3]' \mathbf{Y} \tag{3}$$

Again, it can be shown that the expected values of the slope estimates equal the true slope values:

$$\mathrm{E}\left(\begin{bmatrix}\hat{\boldsymbol{\beta}}_1\\ \hat{\boldsymbol{\beta}}_2\\ \hat{\boldsymbol{\beta}}_3\end{bmatrix}\right)=\begin{bmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\\ \mathbf{0}\end{bmatrix}$$

By contrast, in the omitted variable case, we have

$$\hat{\boldsymbol{\beta}}_1=\left(\mathbf{X}_1'\mathbf{X}_1\right)^{\mathbf{-1}}\mathbf{X}_1'\mathbf{Y} \tag{4}$$

Substituting (1) into (4), we have

$$\hat{\boldsymbol{\beta}}_1=\left(\mathbf{X}_1'\mathbf{X}_1\right)^{\mathbf{-1}}\mathbf{X}_1'\left(\mathbf{X}_1\boldsymbol{\beta}_1+\mathbf{X}_2\boldsymbol{\beta}_2+\mathbf{u}\right)$$

and the expected values of the slope estimates are:

$$\mathrm{E}\left(\hat{\boldsymbol{\beta}}_1\right)=\boldsymbol{\beta}_1+\left(\mathbf{X}_1'\mathbf{X}_1\right)^{\mathbf{-1}}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2\neq\boldsymbol{\beta}_1 \tag{5}$$

From (5), we see that the expected value of the slope estimates in the omitted variable case are biased and that the direction of the bias depends on $\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2$.

We can construct a sequence of omitted variable cases as follows. Let $\{\mathbf{Z}_1, \mathbf{Z}_2, ..., \mathbf{Z}_{2^{k_2}-1}\}$ be the set of all (column-wise unique) subsets of $\mathbf{X}_2$.[2] Let $\tilde{\mathbf{X}}_\mathbf{i}$ be the set of regressors formed by merging $\mathbf{X}_1$ and $\mathbf{X}_2$ and then removing $\mathbf{Z}_\mathbf{i}$ so that, from the superset of regressors formed by merging $\mathbf{X}_1$ and $\mathbf{X}_2$, $\tilde{\mathbf{X}}_\mathbf{i}$ is the set of included regressors and $\mathbf{Z}_\mathbf{i}$ is the corresponding set of excluded regressors. Finally, let $\hat{\boldsymbol{\beta}}_1^i$ be the OLS estimate of $\boldsymbol{\beta}_1$ obtained by regressing $\mathbf{Y}$ on $\tilde{\mathbf{X}}_\mathbf{i}$. The expected value of the mean of the $\hat{\boldsymbol{\beta}}_1^i$ across all $2^{k_2}-1$ regression models is:

$$\mathrm{E}\left(\frac{1}{2^{k_2}-1}\sum_{i=1}^{2^{k_2}-1}\hat{\boldsymbol{\beta}}_1^i\right)=\boldsymbol{\beta}_1+\mathrm{E}\left(\frac{1}{2^{k_2}-1}\sum_{i=1}^{2^{k_2}-1}\left(\tilde{\mathbf{X}}_i'\tilde{\mathbf{X}}_i\right)^{\mathbf{-1}}\tilde{\mathbf{X}}_i'\mathbf{Z}_i\boldsymbol{\beta}_2^i\right) \tag{6}$$

[2] Since $\mathbf{Z}_\mathbf{i}$'s are subsets of $\mathbf{X}_2$, each $\mathbf{Z}_\mathbf{i}$ is Nxj where $j \le k_2$.

From (6), we see that the expected value of the mean of the $\hat{\boldsymbol{\beta}}_{\mathbf{1}}^{i}$ is $\boldsymbol{\beta}_{\mathbf{1}}$ when any of the following conditions are met:

1. For each set of included regressors, there is no covariance between the included regressors and the corresponding set of excluded regressors (i.e., $\tilde{\mathbf{X}}_{\mathbf{i}}^{'}\mathbf{Z}_{\mathbf{i}} = \mathbf{0}\ \forall\ \mathbf{i}$);
2. For some sets of included regressors, there is a non-zero covariance between the included regressors and the corresponding set of excluded regressors (i.e., $\tilde{\mathbf{X}}_{\mathbf{i}}^{'}\mathbf{Z}_{\mathbf{i}} \neq \mathbf{0}\ \forall\ \mathbf{i}$), but the expected value of the covariances is zero (i.e., $\mathrm{E}\left(\tilde{\mathbf{X}}_{\mathbf{i}}^{'}\mathbf{Z}_{\mathbf{i}}\right) = \mathbf{0}$);
3. For some sets of included regressors, there is a non-zero covariance between the included regressors and the corresponding set of excluded regressors (i.e., $\tilde{\mathbf{X}}_{\mathbf{i}}^{'}\mathbf{Z}_{\mathbf{i}} \neq \mathbf{0}\ \forall\ \mathbf{i}$), and the expected value of the covariances is non-zero (i.e., $\mathrm{E}\left(\tilde{\mathbf{X}}_{\mathbf{i}}^{'}\mathbf{Z}_{\mathbf{i}}\right) \neq \mathbf{0}$), but the expected covariance of the excluded regressors with the dependent variable is zero (i.e., $\mathrm{E}\left(\boldsymbol{\beta}_{\mathbf{2}}^{i}\right) = \mathbf{0}$);
4. None of the above holds, but the expected value of the product of (1) the covariances between the included and excluded regressors and (2) the covariance of the excluded regressors with the dependent variable is zero (i.e., $\mathrm{E}\left(\tilde{\mathbf{X}}_{\mathbf{i}}^{'}\mathbf{Z}_{\mathbf{i}}\boldsymbol{\beta}_{\mathbf{i}}\right) = \mathbf{0}$).

The ER procedure relies on the reasonable assumption that, as the data number of factors in the data set increases, conditions (2), (3), and (4) will hold asymptotically. If true, this enables us to construct the cross-model chi-square statistic.

4. The Cross-Model Chi-Square Statistic

Let us assume that the i^{th} (in a set of K) factor, x_i (where x_i is an Nx1 vector) has no structural relationship with an outcome variable y. Consider the equation:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u \tag{7}$$

Under the null hypothesis of $\beta_i = 0$, the square of the ratio of $\hat{\beta}_i$ to its standard error is distributed χ^2 with one degree of freedom. By adding factors to and removing factors (other than x_i) from (7), we can obtain other estimates of β_i. Let $\hat{\beta}_{ij}$ be the j^{th} such estimate of β_i. By looking at all combinations of factors from the superset of K factors, we can construct 2^{K-1} estimates of β_i. Under the null hypothesis that $\beta_i = 0$ and assuming that the $\hat{\beta}_{ij}$ are independent, we have:

$$\sum_{j=1}^{2^{K-1}} \left(\frac{\hat{\beta}_{ij}}{s_{\beta_i}} \right)^2 \sim \chi^2_{2^{K-1}} \tag{8}$$

From (8), we can construct c_i, the cross-model chi-square statistic for the factor x_i:

$$c_i = \frac{1}{2^{K-1}} \sum_{j=1}^{\frac{2^K - 1}{2}} \left(\frac{\hat{\beta}_{ij}}{s_{\beta_i}} \right)^2 \sim \chi^2_1 \tag{9}$$

Given the (typically) large number of degrees of freedom inherent in the ER procedure, it is worth noting that the measure in (8) is likely to be subject to Type II errors. In an attempt to compensate, we divide by the number of degrees of freedom to obtain c_i, a *relative chi-square statistic*, a measure that is less dependent on sample size. Carmines and McIver (1981) and Kline (1998) claim that one should conclude that the data represent a good fit to the hypothesis when the relative chi-square measure is less than 3. Ullman (2001) recommends using a chi-square less than 2.

Because the $\hat{\beta}_{ij}$ are obtained by exploring all combinations of factors from a single superset, one might expect the $\hat{\beta}_{ij}$ to be correlated (particularly when factors are positively correlated), and for the correlation to increase in the presence of stronger multicollinearity among the factors.

5. Monte-Carlo Tests of c_i

To test the ability of the cross-model chi-square statistic to identify factors that might show statistically significant slope coefficients simply by random chance, consider an outcome variable, *Y*, that is determined by three factors as follows

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + u \tag{10}$$

where *u* is an error term satisfying the requirements for the classical linear regression model. Let us randomly generate additional factors X_4 through X_{15}, run the ER procedure and calculate c_i for each of the factors. The following figures show the results of the ER runs. The first set of bars show results for ER runs applied to the superset of factors X_1 through X_4. The results are derived as follows:

1. Generate 500 observations for X_1, randomly selecting observations from the uniform distribution.
2. Generate 500 observations each for X_2 through X_{15} such that $X_i = \gamma_i X_1 + v_i$ where the γ_i are randomly selected from the standard normal distribution and distributed, and v_i are normally distributed with mean zero and variance 0.1. This step creates varying multicollinearity among the factors.
3. Generate *Y* according to (10) where $\alpha = \beta_1 = \beta_2 = \beta_3 = 1$, and *u* is normally distributed with a variance of 1.

4. Run all $2^K - 1 = 15$ regression models to obtain the 2^{K-1} estimates for each β:

   $\hat{\beta}_{1,1}, \ldots, \hat{\beta}_{1,8}, \hat{\beta}_{2,1}, \ldots, \hat{\beta}_{2,8}, \hat{\beta}_{3,1}, \ldots, \hat{\beta}_{3,8}, \hat{\beta}_{4,1}, \ldots, \hat{\beta}_{4,8}$.

5. Calculate c_1, c_2, c_3, and c_4 according to (9).

6. Repeat steps 1 through 5 three-thousand times.

7. Repeat steps 1 through 6, each time increasing the variance of u by 1 until the variance reaches 20.[3]

At the completion of the algorithm, there will be 60,000 measures for each of c_1, c_2, c_3, and c_4 based on (60,000) (15) = 900,000 separate regressions. We then repeat the procedure using a superset of factors X_1 through X_5, then X_1 through X_6, etc. up to X_1 through X_{15}.[4]

Step 2 introduces random multicollinearity among the factors. On average, the multicollinearity of factors with X_1 follows the pattern shown in Figure 3. Approximately half of the correlations with X_1 are positive and half are negative. While the correlations are constructed between X_1 and the other factors, this will also result in the other factors being pair-wise correlated though to a lesser extent (on average) than they are correlated with X_1.

[3] This results in an average R^2 for the estimate of equation (10) of approximately 0.2.

[4] This final pass requires the estimation of 2 billion separate regressions. The entire Monte-Carlo run requires computation equivalent to almost 2,000 CPU-hours.

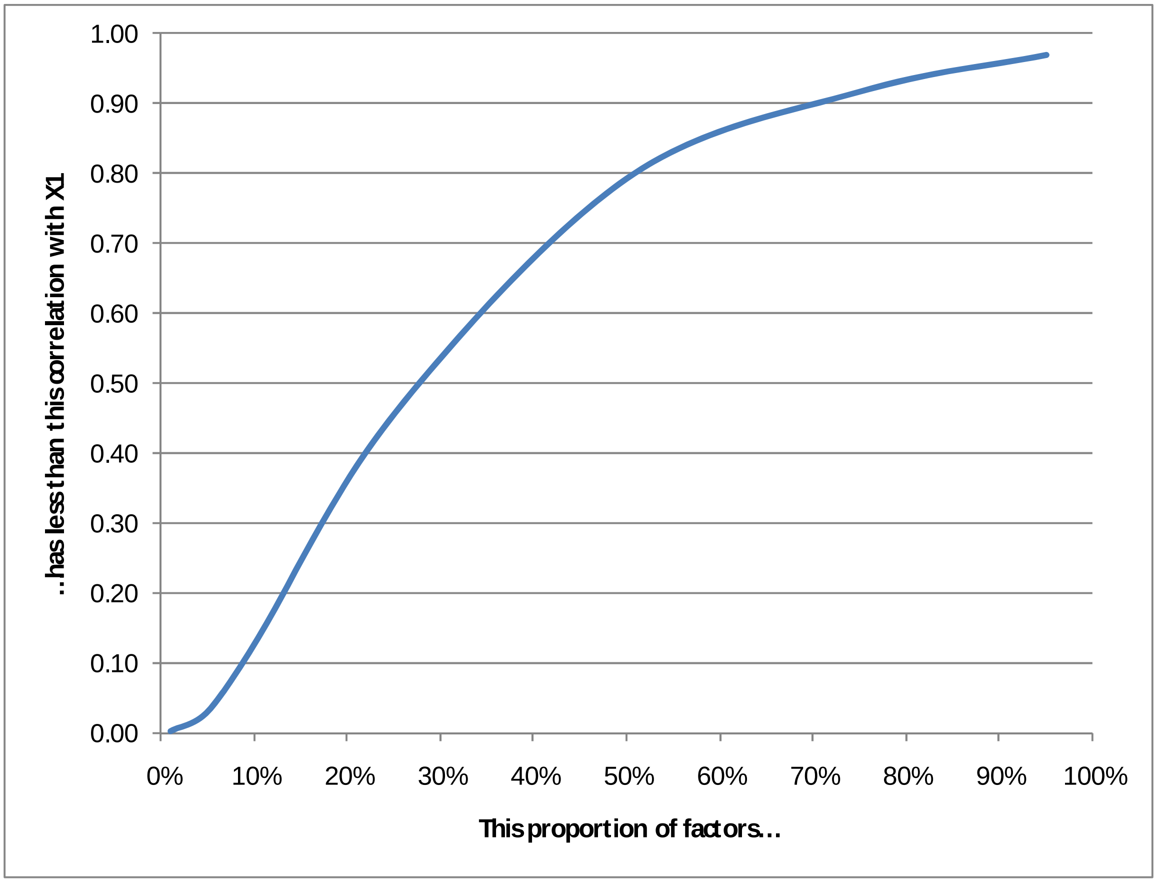

Figure 3. Pattern of Squared Correlations of Factors X_2 through X_K with X_1

Figure 4 shows the results of the Monte-Carlo runs in which the critical value for the c_i is set to 3. For example, when there are four factors in the data set, c_1, c_2, and c_3 pass the significance test slightly over 50% of the time versus 20% for c_4. In other words, for data sets in which three out of four factors determine the outcome variable (and a critical value of 3), the ER procedure will identify the three determining factors 50% of the time and identify the non-determining factor 20% of the time. As the number of factors in the data set increases, the ER procedure better discriminates between the factors that determine the outcome variable and those that might appear significant by random chance alone. The last set of bars in Figure 4 shows the results for data sets in which three out of fifteen factors determine the outcome variable. Here, the ER procedure identifies the determining factors approximately 85% of the time and (erroneously) identifies non-determining factors only slightly more than 10% of the time.

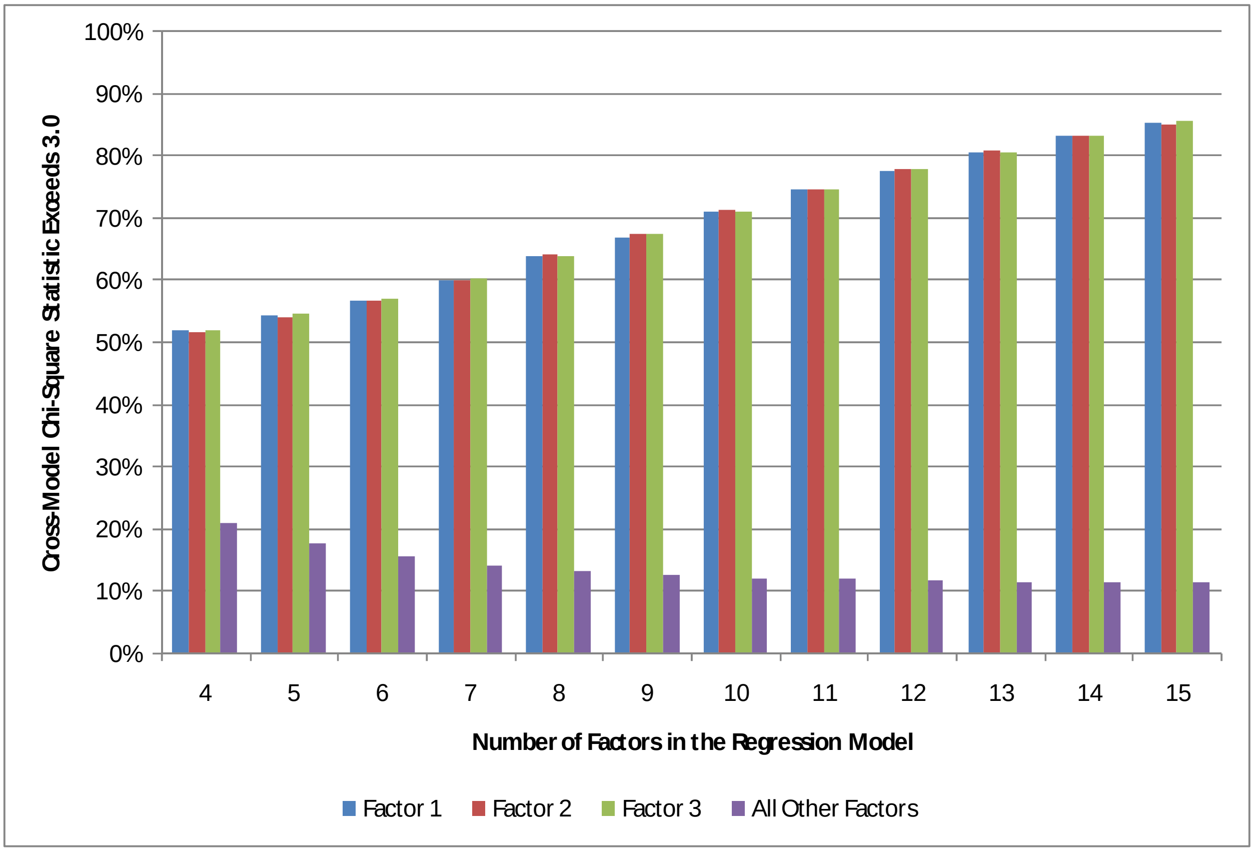

Figure 4. Monte-Carlo Tests of ER Procedure Using Supersets of Data from 4 Through 15 Factors (critical value = 3)

These results are based on the somewhat arbitrary selection of 3 for the relative chi-square critical value. Reducing the critical value to 2 produces the results shown in Figure 5. As expected, reducing the critical value causes the incidence of false positives (where “positive” means “identification of a determining factor”) to approximately 25%, but the incidence of false negatives falls to below 10%. Figure 6 and Figure 7, where the critical value is set to 1.5 and 1, respectively, are shown for comparison. As expected, the marginal gains in the reduction in false negatives (versus Figure 5) appear to be outweighed by the increase in the rate of false positives.

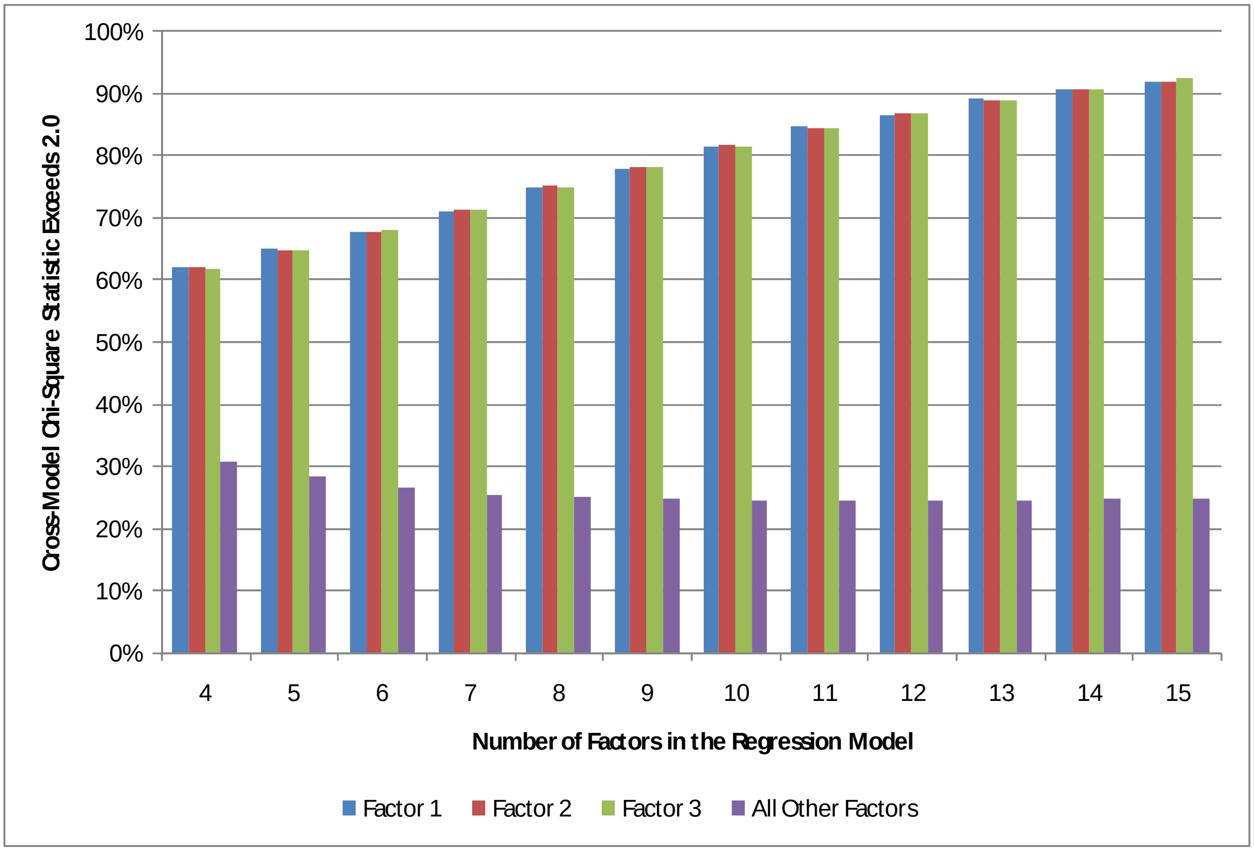

Figure 5. Monte-Carlo Tests of ER Procedure Using Supersets of Data from 4 Through 15 Factors (critical value = 2)

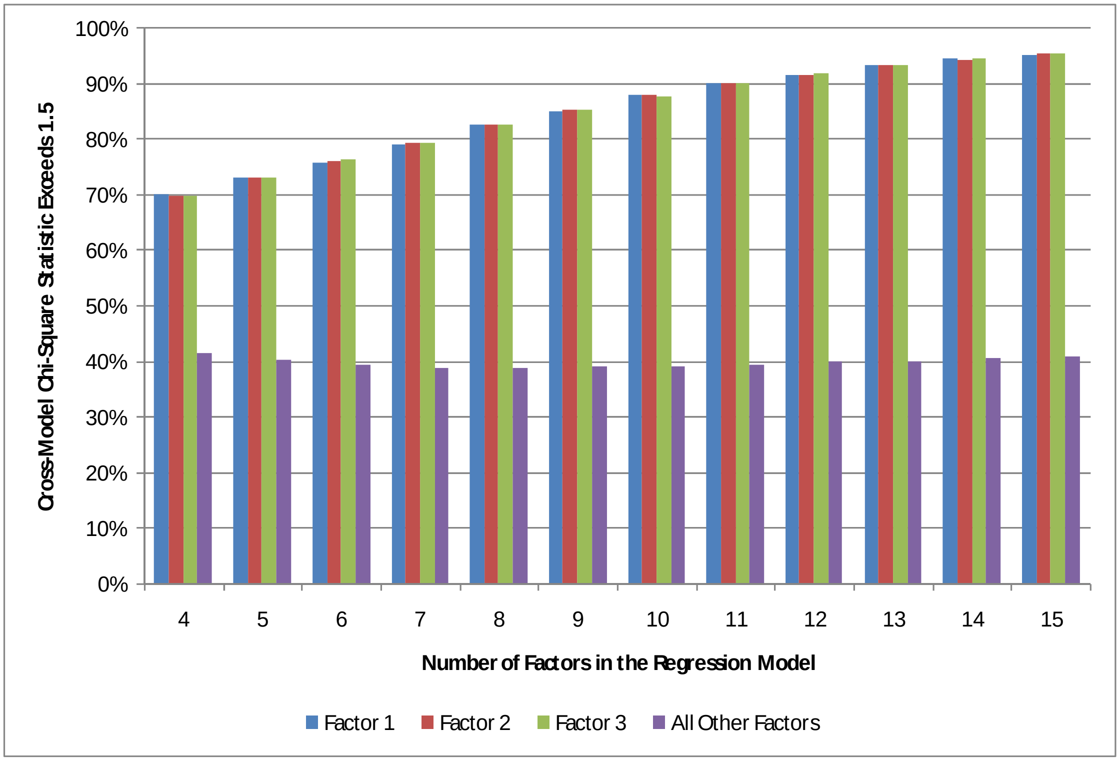

Figure 6. Monte-Carlo Tests of ER Procedure Using Supersets of Data from 4 Through 15 Factors (critical value = 1.5)

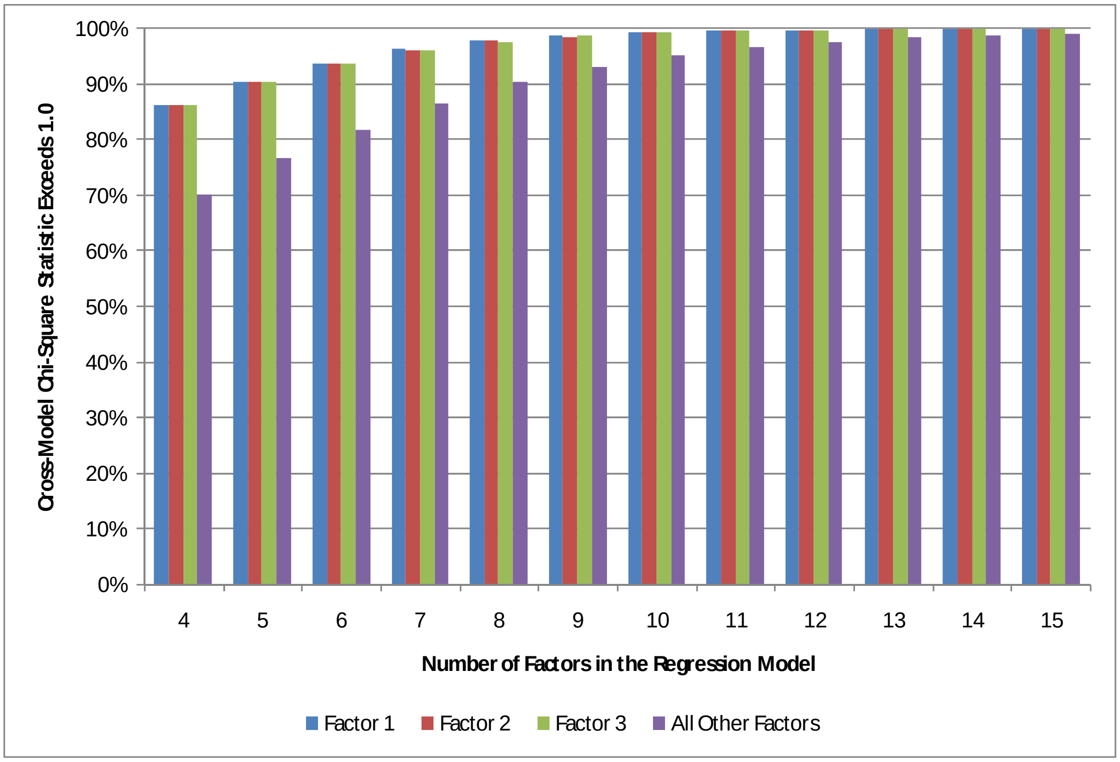

Figure 7. Monte-Carlo Tests of ER Procedure Using Supersets of Data from 4 Through 15 Factors (critical value = 1)

To measure the effect of the deterministic model's goodness of fit on the cross-model chi-square statistic, we can arrange the Monte-Carlo results according to the variance of the error term in (10). The algorithm varies the error term from 1 to 20 in increments of 1. Figure 8 shows the proportion of times that c_i passes the significance threshold of 3 for factors X_1, X_2, and X_3 (combined) for various numbers of factors in the data set and for various levels of error variance. An error variance of 1 corresponds to an R^2 (for the estimate of equation (10)) of approximately 0.93 while an error variance of 20 corresponds to an R^2 of approximately 0.03.

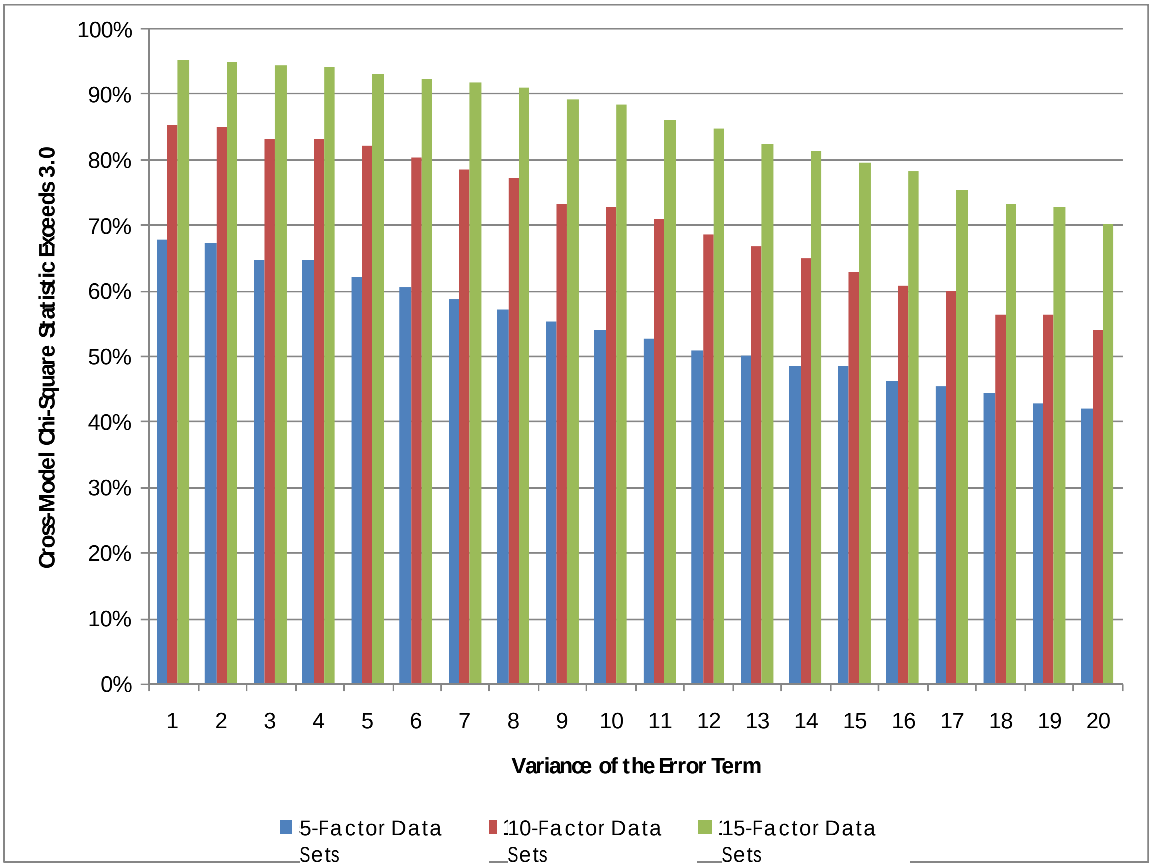

Figure 8. Monte-Carlo Tests of ER Procedure for Factors X_1 through X_3 (combined)

As expected, as the error variance rises in a 15-factor data set, the probability of a false negative rises from approximately 5% (when var(u) = 1) to 30% (when var(u) = 20). Results are markedly worse for data sets with fewer factors. Figure 9 shows results for factors X_4 through

X_{15}, combined. Here, we see that the probability of a false positive rises from under 5% (when var(u) = 1) to almost 20% (when var(u) = 20) for 15-factor data sets. Employing a critical value of 2.0 yields the results in Figure 10 and Figure 11. Comparing Figure 8 and Figure 9 with Figure 10 and Figure 11, we see that employing the critical value of 3 versus 2 cuts in half (approximately) the likelihoods of false positives and negatives.

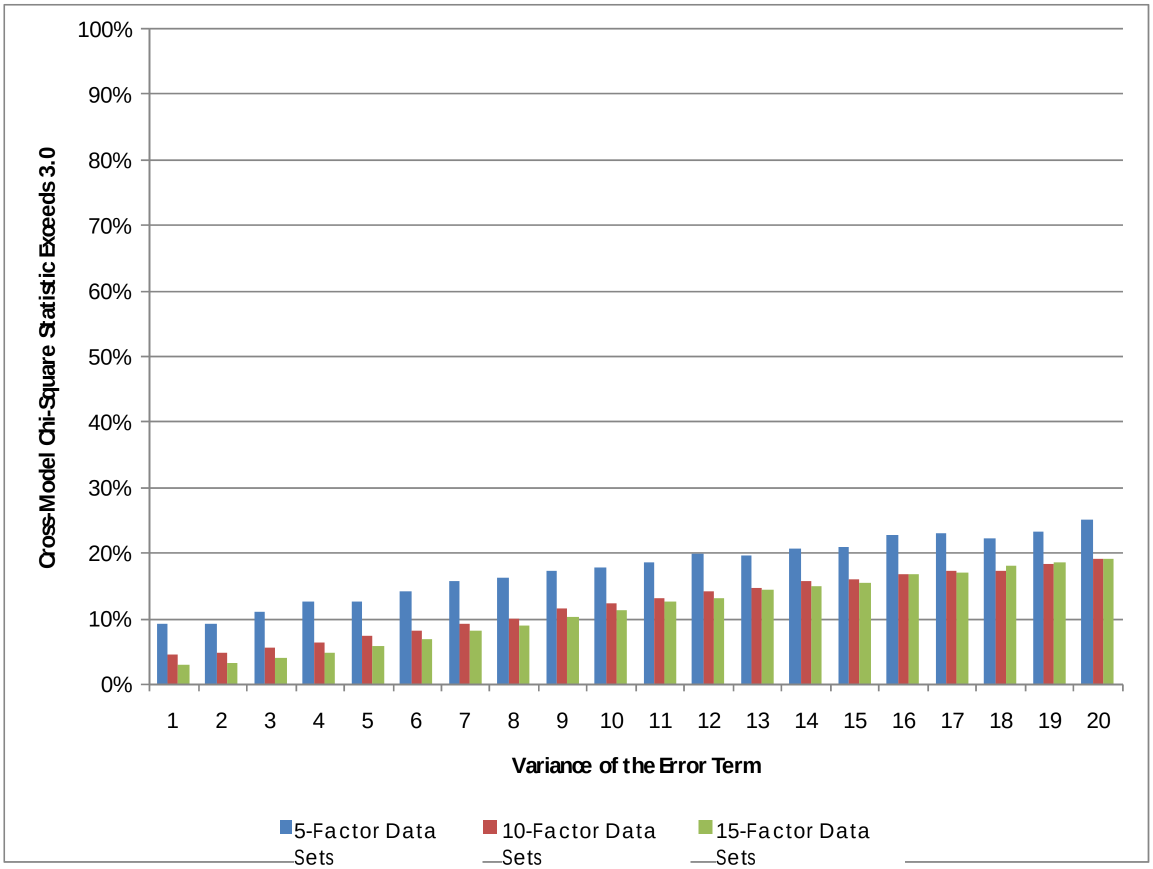

Figure 9. Monte-Carlo Tests of ER Procedure for Factors X_4 through X_5, X_{10}, and X_{15} (combined)

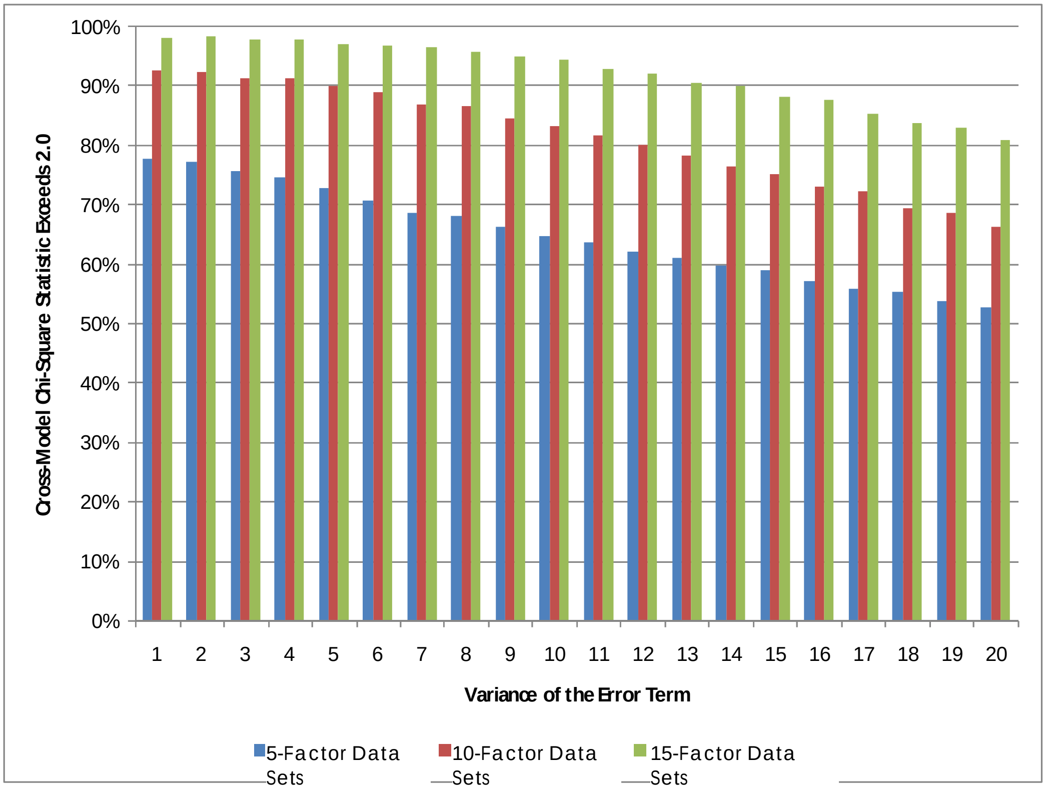

Figure 10. Monte-Carlo Tests of ER Procedure for Factors X_1 through X_3 (combined)

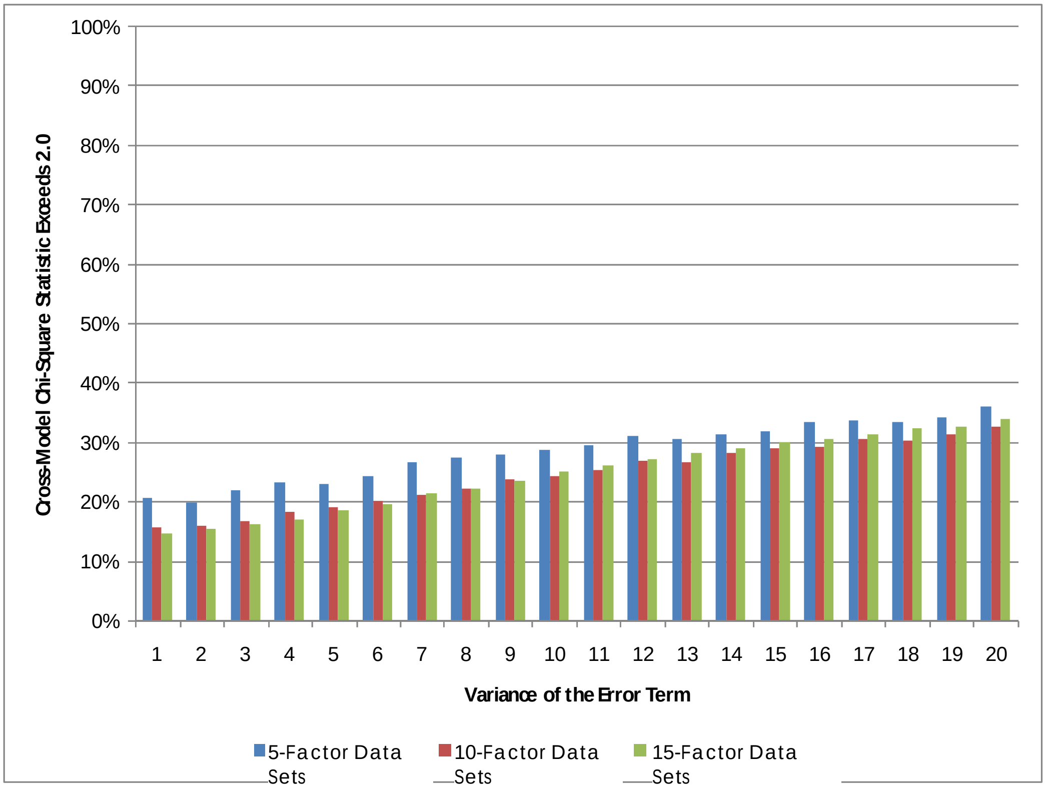

Figure 11. Monte-Carlo Tests of ER Procedure for Factors X_4 through X_5, $X_{10,}$ and X_{15} (combined)

6. Estimated ER (EER)

Even with the application of super computation, large data sets can make ASR and ER impractical. For example, it would take a full year for a top-of-the-line 100,000 node cluster computer to perform ASR/ER on a 50-factor data set. When one considers data mining just simple non-linear transformations of factors (inverse, square, logarithm), the 50-factor data set becomes a 150-factor data set. If one then considers data mining two-factor cross-products (e.g., X_1X_2, X_1X_3, etc.), the 150-factor data set balloons to an 11,175-factor data set. This suggests that super computation alone isn't enough to make ER a universal tool for data mining. One possible approach to using ER with large data sets is to employ *estimated* ER. Estimated ER (EER) randomly selects J (out of a possible $2^K - 1$) models to estimate. Note that EER does not select the *factors* randomly, but selects the *models* randomly. Selecting factors randomly biases the model selection toward models with a total of $K/2$ factors. Selecting models randomly gives each of the $2^K - 1$ models an equal probability of being chosen.

Figure 12 and Figure 13 show the results of EER for various numbers of randomly selected models for 5-, 10-, and 15-Factor data sets. These tests were performed as follows:

1. Generate 500 observations for each of the factors X_1, randomly selecting observations from the uniform distribution.
2. Generate 500 observations each for X_2 through X_K (where K is 5, 10, or 15) such that $X_i = \gamma_i X_1 + v_i$ where the γ_i are randomly selected from the standard normal distribution and distributed, and v_i are normally distributed with mean zero and variance 0.1. This step creates varying multicollinearity among the factors.
3. Generate Y according to (10) where $\alpha = \beta_1 = \beta_2 = \beta_3 = 1$, and u is normally distributed with a variance of 1.

4. Randomly select J models out of the possible $2^K - 1$ regression models to obtain J estimates for each β.
5. Calculate $c_1, c_2, c_3, \ldots, c_K$ according to (9).
6. Repeat steps 1 through 5 three-hundred times.
7. Calculate the percentage of times that the cross-model test statistics for each β exceed the critical value.
8. Repeat steps 1 through 7, for J running from 1 to 100.

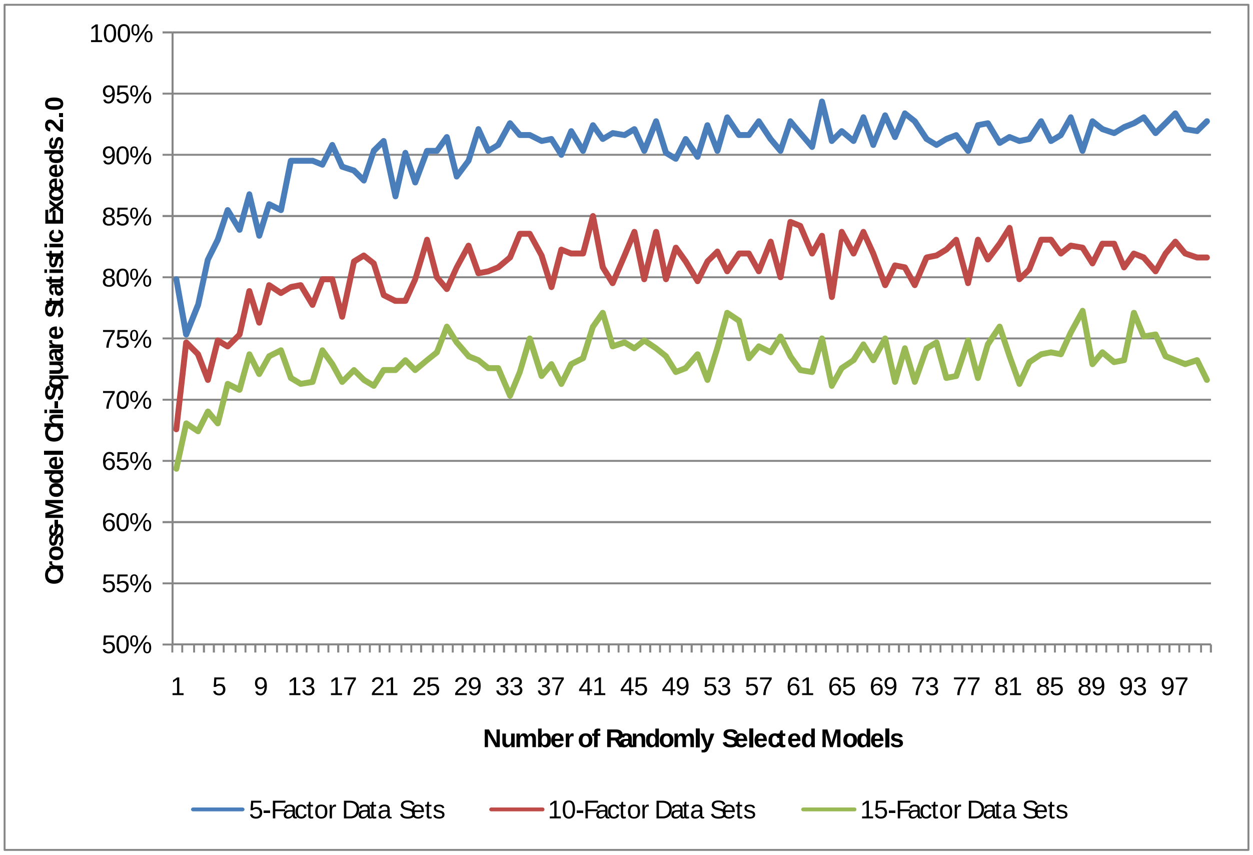

Figure 12. Monte-Carlo Tests of EER Procedure for Factors X_1 through X_3 (combined)

Evidence suggests that smaller data sets are more sensitive to a small number of randomly selected models. For 5-factor data sets, the rate of false negatives (Figure 12) does not stop falling significantly until approximately $J = 30$, and the rate of false positives (Figure 13) does not stop rising significantly until approximately $J = 45$. The rates of false negatives (Figure 12) and false positives (Figure 13) for 10-factor and 15-factor data sets appear to settle for lesser values of J.

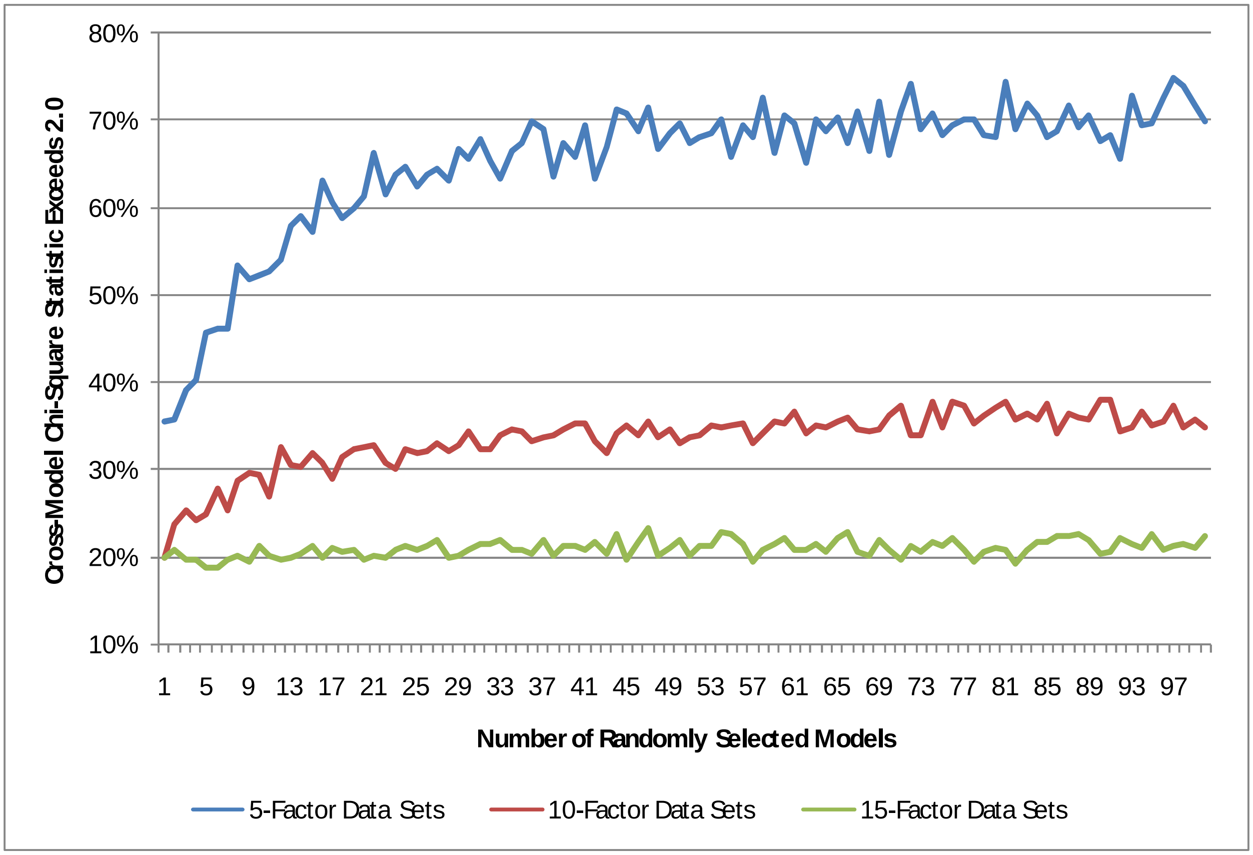

Figure 13. Monte-Carlo Tests of EER Procedure for Factors X_4 through X_5, X_{10}, and X_{15} (combined)

The number of factors in the data set that determine the outcome variable has a greater effect on the number of randomly selected models required in EER. Figure 14 compares results for 5-factor data sets when the outcome variable is a function of only one factor versus being a function of three factors. The vertical axis measures the cross-model chi-squared statistic for the indicated number of randomly selected models divided by the average cross-model chi-squared statistic over all 100 runs. Figure 14 shows that, for 5-factor data sets, as the number of randomly selected models increases, the cross-model chi-squared statistic approaches its mean value for the 100 runs faster when the outcome variable is a function of three factors versus being a function of only one factor. Figure 15 shows similar results for 10-factor data sets.

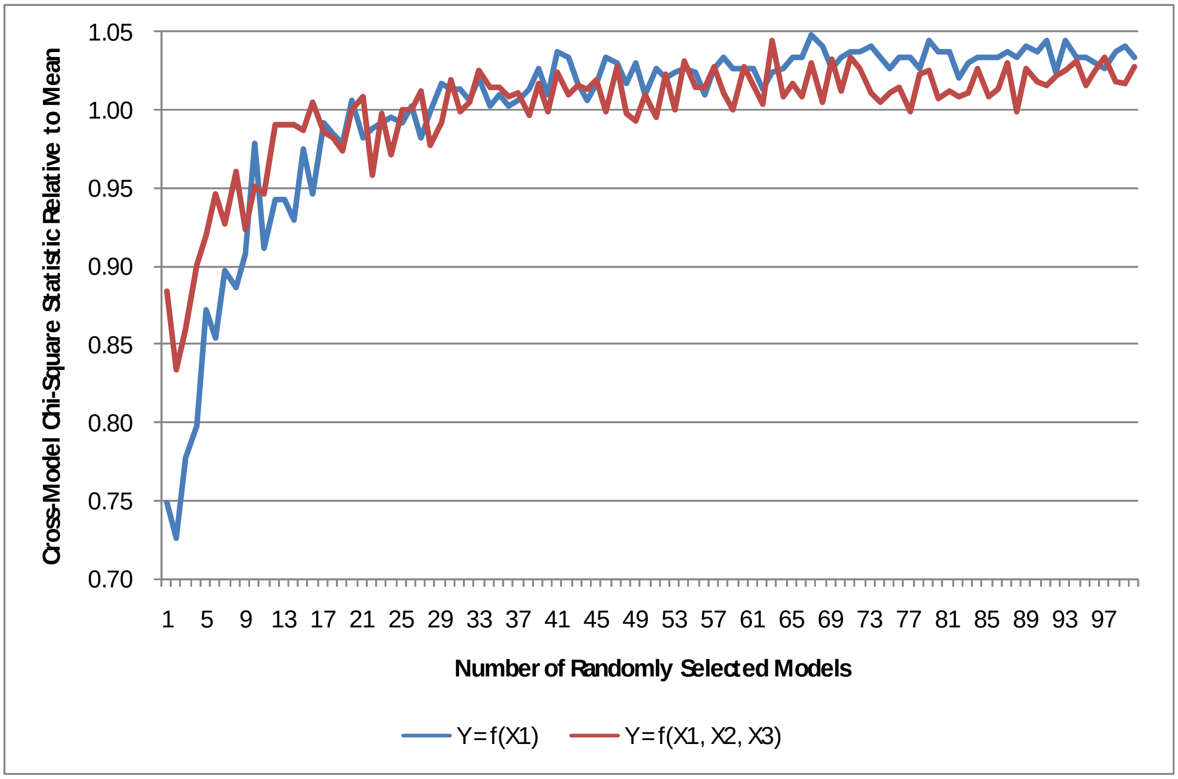

Figure 14. EER Procedure for Factor X_1 and X_1 through X_3 (combined) for 5-Factor Data Sets when Outcome Variable is a Function of One vs. Three Factors

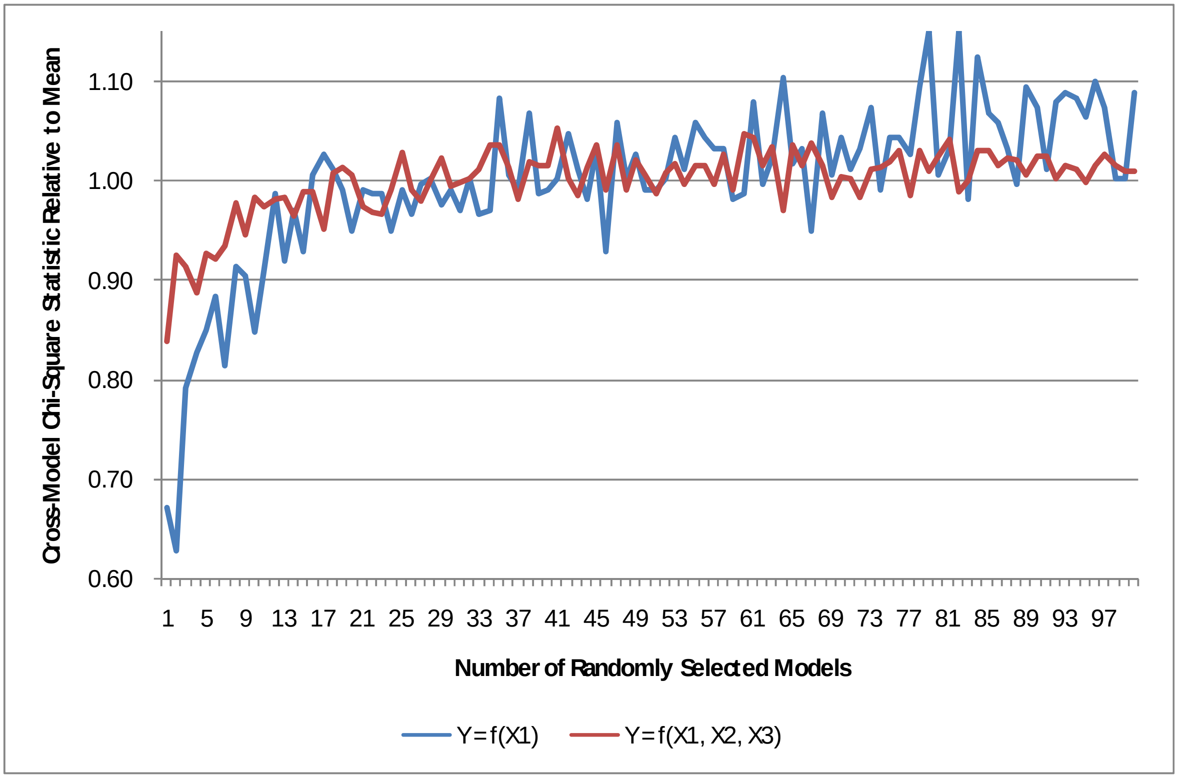

Figure 15. EER Procedure for Factor X_1 and X_1 through X_3 (combined) for 10-Factor Data Sets when Outcome Variable is a Function of One vs. Three Factors

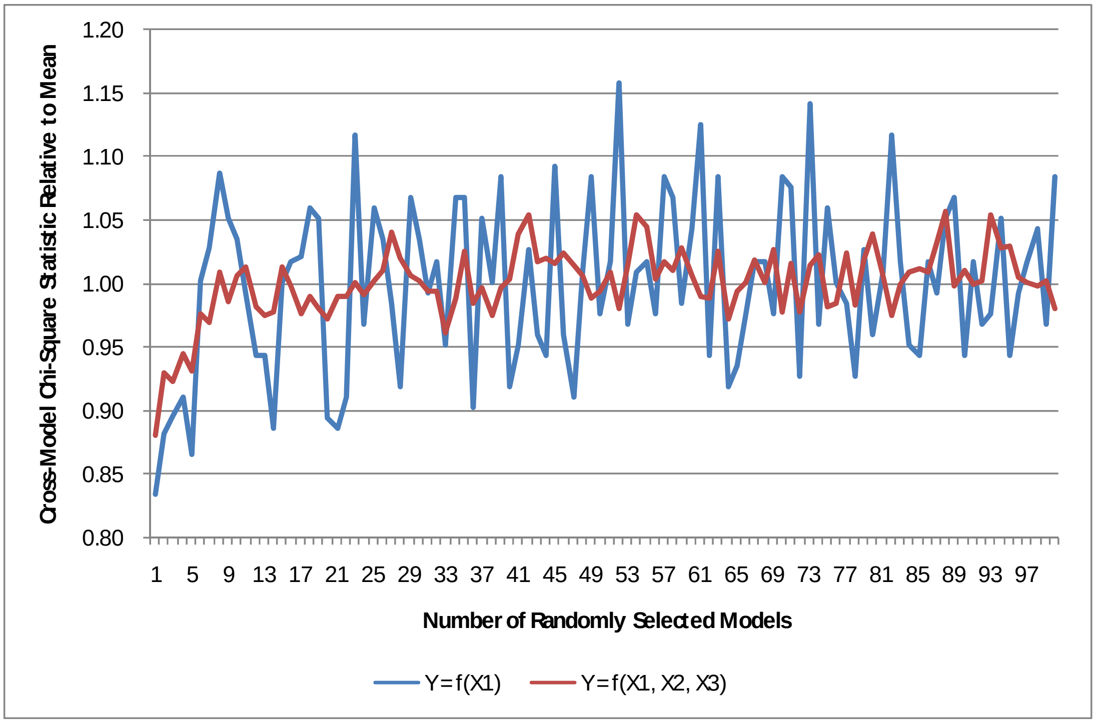

Figure 16. EER Procedure for Factor X_1 and X_1 through X_3 (combined) for 15-Factor Data Sets when Outcome Variable is a Function of One vs. Three Factors

Results in Figure 16 are less compelling, but not contradictory. A second result, common to the large data sets (Figure 15 and Figure 16), is that the variation in the cross-model chi-squared estimates is less when the outcome variance is a function of three versus one factor. Figure 17, Figure 18, and Figure 19 show corresponding results for factors that do not determine the outcome variable. These results suggest that EER may be an adequate procedure for estimating ER for larger data sets.

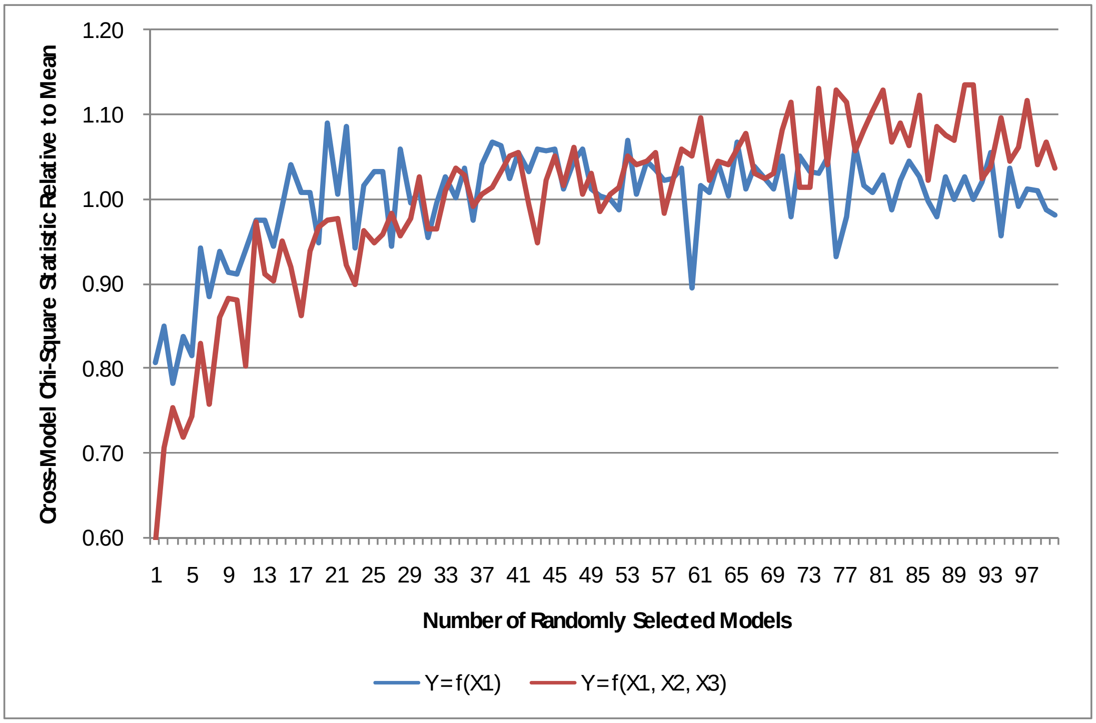

Figure 17. EER Procedure for Factors X_1 through X_{15} (combined) and X_4 through X_{15} (combined) for 5-Factor Data Sets when Outcome Variable is a Function of One vs. Three Factors

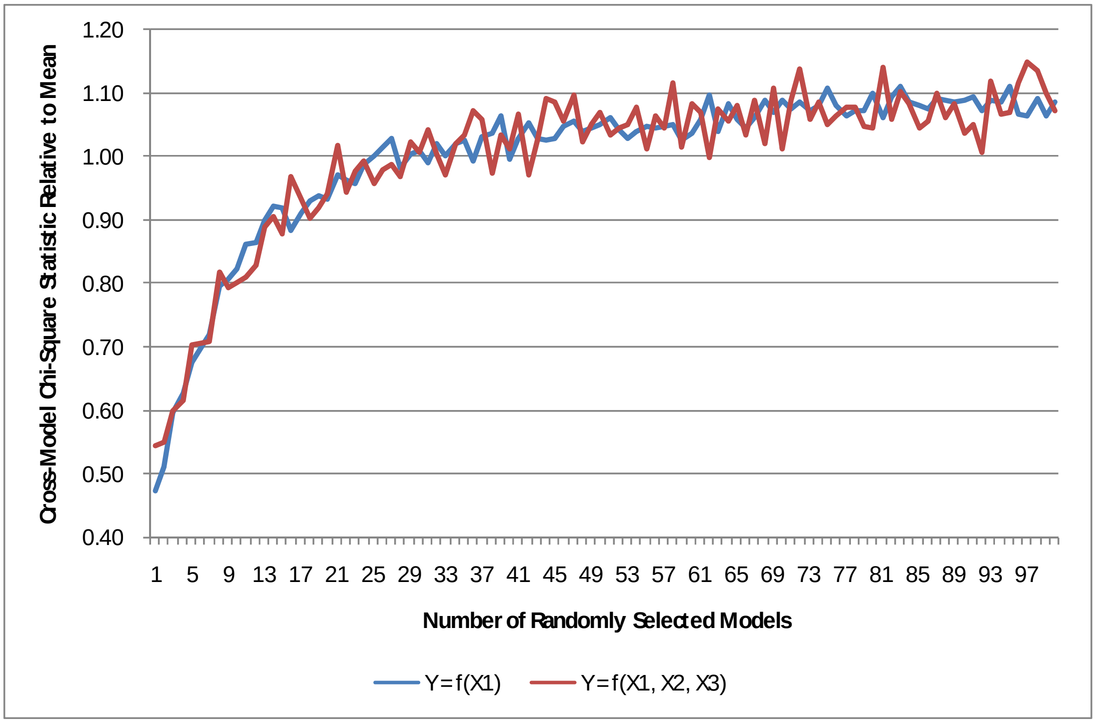

Figure 18. EER Procedure for Factors X_1 through X_{15} (combined) and X_4 through X_{15} (combined) for 10-Factor Data Sets when Outcome Variable is a Function of One vs. Three Factors

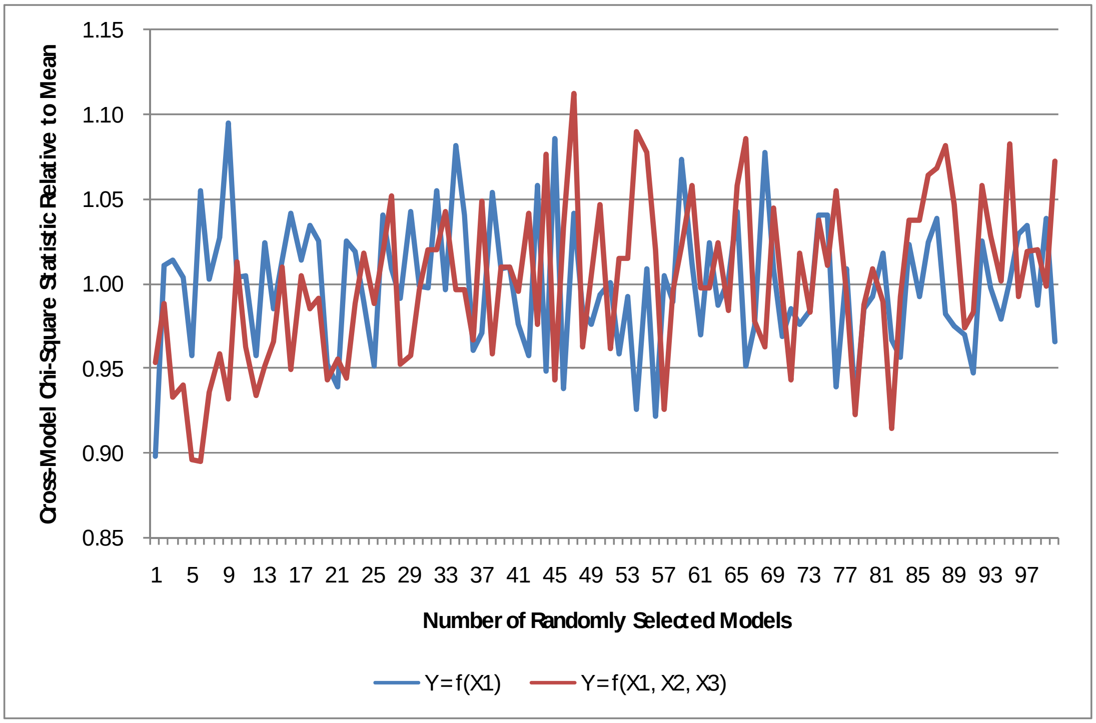

Figure 19. EER Procedure for Factors X_1 through X_{15} (combined) and X_4 through X_{15} (combined) for 15-Factor Data Sets when Outcome Variable is a Function of One vs. Three Factors

7. Comparison to Stepwise and k-Fold Holdout

Kuk (1984) demonstrated that stepwise procedures are inferior to all subsets procedures in data mining proportional hazard models. Logically, the same argument applies to data mining regression models. As stepwise examines only a subset of models and ASR examines all models, the best that stepwise can do is to match ASR. Because stepwise smartly samples on the basis of marginal changes to a fit function, multicollinearity among the factors can cause stepwise to return a solution that is a local, but not global, optimum. What is of interest is a comparison of stepwise to EER because EER, like stepwise, samples the space of possible regression models.

K-fold holdout is less an alternative data mining method than it is an alternative objective. Data mining methods typically have the objective of finding the model that best fits the data (typically measured by improvements to the F-statistic). K-fold holdout offers the alternative objective of maximizing the model's ability to predict observations that were not included in the model estimation (i.e., "held out" observations). The selection of which observations to hold out varies depending on the data set. For example, in the case of time series data, it makes more sense to hold out observations at the end of the data set.

The following tests use the same data set and apply EER, "estimated" all subets regression (EASR) where we conduct a sampling of models rather than examine all possible models, and stepwise. The procedure is as follows:

1. Generate 500 observations for X_1, randomly selecting observations from the uniform distribution.
2. Generate 500 observations each for X_2 through X_{15} such that $X_i = \gamma_i X_1 + v_i$ where the γ_i are randomly selected from the standard normal distribution and distributed, and v_i are

normally distributed with mean zero and variance 0.1. This step creates varying multicollinearity among the factors.

3. Generate Y according to (10) where $\alpha = \beta_1 = \beta_2 = \beta_3 = 1$, and u is normally distributed with a variance of 1.
4. Perform EER and EASR k-fold holdout:
    a. Randomly select 500 models out of the possible $2^{30} - 1$ regression models to obtain J estimates for each β.
    b. Evaluate the k-fold holdout criterion:
    c. Calculate $c_1, c_2, c_3, \ldots, c_{30}$ according to (9).
    d. Mark factors for which $c_i > 2$ as being "selected" by EER.
5. Evaluate the EER models using the k-fold holdout criterion:
    a. For each randomly selected model in step 4a, randomly select 50 observations to exclude.
    b. Estimate the model using the remaining 450 observations.
    c. Use the estimated model to predict the 50 excluded observations.
    d. Calculate the MSE (mean squared error) where

$$MSE = \frac{1}{50} \sum_{excluded\ observations} \left(observation - predicted\ observation \right)^2$$

    e. Over the 500 models randomly selected by EER, identify the one for which the MSE is least. Mark the factors that produce that model as being "selected" by the k-fold holdout criterion.
6. Peform backward stepwise:
    a. Let M be the set of included factors, and N be the set of excluded factors such that $|M| = m,\ |N| = n,$ and $m + n = 30.$

b. Estimate the model $Y = \alpha + \sum_{X_i \in M} \beta_i X_i + u$ and calculate the estimated model's adjusted multiple correlation coefficient, $\overline{R}_0^2$.

c. For each factor X_i, $i = 1,\ldots, m$, move the factor from set M to set N, estimate the model $Y = \alpha + \sum_{X_i \in M} \beta_i X_i + u$, calculate the estimated model's adjusted multiple correlation coefficient, $\overline{R}_i^2$, and then return the factor X_j to M from N. This will result in the set of measures $\overline{R}_1^2,\ldots,\overline{R}_m^2$.

d. Let $\overline{R}_L^2 = \min\left(\overline{R}_0^2,\ldots,\overline{R}_m^2\right)$. Identify the factor whose removal resulted in the measure $\overline{R}_L^2$. Move that factor from set M to set N.

e. Given the new set M, for each factor X_i, $i = 1,\ldots, m$, remove the factor from set M, estimate the model $Y = \alpha + \sum_{X_i \in M} \beta_i X_i + u$, calculate the estimated model's adjusted multiple correlation coefficient, $\overline{R}_i^2$, and then return the factor X_j to M from N. This will result in the set of measures $\overline{R}_1^2,\ldots,\overline{R}_m^2$.

f. For each factor X_i, $i = 1,\ldots, n$, move the factor from N to M, estimate the model $Y = \alpha + \sum_{X_i \in M} \beta_i X_i + u$, calculate the estimated model's adjusted multiple correlation coefficient, $\overline{R}_i^2$, and then return the factor X_i to M from N. This will result in the set of measures $\overline{R}_{m+1}^2,\ldots,\overline{R}_{m+n}^2$.

g. Let $\overline{R}_L^2 = \min\left(\overline{R}_1^2,\ldots,\overline{R}_{m+n}^2\right)$ and $\overline{R}_H^2 = \max\left(\overline{R}_1^2,\ldots,\overline{R}_{m+n}^2\right)$. Identify the factor whose removal resulted in the measure $\overline{R}_L^2$. Move that factor from set M to set N. If $\overline{R}_H^2$ was attained using a factor from N, move that factor from set N to set M.

   h. Repeat steps e through g until there is no further improvement in $\overline{R}_H^2$.

   i. Mark the factors in the model attained in step 5h as being "selected" by stepwise.

7. Repeat steps 1 through 6 six-hundred times.
8. Calculate the percentage of times that each factor is selected by EER, k-fold, and stepwise.

Figure 20 shows the results of this comparison. For 30 factors, where the outcome variable is determined by the first three factors, stepwise correctly identified the first factor slightly less frequently than did EER (43% of the time versus 48% of the time). Stepwise correctly identified the second and third factors slightly more frequently than did EER (87% and 92% of the time versus 81% and 85% of the time). The k-fold criterion when applied to EASR identified the first three factors 53%, 63%, and 63% of the time, respectively. In this, the false positive error rate was comparable for EER and stepwise, and significantly worse for k-fold. EER erroneously identified the remaining factors as being significant, on average, 17% of the time versus 34% of the time for stepwise and 49% of the time for k-fold. This suggests that EER may be more powerful as a tool for eliminating factors from consideration.

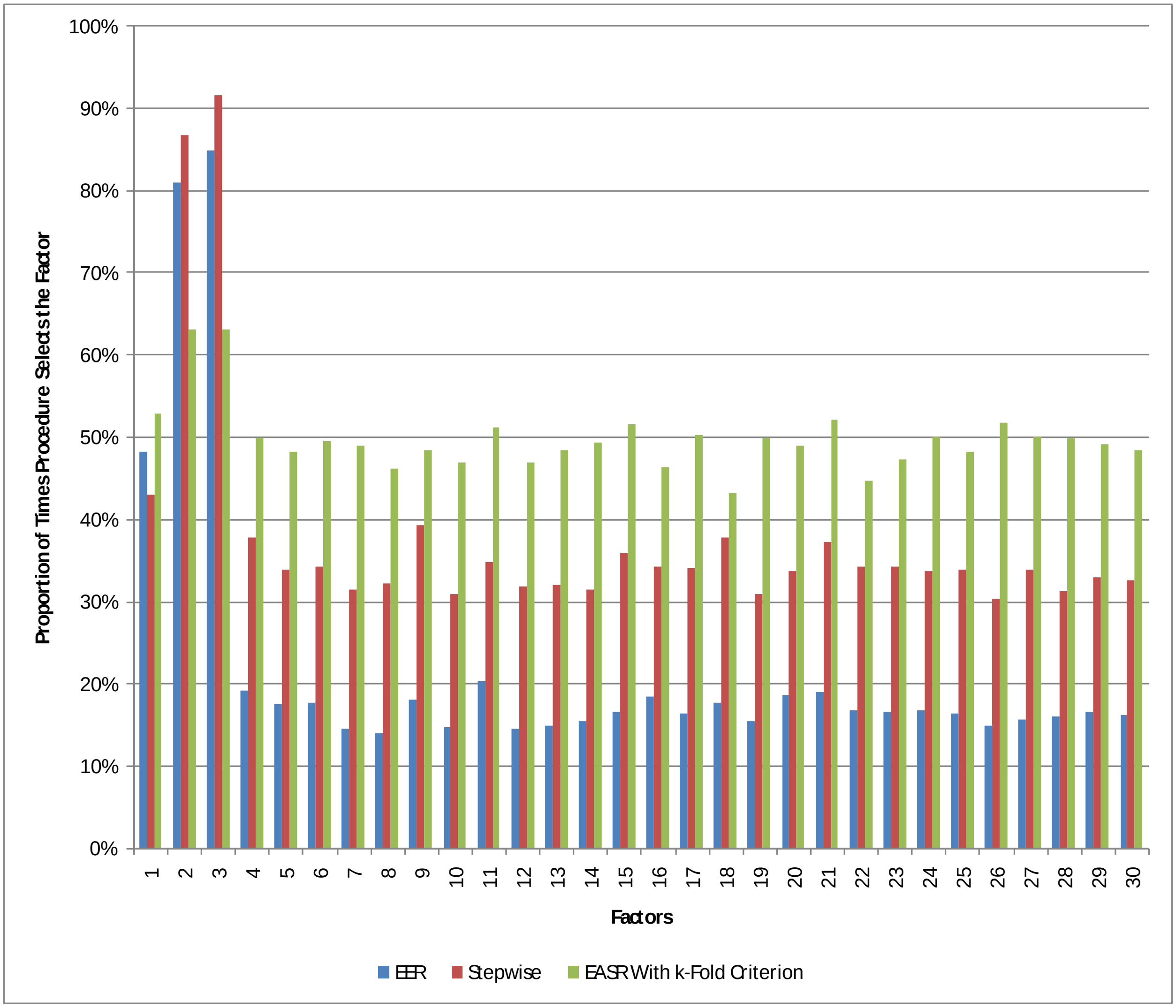

Figure 20. Comparison of EER, Stepwise, and k-Fold Criterion

8. Applicability to Other Procedures and Drawbacks

Ordinary least squares estimators belong to the class of maximum likelihood estimators. The estimates are unbiased (i.e., $\mathrm{E}\left(\hat{\beta}\right)=\beta$), consistent (i.e., $\lim_{N\to\infty}\Pr\left(\left|\hat{\beta}-\beta\right|>\varepsilon\right)=0$ for an arbitrarily small ε), and efficient (i.e., $\operatorname{var}\left(\hat{\beta}\right)<\operatorname{var}\left(\tilde{\beta}\right)$ where $\tilde{\beta}$ is any linear, unbiased estimator of β. The ER procedure relies on the fact that parameter estimators are unbiased in extraneous variable cases though biased in varying directions across the omitted variable cases.

In the case of limited dependent variable models (e.g., logit), parameter estimates are unbiased and consistent when the model is correctly specified. However, in the omitted variable case, logit slope coefficients are biased toward zero. For example, suppose the outcome variable, *Y*, is determined by a latent variable Y^* such that $Y = \begin{cases} 1 \text{ iff } Y^* > 0 \\ 0 \text{ otherwise} \end{cases}$. Suppose also that Y^* is determined by the equation $Y^* = \alpha + \beta_1 X_1 + \beta_2 X_2 + u$. If we omit the factor X_2 from the (logit) regression model, we estimate $Y^* = \alpha + \beta_1^o X_1 + u$. Yatchew and Griliches (1985) show that

$$\beta_1^o = \frac{\beta_1}{\sqrt{1 + \frac{\beta_2^2 \sigma_{X_2}^2}{\sigma_u^2}}} < \beta_1 \tag{11}$$

As the denominator in (11) is strictly greater than zero, the estimator is biased toward zero. More importantly, the biased estimator has the same sign as the unbiased estimator. This violates all four of conditions in (6), any one of which would validate the ER procedure. However, while estimates of deterministic parameters are biased downward, estimates of non-deterministic parameters are randomly biased. For example, if Y^* is determined by the equation $Y^* = \alpha + \beta_1 X_1 + \beta_2 X_2 + u$ and we estimate $Y^* = \alpha + \beta_1^o X_1 + \beta_3^o X_3 + u$, we obtain

$$\beta_1^o = \frac{\beta_1}{\sqrt{1 + \frac{\beta_2^2 \sigma_{X_2}^2}{\sigma_u^2}}} < \beta_1 \tag{12}$$

$$\beta_3^o = \frac{\beta_3}{\sqrt{1 + \frac{\beta_2^2 \sigma_{X_2}^2}{\sigma_u^2}}} = 0 \tag{13}$$

Because X_3 is extraneous, β_3 is zero and so repeated estimates of $\hat{\beta}_3^o$ will be randomly distributed around zero. In short, parameter estimates for deterministic factors will be: (1) unbiased in the correctly specified case, (2) unbiased in the extraneous variable case, and (3) biased toward zero in the omitted variable case. But, parameter estimates for spurious factors will be unbiased toward zero in all three cases. Hence, we can expect the cross-model chi-square statistics to work in the case of logit models, though likely in an asymptotic sense.

One drawback of the ER (and EER) procedure is that the procedure may select a set of factors that, while individually passing the cross-model chi-square test, are not statistically significant when run together in a regression model. One possible explanation is that, within the confines of a single regression model, the error variance is large enough to drown out the explanatory power of a factor but, because the cross-model chi-square statistic is based on cross-model information that has estimated and filtered out the error term, the factor appears significant. This is analogous to the gain in information obtained from employing panel data versus time series data (cf., Davies, 2006). A time series data set may measure the same phenomenon over the same time period as a panel data set, but because the panel data set also measures the phenomenon over multiple cross-sections, information across cross-sections can be used to mitigate the error variance within a given time period.

9. Conclusion

The purpose of this analysis is to build on ASR by proposing a new procedure, ER, that combines ASR with a cross-model chi-square statistic that attempts to distinguish deterministic factors from spurious factors. Recognizing that, for large data sets, even ER is infeasible even with the application of super computation, this analysis further proposes an adaptation on ER, EER, which samples the space of possible regression models.

This analysis (1) proves that, under conditions less stringent than the classical linear model conditions, the cross-model chi-square statistic is a reasonable measure for distinguishing between deterministic and spurious factors, (2) demonstrates via Monte-Carlo studies the likelihood of ER produce false positive and false negative results under conditions of varying numbers of factors in the data set, varying variance of the regression error, and varying choice of critical value, (3) proposes a procedure, EER, that estimates ER results via sampling a subset of the space of possible regression models, (4) demonstrates via Monte-Carlo studies the likelihood of EER producing false positive and false negative results under conditions of varying sample size, and varying number of factors in the deterministic equation, (5) compares EER model selection with backward stepwise and the k-fold criterion by via Monte-Carlo studies that apply the same data sets to all three procedures, and (6) demonstrates that EER avoids false positives and false negatives better than the k-fold criterion, avoids false positives approximately as well as stepwise, and avoids false negatives significantly better than stepwise.

10. References

Carmines, E.G., and J.P. McIver, 1981. Analyzing models with unobserved variables: Analysis of covariance structures, in Bohmstedt. In G.W. and E.F. Borgatta, eds., Social Measurement. Sage Publications: Thousand Oaks, CA. pp. 65-115.

Davies, A., 2006. A framework for decomposing shocks and measuring volatilities derived from multi-dimensional panel data of survey forecasts. International Journal of Forecasting, 22(2): 373-393.

Kline, R.B., 1998. Principles and practice of structural equation modeling. Guilford Press: New York.

Kuk, A.C., 1984. All subsets regression in proportional hazard models. Biometrika, 71(3): 587-592.

Ullman, J.B., 2001. Structural equation modeling. In Tabachnick, B.G. and L.S. Fidell, eds., Using Multivariate Statistics. Allyn and Bacon: Needham Heights, MA. pp. 653-771.

Yatchew, A. and Z. Griliches, 1985. Specification error in probit models. The Review of Economics and Statistics, 67: 134-139.