Estimated Exhaustive Regression

Spencer Marlen-Starr

Graduate Student in the Data Analytics Engineering Master's Program at

George Mason University, Arlington, VA

This paper proposes the Exhaustive Regression (ER) procedure, a supervised learning algorithm for variable selection purposes based on the All Subsets Regression (ASR) algorithm, known properties of Ordinary Least Squares (OLS) estimators, and replacing the standard optimal model selection criteria with a cross-model Chi-Square Statistic used to detect spuriously selected variables. However, this ER algorithm retains the prohibitive runtime problems that prevent the standard ASR algorithm from being used on the types of high dimensional datasets commonly confronting the analyst in practice today. To get around this problem, a further modification, Estimated Exhaustive Regression (EER), a computationally feasible version of ER, is proposed. Its properties in the context of continuous linear regression analysis and modeling are explored via Monte Carlo Simulations. Furthermore, how the EER algorithm compares to LASSO, and both the Backward Elimination and Forward Selection variations of Stepwise Regression, three well known benchmark variable selection algorithms are also explored via Monte Carlo Comparison Experiments on randomly generated synthetic datasets where the true underlying model characterizing each dataset is known in advance. Econometrically speaking, without taking runtime into account, the standard ASR algorithm is superior to LASSO and Stepwise in that it is based on well-known estimator properties and as a result is not an ad-hoc datamining technique like LASSO and Stepwise Regression.

Keywords: Regression, Variable Selection, Supervised Learning, Monte Carlo Simulation

1. Background and Introduction

Regression analysis is used when the data analyst or researcher wants to test an initial hypothesis (generally based on some sort of pre-existing theoretical framework) in a systematic empirical manner against a dataset of observations, whether they be a sample or the entire population of interest. More specifically, what a regression analysis, which in practical circumstances is almost always a multivariate regression analysis rather than a univariate regression situation, seeks to determine what the probability of observing your data on the assumption that your initial hypothesis is true. This approach is sound when you have a suitable initial hypothesis inferred from a reliable body of knowledge (a social scientific or scientific field for example, or from seasoned professionals in your industry). However, Leamer (1978 and 1983) reminds us that when an analyst or researcher lacks an adequate basis for forming an initial hypothesis before confronting the data, so he instead chose to run several dozen or more different possible multivariate regression specifications and chose the specification which best fits the data and achieves the highest overall model statistical significance, it would then be disingenuous for him to pretend as if he had formulate that hypothesis which he generated as his initial hypothesis all along. In this situation, the opposite is the case, and not only is it dishonest as a research practice, but it violates the standard rules of statistical inference.

Machine learning algorithms, formerly known as data mining techniques, are resorted to in regression analysis when it is either not possible, feasible or preferable to test an initial hypothesis which is stated in enough detail to entail a regression specification. That is to say, the ultimate purpose of many if not most machine learning algorithms is to serve as *hypothesis generators*, not necessarily hypothesis validators or falsifiers.

Regression-based machine learning algorithms (i.e. supervised, non-classification ML techniques) are used very often by data analytics engineers, data scientists, and other data analytics professionals, as well as by statistical, econometrics, computer science, and artificial intelligence researchers in universities, because they are preferable to traditional alternatives when doing analyses in the context of hypothesis-less situations, especially on so-called “big” datasets (that is, those with high volume, velocity, and/or variety where by volume refers to the number of columns, not the number of rows).

The novel supervised learning algorithm for the purposes of optimal regressor selection evaluated in this study, Estimated Exhaustive Regression (EER henceforth), was proposed in a working paper by noted econometrician Antony Davies (2008). Its underlying statistical properties are explained and its practical characteristics are explored by comparison with several standard optimal regressor selection algorithms via the accuracy of how accurately it is able to select the true population models describing synthetic datasets which are known in advance by their construction via Monte Carlo.

In contrast to the majority of existing variable selection techniques which are essentially *ad hoc*, either in theory or in practice, EER is based on known properties of Ordinary Least Squared estimators when performing Multiple Linear Regression analyses in the presence of omitted and extraneous variable models and a different selection criteria for the All Subsets Regression (ASR henceforth) procedure (also known as Best Subset Selection) which attempts to distinguish between parameter estimates which are statistically significant due to underlying structural relationships and those which are spurious.

2. All Subsets Regression

To perform ASR, the *best fitting* regression specification is selected the subsets of all possible sizes from the superset of all possible combinations of regressors in the set of k candidate regressors in the given dataset. That is, we fit all k possible regression equations that contain exactly one predictor, then we do so for all models which contain exactly two predictors, and so on and so forth up until $k - 1$ (one must be subtracted to avoid perfect multicollinearity). Thence, we then look at the estimates generated by running all of those models, with the goal of identifying the one that is *best* in a way which is both meaningful and measurable.

The previous point is worth emphasizing here, one major advantage of ASR as compared with almost all other automated variable selection methods is that is attempts to select optimal *models* rather than individual variables one at a time. After all, variable selection is only a means to an end with that end being determining the optimal *overall model* because there is a difference between whether an individual regressor is statistically significant and the overall regression model is statistically significant[1].

The classical All Subsets Regression Algorithm is outlined on the next page:

[1] In the case of standard Multiple Linear Regression Estimation via OLS, the significance of individual regressors is assessed via a *t-test* while the significance of the overall regression is assessed via an *F-test*.

All Subsets Regression Algorithm

```
Step 1. Let $\aleph_0$ denote the null model, i.e. that which contains only the
    intercept (0 regressors). All this model predicts is the sample mean or
    population mean for each observation.
Step 2. For $p = 1, 2, 3, \dots, k$:
    a. Fit all $\binom{k}{p}$ regression models which contain $p$ regressors.
    b. Select the optimal specification among all these $\binom{k}{p}$ models and
       call this model $O_p$. Which model is optimal is typically determined
       by which has the highest R$^2$ or the smallest Residual Sum of
       Squared Errors.
Step 3. Select the single best regression specification among all of the
    local optima $O_0, \dots, O_k$. How to determine which of the local or sub
    optima is the global optimally fit regression can be done in any number
    of ways including, with either Adjusted R$^2$ or Standard R$^2$, the AIC,
    and Cross-Validation being the most common.
```

Theoretically speaking, ASR appears at first glance to be the most honest and straightforward possible automated optimal regression model selection algorithm because it assesses each and every possible regression specification which could be constructed given your dataset one by one in order to determine which best fits the data.

However, ASR's greatest strength also proposes a serious potential weakness of using the standard version of it in practice, namely, it is likely to select a *spurious* model rather than the true underlying population model because it estimates all possible models and the more models are estimated, the more chances there are for one to be significant by random chance alone. Therefore, if we truly want to make it more reliable in practice, we must find a better way to identify included regressors which are likely to be spurious.

3. Exhaustive Regression

3.1 Motivation and Estimator Properties

Suppose that unknown to the analyst, the process that determines the outcome variable is characterized by:

$$\mathrm{Y} = \mathbf{X_1}\boldsymbol{\beta}_\mathbf{1} + \ \mathbf{X_2}\boldsymbol{\beta}_\mathbf{2} + \mathrm{u} \tag{1}$$

where $\mathbf{u}$ is an *N*x1 vector of i.i.d. errors which follow a Gaussian Distribution. There are three possible outcomes in terms of the accuracy of the models fitted to a dataset by any regression which uses an OLS estimation procedure: an omitted variable model, a correctly specified model, and an extraneous variable model. An omitted variable model is a regression equation which is missing at least one of the explanatory variables, and it is also possible for an omitted variable model to include one or more extra candidate variables as well. A correctly specified regression model is one which includes all the variables which truly explain or predict the outcome variable and no other candidate variables. And extraneous variable model includes all explanatory variable, plus at least one spurious variable. Put formally, they can be stated as the following:

The Omitted Variable Case: $\mathbf{Y} = \ \mathbf{X_1}\boldsymbol{\beta}_\mathbf{1} + \mathbf{u_O}$

The Correctly Specified Case: $\mathbf{Y} = \ \mathbf{X_1}\boldsymbol{\beta}_\mathbf{1} + \mathbf{X_2}\boldsymbol{\beta}_\mathbf{2} + \mathbf{u}$

The Extraneous Variable Case: $\mathbf{Y} = \ \mathbf{X_1}\boldsymbol{\beta}_\mathbf{1} + \mathbf{X_2}\boldsymbol{\beta}_\mathbf{2} + \mathbf{X_3}\boldsymbol{\beta}_\mathbf{3} + \mathbf{u_E}$

Firstly, the correctly specified case, the values fitted to the sample observations when running an OLS regression procedure are of the following form:

$$\begin{bmatrix}\hat{\beta}_1 \\ \hat{\beta}_2\end{bmatrix} = \left([\mathbf{X_1} \quad \mathbf{X_2}]'[\mathbf{X_1} \quad \mathbf{X_2}]\right)^{-1} [\mathbf{X_1} \quad \mathbf{X_2}]'\ \mathbf{Y} \tag{2}$$

For both the correctly specified case and the overspecified (extraneous), the expected values of the slope estimates are equal to their population parameters, i.e.

$$\text{CSM: } E\left(\begin{bmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{bmatrix}\right)=\begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix}\quad \text{and}\quad \text{EVM: } E\left(\begin{bmatrix}\hat{\beta}_2\\ \hat{\beta}_2\\ \hat{\beta}_2\end{bmatrix}\right)=E\left(\begin{bmatrix}\beta_1\\ \beta_2\\ \beta_3\end{bmatrix}\right)$$

where the square brackets indicate that it is a partitioned matrix.

In contrast to the previous two cases however, for an omitted variable model we have:

$$\widehat{\boldsymbol{\beta}}_{\mathbf{1}}=(\mathbf{X}_{\mathbf{1}}'\mathbf{X}_{\mathbf{1}})^{-1}\,\mathbf{X}_{\mathbf{1}}'\mathbf{Y} \tag{3}$$

and the expected values for the slope estimates for an OVM are:

$$E\big(\widehat{\boldsymbol{\beta}}_{\mathbf{1}}\big)\;=\boldsymbol{\beta}_{\mathbf{1}}+\;(\mathbf{X}_{\mathbf{1}}'\mathbf{X}_{\mathbf{1}})^{-1}\,\mathbf{X}_{\mathbf{1}}'\mathbf{X}_{\mathbf{2}}\boldsymbol{\beta}_{\mathbf{2}}\neq\boldsymbol{\beta}_{\mathbf{1}} \tag{4}$$

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 $\mathbf{X}_{\mathbf{1}}'\mathbf{X}_{\mathbf{2}}\boldsymbol{\beta}_{\mathbf{2}}$.

3.2 The Cross-Model Chi-Square Statistic

Davies proposed method of identifying and eliminating variables in the selected model which are spurious by the traditional ASR algorithm is to include a *cross-model Chi-Square statistic* as the selection criteria in order to determine the stability in estimates of each parameter across different possible models (rather than the standard methods of R^2, AIC, or cross-validation). This cross-model chi-square test of stability compares parameter estimates for each candidate regressor across all $2^K - 1$ possible models in which that factor appears. Factors whose parameter estimates yield are noticeably and quantifiably different results across different overall regression specifications are identified as spurious.

In the context of a multiple regression model of the standard form:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + u \tag{5}$$

where the null hypothesis states $\forall\, i |\, i \in (1{:}\,k)$, $\beta_i = 0$ and the $\hat{\beta}_{ij}$ are independent, that is, $\forall\, i\, \&\, j |\, i, j \in (1{:}\,k)$; with N observations on each of the k regressors; the *Cross-Model Chi-Square Statistic* for regressor x_i is given by:

$$\sum_{j=1}^{2^{K-1}} \left(\frac{\hat{\beta}_{ij}}{s_{\beta_i}} \right)^2 \sim \chi^2_{2^{K-1}} \tag{6}$$

However, the above version is still highly likely to result in too many Type-II Errors for cases with high degrees of freedom. This danger can be alleviated simply by dividing the above by the degrees of freedom, like so

$$c_i = \sum_{j=1}^{2^{K-1}} \left(\frac{\hat{\beta}_{ij}}{s_{\beta_i}} \right)^2 \sim \chi^2_1 \tag{7}$$

This modification of the classical ASR procedure by way of an alternative criterion for regressor selection is Exhaustive Regression (ER). In terms of the preferred threshold for when a candidate variable has been selected for inclusion by the exhaustive regression procedure, this is fundamentally up to the researcher, however, an previously indicated guidepost according to both Carmines and McIver (1981) and Kline (1998) is that an analyst should only conclude the data represent a good fit to the hypothesis implicit in the estimated regression equation when the *relative cross-model chi-square statistic* is greater than three, i.e., $c_i > 3$.

ER has an important potential downside to note here, according to Davies, "because the $\hat{\beta}_{ij}$ estimates are obtained by exploring all combinations of factors from a single superset, one might expect the $\hat{\beta}_{ij}$ to be positively correlated." Furthermore, one may expect for the strength of their correlation to increase in tandem with the degree of multicollinearity, and this is particularly true when the explanatory variables are positively correlated.

A crucial positive characteristic of the ER procedure which it is important to mention is that its ability to accurately discriminate between factors which actually determine the outcome, i.e., structural factors, and those which appear to be significant by random chance alone, i.e., spurious factors, increases along with the number of candidate factors in the dataset. This is an extremely desirable trait in the context of conducting analyses on "big" datasets which become more and more ubiquitous every year.

4. Estimated Exhaustive Regression

The fundamental problem with using ASR in practice is still present and not even diminished in ER, that problem is that it is computationally impractical to run ASRs for datasets with 35 to 45 or more columns in them and infeasible to do so with 60 or more column datasets because it would take even a modern PC with a high end quad core processor and 32 GB of good quality RAM or more weeks or months, or possibly even a couple of years to run. This is due to what may be termed the combinational explosion, because All Subsets Regressions estimates all $2^K - 1$ potential regression specifications out of the K candidate factors (the number of columns in the dataset minus one) which means the number of regressions which must be estimated double for each increase in K.

The Estimated Exhaustive Regression procedure gets around this by selecting only a random sample J of the $2^K - 1$ possible combinations of the $k$columns, in the dataset. By way of Monte Carlo Simulations, we were able to get a sense of what might be a suitable minimum amount of the J random models (without replacement of course) for datasets with different numbers of candidate predictors K. Surprisingly, as few as only several hundred, or even 100 randomly selected regression specifications can be enough to outperform some benchmark methods as we will see in the following section.

One very important aspect of the EER Algorithm which has already been noted but emphasized is that it randomly selects J models (entire possible regression equations), not variables. Selecting factors randomly would bias the model selection toward models with a total of $\frac{K}{2}$ factors which is arbitrary and thus not correct. Selecting whole models randomly on the other hand gives each of the $2^K - 1$ possible models an equal probability of being chosen.

Estimated Exhaustive Regression Algorithm

```
Repeat each of the following steps from 1…
    Step 1. Randomly chose one out of the $2^K - 1$ possible regression models to obtain estimates for each of the $k$ regressors included in it where $k$ can be anything from 1 to the number of candidate regressors $K$.
    Step 2. Calculate $c_1, c_2, c_3, \dots, c_k$ according to formula (3).
    Step 3. Choose which variables $i$ to include in the model by whether its Chi-Square statistic is above the threshold chosen by the analyst, $c_i > 3$.
… until J where the value of J is chosen by the analyst.
```

5. Comparing the Performance of EER to Common Benchmarks

In this study, the EER procedure is evaluated chiefly via comparison to three popular benchmark methods for automated regressor selection, one modern and two classical. The 1st benchmark regressor selection algorithm is the LASSO Regression algorithm Tibshirani (1996). LASSO stands for the *least absolute shrinkage and selection operator*, it is a member of the family of regression method regularization techniques which employ a penalty used to tune their predictions within their objective functions known as *Shrinkage Methods* [2]. The 2nd benchmark method is Backward Elimination, a form of the Stepwise Regression (SR) family of regressor selection procedures and the 3rd benchmark is Forward Selection, another common regressor selection procedure from the Stepwise Regression family of methods.

5.1 Benchmark Method 1: The Least Absolute Shrinkage and Selection Operator

A LASSO Regression is essentially a modification of to the Sum of Squared Errors formula minimized by a standard linear regression whereby a penalty factor is added in to regularize its predictions. This modification is shown below:

$$SSE_{L_1} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{P}|\beta_j| \tag{7}$$

The λ in equation (7) is the penalty factor is used to correct its predictions by shrinking the absolute value of all estimates obtained by running a model including all candidate regressors by the same amount which begins eliminating candidate regressors after it surpasses a certain threshold. In this sense, the selection criterion employed by LASSO Regression can be both ad hoc and arbitrary, two undesirable properties from a statistical perspective. Despite this, due to its positive track record in practical applications where the interpretability of the selected model is not important, LASSO is

[2] The other main shrinkage methods are Ridge Regression and Elastic Net, with the latter being a hybrid method combining features of LASSO and Ridge Regression.

one of the most popular methods for conducting automated optimal feature selection of predictors among data scientists and machine learning researchers today.

5.2 Benchmark Methods 2 & 3: BE and FS Stepwise Regression

The other two benchmark algorithms whose results are compared with the results of EER are the two standard versions of Stepwise Regression, namely, Backward Elimination Stepwise (wherein the first regression model estimated includes all candidate regressors and regressors are removed one at a time incrementally) and Forward Selection Stepwise (wherein the initial regression model estimated includes no candidate regressors and each regressor is then added to the model one at a time incrementally). SR is one of the oldest automated regressor selection procedures in statistics. It was the main method for this purpose besides ASR used by statisticians prior to the early stages of the modern machine learning revolution back in the 1990s.

Stepwise Regression procedures attempt to select optimal regressors among a "smart" sampling of all possible regression models (as compared to ASR). For example, with only 30 candidate regressors in the given dataset, SR procedures typically only estimate a subset of less than 100 possible models[3] before making its selection. This has the upside of saving on computation time, with the downside being that it can only hope to select *an* (locally) optimal regression rather than *the* global optimum like ASR. Which version of SR is more likely to select the correctly specified regression equation for a given dataset depends on two conditions; the first condition is how many regressors are included in the structural regression equation being modelled and the second is how many candidate

[3] Out of a total of $2^{30} - 1 = 1.07 \times 10^{9}$ or a little over 1 billion regressions.

regressors are in the dataset[4]. The step-by-step algorithms for performing the BE and FS versions of SR are:

Backward Elimination Stepwise Regression Algorithm

```
Step 1. Tune/train the model on the training set using all $P$ candidate
predictors.

Step 2. Calculate the model's overall performance.

Step 3. Rank each candidate regressor/predictor by its importance.

Step 4. **For** *each subset size* $S_i$, $i = 1, 2, \dots, S$ **do**
    5. Keep the $S_i$ most important predictors
    6. Tune/train the model on the training set using $S_i$ predictors
    7. Calculate the model's overall performance
**end**

Step 5. Calculate the performance profile over the $S_i$.

Step 6. Determine the appropriate number of predictors by which $S_i$ are
associated with the best performance (lowest *p-value* among all $p_{S_i} < \alpha$.

Step 7. Fit the final model based on the optimal $S_i$.
```

Forward Selection Stepwise Regression Algorithm

```
Step 1. Create an initial null model containing only an intercept term.

*Repeat all steps below until…*

    **for** each predictor not in the current model **do**
        Create a candidate model by adding it to the current model.
        Use a hypothesis test to estimate its statistical significance.
    **End**
    **if the smallest p-value is less than the inclusion threshold then**
        Update the current model to include the predictor with the highest
        test statistic value, i.e. the lowest *p-value* among all of those which
        are above the statistically significance threshold.
    **Else**
         Stop.
    **End**
*… there are no statistically significant candidate regressors remaining that
are not already included in the model.*
```

[4] The first condition is not known apriori, for if it were, there would obviously be no need to run a statistical learning algorithm to determine the structural regression equation in the first place, while the second needs is known upfront.

Because the Backward Elimination procedure always begins by estimating the full regression model (the one including every candidate regressor), it tends for work better when both the true population model and the dataset are of large data scope. A case of large data scope is where there are a large number of candidate predictors p compared to the number of observations on each of them n (where large typically means greater than 45 or 50). Using the same reasoning in reverse, because the Forward Selection procedure always begins by estimating the null model, i.e., the fitness of just the intercept, it usually works best on for datasets of limited scope and for modelling or forecasting low dimensional processes/phenomena (high K relative to N).

5.3 Results of the Monte Carlo Comparisons of EER to the Benchmarks

Our 1st Benchmark Variable Selection Algorithm, LASSO, only selected 4,937 correctly specified regression models out of a possible 58,500, or about 8.4% of the time.

Backward Elimination Stepwise Regression, our 2nd Benchmark Method, only selected 1,143 correctly specified models for a True Positive Rate of 1.96%, 41,830 (71.5%) of the models it selected we overspecified, and models which failed to include at least one structural variable 15,527 times for a False Negative Rate of 26.5%.

Forward Selection Stepwise Regression, our 3rd Benchmark Method, only selected 1,413 correctly specified models, which is only 2.4% of the time (its True Positive Rate). As for regression selections with all the correct regressors included but also at least one spurious regressor, i.e., overspecified model selections, it selected 45,004 (76.9%) such regression specifications. And lastly, FS Stepwise failed to select all true structural variables (omitting at least one) for 12,083 of the 58,500 dataset, which is 21.6% of them (its False Negative Rate).

The EER procedure with its α significance threshold chosen for all the candidate regressors x_i set to $c_i > 3$ was run with the number of randomly selected models, J estimated set to 6 different sizes in order to assess what may be a suitable level to start out with when using it as an analyst. The results are presented in the table below:

Chi-Square Statistic Threshold for Selection	**J**	**Percentage Underspecified**	**Percentage Correctly Specified**	**Percentage Overspecified**
$\chi_i^2 > 3$	50	0.0054%	0%	99.995%
$\chi_i^2 > 3$	100	0%	0%	100%
$\chi_i^2 > 3$	150	0%	0%	100%
$\chi_i^2 > 3$	200	0%	0%	100%
$\chi_i^2 > 3$	250	0%	0%	100%
$\chi_i^2 > 3$	500	0%	0%	100%

Surprisingly, at least in the case of modelling datasets with 30 candidate regressors and 500 observations on each of them, its performance was excellent for each of the 6 different number of randomly selected models, *J*, estimated. Even when only 50 random regressions were estimated for each dataset, the results were fantastic. However, because the difference in computation time required between running an EER with *J* set to 50 and with it set to 250 or 500 is not substantial, for now, the recommendation for anyone wanting to use EER in practice is to start out by setting *J* equal to 250 or 500.

In order to ascertain whether some correctly specified regressions could be identified by EER by increasing our significance requirement for regressor selection from 3 to something higher, 5, 7, and 10 were tried, each of which for the same six *J* settings. The results we almost identical, so only the exact results for only one of these alternative significance thresholds for selection are reported below:

Chi-Square Statistic Threshold for Selection	J	Percentage Underspecified	Percentage Correctly Specified	Percentage Overspecified
$\chi_i^2 > 7$	50	1.16%	0%	98.84%
$\chi_i^2 > 7$	100	0%	0%	100%
$\chi_i^2 > 7$	150	0%	0%	100%
$\chi_i^2 > 7$	200	0.23%	0%	99.77%
$\chi_i^2 > 7$	250	0.21%	0%	99.79%
$\chi_i^2 > 7$	500	0.16%	0%	99.84%

One particular aspect of the selections made by EER which is more qualitative than the results reported in the previous two tables, but one which I would be remiss if I did not include in this paper is that for the vast majority of the regression models selected by EER, only one extra regressor is included in the output and that single extraneous regressor included is almost always the very next candidate regressor after the last structural regressor. This is in stark contrast to the large number of extraneous variable models selected by both Backward Elimination and Forward Selection. The extra regressors included in their regression selections exhibited no pattern whatsoever and often included many more than one extra candidate factor.

6. Conclusion

While not necessarily ideal if the goal is to maximize the percentage of selecting the correctly specified model only, Estimated Exhaustive Regression has a substantially lower False Negative Rate than LASSO, Backward Elimination Stepwise, or Forward Selection Stepwise Regression. And once again, as long as the standard assumptions underlying Ordinary Least Squares estimators for Multiple Linear Regression Analysis are satisfied, the expected value of the slope estimates returned for extraneous variable models equal their true parameter values[5].

Therefore, while the ideal target is clearly the maximal selection of correctly specified models, one can say that the selection of an extraneous variable regression specification is not incorrect, strictly speaking. As a result, a more robust primary target for the output of any feature selection algorithm which satisfies the standard OLS assumptions should simply be the minimization of incorrectly excluded regressors (False Negatives), i.e. minimizing the likelihood of selecting omitted variable models rather than the target of simultaneously minimizing the likelihood of selecting both omitted and extraneous variable models because the former are much more problematic than the latter. Furthermore, the runtime required for Estimated Exhaustive Regression with increasing J arguments up to 500 is still less than either version of Stepwise while not being substantially larger than LASSO.

[5] Unlike in the case of Correctly Specified Regression Equations, the expectation of the coefficient estimates for Extraneous Regression Models are only equal to the true parameters asymptotically.

For any supervised learning algorithm with feasible computation time requirements to result in anything close to a 0% False Negative Rate on 58,500 randomly generated synthetic datasets, with 117 different underlying probabilistic or statistical conditions and 500 random variations for each of those conditions is incredibly good performance even if it fails to correctly identify any models which are specified perfectly.

7. Drawbacks, Limitations, and Further Work

Further research will have to be done in order to determine how well EER performs in the selection of factors in nonlinear multiple regression models and how that performance compares with the Benchmark Methods included in this study as well as others more specifically suited for nonlinear regression modeling (such as basic neural networks, support vector machines, and deep learning). The same thing will have to be explored for the case of optimal factor selection in classification models[6] as well.

One drawback common to both the ER and EER procedures is that while they are more likely to select an optimal overall model specification than the benchmarks, it is still possible for it to select a set of regressors more based on each of their individual optimality than their collective optimality. Because of this, it remains possible for them to select models, all of those included regressors individually pass the cross-model chi-square test yet are not statistically significant as an overall regression when run together. 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,

[6] Regression Models with either a discrete dependent variable (e.g. logit and probit) or a discrete dependent variable and at least one discrete independent variable as well.

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[7]. However, this is less likely with ER and EER than it is with current Benchmarks.

Finally, much further investigation and analysis will be needed to uncover what is behind the peculiarly consistent, indeed, seemingly predictable way in which Estimated Exhaustive Regression selects one spurious factor in the regression specifications it returns. Perhaps this can be eliminated somehow, in which case, EER could truly become a powerhouse in this domain, or perhaps it is somehow a feature not a bug which helps to explain just how EER appears to be able to avoid committing any Type-II Errors.

[7] This is analogous to the gain in information obtained from employing panel data versus time series data (cf., Davies, 2006).

References

Leamer, E.E, 1983. Specification Searches. Wiley.

Leamer, E.E, 1983. Let's Take the Con out of Econometrics. The American Economic Review, 73: 31-43.

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.

Davies, A., 2008. Exhaustive Regression: An Exploration of Regression-Based Data Mining Techniques Using Super Computation. The Research Program on Forecasting, George Washington University, RPF Working Paper No. 2008-008. http://www.gwu.edu/~forcpgm/2008-008.pdf

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.

Tibshirani, R, 1996. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society, Series B (Methodological), 58: 267-288.

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