Smith *J Big Data* (2018) 5:32
https://doi.org/10.1186/s40537-018-0143-6

Journal of Big Data

SHORT REPORT Open Access

Step away from stepwise

Gary Smith*

*Correspondence:
gsmith@pomona.edu
Department of Economics, Pomona College, 425 N. College Avenue, Claremont, CA 91711, USA

Abstract

Background: Stepwise regression is a popular data-mining tool that uses statistical significance to select the explanatory variables to be used in a multiple-regression model.

Findings: A fundamental problem with stepwise regression is that some real explanatory variables that have causal effects on the dependent variable may happen to not be statistically significant, while nuisance variables may be coincidentally significant. As a result, the model may fit the data well in-sample, but do poorly out-of-sample.

Conclusion: Many Big-Data researchers believe that, the larger the number of possible explanatory variables, the more useful is stepwise regression for selecting explanatory variables. The reality is that stepwise regression is less effective the larger the number of potential explanatory variables. Stepwise regression does not solve the Big-Data problem of too many explanatory variables. Big Data exacerbates the failings of stepwise regression.

Keywords: Stepwise regression, Data mining, Big Data

Introduction

Researchers typically do not know with certainty which explanatory variables ought to be included in their multiple regression models. More than 50 years ago, stepwise regression was proposed as an efficient way to select the most useful explanatory variables. Despite widespread criticism, it never disappeared and has enjoyed a revival as a method for analyzing Big Data, where the number of potential explanatory variables can be very large. This paper uses a series of Monte Carlo simulations to demonstrate that stepwise regression is a poor solution to a surfeit of variables. In fact, the larger the number of potential explanatory variables, the more likely stepwise regression is to be misleading.

The stepwise regression method

Efroymson [1] proposed choosing the explanatory variables for a multiple regression model from a group of candidate variables by going through a series of automated steps. At every step, the candidate variables are evaluated, one by one, typically using the t statistics for the coefficients of the variables being considered.

A forward-selection rule starts with no explanatory variables and then adds variables, one by one, based on which variable is the most statistically significant, until there are no remaining statistically significant variables.

A backward-elimination rule starts with all possible explanatory variables and then discards the least statistically significant variables, one by one. The discarding stops when each variable remaining in the equation is statistically significant. Backward elimination is challenging if there is a large number of candidate variables and impossible if the number of candidate variables is larger than the number of observations.

A bi-directional stepwise procedure is a combination of forward selection and backward elimination. As with forward selection, the procedure starts with no variables and adds variables using a pre-specified criterion. The wrinkle is that, at every step, the procedure also considers the statistical consequences of dropping variables that were previously included. So, a variable might be added in Step 2, dropped in Step 5, and added again in Step 9.

Some researchers use stepwise regression to prune a list of plausible explanatory variables down to a parsimonious collection of the "most useful" variables. Others pay little or no attention to plausibility. They let the stepwise procedure choose their variables for them.

False confidence in stepwise results

Several authors [2–10] have pointed out that standard statistical tests assume a single test of a pre-specified model and are not appropriate when a sequence of steps is used to choose the explanatory variables. The standard errors of the coefficient estimates are underestimated, which makes the confidence intervals too narrow, the t statistics too high, and the p values too low—which leads to overfitting and creates a false confidence in the final model. In 1995, one educational psychology journal announced that authors should not submit papers using stepwise regression [10].

However, stepwise regression remains a popular tool (for example, [11–13]) and most statistical software packages include stepwise regression—which evidently reflects the demand for it and, perversely, may tempt researchers to try it. A survey of papers published in 2004 in three leading ecological and behavioral journals found that 57% of the papers that reported multiple regression results used stepwise regression [7]. A survey of four leading epidemiologic journals found that 20% of the articles published in 2008 used stepwise regression [14]. A study of articles published between 2004 and 2008 in two leading Chinese epidemiology journals found that, of the articles using multiple regression models, 44% used stepwise procedures [15].

Several textbooks endorse stepwise regression [16, 17], including a handbook explicitly devoted to data mining methods [18]. The Chartered Financial Analyst Level II exam includes stepwise regression [19].

Other problems with stepwise regression

One fatal problem that has not been emphasized is that stepwise estimates are not invariant to inconsequential linear transformations. For example, Nobel laureate Milton Friedman [20] proposed this model of consumer spending,

$$C = \alpha + \beta_1 Y + \beta_2 P + \varepsilon$$

where C is spending, Y is current income, and P is permanent income. The idea is that households don't live hand-to-mouth, basing their spending decisions solely on how

much income they are currently earning. They also take into account their average or "permanent" income. The difference between current income and permanent income is labeled transitory income, $T = Y - P$. If we interpret the variables carefully, it doesn't matter which two of these three variables, Y, P, and T, are included in the model and multiple regression estimates will give equivalent estimates and identical spending predictions. Not so with stepwise regression. From a list of candidate explanatory variables that includes, say, Y and T, both might be chosen, but from a list that includes P and T, only P might be chosen.

Similarly, a spending model that uses last year's income and this year's income as explanatory variables should be equivalent to a model that uses last year's income and the change in income from last year to this year. Multiple regression estimates will not be affected; stepwise estimates might.

Thompson [10] argues that another major problem with stepwise regression is that a local optimization obtained by including variables one-by-one is not necessarily a global optimization. For example, selecting a fifth explanatory variable contingent on the four variables that were already chosen does not necessarily select the five variables that give the highest possible R^2.

However, global maximization is not a goal worth seeking. Choosing a model's explanatory variables based on R^2 or statistical significance is treacherous—and this is the most fundamental problem with stepwise regression and the most compelling reason why researchers should stop using it.

Data mining

The traditional statistical analysis of data follows what has come to be known as the scientific method that replaced superstition with scientific knowledge. Based on observation or speculation, the researcher poses a question, such as whether vitamin C reduces the incidence and severity of the common cold. The researcher then gathers data, ideally through a controlled experiment, to test the theory. If there are statistically persuasive differences in the outcomes for those taking vitamin C and those taking a placebo, the study concludes that vitamin C has a statistically significant effect. The researcher uses data to test a theory.

Data mining goes in the other direction, analyzing data without being motivated or encumbered by preconceived theories. Data-mining algorithms are programmed to look for trends, correlations, and other patterns in data. When an interesting pattern is found, the researcher may argue that the data speak for themselves and that is all that needs to be said. We don't need theories—data are sufficient.

In addition to those who believe that theories are unnecessary, some believe that data should be used to discover new theories (for example, [21–23]). The label *knowledge discovery* emphasizes that the goal is a data-driven discovery of new, heretofore, unknown theories. Indeed, committed data-miners view the use of a priori knowledge of the phenomena being modeled as a constraint that limits the possibilities for knowledge discovery [24].

Sagiroglu and Sinanc [25] describe data mining as a quest "to reveal hidden patterns and secret correlations." In the opening lines to a forward for a book on using data

mining for knowledge discovery [26], a computer science professor wrote, without evident irony,

> *"If you torture the data long enough, Nature will confess," said 1991 Nobel-winning economist Ronald Coase. The statement is still true. However, achieving this lofty goal is not easy. First, "long enough" may, in practice, be "too long" in many applications and thus unacceptable. Second, to get "confession" from large data sets one needs to use state-of-the-art "torturing" tools. Third, Nature is very stubborn—not yielding easily or unwilling to reveal its secrets at all.*

The author was apparently unaware of the fact that Coase intended his comment not as a lofty goal, but as a succinct criticism of the practice of ransacking data in search of statistical significance [27].

Variables should be included in a model because, on theoretical grounds, they should be in the model, not based on the size of their *t*-values. The estimated coefficients of the true explanatory variables are biased if variables that belong in the model are excluded, and have enlarged variances if variables that don't belong are included [28, 29].

New life with Big Data

Stepwise regression was born back when computers were much slower than today, but it has become a popular data-mining tool because it is computationally less demanding than a full search over all possible combinations of explanatory variables and, it is hoped, will give a reasonable approximation to the results of a full data-mining search. For instance, Cios et al. [22] recommend stepwise regression as an efficient way of using data mining for knowledge discovery (see also [30–32]).

Suppose that a researcher has 100 possible explanatory variables and wants to choose up to 10 variables to include in a regression model. There are 19.4 trillion possible combinations to choose from. With 1000 possible explanatory variables, there are 2.66×10^{23} combinations of up to 10 variables. With one million possible explanatory variables, the number of possibilities grows to 2.76×10^{53}.

Stepwise regression circumvents the computational burden of trying all possible combinations of explanatory variables, by testing variables, one by one, in each step. The use of forward-selection stepwise regression for identifying the 10 most statistically significant explanatory variables requires only 955 regressions if there are 100 candidate variables, 9955 regressions if there are 1000 candidates, and slightly fewer than 10 million regressions if there are one million candidate variables. This simplification is very appealing, and many researchers working with Big Data have succumbed to the appeal of stepwise regression.

However, the more variables that are considered, the more likely it is that coincidental statistical relationships will be discovered. Calude and Longo [33] prove that

> *the more data, the more arbitrary, meaningless and useless (for future action) correlations will be found in them. Thus, paradoxically, the more information we have, the more difficult is to extract meaning from it. Too much information tends to behave like very little information.*
>
> *If there is a fixed set of true statistical relationships that are useful for making pre-*

dictions, the data deluge necessarily increases the ratio of meaningless statistical relationships to true relationships.

The fundamental problem with the notion that data come before theory is simple: We think that patterns are unusual and therefore meaningful; in Big Data, patterns are inevitable and therefore meaningless.

Stepwise regression steps—indeed leaps—into this trap. It follows automated rules that only consider statistical correlations, with no regard for whether it makes sense to include a potential explanatory variable. It is data without theory. It is data mining on steroids.

Methods

A Monte Carlo simulation model can be used to demonstrate the core problem with stepwise regression and how the problem is exacerbated in large data sets.

Steyerberg et al. [34] argue that stepwise models do poorly in small data sets, an argument they illustrate by applying stepwise regression to subsets of a data set with 4, 8, or 16 explanatory variables (whose estimated coefficients are assumed to be the "true" values). Derksen and Keselman [35] analyze 250 simulations of a Monte Carlo model with 12, 18, or 24 candidate explanatory variables and conclude that stepwise regression often chooses the wrong explanatory variables. Done decades ago, when computer capabilities were modest, these tests were understandably limited to a small number of explanatory variables and simulations.

The simulations here use $n = 10$, 50, 100, 200,250, 500, or 1000 candidate explanatory variables. In each simulation, 200 observations for each candidate variable are determined either by random draws from a normal distribution,

$$\text{no drift:} \quad X_{i,t} = \varepsilon_{i,t} \quad \varepsilon \sim N[0, \sigma_x] \tag{1}$$

or by a Gaussian random walk,

$$\text{drift:} \quad X_{i,t} = X_{i,t-1} + \varepsilon_{i,t} \quad X_{i,0} = 0 \quad \varepsilon \sim N[0, \sigma_x] \tag{2}$$

where ε is normally distributed with mean 0 and standard deviation σ_x. In the non-drift model, the values of each variable are i.i.d; in the drift model, changes in the values of the explanatory variables are i.i.d.

Five randomly selected explanatory variables (the *true* explanatory variables) are used to determine the values of a dependent variable

$$Y_t = \alpha_0 + \sum_{i=1}^{5} \beta_i X_{i,t} + \upsilon_t \quad \upsilon \sim N\left[0, \sigma_y\right] \tag{3}$$

where the value of each coefficient is randomly determined from a uniform distribution ranging from -2 to $+2$, excluding -1 to $+1$, and υ is normally distributed with mean 0 and standard deviation σ_y. The range -1 to $+1$ was excluded so that none of the coefficients of the real variables would be approximately zero. The other $n-5$ candidate variables are *nuisance* variables that were determined independently and have no effect on Y, but might be coincidentally correlated with Y. The simulations use $\sigma_x = 5$ and $\sigma_y = 10$, 20, or 30.

Table 1 Average number of explanatory variables per equation, $\sigma_x = 5$

Candidates	No drift			Drift		
	$\sigma_y = 10$	$\sigma_y = 20$	$\sigma_y = 30$	$\sigma_y = 10$	$\sigma_y = 20$	$\sigma_y = 30$
10	5.25	4.74	3.45	5.59	5.54	5.13
50	7.51	6.99	5.68	8.14	7.81	6.52
100	11.25	10.71	9.44	10.03	9.19	7.31
200	27.67	26.75	25.34	13.03	11.03	8.43
250	56.75	56.78	55.59	14.45	11.88	9.00
500	97.64	96.82	94.91	32.26	21.04	14.85
1000	98.55	97.88	96.92	70.11	57.30	42.79

After generating the data for the explanatory variables and the dependent variable, a stepwise regression procedure was used with the n candidate variables evaluated in random order. At each step, the potential explanatory variable with the lowest two-sided p-value is added to the equation if this p value is less than 0.05. One million simulations were done for each parameterization of the model.

The central question is how effective stepwise regression is at identifying the true variables that determine Y, so that reliable predictions can be made with fresh data. So, in each simulation, 100 observations were used to estimate the stepwise model's coefficients, and the remaining 100 observations were used to test the model's reliability.

In practice, a stepwise regression procedure might sometimes select explanatory variables that, although not directly affecting the dependent variable, are systematically related to variables that do affect the dependent variable. For example, consumer spending depends on income, which is related to years of education. Even if education does not directly influence spending, it is a noisy proxy for income and might find its way into a stepwise regression equation. All the explanatory variables in these Monte Carlo simulations were generated independently (so that there are no proxy variables) in order to focus on the fact that stepwise regression might be fooled by purely coincidental correlations.

While they might be fortuitously correlated with the dependent variable during the estimation period, nuisance variables are useless out-of-sample because they are truly independent of the variable being predicted. The selection of nuisance variables by the stepwise regression procedure gives a false confidence in the estimated model because of the high t values and the boost they provide to R^2.

An extreme case (that did happen in some simulations) is when all of the explanatory variables chosen by the stepwise procedure are nuisance variables. Although there might be a great fit during the estimation period, the prediction errors will be large out-of-sample because the dependent variable will be predicted based solely on the values of irrelevant variables. There are less extreme consequences in less extreme cases, but when nuisance variables are included in the stepwise equation, we should anticipate that the prediction errors will be larger out-of-sample than in-sample.

Table 2 Frequencies for an included variable being a nuisance variable, $\sigma_x = 5$

Candidates	No drift			Drift		
	$\sigma_y = 10$	$\sigma_y = 20$	$\sigma_y = 30$	$\sigma_y = 10$	$\sigma_y = 20$	$\sigma_y = 30$
10	0.048	0.055	0.073	0.106	0.113	0.131
50	0.335	0.371	0.464	0.388	0.429	0.497
100	0.556	0.599	0.693	0.507	0.570	0.645
200	0.820	0.846	0.891	0.628	0.704	0.768
250	0.912	0.927	0.948	0.667	0.745	0.804
500	0.949	0.963	0.975	0.860	0.890	0.909
1000	0.959	0.967	0.986	0.945	0.967	0.970

Table 3 Frequencies for the number of true variables selected, $\sigma_x = 5$ and $\sigma_y = 20$

Candidates	No drift				Drift			
	< 3	3	4	5	< 3	3	4	5
10	0.016	0.082	0.302	0.600	0.009	0.017	0.018	0.955
50	0.025	0.103	0.324	0.549	0.079	0.088	0.090	0.743
100	0.035	0.125	0.341	0.498	0.175	0.145	0.137	0.542
200	0.057	0.167	0.359	0.417	0.331	0.198	0.151	0.320
250	0.049	0.177	0.358	0.416	0.388	0.207	0.148	0.256
500	0.154	0.255	0.329	0.262	0.580	0.203	0.110	0.108
1000	0.320	0.240	0.260	0.180	0.711	0.160	0.088	0.041

Results

Table 1 shows the average number of explanatory variables selected by the stepwise regressions. These are not affected much by the standard deviations, at least for the range of values considered here. For $\sigma_x = 5$, $\sigma_y = 20$, and 100 candidate variables, the average number of explanatory variables chosen by the stepwise regression was 10.71 (no drift) or 9.20 (drift), and the longest stepwise equation among the one million simulations had 32 (no drift) or 29 (drift) explanatory variables. Overall, the average number of explanatory variables chosen by stepwise regression increases with the number of candidate variables.

Table 2 shows the frequency with which a variable included in the final stepwise equation was, in fact, a nuisance variable. The standard deviations do not matter much, but the nuisance probability increases substantially as the number of candidate variables increases. Even with only 100 candidate variables, it is more likely than not that a variable chosen by the stepwise procedure is a nuisance variable, rather than a real variable.

Table 3 shows the frequency with which the true variables were selected. Because the standard deviations do not matter much, results are only shown for the case $\sigma_x = 5$ and $\sigma_y = 20$. For example, in the no-drift model with 100 candidate variables, all five true variables were included 49.5% of the time—which means that at least one true variable was excluded 50.5% of the time. In the drift model with 100 candidate variables, all five true variables were included 53.8% of the time—and at least one true variable was excluded 46.2% of the time. As the number of candidate variables increases, the chances that all five true variables will be included in the stepwise equation falls.

Table 4 In-sample and out-of-sample prediction errors, $\sigma_x = 5$ and $\sigma_y = 20$

Candidates	No drift				Drift			
	MAE		RMSE		MAE		RMSE	
	In	Out	In	Out	In	Out	In	Out
10	15.45	16.87	19.32	21.09	15.40	24.51	19.25	29.81
50	14.42	18.07	18.03	22.60	14.91	43.99	18.65	52.11
100	12.82	19.92	16.03	24.90	14.55	61.09	18.18	71.85
200	7.76	22.27	9.69	31.60	13.90	83.38	17.38	97.69
250	3.53	29.08	4.43	36.35	13.57	91.71	16.90	107.35
500	0.00	29.89	0.00	37.38	11.46	127.26	14.31	148.16
1000	0.00	31.12	0.00	38.28	6.04	182.83	7.57	212.64

Stepwise enthusiasts often claim that adding variables based on statistical significance will improve the model's predictions, by which they mean improve the fit for the data used to estimate the model. However, adding variables does not necessarily help, and may hurt, when a stepwise model is used to make predictions with fresh data.

Table 4 compares the in-sample and out-of-sample prediction errors using the mean absolute error (MAE):

$$MAE = \frac{\sum_{t=1}^{n} \left| \widehat{Y} - Y \right|}{n} \tag{4}$$

and the root mean square error (RMSE):

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n} \left(\widehat{Y} - Y \right)^2}{n}}. \tag{5}$$

The stepwise models consistently did substantially worse out-of-sample than in-sample. As the number of candidate variables increases, the in-sample fit improves, while the out-of-sample fit deteriorates, causing the ratio of the out-of-sample errors to the in-sample errors to balloon.

Out-of-sample validation

A model's weaknesses can be exposed by the deterioration of the model's fit using fresh data. It is therefore reasonable to hold out part of the available data for testing the estimated model [36, 37]. The two parts of the data are labeled *in-sample* and *out-of-sample* or, more recently, *training data* and *validation data*.

It is always a good idea to test a model with fresh data. However, choosing a data-mined model by using a repetitive cycle of in-sample estimation and out-of-sample testing does not guarantee that the best model will be chosen.

Tireless data mining guarantees that some models will fit both parts of the data remarkably well, even if none of the models are meaningful. Just as some models are certain to fit the in-sample data by luck alone, so some models are certain to fit the out-of-sample data as well. Uncovering a model that fits both the in-sample data and the

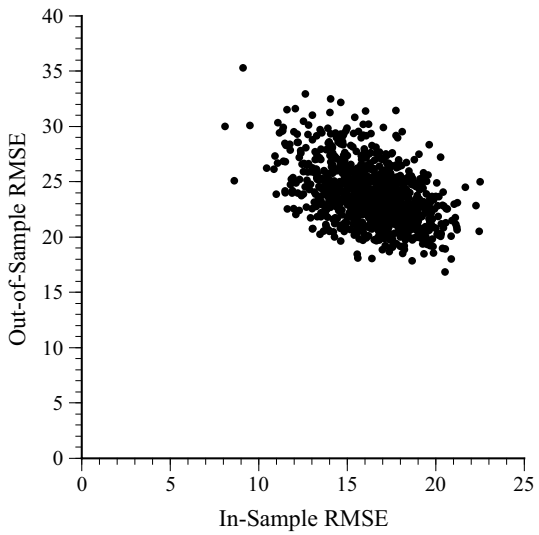

Fig. 1 No-drift model with 100 nuisance variables and no real explanatory variables

out-of-sample data is just another form of data mining. Instead of discovering a model that fits half the data, we discover a model that fits all the data. That doesn't solve the fundamental problem, which is that models that are chosen solely to fit the data, either half the data or all the data, cannot be expected to fit new data nearly as well.

To illustrate this point, the Monte Carlo simulations were redone using $\sigma_x = 5$, $\sigma_y = 20$, and 100 candidate variables, with *all* 100 candidate variables being nuisance variables. This way, one cannot argue that a stepwise model with a good in-sample and out-of-sample fit is a good model. Figure 1 shows the in-sample and out-of-sample RMSEs for the first 10,000 of 1 million simulations with no-drift data. (A figure with all 1 million simulations would be a dense blob). The average RMSE is 16.30 in-sample and 23.76 out-of-sample.

Figure 1 shows that, although the out-of-sample RMSEs are generally larger than the in-sample RMSEs, there are many simulations in which the out-of-sample RMSE is close enough to the in-sample RMSE to suggest that a good model has been discovered—when, in fact, all the models are just coincidental correlations. Specifically, the out-of-sample RMSE is *less* than the in-sample RMSE 2% of the time, and within 10% of the in-sample RMSE 8% of the time. For similar simulations with the drift model, the out-of-sample RMSE is less than the in-sample RMSE 5% of the time, and within 10% of the in-sample RMSE 15% of the time.

A persistent data miner would have no trouble finding a model that performs almost as well out-of-sample as in-sample, even though the model is useless because the variable being predicted is only coincidentally related to the explanatory variables.

Discussion

In addition to stepwise regression, several other feature selection methods have been proposed to deal with the curse of dimensionality, which can be computationally demanding and lead to overfitting and inaccurate out-of-sample predictions due of

the inclusion of nuisance variables. For example, good results have been reported with recursive feature elimination with cross-validation [38, 39] and regularized tree ensembles [40], which are two efficient ways of identifying a parsimonious set of predictors. One of the particular strengths of recursive feature elimination with cross-validation is that a feature selection method is most likely to be successful when it is validated with out-of-sample data.

The stepwise simulations reported here confirm the value of using theoretical arguments or expert opinion to select the initial list of predictors. The stepwise regression models are much more successful when the procedure begins with 5 true variables and 5 nuisance variables than with 5 true variables and hundreds of nuisance variables.

One appealing way to deal with ambiguous theory is to use a Bayesian approach that explicitly allows uncertainty about the relevance of potential predictors and does not force a binary choice between inclusion and exclusion. Bayesian regression combines the data with a prior distribution for the model's parameters by using Bayes' theorem to derive a posterior distribution for the parameters and for predictions made with the model. As the amount of data increases, the posterior means converge to the least squares estimates. The computations can be challenging, but have now become practical. Detailed examples can be found in [41–43].

Ridge regression implicitly uses prior distributions for the coefficients of the explanatory variables that have zero means, identical variances, and are independent [44]. It seems unlikely that the coefficients of predictors chosen on the basis of expert opinion would have prior means of zero. It is more appealing to use explicit priors instead of implicit priors.

Conclusion

Stepwise regression selects explanatory variables for multiple regression models based on their statistical significance. Although it has often been criticized for the misapplication of single-step statistical tests to a multi-step procedure, stepwise regression has become popular with Big Data because it is a very efficient way of choosing a relatively small number of explanatory variables from a vast array of possibilities. The assumption is that the larger the number of possible predictors, the more useful is stepwise regression.

This paper uses Monte Carlo simulations to demonstrate that a stepwise procedure may choose nuisance variables rather than true variables and that the out-of-sample accuracy of the model may be far worse than the in-sample fit. These problems are more likely to be serious when there are a large number of potential predictors. Stepwise regression does not solve the problem of Big Data. Big Data exacerbates the problems of stepwise regression.

Authors' contributions
The author read and approved the final manuscript.

Authors' information
GS received his Ph.D. in Economics from Yale University and was an Assistant Professor there for 7 years. He is now the Fletcher Jones Professor of Economics at Pomona College. He has written (or co-authored) more than 80 academic papers and 13 books. His *Standard Deviation: Flawed Assumptions, Tortured Data, and Other Ways to Lie with Statistics* (Overlook/Duckworth, 2015) was a *London Times* Book of the Week and debunks a variety of dubious and misleading statistical practices. *The AI Delusion* (Oxford University Press, 2018) argues that, in this age of Big Data, the real danger is

not that computers are smarter than us, but that we think computers are smarter than us and, so, trust computers to make important decisions for us.

Acknowledgements
Not applicable.

Competing interests
The author declares that there is no competing interests.

Availability of data and materials
Not applicable (all data are from Monte Carlo simulations; the source code is available).

Consent for publication
Yes.

Ethics approval and consent to participate
Not applicable (no human participants).

Funding
None.

Publisher's Note

Received: 23 May 2018 Accepted: 5 September 2018
Published online: 15 September 2018

References

1. Efroymson MA. Multiple regression analysis. In: Ralston A, Wilf HS, editors. Mathematical methods for digital computers. New York: Wiley; 1960.
2. Thompson B. Why won't stepwise methods die? Meas Eval Couns Dev. 1989;21(4):146–8.
3. Hurvich CM, Tsai CL. The impact of model selection on inference in linear regression. Am Stat. 1990;44(3):214–7.
4. Harrell FE Jr. Regression modeling strategies: with applications to linear models, logistic regression and survival analysis. New York: Springer; 2001.
5. Hendry DF, Krolzig HM. Automatic econometric model selection. London: Timberlake Consultants Press; 2001.
6. Babyak MA. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. Psychosom Med. 2004;66:411–21.
7. Whittingham MJ, Stephens PA, Bradbury RB, Freckleton RP. Why do we still use stepwise modelling in ecology and behaviour? J Anim Ecol. 2006;75(5):1182–9.
8. Castle JL, Fawcett NWP, Hendry DF. Evaluating automatic model selection, Technical Report 474. Oxford: Department of Economics, University of Oxford; 2010.
9. Flom PL, Cassell DL. Stopping stepwise: why stepwise and similar selection methods are bad, and what you should use. In: NESUG 2007 proceedings. 2007.
10. Thompson B. Stepwise regression and stepwise discriminant analysis need not apply here: a guidelines editorial. Educ Psychol Meas. 1995;55:525–34.
11. Marascuilo LA, Serlin RC. Statistical methods for thesocial and behavioral sciences. New York: W. H. Freeman; 1988.
12. Huberty CJ. Problems with stepwise methods—better alternatives. In: Thompson B, editor. Advances in social science methodology, vol. 1. Greenwich: JAI Press; 1989.
13. Vlachopoulou M, Ferryman TA, Zhou N, Tong J. A stepwise regression method for forecasting net interchange schedule. https://doi.org/10.1109/pesmg.2013.6672763. 2013.
14. Walter S, Tiemeier H. Variable selection: current practice in epidemiological studies. Eur J Epidemiol. 2009;24(12):733–6.
15. Liao H, Lynn HS. A survey of variable selection methods in two Chinese epidemiology journals. BMC Med Res Methodol. 2010;10:87. https://doi.org/10.1186/1471-2288-10-87.
16. Rachev ST, Mittnik S, Fabozzi FJ, Focardi SM, Jašić T. Financial econometrics: from basics to advanced modeling techniques. New York: Wiley; 2006.
17. McDonald JH. Handbook of biological statistics. 3rd ed. Baltimore: Sparky House Publishing; 2014.
18. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. 2nd ed. New York: Springer; 2016.
19. Wiley. Wiley 11th hour study guide for level II CFA exam. 2nd ed. New York: Wiley; 2017. p. 31.
20. Friedman M. The permanent income hypothesis: a theory of the consumption function. Princeton: Princeton University Press; 1957.
21. Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. AI Mag. 1996;17(3):37–54.
22. Kecman V. Foreword. In: Cios KJ, Pedrycz W, Swiniarski RW, Kurgan LA, editors. Data mining: a knowledge discovery approach. New York: Springer; 2007.
23. Begoli E, Horsey J. Design principles for effective knowledge discovery from big data. In: Software architecture (WICSA) and European conference on software architecture (ECSA), 2012 joint working IEEE/IFIP conference.
24. Piatetsky-Shapiro G. Knowledge discovery in real databases: a report on the IJCAI-89 workshop. AI Mag. 1991;11(5):68–70.

25. Sagiroglu S, Sinanc D. Big data: a review. In: 2013 international conference on collaboration technologies and systems (CTS). 2013.
26. Kecman V. Foreword. In: Cios KJ, Pedrycz W, Swiniarski RW, Kurgan LA. Data mining: a knowledge discovery approach. New York: Springer; 2007.
27. Tullock G. A comment on Daniel Klein's "A plea to economists who favor liberty". East Econ J. 2001;27(2):203–7.
28. Wooldridge JW. Introductory econometrics: a modern approach. 3rd ed. Mason: Thompson; 2006. p. 94–7.
29. Stock JH, Watson MW. Introduction to econometrics. 2nd ed. Boston: Pearson; 2007. p. 316–9.
30. Hastie T, Tibshirani R, Friedman J. the elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York: Springer; 2009. http://www-stat.stanford.edu/~tibs/ElemStatLearn/download.html.
31. Varian HR. Big data: new tricks for econometrics. J Econ Perspect. 2014;28(2):3–27.
32. Bruce P, Bruce A. Practical statistics for data scientists: 50 essential concepts. Sebastopol: O'Reilly Media; 2017.
33. Calude CS, Longo G. The deluge of spurious correlations in big data. Found Sci. 2016. https://doi.org/10.1007/s10699-016-9489-4.
34. Steyerberg EW, Eijkemans MJC, Habbema JDF. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. J Clin Epidemiol. 1999;52(10):935–42.
35. Derksen S, Keselman HJ. Backward, forward and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables. Br J Math Stat Psychol. 1992;45(2):265–82.
36. Mayers JH, Forgy EW. The development of numerical credit evaluation systems. J Am Stat Assoc. 1963;58(303):799–806.
37. Mark J, Goldberg MA. Multiple regression analysis and mass assessment: a review of the issues. Apprais J. 2001;56:89–109.
38. Guyan I, Weston J, Barnhill S, Vopnik V. Gene selection for cancer classification using support vector machines. Mach Learn. 2002;46:389–422.
39. Mukherjee T, Duckat M, Kumar P, Paquet JD, Rodriguez D, Haulcomb M, George K, Pasiliao E. RSSI-based supervised learning for uncooperative direction-finding. In: Altun Y, editor. Machine learning and knowledge discovery in databases. ECML PKDD 2017, vol. 10536., Lecture Notes in ComputerCham: Springer; 2015.
40. Deng H, Runger G. Feature selection via regularized trees. In: Proceedings of the 2012 international joint conference on neural networks (IJCNN), IEEE; 2012.
41. Box GEP, Tiao GC. Bayesian inference in statistical analysis. New York: Wiley; 1973.
42. Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian data analysis. 2nd ed. Boca Raton: Chapman and Hall/CRC; 2003.
43. Koehrsen W. Introduction to Bayesian linear regression. Towards Data Science. 2018. https://towardsdatascience.com/introduction-to-bayesian-linear-regression-e66e60791ea7.
44. Smith G, Campbell F. A critique of some ridge regression methods. J Am Stat Assoc. 1980;75(369):74–81.