PROPERTIES OF BACKWARD ELIMINATION AND FORWARD SELECTION IN LINEAR REGRESSION

Milan Bašta

Abstract

Using Monte Carlo simulation, we study the properties of backward elimination and forward selection procedures which are widely used in classical linear regression for choosing variables to include in a regression model. Among others, we explore whether the procedures are capable of identifying authentic explanatory variables and eliminating the noise ones, whether classical least-squares inference theory is valid in the final selected model and whether the final selected model is superior to the full model in terms of prediction accuracy. We conclude that the use of the procedures is not advocated if classical inference is to be employed in the final selected model, while the procedures can potentially be useful for predictive purposes in some situations. The findings we have obtained are in agreement with similar studies available in the literature and have important implications for the practice of data analysis. Ignoring them can have severe consequences and can be interpreted as misuse/abuse of statistics.

Key words: backward elimination, forward selection, linear regression, Monte Carlo simulation

JEL Code: C15, C2

Introduction

In linear regression a researcher often faces the situation where she has several explanatory variables at hand and has to decide which of them are *authentic* variables in the sense that the true parameters on the variables in the regression model are *non-zero* and which of them are *noise* variables in the sense that the true parameters on these variables are *zero*. A different objective the researcher can come up against is selecting only a group of explanatory variables as predictors so that the predictions of the response variable based on the corresponding fitted regression model are as accurate as possible. In both the above situations the researcher is confronted with the task of *subset selection*.

Best subset regression – being one of the approaches to subset selection – considers each possible subset, selects the "best" one according to a suitable criterion (such as the adjusted coefficient of determination, AIC, etc.) and fits the regression model using the selected subset.

Forward selection or backward elimination are greedy algorithms that do not search among all possible subsets but consider a path through them and check only the subsets along this path (Friedman et al., 2009).

Forward selection and backward elimination procedures have been subject to various critical remarks and scrutiny in the literature (see, e.g., Harrell, 2001; Wittingham et al., 2006; Derksen and Keselman, 1992; Berk et al., 2010), pointing out, among others, that the test statistics used in the procedures do not have the claimed distributions, p-values do not have their proper meaning, estimated standard errors of the estimated regression parameters are biased low, confidence intervals are narrow, results differ for backward elimination and forward selection procedures and that researchers inappropriately rely on the final model.

Our aim is to inspect various aspects of the backward elimination and forward selection procedures using Monte Carlo simulations. Monte Carlo studies similar to that of ours are available in the literature (e.g., Derksen and Keselman, 1992; Molodkina, 2014), though some of the aspects we examine have not been reported yet. Moreover, the settings of our simulations differ from those presented in the literature.

We introduce the backward elimination and forward selection algorithms in Section 1. Section 2 studies their properties using Monte Carlo simulations. The final section of the paper concludes. The analysis and presentation of the results are rather brief so that they fit the scope and the extent of the papers to the conference proceedings.

1 Backward elimination and forward selection algorithms

We assume one quantitative response variable and a set of k explanatory variables (some of which are authentic and some of which are noise variables). For simplicity, we assume that each of the k explanatory variables is quantitative and represented by a single column in the model matrix. A *full model* is such a model which contains all the k explanatory variables. A *final model* is the model which contains only those explanatory variables that have been selected by the backward elimination (or forward selection) procedure.

We perform *backward elimination* as follows: We start with the full model which includes all the k explanatory variables and estimate its parameters using least squares. Further, we remove the explanatory variable with the largest p-value associated with the corresponding

two-tailed t-test provided that the p-value is above a given threshold α_{remove}, such as 0.05. The model with the remaining $(k-1)$ explanatory variables is re-estimated and, again, the explanatory variable with the largest p-value is removed if the p-value exceeds the threshold. The procedure is repeated until all the explanatory variables remaining in the model have p-values less than or equal to α_{remove}.

We perform *forward selection* as follows: We start with a null model with no explanatory variables. We consecutively estimate the parameters of k straight line regression models using least squares, where each of the k explanatory variables is used as a single explanatory variable in the model. In each of the models, we evaluate the two-tailed t-test which "assesses the relevance of the variable in the model" and save the p-value of the test. Further, the explanatory variable with the lowest p-value is permanently included into the model if this p-value is less than or equal to a given threshold α_{add}, such as 0.05. Afterwards, $(k-1)$ regression models with two explanatory variables are successively estimated using least squares, the first variable being the one which has permanently been included into the model in the previous step, the second variable being successively each of the $(k-1)$ remaining explanatory variables which are not yet permanently included in the model. The p-values associated with the corresponding two-tailed t-tests on these second variables are obtained. Further, the variable with the lowest p-value is permanently included into the model provided the p-value is less than or equal to α_{add}. The above procedure is repeated until the lowest p-value of the two-tailed t-test is larger than α_{add}.

In our context (with quantitative explanatory variables used linearly), it is easy to see that deciding on the removal (or inclusion) of variables using the two-tailed t-tests is equivalent to using one-tailed partial F-tests and is also equivalent to removing (including) variables based on the least increase (largest decrease) of the residual sum of squares. As a result, our backward elimination and forward selection procedures are comparable to those presented e.g. in James et al. (2013), Yan and Gang Su (2009) or Rawlings et al. (1998).

2 Monte Carlo simulation

We assume a linear regression model (with no intercept)

$$Y_i = \beta_1 x_{i1} + \ldots + \beta_m x_{im} + \ldots + \beta_k x_{ik} + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{1}$$

where n is the number of observations, Y_i is the response variable for observation i, $x_{i1}, \ldots, x_{im}, \ldots, x_{ik}$ are known constants (values of k explanatory variables for observation i), ε_i is error for

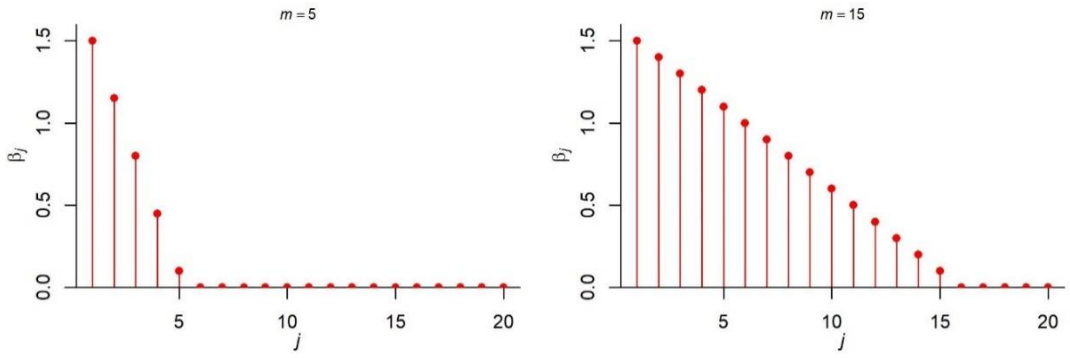
observation *i*. $\beta_1, \ldots, \beta_m, \ldots, \beta_k$ are regression parameters and $1 \leq m \leq k$ is the number of non-zero regression parameters. The model of Equation 1 is the full model, the number of authentic explanatory variables being *m*, the number of noise variables being $k - m$.

The $[\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n]^T$ vector has an *n*-variate normal distribution with zero mean vector and covariance matrix $\sigma^2\mathbf{I}_n$ where $\mathbf{I}_n$ is an identity matrix of order *n* and σ^2 is a constant. The (sample, empirical) variance of the values of each explanatory variable is exactly equal to 1, the (sample, empirical) correlation between values of any pair of distinct explanatory variables being exactly equal to ρ. The histogram of the values of each explanatory variable is close to a Gaussian one.

Further, we focus on backward elimination and set $\alpha_{remove} = 0.05$, $k = 20$ and $\sigma^2 = 6.25$. We explore $2^3 = 8$ possible settings differing in the values of *n*, *m* and ρ. Specifically, we examine:

- two possible values of *n*: $n = 50$ or $n = 300$,
- two possible values of *m:* $m = 5$ or $m = 15$ (the regression parameters $\beta_1, \ldots, \beta_m, \ldots, \beta_k$ for $m = 5$ are presented in the left plot of Figure 1, whereas those for $m = 20$ in the right one),
- and two possible values of ρ: $\rho = 0$ or $\rho = 0.8$.

Fig. 1: Regression parameters for *m* = 5 (left) and *m* = 15 (right)



Source: Author's construction

The above simulation settings were chosen so that we can compare various situations we can encounter in real-life analysis (smaller and larger sample size, few and many authentic explanatory variables, weak and strong correlation between explanatory variables). Moreover, the settings we assume allow us to see the weakness and strengths of the procedures well. Of course, other settings could be of interest such as those differing in the value of σ^2. This is, however, prevented by the extent of the paper and would hinder the clarity of the presentation

of the simulation outcomes. Though not presented in the paper, we have checked that under other simulation settings qualitatively similar conclusions would be reached.

The number of Monte Carlo simulations for each setting is 1000, giving an upper bound on the standard error of relative frequency (i.e., probability estimator) of $0.5/\sqrt{n} = 0.016$. The estimated standard errors of the various bias or mean estimators are bounded from above by 0.06. R software (R Core Team, 2018) has been used to perform the analysis.

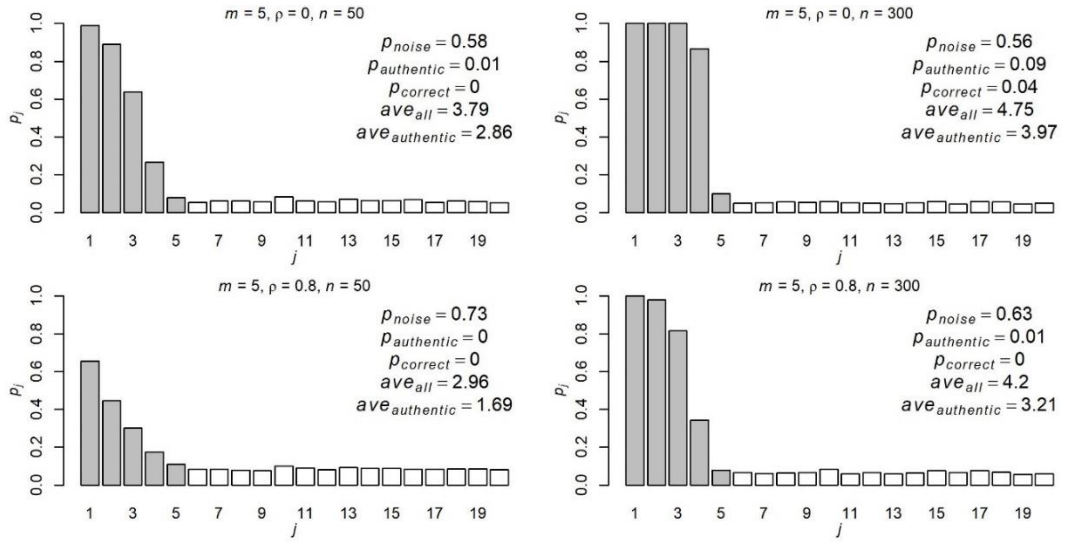
In the next sections various facets of backward elimination are inspected. As the extent of the paper is limited, forward selection is discussed briefly only in the Conclusion.

2.1 Nature of the final models

The following quantities are calculated within the Monte Carlo simulation in this section:

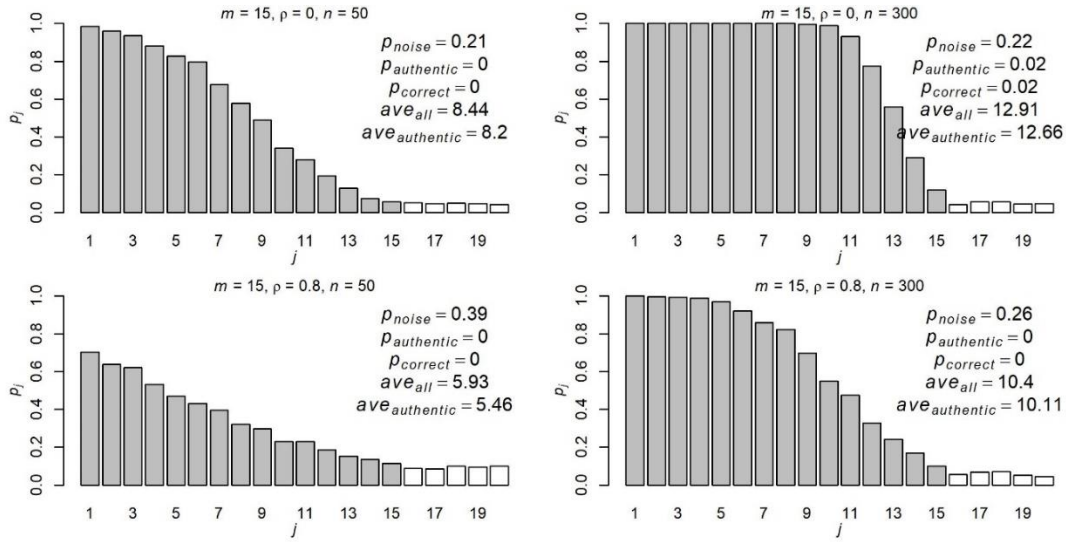
- The relative frequency (RF) of obtaining a final model which contains at least one noise explanatory variable (p_{noise}), RF of obtaining a final with all the authentic explanatory variables ($p_{authentic}$), RF of obtaining a final model with all the authentic explanatory variables and no noise ones ($p_{correct}$), and RF of obtaining a final model which includes the jth explanatory variable (p_j, for $j = 1, \ldots, k$). These relative frequencies are estimates of the corresponding true probabilities and are presented in Figure 2 and 3 (results are rounded to 2 decimal places).
- The average number of all explanatory variables (ave_{all}) and the average number of authentic explanatory variables ($ave_{authentic}$) in the final model. These averages provide us with the estimates of the expected number of all and authentic explanatory variables in the final model and are presented in Figure 2 and 3 (results are rounded to 2 decimal places).

Fig. 2: p_j, p_{noise}, $p_{authentic}$, $p_{correct}$, ave_{all} and $ave_{authentic}$ for various settings with $m = 5$. The grey (white) rectangles correspond to authentic (noise) explanatory variables.



Source: Author's construction

Fig. 3: p_j, p_{noise}, $p_{authentic}$, $p_{correct}$, ave_{all} and $ave_{authentic}$ for various settings with $m = 15$. The grey (white) rectangles correspond to authentic (noise) explanatory variables.



Source: Author's construction

We see that the probability of obtaining a final model with at least one noise explanatory variable depends on the simulation setting and can generally be considerable. On the other hand, the probability of obtaining a model which contains all the authentic explanatory variables or a model which contains all the authentic explanatory variables and no noise one is very tiny for all the settings. The expected number of authentic explanatory variables in the final model differs across the settings and can be below m by a sizeable amount.

2.2 Inference

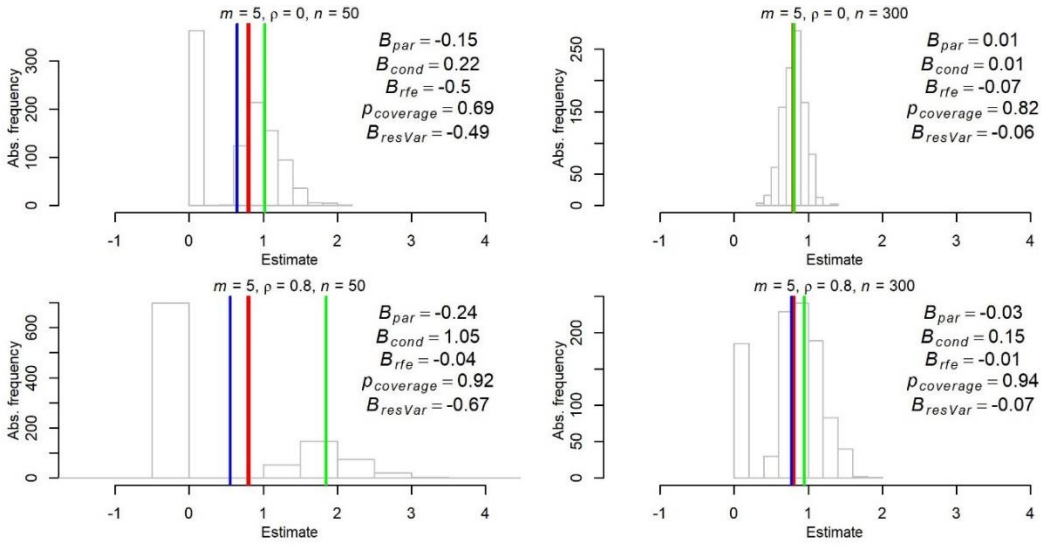
We calculate the following quantities within the Monte Carlo simulation, presenting the results in Figure 4 and 5:

- B_{par}, the difference between the sample average of b_3 (for $m = 5$) or b_8 (for $m = 15$) and $\beta_3 = 0.8$ (for $m = 5$) or $\beta_8 = 0.8$ (for $m = 15$). B_{par} provides us with an estimate of the bias of the regression parameter estimator.
- B_{cond}, the difference between the sample average of b_3 (for $m = 5$) or b_8 (for $m = 15$) and $\beta_3 = 0.8$ (for $m = 5$) or $\beta_8 = 0.8$ (for $m = 15$), calculated only across those final models where the third (eighth, for $m = 15$) explanatory variable was included in the final model. B_{cond} provides us with an estimate of the bias of the regression parameter estimator *given that* the third (eighth) explanatory variable is included in the final model.
- B_{rfe}, defined as the difference between the sample average of regression function estimates at point $[1, 1, \ldots, 1]^T$ and the true regression function at the point. It is an estimate of the bias of regression function estimator at point $[1, 1, \ldots, 1]^T$.
- $p_{coverage}$, the relative frequency of covering the true regression function at point $[1, 1, \ldots, 1]^T$ by a nominal 95% confidence interval for the regression function calculated from the final model according to the classical theory as if the model was prespecified in advance. $p_{coverage}$ can be used to estimate the corresponding probability of covering the true regression function with the confidence interval.
- B_{resVar}, the difference between the sample average of residual variances and $\sigma^2 = 6.25$. It provides an estimate of the bias of residual variance.

The red, blue and green lines in Figure 4 (Figure 5) capture β_3 (β_8), the sample average of b_3 (b_8) and the sample average of b_3 (b_8) given that the third (eighth) explanatory variable is in the final model. The grey histogram in Figure 4 (Figure 5) represent the distribution of b_3 (b_8).

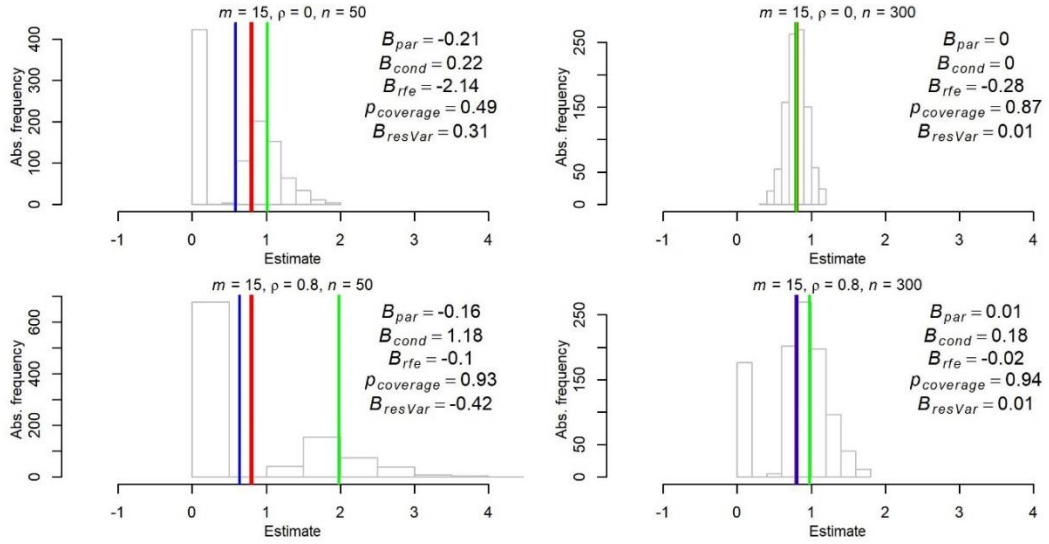
Even though the results vary across simulation settings we can draw some common conclusions. We see that estimators of regression parameters, regression function and error variance are biased. Further, given that an explanatory variable is included in the final model, the estimated effect on the variable is inflated on average. Confidence intervals (for regression function) calculated according to the least squares theory as if the final model was prespecified in advance do not generally have the required nominal coverage.

Fig. 4: B_{par}, B_{cond}, B_{rfe}, $p_{coverage}$ and B_{resVar} for various settings with $m = 5$.



Source: Author's construction

Fig. 5: B_{par}, B_{cond}, B_{rfe}, $p_{coverage}$ and B_{resVar} for various settings with m = 15.



Source: Author's construction

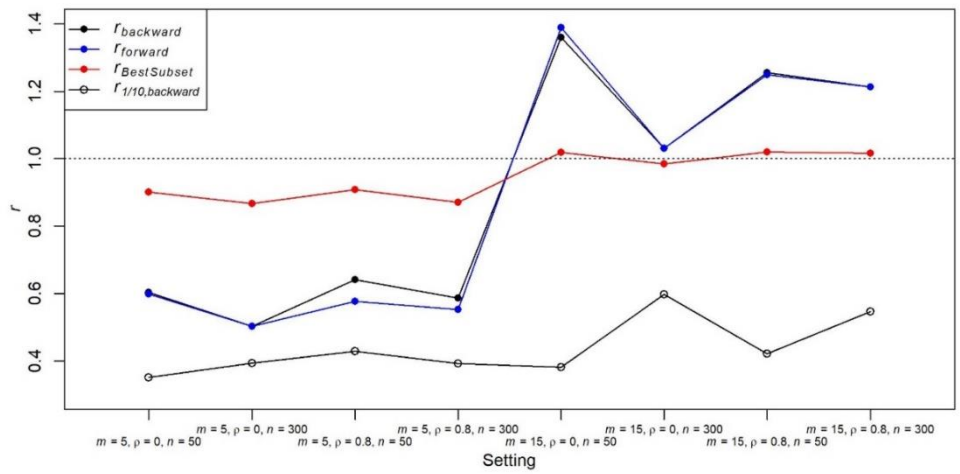
2.3 Mean squared error of regression function estimator

The fitted values from the full as well as from the final model can be considered as estimators of regression function. Let $\Delta_{full} = \sigma^2 k/n$ be the *average* (over n observations) mean squared error of the estimator for the full model and let $\Delta_{backward}$ be the *average* (over n observations) mean squared error of the estimator for the final model from backward elimination. Further, let $R_{backward}$ be the ratio of $\Delta_{backward}$ to Δ_{full}. Values of $R_{backward}$ less than 1 imply that backward elimination performs better (in terms of mean squared error) in regression function estimation than using the full model. Consequently, it also performs better in prediction since the mean

squared error of prediction (at any point) is larger by σ^2 than the mean squared error of regression function estimation regardless of the model used. $R_{backward}$ will be estimated by Monte Carlo simulation, the estimate denoted by $r_{backward}$.

We define $R_{forward}$ and $R_{BestSubset}$ in an analogous way as a ratio of $\Delta_{forward}$ to Δ_{full} and a ratio of $\Delta_{BestSubset}$ to Δ_{full}, where $\Delta_{forward}$ and $\Delta_{BestSubset}$ is the average (over n observations) mean squared error of fitted value for the final model from forward selection procedure and for the final model from best subset regression. In best subset regression, the adjusted coefficient of determination is used as a criterion to select the optimal model. The Monte Carlo estimates of $R_{forward}$ and $R_{BestSubset}$ will be denoted as $r_{forward}$ and $r_{BestSubset}$. $r_{1/10,backward}$ is an estimate of $R_{1/10,backward}$ where $R_{1/10,backward}$ corresponds to the ratio of average mean squared errors (for backward elimination) in the case where one tenth of the original size is used for the true regression parameters.

Fig. 6: $r_{backward}$, $r_{forward}$, $r_{BestSubset}$ *and* $r_{1/10,backward}$ *for various settings*



Source: Author's construction

It follows that backward elimination can improve the prediction accuracy compared to the full model in some situations (many noise variables or variables with weak effects). More scrutiny is, however, needed to understand the benefits of the procedure for prediction objectives.

Conclusion

The belief that backward elimination is capable of detecting a model where all the authentic and no noise variables are included turned out to be faulty. Unfounded is also the conviction that results from classical least squares estimation theory can be enforced on the final model.

Consequently, if inference using classical least squares theory is the goal, the analysis within the full model without any removal of variables is advocated. The only situation we are currently aware of where backward elimination could potentially be fruitful is for improving predictions.

Though not reported explicitly, results analogous to those for backward elimination have been obtained for forward selection. We consider the findings important for statistical data analysis. Ignoring them can have severe consequences comparable to other instances of misuse or misinterpretation of statistics (see, e.g., Goodman, 2008; Ioannidis, 2005).

References

Berk, R., Brown, L., & Zhao, L. (2010). Statistical inference after model selection. *Journal of Quantitative Criminology*, *26*(2), 217-236.

Derksen, S., & Keselman, H. J. (1992). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, *45*(2), 265-282.

Friedman, J., Hastie, T., & Tibshirani, R. (2009). *The elements of statistical learning*. 2nd Edition. New York: Springer series in statistics.

Goodman, S. (2008). A dirty dozen: twelve p-value misconceptions. In *Seminars in hematology*. 45(3), 135-140. Elsevier.

Harrell, F. E. (2001). *Regression modeling strategies, with applications to linear models, survival analysis and logistic regression*. Springer.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, *2*(8), e124.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.

Molodkina, K. (2014). *Krokové metody v lineární regresi a jejich vlastnosti*. Bakalářská práce. MFF UK.

R Core Team (2017), R. *A Language and Environment for Statistical Computing*. Vienna, Austria. https://www.R-project.org/

Rawlings, J. O., Pantula, S. G., & Dickey, D. A. (1998). *Applied regression analysis: a research tool*. Springer.

Whittingham, M., Stephens, P., Bradbury, R., & Freckleton, R. (2006). Why do we still use stepwise modelling in ecology and behaviour? *Journal of animal ecology*, *75*(5), 1182-1189.

Yan, X., & Su, X. (2009). *Linear regression analysis: theory and computing*. World Scientific.

Contact

Milan Bašta

Department of Statistics and Probability, Faculty of Informatics and Statistics, University of Economics in Prague

nám. W. Churchilla 1938/4

130 67 Praha 3 – Žižkov

milan.basta@vse.cz