Commentary on Exact Post-selection Inference for Sequential Regression Procedures

Lawrence D. Brown and Kory D. Johnson[1]

[1] *The Wharton School, University of Pennsylvania*

February 12, 2016

1 Introduction

The authors provide a novel and exciting framework for analyzing conditional selection. Formalizing the steps of a selection procedure as constraints on the response is applicable beyond the linear models theory discussed here and involves a high degree of technical accomplishment. It also raises interesting questions about different approaches to conditional inference. That being said, the usefulness of these tests appears limited in practice.

For ease of exposition, we focus on the forward stepwise case, though the arguments are also applicable to LARS. The authors' propose an "Exact Forward Stepwise" procedure (FS) that assigns new, "exact" p-values to the variables in a standard forward selection algorithm based on the usual ANOVA forward selection. At each stage, the algorithm includes the variable which creates the largest reduction in error sum of squares. After a variable is added, it is assigned a "p-value" by this "exact" procedure. This is a numerical quantity that has a U(0,1) distribution conditional on the sign of the selected variable and the variables that have been previously chosen.

This extends conditional inference ideas and calculations from other recent papers so as to provide p-values for feature selection algorithms. As these algorithms are commonly used to build multiple regression models, one might think that improved p-values would lead to improved model selection, at least in some circumstances; however, the formulation in this paper involves a serious paradox. One needs to begin with a well-specified model selection algorithm and construct a model independent of the exact p-values described in the paper. The exact p-values can be constructed only after the model has been chosen; they cannot validly be used to select the model. If one tries to use them in this way, they become invalid.

While forward stepwise and LARS operate independently of these p-values, one would expect the modeler to want to use the p-values to determine the step at which to "stop" the procedure and provide a final model. Consider the authors' Table 1 (recreated below), which compares their FS p-values to naive forward stepwise p-values. Identifying a final model using such a table requires considering multiple p-values from separate steps of the procedure. Therein lies the problem: the set of exact p-values cannot be used to make decisions, else they are invalid. Even using these p-values as input into a secondary FDR-controlling procedure as in G'Sell et al. (2015) is inappropriate. Only one exact p-value can validly be used, testing one step of a much larger procedure.

It should be noted that the conventional p-values are single-step values. They do not correct for the multiple testing nature of a stepwise procedure. Later in this commentary we recommend modified versions of the p-value calculations that can be validly and directly used for stepwise selection. See procedures (c) and (b) defined below. The procedure ES, defined below, is built on conditional inference logic and could be used to replace FS. For reasons discussed below, however, we do not favor its use.

The columns in Table 1 labeled "JASA" are taken directly from the authors' Table 1. The column labeled "Seq. p-value" contains traditional 2-sided p-values we calculated from our version of the data. Further information on our computations is in the appendix. Note that our p-values are 2-sided, whereas those in the paper are 1-sided. While providing 1-sided p-values may aid numerical comparison to the FS exact p-values, we note that these are one-sided conditional on the sign of the observed effect. Therefore, they are in effect 2-sided p-values and should be compared to ordinary 2-sided p-values. (We believe our

This research was supported in part by an NSF grant.

Table 1: Replicated Stepwise Table

Step	Parameter	Seq. p-value	FS, naive (JASA)	FS, Exact (JASA)
1	lcavol	0.0000	0.000	0.000
2	lweight	0.0003	0.002	0.006
3	svi	0.0424	0.024	0.425
4	lbph	0.0468	0.023	0.168
5	ppg45	0.2304	0.116	0.423
6	lcp	0.0878	0.041	0.273
7	age	0.1459	0.069	0.059
8	gleason	0.8839	0.442	0.156

"Seq. p-value" entries should be twice those in the column "FS, naive (JASA)." They are not exactly so but are numerically close to that.)

The paradox in using the FS p-values is rather subtle, and is easiest to explain in the context of an example. Let X_i be independently distributed $N(\theta_i, 1)$, for $i \in \{1, 2\}$. The forward selection problem is equivalent to determining an order for testing $H_{0,i} : \theta_i = 0$, while controlling false rejections at level α. Since we are performing model selection, a variable is "included" or "added" to the model when the corresponding null hypothesis is rejected. Allowing correlated variables does not change our discussion, it merely complicates the exposition and graphs. Similarly, without loss of generality, let $X_1 > X_2 > 0$.

The authors' FS significance thresholds are given as "FS Step 1" and "FS Step 2" in Figure 1. The conditioning set for both steps of the procedure is the same: $\{X_1 > X_2 > 0\}$. Values to the right of the curve "FS Step 1" (in red) yield p-values below α when testing $H_{0,1}$ while values between "FS Step 1" and the blue, 45° line yield p-values greater than α. Thus, values to the right of FS Step 1 are those for which the statistician using FS p-values would select X_1 with a positive sign at the first step of the selection process. During the second step, values above the curve "FS Step 2" (in gold) are significant at level α, while values below are not. Note that the calculation at the second step does not change depending on whether or not $H_{0,1}$ was rejected.

In order to use the FS p-values as Table 1 would imply, testing $H_{0,2}$ must account for rejecting $H_{0,1}$. Following the methodology of the authors' paper, this requires updating the conditioning set. We propose the following corrected procedure, "Exact Stepwise" (ES), that terminates on the first step in which a corrected, conditional p-values is above α. If $H_{0,2}$ is only tested when $H_{0,1}$ is rejected by FS, then the conditioning set is the region to the right of FS Step 1. Those points to the right of FS Step 1 and outside the convex, parabolic region whose boundary is the curve "ES Step 2" (in green) are those for which the new ES procedure selects X_1 at the first step (with a positive coefficient) and X_2 at the second step (with positive coefficient). It is clear that this correction does not invalidate the authors' methodology, but it does yields different p-values. Furthermore, the new conditioning sets are not polyhedral and need not be convex. In spite of the authors' indirect implication, convex polyhedral conditioning regions are not required for their methodology, although computations are simpler for such regions.

2 Stopping Procedures Using ES P-values

We are also concerned with the counter-intuitive results given when using conditional p-values, even when they are corrected as above. The problem has obvious symmetries such as relabeling variables 1 and 2 or changing their signs. While our new proposal, ES, preserves those symmetries, it does not preserve the natural monotonicity of the problem. For example, there exist values (x_1, x_2) and (x_1', x_2') for which $x_1 \leq x_1'$ and $x_2 \leq x_2'$, but for which ES selects both variables at (x_1, x_2) and no variables at (x_1', x_2'). The authors' FS procedure does not produce as extreme an example since $H_{0,2}$ is tested regardless of the result of testing $H_{0,1}$; however, the significance of the test of $H_{0,1}$ depends on the value of X_2. This is particularly troubling given that X_1 and X_2 are independent.

It is also instructive to compare the rejection regions of the FS and ES procedures to those of more traditional methods (again, see Figure 1). The conventional procedure first adds X_1 if $|X_1| > |X_2|$ and $|X_1| > \Phi^{-1}(\alpha/2)$. It then adds X_2 if, also, $|X_2| > \Phi^{-1}(\alpha/2)$. If $|X_2| > |X_1|$ and $|X_2| > \Phi^{-1}(\alpha/2)$ the first step adds X_2, etc. Relevant portions of the lines in pink form the boundaries of this region. There are multiple classically constructed stepwise regions to control for selection and multiple comparisons:

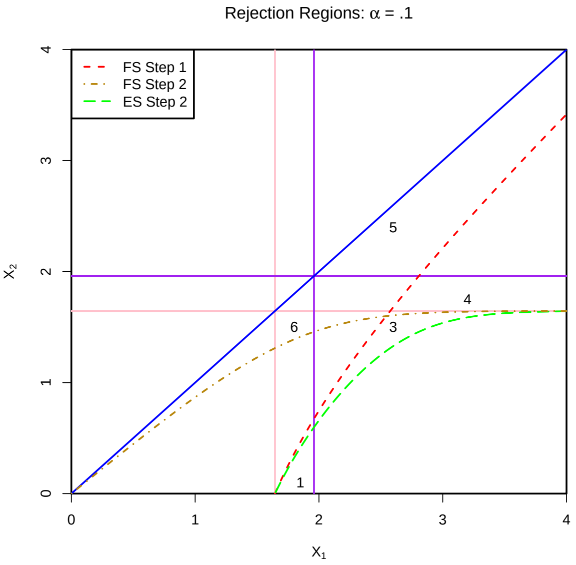

Figure 1: Stepwise rejection regions at $\alpha = .1$. The full picture is symmetric around the x- and y-axes. A corresponding image would be drawn if $X_2 > X_1 > 0$, in which case the graph would be rotated 90° and maintain its symmetries.

a) The pink lines are at $X_1 = 1.645$ and $X_2 = 1.645$. To the right of the blue 45° line, they show regions where the fully classical stepwise procedure would first choose X_1 (to the right of $X_1 = 1.645$) and then X_2 (above $X_2 = 1.645$). This region is not adjusted for multiple comparisons, and hence has an error probability of choosing non-empty models under the null hypothesis that exceeds the nominal level α.

b) Similarly, the purple lines at $X_1 = 1.96$ and $X_2 = 1.96$ bound regions which provide conservative multiple-comparison adjustments based on conventional single-coordinate p-values. This uses the Bonferroni approximation to control for multiple-comparisons. The exact numerical value can be computed from a maximum modulus calculation and is 1.948.

c) A better stepwise procedure controlling for multiple comparisons can be constructed as follows: at each step, choose among the remaining k variables using a p-value threshold such that the null probability of choosing any model is less than or equal to α. As in b), Bonferroni yields the conservative threshold α/k, though an exact calculation is possible when k is small. On the figure, one would include X_1 to the right of $X_1 = 1.96$ (or 1.948 for an exact calculation) and then would include X_2 when X_2 is above $X_2 = 1.645$. At each step of the procedure, the conditional probability under the null hypothesis of continuing with an incorrect rejection is α. This type of procedure was briefly proposed in Buja and Brown (2014). If the goal is to preserve FDR, then one can improve the procedure, and we are currently working on a paper to explain how to do so. A similar procedure can be efficiently computed that controls mFDR (Johnson et al., 2015a).

Many interesting comparisons can be made between FS, ES, and the more conventionally motivated, multiplicity-corrected procedures b) and c). These regions are labeled (numerically) in Figure 1.

1. Consider the triangular region to the right of FS Step 1 and to the left of $X_1 = 1.96$. This is where the ES procedure selects X_1 and the conventionally motivated procedures choose no variables. Heuristically, this seems to be a success for the ES procedure.
2. Within the region described in 1, there is a sliver between FS Step 1 and ES Step 2. Here, ES selects both X_1 and X_2, while the conventional procedures select neither. While this maintains a significance guarantee, this may not be an advantage. These points do have conventional (2-sided) p-values for X_1 that are below α, but the conventional p-value for X_2 is quite large. Selecting X_2 appears to be a mistake.
3. There is a more noticeable triangular region bounded by FS Step 1, ES Step 2, and $X_2 = 1.645$. In this region, the ES procedure selects both X_1 and X_2 but b) and c) select only X_1. For reasons similar to those in 2, the second step of ES appears undesirable. The disadvantage is not as clear, however, since X_2 can have a 2-sided p-value as small as $\alpha = .1$ in this region. The uncorrected FS p-value yields intuitively more satisfactory results in this region.

4. Consider the region between $X_2 = 1.645$ and $X_2 = 1.96$, and to the right of FS Step 1. ES and c) select both variables, but the simpler, conservative procedure b) does not include X_2. The advantage here goes to ES and c).
5. The area between the 45° line and FS Step 1 and to the right of $X_1 = 1.96$ is where b) and c) have a clear advantage in power relative to ES or to a procedure based on FS. In those regions, ES and FS have first step p-values above α and hence do not select any variable, while b) and c) always select X_1 and often select X_2.
6. In the region above FS Step 2 and below $X_2 = 1.645$, X_2 has a significant FS p-value even though its conventional p-value can be close to 1 near the origin.

In summary, b) or c) seem preferable to the ES procedure. The latter does better if the data fall in the small, but not negligible region 1; however, b) and c) produce much more reasonable models in the more noticeable area 5. Procedure c) is preferred to b) because of the difference noted in region 4. The regions 2 and 3 are quite small and nearly negligible in probability. While the ES procedure seems undesirable on these regions, the concern is not important.

As a further comparison, consider the point (4, 3.8). The ES p-values for step 1 and 2 are $\approx .44$ and $\approx .0001$ respectively, while the naive p-values are approximately .0001 for both variables. The decision of ES to stop at step 1 and declare an empty model might well be viewed as embarrassing and subjectively undesirable. This is consistent with the claimed ES p-values though.

In the correlated setting, the interesting simulation in Section 6 strongly suggests that the p-values used in procedures b) and c) can be extremely conservative. Hence the naive scheme suggested above must be modified in order to achieve desirable performance. To decide whether such modification is possible needs further research, and, if possible, further investigation of the geometric structure and stochastic performance of the resulting tests.

For all considered testing methods, when the regressors are correlated the values of regression coefficients depend on which other coefficients are in the current model. Hence a coefficient may have a non-zero value within the currently active set of variables; and so be correctly included into that model at that step. Within a later active model it might then have a value of 0. Thus a correct selection at a given step may become incorrect as the process proceeds, and vice-versa. The related phenomenon of suppression can yield a series of insignificant steps followed by highly significant steps (Johnson et al., 2015b). These issues have important consequences for interpretation of p-values produced in a stepwise routine. Such issues do not occur in the simple model at hand involving independent variables with fixed mean values.

3 Appendix

The sequential p-values were constructed using data downloaded from Robert Tibshirani's website. The p-values computed in Table 1 are computed from the standard F-test with 1 and 58 = 67 - 9 degrees of freedom. As some additional numerical details, note that the MSE from the full model is $\hat{\sigma}^2 = 0.5074$. Thus, for example, the sequential F-value for testing "svi" is 2.1841/.5074 = 4.305 with a t-value of 2.075 $= \sqrt{4.305}$. This has a p-value with 58 degrees of freedom of 0.0426.

FS Step 1 (red curve): If X_1 is chosen before X_2 with a positive sign, the observation lies in the cone $R_1 = \{X_1 > X_2 > 0\}$. In order to have a level α test of $H_0 : \theta_1 = 0$ conditional on $(x_1, x_2) \in R_1$ one must have

$$\theta = \mathbb{P}(X_1 > \tau_1 | (x_1, x_2) \in R_1, X_2 = x_2) \quad \forall x_2.$$

This entails choosing the point via

$$\alpha = \frac{1 - \Phi(x_1)}{1 - \Phi(|x_2|)}. \tag{1}$$

This defines $x_1 = x_1(x_2)$ for the red curve.

FS Step 2 (yellow curve): The conditioning region is the same, so the level α test of $H_0 : \theta_2 = 0$ conditional on $(x_1, x_2) \in R_1$ requires

$$\theta = \mathbb{P}(X_2 > \tau_2 | (x_1, x_2) \in R_1, X_1 = x_1) \quad \forall x_1.$$

This entails choosing the point via

$$\alpha = \frac{\Phi(x_1) - \Phi(X_2)}{\Phi(x_1) - 1/2}. \tag{2}$$

ES Step 2 (green curve): Given $H_0 : \theta_1 = 0$ has been rejected, possible values of (X_1, X_2) lie to the right of FS Step 1. Denote this region as R_2. Now the test $H_0 : \theta_2 = 0$ must satisfy

$$\theta = \mathbb{P}(X_2 > \tau_2 | (x_1, x_2) \in R_2, X_1 = x_1)$$

for all x_1 for which the conditioning region is non-empty. The only change from FS Step 2 is that the conditioned region is a function of x_2. This entails choosing the point $x_2 = x_2(x_1)$ for which

$$\alpha = \frac{\Phi(x_2^*) - \Phi(X_2)}{\Phi(x_2^*) - 1/2}, \tag{3}$$

where x_2^* denotes the value for which $x_1(x_2^*) = x_1$. The computation in equation (3) is facilitated by noting that equation (1) implies

$$\Phi(x_2^*(x1)) = 1 + \frac{\Phi(x_1) - 1}{\theta}. \tag{4}$$

References

Buja, A. and Brown, L. (2014). Discussion: ”A significance test for the lasso”. *Annals of Statistics 2014, Vol. 42, No. 2, 509-517.*

G'Sell, M. G., Wager, S., Chouldechova, A., and Tibshirani, R. (2015). Sequential selection procedures and false discovery rate control. *Journal of the Royal Statistical Society: Series B (Statistical Methodology).*

Johnson, K. D., Stine, R. A., and Foster, D. P. (2015a). Revisiting Alpha-Investing: Conditionally Valid Stepwise Regression. *ArXiv e-prints*. http://arxiv.org/abs/1510.06322.

Johnson, K. D., Stine, R. A., and Foster, D. P. (2015b). Submodularity in Statistics: Comparing the Success of Model Selection Methods. *ArXiv e-prints*. http://arxiv.org/abs/1510.06301.