Comments on the paper, "An adaptive resampling test for detecting the presence of significant predictors" by I. W. McKeague and M. Qian

Lawrence D. Brown* and Daniel McCarthy*

ABSTRACT: This commentary deals with some issues related to the paper by McKeague and Qian (2015). We first discuss their formulation of the problem via a general random covariate model, but with a traditional structure of independent, homoscedastic residuals that are also independent of the covariates. A principle focus of their paper involves testing a null hypothesis of no marginal effect of the linear slope parameters. We present some alternate paths to test this null hypothesis in their formulation and in closely related formulations including an assumption-lean formulation.

* Statistics Department, Wharton School, University of Pennsylvania. Email: lbrown@wharton.upenn.edu and danielmc@wharton.upenn.edu .

KEY WORDS: Bootstrap; Family-wise error rate; Marginal regression; random-covariate regression; Testing a composite null; assumption-lean regression.

Introduction

The following comments relate to material in McKeague and Qian (2015). (Henceforth referred to as M&Q.) We have found this material to be very worthwhile reading and we thank the organizers for the opportunity to provide these comments.

M&Q study a maximum statistic, $\hat{\theta}_n$, that can be used to test for the null hypothesis of no effect of covariates in a linear-model analysis. Their major focus is on the development of a bootstrap style procedure that serves the dual purpose of estimating the distribution of $\hat{\theta}_n$ and then using this estimate as the basis for a test of the null hypothesis. Our perspective is that their procedure is composed of two almost separate components – a simulation of the distribution under the null and a bootstrap estimate that is valid away from the null. Our comments focus on the testing component of their procedure, and on alternate tests for this hypothesis.

The first part of our discussion deals with this perspective of their formulation. It points to concerns that are treated in the remainder of our discussion and, at the end, raises a few questions for the authors. Section 2 of our discussion sketches a generalization of their basic statistical model that we have treated elsewhere in more detail, and suggests that a modification of their test may be suitable also for this generalization. Section 3 discusses a related, though different, test of their null hypothesis that is embedded in Berk, et. al. (2013). Section 4 describes bootstrap ideas that do yield a valid test of the null hypothesis without attempting to estimate the distribution of $\hat{\theta}_n$. That section concludes with some prospective remarks about our ongoing research into bootstrap methods for this and related problems.

1. Formulation

The essential structure of the observed data is implicit in the first sentence of Section 2 and in equation (2) of McKeague and Qian (2015). (Henceforth referred to as M&Q.) The observations are a sample $\{\mathbf{X}_i, Y_i : i = 1,..,n\}$ from a population whose distribution has the property that

(1.1) $$Y = \alpha_0 + \mathbf{X}^T \boldsymbol{\beta} + \varepsilon .$$

Here **X** is a *p*-vector of covariates and $\boldsymbol{\beta}$ is a *p*-vector of parameters. Since **X** is random we refer to this formulation as having a random-covariate structure in contrast to the traditional structure in which the elements of each $\mathbf{X}_i$ are viewed as fixed constants. The marginal distribution of **X** is not known or constrained (except that it is assumed to have a finite covariance matrix). The residual variables $\{\varepsilon_i : i = 1,..,n\}$ are an iid sample and independent of $\{\mathbf{X}_i\}$. Their distribution is not specified in advance, except that they are assumed to have a finite variance, and hence be homoscedastic.

A primary goal of M&Q is to develop a test of the null hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{0}$ in (1.1) that is based on a maximal statistic drawn from simple regressions rather than from the multiple regression analysis. In order to describe this statistic, and to prepare for further discussion, some additional notation may be helpful. The following quantities are discussed in M&Q but not given explicit notations. The population and the sample slope coefficients from the simple (one-dimensional, marginal) regressions are

$$\beta_k^1 = \frac{\operatorname{cov}(\mathbf{X}_k, Y)}{\operatorname{var}(\mathbf{X}_k)},\ \hat{\beta}_k^1 = \frac{\widehat{\operatorname{cov}}(\mathbf{X}_k, Y)}{\widehat{\operatorname{var}}(\mathbf{X}_k)}\ , k = 1,..,p.$$

The corresponding t-statistics are

$$T_k^1 = \frac{\hat{\beta}_k^1}{\widehat{SE}\left(\hat{\beta}_k^1\right)}.$$

Where $\widehat{SE}^2\left(\hat{\beta}_k^1\right) = \sum\left(Y_i - \bar{Y} - \hat{\beta}_k^1\left(X_{ki} - \bar{X}_k\right)\right)^2 \Big/ \sum\left(X_{ki} - \bar{X}_k\right)^2$.

M&Q then define

$$\hat{k}_n = \arg\max\left|\widehat{\operatorname{corr}}(\mathbf{X}_k, Y)\right| = \arg\max\left|T_k^1\right| \tag{1.2}$$

where the last equality results from standard textbook manipulations. They then consider the statistic

$$\hat{\theta}_n = \hat{\beta}_{\hat{k}_n}^1 = \frac{\widehat{\operatorname{cov}}\left(\mathbf{X}_{\hat{k}_n}, Y\right)}{\widehat{\operatorname{var}}\left(\mathbf{X}_{\hat{k}_n}\right)}. \tag{1.3}$$

As M&Q aptly point out, direct bootstrap methods are not appropriate if one wishes to use this statistic to test H_0. (The flaw in attempting to use a bootstrap here is evident in the simulations in Section 4: The conventional bootstrap procedure leads to a strongly anti-conservative test of H_0.) For a test of H_0 one needs the null distribution of $\hat{\theta}_n$. This null distribution cannot accurately be obtained via a standard bootstrap since the distribution of $\hat{\theta}_n$ does not converge uniformly in a $1/\sqrt{n}$ neighborhood of the null. But it is possible to simulate the distribution of $\hat{\theta}_n$ under the null. The simulated random variable is given in the paper as $\mathbf{V}_n^*(\mathbf{0})$. This random quantity depends on an independent simulation of a mean-zero multivariate normal random vector, $\mathbf{Z}(\mathbf{0})$, whose covariance matrix is described in a display in Theorem 1. [The covariance matrix for this vector could be estimated directly from the data. In their computer code (which they have kindly shared with us) M&Q appear to use a bootstrap to estimate this covariance matrix. This may be more accurate in practice than a direct estimate of the covariance, but the paper provides no evidence on this secondary issue.]

The more general bootstrap result in Theorem 2 is the basis for the author's **ART** procedure. This should be viewed as a marriage of the simulation described above and a more standard bootstrap estimate of the distribution of $\sqrt{n}\left(\hat{\theta}_n - \theta_n\right)$. The bootstrap is used when the data convincingly reject the null hypothesis (when

$\max\left(\left|T_n\right|,\left|T_n^*\right|\right) > \lambda_n$) and otherwise the simulation at $\boldsymbol{b}_0 = \boldsymbol{\beta}_0 = \mathbf{0}$ is used. Theorem 2 should not be interpreted as providing a reliable bootstrap estimate for the distribution of $\sqrt{n}\left(\hat{\theta}_n - \theta_n\right)$ at every true parameter (which M&Q correctly remark does not exist). Such a global bootstrap based on this theorem (and hence without knowledge of $\boldsymbol{b}_0$) would require a consistent estimate of $\boldsymbol{b}_0$, but such an estimate does not exist. (One would also need to know, or assume, a-priori that $\boldsymbol{\beta}_0$ satisfies the special assumption of Theorem 1 as expressed in the paragraph above the Theorem. This would not, for example, be the case if the true $\boldsymbol{\beta}$ in (1.1) were sparse.)

We have two additional questions about the formal results.

(i) The **ART** simulation and test is based on the distribution of the selected slope statistic in (1.3). But selection is based on the maximal |t|-statistic as in (1.2). The maximal |t|-statistic is $\left|T_{\hat{k}_n}^1\right|$, and, in general, $\left|T_{\hat{k}_n}^1\right| \neq \hat{\theta}_n$. Such a test statistic would have the advantage of being invariant under coordinate-wise affine transformations of the X-variables, whereas the statistic in (1.3) is not. Did you investigate the performance of such a procedure? In the simulations of Section 4 there would be very little difference in numerical value or performance since the X-coordinates there are independent with equal variance, but differences might be more noticeable in other settings.

(ii) Theorems 1 and 2 are proved within a formulation in which *p* remains fixed as $n \to \infty$. Yet, the simulations in Section 4 address some cases in which *p* is comparable to *n*, or even larger. (It is an advantage of the **ART** procedure over more familiar procedures based on inference about the full vector $\boldsymbol{\beta}$ in (1.1) that it can numerically deal with such situations.) The **ART** procedure appears to perform satisfactorily even in these large *p* cases. Is this perhaps only because the choice of standard normality for the distributions of **X** and ε are so favorable to **ART**, or is this a more general phenomenon? Is there any asymptotic theory to justify the simulation at the heart of **ART** when $p \to \infty$ along with *n*?

2. Assumption-lean Models

Buja, et. al. (2015) contains a detailed exposition of an "assumption-lean" regression formulation. In such a formulation one need only assume that the observed variables are a random sample $\left\{\mathbf{X}_i, Y_i : i = 1,..,n\right\}$ from a joint distribution possessing low-order moments. The target of inference is the population slope; this can be defined in any one of several equivalent ways. A straightforward version is to define the population slope vector via $\boldsymbol{\beta}^{\bullet} = \arg\min_{\gamma} E\left(Y - \mathbf{X}\boldsymbol{\gamma}\right)^2$. An alternate form that is more analogous to (1.1) involves writing

(2.1) $$Y = \alpha_0 + \mathbf{X}^T \beta^{\bullet} + \varepsilon \text{ where } \operatorname{cov}\left(\mathbf{X}, \varepsilon\right) = 0 .$$

(We use the notation $\beta^{\bullet}$ here, rather than just $\boldsymbol{\beta}$ in order to distinguish this formulation from that of (1.1). (2.1) is more general that (1.1), but if the assumptions of (1.1) hold and the population model is full-rank then $\boldsymbol{\beta}^{\bullet} = \boldsymbol{\beta}$ and $\boldsymbol{\beta}^{\bullet 1} \neq \boldsymbol{\beta}^{1}$.)

This formulation shares with (1.1) the feature that **X** is a random vector whose marginal distribution is unknown (except for the existence of low order moments). But it is otherwise much broader and assumption-lean. The randomness of **X** (in both (1.1) and (2.1)) has an important side benefit in that in general it justifies the asymptotic use of an **X**-Y bootstrap such as that in Theorem 2 when $\boldsymbol{\beta}^{\bullet}$ is not in a $\sqrt{n}$ neighborhood of 0, and otherwise satisfies the assumptions in that Theorem. But, as noted above, the core of the **ART** procedure as a test is really a simulation of the null distribution of the statistic $\hat{\theta}_n$. We believe that this simulation should also be valid in the assumption-lean setting of (2.1), and should hence lead to a useful test of $H_0 : \boldsymbol{\beta}^{\bullet} = 0$. Here's why.

Buja, et. al. (op. cit.) describes interpretations and inference for $\boldsymbol{\beta}^{\bullet}$ from data as in (2.1), and several other aspects of such a formulation. The sandwich estimator of Huber (1967) and White (1980a,b; 1982) plays a key role in such inference. For M&Q a key ingredient of the simulation in **ART** is the covariance matrix in Theorem 1 at $\boldsymbol{\beta} = \mathbf{0}$. This relates to a form of the sandwich estimator for the covariance matrix of the vector of marginal sample slopes, $\hat{\boldsymbol{\beta}}^{1}$. The appropriate sandwich estimator would be

$$(2.2) \qquad \left[diag\left(\left\{ \widehat{\operatorname{var}}\left(\mathbf{X}_k\right) \right\} \right) \right]^{-1} \mathbf{M} \left[diag\left(\left\{ \widehat{\operatorname{var}}\left(\mathbf{X}_k\right) \right\} \right) \right]^{-1} \text{ where}$$

$$\mathbf{M}_{k\ell} = n^{-1} \sum_i \left(Y_i - \bar{Y} - \hat{\beta}_k^1 \left(\left(\mathbf{X}_i\right)_k - \bar{\mathbf{X}}_{\bullet k} \right) \right)^2 \left(\left(\left(\mathbf{X}_i\right)_k - \bar{\mathbf{X}}_{\bullet k} \right) \right)^2 .$$

(If $\boldsymbol{\beta}^{1}$ is assumed to lie in a $1/\sqrt{n}$ neighborhood of **0** then the term $\hat{\beta}_k^1 \left(\left(\mathbf{X}_i\right)_k - \bar{\mathbf{X}}_{\bullet k} \right)$ in (2.2) is asymptotically negligible and can be ignored.) The matrix **M** is very similar to the covariance matrix described in Theorem 1; we believe they are asymptotically equivalent when $\boldsymbol{\beta}^{1}$ is assumed to lie in a $1/\sqrt{n}$ neighborhood of **0.** (The inverse diagonal matrix terms do not appear in the covariance expression described in Theorem 1, but are instead accommodated in the first denominator of (4).)

In summary, we believe that the simulation idea embodied within **ART** can be directly applied to testing the null hypothesis $H_0 : \boldsymbol{\beta}^{\bullet} = 0$. Thus the simulation component in **ART** may turn out to be more flexible and robust than appears from the specific formulation via (1.1). (The bootstrap idea in M&Q for the distribution of $\sqrt{n}\left(\hat{\theta}_n - \theta_n\right)$ away from the null and other special points excluded by the assumptions of Theorem 1 is also almost automatically valid under (2.1).)

3. POSI

Berk, et. al. (2013) provides a simultaneous confidence interval procedure for estimates of slope coefficients in a setting like that of (1.1), but with the elements of

X treated as fixed constants, rather than as random variables. To be more precise, the setting for that paper involves observation of an $n \times 1$ vector **Y** satisfying

(3.1) $$\mathbf{Y} = \alpha_0 \mathbf{1} + \mathbf{X}_{n\times p}\boldsymbol{\beta}^{\circ} + \boldsymbol{\varepsilon} \text{ where } \boldsymbol{\varepsilon} \sim N_n\left(0, \sigma^2 \mathbf{I}_{p\times p}\right).$$

Here, the design matrix, $\mathbf{X}_{n\times p}$, is composed of observed constants.

Primary interest in Berk, et. al. (op. cit.) focuses on finding simultaneous confidence levels for the slope coefficients in the family of all possible sub-models composed of subsets of columns of **X**. However that paper also discusses cases in which the family or the targeted slope coefficients are restricted in some fashion. One possibility is to restrict the sub-models to consist of only one predictor at a time. (See possibility (2) in subsection 4.4 of that article.) The family of confidence intervals is thus composed of confidence intervals for simple, marginal regression coefficients. These coefficients can be combined into a vector $\boldsymbol{\beta}^{\circ 1}$, where the notation is analogous to that used following (1.1). Hence the POSI procedure provides a valid test of $H_0^{\circ} : \boldsymbol{\beta}^{\circ 1} = \mathbf{0}$ relative to the model (3.1).

This POSI test is also valid for testing the null hypothesis of M&Q – $H_0 : \boldsymbol{\beta}^1 = \mathbf{0}$ within the model (1.1) with Gaussian errors. (See comment (iii) below on this issue.) In most cases results within (1.1) and (3.1) do not transfer so directly. But in the case of this null hypothesis the direct carryover is justified because under the null hypothesis (but not otherwise) the distribution of **X** is an ancillary statistic. Thus a test that is conditionally valid (i.e., is valid under (3.1)) is also unconditionally valid (i.e., under (1.1)). Buja, et. al. (op. cit.) contains an extensive discussion of ancillarity issues in random design models like (1.1) and (2.1).

The POSI test described here is not a clone of the test provided by the simulation in **ART** because the critical value in Berk, et. al. (op.cit.) is drawn from the distribution of $\left|T^1_{\hat{k}_n}\right|$ rather than from the distribution of $\hat{\theta}_n$ as defined in (1.3). (See our comment 1(i).) In other respects the simulations in the two procedures are very similar. Comment (ii) below points to one additional structural difference but otherwise there seem to be only minor technical differences that are asymptotically insignificant.

Some additional comments may be helpful.

(i) R-code is available for computing the critical constant for the POSI test. This code has an explicit option for the restriction to marginal sub-models as described above. See Buja (2015).

(ii) The POSI test involves an estimate for the residual variance, σ^2, in (3.1). The POSI software and the theory supporting it draw this estimate from the full model Sum of Squares for Error. The simulation portion of **ART** draws an estimate for the analogous purpose via the sandwich style expression in Theorem 1 at $\boldsymbol{\beta} = \mathbf{0}$ combined with (4). This is asymptotically equivalent (under suitable assumptions) to what would appear at the corresponding step of the POSI algorithm if one were to draw the estimate of σ^2 under the assumption that H_0 is true (i.e., from the restricted model rather than from the full model). The POSI software can be modified to proceed in this fashion. Unless *n* – *p* is small we would not recommend proceeding in this fashion because the resulting test will not be similar (even under assumptions of normality). But for small or negative values of *n* – *p* such a path would be desirable.

(iii) The residual distributions in (1.1) are more general than in (3.1); in (3.1) they are required to be Gaussian whereas in (1.1) they need only be iid (with finite variance). Berk, et. al. (op. cit) does not explicitly discuss such an extension of the model in (3.1). However, in retrospect, after reading M&Q we realize that the considerations in the POSI paper appear to be asymptotically valid under such an iid assumption for the coordinates of $\boldsymbol{\varepsilon}$ in (3.1). We conjecture that this is so, and hence that the POSI test of H_0° and H_0 is asymptotically valid.

4. A bootstrap test

A too casual reading of M&Q might incline one to feel that a bootstrap test of their H_0 is not possible unless the test also includes a simulation component, as does their **ART**. This is not so. What is true is that a pure bootstrap estimate of the distribution of a statistic like their $\hat{\theta}_n$ would be flawed, and this would also be the case for a statistic like $\left|T_{\hat{k}_n}^1\right|$ discussed above. A valid bootstrap test of H_0 requires a different structure.

Here is an outline of a simple bootstrap test of H_0 that is (asymptotically) valid under the model (1.1). Consider a family of confidence sets for $\boldsymbol{\beta}^1$ of rectangular form:

(4.1)
$$\mathrm{Rect}_C\left(\hat{\boldsymbol{\beta}}^1\right)=\left\{\boldsymbol{\beta}^1:\left|\boldsymbol{\beta}_k^1-\hat{\boldsymbol{\beta}}_k^1\right|\le C,\ k=1,..,p\right\}$$

.

Use a bootstrap to determine the constant, C_{boot}, for which these rectangles have the desired estimated coverage, $1-\alpha$. Then reject H_0 if $\mathbf{0}\notin\mathrm{Rect}_{C_{boot}}$. This procedure has the desired asymptotic coverage as $n\to\infty$ for fixed *p*, and provides asymptotically satisfactory performance. (There is no claim here of any optimality for this test. The particular form for the rectangles in (4.1) is suggested here only for expositional convenience, and as a parallel to the focus of M&Q on $\hat{\theta}_n$.) In this simple setting the asymptotic properties follow from standard bootstrap theory, but see Buja and Rolke (2015 – and earlier) for a full, general treatment of such procedures.

Although this procedure does not attempt to discover the true distribution of a maximum statistic like $\hat{\theta}_n$ it does involve the distribution of $\hat{\theta}_n-\theta_n$. The form of the rectangles in (4.1) was chosen because of its relation to the simulation in **ART**. Other forms of confidence region may yield more satisfactory performance. For example, one could choose the sides of the rectangle to be proportional to the values of sandwich estimates of $\mathrm{SD}\left(\hat{\boldsymbol{\beta}}_k^1\right)$. This bootstrap procedure can be converted to create yield an asymptotically valid statement about the distribution of $\hat{\theta}_n-\theta_n$ under the null hypothesis. As such, it could be used in place of the simulation in ART involving $\mathbf{V}_n^*(\mathbf{0})$.

Along with collaborators including K. Zhang we are preparing a methodological study of bootstrap confidence intervals for the slope coefficients in the assumption-lean model. The bootstrap estimator we are proposing is a double bootstrap, with the second level of bootstrap improving the calibration of intervals provided by the first level. Asymptotic theory in our study suggests that such a double bootstrap can have better performance than a single bootstrap. (Based on helpful dialog with M&Q we note that that our proposal involves a more sophisticated, and more computer intensive, style of bootstrap than the CPB bootstrap used in their simulations – of course this does not negate the objection to using a bootstrap of any sort to estimate the distribution of a statistic such as $\hat{\theta}_n$.) Our research to date has been focused on intervals for pre-chosen coordinates $\beta_k^{\bullet}$ (or $\beta_k^{\bullet 1}$). But after reading M&Q we realize that the methodology in our study can be adapted to the simultaneous confidence problem described here, and can also yield more evolved forms of confidence rectangles than those in (4.1). We intend to pursue such issues in future.

Both space and time constrain us from going into further detail here. But we are indebted to M&Q for indirectly providing the motivation to study such an issue as well as for the very interesting treatment and results involving their **ART** procedure.

References

Berk, R., Brown, L. D., Buja, A., Zhang, K., and Zhao, L. (2013). Valid Post-selection inference. *Ann. Statist.* **41**, 802-837.

Buja, A. (2015). Software for computing the POSI constant. Follow link at http://www-stat.wharton.upenn.edu/~buja/ .

Buja, A. and Rolke, W. (2015). Calibration for Simultaneity: (Re)Sampling Methods for Simultaneous Inference with Applications to Function Estimation and Functional Data. Available at https://statistics.wharton.upenn.edu/profile/555/research/?pubFilter=publishedPaper.

Buja, A., Berk, R., Brown, L. D., George, E. Pitkin, E., Traskin, M., Zhao, L., and Zhang, K. (2015). Models as approximations – a conspiracy of random regressors and model deviations against classical inference in regression. *Submitted*. Available from http://www-stat.wharton.upenn.edu/~buja/ .

McKeague, I. W. and Qian, M. (2015). An adaptive resampling test for detecting the presence of significant predictors. *Jour. Amer. Statist. Assn.*, to appear