The Canadian Journal of Statistics
Vol. 37, No. 4, 2009, Pages 625-644
La revue canadienne de statistique

Inference after variable selection using restricted permutation methods

Rui WANG and Stephen W. LAGAKOS

Key words and phrases: variable selector; covariates; regression; sample splitting.
MSC 2000: Primary 62G09; secondary 62J05.

Abstract: When confronted with multiple covariates and a response variable, analysts sometimes apply a variable-selection algorithm to the covariate-response data to identify a subset of covariates potentially associated with the response, and then wish to make inferences about parameters in a model for the marginal association between the selected covariates and the response. If an independent data set were available, the parameters of interest could be estimated by using standard inference methods to fit the postulated marginal model to the independent data set. However, when applied to the same data set used by the variable selector, standard ("naive") methods can lead to distorted inferences. The authors develop testing and interval estimation methods for parameters reflecting the marginal association between the selected covariates and response variable, based on the same data set used for variable selection. They provide theoretical justification for the proposed methods, present results to guide their implementation, and use simulations to assess and compare their performance to a sample-splitting approach. The methods are illustrated with data from a recent AIDS study.

1. INTRODUCTION

Let $X_1, X_2, \cdots, X_p$ denote a set of covariates and Y denote some continuous response, and consider the following scenario: a variable selection procedure is applied to n independent and identically distributed copies of $(X_1, \cdots, X_p, Y)$, resulting in a subset, $(X_{j_1}, \cdots, X_{j_s})$, of covariates that appear to be related to Y. One is then interested in making an inference about parameters in a model for the marginal association between $(X_{j_1}, \cdots, X_{j_s})$ and Y; that is, about the parameters that would be estimated if standard methods (e.g., least-squares) were applied to fit the marginal model to m additional independent copies of $(X_{j_1}, \cdots, X_{j_s}, Y)$. In practice, however, an independent data set is often unavailable, and thus one wishes to make inferences about these parameters based on the same data set used by the variable selector. This problem is sometimes referred to as "inference after variable selection".

Past work in inference after variable selection has focused on inferences about the parameters in models for the association between Y and all p candidate covariates. Rather than fitting a single regression model involving all p covariates, the idea is to use a variable selection step to reduce the number of covariates and then base inferences on a more parsimonious model which acts as if the selected covariates can fully explain the association of Y with all p covariates. Several authors have noted that application of standard ("naive") inference methods to the same data set used for variable selection can lead to invalid inferences, including distorted Type I errors, biased estimators, and confidence intervals with distorted coverage probabilities (cf: Miller 1984; Chatfield 1995; Hurvich & Tsai 1990; Zhang 1992; Kabaila 1995; Pötscher & Novák 1998, Danilov & Magnus 2004, Leeb & Pötscher 2005, Kabaila & Leeb 2006, Giri & Kabaila 2008), as well as overly

optimistic prediction errors (cf: Efron 1986; Gong 1986; Breiman 1992; Leeb 2009). Although the bias of naive approaches can become negligible when the probability of selecting all the important covariates approaches 1 as the sample size increases (Pötscher 1991), Chatfield (1995), Leeb & Pötscher (2003, 2005), Leeb (2005), and others have noted that such results are not useful for making inferences in finite sample because severe biases can still occur. Alternative analytical methods, including bootstrap and jackknife, can sometimes reduce, but do not in general eliminate these biases, (cf: Faraway 1992, Veall 1992, Shen, Huang & Ye 2004). Conditions under which these methods are valid have not been identified, and their poor performance in some settings was noted in Freedman, Navidi & Peters (1988), and Dijkstra & Veldkamp (1988). As Leeb & Pötscher (2005) note,"... a proper theory of inference post model selection is only slowly emerging ...". Recently, Shen *et al.* (2004) develop approximate inference methods for the regression setting when the response is normally distributed, a consistent or over-consistent variable selector is used (that is, a variable selector that asymptotically selects or includes the true model with probability 1), and where the parameters of interest correspond to covariates that are not candidate for exclusion by the variable selector. Leeb (2009) proposes a prediction interval which is approximately valid and short with high probability in finite samples for the linear regression setting where the error and also the explanatory variables are jointly normally distributed.

We are unaware of any literature addressing the issue of using the same data set to first select a subset of covariates and then to make inferences for the marginal association between this selected subset and the response. For linear models, the parameters in the marginal model sometimes coincide with those in the model for the association between all covariates and the response, but in general they differ (Cochran 1938, Cox 2007). Naive methods which ignore the fact that the same data set is used both for variable selection and for subsequent inferences are commonly used in practice. As with inferences for the full model following variable selection, naive methods are in general biased in the settings we consider, as we will illustrate. Our interest in inferences about the marginal association is motivated by a common situation in medical research of wishing to understand the relationship between a specific set of covariates and a response, while recognizing that other covariates might also be associated with the response. For example, Lossos *et al.* (2004) examined 36 candidate genes and a disease outcome using a variable selection algorithm to arrive at a subset of 6 genes, and then assessed their association with the disease outcome. Here the scientific goal is to use these selected genes to develop an improved staging system for guiding patient management, and thus the main statistical interest is the marginal association between the selected genes and the clinical outcome, not about whether all the important genes were selected or whether a better variable selector could have been used. Similarly, in determining the genetic correlates of efavirenz hypersusceptibility, Shulman *et al.* (2004) used stepwise regression to identify specific reverse transcriptase (RT) mutations at three codons, and then considered a logistic regression model relating efavirenz hypersusceptibility and the RT mutations at these three codons. Having identified these codons, the scientific question is whether they can be used to guide the use of antiretroviral therapy by identifying efavirenz hypersusceptibility, and thus the statistical focus is the marginal association between these 3 codons and clinical response.

The proposed approach involves transforming the covariate-response data matrix to one that can be partitioned into two components that are independent under a null hypothesis, then forming new matrices by permuting the rows of one component while holding the other component fixed, and then basing inferences on a permutation distribution formed from a specific subset of the resulting matrices. Exact tests and confidence intervals are obtained for some settings and approximate inferences for others. In section 2 we describe the general approach, present the theorem that justifies it, and discuss its implementation. In section 3 we use simulations to assess and compare the performance of the proposed approach to naive methods and to a sample-splitting approach, and in section 4 we illustrate the methods with data from a recent AIDS study. In section 5 we discuss related issues and areas in need of further development. All proofs are included in the Appendix.

2. METHODS

We begin this section by introducing notation and conceptually describing the proposed approach. This is followed by a theorem that justifies the approach and results that guide its implementation.

2.1 Notation

Let $\mathcal{X} = \{X_1, X_2, \cdots, X_p\}$ denote a set of candidate covariates, including interaction terms of interest, and let Y denote some continuous response variable. Suppose that the observations consist of n i.i.d. copies of the random vector $(X_1, X_2, \cdots, X_p, Y)$, let $\mathbf{X}$ and $\mathbf{Y}$ denote the corresponding $n \times p$ matrix and $n \times 1$ vector, and let $\mathbf{D} = (\mathbf{X},\ \mathbf{Y})$. In what follows $g(\cdot)$ is a one-to-one function of $(X_1, \cdots, X_p, Y)$ and for a $n \times (p+1)$ matrix $\mathbf{M}$, we define $g(\mathbf{M})$ (or $g^{-1}(\mathbf{M})$) as the $n \times (p+1)$ matrix with i^{th} row obtained by applying g (or g^{-1}) to the i^{th} row of $\mathbf{M}$. We use "$\stackrel{p}{=}$" to denote "is a row permutation of" and " $\perp$ " to denote "is independent of".

Suppose $\mathcal{R} : \mathbf{D} \rightarrow \mathcal{X}_S \subseteq \mathcal{X}$ denotes a variable selector that maps $\mathbf{D}$ into a subset, $\mathcal{X}_S = \{X_{l_1}, X_{l_2}, \cdots, X_{l_s}\}$, of $\mathcal{X}$. Let $X_S = (X_{l_1}, X_{l_2}, \cdots, X_{l_s})$ denote the corresponding vector of selected covariates and let $X_{\setminus S}$ denote the vector formed by the random variables in $\mathcal{X} \setminus \mathcal{X}_S$. The variable selector $\mathcal{R}$ is arbitrary, including selectors where $\mathcal{X}_S$ is empty with positive probability or where some covariates are selected with probability 1. For example, the variable selector might first assess the univariate association between each of p candidate covariates and Y, and then select all those covariates achieving a significance level below some threshold. Or, if X_1 denotes treatment or exposure in a medical study, the variable selector might be a step-up procedure that identifies a subset of possible prognostic factors, which are then included in a model with X_1 to make inferences about the treatment effect while controlling for the selected prognostic factors, and about the association of the selected prognostic factors with the response. Here X_1 is selected with probability 1, while the selection of each prognostic factor is uncertain.

2.2 Conceptual Description of Approach and Theoretical Justification

Consider testing a hypothesis, H_0, about the marginal association of the selected covariates $\mathcal{X}_S$ with the response Y. The approach consists of the following 2 steps.

Transformation and Partition Step: Based on the hypothesis of interest and conditions on the underlying joint distribution of $(X_1, \cdots, X_p, Y)$, we first identify a one-to-one transformation, $g(\cdot)$, of $(X_1, \cdots, X_p, Y)$, whose $p+1$ components can be partitioned into two nonempty sets of random variables that are independent under H_0. We then use $g(\cdot)$ to form the $n \times (p+1)$ matrix $\tilde{\mathbf{D}} = g(\mathbf{D})$ and partition $\tilde{\mathbf{D}}$ into $\tilde{\mathbf{D}}_P$ and $\tilde{\mathbf{D}}_F$, where the variables comprising $\tilde{\mathbf{D}}_P$ and $\tilde{\mathbf{D}}_F$ are independent under H_0.

Permutation and Restriction Step: Consider the $n!$ matrices of the form $\tilde{\mathbf{D}}(l) = (\tilde{\mathbf{D}}_P^l, \tilde{\mathbf{D}}_F)$, where $\tilde{\mathbf{D}}_P^l \stackrel{p}{=} \tilde{\mathbf{D}}_P$, and identify the subset, Π_R, of matrices for which $\mathcal{R}(g^{-1}(\tilde{\mathbf{D}}(l))) = \mathcal{X}_S$; that is, the subset of matrices $\tilde{\mathbf{D}}(l)$ whose reverse transformation is mapped to $\mathcal{X}_S$ by the variable selector. Let $T = T(\tilde{\mathbf{D}})$ denote some test statistic for H_0. We then compare the observed value of T to the permutation distribution formed by evaluating $T(\tilde{\mathbf{D}}(l))$ for those $\tilde{\mathbf{D}}(l)$ in the restricted set Π_R.

The validity of the proposed approach is based on the following Theorem. In Section 2.3, we will discuss ways to choose the transformation $g(\cdot)$ and the partition based on a given hypothesis of interest and conditions on the underlying joint distribution of $(X_1, \cdots, X_p, Y)$.

THEOREM 1. *Let $\mathbf{D} = (\mathbf{X}, \mathbf{Y})$ denote the $n \times (p+1)$ matrix consisting of n i.i.d. copies of $(X_1, X_2, \cdots, X_p, Y)$ and suppose there exists a one-to-one transformation, $g(\cdot)$, of $(X_1, X_2, \cdots, X_p, Y)$ such that $\tilde{\mathbf{D}} = g(\mathbf{D})$ can be partitioned as $\tilde{\mathbf{D}} = (\tilde{\mathbf{D}}_P, \tilde{\mathbf{D}}_F)$ for some $\tilde{\mathbf{D}}_P$ and $\tilde{\mathbf{D}}_F$ whose elements are independent under H_0. For any $\mathbf{d}$ in the support of $\mathbf{D}$, define $x_S = \mathcal{R}(\mathbf{d})$ and $\tilde{\mathbf{d}} = g(\mathbf{d}) = (\tilde{\mathbf{d}}_P, \tilde{\mathbf{d}}_F)$.*

For each row permutation, $\tilde{\mathbf{d}}_P^l$, of $\tilde{\mathbf{d}}_P$, define $\tilde{\mathbf{d}}(l) = (\tilde{\mathbf{d}}_P^l, \tilde{\mathbf{d}}_F)$ and suppose $g^{-1}(\tilde{\mathbf{d}}(l))$ is in the support of $\mathbf{D}$. Let $\Pi_R = \{\tilde{\mathbf{d}}(l) \mid \mathcal{R}(g^{-1}(\tilde{\mathbf{d}}(l))) = x_S\ \}$. Then under H_0 and for any $\tilde{\mathbf{d}}(l) \in \Pi_R$,

$$P(\tilde{\mathbf{D}} = \tilde{\mathbf{d}}(l) \mid \tilde{\mathbf{D}}_P \stackrel{p}{=} \tilde{\mathbf{d}}_P, \tilde{\mathbf{D}}_F = \tilde{\mathbf{d}}_F, \mathcal{R}(\mathbf{D}) = x_S) = 1/M,$$

where M is the number of matrices in Π_R.

By construction, the unpermuted matrix $\tilde{\mathbf{d}} = g(\mathbf{d})$ is an element of the restricted set Π_R. Theorem 1 shows that under H_0 and conditional on 1) the result of the variable selection, 2) the observed value of $\tilde{\mathbf{D}}_F$, and 3) knowledge of $\tilde{\mathbf{D}}_P$ up to a row permutation, the M matrices that comprise Π_R are equally likely. Thus, the observed value, $\tilde{\mathbf{d}}$, of $\tilde{\mathbf{D}}$ can be viewed as a randomly selected element from the set Π_R. It follows that if $T = T(\tilde{\mathbf{D}})$ is any test statistic, the observed value of T can be viewed under H_0 as a random sample of size 1 from the resulting (computable) permutation distribution of values $\{T(\tilde{\mathbf{d}}(l)) \mid \tilde{\mathbf{d}}(l) \in \Pi_R\}$. This provides the basis for exact inferences about H_0 that correct for variable selection. Note that although the choice of test statistics does not affect Type I error, it does affect power and therefore requires careful consideration. Confidence regions for model parameters can be obtained by inverting the restricted permutation tests; that is, a $(1-\alpha)\%$ confidence region is given by those parameter values that are not rejected at the α level of significance.

To give a flavor of the method, suppose X is scalar, so that $\mathbf{D}$ is the $n \times 2$ matrix with i^{th} row (X_i, Y_i), and assume that the conditional density function of Y, given $X = x$, is $f_0(y - \beta x)$ for some $f_0(\cdot)$, where β is an unknown parameter.

(a) First suppose there is no variable selection, that is, $\mathcal{R}(\mathbf{D}) \equiv \{X\}$, and consider testing the hypothesis $H_0 : \beta = 0$ of no association between X and Y. Let g denote the identity transformation and partition $\tilde{\mathbf{D}} = \mathbf{D}$ by taking $\tilde{\mathbf{D}}_P = \mathbf{X}$ and $\tilde{\mathbf{D}}_F = \mathbf{Y}$. If $\mathbf{X}^l$ denotes a row permutation of $\mathbf{X}$, then $\tilde{\mathbf{D}}(l) = (\mathbf{X}^l, \mathbf{Y})$ and Π_R is the set of all $n!$ such matrices. Let $T(\tilde{\mathbf{D}})$ be some test statistic, say $\mathbf{X}^T\mathbf{Y}$. We can then test H_0 by comparing its observed value to the permutation distribution formed by the $n!$ values of $T(\tilde{\mathbf{D}}(l)) = (\mathbf{X}^l)^T\mathbf{Y}$. The proposed method thus reduces to the classical permutation test for the association between X and Y. Formally, inference is based on the null permutation distribution $P(\tilde{\mathbf{D}} = \tilde{\mathbf{d}}(l) \mid \tilde{\mathbf{d}}_P \stackrel{P}{=} \mathbf{x}, \tilde{\mathbf{d}}_F = \mathbf{y}, \mathcal{R}(\mathbf{D}) = \{X\}) = P(\tilde{\mathbf{D}} = (\mathbf{x}^l, \mathbf{y}) \mid \mathbf{X} \stackrel{p}{=} \mathbf{x}, \mathbf{Y} = \mathbf{y}) = 1/n!$, for each of the $M = n!$ matrices in Π_R.

(b) Now consider use of a variable selector that sometimes selects $\{X\}$ but otherwise does not according to some rule; that is, $\mathcal{R}(\mathbf{D}) = \{X\}$ or $\{\phi\}$. Suppose $\mathcal{R}(\mathbf{D}) = \{X\}$ for a particular $\mathbf{D}$, and consider testing $H_0 : \beta = 0$. We again take g to be the identity transformation so that $\tilde{\mathbf{D}}(l) = (\mathbf{X}^l, \mathbf{Y})$ as before, but now base inferences on only those $\tilde{\mathbf{D}}(l)$ such that $\mathcal{R}(g^{-1}(\tilde{\mathbf{D}}(l))) = \mathcal{R}(\mathbf{X}^l, \mathbf{Y}) = \{X\}$; that is, the method applies the classical permutation test but restricted to only those matrices $(\mathbf{X}^l, \mathbf{Y})$, for which $\mathcal{R}$ selects X.

(c) Finally, consider the same variable selector as in (b) and suppose that $\mathcal{R}(\mathbf{D}) = \{X\}$ for a particular $\mathbf{D}$, but now consider testing $H_0 : \beta = \beta^0$ for some $\beta^0 \neq 0$. Here we can take $g(x, y) = (x, y + \beta_0 x)$, so that $\tilde{\mathbf{D}} = (\mathbf{X}, \mathbf{Y} + \beta^0\mathbf{X})$. Note that the 2 columns of $\tilde{\mathbf{D}}$ are independent under H_0 and that $\tilde{\mathbf{D}}(l) = (\mathbf{X}^l, \mathbf{Y} + \beta^0\mathbf{X})$. The inverse mapping is $g^{-1}(a, b) = (a, b - \beta_0 a)$ and thus $g^{-1}(\tilde{\mathbf{D}}(l)) = (\mathbf{X}^l, \mathbf{Y} + \beta^0(\mathbf{X} - \mathbf{X}^l))$. If $T(\tilde{\mathbf{D}})$ is some test statistic, the observed value $T(\tilde{\mathbf{D}})$ is then compared to the permutation distribution formed by $\{T(\tilde{\mathbf{D}}(l)) \mid \tilde{\mathbf{D}}(l) \in \Pi_R\} = \{T(\tilde{\mathbf{D}}(l)) \mid \mathcal{R}((\mathbf{X}^l, \mathbf{Y} + \beta^0(\mathbf{X} - \mathbf{X}^l))) = X\ \}$. This test could be used to construct a confidence interval for β.

2.3 Finding a Transformation and Partition

2.3.1 General Considerations

In more complex settings than the example discussed above, finding a transformation $g(\cdot)$ that leads to a desired partition may not be obvious. The proposed methods are invariant to reversing the roles of $\tilde{\mathbf{D}}_P$ and $\tilde{\mathbf{D}}_F$, and for a given data matrix $\tilde{\mathbf{D}}$ there could be multiple partitions, as

we illustrate in section 3.1.2. Below we give two general conditions to guide the choice of the transformation and partition. We then discuss the application of these conditions for specific hypotheses when a linear model is assumed for the association between Y and the candidate covariates $X_1, X_2, \cdots, X_p$.

Condition A: (X, Y) can be transformed one-to-one to some $(\tilde{X}, \tilde{Y}) = (\tilde{X}_P, \tilde{X}_F, \tilde{Y})$ such that $\tilde{X}_P \perp (\tilde{X}_F, \tilde{Y})$ under H_0. This can be verified using $\tilde{X}_P \perp \tilde{X}_F$ and $\tilde{Y} \perp \tilde{X}_P \mid \tilde{X}_F$.
With this condition, Theorem 1 holds by taking $\tilde{\mathbf{D}}_P = \tilde{\mathbf{X}}_P$ and $\tilde{\mathbf{D}}_F = (\tilde{\mathbf{X}}_F, \tilde{\mathbf{Y}})$. We refer to the resulting test as restricted permutation test A.

Condition B: (X, Y) can be transformed one-to-one to some $(X, \tilde{Y})$ so that $\tilde{Y} \perp X$ under H_0.
With this condition, Theorem 1 holds by taking $\tilde{\mathbf{D}}_P = \mathbf{X}$ and $\tilde{\mathbf{D}}_F = \tilde{\mathbf{Y}}$. We refer to the resulting test as restricted permutation test B.

2.3.2 Linear Models

Suppose the response Y is related to the candidate covariates $X_1, \cdots, X_p$ by a standard linear model:

$$Y = \beta^* X + \epsilon^*, \tag{2.1}$$

where $\epsilon^* \perp X$ and where some of the components of β^* may be zero. Consider the following conditions:

(C1) $X_S \perp X_{\setminus S}$

(C2) $Y \perp X_{\setminus S} \mid X_S$

(C3) The components of $X_{\setminus S}$ are continuous and $X_{\setminus S} = \gamma_S X_S + \tilde{\epsilon}$, where $\tilde{\epsilon} \perp X_S$. A special case is when $X = (X_1, \cdots, X_p)$ is normally distributed, in which case $\gamma_S = \Sigma_{\setminus S,S}\Sigma_S^{-1}$, where $\Sigma_{\setminus S,S} = cov(X_{\setminus S}, X_S)$ and $\Sigma_S = var(X_S)$.

PROPOSITION 1. *Assume (2.1). If any of (C1), (C2), or (C3) holds, the marginal association between X_S and Y is of the form*

$$Y = \beta_S X_S + \epsilon \tag{2.2}$$

for some β_S, where $\epsilon \perp X_S$. Furthermore, when (C2) holds, $\epsilon \perp X$.

Note that (2.2) refers to the marginal association of X_S and Y, that is, $f(Y \mid X_S)$, and not the association between X_S and Y induced by the variable selector, that is, $f(Y \mid X_S, \mathcal{R}(\mathbf{D}) = \mathcal{X}_S)$. Condition (C1) can be checked empirically, and sometimes covariates can be transformed to make this condition hold approximately. Condition (C2), which indicates that the selected covariates can fully explain the association of Y with $X_1, \cdots, X_p$, might be expected to hold approximately when using a consistent (such as Bayesian Information Criterion) or over-consistent (such as Akaike Information Criterion) variable selector (Akaike 1973, Shibata 1976, Schwarz 1978, Pötscher 1989). Condition (C3) requires that after eliminating the linear effect of X_S from $X_{\setminus S}$, the error term $\tilde{\epsilon}$ is independent of X_S. This can be assessed by calculating the residuals from a least-squares regression of $X_{\setminus S}$ on X_S and plotting these against the covariates in X_S.

Below we give several results specifying conditions under which the proposed methods can be used to make an inference about a global hypothesis (Propositions 2-4) or an individual covariate (Propositions 5-7) from (2.2). Propositions 5-7 generalize in a straightforward way to any subset of the selected covariates (see the Appendix).

Consider the global hypothesis $H_0 : \beta_S = \beta_S^0$ for some β_S^0.

PROPOSITION 2. *Suppose that (C1) holds. Let $g(X, Y) = (X, \tilde{Y})$, where $\tilde{Y} = Y - \beta_S^0 X_S$, and $\tilde{\mathbf{D}}_P = \mathbf{X}_S$, $\tilde{\mathbf{D}}_F = (\mathbf{X}_{\setminus S}, \tilde{\mathbf{Y}})$. Then $g(\cdot)$ and the partition $(\tilde{\mathbf{D}}_P, \tilde{\mathbf{D}}_F)$ satisfy the conditions of Theorem 1.*

PROPOSITION 3. *Suppose that (C2) holds under H_0. Let $g(X, Y) = (X, \tilde{Y})$, where $\tilde{Y} = Y - \beta_S^0 X_S$, and $\tilde{\mathbf{D}}_P = \mathbf{X}$, $\tilde{\mathbf{D}}_F = \tilde{\mathbf{Y}}$. Then $g(\cdot)$ and the partition $(\tilde{\mathbf{D}}_P, \tilde{\mathbf{D}}_F)$ satisfy the conditions of Theorem 1.*

PROPOSITION 4. *Suppose that (C3) holds. Define* $Z_{\setminus S} = X_{\setminus S} - \gamma_S X_S$. *Let* $g(X,Y) = (X_S, Z_{\setminus S}, \tilde{Y})$, *where* $\tilde{Y} = Y - \beta_S^0 X_S$, *and define* $\tilde{\mathbf{D}}_P = \mathbf{X}_S$ *and* $\tilde{\mathbf{D}}_F = (\mathbf{Z}_{\setminus S}, \tilde{\mathbf{Y}})$. *Then* $g(\cdot)$ *and the partition* $(\tilde{\mathbf{D}}_P, \tilde{\mathbf{D}}_F)$ *satisfy the conditions of Theorem 1.*

When multiple covariates are selected, it is often of interest to test a hypothesis about a single covariate. Without loss of generality, let $X_1 \in \mathcal{X}_S$, and consider $H_0 : \beta_1 = \beta_1^0$ for some β_1^0. Below we use $X_{S\setminus 1}$ and $X_{\setminus 1}$ to denote the random vectors formed by the variables in $\mathcal{X}_S \setminus \{X_1\}$ and $\mathcal{X} \setminus \{X_1\}$, respectively.

PROPOSITION 5. *Suppose that (C2) holds under* H_0 *and* $X_1 \perp X_{\setminus 1}$. *Let* $g(X,Y) = (X, \tilde{Y})$, *where* $\tilde{Y} = Y - \beta_1^0 X_1$, *and let* $\tilde{\mathbf{D}}_P = \mathbf{X}_1$, *and* $\tilde{\mathbf{D}}_F = (\mathbf{X}_{\setminus 1}, \tilde{\mathbf{Y}})$. *Then* $g(\cdot)$ *and the partition* $(\tilde{\mathbf{D}}_P, \tilde{\mathbf{D}}_F)$ *satisfy the conditions of Theorem 1.*

PROPOSITION 6. *Suppose that (C2) holds under* H_0, X_1 *is a continuous covariate, and* $X_1 = \gamma_{\setminus 1} X_{\setminus 1} + \epsilon_1$, *where* $\epsilon_1 \perp X_{\setminus 1}$. *Define* $Z_1 = X_1 - \gamma_{\setminus 1} X_{\setminus 1}$. *Let* $g(X,Y) = (Z_1, X_{\setminus 1}, \tilde{Y})$, *where* $\tilde{Y} = Y - \beta_1^0 Z_1$, *and* $\tilde{\mathbf{D}}_P = \mathbf{Z}_1$, $\tilde{\mathbf{D}}_F = (\mathbf{X}_{\setminus 1}, \tilde{\mathbf{Y}})$, *then* $g(\cdot)$ *and the partition* $(\tilde{\mathbf{D}}_P, \tilde{\mathbf{D}}_F)$ *satisfy the conditions of Theorem 1.*

PROPOSITION 7. *Suppose (C3) holds,* X_1 *is continuous and* $X_1 = \delta_{S\setminus 1} X_{S\setminus 1} + \epsilon_1$, *where* $\epsilon_1 \perp X_{S\setminus 1}$. *Define* $Z_1 = X_1 - \delta_{S\setminus 1} X_{S\setminus 1}$, *and* $Z_{\setminus S} = X_{\setminus S} - \gamma_1 Z_1$. *Let* $g(X,Y) = (Z_1, X_{S\setminus 1}, Z_{\setminus S}, \tilde{Y})$, *where* $\tilde{Y} = Y - \beta_1^0 Z_1$, *and let* $\tilde{\mathbf{D}}_P = \mathbf{Z}_1$, *and* $\tilde{\mathbf{D}}_F = (\mathbf{X}_{S\setminus 1}, \mathbf{Z}_{\setminus S}, \tilde{\mathbf{Y}})$. *Then* $g(\cdot)$ *and the partition* $(\tilde{\mathbf{D}}_P, \tilde{\mathbf{D}}_F)$ *satisfy the conditions of Theorem 1. In the special case where* X *is normally distributed,* $\delta_{S\setminus 1} = \Sigma_{1,S\setminus 1}\Sigma_{S\setminus 1}^{-1}$, *where* $\Sigma_{1,S\setminus 1} = cov(X_1, X_{S\setminus 1})$ *and* $\Sigma_{S\setminus 1} = var(X_{S\setminus 1})$.

In practice, the covariances matrices in Propositions 4 and 7 are usually not known, in which case they can be replaced by their empirical counterparts, resulting in an approximate inference. Similarly, in Propositions 4, 6, and 7, the γ_S,$\gamma_{\setminus 1}$, and $\delta_{S\setminus 1}$ can be replaced by their least-squares estimates. The linear representation for X_1 assumed in Propositions 6 and 7 can be assessed empirically by plotting residuals against the values of other covariates.

Each of Propositions 2-7 gives a transformation and partition that satisfy the conditions of Theorem 1 for specific hypotheses and assumptions about the distribution of $(X_1, \cdots, X_p, Y)$. Propositions 2, 4, 5, 6, 7 represent applications of Condition A while Proposition 3 represents an application of Condition B. Inferences then follow from the transformation/partition and permutation/restriction steps described in Section 2.2.

In the above, we began by assuming a linear model relating Y to the entire set of candidate covariates. Alternatively, we could have begun by assuming a linear model for the marginal association between Y and the selected covariates, and then imposed other conditions on the error term of this model and/or on the joint distribution of the covariates to apply the proposed methods.

The proposed methods can sometimes be applied without specifying a particular parametric model. For example, consider the null hypothesis $H_0 : Y \perp X_S$ that Y and X_S are independent. If $X_S \perp X_{\setminus S}$, Condition A holds with $\tilde{X}_P = X_S$, $\tilde{X}_F = X_{\setminus S}$, and $\tilde{Y} = Y$, so that we can apply Theorem 1 with $\tilde{\mathbf{D}}_P = \mathbf{X}_S$ and $\tilde{\mathbf{D}}_F = (\mathbf{X}_{\setminus S}, \mathbf{Y})$. Alternatively, suppose $Y \perp X_{\setminus S} \mid X_S$, which can be viewed as saying that $\mathcal{X}_S$ includes all of the important covariates. Here Condition B would hold under H_0, so that we can apply Theorem 1 with $\tilde{\mathbf{D}}_P = \mathbf{X}$ and $\tilde{\mathbf{D}}_F = \mathbf{Y}$.

3. SIMULATION STUDIES

In this section we use simulations to assess and compare the performance of the proposed methods to the naive and sample-splitting approaches for several specific settings. For a given outcome, x_S, of the variable selection step, we estimate the Type I error and power for the proposed methods and naive approach by first generating independent data sets and then, from among those that result in x_S being selected, computing the proportion that reject the null hypothesis. For the

sample-splitting approach, we first generate independent data sets and randomly split each in half, using the first half for variable selection. From among those data sets that result in x_S, we then use the second half to fit the marginal model and record the proportion that reject the null hypothesis. Confidence intervals are obtained similarly; the coverage probability is estimated by the proportion of the resulting intervals that include the true parameter value. All simulations were done using R, version 2.2.0 or later.

3.1 Linear regression models following variable selection

3.1.1 Parameters of interest

We begin this section with an example to illustrate the difference between the parameters of interest in this paper and those in models for the association between the response variable and all the candidate covariates. Similar to the setting considered by Shen *et al.* (2004), suppose (X_1, X_2, X_3, X_4) follows a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix having $(i,j)^{th}$ element $\rho^{|i-j|}$, and that conditional on (X_1, X_2, X_3, X_4),

$$Y = \beta_1^* X_1 + \beta_2^* X_2 + \beta_3^* X_3 + 0 \cdot X_4 + \epsilon^*, \tag{3.1}$$

where $\epsilon^* \sim N(0,1)$ and independent of (X_1, X_2, X_3, X_4). In Shen *et al.* (2004), the variable selection must always include X_1 and the parameter of interest is β_1^*, regardless of which subset of (X_2, X_3, X_4) is also selected. Our focus is on the marginal association between the selected covariates and the response. When X_1, X_2 and X_3 are selected, the parameters we consider are the same as those in (3.1), but otherwise they differ. For example, when only X_1 is selected, we no longer are interested in β_1^*, the coefficient for X_1 from the linear model that adjusts for X_2 and X_3, but instead in β_1 from the marginal model $Y = \beta_1 X_1 + \epsilon$, where $\beta_1 = \beta_1^* + \rho\beta_2^* + \rho^2\beta_3^*$, $\epsilon = \beta_2^*(X_2 - \rho X_1) + \beta_3^*(X_3 - \rho^2 X_1) + \epsilon^*$, and $X_1 \perp \epsilon$ (Proposition 1).

Table 1: Empirical Type I Errors for Testing $H_0 : \beta_1 = \beta_1^0$ at Nominal .05 Level, Using the Naive, Sample-Splitting, and Restricted Permutation Approaches with Actual (c) and Estimated ($\hat{c}$) Values of c, for Different ρ. Each Entry Based on 2,000 Replications.

$\beta_2^* = .05$					
ρ	β_1^0	Naive	Sample-Splitting	Perm (c)	Perm ($\hat{c}$)
.5	.1275	.13	.057	.049	.054
-.5	.0775	.13	.055	.047	.059
.8	.1464	.12	.053	.048	.061
-.8	.0664	.12	.055	.058	.062
$\beta_2^* = .1$					
ρ	β_1^0	Naive	Sample-Splitting	Perm (c)	Perm ($\hat{c}$)
.5	.1525	.12	.055	.050	.061
-.5	.0525	.13	.059	.042	.069
.8	.1864	.12	.052	.043	.068
-.8	.0264	.12	.056	.046	.061

Suppose $\mathcal{X}_S = \{X_1\}$. Table 1 displays the empirical Type I errors for testing $H_0 : \beta_1 = \beta_1^0$ using the naive approach, the sample-splitting approach, and restricted permutation test A using both actual (c) and estimated ($\hat{c}$) covariances, where $c = \Sigma_{\setminus S,S}\Sigma_S^{-1}$ (Proposition 4). The variable selector is a step-down procedure (step() function in R), starting with the model that includes all p covariates, and using the Akaike Information Criterion (Akaike 1973) to eliminate covariates. Here $\beta_1^* = .1, \beta_2^* = .1, .05$, $\beta_3^* = .01$, $\rho = .5, -.5, .8, -.8$ and, as in Shen and others, $n = 22$. The

Type I error estimates using the restricted permutation test with actual c or sample-splitting are very close to the nominal level of .05; those based on the restricted permutation test using $\hat{c}$ are close to the nominal level and offer a substantial improvement over those obtained from the naive approach.

3.1.2 Testing a global hypothesis

Suppose $(X_1, \cdots, X_5)$ has a multivariate normal distribution with mean $(.1, .2, .3, .4, .5)$, $X_1 \perp X_5$, $(X_1, X_5) \perp (X_2, X_3, X_4)$, $var(X_j) = 1$ for $j = 1, \cdots, 5$, and $corr(X_j, X_k) = .1$ for $j, k = 2, 3, 4$ and $j \neq k$. Assume that the model for Y, given $(X_1, X_2, \cdots, X_5)$, is $Y = .5 + \beta_1^* X_1 + \epsilon^*$, where $\epsilon^* \sim N(0, 1)$ and independent of $(X_1, \cdots, X_5)$. We take $n = 100$ and use the same variable selector as in section 3.1.1.

Suppose $\mathcal{X}_S = \{X_1, X_5\}$. From Proposition 1, the marginal association of (X_1, X_5) and Y is given by $Y = \beta_0 + \beta_1 X_1 + \beta_5 X_5 + \epsilon$. Consider testing the hypothesis $H_0 : \beta_1 = \beta_5 = 0$ using the two degrees of freedom likelihood ratio test statistic. Figure 1 gives the power of restricted permutation test A (Proposition 2), permutation test B (Proposition 3), and sample-splitting for increasing values of β_1 and for $\beta_5 = 0$. Results are based on 1000 simulations for Type I error and 200 simulations for power. The Type I error of the naive test is highly distorted (=0.559), whereas those for the permutation and sample-splitting approaches are very close to the nominal .05 level. The power curves of the two permutation tests are similar, especially in power ranges of typical interest, and larger than that of the sample-splitting approach.

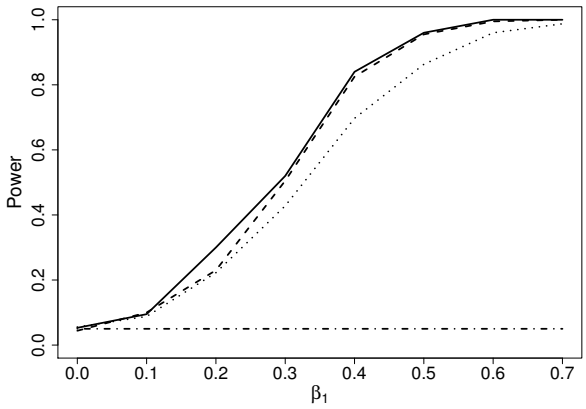

Figure 1: Power Comparison of Restricted Permutation Test A (Dashed), Restricted Permutation Test B (Solid), and Sample-Splitting (Dotted) for Testing $H_0 : \beta_1 = \beta_5 = 0$, for $\beta_1 = 0, .1, \cdots, .7$. The dot-dash line represents a horizontal line at .05.

3.1.3 Testing a subset of the coefficients

Suppose $p = 3$ and $X = (X_1, X_2, X_3)$ has multivariate normal distribution with mean vector $(.1, .2, .3)$ and covariance matrix $\Sigma = (\sigma_{ij})$, where $\sigma_{ii} = 1$ and $\sigma_{ij} = \rho$ for $i \neq j$. Assume the conditional distribution of Y, given (X_1, X_2, X_3), is given by $Y = .5 + .3X_1 + 0X_2 + 0X_3 + \epsilon^*$, where $\epsilon^* \sim N(0, 1)$ and independent of (X_1, X_2, X_3), and consider the same variable selector as section 3.1.1. Suppose $\mathcal{R}$ selects $\{X_1, X_2\}$, and we wish to test hypotheses about parameters from the marginal model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$ (Proposition 1).

To test $H_0 : \beta_1 = \beta_1^0$, we use Proposition 7 and define $Z_1 = X_1 - c_1 X_2$, where $c_1 = cov(X_1, X_2)/var(X_2) = \rho$, and $Z_3 = X_3 - c_2 Z_1$, where $c_2 = cov(X_3, Z_1)/var(Z_1) = \rho/(1+\rho)$. Let $\tilde{Y} = Y - \beta_1^0 Z_1$. To test $H_0 : \beta_2 = \beta_2^0$, we transform X_2 to $Z_2 = X_2 - c_1 X_1$, where $c_1 = cov(X_1, X_2)/var(X_1) = \rho$, and $Z_3 = X_3 - c_2 Z_2$, where $c_2 = cov(X_3, Z_2)/var(Z_2) = \rho/(1+\rho)$, and $\tilde{Y} = Y - \beta_2^0 Z_2$. Note that we use a different transformation for testing β_2 to ensure that this remains the coefficient being tested after the transformation.

Table 2: Empirical Type I Errors for the Naive, Sample-Splitting, and Restricted Permutation Approaches Using Actual (c) and Estimated ($\hat{c}$) Values of ($c = (c_1, c_2)$) for Different Sample Size (n) and $\rho = cov(X_i, X_j)$, $i \neq j$, Based on Nominal .05 Level Tests. Each Entry Based on 2,000 Replications.

$H_0 : \beta_1 = .3$					
n	ρ	Naive	Sample-Splitting	Perm(c)	Perm($\hat{c}$)
50	.6	.15	.048	.054	.043
100	.6	.11	.050	.050	.040
100	0	.031	.052	.053	.048
500	-.2	.061	.060	.045	.042
$H_0 : \beta_2 = 0$					
n	ρ	Naive	Sample-Splitting	Perm(c)	Perm($\hat{c}$)
50	.6	.30	.051	.049	.029
100	.6	.30	.055	.049	.034
100	0	.30	.042	.051	.041
500	-.2	.31	.045	.054	.034

Table 2 gives the empirical Type I errors for the naive test, the restricted permutation test A, and the sample-splitting approach for different choices of n and ρ, and using the actual (c) and empirically-estimated ($\hat{c}$) values of $c = (c_1, c_2)$, based on the test statistic $\sum_{j=1}^{n}(Y_j - \overline{Y})(Z_{ij} - \overline{Z}_i)$, where $i = 1, 2$ refer to Z_1 and Z_2. The permutation test using the actual values of c and the sample-splitting approach lead to Type I errors that are very close to the nominal .05 level. Use of empirical estimates leads to somewhat conservative tests while the naive test gives distorted Type I errors.

3.2 Two-sample problems following variable selection

Suppose that we have p binary covariates $X_1, \cdots, X_p$. Define $p_j = P(X_j = 1)$ and define β_k to be the difference in mean response for the two levels of X_k; that is, $\beta_k = E(Y \mid X_k = 0) - E(Y \mid X_k = 1)$, $k = 1, 2, \cdots, p$. Suppose that $\mathcal{R}(X) = \{X_j\}$ for some $j \in \{1, 2, \cdots, p\}$ and we want to test $H_0 : \beta_j = \beta_j^0$. A natural transformation to eliminate the mean location difference induced by X_j under H_0 is $\tilde{Y} = Y + \beta_j^0 X_j$, since then we have $E(\tilde{Y} \mid X_j = 0) = E(\tilde{Y} \mid X_j = 1)$.

Suppose that $X_3, \cdots, X_p$ are mutually independent and independent of (X_1, X_2), and $corr(X_1, X_2) = \rho$. Suppose that $f(Y \mid X_1, \cdots, X_p) = f(Y \mid X_1)$ and $f(y|X_1 = x) = f_0(y - x\beta)$ for some β and density f_0. It is easily verified that $\beta_1 = \beta$, $\beta_2 = \beta\rho\sqrt{p_1(1-p_1)}/\sqrt{p_2(1-p_2)}$, and $\beta_j = 0$ for $j = 3, 4, \cdots, p$.

When X_1 is selected, Condition (C2) holds and we can apply Proposition 3 (Permutation Test B). Suppose X_j ($j \geq 3$) is selected, Condition (C1) holds and we can apply Proposition 2 (Permutation Test A). Suppose X_2 is selected. When $\rho \neq 0$, X_2 is marginally associated with Y because of its correlation with X_1. Condition (C1) does not hold because X_2 is not independent of X_1. Neither does Condition (C2) hold for this $\tilde{Y}$ because $\tilde{Y}$ is not independent of X ($Var(\tilde{Y} \mid$

$X_2) \neq Var(\tilde{Y})$). We include this case below to assess the robustness of the permutation tests when the conditions for their validity are not satisfied.

3.2.1 Validity and Robustness

Consider the variable selector defined by $\mathcal{R}(\mathbf{D}) = \{X_j\}$ if $V_j = argmax\{V_1, \cdots, V_p\}$, where V_j is the t-test statistic based on the difference between the average values of Y when $X_j = 1$ and $X_j = 0$. We use the usual t-statistic for the permutation tests. For $p = 10$, $p_1 = .1$, $p_j = .5$ for $j \geq 2$, $\epsilon \sim N(0, 1)$, $\rho = .2$, and $n = 100$, Figure 2 displays empirical Type I error rates for testing $H_0 : \beta_j = \beta_j^0$, $(j = 1, 2, 3)$ using the naive approach, permutation tests A and B, and the sample-splitting approach when the selected covariate is X_1 (panel a), X_2 (panel b), or X_3 (panel c), for β_1 varying from 0 to 1. The condition for permutation test A does not hold in panel (a) or (b), and the condition for permutation test B does not hold in panel (b) and (c); hence the respective curves indicate the robustness of these tests to violations of their conditions. In all cases, a nominal .05 level test and 2000 simulations were used.

In panel (a), the naive test is severely biased for smaller values of β_1 and improves for increasing β_1 because the probability of selecting X_1 approaches 1. The Type I error rates for both permutation tests and for the sample-splitting approach are close to the nominal values for all values of β_1, even though the conditions for restricted permutation test A are not met. The small distortion for permutation test A is in part due to the small dependence between X_1 and other covariates.

In panel (b), where X_2 is selected, the naive test has highly distorted Type I error for all β_1. The conditions for both permutation tests are violated, and their Type I errors climb above .05 when $\beta_1 > .5$ ($\beta_2 > .06$), increasing to .15 when $\beta_1 = 1$ ($\beta_2 = .12$). Compared to the case where X_1 is selected, the larger distortion using permutation test A is likely due to the fact that X_2 is dependent on both X_1 and $\tilde{Y}$, while when $\mathcal{X}_S = \{X_1\}$, X_1 is dependent only on X_2.

In panel (c), where X_3 is selected, the curves represent the Type I errors for the hypothesis $H_0 : \beta_3 = 0$. Here the condition for permutation test A is satisfied and the actual Type I error is very close to the nominal value. The condition for permutation test B is not satisfied, resulting in an inflated Type I error which begins for $\beta_1 \approx .4$ and increases to about .15 when $\beta_1 = 1$. This occurs because as β_1 increases, the association between $\tilde{Y} = Y$ and X_1 increases, resulting in stronger dependency between $\tilde{\mathbf{D}}_P = \mathbf{X}$ and $\tilde{\mathbf{D}}_F = \mathbf{Y}$. The naive test is severely distorted for all values of β_1.

We compared the power of permutation test B to the sample-splitting approach for testing $H_0 : \beta_1 = 0$ when X_1 is selected, for β_1 ranging from 0 to 1.0, and for different choices of p_1, p_2, ρ, and n. Results are based on 2000 simulations for Type I error and 1000 simulations for power. In all cases that we examined, the power of the restricted permutation test exceeds that of the sample-splitting approach. A typical pattern of the comparisons is shown in Figure 3 for $p_1 = p_2 = .1$ and varying ρ $(= 0, .2, .5)$ and n $(= 100, 500)$. Very similar results were obtained for other choices of parameters values (data available upon request). Also shown in Figure 3 are the power curves for the unachievable test of association between X_1 and Y without model selection (that is, if we fit the marginal model relating X_1 and Y without a variable selection step). The lower power of the restricted permutation test relative to the test that does not undertake model selection can be viewed as the "cost" of variable selection.

Table 3 summarizes the performance of the nominal 95% confidence intervals obtained from the naive approach, the sample-splitting approach, and from permutation tests A and B when $n = 100,\ 500$ and $\beta_1 = .05,\ .2$. The shaded areas represent settings where the conditions (specified in Propositions 2 and 3) for the restricted permutation tests are satisfied, and thus the confidence intervals are exact. For the remaining values, the corresponding coverage results reflect the robustness of the methods to violations of these conditions. The coverage of the naive confidence intervals is often substantially lower than the nominal value. The coverage of the restricted permutation tests is very close to the nominal value when the conditions for these tests hold, and

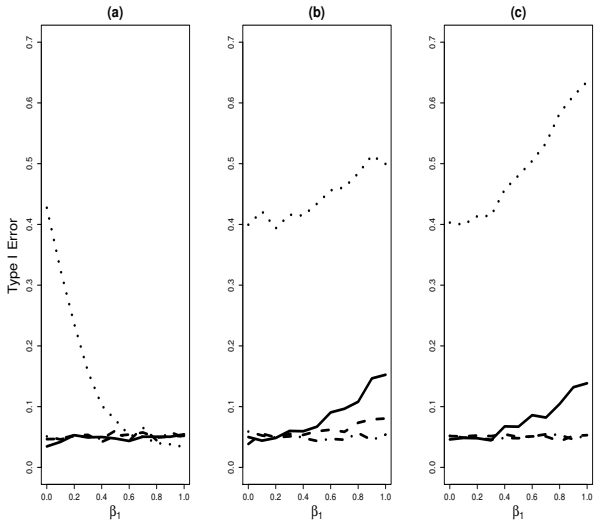

Figure 2: Type I Error Estimates. Panel (a): $\mathcal{X}_S = \{X_1\}$, Testing H_0: $\beta_1 = \beta_1^0$. Panel (b): $\mathcal{X}_S = \{X_2\}$, Testing H_0: $\beta_2 = \beta_2^0$, Where $\beta_2 = 2\rho\beta_1\sqrt{p_1(1-p_1)}$. Panel (c): $\mathcal{X}_S = \{X_3\}$, Testing H_0: $\beta_3 = \beta_3^0$. Horizontal Axis Refers to Different Values of True β_1, Ranging from 0 to 1, in Increments of .1. Dotted: Naive Test; Dashed: Restricted Permutation Test A; Solid: Restricted Permutation Test B; Dotdash: Sample-Splitting.

offer a substantial improvement over naive methods when they are violated. Note also that the width of the restricted permutation test intervals are only moderately greater than those for the naive approach. Because the width of the naive approach coincides with the width for the approach without variable selection, the relative increase in the width of the permutation intervals can be viewed as the "cost" paid for applying this variable selector. While the coverage for the sample-splitting approach is always close to the nominal value, the intervals are wider than those obtained from the restricted permutation methods, reflecting a loss of efficiency because only 50% of the sample was used in their construction.

3.2.2 Two variable selectors

The proposed methods allow comparisons of different variable selectors. Consider the same setting as above but now with $p = 3$ covariates, and with $p_1 = .1$, $p_2 = .2$, $p_3 = .3$, $\rho = .5$, and $\beta_1 = 0, .2$. Let $\mathcal{R}_1$ be the selector used in Section 3.2.1, and let $\mathcal{R}_2$ select those covariates whose marginal association with Y yields a t-statistic greater in absolute value than 1.96. Table 4 gives the actual coverage probabilities for the nominal 95% confidence intervals from the naive approach, the sample-splitting approach, and for permutation test B when $\mathcal{X}_S = \{X_1\}$. The coverage probabilities for the intervals obtained from the restricted permutation methods and the sample-splitting approach are all very close to the nominal levels, while those for the naive approach are lower than the nominal values. The intervals obtained from sample-splitting are wider than those obtained from the restricted permutation methods for both variable selectors. The permutation intervals following use of $\mathcal{R}_2$ are about 10% wider than those following use of $\mathcal{R}_1$. For $\mathcal{R}_2$, where the distortion of the naive intervals is greater, the exact permutation intervals are 16%-18% wider than the naive intervals. In contrast, for $\mathcal{R}_1$, where the naive intervals are less distorted, the exact permutation intervals are only 3%-5% wider. Because the expected width of the naive intervals is the same as that if there were no variable selection, one can view the "cost" of correcting for variable selection as being greater for $\mathcal{R}_2$ than for $\mathcal{R}_1$.

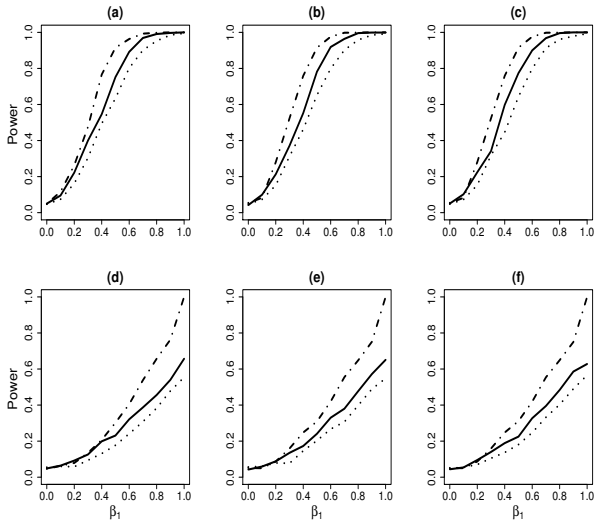

Figure 3: Power Comparison of Permutation Test B (Solid) to Sample-Splitting (Dotted), when $\mathcal{X}_S = \{X_1\}$ and Testing $H_0 : \beta_1 = 0$ in Example 1, for $p_1 = p_2 = .1$, when $n = 500$ (panels (a)-(c)) or 100 (panels (d)-(f)) and $\rho = 0$ (panels (a), (d)), .2 (panels (b), (e)), or .5 (panels (c), (f)). Dotdash Curve Represents Power of Test of $\beta_1 = 0$ without Model Selection.

4. AN EXAMPLE

We illustrate the proposed methods using the results from an immunological study conducted by the AIDS Clinical Trials Group (Malhotra *et al.*, 2004) to assess the association between several immunological markers measured at enrollment and Y, the change in CD4+ T-cell count by study week 24. There were $n = 59$ patients and $p = 15$ covariates consisting of nine immunological markers obtained from flow cytometry, four stimulation indices obtained from lymphocyte proliferation (LP) assays, and two binary variables to describe which of three treatments a patient received. The nine immunological markers were the percentage of CD8 T cells in lymphocyte; the percentages of CD4+ T cells expressing: naive markers (nCD4 %), activation markers, coexpressing CD28+, and coexpressing Fas; and the percentages of CD8+ T cells expressing: naive markers (nCD8%), activation markers, coexpressing CD28+, and coexpressing Fas (CD8+95+%). The four stimulants used in LP assays were baculovirus control protein, Candida, baculovirus-expressed recombinant HIV-1_{LAI} p24, and HIV-1_{MN} gp160.

We considered the variable selector that separately examines the marginal association of each covariate with Y, using least squares, and selects all covariates showing some evidence of being associated with Y, defined as a significance level (p-value) of .10 or less. This led to the selection of 3 covariates: nCD4% (p=.002), nCD8% (p=.036), and CD8+95+% (p=.021). We then considered a linear model for the marginal association of these covariates with Y. As seen in the left side of Table 5, the naive multivariate (least-squares) analysis indicates that Y is significantly associated with nCD4% (p=.005), possibly associated with CD8+95% (p=.09), and not associated with nCD8% (p=.37). The disparate p-values for the univariate and multivariate analysis of nCD8% are likely due to its correlation (.38) with nCD4%. The naive likelihood ratio test of the global hypothesis that none of the three covariates are associated with Y gives $p < .001$.

We subsequently analyzed the marginal association between the three selected covariates and response using the restricted permutation methods (Table 5, right). We employed Proposition 3 to assess these global hypothesis that none were associated with response, using the log-likelihood

Table 3: Empirical Coverage Probabilities and Mean ± SD of Width of Nominal 95% CIs for β_j $(j = 1, 2, 3)$ for Different β_1 and Sample Sizes (n). Based on 500 Replications. Underlined Boldfaced Numbers Indicate Situations Where Conditions for Restricted Permutation Tests Are Met. Numbers in Parentheses Refer to the Frequency of a Specific Variable Selection Outcome.

β_1	n	$\mathcal{X}_S$	β_j	Naive	Sample-Splitting	Perm A	Perm B
.2	500	X_1	.2	95.3%	94.0%	94.8%	**94.6%**
		(.62)		.35 ± .01	.50 ± .02	.41 ± .05	**.40 ± .03**
		X_2	.1	79.5%	94.7%	94.2%;	89.0%
		(.08)		.35 ± .01	.50 ± .02	.43 ± .04	.40 ± .03
		X_3	0	44.0%	94.2%	**94.8%**	87.6%
		(.04)		.35 ± .01	.50 ± .02	**.40± .03**	.40 ± .02
.05	500	X_1	.05	74.5%	94.6%	97.0%	**94.0%**
		(.13)		.35 ± .01	.50 ± .02	.40 ± .03	**.39± .02**
		X_2	.03	63%	94.4%	96.8%	93.8%
		(.09)		.35 ± .01	.50 ± .02	.40 ± .03	.39 ± .02
		X_3	0	59.5%	93.8%	**93.4%**	94.4%
		(.10)		.35 ± .01	.50 ± .02	**.39± .03**	.39 ± .02
.2	100	X_1	.2	86.2%	95.6%	96.0%	**93.6%**
		(.21)		.79 ± .06	1.15 ± .12	.92 ± .10	**.89± .08**
		X_2	.1	66.6%	96.2%	94.6%	92.8%
		(.09)		.79 ± .06	1.15 ± .12	.90 ± .09	.89 ± .07
		X_3	0	57.0%	95.8%	**95.0%**	94.0%
		(.09)		.79 ± .06	1.16 ± .12	**.88± .09**	.89 ± .08
.05	100	X_1	.05	63.6%	96.0%	96.0%	**94.6%**
		(.10)		.79 ± .06	1.15 ± .12	.88 ± .09	**.89± .07**
		X_2	.03	60.5%	94.8%	95.6%	95.2%
		(.10)		.78 ± .06	1.15 ± .12	.88 ± .09	.88 ± .07
		X_3	0	59.9%	95.4%	**94.0%**	95.6%
		(.10)		.79 ± .06	1.15 ± .12	**.89± .08**	.87 ± .07

ratio statistic. Use of Proposition 3, which assumes that the selected covariates can fully explain the association between Y and all the covariates (Condition C2), was motivated by the liberal criterion ($p \leq .10$) used to select covariates. The global permutation test provides evidence (p=.026) against the hypothesis that all regression coefficients were zero. We then employed Proposition 6 to assess the individual covariates. Proposition 6 also assumes that each of the 3 selected covariates can be expressed as a linear combination of the remaining covariates plus an unrelated error term. This assumption was empirically supported by examining scatter-plots of the residuals from each fitted model relating a selected covariate with the remaining covariates. As a test statistic, we used the least-squares estimator of the coefficient of Z_1 (see Proposition 6) in the model relating $\tilde{Y}$ with Z_1 and the other two selected covariates. After correcting for variable selection, the association of nCD4% and response remained significant (p=.027), the association for CD8+95+% became marginally significant (p=.051) and the association for nCD8% remained non-significant. The 95% confidence intervals obtained from the restricted permutation methods are wider than the naive confidence intervals, incorporating the additional uncertainty about the regression coefficients due to variable selection.

With $p = 15$ covariates and only $n = 59$ observations, most analysts would not likely use a sample splitting approach to analyze these data out of concern that some important covariates

Table 4: Empirical Coverage Probabilities and Mean ± SD of the Width of the Nominal 95% CIs for β_1, for Different Variable Selectors. Based on 1000 Replications.

Selector 1 ($\max(\lvert T_j \rvert)$)			
β_1	Naive	Sample-Splitting	Perm B
0			
	86.1%	96%	95.4%
	$1.35 \pm .23$	$1.89 \pm .35$	$1.39 \pm .24$
.2			
	90.5%	95.9%	95.9%
	$1.34 \pm .22$	$1.89 \pm .35$	$1.37 \pm .23$
Selector 2 (all $\lvert T_j \rvert \geq 1.96$)			
β_1	Naive	Sample-Splitting	Perm B
0			
	5.2%	95.7%	95%
	$1.33 \pm .22$	$1.86 \pm .35$	$1.55 \pm .28$
.2			
	71.5%	95.5%	94.4%
	$1.32 \pm .22$	$1.91 \pm .35$	$1.55 \pm .28$

Table 5: Results from the AIDS Study

	Naive		Perm	
Covariate	P-value	95% CI	P-value	95% CI
nCD4 %	.005	(.84, 4.52)	.027	(.32, 5.22)
nCD8 %	.37	(-1.27, 3.37)	.17	(-2.60, 6.64)
CD8+95+ %	.09	(-.18, 2.63)	.051	(-.05, 4.83)
Overall	$< .001$		.026	

might not be selected, and that even if selected, there might be inadequate power to demonstrate their association with Y. Using the same variable selector and randomly splitting the data into variable selection (n=30) and testing (n=29) sets resulted in nCD4% being selected 87% of the time. When the same 3 covariates were selected as with the full data, nCD4% was no longer significant ($p < .05$) 32% of the time.

5. DISCUSSION

Despite the recognized biases resulting from use of naive inference methods that do not account for variable selection, such methods are still commonly used in practice, in large part due to a lack of alternatives. The methods proposed in this paper, based on restricted permutation tests, provide exact or approximate inferences about the marginal association between a subset of covariates and a response variable under specific conditions. The methods are not tied to any specific variable selector and do not require that the inference be restricted to covariates that are not candidates for exclusion by the variable selector. The proposed methods do not apply to all settings, as indicated by the assumptions made in Propositions. However, an advantage is that conditions for their validity can be specified and most can be checked empirically. When a lenient (non-parsimonious) variable selector is used, it is more likely that all important covariates in the full model will be selected. In such cases, the regression coefficient for a covariate in the marginal model will be the same as the coefficient in the full model, and hence the proposed methods can be used to make

inferences on associations in the overall model. However, when some important covariates are not selected, the regression coefficients in the full and marginal models will not, in general, be the same.

The results in Figures 1 and 3 suggest that use of the proposed methods can provide more powerful tests than a sample-splitting approach. Because sample-splitting does not require the assumptions made by the restricted permutation methods, it would be preferred when the sample size is sufficiently large. However, in many settings it may not be feasible to split a sample, in which case the proposed methods offer an appealing alternative. Further investigation of the relative efficiency of these approaches would be worthwhile. For small or moderate sample sizes, a related advantage of the proposed methods over the sample-splitting is that, due to the larger sample size on which they are based, it is more likely that important covariates will be identified.

Several considerations arise in the implementation of the methods. Firstly, rather than enumerating all $n!$ permutations to identify the restricted subset Π_R, it is sufficient to sample from the $n!$ permutations. In the results presented in sections 3 and 4, we sampled enough permutations to yield 1000 that met the restriction criterion, and found that this gave very similar results as when 2000 were used. Secondly, when the variable selector is sequential (cf: DiRienzo *et al.* 2003) or in certain special cases, the variable selector does not need to be fully evaluated on each reverse-transformed permuted dataset to determine whether it is in the restricted set. For example, in section 3.2 when X_3 is selected and the Permutation Test A is applied, $V_1, V_2, V_4, \cdots, V_{10}$ do not change when evaluating the variable selector for the $n!$ permuted datasets, and thus need not be re-computed. Finally, when constructing a confidence region for a parameter, our experience has been that these regions are intervals and thus computation time can be greatly shortened by searching for the interval endpoints rather than evaluating all possible parameters.

When a specific covariate is known to be of interest a priori, the variable selector would be chosen to always include this covariate. For example, in a randomized clinical trial, the covariate denoting the treatment effect would always be included by the variable selector, and other candidate covariates might be selected in the hope of leading to a more efficient inference about the treatment effect. For linear models, the randomization of treatments ensures that the regression coefficient for the treatment effect is the same regardless of which additional covariates are selected. The proposed procedure would produce valid inference for the treatment effect conditional on each of the candidate models, and therefore would also produce valid inference unconditionally. If one performs variable selection and then wishes to make an inference about a covariate that was not selected, the proposed methods also can be used. For example, if the covariate X_2 is selected and one wishes to make an inference about the marginal association between Y and the unselected covariate X_1, the proposed methods still apply, except that now the restriction set will consist of all the permuted data matrices (properly transformed) which lead to the same outcome ($\{X_2\}$) as the original dataset.

Throughout we have assumed that the explanatory variables are random. In some settings, such as when Condition B can be utilized, the proposed methods can be applied directly for fixed $\mathbf{X}$. In other settings, such as when one wishes to make inference about an individual coefficient in the regression models, we believe that the proposed approach can be modified by adapting permutation methods for fixed $\mathbf{X}$ developed for settings where there is no variable selection (cf: Huh & Jhun 2001). However, careful investigation and evaluation of the properties of such methods for fixed $\mathbf{X}$ are needed.

Although we consider linear regression models, the proposed methods can in principle be applied with any model for which an appropriate transformation and partition can be identified. Extensions to making inferences about discrete covariates in linear regression and about parameters in a Cox model for survival data are under development. It would also be useful to undertake more assessments of the robustness of the proposed methods, to develop criteria for selecting a test statistic, and to develop computationally efficient algorithms, especially when computationally-intensive variable selector is employed (cf: DiRienzo *et al.* 2003).

APPENDIX

Proof of Theorem 1: Under H_0, for any $\tilde{\mathbf{d}}(l)$ and $\tilde{\mathbf{d}}(m)$ in Π_R,

$$\begin{aligned}
P(\tilde{\mathbf{D}} = \tilde{\mathbf{d}}(l) \quad | \quad & \tilde{\mathbf{D}}_P \stackrel{p}{=} \tilde{\mathbf{d}}_P, \tilde{\mathbf{D}}_F = \tilde{\mathbf{d}}_F, \mathcal{R}(\mathbf{D}) = x_S) \\
= \quad & P(g^{-1}(\tilde{\mathbf{D}}) = g^{-1}(\tilde{\mathbf{d}}^l_P, \tilde{\mathbf{d}}_F) | \tilde{\mathbf{D}}_P \stackrel{p}{=} \tilde{\mathbf{d}}_P, \tilde{\mathbf{D}}_F = \tilde{\mathbf{d}}_F, \mathcal{R}(\mathbf{D}) = x_S) \\
= \quad & P(\mathbf{D} = g^{-1}(\tilde{\mathbf{d}}^l_P, \tilde{\mathbf{d}}_F) | \tilde{\mathbf{D}}_P \stackrel{p}{=} \tilde{\mathbf{d}}_P, \tilde{\mathbf{D}}_F = \tilde{\mathbf{d}}_F, \mathcal{R}(\mathbf{D}) = x_S) \\
= \quad & \frac{P(\mathbf{D} = g^{-1}(\tilde{\mathbf{d}}^l_P, \tilde{\mathbf{d}}_F), \mathcal{R}(\mathbf{D}) = x_S | \tilde{\mathbf{D}}_P \stackrel{p}{=} \tilde{\mathbf{d}}_P, \tilde{\mathbf{D}}_F = \tilde{\mathbf{d}}_F)}{P(\mathcal{R}(\mathbf{D}) = x_S | \tilde{\mathbf{D}}_P \stackrel{p}{=} \tilde{\mathbf{d}}_P, \tilde{\mathbf{D}}_F = \tilde{\mathbf{d}}_F)} \\
= \quad & \frac{P(\mathbf{D} = g^{-1}(\tilde{\mathbf{d}}^l_P, \tilde{\mathbf{d}}_F), \mathcal{R}(g^{-1}(\tilde{\mathbf{d}}^l_P, \tilde{\mathbf{d}}_F)) = x_S | \tilde{\mathbf{D}}_P \stackrel{p}{=} \tilde{\mathbf{d}}_P, \tilde{\mathbf{D}}_F = \tilde{\mathbf{d}}_F)}{P(\mathcal{R}(\mathbf{D}) = x_S | \tilde{\mathbf{D}}_P \stackrel{p}{=} \tilde{\mathbf{d}}_P, \tilde{\mathbf{D}}_F = \tilde{\mathbf{d}}_F)} \\
= \quad & \frac{P(\mathbf{D} = g^{-1}(\tilde{\mathbf{d}}^m_P, \tilde{\mathbf{d}}_F), \mathcal{R}(g^{-1}(\tilde{\mathbf{d}}^m_P, \tilde{\mathbf{d}}_F)) = x_S | \tilde{\mathbf{D}}_P \stackrel{p}{=} \tilde{\mathbf{d}}_P, \tilde{\mathbf{D}}_F = \tilde{\mathbf{d}}_F)}{P(\mathcal{R}(\mathbf{D}) = x_S | \tilde{\mathbf{D}}_P \stackrel{p}{=} \tilde{\mathbf{d}}_P, \tilde{\mathbf{D}}_F = \tilde{\mathbf{d}}_F)},
\end{aligned}$$

because of the row independence of $\tilde{\mathbf{D}}$ and independence of $\tilde{\mathbf{D}}_P$ and $\tilde{\mathbf{D}}_F$. Thus,

$$\begin{aligned}
P(\tilde{\mathbf{D}} = \tilde{\mathbf{d}}(l) \quad | \quad & \tilde{\mathbf{D}}_P \stackrel{p}{=} \tilde{\mathbf{d}}_P, \tilde{\mathbf{D}}_F = \tilde{\mathbf{d}}_F, \mathcal{R}(\mathbf{D}) = x_S) \\
= \quad & \frac{P(\mathbf{D} = g^{-1}(\tilde{\mathbf{d}}^m_P, \tilde{\mathbf{d}}_F), \mathcal{R}(\mathbf{D}) = x_S | \tilde{\mathbf{D}}_P \stackrel{p}{=} \tilde{\mathbf{d}}_P, \tilde{\mathbf{D}}_F = \tilde{\mathbf{d}}_F)}{P(\mathcal{R}(\mathbf{D}) = x_S | \tilde{\mathbf{D}}_P \stackrel{p}{=} \tilde{\mathbf{d}}_P, \tilde{\mathbf{D}}_F = \tilde{\mathbf{d}}_F)} \\
= \quad & P(g^{-1}(\tilde{\mathbf{D}}) = g^{-1}(\tilde{\mathbf{d}}^m_P, \tilde{\mathbf{d}}_F) | \tilde{\mathbf{D}}_P \stackrel{p}{=} \tilde{\mathbf{d}}_P, \tilde{\mathbf{D}}_F = \tilde{\mathbf{d}}_F, \mathcal{R}(\mathbf{D}) = x_S) \\
= \quad & P(\tilde{\mathbf{D}} = \tilde{\mathbf{d}}(m) | \tilde{\mathbf{D}}_P \stackrel{p}{=} \tilde{\mathbf{d}}_P, \tilde{\mathbf{D}}_F = \tilde{\mathbf{d}}_F, \mathcal{R}(\mathbf{D}) = x_S).
\end{aligned}$$

Since $\tilde{\mathbf{d}}(l)$ and $\tilde{\mathbf{d}}(m)$ are arbitrary, it follows that this common probability equals $1/M$, where M is the number of matrices in Π_R.

Proof of Proposition 1. Rewrite (2.1) as $Y = \beta^*_S X_S + \beta^*_{\setminus S} X_{\setminus S} + \epsilon^* = \beta_S X_S + \epsilon$. Under (C1), $\beta_S = \beta^*_S$ (Cochran 1938) and $\epsilon = \beta^*_{\setminus S} X_{\setminus S} + \epsilon^*$. (C1) and the independence of ϵ^* and $X = (X_S, X_{\setminus S})$ imply that $X_S \perp (X_{\setminus S}, \epsilon^*)$, and hence X_S is independent of ϵ. If (C2) holds, then $\beta^*_{\setminus S} = 0$ and $\epsilon = \epsilon^* \perp X$. If (C3) holds, then $\beta_S = \beta^*_S + \beta^*_{\setminus S}\gamma_S$ and $\epsilon = \beta^*_{\setminus S}\tilde{\epsilon} + \epsilon^*$, where $\tilde{\epsilon} = X_{\setminus S} - \gamma_S X_S$. Since $\epsilon^* \perp (X_S, \tilde{\epsilon})$ and $X_S \perp \tilde{\epsilon}$, we have $X_S \perp (\tilde{\epsilon}, \epsilon^*)$ and therefore $X_S \perp \epsilon$.

In Propositions $2 - 7^*$ below, it is easily verified that g is one-to-one, and that for any $\mathbf{d}$ in the support of $\mathbf{D}$, $g^{-1}(\tilde{\mathbf{d}}(l))$ is in the support of $\mathbf{D}$. Thus it is sufficient to show that the variables consisting of $\tilde{\mathbf{D}}_P$ and $\tilde{\mathbf{D}}_F$ are independent under H_0.

Proof of Proposition 2. Suppose (C1) holds. Under H_0, $\tilde{Y} = Y - \beta_S X_S = \beta^*_{\setminus S} X_{\setminus S} + \epsilon^*$ (see the Proof for Lemma 1), and therefore $\tilde{Y} \perp X_S \mid X_{\setminus S}$. This, combined with $X_S \perp X_{\setminus S}$, yields $X_S \perp (X_{\setminus S}, \tilde{Y})$.

Proof of Proposition 3. Under H_0, if (C2) holds, then $\tilde{Y} = \epsilon^* \perp X$.

Proof of Proposition 4. Suppose (C3) holds. Under H_0, $\tilde{Y} = \beta^*_{\setminus S} Z_{\setminus S} + \epsilon^*$ (see the Proof for Proposition 1), and thus $\tilde{Y} \perp X_S \mid Z_{\setminus S}$. This, in combination with $X_S \perp Z_{\setminus S}$, implies that $X_S \perp (Z_{\setminus S}, \tilde{Y})$.

The following propositions, denoted with asterisks, generalize Propositions 5-7 for testing an individual covariate to any proper subset of the selected covariates. We use X_H, $X_{S\setminus H}$, or $X_{\setminus H}$

to denote the vector formed by the random variables in a subset $\mathcal{X}_H = \{X_{l_1}, \cdots, X_{l_h}\} \subset \mathcal{X}_S$, $\mathcal{X}_S \setminus \mathcal{X}_H$, or $\mathcal{X} \setminus \mathcal{X}_H$. Consider $H_0 : \beta_H = \beta_H^0$, for some β_H^0.

PROPOSITION 5*. *Suppose that (C2) holds under H_0 and $X_H \perp X_{\setminus H}$. Let $g(X,Y) = (X, \tilde{Y})$, where $\tilde{Y} = Y - \beta_H^0 X_H$, and let $\tilde{\mathbf{D}}_P = \mathbf{X}_H$, and $\tilde{\mathbf{D}}_F = (\mathbf{X}_{\setminus H}, \tilde{\mathbf{Y}})$. Then $g(\cdot)$ and the partition $(\tilde{\mathbf{D}}_P, \tilde{\mathbf{D}}_F)$ satisfy the conditions of Theorem 1.*

Proof. Write (2.1) as $Y = \beta_H^* X_H + \beta_{S\setminus H}^* X_{S\setminus H} + \beta_{\setminus S}^* X_{\setminus S} + \epsilon^*$. Since $X_H \perp X_{\setminus H}$, $\beta_H^* = \beta_H$ (Cochran 1938), and when (C2) holds, $\beta_{\setminus S}^* = 0$. Thus, under H_0, $\tilde{Y} = \beta_{S\setminus H}^* X_{S\setminus H} + \epsilon^*$, and therefore $\tilde{Y} \perp X_H \mid X_{\setminus H} = (X_{S\setminus H}, X_{\setminus S})$. This combined with $X_H \perp X_{\setminus H}$ implies that $X_H \perp (X_{\setminus H}, \tilde{Y})$.

PROPOSITION 6*. *Suppose that (C2) holds under H_0 and X_{l_j}, $j = 1, 2 \cdots, h$ are continuous covariates. Suppose that $X_H = \Gamma_{\setminus H} X_{\setminus H} + e$, where $e \perp X_{\setminus H}$. Define $Z_H = X_H - \Gamma_{\setminus H} X_{\setminus H}$. Let $g(X,Y) = (Z_H, X_{\setminus H}, \tilde{Y})$, where $\tilde{Y} = Y - \beta_H^0 Z_H$, and $\tilde{\mathbf{D}}_P = \mathbf{Z}_H$, $\tilde{\mathbf{D}}_F = (\mathbf{X}_{\setminus H}, \tilde{\mathbf{Y}})$. Then $g(\cdot)$ and the partition $(\tilde{\mathbf{D}}_P, \tilde{\mathbf{D}}_F)$ satisfy the conditions of Theorem 1.*

Proof. Under (C2), we can write $Y = \beta_H^* X_H + \beta_{S\setminus H}^* X_{S\setminus H} + \epsilon^* = \beta_H^* Z_H + \beta_H^* \Gamma_{\setminus H} X_{\setminus H} + \beta_{S\setminus H}^* X_{S\setminus H} + \epsilon^* = \beta_H^* Z_H + \alpha X_{\setminus H} + \epsilon^*$, where the elements of α combine coefficient terms from $X_{\setminus H}$ and $X_{S\setminus H}$. Under H_0, $\tilde{Y} = \alpha X_{\setminus H} + \epsilon^*$, and hence $\tilde{Y} \perp Z_H \mid X_{\setminus H}$. Also note that $Z_H \perp X_{\setminus H}$. It follows that $Z_H \perp (X_{\setminus H}, \tilde{Y})$.

PROPOSITION 7*. *Suppose that (C3) holds, $X_{l_1}, \cdots, X_{l_h}$ are continuous, and $X_H = \delta_{S\setminus H} X_{S\setminus H} + \epsilon_H$, where $\epsilon_H \perp X_{S\setminus H}$. Define $Z_H = X_H - \delta_{S\setminus H} X_{S\setminus H}$ and $Z_{\setminus S} = X_{\setminus S} - \gamma_H Z_H$. Let $g(X,Y) = (Z_H, X_{S\setminus H}, Z_{\setminus S}, \tilde{Y})$, where $\tilde{Y} = Y - \beta_H^0 Z_H$, and define $\tilde{\mathbf{D}}_P = \mathbf{Z}_H$ and $\tilde{\mathbf{D}}_F = (\mathbf{X}_{S\setminus H}, \mathbf{Z}_{\setminus S}, \tilde{\mathbf{Y}})$. Then $g(\cdot)$ and the partition $(\tilde{\mathbf{D}}_P, \tilde{\mathbf{D}}_F)$ satisfy the conditions of Theorem 1. In the special case where X is normally distributed, $\delta_{S\setminus H} = \Sigma_{H,S\setminus H} \Sigma_{S\setminus H}^{-1}$, where $\Sigma_{H,S\setminus H} = cov(X_H, X_{S\setminus H})$ and $\Sigma_{S\setminus H} = var(X_{S\setminus H})$.*

Proof. Suppose that (C3) holds. $\tilde{\epsilon} \perp (X_H, X_{S\setminus H})$ implies that $\tilde{\epsilon} \perp (Z_H, X_{S\setminus H})$. This combined with $Z_H \perp X_{S\setminus H}$ (by construction) implies that $Z_H \perp (X_{S\setminus H}, \tilde{\epsilon})$. Note that $Z_{\setminus S} = X_{\setminus S} - \gamma_H Z_H = (\gamma_H \delta_{S\setminus H} + \gamma_{S\setminus H}) X_{S\setminus H} + \tilde{\epsilon}$, therefore $Z_H \perp (X_{S\setminus H}, Z_{\setminus S})$. Under H_0, $\tilde{Y} = (\beta_H^* \delta_{S\setminus H} + \beta_{S\setminus H}^*) X_{S\setminus H} + \beta_{\setminus S}^* Z_{\setminus S} + \epsilon^*$, therefore $\tilde{Y} \perp Z_H \mid (X_{S\setminus H}, Z_{\setminus S})$. Thus $Z_H \perp (X_{S\setminus H}, Z_{\setminus S}, \tilde{Y})$.

ACKNOWLEDGEMENTS

This research was supported by grants from the US National Institute of Allergy and Infectious Diseases. We thank Professor Paul Gustafson, the Associate Editor and a referee for their comments which have led to an improved version of the paper.

REFERENCES

H. Akaike (1973). Information Theory and the Maximum Likelihood Principle. In V. Petrov and F. Csáki (eds), *International Symposium on Information Theory.* Budapest: Akademiai Kiádo, 267–281.

L. Breiman (1992). The Little Bootstrap and Other Methods for Dimensionality Selection in Regression: X-fixed Prediction Error. *Journal of the American Statistical Association*, 419, 739–754.

C. Chatfield (1995). Model Uncertainty, Data Mining and Statistical Inference (with discussion). *Journal of the Royal Statistical Society Series A*, 158, 419–466.

W. G. Cochran (1938). The Omission or Addition of an Independent Variable in Multiple Linear Regression. *Supplement to the Journal of the Royal Statistical Society,* 5, 171–176.

D. R. Cox (2007). On a Generalization of a Result of W. G. Cochran. *Biometrika*, 94, 755–759.

D. L. Danilov & J. R. Magnus (2004). On the Harm That Ignoring Pre-testing Can Cause. *Journal of Econometrics*, 122, 27–46.

Y. K., Dijkstra & J. H. Veldkamp (1988). Data-Driven Selection of Regressors and the Bootstrap. In T.K.Dijkstra (eds), *On Model Uncertainty and Its Statistical Implications,* Berlin: Springer-Verlag.

G. DiRienzo, V. DeGruttola, B. Larder, & K. Hertogs (2003). Non-Parametric Methods to Predict HIV Drug Susceptibility Phenotype from Genotype. *Statistics in Medicine*, 22, 2785–2798.

B. Efron (1986). How Biased is the Apparent Error Rate of a Prediction Rule? *Journal of the American Statistical Association*, 81, 461–470.

J. J. Faraway (1992). On the Cost of Data Analysis. *Journal of Computational and Graphical Statistics*, 1, 213–229.

D. A. Freedman, W. Navidi, & S. C. Peters (1988). On the Impact of Variable Selection in Fitting Regression Equations. In Dijkstra, T.K. (eds), *On Model Uncertainty and Its Statistical Implications*, Berlin: Springer-Verlag.

K. Giri, & P. Kabaila (2008). The Coverage Probability of Confidence Intervals in 2^r Factorial Experiments after Preliminary Hypothesis Testing. *Australian & New Zealand Journal of Statistics,* 50, 69–79.

S. Gong (1986). Cross-Validation, the Jackknife, and the Bootstrap: Excess Error Estimation in Forward Logistic Regression. *Journal of the American Statistical Association,* 81, 108–118.

M. Huh & M. Jhun (2001). Random Permutation Testing in Multiple Linear Regression. *Communications in Statistics – Theory and Methods,* 30, 2023 – 2032.

C. M. Hurvich & C. Tsai (1990). The Impact of Model Selection on Inference in Linear Regression. *The American Statistician*, 44, 214–217.

P. Kabaila (1995). The Effect of Model Selection on Confidence Regions and Prediction Regions. *Econometric Theory* 11, 537-549.
—(1998). Valid Confidence Intervals in Regression After Variable Selection. *Econometric Theory*, 14, 463-482.

P. Kabaila & H. Leeb (2006). On the Large-Sample Minimal Coverage Probability of Confidence Intervals After Model Selection. *Journal of the American Statistical Association,* 101, 619–629.

H. Leeb (2005). The Distribution of a Linear Predictor After Model Selection: Conditional Finite-sample Distributions and Asymptotic Approximations. *Journal of Statistical Planning and Inference*, 134, 64–89.
— (2009). Conditional Predictive Inference Post Model Selection. *The Annals of Statistics,* forthcoming.

H. Leeb, & B. M. Pötscher (2003). The Finite-sample Distribution of Post-model-selection Estimators, and Uniform Versus Non-uniform Approximations. *Econometric Theory*, 19, 100–142.
— (2005). Model Selection and Inference: Facts and Fictions. *Econometric Theory* 21, 21–59.

I. S. Lossos, D. K. Czerwinski, A. A. Alizadeh, M. A. Wechser, R. Tibshirani, D. Botstein,& R. Levy (2004). Prediction of Survival in Diffuse Large-B-Cell Lymphoma Based on the Expression of Six Genes. *New England Journal of Medicine,* 350: 1828–1837.

U. Malhotra, R. J. Bosch, E. Chan, R. Wang, M. A. Fischl, A. C. Collier & M. J. McElrath (2004). Association of T Cell Proliferative Responses and Phenotype with Virus Control in Chronic Progressive HIV-1 Disease. *The Journal of Infectious Diseases*, 189, 515-519.

A. J. Miller (1984). Selection of Subsets of Regression Variables (with discussion). *Journal of the Royal Statistical Society Series A*, 147, 398–425.

B. M. Pötscher (1989). Model Selection Under Nonstationarity: Autoregressive models and Stochastic Linear Regression Models. *Annals of Statistics*, 17, 1257–1274.
—(1991). Effects of Model Selection on Inference. *Econometric Theory*, 7, 163-185.
—(1995). Comment on The Effect of Model Selection on Confidence Regions and Prediction Regions. *Econometric Theory* 11, 550–559.

B. M. Pötscher, & A. J. Novák (1998). The Distribution of Estimators After Model Selection: Large and Small Sample Results. *Journal of Statistical Computation and Simulation*, 60, 19–56.

G. Schwarz (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461–464.

X. Shen, H. Huang, & J. Ye (2004). Inference After Model Selection. *Journal of the American Statistical Association*, 467, 751–762.

N. S. Shulman, R. J. Bosch, J. W. Mellors, M. A. Albrecht, & D. A. Katzenstein (2004). Genetic Correlates of Efavirenz Hypersusceptibility. *AIDS*, 18, 1781–1785.

R. Shibata (1976). Selection of the Order of an Autoregressive Model by Akaike's Information Criterion. *Biometrika* , 63, 117–126.

M. Veall (1992). Bootstrapping the Process of Model Selection: An Econometric Example. *Journal of Applied Econometrics*, 7, 93–99.

P. Zhang (1992). Inference After Variable Selection in Linear Regression Models. *Biometrika*, 79, 741–746.

Received 30 October 2008
Accepted 29 June 2009

Rui WANG: `rwang@hsph.harvard.edu`
Department of Biostatistics, Harvard School of Public Health
Boston, MA 02115, USA

Stephen W. LAGAKOS: `lagakos@hsph.harvard.edu`
Department of Biostatistics, Harvard School of Public Health
Boston, MA 02115, USA