# On Measuring and Correcting the Effects of Data Mining and Model Selection

Jianming Ye

In the theory of linear models, the concept of degrees of freedom plays an important role. This concept is often used for measurement of model complexity, for obtaining an unbiased estimate of the error variance, and for comparison of different models. I have developed a concept of generalized degrees of freedom (GDF) that is applicable to complex modeling procedures. The definition is based on the sum of the sensitivity of each fitted value to perturbation in the corresponding observed value. The concept is nonasymptotic in nature and does not require analytic knowledge of the modeling procedures. The concept of GDF offers a unified framework under which complex and highly irregular modeling procedures can be analyzed in the same way as classical linear models. By using this framework, many difficult problems can be solved easily. For example, one can now measure the number of observations used in a variable selection process. Different modeling procedures, such as a tree-based regression and a projection pursuit regression, can be compared on the basis of their residual sums of squares and the GDF that they cost. I apply the proposed framework to measure the effect of variable selection in linear models, leading to corrections of selection bias in various goodness-of-fit statistics. The theory also has interesting implications for the effect of general model searching by a human modeler.

KEY WORDS: Data mining; Degrees of freedom; Effect of model selection; Goodness-of-fit; Half-normal plot; Nonparametric regression; Sensitivity.

## 1. INTRODUCTION

In the theory of linear models, the concept of degrees of freedom plays an important role. This concept has several different interpretations. The degrees of freedom in regression are the number of variables in the model. Accordingly, degrees of freedom are often used as a model complexity measure in various model selection criteria, such as Akaike information criterion (AIC) (Akaike 1973), $C_p$ (Mallows 1973), and Bayesian information criterion (BIC) (Schwarz 1978), generalized cross-validation (GCV) (Craven and Wahba 1979), and risk inflation criterion (RIC) (George and Foster 1994). Degrees of freedom can also be interpreted as the cost of the estimation process and thus can be used for obtaining an unbiased estimation of the error variance. Finally, the degrees of freedom in regression are the trace of the so-called "hat" matrix; that is, the sum of the sensitivity of each fitted value with respect to the corresponding observed value.

An extension of degrees of freedom to general model structures is useful both practically and theoretically. The last two decades have brought rapid progress in modeling high-dimensional data by means of complex statistical procedures. These procedures typically require minimum assumptions about the structures of the underlying models and try to capture the structures through adaptive fitting. But their flexibility often leads to substantial overfitting. The complex nature of these procedures makes it difficult to study their statistical behavior and to assess their performance objectively.

Traditionally, for general modeling problems, statisticians tend to define degrees of freedom based on the number of parameters in the final model, because of the coincidence of these two quantities in linear models. The focus is often on how to find the correct number of parameters. As an example, in tree-based regression models, the S-PLUS software uses the number of terminal nodes as the degrees of freedom for calculating the error variance (Venables and Ripley 1994). Friedman (1991) and many of the discussants of his paper proposed various definitions of ad hoc degrees of freedom. It has been suggested that the degrees of freedom may depend on more subtle aspects of the modeling procedures. Owen (1991) suggested that searching for a knot cost roughly 3 df.

The major difficulty in handling complex modeling procedures is that the fitted values are often complex, nondifferentiable, or even discontinuous functions of the observed values. For example, in fitting of a linear model with variable selection, a small change in $Y$ may lead to a different selected model, resulting in a discontinuity in the fitted values. Another example is that of the tree-based models (Breiman, Friedman, Olshen, and Stone 1984), which has discontinuous boundaries.

Consider a general modeling process involving application of a sequence of statistical and nonstatistical tools to a dataset in an attempt to obtain a final model. This process is sometimes also called data mining. The main goal of this article is to develop a concept of generalized degrees of freedom (GDF) that is applicable for evaluation of the final model or fits produced by such a process. The GDF is defined as the sum of the sensitivity of each fitted value to perturbations in the corresponding observed value. It is nonasymptotic in nature and thus is free of the sample-size constraint. I show that the GDF can be used as a measure of the complexity of a general modeling procedure, so the pro-

Jianming Ye is Assistant Professor, Graduate School of Business, University of Chicago, IL 60637. The author is grateful for many encouragements and useful comments from Naihua Duan, Edward I. George, Zvi Gilula, Jiayu Li, Albert Madansky, Per A. Mykland, Xiufeng Niu, Art B. Owen, Xiaochuan Pan, Xiaotong Shen, Thomas Severini, Ruey S. Tsay, Grace Wahba, and Wing H. Wong on various versions of this work. Questions, suggestions, and comments from the editor, the associate editor, and four anonymous referees also led to substantial improvement of the article. This research is supported in part by the Faculty Research Fund at the Graduate School of Business, University of Chicago.

Journal of the American Statistical Association
March 1998, Vol. 93, No. 441, Theory and Methods

cedure's goodness of fit can be measured with an extended AIC. One can also view the GDF as the cost of the modeling process, so that under suitable conditions, an unbiased estimate of the error variance can be obtained. This concept allows one to establish a unified theory under which complex modeling procedures can be analyzed in the same way as classical linear models. It allows one to understand modeling procedures in terms of their complexity and tendency to overfit, rather than merely in terms of how well they fit a specific dataset.

There are several differences between GDF and the traditional degrees of freedom. I show that the GDF of a parameter may be substantially larger than 1; there is no longer an exact correspondence between the degrees of freedom and the number of parameters. In general, GDF depends on both the modeling procedure and the underlying true model.

In Section 2 I briefly summarize the concept of degrees of freedom in linear models and motivate its extension to general modeling procedures. A computational procedure is also given. Using GDF, in Section 3 I establish a framework for evaluating and comparing modeling procedures. Some theoretical results are also given. In Section 4 I briefly outline applications of the theory to nonparametric regression procedures.

Model selection is an important tool in statistics. Most modern regression textbooks treat this as a standard topic. Often, inferences are made about the selected model based on the same data, assuminging that the selected model is given a priori. It is well known that goodness-of-fit statistics from this method are often too optimistic. For example, the error variance estimate is often underestimated, and the $R^2$(adj) is often too optimistic. In Section 5 I apply the theory to derive useful corrections that have a nearunbiasedness property.

The half-normal plot (HNP) is a useful graphical tool in the analysis of orthogonal experiments. It is used for selection of orthogonal variables while taking into account the selection effect. In Section 6 I show an interesting connection between the proposed theory and the HNP. The proposed theory formalizes and extends the HNP to very general settings, whereas HNP provides a successful special case of the proposed theory. Some results for the behavior of the GDF in model selection are also given.

Data analysis, as a process, can be viewed as the combination of model searching and estimation. Based on the proposed theory, I provide some general discussion in Section 7 on the effect of model searching by a human modeler.

Throughout, I assume that the response variable is normally distributed.

## 2. GENERALIZED DEGREES OF FREEDOM

In this section I first review the concept of degrees of freedom and its use in linear models. This review motivates the definition of GDF given in the second part of the second. Consider a linear model,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon, \qquad \varepsilon \sim \mathrm{N}(0, \sigma^2\mathbf{I}), \tag{1}$$

where $\mathbf{Y} = (y_1, \dots, y_n)'$ is the response vector, $\boldsymbol{\beta}$ is a $p \times 1$ coefficient vector, $\mathbf{X}_{n\times p}$ is the design matrix with $p$ variables, and $\mathbf{I}$ is the identity matrix. Using the least squares method, I estimate $\boldsymbol{\beta}$ and $\boldsymbol{\mu} = X\boldsymbol{\beta}$ by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

and

$$\hat{\boldsymbol{\mu}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \tag{2}$$

An important result in classical theory is that the scaled residual sum of squares (RSS) follows a chi-squared distribution with $n-p$ df; that is,

$$D = (\mathbf{Y} - \hat{\boldsymbol{\mu}})'(\mathbf{Y} - \hat{\boldsymbol{\mu}})/\sigma^2 \sim \chi^2_{n-p}.$$

From the properties of the chi-squared distribution, $E(D) = n-p$ and $V(D) = 2(n-p)$. This eventually leads to the AIC (Akaike 1973). To measure how well the model captures the underlying structure, consider the quadratic loss,

$$(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})'(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}). \tag{3}$$

Akaike (1973) showed that the quadratic loss can be estimated unbiasedly by

$$(\mathbf{Y} - \hat{\boldsymbol{\mu}})'(\mathbf{Y} - \hat{\boldsymbol{\mu}}) - n\sigma^2 + 2p\sigma^2, \tag{4}$$

which is often referred to as the AIC. The quantity $p$ is the degrees of freedom in regression, which equals the number of variables in (1). Different models can then be compared based on their AIC values. The degrees of freedom can also be used for estimation of the residual variance $\sigma^2$. Assuming that (1) is the *true* model,

$$s^2(\text{adj}) = (\mathbf{Y} - \hat{\boldsymbol{\mu}})'(\mathbf{Y} - \hat{\boldsymbol{\mu}})/(n-p) \tag{5}$$

gives an unbiased estimate of $\sigma^2$; that is, $E[s^2(\text{adj})] = \sigma^2$, where $n-p$ is the degrees of freedom of the residuals. Also,

$$R^2(\text{adj}) = 1 - \frac{(\mathbf{Y} - \hat{\boldsymbol{\mu}})'(\mathbf{Y} - \hat{\boldsymbol{\mu}})/(n-p)}{\mathbf{Y}'\mathbf{Y}/n}, \tag{6}$$

which measures the variation in $\mathbf{Y}$ explained by the model.

To motivate the definition of the GDF, it is useful to note the following coincidence in linear models. Let $H = (h_{ij})_{n\times n} = X(X'X)^{-1}X'$. The degrees of freedom in regression can be reexpressed as

$$p = \mathrm{tr}(H) = \sum_i h_{ii} = \sum_i \frac{\partial \hat{\mu}_i}{\partial y_i}, \tag{7}$$

where $y_i$ and $\hat{\mu}_i$ are the $i$th components of $\mathbf{Y}$ and $\hat{\boldsymbol{\mu}}$. Thus the degrees of freedom are the sum of the sensitivities of the fitted values $\hat{\mu}_i$ with respect to the observed response values $y_i$.

In general, consider a response vector $\mathbf{Y} = (y_1, \dots, y_n)'$,

$$\mathbf{Y} \sim \mathrm{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I}), \tag{8}$$

where $\sigma^2$ is assumed to be known and $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)'$ is an $n \times 1$ mean vector. I define a modeling procedure $\mathcal{M}$ as a mapping from $\mathbf{R}^n$ to $\mathbf{R}^n$ that produces a set of fitted values $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \dots, \hat{\mu}_n)'$ from $\mathbf{Y}$; that is,

$$\mathcal{M}: \mathbf{Y} \to \hat{\boldsymbol{\mu}}. \tag{9}$$

Note that $\boldsymbol{\mu}$ often depends on some observed covariates, and so does the modeling procedure $\mathcal{M}$. This dependence is always assumed and will be implicit. To emphasize the dependence of $\hat{\boldsymbol{\mu}}$ on $\mathbf{Y}$, I often use the notation $\hat\mu_i \equiv \hat\mu_i(\mathbf{Y})$.

*Definition 1.* The GDF for a modeling procedure $\mathcal{M}$ are given by $D(\mathcal{M}) = \sum_{i=1}^n h_i^{\mathcal{M}}(\boldsymbol{\mu})$, where

$$\begin{aligned} h_i^{\mathcal{M}}(\boldsymbol{\mu}) &= \frac{\partial E_\mu[\hat\mu_i(\mathbf{Y})]}{\partial \mu_i} = \lim_{\delta\to 0} E_\mu\left[\frac{\hat\mu_i(\mathbf{Y}+\delta\mathbf{e}_i)-\hat\mu_i(\mathbf{Y})}{\delta}\right] \\ &= \frac{1}{\sigma^2}E[\hat\mu_i(\mathbf{Y})(y_i-\mu_i)] = \frac{1}{\sigma^2}\operatorname{cov}(\hat\mu_i(\mathbf{Y}), y_i-\mu_i), \end{aligned} \tag{10}$$

where $\mathbf{e}_i$ is the $i$th column of the $n \times n$ identity matrix.

The GDF is an extension of (7) to general modeling procedures. It is defined to be the sum of the *average* sensitivities of the fitted value $\hat\mu_i(\mathbf{Y})$ to a small change in $y_i$. Thus it measures the flexibility of the modeling procedure $\mathcal{M}$. If $\mathcal{M}$ is highly flexible, then the fitted values tend to be close to the observed values. Thus the sensitivity of the fitted values to the observed values would be high, and the GDF would be large.

In classical linear models, the GDF reduce to the standard degrees of freedom in (5). If $\mathcal{M}$ is a linear smoother, then $D(\mathcal{M})$ reduces to the trace of the smoothing matrix (Hastie and Tibshirani 1990; Wahba 1983). Efron (1986) obtained the concept of "expected optimism" by using the covariance form of Definition 1. The covariance form is less intuitive and more difficult to analyze and estimate. Girard (1989) used a randomized trace definition for degrees of freedom that is related to my Algorithm 1.

It is important to note that $E_\mu[\hat\mu_i(\mathbf{Y})]$ is an infinitely differentiable function of $\hat\mu$. This has three implications; First, GDF is defined even when $\hat\mu_i$ is highly irregular or even discontinuous; Second, because $h_i^{\mathcal{M}}(\boldsymbol{\mu})$ is also infinitely differentiable, it can be estimated with its value $h_i^{\mathcal{M}}(\mathbf{Y})$. Finally, because $E_\mu[\hat\mu_i(\mathbf{Y})]$ can be viewed as a smoothing of the fitted value $\hat\mu_i$, what is important is the global behavior of $\hat\mu_i$, note the local behavior. The following example illustrates this point.

*Example 1.* Let $y \sim \mathrm{N}(\mu, \sigma^2)$. Consider an estimate for $\mu$,

$$\hat\mu = y + \frac{1}{K}\sin(K^2 y).$$

Note that as $K \to \infty$, $\hat\mu \to y$. The local behavior of the estimate, characterized by $d\hat\mu/dy = 1 + K\cos(K^2 y)$, is highly unstable and is irrelevant for evaluating the estimate. By direct calculation, the GDF is $h(\mu) \to 1$. It is a more meaningful characterization of the behavior of the estimate than $d\hat\mu/dy$.

For linear models and linear smoothers, $D(\mathcal{M})$ is independent of $\mathbf{Y}$ and $\boldsymbol{\mu}$. This is not true in general. $D(\mathcal{M})$ may depend on $\boldsymbol{\mu}$. Although this leads to the need for an estimate, the dependence on $\boldsymbol{\mu}$ is of crucial importance. For example, searching for a knot in a white noise sequence costs approximately 3 df (Owen 1991). However, if the data are not white noise and there is a clear level shift, searching for a knot bears little cost. This point will become more evident in Sections 4 and 5.

To estimate $D(\mathcal{M})$, I use a Monte Carlo method, as follows:

*Algorithm 1*

```
• Repeat $t = 1, \ldots, T$.
• Generate $\Delta_t = (\delta_{t1}, \ldots, \delta_{tn})$ from the density $\prod(1/\tau^n)\phi(\delta_{ti}/\tau)$.
• Evaluate $\hat\mu_i(Y + \Delta_t)$ based on the modeling procedure $\mathcal{M}$.
• Calculate $\hat h_i^{\mathcal{M}}$ as the regression slope from

      $\hat\mu_i(\mathbf{Y} + \Delta_t) = \alpha + \hat h_i \delta_{ti}, \qquad t = 1, \ldots, T,$
```

I estimate $D(\mathcal{M})$ by $\hat D(\mathcal{M}) = \sum_i \hat h_i$.

The foundation of Algorithm 1 is the infinite differentiability of $h_i$ as discussed following Definition 1. The estimated GDF measures the sensitivity of the fitted vector $\hat{\boldsymbol{\mu}}$ with respect to $\mathbf{Y}$ at the *observed value* of $\mathbf{Y}$. Some justifications and theory about the algorithm are given in the Appendix.

The foregoing estimation of $D(\mathcal{M})$ requires repeated applications of the modeling procedure $\mathcal{M}$ to the perturbed dataset $\mathbf{Y} + \Delta$. In the algorithm I generate perturbations from a normal distribution. This is done only for the sake of convenience. Different methods of generating the perturbations may influence the efficiency of estimating the slope, $\hat h_i^{\mathcal{M}}$, but will not lead to any substantial bias.

$\hat D(\mathcal{M})$ depends on a tuning parameter $\tau$, which I refer to as the perturbation size. If $\tau \to 0$, then it can be shown that $\hat D(\mathcal{M})$ is an unbiased estimate for $D(\mathcal{M})$. However, in handling of modeling procedures that have discontinuous fitted values, it is crucial that $\tau/\sigma$ be bounded away from 0. A reasonable choice of $\tau$ is near $.6\sigma$, regardless of the sample size. In my experience, the results are relatively insensitive to the choice of $\tau$ for $\tau \in [.5\sigma, \sigma]$. Some rationales regarding the choice of $\tau$ are also given in the Appendix.

The parameter $T$ determines the number of perturbations. In general, I use $T \geq n$. It may be possible to use a smaller number of perturbations, especially when $n$ is large.

## 3. EVALUATIONS AND COMPARISONS OF COMPLEX MODELING PROCEDURES

The concept of GDF allows complex modeling procedures to be analyzed in a way similar to the analysis of classical linear models. This concept is independent of the sample size constraint and the complexity of the modeling procedure. This section discusses applications of GDF that are parallel to ordinary degrees of freedom.

Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be two different modeling procedures. Let $\hat{\boldsymbol{\mu}}_k$, $k = 1, 2$, be the fitted values for $\mathcal{M}_k$, $k = 1, 2$. Let

$$\mathrm{RSS}(\mathcal{M}_k) = (\mathbf{Y} - \hat{\boldsymbol{\mu}}_k)'(\mathbf{Y} - \hat{\boldsymbol{\mu}}_k) \tag{11}$$

be the sum of squared residuals. Given the GDF $D(\mathcal{M}_k)$, the two modeling procedures can be compared by using the criterion

$$A_e(\mathcal{M}_k) = \mathrm{RSS}(\mathcal{M}_k) - n\sigma^2 + 2D(\mathcal{M}_k)\sigma^2. \tag{12}$$

This is based on the following result.

*Theorem 1.* Given a modeling procedure $\mathcal{M} : \mathbf{Y} \to \hat{\boldsymbol{\mu}}$, and a statistic

$$A(\mathcal{M}, P) = \text{RSS}(\mathcal{M}) - n\sigma^2 + 2P\sigma^2, \tag{13}$$

where $P$ is a constant, the quantity

$$E[(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})'(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) - A(\mathcal{M}, P)]^2$$

is minimized when $P = D(\mathcal{M})$. Consequently, $A_e(\mathcal{M})$ is an unbiased estimate for the quadratic loss $(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})'(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})$.

A proof is provided in the Appendix. From Theorem 1, $A_e(\mathcal{M})$ is the optimal estimate under the quadratic loss. This holds without any condition on the procedure $\mathcal{M}$. Because of its similarity to AIC, I will refer to (11) as the extended AIC (EAIC).

Another useful criterion is the analog to GCV (Craven and Wahba 1979). Define

$$\text{GCV}(\mathcal{M}_k) = \text{RSS}(\mathcal{M}_k)/(n - D(\mathcal{M}_k))^2. \tag{14}$$

The advantage of this criterion is that it does not assume a known $\sigma^2$.

When $\sigma^2$ is unknown, an estimate of it may be obtained by

$$s^2(\text{cor}) = (\mathbf{Y} - \hat{\boldsymbol{\mu}})'(\mathbf{Y} - \hat{\boldsymbol{\mu}})/(n - D(\mathcal{M})). \tag{15}$$

As a goodness-of-fit measure, define

$$R^2(\text{cor}) = 1 - \frac{(\mathbf{Y} - \hat{\boldsymbol{\mu}})'(\mathbf{Y} - \hat{\boldsymbol{\mu}})/(n - D(\mathcal{M})}{\mathbf{Y}'\mathbf{Y}/n,}, \tag{16}$$

where $\bar{y} = \sum_i y_i/n$ is the sample average.

*Theorem 2.* $E[s^2(\text{cor})] = \sigma^2$ if and only if

$$E(\mathbf{Y} - \hat{\boldsymbol{\mu}})'(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) = 0. \tag{17}$$

If $\mathcal{M}$ satisfies (17), then it models the underlying structure $\boldsymbol{\mu}$ "adequately." This assumption is usually not satisfied exactly. But it is a good approximation when $\mathcal{M}$ is flexible enough to fit the underlying model structure. In classical least squares theory, (17) is equivalent to the assumption that (1) is the true model. If $\mathcal{M}$ is inadequate for modeling the underlying model, then $\hat{\sigma}^2$ is usually larger than $\sigma^2$.

*Remark 1.* In Algorithm 1 the perturbation size $\tau$ depends on $\sigma$. This leads to a looping situation for estimating $\sigma^2$. An iterative procedure, starting with an underestimated $\hat{\sigma}$, can be used. With a fixed amount of information about $\mathcal{M}$, there is no better alternative.

*Remark 2.* Assumption (17) implies that

$$\begin{aligned} D(\mathcal{M}) &= \sum_i h_i(\mu) = E(\mathbf{Y} - \boldsymbol{\mu})'(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \\ &= E(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})'(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}). \end{aligned} \tag{18}$$

Suppose that $E\hat{\mu}(\mathbf{Y}) = \boldsymbol{\mu} + \mathbf{B}$, where $\mathbf{B}$ does not depend on $\boldsymbol{\mu}$. Define a modeling procedure $\mathcal{M}^*$ by

$$\mathcal{M}^*\colon \mathbf{Y} \to \hat{\mu}^*(\mathbf{Y}) = \hat{\mu}(\mathbf{Y}) - \mathbf{B}.$$

From Definition 1, $D(\mathcal{M}^*) = D(\mathcal{M})$. This implies that

$$\begin{aligned} &E(\hat{\boldsymbol{\mu}}^* - \boldsymbol{\mu})'(\hat{\boldsymbol{\mu}}^* - \boldsymbol{\mu}) \\ &\quad = E(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})'(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) = E(\hat{\boldsymbol{\mu}}^* - \boldsymbol{\mu})'(\hat{\boldsymbol{\mu}}^* - \boldsymbol{\mu}) + \mathbf{B}'\mathbf{B}. \end{aligned}$$

Thus (17) implies that $\mathbf{B} = \mathbf{0}$. In this sense, the assumption prevents $\mathcal{M}$ from having a systematic bias in estimating $\boldsymbol{\mu}$. Therefore, it is necessary that $\mathcal{M}$ be approximately unbiased; that is,

$$E\hat{\boldsymbol{\mu}} \approx \boldsymbol{\mu}.$$

A phenomenon for general modeling procedures is that $E\hat{\sigma}^2$ may be smaller than $\sigma^2$. This is different from the classical linear model, where it is alway true that $E\hat{\sigma}^2 \geq \sigma^2$. The situation is most extremely in the following example.

*Example 2.* Let $\boldsymbol{\mu} = \mathbf{0}$, so that $\mathbf{Y} = \varepsilon$. Let $\hat{\boldsymbol{\mu}} = \lambda\mathbf{Y}$ for some constant $0 \leq \lambda \leq 1$. It is straightforward that

$$E(\mathbf{Y} - \hat{\boldsymbol{\mu}})'(\mathbf{Y} - \hat{\boldsymbol{\mu}}) = n(1 - \lambda)^2\sigma^2.$$

The GDF are $\sum_i h_i = n\lambda$ and

$$E\hat{\sigma}^2 = (1 - \lambda)\sigma^2.$$

Therefore, $\hat{\sigma}^2/\sigma^2 \to 0$ as $\lambda \to 1$.

A more realistic example is the linear smoother (Hastie and Tibshirani 1990).

*Example 3.* Let $\boldsymbol{\mu} = \mathbf{0}$ and $\hat{\boldsymbol{\mu}} = M\mathbf{Y}$, where $M$ is a symmetric, nonnegative definite square matrix. Assume that $M$ has an eigendecomposition $M = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}'$, where $\mathbf{U}$ is a unitary matrix and $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots \lambda_n)$ is a diagonal matrix with $\lambda_i \geq 0$, $i = 1, \ldots, n$. By some algebra,

$$E\hat{\sigma}^2 = \frac{\sum_i(1 - \lambda_i)^2}{\sum_i(1 - \lambda_i)}\sigma^2.$$

If the $\lambda_i$'s are close to either 0 or 1, with at least some of them close to 0, then $E\hat{\sigma}^2 \approx \sigma^2$. As a nontrivial example, let $\lambda_i = (n - i)/n$, in which case $E\hat{\sigma}^2 \approx 2\sigma^2/3$.

These two examples suggest that if there is a shrinkage effect in the modeling procedure, the estimated variance when GDF are used may underestimate the true variance.

## 4. APPLICATIONS TO NONPARAMETRIC REGRESSION

The framework proposed in Section 3 has many potential applications. For example, EAIC or GCV can be used to compare the performance of a nonparametric regression to a linear model as a way of diagnosing the adequacy of the linear model (Eubank and Spiegelman 1990; Staniswalis and Severini 1991), or of selecting the most suitable nonparametric regression procedures among various alternatives, such as classification and regression trees (CART), projection pursuit regression (PPR), and artificial neural network (ANN). The criteria can also be used for selection of variables in nonparametric regression settings (Bickel and Zhang 1992; Zhang 1991) by treating variable selection as a special case of selecting modeling procedures. Other applications include the evaluation of the effect induced by various selection procedures, such as variable selection in linear models (see Sec. 5), and bandwidth selection in nonparametric regression.

In this section I illustrate the behavior of GDF in the context of CART. The computation of the GDF is based on Algorithm 1. CART and multivariate adaptive regression splines (MARS, Friedman 1991) are powerful nonparametric regression tools using recursive partitioning of the space of independent variables. I study CART only because it is more difficult to analyze due to its discontinuity. (Detailed discussions on CART can be found in Breiman et al. 1984, Chipman, George, and McCullogh 1996), and Clark and Pregibon 1992).

We use the implementation in the S-PLUS software, *tree* ( ), as described by Clark and Pregibon (1992) and Venables and Ripley (1994). In S-PLUS, the function *summary* (*tree.object*) provides two statistics: the size (number of terminal nodes) and the deviance (RSS) of the selected tree. Assuming that the degrees of freedom equal the number of terminal nodes, $N_t$, S-PLUS gives the following estimate of $\sigma^2$:

$$\hat{\sigma}_o^2 = \frac{(\mathbf{Y}-\hat{\boldsymbol{\mu}})'(\mathbf{Y}-\hat{\boldsymbol{\mu}})}{n-N_t}. \tag{19}$$

I show that the GDF of a tree are often several times $N_t$ and are usually not a deterministic function of $N_t$. The estimate (19) tends to severely underestimate $\sigma^2$. Let $g$ be the GDF of the fitting procedure. It can be shown that if a CART overfits the data, then the variance estimate when GDF is used,

$$\hat{\sigma}^2 = \frac{(\mathbf{Y}-\hat{\boldsymbol{\mu}})'(\mathbf{Y}-\hat{\boldsymbol{\mu}})}{n-g}, \tag{20}$$

gives an unbiased estimate of the error variance.

In the simulations all $x$ variables are generated independently from uniform (0, 1). I consider the following models:

$$N_2\colon\ y = 0\cdot x_1 + 0\cdot x_2 + \varepsilon,$$

$$N_{10}\colon\ y = 0\cdot(x_1+\cdots+x_{10}) + \varepsilon,$$

$$T_{10}\colon\ y = s(x_1,x_2) + 0\cdot(x_3+\cdots+x_{10}) + \varepsilon.$$

The function $s(x_1,x_2)$ is given in Figure 1, and $\varepsilon \sim \mathrm{N}(0,\sigma^2)$, $\sigma^2 = .25$. Model $N_2$ has two variables, whereas models $N_{10}$ and $T_{10}$ have 10 variables. The number of observations is $n = 100$ for all cases.

Table 1 gives the size, GDF, and error variance estimate for trees. The sizes are fixed at 19 nodes. The estimated variance is given in terms of the simulated ratio $\hat{\sigma}_o^2/s^2$ and $\hat{\sigma}^2/s^2$, where $s^2 = 1/N\sum_i \varepsilon_i^2$. I use ratios because $s^2$ is the variance estimate when the true errors are known. The standard deviations of the ratios are also given.

The estimate $\hat{\sigma}_o^2$ biases downward substantially and equals only about $.25\sigma^2$. The new variance estimate, $\hat{\sigma}^2$, is almost unbiased for all three cases in Table 1. The GDF are larger for model $N_{10}$ than for model $N_2$ due to more independent variables, even though the sizes are the same. With 100 observations and 10 variables, a tree of 19 nodes uses up to 79 df.

Figure 2 shows the behavior of GDF and $\hat{\sigma}^2$ for fitted trees of different sizes. The estimate $\hat{\sigma}^2$ stabilizes near $\sigma^2 = .25$ when the tree is large enough to fit the true models. Simulation results show that $\hat{\sigma}^2$ given by (20) behaves quite well for individual datasets.

The GDF depend on the underlying true model. Table 2 displays GDF for models $N_{10}$ and $T_{10}$. Fitting a tree to a pure noise dataset costs substantially more GDF than fitting

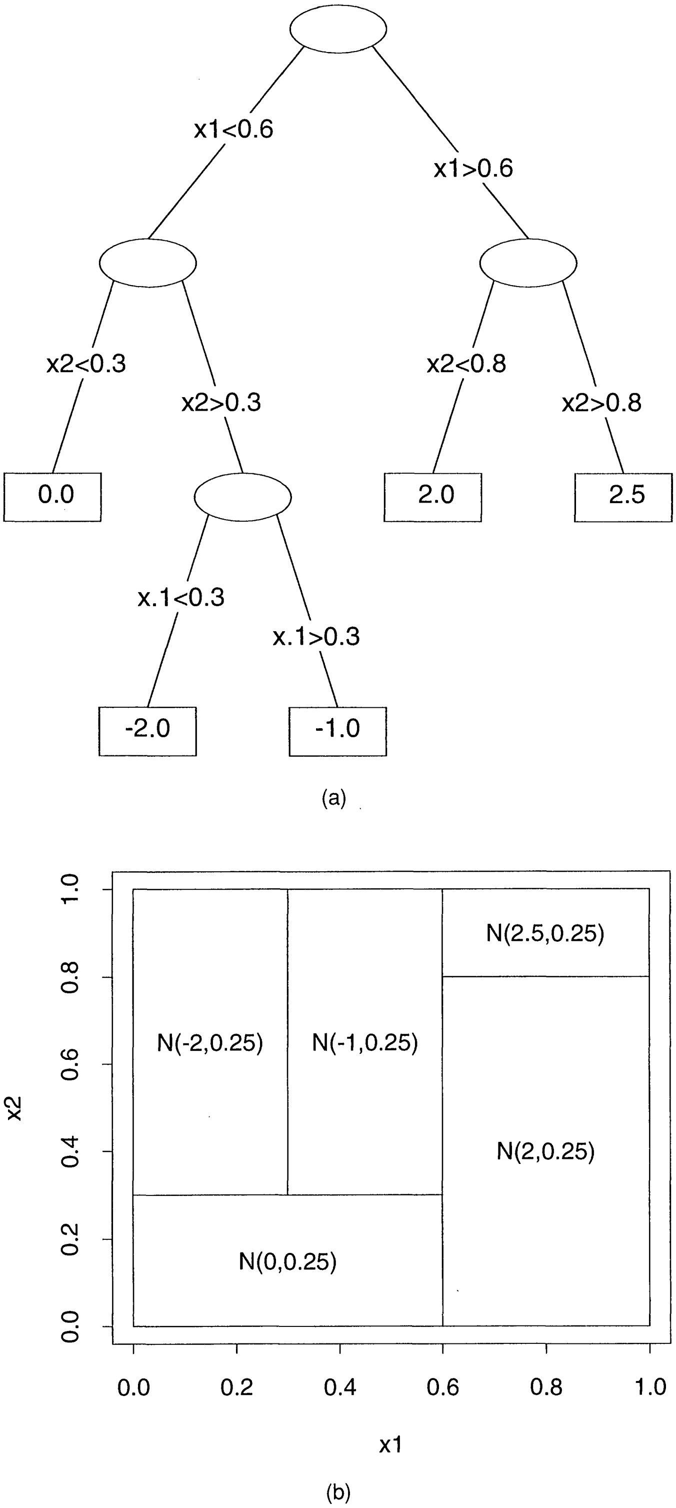

*Figure 1. Two Perspectives of a Sample Tree With Five Terminal Nodes. (a) A classification structure; (b) A step function perspective.*

*Table 1. Variance Estimate in CART With Fixed Number of Terminal Nodes*

| *Model* | *Nodes* | *GDF* | $\hat{\sigma}_0^2/s^2$ | $\hat{\sigma}^2/s^2$ |
|---|---|---|---|---|
| $N_2$ | 19 | 65.2(.78) | .43(.05) | 1.01(.15) |
| $N_{10}$ | 19 | 79.1(.44) | .25(.04) | .97(.19) |
| $T_{10}$ | 19 | 75.9(1.14) | .31(.04) | 1.04(.20) |

NOTE: $n$ = 100, $\sigma$ = .5, $\tau$ = .5$\sigma$; results are based on 100 simulations.

to a dataset with a clear structure for the first several nodes. Searching for one node in a 10-dimensional space costs as many as 15 GDF when the data are pure noise.

If the fitted tree is large enough, the GDF is stable across different simulations (its variance is negligible), as shown in Figure 2a and c. This observation is different from that of Gu (1996), who found that the degrees of freedom are unstable across the simulations in spline smoothing. If the tree is too small to fit the underlying structure, then the variance of GDF can be considerably larger.

The GDF can be used in EAIC and GCV for selecting the size for an estimated tree. In the context of a Bayesian CART (Chipman et al. 1996), EAIC and GCV can be used for selecting the prior distribution, leading to an empirical Bayes estimation procedure.

## 5. CORRECTING BIAS FROM MODEL SELECTION

Model selection is an area that has undergone extensive research in the last three decades. To evaluate a selected model, one often assumes that the selected model is given a priori. It is well known that such a method leads to overly optimistic goodness-of-fit statistics. In this section I apply the concept of GDF to derive a systematic way of handling this problem.

Consider a linear model that has $q$ variables:

$$\mathbf{Y} = \sum_{j=1}^{q} \beta_j \mathbf{x}_j + \varepsilon, \qquad \varepsilon \sim \mathrm{N}(0, \mathbf{I}). \tag{21}$$

Let $M = \{\mathbf{x}_{j_1}, \ldots, \mathbf{x}_{j_k}\}$ be a subset of $\{\mathbf{x}_1, \ldots, \mathbf{x}_q\}$. Let the corresponding fitted value be

$$\hat{\mathbf{Y}}_M = \sum_{h=1}^{k} \beta_{j_h} \mathbf{x}_{j_h} \tag{22}$$

For a fixed $0 \le k \le q$, select $M_k$ as

$$M_k = \mathrm{argmin}_M (\hat{\mathbf{Y}}_M - \mathbf{Y})'(\hat{\mathbf{Y}}_M - \mathbf{Y}).$$

To evaluate the goodness of fit of $M_k$, it is often assumed to be given a priori. This method may lead to substantial bias. For example, AIC, as defined by (3), underestimates the quadratic loss, and $s^2$(adj), as given in (4), underestimates $\sigma^2$.

To correct for this bias, the selection process must be taken into account. Define a modeling procedure $\mathcal{M}_k$ as a combination of model selection and fitting as a mapping from $\mathbf{Y} = (y_1, \ldots, y_n)$ to $\hat{\mathbf{Y}}_{M_k} = (\hat{y}_1, \ldots, \hat{y}_n)$:

$$\mathcal{M}_k\colon \mathbf{Y} \xrightarrow{\text{selection}} M_k(\mathbf{Y}) \xrightarrow{\text{fitting } M_k(\mathbf{Y})} \hat{\mathbf{Y}}_{M_k}. \tag{23}$$

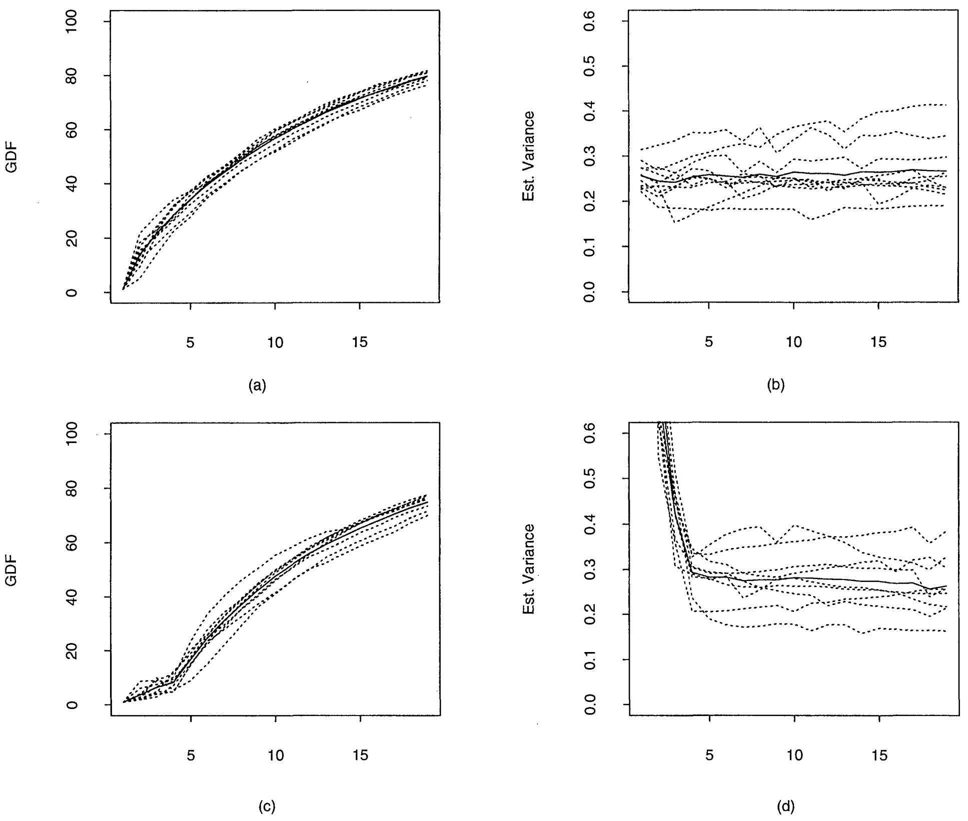

*Figure 2. The GDF and $\hat{\sigma}^2$ for the Estimated Trees of Different Sizes. (a) and (b) Ten simulations based on model $N_{10}$; (c) and (d) ten simulations based on $T_{10}$. Results are obtained wtih $n$ = 100, $\sigma$ = 0.5, and $\tau$ = .5$\sigma$.*

Table 2. GDF for Two Different Models

| Model | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $N_{10}$ | 14.8 | 22.6 | 29.2 | 34.7 | 39.8 |
| $T_{10}$ | 2.7 | 5.6 | 8.18 | 17.6 | 25.4 |

NOTE: The settings are the same as in Table 1.

Let $D(k) = D(\mathcal{M}_k)$. Then (12), (14), (15), and (16), can be applied to $\mathcal{M}_k$ to obtain an unbiased assessment of the model $M_k$.

The following example illustrates the bias induced by selection and the effectiveness of the correction derived from GDF.

*Example 4.* Data are generated from a linear model

$$y_i = \alpha + \sum_{j=1}^{q} \beta_i x_{ij} + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where $q = 20$, $n = q + 2 = 22$, and $\varepsilon_i$'s are iid as N(0, 1); $\alpha = 0$ is set in all cases. The coefficient vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{20})$ is set to take one of the following two values:

$$\boldsymbol{\beta}^{(0)} = 0_{20} : \beta_1 = \cdots = \beta_{20} = 0;$$

or

$$\boldsymbol{\beta}^{(1)} = 2_5 0_{15} : \beta_1 = \cdots = \beta_5 = 2, \quad \beta_6 = \cdots = \beta_{20} = 0.$$

Let $\mathbf{x}_{i\cdot} = (x_{i1}, \ldots, x_{ip})$ be the $i$th row of the matrix $\mathbf{X} = (x_{ij})_{22\times 20}$. $\mathbf{X}$ is generated from

$$\mathbf{x}_{i\cdot} \overset{\text{iid}}{\sim} \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}), \qquad i = 1, \ldots, n,$$

where $\boldsymbol{\Sigma} = (1 - \rho)\mathbf{I} + \rho\mathbf{J}$, $\mathbf{I}$ is an identity matrix and $\mathbf{J}$ is a matrix of 1s.

All cases use the perturbation size $\tau = .5\sigma$, with the number of perturbations $T = 100$. The results are not sensitive to $\tau$ for $\tau \in [.5\sigma, \sigma]$. The intercept is always included in selected models.

Table 3 illustrates the bias from ignoring the selection process and performance of the GDF. The rows "loss," "AIC," "$s^2$(adj)," and "$R^2$(adj)" are defined by (3)–(6), using nominal degrees of freedom. Therefore, these quantities are evaluated as if the selected model were given a prior, "EAIC," "$s^2$(cor)," and "$R^2$(cor)" are calculated based on (12), (14), and (15) by using GDF of the combined process (23).

In general, the GDF are substantially larger than the nominal degrees of freedom. In case 1 the selection process of the best five variables costs $14.1 - 6 = 8.1$ df, which is 135% of the nominal degrees of freedom. EAIC estimates the true loss much better than AIC, because it is approximately unbiased.

Table 3 shows that $s^2$(adj) becomes substantially smaller than $\sigma^2 = 1$ as $k$ increases. $R^2$(adj) increases as $k$ increases. In case 1 the average $R^2$(adj) reaches 77% even though $\boldsymbol{\beta} = \mathbf{0}$. The estimates using GDF behave substantially better. The $s^2$(cor) is nearly unbiased for $\sigma^2$ when $k \geq 0$ in case 1 and $k \geq 5$ in case 2. In case 1 $R^2$(cor) is near 0 for all $k$. For case 2 $R^2$(cor) stabilizes when $k \geq 5$, because the true model has five variables with nonzero coefficients.

In case 1 AIC selects three variables on average. The overestimation of the model size by AIC here is mainly due to the bias induced by ignoring that $M_k$ is the best selected model for each $k$. This is quite different from the overestimation problem of Zhang (1992), where AIC is unbiased, because he assumes that the models under selection forms a nested sequence, as in an autoregression, so that there is only one model for each given dimensionality. The reason for overestimation by Zhang (1992) derives from the inconsistency of AIC, not from the selection bias.

Table 4 illustrates the performance of EAIC, AIC, $\mathrm{GCV}_{\mathrm{gdf}}$, and $\mathrm{GCF}_{\mathrm{df}}$ as variable selection procedures. The table gives the true loss of the models selected by each criterion, the average number of variables in the selected model, and the variance estimates when degrees of freedom and GDF are used. For the null model $\boldsymbol{\beta} = 0_{20}$, EAIC selects about .4 variable on average, whereas the model selected by AIC has three variables. In all cases simulated, the dimensionality selected by minimizing EAIC has a smaller average true loss than that selected by minimizing AIC. The improvement can often be substantial, especially in cases where the true model contains relatively few strong variables.

$\mathrm{GCV}_{\mathrm{gdf}}$ and $\mathrm{GCV}_{\mathrm{df}}$ are the GCV values using GDF and nominal degrees of freedom. $\mathrm{GCV}_{\mathrm{gdf}}$ performs well if the true model has only a few strong variables, but may behave erratically if the true model is close to the full model. The reason is that GCV relies implicitly on an estimated $\sigma^2$; see (39) later. When the true model contains only a few variables, the error variance can be estimated well and $\mathrm{GCV}_{\mathrm{gdf}}$ performs well. If there are almost no degree of freedom in the residuals for error variance estimation, then

Table 3. Goodness-of-Fit Statistics of a Selected Model: Bias and the Corresponding Corrections

| | k (Dimensionality of the model) | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 5 | 6 | 10 | 15 | 20 |
| | | | Case 1 ($\beta$ = $0_{20}$) | | | |
| GDF | 5.23 | 14.10 | 15.37 | 18.75 | 20.56 | 21.00 |
| AIC | −1.28 | −2.21 | −1.48 | 3.24 | 11.37 | 20.96 |
| Loss | 5.24 | 14.17 | 15.44 | 18.72 | 20.59 | 21.00 |
| EAIC | 5.18 | 13.99 | 15.26 | 18.74 | 20.48 | 20.95 |
| $s^2$(adj) | .84 | .49 | .44 | .29 | .23 | .96 |
| $s^2$(cor) | 1.00 | 1.02 | 1.03 | 1.10 | 1.00 | .91 |
| $R^2$(adj) | .16 | .51 | .56 | .71 | .77 | .05 |
| $R^2$(cor) | 0 | −.02 | −.03 | −.09 | .01 | .09 |
| | | | Case 2 ($\beta$ = $2_5 0_{15}$) | | | |
| GDF | 10.98 | 6.32 | 10.25 | 17.43 | 20.51 | 21.00 |
| AIC | 347 | 6.22 | 4.20 | 4.79 | 11.59 | 20.99 |
| Loss | 362 | 6.27 | 10.33 | 17.76 | 20.84 | 21.33 |
| EAIC | 365 | 6.86 | 10.70 | 17.66 | 20.61 | 21.00 |
| $s^2$(adj) | 18.23 | 1.01 | .81 | .44 | .26 | .99 |
| $s^2$(cor) | 88.96 | 1.05 | 1.06 | 1.18 | 1.14 | .94 |
| $R^2$(adj) | .69 | .98 | .99 | .99 | 1.00 | .98 |
| $R^2$(cor) | −.53 | .98 | .98 | .98 | .98 | .98 |

NOTE: $n$ = 22, $q$ = 20, $\rho$ = .5, $\sigma$ = 1, and $\tau$ = .5. Results are based on 1,000 simulations with the same design.

*Table 4. Dimension Selection: Comparisons of Criteria*

| | $\rho = .5$ | | $\rho = 0$ | |
|---|---|---|---|---|
| $\beta$ | $0_{20}$ | $2_5 0_{15}$ | $0_{20}$ | $2_5 0_{15}$ |
| | *True loss* | | | |
| EAIC | 3.39 | 8.74 | 3.17 | 8.26 |
| AIC | 11.67 | 14.75 | 11.86 | 14.50 |
| $\mathrm{GCV}_{\mathrm{gdf}}$ | 6.82 | 11.59 | 6.72 | 11.24 |
| $\mathrm{GCV}^*_{\mathrm{gdf}}$ | 4.17 | 9.26 | 4.05 | 8.74 |
| $\mathrm{GCV}_{\mathrm{df}}$ | 18.92 | 19.65 | 18.95 | 19.53 |
| $\mathrm{GCV}^*_{\mathrm{df}}$ | 18.49 | 19.06 | 18.53 | 18.92 |
| | *Average number of variables selected* | | | |
| EAIC | .40 | 5.50 | .36 | 5.40 |
| AIC | 2.98 | 7.39 | 3.09 | 7.36 |
| $\mathrm{GCV}_{\mathrm{gdf}}$ | 3.82 | 8.40 | 3.79 | 8.38 |
| $\mathrm{GCV}^*_{\mathrm{gdf}}$ | .84 | 5.76 | .84 | 5.70 |
| $\mathrm{GCV}_{\mathrm{df}}$ | 11.53 | 13.34 | 11.49 | 13.27 |
| $\mathrm{GCV}^*_{\mathrm{df}}$ | 10.09 | 11.57 | 10.09 | 11.60 |
| | *$s^2$(adj)* | | | |
| $\mathrm{GCV}_{\mathrm{df}}$ | .22 | .25 | .22 | .26 |
| $\mathrm{GCV}^*_{\mathrm{df}}$ | .26 | .31 | .25 | .31 |
| | *$s^2$(cor)* | | | |
| $\mathrm{GCV}_{\mathrm{gdf}}$ | .75 | .72 | .75 | .72 |
| $\mathrm{GCV}^*_{gdf}$ | .88 | .88 | .88 | .88 |
| $\mathrm{GCV}_{\mathrm{df}}$ | .54 | .59 | .54 | .57 |
| $\mathrm{GCV}^*_{\mathrm{df}}$ | .65 | .78 | .65 | .74 |

NOTE: $n$ = 22, $p$ = 20, $\sigma$ = 1, and $\tau$ = .5. Results are from 1,000 simulations with the same design.

the dimensionality selection procedure tends to behave erratically. The criteria $\mathrm{GCV}^*_{\mathrm{gdf}}$ and $\mathrm{GCV}^*_{\mathrm{df}}$ are the same as $\mathrm{GCV}_{\mathrm{gdf}}$ and $\mathrm{GCV}_{\mathrm{df}}$, except that the maximum order is limited to 15 variables. This avoids using GCV when the degrees of freedom are almost exhausted. The performance of $\mathrm{GCV}^*_{\mathrm{gdf}}$ is better than that of $\mathrm{GCV}_{\mathrm{gdf}}$ when the true model contains relatively few variables.

The variance estimate based on the model selected by $\mathrm{GCV}_{\mathrm{df}}$ substantially underestimates the true variance, even with $s^2$(cor). Using $s^2$(cor) with the model selected by $\mathrm{GCV}^*_{\mathrm{gdf}}$ gives a substantially less-biased estimate.

## 6. GENERALIZED DEGREES OF FREEDOM AND THE HALF-NORMAL PLOT

In this section I provide some theoretical results on the GDF-based methods for the selection bias problem as given in Section 5. I show that these methods are deeply connected to the HNP, a frequently used graphical tool in the analysis of orthogonal experiments. This connection provides two important points: First, it establishes GDF and the associated framework as a formalization and extension of the HNP to very general settings. Second, on the other hand, the widespread use of HNP in orthogonal experiments provides a successful case for GDF.

### 6.1 The Behavior of GDF

In a classical linear model with no variable selection, the number of degrees of freedom in regression equals the number of variables in the model. It is independent of the true model; that is, independent of the value of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)$. It is always nonnegative and is no more than the total number of observations. It is also independent of the design matrices $\mathbf{X}$ and $\mathbf{X}$. The GDF behave quite differently from the classical degrees of freedom. In general, none of the aforementioned properties is guaranteed. In this section we obtain some results on the behavior of $D(k)$.

*Theorem 3.* Assume that $\boldsymbol{\beta} = \mathbf{0}$. Then

a. for every given $k < n$, $M_k$ satisfies the assumption (17). Thus

$$E[s^2(\mathrm{cor})] = \sigma^2.$$

Equivalently,

$$D(k) = n - E(\hat{\mathbf{Y}}_{M_k} - \mathbf{Y})'(\hat{\mathbf{Y}}_{M_k} - \mathbf{Y})/\sigma^2. \quad (24)$$

b. $D(k)$ is a nondecreasing function of $k$. Moreover,

$$k \le D(k) \le \min(n, q). \quad (25)$$

*Corollary 1.* Assume that $\boldsymbol{\beta} = \mathbf{0}$; then

$$s_k^2(\mathrm{adj}) = (\hat{\mathbf{Y}}_{M_k} - \mathbf{Y})'(\hat{\mathbf{Y}}_{M_k} - \mathbf{Y})/(n - k) \quad (26)$$

underestimates $\sigma^2$. Moreover,

$$1 - \frac{q}{n} \le \frac{E[s_k^2(\mathrm{adj})]}{\sigma^2} \le 1. \quad (27)$$

*Corollary 2.* If $\beta_1, \ldots, \beta_l \to \infty$, $\beta_{l+1} = \cdots = \beta_q = 0$, and $n > l$, then

a. $D(l) \to l$ and
b. For all $k \ge l$, $s_k^2(\mathrm{cor}) \to \sigma^2$.

Theorem 3 gives the limits of the GDF in the case in which the selection effect is "pure" in the sense that $\boldsymbol{\beta} = \mathbf{0}$. Under the assumptions of Theorem 3, the GDF are nonnegative and are no more than the total number of variables. Corollary 1 gives the lower bound of underestimation of $\hat{\sigma}_a^2$. It is clear that the maximum amount of underestimation goes to 0 at the rate of $n^{-1}$ as $n \to \infty$. Thus the correction is of interest only when is comparable to or smaller than $q$. Corollary 2 shows that the GDF automatically reduce to degrees of freedom if there is little uncertainty in the selection. The following example gives a scenario in which $D(k)$ is unbounded.

*Example 5.* Consider the model (21) with $q = 2$ and $n = 2$. Assume that $\mathbf{x}_1$ and $\mathbf{x}_2$ are orthogonal, with $\mathbf{x}_i'\mathbf{x}_i = 1$, $i = 1, 2$. Assume further that $\beta_1 = \beta_2 = \beta$. This is when the model uncertainty is the greatest. Thus $\hat{\beta}_i \sim \mathrm{N}(\beta, 1)$, $i = 1, 2$ are independent random variables. Rearrange $\hat{\beta}_i$, $i = 1, 2$, into $\hat{\beta}_{j_1}$ and $\hat{\beta}_{j_2}$ so that $|\hat{\beta}_{j_1}| > |\hat{\beta}_{j_2}|$. From the Appendix, we have

$$\begin{aligned} D(1) &= 2 - E[\hat{\beta}_{j_2}(\hat{\beta}_{j_2} - \beta)] \\ &= 2 - E(\hat{\beta}_{j_2} - \beta)^2 + \beta(\beta - E\hat{\beta}_{j_2}). \end{aligned} \quad (28)$$

Let $\beta \to +\infty$. One can show that $E(\hat{\beta}_{j_2} - \beta)^2 \to c_1 > 0$ and $(\beta - E\hat{\beta}_{j_2}) \to c_2 > 0$, where $c_1 = 1$ and $c_2 = .564$ are constants. Thus $D(1) \approx 1 + .564\beta$, which goes to $\infty$ as $\beta \to \infty$.

In this example, if $\beta$ is large, then both variables have a significant effect on $\mathbf{Y}$. By setting $k=1$, we force the selected model to leave out one of the important variables. Thus the GDF may be high when the selected model is unable to include all of the important variables. In this case, the GDF serve as a useful indicator for whether the selected model is adequate. In this example, $D(1)$ is not useful for estimating $\sigma^2$, because $D(1) > n$ as $\beta \to \infty$, and assumption (3.7) is violated. No variance estimator can be unbiased unless all the important variables are included in the selected model. This is why assumption (17) is important and is not used only for mathematical convenience.

The fact that $\hat{D}(k)$ may be greater than $n$ creates some problems, because the variance estimator and the estimated $R^2$ may be negative. Thus to rectify the problem, we set

$$D^*(k) = \begin{cases} \hat{D}(k) & \text{if } D(k) < \min(q,n) \\ \min(q,n) & \text{otherwise} \end{cases} \qquad (29)$$

in (15) and (16).

*Theorem 4.* Assume that $\boldsymbol{\beta} = \mathbf{0}$, $n \geq q$.

a. If $\mathbf{X}$ is orthogonal, then

$$D(k) = \sum_{i=1}^{k} \chi^2_{(i)} \qquad (30)$$

where $\chi^2_{(1)} > \cdots > \chi^2_{(q)}$ is the expected value of the order statistics of $q$ observations from a $\chi^2_1$ random variable.

b. In general, without assuming orthogonality, we have

$$D(1) \leq \chi^2_{(1)}. \qquad (31)$$

The equality is achieved when $\mathbf{X}$ is an orthogonal matrix.

*Conjecture.* Assume that $\boldsymbol{\beta} = \mathbf{0}$, $n \geq q$. Then

$$D(k) \leq \sum_{i=1}^{k} \chi^2_{(i)}. \qquad (32)$$

Theorem 4(a) gives an explicit expression for the GDF in an important special case. Theorem 4(b) suggest that for $k=1$, the searching cost (GDF) is highest when the variables are orthogonal. For $k > 1$, the conclusion is less clear. The right side of (32) may also be used as a convenient bound for GDF.

### 6.2 Extended AIC, Generalized Cross-Validation, and the Half-Normal Plot

In the analysis of (fractional) factorial design of experiments (DOE), the number of possible effects (variables) is typically equal to or greater than the number of observations $n$. Assume that the response $\mathbf{Y}$ follows a linear model,

$$\mathbf{Y} = \sum_{i=1}^{n} \beta_i \mathbf{x}_i + \varepsilon, \qquad \varepsilon \sim \mathrm{N}(0, \sigma^2 \mathbf{I}), \qquad (33)$$

where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ is an $n \times n$ orthonormal matrix so that $\mathbf{X}'\mathbf{X} = \mathbf{I}$. Even though (33) is a standard linear model, it is rare in the DOE literature to use AIC to identify the existence of a nonnull effect, that is, to select the variables from $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Instead, an intuitive graphical tool, the HNP, is often used (Daniel 1959; Zahn 1975a,b). In this section we show that there is a substantial difference between AIC and HNP. The difference lies in the selection bias that AIC ignores. EAIC, which takes into account the selection effect, is shown to be equivalent to HNP.

Because $\mathbf{X}$ is orthonormal, we have $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_n) = \mathbf{X}'\mathbf{Y} \sim \mathrm{N}(\boldsymbol{\beta}, \sigma^2 I)$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)$. Suppose that $(j_1, \ldots, j_n)$ is a permutation of $(1, \ldots, n)$ so that $|\hat{\beta}_{j_1}| \geq \cdots \geq |\hat{\beta}_{j_n}|$. We also denote $(\hat{\beta}_{j_1}, \ldots, \hat{\beta}_{j_n})$ by $(\hat{\beta}_{(1)}, \ldots, \hat{\beta}_{(n)})$. Let $M_k = \{j_1, \ldots, j_k\}$ be the model with $k$ largest coefficients. Note that $\mathrm{RSS}(M_k) = \sum_{i=k+1}^{n} \hat{\beta}^2_{(i)}$.

*Proposition 1.* The model selected by minimizing (4) contains all the variables $\mathbf{x}_i$ that satisfy

$$|\hat{\beta}_i| > \sqrt{2}\sigma. \qquad (34)$$

On average, AIC selects 15.7% of the candidate variables when $\boldsymbol{\beta} = \mathbf{0}$.

If $\boldsymbol{\beta} = \mathbf{0}$, then $\hat{\boldsymbol{\beta}} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I})$. Thus $|\hat{\beta}_1|/\sigma, \ldots, |\hat{\beta}_n|/\sigma$ are iid with a standard half-normal distribution. Let $Z_{(1)} > \cdots > Z_{(n)}$ be the expected values of the $n$-order statistics from a standard half-normal distribution. Then $\boldsymbol{\beta} = \mathbf{0}$ implies that $E|\hat{\beta}_{(i)}| = \sigma Z_{(i)}$. An HNP plots $|\hat{\beta}_{(i)}|$, $i = 1, \ldots, n$ against $Z_{(i)}$, $i = 1, \ldots, n$. If $\boldsymbol{\beta} = \mathbf{0}$, then the points should follow approximately a straight line with slope $\sigma$. A variable $\mathbf{x}_i$ is selected only if the point $(Z_{(i)}, |\hat{\beta}_{(i)}|)$ is well above the line; that is,

$$|\hat{\beta}_{(i)}| \gg \sigma Z_{(i)}. \qquad (35)$$

For $i = 1$, because $Z_{(1)} \to \sqrt{2\log(q)}$ as the number of variables $q \to \infty$, (35) is essentially the risk inflation criterion of George and Foster (1994) and the hard thresholding of Donoho and Johnstone (1994). Note that (35) differs significantly from (34).

From Theorem 4, if $\boldsymbol{\beta} = \mathbf{0}$, then the GDF for $\mathbf{M}_k$ is $D(k) = \sum_i \chi^2_{(i)}$. Thus the EAIC value for $M_k$ is

$$A_e(k) = \sum_{i=k+1}^{n} \hat{\beta}^2_{(i)} + 2\sigma^2 \sum_{i=1}^{k} \chi^2_{(i)}. \qquad (36)$$

This implies that $A_e(k) > A_e(k+1)$ iff $\hat{\beta}_{(k)} > 2\chi^2_{(k)}\sigma$, which defines the criterion for selecting the optimal $k$. Thus the variable corresponding to the coefficient $\hat{\beta}_{(k)}$ would be included in the selected model only if

$$|\hat{\beta}_{(k)}| > \sqrt{2\chi^2_{(k)}}\sigma \approx \sqrt{2}Z_{(k)}\sigma. \qquad (37)$$

By taking into account the selection bias, the criterion (37) is similar to the HNP criterion (35).

A similar result can be shown for GCV when we use GDF. The GCV criterion selects $M_k$ over $M_{k-1}$ if

$$\frac{\sum_{i=k}^{n} \hat{\beta}^2_{(i)}}{(n - \sum_{i=k}^{n} \chi^2_{(i)})^2} > \frac{\sum_{i=k+1}^{n} \hat{\beta}^2_{(i)}}{(n - \sum_{i=k+1}^{n} \chi^2_{(i)})^2}. \qquad (38)$$

By some algebra, (38) implies

$$|\hat\beta_{(k)}| > \hat\sigma_{(k)}\chi_{(k)}\sqrt{2+\frac{\chi^2_{(k)}}{\sum_{i=k+1}^n \chi^2_{(i)}}}, \tag{39}$$

where $\hat\sigma_{(k)}$ is the estimated standard deviation using (17). This expression is similar to (35) and (37). The difference is that an estimated standard deviation $\hat\sigma_{(k)}$, based on $M_k$, is used. GCV tends to select fewer variables than EAIC when $\hat\sigma_{(k)} \approx \sigma$, because $\chi^2_{(k)}/\sum_{i=k+1}^n \chi^2_{(i)} > 0$.

### 6.3 Variance Estimates and Half-Normal Plot

In the literature on linear models, estimating the error variance $\sigma^2$ requires that the number of observations $n$ be greater than the number of variables. But in the context of orthogonal experiments, such a requirement is not necessary. Several articles, including those of Zahn (1975a,b), have studied the estimation of $\sigma^2$ when the number of observations is equal to the number of variables.

The reason for the existence of such estimates is that, even though there may be a large number of candidate variables, we assume that only a few of them have an effect on the response. In this section we establish the approximate equivalence of (14) and the variance estimate by using HNP in an orthogonal design. Thus (14) extends such results to the general cases under a similar assumption.

For a given $k$, if $\boldsymbol\beta = \mathbf{0}$, then the points $(Z_{(i)}, |\hat\beta_{(i)}|)$, $i = 1,\ldots,n$ follow approximately a straight line with slope $\sigma$. This gives rise to an estimate for $\sigma$,

$$\tilde\sigma_k = \frac{\sum_{i=k+1}^n |\hat\beta_{(i)}|Z_{(i)}}{\sum_{i=k+1}^n Z^2_{(k)}}. \tag{40}$$

Because $E|\hat\beta_{(i)}| = \sigma Z_{(i)} \approx \sigma\tilde Z_{(i)}$, we have

$$\tilde\sigma^2_k \approx \frac{\sum_{i=k+1}^n \hat\beta^2_{(i)}}{\sum_{i=k+1}^n Z^2_{(k)}} \approx \frac{\sum_{i=k+1}^n \hat\beta^2_{(i)}}{\sum_{i=k+1}^n \chi^2_{(k)}}, \tag{41}$$

where the right side equals

$$s^2_k(\text{cor}) = \frac{\sum_{i=k+1}^n \hat\beta^2_{(i)}}{n-\sum_{i=1}^k \chi^2_{(k)}} \tag{42}$$

due the orthogonality of $\mathbf{X}$ and the fact that $\sum_{i=1}^n \chi^2_{(k)} = n$. Thus the HNP estimate $\tilde\sigma^2_k$ and $s^2_k(\text{cor})$ are essentially the same.

The following example illustrates a scenario in which the error variance is not estimable.

*Example 6.* Assume that $X$ is orthonormal and that $\boldsymbol\beta$ is a realization from $\sim \mathrm{N}(\mathbf{0}, \sigma_o^2\mathbf{I})$. Then $\hat{\boldsymbol\beta} \sim \mathrm{N}(\mathbf{0}, (\sigma_o^2+\sigma^2)\mathbf{I})$ and

$$s^2_r(\text{cor}) \approx \sigma_o^2 + \sigma^2$$

for all $r$; that is, it gives a biased estimate. The GCV criterion tends to select a null model. Thus, unless we can assume that $\mathbf{Y}$ is independent of some of the candidate variables, the variance estimates, including (14) and the HNP estimate, fail.

## 7. DISCUSSION: THE COST OF DATA ANALYSIS

A good statistical analysis almost always involves careful graphical data exploration and model diagnostics for identification of a good model structure before estimation. Thus data analysis is, at a deeper level, a type of model selection. The goodness-of-fit statistics of the resulting model thus can be too optimistic. Faraway (1992) made an attempt to measure such a bias using the bootstrap. Draper (1995) discussed the need to incorporate structure uncertainty into a Bayesian paradigm.

Consider a data analysis as a modeling procedure that produces $\hat{\mathbf{Y}}$ from $\mathbf{Y}$, theoretically we can apply the framework proposed in this article for making a proper correction to the goodness-of-fit statistics. But this is not feasible practically. Because a human modeler learns from experience, and thus analysis of one dataset would often affect the analysis of the next dataset, it is not possible to repeat exactly the modeling procedure on a perturbed dataset, as required in Section 3, for computing the GDF. There are also many other practical difficulties.

The results in this article have important philosophic implications however. Corollary 2 states that if a group of variables is highly significant, then the GDF due to searching are almost 0. Table 2 shows that the identification of a tree with five terminal nodes costs much less when the underlying structure is not nonnull than when there is no structure. The implication is that the identification of a clear structure bears little cost, whereas searching through white noise has a heavy cost.

To be more precise, we say that an identified structure is stable if the same structure would be identified by use of perturbed responses $\mathbf{Y}+\boldsymbol\delta$, where $\boldsymbol\delta \sim \mathrm{N}(\mathbf{0}, \tau^2\mathbf{I})$, with $\tau^2$ comparable to the actual error variance $\sigma^2$. From Algorithm 1, the GDF of the modeling procedure would be the same as when the structure is given a priori. Therefore, identification of the structure bears no cost. On the other hand, if an identified structure changes when the response is perturbed, then there could be a large searching cost.

## APPENDIX: PROOFS

### Theory about Algorithm 1

Let $\phi(\cdot)$ be an $n$-variate standard normal density function. We consider the estimate of $D(\mathcal{M})$ to be of the form

$$\hat D(\mathcal{M}) = \sum \frac{\partial E^*_{\mathbf{Y}}[\hat\mu_i(\mathbf{Y}^*)]}{\partial y_i} = \frac{1}{\tau^2}\int \boldsymbol\delta'\hat{\boldsymbol\mu}(\mathbf{Y}+\boldsymbol\delta)\frac{1}{\tau}\phi\left(\frac{\boldsymbol\delta}{\tau}\right) d\boldsymbol\delta, \tag{A.1}$$

where the expectation $E^*_{\mathbf{Y}}$ is taken under the distribution $\mathbf{Y}^* \sim \mathrm{N}(\mathbf{Y}, \tau^2\mathbf{I})$. Note that

$$\begin{aligned} E\hat D(\mathcal{M}) &= \sum \frac{\partial E^{**}_{\mathbf{Y}}[\hat\mu_i(\mathbf{Y}^{**})]}{\partial \mu_i} \\ &= \frac{1}{\sqrt{\sigma^2+\tau^2}}\int \boldsymbol\varepsilon'\hat{\boldsymbol\mu}(\boldsymbol\mu+\boldsymbol\varepsilon)\frac{1}{\sqrt{\sigma^2+\tau^2}}\phi\left(\frac{\boldsymbol\varepsilon}{\sqrt{\sigma^2+\tau^2}}\right) d\boldsymbol\varepsilon, \end{aligned} \tag{A.2}$$

where the expectation $E^{**}$ is taken with respect to the distribution $\mathbf{Y}^{**} \sim \mathrm{N}(\boldsymbol\mu, (\tau^2+\sigma^2)\mathbf{I})$. From (A.2), we have

*Proposition 2.* $\lim_{\tau\to 0} E_\mu \hat{D}(\mathcal{M}) = D(\mathcal{M})$.

Note that Proposition 2 is practically irrelevant. Consider the decomposition

$$E[\hat{D}(\mathcal{M}) - D(\mathcal{M})] = \mathrm{var}(\hat{D}(\mathcal{M})) + [E\hat{D}(\mathcal{M}) - D(\mathcal{M})]^2. \quad \text{(A.3)}$$

As $\tau \to 0$, $\mathrm{var}(\hat{D}(\mathcal{M}))$ goes to $\infty$ when $\hat{\boldsymbol{\mu}}(\mathbf{Y})$ is highly nonlinear locally, as in Example 1, or discontinuous. Thus it is often desirable to avoid $\tau \sim 0$. On the other hand, if $\tau \gg \sigma$, then the bias $[E\hat{D}(\mathcal{M}) - D(\mathcal{M})]^2$ may increase. Empirically, the GDF is not sensitive to $\tau$ for $\tau \in [.5\sigma, \sigma]$.

Algorithm 1 uses a regression method to implement (A.1). Note that as $T \to \infty$,

$$\hat{h}_i = \frac{\sum_{t=1}^{T} (\delta_{it} - \bar{\delta}_i) \hat{\mu}_i^m (\mathbf{Y} + \boldsymbol{\delta}_t)}{\sum_{t=1}^{T} (\delta_{it} - \bar{\delta}_i)^2} \to \frac{1}{\tau^2} \int \hat{\mu}_i(\mathbf{Y} + \boldsymbol{\delta}) \frac{\delta_i}{\tau} \phi\left(\frac{\boldsymbol{\delta}}{\tau}\right) d\boldsymbol{\delta},$$

which is the $i$th term in (A.1). The regression algorithm is more efficient than a Monte Carlo integration because it is conditioned on the actual realizations of $\boldsymbol{\delta}$.

## Proof of Theorem 1

Note that

$$(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})'(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) = \mathrm{RSS}(\mathcal{M}_k) - \varepsilon'\varepsilon + 2\varepsilon'(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}). \quad \text{(A.4)}$$

Thus the value of $P$ that minimizes (13) is

$$P_o = E\varepsilon'(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})/\sigma^2 = \sum E\varepsilon_i \hat{\mu}_i / \sigma^2. \quad \text{(A.5)}$$

By directly differentiating $E\hat{\mu}_i$, we can see that $P_o = D(\mathcal{M})$.

## Proof of Theorem 2

Consider the identity

$$\begin{aligned} n\sigma^2 &= E(\mathbf{Y} - \mu)'(\mathbf{Y} - \mu) \\ &= E(\mathbf{Y} - \hat{\boldsymbol{\mu}})'(\mathbf{Y} - \hat{\boldsymbol{\mu}}) + 2E(\mathbf{Y} - \hat{\boldsymbol{\mu}})'(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \\ &\quad + E(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})'(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}). \end{aligned} \quad \text{(A.6)}$$

By taking the expectation of (A.4), we have

$$E(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})'(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) = E(\mathbf{Y} - \hat{\boldsymbol{\mu}})'(\mathbf{Y} - \hat{\boldsymbol{\mu}}) - n\sigma^2 + 2\sigma^2 D(\mathcal{M}). \quad \text{(A.7)}$$

Combining (A.6) and (A.7), we have

$$E(\mathbf{Y} - \hat{\boldsymbol{\mu}})'(\mathbf{Y} - \hat{\boldsymbol{\mu}}) = (n - D(\mathcal{M}))\sigma^2 - E(\mathbf{Y} - \hat{\boldsymbol{\mu}})'(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}).$$

Thus

$$E\hat{\sigma}^2 = \frac{E(\mathbf{Y} - \hat{\boldsymbol{\mu}})'(\mathbf{Y} - \hat{\boldsymbol{\mu}})}{n - D(\mathcal{M})} = \sigma^2$$

if and only if (17) holds.

## Proof of Theorem 3

(a) To show that $\hat{\sigma}_k^2$ is unbiased, it is necessary only to show that $E(\mathbf{Y} - \hat{\mathbf{Y}}_{M_k})'(\hat{\mathbf{Y}}_{M_k} - \boldsymbol{\mu}) = 0$, where $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} = \mathbf{0}$ because $\boldsymbol{\beta} = \mathbf{0}$. Let $\hat{\mathbf{Y}}_{M_k} = H_k \mathbf{Y}$, where $H_k$ may depend on $\mathbf{Y}$. From the properties of the least squares estimates, we have $(\mathbf{Y} - \hat{\mathbf{Y}}_{M_k})'\hat{\mathbf{Y}}_{M_k} \equiv 0$, independent of the model selected. Because $X$ is always included in regression, we also have $(I - H_k)'X \equiv 0$, independent of the model selected. Therefore, $(\mathbf{Y} - \hat{\mathbf{Y}}_{M_k})'\boldsymbol{\mu} = \mathbf{Y}(I - H_k)'X\boldsymbol{\beta} \equiv 0$. This completes the proof.

(b) Let $M_k^o$ be a given model with $k$ selected variables from $\mathbf{X}$, and let $M_q$ be the full model. We have

$$\begin{aligned} (\mathbf{Y} - \hat{\mathbf{Y}}_{M_k^o})'(\mathbf{Y} - \hat{\mathbf{Y}}_{M_k}^o) &\geq (\mathbf{Y} - \hat{\mathbf{Y}}_{M_k})'(\mathbf{Y} - \hat{\mathbf{Y}}_{M_k}) \\ &\geq \min\{(\mathbf{Y} - \hat{\mathbf{Y}}_{M_q})'(\mathbf{Y} - \hat{\mathbf{Y}}_{M_q}), 0\}. \end{aligned}$$

Combining this with (19) gives the desired result.

## Proof of Corollary 1

This result follows from $E(\hat{\sigma}_k^2) = \sigma^2$ and Theorem 3(b).

## Proof of Corollary 2

Let $S_t$ be the model with minimum RSS among all models of the form

$$\mathbf{Y} = \sum_{j=1}^{l} \beta_j \mathbf{x}_j + \sum_{h=1}^{t} \beta_{j_h} \mathbf{x}_{j_h},$$

where $j_h > l$, $h = 1, \ldots, t$. Under the given assumptions, $P(M_k = S_{k-l}) \to 1$. The desired results are obtained after Theorem 3 is applied to $S_{k-l}$.

## A.1 Proof of Example 5

We use the unbiasedness of the EAIC to obtain $D(1)$. Let $\hat{\mathbf{Y}} = \hat{\beta}_{j_1}\mathbf{x}_{j_1}$ and $\boldsymbol{\mu} = \beta\mathbf{x}_1 + \beta\mathbf{x}_2$. Because $\sigma^2 = 1$, we have

$$\begin{aligned} 2D(1) &= E[(\hat{\mathbf{Y}} - \boldsymbol{\mu})'(\hat{\mathbf{Y}} - \boldsymbol{\mu})] - E[(\hat{\mathbf{Y}} - \mathbf{Y})'(\hat{\mathbf{Y}} - \mathbf{Y})] + 2 \\ &= E[(\hat{\beta}_{j_1} - \beta)^2 + \beta^2] - E\hat{\beta}_{j_2}^2 + 2 \\ &= E\left[\sum_{i=1}^{2} (\hat{\beta}_{j_i} - \beta)^2\right] + \beta^2 - E(\hat{\beta}_{j_2} - \beta)^2 - E\hat{\beta}_{j_2}^2 + 2 \\ &= 4 - 2E[\hat{\beta}_{j_2}(\hat{\beta}_{j_2} - \beta)], \end{aligned}$$

which gives the desired result.

## Proof of Theorem 4

(a) Without loss of generality, assume that $\mathbf{X}'\mathbf{X} = \mathbf{I}$, where $\mathbf{I}$ is an identity matrix. Let $\hat{\beta}_i = \mathbf{x}_i'\mathbf{Y}$. Let $(j_1, \ldots, j_q)$ be a permutation of $(1, \ldots, q)$, so that $|\hat{\beta}_{j_1}| \geq \ldots \geq |\hat{\beta}_{j_q}|$. The model $M_k$ contains the variables in $\mathbf{X}$ and the variables $\mathbf{x}_{j_1}, \ldots, \mathbf{x}_{j_k}$. Because $\boldsymbol{\beta} = \mathbf{0}$, we have

$$(\mathbf{Y} - \hat{\mathbf{Y}}_{M_k})'(\mathbf{Y} - \hat{\mathbf{Y}}_{M_k}) = (\mathbf{Y} - \hat{\mathbf{Y}}_{M_q})'(\mathbf{Y} - \hat{\mathbf{Y}}_{M_q}) + \sum_{i=k+1}^{q} \hat{\beta}_{j(i)}^2.$$

From the fact that $\hat{\beta}_i^2/\sigma^2 \sim \chi_1^2$, we have

$$E(\mathbf{Y} - \hat{\mathbf{Y}}_{M_k})'(\mathbf{Y} - \hat{\mathbf{Y}}_{M_k})/\sigma^2 = n - \sum_{i=1}^{k} \chi_{(i)}^2.$$

An application of Theorem 3(a) gives the desired result.

(b) Let $\hat{\mathbf{Y}}_i = \hat{\beta}_i \mathbf{x}_i$ be the vector of fitted values using the variable $\mathbf{x}_i$. By the least squares theory, we have $(\mathbf{Y} - \hat{\mathbf{Y}}_i)'(\mathbf{Y} - \hat{\mathbf{Y}}_i) = \mathbf{Y}'\mathbf{Y} - \hat{\beta}_i^2$. Hence

$$\min_i (\mathbf{Y} - \hat{\mathbf{Y}}_i)'(\mathbf{Y} - \hat{\mathbf{Y}}_i) = n - \max_i [\hat{\beta}_i^2].$$

By a theorem of Slepian (see Tong 1980, thm. 2.1.1), $E[\max_i(\hat{\beta}_i^2)]$ achieves its maxima, $\chi_{(1)}^2$, when $\boldsymbol{\beta}_i$, $i = 1, \ldots, q$ are independent; that is, when $\mathbf{x}_i$, $i = 1, \ldots, q$ are orthogonal.

*[Received December 1996. Revised August 1997.]*

## REFERENCES

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

Bickel, P. and Zhang, P. (1992), "Variable Selection in Nonparametric Regression with Categorical Covariates," *Journal of the American Statistical Asociation*, 87, 90–97.

Breiman, L. (1992), "The Little Bootstrap and Other Methods for Dimensionality Selection in Regression: X-Fixed Prediction Error," *Journal of the American Statistical Association*, 87, 738–754.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984), *Classification and Regression Trees*, Belmont, CA: Wadsworth.

Chipman, H., George, E. I., and McCullogh, R. E. (1996), "Bayesian CART," Technical Report, University of Chicago, Graduate School of Business.

Clark, L. A., and Pregibon, D. (1992), "Tree-Based Models," in *Statistical Models in S*, eds. J. M. Chambers, and T. J. Hastie, New York: Chapman and Hall.

Craven, P., and Wahba, G. (1979), "Smoothing Noisy Data With Spline Functions," *Numerical Mathematics*, 31, 377–403.

Donoho, D. L., and Johnstone, I. M. (1994), "Ideal Spatial Adaptation by Wavelet Shrinkage," *Biometrika*, 81, 425–456.

Draper, D. (1995), "Assessment and Propagation of Model Uncertainty" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 57, 45–98.

Efron, B. (1983), "Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation," *Journal of the American Statistical Association*, 78, 316–331.

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

Eubank, R. L., and Spiegelman, C. H. (1990), "Testing the Goodness of Fit of a Linear Model via Nonparametric Regression Techniques," *Journal of the American Statistical Association*, 85, 387–392.

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

Friedman, J. H. (1991), "Multivariate Adaptive Regression Splines," (with discussion), *The Annals of Statistics*, 19, 1–67.

George, E. I., and Foster, D. P. (1994), "The Risk Inflation Criterion for Multiple Regression," *The Annals of Statistics*, 22, 1947–1975.

Gu, C. (1996), "Model Indexing and Smoothing Parameter Selection in Nonparametric Function Estimation," Technical Report, Purdue University, Dept. of Statistics.

Hastie, T. J., and Tibshirani, R. J. (1990), *Generalized Additive Models*, London: Chapman and Hall.

Mallows, C. L. (1973), "Some Comments on $C_p$," *Technics*, 15, 661–675.

Miller, A. J. (1990), *Subset Selection in Regression*, London: Chapman and Hall.

Owen, A. (1991), "Comment on 'Multivariate Adaptive Regression Splines'," by J. H. Friedman, *The Annals of Statistics*, 19, 102–112.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

Staniswalis, J. G., and Severini, T. A. (1991), "Diagnostics for Assessing Regression Models," *Journal of the American Statistical Association*, 86, 684–692.

Venables, W. N., and Ripley, B. D. (1994), *Modern Applied Statistics with S-Plus*, New York: Springer-Verlag.

Wahba, G. (1983), "Bayesian 'Confidence Intervals' for the Cross-Validated Smoothing Spline," *Journal of Royal Statistical Society*, Ser. B, 45, 133–150.

Zahn, D. (1975a), "Modifications of and Revised Critical Values for the Half-Normal Plot," *Technics*, 17, 189–200.

——— (1975b), "An Empirical Study of the Half-Normal Plot," *Technics*, 17, 201–211.

Zhang, P. (1991), "Variable Selection in Nonparametric Regression With Continuous Covariates," *The Annals of Statistics*, 19, 1869–1882.

——— (1992), "On the Distribution Properties of Model Selection Criteria," *Journal of the American Statistical Association*, 87, 732–737.