CROSS-VALIDATION, THE JACKKNIFE, AND THE BOOTSTRAP: EXCESS ERROR ESTIMATION IN FORWARD LOGISTIC REGRESSION

BY

GAIL GONG

TECHNICAL REPORT NO. 192
AUGUST 1982

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT MCS80-24649

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
CROSS-VALIDATION, THE JACKKNIFE, AND THE BOOTSTRAP:
EXCESS ERROR ESTIMATION IN FORWARD LOGISTIC REGRESSION

BY

GAIL GONG

TECHNICAL REPORT NO. 192
August 1982

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT MCS80-24649

Also prepared under the Auspices of Public Health Service Grant
5 R01 GM21215-08 and issued as Technical Report No. 80 - Stanford
University, Division of Biostatistics.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
CROSS-VALIDATION, THE JACKKNIFE, AND THE BOOTSTRAP: EXCESS ERROR ESTIMATION IN FORWARD LOGISTIC REGRESSION

by

Gail Gong

Abstract

Given a prediction rule based on a set of patients, what is the probability of incorrectly predicting the outcome of a new patient? Call this probability the true error. An optimistic estimate is the apparent error, or the proportion of incorrect predictions on the original set of patients, and it is the goal of this paper to study estimates of the excess error, or the difference between the true and apparent errors.

We consider three estimates of the excess error: cross-validation, the jackknife, and the bootstrap. Under regularity conditions, we show that cross-validation and the jackknife are close - their difference is $O(n^{-2})$. We also compare, using both simulations and real data, the three estimates for a specific prediction rule. When the prediction rule is allowed to be complicated, overfitting becomes a real danger, and excess error estimation becomes important. The prediction rule we choose to study is moderately complicated, involving a variable selection procedure based on forward logistic regression.

Key Words: Prediction, error rate estimation, bootstrap, jackknife, cross-validation, forward logistic regression, variables selection.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Introduction</td>
<td>1</td>
</tr>
<tr>
<td>II. Definitions</td>
<td>5</td>
</tr>
<tr>
<td>III. The Asymptotic Relationship Between Cross-Validation and the Jackknife</td>
<td>12</td>
</tr>
<tr>
<td>VI. Checking Regularity Conditions</td>
<td>20</td>
</tr>
<tr>
<td>V. Chronic Hepatitis - An Example</td>
<td>36</td>
</tr>
<tr>
<td>VI. The Performance of Cross-Validation, the Jackknife and the Bootstrap in Simulations</td>
<td>43</td>
</tr>
<tr>
<td>VII. The Relationship Between Cross-Validation and the Jackknife in Simulations</td>
<td>52</td>
</tr>
<tr>
<td>VIII. Some Remarks about the Bootstrap</td>
<td>54</td>
</tr>
</tbody>
</table>

Appendix  
References  

82
I. Introduction

Cross-validation is a technique for assessing prediction rules. Suppose we are in a prediction setting where we observe $n$ patients

$$x_1 = (t_1, y_1), \ldots, x_n = (t_n, y_n),$$

where $y_i$ is a binary variable indicating whether or not the $i$th patient dies of chronic hepatitis and where $t_i$ is a vector of explanatory variables describing the medical history and examinations of the $i$th patient. These $n$ patients are called the training sample, and they are used to build a prediction rule $\tilde{\eta}_x$, where $x = (x_1, \ldots, x_n)'$. The rule $\tilde{\eta}_x$ takes on values "death" or "not death". Given a new patient whose medical history and examinations are summarized by the vector $t_0$, we predict whether or not he will die of chronic hepatitis by $\tilde{\eta}_x(t_0)$. Allowing for the prediction rule to be very complicated, perhaps gotten by transforming and choosing from many variables and estimating parameters, we want to know: What is the error rate, or the probability of predicting a future observation incorrectly?

A possible estimate of the error rate is the proportion of errors that $\tilde{\eta}_x$ makes when applied to the original observations $x_1, \ldots, x_n$. Because the same observations are used for both forming and assessing the prediction rule, this proportion, which we call the apparent error, underestimates the error rate. The apparent error is biased downward.

To get an unbiased estimate, we might use the proportion of incorrect predictions on $N$ new patients. If these new observations
are unobtainable, assign half of the patients to the training sample and predict on the other half. This procedure gives good answers to the wrong problem; it evaluates the performance of a prediction rule based on \( \frac{n}{2} \) patients. Presumably, a prediction rule based on \( n \) patients would perform better, and any experimenter would prefer using the better prediction rule. To get closer to answering the desired question, let the training sample omit only 1 patient, say the first one. Count the number (0 or 1) of errors that the resulting prediction rule makes when it tries to predict the omitted patient. Repeat, replacing the first patient and omitting the second patient from the training sample, and again count the number of errors that the resulting prediction rule makes on the omitted patient. Continue in this way until each patient has been omitted exactly once. We will have constructed \( n \) prediction rules and predicted the outcome of \( n \) patients. The proportion of errors made in these \( n \) predictions is the cross-validation estimate of incorrect prediction.

Stone (1974) is a key reference on cross-validation and has a good historical account. See also Geisser (1975). Hills (1965) discusses estimates of the error rate of the usual Fisher discriminant in the case of two normal populations having common covariance matrix and differing means. He shows that the apparent error estimate can be seriously biased when the number of parameters requiring estimation is large relative to the number of observations. On the other hand, the cross-validation estimate is approximately unbiased, and using a result from Lachenbruch (1965), Hills gives an approximate distribution of the cross-validation estimate.
Half-samples have been used by Mosteller and Wallace (1963) to assess some ad hoc rules for predicting the authorship of The Federalist Papers, and by Snee (1977) to evaluate the predictive abilities of competing models. Also, Stevens (1978) gives an interesting application of half-samples to Kolmogorov-Smirnov tests.

In addition to cross-validation, Efron (1980) discusses two other methods, the jackknife and the bootstrap, for estimating error rates. This paper is an expansion of Efron's work. In Chapter II, we make precise the setting and definitions suggested in the chronic hepatitis example above. Efron proposes an asymptotic relation between the cross-validation and the jackknife and shows his conjecture is true in the case of linear regression. A general proof of his conjecture appears in Chapter III. Regularity conditions are required; to see that these regularity conditions are reasonable, we show in Chapter IV that they hold for a specific example.

Error rate estimation is important when the training sample is small relative to the number of parameters requiring estimation. It is in this situation that the apparent error is seriously biased. In the chronic hepatitis example, if the dimension of $t_i$ is large, we might use a prediction rule that selects a subset of the variables which hopefully are strong predictors. What is the error rate of a prediction rule which involves variable selection? Specifically, we consider a prediction rule which uses forward logistic regression.

In Chapter V, we apply this prediction rule to some chronic hepatitis data collected at Stanford Hospital, and in Chapter VI we use simulations
to compare the performance of cross-validation, the jackknife, and
the bootstrap as estimators of the error rate.

In Chapter VII, we return briefly to the results of Chapter III and
verify that the relationship between the jackknife and cross-validation
holds in simulations. Chapter VIII concludes with some remarks on the
bootstrap.

We remark that cross-validation, the jackknife, and the bootstrap
are nonparametric techniques. In a simple problem such as the one in
which the covariates are normal and the prediction rule is Fisher
discrimination, we would expect the usual parametric estimates of error
to outperform nonparametric techniques. See Lachenbruch and Mickey
(1968) and McLachlan (1980). However, traditional theoretical solu-
tions are hopeless in the chronic hepatitis problem where normality is
blatantly violated and the prediction rule is complicated. Cross-
validation, the jackknife, and the bootstrap are unfazed by such a
problem.
II. Definitions

Let \( x_1, \ldots, x_n \) be iid from an unknown distribution \( F \), and suppose \( x_i = (t_i, y_i) \), where \( t_i \) is a p-dimensional row vector of real-valued explanatory variables and \( y_i \) is a real-valued response. Let \( \mathbf{x} = (x_1, \ldots, x_n)' \) and \( \hat{F} \) be the empirical distribution which puts mass \( 1/n \) at each of the points in \( \mathbf{x} \). Form the prediction rule \( \eta_{\hat{F}} \); having available the explanatory vector \( t_0 \) of a new observation, we predict the corresponding response to be \( \eta_{\hat{F}}(t_0) \). We are assuming that the prediction rule depends on the training sample \( \mathbf{x} \) only through \( \hat{F} \). Let \( Q(y_0, \eta) \) measure the error between the observed value \( y \) and its predicted value \( \eta \). Define the 

\[
q = q(\hat{F}, F) = E_{x_0 \sim F} Q(y_0, \eta_{\hat{F}}(t_0)) ,
\]

where the expectation is over \( x_0 = (t_0, y_0) \) from \( F \). The first argument of \( q \) refers to the training sample while the second refers to the distribution over which the expectation is taken. A possible estimate of \( q(\hat{F}, F) \) is the apparent error of \( \eta_{\hat{F}} \):

\[
\hat{q}_{\text{app}} = q(\hat{F}, \hat{F}) = E_{x_0 \sim \hat{F}} Q(y_0, \eta_{\hat{F}}(t_0)) = \frac{1}{n} \sum_{i=1}^{n} Q(y_i, \eta_{\hat{F}}(t_i)) .
\]

Because \( \eta_{\hat{F}} \) is usually chosen to fit \( \hat{F} \), the apparent error \( q(\hat{F}, \hat{F}) \) is typically smaller than the true error \( q(\hat{F}, F) \). We call the difference

5
\[ R(\hat{F},F) = q(\hat{F},F) - q(F,F) \]

\[ = E_{x \sim F} Q(y_i, \eta_{\hat{F}}(t_0)) - E_{x \sim \hat{F}} Q(y_0, \eta_{\hat{F}}(t_0)) , \]

the excess error of \( \eta_{\hat{F}} \). We want to estimate the expected excess error

\[ r = E_{\hat{F} \sim F} R(\hat{F},F) \]

where the expectation is taken over the random variable \( \hat{F} \) which is gotten from \( x \sim \) generated by \( F \).

The definitions here apply to quite general situations. In the chronic hepatitis example, \( y \) and \( \eta \) are both binary

\[ y_i = \begin{cases} 1 & \text{if the ith patient dies} \\ 0 & \text{otherwise} \end{cases} , \]

and

\[ Q(y,\eta) = \begin{cases} 1 & \text{if } y \neq \eta \\ 0 & \text{otherwise} \end{cases} , \]

counts the number of incorrect predictions. We may be interested in other measures of error, for example,

\[ Q(y,\eta) = \begin{cases} 1 & \text{if } y = 1 \text{ and } \eta = 0 \\ 0 & \text{otherwise} \end{cases} , \]

which counts the number of times a patient who will die of chronic hepatitis is predicted incorrectly. The prediction rule \( \eta_{\hat{F}} \) may be
gotten by logistic regression, discriminant analysis, or any other
more-or-less ad hoc procedure. The definitions also apply to the
usual linear regression situation in which \( y \) and \( \eta \) take real val-
ues and the measure of error

\[
Q(y, \eta) = (y - \eta)^2
\]

corresponds to the sum of squares.

We will consider three estimates, the bootstrap, the jackknife,
and cross-validation, of the expected excess error.

To motivate the bootstrap estimate, first suppose that we know
the true distribution \( F \). Then, in principle, we could perform the
following simulation.

(A) Generate \( x_1^*, \ldots, x_n^* \), a random sample from \( F \). Let \( x^* = (x_1^*, \ldots, x_n^*)' \) and \( \hat{F}^* \) be the empirical distribution of \( x^* \).

(B) Construct \( \hat{\eta}_{\hat{F}^*} \), the prediction rule formed from the training
sample \( x^* \).

(C) Calculate

\[
q(\hat{F}^*, \hat{F}^*) = \frac{1}{n} \sum_{i=1}^{n} Q(y_i^*, \hat{\eta}_{\hat{F}^*}(t_i^*)) ,
\]

the apparent error of \( \hat{\eta}_{\hat{F}^*} \).

(D) Generate \( x_{n+1}^*, \ldots, x_{n+N}^* \) iid from \( F \) and also independent of
\( x^* \), and calculate
\[ q(\hat{F}^*, F) \approx \frac{1}{N} \sum_{i=1}^{N} Q(y_{n+i}^*, \eta_{\hat{F}^*}(t_{n+i}^*)) , \]

an approximation to the true error of \( \eta_{\hat{F}^*} \). The approximation improves as \( N \) increases.

(E) Form the excess error of \( \eta_{\hat{F}^*} \)

\[ R^* = q(F^*, F) - q(\hat{F}^*, \hat{F}^*) . \]

Repeat (A)-(E) \( B \) times to get \( R_1^*, \ldots, R_B^* \). Then

\[ r \approx \frac{1}{B} \sum_{b=1}^{B} R_b^* \]

approximates the expected excess error. The approximation improves as \( B \) increases.

Not knowing \( F \), we use an estimate \( \hat{F} \) of \( F \). This is the underlying idea of the bootstrap. We state the bootstrap algorithm explicitly.

(A') Generate \( x_1^*, \ldots, x_n^* \), a random sample from \( \hat{F} \). That is, \( x_1^*, \ldots, x_n^* \) is a sample with replacement from \( x_1, \ldots, x_n \).

Let \( \hat{x}^* = (x_1^*, \ldots, x_n^*)' \) and \( \hat{F}^* \) be the empirical distribution of \( \hat{x}^* \).

(B') Construct \( \eta_{\hat{F}^*} \), the prediction rule formed from the training sample \( \hat{x}^* \).
(C') Calculate

\[ q(\hat{F}^*, \hat{F}^*) = \frac{1}{n} \sum_{i=1}^{n} Q(y_{i1}, \eta_{\hat{F}}(t_{i1}^*)) . \]

(D') Calculate

\[ q(\hat{F}^*, \hat{F}^*) = \frac{1}{n} \sum_{i=1}^{n} Q(y_{i1}, \eta_{\hat{F}}(t_{i1}^*)) . \]

(E') Form

\[ R^* = q(\hat{F}^*, \hat{F}^*) - q(\hat{F}^*, \hat{F}^*) . \]

Repeat (A')-(E') a large number B times to get \( R_1^*, \ldots, R_B^* \), and define the bootstrap estimate of expected excess error to be

\[ \hat{r}_{\text{boot}} = \frac{1}{B} \sum_{b=1}^{B} R_b^* . \]

More briefly, the bootstrap estimate of the expected excess error \( E_{\hat{F} \sim F} R(\hat{F}, F) \) where

\[ R(\hat{F}, F) = E_{X_0 \sim F} Q(y_0, \eta_{\hat{F}}(t_0)) - E_{X_0 \sim F} Q(y_0, \eta_{\hat{F}}(t_0)) , \]

is

\[ \hat{r}_{\text{boot}} = E_{\hat{F}^* \sim \hat{F}} R(\hat{F}^*, \hat{F}) \]

where
\[ R(\hat{F}^*, \hat{F}) = E_{\mathcal{X}_0 \sim \hat{F}} Q(y_0, \eta_{\hat{F}^*}(t_0)) - E_{\mathcal{X}_0 \sim \hat{F}^*} Q(y_0, \eta_{\hat{F}^*}(t_0)) \]

\[ = \frac{1}{n} \sum_{i=1}^{n} Q(y_i, \eta_{\hat{F}^*}(t_i)) - \frac{1}{n} \sum_{i=1}^{n} Q(y_i^*, \eta_{\hat{F}^*}(t_i^*)) . \]

Notice that we have merely replaced \( F \) with \( \hat{F} \) and \( \hat{F}^* \) in the expression for \( R(\hat{F}, F) \).

We define the jackknife estimate of expected excess error to be

\[ \hat{r}_{jack} = (n-1)(R(\cdot) - \hat{R}) \]

where

\[ \mathcal{X}(i) = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) \]

is the vector of observations with the \( i \)th one deleted, and \( \hat{F}^{(i)} \) is the empirical distribution of \( \mathcal{X}(i) \), putting mass \( 1/(n-1) \) at each of the points in \( \mathcal{X}(i) \), and

\[ R_{(i)} = R(\hat{F}^{(i)}, \hat{F}) , \]

\[ R_{(\cdot)} = \frac{1}{n} \sum_{i=1}^{n} R_{(i)} , \]

\[ \hat{R} = R(\hat{F}, \hat{F}) . \]

Because
\[ R_{(i)} = R(\hat{\beta}(i), \hat{\beta}) \]

\[ = E_{x_0 \sim \hat{\beta}} Q(y_0, \eta_{\hat{\beta}(i)}(t_0)) - E_{x_0 \sim \hat{\beta}(i)} Q(y_0, \eta_{\hat{\beta}(i)}(t_0)) \]

\[ = \frac{1}{n} \sum_{j=1}^{n} Q(y_j, \eta_{\hat{\beta}(i)}(t_j)) - \frac{1}{n-1} \sum_{j \neq i} Q(y_j, \eta_{\hat{\beta}(i)}(t_j)) \]

\[ = \frac{1}{n-1} \left[ Q(y_i, \eta_{\hat{\beta}(i)}(t_i)) - \frac{1}{n} \sum_{j=1}^{n} Q(y_j, \eta_{\hat{\beta}(i)}(t_j)) \right] , \]

and because \( \hat{R} = 0 \), then we may re-express the jackknife estimate

\[ \hat{r}_{\text{jack}} = \frac{1}{n} \sum_{i=1}^{n} Q(y_i, \eta_{\hat{\beta}(i)}(t_i)) - \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} Q(y_j, \eta_{\hat{\beta}(i)}(t_j)) . \]

The cross-validation estimate of expected excess error is

\[ \hat{r}_{\text{cross}} = \frac{1}{n} \sum_{i=1}^{n} Q(y_i, \eta_{\hat{\beta}(i)}(t_i)) - \frac{1}{n} \sum_{i=1}^{n} Q(y_i, \eta_{\hat{\beta}(i)}(t_i)) . \]

The definitions made here are those of Efron (1982a). Another important reference on the bootstrap is Efron (1979), and an excellent reference on the jackknife is Miller (1974).
### III. The Asymptotic Relationship Between Cross-Validation and the Jackknife

Consider the linear model

\[ y = t\beta + \varepsilon \]

where \( \beta' = (\beta_1, \ldots, \beta_p) \) and

\[
\begin{align*}
\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, & \quad t = \begin{pmatrix} t_1 \\ \vdots \\ t_n \end{pmatrix}, & \quad \begin{pmatrix} t_{11} & \cdots & t_{1p} \\ \vdots & \ddots & \vdots \\ t_{n1} & \cdots & t_{np} \end{pmatrix}, & \quad \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.
\end{align*}
\]

Form the prediction rule

\[ \eta_F(t_0) = t_0\hat{\beta}, \]

where \( \hat{\beta} \) is the usual least squares estimate

\[ \hat{\beta} = (t't)^{-1}t'y; \]

and let the measure of error be

\[ Q(y, \eta) = (y - \eta)^2. \]

Then the true error of \( \eta_F \) is

\[ q(\hat{F}, F) = E_{x_0,F} Q(y_0, \eta_F(t_0)) = E_{x_0,F} (y_0 - \eta_F(t_0))^2, \]

and the apparent error,
\[ q(\hat{F}, \hat{F}) = E_{x_0 \sim \hat{F}} Q(y_0, \eta_\hat{F}(t_0)) = \frac{1}{n} \sum_{i=1}^{n} (y_i - t_i \hat{\beta})^2 \]

is the mean square residual. Efron (1980) shows that in this linear model context,

\[ \hat{r}_{\text{jack}} - \hat{r}_{\text{cross}} = O\left(\frac{1}{n^2}\right). \]

We will show that under regularity conditions, this result is general.

Before stating the general result in Proposition 1, we will give the precise regularity conditions and a lemma which will be used in the proof of the proposition. In Chapter IV, we will show that the linear model example satisfies the regularity conditions, and therefore, the set of problems to which the proposition applies is nonempty.

**Regularity Conditions 1.** Introduce the notation

\[ (3.1) \quad S(\hat{F}, x_0) = Q(y_0, \eta_\hat{F}(t_0)). \]

Holding \( x_0 \) fixed, \( S \) is a functional on a class of distributions. If we consider only those distributions \( \tilde{F}^* \) which restrict their support to \( \{x_1, \ldots, x_n\} \), then we may define the function \( s(\cdot, x_0) \) on

\[ \mathcal{P}^* = \left\{ p^*: p^*_\alpha = \frac{m^\alpha}{n} \land \sum_{\alpha=1}^{n} m^\alpha = n \right\} \]

a subset of \( \mathbb{R}^n \) by
(3.2) \[ s(P^*, x_0) = S(F^*, x_0) \]

where

\[ P^* = \begin{pmatrix} P_1^* \\ \vdots \\ P_n^* \end{pmatrix} = \begin{pmatrix} F^*(x_1) \\ \vdots \\ F^*(x_n) \end{pmatrix}. \]

Efron calls \( P^* \) the resampling vector of \( F^* \) and \( P^* \) the bootstrap resampling space. Special resampling vectors are

\[ p^0 = \begin{pmatrix} \hat{F}(x_1) \\ \vdots \\ \hat{F}(x_n) \end{pmatrix} = \frac{1}{n} \mathbf{1} \]

and

\[ p(j) = \begin{pmatrix} \hat{F}^{(j)}(x_1) \\ \vdots \\ \hat{F}^{(j)}(x_n) \end{pmatrix} = \frac{1}{n-1} (1 - \varepsilon_j) \]

where \( \mathbf{1} \) is the vector of all components equal to one, and \( \varepsilon_j \) is the vector whose only nonzero component is the \( j \)th component equal to one.

Assume that \( s(\cdot, x_0) \) can be extended to the \( n \)-dimensional simplex

\[ \mathcal{P} = \left\{ P : p_\alpha \geq 0 \quad \text{and} \quad \sum_{\alpha=1}^{n} p_\alpha = 1 \right\}. \]
Further, to facilitate differentiation, extend $s(\cdot, x_0)$ to $\mathbb{R}^n$ by defining

$$s(P, x_0) = s\left(\frac{P}{\|P\|}, x_0\right).$$

Define

$$v_{ij}^{\alpha \gamma} = \left. \frac{\partial^2}{\partial P_\alpha \partial P_\gamma} s(P, x_1) \right|_{P = p^i j},$$

where $p^{ij}$ is any point between $p^0$ and $p^{(j)}$, and assume

$$\frac{1}{n^2} \sum_{ij} v_{ij}^{jj} < \infty \text{ a.s.},$$

$$\frac{1}{n^3} \sum_{ij\alpha} v_{ij}^{\alpha j} < \infty \text{ a.s.},$$

$$\frac{1}{n^4} \sum_{ij\alpha\gamma} v_{ij}^{\alpha \gamma} < \infty \text{ a.s.}.$$

(3.3)

**Lemma 1.** If $s$ is a function on $\mathbb{R}^n$ such that

$$s(P) = s\left(\frac{P}{\|P\|}\right),$$

then

$$s(p^0) - s(p^{(j)}) = \frac{1}{n-1} u_j - \frac{1}{2(n-1)^2} v_{jj}^j,$$

(3.5)

where for $\alpha = 1, \ldots, n$ and $\gamma = 1, \ldots, n$, 

where for $\alpha = 1, \ldots, n$ and $\gamma = 1, \ldots, n$, 

15
\[ U_\alpha = \left. \frac{\partial^2}{\partial P_\alpha \partial P_\gamma} s(P) \right|_{p=p^0} , \]

\begin{equation}
(3.6) \quad v^j_{\alpha \gamma} = \left. \frac{\partial^2}{\partial P_\alpha \partial P_\gamma} s(P) \right|_{p=p^j} ,
\end{equation}

\[ v^j_{\alpha \gamma} = v^j_{\alpha \gamma} - \frac{1}{n} \sum_{\gamma=1}^{n} v^j_{\alpha \gamma} - \frac{1}{n} \sum_{\alpha=1}^{n} v^j_{\alpha \gamma} + \frac{1}{n^2} \sum_{\alpha=1}^{n} v^j_{\alpha \gamma} , \]

and \( p^j \) lies between \( p^0 \) and \( p(j) \). Also

\begin{equation}
(3.7) \quad \sum_{\alpha=1}^{n} U_\alpha = 0 ,
\end{equation}

\begin{equation}
(3.8) \quad \sum_{\gamma=1}^{n} v^j_{\alpha \gamma} = \sum_{\alpha=1}^{n} v^j_{\alpha \gamma} = 0 .
\end{equation}

**Proof of Lemma 1.** In matrix notation, (3.6) becomes

\[ U = \nabla s(P^0) , \]

\[ v^j = \nabla v s(P^j) , \]

\[ v^j = v^j - \frac{1}{n} \sum_{\gamma=1}^{n} v^j_{\gamma} - \frac{1}{n} \sum_{\alpha=1}^{n} v^j_{\alpha} + \frac{1}{n^2} \sum_{\alpha=1}^{n} v^j_{\alpha} . \]

Using \( \sum_{\gamma=1}^{n} = n \), we easily see that

\[ (v^j)^{\gamma}_{\gamma} = 0 , \]

which is (3.8). Also, using the chain rule \( (ds(P+\varepsilon R)/d\varepsilon)|_{\varepsilon=0} = \nabla s(P) R \) , and using (3.4), \( s(P) = s(cP) \) for any constant \( c \) , and so we have

\[ \nabla s(P) P = \frac{d}{d\varepsilon} s(P+\varepsilon P)|_{\varepsilon=0} = 0 . \]
Specializing to \( P = p^0 \) gives
\[
\frac{1}{n} \sum_{\alpha=1}^{n} U_\alpha = U' p^0 = \nabla' s(p^0) \cdot p^0 = 0 ,
\]
which is (3.7).

To show (3.5) we expand \( s(p^{(j)}) \) about \( p^0 \):
\[
s(p^{(j)}) - s(p^0) = (p^{(j)} - p^0)' U + \frac{1}{2} (p^{(j)} - p^0)' \nabla^j (p^{(j)} - p^0) .
\]
Because \( (p^{(j)} - p^0)' 1 = 0 \),
\[
(p^{(j)} - p^0)' V(p^{(j)} - p^0) = (p^{(j)} - p^0)' v^j (p^{(j)} - p^0) ,
\]
which gives
\[
s(p^{(j)}) - s(p^0) = (p^{(j)} - p^0)' U + \frac{1}{2} (p^{(j)} - p^0)' v^j (p^{(j)} - p^0) .
\]

Finally, because \( p^{(j)} - p^0 = \frac{1}{n-1} (\frac{1}{n} 1 - e_j) \),
\[
s(p^{(j)}) - s(p^0) = \frac{1}{n-1} (\frac{1}{n} 1 - e_j)' U + \frac{1}{2(n-1)^2} (\frac{1}{n} 1 - e_j)' v^j (\frac{1}{n} 1 - e_j)
\]
\[
= -\frac{1}{n-1} U_j + \frac{1}{2(n-1)^2} v^j_{jj} ,
\]
which is (3.5).

**Proposition 1.** Define
\[
\Omega_n = \hat{r}_{\text{jack}} - \hat{r}_{\text{cross}}
\]
\[
= \frac{1}{n^2} \sum_{ij} [Q(y_i, \eta_{\hat{p}}(t_i)) - Q(y_i, \eta_{\hat{p}}(j)(t_i))]
\]
Then under the Regularity Conditions 1, $\mathcal{D}_n = 0 \frac{1}{n^2}$.

Proof of Proposition 1. Under (3.1) and (3.2) and denoting $s(P, x_i) = s^i(P)$,

$$\mathcal{D}_n = \frac{1}{n^2} \sum_{i,j} \left[ Q(y_i, \eta_{\hat{P}}(t_i)) - Q(y_i, \eta_{\hat{P}}(j)(t_i)) \right],$$

$$= \frac{1}{n^2} \sum_{i,j} \left[ S(\hat{P}, x_i) - S(\hat{P}(j), x_i) \right],$$

$$= \frac{1}{n^2} \sum_{i,j} \left[ s(0, x_i) - s(P(j), x_i) \right],$$

$$= \frac{1}{n^2} \sum_{i,j} \left[ s^i(0) - s^i(P(j)) \right],$$

and using Lemma 1 and obvious extensions of the notation,

$$\mathcal{D}_n = \frac{1}{n^2} \sum_{i,j} \left[ \frac{1}{n-1} u^i_{\gamma j} - \frac{1}{2(n-1)^2} v^i_{\gamma j} \right]$$

$$= \frac{1}{n^2} \left[ \frac{1}{n-1} \sum_{i,j} u^i_{\gamma j} - \frac{1}{2(n-1)^2} \sum_{i,j} v^i_{\gamma j} \right].$$

The first sum in the bracket vanishes by (3.7)

$$\sum_{i,j} u^i_{\gamma j} = \sum_{i,j} u^i_{\gamma j} = 0,$$

and the second sum can be expanded

$$\sum_{i,j} v^i_{\gamma j} = \frac{1}{n^2} \sum_{i,j} (v^i_{\gamma j} - 2v^i_{\alpha j} + v^i_{\alpha \gamma}).$$

Therefore,
\[ \omega_n = \frac{-1}{2(n-1)^2} \sum_{ij \alpha \gamma} \frac{1}{n^4} \left( v_{ij} - 2v_{ij}^{\alpha j} + v_{ij}^{i \gamma} \right), \]

and by (3.3), \( \omega_n = O\left(\frac{1}{n^2}\right). \)
IV. Checking Regularity Conditions

To see that Regularity Conditions 1 are reasonable, we show in this chapter that they hold for a specific example. Although in later chapters we will be concerned with a prediction rule \( \eta_0 \) which is
based on forward logistic regression, here we check Regularity Conditions 1 for a much simpler prediction rule.

We return to the linear model

\[
y = t \beta + \epsilon .
\]

Recall that \( x_1 = (t_1, y_1), \ldots, x_n = (t_n, y_n) \) are iid from \( F \). Here, we also assume

\[
E(t_1) = 0 ,
\]

\[
\text{Cov}(t_1) = E(t_1 t_1) = \Sigma > 0 ,
\]

\[
E(t_{ik}) < \infty ,
\]

\[
E(\epsilon | t) = 0 ,
\]

\[
\text{Cov}(\epsilon | t) = \sigma^2 I_n .
\]

We can rewrite the prediction rule \( \eta_0(t_0) = t_0 \hat{\beta} \) to exhibit explicitly that it is a functional of \( \hat{F} \):

\[
\eta_0(t_0) = t_0 \hat{\beta}
\]

\[
= t_0 (t' t)^{-1} t' y
\]

\[
= t_0 \left( \sum_{\alpha=1}^{n} t_\alpha t_\alpha p_\alpha \right)^{-1} \left( \sum_{\alpha=1}^{n} t_\alpha y_\alpha p_\alpha \right)
\]

\[
= t_0 \left( \int t_0' t_0 d\hat{F}(x_0) \right)^{-1} \left( \int t_0' y_0 d\hat{F}(x_0) \right).
\]
Defining
\[ \eta_p(t_0) = t_0 \left( \sum_{\alpha=1}^{n} t^\alpha t^\alpha p^\alpha \right)^{-1} \left( \sum_{\alpha=1}^{n} t^\alpha y^\alpha p^\alpha \right), \]
the function
\[ s(P, x_0) = (y_0 - \eta_p(t_0))^2 \]
defined on the space \(\mathbb{R}^n\) is an obvious extension of
\[ s(P^*, x_0) = S(F^*, x_0) = (y_0 - \eta_{P^*}(t_0))^2 \]
defined on the bootstrap resampling space \(P^*\), where \(P^* = (F^*(x_1), \ldots, F^*(x_n))^t\). Denoting
\[ s^i(P) = s(P, x_i), \]
\[ v^{ij}_{\alpha\gamma} = \left. \frac{\partial^2}{\partial p^\alpha \partial p^\gamma} s^i(P) \right|_{P=P^{ij}}, \]
where \(P^{ij}\) lies between \(P^0\) and \(P^{(j)}\), we want to show
\[ (4.1) \quad \frac{1}{n^2} \sum_{ij} v^{ij} < \infty, \quad \frac{1}{n^3} \sum_{ij\alpha} v^{ij}_{\alpha j} < \infty, \quad \frac{1}{n^4} \sum_{ij\alpha\gamma} v^{ij}_{\alpha\gamma} < \infty \]
almost surely.

**Lemma 1.** Define
\[ A_p = \sum_{\alpha=1}^{n} t_{\alpha}^\prime t_{\alpha} P_{\alpha} \]

\[ B_p = \sum_{\alpha=1}^{n} t_{\alpha}^\prime y_{\alpha} P_{\alpha} \]

\[ q_{\alpha Y}^+ = t_{\alpha} A_{p+1}^t \]

\[ d_{\alpha}^+ = y_{\alpha} - \eta_{p+1}(t_{\alpha}) \]

where \( p_{ij} = p_{ij}^+ \). Then

\[ v_{\alpha Y}^{ij} = 2(q_{i\alpha}^+ q_{i\gamma}^+ d_{\alpha Y}^+ + q_{\alpha Y}^+ d_{i\gamma Y}^+ + q_{\gamma Y}^+ q_{\gamma \alpha}^+ d_{i \alpha}^+) \]

and so

\[ v_{jj}^{ij} = 2((q_{i j}^+ d_{j}^+) + 2q_{i j}^+ q_{j j}^+ d_{1 j}^+) \]

\[ v_{\alpha j}^{ij} = 2(q_{\alpha i}^+ q_{i j}^+ d_{\alpha j}^+ + q_{\alpha j}^+ d_{1 j}^+ + q_{j i}^+ q_{j \alpha}^+ d_{1 \alpha}^+) \]

The proof of Lemma 1, a straightforward computation of derivatives, appears in the Appendix.

To show (4.1), we want to show

\[ \frac{1}{n^4} \sum_{i,j,a,Y} \phi_{n i j a Y} < \infty \text{ a.s. ,} \]

where \( \phi_{n i j a Y} \) is any one of the following eight products
(a) \((q_{ij}^+ d_{ij}^+)^2\),

(b) \(q_{ij}^+ q_{j}^+ d_{ij}^+ d_{ij}^+\),

(c) \(q_{\alpha i}^+ q_{ij}^+ d_{\alpha}^+ d_{ij}^+\),

(d) \(q_{\alpha i}^+ q_{\alpha j}^+ d_{i}^+ d_{ij}^+\),

(4.4)

(e) \(q_{ji}^+ q_{j}^+ d_{ij}^+ d_{ij}^+\),

(f) \(q_{\alpha i}^+ q_{ij}^+ d_{\alpha}^+ d_{ij}^+\),

(g) \(q_{\alpha i}^+ q_{\alpha j}^+ d_{ij}^+ d_{ij}^+\),

(h) \(q_{\alpha i}^+ q_{\alpha j}^+ d_{ij}^+ d_{ij}^+\).

The proof of (4.3) will follow a series of lemmas, some of which appear painfully complicated though not conceptually difficult. To get the main idea consider the case when \(\phi_{nij\alpha\gamma}\) is (4.4a). Since \(p^+ \approx p^0\) for large \(n\), then \(q_{ij}^+ \approx q_{ij}^0 \equiv t_i A_{-1}^{-1} t_j^t\). Also, \(\eta_{p^+}(t_0) \rightarrow t_0^\beta \equiv \eta(t_0)\) and so \(d_{ij}^+ \equiv y_j^\alpha - \eta(t_j)\). Therefore,

\[
\phi_{nij\alpha\gamma} = (q_{ij}^+ d_{ij}^+)^2 = (q_{ij}^0 d_{ij}^0)^2 + \Delta_{nij},
\]

where \(\Delta_{nij}\) is small, and so

\[
\frac{1}{n^2} \sum_{ij \alpha \gamma} \phi_{nij\alpha\gamma} = \frac{1}{n^2} \sum_{ij} (q_{ij}^0 d_{ij}^0)^2 + \frac{1}{n^2} \sum_{ij} \Delta_{nij},
\]

23
where
\[
\frac{1}{n^2} \sum_{i,j} (q^0_{ij}d_{ij})^2 = \frac{1}{n} \sum_{j=1}^{n} d_j^2 t_j A^{-1} p_0 \left( \frac{1}{n} \sum_{i=1}^{n} t_i't_i A^{-1} p_0 t_j' \right)
\]
\[
= \frac{1}{n} \sum_{j=1}^{n} d_j^2 t_j A^{-1} p_0 t_j'
\]
\[
= \text{trace} \left( \frac{1}{n} \sum_{j=1}^{n} d_j^2 t_j't_j A^{-1} \right)
\]
\[
a.s.
\Rightarrow \text{trace}(\sigma^2 \Sigma^{-1})
\]

The convergence follows from the strong law together with
\[
E(d_1^2t_1't_1) = E(E(d_1^2|t_1)t_1't_1)
\]
\[
= E(\sigma^2 t_1't_1)
\]
\[
= \sigma^2 \Sigma
\]

Since \( \Delta_{nij} \) is small for sufficiently large \( n \), we hope that
\[
\frac{1}{n^2} \sum_{i,j} \Delta_{nij} < \infty
\]

**Lemma 2.** Defining
\[
d_\alpha^0 = y_\alpha - n_\alpha(t_\alpha),
\]
\[
d_\alpha^0 = t_\alpha A^{-1} p_0 t_\alpha'
\]
we have

24
\[ a_\alpha^+ = a_\alpha^0 + R_{nij} q_{\alpha j}^0 q_{\alpha j}^0 , \]

(4.5)

\[ q_{\alpha \gamma}^+ = \frac{1}{n g_{ni}} (q_{\alpha \gamma}^0 + R_{nij} q_{\alpha j}^0 q_{\gamma j}^0) \]

where

\[ R_{nij} = \frac{\frac{r_{ni}}{n}}{1 - \frac{r_{ni}}{n} q_{jj}^0} , \]

\[ r_{ni} \in [0,1] , \]

\[ \frac{1}{n g_{ni}} \in \left[ \frac{n-1}{n}, 1 \right] , \]

**Proof.** Since \( p^+ \) lies between \( p^0 \) and \( p^{(j)} \), then for some \( \theta_i \in [0,1] \),

\[ p^+ = \theta_i p^0 + (1-\theta_i)p^{(j)} \]

(4.6)

\[ = g_{ni}^1 - h_{ni} e_j , \]

where

\[ g_{ni} = \frac{1}{n-1} (1 - \frac{\theta_i}{n}) , \]

\[ h_{ni} = \frac{1}{n-1} (1-\theta_i) . \]

Substituting (4.6) into our formulas for \( A_{p^+} \) and \( B_{p^+} \) (4.2) we get

(4.7)

\[ A_{p^+} = n g_{ni} \left( A_{p^0}^0 - \frac{r_{ni}}{n} t^t j \right) \]

25
\begin{equation}
B_{p}^\dagger = ng_{ni} \left( B_{p0} - \frac{r_{ni}}{n} t_{j}^{\prime} j \right),
\end{equation}

where \( r_{ni} = h_{ni} / g_{ni} \in [0,1] \). Recall Woodbury's formula

\[
(M + uv')^{-1} = M^{-1} - \frac{M^{-1} uv' M^{-1}}{1 + v' M^{-1} u},
\]

where \( M \) and \( M + uv' \) are invertible matrices and \( u \) and \( v \) are vectors. Substituting \( M = A_{p0} \), \( u = -\frac{r_{ni}}{n} t_{j}^{\prime} \), and \( v = t_{j}^{\prime} \), and using (4.7)

\begin{equation}
A_{p}^{-1} = \frac{1}{ng_{ni}} (A_{p0}^{-1} + R_{ni} A_{p0}^{-1} t_{j}^{\prime} t_{j} A_{p0}^{-1}).
\end{equation}

Now to get the desired results, substitute (4.8) and (4.9) into

\[
q_{\alpha Y}^{\dagger} = t_{\alpha} A_{p}^{-1} t_{Y}^{\prime},
\]

\[
d_{\alpha}^{\dagger} = y_{\alpha} - \eta_{p} (t_{\alpha}^{\prime}) = y_{\alpha} - t_{\alpha} A_{p}^{-1} B_{p}^{\dagger}.
\]

For more details, see the Appendix.

In subsequent lemmas, we will find the concept of matrix norms to be highly useful. For \( u \in \mathbb{R}^{p} \), let \( \|u\| = (u'u)^{1/2} \) be the usual Euclidean norm of \( u \). For \( A \in \mathbb{R}^{p \times p} \), define the "matrix norm" of \( A \) to be

\[
\|A\| = \max_{\|u\| = 1} \|Au\|
\]

A nice discussion of matrix norms can be found in Stewart (1973). We list some useful properties.
(a) \[ \| A \| > 0 \] for \( A \neq 0 \) and \( \| A \| = 0 \) implies \( A = 0 \).

(b) \[ \| \kappa A \| = |\kappa| \| A \| \] for any \( \kappa \in \mathbb{R} \).

(c) \[ \| A + B \| \leq \| A \| + \| B \| . \]

(d) \[ \| AB \| \leq \| A \| \cdot \| B \| . \]

(e) \[ \| Au \| \leq \| A \| \cdot \| u \| . \]

(f) \[ |u^t A v| \leq \| u \| \cdot \| A \| \cdot \| v \| \]

(g) \[ \| A \| = \text{(maximum eigenvalue of } A^t A) \]

(h) If \( \| A \| < 1 \), then \( I + A \) is nonsingular and

\[ \| (I + A)^{-1} \| \leq \frac{1}{1 - \| A \|} . \]

Let \( A_n, A \) be \( p \times p \) random matrices. We will say \( A_n \overset{a.s.}{\rightarrow} A \)
if each element converges almost surely. Since for any \( i, j \in \{1, \ldots, p\} \),

\[ |(A_n - A)_{ij}| = |e_i^t (A_n - A) e_j| \]

\[ \leq \| e_i \| \cdot \| A_n - A \| \cdot \| e_j \| \]

\[ = \| A_n - A \| \]

and for some \( u \in \mathbb{R}^p \),
\[ \| A_n - A \|^2 = \| (A_n - A)u \|^2 \]

\[ = \sum_{i=1}^{P} \left( \sum_{j=1}^{P} (A_n - A)_{ij} u_j \right)^2, \]

we see that \( A_n \overset{a.s.}{\rightarrow} A \) whenever \( \| A_n - A \| \overset{a.s.}{\rightarrow} 0 \). Also, because

\[ \| A_n \| \leq \| A_n - A \| + \| A \|, \]

\[ \| A \| \leq \| A_n - A \| + \| A_n \|, \]

\[ \| \| A_n \| - \| A \| \| < \| A_n - A \|, \]

if \( A_n \overset{a.s.}{\rightarrow} A \), then \( \| A_n \| \overset{a.s.}{\rightarrow} \| A \| \).

The next lemma may seem unmotivated; the reader may wish to jump to Lemma 4, referring to Lemma 3 as the need arises.

**Lemma 3.** Let \( \varepsilon \in (0,1) \) and define

\[ \kappa_1 = \| \varepsilon^{-1} \| + 1, \]

\[ \kappa_2 = \min \left( 1, \frac{1}{\kappa_1} \right), \]

\[ \kappa_3 = \frac{1}{1 - \varepsilon}, \]

\[ \delta_\alpha = |d_\alpha|, \]

\[ \tau_\alpha = \| t_\alpha \|, \]

and also,
\[ \Omega_2 = \{ \omega : \exists n_2(\omega) \forall n > n_2(\omega) \ [ \| A^{-1} \| < \kappa_1 & \| \hat{\beta} - \beta \| < \varepsilon ] \} , \]

\[ \Omega_3 = \{ \omega : \exists j_0(\omega) \forall j > j_0(\omega) \ [ \tau_j^2 < \varepsilon j \kappa_2 & \delta_j^2 < \varepsilon j ] \} , \]

\[ \Omega_4 = \{ \omega : \exists n_4(\omega) \forall n > n_4(\omega) \forall j \leq n \ [ \kappa_1 \tau_j^2 < \varepsilon n ] \} , \]

\[ \Omega_5 = \left\{ \frac{1}{n} \sum_{\alpha=1}^{n} \tau_\alpha^2, \frac{1}{n} \sum_{\alpha=1}^{n} \tau_\alpha \delta_\alpha, \frac{1}{n} \sum_{\alpha=1}^{n} \tau_\alpha^4, \frac{1}{n} \sum_{\alpha=1}^{n} \tau_\alpha^3 \delta_\alpha, \frac{1}{n} \sum_{\alpha=1}^{n} \tau_\alpha^2 \delta_\alpha^2 \text{ all converge} \right\} , \]

\[ \Omega_1 = \Omega_2 \cap \Omega_3 \cap \Omega_4 \cap \Omega_5 . \]

Then \( P(\Omega_1) = 1. \)

\textbf{Proof.} Because

\[ E \tau_1^2 = E t_1 t_1' = \text{trace} E t_1 t_1' = \text{trace} \Sigma < \infty , \]

\[ E \delta_1^2 = E \{ E(\delta_1^2 | t_1) \} = E \delta_1^2 = \sigma^2 < \infty , \]

\[ E |\tau_1 \delta_1| \leq (E \tau_1^2 E \delta_1^2)^{1/2} < \infty , \]

\[ E \tau_1^4 = E (t_1 t_1')^2 = E \sum_{k=1}^{p} \sum_{k'=1}^{p} t_k t_1 k' t_{1k'} < \infty , \]

\[ E |\tau_1^3 \delta_1| = E(|\tau_1|^3 E(\delta_1^3 | t_1)) \leq E|\tau_1|^3 \sigma < \infty , \]

then by the strong law, \( P(\Omega_5) = 1. \) Also by the strong law,

\( P(\Omega_2) = 1. \) Because
\[
\sum_{j=1}^{\infty} P\{\tau_j^2 > \varepsilon j \kappa_2^2\} \leq \frac{E \tau_j^2}{\varepsilon \kappa_2^2} < \infty,
\]
\[
\sum_{j=1}^{\infty} P\{\delta_j^2 > \varepsilon j\} \leq \frac{E \delta_j^2}{\varepsilon} < \infty,
\]

then by the Borel-Cantelli Lemma, \( P(\Omega_2) = 1 \). To show \( P(\Omega_4) = 1 \), we show \( \Omega_3 \subseteq \Omega_4 \). Let \( \omega \in \Omega_3 \), and let \( j_0(\omega) \) be as specified in the definition of \( \Omega_3 \). Define
\[
n_4(\omega) = \max\left(j_0(\omega), \frac{\tau_1^2}{\kappa_1}, \frac{\tau_2^2}{\kappa_1}, \ldots, \frac{\tau_j^2}{\kappa_1}\right).
\]

Then \( \forall n > n_4 \), if \( j < j_0 \), \( \kappa_1 \tau_j^2 < \varepsilon n \) and if \( j_0 < j < n \), since \( \omega \in \Omega_3 \), we have \( \kappa_1 \tau_j^2 < \kappa_1 \varepsilon j \kappa_2 < \varepsilon n \).

**Lemma 4.** For any \( \omega \in \Omega_4 \) there exists an \( n_1(\omega) \) such that \( \forall n > n_1 \) and \( \forall j \leq n \) and \( \forall i, \alpha, \gamma \),
\[
|d_\alpha^+| \leq \delta_\alpha + \tau_\alpha \mu_j
\]
\[
|q_{\alpha \gamma}^+| \leq \tau_\alpha \tau_\gamma \nu_j
\]

where
\[
\mu_j = \varepsilon + \frac{1}{n} \kappa_1 \kappa_3 T_j(\delta_j + \varepsilon \tau_j),
\]
\[
\nu_j = \kappa_1 (1 + \frac{1}{n} \kappa_1 \kappa_3 \tau_j^2),
\]

and \( \kappa_1, \kappa_3, \delta_\alpha, \tau_\alpha \) are defined in Lemma 3. Notice that both \( \mu_j \) and \( \nu_j \) depend on \( n \).
Proof. Let $\omega \in \Omega_1$ and let $n_1(\omega) = \max(n_2(\omega), n_4(\omega))$ where $n_2(\omega)$ and $n_4(\omega)$ are as specified in the definitions of $\Omega_2$ and $\Omega_4$ respectively. Let $n > n_1$. We show for $j \leq n$,

\[(4.10)\]
\[|R_{nij}| < \frac{\kappa_3}{n}.\]

Because

\[q_{jj}^0 = t_j A^{-1} p_0 t_j' \leq \|A^{-1} t_j\|^2 \leq \kappa_1 \|t_j\|^2 < \varepsilon n\]

(the first inequality follows from matrix norms properties, the second and third inequalities follow from the definitions in Lemma 3 of $\Omega_2$ and $\Omega_4$ respectively) then $q_{jj}^0/n < \varepsilon$, and

\[R_{nij} = \frac{r_{ni}/n}{1 - r_{ni}q_{jj}^0/n},\]

is a continuous and increasing function of $r_{ni}$, taking values in $[0, \frac{1/n}{1-q_{jj}^0/n}]$ when $r_{ni} \in [0,1]$. Therefore,

\[|R_{nij}| \leq \frac{1/n}{1 - q_{jj}^0/n} < \frac{1/n}{1-\varepsilon} = \frac{\kappa_3}{n}.\]

Also, it is easy to show

\[(4.11)\]
\[\delta_\alpha^0 \leq \delta_\alpha + \varepsilon \tau_\alpha,\]

\[|q_{\alpha\gamma}^0| \leq \kappa_1 \tau_\alpha \tau_\gamma.\]

Substituting (4.10) and (4.11) into (4.5) gives the desired results.
Using Lemma 4 we can give bounds for each product in (4.4). Notice that the index \( j \) has a special role whereas \( i, \alpha, \gamma \) may be interchanged. Therefore, the bound of (4.4d) looks like that of (4.4c) and the bounds of both (4.4g) and (4.4h) look like that of (4.4f). The proof of the following lemma, a straightforward but tedious application of Lemma 4, is relegated to the Appendix.

**Lemma 5.** For any \( \omega \in \Omega_1 \), there exists an \( n_1(\omega) \) such that \( \forall n > n_1 \) and \( \forall j \leq n \) and \( \forall i, \alpha, \gamma \),

\[
(q^+_i d^+_j)^2 \leq (\tau^2_i \tau^2_j \delta_j) v^2_j + 2(\tau^2_i \tau^2_j \delta_j) \mu_j v^2_j + (\tau^2_i \tau^4_j) \mu_j^2 v^2_j,
\]

\[
|q^+_i qq^+_j \delta^+_i \delta^+_j| \leq (\tau^2_i \tau^2_j \delta_j) v^2_j + (\tau^2_i \tau^2_j \delta_j) \mu_j v^2_j + (\tau^2_i \tau^2_j \delta_j) \mu_j^2 v^2_j,
\]

\[
|q^+_i q^+_j d^+_j d^+_j| \leq (\tau^2_i \tau^2_j \delta_j) v^2_j + (\tau^2_i \tau^2_j \delta_j) \mu_j v^2_j + (\tau^2_i \tau^2_j \delta_j) \mu_j^2 v^2_j,
\]

\[
|q^+_i q^+_j d^+_j d^+_j| \leq (\tau^2_i \tau^2_j \delta_j) v^2_j + (\tau^2_i \tau^2_j \delta_j) \mu_j v^2_j + (\tau^2_i \tau^2_j \delta_j) \mu_j^2 v^2_j,
\]

\[
|q^+_i q^+_j d^+_j d^+_j| \leq (\tau^2_i \tau^2_j \delta_j) v^2_j + (\tau^2_i \tau^2_j \delta_j) \mu_j v^2_j + (\tau^2_i \tau^2_j \delta_j) \mu_j^2 v^2_j,
\]
\[ |q_{i\alpha}^+ q_{j\gamma}^+ d_{i\alpha}^+ d_{j\gamma}^+| \leq (\tau_1^2)(\tau_2\delta_2)(\tau_3\delta_3)v_j^2 + (\tau_1^2)(\tau_2\delta_2)(\tau_3\mu_j v_j^2) + (\tau_1^2)(\tau_2\delta_2)(\tau_3\mu_j)^2 v_j^2. \]

Recall that we want to show (4.3) for \( \phi_{nij\alpha\gamma} \) being any one of (4.4). It now suffices to show

\[ \frac{1}{n} \sum_{ij\alpha\gamma} \xi_{nij\alpha\gamma} < \infty \text{ a.s.,} \]

where \( \xi_{nij\alpha\gamma} \) is any term of any bound in Lemma 5. Look at \( \xi_{nij\alpha\gamma} = (\tau_1^2)(\tau_2\delta_2 v_j^2) \), the first term of the first bound in Lemma 5.

\[ \frac{1}{n} \sum_{ij\alpha\gamma} \xi_{nij\alpha\gamma} = \left( \frac{1}{n} \sum_{i=1}^{n} \tau_1^2 \right) \left( \frac{1}{n} \sum_{j=1}^{n} \tau_2^2 \delta_2 v_j^2 \right). \]

The first term on the right hand side converges if \( \omega \in \Omega_5 \) and since \( v_j = \kappa_1 \left( 1 + \frac{1}{n} \kappa_1 \kappa_3 \tau_j^2 \right) \), we hope for sufficiently large \( n \), the second term is approximately

\[ \frac{1}{n} \sum_{j=1}^{n} \tau_2^2 \delta_2 v_j^2 \kappa_1, \]

which converges if \( \omega \in \Omega_5 \). Making the above idea precise is the object of our last lemma.

\textbf{Lemma 6.} Define

\[ \psi_{nij\alpha\gamma} = \xi_{nij\alpha\gamma} I\{\tau_j^2 < \varepsilon \kappa_2 \delta_j < \varepsilon j \}, \]

where \( \xi_{nij\alpha\gamma} \) is any term of any bound in Lemma 5. If \( \omega \in \Omega_1 \),
then
\begin{equation}
\frac{1}{n^4} \sum_{ij\alpha\gamma=1}^{n} (\xi_{nij\alpha\gamma} - \psi_{nij\alpha\gamma}) \xrightarrow{a.s.} 0, \end{equation}

and
\begin{equation}
\frac{1}{n^4} \sum_{ij\alpha\gamma=1}^{n} \psi_{nij\alpha\gamma} < \infty, \end{equation}

and therefore,
\begin{equation}
\frac{1}{n^4} \sum_{ij\alpha\gamma=1}^{n} \xi_{nij\alpha\gamma} \xrightarrow{a.s.} 0. \end{equation}

**Proof.** Let $\omega \in \Omega_1$ and $j_0(\omega)$ be as specified in the definition of $\Omega_3$. For $n > j_0$,
\begin{equation}
\frac{1}{n^4} \sum_{ij\alpha\gamma=1}^{n} (\xi_{nij\alpha\gamma} - \psi_{nij\alpha\gamma}) = \frac{1}{n^3} \sum_{i\alpha\gamma=1}^{n} \frac{1}{n} \sum_{j=1}^{j_0} \xi_{nij\alpha\gamma}. \end{equation}

Consider the example $\xi_{nij\alpha\gamma} = (\tau_i^2)(\tau_j^2)(\sigma_j^2)(\nu_j^2)$, the first term of the first bound in Lemma 5.
\begin{equation}
\frac{1}{n^3} \sum_{i\alpha\gamma=1}^{n} \frac{1}{n} \sum_{j=1}^{j_0} \xi_{nij\alpha\gamma} = \left(\frac{1}{n} \sum_{i=1}^{n} \tau_i^2\right)\left(\frac{1}{n} \sum_{j=1}^{j_0} \tau_j^2\right)\left(\frac{1}{n} \sum_{j=1}^{j_0} \sigma_j^2\right)\left(\frac{1}{n} \sum_{j=1}^{j_0} \nu_j^2\right). \end{equation}

Since $\nu_j$ decreases with $n$ (also $\mu_j$ decreases with $n$), then
\begin{equation}
\frac{1}{n} \sum_{j=1}^{j_0} \tau_j^2\sigma_j^2\nu_j^2 \to 0, \end{equation}

and since $\omega \in \Omega_5$ implies $\frac{1}{n} \sum_{i=1}^{n} \tau_i^2$ converges, we have (4.12) for this example of $\xi_{nij\alpha\gamma}$. Other possibilities may be treated similarly.
Notice that if \( \tau_j^2 < \varepsilon j \kappa_2 \) and \( \delta_j^2 < \varepsilon j \), then \( \tau_j \delta_j < \varepsilon j \sqrt{\kappa_2} \), and for \( j \leq n \),

\[
\mu_j = \varepsilon + \frac{1}{n} \kappa_1 \kappa_3 \tau_j (\delta_j + \varepsilon \tau_j) \\
\leq \varepsilon + \frac{1}{n} \kappa_1 \kappa_3 (\varepsilon j \sqrt{\kappa_2} + \varepsilon^2 j \kappa_2) \\
\leq \varepsilon + \kappa_1 \kappa_3 (\varepsilon \sqrt{\kappa_2} + \varepsilon^2 \kappa_2) \equiv \mu ,
\]

and

\[
\nu_j = \kappa_1 (1 + \frac{1}{n} \kappa_1 \kappa_3 \tau_j^2) \\
\leq \kappa_1 (1 + \frac{1}{n} \kappa_1 \kappa_3 \varepsilon j \kappa_2) \\
\leq \kappa_1 (1 + \kappa_1 \kappa_2 \kappa_3 \varepsilon) \equiv \nu .
\]

For the example \( \xi_{nij\alpha \gamma} = (\tau_i^2) (\tau_j^2 \delta_j^2) \nu_j^2 \),

\[
\psi_{nij\alpha \gamma} = (\tau_i^2) (\tau_j^2 \delta_j^2) \nu_j^2 \ \mathbb{I}\{\tau_j^2 < \varepsilon j \kappa_2 \ \& \ \delta_j^2 < \varepsilon j\} \\
\leq (\tau_i^2) (\tau_j^2 \delta_j^2) \nu^2 ,
\]

and so

\[
\frac{1}{n} \sum_{i,j} \psi_{nij\alpha \gamma} \leq \left( \frac{1}{n} \sum_{i=1}^{n} \tau_i^2 \right) \left( \frac{1}{n} \sum_{j=1}^{n} \tau_j^2 \delta_j^2 \right) \nu^2 ,
\]

which converges since \( \omega \in \Omega \). Other possibilities of \( \xi_{nij\alpha \gamma} \) may be similarly treated.
V. Chronic Hepatitis - An Example

We now discuss a real prediction rule. Dr. Peter Gregory of Stanford Hospital observed \( n = 155 \) chronic hepatitis patients, of which 33 died from the disease. On each patient were recorded \( p = 19 \) covariates summarizing medical history, physical examinations; x-rays, liver function tests, and biopsies. An effective prediction rule, based on these 19 covariates, was desired to identify future patients at high risk. Such patients require more aggressive treatment.

Display 1 shows the data for the last 11 patients. Negative numbers indicate missing observations which were replaced by column averages before beginning any analysis of the data.

Gregory used a prediction rule based on forward logistic regression. Recall that the logistic model assumes \((t_1, y_1), \ldots, (t_n, y_n)\) are independent and identically distributed such that conditional on \( t_i \), \( y_i \) is Bernoulli with probability of success \( \theta(t_i) \), where \( \theta(t_0) \) is the logistic function of the linear combination \( \beta_0 + t_0 \beta \):

\[
\theta(t_0) = \frac{\exp(\beta_0 + t_0 \beta)}{1 + \exp(\beta_0 + t_0 \beta)},
\]

or equivalently,

\[
\log \left[ \frac{\theta(t_0)}{1 - \theta(t_0)} \right] = \beta_0 + t_0 \beta,
\]

and where \( \beta \) is a column vector of \( p \) elements. If \((\hat{\beta}_0, \hat{\beta})\) is an estimate of \((\beta_0, \beta)\), then
\[
\hat{\theta}(t_0) = \frac{\exp(\hat{\beta}_0 + t_0\hat{\beta})}{1 + \exp(\hat{\beta}_0 + t_0\hat{\beta})},
\]

is an estimate of \( \theta(t_0) \), and a reasonable rule \( \eta^*_\beta \) would predict death if the estimated probability of death were greater than \( \frac{1}{2} \):

\[
\eta^*_\beta(t_0) = \begin{cases} 
1 & \text{if } \hat{\theta}(t_0) \geq \frac{1}{2} \\
0 & \text{otherwise}
\end{cases}
\]

Gregory's rule estimates \( (\beta_0, \beta) \) in the following way:

(A) Perform an initial screening of the variables by testing for \( j = 1, \ldots, p \) separately at level \( \alpha = 0.05 \)

\[ H_0 : \beta_j = 0, \]

in the logistic model

\[
\log \left[ \frac{\theta(t_0)}{1 - \theta(t_0)} \right] = \beta_0 + t_0 \beta_j.
\]

Retain only those variables \( j \) for which the test is significant. Thirteen variables

\[ 17, 12, 14, 11, 13, 19, 6, 5, 18, 10, 1, 4, 2, \]

in increasing order of p-value, and therefore decreasing order of significance, were retained.

(B) To the variables that survive the first screening, apply forward logistic regression which adds variables one at a time in the following way. Assuming variables \( j_1, j_2, \ldots, j_{p_1} \) are already added, test
\[ H_0 : \beta_j = 0 , \]
in the logistic model
\[ \log \left[ \frac{\theta(t_0)}{1 - \theta(t_0)} \right] = \beta_0 + \sum_{k=1}^{p_1} t_{0j_k} \beta_j + t_{0j} \beta_j , \]
for \( j \neq j_1, j_2, \ldots, j_{p_1} \). Use Rao's (1973) efficient score test which requires calculating the maximum likelihood only under \( H_0 \). If at least one of these tests is significant at \( \alpha = 0.05 \), let \( j_{p_1+1} \) correspond to the most significant test. Stop when no further test achieves significance \( \alpha = 0.05 \). When applied to the thirteen variables above, forward logistic regression chose variables

17, 11, 14, 2.

(C) Let \( (\hat{\beta}_0, \hat{\beta}) \) be the mle based on the logistic model consisting only of the variables chosen by forward logistic regression together with the constant. It turned out that
\[
\begin{align*}
\hat{\beta}_0 &= 12.17 , \\
\hat{\beta}_{17} &= -1.83 , \\
\hat{\beta}_{11} &= -1.58 , \\
\hat{\beta}_{14} &= 0.56 , \\
\hat{\beta}_2 &= -5.17 , 
\end{align*}
\]

(5.3)

Gregory's prediction rule was
\[
\eta_\hat{\theta}(t_0) = \begin{cases} 
1 & \text{if } \hat{\theta}(t_0) \geq \frac{1}{2} , \\
0 & \text{otherwise} 
\end{cases}
\]

(5.4)

where \( \hat{\theta}(t_0) \) is defined by
\[
\log \left[ \frac{\hat{\theta}(t_0)}{1 - \hat{\theta}(t_0)} \right] = 12.17 - 1.83 \, t_{0,17} - 1.58 \, t_{0,11} \\
+ 0.56 \, t_{0,14} - 5.17 \, t_{0,2}.
\]

Since variables 17, 11, 14, and 2 are albumin, spiders, bilirubin, and sex, we see that Gregory's rule predicts that a patient with low albumin, low spiders, high bilirubin, and male sex is more likely to die. The rule is sensible according to the variable values associated with high risk in Display 1. The apparent error was \( \hat{\eta}_{\text{app}} = 0.156 \).

Display 2 shows a histogram of \( B = 400 \) bootstrap replications of \( R^* = R(\hat{F}^*, \hat{F}) \). Recall that each \( R^* \) was calculated using (2.1)-(2.3) where \( \eta_{F^*} \) is the prediction rule gotten by applying Gregory's rule (A)-(C) and (5.1)-(5.2) to a bootstrap sample \( x^*_1, x^*_2, \ldots, x^*_n \). That is, to compute \( R^*_1, R^*_2, \ldots, R^*_B \), we needed to perform \( B = 400 \) forward logistic regressions. The bootstrap estimate of expected excess error was

\[
\hat{r}_{\text{boot}} = \frac{1}{B} \sum_{b=1}^{B} R^*_b = 0.039.
\]

Also, the jackknife and cross-validation estimates were calculated to be

\[
\hat{r}_{\text{jack}} = 0.023,
\]

\[
\hat{r}_{\text{cross}} = 0.019.
\]

Adding expected excess error estimates to the apparent error gives bias-corrected estimates of the error:
\[ \hat{q}_{\text{boot}} = 0.175 , \]
\[ \hat{q}_{\text{jack}} = 0.159 , \]
\[ \hat{q}_{\text{cross}} = 0.145 . \]

Fortran programs for performing the above calculations and the ones in the following chapter were developed on a PDP-11/34 minicomputer. The cross-validation and jackknife estimates were computed in 1\(\frac{1}{2}\) hours, while the 400 bootstrap replications required under 6 hours.

Are \( B = 400 \) bootstrap replications enough? Components of variance analyses for computing appropriate \( B \)'s are possible in simulations, but not on real data. In practical applications, a simple approach looks at

\[ \hat{r}_B = \frac{1}{B} \sum_{b=1}^{B} R_b^* , \]

for increasing values of \( B \) and stops when only small changes occur.

For the chronic hepatitis data,

\[ \hat{r}_{100} = 0.0390 , \]
\[ \hat{r}_{200} = 0.0393 , \]
\[ \hat{r}_{300} = 0.0392 , \]
\[ \hat{r}_{400} = 0.0393 . \]

We expect that increasing \( B \) beyond 400 would change \( \hat{r}_B \) little.

For another way to answer the question, notice \( R_1^* , R_2^* , \ldots , R_B^* \) is a random sample from a population with mean
\[ E_{P* \sim \hat{P}} R(\hat{P}^*, \hat{P}) = \hat{r}_{\text{boot}} = \hat{r}_{\infty}, \]

and variance, say \( \sigma^2 \). Display 2 shows that this population is close to normal, so

\[ |\hat{r}_{400} - \hat{r}_{\infty}| \leq \frac{2\sigma}{\sqrt{400}}, \]

with high probability. Approximating \( \sigma^2 \) with

\[ \hat{\sigma}^2_{400} = \frac{1}{400-1} \sum_{b=1}^{400} (R_b^* - \hat{r}_{400})^2 = (.027)^2 \]

gives

\[ |\hat{r}_{400} - \hat{r}_{\infty}| \leq \frac{2(.027)}{\sqrt{400}} = .0027 \]

and so with high probability, \( \hat{r}_{400} \) is within 0.0027 of \( \hat{r}_{\infty} = \hat{r}_{\text{boot}} \).

Before leaving the chronic hepatitis data, we mention that other prediction rules might be used. Examples include more complicated forms of variable selection such as best subset regression and alternative models such as discriminant analysis. Jerome Friedman applied recursive partitioning to these data and got the binary-decision tree in Display 3 for predicting death in chronic hepatitis patients. Each patient travels down the tree according to his covariate values. For example, a patient with large protime and small bilirubin would travel down the right branch of the top node and then the left branch of the next node to arrive in terminal node T4. The tree predicts that patients arriving in terminal nodes T1, T5, and T7 to die while those arriving in terminal nodes T2, T3, T4, and T6 to live. We will not
discuss here Friedman's procedure for building the tree. See Friedman (1977). It is interesting to note that the variables, protime, bilirubin, alkaline phosphate, albumin, varices, and age, used in Friedman's rule correspond only marginally with albumin, spiders, bilirubin and sex, the variables used in Gregory's rule. Friedman's tree had apparent error $\hat{q}_{\text{app}} = 0.10$ and bias-corrected estimates of error $\hat{q}_{\text{boot}} = 0.15$ (with $B = 30$) and $\hat{q}_{\text{cross}} = 0.12$.

We chose to focus our attention on the rule based on forward logistic regression because it is a real rule, one which was proposed and used by the experimenter, Dr. Gregory. The question of choosing an optimal prediction rule was not our goal.
VI. The Performance of Cross-Validation, the Jackknife, and the Bootstrap in Simulations

In the previous chapter, we saw the bootstrap, jackknife, and cross-validation estimates of expected excess error for Gregory's rule, which involves an initial screening of variables followed by forward logistic regression. These estimates give bias corrections to the apparent error. Do these corrections offer real improvements, and if so, which correction should we use? To help answer these questions, we performed some simulations.

In these simulations, the underlying model was the logistic model which assumes \( x_1 = (t_1, y_1), \ldots, x_n = (t_n, y_n) \) are independent and identically distributed such that conditional on \( t_i \), \( y_i \) is Bernoulli with probability of success

\[
\theta(t_i) = \frac{\exp(\beta_0 + t_i \beta)}{1 + \exp(\beta_0 + t_i \beta)}
\]

and such that \( t_i = (t_{i1}, \ldots, t_{ip}) \) is \( p \)-variate normal with zero mean and a specified covariance structure \( \Sigma \).

We performed two sets of simulations. In the first set, simulations 1.1, 1.2, 1.3, we let the sample sizes be respectively \( n = 20, 40, 60 \), the dimension of \( t_i \) be \( p = 4 \), and

\[
\Sigma = \begin{pmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & \tau & 0 \\
0 & \tau & 1 & 0 \\
0 & 0 & 0 & 1
\end{pmatrix}, \quad \beta_0 = 0, \quad \beta = \begin{pmatrix} 1 \\ 2 \\ 0 \\ 0 \end{pmatrix}
\]
where \( \tau = 0.86 \). We would expect a good prediction rule to choose variables \( t_1 \) and \( t_2 \), and because of the correlation between variables \( t_2 \) and \( t_3 \), a prediction rule choosing \( t_1 \) and \( t_3 \) would probably not be too bad. In the second set of simulations, simulations 2.1, 2.2, 2.3, the sample sizes were again \( n = 20, 40, 60 \), the dimension of \( t_1 \) was increased to \( p = 6 \) and

\[
\Sigma = \begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & \tau & 0 \\
0 & 0 & 0 & \tau & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 \\
\end{pmatrix}, \quad \beta_0 = 0, \quad \beta = \begin{pmatrix}
1 \\
1 \\
2 \\
0 \\
0 \\
0 \\
\end{pmatrix}.
\]

Before discussing the results of the simulations, we pause to emphasize the distinction between the excess error and the expected excess error. In the chronic hepatitis problem, the training sample which Dr. Gregory observed led to a particular prediction rule (5.4)-(5.5). The excess error of this prediction rule is the difference between the true error of this rule and the apparent error of this rule. The expected excess error averages the excess error over the many training samples that Dr. Gregory might have observed, and therefore, over many prediction rules. Recall that we have been thinking of \( \hat{r}_{\text{cross}} \), \( \hat{r}_{\text{jack}} \), and \( \hat{r}_{\text{boot}} \) as estimates of the expected excess error, but Dr. Gregory would much rather know the excess error. Therefore, we will compare the performance of \( \hat{r}_{\text{cross}} \), \( \hat{r}_{\text{jack}} \), and \( \hat{r}_{\text{boot}} \) as estimates of
both the excess error and expected excess error by defining two mean square errors:

\[ \text{MSE}_1(\hat{r}) = E[\hat{r} - R]^2 \]

the mean square error about the excess error, and

\[ \text{MSE}_2(\hat{r}) = E[\hat{r} - E(R)]^2 \]

the mean square error about the expected excess error.

Notice that since

\[ E[\hat{r} - R]^2 = E[(\hat{q}_{\text{app}} + \hat{r}) - (\hat{q}_{\text{app}} + R)]^2 \]

MSE_1(\hat{r}) also measures the performance of the bias-corrected estimate \( \hat{q}_{\text{app}} + \hat{r} \) as an estimator of the true error \( \hat{q}_{\text{app}} + R \).

It is perhaps unfair to think of \( \hat{r}_{\text{cross}}, \hat{r}_{\text{jack}}, \) and \( \hat{r}_{\text{boot}} \) as estimators of the excess error. A simple analogy may be helpful. Suppose \( X \) is an observation from the distribution \( F_\xi \), and \( T(X) \) estimates \( \xi \). The bias is the expected difference \( E[T(X) - \xi] \) and is analogous to the expected excess error. Getting a good estimate of the bias is sometimes possible, but getting a good estimate of the difference \( T(X) - \xi \) would be equivalent to knowing \( \xi \). The difference \( T(X) - \xi \) is analogous to the excess error.

Turn now to the simulation results. Each simulation consisted of 400 experiments. In the first experiment of Simulation 1.1, a training sample, \( x_1 = (t_1, y_1), \ldots, x_{20} = (t_{20}, y_{20}) \), was generated such that \( t_i \) is a 4-dimensional normal vector with zero mean and covariance \( \Sigma \) given in (6.2) and \( y_i \) conditional on \( t_i \) is Bernoulli with probability of success \( \theta(t_i) \) given in (6.1) and (6.2).
Gregory's rule applied to this training sample chose variables $t_1$ and $t_2$ with $\hat{\beta}_0 = -0.30$, $\hat{\beta}_1 = 1.51$, and $\hat{\beta}_2 = 2.85$, and had an apparent error of 0.15.

Details of the first 20 experiments appear in Display 4 and the results of all 400 experiments of Simulation 1.1 are summarized in Display 5. In the first 20 experiments, the bootstrap was closest to the true excess 14 times. The excess error, averaged over 400 experiments, was 0.10 and gives an approximation to the expected excess error. In the first 20 experiments, the bootstrap was closest to the expected excess error 16 times.

Turning to Display 5, we see that since

$$E(\hat{r}_{cross}) = E(\hat{r}_{jack}) = E(R) = 0.10,$$

$\hat{r}_{cross}$ and $\hat{r}_{jack}$ are unbiased estimates of the expected excess error $E(R)$, while $\hat{r}_{boot}$ with expectation

$$E(\hat{r}_{boot}) = 0.08$$

is biased downwards slightly. However, $\hat{r}_{cross}$ and $\hat{r}_{jack}$ have enormous standard deviations 0.11 and 0.10 compared to 0.02, the standard deviation of $\hat{r}_{boot}$. To more easily compare $\hat{r}_{cross}$, $\hat{r}_{jack}$, and $\hat{r}_{boot}$ we introduce the "estimators"

$$\hat{r}_{apparent} = 0,$$

the zero-correction estimate corresponding to the apparent error, and

$$\hat{r}_{ideal} = E(R),$$

the best constant estimate if we knew the expected excess error $E(R)$. We see in Display 5 that
\[ \sqrt{\text{MSE}_1(\hat{r}_{\text{ideal}})} < \sqrt{\text{MSE}_1(\hat{r}_{\text{boot}})} \]

\[ < \sqrt{\text{MSE}_1(\hat{r}_{\text{apparent}})} \approx \sqrt{\text{MSE}_1(\hat{r}_{\text{cross}})} \approx \sqrt{\text{MSE}_1(\hat{r}_{\text{jack}})} . \]

with \( \sqrt{\text{MSE}_1(\hat{r}_{\text{boot}})} \) being about one-third the distance between \( \sqrt{\text{MSE}_1(\hat{r}_{\text{ideal}})} \) and \( \sqrt{\text{MSE}_1(\hat{r}_{\text{apparent}})} \). The same ordering holds for \( \sqrt{\text{MSE}_2} \).

Displays 6 and 7 summarize the results of Simulations 1.2 and 1.3. Recall that Simulations 1.1, 1.2, and 1.3 had the same underlying distribution but differing sample sizes \( n = 20, 40, \) and \( 60 \) respectively. Notice that as sample size increases, the expected excess error decreases as does the mean square error of the apparent error. In the second set of simulations, Simulations 2.1, 2.2, and 2.3, the sample sizes were again \( n = 20, 40, \) and \( 60 \), while the dimension of \( t_1 \) was increased to \( p = 6 \) and \( \Sigma, \beta_0, \) and \( \beta \) given in (6.3). Displays 8, 9, and 10 show that as in the first set, the expected true excess and the mean square error of the apparent error both decrease as the sample size increases. For large sample sizes, bias corrections to the apparent error become unimportant.

It is interesting to note that the ordering of the root mean square error of the five estimates noticed in Simulation 1.1 holds also in the other five simulations. That is,

\[ \sqrt{\text{MSE}(\hat{r}_{\text{apparent}})} \sim \sqrt{\text{MSE}(\hat{r}_{\text{cross}})} \sim \sqrt{\text{MSE}(\hat{r}_{\text{jack}})} , \]

and \( \sqrt{\text{MSE}(\hat{r}_{\text{boot}})} \) is about one-third the distance between
\( \sqrt{\text{MSE}(\hat{r}_{\text{ideal}})} \) and \( \sqrt{\text{MSE}(\hat{r}_{\text{apparent}})} \). Cross-validation and the jackknife offer no improvement over the apparent error, while the improvement given by the bootstrap is substantial.

The superiority of the bootstrap over cross-validation has been observed in other problems. Efron (1982b) discusses estimates of excess error including cross-validation and the bootstrap and performs several simulations with flavor similar to ours. We report on only one of these simulations here. When the prediction rule is the usual Fisher discriminant and the training sample consists of 14 observations which are equally likely from \( h((-1/2,0), I) \) or \( h((1/2,0), I) \), then the root mean square errors of the apparent, cross-validation, bootstrap, and ideal estimates of excess error are respectively 0.149, 0.144, 0.134, 0.114 with rescaled \( \sqrt{\text{MSE}_1} \)'s 1, 0.85, 0.58, 0.

Throughout our simulations, we used \( B = 100 \) bootstrap replications for each experiment. We argue here that this choice of \( B \) is sufficiently large. Let

\[
\hat{r}_B = \frac{1}{B} \sum_{b=1}^{B} R_b^*,
\]

and denote the mean square error about the true excess by

\[
\text{M}(B) = \text{MSE}_1(\hat{r}_B).
\]

Then

48
\[ M(B) = E[\hat{\tau}_B - R]^2 \]

\[ = E[(\hat{\tau}_B - \hat{\tau}_\infty) + (\hat{\tau}_\infty - R)]^2 \]

\[ = E(\hat{\tau}_B - \hat{\tau}_\infty)^2 + 2E(\hat{\tau}_B - \hat{\tau}_\infty)(\hat{\tau}_\infty - R) + E(\hat{\tau}_\infty - R)^2. \]

Let us condition of \( \hat{F} \), the empirical distribution. Then \( R_1^*, ..., R_B^* \) are iid random variables from a distribution with mean \( \mu(\hat{F}) = \hat{\tau}_\infty \) and variance \( \sigma^2(\hat{F}) \), and so

\[ \hat{\tau}_B = \frac{1}{B} \sum_{b=1}^{B} R_B^*, \]

has mean \( \hat{\tau}_\infty \) and variance \( \sigma^2(\hat{F})/B \). Using this conditioning we see that the cross-term in (6.4) vanishes

\[ E(\hat{\tau}_B - \hat{\tau}_\infty)(\hat{\tau}_\infty - R) = E \ E_{\hat{F}}(\hat{\tau}_B - \hat{\tau}_\infty | \hat{F})(\hat{\tau}_\infty - R) \]

\[ = 0, \]

and

\[ E(\hat{\tau}_B - \hat{\tau}_\infty)^2 = E \ E_{\hat{F}}((\hat{\tau}_B - \hat{\tau}_\infty)^2 | \hat{F}) \]

\[ = E \frac{\sigma^2(\hat{F})}{B} \]

\[ = \frac{1}{B} E \sigma^2(\hat{F}). \]

The mean square error can be rewritten

\[ M(B) = \frac{1}{B} E \sigma^2(\hat{F}) + E(\hat{\tau}_\infty - R)^2. \]
In addition to performing Simulation 1.1 with $B = 100$, we also ran it for $B = 400$ to get

$$M(100) = 0.01241651,$$
$$M(400) = 0.01238761.$$ 

Substituting into (6.5) and solving, we get

$$M(\infty) = E[\hat{r}_\infty - R]^2 = 0.01237798,$$
$$\sqrt{M(\infty)} = 0.1113 \approx 0.1114 \approx \sqrt{M(100)}.$$ 

If we are interested in comparing root mean square errors, we need not perform more than $B = 100$ bootstrap replications.

If we are interested in the results of a particular experiment, say the first one in Simulation 1.1, we can perform analyses similar to those made in the chronic hepatitis experiment. First, we can look at changes in $\hat{r}_B$ as $B$ grows. For the first experiment of Simulation 1.1 we got

$$\hat{r}_{100} = 0.1060,$$
$$\hat{r}_{200} = 0.1085,$$
$$\hat{r}_{300} = 0.1050,$$
$$\hat{r}_{400} = 0.1054.$$ 

50
and we feel comfortable in believing \( \hat{r}_{\text{boot}} \) is between 0.10 and 0.11.
Secondly, we can argue that with high probability,

\[
\left| \hat{r}_{100} - \hat{r}_\infty \right| \leq \frac{2(.0656)}{\sqrt{100}} = 0.013,
\]

and so \( \hat{r}_{100} \) is within 0.013 of \( \hat{r}_\infty = \hat{r}_{\text{boot}} \).

We included 400 experiments in each simulation and therefore are using the approximation

\[
\text{MSE}_1(\hat{r}) \equiv \mathbb{E}[\hat{r} - R]^2
\]

\[
\approx \frac{1}{400} \sum_{e=1}^{400} [\hat{r}_e - R_e]^2
\]

where \( \hat{r}_e \) and \( R_e \) are the estimate and true excess of the \( e \)-th experiment. Displays 11 and 12 show 95 percent individual confidence intervals for the rescaled \( \sqrt{\text{MSE}_1} \) and \( \sqrt{\text{MSE}_2} \). Shorter intervals would be preferable but obtaining them would be time-consuming. Four-hundred experiments of Simulation 1.1 with \( p = 4, n = 20 \), and \( B = 100 \) took 16 computer hours while 400 experiments of Simulation 2.3 with \( p = 16, n = 60 \), and \( B = 100 \) took 72 hours. Halving the length of the confidence intervals in Displays 11 and 12 would require four times the number of experiments and four times the amount of computer time.

Confidence intervals aside, the constant relation of the five estimates as measured by either \( \sqrt{\text{MSE}_1} \) or \( \sqrt{\text{MSE}_2} \) is striking. Cross-validation and the jackknife do no better than the apparent error, while the bootstrap offers substantial improvement.
Recall from Chapter III that under suitable regularity conditions \( \hat{r}_{\text{cross}} \) and \( \hat{r}_{\text{jack}} \) are close. Regularity conditions are impossible to verify for Gregory's complicated prediction rule, but we can look at the simulations in Chapter VI to check the relationship between \( \hat{r}_{\text{cross}} \) and \( \hat{r}_{\text{jack}} \). Recall that Simulations 1.1, 1.2, and 1.3 assume the logistic model where \((t_i, y_i), \ldots, (t_n, y_n)\) are iid such that \( y_i \) conditional on \( t_i \) is Bernoulli with probability of death \( \theta(t_i) \) given in (6.1) and (6.2) and such that \( t_i \) is 4-variate normal with covariance \( \Sigma \) given in (6.2). The sample sizes of Simulations 1.1, 1.2, and 1.3 are respectively \( n = 20, 40, \) and \( 60 \).

Display 13 shows that the plot of \( \hat{r}_{\text{cross}} \) and \( \hat{r}_{\text{jack}} \) for the first 100 experiments of Simulation 1.1 lies close to the line with slope 1, while \( \hat{r}_{\text{cross}} \) and \( \hat{r}_{\text{boot}} \) share no such relationship. The correlations of all 400 experiments

\[
\rho(\hat{r}_{\text{cross}}, \hat{r}_{\text{jack}}) = 0.969 ,
\]

\[
\rho(\hat{r}_{\text{cross}}, \hat{r}_{\text{boot}}) = 0.563 ,
\]

reflect our findings in the plots. Even for a sample size as small as \( n = 20 \), \( \hat{r}_{\text{cross}} \) and \( \hat{r}_{\text{jack}} \) are close.

Displays 14 and 15 give plots for Simulations 1.2 and 1.3. As the sample size \( n \) increases, the excess error as well as the estimates of excess error decrease, and therefore for better comparisons,
the plots are rescaled so that the range in each plot is proportional to $E(\hat{r}_{\text{cross}})$. Displays 14 and 15 exhibit behavior similar to that of Display 13. Correlations for Simulations 1.1, 1.2, and 1.3 appear in Display 16. For completeness, Display 16 also includes correlations for Simulations 2.1, 2.2, and 2.3 described in Chapter VI. In every simulation considered $\hat{r}_{\text{cross}}$ and $\hat{r}_{\text{jack}}$ are close.
VIII. Some Remarks About the Bootstrap

The bootstrap is a powerful tool that can squeeze a surprising amount of information from the data. Until now, we have seen it only as a point estimate of the excess error. We conclude with examples of other uses of the bootstrap in the chronic hepatitis problem.

The density of \( R^* = R(\hat{F}^*, \hat{F}) \) with \( \hat{F} \) fixed, estimates the density of \( R(\hat{F}, F) \) considered as a function of the random quantity \( \hat{F} \), so the histogram in Display 2 of \( B = 400 \) replications of \( R^* \) estimates the density of \( R(\hat{F}, F) \). Since the 0.05 and 0.95 quantiles are \( R^*_{(.05)} = -0.006 \) and \( R^*_{(.95)} = 0.084 \), then in 90 percent of the experiments which Dr. Gregory might perform, we expect the excess error to lie in \((-0.006, 0.084)\). So in addition to giving a point estimate of the excess error, the bootstrap can give a confidence interval.

If a new patient is observed with covariates \( t_0 \), Gregory's rule predicts the outcome of the new patient according to the estimated probability of death

\[
\hat{\theta}(t_0) = \frac{\exp(\hat{\beta}_0 + t_0 \hat{\beta})}{1 + \exp(\hat{\beta}_0 + t_0 \hat{\beta})},
\]

where \((\hat{\beta}_0, \hat{\beta})\) is given in (5.3). For example, a new patient with the same covariates \( t_0 \) as Patient 145 in the training sample (Display 1) would have estimated probability of death 0.89. We can use the bootstrap to get a confidence interval for the probability of death. Actually, because it is more symmetrical, we get a confidence interval for the "logit score"
\[
\lambda(t_0) = \log \frac{\theta(t_0)}{1 - \theta(t_0)}.
\]

A simple transformation gives a confidence interval for \(\theta(t_0)\) the probability of death. In producing \(R_1^*, \ldots, R_B^*\) in Chapter V, we performed \(B = 400\) logistic regressions and got 400 bootstrap models \((\hat{\beta}_0^*, \hat{\beta}_1^*)_1, \ldots, (\hat{\beta}_0^*, \hat{\beta}_1^*)_B\). Compute

\[
\lambda_b^* = \hat{\beta}_0^* b + t_0 \hat{\beta}_1^*, \quad b = 1, \ldots, B.
\]

Display 17 shows a histogram of \(\lambda_1^*, \ldots, \lambda_B^*\) with 0.05 and 0.95 quantiles \(\lambda_{.05}^* = -0.516\) and \(\lambda_{.95}^* = 3.964\). Therefore, a 90 percent confidence interval for \(\lambda(t_0)\) is \((-0.5159, 3.9639)\) and a 90 percent confidence interval for \(\theta(t_0)\) the probability of death is \((0.37, 0.98)\).

We are using here the percentile method introduced by Efron (1982a).

Another approach assumes that

\[
\hat{\lambda}(t_0) = \log \frac{\hat{\theta}(t_0)}{1 - \hat{\theta}(t_0)}
\]
is approximately normally distributed with mean \(\lambda(t_0)\) and some variance \(\sigma_P(t_0)\). Bootstrap ideas give an estimate \(\sigma_P(t_0)\) of the variance and therefore confidence intervals for \(\lambda(t_0)\). Efron also discusses a third approach, which he calls the bias-corrected percentile method, for constructing confidence intervals. We only mention here that these bootstrap procedures promise to give realistic inference on the probability of death; comparison of these procedures is the subject of future research.
Our final example shows the real ability of the bootstrap for squeezing information from the data. Forward logistic regression chooses one of many models. What is the variability of the model? Put another way, we saw that Gregory's rule chose the model (5.3) with covariates 17, 11, 4, 2. If Dr. Gregory repeated his entire experiment, would forward logistic regression choose a similar model? The $B = 400$ bootstrap models $(\hat{\beta}_1^*, \hat{\beta}_2^*)_1, \ldots, (\hat{\beta}_0^*, \hat{\beta}_2^*)_B$ can give us a clue. Display 18 shows the covariates with nonzero coefficients for the first 50 bootstrap models. The high degree of variability of the covariates chosen may seem surprising. In an experiment such as Dr. Gregory's in which the causative relationship between covariates and the outcome of a disease is not understood, a usual approach invokes a variable selection technique to choose a model. It is tempting to say that the covariates chosen in the model are special. However, as Display 18 shows, such interpretations can be far from correct. The bootstrap can say something about the variability of the model chosen by forward logistic regression. Current research focuses on this idea.
Display 1: Data for the last 11 chronic hepatitis patients. Negative numbers indicate missing observations. Values in parentheses are associated with high risk.

<table>
<thead>
<tr>
<th>id</th>
<th>y</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
</tr>
</thead>
<tbody>
<tr>
<td>145</td>
<td>1</td>
<td>45</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1.90</td>
<td>-1</td>
<td>114</td>
<td>2.4</td>
<td>-1</td>
<td>2</td>
</tr>
<tr>
<td>146</td>
<td>0</td>
<td>31</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1.20</td>
<td>75</td>
<td>193</td>
<td>4.2</td>
<td>54</td>
<td>2</td>
</tr>
<tr>
<td>147</td>
<td>1</td>
<td>41</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4.20</td>
<td>65</td>
<td>120</td>
<td>3.4</td>
<td>-1</td>
<td>2</td>
</tr>
<tr>
<td>148</td>
<td>1</td>
<td>70</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>1.70</td>
<td>109</td>
<td>528</td>
<td>2.8</td>
<td>35</td>
<td>2</td>
</tr>
<tr>
<td>149</td>
<td>0</td>
<td>20</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0.90</td>
<td>89</td>
<td>152</td>
<td>4.0</td>
<td>-1</td>
</tr>
<tr>
<td>150</td>
<td>0</td>
<td>36</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0.60</td>
<td>120</td>
<td>30</td>
<td>4.0</td>
<td>-1</td>
<td>2</td>
</tr>
<tr>
<td>151</td>
<td>1</td>
<td>46</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>7.60</td>
<td>-1</td>
<td>242</td>
<td>3.3</td>
<td>50</td>
<td>2</td>
</tr>
<tr>
<td>152</td>
<td>0</td>
<td>44</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0.90</td>
<td>126</td>
<td>142</td>
<td>4.3</td>
<td>-1</td>
<td>2</td>
</tr>
<tr>
<td>153</td>
<td>0</td>
<td>61</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0.80</td>
<td>95</td>
<td>20</td>
<td>4.1</td>
<td>-1</td>
<td>2</td>
</tr>
<tr>
<td>154</td>
<td>0</td>
<td>53</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1.50</td>
<td>84</td>
<td>19</td>
<td>4.1</td>
<td>48</td>
<td>2</td>
</tr>
<tr>
<td>155</td>
<td>1</td>
<td>43</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1.20</td>
<td>100</td>
<td>19</td>
<td>3.1</td>
<td>42</td>
<td>2</td>
</tr>
</tbody>
</table>
Display 2: A histogram of 400 bootstrap replications of $R^*$ used in estimating the expected excess error of Gregory's rule for predicting death in chronic hepatitis. Values of $R^*$ range from -0.045 to 0.116 with mean 0.039, standard deviation 0.027 and quantiles $R^*_{.05} = -0.006$ and $R^*_{.95} = 0.084$. 

```
  0  10  20  30  40  50
-------------------|-------------------|-------------------|
-0.052 |
-0.045 |*
-0.039 |
-0.032 |***
-0.026 |*****
-0.019 |**
-0.013 |****
-0.006 |********
 0.000 |***********
 0.006 |*************
 0.013 |***************
 0.019 |***************
 0.026 |***************
 0.032 |***************
 0.039 |***************
 0.045 |***************
 0.052 |***************
 0.058 |***************
 0.064 |***************
 0.071 |***************
 0.077 |***************
 0.084 |***************
 0.090 |****
 0.097 |**
 0.103 |*
 0.110 |**
 0.116 |*
 0.129 |
-------------------|-------------------|-------------------|
```
Display 3: A binary-decision tree for predicting death in chronic hepatitis patients. Travel down the tree according to patient's covariates, choosing the right branch of a node if the covariate named at that node is large. Patients reaching a terminal node marked 1 are predicted to die. In the training sample, 15 patients were in terminal node T1, 2 who lived and 13 who died.
Display 4: The results of the first 20 experiments of Simulation 1.1. The columns labeled Cross-Validation, Jackknife, and Bootstrap are estimates of the expected excess. In each experiment, the estimate closest to the True Excess is starred, and the estimate closest to the expected excess error 0.0985 is daggered.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Apparent Error</th>
<th>True Excess</th>
<th>Cross-Validation</th>
<th>Jackknife</th>
<th>Bootstrap</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1500</td>
<td>0.0590</td>
<td>0.1500</td>
<td>0.1075</td>
<td>0.1060*†</td>
</tr>
<tr>
<td>2</td>
<td>0.2000</td>
<td>0.0580</td>
<td>0.0000</td>
<td>0.0075</td>
<td>0.0775*†</td>
</tr>
<tr>
<td>3</td>
<td>0.3000</td>
<td>-0.0850</td>
<td>0.1500</td>
<td>0.1325</td>
<td>0.1175*†</td>
</tr>
<tr>
<td>4</td>
<td>0.1000</td>
<td>0.1670</td>
<td>0.0000</td>
<td>0.0050</td>
<td>0.0505*†</td>
</tr>
<tr>
<td>5</td>
<td>0.1000</td>
<td>0.1390</td>
<td>0.0500*†</td>
<td>0.0500*†</td>
<td>0.0265</td>
</tr>
<tr>
<td>6</td>
<td>0.2000</td>
<td>0.1222</td>
<td>0.0500</td>
<td>0.0425</td>
<td>0.0835*†</td>
</tr>
<tr>
<td>7</td>
<td>0.2000</td>
<td>0.0514</td>
<td>0.2500</td>
<td>0.1725</td>
<td>0.0980*†</td>
</tr>
<tr>
<td>8</td>
<td>0.0000</td>
<td>0.2518</td>
<td>0.1500*†</td>
<td>0.1325</td>
<td>0.0495</td>
</tr>
<tr>
<td>9</td>
<td>0.1500</td>
<td>0.1326</td>
<td>0.0500</td>
<td>0.0450</td>
<td>0.0665*†</td>
</tr>
<tr>
<td>10</td>
<td>0.2500</td>
<td>-0.0044</td>
<td>0.2000</td>
<td>0.1925</td>
<td>0.1080*†</td>
</tr>
<tr>
<td>11</td>
<td>0.2000</td>
<td>0.0316</td>
<td>0.0000*</td>
<td>-0.0050</td>
<td>0.0705†</td>
</tr>
<tr>
<td>12</td>
<td>0.3500</td>
<td>-0.1020</td>
<td>0.2000</td>
<td>0.2225</td>
<td>0.1120*†</td>
</tr>
<tr>
<td>13</td>
<td>0.2000</td>
<td>0.0430</td>
<td>0.0500*</td>
<td>0.0575</td>
<td>0.0880†</td>
</tr>
<tr>
<td>14</td>
<td>0.1500</td>
<td>0.1112</td>
<td>0.0500</td>
<td>0.0375</td>
<td>0.0655*†</td>
</tr>
<tr>
<td>15</td>
<td>0.1500</td>
<td>0.2878</td>
<td>0.0000</td>
<td>0.0025</td>
<td>0.0535*†</td>
</tr>
<tr>
<td>16</td>
<td>0.1500</td>
<td>0.1152</td>
<td>0.1500*†</td>
<td>0.1575</td>
<td>0.0725</td>
</tr>
<tr>
<td>17</td>
<td>0.2000</td>
<td>0.1882</td>
<td>0.0500</td>
<td>0.0625</td>
<td>0.0700*†</td>
</tr>
<tr>
<td>18</td>
<td>0.2500</td>
<td>0.0178</td>
<td>0.0000</td>
<td>0.0050*</td>
<td>0.0705†</td>
</tr>
<tr>
<td>19</td>
<td>0.4500</td>
<td>-0.2068</td>
<td>0.1500</td>
<td>0.1600</td>
<td>0.1320*†</td>
</tr>
<tr>
<td>20</td>
<td>0.1500</td>
<td>0.0698</td>
<td>0.1000†</td>
<td>0.1150</td>
<td>0.0835*</td>
</tr>
</tbody>
</table>
Display 5: The results of 400 experiments of Simulation 1.1. $\text{MSE}_1$ is the mean square error about the true excess and $\text{MSE}_2$ is the mean square error about the expected excess error. The expected excess error is $E(\hat{r})$ for ideal. Rescaled root mean square errors take the root mean square error of ideal to be 0 and that of apparent to be 1.

<table>
<thead>
<tr>
<th>r</th>
<th>E(\hat{r})</th>
<th>SD(\hat{r})</th>
<th>$\sqrt{\text{MSE}_1(\hat{r})}$</th>
<th>rescaled</th>
<th>$\sqrt{\text{MSE}_2(\hat{r})}$</th>
<th>rescaled</th>
</tr>
</thead>
<tbody>
<tr>
<td>apparent</td>
<td>.0000</td>
<td>.0000</td>
<td>.1381</td>
<td>1.0000</td>
<td>.0985</td>
<td>1.0000</td>
</tr>
<tr>
<td>cross</td>
<td>.1028</td>
<td>.1111</td>
<td>.1473</td>
<td>1.2222</td>
<td>.1110</td>
<td>1.1269</td>
</tr>
<tr>
<td>jack</td>
<td>.0955</td>
<td>.0950</td>
<td>.1400</td>
<td>1.0459</td>
<td>.0949</td>
<td>.9635</td>
</tr>
<tr>
<td>boot</td>
<td>.0811</td>
<td>.0247</td>
<td>.1114</td>
<td>.3551</td>
<td>.0302</td>
<td>.3066</td>
</tr>
<tr>
<td>ideal</td>
<td>.0985</td>
<td>.0000</td>
<td>.0967</td>
<td>.0000</td>
<td>.0000</td>
<td>.0000</td>
</tr>
</tbody>
</table>
Display 6: The results of 400 experiments of Simulation 1.2.

<table>
<thead>
<tr>
<th>$\hat{r}$</th>
<th>E($\hat{r}$)</th>
<th>SD($\hat{r}$)</th>
<th>$\sqrt{\text{MSE}_1(\hat{r})}$</th>
<th>rescaled</th>
<th>$\sqrt{\text{MSE}_2(\hat{r})}$</th>
<th>rescaled</th>
</tr>
</thead>
<tbody>
<tr>
<td>apparent</td>
<td>.0000</td>
<td>.0000</td>
<td>.0829</td>
<td>1.0000</td>
<td>.0477</td>
<td>1.0000</td>
</tr>
<tr>
<td>cross</td>
<td>.0509</td>
<td>.0498</td>
<td>.0828</td>
<td>.9934</td>
<td>.0499</td>
<td>1.0461</td>
</tr>
<tr>
<td>jack</td>
<td>.0493</td>
<td>.0437</td>
<td>.0827</td>
<td>.9868</td>
<td>.0437</td>
<td>.9161</td>
</tr>
<tr>
<td>boot</td>
<td>.0480</td>
<td>.0142</td>
<td>.0752</td>
<td>.4901</td>
<td>.0141</td>
<td>.2956</td>
</tr>
<tr>
<td>ideal</td>
<td>.0477</td>
<td>.0000</td>
<td>.0678</td>
<td>.0000</td>
<td>.0000</td>
<td>.0000</td>
</tr>
</tbody>
</table>
Display 7: The results of 400 experiments of Simulation 1.3.

<table>
<thead>
<tr>
<th>$\hat{r}$</th>
<th>$E(\hat{r})$</th>
<th>$SD(\hat{r})$</th>
<th>$\sqrt{MSE_1(\hat{r})}$</th>
<th>rescaled</th>
<th>$\sqrt{MSE_2(\hat{r})}$</th>
<th>rescaled</th>
</tr>
</thead>
<tbody>
<tr>
<td>apparent</td>
<td>.0000</td>
<td>.0000</td>
<td>.0608</td>
<td>1.0000</td>
<td>.0300</td>
<td>1.0000</td>
</tr>
<tr>
<td>cross</td>
<td>.0318</td>
<td>.0322</td>
<td>.0611</td>
<td>1.0375</td>
<td>.0322</td>
<td>1.0733</td>
</tr>
<tr>
<td>jack</td>
<td>.0310</td>
<td>.0285</td>
<td>.0609</td>
<td>1.0125</td>
<td>.0285</td>
<td>.9500</td>
</tr>
<tr>
<td>boot</td>
<td>.0328</td>
<td>.0092</td>
<td>.0567</td>
<td>.4875</td>
<td>.0096</td>
<td>.3200</td>
</tr>
<tr>
<td>ideal</td>
<td>.0300</td>
<td>.0000</td>
<td>.0528</td>
<td>.0000</td>
<td>.0000</td>
<td>.0000</td>
</tr>
</tbody>
</table>

63
Display 8: The results of 400 experiments of Simulation 2.1.

<table>
<thead>
<tr>
<th>( \hat{r} )</th>
<th>( E(\hat{r}) )</th>
<th>( SD(\hat{r}) )</th>
<th>( \sqrt{\text{MSE}_1(\hat{r})} )</th>
<th>rescaled</th>
<th>( \sqrt{\text{MSE}_2(\hat{r})} )</th>
<th>rescaled</th>
</tr>
</thead>
<tbody>
<tr>
<td>apparent</td>
<td>.0000</td>
<td>.0000</td>
<td>.1768</td>
<td>1.0000</td>
<td>.1465</td>
<td>1.0000</td>
</tr>
<tr>
<td>cross</td>
<td>.1584</td>
<td>.1382</td>
<td>.1749</td>
<td>.9756</td>
<td>.1386</td>
<td>.9461</td>
</tr>
<tr>
<td>jack</td>
<td>.1458</td>
<td>.1182</td>
<td>.1643</td>
<td>.8395</td>
<td>.1180</td>
<td>.8055</td>
</tr>
<tr>
<td>boot</td>
<td>.0989</td>
<td>.0263</td>
<td>.1237</td>
<td>.3184</td>
<td>.0544</td>
<td>.3713</td>
</tr>
<tr>
<td>ideal</td>
<td>.1465</td>
<td>.0000</td>
<td>.0989</td>
<td>.0000</td>
<td>.0000</td>
<td>.0000</td>
</tr>
</tbody>
</table>
Display 9: The results of 400 experiments of Simulation 2.2.

<table>
<thead>
<tr>
<th>$\hat{r}$</th>
<th>$E(\hat{r})$</th>
<th>$SD(\hat{r})$</th>
<th>$\sqrt{MSE_1(\hat{r})}$</th>
<th>rescaled</th>
<th>$\sqrt{MSE_2(\hat{r})}$</th>
<th>rescaled</th>
</tr>
</thead>
<tbody>
<tr>
<td>apparent</td>
<td>.0000</td>
<td>.0000</td>
<td>.1036</td>
<td>1.0000</td>
<td>.0797</td>
<td>1.0000</td>
</tr>
<tr>
<td>cross</td>
<td>.0868</td>
<td>.0676</td>
<td>.0984</td>
<td>.8606</td>
<td>.0679</td>
<td>.8519</td>
</tr>
<tr>
<td>jack</td>
<td>.0850</td>
<td>.0618</td>
<td>.0978</td>
<td>.8445</td>
<td>.0619</td>
<td>.7767</td>
</tr>
<tr>
<td>boot</td>
<td>.0686</td>
<td>.0148</td>
<td>.0751</td>
<td>.2359</td>
<td>.0185</td>
<td>.2321</td>
</tr>
<tr>
<td>ideal</td>
<td>.0797</td>
<td>.0000</td>
<td>.0663</td>
<td>.0000</td>
<td>.0000</td>
<td>.0000</td>
</tr>
</tbody>
</table>
Display 10: The results of 400 experiments of Simulation 2.3.

<table>
<thead>
<tr>
<th>$\hat{r}$</th>
<th>$E(\hat{r})$</th>
<th>$SD(\hat{r})$</th>
<th>$\sqrt{MSE_1(\hat{r})}$</th>
<th>rescaled</th>
<th>$\sqrt{MSE_2(\hat{r})}$</th>
<th>rescaled</th>
</tr>
</thead>
<tbody>
<tr>
<td>apparent</td>
<td>.0000</td>
<td>.0000</td>
<td>.0784</td>
<td>1.0000</td>
<td>.0588</td>
<td>1.0000</td>
</tr>
<tr>
<td>cross</td>
<td>.0570</td>
<td>.0444</td>
<td>.0679</td>
<td>.6038</td>
<td>.0443</td>
<td>.7534</td>
</tr>
<tr>
<td>jack</td>
<td>.0572</td>
<td>.0405</td>
<td>.0679</td>
<td>.6038</td>
<td>.0405</td>
<td>.6888</td>
</tr>
<tr>
<td>boot</td>
<td>.0513</td>
<td>.0103</td>
<td>.0579</td>
<td>.2264</td>
<td>.0127</td>
<td>.2160</td>
</tr>
<tr>
<td>ideal</td>
<td>.0588</td>
<td>.0000</td>
<td>.0519</td>
<td>.0000</td>
<td>.0000</td>
<td>.0000</td>
</tr>
</tbody>
</table>
Display 11: 95 percent (non-simultaneous) confidence intervals for rescaled $\sqrt{\text{MSE}_1}$ of Apparent, Cross-Validation, Jackknife, Bootstrap, and Ideal estimates of the excess error.
Display 12: 95 percent (non-simultaneous) confidence intervals for rescaled $\sqrt{\text{MSE}_2}$ of Apparent, Cross-Validation, Jackknife, Bootstrap, and Ideal estimates of the expected excess error.
Display 13: Scatterplots of $\hat{r}_{\text{cross}}$ vs. $\hat{r}_{\text{jack}}$ and $\hat{r}_{\text{cross}}$ vs. $\hat{r}_{\text{boot}}$ for the first 100 experiments of Simulation 1.1. (A = 1 observation, B = 2 observations, etc.)
Display 14: Scatterplots of $\hat{r}_{\text{cross}}$ vs. $\hat{r}_{\text{jack}}$ and $\hat{r}_{\text{cross}}$ vs. $\hat{r}_{\text{boot}}$ for the first 100 experiments of Simulation 1.2. (A = 1 observation, B = 2 observations, etc.)
Display 15: Scatterplots of \( \hat{r}_{\text{cross}} \) and \( \hat{r}_{\text{jack}} \) and \( \hat{r}_{\text{cross}} \) vs. \( \hat{r}_{\text{boot}} \) for the first 100 experiments of Simulation 1.3. (A = 1 observation, B = 2 observations, etc.).
Display 16: Correlations to verify the asymptotic relationship between the cross-validation and jackknife estimates of expected excess error.

<table>
<thead>
<tr>
<th>Simulation</th>
<th>$\rho(r_{\text{cross}}, r_{\text{jack}})$</th>
<th>$\rho(r_{\text{cross}}, r_{\text{boot}})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>.969</td>
<td>.563</td>
</tr>
<tr>
<td>1.2</td>
<td>.963</td>
<td>.430</td>
</tr>
<tr>
<td>1.3</td>
<td>.967</td>
<td>.314</td>
</tr>
<tr>
<td>2.1</td>
<td>.969</td>
<td>.524</td>
</tr>
<tr>
<td>2.2</td>
<td>.965</td>
<td>.476</td>
</tr>
<tr>
<td>2.2</td>
<td>.972</td>
<td>.299</td>
</tr>
</tbody>
</table>
Display 17: A histogram of 400 bootstrap replications of $\lambda^*(t_0)$ where $t_0$ is the covariates of Patient 145. Values of $\lambda^*$ range from -3.873 to 12.332 with mean 1.774, standard deviation 1.478 and quantiles $\lambda^*(.05) = -0.516$ and $\lambda^*(.95) = 3.964$. 
**Display 18:** Variables chosen by Gregory's rule in the first 50 bootstrap samples of the chronic hepatitis data.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>14</th>
<th>17</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>17</td>
<td>11</td>
<td>19</td>
<td>2</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>17</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>11</td>
<td>17</td>
<td>1</td>
<td>2</td>
<td>14</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>12</td>
<td>15</td>
<td>1</td>
<td>14</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>14</td>
<td>12</td>
<td>19</td>
<td>17</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>14</td>
<td>17</td>
<td>1</td>
<td>11</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>17</td>
<td>13</td>
<td>6</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>12</td>
<td>18</td>
<td>11</td>
<td>8</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>17</td>
<td>13</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>12</td>
<td>17</td>
<td>11</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>17</td>
<td>14</td>
<td>11</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>14</td>
<td>12</td>
<td>11</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>12</td>
<td>11</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>14</td>
<td>17</td>
<td>5</td>
<td>10</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>14</td>
<td>17</td>
<td>19</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>12</td>
<td>5</td>
<td>18</td>
<td>19</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>12</td>
<td>19</td>
<td>1</td>
<td>14</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>12</td>
<td>13</td>
<td>1</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>17</td>
<td>11</td>
<td>14</td>
<td>18</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>21</td>
<td>17</td>
<td>11</td>
<td>2</td>
<td>1</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>22</td>
<td>14</td>
<td>17</td>
<td>6</td>
<td>18</td>
<td>19</td>
<td>7</td>
</tr>
<tr>
<td>23</td>
<td>12</td>
<td>14</td>
<td>11</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>14</td>
<td>17</td>
<td>2</td>
<td>1</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>11</td>
<td>17</td>
<td>18</td>
<td>5</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>14</td>
<td>18</td>
<td>1</td>
<td>11</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>12</td>
<td>11</td>
<td>17</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>12</td>
<td>18</td>
<td>6</td>
<td>19</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>17</td>
<td>18</td>
<td>15</td>
<td>11</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>30</td>
<td>17</td>
<td>6</td>
<td>2</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>12</td>
<td>11</td>
<td>18</td>
<td>19</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>14</td>
<td>12</td>
<td>10</td>
<td>2</td>
<td>17</td>
<td>5</td>
</tr>
<tr>
<td>33</td>
<td>13</td>
<td>1</td>
<td>6</td>
<td>2</td>
<td>11</td>
<td>5</td>
</tr>
<tr>
<td>34</td>
<td>12</td>
<td>14</td>
<td>3</td>
<td>1</td>
<td>18</td>
<td>2</td>
</tr>
<tr>
<td>35</td>
<td>14</td>
<td>11</td>
<td>12</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>17</td>
<td>11</td>
<td>18</td>
<td>2</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>37</td>
<td>14</td>
<td>13</td>
<td>3</td>
<td>1</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>14</td>
<td>18</td>
<td>6</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>12</td>
<td>11</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>12</td>
<td>11</td>
<td>14</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>11</td>
<td>17</td>
<td>2</td>
<td>1</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>17</td>
<td>11</td>
<td>1</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>12</td>
<td>11</td>
<td>14</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>11</td>
<td>17</td>
<td>18</td>
<td>2</td>
<td>1</td>
<td>14</td>
</tr>
<tr>
<td>45</td>
<td>14</td>
<td>17</td>
<td>6</td>
<td>10</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>46</td>
<td>11</td>
<td>12</td>
<td>1</td>
<td>2</td>
<td>18</td>
<td>6</td>
</tr>
<tr>
<td>47</td>
<td>12</td>
<td>14</td>
<td>6</td>
<td>18</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>48</td>
<td>14</td>
<td>1</td>
<td>11</td>
<td>2</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>49</td>
<td>12</td>
<td>14</td>
<td>3</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>14</td>
<td>11</td>
<td>17</td>
<td>3</td>
<td>12</td>
<td></td>
</tr>
</tbody>
</table>

74
APPENDIX

Proofs of Lemmas in Chapter IV.

Proof of Lemma 1.

We just take derivatives. Here only, let \( q_{\alpha \gamma} = t_{\alpha} A_p^{-1} t_{\gamma} \) and \( d_{\alpha} = y_{\alpha} - \eta_p(t_{\alpha}) \). Differentiating \( A_p = \Sigma t_{\alpha} t_{\alpha} p_{\alpha} \) and \( B_p = \Sigma t_{\alpha} y_{\alpha} p_{\alpha} \),

\[
\frac{\partial A_p}{\partial p_{\alpha}} = t_{\alpha} t_{\alpha},
\]

\[
\frac{\partial B_p}{\partial p_{\alpha}} = t_{\alpha} y_{\alpha}.
\]

Differentiating both sides of \( I = A_p^{-1} A_p \) gives

\[
0 = \frac{\partial A_p^{-1}}{\partial p_{\alpha}} A_p + A_p^{-1} \frac{\partial A_p}{\partial p_{\alpha}},
\]

and solving for \( \frac{\partial A_p^{-1}}{\partial p_{\alpha}} \),

\[
\frac{\partial A_p^{-1}}{\partial p_{\alpha}} = -A_p^{-1} \frac{\partial A_p}{\partial p_{\alpha}} A_p^{-1},
\]

\[
= -A_p^{-1} t_{\alpha} t_{\alpha} A_p^{-1}.
\]

Differentiating \( \eta_p(t_{i}) = t_{i} A_p^{-1} B_p \), \( d_{\gamma} = y_{\gamma} - \eta_p(t_{\gamma}) \), and \( q_{i\gamma} = t_{i} A_p^{-1} t_{\gamma} \),
\[
\frac{\partial \eta_p(t_i)}{\partial P_\alpha} = t_i \left( A_p^{-1} \frac{\partial P}{\partial P_\alpha} B_p + A_p^{-1} \frac{\partial B_p}{\partial P_\alpha} \right),
\]

\[
= t_i \left( A_p^{-1} t_\alpha A_p^{-1} B_p + A_p^{-1} t_\alpha y_\alpha \right),
\]

\[
= t_i A_p^{-1} t_\alpha (y_\alpha - t_\alpha A_p^{-1} B_p),
\]

\[
= t_i A_p^{-1} t_\alpha (y_\alpha - \eta_p(t_\alpha))
= q_{i\alpha} d_\alpha,
\]

\[
\frac{\partial d_\gamma}{\partial P_\alpha} = - \frac{\partial \eta_p(t_\gamma)}{\partial P_\alpha},
\]

\[
= -q_{j\alpha} d_\alpha,
\]

\[
\frac{\partial q_{i\gamma}}{\partial P_\alpha} = t_i A_p^{-1} \frac{\partial A_p}{\partial P_\alpha},
\]

\[
= -t_i A_p^{-1} t_\alpha A_p^{-1} t_\gamma,
\]

\[
= -q_{i\alpha} q_{\alpha\gamma}.
\]

Finally, differentiating \( s^i(P) = (y_i - \eta_p(t_i))^2 = d_i^2 \) twice gives

\[
\frac{\partial s^i(P)}{\partial P_\alpha} = -2d_i \frac{\partial d_i}{\partial P_\alpha},
\]

\[
= -2q_{i\alpha} d_i d_\alpha,
\]

76
\[
\frac{\partial^2 s_i^i(P)}{\partial p^\alpha_\alpha \partial p^\gamma_\gamma} = \frac{\partial}{\partial p^\alpha_\alpha} \left( -2 q_{i\gamma} d_{i\gamma} \right),
\]

\[
= 2(q_{i\alpha} q_{\alpha\gamma} d_{i\gamma} + q_{i\gamma} q_{i\alpha} d_{\alpha\gamma} + q_{i\gamma} q_{\alpha\gamma} d_{i\alpha}).
\]

**Detailed proof of Lemma 2.**

To see (4.6), (drop subscript i in \( \theta_i \))

\[
p^\dagger = \theta p^0 + (1-\theta)p^{(j)},
\]

\[
= \frac{\theta}{n} \frac{1}{1} + \frac{(1-\theta)}{n-1} (1 - e_j),
\]

\[
= \left( \frac{\theta}{n} + \frac{(1-\theta)}{n-1} \right) - \frac{1-\theta}{n-1} e_j,
\]

\[
= \frac{1}{n-1} \left( 1 - \frac{\theta}{n} \right) - \frac{1}{n-1} (1-\theta) e_j.
\]

To see (4.7) and (4.8),

\[
A_{p^\dagger} = \Sigma t^\dagger_\alpha t_\alpha a_\alpha,
\]

\[
= \Sigma t^\dagger_\alpha t_\alpha (g_{ni} \frac{1}{n} - h_{ni} e_j)_\alpha,
\]

\[
= g_{ni} \Sigma t^\dagger_\alpha t_\alpha - h_{ni} t^\dagger_{j} t_j,
\]

\[
= ng_{ni}(A_{p^0} - \frac{r_{ni}}{n} t^\dagger_{j} t_j),
\]

\[
B_{p^\dagger} = \Sigma t^\dagger_\gamma a_\gamma a_\alpha,
\]

\[
= ng_{ni}(B_{p^0} - \frac{r_{ni}}{n} t^\dagger_{j} y_j).
\]
To see that $r_{ni} = \frac{h_{ni}}{g_{ni}} = \frac{1 - \theta_i}{1 - \theta_i/n} \in [0,1]$, notice that $f(\theta) = \frac{1 - \theta}{1 - \theta/n}$ is a decreasing continuous function on $\theta \in [0,1]$ and takes values in $[0,1]$. To see (4.9),

$$(A_p - \frac{r_{ni}}{n} t_j t_j')^{-1} = A_p^{-1} + \frac{r_{ni}}{n} A_p^{-1} t_j t_j A_p^{-1} \frac{1 - \frac{r_{ni}}{n} t_j A_p^{-1} t_j}{1 - \frac{r_{ni}}{n} t_j A_p^{-1} t_j} ,$$

$$= A_p^{-1} + R_{nij} A_p^{-1} t_j t_j A_p^{-1} .$$

To get (4.5),

$$q_{\alpha Y}^+ = t_{\alpha} A_p^{-1} t_{\gamma}$$

$$= \frac{1}{ng_{ni}} A_p^{-1} t_{\alpha} (A_p^{-1} + R_{nij} A_p^{-1} t_j A_p^{-1}) t_{\gamma}$$

$$= \frac{1}{ng_{ni}} q_{\alpha Y}^0 + R_{nij} q_{\alpha j}^0 q_{j Y}^0 .$$
\[ \eta_{p}^{\dagger}(t_{\alpha}) = t_{\alpha} A_{p}^{-1} B_{p}^{\dagger}, \]

\[ = t_{\alpha} \left( A_{p0}^{-1} + \frac{R_{nij}}{n} A_{p0}^{-1} t_{j} A_{p0}^{-1} \right) \left( B_{p0} - \frac{r_{ni}}{n} t_{j} y_{j} \right), \]

\[ = t_{\alpha} \left[ A_{p0}^{-1} B_{p0} - \frac{r_{ni}}{n} A_{p0}^{-1} t_{j} y_{j} + R_{nij} A_{p0}^{-1} t_{j} A_{p0}^{-1} \left( B_{p0} - \frac{r_{ni}}{n} t_{j} y_{j} \right) \right], \]

\[ = t_{\alpha} A_{p0}^{-1} B_{p0}, \]

\[ + \frac{t_{\alpha}}{1 - \frac{r_{ni}}{n} q_{jj}^{0}} \left[ -(1 - \frac{r_{ni}}{n} q_{jj}^{0}) (\frac{r_{ni}}{n} A_{p0}^{-1} t_{j} y_{j}) \right. \]

\[ + \left. \left( \frac{r_{ni}}{n} A_{p0}^{-1} t_{j} A_{p0}^{-1} \right) \left( B_{p0} - \frac{r_{ni}}{n} t_{j} y_{j} \right) \right], \]

\[ = t_{\alpha} A_{p0}^{-1} B_{p0} + \frac{t_{\alpha}}{1 - \frac{r_{ni}}{n} q_{jj}^{0}} \left[ -\frac{r_{ni}}{n} A_{p0}^{-1} t_{j} y_{j} + \frac{r_{ni}}{n} A_{p0}^{-1} t_{j} A_{p0}^{-1} B_{p0} \right], \]

\[ = t_{\alpha} A_{p0}^{-1} B_{p0} - R_{nij} t_{\alpha} A_{p0}^{-1} t_{j} y_{j} - t_{j} A_{p0}^{-1} B_{p0}, \]

\[ = \eta_{p0}(t_{\alpha}) - R_{nij} q_{\alpha j}^{0} d_{j}^{0}, \]

\[ d_{\alpha}^{\dagger} = y_{\alpha} - \eta_{p}^{\dagger}(t_{\alpha}), \]

\[ = y_{\alpha} - \eta_{p0}(t_{\alpha}) + R_{nij} q_{\alpha j}^{0} d_{j}^{0}, \]

\[ = d_{\alpha}^{0} + R_{nij} q_{\alpha j}^{0} d_{j}^{0}. \]

**Proof of Lemma 5.**

From Lemma 4,
\[ |d^\dagger_\alpha| \leq \delta_\alpha + \tau_\alpha \mu_j, \]
\[ |q^\dagger_{\alpha \gamma}| \leq \tau_\alpha \tau_\gamma \nu_j. \]

Therefore,

\[ (q^\dagger_{ij} d^\dagger_j)^2 \leq \tau^2_i \tau^2_j \nu_j^2 (\delta_j + \tau_j \mu_j)^2, \]
\[ = \tau^2_i \tau^2_j (\delta_j^2 + 2 \delta_j \tau_j \mu_j + \tau_j^2 \mu_j^2) \nu_j^2 \]
\[ = (\tau^2_i) (\tau^2_j \delta_j^2) \nu_j^2 + 2 (\tau^2_i) (\tau^3_j \delta_j) \nu_j^2 + (\tau^2_i) (\tau^4_j) \mu_j^2 \nu_j^2, \]
\[ |q^\dagger_{ij} q^\dagger_{jj} d^\dagger_i d^\dagger_j| \leq \tau_i \tau_j \nu_j^2 (\delta_i + \tau_1 \mu_j) (\delta_j + \tau_1 \mu_j), \]
\[ = \tau_i \tau_j (\delta_i \delta_j + \delta_i \tau_j \mu_j + \tau_1 \delta_j \mu_j + \tau_1 \tau_j \mu_j^2) \nu_j^2, \]
\[ = (\tau_1 \delta_i) (\tau^3_j \delta_j) \nu_j^2 + (\tau_1 \delta_i) (\tau^4_j) \mu_j \nu_j^2 \]
\[ + (\tau^2_i) (\tau^3_j \delta_j) \nu_j^2 + (\tau^2_i) \mu_j \nu_j^2, \]
\[ |q^\dagger_{i\alpha} q^\dagger_{ij} d^\dagger_\alpha d^\dagger_j| \leq \tau^2_i \tau_\alpha \tau_j \nu_j^2 (\delta_\alpha + \tau_\alpha \mu_j) (\delta_j + \tau_\alpha \mu_j), \]
\[ = \tau^2_i \tau_\alpha \tau_j (\delta_\alpha \delta_j + \delta_\alpha \tau_j \mu_j + \tau_\alpha \delta_j \mu_j + \tau_\alpha \tau_j \mu_j^2) \nu_j^2, \]
\[ = (\tau^2_i) (\tau_\alpha \delta_\alpha) (\tau_j \delta_j) \nu_j^2 + (\tau^2_i) (\tau_\alpha \delta_\alpha) (\tau_j \mu_j) \nu_j^2 \]
\[ + (\tau^2_i) (\tau^2_\alpha) (\tau_j \delta_j) \nu_j^2 + (\tau^2_i) (\tau^2_\alpha) \mu_j \nu_j^2, \]
\[ |q_j q_j a_\alpha a_\alpha^\dagger a_\alpha^\dagger| \leq \tau_i \tau_i \tau_j \nu_i^2 (\delta_i + \tau_i \mu_j)(\delta_\alpha + \tau_\alpha \mu_j), \]

\[ = \tau_i \tau_i (\delta_i + \tau_i \mu_j)(\delta_\alpha + \tau_\alpha \mu_j) \tau_j \nu_j^2, \]

\[ = \tau_i \tau_\alpha (\delta_i \delta_\alpha + \delta_i \tau_\alpha \mu_j + \tau_i \delta_\alpha \mu_j + \tau_i \tau_\alpha \mu_j^2) \tau_j \nu_j^2, \]

\[ = (\tau_i \delta_i) (\tau_\alpha \delta_\alpha) (\tau_j \nu_j^2) + (\tau_i \delta_i) (\tau_i \tau_\alpha \mu_j \nu_j^2), \]

\[ + (\tau_i \delta_i) (\tau_\alpha \delta_\alpha) (\tau_j \nu_j^2) + (\tau_i \delta_i) (\tau_i \tau_\alpha \mu_j \nu_j^2), \]

\[ |q_j q_j a_\alpha a_\alpha^\dagger a_\alpha^\dagger| \leq \tau_i \tau_i \tau_j \nu_j^2 (\delta_\alpha + \tau_\alpha \mu_j)(\delta_\gamma + \tau_\gamma \mu_j), \]

\[ = \tau_i \tau_i \tau_\gamma (\delta_\alpha \delta_\gamma + \delta_\alpha \tau_\gamma \mu_j + \tau_\alpha \delta_\gamma \mu_j + \tau_\alpha \tau_\gamma \mu_j^2) \nu_j^2, \]

\[ = \tau_i \tau_i (\tau_\alpha \delta_\alpha) (\tau_\gamma \delta_\gamma) \nu_j^2 + (\tau_i \tau_i) (\tau_\alpha \delta_\alpha) (\tau_\gamma \nu_j^2), \]

\[ + \tau_i \tau_i (\tau_\alpha \delta_\alpha) (\tau_\gamma \delta_\gamma) \nu_j^2 + (\tau_i \tau_i) (\tau_\alpha \nu_j^2). \]
REFERENCES


Efron, B. (1982a). The jackknife, the bootstrap, and other resampling plans. To appear in SIAM.


