

## **Abstract**

WU, YUJUN. Controlling variable selection by the addition of pseudo-variables. (Under the direction of Dr. Dennis D. Boos and Dr. Leonard A. Stefanski)

Many variable selection procedures have been developed in the literature for linear regression models. We propose a new and general approach, the False Selection Rate (FSR) method, to control variable selection with the advantage of being applicable to a broader class of regression models; for example, binary regression, Poisson regression, etc. By adding a number of pseudo-variables to the real set of data and monitoring the proportion of pseudo-variables falsely selected in the model, we are able to control the model false selection rate, selecting as many important variables as possible while selecting a relatively low proportion of false important ones. We focus on forward selection because it is applicable in the case where there are more variables than observations. Due to the difficulty of obtaining analytical results, we study our approach by Monte Carlo and compare it with a variety of commonly used procedures. We first focus on linear regression models, and then extend the approach to logistic regression models. The new method is illustrated on four real data sets.

# Controlling Variable Selection By the Addition of Pseudo-Variables

by

**Yujun Wu**

A Dissertation

submitted to the advisory committee on graduate studies of

North Carolina State University

in partial fulfillment of the requirements

for the Degree of Doctor of Philosophy

**DEPARTMENT OF STATISTICS**

Raleigh, NC

August, 2004

**APPROVED BY:**

---

Dennis D. Boos  
Co-Chair of Advisory Committee

---

Leonard A. Stefanski  
Co-Chair of Advisory Committee

---

Marc G. Genton

---

Hao Helen Zhang

*To Lin and my parents*

## Biography

Yujun Wu was born in Xiushan, Chongqing, China on April 5, 1974. He attended Peking University in Beijing, China in 1992, where he earned a Bachelor of Science degree in Probability and Statistics in 1997. He then worked as a software engineer in Institute of Computer Science & Technology, Peking University, until 2000. Thereafter, he came to North Carolina State University and received a Master of Science degree in statistics in May, 2002. He continued his studies towards a Ph.D degree in statistics at North Carolina State University.

## Acknowledgements

I would like to express my deepest gratitude to both of my advisors: Dr. Dennis D. Boos and Dr. Leonard A. Stefanski. They have been advising me throughout this work and providing valuable ideas, insightful comments, and continuous encouragement. Without their direction, this work could not have become a reality. It was a privilege and a pleasure to be their student, and this valuable experience will definitely benefit my future researches.

Thanks to Dr. Marc G. Genton and Dr. Hao Helen Zhang for being on my Ph.D committee, and for the time they dedicated to reviewing my thesis. Their helpful suggestions have made this dissertation much better. Thanks to Dr. Zhao-Bang Zeng for being on my Ph.D prelim oral exam committee and providing me with data. It was regretful that he was not able to attend my final oral. Thanks also to Terry Byron who kindly assisted me in using the department cluster servers for my computation and make my work easy. I would also thank Dr. John Bishir for his kindly agreement to represent the graduate school on my committee.

Special thanks to my wife and my parents. Their love and support have carried me to where I am today.

# Contents

<b>List of Tables</b>	<b>vii</b>
<b>List of Figures</b>	<b>x</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Interpretation and Prediction . . . . .	2
1.2 Algorithms for Selecting Subsets . . . . .	3
1.3 Selection Bias . . . . .	6
<b>2 Selection Criteria</b>	<b>9</b>
2.1 Prediction Criteria . . . . .	10
2.2 Data Resampling Selection . . . . .	12
2.3 Information-based Criteria . . . . .	15
2.4 Bayesian Variable Selection . . . . .	17
<b>3 Controlling Variable Selection</b>	<b>19</b>
3.1 Motivation . . . . .	19
3.2 False Selection Rate Variable Selection . . . . .	22
3.3 Pseudo-Variables Generation . . . . .	34
3.4 Some Comments . . . . .	36
<b>4 Simulation – In Linear Regression Models</b>	<b>40</b>
4.1 Simulation Design . . . . .	40
4.2 Factors Affecting the FSR Procedure . . . . .	43
4.2.1 Methods for Generating Pseudo-Variables . . . . .	43
4.2.2 The Choice of $k_p$ . . . . .	47
4.2.3 The Effect of Target False Selection Rate $\gamma$ . . . . .	47
4.2.4 The Monte Carlo Estimated False Selection Rate . . . . .	49
4.2.5 The Performance of $\hat{\alpha}_0$ . . . . .	49
4.3 Comparison to Other Selection Criteria . . . . .	53
4.4 Comparison to the Lasso . . . . .	60
<b>5 Application to Logistic Regression Models</b>	<b>65</b>
5.1 Forward Selection In Logistic Regression . . . . .	66
5.2 Simulation Study . . . . .	67
5.2.1 Simulation Design . . . . .	67
5.2.2 Simulation Results . . . . .	69

<b>6</b>	<b>Applications</b>	<b>76</b>
6.1	Wing Shape Data . . . . .	76
6.2	Diabetes Data . . . . .	79
6.3	The TPLO Study . . . . .	82
6.4	The UMARU IMPACT Study . . . . .	84
<b>7</b>	<b>Discussion and Conclusion</b>	<b>88</b>
<b>A</b>	<b>Tables of Simulation Results</b>	<b>91</b>
A.1	Results for Linear Regression Models . . . . .	91
A.2	Results for Logistic Regression Models . . . . .	91
<b>B</b>	<b>Derivation of <math>\beta_0</math> in Logistic Regression Simulation Design</b>	<b>109</b>
<b>C</b>	<b>Estimating Model Size</b>	<b>111</b>
C.1	Based on Student's $t$ Test . . . . .	112
C.2	"Capture-Recapture" Method . . . . .	113
C.3	Least Squares Method . . . . .	114
C.4	Isotonic Method . . . . .	114
	<b>Bibliography</b>	<b>115</b>

# List of Tables

3.1	Possible variables in the selected subset . . . . .	22
4.1	Average model error for the FSR method with $\gamma = 0.05$ for the four ways of generating pseudo-variables, ( $n = 150$ , $k_T = 21$ , and $k_P = 21$ ) . . . . .	45
4.2	Average model size for the FSR method with $\gamma = 0.05$ for the four ways of generating pseudo-variables, ( $n = 150$ , $k_T = 21$ , and $k_P = 21$ ) . . . . .	45
4.3	Average model error for the FSR method with $\gamma = 0.05$ for the four ways of generating pseudo-variables, ( $n = 50$ , $k_T = 21$ , and $k_P = 21$ ) . . . . .	46
4.4	Average model size for the FSR method with $\gamma = 0.05$ for the four ways of generating pseudo-variables, ( $n = 50$ , $k_T = 21$ , and $k_P = 21$ ) . . . . .	46
4.5	Average model error for the FSR method with $\gamma = 0.05$ when choosing different $k_P$ , ( $n = 150$ and $k_T = 21$ ) . . . . .	47
4.6	Average model size for the FSR method with $\gamma = 0.05$ when choosing different $k_P$ , ( $n = 150$ and $k_T = 21$ ) . . . . .	48
6.1	Description of markers for the wing shape data . . . . .	77
6.2	Description of variables in the diabetes data . . . . .	79
6.3	Variables selected by the FSR method and LARS for the diabetes data . . . . .	81
6.4	Description of variables in the TPLO study . . . . .	83
6.5	Description of variables in the UMARU IMPACT study . . . . .	85
A.1a	Average model error for Best, FSR, Tune, $C_p$ , and LB for linear regression ( $n = 50$ and $k_T = 21$ ) . . . . .	92
A.1b	Average model size for Best, FSR, Tune, $C_p$ , and LB for linear regression ( $n = 50$ and $k_T = 21$ ) . . . . .	92
A.2a	Average model error for Best, FSR, Tune, $C_p$ , and LB for linear regression ( $n = 150$ and $k_T = 21$ ) . . . . .	93
A.2b	Average model size for Best, FSR, Tune, $C_p$ , and LB for linear regression ( $n = 150$ and $k_T = 21$ ) . . . . .	93
A.3a	Average model error for Best, FSR, Tune, $C_p$ , and LB for linear regression ( $n = 500$ and $k_T = 21$ ) . . . . .	94
A.3b	Average model size for Best, FSR, Tune, $C_p$ , and LB for linear regression ( $n = 500$ and $k_T = 21$ ) . . . . .	94
A.4a	Average model error for Best, FSR, Tune, $C_p$ , and LB for linear regression ( $n = 150$ and $k_T = 42$ ) . . . . .	95
A.4b	Average model size for Best, FSR, Tune, $C_p$ , and LB for linear regression ( $n = 150$ and $k_T = 42$ ) . . . . .	95
A.5a	Average model error for FSR and Lasso for linear regression ( $n = 150$ , $k_T = 21$ and $\epsilon \sim N(0, 1)$ ) . . . . .	96



A.5b Average model size for FSR and Lasso for linear regression ( $n = 150$ , $k_T = 21$ , and $\epsilon \sim N(0, 1)$ ) . . . . .	96
A.5c The Monte Carlo estimated $\gamma$ for FSR and Lasso for linear regression ( $n = 150$ , $k_T = 21$ , and $\epsilon \sim N(0, 1)$ ) . . . . .	96
A.6a Average model error for FSR and Lasso for linear regression ( $n = 150$ , $k_T = 21$ , and $\epsilon \sim \exp(1) - 1$ ) . . . . .	97
A.6b Average model size for FSR and Lasso for linear regression ( $n = 150$ , $k_T = 21$ , and $\epsilon \sim \exp(1) - 1$ ) . . . . .	97
A.6c The Monte Carlo estimated $\gamma$ for FSR and Lasso for linear regression ( $n = 150$ , $k_T = 21$ , and $\epsilon \sim \exp(1) - 1$ ) . . . . .	97
A.7a Average SME for FSR, AIC, and BIC for logistic regression ( $n = 150$ , $P = 0.1$ , $R^2 = 0.75$ ) . . . . .	98
A.7b Average LE for FSR, AIC, and BIC for logistic regression ( $n = 150$ , $P = 0.1$ , $R^2 = 0.75$ ) . . . . .	98
A.7c Average model size for FSR, AIC, and BIC for logistic regression ( $n = 150$ , $P = 0.1$ , $R^2 = 0.75$ ) . . . . .	98
A.8a Average SME for FSR, AIC, and BIC for logistic regression ( $n = 500$ , $P = 0.1$ , $R^2 = 0.75$ ) . . . . .	99
A.8b Average LE for FSR, AIC, and BIC for logistic regression ( $n = 500$ , $P = 0.1$ , $R^2 = 0.75$ ) . . . . .	99
A.8c Average model size for FSR, AIC, and BIC for logistic regression ( $n = 500$ , $P = 0.1$ , $R^2 = 0.75$ ) . . . . .	99
A.9a Average SME for FSR, AIC, and BIC for logistic regression ( $n = 150$ , $P = 0.5$ , $R^2 = 0.75$ ) . . . . .	100
A.9b Average LE for FSR, AIC, and BIC for logistic regression ( $n = 150$ , $P = 0.5$ , $R^2 = 0.75$ ) . . . . .	100
A.9c Average model size for FSR, AIC, and BIC for logistic regression ( $n = 150$ , $P = 0.5$ , $R^2 = 0.75$ ) . . . . .	100
A.10a Average SME for FSR, AIC, and BIC for logistic regression ( $n = 500$ , $P = 0.5$ , $R^2 = 0.75$ ) . . . . .	101
A.10b Average LE for FSR, AIC, and BIC for logistic regression ( $n = 500$ , $P = 0.5$ , $R^2 = 0.75$ ) . . . . .	101
A.10c Average model size for FSR, AIC, and BIC for logistic regression ( $n = 500$ , $P = 0.5$ , $R^2 = 0.75$ ) . . . . .	101
A.11a Average SME for FSR, AIC, and BIC for logistic regression ( $n = 150$ , $P = 0.1$ , $R^2 = 0.35$ ) . . . . .	102
A.11b Average LE for FSR, AIC, and BIC for logistic regression ( $n = 150$ , $P = 0.1$ , $R^2 = 0.35$ ) . . . . .	102
A.11c Average model size for FSR, AIC, and BIC for logistic regression ( $n = 150$ , $P = 0.1$ , $R^2 = 0.35$ ) . . . . .	102
A.12a Average SME for FSR, AIC, and BIC for logistic regression ( $n = 500$ , $P = 0.1$ , $R^2 = 0.35$ ) . . . . .	103
A.12b Average LE for FSR, AIC, and BIC for logistic regression ( $n = 500$ , $P = 0.1$ , $R^2 = 0.35$ ) . . . . .	103
A.12c Average model size for FSR, AIC, and BIC for logistic regression ( $n = 500$ , $P = 0.1$ , $R^2 = 0.35$ ) . . . . .	103
A.13a Average SME for FSR, AIC, and BIC for logistic regression ( $n = 150$ , $P = 0.5$ , $R^2 = 0.35$ ) . . . . .	104

A.13b	Average LE for FSR, AIC, and BIC for logistic regression ( $n = 150, P = 0.5, R^2 = 0.35$ ) . . . . .	104
A.13c	Average model size for FSR, AIC, and BIC for logistic regression ( $n = 150, P = 0.5, R^2 = 0.35$ ) . . . . .	104
A.14a	Average SME for FSR, AIC, and BIC for logistic regression ( $n = 500, P = 0.5, R^2 = 0.35$ ) . . . . .	105
A.14b	Average LE for FSR, AIC, and BIC for logistic regression ( $n = 500, P = 0.5, R^2 = 0.35$ ) . . . . .	105
A.14c	Average model size for FSR, AIC, and BIC for logistic regression ( $n = 500, P = 0.5, R^2 = 0.35$ ) . . . . .	105
A.15a	The Monte Carlo estimated $\gamma$ for FSR, AIC and BIC for logistic regression ( $n = 150, P = 0.1, R^2 = 0.75$ ) . . . . .	106
A.15b	The Monte Carlo estimated $\gamma$ for FSR, AIC and BIC for logistic regression ( $n = 500, P = 0.1, R^2 = 0.75$ ) . . . . .	106
A.15c	The Monte Carlo estimated $\gamma$ for FSR, AIC and BIC for logistic regression ( $n = 150, P = 0.5, R^2 = 0.75$ ) . . . . .	106
A.15d	The Monte Carlo estimated $\gamma$ for FSR, AIC and BIC for logistic regression ( $n = 500, P = 0.5, R^2 = 0.75$ ) . . . . .	107
A.15e	The Monte Carlo estimated $\gamma$ for FSR, AIC and BIC for logistic regression ( $n = 150, P = 0.1, R^2 = 0.35$ ) . . . . .	107
A.15f	The Monte Carlo estimated $\gamma$ for FSR, AIC and BIC for logistic regression ( $n = 500, P = 0.1, R^2 = 0.35$ ) . . . . .	107
A.15g	The Monte Carlo estimated $\gamma$ for FSR, AIC and BIC for logistic regression ( $n = 150, P = 0.5, R^2 = 0.35$ ) . . . . .	108
A.15h	The Monte Carlo estimated $\gamma$ for FSR, AIC and BIC for logistic regression ( $n = 500, P = 0.5, R^2 = 0.35$ ) . . . . .	108

# List of Figures

3.1	Illustration of the tuning procedure for finding $\hat{\alpha}_0$ . . . . .	25
3.2	Graphical goodness-of-fit to $\bar{\eta}_p(\alpha)$ by formula in equation (3.9) when the predictors are uncorrelated . . . . .	29
3.3	Graphical goodness-of-fit to $\bar{\eta}_p(\alpha)$ by formula in equation (3.9) when the predictors are correlated . . . . .	30
3.4	Tuning results for different $k_p$ ( $k_T = 21$ ) . . . . .	35
3.5	Illustration of adjustment of the cut-off $c$ for different true model sizes. . . . .	38
4.1	Values of regression coefficients, $\beta_1, \beta_2, \dots, \beta_{21}$ , ( $n = 150, \rho = 0$ ) . . . . .	42
4.2	Average model error for the FSR method with different values $\gamma$ . . . . .	50
4.3	Average model size for the FSR method with different values $\gamma$ . . . . .	50
4.4	The Monte Carlo estimated false selection rate $\gamma$ for the FSR method with $\gamma = 0.05$ for different sample sizes, ( $k_T = 21$ and $k_p = 21$ ) . . . . .	51
4.5	The boxplot of the $\hat{\alpha}_0$ found by the FSR method with $\gamma = 0.05$ , ( $k_T = 21$ and $k_p = 21$ ) . . . . .	52
4.6	The ratios of average ME of Best to average ME of FSR, Tune, $C_p$ and LB when $k_T = 21$ , $n = 50$ and $n = 150$ . . . . .	56
4.7	Average model size for Best, FSR, Tune, $C_p$ , and LB when $k_T = 21$ , $n = 50$ and $n = 150$ . . . . .	57
4.8	The ratios of average ME of Best to average ME of FSR, Tune, $C_p$ and LB when $k_T = 21$ and $n = 500$ . . . . .	58
4.9	Average model size for Best, FSR, Tune, $C_p$ , and LB when $k_T = 21$ and $n = 500$ . . . . .	58
4.10	The ratios of average ME of Best to average ME of FSR, Tune, $C_p$ and LB when $k_T = 42$ and $n = 150$ . . . . .	59
4.11	Average model size for Best, FSR, Tune, $C_p$ , and LB when $k_T = 42$ and $n = 150$ . . . . .	59
4.12	Average model error for the FSR method and the Lasso . . . . .	62
4.13	Average model size for the FSR method and the Lasso . . . . .	63
4.14	The Monte Carlo estimated false selection rate $\gamma$ for the FSR method and the Lasso . . . . .	64
5.1	Average SME for the FSR method, AIC, and BIC for logistic regression when $R^2 = 0.75$ . . . . .	70
5.2	Average LE for the FSR method, AIC, and BIC for logistic regression when $R^2 = 0.75$ . . . . .	71
5.3	The Monte Carlo estimated false selection rate $\gamma$ for the FSR method, AIC, and BIC for logistic regression when $R^2 = 0.75$ , and the black solid line corresponds to value 0.05 . . . . .	72

5.4	Average SME for the FSR method, AIC, and BIC for logistic regression when $R^2 = 0.35$ . . . . .	73
5.5	Average LE for the FSR method, AIC, and BIC for logistic regression when $R^2 = 0.35$ . . . . .	74
5.6	The Monte Carlo estimated false selection rate $\gamma$ for the FSR method, AIC, and BIC for logistic regression when $R^2 = 0.35$ , and the black solid line corresponds to value 0.05 . . . . .	75

# Chapter 1

## Introduction

The variable selection problem, often referred to as the subset selection problem, has long been of interest in statistical applications. Suppose that we have data consisting of a dependent variable, denoted by  $Y$ , and a large collection of potential explanatory variables or predictors, denoted by  $X_1, \dots, X_M$ . Usually, in a regression problem, a fitted model based on a subset of  $X$ -variables is less variable and certainly simpler than a fitted model that uses the full set of  $X$ -variables. The variable selection problem is how to pick the subset of predictor variables that result in the “best” model. In fact, this problem is equivalent to a model selection problem in which each model corresponds to a particular subset.

Considering the familiar linear regression context, we have the model

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\epsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad (1.1)$$

where  $\mathbf{Y} = (y_1, y_2, \dots, y_n)^T$  is a vector of  $n$  observations,  $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_M)$  is the  $n \times M$  design matrix,  $\boldsymbol{\mu} = E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$  is a prediction function of  $\mathbf{Y}$  in terms of  $\mathbf{X}$ ,  $\boldsymbol{\beta}$  is a  $M \times 1$  vector of regression coefficients possibly with some elements zero, and  $\boldsymbol{\epsilon}$  has  $n$ -dimensional multivariate normal distribution  $N_n(0, \sigma^2 \mathbf{I})$ . The predictors with nonzero coefficients are called *important* variables. Otherwise, they are called *unimportant* variables. It is of interest to select the correct important variables and to give some estimate of the

predictive capability of the model based on this selection. Many selection procedures have been developed in the literature for linear regression models. We propose a new and general approach, the False Selection Rate (FSR) procedure, to control variable selection with the advantage of being applicable to a broad class of regression models; for example, binary regression, poisson regression, etc. We first focus on linear regression models, and then extend the approach to logistic regression models.

This thesis is organized as follows. The remainder of Chapter 1 presents an introduction to the problem of variable selection in regression and describes the most popular procedures for searching important subsets. In Chapter 2, we briefly review the literature about commonly used model selection criteria in choosing the “best” subset from the candidates. Our proposed method for controlling variable selection is described in Chapter 3. In Chapter 4, we carry out Monte Carlo simulations to study the behavior of our new method in the context of the linear regression model. The performance of our approach in logistic regression is studied by simulations in Chapter 5. In Chapter 6, our approach is illustrated on some real examples. Finally, conclusions are given in Chapter 7. The tables summarizing the simulation results and our previous approaches to estimating model size are presented in Appendices.

## **1.1 Interpretation and Prediction**

Interpretation and prediction are two main aspects of variable selection. With a large number of predictors, in order to make the model selected meaningful, we might prefer to sacrifice some of them that exhibit small effects and obtain the “big picture.” A less complex model will often give a more understandable explanation of the effects of the predictors on the dependent variable.

Since the true relationship between the  $X$ - and  $Y$ - variables is unknown in most practical cases, we often use the observed data to select the variables and to calibrate the relationship that is used for future prediction. The overall prediction accuracy of the selected model usually involves a trade-off between bias and variance. Adding additional variables to the model will reduce the bias of prediction and possibly enhance the predictive capabilities, but on the other hand it will generally also increase the predictive variation due to the increase in variance from estimating the regression parameter of the new variables. It is well known that the variance of the predicted values is proportional to the number of variables used in the prediction. If a variable has a strong effect, then the benefit from bias reduction by introducing it into the model may exceed the increase in prediction variance. Otherwise, it will be wise to exclude it from the model due to the increase in variance. Balancing the reduction in bias against the increase in variance is the key problem. Much of the work on variable selection has been devoted to trying to handle this bias and variance trade-off, and trying to find a good selection criteria to choose the optimal number of predictors. Another factor affecting the prediction variance is the sample size  $n$ . Miller (2002, P.5) pointed out “In most practical cases the prediction variance will be of the order  $n^{-1}$  while the biases from omitting variables will be of order 1.” Consequently, increasing the number of observations used to calibrate the prediction function will usually help to reduce the prediction variance.

## 1.2 Algorithms for Selecting Subsets

For searching the important subsets of variables whose model fits the data fairly well, we need a search strategy and computational algorithm. In a regression problem, a number of selection procedures have been developed. Some commonly used approaches

are *all subsets*, *forward selection*, *backward elimination*, and *stepwise regression*. Further modifications to them have been also proposed.

For each subset size of  $1, 2, \dots, M$ , all subsets searches through all possible subsets and chooses the one with the smallest residual sum of squares. An advantage of this exhaustive search is that it will not miss finding the “best-fitting” subset. If  $M$  is not too large, this exhaustive selection procedure is feasible and often recommended. Garside (1965, 1971 a,b), Furnival (1971), Morgan and Tatar (1972), and Furnival and Wilson (1974) have published algorithms for performing all-subset model evaluation. However, a drawback of this selection procedure is that it is very time consuming. When the number of predictors is too large, specifically more than 40, it is not feasible for standard computer programs to carry out all possible subsets evaluation. This computational difficulty prevents the all subsets algorithm from being widely used because in practice we often have a large collection of predictors.

Instead of searching through all subsets, stepwise methods have been proposed for evaluating a small number of subsets. The two basic ideas are called forward selection and backward elimination, which either add or delete variables one at a time. Starting with no variables in the model, forward selection adds one variable at a time until either all variables are included or until a stopping criterion is satisfied. At each step, the method adds the variable that makes the largest contribution to the model if it is included compared to other variables eligible for inclusion. In contrast, backward elimination goes in the opposite way. Beginning with the full model, this method eliminates one variable at a time. At any step, it excludes the variable yielding the smallest contribution from the model. The variables are deleted one by one until none is left in the model or until a stopping criterion is satisfied. Both forward selection and backward elimination proceed in only one direction: forward or backward. As a combination of the two procedures, stepwise regression, which is often referred to an algorithm proposed by



Efroymson (1960), selects the variables back and forth. Starting with the null model, at each step, this selection method adds a variable into the model as in forward selection, and then tests if any of the previously selected variables can be removed as in backward elimination. This procedure is repeated until no further additions or deletions are possible according to a particular criterion. Similar to Efroymson’s algorithm in incorporating the forward selection and backward elimination, Broersen (1986) proposed a variation with “stepwise directed search” that can select automatically the subset with the smallest  $C_p$  (see Section 2.1 for a discussion of  $C_p$ ).

Stepwise procedures are highly recommended for data sets with large number of predictor variables, and have been implemented in all major statistical software packages. Particularly, forward selection is probably the most pervasive subset selection methods used, because it is straightforward and still feasible even when there are more variables than observations while other selection procedures, like all subsets and backward elimination, fail. Efron et al. (2004) propose another promising model selection method, *forward stagewise*, which is a much more cautious version of forward selection. This method fixes the problem faced by forward selection that an important variable is often ignored when it is correlated with some variables already in the model. It was the original motivation for the new model selection algorithm, *Least Angle Regression* (LARS), by Efron et al. (2004). A simple LARS modification efficiently implements forward stagewise linear regression, greatly reducing the computational burden.

Due to the small proportion of subsets examined, stepwise is feasible even when the number of explanatory variables is huge. However, stepwise methods do not guarantee finding the best-fitting subset. Miller (2002) emphasized that there is no reason that the best-fitting subset of  $p$  variables should contain the best-fitting subset of  $p - 1$  variables; sometimes the two have no variables in common. In such cases, it is very possible for “cheap” selection methods to miss the best model because of the restriction of adding or

removing one variable at a time. To reduce this effect, algorithms which add or remove two or more variables at each step were discussed by Miller (2002, Chapter 3). Replacing two or more variables at a time is more likely to find the best-fitting subsets than replacing one variable at a time. Such selection methods are usually feasible even with hundreds of predictor variables, but the searching time is still so long that they are not used widely in practice. The “cheap” stepwise selection procedures are also criticized because of their implication of an order of importance of the variables, which can be misleading because the importance of a variable could be altered in the presence of other ones. Hocking (1976) pointed out that the first variable selected by forward selection can be deleted first by backward elimination.

### **1.3 Selection Bias**

Selection bias is introduced when the same data are used for both selecting the subset of variables in the model and for estimating their regression coefficients. Any data-driven procedure inevitably has this problem. Lane and Dietrich (1976) carried out a simulation to study such bias. They found that, for some of the regression coefficients, the average estimated values when the variables were selected were double their true values. Zhang (1992, p.741) warned that if the model is built by a data-driven selection criterion, then standard statistical inferences are both logically unsound and practically misleading. More inferential problems associated with data-driven model building are discussed extensively by Chatfield (1995). Actually, different selection procedures produce different amounts of selection bias. That is, selection bias is a function of the selection method. Miller (2002, P.165) pointed out that “It can usually be anticipated that the more extensive the search for the chosen model, the more extreme the data values must be to

satisfy the conditions and hence, the greater the selection bias.” From this point of view, all subsets selection can be expected to yield larger bias than stepwise methods in which a much smaller number of models are compared.

Selection bias presents a dilemma. On one hand, we would like to choose a good model which can give good prediction on the basis of the information provided by the data under study; on the other hand, the model selected should not be dependent on the given data to which the model will be applied in order to obtain unbiased parameter estimates. To reduce the effects of selection bias, a number of methods are discussed in the literature. One procedure is to use ‘shrunk’ regression estimators instead of least squares estimators. Ridge regression proposed by Hoerl and Kennard (1970 a,b) and LASSO introduced by Tibshirani (1996) are two kinds of popular shrinkage methods. By putting a constraint on the regression coefficient, these procedures shrink the coefficient estimators toward zero. This results in the estimated coefficients of unimportant variables to be close to zero or even equal to zero. A suitable choice of the parameter controlling the amount of shrinkage is chosen such that some kind of optimum is reached. However, because only one parameter is used to control the amount of shrinkage, the selection bias cannot be eliminated or reduced simultaneously in all the regression parameters. Also, shrinkage methods tend to leave too many variables in the model such that the scientific insight into the  $X - Y$  relationship becomes difficult. Another way to completely eliminate selection bias is to use independent data sets for model selection and model estimation. That is, we can randomly split the data into two parts with one used for selecting the subset of variables and the other for estimating the regression coefficients. Doing this requires that the sample size is large enough. However, in practice we are often not so lucky. In fact, splitting the data will reduce the information used for selection and estimation, and hence produce less efficient results. Roeder (1991) reported that prediction based on using half of the data for model selection and the other half for model

estimation was inferior to prediction in which the whole data were used for both selection and estimation. Due to the difficulty in handling selection bias, there is no satisfactory general solution. Most subset selection procedures usually ignore this bias.

# Chapter 2

## Selection Criteria

The procedures of subset selection, such as *all subsets*, *forward selection*, and *backward elimination*, generate a sequence of subsets of  $\{X_1, X_2, \dots, X_M\}$  of dimension  $1, 2, \dots, M$ , each of which is the “best” of its size according to a particular searching criterion. Then given these candidate subsets, the problem is to choose one yielding optimum or near optimum performance. In other words, an appropriate selection criteria is required to determine which of the  $M$  subsets to use. A number of prevalent methods used in stepwise methods are  $F$ -to-enter,  $F$ -to-delete, and adjusted  $R^2$ . Bendel and Afifi (1977) provided an early comparison of stopping rules in forward “stepwise” regression. Most common selection criteria combine statistical measures with penalties for increasing complexity (number of predictors). Good reviews can be found in Hocking (1976), Thompson (1978), Rao and Wu (2001), and Miller (2002). Basically, we can broadly divide these criteria into four classes :

- Prediction criteria;
- Data resampling methods;
- Information-based criteria;
- Bayesian variable selection.

A brief discussion of them and their relationship are given in the next section. Unless indicated otherwise, the following discussion is restricted to linear regression models.

## 2.1 Prediction Criteria

Prediction is an important criterion in practice. The prediction accuracy is often evaluated by prediction error (PE), which is defined as

$$\text{PE}(\hat{\boldsymbol{\mu}}) = \frac{1}{n} E \|\mathbf{Y}^{new} - \hat{\boldsymbol{\mu}}(\mathbf{X})\|^2, \quad (2.1)$$

where the expectation is taken over  $\mathbf{Y}^{new}$  only, where  $\mathbf{Y}^{new}$  is a new random vector with the same distribution as the original  $\mathbf{Y}$ ,  $\hat{\boldsymbol{\mu}}$  is the prediction equation derived from the present data, and  $\|\cdot\|$  refers to Euclidean norm. In this definition, the prediction error represents the average error in predicting  $\mathbf{Y}$  from  $\mathbf{X}$  for future cases based on the calibrated model. In the context of linear model (1.1), (2.1) can be also written as

$$\text{PE}(\hat{\boldsymbol{\mu}}) = \sigma^2 + \frac{1}{n} \|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|^2. \quad (2.2)$$

It is straightforward to see that the prediction error consists of two parts: the first term is inherited from the random error; the second term is due to lack of fit to the true model. The second term is called the model error (ME)

$$\text{ME} = \frac{1}{n} \|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|^2. \quad (2.3)$$

Breiman (1992) defined PE and ME without taking the average over  $n$ , in other words, his definitions are  $n$  times (2.2) and (2.3).

From the prediction point of view, we should select the model in which the PE is small. Notice that the first term in PE is not affected by the variable selection procedure, and hence can be ignored and only ME is considered. However, values of ME cannot be

computed directly because the underlying true  $\boldsymbol{\mu}$  is unknown. In order to deal with this difficulty, many procedures have been developed to construct an estimate  $\widehat{\text{ME}}$ , and then select a model by minimizing it. Perhaps the most widely used of these criteria is to minimize the statistic  $C_p$  proposed by Mallows (1973). Mallows'  $C_p$  is defined as

$$C_p = \frac{RSS_p}{\hat{\sigma}_F^2} - n + 2p,$$

where  $RSS_p$  is the residual sum of squares from the least squares fit of  $p$  predictors, and  $\hat{\sigma}_F^2 = RSS_F/(n-M)$  is an unbiased estimate of  $\sigma^2$ , where  $RSS_F$  is the full model residual sum of squares and  $n > M$ . Mallows (1995) discussed minimizing  $C_p$  in the case of a large number of predictors and in the presence of many weak variables. He pointed out that  $C_p$  has the property that  $E(n\text{ME}) = E(\hat{\sigma}_F^2 C_p)$ . In this sense,  $\hat{\sigma}_F^2 C_p/n$  can be used to estimate expected ME, and minimizing  $C_p$  is approximately equivalent to minimize ME. However, unless the subset is selected independently of the observed data,  $E(\text{ME})$  will not be estimated accurately because of the selection bias. As a result, Mallows warned that the selection based on minimum  $C_p$  can be misleading and can give much worse prediction than if there were no selection at all and all predictors are used.

Mallows'  $C_p$  has received considerable attention since it was proposed, and a number of modifications have been suggested to improve its performance. Gilmour (1996) suggested to adjust the  $C_p$  slightly as

$$\overline{C}_p = C_p - \frac{2(M-p)}{n-M-2}, \quad n-M-2 > 0.$$

The aim is to make the expected value of  $\overline{C}_p$  equal to  $p$ , even for models that include all important predictors while  $E(C_p)$  is more likely less than  $p$ . He also suggested that it is worthwhile to apply hypothesis tests to the differences between successive values of  $C_p$  or adjusted  $\overline{C}_p$  instead of minimizing  $C_p$  in order to keep the model selected from overfitting. A further modification to achieve robustness of Mallows'  $C_p$  has been proposed

by Ronchetti and Staudte (1994).

## 2.2 Data Resampling Selection

In recent years, data resampling methods have received increasing attention and applied widely to variable selection. The most common methods advocated are cross-validation, bootstrap and little bootstrap. They estimate the model error for each candidate subset of variables by resampling the data in a certain way, and select the subset by minimizing the estimated model error. The selection is still according to the predictive ability of the model as Mallows'  $C_p$ , but the data resampling methods reduce the large biases in the estimate of model error.

### Cross-validation

The basic idea behind cross-validation is to use part of the data to fit the model, and use the rest to evaluate the constructed model. Generally,  $V$ -fold cross-validation divides the data set randomly into  $V$  equal or nearly equal parts. Denote these parts by  $\Gamma_1, \dots, \Gamma_V$  and let  $\Gamma^{(v)}$  denote the data with  $\Gamma_v$  deleted,  $v = 1, \dots, V$ . Suppose that we have a sequence of candidate subset models  $\xi_0, \xi_1, \dots, \xi_M$  with size  $0, 1, \dots, M$ . Given a subset  $\xi_J$ , we carry out the same selection procedure on  $\Gamma^{(v)}$ , and select a subset model  $\xi_J^{(v)}$  which has the same size  $J$  but with no guarantee of the same variables as  $\xi_J$ .  $\Gamma_v$  is then used for assessing the predictive ability of model  $\xi_J^{(v)}$ . This is done in turn for each  $v = 1, 2, \dots, V$ . The prediction error for the candidate model  $\xi_J$  is then estimated by

$$\widehat{\text{PE}}(J) = \frac{1}{n} \sum_{v=1}^V \sum_{(y_i, \mathbf{x}_i) \in \Gamma_v} (y_i - \hat{\mu}_v(\mathbf{x}_i, \xi_J^{(v)}))^2,$$



where  $\hat{\mu}_v(\mathbf{x}, \xi_J^{(v)})$  is the predictor of  $\Gamma_v$  for submodel  $\xi_J^{(v)}$  fitted from  $\Gamma^{(v)}$ . According to equation (2.2), the model error for model  $\xi_J$  can be estimated by

$$\widehat{\text{ME}}(J) = \widehat{\text{PE}}(J) - \hat{\sigma}_F^2. \quad (2.4)$$

The variable selection of cross-validation is then based on minimizing  $\widehat{\text{ME}}(J)$ .

Five-fold cross-validation is recommended by Breiman and Spector (1992) for subset selection and evaluation. That is, we let  $V = 5$ , and each time leave out 20% of data for prediction assessment and use the remaining 80% for model selection. Leaving out one case at a time ( $V = n$ ) is asymptotically equivalent to many other stopping criteria such as Akaike information criterion (AIC) and Mallows'  $C_p$ . However, the simulation results by Breiman and Spector (1992) show that although  $V = n$  has a good global estimate of ME, it does not perform as well as leaving 20% out at subset selection and evaluation. An important theoretical contribution to cross-validation was made by Shao (1993). He pointed out that leaving out one-at-a-time is asymptotically inconsistent in the sense that it does not converge to the correct model, but rather tends to include too many variables as the total number of observations  $n$  goes to infinity. He shows that this deficiency can be rectified by leaving  $n_v$  cases out each time, where  $n_v$  satisfies  $n_v/n \rightarrow 1$  as  $n \rightarrow \infty$ . A robust version of cross-validation was developed by Ronchetti et al. (1997).

## Bootstrap

As for application of bootstrap to regression subset selection, there are usually two ways of generating bootstrap observations:

1. Bootstrapping residuals. Fit the full model and get the estimate of regression coefficients,  $\hat{\beta}$ , and the residuals,  $\hat{\epsilon}_i$ ,  $i = 1, \dots, n$ . Studentize these residuals using  $\hat{\epsilon}_i^* = \hat{\epsilon}_i / \sqrt{1 - d_i}$ , where the  $d_i$  are the  $i$ th diagonal elements of  $\mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ .

Then, for each case  $\mathbf{x}_i$ , the new response  $y_i$  is generated by  $\mathbf{x}_i\hat{\boldsymbol{\beta}} + \tilde{\epsilon}_i^*$ , where  $\tilde{\epsilon}_i^*$  is sampled from  $\{\hat{\epsilon}_1^*, \dots, \hat{\epsilon}_n^*\}$  with replacement. This method is suitable for the case of a fixed design  $\mathbf{X}$ .

2. Bootstrapping pairs. Generate the bootstrap data by sampling with replacement  $n$  times from the original  $(\mathbf{x}_i, y_i)$  pairs. This method is more appropriate for the case of random design  $\mathbf{X}$ , but is attractive because it makes no model assumption.

Efron (1982, 1983) stated how to derive the bootstrap estimate of the prediction error PE in (2.1). Correspondingly, the model error in (2.3) can be estimated by subtracting  $\hat{\sigma}_F^2$  from the estimate of PE. Shao (1996) pointed out that, for the traditional bootstrap sampling plan drawing  $n$  bootstrap observations, the probability of selecting the correct subset of variables does not converge to 1 as the number of observation  $n \rightarrow \infty$ . He suggested rectifying this inconsistency by drawing smaller bootstrap observations for bootstrapping pairs, and modifying the bootstrap sampling procedure by increasing the variability among the bootstrap observations for bootstrapping residuals.

## Little Bootstrap

The little bootstrap variable selection procedure was introduced by Breiman (1992). For a candidate subset model  $J$ , the model error can be written as

$$\text{ME}(J) = RSS_J - RSS_F + \|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_F\|^2 - 2(\boldsymbol{\epsilon}, \hat{\boldsymbol{\mu}}_F - \hat{\boldsymbol{\mu}}_J),$$

where  $RSS_J$  and  $\hat{\boldsymbol{\mu}}_J$  are the least-squares residual sum of squares and prediction equation, respectively, from fitting submodel  $J$ ,  $RSS_F$  and  $\hat{\boldsymbol{\mu}}_F$  are calculated from full model fitting, and  $(\cdot, \cdot)$  indicates the inner product. The term  $\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_F\|^2$  can be estimated by  $M\hat{\sigma}_F^2$ . To estimate the last term, the little bootstrap procedure is applied. First, the dependent

variable  $y_i$  is contaminated by adding additional noise, i.e.,

$$\tilde{y}_i = y_i + \tilde{\epsilon}_i,$$

where  $\tilde{\epsilon}_i$  i.i.d.  $N(0, t^2 \hat{\sigma}_F^2)$ ,  $t > 0$ . Then, one performs the same chosen subset selection procedure on the new data  $(\tilde{y}_i, \mathbf{x}_i)$ ,  $i = 1, \dots, n$ , and identifies a subset  $\tilde{J}$  that has the same size as subset  $J$ . There is no guarantee that  $\tilde{J}$  and  $J$  are the same. Calculate

$$\frac{1}{t^2}(\tilde{\epsilon}, \tilde{\boldsymbol{\mu}}_F - \tilde{\boldsymbol{\mu}}_{\tilde{J}}), \quad (2.5)$$

where  $\tilde{\boldsymbol{\mu}}_F$  and  $\tilde{\boldsymbol{\mu}}_{\tilde{J}}$  are based on the new data. Repeat this procedure a number of times  $B$ , and denote the average of quantities (2.5) as  $D(J)$ . Breiman (1992) showed that, when  $t$  is small,

$$\frac{1}{t^2}E(\tilde{\epsilon}, \tilde{\boldsymbol{\mu}}_F - \tilde{\boldsymbol{\mu}}_{\tilde{J}}) \approx E(\epsilon, \hat{\boldsymbol{\mu}}_F - \hat{\boldsymbol{\mu}}_J).$$

Therefore, the little bootstrap  $\text{ME}(J)$  estimate is taken as

$$\widehat{\text{ME}}(J) = \text{RSS}_J - \text{RSS}_F + M\hat{\sigma}_F^2 - 2D(J).$$

By his simulation, Breiman suggested setting  $B = 40$  and  $t = 0.6$ . The little bootstrap gives almost unbiased estimates of the submodel  $\text{ME}(J)$ . The model minimizing  $\widehat{\text{ME}}(J)$  is then selected.

## 2.3 Information-based Criteria

Another widely used family of criteria are those based on likelihood or information measures. Among them, the two most popular criteria are the Akaike Information Criterion (AIC) proposed by Akaike (1973, 1977) and the Bayesian Information criterion (BIC) by Schwarz (1978). The AIC selects the model that minimizes

$$\text{AIC}_j = -2L_j + 2p_j, \quad (2.6)$$

where  $L_j$  is the maximum log-likelihood of the  $j$ th model, and  $p_j$  represents the complexity of the model, which is the number of predictors in the model. BIC has the same form as AIC except that the log-likelihood is penalized by  $\log(n)$  rather than by 2. That is, the BIC selects the model that minimizes

$$\text{BIC}_j = -2L_j + p_j \log(n). \quad (2.7)$$

Asymptotic properties (i.e., consistency) of AIC and BIC have been well studied and compared in the literature. Stone (1977, 1979) showed the asymptotic equivalence of the AIC and cross-validation, and made an illuminating comparison of AIC and BIC. Shibata (1981) claimed that AIC is consistent if the true model size increases with  $n$  at an appropriate rate. When the true model is fixed as  $n$  goes to infinity, Haughton (1988) showed that BIC is consistent. Hurvich and Tsai (1989) derived a bias correction to the AIC, called  $\text{AIC}_C$ , which has the form

$$\text{AIC}_C = \text{AIC} + \frac{2(M+1)(M+2)}{n-M-2},$$

where  $M$  is the number of explanatory variables.  $\text{AIC}_C$  is asymptotically equivalent to AIC but similarly not consistent. It performs well when the sample size is small, or when the true model is of finite dimension.

AIC and BIC are generally used for any model selection. For the linear model (1.1), both AIC and BIC are special cases of the generalized information criterion (GIC), which selects a model by minimizing

$$\text{GIC}_j = \frac{RSS_j}{\hat{\sigma}_F^2} + \lambda p_j, \quad (2.8)$$

where  $RSS_j$  is the residual sum of squares from fitting the  $j$ th model, and  $\lambda$  is a penalty parameter on the complexity of the model  $p_j$ . The AIC is asymptotically equivalent to the GIC with  $\lambda = 2$ , whereas asymptotically BIC is obtained by setting  $\lambda = \log(n)$ .

When  $\lambda = 2$ , the GIC is equal to the well-known Minimum  $C_p$  method.

## 2.4 Bayesian Variable Selection

A dramatically increasing level of Bayesian activity has been seen in the last decade, and a large number of papers were devoted to the applications of Bayesian methods to variable selection. The fully Bayesian selection approach is as follows. Suppose that we have a set of models,  $M_1, M_2, \dots, M_K$ , each of which corresponds to a subset of variables  $\{X_1, X_2, \dots, X_M\}$ , and  $K$  could equal  $2^M$ . Two prior probabilities are assumed. That is, the prior for the model, denoted by  $\pi(M_\kappa)$ , and the prior for the parameters in that model, denoted by  $\pi(\boldsymbol{\beta}_\kappa|M_\kappa)$ ,  $\kappa = 1, \dots, K$ . The posterior probability for model  $\kappa$  is then calculated by

$$\pi(M_\kappa|\mathbf{Y}, \mathbf{X}) = \frac{p_\kappa(\mathbf{Y}|\mathbf{X})\pi(M_\kappa)}{\sum_{\kappa=1}^K p_\kappa(\mathbf{Y}|\mathbf{X})\pi(M_\kappa)}, \quad (2.9)$$

where

$$p_\kappa(\mathbf{Y}|\mathbf{X}) = \int p_\kappa(\mathbf{Y}|\mathbf{X}, \boldsymbol{\beta}_\kappa)\pi(\boldsymbol{\beta}_\kappa|M_\kappa)d\boldsymbol{\beta}_\kappa,$$

and  $p_\kappa(\mathbf{Y}|\mathbf{X}, \boldsymbol{\beta}_\kappa)$  is the joint density of  $\mathbf{Y}$  for model  $M_\kappa$ . The selection is then based on the posterior model probabilities  $\pi(M_\kappa|\mathbf{Y}, \mathbf{X})$ ; for example, pick the model giving the highest posterior probability.

The obvious problem faced by Bayesian variable selection method is how to choose the two priors. The specification of priors strongly influences the posterior probabilities. For this reason, much of the work on Bayesian variable selection procedure has been devoted to finding “good” priors with an attempt to reduce this effect. An early paper by Lindley (1968) proposed to use uniform priors and a cost function for selection. Mitchell and Beauchamp (1988) suggested using “spike and slab” priors. Some other criteria were also developed for Bayesian variable selection with the goal of avoiding the prior selection

difficulties, such as the default Bayes factor criteria by Berger and Pericchi (1996a, b), and the predictive criteria by Laud and Ibrahim (1995). Another difficulty in Bayesian selection is posterior computation, especially when the set of models is large. Markov chain Monte Carlo (MCMC) is used to reduce this difficulty and has become the most popular method of Bayesian computation.

Another development in this context has been the use of model averaging rather than just a single model. The reason is that it is very possible that the model selected could come out best just by chance based on the given data and chosen procedure. From the Bayesian point of view, each model is associated with a probability. It is proposed to take the posterior model as a weighted average of all of the models with some models given moderately large weights and most of them given very small weights based on the posterior probabilities. The model averaging improves the predictive ability on variable selection, but it does not reduce the number of variables.

More recent discussion on Bayesian variable selection can be seen in George and McCulloch (1993), Clyde et al. (1998), Chipman et al. (2001), Berger and Pericchi (2001), and Miller (2002, Chapter 7). Its application for logistic regression models is discussed by Chen et al. (1999).

# Chapter 3

## Controlling Variable Selection

### 3.1 Motivation

Miller (2002, Chapter 4) mentions the use of adding artificial variables to the original set of predictor variables and then checking the first appearance of them in the selected model. His idea of adding artificial variables is attractive, but only checking the first appearance of these variables can lead to variable results, and the selection is very sensitive to the number of artificial variables added. Rather than considering the first appearance of the artificial variables, we are thinking about the problem in a different way.

Suppose that we augment the set of original variables with a number of artificial variables that have no relationship with the responses by generation. We call these artificial variables *pseudo-variables*. Intuitively, a good selection criterion should not include too many of the pseudo-variables in the model selected. If a procedure selects a model containing a lot of pseudo-variables, then this procedure is too generous at including variables that should be excluded, and hence the selection procedure tends to overfit. On the other hand, if a procedure never selects pseudo-variables, then the selection procedure is too “ruthless” at excluding variables and will seldom include weak important variables for which the regression coefficients are small but non-zero. Hence it tends to underfit the model. As a result, the number of pseudo-variables falsely selected

is an indicator of model overfitting or underfitting for any selection procedure. Motivated by this idea, we develop a new approach to variable selection based on monitoring and controlling the number of pseudo variables added to the real set of explanatory variables.

We focus on forward selection because it is applicable in the case where there are more predictors than observations ( $M > n$ ). At each step, forward selection includes a variable based on its “importance”, which is measured by the statistical significance of the coefficient for the variable. For example, in linear regression, because the errors are assumed to be normally distributed, an  $F$ -test is carried out. In particular, given  $p$  variables (including an intercept) already in the model, for each of the variables not in the model, an  $F$  statistic is calculated by

$$F = \frac{RSS_p - RSS_{p+1}}{RSS_{p+1}/(n - p - 1)}, \quad (3.1)$$

where  $RSS_p$  is the residual sum of squares for the previously selected  $p$  variables,  $RSS_{p+1}$  is the residual sum of squares when an additional variable is included. Under the test of the null hypothesis that the coefficient of the  $(p + 1)$ th variable is 0, if we assume no selection bias, the associated  $P$ -value is therefore

$$P\text{-value} = \Pr(F_{1,n-p-1} > F|F). \quad (3.2)$$

The most important variable is the one with the smallest  $P$ -value. Thus, the forward selection procedure will include the variable with the minimum  $P$ -value. Forward selection can be easily implemented in SAS for a variety of regression models. For example, with the selection option set to “FORWARD”, SAS PROC REG carries out forward selection in the linear regression context. A crucial aspect of using forward selection is to choose a suitable significance level  $\alpha$  to control the entry into the model. Specifically, suppose that at the  $i$ th step the  $P$ -value associated with the variable selected is denoted by  $P_{\text{MIN}}^{(i)}$ . Then, the selection will be terminated whenever  $P_{\text{MIN}}^{(i)}$  exceeds the specified level  $\alpha$ . The



choice of  $\alpha$  sets a threshold for judging the importance of entering variables and determines how many variables eventually are contained into the model. When  $\alpha$  is large, the method tends to include too many variables into the model. On the other hand, a small  $\alpha$  tends to result in an underfitted model. In SAS, “SLENTY” in the procedures specifies the  $\alpha$  in forward selection.

For choosing  $\alpha$ , the most widely used level is 0.1 or 0.05. However, it is easily seen that neither of these two levels are good choices because they don’t reflect the dependence of  $\alpha$  on the particular data set. In fact, when there are more important variables, a larger  $\alpha$  level should be used to give each important variable a chance of being selected comparable to that when there are a smaller number of important variables. Also, the overall power as well as the type I error rate are unknown due to the fact that the order of entry of the variables differs with observations. Luo, Boos and Stefanski (2002) proposed an approach to tune the stopping rule  $\alpha$  based on adding additional noise to the response variable with the hope of finding an unbiased estimator of the total error variance  $\sigma^2$ . Their method is restricted to additive error models.

Here we develop a new method, the False Selection Rate (FSR) procedure, for estimating the optimal level  $\alpha$  based on augmenting the original explanatory variables with a number of pseudo-variables. By controlling the proportion of pseudo-variables falsely selected in the model, we are able to control the model false selection rate, keeping a low proportion of unimportant variables in the model. One advantage of this new procedure is that it is applicable to a broader class of regression models; for example, binary regression, poisson regression, etc.

## 3.2 False Selection Rate Variable Selection

For convenience, let us define the following notation.

$k_T$  = the total number of original predictor variables. This was  $M$  before.

$k_I$  = the number of important original predictor variables ( $\beta \neq 0$ ) .

$k_U$  = the number of unimportant original predictor variables ( $\beta = 0$ ).

$k_P$  = the number of generated pseudo-variables.

$X_1, \dots, X_{k_T}$  = original predictor variables, not including the intercept.

$Z_1, \dots, Z_{k_P}$  = generated pseudo-variables.

Obviously,  $k_I + k_U = k_T$ . Here,  $k_T$  and  $k_P$  are known, and  $k_I$  and  $k_U$  are unknown. We create a set of  $k_P$  pseudo-variables that are independent of the response variable. Generation of the pseudo-variables will be discussed later in Section 3.3. Given a significance level  $\alpha$ , for the augmented set of data, the subset selected by forward selection usually includes three types of variables: important original variables, unimportant original variables and generated pseudo-variables. The possible variables in the selected subset are displayed in Table 3.1.

Table 3.1: Possible variables in the selected subset

Original Variables		Pseudo-Variables	Total
Important	Unimportant		
$S_I(\alpha)$	$S_U(\alpha)$	$S_P(\alpha)$	$S(\alpha)$

For example,  $S_P(\alpha)$  is the number of pseudo-variables contained in the selected subset, while  $S(\alpha)$  is the total number of variables selected, i.e.,  $S(\alpha) = S_I(\alpha) + S_U(\alpha) + S_P(\alpha)$ . It is obvious that these two numbers are functions of  $\alpha$  with their values increasing as  $\alpha$  increases. Specifically, with intercept included and considered as an important variable,

we have  $S(0) = 1$  and  $S(1) = k_T + k_P + 1$ , and  $S_P(0) = 0$  and  $S_P(1) = k_P$ . We denote the proportion of pseudo-variables among the selected variables by

$$\begin{aligned}\eta_P(\alpha) &= \frac{\# \text{ pseudo-variables selected}}{\# \text{ total variables selected}} \\ &= \frac{S_P(\alpha)}{S(\alpha)},\end{aligned}\tag{3.3}$$

where symbol  $\#$  means the number.  $\eta_P(\alpha)$  is to quantify the extent of falsely selected pseudo-variables. To reduce the variation in  $\eta_P(\alpha)$ , we repeatedly generate sets of  $k_P$  pseudo-variables  $B$  times, for example,  $B = 500$ , and then take the averages of the numerator and denominator of  $\eta_P(\alpha)$  over the  $B$  times generation, respectively. This results in the Monte Carlo averaged version of  $\eta_P(\alpha)$ ,

$$\bar{\eta}_P(\alpha) = \frac{\bar{S}_P(\alpha)}{\bar{S}(\alpha)},\tag{3.4}$$

where

$$\bar{S}_P(\alpha) = \frac{1}{B} \sum_{j=1}^B S_{P,j}(\alpha) \quad \text{and} \quad \bar{S}(\alpha) = \frac{1}{B} \sum_{j=1}^B S_j(\alpha),$$

and  $S_j(\alpha)$  and  $S_{P,j}(\alpha)$  are the total number of selected variables and the number of pseudo-variables selected for the  $j$ th generation of pseudo-variables, respectively,  $j = 1, \dots, B$ . Suppose that  $\alpha_0$  denotes the optimal significance level that results in the subset whose model is the best in the sense of neither overfitting nor underfitting. By controlling the averaged proportion of pseudo-variables falsely selected,  $\bar{\eta}_P(\alpha)$ , we are able to find an estimate of  $\alpha_0$ . Particularly, by specifying a suitable cut-off  $c$ , we estimate  $\alpha_0$  by  $\hat{\alpha}_0$  solving the equation,

$$\bar{\eta}_P(\hat{\alpha}_0) = c.$$

On average, the function  $\bar{\eta}_P(\alpha)$  tends to be monotone increasing for small  $\alpha$  but there is no guarantee of monotonicity for particular data sets. The function  $\bar{\eta}_P(\alpha)$  may be also not continuous. Thus, to ensure a unique solution for any given data set, we choose the

smallest  $\alpha$  satisfying  $\bar{\eta}_P(\alpha) \geq c$ . That is

$$\hat{\alpha}_0 = \text{Min}\{\alpha : \bar{\eta}_P(\alpha) \geq c\}. \quad (3.5)$$

In practice, we search  $\hat{\alpha}_0$  over a set of values, denoted by  $\mathcal{A}$ . Typically, we take  $\mathcal{A} = \{0.002, 0.004, 0.006, \dots, 0.2\}$ . For each  $\alpha$  in  $\mathcal{A}$ ,  $\bar{\eta}_P(\alpha)$  is calculated, and then we take  $\hat{\alpha}_0$  to be the smallest value that has  $\bar{\eta}_P(\alpha)$  more than or equal to  $c$ , i.e.,

$$\hat{\alpha}_0 = \text{Min}\{\alpha : \bar{\eta}_P(\alpha) \geq c, \alpha \in \mathcal{A}\}. \quad (3.6)$$

This tuning procedure is illustrated in Figure 3.1, in which the curve denoted by symbol “+” stands for the averaged proportion  $\bar{\eta}_P(\alpha)$ , and the solid horizontal line corresponds to the cut-off  $c$  that is 0.058 in this example. The point marked by  $\oplus$  has coordinates  $(\hat{\alpha}_0, \bar{\eta}_P(\hat{\alpha}_0))$ , where  $\bar{\eta}_P(\hat{\alpha}_0) \geq 0.058$  and  $\hat{\alpha}_0 = 0.018$ . The value of  $\hat{\alpha}_0$ , 0.018, is then used as the “SLENTY” in a final application of forward selection on the true set of data.

This raises two obvious questions, namely how to choose the cut-off  $c$  and how many pseudo-variables  $k_P$  should be created. It is easily seen that the specification of  $c$  plays a key role in the tuning procedure. Also, the sensitivity of the procedure to the number of pseudo-variables generated is of interest. Before answering the two questions, we first introduce an important definition, the false selection rate. This definition is about the final model selected that includes only the important and unimportant original variables, and is not related to the pseudo-variables.

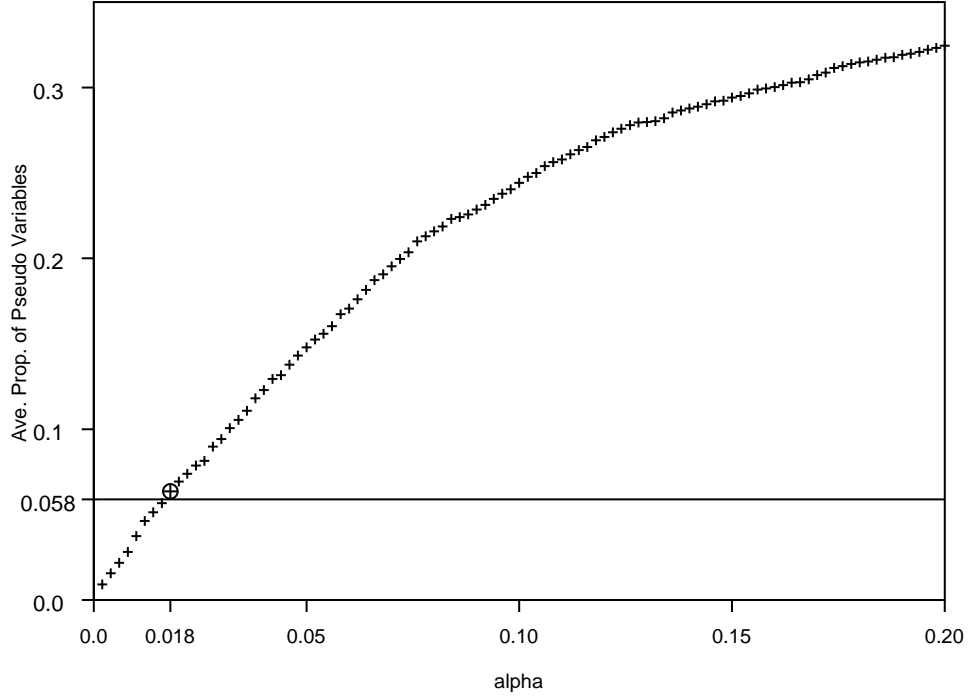


Figure 3.1: Illustration of the tuning procedure for finding  $\hat{\alpha}_0$   
( $c = 0.058$  and  $\hat{\alpha}_0 = 0.018$ )

**Definition 3.1** Suppose that in the final model selected, the number of important original variables is denoted by  $I(\mathbf{Y}, \mathbf{X})$ , where  $\mathbf{Y}$  is the vector of responses and  $\mathbf{X}$  is the matrix made up from the original predictor variables, and the number of unimportant original variables is denoted by  $U(\mathbf{Y}, \mathbf{X})$ . The model *false selection rate*, denoted by  $\gamma$ , is defined to be the expected proportion of falsely selected unimportant original variables among the total variables selected, i.e.,

$$\begin{aligned} \gamma &= \text{E} \left( \frac{\# \text{ falsely selected unimportant original variables}}{\# \text{ total selected original variables}} \right) \\ &= \text{E} \left( \frac{U(\mathbf{Y}, \mathbf{X})}{U(\mathbf{Y}, \mathbf{X}) + I(\mathbf{Y}, \mathbf{X})} \right). \end{aligned}$$

Notice the difference between  $\gamma$  and  $\bar{\eta}_P(\alpha)$ . The former is based on the selection from only the original set of variables, whereas the latter is based on the selection from the augmented data including the pseudo-variables. The model false selection rate  $\gamma$  reflects the probability of the selection method to falsely select unimportant predictors. Thus, it is meaningful to control it to a suitable value. For example, if  $\gamma$  is equal to 0.05, then this implies that among all the variables selected, about 5% of them are not really important (i.e., falsely selected). In this sense,  $\gamma$  can be used as a target for controlling variable selection. It can help to control the type I error rate.

Now, let us go back to our previous two issues of determining the value of the cut-off  $c$  and the number of pseudo-variables  $k_P$ . To address these issues, we derive a relationship between the false selection rate  $\gamma$  and the cut-off  $c$  as follows based on the proportion  $\bar{\eta}_P(\alpha)$ .

As we described in Section 3.1, forward selection chooses the variables according to their test  $P$ -values. Provided that the  $k_I$  important real variables are already in the model, for the unimportant real variables and generated pseudo-variables, their  $P$ -values to enter are distributed as Uniform(0,1) because their underlying coefficients are zero. Thus, given a significance level  $\alpha$ , the probability of an unimportant real variable or a pseudo-variable entering into the model is approximately equal to  $\alpha$ . In addition, suppose that forward selection includes each important real variable  $X_i$  with a probability  $\lambda_i$ ,  $i = 1, \dots, k_I$ , and that the average of these probabilities is denoted by  $\lambda$ , i.e.,

$$\lambda = \frac{1}{k_I} \sum_{i=1}^{k_I} \lambda_i.$$

Here,  $\lambda$  may depend on  $\alpha$ , i.e.,  $\lambda = \lambda(\alpha)$ . However,  $\lambda$  does not vary much with  $\alpha$  over a limited range of “small”  $\alpha$  values, e.g.,  $\alpha \in \mathcal{A} = \{0.002, 0.004, \dots, 0.2\}$ . Thus we can essentially regard it as a constant. Based on above selection probabilities, we can easily

derive the expectation of  $S(\alpha)$  (including the intercept) in Table 3.1 as

$$\begin{aligned}
E(S(\alpha)) &= E\left(\sum_{i=1}^{k_T} I(X_i) + \sum_{i=1}^{k_P} I(Z_i) + 1\right) \\
&= \sum_{i=1}^{k_T} E(I(X_i)) + \sum_{i=1}^{k_P} E(I(Z_i)) + 1 \\
&= k_I\lambda + k_U\alpha + k_P\alpha + 1,
\end{aligned}$$

where

$$I(w) = \begin{cases} 1, & \text{if variable } w \text{ is selected;} \\ 0, & \text{else.} \end{cases}$$

Because the average  $\bar{S}(\alpha)$  is an unbiased estimate of  $E(S(\alpha))$ , we have

$$\bar{S}(\alpha) \approx k_I\lambda + k_U\alpha + k_P\alpha + 1. \quad (3.7)$$

Similarly, the number of pseudo-variables selected  $S_P(\alpha)$  in Table 3.1 has expectation

$$E(S_P(\alpha)) = k_P\alpha,$$

which is approximated by  $\bar{S}_P$ , i.e.,

$$\bar{S}_P \approx k_P\alpha. \quad (3.8)$$

Therefore, under (3.7) and (3.8), the proportion  $\bar{\eta}_P(\alpha)$  given in (3.4) is such that

$$\bar{\eta}_P(\alpha) \approx \frac{k_P\alpha}{k_I\lambda + k_U\alpha + k_P\alpha + 1}. \quad (3.9)$$

Intuitively,  $\bar{\eta}_P(\alpha)$  should increase as  $k_P$  increases, and decrease as  $k_I$  increases. It is easily seen that the derived approximating formula of  $\bar{\eta}_P(\alpha)$  in equation (3.9) reflects these changes. Figure 3.2 exhibits four examples of the graphical view of goodness-of-fit to  $\bar{\eta}_P(\alpha)$  based on equation (3.9). From top to bottom and left to right, the four pictures show the graphs for the H1, H2, H3 and H4 models in our later simulations in Chapter 4 whose underlying model sizes are 2, 6, 10, and 14, respectively. In these examples,

the data sets have sample size 150 and 21 uncorrelated predictor variables (see Section 4.1 for details). According to the generation of data, the important predictor variables are specified with different regression coefficients leading to different contributions to the response variable. In each picture, the dotted curve indicates the average of  $\bar{\eta}_p(\alpha)$  taken over 100 replicated data sets, and the fitted solid curve is using the formula in equation (3.9) where the  $\lambda$  is replaced by its estimate from least squares fitting. Furthermore, Figure 3.3 shows another four examples with the same situations as in Figure 3.2 except that the predictors are correlated. From these graphs, we observe that the model from equation (3.9) fits very well, which lends support to the approximating formula for  $\bar{\eta}_p(\alpha)$  in equation (3.9).

According to the formula in equation (3.9), the  $\hat{\alpha}_0$  found by (3.5) should approximately satisfy

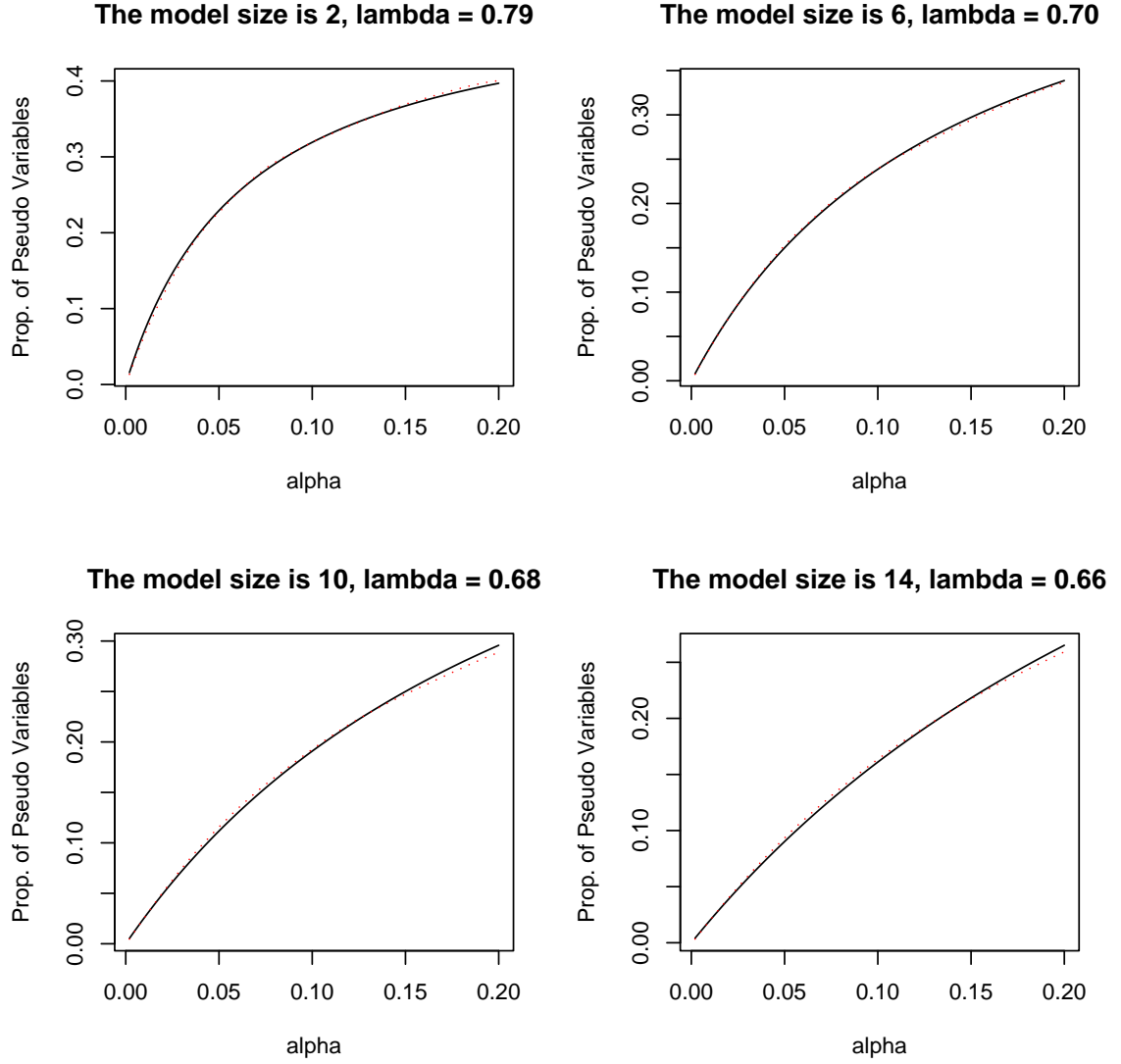
$$\frac{k_P \hat{\alpha}_0}{k_I \lambda + k_U \hat{\alpha}_0 + k_P \hat{\alpha}_0 + 1} \approx c. \quad (3.10)$$

When the  $\hat{\alpha}_0$  is used in a final forward selection on the original set of predictors, by the same argument for deriving equation (3.9), the model false selection rate  $\gamma$  in Definition 3.1 can be approximately expressed as

$$\begin{aligned} \gamma &\approx \frac{E(\# \text{ falsely selected unimportant original variables})}{E(\# \text{ total selected original variables})} \\ &\approx \frac{k_U \hat{\alpha}_0}{k_I \lambda + k_U \hat{\alpha}_0 + 1}. \end{aligned} \quad (3.11)$$

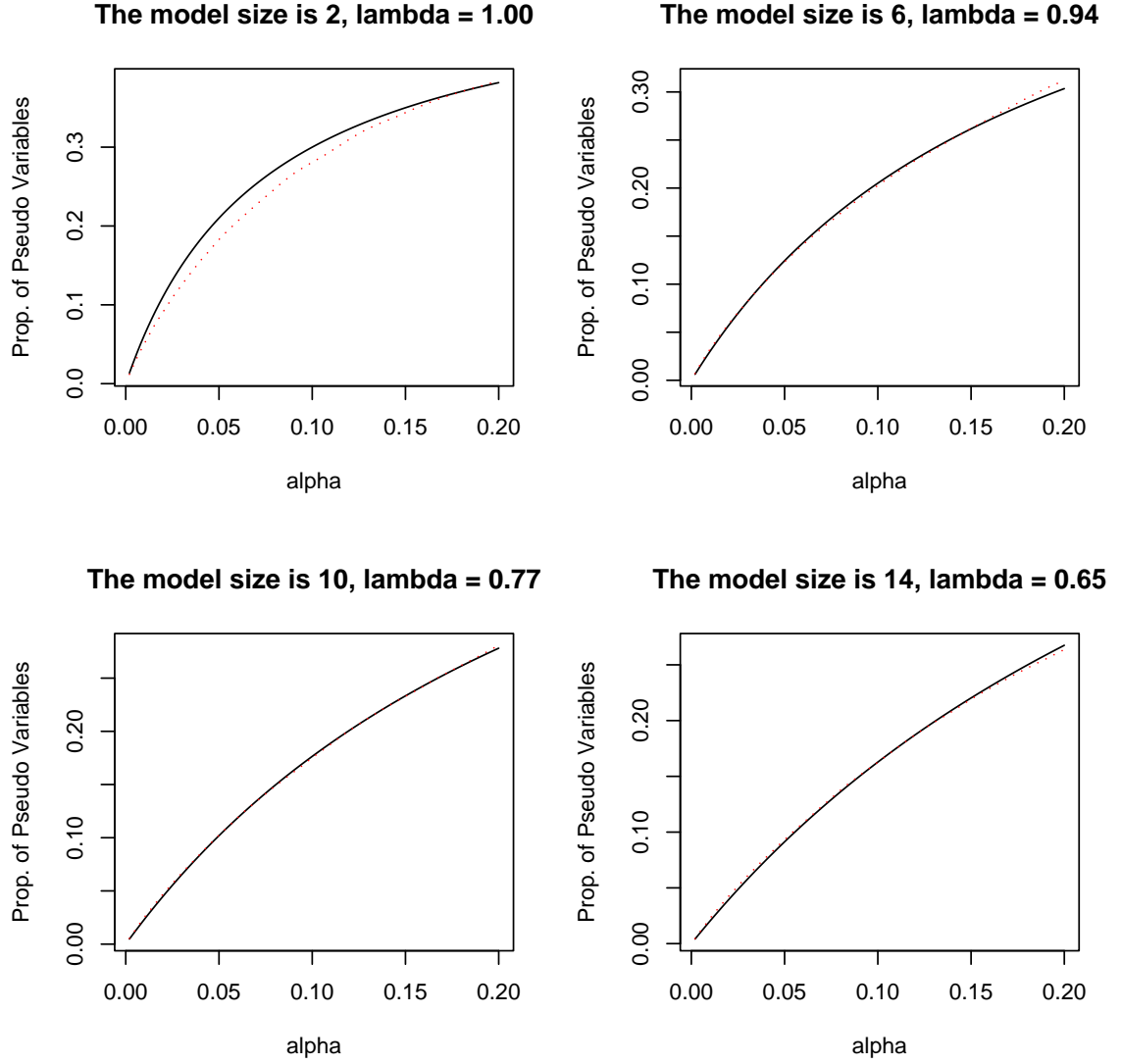
With the combination of the two equations (3.10) and (3.11), the following theorem building up the relationship between the cut-off  $c$  and the false selection rate  $\gamma$  can be stated.





$\cdots$  : the average of  $\bar{\eta}_p(\alpha)$  taken over 100 replicated simulated data sets.  
 $—$  : the fitted curve by (3.9) based on least squares estimate of  $\lambda$ .

Figure 3.2: Graphical goodness-of-fit to  $\bar{\eta}_p(\alpha)$  by formula in equation (3.9) when the predictors are uncorrelated



$\cdots$  : the average of  $\bar{\eta}_p(\alpha)$  taken over 100 replicated simulated data sets.  
 $—$  : the fitted curve by (3.9) based on least squares estimate of  $\lambda$ .

Figure 3.3: Graphical goodness-of-fit to  $\bar{\eta}_p(\alpha)$  by formula in equation (3.9) when the predictors are correlated

**Theorem 3.1** Assume that  $k_U > 0$  and  $0 < \hat{\alpha}_0 \leq 1$ . Under the equations (3.10) and (3.11), the cut-off  $c$  and the false selection rate  $\gamma$  have the relationship

$$c \approx \frac{\gamma k_P}{\gamma k_P + k_U}. \quad (3.12)$$

**Proof.** From the equation (3.11), we have

$$\begin{aligned} \frac{1}{\gamma} &\approx \frac{k_I \lambda + k_U \hat{\alpha}_0 + 1}{k_U \hat{\alpha}_0} \\ &\approx 1 + \frac{k_I \lambda + 1}{k_U \hat{\alpha}_0} \end{aligned} \quad (3.13)$$

On the other hand, the equation (3.10) can be rewritten as

$$c \approx \frac{\frac{k_P}{k_U}}{\frac{k_P}{k_U} + 1 + \frac{k_I \lambda + 1}{k_U \hat{\alpha}_0}} \quad (3.14)$$

Substituting (3.13) into (3.14) produces

$$\begin{aligned} c &\approx \frac{\frac{k_P}{k_U}}{\frac{k_P}{k_U} + \frac{1}{\gamma}} \\ &\approx \frac{\gamma k_P}{\gamma k_P + k_U} \end{aligned}$$

♠

With this theorem we are able to resolve our two problems: determination of  $c$  and  $k_P$ . Theorem 3.1 points out that the cut-off  $c$  can be defined in terms of the model false selection rate  $\gamma$  based on the equation (3.12). Then, with a target false selection rate  $\gamma$  specified, e.g.,  $\gamma = 0.05$ , the value of the cut-off  $c$  can be determined. This process is similar to finding the rejection region in hypothesis testing. In that situation, we specify a test size, and then find the corresponding test rejection region. Notice that, in equation (3.12), the value of  $k_U$  is unknown, which gives us a challenge to determine  $c$ . In an attempt to handle this difficulty, we consider an iteration procedure. First, we assume

that all of the real predictors are unimportant, and hence  $k_U$  should equal to  $k_T$  that is known. An initial value of  $c$ , say  $c^{(0)}$ , can be obtained from equation (3.12) with  $k_U$  replace by  $k_T$  and used as a threshold for  $\bar{\eta}_p(\alpha)$  that allows us to tune the  $\alpha$  and select the model. Naturally, the size of the model selected can be used as an estimate of  $k_I$ , denoted by  $\hat{k}_I^{(1)}$ , which yields an estimate of  $k_U$  as,  $\hat{k}_U^{(1)} = k_T - \hat{k}_I^{(1)}$ . Then, replacing  $k_U$  with  $\hat{k}_U^{(1)}$  in (3.12) produces an updated value of  $c$ , say  $c^{(1)}$ , and we use it to tune the  $\alpha$  again. Based on the selected model, we find an updated estimate of  $k_U$ ,  $\hat{k}_U^{(2)}$ , which leads to a new  $c$ , say  $c^{(2)}$ . This process is repeated until there is no change in the estimate of  $k_U$ . Notice that at each iteration the estimate of  $k_U$ ,  $\hat{k}_U^{(i)}$ , is derived by subtracting the size of the selected model,  $\hat{k}_I^{(i)}$ , from the total number of real variables,  $k_T$ , i.e.,  $\hat{k}_U^{(i)} = k_T - \hat{k}_I^{(i)}$ . On the other hand, starting with  $\hat{k}_I^{(0)} = 0$ ,  $\hat{k}_I^{(i)}$  is always no less than  $\hat{k}_I^{(i-1)}$  (i.e.,  $\hat{k}_I^{(i)} \geq \hat{k}_I^{(i-1)}$ ), because the cut-off  $c$  is adjusted up at each iteration which results in a larger  $\hat{\alpha}_0$ . Thus, we have  $\hat{k}_U^{(i)} \leq \hat{k}_U^{(i-1)}$ , indicating that  $\hat{k}_U^{(i)}$  is monotonously decreasing. In addition,  $\hat{k}_U^{(i)}$  will never go below 0. Therefore, we anticipate that the convergence in  $\hat{k}_U^{(i)}$  will occur very quickly, at most after  $k_T$  iterations. Actually, in our later simulation studies,  $\hat{k}_U^{(i)}$  converges in just 2 or 3 iterations. By specifying a false selection rate  $\gamma$ , the cut-off  $c$  is initialized and adjusted for finding the optimal  $\alpha$ . Because we are targeting the false selection rate for variable selection, we call this method the “False Selection Rate (FSR) Procedure.” The procedure is fully detailed in Algorithm 4.1 on the next page.

---

**Algorithm 4.1 False Selection Rate (FSR) Procedure**


---

1. Pick a target false selection rate  $\gamma$ , e.g.,  $\gamma = 0.05$ .
2. Generate sets of  $k_P$  pseudo-variables  $B$  times. For  $\alpha$  in  $\mathcal{A}$ , i.e.,  $\mathcal{A} = \{0.002, 0.004, \dots, 0.2\}$ , count the average number of pseudo-variables selected,  $\bar{S}_P(\alpha)$ , and the average number of total selected variables,  $\bar{S}(\alpha)$ , and calculate

$$\bar{\eta}_P(\alpha) = \frac{\bar{S}_P(\alpha)}{\bar{S}(\alpha)}.$$

3. Obtain an initial value of cut-off,  $c^{(0)}$ , from the formula

$$c^{(0)} = \frac{\gamma k_P}{\gamma k_P + k_T}.$$

Determine  $\hat{\alpha}_0^{(0)}$  as follows:

$$\hat{\alpha}_0^{(0)} = \text{Min}\{\alpha : \bar{\eta}_P(\alpha) \geq c^{(0)}, \alpha \in \mathcal{A}\}.$$

4. Run forward selection on the original set of variables (no pseudo-variables) using  $\hat{\alpha}_0^{(0)}$ . Denote the size of the model selected by  $\hat{k}_I^{(0)}$  and set  $\hat{k}_U^{(0)} = k_T - \hat{k}_I^{(0)}$ .
5. Update the cut-off by

$$c^{(1)} = \frac{\gamma k_P}{\gamma k_P + \hat{k}_U^{(0)}},$$

and then find

$$\hat{\alpha}_0^{(1)} = \text{Min}\{\alpha : \bar{\eta}_P(\alpha) \geq c^{(1)}, \alpha \in \mathcal{A}\}.$$

6. Go back to Step 4 and iterate until there is no change in  $\hat{k}_U^{(i)}$ . The final  $\hat{\alpha}_0^{(i)}$  is used in a final forward selection on the original set of data.
-

As for choosing the  $k_P$ , Theorem 3.1 also indicates that the number of pseudo-variables  $k_P$  should not matter much because the cut-off  $c$  is simultaneously adjusted for tuning  $\alpha$  as  $k_P$  changes. Specifically, when we add more (or less) pseudo-variables, the curve of  $\bar{\eta}_P$  will become steeper (or flatter), but larger (or smaller) values of the cut-off  $c$  are obtained from equation (3.12). As a result, we may expect little change in the  $\hat{\alpha}_0$ . For example, in Figure 3.4, the top curve (denoted by “ $\star$ ”) stands for  $k_P = 42$ , the middle curve (denoted by “ $+$ ”) represents  $k_P = 21$ , and the bottom curve (denoted by “ $.$ ”) represents  $k_P = 10$ . The underlying data set includes 21 predictors among which 6 variables are important. The final cut-off  $c$  after iterations is 0.116 for  $k_P = 42$ , 0.062 for  $k_P = 21$ , and 0.03 for  $k_P = 10$ . We see that the three choices of  $k_P$  lead to almost identical  $\hat{\alpha}_0$ , i.e.,  $\hat{\alpha}_0 = 0.024$  when  $k_P = 42$ , and  $\hat{\alpha}_0 = 0.026$  when  $k_P = 10$  and 21. In our later simulation studies in Section 4.2, robustness of the FSR method to the  $k_P$  value is further verified.

### 3.3 Pseudo-Variables Generation

The basic requirement for the pseudo-variables is that they should be independent of the original response variable. This is to guarantee that the pseudo-variables are “unimportant,” and that their proportion in the model selected truly reflects the variable selection method’s tendency to overfit or underfit the model. We propose four ways of generating pseudo-variables:

**Method 1.** Generate independently from the standard normal distribution  $N(0,1)$ .

**Method 1a.** Regress each of the pseudo-variables generated in Method 1 on all the original predictor variables linearly, and use the regression residuals as the new pseudo-variables. This forces correlations to be zero.

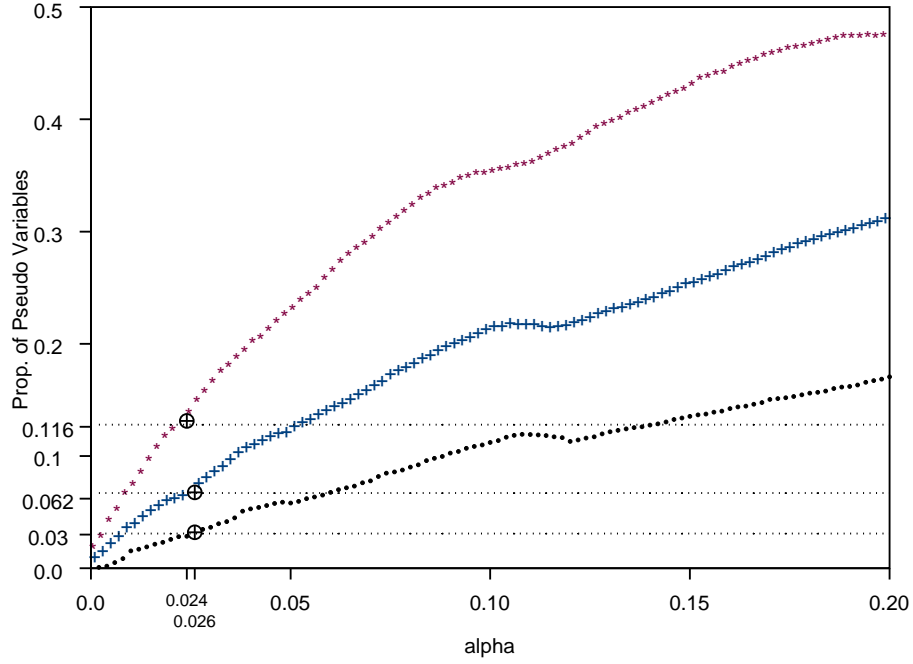


Figure 3.4: Tuning results for different  $k_P$  ( $k_T = 21$ )  
“\*\*\*”:  $k_P = 42$  , “+++”:  $k_P = 21$  , “...”:  $k_P = 10$

**Method 2.** Permute the rows of the design matrix  $\mathbf{X}$ , where  $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_{k_T})_{n \times k_T}$ , and the pseudo-variables are the columns of the permuted matrix.

**Method 2a.** Same as in Method 1a except that the initial pseudo-variables are generated as in Method 2.

Method 1 is the simplest way to create any number of pseudo-variables, but a drawback is that the variables are normally distributed. With real data the true predictor variables often are not approximately normal. Hence, the large difference in the variable distribution might be a concern in regression. Method 2 generates the pseudo-variables by permutation. Randomly permuting the rows of the true design matrix creates a set

of pseudo-variables that are independent of the response yet have the same marginal distribution as the true predictor variables. This makes permutation attractive. However, it can happen by chance that the pseudo-variables generated in Method 1 or 2 are correlated with the original predictor variables, especially when the sample size is small. Thus, forward selection might replace the positions of important real variables with their correlated pseudo-variables. This will influence the FSR procedure and cause it to be misleading. Methods 1a and 2a eliminate this possibility. As we know, in linear regression, the residual is always orthogonal to the regressors. After replacing the initial pseudo-variable with its residual from linear regression on all the original predictors, we eliminate the linear relationship between the pseudo and the original variables, and reduce the variation produced in generation of the pseudo-variables. We will make a comparison among the four methods in later simulation studies in Chapter 4.

### 3.4 Some Comments

When the number of candidate predictors is large and there are many weak important variables, it is very possible in forward selection that the unimportant variables enter the model before some important ones due to random variation. It is impossible for any stopping criteria to exclude all the unimportant variables from the model while all the important are included. Obviously, it is not wise to keep adding variables until all the important variables are included, because there will be too many unimportant variables in the model. The goal is therefore to select as many important variables as possible, while incurring a relatively low proportion of false important ones. Of course, we do not know which of the original predictors are unimportant, and thus cannot evaluate the false selection rate  $\gamma$  directly. However, by Theorem 3.1,  $\gamma$  is connected with the cut-off



$c$  that is used as a threshold for the average proportions  $\bar{\eta}_p(\alpha)$ . Our selection procedure tunes  $\alpha$  to control  $\bar{\eta}_p(\alpha)$  based on the  $c$ , and thus is trying to achieve the target false selection rate, balancing the number of truly important predictors and the number of false important ones in the selected model.

Usually,  $\bar{\eta}_p(\alpha)$  is an increasing function of  $\alpha$ , and its rate of increase depends on the underlying model size. When the true model size is large, i.e.,  $k_I$  is large,  $\bar{\eta}_p(\alpha)$  is flatter because the pseudo-variables tend to enter the model only for large  $\alpha$ . For small true models it is steeper. This property ensures that the cut-off  $c$  is self-correcting for the true model size. That is, keeping the same  $c$ , we tend to obtain small estimates  $\hat{\alpha}_0$  for small size models, and large  $\hat{\alpha}_0$  when the model size increases. In our method, we also do some adjustment for the  $c$  based on the estimate of the number of unimportant real predictors  $\hat{k}_U$ . The effect of this procedure is to raise  $\hat{\alpha}_0$  a bit when we suspect  $k_I$  is large in order to keep the selected model from being underfitted. Otherwise, it does not change much. These features are illustrated in Figure 3.5. In these examples, there are 21 candidate predictors in the underlying data sets, and  $\bar{\eta}_p(\alpha)$  are calculated based on generating 21 pseudo-variables 500 times. The top picture refers to the model whose true size is 2, i.e., there are only 2 important predictor variables, whereas the bottom picture stands for the model whose true size is 14. Clearly,  $\bar{\eta}_p(\alpha)$  in the top graph is steeper than those in the bottom graph. It is also easily seen that, when the true model size is 2, the final cut-off  $c^{(i)}$  is adjusted only a small amount from its initial value  $c^{(0)}$  (i.e.,  $c^{(0)} = 0.048, c^{(i)} = 0.053$ ), and the same  $\hat{\alpha}_0$  (i.e.,  $\hat{\alpha}_0 = 0.008$ ) is determined for both values of  $c$ . However, when the true model size is 14, the final cut-off  $c^{(i)}$  is almost doubled (i.e.,  $c^{(0)} = 0.048, c^{(i)} = 0.081$ ), correspondingly, the  $\hat{\alpha}_0$  changes from 0.024 to 0.042. In these graphs, the points marked by  $\oplus$  indicate the  $\bar{\eta}_p(\hat{\alpha}_0)$  satisfying the tuning criteria (3.6).

The FSR procedure depends only on the forward selection strategy and generation

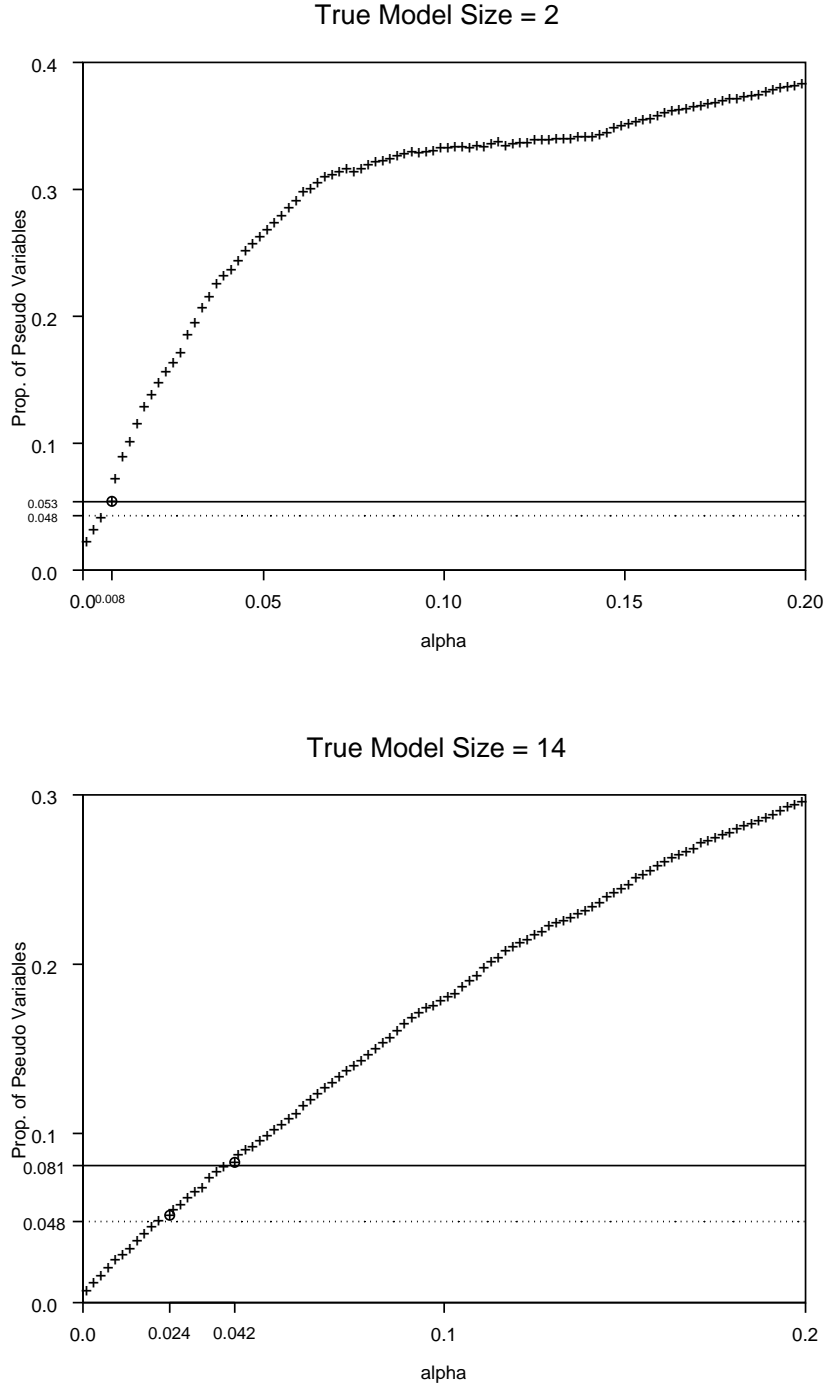


Figure 3.5: Illustration of adjustment of the cut-off  $c$  for different true model sizes.

+++ :  $\bar{\eta}_P(\alpha)$ , proportions of pseudo-variables falsely selected  
... : The dotted line corresponds to the initial  $c^{(0)}$   
— : The solid line corresponds to the final adjusted  $c^{(i)}$

of pseudo-variables. In particular it is not specific to linear models. As a result, it can be extended to other regression models, such as poisson regression, logistic regression, proportional hazards regression, etc. Besides forward selection, other stepwise features, like backward elimination, can also be employed for the FSR procedure.

# Chapter 4

## Simulation – In Linear Regression Models

Due to the difficulty of obtaining analytical results, we study our approach by Monte Carlo and compare it with a variety of commonly used selection procedures. In this Chapter, we consider the common linear regression models. For simplicity of notation, we will use FSR to refer to our false selection rate procedure in the figures and tables in the future. The SAS is used to implement the method.

### 4.1 Simulation Design

Luo et al. (2003) developed a comprehensive simulation design in studying the performance of their estimators, to be called Tune. Their design in turn was based on one used by Tibshirani and Knight (1999). Our design uses the same simulated data sets from Luo et al. (2003) in order to facilitate comparison with their results. Each data set contains 21 predictors, generated from a multivariate normal distribution with mean 0 and an autoregressive covariance structure where the covariance between any two predictors  $x_i$  and  $x_j$  is equal to  $\rho^{|i-j|}$ , with  $\rho = 0$  and 0.7, respectively. The generated design matrix  $\mathbf{X}$  is then held fixed for all runs. The response variable is generated from the linear

regression model  $y_i = \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i$ ,  $i = 1, \dots, n$ , where  $\mathbf{x}_i$  is the  $i$ th row of the design matrix  $\mathbf{X}$ ,  $\boldsymbol{\beta}$  is the  $21 \times 1$  regression coefficients vector, and the random errors  $\epsilon_i$  independently follow a standard normal distribution  $N(0, 1)$ . The coefficients  $\boldsymbol{\beta}$  are a little complex and so will be described in the following two stages.

1. The nonzero coefficients are clustered around two variables:  $x_7$  and  $x_{14}$ , with their initial values given by

$$\beta_{7+j} = (h-j)^2, \quad |j| < h \quad (4.1)$$

$$\beta_{14+j} = (h-j)^2, \quad |j| < h \quad (4.2)$$

where  $h$  has values 1, 2, 3, and 4. All other coefficients are zero.

2. The coefficients are then multiplied by a common constant to make the theoretical model  $R^2$  equal to 0.75, where theoretical  $R^2$  is defined as

$$R^2 = \frac{(\mathbf{X}\boldsymbol{\beta})^T(\mathbf{X}\boldsymbol{\beta})}{(\mathbf{X}\boldsymbol{\beta})^T(\mathbf{X}\boldsymbol{\beta}) + n\sigma^2}. \quad (4.3)$$

The four different values of  $h$  in equations (4.1) and (4.2) result in models with 2, 6, 10 and 14 nonzero coefficients. We designate these four different sets of coefficients by H1, H2, H3 and H4. In each model, the nonzero coefficients are specified with different values such that the associated variables make different contribution to the response variable. Figure 4.1 shows the values of regression coefficients,  $\beta_1, \beta_2, \dots, \beta_{21}$ , for different models for the case of  $n = 150$  and  $\rho = 0$ . In other cases with different sample size and value  $\rho$  a similar pattern can be observed. For model H1, there are two strong variables. At the other extreme, model H4, each cluster contains at least two weak variables. In addition, we also include the null model for which all coefficients are zero, denoted by H0. Each data set has either 50, 150 and 500 observations. In each run of combination of sample size  $n$  and  $\rho$  there are 100 repetitions for each H model.

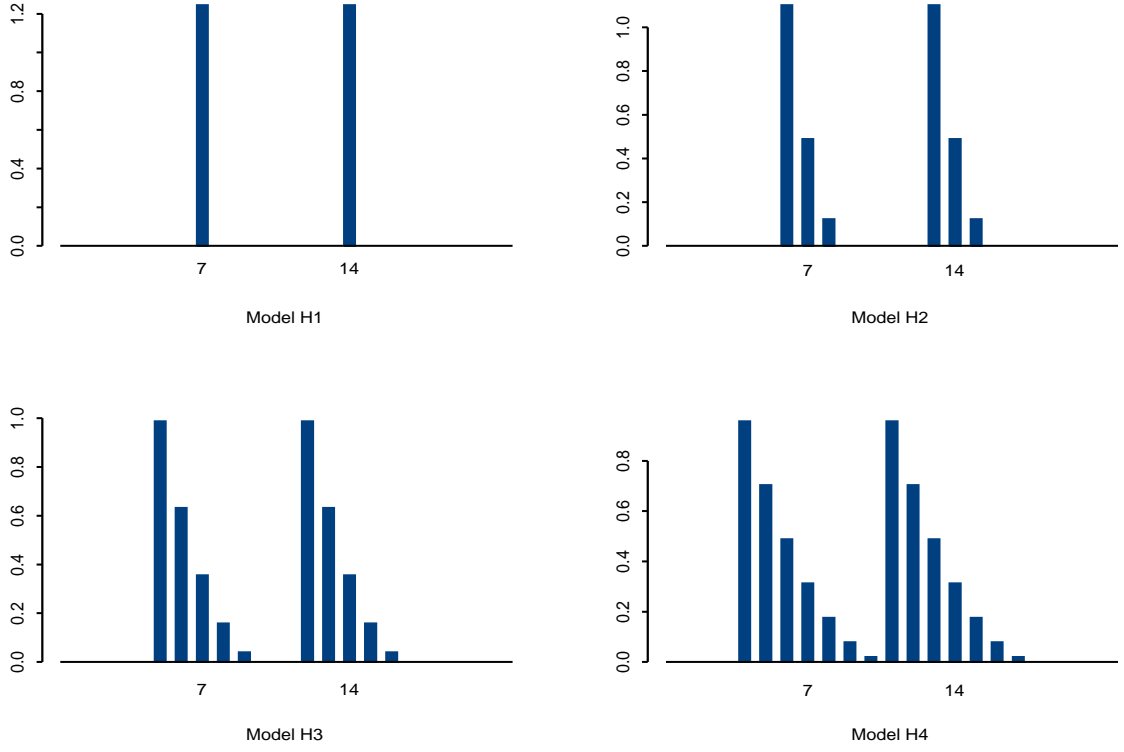


Figure 4.1: Values of regression coefficients,  $\beta_1, \beta_2, \dots, \beta_{21}$ , ( $n = 150, \rho = 0$ )

For selection method evaluation, i.e., how good a model does it select, the average model error for the selected model, where the model error is defined in equation (2.3) in Chapter 2, is used as the primary criterion (the smaller, the better). Average model sizes are also presented to assess how well the selection procedure identifies the correct dimensionality.

## 4.2 Factors Affecting the FSR Procedure

Recall that the FSR procedure tunes  $\alpha$  by adding pseudo-variables to the real set of predictors. Thus, the generation of pseudo-variables plays an important role in this procedure, and the number of pseudo-variables and their distribution are of interest. Also recall that the model false selection rate  $\gamma$  is the target control of our method. It is of interest to know whether or not the method achieves its specified target and the effect of different choices of  $\gamma$  on model selection. These aspects will be studied in the following Monte Carlo simulations.

### 4.2.1 Methods for Generating Pseudo-Variables

As was discussed in Section 3.3, there are four ways of generating pseudo-variables: **Method 1**, **1a**, **2**, and **2a**. Method 1 and 2 generate pseudo variables by drawing independently from  $N(0, 1)$  and permuting the rows of design matrix, respectively. Method 1a then takes the residuals after regressing each of the pseudo-variables generated in Method 1 to all the original variables linearly. Similarly, Method 2a also takes the residuals but based on pseudo-variables from Method 2. In this section, the four methods are studied. We choose data sets having sample sizes 50 and 150. Without loss of generality, we generate a set of 21 pseudo-variables, equal to the same number as that of the  $X$  variables (i.e.,  $k_P = k_T = 21$ ), and repeat the generation 500 times (i.e.,  $B = 500$ ). The target  $\gamma$  is set to be 0.05.

Tables 4.1 – 4.4 exhibit the results of average model error and average model size. In these tables, the numbers in parentheses are standard errors, and “Method” indicates the ways of generating pseudo-variables. As for the average model error, clearly, when sample size is 150, there is little difference observed among these four methods. However, when

sample size goes down to 50, we see that, although Methods 1a and 2a increase the average model error a little for H0 and H1 models, they reduce the average model error to a large extent for H3 and H4 models, especially when  $\rho = 0$ . It is obvious that, by using Methods 1a and 2a, the gain for models with large model size fairly exceeds the loss for models with small model size. In fact, the bad performance of Methods 1 and 2 for small sample size may be caused by the possible correlations among the pseudo-variables and original variables, while Methods 1a and 2a eliminate the correlations. Reflected in the average model size, Method 1a and 2a tend to overestimate model size, resulting in overfitting. In addition, for Method 2a, the pseudo-variables are generated based on permuting the whole row of the design matrix. This results in the same correlation structure among the original predictor variables retained by the set of pseudo-variables. Therefore, we would anticipate that Method 2a should outperform Method 1a when predictors are correlated, because Method 1a can only generate independent pseudo-variables no matter whether or not the original predictor variables are correlated. This is seen in the simulation results. In Tables 4.1 and 4.3, we find that when  $X$  variables are correlated, i.e.,  $\rho = 0.7$ , Method 2a produces smaller average model errors for H3 and H4 models than Method 1a. Corresponding to the average model size, Method 2a overfits relative to Method 1a. Consequently, we recommend using Method 2a to generate pseudo-variables. Unless indicated otherwise, Method 2a will be used to generate pseudo-variables in all of the remaining simulations.



Table 4.1: Average model error for the FSR method with  $\gamma = 0.05$  for the four ways of generating pseudo-variables, ( $n = 150$ ,  $k_T = 21$ , and  $k_P = 21$ )

$\rho$	Method	H0	H1	H2	H3	H4
0	1	0.011(0.002)	0.030(0.003)	0.076(0.004)	0.105(0.004)	0.136(0.005)
	1a	0.013(0.003)	0.031(0.003)	0.077(0.004)	0.106(0.004)	0.135(0.005)
	2	0.012(0.002)	0.029(0.003)	0.076(0.004)	0.105(0.004)	0.137(0.005)
	2a	0.014(0.003)	0.031(0.003)	0.078(0.004)	0.106(0.004)	0.134(0.005)
0.7	1	0.009(0.002)	0.030(0.003)	0.086(0.005)	0.113(0.005)	0.148(0.006)
	1a	0.010(0.002)	0.031(0.003)	0.085(0.005)	0.113(0.005)	0.141(0.006)
	2	0.010(0.002)	0.032(0.003)	0.087(0.005)	0.110(0.005)	0.136(0.006)
	2a	0.010(0.002)	0.033(0.003)	0.085(0.005)	0.109(0.004)	0.134(0.006)

Table 4.2: Average model size for the FSR method with  $\gamma = 0.05$  for the four ways of generating pseudo-variables, ( $n = 150$ ,  $k_T = 21$ , and  $k_P = 21$ )

$\rho$	Method	H0	H1	H2	H3	H4
0	1	0.09(0.03)	2.18(0.05)	4.63(0.09)	7.28(0.11)	9.74(0.14)
	1a	0.11(0.04)	2.21(0.05)	4.70(0.09)	7.36(0.13)	9.84(0.14)
	2	0.09(0.03)	2.17(0.04)	4.65(0.09)	7.24(0.11)	9.73(0.14)
	2a	0.12(0.04)	2.20(0.05)	4.74(0.09)	7.34(0.12)	9.83(0.14)
0.7	1	0.06(0.02)	2.15(0.05)	4.00(0.08)	5.37(0.08)	5.93(0.11)
	1a	0.08(0.03)	2.16(0.05)	4.15(0.08)	5.48(0.09)	6.13(0.11)
	2	0.07(0.03)	2.19(0.05)	4.14(0.08)	5.61(0.10)	6.25(0.11)
	2a	0.08(0.03)	2.21(0.05)	4.20(0.08)	5.72(0.11)	6.36(0.10)

Table 4.3: Average model error for the FSR method with  $\gamma = 0.05$  for the four ways of generating pseudo-variables, ( $n = 50$ ,  $k_T = 21$ , and  $k_P = 21$ )

$\rho$	Method	H0	H1	H2	H3	H4
0	1	0.033(0.005)	0.102(0.011)	0.299(0.015)	0.447(0.021)	0.759(0.042)
	1a	0.054(0.009)	0.129(0.013)	0.294(0.016)	0.403(0.019)	0.491(0.018)
	2	0.036(0.005)	0.101(0.011)	0.303(0.015)	0.446(0.021)	0.775(0.042)
	2a	0.053(0.009)	0.129(0.013)	0.293(0.016)	0.405(0.018)	0.488(0.018)
0.7	1	0.025(0.004)	0.103(0.011)	0.299(0.011)	0.382(0.016)	0.406(0.015)
	1a	0.037(0.007)	0.122(0.012)	0.274(0.012)	0.346(0.016)	0.377(0.015)
	2	0.026(0.004)	0.108(0.012)	0.298(0.012)	0.369(0.016)	0.399(0.015)
	2a	0.041(0.008)	0.129(0.013)	0.271(0.013)	0.347(0.016)	0.367(0.014)

Table 4.4: Average model size for the FSR method with  $\gamma = 0.05$  for the four ways of generating pseudo-variables, ( $n = 50$ ,  $k_T = 21$ , and  $k_P = 21$ )

$\rho$	Method	H0	H1	H2	H3	H4
0	1	0.08(0.03)	2.27(0.06)	3.61(0.10)	4.51(0.12)	4.67(0.18)
	1a	0.24(0.06)	2.52(0.09)	4.08(0.11)	5.33(0.14)	6.60(0.15)
	2	0.10(0.03)	2.27(0.06)	3.60(0.10)	4.56(0.12)	4.63(0.19)
	2a	0.23(0.06)	2.52(0.09)	4.07(0.11)	5.31(0.15)	6.62(0.15)
0.7	1	0.01(0.01)	2.23(0.05)	2.60(0.07)	3.57(0.10)	2.95(0.10)
	1a	0.11(0.05)	2.37(0.07)	2.87(0.08)	4.01(0.11)	3.46(0.13)
	2	0.02(0.02)	2.26(0.05)	2.70(0.08)	3.73(0.11)	3.02(0.10)
	2a	0.16(0.07)	2.44(0.07)	3.02(0.10)	4.07(0.11)	3.63(0.13)

### 4.2.2 The Choice of $k_P$

In this section, we report results from a simulation study designed to study the sensitivity of the FSR procedure to choice of  $k_P$ . In Section 3.2, we have discussed heuristically that the number of generated pseudo-variables should not matter for tuning  $\alpha$ . Here, we exhibit this by simulation studies. We chose  $n = 150$  sample size data sets and three values of  $k_P$ : 10, 21 and 42. As was recommended in the last section, Method 2a is used for generation of pseudo-variables, and we repeat generation 500 times, keeping the false selection rate  $\gamma = 0.05$ . The results of average model error and average model size are displayed in Tables 4.5 and 4.6. The three values of  $k_P$  produce nearly identical results, indicating that the FSR procedure is robust to the number of generated pseudo-variables. Henceforth we use  $k_P = k_T$ .

Table 4.5: Average model error for the FSR method with  $\gamma = 0.05$  when choosing different  $k_P$ , ( $n = 150$  and  $k_T = 21$ )

$\rho$	$k_P$	H0	H1	H2	H3	H4
0	10	0.014(0.003)	0.030(0.003)	0.077(0.004)	0.106(0.004)	0.135(0.005)
	21	0.014(0.003)	0.031(0.003)	0.078(0.004)	0.106(0.004)	0.134(0.005)
	42	0.014(0.003)	0.030(0.003)	0.077(0.004)	0.106(0.004)	0.135(0.005)
0.7	10	0.011(0.002)	0.033(0.003)	0.084(0.005)	0.109(0.004)	0.137(0.006)
	21	0.010(0.002)	0.033(0.003)	0.085(0.005)	0.109(0.004)	0.134(0.006)
	42	0.011(0.002)	0.033(0.003)	0.085(0.005)	0.109(0.004)	0.134(0.006)

### 4.2.3 The Effect of Target False Selection Rate $\gamma$

In our selection procedure, we determine the value of cut-off  $c$  by specifying a target model false selection rate  $\gamma$ . We anticipate that increasing  $\gamma$  will lead to a larger value  $c$

Table 4.6: Average model size for the FSR method with  $\gamma = 0.05$  when choosing different  $k_P$ , ( $n = 150$  and  $k_T = 21$ )

$\rho$	$k_P$	H0	H1	H2	H3	H4
0	10	0.11(0.04)	2.19(0.05)	4.71(0.09)	7.32(0.12)	9.82(0.14)
	21	0.11(0.04)	2.20(0.05)	4.75(0.09)	7.34(0.12)	9.81(0.14)
	42	0.12(0.04)	2.19(0.05)	4.74(0.09)	7.33(0.12)	9.81(0.14)
0.7	10	0.07(0.03)	2.16(0.05)	4.15(0.08)	5.47(0.09)	6.11(0.11)
	21	0.08(0.03)	2.17(0.05)	4.12(0.08)	5.45(0.09)	6.11(0.11)
	42	0.08(0.03)	2.16(0.05)	4.08(0.08)	5.48(0.09)	6.19(0.10)

and hence to a larger  $\hat{\alpha}_0$ , and thus there will be more variables included into the model. On the other hand, reducing  $\gamma$  will result in less variables selected. To look at the effect on model selection when moving  $\gamma$  up or down, we ran the sample size 150 cases using  $\gamma = 0.01, 0.05$ , and  $0.1$ . Figures 4.2 and 4.3 show plots of average model error and average model size versus model. Examination of Figure 4.3 reveals that, as was expected, average model size increases as  $\gamma$  increases. In Figure 4.2 we see that, for H0 and H1 models, average model error increases as  $\gamma$  increases due to overfitting. On the other hand, for H3 and H4 models, larger  $\gamma$  tends to keep the model from being underfitted, and hence reduces average model error. Among the three values,  $\gamma = 0.05$  shows the best global performance, because in this case the prediction accuracy is not reduced much for models that have small underlying model size (e.g., H0 and H1 models) compared to  $\gamma = 0.01$ , while it gets close to the prediction accuracy when  $\gamma = 0.1$  for large size models (e.g., H3 and H4 models). In our later simulation studies, we will use  $\gamma = 0.05$ . In practice, the users have the flexibility to choose  $\gamma$ , controlling what percentage of the variables selected is not important.

#### 4.2.4 The Monte Carlo Estimated False Selection Rate

The target of the FSR method is the model false selection rate. The major task is to keep the selection procedure from including too many unimportant variables in the model while identifying as many important variables as possible. We want to check whether or not the procedure achieve this goal. Figure 4.4 shows plots of Monte Carlo estimated model false selection rates,  $\hat{\gamma}$ , when the  $\gamma = 0.05$  for different sample sizes (i.e.,  $n = 50, 150$ , and  $500$ ). The top picture gives the estimated false selection rate when  $\rho = 0$ , and the bottom one gives the results when  $\rho = 0.7$ . In both graphs, the horizontal solid lines correspond to  $0.05$ . The appealing result is that the estimated false selection rates,  $\hat{\gamma}$ , are quickly approaching the target as the sample sizes increase, and get very close to  $0.05$  when the sample sizes are up to  $500$ .

#### 4.2.5 The Performance of $\hat{\alpha}_0$

The FSR procedure tunes the significance level  $\alpha$  in forward selection such that a good bias-variance trade-off can be obtained. Intuitively, if the underlying true model size is large, we would anticipate a large  $\alpha$  level being specified. The underlying idea is that, with more important variables competing to enter the model, any particular one has less chance of selection. In this situation, a large  $\alpha$  level should be used to give each important variable a large chance of being selected such that no important variables is missed. Otherwise, the method will have small power in identifying the important variables, and the selected model tends to be underfitted. Figure 4.5 gives the boxplots of  $\hat{\alpha}_0$  found by the FSR procedure under different models for different sample sizes and  $\rho$ . On average, the values of  $\hat{\alpha}_0$  increase as the underlying model size becomes large. Thus the FSR procedure tunes  $\alpha$  as expected.

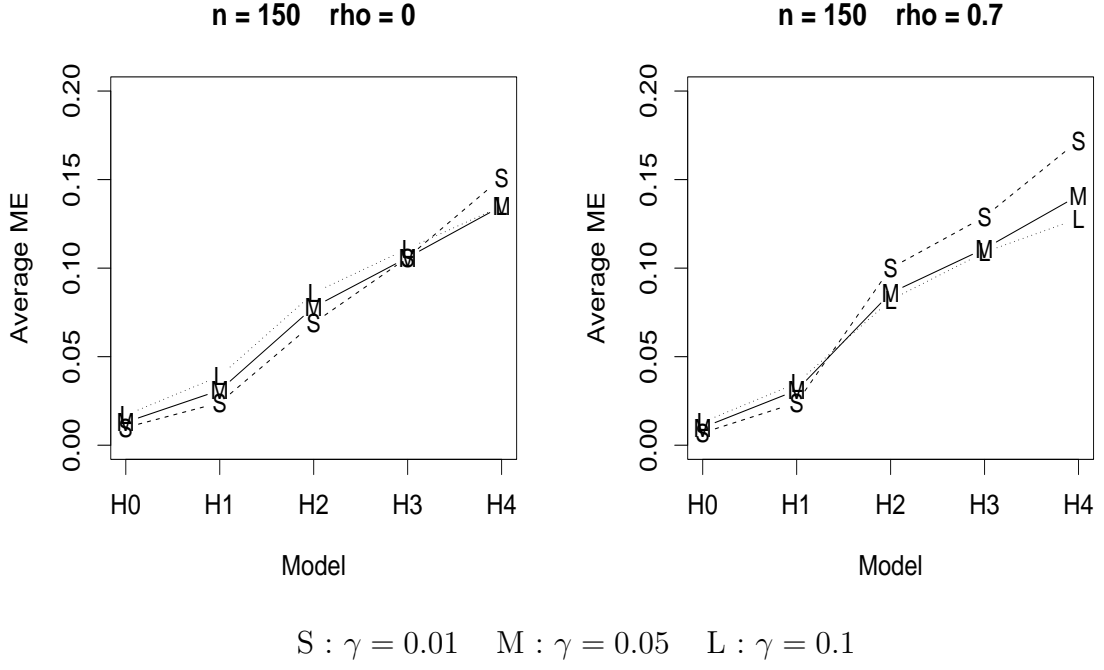


Figure 4.2: Average model error for the FSR method with different values  $\gamma$  ( $n = 150$ ,  $k_T = 21$ , and  $k_P = 21$ )

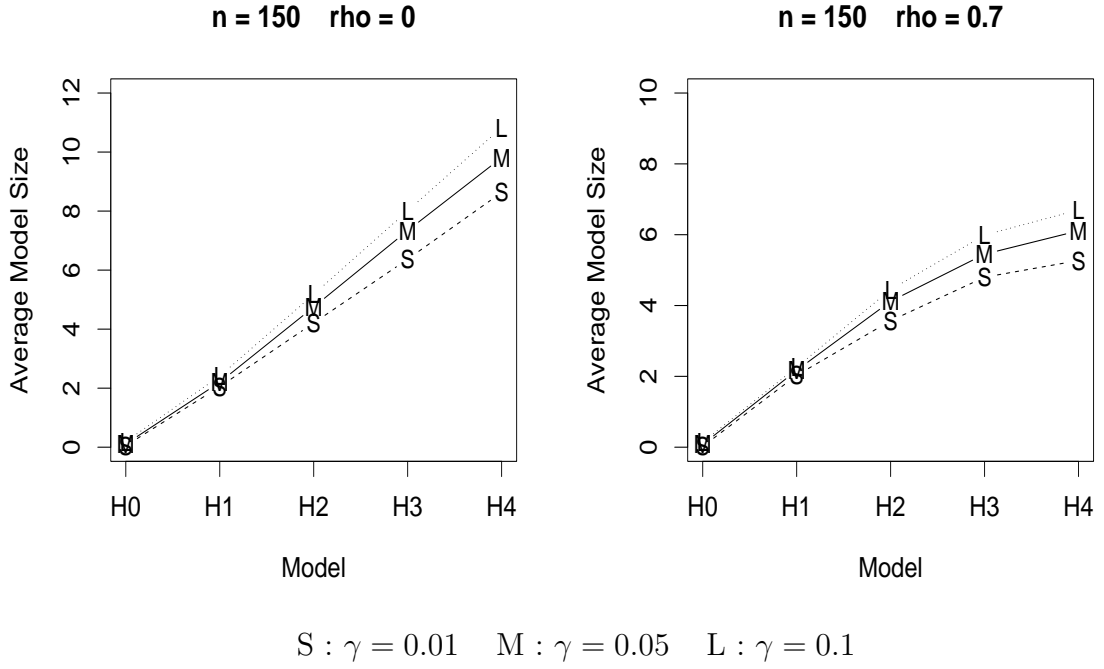
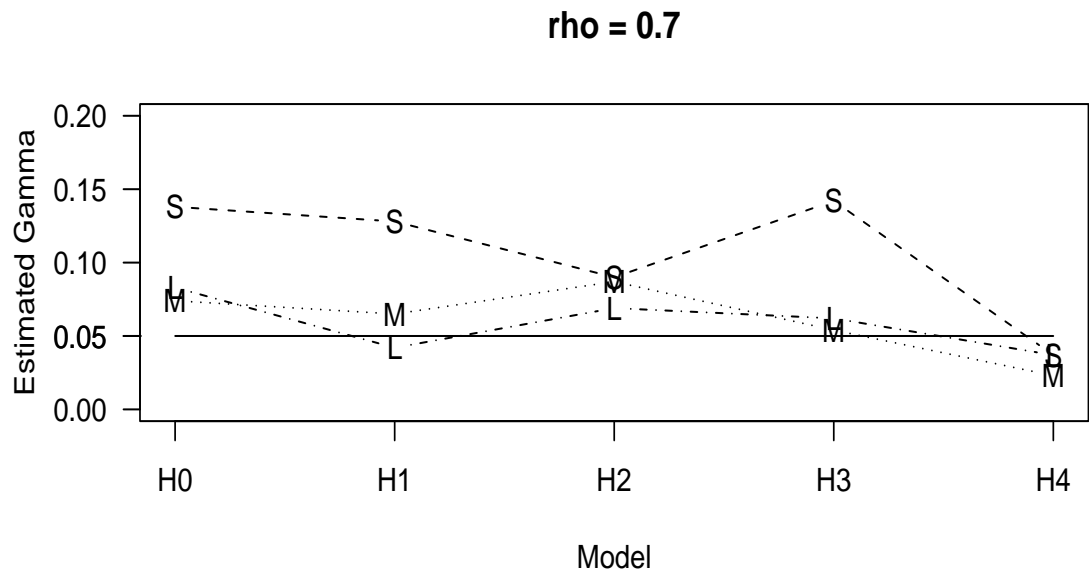
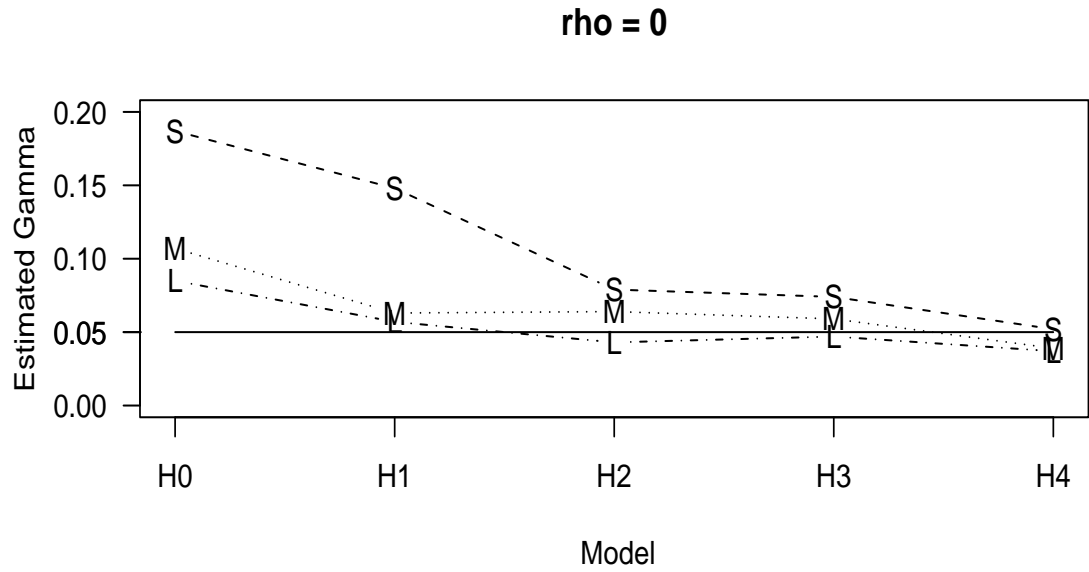


Figure 4.3: Average model size for the FSR method with different values  $\gamma$  ( $n = 150$ ,  $k_T = 21$ , and  $k_P = 21$ )



S :  $n = 50$     M :  $n = 150$     L :  $n = 500$

Figure 4.4: The Monte Carlo estimated false selection rate  $\gamma$  for the FSR method with  $\gamma = 0.05$  for different sample sizes, ( $k_T = 21$  and  $k_P = 21$ )

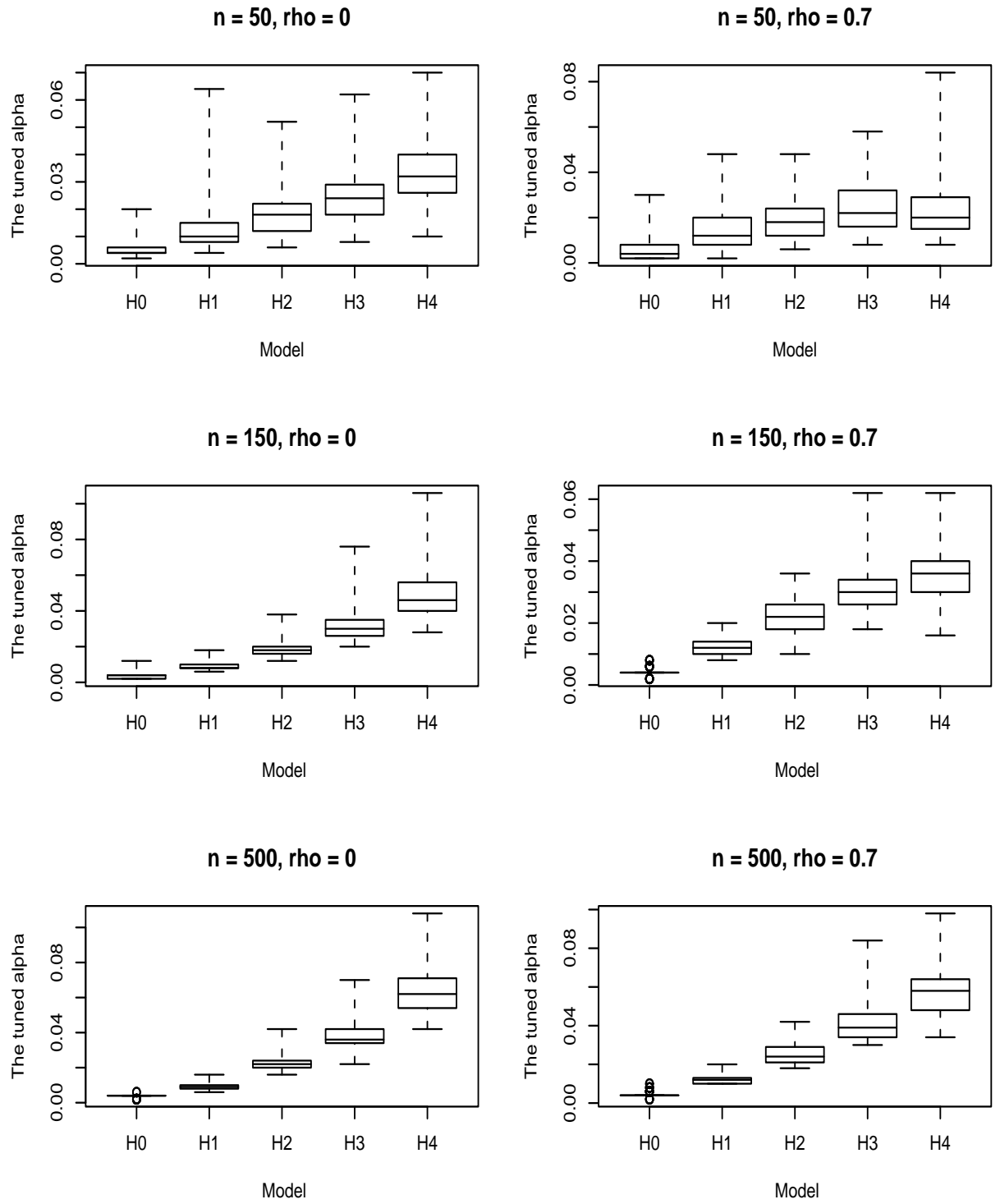


Figure 4.5: The boxplot of the  $\hat{\alpha}_0$  found by the FSR method with  $\gamma = 0.05$ , ( $k_T = 21$  and  $k_P = 21$ )



### 4.3 Comparison to Other Selection Criteria

To further study the performance of the FSR procedure, we compare it with Mallows' Minimum  $C_p$ , Breiman's little bootstrap, and the tuning selection procedure proposed by Luo et al. (2003). The best selection method, which selects the model having minimum model error from the forward selection candidate models based on "inside" knowledge of the true model error, is referred to as a "gold standard" for comparisons. It indicates the best that a forward selection procedure can do in prediction. For convenience, mnemonics for each of the selection procedures compared are defined as follows

Best : The best selection method. ("gold standard")

FSR : Our False Selection Rate procedure.

Tune : Luo's tuning procedure.

$C_p$  : Mallows' Minimum  $C_p$ .

LB : The little bootstrap.

Although  $C_p$  is based on all subsets searching, all other selection procedures are based on the forward selection algorithm. In LB, we let  $t = 0.6$  and  $B=40$  as suggested by Breiman (1992). In our FSR method, we let  $k_P = k_T$ ,  $B = 500$  and  $\gamma = 0.05$  as suggested in previous simulations.

We first concentrate on the results of sample size 50 and 150. The results (average model error and average model size) of the simulations are summarized in Tables A.1a - A.2b in Appendix A, where the average standard error of the entries over all five models for each method is given in the last column. The best way to understand the global behavior of the selection methods is to look at the graphs in Figure 4.6. The ratios of average ME of Best to average ME of FSR, Tune,  $C_p$  and LB are calculated and plotted versus model. Of course, the larger the ratio, the better is the selection in terms of prediction. The most striking result is that FSR performs substantially better than  $C_p$

and LB when the models are H0, H1, and H2. Also, for model H3, it turns out that FSR wins almost all the cases except for the case of  $n = 50$  and  $\rho = 0$ , in which it loses to  $C_p$  but still beats LB. For model H4, although FSR loses obviously to these two methods still at sample size 50 and  $\rho = 0$ , it gives fairly competitive results when  $\rho = 0.7$  or  $n = 150$ . Another striking result is that FSR is efficiently equivalent to Tune when the sample size is 150, but performs better than Tune at sample size 50 (i.e., nicely beats Tune when  $\rho = 0.7$ ). The above results are also reflected in the comparison of the average model size in Figure 4.7, in which the graphs are the average of model size estimates over the 100 replications plotted as a function of the H models. The solid lines are the average model sizes of Best. It is clear that  $C_p$  and LB always select dimensionalities larger than the optimal selection by Best, which indicates that they tend to overfit the model. The overfitting is considerable when the true model size is small, e.g., for model H0, H1 and H2, and this results in the model selected being penalized by larger average model error. This overfitting becomes less damaging if there are many weak variables, because it will then include some of the weak variables. For this reason,  $C_p$  and LB have good performance in model H4 that includes many weak variables. In contrast, FSR and Tune always select the most parsimonious models and hence are the best in reducing complexity. In fact, they select the optimal dimensionalities for small size models (e.g., for model H0, H1 and H2) as Best does, but underfit the model a bit for large size model (e.g., for model H3 and H4).

A similar pattern can be observed at sample size 500. The results of average model error and average model size are summarized in Tables A.3a and A.3b in Appendix A. The ratios of average ME of Best to average ME of FSR, Tune,  $C_p$  and LB are plotted in Figure 4.8. These results show that when the sample size becomes large all the methods improve and have reduced average model errors, so there is not much difference observed among them. But, we can still see that both FSR and Tune outperform  $C_p$  and LB. As

for the dimensionality selection, Figure 4.9 shows the graphs of the average model sizes, for which the solid lines still correspond to Best. It turns out that FSR is almost as good as Best in selecting the dimensionality, and that Tune also performs well but is a little more conservative than FSR for the H4 model. Furthermore, obvious overfitting is observed for  $C_p$  and LB whose curves are always above Best's curve.

In addition, to study the effect in the different methods when more unimportant variables are included, we ran the sample size 150 cases increasing the number of predictors in the data sets to 42 by adding the squares of the original 21 variables. We give Table A.4a and A.4b in Appendix A to record the average ME and average model size, respectively. Also, Figure 4.10 describes graphically the results of relative average ME with respect to Best, and Figure 4.11 shows the graphs of the average model size. The major result is that  $C_p$  and LB are adversely affected by the additional 21 predictors to a large extent, while FSR and Tune are less affected. Comparing Table A.2a to A.4a reveals that for  $C_p$  and LB the average model errors increase substantially when 21 additional predictors are included. This implies a sensitivity of the two methods to the number of unimportant variables in the given data. As for FSR and Tune, they exhibit relatively stable performance with only a little loss in predictive accuracy. Particularly, as predictors increases from 21 to 42, FSR is performing a little bit better than Tune. Because FSR targets the model false selection rate, it is reasonable to expect it to select smaller model when there are more unimportant variables, whereas other methods tend to choose larger models.

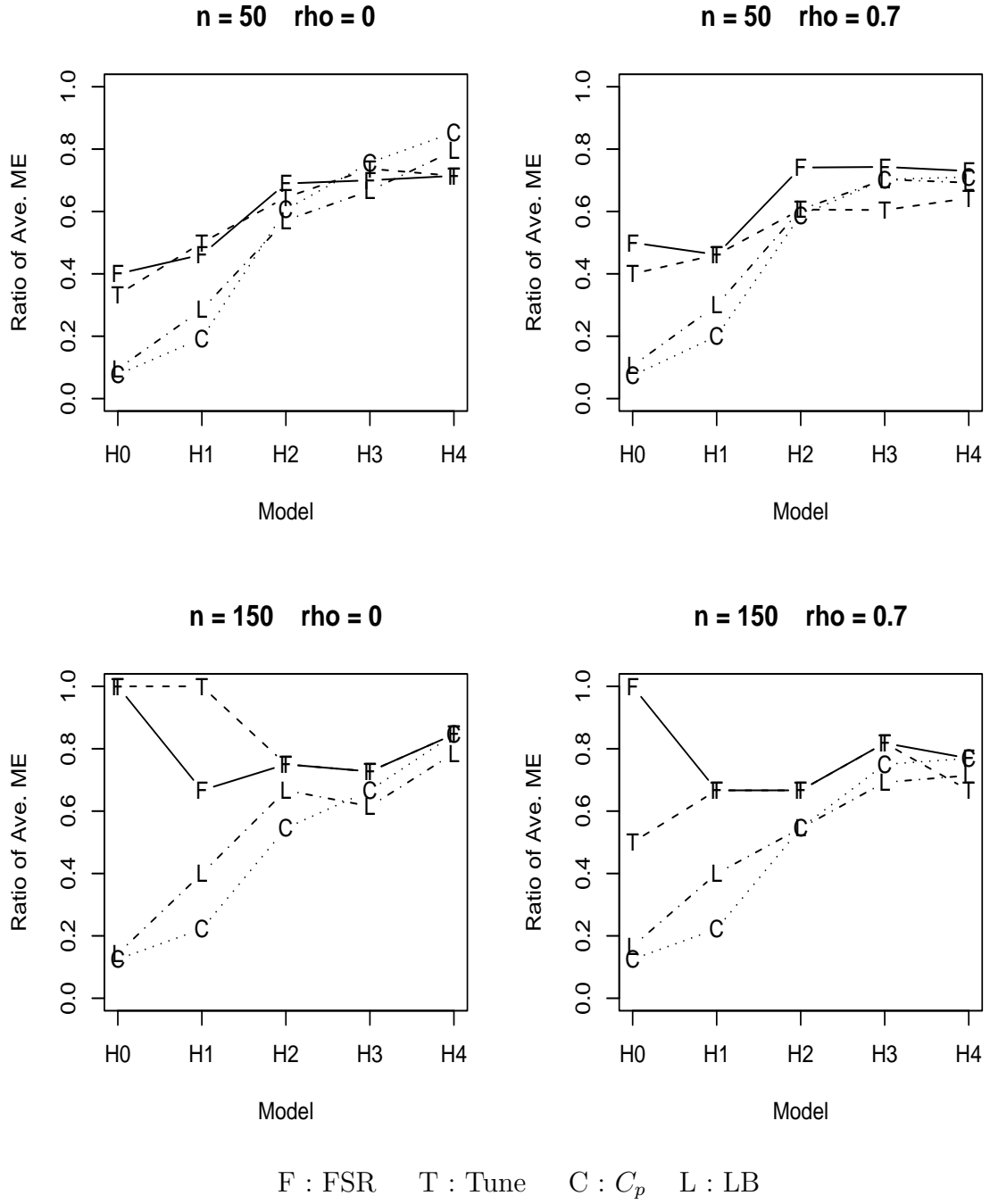


Figure 4.6: The ratios of average ME of Best to average ME of FSR, Tune,  $C_p$  and LB when  $k_T = 21$ ,  $n = 50$  and  $n = 150$

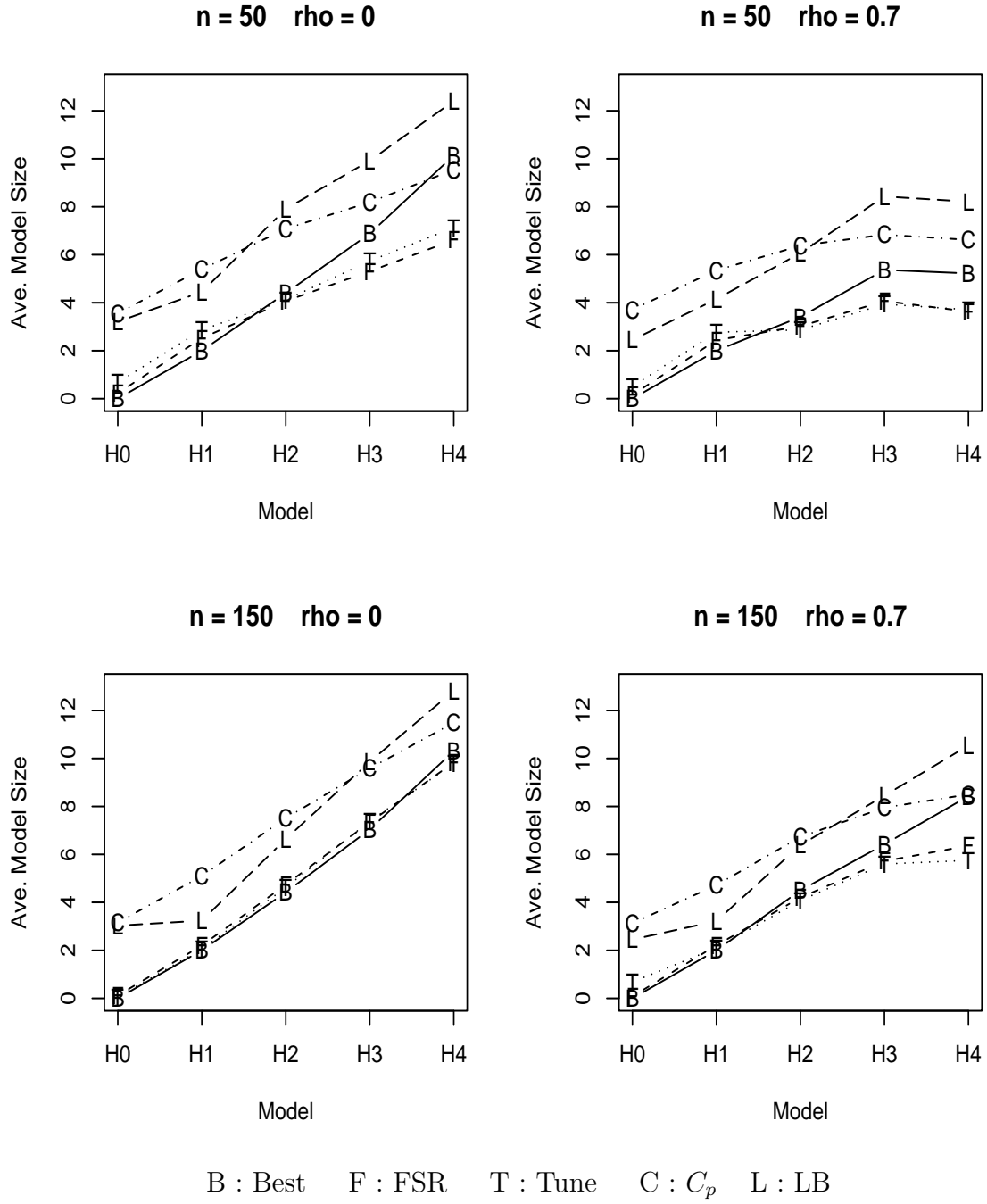


Figure 4.7: Average model size for Best, FSR, Tune,  $C_p$ , and LB when  $k_T = 21$ ,  $n = 50$  and  $n = 150$

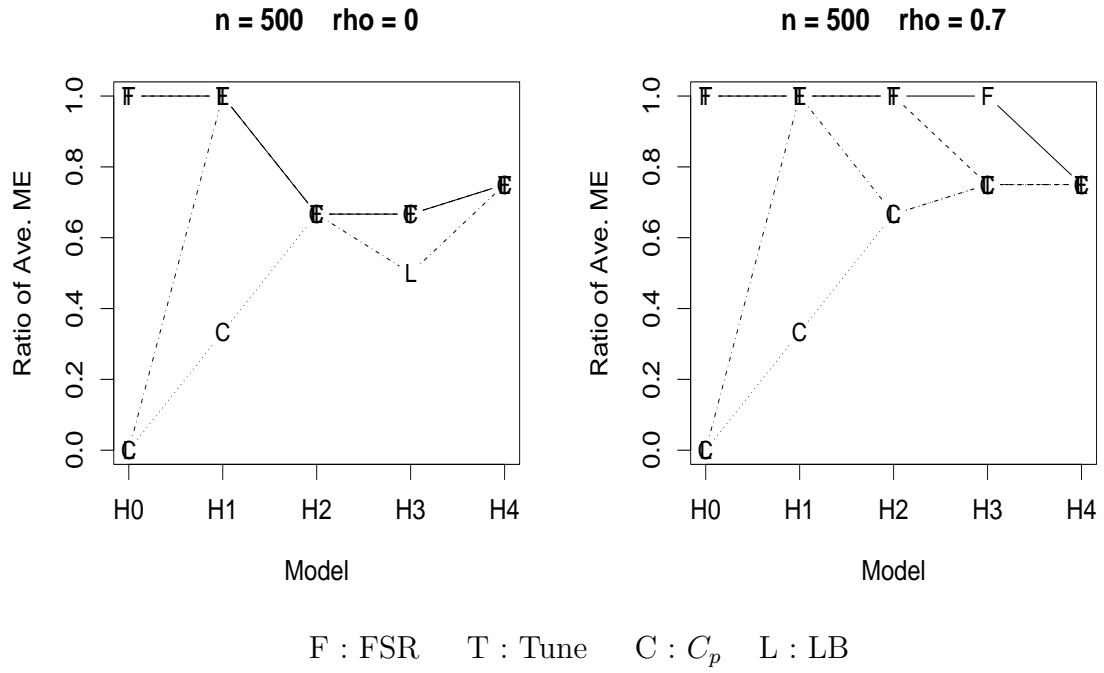


Figure 4.8: The ratios of average ME of Best to average ME of FSR, Tune,  $C_p$  and LB when  $k_T = 21$  and  $n = 500$

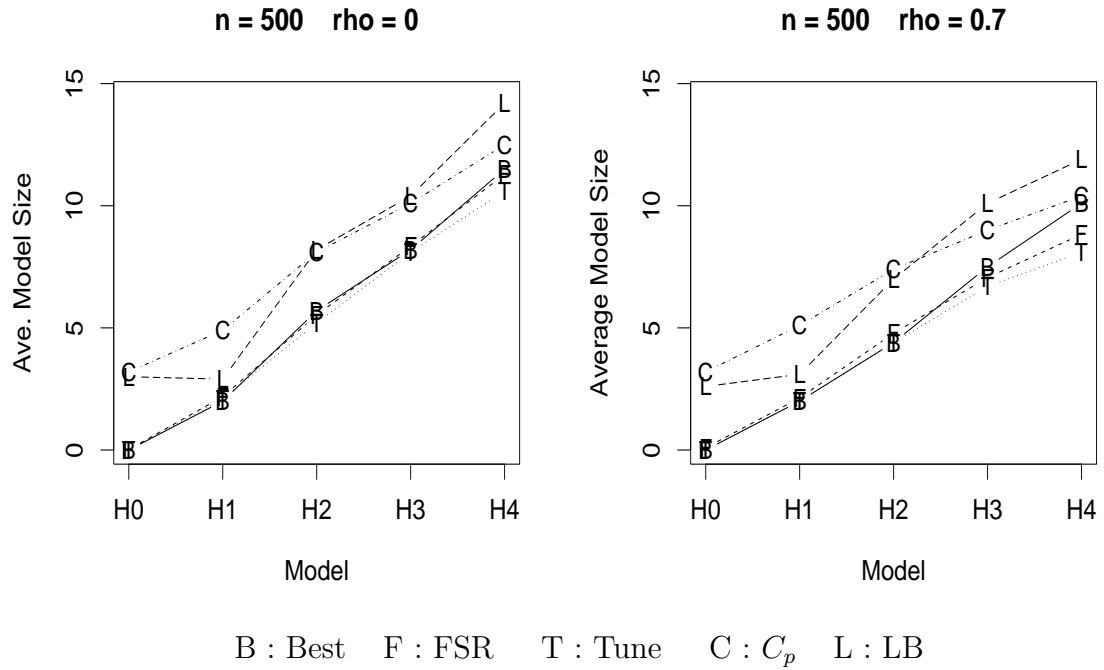


Figure 4.9: Average model size for Best, FSR, Tune,  $C_p$ , and LB when  $k_T = 21$  and  $n = 500$

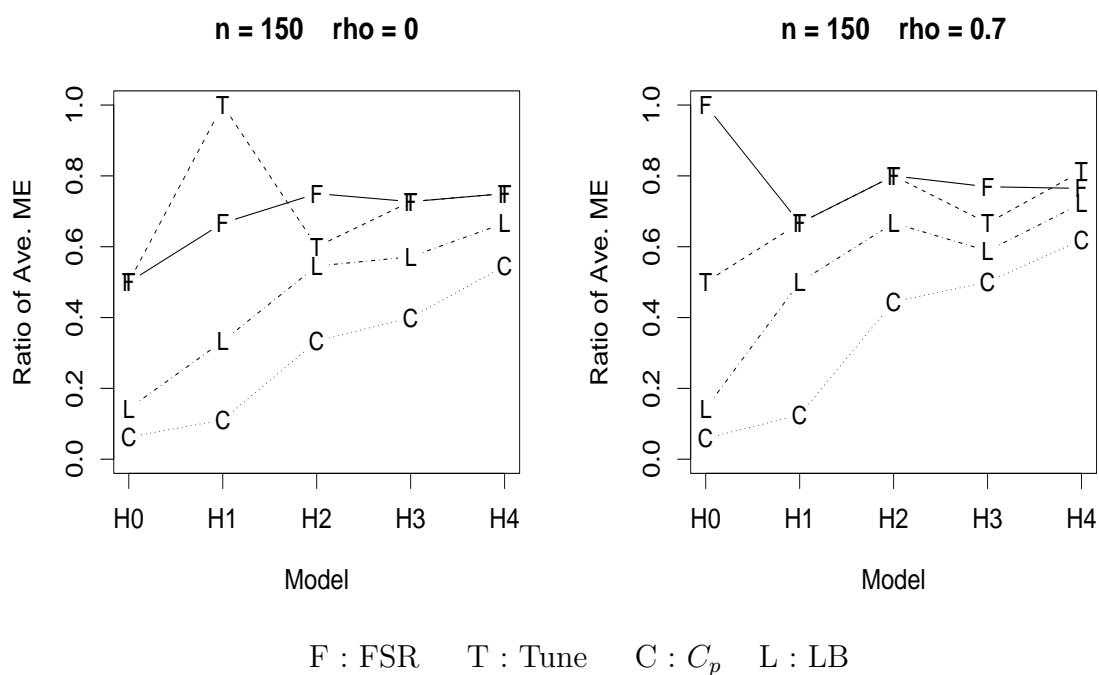


Figure 4.10: The ratios of average ME of Best to average ME of FSR, Tune,  $C_p$  and LB when  $k_T = 42$  and  $n = 150$

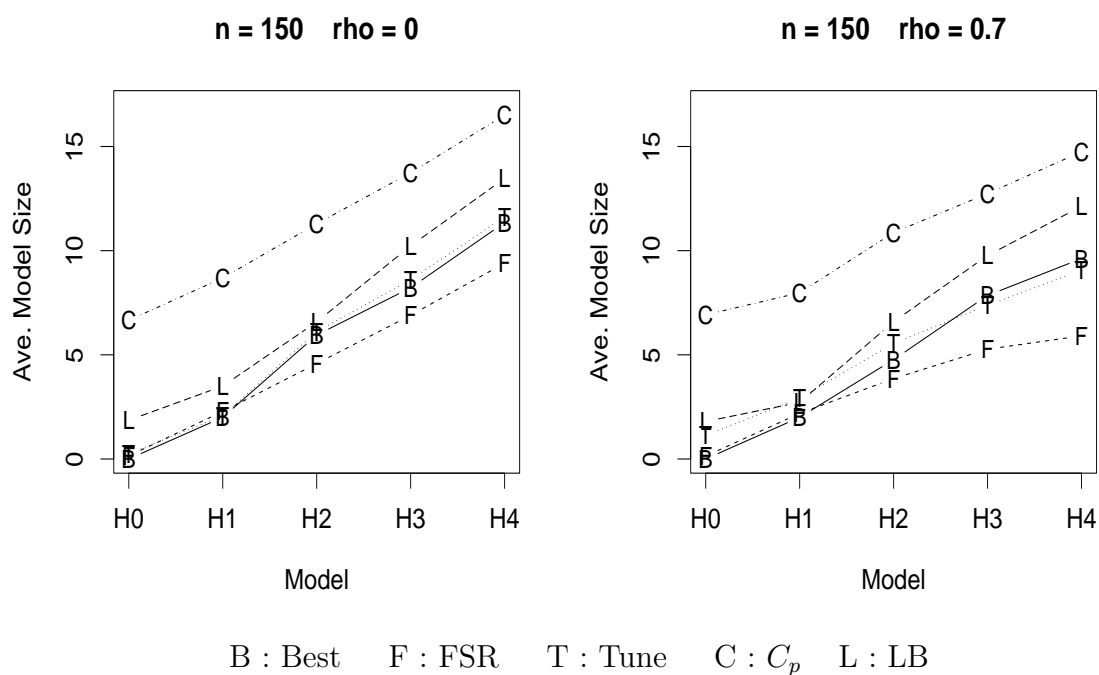


Figure 4.11: Average model size for Best, FSR, Tune,  $C_p$ , and LB when  $k_T = 42$  and  $n = 150$

## 4.4 Comparison to the Lasso

In recent years, the Lasso proposed by Tibshirani (1996) has received increasing interest. The underlying idea is to minimize the residual sum of squares under a constraint on the sum of the absolute value of the coefficients to be less than a constant. The Lasso exhibits the stability of ridge regression because of its shrinkage of the coefficients. On the other hand, it makes some of the coefficients exactly 0, and hence results in subset selection and produces interpretable models. In this section, we intend to compare our FSR method with the Lasso. We choose data sets having sample size 150 and 21 predictors. In addition to the normal assumption of the response variable (i.e. the random error  $\epsilon_i$  are assumed to be independently distributed as standard normal distribution), we consider a skewed distribution. In particular, we also assume that the error  $\epsilon_i$  are independently taken from the exponential distribution  $\exp(1)-1$ , for which we still have mean  $E(\epsilon_i) = 0$  and variance  $\text{Var}(\epsilon_i) = 1$ .

The average model errors and average model sizes are plotted in Figure 4.12 and Figure 4.13, respectively. Also, the Monte Carlo estimated false selection rates are plotted in Figure 4.14, in which the solid line corresponds to the value 0.05. In these figures, the top two pictures are in the cases when the random error  $\epsilon_i$  are taken from standard normal distribution  $N(0, 1)$ , and the bottom two show the results when  $\epsilon_i$  are from the skewed exponential distribution  $\exp(1)-1$ . The detailed results are summarized in Table A.5a – A.6c in Appendix A. The results indicate that, when the  $X$ -variables are independent (i.e.,  $\rho = 0$ ), the FSR procedure exhibits a significant advantage over the Lasso in terms of the prediction accuracy. When the  $X$ -variables are correlated, the FSR method does not do as well as the Lasso for model H3 and H4, but it still performs better on H1 model. As for average model size, Figure 4.13 shows that the Lasso always choose more variables than the FSR method. In fact, the Lasso tends to select so many unimportant



variables into the model such that the model false selection rates are extremely high, whereas the FSR method keep the false selection rate around its target level 0.05.

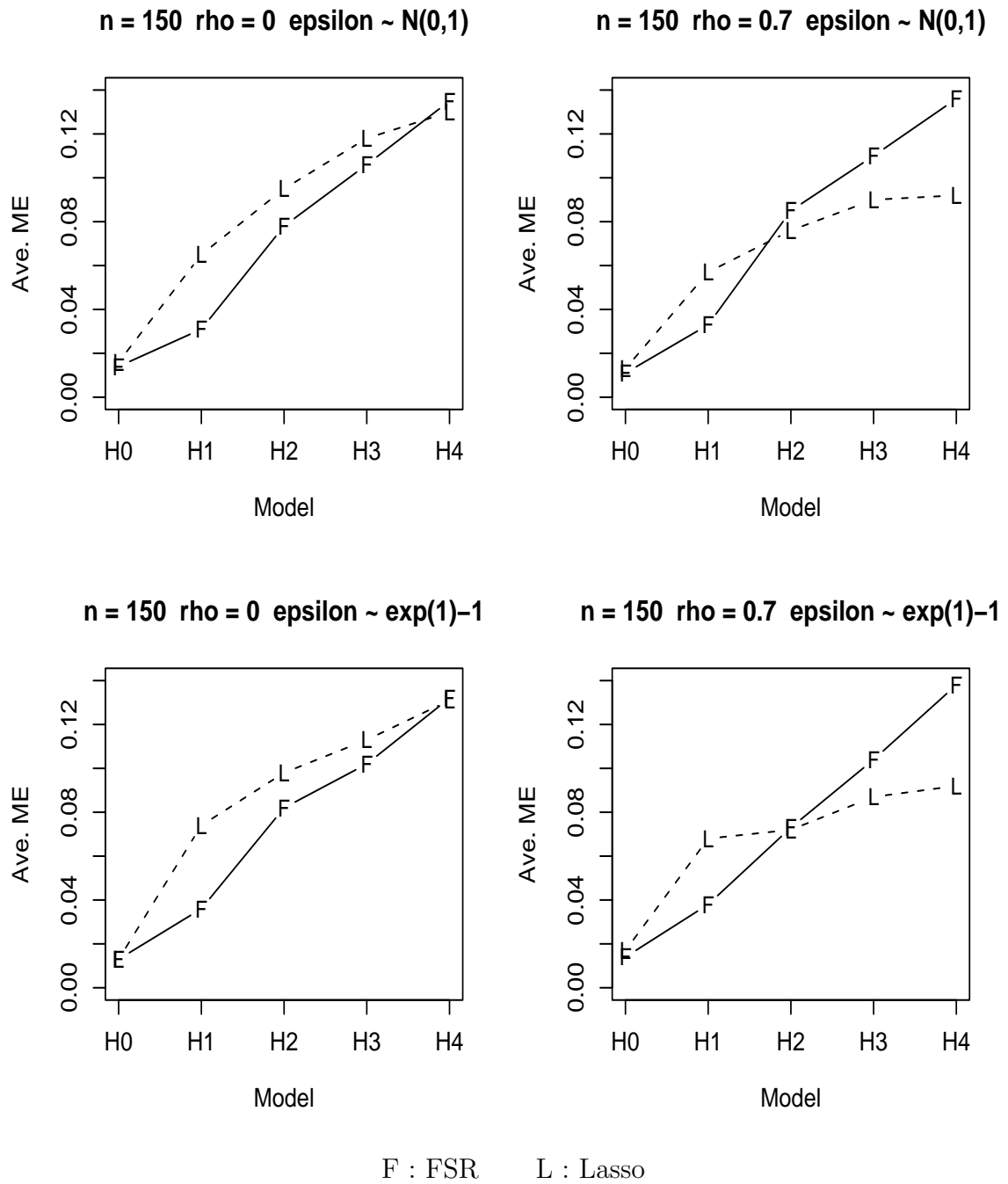


Figure 4.12: Average model error for the FSR method and the Lasso

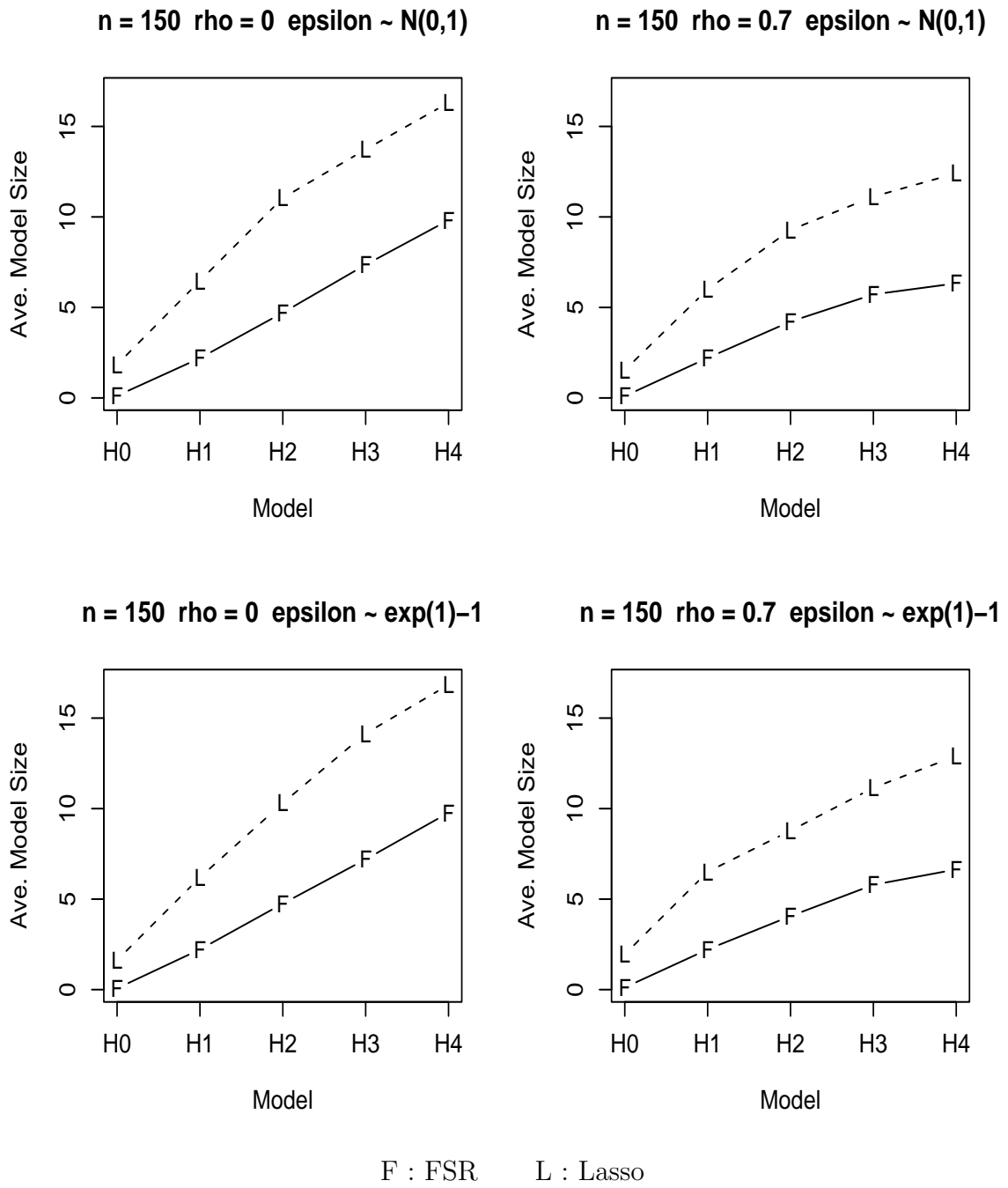


Figure 4.13: Average model size for the FSR method and the Lasso

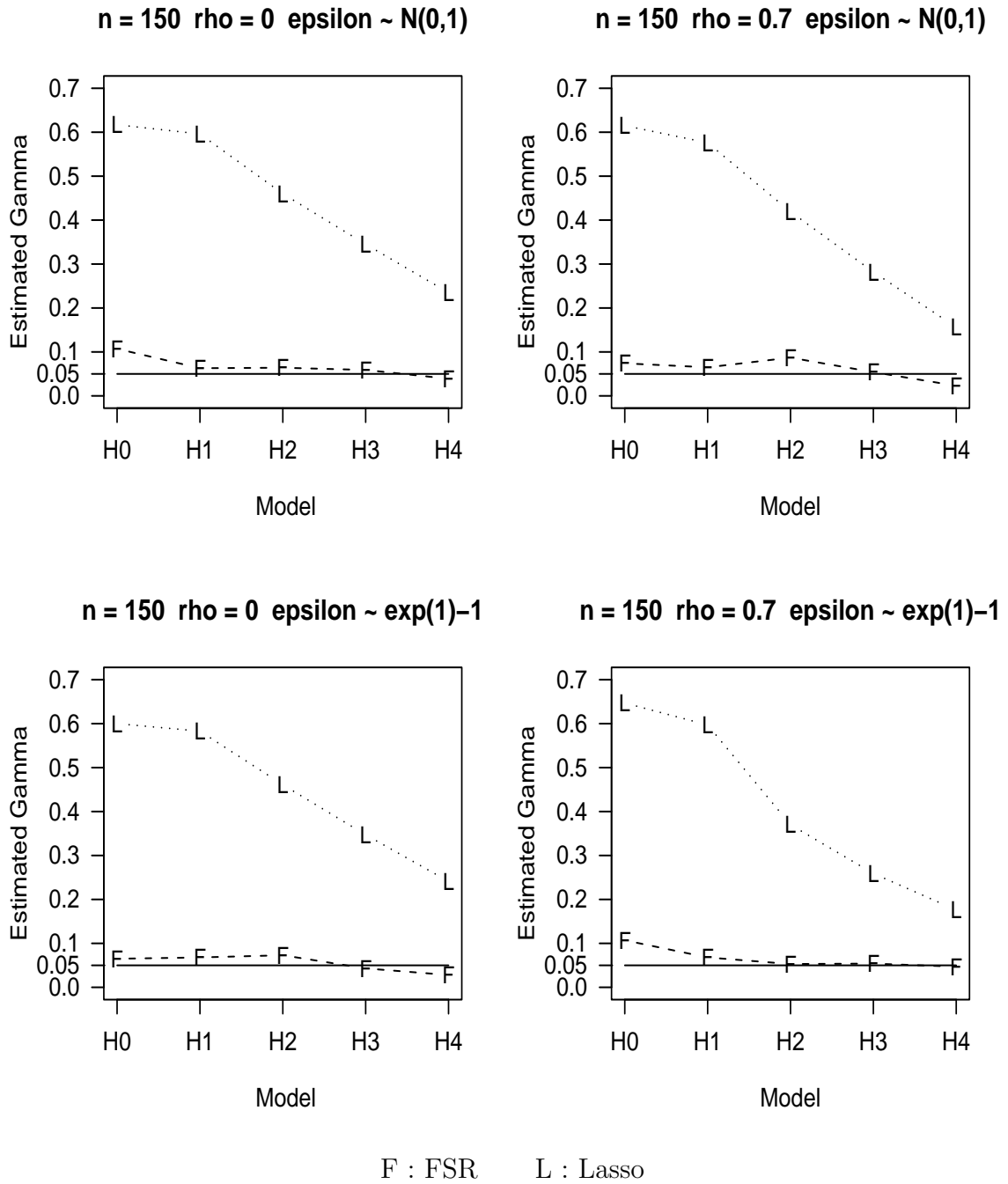


Figure 4.14: The Monte Carlo estimated false selection rate  $\gamma$  for the FSR method and the Lasso

## Chapter 5

# Application to Logistic Regression Models

There exist many methods to select variables for linear regression models, but not as much attention has been paid to logistic regression models, which are widely used in many fields, such as biomedical research, business and finance, and ecology. Recall that the FSR procedure is not specific to linear regression models, but has the advantage of being applicable to a broader class of regression models. In this chapter, we focus on its application in logistic regression models. The logistic regression models have binary response variable  $Y$ , and  $\Pr(Y = 1|X) = \pi(X)$  where

$$\ln \left[ \frac{\pi(X)}{1 - \pi(X)} \right] = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_{k_T}.$$

The conditional distribution of the binary response variable follows a binomial distribution with probability given by the conditional mean,  $\pi(X)$ . The subset searching methods, like forward selection and backward elimination, can easily accommodate logistic regression models, and a model is selected by specifying a significance level. Other selection criteria besides significance tests are information-based, among which AIC and BIC are the best known.

## 5.1 Forward Selection In Logistic Regression

In linear regression, an  $F$ -test is used for testing the “importance” of a variable entering the model at each step of forward selection. In logistic regression, the most convenient test to use is the score test because the maximum likelihood estimates for potential variables are not required, and forward selection includes the variable with the smallest  $P$ -value.

Suppose that intercept and  $t - 1$  explanatory variables have been entered into the model and that  $X_t$  is another explanatory variable to be included. Let the regression parameter vector of the  $t + 1$  variables (including the intercept) be denoted by  $\theta = (\beta_0, \beta_1, \dots, \beta_t)^T$ , where  $\beta_0$  is related to the intercept. Let  $L(\theta)$  be the value of the log-likelihood for the model containing the  $t + 1$  variables. Let  $\mathbf{U}(\theta)$  be the vector of first partial derivatives of the log-likelihood  $L(\theta)$  with respect to the parameter vector  $\theta$ , i.e.,  $\mathbf{U}(\theta) = \partial L(\theta) / \partial \theta$ , and let  $\mathbf{H}(\theta)$  be the matrix of second partial derivatives of  $L(\theta)$ , i.e.,  $\mathbf{H}(\theta) = \partial^2 L(\theta) / \partial \theta^2$ . That is,  $\mathbf{U}(\theta)$  is the gradient vector, and  $\mathbf{H}(\theta)$  is the Hessian matrix. Let  $\mathbf{I}(\theta)$  be either  $-\mathbf{H}(\theta)$  or its expected value. We consider the null hypothesis  $H_0 : \beta_t = 0$ . The score test statistic for testing  $H_0$  is defined by

$$G = \mathbf{U}^T(\hat{\theta}_0) \mathbf{I}(\hat{\theta}_0) \mathbf{U}(\hat{\theta}_0),$$

where  $\hat{\theta}_0$  is the maximum likelihood estimate of  $\theta$  under the null hypothesis  $H_0$ . Under  $H_0$ , the  $G$  has an asymptotic Chi-squared distribution with one degree of freedom, say  $\chi^2(1)$ . The test  $P$ -value is therefore

$$P\text{-value} = \int_G^\infty f(t) dt,$$

where  $f(t)$  is the Chi-squared density with one degree of freedom, i.e.,

$$f(t) = \frac{t^{(1/2)-1} e^{-t/2}}{\Gamma(1/2) 2^{1/2}}.$$

It is obvious that for the score test only the maximum likelihood estimates for the current variables in the model is computed, and we do not have to compute the maximum likelihood estimates for all the variables not in the model. Thus, the score tests are preferred in forward selection for logistic regression. In SAS, PROC LOGISTIC uses the score tests to select new variable in forward selection.

## 5.2 Simulation Study

To study the behavior of the FSR procedure in logistic regression, we carried out a Monte Carlo simulation and compare it with AIC and BIC used with forward selection. We use SAS to do the simulation.

### 5.2.1 Simulation Design

For convenience, we use the same design matrix  $\mathbf{X}$ , including  $\rho = 0$  and  $\rho = 0.7$ , with sample size  $n = 150$  and  $n = 500$  as found in Chapter 4. The binary response variable,  $y$ , is generated at two stages as follows.

1. We generate the variable  $y_i^*$  by

$$y_i^* = \mathbf{x}_i \boldsymbol{\beta} + \epsilon_i \quad i = 1, \dots, n,$$

where  $\mathbf{x}_i$  is the  $i$ th row of the design matrix  $\mathbf{X}$ ,  $\boldsymbol{\beta}$  is the  $21 \times 1$  regression coefficients vector, and  $\epsilon_i$  are independently drawn from a logistic distribution with mean zero and variance 1, i.e.,

$$\epsilon_i = \frac{\sqrt{3}}{\pi} \ln \left( \frac{U}{1 - U} \right),$$

where  $U$  has a uniform distribution  $(0, 1)$ . The coefficients  $\boldsymbol{\beta}$  are generated in the same way as in Section 4.1 and lead to four H models (i.e., H1, H2, H3, and H4).

We force the model theoretical  $R^2$  defined in (4.3) to be equal to 0.35 and 0.75 in order to study the performance under different latent variable scales. We also include the null model H0 for which  $\beta = 0$ .

2. The binary response variable  $y_i$  are then generated by

$$y_i = \begin{cases} 1, & y_i^* > \beta_0; \\ 0, & y_i^* < \beta_0, \end{cases}$$

where the value of  $\beta_0$  is determined by constraining the unconditional probability that  $Y_i$  is 1 to be 0.1 and 0.5, i.e.,

$$Pr(Y_i = 1) = P, \quad P = 0.1, 0.5.$$

Details of the method used to determine  $\beta$  are given in Appendix B.

As before, for each model, we carry out 100 replications.

To evaluate the model selection, we propose two criteria, Standardized Model Error (SME), and Logit Error (LE). Standardized model error is defined as

$$SME = \frac{1}{n} \sum_{i=1}^n \frac{(\pi_i - \hat{\pi}_i)^2}{\pi_i(1 - \pi_i)},$$

where  $\pi_i = Pr(Y_i = 1|\mathbf{x}_i)$ , and  $\hat{\pi}_i$  is its estimate based on the selected model. The logit error is defined by

$$LE = \frac{1}{n} \sum_{i=1}^n \left[ \ln \left( \frac{\hat{\pi}_i}{1 - \hat{\pi}_i} \right) - \ln \left( \frac{\pi_i}{1 - \pi_i} \right) \right]^2,$$

and thus reflects prediction accuracy in the linear part of the model. Small values of SME and LE are desirable. In addition to SME and LE, average dimension selected, and false selection rate are used for comparison.



### 5.2.2 Simulation Results

We first concentrate on the results when  $R^2 = 0.75$ . The average SME and LE are plotted in Figure 5.1 and 5.2. The clearest result is that the FSR method is generally superior to AIC. AIC produces generally larger average LE across the models, and has larger average SME for models H0, H1, and H2. Only for model H4 is AIC a little better than the FSR method in terms of average SME. As for comparison between the FSR method and BIC, in terms of average LE, the FSR method outperforms BIC in the cases  $n = 150$  and  $P = 0.1$ , and is essentially equivalent to BIC in the other cases. For average SME, the results are competitive between FSR method and BIC. We now turn our attention to the cases where  $R^2 = 0.35$ . Figure 5.4 and 5.5 show the plots of the average SME and LE. A similar pattern is observed in this situation. The detailed results of average SME and LE and average model size are listed in Tables A.7a - A.14c in Appendix A. It is seen that both the FSR method and BIC tend to select parsimonious models, while AIC tends to overfit the models.

As for comparison of the false selection rates estimated by Monte Carlo, we give Tables A.15a - A.15h in Appendix A to summarize the results. Figure 5.3 and 5.6 exhibit the plots. It is obvious that the FSR method is best in that it retains the false selection rate close to 0.05 across the models. AIC has fairly high false selection rate, which indicates its high probability to select too many noise variables. In the cases  $n = 500$ , BIC performs as well as FSR method except for model H0, but becomes inferior with larger false selection rate for the cases where  $n = 150$ .

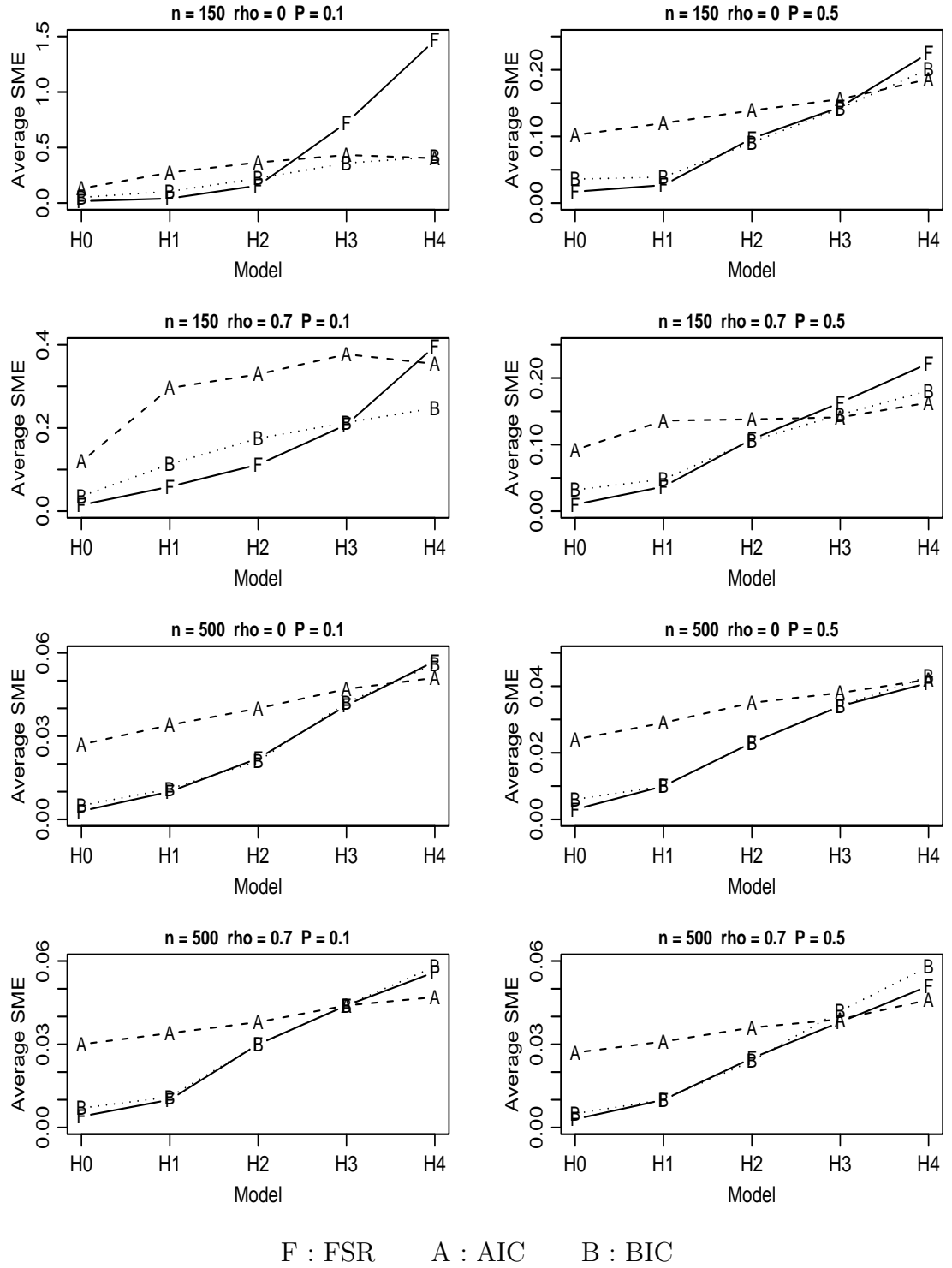


Figure 5.1: Average SME for the FSR method, AIC, and BIC for logistic regression when  $R^2 = 0.75$

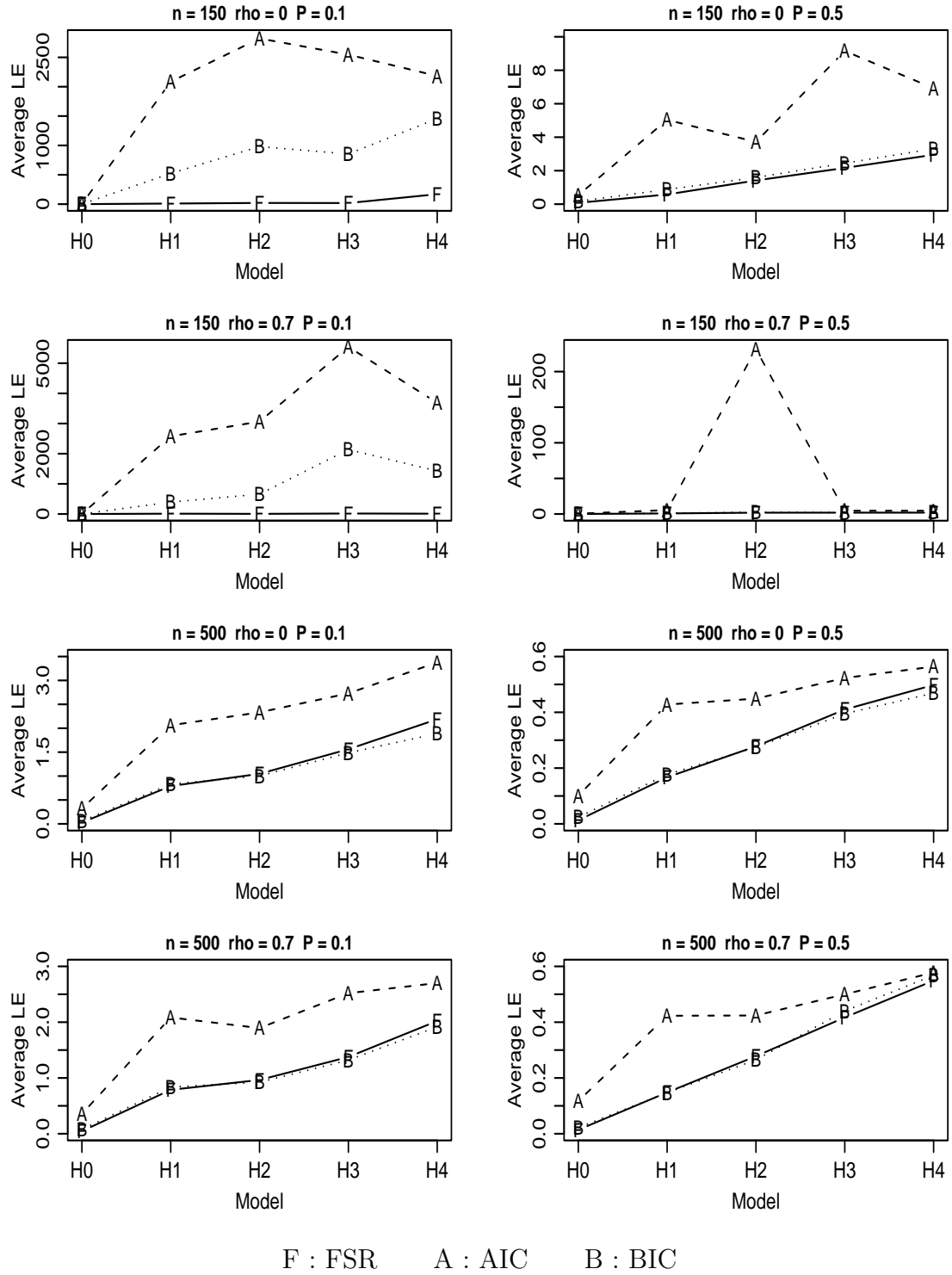


Figure 5.2: Average LE for the FSR method, AIC, and BIC for logistic regression when  $R^2 = 0.75$

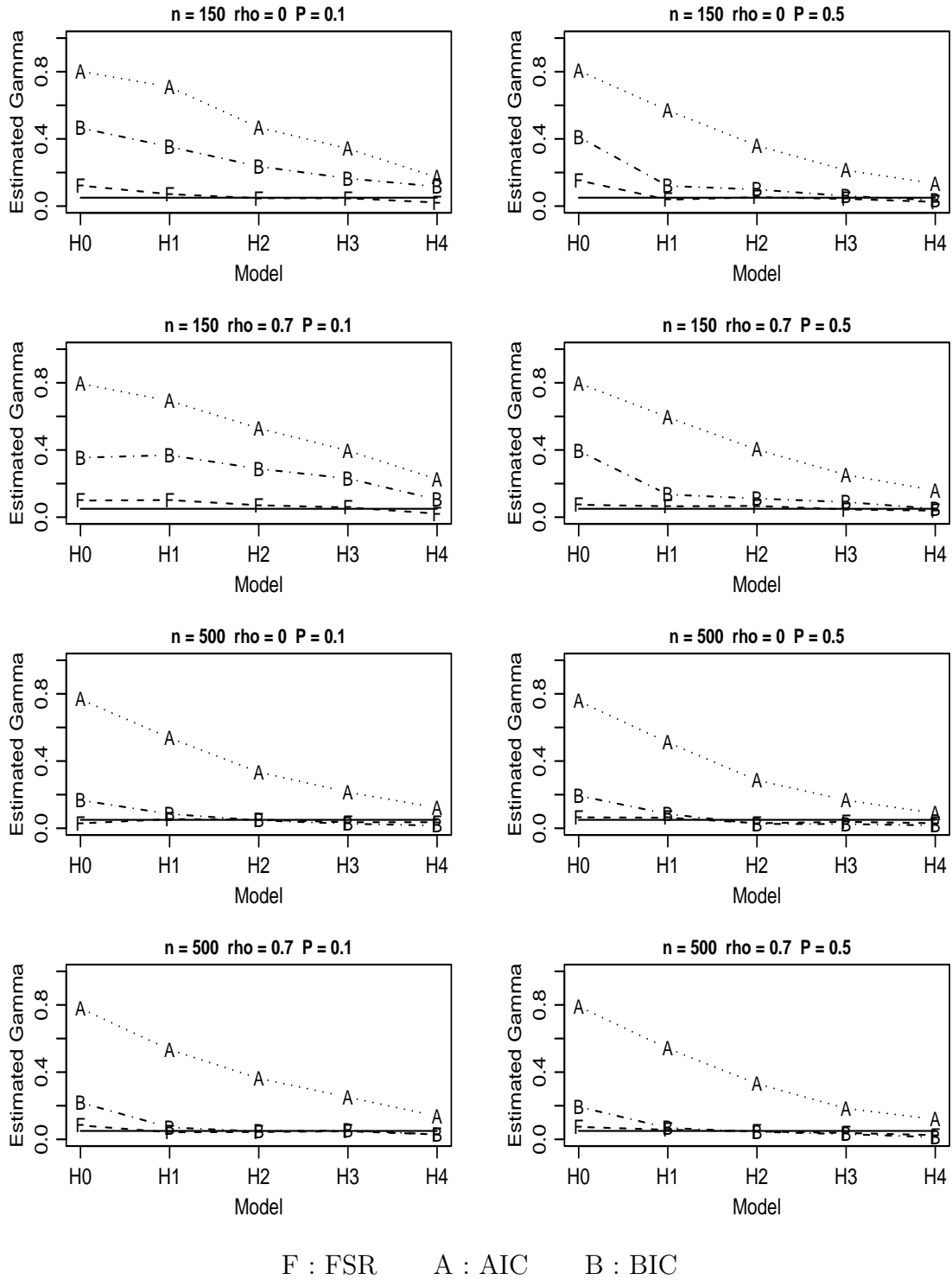


Figure 5.3: The Monte Carlo estimated false selection rate  $\gamma$  for the FSR method, AIC, and BIC for logistic regression when  $R^2 = 0.75$ , and the black solid line corresponds to value 0.05

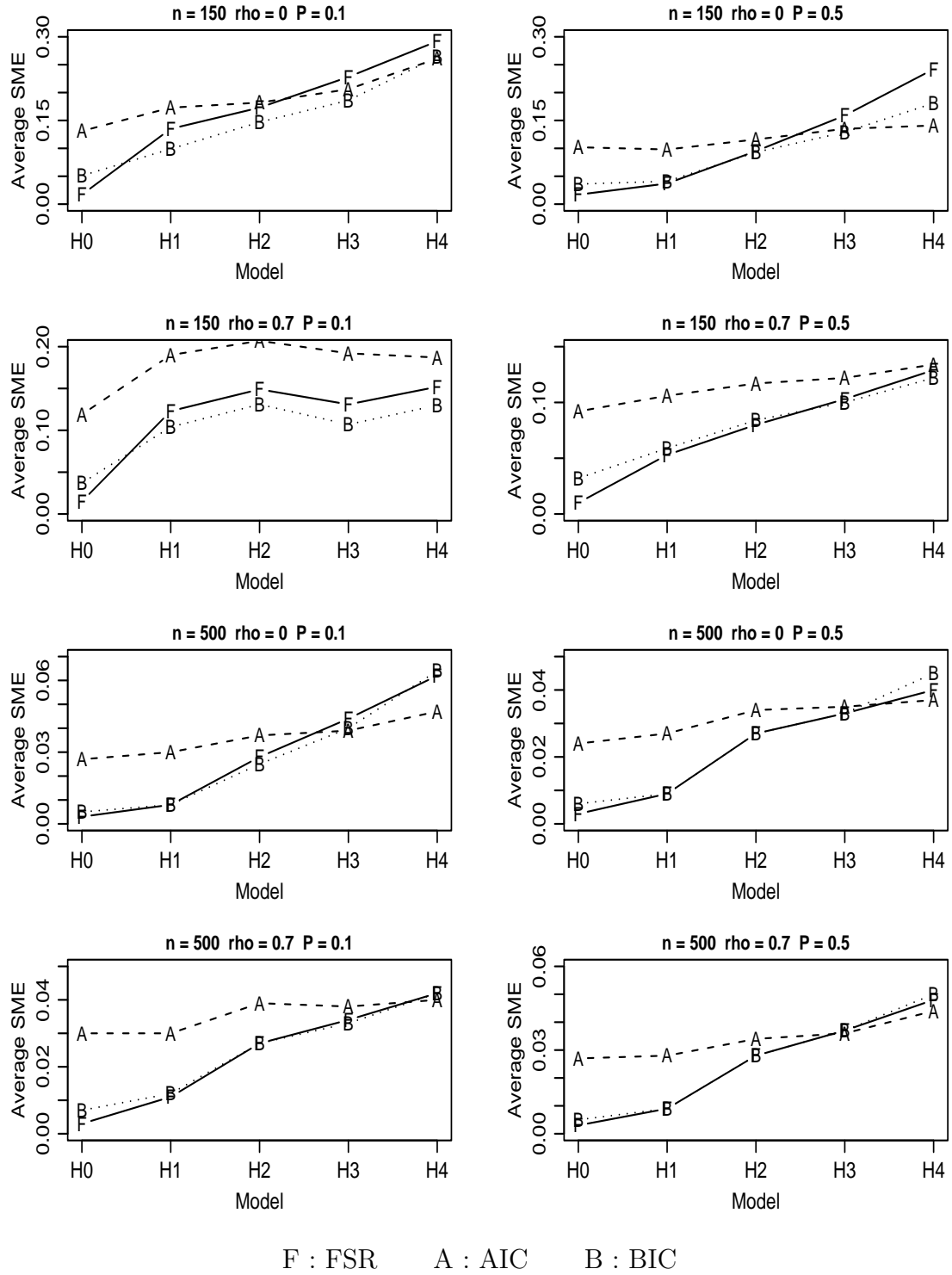


Figure 5.4: Average SME for the FSR method, AIC, and BIC for logistic regression when  $R^2 = 0.35$

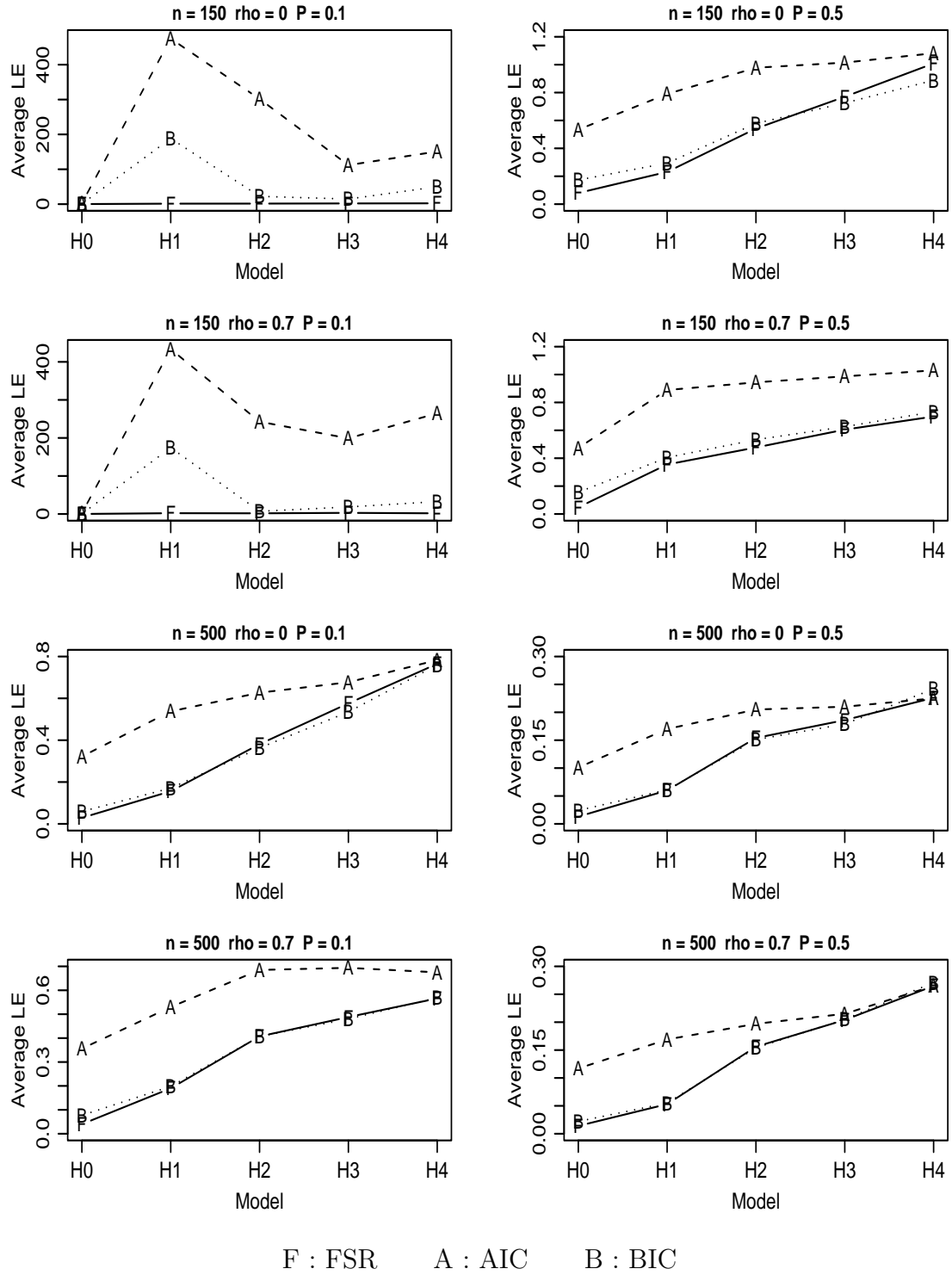


Figure 5.5: Average LE for the FSR method, AIC, and BIC for logistic regression when  $R^2 = 0.35$

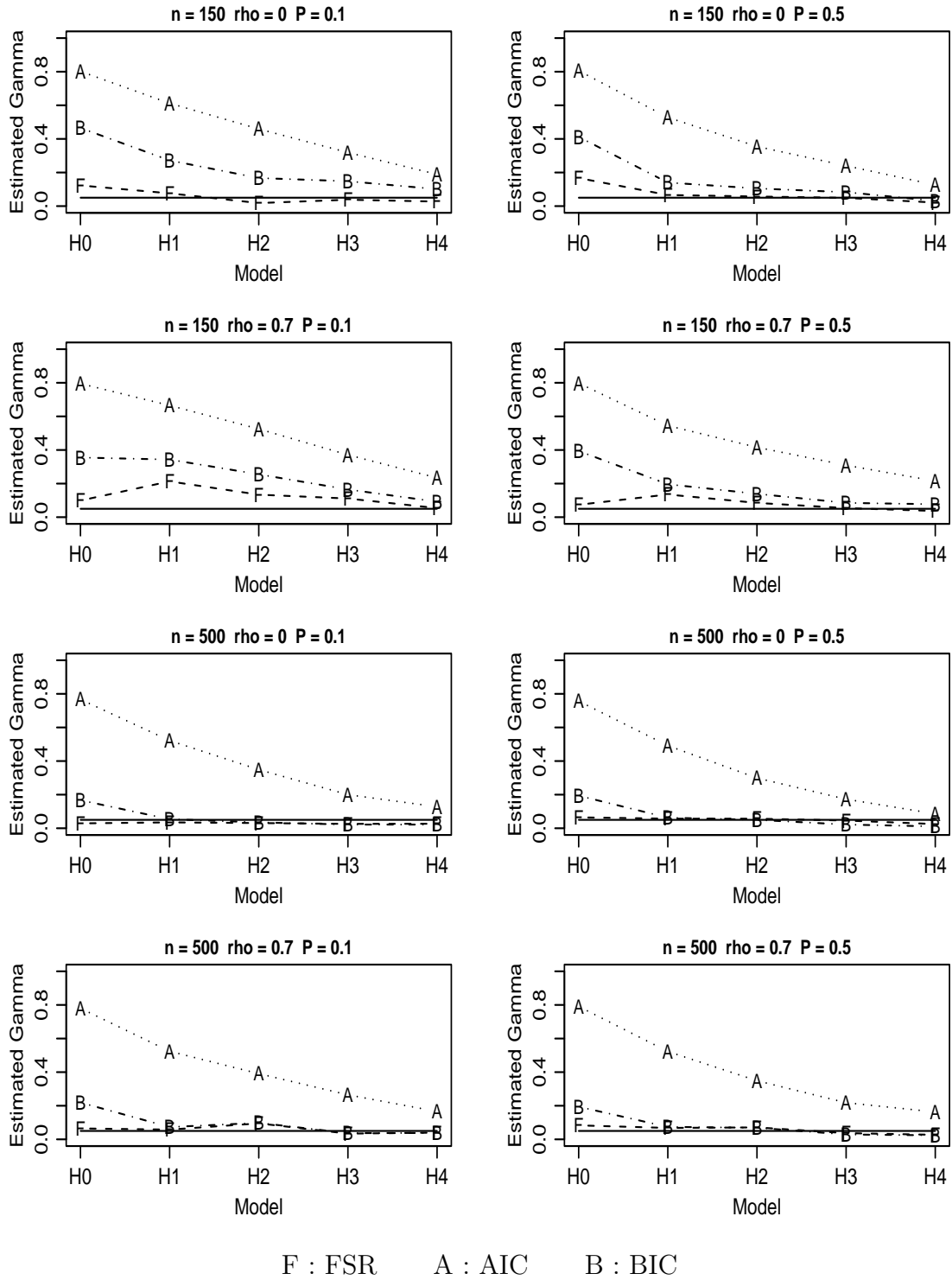


Figure 5.6: The Monte Carlo estimated false selection rate  $\gamma$  for the FSR method, AIC, and BIC for logistic regression when  $R^2 = 0.35$ , and the black solid line corresponds to value 0.05

# Chapter 6

## Applications

Four real data sets are used for the purpose of illustrating the FSR procedure: wing shape, diabetes, dog surgery TPLO, and drug abuse data. Linear regression models are estimated for the first two data sets, and logistic regression models are used for the last two. In the FSR method, we choose the target false selection rate  $\gamma = 0.05$ , generate the same number of pseudo-variables as found in the set of original variables, and repeat generation of pseudo-variables 500 times as we did in the simulation studies. The pseudo-variables are generated using Method 2a (residuals after permutation) discussed in Section 3.3. PROC REG and PROC LOGISTIC in SAS are used to carry out forward selection.

### 6.1 Wing Shape Data

The data come from a study analyzing polygenes affecting wing shape on chromosome 2 in *Drosophila melanogaster*. Details of the study are described in Weber et al. (2001). This data set contains genotypes (1 or 0 for two homozygotes) of 48 markers and the phenotype of one trait (wing size index) for 701 recombinant inbred lines. The markers are listed in Table 6.1, where cM gives the linkage map of markers based on the standard *Drosophila*



map. An interesting problem is to identify markers that are linked to quantitative trait loci (QTL).

Table 6.1: Description of markers for the wing shape data

ID	Marker	cM	ID	Marker	cM	ID	Marker	cM
1	22A2	2.2	17	37B8	53.7	33	46F1	60.0
2	22B2	2.4	18	37D1	53.9	34	47B4	60.3
3	23B1	6.6	19	38B1	54.4	35	47D1	60.9
4	23C3	7.5	20	38D1	54.6	36	48D1	62.5
5	24C1	10.3	21	38E1	54.7	37	49E4	66.7
6	24D4	11.2	22	39E1	54.9	38	50D1	70.8
7	24E5	12.0	23	41D1	55.5	39	51D6	73.5
8	25A1	13.0	24	42B1	55.6	40	52A1	74.6
9	25A3	13.5	25	42B2	55.7	41	55C4	83.8
10	27B1	19.8	26	42C1	55.8	42	56B7	87.5
11	29F1	30.4	27	42C3	55.9	43	56F3	90.0
12	32C1	42.9	28	43A1	56.4	44	57E6	98.1
13	33A1	44.3	29	43B3	56.5	45	58D1	99.2
14	34A5	46.2	30	44F1	58.3	46	59A1	101.7
15	34B2	46.5	31	46C1	59.7	47	59B8	102.3
16	35B6	50.4	32	46C6	59.8	48	60F1	108.1

A simple linear model with multiple regression on markers is applied, i.e.,

$$Y_i = \beta_0 + \sum_j X_{ij}\beta_j + \epsilon_i, \quad i = 1, \dots, 701, \quad j = 1, \dots, 48,$$

where  $i$  indexes recombinant inbred lines,  $X_{ij}$  is the genotype values of the  $j$ th markers with  $j$  indexing markers,  $Y_i$  is the trait value,  $\epsilon_i$  is the residual, and  $\beta_0$  and  $\beta_j$  are the regression parameters. The FSR method chooses  $\hat{\alpha}_0 = 0.11$ . Forward selection produces the following sequence,

### Summary of Forward Selection

Step	Variable Entered	Model R-Square	Fvalue	Pr > F
1	34A5	0.5148	741.53	<.0001
2	23C3	0.6431	251.06	<.0001
3	56B7	0.8141	640.96	<.0001
4	42C3	0.8678	283.05	<.0001
5	59A1	0.8954	183.15	<.0001
6	22A2	0.9119	130.13	<.0001
7	27B1	0.9226	95.33	<.0001
8	50D1	0.9273	44.79	<.0001
9	24E5	0.9310	37.18	<.0001
10	60F1	0.9326	16.64	<.0001
11	46C1	0.9337	11.34	0.0008
12	37B8	0.9345	8.03	0.0047
13	29F1	0.9353	8.65	0.0034
14	58D1	0.9359	7.08	0.0080
15	33A1	0.9363	3.66	0.0561
16	25A3	0.9366	3.32	0.0690
17	52A1	0.9369	3.10	0.0790
-----				
18	57E6	0.9371	2.31	0.1294
19	22B2	0.9373	2.01	0.1566
...				

The last column shows the  $P$ -values for the variables entering into the model. In step 18, the  $P$ -value for entering is 0.1294, which is larger than 0.11, whereas the  $P$ -value in step 17 is 0.0790, which is less than 0.11. Therefore, forward selection stops in step 17 and the final model includes the variables entering in the first 17 steps. In summary, the following 17 markers are identified to be statistically significant:

22A2 23C3 24E5 25A3 27B1 29F1 33A1 34A5 37B8  
 42C3 46C1 50D1 52A1 56B7 58D1 59A1 60F1

These markers explain 93.4% of the total phenotypic variance.

## 6.2 Diabetes Data

This data set is used as an example in the paper by Efron et al. (2004) to illustrate their new model selection method, LARS. Ten baseline variables were obtained for each of  $n = 442$  diabetes patients. The response of interest is a quantitative measure of disease progression one year after baseline. The description of the variables is listed in Table 6.2. A statistical model is desired to produce accurate baseline predictions of response for future patients. The important baseline factors in disease progression suggested by the form of the model are of interest.

Table 6.2: Description of variables in the diabetes data

Variable	Name	Description
1	AGE	Patients' age
2	SEX	Patients' sex
3	BMI	Body mass index
4	BP	Average blood pressure
5	S1	Blood serum measurements 1
6	S2	Blood serum measurements 2
7	S3	Blood serum measurements 3
8	S4	Blood serum measurements 4
9	S5	Blood serum measurements 5
10	S6	Blood serum measurements 6
11	DISEASE	Disease progression one year after baseline

The data are first standardized as done by Efron et al. (2004) such that the baseline variables have mean 0 and unit length, and that the response has mean 0. Then, a linear regression model with multiple regression on the baseline variables is applied. With only 10 main effects considered, forward selection produce the sequence

Summary of Forward Selection

Step	Variable Entered	Model R-Square	Fvalue	Pr > F
1	BMI	0.3439	230.65	<.0001
2	S5	0.4595	93.86	<.0001
3	BP	0.4801	17.35	<.0001
4	S1	0.4920	10.27	0.0015
5	SEX	0.4999	6.84	0.0092
6	S2	0.5149	13.47	0.0003
-----				
7	S4	0.5163	1.26	0.2619
8	S6	0.5175	1.06	0.3040
9	S3	0.5177	0.22	0.6386
10	AGE	0.5177	0.03	0.8670
...				

The FSR method produces  $\hat{\alpha} = 0.112$ , and hence forward selection stops in step 6. The final model therefore includes 6 covariates. We then consider the quadratic model with 10 main effects, 45 two-way interactions, and 9 quadratic terms (each baseline variable except the dichotomous variable SEX). We do not enforce the hierarchy principle, and thus the interaction effect can be entered in the model before both main effects. Forward selection produces the sequence

#### Summary of Forward Selection

Step	Variable Entered	Model R-Square	Fvalue	Pr > F
1	BMI	0.3439	230.65	<.0001
2	S5	0.4595	93.86	<.0001
3	BP	0.4801	17.35	<.0001
4	AGE*SEX	0.4957	13.56	0.0003
5	BMI*BP	0.5066	9.60	0.0021
6	S3	0.5166	9.00	0.0029
7	SEX	0.5340	16.23	<.0001
-----				
8	S6*S6	0.5399	5.53	0.0192
9	AGE*AGE	0.5426	2.58	0.1091

10	BP*S6	0.5446	1.88	0.1705
...				

Our method produces  $\hat{\alpha} = 0.012$ , and selects 7 variables in the model. For comparison with LARS, Table 6.3 shows the variables included in the model by FSR and LARS. The variables are listed in the order of entering the model. LARS is using a  $C_p$ -type selection criterion to decide the stopping place. For a main effects model, FSR and LARS produce very similar models whose model  $R^2$  are almost identical. For a quadratic model, the variables selected by FSR are all included in the LARS model. The model  $R^2$  for LARS is a little bit higher than that for FSR. Notice that in this situation FSR only includes about half of the variables selected by LARS. This is reasonable because FSR controls the false selection rate, and tends to select smaller models when there are more unimportant variables. Note that we did not enforce a hierarchy principle where main effects would be entered before quadratic effects.

Table 6.3: Variables selected by the FSR method and LARS for the diabetes data

Model	Method	Variables in the model	Model $R^2$
Main Effects	FSR	BMI, S5, BP, S1, SEX, S2	0.5149
	LARS	S5, BMI, BP, S3, SEX, S6, S1	0.5146
Quadratic	FSR	BMI, S5, BP, AGE*SEX, BMI*BP, S3, SEX	0.5340
	LARS	BMI, S5, BP, S3, BMI*BP, AGE*SEX, S6 <sup>2</sup> , BMI <sup>2</sup> , AGE*BP, AGE*S6, SEX, S6, AGE*S5, AGE <sup>2</sup> , SEX*BP, BP*S3	0.5493

### 6.3 The TPLO Study

Cranial cruciate ligament (CrCL) rupture is a common orthopedic problem in dogs. Tibial Plateau Leveling Osteotomy (TPLO) is a surgical technique that treats CrCL disease by providing functional stability of the stifle joint during weight bearing by decreasing cranial tibial thrust. A retrospective study was conducted for the purpose of investigating radiographic changes of the tibial tuberosity following TPLO surgery and identifying clinical findings and risk factors associated with such changes. The investigators reviewed medical records of 243 dogs that underwent TPLO surgery between March 2000 and February 2002, among which 57 cases were excluded from the study because reevaluation radiographs were not available. The resulting group of 186 dogs included 219 stifles, and TPLO surgery was performed on every stifle reviewed. Reevaluation radiographs were examined for evidence of radiographic changes (fracture and caudal rotation of the proximal tibial tuberosity or non-displaced fracture of the tibial tuberosity) in the tibial tuberosity. Whether or not radiographic lucency/fracture were seen in the reevaluation radiographs (variable LUCENCY in Table 6.4) is the main end point of this study. A total of 19 of the 219 subjects (8.68%) exhibited radiographic changes in the tibial tuberosity. The seven possible risk factors surveyed for fracture of the tibial tuberosity are described in Table 6.4. Logistic regression (PROC LOGISTIC in SAS) was applied for data analysis and assessment of the risk factors and the two-way interactions for these end points. Rather than use variable TTAREA and TTWIDTH directly, we take the inverses of them as  $1/TTAREA$  and  $1/TTWIDTH$ , denoted by RAREA and RWIDTH, separately.

With all two-way interactions included, the number of potential predictors is 28 (7 main variables plus 21 two-way interactions). Thus, we generate 28 pseudo-variables. When we check the two-way interactions, the hierarchy principle is enforced in forward

Table 6.4: Description of variables in the TPLO study

Variable	Name	Description
1	LUCENCY	radiographic changes in the tibial tuberosity (1 = radiographic lucency/fracture; 0 = no changes)
2	BLADE	The size of blade
3	TTAREA	approximate tibial tuberosity area
4	TTWIDTH	Approximate average tibial tuberosity width
5	AGE	Age at the time of surgery
6	WT	Body weight at the time of surgery
7	UNIBI	Whether unilateral or single session bilateral surgery had been performed
8	PIN	Location of the anti-rotational pin

selection (i.e., HIER = single in SAS PROC LOGISTIC) such that the interaction effect can enter the model only when both main effects have already been entered. (We enforce the hierarchy principle here because the study investigators were asked by a referee to do this.) Correspondingly, the generation of pseudo-variables is modified to ensure the same structure in the set of pseudo-variables as in the set of original variables. In particular, we first generate 7 main pseudo-variables based on the 7 main original variables, and then by taking the two-way interactions of the main pseudo-variables we produce the other 21 pseudo-variables. In this way, under the enforcement of hierarchy, the interaction pseudo-variables will not enter the model unless both main pseudo-variables are in the model already. The final  $\hat{\alpha}_0$  found by FSR is 0.06. The sequence produced by forward selection is listed as follows.

#### Summary of Forward Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq
------	-------------------	----	--------------	---------------------	------------

1	unibi	1	1	13.7217	0.0002
2	pin	1	2	7.5999	0.0058
3	rwidth	1	3	4.7666	0.0290
4	wt	1	4	4.7077	0.0300
5	age	1	5	4.4516	0.0349
-----					
6	rwidth*age	1	6	1.1909	0.2752
7	age*wt	1	7	3.0756	0.0795
8	pin*unibi	1	8	1.0616	0.3029
9	rwidth*unibi	1	9	1.2726	0.2593
10	rarea	1	10	0.5819	0.4456
11	age*rarea	1	11	3.0063	0.0829
...					

Forward selection with  $\alpha = 0.06$  stops after step 5 and the final model includes 5 variables: UNIBI, PIN, RWIDTH, WT, and AGE. As a result, whether unilateral or single session bilateral surgery had been performed, location of the anti-rotational pin, average tibial tuberosity width, body weight, and age are identified as significant risk factors for fractures. No interactions are identified to affect the radiographic changes.

## 6.4 The UMARU IMPACT Study

This data set, given by Hosmer and Lemeshow (2000), is a subset of data from the University of Massachusetts Aids Research Unit (UMARU) IMPACT Study (UIS). Basically, this study is comprised of two concurrent randomized trials of residential treatment for drug abuse. There are two treatment program sites, referred to as A and B in this text. The trial at site A was a comparison of 3- and 6-month modified therapeutic communities that incorporated elements of health education and relapse prevention. In the trial at site B, clients were randomized to receive either a 6- or 12-month therapeutic community program. For simplifying the study, the planned duration is represented as short versus long at each site in this text. In this data set, there are 9 variables that are described



in Table 6.5. The original purpose of the study was to compare treatment programs of different planned durations designed to reduce drug abuse. One outcome of considerable public health interest is whether or not a subject remained drug free for at least one year from randomization to treatment (DFREE in Table 6.5). There are 575 subjects contained in the data set, among which a total of 147 (25.57%) remained drug free for at least one year. Considering the dichotomous observations for the outcome of interest, we use logistic regression to study the relationship between variable DFREE and the other 7 variables (exclude variable ID). It is of interest to identify the risk variables associated with DFREE.

Table 6.5: Description of variables in the UMARU IMPACT study

Variable	Name	Codes/Values	Description
1	ID	1-575	Identification Code
2	AGE	years	Age at Enrollment
3	BECK	0.000-54.000	Beck Depression Score at Admission
4	IVHX	1=Never, 2=Previous 3 = Recent	IV Drug Use History at Admission
5	NDRUGTX	0-40	Number of Prior Drug Treatments
6	RACE	0 = White, 1 = Other	Subject's Race
7	TREAT	0 = Short, 1 = Long	Treatment Randomization Assignment
8	SITE	0 = A, 1 = B	Treatment Site
9	DFREE	1 = Remained Drug Free 0 = Otherwise	Returned to Drug Use Prior to the Scheduled End of the Treatment Program

As mentioned by Hosmer and Lemeshow (2000), RACE is a clinically important

variable, and hence it is included into the model. Also, we keep variable SITE in the model due to the design of the study, i.e., subjects were randomized to treatment within site. After including RACE and SITE into the model, we then run forward selection on the remaining 5 main variables (i.e., AGE, BECK, IVHX, NDRUGTX, and TREAT) and all possible two-way interactions. Like the dog TPLO study, we enforce the hierarchy principle so that main variables enter the model before their interactions. Notice that IVHX is a categorical variable with three levels. We use three design variables, IVHX\_1, IVHX\_2 and IVHX\_3, to replace it so that each one corresponds to a level, i.e.,

$$\text{IVHX\_1} = 1, \text{ if IVHX} = 1; 0, \text{ otherwise,}$$

$$\text{IVHX\_2} = 1, \text{ if IVHX} = 2; 0, \text{ otherwise,}$$

$$\text{IVHX\_3} = 1, \text{ if IVHX} = 3; 0, \text{ otherwise.}$$

We let forward selection decide by itself which design variables to choose. Once any two of them are in the model the third will never enter. Because we are doing forward selection, we do not run into the problem of a non full rank model. The summary of forward selection is as follows.

#### Summary of Forward Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq
1	ivhx_1	1	3	9.0722	0.0026
2	age	1	4	6.3669	0.0116
3	race*site	1	5	7.4846	0.0062
4	ndrugtx	1	6	6.4501	0.0111
5	treat	1	7	5.7759	0.0162
6	ndrugtx*age	1	8	4.6635	0.0308
<hr/>					
7	race*ndrugtx	1	9	2.5322	0.1115
8	treat*age	1	10	1.9107	0.1669
9	ivhx_1*age	1	11	1.8462	0.1742
10	race*treat	1	12	1.5627	0.2113

11	treat*ivhx_1	1	13	1.4014	0.2365
12	site*ivhx_1	1	14	0.5572	0.4554
13	ndrugtx*ivhx_1	1	15	0.5153	0.4728
14	site*ndrugtx	1	16	0.2962	0.5863
...					

For our FSR method, we generate the pseudo-variables in the same way as we did in Section 6.3 to enforce the same structures in pseudo-variables and real variables. Thus, we generated five pseudo-variables and formed their interactions. The  $\hat{\alpha}_0$  found is 0.064, and hence forward selection stops in step 6. Then, the final model will include the variables entering in the first 6 steps of forward selection, plus RACE and SITE which were forced into the model first. In summary, the risk main variables identified to be associated with DFREE are RACE, SITE, AGE, NDRUGTX, TREAT, and IVHX\_1. And the risk interactions identified are RACE×SITE and NDRUGTX×AGE. This result is consistent with that given by Hosmer and Lemeshow (2000).

# Chapter 7

## Discussion and Conclusion

The focus of this research has been to develop a general approach to controlling variable selection in regression settings. We have made no assumptions about the type of models to which the method is specified. It should apply to a broad range of situations that use forward selection. We have studied its behavior in linear regression and logistic regression by simulation studies in Chapter 4 and 5. In terms of prediction accuracy, our FSR method exhibits significant advantage over other commonly used traditional selection criteria, like  $C_p$  and AIC, except that it appears a bit conservative when there are many weak important variables. The FSR method tends to select a parsimonious model resulting in an easy explanation of the data. It has been applied to four real data in Chapter 6.

An appealing aspect of our approach is its control of the model false selection rate. The goal is to maintain a relatively small percentage of unimportant variables (i.e., false selected), while selecting as many important variables as possible. The model false selection rate is particularly meaningful in certain bioinformatics and statistical genetics applications. For example, in Quantitative Trait Loci (QTL) mapping, a major concern is determining how many of the markers (predictors) contained in the selected model are correct. Also, for identification of useful genes, if a high proportion of the identified genes are not really important, that will waste millions of dollars and precious time when

the scientists check the selected genes in lab but find them useless. The model false selection rate gives the users a meaningful error measure on their selection. In the FSR method, the false selection rate is set to a value  $\gamma$ , and users are able to move it up or down depending on whether they want their procedure to tend towards overfitting or underfitting. In the simulation studies, we have shown that the method performs well in achieving its target control over the false selection rate. It turns out that other commonly used selection criteria, like AIC and BIC, lead to higher false selection rates. Although they may have larger power to identify the weak important variables, they have a tendency to select too many noise variables.

For forward selection, the naive selection criterion is to choose a fixed significance level  $\alpha$ , e.g.,  $\alpha = 0.1$ . It is easily seen that our method is superior to the naive method in that it searches for the optimal level  $\alpha$  for a given data set. Before finding the current procedure to tune the parameter  $\alpha$ , we originally thought to estimate the underlying dimension of the true model. To this end, a number of methods were developed (see Appendix C for details). However, our preliminary simulation studies reveal that they do not work as well as our current tuning procedure. It seems that estimating model size is not as feasible as estimating the model false selection rate.

The idea behind the FSR method is to generate a set of noise variables, and then monitor the number of them selected in the model. The false selection rate is indicative of model overfitting and underfitting, such that we can find a model with the best bias and variance trade-off. A nice property of our method is its robustness to the number of pseudo-variables generated. In fact, this idea of adding pseudo-variables is not restricted to forward selection; instead it can be applied more widely. For example, we can use it in backward elimination or even tree pruning.

In practice, variable selection is a very complicated problem due to the complexity and large variation of real data. Successful modeling of a complex data set is part science,

part statistical methods, and part experience and common sense. Our work is to provide the users with a statistical method that, when applied thoughtfully, will help select the “best” model within the constraints of the available data.

# Appendix A

## Tables of Simulation Results

### A.1 Results for Linear Regression Models

1. The results of average ME and model size for Best, FSR, Tune,  $C_p$ , and LB: Tables A.1a - A.4b
2. The results of average ME, model size and Monte Carlo estimated  $\gamma$  for FSR and Lasso: Tables A.5a - A.6c

### A.2 Results for Logistic Regression Models

1. The results of average SME, LE and mode size for FSR, AIC, and BIC when  $P = 0.1$  and  $R^2 = 0.75$ : Tables A.7a - A.8c
2. The results of average SME, LE and mode size for FSR, AIC, and BIC when  $P = 0.5$  and  $R^2 = 0.75$ : Tables A.9a - A.10c
3. The results of average SME, LE and mode size for FSR, AIC, and BIC when  $P = 0.1$  and  $R^2 = 0.35$ : Tables A.11a - A.12c
4. The results of average SME, LE and mode size for FSR, AIC, and BIC when  $P = 0.5$  and  $R^2 = 0.35$ : Tables A.13a - A.14c
5. The results of Monte Carlo estimated  $\gamma$  for FSR, AIC, and BIC : Tables A.15a - A.15h

Table A.1a: Average model error for Best, FSR, Tune,  $C_p$ , and LB for linear regression ( $n = 50$  and  $k_T = 21$ )

$\rho$	Method	H0	H1	H2	H3	H4	Ave. Std. Err.
0	Best	0.02	0.06	0.20	0.28	0.35	0.009
	FSR	0.05	0.13	0.29	0.40	0.49	0.015
	Tune	0.06	0.12	0.31	0.38	0.49	0.022
	$C_p$	0.26	0.31	0.33	0.37	0.41	0.016
	LB	0.21	0.21	0.35	0.42	0.44	0.017
0.7	Best	0.02	0.06	0.20	0.26	0.27	0.008
	FSR	0.04	0.13	0.27	0.35	0.37	0.012
	Tune	0.05	0.13	0.33	0.43	0.42	0.015
	$C_p$	0.27	0.30	0.34	0.37	0.38	0.014
	LB	0.19	0.20	0.33	0.37	0.39	0.016

Table A.1b: Average model size for Best, FSR, Tune,  $C_p$ , and LB for linear regression ( $n = 50$  and  $k_T = 21$ )

$\rho$	Method	H0	H1	H2	H3	H4	Ave. Std. Err.
0	True	0.00	2.00	6.00	10.00	14.00	0.00
	Best	0.00	2.00	4.40	6.88	10.14	0.12
	Control	0.25	2.50	4.07	5.31	6.59	0.11
	Tune	0.68	2.86	4.12	5.74	7.13	0.27
	$C_p$	3.56	5.39	7.08	8.19	9.54	0.20
	LB	3.21	4.45	7.91	9.92	12.41	0.46
0.7	True	0.00	2.00	6.00	10.00	14.00	0.00
	Best	0.00	2.00	3.41	5.37	5.22	0.12
	Control	0.12	2.36	2.87	4.00	3.44	0.09
	Tune	0.51	2.77	2.87	3.94	3.71	0.23
	$C_p$	3.71	5.35	6.38	6.84	6.62	0.20
	LB	2.45	4.15	6.07	8.43	8.21	0.47



Table A.2a: Average model error for Best, FSR, Tune,  $C_p$ , and LB for linear regression ( $n = 150$  and  $k_T = 21$ )

$\rho$	Method	H0	H1	H2	H3	H4	Ave. Std. Err.
0	Best	0.01	0.02	0.06	0.08	0.11	0.003
	FSR	0.01	0.03	0.08	0.11	0.13	0.004
	Tune	0.01	0.02	0.08	0.11	0.13	0.004
	$C_p$	0.08	0.09	0.11	0.12	0.13	0.005
	LB	0.07	0.05	0.09	0.13	0.14	0.005
0.7	Best	0.01	0.02	0.06	0.09	0.10	0.003
	FSR	0.01	0.03	0.09	0.11	0.13	0.004
	Tune	0.02	0.03	0.09	0.11	0.15	0.005
	$C_p$	0.08	0.09	0.11	0.12	0.13	0.005
	LB	0.06	0.05	0.11	0.13	0.14	0.005

Table A.2b: Average model size for Best, FSR, Tune,  $C_p$ , and LB for linear regression ( $n = 150$  and  $k_T = 21$ )

$\rho$	Method	H0	H1	H2	H3	H4	Ave. Std. Err.
0	True	0.00	2.00	6.00	10.00	14.00	0.00
	Best	0.00	2.00	4.42	7.04	10.31	0.06
	FSR	0.12	2.20	4.74	7.34	9.83	0.09
	Tune	0.06	2.05	4.62	7.39	9.79	0.08
	$C_p$	3.16	5.10	7.51	9.59	11.49	0.15
	LB	3.03	3.24	6.64	9.86	12.82	0.36
0.7	True	0.00	2.00	6.00	10.00	14.00	0.00
	Best	0.00	2.00	4.50	6.42	8.43	0.13
	FSR	0.08	2.21	4.20	5.72	6.36	0.07
	Tune	0.68	2.10	4.08	5.60	5.75	0.14
	$C_p$	3.13	4.72	6.73	7.95	8.50	0.16
	LB	2.45	3.22	6.39	8.45	10.54	0.40

Table A.3a: Average model error for Best, FSR, Tune,  $C_p$ , and LB for linear regression ( $n = 500$  and  $k_T = 21$ )

$\rho$	Method	H0	H1	H2	H3	H4	Ave. Std. Err.
0	Best	0.00	0.01	0.02	0.02	0.03	0.001
	FSR	0.00	0.01	0.03	0.03	0.04	0.001
	Tune	0.00	0.01	0.03	0.03	0.04	0.001
	$C_p$	0.03	0.03	0.03	0.03	0.04	0.001
	LB	0.02	0.01	0.03	0.04	0.04	0.002
0.7	Best	0.00	0.01	0.02	0.03	0.03	0.001
	FSR	0.00	0.01	0.02	0.03	0.04	0.001
	Tune	0.00	0.01	0.02	0.04	0.04	0.001
	$C_p$	0.03	0.03	0.03	0.04	0.04	0.001
	LB	0.02	0.01	0.03	0.04	0.04	0.001

Table A.3b: Average model size for Best, FSR, Tune,  $C_p$ , and LB for linear regression ( $n = 500$  and  $k_T = 21$ )

$\rho$	Method	H0	H1	H2	H3	H4	Ave. Std. Err.
0	True	0.00	2.00	6.00	10.00	14.00	0.00
	Best	0.00	2.00	5.70	8.20	11.50	0.05
	FSR	0.02	2.18	5.53	8.34	11.29	0.07
	Tune	0.00	2.10	5.20	8.10	10.60	0.07
	$C_p$	3.20	4.90	8.10	10.10	12.50	0.12
	LB	3.00	2.90	8.20	10.40	14.20	0.32
0.7	True	0.00	2.00	6.00	10.00	14.00	0.00
	Best	0.00	2.00	4.40	7.50	10.10	0.08
	FSR	0.09	2.13	4.78	7.06	8.83	0.07
	Tune	0.00	2.00	4.40	6.70	8.10	0.07
	$C_p$	3.20	5.10	7.40	9.00	10.40	0.16
	LB	2.60	3.10	7.00	10.10	11.90	0.35

Table A.4a: Average model error for Best, FSR, Tune,  $C_p$ , and LB for linear regression ( $n = 150$  and  $k_T = 42$ )

$\rho$	Method	H0	H1	H2	H3	H4	Ave. Std. Err.
0	Best	0.01	0.02	0.06	0.08	0.12	0.003
	FSR	0.02	0.03	0.08	0.11	0.16	0.005
	Tune	0.02	0.02	0.10	0.11	0.16	0.005
	$C_p$	0.16	0.18	0.18	0.20	0.22	0.007
	LB	0.07	0.06	0.11	0.14	0.18	0.008
0.7	Best	0.01	0.02	0.08	0.10	0.13	0.003
	FSR	0.01	0.03	0.10	0.13	0.17	0.005
	Tune	0.02	0.03	0.10	0.15	0.16	0.006
	$C_p$	0.17	0.16	0.18	0.20	0.21	0.007
	LB	0.07	0.04	0.12	0.17	0.18	0.005

Table A.4b: Average model size for Best, FSR, Tune,  $C_p$ , and LB for linear regression ( $n = 150$  and  $k_T = 42$ )

$\rho$	Method	H0	H1	H2	H3	H4	Ave. Std. Err.
0	True	0.00	2.00	6.00	10.00	14.00	0.00
	Best	0.00	2.00	5.95	8.22	11.35	0.04
	FSR	0.15	2.28	4.58	6.90	9.40	0.09
	Tune	0.25	2.07	6.12	8.60	11.63	0.09
	$C_p$	6.66	8.69	11.29	13.74	16.51	0.24
	LB	1.86	3.50	6.58	10.23	13.47	0.41
0.7	True	0.00	2.00	6.00	10.00	14.00	0.00
	Best	0.00	2.00	4.77	7.85	9.62	0.11
	FSR	0.14	2.19	3.86	5.27	5.91	0.08
	Tune	1.13	2.92	5.55	7.38	9.06	0.35
	$C_p$	6.92	7.98	10.87	12.74	14.75	0.26
	LB	1.80	2.72	6.59	9.77	12.13	0.52

Table A.5a: Average model error for FSR and Lasso for linear regression ( $n = 150$ ,  $k_T = 21$  and  $\epsilon \sim N(0, 1)$ )

$\rho$	Method	H0	H1	H2	H3	H4
0	FSR	0.014(0.003)	0.031(0.003)	0.078(0.004)	0.106(0.004)	0.134(0.005)
	Lasso	0.016(0.003)	0.065(0.004)	0.095(0.004)	0.118(0.005)	0.130(0.004)
0.7	FSR	0.010(0.002)	0.033(0.003)	0.085(0.005)	0.109(0.004)	0.134(0.006)
	Lasso	0.013(0.001)	0.057(0.003)	0.076(0.003)	0.090(0.004)	0.092(0.004)

Table A.5b: Average model size for FSR and Lasso for linear regression ( $n = 150$ ,  $k_T = 21$ , and  $\epsilon \sim N(0, 1)$ )

$\rho$	Method	H0 ( $k_I = 0$ )	H1 ( $k_I = 2$ )	H2 ( $k_I = 6$ )	H3 ( $k_I = 10$ )	H4 ( $k_I = 14$ )
0	FSR	0.12(0.04)	2.20(0.05)	4.74(0.09)	7.34(0.13)	9.83(0.14)
	Lasso	1.84(0.30)	6.47(0.33)	11.05(0.31)	13.73(0.28)	16.31(0.24)
0.7	FSR	0.08(0.03)	2.21(0.05)	4.20(0.08)	5.72(0.11)	6.36(0.11)
	Lasso	1.54(0.23)	6.00(0.26)	9.28(0.30)	11.12(0.26)	12.41(0.20)

Table A.5c: The Monte Carlo estimated  $\gamma$  for FSR and Lasso for linear regression ( $n = 150$ ,  $k_T = 21$ , and  $\epsilon \sim N(0, 1)$ )

$\rho$	Method	H0	H1	H2	H3	H4
0	FSR	0.107	0.063	0.064	0.059	0.039
	Lasso	0.617	0.596	0.460	0.345	0.235
0.7	FSR	0.074	0.065	0.087	0.054	0.023
	Lasso	0.615	0.576	0.419	0.282	0.156

Table A.6a: Average model error for FSR and Lasso for linear regression ( $n = 150$ ,  $k_T = 21$ , and  $\epsilon \sim \exp(1) - 1$ )

$\rho$	Method	H0	H1	H2	H3	H4
0	FSR	0.013(0.003)	0.036(0.004)	0.082(0.004)	0.102(0.005)	0.132(0.006)
	Lasso	0.013(0.002)	0.074(0.005)	0.098(0.005)	0.113(0.004)	0.131(0.006)
0.7	FSR	0.014(0.003)	0.038(0.004)	0.073(0.005)	0.104(0.005)	0.138(0.005)
	Lasso	0.017(0.003)	0.068(0.005)	0.072(0.004)	0.087(0.005)	0.092(0.004)

Table A.6b: Average model size for FSR and Lasso for linear regression ( $n = 150$ ,  $k_T = 21$ , and  $\epsilon \sim \exp(1) - 1$ )

$\rho$	Method	H0 ( $k_I = 0$ )	H1 ( $k_I = 2$ )	H2 ( $k_I = 6$ )	H3 ( $k_I = 10$ )	H4 ( $k_I = 14$ )
0	FSR	0.07(0.03)	2.22(0.06)	4.75(0.09)	7.20(0.11)	9.77(0.14)
	Lasso	1.63(0.24)	6.20(0.32)	10.31(0.31)	14.15(0.27)	16.84(0.24)
0.7	FSR	0.12(0.04)	2.22(0.05)	4.06(0.07)	5.79(0.11)	6.64(0.13)
	Lasso	1.95(0.31)	6.51(0.30)	8.78(0.28)	11.15(0.28)	12.91(0.22)

Table A.6c: The Monte Carlo estimated  $\gamma$  for FSR and Lasso for linear regression ( $n = 150$ ,  $k_T = 21$ , and  $\epsilon \sim \exp(1) - 1$ )

$\rho$	Method	H0	H1	H2	H3	H4
0	FSR	0.065	0.068	0.073	0.043	0.028
	Lasso	0.600	0.583	0.461	0.347	0.241
0.7	FSR	0.107	0.068	0.053	0.054	0.047
	Lasso	0.647	0.597	0.370	0.259	0.176

Table A.7a: Average SME for FSR, AIC, and BIC for logistic regression ( $n = 150, P = 0.1, R^2 = 0.75$ )

$\rho$	Method	H0	H1	H2	H3	H4
0	FSR	0.017(0.004)	0.040(0.006)	0.158(0.012)	0.724(0.141)	1.466(0.253)
	AIC	0.131(0.009)	0.275(0.018)	0.366(0.033)	0.433(0.052)	0.404(0.038)
	BIC	0.051(0.006)	0.105(0.012)	0.223(0.024)	0.360(0.051)	0.416(0.041)
0.7	FSR	0.015(0.003)	0.059(0.008)	0.112(0.015)	0.208(0.018)	0.396(0.109)
	AIC	0.119(0.009)	0.297(0.031)	0.329(0.025)	0.378(0.046)	0.354(0.034)
	BIC	0.037(0.005)	0.113(0.014)	0.176(0.019)	0.214(0.014)	0.248(0.033)

Table A.7b: Average LE for FSR, AIC, and BIC for logistic regression ( $n = 150, P = 0.1, R^2 = 0.75$ )

$\rho$	Method	H0	H1	H2	H3	H4
0	FSR	0.22(0.06)	9.94(4.44)	20.46(6.75)	18.51(6.25)	169.43(136.61)
	AIC	3.06(0.42)	2084.9(391.5)	2821.6(598.4)	2545.7(515.2)	2182.6(392.7)
	BIC	0.91(0.15)	523.28(276.52)	980.74(292.33)	854.14(310.21)	1457.4(333.1)
0.7	FSR	0.16(0.04)	10.38(5.40)	4.60(0.99)	16.75(6.46)	10.41(4.23)
	AIC	2.55(0.39)	2577.8(765.6)	3065.1(764.2)	5541.1(963.2)	3687.0(816.3)
	BIC	0.48(0.07)	394.16(112.99)	658.08(321.36)	2137.6(690.2)	1425.3(576.6)

Table A.7c: Average model size for FSR, AIC, and BIC for logistic regression ( $n = 150, P = 0.1, R^2 = 0.75$ )

$\rho$	Method	H0 ( $k_I = 0$ )	H1 ( $k_I = 2$ )	H2 ( $k_I = 6$ )	H3 ( $k_I = 10$ )	H4 ( $k_I = 14$ )
0	FSR	0.14(0.04)	2.24(0.05)	3.08(0.10)	3.49(0.13)	4.23(0.14)
	AIC	4.07(0.23)	9.31(0.43)	9.31(0.35)	10.62(0.36)	10.65(0.27)
	BIC	0.87(0.11)	3.64(0.22)	4.97(0.25)	5.90(0.23)	8.16(0.27)
0.7	FSR	0.11(0.04)	2.30(0.06)	2.28(0.05)	2.56(0.08)	2.73(0.07)
	AIC	3.90(0.27)	8.66(0.43)	9.07(0.39)	9.54(0.35)	9.65(0.38)
	BIC	0.55(0.08)	3.71(0.24)	3.89(0.25)	4.43(0.26)	4.79(0.24)

Table A.8a: Average SME for FSR, AIC, and BIC for logistic regression ( $n = 500, P = 0.1, R^2 = 0.75$ )

$\rho$	Method	H0	H1	H2	H3	H4
0	FSR	0.003(0.001)	0.010(0.001)	0.022(0.002)	0.041(0.002)	0.057(0.003)
	AIC	0.027(0.002)	0.034(0.002)	0.040(0.002)	0.047(0.002)	0.051(0.002)
	BIC	0.005(0.001)	0.011(0.001)	0.021(0.002)	0.042(0.003)	0.056(0.003)
0.7	FSR	0.004(0.001)	0.010(0.001)	0.030(0.002)	0.044(0.003)	0.056(0.003)
	AIC	0.030(0.002)	0.034(0.002)	0.038(0.002)	0.044(0.002)	0.047(0.002)
	BIC	0.007(0.001)	0.011(0.001)	0.030(0.002)	0.044(0.003)	0.058(0.003)

Table A.8b: Average LE for FSR, AIC, and BIC for logistic regression ( $n = 500, P = 0.1, R^2 = 0.75$ )

$\rho$	Method	H0	H1	H2	H3	H4
0	FSR	0.03(0.01)	0.79(0.13)	1.05(0.18)	1.56(0.17)	2.19(0.25)
	AIC	0.32(0.02)	2.06(0.29)	2.33(0.27)	2.73(0.30)	3.38(0.42)
	BIC	0.06(0.01)	0.83(0.14)	1.00(0.18)	1.49(0.15)	1.89(0.19)
0.7	FSR	0.05(0.01)	0.79(0.11)	0.96(0.07)	1.38(0.09)	2.02(0.26)
	AIC	0.36(0.02)	2.08(0.23)	1.89(0.20)	2.52(0.22)	2.70(0.33)
	BIC	0.08(0.01)	0.84(0.12)	0.92(0.06)	1.32(0.09)	1.92(0.22)

Table A.8c: Average model size for FSR, AIC, and BIC for logistic regression ( $n = 500, P = 0.1, R^2 = 0.75$ )

$\rho$	Method	H0 ( $k_I = 0$ )	H1 ( $k_I = 2$ )	H2 ( $k_I = 6$ )	H3 ( $k_I = 10$ )	H4 ( $k_I = 14$ )
0	FSR	0.03(0.02)	2.20(0.06)	4.34(0.07)	6.14(0.10)	7.80(0.15)
	AIC	3.30(0.19)	5.50(0.18)	7.54(0.16)	9.45(0.17)	10.93(0.20)
	BIC	0.20(0.04)	2.28(0.06)	4.32(0.06)	5.98(0.09)	7.20(0.11)
0.7	FSR	0.10(0.03)	2.18(0.04)	3.20(0.10)	4.40(0.09)	4.79(0.09)
	AIC	3.55(0.18)	5.44(0.20)	6.63(0.19)	7.88(0.21)	7.73(0.19)
	BIC	0.28(0.06)	2.23(0.05)	3.24(0.09)	4.24(0.08)	4.59(0.07)

Table A.9a: Average SME for FSR, AIC, and BIC for logistic regression ( $n = 150, P = 0.5, R^2 = 0.75$ )

$\rho$	Method	H0	H1	H2	H3	H4
0	FSR	0.017(0.003)	0.027(0.003)	0.097(0.009)	0.144(0.007)	0.225(0.013)
	AIC	0.102(0.005)	0.120(0.007)	0.139(0.008)	0.156(0.008)	0.186(0.012)
	BIC	0.036(0.004)	0.039(0.004)	0.091(0.007)	0.142(0.007)	0.201(0.013)
0.7	FSR	0.010(0.003)	0.037(0.004)	0.108(0.006)	0.163(0.014)	0.222(0.018)
	AIC	0.092(0.005)	0.136(0.009)	0.138(0.007)	0.141(0.006)	0.163(0.008)
	BIC	0.032(0.004)	0.048(0.005)	0.106(0.006)	0.145(0.010)	0.181(0.010)

Table A.9b: Average LE for FSR, AIC, and BIC for logistic regression ( $n = 150, P = 0.5, R^2 = 0.75$ )

$\rho$	Method	H0	H1	H2	H3	H4
0	FSR	0.08(0.01)	0.58(0.07)	1.42(0.16)	2.16(0.20)	2.95(0.30)
	AIC	0.53(0.03)	5.04(1.02)	3.72(0.43)	9.15(3.15)	6.93(1.21)
	BIC	0.17(0.02)	0.88(0.12)	1.59(0.18)	2.45(0.24)	3.31(0.36)
0.7	FSR	0.05(0.01)	0.80(0.15)	1.87(0.25)	1.89(0.20)	1.94(0.10)
	AIC	0.47(0.03)	5.47(1.26)	231.30(222.51)	4.59(0.83)	4.56(0.73)
	BIC	0.16(0.02)	1.17(0.28)	2.46(0.72)	1.97(0.21)	2.19(0.30)

Table A.9c: Average model size for FSR, AIC, and BIC for logistic regression ( $n = 150, P = 0.5, R^2 = 0.75$ )

$\rho$	Method	H0 ( $k_I = 0$ )	H1 ( $k_I = 2$ )	H2 ( $k_I = 6$ )	H3 ( $k_I = 10$ )	H4 ( $k_I = 14$ )
0	FSR	0.20(0.05)	2.11(0.03)	4.07(0.08)	5.71(0.12)	7.15(0.14)
	AIC	4.15(0.17)	5.98(0.23)	7.63(0.21)	9.36(0.23)	11.16(0.18)
	BIC	0.70(0.10)	2.41(0.07)	4.55(0.09)	6.12(0.12)	7.60(0.14)
0.7	FSR	0.08(0.04)	2.23(0.06)	3.01(0.09)	4.09(0.10)	4.34(0.10)
	AIC	3.96(0.24)	6.41(0.28)	7.13(0.20)	7.62(0.21)	8.33(0.23)
	BIC	0.65(0.10)	2.47(0.08)	3.49(0.11)	4.53(0.11)	4.79(0.11)



Table A.10a: Average SME for FSR, AIC, and BIC for logistic regression ( $n = 500, P = 0.5, R^2 = 0.75$ )

$\rho$	Method	H0	H1	H2	H3	H4
0	FSR	0.003(0.001)	0.010(0.001)	0.023(0.001)	0.034(0.001)	0.041(0.002)
	AIC	0.024(0.001)	0.029(0.002)	0.035(0.002)	0.038(0.002)	0.042(0.002)
	BIC	0.006(0.001)	0.010(0.001)	0.023(0.001)	0.034(0.001)	0.043(0.002)
0.7	FSR	0.003(0.001)	0.010(0.001)	0.025(0.002)	0.038(0.002)	0.051(0.002)
	AIC	0.027(0.002)	0.031(0.002)	0.036(0.002)	0.039(0.002)	0.046(0.002)
	BIC	0.005(0.001)	0.010(0.001)	0.024(0.002)	0.042(0.002)	0.058(0.003)

Table A.10b: Average LE for FSR, AIC, and BIC for logistic regression ( $n = 500, P = 0.5, R^2 = 0.75$ )

$\rho$	Method	H0	H1	H2	H3	H4
0	FSR	0.01(0.01)	0.17(0.02)	0.28(0.02)	0.41(0.03)	0.49(0.04)
	AIC	0.10(0.01)	0.43(0.04)	0.45(0.03)	0.52(0.04)	0.56(0.05)
	BIC	0.02(0.01)	0.18(0.02)	0.28(0.02)	0.40(0.03)	0.47(0.03)
0.7	FSR	0.01(0.01)	0.15(0.01)	0.28(0.02)	0.42(0.02)	0.55(0.03)
	AIC	0.12(0.01)	0.42(0.03)	0.42(0.03)	0.50(0.03)	0.58(0.04)
	BIC	0.02(0.01)	0.15(0.01)	0.27(0.02)	0.44(0.02)	0.57(0.03)

Table A.10c: Average model size for FSR, AIC, and BIC for logistic regression ( $n = 500, P = 0.5, R^2 = 0.75$ )

$\rho$	Method	H0 ( $k_I = 0$ )	H1 ( $k_I = 2$ )	H2 ( $k_I = 6$ )	H3 ( $k_I = 10$ )	H4 ( $k_I = 14$ )
0	FSR	0.07(0.03)	2.22(0.05)	4.58(0.08)	7.14(0.10)	10.08(0.12)
	AIC	3.11(0.16)	5.13(0.19)	7.42(0.17)	9.72(0.16)	11.89(0.15)
	BIC	0.24(0.05)	2.28(0.06)	4.48(0.06)	6.85(0.10)	9.22(0.10)
0.7	FSR	0.09(0.03)	2.22(0.06)	4.33(0.07)	5.68(0.08)	6.84(0.12)
	AIC	3.79(0.21)	5.54(0.20)	7.26(0.19)	8.30(0.18)	9.24(0.16)
	BIC	0.24(0.05)	2.22(0.05)	4.12(0.06)	5.36(0.07)	6.10(0.08)

Table A.11a: Average SME for FSR, AIC, and BIC for logistic regression ( $n = 150, P = 0.1, R^2 = 0.35$ )

$\rho$	Method	H0	H1	H2	H3	H4
0	FSR	0.017(0.004)	0.135(0.017)	0.173(0.019)	0.228(0.013)	0.292(0.014)
	AIC	0.131(0.009)	0.173(0.013)	0.182(0.013)	0.206(0.012)	0.261(0.020)
	BIC	0.051(0.006)	0.099(0.011)	0.147(0.017)	0.186(0.011)	0.264(0.017)
0.7	FSR	0.015(0.003)	0.123(0.011)	0.149(0.013)	0.131(0.009)	0.152(0.010)
	AIC	0.119(0.009)	0.190(0.014)	0.207(0.018)	0.192(0.026)	0.187(0.014)
	BIC	0.037(0.005)	0.104(0.010)	0.131(0.011)	0.107(0.008)	0.130(0.010)

Table A.11b: Average LE for FSR, AIC, and BIC for logistic regression ( $n = 150, P = 0.1, R^2 = 0.35$ )

$\rho$	Method	H0	H1	H2	H3	H4
0	FSR	0.22(0.05)	1.35(0.11)	1.53(0.10)	2.02(0.15)	2.37(0.16)
	AIC	3.06(0.42)	474.73(248.88)	301.69(145.45)	111.91(59.68)	150.95(53.17)
	BIC	0.91(0.15)	188.38(111.46)	22.38(13.20)	14.62(8.34)	49.95(34.37)
0.7	FSR	0.16(0.03)	2.13(0.43)	1.81(0.20)	2.98(1.11)	1.82(0.16)
	AIC	2.55(0.39)	433.37(210.30)	242.81(109.20)	198.95(120.04)	264.79(141.92)
	BIC	0.48(0.07)	174.51(166.66)	7.05(4.83)	18.53(9.70)	32.65(21.75)

Table A.11c: Average model size for FSR, AIC, and BIC for logistic regression ( $n = 150, P = 0.1, R^2 = 0.35$ )

$\rho$	Method	H0 ( $k_I = 0$ )	H1 ( $k_I = 2$ )	H2 ( $k_I = 6$ )	H3 ( $k_I = 10$ )	H4 ( $k_I = 14$ )
0	FSR	0.14(0.04)	1.36(0.09)	1.23(0.10)	1.10(0.11)	1.22(0.13)
	AIC	4.07(0.23)	6.47(0.30)	6.93(0.34)	7.56(0.30)	8.31(0.30)
	BIC	0.87(0.11)	2.71(0.21)	2.41(0.16)	2.79(0.20)	3.16(0.24)
0.7	FSR	0.11(0.03)	1.49(0.08)	1.46(0.08)	1.61(0.08)	1.61(0.09)
	AIC	3.90(0.27)	6.68(0.38)	7.05(0.41)	6.39(0.33)	6.34(0.40)
	BIC	0.55(0.08)	2.41(0.17)	2.33(0.13)	2.28(0.11)	2.46(0.16)

Table A.12a: Average SME for FSR, AIC, and BIC for logistic regression ( $n = 500, P = 0.1, R^2 = 0.35$ )

$\rho$	Method	H0	H1	H2	H3	H4
0	FSR	0.003(0.001)	0.008(0.001)	0.028(0.002)	0.044(0.004)	0.062(0.004)
	AIC	0.027(0.002)	0.030(0.002)	0.037(0.002)	0.039(0.002)	0.047(0.002)
	BIC	0.005(0.001)	0.008(0.001)	0.025(0.001)	0.040(0.002)	0.064(0.004)
0.7	FSR	0.003(0.001)	0.011(0.002)	0.027(0.002)	0.034(0.001)	0.042(0.002)
	AIC	0.030(0.002)	0.030(0.002)	0.039(0.002)	0.038(0.002)	0.040(0.002)
	BIC	0.007(0.001)	0.012(0.002)	0.027(0.002)	0.033(0.001)	0.042(0.002)

Table A.12b: Average LE for FSR, AIC, and BIC for logistic regression ( $n = 500, P = 0.1, R^2 = 0.35$ )

$\rho$	Method	H0	H1	H2	H3	H4
0	FSR	0.03(0.006)	0.16(0.02)	0.38(0.02)	0.58(0.03)	0.77(0.04)
	AIC	0.32(0.021)	0.54(0.04)	0.63(0.04)	0.68(0.05)	0.78(0.05)
	BIC	0.06(0.01)	0.17(0.02)	0.36(0.02)	0.54(0.03)	0.76(0.04)
0.7	FSR	0.04(0.01)	0.19(0.02)	0.41(0.03)	0.49(0.02)	0.57(0.03)
	AIC	0.36(0.02)	0.53(0.04)	0.69(0.05)	0.69(0.05)	0.68(0.04)
	BIC	0.08(0.01)	0.20(0.02)	0.41(0.03)	0.48(0.02)	0.57(0.03)

Table A.12c: Average model size for FSR, AIC, and BIC for logistic regression ( $n = 500, P = 0.1, R^2 = 0.35$ )

$\rho$	Method	H0 ( $k_I = 0$ )	H1 ( $k_I = 2$ )	H2 ( $k_I = 6$ )	H3 ( $k_I = 10$ )	H4 ( $k_I = 14$ )
0	FSR	0.03(0.02)	2.11(0.03)	2.86(0.10)	4.03(0.12)	4.66(0.16)
	AIC	3.30(0.19)	5.27(0.18)	6.77(0.18)	7.73(0.17)	8.87(0.17)
	BIC	0.20(0.04)	2.17(0.04)	3.01(0.09)	4.02(0.10)	4.46(0.12)
0.7	FSR	0.07(0.03)	2.13(0.04)	2.38(0.07)	2.50(0.08)	2.75(0.08)
	AIC	3.55(0.18)	5.24(0.21)	5.58(0.19)	6.02(0.17)	6.04(0.17)
	BIC	0.28(0.06)	2.17(0.04)	2.38(0.06)	2.50(0.07)	2.71(0.07)

Table A.13a: Average SME for FSR, AIC, and BIC for logistic regression ( $n = 150, P = 0.5, R^2 = 0.35$ )

$\rho$	Method	H0	H1	H2	H3	H4
0	FSR	0.017(0.003)	0.037(0.006)	0.095(0.006)	0.159(0.009)	0.242(0.015)
	AIC	0.102(0.005)	0.098(0.005)	0.116(0.004)	0.135(0.005)	0.141(0.005)
	BIC	0.036(0.004)	0.041(0.005)	0.093(0.005)	0.129(0.006)	0.181(0.011)
0.7	FSR	0.010(0.003)	0.053(0.006)	0.080(0.005)	0.103(0.005)	0.129(0.005)
	AIC	0.092(0.005)	0.106(0.007)	0.117(0.005)	0.122(0.005)	0.134(0.005)
	BIC	0.032(0.004)	0.059(0.006)	0.084(0.005)	0.100(0.004)	0.122(0.004)

Table A.13b: Average LE for FSR, AIC, and BIC for logistic regression ( $n = 150, P = 0.5, R^2 = 0.35$ )

$\rho$	Method	H0	H1	H2	H3	H4
0	FSR	0.08(0.01)	0.23(0.03)	0.54(0.04)	0.77(0.03)	1.01(0.04)
	AIC	0.53(0.03)	0.79(0.06)	0.98(0.10)	1.02(0.06)	1.08(0.07)
	BIC	0.17(0.02)	0.29(0.04)	0.58(0.04)	0.73(0.03)	0.89(0.04)
0.7	FSR	0.05(0.01)	0.36(0.04)	0.48(0.04)	0.61(0.04)	0.70(0.03)
	AIC	0.47(0.03)	0.89(0.07)	0.94(0.08)	0.99(0.07)	1.03(0.06)
	BIC	0.16(0.02)	0.41(0.04)	0.53(0.04)	0.63(0.04)	0.73(0.04)

Table A.13c: Average model size for FSR, AIC, and BIC for logistic regression ( $n = 150, P = 0.5, R^2 = 0.35$ )

$\rho$	Method	H0 ( $k_I = 0$ )	H1 ( $k_I = 2$ )	H2 ( $k_I = 6$ )	H3 ( $k_I = 10$ )	H4 ( $k_I = 14$ )
0	FSR	0.20(0.05)	2.18(0.06)	2.71(0.10)	3.48(0.11)	3.89(0.17)
	AIC	4.15(0.17)	5.35(0.17)	6.68(0.19)	7.71(0.17)	8.77(0.19)
	BIC	0.70(0.10)	2.49(0.08)	3.37(0.10)	4.31(0.12)	4.90(0.14)
0.7	FSR	0.08(0.04)	2.25(0.07)	2.19(0.05)	2.36(0.07)	2.48(0.07)
	AIC	3.96(0.24)	5.38(0.24)	5.61(0.21)	6.11(0.22)	6.27(0.19)
	BIC	0.65(0.10)	2.52(0.08)	2.53(0.08)	2.65(0.08)	3.01(0.10)

Table A.14a: Average SME for FSR, AIC, and BIC for logistic regression ( $n = 500, P = 0.5, R^2 = 0.35$ )

$\rho$	Method	H0	H1	H2	H3	H4
0	FSR	0.003(0.001)	0.009(0.001)	0.027(0.002)	0.033(0.002)	0.040(0.002)
	AIC	0.024(0.001)	0.027(0.001)	0.034(0.002)	0.035(0.002)	0.037(0.002)
	BIC	0.006(0.001)	0.009(0.001)	0.027(0.002)	0.033(0.002)	0.045(0.002)
0.7	FSR	0.003(0.001)	0.009(0.001)	0.028(0.002)	0.037(0.002)	0.048(0.002)
	AIC	0.027(0.002)	0.028(0.002)	0.034(0.002)	0.036(0.002)	0.044(0.002)
	BIC	0.005(0.001)	0.009(0.001)	0.028(0.002)	0.037(0.002)	0.050(0.002)

Table A.14b: Average LE for FSR, AIC, and BIC for logistic regression ( $n = 500, P = 0.5, R^2 = 0.35$ )

$\rho$	Method	H0	H1	H2	H3	H4
0	FSR	0.01(0.00)	0.06(0.01)	0.15(0.01)	0.19(0.01)	0.23(0.01)
	AIC	0.10(0.01)	0.17(0.01)	0.21(0.01)	0.21(0.01)	0.23(0.01)
	BIC	0.02(0.00)	0.06(0.01)	0.15(0.01)	0.18(0.01)	0.24(0.01)
0.7	FSR	0.01(0.00)	0.05(0.00)	0.16(0.01)	0.20(0.01)	0.27(0.01)
	AIC	0.12(0.01)	0.17(0.01)	0.20(0.01)	0.22(0.01)	0.27(0.01)
	BIC	0.02(0.00)	0.06(0.00)	0.15(0.01)	0.21(0.01)	0.27(0.01)

Table A.14c: Average model size for FSR, AIC, and BIC for logistic regression ( $n = 500, P = 0.5, R^2 = 0.35$ )

$\rho$	Method	H0 ( $k_I = 0$ )	H1 ( $k_I = 2$ )	H2 ( $k_I = 6$ )	H3 ( $k_I = 10$ )	H4 ( $k_I = 14$ )
0	FSR	0.07(0.03)	2.18(0.05)	4.09(0.10)	5.81(0.11)	7.74(0.15)
	AIC	3.11(0.16)	4.90(0.15)	6.80(0.17)	8.42(0.16)	10.52(0.14)
	BIC	0.24(0.05)	2.20(0.05)	4.05(0.08)	5.47(0.08)	6.89(0.11)
0.7	FSR	0.09(0.03)	2.22(0.05)	2.94(0.09)	3.93(0.10)	4.18(0.09)
	AIC	3.79(0.21)	5.29(0.18)	6.26(0.18)	6.81(0.16)	7.35(0.16)
	BIC	0.24(0.05)	2.24(0.05)	2.92(0.08)	3.78(0.08)	4.04(0.08)

Table A.15a: The Monte Carlo estimated  $\gamma$  for FSR, AIC and BIC for logistic regression ( $n = 150, P = 0.1, R^2 = 0.75$ )

$\rho$	Method	H0	H1	H2	H3	H4
0	FSR	0.122	0.071	0.047	0.047	0.021
	AIC	0.802	0.709	0.468	0.342	0.171
	BIC	0.465	0.353	0.236	0.165	0.116
0.7	FSR	0.099	0.102	0.071	0.058	0.021
	AIC	0.795	0.692	0.529	0.396	0.226
	BIC	0.354	0.369	0.288	0.230	0.105

Table A.15b: The Monte Carlo estimated  $\gamma$  for FSR, AIC and BIC for logistic regression ( $n = 500, P = 0.1, R^2 = 0.75$ )

$\rho$	Method	H0	H1	H2	H3	H4
0	FSR	0.029	0.054	0.051	0.036	0.036
	AIC	0.767	0.538	0.334	0.215	0.122
	BIC	0.167	0.085	0.047	0.027	0.016
0.7	FSR	0.083	0.042	0.044	0.048	0.029
	AIC	0.780	0.534	0.363	0.251	0.138
	BIC	0.219	0.071	0.047	0.053	0.029

Table A.15c: The Monte Carlo estimated  $\gamma$  for FSR, AIC and BIC for logistic regression ( $n = 150, P = 0.5, R^2 = 0.75$ )

$\rho$	Method	H0	H1	H2	H3	H4
0	FSR	0.153	0.038	0.053	0.043	0.025
	AIC	0.806	0.570	0.360	0.213	0.134
	BIC	0.412	0.120	0.099	0.060	0.035
0.7	FSR	0.074	0.065	0.067	0.046	0.039
	AIC	0.798	0.595	0.402	0.252	0.159
	BIC	0.394	0.135	0.111	0.090	0.050

Table A.15d: The Monte Carlo estimated  $\gamma$  for FSR, AIC and BIC for logistic regression ( $n = 500, P = 0.5, R^2 = 0.75$ )

$\rho$	Method	H0	H1	H2	H3	H4
0	FSR	0.065	0.063	0.030	0.038	0.030
	AIC	0.757	0.511	0.286	0.167	0.089
	BIC	0.194	0.085	0.026	0.025	0.017
0.7	FSR	0.074	0.057	0.047	0.038	0.025
	AIC	0.791	0.541	0.331	0.184	0.119
	BIC	0.194	0.068	0.045	0.031	0.013

Table A.15e: The Monte Carlo estimated  $\gamma$  for FSR, AIC and BIC for logistic regression ( $n = 150, P = 0.1, R^2 = 0.35$ )

$\rho$	Method	H0	H1	H2	H3	H4
0	FSR	0.123	0.076	0.018	0.038	0.027
	AIC	0.802	0.612	0.459	0.319	0.189
	BIC	0.465	0.272	0.167	0.148	0.101
0.7	FSR	0.099	0.213	0.134	0.111	0.054
	AIC	0.795	0.665	0.525	0.371	0.234
	BIC	0.355	0.343	0.255	0.165	0.092

Table A.15f: The Monte Carlo estimated  $\gamma$  for FSR, AIC and BIC for logistic regression ( $n = 500, P = 0.1, R^2 = 0.35$ )

$\rho$	Method	H0	H1	H2	H3	H4
0	FSR	0.029	0.035	0.031	0.026	0.027
	AIC	0.767	0.522	0.349	0.200	0.128
	BIC	0.167	0.054	0.035	0.022	0.022
0.7	FSR	0.065	0.058	0.095	0.034	0.040
	AIC	0.780	0.524	0.391	0.265	0.166
	BIC	0.219	0.073	0.098	0.037	0.038

Table A.15g: The Monte Carlo estimated  $\gamma$  for FSR, AIC and BIC for logistic regression ( $n = 150, P = 0.5, R^2 = 0.35$ )

$\rho$	Method	H0	H1	H2	H3	H4
0	FSR	0.167	0.066	0.057	0.049	0.021
	AIC	0.806	0.528	0.355	0.240	0.126
	BIC	0.412	0.140	0.105	0.083	0.031
0.7	FSR	0.074	0.135	0.085	0.054	0.037
	AIC	0.798	0.545	0.416	0.309	0.215
	BIC	0.394	0.196	0.139	0.085	0.077

Table A.15h: The Monte Carlo estimated  $\gamma$  for FSR, AIC and BIC for logistic regression ( $n = 500, P = 0.5, R^2 = 0.35$ )

$\rho$	Method	H0	H1	H2	H3	H4
0	FSR	0.065	0.057	0.057	0.046	0.025
	AIC	0.757	0.492	0.299	0.173	0.086
	BIC	0.194	0.063	0.050	0.022	0.011
0.7	FSR	0.083	0.068	0.071	0.037	0.027
	AIC	0.791	0.523	0.349	0.218	0.163
	BIC	0.194	0.074	0.069	0.029	0.024



## Appendix B

# Derivation of $\beta_0$ in Logistic Regression Simulation Design

A crucial step of generating the binary response,  $Y_i$ , in Section 5.2.1 is to determine the value of  $\beta_0$  that ensures the unconditional probability that of  $Y_i = 1$  be a certain value. The details are described as follows.

By generation, the probability that  $Y_i = 1$  conditional on  $\mathbf{x}_i$  is such that

$$\begin{aligned} Pr(Y_i = 1|\mathbf{x}_i) &= Pr(Y_i^* > \beta_0|\mathbf{x}_i) \\ &= Pr(\mathbf{x}_i\boldsymbol{\beta} + \epsilon_i > \beta_0|\mathbf{x}_i) \\ &= Pr(\epsilon_i > \beta_0 - \mathbf{x}_i\boldsymbol{\beta}|\mathbf{x}_i) \\ &= 1 - Pr(\epsilon_i \leq \beta_0 - \mathbf{x}_i\boldsymbol{\beta}|\mathbf{x}_i) \end{aligned}$$

Letting  $\epsilon_i$  belong to a logistic distribution with mean 0 and variance 1, we have

$$Pr(\epsilon_i \leq \beta_0 - \mathbf{x}_i\boldsymbol{\beta}|\mathbf{x}_i) = \frac{1}{1 + e^{-\frac{\pi}{\sqrt{3}}\beta_0 + \frac{\pi}{\sqrt{3}}\mathbf{x}_i\boldsymbol{\beta}}}$$

Thus,

$$\begin{aligned} Pr(Y_i = 1|\mathbf{x}_i) &= 1 - \frac{1}{1 + e^{-\frac{\pi}{\sqrt{3}}\beta_0 + \frac{\pi}{\sqrt{3}}\mathbf{x}_i\boldsymbol{\beta}}} \\ &= \frac{e^{-\frac{\pi}{\sqrt{3}}\beta_0 + \frac{\pi}{\sqrt{3}}\mathbf{x}_i\boldsymbol{\beta}}}{1 + e^{-\frac{\pi}{\sqrt{3}}\beta_0 + \frac{\pi}{\sqrt{3}}\mathbf{x}_i\boldsymbol{\beta}}} \end{aligned}$$

Therefore, given fixed design matrix  $\mathbf{X}$ , we have

$$\begin{aligned} Pr(Y_i = 1) &= \frac{1}{n} \sum_{i=1}^n Pr(Y_i = 1 | \mathbf{x}_i) \\ &= \frac{1}{n} \sum_{i=1}^n \frac{e^{-\frac{\pi}{\sqrt{3}}\beta_0 + \frac{\pi}{\sqrt{3}}\mathbf{x}_i\beta}}{1 + e^{-\frac{\pi}{\sqrt{3}}\beta_0 + \frac{\pi}{\sqrt{3}}\mathbf{x}_i\beta}}. \end{aligned}$$

$\beta_0$  is then obtained by solving

$$\frac{1}{n} \sum_{i=1}^n \frac{e^{-\frac{\pi}{\sqrt{3}}\beta_0 + \frac{\pi}{\sqrt{3}}\mathbf{x}_i\beta}}{1 + e^{-\frac{\pi}{\sqrt{3}}\beta_0 + \frac{\pi}{\sqrt{3}}\mathbf{x}_i\beta}} = P,$$

where  $P = 0.1$  and  $0.5$ . Because solving the above equation analytically is difficult, we instead used a numerical approach.

# Appendix C

## Estimating Model Size

Originally, we were trying to estimate the underlying dimension of the true model,  $k_I$ . Although our methods did not show satisfactory performance in preliminary simulation studies, they illustrate some reasonable approaches. It is worthwhile to mention them.

Recall that we generate a set of  $k_P$  pseudo-variables  $B$  times. Let us use  $S_i$  to denote the total number of pseudo-variables entering at the  $i$ th step of  $B$  times forward selection,  $i = 1, 2, \dots, k_T + k_P$ . Notice that  $S_i$  is based on the  $i$ th step of forward selection, and is not related to the significance level  $\alpha$ . Then, we compute the proportion of  $B$  times that a pseudo-variable enters at step  $i$ , i.e.,

$$P_i = \frac{S_i}{B}. \quad (\text{C.1})$$

The probability that an entering variable is a pseudo-variable should be essentially 0 for the first  $k_I$  variables entered provided that the important variables always enter first into the model. Given that the  $k_I$  important variables have been entered, all the other variables have equal status and the probability of a pseudo-variable entering should be  $\frac{k_P}{k_U + k_P}$ . In this sense, we would anticipate that  $P_i$  should be very small when  $i \leq k_I$ , and should jump to  $\frac{k_P}{k_U + k_P}$  when  $i \geq k_I$ . This motivates us to estimate  $k_I$  based on studying the plot of  $P_i$  versus  $i$ . Our methods are as follows.

## C.1 Based on Student's $t$ Test

Given a  $j$  in  $\{1, 2, \dots, k_T\}$ , we divide  $P_i$  into two groups:  $g_1(j) = \{P_i, i \leq j\}$ ;  $g_2(j) = \{P_i, i > j\}$ . Clearly, the size of  $g_1(j)$  is  $j$  and the size of  $g_2(j)$  is  $k_T + k_P - j$ . A Student's  $t$  test statistic testing the equality of means of the two groups is then calculated as

$$t(j) = \frac{|\bar{P}_1(j) - \bar{P}_2(j)|}{\sqrt{\frac{s_1^2(j)}{j} + \frac{s_2^2(j)}{k_T + k_P - j}}},$$

where  $\bar{P}_k(j)$  and  $s_k^2(j)$  are the sample mean and variance of  $P_i$  in group  $g_k(j)$ , respectively,  $k = 1, 2$ . When  $j = 1$ , we take  $s_1^2(j) = 0$ . We then search for the  $j^*$  that maximizes  $t(j)$ , i.e.,

$$t(j^*) = \max_{j \in \{1, \dots, k_T\}} \{t(j)\},$$

and use it as an estimate of  $k_I$ .

Rather than basing  $j^*$  on the test statistics, a modification was made using the P-values. For each  $t(j)$  with  $j > 1$ , the degree of freedom is obtained by

$$df(j) = \frac{\left(\frac{s_1^2(j)}{j} + \frac{s_2^2(j)}{k_T + k_P - j}\right)^2}{\frac{\left(\frac{s_1^2(j)}{j}\right)^2}{j-1} + \frac{\left(\frac{s_2^2(j)}{k_T + k_P - j}\right)^2}{k_T + k_P - j - 1}}, \quad j > 1.$$

If we denote the corresponding P-value by  $pv(j)$ , then  $k_I$  is then estimated by  $j'$  satisfying

$$pv(j') = \min_{j \in \{2, \dots, k_T\}} \{pv(j)\}.$$

An obvious drawback of this method is that  $j'$  is always larger than 1 while  $k_I$  could be 1 or even 0.

## C.2 “Capture-Recapture” Method

Suppose that  $\bar{P}$  denotes the average proportions of pseudo-variables entering in last  $k_P$  steps, i.e.,

$$\bar{P}_P = \frac{1}{k_P} \sum_{i=k_T+1}^{k_T+k_P} P_i.$$

Recall that, after the first  $k_I$  steps, the probability of a pseudo-variable entering should be  $\frac{k_P}{k_U+k_P}$ . Hence, it is natural to expect that

$$\begin{aligned} \bar{P}_P &\approx \frac{k_P}{k_U + k_P} \\ &= \frac{k_P}{k_P + k_T - k_I}. \end{aligned}$$

Solving the above equation yields an estimator of  $k_I$ , which is

$$\widehat{k}_I = k_T + k_P - \frac{k_P}{\bar{P}_P}.$$

This process is very similar to what are called Capture-Recapture methods.

Such “Capture-Recapture” estimates can be further modified in two stages. First, we estimate  $k_I$  based on student’s  $t$  test statistics or P-values and obtain  $\widehat{k}_I^{(0)}$ . Then, we take the average of  $P_i$ ’s in the last  $k_T + k_P - \widehat{k}_I^{(0)}$  steps, i.e.,

$$\bar{P}_P^* = \frac{1}{k_T + k_P - \widehat{k}_I^{(0)}} \sum_{i=\widehat{k}_I^{(0)}+1}^{k_T+k_P} P_i.$$

By the same argument as before, we obtain an estimate of  $k_I$  equal to

$$\widehat{k}_I = k_T + k_P - \frac{k_P}{\bar{P}_P^*},$$

### C.3 Least Squares Method

Basically, we postulate a model for the mean of  $P_i$  in terms of  $k_I$  and one other parameter  $k$ , i.e.,

$$E(P_i) = \begin{cases} 0, & 1 \leq i \leq k_I - k; \\ \frac{1}{2k} \left( \frac{k_P}{k_P + k_T - k_I} \right) (i - k_I + k), & k_I - k < i < k_I + k; \\ \frac{k_P}{k_P + k_T - k_I}, & k_I + k \leq i \leq k_T + k_P, \end{cases}$$

where  $k$  indicates a non-negative integer. Typically, we limit  $k$  to be no more than 3. Define the least squares criterion between the  $P_i$  and the mean model to be  $SSE$ , which has the form of

$$SSE = \sum_{i=1}^{k_T + k_P} (P_i - E(P_i))^2$$

Searching through  $k \in \{0, 1, 2, 3\}$  and  $k_I \in \{0, \dots, k_T\}$  under the constraint  $k \leq k_I$ , we are able to find  $\hat{k}$  and  $\hat{k}_I$  such that  $SSE$  is minimized. Then,  $\hat{k}_I$  is used as an estimate of  $k_I$ .

### C.4 Isotonic Method

Another way to model the  $P_i$  is to use isotonic smoothing splines. We have the following rule for estimating  $k_I$  based on the isotonic estimates. Suppose that  $\hat{P}_i^{(iso)}$  denotes the  $i$ th isotonic estimate of  $P_i$ , and that  $\delta(i)$  denotes the difference between two successive isotonic estimates  $\hat{P}_i^{(iso)}$  and  $\hat{P}_{i+1}^{(iso)}$ , i.e.,  $\delta(i) = \hat{P}_{i+1}^{(iso)} - \hat{P}_i^{(iso)}$ ,  $i = 0, 1, \dots, k_T$ . We assume that  $\hat{P}_0^{(iso)} = 0$ . We search for the value of  $i$  that maximizes  $\delta(i)$  over the set  $\{0, 1, \dots, k_T\}$ , and then take it as an estimate of  $k_I$ .

# Bibliography

- Akaike (1973), “Maximum Likelihood Identification of Gaussian Autoregressive Moving Average Models,” *Biometrika*, 60, 255-265.
- (1977), “On Entropy Maximization Principle,” *Applications of Statistics* (Krishnaiah, P. R., ed.), North Holland: Amsterdam, pp. 27-41.
- Bendel, R. B. and Afifi, A. A. (1977), “Comparison of Stopping Rules in Forward “Step-wise” Regression,” *J. Amer. Statist. Assoc.*, 72, 46-53.
- Berger, J. O. and Pericchi, L. R. (1996a), “The Intrinsic Bayes Factor for Model Selection and Prediction,” *J. Amer. Statist. Assoc.*, 91, 109-122.
- (1996b), “The Intrinsic Bayes Factor for Linear Models,” *Bayesian Statistics 5*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, London: Oxford University Press, pp. 25-44.
- (2001), “Objective Bayesian Methods for Model Selection: Introduction and Comparison (with discussion),” *Institute of Mathematical Statistics Lecture Notes - Monograph Series* (Lahiri, P., ed.), vol. 38, 135-207.
- Breiman, L. (1992), “The Little Bootstrap and Other Methods for Dimensionality Selec-

- tion in Regression:  $X$ -fixed Prediction Error,” *J. Am. Stat. Assoc.*, 87, 738-754.
- Breiman, L. and Spector, P. (1992), “Submodel Selection and Evaluation in Regression: the  $X$ -random Case,” *Int. Statist. Review*, 60, 291-319.
- Broersen, P. M. T. (1986), “Subset Regression with Stepwise Directed Search,” *Appl. Statist.*, 35, 168-177.
- Chen, M.-H., Ibrahim, J.G., and Yiannoutsos, C. (1999), “Prior Elicitation, Variable Selection, and Bayesian Computation for Logistic Regression Models,” *J. Roy. Statist. Soc., B*, 61, 223-242.
- Chipman, H. George, E. I. and McCulloch, R. E. (2001), “The Practical Implementation of Bayesian Model Selection (with discussion),” *Institute of Mathematical Statistics Lecture Notes - Monograph Serires* (Lahiri, P., ed.), vol. 38, 65-134.
- Chatfield, C. (1995), “Model Uncertainty, Data Mining and Statistical Inference (with discussion),” *J. Roy. Statist. Soc., A*, 158, 419-466.
- Clyde, M., Parmigiani, G., and Vidakovic, B. (1998), “Multiple Shrinkage and Subset Selection in Wavelets,” *Biometrika*, 85, 391-402.
- Efron, B. (1982), *The Jackknife, the Bootstrap, and Other Resampling Plans*, Philadelphia:SIAM.
- (1983), “Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation,” *J. Amer. Statist. Assoc.*, 78, 316-331.
- Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004), “Least Angle Regression,” *The Annals of Statistics*, 32, No. 2, 407-499.



- Efroymson, M. A. (1960), "Multiple Regression Analysis," *Mathematical Methods for Digital Computers* (Ralston, A. and Wilf, H. S., ed.), vol. 1, Wiley: New York, pp. 191-203.
- Furnival, G. M. (1971), "All Possible Regressions with Less Computation," *Technometrics*, 13, 403-408.
- Furnival, G. M. and Wilson, R. W. (1974), "Regression by Leaps and Bounds," *Technometrics*, 16, 499-511.
- Garside, M. J. (1965), "The Best Subset in Multiple Regression Analysis," *Appl. Statist.*, 14, 196-200.
- (1971a), "Some Computational Procedures for the Best Subset Problem," *Appl. Statist.*, 20, 8-15.
- (1971b), "Algorithm AS38: Best Subset Search," *Appl. Statist.*, 20, 112-115.
- George, E. I. and McCulloch, R. E. (1993), "Variable Selection via Gibbs Sampling," *J. Amer. Statist. Assoc.*, 88, 881-889.
- Gilmour, S. G. (1996), "The Interpretation of Mallows's  $C_p$ -statistic," *The Statistician*, 45, 49-56.
- Haughton, D. (1988), "On the Choice of a Model to Fit Data From an Exponential Family," *Ann. Statist.*, 16, 342-355.
- Hoerl, A. E. and Kennard, R. W. (1970a), "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, 12, 55-67.
- (1970b), "Ridge Regression: Applications to Nonorthogonal Problems," *Techno-*

*metrics*, 12, 69-82.

- Hosmer, D. W. and Lemeshow, S. (2000), *Applied Logistic Regression*, John Wiley & Sons, Inc.
- Hurvich, C. M. and Tsai, C-L (1989), "Regression and Time Series Model Selection in Small Samples," *Biometrika*, 76, 297-307.
- Lane, L. J. and Dietrich, D. L. (1976), "Bias of Selected Coefficients in Stepwise regression," *Proc. Statist. Comput. Section, Amer. Statist. Assoc.*, pp. 196-200.
- Laud, P. W. and Ibrahim, J. G. (1995), "Predictive Model Selection," *J. Roy. Stat. Soc., Series B*, 57, 247-262.
- Lindley, D. V. (1968), "The Choice of Variables in Multiple Regression (with discussion)," *J. Roy. Statist. Soc., Series B*, 30, 31-66.
- Luo, X., Boos, D. D. and Stefanski, L. A. (2003), "Tuning Variable Selection Procedure," (upcoming)
- Mallows, C. L. (1973), "Some Comments on  $C_p$ ," *Technometrics*, 15, 661-675.
- (1995), "More Comments on  $C_p$ ," *Technometrics*, 37, 362-372.
- Mitchell, T. J. and Beauchamp, J. J. (1988), "Bayesian Variable Selection in Linear Regression (with discussion)," *J. Amer. Statist. Assoc.*, 83, 1023-1036.
- Miller, A. J. (2002), *Subset Selection in Regression*, London: Chapman & Hall.
- Morgan, J. A. and Tatar, J. F. (1972), "Calculation of the Residual Sum of Squares for All Possible Regressions," *Technometrics*, 14, 317-325.
- Rao, C. R. and Wu, Y. (2001), "On Model Selection (with discussion)," *Institute of*

- Mathematical Statistics Lecture Notes - Monograph Series* (Lahiri, P., ed.), vol. 38, 1-64.
- Roecker, E. B. (1991), "Prediction Error and Its Estimation for Subset-selected Models," *Technometrics*, 33, 459-468.
- Ronchetti, E. and Staudte, R. G. (1994), "A Robust Version of Mallows's  $C_p$ ," *J. Amer. Statist. Assoc.*, 89, 550-559.
- Ronchetti, E., Field, C. and Blanchard, W. (1997), "Robust Linear Model Selection by Cross-validation," *J. Amer. Statist. Assoc.*, 92, 1017-1023.
- Schwarz, G. (1978), "Estimating the Dimension of a Model," *Ann. Statist.*, 6, 461-464.
- Shao, J. (1993), "Linear Model Selection by Cross-Validation," *J. Amer. Statist. Assoc.*, 88, 486-494.
- Shao, J. (1996), "Bootstrap Model Selection," *J. Amer. Statist. Assoc.*, 91, 655-665.
- Shibata, R. (1981), "An Optimal Selection of Regression Variables," *Biometrika*, 68, 45-54.
- Stone, M. (1977), "An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion," *J. Roy. Statist. Soc., Series B*, 39, 44-47.
- (1979), "Comments on Model Selection Criteria of Akaike and Schwarz," *J. Roy. Statist. Soc., Series B*, 41, 276-278.
- Thompson, M. L. (1978), "Selection of Variables in Multiple Regression," *Internat. Statist. Rev.*, 46, 1-19 and 129-146.
- Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *J. Roy.*

*Statist. Soc., B*, 58, 267-288.

Tibshirani, R. and Knight, K. (1999), “The Covariance Inflation Criterion for Adaptive Model Selection,” *J. R. Statist. Soc., B*, 61, 529-546.

Weber, K., R. Eisman, S. Higgins, L. Morey, A. Patty, M. Tausek and Z.-B. Zeng (2001), “An Analysis of Polygenes Affecting Wing Shape on Chromosome 2 in *Drosophila Melanogaster*,” *Genetics*, 159, 1045-1057.

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