Topics in Modern Regression

(Thesis format: Integrated-Article)

by

Muslim-Mohammad Nagham

Graduate Program
in
Statistics

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Statistics

Faculty of Graduate Studies
The University of Western Ontario
London, Ontario, Canada

ABSTRACT

This thesis provides an overview of the state-of-the art in three current topics in regression analysis. These topics are:

- model selection methods in regression,
- regularization path for logistic regression by l_1- penalty or LASSO,
- regression discontinuity analysis.

Model selection is the task of choosing a best model for the given **finite** data. I introduce some of the methods for model selection in Chapter One such as subset selection, best subset selection by leaps and bound algorithm, and bootstrap model selection. Then I give an overview of the shrinkage methods and least angle regression algorithm.

Logistic regression is widely used as the method of analysis in a situation where the outcome variable is discrete; binary or dichotomous. I introduce logistic regression in Chapter Two. In addition, I give an over view for the Quadratic Lower Bound Algorithm QLB Tian et al. (2008) an efficient methods for estimating constrained parameters with application to lasso logistic regression. Using the proposed QLB algorithm I implemented the R-code for this algorithm and the R-code for the Pseudo-Newton method, which is faster than the fastest QLB algorithm.

Finally, I present the regression discontinuity (RD) designs in Chapter Three with its statistical analysis and the model specification problem.

KEY WORDS: LARS -algorithm, QLB-algorithm, Branch-and-Bound algorithm, RD-design.

ACKNOWLEDGEMENTS

First and foremost I would like to express my sincere gratitude to my supervisor Professor A. Ian McLeod for his invaluable guidance and generous support throughout the course of my study at Western. I would also like to thank my thesis examiners, Professors ..., ..., ... and ....for carefully reading this thesis and helpful comments. I am also grateful to all faculty, staff and fellow students at the Department of Statistical and Actuarial Sciences for their encouragement. The author would also like to acknowledge the financial support of the Department of Statistical and Actuarial Sciences and the Faculty of Graduate Studies and the Board of the Ontario Graduate Scholarships in Science and Technology. Finally, I would like to thank my family for their patience and love that helped me to reach this point.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Model selection is a difficult and important problem that is an important goal for many types of statistical modelings. The difficulty of this problem is increased in actual applications a true model may not exist. Given a data set, you can fit thousands of models, but how do you choose the best? This thesis introduces some of the model selection methods, especially in Linear model and extend it to some of the methods for logistic regression.

Chapter One, gives an application for subset selection methods. We start with Stepwise Procedures such as forward, backward, hybrid, and forward stagewise. These methods typically produce a model that is interpretable but has high variance so, they do not reduce the prediction error of the full model. Then we give an over view to All Subsets Selection where the search for the best subsets with minimum RSS or any other criteria can be done by computing all possible regressions but the amount of computation required can be formidable. Also, we introduce some subset selection criteria that used in model selection such as Mallow's C_p (Mallows, 1973), Akaike Information Criterion AIC (Akaike, 1974), Bayesian Information Criterion BIC (Schwarz, 1978), Extended Bayesian Information Criterion BIC_γ (Chen and Chen, 2008), Another Extended Bayesian Information Criterion BIC_q (Xu and McLeod, 2009). Then we introduce Tournament Screening TS (Chen and Chen, 2009)when a small sample size, n, and extremely high-dimensional features or covariates space, p, are used. Moreover, we give an application for subset selection by Cross Validation such as

Delete-d, K-fold, and LOOCV. Furthermore, we present the best subset by leaps and bound algorithm (Furnival and Wilson, 1974). The Leaps-and-Bounds strategy derives the best models for each number of variables with out examining all possible subsets. In addition, we present BestReg package in R (McLeod and Xu, 2009). BestReg package utilizes the regsubset function in leaps package to find the models with smallest sum of squares for size $k = 0, 1, ..., p$. We introduce a modified bootstrap variable/model selection procedure by (Shao, 1996), in linear models, extended to more complicated problems such as the nonlinear models, generalized linear models, and autoregressive time series.

Chapter Two, presents the idea of shrinkage methods where coefficients are shrinkage toward zero or exactly zero (Hastie et al., 2009). Ridge regression minimizes the residual sum of squares together with the penalty term. LASSO penalizes by absolute norm of the coefficients. The LASSO is a constrained version of OLS or Ordinary least squares. Finally, we give an overview on the least angle regression (Hastie et al., 2009), which is a new model selection algorithm related to forward selection.

In Chapter Three, we introduce the logistic regression which is widely used as the method of analysis in situation where the response variable is discrete binary or dichotomous (Hosmer et al., 2000). We present the QLB-algorithm and Pseudo-Newton algorithm (Tian et al., 2008) for regularized logistic regression and we implemented the R-code for these algorithms.

In Chapter Four, we introduce Regression Discontinuity Analysis and its assumptions to build an appropriate model which received a lot of attention recently (Lawrence, 1995; William and Trochim, 2006). The main concern in regression discontinuity analysis is model specification. When you are misspecified the statistical model your estimates of the treatment effect is likely to be biased. However, in typical regression our main concerned is about the variables that we should include in the

model (William and Trochim, 1984).

Chapter 2

Subset Selection

2.1 Introduction

In regression problems, a vector, X, of p-explanatory variables possibly related to a response variable, Y. These explanatory variables may not contribute equally well or may not contribute at all to the fitted model. In this case, we are looking to select the model with the least variables and still the best model. The goal of any model-building technique is to find the best fit that describes the relationship between the outcome or response variable, and the explanatory variables or predictors/covariates. The goal in using model selection methods is to find an accurate, parsimonious, interpretable, and stable model. In this chapter we introduce a number of methods for subset selection which are used to produce the best model by retaining only a subset of variables and eliminating the rest (Hastie et al., 2009; Furnival, 1971). Because of this discrete process, these methods usually have a high variance, and they do not reduce the prediction error for the full model but, they provide interpretable models. There are several methods that can be used (Hocking and Leslie, 1967). This chapter consists of five sections.

In Section (2.2), we start with stepwise selection methods, which compare models that have the same number of parameters, and they stop the process when adding or excluding any of the variables does not improve the fits. Forward Selection starts with the intercept only, and, step by step, builds up the model by including the variable that improves the fit the most. Backward elimination starts with the full model, and,

step by step removes the variable that improves the fit the least. Hybrid stepwise selection methods use both forward and backward moves. Forward stagewise is an attractive version of forward selection.

In Section 2.3 we give an over view to All Subsets Selection. The search for the best subsets with minimum RSS or any other criteria can be done by computing all possible regressions but the amount of computation required can be formidable.

Moving on, we introduce methods that works without stopping rule but, using some specific criterion such as Mallow's C_p , Akaike Information Criterion AIC (Akaike, 1974), and Bayesian Information Criterion BIC (Schwarz, 1978), Extended Bayesian Information Criterion BIC_γ (Chen and Chen, 2008), Another Extended Bayesian Information Criterion BIC_q (Xu and McLeod, 2009).

In Section 2.4.6, we introduce a tournament screening approach for subset selection when a small sample size, n, and extremely high-dimensional features or covariates space, p, are used. But only a small number of these covariates are related to the response variable. The TS reduces the dimensionality of original feature space $p \gg n$ to $K < n$. Then any model selection method can be used to select the final mode.

In Section (2.5), we introduce cross-validation approaches to model selection such as Delete-d, K-fold, and LOOCV which have been widely used.

In Section (2.6), we present the best-subset selection by Leaps and Bounds Algorithm, that derives the best models for each number of variables with out examining all possible subsets (Furnival and Wilson, 1974). It uses the a branch and bound strategy.

In Section (2.7), we present BestReg package in R (McLeod and Xu, 2009). BestReg package utilizes the regsubset function in leaps package to find the models with smallest sum of squares for size $k = 0, 1, ..., p$.

In Section (2.8), we introduce a modified bootstrap variable/model selection procedure that select a subset from p explanatory variables, x, possibly related to a response variable, y, by minimizing bootstrap estimates of the prediction error (Shao, 1996).

2.2 Stepwise Procedures

Stepwise procedures compare models that have the same number of parameters and they stop the process when including or dropping any of the variables does not improve the fit substantially.

2.2.1 Forward Selection or Forward Stepwise Selection

Forward selection starts with the intercept and step by step build up the model by including the variables that improves the fit the most. The algorithm is described for multiple linear regression; however, stepwise model selection may be used with many other types of statistical models. The algorithm is defined as follow:

- Starts with the intercept and select the variable that improves the fit the most. This may be done using the variable that results in the most significant F test.
- Calculate partial F values for all combinations of this variable with the remaining variables in order to find the best pair of variables.
- Add the variable with the most significant F-test.
- The algorithm stops when the most significant F-statistic does not exceed a pre-defined limit. In R, this pre-defined limit is 4.

More precisely, assume at stage m the variables $(X_j : j \in J_m)$ are already chosen. Let SSE_m be the residual sum of squares from fitting the model

$$Y = \sum_{j \in J_m} \theta_j X_j + \text{error}.$$

For each variable not in the model, X_k, $k \notin J_m$, let $\text{SSE}_m(k)$ be the residual sum of squares from fitting the model

$$Y = \sum_{j \in J_m} \theta_j X_j + \theta_k X_k + \text{error},$$

and define,

$$F_m(k) = (\text{SSE}_m - \text{SSE}_m(k))/(\text{SSE}_m(k)/(n-k-2)). \tag{2.1}$$

If there is $k \notin J_m$ such that $F_m(k)$ >pre-defined limit then include the variable X_k that maximizes $F_m(k)$ otherwise, stop the process. Note that forward selection can be overly aggressive in selection in the respect that if X_j is already included in a model, forward selection primarily adds variables orthogonal to X_j, thus ignoring possibly useful variables that are correlated with X_j.

2.2.2 Backward Elimination or Pruning of Variables

Backward stepwise selection starts with the full model, and, step by step removes the variable that improves the fit the least. For the linear regression model, backward elimination uses the F-ratio to delete variables. Step by step, we delete the variable that gives the smallest value of F, if it falls below a pre-defined limit. stopping when all explanatory variables give value of F-ratio grater than pre-defined limit. Backward selection works only when $n > p$, while Forward selection may be useful when $p \geq n$ provided numerical problems do not arise due to multicollinearity. Note that Backward elimination may be very computationally expensive if there are many

variables. But Backward Elimination is preferable to for Forward Selection if $p < n$ because it will often select a better model (Hastie et al., 2009).

2.2.3 Hybrid Stepwise Selection

Hybrid stepwise selection uses both forward and backward moves; at each step, we add or delete one variable depending on pre-defined rule to determine when we should use an add or drop move. The algorithm is defined as follow:

- Use the independent variable with the highest correlation with the response variable as the starting variable.

- Calculate partial F values for all combinations of this variable with the remaining variables. Add the one with the highest F values.

- Calculate partial F values for all variable in the current model and remove any variable which falls below a pre-defined limit.

- Repeat the process until the most significant F-statistic does not exceed a predefined limit.

2.2.4 Forward Stagewise (FS)

Forward stagewise is a version of forward selection, but behaves less greedy by producing less nested sequence of models, and it requirers many small steps before getting to the final model. This algorithm produce more accurate prediction than forward

selection, which may eliminate a useful predictor by using large steps (Hastie et al., 2009). The algorithm works as follow:

- Starts only with the intercept equal to $\bar{y}$ in the model, and all the coefficients are equal to zero, $\theta_1 = \theta_2 = ... = \theta_P = 0$.

- Starts with residual equal to $r = y - \bar{y}$.

- Selects the first predictors as forward selections, such as, x_j the one with the highest absolute current correlation c_j, with r. If $\hat{\theta}$ is the current stagewise estimate, let c_j be the vector of current correlations.

$$c_j = \text{Corr}\,(r, x_j),$$

$$c_j = \acute{x}_j(r - \hat{\theta}),$$

  leaving the residual vector as a response variable.

- In the direction of the greatest current correlation takes small step $0 < \epsilon < |c_j|$ to select x_j the one with the largest current correlation with ϵ some small constant.

$$\hat{j} = \arg\max |c_j|.$$

- Update $\hat{\theta}$ by

$$\hat{\theta} \longleftarrow \hat{\theta} + \epsilon.\text{sign}\,(c_{\hat{j}})x_{\hat{j}}.$$

- Update r

$$r \longleftarrow r - \epsilon.\text{sign}\,(c_{\hat{j}})x_{\hat{j}}.$$

- Repeat the process until no correlation is found between the update r and the predictors.

2.3 All Subsets Selection

The search for the best subsets with minimum RSS or any other criteria can be done by computing all possible regressions (Morgan and Tatar, 1972) but, the amount of computation required can be formidable. There are 2^p possible subsets therefore, the cost in computer time will be impractical specially when $p \geq 50$ since all possible models are considered. For $p \sim 10$ or perhaps a little larger a direct enumeration of all 2^p subsets is quite feasible. Letting $i = 0, 1, ..., 2^p - 1$ represent the index for the model. The inputs to include in the i-th model are determined by the base 2 expansion of i. For example, let $p = 8$ and $i = 217$, we see that this corresponds to the model using $1, 2, 4, 5$, and 8.

$$i = \text{IntegerDigits } [217, 2, 8]$$
$$\{1, 1, 0, 1, 1, 0, 0, 1\}$$

2.4 Subset Selection Criteria

2.4.1 $C_p$method (Mallows, 1973)

The model with the lowest C_p value approximately equal to p is the best model.

$$C_p = \frac{\text{SSE}}{\text{MSE}} - n + 2p, \tag{2.2}$$

where SSE is the residual sum of squares for the model with $p-1$ variables, MSE is the residual means squares for the full model, n denotes number of observations, p denotes number of variables in the model +1.

2.4.2 Akaike Information Criterion AIC (Akaike, 1974)

The model with lowest AIC is the best model.

$$\text{AIC} = -2\log L + 2p, \tag{2.3}$$

where L : Likelihood MLE, p : number of parameters, note that the error is normal $(i.i.d)$. For a linear regression model, the AIC may be equivalently written,

$$\text{AIC} = \text{SSE} + 2p. \tag{2.4}$$

where n denotes the number of observations.
Note that AIC is asymptotically equivalent to C_p (Lahiri, 1999).

2.4.3 Bayesian Information Criterion BIC (Schwarz, 1978)

The model with lowest BIC is the best model.

$$\text{BIC} = -2\log L + \log(n)\ p, \tag{2.5}$$

where L is Likelihood MLE, p denotes number of parameters, n denotes number of observations. For a linear regression model,

$$\text{BIC} = \text{SSE} + p\log(n). \tag{2.6}$$

2.4.4 Extended Bayesian Information Criterion BIC_γ (Chen and Chen, 2008)

When we have a moderate sample size but a huge number of covariates p, the BIC will select models with too many parameters (Chen and Chen, 2008), suggested a prior

uniform of models with fixed size instead of a prior uniform of all possible models. An extended information criterion score can be written,

$$\mathrm{BIC}_\gamma = -2\log L + k\log(n) + 2\ \gamma\ log\ \binom{k}{p}, \tag{2.7}$$

where L is Likelihood MLE, $\log L = -(n/2)\log(\mathrm{SSE}/n)$, k denotes number of parameters in the model, n denotes number of observations, p is the number of possible input variables not counting the bias or intercept term. Notice that, $k = 0$ corresponding to only an intercept term, $k = p$ corresponding to using all parameters in the model. The BIC_γ are shown to be consistent and particularly useful even when the covariates are heavily collinear.

Note that, when $p \gg n$, we can not calculate BIC_γ directly. Chen and Chen (2009) proposed tournament screening approach that can be used to calculate BIC_γ. In the tournament screening approach Section 2.4.6, the penalized likelihood technique (Tibshirani, 1996) is used to order models and BIC_γ is then used to make the final selection.

2.4.5 Another Extended Bayesian Information Criterion BIC_q (Xu and McLeod, 2009)

Xu and McLeod (2009) suggested a Bernoulli prior for the parameters with a probability $q \in (0, 1)$. The extended information criterion can be written,

$$\mathrm{BIC}_q = -2\log L + k\log(n) + 2\ \gamma\ \log\ q\ /(1-p), \tag{2.8}$$

when $q = 1/2$, the BIC_q is equivalent to the BIC (Xu and McLeod, 2009). The full model with all covariates is selected when $q = 1$. No parameters are selected when $q = 0$ because the penalty is taken to be $-\infty$.

2.4.6 Tournament Screening TS (Chen and Chen, 2009)

Chen and Chen (2009) proposed a new approach for subset selection when a small sample size, n, and extremely high-dimensional features or covariates space, p, are used. But only a small number of these covariates are related to the response variable such as, a microarry tumor studies. The TS reduces the dimensionality of original feature space $p \gg n$ to $K < n$. Then any model selection methods can be used in the final selection. Chen and Chen (2009) preferred to work with the penalized likelihood technique (Tibshirani, 1996) in the final selection.

The basic idea of the tournament screening is that the set of all covariates or features is randomly partitioned into non-overlapping groups of approximately equal size. Then, applied a penalized likelihood technique (Tibshirani, 1996) to each group of covariates. A given number, the non-zero components, of the fitted are selected and then pooled together. The dimensionality of the feature space is further reduced. Repeat the process until you get the desirable dimension of the feature space.

2.5 Subset Selection by Cross Validation

The simplest method for estimating predication error is cross-validation (Hastie et al., 2009). For the model $y = \sum_{j \in J} \theta_j x_j$, the predication error is defined as

$$\text{PE} = \text{E}\ (y^* - \sum_{j \in J} \hat{\theta}_j x_j^*)^2, \tag{2.9}$$

where $\hat{\theta}$ is given by the fit and the expectation is with respect to a new observation (x^*, y^*). Since we do not know the joint distribution of (x^*, y^*), we need to estimate this quantity. If we had additional data $(x_i^*, y_i^*) : i = 1...m$, not used in fitting the model, then the estimate of the prediction error is

$$\mathrm{PE} = \mathrm{E}(y_i^* - \sum_{j \in J} \hat{\theta}_j x_{i,j}^*)^2. \tag{2.10}$$

The data used to fit the model is called training data and the data used to estimate the predication error is called test or validation data. Divided a Given a single dataset to build a model into training and testing sets. Cross-Validation approaches to model selection have been widely used.

2.5.1 Delete-d Cross-Validation

Shao (1993) proposed a delete-d method for model selection in linear regression which is consistent in the sense that the probability of selecting the model with the best predictive ability does converge to 1 as $n \longrightarrow \infty$ and d increase with n . As the validation set a random sample of size d are used and in this way, many validation sets are generated. As the training set the remaining part of the data is used. If enough validation sets are used LOOCV Section (2.5.3) will give the same result as the delete -d when d = 1. Shao (1997) recommends a much larger cross-validation sample than is used in K-folder CV. Shao (1997), suggested "$\lambda_n = \log n$", and

$$\mathrm{d} = n(1 - (\log n - 1)^{-1}), \tag{2.11}$$

where n is the number of observations.

2.5.2 K-fold Cross-Validation

In K-fold cross-validation, (Hastie et al., 2009), a sample of data is randomly partitioned into subsamples approximately equal size. A single subsample is retained as the validation data for testing the model. The remaining $K - 1$ subsamples are used as training data. The process is repeated K times. The results for the K folds can

be averaged or combined to produce a single estimation. With K-fold CV, (Hastie et al., 2009), suggest using "one-standard deviation". For each of the K validation samples, the validation sum-of-squares is computed

$$S_k = \sum_{i \in \prod_k} \lim(\hat{e}^{(-k)i})^2, \tag{2.12}$$

where $\hat{e}^{(-k)i}$ denotes the prediction error when the kth validation sample is removed. Fit the model for the remainder of the data and predict the observation $i \in \prod_k$ in the validation sample. The final cross-validation score is

$$\text{CV} = \frac{1}{n} \sum_{k=1}^{K} \lim S_k, \tag{2.13}$$

where n is the number of observations. The cross-validation mean-square error, $\text{CV}_k = S_k/N_k$, where N_k denotes the number of observations in the kth validation sample can be calculated for each validation sample. For $\text{CV}_1, ..., \text{CV}_K$ let s be the sample variance. Then using the one-standard-deviation rule, $\text{CV} \pm 0.5s$, is the interval estimate of CV. For model selection, this suggests that the most parsimonious adequate model will correspond to the model with the largest CV which still inside this interval.

2.5.3 Leave-One-Out Cross-Validation

In Leave-One-Out CV, remove one observation, say the ith, and the fit is computed using all the data except the ith. The prediction error, $\hat{e}_{(i)}$ for the missing observation is calculated. The cross-validation process is then repeated for all observations $i = 1, ...n$ and the prediction error sum of squares is calculated,

$$\text{PRESS} = \sum_{i=1}^{n} \hat{e}_{(i)}^2 \tag{2.14}$$

DM (1971), proposed PRESS method, which can be used very efficiently to compute LOOCV in linear regression contest, $\hat{e}_{(i)} = \hat{e}_i$, where $\hat{e}_i$ is the usual regression residual and $h_{i,i}$ is the ith element of the diagonal of the hat matrix $H = X(\acute{X}X)^{-1}\acute{X}$. LOOCV is asymptotically equivalent to the AIC (Stone, 1977).

2.6 Leaps and Bounds Algorithm

Leaps-and-Bounds is a widely used algorithm for best subset selection. The Leaps-and-Bounds strategy derives the best models for each number of variables with out examining all possible subsets. It uses a Branch and Bound strategy. The method works by generating two trees (Furnival and Wilson, 1974). The first one, the bounds tree, provides half of the models that include the last variable. The bound tree is obtained by eliminating all pivots on the last variable from the original inverse tree. A Gaussian elimination step is used to obtain a child from parent node. The second tree, is called the regression tree, provides the other half of the models that do not includes the last variable and this need not be an inverse tree. By moving from one node to another the models are generated and at each time a model is obtained by an updated matrix inverse computation. The Leaps-and-Bounds algorithm works out the branches of the full inverse tree by simultaneously traverses of the regression tree and the bound tree (Huo and Ni, 2006). In other words, the Leaps-and-Bounds algorithm scans the full inverses tree through all the subsets, simultaneously, leaping over those that are not the optimal subsets. A dot notation is used for labeling the nodes. The integers listed before the dot represent the explanatory variables present in the sub-matrix for which the pivots have not been performed; the subscripts after the dot represents the variables on which pivots have been performed. Deletions are indicated by missing subscripts, rows and columns associated with those variables have been

deleted. For example, the submatrix 4.2 has been obtained from the original matrix $(X(1), X(2), X(3), X(4))$ by deleting $X(1), X(3)$ and pivoting on $X(2)$.

2.6.1 The Regression Tree

The sequences of pivots of the Leaps-and-Bounds algorithm are derived from the binary tree of Figure (2.1). The tree is constructed by starting at the root with the original matrix and **splitting** the matrix into two new sub-matrices on the interior nodes. The first one is obtained by pivoting on the first variable. The second one is obtained by deleting the row and column associated with that variable. The process is repeated until all variables have been treated either by deletion or by pivoting. Note that, each leaf or terminal node is one of the 2^p possible regressions including the null regression where p is the number of covariates. For example, in Figure (2.2) we have 4-variables, so we have in result 16 possible regressions including the null regression. There are several algorithms for computing the residual sums of squares for all possible regressions with minimum arithmetic (Furnival and Wilson, 1974) their recursive applications define trees structure, and two of these algorithms can be combined to form the Leaps-and-Bounds technique. The solution of Leaps-and-Bound requires two trees-one for bounds or inverse tree and one for regressions which is one of the four following trees. In the following trees deletions are implied and interior and terminal nodes represent regressions.

- The Natural and Lexicographic Tree.

  Search the tree of Figure (2.3) horizontally, level by level, from top to bottom this process produces the regressions in a natural order all one-variable regression, followed by all two and three variable regressions Table (2.1). Search the

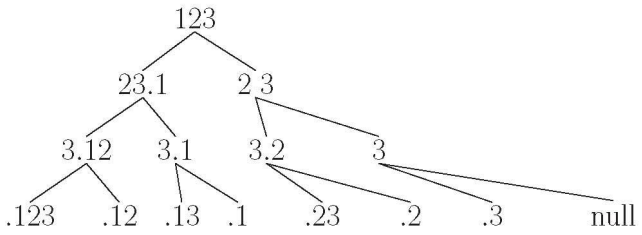

Figure 2.1: The Regression tree for 3-variable

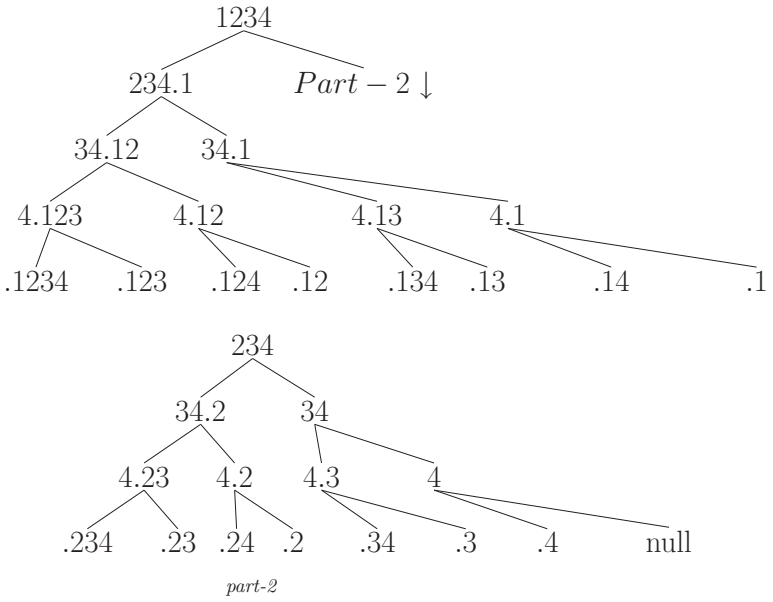

Figure 2.2: The Regression tree for 4-variable

Table 2.1: Sequences of regressions for 3-variables.

Natural	*Lexicog.*	*Binary*	*familial*
1	1	1	1
2	12	2	2
3	123	12	3
12	13	3	12
13	2	13	13
23	23	23	23
123	3	123	123

tree of Figure (2.4) vertically branch by branch beginning at the root and moving from father to older son at an interior node. At the terminal node, move to the next young brother, or if there is no brother, to the father's next young brother and so on. This process produces the regressions in a lexicographic order Table (2.2).

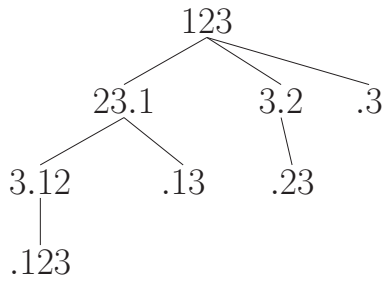

Figure 2.3: The Natural and Lexicographic tree for 3-variable

- The Binary and Familiar Tree.

  Search the tree of Figure (2.5) vertically, the regressions are produced in a binary order. Search the tree of Figure (2.5) both horizontally and vertically, the regressions are produced in a familial order. Move from father to older son, as

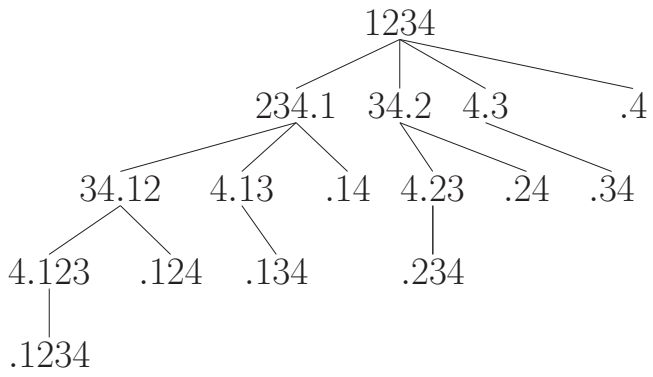

Figure 2.4: The Natural and Lexicographic tree for 4-variable

Table 2.2: Sequences of regressions for 4-variables.

Natural	*Lexicog.*	*Binary*	*familial*
1	1	1	1
2	12	2	2
3	123	12	3
4	1234	3	4
12	124	13	12
13	13	23	13
14	134	123	23
23	14	4	123
24	2	14	14
34	23	24	24
123	234	124	34
124	24	34	124
134	3	134	134
234	34	234	234
1234	4	1234	1234

described before but, when a node is visited, the sons of that node are listed in order of age from oldest to youngest Table (2.1). Search the tree vertically Figure (2.6), the regressions are produced in a binary order Table (2.2).

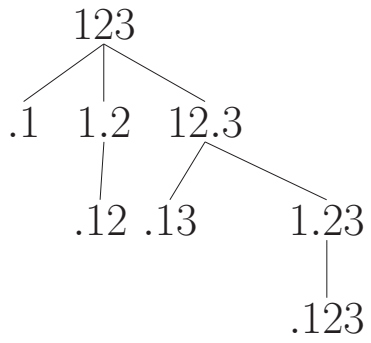

Figure 2.5: The Binary and Familial tree for 3-variable

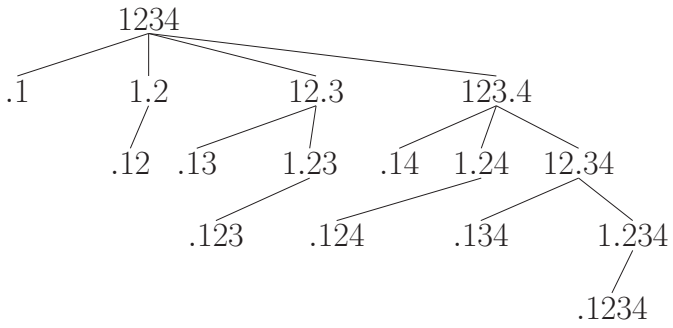

Figure 2.6: The Binary and Familial tree for 4-variable

2.6.2 The Inverse Tree

The inverse tree in Figure (2.7) and (2.8) can be constructed as follow:

- The root node is the set of all the covariates $\{1, 2, ..., p\}$.

- Level one is obtained by removing one covariate at a time from the set of all covariates (p) by a decreasing order $p, p-1, p-2, ..., 1$. At the end of this level we will have n ordered children.

- All the others levels are obtained by deleting one covariates from the subset associated with its parent node. For example, if we have the node associated with the subset $(i_1, i_2, ...i_k)$, where $i_1 < i_2 < ... < i_k$ and $i \geq 1$, of the jth child. The other levels are made by deleting one covariate at a time with order $(i_k, i_{k-1}, i_{k-2}, ..., i_{k+2-j})$.

- The growth of the tree is stopped when we obtained the subsets of one covariate.

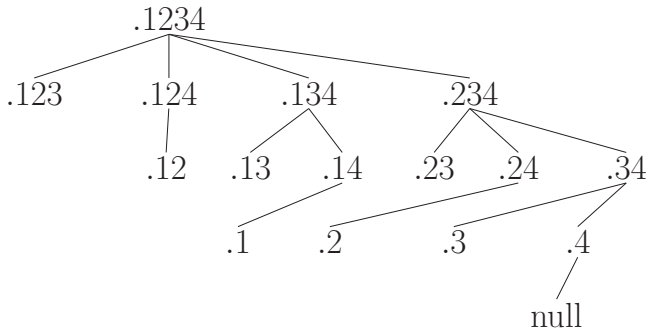

Figure 2.7: The Inverse tree for 4-variable

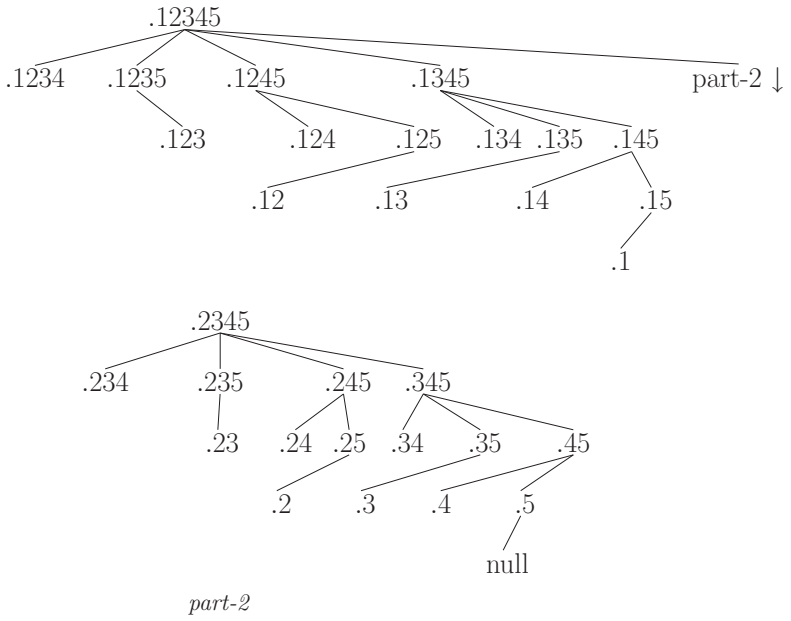

Figure 2.8: The Inverse tree for 5-variable

2.6.3 Branch and Bound Algorithm

Land and Doig (1960) first proposed Branch and Bound method for linear programming. Branch and Bound is a general algorithm for optimization problems. It requires the following steps:

- The fist step is called branching.

  A splitting procedure, that given a set of candidates K, returns smaller sets $K_1, K_2, ...$ whose union are K. To find the minimum value of the function $f(x)$ over K is $\min\{k_1, k_2, ...\}$, where each k_J is the minimum of $f(x)$ within K_j. The subsets of K are the nodes of a search tree or regression tree.

- The second step is called bounding.

  Computes upper and lower bound for the minimum $\{k_1, k_2, ...\}$.

- The third step is called pruning.

  For some tree node A, if the lower bound is grater than the upper bound for some other node B, then discarde A from the search. Also, any node can be discarded if its lower bound is grater than a global variable S, that records the minimum upper bound that have been scanned up to this point. The procedure stops when the candidate set K become a single elements; or also when the upper bound for set K matches the Lower bound.

Branch and Bound algorithm computes the best subset regression models by searching the whole regression tree that generates all possible subset models. However, a number of authors have described procedures for finding the best subset regressions with out computing all possible regressions (Hocking, 2003). All these methods are based on $\text{RSS}(A) \leq \text{RSS}(B)$, where B is subset of A, and A is any set of independent variables ($B \subseteq A$). In other word, deleting variables from the regression will not reduce the residual sum of squares for the regression. Figure (2.9) shows the bound tree for 4-variables.

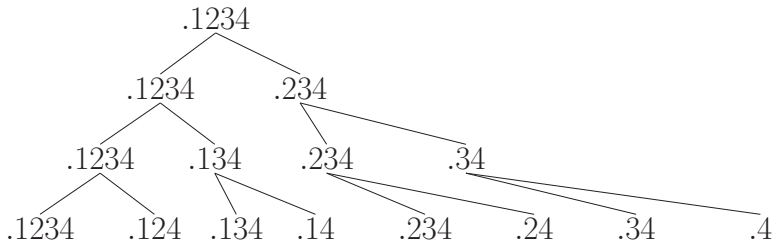

Figure 2.9: The Bound tree

2.6.4 Pair Tree

The Leaps-and-Bounds algorithm is based on two trees, Regression tree and Bound tree. The following pair trees construct the Regression and the Bound pairs for the Leaps-and-Bound subsets. For $p = 2$ the Pair tree is

$$(\emptyset, \{12\})$$
$$\downarrow$$
$$(\{1\}, \{2\})$$

For $p = 3$

$$(\emptyset, \{123\})$$
$$\swarrow \quad \searrow$$
$$(\{12\}, \{13\}) \qquad (\{1\}, \{23\})$$
$$\downarrow$$
$$(\{1\}, \{2\})$$

For $p = 4$

$$(\emptyset \ , \ \{1\ 2\ 3\ 4\})$$
$$\swarrow \qquad \downarrow \qquad \searrow$$
$$(\{123\}, \{124\}) \qquad (\{12\}, \{134\}) \qquad (\{1\}, \{234\})$$
$$\downarrow \qquad \swarrow \quad \searrow$$
$$(\{13\}, \{14\}) \qquad (\{23\}, \{24\}) \qquad (\{2\}, \{34\})$$
$$\downarrow$$
$$(\{3\}, \{4\})$$

2.6.5 Optimality Test

At each step in of the computation, Leaps-and-Bounds algorithm produce both bound and regression (RSS) by parallel pivots on sub-matrices from the product matrix and its inverse Table (2.3). Which is equivalent to scanning through the pair tree according to the following scheme:

- For the Pair tree, calculate the residual sum of squares for all subsets in the root node and the nodes in level 1.
- Let A, B be an intermediate node. Let $|.|$ be the size of subset $|A| \leq |B|$.
- Let RSS(m) denote the minimum residual sum of squares of all the m-subsets that have been scanned so far, where $|A| \leq m \leq |B|$.
- Skip all the descendants of node (A, B) when RSS($|A|$) $\leq$ RSS(B). Because all the descendant of node (A, B) will have residual sum of squares bigger than RSS(m).
- Skip the first m children of node (A, B) if RSS($|B|-m$) $\leq$ RSS(B) $<$ RSS($|B|-m-1$) for specific m, where $(1 \leq m \leq (|B| - |A| - 1)$.
- If RSS(B) $<$ RSS($|B| - 1$), then do not skip any child from (A, B).

We discuss the procedure of Leaps-and-Bounds algorithm for $p = 5$ (Furnival and Wilson, 1974) as follow

- The sequence product of the traverse in Table (2.3) is the Lexicographic algorithm.
- The asterisk in the Table (2.3) indicates failure of the test to leap.
- When the test of leaping is successful, the increment for the stage counter is 2^{p-k-1} where k is the pivot index.
- You always must performed the pivots of stage zero.
- Note that, the inverse source always bounds the regressions that can be produced from it and from the product source.
- At stage 1, the inverse source is the submatrix .1245 and the product source is the submatrix 4.12. The RSS(612)of the submatrix .123 is larger than the RSS(596) of the submatrix 1245. The asterisk indicates that our test fails.
- At stage 2, the first pivot, the inverse source is the submatrix .1345. Again the inverse source always bounds the regressions that can be produced from it and from the product source. The RSS (615) of the submatrix .12 is larger than the RSS(605) of the submatrix .1345, our test fails.
- At Stage 2, the second pivot, the RSS (597) of the submatrix .125 is smaller than RSS(605) of the submatrix .1345 therefore, we can leap by skipping the last pivot. The stage increment is 2^{5-2-1} or 1.
- At stage 4, the fist pivot, the RSS(668) of the submatrix .1 is larger than the RSS(660) of the submatrix .2345, our test fails.
- At stage 4, the second pivot, the RSS(615) of the submatrix .12 is smaller than RSS(660) of the submatrix .2345, our test is successful therefore, the regression from the submatrix .2345 and 34.2 need not be produced.

Table 2.3: Order of computations for Leap and Bound algorithm

Stage Number	*Pivot Index*	*Product Source :*	*Regr.*	*Traverse RSS*	*Inverse Source :*	*Regr.*	*Traverse RSS*	*Stage Bound*
0	$*$ 1	1234.	234.1	668	.12345	.2345	660	
	$*$ 2	234.1	34.12	615	.12345	.1345	605	
	$*$ 3	34.12	4.123	612	.12345	.1245	596	
	$*$ 4	4.123	.1234	592	.12345	.1235	596	
1	$*$ 4	4.12	.124	615	.1245	.125	597	596
2	$*$ 3	34.1	4.13	641	.1345	.145	618	605
	4	4.13	.134	612	.1345	.135	618	
3	4	4.1	.14	648	.145	.15	618	618
4	$*$ 2	234.	34.2	702	.2345	.345	720	660
	3	34.2	4.23	673	.2345	.245	667	
	4	4.23	.234	664	.2345	.235	666	
5	4	4.2	.24	685	.245	.25	675	667
6	3	34.	4.3	746	.345	.45	736	720
	4	4.3	.34	727	.345	.35	732	
7	4	4.	.4	792	.45	.5	799	736

- At stage 6, the RSS(668) of the submatrix .1 is smaller than RSS(720) of the submatrix (.345), our test is successful- permits a leaps of two stages.
- The search is completed after evaluating 6 of the 22 possible subsets of stage $1-7$.

2.7 BestReg: Best Subset linear regression (McLeod and Xu, 2009)

Lumley and Miller (2004) implemented the package leaps in R to solve the Branch-and-Bound algorithm with other subset selection algorithms. When $p \leq 25$ the leaps

function, regsubset, can determined the best model of size k, $k = 1, ..., p$ in few seconds, while when p is large enough such as $p \geq 100$ the problem will be more complicated since the computer time grows exponentially with p, and in this case there are 2^{100} all possible regressions. (McLeod and Xu, 2009) implemented the package BestReg in R, that utilizes the regsubset function in leaps package to find the models with smallest sum of squares for size $k = 0, 1, ..., p$. Cross-Validation or information criterion such as AIC, BIC, BIC_q, and BIC_γ is used to find the best model. For Cross-Validation (McLeod and Xu, 2009), implemented the Delete- d, K-fold, and LOOCV in the BestReg package.

2.8 Bootstrap Model Selection

A Bootstrap Variable/Model Selection procedure is to select a subset from p explanatory variables (x) that possibly related to a response variable (y) by minimizing bootstrap estimates of the prediction error that is constructed based on a data set of size n (Efron, 1982) and (Efron, 1979). This bootstrap procedure is inconsistent because the probability of selecting the optimal subset of variables does not converge to 1 as $n \longrightarrow \infty$. (Shao, 1996), and (Dongsheng and Shao, 1995) proposed two consistent modification of bootstrap selection procedures in linear models, extended to more complicated problems such as the nonlinear models, generalized linear models, and autoregressive time series. First, for bootstrapping pairs (x, y), where $m < n$, he suggest to generate m pairs of bootstrap data. Second, for bootstrapping residual, he suggest multiplying the residuals by factor $\sqrt{(n/m)}$, where $(m/n) \longrightarrow 0$ as $m \longrightarrow \infty$. The bootstrap selection procedures are asymptotically equivalent to the selection procedures using information criterion, C_p.

2.8.1 Linear Models

Assume that $X = (x_1, x_2, ..., x_n)'$ is full rank and

$$\mu_i = E(y_i|x_i) = x_i'\theta, \qquad Var(y_i|x_i) = \sigma^2,$$

$$i = 1, ..., n, \tag{2.15}$$

where (x_i, y_i), $i = 1, 2, ..., n$ are i.i.d data set, x_i is the ith value of a p vector of the explanatory variable, y_i is the response variable at x_i, and θ is a p vector of unknown parameters.

A model α of the subset of P of size p_α, is

$$\mu_{i\alpha} = \mathrm{E}(y_i|x_i) = x_{i\alpha}'\theta_\alpha, \qquad \mathrm{Var}(y_i|x_i) = \sigma^2,$$

$$i = 1, ..., n, \tag{2.16}$$

where α is a subset of $\{1, 2, ..., p\}$ and $x_{i\alpha}$ (or θ_α) is a subvector of x_i (or θ). Use least squares method to fit the model α

$$\hat{\theta}_\alpha = (X_\alpha' X_\alpha)^{-1} X_\alpha' Y,$$

where $X_\alpha = (x_{1\alpha}, ..., x_{n\alpha})'$ and $Y = (y_1, ..., y_n)'$. The efficiency of model α can be measured by the average loss,

$$\begin{aligned} L_n(\alpha) &= \frac{1}{n}\sum_{i=1}^{n}(\mu_i - x_{i\alpha}'\hat{\theta}_\alpha)^2, \\ &= \frac{||\mu - \hat{\mu}_\alpha||^2}{n}, \end{aligned}$$

where $\mu = (\mu_1, ..., \mu_n)'$, $\hat{\mu}_\alpha = x_\alpha\hat{\theta}_\alpha$ and $||b|| = \sqrt{b'b}$ for any b vector. Select a model over all $\alpha \in A$ so that $L_n(\alpha)$ may be as small as possible, where A is a collection of some subsets of $\{1, ..p\}$ this is equivalent to selecting a model with the best prediction ability.

$$\Gamma_n(\alpha) = E[\frac{1}{n} \sum_{i=1}^{n}(w_i - \acute{x}_{i\alpha}\hat{\theta}_\alpha)^2|y,x], \tag{2.17}$$
$$= \sigma^2 + L_n(\alpha),$$

where $\Gamma_n(\alpha)$ is the average conditional expected loss in prediction, w_i is a future response at x_i, and assume that w_i are independent of the y_i. Let $\epsilon = y - \mu$, $H_\alpha = X_\alpha(\acute{X_\alpha}X_\alpha)^{-1}X_\alpha$ and

$$\Delta_n(\alpha) = \frac{||\mu - H_\alpha\mu||^2}{n}.$$

Then

$$L_n(\alpha) = \Delta_n(\alpha) - \frac{2(\mu - H_\alpha\mu)\acute{}\epsilon}{n} + \frac{||H_\alpha\epsilon||^2}{n}. \tag{2.18}$$

For given α, model α is called a correct model if θ_α contains all nonzero components of θ, then $\acute{x_i}\theta = \acute{x_{i\alpha}}\theta_\alpha$, for any x_i. If α is the correct model, then $\mu = X\beta = X_\alpha\theta_\alpha = H_\alpha\mu$, $\Delta_n(\alpha) = 0$, and

$$L_n(\alpha) = \frac{||H_\alpha\epsilon||^2}{n}. \tag{2.19}$$

Let α_0 be the subset corresponding to the correct model with smallest size, then

$$\liminf_{n\longrightarrow\infty} \quad \triangle(\alpha) > 0 \text{ for any incorrect model } \alpha \quad ,$$

and

$$\lim_{n\longrightarrow\infty} P\{L_n(\alpha_0) = \min_{\alpha\in A} L_n(\alpha)\} = 1$$

Model α_0 minimizes $L_n(\alpha)$ over $\alpha \in A$ as $n \longrightarrow \infty$, therefore , it is the optimal model. The model selection procedure is said to be consistent if

$$\lim_{n\longrightarrow\infty} P(\hat{\alpha} = \alpha_0) = 1,$$

where $\hat{\alpha}$ is the estimate of α based on model selection procedure. The optimal α_0 must be estimated because $L_n(\alpha)$ contains the unknown parameter θ.

2.8.2 Bootstrap Selection Procedures

There are two ways of generating bootstrap observations for linear model or bootstrap estimators of α_0:

- Bootstrapping Paris (Efron, 1982).

  Generate $\{(x_i^*, y_i^*), i = 1, ..., n\}$ i.i.d bootstrap data from the empirical distribution $\hat{F}$ putting mass n^{-1} on each pair (x_i, y_i). The bootstrap of $\hat{\theta_\alpha}$ is

$$\tilde{\theta_\alpha^*} = (\acute{X}_\alpha^* X_\alpha^*)^{-1} \acute{X}_\alpha^* Y^*, \tag{2.20}$$

  where $X^* = (x_1^*, ..., x_n^*)\acute{}$, $Y^* = (y_1^*, ..., y_n^*)\acute{}$, $\acute{X}X \longrightarrow \infty$, $\acute{X}^*X^* \longrightarrow \infty$ almost surly , and $\tilde{\theta_\alpha^*}$ can be replaced by $\hat{\theta_\alpha}$ in the situation where $(\acute{X}_\alpha^* X_\alpha^*)^{-1}$ does not exist.

- Bootstrapping Residuals (Efron, 1983).

  $\hat{\theta}$ is the least square estimate (LSE) under linear model. The ith residual is $r_i = y_i - \acute{x_i}\hat{\theta}$. Generate $\epsilon_1^*, ..., \epsilon_n^*$ i.i.d from the empirical distribution that puts mass n^{-1} on $(r_i - \bar{r})/\sqrt{(1-p)/n}$, $i = 1, 2, ..., n$, where $\bar{r}$ is the average of the r_i. The bootstrap analog of $\hat{\theta_\alpha}$ is

$$\hat{\theta}_\alpha^* = (\acute{X}_\alpha X_\alpha)^{-1} \acute{X}_\alpha y_\alpha^*, \tag{2.21}$$

  where $Y_\alpha^* = (y_{1\alpha}^*, ..., y_{n\alpha}^*)\acute{}$, $y_{i\alpha}^* = \acute{x}_{i\alpha}\hat{\theta}_\alpha + \epsilon_i^*$, and $\{(x_{i\alpha}, y_{i\alpha}^*), i = 1, ..., n\}$ are the bootstrap observations under model α. The bootstrap estimate of the mean of the prediction error $\Gamma_n(\alpha)$ is

$$E(\Gamma_n(\alpha) = E[\frac{||Y - X_\alpha\hat{\theta}_\alpha||^2}{n}] + e_n(\alpha),$$

  where

$$e_n(\alpha) = E[\Gamma_n(\alpha) - \frac{||Y - X_\alpha\hat{\theta}_\alpha||^2}{n}] \quad .$$

The bootstrap estimate of $E[\Gamma_n(\alpha)]$ is

$$\hat{\Gamma}_n(\alpha) = \frac{||Y - X_\alpha \hat{\theta}_\alpha||^2}{n} + \hat{e}_n(\alpha).$$

This estimate is almost unbiased, but this procedure is inconsistent unless $\alpha_0 = \{1, ..., p\}$ is only the correct model. The bootstrap estimate of $e_n(\alpha)$ is

$$\hat{e}_n(\alpha) = E_*[\frac{||Y - X_\alpha \hat{\theta}^*_\alpha||^2}{n} - \frac{||Y^* - X^*_\alpha \hat{\theta}^*_\alpha||^2}{n}],$$

where $\theta^*_\alpha = \hat{\theta}^*_\alpha$ or $\tilde{\theta}^*_\alpha$, and E_* is the expectation with respect to bootstrap sampling(pairs, residual). For bootstrapping residuals $X^*_\alpha = X_\alpha$ and $Y^* = Y^*_\alpha$.

2.8.3 Modified Bootstrap Selection Procedures for Linear Model

Assume that the largest subset of the explanatory variables is α_p. For any $\alpha \in A$

$$D_n(\alpha) = \mathrm{E}(\Gamma_n(\alpha) - \Gamma_n(\alpha_p))$$

$$D_n(\alpha) = E(L_n(\alpha) - L_n(\alpha_p)). \tag{2.22}$$

Only in the case where α_p is the only correct model, we can find a consistent estimator $\hat{D}_n(\alpha)$ of $D_n(\alpha)$ such that

$$\frac{\hat{D}_n(\alpha)}{D_n(\alpha)} \longrightarrow \quad 1 \quad \textit{in probability}, \quad \alpha \in A.$$

Let $\{m_n\}$ be a sequence of integers such that $\lim_{n\longrightarrow\infty} m_n = \infty$, $\lim_{n\longrightarrow\infty} \frac{m_n}{n} = 0$, and $\lim_{n\longrightarrow\infty} P[\hat{\alpha}_{n,m} = \alpha_0] = 1$. There is one restriction on m is that p/m should be reasonably small; we should choose an m so that the least squares fitting of regression model with p regressors does not have too high variability.

First, for bootstrapping pairs, $m < n$, bootstrap estimator of $E[\Gamma_m(\alpha)]$ is

$$\hat{\Gamma}_{n,m}(\alpha) = E_*\,[\frac{||Y - X_\alpha \tilde{\theta}^*_{\alpha,m}||^2}{n}], \tag{2.23}$$

where $\tilde{\theta}^*_{\alpha,m}$ is the bootstrap of $\hat{\theta}_\alpha$ based on m i.i.d pairs (x^*_i, y^*_i) generated from the empirical distribution putting mass n^{-1} on (x_i, y_i), $i = 1, 2, ..., n$; that is

$$\tilde{\theta}^*_{\alpha,m} = (\sum_{i=1}^{m} X^*_{i\alpha} \acute{X}^*_{i\alpha})^{-1} \sum_{i=1}^{m} X^*_{i\alpha} y^*_i, \tag{2.24}$$

we can modify the procedure in bootstrapping pairs by selecting the model $\alpha_{n,m} \in A$ that minimizes $\hat{\Gamma}_{n,m}(\alpha)$.

Second, for bootstrapping residuals, and for a special case where $(x_i = \frac{i}{n})$, $m < n$ and the fact that

$$E_* \, \tilde{\theta}^*_{\alpha,m} \approx E_* \, \tilde{\theta}^*_\alpha, \qquad Var_* \, \tilde{\theta}^*_{\alpha,m} \approx \tfrac{n}{m} Var_* \, \tilde{\theta}^*_\alpha,$$

where $\tilde{\theta}^*_{\alpha,m}$ and $\tilde{\theta}^*_\alpha$ are defined in equation (2.24) and equation (2.20), respectively. A modified bootstrap model selection procedure in bootstrapping residuals can be done by multiplying a factor $\sqrt{n/m}$ to the values from which the bootstrap data are generated. The estimate of $E[\Gamma_m(\alpha)]$ is

$$\hat{\Gamma}_{n,m}(\alpha) = E_* \, \frac{||Y - X_\alpha \hat{\theta}^*_{\alpha,m}||^2}{n},$$

where

$$\hat{\theta}^*_{\alpha,m} = (\acute{X}_\alpha X_\alpha)^{-1} \sum_{i=1}^{n} x_{i\alpha} y^*_{i\alpha},$$

and $y^*_{i\alpha} = \acute{x}_{i\alpha} \hat{\theta}_\alpha + \epsilon^*_i$ and ϵ^*_i, $i = 1, ..., n$, be i.i.d from the distribution that puts mass n^{-1} on each $\sqrt{\frac{n}{m}}(r_i - \bar{r})\sqrt{\frac{1-p}{n}}$, $i = 1, ..., n$. Under the linear model

$$\hat{\Gamma}_{n,m}(\alpha) = \frac{||Y - X_\alpha \hat{\theta}_\alpha||^2}{n} + \frac{p_\alpha}{m(n-p)} \sum_{i=1}^{n} (r_i - \bar{r})^2.$$

2.8.4 Generalized Linear Models

McCullagh and Nelder (1989), have provided many examples of generalized linear models, such as logit models, log-linear models, gamma-distributed data models, and survival data models. Let $y_1, ..., y_n$ are the independent response variables, and

$$\mu_i = E(y_i|x_i) = \mu(\eta_i), \qquad Var(y_i|x_i) = \sigma_i^2 = \phi\dot{\mu}(\eta_i),$$

$$i = 1, ..., n, \tag{2.25}$$

where $\dot{\mu}(\eta) > 0$ is the derivative of known differentiable function $\mu(\eta)$; $\phi > 0$ is an unknown scale parameter; the values of explanatory variables by a continuously differentiable link function f is

$$f(\mu(\eta_i)) = \acute{x}_i\theta; \tag{2.26}$$

where θ is a p vector of unknown parameters. Let A be collection of subsets of $\{1, ..., p\}$ and let

$\mu_i = \mu(\eta_{i\alpha})$, $\sigma_i^2 = \phi\ \dot{\mu}(\eta_{i\alpha})$, $\eta_{i\alpha} = (f\ o\ \mu)^{-1}(\acute{x}_{i\alpha}\theta_\alpha)$, $i = 1, 2, ..., n$ be the model corresponding to α, where $x_{i\alpha}$ and θ_α are defined the same as before and, the optimal model is still the correct model with the smallest size. Since the distribution of y_i in equation (2.25) and equation(2.26), is not specified, so we may not be able to obtain the maximum likelihood estimator of θ_α therefore, we follow the general estimation equation approach. That is, under model α, θ_α is estimated by $\hat{\theta}_\alpha$, a solution of

$$\sum_{i=1}^{n} x_{i\alpha}\ \Psi(\acute{x}_{i\alpha}\gamma)\ [y_i - f^{-1}(x_{i\alpha}\gamma)] = 0,$$

where $\hat{\theta}_\alpha$ is a weighted least squares estimator of θ_α, and Ψ is the first-order derivative of $(fo\mu)^{-1}$. We can use the same modified bootstrap model selection procedures that we discussed in linear model for selecting a model from A by select a model that minimizes

$$\hat{\Gamma}_{n,m}(\alpha) = E_* \sum_{i=1}^{n} \frac{[y_i - \mu(\hat{\eta}^*_{i\alpha})]^2}{n\ \upsilon_{i\alpha}},$$

over $\alpha \in A$, where

$\upsilon_{i\alpha} \equiv \dot{\mu}(\hat{\eta}_{i\alpha})$, $\hat{\eta}_{i\alpha} = (f\ o\ \mu)^{-1}(\acute{x}_{i\alpha}\hat{\theta}_\alpha)$, $\hat{\eta}^*_{i\alpha} = (f\ o\ \mu)^{-1}(\acute{x}_{i\alpha}\hat{\theta}^*_\alpha)$

and θ^*_α is bootstrap of $\hat{\theta}_0$ obtained by bootstrapping (residual, pairs).

- Bootstrapping Residuals.

  We generate i.i.d $\epsilon^*_1, ..., \epsilon^*_n$ from the distribution putting mass n^{-1} to $\sqrt{\frac{n}{m}(r_i - \bar{r})}$, where $r_i = [y_i - f^{-1}\ (\acute{x}_i\hat{\theta})]/\sqrt{\upsilon_i}$, where $\hat{\theta}$ and $\upsilon_i$are $\hat{\theta}_\alpha$ and $\upsilon_{i\alpha}$ with $\alpha = \{1, ..., p\}$. θ^*_α can be defined as the linear bootstrap estimator

$$\theta^*_\alpha = \hat{\theta}_\alpha - \hat{M}^{-1}_\alpha \sum_{i=1}^{n} x_{i\alpha}\Psi(\acute{x}_{i\alpha}\hat{\theta}_\alpha)\sqrt{\upsilon_{i\alpha}\epsilon^*_i},$$

  where

$$\hat{M_\alpha} = \sum_{i=1}^{n} \Psi^2(\acute{x}_{i\alpha}\hat{\theta})\ \upsilon_{i\alpha}\ x_{i\alpha}\acute{x}_{i\alpha}.$$

- Bootstrapping Pairs.

  We generate i.i.d Paris $(x^*_1, y^*_1), ..., (x^*_m, y^*_m)$ from the distribution putting mass n^{-1} to each (x_i, y_i), and let θ^*_α to be the linear bootstrap estimator

$$\tilde{\theta}^*_\alpha = \hat{\theta}_\alpha - \ \hat{M}^{-1}_\alpha \sum_{i=1}^{n} x^*_{i\alpha}\Psi(\acute{x}^*_{i\alpha}\hat{\theta}_\alpha)[y^*_i - f^{-1}(\acute{x}^*_{i\alpha}\hat{\theta}_\alpha)].$$

2.9 Illustrative Examples

2.9.1 Detroit homicide data for 1961-73 used in the book Subset Regression by (Miller, 2002)

The data are on the homicide rate in Detroit for the years 1961 to 1973. The data were originally collected and discussed by (Fisher, 1976) but the complete data set first appeared in ((Gunst and Mason, 1980), Appendix A). Miller (2002) discusses this data set throughout his book 'Subset selection in regression'. There were 13 observations and only the first 11 variables were used in Miller's analysis as predictors. The outcome is the number of homicides per 100000 of population; the predictors: $x_1 \equiv$ Full-time police per 100000 population, $x_2 \equiv$ Percent unemployed in the population, $x_3 \equiv$ Number of manufacturing workers in thousands, $x_4 \equiv$ Number of handgun

licences per 100000 population, $x_5 \equiv$ Number of handgun registrations per 100000 population, $x_6 \equiv$ Percent homicides cleared by arrests, $x_7 \equiv$ Number of white males in the population, $x_8 \equiv$ Number of non-manufacturing workers in thousands, $x_9 \equiv$ Number of government workers in thousands, $x_10 \equiv$ Average hourly earnings, $x_11 \equiv$ Average weekly earnings, $x_12 \equiv$ Death rate in accidents per 100000 population, $x_13 \equiv$ Number of assaults per 100000 population, $Y \equiv$ Number of homicides per 100000 of population. In this example none of the stepwise methods has performed well in finding the best-fitting subsets of three or four variables. Table (2.4) shows the performance of forward selection, backward elimination, stagewise procedure and best subset regression by searching all possible regression. For convenience we have labeled the input variables 1 through 11 to be consistent with the notation used in (Miller, 2002). The best fitting subset regression with these 11 variables, uses only 3 inputs variables numbered (2,4,11) and has a residual sum of squares of 6.77. Table (2.5) shows the RSS's for different combinations of variables 2, 4, and 11. Forward selection produces a best fit with 3 inputs variables numbered (4,6,10) with residual sum of squares 21.19. Backward selection produces a best fit with 3 inputs variables numbered (3,4,11) with residual sum of squares 23.51 and stagewise methods produce similar results. It is remarkable that there is such a big difference. Note that the usual forward and backward selection algorithms may fail since the linear regression using 11 variables gives essentially a perfect fit.

2.9.2 Passenger Car Mileage

The data for this example come from a study by (Heavenrich et al., 1991). Variation in gasoline mileage among makes and models of automobiles is influenced substantially by the weight and horsepower of the vehicles. The variables are cubic feet of cab space vol, engine horsepower hp, average miles per gallon mpg, top speed miles per

Table 2.4: RSS's for subsets of variables for DETROIT data set. The numbers in brackets are the number of the selected variables

$No.of vars.$	$Forward$	$Backward$	$Stagwise$	$All subset$
2	33.83	134.0	33.83	33.83
	$(4, 6)$	$(4, 11)$	$(4, 6)$	$(4, 6)$
3	21.19	23.51	21.19	6.77
	$(4, 6, 10)$	$(3, 4, 11)$	$(4, 6, 10)$	$(2, 4, 11)$
4	13.32	10.67	13.32	3.79
	$(1, 4, 6, 10)$	$(3, 4, 8, 11)$	$(1, 4, 6, 10)$	$(2, 4, 6, 11)$

Table 2.5: RSS's of DETROIT data set for combinations of variables numbered (2, 4, 11)

$Variable$	2	4	11	$2, 4$	$2, 11$	$4, 11$	$2, 4, 11$
RSS	3080	1522	680	1158	652	134	6.77

hour sp, and vehicle weight/100 lb wt. The number of observation is 82. To predict a car's gas consumption mpg based on vol, hp, sp, and wt we us mpg as the response variable and the other four variables as predictors. Figure 2.6 is a scatterplot matrix showing every pairwise plot between the variables.

We see that hp, sp, and wt are correlated, and also vol and wt. We see also that the response variable mpg is polynomial in hp and sp. We fit a linear model to the data as shown in Table (2.6), as we see that the variable vol is not a significant variable in the mode.

By using some of the methods for model selection we see the following: Forward Stepwise selection chooses the full model buy using the AIC criterion where the AIC = 217.3. Backward and Hybrid Stepwise Selections both choose the model without vol where AIC = 215.8. Best subset selection chooses the model without vol using C_p Mallow's.

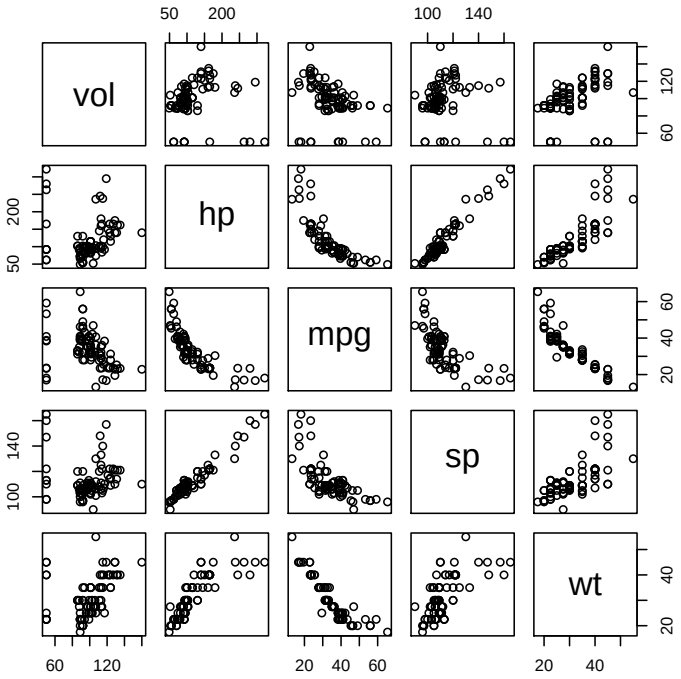

Figure 2.10: A scatterplot matrix of the Passenger Car Mileage data.

Table 2.6: Results from a Linear Regression fit to the Passenger Car Mileage, mpg.

Coefficients:	Estimate	Std.Error	t-value	$\Pr(>\|t\|)$	
(Intercept)	192.43775	23.53161	8.178	$4.62e-12$	$***$
vol	-0.01565	0.02283	-0.685	0.495	
hp	0.39221	0.08141	4.818	$7.13e-06$	$***$
sp	-1.29482	0.24477	-5.290	$1.11e-06$	$***$
wt	-1.85980	0.21336	-8.717	$4.22e-13$	$***$

2.10 Review

We have described a number of approaches to variable subset selection with linear regression. Stepwise procedures such as forward, backward, Hybrid stepwise, and forward-stagewise regression (FS), compare models that have the same number of parameters and they stop the process when including or dropping any of the variables does not improve the fit substantially. These methods typically produce a model that is interpretable but has high variance so, they do not reduce the prediction error of the full model.

The subset selection criteria we have considered such as Mallow's C_p (Mallows, 1973), Akaike Information Criterion (AIC) (Akaike, 1974), Bayesian Information Criterion (BIC) (Schwarz, 1978), Extended Bayesian Information Criterion BIC_γ (Chen and Chen, 2008), Another Extended Bayesian Information Criterion BIC_q (Xu and McLeod, 2009), are all monotone function of the residual sum of the square.

Then we give an over view to All Subsets Selection where the search for the best subsets with minimum RSS or any other criteria can be done by computing all possible regressions but the amount of computation required can be formidable.

Moreover, we have described the tournament screening approach for subset selection when a small sample size, n, and extremely high-dimensional features or covariates space, p, are used.

Also we have introduced subset selection by Cross Validation which is a statistical method for validating a predictive model. A data set partitioned into subsets, a subset of the data are held out, to be used as validating sets; a model is fit to the remaining data as training set and used to predict for the validation set. The results across the validation sets can be averaged or combined to produce a measure of predication accuracy. One form of Cross Validation is LOOCV, leaves out a single observation

at a time. Another, K-fold cross-validation, splits the data into K subsets; each is held out as the validation set. Finally, Delet-d cross validation, the best method for use with model selection in linear regression Shao (1993), uses random sample of size d as a validation set with larger cross-validation sample than is used in K-fold cross validation.

An efficient algorithm, the Leaps and Bounds procedure for finding the best subset regression gives smallest residual sum of squares has intrusted.

In addition, we present BestReg package in R (McLeod and Xu, 2009). BestReg package utilizes the regsubset function in leaps package to find the models with smallest sum of squares for size $k = 0, 1, ..., p$.

Finally, we have described the two consistent modifications of bootstrap selection procedures in linear models developed by (Shao, 1996). Then Shao extended this procedure to more complicated problems such as the nonlinear models, generalized linear models, and autoregressive time series. This procedure select a subset from p explanatory variables that related to a response variable by minimizing bootstrap estimates of the prediction error.

Chapter 3

Regularization Methods

3.1 Introduction

In these methods, coefficients are shrinkage rather than make a choice to include or remove them from the model. This process is more continuous therefore, we get lower variance than by subset selection and also reduce the prediction error of the full model. Shrinkage often improves prediction accuracy, trading off decreased variance for increased biased discussed in (Hastie et al., 2009). These methods are also called Shrinkage Methods. In these methods, we standardize the variable in advance because these methods work on the magnitude of the coefficients. Suppose $y = (y_1, y_2, ..., y_n)^{'}$ is the response vector. $x_j = (x_{1j}, x_{2j}, ..., x_{nj})^{'}, j = 1, 2, ..p$ are the linearly independent predictors. $X = (x_1, x_2, ..., x_p)$ is predictor matrix. This chapter consists of three sections.

Section (3.2), presents one of the shrinkage methods. Ridge regression penalizes by the sum-of-squares of the parameters.

Section (3.3), presents the Least Absolute Shrinkage and Selection Operator. LASSO penalizes by absolute value of the parameter. In addition, we give an overview on the degrees of freedom of LASSO (Zou et al., 2007).

In Section (3.4), we introduce the least angle regression (LARS) (Efron et al., 2004), which is a new model selection algorithm related to forward selection. Then we move on to the modification on LARS to get LASSO.

3.2 Ridge Regression (Miller, 2002)

Ridge regression was first introduced in statistics by (Hoerl and Kennard, 1970). Ridge regression minimizes the residual sum of squares together with the penalty term. Penalizing by the sum-of-squares of the coefficients will enhance the coefficients estimates and reduce the variance, especially when there are many variables correlated in the model. Ridge regression (Miller, 2002), includes all predictors in the model but with smaller coefficients.

$$\hat{\theta}^{\text{ridge}} = \arg\min_{\theta} \quad [\ \sum_{i=1}^{n}(y_i - \theta_0 - \sum_{j=1}^{n}\theta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}\theta_j^2\ ], \tag{3.1}$$

Where $\lambda \geq 0$ is a tuning parameter that controls the amount of shrinkage. When λ is larger, we get more $\theta_j,\ j = 1, 2, .., p$ are shrinkage toward zero. Note that we usually fit the model without an intercept, and we assume that $y_i,\ x_{ij}$ have been normalized. An equivalent way to write equation (3.1) is

$$\hat{\theta}^{\text{ridge}} = \arg\min_{\theta} \quad [\ \sum_{i=1}^{n}(y_i - \theta_0 - \sum_{j=1}^{p}\theta_j x_{ij})^2], \tag{3.2}$$

subject to

$$\sum_{j=1}^{p}{\theta_j}^2 \leq u,$$

where u is a tuning parameter and is in correspondence with λ, the larger the λ the smaller the u.

We also can write equation (3.1) in a matrix form as follow:

$$\hat{\theta}^{\text{ridge}} = (X^T X + \lambda I)^{-1} X^T y, \tag{3.3}$$

where I is p x p identity matrix. By using a quadratic penalty $\theta^T\theta$, the solution of ridge regression will be a linear function of y, so even if X^TX is not of full rank, the problem will be nonsingular, because the ridge solution will add a positive constant to the diagonal of X^TX before getting the inverse. Ridge regression is a continuous process shrinks coefficients toward zero but does not set them to zero, so it keeps all variables in the model. Therefore, we can not easily interpret the ridge model, but the ridge model is more stable than subset models.

3.3 Least Absolute Shrinkage and Selection Operator (Tibshirani, 1996)

Tibshirani (1996) proposed a new method for variable selection that produces an accurate, stable, and parsimonious model called lasso. It penalizes by absolute norm of the coefficients. The LASSO is a constrained version of OLS (Ordinary least squares).

$$\hat{\theta}^{\text{LASSO}} = \arg\min_{\theta} \ [\ \sum_{i=1}^{n}(y_i - \theta_0 - \sum_{j=1}^{p}\theta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}|\theta_j|\ ], \tag{3.4}$$

where $\lambda > 0$ is a smoothing or regularization parameter that controls the amount of shrinkage.

An equivalent way to write equation (3.4) is

$$\hat{\theta}^{\text{LASSO}} = \arg\min_{\theta} \ [\ \sum_{i=1}^{n}(y_i - \theta_0 - \sum_{j=1}^{p}\theta_j x_{ij})^2], \tag{3.5}$$

subject to

$$\sum_{j=1}^{p}|\theta_j| \leq u,$$

where u is a tuning parameter and there is a relation between u and λ.

If u is large enough, then we get the usual least squares estimates. In contrast, using a small value of u will shrink the coefficients estimates by setting some of them equal to zero.

We assume that y_i and x_{ij} have been normalized. For all $\lambda \geq 0$, the solution of θ_0 is $\hat{\theta_0} = \bar{y}$, where $\bar{y} = \sum_{i=1}^{n} y_i/n$. Lasso does model selection by setting some of the variables to zero and that produce an easily interpretable model. Therefore, LASSO combines the best feature of ridge regression and subset selection. Precisely, LASSO does not depend on subsets but on a continuous shrinkage operation that can set some of the coefficients to zero. Continuous shrinkage often improves prediction accuracy, trading off decreased variance for increased biased. Note that the lasso estimate is a non-linear and non-differentiable function of the response values; therefore we need a quadratic optimization or iterative techniques to solve equation (3.4). (Tibshirani, 1996) suggested to solve equation (3.5) by starting from the overall least squares estimate, introducing the constraints sequentially, searching for a feasible solution satisfying Kuhn-Tucher conditions (Lawson and Hanson, 1974). Efron et al. (2004) proposed another model selection algorithm, ***Least Angle Regression*** (LARS), which can be modified to solve the whole Lasso.

3.3.1 Estimation of the Tuning Parameter u

We describe two methods for the estimation of the lasso parameter u: Cross-Validation and Generalized Cross-Validation. In these methods we assumed that the observations (x, y) are drown from some unknown distribution. Assume a model $y = f(x)+\epsilon$ where $E(\epsilon) = 0$ and $\text{Var}(\epsilon) = \sigma^2$. The mean squared error for an estimate $\hat{f}(x)$ of $f(x)$, is defined by

$$\text{MSE} = \text{E}(\hat{f}(x) - f(x))^2.$$

The predication error of $\hat{f}(x)$ is given by

$$\text{PE} = \text{E}(Y - \hat{f}(x))^2$$

$$\text{PE} = \text{MSE} + \sigma^2.$$

For the linear models $f(x) = \acute{x}\beta$ and the mean squared error has the simple form $\text{MSE} = (\hat{\theta} - \theta)^{'} V (\hat{\theta} - \theta)$, where V is the population covariance matrix of x.

(1) Cross - Validation method.

- Let $s = \frac{u}{\sum \hat{\theta}_j^o}$ be the normalized parameter where $\hat{\theta}_j^o$ is the full least squares estimates.

- Estimate the prediction error for the lasso procedure over a grid of values of s from 0 to 1 inclusive by five fold cross-validation as described in (Efron et al., 2004).

- Selectee the value of $\hat{s}$ that yield to the lowest estimated PE..

(2) Generalized Cross - Validation method.

- Write the constraint $\sum_{j=1}^{p} |\theta_j| \leq u$ as $\sum_{j=1}^{p} \frac{\theta_j^2}{|\theta_j|} \leq u$. This constraint is equivalent to adding $\lambda \sum_{j=1}^{p} \frac{\theta_j^2}{|\theta_j|}$ to the residual sum of squares, with λ depending on u.

- Write the constrained solution $\tilde{\theta}$ as the ridge regression estimator

$$\tilde{\theta} = (\acute{X}X + \lambda\, W^{-})^{-1}\, \acute{X}\, Y,$$

where $W = diag(|\tilde{\theta_j}|)$ and W^- denotes a generalized inverse.

- Calculate

$$\text{GCV}(u) = \frac{1}{n} \frac{\text{RSS}(u)}{(1 - p(u)/n)^2},$$

where $p(u)$ is the number of effective parameters in the constrained fit $\tilde{\theta}$

$$p(u) = \text{tr}(X(\acute{X}X + \lambda W^-)^{-1} \acute{X}),$$

and $\text{RSS}(u)$ is the residual sum of squares for the constrained fit with constraint u.

3.3.2 On the Degrees of Freedom of the LASSO (Zou et al., 2007)

Zou et al. (2007) proved that the number of nonzero coefficients is an unbiased estimate for the degrees of freedom of lasso. This unbiased estimator is also consistent, and can be used to construct C_p, BIC type model selection criteria. This is a finite sample, and it is true as long as the predictor matrix X is a full tank. This result provides the optimal lasso fit through the LARS algorithm. Suppose
$\hat{U}_\lambda(y)$ is the lasso fit, Z_λ is the active set of $\hat{\beta}_\lambda$, where $Z = [\ j : sign(\beta)_j \neq 0\ ]$.
Therefore $\text{df}(\hat{U}_\lambda) = E|Z_\lambda|$, where $rank(X) = p$, X is full rank.
Then, $\hat{\text{df}}(\hat{U}_\lambda) = |Z_\lambda|$
Given any set of data, this method works as follow:

- Compute lasso through LARS algorithm.
- $\hat{\text{df}}(\lambda) = |Z_\lambda|$.

In general, for any model, Stein's Unbiased Risk Estimation (SURE)theory (Stein, 1981) provides a definition of the degrees of freedom. Efron et al. (2004) has shown that

$$\mathrm{df}(\hat{u}) = \sum_{i=1}^{n} \mathrm{Cov}\left(\hat{u}_i, y_i\right)/\sigma^2 \tag{3.6}$$

where $\hat{u} = \delta(y)$ is the fit of the model. $y \sim (\mu, \sigma^2\, I)$,where μ is the true mean, σ^2 is the common variance.

Through the SURE theory, the best unbiased estimate for $\mathrm{df}(\hat{u}_\lambda)$ should provide an unbiased estimate for the prediction error of $\hat{u}_\lambda$. By using the covariance penalty method Mallow's C_p

$$C_p(\hat{u}) = \frac{\|y - \hat{u}\|^2}{n} + \frac{2}{n}\, \hat{\mathrm{df}}(\hat{u})\sigma^2. \tag{3.7}$$

Where the C_p and AIC give the same result. The BIC for LASSO is

$$\mathrm{BIC}(\hat{u}) = \frac{\|y - \hat{u}\|^2}{n\sigma^2} + \frac{\log(n)}{n}\, \hat{\mathrm{df}}(\hat{u}). \tag{3.8}$$

We need to find the optimal λ, λ^* to get the optimal LASSO model.

$$\lambda^* = \arg\min_{\lambda}\ \{\frac{\|y - \hat{u}_\lambda\|^2}{n\sigma^2} + \frac{A}{n}\, \hat{\mathrm{df}}(\lambda)\}, \tag{3.9}$$

where $A = 2$ for AIC and, $A = \log(n)$ for BIC.

We are looking for the optimal λ that minimizes AIC or BIC. Which of AIC or BIC criteria is preferred to get the optimal λ? Use AIC when you are looking for the model with the optimal prediction performance, and use BIC when your concern is the sparsity of the model.

3.4 Least Angle Regression (Efron et al., 2004)

Efron et al. (2004), produced a new technique that derived from a stagewise procedure that reduced the small steps toward the final model, but not as much as forward selection. LARS needs just P (number of covarates) steps to get to the final least square- estimates while LASSO can have more than p-steps. However, the two results are almost identical. LARS algorithm works as follow:

- Standardize predictors to have mean 0, and variance 1.
- Start with only the intercept in the model and all the coefficients equal to zero ($\theta_1 = \theta_2 = ... = \theta_p \ = 0$).
- Start with the residual equal to r.

$$r = y - \bar{y}.$$

- Select the first variable x_j the one with the highest current correlation (c_j) with r. Performing the active set Z_k during the step k leaving the residual vector as responds variable.

$$c_j = \text{Corr}(y, x_j),$$

$$c_j = \acute{x}_j(y - \hat{\theta}).$$

- Move θ_j from 0 in the direction of its least-square coefficient until other predictors, that are not in the active set Z_k for example, (x_l) which has correlation as x_j with the current residual.
- In the direction of their joint least squares coefficients, move both θ_j and θ_l until other predictors one or more, for example, (x_h) that is not in the active set Z_k has as much correlation with the current residual as x_j.

- Repeat the process until all the predictors have entered the model, resulting in the full least-squares solutions.

Note that in LARS , the predictors will stay forever in the active set when they are added. Efron et al. (2004) applied a simple modification of the LARS algorithm to implement the lasso. LARS modification is faster than the traditional lasso, and it finds all the possible lasso estimates for the model. We can perform this restriction on LARS to get lasso which is dropping the variable from the active set of variables if its coefficients hits zero and recompute the current joint least squares direction. The least-angle regression and LASSO produce more accurate, stable, and interpretable predictions compared to other variable selection procedures such as stepwise or ridge regression.

3.5 Illustrative example

3.5.1 Passenger Car Mileage

We use the same data set that we used in Example (2.9.2). We performed a Regularization Method on this data set. Ridge regression shrinkage the variables toward zero and keeps all the variables as shown in Figure (3.1).

The lines in the Figure (3.1) represent the $\hat{\theta}^{\text{ridge}}$ as a function of λ for each of the independent variables. LASSO shrinkage some of the variables toward zero and set the other variables to zero as shown in Figure (3.2).

The lines in the Figure (3.2) represent the $\hat{\theta}^{\text{LASSO}}$ as a function of u for each of the independent variables.

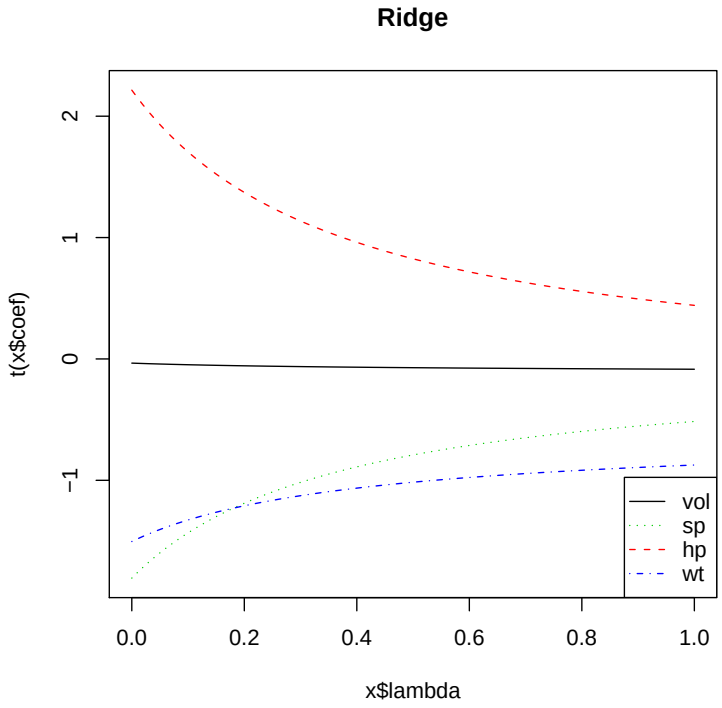

Figure 3.1: Ridge Regression. In black is the line representing $\hat{\theta}^{\text{ridge}}_{\text{vol}}$ as a function of λ ; same for $\hat{\theta}^{\text{ridge}}_{\text{hp}}$ red, $\hat{\theta}^{\text{ridge}}_{\text{sp}}$ green, and $\hat{\theta}^{\text{ridge}}_{\text{wt}}$ blue.

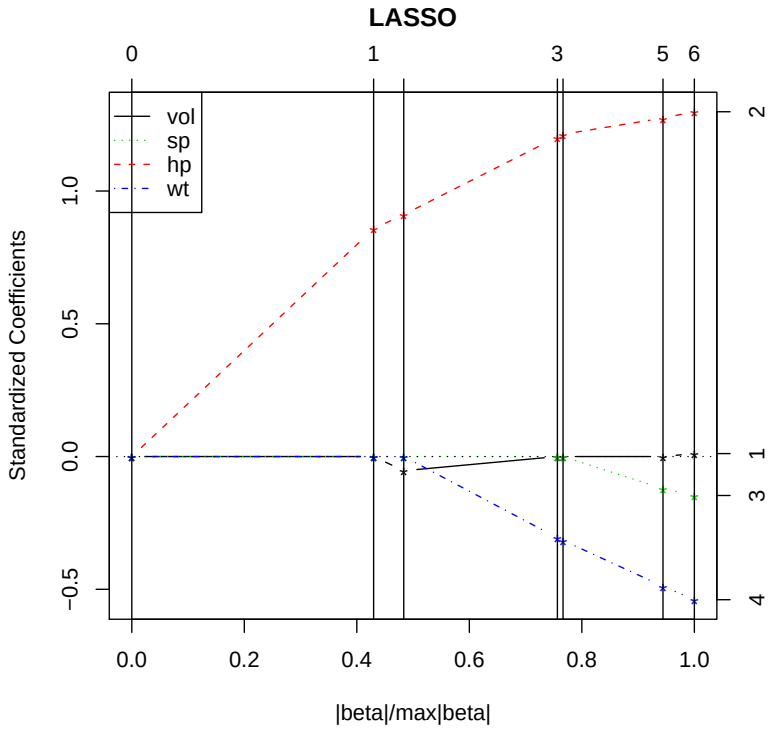

Figure 3.2: LASSO Regression. In black is the line representing $\hat{\theta}_{\text{vol}}^{\text{LASSO}}$ as a function of u; same for $\hat{\theta}_{\text{hp}}^{\text{LASSO}}$ red, $\hat{\theta}_{\text{sp}}^{\text{LASSO}}$ green, and $\hat{\theta}_{\text{wt}}^{\text{LASSO}}$ blue.

3.6 Review

We have introduced the shrinkage methods. This process is more continuous than subset selection and does not suffer a lot from high variability as subset selection.

Ridge regression (Hoerl and Kennard, 1970) minimizes the residual sum of squares subject to abound on the l_2-norm of the coefficients. Ridge regression keeps all the predictors in the model therefore, it can not produce a parsimonious model.

The LASSO (Tibshirani, 1996) penalized least squares method imposing an l_1-penalty on the regression coefficients. The LASSO does both continuous shrinkage and automatic variable selection simultaneously.

Finally, we have described a new model selection algorithm Least Angle Regression LARS with its modification to implement LASSO with less computer time than previous methods. LARS a technique that derived from a stagewise procedure that reduced the small steps toward the final model, but not as much as forward selection.

Chapter 4

Application of lasso in logistic regression

4.1 Introduction

Logistic regression is widely used as the method of analysis in a situation where the outcome variable is discrete binary or dichotomous. Typically logistic regression models are used to predict the probability of occurrence of an event.

Section One, presents the logistic regression model where the outcome variable is binary or dichotomous. In Section Two, we give an overview of the efficient methods for estimating constrained parameters with application to lasso logistic regression (Tian et al., 2008). We implemented the R code for the both procurers: the Faster Quadratic Lower-Bound QLB algorithm for estimation in lasso logistic regression where the convergence is not generally ensured and the Pseudo-Newton method, which is faster than the Fastest Quadratic Lower-Bound algorithm. In addition, we calculated the bootstrap variance estimation. Finally, we tried to use the faster QLB algorithm in the case when $p > n$ the number of the parameter is bigger than the number of the observation, but it did not work well.

4.2 Logistic Regression

Logistic regression is a model used to predict a discrete outcome by using a set of predictor variables that maybe numerical, categorical or any other form (Hosmer et al., 2000) and (Agresti, 1996). The response variable is usually binary or dichotomous,

such as having heart disease or not, success or failure. However logistic regression can be used in cases where the response variable has more than two cases, which is called multinomial. In this section, we are going to focus our work on cases where the response variable is binary. In other words, the response variable can take the value 1 with probability of success p, or 0 with probability of failure q. Logistic regression is part of a class of models called generalized linear models, and it is widely used in the medical area, social sciences, and business. We prefer to use logistic regression rather than linear regression when the outcome is binary because, the predicted values will become greater than 1 and less than 0, with linear regression, and this is not acceptable. In addition, the variance in the binary variable is pq, while the variance of y across x should be constant in the assumption of linear regression. Finally, it is a function that is easy to used and interpret.

4.2.1 The Model

In logistic regression, there is no linear relationship between the predictor and the response variables. The logistic curve has the S-shape, which relates the explanatory variables that take any value from $-\infty$ to ∞, as an input to the rolling mean of the response variable $\pi(x)$ that take values between 0 and 1 as an out put. There are two cases for the logistic regression:

4.2.1.1 Univariate Case

There is only one independent variable. Suppose, $\pi(x) = E(y|x)$ is the conditional mean of y given x then,

$$\pi(x) = \frac{\exp^{a+\theta x}}{1 + \exp^{a+\theta x}}, \tag{4.1}$$

where a is the constant of the equation, θ is the coefficient of the predictor. Transformation of $\pi(x)$ is the logit transformation

$$f(x) = \ln\{ \frac{\pi(x)}{1-\pi(x)} \}, \tag{4.2}$$

$$f(x) = a + \theta x, \tag{4.3}$$

which is linear in the parameters.

4.2.1.2 Multiple Case

Suppose we have p independent variables denoted by $X^T = (x_1, x_2, ..., x_p)$, then the logit transformation is:

$$f(x) = a + \theta_1 x_1 + \theta_2 x_2 + ... + \theta_p x_p, \tag{4.4}$$

$$\pi(x) = \frac{\exp^{f(x)}}{1 + \exp^{f(x)}}. \tag{4.5}$$

4.2.2 Fitting Logistic Regression Models

Maximum likelihood estimation MLEia a popular method used for fitting logistic regression models (Dobson, 2002). The maximum likelihood estimation picks the parameters that maximize the probability of the sample data. For the two cases of logistic regression, suppose we have n independent observation $\{(x_i, y_i), i = 1, 2, ..., n\}$ the likelihood function is given by

$$l(\theta) = \prod_{i=1}^{n} \Lambda(x_i), \tag{4.6}$$

where

$$\Lambda(x_i) = \pi(x_i)^{y_i}\ [1 - \pi(x_i)]^{1-y_i}, \tag{4.7}$$

where $\pi(x_i)$ is defined as equation (4.1) in the univariate case or as equation (4.5) in the multiple case. The logarithmic likelihood function is given by

$$L(\theta) = \log\left(l(\theta)\right) = \sum_{i=1}^{n}\{\ y_i \log \pi(x_i) + (1 - y_i) \log\left(1 - \pi(x_i)\right)\ \}. \tag{4.8}$$

We maximize $L(\theta)$, as follows:

- Set the result of the derivatives to zero, and get the score equations,

$$\sum_{i=1}^{n}\{\ y_i - \pi(x_i)\} = 0,$$

the first score equation specifies that $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \pi(x_i)$, because the first component of x_i is 1 and,

$$\sum_{i=1}^{n} x_i(y_i - \pi(x_i)) = 0. \tag{4.9}$$

- Use the Newton-Raphson algorithm to solve equation (4.9). Calculate the second-derivative (Hessian matrix)

$$\frac{\partial^2 L(\theta)}{\partial\theta\partial\acute{\theta}} = -\sum_{i=1}^{n} x_i\acute{x}_i\ \pi(xi)(1 - \pi(xi)). \tag{4.10}$$

- Update the coefficients, a single Newton update is:

$$\theta^{\text{new}} = \theta^{\text{old}} - \left(\frac{\partial^2 L(\theta)}{\partial\theta\partial\acute{\theta}}\right)^{-1} \frac{\partial L(\theta)}{\partial\theta}. \tag{4.11}$$

- Repeat the process until θ^{old} is close enough to θ^{new}.

We can write the above in the matrix notation as follows:

Y is the response vector of y_i, X denotes the n x $(p+1)$ matrix of x_i, π is the vector of fitted response probabilities, K: n x n diagonal matrix of weights with the diagonal element $\pi(x_i, \theta^{\text{old}})(1 - \pi(x_i, \theta^{\text{old}}))$

$$\frac{\partial L(\theta)}{\partial \theta} = \acute{X}(y - \pi)$$
$$\frac{\partial^2 L(\theta)}{\partial \theta \partial \acute{\theta}} = -\acute{X}\ K\ X.$$

The Newton steps is:

$$\theta^{\text{new}} = \theta^{\text{old}} + (\acute{X}KX)^{-1}\acute{X}(y - p), \tag{4.12}$$

$$\theta^{\text{new}} = (\acute{X}KX)^{-1}\acute{X}K(X\theta^{\text{old}} + K^{-1}(y - p)), \tag{4.13}$$

$$\theta^{\text{new}} = (\acute{X}KX)^{-1}\acute{X}Kw, \tag{4.14}$$

where,

$$w = X\theta^{\text{old}} + K^{-1}(y - p). \tag{4.15}$$

This algorithm is called ***Iteratively Reweighed Least Squares*** (IRLS.)

4.3 Quadratic Lower-Bound QLB Algorithm (Tian et al., 2008)

For optimization problems with box or linear inequality constraints, (Tian et al., 2008) have developed a quadratic lower-bound QLB algorithm, then generalized this algorithm to penalized problems, the faster QLB algorithm, in which the penalty function may not be totally differentiable. This algorithm is used for estimation in *lasso logistic regression*, where the convergence is not totally ensured. They also have

developed a Pseudo-Newton method that combines the good features of the QLB and Newton methods which are simplicity and fast convergence. They used these algorithms with independent binary data.

4.3.1 QLB Algorithm for Optimization with box or Linear inequality Constraints

The algorithm consists of two steps:

- T-step : find GM, Global Majorization Matrix, B, so that
  $B = \acute{C}C$ where C is an upper triangular matrix calculated via Cholesky decomposition.
- M-step : update the current estimate by using some built-in SPLUS functions (Venables and Ripley, 2002).

The algorithm is working as follow:
Defined a positive defined matrix $B > 0$, and B is independent of θ, so that $\forall\, \theta \in \mathbf{R^q}$:

$$B \;\geq\; -\nabla^2 l(\theta), \tag{4.16}$$

where $\nabla^2 l(\theta)$ is the Hessian matrix, and $B = \acute{C}C$ where C is an upper triangular matrix calculated via Cholesky decomposition. Let ξ be a q-vector depend on B and θ^t where $\theta^t \in [a, b]$,

$$\xi = \xi(B, \theta^{(t)}) = (C^{-1}\acute{)}\Delta l(\theta^{(t)}) + C\theta^{(t)}. \tag{4.17}$$

We want to find constrained MLE $\hat{\theta}$

$$\hat{\theta} = \arg \max_{\theta \epsilon [a,b]} \; l(\theta), \tag{4.18}$$

where

$$l(\theta) = \sum_{i=1}^{m} \{\ y_i(\acute{x}_{(i)}\theta) - n_i\ \ln[1 + \exp(\acute{x}_{(i)}\theta]\ \}. \tag{4.19}$$

$l(\theta)$ is twice continuously differentiable and concave function. $y_i \sim Binomal(n_i, p_i)$, $\text{logit}(p_i) = \acute{x}_i\theta$, $1 \leq i \leq m$ and y_i are independent and denote the number of subjects with positive response in n_i trials.

Since the QLB algorithm has ascent property that is the likelihood increases in each QLB iteration, then finding equation (4.18) is similar to iterative finding

$$\theta^{(t+1)} = \arg \max_{\theta \epsilon [a,b]} Q(\theta|\theta^{(t)}), \tag{4.20}$$

where for any θ and $\acute{\theta} \epsilon\ R^q$

$$Q(\theta|\acute{\theta}) = l(\acute{\theta}) + (\theta - \acute{\theta}\ \acute{)}\nabla l(\acute{\theta}) - 0.5\ (\theta - \acute{\theta}\ \acute{)}B(\theta - \acute{\theta}). \tag{4.21}$$

The equation (4.20) becomes

$$\theta^{(t+1)} = \arg \min_{\theta \in [a,b]}\ \|\xi(\beta, \theta^{(t)})\ -\ C\theta\|^2. \tag{4.22}$$

To update $\theta(t)$ through equation (4.22), we can use some built-in SPLUS functions like nnls .fit, nonnegative least squares, and nlregb , nonlinear least squares subject to box constraints, to perform M-step.

4.3.2 QLB Algorithm for Penalized Problems or The Faster QLB algorithm

We want to find penalized MLE $\tilde{\theta}$ so that

$$\tilde{\theta} = \arg\max_{\theta} \ l_{\lambda}(\theta) = \arg\max_{\theta}\{ \ l(\theta) - \lambda J_1(\theta) \ \}, \tag{4.23}$$

where

$J_1(\theta)$ is the penalty function, $\lambda > 0$ is the smoothing parameter.

Using QLB algorithm, we can find $\tilde{\theta}$ by iteratively calculating

$$\theta^{(t+1)} = \arg\max_{\theta} \ Q_{\lambda}(\theta|\theta^{(t)}) = \arg\max_{\theta} \ \{ \ Q(\theta|\theta^{t)}) - \lambda J_1(\theta) \ \}, \tag{4.24}$$

where Q is the same as equation (4.21).

4.3.3 A Pseudo-Newton method

The QLB algorithm is usually criticized for its slow convergence in high-dimensional data analysis. Therefore, (Tian et al., 2008) developed a Pseudo-Newton method which retains the speed of convergence of the Newton method and the simplicity of QLB algorithm. In this method we want to find

$$\hat{\theta} = \arg\max_{\theta\epsilon \ [a,b]} \ l(\theta). \tag{4.25}$$

By using the Newton method, finding equation (4.25) is similar to iterative finding

$$\theta^{(t+1)} = \arg\min_{\theta \ \epsilon \ [a,b]} \ \|\xi(-\nabla^2 \ l(\theta)^{(t)}, \theta^{(t)}) - C^{(t)}\theta\|^2, \tag{4.26}$$

the Cholesky decomposition should be calculated at each iteration in equation (4.27),

$$-\nabla^2\ l(\theta^{(t)}) = (C^{(t)}\ \acute{)}\ C^{(t)}, \tag{4.27}$$

If $-\nabla^2\ l(\theta^{(t)})$ in equation (4.27) replaced with a surrogate matrix $\hat{\beta}^u > 0$, then Pseudo-Newton algorithm is defined by the following iteration:

$$\theta^{(t+1)} = \arg \min_{\theta\ \epsilon\ [a,b]}\ \|\xi(\hat{B}^u, \theta^{(t)}) - C_*\theta\|^2, \tag{4.28}$$

where

$$\hat{B}^u = \acute{C}_*\ C_*, \tag{4.29}$$

is the expected information which is calculated only once, $\hat{\theta}^u$ is unconstrained MLE of θ in equation (4.18).

4.3.4 The Faster QLB and the Pseudo-Newton algorithms in Logistic regression with constraints

The first step in the QLB algorithm is to find the GM matrix B so that $B > 0$ and B does not depend on θ and satisfying the equation (4.16). For each i, $0.25 \geq p_i(1-p_i)$ therefore, for the Fastest QLB algorithm the smallest GM matrix (Bohning and Lindsay, 1988) for the logistic regression is

$$B = (\frac{1}{4})\ \acute{X}\ N\ X, \tag{4.30}$$

where $N = \text{diag } (n_1, n_2, ..., n_m)$, $\acute{X} = (x_{(1)}, ...x_{(m)})$ $\qquad i = 1, ..., m$, and

$$\Delta l(\theta) = \sum_{i=1}^{m} (y_i - n_i p_i) = \acute{X}(y - N_p),$$

$$-\Delta^2 l(\theta) = \sum_{i=1}^{m} n_i p_i (1 - p_i) x_i \acute{x_i} = \acute{X} N D X.$$

For Pseudo-Newton algorithm

$$\hat{B}^u = \acute{X}\ N\ \text{diag}(\hat{d}_i^u, ..., \hat{d}_m^u)\ X, \tag{4.31}$$

where

$$D = \text{diag } (p_1\ (1 - p_1)....p_m(1 - p_m)), \tag{4.32}$$

$$\hat{p}_i^u = \frac{\exp\ \{\acute{x}_{(i)}\ \hat{\theta}^u\}}{1 + \exp\{\acute{x}_{(i)}\ \hat{\theta}^u\}}, \tag{4.33}$$

where $\hat{\theta}^u$ is the unconstrained MLE of θ in the logistic model equation (4.19), and $\hat{d}_i^u = \hat{p}_i^u\ (1 - \hat{p}_i^u)$ if $\hat{p}_i^u\ \in\ (0, 1)$, and $\hat{d}_i^u = 0.25$ otherwise, $i = 1, ..., m$.

For a logistic model, the LASSO regression is to find

$$\hat{\theta}^{\text{LASSO}} = \arg\max_{\theta}\ \{\ l(\theta) -\ \lambda\ \sum_{j=1}^{q} |\theta_j|\ \}, \tag{4.34}$$

where $\lambda > 0$ is a smoothing parameter. By using the faster QLB algorithm, we can find equation (4.34) by iteratively calculating

$$\theta^{(t+1)} = \arg\min_{\theta}\ \{\ \|\ \xi(B, \theta^{(t)}) -\ C\theta\|^2\ +\ \lambda\ \sum_{j=1}^{q} |\theta_j|\}, \tag{4.35}$$

where $\hat{\theta}^u$ is the unconstrained MLE of θ in the logistic model, $v = (v_1, ..., v_q\acute{)}$ is a sign vector, $v_j = \text{sign}\,(\hat{\theta}_j^u) = \ +1\ ,\ 0\ ,\ or -1$ corresponding to positive, zero, or

negative values of $\hat{\theta}_j^u$. $\xi(B, \theta^{(t)})$ and B are defined in equation (4.17) and equation (4.30), respectively.

From the property of the lasso solution, we know that $\hat{\theta}^{\text{LASSO}}$ and $\hat{\theta}^u$ share signs (Efron et al., 2004)

$$\theta^{(t+1)} = \text{diag } (v)\ B^{(t+1)},$$

and

$$B^{(t+1)} = \arg \min_{B\ \epsilon\ R_+^q} \parallel \eta(B, \theta^{(t)}) - Z\ B\parallel^2, \tag{4.36}$$

$$\eta(B, \theta^{(t)}) = (\acute{Z})^{-1}\ \{\ \text{diag } (v\ \acute{C}\ \xi(B, \theta^{(t)}))\ -\ 0.5\lambda\}, \tag{4.37}$$

where Z can be obtained via the Cholesky decomposition $\acute{Z}\ Z = \text{diag}(v)\ \acute{C}\ C\ \text{diag}(v)$. Because equation (4.36) is a quadratic optimization problem with non-negative constraints, we can use nnls.fit function in S-plus or nnls in R to solve it. Note that we can get, $\hat{\lambda}^{opt}$, the optimal smoothing parameters by minimizing an approximate generalized cross-validation GCV statistic (Craven and Wahba, 1979) where

$$GCV = \frac{-l(\hat{\theta}_\lambda^{\text{LASSO}})}{m[1-\frac{e(\lambda)}{m}]^2},$$

where

$$e(\lambda) = tr[\ X(\acute{X}NDX + \lambda W^-)^{-1}\acute{X}ND], \tag{4.38}$$

is the effective number of parameters.

$W = diag(|\hat{\theta}_\lambda^{\text{LASSO}}|)$, where W^- is the *Moore-Penrose generalized inverse* of W.

4.3.5 R-Code for fastest QLB algorithm and Pseudo-Newton algorithm

Tian et al. (2008) used SPLUS and some of its built-in functions with these algorithms. We implemented the R-codes for the fastest QLB algorithm and Pseudo-Newton algorithm; then, we used these codes with the same data set they used for lasso logistic regression, which is the Kyphosis data Example **??**and we obtained the same result as (Tian et al., 2008) found.

4.3.6 Simulated study

We have simulated vector y (1000×1) binary data $(0, 1)$ to represent the response variable, and a matrix x (1000×2000) binomial as the explanatory variables. Then we followed all the steps we discussed above to get the QLB algorithm and Pseudo-Newton algorithm with the case $p > n$ (number of parameters $>$ number of observation), but the algorithms did not work well because all the values for the unconstrain $\theta \geq 1000$ are missing. The lasso selects at most n=1000 variables before it saturates, therefore we could not find the optimal λ via the GCV method; also did not get a positive definite matrix when we calculated the Cholesky decomposition, which has effect on the other functions that we used for this method.

4.4 Illustrative Examples

4.4.1 South African Heart Disease

In this example we present an analysis of binary data to demonstrate the statistical use of logistic regression model. The scatterplot in Figure(4.1) is a retrospective sample of white males between 15 and 64, in a heart-disease high-risk region of the Western Cape, South Africa. The aim of the study was to establish the intensity of ischemic heart disease risk factors in the high-incidence region. The response variable,

Table 4.1: Results from a logistic Regression fit for to the South African heart disease data,

Coefficients:	Estimate	Std.Error	Z Score
(Intercept)	−0.5080306	0.2045921	−2.483139
sbp	0.0013387	0.0010581	1.265192
tobacco	0.0165841	0.0048610	3.411664
ldl	0.0331791	0.0106747	3.1082
adiposity	0.0023026	0.0047696	0.4827659
famhist	0.1734290	0.0412746	4.201834
typea	0.0060817	0.0020369	2.985763
obesity	−0.0111711	0.0070315	−1.588722
alcohol	−0.0002364	0.0008284	−0.2853694
age	0.0068440	0.0019868	3.444735

chd, is the presence or absence of myocardial infarction. The independent variables sbp: systolic blood pressure, tobacco: cumulative tobacco (kg), ldl: low densiity lipoprotein cholesterol, adiposity, famhist: family history of heart disease (Present, Absent), typea: type-A behavior, obesity, alcohol: current alcohol consumption, age: age at onset. A sample of size 463 are used. There are roughly two controls per case of chd. Many of the chd positive men have undergone blood pressure reduction treatment and other programs to reduce their risk factors after their chd event. In some cases the measurements were made after these treatments. These data are taken from a larger dataset, described in (Rousseauw et al., 1983). Table (4.1) shows the results from a logistic-regression model fitted by maximum likelihood. This summary includes Z scores for each of the coefficients in the mode. Any coefficients whose Z score is nonsignificant, can be dropped from the model. A Z score significant at the level %5 if it is grater than approximately 2 in absolute value.

From Table (4.1) we can see that systolic blood pressure, sbp, is not significant !

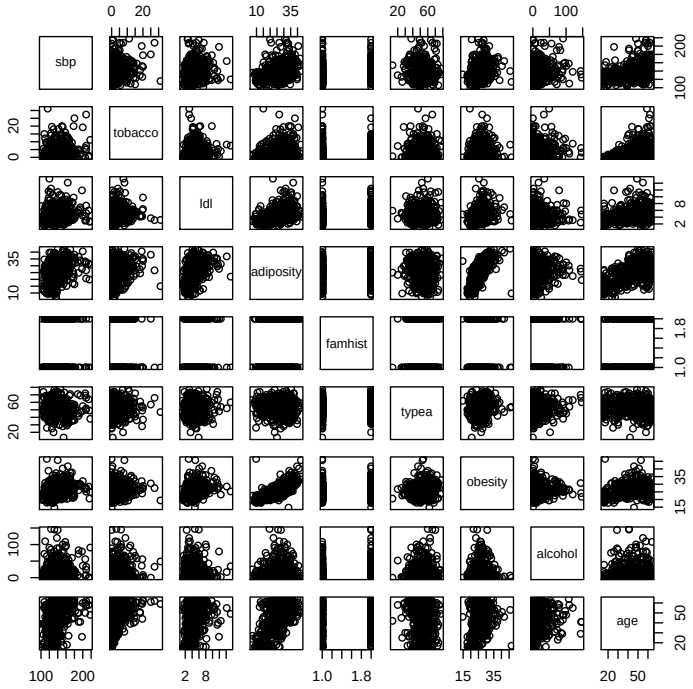

Figure 4.1: A scatterplot matrix of the South African Heart Disease data.

Table 4.2: Results from Stepwise logistic Regression fit for to South African heart disease data,

Coefficients:	Estimate	Std.Error	Z Score
(Intercept)	−0.237407	0.068561	−3.462712
tobacco	0.017263	0.004815	3.585254
ldl	0.032533	0.010076	3.228761
famhist	0.178173	0.041395	4.304215
age	0.006836	0.001603	4.264504

Nor adiposity, alcohol, and obesity which is with negative sign. However, both sbp and obesity are significant on their own and with positive sign. This is a result of the correlation between the set of predictors. In Table (4.2) we did model selection by dropping the least significant coefficient, and refit the logistic model. Repeat the process above until no further terms can be dropped. All Z score grater than approximately 2 in absolute value.

4.4.2 Kyphosis data

These data are taken from a dataset, described in (Hastie and Tibshirani, 1990). The outcome is the status of Kyphosis (1=present, 0=absent); the predictors: $X_1 \equiv$ age in months at time of the operation, $X_2 \equiv$ number of vertebrae levels, $X_3 \equiv$ starting vertebrae level. The data set is for 83 laminectomy patients. We want to determine the risk factor for Kyphosis in this study. We did not include the interaction effects, as (Tian et al., 2008) did, and we included the three main effects and their quadratic effects since the predictor effects are known to be non-linear. The full logistic model is

$$\text{logit}\{Pr(Y=1)\} = \theta_0 + \sum_{j=1}^{3}\theta_j\ x_j\ +\ \sum_{j=1}^{3}\theta_{3+j}\ x_j^2. \tag{4.39}$$

To get the Faster QLB algorithm, we did the following:

- Standardized the data.
- Calculate the unconstrained MLE.

$$\hat{\theta}^u = (\ -2.642, 0.827, 0.767, -2.269, -1.541, 0.032, -1.158)$$

- Find the sign vector.

  $(\ -1,\ 1,\ 1,\ -1,\ -1,\ 1,\ -1)$
- Use the assumption that $\hat{\theta}^{LASSO}$ and $\hat{\theta}^u$ share signs.
- Find $\hat{\lambda}^{optimal}$ via Generalized Cross Validation GCV static . The optimal$\hat{\lambda}^{opt} = 0.351$, and the plot of generalized cross-validation is shown in Figure (4.2).
- Use $\theta^{(0)} = v$ as initial value.
- Find GM matrix B via equation (4.29).
- Use some built-in R functions; nnls defines the Lawson-Hanson NNLS algorithm for non-negative least squares that solves the least squares problem $Ax = b$ with the constraint $x \geq 0$, and nlminb which is a function for unconstrained and constrained optimization using PORT routines to calculate equation (4.34). The lasso solution is:
  $(\ -2.264,\ 0.6581,\ 0.692,\ -1.837,\ -1.252,\ 0.004,\ -0.859)$
  The monotone convergence of the algorithm is shown in the Figure (4.3).

- We calculate the standard error with 1000 bootstrap replications

  $(\ 0.557, 0.482, 0.466, 0.624, 0.607, 0.623, 0.489)$

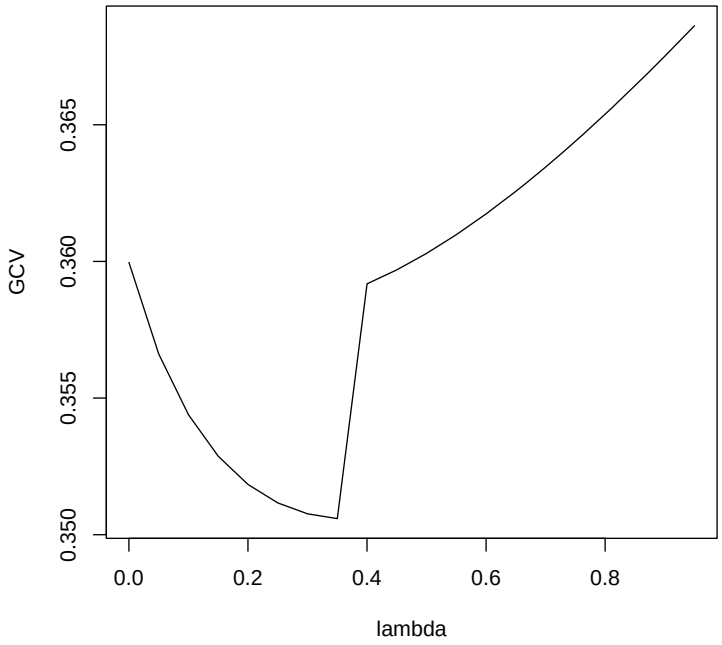

Figure 4.2: Plot of generalized cross-validation for kyphosis data

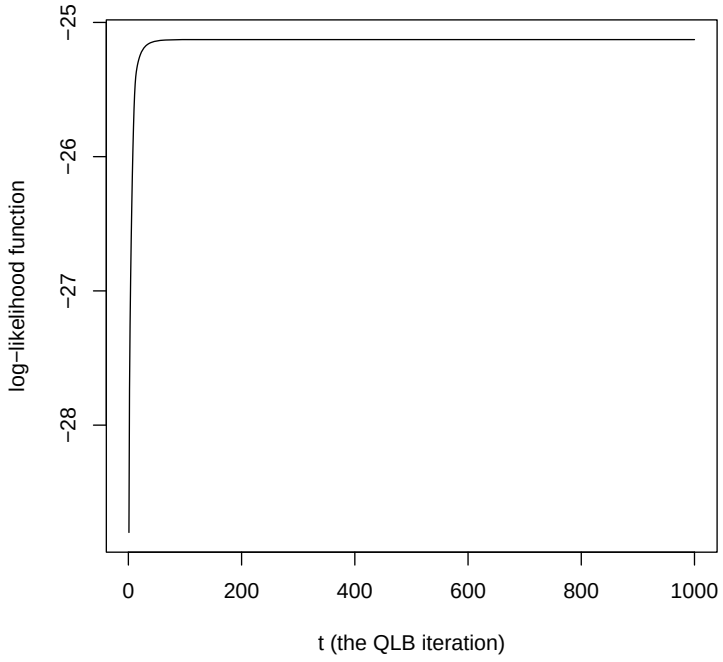

Figure 4.3: The monotone convergence of the Faster QLB algorithm for Kyphosis data

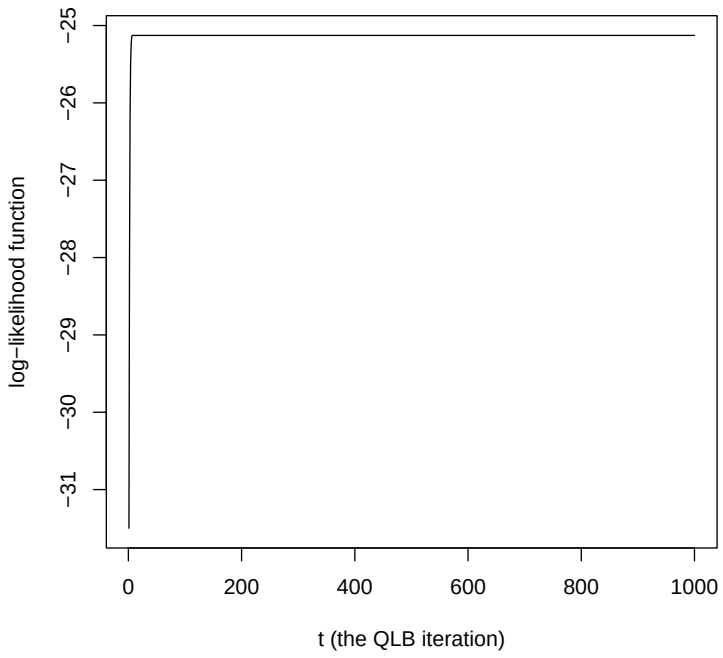

Figure 4.4: The monotone convergence of the Psude-Newton for Kyphosis data

To get the Pseudo-Newton algorithm, we can use the same process above, except use $\hat{B}^u$ in equation (4.31) instead of B in equation (4.30). We used the same set of data above with this method, which they did not use, and we found that the Pseudo-Newton algorithm is faster in convergence than the faster QLB algorithm. The lasso solution is

$$(-2.433, 0.735, 0.725, -2.035, -1.381, 0.018, -0.997).$$

The monotone convergence of the Psude-Newton is shown in Figure (4.4).

4.5 Review

We have presented the logistic regression model where the outcome variable is binary or dichotomous. In Example 4.4.1 we did an analysis of binary data, South African Heart Disease, to demonstrate the statistical use of logistic regression model also, we use forward selection procedure to find the best model.

We have investigated some of the methods in lasso logistic regression, such as the Quadratic Lower-Bound (QLB) Algorithm, and Pseudo-Newton algorithm. We implemented the R-cods for the both QLB and Pseudo-Newton algorithms. We tried to use the two methods in the case where $p \gg n$), but the algorithms did not work well because we obtained a singular matrix for x.

Chapter 5

Regression Discontinuity Analysis

5.1 Introduction

Regression discontinuity (RD) designs were first introduced by (Thistlethwaite and Campbell, 1960) and are recently receiving more attention. The recent monograph by Shadish et al. (2002) was reviewed favorably by Anderson-Cook (2005) in the prestigious statistical journal *Journal of the American Statistical Association* and a special issue of a major econometric journal was devoted to the topic of RD (Imbens and Lemieux, 2008). As noted by Anderson-Cook (2005), many statisticians feel that the only valid methodology for making scientific causal inferences are Randomized Designs. Undoubtedly randomized experimental designs were one of the greatest scientific and technological inventions of the 20th century but much of the data available to social scientists as well as in finance, medicine and epidemiology are observational. Great philosophical progress has been made by Pearl (2000) on elucidating the nature of causality and describing and qualifying when causal inferences may be drawn from observation data. Application papers using RD have appeared in the *Journal of the American Statistical Association* and theoretical contributions to RD have appears in the *Annals of Statistics*, so RD can most definitely be considered mainstream statistics even it is not widely taught in Statistics Departments.

There were several reasons that the regression discontinuity design did receive little attention until recently. The design was not well understood, because the program and comparison groups are chosen differently from the other multiple group designs.

If the program group consists of high pretest scorers, the comparison group will be lower scorers, and vice versa. Another difficulty is that it is not easy to implement the RD design. There may be political or social difficulties that threaten the correct execution of the design. Finally, the statistical analysis of the regression discontinuity design is not always easy since it may be difficult to select the appropriate statistical model. Despite all the reasons we discussed, the RD design is important and widely used as an applied social research technique because of it ability to accomplish the political and social goals of allocating scarce resources to those that need them most (van der Klaauw, 2008).

There are different structures of the regression discontinuity design. Two important RD's are the Two-Group Pretest-Posttest Design and the Nonequivalent Group Design. In their most basic form, we will need a statistical model that includes a term for the pretests, posttest, and a dummy-coded variable to represent the assignment status of the person in the study such as received the program or did not received the program (Lee, 2008). In typical regression analysis, we usually concerned about the variables that we should included them in the model, and the nature of the functional form. However, the main problem in regression discontinuity analysis is model specification because the measure of included are determined by the design (Bloob et al., 2005). The Regression-Discontinuity Design starts with a selection criterion that separates people into two groups on the basis of some measurement such as achievement then, test wither some invention or program has any impact on the outcome measure. This chapter consists of two sections.

Section One, introduces the basic regression-discontinuity design with it's assumptions. Moreover, we discuss the differences between the fuzzy and sharp regression-discontinuity design.

Section Two presents the statistical analysis of the regression-discontinuity design

(Imbens and Lemieux, 2008). In addition, we introduce the model specification, the curvilinearity problem, the analysis steps, and the multiple cutoff points.

5.2 The Basic Regression-Discontinuity Design

In regression discontinuity design all persons are assigned to program or comparison group according to a cutoff score on the preprogram measure **pretest**. Thus, all persons scoring on one side of the cutoff score are assign to one side group for example, program group while persons who scoring on the other side are assigned to the other group such as comparison group. This is the case in medicine, for example, this design is applicable for assessing the effect of giving a new surgical procedure to all patients who exceed a certain score on a presurgical measure of severity of illness. There are several assumption which must be made in order for the analytic model to be appropriate:

- The cutoff criterion must be followed.

  The regression discontinuity design assume a perfect assignment no misassignment relative to the cutoff unless it is know to be random. Persons scoring on one side of the cutoff score are assign to one group not be placed in the other. Misassignment can arise from a different reasons. First, political. Persons who are able to misassign them into a desirable group or out of an undesirable one. Second, administrative error, administrators do not adhere to the cutoff criterion.

  Campbell (1969), termed the misassignment relative to the cutoff score by ***fuzzy*** regression-discontinuity. This types of misassignment unless is know to be ran-

dom yield to biased estimates of the effect of the program (Goldberger, 1972). Analysis in the presence of fuzzy cutoff points will result in two problems:

- Effects the significant of the test.

  The two samples of disturbance terms for the study cannot considered to be random samples. Therefore, we loss the justification for applying statistical theory.

- The overlap between the groups can gives a different result for the effect of the treatment.

The term ***sharp*** regression-discontinuity design is used when there is no mis-assignment.

- One Factor.

  The second assumption is that only one factor would result in a discontinuity in the pre-post relationship at the cutoff point which is the program effect. If there are more than one factor this will lead to misunderstand which factor is the result of an observed effect. It may be due to the program effect or partially to some other factor consequently, the result may not be valid.

- Continuous Pretest Distribution.

  The cutoff point determined the division between the two groups: control and program which both must come from a single continuous pretest distribution. In some cases, there are intact groups such as two groups of patients or students from two different geographical regions which coincidental divide on some measure so as to imply some cutoff. This kind of groups could reflect a selection bias

because they are different naturally at the cutoff prior to the program which introduce some discontinuities at this point.

- In both program and comparison groups, there must be a sufficient number of points in order to obtain the regression lines.

- The true pre-post distribution must be describable as polynomial in x. The model is misspecified and the estimates of the program effect are not accurate if the true model is not polynomial for example, exponential or logarithmic. Therefor, this data should be transformed to a polynomial distribution prior to analysis however, the model will be more problematic to interpret.

5.2.1 The statistical Analysis of the Regression-Discontinuity Design

The analytic model has been described in (William and Trochim, 1984) presented as follow:

$$y_i = \alpha_0 + \alpha_1 \tilde{x_i} + \alpha_2 z_i + \alpha_3 \tilde{x_i} z_i + ... + \alpha_{n-1} \tilde{x_i}^k z_i + \alpha_n \tilde{x_i}^k z_i + e_i, \qquad (5.1)$$

where:

x_i is the preprogram measure, pretest for individual i.

y_i is the post program measure, posttest for individual i.

$$\tilde{x_i} = x_i - x_{ct}, \qquad (5.2)$$

where x_{ct} is the value of the cutoff, k is the degree of the polynomial, α_0 is the intercept at cutoff for comparison group, α_1 is the linear slope parameter, α_2 is the estimate of the main effect of the program, α_n parameter for the kth polynomial or interaction terms, e_i is the random error.

The hypothesis of interest is:

$$H_0 : \alpha_2 = 0 \quad v.s \quad H_1 : \alpha_2 \neq 0.$$

The mechanics of this model works as in the example follow:

- Suppose that the true function is linear.

  The model can be written as follow:

$$y_i = \alpha_0 + \alpha_1 x_i + \alpha_2 z_i. \tag{5.3}$$

  The comparison group line would be:

$$y_i = \alpha_0 + \alpha_1 \tilde{x_i}, \tag{5.4}$$

  the estimate where this line intersect the cutoff in $y_{ct} = \alpha_0$ and $\tilde{x_i} = 0$ at the cutoff point.

  The program group line would be:

$$y_i = \alpha_0 + \alpha_1 \tilde{x_i} + \alpha_2, \tag{5.5}$$

  the estimate where this line intersects the cutoff in $y_p = \alpha_0 + \alpha_2$.

  The main effect is defined as the vertical difference between the lines at the cutoff.

$$y_p - y_{ct} = (\alpha_0 + \alpha_2) - \alpha_0 = \alpha_2, \tag{5.6}$$

  therefore, the main program effect would be α_2.

- Suppose that the true function includes both a main and interaction effect.

  The model can be written as follow:

$$y_i = \alpha_0 + \alpha_1 x_i + \alpha_2 z_i + \alpha_3 x_i z_i. \tag{5.7}$$

  The comparison group line would be:

$$y_i = \alpha_0 + \alpha_1 \tilde{x_i}, \tag{5.8}$$

  the estimate where this line intersect the cutoff in $y_{ct} = \alpha_0$ and $\tilde{x_i} = 0$ at the cutoff point.

  The program group line would be:

$$y_i = \alpha_0 + \alpha_1 \tilde{x_i} + \alpha_2 + \alpha_3 \tilde{x_i}, \tag{5.9}$$

  the estimate where this line intersects the cutoff in $y_p = \alpha_0 + \alpha_2 + \alpha_3 \tilde{x_i}$.

  The difference between these lines at the cutoff is:

$$y_p - y_{ct} = (\alpha_0 + \alpha_2 + \alpha_3 \tilde{x_i}) - \alpha_0$$

$$y_p - y_{ct} = \alpha_2 + \alpha_3 \tilde{x_i}, \tag{5.10}$$

  where α_2 is the main effect of the program and α_3 is the difference in slopes between the lines of the two groups. α_0 is where the comparison group line hits the cutoff. $\alpha_2 + \alpha_3$ is the program group cutoff intercept.

5.2.2 Model Specification

The main goal in regression-discontinuity is to obtain an unbiased and statistically efficient estimate of program effect. To achieve this goal, the subset of variables that

is selected from the general model must describes the true pre-post relationship accurately. There are three types of specifications:

- Exactly specify the true model. Specify the model where there are no unnecessary terms in it. In practices exact specification is hard to obtain. When we exactly specify the true model we get an unbiased and sufficient estimates of the treatment effect as shown in the example follow:

  If the true model is:

  $$y_i = \alpha_0 + \alpha_1 x_i + \alpha_2 z_i + \alpha_3 x_i z_i$$

  We fit the model :

  $$y_i = \alpha_0 + \alpha_1 x_i + \alpha_2 z_i + \alpha_3 x_i z_i + e_i$$

  We obtain an unbiased and efficient estimate.

- Overspecify the true model. Specify the model where there are unnecessary, extra, terms. In this case, our estimate will be unbiased because we included all the necessary terms, but not efficient because of the extra terms. Thus, it is hard to us to see the treatment effect even if it exists as shown in the example follow:

  If the true model is:

  $$y_i = \alpha_0 + \alpha_1 x_i + \alpha_2 z_i$$

  We fit the model :

$$y_i = \alpha_0 + \alpha_1 x_i + \alpha_2 z_i + \alpha_3 x_i z_i + e_i$$

We obtain an unbiased ,but an efficient estimate.

- Underspecify the true model. Specify the model where we excluded some necessary terms as shown in the example follow:

  If the true model is:

$$y_i = \alpha_0 + \alpha_1 x_i + \alpha_2 z_i + \alpha_3 x_i z_i$$

  We fit the model :

$$y_i = \alpha_0 + \alpha_1 x_i + \alpha_2 z_i + e_i$$

  We obtain both biased and inefficient estimate.

In Practice, it is difficult to specify the exact true model. Therefore, we prefer to overspecify the true model rather than underspecify. Then, in the analysis, gradually remove higher-order terms until the model diagnostic for example, residual plots indicate that the model fits poorly.

5.2.3 The Curvilinearity Problem

Misspecify the statistical model, is the main problem in analyzing data from the regression-discontinuity design because this will lead to get a biased estimates of the treatment effect. If the true pre-post relationship is curviline and we fit a straightline model to the data the conclusion of the treatment effect will be inaccurate, made a

difference when it did not. Suppose that there is no jump or discontinuity in the data when we plot the curveline. However, there is a cutoff value.

- Force the slop of the program group and the slop of the pretest group to be equal. Use the model with out any interaction between the program and pretest group. The straight line model suggests that there is a jump at the cutoff however,there is no jump in the true data .

- Allowing the slop of the program group and the pretest group to be differ. Use the model with interaction between the program and the pretest groups. The straight line model suggests that there is a jump at the cutoff, but the pseudo-effect in this case is smaller than the first case.

5.2.4 The Analysis Steps

The analysis of the basic regression-discontinuity design consist of the following steps:

- Ordered the data a cording to the pretest values (x_i).

- Find the cutoff point for example the median of the pretest (x_i).

- Assign the dummy-coded variable z where,
  z_i=1 receiving the treatment (program group).
  z_i=0 comparison group.

- Subtract the cutoff value from each pretest score in order to set the intercept equal to the cutoff value. The modified pretest ($\tilde{x_i}$) will be equal to 0 at the

cutoff value making the cutoff the intercept point because the intercept is by definition the value of y when $x = 0$.

$$\tilde{x_i} = x_i - x_{ct}, \tag{5.11}$$

where x_{ct} denoted the cutoff value.

- Plot the pre-post relationship to determined if there is a discontinuity in the data. Also, count the number of times the distribution *flexes* or *bends*.

- Use the rule of thumb. Go two orders of polynomial higher than the numbers of flexion points that you obtained in step five. For example, if the bivariate distribution shows no ***flexes*** or ***bends*** points, use second-order $(0 + 2)$ polynomial transformations. The first-order polynomial already exist in the model (x). You have to create the second-order polynomial by squaring x to obtain x^2 also, you have to create the interaction term by multiplying the polynomial by z.

- Use any multiple regression program. If there is a discontinuity at the cutoff, it would be estimated by the coefficient associated with the z term. Test the significance of the coefficient by a standard t-test. If the coefficient is highly significant, conclude that there is a program effect.

- Remove the an unnecessary terms from the overspecified model. Examine the significant of the coefficients of the highest-order term in the current model, the goodness-of-fit measure, and the pattern of residuals. If these measures indicate not a significant coefficients and poorly fitting model, drop the highest-order term and repeat the analysis until all the coefficients will be significant. The

final model may still contain unnecessary terms, but they are less than before and their efficiency will be grater.

5.3 Multiple Cutoff Points

Sometimes we need more than one cutoff value for assignment. In this situation it will be two programs group and one control group. This will be powerful tool when the program could applied to those who needed the most. The comparison condition to those who do not need it, and the other program to persons falling in between these two groups. The analysis of this design is similar to the case of one cutoff point except we need two assignment (z) variables. Assume a linear relationship between the pre-pos data and there is no interaction effect. The analysis mode would be

$$y_i = \alpha_0 + \alpha_1 x_i + \alpha_2 z_{1_i} + \alpha_3 z_{2_i} + e_i,$$

where all terms are as equation (4.1) except

z_{1_i}=1 if program 1; 0 otherwise, z_{2_i}=1 if program 2; 0 otherwise, α_2 is the different between program 1 and the comparison group, α_3 is the different between program 2 and the comparison group, and $\alpha_2 - \alpha_3$ is the different between program 1 and program 2.

In the cases of two cutoff points, only one cutoff point can be subtracted from the pretest to obtain ($\tilde{x_i}$) term. For example, if the cutoff point that separates the program 1 and program 2 groups is subtracted from the pretest value, then α_2 would estimate the vertical distance between the lines of the two groups at the subtracted cutoff value. In the multiple cutoff points the model specification problem will be more complicated than one cutoff point.

Table 5.1: Data for Hypothetical Sharp Regression Discontinuity Analysis

	school									
Title	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*
Rating	10	11	12	13	14	15	16	17	18	19
Treatment	1	1	1	1	1	0	0	0	0	0
Outcome	20	25	30	35	40	30	35	40	45	50

5.4 Illustrative Examples

5.4.1 The Impact of Receiving an Extra Counseling Program on Student Achievement

For a hypothetically data, suppose we want to study the impact of receiving an extra counseling program on student achievement. These data are taken from (Bloom and Kemple, 2005). The example assume that the out comes are a linear function of rating and the ratings are set so that funding is targeted to school with lowest rate. The district will award schools who their scoring bellow a given rate, while those above that score are not. Table (5.1) shows the case of sharp regression-discontinuity design where all schools are assigned perfectly without any exception to treatment group, receiving the counselor program, or comparison group by the rating index (cutoff point). Table (5.2) shows the case of fuzzy regression-discontinuity design where one school assigned to the treatment group by the rating index did not receive treatment and one school assigned to the comparison group by the rating index did receive treatment. This could occurs if consideration other than the rating index caused the district officials to choose some schools over others for the program. Figure (5.1) and Figure (5.2) show the different of the main effect between the two cases. The discontinuity at the cutoff point will be muted because of the misassignment.

Table 5.2: Data for Hypothetical Fuzzy Regression Discontinuity Analysis

	school									
Title	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*
Rating	10	11	12	13	14	15	16	17	18	19
Treatment	1	**0**	1	1	1	0	0	**1**	0	0
Outcome	20	25	30	35	40	27	32	37	42	47

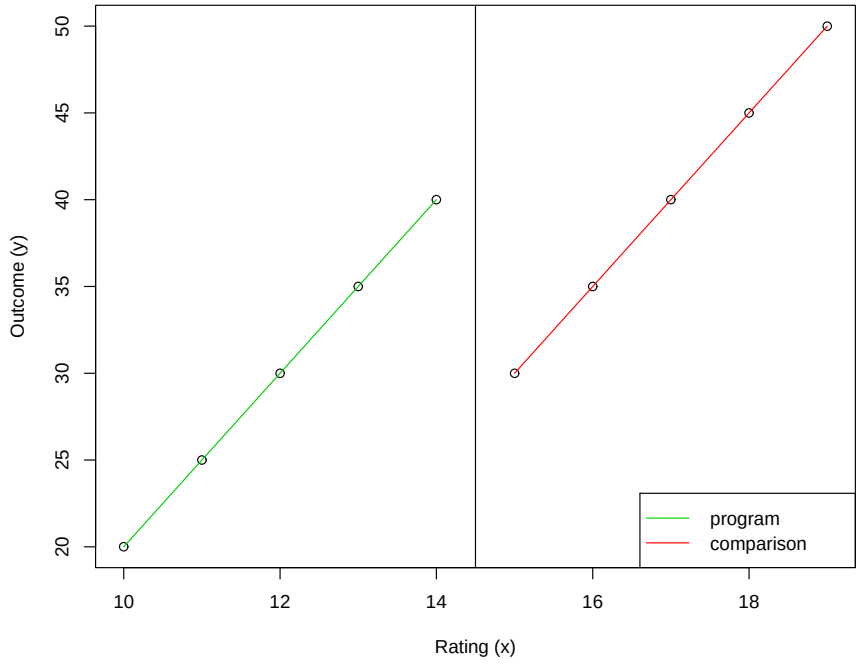

Figure 5.1: Change At The Margin With a Sharp Regression Discontinuity

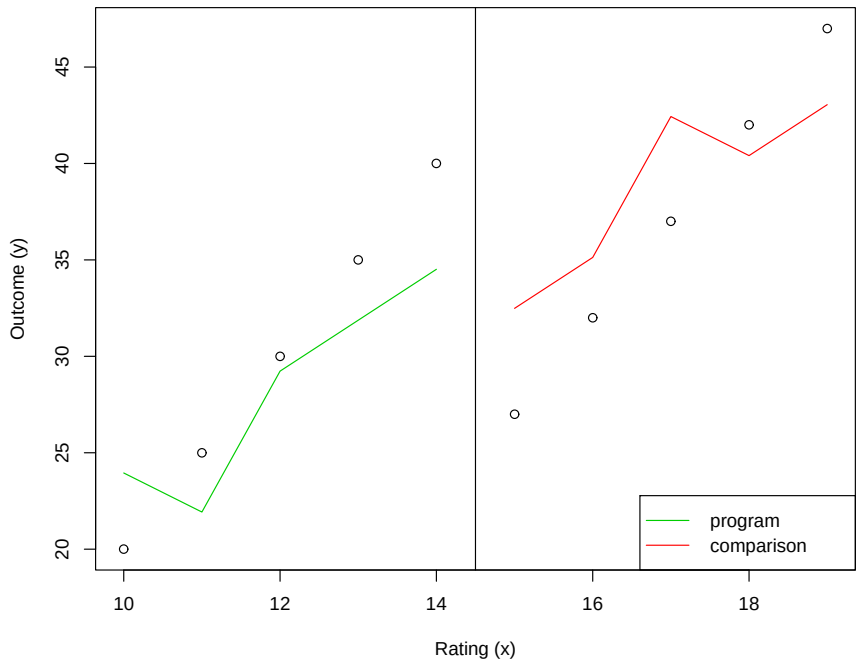

Figure 5.2: Change At The Margin With a Fuzzy Regression Discontinuity

5.4.2 Reducing Delinquency of Female Teenagers

As an artificial example of the regression-discontinuity design, we introduced the example in (Manly, 1992). The problem concerns girls with poor home conditions, so that extra counselling is introduced to the girls with poor conditions. Table(5.3) shows the scores of x and y where, x is the scale of the home condition where high scores indicate poor conditions that is the value of x above the median, y represents the delinquency which is determined for all girls after six months. Figure (5.3) shows the home condition scores (x), plotted against the observed delinquency scores. After ordering the data according to the value of x, and assigning the the value of z where $z = 1$ for the girls that received counselling the value of x is grater then the median of all x, and $z = 0$ for the other girls as shown in Table (5.4).

The model is simple linear regression,

$$y_i = \alpha_0 + \alpha_1 x_i + \alpha_2 z_i + \epsilon,$$

where: α_0 represent the constant term when there is no counselling, $\alpha_0 + \alpha_2$ represent the constant term when there is counselling.

The fitted regression equation is :

$$\hat{y} = 18.79 + 1.70x - 2.02z.$$

The estimated standard errors associated with the coefficients of x is (0.72) and with z is (1.65) by using the software R. The t-statistics for the coefficients of x is (2.39) and for z is (-1.22) each with 37 degrees of freedom. The coefficient of x is significant at the 5% level, but the coefficient of z is not significant at the same level. Therefore, we can conclude that the level of delinquency is related to home condition, but there is no enough evident to support the idea that counseling is reducing delinquency.

Table 5.3: The data of assess the effect of counseling on problem teenage girls (x=index of home condition, y=delinquency score after six months

	No Couns.		Counseling	
No	x	y	x	y
1	4.3	28.2	6.0	27.3
2	3.7	21.9	5.9	30.3
3	4.9	25.9	6.5	26.7
4	4.1	23.0	6	25.4
5	4.4	31.1	6.1	24.7
6	5.0	30.3	6.7	30.9
7	5.1	28.8	7.3	25.7
8	4.2	23.2	6	20.3
9	4.5	26.6	6.2	25.8
10	2.5	26.8	5.1	26.9
11	3.0	25.4	5.3	31.6
12	3.6	20.9	5.6	23.1
13	3.7	24.6	5.7	28.0
14	3.8	24.4	5.9	27.5
15	4.5	29.9	6.2	27.7
16	3.3	21.6	5.4	23.2
17	3.3	24.0	5.4	25.0
18	4.7	28.3	6.3	30.6
19	4.7	27.1	6.4	31.3
20	5.0	24.1	6.9	30.0

Table 5.4: The data of assess the effect of counseling on problem teenage girls after ordered x and assigned the value of z

	No Counseling			Counseling		
No	x	y	z	x	y	z
1	2.5	26.8	0	5.1	26.9	1
2	3.0	25.4	0	5.3	31.6	1
3	3.3	21.6	0	5.4	23.2	1
4	3.3	24.0	0	5.4	25.0	1
5	3.6	20.9	0	5.6	23.1	1
6	3.7	24.6	0	5.7	28.0	1
7	3.7	21.9	0	5.9	30.3	1
8	3.8	24.4	0	5.9	27.5	1
9	4.1	23.0	0	6	25.4	1
10	4.2	23.2	0	6	20.3	1
11	4.3	28.2	0	6.0	27.3	1
12	4.4	31.1	0	6.1	24.7	1
13	4.5	26.6	0	6.2	25.8	1
14	4.5	29.9	0	6.2	27.7	1
15	4.7	28.3	0	6.3	30.6	1
16	4.7	27.1	0	6.4	31.3	1
17	4.9	25.9	0	6.5	26.7	1
18	5.0	30.3	0	6.7	30.9	1
19	5.0	24.1	0	6.9	30.0	1
20	5.1	28.8	0	7.3	25.7	1

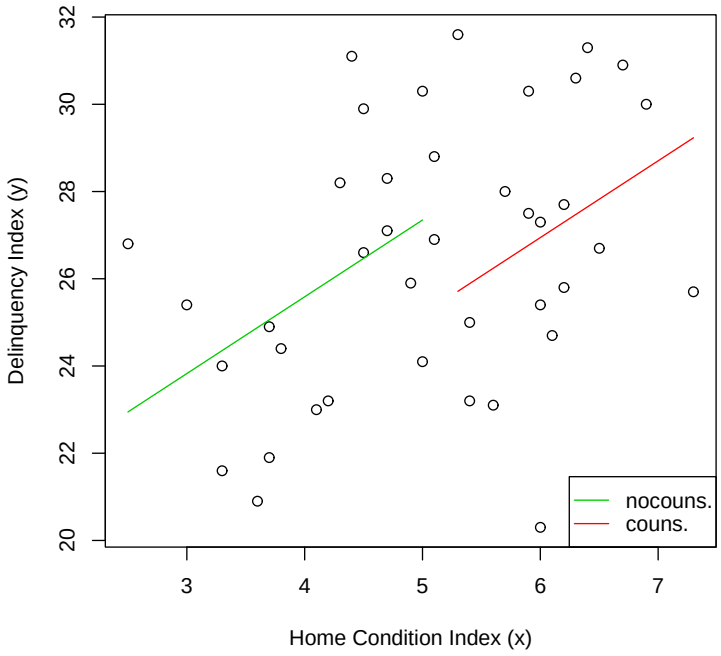

Figure 5.3: Delinquency scores plotted against home conditions for problem teenage girls. The lines shown are the expected frequencies from the fitted regression equation

5.5 Review

We have introduced the regression-discontinuity design and its assumption to build an appropriate model. In addition, we introduced the misspecification, and the curvilinearity problems. The regression-discontinuity design is useful to study any program or procedure that is given out on the basis of need or deserve.

Model misspecification is the most important problem in analyzing data from RD design because it leads to wrong conclusion about the treatments effect. If a linear relation is assumed, the result would threatened by possibility of nonlinearities. In some situations, one could use more than one cutoff value for assignment however, the model specification problem will be more complicated.

REFERENCES

A. Agresti. *An Introduction to Categorical Data Analysis*. John Wiley and Sons,Inc, 1996.

H. Akaike. A new look at the statistical model identification. *IEEE Trans. Automatic Control*, 19(6):716–723, 1974. ISSN 0018-9286, 1558-2523 (online). System identification and time-series analysis.

C. M. Anderson-Cook. Experimental and quasi-experimental designs for generalized causal inference. william r. shadish, thomas d. cook, and donald t. campbell. *Journal of the American Statistical Association*, 100:708–708, June 2005. URL `http://ideas.repec.org/a/bes/jnlasa/v100y2005p708-708.html`.

H. Bloob, J. Kemple, B. Games, and R. Jacob. Using regression discontinuity analysis to measure the impacts of reading first. Montreal Canada, 2005. The annual conference of the American Education Research Association.

H. S. Bloom and J. Kemple. Using regression discontinuity analysis to measure the impacts of reading first. *The Annual Conference of the American Educational Research Association*, 2005.

D. Bohning and B. Lindsay. Monotonicity of quadratic-approximation algorithms. *Annals of the Institute of Statistical Mathematics*, 40:641–663, 1988.

Campbell. Reforms as experiments. *American Psychologist*, 24:409–429, 1969.

J. Chen and Z. Chen. Extended bayesian information criteria for model selection with large model space. *Biometrika*, 95:579–771, 2008.

J. Chen and Z. Chen. Tournament screening cum ebic for feature selection with high-dimensional feature spaces. *Science in China Series A-Mathematics*, 52:1327–1341, 2009.

P. Craven and G. Wahba. Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation references. *Numerische Mathematik*, 31:377–403, 1979.

A. DM. Mean square error of prediction as a criterion for selecting variables. *Technometrics*, 13:459–475, 1971.

J. Dobson. *An Introduction to Generalized Linear Models*. Chapman and Hall/CRC, 2nd edition, 2002.

T. Dongsheng and J. Shao. *The Jackknife and Bootstrap*. Springer, 1995.

B. Efron. Bootstrap methods: Another look at the jackknife. *Journal of the American Statistical Association*, 78:316–333, 1983.

B. Efron. *The jackknife, the Bootstrap, and other Reaampling Plans*. Philadelphia: SIAM, 1982.

B. Efron. Estimating the error rate of a prediction rule: Improvement on cross-validation. *Thae Annals of statistics*, 78:1–26, 1979.

B. Efron, T. Hastie, I. Johnstone, and J. Tibshirani. Least angle regression. *The Annals of Statistics*, 32:407–499, 2004.

J. Fisher. Homicide in detroit: The role of firearms. *Criminology*, 14:387–400, 1976.

G. Furnival. All possible regressions with less computation. *Technometrics*, 13:403–408, 1971.

G. Furnival and R. Wilson. Regression by leaps and bounds. *Technometrics*, 16(4): 499–511, 1974.

A. Goldberger. Selection bias in evaluating treatment effects: Some formal illustrations. *Madison: University of Wisconsin, Institute for Research on Poverty*, 24 (123–172), 1972.

R. Gunst and Mason. *Regression analysis and its Application: A Data-Oriented Approach*. Marcel Dekker, 1980.

T. Hastie and R. Tibshirani. *Generalized Additive Models*. Chapman and Hall, New York, 1990.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer-Verlag, New York, 2nd edition, 2009.

R. Heavenrich, J. Murrell, and K. Hellman. Light duty automotive technology and fuel economy trends through. *U.S. Environmental Protection Agency*, EPA/AA/CTAB/91-02, 1991.

R. Hocking. *Methods and Applications of LinearModels: Regression and the Analysis of Variance*. John Wiley and Sons, 2nd edition, 2003.

R. Hocking and R. Leslie. Selection of the best subset in regression analysis. *Technometrics*, 9:531–540, 1967.

A. Hoerl and R. Kennard. Ridge regression: Applications to nonorthogonal problems. *Technometrics*, 12:69–82, 1970.

Hosmer, DW., and Lemeshow. *Applied Logistic Regression*. John Wiley and Sons, USA, 2000.

X. Huo and X. Ni. Regressions by enhanced leaps-and-bounds via optimality tests. Technical report, Georgia Institute of Technology, 2006. URL http://www2.isye.gatech.edu/statistics/papers/06-05.pdf.

G. Imbens and T. Lemieux. Special issue editors' introduction: The regression discontinuity design–theory and applications. *Journal of Econometrics*, 142(2): 611 – 614, 2008. ISSN 0304-4076. doi: DOI: 10.1016/j.jeconom.2007.05.008. URL http://www.sciencedirect.com/science/article/B6VC0-4NT9GJ9-1/2/4365ce4e015f5b9be The regression discontinuity design: Theory and applications.

P. Lahiri. *Model Selection*. Instituate of Mathmatical Statistics, U.S.A, 1999.

A. Land and A. Doig. An automatic method for solving discrete programming problems. *Econometrica*, 28:497–520, 1960.

B. Lawrence. *Impact Analysis for Program Evaluation*. Sage Publication, Inc., California, 2nd edition, 1995.

C. Lawson and R. Hanson. *Solving Least Squares Problems*. Prentice Hall, 1974.

H. B. Lee. Analyzing data from a regression discontinuity study: A research note. *Journal of Research Methods and Methodological*, 2, 2008.

T. Lumley and A. Miller. Leaps: Regression subset selection. *R package version* 2.7, 2004.

C. L. Mallows. Some comments on C_p. *Technometrics*, 15:661–675, 1973.

B. F. J. Manly. *The Design and Analysis of Research Studies*. Cambridge University Press, 1992.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall/CRC, 2nd edition, 1989.

A. I. McLeod and C. Xu. Bestreg: Best subset linear regression. *Submitted for publication*, 2009.

J. Miller. *Subset Selection in Regression*. CRC Press, 2nd edition, 2002.

J. A. Morgan and J. F. Tatar. Calculation of the residual sum of squares for all possible regressions. *Technometrics*, 14:317–325, 1972.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, 2000.

J. Rousseauw, J. du Plessis, A. Benade, P. Jordaan, J. Kotze, P. Jooste, and J. Ferreira. Coronary risk factor screening in three rural communities. *South African Medical Journal*, 64(430–436), 1983.

G. Schwarz. Estimation the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

W. R. Shadish, T. D. Cook, and D. T. Campbell. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton-Mifflin, Boston, 2002. ISBN 0-395-61556-9.

J. Shao. Linear model selection by cross-validation. *Journal of the American Statistical Association*, 88(422), 1993.

J. Shao. An asympotic thery for linear model selection (with discussion). *Statistic Sinica*, 7:221–262, 1997.

J. Shao. Bootstrap model selection. *Journal of the American Statistical Association*, 91, 1996.

C. Stein. Estimation of a multivariate normal distribution. *Ann. Statist*, 9:1135–1151, 1981.

M. Stone. An asymptotic equivalence of chice of model by cross-validation and akaike's criterion. *J. Royal. Statist*, 39:44–47, 1977.

D. Thistlethwaite and D. Campbell. Regression-discontinuity analysis: An alternative to the ex post factor experiment. *Journal of Educational Psychology*, 51:309–317, 1960.

Tian, Tang, Fang, and Tan. Efficient methods for estimating constrained parameters with applications to lasso logistic regression. *Computational Statistics and Data Analysis*, 52(7):3528–3542, 2008.

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal. Statist*, 58 (1):267–288, 1996.

W. van der Klaauw. Regression-discontinuity analysis. *Federal Reserve Bank of New York*, 2008.

W. Venables and B. Ripley. *Modern Applied Statistics with S*. Springer, 4th edition, 2002.

M. William and K. Trochim. Regression-discontinuity analysis. *Web center for social research methods*, 2006.

M. William and K. Trochim. *Research Design for Program Evaluation the Regression-Discontinuity Approach*. Sage Publication, Inc., California, 1984.

C. Xu and A. McLeod. Another extended bayesian information criterion for model selection. *Submitted for publication*, 2009.

H. Zou, T. Hastie, and J. Tibshirani. On the degrees of freedom of the lasso. *The Annals of Statistics*, 35:2173–2192, 2007.