CENTRAL DOUBLE CROSS-VALIDATION FOR ESTIMATING PARAMETERS IN REGRESSION MODELS

CHYE ROU SHI

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Master of Science (Mathematics)

Faculty of Science
Universiti Teknologi Malaysia

JULY 2016

ACKNOWLEDGEMENT

I am ever grateful to God who always is beside me, and has guided me with wisdom and courage to finish this thesis. This thesis is the result of three years of work whereby I have been accompanied and supported by many people. It is a pleasure to have this opportunity to express my gratitude to all of them.

First and foremost, words cannot express my gratitude to my supervisor, Assoc. Prof. Dr. Robiah Adnan for her encouragement, patience, and critical reviews of this manuscript. Without her guidance and help, I could never accomplish this difficult task. Being your student is a valuable experience for my future career, and you are my role model in teaching and consulting work. I am really glad that I have come to get to know Assoc. Prof. Dr. Robiah Adnan in my life.

I would like to thank all the faculty members at the Department of Science who have made my time at UTM more productive than I could have ever hoped for. I appreciate them for providing me with a friendly atmosphere and the resources to complete the program.

I owe my loving thanks to my father, mother and family, who have given me unlimited support and love, not just during my graduate study but for my whole life. Without their love, dedication and sacrifice, I would not have made it this far. Last, but definitely not least, many thanks are due to everyone who helped me during my time at Universiti Teknologi Malaysia.

ABSTRACT

The ridge regression, lasso, elastic net, forward stagewise regression and the least angle regression require a solution path and tuning parameter, λ, to estimate the coefficient vector. Therefore, it is crucial to find the ideal λ. Cross-validation (CV) is the most widely utilized method for choosing the ideal tuning parameter from the solution path. CV is essentially the breaking down of the original sample into two parts. One part is used to develop the regression equation. The regression equation is then applied to the other part to evaluate the risk of every model. Consequently, the final model is the model with smallest estimated risk. However, CV does not provide consistent results because it has overfitting and underfitting effects during the model selection. In the present study, a new method for estimating parameter in best-subset regression called central double cross-validation (CDCV) is proposed. In this method, the CV is run twice with different number of folds. Therefore, CDCV maximizes the usage of available data, enhances the model selection performance and builds a new stable CV curve. The final model with an error of less than π standard error above the smallest CV error is chosen. The CDCV was compared to existing CV methods in determining the correct model via a simulation study with different sample size and correlation settings. Simulation study indicates that the proposed CDCV method has the highest percentage of obtaining the right model and the lowest Bayesian information criterion (BIC) value across multiple simulated study settings. The results showed that, CDCV has the ability to select the right model correctly and prevent the model from underfitting and overfitting. Therefore, CDCV is recommended as a good alternative to the existing methods in the simulation settings.

ABSTRAK

Regresi batasan, penjerat, jaringan kenyal, regresi berperingkat ke hadapan dan regresi sudut terkecil memerlukan lintasan penyelesaian dan parameterpenalaan, λ, untuk menganggarkan koefisien vektor regresi.Oleh itu, pencarian parameter penalaan λ yang ungguladalah sangat kritikal. Pengesahan silang(CV) adalah kaedah yang paling meluas digunakan untuk memilih parameter penalaan unggul daripada lintasan penyelesaian. Pada asasnya CV adalah pemecahan data kepada dua bahagian. Satu bahagian digunakan untuk membangunkan persamaan regresi. Persamaan regresi itu kemudian diaplikasikan pada satu lagi bahagian bagi menilai risiko pada setiap model. Kerananya, model terakhir adalah model yang mempunyai anggaran risiko yang terkecil. Walau bagaimanapun, CV tidak menghasilkan keputusan yang konsisten kerana ia mempunyai kesan suaian-terlebih dan suaian-terkurang semasa pemilihan model. Dalam kajian ini, satu kaedah baharu untuk menganggarkan parameter dalam regresi subset terbaik yang dipanggil pengesahan silang gandadua pusat(CDCV) dicadangkan.Dalamkaedah ini, CV dijalankan dua kali dengan bilanganlipatan yang berbeza. Oleh sebab itu, CDVC akan memaksimumkan penggunaan data yang sedia ada, meningkatkan prestasi pilihan model dan membina keluk CV stabil yang baharu. Model terakhir yang mana ralatnyaadalah kurang daripada π ralat piawai di atas ralat CV yang paling kecil akan dipilih. CDCVtelah dibandingkandengan kaedah CV sedia adauntuk menentukan model yang betul melalui kajian simulasi dengan saiz sampel dan tetapan korelasi yang berbeza. Kajian simulasi menunjukkan bahawa kaedah CDCV yang dicadangkan mempunyai peratusan tertinggi dalam mendapatkan model yang betul dan kriteria maklumat Bayesan(BIC) yang terendah, merentasi pelbagai tetapan dalam kajian simulasi. Hasil kajian menunjukkan bahawa CDCV boleh memilih model yang tepat dengan betul dan mengatasi suaian-terlebih dan suaian-terkurang. Oleh sebab itu CDCV disyorkan sebagai satu alternatif yang baik kepada kaedah yang sedia ada dalam tetapankajian simulasi yang dijalankan.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Introduction and Background of the Problem

Multiple regression is frequently used to investigate the relationship between the predictor variables and the response variable. There are many types of regression methods for example, Ordinary least squares (OLS) regression, best-subset regression, Forward selection, Backward elimination, Stepwise, Ridge regression, Bridge regression, Garotte, Lasso, LARS, Pathseeker and Elastic Net.

Shrinkage method such as Ridge regression, Lasso and Elastic Net require a solution path to estimate the coefficient vector. Since the tuning parameter λ is utilized to select one estimator from the solution path, finding the ideal tuning parameter λ is very important.

There are numerous approaches such as Bayesian information criterion (BIC), Akaike information criterion (AIC), R^2_{Adj} and C_p for finding the ideal tuning parameter λ. However, the Cross-validation (CV) is the most broadly utilized method due to its simplicity and its universality. In particular, the CV splits the original data into multiple parts and develop the regression equation using one part of the data. Then the regression equation is applied to the remaining data to evaluate the risk of every model. Consequently the final model is the model with the smallest estimated risk. CV can be applied in almost any algorithm for any structure. For example, the R^2_{Adj}, C_p, and BIC statistics require information of the quantity of parameter p, while CV does not. In addition CV not only tends to give a comparable solution for simple problems but it is also a good solution for complex situations, such as when the quantity of parameter p is unknown. For instance, the addition of a constant to any of the measures would not change the resulting chosen model. However, for many adaptive nonlinear techniques, it is difficult to estimate the effective number of parameters. Thus, the method like AIC is infeasible and leaves us with CV as a more favourable method of choice. Secondly, benefit of CV is its robustness. The C_p and BIC require a proper working model to estimate σ^2. Otherwise, it has a suboptimal performance [1]. While CV doesn't use σ^2 to build the tuning curve, if the models being assessed are far from correct, CV still works well [2, 3]. Thirdly, AIC prefers to pick models which are overfitted as $n \rightarrow \infty$, whereas BIC is more likely to pick the too parsimonious models if sample size is too big.

Many types of CV utilize the similar idea of creating separate folds to choose the tuning parameter. However CV uses only a subset of the observations for training in each round, so CV typically delivers pessimistic accuracy estimations. Consequently, the quantity of folds affect the bias and variance in CV. For example, utilizing a small quantity of folds leads to smaller variance but large bias, whereas, CV with larger quantity of folds, - makes the variance of the estimators larger, but bias smaller. Another weakness of CV is a large portion of the CV error curves are very flat over large ranges close to their minimum and this makes it more difficult to choose the perfect tuning parameter λ_{OPT}.

1.2 Statement of the Problem

CV is the most widely used method to select the optimal tuning parameter to estimate the coefficient vector, yet it also has some restrictions.

First, the CV does not provide consistent model selection results because CV has an overfitting and underfitting effect during the model selection. The overfitted model will frequently result in variances of estimated parameter that are more than the model which contains small quantity of predictor. Moreover, it is hard to maintain a model with too many predictor variables. Thus the underfitted model will remove the key predictor variables and will reduce the predictor power of the model that leads to biased estimates of the regression coefficients.

Next CV is normally utilized to validate every model, and then select the model that has the lowest CV error. It expects model with lower generalization error to have lower CV error. However, Ng [4] shows that the idea may not be right because lower expected generalization error may not always come from the model with lower CV error.

Moreover many of the CV curves are very flat over large ranges close to their lowest value. It is most difficult to choose the ideal λ_{OPT} for CV.

1.3 Objectives of the Study

The objectives of this research are:

1. To modify the existing CV in order to avoid overfitting and underfitting the model runs the CV twice with different number of folds and choose the ideal tuning parameter with new stable CV curve.
2. To simulate data with different sample sizes and correlation settings to investigate the effect of different CV methods.
3. To compare the proposed method against other CV methods in order to determining correct model.

1.4 Scope of the Study

Compare the proposed method, Central Double Cross-Validation (CDCV) with current methods: K-fold Cross-Validation (K-fold CV), One-standard-error rule of K-fold CV (K-fold CV(SE=1)) and Leave-n_v-out CV (CV(n_v)). Four principal assumptions which justify the utilization of multiple linear regression models are:

- linearity of the relationship between predictor variables and response variable
- independence of the errors (no serial correlation)
- homoscedasticity (constant variance) of the errors
- normality of the error distribution.

Simulation studies were done to evaluate the effect of multicollinearity on the estimation of the parameters at various levels of multicollinearity.

Simulation studies were done using statistical environment R. For the development of the current methods, the *bestglm* package was used. For the CDCV method, the *bestglm* package was used with modification of the algorithm.

1.5 Significance of the Study

Cross-validation (CV) is the most broadly utilized method in order to choose the ideal tuning parameter from solution path. The aim of this study is to modify the current CV method to avoid overfitting and underfitting problem during the model selection. Thus, we proposed a new method for estimating parameter in best-subset regression called Central Double Cross-Validation (CDCV). Simulation study indicates that the proposed method, CDCV, has the highest percentage of obtaining the right model across multiple simulated study settings. Therefore CDCV can help analysts to select the correct regression models with a reasonable number of predictor variables.

1.6 Summary and Outline of the Study

Chapter 2 covers literature review on published work done recently concerning the CV. The explanation and methodology of current CV methods and the proposed method are shown in Chapter 3. Chapter 4 discusses the structure of simulation studies and analyzes the performance of current CV and CDCV in terms of percentage of determining the right model and lower BIC value via a simulation study. Next in Chapter 4 compares and discusses the performance measures used in this research. Finally, in chapter 6 discussion, conclusion and the suggestion for further research are given.

1.7 Linear Regression

1.7.1 Introduction

This section covers the basic concept of linear regression. Section 1.7.2 explains the effect of overfitting and underfitting during model selection. This section also explains how number of variables in the model affect the model's predictive ability, bias and variance. Section 1.7.3 introduces some methods that are used to select model. These methods can be divided into three classes: goodness-of-fit measures, estimating distributional discrepancies and criteria based on prediction error.

1.7.1.1 Ordinary Least Squares

The Ordinary Least Squares (OLS) is a widely accepted method generally utilized to estimate parameters to fit data. Analysts normally utilized OLS to build a linear regression model [5].

The objective of doing regression is to get the estimate of the regression coefficients $\boldsymbol{\beta}$ for the equation of linear regression model as in equation (2.1).

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1.1}$$

where matrices and vectors are represented by bold uppercase letters. Response vector, $\mathbf{Y} \in \mathbb{R}^{n \times 1}$; matrix of predictors, $\mathbf{X} \in \mathbb{R}^{n \times p}$; error term, $\boldsymbol{\varepsilon} \in \mathbb{R}^{n \times 1}$. Sample size is denoted by n and number of parameters is denoted by p.

Estimate $\boldsymbol{\beta}$ by minimizing the sum of the squared differences between the response values and those predicted by the equation with the least squares criterion:

$$\min_{\hat{\boldsymbol{\beta}}} \sum_{i=1}^{n} \mathbf{e}_i^2 \tag{1.2}$$

and,

$$\mathbf{e}_i = \mathbf{Y}_i - \mathbf{X}_i'\boldsymbol{\beta}, \quad i = 1, 2, ..., n \tag{1.3}$$

where residual is denoted by $\mathbf{e}_i$ and transpose matrix of $\mathbf{X}$ is denoted by $\mathbf{X}_i'$.

Putting it in another way, the coefficients for the least squares regression are chosen to minimize the sum of squares of the residuals. The model function is given by:

$$\begin{aligned}
F(\boldsymbol{\beta}) &= \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} \\
&= (\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) \\
&= \mathbf{Y}'\mathbf{Y}-\boldsymbol{\beta}'\mathbf{X}'\mathbf{Y}-\mathbf{Y}'\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\
&= \mathbf{Y}'\mathbf{Y}+\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}-2\boldsymbol{\beta}'\mathbf{X}'\mathbf{Y} \qquad (1.4)
\end{aligned}$$

Differentiate with respect to $\boldsymbol{\beta}$ to minimize $F(\boldsymbol{\beta})$ and sets it to zero:

$$\frac{\partial F(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\beta}} = 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} - 2\mathbf{X}'\mathbf{Y} = 0$$

The normal equations of the least squares is,

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y}$$

If the $(\mathbf{X}'\mathbf{X})^{-1}$ exists, then $\hat{\boldsymbol{\beta}}$ is called the Ordinary Least Squares estimate of $\boldsymbol{\beta}$,

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\boldsymbol{\beta}}_0 \\ \hat{\boldsymbol{\beta}}_1 \\ \vdots \\ \hat{\boldsymbol{\beta}}_p \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \qquad (1.5)$$

This model is not only used to investigate and model the relationship of the response and predictor variables but also used to identify significant predictor variables. Therefore this model can also be used for predicting the value of the response variable.

1.7.1.2 Analysis of Variance (ANOVA) Approach to Regression Analysis

The ANOVA methodology is based on the partitioning of sums of squares and degree of freedom (*df*) associated with the response variable Y [5]. The variation is customarily measured in terms of the deviations of the observed response value, Y_i and their mean $\bar{Y}$:

$$Y_i - \bar{Y} \tag{1.6}$$

The total variation (*SSTO*), is the sum of squared deviation and defined as:

$$SSTO = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \tag{1.7}$$

The *SSTO* increases when the variance among the Y_i increases. This is because $SSTO$ is the total sum of squares in the response about the mean. The *SSTO* = 0 when all Y_i are the same. Once the predictor variable X is utilized, the variation indicating the uncertainty related to the variable Y is equivalent to Y_i data around the fitted regression line:

$$Y_i - \hat{Y}_i \tag{1.8}$$

where $\hat{Y}_i$ is the i^{th} fitted response.

The measure of variation in the Y_i observations that is present when the predictor variable X is taken into account is the sum of the squared deviations where the equation is defined in equation (1.8), which is the SSE.

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{1.9}$$

Next, SSE denotes the error sum of squares. The $SSE = 0$ when all observed response value fall on the fitted regression line. As the variation of the Y_i around the fitted regression line increases, the SSE also increases. The equation for the regression sum of squares (SSR) is:

$$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \tag{1.10}$$

Note that SSR is a sum of squared deviations, the deviations being:

$$\hat{Y}_i - \bar{Y} \tag{1.11}$$

Each deviation is basically the difference between the fitted value on the regression line and the mean of the fitted values $\bar{Y}$. If the slope of the regression line is equal to zero so that $\hat{Y}_i - \bar{Y} \equiv 0$, then $SSR = 0$. Else, SSR is greater than zero.

SSR may be seen as a measure of that part of the variability of the Y_i which is connected with the regression line. The SSR is in association with $SSTO$. When the SSR is large, the more imperative is the effect of the regression connection in representing the total variation in Y_i.

1.7.1.3 Stepwise Methods

There are two weaknesses of OLS.

Firstly, OLS usually has low bias yet large variance, shrinking or changing some coefficients to zero might improve the accuracy of prediction. Thus, sacrifice a little bit of bias will lower the variance of the predicted values, and thus, enhance the prediction accuracy on the whole.

Subsequently, certain regression models possess a large quantity of predictor variables, where user frequently might be keen to establish a smaller model which show the greatest effects. As a result, users are willing to sacrifice some minor details to a bigger picture of the model.

Although there are a variety of approaches that can be used for model selection, the focus of this research is on best-subset regression.

Since the quantity of conceivable models, 2^{p-1}, which grows quickly with the increase in the number of predictors, assessing the majority of the conceivable choices can be an overwhelming process. Many automatic computer methodologies have been developed to simplify the task. In this thesis, the focus is on the best-subset regression. Efficient algorithms have been created in the best-subset according to a specified criterion which can identify the best subsets without requiring the fitting of all of the possible subset regression models. Indeed, these methods require the calculation of only small portion of all possible regression models. A case in point is, if the C_p criterion is to be employed and the five best subsets according to this criterion are to be identified, these algorithms search for five subsets of X variables with the smallest C_p value utilizing much less computational effort than running all of the possible subset regression models. These algorithms do not just give the best subsets according to the specified criterion, but also frequently recognize a few good subsets for each conceivable quantity of X variables in the model to give the user additional helpful information in final regression model.

1.7.2 Deletion of Predictor Variables

In exploratory observational studies, after the screening process, the model would still include a large number of predictor variables. Another problem is that the model that contains large quantity of predictor variables frequently has multicollinearity issue. Subsequently, the user will hope to lessen the number of predictor variables utilized in the final model [5]. A brief explanation is given below.

First, the model which contains a limited number of predictor variables may be less demanding to work with and miss some truly informative variables. Second it is hard to maintain a regression model with too many predictor variables. Last, if the model includes many highly correlated predictor variables, the sampling variation of the regression coefficients may increase, and thus decreases the model's descriptive capacities, and consequently increases the problem of round off errors and the performance of model to predict new observation becomes lower. In conclusion, the variances of the fitted values, $\sigma^2\hat{Y}_i$ will become bigger when more useless predictor variables stay in the model.

Subsequently, once the user has decided upon the functional form of the regression relations and whether any interaction terms are to be included, the following step in various exploratory observational studies is to identify a couple of good subsets of X variables for further study. The model should include the potential predictor variables in first-order form, higher-order form, and interaction terms.

The identification of good subsets is promisingly useful in deciding which predictor variables are to be included in the final regression model. The influence of suitable functional and interaction relationship for these variables generally comprise the main challenges in regression analysis.

How to determine the "best" model may depend on the user's purpose because application of regression model is very wide. For example, a descriptive use of a regression model will typically emphasize precise estimation of the regression coefficients, whereas a predictive use will focus on the prediction errors. Frequently, diverse subsets of the pool of potential predictor variables will best serve these varying purposes. Even for a given purpose, sometime we found that few subsets are also similarly "good" according to a given criterion, and the decision among these good subsets needs to be made on the basis of extra contemplations.

The identification of good subsets should be done with careful consideration. Removing the key predictor variables can reduce the predictor power of the model and bring about biased estimates of regression coefficients, mean responses, and model's predictive capability, and biased estimates of the error variance. The bias in these estimates is related to the fact that with observational data, the error terms in an underfitted regression model may reflect nonrandom effects of the predictor variables not incorporated in the regression model. It is imperative to exclude predictors which are sometimes called latent predictor variables.

Next, the overfitted model will frequently result in variances of estimated parameter that are more than the model which contain small quantity of predictor variables.

1.7.2.1 Curve Fitting

Gruber [6] pointed out that model which contains too many predictor variables has smaller bias and large variance. Due to the flexibility of the functional form being applied to fit the model, noise was also detected simultaneously with the data, resulting in unstable predictions. However, the model which contains a limited quantity predictor variables, has smaller variance yet large bias because it capture the curvatures of the true function f.

Their researches show that the number of variables in the model affect the model's predictive ability.

1.7.2.2 Prediction Error and Model Error

In the prediction problem, an initial data set (x_i, Y_i), $i = 1, ..., n$, consisting of $n(p+1)$- dimensional multivariate observations, comprised of a response Y_i, and a p-dimensional vector of predictor variables x_i. It is assumed that the first element of every x_i is 1, correspond to the constant term in the regression model.

Assume this data set which is occasionally named as training set is employ to predict the responses $Y_{0i}, i = 1, ..., m$, corresponding to m new vectors X_{0i}. The main matter is to estimate the mean μ_{0i} of the response Y_{0i}.

Let $\mathbf{Y}' = (Y_1, ..., Y_n)$ and $\mathbf{Y}_0' = (Y_{01}, ..., Y_{0m})$. $\mathbf{Y}$ and $\mathbf{Y}_0$ assumed to have covariance matrices $\sigma^2\mathbf{I}_n$ and $\sigma^2\mathbf{I}_m$ (where $\mathbf{I}_n$ and $\mathbf{I}_m$ are identity matrix), respectively, and that $\mathbf{Y}$ and $\mathbf{Y}_0$ are independent with the same probability structure; if $x_i = x_{j0}$, then $E[Y_i] = E[Y_{0j}]$. Also, let

$$X = \begin{bmatrix} x_1' \\ \vdots \\ x_n' \end{bmatrix} \text{and} \quad X_0 = \begin{bmatrix} x_{01}' \\ \vdots \\ x_{0m}' \end{bmatrix}$$

Estimate $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ from the training set. The least squares predictor of $\mathbf{Y}_0$ is $\mathbf{X}_0\hat{\boldsymbol{\beta}}$. The sum of the squared of prediction errors is

$$\sum_{i=1}^{m}(Y_{0i} - x_{0i}'\hat{\beta})^2 = ||\mathbf{Y}_0 - \mathbf{X}_0\hat{\boldsymbol{\beta}}||^2, \tag{1.12}$$

which can be written as

$$\begin{aligned} ||\mathbf{Y}_0 - \boldsymbol{\mu}_0 + \boldsymbol{\mu}_0 - \mathbf{X}_0\hat{\boldsymbol{\beta}}||^2 = & \; ||\mathbf{Y}_0 - \boldsymbol{\mu}_0||^2 + ||\boldsymbol{\mu}_0 - \mathbf{X}_0\hat{\boldsymbol{\beta}}||^2 + \\ & 2(\mathbf{Y}_0 - \boldsymbol{\mu}_0)'(\boldsymbol{\mu}_0 - \mathbf{X}_0\hat{\boldsymbol{\beta}}) \end{aligned} \tag{1.13}$$

If take expectations over the new data only, the quantity obtained is known as prediction error (PE):

$$\begin{aligned} PE &= E_{\mathbf{Y}_0}[||\mathbf{Y}_0 - \mathbf{X}_0\hat{\boldsymbol{\beta}}||^2] \\ &= E_{\mathbf{Y}_0}[||\mathbf{Y}_0 - \boldsymbol{\mu}_0||^2] + ||\boldsymbol{\mu}_0 - \mathbf{X}_0\hat{\boldsymbol{\beta}}||^2 \\ &= E_{\mathbf{Y}_0}\left[\sum_{i=1}^{m}(\mathbf{Y}_{i0} - \boldsymbol{\mu}_{i0})^2\right] + ||\boldsymbol{\mu}_0 - \mathbf{X}_0\hat{\boldsymbol{\beta}}||^2 \\ &= m\sigma^2 + ||\boldsymbol{\mu}_0 - \mathbf{X}_0\hat{\boldsymbol{\beta}}||^2. \end{aligned} \tag{1.14}$$

where $m\sigma^2$ reflects the underlying variability of the data, $||\boldsymbol{\mu}_0 - \mathbf{X}_0\hat{\boldsymbol{\beta}}||^2$ measures how well the linear model represented by $\mathbf{X}_0$ estimates the mean $\boldsymbol{\mu}_0$ of the new responses $\mathbf{Y}_0$.

In equation (1.14), the cross-product term vanishes because Y and Y_0 are independent and

$$\begin{aligned} E_{\mathbf{Y}_0}[(\mathbf{Y}_0 - \boldsymbol{\mu}_0)^2(\boldsymbol{\mu}_0 - \mathbf{X}_0\hat{\boldsymbol{\beta}})] &= (E_{\mathbf{Y}_0}[\mathbf{Y}_0] - \boldsymbol{\mu}_0')(\boldsymbol{\mu}_0 - \mathbf{X}_0\hat{\boldsymbol{\beta}}) \\ &= (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_0)'(\boldsymbol{\mu}_0 - \mathbf{X}_0\hat{\boldsymbol{\beta}}) \\ &= 0. \end{aligned}$$

Equation (1.14) shows that the PE is the sum of $m\sigma^2$ and a term $||\boldsymbol{\mu}_0 - \mathbf{X}_0\hat{\boldsymbol{\beta}}||^2$. This term $||\boldsymbol{\mu}_0 - \mathbf{X}_0\hat{\boldsymbol{\beta}}||^2$ is called model error (ME) and it is a vital quantity for measuring model's predictive capacity.

$$PE = m\sigma^2 + ME, \tag{1.15}$$

where

$$ME = ||\boldsymbol{\mu}_0 - \mathbf{X}_0\hat{\boldsymbol{\beta}}||^2. \tag{1.16}$$

The ME depends on the regression matrix $\mathbf{X}_0$.

If $\mathbf{X}_0 = \mathbf{X}$, this indicates that $m = n$ and $\boldsymbol{\mu}_0 = E[\mathbf{Y}] = \boldsymbol{\mu}$. Suppose the probability structures of the new observations are equal to old observations. Therefore, let $\boldsymbol{\varepsilon} = \mathbf{Y} - \boldsymbol{\mu}$ and $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, hence a simple equation for the ME can be defined as the following given that the cross-product term disappears again as $\mathbf{P}^2 = \mathbf{P}$.

$$\begin{aligned} ME &= ||\boldsymbol{\mu} - \mathbf{X}\hat{\boldsymbol{\beta}}||^2 \\ &= ||\boldsymbol{\mu} - \mathbf{P}\mathbf{Y}||^2 \\ &= ||\boldsymbol{\mu} - \mathbf{P}(\boldsymbol{\mu} + \boldsymbol{\varepsilon})||^2 \\ &= ||(\mathbf{I}_n - \mathbf{P})\boldsymbol{\mu} - \mathbf{P}\boldsymbol{\varepsilon}||^2 \\ &= ||(\mathbf{I}_n - \mathbf{P})\boldsymbol{\mu}||^2 + ||\mathbf{P}\boldsymbol{\varepsilon}||^2 \\ &= \boldsymbol{\mu}'(\mathbf{I}_n - \mathbf{P})\boldsymbol{\mu} + \boldsymbol{\varepsilon}'\mathbf{P}\boldsymbol{\varepsilon}, \end{aligned}$$

The expected model error $E[ME]$ is

$$\begin{aligned} E[ME] &= \boldsymbol{\mu}'(\mathbf{I}_n - \mathbf{P})\boldsymbol{\mu} + E[\varepsilon'\mathbf{P}\varepsilon] \\ &= \boldsymbol{\mu}'(\mathbf{I}_n - \mathbf{P})\boldsymbol{\mu} + \sigma^2 tr(\mathbf{P}) \\ &= \boldsymbol{\mu}'(\mathbf{I}_n - \mathbf{P})\boldsymbol{\mu} + \sigma^2 p. \end{aligned} \tag{1.17}$$

Using equation (1.15) and (1.17), the corresponding formula for the expected PE when $\mathbf{X}_0 = \mathbf{X}$ is

$$\begin{aligned} E[PE] &= E[n\sigma^2 + ME] \\ &= (n+p)\sigma^2 + \boldsymbol{\mu}'(\mathbf{I}_n - \mathbf{P})\boldsymbol{\mu}, \end{aligned} \tag{1.18}$$

Now define the total bias and total variance of the predictor $\mathbf{X}\hat{\boldsymbol{\beta}}$ by

$$\text{TOTAL BIAS} = ||\boldsymbol{\mu} - E[\mathbf{X}\hat{\boldsymbol{\beta}}]||$$

and

$$\text{TOTAL VARIANCE} = tr(Var[\mathbf{X}\hat{\boldsymbol{\beta}}]) = \sigma^2 tr(\mathbf{P}).$$

The first term of equation (1.17) is only the square of the total bias,

$$||\mu - E[\mathbf{X}\hat{\boldsymbol{\beta}}]||^2 = ||\boldsymbol{\mu} - (\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}E[\mathbf{Y}]||^2 = ||(\mathbf{I}_n - \mathbf{P})\boldsymbol{\mu}||^2 = \boldsymbol{\mu}'(\mathbf{I}_n - \mathbf{P})\mu.$$

The second term of equation (1.17) indicates the total variance. Subsequently equation (1.17) expresses the expected ME as the addition of the total of the measuring bias and the measuring variability of the predictor. Thus, when the variability becomes bigger yet the bias becomes smaller when the model includes new variables. The best model, having minimum expected ME, will have negotiations amid these conflicting prerequisites of small bias and low variability.

Consequently, the bias and variance are affected by the quantity of variables in the model. Quantity of variables and variance both tend to increase together, whereas the bias decreases as the quantity of model increases.

1.7.3 Choosing the Best Subset

Section 1.7.2 shows that utilizing a model which includes all X variables, will cause the predictive ability to reduce. An alternative way to improve model's predictive ability is to utilize one subset of the predictor variables and apply a least squares predictor based on the chosen subset. This section discuss the methodology to identify the best subset. Use a criterion to evaluate the performance of each model, then select the model that optimizes the criterion.

Numerous approaches are utilized to select model. We split these approaches into the accompanying classes: goodness-of-fit measures, estimating distributional discrepancies and criteria based on prediction error.

1.7.3.1 Goodness-of-Fit Criteria

The R^2 criterion is known as the coefficient of determination and defined as follows [5]:

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} \tag{1.19}$$

where *SSTO* is the total sum of squares and SSE is the error sum of squares.

In order to recognize a few good subsets of X variables, subsets with high R^2 are chosen.

The R^2 criterion and error sum of squares SSE provide the same information. Since the denominator $SSTO$ is a constant for all possible regression models, when R^2 becomes bigger, the SSE becomes smaller. Therefore the model is better when the SSE is at low level.

The R_p^2 criterion is not intended to identify the subsets that maximize this criterion because R^2 can never diminish when more X variables are included in the model. Consequently R^2 will be maximum when all P-1 potential X variables are added in the regression model. The expectation in utilizing the R^2 criterion is to discover the point where adding more X parameters does not increase R^2 significantly. Obviously, the determination of significant increment in R^2 is judgmental in nature.

Since R^2 does not take note of the quantity of variables in the model and since max (R^2) can never decreases as p increases, the adjusted coefficient of multiple determination $R^2_{(Adj)}$, adjusts R^2 by dividing each sum of squares by its associated *df*:

$$\begin{aligned} R_{(Adj)}{}^2 &= 1 - \frac{SSE/(n-p)}{SSTO/(n-1)} \\ &= 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE}{SSTO} \\ & 1 - \frac{MSE}{SSTO/(n-1)} \end{aligned} \tag{1.20}$$

This coefficient takes the quantity of parameters in the regression model into account through the degrees of freedom (*df*). From equation (1.20) the $R^2_{(Adj)}$ increases if and only if MSE diminishes since SSTO$/(n-1)$ is fixed for the given Y data. Subsequently, $R^2_{(Adj)}$ is equivalent to utilizing MSE, so $R^2_{(Adj)}$ might be utilized to identify the subsets that maximize this criterion. The largest $R^2_{(Adj)}$ for a given number of parameters in the model, max $(R^2_{(Adj)})$, can indeed decreases as p increases. This happen when the increase in max (R^2) turns out to be too small to the point that it is not adequate to offset the loss of an extra *df*. Users of the $R^2_{(Adj)}$ criterion seek to find a few subsets for which $R^2_{(Adj)}$ is at the maximum or so near to the maximum that adding more variables is not beneficial.

1.7.3.2 Estimating Distributional Discrepancies - AIC_p and BIC_p Criterion

$R^2_{(Adj,p)}$ is a model selection criteria that penalizes models having large quantity of predictors. While, Akaike's information criterion (AIC_p) and Bayesian information criterion (BIC_p) are well known methods and these methods give penalties for adding predictors [7].

For a regression model with Gaussian errors, the AIC_p is defined as [8]

$$AIC_p = n\ln(SSE_p) - n\ln(n) + 2p \tag{1.21}$$

and the BIC_p is defined as [8]

$$BIC_p = n\ln(SSE_p) - n\ln(n) + [\ln(n)]p \tag{1.22}$$

where p is the number of parameters, n is the size of sample and SSE_p is SSE for the model being considered.

Notice that for both of these measures, the first term is $n \ln SSE_p$ which will decrease when p increases. The second term is fixed, and the third term increases proportionately with p. Model with small SSE_p function well by these criteria, as long as the third term of AIC$_p$ and BIC$_p$ are kept small. If $n \geq 8$, the penalty for BIC$_p$ is bigger than that for AIC$_p$; henceforth the BIC$_p$ criterion prefer more parsimonious models. In general, for these two measurements, model with smaller AIC and BIC are preferred.

1.7.3.3 Criteria based on prediction error - Mallows' C_p

The Mallows' C_p is defined as [7]:

$$(SSE_p/MSE_m) - (n - 2p)$$

where SSE_p is SSE for the model being considered, MSE_m is the mean square error for the full model, n is the sample size, and p is the number of parameters in the model, including the constant.

In general, models with small Mallows' C_p and near to p will do well. The reason is the model is reasonably precise in evaluating the true regression coefficients and predicting future observation, the smaller the value of C_p, the better the model predict future observations. However, the model's predictive ability and bias will be poor, when value of C_p is more than p.