A Price Prediction Method In Real Estate Market

by

Heng Li

B.Eng in Civil Engineering, Wuhan University (2010)

Submitted to the Department of Civil and Environmental Engineering
in partial fulfillment of the requirements for the degree of

Master of Science in Civil and Environmental Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2016

Author ...... Signature redacted ....
Department of Civil and Environmental Engineering
January 14, 2016

Certified by ...... Signature redacted ....
Jerome J. Connor
Professor Emeritus of Civil and Environmental Engineering
Thesis Supervisor

Accepted by ...... Signature redacted ....
Heidi Nepf
Donald and Martha Harleman Professor of Civil and Environmental
Engineering Chair, Graduate Program Committee

A Price Prediction Method In Real Estate Market

by

Heng Li

Submitted to the Department of Civil and Environmental Engineering
on January 14, 2016, in partial fulfillment of the
requirements for the degree of
Master of Science in Civil and Environmental Engineering

Abstract

Current housing price prediction usually employs hedonic or repeat-sales models. The objective is to build a statistical model which is more focused on statistic methods. Neither ordinary nor regularized regression model haven been applied to the field of real estate, even though they are rather well-known statistical procedures. This thesis concludes lots of ordinary and regularized regression models. A theoretical review was performed for these models, and Boston Housing data was used to evaluate their performance. The results were found to be reasonable, from a statistical perspective.

Thesis Supervisor: Jerome J. Connor
Title: Professor Emeritus of Civil and Environmental Engineering

Acknowledgments

I realize that I could not finish this thesis without the help of these people:

I want to thank my thesis supervisor and reader, Dr. Jerome J. Connor, for offering me generous guidance, support and suggestions, which made it possible for me to finish a project that I was interested in.

I would also like to express my gratitude to my academic adviser, Dr. Haoxiang Zhu, for inspiring me to come up with a topic that combine my knowledge in numerical field and civil engineering discipline together.

I am grateful to teachers who encourage me to try to be a better man.

I am thankful for my family and friends for their constant support all the time.

Contents

1 Introduction **13**

1.1 Significance of price modeling in real estate 13

1.2 A Brief Literature Review of Real Estate Pricing Methods 15

1.3 The Problem of Existing Real Estate Price Indices Construction Methods and Present Status of Related Research 16

1.4 Thinking Statistically 18

1.5 Motivation description 20

1.6 Outline of my thesis 21

2 Methodology **23**

2.1 Traditional regression model 23

2.1.1 Formal model description 24

2.1.2 Estimation of parameters 25

2.1.3 Prediction schema 26

2.1.4 Selection of variables 26

2.1.5 Potential problem 28

2.2 Regularized regression 29

2.2.1 Ridge regression 29

2.2.2 LASSO regression 30

2.2.3 Elastic net regression 32

2.2.4 Comparison argument 32

2.3 Model diagnosis 33

2.3.1 Linearity of regression function 34

2.3.2 Normality assumption 34

2.3.3 Homogeneity of variance 34

2.3.4 Independence among samples 35

2.3.5 Outliers detection 35

3 Analysis of Boston Housing Data **37**

3.1 Preliminary Analysis 37

3.1.1 Data set description 37

3.1.2 Numerical and graphical exploration 38

3.2 Fit general linear regression model 39

3.2.1 Relationship exploratory 39

3.2.2 Variable selection 40

3.2.3 Model fitting, diagnosis and remedy 41

3.2.4 Prediction result 42

3.2.5 Regularized regression 43

4 Discussion **47**

4.1 Conclusion 47

4.2 Future work 48

A Tables **49**

B Figures **55**

List of Figures

B-1 Prototype of linear and nonlinear regression function residual plots . 56

B-2 Prototype of normal and non-normal regression function residual plots 57

B-3 Scatter plot matrix of highly correlated pairs 58

B-4 Scatter plot of 'MEDV' with potential predictor variables 59

B-5 Histogram of 'MEDV' for 'CHAS'=0 and 1 separately 60

B-6 'Best' subset algorithm outcomes 61

B-7 Graphical diagnosis 62

B-8 Prediction result from linear regression model 63

B-9 Cross-validation for Ridge regression 64

B-10 Prediction result from Ridge regression model 65

B-11 Cross-validation for LASSO regression 66

B-12 Prediction result from LASSO regression model 67

B-13 2-D Cross-validation for Elastic net 68

B-14 Prediction result from Elastic net model 69

List of Tables

A.1 Numerical description of 14 variables in the data set 50

A.2 Correlation matrix for high correlated pairs 51

A.3 Summary table of fitted linear regression model 52

A.4 Relative importance of each predictor variable in linear regression model 53

Chapter 1

Introduction

The major focus of this chapter is on explaining the motivation to model real estate price, providing a comprehensive review of previous work and providing a big picture about how the thesis is organized. Specifically, the necessity and advantages of real estate price modeling is clarified in section 1.1; statistical thought and methodology review are elucidated in section 1.2; finally, section 1.3 is the outline of the whole thesis.

1.1 Significance of price modeling in real estate

Accurate real estate pricing is one of the key issues for countries all over the world.

Firstly, at the microeconomic level, real estate, no matter it is a house or an apartment, is the most expensive expenditure that the majority of people will make during their lifetime. Current market housing prices may cause people to be misled by actions of sellers and other bidders, such as price manipulation and shill bidding. What's more, accurate price models are also desired by real estate agents, whose objective is to sell houses both profitably and quickly. As it is unrealistic to achieve both goals to a high degree, they want to find a balance between profit and speed. A real estate pricing tool is of great help. The value of information in real estate transaction to agents is analytically explained in Levitt and Syverson (2008) [33].

Secondly, at the macroeconomic level, real estate is closely related to the national economic trends and stability of the financial market. In the first place, as argued by Fama and Schwert (1977) [17], Wang and He (2005) [58], inflation has something to do with real estate prices. Real estate price has an effect not only on price level of the whole market but also on people's savings behavior. And we all know that prices and savings effect each other heavily. Generally, saving level becomes less with the increase of retail price and people tend to have more savings if prices are lower. And when saving proportion becomes too small and prices go up dramatically, gradually inflation occurs. As a result, even though the influence that real estate price has on inflation is not substantial in a short term of time, the effect is substantial in the long run. In the meanwhile, Miller et al. (1988) [39] claims that real estate prices have a certain effect on the monetary exchange rate, taking the Japanese yen as an illustration. They conclude that upvaluation of money and high real estate price move hand in hand. So a comprehensive understanding of real estate prices is beneficial to government, by allowing them to supervise and control the whole market effectively.

Finally, the real estate pricing issue also interests academic researchers, as real estate data modeling is challenging and motivates new methodology which makes relevant theory more complete and solid. Here is a list of famous literature regarding real estate pricing index: textbooks by Brueggeman and Fishers (1993) [7], DiPasquale and Wheaton (1996) [15] and Miles et al. (2000) [38] provide comprehensive topics in real estate finance; Castle and Hoch (1982) [10] focus on farm real estate price, while Geltner et al. (2013) [24] concerns commercial real estate; index-based futures and option markets concepts is introduced to real estate by Case Jr et al. (1993) [9] and studied further by Grenadier (1996) [29]; Hong Kong is taken as an example in Chau et al. (2005) [11], while Geneva, Switzerland is analyzed by Hoesil et al. [31]; spatial analysis is applied to real estate analysis by Pace et al. (1998) [44]; dynamics of real estate is discussed in Case and Quigley (1991) [8]; the role of speculation in real estate field can be found be Malpezzi and Wachter (2005) [36]; Yavas and Yang (2005) [60] studies the strategic role of listing price in marketing; thin market real estate price index is presented in Schwann (1998) [51]; Gilberto (1980) [26] investigates real estate returns and equity; Geltner et al. (2003) [23] proposed appraisal smoothing to price discovery; Barlowe (1978) [4] investigate the economics in real estate; alternative real estate building techniques are mentioned in Palmquist (1980) [45].

1.2 A Brief Literature Review of Real Estate Pricing Methods

Since residential properties occupy a large proportion of single-family's wealth, accurate measurement of real estate price indices attracts great attention to real estate researchers and practitioners. Besides that, real estate properties frequently are treated as investment portfolios candidates. Therefore, the prices indices of real estate are essential for investors' decisions in risk hedging, performance evaluation and investment strategy. However, infrequent transactions in real estate market make the predicting of housing price much more complex compared with financial stock market.

The most popular methods to construct real estate assets price indices are divided into two categories: a hedonic regression model (an adjusted-quality index) and a repeat-sales model (a constant-quality index).

With the availability of housing transactions data, hedonic regression method could be applied into real estate market, which first is used in automobiles industry. Then due to the pioneering work done by Rosen (1974), the hedonic pricing method starts to be widely adopted in real estate market.

Hedonic pricing model is an implicit pricing method, which evaluates and forecasts the value of real estate assets by the implicit price of the key underlying characteristics. Most researchers define the key characteristics as 'Structural characteristics', 'Locational characteristics' and 'Environmental characteristics'. Then lots of researchers conjecture their own type hedonic pricing functions; nevertheless, all these functions have a common pattern–that is–they all weight the set of key characteristics in some function forms. And the assumption of key characteristics is of great importance, which would determine the accuracy of price indices (Beamonte and Gargallo 2013). During the past several decades, it has been widely acknowledged that hedonic regression modelling is appropriate to account for the important determinants of price variation (Kain and Quigley 1970). The usual process of this method is by considering the changed characteristics and temporal variation to disaggregate the hedonic pricing function components. A standard or quality-adjusted real estate price index would then be constructed from the statistical results (Bryan and Colwell1982).

Groundbreaking paper written by Bailey, Muth and Nourse (1963) issues an alternative method called as a repeat-sales method to estimate a constant quality real estate price index. Since quality differences make prediction of price indices of real estate market difficult, the repeat-sales method could avoid these difficulties to eliminate quality differences by using prices at different points in time for the same property, provided that property characteristics are unchanged between two sales. Therefore, those properties that have changed substantially between two sales must be excluded from data samples. The repeat-sales method has already become the most popular approach to construct real estate price indices for its advantages. The power of this method is that it could compare constant quality properties across different periods and avoid characteristics selection and function form selection versus hedonic model (Dombrow, Knight and Sirmans 1997).

1.3 The Problem of Existing Real Estate Price Indices Construction Methods and Present Status of Related Research

Economists have long recognized that housing markets are geographically localized (Case,Pollakowski and Wachter 1991). Therefore, real estate market price indices are typically assumed to predict for certain geographic area within county area or metropolitan area, which only valid for that particular area. Another crucial issue of real estate market differing with financial market is assumption of imperfections in real estate, (Geltner and Kluger 1998).

There are many other methodologies to construct real estate market price indices except for hedonic regressions and repeat-sales analysis (Englund, Quigley and Redfeam 1998; Case and Quigley 1991). However, hedonic and repeat-sales models are the fundamental of all other evolution models.

Unfortunately, Both of these two models have their own flaws.

Since the hedonic regression model uses data on a vector of key characteristics to control property quality, this method needs large amounts of data across different periods. Then hedonic regression method would only be limited to construct real estate price indices in some metropolitan areas. To obtain unbiased predictions of real estate prices, it is critical to define both the set of key characteristics and hedonic functional form correctly. Nevertheless, selecting the set of regression factors is subjective and flexible because many attributes would have an influence on the transaction price of housing (Clapp and Giaccotto 1992). Therefore, there are some criticisms of hedonic pricing model, which mainly are referred to the estimation of hedonic pricing function. Halvorsen and Palmquist (1980) proposed that two aspects must be fulfilled to guarantee the accuracy of price indices: 1) Inclusion of a correct set of property characteristics in the regression; 2) Selection of appropriate functional form for estimating equation. Lack of standard selection techniques may fail to include some influential attributes in the function. Finally, large and costly data sets with actual sales prices and property characteristics result in this method less appealing (Meese and Wallace 1991).

Due to the difficulty of accurately specifying the hedonic pricing model, some researchers have issued the repeat-sales method to avoid the bias from an imperfect assumed hedonic model, in the meanwhile, taken advantage of the constant-quality controls by repeat transactions of the same property. The repeat-sales method could avoid the differentiated characteristics of real estate properties by basing the price index on sales price of the same property at different time, which means that real estate properties of the selected sample would be transacted at least twice during the supposed period. With the elimination of characteristics and pricing function selection, repeat-sales method has gradually dominated in establishing real estate price indices. However the repeat-sales method also has some flaws. Since the selected properties must be transacted at least twice, those properties sold only once must be discarded from the sample, which exclude the majority of transactions data. Therefore, it fails to utilize the full information in real estate market, which may affect the results. More importantly, even the remaining sample would be biased with the fact that the attributes of the properties may change between two sale dates, such as aging, replacement and rehabilitation. Although it may be able to eliminate those physical characteristics that have changed between sales from statistical analysis, it is difficult to identify properties whose locational attributes have changed (Case and Quigley 1991).

1.4 Thinking Statistically

Statistics, as an important branch of Mathematical science, provides logical methodology to explore valuable information contained in data set, textbook Friedman et al. (2001) [20] is a good resource to refer to. Even though there are numerous topics being covered in the field of statistics, including survival data analysis (Cox and Oakes (1984) [13]), time series analysis (Shumway and Stoffer (2008)[12]), regression model (Neter et al. (1963) [42]; Myers (1990) [41]), functional data analysis (Ramsay (2006) [49]), machine learning (Bishop (2006) [6]), Bayesian analysis (Berger (2013) [5]) and so forth, they all follow the universal idea that the ultimate goal of statistical procedures is to infer truth of the whole population from limited information provided by a small proportion of all population, also called samples, with a certain degree of confidence. Namely, statistical procedures trade in a little bit certainty of results, seeking tremendous savings in time and energy to collect data, storage cost and computation cost to analyze data.

Because of its tremendous power in delving into data deeply and dig valuable information about the population out, statistics has been widely employed in numerous fields, such as Finance, Biology information, Computer Science, Public Health, Pharmaceutical industry, History analysis, Insurance science, Chemistry and so on. Application to real estate is far away from mature enough, even though a certain amount of literature has taken efforts to combine statistics science and real estate together: a five equation recursive methodology regarding prediction for US form real estate is given in Tweeten and Martin (1966)[56]; time series method are utilized to figure relationship between real estate and price out in Cheng et al. (2008) [12]; index created from sales prices of the same property at different times are considered by regression model in Bailey et al. (1963) [3], and Quigley (1995) [48] investigate the same topic by a simple hybrid model; econometric analysis in Straszheim (1975) [53] provides a new prospective; statistical procedures in spatial analysis are applied to real estate problem to recover spatial information by Roehner (1999) [50], Dubin et al. (1999) [16]; outliers properties in terms of real estate prices are being detected by Ashefelter and Genesove (1992) [1].

To the best of my knowledge, regression model, especially regularized regression model has not been performed in real estate data analysis up to now, even though it has been well developed by previous studies and been implemented in other scientific and empirical fields: general aspects in regression model is stated by Mosteller and Tukey (1977) [40]; model estimation and prediction in Fu (1998) [21]; variable selection techniques are given in Kuo and Mallick (1998) [32], Geveke et al. (1996) [25]; remedies for outliers are formalized in Lletí et al. (2005) [34]; regularized regression models are argued by Tibshirani (2011) [55], Tibshirani (1996) [54], Zou and Hastie (2005) [61], De Mol et al. (2009) [14], Owen (2007) [43], Wu et al. (2006) [59], Wang et al. (2007) [57], Goodhill and Willshaw (1990) [27], Friedman et al. (2009) [19], Polson and Scott (2012) [46] and so forth; overfitting issue and possible remedies are discussed in Babyak (2004) [2] and Hawkins (2004) [30]; several literature including Graham (2003) [28] and Farrar and Glauber (1967) [18] concern the illposed problem called multicollinearity; Graham (2003) [28] applies regression methodology to ecological problems; Mahon (1996) [35] utilizes regression techniques to geochemistry; Price (1977) [47] solves an nonexperimental data problem by statistical regression model.

1.5 Motivation description

This thesis is motivated by a real estate data set in real practice, called Boston Housing data, where 14 features, including median sell price of owner-occupied properties, crime rate by town, residential land zoned for lots over 25,000 sq.ft proportion, non-retail acres proportion, whether or not bounds Charles River, nitric oxides concentration, average number of rooms, age index, distance to five Boston employment centers, accessibility to radial highways index, tax rate index, pupil-teacher ratio, black population index and lower status population proportion, are available for 506 US cencus tractors near Boston. One of the most straightforward information that we what to extract from this data set is how property prices are affected by the other 13 features, i.e, how property prices can be predicted basing upon information contained in the other 13 features.

Regression models, one of the most powerful tools in statistics, focus on learning the relationship among multiple variables statistically by coming up a functional relationship between two or more features such that the feature that people care mostly can be predicted from the other or the others. As mentioned above, the scientific goal here is to achieve accurate enough predictions of median property price of tracts basing upon all approachable features, so regression model in statistics science is one of the most reasonable solvents for this problem.

Even though the most simplest version of regression model works pretty well in most circumstance, it suffers from several severe drawbacks, such as a certain amount number of samples are required, multicollinearity, over-fitting and other illposed problems may occur and make regression model almost useless, in terms of prediction power, and so forth. Under circumstance like these, a more advanced version of regression model, regularized regression, is sometimes a better choice. As it overcomes these challenges by trading in accuracy of estimation and yields a effective solution, when one or more of these problems occurs. When analyzing the Boston Housing data, regularization regressions are proposed to be candidate models that should be considered, as multicollinearity, i.e, overlapped information may provided by the 13 features with a high probability, for example, nitric oxides concentration and age index are probably positively correlated, as old houses are built with materials made by old technology, such that they are more likely to contain more harmful chemicals then newer properties.

1.6 Outline of my thesis

In this section, a big picture of the whole thesis is listed. The following chapters will be presented in the following order:

Chapter 2 is the theory part, it introduces regression model and lists all relevant topics that we may meet with during our implementing process: how regression model can be formally built by mathematical formula, stimation methodology for unknown values in our model, how the estimated should be utilized to provide useful information, how to select variables among all provided variables, especially when the number of variables are extremely large and potential challenges in regression model are discussed in section 2.1; regularization techniques, namely, Ridge regression, LASSO regression and Elastic net, are also introduced with detailed argument about how each of them can be implemented in section 2.2; finally, model diagnosis for assumptions of regression model, which plays a key role in validating our model, is also discussed thoroughly in section 2.3.

Chapter 3 is the data modeling part, where the whole statistical modeling process is recorded with great details: data analysis procedure starts with numerical and graphical exploratory analysis, following the tradition in statistics; next variable selection process and estimating results for linear regression model are offered; in the meantime, diagnosis plots and interpretation of fitted regression model come after that; finally, prediction performance for ordinary regression and regularized regression are compared, regarding to estimation power and prediction power defined in Chapter 3.

Chapter 4 is the summary part, where conclusions are made to summarize major results achieved in this project and several future work are suggested based on comprehension of this problem.

Chapter 2

Methodology

As the scientific goal of our project is to help people in real estate market, both sellers and buyers, to achieve statistically accurate predictions of trading prices basing upon all approachable features, regression model in statistics science is one of the most reasonable solvents for this problem. The underlying idea of regression model is to come up a functional relationship between two or more features such that the feature that people care mostly can be predicted from the other or the others. The whole chapter here presents theory of regression formally by introducing traditional regression in section 2.1 and arguing several regularization techniques in section 2.2.

2.1 Traditional regression model

It's well-known that the concept of regression was initially introduced by Galton, when he focused on analyzing the inherit behavior of sweet peas in the year of 1894 [22]. Since then, a vast of literature pointing at investigating different aspects of regression model has been providing by researchers and practitioners. The review papers by Stanton et al. (2001) [52] and Barnes (1998) [3] go through the historical development of regression briefly and textbook by Neter et al. (1996) [42] offer an comprehensive prospective of regression model in detail.

Even though there are different kinds of regression model, including linear regression, non-linear regression and generalized regression family composing of logistic regression and Poisson regression, we will discuss linear regression model merely in this the several following sections, not only because that linear regression is the most basic and popular one, but also it is the most convenient model among those who can handle continuous quantitative variable 'price'.

2.1.1 Formal model description

First of all, we introduces some notations and statistical terms: the feature that we want to predict is called response variable, denoted as Y, while all other features are predictive variables, denoted as $X_1, ..., X_p$, i.e., we believe p features $X_1, ..., X_p$ are mainly responsible for determination of Y. However, it's apparent that $X_1, ..., X_p$ cannot explain changes of response variable Y thoroughly, so a random error term ϵ should be added to the model to represent all the other features that may affect values of Y. Then theoretical linear regression model is

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \epsilon$$

where Y and $X_1, ..., X_p$ are given constants in the problem; $\beta_0, ..., \beta_p$ are unknown parameters that we want to estimate, and they play an essential role in regression, as we will be able to make predictions by our model, once we have their estimates; ϵ are usually assumed to be Gaussian$(0, \sigma^2)$, the most common statistical distribution that is able to model real data with a high degree of accuracy and is easy to perform statistical inference with.

For the linear regression model formalized above, we should interpret meanings of parameters $\beta_0, ..., \beta_p$ as follows: β_j, $j = 1, ..., p$ quantifies the expected increases of Y when X_j increases by one measurement unit, given all the other features $X_1, ..., X_{j-1}, X_{j+1}, ...X_p$ stay the unchanged; β_0 equals the expected value of Y when all predict variables are zero. Based on this knowledge, the following facts are straightforward: feature X_j doesn't have any influence on the response variable Y, if and only if $\beta_j = 0$; positive β_j reflects positive relationship, i.e., value of Y will be higher if the corresponding value of X_j is larger; similarly negative β_j indicates negative relationship between Y and X_j.

2.1.2 Estimation of parameters

All relevant unknown parameters in linear regression model are coefficients $\beta_0, ..., \beta_p$ and σ, variance of random error. And they are required to be estimated in order to implement our model meaningfully. Following the rule of thumb in statistics, we use lower-case letters to represent sample. Let's assume we totally have n identically independently distributed samples, $(y_i, x_{1i}, ..., x_{pi})$ for $i = 1, ..., n$, i.e.

$$\vec{y} \ = \ x\vec{\beta} + \vec{\epsilon}$$

where $\vec{y} = (y_1, ..., y_n)^T$, $\vec{\beta} = (\beta_0, \beta_1, ..., \beta_p)^T$, $\vec{\epsilon} = (\epsilon_1, ..., \epsilon_n)^T$ and matrix $x = (x_1, ..., x_n)^T$, with $x_i = (x_{1i}, ..., x_{1i})^T$.

According to common sense, we hope that our estimated response vector $\vec{\hat{y}} = x\vec{\hat{\beta}}$ as close as to the true response vector $\vec{y}$ as possible, i.e., $\vec{\hat{\beta}}$ minimizes

$$||\vec{y} - x\vec{\beta}||_2 \ = \ (\vec{y} - x\vec{\beta})^T(\vec{y} - x\vec{\beta})$$

Matrix and vector differentiation yield the result $\vec{\hat{\beta}} = (x^T x)^{-1} x^T \vec{y}$ instantly. This algorithm is called Lease Square Estimate (LSE; Mardquardt (1963) [37]).

As for an estimate of σ, a measure of inconsistency between observation $\vec{y}$ and estimation $\vec{\hat{y}}$, is found to be

$$\hat{\sigma}^2 \ = \ \frac{(\vec{y} - x\vec{\beta})^T(\vec{y} - x\vec{\beta})}{n - (p + 1)}$$

Once appropriate estimates for all parameters in our regression models are allowed, we can next utilize it to make prediction for any new subjects.

2.1.3 Prediction schema

In order to make predict y^{pred} for a new subject, our model requires $(x_1^{new}, ..., x_p^{new})^T$ are all given constants, then the naive predictor

$$y^{pred} \quad = (x_1^{new}, ..., x_p^{new})^T \vec{\hat{\beta}}$$

has been proven to be the best linear unbiased predictor (BLUP) of y given the fact that $(x_1, ..., x_p) = (x_1^{new}, ..., x_p^{new})^T$. Note that the statistical term 'unbiase' means that expectation of y^{pred} equals to its true value, i.e., the predictor is rather precise.

2.1.4 Selection of variables

In subsection 2.1.1-2.1.3, we develop our analysis basing on fact that we have get a good knowledge of which predictor variables should be included in our linear regression model. However, in real practice, such as our data analysis in section 3, we usually have no idea about which features should be taken into account, even though relevant background may provide us some useful points. Consequently, statistically reasonable procedures should be employed to perform selection of variables before fitting the model.

First of all, criteria should be chosen to quantify the quality of different models. MSE, AIC and R^2 are three of the most commonly used statistical rule, where

$$\begin{aligned} MSE &= \frac{SSE}{n-(p+1)} = \frac{(\vec{y} - x\vec{\beta})^T(\vec{y} - x\vec{\beta})}{n-(p+1)} \\ AIC &= n \ln SSE - n \ln n + 2p \\ R^2 &= 1 - \frac{SSE}{(\vec{y} - \vec{\bar{y}})^T(\vec{y} - \vec{\bar{y}})} \end{aligned}$$

As smaller SSE value indicates better models, we can easily find that we should choose model with small MSE, small AIC and large R^2 respectively.

Next, we can start to conduct variable selection. The most self-evident and accurate method is that we fit regression model on all possible subsets of predictor variables and choose the one that yield smallest AIC value, if we use AIC criteria as an example. This is often referred as 'Best' subsets algorithm. However, this method will be rather time-wasting, when number of predictor variables are extremely large. Under this circumstance, we could get a balance between accuracy and efficiency by adopting one of the following three procedures: forward stepwise algorithm, backward stepwise algorithm and hybrid selection algorithm.

To be more specific, for forward stepwise algorithm:

1. For each of the potential predictor variables X_j $j = 1, ..., W$, a linear regression model containing only intercept and X_j is fitted, then the predictor variables with the largest $t_j^* = \frac{\hat{\beta}_j}{s(\hat{\beta}_j)}$ (this is a statistic for testing whether or not $\beta_j = 0$) value is retained in the model. Let's denote it as X_{r_1}.

2. Next, the routine now fits linear regression model with intercept, X_{r_1} and one of the remaining $W - 1$ variable. Similarly, we choose the candidate who has the largest t_j^*, for $j = 1, .., r_1 - 1, r_1 + 1, ..., W$. Suppose X_{r_2} is added into our model.

3. After step 2, we want to check if any of the other variables that have already been in the model should be deleted by calculating P-value of statistic and compare it with nominal significance level α (often assumed to be 0.05). If the P-value is larger than α, this variable is retained and we move forward; otherwise, it is deleted and we keep going.

4. We now examine which variable is the next candidate to put it, then examine whether the candidate to remove should be taken out, and so on until no further variables can either be added or removed.

When it comes to backward stepwise algorithm, the fundamental intuition is the same as forward stepwise algorithm, except that we start with the largest model: after fitting regression model with all potential predictor variables X_j, $j = 1, ..., W$, we identify the predictor variables with the largest P-value for testing the null hypothesis that $\beta_j = 0$ against the alternative $\beta_j \neq 0$, then if this P-value is larger than pre-specified level α, we remove it; otherwise, selection is stopped and we should use current model. This process continues until no further predictor variables can be deleted.

These two traditional variable selection procedure has the advantage of quick implement, however, they have been shown to have some drawbacks: firstly, it's possible for them to miss the 'optimal' model, as they only add or deleting one variable at a time; secondly, they tend to select models that are smaller than ones that are desirable for prediction purposes; finally, there is no way that we could control which variables can certainly been selected, so they may not capture features of interests appropriately and are not really helpful investigate the problem of interest. To overcome these disadvantages, a hybrid selection algorithm, named Bidirectional elimination, offers another option. It tests variables of included and excluded at every single step.

2.1.5 Potential problem

Even though all assumptions for linear regression model, including linearity relationship, identically independently distributed Gaussian random errors, are satisfied, we may still be faced with some technical difficulty, when one or more of the following situations occur.

1. Multicollinearity, namely two or more predictor variables are highly linearly correlated (correlation coefficient levels are greater than 0.7), is widely believed to make estimation of coefficients $\vec{\beta}$ change erratically to small changes in data used to perform estimation, which is caused by the fact that highly correlated variables contribute duplicate information.

2. Small number of observations, $n < p$, also rises challenges to regression model. Using a smaller observations to estimate a larger number of unknown parameters is unapproachable not only mathematically but also commonsensibly. Under this circumstance, traditional regression model cannot be fitted appropriately.

3. Overfitting is another severe problem when it occurs, as the model fits the

provided observations perfectly and predictions for new subjects are not reliable. And then the fitted regression model is almost useless for prediction purposes.

Among all the possible remedy for these problems, the most regular one is regularized regression, which will be discussed in detail in section 2.2.

2.2 Regularized regression

Regularized regression model is motivated by the phenomenon that some elements of the estimated coefficients $\vec{\hat{\beta}}$ are extremely large, when one or more ill-posed problems, including multicollinearity, small sample size n and overfitting, arises. The core intuition behind regularized regression is control the magnitude of coefficients $\vec{\beta}$ in addition to minimization of $SSE = (\vec{y} - x\vec{\beta})^T(\vec{y} - x\vec{\beta})$ by introducing certain mathematical formula to penalize size of $\vec{\beta}$, i.e., we want to minimize

$$SSE + \lambda Penalty(\vec{\beta}) \quad = (\vec{y} - x\vec{\beta})^T(\vec{y} - x\vec{\beta}) + \lambda Penalty(\vec{\beta})$$

where λ is a positive parameter to quantify the size of penalty and $Penalty(\dot{)}$ is an function of $\vec{\beta}$. And various choice of function $Penalty(\dot{)}$ results in different regularized regression model.

2.2.1 Ridge regression

Ridge regression is proposed by various researchers in different contexts and became well-known after Andrey Tikhonov published his paper in 1977 and David L. Phillips's paper came out. So the penalty term is also referred as Tikhonov Phillips regularization.

In Ridge regression, L^2 norm of in R^2 vector space is employed in the penalty term,

$$\lambda Penalty(\vec{\beta}) \quad = \lambda||\vec{\beta}||_2^2 = \lambda \textstyle\sum_{j=1}^{p} \beta_j^2$$

and instead of minimizing SSE,

$$SSE + \lambda||\vec{\beta}||_2^2 \ = SSE + \lambda \textstyle\sum_{j=1}^{p} \beta_j^2$$

is our target function to minimize.

Mathematical computation similar to the differentiation of vector and matrix technique used for ordinary least square estimation, we have

$$\hat{\vec{\beta}}_{ridge} = (x^T x + \hat{\lambda} I)^{-1} x^T y$$

where $\hat{\lambda}$, estimate of the biasing constant λ, plays an important role in balancing the size of SSE and magnitude of $\vec{\beta}$. It can be usually obtained by one the the following two procedures:

1. k-fold cross-validation: firstly, the whole sample should be divided into k groups with equal size randomly; we next specify a sequence of candidate λ; then for any given λ_0 in the sequence, we use all samples except ones in the first group to fit ridge regression model with parameter λ_0, then we predict response values of using data in the first group, obtaining a error measure $SSE_{\lambda_0,1}$; we follow the same process and go through all groups, achieving $SSE_{\lambda_0,1}, ..., SSE_{\lambda_0,k}$ and their mean value $MSSE_{\lambda_0}$ is a measurement of error for λ_0. After calculating the error term for all λ in the sequence, we determine $\hat{\lambda}$ equals to the one with smallest error values.

2. The second method is based on the ridge trace and the variance inflation factors, which is a judgmental procedure. The rule is that one should choose the smallest value of λ, where regression coefficients $\vec{\beta}$ first become stables in the ridge trace and the variance inflation factors have become sufficiently small.

2.2.2 LASSO regression

LASSO regression comes after ridge regression and is proposed by Tibshirani (1996). He argued that L^1 norm of $\vec{\beta}$ provides an alternative solvent. Then the penalty function is now

$$\lambda Penalty(\vec{\beta}) \quad = \lambda||\vec{\beta}||_1 = \lambda \textstyle\sum_{j=1}^{p} |\beta_j|$$

and our goal is to find $\vec{\beta}$ such that

$$SSE + \lambda||\vec{\beta}||_1 \quad = SSE + \lambda \textstyle\sum_{j=1}^{p} |\beta_j|$$

is minimized.

Similarly, parameter λ controls strength of penalty: if $\lambda = 0$, it returns back to linear regression model; if $\lambda = \infty$, all coefficients are zero.

The explicit formula for the estimate of $\vec{\beta}$ is found to be

$$\hat{\beta}_j(LASSO) = (\hat{\beta}_j - \frac{\lambda}{2})_{+} sgn[\hat{\beta}_j]$$

where $\hat{\beta}_j$ comes from ordinary least square estimate and $sgn(x) = 1$ if $x > 0$; $sgn(x) = 0$ if $x = 0$; $sgn(x) = -1$ otherwise. To implement this estimate, λ is required to be specified ahead and the k-fold cross validation approach discussed in Ridge regression section is also applicable here.

There is an alternative method to get $\vec{\hat{\beta}}$ for LASSO. A so-called Orthogonal Matching Pursuit (OMP) method provides the estimate algorithm:

1. Set $r = \vec{y}$ and $\beta_j = 0$ for all $j = 1, ..., p$

2. For $i = 1, ..., n$: set $x_j = argmax_{x_{j'} \in X}| < r, x_{j'} > |$; set $\beta_j = argmin_{\gamma}||r - x_j\gamma||^2 + \lambda|\gamma|$; set $r = r - x_j\beta_j$

3. Return $\vec{\beta}$

As for penalty λ, it can be estimated by the k-fold cross validation approach discussed in Ridge regression section.

2.2.3 Elastic net regression

Elastic net regression is a hybrid of Ridge regression and LASSO regression. Being firstly investigated by Zou and Hastie (2005), Elastic net regression assumes the regularization term equals

$$\lambda Penalty(\vec{\beta}) \quad = \lambda\tfrac{1-\alpha}{2}||\vec{\beta}||_2^2 + \lambda\alpha||\vec{\beta}||_1 = \lambda\tfrac{1-\alpha}{2}\textstyle\sum_{j=1}^p \beta_j^2 + \lambda\alpha\sum_{j=1}^p |\beta_j|$$

and function

$$SSE + \lambda\frac{1-\alpha}{2}||\vec{\beta}||_2^2 + \lambda\alpha||\vec{\beta}||_1 \quad = SSE + \lambda\tfrac{1-\alpha}{2}\textstyle\sum_{j=1}^p \beta_j^2 + \lambda\alpha\sum_{j=1}^p |\beta_j|$$

with parameters $\vec{\beta}$, λ and α is our target function to minimize. And α has the value between 0 and 1 and it quantifies the proportion L^1 norm regularization; the function of λ is the same as those in Ridge regression and LASSO regression.

The explicit formula for the estimate of $\vec{\beta}$ is found to be

$$\hat{\beta}_j(Elastic) = \frac{(\hat{\beta}_j - \frac{\lambda\frac{1-\alpha}{2}}{2})_+}{1+\lambda\alpha} sgn[\hat{\beta}_j]$$

where $\hat{\beta}_j$ comes from ordinary least square estimate and $sgn(x) = 1$ if $x > 0$; $sgn(x) = 0$ if $x = 0$; $sgn(x) = -1$ otherwise. To implement this estimate, λ and α should be estimated by k-fold cross validation approach discussed in Ridge regression section is also applicable here.

2.2.4 Comparison argument

Ridge regression is regularized by L^2, which controls the magnitude of parameters $\vec{\beta}$ efficiently. In real practice, as Ridge regression shrinks the coefficients towards zero, it performs rather well when there is a subset of true coefficients that are small or even zero, while it does not as well when all of the true coefficients are moderately large.

As the nature of the L^1-typed penalty, some coefficients in LASSO regression are shrunken to zero exactly, which makes LASSO substantially different from Ridge regression, as it is capable of performing variable selection and regularization at the same time. This feature becomes extremely helpful to introduce sparsity in regression model.

Elastic net combines superiority of Ridge regression and LASSO regression. But it introduces two tuning parameters, which not only makes the model more complex, but may also make the computation more expansive.

When solving real world problems, we cannot say which one is the best all the time. It depends on the question of interest and property of data.

2.3 Model diagnosis

After variable selection procedure and fitting either linear regression model or one of the regularized regression model with selected variables, certain examination should be executed to confirm the validity of our model, both numerically and graphically.

Basically, there are five assumptions in both ordinary linear regression model and regularized linear regression model: regression function, the functional relationship between response variable and predictor variables, is linear; all random errors ϵ_i, $i = 1, .., n$ are normally distributed; variance of random errors are the same; random errors, consequently, response variables y_i, are statistically independent, namely, the value of y_{j_1} has no effect on value of y_{j_2}, as long as $j_1 \neq j_2$. In this section, we will discuss relevant examination techniques for these assumptions one by one.

Before talking about the techniques formally, we firstly introduce an essential statistical concept. As all the model assumptions are related to random errors in a certain degree, so the sample version of random errors, residuals

$$r_i = y_i - \hat{y}_i = y_i - x_i\vec{\hat{\beta}}$$

undoubtedly serve a key part in model diagnosis.

2.3.1 Linearity of regression function

For the linearity assumption, a scatter plot with fitted response $\hat{y_i}$ represented in x-axis and corresponding residuals in y-axis, called residual plot, is the most regular graphic diagnostic method. If the points scattered randomly around the horizontal line $y = 0$ without any systematic pattern showing up, linear regression function is then an reasonable assumption; otherwise, the fitted linear function might provide misleading predicted information. The Prototypes of linear and nonlinear residual plots are provided in Figure B-1.

2.3.2 Normality assumption

Normality assumption in regression model can be justified by two methods, one graphical method and one formal statistical testing method.

The graphical technique is $Q - Q$ plot, where Q is short for quantile. In statistics, for any constant α between 0 and 1, α-quantile is the number so that the probability that the variable greater than this number equals to α. When checking if two variables have the same distribution, we can confirm this argument by examining if the quantiles for $\alpha \in [0, 1]$ are all the same. So in $Q - Q$ plot, where x-axis present all quantiles for Gaussian distribution and y-axis present all quantiles of our residuals, if we scatters form a straight line, then normality assumption is satisfied. The Prototypes of Gaussian and non-Gaussian $Q - Q$ plots are provided in Figure B-2.

The graphic technique is very convenient to implement and provide a relatively correct general idea, however, sometimes, practitioners also take efforts to pursue an non-subjective answer. Several famous normality test, such as Kolmogorov-Simirnov test and Shapiro-Wilk test and so forth, are suitable choices. More details of these tests can be found in any regression textbook.

2.3.3 Homogeneity of variance

We will focus on the graphical method in checking the homogeneity assumption, even though several formal hypothesis testing procedures (such as Brown-Forsythe test, Breusch-Pagan test and so on) are available.

We again use exactly the same residual plot mentioned in section 2.3.1. And if the residuals distributed randomly within a horizontal band around the linear $y = 0$, we claim that the variances of random errors are constant; otherwise, homogeneous variance is violated. Again Figure B-3 provide prototypes of constant and non-constant residual plots.

2.3.4 Independence among samples

Independence among subjects are generally related temporally and (or) geographically. Let's use time dependence detection as an example for illustration purpose. A scatter plot with 'time' being x-axis and 'residual' being y-axis is the major tool here. If there is an apparent systematic pattern in the scatter plot, then the samples are not independent; otherwise, we have confidence to claim that samples are independent.

2.3.5 Outliers detection

A slight different version of residual plot works well in terms of detecting possible outliers. The y-axis is now the normalized residual, i.e., $r_{i;normalized} = \frac{r_i - mean(r_i)}{var(r_i)}$, instead of the original residuals. And we outliers are those subjects with absolute normalized residual greater than 4.

In chapter 3, we will apply all statistical theory presented here to our 'Boston Housing' data set.

Chapter 3

Analysis of Boston Housing Data

After stating relevant theory about regression, all these data modeling techniques are now applied to our Boston Housing data set, seeking statistically reasonable solution to the fundamental questions in real estate market: 'What factors should be taken into consideration when predicting the price of houses?' and more importantly, 'How could rational prediction be predicted basing upon these factors?'

The whole chapter will be organized as follows: in section 3.1, preliminary analysis is performed by offering a detailed description of our data set and conducting numerical and graphical exploration in section 3.1; then ordinary linear regression model fits our data in section 3.2; the effectiveness of general linear regression model is examined in the next section; finally, regularization discussed in section 2.2 is added to general regression on section 3.4.

3.1 Preliminary Analysis

3.1.1 Data set description

This data set is consisted of housing data for 506 US census tracts in Boston area, with 14 features provided for each tract: variable 'crim' is per capita crime by town; 'zn' means proportion of residential land zoned for lots over 25,000 square foot; 'INDUS' is proportion of non-retail business acres per town; 'CHAS' is an indicator variable about whether the tract bounds Charles River; ‘nox’ gives nitric oxides concentration measured in parts per 10 million; ‘rooms’ equals average number of rooms per house; ‘age’ represents how many of non-rental unites were built before the year of 1940; ‘distance’ is a weighted distances to five Boston employment centers; the next feature is index of accessibility to radial highway and named as ‘RAD’; ‘TAX’ gives full-value property-tax rate every \$ 10,000; ‘PTRATIO’is ratio of pupil and teacher in town; ‘B’ is related to proportion of blacks (denoted by BK) by $B = 1000(Bk - 0.63)^2$; next, ‘LSTAT’ announce percentage of lower status people per town; finally ‘MEDV’ shows median value of owner-occupied houses in unit of \$1,000.

3.1.2 Numerical and graphical exploration

Exploring the data set in advance before acting data modeling provides more insights and is definitely beneficial for analysis afterwards.

Firstly, numerical summaries in statistics, including mean, standard deviation and five number summary (minimum, Q1, medium, Q3, maximum) are helpful to provide distribution feature of each of the 14 variables respectively. Table A-1 report these summary statistics: ‘CRIM’ is highly skewed to the right, i.e., crime rates of most tracts are rather small (75% of them are between 0.00632 % and 3.677083 %), while there are several extremely high values (maximum is as high as 88.9762 %); ‘ZN’ also skews to the right with a high degree; distributions of ‘INDUS’, ‘CHAS’, ‘NOX’, ‘RM’ and ‘AGE’ are relatively symmetric; ‘DIS’ skewed to the right a little bit; ‘RAD’ skewed to the right moderately; ‘TAX’ is symmetrically distributed; distribution of ‘PTRARIO’ is rather symmetric; ‘B’ skewed to left heavily, namely most tracts have Black race percentage well different from 0.63, while a few tracts have approximate 63 % black people; ‘LSTAT’ is skewed to the right; ‘MEDV’ is skewed to the right a little bit but it is rather close to a symmetric distribution.

After exploring each variables separately, investigation of their pairwise relationship is conducted by calculating correlation coefficients, a numerical constant that quantifies strength of linearity between two random variables, with larger absolute value indicating higher level of positive or negative linear relationship. Table A-2 presents correlation coefficients with high values and it can be found that 'INDUS', 'NOX', 'RM', 'AGE', 'DIS', 'RAD' and 'TAX' are linearly correlated heavily. So multicollinearity is an important point that we need to keep in mind when fitting regression models. Scatter plot matrix in Table B-3 also justifies this argument from an more straightforward perspective, as it can be seen clearly that there are strong pairwise linear relationship among those variables.

3.2 Fit general linear regression model

In order to check the effectiveness of our fitted model, we divide our total 506 sample into two groups: the first group contains 337 randomly selected tracts (takes up roughly 2/3 percent of 506) and called training data, on which estimation of parameters in proposed model is based; the remaining subjects in the second group and they are testing data, with which the predict power of the fitted model can be evaluated. Note that this partition methodology and the number 2/3 are almost rule of thumb in statistics and machine learning field.

This project aiming at building a statistically significant model to predict median price level of any new census tracts basing upon one or more factors among the 13 features mentioned in our data set. As a result, regression model seems to be a perfect scientific tool to meet our need. It can be noted immediately that 'MEDV', median value of owner-occupied houses in unit of $1,000, is the response variable and all the other 13 variables are candidate predictor variables.

Detailed model building process and prediction results are demonstrated in this section as follows.

3.2.1 Relationship exploratory

Figure B-4 is a collection of scatter plots of 'MEDV' and each of the 13 candidate predictor variables basing on our training data ('CHAS' is omitted as it is a categorical variable which only have values 0 and 1, i.e., scatter plot will be of little help for analyzing and a side-by-side histogram is plotted to explore the effect of 'CHAS' on 'MEDV' in Figure B-5) and it can be found: 'INDUS', 'NOX', 'RM', 'TAX', 'PTRATIO' and 'LSTAT' have a clear linear type relationship with 'MEDV'; even though the linear relationship between 'MEDV' and 'CRIM', 'ZN', 'AGE', 'DIS', 'RAD' and 'B' correspondingly are not as strong as those argued above, they are still strong enough to justify that linear terms are reasonable choices in the regression function. What's more, Figure B-5 illustrates that shape of distributions of price for tracts bound Charles River and those do not bound are rather close and the group with 'CHAS'=1 shifts to higher value direction by about $5,000. As a result, linear 'CHAS' term is also reasonable.

Based on all argument above, it can be confidently claimed that linear regression model is a rational start for the modeling process.

3.2.2 Variable selection

As noted in section 3.1.2, several variables are highly linearly correlated and may contribute duplicated information to regression model, variable selection is a required step that we need to take.

Implementing forward stepwise, backward stepwise and hybrid selection procedures presented in section 2.1.4 on the training data set results in following final linear regression models:

1. Forward stepwise selection: all 13 predictor variables should be in final linear regression model.

2. Backward stepwise selection: 'CRIM', 'ZN', 'CHAS', 'NOX', 'RM', 'DIS', 'RAD', 'TAX', 'PTRATIO', 'B' and 'LSTAT' are included.

3. Hybrid selection: 'CRIM', 'ZN', 'CHAS', 'NOX', 'RM', 'DIS', 'RAD', 'TAX', 'PTRATIO', 'B' and 'LSTAT' are included.

Note that backward selection and hybrid selection point at the same final model, where 'INDUS' and 'AGE' being excluded, while forward selection includes all the predictor variables. Both of them will be regarded as candidate final models.

The 'Best' subsets algorithm in section 2.1.4 is also acted and corresponding outcomes are summarized in Figure B-6. x-axis in the figure lists names of all the 13 predictor variables, y-axis provides R^2 (a popular measurement for quality of regression models mentioned in section 2.1.4) for the best models with number of predictor variables equals to 1, 2, ..., 13, white blocks mean that the corresponding variables are not in the model. It is not hard to noticed that full model, model that omits 'AGE' and model that excludes both 'INDUS' and 'AGE' are the best models in terms of R^2 (they all reach the largest 0.79 value). Then the model that excludes both 'INDUS' and 'AGE', coinciding with results from backward selection and hybrid selection, is preferred, as simpler models tend to be better in statistics and they tend to do better jobs when being generalized to new data.

Given all these analysis, the final regression model can be announced:

$$y_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_{11} x_{11i} + \epsilon_i$$

for $i = 1, ..., 337$, where x_{1i} are 'CRIM' values within training data; x_{2i} are 'ZN'; x_{3i} mean 'CHAS'; x_{4i} represent 'NOX' values; x_{5i} denote 'RM'; x_{6i} are values of 'DIS'; x_{7i} are 'RAD'; 'TAX' is denoted by x_{8i}; x_{9i} provide 'PTRATIO'; x_{10i} are 'B'; 'LSTAT' is featured by x_{11i}.

3.2.3 Model fitting, diagnosis and remedy

Ordinary lease square estimation method is adopted to fit the training data, summary of fitted results is offered in Table A-3. Indicated by the fitted results, it can be found that: median value of owner-occupied homes in \$1000 increases with the increase of proportion of residential land zoned for lots over 25,000 sq.ft, the fact that it bounds Charles River, number of rooms, accessibility to radial highways, dispersion between black population ratio to 63%, while median price will decrease, when crime rate is higher, nitric oxides concentration is denser, distance to employment centers is longer, tax rate is higher, pupil and teacher rate is higher, lower status population ratio is higher. As all these results are consistent with common sense, the model fitting are quite reasonable from this point of view.

Next, the relative importance for each Predictor is investigated, with results being reported in Table A-4: the proportion of variance of response variable explained by the final linear model with 11 selected predictor variable is 78.57%; using relative importance calculated by the first method 'lmg' as an example, 'LSTAT', 'RM' and 'PTRATIO' are of the greatest significance to the real estate price, and 29.18%, 26.62% and 10.92% are their index of importance respectively; the remaining 8 predictor variables affects variation of median housing price at similar level of degree, around $3-5\%$.

Another substantial fact is that the so-called residual standard error is 4.44, which indicates that the difference between estimated median value $x_i\vec{\hat{\beta}}$ and its true value is $\$4,440$ on average. Note that residual standard error can be treated as a estimation error measurement for regression models.

After interpreting the statistical output of our final regression model, appropriateness of our final model for training data sets is now being checked. All graphical diagnosis can be found in Figure B-7. Given these plots, it can be seen that: linear regression function, constant variance and normality assumptions are met by our training data set at high level, as residual plots are relatively random scattered within a horizontal band around 0 and $Q-Q$ plot has the shape of a straight line; however, an outlier, the 369^{th} observation in original data set, is detected. After taking a close look at this observation, I decide not to delete this subject, as there is not apparent evidence to suggest that this subjects is not recorded correctly and deleting it may lose some important point.

3.2.4 Prediction result

Now let's apply the estimated final model to the testing data set to evaluate the prediction power of the final linear regression model. Visualization of prediction results can be found in Figure B-8. The predicted price line and true price line are close to each other, even though they do not coincide perfectly. Numerically, the mean predictive error (MPE) should be calculated by the following formula

$$MPE = \frac{\sum_{i=1}^{n_{test}} (y_i - \hat{y}_i^{pred})^2}{n_{test}}$$

which serves as a numerical measurement of model's prediction power. The MPE here has the value of 29.06.

3.2.5 Regularized regression

In this part, regularized regression explained in section 2.2, including Ridge regression, LASSO regression and Elastic net, are conducted to see if the final linear regression model can be improved by introducing regularization term.

First of all, for Ridge regression, cross-validation technique is employed to find the optimal regularization parameter λ among sequence 0 to 10 with increment 0.001, resulting in $\lambda = 3.149$. This result is also demonstrated by Figure B-9, the functional relationship between λ and cross-validation error term GCV, with x-axis denoting value of λ and y-axis representing cross-validation error term GCV. Using this regularization value, we fit our final regression again. Visualization of prediction results can be found in Figure B-10. The mean predictive error (MPE) is 28.92, which is improved a little bit than regular linear regression model and the residual standard error found to be is 3.89, which indicates that the difference between estimated median value $x_i\vec{\beta}$ and its true value is now $3,890 on average.

Secondly, cross-validation result (recorded as Figure B-11) for λ sequence 0 to 10 with increment 0.001 suggests that no L^1 type regularization is the best, i.e., $\lambda = 0$ yields smallest cross-validation error. However, for illustration purpose, we set λ equals to 0.01 and use it to fit LASSO regression model. Prediction curve and true curves are given in Figure B-12. MPE for LASSO regression is 29.00 and the residual standard error is 4.41. Even though, there is a little bit improvement for regression model, by using LASSO regularization term, the improvement is smaller than Ridge regression. This outcome makes a lot of sense, as LASSO is mainly good at modeling regression model with sparsity in coefficients and our problem doesn't have sparse features.

Finally, in order to fit Elastic net model, similarly, regularization parameters λ and α are required to be approximated in advance. However, instead of one dimensional cross-validation utilized in Ridge regression and LASSO regression, two dimensional cross-validation gets involved for Elastic net model. In order to search for optimal values of λ and α simultaneously, for each α value in sequence $\{0.01, 0.02, ..., 0.99\}$, cross-validation error term cmv for each λ in sequence $\{0, 0.001, 0.002, 0.003, ..., 10\}$ should be calculated , then among the 99001 combinations of different α and λ, the one with smallest mean error measurement is chosen, i.e., $\alpha = 0.42$ and $\lambda = 0.004$. Visualization of this result is Figure B-13, a three dimensional plots displaying mean errors for each combination of α and λ in two corresponding pre-specified sequences, where α is denoted by x-axis, λ is in y-axis and cross-validation error term CMV is represented in z-axis. And it can be seen that $(x, y) = (0.42, 0.004)$ minimizes value of z approximately. Once these regularization parameters are determined, Elastic model can then be fitted and corresponding prediction results can be achieved. True curves and estimated curves are in Figure B-14 and the MPE for Elastic net model is 26.85 and the residual standard error found to be is 3.44, both of which are the best among all models.

Given all argument above, it is clear that linear regression of 11 selected predictor variables has the ability of providing statistical reasonable model fitting results that are also sensible in terms of common sense: firstly, remember that in the fitted model, median value of owner-occupied homes in \$1000 are positively related with good features, including proportion of residential land zoned for lots over 25,000 sq.ft, the fact that it bounds Charles River, number of rooms, accessibility to radial highways, dispersion between black population ratio to 63%, while median price is negatively related with the remaining bad features; secondly, estimation power of the final regression model, i.e., the residual standard error is rather small; finally, prediction power of the final model, measured in terms of the mean predictive error (MPE) is also within the reasonable range. As for regularized regression, they do introduce smaller residual standard errors and MPE's, even though the improvements are not extremely big. And among the three regularization term, Elastic Net performs best, with Ridge regressing being the next and LASSO being the last one. More discussion about this results will be presented in the following chapter.

Chapter 4

Discussion

In this final chapter, the whole thesis is summarized by the proceeding two parts: remarks concerning the result of real data analysis conducted in chapter 3 serves as conclusions in section 4.1 ; several possible future works regarding this project are listed in section 4.2.

4.1 Conclusion

Based on the arguments in all previous three chapters, the following conclusions are identical:

Firstly, model fitting result from linear regression method makes sense: after preliminary statistical analysis by numerical and graphical examination, linear regression model was proven to be logical methodology for analysis Boston Housing data; Next, linear regression fitting results were reasonable enough, as interpretation of both the estimated model and influence index are rational in terms of common sense; what's more, different variable selection procedures were compared to ensure that our final model was appropriate; finally, detailed model diagnosis, regarding all five assumptions, which are linearity, normality, independent observations, constant variance, no outliers, was performed to validate our model.

Secondly, effects of adding regularization on the linear regression model varies: Ridge regression and Elastic net both improve the original linear regression model in terms of average estimation error, i.e., residual standard error, and prediction power, mean prediction error (MPE), while the improvement from Elastic net, the hybrid of Ridge and LASSO regularization, was larger than Ridge regression. These result make sense, as the Boston Housing data is of good statistical property: sample size is 506, large enough for estimating 13 unknown parameters in the final model; multicollinearity problem is not severe at all after 'INDUS' and 'AGE' being removed from the linear regression model; over-fitting is not a problem for the Boston Housing data. As a result, regularized regression, which is to fix these issues, will improve the model by a rather limited level. Even though these improvements are kind of small, the proposed regularized regression model is still meaningful: it did provide better modeling results; it may yield significant improved results for other problem in the field of real estate.

4.2 Future work

Some possible future works that are relevant to this project are listed below:

1. Extending linear regression model by other modern regularization term, such as the new hybrid term offered in Owen (2007) [43], could be implemented to see if it provides better model fitting results.

2. The effect of partition into training data and testing data could be studied by repeat the random partition process for a large number of times, for each partition, regression model should be fitted. Then examination about the variation of fitted models should be able to answer this question appropriately.

3. Applying our methodology to other real estate data will be helpful to confirm its effectiveness and generality.

4. The effect of partition into training data and testing data could be studied by repeat the random partition process for a large number of times, for each partition, regression model should be fitted. Then examination about the variation of fitted models should be able to answer this question appropriately.

Appendix A

Tables

Table A.1: Numerical description of 14 variables in the data set

variable	mean	standard deviation	min	Q1	medium	Q3	maximum
CRIM	3.6135	8.6015	0.00632	0.082045	0.25651	3.677083	88.9762
ZN	11.3636	23.3225	0	0	0	12.5	100
INDUS	11.1368	6.8604	0.46	5.19	9.69	18.10	27.74
CHAS	0.0692	0.2540	0	0	0	0	1
NOX	0.5547	0.1159	0.385	0.449	0.538	0.624	0.871
RM	6.2846	0.7026	3.561	5.8855	6.2085	6.6235	8.7800
AGE	68.5749	28.1489	2.9	45.025	77.5	94.075	100
DIS	3.7950	2.1057	1.1296	2.100175	3.20745	5.188425	12.1265
RAD	9.5494	8.7073	1	4	5	24	24
TAX	408.2372	168.5371	187	279	330	666	711
PTRATIO	18.4555	2.1649	12.6	17.40	19.05	20.20	22.00
B	356.6740	91.2949	0.32	375.3775	391.44	396.2250	396.9000
LSTAT	12.6531	7.1411	1.73	6.950	11.36	16.955	37.970
MEDV	22.5328	9.1971	5	17.025	21.2	25	50

Table A.2: Correlation matrix for high correlated pairs

	INDUS	NOX	RM	AGE	DIS	RAD	TAX
INDUS	1	0.76365	-0.39168	0.64478	-0.70803	0.59513	0.72076
NOX	0.76365	1	-0.30219	0.73147	-0.76923	0.61144	0.66802
RM	-0.39168	-0.30219	1	-0.24026	0.20525	-0.20985	-0.29205
AGE	0.64478	0.73147	-0.24026	1	-0.74788	0.45602	0.50646
DIS	-0.70803	-0.76923	0.20525	-0.74788	1	-0.49459	-0.53443
RAD	0.59513	0.61144	-0.20985	0.45602	-0.49459	1	0.91023
TAX	0.72076	0.66802	-0.29205	0.50646	77.5	0.91023	1

Table A.3: Summary table of fitted linear regression model

	Estimate	Std.Error	t value	p-value
Intercept	30.820901	6.138125	5.021	8.48e-07
CRIM	-0.141173	0.037075	-3.808	0.000168
ZN	0.041071	0.014961	2.745	0.006385
CHAS	2.721785	1.002930	2.714	0.007006
NOX	-13.623913	4.193561	-3.249	0.001280
RM	4.195958	0.463339	9.056	$< 2e-16$
DIS	-1.510601	0.217961	-6.931	2.26e-11
RAD	0.328993	0.074892	4.393	1.52e-05
TAX	-0.011903	0.003990	-2.983	0.003072
PTRATIO	-0.963671	0.148467	-6.491	3.19e-10
B	0.013874	0.003193	4.345	1.87e-05
LSTAT	-0.554977	0.054009	-10.276	$< 2e-16$

Table A.4: Relative importance of each predictor variable in linear regression model

	lmg	last	first	pratt
CRIM	0.04247827	0.03974715	0.06303720	0.06472124
ZN	0.04237676	0.02065840	0.06071442	0.04983951
CHAS	0.02126913	0.02018964	0.01632610	0.01831275
NOX	0.05348607	0.02893342	0.08023018	0.09587225
RM	0.26619849	0.22481586	0.20552169	0.28811442
DIS	0.03729336	0.13167489	0.02593350	-0.10399738
RAD	0.03775266	0.05290042	0.06403365	-0.15652559
TAX	0.05751941	0.02439086	0.09394073	0.13291788
PTRATIO	0.10919639	0.11549472	0.11045553	0.14960897
B	0.04059905	0.05174211	0.04669215	0.05501827
LSTAT	0.29183040	0.28945253	0.23311484	0.40611769

Appendix B

Figures

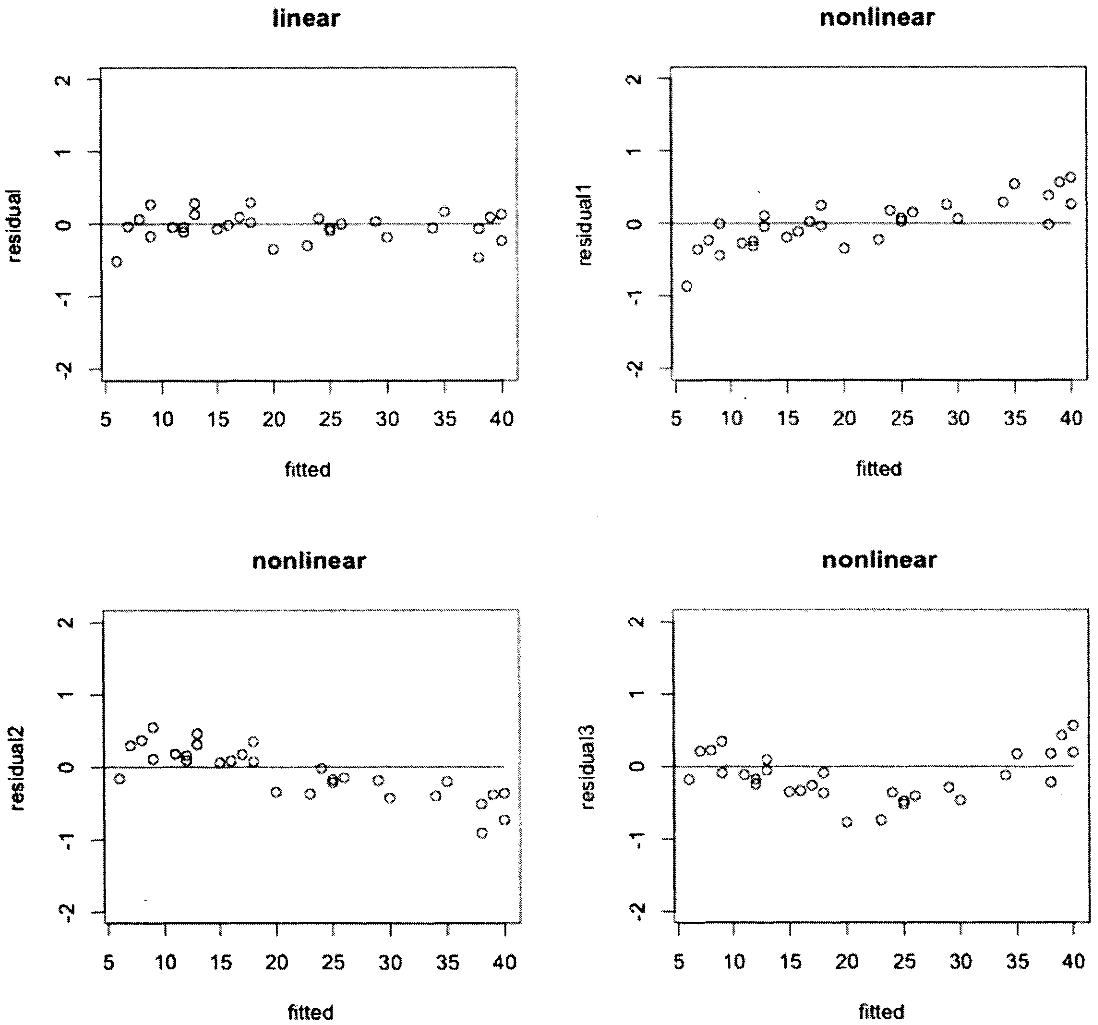

Figure B-1: Prototype of linear and nonlinear regression function residual plots

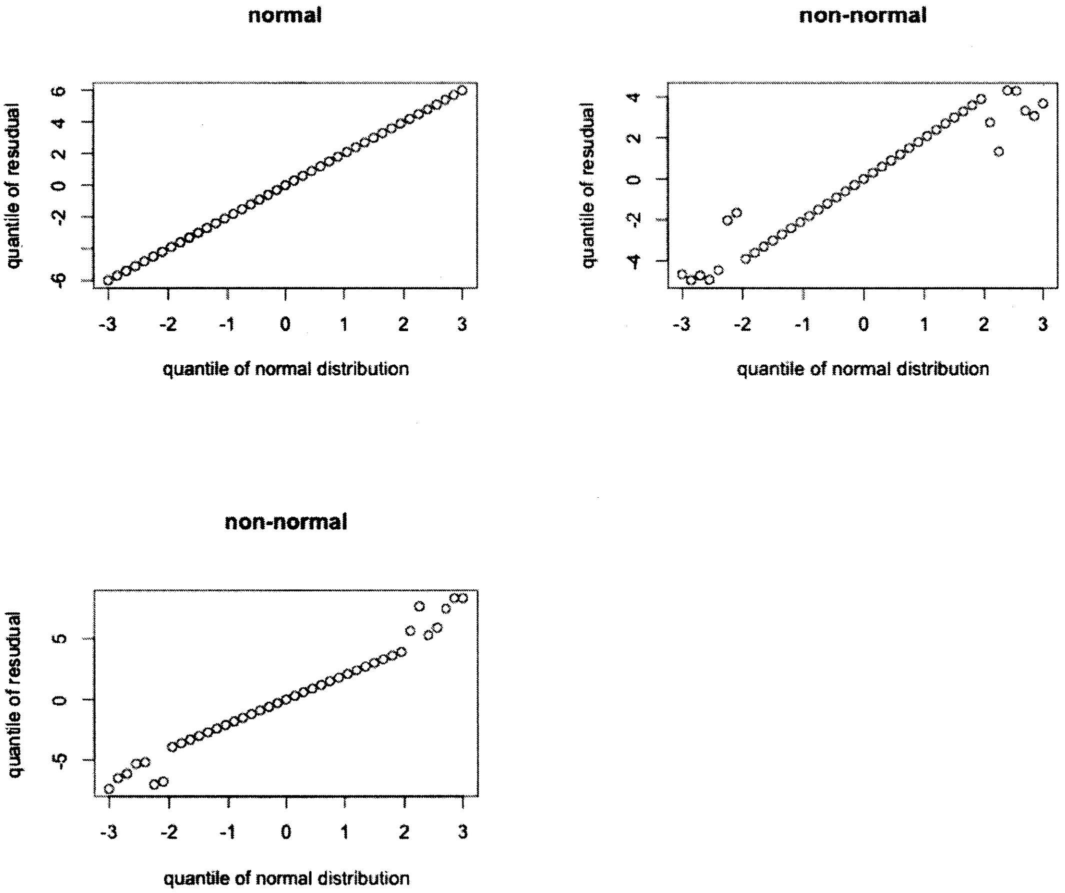

Figure B-2: Prototype of normal and non-normal regression function residual plots

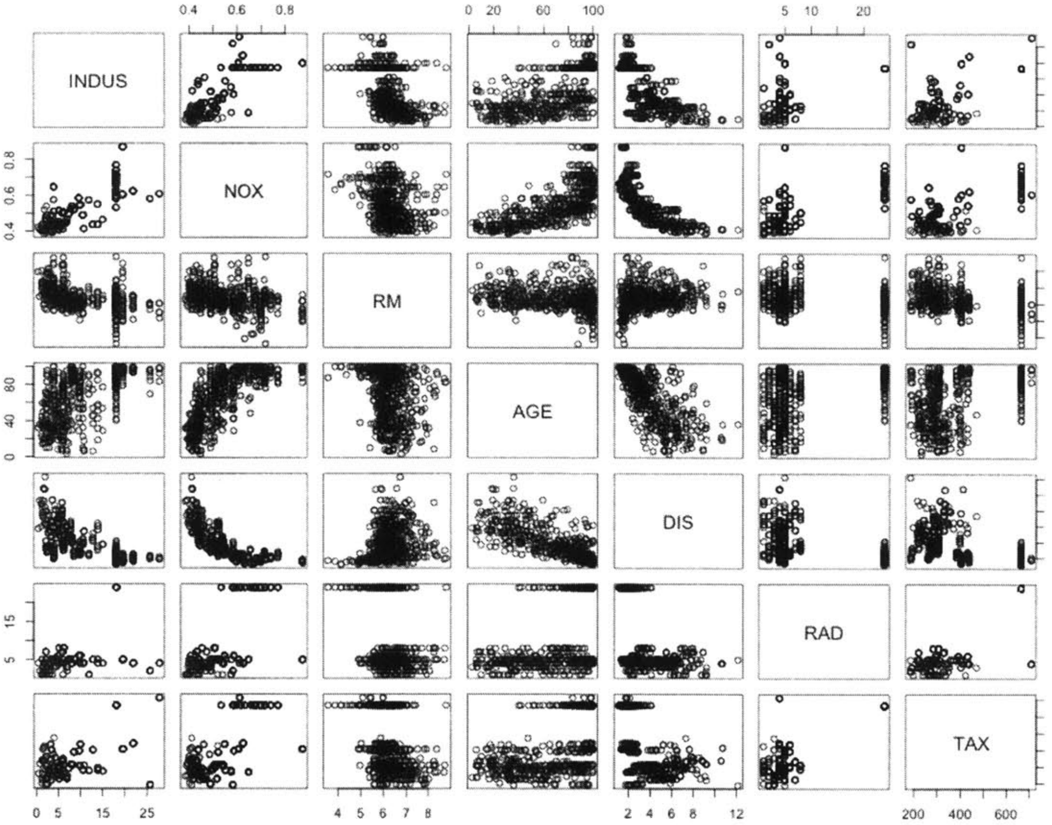

Figure B-3: Scatter plot matrix of highly correlated pairs

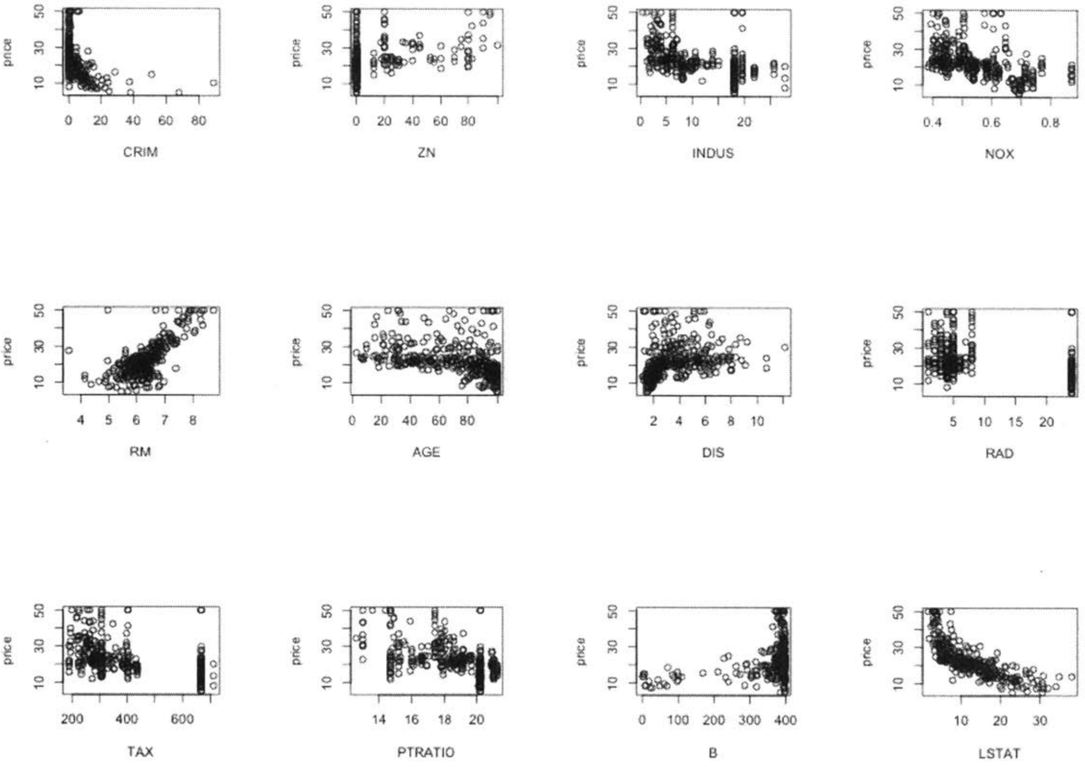

Figure B-4: Scatter plot of 'MEDV' with potential predictor variables

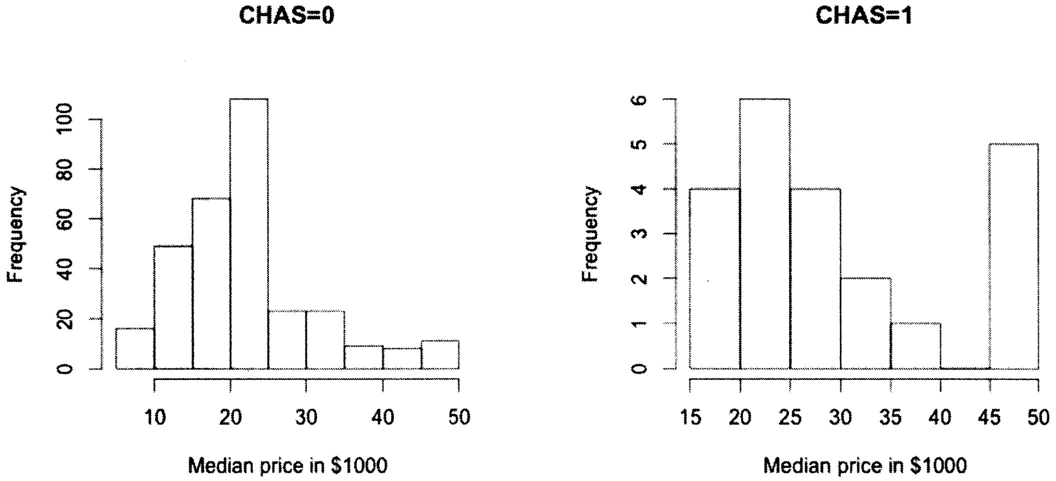

Figure B-5: Histogram of 'MEDV' for 'CHAS'=0 and 1 separately

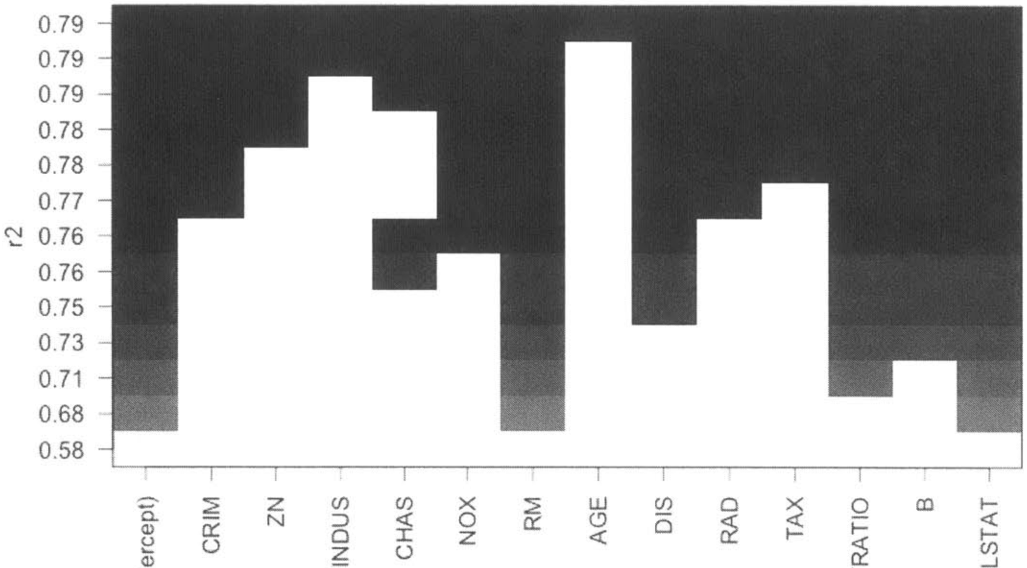

Figure B-6: 'Best' subset algorithm outcomes

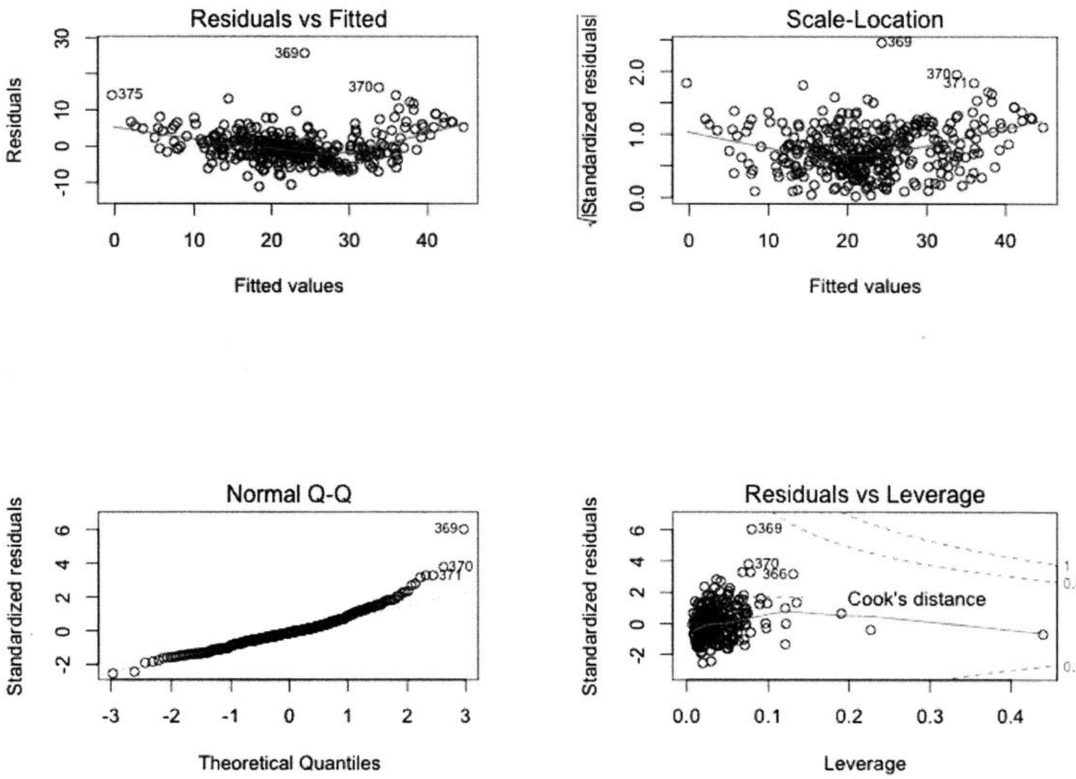

Figure B-7: Graphical diagnosis

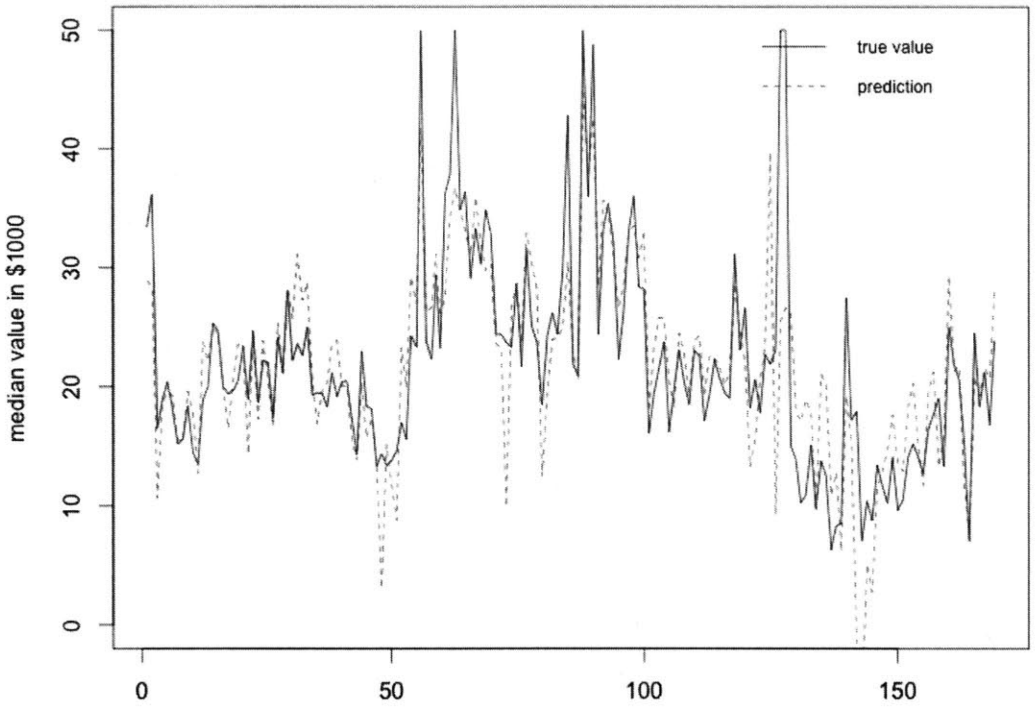

Figure B-8: Prediction result from linear regression model

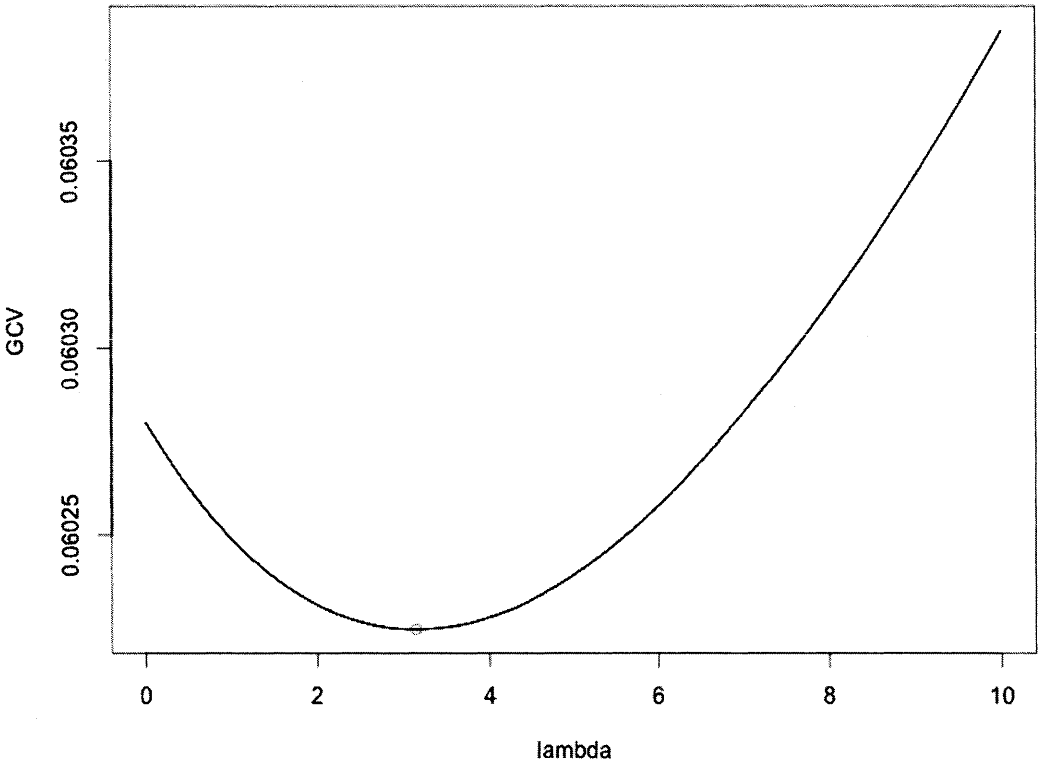

Figure B-9: Cross-validation for Ridge regression

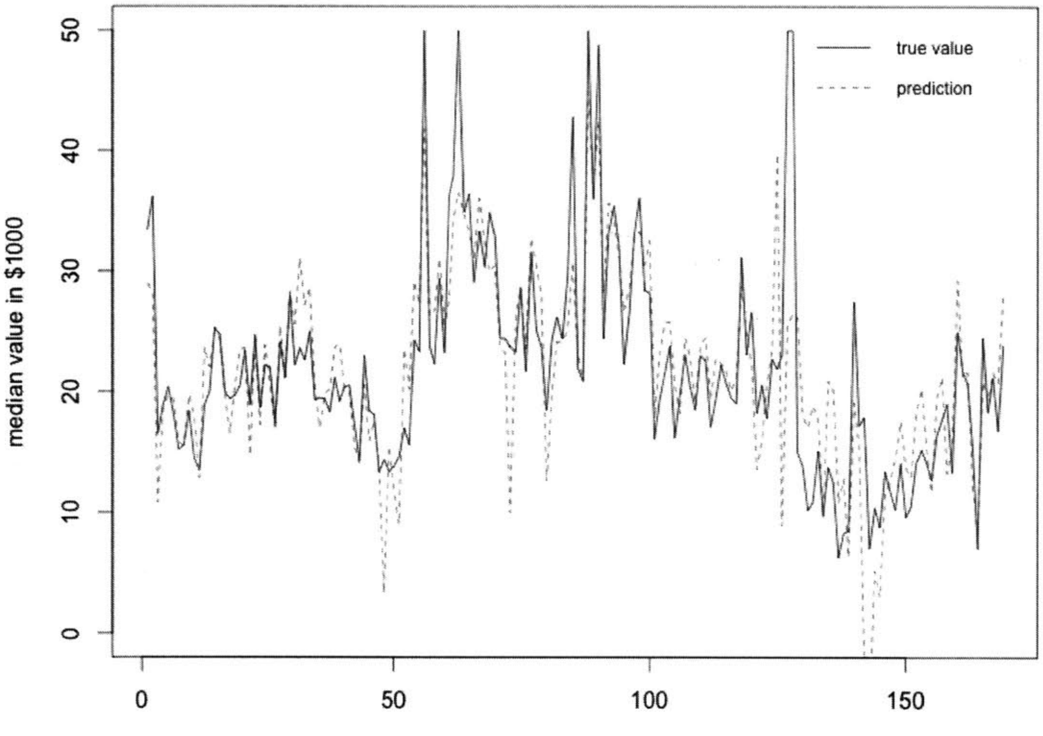

Figure B-10: Prediction result from Ridge regression model

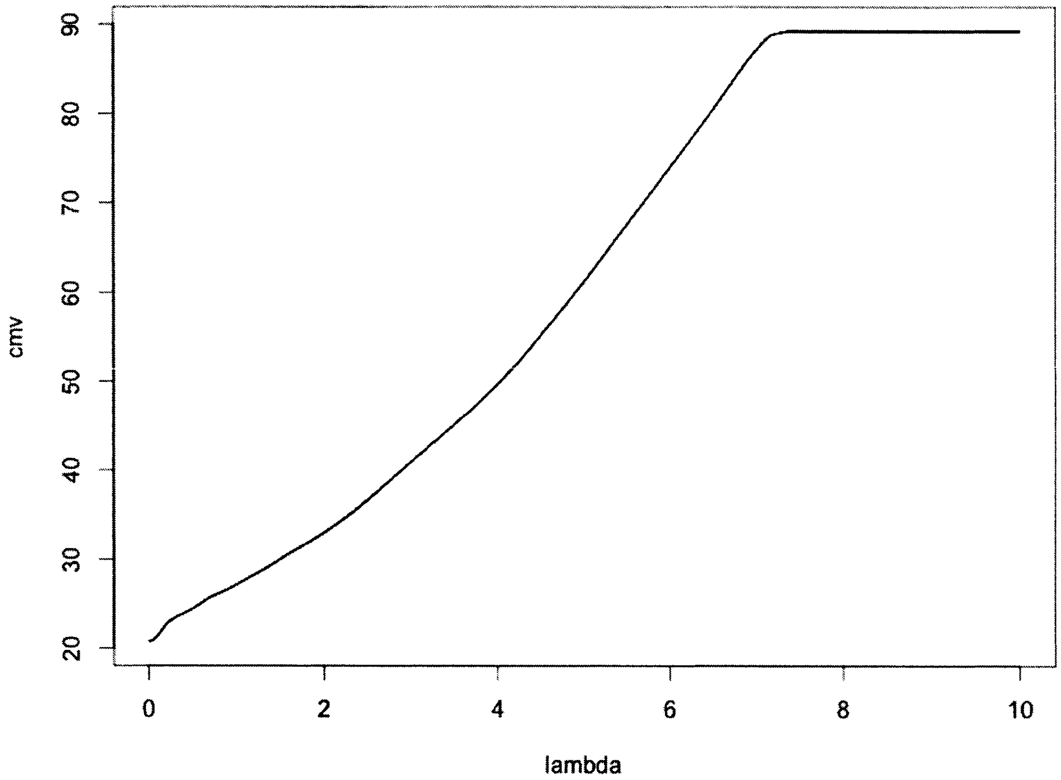

Figure B-11: Cross-validation for LASSO regression

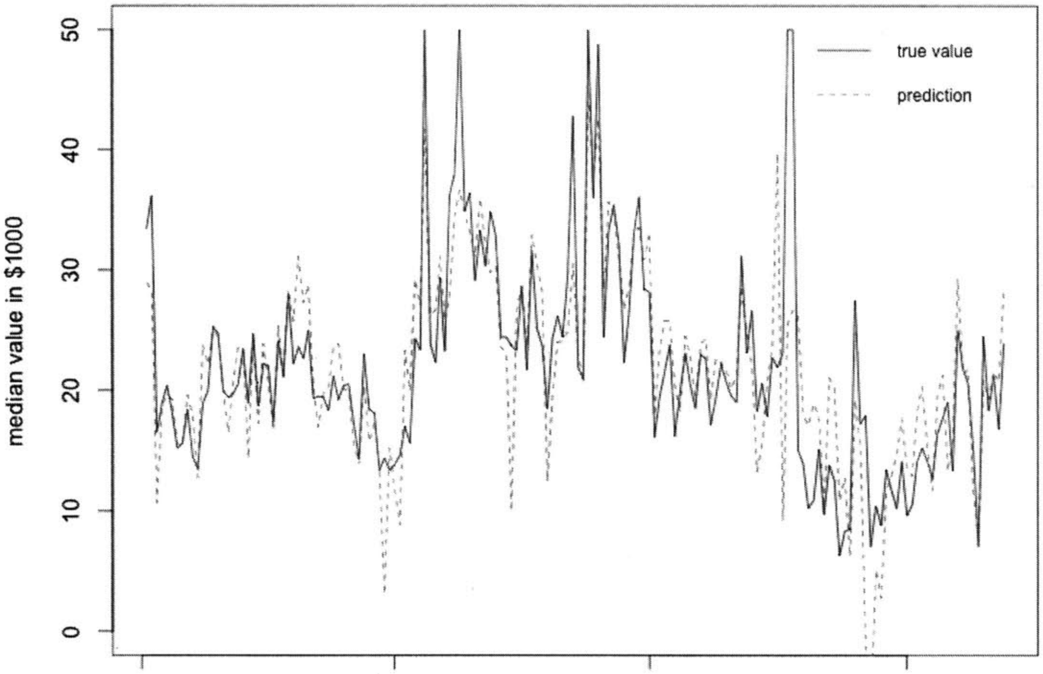

Figure B-12: Prediction result from LASSO regression model

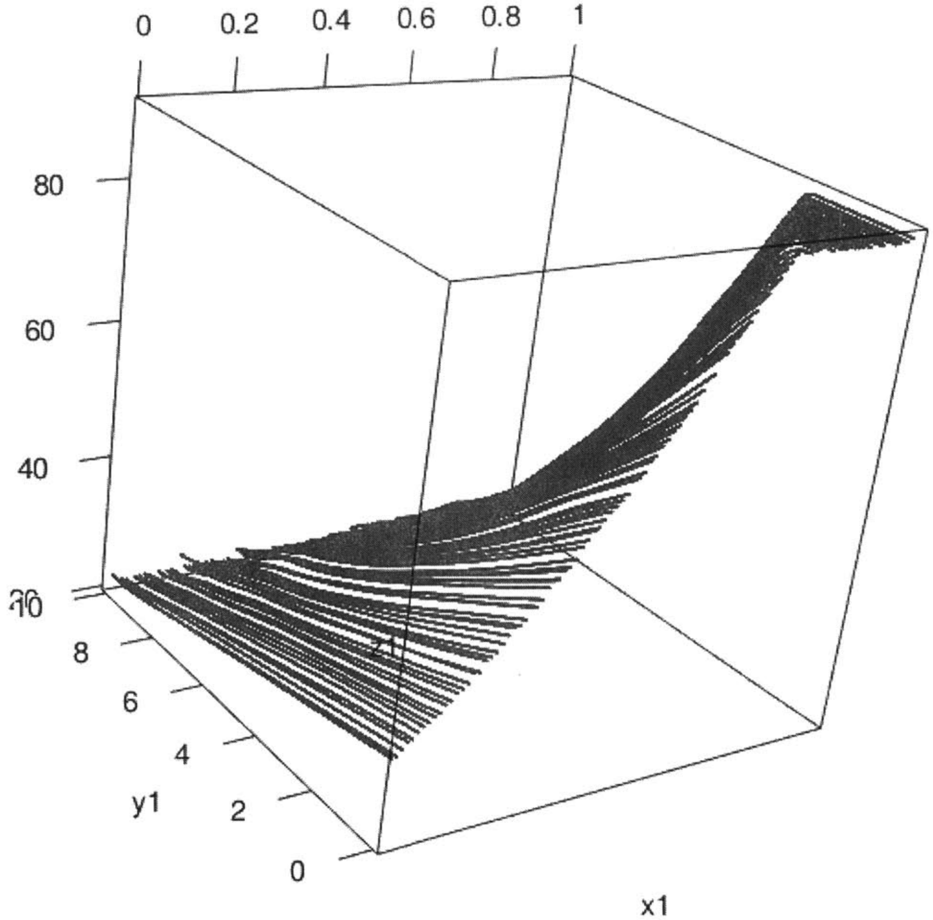

Figure B-13: 2-D Cross-validation for Elastic net

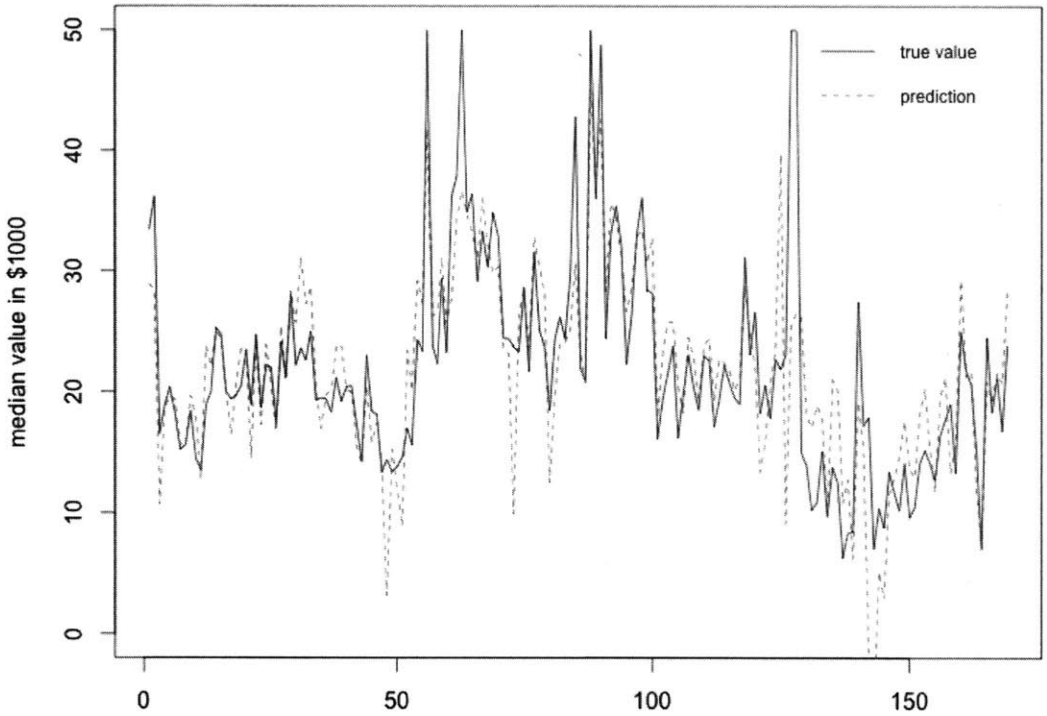

Figure B-14: Prediction result from Elastic net model

Bibliography

[1] Orley Ashenfelter and David Genesove. Testing for price anomalies in real estate auctions. Technical report, National Bureau of Economic Research, 1992.

[2] Michael A Babyak. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic medicine*, 66(3):411–421, 2004.

[3] Martin J Bailey, Richard F Muth, and Hugh O Nourse. A regression method for real estate price index construction. *Journal of the American Statistical Association*, 58(304):933–942, 1963.

[4] Raleigh Barlowe. Land resource economics: The economics of real estate. 1978.

[5] James O Berger. *Statistical decision theory and Bayesian analysis.* Springer Science & Business Media, 2013.

[6] Christopher M Bishop. *Pattern recognition and machine learning.* springer, 2006.

[7] William B Brueggeman and Jeffrey D Fisher. *Real estate finance and investments.* Irwin Homewood, IL, 1993.

[8] Bradford Case and John M Quigley. The dynamics of real estate prices. *The Review of Economics and Statistics*, pages 50–58, 1991.

[9] Karl E Case Jr, Robert J Shiller, and Allan N Weiss. Index-based futures and options markets in real estate. *The Journal of Portfolio Management*, 19(2):83–92, 1993.

[10] Emery N Castle and Irving Hoch. Farm real estate price components, 1920–78. *American Journal of Agricultural Economics*, 64(1):8–18, 1982.

[11] K Chau, S Wong, C Yiu, and H Leung. Real estate price indices in hong kong. *Journal of Real Estate Literature*, 13(3):337–356, 2005.

[12] Ping Cheng, Zhenguo Lin, and Yingchun Liu. A model of time-on-market and real estate price under sequential search with recall. *Real Estate Economics*, 36(4):813–843, 2008.

[13] David Roxbee Cox and David Oakes. *Analysis of survival data*, volume 21. CRC Press, 1984.

[14] Christine De Mol, Ernesto De Vito, and Lorenzo Rosasco. Elastic-net regularization in learning theory. *Journal of Complexity*, 25(2):201–230, 2009.

[15] Denise DiPasquale and William C Wheaton. *Urban economics and real estate markets*, volume 23. Prentice Hall Englewood Cliffs, NJ, 1996.

[16] Robin Dubin, Kelley Pace, and Thomas Thibodeau. Spatial autoregression techniques for real estate data. *Journal of Real Estate Literature*, 7(1):79–95, 1999.

[17] Eugene F Fama and G William Schwert. Asset returns and inflation. *Journal of financial economics*, 5(2):115–146, 1977.

[18] Donald E Farrar and Robert R Glauber. Multicollinearity in regression analysis: the problem revisited. *The Review of Economic and Statistics*, pages 92–107, 1967.

[19] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. glmnet: Lasso and elastic-net regularized generalized linear models. *R package version*, 1, 2009.

[20] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

[21] Wenjiang J Fu. Penalized regressions: the bridge versus the lasso. *Journal of computational and graphical statistics*, 7(3):397–416, 1998.

[22] Francis Galton. *Natural inheritance.* Macmillan, 1894.

[23] David Geltner, Bryan D MacGregor, and Gregory M Schwann. Appraisal smoothing and price discovery in real estate markets. *Urban Studies*, 40(5-6):1047–1064, 2003.

[24] David Geltner, Norman Miller, Jim Clayton, and PMA Eichholtz. Commercial real estate analysis and investments. 2013.

[25] John Geweke et al. Variable selection and model comparison in regression. *Bayesian statistics*, 5:609–620, 1996.

[26] Michael Giliberto. Equity real estate investment trusts and real estate returns. *Journal of Real Estate Research*, 5(2):259–263, 1990.

[27] Geoffrey J Goodhill and DJ Willshaw. Application of the elastic net algorithm to the formation of ocular dominance stripes. *Network: Computation in Neural Systems*, 1(1):41–59, 1990.

[28] Michael H Graham. Confronting multicollinearity in ecological multiple regression. *Ecology*, 84(11):2809–2815, 2003.

[29] Steven R Grenadier. The strategic exercise of options: Development cascades and overbuilding in real estate markets. *The Journal of Finance*, 51(5):1653–1679, 1996.

[30] Douglas M Hawkins. The problem of overfitting. *Journal of chemical information and computer sciences*, 44(1):1–12, 2004.

[31] Martin Hoesli, Carmelo Giaccotto, and Philippe Favarger. Three new real estate price indices for geneva, switzerland. *The Journal of Real Estate Finance and Economics*, 15(1):93–109, 1997.

[32] Lynn Kuo and Bani Mallick. Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 65–81, 1998.

[33] Steven D Levitt and Chad Syverson. Market distortions when agents are better informed: The value of information in real estate transactions. *The Review of Economics and Statistics*, 90(4):599–611, 2008.

[34] R Lletí, E Melêndez, MC Ortiz, LA Sarabia, and MS Sánchez. Outliers in partial least squares regression: Application to calibration of wine grade with mean infrared data. *Analytica chimica acta*, 544(1):60–70, 2005.

[35] Keith I Mahon. The new ÃŠyorkÃŞ regression: Application of an improved statistical method to geochemistry. *International Geology Review*, 38(4):293–303, 1996.

[36] Stephen Malpezzi and Susan Wachter. The role of speculation in real estate cycles. *Journal of Real Estate Literature*, 13(2):141–164, 2005.

[37] DW Mardquardt. An algorithm for least square estimation of parameters. *J. Soc. Ind. Appl. Math*, 11:431, 1963.

[38] Mike E Miles, Gayle Berens, and Marc Allan Weiss. *Real estate development: principles and process.* Urban Land Inst, 2000.

[39] Norman Miller, Michael Sklarz, and Nicholas Real. Japanese purchases, exchange rates and speculation in residential real estate markets. *Journal of Real Estate Research*, 3(3):39–49, 1988.

[40] Frederick Mosteller and John Wilder Tukey. Data analysis and regression: a second course in statistics. *Addison-Wesley Series in Behavioral Science: Quantitative Methods*, 1977.

[41] Raymond H Myers. *Classical and modern regression with applications*, volume 2. Duxbury Press Belmont, CA, 1990.

[42] John Neter, Michael H Kutner, Christopher J Nachtsheim, and William Wasserman. *Applied linear statistical models*, volume 4. Irwin Chicago, 1996.

[43] Art B Owen. A robust hybrid of lasso and ridge regression. *Contemporary Mathematics*, 443:59–72, 2007.

[44] R Kelley Pace, Ronald Barry, and CF Sirmans. Spatial statistics and real estate. *The Journal of Real Estate Finance and Economics*, 17(1):5–13, 1998.

[45] Raymond B Palmquist. Alternative techniques for developing real estate price indexes. *The Review of Economics and Statistics*, pages 442–448, 1980.

[46] Nicholas G Polson and James G Scott. Local shrinkage rules, lévy processes and regularized regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):287–311, 2012.

[47] Bertram Price. Ridge regression: Application to nonexperimental data. *Psychological Bulletin*, 84(4):759, 1977.

[48] John M Quigley. A simple hybrid model for estimating real estate price indexes. *Journal of Housing Economics*, 4(1):1–12, 1995.

[49] James O Ramsay. *Functional data analysis*. Wiley Online Library, 2006.

[50] Bertrand M Roehner. Spatial analysis of real estate price bubbles: Paris, 1984–1993. *Regional science and urban economics*, 29(1):73–88, 1999.

[51] Gregory M Schwann. A real estate price index for thin markets. *The Journal of Real Estate Finance and Economics*, 16(3):269–287, 1998.

[52] JM Stanton et al. A brief history of linear regression for statistics instructors. *Journal of Statistics Education*, 9(3), 2001.

[53] Mahlon R Straszheim. An econometric analysis of the urban housing market. *NBER Books*, 1975.

[54] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[55] Robert Tibshirani. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):273–282, 2011.

[56] Luther G Tweeten and James E Martin. A methodology for predicting us farm real estate price variation. *Journal of Farm Economics*, pages 378–393, 1966.

[57] Hansheng Wang, Guodong Li, and Guohua Jiang. Robust regression shrinkage and consistent variable selection through the lad-lasso. *Journal of Business & Economic Statistics*, 25(3):347–355, 2007.

[58] Wei-an WANG and Cong HE. Real estate price and inflation expectation [j]. *Journal of Finance and Economics*, 12:005, 2005.

[59] Qiang Wu, Yiming Ying, and Ding-Xuan Zhou. Learning rates of least-square regularized regression. *Foundations of Computational Mathematics*, 6(2):171–192, 2006.

[60] Abdullah Yavas and Shiawee Yang. The strategic role of listing price in marketing real estate: theory and evidence. *Real Estate Economics*, 23(3):347–368, 1995.

[61] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.