

Preface

In statistics, regression analysis consists of techniques for modeling the relationship between a dependent variable (also called response variable) and one or more independent variables (also known as explanatory variables or predictors). In regression, the dependent variable is modeled as a function of independent variables, corresponding regression parameters (coefficients), and a random error term which represents variation in the dependent variable unexplained by the function of the dependent variables and coefficients. In linear regression the dependent variable is modeled as a linear function of a set of regression parameters and a random error. The parameters need to be estimated so that the model gives the “ best fit ” to the data. The parameters are estimated based on predefined criterion. The most commonly used criterion is the least squares method, but other criteria have also been used that will result in different estimators of the regression parameters. The statistical properties of the estimator derived using different criteria will be different from the estimator using the least squares principle. In this book the least squares principle will be utilized to derive estimates of the regression parameters. If a regression model adequately reflects the true relationship between the response variable and independent variables, this model can be used for predicting dependent variable, identifying important independent variables, and establishing desired causal relationship between the response variable and independent variables.

To perform regression analysis, an investigator often assembles data on underlying variables of interest and employs regression model to estimate the quantitative causal effect of the independent variables to the response variable. The investigator also typically assesses the “ statistical significance ” of the estimated relationship between the independent variables and depen-

dent variable, that is, the degree of confidence on how the true relationship is close to the estimated statistical relationship.

Regression analysis is a process used to estimate a function which predicts value of response variable in terms of values of other independent variables. If the regression function is determined only through a set of parameters the type of regression is the parametric regression. Many methods have been developed to determine various parametric relationships between response variable and independent variables. These methods typically depend on the form of parametric regression function and the distribution of the error term in a regression model. For example, linear regression, logistic regression, Poisson regression, and probit regression, etc. These particular regression models assume different regression functions and error terms from corresponding underline distributions. A generalization of linear regression models has been formalized in the “generalized linear model” and it requires to specify a link function which provides the relationship between the linear predictor and the mean of the distribution function.

The regression model often relies heavily on the underlying assumptions being satisfied. Regression analysis has been criticized as being misused for these purposes in many cases where the appropriate assumptions cannot be verified to hold. One important factor for such criticism is due to the fact that a regression model is easier to be criticized than to find a method to fit a regression model (Cook and Weisberg (1982)). However, checking model assumptions should never be overlooked in regression analysis.

By saying much about regression model we would like to go back to the purpose of this book. The goal of the book is to provide a comprehensive, one-semester textbook in the area of regression analysis. The book includes carefully selected topics and will not assume to serve as a complete reference book in the area of regression analysis, but rather as an easy-to-read textbook to provide readers, particularly the graduate students majoring in either statistics or biostatistics, or those who use regression analysis substantially in their subject fields, the fundamental theories on regression analysis, methods for regression model diagnosis, and computing techniques in regression. In addition to carefully selected classical topics for regression analysis, we also include some recent developments in the area of regression analysis such as the least absolute shrinkage and selection operator (LASSO) proposed by Tibshirani (1996) and Bayes averaging method.

The topics on regression analysis covered in this book are distributed among 9 chapters. Chapter 1 briefly introduces the basic concept of regression and defines the linear regression model. Chapters 2 and 3 cover the simple linear regression and multiple linear regression. Although the simple linear regression is a special case of the multiple linear regression, we present it without using matrix and give detailed derivations that highlight the fundamental concepts in linear regression. The presentation of multiple regression focus on the concept of vector space, linear projection, and linear hypothesis test. The theory of matrix is used extensively for the proofs of the statistical properties of linear regression model. Chapters 4 through 6 discuss the diagnosis of linear regression model. These chapters cover outlier detection, influential observations identification, collinearity, confounding, regression on dummy variables, checking for equal variance assumption, graphical display of residual diagnosis, and variable transformation technique in linear regression analysis. Chapters 7 and 8 provide further discussions on the generalizations of the ordinary least squares estimation in linear regression. In these two chapters we discuss how to extend the regression model to situation where the equal variance assumption on the error term fails. To model the regression data with unequal variance the generalized least squares method is introduced. In Chapter 7, two shrinkage estimators, the ridge regression and the LASSO are introduced and discussed. A brief discussion on the least squares method for nonlinear regression is also included. Chapter 8 briefly introduces the generalized linear models. In particular, the Poisson Regression for count data and the logistic regression for binary data are discussed. Chapter 9 briefly discussed the Bayesian linear regression models. The Bayes averaging method is introduced and discussed.

The purpose of including these topics in the book is to foster a better understanding of regression modeling. Although these topics and techniques are presented largely for regression, the ideas behind these topics and the techniques are also applicable in other areas of statistical modeling. The topics presented in the book cover fundamental theories in linear regression analysis and we think that they are the most useful and relevant to the future research into this area. A thorough understanding of the basic theories, model diagnosis, and computing techniques in linear regression analysis is necessary for those who would like to learn statistics either as a discipline or as a substantial tool in their subject field. To this end, we provide detailed proofs of fundamental theories related to linear regression modeling, diagnosis, and computing so that readers can understand

the methods in regression analysis and actually model the data using the methods presented in the book.

To enable the book serves the intended purpose as a graduate textbook for regression analysis, in addition to detailed proofs, we also include many examples to illustrate relevant computing techniques in regression analysis and diagnosis. We hope that this would increase the readability and help to understand the regression methods for students who expect a thorough understanding of regression methods and know how to use these methods to solve for practical problems. In addition, we tried to avoid an oversized-textbook so that it can be taught in one semester. We do not intend to write a complete reference book for regression analysis because it will require a significantly larger volume of the book and may not be suitable for a textbook of regression course. In our practice we realize that graduate students often feel overwhelming when try to read an oversized textbook. Therefore, we focus on presenting fundamental theories and detailed derivations that can highlight the most important methods and techniques in linear regression analysis.

Most computational examples of regression analysis and diagnosis in the book use one of popular software package the Statistical Analysis System (SAS), although readers are not discouraged to use other statistical software packages in their subject area. Including illustrative SAS programs for the regression analysis and diagnosis in the book is to help readers to become familiar with various computing techniques that are necessary to regression analysis. In addition, the SAS Output Delivery System (ODS) is introduced to enable readers to generate output tables and figures in a desired format. These illustrative programs are often arranged in the end of each chapter with brief explanations. In addition to the SAS, we also briefly introduce the software R which is a freeware. R has many user-written functions for implementation of various statistical methods including regression. These functions are similar to the built-in functions in the commercial software package S-PLUS. We provide some programs in R to produce desired regression diagnosis graphs. Readers are encouraged to learn how to use the software R to perform regression analysis, diagnosis, and producing graphs.

X. Yan and X. G. Su