

Chapter 1

Introduction

1.1 Regression Model

Researchers are often interested in the relationships between one variable and several other variables. For example, does smoking cause lung cancer? Following Table 1.1 summarizes a study carried out by government statisticians in England. The data concern 25 occupational groups and are condensed from data on thousands of individual men. One variable is smoking ratio which is a measure of the number of cigarettes smoked per day by men in each occupation relative to the number smoked by all men of the same age. Another variable is the standardized mortality ratio. To answer the question that does smoking cause cancer we may like to know the relationship between the derived mortality ratio and smoking ratio. This falls into the scope of regression analysis. Data from a scientific

Table 1.1 Smoking and Mortality Data

Smoking	77	112	137	113	117	110	94	125	116	133
Mortality	84	96	116	144	123	139	128	113	155	146
Smoking	102	115	111	105	93	87	88	91	102	100
Mortality	101	128	118	115	113	79	104	85	88	120
Smoking	91	76	104	66	107					
Mortality	104	60	129	51	86					

experiment often lead to ask whether there is a causal relationship between two or more variables. Regression analysis is the statistical method for investigating such relationship. It is probably one of the oldest topics in the area of mathematical statistics dating back to about two hundred years

ago. The earliest form of the linear regression was the least squares method, which was published by Legendre in 1805, and by Gauss in 1809. The term “least squares” is from Legendre’s term. Legendre and Gauss both applied the method to the problem of determining, from astronomical observations, the orbits of bodies about the sun. Euler had worked on the same problem (1748) without success. Gauss published a further development of the theory of least squares in 1821, including a version of the today’s well-known Gauss-Markov theorem, which is a fundamental theorem in the area of the general linear models.

What is a statistical model? A statistical model is a simple description of a state or process. “A model is neither a hypothesis nor a theory. Unlike scientific hypotheses, a model is not verifiable directly by an experiment. For all models of true or false, the validation of a model is not that it is “true” but that it generates good testable hypotheses relevant to important problems.” (R. Levins, *Am. Scientist* 54: 421-31, 1966)

Linear regression requires that model is linear in regression parameters. Regression analysis is the method to discover the relationship between one or more response variables (also called dependent variables, explained variables, predicted variables, or regressands, usually denoted by y) and the predictors (also called independent variables, explanatory variables, control variables, or regressors, usually denoted by x_1, x_2, \dots, x_p).

There are three types of regression. The first is the simple linear regression. The simple linear regression is for modeling the linear relationship between two variables. One of them is the dependent variable y and another is the independent variable x . For example, the simple linear regression can model the relationship between muscle strength (y) and lean body mass (x). The simple regression model is often written as the following form

$$y = \beta_0 + \beta_1 x + \varepsilon, \tag{1.1}$$

where y is the dependent variable, β_0 is y intercept, β_1 is the gradient or the slope of the regression line, x is the independent variable, and ε is the random error. It is usually assumed that error ε is normally distributed with $E(\varepsilon) = 0$ and a constant variance $\text{Var}(\varepsilon) = \sigma^2$ in the simple linear regression.

The second type of regression is the multiple linear regression which is a linear regression model with one dependent variable and more than one

independent variables. The multiple linear regression assumes that the response variable is a linear function of the model parameters and there are more than one independent variables in the model. The general form of the multiple linear regression model is as follows:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon, \quad (1.2)$$

where y is dependent variable, $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ are regression coefficients, and x_1, x_2, \dots, x_n are independent variables in the model. In the classical regression setting it is usually assumed that the error term ε follows the normal distribution with $E(\varepsilon) = 0$ and a constant variance $\text{Var}(\varepsilon) = \sigma^2$.

Simple linear regression is to investigate the linear relationship between one dependent variable and one independent variable, while the multiple linear regression focuses on the linear relationship between one dependent variable and more than one independent variables. The multiple linear regression involves more issues than the simple linear regression such as collinearity, variance inflation, graphical display of regression diagnosis, and detection of regression outlier and influential observation.

The third type of regression is nonlinear regression, which assumes that the relationship between dependent variable and independent variables is not linear in regression parameters. Example of nonlinear regression model (growth model) may be written as

$$y = \frac{\alpha}{1 + e^{\beta t}} + \varepsilon, \quad (1.3)$$

where y is the growth of a particular organism as a function of time t , α and β are model parameters, and ε is the random error. Nonlinear regression model is more complicated than linear regression model in terms of estimation of model parameters, model selection, model diagnosis, variable selection, outlier detection, or influential observation identification. General theory of the nonlinear regression is beyond the scope of this book and will not be discussed in detail. However, in addition to the linear regression model we will discuss some generalized linear models. In particular, we will introduce and discuss two important generalized linear models, logistic regression model for binary data and log-linear regression model for count data in Chapter 8.

1.2 Goals of Regression Analysis

Regression analysis is one of the most commonly used statistical methods in practice. Applications of regression analysis can be found in many scientific fields including medicine, biology, agriculture, economics, engineering, sociology, geology, etc. The purposes of regression analysis are three-folds:

- (1) Establish a casual relationship between response variable y and regressors x_1, x_2, \dots, x_n .
- (2) Predict y based on a set of values of x_1, x_2, \dots, x_n .
- (3) Screen variables x_1, x_2, \dots, x_n to identify which variables are more important than others to explain the response variable y so that the causal relationship can be determined more efficiently and accurately.

An analyst often follows, but not limited, the following procedures in the regression analysis.

- (1) The first and most important step is to understand the real-life problem which is often fallen into a specific scientific field. Carefully determine whether the scientific question falls into scope of regression analysis.
- (2) Define a regression model which may be written as

Response variable = a function of regressors + random error,
or simply in a mathematical format

$$y = f(x_1, x_2, \dots, x_p) + \varepsilon.$$

You may utilize a well-accepted model in a specific scientific field or try to define your own model based upon a sound scientific judgement by yourself or expert scientists in the subject area. Defining a model is often a joint effort among statisticians and experts in a scientific discipline. Choosing an appropriate model is the first step for statistical modeling and often involve further refinement procedures.

- (3) Make distributional assumptions on the random error ε in the regression model. These assumptions need to be verified or tested by the data collected from experiment. The assumptions are often the basis upon which the regression model may be solved and statistical inference is drawn.
- (4) Collect data y and x_1, x_2, \dots, x_p . This data collection step usually involves substantial work that includes, but not limited to, experimental design, sample size determination, database design, data cleaning,

and derivations of analysis variables that will be used in the statistical analysis. In many real-life applications, this is a crucial step that often involve significant amount of work.

- (5) According to the software used in the analysis, create data sets in an appropriate format that are easy to be read into a chosen software. In addition, it often needs to create more specific analysis data sets for planned or exploratory statistical analysis.
- (6) Carefully evaluate whether or not the selected model is appropriate for answering the desired scientific questions. Various diagnosis methods may be used to evaluate the performance of the selected statistical model. It should be kept in mind that the model diagnosis is for the judgment of whether the selected statistical model is a sound model that can answer the desired scientific questions.
- (7) If the model is deemed to be appropriate according to a well accepted model diagnosis criteria, it may be used to answer the desired scientific questions; otherwise, the model is subject to refinement or modification. Several iterations of model selection, model diagnosis, and model refinement may be necessary and very common in practice.

1.3 Statistical Computing in Regression Analysis

After a linear regression model is chosen and a database is created, the next step is statistical computing. The purposes of the statistical computing are to solve for the actual model parameters and to conduct model diagnosis. Various user-friendly statistical softwares have been developed to make the regression analysis easier and more efficient.

Statistical Analysis System (SAS) developed by SAS Institute, Inc. is one of the popular softwares which can be used to perform regression analysis. The SAS System is an integrated system of software products that enables users to perform data entry and data management, to produce statistical graphics, to conduct wide range of statistical analyses, to retrieve data from data warehouse (extract, transform, load) platform, and to provide dynamic interface to other software, etc. One great feature of SAS is that many standard statistical methods have been integrated into various SAS procedures that enable analysts easily find desired solutions without writing source code from the original algorithms of statistical methods. The SAS “macro” language allows user to write subroutines to perform particular user-defined statistical analysis. SAS compiles and runs on UNIX platform

and Windows operating system.

Software S-PLUS developed by Insightful Inc. is another one of the most popular softwares that have been used substantially by analysts in various scientific fields. This software is a rigorous computing tool covering a broad range of methods in statistics. Various built-in S-PLUS functions have been developed that enable users to perform statistical analysis and generate analysis graphics conveniently and efficiently. The S-PLUS offers a wide collection of specialized modules that provide additional functionality to the S-PLUS in areas such as: volatility forecasting, optimization, clinical trials analysis, environmental statistics, and spatial data analysis, data mining. In addition, user can write S-PLUS programs or functions using the S-language to perform statistical analysis of specific needs. The S-PLUS compiles and runs on UNIX platform and Windows operating system.

Statistical Package for Social Sciences (SPSS) is also one of the most widely used softwares for the statistical analysis in the area of social sciences. It is one of the preferred softwares used by market researchers, health researchers, survey companies, government, education researchers, among others. In addition to statistical analysis, data management (case selection, file reshaping, creating derived data) and data documentation (a metadata dictionary is stored) are features of the SPSS. Many features of SPSS are accessible via pull-down menus or can be programmed with a proprietary 4GL command syntax language. Additionally, a “macro” language can be used to write command language subroutines to facilitate special needs of user-desired statistical analysis.

Another popular software that can be used for various statistical analyses is R. R is a language and environment for statistical computing and graphics. It is a GNU project similar to the S language and environment. R can be considered as a different implementation of S language. There are some important differences, but much code written for S language runs unaltered under R. The S language is often the vehicle of choice for research in statistical methodology, and R provides an open source route to participation in that activity. One of R’s strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. Great care has been taken over the defaults for the design choices in graphics, but user retains full control. R is available as free software under the terms of the Free Software Foundation’s GNU General Public License in source code form. It compiles and runs on a wide

variety of UNIX platforms and similar system Linux, as well as Windows.

Regression analysis can be performed using various softwares such as SAS, S-PLUS, R, or SPSS. In this book we choose the software SAS to illustrate the computing techniques in regression analysis and diagnosis. Extensive examples are provided in the book to enable readers to become familiar with regression analysis and diagnosis using SAS. We also provide some examples of regression graphic plots using the software R. However, readers are not discouraged to use other softwares to perform regression analysis and diagnosis.