Chapter 1

Examples and Limits of the GLM

1.1 Motivation

In this chapter we do three things. First, we describe the structure of the book and give suggestions on how a student might use it. Second, we review basic statistical ideas in order to introduce the definition of the General Linear Model (GLM). Third, and finally, we provide examples that demonstrate some of the value of the General Linear Model.

This book contains five modules:

I. Basic theory (chapters 1–3)
II. Multiple regression (chapters 4–6)
III. Model building and evaluation (chapters 7–11)
IV. ANOVA (chapters 12–15)
V. General topics, including ANCOVA and power (chapters 16–17).

The first module contains essentially all of the theory for the entire book. Hence the student should not be discouraged or misled by the initial emphasis on formulas. The remainder of the book centers on applications, driven by data and the need to answer particular scientific questions. The beauty and the value of the General Linear Model arise from its wide range of application. As one learns more, one has to remember less.

The exercises provide an essential part of this book. The authors believe that effective learning requires completing all of the exercises in this book. Except for those in chapters 1 and 17, the exercises allow students to use data from their own field. Appendix D contains many alternate sets of data. All numeric examples in the text have been computed with simulated data in order to ensure knowledge of the underlying model. In contrast, all exercises use “real” data, collected in the course of scientific research. Only real data have the richness that creates the wonderful excitement and uncertainty inherent to the scientific process.

1.2 A Review of Basic Statistical Ideas

1.2.1 Variables

A statistical model describes relationships between *variables*. The *scale of measurement* (Stevens 1946, 1951) of a variable constrains and guides statistical analysis and statistical modeling. Four types of scales may be distinguished. *Nominal* scales have values that only name objects. For example, a person's blood type, A, B, O, and so on, provides a nominal scale. An *ordinal* scale gives order to objects. For example, the finishing position of runners in a 100m dash only ranks the competitors. An *interval* scale has values for which distances between them carry meaning. For example, measuring the decrease in light level during a solar eclipse provides an interval scale. A *ratio* scale has values whose ratios carry meaning. For example, a person's body weight measured in kilograms provides a ratio scale.

The interpretation of a zero value reflects the distinction between an interval and ratio scale. For example, measuring temperature in degrees centigrade reflects an interval scale and an arbitrary zero point. In contrast, measuring temperature on the kelvin scale reflects a ratio scale with a "true" zero, indicating none of the property. Absolute zero (0° kelvin) corresponds to cessation of all molecular motion, while 0° centigrade merely corresponds to the freezing point of water. Velleman and Wilkinson (1993) provided useful discussion and a well-reasoned warning about taking a description of scale too seriously in choosing a data analysis.

Certain other descriptors are often associated with the various types of variables: categorical (nominal), rank (ordinal), and continuous (interval or ratio). In the context of statistical analysis, the value of the distinction between interval and ratio scales lies in the customary need to transform ratio scale variables. The error variance of a measurement of biological extent, such as mass, tends to be proportional to the size of the measurement. Consequently, as the range of values for a ratio scale increases, the likelihood of needing to transform the data in order to homogenize variance increases.

Building models may involve a number of types of variables. The *response* variable reflects the outcome to be modeled. Any variable used as input to the model provides a *predictor*. Such predictors may involve *fixed* values, like gender, and therefore may be referred to as *fixed effects* (or fixed predictors). In contrast, predictors with *random* values, such as familial likeness, are referred to as *random effects* (or random predictors). In many cases only the impact of such a random effect on the response can be observed. In some cases, such values are assumed to follow a Gaussian distribution with mean zero. Nuisance variables do not lie at the focus of the current research, yet they contribute to predicting the response. Synonyms for nuisance include *control* (a term often used in the behavioral sciences) and *confounder* (a term often used in epidemiology).

An *experiment* involves random assignment to a predictor value (called the *treatment level*), while an *observational* study (for example, a survey) does not. The terms *independent* variable and *dependent* variable are reserved to describe the predictor and response variables for an experiment.

The practice of statistics begins with the definition of a *population*, defined as any set of interest. Any subset of a population forms a *sample*. Note that no standard of sampling quality is implied. Everyone in fifth grade in a single city provides a sample of all fifth graders in the nation, but perhaps a misleading one. Similarly, a *parameter* defines any property of a population, while a *statistic* defines any property of a sample.

The axiomatic mathematical system referred to as *probability theory* provides the tools needed for describing the likelihood of observing a particular sample (one *event* in the population) and associated statistic. The validity of such calculations depends strongly on the accuracy of the assumptions about the population, the *model*. A careful construction of probability theory starts with a detailed consideration of sets and properties of collections of sets. A *random variable* maps sets (events) into real numbers, and thereby defines a set function. As with any function, each input produces only one output, although more than one input may produce the same output. These definitions are intended to remind the reader of the feel of the theory. Many technical issues are ignored here for the sake of brevity.

1.2.2 Statistical Activities

One common notation for the study of probability uses Greek letters for parameters, uppercase Roman letters for random variables, and lowercase Roman letters for realizations of random variables (particular sample values). This convention conflicts with the need to distinguish between scalars, vectors, and matrices. The dominance of matrix expressions throughout this book means that the latter requirement dominates. Standard probability notation is used only where it does not conflict with matrix properties. Hence the reader must often distinguish fixed from random, known from unknown, observed from unobserved, and observable from unobservable via the context of the discussion. When in doubt about a particular item, simply search backward in the text to discover where the concept was introduced, or consult the index.

Estimation provides a value, $\hat{\theta}$, the estimate, thought to be a good approximation of a parameter, θ. Many optimality criteria for estimation have been proposed, including consistency ($\hat{\theta} \rightarrow \theta$ as $N \rightarrow \infty$), unbiased ($\mathcal{E}(\hat{\theta}) = \theta$ for any N), least squares ($\hat{\theta}$ minimizes $\sum_{i=1}^{N} \hat{e}_i^2$), and maximum likelihood ($\hat{\theta}$ maximizes the probability of the sample). The choice of criterion and estimator depends upon the nature of the parameter and the purpose of the analysis.

In the context of current statistical practice, testing a hypothesis corresponds to specifying the probability of some guess (hypothesis) about a state of nature (a parameter). Hypothesis testing comprises one part of statistical inference, as does the topic of erecting confidence intervals.

Exploratory analysis includes any result that depends on the data at hand. Confirmatory analysis requires choosing the variables, model, and hypothesis test without knowledge of the data at hand. See Muller, Barton, and Benignus 1984 for further discussion of the distinction.

Unlike probability theory, all of statistical theory and practice concerns either estimation or inference. Recognition of the importance of personal and institutional values in such activities led to the study of decision theory, which blends mathematics, philosophy (including especially ethics), economics, and psychology. As of this writing, the most common approaches to making decisions with statistical tools may be described as Neyman-Pearson, Bayesian, or Fisherian (each name refers to statisticians identified with the approach). The most popular, the Neyman-Pearson approach, is used throughout this book. Consult Dawid 1983, Fraser 1983, and Koch and Gillings 1983 for further discussion.

Each approach has appealing properties and disturbing limitations. Hopefully more general methods will be developed and resolve the conflicts plaguing discussion of statistical practice. Until that time, the reader should recognize that the richness of real data usually provides many paths to the proper conclusion. Statistics provides tools for the scientist. At the end of analysis, the scientist must decide, Does this drug help? Will this airplane fly? The accumulation and quantification of information across studies remains at the heart of science, yet this practice has only recently gained the attention of statisticians.

1.2.3 Error Rates: Evaluating Tests

One traditional interpretation of the Neyman-Pearson approach considers four decision probabilities, as represented in Table 1.2.1. Each concerns either the null hypothesis, H_0, or the alternative, H_A. This format is mostly ignored in this book. The concepts of type I and type II errors remain useful and in vogue. However, current practice in mathematical

TABLE 1.2.1 Traditional Approach (Not Used in This Text)

Truth	Decision	
	H_0: No Effect	H_A: Effect
H_0: No Effect	Pr{Correct Negative} = $(1 - \alpha)$	Pr{False Positive} = Pr{Type I error} = α
H_A: Effect	Pr{False Negative} = Pr{Type II error} = β	Pr{Correct Positive} = $(1 - \beta)$ = Power

statistics involves a more general definition for power (Hogg and Craig 1995, 285). Statistical power is defined as the probability of rejecting the null hypothesis, whether or not the null is true. For the tests considered in this book, power under the null becomes α. This more general approach has philosophical and mathematical advantages. See chapter 17 for a detailed introduction to power analysis of linear models.

1.3 GLM Definition

The name *General Linear Model* (GLM) represents its features: *General* indicates the wide applicability of the model to problems of estimation and testing of hypotheses about parameters. *Linear* refers to the regression function, $\mathcal{E}[y_i|\text{row}_i(\boldsymbol{X})]$, being a linear function of the parameters. A *model* provides a description of the relationship between (one) response and (many) predictor variables. In this book we consider only models with a single response, although such univariate models may include many predictors. The model is a statistical model, expressing observable random variables as a function of unobservable constants and unobservable random variables.

The simplest model function has one predictor:

$$y_i = \beta_0 + x_i\beta_1 + e_i\,. \tag{1.3.1}$$

The corresponding regression function takes the form

$$\mathcal{E}y_i = \beta_0 + x_i\beta_1. \tag{1.3.2}$$

Note that a *function* maps each input to one and only one output, although different inputs may lead to the same output. A function may be many to one, or one to one, but not one to many.

1.4 GLM Examples

It has been suggested that the distance from one outstretched fingertip to the other equals one's height. This corresponds to the simple statement

$$\text{Wingspan} = \text{Height}, \tag{1.4.1}$$

which may be seen to be a special case of (1.3.2), with $\beta_0 = 0$, $\beta_1 = 1$, and no error. Consider

$$\begin{aligned}\mathcal{E}y_i &= \qquad x_i \\ &= 0 \;\; + x_i \cdot 1 \\ &= \beta_0 + x_i\beta_1.\end{aligned} \tag{1.4.2}$$

Allowing for an unobservable, additive random error, specific to the person, leads to a special case of model (1.3.1), namely,

$$y_i = 0 + x_i \cdot 1 + e_i. \tag{1.4.3}$$

An obvious question arises: how might one judge the validity of reducing model (1.3.1) to (1.4.3)? This corresponds to testing $H_0 : \beta_0 = 0$ and $H_0 : \beta_1 = 1$ simultaneously. An obvious start would be to collect observations (Wingspan and Height measurements for each person) and plot the measurement pairs.

The simple model described above may be generalized in many ways:

Example 2: Add gender as a predictor (which allows distinct intercepts).
Example 3: Allow different slopes for different genders.
Example 4: Add age as a predictor.
Example 5: Use only ethnic origin as a predictor.

The corresponding models may be described in the following ways:

Model for Example 2: analysis of covariance (ANCOVA).
Model for Example 3: full model in every cell.
Model for Example 4: multiple regression with two continuous predictors.
Model for Example 5: one-way analysis of variance (ANOVA).

1.5 Student Goals

This book was written to help the student of linear models achieve two related goals. The first goal seems obvious: learn to analyze and interpret linear models of interval-scale responses. Equally important, the student should strive to understand the associated theory well enough to know when *not* to apply the methods.

The structure of the book helps to meet these goals. In order to provide a compact yet careful treatment of both practice and theory, essentially all proofs are omitted. The text of the book may thus serve as a convenient reference for a data analyst familiar with the GLM. However, in order to learn the material, the reader must also work all of the exercises (except for the few indicated as optional). Although numerous and time consuming, the exercises are condensed and sequenced so as to minimize the total amount of work. Consequently, skipping exercises sometimes makes later ones unnecessarily more difficult. Only data from actual scientific research has been used for exercises, while the few examples in the text all use simulated data. Using simulated data ensures the validity of any claims about the examples. However, simulated data never create the challenge or provide the enjoyment of real data. The exercises are designed to allow substitution of data from the student's own research; likewise, the instructor can tailor the work to student needs.

1.6 Homework Exercises

1.1 Consider consulting with a nephrologist (physician specialized in kidney disease) who asks for assistance in understanding a number of basic concepts in statistics. Throughout this homework, use examples and situations relevant to the person with whom you are consulting.

Give examples of the following types of measurement scales.

1.1.1 Nominal

1.1.2 Ordinal

1.1.3 Interval

1.1.4 Ratio

1.2 The NIDDK (U.S. National Institute of Diabetes, Digestive, and Kidney Diseases) has an ongoing study of the natural history of diabetes. Adults were enrolled at a number of sites, scattered across the United States, with varying demographics. Particular interest centers on the impact of lifestyle and diet. Subjects are examined repeatedly over time, for many years.

1.2.1 Describe the apparent target population of interest for the sponsors.

1.2.2 What samples were taken?

1.2.3 Describe one parameter that seemed of interest to the sponsors.

1.2.4 What statistic might be sensibly used to estimate the parameter?

1.3 Consider a nephrologist studying the impact of adding a new drug to current therapy for the treatment of a glomerulopathy (disease of the filter system in the kidney). The physician decides to measure creatinine level in the blood prior to treatment and again after the treatment, with the aim of reducing creatinine level.

1.3.1 In testing the hypothesis of equal decrement in both groups, describe the decision made (in terms of patient health) if a type I error has been made.

1.3.2 Similarly, describe the decision made if a type II error has been made.

1.3.3 Similarly, explain the power of the test.