Statistical Modeling for the Biological Sciences

Eric M Reyes

Updated: 02 March 2024

Table of contents

Preface

The biological sciences often yield data which present unique challenges to analysis. This text introduces these challenges and the statistical methods employed to overcome them. We begin with an introduction to the use of statistical regression models and then explore how such models can be altered to account for various features in the data. This could include non-linear or categorical response variables, censored survival (or reliability) data, or repeated measurements on the same subject. We touch on additional topics, such as study design and power, drawing causal conclusions from observational data, missing data, and general modeling techniques throughout.

This text is applied, focusing primarily on knowing when various modeling strategies are appropriate and how to interpret their results. This course surveys many different analysis strategies under a statistical modeling framework; we leave a thorough treatment of each topic to other authors. Our primary aim is to enable readers to evaluate the strength of evidence presented in the literature within their own field of study.

As with any text in statistics, we seek to develop your statistical literacy and statistical reasoning.

Part I

Unit I: Review of Statistics and Probability

We assume that you are familiar with performing statistical inference at the introductory level; this includes graphical and numerical summaries, inference (confidence intervals and hypothesis testing) for a mean response, simple linear regression to characterize the relationship between two quantitative variables, and analysis of variance for comparing the mean response across groups. We also assume this introduction to statistical inference includes major concepts like the importance of study design in interpreting results, modeling the sampling distribution of a statistic (or standardized test statistic) using a classical approach (probability) or a modern empirical approach (bootstrapping). When viewed as a list of topics like this, the introductory course can feel overwhelming. In this first unit, we provide a brief review of these topics through the lens of a unifying framework for inference. Our goal is to provide a "story" that will be further developed in the remainder of the text.

We also provide a brief introduction to the essential elements of probability. Probability is the field within mathematics that studies and models random processes. In contrast, Statistics is a discipline separate from mathematics that uses data to make inference on a population. Like many other disciplines (e.g., Engineering and the Sciences), while Statistics is a separate discipline, the theory underlying the discipline relies heavily on mathematics; for Statistics, probability plays a pivotal role. We personally favor introducing statistical concepts with minimal reference to probability, instead choosing to build on students' intuition of probability. In line with that philosophy, we will strive to introduce statistical approaches within the biological sciences with minimal probability. However, we do need a little more machinery to address these topics than is necessary for an introductory course. As a result, this unit includes elements of probability essential to our future development of statistical methods. We choose to place it here so that it is easily referenced from multiple units in the future and to emphasize that probability is separate from statistics.

1 Overview of the Statistical Process

Research is about telling a story, and good data presentation and statistical inference can help tell that story in a compelling way. This chapter cannot replace an introductory course on statistical analysis. We strive to give practical advice for data storage, presentation, and analysis while presenting a framework for inference; it is within this context that we review terminology fundamental to our study of statistical models in the biological sciences.

1.1 Overview of Drawing Inference

Every research question posed is trying to characterize a **population**.

Definition 1.1 (Population)**.** The collection of subjects we would like to say something about.

It is often impossible (or impractical) to observe the entire population. Instead, we make observations on a subset of the population; this smaller group is known as the **sample**.

Definition 1.2 (Sample)**.** The collection of subjects for which we actually obtain measurements (data).

> **Note**
>
> We acknowledge the weight of the term "subject" when discussing human participants. Medical research has not been immune to unjust practices exploiting marginalized groups within society. While the term persists in the description of research practices in general, we opt for the term "participants" when describing those individuals who actually participate in a study.
> While the semantics may seem a small component, this small shift adds a human element to the analysis. It is important to remember that each observation within the data has a story, and when those stories represent the lives of others, they deserve our full respect.

For each subject within the sample, we obtain a collection of measurements, which form our data. This could be the result, for example, of a survey, examination of medical records, or a prospective study which follows subjects for a lengthy period of time. The goal of statistical modeling is to use the sample (the group we actually observe) to say something about the

population of interest (the group we wish we had observed); this process is known as **statistical inference** and is illustrated in Figure 1.1.

Definition 1.3 (Statistical Inference)**.** The process of using a sample to characterize some aspect of the underlying population.

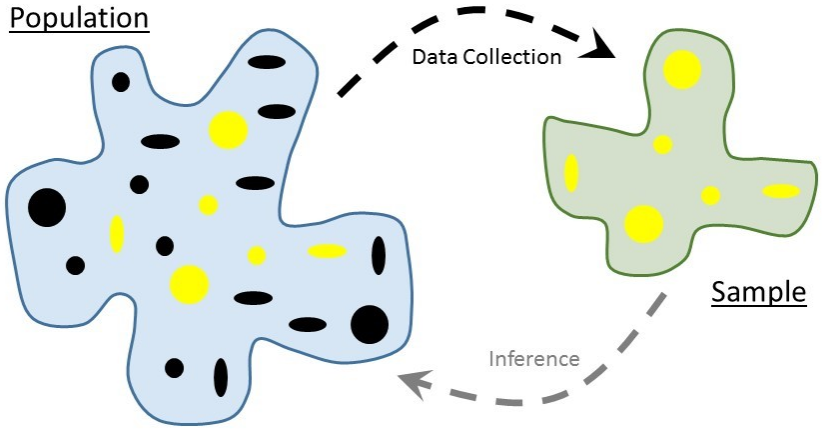


Figure 1.1: Illustration of the statistical process.

1.2 Data Storage

Each measurement, or piece of information, you record for a subject is a different **variable**.

Definition 1.4 (Variable)**.** A measurement, or category, describing some aspect of the subject.

In order to conduct analysis, it is best to adhere to "tidy data principles" (Wickham 2014) when storing data. In brief:

1. Each column contains a unique variable.
2. Each record (or row in the data set) corresponds to a different observation of the variable(s). If each subject is only measured once (a single survey for each subject, for example), each record will correspond to a different subject. If, on the other hand, each subject is measured multiple times (the same survey is given prior to an appointment and at a specified follow-up period, for example), there may be multiple records which correspond to the same subject, but each record corresponds to a unique observation.

Table 1.1: Example of storying data according to "tidy data" principles. Data is from a hypothetical study.

Subject ID	Education	Age (yrs)	Parity	Number of Miscarriages	Treatment Group
1089	0-5yrs	28	6	0	Active Treatment
1160	0-5yrs	36	1	0	Active Treatment
1025	0-5yrs	34	6	0	Active Treatment
1035	0-5yrs	32	4	1	Active Treatment
1112	6-11yrs	32	3	0	Active Treatment
1030	6-11yrs	33	4	1	Active Treatment
1159	0-5yrs	26	6	2	Placebo
1207	0-5yrs	42	1	0	Placebo
1179	0-5yrs	39	6	0	Placebo
1014	0-5yrs	34	4	0	Placebo
1195	6-11yrs	35	3	1	Placebo
1170	6-11yrs	36	4	1	Placebo

3. If you have multiple data sets, there should be a variable in the table that allows the various tables to be linked (subject identifier). For larger more complex studies, for example, you may have one table that has the demographic information of subjects and a separate table which contains the lab results for the subjects.
4. The first row in the data set should have the names of each variable.

The above description eliminates a common method of data storage — placing different groups in different spreadsheets. All observations should be stored together. The first few records of a hypothetical data set are illustrated in Table 1.1.

Once your data has been placed in a spreadsheet, it should be kept separate from the analysis. Any changes to the data should be done using your analysis file so that those changes are clearly documented alongside the analysis. While it may be easy, it is poor practice to include graphics and numeric summaries in the same spreadsheet as the data. If you want your data to be *portable* (easily opened by any spreadsheet or analysis software package), save your data as a comma separated file (CSV).

1.3 Tabular Data Presentation

If you have several variables you want to summarize, this is probably best done using a table. For example, you may want to summarize the demographics of the subjects in your study across each treatment group. How a variable is summarized depends on its type. **Qualitative** (or **categorical**) variables define a grouping or categorization of a subject (e.g., race, treatment

group, etc.). When summarizing qualitative data, we generally report the number of subjects in each group and the corresponding percentage of the sample.

Definition 1.5 (Categorical Variable)**.** Also called a "qualitative variable," a measurement on a subject which denotes a grouping or categorization.

Quantitative (or **numeric**) variables are those measurements for which arithmetic makes sense (e.g., heart rate, age). These variables are generally summarized by reporting both a measure of location and spread; this could be mean and standard deviation or median and interquartile range.

Definition 1.6 (Numeric Variable)**.** Also called a "quantitative variable," a measurement on a subject which takes on a numeric value *and* for which ordinary arithmetic makes sense.

If you are not comparing groups of subjects, it is reasonable to report results for the entire sample. If the goal of your research is to compare groups (such as rural vs. urban residents), we typically summarize data within each group and present the comparisons side by side. Table 1.2 summarizes the data from our hypothetical study, allowing the reader to compare the treatment and placebo groups. Notice that while we might think of the number of miscarriages as being a numeric variable, when there are only a small number of possibilities, we might treat that variable as categorical for the purposes of summarizing it.

Table 1.2: Summary of patient characteristics from our hypothetical study.

Characteristic	**Active Treatment**, N = 165	**Placebo**, N = 83
Education		
0-5yrs	8 (4.8%)	4 (4.8%)
6-11yrs	80 (48%)	40 (48%)
12+ yrs	77 (47%)	39 (47%)
Age	30.51 (2.67)	31.53 (5.28)
Parity	2.08 (1.24)	2.11 (1.28)
Number of Miscarriages		
0	113 (68%)	28 (34%)
1	40 (24%)	31 (37%)
2	12 (7.3%)	24 (29%)

When reporting numerical summaries within the body of your report, it is good to keep the same format as you adopt in the table; for example, summarizing a qualitative variables with N (%).

Statistics is generally concerned with explaining the variation in a variable, and that is characterized by its **distribution**. When we summarize a variable, whether numerically or graphically, we are actually summarizing this distribution.

Definition 1.7 (Distribution)**.** The pattern of variability corresponding to a set of values.

1.4 Graphical Data Presentation

As the saying goes, a picture is worth 1000 words. Each graphic you construct, however, should add value to the story you are telling. We primarily reserve graphics for conveying a message about our primary **response**.

Definition 1.8 (Response Variable)**.** Also called the "outcome," this is the primary variable of interest in the research question; it is the variable we either want to explain or predict.

As with tabular data presentation, our approach to graphical presentation depends on the type of variable being summarized. For example, while a scatter-plot is well suited for examining the relationship between two quantitative variables, side-by-side box-plots are better suited for examining the relationship between a quantitative response and a categorical predictor.

> **Note**
>
> Following best practices in the research community, we recommend the use of a *bar chart* instead of a *pie chart* when examining a categorical response. Bar charts are often less cluttered and more clearly communicate the same information.

Figure 1.2 illustrates two graphics (one for a qualitative and one for a quantitative response); again, in practice, your graphics should be driven by your research question.

Notice that the left panel of the graphic makes use of bar charts to compare a qualitative variable (number of miscarriages) across a second qualitative variable (treatment group). The use of color here is important because it brings out additional features that do not appear on the x- or y-axis. The right panel of the graphic makes use of box-plots (with jitter-plots overlaid) to compare a quantitative variable (age of the patient) across the qualitative variable (treatment group). One idea worth discussing here is that a graphical summary of a quantitative variable should always portray both *location* and *spread*. Notice that in the right panel in Figure 1.2, we see that the ages of patients receiving placebo are comparable (in location) to that of those receiving the active treatment; however, the variability in the ages of patients receiving placebo is much larger compared to those receiving the active treatment. Compare this to Figure 1.3, which only summarizes location with no sense of spread; while this is a popular default graphic in some software, it does not adequately allow a reader to determine the size of the effect relative to the variability in the data.

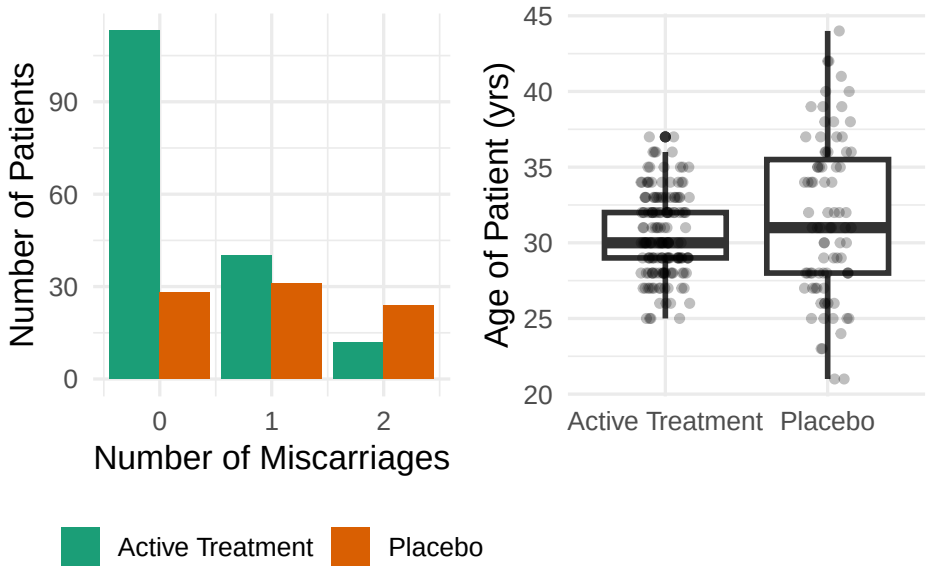


Figure 1.2: Two graphical presentations of data from a hypothetical study, adhering to good graphical practices.

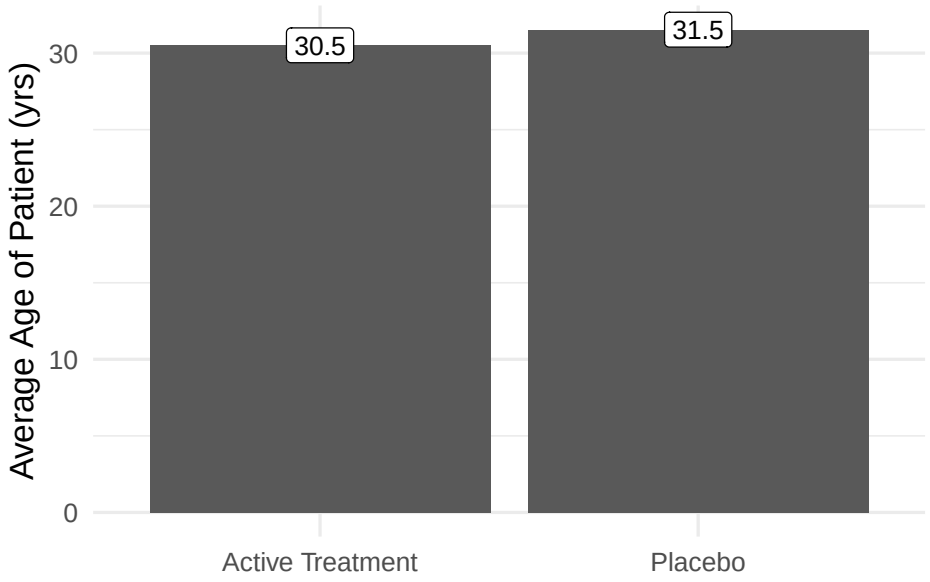


Figure 1.3: Inappropriate graphical presentation from a hypothetical study as it ignores a sense of variability in the data.

1.5 Basic Terminology for Statistical Tests

In some cases, summarizing the data numerically and graphically is sufficient for telling a compelling story. Often, however, the summaries are accompanied by a statistical analysis. Regardless of the simplicity (or complexity) of the statistical procedure, there are a few fundamental ideas which are common to all methods.

A **statistic** (summary of data) is a point estimate of a **parameter** (corresponding value in the population of interest). For example, the value 2.08 in Table 1.2 is the average number of children among those women in the study who received the active treatment; but, it *estimates* the average number of children among all women in the population who receive the active treatment.

Definition 1.9 (Parameter)**.** Numeric quantity which summarizes the distribution of a variable within the *population* of interest. Generally denoted by Greek letters in statistical formulas.

Definition 1.10 (Statistic)**.** Numeric quantity which summarizes the distribution of a variable within the observed *sample*.

Instead of estimating a parameter with this single value, we can estimate the parameter with a **confidence interval** (a 95% confidence interval is standard practice).

Definition 1.11 (Confidence Interval)**.** An interval (range of values) estimate of a parameter that incorporates the variability in the statistic. The process of constructing k% confidence intervals results in them containing the parameter of interest in k% of *repeated* studies. The value of k is called the *confidence level*.

It is important to recognize that the entire interval is our estimate. In text, we generally report the point estimate with the 95% confidence interval in parentheses. For example,

> The probability of a miscarriage is 0.32 (95% CI: [0.25, 0.39]) for women given the active treatment.

There are several common misinterpretations of a confidence interval; generally, these do not enter the literature because we avoid interpreting the interval directly and simply state it and discuss its implications (as above). For completeness, however, it is best to think of a confidence interval as giving all the reasonable values of the parameter based on the data observed.

While confidence intervals estimate an effect, a **p-value** quantifies the amount of evidence in the data against the lack of an effect.

Definition 1.12 (P-Value)**.** The probability, assuming the null hypothesis is true, that we would observe a statistic, from sampling variability alone, as extreme or more so as that observed in our sample. This quantifies the strength of evidence against the null hypothesis. Smaller values indicate stronger evidence.

We generally report a p-value to 3 decimal places (with values less than 0.001 being written as "< 0.001"). It is best to state p-values alongside the conclusion. For example,

> There is strong evidence ($p < 0.001$) that the active treatment reduces the risk of a miscarriage.

Caution

There are two very important things to keep in mind when examining a p-value:

1. A small p-value does not imply the effect is clinically relevant/important. It simply indicates that we are able to statistically discern an effect/difference is present.
2. A large p-value does not imply there is no effect/difference. It simply indicates that we cannot statistically discern the presence of an effect/difference.

For these reasons, a p-value should always be accompanied by either a confidence interval (preferred when possible) or a point estimate of the effect to allow readers to determine if the impact is clinically relevant.

When interpreting statistical results, the design of the study plays a role. In particular, we can only conclude a causal relationship when the data is from a **randomized clinical trial**. When your data is from an **observational study**, any group comparisons are subject to **confounding**.

Definition 1.13 (Randomized Clinical Trial)**.** Also called a "controlled experiment," a study in which each participant is randomly assigned to one of the groups being compared in the study.

Definition 1.14 (Observational Study)**.** A study in which each participant "self-selects" into one of groups being compared in the study. The phrase "self-selects" is used very loosely here and can include studies in which the groups are defined by an inherent characteristic, the groups are determined according to a non-random mechanism, and each participant chooses the group to which they belong.

Definition 1.15 (Confounding)**.** When the effect of a variable on the response is misrepresented due to the presence of a third, potentially unobserved, variable known as a confounder.

Example 1.1 (Dental Health and Cardiovascular Disease)**.** It has been suggested that brushing your teeth twice a day for at least two minutes may lower the risk of cardiovascular diseases[1].

However, most of these results are from large surveys, which are observational studies. It is quite plausible that those who take excellent care of their teeth tend to be health-conscious individuals, and health-conscious individuals are more likely to have healthy diets and exercise regularly, both of which decrease the risk of cardiovascular diseases.

In Example 1.1, being health-conscious is a confounder because it is associated with *both* the factor under study (brushing behavior) *and* the outcome of interest (cardiovascular disease); see Figure 1.4. In order to establish a causal link between brushing and the risk of cardiovascular disease, we could conduct a clinical trial in which we randomize patients to a brushing routine and then track their long-term cardiovascular health; in this design, the link between the confounder and the treatment group is broken, allowing us to make a causal conclusion. While clinical trials allow for causal conclusions, they are not always feasible or practical; observational studies allow us to add to the body of knowledge in such situations. There are some methods for addressing confounding in observational studies through statistical analysis, but such methods often require a large sample and more advanced methodology.

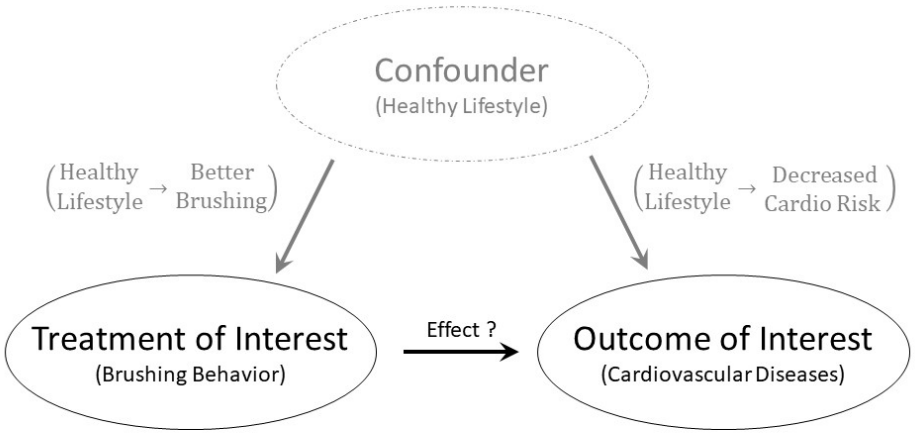


Figure 1.4: Illustration of confounding in observational studies.

[1] https://www.heart.org/en/news/2018/11/07/bad-tooth-brushing-habits-tied-to-higher-heart-risk

1.6 A Note on Codebooks

A dataset on its own is meaningless if you cannot understand what the values represent. *Before* you access a dataset, you should always review any available **codebook**.

Definition 1.16 (Codebook)**.** Also called a "data dictionary," a codebook provides complete information regarding the variables contained within a dataset.

Some codebooks are excellent, with detailed descriptions of how the variables were collected and appropriate units. Other codebooks give only an indication of what each variable represents. Whenever you are working with previously collected data, reviewing a codebook is the first step; and, you should be prepared to revisit the codebook often throughout an analysis. When you are collecting your own dataset, constructing a codebook is essential for others to make use of your data.

2 Distributional Quartet

Any good statistical analysis moves between four key distributions — what we refer to as the *Distributional Quartet*. While not always explicitly discussed, these distributions are always present in an analysis. Understanding their role is important to implementing and interpreting an analysis.

We begin by considering the following example from Rosner (2006).

Example 2.1 (Blood Pressure when Lying Down)**.** Blood pressure is one metric for the health of your heart. A blood pressure reading includes two numbers — the systolic blood pressure (the "top number," measures the amount of pressure in your arteries when your heart contracts) and the diastolic blood pressure (the "bottom number," measures the amount of pressure in your arteries when your heart is between beats).

An individual does not have a single blood pressure reading; our blood pressure fluctuates as a result of activity as well as our position. In a study examining the impact of position on blood pressure, 32 participants had their blood pressure measured while lying down with their arms at their sides.

Stated simply, the discipline of statistics is about using data to say something about a process that characterizes a population. Our analysis, therefore, begins with the **Distribution of the Population**.

> **!** Distribution of the Population
>
> The pattern of variability in values of a variable across individuals of the population. The shape of this distribution is governed by unknown parameters. While we generally do not know the shape of this distribution, we may occasionally posit a model for it.

We are interested in using the data from this study to characterize the systolic blood pressure of individuals when in this recumbent position, with their arm at their side. Of course, we are unable to assess the blood pressure of all individuals in the world. Therefore, we do not know what the distribution of systolic blood pressure measurements is for this population. However, we *might* posit a model for this distribution. Such models, which must account for the variability among the population, are studied in probability theory, which we consider in the next section. For now, it suffices to imagine the distribution of the population graphically; it is characterized by the unknown parameters (Figure 2.1).

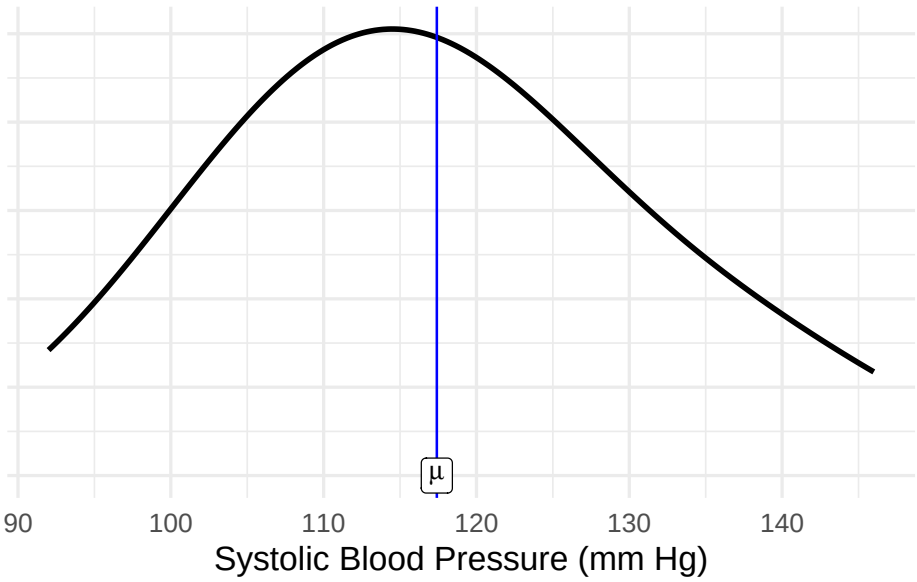


Figure 2.1: Hypothetical model for the distribution of systolic blood pressure within the population. The unknown population mean is denoted on the graphic.

The data we actually observe comes from the sample, and we will use this smaller group to say something about the underlying population. It is the distribution of sample which we are summarizing each time we construct a graphic.

! Distribution of the Sample

The pattern of variability in values of a variable across individuals of the sample. This is typically summarized graphically and numerically.

Figure 2.2 summarizes the sample using a histogram. If our sample is collected well, then it should be representative of the population, meaning that the distribution of the sample should reflect the characteristics of the (unobserved) distribution of the population. The location, spread, and shape should all reflect what we might see within the population.

Examining the sample is critical to understanding the story in the data. We must remember, however, that the statistics we compute using our sample are dependent upon the data we observed. If we were to repeat the sampling process (collect new data to answer the same question), our statistics would change. Good inference requires us to acknowledge this sampling variability and incorporate it when making statements about the population.

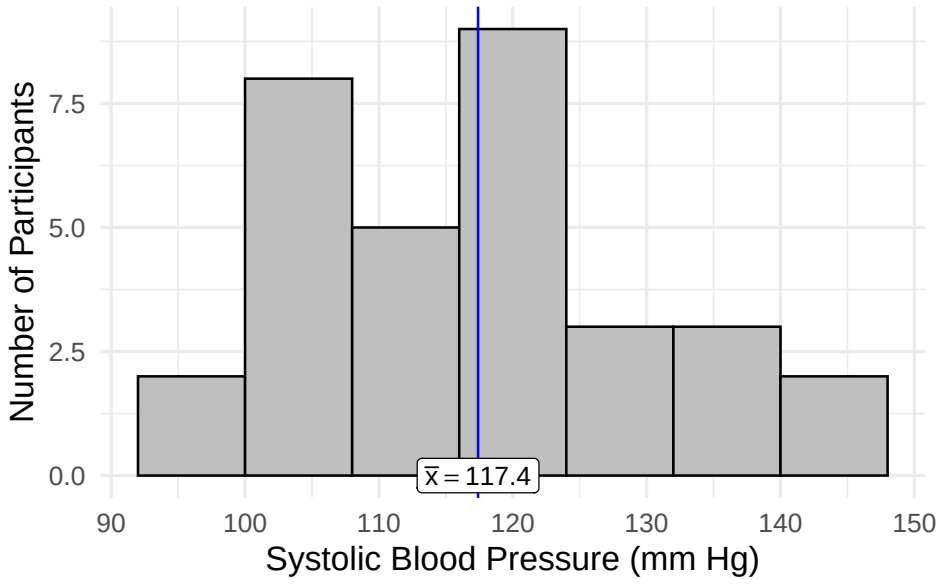


Figure 2.2: Distribution of systolic blood pressure within the observed sample of 32 participants.

> **!** Sampling Distribution
>
> The pattern of variability in values of a *statistic* (or standardized statistic) across repeated samples of the same size from the population. This must be modeled in practice.

As we are generally unable to perform replicate studies, we model the sampling distribution using the data available (future chapters will discuss the various methods available for modeling this distribution). The model for the sampling distribution allows us to determine values of the parameter for which the data is consistent. That is, it allows us to compute a confidence interval to estimate a parameter of interest. For example, a 95% CI for the mean systolic blood pressure (mm Hg, when recumbent with arm at their side), based on our data available, is (113.1, 121.9); this is illustrated in Figure 2.3 alongside the model for the sampling distribution of the sample mean systolic blood pressure from which the CI was computed.

While statisticians generally have a preference toward estimation (and therefore reporting confidence intervals), scientists often have specific research questions they would like to address. When this is the case, the scientist may want to quantify the strength of evidence in the sample against some specified hypothesis. In order to determine how rare our data is, we must know what we should expect under the specified hypothesis.

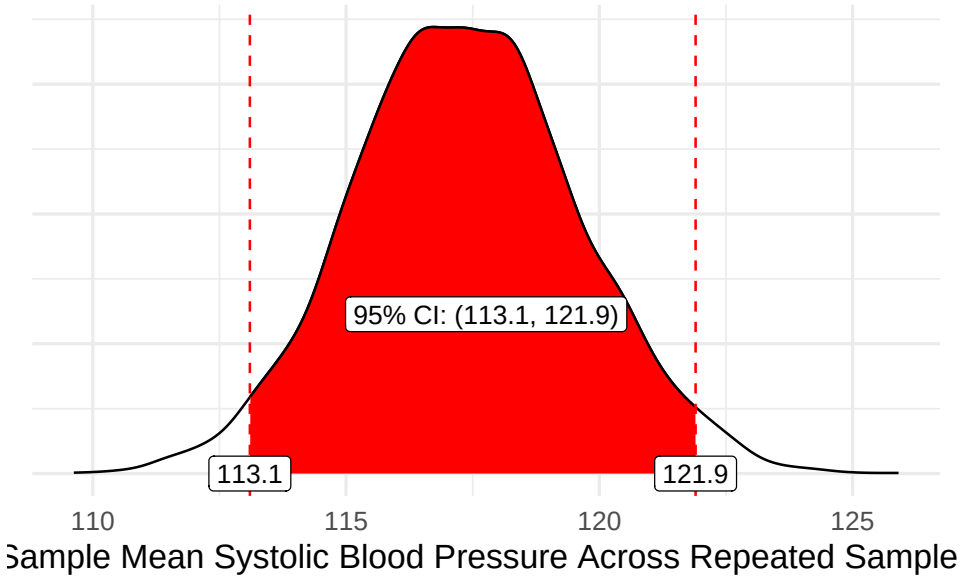


Figure 2.3: Empirical model for the sampling distribution of mean systolic blood pressure for a sample of 32 participants. A 95% confidence interval is also illustrated.

> **!** Null Distribution
>
> The sampling distribution of a statistic (or standardized "test" statistic) *when the null hypothesis is true*. This must be modeled in practice.

The null distribution effectively tells us what values of the statistic we would expect to see if the null hypothesis were true. If our observed statistic seems plausible according to the null distribution, then it follows that the null hypothesis is reasonable. If, on the other hand, our observed statistic is unexpected according to the null distribution, then we have evidence that the null hypothesis is false (and therefore that the alternative is true). That is, we are able to statistically discern the difference between our data and what we would have expected to see under the null hypothesis. Note that these are our only two potential conclusions. The strength of the evidence is quantified by the p-value. For example, suppose we are interested in using our data to test the following set of hypotheses:

$$H_0 : \mu = 120 \qquad \text{vs.} \qquad H_1 : \mu \neq 120,$$

where μ represents the average systolic blood pressure of an individual when lying down.

That is, we are interested in determining if there is evidence that, on average, the systolic blood pressure (recumbent with arm at side) differs from 120 mm Hg. According to the data, it is reasonable (p = 0.247) that the average systolic blood pressure is 120 mm Hg.

The computation of the p-value is illustrated in Figure 2.4 alongside the model for the null distribution for this particular hypothesis.

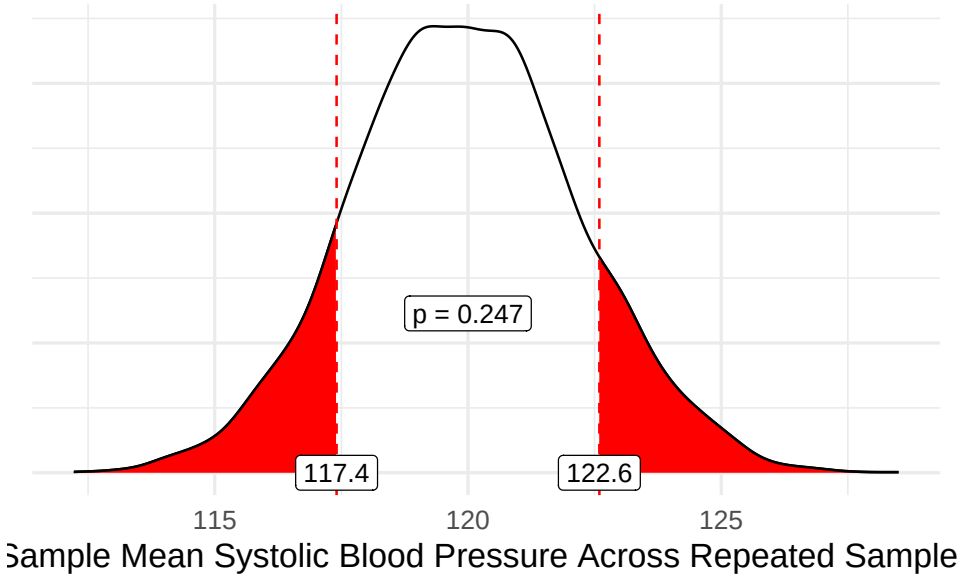


Figure 2.4: Empirical model for the null distribution of mean systolic blood pressure within the sample of 32 participants when the null hypothesis is that the mean systolic blood pressure is 120 mm Hg. The p-value is also illustrated.

Our focus in this short review is not this specific analysis. The emphasis here is on the *process* — the use of the four distributions used in a statistical analysis. They allow us to take the sample and make inference on the underlying population. As we move forward in the course, we will study more sophisticated models. However, behind the scenes, the procedures are always bouncing between these four distributions in order to allow us to make inference.

Note

Note that we construct a *model* for the sampling distribution of a statistic or a *model* for the null distribution of a (standardized) statistic. All models are simplifications of complex processes; and, the validity of each model requires certain conditions be true about the data generating process.
Generally, we must make assumptions about the data generating process. Therefore, the reliability of these models, and consequently our analysis, depends on whether these assumptions are reasonable. We must always keep in mind that our analysis is subject to the assumptions we make.

3 Essential Probability

The discipline of Statistics uses data to make inference on a population. In turn, statistical theory is built on probability — the discipline of mathematics that studies and models random processes. While we do not need to be experts in probability to be practitioners of statistical methodology, a foundation in models from probability is helpful for seeing common threads in statistical modeling. This chapter provides a brief introduction to the most relevant aspects of probability theory necessary for engaging with the remainder of the text.

3.1 Density Functions as Models

Any process for which the outcome cannot be predicted with certainty is a random process. Typically, probability is taught from a mathematical perspective, with a goal of constructing a coherent and complete framework for characterizing such random processes. Here, our goal is to introduce key probability concepts by relating them to their data-centric analogues. That is, we want to think of probability in light of how we will use it in statistical analysis.

Each time we collect data, we can think of each observation as the result of a random process. These observations are recorded as variables in our dataset. In probability, a random variable is used to represent a measurement that results from a random process. Just as we have both *quantitative* and *qualitative* variables, there are *continuous* and *discrete* random variables.

Definition 3.1 (Random Variable)**.** A random variable represents a measurement that will be collected and for which the value cannot be predicted with certainty; they are generally represented with a capital letter. Continuous random variables represent quantitative measurements while discrete random variables represent qualitative measurements.

Consider measuring a single variable on a sample of n participants. Then, we might represent the measurements we will obtain as $X_1, X_2, \ldots, X_n$.

> **Note**
>
> There are many ways to interpret probability. In classical ("frequentist") statistics, we think of probability as the likelihood of an event over repeated experimentation. Therefore, probability does not describe events that have already occurred; we can only describe the likelihood of future events.

Each of our random variables $X_1, X_2, \ldots, X_n$ will be observations from some underlying population. As we described in previous chapters, the distribution of the population is unknown. However, we might posit a model for this distribution. This is our primary use of probability theory in statistics — to model distributions. The most common way to represent a probability model is through its density function.

Definition 3.2 (Density Function)**.** A density function f relates the potential values of a random variable X with the probability those values occur. For a *continuous* random variable, the probability the random variable X falls within an interval (a, b) is given by

$$Pr(a \leq X \leq b) = \int_a^b f(x)dx.$$

For a *discrete* random variable, the probability the random variable X is equal to the value u is given by

$$Pr(X = u) = f(u).$$

> **Note**
>
> In a probability course, there is often a distinction made between probability density functions (continuous random variables) and probability mass functions (discrete random variables). We do not make this distinction and instead rely on the context to determine whether we are dealing with a continuous or discrete random variable.

With few exceptions, we will be working with continuous random variables. As a result, the density function is a smooth function over some region, and the actual value of the function is not interpretable; instead, we obtain probabilities by computing the area under the curve. Again, drawing connections to data analysis, we can think of a density function as a mathematical formula representing a smooth histogram. The area under the curve for any region gives the proportion of the population which has a value in that region. That is, we get the probability that a random variable will be in an interval by integrating the density function over that interval. Figure 3.1 illustrates this idea; we have a hypothetical dataset that has been summarized using a histogram; we overlay a density function (with the corresponding mathematical model that describes this density function). The figure shows how the sample (summarized in the histogram) is approximating the population (the density function).

Especially for visualization, the density function is the most common way of characterizing a probability model. However, computing the probability using the density is problematic due to the integration required. Many software programs address this by working with the cumulative distribution function (CDF).

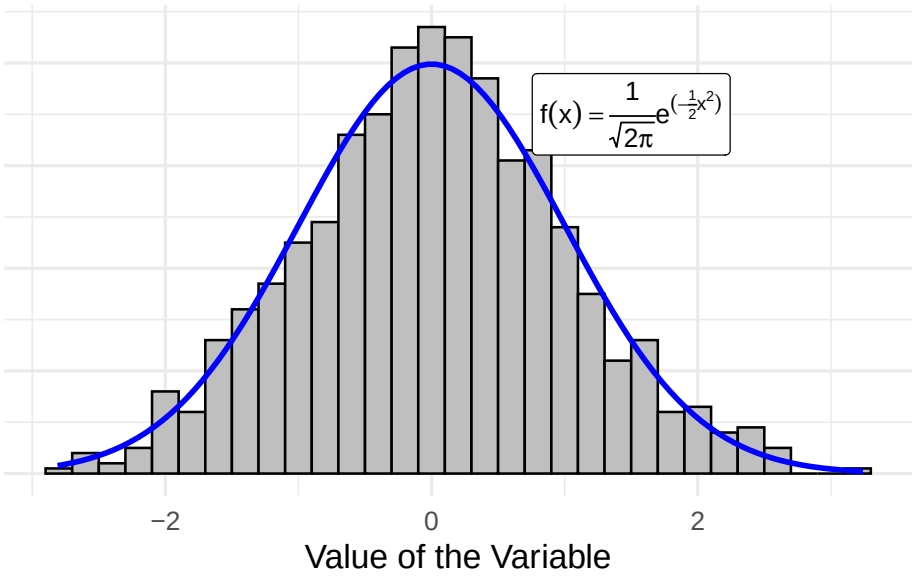


Figure 3.1: Illustration of a density function representing the distribution of the population and a histogram from a representative sample.

Definition 3.3 (Cumulative Distribution Function (CDF))**.** Let X be a random variable; the cumulative distribution function (CDF) is defined as

$$F(u) = Pr(X \leq u).$$

For a continuous random variable, we have that

$$F(u) = \int_{-\infty}^{u} f(x)dx$$

implying that the density function is the derivative of the CDF. For a discrete random variable

$$F(u) = \sum_{x \leq u} f(x).$$

Working with the CDF improves computation because it avoids the need to integrate each time; instead, the integral is computed once (and stored internally in the computer) and we use the result to compute probabilities directly.

Big Idea

Density functions are the mathematical models for distributions; they link values of the variable with the likelihood of occurrence. However, for computational reasons, we often work with the cumulative distribution function which provides the probability a random variable is less than or equal to a value.

3.2 Summarizing Distributions (Parameters)

Most scientific questions are focused on the location or spread of a distribution. For example, we are interested in estimating the average yield of a crop, or the variance in the amount of sleep among college students. Introductory statistics introduces summaries of location and spread within the sample (e.g., sample mean for location and sample variance for spread). Analogous summaries exist for density functions.

In particular, the mean of a random variable (denoted by $E(X)$) and the variance of a random variable (denoted by $Var(X)$) are measures of the location and spread, respectively, of the distribution represented by its corresponding density function. When the density function is a model for the population, these represent the parameters of the population — the same parameters we estimate and make inference on using our data analysis. For completeness, we present the computational formulas for the mean and variance of a random variable, but we do not make use of these formulas moving forward. Instead, we simply note that these formulas are similar to their sample counterparts.

Definition 3.4 (Mean and Variance of a Random Variable)**.** Suppose X is a random variable with density function f. If X is a continuous random variable, then the mean and variance are given by

$$E(X) = \int xf(x)dx$$
$$Var(X) = \int (x - E(X))^2 f(x)dx.$$

If X is a discrete random variable, then the mean and variance are given by

$$E(X) = \sum xf(x)$$
$$Var(X) = \sum (x - E(X))^2 f(x).$$

As we have stated, the distribution of the population is generally unknown. If we were able to fully specify the density function for the population, then there would be no need for statistical analysis. Instead, the model is generally posited up to some unknown values (parameters). For example, a researcher might posit that within the population, the time until a medical device fails could be modeled using the density

$$f(x) = \frac{1}{\mu} e^{-\frac{x}{\mu}} \qquad x > 0.$$

Here, the researcher has really posited a form of the model, but not the exact model as μ is unknown. The value μ represents the average response (which could be confirmed using the formulas in the above definition). In such cases, making inference on the parameters allows us to characterize the distribution of the population.

> **Big Idea**
>
> When a probability model is specified for a population, it is generally specified up to some unknown parameter(s). Making inference on the unknown parameter(s) therefore characterizes the distribution — characterizes the manner in which the response varies across individuals in the population.

3.3 Specific Models for Populations

While we could posit any non-negative function as a model for a density function, there are some models that are very common. The most common model for the population of a continuous random variable is the Normal distribution.

Definition 3.5 (Normal (Gaussian) Distribution)**.** Let X be a continuous random variable. X is said to have a Normal (or Gaussian) distribution if the density is given by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \qquad -\infty < x < \infty,$$

where μ is any real number and $\sigma^2 > 0$.

- $E(X) = \mu$
- $Var(X) = \sigma^2$

We write $X \sim N(\mu, \sigma^2)$, which is read "X has a Normal distribution with mean μ and variance σ^2." This short-hand implies the density above.

This model is a bell-shaped distribution centered at the mean μ. While this is a common model, it should not be assumed by default. In future chapters, we will consider methods for assessing whether assuming a Normal distribution is reasonable.

When a response is binary (assumes one of two values), it is a Bernoulli distribution. In order to make use of this distribution, we typically define one of the two possible outcomes as a "success" and the other as a "failure." For example,

$$X = \begin{cases} 1 & \text{if a success is observed} \\ 0 & \text{if a success is not observed.} \end{cases}$$

Definition 3.6 (Bernoulli Distribution)**.** Let X be a discrete random variable taking the value 0 or 1. X is said to have a Bernoulli distribution with density

$$f(x) = \theta^x(1-\theta)^{1-x} \qquad x \in \{0, 1\},$$

where $0 < \theta < 1$ is the probability that X takes the value 1.

- $E(X) = \theta$
- $Var(X) = \theta(1-\theta)$

We write $X \sim Ber(\theta)$, which is read "X has a Bernoulli distribution with probability θ."

> **Note**
>
> A generalization of the Bernoulli distribution is the Binomial distribution. So, we sometimes hear people refer to a Bernoulli distribution as "a Binomial distribution with a single event."

3.4 Models for Sampling Distributions and Null Distributions

A statistical analysis does not exist in a vacuum. Instead, based on the context of the study, we make assumptions about the process which generated the data. The conditions we are willing to assume govern how we model the sampling distribution or null distribution. Occasionally, we can lean on statistical theory to say how the sampling distribution or null distribution will behave. That is, under certain conditions, statistical theory tells us what the appropriate model is. In these situations, there are some common models.

The t-distribution is a bell-shaped distribution, similar to the Normal distribution but with wider tails. It has a single parameter, known as the degrees of freedom. Note that unlike many other distributions, this parameter (the degrees of freedom) is not associated with the location of the distribution. Instead, the parameter governs the spread (but is not the variance).

Definition 3.7 (t-Distribution)**.** Let X be a continuous random variable. X is said to have a t-distribution if the density is given by

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \qquad x>0$$

where $\nu > 0$ is the degrees of freedom.

We write $X \sim t_\nu$, which is read "X has a t-distribution with ν degrees of freedom."

The Chi-Square distribution is a skewed distribution (looks like a giant slide). It has a single parameter, known as the degrees of freedom. The degrees of freedom for this distribution characterize both the location and spread simultaneously.

Definition 3.8 (Chi-Square Distribution)**.** Let X be a continuous random variable. X is said to have a Chi-Square distribution if the density is given by

$$f(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)}\, x^{\nu/2-1}e^{-x/2} \qquad x>0,$$

where $\nu > 0$ is the degrees of freedom.

We write $X \sim \chi^2_\nu$, which is read "X has a Chi-Square distribution with ν degrees of freedom."

The F-distribution is a skewed distribution. It has two parameters, known as the numerator and denominator degrees of freedom. While neither variable is the mean or variance, together these two parameters characterize both the location and the spread.

Definition 3.9 (F-Distribution)**.** Let X be a continuous random variable. X is said to have an F-distribution if the density is given by

$$f(x) = \frac{\Gamma((r+s)/2)}{(\Gamma(r/2)\Gamma(s/2))}(r/s)^{(r/2)}x^{(r/2-1)}(1+(r/s)x)^{-(r+s)/2} \qquad x>0,$$

where $r, s > 0$ are the numerator and denominator degrees of freedom, respectively.

We write $X \sim F_{r,s}$, which is read "X has an F-distribution with r numerator degrees of freedom and s denominator degrees of freedom."

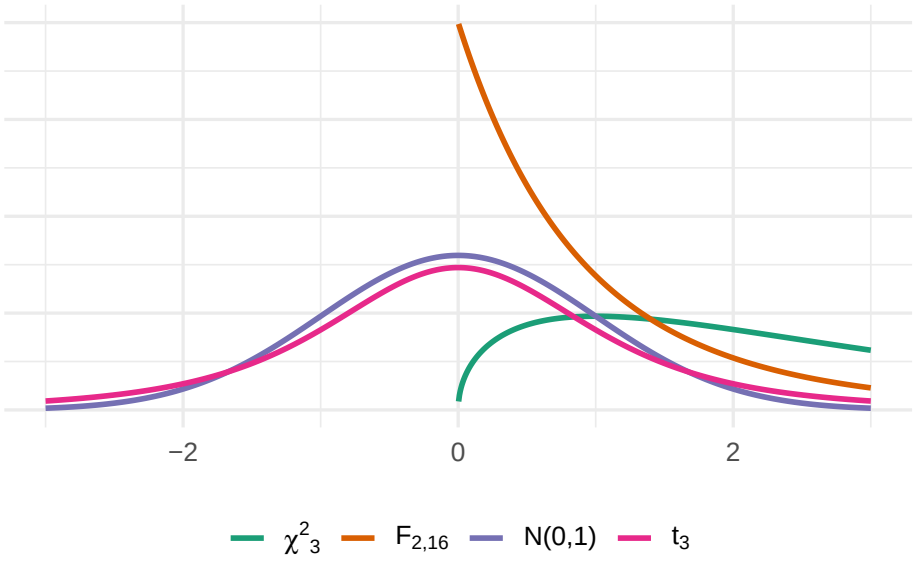


Figure 3.2: Comparison of various common distributions.

The formulas above are ugly, but we will not be working with them directly. Instead, statistical software has these distributions embedded. The key idea here is that when we know the model for a sampling distribution, we are able to rely on that model in order to obtain confidence intervals. And, when we have a model for the null distribution, we are able to rely on that model to obtain p-values. These models are behind default implementations of statistical methods in software.

Big Idea

Some probability models occur so frequently that we give them names for easy reference. Some models are common for modeling the population, in which case they are defined in terms of unknown parameters to be estimated. Some models are common for modeling sampling distributions or null distributions, in which case their form will be explicitly determined according to statistical theory.

Part II

Unit II: General Linear Model

The general linear model, also known as multiple linear regression, provides a framework appropriate for modeling a continuous outcome (response) as a function of several predictors. We introduce the framework and illustrate how it unifies the methods typically discussed in an introductory statistics course.

4 General Linear Model Framework

The most interesting scientific questions involve characterizing the relationship between a response and some predictor. And, we know that these relationships do not exist in a vacuum. The response we observe is typically the result of a complex data generating process involving several potential predictors (or features/characteristics) of the subjects in the population. In order to incorporate these additional features, we need *multivariable* models.

The development of a model should not be divorced from its intended use, and in general, there are three uses for multivariable models. That is, the majority of scientific questions can be categorized into one of three groups: prediction, isolating an effect, or studying the interplay between variables.

Big Idea

There are primarily three uses for a multivariable model

- **Prediction**: modeling a relationship for the purpose of estimating a future occurrence given new data.
- **Isolating an Effect**: describing the relationship between a response and predictor after accounting for the influence of other predictors measured.
- **Studying the Interplay**: examining how the relationship of two variables is impacted by the value of a third variable.

While we introduce these elements in the context of the general linear model, note that these uses carry over into other regression models we will examine.

Consider a gardener studying two common organic fertilizers. She could have the following questions in mind:

A. What do I anticipate the yield of tomatoes to be next summer when using cow manure?
B. Does bat guano tend to result in higher tomato yields compared with cow manure after accounting for any impact on yield that results from the amount of water the plants receive?
C. Does the efficacy of bat guano (compared with cow manure) depend on the amount of sunlight the plants receive?

The first question is an example of prediction; given the fertilizer applied (as well as potentially other characteristics of the garden), what does she expect the results to be in the future? The

second question examines the impact (or effect) of the fertilizer *above and beyond* any impact of watering; she is interested in *isolating* the effect of fertilizer from the effect of watering. In the last question, she is not only interested in the effect of the fertilizer on the yield, but she wants to acknowledge that this impact could depend on a third variable (sunlight); for example, bat guano may be superior under low light settings but inferior under lots of sunlight. This is an example of the interplay between the fertilizer and the sunlight.

In each of these objectives, there are multiple things at play, requiring modeling techniques that account for multiple predictors simultaneously. The general linear model views the response as a being the result of a linear combination of several variables; our measurement of this linear combination is then subject to error. Specifically, the framework generalizes the simple linear regression model studied in introductory statistics to characterize the average response as a function of several variables simultaneously.

Definition 4.1 (General Linear Model)**.** The general linear model views the response (outcome) as a linear combination of several predictors:

$$\begin{aligned}(\text{Response})_i &= \beta_0 + \beta_1(\text{Predictor 1})_i + \beta_2(\text{Predictor 2})_i + \cdots + \beta_p(\text{Predictor } p)_i + \varepsilon_i \\ &= \beta_0 + \sum_{j=1}^{p} \beta_j(\text{Predictor } j)_i + \varepsilon_i\end{aligned}$$

where n is the number of subjects in the sample, $p < n$ is the number of predictors in the model, and ε_i is a random variable that captures the error in the response.

i Note

Many texts use y_i to denote the response of the i-th observation and $x_{j,i}$ to denote the value of the j-th predictor for the i-th subject, resulting in the general linear model having the form

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{j,i} + \varepsilon_i.$$

This notation is helpful when discussing the underlying mathematics, but we prefer being more explicit in identifying the response and predictors when discussing the model itself.

i Note

Some disciplines refer to the response/outcome as the "dependent variable" and the predictors as "independent variables," but we find this language a bit dated.
We will use the terms "predictor" and "covariate" interchangeably, while some disciplines distinguish between categorical predictors as factors and continuous predictors as

covariates (variables that "co-vary" with the factor of interest).

> **Note**
>
> While not a theoretical requirement, we will only consider the case where $p < n$, which is common in many disciplines. One discipline in which this is often not valid is genetics. Special methods are required in such "high dimensional" settings that are beyond the scope of this text.

The general linear model has two distinct components — a deterministic component (the linear combination of the predictors) and a stochastic component (the error term). We can think of the error term as the "junk drawer" for the model, capturing anything not explained by the deterministic portion of the model. The error could include systematic error in measuring the response, biological error contributing to the fact that two subjects with the same values of the predictors have different responses, etc.
:::{.callout-caution} We stress that Definition 4.1 is a *model*. Like all models, it is a simple representation of a complex process. It is something we posit characterizes the underlying data generating process. :::

The key feature of this model is that it relates the response to several predictors *simultaneously*. However, this model is currently comprised of unknown parameters (the coefficients $\beta_1, \beta_2, \ldots, \beta_p$). For it to be useful in practice, we need estimates of these parameters.

4.1 Parameter Estimation

The coefficients in front of each predictor act as parameters in the model, as they are unknown and characterize the distribution of the response in some way. Our goal is to construct estimates of these unknown quantities. The most common method of estimation is the method of least squares.

Definition 4.2 (Least Squares Estimation)**.** The method of least squares may be used to estimate the coefficients (parameters) of a linear model. In particular, we choose the values of the coefficients that minimize

$$\sum_{i=1}^{n} \left((\text{Response})_i - \beta_0 - \sum_{j=1}^{p} \beta_j (\text{Predictor } j)_i \right)^2 .$$

The resulting "least squares" estimates are denoted $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$.

It is important to remember that the method of least squares results in *estimates* of the parameters. We are not "solving" for the parameters; the parameters will always remain unknown quantities. We are using data to estimate the parameters. There is really nothing statistical about least squares. It is simply an optimization problem — choosing coefficients to minimize some criteria. Of course, we do not determine these estimates by hand; instead, we rely on statistical software.

We cannot stress enough that the act of obtaining these estimates is simply an optimization exercise. While a computer can provide these estimates, we cannot yet even interpret these estimates without further assumptions on the model. This is where it becomes a *statistical* problem — specifying the conditions required for the purpose of making inference on the unknown parameters.

4.2 Conditions on the Model

The act of estimation alone is really a mathematical problem. Being able to describe the properties of those estimates, quantify the variability in those estimates, and use those estimates to make inference on the population parameters is where we enter statistics. Whenever a random variable is present in a model, inference requires us to make assumptions about its underlying distribution. As analysts, we balance making inference easy mathematically by making more assumptions (adding more structure to the model) and making the model more flexible (not making the model too restrictive).

Most software, by default, places four conditions on the distribution of the error term in the model. We refer to this collection of conditions as the "classical regression model."

Definition 4.3 (Classical Regression Model)**.** In the "classical regression model," we place the following four conditions on the distribution of the error ε_i:

1. The average error across all levels of the predictors is 0; mathematically, we write $E\left(\varepsilon_i \mid (\text{Predictors 1 - } p)_i\right) = 0$.
2. The variance of the errors is constant across all levels of the predictors; mathematically, we write $Var\left(\varepsilon_i \mid (\text{Predictors 1 - } p)_i\right) = \sigma^2$ for some unknown constant $\sigma^2 > 0$. This is sometimes referred to as homoskedasticity.
3. The error terms are independent; in particular, the magnitude of the error for one observation does not influence the magnitude of the error for any other observation.
4. The distribution of the errors follows a Normal distribution with the above mean and variance.

It would be a mistake to consider the above conditions only from a probabilistic perspective; wrestling with what these mean in practice is critical to understanding the model.

The first condition says the structure of the model is correct; that is, no variables were omitted and the functional form of the response is really determined by a linear combination of the predictors. Violations of this assumption are very serious and indicate a different model structure is needed. Essentially, if we believe this condition is not met, it means we should revisit the science and rationale behind the proposed model because it is likely invalid.

The second condition considers the precision with which the response is measured. The condition asserts that this precision is consistent across all possible values for the predictors. For example, consider the academic performance of two classes; this condition prohibits cases in which the grades for one class have a wider range than the grades for the other.

The third condition eliminates data for which measurements are related beyond sharing common values of the predictors in the model. For example, suppose we are modeling the height of a tree as a function of its age. All trees of a similar age may be "related" in the sense that we expect them to have similar heights; the model allows this. However, it does not allow for trees being "related" in the sense that trees in a similar region will share a similar height due to differences in resources among regions; this is prohibited because "region" is not captured by the model. In the biological sciences, this condition is often called into question when we take repeated measurements on subjects or when observations are measured close together in time. This type of data will be addressed later in the text (Chapter 13).

The last condition is a strong one; it states that we are able to fully characterize the distribution of the error terms. While the other conditions describe certain characteristics of the distribution, this says we know the exact form of the distribution. Historically, this condition was imposed to ensure the error terms were well behaved (and because the probability theory worked out nicely).

Statistics courses (especially the introductory course) focus on these four conditions on the error. However, the classical framework typically also imposes additional conditions on the predictors.

Definition 4.4 (Classical Regression (Conditions on Predictors))**.** The classical regression model (Definition 4.3) places the following conditions on the predictors:

1. Each predictor is measured without error.
2. Each predictor has an additive linear effect on the response.

The first condition states that there cannot be any noise present in the measurement of the *predictors*. For example, imagine modeling the length (or height) of infants as a function of their age. When the doctor asks for the age of the child, we are assuming that this age can be computed/measured without error. This seems reasonable when the predictor is age. However, consider using the temperature of the infant as a predictor in the model; if the thermometer is only accurate to within 2 tenths of a degree, than we may believe that the body temperature is measured with error. Addressing measurement error in models is beyond the scope of this text, and it is in general a difficult problem. Typically, even if a predictor is potentially measured

with error, we are able to assume the error is negligible compared to the amount of error in the response. Throughout the text, we will assume all predictors are measured without error.

The second condition on the predictors is very closely related to the condition on the errors that the mean of the errors is 0. If we are empirically building a model and find evidence that the model has been mis-specified, it is generally a result of the predictors not having a linear relationship with the response.

4.3 Alternate Characterization of the Model

Recall that a distribution is just the pattern of variability among the values of a variable; that is, a distribution describes how values differ from one another. Chapter 3 presented probability tools that can be used to model these distributions. We saw that it is possible to specify these models up to some unknown parameters; for example, we may write $X \sim N(\mu, \sigma^2)$ in order to say the density of the random variable X can be modeled using the following mathematical formula:

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}.$$

We often think about these parameters as being a single value, but nothing prohibits that value from being described by a function of variables. That is, we could let $\mu = g(\text{Predictors})$ for some function g. In fact, the conditions on the error term specified in the previous section lead us to an alternate characterization of the general linear model.

Definition 4.5 (Alternate Characterization of the Classical Regression Model)**.** Under the classical regression conditions on the error term (see Definition 4.3), we can characterize the classical regression model as

$$(\text{Response})_i \mid (\text{Predictors 1 through } p)_i \overset{\text{Ind}}{\sim} N\left(\beta_0 + \sum_{j=1}^{p} \beta_j (\text{Predictor } j)_i, \sigma^2\right).$$

Here, the symbol | is read "given" and means that the distribution of the response is specified after knowing the values of the predictors. That is, the distribution of the response depends on these variables.

The alternate characterization of the regression model in Definition 4.5 is particularly useful in statistical theory, but that is not why we mention it here. We mention this form because it sheds light on the true nature of regression models (beyond just the classical regression model) — regression models characterize the distribution of the response.

Big Idea

Regression models allow the parameters characterizing the distribution of the population to depend on the predictors through some function.

Closely examining Definition 4.5, we see that the deterministic portion of the general linear model is actually characterizing the *mean* of the response (for specified values of the predictors). In fact, this realization is actually the direct result of the first ("mean 0") condition we placed on the error terms. This is what allows us to begin interpreting the parameters in the model.

4.4 Interpretation of Parameters

When we assume that the error in the response, on average, is 0 for all values of the predictor, we are really saying that the deterministic portion of the model defines the mean response. We see this in the alternate characterization of the regression model above where μ in the Gaussian (Normal) model is replaced by

$$\beta_0 + \sum_{j=1}^{p} \beta_j (\text{Predictor } j)_i.$$

Notice what happens if we plug zero in for *every* predictor:

$$\beta_0 + \sum_{j=1}^{p} \beta_j (0) = \beta_0.$$

Since this deterministic portion specifies the average response, then we see that the average response is β_0 when all predictors have the value zero.

Definition 4.6 (Intercept)**.** The population intercept, denoted β_0, is the *mean* response when all predictors take the value zero.

We should point out that while this is the correct interpretation, it may not always make sense in context. For example, if we are modeling the heart rate of patients as a function of their body temperature and weight; the model would have the form

$$(\text{Heart Rate})_i = \beta_0 + \beta_1 (\text{Body Temperature})_i + \beta_2 (\text{Weight})_i + \varepsilon_i.$$

Based on Definition 4.6, we would interpret the intercept in this model as the average heart rate for individuals with a body temperature of zero degrees and a weight of zero pounds;

as this group of individuals does not exist, the interpretation does not make sense in this context.

We now turn to considering an interpretation for the slope. Consider two groups of individuals:

- Group 1 has the value a for the first predictor and value x_j for Predictor j (for $j = 2, \ldots, p$).
- Group 2 has the value $a + 1$ for the first predictor and value x_j for Predictor j for $j = 2, \ldots, p$.

That is, the only way the two groups differ is that Group 2 has increased the value of the first predictor by 1. From our model, we have that the average response for Group 1 is

$$\beta_0 + \beta_1 a + \sum_{j=2}^{p} \beta_j x_j.$$

The average response for Group 2 is

$$\beta_0 + \beta_1 (a + 1) \sum_{j=2}^{p} \beta_j x_j.$$

Consider taking the difference in these two *mean* responses (Group 2 minus Group 1):

$$\beta_0 + \beta_1 (a + 1) \sum_{j=2}^{p} \beta_j x_j - \left(\beta_0 + \beta_1 a + \sum_{j=2}^{p} \beta_j x_j \right) = \beta_1.$$

That is, the slope is the difference in the *mean* response between the two groups.

Definition 4.7 (Slope)**.** The coefficient for the j-th predictor, denoted β_j, is the change in the mean response associated with a one unit increase in Predictor j, *holding all other predictors fixed.*

The last part of Definition 4.7 is a critical part of the interpretation, and it is critical to the full utility of regression models. Again, while holding all other predictors fixed may not be practically feasible (for example, could we really increase an individual's height without also increasing their weight), it allows us to investigate the impact of a predictor separate from other variables.

Interpretation of the parameters is a large step beyond simply estimating the parameters. However, we still have not developed the tools to do much beyond estimation. We now turn our attention to inference.

4.5 Inference About the Mean Parameters

As suggested in Chapter 2, the key to making formal inference on the parameters of a population is to develop a model for the sampling distribution (or null distribution) of the corresponding statistics. Under the classical regression conditions of Definition 4.3, we are able to form an exact model for the sampling distribution of the least squares estimates.

> **Note**
>
> While beyond the scope of this course, it can be shown that the least squares estimates of the parameters are linear combinations of the observed responses. This, combined with the modeling assumptions, allows us to construct a model for the sampling distribution of the estimates.

Definition 4.8 (Sampling Distribution of the Least Squares Estimates)**.** Under the classical regression conditions (Definition 4.3), we have that

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{Var\left(\hat{\beta}_j\right)}} \sim t_{n-p-1}.$$

The denominator $\sqrt{Var\left(\hat{\beta}_j\right)}$ is known as the *standard error* of the estimate $\hat{\beta}_j$. This formula holds for all $j = 0, 1, \ldots, p$.

Definition 4.8 states the standardized difference between our estimate and the parameter follows a t-distribution, where the degrees of freedom depend on the sample size and the number of parameters in the model. The specific model is not as important as knowing that under the classical regression conditions, an exact model is known. Nearly every software package that implements regression does so under the classical regression conditions, and the inference is based on the above model for the sampling distribution.

The detail-oriented reader will note that we did not include a formula for the standard error of an estimate. The formula is beyond the scope of this course, but it is a function of the values of the predictor as well as the variability in the error term. You see, the moment we specified the second condition ("constant variance"), we introduced another parameter: σ^2. The parameter σ^2 does not govern the mean response; so, it tends to be of less direct interest for our purposes. Instead, it characterizes the variability in the response (for a given set of predictors), and it plays a role in inference (as we see in the above model for the sampling distribution of the least squares estimates of the parameters in the mean model). It will therefore play a role in computing confidence intervals and p-values. Since it is unknown, it must also be estimated.

Definition 4.9 (Estimate of the Variance of the Errors)**.** The unknown variance in the linear model, which captures the variability in the response for any set of predictors (also called the residual variance), is estimated by

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}\left((\text{Response})_i - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j(\text{Predictor } j)_i\right)^2.$$

Note that the estimate of the variance depends upon the least squares estimates. Of more interest is that the scaling factor $(n-p-1)$ is the same as the degrees of freedom for the sampling distribution; that is not an accident.

A model for the sampling distribution is the holy grail of statistical inference. It can be updated to determine the model for the null distribution. And, once you have a model for the sampling distribution in hand, you can wield it to construct a confidence interval (and null distributions to yield p-values).

Definition 4.10 (Confidence Interval for Parameters Under Classical Model)**.** Under the classical regression conditions (Definition 4.3), a $100c\%$ confidence interval for the parameter β_j is given by

$$\hat{\beta}_j \pm t_{n-p-1,0.5(1+c)}\sqrt{Var\left(\hat{\beta}_j\right)}.$$

where $t_{n-p-1,0.5(1+c)}$ is the $0.5(1+c)$ quantile from the t_{n-p-1} distribution, known as the critical value for the confidence interval.

Like many confidence intervals, the idea is that we are grabbing the middle portion of the model for the sampling distribution. The confidence interval represents the values of the parameter for which the data is consistent — the reasonable values of the parameter based on the observed data. Also note that this confidence interval is specified for each parameter individually.

> **Note**
>
> For large values of n relative to p, the critical value for a 95% confidence interval is approximately 1.96. Hence, a rough confidence interval is therefore 2 standard errors in either direction of the point estimate.

Definition 4.11 (P-Value for Testing if Parameter Belongs in Model Under Classical Model)**.** Under the classical regression conditions (Definition 4.3), the p-value for testing the hypotheses

$$H_0 : \beta_j = 0 \qquad \text{vs.} \qquad H_1 : \beta_j \neq 0$$

is given by

$$Pr\left(|T| > \left|\frac{\hat{\beta}_j}{\sqrt{Var\left(\hat{\beta}_j\right)}}\right|\right)$$

where $T \sim t_{n-p-1}$.

Definition 4.11 highlights that the null distribution for the standardized ratio is developed by taking the model for the sampling distribution and enforcing the null hypothesis $\left(\beta_j = 0\right)$. Using this null distribution, our p-value then summarizes how likely it is we would obtain a value of the standardized statistic at least as large of that observed by chance alone when the null hypothesis is true.

The interpretation of the confidence interval and p-value follows the interpretation of the confidence intervals and p-values computed in an introductory course (and reviewed in Chapter 1. This section just establishes that the conditions we placed on the error term yield explicit formulas for their computation (even if these formulas are implemented in the background of the software).

The framework introduced here provides the basics for making inference using a statistical model. As we consider more flexible modeling strategies, these key concepts do not leave us. We need a model for the sampling distribution or null distribution in order to make inference. And, the model for that distribution depends on the conditions we are willing to make.

> **Big Idea**
>
> A model for the sampling distribution (and/or null distribution) is needed for making inference, and that model depends on the conditions we are willing to impose on the model for the data generating process.

5 Assessing the Conditions for the General Linear Model

In the previous chapter, we presented a model for the sampling distribution (Definition 4.8) of the least squares estimates. This model allows us to make inference on the unknown parameters that govern the mean response of the general linear model (Definition 4.1). However, our model for the sampling distribution presented assumes all the conditions for the classical regression model (Definition 4.3) hold. We should not blindly make assumptions. Instead, we should ensure our data is consistent with any conditions we impose.

The majority of the conditions in the classical regression model are placed on the error term, a random variable that we never observe in practice. This means that the conditions cannot be assessed using the errors directly. Instead, modeling conditions are assessed graphically using residuals.

Definition 5.1 (Residual)**.** A residual for the i-th observation is the difference between an observed value and the predicted response:

$$\begin{aligned}(\text{Residual})_i &= (\text{Observed Response})_i - (\text{Predicted Response})_i \\ &= (\text{Response})_i - \left(\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j (\text{Predictor } j)_i\right).\end{aligned}$$

Big Idea

If a condition holds on the distribution of the errors, we expect the residuals to adopt specific behavior. Therefore, while the conditions are placed on the error, they are *assessed* using the residuals. The conditions, however, are *not* about the residuals.

It is important to assess the conditions we place on the model as they determine the model for the sampling distribution (and null distribution). That is, if the conditions we have assumed are incorrect, our p-values and confidence intervals will be invalid.

Table 5.1: Method of graphical assessment for the conditions of the classical regression model.

Condition	Graphical Assessment
Error is 0, on average, for all predictors	Residual vs. Predicted Values
Errors are independent	Time-series Plot of Residuals
Homoskedasticity	Residual vs. Predicted Values
Errors are Normally distributed	Probability Plot of Residuals
No Measurement Error	None (discipline expertise)
Predictor enters linearly	Residual vs. Predictor

Big Idea

The assumptions we are willing to make about a data generating process (in particular, the random component of our regression model) determine the form of the model for the sampling distributions (null distributions) of the resulting estimates (standardized statistics).

Note

You will notice we often jump between "conditions" and "assumptions" as if they are interchangeable. They are often used synonymously in the literature. However, as a distinction, *conditions* are the mathematical properties that must be met in order to justify the statistical theory; in practice, we make *assumptions* about which conditions we believe are reasonable.
We can never prove a condition holds; therefore, we must always make assumptions.

As a general rule, if our data is consistent with the conditions we have assumed, then the error term is just noise; that is, it should not have any signal left. We would therefore expect the residuals to lack any patterns; if we find patterns in the residuals, it suggests there is some structure our model has ignored. While there are many methods available for detecting patterns in the residuals, we prefer a graphical approach. Since we cannot verify a condition holds, we will always be making assumptions. By taking a graphical approach to assessment, we are embracing the subjective nature of such investigations; other approaches can give the appearance that there is more certainty in the conclusion than actually exists.

Table 5.1 aligns the conditions we place on the model with the graphic used for assessment.

The conditions on the error are assessed in the same way they are in simple linear regression, covered in an introductory course. So, we only briefly review them here.

Using the Plot of the Residuals vs. Predicted Values

This plot is used to assess both the "mean 0" and "constant variance" conditions. We are looking for trends in the *location* and in the *spread*, respectively, to assess these conditions.

If we observe a trend in the *location* of the residuals as we move left-to-right on the graphic, we have evidence that the "mean 0" condition is violated. If we observe a trend in the *spread* of the residuals as we move left-to-right across the graphic, we have evidence that the "constant variance" condition is violated. As we move forward, we will examine techniques to relax the condition of constant variance. While there is no "fix" for violations of the "mean 0" condition, we will study scenarios for which the response and predictors are not linearly related. What is amazing about this graphic is that we have reduced a multi-dimensional problem to two dimensions.

Using the Time-Series Plot of the Residuals

This graph plots the residuals against the order in which the data was collected. Any trends in the residuals (in location *or* spread) indicates evidence the errors collected close together in time may be associated in some way, violating the assumption of independence.

The creation of a time-series plot is only reasonable/useful if we know the order in which the data was collected. A cross-sectional analysis (single snapshot in time), for example has no natural ordering. It is important to note that a time-series plot only allows us to examine dependence as a result of time. If the errors are correlated due to some other factor (for example, geographical location or family groups), the violation may not be detected. When we are able to identify the dependence structure, we can incorporate it into the model, as discussed in later units (Chapter 13). Regardless of whether the graphic can be constructed, a thorough assessment of independence always requires discipline expertise and a critical review of the data collection plan.

Using the Probability Plot of the Residuals

Also called a "QQ Plot," a probability plot examines the relationship between the observed residuals and the expected (or "theoretical") values we would expect if the errors followed a Normal distribution. Any departures from a straight line indicate evidence the errors do not follow a Normal distribution.

The assumption of Normality is the strongest of the four conditions; that is, this condition goes beyond characterizing an aspect of the error distribution to specifying the functional family to which the distribution belongs. As we will see, this condition is the easiest to relax.

Recall that in addition to conditions on the stochastic portion of the model, we have considered

conditions on the predictors as well (Definition 4.4).

> **Assessing Measurement Error in the Predictors**
>
> Determining whether a predictor is subject to measurement error relies on discipline expertise.

If the predictors are measured with (non-negligible) error, our estimates are biased and therefore unreliable. Methods for addressing predictors that are subject to measurement error are beyond the scope of the text.

Requiring that the predictors enter the model linearly is a refinement of the "mean 0" condition; that is, one way that we often misspecify the deterministic portion of the model is to attempt to model curvature in the data using a line. It should not be a surprise then that assessing the linearity of the predictors is similar to assessing the "mean 0" condition.

> **Assessing Linearity of the Predictors**
>
> If the relationship between the response and predictor is adequately explained by a linear relationship, then there should not be any structure in the *location* of the residuals when examined against the predictor.

That is, when assessing the "mean 0" condition, we examine the residuals against the predicted values; when assessing the linearity condition, we examine the residuals against each *quantitative* predictor.

> **Note**
>
> We will discuss the incorporation of categorical predictors in the next chapter; however, we note that the linearity assumption only applies to quantitative predictors.

The data will not always be consistent with the conditions we would like to place on the model. Proceeding as if the conditions are reasonable when they are not can lead to invalid inference and incorrect conclusions. Discarding the results misses out on potential insights the data offers. Fortunately, the modeling framework is flexible enough to be relaxed to address violations of the conditions, which we examine toward the end of this unit in the text.

6 The General Linear Model as a Unifying Framework

The general linear model (Definition 4.1) is much more flexible and powerful than we might initially imagine. The full flexibility of this modeling framework is explored in the next unit. It should be clear that simple linear regression is a special case of the general linear model for which we only have a single predictor. In the remainder of this chapter, we outline how the methods typically studied in an introductory course also relate to the general linear model framework.

6.1 One Sample Inference

Perhaps the most cited analysis technique from an introductory statistics course is the "1-sample t-test." In brief, this test considers the hypotheses

$$H_0 : \mu = \mu_0 \qquad \text{vs.} \qquad H_1 : \mu \neq \mu_0$$

where μ is the average response in the population. These hypotheses are making inference on the mean response of a single population (or "one sample"). A classical introductory statistics course will introduce the test statistic

$$T^* = \frac{\sqrt{n}\,(\bar{y} - \mu_0)}{s}$$

where $\bar{y}$ and s represent the sample mean and sample standard deviation, respectively, of the response. The one-sample t-test proceeds to model the (null) distribution of T^* as a t-distribution with $n - 1$ degrees of freedom. This test assumes that

1. The response for one observation is independent of the response for all other observations.
2. The responses are identically distributed.
3. The responses follow a Normal distribution.

We can recover the same analysis using the general linear model. Consider the model

$$(\text{Response})_i = \mu + \varepsilon_i.$$

This is sometimes referred to as the "intercept-only" model. It can be shown that the least-squares estimate of μ is the sample mean of the response. Similarly, our estimate of σ^2 is the sample variance. Further, we have that

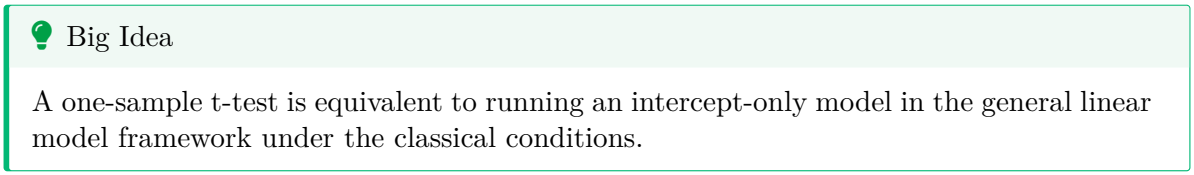
$$Var\left(\hat{\mu}\right) = \frac{s^2}{n},$$

meaning the standardized statistic in Definition 4.8 is equivalent to the ratio taught in the introductory statistics course.

Having the same standardized statistic is one thing, but the analysis is only equivalent if the conditions imposed also agree. If you consider the classical regression conditions (Definition 4.3), then given that there is no predictor in the model, the conditions would translate to

1. The response for one observation is independent of the response for all other observations.
2. The responses are identically distributed.
3. The responses follow a Normal distribution.

You may wonder where the "mean 0" condition went. Assuming the error is 0 on average is equivalent to assuming the deterministic portion of the model is correctly specified. In the case when the deterministic model does not have a predictor, we are essentially assuming that μ is the average (the sample is representative of the underlying population). In particular, since we made no simplifying assumptions about the structure of the relationship between the response and predictor (since there is no predictor), those simplifying assumptions cannot be incorrect.



6.2 Paired t-Test

Most analyses covered in an introductory class assume independence between all observations. The most notable exception is the "paired t-test." In this scenario, we collect a total of $2n$ observations, but for a total of n independent pairs. For example, n participants complete a pre and post test. Or, we measure a response on both the left and right eye for n participants.

In each of these examples, there are two measurements for each of the n participants that we believe are related (general ability with exams means some students will naturally score higher; genetics result in some individuals having better vision; etc.). The typical way of addressing this problem in the introductory course is to take the difference in the response within each pair, and then conduct a one-sample t-test. That is, we consider

$$T^* = \frac{\sqrt{n}\left(\bar{y}_d - \mu_{d,0}\right)}{s_d},$$

where $\bar{y}_d$ is the sample mean of the differences in the response, s_d is the sample standard deviation of the differences, and n is the number of paired observations.

Since this is a one sample test, building on our discussion earlier in this chapter, we have that an equivalent approach is to consider the model

$$(\text{Difference in Responses})_i = \mu + \varepsilon_i$$

with the classical conditions imposed.

The paired t-test is actually a special case of a "repeated measures ANOVA," which can also be viewed as a special case of the general linear model. We discuss repeated measures and appropriate approaches for analysis in a future unit.

> **Big Idea**
>
> A paired t-test is equivalent to running an intercept-only model, with the pairwise differences as the response, in the general linear model framework under the classical conditions.

6.3 Group Comparisons

Suppose we are interested in comparing the average response across two or more groups. If there are only two groups, our hypotheses take the form

$$H_0 : \mu_1 = \mu_2 \qquad \text{vs.} \qquad H_1 : \mu_1 \neq \mu_2,$$

where μ_1 and μ_2 represent the average response for each of the two groups. The two-sample t-test is typically discussed to address this question where

$$T^* = \frac{(\bar{y}_2 - \bar{y}_1)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

where $\bar{y}_j$ and s_j^2 are the sample mean and sample variance of the response within group j ($j = 1, 2$), respectively. A t-distribution is used to model the distribution of this standardized statistic under the following conditions:

1. The response from one observation is independent of the response of all other observations (implying independence both within and between groups).
2. The responses *within* a group are identically distributed.
3. The responses within each group follow a Normal distribution.

When there are three or more groups, our hypotheses take the form

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \qquad \text{vs.} \qquad H_1 : \text{at least 1 } \mu_j \text{ differs,}$$

where μ_j represents the average response for group j, $j = 1, 2, \ldots, k$. Analysis of variance (ANOVA) is typically used to address this question. Typically, ANOVA imposes the following conditions:

1. The response from one observation is independent of the response of all other observations (implying independence both within and between groups).
2. The variability in the response within a group is the same for each group.
3. The responses within each group follow a Normal distribution.

It would seem that the two-group comparison is a special case of ANOVA; however, as stated above, they would yield slightly different inference because the conditions imposed differ. Specifically, ANOVA assumes the variance of the response in a group is the same across groups; however, the two-sample t-test allows the variance of the response to differ across groups.

> **Note**
>
> There is a version of the two-sample t-test which uses the "pooled" sample variance; in that case, the variance of the response is assumed to be the same within each group and the two-sample t-test is a special case of ANOVA.

It would therefore seem that there is no way these methods relate to the general linear model framework. However, the general linear model framework does encompass both approaches. An ANOVA can be accomplished by inserting a single categorical predictor into the model capturing the grouping structure (see Chapter 8). Since the classical regression assumptions correspond to ANOVA, we recover the same inference. The two-group comparison requires that we additionally relax the constant variance condition, which we discuss in Chapter 17 (while discussed in the context of non-linear models, the same techniques apply to the general linear model as well).

Big Idea

ANOVA is equivalent to running the general linear model with a single categorical predictor under the classical conditions.

Part III

Unit III: General Modeling Techniques

In the previous unit, we introduced the general linear model. In this unit, we use the general linear model as the backdrop for introducing several modeling strategies that allow us to address common questions in scientific research. While these strategies are introduced in the context of the linear model, they can be applied with all the models discussed in the text.

7 Addressing Relationships Between Predictors

Chapter 4 introduced the framework of the general linear model, including the interpretation of the parameters (Definition 4.7 and Definition 4.6). While the mathematical structure of the general linear model may be fascinating to some, we are particularly interested in the model's utility to address scientific questions. Therefore, for our purposes, the interpretation of the parameters is of utmost importance. And, immediately from the interpretation of the slope coefficients, the linear model framework allows to isolate the effect of one variable while holding all other variables constant. This has implications on the conclusions we can draw from the model.

7.1 Adjusting for Confounders

Scientific studies can be roughly categorized as being an observational study (Definition 1.14) or a controlled experiment (also called a randomized clinical trial, Definition 1.13). When our scientific question centers on the relationship between a response and a predictor, and the values of the predictor have not been randomly assigned to subjects, the (observational) study is subject to confounding (Definition 1.15). It is the potential for confounding that leads to the often cited "correlation does not imply causation." It does not take much to see that this is extremely limiting. We can imagine randomizing subjects to different levels of a categorical predictor, but randomizing subjects to a quantitative predictor would require very large sample sizes. For example, imagine randomly allocating subjects to the amount of water they consume in their diet each day — think of all the amounts of water you might want to consider. Further, it would rule out ever being able to causally link a response with a quantitative predictor which represents an inherit characteristic. For example, randomly allocating a participant to a specific height would require adding or removing bone mass to alter their height — which we hope goes without saying is unethical and unacceptable (not to mention just disturbing).

Multivariable models, however, naturally address confounding because of the interpretation of the coefficients: *holding all other predictors fixed* (we said this would be a crucial phrase). This interpretation suggests the coefficient attached to a predictor is quantifying the effect of the predictor on the response, isolated from all other predictors in the model. By isolating the predictor's effect, we are able to see its impact on the response beyond the impact of any other predictors. We often say that the model is "adjusted" for the other predictors.

Adjusted Models

An "adjusted" model just means that we constructed a multivariable model. The relationship between the response and the predictor of interest is "adjusted" by the other predictors that appear in the model.

Think again about the importance of random allocation in allowing for causal interpretations in controlled experiments. The random allocation of subjects to groups means that the groups are similar *with respect to all other variables* — the only difference between the groups is the treatment received. The "holding all other predictors fixed" phrase accomplishes something similar.
We know that the deterministic portion of the general linear model characterizes the mean response. That is, the deterministic portion tells us the average value of the response we should expect among the *group* of individuals with a particular value of the predictors. Therefore, when we increase a predictor by one unit *holding all other predictors fixed*, we are creating two groups where the only difference is that increase of one unit. The two groups are similar *with respect to all other predictors in the model*.

This seems almost too good to be true. By simply expanding our model to incorporate other predictors, we have addressed the issue of confounding, and we have gotten back a causal interpretation of the impact of the predictor on the response. But, it is important to note the differences between a controlled experiment and the interpretation provided by a multivariable model:

- Controlled experiment: the groups are similar *with respect to all other variables*.
- Multivariable model: the groups are similar *with respect to all other predictors*.

The difference in the language is subtle but important. We can only adjust for predictors that we observe and put in the model. That is, a multivariable model does not automatically solve all confounding issues; it ensures that any potential confounding cannot be the result of the other predictors in the model. This allows us to make causal conclusions if we can assume that all potential confounders are present as other predictors in the model.

Big Idea

A multivariable model allows us to isolate the effect of one variable from the other predictors; this allows us to state that those other predictors are not contributing to any potential confounding between the variable of interest and the response.

7.2 Multicollinearity

A confounder masks or exaggerates the effect of one predictor on the response. In order for confounding to exist, the confounder must be related to both the response and the predictor. A distinctly unique, but often confused topic, is that of multicollinearity.

Definition 7.1 (Multicollinearity)**.** When two predictors are highly correlated with one another, we say that there is multicollinearity in the model.

To understand the impact, we consider a very simple hypothetical example. Suppose we are interested in modeling the number of steps taken over the course of a typical day (as recorded by popular fitness trackers) using the participant's height and stride length. Our model would have the form

$$(\text{Number of Steps})_i = \beta_0 + \beta_1(\text{Height})_i + \beta_2(\text{Stride Length})_i + \varepsilon_i.$$

Suppose that we test the following two sets of hypotheses:

$$\begin{aligned} H_0 &: \beta_1 = 0 \qquad \text{vs.} \qquad H_1 : \beta_1 \neq 0 \\ H_0 &: \beta_2 = 0 \qquad \text{vs.} \qquad H_1 : \beta_2 \neq 0. \end{aligned}$$

Further, suppose that we have a large p-value for each of these tests.

How would we interpret the large p-value for the first test? It would tell us that there is no evidence of a relationship between the average number of steps taken and the participant's height, *holding their stride length fixed*. The idea of holding all other predictors fixed plays an important part here. The large p-value for this first hypothesis tells us that there is no evidence the participant's height is helpful for predicting the number of steps taken *after accounting for their stride length*. Similarly, the large p-value for the second hypothesis tells us that there is no evidence the participant's stride length is helpful for predicting the number of steps taken *after accounting for their height*. However, we know that a person's stride is very correlated with their height — those who are taller have longer legs and tend to have a longer stride. That is, there is not much information that a person's stride length will tell us that their height does not already convey; that explains the results we are seeing here. It is not that the height is not associated with the number of steps taken; it is that it is not helpful *above and beyond* knowing the participant's stride length. This is multicollinearity.

Multicollinearity can make inference on a single predictor misleading. The p-value is not incorrect; we just need to remember that it is testing whether the term belongs after accounting for other predictors in the model. That is, the regression model is trying to isolate the effect above and beyond that of other predictors in the model.

> **Note**
>
> A tell-tale sign of multicollinearity between two predictors is that alone, each is significantly associated with the response; but, when both are placed in the model, neither appears significant.

Multicollinearity is not necessarily a problem. If our primary aim in constructing the regression model is prediction, then we are not concerned with the inference on each parameter individually. The estimates of the parameters are valid; as a result, we can leave both predictors in the model. However, if our goal is inference on the parameters to further characterize the relationship, then the standard errors are misleading (leading to unreliable p-values and confidence intervals). The solution is to remove the predictor that is less likely to be on the "causal pathway" which requires discipline expertise.

> **Big Idea**
>
> When two predictors capture the same information, placing both in the model can lead to misleading p-values and confidence intervals for each predictor. However, if the primary aim is prediction, this is not generally a concern.

8 Incorporating Categorical Predictors

The general linear model framework (Definition 4.1) is quite flexible; in particular, it allows us to consider not only quantitative predictors, but also categorical (qualitative) predictors. Our strategy is to define a new set of quantitative variables which capture the group membership appropriately.

Example 8.1 (Perceived Stress)**.** The *Perceived Stress Scale (PSS)* is a widely used psychological instrument for measuring the perception of stress. Subjects answer ten short questions regarding the degree to which situations in their life are viewed as stressful, and the responses are codified into a score between 0 and 40 (higher values indicate higher stress). Suppose we were interested in modeling the PSS score among college students as a function of their class standing (Freshman, Sophomore, Junior, Senior) and the number of hours of sleep the student reports getting on a typical night. The first few records in our data might hypothetically look like that illustrated in Table 8.1.

It does not take long to recognize that forming a model like

$$(\text{PSS Score})_i = \beta_0 + \beta_1(\text{Hours Sleep})_i + \beta_2(\text{Class Standing})_i + \varepsilon_i$$

does not work. Since class standing is a categorical variable, plugging in does not make sense; that is, what does it mean to multiply β_2 by "Junior"? We need a way of somehow bringing the categorical predictor into the linear model. Before stating our approach, let's first consider two common naive approaches:

- Replace each level of the categorical predictor with a number: convert Freshman to 1, Sophomore to 2, Junior to 3, and Senior to 4; enter this numeric variable into a regression model.
- Construct different data sets for each level of the categorical predictor: four data sets in this case with one for Freshman, one for Sophomore, one for Junior, and one for Senior; conduct a different analysis on each set of data.

The first approach solves the "number times word" problem. We could certainly fit such a model. However, this approach is limiting. It assumes a linear trend across the levels of the categorical predictor. Are we sure that the stress either increases or decreases as the class standing increases? Do we want to allow the stress to be highest during the sophomore year?

Table 8.1: Hypothetical data on stress in college students.

Subject ID	PSS	Hours Sleep	Class Standing
1415	14	7.5	Freshman
1463	25	8.5	Senior
1179	26	7.0	Junior
1526	27	8.5	Senior
1195	5	8.0	Sophomore
1938	27	5.0	Freshman
1818	28	5.5	Junior
1118	9	4.0	Freshman
1299	29	9.0	Freshman
1229	35	7.0	Sophomore

More problematic are categorical predictors which have no natural ordering (e.g., eye color); how do we determine the mapping from text to numbers in that case?

The second approach sounds reasonable at first glance. It would yield four different models:

$$\begin{aligned}
\text{Model 1 :}&(\text{PSS Score})_i = \gamma_{\text{FR}} + \alpha_{\text{FR}}(\text{Hours Sleep})_i + \varepsilon_{1,i} \\
\text{Model 2 :}&(\text{PSS Score})_i = \gamma_{\text{SO}} + \alpha_{\text{SO}}(\text{Hours Sleep})_i + \varepsilon_{2,i} \\
\text{Model 3 :}&(\text{PSS Score})_i = \gamma_{\text{JR}} + \alpha_{\text{JR}}(\text{Hours Sleep})_i + \varepsilon_{3,i} \\
\text{Model 4 :}&(\text{PSS Score})_i = \gamma_{\text{SR}} + \alpha_{\text{SR}}(\text{Hours Sleep})_i + \varepsilon_{4,i}.
\end{aligned}$$

In these models, we have different parameters for each group. The problem is that we no longer have a single estimate for the impact of the number of hours of sleep; we have a different estimate for each group. Further, we would have a different estimate of the residual variance for each model, which would not align with the condition of assuming the variance is constant for all values of the predictors. This approach diminishes the power of the study, and it does not make it easy to address some questions of interest (such as, do freshman and sophomores differ in their PSS score?).

Neither of these approaches seems to fully capture our goal. Instead, we create multiple new variables that capture the qualitative grouping. Consider defining new variables as follows

Table 8.2: Hypothetical data on stress in college students augmented to include additional variables capturing the class standing.

Subject ID	PSS	Hours Sleep	Class Standing	Sophomore	Junior	Senior
1415	14	7.5	Freshman	0	0	0
1463	25	8.5	Senior	0	0	1
1179	26	7.0	Junior	0	1	0
1526	27	8.5	Senior	0	0	1
1195	5	8.0	Sophomore	1	0	0
1938	27	5.0	Freshman	0	0	0
1818	28	5.5	Junior	0	1	0
1118	9	4.0	Freshman	0	0	0
1299	29	9.0	Freshman	0	0	0
1229	35	7.0	Sophomore	1	0	0

$$(\text{Sophomore})_i = \begin{cases} 1 & \text{if i-th subject is a sophomore} \\ 0 & \text{otherwise} \end{cases}$$

$$(\text{Junior})_i = \begin{cases} 1 & \text{if i-th subject is a junior} \\ 0 & \text{otherwise} \end{cases}$$

$$(\text{Senior})_i = \begin{cases} 1 & \text{if i-th subject is a senior} \\ 0 & \text{otherwise.} \end{cases}$$

Augmenting our original data set with these new predictors would result in the data set illustrated in Table 8.2.

Using these additional variables, consider the model

$$\begin{aligned}(\text{PSS Score})_i = \beta_0 + \beta_1(\text{Hours Sleep})_i + \beta_2(\text{Sophomore})_i \\ + \beta_3(\text{Junior})_i + \beta_4(\text{Senior})_i + \varepsilon_i.\end{aligned}$$

This model embeds the grouping structure while only having one parameter for the effect of sleep on the PSS score and one parameter for the residual variance. You might at first think "what happened to freshman?" To see what is really happening with this model, think about what the structure provides us. Suppose we are considering freshman students who get x hours of sleep; for this group of students, the value of the variables `Sophomore`, `Junior`, and `Senior` are all zero. Plugging into the deterministic portion of the model, we find that the average PSS score for this group is

$$\beta_0 + \beta_1 x + \beta_2(0) + \beta_3(0) + \beta_4(0) = \beta_0 + \beta_1 x.$$

The fact that each of the variables `Sophomore`, `Junior`, and `Senior` take either the value 0 or 1 makes the arithmetic work out nicely. We can easily write down the average PSS score for each group for a specific number of hours of sleep:

$$\begin{aligned}
E\left[\text{PSS Score} \mid \text{Hours Sleep, Freshman}\right] &= \beta_0 + \beta_1(\text{Hours Sleep}) \\
E\left[\text{PSS Score} \mid \text{Hours Sleep, Sophomore}\right] &= (\beta_0 + \beta_2) + \beta_1(\text{Hours Sleep}) \\
E\left[\text{PSS Score} \mid \text{Hours Sleep, Junior}\right] &= (\beta_0 + \beta_3) + \beta_1(\text{Hours Sleep}) \\
E\left[\text{PSS Score} \mid \text{Hours Sleep, Senior}\right] &= (\beta_0 + \beta_4) + \beta_1(\text{Hours Sleep}).
\end{aligned}$$

So, freshman did not disappear from the model; they were there all along in the intercept. This strategy relies on capturing the grouping structure through a series of binary (0 or 1) variables, known as indicator variables.

Definition 8.1 (Indicator Variables)**.** Also called "dummy variables," these are a set of binary variables that capture the grouping defined by a categorical variable for regression modeling.

Indicator variables are like light switches that click on or off in order to specify that a particular subject (or population of subjects) with the corresponding characteristic is being considered. The way that we have defined these variables ensures that no two light switches are on at the same time; each subject is a member of exactly one group (a student must have a class standing and cannot have two class standings simultaneously; that is, a student cannot be classified as both a freshman and a sophomore).

> **Note**
>
> A categorical predictor with k groups/levels requires $k - 1$ indicator variables to fully capture the grouping structure.

One group (known as the reference group) will always be captured by the intercept term; the choice of this group is arbitrary and is often chosen by the software package (perhaps alphabetically, for example). Note that if the only difference between two models is the choice of the reference group, the models result in equivalent inference (though the interpretation of the parameters differs).

Definition 8.2 (Reference Group)**.** The group defined by having all indicator variables for a particular categorical variable set to zero.

Recall that provided the "mean 0" condition on the error holds, we have an interpretation of the coefficients in the model (Definition 4.7). This yields a nice interpretation of the coefficients for indicator variables.

Interpretation of Coefficient for Indicator

Let β be the parameter corresponding to an indicator variable in a linear model; then, β is the difference in the *average* response between the group defined by that indicator taking the value 1 and the reference group *holding all other predictors fixed*.

This does create a situation we have not yet encountered. Suppose we are interested in determining if the PSS score is associated with class standing after accounting for the hours of sleep the student gets on a typical night. The hypothesis is no longer of the form

$$H_0 : \beta_j = 0 \qquad \text{vs.} \qquad H_1 : \beta_j \neq 0$$

for some j. Instead, there are several predictors in the model which capture class standing. We need to instead consider testing multiple parameters *simultanesously*.

Big Idea

The statistical significance of a categorical predictor is assessed by testing if *all* corresponding indicator variables are simultaneously 0.

In our case, we would be testing a hypothesis of the form

$$H_0 : \beta_2 = \beta_3 = \beta_4 = 0 \qquad \text{vs.} \qquad H_1 : \text{at least one of these } \beta_j \text{ not equal to } 0.$$

We will address hypotheses of this form in Chapter 10.

We end this section by stating that while our discussion centered on the inclusion of categorical predictors in the linear model, this is a general modeling technique. Regardless of the type of regression model, categorical predictors can be included through the use of indicator variables.

9 Allowing Effect Modification (Interaction Terms)

Chapter 4 outlined three broad uses for creating a multivariable model. Chapter 7 highlighted the benefit of *isolating* an effect. In this chapter, we discuss how to examine the *interplay* between two predictors. That is, we allow the effect of one predictor to be modified (or dependent upon) the value of another predictor. This is done through "interaction" terms in the model

> **Warning**
>
> "Interactions" are *not* about the relationship between predictors. Our goal is still to examine the relationship between the response and the predictor. An interaction allows that relationship to depend upon another predictor.

Consider a linear model which views the response as a function of two quantitative predictors and a categorical variable with only two groups; that is,

$$(\text{Response})_i = \beta_0 + \beta_1(\text{Predictor 1})_i + \beta_2(\text{Predictor 2})_i + \beta_3(\text{Group B})_i + \varepsilon_i \tag{9.1}$$

where

$$(\text{Group B})_i = \begin{cases} 1 & \text{if i-th participant belongs to Group B} \\ 0 & \text{if i-th participant belongs to Group A.} \end{cases}$$

This model says there is a single effect of the second predictor on the response; that is, the effect of Predictor 2 is the same for participants in Group A as it is for those in Group B. What if we believe the effect is of Predictor 2 differs in the two groups? A naive approach would be to consider two separate models:

$$\begin{aligned} &\text{Model 1 (Group A Only)}: \quad (\text{Response})_i = \alpha_0 + \alpha_1(\text{Predictor 1})_i + \alpha_2(\text{Predictor 2})_i + \varepsilon_{1,i} \\ &\text{Model 2 (Group B Only)}: \quad (\text{Response})_i = \gamma_0 + \gamma_1(\text{Predictor 1})_i + \gamma_2(\text{Predictor 2})_i + \varepsilon_{2,i}. \end{aligned}$$

This creates some of the same problems we saw with creating multiple models in order to address categorical predictors (see Chapter 8). In particular,

1. We obtain two estimates of residual variance, but neither estimate is using the full data.
2. While we accomplished our goal of letting the effect of Predictor 2 differ in each group, we also allowed the effect of Predictor 1 to differ in each group. If we believe the effect of Predictor 1 is the same across the two groups, allowing it to be estimated separately in each group is inefficient.
3. It makes it difficult to clearly conduct a hypothesis test to determine if there is evidence the effects are actually different; that is, it is harder to quantify the evidence that $\alpha_2 \neq \gamma_2$.

We seek a modeling structure that allows the effect of the a predictor (Predictor 2 in the model above) to change for each value of the second predictor (the group structure in this case). That is, we would like to capture the possibility illustrated in Figure 9.1.

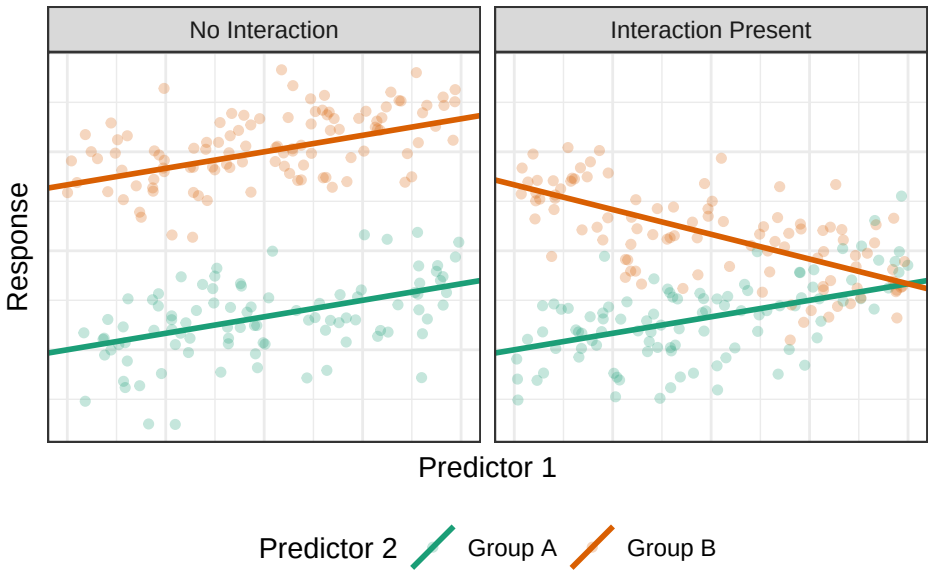


Figure 9.1: Illustration of the interaction of a quantitative predictor and a qualitative predictor when predicting a quantitative response.

Consider the following model:

$$\begin{aligned}(\text{Response})_i = \beta_0 + \beta_1(\text{Predictor 1})_i + \beta_2(\text{Predictor 2})_i \\ + \beta_3(\text{Group B})_i + \beta_4(\text{Predictor 2})_i(\text{Group B})_i + \varepsilon_i\end{aligned} \tag{9.2}$$

To motivate the development of this model, let's return to Equation 9.1. Remember, our goal is to allow the effect of Predictor 2, captured by β_2, to depend on the group to which the participant belongs. Imagine replacing β_2 with

$$\beta_2 = \eta_0 + \eta_1(\text{Group B})_i.$$

Notice that this says the *effect* β_2 can depend on whether the participant is in Group B; if the participant is in Group A, then the indicator is 0 and $\beta_2 = \eta_0$. If, on the other hand, the participant is in Group B, then the indicator is 1 and $\beta_2 = \eta_0 + \eta_1$. Substituting this in, we have

$$\begin{aligned}(\text{Response})_i &= \beta_0 + \beta_1(\text{Predictor 1})_i + (\eta_0 + \eta_1(\text{Group B})_i)\,(\text{Predictor 2})_i \\ &\quad + \beta_3(\text{Group B})_i + \varepsilon_i \\ &= \beta_0 + \beta_1(\text{Predictor 1})_i + \eta_0(\text{Predictor 2})_i \\ &\quad + \eta_1(\text{Predictor 2})_i(\text{Group B})_i + \beta_3(\text{Group B})_i + \varepsilon_i.\end{aligned}$$

Recognizing that the choice of Greek letters η_0 and η_1 are arbitrary, we have the same model as in Equation 9.2.

Equation 9.2 adds another variable to the model that is the product of the second predictor and the group indicator. Let's examine the structure that is provided here. Suppose we are interested in examining the mean response for subjects in Group A; the value of the group indicator is 0 for these subjects, leading to

$$E\,[\text{Response} \mid \text{Predictors, Group A}] = \beta_0 + \beta_1(\text{Predictor 1}) + \beta_2(\text{Predictor 2}).$$

For subjects in Group B, the value of group indicator is 1, leading to the mean response

$$E\,[\text{Response} \mid \text{Predictors, Group B}] = (\beta_0 + \beta_3) + \beta_1(\text{Predictor 1}) + (\beta_2 + \beta_4)\,(\text{Predictor 2}).$$

This allows both the intercept and the slope associated with Predictor 2 to differ between the two groups. That is, it allows not only a "bump" for being in Group B to the mean response, but it also allows the *effect* of the second predictor to differ for the two groups.

> **Big Idea**
>
> In order to capture complex modeling structures, we embed those structures in a large model as opposed to fitting several smaller models in different subgroups of the population.

We prefer the strategy in Equation 9.2 to performing a subgroup analysis.

Definition 9.1 (Subgroup Analysis)**.** Refers to repeating a specified analysis (e.g., regression model) within various levels of a categorical predictor.

- This will appropriately estimate the effect modification.

- This results in a loss of information because *all parameters* are forced to vary across the subgroups.

The new predictor added to our model in Equation 9.2 (the product term) is known as an interaction term.

Definition 9.2 (Interaction)**.** An interaction term allows the effect of a predictor on the response to depend on the value of a second predictor (capturing an effect modification).

- The interaction term is created by adding the product of the two predictors under consideration to the model.

> **Note**
>
> While we have illustrated the use of interaction terms using a quantitative and categorical predictor, interactions can be used with any type of predictors. For example, we could have considered the product of Predictors 1 and 2 in Equation 9.1 if we had wanted to.

While we have illustrated the use of interaction terms in a linear model, this is a general modeling technique that can be extended to other forms of regression models.

10 General Linear Hypothesis Test

We previously discussed a model for the sampling distribution of the parameter estimates (Definition 4.8) that allows for making inference on individual parameters; that is, we can test hypotheses of the form

$$H_0 : \beta_j = 0 \qquad \text{vs.} \qquad H_1 : \beta_j \neq 0.$$

However, we do not yet have a way of testing more complex hypotheses that occur regularly in scientific research. For example, testing whether the response is associated a categorical predictor involves determining if there is evidence that any of the coefficients associated with the indicator variables differs from zero. Such simultaneous tests fall under the General Linear Hypothesis framework.

While we write hypotheses as statements about parameters, we should keep in mind that they are really comparisons of alternative models for the response.

Big Idea

Hypothesis testing is a way of determining if a simpler model (under the null hypothesis) is sufficient for explaining the variability in the response or if a more complex model (under the alternative hypothesis) is necessary. The simpler model is the result of placing *constraints* on the complex model.
Statistical inference then allows us to determine if we can discern the difference between data generated under the simpler model and that observed. That is, we seek evidence that a more complex model is needed to capture the observed variability.

The vast majority of scientific questions can be framed as a hypothesis which places a constraint on a more complex model for the data generating process. These constraints can in turn often be written as a linear combination of the parameters.

Note

For those less familiar with matrix algebra, a linear combination simply a sum of parameters that have been multiplied by constants. We can write a linear combination as the product of two vectors, and a series of linear combinations is the product of a matrix and a vector. Let $\mathbf{K}$ be a 2-by-3 matrix defined as

$$\mathbf{K} = \begin{pmatrix} K_{1,1} & K_{1,2} & K_{1,3} \\ K_{2,1} & K_{2,2} & K_{2,3} \end{pmatrix}.$$

Let β be a column vector of length 3 defined as

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}.$$

The product $\mathbf{K}\beta$ is defined as

$$\mathbf{K}\beta = \begin{pmatrix} K_{1,1}\beta_1 + K_{1,2}\beta_2 + K_{1,3}\beta_3 \\ K_{2,1}\beta_1 + K_{2,2}\beta_2 + K_{2,3}\beta_3 \end{pmatrix}$$

which is a vector of length 2. Each element in $\mathbf{K}\beta$ is a linear combination of the elements of β.

Definition 10.1 (General Linear Hypothesis)**.** The general linear hypothesis framework refers to testing hypotheses of the form

$$H_0 : \mathbf{K}\beta = \mathbf{m} \qquad \text{vs.} \qquad H_1 : \mathbf{K}\beta \neq \mathbf{m}$$

where

- β is the $(p+1)$-length vector of the parameters (includes the intercept),
- $\mathbf{K}$ is an r-by-$(p+1)$ matrix that specifies the linear combinations defining the hypothesis of interest, and
- $\mathbf{m}$ is a vector of length r specifying the null values, the value of each linear combination under the null hypothesis (often a vector of 0's).

Before discussing inference for this hypothesis, we discuss the most common use of this framework. Consider the following linear model:

$$(\text{Response})_i = \beta_0 + \beta_1(\text{Predictor 1})_i + \beta_2(\text{Predictor 2})_i + \varepsilon_i.$$

Suppose we are interested in testing the following hypotheses:

$$H_0 : \beta_1 = \beta_2 = 0 \qquad \text{vs.} \qquad H_1 : \text{At least one } \beta_j \text{ not equal to 0.}$$

To express this in the general linear hypothesis framework, we must identify the matrix $\mathbf{K}$ and the vector $\mathbf{m}$. To do this, note that we can rewrite the null hypothesis as

$$H_0 : \beta_1 = 0 \quad \text{and} \quad \beta_2 = 0.$$

Now, we identify the parameter vector as

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}.$$

There are actually several choices for $\mathbf{K}$, but we select the most straight-forward that corresponds to how we rewrote the null hypothesis:

$$\mathbf{K} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{with} \quad \mathbf{m} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

That is, the null hypothesis above can be stated as

$$H_0 : \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \beta = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Notice that each row of this hypothesis corresponds to one of the equalities expressed in the original statement of this hypothesis.

> **Note**
>
> When developing the matrix $\mathbf{K}$, the number of rows corresponds to the number of equal signs in the null hypothesis.

The general linear hypothesis allows us to say something about multiple parameters (or combinations of parameters) *simultaneously*. Each linear combination is not a separate hypothesis; together, they form a "joint" hypothesis. That is, we should think of each linear combination defined by the rows of $\mathbf{K}$ as "and" statements; we want every statement to be true at the same time. The framework is extremely flexible and can be used across several types of statistical models. It allows us to write the hypotheses compactly, mostly for communicating them to a computer. However, the framework alone does not produce p-values for such tests. In order to obtain a p-value, we need a model for the null distribution.

As the hypothesis involves many parameters, we cannot (and should not) test each statement separately. To fully understand that statement requires a background in statistical theory. We hand-wave this by saying that our parameter estimates are related. This is somewhat intuitive. Imagine trying to develop a line that runs through a cloud of points (see Figure 10.1); if we constrain the line to go through the "middle" of the data (the point represented by the average

of the predictor and the average of the response), then changing the slope of the line will necessarily change the intercept of the line. We extend this intuition by claiming that the estimate of one coefficient is related to the estimate of the other parameters in a model.

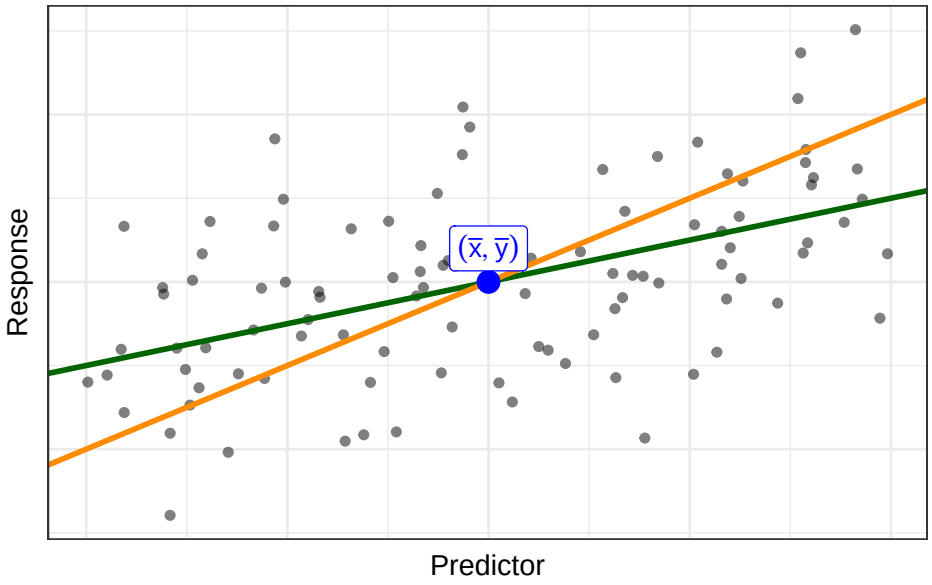


Figure 10.1: Illustration of the relationship between parameter estimates.

We know that the standard error is a measure of the variability in an estimate; we could just as easily use the variance (the square of the standard error). A convenient measure of the relationship between the estimates is known as the covariance. We then store all the information about how the parameter estimates vary and co-vary in the variance-covariance matrix.

Definition 10.2 (Variance-Covariance Matrix)**.** Let β represent the $(p+1)$-length vector of the parameters and $\hat{\beta}$ represent the $(p+1)$ vector of the parameter *estimates*. The variance-covariance matrix of the parameter estimates is the $(p+1)$-by-$(p+1)$ matrix Σ where

- the j-th diagonal element contains $Var\left(\hat{\beta}_j\right)$, and
- the (i,j)-th element contains the covariance between $\hat{\beta}_i$ and $\hat{\beta}_j$.

The variance-covariance matrix is an extremely important concept in statistical theory; here, we simply note that it contains information on the structure of how the estimates are related to one another. We further note that this is computed automatically in most software. We are now in a place to discuss inference for the general linear hypothesis.

Definition 10.3 (Model for the Null Distribution with the General Linear Hypothesis)**.** Let $\widehat{\beta}$ be the $(p+1)$ vector of estimates for the parameter vector β, and let the estimates have variance-covariance matrix Σ. Assuming the null hypothesis

$$H_0 : \mathbf{K}\beta = \mathbf{m}$$

is true, under the conditions of the classical regression model (Definition 4.3)

$$(1/r)\left(\mathbf{K}\hat{\beta} - \mathbf{m}\right)^{\top}\left(\mathbf{K}\widehat{\Sigma}\mathbf{K}^{\top}\right)^{-1}\left(\mathbf{K}\hat{\beta} - \mathbf{m}\right) \sim F_{r,n-p-1}.$$

As with previous results about the sampling and/or null distribution, the specifics of the above result are not as important as understanding there exists a standardized statistic for which the null distribution can be modeled explicitly, as an F-distribution with r numerator degrees of freedom and $n-p-1$ denominator degrees of freedom, and that this model depends on specific conditions. While the denominator degrees of freedom are associated with the scaling term for the residual variance estimate, the numerator degrees of freedom are associated with the complexity of the hypothesis (the number of rows of $\mathbf{K}$).

While the theory that provides the above results holds only for the linear model under the classical regression model, the approach we have outlined here will allow us to provide general results which are applicable under many types of regression models.

11 Large Sample Theory

The classical regression model (Definition 4.3) imposes several conditions on the distribution of the error term. These conditions are needed to depend on the model for the sampling distribution of the parameter estimates we developed in that section (Definition 4.8). However, these conditions are not always reasonable. Fortunately, many of the conditions can be relaxed. In this section, we consider relaxing the "Normality" condition. Changing the conditions we impose impacts how we model the sampling distribution of our estimates (and therefore impacts confidence intervals and p-values).

11.1 Two Types of Models

Models for the data generating process are broadly characterized into one of three groups: parametric, semiparametric, and nonparametric.

Definition 11.1 (Parametric Model)**.** A parametric model characterizes the distribution of the response using a finite set of parameters; for our purposes, this generally means the model *fully* characterize the distribution of the response given the predictors.

Definition 11.2 (Nonparametric Model)**.** A nonparametric model is unable to characterize the response using a finite set of parameters; for our purposes, this generally means the model makes no assumptions about the structure of the underlying distribution of the response given the predictors. Only minimal assumptions (such as independence between observations) are imposed.

Definition 11.3 (Semiparametric Model)**.** A semiparametric model specifies some components of the underlying distribution of the response using a finite set of parameters but does not fully characterize the distribution. This generally means that we may specify the mean and/or variance of the response given the predictors, but we do not characterize the distributional family of the response.

> **Note**
>
> Technically, semiparametric models are a subset nonparametric models. However, semiparametric models are often considered a distinct type of model because they have el-

ements of both parametric models (there are some parameters to be estimated) and nonparametric models (completely data-driven).

Parametric models make strong assumptions, but often make the analysis straight-forward as we are able to make use of a large class of results from statistical theory. This is very useful when we have a small sample size in particular. Nonparametric models are extremely flexible, but they require substantially large sample sizes. Semiparametric models, in turn, often find a the "sweet spot." They require larger samples than a parametric model, but not as large as a nonparametric model. Further, the scientific question of interest is still represented through statements about parameters in the model, linking the interpretations more directly with practical application and making the models more interpretable.

The alternate characterization of the classical regression model (Definition 4.5) reveals that the classical regression model is a parametric model. It completely characterizes the distribution of the response:

$$(\text{Response})_i \mid (\text{Predictors 1 through } p)_i \overset{\text{Ind}}{\sim} N\left(\beta_0 + \sum_{j=1}^{p} \beta_j (\text{Predictor } j)_i, \sigma^2\right).$$

Consider the aspects being *structure* specified here:

1. The mean response is given as a linear combination of the p predictors.
2. The variance of the response is constant for all combinations of the predictors.
3. Given the predictors, the response follows a Normal distribution.

Each of the aspects of this structure can be very useful when working with statistical theory; however, the questions often posed by researchers do not concern the form of the distribution of the response but only some aspect of the distribution. For example, the hypotheses we have considered thus far in the text surround the parameters of the mean model; that is, we have concerned ourselves only with questions regarding the *mean* response. That is, whether we model the distribution of the response using a Normal distribution or some other is not of interest; scientifically, we are interested in the mean response. If we are willing to relax the distributional form of the response, we are led to positing a semiparametric model.

Definition 11.4 (Semiparametric Linear Model)**.** Suppose we no longer require that the error terms follow a Normal distribution; however, we do continue to impose the remaining conditions of the classical regression model. Then, our model could be written as

$$\begin{aligned} E\left[(\text{Response})_i \mid (\text{Predictors 1 through } p)_i\right] &= \beta_0 + \sum_{j=1}^{p} \beta_j (\text{Predictor } j)_i \\ Var\left[(\text{Response})_i \mid (\text{Predictors 1 through } p)_i\right] &= \sigma^2 \end{aligned}$$

where the responses are independent of one another given the predictors.

Notice that this version of the model only specifies some aspects of the response distribution; it specifies that the mean is a linear combination of the predictors, and it specifies that the variance is constant. However, it does not specify the functional form of the distribution.

11.2 Large Sample Results

The primary benefit of a parametric model is that often the distributional assumption trickles through the analysis and allows us to exactly specify the model for the sampling distribution of the parameter estimates (Definition 4.8, for example). When we move to a semiparametric model, we need additional tools to allow us to model the sampling distribution of our estimates. Large sample theory is one such tool.

Definition 11.5 (Large Sample Theory)**.** The phrase "large sample theory" (or "asymptotics") is used to describe a scenario when the model for the sampling distribution (or null distribution) of an estimate (or standardized statistic) can be approximated as the sample size becomes infinitely large. That is, as the sample size approaches infinity, the sampling distribution (or null distribution) can be easily modeled using a known probability distribution.

Perhaps the most well-known example of large sample theory is the Central Limit Theorem encountered in introductory statistics.

Definition 11.6 (Central Limit Theorem)**.** Let $Y_1, Y_2, \ldots, Y_n$ be independent and identically distributed random variables with finite mean μ and variance σ^2. Then, as n approaches infinity, the distribution of the ratio

$$\frac{\sqrt{n}\left(\bar{Y} - \mu\right)}{\sigma}$$

approaches that of a Standard Normal random variable.

Rephrasing in the language of this text, Definition 11.6 states that as the sample size gets large, the standardized distance between the average response observed and the true average response can be modeled using a Normal distribution with mean 0 and variance 1. Notice that the theorem does not specify the distribution of the response Y; it only specifies the mean and variance. That is, we began with a semiparametric model for the data generating process (the distribution of the response is only partially specified) and yet obtained a model for the sampling distribution or our statistic of interest. We exchanged the condition that the response follow a Normal distribution (a more classical approach) for the condition that

"the sample size be sufficiently large" (the large sample theory approach); under thess revised conditions, we have a model for the sampling distribution of our parameter estimate.

It turns out that similar results can be derived for the semiparametric linear model. That is, in large samples, we can approximate the sampling distribution of our estimates and the null distribution of our standardized statistics.

Definition 11.7 (Large Sample Model for the Sampling Distribution of the Least Squares Estimates)**.** Suppose the classical regression conditions hold, with the exception of the errors following a Normal distribution. As the sample size gets large, we have that the distribution of the ratio

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{Var\left(\hat{\beta}_j\right)}} \sim N(0,1)$$

for all $j = 0, 1, ..., p$. Further, under the null hypothesis

$$H_0 : \mathbf{K}\beta = \mathbf{m}$$

we have

$$\left(\mathbf{K}\hat{\beta} - \mathbf{m}\right)^\top \left(\mathbf{K}\widehat{\Sigma}\mathbf{K}^\top\right)^{-1} \left(\mathbf{K}\hat{\beta} - \mathbf{m}\right) \sim \chi^2_r.$$

Notice that our statistics and standardized statistic have a similar form as before; the difference is the probability model being used. In place of a t-distribution, we have a Normal distribution. In place of the F-distribution, we have a Chi-Square distribution. These large sample results allow us to perform inference even if we are unwilling to assume the errors follow a Normal distribution.

It is natural to ask how large of a sample size is required for these models to be reasonable; there is no simple answer. Empirical studies suggest that in practice, if we have at least 30 degrees of freedom for estimating the error term, these results are often reasonable. However, empirical studies also demonstrate that it does depend on the tail behavior of the underlying population distribution; there are some populations that begin to mimic these results at samples of size 10 and others that require samples of size 10000. In practice, this is an *assumption* the analyst must determine whether to adopt.

Not all software implements methods for relying on these large sample results. However, as the sample size gets large, it turns out that classical inference and the large sample results coincide. That is, the confidence intervals and p-values we would compute using the large sample models and those obtained assuming the classical regression model are nearly identical. Therefore, in practice, when the sample size is large, we can rely on the default output even if we are unwilling to assume the errors follow a Normal distribution.

11.3 Residual Bootstrap

An alternative to large sample theory is building an empirical model for the sampling distribution (or null distribution) when working with a semiparametric model; one process for this is known as bootstrapping.

Definition 11.8 (Bootstrapping)**.** A process of constructing a sampling distribution of the parameter estimates through resampling. The observed data is resampled repeatedly, and the parameters of interest are estimated in each resample. The distribution of these estimates across the resamples is then used as an empirical model of the corresponding sampling distributions.

There are several bootstrapping algorithms; the most foundational for regression modeling is the residual bootstrap.

Definition 11.9 (Residual Bootstrap)**.** Suppose we observe a sample of size n and use the data to compute the least squares estimates $\widehat{\beta}$ for the parameters in the model

$$(\text{Response})_i = \beta_0 + \sum_{j=1}^{p} \beta_j (\text{Predictor } j)_i + \varepsilon_i.$$

The residual bootstrap proceeds according to the following algorithm:

1. Compute the residuals

$$(\text{Residuals})_i = (\text{Response})_i - (\text{Predicted Response})_i$$

2. Take a random sample of size n (with replacement) of the residuals; call these values $e_1^*, \ldots, e_n^*$.
3. Form "new" responses $y_1^*, \ldots, y_n^*$ according to

$$y_i^* = \widehat{\beta}_0 + \sum_{j=1}^{p} \widehat{\beta}_j (\text{Predictor } j)_i + e_i^*.$$

4. Obtain the least squares estimates $\widehat{\alpha}$ by finding the values of α that minimize

$$\sum_{i=1}^{n} \left(y_i^* - \alpha_0 - \sum_{j=1}^{p} \alpha_j (\text{Predictor } j)_i \right)^2.$$

5. Repeat steps 2-4 m times.

We often take m to be large (at least 1000). After each pass through the algorithm, we retain the least squares estimates $\hat{\alpha}$ from the resample. The distribution of these estimates across the resamples is a good empirical model for the sampling distribution of the original least squares estimates.

While the residual bootstrap is the foundation of many similar algorithms, it is perhaps not as easy to understand as the case-resampling bootstrap.

Definition 11.10 (Case Resampling Bootstrap)**.** Suppose we observe a sample of size n and use the data to compute the least squares estimates $\hat{\beta}$ for the parameters in the model

$$(\text{Response})_i = \beta_0 + \sum_{j=1}^{p} \beta_j (\text{Predictor } j)_i + \varepsilon_i.$$

The case resampling bootstrap proceeds according to the following algorithm:

1. Take a random sample of size n (with replacement) of the raw data (keeping all variables from the same observation together); denote the i-th selected response and predictors $(\text{Response})_i^*$ and $(\text{Predictor } j)_i^*$, respectively.
2. Obtain the least squares estimates $\hat{\alpha}$ by finding the values of α that minimize

$$\sum_{i=1}^{n} \left((\text{Response})_i^* - \alpha_0 - \sum_{j=1}^{p} \alpha_j (\text{Predictor } j)_i^* \right)^2.$$

3. Repeat steps 1-2 m times.

We often take m to be large (at least 1000). After each pass through the algorithm, we retain the least squares estimates $\hat{\alpha}$ from the resample. The distribution of these estimates across the resamples is a good empirical model for the sampling distribution of the original least squares estimates.

The case-resampling bootstrap procedure is easier to visualize as we are resampling the data observed. The resampling in the residual bootstrap is a bit more indirect as the residuals are what is resampled; this mimics generating new observations by "jittering" points away from the estimated regression line. In both algorithms, the same model is refit on each resample producing new estimates. The collection/distribution of these estimates across the m resamples is our model for the sampling distribution (which could be visualized using a histogram, for example).

The theoretical underpinnings of bootstrapping (and how it is implemented efficiently in software) is beyond the scope of this text. What we emphasize is that through this process, we

construct a model for the sampling distribution of the estimates, and that allows us to compute confidence intervals. Further, the residual bootstrap requires the same conditions as the classical regression model, with the exception of requiring the errors to follow a Normal distribution. That is, it has the same conditions as we stated for our semiparametric regression model above.

> **Note**
>
> Technically, the case-resampling bootstrap and the residual bootstrap require different conditions, with the case-resampling bootstrap being less restrictive. However, at this point, we do not make a distinction between which bootstrap algorithm is utilized. We will discuss the benefits of case-resampling later in the text.

Bootstrapping is more computationally burdensome than large sample theory, but it does not rely on the sample size being "large enough."

> **Note**
>
> While theoretically bootstrapping can be used with any sample size, it has been shown to yield more reliable results in large samples.

> **Note**
>
> While we have focused on the use of bootstrapping for computing a confidence interval, the same procedures can be adapted to compute p-values for hypothesis tests as well.

11.4 Choosing a Path

We have discussed two alternatives to using inference results from the classical regression model when we are unwilling to assume the errors follow a Normal distribution. Further, we have discussed ways to determine if the data is consistent with the assumption that the errors have a Normal distribution (Chapter 5). As we have seen throughout this unit, while these approaches were discussed in the context of the linear model, they illustrate a concept that holds across many types of regression models — there are essentially three ways to build a model for the sampling distribution of our parameter estimates.

> **Big Idea**
>
> There are three options for modeling the sampling (null) distribution of a parameter estimate (standardized statistic):

1. Exact Probability Theory: often the result of assuming a parametric model, the sampling distribution of the resulting parameter estimates is derived explicitly using probability theory.
2. Large Sample Theory: often employed in semiparametric models, the sampling distribution of the resulting parameter estimates can be approximated as the sample size gets large.
3. Empirical: often employed in semiparametric models, the sampling distribution of the resulting parameter estimates is modeled through resampling.

We will see as we move throughout the text that we often move between these various approaches. However, which approach we take is governed by the conditions we are willing to impose on the data generating process.

We end with a common question: if we are able to model the sampling distribution without fewer conditions, why would we not always take that approach? The closer the conditions are to the true data generating process, the more powerful our analysis; that is, if the errors are truly Normally distributed, then imposing that condition will make it more likely for us to find a signal that really exists. So, we battle the tension of a more powerful analysis with one that is more flexible. We adhere to the belief that we should choose the approach that is most consistent with the available data.

12 Modeling Curvature (Splines)

In addition to conditions on the error term, the classical regression model (Definition 4.3) requires that the predictors enter the model linearly. It is often the case, however, that the relationship between the response and a predictor is not linear, even after accounting for other predictors. Ignoring this curvature, essentially leaving the deterministic portion of the model incorrectly specified, can result in incorrect conclusions. Fortunately, the linear model framework is flexible enough to model curvature. The apparent contradiction that a "linear" model can address curvature comes from a misunderstanding of what it means to be a "linear" model.

Definition 12.1 (Linear Model)**.** A model is said to be linear if it can be expressed as a linear combination of the *parameters*. That is, the linearity does not refer to the form of the predictors but the form of the parameters.

The beauty of this understanding of linearity is that it allows us to capture curvature, provided that we can represent that curvature through the addition of additional predictors. As a simple example, suppose we have a response that has a parabolic relationship with a predictor (Figure 12.1).

Such a relationship could be captured by the *linear* model

$$(\text{Response})_i = \beta_0 + \beta_1 (\text{Predictor})_i^2 + \varepsilon_i.$$

The mean response will clearly generate curvature, but the deterministic portion is linear in the parameters; let $\mathbf{x}_i$ represent the vector of all predictors for the i-th observation, including the intercept. In this case we have

$$\mathbf{x}_i = \begin{pmatrix} 1 \\ (\text{Predictor})_i \end{pmatrix}.$$

And, let β represent the parameter vector

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}.$$

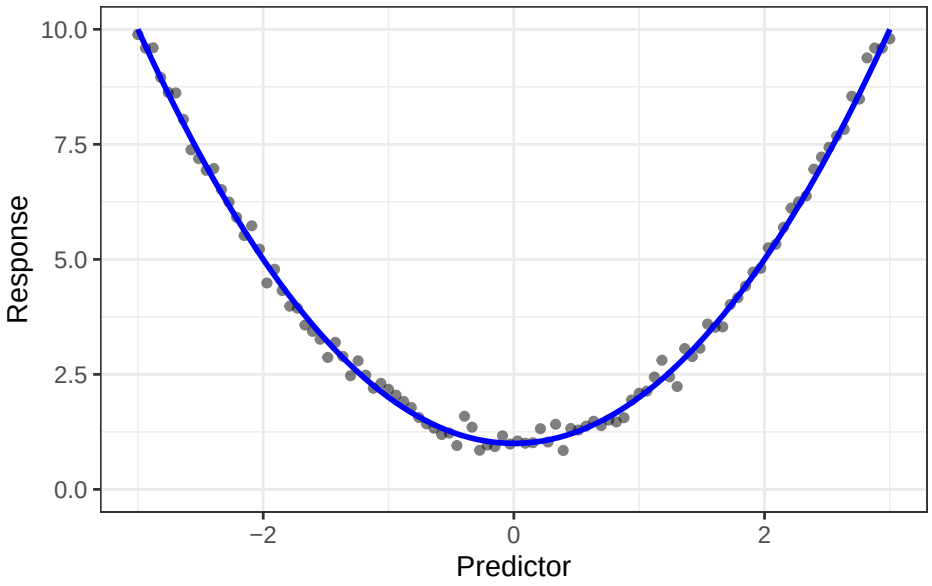


Figure 12.1: Illustration of a parabolic relationship between a response and a predictor.

Then, we can write the above model as

$$(\text{Response})_i = \mathbf{x}_i^\top \beta + \varepsilon_i$$

where we have that the deterministic portion is the product of two vectors; being able to express the model in this form satisfies the definition of a linear model.

> **Note**
>
> For those not as comfortable with matrix algebra, essentially, the model will be linear in the parameters as long as we can express it in the form
>
> $$(\text{Response})_i = \beta_0 + \sum_{j=1}^{p} \beta_j (\text{Something not involving parameters})_i + \varepsilon_i$$
>
> even if the "something not involving parameters" consists of nonlinear transformations of variables in our data set.

We are interested in investigating transformations of the predictors to add to the model that would capture curvature. In the above example, we knew the form of the curvature we wanted to model (a parabola shifted up from the origin). In practice, we often will not know the form of the curvature, just that it exists (from the plots of the residuals against the predictor). And, that curvature may not be modeled well with a high-degree polynomial (or would require a

polynomial of such a high degree it would not be practical). In such cases, splines are very useful.

Definition 12.2 (Spline)**.** A spline is a continuous piecewise polynomial used to model curvature. The points that define the piecewise components are called *knot points*; the functional form is allowed to change at the knot points.

Definition 12.3 (Linear Spline)**.** A linear spline is a continuous piecewise linear function.

A linear spline is perfect for capturing relationships which appear to be linear over regions, but for which the relationship is different in each of those regions. For example, a "V" relationship would suggest that as the predictor increases, the response tends to decrease up to a point (the bottom of the V); past that point (which is the "knot point" here), as the predictor increases, the response tends to increase as well. A linear spline can be placed into the linear model framework.

i Formula for Linear Spline

A response can be related to a predictor using a linear spline with k knot points, call them $t_1, t_2, \ldots, t_k$ using the following formula:

$$(\text{Response})_i = \beta_0 + \beta_1(\text{Predictor})_i + \sum_{j=1}^{k} \beta_{j+1} \left((\text{Predictor})_i - t_j\right)_+ + \varepsilon_i \tag{12.1}$$

where u_+ takes the value u when $u > 0$ and takes the value 0 otherwise. Capturing curvature using a linear spline with k knot points requires k additional terms.

Figure 12.2 illustrates a linear spline with two knot points.

i Note

In practice, the knot points for a linear spline are generally determined by the discipline expert based on some scientific reason why the relationship might change at that particular value of the predictor.

While the above definition of the linear spline considers only a single predictor, we can add a spline to a model which has additional predictors. That is, we could consider a model like

$$(\text{Response})_i = \beta_0 + \beta_1(\text{Predictor 1})_i + \beta_2 \left((\text{Predictor 1})_i - t_1\right)_+ + \beta_3(\text{Predictor 2})_i + \varepsilon_i.$$

Here, we have placed a linear spline with a single knot point on the first predictor, but the second predictor enters the model linearly.

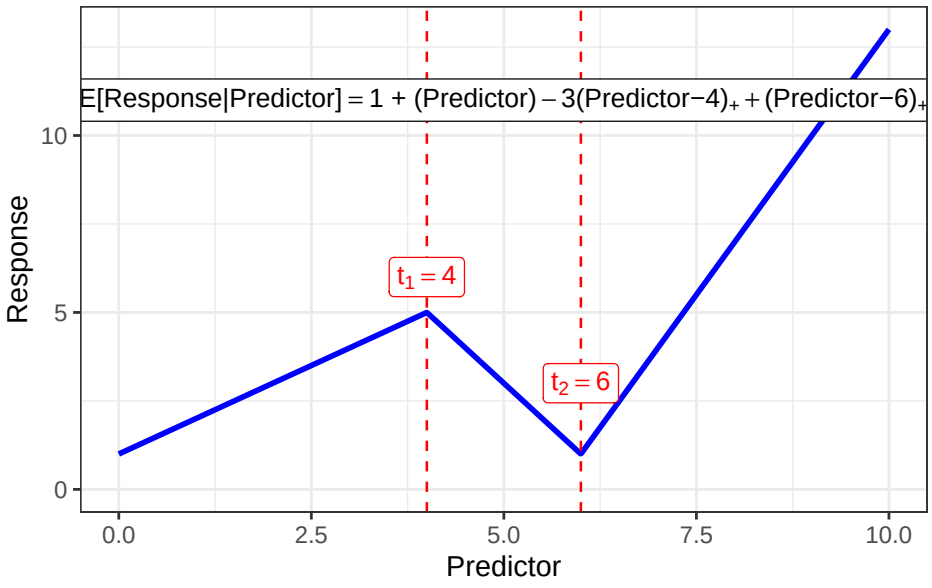


Figure 12.2: Illustration of a linear spline with two know points.

Once we have the additional elements in the model to capture the curvature, we can actually perform a hypothesis test to determine if the additional complexity is needed. Consider Equation 12.1, the hypothesis

$$H_0 : \beta_2 = \beta_3 = ... = \beta_{k+1} = 0$$

imposes the constraint that each of the spline components be removed from the model. That is, under this hypothesis, a linear relationship is sufficient for modeling the relationship between the response and the predictor.

There are plenty of forms of curvature which would not be captured by a linear spline. When we have a more complex relationship that needs to be modeled, we use a restricted cubic spline.

Definition 12.4 (Restricted Cubic Spline)**.** A restricted cubic spline is a continuous function comprised of piecewise cubic polynomials for which the tails of the spline have been restricted to be linear.

Restricted cubic splines (related to natural splines in the computational science community) are smooth at the knot points (meaning they have nice mathematical properties). Further, it has been shown empirically that restricted cubic splines are often flexible enough to approximate nearly any nonlinear relationship. As flexible as they are, what is really amazing is that we can embed restricted cubic splines into the linear model framework.

Formula for Resctricted Cubic Spline

A response can be related to a predictor using a restricted cubic spline with k knot points, call them $t_1, t_2, \ldots, t_k$ using the following formula:

$$(\text{Response})_i = \beta_0 + \beta_1(\text{Predictor})_i + \sum_{j=1}^{k-2} \beta_{j+1} x_{j,i} + \varepsilon_i$$

where

$$\begin{aligned} x_{j,i} = &\left((\text{Predictor})_i - t_j\right)_+^3 - \frac{\left((\text{Predictor})_i - t_{k-1}\right)_+^3 (t_k - t_j)}{t_k - t_{k-1}} \\ &+ \frac{\left((\text{Predictor})_i - t_k\right)_+^3 (t_{k-1} - t_j)}{t_k - t_{k-1}} \end{aligned} \tag{12.2}$$

and u_+ is a function taking the value u when $u > 0$ and the value 0 otherwise. Capturing curvature using a restricted cubic spline with k knot points requires $k - 2$ additional terms.

Empirical studies have shown that generally only $k = 5$ knot points are needed, and these are set at the 5-th, 27.5-th, 50-th, 72.5-th, and 95-th percentiles. This ensures there is enough data in each region to appropriately capture the curvature.

Similar to linear splines, we can add a restricted cubic spline for a variable to a model which has additional predictors. For example,

$$\begin{aligned} (\text{Response})_i = \beta_0 &+ \beta_1(\text{Predictor 1})_i + \beta_2 \left((\text{Predictor 1})_i - t_1\right)_+^3 \\ &- \beta_2 \frac{\left((\text{Predictor 1})_i - t_2\right)_+^3 (t_3 - t_1)}{t_3 - t_2} + \beta_2 \frac{\left((\text{Predictor 1})_i - t_3\right)_+^3 (t_2 - t_1)}{t_3 - t_2} \\ &+ \beta_3(\text{Predictor 2})_i + \varepsilon_i \end{aligned}$$

captures curvature in the first predictor using a restricted cubic spline with three knot points while allowing the second predictor to enter the model linearly. We note that while it appears this model is much more complex, only a single additional term is needed to capture the curvature on the first predictor.

Once we have the additional elements in the model to capture the curvature with the spline, we can actually perform a hypothesis test to determine if the additional complexity is needed. Consider the model in Equation 12.2, the hypothesis

$$H_0 : \beta_2 = \beta_3 = \ldots = \beta_{k-1} = 0$$

imposes the constraint that each of the spline components be removed from the model. That is, under this hypothesis, a linear relationship is sufficient for modeling the relationship between the response and the predictor.

Figure 12.3 illustrates a restricted cubic spline with five knot points. We point out that there is quite a bit of curvature here, and yet this is captured by a linear model!

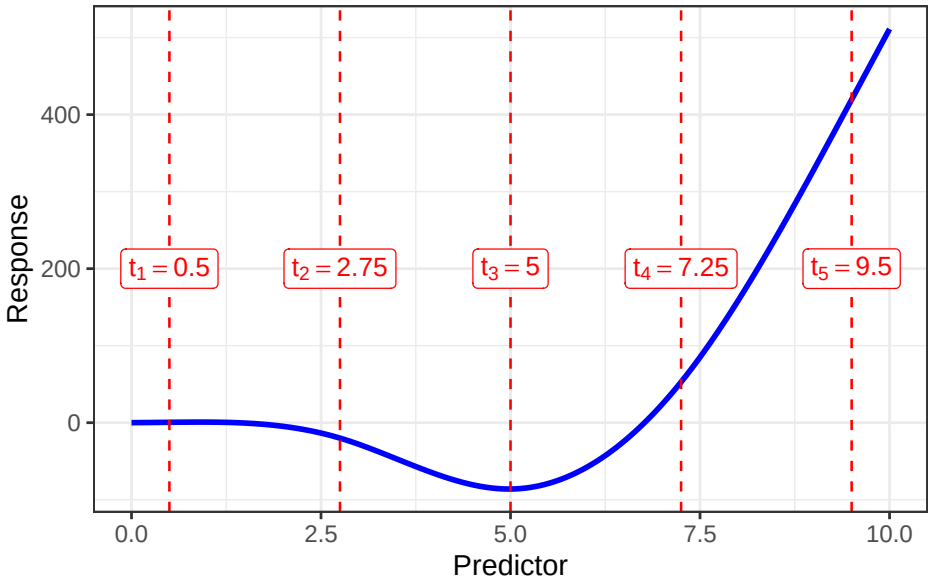


Figure 12.3: Illustration of a restricted cubic spline with five knot points.

Both types of splines can be fit using standard software if we are willing to program the above formulas; however, statistical software often has a direct implementation.

There are nonparametric approaches to capturing curvature as well ("loess" curves are popular choices). We prefer splines to nonparametric approaches as splines require less computation and can be easily implemented in any software. Further, splines can be placed in a semi-parametric model allowing us to capture complex curvature without extremely large sample sizes. Finally, since splines can be implemented within a linear model, we can easily test to determine if the additional complexity is needed.

Linear splines allow for very easy interpretation since the relationships are linear in each region. Restricted cubic splines, in contrast, have intractable interpretations but provide a lot of flexibility. In general, if the predictor that requires a spline is the primary variable of interest, we recommend trying a linear spline. If however, the predictor is just being adjusted for in the model, and the curvature is not of primary importance, using a restricted cubic spline with five knot points is typically sufficient.

> **Note**
>
> When a categorical predictor is modeled using a series of indicator variables, linearity cannot be violated for these components. This is known as a "saturated" model because no simplifying assumptions about the structure have been enforced. As a result, violations of linearity and their subsequent adjustments (like splines) are only considered for quantitative predictors.

As in the previous chapters of this unit, we introduced splines in the context of the linear model. However, they can be incorporated into a vast number of regression models.

Part IV

Unit IV: Models for Repeated Measures

The conditions we impose on the data generating process impact the model of the sampling distribution of our parameter estimates, and the model for the sampling distribution is necessary for performing inference. At times, the conditions we are willing to impose are driven by the data collection procedure. In this unit, we consider data that result from study designs that lead to collections of observations being associated with one another (beyond what can be explained by the predictors in the model). This correlation between observations must be appropriately modeled if we want to model the sampling distribution of our parameter estimates.

Specifically, this unit considers the topic of repeatedly measuring the response. This could be the result of a study design that requires the response be measured routinely on the same participants (a "pre-post" study, for example); or, it could be the result of a study design that first enrolls entire families and then records the response for each member of the family. In the first example, it is reasonable to believe the responses recorded on the same participant are associated with one another in some way; in the second example, it is reasonable to believe responses recorded on members of the same family are associated with one another in some way. Understanding how to incorporate this relationship in the model allows us to conduct such studies, which often have more power to detect effects of interest.

13 The Language of Repeated Measures

13.1 Importance of Study Design

Study design is too often separated from the statistical analysis that follows. However, in addition to informing the conclusions we draw regarding the data, the study design helps in choosing an appropriate analysis to address the question of interest. As an example, consider the following study reported in Vittinghoff et al. (2012).

Example 13.1 (Digestive Enzymes)**.** The ability of the bowels to properly absorb nutrients can be impacted by a lack of digestive enzymes. This presents as excess fat in the feces, which is in turn treated with pancreatic enzyme supplements. A study was conducted comparing three forms of a particular enzyme supplement; these were compared to no supplement (placebo) as a control. Participants were given the supplement to take for a specified length of time; then, the amount of fecal fat (g/day) present was recorded. Interest is in determining if the amount of fecal fat produced, on average, differs for any of the treatments.

Suppose we are given the data in Table 13.1 to address the question posed in Example 13.1.

Consider the following generalized linear model to describe the data generating process:

$$(\text{Fat})_i = \beta_0 + \beta_1(\text{Tablet})_i + \beta_2(\text{Coated})_i + \beta_3(\text{Uncoated})_i + \varepsilon_i, \tag{13.1}$$

where

Table 13.1: Fecal fat (g/day) present in participants given a pancreatic enzyme supplement.

Placebo	44.5	33.0	19.1	9.4	71.3	51.2
Tablet	7.3	21.0	5.0	4.6	23.3	38.0
Coated Capsule	12.4	25.6	22.0	5.8	68.2	52.6
Uncoated Capsule	3.4	23.1	11.8	4.6	25.6	36.0

$$(\text{Tablet})_i = \begin{cases} 1 & \text{if i-th observation corresponds to supplement in tablet form} \\ 0 & \text{otherwise} \end{cases}$$

$$(\text{Coated})_i = \begin{cases} 1 & \text{if i-th observation corresponds to supplement in coated form} \\ 0 & \text{otherwise} \end{cases}$$

$$(\text{Uncoated})_i = \begin{cases} 1 & \text{if i-th observation corresponds to supplement in uncoated form} \\ 0 & \text{otherwise} \end{cases}$$

are indicator variables capturing the impact of the categorical treatment group. Our question of interest is captured by the testing

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \qquad \text{vs.} \qquad H_1 : \text{At least one } \beta_j \text{ differs.}$$

This analysis demonstrates no evidence (p = 0.168) the average amount of fecal fat differs for any of the supplement forms (see Figure 13.1).

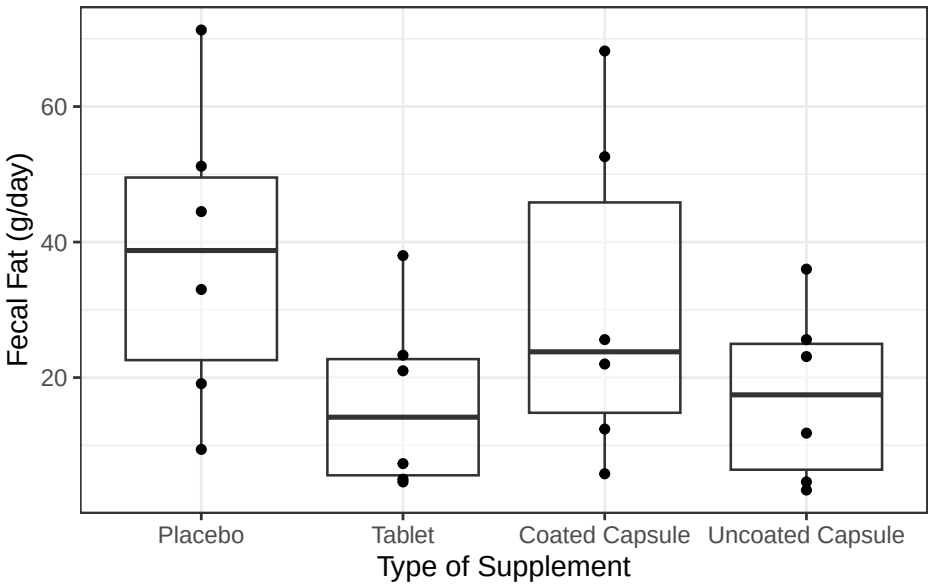


Figure 13.1: Fecal fat (g/day) present in participants given a pancreatic enzyme supplement when we assume 24 independent participants.

Of course, the above analysis is predicated on the data being consistent with the conditions for the classical regression model. For example, it seems reasonable to assume that the fecal fat present in one subject is independent of the fecal fat present in any other subject once we have accounted for the type of supplement. Therefore, it seems reasonable that any two fecal

Table 13.2: Fecal fat (g/day) of six participants enrolled in a cross-over study examining four types of enzyme supplement.

	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5	Subject 6
Placebo	44.5	33.0	19.1	9.4	71.3	51.2
Tablet	7.3	21.0	5.0	4.6	23.3	38.0
Coated Capsule	12.4	25.6	22.0	5.8	68.2	52.6
Uncoated Capsule	3.4	23.1	11.8	4.6	25.6	36.0

fat measurements above are independent once we have accounted for the type of supplement received. This, however, follows from how we assumed the data was collected — that each measurement is from a different subject.

Let's reconsider the above example but add a little more context to the study design.

Example 13.2 (Example 13.1 Continued (Expanded Context))**.** The study described in Example 13.1 was actually conducted as a cross-over study enrolling six participants. Each participant was given one form of the enzyme (determined randomly) and followed for a specified period of time at which point the amount of fecal fat (g/day) present was obtained. Following a substantial wash-out period the participant was assigned a different form of the supplement. This continued until each participant had been assigned to all four supplement forms.

Interest is in determining if the amount of fecal fat produced, on average, differed for any of the treatments. However, it is known that the amount of fecal fat present can vary substantially from one individual to another due to dietary preferences; researchers would like to account for the variation in the fecal fat across subjects when performing the analysis.

Table 13.2 provides the same data as Table 13.1 but with the additional context given in Example 13.2 — specifically, that six participants underwent each of the four treatments.

The study design results in several measurements being taken on each subject; further, the researchers believe that amount of fecal fat present can vary substantially from one individual to another. This additional information suggests that observations from the same individual are associated in some way; it is therefore unreasonable to assume the errors in Equation 13.1 are independent of one another.

> **Note**
>
> For those familiar, you may recognize the cross-over design described in Example 13.2 as a "randomized complete block design." Others may be familiar with the concept of "paired data," of which Example 13.2 is a generalization. While the idea is similar, the methods discussed in this text are a more inclusive approach.

Why would recording multiple observations on the same subject impact the condition of independence? A cursory look at the data confirms the researchers' beliefs: the values of fecal fat vary greatly from one participant to another, ranging from 9 to 71 g/day in some cases. However, the values of fecal fat recorded for a single participant do not tend to vary to such a degree. That is, the variability *between* participants is substantially larger than the variability *within* a participant. As suggested by the researchers, this could be explained by differences in participant diets. It is difficult to detect differences between the supplement types since the differences between subjects is so much larger.

To better understand why researchers would design such a study, we review attributes of good study design. Generally speaking, there are three components to any well-designed study: replication, randomization, and reduction of extraneous noise.

> ⚠ Warning
>
> A study is not poor just because it lacks one of these elements. That is, a study can provide meaningful insights even if it does not make use of each of these elements; every study is unique and should be designed to address the research objective. These elements are simply helpful in creating study designs.

Definition 13.1 (Replication)**.** Replication results from taking measurements on different units (or subjects) for which you expect the results to be similar. That is, any variability across the units is due to natural variability within the population.

> ⚠ Warning
>
> The term "replication" is also used in the context of discussing whether the results of a study are replicable. While our use of the term is about replicating a measurement process within a study, this does not downplay the importance of replicating an entire study.

Replication allows us to estimate subject-to-subject variability. Our intuition is that more data is better; in fact, increasing the sample size (the number of *unique* subjects on which we collect data) will result in less variability in our estimates. That is, the sampling distribution for the parameter estimates will be narrower. Increased replication also leads to increased power — the ability to detect a signal when it really exists (as the null distribution will also be narrowed).

Definition 13.2 (Randomization)**.** Randomization can refer to random *selection* or random *allocation*.

Random selection refers to the use of a random mechanism to select units from the population. Random selection minimizes bias.

Random allocation refers to the use of a random mechanism when assigning units to a specific treatment group in a controlled experiment. Random allocation eliminates confounding and permits causal interpretations.

There are many forms of random sampling. Some sampling schemes ensure each collection of subjects is equally likely (e.g., simple random sample), while others over-sample from under-represented subpopulations (e.g., stratified random sample). Each scheme shares the goal of eliminating bias, making the data more representative of the target population. While random sampling is the ideal, it is not always feasible. In clinical trials, for example, patients must elect to participate, thereby making the sample not random. When a random sample is not possible, summarizing the data to ensure it is representative of the target population is critical.

There are many forms of random allocation. Some randomization schemes ensure each treatment group is equally likely, while others assign participants to the active treatment with a higher probability than to the placebo. Other randomization schemes, like that mentioned in Example 13.2 randomizes the *order* of treatments. Each scheme shares the goal of eliminating confounding, allowing for causal interpretations. Whenever possible, random allocation is utilized, but it is not always feasible. Perhaps most famously, it would be unethical to conduct a randomized controlled trial to investigate the link between smoking and cancer. Therefore, observational studies were utilized to establish this link.

Definition 13.3 (Reduction of Noise)**.** Reducing extraneous sources of variability can be accomplished by fixing extraneous variables or through blocking. These actions reduce the number of differences between the units under study.

> **i** Tension between Lab Settings and Reality
>
> Scientists and engineers are trained to control unwanted sources of variability (or sources of error in the data generating process). This creates a tension between what is observed in the study (under "lab" settings) and what is observed in practice (under "real-world" settings). This tension always exists, and the proper balance depends on the goals of the researchers.

Intuitively, the less variation in the response, the easier it is to detect a signal. This leads naturally to saying that we could eliminate extraneous variability if the groups were *identical*; that results in using the same subjects in multiple groups, resulting in taking repeated measurements on the subjects — blocking on the participant.

Definition 13.4 (Blocking)**.** Blocking is a way of minimizing the variability contributed by an inherent characteristic that results in dependent observations. In some cases, the blocks are the unit of observation which is sampled from a larger population, and multiple observations are taken on each unit. In other cases, the blocks are formed by grouping the units of observations

according to an inherent characteristic; in these cases that shared characteristic can be thought of having a value that was sampled from a larger population.

In both cases, the observed blocks can be thought of as a random sample; within each block, we have multiple observations, and the observations from the same block are more similar than observations from different blocks.

Blocking is useful when we can identify the nuisance characteristic in advance of data collection. Blocking is a way of ensuring the treatment groups are similar because the same subjects (with respect to this particular characteristic) end up in each treatment group. We know that random allocation eliminates confounding because it ensures that, on average, treatment groups are similar. Blocking alone does not eliminate confounding; it must be combined with randomization. However, if we account for the blocking in the analysis, we are able to sharpen our estimates because not only has balance occurred, we are able to explain a portion of the variability within each treatment group.

13.2 Studies with Repeated Measures

Studies that have repeated measurements taken on subjects typically violate the condition of independence. While the above design concepts apply broadly, the following terminology is specific to such studies.

Definition 13.5 (Repeated Measures)**.** The phrase "repeated measures" refers to data for which the observed responses can be grouped based on some nuisance variable (typically the participant), and this grouping captures some inherent characteristic such that observations within a group tend to be more alike than observations across groups.

> **Note**
>
> The "paired data" setting (typically studied alongside the "paired t-test") is a special case of using blocking with two observations per block, resulting in repeated measures.

As discussed in Example 13.2, the large variability in fecal fat across participants relative to the variability in fecal fat within participants can mask the treatment effect if not accounted for appropriately. From a more theoretical perspective, this difference in the variability induces a correlation structure among the error in the responses.

Definition 13.6 (Correlation Structure)**.** The correlation structure quantifies the strength and direction of the relationship between the errors in the observed responses.

Big Idea

Ignoring the correlation structure does not tend to affect the parameter estimates, but it often affects the resulting standard errors, thereby impacting confidence intervals and p-values.

Unfortunately, there is no way to predict the impact of ignoring the correlation structure on the standard errors of our estimates. As a result, ignoring the correlation structure could result in confidence intervals that are too wide or too narrow (and p-values that are too large or too small). In short, ignoring the correlation structure in the data can result in inappropriate inference. In order to obtain appropriate inference, we need to account for the structure in our model!

Big Idea

When the data is correlated, it must be taken into account in all aspects of the analysis, from graphics to inference.

In order to illustrate the impact of the correlation structure on an analysis, let us revisit visualizing the data from the digestive enzymes study. Recall that Figure 13.1 visualized the data from Example 13.1 when we assumed the data came from 24 independent participants; Figure 13.2 visualizes the same data with the context from Example 13.2 included — namely that there were only six participants, each measured under multiple treatments. While there are some exceptions, notice how the lines do not "mix" often; that is, some participants tend to have less fecal fat than others, regardless of the form of the supplement given. The differences between Figure 13.1 and Figure 13.2 highlights that the differences in the overall average response for each supplement type is small relative to the variability across participants; as a result, focusing on the overall average response alone (Figure 13.1) make it difficult to detect differences between the supplement types. However, if we examine the differences in the supplement types *within each participant* (Figure 13.2), we can see that for nearly every participant, the fecal fat is reduced with the supplement (compared to placebo). It may not be immediately obvious if there is a difference among the form of the supplement, but it seems clear that having the supplement is better than not having it. Accounting for the variability in the response across participants changes our conclusions.

Notice that when examining Figure 13.2, we were not interested in comparing one participant to another; our focus was still on comparing supplement types. The additional grouping was simply to account for the relationship between the observations. Not all "grouping" variables are the same. We think about variables differently depending on their role in the model for the data generating process. Loosely, we can categorize factors as either fixed or random effects.

Definition 13.7 (Fixed Effect)**.** Fixed effects are terms in the model for which we are interested in both the specific grouping levels, and we are interested in characterizing the relation-

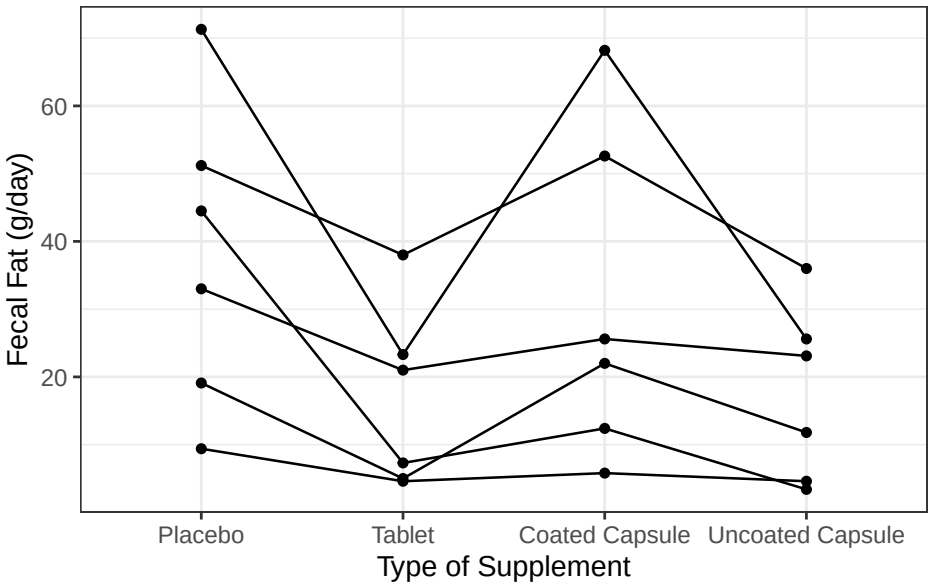


Figure 13.2: Fecal fat (g/day) of six participants enrolled in a cross-over study examining four types of enzyme supplement. Data from the same participant connected to illustrate the correlation structure.

ship between these levels and the response.

The treatment/factor of interest is nearly always a fixed effect. In addition, fixed effects can include anything we want to account for in the model in such a way that if we were to repeat the study, the same levels would be visible again. In Example 13.2, the form of the supplement is the fixed effect. If we were to repeat the study, we would still expect to use these same four levels (placebo, tablet, coated and uncoated capsule). And, we are interested in examining the impact of supplement type on the resulting fecal fat.

Definition 13.8 (Random Effect)**.** Random effects are terms in the model that capture the correlation induced due to an inherent characteristic that varies across the population. We are *not* interested in the specific grouping levels, and we either are not interested in the relationship with the response.

We are rarely interested in saying how random effects impact the response; so, the treatment/factor of interest is rarely a random effect. In Example 13.2, the participant is a random effect. If we were to repeat the study, it is unlikely we would use these same six individuals. Instead of quantifying the difference between participants, we wanted to think about this variable because groups of observations on the same participant are more alike than observations across participants.

> **Note**
>
> In order to distinguish fixed and random effects in a model, think about repeating the study; would you care if the levels of the factor were to change? Are you interested in comparing the first level to the second? If you answer "yes" to these questions, the factor is most likely a fixed effect.

The impacts of fixed and random effects in modeling are studied in Chapter 14. In brief, regression modeling is about partitioning variability. When we are able to identify additional sources of variability (like how the response varies across individuals in the population), we are able to improve our estimation of the effects of interest.

In Example 13.2, each observation is from a unique combination of the participant on which it was observed and the treatment assigned at that time (a cross-over study). This is not the only form of a study that can result in repeated measurements. Other common study designs include longitudinal studies, cross sectional studies with subsampling, and studies that utilize cluster samples.

Definition 13.9 (Cross-Over Study)**.** A cross-over study exposes each participant to multiple treatments. Whenever possible, the order of the treatments is randomly determined. This is equivalent to a randomized complete block design in which the blocks are the participants. When the treatments are believed to have a lingering effect, a wash-out period between treatments is used to minimize the impact of previous treatments on the treatment the participant is currently being exposed to.

Definition 13.10 (Randomized Complete Block Design)**.** A randomized complete block design is an example of a controlled experiment utilizing blocking. Each treatment is randomized to observations within blocks in such a way that every treatment is present within the block and the same number of observations are assigned to each treatment within each block.

Definition 13.11 (Longitudinal Study)**.** A longitudinal study repeatedly measures the response on each subject at various points in time.

All clinical trials follow subjects over time; a longitudinal study measures the response of interest multiple times over the course of the trial, resulting in repeated measures. In a longitudinal study, interest is often in modeling the overall trajectory across subjects instead of the trajectory for subjects individually. We are generally interested in modeling the trajectory of the response over the time interval. For example, we may be interested in the size of a tumor each month for the first year after being treated with radiation.

> **Note**
>
> There are many similarities between longitudinal studies and time-series data as each follows data over time. We do see some differences. Time-series data is often focused on business applications while longitudinal studies are more common in the biological sciences. Time-series data often models a single "stream" that is quite long. Longitudinal studies have several "streams" (one for each subject), but these tend to be a bit shorter as we do not have constant follow-up. In time-series data, it is often believed that the previous response is useful in predicting the next response; in longitudinal data, we do believe there is correlation among the errors in the model, but we do not generally use the value of the previous observation itself in making the next prediction but instead model with time as the predictor.

Definition 13.12 (Cross Sectional Study)**.** A cross sectional study considers data from a single snapshot in time.

Cross sectional studies are generally what we imagine when we first learn about data collection in an introductory course. We note that individual observations in a cross sectional study may not have been taken at exactly the same time, but it is believed that time does not have an impact. For example, we might record the size of a tumor one month after treatment with radiation. While the study might enroll participants over the course of two years, we record the measurement at one snapshot in time (one month after treatment) on each participant, and we believe the participants from each year of the study should be representative of the same group.

Longitudinal studies clearly involve repeated measures. Whether a cross sectional study has repeated measures depends on the study design. In the biological sciences, it is common to employ subsampling in cross sectional studies.

Definition 13.13 (Subsampling)**.** Subsampling occurs when several measurements are taken on each subject under the same treatment, possibly at unique locations.

In a study with subsampling, the unit of observation is the subject itself. As an example, a person may be assigned to a particular treatment to improve eyesight. The subject's eyesight is then measured in each eye; the person is the unit of measurement but we obtain two measurements of the response (eyesight), one corresponding to each eye. This is sometimes referred to as "pseudo-replication," and many researchers can mistakenly believe they have a larger sample size than exists in reality; the observations recorded on the same subject are related and should not be treated as independent observations. While it is common to average the subsamples, doing so results in a loss of information compared to modeling the correlation structure on the original data.

Definition 13.14 (Cluster Samples)**.** Stratified sampling divides a population into groups and samples from within each group; in contrast, cluster sampling divides the population into groups and randomly samples a few groups and takes measurements from within the group.

Observations from the same cluster are typically related. For example, if we are studying the nutrition levels of citizens within a particular state, we may sample specific counties first, and then sample participants from within the chosen counties. Citizens from the same county may be related since they have similar access to healthy food options.

Regardless of the study design, when we recognize clusters of observations which have some relationship beyond that explained by the fixed effects of interest, we need to model that correlation structure.

> **Note**
>
> When we use the phrase "repeated measures," we mean repeatedly measuring the same response. All regression models make use of multiple variables measured on the same subject. The models discussed in this unit refer to measuring the same response repeatedly.

When representing data from a repeated-measures study, it is useful to convey the correlation structure in the graphic as well (as we did in Figure 13.2). This is not always straight-forward. When the number of subjects in the study is not overwhelming, a spaghetti plot can be useful. These are particularly useful for studies that follow subjects over time and can be helpful in illustrating the trend over time.

Definition 13.15 (Spaghetti Plot)**.** A spaghetti plot is a scatterplot that displays the trends within a subject, highlighting the correlation structure by connecting points from the same subject.

Other times, using color or other aesthetics is needed to illustrate the correlation in the data.

14 Mixed Effects Models

Chapter 13 discussed how studies with repeated measures induce a correlation structure on the data. In this chapter and the next, we consider two approaches for modeling repeated measures data. In this chapter, we focus on a flexible modeling framework that builds up the data generating process in stages, recognizing the relationship between observations at each stage.

> **Note**
>
> While we discuss these methods in the context of a linear model, these methods can be extended to other modeling frameworks.

14.1 Partitioning Variability

In order to motivate the modeling approach developed in this chapter, we first discuss the various reasons the value of the response is not the same for all observations. We can view regression models as trying to explain why the response values differ across observations. The more reasons we can put in place, the more variability we are able to explain. By naming these sources of variability, we are able to construct a corresponding model by building it up in stages.

> **Big Idea**
>
> Statistical models partition the variability in the response.

While our discussion generalizes to many types of studies, it helps to imagine measuring the response of interest at several points across time for several subjects (Figure 14.1). For example, imagine tracking a child's weight as they age. We can imagine an overall trend — as a child ages, their weight increases.

Further, we may want to allow this trend to depend on a key factors (or other covariates). For example, we might posit that the child's weight is higher for those in one medical treatment group compared to another. Our research questions are generally at this stage — characterizing the impacts of fixed effects on the response (in this case, time and treatment groups).

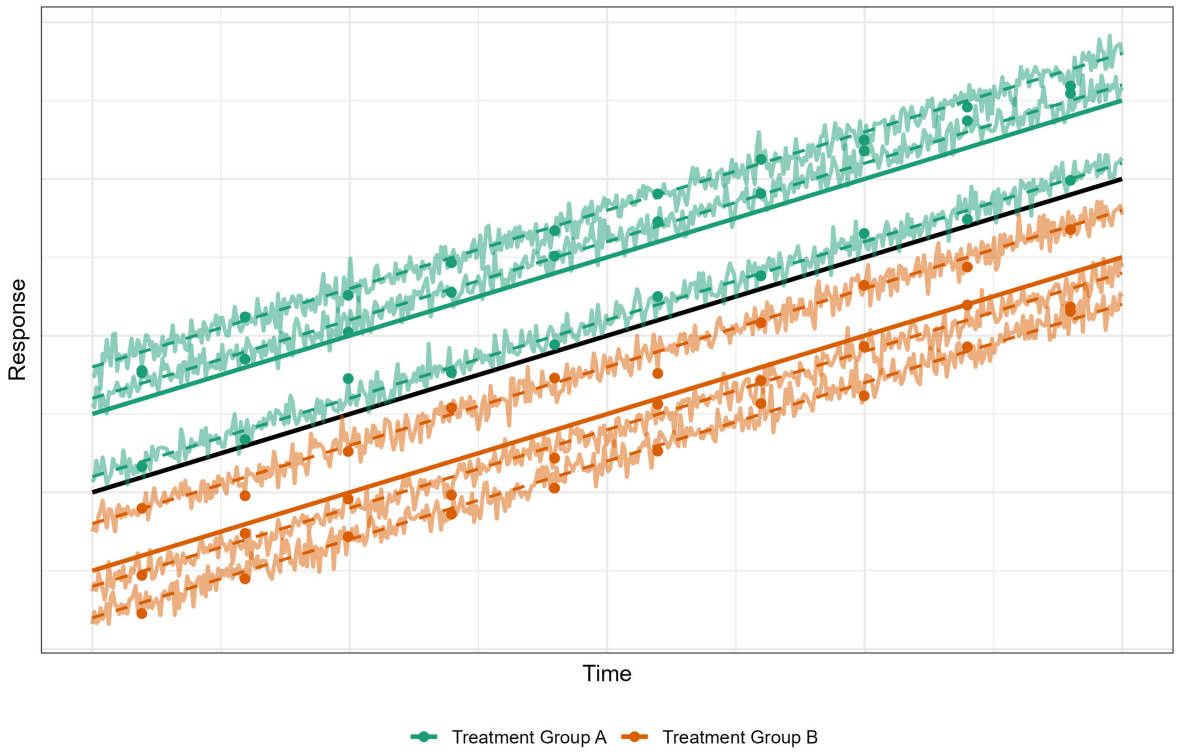


Figure 14.1: Illustration of various sources of variability in the data generating process.

Of course, the trajectory of any particular subject (across their repeated observations) will differ from the overall trajectory (and differ from one subject to another). Some children naturally have larger or smaller weights than others. At any point in time, a subject may have a trajectory that is above average; others will have a trajectory that is below average. This vertical shift or "bump" in the position of the trajectory captures the biological variation between subjects. All observations measured on the same subject will share a similar "bump," creating a relationship between observations. This between-subject variability is primarily what contributes to the correlation structure in the response.

For any subject, the actual response is likely to not lie directly on their individual trajectory. For example, a child's growth may not follow a smooth growth-curve even if we model it in that way. This is the result of natural biological fluctuations within the subject. As such, observations measured close together in time can tend to be more alike than those measured further apart in time. As an example, if a child's weight is slightly above their individual trajectory at this moment, it is likely to be above their trajectory an hour from now. However, if a child's weight is slightly above their individual trajectory at this moment, it does not really tell us anything about whether their weight will be slightly above or below their individual trajectory a year from now. Therefore, magnitudes of this within-subject source of variability are thought to be related when close in time and independent otherwise.

Finally, it is unlikely that the observed response is equal to the *actual* response as a result of measurement error. The weight of a child will be subject to the accuracy and precision of the scale, whether the subject was measured with or without clothing and shoes, etc. The magnitude of such errors are thought to be independent of one another.

What we are seeing in this discussion is that the "error" term we considered in the linear model of the previous units is actually the result of several sources of variability.

Big Idea

Observations from the same block tend to be high or low (relative to the average) together. These blocks could be due to repeated measures on the same subject or observations clustered together due to some other characteristic. While this "between-subject" variability induces a correlation structure among observations from the same block, it is common to think that observations from different blocks are independent.
Observations from within a block recorded close together in time are likely to be related, inducing a "within-subject" correlation. It is common to think that observations far apart in time are independent.

Taking into account all sources of variability can result in a very complex model (and this continues to be an area of active research). In practice, we can make simplifying assumptions about the data generating process that allows us to rely on a simpler construct. For example, we may assume the data are collected far enough apart in time so that the within-subject correlation is negligible.

14.2 Model Formulation

While our discussion above illustrates how the various sources of error/variability build on one another to create the observed data, we really worked backward. That is, we started with the overall trend and decomposed it to arrive at the data observed. When we model, we want to work in the other direction, building the model in stages. This is known as a hierarchical model.

Definition 14.1 (Hierarchical Model)**.** A hierarchical model breaks the data generating process into smaller stages and posits a model for each stage. The stages are determined by defining a hierarchy of units and thereby capturing the sources of variability.

While a hierarchical model could have an arbitrary number of stages, for a large number of applications, it suffices to consider the model being composed of only two stages: the individual-level (or within subject) and the population-level (or between subject).

> **Note**
>
> Remember, repeated measures can be the result of clusters of observations; so, the term "within-subject" should be interpreted as "within-block."

The individual-level stage posits a model for the observations within a subject (or block). That is, this model characterizes the relationship between the response and only those variables that change across observations on a single subject. Conceptually, this is the model we would construct if we only had data for a single subject. As a result, this model only includes "within-individual" predictors — those that vary within a subject. In biological settings, this is most common when measurements are taken across time; for example, following a child over several years, the weight and height of the child will change. However, it is unlikely that the highest level of education achieved by the child's parents is likely to change over this time frame. Therefore, when modeling the weight of the child, time and the child's height would be within-individual predictors, and the education level of the parents would be a "between-subject" predictor — those that change from one subject to another but remain constant for all observations from the same subject.

Definition 14.2 (Individual-Level Model)**.** The individual-level model characterizes the response for the i-th subject (or block) only.

Consider the study to investigate the impact of four methods of delivering a pancreatic enzyme supplement (Example 13.2). The individual-level model would characterize the fecal fat (response) observed within each subject. There are only a few variables in this data: the fecal fat, the supplement type, and the participant identification. The fecal fat is the response. The

supplement type is the factor of interest, and in this case, happens to be an individual-level variable as it changes *within* each participant.

Our individual-level model can make use of only the individual-level predictors; in this case, it can only depend on the supplement type. We would like to allow the fecal fat observed to depend on the type of supplement; we also acknowledge the potential for measurement error. This leads to a model of the form

$$(\text{Fecal Fat})_i = \alpha_{0,i} + \alpha_{1,i}(\text{Tablet})_{i,j} + \alpha_{2,i}(\text{Coated})_{i,j} + \alpha_{3,i}(\text{Uncoated})_{i,j} + \varepsilon_{i,j} \tag{14.1}$$

where here i is indexing the subject and j the observation within each subject. If we only had data on a single subject, we would compare the fecal fat for this subject under each of the supplement types; that is exactly what Equation 14.1 does (while allowing for measurement error). We must describe the distribution of random variables, such as the error term in Equation 14.1, that appear in the model. It is common to assume these errors are independent and follow a Normal distribution; specifically,

$$\varepsilon_{i,j} \overset{\text{IID}}{\sim} N\left(0, \sigma^2\right).$$

What is unique about this model formulation is at this point, the *parameters* themselves (the α terms in Equation 14.1) are allowed to vary across subjects. How those vary is determined by the population-level model.

Definition 14.3 (Population-Level Model)**.** The population-level model characterizes how the *parameters* of the individual-level model vary across subjects (or blocks) in the population.

While we do not actually proceed in this way, conceptually, the population-level model is constructed by (a) computing the parameter estimates from the individual-level model, (b) treating those estimates as responses in a new model, and (c) modeling those estimates as functions of between-subject predictors.

This second-stage model makes use of between-subject predictors. For Example 13.2, the participant identification is the random effect and is constant across all observations from the same subject (and is therefore a between-subject variable). To construct the population-level model, we need to think about how the α terms from Equation 14.1 might (if at all) vary across individuals in the population. There is no single answer to how this model is constructed; instead, the model should communicate the researchers' beliefs about the data generating process. For example, suppose we are willing to believe the following:

- The delivery of the supplement affects each individual in a similar way. That is, if the tablet is best for one person, we believe it is best for everyone. And, the only reason it may not appear this way in the observed data is due to random noise (which we captured by $\varepsilon_{i,j}$ in the individual-level model).

- The baseline level of fat in each person is fundamentally different. That is, due to diet, genetics etcetera, each person has a unique level of fat. Some people have more than others, and this amount varies randomly in the population, but it is centered on some value.

Again, we might disagree on whether these two assumptions are appropriate, but once we agree on these assumptions, it guides the model. The first assumption above says that the differences due to the supplement types (the treatment effects) in the individual-level model $(\alpha_{1,i}, \alpha_{2,i}, \alpha_{3,i})$ are the same for every individual. That is,

$$\begin{aligned} \alpha_{1,i} &= \beta_1 \\ \alpha_{2,i} &= \beta_2 \\ \alpha_{3,i} &= \beta_3, \end{aligned} \tag{14.2}$$

where $\beta_1, \beta_2, \beta_3$ are unknown parameters. These parameters capture the difference in the fecal fat, on average, between supplement types.

The second assumption says that the the baseline level of fat under the placebo arm, captured by the intercept term $\alpha_{0,i}$ in Equation 14.1, differs across subjects randomly. That is,

$$\begin{aligned} \alpha_{0,i} &= \beta_0 + b_{0,i} \\ b_{0,i} &\sim N\left(0, \sigma_0^2\right) \end{aligned} \tag{14.3}$$

where β_0 and σ_0^2 are unknown parameters. While β_0 captures the overall fecal fat, on average, under the placebo arm, σ_0^2 captures the *variability* in the average fecal fat across individuals in the population taking the placebo. Together, Equation 14.2 and Equation 14.3 create the population-level model.

In this population-level model, we are allowing the intercept term to have a random component as a result of the variability across subjects (hence the name "random effect" for the subject identifier in this case). However, the impact of the treatments are fixed for all subjects (hence the name "fixed effect" for the factor of interest in this case).

Technically, the individual-level and population-level models together fully specify the hierarchical model for the data generating process. However, combining these into a single equation can be instructive. Substituting the population-level model (Equation 14.2 and Equation 14.3) into the individual-level model (Equation 14.1), we have that the response is characterized by

$$\begin{aligned} (\text{Fecal Fat})_{i,j} &= \left(\beta_0 + b_{0,i}\right) + \beta_1(\text{Tablet})_{i,j} + \beta_2(\text{Coated})_{i,j} + \beta_3(\text{Uncoated})_{i,j} + \varepsilon_{i,j} \\ \varepsilon_{i,j} &\overset{\text{IID}}{\sim} N\left(0, \sigma^2\right) \\ b_{0,i} &\overset{\text{IID}}{\sim} N\left(0, \sigma_0^2\right). \end{aligned} \tag{14.4}$$

The idea that some of our effects (coefficients) in a model are fixed (not allowed to vary across individuals) and some are random (allowed to vary across individuals) leads to our description of this approach as a mixed-effects model.

Definition 14.4 (Mixed-Effects Model)**.** A mixed-effects model denotes a hierarchical model for which some effects are fixed (not allowed to vary across subjects) and others are random (allowed to vary across subjects).

> **Note**
>
> Specifying the mixed-effects model as a single model can help with specifying the model in statistical software.

Despite the fully parametric nature of a mixed-effects model, inference on the fixed effects is generally carried out using large-sample theory. Generally, we do not test the random effects; instead, we leave them in the model to capture the correlation we believe is present in the data. Recall that we concluded there was no evidence ($p = 0.168$) the average fecal fat was associated with the type of supplement when we ignored the correlation in the data. However, if we account for the correlation using the mixed-effects model in Equation 14.4, we have strong evidence ($p < 0.001$) the average fecal fat differs for at least one of the supplement types. Correctly accounting for the correlation structure resulted in a more powerful analysis.

> **Warning**
>
> Proper inference in mixed-effects models is debated among statisticians. The two leading statistical software packages (SAS and R) disagree on implementation. SAS provides default p-values for each fixed-effect parameter; R does not.

> **Considerations when Building a Mixed-Effects Model**
>
> Constructing a mixed-effects model can feel overwhelming at times. We urge you to keep the following ideas in mind:
>
> - Construct a model that preserves the behavior at the individual-level. This is generally the stage in which we have more scientific intuition.
> - Allow the behavior to vary across subjects, which corresponds to allowing the parameters to vary. Choose which parameters to vary based on discipline expertise.
> - The variation in the parameters naturally induces a correlation structure in the data.

In the above discussion, we considered the error term at the individual-level model to be independent and identically distributed. In theory, we could allow a correlation to exist here

as well; for example, we may want observations closer together in time to be correlated. This can dramatically increase the complexity of the model fit and often requires custom software.

Before closing this chapter, we address the conditions of a mixed-effects model. We have the same conditions on the individual-level model as we do in classical regression models (Definition 4.3). Further, those conditions can be assessed and relaxed in the same way. However, care must be taken if bootstrapping is used to relax the Normality condition; often times, the bootstrap algorithms need to be custom-written to take advantage of the hierarchy in the data. The conditions on the random effects are not easily assessed. In particular, the distributional assumption of Normality for the random effects is rarely questioned.

In this text, our emphasis is on understanding and interpreting such models. Other texts consider the theoretical underpinnings of such model in more detail.

15 Generalized Estimating Equations

Chapter 13 discussed how studies with repeated measures induce a correlation structure on the data. The previous chapter addressed this correlation structure by developing a hierarchical model in stages, capturing the correlation structure as a by-product. That is, we did not model the correlation directly; instead, by first describing the individual-level model and then allowing the parameters of that model to vary across individuals in the population, the correlation structure was handled naturally. In this chapter, we consider an alternate approach where we focus on modeling the overall average trajectory and the the correlation structure directly. While very general, this approach is particularly popular in longitudinal studies (Definition 13.11).

15.1 Correlation Structrues

Chapter 13 defined the correlation structure as a summary of the relationship among the errors in the response. In a mixed effects model (Definition 14.4), we considered the various sources of variability as contributing to the correlation structure; in this chapter, we are interested in modeling the structure directly. As a result, we are interested in the overall impact of the sources of variability on this structure. By specifying this structure, at least approximately, we are able to adjust the standard errors of our parameter estimates to obtain appropriate inference.

We can think of the correlation on the error terms of our model as a combination of between-subject and within-subject sources of variability. While we may not be discussing the specific sources of variability, they are just as important as before to help us determine an appropriate form of the correlation structure. We are generally willing to assume that observations from different subjects (or blocks) are independent; therefore, when we describe the correlation structure of the errors, we need only focus on the correlation of the observations from the same subject. Further, we assume that the correlation structure is the same for every subject. Therefore, there is only one correlation structure to be specified, and it will be shared across all subjects.

Recall from your introductory course that the correlation coefficient captures the strength and direction of the linear relationship between two variables and must be a value between -1 and 1. If each subject has five observations (for example), then we need to describe the relationship

between any pair of these five observations. That is, we need $\binom{5}{2} = 10$ correlation coefficients. We store these in a matrix

$$\Gamma = \begin{pmatrix} 1 & \rho_{1,2} & \rho_{1,3} & \rho_{1,4} & \rho_{1,5} \\ \cdot & 1 & \rho_{2,3} & \rho_{2,4} & \rho_{2,5} \\ \cdot & \cdot & 1 & \rho_{3,4} & \rho_{3,5} \\ \cdot & \cdot & \cdot & 1 & \rho_{4,5} \\ \cdot & \cdot & \cdot & \cdot & 1 \end{pmatrix}, \tag{15.1}$$

where the lower-half is determined from the upper-half since the correlation between the i-th and j-th observations is the same as the correlation between the j-th and i-th observations; that is, $\rho_{i,j} = \rho_{j,i}$.

Properties of Correlation Matrices

Every correlation matrix has the following properties:

1. It is a square matrix, and the dimension is determined by the number of observations within a subject.
2. It is symmetric (the transpose is the same as the original matrix). That is, $\rho_{i,j} = \rho_{j,i}$.
3. The diagonal entries are always 1; any value is perfectly correlated with itself.
4. All off-diagonal elements must be between -1 and 1.

A correlation matrix is very similar to a variance-covariance matrix; in fact, we can think of a correlation matrix as a standardized variance-covariance matrix.

The correlation matrix in Equation 15.1 makes no assumptions about the structure of the off-diagonal elements (other than they are values between -1 and 1). This is known as an unstructured form.

Definition 15.1 (Unstructured Correlation Structure)**.** An unstructured correlation structure suggests that the correlation between any two errors within a subject can take on any value. We only require that it be a valid correlation matrix.

If we think of each correlation as an additional parameter to estimate, then we have just specified an additional $\binom{m}{2}$ parameters to our model, where m is the number of repeated observations on a subject. We are essentially choosing not to place any structure on the correlation matrix and allow the data to completely determine the structure. This can be useful if we have no intuition about the sources of variability; however, it requires a lot of data as we have added a large number of parameters to the model.

As in any model, there is tension between specifying a model which is flexible and one which is more tractable. We often impose some simplifying structure on the correlation matrix. While

there are several possible structures, we discuss the most common. On the other extreme from the unstructured correlation matrix discussed above is to assume the observations within a subject are independent of one another.

Definition 15.2 (Independence Correlation Structure)**.** An independence correlation structure suggests there is no correlation among any of the error terms within a subject. If there are five observations within a block, this has the form

$$\Gamma = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ \cdot & 1 & 0 & 0 & 0 \\ \cdot & \cdot & 1 & 0 & 0 \\ \cdot & \cdot & \cdot & 1 & 0 \\ \cdot & \cdot & \cdot & \cdot & 1 \end{pmatrix}.$$

While we already assume that observations between subjects are independent, the independence structure goes further and essentially says all observations are independent. At first glance, this would seem to revert back to the classical regression model, which we have already established is inappropriate for repeated measures. However, we will argue later that using such a structure does have some differences.

When we feel that observations from the same subject are associated primarily because they are from the same subject, and that the order of the observations within the subject is irrelevant, a compound symmetric correlation structure is appropriate.

Definition 15.3 (Compound Symmetric Correlation Structure)**.** A compound symmetric correlation structure, also known as an *exchangeable* correlation structure, suggests the correlation between any two errors within a subject is equal. If there are five observations within a block, this has the form

$$\Gamma = \begin{pmatrix} 1 & \rho & \rho & \rho & \rho \\ \cdot & 1 & \rho & \rho & \rho \\ \cdot & \cdot & 1 & \rho & \rho \\ \cdot & \cdot & \cdot & 1 & \rho \\ \cdot & \cdot & \cdot & \cdot & 1 \end{pmatrix}.$$

The compound symmetric structure adds only one additional parameter to our model and actually models well many scenarios. When we do believe that the order of the observations within a subject is important, and that observations occurring closer together (generally in time) are more highly correlated than observations further apart in time, an autoregressive structure is appropriate.

Definition 15.4 (Autoregressive Correlation Structure)**.** An autoregressive correlation structure suggests the correlation between two observations diminishes as the observations get further apart (generally, further apart in time). We generally only consider the autoregressive structure of degree 1 here; if there are five observations within a block, this has the form

$$\Gamma = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \cdot & 1 & \rho & \rho^2 & \rho^3 \\ \cdot & \cdot & 1 & \rho & \rho^2 \\ \cdot & \cdot & \cdot & 1 & \rho \\ \cdot & \cdot & \cdot & \cdot & 1 \end{pmatrix}.$$

The autoregressive structure is borrowed from the time-series literature. It is primarily useful when we are taking observations somewhat close together in time. Like the compound symmetric structure, it only adds a single parameter to the model.

Regardless of which of the structures we believe is beneath the data, we also assume stationarity.

Definition 15.5 (Stationarity)**.** The assumption of stationarity states that the correlation structure does not depend on time, only the distance between the observations.

Essentially, stationarity says that at no point does the structure evolve as the study continues; that is, we cannot include a parameter such as $\rho^{(\text{time})}$.

Choosing an appropriate structure is often guided by discipline expertise regarding how the sources of variability combine and impact the relationship between the responses. When multiple sources of variability have competing structures, we generally adopt the structure of the "dominant" source. However, it turns out that the choice of the structure need not have a large impact on the analysis — simply indicating in the analysis that there is a potential for correlation can be sufficient. This is the idea behind the approach we describe.

15.2 The Key to Success of Generalized Estimating Equations

In the previous section, we considered models for the correlation structure that result from the combination of the various sources of variability in the data generating process. This structure will be used within the generalized estimating equation (GEE) approach.

Definition 15.6 (Generalized Estimating Equations (GEE))**.** Generalized estimating equations can be used to estimate the parameters of a model while accounting for the correlation among observations. In addition to specifying a model for the overall average response, a "working" structure is specified for the correlation of observations from the same subject. The working structure is updated during the estimation process and used to adjust the standard errors of the parameter estimates in the mean model.

Recall that our inference on the parameters requires us to compute the variance-covariance matrix of the corresponding parameter estimates. When a model is estimated using generalized estimating equations, the model we specify for the correlation structure is known as the "working" correlation matrix; this is then updated using the observed data when computing the variance-covariance matrix. As a result, the variance-covariance matrix we use is not based solely on the specified model but is a blend of the model specified and the observed data; this is known as the robust sandwich estimator.

Definition 15.7 (Robust Sandwich Estimator)**.** The robust sandwich estimator of the variance-covariance matrix of the parameter estimates from the mean model balances the relationship between the parameter estimates specified by the model (and the "working" correlation matrix) with the relationship suggested by the observed data. Specifically, it has the form

$$\widehat{\Sigma} = \widehat{\mathbf{U}}\widehat{\mathbf{U}}^{-1/2}\mathbf{R}\widehat{\mathbf{U}}^{-1/2}\widehat{\mathbf{U}}$$

where $\mathbf{U}$ represents the model-based variance-covariance matrix if the structure specified by the working correlation matrix were completely correct, and $\mathbf{R}$ represents the correction factor estimated from the residuals (an empirical estimate).

> **Big Idea**
>
> The use of the robust sandwich variance-covariance estimator is what makes the GEE approach unique and so powerful.

While the structure of $\mathbf{U}$ is beyond the scope of this text, we can think of it as what the computer does by default when we specify a model under the classical conditions. Essentially, the use of the robust sandwich estimator in the GEE framework means our posited correlation structure need not be correct; it is okay if $\mathbf{U}$ is wrong. With enough data, the inference will be the same regardless of the structure we choose. What we are really specifying is that there is a potential for correlation among these observations. Of course, the better the specified model for the correlation structure, the less adjustment that is needed and the more powerful the results.

> **Note**
>
> It is the use of the robust-sandwich estimator that makes specifying the "independent" correlation structure different than assuming the classical regression model. In classical regression, inference is based on assuming independence. In a GEE framework, the correlation structure will be updated after we assume independence.

> **Note**
>
> While we are discussing the use of the robust-sandwich estimator as a way of adjusting for the correlation present, we note that this will also adjust for violations in constant variance as a result.

Consider the study to investigate the impact of four methods of delivering a pancreatic enzyme supplement (Example 13.2). In the GEE approach, we focus on specifying a model for the overall mean response; that is, our modeling is very similar to what we would do if we ignored the correlation altogether:

$$(\text{Fecal Fat})_{i,j} = \beta_0 + \beta_1(\text{Tablet})_{i,j} + \beta_2(\text{Coated})_{i,j} + \beta_3(\text{Uncoated})_{i,j} + \varepsilon_{i,j}. \tag{15.2}$$

Notice that our model does acknowledge that multiple observations were collected on each subject (we used both an i and j index). The difference between the GEE approach and the classical approach is that we now specify a correlation structure for the errors. As described in Example 13.2, there are four observations within each participant. Since we believe that diet is the primary reason for differences in fecal fat across individuals, the primary source of variability is the participant. This leads us to posit the following compound symmetric working correlation structure:

$$\Gamma = \begin{pmatrix} 1 & \rho & \rho & \rho \\ \cdot & 1 & \rho & \rho \\ \cdot & \cdot & 1 & \rho \\ \cdot & \cdot & \cdot & 1 \end{pmatrix}.$$

We would then be sure to use the robust-sandwich estimator when computing the standard errors of the parameter estimates in Equation 15.2.

When we fit a model using generalized estimating equations, we are fitting a semiparametric model — specifying the model for the mean response and the correlation structure. Inference on parameters is then carried out using large-sample theory. Recall that we concluded there was no evidence ($p = 0.168$) the average fecal fat was associated with the type of supplement when we ignored the correlation in the data. However, if we account for the correlation estimating the parameters using generalized estimating equations with a compound-symmetric working correlation matrix, we have strong evidence ($p = 0.003$) the average fecal fat differs for at least one of the supplement types. Correctly accounting for the correlation structure resulted in a more powerful analysis.

15.3 Comparison of GEE and Mixed Effects Approaches

While both mixed effects models and estimation via generalized estimating equations account for the correlation structure, the two approaches differ in many ways. The mixed effects modeling approach is fully parametric, while estimating via GEE is semiparametric.

More broadly, these represent two different approaches to repeated measures data: subject-specific and population-averaged.

Definition 15.8 (Subject Specific Models)**.** Also known as conditional modeling, the subject-specific approach models at the subject-level and addresses the correlation indirectly through the inclusion of random effects.

Definition 15.9 (Population Averaged Models)**.** Also known as marginal modeling, the population-averaged approach posits a model for the mean response directly and addresses the correlation through directly modeling its structure.

If you have more intuition about how the response should be characterized at the individual level, then the subject-specific (mixed-effects) approach will be more tractable. If you have more intuition about how the response should be characterized on average across individuals, then the population-averaged (estimated with GEE's) approach will be more tractable. The GEE approach is fairly robust to misspecifications of the working correlation structure and constant variance conditions. The mixed-effects model approach allows us to quantify the variability in an effect across subjects in the population.

Part V

Unit V: Nonlinear Models

Chapter 12 considered an approach for modeling curvature between the response and predictor(s). That curvature was captured by functions that were linear in the *parameters*, exploiting the flexibility in the general linear model framework. But, these models were largely empirical. That is, while the form of the model was certainly influenced by the research question of interest, any curvature was primarily suggested by the data itself. When we begin the modeling process with a mathematical representation of the scientific processes (such as those in chemical engineering, ecology, cellular biology, etc.), the resulting model may be nonlinear in the parameters.

In the previous units of the text, we have assumed the response to be a quantitative variable. However, when the response is a categorical variable, the most appropriate models are often nonlinear in the parameters.

In each of these settings, while a general linear model may be constructed that would capture the curvature in the data, a nonlinear model better represents the data generating process. Further, the scientific questions of interest can often be represented by the parameters of that nonlinear model. In this unit, we introduce a framework for making inference on the parameters of nonlinear models.

16 Nonlinear Model Framework

There are several disciplines that routinely use nonlinear models; one such discipline is pharmacokinetics (the study of how medications are processed by the body).

Example 16.1 (Pharmacokinetics of Theophylline)**.** An early-phase clinical study was conducted to assess how Theophylline (an anti-asthmatic agent) is absorbed and eliminated from the human body. A single subject was given an oral dose of 4 mg of the drug, and 11 blood samples were taken over the course of a 24-hour period to determine the concentration of the drug in the body. A graphical summary of the resulting data is presented in Figure 16.1.

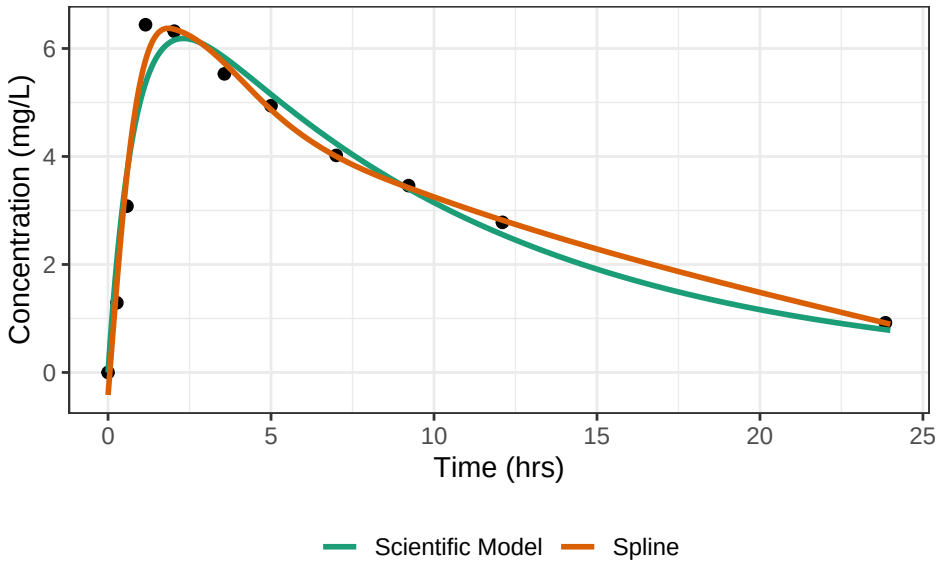


Figure 16.1: Concentration of Theophylline in the blood stream of a patient over a 24-hour period.

Given the topics discussed thus far in the text, it would be reasonable to consider using a flexible spline to model the curvature in the response over time. Based on Figure 16.1, qualitatively, using a spline appears to perform rather well. While this approach fits the data well and would allow us to construct good predictions, using a spline does not allow us to easily address the research objective: quantify how quickly the drug is absorbed by and eliminated

from the body. That is, this initial modeling approach does not lend itself to expressing the question of interest as a statement about the parameters in the model. It also separates the data analysis from the scientific modeling that suggests the concentration of Theophylline in the body at any point t in time is given by

$$C(t) = \frac{k_a D}{(\beta/k_e)(k_a - k_e)} \left(e^{-k_e t} - e^{-k_a t}\right), \tag{16.1}$$

where k_a (the absorption rate), k_e (the elimination rate), and β (the clearance) are the unknown parameters characterizing the pharmacokinetics of Theophylline, and D is the known dosage given.

Big Idea

Nonlinear models are often the result of embedding a scientific model in a statistical framework.

Scientific Model for Theophylline

We illustrate how the scientific model for the pharmicokinetics of Theophylline suggests a nonlinear model. Readers not familiar with differential equations can skip this without loss of continuity.
Researchers believe that the absorption and elimination of Theophylline can be modeled using a one-compartment open model with first order absorption, represented by Figure 16.2. The box represents the "blood compartment" (blood stream). The drug is absorbed into the blood stream through the gut; it is then metabolized by the liver and excreted by the kidneys.

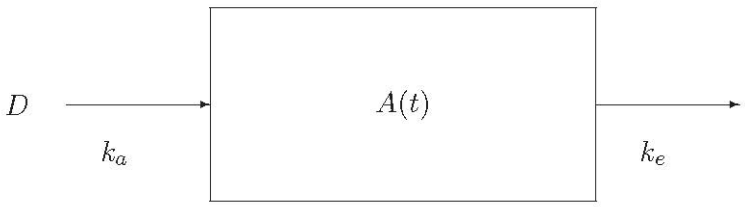


Figure 16.2: Scientific model for pharmicokinetics of Theophylline.

Mathematical modeling allows us to characterize the concentration of Theophylline in the blood over time as a function of the elimination rate, the absorption rate, and the clearance rate (measures volume of blood cleared of the drug per unit time). Let D represent the oral dose given at time $t = 0$; as the subject has not been given the

treatment prior, it is reasonable to assume that the amount of drug in the blood is initially 0 mg; that is, $A(0) = 0$. Further, researchers believe the body can absorb the entire dose D; therefore, the amount of drug at the absorption site $A_a(t)$ is initially 4 mg; that is, $A_a(0) = D = 4$. Letting k_a and k_e represent the absorption and elimination rates, respectively, we have the following mathematical model corresponding to the above scientific model:

$$\frac{d}{dt}A(t) = k_a A_a(t) - k_e A(t)$$
$$\frac{d}{dt}A_a(t) = -k_a A_a(t).$$

Scientists believe that the amount of drug in the blood stream at any time is proportional to the concentration of the drug at that time $C(t)$. That is, $A(t) = VC(t)$, where V represents the volume of the blood compartment. The above system of differential equations can be solved using Laplace transforms. Letting $V = \beta/k_e$ where β is the clearance rate, we are led to the following solution:

$$C(t) = \frac{k_a D}{(\beta/k_e)(k_a - k_e)}\left(e^{-k_e t} - e^{-k_a t}\right).$$

What we want to emphasize is that the form of the model for the concentration of Theophylline in the body was not developed empirically using statistical modeling techniques; instead, it was derived through mathematical modeling of the drug itself from scientific principles. The parameters of the model k_a, k_e, and β are unknown, but they govern the process. More, these parameters are directly related to the scientific question of interest. We can now embed this scientific model into a statistical framework (accounting for sources of variability) in order to make inference on the parameters.

Notice that there is no way of rewriting the model in Equation 16.1 as a linear combination of the parameters k_a, k_e, and β; that is, there is no vector $\mathbf{x}$ such that

$$C(t) = \mathbf{x}^\top \begin{pmatrix} k_a \\ k_e \\ \beta \end{pmatrix}.$$

Therefore, our model for the concentration of Theophylline is a nonlinear model.

Definition 16.1 (Nonlinear Model)**.** A model is said to be nonlinear if it cannot be written as a linear combination of the parameters.

> **Note**
>
> When your model has more parameters than predictors, you likely have a model that is nonlinear in the parameters.

16.1 Nonlinear Regression Model

Generalizing our approach with the general linear model, we consider a semiparametric modeling perspective for most nonlinear models. Under this approach, we specify the model for the mean response and the variability of the response (given the predictors). We then use the method of least squares to obtain our estimates of the parameters, and we rely on large sample theory to characterize the sampling distribution of these parameter estimates. This framework turns out to be quite flexible.

Our semiparametric approach, also known as a "moment model," focuses on specifying the mean and variance of the response given the predictors.

Definition 16.2 (Semiparametric Nonlinear Model)**.** A semiparametric nonlinear model specifies the *mean* and *variance* of the response given the predictors; we write

$$\begin{aligned} E\left[(\text{Response})_i \mid (\text{Predictors})_i\right] &= f\left((\text{Predictors})_i, \beta\right) \\ Var\left[(\text{Response})_i \mid (\text{Predictors})_i\right] &= \sigma^2 \end{aligned}$$

where $f(\cdot)$ is referred to as the mean response function.

For Example 16.1, our nonlinear model would have the form

$$\begin{aligned} E\left[(\text{Concentration})_i \mid (\text{Time})_i\right] &= \frac{4k_a}{(\beta/k_e)\left(k_a - k_e\right)}\left(e^{-k_e(\text{Time})_i} - e^{-k_a(\text{Time})_i}\right) \\ Var\left[(\text{Concentration})_i \mid (\text{Time})_i\right] &= \sigma^2. \end{aligned} \tag{16.2}$$

Notice that under this specification there is no "error" term. We only concern ourselves with positing the mean and variance; not only do we not specify the distribution of the "errors," we do not even make use of them in our conceptualization of the model.

> **Note**
>
> Some authors choose to write nonlinear models as
>
> $$(\text{Response})_i = f\left((\text{Predictors})_i, \beta\right) + \varepsilon_i,$$

and sometimes go on to assume $\varepsilon_i \overset{\text{IID}}{\sim} N(0, \sigma^2)$. However, this approach is not as flexible. First, it requires the response to be quantitative. By only specifying the mean and variance of the response, our approach generalizes to accommodate categorical responses (though we focus on quantitative responses for the majority of this unit). Second, there is little to be gained by a fully parametric approach as this does not avoid the requirements for large sample theory as it does in the linear model case. That is, even if we assume the errors follow a Normal distribution, we must rely on large sample theory to model the sampling distribution of the resulting parameter estimates.

Finally, we note that in our specification above, we have assumed the variability of the response is constant for all values of the predictor. We will relax this condition in Chapter 17.

Note

In some cases, the mean response function $f(\cdot)$ lends itself to a transformation that results in a linear model. For example, if $f(x, \beta) = \beta_0 e^{\beta_1 x}$, then taking the natural logarithm of both sides results in a linear model:

$$\ln(f(x, \beta)) = \ln(\beta_0) + \beta_1 x.$$

As a result, in some disciplines, it is standard practice to consider such transformation prior to modeling. There is not widespread agreement in the statistical community on such transformations.

Applying some function so that the model is linear in the parameters (or at least, some function of the parameters) allows standard software to be used. As scientists are familiar with approaches for fitting linear models, this often makes it easy to put in a context that is understood. However, such transformations may destroy other modeling conditions, such as that of constant variance (or any distributional assumptions, if applicable). Transformations can also result in the model no longer being parameterized by the scientific quantities of interest. As scientific principles often directed the development of the nonlinear model, this is the scale on which scientists have intuition (it may be more difficult to think on the logarithmic-scale, for example), meaning scientists have less intuition on the transformed scale. After transformation, the model may no longer characterize the average response. Finally, there are some models for which such a transformation is not possible. As a result, we argue for using the nonlinear regression framework and studying methods for fitting such models.

Warning

There is a difference between taking nonlinear transformations of a predictor to fit curvature in a linear model and taking nonlinear transformations of a model (the response

often) in order to convert a nonlinear model into a model which is linear with respect to some function of the parameters.

16.2 Estimation

The method of least squares for nonlinear models is similar to the process for linear models. We choose the values of the parameters that ensures the predicted mean function is as "close" to the observed responses as possible. While we do not have an error term in the model, we can still consider the residuals when defining how "close" our predicted mean response is to the observed responses. The least squares estimates minimize

$$\sum_{i=1}^{n} \left[(\text{Response})_i - f\left((\text{Predictors})_i, \beta\right)\right]^2.$$

In the nonlinear model literature, this is often referred to as the *ordinary* least squares estimator.

Note

If we take $f(\cdot)$ to be a linear function in the parameters,

$$f\left((\text{Predictors})_i, \beta\right) = \sum_{j=1}^{p} (\text{Predictor } j)_i \beta_j,$$

the linear model framework is a special case of the nonlinear regression model.

The least squares estimates are computed numerically (for details, see Chapter 20).

Given parameter estimates, we can then estimate the residual variance:

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^{n} \left[(\text{Response})_i - f\left((\text{Predictors})_i, \hat{\beta}\right)\right]^2$$

where p represents the number of *parameters* in the model, which need not correspond to the number of predictors in the model.

As with previous models discussed in the text, the process of estimating the parameters is a mathematical problem. It is the process of making inference where we move into the realm of statistics.

16.3 Inference on the Parameters

In order to make inference about the parameters, we need a model for the sampling distribution of the parameter estimates. Unlike with the general linear model, we cannot rely on probability theory to obtain exact models for the sampling distributions. Therefore, we rely on large sample theory and empirical models (Chapter 11).

Definition 16.3 (Large Sample Model for the Sampling Distribution of the Least Squares Estimates in Nonlinear Models)**.** Consider a nonlinear model as described in Definition 16.2. Assuming the form of the model is correctly specified, as the sample size gets large, we have that

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{Var\left(\hat{\beta}_j\right)}} \sim N(0,1)$$

for all $j = 1, \ldots, p$. Further, under the null hypothesis

$$H_0 : \mathbf{K}\beta = \mathbf{m}$$

we have that

$$\left(\mathbf{K}\hat{\beta} - \mathbf{m}\right)^\top \left(\mathbf{K}\widehat{\Sigma}\mathbf{K}^\top\right)^{-1} \left(\mathbf{K}\hat{\beta} - \mathbf{m}\right) \sim \chi^2_r$$

where r is the rank (number of rows) of $\mathbf{K}$ and $\widehat{\Sigma}$ is the estimated variance-covariance matrix of the parameter estimates.

> **Note**
>
> Large sample theory is often relied on regardless of the sample size. It is often the case that there is so little error in the responses that the sampling distributions of the estimates comes close to these asymptotic approximations even in small samples.
> Making additional distributional assumptions does not avoid the need for large sample theory.

The above results allow us to not only construct confidence intervals, but we can also make use of the general linear hypothesis testing framework (Chapter 10) for testing specific hypotheses. That is, our inference is not all that different than under the general linear model framework once we have estimates for the parameters and estimates for their standard errors.

16.4 Allowing Relationships to Vary Across Groups

Modeling involves positing a relationship between the response and the predictors. Suppose we believe the form of the model is similar for all subjects in a population, but the specific parameters may differ across sub-populations. By including interaction terms, we can allow the relationship to vary across sub-populations.

> **Big Idea**
>
> Interaction terms allow a parameter to depend on the value of another variable.

The idea of allowing a relationship to depend upon a predictor is something we have studied in each of the previous units. The same ideas apply in nonlinear models; the difference is that computationally, it requires more to specify the model appropriately in the computer.

> **General Approach for Including Interactions**
>
> In order to allow parameters in a model vary across sub-populations, consider the following:
>
> - Refine your question in terms of which parameter(s) will be allowed to vary, and which parameter(s) will remain fixed for all groups.
> - Use an indicator variable to capture the grouping structure.
> - Embed your research question into the general linear hypothesis framework.
>
> Additionally, when fitting the model using software, starting estimates are typically needed (see Chapter 20); these can be obtained by fitting separate models to subsets of the data; that is, you can fit the model in each subpopulation and use the resulting estimates as starting estimates for the actual model of interest.
> These considerations are quite general and can be widely applied.

Example 16.2 (Bacteria Growth)**.** Suppose we believe a particular bacteria grows exponentially. That is, we believe that, on average, the response is given by

$$E\left[(\text{Response})_i \mid (\text{Time})_i\right] = \alpha_0 + e^{\alpha_1 (\text{Time})_i}$$

at each observed time.

Researchers are considering two different media for culturing the bacteria. Before proceeding to production, they would like to determine if there is evidence that the growth trajectory of the bacteria differs between the two media.

While wells containing the bacteria are seeded at the same amount regardless of the media (same amount of bacteria to begin with), we believe the media might result in different growth rates. As a result, we expect the value of α_0 to be the same for both media, but the value of α_1 could differ for the each of the two media.

To develop a model consistent with the researchers' beliefs in Example 16.2, we consider an interaction term. Specifically, define an indicator variable

$$(\text{Media B})_i = \begin{cases} 1 & \text{if i-th observation corresponds to bacteria grown in media B} \\ 0 & \text{if i-th observation corresponds to bacteria grown in media A.} \end{cases}$$

Now, consider the following nonlinear model:

$$E\left[(\text{Response})_i \mid (\text{Time})_i, (\text{Media})_i\right] = \beta_0 e^{(\beta_1+\beta_2(\text{Media B})_i)(\text{Time})_i} = \beta_0 e^{\beta_1(\text{Time})_i+\beta_2(\text{Media B})_i(\text{Time})_i}$$
$$Var\left[(\text{Resposne})_i \mid (\text{Time})_i, (\text{Media})_i\right] = \sigma^2.$$

This model has an interaction between the indicator variable for Media B and the time; however, there is no "main effect" of the media; that is, there is no term that has $\beta_3(\text{Media B})_i$. This single model implies expected responses for each media:

$$\text{Media A} : E\left[(\text{Response})_i \mid (\text{Time})_i\right] = \beta_0 + e^{\beta_1(\text{Time})_i}$$
$$\text{Media B} : E\left[(\text{Response})_i \mid (\text{Time})_i\right] = \beta_0 + e^{(\beta_1+\beta_2)(\text{Time})_i}.$$

Notice this model maintains the assumptions we had about the process:

- The initial state is the same under both media (β_0).
- The growth rate is allowed to differ under the two media (β_1 vs. $\beta_1 + \beta_2$).

As discussed when we introduced interaction terms in linear models, the benefits of using interaction terms instead of doing a subgroup analysis is that we can easily control which parameters are allowed to differ, we make use of the constant variance condition gaining power, and our question of interest is embedded in the model. To illustrate this last point, notice that the hypotheses

$$H_0 : \gamma_2 = 0 \qquad \text{vs.} \qquad H_1 : \gamma_2 \neq 0$$

test whether the growth rate actually differs between the two media.

17 Relaxing the Constant Variance Condition

We proposed a semiparametric approach for specifying nonlinear models (Definition 16.2). In this chapter, we first discuss the conditions imposed in this approach; then, we consider how these conditions might be relaxed. Specifically, we discuss methods for relaxing the "constant variance" condition.

> **Note**
>
> While we introduce methods for relaxing the "constant variance" condition within the context of nonlinear models, these methods are applicable to a wider range of models including linear models and models for repeated measures data.

17.1 Conditions for Nonlinear Models

Chapter 16 introduced the nonlinear modeling framework, a semiparametric approach for modeling the data generating process, and methods for estimating the parameters of the model. While a semiparametric model is quite flexible, its specification does place certain conditions on the data generating process.

> **Conditions for the Semiparametric Nonlinear Model**
>
> A semiparametric model places the following conditions on the data generating process:
>
> - The mean response function is correctly specified.
> - Given the value of the predictors, the response for one observation is independent of the response for all other observations.
> - The variability of the response is the same for all values of the predictor (also known as "homoskedasticity" or "constant variance").

These are essentially the same first three conditions we imposed in the general linear model framework (Definition 4.3); since the linear model is a special case of the nonlinear model, it should not come as a surprise that we impose the same conditions. As the conditions imposed on our nonlinear model are the same as those imposed in the general linear model framework, the conditions can be assessed in the same way — residual graphics. We can define the residual for the i-th observed value as

$$\begin{aligned}(\text{Residual})_i &= (\text{Observed Response})_i - (\text{Predicted Mean Response})_i \\ &= (\text{Observed Response})_i - f\left((\text{Predictors})_i, \widehat{\beta}\right).\end{aligned}$$

A plot of the residuals against the predicted values can be used to assess whether the mean response function is correctly specified. If this condition is met, we would expect the residuals to balance around zero for all predicted values. Recall that for nonlinear models, the mean response function is often the result of scientific theory; therefore, assessing this "mean-0" condition allows us to assess if the process we are observing is behaving according to the scientific theory underlying the model.

When the order in which the data was collected is known, we can use a time-series plot of the residuals to assess whether it is reasonable to assume the responses are independent of one another. If the response, given the predictors, of one observation is independent of the response for all other observations, we would not expect to observe any trends in the location or spread of the time-series plot. As with the linear model, we note that this graphic can only detect dependence across time; we should rely on the context to determine if there are additional reasons to suspect dependence (such as repeated measures).

Finally, we can assess the condition imposed on the variability using a plot of the residuals against the predicted values. If the variance of the response has been correctly specified, then we would expect the spread of the residuals to remain fairly constant as we move left-to-right across the plot.

> **Note**
>
> Nonlinear modeling applications often have small samples. As a result, it can be difficult to assess constant variance from the plot of the residuals against the fitted values. A "trick" is to plot the absolute value of the residuals against the fitted values. This "doubles" the visual information in the graphic and can allow us to more easily pick up trends in the spread.

> **Note**
>
> While we do not assume the response (conditional on the predictors) follows a Normal distribution, if we had, we could assess this condition using a probability plot of the residuals.

Example 17.1 (Pharmacokinetics of Indomethacin)**.** Indomethacin is a non-steroidal anti-inflammatory (NSAID) pain reliever used to treat severe pain and prevent premature labor in some cases. A study was conducted to examine the pharmacokinetic properties of the drug. Indomethacin is given as an IV-bolus and travels through the blood and deeper tissues.

Scientists model this as a "two-compartment open model." This approach leads to the following nonlinear model for the concentration $C(t)$ of the drug at any time t:

$$C(t) = \beta_1 e^{-\beta_2 t} + \beta_3 e^{-\beta_4 t}$$

which is also known as the bi-exponential model with four parameters, $\beta_1, \beta_2, \beta_3$, and β_4.

Blood samples were taken periodically (over the course of 8 hours) from a single subject after being given an IV-bolus of the drug. The data is shown in Figure 17.1.

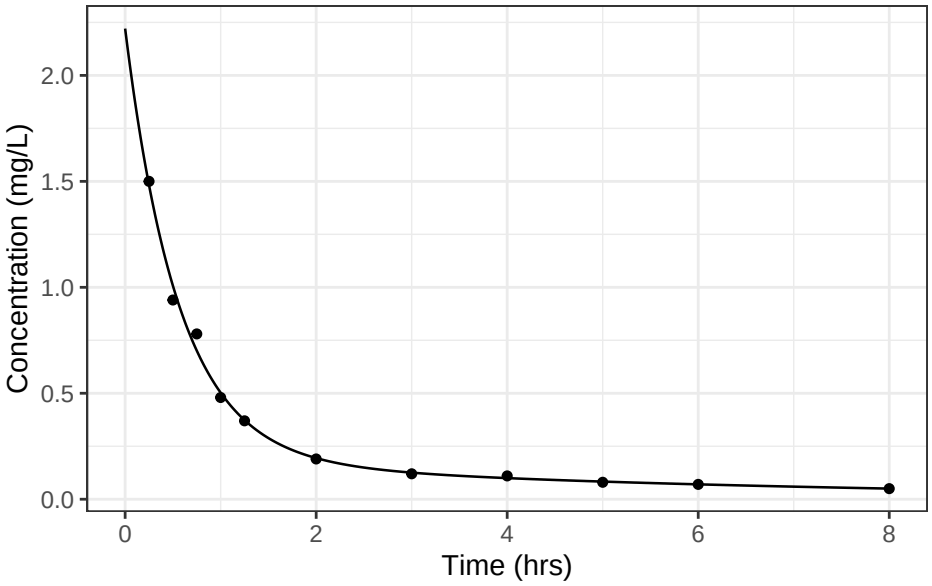


Figure 17.1: Concentration of Indomethacin in the bloodstream for a single subject after being given an IV-bolus of the drug. The estimated bi-esponential model is overlaid.

Given the researchers' beliefs, we might posit the following nonlinear model:

$$\begin{aligned} E\left[(\text{Concentration})_i \mid (\text{Time})_i\right] &= \beta_1 e^{-\beta_2 (\text{Time})_i} + \beta_3 e^{-\beta_4 (\text{Time})_i} \\ Var\left[(\text{Concentration})_i \mid (\text{Time})_i\right] &= \sigma^2. \end{aligned} \tag{17.1}$$

The unknown parameters are estimated using the method of least squares (Chapter 20); and, we can assess the conditions using the residuals. Figure 17.2 plots the residuals against the predicted values; and, it plots the absolute value of the residuals against the fitted values. In the first panel, we see the residuals tend to balance around 0 at all predicted responses; that is, the data is consistent with the two-compartment open model suggested by researchers (that led to the bi-exponential nonlinear model). However, in both the first and second panel, it is

clear that the spread of the residuals increases as the predicted concentration increases. That is, we measure small concentrations with more precision than large concentrations.

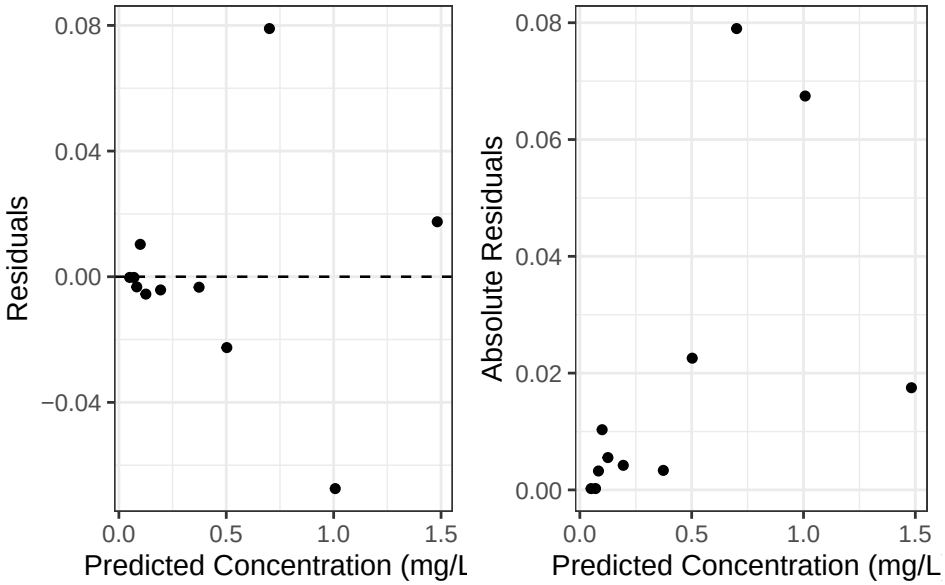


Figure 17.2: Plot of the residuals against the predicted values for data from a study of the pharmacokinetics of Indothemacin.

Recall that misspecifying the variance does not mean that our parameter *estimates* are inappropriate. However, incorrectly assuming the variance is constant results in an inappropriate model for the sampling distribution of those estimates, which in turn means our confidence intervals and p-values may be unreliable.

In Chapter 15, we mentioned that using the robust sandwich estimator of the variance-covariance matrix adjusts for departures from constant variance. While this was done in the context of repeated measures, the robust sandwich estimator can be employed in settings where we know each observation is independent.

> **Note**
>
> The robust sandwich estimator is sometimes referred to as the Huber sandwich estimator, the White estimator, or the Huber-White estimator.

In the remainder of this chapter, we consider two additional approaches to addressing departures from homoskedasticity.

17.2 Modeling the Variance

Notice that to assess the condition of constant variance, we plot the residuals against the predicted values. This is not necessarily intuitive. In fact, the condition states that the variability of the response is constant for all values of the predictors; so, it may seem more reasonable to plot the residuals against each predictor. In fact, this is sometimes taught in other texts and routinely done by analysts, and there is nothing wrong with that approach. We advocate for plotting the residuals against the predicted values because it highlights a common phenomena — the variability of the response often depends on the *value* of the response.

In our discussion in the previous section, we noted that Figure 17.2 illustrates that as the concentration increases, the variability in the concentration tends to increase as well. That is, we have much more precision when measuring small concentrations, and we have much less precision when measuring large concentrations. When we have some sense of how the variability behaves, especially as a function of the mean response, we can model that structure.

> Big Idea
>
> When we can specify how a component of the response distribution (such as the mean or variance) behaves, we can incorprate it into the model.

Definition 17.1 (Generalized Least Squares)**.** The semiparametric nonlinear model can be generalized to capture non-constant variance. Specifically, we specify the *mean* and *variance* of the response given the predictors

$$\begin{aligned} E\left[(\text{Response})_i \mid (\text{Predictors})_i\right] &= f\left((\text{Predictors})_i, \beta\right) \\ Var\left[(\text{Response})_i \mid (\text{Predictors})_i\right] &= g\left((\text{Predictors})_i, \beta, \gamma\right). \end{aligned}$$

Such a model is fit with the method of generalized least squares (as opposed to "ordinary" least squares) in which we alternate between (a) minimizing a weighted distance between the observed response and the mean function and (b) minimizing the distance between the squared residuals and the variance function. That is, we minimize

$$\begin{aligned} &\sum_{i=1}^{n} \frac{1}{g\left((\text{Predictors})_i, \beta, \gamma\right)} \left[(\text{Response})_i - f\left((\text{Predictors})_i, \beta\right)\right]^2 \\ &\sum_{i=1}^{n} \left(g\left((\text{Predictors})_i, \beta, \gamma\right) - \left[(\text{Response})_i - f\left((\text{Predictors})_i, \beta\right)\right]^2\right)^2. \end{aligned}$$

Essentially, this approach down-weights responses which have less precision when fitting the model. As an iterative process, it allows the parameters in the mean model to be updated based on the variance estimates, and the estimates in the variance function to be updated based on

the mean estimates. Further, allowing the variance function to depend on the parameters in the mean response function captures behaviors like that in the Indomethacin example where the variability is dependent upon the values of the response. A popular model is the power of the mean model.

Definition 17.2 (Power of the Mean Model)**.** The power of the mean model allows the variance to be specified as a power of the mean response function. Specifically, we consider

$$E\left[(\text{Response})_i \mid (\text{Predictors})_i\right] = f\left((\text{Predictors})_i, \beta\right)$$
$$Var\left[(\text{Response})_i \mid (\text{Predictors})_i\right] = \sigma^2\left[f\left((\text{Predictors})_i, \beta\right)\right]^{2\theta}$$

The details of implementing generalized least squares are beyond the scope of this text. We simply mention that this is a method for addressing heteroskedasticity (non-constant variance). Implementing generalized least squares results in appropriate inference about the parameters in the mean model. Large sample theory is relied upon to develop these sampling distributions.

17.3 Wild Bootstrap

Chapter 11 introduced the residual bootstrap as a method of relaxing distributional conditions. Specifically, the residual bootstrap avoided the need to assume the errors in a linear model followed a Normal distribution. However, it still required assuming the variability of the errors was constant; that is what allowed us to resample from the residuals during the algorithm. In this section, we extend these ideas to overcome heteroskedasticity in nonlinear models.

At first glance, it is not obvious why an additional algorithm for bootstrapping is necessary. In addition to the residual bootstrap, we discussed case-resampling as a method of bootstrapping. Case-resampling does not require that we assume the variability of the response is constant for all values of the predictor. However, the performance of this particular algorithm can be quite poor in nonlinear settings due to the lower sample sizes that are common. In particular, fitting a nonlinear model can be very unstable. If, when resampling the cases, we do not observe data in key regions that define the curvature, the numerical algorithms underlying the minimization can fail to converge to a solution. As an example, consider the Indomethacin data in Figure 17.1. Suppose that we performed a single bootstrap resample in which we resampled $n = 11$ observations, with resampling, at random. It is quite possible that we end up with a resample containing only data between times 2 and 8 (see Figure 17.3).

This bootstrap resample inadvertently misses the concentrations measured early in time which define the curvature in the data. As a result, attempting to fit the bi-exponential model in Equation 17.1 fails. This failure is not the result of specific software limitations; the failure is a result of not having adequate data to capture the curvature expressed in the model.

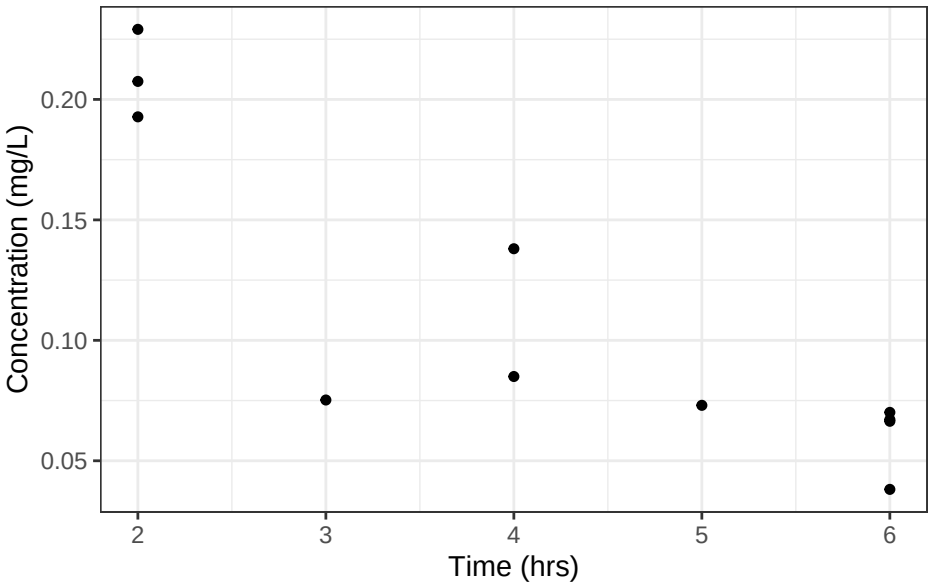


Figure 17.3: Potential bootstrap resample of the Indomethacin data resulting from case resampling. In this resample, the primary curvature present in the original data is lost.

So, while case-resampling avoids assuming constant variance, it can fail spectacularly in some nonlinear models. The residual bootstrap, on the other hand, assumes constant variance, but it maintains the curvature as the same predictor values are used when constructing each bootstrap resample. Pseudo-responses are generated in each bootstrap resample by adding a sample of the residuals to the fitted values ("jittering" the fitted line so to speak). The wild bootstrap is an alteration of the residual bootstrap which relaxes the assumption of constant variance while maintaining this beneficial property of the residual bootstrap.

Definition 17.3 (Wild Bootstrap)**.** Suppose we observe a sample of size n and use it to fit the mean model (linear or nonlinear)

$$E\left[(\text{Response})_i \mid (\text{Predictors})_i\right] = f\left((\text{Predictors})_i, \beta\right)$$

to obtain the ordinary least squares estimates $\widehat{\beta}$. The wild bootstrap proceeds along the following algorithm:

1. Compute the residuals

$$(\text{Residual})_i = (\text{Response})_i - f\left((\text{Predictors})_i, \widehat{\beta}\right)$$

2. Construct new pseudo-residuals $e_1^*, \ldots, e_n^*$ by multiplying each residual by a random variable U such that $E(U_i) = 0$ and $Var(U_i) = 1$, for example $U_i \sim N(0, 1)$:

$$e_i^* = U_i(\text{Residual})_i$$

3. Form "new" responses $y_1^*, \ldots, y_n^*$ according to

$$y_i^* = f\left((\text{Predictors})_i, \hat{\beta}\right) + e_i^*.$$

4. Obtain the least squares estimates $\hat{\alpha}$ by finding the values of α which minimize

$$\sum_{i=1}^{n} \left(y_i^* - f\left((\text{Predictors})_i, \alpha\right)\right)^2.$$

5. Repeat steps 2-4 m times.

We often take m to be large (at least 1000). After each pass through the algorithm, we retain the least squares estimates $\hat{\alpha}$ from the resample. The distribution of the estimates across the resamples is a good empirical model for the sampling distribution of the original least squares estimates.

The wild bootstrap alters the residuals (as opposed to resampling them as in the residual bootstrap) to mimic the variability of the response being potentially unique for each observation. The theoretical underpinnings are beyond the scope of this text, but intuitively, we are adding noise to the line ("jittering" the predicted model) such that the noise has mean zero (meaning the model is still correctly specified) and has variance of the same magnitude as the original observation (captured by the magnitude of the residual).

This process is computationally intensive but can dramatically improve inference when necessary. For the data from Example 17.1, Figure 17.4 compares the model for the sampling distribution of $\hat{\beta}_3$ assuming constant variance and after implementing a wild bootstrap. Assuming constant variance, our 95% CI for β_3 is (-0.03, 0.41); after implementing a wild bootstrap, we have a 95% CI for β_3 of (0.11, 0.40).

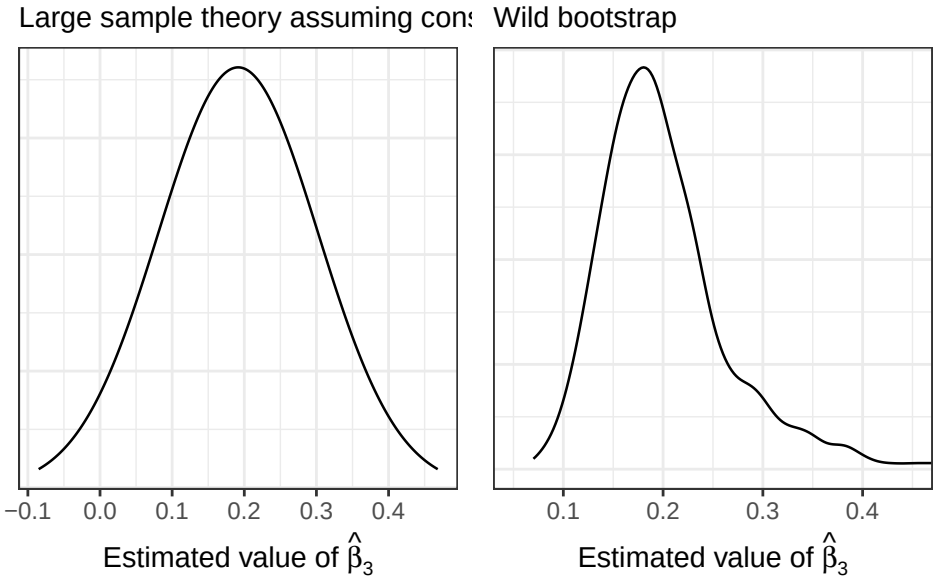


Figure 17.4: Comparison of two models for the sampling distribution of a parameter in the bi-exponential model fit to data examining the pharmacokinetics of Indomethacin. The left panel is the model for the sampling distribution based on large sample theory when we assume the variance is constant; the right panel is the model for the sampling distribution based on a wild bootstrap.

18 Logistic Regression

When the response is binary (taking one of only two values), the models we have discussed so far in this text are inappropriate. To see some of the complications, consider trying to characterize the impact of lifestyle choices on whether an individual is diagnosed with cancer. The response (is a person diagnosed with cancer, yes/no) does not readily fit into a framework such as

$$(\text{Response})_i = f\left((\text{Predictors})_i, \beta\right) + \varepsilon_i.$$

To begin with, the left-hand side is not a number, but a categorical variable. We could potentially address this using an indicator variable:

$$(\text{Cancer Diagnosis})_i = \begin{cases} 1 & \text{if i-th subject diagnosed with cancer} \\ 0 & \text{otherwise} \end{cases};$$

we could use this indicator variable as the response. However, we now have another issue; the right-hand side of the model would need to only return either a 0 or a 1. The response will *never* be 0.96; as a result, the idea of ε_i being errors that "jitter" the observed response from some overall mean response is no longer reasonable.

Recall that in developing our approach to nonlinear models, we specifically did not consider the "signal plus noise" approach and instead chose to specify a semiparametric model (Definition 16.2). This approach is general enough to permit binary responses. Specifically, when the response is binary, the most common technique is logistic regression, which is the focus of this chapter.

> **Note**
>
> Many other texts consider logistic regression to be a separate topic from nonlinear models; however, the structure of the logistic regression model is nonlinear in the parameters. We present it in this unit as a way of linking to ideas we have already discussed.

18.1 Considerations for a Binary Response

The benefit of the nonlinear model framework we outlined in Chapter 16 is that it emphasizes that regression models simply characterize the aspects of the distribution of the response we are confident about. When the response is binary, we actually know quite a bit about the distribution of the response.

Chapter 3 introduced common models for the distribution of a random variable. When the random variable is quantitative, there are many potential distributional models. However, when the response is binary, there is only a single model for characterizing the distribution: the Bernoulli distribution (Definition 3.6). Therefore, we can leverage that in developing a model for a binary response.

In particular, the Bernoulli distribution states that the mean response p represents the probability the response takes the value 1 (representing a "success"); and, p is constrained to be between 0 and 1. Further, the variance of the response is determined by the mean response: $p(1-p)$. Of course, the definition of the Bernoulli distribution considers the parameter p to be a single value. As we have seen, regression models allow the distribution to depend on additional predictors.

Placing this in the nonlinear model framework, we might say

$$\begin{aligned} E\left[(\text{Response})_i \mid (\text{Predictors})_i\right] &= f\left((\text{Predictors})_i, \beta\right) \\ Var\left[(\text{Response})_i \mid (\text{Predictors})_i\right] &= f\left((\text{Predictors})_i, \beta\right)\left(1 - f\left((\text{Predictors})_i, \beta\right)\right). \end{aligned}$$

The mean response function $f(\cdot)$ is allowing the mean response (p in the Bernoulli distribution) to depend upon the predictors. And, once the mean response function is specified, the variance is known (through the relationship $p(1-p)$). However, since we know that the response is binary, we can go further and say not only is this an appropriate mean and variance function, but we know the distribution as well; that is,

$$(\text{Response})_i \mid (\text{Predictors})_i \overset{\text{Ind}}{\sim} Ber\left[f\left((\text{Predictors})_i, \beta\right)\right],$$

where we are assuming that each response is independent of all others. Remember, nothing about the nonlinear modeling framework prohibited making distributional assumptions; we just often were unwilling to. Here, we know the distribution; so, we include it as part of the model.

Of course, the nature of the binary response impacts our choice of the mean response function $f(\cdot)$. In particular, the Bernoulli distributional model tells us that the mean response represents the probability the response takes the value 1. In our working example, this would be the probability of a subject receiving a cancer diagnosis given the value of the predictors. And, we

know that probabilities must be between 0 and 1. Therefore, we must choose a mean response function $f(\cdot)$ which has a range on the interval 0 to 1.

This last point is what prohibits using linear regression with a binary response. The mean response function should represent a probability, but it is entirely likely that linear regression will result in probabilities less than 0 or larger than 1. Therefore, any reasonable choice of $f(\cdot)$ will be nonlinear in the parameters.

While technically any function $f(\cdot)$ that has a range on 0 to 1 is possible, one choice has dominated the literature in applied sciences for many years.

18.2 The Logistic Regression Model

If we want a large class of functions $f(\cdot)$ that have a range on 0 to 1, we need only look to any cumulative distribution function (Definition 3.3). The most popular choice in practice is the cumulative distribution function of the Logistic distribution:

$$f(x) = \frac{e^x}{1+e^x}.$$

This leads to the logistic regression model.

Definition 18.1 (Logistic Regression Model)**.** A model for binary responses where the response, given the predictors, has a Bernoulli distribution such that

$$Pr\left((\text{Response})_i = 1 \mid (\text{Predictors})_i\right) = \frac{e^{\beta_0 + \sum_{j=1}^{p} \beta_j (\text{Predictor } j)_i}}{1 + e^{\beta_0 + \sum_{j=1}^{p} \beta_j (\text{Predictor } j)_i}} \tag{18.1}$$

and all responses are independent of one another.

Notice that we took the term in the exponents of Equation A.1 to be

$$\beta_0 + \sum_{j=1}^{p} \beta_j (\text{Predictor } j)_i,$$

a linear combination of the parameters. While not a requirement, this is common in practice since any additional curvature could be captured through flexible modeling techniques like splines. As a result, this term in the exponent is often referred to as the "linear predictor."

> **Warning**
>
> Do not be fooled by the linear combination of the parameters in the linear predictor. The logistic regression model is nonlinear in the parameters since the linear predictor appears in an exponent.

As stated above, by specifying the mean response function in Equation A.1, we have also specified the variance of the response. It will have the form

$$\left(\frac{e^{\beta_0+\sum_{j=1}^{p}\beta_j(\text{Predictor } j)_i}}{1+e^{\beta_0+\sum_{j=1}^{p}\beta_j(\text{Predictor } j)_i}}\right)\left(1-\frac{e^{\beta_0+\sum_{j=1}^{p}\beta_j(\text{Predictor } j)_i}}{1+e^{\beta_0+\sum_{j=1}^{p}\beta_j(\text{Predictor } j)_i}}\right).$$

We point this out to emphasize that the variance is not constant! Instead of addressing the non-constant variance through the wild bootstrap, we are modeling the structure of the variance directly.

> **Note**
>
> While not presented this way, it is possible to envision a binary response as the result of a latent (unobserved) quantitative response. For example, whether a student "graduates with honors" is a binary response (they either do or do not); but, it is the result of discretizing a quantitative measure (the student's GPA). Thinking of the observed binary response as a discretation of some *unobserved* quantitative measure, with proper assumptions on the error term, results in the logistic regression model.

18.3 Estimation of the Parameters

The logistic regression model not only specifies the form of the mean and variance of the response; it also specifies the distributional model. As a result, we could specify the density function of the response given the predictors. Proceeding by estimating the parameters using least squares (as advocated in Chapter 16) would actually ignore this additional information. When a parametric model is specified, we should take advantage of the additional structure (knowing the form of the density function) when estimating the parameters. This is accomplished through likelihood theory.

While a full development of likelihood theory is beyond the scope of this text, we motivate its use. In a probability course, the density function of a random variable is fully known, and we use it to compute the probability of the random variable taking on specific values. In a statistics course, we work in reverse. We have already observed specific outcomes; but, the density function is not fully known (as the parameters are unknown). We want to choose

values of the unknown parameters that would result in a density function making the observed data as likely as possible.

Consider a specific example. Suppose we *spin* a penny 100 times and observe it landing "tails-side up" in 82 of those trials. If you had to guess at the true probability of a penny landing "tails-side up" when spun, what would you guess based on this data? Putting it into our logistic model framework, consider the indicator

$$(\text{Tails})_i = \begin{cases} 1 & \text{if i-th spin lands tails-side up} \\ 0 & \text{otherwise;} \end{cases}$$

then, our logistic model (with no predictors) has the form

$$Pr\left((\text{Tails})_i = 1\right) = \frac{e^{\beta_0}}{1 + e^{\beta_0}}.$$

We want to choose a value of β_0 which makes it as likely as possible (maximizes the probability) that in a new sample of 100 spun pennies, 82 would land tails-side up (matching what we observed in the sample). That is, we want to choose the value of the parameter that maximizes the probability of seeing our observed data. Since the observed data is the only information we have about the process, we want a model that aligns with this data as closely as possible; or more accurately, we want the data to align with the model as closely as possible. Hopefully, it is intuitive that if *in reality* (where "reality" depends on the value of the unknown parameter β_0), a penny lands "tails-side up" 82% of the time, that makes this observed data much more likely than if *in reality*, it lands "tails-side up" only 50% of the time. Therefore, we would want the above probability to be equal to 0.82, leading to an estimate of β_0 of 1.516.

To help with visualizing this process, Figure 18.1 reports the probability of observing 82 coins (out of a sample of 100) land "tails-side up" as the value of β_0 changes. The likelihood is maximized when we set the true value of a "tails-side up" at being 0.82 (corresponding to $\beta_0 = 1.516$). Other values can make the data likely, but not as likely as that value.

We generalize this to saying that for a fully parametric nonlinear model (such as logistic regression), it is best to choose the values of the parameters that maximize the likelihood function.

Definition 18.2 (Likelihood Function)**.** For a fully parametric model, the likelihood function $\mathcal{L}(\beta, \text{Observed Data})$ captures how likely the observed data is to be realized in a future study under a specific set of parameters. This is directly related to the density function of the parametric model assumed.

Definition 18.3 (Maximum Likelihood Estimation)**.** The method of maximum likelihood estimation chooses parameter estimates to maximize the likelihood function under an assumed parametric model. The resulting estimates are known as maximum likelihood estimates.

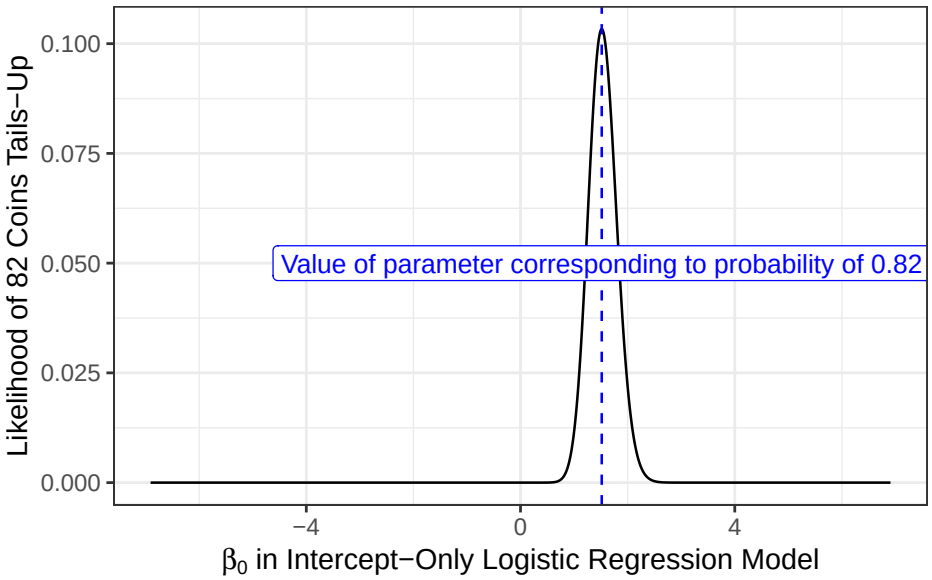


Figure 18.1: Likelihood of observing 82 coins land tails-side up when spinning 100 independent pennies. The likelihood is computed for various values of the parameter governing an intercept-only logistic regression model.

While the actual form is not critical to our exposition, for completeness, the likelihood function corresponding to logistic regression is

$$\prod_{i=1}^{n} \left(\frac{e^{\beta_0 + \sum_{j=1}^{p} \beta_j (\text{Predictor } j)_i}}{1 + e^{\beta_0 + \sum_{j=1}^{p} \beta_j (\text{Predictor } j)_i}} \right)^{(\text{Response})_i} \left(1 - \frac{e^{\beta_0 + \sum_{j=1}^{p} \beta_j (\text{Predictor } j)_i}}{1 + e^{\beta_0 + \sum_{j=1}^{p} \beta_j (\text{Predictor } j)_i}} \right)^{1-(\text{Response})_i}$$

where $(\text{Response})_i$ is an indicator value taking the value 1 or 0. Maximizing this likelihood is done numerically. While the details of this process are beyond the scope of this text, the procedure is similar to the least-squares procedure discussed in Chapter 20.

18.4 Inference on the Parameters

In order to make inference about the parameters, we need a model for the sampling distribution of the parameter estimates. Likelihood theory provides results for modeling the sampling distribution of maximum likelihood estimates. Generally, these results rely on large sample theory (though empirical models could be developed).

Definition 18.4 (Large Sample Sampling Distribution of Parameter Estimates in Logistic Regression)**.** Consider the logistic regression model in Definition 18.1. Assuming the form of the model is correctly specified with parameter vector β, as the sample size gets large, we have that

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{Var\left(\hat{\beta}_j\right)}} \sim N(0,1)$$

for all $j = 1, \ldots, p$. Further, under the null hypothesis

$$H_0 : \mathbf{K}\beta = \mathbf{m}$$

we have

$$\left(\mathbf{K}\hat{\beta} - \mathbf{m}\right)^\top \left(\mathbf{K}\widehat{\Sigma}\mathbf{K}^\top\right)^{-1} \left(\mathbf{K}\hat{\beta} - \mathbf{m}\right) \sim \chi^2_r$$

where r is the rank (number of rows) of $\mathbf{K}$.

> **Note**
>
> Though the above results require large sample sizes, generally fitting a logistic regression model itself requires a relatively large sample size (the less variability in the response, the harder it is to fit a model). As a result, being able to estimate the parameters often means the sample size is large enough to rely on the default inference.

As stated in Chapter 16, adding distributional assumptions does not avoid the need for large sample inference. The difference is primarily in the estimation process (least squares compared to maximum likelihood). When we do know the distributional model (as in the case of logistic regression), it turns out maximum likelihood estimation is optimal.

> **Warning**
>
> When we have a binary response, we *know* it has a Bernoulli distribution. As a result, we do not need to posit a model for the distribution. However, that does not guarantee our model is specified correctly in logistic regression because we may have misspecified the mean response function.

The above results allow us to not only construct confidence intervals, but we can also make use of the general linear hypothesis testing framework for testing specific hypotheses. That is, our inference is not all that different than under the linear model framework once we have estimates for the parameters and estimates for their standard errors.

18.5 Interpretation of Parameters

The parameters in a logistic regression model have a nice interpretation; however, that interpretation is not a natural interpretation for most individuals. In order to understand what is happening, we need to think in terms of odds of an event instead of probability of an event.

Definition 18.5 (Odds)**.** The odds of an event with probability p is defined as

$$\frac{p}{1-p}.$$

We often hear odds presented in terms of integers. Such as, "there are 3-to-1 odds the event will occur." This would mean that out of four trials, we would expect the event to happen 3 times; this corresponds to $p = 0.75$ and therefore $1 - p = 0.25$ giving an odds of 3. While many of us think in terms of probabilities, clinicians tend to think in terms of the odds of an event. As a result, it is natural to clinicians to compare the odds of an event under two scenarios instead of comparing the probability of an event under two scenarios.

Definition 18.6 (Odds Ratio)**.** The odds ratio is a method of comparing two events; typically, it is formed by the ratio of the odds of the same event under two different scenarios. Let p_1 be the probability of the event under scenario 1 and let p_2 be the probability of an event under scenario 2; then, the odds of the event under scenario 1 are

$$\gamma_1 = \frac{p_1}{1-p_1},$$

and the odds of the event under scenario 2 are

$$\gamma_2 = \frac{p_2}{1-p_2}.$$

The odds ratio comparing scenario 1 to scenario 2 is

$$OR = \frac{\gamma_1}{\gamma_2} = \left(\frac{p_1}{1-p_1}\right)\left(\frac{1-p_2}{p_2}\right).$$

If the odds of the event are the same under both scenarios, we obtain an odds ratio of 1. Odds ratios larger than 1 indicate that the event is more likely to occur (has greater odds) under scenario 1. Odds ratios less than 1 indicate that the event is less likely to occur (has lower odds) under scenario 1.

Warning

The odds ratio should not be confused with the relative risk. The relative risk of an event is the ratio of the probabilities under two scenarios. That is, under the setup of Definition 18.6, the relative risk comparing scenario 1 to scenario 2 is

$$\frac{p_1}{p_2}.$$

A relative risk of 4 says that the *probability* of the event is 4 times as large under scenario 1. An odds ratio of 4 says that the *odds* of an event are 4 times as large under scenario 1.

Now, let's return to our logistic regression model. Consider a model with two predictors:

$$Pr\left((\text{Response})_i = 1 \mid (\text{Predictors})_i\right) = \frac{\exp\left\{\beta_0 + \beta_1(\text{Predictor 1})_i + \beta_2(\text{Predictor 2})_i\right\}}{1 + \exp\left\{\beta_0 + \beta_1(\text{Predictor 1})_i + \beta_2(\text{Predictor 2})_i\right\}}.$$

Consider the group of subjects where Predictor 1 takes the value a and Predictor 2 takes the value b. Then, the probability the response takes the value 1 in this group is

$$p_a = \frac{e^{\beta_0+\beta_1 a+\beta_2 b}}{1 + e^{\beta_0+\beta_1 a+\beta_2 b}},$$

and the odds of the response taking the value 1 are

$$\frac{p_a}{1-p_a} = e^{\beta_0+\beta_1 a+\beta_2 b}.$$

Now, consider the group of subjects where Predictor 1 takes the value $a+1$ and Predictor 2 takes the value b. Then, following the above process, we have that the probability the response takes the value 1 in this group is

$$p_{a+1} = \frac{e^{\beta_0+\beta_1(a+1)+\beta_2 b}}{1 + e^{\beta_0+\beta_1(a+1)+\beta_2 b}},$$

and the odds of the response taking the value 1 are

$$\frac{p_{a+1}}{1-p_{a+1}} = e^{\beta_0+\beta_1(a+1)+\beta_2 b}.$$

Therefore, the odds ratio for the group with an increase in Predictor 1 relative to the other group is

$$\left(\frac{p_{a+1}}{1-p_{a+1}}\right)\left(\frac{1-p_a}{p_a}\right) = e^{\beta_1}.$$

That is, the parameters in the logistic regression model are directly related to the odds ratio.

Definition 18.7 (Interpretation of Parameters in a Logistic Regression Model)**.** Let β_j be the parameter associated with the j-th predictor in a logistic regression model (Definition 18.1). Then, β_j represents the log-OR ("log odds ratio") associated with a one-unit increase in the j-th predictor holding all other predictors fixed.

That is, exponentiating the j-th coefficient gives the odds ratio comparing the odds of the event under two scenarios: when the j-th predictor is increased by 1 unit relative to leaving it alone. Notice that unlike the linear model, increasing the j-th predictor by 1 unit does not result in an additive effect on the mean response (the probability of the response occurring in this case). Instead, it has an additive effect on the log odds.

> **Note**
>
> When we use the word "log" throughout the text, we are always referring to the natural logarithm.

If a parameter in our model is 0, it will result in an odds ratio of 1, indicating no association between the response and predictor. A parameter larger than 0 results in an odds ratio larger than 1, indicating that the likelihood of the response increases as the predictor increases. A parameter smaller than 0 results in an odds ratio less than 1, indicating the likelihood of the response decreases as the predictor increases.

19 Model Selection

As we discussed in Chapter 10, hypothesis testing is really a formal way of comparing two models — a simplified model and a more complex model. However, this form of model comparison requires the simple model to be a special case of the more complex model resulting from constraining the parameters in some way. This is often referred to as "nested testing." In some applications, we may be interested in comparing two competing (yet equally complex) models. In such cases, the general linear hypothesis testing framework does not work; as an alternative, we consider likelihood based model selection criteria.

There are several ways to quantify how well a model fits a set of data, or more accurately, how closely the data fits a particular model. The most well-known is R-squared.

Definition 19.1 (R-squared)**.** The proportion of the variability in the response explained by the model. When the response is quantitative, R-squared is defined as

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n} (\text{Residuals})_i^2}{\sum_{i=1}^{n} \left((\text{Response})_i - (\text{Overall Average Response}) \right)^2}.$$

When the response is categorical (such as logistic regression), there is no unique definition of R-squared.

R-squared has a nice interpretation and is often used to quantify the quality of the model fit. However, it is not a great criteria for comparing models. It can be proven that R-squared will continue to increase as the model becomes more complex. In fact, in many situations it is possible to add enough terms that you will obtain an R-squared value of 1 indicating perfect fit (visually, think about connecting the dots). However, this is the result of overfitting.

Definition 19.2 (Overfitting)**.** Overfitting refers to constructing a model that accurately predicts the observed data at the expense of accurate variance estimation and model parsimony. As a result, the model will have poor prediction for future observations.

Remember, models are never correct; by nature, they are a simplification of a complex process. We do, however, want our models to be useful. Creating a model that predicts well in the sample but performs poorly outside the sample is not helpful, and it could lead to incorrectly characterizing the scientific behavior being observed. As a result, we want a different metric for assessing model performance when comparing models.

> **Note**
>
> There are a couple "adjusted R-squared" metrics which do take into account model complexity. We do not discuss them in this text as they tend to be reserved for linear models and it is not obvious how to generalize them to other model settings.

As introduced in Chapter 18, when we define a fully parametric model, we can use likelihood theory to fit the model. In particular, the maximum likelihood estimates are the values of the parameters that maximize the likelihood function (Definition 18.2). Larger values of the likelihood (under an estimated model) indicate stronger agreement with the data. That is, larger values of the likelihood indicate a better fit. This suggests a general metric for assessing models regardless of their complexity; such metrics are known as information criteria.

Definition 19.3 (Information Criteria)**.** Information criteria refer to metrics that balance the model fit (through the likelihood function) with the model complexity through some penalty term.

Information criteria do not have an intuitive scale like R-squared; therefore, the values returned from these metrics have no natural interpretation. However, they do allow for comparisons between arbitrary models.

> **Warning**
>
> Information criteria rely on the likelihood; as a result, they only truly make sense when we are willing to assume the parametric model is appropriate. The computer can always compute them, but we should understand that they assume some distributional model for the response.

While there have been several information criteria proposed (and more continue to be developed), there are two that are very well known and are often defaults in software: AIC and BIC.

Definition 19.4 (AIC)**.** Given a model $\mathcal{M}$, Akaike's information criterion (AIC) is defined as

$$\text{AIC}_{\mathcal{M}} = -2\ln\left(\widehat{L}_{\mathcal{M}}\right) + 2p_{\mathcal{M}}$$

where $\widehat{L}_{\mathcal{M}}$ represents the likelihood function corresponding to model $\mathcal{M}$ when evaluated at the maximum likelihood estimates, and $p_{\mathcal{M}}$ is the number of parameters in $\mathcal{M}$.

Larger values of the likelihood function indicate better agreement; similarly, we prefer parsimonious models when possible. Therefore, *lower* values of AIC are preferred. AIC favors models that fit the data well (high likelihood) but penalizes for overly complex models (too

many parameters). Of course, this penalty is the same regardless of the sample size. This inspires BIC.

Definition 19.5 (BIC)**.** Given a model $\mathcal{M}$, Schwarz's Bayesian information criterion (BIC) is defined as

$$\text{BIC}_{\mathcal{M}} = -2\ln\left(\widehat{L}_{\mathcal{M}}\right) + p_{\mathcal{M}}\ln(n).$$

Note that the formula for BIC is nearly identical to that of AIC. As with AIC, lower values of BIC are preferred. With information criteria, as there is no natural scale, it only makes sense to compare models on the same set of data (you cannot compare the AIC from one study to the AIC of a different study). There is also no p-value for determining if the difference in AIC (or BIC) values for two models is "significant." Lower is simply better; however, some people do advocate various rules of thumb. For our purpose, if the values between two models are at all close, it should be a subject-matter decision: which model better explains the scientific process? Alternatively, if it is close, choose the model that is easier to interpret.

20 Estimation Details for Nonlinear Models

We have stated that numerical methods are needed to obtain estimates of the parameters for nonlinear models. In this chapter, we introduce the primary algorithm used for obtaining these estimates and sketch out some ideas for how this is implemented in practice. Readers may skip this chapter without loss of continuity.

When we begin working with nonlinear models, we will often come across computational issues. In order to begin diagnosing these issues, we must have some understanding of the underlying algorithm. Our objective is to find the values of the parameters β such that the mean function is close to the observed response. One way of defining "close" is to consider minimizing the sum of squared residuals. That is, we desire the values of β that minimize

$$\sum_{i=1}^{n} \left((\text{Response})_i - f\left((\text{Predictors})_i, \beta \right) \right)^2 . \tag{20.1}$$

We called the value that minimizes this objective function the ordinary least squares estimate.

> **Warning**
>
> This chapter applies to ordinary least squares. Logistic regression, which makes use of maximum likelihood estimation has a slightly different objective function:
>
> $$\sum_{i=1}^{n} (\text{Response})_i \log\left(\frac{e^{\beta_0 + \sum_{j=1}^{p} \beta_j (\text{Predictor } j)_i}}{1 + e^{\beta_0 + \sum_{j=1}^{p} \beta_j (\text{Predictor } j)_i}} \right) + \sum_{i=1}^{n} (1 - (\text{Response})_i) \log\left(1 - \frac{e^{\beta_0 + \sum_{j=1}^{p} \beta_j (\text{Predictor } j)_i}}{1 + e^{\beta_0 + \sum_{j=1}^{p} \beta_j (\text{Predictor } j)_i}} \right)$$
>
> where $(\text{Response})_i$ is an indicator function taking the value 1 when the response is a "success" and 0 otherwise.

For the remainder of this chapter, let y_i denote the response and $\mathbf{x}_i$ the vector of predictors for the i-th observation. Then, minimizing the objective function in Equation 20.1 is equivalent to finding the values of β that solve the system of equations given by

$$\mathbf{0} = \sum_{i=1}^{n} \left(y_i - f\left(\mathbf{x}_i, \beta \right) \right) f_\beta \left(\mathbf{x}_i, \beta \right) , \tag{20.2}$$

where f_β represents the gradient vector (with respect to the parameters) of the mean function f. Solving this system of equations is our primary objective. This is typically done via the Gauss-Newton Method.

Definition 20.1 (Gauss-Newton Method)**.** The Gauss-Newton method is the most commonly used optimization routine for statistical applications; this method relies on linearizing the problem using a Taylor Series approximation.

Observe that for a value β^* close to β, a first-order Taylor Series approximation gives

$$\begin{aligned} f\left(\mathbf{x}_i, \beta\right) &\approx f\left(\mathbf{x}_i, \beta^*\right) + f_\beta^\top\left(\mathbf{x}_i, \beta^*\right)\left(\beta - \beta^*\right) \\ f_\beta\left(\mathbf{x}_i, \beta\right) &\approx f_\beta\left(\mathbf{x}_i, \beta^*\right) + f_{\beta\beta}\left(\mathbf{x}_i, \beta^*\right)\left(\beta - \beta^*\right) \end{aligned}$$

where $f_{\beta\beta}$ is a p-by-p matrix of second partial derivatives. We appeal to a linear approximation because if β^* is sufficiently close to β, then higher order terms are negligible.

Now, we substitute these approximations into Equation 20.2, the system of equations we are solving. This results in four terms, of which the last is small because it has quadratic terms, and the third is small because on average, we expect $y_i - f(\mathbf{x}_i, \beta^*)$ to be small, which when multiplied by $(\beta - \beta^*)$ is really small. Thus, we are left with the approximation

$$\begin{aligned} \mathbf{0} &= \sum_{i=1}^{n}\left(y_i - f\left(\mathbf{x}_i, \beta\right)\right) f_\beta\left(\mathbf{x}_i, \beta\right) \\ &\approx \sum_{i=1}^{n}\left(y_i - f(\mathbf{x}_i, \beta^*)\right) f_\beta(\mathbf{x}_i, \beta^*) \\ &\quad - \sum_{i=1}^{n} f_\beta\left(\mathbf{x}_i, \beta^*\right) f_\beta^\top\left(\mathbf{x}_i, \beta^*\right)\left(\beta - \beta^*\right), \end{aligned}$$

which suggests

$$\sum_{i=1}^{n} f_\beta\left(\mathbf{x}_i, \beta^*\right) f_\beta^\top\left(\mathbf{x}_i, \beta^*\right)\left(\beta - \beta^*\right) \approx \sum_{i=1}^{n}\left(y_i - f(\mathbf{x}_i, \beta^*)\right) f_\beta(\mathbf{x}_i, \beta^*).$$

Solving this for β, we obtain

$$\beta = \beta^* + \left[\sum_{i=1}^{n} f_\beta\left(\mathbf{x}_i, \beta^*\right) f_\beta^\top\left(\mathbf{x}_i, \beta^*\right)\right]^{-1} \sum_{i=1}^{n}\left(y_i - f(\mathbf{x}_i, \beta^*)\right) f_\beta(\mathbf{x}_i, \beta^*).$$

That is, we have an iterative scheme for updating estimates of β given a current estimate. Given initial estimates $\beta^{(0)}$, we perform these iterative updates, terminating the algorithm when two successive iterations yield estimates which are close (iterations converged).

One of the key lessons here is that starting estimates of the parameters are always needed. This is always the case when numerical methods are used. There are no all-encompassing strategies here; determining starting values is unique to each problem. Often times, starting values can be obtained by making large simplifications or transformations to the model to get a rough idea of the estimates.

As an example, consider the Michaelis-Menten model for the mean response:

$$E\left((\text{Response})_i \mid (\text{Predictor})_i\right) = f\left((\text{Predictor})_i, \theta\right) = \frac{\theta_1(\text{Predictor})_i}{\theta_2 + (\text{Predictor})_i}.$$

In this model, θ_1 represents the maximum possible response and θ_2 the shape parameter known as the inverse affinity. Starting values for this model are obtained by considering a transformation. Specifically observe that if we assume the observed response is near the mean response given the corresponding value of the predictor, we have that

$$(\text{Response})_i \approx \frac{\theta_1(\text{Predictor})_i}{\theta_2 + (\text{Predictor})_i}.$$

Taking the inverse of the observed response suggests

$$\frac{1}{(\text{Response})_i} \approx \frac{\theta_2 + (\text{Predictor})_i}{\theta_1(\text{Predictor})_i} = \frac{\theta_2}{\theta_1}\left(\frac{1}{(\text{Predictor})_i}\right) + \frac{1}{\theta_1}.$$

This motivates considering a linear model of the form

$$(\text{Inverse Response})_i = \beta_0 + \beta_1(\text{Inverse Predictor})_i + \varepsilon_i.$$

The estimate of β_0 will allow us to compute an estimate for θ_1 where

$$\hat{\theta}_1 = \frac{1}{\hat{\beta}_0},$$

and the estimate of β_1 will allow us to compute an estimate for θ_2 where

$$\hat{\theta}_2 = \hat{\beta}_1\hat{\theta}_1.$$

We now have starting estimates that would allow us to resume fitting the nonlinear model on the original scale.

21 Nonlinear Models with Repeated Measures

Just as the linear model could be extended to account for repeated measures, nonlinear models can be extended as well. This is most commonly done through the mixed models framework (Chapter 14). In this chapter we provide some guidance for extending the nonlinear modeling framework for those interested. Readers may skip this chapter without loss of continuity.

Recall the Theophylline example (Example 16.1) we used to introduce nonlinear models at the beginning of this unit. Researchers were interested in studying the pharmacokinetics of the anti-asthmatic agent. In the original example, we considered blood samples taken from a single individual over a 24-hour period. In reality, the study enrolled 12 subjects, and each had blood samples taken over a 24-hour period (Figure 21.1).

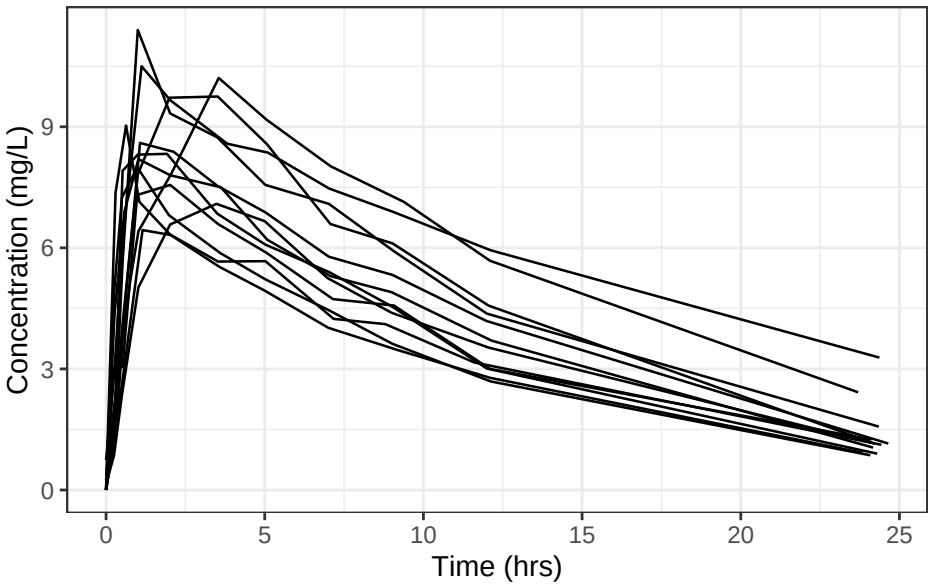


Figure 21.1: Concentration of Theophylline in the blood stream of a several patients over a 24-hour period.

The goal of the study, and the goal of most pharmacokinetics studies, is to understand the way the body processes the drug *across the population*. That is, researchers were interested in how the concentration-time profiles (through their parameters) varied across individuals.

Since we have multiple measurements for each of the 12 subjects, we have repeated measures data.

Notice that the goal here is not to model the average trend in the population but to characterize the average (and even the variability) of in the parameters across individuals in the population. This is not the same as a confidence interval for the parameter (which would still be describing the average value); researchers want to know how different the parameters can be across individuals. This is the perfect set-up for a subject-specific modeling approach to the repeated measures.

Recall that within a subject, researchers believe the one-compartment model with first-order absorption is an appropriate scientific model. Therefore, our individual-level model has the form

$$(\text{Concentration})_{i,j} = \frac{k_{a,i} D_i}{\left(\beta_i / k_{e,i}\right)\left(k_{a,i} - k_{e,i}\right)} \left(e^{-k_{e,i} t_{i,j}} - e^{-k_{a,i} t_{i,j}}\right) + \varepsilon_{i,j},$$

where the parameters retain the same interpretations they had previously. While the form is held the same across all subjects, the specific values of the parameters can vary from one subject to another. The population-level model *could* have the form

$$\begin{aligned} k_{a,i} &= k_a + b_{1,i} \\ k_{e,i} &= k_e + b_{2,i} \\ \beta_i &= \beta + b_{3,i}, \end{aligned}$$

where we assume each $b_{k,i} \overset{\text{IID}}{\sim} N\left(0, \sigma_k^2\right)$ for each k and assume all terms are independent. That is, we are allowing each parameter to have a random effect that allows them to vary across the population. The parameter k_a, for example, represents the average absorption rate across the population and σ_1^2 would capture the variability of the absorption rate across the population.

Fitting this model to the data illustrated in Figure 21.1, we estimate that the average clearance rate in the population is 0.04. A confidence interval for the clearance rate is not of interest, as it would still be estimating the *average* clearance rate in the population. Based on the figure, the clearance rate differs across individuals; specifically, we estimate the clearance rate may vary between 0.02 and 0.06 among 95% of the population.

We could generalize the population-level model further to allow these parameters to depend upon other predictors (such as the weight of the subject).

These models are commonly fit (using specialized statistical software), but they continue to be the focus of research as there is not complete agreement on modeling constraints and how to relax conditions placed on the model. Unlike the linear case, it is not easy to determine the

overall average trend given the subject-level model (though it is theoretically feasible). The nonlinearity makes several aspects of this problem quite challenging.

Part VI

Unit VI: Survival Analysis

Survival analysis (often referred to as "reliability analysis" in engineering) refers to the statistical analysis of studies for which the response is the time until an event occurs, such as the time until a device fails or the time until a subject dies. Studies involving such an endpoint often involve censoring — the event of interest is not observed for every subject during the study period. In such situations, we must develop special models in order to obtain proper inference. In this unit, we discuss the challenges associated with censoring and how we address these challenges when performing inference and when developing models.

22 The Language of Survival Analysis

Models for survival analysis, in contrary to other models discussed in this text, are not generally concerned with the average response. Instead, they rely on different characterizations of the distribution of the response. While Chapter 3 introduced basic methods for characterizing the distribution of a random variable, we need to extend these ideas in order to discuss survival analysis.

Let T represent the time until an event occurs; as we cannot predict this time with certainty, T is a random variable. Since it is a time, it could take on any value larger than 0; and, its distribution can be characterized through the probability density function $f(t)$ (Definition 3.2), which links the values it can assume with the corresponding likelihood they occur.

Example 22.1 (Carcinogen Exposure in Rats)**.** Consider a study in which rats were exposed to a carcinogen and then monitored closely. Suppose the distribution of the time T (in days) between exposure and the development of a tumor can be modeled with the following probability density function:

$$f(t) = \frac{1}{10}e^{-t/10} \qquad t > 0.$$

Figure 22.1 plots the density function from Example 22.1. We see that smaller times are more likely (since they correspond to higher values of the density function) than larger times. That is, we would expect a tumor to develop relatively quickly in these rats.

Survival analysis concerns itself primarily with the likelihood of remaining "event-free." For example, for the Carcinogen example, we might be interested in the proportion of rats that are still tumor free through 7 days; this translates to computing

$$Pr(T > 7) = \int_7^\infty f(t)dt.$$

Notice that this is the complement of the cumulative distribution function (Definition 3.3); this is known as the survival function.

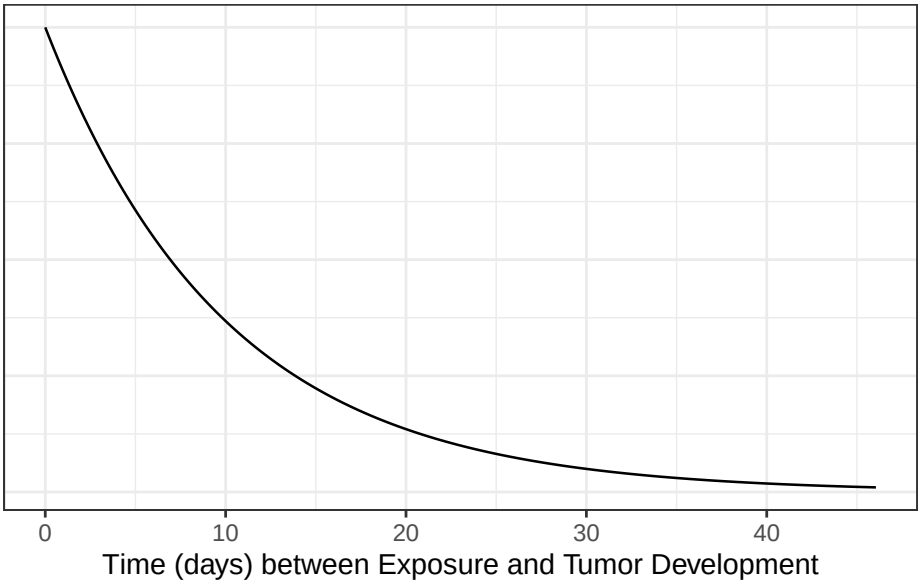


Figure 22.1: Density function from a hypothetical study investigating the time until a tumor develops in rats exposed to a carcinogen.

Definition 22.1 (Survival Function)**.** Let T be a random variable; the survival function $S(u)$ is defined as

$$S(t) = Pr(T > t) = 1 - F(t),$$

capturing the probability of failing *after* a time, where $F(t)$ is the CDF of T.

For a continuous random variable, we have that

$$S(t) = \int_t^\infty f(u)du$$

implying that $f(t) = -\frac{d}{dt}S(t)$.

It should be clear from the definition that survival functions range between 0 and 1, must have $S(0) = 1$, and diminish towards 0 as time increases. In our example, this translates to rats being alive at the start of the study and all rats will develop a tumor at some point after exposure.

In addition to the likelihood of being event-free past some point, we may want to know the likelihood of experiencing the event in the next unit of time *given* the subject has been event-free up to that point. For example, what proportion of rats in the Carcinogen example will

develop a tumor within the next day *given* the rats were tumor-free through seven days? The additional information (following the word "given") changes the computation; we are not interested in $Pr(7 < T \leq 8)$ because the additional information says we are only interested in the subset of rats which are tumor-free through 7 days. We express this as $Pr(T \leq 8 \mid T > 7)$. Since we are only interested in a subset of rats, we need to rescale (see Figure 22.2); this results in

$$Pr(T \leq 8 \mid T > 7) = \frac{Pr(7 < T \leq 8)}{Pr(T > 7)} = 1 - \frac{S(8)}{S(7)}.$$

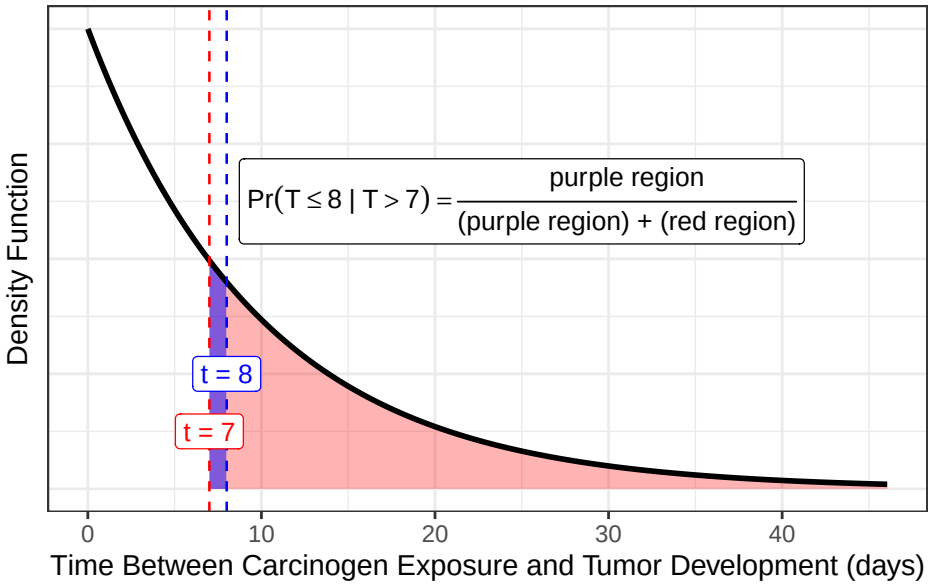


Figure 22.2: Illustration of conditional probability.

This captures the event rate in the next time interval among survivors; this is known as the mortality rate.

Definition 22.2 (Mortality Rate)**.** For any particular time t, the morality rate $m(t)$ is the proportion of the population that experiences the event between times t and $t+1$, among individuals that are event-free at time t. That is,

$$m(t) = Pr(t \leq T \leq t+1 \mid T > t) = 1 - \frac{S(t+1)}{S(t)}.$$

The mortality rate considers a step of a single unit of time. Imagine taking a step of size h, considering the proportion of deaths *per unit time* with this step size, and then allowing

this step size to become very small (mathematically, taking a limit). This would describe the instantaneous mortality rate, known as the hazard.

Definition 22.3 (Hazard Function)**.** For any time t, the hazard function $\lambda(t)$ is the instantaneous mortality rate per unit time:

$$\lambda(t) = \lim_{h \to 0} \frac{Pr(t \leq T \leq t + h \mid T > t)}{h} = \frac{f(t)}{S(t)} = -\frac{\frac{d}{dt}S(t)}{S(t)} = -\frac{d}{dt}\log(S(t)).$$

> **Warning**
>
> While the hazard function is related to probabilities, it is *not* a probability. As a result, it can take on values much larger than 1.

Larger values of the hazard function indicate a higher likelihood of the event. Given that each of these quantities (survival, mortality, hazard) is related to the density function, they are not technically necessary. However, as we will see, it turns out that it is sometimes easier to model on the hazard scale than modeling the density function. What is important to see is that if we are able to characterize the hazard function, we have also characterized the survival function and therefore the entire distribution.

The last component of the above definition emphasizes the relationship between the survival function and the hazard function. Inverting this relationship provides another quantity often referenced in survival analysis.

Definition 22.4 (Cumulative Hazard)**.** For any time t, we have that

$$S(t) = e^{-\int_0^t \lambda(u)du}$$

where

$$\Lambda(t) = -\int_0^t \lambda(u)du$$

is known as the cumulative hazard function.

There are many named probability models used to characterize the distribution of an event time (Exponential, Gamma, Weibull, etc.). There are a few commonalities; each is defined over non-negative values as they are modeling the time until an event. Second, each tends to be right-skewed (the distribution has a "slide" shape with a long tail toward higher values). This is why it is often more common to report the median survival time instead of the average

survival time, as the average may be a misleading estimate of the center of the distribution. Such parametric models are quite common in engineering disciplines; however, among the biological sciences, it is more common to adopt a semiparametric approach.

23 Censoring

In an ideal world, survival analysis would not differ from modeling any other quantitative response (in this case, the time to an event). The problem is that we often do not observe the response on a subject prior to the end of the study but instead only have partial information on the response; this phenomena requires additional considerations. Consider the following example.

Example 23.1 (Hypertension)**.** A study was conducted to examine the efficacy of a new anti-hypertensive medication. A cohort of 146 patients with a previous history of heart disease were treated and then followed over the next 10 years. The primary event of interest was death, which was grouped into one year intervals.

As you might imagine, following patients over the span of years can be challenging. Occasionally, subjects are "lost to follow-up." This can happen for a number of reasons. It could be the subjects withdrew their consent for the study, prohibiting researchers from further contact. As a result, from the time the subject withdrawals from the study, we are no longer able to ascertain whether the subject has experienced the event; all we know is that they were event-free up until the time they withdrew consent. Other subjects might die during the study; however, their cause of death was not the result of hypertension but something unrelated (such as a car accident). Again, we do not get to see when they would have died from the underlying condition but instead only know they were event-free up until they died from these external circumstances. Finally, some patients may remain event-free throughout the entire study; and, at the end of the study are still event-free. We do not get to see when these patients would experience the event; we only know it is after the study. The central question in survival analysis is how to handle such partial information. Excluding the patients from the analysis throws out valuable information and can bias the results. But, pretending to know their event-time adds a certainty not present in the data and will also bias the results.

There are two important issues in biological studies where the primary endpoint of interest is the time to an event:

- The event has not occurred at the time of analysis, resulting in only partial information on the response.
- The length of follow-up varies due to staggered entry into the study.

The second issue listed is not as much of a concern in controlled lab experiments where all subjects can be studied over the same interval of time. However, in larger clinical trials, it takes time to enroll patients, and as a result, they enter the study at different points. Since studies are often funded for a fixed duration, the result is that each patient could potentially be followed for a different length of time. In order to address this, we often "start the clock" when a subject enters the study and think in terms of patient time (instead of time since birth, for example). Since entry into the study often involves a treatment at some point, it can be helpful to use this as a reference point.

> **Note**
>
> The time at which a patient enters a study is often referred to as "baseline;" in survival analysis, this is also generally the point we think of as time $t = 0$.

The first issue is a larger concern known as censoring.

Definition 23.1 (Censored Data)**.** Censored data is a special case of missing data for which a bound on the missing value is known. In survival analysis, the response of interest (time to an event) is subject to censoring.

> **Warning**
>
> Censoring is not the same as "missing data" in the literature. With censoring, we have partial information regarding the value that is unobserved.

It is often impossible or impractical to observe the time to an event on all subjects. There are three primary causes for censoring, which we illustrated in the above discussion of Example 23.1:

- End of the study; common in clinical trials, resulting from both staggered entry into the study and financial constraints preventing indefinite follow-up. At the end of the study, all subjects who have not experienced the event are censored.
- "Lost to follow-up," which is common in studies involving human subjects. Living subjects, especially humans, do not always behave as you expect. As an example, humans can change their mind about participating in a study and withdraw consent.
- Competing risks. This is an outcome that prevents observing the event of interest, which is common with ill patients. As an example, a cancer patient may die preventing our ability to observe if they experience a heart attack (the endpoint of interest).

While many scenarios may result in censored data, we can construct a taxonomy of types of censoring (Figure 23.1).

In order to illustrate the various types of censoring, we consider variations on the design of the Carcinogen example from the previous chapter (Example 22.1). First, consider a study design

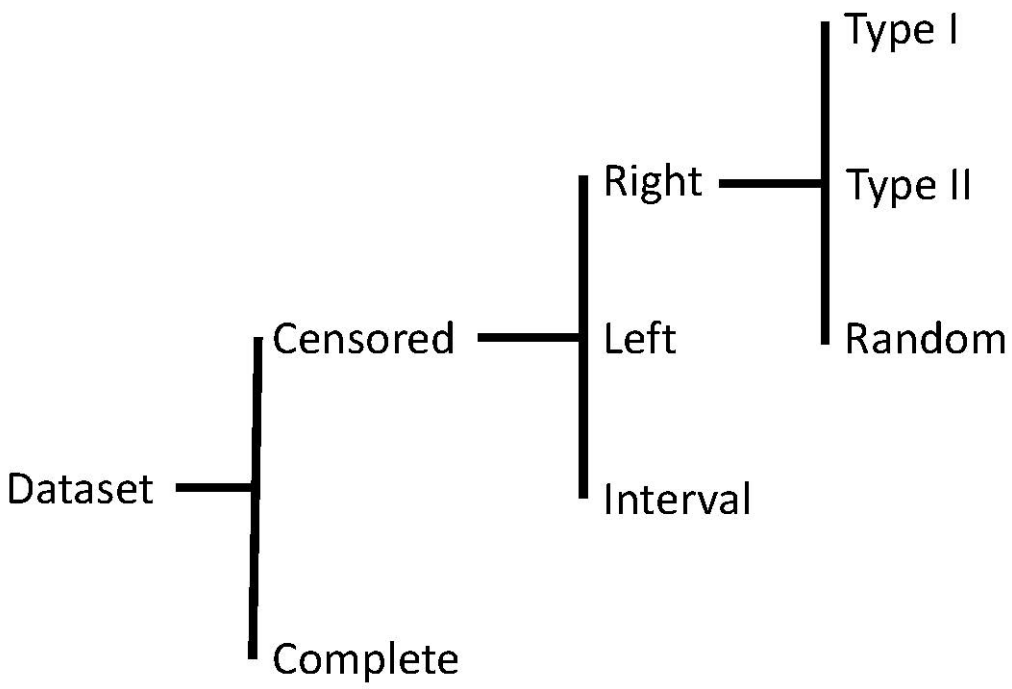


Figure 23.1: Taxonomy of various types of censoring.

in which a sample of n rats are exposed at the same time point. The rats are followed for a period of two weeks, at which point the study is ended. This design emphasizes a situation in which the duration of the study is the primary constraint. If all rats develop a tumor during this two-week period, then the data set is complete.

Definition 23.2 (Complete Data)**.** This term describes a data set for which the event is observed on all subjects.

When the data is complete, survival analysis reduces to a standard estimation problem that could be addressed using methods previously described in this text.

Suppose, however, that some rats remain tumor-free at the end of the studied (Figure 23.2). These rats are censored. Since all rats began the study at the same point in time, any observation censored is censored at the same time point. In this case, we only know that the time until the rat develops the tumor is at least two weeks.

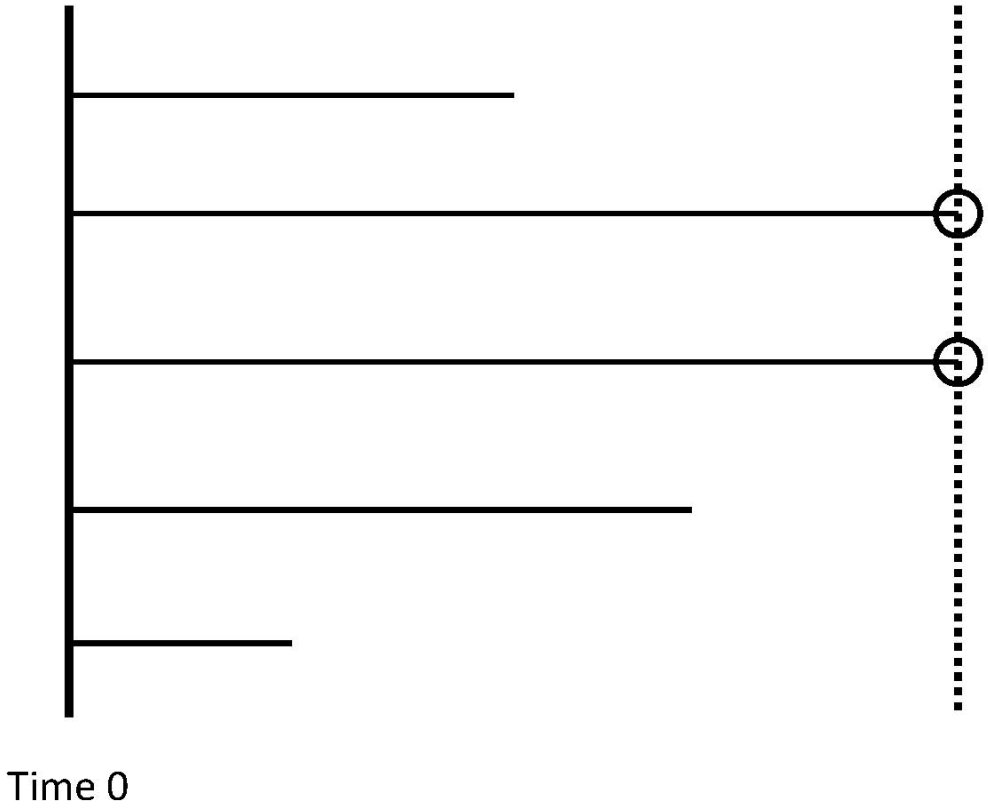


Figure 23.2: Illustration of Type-I, right censoring.

Definition 23.3 (Right Censoring)**.** Right censoring refers to scenarios when a *lower* bound is known on the response.

Definition 23.4 (Type I Censoring)**.** Type-I censoring is a form of right censoring where the only source of censoring is the end of the study, for which the duration was pre-determined. Therefore, the time at which subjects are censored is the pre-determined study duration.

In this study, the number of rats that remain tumor-free is random, and out of the control of the researchers. We contrast that with a design that says that each of the n rats is exposed at the same time and the study continues until the first r rats develop a tumor (where r is chosen ahead of time). In this case, the length of the study is random. This study design might be chosen when running the study itself is cheap once the subjects are obtained; it ensures an adequate number of events to power the study. All rats remaining tumor-free past the r-th rat will be censored (Figure 23.3).

Definition 23.5 (Type II Censoring)**.** Type-II censoring is a form of right censoring where the only source of censoring is the end of the study, which is determined when the r-th event occurs and r is pre-determined. Therefore, the time at which subjects are censored is determined by the r-th event.

Type I and Type II censoring tend to occur in controlled settings; in large-scale clinical trials involving human subjects, the source of censoring cannot be controlled. For example, suppose the particular breed of rat being studied is difficult to obtain. As a result, we are unable to obtain the n rats at the same time. Instead, each rat is obtained as it becomes available; the rat is then exposed to the carcinogen and followed for as long as possible (until it develops a tumor or the study ends). The study will end two months after obtaining the first rat. This design results in staggered entry into the study, which as described above can be accounted for by thinking of "time" as "time since exposure" instead of "time since study began." Many rats will experience the event. Others will not experience the event at the time the study ends. However, the length of time between entering the study and the study ending differs across those rats which are censored (as some of those rats entered the study earlier than others). Further, it is possible some rats are censored prior to the end of the study because they experience organ failure and die prior to the development of a tumor. The key observation is that the censoring times may differ for each rat (see Figure 23.4).

Definition 23.6 (Random Censoring)**.** Random censoring is a form of right censoring when subjects are withdrawn from the study at any time. It is typically assumed that the event time and the censoring time are independent of one another.

The assumption of independence between the censoring and event times is important. If only the healthiest patients are censored, then it is difficult to estimate the survival probability correctly; we assume the reason each patient is censored has nothing to do with their underlying

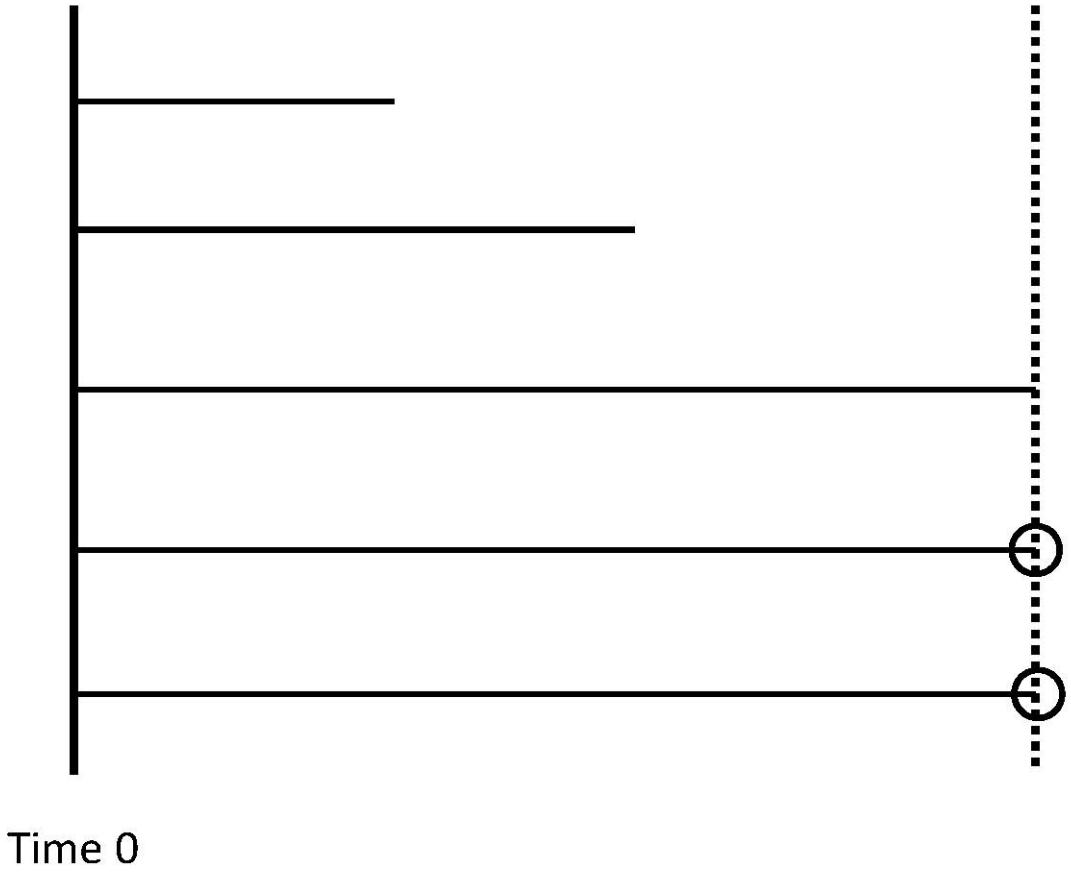


Figure 23.3: Illustration of Type-II, right censoring.

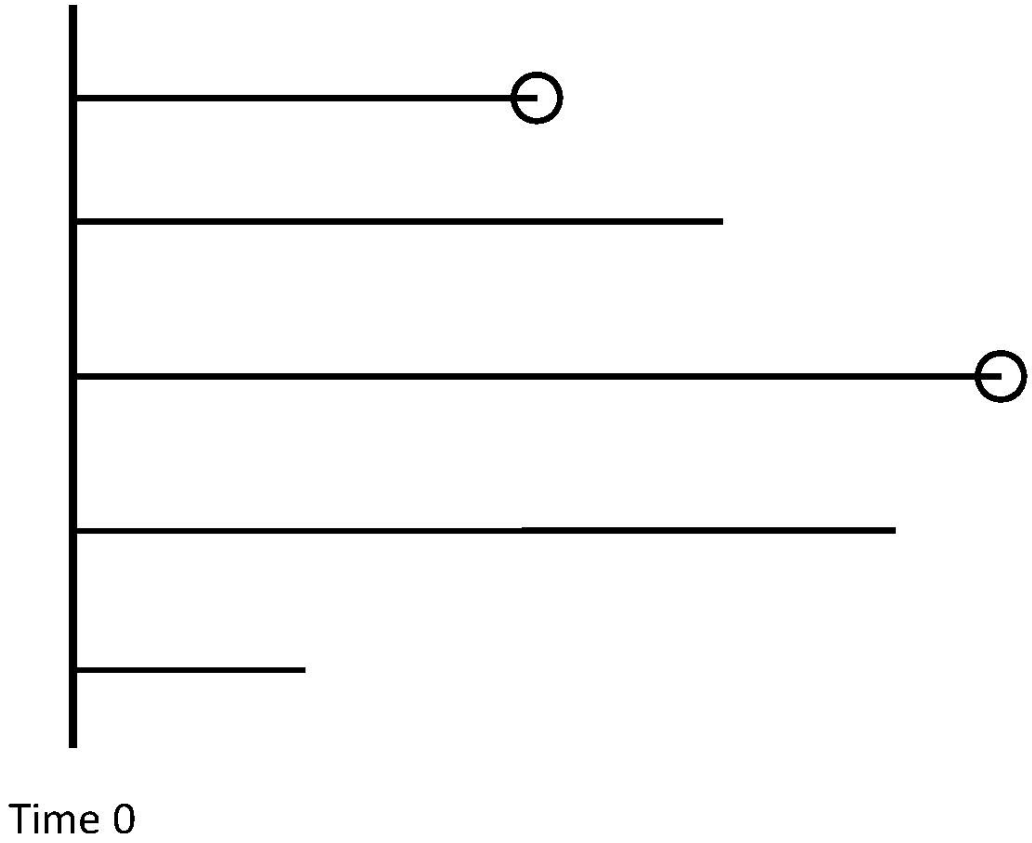


Figure 23.4: Illustration of random right censoring.

survival. This is fundamentally different than Type I or Type II censoring where the healthiest subjects are those that are censored. By far, the most common type of censoring is random right censoring.

In each of the above cases, a lower bound was known on the event time, but there are other possibilities. Suppose the sample of n rats were exposed to the carcinogen at the same moment; however, this occurred at an off-site location. The rats are then transported to the lab, which takes several days. Upon arrival, it is discovered that m rats have already developed a tumor. For these rats, an upper bound is known on the event time.

Definition 23.7 (Left Censoring)**.** Left censoring refers to scenarios when an *upper* bound is known on the response.

Finally, we consider the case in which the n rats are exposed to a carcinogen; the rats are assessed once weekly to determine if a tumor has developed. At the end of each week, the number who have developed a tumor is noted (Figure 23.5). As a result, we do not know the exact day on which the tumor developed; we only know it occurred sometime within the last week. This creates both an upper and lower bound on the event time.

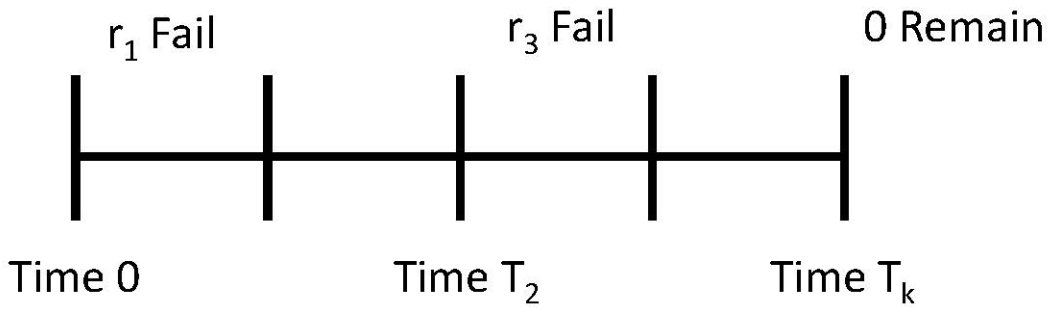


Figure 23.5: Illustration of interval censoring.

Definition 23.8 (Interval Censoring)**.** Interval censoring refers to scenarios when the event time is known to have occurred within some interval, but the exact time is unknown.

Interval censoring results when only periodic assessment can be performed. Of course, it is impossible to measure an event time with infinite precision; therefore, it would seem all events are interval censored. In practice, we consider interval censoring to occur when the interval is larger than the smallest unit of time we would like to consider. For example, suppose for Example 22.1, we are interested in the number of days until the rat develops a tumor. In that case, daily assessments on the rats, while periodic, would not result in interval censoring;

however, weekly assessments would. In contrast, if we were interested in the number of hours until the rat develops a tumor, daily assessments would result in interval censoring.

Regardless of the type of censoring, it cannot be ignored. Due to censoring, we do not observe the survival time on all subjects, and this impacts the likelihood (Definition 18.2). Consider the case of right censoring; in that case, we observe the smaller of the survival time and the censoring time on each subject. We also are able to note whether our observation was the survival time or the censoring time. These together give us information about the response.

Definition 23.9 (Event Time and Censoring Indicator)**.** Let T_i and C_i represent the survival time and the censoring time for the i-th subject. When data is subject to right censoring, we observe

$$X_i = \min\{T_i, C_i\}$$

which is known as the *event* time. We also observe whether this observation was triggered by the actual event or censoring:

$$\delta_i = \begin{cases} 1 & \text{if } T_i \leq C_i \\ 0 & \text{if } T_i > C_i. \end{cases}$$

This is known as the *censoring indicator*.

Warning

Note that counter-intuitively, the "censoring indicator" actually indicates when the survival time is observed, not when an event is censored.

Big Idea

Any analysis of survival data must account for the survival times and the censoring times; however, neither is directly observed on all subjects. As a result, standard methods are not applicable.

24 Basic Estimation and Inference

There are some studies where we are interested in characterizing the overall survival; other studies seek to compare the survival across a small number of groups. In these situations, when the relation between survival and multiple characteristics of the subject is *not* of primary interest, modeling is not necessary. The methods discussed in this chapter are analogous to computing the sample mean, constructing a boxplot, or conducting a "one-sample t-test" in an introductory class. However, having a grasp of these methods can help us better understand the benefits of modeling that we discuss in the next chapter.

As we have discussed in other settings, there are essentially three approaches to estimation: parametric, semiparametric, and nonparametric. This is true with time-to-event data as well. If we are willing to posit a model for the distribution of the survival times and the censoring times, we could potentially use likelihood methods to estimate the unknown parameters. While more popular among engineering disciplines, the biological sciences prefer semiparametric and nonparametic approaches. The methods discussed in this chapter are nonparametric in nature, while the model discussed in the next chapter is semiparametric.

24.1 Life-Table Methods

In some cohort studies, where a large group is followed over the course of time with periodic "check-ins," the exact event times are known to fall with key intervals. The origins of survival analysis are steeped in such studies. In such cases, life-table methods are used to characterize survival.

Definition 24.1 (Life Table)**.** Life tables are a method of estimating overall survival over key intervals of time, generally constructed for a single population.

In order to illustrate the considerations in life-table methods, consider the following example.

Example 24.1 (Hypertension, Revisited)**.** A study was conducted to examine the efficacy of a new anti-hypertensive medication. A cohort of 146 patients with a previous history of heart disease were treated and then followed over the next 10 years. The primary event of interest was death. A summary of the data is provided in Table 24.1.

Table 24.1: Summary of event times for a study investigating a new medication for hypertension.

Years from Baseline	Number at Risk	Number of Deaths	Number Censored
1	146	27	3
2	116	18	10
3	88	21	10
4	57	9	3
5	45	1	3
6	41	2	11
7	28	3	5
8	20	1	8
9	11	2	1
10	8	2	6

Let's consider how we should estimate the survival at time $t = 5$; that is, what proportion of individuals survive past 5 years in the study? Notice that we observe 76 deaths during the first 5 years of follow-up; therefore, an initial guess might be

$$S(5) = 1 - \frac{\text{76 deaths over 5 years}}{\text{146 individuals}} = 0.479.$$

However, this assumes that every subject censored during the study lived the full five years. Remember, we do not know why the 3 subjects who were censored during the first year were lost to follow-up. All we know is that they were alive at the beginning of the study; we do not know if they died during the first year or survived through the end of the study. Assuming that the 29 censored individuals all survived through 5 years is quite optimistic, meaning that our estimate of survival is biased on the high-sided.

In order to correct this optimistic perspective, we might consider removing the censored subjects:

$$S(5) = 1 - \frac{\text{76 deaths over 5 years}}{\text{146 individuals} - \text{29 withdrawn}} = 0.350.$$

This essentially assumes the censored individuals never existed; however, we know that 3 of the subjects, for example, survived at least through the first 4 years. Excluding this information results in a pessimistic estimate, meaning our estimate of survival is biased on the low-side.

While these two extremes were illustrated over a five-year span, the same concerns exist on any single interval. For example, assuming the 3 subjects censored during the first year all survived the first year is optimistic; assuming they never existed is pessimistic. Life-table computations

balance these two extremes. Our critical assumption is that the censoring occurs uniformly over each interval.

Consider the *mortality* rate within the first interval, assuming subjects censored are done so uniformly over that first year. We can think of this as saying half the subjects who were censored should be removed from the study as if they never existed in that interval, and the other half survived to the end of the interval:

$$\widehat{m}(1) = \frac{27}{146 - 1.5} = 0.187.$$

This is an estimate of the probability a patient in this cohort dies during the first year. The survival probability during this interval is the probability of *not* dying during the first year:

$$\widehat{S}(1) = 1 - \widehat{m}(1) = 0.813.$$

We can apply this adjustment on each interval. That is, the mortality within an interval is computed by considering the number of deaths over that interval as a fraction of the number of people at risk who entered that interval and assuming those censored exited uniformly over the interval. To survive to the end of any interval, you must have survived all previous intervals. This is captured in the following life-table computations.

Life Table Computations

Let d_t represent the number of subjects that experience the event during the t-th interval. Let w_t represent the number of subjects censored during the t-th interval. And, let n_t represent the number of subjects at risk at the start of the t-th interval.
Assuming censoring occurs uniformly over the interval, the estimated *mortality* for the t-th interval is defined as

$$\widehat{m}(t) = \frac{d_t}{n_t - \frac{w_t}{2}}.$$

The estimated survival at the end of the t-th interval is given by

$$\widehat{S}(t) = \prod_{k=1}^{t} \left[1 - \widehat{m}(k)\right].$$

We emphasize that we are *estimating* survival; the true survival is unknown. Of course, $\widehat{S}(0) = 1$ since we begin with living subjects. Computing the survival as a product of interval-specific survival can be derived through a series of conditional probability statements; however, it captures the idea that a subject must survive each interval in turn. By looking one interval at a time, we reduce the bias in our estimate of survival. Employing these computations for Example 24.1 results in an estimated 5-year survival of $\widehat{S}(5) = 0.417$.

Of course, a point estimate does not allow us to make inference on the true survival. In order to perform inference, we need to model the sampling distribution of our estimate.

Definition 24.2 (Model for the Sampling Distribution of Life-Table Estimates)**.** Consider estimating the survival $S(t)$ at time t using life-table estimate $\widehat{S}(t)$. As the sample size increases, we have that

$$\frac{\widehat{S}(t) - S(t)}{\hat{\sigma}^2} \sim N(0,1),$$

where

$$\hat{\sigma} = \widehat{S}(t)\sqrt{\sum_{i=1}^{t} \frac{d_t}{(n_t - w_t/2)\,(n_t - w_t/2 - d_t)}}$$

and the estimate $\widehat{S}(0) = 1$ has no error.

Given the model for the sampling distribution, it is possible to construct confidence intervals. However, using the classical

$$\widehat{S}(t) \pm (1.96)\hat{\sigma}$$

to compute a 95% confidence interval can result in bounds that extend beyond 0 or 1. Since $S(t)$ is a probability, such bounds are unreasonable. One approach is to simply truncate the confidence limits at 0 or 1. Other approaches require computing the confidence interval on the log-scale and then transforming back to the original scale of interest.

We note that this model for the sampling distribution (and the resulting confidence intervals) are developed point-wise. That is, this is different than a confidence band meant to encompass the entire curve. For those unfamiliar with statistical theory, it is sufficient to keep in mind that the bounds were generated for a specific point in time t.

Life-table methods are somewhat limited in their use as they apply to a single cohort and the survival times are grouped within intervals. However, the ideas discussed here are important for generalizing to other settings, as we discuss in the next section.

24.2 Kaplan-Meier Estimation

The life-table estimation approach helps to highlight the key considerations when working with time-to-event data. To remove bias, we must carefully consider how the censored data is addressed. While life-table methods are not always applicable, we can generalize the results by considering how we would apply life-table methods with smaller and smaller intervals of time. Intuitively, the smaller we can make the interval, the less bias our estimates will have. To that end, we shrink the interval under consideration until it includes only a single event time; then, we apply the life-table approach. This is known as the Kaplan-Meier estimate.

Definition 24.3 (Kaplan-Meier Estimator)**.** Also known as the product-limit estimator, the Kaplan-Meier estimator is the limit of the life-table estimate as we allow the intervals to shrink to a single event time:

$$\widehat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_{t_i}}{n_{t_i}}\right)$$

where t_i is the i-th survival time where the event of interest was observed.

There are a couple of things to note about this definition. Keep in mind that shrinking an interval to have no width is relative to what we consider the unit of time; if we are measuring survival time in days, then an interval only large enough to include a single event time would capture a single day. If survival time is measured in weeks, the interval would shrink to capture a single week. In the Kaplan-Meier estimator, the product is taken over all observed survival times prior to t. We emphasize that this does *not* include censoring times.

> ⚠ Warning
>
> The product-limit estimator does not change when an individual is censored. It is only updated when an individual experiences the event of interest.

How then is censoring accounted for if the estimator only updated when an event is observed? Each time an event does occur and we update the estimator, we are adjusting the number of subjects at risk at that event time. When determining how many subjects are at risk, we do not consider previously censored subjects. This is similar to how the life-table estimate subtracted out the number of risk within each interval. Except, since our "interval" is a single event time, we do not need to think about those subjects being uniformly distributed over the interval; they all share the same time.

Kaplan-Meier curves (survival curves estimated using the product-limit estimator) are step-functions, with a step being taken at each time an individual subject experiences the event. Figure 24.1 gives an illustration of what these curves look like. Notice that if the last subjects

in the sample are censored (which occurred in Group A of Figure 24.1), then the survival probability will not drop to 0. In large scale-studies, it is not uncommon for the study to end with several patients remaining, meaning the survival curves stay relatively high.

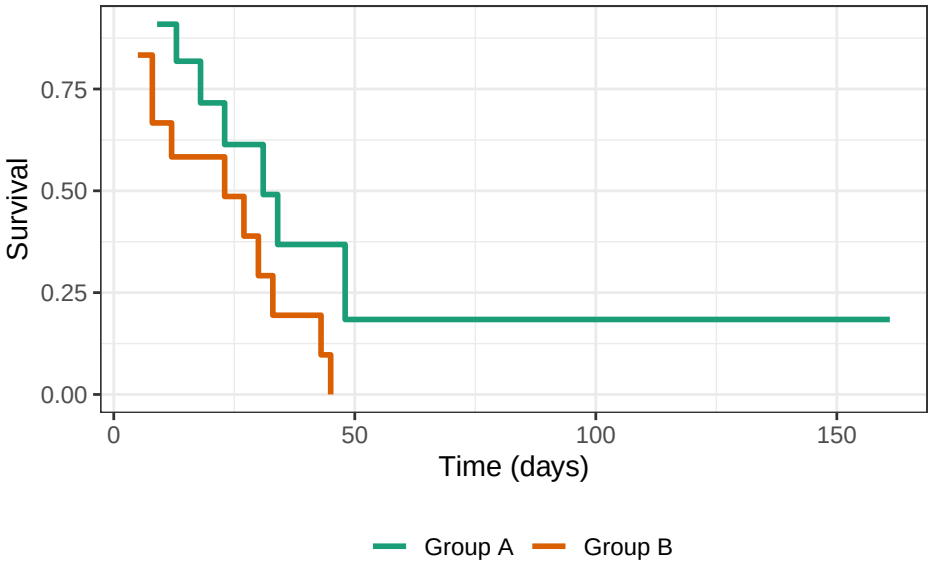


Figure 24.1: Illustration of Kaplan-Meier curves for two groups from a hypothetical study.

Of course, a point estimate does not allow us to make inference on the true survival. In order to perform inference, we need to model the sampling distribution of our estimate.

Definition 24.4 (Model for the Sampling Distribution of the Kaplan-Meier Estimator)**.** Consider estimating the survival $S(t)$ at time t using the Kaplan-Meier estimator $\widehat{S}(t)$. As the sample size increases, we have that

$$\frac{\widehat{S}(t) - S(t)}{\hat{\sigma}^2} \sim N(0, 1)$$

where

$$\hat{\sigma} = \widehat{S}(t) \sqrt{\sum_{t_i \leq t} \frac{d_{t_i}}{n_{t_i}\left(n_{t_i} - d_{t_i}\right)}}$$

and the estimate $\widehat{S}(0) = 1$ has no error.

As with life-table estimates, creating a classical confidence interval using the above standard error can result in confidence limits below 0 or above 1. It is common to construct the confidence interval on the hazard scale and then make a transformation back to the survival scale, which ensures confidence limits that are reasonable. This is the default approach in many statistical software packages.

> Big Idea
>
> Due to the censoring, visualizing survival data involves comparing the survival functions. Kaplan-Meier curves are preferred over summaries of only the observed survival times.

24.3 Log-Rank Test

Estimating a survival curve is useful for describing the survival experience of a population; however, we are often interested in formally comparing the survival of two populations. One method of comparing two or more groups is the log-rank test.

Definition 24.5 (Log-Rank Test)**.** The log-rank test formally compares k survival curves by testing the hypotheses

$$H_0 : S_1(t) = S_2(t) = \cdots = S_k(t) \ \forall t \quad \text{vs.}$$
$$H_1 : \text{At least one } S_k \text{ differs for at least one } t.$$

Notice that the above test is looking for *any* difference whatsoever in the curves; that is, this is not a point-wise comparison but a comparison of the entire curve. The log-rank test compares the expected number of events at a given time with the observed number of events. Tests of this form are often referred to as Chi-Squared tests as the test statistic generally has a Chi-Square distribution in large sample sizes.

While the mathematical details for implementing the log-rank test are beyond the scope of the test, we note that the asymptotic distribution depends on the number of events observed, not the sample size. That is, power calculations are based on the number of events expected.

While the log-rank test allows for comparisons across groups, it does not allow us to quantify the effect of a predictor on the survival. In order to characterize the impact of a predictor on survival or to account for multiple predictors, we need a modeling approach for survival analysis.

25 Proportional Hazards Model

Regression models allow us to quantify the effect of a set of predictors on the distribution of the response. While there are various regression methods for survival analysis, perhaps the most common is the Cox Proportional Hazards model, which we discuss in this chapter.

As the name of this chapter hints at, our modeling approach depends on the assumption of proportional hazards.

Definition 25.1 (Proportional Hazards)**.** Let $\lambda_1(t)$ and $\lambda_2(t)$ represent the hazard functions for two different groups. The assumption of proportional hazards states that

$$\frac{\lambda_2(t)}{\lambda_1(t)} = e^{\gamma}$$

for some fixed γ. That is, the ratio of the hazard functions does not depend on time.

The reason for choosing to exponentiate the constant in Definition 25.1 is because the hazard ratio must always be positive (since each hazard is positive for all values of t). This allows γ, the natural logarithm of the hazard ratio, to play a role similar to the log-odds ratio (from logistic regression in Chapter 18). When $\gamma = 0$, the hazard ratio is 1 and the two hazard functions are equal across all time (meaning the corresponding survival curves are equal across all time). When $\gamma > 0$, the hazard ratio is larger than 1 and group 2 is more likely to experience the event (survival curve sits below) compared with group 1. When $\gamma < 0$, the hazard ratio is below 1 and group 2 is less likely to experience the event (survival curve sits above) compared with group 1.

> **Warning**
>
> Proportional hazards is an assumption. It is not guaranteed to hold in any setting; however, it is a useful simplifying assumption and often does hold at least reasonably well.

Early in this unit, we alluded to the idea that some characterizations of the distribution are easier to model than others. When censoring is present, it turns out that modeling the hazard turns out to be easier than modeling the survival function directly. Therefore, we want our model to allow the hazard function to depend upon predictors.

> **Note**
>
> Remember, a key idea in regression modeling is characterizing the distribution of the response through its parameters. We have spent a great deal of time in the text characterizing the mean and variance of the response with our regression models. Here, instead of the mean survival time, we are modeling the hazard function.

> **Note**
>
> Note that unlike modeling the mean response in which our model specifies a single value (the mean response) for a given set of predictors, in survival analysis, we are modeling an entire function. That is, for a given set of predictors, we are specifying the hazard *function* over time.

We could model the hazard function under parametric assumptions (assuming a particular distribution for the event times and censor times, for example). However, Cox developed a semiparametric model that dominates the literature in the biological sciences.

Definition 25.2 (Cox Proportional Hazards Model)**.** The proportional hazards model (or Cox proportional hazards model, or PH model, or Cox PH model) is a model for the hazard function that enforces the assumption of proportional hazards; it has the following form:

$$\lambda\left(t \mid (\text{Predictors})_i\right) = \lambda_0(t) e^{\sum_{j=1}^{p} \beta_j (\text{Predictor } j)_i},$$

where the form of $\lambda_0(t)$, known as the baseline hazard, is not specified.

This model separates the hazard function into a function of time alone (the baseline hazard) and a function of the predictors alone (the exponent). Since the baseline hazard is not specified, this is a semiparametric model; if this form were specified, we would have a fully parametric model. The baseline hazard function represents the hazard function when all predictors take the value of 0. That is, instead of a single intercept, we have a type of "intercept-function."

As a result of the product of the baseline hazard with the exponential term, the predictors serve to scale the hazard function, making this a multiplicative model instead of additive. The name comes from the fact that the model enforces the assumption of proportional hazards. To see this, consider a simple model with only two predictors. Suppose one group of subjects has predictor values a and b, respectively. Then, their hazard function has the form

$$\lambda_1(t \mid \text{Predictors}) = \lambda_0(t) e^{\beta_1 a + \beta_2 b}.$$

Let a second group of subjects have predictor values $a+1$ and b, increasing the first predictor by 1 unit. Observe that their hazard function is

$$\lambda_2(t \mid \text{Predictors}) = \lambda_0(t)e^{\beta_1(a+1)+\beta_2 b}.$$

The baseline hazard function does not depend on the predictor values and so is shared between both groups. Now, the hazard ratio of group 2 compared with group 1 is

$$\begin{aligned}\text{HR} &= \frac{\lambda_2(t \mid \text{Predictors})}{\lambda_1(t \mid \text{Predictors})} \\ &= \frac{\lambda_0(t)e^{\beta_1 a+\beta_1+\beta_2 b}}{\lambda_0(t)e^{\beta_1 a+\beta_2 b}} \\ &= e^{\beta_1},\end{aligned}$$

which does not depend on time t and therefore suggests the hazard functions are proportional across time. Further, this derivation provides an interpretation for the parameters in the model.

Definition 25.3 (Interpretation of Parameters for the Proportional Hazards Model)**.** Consider modeling the hazard function using the proportional hazards model of Definition 25.2. The coefficient on the j-th predictor β_j is the log-hazard ratio associated with a one-unit increase in the j predictor, holding all other predictors fixed.

Of course, we do not observe the actual values of the parameters; therefore, we estimate the parameters and therefore estimate the hazard function. This is done via a partial likelihood, the details of which are beyond the scope of this text; it was this partial likelihood, however, that made the above model useful in practice. Partial likelihood allows us to estimate the parameters, and even model the sampling distribution of these estimates, without needing to specify the baseline hazard function!

The simplicity of the Cox PH model has led to wide use. However, its apparent simplicity is also the source of many common mistakes when using the model in practice.

i Common Errors when Working with the Proportional Hazards Model

The following are common mistakes made by practitioners working with the proportional hazards model (Definition 25.2).

- Using predictors that were gathered after baseline or the assignment of the treatment of interest. If we observe something after we "start the clock," then its very observation implies the subject is event-free at its observation.
- Forgetting the proportional hazards assumption implies one curve is superior regardless of time. If we estimate the survival curves from a proportional hazards

model, one treatment group will always be superior to the other (though the difference may be statistically indiscernible). This is because the proportional hazards assumption states that one survival curve always sits above the other.
- Neglecting that this framework can be generalized to handle repeated outcomes and time-dependent predictors. In our formulation, we do not allow the predictors to depend on time, but this framework generalizes nicely to do so (or multiple types of events). Some researchers perform unnecessary simplifications when the method should be generalized instead.

We have seen the appeal of semi-parametric models throughout the course. Given enough data, they often provide efficient estimation without requiring many conditions on the data generating process. While the proportional hazards model does avoid explicitly specifying the form of the survival distribution, it does carry certain conditions.

Conditions of the Proportional Hazards Model

When we model the hazard using the proportional hazards model (Definition 25.2), we impose the following conditions:

- The portion of the model involving the predictors is correctly specified.
- The predictors are linearly related to the log-hazard function.
- The affect of changing a predictor in isolation results in proportional hazards; this is enforced by the model. It implies that survival curves should not cross for different groups.
- The censoring time is independent of the survival time.

As with other forms of regression models, residual plots can be created to assess the first three conditions. The final condition must be assessed by discipline expertise. The complication is that there is not a single definition of a residual with time-to-event data. And, different types of residuals are useful for assessing different conditions. While beyond the scope of this text, we want to note that there are methods for relaxing many of these conditions while still remaining in the same general framework, making the proportional hazards model extremely flexible. For example, we could use splines to relax the "linearity" condition stated above.

References

Rosner, Bernard. 2006. *Fundamentals of Biostatistics*. 6th ed. CA: Thomson-Brooks/Cole.

Vittinghoff, Eric, David V. Glidden, Stephen C. Shiboski, and Charles E. McCulloch. 2012. *Regression Methods in Biostatistics: Linear Logistic, Survival, and Repeated Measures Models*. 2nd ed. NY: Springer-Verlag.

Wickham, Hadley. 2014. "Tidy Data." *Journal of Statistical Software* 59 (10): 1–23.

A Glossary

The following key terms were defined in the text; each term is presented with a link to where the term was first encountered in the text.

AIC (Definition 19.4) Given a model $\mathcal{M}$, Akaike's information criterion (AIC) is defined as

$$\text{AIC}_{\mathcal{M}} = -2\ln\left(\widehat{L}_{\mathcal{M}}\right) + 2p_{\mathcal{M}}$$

where $\widehat{L}_{\mathcal{M}}$ represents the likelihood function corresponding to model $\mathcal{M}$ when evaluated at the maximum likelihood estimates, and $p_{\mathcal{M}}$ is the number of parameters in $\mathcal{M}$.

Alternate Characterization of the Classical Regression Model (Definition 4.5) Under the classical regression conditions on the error term (see Definition 4.3), we can characterize the classical regression model as

$$(\text{Response})_i \mid (\text{Predictors 1 through } p)_i \overset{\text{Ind}}{\sim} N\left(\beta_0 + \sum_{j=1}^{p} \beta_j (\text{Predictor } j)_i, \sigma^2\right).$$

Here, the symbol | is read "given" and means that the distribution of the response is specified after knowing the values of the predictors. That is, the distribution of the response depends on these variables.

Autoregressive Correlation Structure (Definition 15.4) An autoregressive correlation structure suggests the correlation between two observations diminishes as the observations get further apart (generally, further apart in time). We generally only consider the autoregressive structure of degree 1 here; if there are five observations within a block, this has the form

$$\Gamma = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \cdot & 1 & \rho & \rho^2 & \rho^3 \\ \cdot & \cdot & 1 & \rho & \rho^2 \\ \cdot & \cdot & \cdot & 1 & \rho \\ \cdot & \cdot & \cdot & \cdot & 1 \end{pmatrix}.$$

BIC (Definition 19.5) Given a model $\mathcal{M}$, Schwarz's Bayesian information criterion (BIC) is defined as

$$\text{BIC}_{\mathcal{M}} = -2\ln\left(\widehat{L}_{\mathcal{M}}\right) + p_{\mathcal{M}}\ln(n).$$

Bernoulli Distribution (Definition 3.6) Let X be a discrete random variable taking the value 0 or 1. X is said to have a Bernoulli distribution with density

$$f(x) = \theta^x(1-\theta)^{1-x} \qquad x \in \{0,1\},$$

where $0 < \theta < 1$ is the probability that X takes the value 1.

- $E(X) = \theta$
- $Var(X) = \theta(1-\theta)$

We write $X \sim Ber(\theta)$, which is read "X has a Bernoulli distribution with probability θ."

Blocking (Definition 13.4) Blocking is a way of minimizing the variability contributed by an inherent characteristic that results in dependent observations. In some cases, the blocks are the unit of observation which is sampled from a larger population, and multiple observations are taken on each unit. In other cases, the blocks are formed by grouping the units of observations according to an inherent characteristic; in these cases that shared characteristic can be thought of having a value that was sampled from a larger population.

In both cases, the observed blocks can be thought of as a random sample; within each block, we have multiple observations, and the observations from the same block are more similar than observations from different blocks.

Bootstrapping (Definition 11.8) A process of constructing a sampling distribution of the parameter estimates through resampling. The observed data is resampled repeatedly, and the parameters of interest are estimated in each resample. The distribution of these estimates across the resamples is then used as an empirical model of the corresponding sampling distributions.

Case Resampling Bootstrap (Definition 11.10) Suppose we observe a sample of size n and use the data to compute the least squares estimates $\widehat{\beta}$ for the parameters in the model

$$(\text{Response})_i = \beta_0 + \sum_{j=1}^{p} \beta_j(\text{Predictor } j)_i + \varepsilon_i.$$

The case resampling bootstrap proceeds according to the following algorithm:

1. Take a random sample of size n (with replacement) of the raw data (keeping all variables from the same observation together); denote the i-th selected response and predictors $(\text{Response})_i^*$ and $(\text{Predictor } j)_i^*$, respectively.
2. Obtain the least squares estimates $\hat{\alpha}$ by finding the values of α that minimize

$$\sum_{i=1}^{n} \left((\text{Response})_i^* - \alpha_0 - \sum_{j=1}^{p} \alpha_j (\text{Predictor } j)_i^* \right)^2.$$

3. Repeat steps 1-2 m times.

We often take m to be large (at least 1000). After each pass through the algorithm, we retain the least squares estimates $\hat{\alpha}$ from the resample. The distribution of these estimates across the resamples is a good empirical model for the sampling distribution of the original least squares estimates.

Categorical Variable (Definition 1.5) Also called a "qualitative variable," a measurement on a subject which denotes a grouping or categorization.

Censored Data (Definition 23.1) Censored data is a special case of missing data for which a bound on the missing value is known. In survival analysis, the response of interest (time to an event) is subject to censoring.

Central Limit Theorem (Definition 11.6) Let $Y_1, Y_2, \ldots, Y_n$ be independent and identically distributed random variables with finite mean μ and variance σ^2. Then, as n approaches infinity, the distribution of the ratio

$$\frac{\sqrt{n}\left(\bar{Y} - \mu\right)}{\sigma}$$

approaches that of a Standard Normal random variable.

Chi-Square Distribution (Definition 3.8) Let X be a continuous random variable. X is said to have a Chi-Square distribution if the density is given by

$$f(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)}\, x^{\nu/2-1} e^{-x/2} \qquad x > 0,$$

where $\nu > 0$ is the degrees of freedom.

We write $X \sim \chi^2_\nu$, which is read "X has a Chi-Square distribution with ν degrees of freedom."

Classical Regression (Conditions on Predictors) (Definition 4.4) The classical regression model (Definition 4.3) places the following conditions on the predictors:

1. Each predictor is measured without error.

2. Each predictor has an additive linear effect on the response.

Classical Regression Model (Definition 4.3) In the "classical regression model," we place the following four conditions on the distribution of the error ε_i:

1. The average error across all levels of the predictors is 0; mathematically, we write $E\left(\varepsilon_i \mid (\text{Predictors 1 - } p)_i\right) = 0$.
2. The variance of the errors is constant across all levels of the predictors; mathematically, we write $Var\left(\varepsilon_i \mid (\text{Predictors 1 - } p)_i\right) = \sigma^2$ for some unknown constant $\sigma^2 > 0$. This is sometimes referred to as homoskedasticity.
3. The error terms are independent; in particular, the magnitude of the error for one observation does not influence the magnitude of the error for any other observation.
4. The distribution of the errors follows a Normal distribution with the above mean and variance.

Cluster Samples (Definition 13.14) Stratified sampling divides a population into groups and samples from within each group; in contrast, cluster sampling divides the population into groups and randomly samples a few groups and takes measurements from within the group.

Codebook (Definition 1.16) Also called a "data dictionary," a codebook provides complete information regarding the variables contained within a dataset.

Complete Data (Definition 23.2) This term describes a data set for which the event is observed on all subjects.

Compound Symmetric Correlation Structure (Definition 15.3) A compound symmetric correlation structure, also known as an *exchangeable* correlation structure, suggests the correlation between any two errors within a subject is equal. If there are five observations within a block, this has the form

$$\Gamma = \begin{pmatrix} 1 & \rho & \rho & \rho & \rho \\ \cdot & 1 & \rho & \rho & \rho \\ \cdot & \cdot & 1 & \rho & \rho \\ \cdot & \cdot & \cdot & 1 & \rho \\ \cdot & \cdot & \cdot & \cdot & 1 \end{pmatrix}.$$

Confidence Interval (Definition 1.11) An interval (range of values) estimate of a parameter that incorporates the variability in the statistic. The process of constructing $k\%$ confidence intervals results in them containing the parameter of interest in $k\%$ of *repeated* studies. The value of k is called the *confidence level*.

Confidence Interval for Parameters Under Classical Model (Definition 4.10) Under the classical regression conditions (Definition 4.3), a $100c\%$ confidence interval for the parameter β_j is given by

$$\hat{\beta}_j \pm t_{n-p-1,0.5(1+c)} \sqrt{Var\left(\hat{\beta}_j\right)}.$$

where $t_{n-p-1,0.5(1+c)}$ is the $0.5(1+c)$ quantile from the t_{n-p-1} distribution, known as the critical value for the confidence interval.

Confounding (Definition 1.15) When the effect of a variable on the response is misrepresented due to the presence of a third, potentially unobserved, variable known as a confounder.

Correlation Structure (Definition 13.6) The correlation structure quantifies the strength and direction of the relationship between the errors in the observed responses.

Cox Proportional Hazards Model (Definition 25.2) The proportional hazards model (or Cox proportional hazards model, or PH model, or Cox PH model) is a model for the hazard function that enforces the assumption of proportional hazards; it has the following form:

$$\lambda\left(t \mid (\text{Predictors})_i\right) = \lambda_0(t) e^{\sum_{j=1}^{p} \beta_j (\text{Predictor } j)_i},$$

where the form of $\lambda_0(t)$, known as the baseline hazard, is not specified.

Cross Sectional Study (Definition 13.12) A cross sectional study considers data from a single snapshot in time.

Cross-Over Study (Definition 13.9) A cross-over study exposes each participant to multiple treatments. Whenever possible, the order of the treatments is randomly determined. This is equivalent to a randomized complete block design in which the blocks are the participants. When the treatments are believed to have a lingering effect, a wash-out period between treatments is used to minimize the impact of previous treatments on the treatment the participant is currently being exposed to.

Cumulative Distribution Function (CDF) (Definition 3.3) Let X be a random variable; the cumulative distribution function (CDF) is defined as

$$F(u) = Pr(X \leq u).$$

For a continuous random variable, we have that

$$F(u) = \int_{-\infty}^{u} f(x)dx$$

implying that the density function is the derivative of the CDF. For a discrete random variable

$$F(u) = \sum_{x \leq u} f(x).$$

Cumulative Hazard (Definition 22.4) For any time t, we have that

$$S(t) = e^{-\int_0^t \lambda(u)du}$$

where

$$\Lambda(t) = -\int_0^t \lambda(u)du$$

is known as the cumulative hazard function.

Density Function (Definition 3.2) A density function f relates the potential values of a random variable X with the probability those values occur. For a *continuous* random variable, the probability the random variable X falls within an interval (a, b) is given by

$$Pr(a \leq X \leq b) = \int_a^b f(x)dx.$$

For a *discrete* random variable, the probability the random variable X is equal to the value u is given by

$$Pr(X = u) = f(u).$$

Distribution (Definition 1.7) The pattern of variability corresponding to a set of values.

Estimate of the Variance of the Errors (Definition 4.9) The unknown variance in the linear model, which captures the variability in the response for any set of predictors (also called the residual variance), is estimated by

$$\hat{\sigma}^2 = \frac{1}{n-p-1} \sum_{i=1}^{n} \left((\text{Response})_i - \hat{\beta}_0 - \sum_{j=1}^{p} \hat{\beta}_j (\text{Predictor } j)_i \right)^2.$$

Event Time and Censoring Indicator (Definition 23.9) Let T_i and C_i represent the survival time and the censoring time for the i-th subject. When data is subject to right censoring, we observe

$$X_i = \min\{T_i, C_i\}$$

which is known as the *event* time. We also observe whether this observation was triggered by the actual event or censoring:

$$\delta_i = \begin{cases} 1 & \text{if } T_i \leq C_i \\ 0 & \text{if } T_i > C_i. \end{cases}$$

This is known as the *censoring indicator*.

F-Distribution (Definition 3.9) Let X be a continuous random variable. X is said to have an F-distribution if the density is given by

$$f(x) = \frac{\Gamma((r+s)/2)}{(\Gamma(r/2)\Gamma(s/2))}(r/s)^{(r/2)}x^{(r/2-1)}(1+(r/s)x)^{-(r+s)/2} \qquad x > 0,$$

where $r, s > 0$ are the numerator and denominator degrees of freedom, respectively.

We write $X \sim F_{r,s}$, which is read "X has an F-distribution with r numerator degrees of freedom and s denominator degrees of freedom."

Fixed Effect (Definition 13.7) Fixed effects are terms in the model for which we are interested in both the specific grouping levels, and we are interested in characterizing the relationship between these levels and the response.

Gauss-Newton Method (Definition 20.1) The Gauss-Newton method is the most commonly used optimization routine for statistical applications; this method relies on linearizing the problem using a Taylor Series approximation.

General Linear Hypothesis (Definition 10.1) The general linear hypothesis framework refers to testing hypotheses of the form

$$H_0 : \mathbf{K}\beta = \mathbf{m} \qquad \text{vs.} \qquad H_1 : \mathbf{K}\beta \neq \mathbf{m}$$

where

- β is the $(p+1)$-length vector of the parameters (includes the intercept),
- $\mathbf{K}$ is an r-by-$(p+1)$ matrix that specifies the linear combinations defining the hypothesis of interest, and
- $\mathbf{m}$ is a vector of length r specifying the null values, the value of each linear combination under the null hypothesis (often a vector of 0's).

General Linear Model (Definition 4.1) The general linear model views the response (outcome) as a linear combination of several predictors:

$$\begin{aligned}(\text{Response})_i &= \beta_0 + \beta_1(\text{Predictor 1})_i + \beta_2(\text{Predictor 2})_i + \cdots + \beta_p(\text{Predictor } p)_i + \varepsilon_i \\ &= \beta_0 + \sum_{j=1}^{p} \beta_j(\text{Predictor } j)_i + \varepsilon_i\end{aligned}$$

where n is the number of subjects in the sample, $p < n$ is the number of predictors in the model, and ε_i is a random variable that captures the error in the response.

Generalized Estimating Equations (GEE) (Definition 15.6) Generalized estimating equations can be used to estimate the parameters of a model while accounting for the correlation among observations. In addition to specifying a model for the overall average response, a "working" structure is specified for the correlation of observations from the same subject. The working structure is updated during the estimation process and used to adjust the standard errors of the parameter estimates in the mean model.

Generalized Least Squares (Definition 17.1) The semiparametric nonlinear model can be generalized to capture non-constant variance. Specifically, we specify the *mean* and *variance* of the response given the predictors

$$\begin{aligned}E\left[(\text{Response})_i \mid (\text{Predictors})_i\right] &= f\left((\text{Predictors})_i, \beta\right) \\ Var\left[(\text{Response})_i \mid (\text{Predictors})_i\right] &= g\left((\text{Predictors})_i, \beta, \gamma\right).\end{aligned}$$

Such a model is fit with the method of generalized least squares (as opposed to "ordinary" least squares) in which we alternate between (a) minimizing a weighted distance between the observed response and the mean function and (b) minimizing the distance between the squared residuals and the variance function. That is, we minimize

$$\begin{aligned}&\sum_{i=1}^{n} \frac{1}{g\left((\text{Predictors})_i, \beta, \gamma\right)} \left[(\text{Response})_i - f\left((\text{Predictors})_i, \beta\right)\right]^2 \\ &\sum_{i=1}^{n} \left(g\left((\text{Predictors})_i, \beta, \gamma\right) - \left[(\text{Response})_i - f\left((\text{Predictors})_i, \beta\right)\right]^2\right)^2.\end{aligned}$$

Hazard Function (Definition 22.3) For any time t, the hazard function $\lambda(t)$ is the instantaneous mortality rate per unit time:

$$\lambda(t) = \lim_{h \to 0} \frac{Pr(t \leq T \leq t + h \mid T > t)}{h} = \frac{f(t)}{S(t)} = -\frac{\frac{d}{dt}S(t)}{S(t)} = -\frac{d}{dt}\log(S(t)).$$

Hierarchical Model (Definition 14.1) A hierarchical model breaks the data generating process into smaller stages and posits a model for each stage. The stages are determined by defining a hierarchy of units and thereby capturing the sources of variability.

Independence Correlation Structure (Definition 15.2) An independence correlation structure suggests there is no correlation among any of the error terms within a subject. If there are five observations within a block, this has the form

$$\Gamma = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ \cdot & 1 & 0 & 0 & 0 \\ \cdot & \cdot & 1 & 0 & 0 \\ \cdot & \cdot & \cdot & 1 & 0 \\ \cdot & \cdot & \cdot & \cdot & 1 \end{pmatrix}.$$

Indicator Variables (Definition 8.1) Also called "dummy variables," these are a set of binary variables that capture the grouping defined by a categorical variable for regression modeling.

Individual-Level Model (Definition 14.2) The individual-level model characterizes the response for the i-th subject (or block) only.

Information Criteria (Definition 19.3) Information criteria refer to metrics that balance the model fit (through the likelihood function) with the model complexity through some penalty term.

Interaction (Definition 9.2) An interaction term allows the effect of a predictor on the response to depend on the value of a second predictor (capturing an effect modification).

- The interaction term is created by adding the product of the two predictors under consideration to the model.

Intercept (Definition 4.6) The population intercept, denoted β_0, is the *mean* response when all predictors take the value zero.

Interpretation of Parameters for the Proportional Hazards Model (Definition 25.3) Consider modeling the hazard function using the proportional hazards model of Definition 25.2. The coefficient on the j-th predictor β_j is the log-hazard ratio associated with a one-unit increase in the j predictor, holding all other predictors fixed.

Interpretation of Parameters in a Logistic Regression Model (Definition 18.7) Let β_j be the parameter associated with the j-th predictor in a logistic regression model (Definition 18.1). Then, β_j represents the log-OR ("log odds ratio") associated with a one-unit increase in the j-th predictor holding all other predictors fixed.

Interval Censoring (Definition 23.8) Interval censoring refers to scenarios when the event time is known to have occurred within some interval, but the exact time is unknown.

Kaplan-Meier Estimator (Definition 24.3) Also known as the product-limit estimator, the Kaplan-Meier estimator is the limit of the life-table estimate as we allow the intervals to shrink to a single event time:

$$\widehat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_{t_i}}{n_{t_i}}\right)$$

where t_i is the i-th survival time where the event of interest was observed.

Large Sample Model for the Sampling Distribution of the Least Squares Estimates (Definition 11.7)
Suppose the classical regression conditions hold, with the exception of the errors following a Normal distribution. As the sample size gets large, we have that the distribution of the ratio

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{Var\left(\hat{\beta}_j\right)}} \sim N(0,1)$$

for all $j = 0, 1, \ldots, p$. Further, under the null hypothesis

$$H_0 : \mathbf{K}\beta = \mathbf{m}$$

we have

$$\left(\mathbf{K}\hat{\beta} - \mathbf{m}\right)^\top \left(\mathbf{K}\widehat{\Sigma}\mathbf{K}^\top\right)^{-1} \left(\mathbf{K}\hat{\beta} - \mathbf{m}\right) \sim \chi^2_r.$$

Large Sample Model for the Sampling Distribution of the Least Squares Estimates in Nonlinear Models (I
Consider a nonlinear model as described in Definition 16.2. Assuming the form of the model is correctly specified, as the sample size gets large, we have that

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{Var\left(\hat{\beta}_j\right)}} \sim N(0,1)$$

for all $j = 1, \ldots, p$. Further, under the null hypothesis

$$H_0 : \mathbf{K}\beta = \mathbf{m}$$

we have that

$$\left(\mathbf{K}\hat{\beta} - \mathbf{m}\right)^\top \left(\mathbf{K}\widehat{\Sigma}\mathbf{K}^\top\right)^{-1} \left(\mathbf{K}\hat{\beta} - \mathbf{m}\right) \sim \chi^2_r$$

where r is the rank (number of rows) of $\mathbf{K}$ and $\widehat{\Sigma}$ is the estimated variance-covariance matrix of the parameter estimates.

Large Sample Sampling Distribution of Parameter Estimates in Logistic Regression (Definition 18.4) Consider the logistic regression model in Definition 18.1. Assuming the form of the model is correctly specified with parameter vector β, as the sample size gets large, we have that

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{Var\left(\hat{\beta}_j\right)}} \sim N(0,1)$$

for all $j = 1, \ldots, p$. Further, under the null hypothesis

$$H_0 : \mathbf{K}\beta = \mathbf{m}$$

we have

$$\left(\mathbf{K}\hat{\beta} - \mathbf{m}\right)^\top \left(\mathbf{K}\widehat{\Sigma}\mathbf{K}^\top\right)^{-1} \left(\mathbf{K}\hat{\beta} - \mathbf{m}\right) \sim \chi^2_r$$

where r is the rank (number of rows) of $\mathbf{K}$.

Large Sample Theory (Definition 11.5) The phrase "large sample theory" (or "asymptotics") is used to describe a scenario when the model for the sampling distribution (or null distribution) of an estimate (or standardized statistic) can be approximated as the sample size becomes infinitely large. That is, as the sample size approaches infinity, the sampling distribution (or null distribution) can be easily modeled using a known probability distribution.

Least Squares Estimation (Definition 4.2) The method of least squares may be used to estimate the coefficients (parameters) of a linear model. In particular, we choose the values of the coefficients that minimize

$$\sum_{i=1}^{n} \left((\text{Response})_i - \beta_0 - \sum_{j=1}^{p} \beta_j (\text{Predictor } j)_i \right)^2 .$$

The resulting "least squares" estimates are denoted $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$.

Left Censoring (Definition 23.7) Left censoring refers to scenarios when an *upper* bound is known on the response.

Life Table (Definition 24.1) Life tables are a method of estimating overall survival over key intervals of time, generally constructed for a single population.

Likelihood Function (Definition 18.2) For a fully parametric model, the likelihood function $\mathcal{L}(\beta, \text{Observed Data})$ captures how likely the observed data is to be realized in a future study under a specific set of parameters. This is directly related to the density function of the parametric model assumed.

Linear Model (Definition 12.1) A model is said to be linear if it can be expressed as a linear combination of the *parameters*. That is, the linearity does not refer to the form of the predictors but the form of the parameters.

Linear Spline (Definition 12.3) A linear spline is a continuous piecewise linear function.

Log-Rank Test (Definition 24.5) The log-rank test formally compares k survival curves by testing the hypotheses

$$\begin{aligned} &H_0 : S_1(t) = S_2(t) = \cdots = S_k(t) \ \forall t \quad \text{vs.} \\ &H_1 : \text{At least one } S_k \text{ differs for at least one } t. \end{aligned}$$

Logistic Regression Model (Definition 18.1) A model for binary responses where the response, given the predictors, has a Bernoulli distribution such that

$$Pr\left((\text{Response})_i = 1 \mid (\text{Predictors})_i\right) = \frac{e^{\beta_0 + \sum_{j=1}^{p} \beta_j (\text{Predictor } j)_i}}{1 + e^{\beta_0 + \sum_{j=1}^{p} \beta_j (\text{Predictor } j)_i}} \tag{A.1}$$

and all responses are independent of one another.

Longitudinal Study (Definition 13.11) A longitudinal study repeatedly measures the response on each subject at various points in time.

Maximum Likelihood Estimation (Definition 18.3) The method of maximum likelihood estimation chooses parameter estimates to maximize the likelihood function under an assumed parametric model. The resulting estimates are known as maximum likelihood estimates.

Mean and Variance of a Random Variable (Definition 3.4) Suppose X is a random variable with density function f. If X is a continuous random variable, then the mean and variance are given by

$$\begin{aligned} E(X) &= \int x f(x) dx \\ Var(X) &= \int (x - E(X))^2 f(x) dx. \end{aligned}$$

If X is a discrete random variable, then the mean and variance are given by

$$\begin{aligned} E(X) &= \sum x f(x) \\ Var(X) &= \sum (x - E(X))^2 f(x). \end{aligned}$$

Mixed-Effects Model (Definition 14.4) A mixed-effects model denotes a hierarchical model for which some effects are fixed (not allowed to vary across subjects) and others are random (allowed to vary across subjects).

Model for the Null Distribution with the General Linear Hypothesis (Definition 10.3) Let $\hat{\beta}$ be the $(p+1)$ vector of estimates for the parameter vector β, and let the estimates have variance-covariance matrix Σ. Assuming the null hypothesis

$$H_0 : \mathbf{K}\beta = \mathbf{m}$$

is true, under the conditions of the classical regression model (Definition 4.3)

$$(1/r)\left(\mathbf{K}\hat{\beta} - \mathbf{m}\right)^{\top}\left(\mathbf{K}\widehat{\Sigma}\mathbf{K}^{\top}\right)^{-1}\left(\mathbf{K}\hat{\beta} - \mathbf{m}\right) \sim F_{r,n-p-1}.$$

Model for the Sampling Distribution of Life-Table Estimates (Definition 24.2) Consider estimating the survival $S(t)$ at time t using life-table estimate $\widehat{S}(t)$. As the sample size increases, we have that

$$\frac{\widehat{S}(t) - S(t)}{\hat{\sigma}^2} \sim N(0,1),$$

where

$$\hat{\sigma} = \widehat{S}(t)\sqrt{\sum_{i=1}^{t}\frac{d_t}{(n_t - w_t/2)\,(n_t - w_t/2 - d_t)}}$$

and the estimate $\widehat{S}(0) = 1$ has no error.

Model for the Sampling Distribution of the Kaplan-Meier Estimator (Definition 24.4) Consider estimating the survival $S(t)$ at time t using the Kaplan-Meier estimator $\widehat{S}(t)$. As the sample size increases, we have that

$$\frac{\widehat{S}(t) - S(t)}{\hat{\sigma}^2} \sim N(0,1)$$

where

$$\hat{\sigma} = \widehat{S}(t)\sqrt{\sum_{t_i \le t}\frac{d_{t_i}}{n_{t_i}\left(n_{t_i} - d_{t_i}\right)}}$$

and the estimate $\widehat{S}(0) = 1$ has no error.

Mortality Rate (Definition 22.2) For any particular time t, the morality rate $m(t)$ is the proportion of the population that experiences the event between times t and $t+1$, among individuals that are event-free at time t. That is,

$$m(t) = Pr(t \leq T \leq t+1 \mid T > t) = 1 - \frac{S(t+1)}{S(t)}.$$

Multicollinearity (Definition 7.1) When two predictors are highly correlated with one another, we say that there is multicollinearity in the model.

Nonlinear Model (Definition 16.1) A model is said to be nonlinear if it cannot be written as a linear combination of the parameters.

Nonparametric Model (Definition 11.2) A nonparametric model is unable to characterize the response using a finite set of parameters; for our purposes, this generally means the model makes no assumptions about the structure of the underlying distribution of the response given the predictors. Only minimal assumptions (such as independence between observations) are imposed.

Normal (Gaussian) Distribution (Definition 3.5) Let X be a continuous random variable. X is said to have a Normal (or Gaussian) distribution if the density is given by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \qquad -\infty < x < \infty,$$

where μ is any real number and $\sigma^2 > 0$.

- $E(X) = \mu$
- $Var(X) = \sigma^2$

We write $X \sim N(\mu, \sigma^2)$, which is read "X has a Normal distribution with mean μ and variance σ^2." This short-hand implies the density above.

Numeric Variable (Definition 1.6) Also called a "quantitative variable," a measurement on a subject which takes on a numeric value *and* for which ordinary arithmetic makes sense.

Observational Study (Definition 1.14) A study in which each participant "self-selects" into one of groups being compared in the study. The phrase "self-selects" is used very loosely here and can include studies in which the groups are defined by an inherent characteristic, the groups are determined according to a non-random mechanism, and each participant chooses the group to which they belong.

Odds (Definition 18.5) The odds of an event with probability p is defined as

$$\frac{p}{1-p}.$$

Odds Ratio (Definition 18.6) The odds ratio is a method of comparing two events; typically, it is formed by the ratio of the odds of the same event under two different scenarios. Let p_1 be the probability of the event under scenario 1 and let p_2 be the probability of an event under scenario 2; then, the odds of the event under scenario 1 are

$$\gamma_1 = \frac{p_1}{1-p_1},$$

and the odds of the event under scenario 2 are

$$\gamma_2 = \frac{p_2}{1-p_2}.$$

The odds ratio comparing scenario 1 to scenario 2 is

$$OR = \frac{\gamma_1}{\gamma_2} = \left(\frac{p_1}{1-p_1}\right)\left(\frac{1-p_2}{p_2}\right).$$

Overfitting (Definition 19.2) Overfitting refers to constructing a model that accurately predicts the observed data at the expense of accurate variance estimation and model parsimony. As a result, the model will have poor prediction for future observations.

P-Value (Definition 1.12) The probability, assuming the null hypothesis is true, that we would observe a statistic, from sampling variability alone, as extreme or more so as that observed in our sample. This quantifies the strength of evidence against the null hypothesis. Smaller values indicate stronger evidence.

P-Value for Testing if Parameter Belongs in Model Under Classical Model (Definition 4.11) Under the classical regression conditions (Definition 4.3), the p-value for testing the hypotheses

$$H_0 : \beta_j = 0 \qquad \text{vs.} \qquad H_1 : \beta_j \neq 0$$

is given by

$$Pr\left(|T| > \left|\frac{\hat{\beta}_j}{\sqrt{Var\left(\hat{\beta}_j\right)}}\right|\right)$$

where $T \sim t_{n-p-1}$.

Parameter (Definition 1.9) Numeric quantity which summarizes the distribution of a variable within the *population* of interest. Generally denoted by Greek letters in statistical formulas.

Parametric Model (Definition 11.1) A parametric model characterizes the distribution of the response using a finite set of parameters; for our purposes, this generally means the model *fully* characterize the distribution of the response given the predictors.

Population (Definition 1.1) The collection of subjects we would like to say something about.

Population Averaged Models (Definition 15.9) Also known as marginal modeling, the population-averaged approach posits a model for the mean response directly and addresses the correlation through directly modeling its structure.

Population-Level Model (Definition 14.3) The population-level model characterizes how the *parameters* of the individual-level model vary across subjects (or blocks) in the population.

Power of the Mean Model (Definition 17.2) The power of the mean model allows the variance to be specified as a power of the mean response function. Specifically, we consider

$$E\left[(\text{Response})_i \mid (\text{Predictors})_i\right] = f\left((\text{Predictors})_i, \beta\right)$$
$$Var\left[(\text{Response})_i \mid (\text{Predictors})_i\right] = \sigma^2\left[f\left((\text{Predictors})_i, \beta\right)\right]^{2\theta}$$

Proportional Hazards (Definition 25.1) Let $\lambda_1(t)$ and $\lambda_2(t)$ represent the hazard functions for two different groups. The assumption of proportional hazards states that

$$\frac{\lambda_2(t)}{\lambda_1(t)} = e^{\gamma}$$

for some fixed γ. That is, the ratio of the hazard functions does not depend on time.

R-squared (Definition 19.1) The proportion of the variability in the response explained by the model. When the response is quantitative, R-squared is defined as

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n} (\text{Residuals})_i^2}{\sum_{i=1}^{n} \left((\text{Response})_i - (\text{Overall Average Response})\right)^2}.$$

When the response is categorical (such as logistic regression), there is no unique definition of R-squared.

Random Censoring (Definition 23.6) Random censoring is a form of right censoring when subjects are withdrawn from the study at any time. It is typically assumed that the event time and the censoring time are independent of one another.

Random Effect (Definition 13.8) Random effects are terms in the model that capture the correlation induced due to an inherent characteristic that varies across the population. We are *not* interested in the specific grouping levels, and we either are not interested in the relationship with the response.

Random Variable (Definition 3.1) A random variable represents a measurement that will be collected and for which the value cannot be predicted with certainty; they are generally represented with a capital letter. Continuous random variables represent quantitative measurements while discrete random variables represent qualitative measurements.

Randomization (Definition 13.2) Randomization can refer to random *selection* or random *allocation*.

Random selection refers to the use of a random mechanism to select units from the population. Random selection minimizes bias.

Random allocation refers to the use of a random mechanism when assigning units to a specific treatment group in a controlled experiment. Random allocation eliminates confounding and permits causal interpretations.

Randomized Clinical Trial (Definition 1.13) Also called a "controlled experiment," a study in which each participant is randomly assigned to one of the groups being compared in the study.

Randomized Complete Block Design (Definition 13.10) A randomized complete block design is an example of a controlled experiment utilizing blocking. Each treatment is randomized to observations within blocks in such a way that every treatment is present within the block and the same number of observations are assigned to each treatment within each block.

Reduction of Noise (Definition 13.3) Reducing extraneous sources of variability can be accomplished by fixing extraneous variables or through blocking. These actions reduce the number of differences between the units under study.

Reference Group (Definition 8.2) The group defined by having all indicator variables for a particular categorical variable set to zero.

Repeated Measures (Definition 13.5) The phrase "repeated measures" refers to data for which the observed responses can be grouped based on some nuisance variable (typically the participant), and this grouping captures some inherent characteristic such that observations within a group tend to be more alike than observations across groups.

Replication (Definition 13.1) Replication results from taking measurements on different units (or subjects) for which you expect the results to be similar. That is, any variability across the units is due to natural variability within the population.

Residual (Definition 5.1) A residual for the i-th observation is the difference between an observed value and the predicted response:

$$\begin{aligned}(\text{Residual})_i &= (\text{Observed Response})_i - (\text{Predicted Response})_i \\ &= (\text{Response})_i - \left(\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j (\text{Predictor } j)_i\right).\end{aligned}$$

Residual Bootstrap (Definition 11.9) Suppose we observe a sample of size n and use the data to compute the least squares estimates $\widehat{\beta}$ for the parameters in the model

$$(\text{Response})_i = \beta_0 + \sum_{j=1}^{p} \beta_j (\text{Predictor } j)_i + \varepsilon_i.$$

The residual bootstrap proceeds according to the following algorithm:

1. Compute the residuals

$$(\text{Residuals})_i = (\text{Response})_i - (\text{Predicted Response})_i$$

2. Take a random sample of size n (with replacement) of the residuals; call these values $e_1^*, \ldots, e_n^*$.
3. Form "new" responses $y_1^*, \ldots, y_n^*$ according to

$$y_i^* = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j (\text{Predictor } j)_i + e_i^*.$$

4. Obtain the least squares estimates $\hat{\alpha}$ by finding the values of α that minimize

$$\sum_{i=1}^{n} \left(y_i^* - \alpha_0 - \sum_{j=1}^{p} \alpha_j (\text{Predictor } j)_i \right)^2.$$

5. Repeat steps 2-4 m times.

We often take m to be large (at least 1000). After each pass through the algorithm, we retain the least squares estimates $\hat{\alpha}$ from the resample. The distribution of these estimates across the resamples is a good empirical model for the sampling distribution of the original least squares estimates.

Response Variable (Definition 1.8) Also called the "outcome," this is the primary variable of interest in the research question; it is the variable we either want to explain or predict.

Restricted Cubic Spline (Definition 12.4) A restricted cubic spline is a continuous function comprised of piecewise cubic polynomials for which the tails of the spline have been restricted to be linear.

Right Censoring (Definition 23.3) Right censoring refers to scenarios when a *lower* bound is known on the response.

Robust Sandwich Estimator (Definition 15.7) The robust sandwich estimator of the variance-covariance matrix of the parameter estimates from the mean model balances the relationship between the parameter estimates specified by the model (and the "working" correlation matrix) with the relationship suggested by the observed data. Specifically, it has the form

$$\widehat{\Sigma} = \widehat{\mathbf{U}}\widehat{\mathbf{U}}^{-1/2}\mathbf{R}\widehat{\mathbf{U}}^{-1/2}\widehat{\mathbf{U}}$$

where $\mathbf{U}$ represents the model-based variance-covariance matrix if the structure specified by the working correlation matrix were completely correct, and $\mathbf{R}$ represents the correction factor estimated from the residuals (an empirical estimate).

Sample (Definition 1.2) The collection of subjects for which we actually obtain measurements (data).

Sampling Distribution of the Least Squares Estimates (Definition 4.8) Under the classical regression conditions (Definition 4.3), we have that

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{Var\left(\hat{\beta}_j\right)}} \sim t_{n-p-1}.$$

The denominator $\sqrt{Var\left(\hat{\beta}_j\right)}$ is known as the *standard error* of the estimate $\hat{\beta}_j$. This formula holds for all $j = 0, 1, \ldots, p$.

Semiparametric Linear Model (Definition 11.4) Suppose we no longer require that the error terms follow a Normal distribution; however, we do continue to impose the remaining conditions of the classical regression model. Then, our model could be written as

$$\begin{aligned} E\left[(\text{Response})_i \mid (\text{Predictors 1 through } p)_i\right] &= \beta_0 + \sum_{j=1}^{p} \beta_j (\text{Predictor } j)_i \\ Var\left[(\text{Response})_i \mid (\text{Predictors 1 through } p)_i\right] &= \sigma^2 \end{aligned}$$

where the responses are independent of one another given the predictors.

Semiparametric Model (Definition 11.3) A semiparametric model specifies some components of the underlying distribution of the response using a finite set of parameters but does not fully characterize the distribution. This generally means that we may specify the mean and/or variance of the response given the predictors, but we do not characterize the distributional family of the response.

Semiparametric Nonlinear Model (Definition 16.2) A semiparametric nonlinear model specifies the *mean* and *variance* of the response given the predictors; we write

$$\begin{aligned} E\left[(\text{Response})_i \mid (\text{Predictors})_i\right] &= f\left((\text{Predictors})_i, \beta\right) \\ Var\left[(\text{Response})_i \mid (\text{Predictors})_i\right] &= \sigma^2 \end{aligned}$$

where $f(\cdot)$ is referred to as the mean response function.

Slope (Definition 4.7) The coefficient for the j-th predictor, denoted β_j, is the change in the mean response associated with a one unit increase in Predictor j, *holding all other predictors fixed.*

Spaghetti Plot (Definition 13.15) A spaghetti plot is a scatterplot that displays the trends within a subject, highlighting the correlation structure by connecting points from the same subject.

Spline (Definition 12.2) A spline is a continuous piecewise polynomial used to model curvature. The points that define the piecewise components are called *knot points*; the functional form is allowed to change at the knot points.

Stationarity (Definition 15.5) The assumption of stationarity states that the correlation structure does not depend on time, only the distance between the observations.

Statistic (Definition 1.10) Numeric quantity which summarizes the distribution of a variable within the observed *sample.*

Statistical Inference (Definition 1.3) The process of using a sample to characterize some aspect of the underlying population.

Subgroup Analysis (Definition 9.1) Refers to repeating a specified analysis (e.g., regression model) within various levels of a categorical predictor.

- This will appropriately estimate the effect modification.
- This results in a loss of information because *all parameters* are forced to vary across the subgroups.

Subject Specific Models (Definition 15.8) Also known as conditional modeling, the subject-specific approach models at the subject-level and addresses the correlation indirectly through the inclusion of random effects.

Subsampling (Definition 13.13) Subsampling occurs when several measurements are taken on each subject under the same treatment, possibly at unique locations.

Survival Function (Definition 22.1) Let T be a random variable; the survival function $S(u)$ is defined as

$$S(t) = Pr(T > t) = 1 - F(t),$$

capturing the probability of failing *after* a time, where $F(t)$ is the CDF of T.

For a continuous random variable, we have that

$$S(t) = \int_t^\infty f(u)du$$

implying that $f(t) = -\frac{d}{dt}S(t)$.

Type I Censoring (Definition 23.4) Type-I censoring is a form of right censoring where the only source of censoring is the end of the study, for which the duration was pre-determined. Therefore, the time at which subjects are censored is the pre-determined study duration.

Type II Censoring (Definition 23.5) Type-II censoring is a form of right censoring where the only source of censoring is the end of the study, which is determined when the r-th event occurs and r is pre-determined. Therefore, the time at which subjects are censored is determined by the r-th event.

Unstructured Correlation Structure (Definition 15.1) An unstructured correlation structure suggests that the correlation between any two errors within a subject can take on any value. We only require that it be a valid correlation matrix.

Variable (Definition 1.4) A measurement, or category, describing some aspect of the subject.

Variance-Covariance Matrix (Definition 10.2) Let β represent the $(p+1)$-length vector of the parameters and $\hat{\beta}$ represent the $(p+1)$ vector of the parameter *estimates*. The variance-covariance matrix of the parameter estimates is the $(p+1)$-by-$(p+1)$ matrix Σ where

- the j-th diagonal element contains $Var\left(\hat{\beta}_j\right)$, and
- the (i,j)-th element contains the covariance between $\hat{\beta}_i$ and $\hat{\beta}_j$.

Wild Bootstrap (Definition 17.3) Suppose we observe a sample of size n and use it to fit the mean model (linear or nonlinear)

$$E\left[(\text{Response})_i \mid (\text{Predictors})_i\right] = f\left((\text{Predictors})_i, \beta\right)$$

to obtain the ordinary least squares estimates $\hat{\beta}$. The wild bootstrap proceeds along the following algorithm:

1. Compute the residuals
$$(\text{Residual})_i = (\text{Response})_i - f\left((\text{Predictors})_i, \hat{\beta}\right)$$
2. Construct new pseudo-residuals $e_1^*, \ldots, e_n^*$ by multiplying each residual by a random variable U such that $E\left(U_i\right) = 0$ and $Var\left(U_i\right) = 1$, for example $U_i \sim N(0,1)$:
$$e_i^* = U_i(\text{Residual})_i$$
3. Form "new" responses $y_1^*, \ldots, y_n^*$ according to
$$y_i^* = f\left((\text{Predictors})_i, \hat{\beta}\right) + e_i^*.$$
4. Obtain the least squares estimates $\hat{\alpha}$ by finding the values of α which minimize
$$\sum_{i=1}^{n} \left(y_i^* - f\left((\text{Predictors})_i, \alpha\right)\right)^2.$$

5. Repeat steps 2-4 m times.

We often take m to be large (at least 1000). After each pass through the algorithm, we retain the least squares estimates $\hat{\alpha}$ from the resample. The distribution of the estimates across the resamples is a good empirical model for the sampling distribution of the original least squares estimates.

t-Distribution (Definition 3.7) Let X be a continuous random variable. X is said to have a t-distribution if the density is given by

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \qquad x > 0$$

where $\nu > 0$ is the degrees of freedom.

We write $X \sim t_\nu$, which is read "X has a t-distribution with ν degrees of freedom."