CHAPTER 1

Introduction to Linear and Generalized Linear Models

This is a book about linear models and generalized linear models. As the names suggest, the linear model is a special case of the generalized linear model. In this first chapter, we define generalized linear models, and in doing so we also introduce the linear model.

Chapters 2 and 3 focus on the linear model. Chapter 2 introduces the least squares method for fitting the model, and Chapter 3 presents statistical inference under the assumption of a normal distribution for the response variable. Chapter 4 presents analogous model-fitting and inferential results for the generalized linear model. This generalization enables us to model non-normal responses, such as categorical data and count data.

The remainder of the book presents the most important generalized linear models. Chapter 5 focuses on models that assume a binomial distribution for the response variable. These apply to binary data, such as “success” and “failure” for possible outcomes in a medical trial or “favor” and “oppose” for possible responses in a sample survey. Chapter 6 extends the models to multicategory responses, assuming a multinomial distribution. Chapter 7 introduces models that assume a Poisson or negative binomial distribution for the response variable. These apply to count data, such as observations in a health survey on the number of respondent visits in the past year to a doctor. Chapter 8 presents ways of weakening distributional assumptions in generalized linear models, introducing quasi-likelihood methods that merely focus on the mean and variance of the response distribution. Chapters 1–8 assume independent observations. Chapter 9 generalizes the models further to permit correlated observations, such as in handling multivariate responses. Chapters 1–9 use the traditional frequentist approach to statistical inference, assuming probability distributions for the response variables but treating model parameters as fixed, unknown values. Chapter 10 presents the Bayesian approach for linear models and generalized linear models, which treats the model parameters as random variables having their
own distributions. The final chapter introduces extensions of the models that handle more complex situations, such as high-dimensional settings in which models have enormous numbers of parameters.

1.1 COMPONENTS OF A GENERALIZED LINEAR MODEL

The ordinary linear regression model uses linearity to describe the relationship between the mean of the response variable and a set of explanatory variables, with inference assuming that the response distribution is normal. Generalized linear models (GLMs) extend standard linear regression models to encompass non-normal response distributions and possibly nonlinear functions of the mean. They have three components.

- **Random component**: This specifies the response variable $y$ and its probability distribution. The observations $y = (y_1, \ldots, y_n)^T$ on that distribution are treated as independent.

- **Linear predictor**: For a parameter vector $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$ and an $n \times p$ model matrix $X$ that contains values of $p$ explanatory variables for the $n$ observations, the linear predictor is $X\beta$.

- **Link function**: This is a function $g$ applied to each component of $E(y)$ that relates it to the linear predictor,

$$g(E(y)) = X\beta.$$ 

Next we present more detail about each component of a GLM.

1.1.1 Random Component of a GLM

The random component of a GLM consists of a response variable $y$ with independent observations $(y_1, \ldots, y_n)$ having probability density or mass function for a distribution in the exponential family. In Chapter 4 we review this family of distributions, which has several appealing properties. For example, $\sum_i y_i$ is a sufficient statistic for its parameter, and regularity conditions (such as differentiation passing under an integral sign) are satisfied for derivations of properties such as optimal large-sample performance of maximum likelihood (ML) estimators.

By restricting GLMs to exponential family distributions, we obtain general expressions for the model likelihood equations, the asymptotic distributions of estimators for model parameters, and an algorithm for fitting the models. For now, it suffices to say that the distributions most commonly used in Statistics, such as the normal, binomial, and Poisson, are exponential family distributions.

The superscript $T$ on a vector or matrix denotes the transpose; for example, here $y$ is a column vector. Our notation makes no distinction between random variables and their observed values; this is generally clear from the context.
COMPONENTS OF A GENERALIZED LINEAR MODEL

1.1.2 Linear Predictor of a GLM

For observation \( i, \, i = 1, \ldots, n \), let \( x_{ij} \) denote the value of explanatory variable \( x_j \), \( j = 1, \ldots, p \). Let \( x_i = (x_{i1}, \ldots, x_{ip}) \). Usually, we set \( x_{i0} = 1 \) or let the first variable have index 0 with \( x_{i0} = 1 \), so it serves as the coefficient of an intercept term in the model. The linear predictor of a GLM relates parameters \{\eta_i\} pertaining to \{E(y_i)\} to the explanatory variables \( x_1, \ldots, x_p \) using a linear combination of them,

\[
\eta_i = \sum_{j=1}^{p} \beta_j x_{ij}, \quad i = 1, \ldots, n.
\]

The labeling of \( \sum_{j=1}^{p} \beta_j x_{ij} \) as a linear predictor reflects that this expression is linear in the parameters. The explanatory variables themselves can be nonlinear functions of underlying variables, such as an interaction term (e.g., \( x_{i3} = x_{i1} x_{i2} \)) or a quadratic term (e.g., \( x_{i2} = x_{i1}^2 \)).

In matrix form, we express the linear predictor as

\[
\eta = X \beta,
\]

where \( \eta = (\eta_1, \ldots, \eta_n)^T \), \( \beta \) is the \( p \times 1 \) column vector of model parameters, and \( X \) is the \( n \times p \) matrix of explanatory variable values \{\( x_{ij} \)\}. The matrix \( X \) is called the model matrix. In experimental studies, it is also often called the design matrix. It has \( n \) rows, one for each observation, and \( p \) columns, one for each parameter in \( \beta \). In practice, usually \( p \leq n \), the goal of model parsimony being to summarize the data using a considerably smaller number of parameters.

GLMs treat \( y_i \) as random and \( x_i \) as fixed. Because of this, the linear predictor is sometimes called the systematic component. In practice \( x_i \) is itself often random, such as in sample surveys and other observational studies. In this book, we condition on its observed values in conducting statistical inference about effects of the explanatory variables.

1.1.3 Link Function of a GLM

The third component of a GLM, the link function, connects the random component with the linear predictor. Let \( \mu_i = E(y_i), \, i = 1, \ldots, n \). The GLM links \( \eta_i \) to \( \mu_i \) by \( \eta_i = g(\mu_i) \), where the link function \( g(\cdot) \) is a monotonic, differentiable function. Thus, \( g \) links \( \mu_i \) to explanatory variables through the formula:

\[
g(\mu_i) = \sum_{j=1}^{p} \beta_j x_{ij}, \quad i = 1, \ldots, n.
\] (1.1)

In the exponential family representation of a distribution, a certain parameter serves as its natural parameter. This parameter is the mean for a normal distribution, the log of the odds for a binomial distribution, and the log of the mean for a Poisson distribution. The link function \( g \) that transforms \( \mu_i \) to the natural parameter is called the canonical link. This link function, which equates the natural parameter with the
linear predictor, generates the most commonly used GLMs. Certain simplifications result when the GLM uses the canonical link function. For example, the model has a concave log-likelihood function and simple sufficient statistics and likelihood equations.

1.1.4 A GLM with Identity Link Function is a “Linear Model”

The link function \( g(\mu_i) = \mu_i \) is called the identity link function. It has \( \eta_i = \mu_i \). A GLM that uses the identity link function is called a linear model. It equates the linear predictor to the mean itself. This GLM has

\[
\mu_i = \sum_{j=1}^{p} \beta_j x_{ij}, \quad i = 1, \ldots, n.
\]

The standard version of this, which we refer to as the ordinary linear model, assumes that the observations have constant variance, called homoscedasticity. An alternative way to express the ordinary linear model is

\[
y_i = \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i,
\]

where the “error term” \( \epsilon_i \) has \( E(\epsilon_i) = 0 \) and \( \text{var}(\epsilon_i) = \sigma^2, i = 1, \ldots, n \). This is natural for the identity link and normal responses but not for most GLMs.

In summary, ordinary linear models equate the linear predictor directly to the mean of a response variable \( y \) and assume constant variance for that response. The normal linear model also assumes normality. By contrast, a GLM is an extension that equates the linear predictor to a link-function-transformed mean of \( y \), and assumes a distribution for \( y \) that need not be normal but is in the exponential family. We next illustrate the three components of a GLM by introducing three of the most important GLMs.

1.1.5 GLMs for Normal, Binomial, and Poisson Responses

The class of GLMs includes models for continuous response variables. Most important are ordinary normal linear models. Such models assume a normal distribution for the random component, \( y_i \sim N(\mu_i, \sigma^2) \) for \( i = 1, \ldots, n \). The natural parameter for a normal distribution is the mean. So, the canonical link function for a normal GLM is the identity link, and the GLM is then merely a linear model. In particular, standard regression and analysis of variance (ANOVA) models are GLMs assuming a normal random component and using the identity link function. Chapter 3 develops statistical inference for such normal linear models. Chapter 2 presents model fitting for linear models and shows this does not require the normality assumption.

Many response variables are binary. We represent the “success” and “failure” outcomes, such as “favor” and “oppose” responses to a survey question about legalizing
same-sex marriage, by 1 and 0. A Bernoulli trial for observation \(i\) has probabilities
\[ P(y_i = 1) = \pi_i \text{ and } P(y_i = 0) = 1 - \pi_i, \]
for which \(\mu_i = \pi_i\). This is the special case of the binomial distribution with the number of trials \(n_i = 1\). The natural parameter for the binomial distribution is \(\log[\mu_i/(1 - \mu_i)]\). This is the log odds of response outcome 1, the so-called logit of \(\mu_i\). The logit is the canonical link function for binary random components. GLMs using the logit link have the form:

\[
\log \left( \frac{\mu_i}{1 - \mu_i} \right) = \sum_{j=1}^{p} \beta_j x_{ij}, \quad i = 1, \ldots, n.
\]

They are called logistic regression models, or sometimes simply logit models. Chapter 5 presents such models. Chapter 6 introduces generalized logit models for multinomial random components, for handling categorical response variables that have more than two outcome categories.

Some response variables have counts as their possible outcomes. In a criminal justice study, for instance, each observation might be the number of times a person has been arrested. Counts also occur as entries in contingency tables. The simplest probability distribution for count data is the Poisson. It has natural parameter \(\log \mu_i\), so the canonical link function is the log link, \(\eta_i = \log \mu_i\). The model using this link function is

\[
\log \mu_i = \sum_{j=1}^{p} \beta_j x_{ij}, \quad i = 1, \ldots, n.
\]

Presented in Chapter 7, it is called a Poisson loglinear model. We will see there that a more flexible model for count data assumes a negative binomial distribution for \(y_i\).

Table 1.1 lists some GLMs presented in Chapters 2–7. Chapter 4 presents basic results for GLMs, such as likelihood equations, ways of finding the ML estimates, and large-sample distributions for the ML estimators.

### 1.1.6 Advantages of GLMs versus Transforming the Data

A traditional way to model data, introduced long before GLMs, transforms \(y\) so that it has approximately a normal conditional distribution with constant variance. Then, the least squares fitting method and subsequent inference for ordinary normal linear

<table>
<thead>
<tr>
<th>Random Component</th>
<th>Link Function</th>
<th>Model</th>
<th>Chapters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>Identity</td>
<td>Regression</td>
<td>2 and 3</td>
</tr>
<tr>
<td>Exponential family</td>
<td>Any</td>
<td>Generalized linear model</td>
<td>4</td>
</tr>
<tr>
<td>Binomial</td>
<td>Logit</td>
<td>Logistic regression</td>
<td>5</td>
</tr>
<tr>
<td>Multinomial</td>
<td>Generalized logits</td>
<td>Multinomial response</td>
<td>6</td>
</tr>
<tr>
<td>Poisson</td>
<td>Log</td>
<td>Loglinear</td>
<td>7</td>
</tr>
</tbody>
</table>

Chapter 4 presents an overview of GLMs, and the other chapters present special cases.
models presented in the next two chapters are applicable on the transformed scale. For example, with count data that have a Poisson distribution, the distribution is skewed to the right with variance equal to the mean, but $\sqrt{y}$ has a more nearly normal distribution with variance approximately equal to 1/4. For most data, however, it is challenging to find a transformation that provides both approximate normality and constant variance. The best transformation to achieve normality typically differs from the best transformation to achieve constant variance.

With GLMs, by contrast, the choice of link function is separate from the choice of random component. If a link function is useful in the sense that a linear model with the explanatory variables is plausible for that link, it is not necessary that it also stabilizes variance or produces normality. This is because the fitting process maximizes the likelihood for the choice of probability distribution for $y$, and that choice is not restricted to normality.

Let $g$ denote a function, such as the log function, that is a link function in the GLM approach or a transformation function in the transformed-data approach. An advantage of the GLM formulation is that the model parameters describe $g[E(y_i)]$, rather than $E[g(y_i)]$ as in the transformed-data approach. With the GLM approach, those parameters also describe effects of explanatory variables on $E(y_i)$, after applying the inverse function for $g$. Such effects are usually more relevant than effects of explanatory variables on $E(g(y_i))$. For example, with $g$ as the log function, a GLM with $\log[E(y_i)] = \beta_0 + \beta_1 x_{i1}$ translates to an exponential model for the mean, $E(y_i) = \exp(\beta_0 + \beta_1 x_{i1})$, but the transformed-data model $E[\log(y_i)] = \beta_0 + \beta_1 x_{i1}$ does not translate to exact information about $E(y_i)$ or the effect of $x_{i1}$ on $E(y_i)$. Also, the preferred transform is often not defined on the boundary of the sample space, such as the log transform with a count or a proportion of zero.

GLMs provide a unified theory of modeling that encompasses the most important models for continuous and discrete response variables. Models studied in this text are GLMs with normal, binomial, or Poisson random component, or with extended versions of these distributions such as the multinomial and negative binomial, or multivariate extensions of GLMs. The ML parameter estimates are computed with an algorithm that iteratively uses a weighted version of least squares. The same algorithm applies to the entire exponential family of response distributions, for any choice of link function.

1.2 QUANTITATIVE/QUALITATIVE EXPLANATORY VARIABLES AND INTERPRETING EFFECTS

So far we have learned that a GLM consists of a random component that identifies the response variable and its distribution, a linear predictor that specifies the explanatory variables, and a link function that connects them. We now take a closer look at the form of the linear predictor.

\footnote{We are not stating that a model for log-transformed data is never relevant; modeling the mean on the original scale may be misleading when the response distribution is very highly skewed and has many outliers.}
1.2.1 Quantitative and Qualitative Variables in Linear Predictors

Explanatory variables in a GLM can be

- quantitative, such as in simple linear regression models.
- qualitative factors, such as in analysis of variance (ANOVA) models.
- mixed, such as an interaction term that is the product of a quantitative explanatory variable and a qualitative factor.

For example, suppose observation $i$ measures an individual’s annual income $y_i$, number of years of job experience $x_{i1}$, and gender $x_{i2}$ ($1 = \text{female}$, $0 = \text{male}$). The linear model with linear predictor

$$
\mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2}
$$

has quantitative $x_{i1}$, qualitative $x_{i2}$, and mixed $x_{i3} = x_{i1} x_{i2}$ for an interaction term. As Figure 1.1 illustrates, this model corresponds to straight lines $\mu_i = \beta_0 + \beta_1 x_{i1}$ for males and $\mu_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_{i1}$ for females. With an interaction term relating two variables, the effect of one variable changes according to the level of the other. For example, with this model, the effect of job experience on mean annual income has slope $\beta_1$ for males and $\beta_1 + \beta_3$ for females. The special case, $\beta_3 = 0$, of a lack of interaction corresponds to parallel lines relating mean income to job experience for females and males. The further special case also having $\beta_2 = 0$ corresponds to identical lines for females and males. When we use the model to compare mean incomes for females and males while accounting for the number of years of job experience as a covariate, it is called an analysis of covariance model.

![Figure 1.1 Portrayal of linear predictor with quantitative and qualitative explanatory variables.](image-url)
A quantitative explanatory variable $x$ is represented by a single $\beta x$ term in the linear predictor and a single column in the model matrix $X$. A qualitative explanatory variable having $c$ categories can be represented by $c - 1$ indicator variables and terms in the linear predictor and $c - 1$ columns in the model matrix $X$. The R software uses as default the “first-category-baseline” parameterization, which constructs indicators for categories $2, \ldots, c$. Their parameter coefficients provide contrasts with category 1. For example, suppose racial–ethnic status is an explanatory variable with $c = 3$ categories, (black, Hispanic, white). A model relating mean income to racial–ethnic status could use

$$
\mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}
$$

with $x_{i1} = 1$ for Hispanics and 0 otherwise, $x_{i2} = 1$ for whites and 0 otherwise, and $x_{i1} = x_{i2} = 0$ for blacks. Then $\beta_1$ is the difference between the mean income for Hispanics and the mean income for blacks, $\beta_2$ is the difference between the mean income for whites and the mean income for blacks, and $\beta_1 - \beta_2$ is the difference between the mean income for Hispanics and the mean income for whites. Some other software, such as SAS, uses an alternative “last-category-baseline” default parameterization, which constructs indicators for categories $1, \ldots, c - 1$. Its parameters then provide contrasts with category $c$. All such possible choices are equivalent, in terms of having the same model fit.

Shorthand notation can represent terms (variables and their coefficients) in symbols used for linear predictors. A quantitative effect $\beta x$ is denoted by $X$, and a qualitative effect is denoted by a letter near the beginning of the alphabet, such as $A$ or $B$. An interaction is represented\(^3\) by a product of such terms, such as $A.B$ or $A.X$. The period represents forming component-wise product vectors of constituent columns from the model matrix. The crossing operator $A*B$ denotes $A + B + A.B$. Nesting of categories of $B$ within categories of $A$ (e.g., factor $A$ is states, and factor $B$ is counties within those states) is represented by $A/B = A + A.B$, or sometimes by $A + B(A)$. An intercept term is represented by 1, but this is usually assumed to be in the model unless specified otherwise. Table 1.2 illustrates some simple types of linear predictors and lists the names of normal linear models that equate the mean of the response distribution to that linear predictor.

<table>
<thead>
<tr>
<th>Linear Predictor</th>
<th>Name of Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1 + X_2 + X_3 + \cdots$</td>
<td>Multiple regression</td>
</tr>
<tr>
<td>$A$</td>
<td>One-way ANOVA</td>
</tr>
<tr>
<td>$A + B$</td>
<td>Two-way ANOVA, no interaction</td>
</tr>
<tr>
<td>$A + B + A.B$</td>
<td>Two-way ANOVA, interaction</td>
</tr>
<tr>
<td>$A + X$ or $A + X + A.X$</td>
<td>Analysis of covariance</td>
</tr>
</tbody>
</table>

\(^3\)In R, a colon is used, such as $A:B$. 

---

Table 1.2 Types of Linear Predictors for Normal Linear Models

---

INTRODUCTION TO LINEAR AND GENERALIZED LINEAR MODELS
1.2.2 Interval, Nominal, and Ordinal Variables

Quantitative variables are said to be measured on an *interval scale*, because numerical intervals separate levels on the scale. They are sometimes called *interval variables*. A qualitative variable, as represented in a model by a set of indicator variables, has categories that are treated as unordered. Such a categorical variable is called a *nominal variable*.

By contrast, a categorical variable whose categories have a natural ordering is referred to as *ordinal*. For example, attained education might be measured with the categories (<high school, high school graduate, college graduate, postgraduate degree>). Ordinal explanatory variables can be treated as qualitative by ignoring the ordering and using a set of indicator variables. Alternatively, they can be treated as quantitative by assigning monotone scores to the categories and using a single $\beta x$ term in the linear predictor. This is often done when we expect $E(y)$ to progressively increase, or progressively decrease, as we move in order across those ordered categories.

1.2.3 Interpreting Effects in Linear Models

How do we interpret the $\beta$ coefficients in the linear predictors of GLMs? Suppose the response variable is a college student’s math achievement test score $y_i$, and we fit the linear model having $x_{i1} = $ the student’s number of years of math education as an explanatory variable, $\mu_i = \beta_0 + \beta_1 x_{i1}$. Since $\beta_1$ is the slope of a straight line, we might say, “If the model holds, a one-year increase in math education corresponds to a change of $\beta_1$ in the expected math achievement test score.” However, this may suggest the inappropriate causal conclusion that if a student attains another year of math education, her or his math achievement test score is expected to change by $\beta_1$.

To validly make such a conclusion, we would need to conduct an experiment that adds a year of math education for each student and then observes the results. Otherwise, a higher mean test score at a higher math education level (if $\beta_1 > 0$) could at least partly reflect the correlation of several other variables with both test score and math education level, such as parents’ attained educational levels, the student’s IQ, GPA, number of years of science courses, etc. Here is a more appropriate interpretation: If the model holds, when we compare the subpopulation of students having a certain number of years of math education with the subpopulation having one fewer year of math education, the difference in the means of their math achievement test scores is $\beta_1$.

Now suppose the model adds $x_{i2} = $ age of student and $x_{i3} = $ mother’s number of years of math education,

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}. $$

Since $\beta_1 = \partial \mu_i / \partial x_{i1}$, we might say, “The difference between the mean math achievement test score of a subpopulation of students having a certain number of years of math education and a subpopulation having one fewer year of math education equals $\beta_1$, when we keep constant the student’s age and the mother’s math education.” Controlling variables is possible in designed experiments. But it is unnatural and
possibly inconsistent with the data for many observational studies to envision increasing one explanatory variable while keeping all the others fixed. For example, \(x_1\) and \(x_2\) are likely to be positively correlated, so increases in \(x_1\) naturally tend to occur with increases in \(x_2\). In some datasets, one might not even observe a 1-unit range in an explanatory variable when the other explanatory variables are all held constant. A better interpretation is this: “The difference between the mean math achievement test score of a subpopulation of students having a certain number of years of math education and a subpopulation having one fewer year equals \(\beta_1\), when both subpopulations have the same value for \(\beta_2\) \(x_2\) + \(\beta_3\) \(x_3\).” More concisely we might say, “The effect of the number of years of math education on the mean math achievement test score equals \(\beta_1\), adjusting for student’s age and mother’s math education.” When the model also has a qualitative factor, such as \(x_4\) = gender (1 = female, 0 = male), then \(\beta_4\) is the difference between the mean math achievement test scores for female and male students, adjusting for the other explanatory variables in the model. Analogous interpretations apply to GLMs for a link-transformed mean.

The effect \(\beta_1\) in the equation with a sole explanatory variable is usually not the same as \(\beta_1\) in the equation with multiple explanatory variables, because of factors such as confounding. The effect of \(x_1\) on \(E(y)\) will usually differ if we ignore other variables than if we adjust for them, especially in observational studies containing “lurking variables” that are associated both with \(y\) and with \(x_1\). To highlight such a distinction, it is sometimes helpful to use different notation\(^5\) for the model with multiple explanatory variables, such as

\[
\mu_i = \beta_0 + \beta_{123} x_{1i} + \beta_{213} x_{2i} + \beta_{312} x_{3i},
\]

where \(\beta_{jkl}\) denotes the effect of \(x_j\) on \(y\) after adjusting for \(x_k\) and \(x_l\).

Some other caveats: In practice, such interpretations use an estimated linear predictor, so we replace “mean” by “estimated mean.” Depending on the units of measurement, an effect may be more relevant when expressed with changes other than one unit. When an explanatory variable also occurs in an interaction, then its effect should be summarized separately at different levels of the interacting variable. Finally, for GLMs with nonidentity link function, interpretation is more difficult because \(\beta_j\) refers to the effect on \(g(\mu_i)\) rather than \(\mu_i\). In later chapters we will present interpretations for various link functions.

### 1.3 MODEL MATRICES AND MODEL VECTOR SPACES

For the data vector \(y\) with \(\mu = E(y)\), consider the GLM \(\eta = X\beta\) with link function \(g\) and transformed mean values \(\eta = g(\mu)\). For this GLM, \(y\), \(\mu\), and \(\eta\) are points in \(n\)-dimensional Euclidean space, denoted by \(\mathbb{R}^n\).

\(^4\)For linear models, Section 2.5.6 gives a technical definition of adjusting, based on removing effects of \(x_2\) and \(x_3\) by regressing both \(y\) and \(x_4\) on them.

\(^5\)Yule (1907) introduced such notation in a landmark article on regression modeling.
MODEL MATRICES AND MODEL VECTOR SPACES

1.3.1 Model Matrices Induce Model Vector Spaces

Geometrically, model matrices of GLMs naturally induce vector spaces that determine the possible $\mu$ for a model. Recall that a vector space $S$ is such that if $u$ and $v$ are elements in $S$, then so are $u + v$ and $cu$ for any constant $c$.

For a particular $n \times p$ model matrix $X$, the values of $X\beta$ for all possible vectors $\beta$ of model parameters generate a vector space that is a linear subspace of $\mathbb{R}^n$. For all possible $\beta$, $\eta = X\beta$ traces out the vector space spanned by the columns of $X$, that is, the set of all possible linear combinations of the columns of $X$. This is the column space of $X$, which we denote by $C(X)$:

$$C(X) = \{\eta : \text{there is a } \beta \text{ such that } \eta = X\beta\}.$$ 

In the context of GLMs, we refer to the vector space $C(X)$ as the model space. The $\eta$, and hence the $\mu$, that are possible for a particular GLM are determined by the columns of $X$.

Two models with model matrices $X_a$ and $X_b$ are equivalent if $C(X_a) = C(X_b)$. The matrices $X_a$ and $X_b$ could be different because of a change of units of an explanatory variable (e.g., pounds to kilograms), or a change in the way of specifying indicator variables for a qualitative predictor. On the other hand, if the model with model matrix $X_a$ is a special case of the model with model matrix $X_b$, for example, with $X_a$ obtained by deleting one or more of the columns of $X_b$, then the model space $C(X_a)$ is a vector subspace of the model space $C(X_b)$.

1.3.2 Dimension of Model Space Equals Rank of Model Matrix

Recall that the rank of a matrix $X$ is the number of vectors in a basis for $C(X)$, which is a set of linearly independent vectors whose linear combinations generate $C(X)$. Equivalently, the rank is the number of linearly independent columns (or rows) of $X$. The dimension of the model space $C(X)$ of $\eta$ values, denoted by $\dim[C(X)]$, is defined to be the rank of $X$. In all but the final chapter of this book, we assume $p \leq n$, so the model space has dimension no greater than $p$. We say that $X$ has full rank when $\text{rank}(X) = p$.

When $X$ has less than full rank, the columns of $X$ are linearly dependent, with any one column being a linear combination of the other columns. That is, there exist linear combinations of the columns that yield the $\mathbf{0}$ vector. There are then nonzero $p \times 1$ vectors $\zeta$ such that $X\zeta = \mathbf{0}$. Such vectors make up the null space of the model matrix,

$$N(X) = \{\zeta : X\zeta = \mathbf{0}\}.$$ 

When $X$ has full rank, then $\dim[N(X)] = 0$. Then, no nonzero combinations of the columns of $X$ yield $\mathbf{0}$, and $N(X)$ consists solely of the $p \times 1$ zero vector, $\mathbf{0} = (0, 0, \ldots, 0)^T$. Generally,

$$\dim[C(X)] + \dim[N(X)] = p.$$
When \( X \) has less than full rank, we will see that the model parameters \( \beta \) are not well defined. Then there is said to be aliasing of the parameters. In one way this can happen, called extrinsic aliasing, an anomaly of the data causes the linear dependence, such as when the values for one predictor are a linear combination of values for the other predictors (i.e., perfect collinearity). Another way, called intrinsic aliasing, arises when the linear predictor contains inherent redundancies, such as when (in addition to the usual intercept term) we use an indicator variable for each category of a qualitative predictor. The following example illustrates.

### 1.3.3 Example: The One-Way Layout

Many research studies have the central goal of comparing response distributions for different groups, such as comparing life-length distributions of lung cancer patients under two treatments, comparing mean crop yields for three fertilizers, or comparing mean incomes on the first job for graduating students with various majors. For \( c \) groups of independent observations, let \( y_{ij} \) denote response observation \( j \) in group \( i \), for \( i = 1, \ldots, c \) and \( j = 1, \ldots, n_i \). This data structure is called the one-way layout.

We regard the groups as \( c \) categories of a qualitative factor. For \( \mu_{ij} = E(y_{ij}) \), the GLM has linear predictor,

\[
\begin{align*}
g(\mu_{ij}) &= \beta_0 + \beta_i.
\end{align*}
\]

Let \( \mu_i \) denote the common value of \( \{\mu_{ij}, j = 1, \ldots, n_i\} \), for \( i = 1, \ldots, c \). For the identity link function and an assumption of normality for the random component, this model is the basis of the one-way ANOVA significance test of \( H_0: \mu_1 = \cdots = \mu_c \), which we develop in Section 3.2. This hypothesis corresponds to the special case of the model in which \( \beta_1 = \cdots = \beta_c \).

Let \( y = (y_{11}, \ldots, y_{1n_1}, \ldots, y_{c1}, \ldots, y_{cn_c})^T \) and \( \beta = (\beta_0, \beta_1, \ldots, \beta_c)^T \). Let \( \mathbf{1}_{n_i} \) denote the \( n_i \times 1 \) column vector consisting of \( n_i \) entries of 1, and likewise for \( \mathbf{0}_{n_i} \). For the one-way layout, the model matrix \( X \) for the linear predictor \( X \beta \) in the GLM expression \( g(\mu) = X \beta \) that represents \( g(\mu_{ij}) = \beta_0 + \beta_i \) is

\[
X = \begin{pmatrix}
1_{n_1} & 1_{n_1} & 0_{n_1} & \cdots & 0_{n_1} \\
1_{n_2} & 0_{n_2} & 1_{n_2} & \cdots & 0_{n_2} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1_{n_c} & 0_{n_c} & 0_{n_c} & \cdots & 1_{n_c}
\end{pmatrix}.
\]

This matrix has dimension \( n \times p \) with \( n = n_1 + \cdots + n_c \) and \( p = c + 1 \).

Equivalently, this parameterization corresponds to indexing the observations as \( y_h \) for \( h = 1, \ldots, n \), defining indicator variables \( x_{hi} = 1 \) when observation \( h \) is in group \( i \) and \( x_{hi} = 0 \) otherwise, for \( i = 1, \ldots, c \), and expressing the linear predictor for the link function \( g \) applied to \( E(y_h) = \mu_h \) as

\[
g(\mu_h) = \beta_0 + \beta_1 x_{h1} + \cdots + \beta_c x_{hc}.
\]

In either case, the indicator variables whose coefficients are \( \{\beta_1, \ldots, \beta_c\} \) add up to the vector \( \mathbf{1}_{n_r} \). That vector, which is the first column of \( X \), has coefficient that is
the intercept term \( \beta_0 \). The columns of \( X \) are linearly dependent, because columns 2 through \( c+1 \) add up to column 1. Here \( \beta_0 \) is intrinsically aliased with \( \sum_{i=1}^{c} \beta_i \). The parameter \( \beta_0 \) is marginal to \{\( \beta_1, \ldots, \beta_c \}\}, in the sense that the column space for the coefficient of \( \beta_0 \) in the model lies wholly in the column space for the vector coefficients of \{\( \beta_1, \ldots, \beta_c \}\}. So, \( \beta_0 \) is redundant in any explanation of the structure of the linear predictor.

Because of the linear dependence of the columns of \( X \), this matrix does not have full rank. But we can achieve full rank merely by dropping one column of \( X \), because we need only \( c-1 \) indicators to represent a \( c \)-category explanatory variable. This model with one less parameter has the same column space for the reduced model matrix.

### 1.4 IDENTIFIABILITY AND ESTIMABILITY

In the one-way layout example, let \( d \) denote any constant. Suppose we transform the parameters \( \beta \) to a new set,

\[
\beta^* = (\beta^*_0, \beta^*_1, \ldots, \beta^*_c)^T = (\beta_0 + d, \beta_1 - d, \ldots, \beta_c - d)^T.
\]

The linear predictor with this new set of parameters is

\[
g(\mu_{ij}) = \beta^*_0 + \beta^*_i = (\beta_0 + d) + (\beta_i - d) = \beta_0 + \beta_i.
\]

That is, the linear predictor \( X\beta \) for \( g(\mu) \) is exactly the same, for any value of \( d \). So, for the model as specified with \( c+1 \) parameters, the parameter values are not unique.

#### 1.4.1 Identifiability of GLM Model Parameters

For this model, because the value for \( \beta \) is not unique, we cannot estimate \( \beta \) uniquely even if we have an infinite amount of data. Whether we assume normality or some other distribution for \( y \), the likelihood equations have infinitely many solutions. When the model matrix is not of full rank, \( \beta \) is not identifiable.

**Definition.** For a GLM with linear predictor \( X\beta \), the parameter vector \( \beta \) is identifiable if whenever \( \beta^* \neq \beta \), then \( X\beta^* \neq X\beta \).

Equivalently, \( \beta \) is identifiable if \( X\beta^* = X\beta \) implies that \( \beta^* = \beta \), so this definition tells us that if we know \( g(\mu) = X\beta \) (and hence if we know \( \mu \) satisfying the model), then we can also determine \( \beta \).

For the parameterization just given for the one-way layout, \( \beta \) is not identifiable, because \( \beta = (\beta_0, \beta_1, \ldots, \beta_c)^T \) and \( \beta^* = (\beta_0 + d, \beta_1 - d, \ldots, \beta_c - d)^T \) do not have different linear predictor values. In such cases, we can obtain identifiability and eliminate the intrinsic aliasing among the parameters by redefining the linear predictor with fewer parameters. Then, different \( \beta \) values have different linear predictor values \( X\beta \), and estimation of \( \beta \) is possible.
For the one-way layout, we can either drop a parameter or add a linear constraint. That is, in \( g(\mu_{ij}) = \beta_0 + \beta_i \), we might set \( \beta_1 = 0 \) or \( \beta_c = 0 \) or \( \sum_i \beta_1 = 0 \) or \( \sum_i n_i \beta_1 = 0 \). With the first-category-baseline constraint \( \beta_1 = 0 \), we express the model as \( g(\mu) = X\beta \) with

\[
X\beta = \begin{pmatrix}
1_{n_1} & 0_{n_1} & 0_{n_1} & \cdots & 0_{n_1} \\
1_{n_2} & 1_{n_2} & 0_{n_2} & \cdots & 0_{n_2} \\
1_{n_3} & 0_{n_3} & 1_{n_3} & \cdots & 0_{n_3} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
n_{c} & 0_{n_c} & 0_{n_c} & \cdots & 1_{n_c}
\end{pmatrix}
\begin{pmatrix}
\beta_0 \\
\beta_2 \\
\beta_3 \\
\vdots \\
\beta_c
\end{pmatrix}.
\]

When used with the identity link function, this expression states that \( \mu_1 = \beta_0 \) (from the first \( n_1 \) rows of \( X \)), and for \( i > 1 \), \( \mu_i = \beta_0 + \beta_i \) (from the \( n_i \) rows of \( X \) in set \( i \)). Thus, the model parameters then represent \( \beta_0 = \mu_1 \) and \( (\beta_i = \mu_i - \mu_1) \). Under the last-category-baseline constraint \( \beta_c = 0 \), the parameters are \( \beta_0 = \mu_c \), \( (\beta_i = \mu_i - \mu_c) \). Under the constraint \( \sum_i n_i \beta_1 = 0 \), the parameters are \( \beta_0 = \bar{\mu} \) and \( (\beta_i = \mu_i - \bar{\mu}) \), where \( \bar{\mu} = (\sum_i n_i \mu_i)/n \).

A slightly more general definition of identifiability refers instead to linear combinations \( \epsilon^T \beta \) of parameters. It states that \( \epsilon^T \beta \) is identifiable if whenever \( \epsilon^T \beta^* \neq \epsilon^T \beta \), then \( X\beta^* \neq X\beta \). This definition permits a subset of the terms in \( \beta \) to be identifiable, rather than treating the entire \( \beta \) as identifiable or nonidentifiable. For example, suppose we extend the model for the one-way layout to include a quantitative explanatory variable taking value \( x_{ij} \) for observation \( j \) in group \( i \), yielding the analysis of covariance model

\[
g(\mu_{ij}) = \beta_0 + \beta_i + \gamma x_{ij}.
\]

Then, without a constraint on \( \{\beta_i\} \) or \( \beta_0 \), according to this definition \( \{\beta_i\} \) and \( \beta_0 \) are not identifiable, but \( \gamma \) is identifiable. Here, taking \( \epsilon^T \beta = \gamma \), different values of \( \epsilon^T \beta \) yield different values of \( X\beta \).

### 1.4.2 Estimability in Linear Models

In a non-full-rank model specification, some quantities are unaffected by the parameter nonidentifiability and can be estimated. In a linear model, the adjective *estimable* refers to certain quantities that can be estimated in an unbiased manner.

**Definition.** In a linear model \( E(y) = X\beta \), the quantity \( \epsilon^T \beta \) is *estimable* if there exist coefficients \( a \) such that \( E(a^T y) = \epsilon^T \beta \).

That is, some linear combination of the observations estimates \( \epsilon^T \beta \) unbiasedly.

We show now that if \( \epsilon^T \beta \) can be expressed as a linear combination of means, it is estimable. Recall that \( x_i \) denotes row \( i \) of the model matrix \( X \), corresponding to observation \( y_i \), for which \( E(y_i) = x_i \beta \). Letting \( \epsilon^T = x_i \) and taking \( a \) to be identically 0 except for a 1 in position \( i \), we have \( E(a^T y) = E(y_i) = x_i \beta = \epsilon^T \beta \) for all \( \beta \). So \( E(y_i) = \epsilon^T \beta \)
Example: Using Software to Fit a GLM

1.5 Example: Using Software to Fit a GLM

General-purpose statistical software packages, such as R, SAS, Stata, and SPSS, can fit linear models and GLMs. In each chapter of this book, we introduce an example to illustrate the concepts of that chapter. We show R code and output, but the choice of software is less important than understanding how to interpret the output, which is similar with different packages.

In R, the \texttt{lm} function fits and performs inference for normal linear models, and the \texttt{glm} function does this for GLMs\footnote{For “big data,” the \texttt{biglm} package in R has functions that fit linear models and GLMs using an iterative algorithm that processes the data in chunks.}. When the \texttt{glm} function assumes the normal distribution for $y$ and uses the identity link function, it provides the same fit as the \texttt{lm} function.

1.5.1 Example: Male Satellites for Female Horseshoe Crabs

We use software to specify and fit linear models and GLMs with data from a study of female horseshoe crabs\footnote{See \url{http://en.wikipedia.org/wiki/Horseshoe_crab} and horseshoecrab.org for details about horseshoe crabs, including pictures of their mating.} on an island in the Gulf of Mexico. During spawning season,
INTRODUCTION TO LINEAR AND GENERALIZED LINEAR MODELS

Table 1.3 Number of Male Satellites (y) by Female Crab’s Characteristics

| y  | C   | S   | W   | Wt  | y  | C   | S   | W   | Wt  | y  | C   | S   | W   | Wt  |
|----|-----|-----|-----|-----|----|-----|-----|-----|-----|----|-----|-----|-----|-----|-----|
| 8  | 2   | 3   | 28.3| 3.05| 0  | 3   | 3   | 22.5| 1.55| 9  | 1   | 1   | 26.0| 2.30|
| 4  | 3   | 3   | 26.0| 2.60| 0  | 2   | 3   | 23.8| 2.10| 0  | 3   | 2   | 24.7| 1.90|
| 0  | 3   | 3   | 25.6| 2.15| 0  | 3   | 3   | 24.3| 2.15| 0  | 2   | 3   | 25.8| 2.65|
| 0  | 4   | 2   | 21.0| 1.85| 14 | 2   | 1   | 26.0| 2.30| 8  | 1   | 1   | 27.1| 2.95|

Source: The data are courtesy of Jane Brockmann, University of Florida. The study is described in Ethology 102: 1–21 (1996). Complete data (n = 173) are in file Crabs.dat at the text website, www.stat.ufl.edu/~aa/glm/data.

C, color (1, medium light; 2, medium; 3, medium dark; 4, dark); S, spine condition (1, both good; 2, one worn or broken; 3, both worn or broken); W, carapace width (cm); Wt, weight (kg).

A female migrates to the shore to breed. With a male attached to her posterior spine, she burrows into the sand and lays clusters of eggs. The eggs are fertilized externally, in the sand beneath the pair. During spawning, other male crabs may cluster around the pair and may also fertilize the eggs. These male crabs are called satellites.

The response outcome for each of the n = 173 female crabs is her y = number of satellites. Explanatory variables are the female crab’s color, spine condition, weight, and carapace width. Table 1.3 shows a small portion of the data and the categories for color and spine condition. As you read through the discussion below, we suggest that you download the data from the text website and practice data analysis by replicating these analyses and conduct others that occur to you (including additional plots) using R or your preferred software.

We now fit some linear models and GLMs to these data. Since the data are counts, the Poisson might be the first distribution you would consider for modeling y.

```r
> Crabs <- read.table("Crabs.dat", header=T)
> attach(Crabs)
> mean(y); var(y)
[1] 2.9191
[1] 9.9120
> hist(y) # Provides a histogram display
> table(y) # Shows frequency distribution for y values
   0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15
  62 16  9 19 15 13  4  6  3  3  1  1  1  1
> fit.pois <- glm(y ~ 1, family = poisson(link=identity), data=Crabs)
> summary(fit.pois) # y ~ 1 puts only an intercept in model
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.9191     0.1299  22.47  <2e-16
```

Fitting the Poisson distribution with a GLM containing only an intercept and using the identity link function gives an estimated Poisson mean that is the sample mean 2.92, for reasons we will see in Chapter 7 on models for count data. However, the Poisson mean equals its variance, and the mode is the integer part of the mean. The
sample variance of 9.92 and the strong mode at 0 shown by the frequency distribution suggest that a Poisson assumption is inappropriate for the marginal distribution of $y$. We study more appropriate distributions for the counts in Chapter 7.

### 1.5.2 Linear Model Using Weight to Predict Satellite Counts

Of the explanatory variables, two are quantitative (width and weight) and two are ordinal categorical (color and spine condition). We begin by illustrating the use of a quantitative explanatory variable. Weight and width are very highly positively correlated, and for illustrative purposes we will use weight, in kilograms, as an explanatory variable. We first find some simple descriptive statistics:

```r
> mean(weight); sd(weight); quantile(weight, c(0, 0.25, 0.50, 0.75, 1))
[1] 2.4372
[1] 0.5770
 0% 25% 50% 75% 100% # minimum, quartiles, and maximum
1.20 2.00 2.35 2.85 5.20
> plot(weight, y) # Scatterplot of y and x = weight
```

The quantiles reveal a relatively large maximum weight, which the scatterplot in Figure 1.2 of the number of satellites against weight also highlights. That plot shows there is not a clear trend in the relation.

We next fit the linear model having a straight-line relationship between $E(y)$ and $x = weight$.

```r
> fit.weight <- lm(y ~ weight, data=Crabs)
> summary(fit.weight)

Estimate  Std. Error  t value  Pr(>|t|)
(Intercept) -1.9911  0.9710   -2.050   0.0418
weight       2.0147  0.3878    5.196 5.75e-07

---

> fit.weight2 <- glm(y ~ weight, family=gaussian(link=identity), + data=Crabs)
> summary(fit.weight2)

Estimate  Std. Error  t value  Pr(>|t|)
(Intercept) -1.9911  0.9710   -2.050   0.0418
weight       2.0147  0.3878    5.196 5.75e-07
```

The fit of an ordinary linear model is the same as the fit of the GLM using normal (Gaussian family) random component with identity link function. The fit $\hat{\mu}_i = -1.991 + 2.015x_i$, with positive estimated slope, suggests that heavier female crabs tend to have more satellites. Figure 1.2 shows the fitted line superimposed on the scatterplot.
For linear modeling, it is most common to assume a normal response distribution, with constant variance. This is not ideal for the horseshoe crab satellite counts, since they are discrete and since count data usually have variability that increases as the mean does. However, the normal assumption has the flexibility, compared with the Poisson, that the variance is not required to equal the mean. In any case, Chapter 2 shows that the linear model fit does not require an assumption of normality.

### 1.5.3 Comparing Mean Numbers of Satellites by Crab Color

To illustrate the use of a qualitative explanatory variable, we next compare the mean satellite counts for the categories of color. Color is a surrogate for the age of the crab, as older crabs tend to have a darker color. It has five categories, but no observations fell in the “light” color. Let us look at the category counts and the sample mean and variance of the number of satellites for each color category.

```r
> table(color)
color # 1 = medium light, 2 = medium, 3 = medium dark, 4 = dark 1 2 3 4
12 95 44 22
> cbind(by(y,color,mean), by(y,color,var))
[,1]   [,2]
1 4.0833 9.7197 # color 1 crabs have mean(y) = 4.08, var(y) = 9.72
2 3.2947 10.2739
3 2.2273 6.7378
4 2.0455 13.0931
```

The majority of the crabs are of medium color, and the mean response decreases as the color gets darker. There is evidence of too much variability for a Poisson distribution to be realistic for y, conditional on color.
EXAMPLE: USING SOFTWARE TO FIT A GLM

We next fit the linear model for a one-way layout with color as a qualitative explanatory factor. By default, without specification of a distribution and link function, the R `glm` function fits the normal linear model:

```r
> fit.color <- glm(y ~ factor(color)) # normal dist. is default
> summary(fit.color)

Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.0833 0.8985 4.544 1.05e-05
factor(color)2 -0.7886 0.9536 -0.827 0.4094
factor(color)3 -1.8561 1.0137 -1.831 0.0689
factor(color)4 -2.0379 1.1170 -1.824 0.0699
```

The output does not report a separate estimate for the first category of color, because that parameter is aliased with the other color parameters. To achieve identifiability, R specifies first-category-baseline indicator variables (i.e., for all but the first category). In fact, $\hat{\beta}_0 = \bar{y}_1$, $\hat{\beta}_2 = \bar{y}_2 - \bar{y}_1$, $\hat{\beta}_3 = \bar{y}_3 - \bar{y}_1$, and $\hat{\beta}_4 = \bar{y}_4 - \bar{y}_1$.

If we instead assume a Poisson distribution for the conditional distribution of the response variable, we find:

```r
> fit.color2 <- glm(y ~ factor(color), family=poisson(link=identity))
> summary(fit.color2)

Estimate Std. Error z value Pr(>|t|)
(Intercept) 4.0833 0.5833 7.000 2.56e-12
factor(color)2 -0.7886 0.6123 -1.288 0.19780
factor(color)3 -1.8561 0.6252 -2.969 0.00299
factor(color)4 -2.0379 0.6582 -3.096 0.00196
```

The estimates are the same, because the Poisson distribution also has sample means as ML estimates of $\{\mu_i\}$ for a model with a single factor predictor. However, the standard error values are much smaller than under the normal assumption. Why do you think this is? Do you think they are trustworthy?

Finally, we illustrate the simultaneous use of quantitative and qualitative explanatory variables by including both weight and color in the normal model’s linear predictor:

```r
> fit.weight.color <- glm(y ~ weight + factor(color))
> summary(fit.weight.color)

Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.8232 1.3549 -0.608 0.544
weight 1.8662 0.4018 4.645 6.84e-06
factor(color)2 -0.6181 0.9011 -0.686 0.494
factor(color)3 -1.2404 0.9662 -1.284 0.201
factor(color)4 -1.1882 1.0704 -1.110 0.269
```
Let us consider the model for this analysis and its model matrix. For response \( y_i \) for female crab \( i \), let \( x_{i1} \) denote weight, and let \( x_{ij} = 1 \) when the crab has color \( j \) and \( x_{ij} = 0 \) otherwise, for \( j = 2, 3, 4 \). Then, the model has linear predictor

\[
\mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}.
\]

The model has the form \( \mu = E(y) = X\beta \) with, using some of the observations shown in Table 1.3,

\[
y = \begin{pmatrix} 8 \\ 0 \\ 9 \\ 4 \\ \vdots \end{pmatrix}, \quad X\beta = \begin{pmatrix} 1 & 3.05 & 1 & 0 & 0 \\ 1 & 1.55 & 0 & 1 & 0 \\ 1 & 2.30 & 0 & 0 & 0 \\ 1 & 2.60 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}.
\]

From \( \hat{\beta}_1 = 1.866 \), for crabs of a particular color that differ by a kilogram of weight, the estimated mean number of satellites is nearly 2 higher for the heavier crabs. As an exercise, construct a plot of the fit and interpret the color coefficients.

We could also introduce an interaction term, letting the effect of weight vary by color. However, even for the simple models fitted, we have ignored a notable outlier—the exceptionally heavy crab weighing 5.2 kg. As an exercise, you can redo the analyses without that observation to check whether results are much influenced by it. We’ll develop better models for these data in Chapter 7.
EXERCISES

link function \( g, \frac{\partial \mu_i}{\partial x_{ij}} \) has value depending on \( g \) and \( \mu_i \) as well as \( \beta_j \). For sample data and a GLM fit, one way to summarize partial effect \( j \), adjusting for the other explanatory variables, is by \( \frac{1}{n} \sum \left( \frac{\partial \hat{\mu}_i}{\partial x_{ij}} \right) \), averaging over the \( n \) sample settings. For example, for a Poisson loglinear model, \( \frac{1}{n} \sum \left( \frac{\partial \hat{\mu}_i}{\partial x_{ij}} \right) = \hat{\beta}_j \) (Exercise 7.9).

1.6 Average causal effect: Denote two groups to be compared by \( x_1 = 0 \) and \( x_1 = 1 \). For GLMs, an alternative effect summary is the average causal effect,

\[
\frac{1}{n} \sum_{i=1}^{n} \left[ E(y_i | x_{i1} = 1, x_{i2}, \ldots, x_{ip}) - E(y_i | x_{i1} = 0, x_{i2}, \ldots, x_{ip}) \right].
\]

This uses, for each observation \( i \), the expected response for its values of \( x_{i2}, \ldots, x_{ip} \) if that observation were in group 1 and if that observation were in group 0. For a particular model fit, the sample version estimates the difference between the overall means if all subjects sampled were in group 1 and if all subjects sampled were in group 0. For observational data, this mimics a counterfactual measure to estimate if we could instead conduct an experiment and observe subjects under each treatment group, rather than have half the observations missing. See Gelman and Hill (2006, Chapters 9 and 10), Rubin (1974), and Rosenbaum and Rubin (1983).

EXERCISES

1.1 Suppose that \( y_i \) has a \( N(\mu_i, \sigma^2) \) distribution, \( i = 1, \ldots, n \). Formulate the normal linear model as a special case of a GLM, specifying the random component, linear predictor, and link function.

1.2 Link function of a GLM:
   a. Describe the purpose of the link function \( g \).
   b. The identity link is the standard one with normal responses but is not often used with binary or count responses. Why do you think this is?

1.3 What do you think are the advantages and disadvantages of treating an ordinal explanatory variable as (a) quantitative, (b) qualitative?

1.4 Extend the model in Section 1.2.1 relating income to racial–ethnic status to include education and interaction explanatory terms. Explain how to interpret parameters when software constructs the indicators using (a) first-category-baseline coding, (b) last-category-baseline coding.

1.5 Suppose you standardize the response and explanatory variables before fitting a linear model (i.e., subtract the means and divide by the standard deviations). Explain how to interpret the resulting standardized regression coefficients.

1.6 When \( X \) has full rank \( p \), explain why the null space of \( X \) consists only of the 0 vector.
1.7 For any linear model \( \mathbf{\mu} = X\mathbf{\beta} \), is the origin \( \mathbf{0} \) in the model space \( \text{C}(X) \)? Why or why not?

1.8 A model \( M \) has model matrix \( X \). A simpler model \( M' \) results from removing the final term in \( M \), and hence has model matrix \( X' \) that deletes the final column from \( X \). From the definition of a column space, explain why \( \text{C}(X') \) is contained in \( \text{C}(X) \).

1.9 For the normal linear model, explain why the expression \( y_i = \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i \) with \( \epsilon_i \sim N(0, \sigma^2) \) is equivalent to \( y_i \sim N(\sum_{j=1}^{p} \beta_j x_{ij}, \sigma^2) \).

1.10 GLMs normally use a hierarchical structure by which the presence of a higher-order term implies also including the lower-order terms. Explain why this is sensible, by showing that (a) a model that includes an \( x^2 \) explanatory variable but not \( x \) makes a strong assumption about where the maximum or minimum of \( E(y) \) occurs, (b) a model that includes \( x_1 x_2 \) but not \( x_1 \) makes a strong assumption about the effect of \( x_1 \) when \( x_2 = 0 \).

1.11 Show the form of \( X\mathbf{\beta} \) for the linear model for the one-way layout, \( E(y_{ij}) = \beta_0 + \beta_i \), using a full-rank model matrix \( X \) by employing the constraint \( \sum_i \beta_i = 0 \) to make parameters identifiable.

1.12 Consider the model for the two-way layout for qualitative factors \( A \) and \( B \),

\[
E(y_{ijk}) = \beta_0 + \beta_i + \gamma_j,
\]

for \( i = 1, \ldots, r, j = 1, \ldots, c, \text{ and } k = 1, \ldots, n \). This model is balanced, having an equal sample size \( n \) in each of the \( rc \) cells, and assumes an absence of interaction between \( A \) and \( B \) in their effects on \( y \).

a. For the model as stated, is the parameter vector identifiable? Why or why not?

b. Give an example of a quantity that is (i) not estimable, (ii) estimable. In each case, explain your reasoning.

1.13 Consider the model for the two-way layout shown in the previous exercise. Suppose \( r = 2, c = 3, \text{ and } n = 2 \).

a. Show the form of a full-rank model matrix \( X \) and corresponding parameter vector \( \mathbf{\beta} \) for the model, constraining \( \beta_1 = \gamma_1 = 0 \) to make \( \mathbf{\beta} \) identifiable. Explain how to interpret the elements of \( \hat{\mathbf{\beta}} \).

b. Show the form of a full-rank model matrix and corresponding parameter vector \( \mathbf{\beta} \) when you constrain \( \sum_i \beta_i = 0 \) and \( \sum_j \gamma_j = 0 \) to make \( \mathbf{\beta} \) identifiable. Explain how to interpret the elements of \( \hat{\mathbf{\beta}} \).

c. In the full-rank case, what is the rank of \( X \)?
EXERCISES

1.14 For the model in the previous exercise with constraints $\beta_1 = \gamma_1 = 0$, generalize the model by adding an interaction term $\delta_{ij}$.
   a. Show the new full-rank model matrix. Specify the constraints that $\{\delta_{ij}\}$ satisfy. Indicate how many parameters the $\delta_{ij}$ term represents in $\beta$.
   b. Show how to write the linear predictor using indicator variables for the factor categories, with the model parameters as coefficients of those indicators and the interaction parameters as coefficients of products of indicators.

1.15 Refer to Exercise 1.12. Now suppose $r = 2$ and $c = 4$, but observations for the first two levels of $B$ occur only at the first level of $A$, and observations for the last two levels of $B$ occur only at the second level of $A$. In the corresponding model, $E(y_{ijk}) = \beta_0 + \beta_i + \gamma_{(j)}$, $B$ is said to be nested within $A$. Specify a full-rank model matrix $X$, and indicate its rank.

1.16 Explain why the vector space of $p \times 1$ vectors $\ell$ such that $\ell^T \beta$ is estimable is $C(X^T)$.

1.17 If $A$ is a nonsingular matrix, show that $C(X) = C(XA)$. (If two full-rank model matrices correspond to equivalent models, then one model matrix is the other multiplied by a nonsingular matrix.)

1.18 For the linear model for the one-way layout, Section 1.4.1 showed the model matrix that makes parameters identifiable by setting $\beta_1 = 0$. Call this model matrix $X_1$.
   a. Suppose we instead obtain identifiability by imposing the constraint $\beta_c = 0$. Show the model matrix, say $X_c$.
   b. Show how to obtain $X_1$ as a linear transformation of $X_c$.

1.19 Consider the analysis of covariance model without interaction, denoted by $1 + X + A$.
   a. Write the formula for the model in such a way that the parameters are not identifiable. Show the corresponding model matrix.
   b. For the model parameters in (a), give an example of a characteristic that is (i) estimable, (ii) not estimable.
   c. Now express the model so that the parameters are identifiable. Explain how to interpret them. Show the model matrix when $A$ has three groups, each containing two observations.

1.20 Show the first five rows of the model matrix for (a) the linear model for the horseshoe crabs in Section 1.5.2, (b) the model for a one-way layout in Section 1.5.3, (c) the model containing both weight and color predictors.
1.21 Littell et al. (2000) described a pharmaceutical clinical trial in which 24 patients were randomly assigned to each of three treatment groups (drug A, drug B, placebo) and compared on a measure of respiratory ability (FEV1 = forced expiratory volume in 1 second, in liters). The data file \texttt{FEV.dat} at \url{www.stat.ufl.edu/~aa/glm/data} has the form shown in Table 1.4. Here, we let \( y \) be the response after 1 hour of treatment (variable \texttt{fev1} in the data file), \( x_1 \) = the baseline measurement prior to administering the drug (variable \texttt{base} in the data file), and \( x_2 \) = drug (qualitative with labels \( a \), \( b \), \( p \) in the data file). Download the data and fit the linear model for \( y \) with explanatory variables (a) \( x_1 \), (b) \( x_2 \), (c) both \( x_1 \) and \( x_2 \). Interpret model parameter estimates in each case.

Table 1.4 Part of FEV Clinical Trial Data File for Exercise 1.21

<table>
<thead>
<tr>
<th>Patient</th>
<th>Base</th>
<th>fev1</th>
<th>fev2</th>
<th>fev3</th>
<th>fev4</th>
<th>fev5</th>
<th>fev6</th>
<th>fev7</th>
<th>fev8</th>
<th>Drug</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>2.46</td>
<td>2.68</td>
<td>2.76</td>
<td>2.50</td>
<td>2.30</td>
<td>2.14</td>
<td>2.40</td>
<td>2.33</td>
<td>2.20</td>
<td>a</td>
</tr>
<tr>
<td>02</td>
<td>3.50</td>
<td>3.95</td>
<td>3.65</td>
<td>2.93</td>
<td>2.53</td>
<td>3.04</td>
<td>3.37</td>
<td>3.14</td>
<td>2.62</td>
<td>a</td>
</tr>
<tr>
<td>03</td>
<td>1.96</td>
<td>2.28</td>
<td>2.34</td>
<td>2.29</td>
<td>2.43</td>
<td>2.06</td>
<td>2.18</td>
<td>2.28</td>
<td>2.29</td>
<td>a</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>72</td>
<td>2.88</td>
<td>3.04</td>
<td>3.00</td>
<td>3.24</td>
<td>3.37</td>
<td>2.69</td>
<td>2.89</td>
<td>2.89</td>
<td>2.76</td>
<td>p</td>
</tr>
</tbody>
</table>

Complete data (file \texttt{FEV.dat}) are at the text website \url{www.stat.ufl.edu/~aa/glm/data}.

1.22 Refer to the analyses in Section 1.5.3 for the horseshoe crab satellites.

a. With color alone as a predictor, why are standard errors much smaller for a Poisson model than for a normal model? Out of these two very imperfect models, which do you trust more for judging significance of the estimates of the color effects? Why?

b. Download the data (file \texttt{Crabs.dat}) from \url{www.stat.ufl.edu/~aa/glm/data}. When weight is also a predictor, identify an outlying observation. Refit the model with color and weight predictors without that observation. Compare results, to investigate the sensitivity of the results to this outlier.

1.23 Another horseshoe crab dataset \(^9\) (\texttt{Crabs2.dat} at \url{www.stat.ufl.edu/~aa/glm/data}) comes from a study of factors that affect sperm traits of male crabs. A response variable, \textit{SpermTotal}, is measured as the log of the total number of sperm in an ejaculate. It has mean 19.3 and standard deviation 2.0. Two explanatory variables are the crab’s carapace width (in centimeters, with mean 18.6 and standard deviation 3.0) and color (1 = dark, 2 = medium, 3 = medium).

\(^8\) Thanks to Ramon Littell for making these data available.
\(^9\) Thanks to Jane Brockmann and Dan Sasson for making these data available.
EXERCISES

3 = light). Explain how to interpret the estimates in the following table. Is the model fitted equivalent to a GLM with the log link for the expected number of sperm? Why or why not?

```
> summary(lm(SpermTotal ~ CW + factor(Color)))
Coefficients:           Estimate Std. Error t value Pr(>|t|)
(Intercept)    11.366      0.638   17.822  < 2e-16
CW             0.391       0.034   11.651  < 2e-16
factor(Color)2 0.809       0.246    3.292  0.00114
factor(Color)3 1.149       0.271    4.239  3.14e-05
```

1.24 For 72 young girls suffering from anorexia, the Anorexia.dat file at the text website shows their weights before and after an experimental period. Table 1.5 shows the format of the data. The girls were randomly assigned to receive one of three therapies during this period. A control group received the standard therapy, which was compared to family therapy and cognitive behavioral therapy. Download the data and fit a linear model relating the weight after the experimental period to the initial weight and the therapy. Interpret estimates.

| Table 1.5  Weights of Anorexic Girls, in Pounds, Before and After Receiving One of Three Therapies |
|------------------|-------------------|-------------------|-------------------|
| Cognitive Behavioral | Family Therapy | Control | |
| Weight Before | Weight After | Weight Before | Weight After | Weight Before | Weight After |
| 80.5 | 82.2 | 83.8 | 95.2 | 80.7 | 80.2 |
| 84.9 | 85.6 | 83.3 | 94.3 | 89.4 | 80.1 |
| 81.5 | 81.4 | 86.0 | 91.5 | 91.8 | 86.4 |

Source: Thanks to Brian Everitt for these data. Complete data are at text website.