# Chapter 1

# Preliminaries

The purpose of an exploration of data may be rather limited and ad hoc, or the purpose may be more general, perhaps to gain understanding of some natural phenomenon. The questions addressed may be somewhat open-ended. The process of understanding often begins with general questions about the structure of the data. At any stage of the analysis, our understanding is facilitated by means of a *model*.

A model is a description that embodies our current understanding of a phenomenon. In an operational sense, we can formulate a model either as a description of a *data generating process*, or as a prescription for processing data. The model is often expressed as a set of equations that relate data elements to each other. It may include probability distributions for the data elements. If any of the data elements are considered to be realizations of random variables, the model is a *stochastic model*.

A model should not limit our analysis; rather the model should be able to evolve. The process of understanding involves successive refinements of the model. The refinements proceed from vague models to more specific ones. An exploratory data analysis may begin by mining the data to identify interesting properties. These properties generally raise questions that are to be explored further.

A class of models may have a common form, within which the members of the class are distinguished by values of *parameters*. For example, the class of normal probability distributions has a single form of a probability density function that has two parameters. If this form of model is chosen to represent the properties of a dataset, we may seek confidence intervals for values of the two parameters, or we may perform statistical tests of hypothesized values of the parameters.

In models that are not as mathematically tractable as the normal probability model — and many realistic models are not — computationally-intensive methods involving simulations, resamplings, and multiple views may be used to make inferences about the parameters of a model.

## 1.1 Discovering Structure: Data Structures and Structure in Data

The components of statistical datasets are "observations" and "variables". In general, "data structures" are ways of organizing data to take advantage of the relationships among the variables constituting the dataset. Data structures may express hierarchical relationships, crossed relationships (as in "relational" databases), or more complicated aspects of the data (as in "object-oriented" databases).

In data analysis, "structure in the data" is of interest. Structure in the data includes such nonparametric features as modes, gaps, or clusters in the data, the symmetry of the data, and other general aspects of the shape of the data. Because many classical techniques of statistical analysis rely on an assumption of normality of the data, the most interesting structure in the data may be those aspects of the data that deviate most from normality.

Sometimes it is possible to express the structure in the data in terms of mathematical models. Prior to doing this, graphical displays may be used to discover qualitative structure in the data. Patterns observed in the data may suggest explicit statements of the structure or of relationships among the variables on the dataset. The process of building models of relationships is an iterative one, and graphical displays are useful throughout the process. Graphs comparing data and the fitted models are used to refine the models.

### Multiple Analyses and Multiple Views

Effective use of graphics often requires multiple views. For multivariate data, plots of individual variables or combinations of variables can be produced quickly and used to get a general idea of the properties of the data. The data should be inspected from various perspectives. Instead of a single histogram to depict the general shape of univariate data, for example, multiple histograms with different bin widths and different bin locations may provide more insight.

Sometimes a few data points in a display can completely obscure interesting structure in the other data points. A zooming window to restrict the scope of the display and simultaneously restore the scale to an appropriate viewing size can reveal structure. A zooming window can be used with any graphics software whether the software supports it or not; zooming can be accomplished by deletion of the points in the dataset outside of the window.

Scaling the axes can also be used effectively to reveal structure. The relative scales is called the "aspect ratio". In Figure 1.1, which is a plot of a bivariate dataset, we form a zooming window that deletes a single observation. The greater magnification and the changed aspect ratio clearly shows a relationship between $X$ and $Y$ in a region close to the origin that may not hold for the full range of data. A simple statement of this relationship would not extrapolate outside the window to the outlying point.

The use of a zooming window is not "deletion of outliers"; it is focusing

in on a subset of the data, and is done independently of whatever is believed about the data outside of the window.

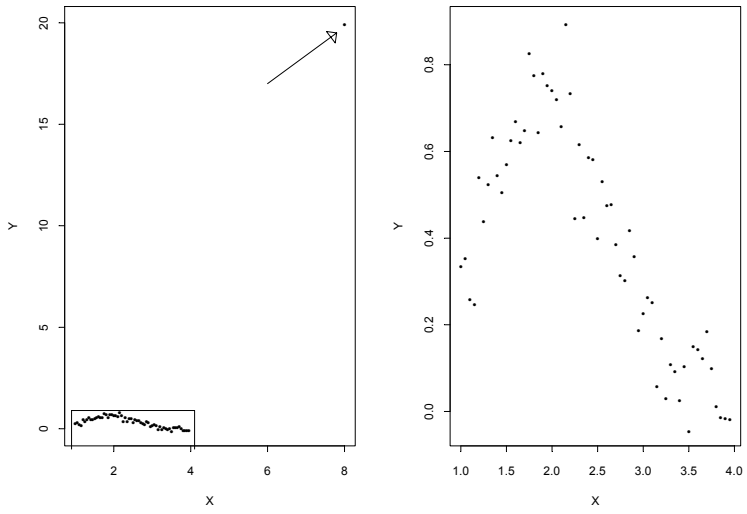


Figure 1.1: Scales Matter

One type of structure that may go undetected is that arising from the order in which the data were collected. For data that are recognized as a time series by the analyst, this is obviously not a problem, but often there is a time-dependency in the data that is not recognized immediately. "Time" or "location" may not be an explicit variable on the data set, even though it may be an important variable. The index of the observation within the dataset may be a surrogate variable for time, and characteristics of the data may vary as the index varies. Often it is useful to make plots in which one axis is the index number of the observations. More subtle time-dependencies are those in which the values of the variables are not directly related to time, but relationships among variables are changing over time. The identification of such time-dependencies is much more difficult, and often requires fitting a model and plotting residuals. Another strictly graphical way of observing changes in relationships over time is by use of a sequence of graphical displays.

### Simple Plots May Reveal the Unexpected

A simple plot of the data will often reveal structure or other characteristics of the data that numerical summaries do not.

An important property of data that is often easily seen in a graph is the unit of measurements. Data on continuous variables are often rounded or measured

on a coarse grid. This may indicate other problems in the collection of the data. The horizontal lines in Figure 1.2 indicate that the data do not come from a continuous distribution. Whether or not we can use methods of data analysis that assume continuity depends on the coarseness of the grid or measurement; that is, on the extent to which the data are discrete or the extent to which they have been discretized.

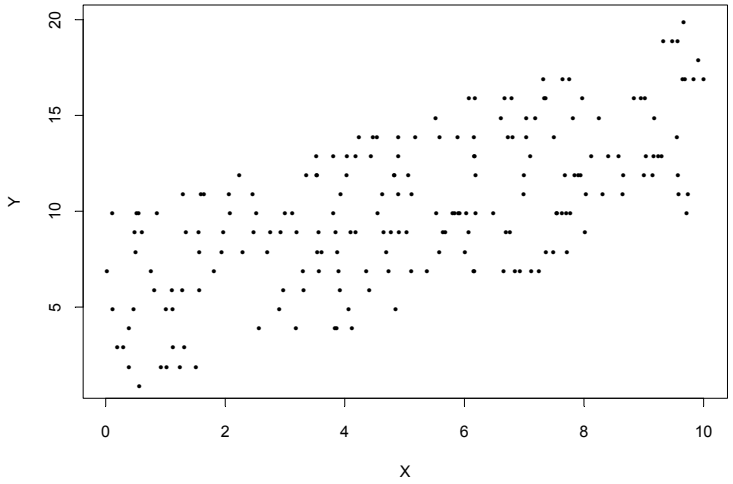


Figure 1.2: Discrete Data, Rounded Data, or Data Measured Imprecisely

We discuss graphics further in Chapter 7. The emphasis is on the use of graphics for discovery. The field of statistical graphics is much broader, of course, and includes many issues of design of graphical displays for conveying (rather than discovering) information.

## 1.2 Modeling and Computational Inference

The process of building models involves successive refinements. The evolution of the models proceeds from vague tentative models to more complete ones, and our understanding of the process being modeled grows in this process.

The usual statements about statistical methods regarding bias, variance, and so on are made in the context of a model. It is not possible to measure bias or variance of a procedure to *select* a model, except in the relatively simple case of selection from some well-defined and simple set of possible models. Only within the context of rigid assumptions (a "metamodel") can we do a precise statistical analysis of model selection. Even the simple cases of selection of variables in

linear regression analysis under the usual assumptions about the distribution of residuals (and this is a highly idealized situation), present more problems to the analyst than are generally recognized. See Kennedy and Bancroft (1971) and Speed and Yu (1993), for example, for some discussions of these kinds of problems in regression model building.

### Descriptive Statistics, Inferential Statistics, and Model Building

We can distinguish statistical activities that involve

- data collection,
- descriptions of a given dataset,
- inference within the context of a model or family of models, and
- model selection.

In any given application, it is likely that all of these activities will come into play. Sometimes (and often, ideally!) a statistician can specify how data are to be collected, either in surveys or in experiments. We will not be concerned with this aspect of the process in this text.

Once data are available, either from a survey or designed experiment, or just observational data, a statistical analysis begins by consideration of general descriptions of the dataset. These descriptions include ensemble characteristics, such as averages and spreads, and identification of extreme points. The descriptions are in the form of various summary statistics and of graphical displays. The descriptive analyses may be computationally intensive for large datasets, especially if there is a large number of variables. The computationally-intensive approach also involves multiple views of the data, including consideration of a large number transformations of the data. We discuss these methods in Chapters 5 and 7, and in Part II.

A stochastic model is often expressed as a probability density function or as a cumulative distribution function of a random variable. In a simple linear regression model with normal errors,

$$Y = \beta_0 + \beta_1 x + E,$$

for example, the model may be expressed by use of the probability density function for the random variable $E$. (Notice that $Y$ and $E$ are written in upper case because they represent random variables.) The probability density function for $Y$ is

$$p(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(y-\beta_0-\beta_1 x)^2/(2\sigma^2)}.$$

In this model, $x$ is an observable covariate; $\sigma$, $\beta_0$, and $\beta_1$ are unobservable (and, generally, unknown) parameters; and 2 and $\pi$ are constants. Statistical inference about parameters includes estimation or tests of their values or statements

about their probability distributions based on observations of the elements of the model.

The elements of a stochastic model include observable random variables, observable covariates, unobservable parameters, and constants. Some random variables in the model may be considered to be "responses". The covariates may be considered to affect the response; they may or may not be random variables. The parameters are variable within a class of models, but for a specific data model the parameters are constants. The parameters may be considered to be unobservable random variables, and in that sense, a specific data model is defined by a realization of the parameter random variable. In the model, written as

$$Y = f(x; \beta) + E,$$

we identify a "systematic component", $f(x; \beta)$, and a "random component", $E$. The selection of an appropriate model may be very difficult, and almost always involves not only questions of how well the model corresponds to the observed data, but also how tractable is the model. The methods of computational statistics allow a much wider range of tractability than can be comtemplated in mathematical statistics.

Statistical analyses generally are undertaken with the purpose of making a decision about a dataset or about a population from which a sample dataset is available, or in making a prediction about a future event. Much of the theory of statistics developed during the middle third of the twentieth century was concerned with formal inference; that is, use of a sample to make decisions about stochastic models based on probabilities that would result if a given model was indeed the data generating process. The heuristic paradigm calls for rejection of a model if the probability is small that data arising from the model would be similar to the observed sample. This process can be quite tedious because of the wide range of models that should be explored, and because some of the models may not yield mathematically tractable estimators or test statistics. Computationally-intensive methods include exploration of a range of models, many of which may be mathematically intractable.

In a different approach employing the same paradigm, the statistical methods may involve direct simulation of the hypothesized data generating process, rather than formal computations of probabilities that would result under a given model of the data generating process. We refer to this approach as *computational inference*. We discuss methods of computational inference in Chapters 2, 3, and 4. In a variation of computational inference, we may not even attempt to develop a model of the data generating process; rather, we build decision rules directly from the data. This is often the approach in clustering and classification, which we discuss in Chapter 10. Computational inference is rooted in classical statistical inference. In subsequent sections of the current chapter we discuss general techniques used in statistical inference.

## 1.3 The Role of the Empirical Cumulative Distribution Function

Methods of statistical inference are based on an assumption (often implicit) that a discrete uniform distribution with mass points at the observed values of ta random sample is asymptotically the same as the distribution governing the data generating process. Thus, the distribution function of this discrete uniform distribution is a model of the distribution function of the data generating process.

For a given set of univariate data, $y_1, y_2, \ldots, y_n$, the *empirical cumulative distribution function*, or *ECDF*, is

$$P_n(y) = \frac{\#\{y_i, \text{ s.t. } y_i \leq y\}}{n}.$$

The ECDF is the basic function used in many methods of computational inference.

Although the ECDF has similar definitions for univariate and multivariate random variables, it is most useful in the univariate case. An equivalent expression for univariate random variables, in terms of intervals on the real line, is

$$P_n(y) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}_{(-\infty,y]}(y_i), \tag{1.1}$$

where I is the indicator function. (See page 363 for the definition and some of the properties of the indicator function. The measure $\mathrm{dI}_{(-\infty,a]}(x)$, which we use in equation (1.6) for example, is particularly interesting.)

It is easy to see that the ECDF is pointwise unbiased for the CDF, that is, for given $y$,

$$\begin{aligned} \mathrm{E}\big(P_n(y)\big) &= \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} \mathrm{I}_{(-\infty,y]}(Y_i)\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left(\mathrm{I}_{(-\infty,y]}(Y_i)\right) \\ &= \Pr\left(Y \leq y\right) \\ &= P(y). \end{aligned} \tag{1.2}$$

Similarly, we find

$$\mathrm{V}\big(P_n(y)\big) = P(y)\big(1 - P(y)\big)/n; \tag{1.3}$$

indeed, at a fixed point $y$, $nP_n(y)$ is a binomial random variable with parameters $n$ and $\pi = P(y)$. Because $P_n$ is a function of the order statistics, which form a complete sufficient statistic for $P$, there is no unbiased estimator of $P(y)$ with smaller variance.

We also define the *empirical probability density function* (EPDF) as the derivative of the ECDF:

$$p_n(y) = \frac{1}{n}\sum_{i=1}^{n}\delta(y - y_i), \tag{1.4}$$

where $\delta$ is the Dirac delta function. The EPDF is just a series of spikes at points corresponding to the observed values. It is not as useful as the ECDF. It is, however, unbiased at any point for the probability density function at that point.

The ECDF and the EPDF can be used as estimators of the corresponding population functions, but there are better estimators. See Chapter 9.

### Statistical Functions of the CDF and the ECDF

In many models of interest, a parameter can be expressed as a functional of the probability density function or of the cumulative distribution function of a random variable in the model. The mean of a distribution, for example, can be expressed as a functional $\Theta$ of the CDF $P$:

$$\Theta(P) = \int_{\mathbb{R}^d} y \,\mathrm{d}P(y). \tag{1.5}$$

A functional that defines a parameter is called a *statistical function*.

### Estimation of Statistical Functions

A common task in statistics is to use a random sample to estimate the parameters of a probability distribution. If the statistic $T$ from a random sample is used to estimate the parameter $\theta$, we measure the performance of $T$ by the magnitude of the bias,

$$|\mathrm{E}(T) - \theta|,$$

by the variance,

$$\mathrm{V}(T) = \mathrm{E}\Big(\big(T - \mathrm{E}(T)\big)^2\Big),$$

by the mean squared error,

$$\mathrm{E}\Big((T - \theta)^2\Big),$$

and by other expected values of measures of the distance from $T$ to $\theta$. (These expressions are for the scalar case, but similar expressions apply to vectors $T$ and $\theta$, in which case the bias is a vector, the variance is the variance-covariance matrix, and the mean squared error is a dot product, hence, a scalar.)

If $\mathrm{E}(T) = \theta$, $T$ is unbiased for $\theta$. For sample size $n$, if $\mathrm{E}(T) = \theta + \mathrm{O}(n^{-1/2})$, $T$ is said to be *first-order accurate* for $\theta$; if $\mathrm{E}(T) = \theta + \mathrm{O}(n^{-1})$, it is *second-order accurate*. (See page 363 for definition of $\mathrm{O}(\cdot)$. Convergence of $\mathrm{E}(T)$ can also be expressed as a stochastic convergence of $T$, in which case we use the notation $\mathrm{O_P}(\cdot)$.)

The order of the mean squared error is an important characteristic of an estimator. For good estimators of location, the order of the mean squared error is typically $\mathrm{O}(n^{-1})$. Good estimators of probability densities, however, typically have mean squared errors of at least order $\mathrm{O}(n^{-4/5})$ (see Chapter 9).

### Estimation Using the ECDF

There are many ways to construct an estimator and to make inferences about the population. In the univariate case especially, we often use data to make inferences about a parameter by applying the statistical function to the ECDF. An estimator of a parameter that is defined in this way is called a *plug-in estimator*. A plug-in estimator for a given parameter is the same functional of the ECDF as the parameter is of the CDF.

For the mean of the model, for example, we use the estimate that is the same functional of the ECDF as the population mean in equation (1.5),

$$\begin{aligned}
\Theta(P_n) &= \int_{-\infty}^{\infty} y \,\mathrm{d}P_n(y) \\
&= \int_{-\infty}^{\infty} y \,\mathrm{d}\frac{1}{n}\sum_{i=1}^{n} \mathrm{I}_{(-\infty,y]}(y_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty} y \,\mathrm{d}\mathrm{I}_{(-\infty,y]}(y_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} y_i \\
&= \bar{y}. \qquad (1.6)
\end{aligned}$$

The sample mean is thus a plug-in estimator of the population mean. Such an estimator is called a *method of moments estimator*. This is an important type of plug-in estimator. The method of moments results in estimates of the parameters $\mathrm{E}(Y^r)$ that are the corresponding sample moments.

Statistical properties of plug-in estimators are generally relatively easy to determine. In some cases, the statistical properties, such as expectation and variance, are optimal in some sense.

In addition to estimation based on the ECDF, other methods of computational statistics make use of the ECDF. In some cases, such as in bootstrap methods, the ECDF is a surrogate for the CDF. In other cases, such as Monte Carlo methods, an ECDF for an estimator is constructed by repeated sampling, and that ECDF is used to make inferences using the observed value of the estimator from the given sample.

Viewed as a statistical function, $\Theta$ denotes a specific functional form. Any functional of the ECDF is a function of the data, so we may also use the notation $\Theta(Y_1, Y_2, \ldots, Y_n)$. Often, however, the notation is cleaner if we use another letter to denote the function of the data; for example, $T(Y_1, Y_2, \ldots, Y_n)$, even

if it might be the case that

$$T(Y_1, Y_2, \ldots, Y_n) = \Theta(P_n).$$

We will also often use the same letter that denotes the functional of the sample to represent the random variable computed from a random sample; that is, we may write

$$T = T(Y_1, Y_2, \ldots, Y_n).$$

As usual, we will use $t$ to denote a realization of the random variable $T$.

Use of the ECDF in statistical inference does not require many assumptions about the distribution. Other methods discussed below are based on information or assumptions about the data generating process.

### Empirical Quantiles

For $\alpha \in (0, 1)$, the $\alpha$ *quantile* of the distribution with CDF $P$ is the value $y_{(\alpha)}$ such that $P(y_{(\alpha)}) = \alpha$. (For a univariate random variable, this is a single point. For a $d$-variate random variable, it is a $(d-1)$-dimensional object that is generally nonunique.) For a discrete distribution the quantile may not exist for a given value of $\alpha$.

If $P(y) = \alpha$ we say the *quantile of* $y$ *is* $\alpha$.

If the underlying distribution is discrete, the above definition of a quantile applied to the ECDF is also meaningful. If the distribution is continuous, however, it is likely that the range of the distribution extends beyond the smallest and largest observed values. For a sample from a continuous distribution, the definition of a quantile applied to the ECDF leads to a quantile of 0 for the smallest sample value $y_{(1)}$, and a quantile of 1 for the largest sample value $y_{(n)}$. These values for quantiles are not so useful if the distribution is continuous. We define the *empirical quantile*, or *sample quantile*, corresponding to the $i^{\text{th}}$ order statistic, $y_{(i)}$, in a sample of size $n$ as

$$\frac{i - \iota}{n + \nu}, \tag{1.7}$$

for $\iota, \nu \in [0, \frac{1}{2}]$. Values of $\iota$ and $\nu$ that make the empirical quantiles of a random sample correspond closely to those of the population depend on the distribution of the population, which, of course, is generally unknown. A certain symmetry may be imposed by requiring $\nu = 1 - 2\iota$. Common choices are $\iota = \frac{1}{2}$ and $\nu = 0$.

We use empirical quantiles in Monte Carlo inference, in nonparametric inference, and in graphical displays for comparing a sample with a standard distribution or with another sample. Empirical quantiles can be used as estimators of the population quantiles, but there are other estimators. Some, such as the Kaigh-Lachenbruch estimator and the Harrell-Davis estimator (see Kaigh and Lachenbruch, 1982, and Harrell and Davis, 1982), use a weighted combination of multiple data points instead of just a single one, as in the simple estimators

above. See Dielman, Lowry, and Pfaffenberger (1994) for comparisons of various quantile estimators. If a covariate is available, it may be possible to use it to improve the quantile estimate. This is often the case in simulation studies. See Hesterberg and Nelson (1998) for discussion of this technique.

## 1.4 The Role of Optimization in Inference

Important classes of estimators are defined as the point at which some function that involves the parameter and the random variable achieves an optimum. There are, of course, many functions that involve the parameter and the random variable; an example is the probability density. In the use of optimization in inference, once the objective function is chosen (it must have a known form), observations on the random variable are taken and then considered to be fixed, and the parameter in the function is considered to be a variable. The function is then optimized with respect to the parameter variable. The nature of the function determines the meaning of "optimized"; if the function is the probability density, for example, "optimized" would logically mean "maximized". (This leads to maximum likelihood estimation, which we discuss below.)

In discussing this approach to estimation, we must be careful to distinguish between a symbol that represents a fixed parameter and a symbol that represents a "variable" parameter. When we denote a probability density function as $p(y \mid \theta)$, we generally expect "$\theta$" to represent a fixed, but possibly unknown, parameter. In an estimation method that involves optimizing this function, however, "$\theta$" is a variable placeholder. In the following discussion, we will generally consider $\theta$ to be a variable. We will use $\theta_*$ to represent the true value of the parameter on which the random variable observed is conditioned. We may also use $\theta_0$, $\theta_1$, and so on to represent specific fixed values of the variable. In an iterative algorithm, we use $\theta^{(k)}$ to represent a fixed value in the $k^{\text{th}}$ iteration.

### Estimation by Minimizing Residuals

In many applications we can express the expected value of a random variable as a function of a parameter (which might be a vector, of course):

$$\mathrm{E}(Y) = f(\theta_*). \tag{1.8}$$

The expectation may also involve covariates, so in general we may write $f(x, \theta_*)$. The standard linear regression model is an example: $\mathrm{E}(Y) = x^{\mathrm{T}}\beta$. If the covariates are observable, they can be subsumed into $f(\theta)$.

The more difficult and interesting problems of course involve the determination of the form of the function $f(\theta)$. In these sections, however, we concentrate on the simpler problem of determining an appropriate value of $\theta$, assuming the form is known.

If we can obtain observations $y_1, y_2, \ldots, y_n$ on $Y$ (and observations on the covariates if there are any), a reasonable estimator of $\theta_*$ is a value $\widehat{\theta}$ that

minimizes the residuals,

$$r_i(\theta) = y_i - f(\theta), \tag{1.9}$$

over all possible choices of $\theta$. This is a logical approach because we expect the observed $y$'s to be close to $f(\theta_*)$.

There are, of course, several ways we could reasonably "minimize the residuals". In general, we seek to minimize some norm of $r(\theta)$, the $n$-vector of residuals. The optimization problem is

$$\min_{\theta} \|r(\theta)\|. \tag{1.10}$$

We often choose the norm as the $\mathrm{L}_p$ norm, so we minimize a function of an $\mathrm{L}_p$ norm of the residuals,

$$s_p(\theta) = \sum_{i=1}^{n} |y_i - f(\theta)|^p, \tag{1.11}$$

for some $p > 1$, to obtain an $\mathrm{L}_p$ *estimator*. Simple choices are the sum of the absolute values and the sum of the squares. The latter choice yields the *least squares estimator*. More generally, we could minimize

$$s_\rho(\theta) = \sum_{i=1}^{n} \rho(y_i - f(\theta))$$

for some nonnegative function $\rho(\cdot)$, to obtain an *"M estimator"*. (The name comes from the similarity of this objective function to the objective function for some maximum likelihood estimators.)

Standard techniques for optimization can be used to determine estimates that minimize various functions of the residuals, that is, for some appropriate function of the residuals $s(\cdot)$, to solve

$$\min_{\theta} s(\theta). \tag{1.12}$$

Except for special forms of the objective function, the algorithms to solve (1.12) are iterative. If $s$ is twice differentiable, one algorithm is Newton's method, in which the minimizing value of $\theta$, $\widehat{\theta}$, is obtained as a limit of the iterates

$$\theta^{(k)} = \theta^{(k-1)} - \Big(\mathrm{H}_s\big(\theta^{(k-1)}\big)\Big)^{-1} \nabla s\big(\theta^{(k-1)}\big), \tag{1.13}$$

where $\mathrm{H}_s(\theta)$ denotes the Hessian of $s$ and $\nabla s(\theta)$ denotes the gradient of $s$, both evaluated at $\theta$. (Newton's method is sometimes called the Newton-Raphson method, for no apparent reason.)

The function $s(\cdot)$ is usually chosen to be differentiable, at least piecewise.

For various computational considerations, instead of the exact Hessian, a matrix $\widetilde{H}_s$ approximating the Hessian is often used. In this case the technique is called a quasi-Newton method.

Newton's method or a quasi-Newton method often overshoots the best step. The *direction*

$$\theta^{(k)} - \theta^{(k-1)}$$

may be the best direction, but the distance

$$\|\theta^{(k)} - \theta^{(k-1)}\|$$

may be too great. A variety of methods using Newton-like iterations involve a system of equations of the form

$$\widetilde{H}_s(\theta)\, d = \nabla s(\theta). \tag{1.14}$$

These equations are solved for the direction $d$, and the new point is taken as the old $\theta$ plus $\alpha d$, for some damping factor $\alpha$.

There are various ways of deciding when an iterative optimization algorithm has converged. In general, convergence criteria are based on the size of the change in $\theta^{(k)}$ from $\theta^{(k-1)}$, or the size of the change in $s(\theta^{(k)})$ from $s(\theta^{(k-1)})$. See Kennedy and Gentle (1980), page 435 and following, for discussion of termination criteria in multivariate optimization.

### Statistical Properties of Minimum-Residual Estimators

It is generally difficult to determine the variance or other high-order statistical properties of an estimator defined as above; that is, defined as the minimizer of some function of the residuals. In many cases all that is possible is to approximate the variance of the estimator in terms of some relationship that holds for a normal distribution. (In robust statistical methods, for example, it is common to see a "scale estimate" expressed in terms of some mysterious constant times a function of some transformation of the residuals.)

There are two issues that affect both the computational method and the statistical properties of the estimator defined as the solution to the optimization problem. One consideration has to do with the acceptable values of the parameter $\theta$. In order for the model to make sense, it may be necessary that the parameter be in some restricted range. In some models, a parameter must be positive, for example. In these cases the optimization problem has constraints. Such a problem is more difficult to solve than an unconstrained problem. Statistical properties of the solution are also more difficult to determine. More extreme cases of restrictions on the parameter may require the parameter to take values in a countable set. Obviously, in such cases Newton's method cannot be used. Rather, a combinatorial optimization algorithm must be used. A function that is not differentiable also presents problems for the optimization algorithm.

Secondly, it may turn out that the optimization problem (1.12) has local minima. This depends on the nature of the function $f(\cdot)$ in equation (1.8). Local minima present problems for the computation of the solution, because the algorithm may get stuck in a local optimum. Local minima also present

conceptual problems concerning the appropriateness of the estimation criterion itself. So long as there is a unique global optimum, it seems reasonable to seek it and to ignore local optima. It is not so clear what to do if there are multiple points at which the global optimum is attained.

**Least-Squares Estimation**

Least-squares estimators are generally more tractable than estimators based on other functions of the residuals. They are more tractable both in terms of solving the optimization problem to obtain the estimate and in approximating statistical properties of the estimators, such as their variances.

Assume $\theta$ is an $m$-vector and assume $f(\cdot)$ is a smooth function. Letting $y$ be the $n$-vector of observations, we can write the least squares objective function corresponding to equation (1.11) as

$$s(\theta) = \big(r(\theta)\big)^{\mathrm{T}} r(\theta), \tag{1.15}$$

where the superscript T indicates the transpose of a vector or of a matrix.

The gradient and the Hessian for a least squares problem have special structures that involve the Jacobian of the residuals, $\mathrm{J}_r(\theta)$. The gradient of $s$ is

$$\nabla s(\theta) = \big(\mathrm{J}_r(\theta)\big)^{\mathrm{T}} r(\theta). \tag{1.16}$$

Taking derivatives of $\nabla s(\theta)$, we see the Hessian of $s$ can be written in terms of the Jacobian of $r$ and the individual residuals:

$$\mathrm{H}_s(\theta) = \big(\mathrm{J}_r(\theta)\big)^{\mathrm{T}} \mathrm{J}_r(\theta) + \sum_{i=1}^{n} r_i(\theta) \mathrm{H}_{r_i}(\theta). \tag{1.17}$$

In the vicinity of the solution $\widehat{\theta}$, the residuals $r_i(\theta)$ should be small, and $\mathrm{H}_s(\theta)$ may be approximated by neglecting the second term:

$$\mathrm{H}_s(\theta) \approx \big(\mathrm{J}_r(\theta)\big)^{\mathrm{T}} \mathrm{J}_r(\theta).$$

Using (1.16) and this approximation for (1.17) in the gradient descent equation (1.14), we have the system of equations

$$\big(\mathrm{J}_r(\theta^{(k-1)})\big)^{\mathrm{T}} \mathrm{J}_r(\theta^{(k-1)})\, d^{(k)} \;=\; -\big(\mathrm{J}_r(\theta^{(k-1)})\big)^{\mathrm{T}} r(\theta^{(k-1)}) \tag{1.18}$$

to be solved for $d^{(k)}$, where

$$d^{(k)} \;\propto\; \theta^{(k)} - \theta^{(k-1)}.$$

It is clear that the solution $d^{(k)}$ is a descent direction; that is, if $\nabla s(\theta^{(k-1)}) \neq 0$,

$$\begin{aligned} (d^{(k)})^{\mathrm{T}} \nabla s(\theta^{(k-1)}) &= -\Big(\big(\mathrm{J}_r(\theta^{(k-1)})\big)^{\mathrm{T}} d^{(k)}\Big)^{\mathrm{T}} \big(\mathrm{J}_r(\theta^{(k-1)})\big)^{\mathrm{T}} d^{(k)} \\ &< 0. \end{aligned}$$

The update step is determined by a line search in the appropriate direction:

$$\theta^{(k)} - \theta^{(k-1)} = \alpha^{(k)} d^{(k)}.$$

This method is called the *Gauss-Newton algorithm*. (The method is also sometimes called the "modified Gauss-Newton algorithm", because many years ago no damping was used in the Gauss-Newton algorithm, and $\alpha^{(k)}$ was taken as the constant 1. Without an adjustment to the step, the Gauss-Newton method tends to overshoot the minimum in the direction $d^{(k)}$.) In practice, rather than a full search to determine the best value of $\alpha^{(k)}$, we just consider the sequence of values $1, \frac{1}{2}, \frac{1}{4}, \ldots$ and take the largest value so that $s(\theta^{(k)}) < s(\theta^{(k-1)})$. The algorithm terminates when the change is small.

If the residuals are not small or if $\mathrm{J}_r(\theta^{(k)})$ is poorly conditioned, the Gauss-Newton method can perform very poorly. One possibility is to add a conditioning matrix to the coefficient matrix in equation (1.18). A simple choice is $\tau^{(k)} I_m$, and the equation for the update becomes

$$\left(\left(\mathrm{J}_r(\theta^{(k-1)})\right)^{\mathrm{T}} \mathrm{J}_r(\theta^{(k-1)}) + \tau^{(k)} I_m\right) d^{(k)} \;=\; -\left(\mathrm{J}_r(\theta^{(k-1)})\right)^{\mathrm{T}} r(\theta^{(k-1)}),$$

where $I_m$ is the $m \times m$ identity matrix. A better choice may be an $m \times m$ scaling matrix, $S^{(k)}$, that takes into account the variability in the columns of $\mathrm{J}_r(\theta^{(k-1)})$; hence, we have for the update

$$\left(\left(\mathrm{J}_r(\theta^{(k-1)})\right)^{\mathrm{T}} \mathrm{J}_r(\theta^{(k-1)}) + \lambda^{(k)} \left(S^{(k)}\right)^{\mathrm{T}} S^{(k)}\right) d^{(k)} \;=\; -\left(\mathrm{J}_r(\theta^{(k-1)})\right)^{\mathrm{T}} r(\theta^{(k-1)}). \tag{1.19}$$

The basic requirement for the matrix $\left(S^{(k)}\right)^{\mathrm{T}} S^{(k)}$ is that it improve the condition of the coefficient matrix. There are various way of choosing this matrix. One is to transform the matrix $\left(\mathrm{J}_r(\theta^{(k-1)})\right)^{\mathrm{T}} \mathrm{J}_r(\theta^{(k-1)})$ so it has 1's along the diagonal (this is equivalent to forming a correlation matrix from a variance-covariance matrix), and to use the scaling vector to form $S^{(k)}$. The nonnegative factor $\lambda^{(k)}$ can be chosen to control the extent of the adjustment. The sequence $\lambda^{(k)}$ must go to 0 for the algorithm to converge.

Equation (1.19) can be thought of as a Lagrange multiplier formulation of the constrained problem,

$$\begin{aligned} \min_x \quad & \tfrac{1}{2} \left\| \mathrm{J}_r(\theta^{(k-1)}) x + r(\theta^{(k-1)}) \right\| \\ \text{s.t.} \quad & \left\| S^{(k)} x \right\| \leq \delta_k. \end{aligned} \tag{1.20}$$

The Lagrange multiplier $\lambda^{(k)}$ is zero if $d^{(k)}$ from equation (1.18) satisfies $\|d^{(k)}\| \leq \delta_k$; otherwise, it is chosen so that $\left\|S^{(k)} d^{(k)}\right\| = \delta_k$.

Use of an adjustment such as in equation (1.19) is called the *Levenberg-Marquardt algorithm*. This is probably the most widely used method for nonlinear least squares.

### Variance of Least-Squares Estimators

If the distribution of $Y$ has finite moments, the sample mean $\overline{Y}$ is a consistent estimator of $f(\theta_*)$. Furthermore, the minimum residual norm $\left(r(\widehat{\theta})\right)^{\mathrm{T}} r(\widehat{\theta})$ divided by $(n-m)$ is a consistent estimator of the variance of $Y$, say $\sigma^2$, that is

$$\sigma^2 = \mathrm{E}(Y - f(\theta))^2.$$

We denote this estimator as $\widehat{\sigma^2}$:

$$\widehat{\sigma^2} = \left(r(\widehat{\theta})\right)^{\mathrm{T}} r(\widehat{\theta})/(n-m).$$

The variance of the least squares estimator $\widehat{\theta}$, however, is not easy to work out, except in special cases. In the simplest case, $f$ is linear and $Y$ has a normal distribution, and we have the familiar linear regression estimates of $\theta$ and of the variance of the estimator of $\theta$. The estimator of the variance is $\widehat{\sigma^2}/(n-m)$, where $\widehat{\sigma^2}$ is an estimator of the varaince of the residuals.

Without the linearity property, however, even with the assumption of normality, it may not be possible to write a simple expression for the variance-covariance matrix of an estimator that is defined as the solution to the least-squares optimization problem. Using a linear approximation, however, we may estimate an approximate variance-covariance matrix for $\widehat{\theta}$ as

$$\left(\left(\mathrm{J}_r(\widehat{\theta})\right)^{\mathrm{T}} \mathrm{J}_r(\widehat{\theta})\right)^{-1} \widehat{\sigma^2}. \tag{1.21}$$

Compare this linear approximation to the expression for the estimated variance-covariance matrix of the least-squares estimator $\widehat{\beta}$ in the linear regression model $\mathrm{E}(Y) = X\beta$, in which $\mathrm{J}_r(\widehat{\beta})$ is just $X$. The estimate of $\sigma^2$ is taken as the sum of the squared residuals, divided by $n-m$, where $m$ is the number of estimated elements in $\theta$.

If the residuals are small, the Hessian is approximately equal to the cross-product of the Jacobian as we see from equation (1.17), and so an alternate expression for the estimated variance-covariance matrix is

$$\left(\mathrm{H}_r(\widehat{\theta})\right)^{-1} \widehat{\sigma^2}. \tag{1.22}$$

This latter expression is more useful if Newton's method or a quasi-Newton method is used instead of the Gauss-Newton method for the solution of the least squares problem, because in these methods the Hessian or an approximate Hessian is used in the computations.

### Iteratively Reweighted Least Squares

Often in applications, the residuals in equation (1.9) are not given equal weight for estimating $\theta$. This may be because the reliability or precision of the observations may be different. For *weighted least squares*, instead of (1.15) we have

the objective function

$$s_w(\theta) = \sum_{i=1}^{n} w_i\big(r_i(\theta)\big)^2. \tag{1.23}$$

The weights add no complexity to the problem, and the Gauss-Newton methods of the previous section apply immediately, with

$$\tilde{r}(\theta) = Wr(\theta),$$

where $W$ is a diagonal matrix containing the weights.

The simplicity of the computations for weighted least squares suggests a more general usage of the method. Suppose we are to minimize some other $L_p$ norm of the residuals $r_i$, as in equation (1.11). The objective function can be written as

$$s_p(\theta) = \sum_{i=1}^{n} \frac{1}{|y_i - f(\theta)|^{2-p}} \, |y_i - f(\theta)|^2 \tag{1.24}$$

This leads to an iteration on the least squares solutions. Beginning with $y_i - f(\theta^{(0)}) = 1$, we form the recursion that results from the approximation

$$s_p(\theta^{(k)}) \approx \sum_{i=1}^{n} \frac{1}{\left|y_i - f(\theta^{(k-1)})\right|^{2-p}} \left|y_i - f(\theta^{(k)})\right|^2.$$

Hence, we solve a weighted least squares problem, and then form a new weighted least squares problem using the residuals from the previous problem. This method is called *iteratively reweighted least squares* or IRLS. The iterations over the residuals are outside the loops of iterations to solve the least squares problems, so in nonlinear least squares, IRLS results in nested iterations.

There are some problems with the use of reciprocals of powers of residuals as weights. The most obvious problem arises from very small residuals. This is usually handled by use of a fixed large number as the weight.

Iteratively reweighted least squares can also be applied to other norms,

$$s_\rho(\theta) = \sum_{i=1}^{n} \rho\big(y_i - f(\theta)\big),$$

but the approximations for the updates may not be as good.

### Estimation by Maximum Likelihood

One of the most commonly-used approaches to statistical estimation is *maximum likelihood*. The concept has an intuitive appeal, and the estimators based on this approach have a number of desirable mathematical properties, at least for broad classes of distributions.

Given a sample $y_1, y_2, \ldots, y_n$ from a distribution with probability density $p(y \mid \theta_*)$, a reasonable estimate of $\theta$ is the value that maximizes the joint

density with variable $\theta$ at the observed sample value: $\prod_i p(y_i \,|\, \theta)$. We define the *likelihood function* as a function of a variable in place of the parameter:

$$L_n(\theta\,;\, y) = \prod_{i=1}^{n} p(y_i \mid \theta). \tag{1.25}$$

Note the reversal in roles of variables and parameters. For a discrete distribution, the likelihood is defined with the probability mass function in place of the density in equation (1.25).

The more difficult and interesting problems of course involve the determination of the form of the function $p(y_i \mid \theta)$. In these sections, however, we concentrate on the simpler problem of determining an appropriate value of $\theta$, assuming the form is known.

The value of $\theta$ for which $L$ attains its maximum value is the *maximum likelihood estimate* (MLE) of $\theta_*$ for the given data, $y$. The data, that is, the realizations of the variables in the density function, are considered as fixed and the parameters are considered as variables of the optimization problem,

$$\max_{\theta} L_n(\theta\,;\, y). \tag{1.26}$$

This optimization problem can be much more difficult than the optimization problem (1.10) that results from an estimation approach based on minimization of some norm of a residual vector. As we dicussed in that case, there can be both computational and statistical problems associated either with restrictions on the set of possible parameter values or with the existence of local optima of the objective function. These problems also occur in maximum likelihood estimation. Applying constraints in the optimization problem so as to force the solution to be within the set of possible parameter values is called *restricted maximum likelihood estimation*, or REML estimation. In addition to these two types of problems, the likelihood function may not even be bounded. The conceptual difficulties resulting from an unbounded likelihood are much deeper. In practice, for computing estimates in the unbounded case, the general likelihood principle may be retained, and the optimization problem redefined to include a penalty that keeps the function bounded. Adding a penalty so as to form a bounded objective function in the optimization problem, or so as to dampen the solution is called *penalized maximum likelihood estimation*.

For a broad class of distributions, the maximum likelihood criterion yields estimators with good statistical properties. The conditions that guarantee certain optimality properties are called the "regular case". The general theory of the regular case is discussed in a number of texts, such as Lehmann and Casella (1998). Various nonregular cases are discussed by Cheng and Traylor (1995).

While in practice the functions of residuals that are minimized are almost always differentiable, and the optimum occurs at a stationary point, this is often not the case in maximum likelihood estimation. A standard example in

which the MLE does not occur at a stationary point is a distribution in which the range depends on the parameter, and the simplest such distribution is the uniform $\mathrm{U}(0, \theta)$. In this case, the MLE is the max order statistic.

An important family of probability distributions are those whose probability densities are members of the *exponential family*, that is, densities of the form

$$\begin{aligned} p(y \,|\, \theta) &= h(y) \exp\Big(\theta^{\mathrm{T}} g(y) - a(\theta)\Big), \quad \text{if} \quad y \in \mathcal{Y}, \\ &= 0, \quad \text{otherwise}, \end{aligned} \tag{1.27}$$

where $\mathcal{Y}$ is some set, $\theta$ is an $m$-vector, and $g(\cdot)$ is an $m$-vector-valued function. Maximum likelihood estimation is particularly straightforward for distributions in the exponential family. Whenever $\mathcal{Y}$ does not depend on $\theta$, and $g(\cdot)$ and $a(\cdot)$ are sufficiently smooth, the MLE has certain optimal statistical properties. This family of probability distributions includes many of the familiar distributions, such as the normal, the binomial, the Poisson, the gamma, the Pareto, and the negative binomial.

The *log-likelihood function*,

$$l_{L_n}(\theta\,;\, y) = \log L_n(\theta\,;\, y), \tag{1.28}$$

is a sum rather than a product. The form of the log-likelihood in the exponential family is particularly simple:

$$l_{L_n}(\theta\,;\, y) = \sum_{i=1}^{n} \theta^{\mathrm{T}} g(y_i) - n\, a(\theta) + c,$$

where $c$ depends on the $y_i$, but is constant with respect to the variable of interest.

The logarithm is monotone, so the optimization problem (1.26) can be solved by solving the maximization problem with the log-likelihood function:

$$\max_{\theta} \; l_{L_n}(\theta\,;\, y). \tag{1.29}$$

In the following discussion we will find it convenient to drop the subscript $n$ in the notation for the likelihood and the log-likelihood. We will also often work with the likelihood and log-likelihood as if there is only one observation. (A general definition of a likelihood function is any nonnegative function that is proportional to the density or the probability mass function; that is, it is the same as the density or the probability mass function except that the arguments are switched, and its integral or sum over the domain of the random variable need not be 1.)

If the likelihood is twice differentiable and if the range does not depend on the parameter, Newton's method (see equation (1.14)) could be used to solve (1.29). Newton's equation

$$\mathrm{H}_{l_L}(\theta^{(k-1)}\,;\, y)\, d^{(k)} = \nabla l_L(\theta^{(k-1)}\,;\, y) \tag{1.30}$$

is used to determine the step direction in the $k^{\text{th}}$ iteration. A quasi-Newton method, as we mentioned on page 17, uses a matrix $\widetilde{H}_{l_L}(\theta^{(k-1)})$ in place of the Hessian $\mathrm{H}_{l_L}(\theta^{(k-1)})$.

The log-likelihood function relates directly to useful concepts in statistical inference. If it exists, the derivative of the log-likelihood is the relative rate of change, with respect to the parameter placeholder $\theta$, of the probability density function at a fixed observation. If $\theta$ is a scalar, some positive function of the derivative such as its square or its absolute value is obviously a measure of the effect of change in the parameter, or of change in the estimate of the parameter. More generally, an outer product of the derivative with itself is a useful measure of the changes in the components of the parameter at any given point in the parameter space:

$$\nabla l_L\big(\theta\,;\,y\big)\;\big(\nabla l_L\big(\theta\,;\,y\big)\big)^{\mathrm{T}}.$$

The average of this quantity with respect to the probability density of the random variable $Y$,

$$I(\theta\,|\,Y) = \mathrm{E}_\theta\Big(\nabla l_L\big(\theta\;|\;Y\big)\;\big(\nabla l_L\big(\theta\;|\;Y\big)\big)^{\mathrm{T}}\Big), \tag{1.31}$$

is called the *information matrix*, or the Fisher information matrix, that an observation on $Y$ contains about the parameter $\theta$.

The optimization problem (1.26) or (1.29) can be solved by Newton's method, equation (1.13) on page 16, or by a quasi-Newton method. (We should first note that this is a maximization problem, and so the signs are reversed from our previous discussion of a minimization problem.)

If $\theta$ is a scalar, the square of the first derivative is the negative of the second derivative,

$$\left(\frac{\partial}{\partial\theta} l_L(\theta\,;\,y)\right)^2 \;=\; -\,\frac{\partial^2}{\partial\theta^2} l_L(\theta\,;\,y),$$

or, in general,

$$\nabla l_L\big(\theta\,;\,y\big)\;\big(\nabla l_L\big(\theta\,;\,y\big)\big)^{\mathrm{T}} \;=\; -\,\mathrm{H}_{l_L}\big(\theta\,;\,y\big). \tag{1.32}$$

This is interesting because the second derivative, or an approximation of it, is used in a Newton-like method to solve the maximization problem.

A common quasi-Newton method for optimizing $l_L(\theta\,;\,y)$ is *Fisher scoring*, in which the Hessian in Newton's method is replaced by its expected value. The expected value can be replaced by an estimate, such as the sample mean. The iterates then are

$$\theta^{(k)} = \theta^{(k-1)} - \Big(\widetilde{E}\big(\theta^{(k-1)}\big)\Big)^{-1} \nabla l_L\big(\theta^{(k-1)}\,;\,y\big), \tag{1.33}$$

where $\widetilde{E}(\theta^{(k-1)})$ is an estimate or an approximation of

$$\mathrm{E}\Big(\mathrm{H}_{l_L}\big(\theta^{(k-1)}\;|\;Y\big)\Big), \tag{1.34}$$

which is itself an approximation of $\mathrm{E}_{\theta_*}(\mathrm{H}_{l_L}(\theta \mid Y))$. By equation (1.32) this is the negative of the Fisher information matrix *if* the differentiation and expectation operators can be interchanged. (This is one of the "regularity conditions" we alluded to earlier.) The most common practice is to take $\widetilde{E}(\theta^{(k-1)})$ as the Hessian evaluated at the current value of the interations on $\theta$; that is, as $\mathrm{H}_{l_L}(\theta^{(k-1)}\ ;\ y)$. This is called the *observed* information matrix.

In some cases a covariate $x_i$ may be associated with the observed $y_i$, and the distribution of $Y$ with given covariate $x_i$ has a parameter $\mu$ that is a function of $x_i$ and $\theta$. (The linear regression model is an example, with $\mu_i = x_i^{\mathrm{T}}\theta$.) We may in general write $\mu = x_i(\theta)$. In these cases another quasi-Netwon method may be useful. The Hessian in equation (1.30) is replaced by

$$\Big(X(\theta^{(k-1)})\Big)^{\mathrm{T}}\ K(\theta^{(k-1)})\ X(\theta^{(k-1)}), \tag{1.35}$$

where $K(\theta^{(k-1)})$ is a positive definite matrix that may depend on the current value $\theta^{(k-1)}$. (Again, think of this in the context of a regression model, but not necessarily linear regression.) This method was suggested by Jörgensen (1984), and is called the *Delta algorithm*, because of its similarity to the delta method for approximating a variance-covariance matrix (described on page 30).

In some cases, when $\theta$ is a vector, the optimization problem (1.26) or (1.29) can be solved by alternating iterations on the elements of $\theta$. In this approach, iterations based on equations such as (1.30), are

$$\widetilde{H}_{l_L}\Big(\theta_i^{(k-1)}\ ;\ \theta_j^{(k-1)}, y\Big)\, d_i^{(k)} = \nabla l_{l_L}\Big(\theta_i^{(k-1)}\ ;\ \theta_j^{(k-1)}, y\Big), \tag{1.36}$$

where $\theta = (\theta_i, \theta_j)$ (or $(\theta_j, \theta_i)$), and $d_i$ is the update direction for $\theta_i$, and $\theta_j$ is considered to be constant in this step. In the next step the indices $i$ and $j$ are exchanged. This is called component-wise optimization. For some objective functions, the optimal value of $\theta_i$ for fixed $\theta_j$ can be determined in closed form. In such cases, component-wise optimization may be the best method.

Sometimes we may be interested in the MLE of $\theta_i$ given a fixed value of $\theta_j$, so the iterations do not involve an interchange of $i$ and $j$, as in component-wise optimization. Separating the arguments of the likelihood or log-likelihood function in this manner leads to what is called *profile likelihood*, or *concentrated likelihood*.

As a purely computational device, the separation of $\theta$ into smaller vectors makes for a smaller optimization problem for which the number of computations are reduced by more than a linear amount. The iterations tend to ziqzag toward the solution, so convergence may be quite slow. If, however, the Hessian is block diagonal, or almost block diagonal (with sparse off-diagonal submatrices), two succesive steps of the alternating method are essentially equivalent to one step with the full $\theta$. The rate of convergence would be the same as that with the full $\theta$. Because the total number of computations in the two steps is less than the number of computations in a single step with a full $\theta$, the method may be more efficient in this case.

### Statistical Properties of MLE

Under suitable regularity conditions we referred to earlier, maximum likelihood estimators have a number of desirable properties. For most distributions used as models in practical applications, the MLE are consistent. Furthermore, in those cases, the MLE $\widehat{\theta}$ is asymptotically normal (with mean $\theta_*$) and variance-covariance matrix

$$\left(\mathrm{E}_{\theta_*}\left(-\mathrm{H}_{l_L}\left(\theta_* \mid Y\right)\right)\right)^{-1}, \tag{1.37}$$

that is, the inverse of the Fisher information matrix. A consistent estimator of the variance-covariance matrix is the Hessian at $\widehat{\theta}$. (Note that there are two kinds of asymptotic properties and convergence issues: some involve the iterative algorithm, and the others are the usual statistical asymptotics in terms of the sample size.)

### EM Methods

As we mentioned above, the computational burden in a single iteration for solving the MLE optimization problem can be reduced by more than a linear amount by separating $\theta$ into two subvectors. The MLE is then computed by alternating between computations involving the two subvectors, and the iterations proceed in a zigzag path to the solution. Each of the individual sequence of iterations is simpler than the sequence of iterations on the full $\theta$.

Another alternating method that arises from an entirely different approach alternates between updating $\theta^{(k)}$ using maximum likelihood and conditional expected values. This method is called the *EM method* because the alternating steps involve an expectation and a maximization. The method was described and analyzed by Dempster, Laird, and Rubin (1977). Many additional details and alternatives are discussed by McLachlan and Krishnan (1997) who also work through about thirty examples of applications of the EM algorithm.

The EM methods can be explained most easily in terms of a random sample that consists of two components, one observed and one unobserved, or missing. A simple example of missing data occurs in life-testing, when, for example, a number of electrical units are switched on and the time when each fails is recorded. In such an experiment it is usually necessary to curtail the recordings prior to the failure of all units. The failure times of the units still working are unobserved. The data are said to be *left censored*. The number of censored observations and the time of the censoring obviously provide information about the distribution of the failure times.

The missing data can be missing observations on the same random variable as the random variable yielding the observed sample, as in the case of the censoring example; or the missing data can be from a different random variable that is related somehow to the random variable observed.

Many common applications of EM methods do involve missing-data problems, but this is not necessary. Often, an EM method can be constructed based

on an artificial "missing" random variable to supplement the observable data.

Let $Y = (U, V)$, and assume we have observations on $U$, but not on $V$. We wish to estimate the parameter $\theta$, which figures in the distribution of both components of $Y$. An EM method uses the observations on $U$ to obtain a value of $\theta^{(k)}$ that increases the likelihood, and then uses an expectation based on $V$ that increases the likelihood further.

Let $L_c(\theta\ ;\ u, v)$ and $l_{L_c}(\theta\ ;\ u, v)$ denote respectively the likelihood and the log-likelihood for the complete sample. The likelihood for the observed $U$ is

$$L(\theta\ ;\ u) = \int L_c(\theta\ ;\ u, v)\,\mathrm{d}v,$$

and $l_L(\theta\ ;\ u) = \log L(\theta\ ;\ u)$. The EM approach to maximizing $L(\theta\ ;\ u)$ has two steps that begin with a value $\theta^{(0)}$. The steps are iterated until convergence.

- E step - compute $q^{(k)}(\theta) = \mathrm{E}_{V|u,\theta^{(k-1)}}\left(l_{L_c}(\theta \mid u, V)\right)$
- M step - determine $\theta^{(k)}$ so as to maximize $q^{(k)}(\theta)$, subject to any constraints on acceptable values of $\theta$.

The sequence $\theta^{(1)}, \theta^{(2)}, \ldots$ converges to a local maximum of the observed-data likelihood $L(\theta\ ;\ u)$ under fairly general conditions (including, of course, the nonexistence of a local maximum near enough to $\theta^{(0)}$). See Wu (1983) for discussion of the convergence conditions. The EM method can be very slow to converge, however.

As an example of the EM method, consider an experiment described by Flury and Zoppè (2000). It is assumed that the lifetime of light bulbs follow an exponential distribution with mean $\theta$. To estimate $\theta$, $n$ light bulbs were put on test until they all failed. Their failure times were recorded as $u_1, u_2, \ldots, u_n$. In a separate test, $m$ bulbs were put on test, but the individual failure times were not recorded; only the number of bulbs, $r$, that had failed at time $t$ was recorded. The missing data are the failure times of the bulbs in the second experiment, $v_1, v_2, \ldots, v_m$. We have

$$l_{L_c}(\theta\ ;\ u, v) = -n(\log\theta + \bar{u}/\theta) - \sum_{i=1}^{m}(\log\theta + v_i/\theta).$$

The expected value, $\mathrm{E}_{V|u,\theta^{(k-1)}}$, of this is

$$q^{(k)}(\theta) = -(n+m)\log\theta - \frac{1}{\theta}\left(n\bar{u} + (m-r)(t + \theta^{(k-1)}) + r(\theta^{(k-1)} - th^{(k-1)})\right),$$

where the hazard $h^{(k-1)}$ is given by

$$h^{(k-1)} = \frac{\mathrm{e}^{t/\theta^{(k-1)}}}{1 - \mathrm{e}^{t/\theta^{(k-1)}}}.$$

The $k^{\text{th}}$ M step determines the maximum, which, given $\theta^{(k-1)}$, occurs at

$$\theta^{(k)} = \frac{1}{n+m} n\bar{u} + (m-r)(t + \theta^{(k-1)}) + r(\theta^{(k-1)} - th^{(k-1)}).$$

Starting with a positive number $\theta^{(0)}$, this equation is iterated until convergence.

This example is interesting because if we assume the distribution of the light bulbs is uniform, $\mathrm{U}(0, \theta)$, (such bulbs are called "heavybulbs"!) the EM algorithm cannot be applied. As we have pointed out above, maximum likelihood methods must be used with some care whenever the range of the distribution depends on the parameter. In this case, however, there is another problem. It is in computing $q^{(k)}(\theta)$, which does not exist for $\theta < \theta^{(k-1)}$.

Although in the paper that first provided a solid description of the EM method (Dempster, Laird, and Rubin, 1977), specific techniques were used for the computations in the two steps, it is not necessary for the EM method to use those same inner-loop algorithms. There are various other ways each of these computations can be performed. A number of papers since 1977 have suggested specific methods for the computations and have given new names to methods based on those inner-loop computations.

For the expectation step there are not so many choices. In the happy case of an exponential family or some other nice distributions, the expectation can be computed in closed form. Otherwise, computing the expectation is a numerical quadrature problem. There are various procedures for quadrature, including Monte Carlo (see page 51). Wei and Tanner (1990) call an EM method that uses Monte Carlo to evaluate the expectation an MCEM method. (If a Newton-Cotes method is used, however, we do not call it an NCEM method.) The additional Monte Carlo computations add a lot to the overall time required for convergence of the EM method. Even the variance-reducing methods discussed in Section 2.6 can do little to speed up the method. An additional problem may be that the distribution of $Y$ is difficult to simulate. The versatile Gibbs method (page 48) is often useful in this context (see Chan and Ledolter, 1995). The convergence criterion for optimization methods that involve Monte Carlo generally should be tighter than those for deterministic methods.

For the maximization step there are more choices, as we have seen in the discussion of maximum likelihood estimation above.

For the maximization step, Dempster, Laird, and Rubin (1977) suggested requiring only an increase in the expected value; that is, take $\theta^{(k)}$ so that $q_k(\theta^{(k)}) \geq q_{k-1}(\theta^{(k-1)})$. This is called a generalized EM algorithm, or GEM. Rai and Matthews (1993) suggest taking $\theta^{(k)}$ as the point resulting from a single Newton step, and called this method EM1.

Meng and Rubin (1993) describe a GEM algorithm in which the M-step is a component-wise maximization, as in the update step of equation (1.36) on page 25; that is, if $\theta = (\theta_1, \theta_2)$, first, $\theta_1^{(k)}$ is determined so as to maximize $q$ subject to the constraint $\theta_2 = \theta_2^{(k-1)}$; then $\theta_2^{(k)}$ is determined so as to maximize $q$ subject to the constraint $\theta_1 = \theta_1^{(k)}$. They call this an expectation conditional

maximization, or ECM, algorithm. This sometimes simplifies the maximization problem so that it can be done in closed form. Jamshidian and Jennrich (1993) discuss acceleration of the EM algorithm, using conjugate gradient methods, and by using quasi-Newton methods (Jamshidian and Jennrich, 1997).

Kim and Taylor (1995) describe an EM method when there are linear restrictions on the parameters.

As is usual for estimators defined as solutions to optimization problems, we may have some difficulty in determining the statistical properties of the estimators. Louis (1982) suggested a method of estimating the variance-covariance matrix of the estimator by use of the gradient and Hessian of the complete-data log-likelihood, $l_{L_c}(\theta\ ;\ u, v)$. Meng and Rubin (1991), use a "supplemented" EM method, SEM, for estimation of the variance-covariance matrix. Kim and Taylor (1995) also described ways of estimating the variance-covariance matrix using computations that are part of the EM steps.

It is interesting to note that under certain assumptions on the distribution, the iteratively reweighted least squares method discussed on page 21 can be formulated as an EM method. (See Dempster, Laird, and Rubin, 1980.)

## 1.5 Inference about Functions

### Functions of Parameters and Functions of Estimators

Suppose, instead of estimating the parameter $\theta$, we wish to estimate $g(\theta)$, where $g(\cdot)$ is some function. If the function $g(\cdot)$ is monotonic or has certain other properties estimators, it may be the case that the estimator that results from the minimum residuals principle or from the maximum likelihood principle is invariate; that is, the estimator of $g(\theta)$ is merely the function $g(\cdot)$ evaluated at the solution to the optimization problem for estimating $\theta$. The statistical properties of a $T$ for estimating $\theta$, however, do not necessarily carry over to $g(T)$ for estimating $g(\theta)$.

As an example of why a function of an unbiased estimator may not be unbiased, consider a simple case in which $T$ and $g(T)$ are scalars. Let $R = g(T)$ and consider $\mathrm{E}(R)$ and $g(\mathrm{E}(T))$ in the case in which $g$ is a convex function. (A function $g$ is a *convex function* if for any two points $x$ and $y$ in the domain of $g$, $g(\frac{1}{2}(x+y)) \leq \frac{1}{2}(g(x)+g(y))$.) In this case, obviously

$$\mathrm{E}(R) \leq g(\mathrm{E}(T)), \tag{1.38}$$

so $R$ is biased for $g(\theta)$. (This relation is *Jensen's inequality*.) An opposite inequality obviously also applies to a concave function, in which case the bias is positive.

It is often possible to adjust $R$ to be unbiased for $g(\theta)$; and properties of $T$, such as sufficiency for $\theta$, may carry over to the adjusted $R$. Some of the applications of the jackknife and the bootstrap that we discuss later are in making adjustments to estimators of $g(\theta)$ that are based on estimators of $\theta$.

The variance of $R = g(T)$ can often be approximated in terms of the variance of $T$. Let $T$ and $\theta$ be $m$-vectors and let $R$ be a $k$-vector. In a simple but common case, we may know that $T$ in a sample of size $n$ has an approximate normal distribution with mean $\theta$ and some variance-covariance matrix, say $\mathrm{V}(T)$, and $g$ is a smooth function (that is, it can be approximated by a truncated Taylor series about $\theta$):

$$\begin{aligned} R_i &= g_i(T) \\ &\approx g_i(\theta) + \mathrm{J}_{g_i}(\theta)(T-\theta) + \frac{1}{2}(T-\theta)^{\mathrm{T}}\mathrm{H}_{g_i}(\theta)(T-\theta). \end{aligned}$$

Because the variance of $T$ is $\mathrm{O}(n^{-1})$, the remaining terms in the expansion go to zero in probability at the rate of at least $n^{-1}$.

This yields the approximations

$$\mathrm{E}(R) \approx g(\theta) \tag{1.39}$$

and

$$\mathrm{V}(R) \approx \mathrm{J}_g(\theta)\,\mathrm{V}(T)\,\big(\mathrm{J}_g(\theta)\big)^{\mathrm{T}}. \tag{1.40}$$

This method of approximation of the variance is called the *delta method*.

A common form of a simple estimator that may be difficult to analyze, and which may have unexpected properties, is a ratio of two statistics,

$$R = \frac{T}{S},$$

where $S$ is a scalar. An example is a studentized statistic, in which $T$ is a sample mean and $S$ is a function of squared deviations. If the underlying distribution is normal, a statistic of this form may have a wellknown and tractable distribution, in particular if $T$ is a mean and $S$ is a function of an independent chi-squared random variable, the distribution is that of a Student's $t$. If the underlying distribution has heavy tails, however, the distribution of $R$ may have unexpectedly light tails. An asymmetric underlying distribution may also cause the distribution of $R$ to be very different from a Student's $t$ distribution. If the underlying distribution is positively skewed, the distribution of $R$ may be negatively skewed (see Exercise 1.9).

### Linear Estimators

A functional $\Theta$ is *linear* if, for any two functions $f$ and $g$ in the domain of $\Theta$ and any real number $a$,

$$\Theta(af + g) = a\Theta(f) + \Theta(g).$$

A statistic is linear if it is a linear functional of the ECDF. A linear statistic can be computed from a sample using an online algorithm, and linear statistics from two samples can be combined by addition. Strictly speaking, this definition excludes statistics such as means, but such statistics are *essentially linear* in the sense that they can be combined by a linear combination if the sample sizes are known.

## 1.6 Probability Statements in Statistical Inference

There are two instances in statistical inference in which statements about probability are associated with the decisions of the inferential methods. In hypothesis testing, under assumptions about the distributions, we base our inferential methods on probabilities of two types of errors. In confidence intervals the decisions are associated with probability statements about coverage of the parameters. In computational inference, probabilities associated with hypothesis tests or confidence intervals are estimated by simulation of an hypothesized data generating process or by resampling of an observed sample.

### Tests of Hypotheses

Often statistical inference involves testing a "null" hypothesis, $H_0$, about the parameter. In a simple case, for example, we may test the hypothesis

$$H_0 : \quad \theta = \theta_0$$

versus an alternative hypothesis that $\theta$ takes on some other value or is in some set that does not include $\theta_0$. The straightforward way of performing the test involves use of a test statistic, $T$, computed from a random sample of data. Associated with $T$ is a rejection region $C$, such that if the null hypothesis is true, $\Pr(T \in C)$ is some preassigned (small) value, $\alpha$, and $\Pr(T \in C)$ is greater than $\alpha$ if the null hypothesis is not true. Thus, $C$ is a region of more "extreme" values of the test statistic if the null hypothesis is true. If $T \in C$, the null hypothesis is rejected. It is desirable that the test have a high probability of rejecting the null hypothesis if indeed the null hypothesis is not true. The probability of rejection of the null hypothesis is called the power of the test.

A procedure for testing that is mechanically equivalent to this is to compute the test statistic $t$ and then to determine the probability that $T$ is more extreme than $t$. In this approach, the realized value of the test statistic determines a region $C_t$ of more extreme values. The probability that the test statistic is in $C_t$ if the null hypothesis is true, $\Pr(T \in C_t)$, is called the "p-value" or "significance level" of the realized test statistic.

If the distribution of $T$ under the null hypothesis is known, the critical region or the p-value can be determined. If the distribution of $T$ is not known, some other approach must be used. A common method is to use some approximation to the distribution. The objective is to approximate a quantile of $T$ under the null hypothesis. The approximation is often based on an asymptotic distribution of the test statistic. In Monte Carlo tests, discussed in Section 2.3, the quantile of $T$ is estimated by simulation of the distribution.

**Confidence Intervals**

Our usual notion of a confidence interval relies on a frequency approach to probability, and it leads to the definition of a $1-\alpha$ confidence interval for the (scalar) parameter $\theta$ as the random interval $(T_L, T_U)$, that has the property

$$\Pr(T_L \leq \theta \leq T_U) = 1-\alpha. \tag{1.41}$$

This is also called a $(1-\alpha)100\%$ confidence interval. The interval $(T_L, T_U)$ is not uniquely determined.

The concept extends easily to vector-valued parameters. Rather than taking vectors $T_L$ and $T_U$, however, we generally define an ellipsoidal region, whose shape is determined by the covariances of the estimators.

A realization of the random interval, say $(t_L, t_U)$, is also called a confidence interval. Although it may seem natural to state that the "probability that $\theta$ is in $(t_L, t_U)$ is $1-\alpha$", this statement can be misleading unless a certain underlying probability structure is assumed.

In practice, the interval is usually specified with respect to an estimator of $\theta$, $T$. If we know the sampling distribution of $T-\theta$, we may determine $c_1$ and $c_2$ such that

$$\Pr(c_1 \leq T-\theta \leq c_2) = 1-\alpha; \tag{1.42}$$

and hence

$$\Pr(T-c_2 \leq \theta \leq T-c_1) = 1-\alpha.$$

If either $T_L$ or $T_U$ in (1.41) is infinite or corresponds to a bound on acceptable values of $\theta$, the confidence interval is one-sided. For two-sided confidence intervals, we may seek to make the probability on either side of $T$ to be equal, to make $c_1 = -c_2$, and/or to minimize $|c_1|$ or $|c_2|$. This is similar in spirit to seeking an estimator with small variance.

For forming confidence intervals, we generally use a function of the sample that also involves the parameter of interest, $f(T,\theta)$. The confidence interval is then formed by separating the parameter from the sample values.

A class of functions that are particularly useful for forming confidence intervals are called *pivotal* values, or pivotal functions. A function $f(T,\theta)$ is said to be a pivotal function if its distribution does not depend on any unknown parameters. This allows exact confidence intervals to be formed for the parameter $\theta$. We first form

$$\Pr\Big(f_{(\alpha/2)} \leq f(T,\theta) \leq f_{(1-\alpha/2)}\Big) = 1-\alpha, \tag{1.43}$$

where $f_{(\alpha/2)}$ and $f_{(1-\alpha/2)}$ are quantiles of the distribution of $f(T,\theta)$; that is,

$$\Pr(f(T,\theta) \leq f_{(\pi)}) = \pi.$$

If, as in the case considered above, $f(T,\theta) = T-\theta$, the resulting confidence interval has the form

$$\Pr\Big(T-f_{(1-\alpha/2)} \leq \theta \leq T-f_{(\alpha/2)}\Big) = 1-\alpha.$$

For example, suppose $Y_1, Y_2, \ldots, Y_n$ is a random sample from a $\mathrm{N}(\mu, \sigma^2)$ distribution, and $\overline{Y}$ is the sample mean. The quantity

$$f(\overline{Y}, \mu) = \frac{\sqrt{n(n-1)}\,(\overline{Y} - \mu)}{\sqrt{\sum \left(Y_i - \overline{Y}\right)^2}} \tag{1.44}$$

has a Student's $t$ distribution with $n-1$ degrees of freedom, no matter what is the value of $\sigma^2$. This is one of the most commonly-used pivotal values.

The pivotal value in equation (1.44) can be used to form a confidence value for $\theta$ by first writing

$$\Pr\left(t_{(\alpha/2)} \;\leq\; f(\overline{Y}, \mu) \;\leq\; t_{(1-\alpha/2)}\right) \;=\; 1 - \alpha,$$

where $t_{(\pi)}$ is a percentile from the Student's $t$ distribution. Then, after making substitutions for $f(\overline{Y}, \mu)$, we form the familiar confidence interval for $\mu$:

$$\left(\overline{Y} - t_{(1-\alpha/2)}\, s/\sqrt{n}, \quad \overline{Y} - t_{(\alpha/2)}\, s/\sqrt{n}\right), \tag{1.45}$$

where $s^2$ is the usual sample variance, $\sum (Y_i - \overline{Y})^2/(n-1)$.

Other similar pivotal values have $F$ distributions. For example, consider the usual linear regression model in which the $n$-vector random variable $Y$ has a $\mathrm{N}_n(X\beta, \sigma^2 I)$ distribution, where $X$ is an $n \times m$ known matrix, and the $m$-vector $\beta$ and the scalar $\sigma^2$ are unknown. A pivotal value useful in making inferences about $\beta$ is

$$g(\widehat{\beta}, \beta) = \frac{\left(X(\widehat{\beta} - \beta)\right)^{\mathrm{T}} X(\widehat{\beta} - \beta)/m}{(Y - X\widehat{\beta})^{\mathrm{T}}(Y - X\widehat{\beta})/(n-m)}, \tag{1.46}$$

where

$$\widehat{\beta} = (X^{\mathrm{T}}X)^{+}X^{\mathrm{T}}Y.$$

The random variable $g(\widehat{\beta}, \beta)$ for any finite value of $\sigma^2$ has an $F$ distribution with $m$ and $n-m$ degrees of freedom.

For a given parameter and family of distributions there may be multiple pivotal values. For purposes of statistical inference, such considerations as unbiasedness and minimum variance may guide the choice of a pivotal value to use. Alternatively, it may not be possible to identify a pivotal quantity for a particular parameter. In that case, we may seek an approximate pivot. A function is asymptotically pivotal if a sequence of linear transformations of the function is pivotal in the limit as $n \to \infty$.

If the distribution of $T$ is known, $c_1$ and $c_2$ in equation (1.42) can be determined. If the distribution of $T$ is not known, some other approach must be used. A common method is to use some numerical approximation to the distribution. Another method, discussed in Section 4.3, is to use "bootstrap" samples from the ECDF.

# Exercises

1.1. (a) How would you describe, in nontechnical terms, the structure of the dataset displayed in Figure 1.1, page 7?

(b) How would you describe the structure of the dataset in more precise mathematical terms? (Obviously, without having the actual data, your equations must contain unknown quantities. The question is meant to make you think about *how* you would do this — that is, what would be the components of your model.)

1.2. Show that the variance of the ECDF at a point $y$ is the expression in equation (1.3) on page 11. *Hint:* Use the definition of the variance in terms of expected values, and represent $\mathrm{E}\Big(\big(P_n(y)\big)^2\Big)$ in a manner similar to how $\mathrm{E}\big(P_n(y)\big)$ was represented in equations (1.2).

1.3. The variance functional.

(a) Express the variance of a random variable as a functional of its CDF as was done in equation (1.5) for the mean.

(b) What is the same functional of the ECDF?

(c) What is the plug-in estimate of the variance?

(d) What are the statistical properties of the plug-in estimate of the variance? (Is it unbiased? Is it consistent? Is it an MLE? etc.)

1.4. Assume a random sample of size 10 from a normal distribution. With $\nu = 1-2\iota$ in equation (1.7), determine the value of $\iota$ that makes the empirical quantile of the $9^{\text{th}}$ order statistic be unbiased for the normal quantile corresponding to 0.90.

1.5. Give examples of

(a) a parameter that is defined by a linear functional of the distribution function, and

(b) a parameter that is not a linear functional of the distribution function.

(c) Is the variance a linear functional of the distribution function?

1.6. Consider the least-squares estimator of $\beta$ in the usual linear regression model, $\mathrm{E}(Y) = X\beta$.

(a) Use expression (1.21), page 20, to derive the variance-covariance matrix for the estimator.

(b) Use expression (1.22) to derive the variance-covariance matrix for the estimator.

1.7. Assume a random sample $y_1, \ldots, y_n$ from a gamma distribution with parameters $\alpha$ and $\beta$.

(a) What are the least-squares estimates of $\alpha$ and $\beta$? (Recall $\mathrm{E}(Y) = \alpha\beta$ and $\mathrm{V}(Y) = \alpha\beta^2$.)

(b) Write a function in a language such as R, Matlab, or Fortran that accepts a sample of size $n$ and computes the least squares estimator of $\alpha$ and $\beta$ and computes an approximation of the variance-covariance matrix using both expression (1.21), page 20, and expression (1.22).

(c) Try out your program in Exercise 1.7b by generating a sample of size 500 from a gamma(2,3) distribution and then computing the estimates. (See Appendix B for information on software for generating random deviates.)

(d) Formulate the optimization problem for determining the MLE of $\alpha$ and $\beta$. Does this problem have a closed-form solution?

(e) Write a function in a language such as R, Matlab, or Fortran that accepts a sample of size $n$ and computes the least squares estimator of $\alpha$ and $\beta$ and computes an approximation of the variance-covariance matrix using expression (1.37), page 26.

(f) Try out your program in Exercise 1.7e by computing the estimates from an artificial sample of size 500 from a gamma(2,3) distribution.

1.8. For the random variable $Y$ with a distribution in the exponential family and whose density is expressed in the form of equation (1.27), page 23, and assuming that the first two moments of $g(Y)$ exist and $a(\cdot)$ is twice differentiable, show that

$$\mathrm{E}(g(Y)) = \nabla a(\theta)$$

and

$$\mathrm{V}(g(Y)) = \mathrm{H}_a(\theta).$$

*Hint:* First show that

$$\mathrm{E}(\nabla \log(p(Y \mid \theta))) = 0,$$

where the differentiation is with respect to $\theta$.

1.9. Assume $\{X_1, X_2\}$ is a random sample of size 2 from an exponential distribution with parameter $\theta$. Consider the random variable formed as a Student's $t$:

$$T = \frac{\overline{X} - \theta}{\sqrt{S^2/2}},$$

where $\overline{X}$ is the sample mean and $S^2$ is the sample variance,

$$\frac{1}{n-1}\sum (X_i - \overline{X})^2.$$

(Note $n = 2$.)

(a) Show that the distribution of $T$ is negatively skewed (although the distribution of $X$ is positively skewed).

(b) Give a heuristic explanation of the negative skewness of $T$.

1.10. A function $T$ of a random variable $X$ with distribution parametrized by $\theta$ is said to be *sufficient* for $\theta$ if the conditional distribution of $X$ given $T(X)$ does not depend on $\theta$. Discuss (compare and contrast) pivotal and sufficient functions. (Start with the basics: Are they statistics? In what way do they both depend on some universe of discourse, that is, on some family of distributions?)

1.11. Use the pivotal value $g(\widehat{\beta}, \beta)$ in equation (1.46), page 33, to form a $(1-\alpha)100\%$ confidence region for $\beta$ in the usual linear regression model.

1.12. Assume a random sample $y_1, \ldots, x_n$ from a normal distribution with mean $\mu$ and variance $\sigma^2$. Determine an unbiased estimator of $\sigma$, based on the sample variance, $s^2$. (Note that $s^2$ is sufficient and unbiased for $\sigma^2$.)

http://www.springer.com/978-0-387-95489-9

Elements of Computational Statistics

Gentle, J.E.

2002, XVIII, 420 p., Hardcover
ISBN: 978-0-387-95489-9