2

Maximum Likelihood Estimation

Maximum likelihood is a very general method for estimation of model parameters. It has good properties in large samples and when a valid model is used. Therefore it has to be accompanied by a method that addresses model uncertainty. In this chapter, we give details of the method of maximum likelihood and compare two approaches to dealing with model uncertainty—selecting a model and combining estimators based on the alternative models.

2.1 Likelihood

The *likelihood* is defined as the joint density or probability of the outcomes, with the roles of the values of the outcomes $\mathbf{y}$ and the values of the parameters $\boldsymbol{\theta}$ interchanged. Thus, let $f(\mathbf{y}; \boldsymbol{\theta})$ be a class of joint densities with the parameter vector $\boldsymbol{\theta}$ in a set (parameter space) $\boldsymbol{\Theta}$. For each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $f(\mathbf{y}; \boldsymbol{\theta})$ is a joint density of a continuous distribution. It will be expedient to use the same notation for joint probabilities of discrete distributions; for them, $f(\mathbf{y}; \boldsymbol{\theta}) = \mathrm{P}(\mathbf{Y} = \mathbf{y}; \boldsymbol{\theta})$, where $\mathbf{Y}$ is the random vector of the outcomes. The likelihood is defined, after recording the values $\mathbf{y}$ of the vector $\mathbf{Y}$, as the function

$$L(\boldsymbol{\theta}; \mathbf{y}) \;=\; f(\mathbf{y}; \boldsymbol{\theta}), \tag{2.1}$$

with $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ as its argument. This definition reflects the task at hand. Having observed, and therefore fixed, $\mathbf{Y}$ at $\mathbf{y}$, we consider all possible values of $\boldsymbol{\theta}$, intending to estimate the value that underlies the observed process, delineate a plausible range of values of $\boldsymbol{\theta}$, or make an inference about $\boldsymbol{\theta}$ that is formulated in some other way. When we planned the study, we considered the configurations of values of $\mathbf{y}$ that might arise and how likely they are to arise if the studied process is governed by a particular joint density f; that is, we temporarily fixed $\boldsymbol{\theta}$ and explored the plausible outcomes $\mathbf{y}$. After observing $\mathbf{y}$, we now want to identify values of $\boldsymbol{\theta}$ that are compatible with the vector of outcomes $\mathbf{y}$, pursuing the unattainable ideal of identifying the value of $\boldsymbol{\theta}$ that governs the process of generating $\mathbf{Y}$.

The maximum likelihood estimator of $\boldsymbol{\theta}$ for the model given by the joint densities or probabilities $f(\mathbf{y}; \boldsymbol{\theta})$, with $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, is defined as the value of $\boldsymbol{\theta}$ at which the corresponding likelihood $L(\boldsymbol{\theta}; \mathbf{y})$ attains its maximum:

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}; \mathbf{y}).$$

This definition is not complete because there is no guarantee that such a maximum exists or, when it does exist, it is unique. However, in many settings this definition turns out to be very useful and constructive, yielding an estimator with good properties.

The principal theoretical results about the efficiency of the maximum likelihood estimators relate to *asymptotic* settings, corresponding, roughly speaking, to large sample sizes or increasing amounts of information. A ubiquitous caveat associated with all the results is that the model has to be valid; it has to contain the distribution according to which the outcomes are generated. Another important assumption is that the likelihood L is a smooth function of the parameter vector $\boldsymbol{\theta}$, usually interpreted as L being twice differentiable with all its second-order partial differentials continuous and bounded. Further, the distributions are distinct; if two distributions in the class coincide, then so do the values of their parameter vectors $\boldsymbol{\theta}$. In most practical settings, these conditions are satisfied, as a small change in the values of the parameters $\boldsymbol{\theta}$ corresponds to small changes in the likelihood L. Further, the parameter vector $\boldsymbol{\theta}$ must be in the interior of the parameter space $\boldsymbol{\Theta}$. Notable cases in which this condition is not satisfied include a zero variance and constraints, such as $\theta_1 \geq \theta_2$, when the data-generating process satisfies the identity $\theta_1 = \theta_2$.

Suppose the outcomes $\mathbf{y}$ are conditionally independent given the values of the other (observed) variables, a matrix $\mathbf{X}$, so that the model for them can be expressed as

$$f(\mathbf{y}; \mathbf{X}, \boldsymbol{\theta}) = \prod_{j=1}^{n} f(y_j; \mathbf{x}_j; \boldsymbol{\theta})$$

($\mathbf{x}_j$ is the jth row of $\mathbf{X}$). The corresponding likelihood is

$$L(\boldsymbol{\theta}; \mathbf{y}, \mathbf{X}) = \prod_{j=1}^{n} f(y_j; \mathbf{x}_j; \boldsymbol{\theta}). \tag{2.2}$$

A standard approach to maximising this likelihood searches for values of $\hat{\boldsymbol{\theta}}$ for which the partial derivatives of L vanish. Instead of L it is more convenient to work with its logarithm, called the *log-likelihood*,

$$l(\boldsymbol{\theta}; \mathbf{y}, \mathbf{X}) = \log\{L(\boldsymbol{\theta}; \mathbf{y}, \mathbf{X})\},$$

because the product in the likelihood L converts to a summation in l,

$$l(\boldsymbol{\theta}; \mathbf{y}, \mathbf{X}) = \sum_{j=1}^{n} \log\{f(y_j; \mathbf{x}_j; \boldsymbol{\theta})\}.$$

This simplifies the differentiation;

$$\frac{\partial l}{\partial \boldsymbol{\theta}} = \sum_{j=1}^{n} \frac{\partial f}{\partial \boldsymbol{\theta}}(y_j \,; \mathbf{x}_j \,; \bullet)$$

for the likelihood in (2.2). The black disc indicates the argument over which the differentiation is carried out.

The vector of the first-order partial differentials of the log-likelihood is called the *score* vector; we denote it by $\mathbf{s}(\boldsymbol{\theta})$, with further arguments (as in l) if it is necessary to avoid any ambiguity. Thus, the maximum likelihood estimator should be sought among the roots of the score vector, solutions of the equation $\mathbf{s}(\boldsymbol{\theta}) = \mathbf{0}$, where the score vector is not defined, and on the boundary of the parameter space $\boldsymbol{\Theta}$. The matrix of the negative second-order partial differentials, defined as the matrix with elements

$$-\frac{\partial l^2}{\partial \theta_k \, \partial \theta_h}$$

as functions of $\boldsymbol{\theta}$ and $\mathbf{y}$, is called the *observed information matrix*. The expectation of the observed information matrix, with elements

$$-\mathrm{E}\left(\frac{\partial l^2}{\partial \theta_k \, \partial \theta_h}\right)$$

as functions of $\boldsymbol{\theta}$, is called the *expected information matrix*. It is denoted by $\mathcal{I}(\boldsymbol{\theta}, \boldsymbol{\theta})$. The argument $\boldsymbol{\theta}$ appears twice, so that we can use the notation also for submatrices of $\mathcal{I}$. For example, the vector $\boldsymbol{\theta}$ may be split into a subvector $\boldsymbol{\theta}_1$ of parameters of interest and subvector $\boldsymbol{\theta}_2$ of nuisance parameters; $\boldsymbol{\theta} = \left(\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top\right)^\top$. Then the complete notation for the square submatrix of $\mathcal{I}$ that corresponds to $\boldsymbol{\theta}_1$ is $\mathcal{I}\left(\boldsymbol{\theta}_1 \,, \boldsymbol{\theta}_1 \,; \boldsymbol{\theta}_1 \,, \boldsymbol{\theta}_2\right)$, emphasising that the submatrix depends on the entire parameter vector.

Example 1. Ordinary Regression. Suppose the vector of outcomes $\mathbf{y}$ is generated according to the ordinary regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\,,$$

where $\mathbf{X}$ is a matrix of covariates, with $\mathbf{1}$ as its first column, and the deviations $\boldsymbol{\varepsilon}$ are distributed according to $\mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ for a positive σ^2. We derive the maximum likelihood estimators of $\boldsymbol{\beta}$ and σ^2. We assume that $\mathbf{X}$ is of full rank p, the number of its columns, and $p < n$, so that the $p \times p$ matrix $\mathbf{X}^\top \mathbf{X}$ is nonsingular. The log-likelihood for this model is

$$l(\boldsymbol{\beta}, \sigma^2; \mathbf{y}, \mathbf{X}) = -\frac{1}{2}\left\{n \log(2\pi\sigma^2) + \frac{1}{\sigma^2}\mathbf{e}^\top \mathbf{e}\right\}, \qquad (2.3)$$

where $\mathbf{e} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$. This vector $\mathbf{e}$ differs from $\boldsymbol{\varepsilon}$; $\mathbf{e}$ is a function of $\boldsymbol{\beta}$, whereas $\boldsymbol{\varepsilon}$ is the value of $\mathbf{e}$ for the population ('true') value of $\boldsymbol{\beta}$. The log-likelihood in

(2.3) is a quadratic function of $\boldsymbol{\beta}$, involved in $\mathbf{e}$, so its extremes are easy to find. As $\partial\mathbf{e}/\partial\boldsymbol{\beta} = -\mathbf{X}$, we have

$$\frac{\partial l}{\partial\boldsymbol{\beta}} = \frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{e}\,,$$

and this score vector has the unique root

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}\,,$$

irrespective of the value of σ^2. (The inverse is well defined because $\mathbf{X}$ has full column rank.) This coincides with the ordinary least squares estimator derived in Section 1.2.

The observed information matrix for $\boldsymbol{\beta}$ is obtained by differentiating the score vector $\partial l/\partial\boldsymbol{\beta}$, yielding

$$-\frac{\partial l^2}{\partial\boldsymbol{\beta}\,\partial\boldsymbol{\beta}^\top} = \frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X}\,.$$

Being constant (not depending on $\mathbf{y}$), it is also equal to the expected information matrix. Its inverse, assuming that $\mathbf{X}^\top\mathbf{X}$ is nonsingular, is the sampling variance matrix of $\hat{\boldsymbol{\beta}}$. The two related results, that the maximum likelihood estimator is efficient and that the inverse of its information matrix is equal to the sampling variance matrix, hold more generally, but, unlike for ordinary regression, they do only approximately and with some qualifications. These are discussed in Section 2.1.2.

The maximum likelihood estimator of σ^2 is obtained by finding the root of the score function for σ^2:

$$-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\mathbf{e}^\top\mathbf{e} = 0\,,$$

that is,

$$\hat{\sigma}^2 = \frac{1}{n}\hat{\mathbf{e}}^\top\hat{\mathbf{e}} = \frac{1}{n}\left(\mathbf{y}^\top\mathbf{y} - \mathbf{y}^\top\mathbf{X}\hat{\boldsymbol{\beta}}\right)\,,$$

where $\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ is the vector of residuals. This estimator differs from its ordinary least squares counterpart by its denominator (n instead of $n - p$). It is biased; $\mathrm{E}(\hat{\sigma}^2/\sigma^2) = (n - p)/n$. This might appear as a deficiency of the maximum likelihood, although the bias of $\hat{\sigma}^2$ is small for large n.

2.1.1 Consistency

Consistency is a property of an estimator that it would recover the value of the target if it were based on many observations. To formalise this, we represent the idea of many observations by a sequence of sets of observations and models for them, with sample sizes increasing beyond all bounds. The sets, as well as observations within each set, are mutually independent. Each set

is associated with a different model, but the models share the same vector of model parameters. The simplest, yet still quite general, setting has the same density (or probability) conditional on some covariates $\mathbf{x}$ with observation-specific values, $f(\mathbf{y}; \boldsymbol{\theta}, \mathbf{x})$.

To avoid contorted verbal expressions, we refer to the sequence of (univariate) estimators $\hat{\theta}_n$ based on the nth set of observations $\mathbf{y}_n$ as a single estimator. Consistency of such an estimator $\hat{\theta}$ of a target θ is defined as convergence of $\hat{\theta}_n$ to the target θ as $n \to +\infty$. For distributions, there are several definitions of convergence, and each corresponds to a different definition of consistency. In practice, these differences are not important, and we can focus on weak convergence, defined as convergence of the distribution functions of $\hat{\theta}_n$ to the degenerate distribution with all its mass (a single jump) at θ.

An important result about maximum likelihood estimators is that under some regularity conditions they are consistent. The regularity conditions include smoothness of the likelihood, its distinctness for each vector of model parameters and finite dimensionality of the parameter space, independent of the sample size. This result has extensions in several directions. First, univariate outcomes can be replaced by multivariate ones. Next, some correlation among the observations can be allowed, so long as it is distant from ± 1. And finally, the parameter space may be expanding with the sample size, but its dimension has to grow at a rate much slower than n.

Consistency of the maximum likelihood estimator is a key condition for deriving other properties of maximum likelihood estimators that are of practical importance.

2.1.2 Asymptotic Efficiency and Normality

Asymptotic efficiency and asymptotic normality are key properties of maximum likelihood estimators. The qualifier *asymptotic* refers to properties in the limit as the sample size increases above all bounds. Asymptotic efficiency supports the everyday application of maximum likelihood estimators, and asymptotic normality enables us to make a convenient reference to a familiar distribution.

For a set of many conditionally independent outcomes (large sample size n), given covariates and a finite-dimensional set of parameters $\boldsymbol{\theta}$, the maximum likelihood estimator is approximately unbiased, and its distribution is well approximated by the normal distribution with sampling variance matrix equal to the inverse of the expected information matrix. This result is referred to as *asymptotic normality*. Further, the maximum likelihood estimator is *asymptotically efficient* and, asymptotically, the sampling variance of the estimator is equal to the corresponding diagonal element of the inverse of the expected information matrix. That is, for large n, there are no estimators substantially more efficient than the maximum likelihood estimator. This result is the main underpinning of maximum likelihood estimation. In the

next section, we construct estimators that are more efficient than maximum likelihood, but not substantially so for large sample sizes.

The *Cramér–Rao inequality* is a powerful result that relates to all unbiased estimators. It gives a lower bound for the variance of an unbiased estimator. Let $l(\theta; \mathbf{y})$ be a log-likelihood and $s(\theta; \mathbf{y})$ and $\mathcal{I}(\theta)$ the corresponding score function and expected information. Suppose l satisfies the regularity conditions listed earlier. Then any unbiased estimator $\hat{\theta}$ of θ satisfies the inequality

$$\operatorname{var}\left(\hat{\theta}\right) \geq \frac{1}{\mathcal{I}(\theta)} \, ; \tag{2.4}$$

that is, there are no unbiased estimators that are more efficient than the maximum likelihood estimator.

A more general result related to (2.4) states that, under the same regularity conditions, any estimator $\hat{\theta}$ of θ, with bias $\mathrm{B}(\hat{\theta}; \theta)$, satisfies the inequality

$$\operatorname{var}\left(\hat{\theta}\right) \geq \frac{\left\{1 + \mathrm{B}'(\hat{\theta}; \theta)\right\}^2}{\mathcal{I}(\theta)} \, . \tag{2.5}$$

Hence, there may be biased estimators with smaller MSE than any unbiased estimator. This may at first appear as a contradiction, because any estimator $\hat{\theta}$ might be improved by removing its bias. However, the bias itself has to be estimated, and so its removal may be accompanied by a variance inflation. The inequality in (2.5) suggests that efficient biased estimators should be sought among those with bias $\mathrm{B}(\hat{\theta}; \theta)$ that is a decreasing function of θ. Some shrinkage estimators have this property, so it makes sense to search for improvement on maximum likelihood estimators among them; see Example 4.

The Cramér–Rao inequality (2.4) justifies our focus on maximum likelihood only for large samples, when unbiasedness is essential for efficiency. There is no clinical formula that would arbitrate whether a given sample size in a particular setting is large enough for a specified purpose. Also, small-sample behaviour of the maximum likelihood estimator may differ from what the asymptotic expression would suggest. A further difficulty is that all the results associated with maximum likelihood estimation are subject to the caveat of working with the appropriate model. Sufficiently large samples can come close to confirming that a particular model is appropriate, but model uncertainty has to be reckoned with in small or moderate samples. The aim of achieving the lower bound in (2.4) may either be too optimistic or not particularly attractive because too complex a model has to be specified. In any case, equality in (2.4) and (2.5) is attained only when the estimator $\hat{\theta}$ is a linear function of the score $\partial l/\partial \theta$.

Asymptotic normality and efficiency of the maximum likelihood estimator confer the central role on the normal distribution in statistics. The proof of asymptotic normality relies on the weak law of large numbers, which confers a similar role on the normal distribution in probability theory. We are rather

fortunate that the normal distribution is relatively easy to handle, it has a comprehensive generalisation to many dimensions that is closed with respect to addition, taking margins and conditioning.

The results of asymptotic normality and efficiency have been extended to settings other than those of independent and conditionally identically distributed outcomes, such as for correlated observations and observations that do not have identical distributions, even after conditioning on the values of the covariates in regression or similar quantities. Features common to these extensions are that none of the observations and groups of observations that have finite sizes make an unbounded (disproportionately large) contribution to the expected information. A complication in formulating the assumptions is that asymptotics requires a much more careful definition than for independent observations. For example, for the random-effects ANOVA, the number of clusters should diverge to infinity, but the fraction n_k/n of the sample size of each cluster k and the overall sample size (the representation of cluster k) should converge to zero in such a way that even $n_k/\sqrt{n}$ converges.

The assumptions necessary for these results are that the log-likelihood is smooth, with all its second-order partial differentials continuous, the expected information matrix exists, the value of the parameter vector is in the interior of the parameter space, and the distributions constituting the model are distinct and contain the data-generating ('true') distribution. Further, all the eigenvalues of the expected information matrix diverge to $+\infty$ as $n \to \infty$. These are the regularity conditions referred to earlier.

In brief, maximum likelihood has no competitor for large samples. The theoretical results provide no formula for establishing what constitutes a large enough sample in any particular setting. Often only trial and error with simulations can provide an indication for how close we are to asymptotics. For some simple models, such as ordinary regression, maximum likelihood estimators coincide with established estimators or are very close to them. Together with its universality, this gives maximum likelihood a strong appeal and justifies its role as the workhorse of statistical analysis. With small or moderate sample sizes, maximum likelihood is applied as a default when there is no obvious alternative. In evaluating maximum likelihood estimators we may have to call on (iterative) numerical methods for maximisation of real functions.

In Chapter 1, we came across examples in which $\hat{\theta}$ was an unbiased and efficient estimator of a parameter θ, yet a monotone nonlinear transformation $g(\hat{\theta})$ was neither unbiased nor efficient for $g(\theta)$. Maximum likelihood has the converse property. If $\hat{\theta}$ is the maximum likelihood estimator of θ, then $g(\hat{\theta})$ is the maximum likelihood estimator of $g(\theta)$ for any monotone (one-to-one) transformation g. This may at first appear to be a very convenient property. However, it implies that not all maximum likelihood estimators are efficient. A maximum likelihood estimator $\hat{\theta}$ is efficient only *asymptotically* (assuming that the regularity conditions apply). As the sample size diverges, the

sampling variance of $\hat{\theta}$ diminishes; only when it vanishes can the operations of estimation and (nonlinear) monotone transformation be interchanged.

Under regularity conditions, the bias of a maximum likelihood estimator converges to zero. We say that such estimators are asymptotically unbiased. Together with the sampling variance converging to zero (as $n \to \infty$), this is equivalent to consistency, with the appropriate definition of convergence (convergence in expectation). Consistency is a valuable property in connection with large samples but is not particularly relevant otherwise, when the sampling variance is the dominating contributor to the mean squared error (MSE).

2.2 Sufficient Statistics

The log-likelihood function $l(\boldsymbol{\theta}; \mathbf{y})$ sometimes depends on the outcomes $\mathbf{y}$ only through one or a few summaries of $\mathbf{y}$; $l(\boldsymbol{\theta}; \mathbf{y}) = l\{\boldsymbol{\theta}; \mathbf{u}(\mathbf{y})\}$. To evaluate such a likelihood, we do not have to provide the n-dimensional data vector $\mathbf{y}$; it suffices to provide the summaries $\mathbf{u}(\mathbf{y})$. A set of summaries that enables the evaluation of the log-likelihood is called a *set of sufficient statistics*. When there is such a set of statistics the score vector depends on $\mathbf{y}$ also only through them: $\mathbf{s}(\boldsymbol{\theta}; \mathbf{y}) = \mathbf{s}\{\boldsymbol{\theta}; \mathbf{u}(\mathbf{y})\}$. As the main use of the likelihood is for its maximisation with respect to the vector of its parameters $\boldsymbol{\theta}$, we do not have to be concerned with its factors that do not involve $\boldsymbol{\theta}$, even if they involve $\mathbf{y}$. A set of sufficient statistics can be motivated as a condensed version of the data that is complete; any additional statistic (data summary) would be redundant for maximum likelihood estimation.

A set of sufficient statistics is qualified by the model (a class of distributions) and the vector of its parameters. Formally, it is defined as follows. A set of statistics $\mathbf{u}$ is sufficient for a parameter vector $\boldsymbol{\theta}$ in a model if the conditional distribution of $\mathbf{y}$ given the value of $\mathbf{u} = \mathbf{u}(\mathbf{y})$, $(\mathbf{y} \mid \mathbf{u})$, does not depend on $\boldsymbol{\theta}$. This conditional distribution may depend on model parameters that are not included in $\boldsymbol{\theta}$. Also, it need not depend on $\boldsymbol{\theta}$ in a particular model, but may depend on it in a more general model.

Checking that a vector $\mathbf{u}$ is sufficient by applying this definition directly is often tedious because it entails derivation of the (joint) density of $\mathbf{u}$. A much more practical equivalent definition refers to the form of the likelihood. A random vector $\mathbf{u}$ is sufficient for a parameter vector $\boldsymbol{\theta}$ in a model if and only if the log-likelihood can be expressed as

$$l(\boldsymbol{\theta}; \mathbf{y}) \,=\, l_1\{\boldsymbol{\theta}; \mathbf{u}(\mathbf{y})\} + l_2(\mathbf{y}) \,. \tag{2.6}$$

This equivalence is known as the *factorisation theorem*, referring to the factors $\exp(l_1)$ and $\exp(l_2)$ of the likelihood $\exp(l)$. Note that maximising the likelihood is equivalent to maximising the 'essential' factor $l_1(\boldsymbol{\theta}, \mathbf{u})$ and is related to the problem of finding the roots of $\mathbf{s}(\boldsymbol{\theta}) = \partial l_1(\boldsymbol{\theta}, \mathbf{u})/\partial \boldsymbol{\theta}$.

For example, $\mathbf{X}^\top \mathbf{y}$ is a set of sufficient statistics for $\boldsymbol{\beta}$ in the ordinary regression, and when supplemented with $\mathbf{y}^\top \mathbf{y}$ it is sufficient also for σ^2. This is obvious from the expression

$$ l \; = \; -\frac{1}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\,\boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} + \frac{1}{\sigma^2}\,\mathbf{y}^\top \mathbf{X}\boldsymbol{\beta} - \frac{1}{2\sigma^2}\,\mathbf{y}^\top \mathbf{y}\,; $$

the first three terms on the right-hand side do not depend on $\mathbf{y}$, and the last term does not involve $\boldsymbol{\beta}$ but does involve σ^2. A set of sufficient statistics $\mathbf{u}(\mathbf{y})$ is not unique because sufficiency is retained, for instance, when the components of $\mathbf{u}(\mathbf{y})$ are subjected to strictly monotone transformations or when further summaries are added to $\mathbf{u}(\mathbf{y})$. In particular, $\mathbf{y}$ itself is a set of sufficient statistics.

A set of sufficient statistics $\mathbf{u}$ that has m components is said to be *minimal* if for every transformation f from $\mathcal{R}^m$ to $\mathcal{R}^{m'}$, $m' < m$, $f(\mathbf{u})$ is not sufficient. That is, a set of sufficient statistics is minimal if all of its reductions to fewer statistics are not sufficient. For example, if we drop one of the components of a set of minimal sufficient statistics or replace a pair of them by their total, the result is not a set of sufficient statistics. In ordinary regression with $\mathbf{X}$ of full column rank, $\mathbf{y}$ as a set of statistics is not minimal sufficient because $(\mathbf{X} \; \mathbf{y})^\top \mathbf{y}$ is a reduction of $\mathbf{y}$ to fewer dimensions. A set of sufficient statistics $\mathbf{u}(\mathbf{y})$ is said to be *linear* if l_1 in the factorisation (2.6) is a linear function of $\mathbf{u}$.

The importance of sufficient statistics is that they reduce the range of data summaries that have to be considered for any inference about the model parameters. Thus, immediately after data collection we can reduce the outcomes $\mathbf{y}$ to the statistics $\mathbf{u}(\mathbf{y})$ without discarding any relevant information. This is particularly valuable when $\mathbf{u}$ has only a small number of components, and their number does not depend on the sample size n. In iterative procedures for maximising the likelihood, we do not have to work with the entire vector $\mathbf{y}$ in each iteration if we evaluate a vector of sufficient statistics before the first iteration. We can consider similar summaries for $(\mathbf{X} \; \mathbf{y})$, formally by regarding the covariates $\mathbf{X}$ also as outcomes. With such summaries, we could conduct the analysis without requiring any access to $\mathbf{X}$ or $\mathbf{y}$. For example, the ordinary least squares requires the summaries in $(\mathbf{X} \; \mathbf{y})^\top (\mathbf{X} \; \mathbf{y})$.

A theoretical result supporting our focus on sufficient statistics is the *Rao–Blackwell theorem*. It states that any estimator $\hat{\theta}$ of a model parameter θ is at least as efficient as the conditional expectation $\mathrm{E}(\hat{\theta} \,|\, \mathbf{u})$ where $\mathbf{u}$ is a set of sufficient statistics. Thus, any estimator of θ can be associated with an estimator that is at least as efficient and depends on $\mathbf{y}$ only through $\mathbf{u}$.

Example 2. Exponential Distributions. The class of exponential distributions is given by the densities

$$ f(x;\theta) \; = \; \theta\,\exp(-\theta x) $$

for argument $x > 0$ and parameter $\theta > 0$. We derive the maximum likelihood estimator of θ based on a random sample $\mathbf{x}$ of size n from an exponential

distribution. The log-likelihood is equal to

$$l(\theta; \mathbf{x}) = n \log(\theta) - \theta x_+ ,$$

where $x_+ = \mathbf{x}^\top \mathbf{1}$ is the sample total. Thus, x_+ is a linear sufficient statistic. Of course, it is minimal. The score for θ is equal to

$$l'(\theta; \mathbf{x}) = \frac{n}{\theta} - x_+ ,$$

so the maximum likelihood estimator is $\hat{\theta} = 1/\bar{x}$, the reciprocal of the sample mean.

The sample total x_+ has the gamma distribution with parameters θ and n:

$$f_n(u) = \frac{\theta^n x^{n-1}}{\Gamma(n)} \exp(-\theta x) .$$

This can be proved by induction. For $n = 1$ the result is obvious. Assuming the result for a given n, the density of the total of $n + 1$ random values is

$$\int_0^x \frac{\theta^n y^{n-1}}{\Gamma(n)} \exp(-\theta y) \theta \exp\{-\theta(x - y)\} \mathrm{d}y = \frac{1}{n} \frac{\theta^{n+1}}{\Gamma(n)} \exp(-\theta x) \left[y^n \right]_0^x ,$$

from which the result follows immediately.

The expectation of the reciprocal total $1/x_+$, assuming that $n > 1$, is

$$\mathrm{E}\left(\frac{1}{X_+}\right) = \int_0^{+\infty} \frac{\theta^n x^{n-2}}{\Gamma(n)} \exp(-\theta x) \, \mathrm{d}x$$

$$= \frac{\theta}{n-1} \int_0^{+\infty} \frac{\theta^{n-1} x^{n-2}}{\Gamma(n-1)} \exp(-\theta x) \, \mathrm{d}x = \frac{\theta}{n-1} ,$$

after realising that the latter integrand is the density of a gamma distribution. By similar steps, we obtain the identity

$$\mathrm{E}\left(\frac{1}{X_+^2}\right) = \frac{\theta^2}{(n-2)(n-1)} ,$$

so long as $n > 2$. Hence the maximum likelihood estimator $\hat{\theta} = n/X_+$ has the expectation $\{1 + 1/(n-1)\}\theta$, that is, bias $\theta/(n-1)$, and MSE

$$\mathrm{MSE}\left(\frac{n}{X_+}; \theta\right) = \frac{n^2 \theta^2}{(n-2)(n-1)^2} + \frac{\theta^2}{(n-1)^2}$$

$$= \frac{(n+2)\theta^2}{(n-2)(n-1)} . \tag{2.7}$$

The estimator n/X_+ is biased; the bias can be eliminated by replacing the numerator n with $n - 1$. This is more efficient than the maximum likelihood estimator, as

$$\mathrm{var}\left(\frac{n-1}{X_+}\right) = \frac{\theta^2}{n-2} \; ;$$

compare with (2.7). In this example, we eliminated the bias and at the same time reduced the MSE. More precisely, we simultaneously reduced the (squared) bias and the variance. In some cases such a 'correction for bias' is counterproductive—it is accompanied by variance inflation that results in a net increase of MSE. Note, however, that the bias of n/X_+ converges to zero as $n \to \infty$, as does the gain in efficiency of $(n-1)/X_+$ over n/X_+. Asymptotically, n/X_+ is unbiased and efficient.

Next we consider maximum likelihood estimation of the expectation $1/\theta$. The estimator is the sample mean $\bar{X}$. Its expectation and variance are $1/\theta$ and $1/(n\theta^2)$, respectively, derived immediately from the expectation and variance of a random draw. We explore the estimators $c\bar{X}$ for positive constants c. Their biases are $(1-c)/\theta$ and MSEs

$$\frac{c^2}{n\theta^2} + \frac{(1-c)^2}{\theta^2} = \frac{c^2}{\theta^2}\left(1 + \frac{1}{n}\right) - 2\frac{c}{\theta^2} + \frac{1}{\theta^2} \; .$$

This quadratic function of c attains its minimum at

$$c^* = \frac{n}{n+1} \; .$$

The corresponding estimator, $n\bar{X}/(n+1)$, is biased,

$$\mathrm{B}\left\{\frac{n}{n+1}\bar{X}; \frac{1}{\theta}\right\} = \frac{1}{(n+1)\theta} \; ,$$

but its MSE,

$$\mathrm{MSE}\left(\frac{n}{n+1}\bar{X}; \frac{1}{\theta}\right) = \frac{1}{(n+1)\theta^2} \; ,$$

is smaller than for the maximum likelihood estimator $\bar{X}$. Asymptotically, as $n \to \infty$, the gain vanishes. But in practice we work (almost) exclusively with finite samples.

Example 3. Continuous Uniform Distributions. Let $x_1, x_2, \ldots, x_n$ be a random sample from the continuous uniform distribution on $(0, \theta)$, with an unknown positive parameter θ. We derive the maximum likelihood estimator of θ, and show that it is not efficient, not even asymptotically. Denote by $x_{\max}$ the largest outcome x_j, $j = 1, \ldots, n$, and by $X_{\max}$ its random-variable counterpart. The joint density of the outcomes is

$$f(\mathbf{x}; \theta) = \frac{1}{\theta^n} I\left(x_{\max} < \theta\right) \; ,$$

where I is the indicator function, equal to unity when its argument is true and to zero when it is false. This density is positive when θ is greater than

or equal to all x_j, that is, when $\theta \geq x_{\max}$. The log-likelihood for $\mathbf{x}$ is defined only when $\theta > x_{\max}$. Then

$$L(\theta; \mathbf{x}) = -n \log(\theta);$$

$x_{\max}$ is a linear minimal sufficient statistic. The log-likelihood is a decreasing function for all $\theta > x_{\max}$. Therefore, its maximum is at $\hat{\theta} = x_{\max}$. Note that the likelihood is not continuous at this point.

The distribution of this estimator is derived as the probability that no outcome x_j exceeds x:

$$P\left(X_{\max} \leq x\right) = \prod_{j=1}^{n} P(X_j < x) = \frac{x^n}{\theta^n}.$$

The corresponding density is

$$f(x) = \frac{nx^{n-1}}{\theta^n}.$$

The expectation and variance of this distribution are

$$\mathrm{E}\left(X_{\max}\right) = n \int_0^\theta \frac{x^n}{\theta^n}\, \mathrm{d}x \quad = \frac{n}{n+1}\,\theta,$$

$$\mathrm{var}\left(X_{\max}\right) = n \int_0^\theta \frac{x^{n+1}}{\theta^n}\mathrm{d}x - \left\{\mathrm{E}\left(X_{\max}\right)\right\}^2$$

$$= \theta^2 \left\{\frac{n}{n+2} - \frac{n^2}{(n+1)^2}\right\} \quad = \frac{n\theta^2}{(n+1)^2(n+2)}.$$

Therefore,

$$\mathrm{MSE}\left(X_{\max}; \theta\right) = \frac{n\theta^2}{(n+1)^2(n+2)} + \frac{\theta^2}{(n+1)^2} \quad = \frac{2\theta^2}{(n+1)(n+2)}.$$

We can eliminate the bias of $X_{\max}$ by multiplying the estimator by $1 + 1/n$. The sampling variance of the resulting estimator is

$$\mathrm{var}\left(\frac{n+1}{n}\, X_{\max}\right) = \frac{\theta^2}{n(n+2)}.$$

For large n, this is only about half of the MSE of the maximum likelihood estimator $X_{\max}$. In this example, the maximum likelihood theory breaks down because the likelihood is not smooth in the neighbourhood of θ.

2.3 Synthetic Estimation

In this section, we describe a generalisation of the synthetic estimator defined in Sections 1.1.1 and 1.2.1. We contrast it with model selection, selecting one of the candidate models.

Suppose there are $M + 1$ candidate models for a particular study; the models are indexed by integers $0, 1, \ldots, M$. Let $\hat{\theta}_m$ be the estimator derived under the assumption that model m is valid. Further, suppose $\hat{\theta}_0$ is unbiased, irrespective of which model is valid, and the estimators $\hat{\theta}_m$ are not linearly dependent, so that a combination $b_0\hat{\theta}_0 + b_1\hat{\theta}_1 + \cdots + b_M\hat{\theta}_M$ has zero variance only when all the coefficients b_m vanish.

A setting for which the general result that is derived next is intended in particular is that model 0 is a general model and all the other models are its submodels. In this setting, we say that model 0 is a *supermodel* or an *envelope* of models $1, \ldots, M$; model 0 contains their union. Model 0 is assumed to be valid at the outset, prior to data inspection. Denote by $\hat{\boldsymbol{\theta}}$ the vector of estimators $\hat{\theta}_m$, $m = 1, \ldots, M$, by $\mathbf{V}$ its variance matrix, by $\mathbf{B}$ the vector of its biases in estimating $\theta\mathbf{1}$ and by $\mathbf{C}$ the vector of the covariances of $\hat{\boldsymbol{\theta}}$ with $\hat{\theta}_0$:

$$\hat{\boldsymbol{\theta}} = \left(\hat{\theta}_1, \ldots, \hat{\theta}_M\right)^\top ,$$
$$\mathbf{V} = \mathrm{var}\left(\hat{\boldsymbol{\theta}}\right) ,$$
$$\mathbf{C} = \mathrm{cov}\left(\hat{\boldsymbol{\theta}}, \hat{\theta}_0\right) ,$$
$$\mathbf{B} = \mathrm{E}\left(\hat{\boldsymbol{\theta}}\right) - \boldsymbol{\theta} . \tag{2.8}$$

These (co-)variances and biases are evaluated assuming model 0. Note that $\hat{\theta}_0$ is not involved in $\hat{\boldsymbol{\theta}}$, $\mathbf{V}$, or $\mathbf{B}$. Let $V_0 = \mathrm{var}(\hat{\theta}_0)$.

The ideal composition of the estimators $\hat{\theta}_m$, $m = 0, 1, \ldots, M$, is defined as their convex combination with the smallest MSE. We show later that this combination is

$$\tilde{\theta}^* = \left(1 - \mathbf{b}^{*\top}\mathbf{1}\right)\hat{\theta}_0 + \mathbf{b}^{*\top}\hat{\boldsymbol{\theta}} , \tag{2.9}$$

where $\mathbf{b}^* = \mathbf{Q}^{-1}\mathbf{P}$, with

$$\mathbf{Q} = \mathrm{E}\left\{\left(\hat{\boldsymbol{\theta}} - \hat{\theta}_0\mathbf{1}\right)\left(\hat{\boldsymbol{\theta}} - \hat{\theta}_0\mathbf{1}\right)^\top\right\} ,$$
$$\mathbf{P} = \mathrm{cov}\left(\hat{\theta}_0\mathbf{1} - \hat{\boldsymbol{\theta}}, \hat{\theta}_0\right) .$$

The matrix $\mathbf{Q}$ is positive definite.

To prove this assertion, we consider the composition

$$\tilde{\theta} = \left(1 - \mathbf{b}^\top\mathbf{1}\right)\hat{\theta}_0 + \mathbf{b}^\top\hat{\boldsymbol{\theta}}$$

for arbitrary $M \times 1$ vector $\mathbf{b}$. Its MSE in estimating θ is

$$\mathrm{MSE}\left(\tilde{\theta}; \theta\right) = \left(1 - \mathbf{b}^\top\mathbf{1}\right)^2 V_0 + \mathbf{b}^\top\mathbf{V}\mathbf{b} + 2\left(1 - \mathbf{b}^\top\mathbf{1}\right)\mathbf{b}^\top\mathbf{C} + \mathbf{b}^\top\mathbf{B}\mathbf{B}^\top\mathbf{b} .$$

This is a quadratic function of $\mathbf{b}$, with its matrix-quadratic term equal to

$$V_0 \mathbf{1}\mathbf{1}^\top + \mathbf{V} - \mathbf{1}\mathbf{C}^\top - \mathbf{C}\mathbf{1}^\top + \mathbf{B}\mathbf{B}^\top = \mathbf{Q},$$

and so it is positive definite. Therefore $\mathrm{MSE}(\tilde{\theta}; \theta)$ has a unique minimum, and it can be found as the root of the vector of first-order partial differentials. Elementary matrix operations yield the identity

$$\frac{1}{2} \frac{\partial \mathrm{MSE}\left(\tilde{\theta}; \theta\right)}{\partial \mathbf{b}} = \left(V_0 \mathbf{1}\mathbf{1}^\top + \mathbf{V} - \mathbf{1}\mathbf{C}^\top - \mathbf{C}\mathbf{1}^\top + \mathbf{B}\mathbf{B}^\top\right) \mathbf{b} + \mathbf{C} - V_0 \mathbf{1}$$
$$= \mathbf{Q}\mathbf{b} - \mathbf{P}.$$

Hence the ideal synthetic estimator has the vector of coefficients $\mathbf{b}^* = \mathbf{Q}^{-1}\mathbf{P}$. Its MSE is

$$\mathrm{MSE}\left(\tilde{\theta}^*; \theta\right) = V_0 - \mathbf{P}^\top \mathbf{Q}^{-1} \mathbf{P}. \tag{2.10}$$

If the matrix $\mathbf{Q}$ and vector $\mathbf{P}$ were known, $\tilde{\theta}^*$ would be more efficient than any of the estimators $\hat{\theta}_m$, because the latter correspond to particular choices of $\mathbf{b}$, equal to $\mathbf{0}_M$ for $m = 0$ and to the unit vectors which comprise $M-1$ zeros except for unity in location m for $m = 1, \dots, M$. The MSE in (2.10) could be attained only if the matrices $\mathbf{Q}$ and $\mathbf{P}$ were known. In practice, the vector $\mathbf{b}^*$ is estimated, eroding some of the advantage of the composition $\tilde{\theta} = \tilde{\theta}(\mathbf{b}^*)$ over any one of the estimators $\hat{\theta}_m$. The composition $\tilde{\theta}$ can be defined for any collection of estimators $\hat{\theta}_m$; neither of them has to be maximum likelihood, although $\hat{\theta}_0$ has to be unbiased, or its bias should be very small. In the context of maximum likelihood or other estimators that are connected with a model, we refer to $\hat{\theta}_m$, $m = 0, 1, \dots, M$, as *single-model-based* and, because they contribute to $\tilde{\theta}^*$, as its *constituent* estimators. Note that these estimators may have good properties when the model they are derived for is valid. However, derivation of the synthetic estimator is based on their properties (joint distribution) when only the a priori specified (designated) model 0 is valid.

Example 4. Variance Estimation. We explore estimation of the variance of a random sample from a centred normal distribution $\mathcal{N}(0, \sigma^2)$. The mean square of the observations,

$$S = \frac{1}{n}\left(X_1^2 + \cdots + X_n^2\right),$$

is an obvious candidate. Its distribution is related to the χ^2 distribution with n degrees of freedom. This distribution is defined by the sum of squares of a sequence of n independent variables, each with standard normal distribution; see Exercise 1.13. From the properties of $\mathcal{N}(0, 1)$, it is easy to derive that $\mathrm{E}\left(\chi_n^2\right) = n$ and $\mathrm{var}\left(\chi_n^2\right) = 2n$. The number of generating variables (degrees of freedom), n, is indicated by the subscript.

We have

$$\frac{n}{\sigma^2} S \sim \chi_n^2,$$

confirming that S is unbiased for σ^2; further, $\text{var}(S) = 2\sigma^4/n$. As a much less credible alternative, consider the constant zero. Its bias is σ^2 and MSE is σ^4. For $n = 1$, zero is more efficient than S, and for $n = 2$ their MSEs coincide. For greater n, S is more efficient. The synthesis of S and zero corresponds to an estimator cS with suitably chosen $c > 0$. For $c < 1$, this can be interpreted as a *shrinkage*, pulling the 'original' unbiased estimator closer to zero. The minimum of the MSE,

$$\text{MSE}\left(cS; \sigma^2\right) = \left\{(1-c)^2 + \frac{2c^2}{n}\right\}\sigma^4,$$

is attained for $c^* = n/(n+2)$, when the MSE is $2\sigma^4/(n+2)$. Thus, a small bias, $\text{E}(c^*S) - \sigma^2 = 2\sigma^2/(n+2)$, is accompanied by a $(1+2/n)$-fold reduction of the MSE. This is modest for large n, but far from trivial for small n.

We come across χ^2 distributions frequently in variance estimation. For example, the ordinary least squares estimator of the residual variance in linear regression, $\hat{\sigma}^2 = \mathbf{e}^\top \mathbf{e}/(n-p)$, has a scaled χ^2 distribution; $(n-p)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p}$. Standard textbooks emphasise unbiased estimation; for small sample sizes we can do a bit better.

2.4 Model Selection

An alternative to synthetic estimation commits us to one of the single-model-based estimators, the selection of which is also based on the data. The importance of such a selection arises from the caveat of the maximum likelihood estimator (in addition to asymptotics)—when the model is valid the estimator is (asymptotically) efficient. Other model-based estimators are subject to similar caveats. Note that the theory does not state that the estimator is not efficient when the model is not valid.

We could protect our inferences against the lack of validity by defining very general models. Although they still cannot ensure validity, they do no harm to the chances of attaining this goal; if a narrower model is valid, then so is its generalisation. If we are committed to basing our inferences on a single model we have a strong incentive to select a narrower model because estimation is in general more efficient in a smaller parameter space, so long as it contains the parameter vector that governs the studied process. However, as we make our model narrower, we may drop the data-generating distribution, ending up with an invalid model.

A procedure for narrowing the model on which we base our inferences is called *model selection*. Model selection entails uncertainty; a different model may be selected in a replication. A typical model selection procedure arbitrates between two models, A and B, where B is a submodel of A. A statistic $t(\mathbf{y})$ is defined, together with a *critical value* t^*. If the realised value of $t(\mathbf{y})$ falls below t^*, model B is adopted; otherwise model A is adopted. Following the

selection (adoption) of either model, further model selection procedures may be applied. A collection of such procedures is called *multistage*. Instead of a critical value, a *critical region*, denoted by $\mathcal{C}$, may be specified. This may be an interval, such as (t_L^*, t_U^*) or its complement, with respective lower and upper limits t_L^* and t_U^*, but in principle any division of the support of the statistic $t(\mathbf{y})$ can be declared a critical region.

There is no straightforward recipe for choosing the statistic $t(\mathbf{y})$, but there are several well-founded criteria for assessing them. In many settings, model A is defined by a parameter space $\boldsymbol{\Theta}$ and model B by its subspace, such as $\boldsymbol{\Theta}_B = (\boldsymbol{\theta}; \theta_1 = 0)$, constraining the first component of $\boldsymbol{\theta}$ to a specific value. This setting is made much more general by allowing transformations of $\boldsymbol{\theta}$ or by imposing a constraint on a subvector of $\boldsymbol{\theta}$; $\boldsymbol{\Theta}_B = \{\boldsymbol{\theta}; g(\boldsymbol{\theta}) = \mathbf{0}\}$. The constraint function g is usually linear.

For the constraint $\theta_1 = 0$, we may use an estimator $\hat{\theta}_1$ of θ_1; the critical region for θ_1 comprises values of $\hat{\theta}_1$ that are distant from the 'special' value 0. The choice of the critical region can be guided by the desire to minimise the probability of making an erroneous choice. In one view, model A is always a correct choice because we assume it to be valid. However, when model B is also valid, we would regard the choice of A as an error, because we would forego some gains in efficiency by not using the narrower model. Therefore, we choose the critical region so that the probability of (inappropriately) selecting A when B is valid does not exceed a small value, such as $\alpha = 0.05$. Note that this probability is conditional or, more accurately, hypothetical, because it refers to the setting of model B, which need not be valid. The critical region is usually set to one or both tails of the hypothetical distribution of $t(\mathbf{y})$; t_L^* and t_U^* are set so that

$$\mathrm{P}\left\{t(\mathbf{y}) \notin (t_L^*; t_U^*) \,|\, \mathrm{B}\right\} \leq \alpha, \qquad (2.11)$$

or equal to α when it can be arranged. Special cases of practical importance are $t_L^* = -t_U^*$ (symmetric critical region), and $t_L^* = -\infty$ or $t_U^* = +\infty$ (one-sided critical regions). Another choice includes in the critical region the part of the support of $t(\mathbf{y})$ that has the smallest density; that is, a constant c is sought for which

$$\mathrm{P}\left[f\{t(\mathbf{y})\} < c \,|\, \mathrm{B}\right] \leq \alpha,$$

where f is the density of $t(\mathbf{y})$. When the density f is symmetric and increasing to the left and decreasing to the right of its single mode at zero, this criterion coincides with that in (2.11) with $t_L^* = -t_U^*$. Values of $\hat{\theta}_1$ may fall in the critical region even when $\theta_1 \neq 0$, but the probability of this event, when $\theta_1 = 0$, is small, so such cases can be regarded as exceptional.

A critical region is difficult to choose when the probability in (2.11) depends on $\boldsymbol{\theta}$ or, more accurately, on $\boldsymbol{\theta}_{-1}$, the subvector of $\boldsymbol{\theta}$ with the constrained component θ_1 removed. It may be opportunistic to simplify our task by choosing a statistic $t(\mathbf{y})$ that depends only on $\hat{\theta}_1$. However, this goal should

be subordinated to the principal purpose, appropriate model selection, or selection that results in an efficient estimator.

When choosing A or B, two kinds of error may be committed. One, described by (2.11), is a failure to narrow down to model B. The other is the inappropriate choice of B when B is not valid. The probability of such an error usually depends on $\boldsymbol{\theta}$, and almost always on θ_1. For example, it is close to $1 - \alpha$ when θ_1 is close to zero and the likelihood corresponding to model A is smooth. The probability of appropriately selecting model A,

$$\mathrm{P}\left\{t(\mathbf{y}) \in \mathcal{C}\right\},$$

as a function of $\boldsymbol{\theta}$, is called the *power* of the selection; it is denoted by β. (Note the potential conflict with the notation for regression parameters.)

The ideal choice of t and t^* for a given probability α is such that it has the highest possible power. As power is a function, procedures for model selection can be compared only partially. Some procedures may be more 'powerful' in certain regions of the parameter space $\boldsymbol{\Theta}$ and less powerful elsewhere. A procedure is called unbiased if its power exceeds α for all $\boldsymbol{\theta}$ associated with the complement of B in A (denoted by B \ A) and is smaller than or equal to α for any $\boldsymbol{\theta}$ when model B is valid. One procedure is said to be uniformly more powerful than another if it has a greater power (is more powerful) for any value of the parameter vector $\boldsymbol{\theta}$ in B \ A.

Example 5. ANOVA. Suppose outcomes y_{jk}, $j = 1, \ldots, n_k$, $k = 1, 2$, with $n_1 = 5$ and $n_2 = 50$, are generated according to the ANOVA model introduced in Section 1.1. We discuss estimation of the mean μ_1 for group $k = 1$. We consider the general model (A) in which the two groups have unrelated means μ_1 and μ_2 and its submodel (B) defined by the constraint $\mu_1 = \mu_2$. In both models we assume that the two groups have the same within-group variance σ_{W}^2, assumed to be known.

We base the model choice on the difference of the sample means $\Delta\hat{\mu} = \bar{y}_1 - \bar{y}_2$, an unbiased estimator of the population difference $\Delta\mu = \mu_1 - \mu_2$. If $|\Delta\hat{\mu}|$ is large we choose model A. If model B is valid, $\Delta\hat{\mu}$ is distributed according to $\mathcal{N}(0, g\sigma_{\mathrm{W}}^2)$, where $g = 1/n_1 + 1/n_2$. We set the critical region to $(-t^*, t^*)$, where t^* is the $(1 - \frac{1}{2}\alpha)$-quantile of $\mathcal{N}(0, g\sigma_{\mathrm{W}}^2)$; that is,

$$t^* = \sigma_{\mathrm{W}}\sqrt{g}\,\Phi^{-1}\left(1 - \tfrac{1}{2}\alpha\right),$$

where Φ is the distribution function of $\mathcal{N}(0, 1)$. Thus, we choose A, and with it $\hat{\mu}_1 = \bar{y}_1$ if $|\Delta\hat{\mu}| > t^*$, and B and $\hat{\mu}_1 = \bar{y} = (n_1\bar{y}_1 + n_2\bar{y}_2)/n$ otherwise.

The power of this selection procedure is

$$\mathrm{P}(|\Delta\hat{\mu}| > t^*; \Delta\mu) = \Phi\left(\frac{-t^* - \Delta\mu}{\sigma_{\mathrm{W}}\sqrt{g}}\right) + 1 - \Phi\left(\frac{t^* - \Delta\mu}{\sigma_{\mathrm{W}}\sqrt{g}}\right),$$

calculating the probabilities separately for negative and positive values of $\Delta\hat{\mu}$, distributed according to $\mathcal{N}(\Delta\mu, g\sigma_{\mathrm{W}}^2)$.

The model selection procedure defined by $\Delta\hat\mu$ and t^* is unbiased. This can be proved by differentiating the power function with respect to the difference $\Delta\mu$:

$$\frac{\partial \mathrm{P}(|\,\Delta\hat\mu\,| \,>\, t^*;\,\Delta\mu)}{\partial\Delta\mu} \;=\; \frac{1}{\sigma_{\mathrm{W}}\sqrt{g}}\left\{-\phi\left(\frac{-t^*-\Delta\mu}{\sigma_{\mathrm{W}}\sqrt{g}}\right) + \phi\left(\frac{t^*-\Delta\mu}{\sigma_{\mathrm{W}}\sqrt{g}}\right)\right\}.$$

As the density $\phi(x)$ of $\mathcal{N}(0,1)$ is symmetric and monotone for negative and positive values of x, the differential vanishes only when $\Delta\mu = 0$. So the power, as a function of $\Delta\mu$, attains its extreme at $\Delta\mu = 0$, and it is easy to check that this extreme is a minimum. Therefore, when $\Delta\mu \neq 0$, the power exceeds α.

Two model selection procedures are said to be *equivalent* if they yield the same decision (A or B) for every possible outcome $\mathbf{y}$. In particular, a procedure based on statistic t and critical value t^* is equivalent to the procedure based on $u(t)$ and $u(t^*)$ for any increasing function u.

2.4.1 Hypothesis Testing

The model selection procedure described in the previous section is also referred to as *hypothesis testing*. In its terminology, model B corresponds to the *null-hypothesis* and model A to the *alternative* hypothesis. We assume that the null-hypothesis is valid and abandon it in favour of the alternative only if the outcomes present sufficient evidence against B. That is, the null-hypothesis is regarded as the status quo and is overturned only when successfully challenged.

To test a particular hypothesis, a *test statistic* t is defined, together with its critical region, just like for model selection. When the realised value of the test statistic, $t(\mathbf{y})$, falls in the critical region, we regard it as evidence against the null-hypothesis. When $t(\mathbf{y})$ is outside the critical region, we adhere to the status quo, although the logically appropriate conclusion is that 'we do not know' whether the null-hypothesis is valid—we adhere to the null-hypothesis by default, not as a result of any evidence that confirms it. After adopting the null-hypothesis as an assumption, we have found no statistical contradiction with it—we have merely *failed to reject it*.

Instead of defining a test statistic $t(\mathbf{y})$ it suffices to split the outcome space, the set of all possible outcomes $\mathbf{y}$, into two subsets and designate one of them the critical region. Although this is a more general definition of a hypothesis test, it is of little practical use. Similarly, any subset of the parameter space $\boldsymbol{\Theta}$ can be declared as the null-hypothesis and its complement as the alternative. However, many practical settings can be expressed in the narrower format of the null-hypothesis being a subspace of the parameter space, defined by constraining one or several components of $\boldsymbol{\theta}$ to default (special) values.

One notable exception arises when the null-hypothesis has the form $\theta_1 > \theta_1^*$ or a similar inequality for one or several components of $\boldsymbol{\theta}$. Such a hypothesis

is called *one-sided*. For instance, in Example 5, adapted for hypothesis testing, we could replace the null-hypothesis defined by $\mu_1 = \mu_2$ with $\mu_1 < \mu_2$. We can use the same test statistic, $\Delta\hat{\mu}$, but a more suitable critical region is an interval $(t^*, +\infty)$, so that the hypothesis would be rejected for large (positive) values of $\Delta\hat{\mu}$. We choose the limit t^* as

$$t^* = \sigma_{\mathrm{W}}\sqrt{g}\,\Phi^{-1}(1 - \alpha)\,,$$

so that the probability of rejecting the null-hypothesis under the borderline assumption $\mu_1 = \mu_2$ is equal to α. It is easy to check that when $\mu_1 < \mu_2$, the probability of rejecting the null-hypothesis is smaller than α and that the power exceeds α whenever the alternative is valid. Thus, the hypothesis test based on $\Delta\hat{\mu}$ and t^* is unbiased.

2.4.2 Inference Following Model Selection

Having selected a model, A or B, we may apply further selection procedures, each time pitting the previous 'winner' against a new challenger model. At the end of such a string of selections we come to the concluding part of inference, a statement about a quantity of interest θ. We might regard the 'winning' model as valid and formulate all inferences assuming so. The profound drawback of this approach is that we ignore *model uncertainty.* In each model selection step, we may have been led astray by an inappropriate selection, not by our (the analyst's) fault, but by the inherent nature of each model selection step that it does not yield the 'correct' answer with certainty. This is easy to confirm by simulations, replicating the process of generating data and model selection.

Figure 2.1 provides an illustration based on the setting of Example 5, with $\mu_1 = 1$, $\mu_2 = 0$, $\sigma_{\mathrm{W}}^2 = 1$, $n_1 = 5$, and $n_2 = 50$. The left-hand panel, containing the plot of a simple random sample of 100 pairs of simulated values of the within-group sample means $\hat{\mu}_1$ and $\hat{\mu}_2$, shows that the hypothesis of equal means is rejected mostly when $\hat{\mu}_1$ is greater than μ_1. This should come as no surprise, because $\hat{\mu}_2$ has a small sampling variance $(1/50)$, and so the outcome of the hypothesis test depends mainly on the value of $\hat{\mu}_1$.

In the right-hand plot, the empirical distribution of the estimator $\hat{\mu}_1^\dagger$ that is based on the selected model is drawn. The distribution is distinctly bimodal, with the two mounds corresponding to rejection of the null-hypothesis (using $\hat{\mu}_1$, the shaded part of the histogram) and estimating μ_1 by $\hat{\mu}$. The solid vertical line indicates the target $(\mu_1 = 1)$ and the vertical dashes the expectation $\mathrm{E}(\hat{\mu}_1^\dagger)$. The MSEs of the estimators we considered are: $\mathrm{var}(\hat{\mu}_1) = \sqrt{1/5} = 0.45$, $\mathrm{MSE}(\hat{\mu}; \mu_1) = \sqrt{1/55 + (50/55)^2} = 0.92$, and $\mathrm{MSE}(\hat{\mu}_1^\dagger; \mu_1) = 0.46$, the last established by simulations. Note that the two mounds in the right-hand panel deviate from the normal distribution; they are *conditional* distributions of the form $(Y \mid I)$ where Y is normally distributed and I is a dichotomous variable correlated with Y. Thus, we set out with two normally distributed estimators,
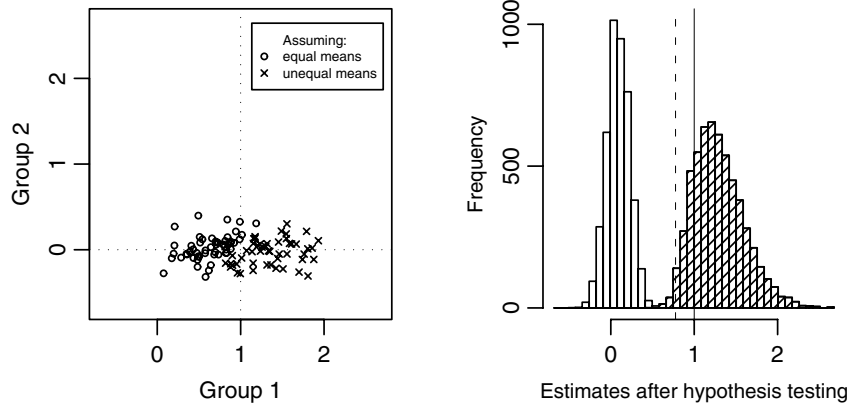
Fig. 2.1. Estimates based on hypothesis testing; the setting of Example 5 ($\mu_1 = 1$, $\mu_2 = 0$, $n_1 = 5$, $n_2 = 50$, and $\sigma^2 = 1$). In the left-hand panel, a random sample of simulated estimates of $\hat{\mu}_1$ and $\hat{\mu}_2$ is plotted; the right-hand panel is the histogram of the estimates of μ_1 after testing the hypothesis that $\mu_1 = \mu_2$; the values obtained after rejecting the null-hypothesis are represented by shading.

$\hat{\mu}_1$ and $\hat{\mu}$, attempting to use the better of them. In this task we have almost succeeded (0.46 vs. 0.45), but we ended up with an estimator that is biased and distinctly not normally distributed.

In the 10 000 replications that are summarised in Figure 2.1, the null-hypothesis was rejected in 5720 instances. If model uncertainty is ignored in these cases the standard error of $\sqrt{1/5} = 0.45$ would be reported, with a reference to $\mathcal{N}(0, 1/5)$, whereas in the remaining 4280 instances, $\sqrt{1/50} = 0.14$ would be reported. Thus the standard error would be substantially underestimated.

In summary, the process of model selection has a nontrivial impact not only on which model is selected and with what probability, but also on the distribution of the estimator, on the quality of the inference made, and on the assessment of this quality. We can rephrase this problem in the language of hypothesis testing as follows. Whichever test we apply, we cannot proceed by pretending that the outcome of the hypothesis test was known in advance of data inspection. We cannot ignore the uncertainty of the steps taken prior to the concluding act of applying the estimator that corresponds to the selected model.

The problem illustrated in Figure 2.1 cannot be resolved by applying a different test (different test statistic or different critical region). It does not appear for all configurations of means μ_1 and μ_2 and sample sizes n_1 and n_2, because in some settings the 'correct' decision is made by hypothesis testing with high probability. However, without knowing the values of μ_1 and μ_2, we are exposed to the risk of poor performance of the two-stage estimator. (The

two stages are model selection and evaluation of the estimator associated with the selected model.)

A more complete overview of the problem is given in the next section where two other model-selection criteria are introduced.

2.5 Model Selection Criteria Related to Likelihood

This section describes some common model selection criteria. They are all related to likelihood and to one test statistic in particular, the *likelihood ratio*. Throughout, we consider the setting with a general model A, assumed to be valid, with a p-dimensional parameter space $\boldsymbol{\Theta}$, and its submodel B defined by constraining the parameter vector $\boldsymbol{\theta}$ to a $(p-r)$-dimensional subspace of $\boldsymbol{\Theta}$, by means of a constraint (to zero) on each of r components of $\boldsymbol{\theta}$.

Let l_A and l_B be the maxima of the log-likelihood under the respective models A and B. That is, $l_A = l(\hat{\boldsymbol{\theta}}_A \,; \mathbf{y})$ and $l_B = l(\hat{\boldsymbol{\theta}}_B \,; \mathbf{y})$, where $\hat{\boldsymbol{\theta}}_A$ and $\hat{\boldsymbol{\theta}}_B$ are the maximum likelihood estimators under the respective models A and B. The likelihood ratio statistic is defined as $\Delta l = 2(l_A - l_B)$. Its practical importance is that the asymptotic sampling distribution of Δl is χ^2 with r degrees of freedom. This motivates the likelihood ratio test, which selects the general model A when Δl exceeds the 95th percentile of the χ_r^2 distribution. This test can be used for model selection, selecting the submodel B when Δl falls short of the 95th percentile of the χ_r^2 distribution. Under the regularity conditions, the likelihood ratio test based on Δl is asymptotically most powerful; that is, as the sample size increases, any other test of the same hypothesis is at best only slightly more powerful.

The test and the model selection procedure can be interpreted as follows. We strive for *model adequacy*, to obtain as high a (log-)likelihood as possible, while pursuing *model parsimony*, to use models with as few parameters (dimensions of the parameter space) as possible. Therefore, we select the more general model A only when it yields a much higher likelihood than the submodel B does. The likelihood ratio requires fitting both models A and B; this is a drawback only for some very complex models and large-scale datasets, because usually both estimation procedures require the same software.

The score test of the hypothesis $\theta_1 = 0$ is defined as

$$ s = \frac{\hat{\theta}_1}{\sqrt{\mathcal{I}(\theta_1)}} \,, $$

where $\mathcal{I}$ is the diagonal element of the information matrix that corresponds to parameter θ_1 and is evaluated at the default value $\theta_1 = 0$. Its asymptotic (large-sample) distribution, assuming model B, is standard normal. The general model is adopted when the value of $|s|$ is large, $|s| > \Phi^{-1}(1 - \alpha/2)$, so s^2 is an alternative form of the test statistic and its distribution under model B is χ_1^2, approximately, in large samples. The score test can be regarded as

comparing a model and its submodel defined by the constraint $\theta_1 = 0$. It requires fitting only the general model.

The *Akaike information criterion* (AIC) is an adjustment of the likelihood ratio test statistic. It is based on the statistic $2(l_A - l_B + r)$; it selects the submodel B when its value falls short of the 95th percentile of the χ_r^2 distribution. It corrects some of the deficiencies of the likelihood ratio criterion (for model selection) but is subject to uncertainty, just like any other procedure.

The *Bayesian information criterion* (BIC) adjusts the likelihood ratio statistic by $2r \log(n)$, where n is the sample size, so it is more likely to prefer the submodel B in all but very small datasets.

We revisit Example 5 with the ANOVA setting with two groups. The likelihood ratio test is equivalent to the test conducted in ANOVA. This is shown by elementary operations:

$$l_A - l_B = -\frac{1}{2\sigma_W^2} \sum_{k=1}^{2} \sum_{j=1}^{n_k} (y_{jk} - \hat{\mu}_k)^2 + \frac{1}{2\sigma_W^2} \sum_{k=1}^{2} \sum_{j=1}^{n_k} (y_{jk} - \hat{\mu})^2$$

$$= \frac{1}{2\sigma_W^2} \sum_{k=1}^{2} n_k (\hat{\mu}_k - \hat{\mu})^2$$

$$= \frac{n_1 n_2}{n\sigma_W^2} (\hat{\mu}_1 - \hat{\mu}_2)^2 .$$

This statistic is an increasing function of $\Delta\hat{\mu} = |\hat{\mu}_1 - \hat{\mu}_2|$.

Figure 2.2 summarises the selected-model-based estimators for the range of differences $\Delta\mu \in (0, 3)$. In panel A, each curve represents the bias $B(\hat{\mu}_1 ; \mu_1)$ as a function of $\mu_2 - \mu_1$. Each curve is based on empirical evaluation (50 000 replicates) for the 31 points $0.0, 0.1, \ldots, 3.0$. The biases and MSEs are plotted in the respective panels A and B for the probabilities α equal to 0.005, 0.01, 0.025, and 0.05. The MSEs in panel B and the powers of selection in panel C are constructed similarly. For the synthetic estimator, no model selection takes place, so the power of selection is not defined. In panel C, the synthetic estimator is represented by the empirical mean of the shrinkage coefficient.

As MSE is our criterion for efficiency, panel B is key. The MSEs of the model-based estimators are constant for model A, equal to 0.2, and quadratic for model B, equal to $0.02 + \Delta\mu^2$. The estimator $\hat{\mu}_B = \hat{\mu}$ based on B is inefficient for all but very small differences $\Delta\mu$. The estimators based on model selection criteria (LR, AIC and BIC) are also efficient for small values of $\Delta\mu$ but are very inefficient for a wide range of values of $\Delta\mu$. For large values of $\Delta\mu$ they are almost as efficient as estimator $\hat{\mu}_A$.

Except for estimator $\hat{\mu}_A = \hat{\mu}_1$, the synthetic estimator has by far the smallest maximum MSE over the range $(0, 3)$. It is not uniformly more efficient than the selected-model-based estimators, but it does not have their glaring weaknesses. It is least efficient when $\Delta\mu$ is small—when the estimation problem is, in a way, the easiest. Similarly, for large $\Delta\mu$, when the data should
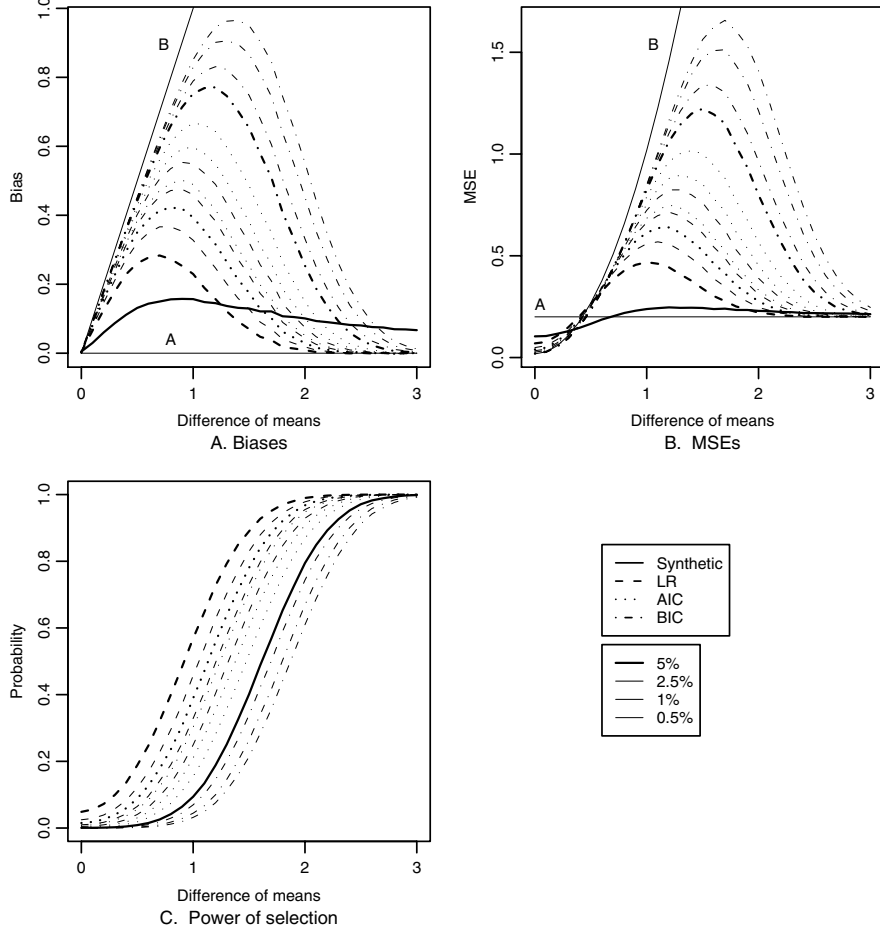
Fig. 2.2. The biases, MSEs, and powers of selection using likelihood ratio (LR), AIC and BIC, and synthesis, using a range of probabilities α, with the ANOVA setting of Example 5: $n_1 = 5$, $n_2 = 50$, $\mu_1 = 0$, $\sigma^2 = 1$, and μ_2 in the range $(0, 3)$. For synthesis, the power of selection in panel C is replaced by the mean of the shrinkage coefficient. A and B in panel B denote the MSEs of the estimators based on the respective models A and B.

strongly indicate that $\Delta\mu$ is positive, the synthetic estimator is not the most efficient, but the MSEs of the competing estimators differ little. Thus, the main strength of the synthetic estimator is that it does not have any weaknesses; for no values of $\Delta\mu$ is its performance very poor. Panel A indicates that the bias is a substantial contributor to the MSE for the selected-model-based estimators. Panel C shows that the power of selection is related to the MSE. Using the mean shrinkage coefficient as a substitute for the power for the
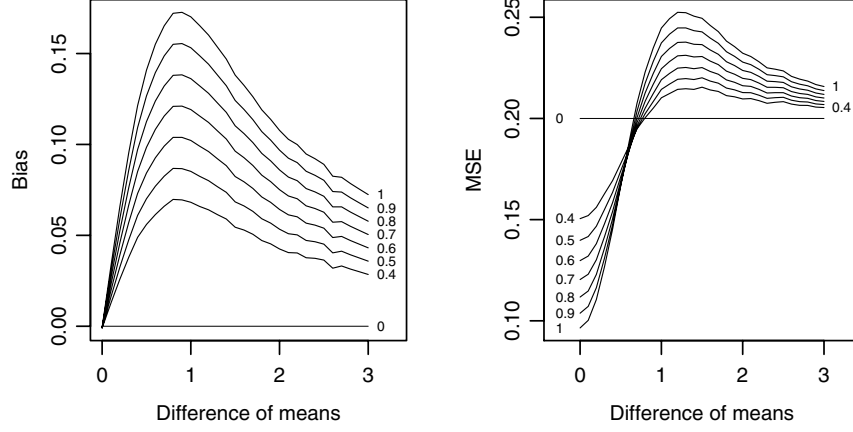
Fig. 2.3. The MSEs of the synthetic estimators with reduced shrinkage.

synthetic estimator, we see that appropriate selection with high probability does not imply small MSE.

Recall that in Section 1.1.1 we constructed an estimator uniformly more efficient than both $\hat{\mu}_1$ and $\hat{\mu}$ but it required information about the largest plausible value of $\Delta\mu$. Such information would be difficult to incorporate in the selected-model-based estimators.

All the estimators can be regarded as trading off small MSE when $|\Delta\mu|$ is small against large MSE when $|\Delta\mu|$ is large. Which estimator is best for the purpose? The answer lies in declaring *our* purpose; we have a freedom (and responsibility) to do it to suit our needs. This is rarely straightforward even for a single party. When several parties have a stake in the outcome of the study, some compromise is necessary between the parties' objectives, which may be in mutual conflict. But we should never shy away from exploring a range of purposes and then settle on an estimator that represents their compromise. Such an exploration should not look at the possible results (estimates) but at properties of estimators, MSEs as functions of parameters and design settings.

The synthetic estimator is attractive if we wish to minimise the maximum of $\mathrm{MSE}(\hat{\mu};\mu_1)$, which we interpret as having no weaknesses. However, by this criterion the sample mean estimator $\hat{\mu}_1$ is superior, negating all our efforts to improve on it. We can explore some variations of the synthetic estimator, by reducing its shrinkage coefficient. As a result, the estimator is improved for large values of $\Delta\mu$, at the price or reduced efficiency for small values of $\Delta\mu$. Figure 2.3 presents the results graphically, using the estimators $\tilde{\mu}_1(b)$ with the shrinkage coefficients $b = rg_1/(g_1 + \hat{\gamma}^2)$, for $r = 0.4, 0.5, \ldots, 1.0$.

The diagram shows that as we reduce the coefficient b we also reduce the bias uniformly and trade off efficiency at small values of $\Delta\mu$ for improvement at large values of $\Delta\mu$. (The breakpoint at which all the plotted functions

intersect is around $\Delta = 0.7$.) At the extreme, for $r = 0$, we match the even performance of $\hat{\mu}_1$.

This extended example should in no way be interpreted as evidence of superiority of the synthetic estimator over selected-model-based estimators. Instead, the example outlines how alternative estimators can be explored and what entails the decision to choose one of them. First, the search is rarely complete and definite, because we can rarely identify an estimator that is uniformly more efficient than all its competitors. Second, we have to weigh carefully the advantages and drawbacks of the alternative estimators and, possibly, supplement our criteria with what we regard as 'good' estimation. And finally, the search may indicate how prior information, additional to the analysed dataset, can make the search more effective. For example, if in Example 5 we knew that $|\Delta| < 0.5$ we would zoom in our attention on the appropriate part of panel B in Figure 2.2.

Nevertheless, some general comments can be made about the two classes of estimators, synthetic and selected-model-based. By model selection, we aim to match the most efficient of the competing estimators. This 'ambition' is not achieved because the selection is imperfect. Synthesis aims higher, to outperform each of the constituent estimators. It fails to achieve this goal because the ideal shrinkage coefficient can only be estimated.

For two constituent estimators (models), both classes of estimators have the form

$$\tilde{\theta} = \left(1 - \hat{B}_\theta\right)\hat{\theta}_1 + \hat{B}_\theta\,\hat{\theta}_2\,. \tag{2.12}$$

With model selection, $\hat{B}_\theta$ 'estimates' the model to be used ($\hat{B}_\theta$ is a binary variable, with possible values 0 and 1), whereas with synthesis it estimates the ideal combination of the constituent estimators. The description by (2.12) has an obvious extension to more than two constituent estimators for synthesis:

$$\tilde{\theta} = \sum_{m=0}^{M} \hat{B}_{\theta,m}\,\hat{\theta}_m\,,$$

with the constraint that $\hat{B}_{\theta,0} + \hat{B}_{\theta,1} + \cdots + \hat{B}_{\theta,M} = 1$. Model selection entails the additional constraint that $(\hat{B}_{\theta,0}\,,\hat{B}_{\theta,1}\,,\ldots,\hat{B}_{\theta,M})$ is multinomial—it contains M zeros and one unity.

Although model selection usually proceeds by choices within pairs, leading to multistage selection, it is equivalent to a single-stage selection from among several models. Synthesis can also be conducted in stages, for example, first by combining estimators A and B, then by combining their synthesis with C, and so on. At first, this might seem not to be useful because the constituent estimators could be combined directly. However, synthesis of many estimators is problematic because the matrix $\mathbf{Q}_\theta^{-1}$ may be estimated inefficiently even when $\mathbf{Q}_\theta$ is estimated efficiently, especially when $\mathbf{Q}_\theta$ or $\hat{\mathbf{Q}}_\theta$ is close to singularity. In such a case, some estimators (models) can be discarded from the

synthetic estimator because they can themselves almost be combined from the other constituent estimators.

As the number of observations increases and the parameter space is unaltered, the probability of selecting the appropriate model increases and the shrinkage coefficients converge to a unit vector $(0, \ldots, 0, 1, 0, \ldots, 0)$ or the zero vector $\mathbf{0}$, so that synthesis is based on a single model. Thus, asymptotically, model selection is not an issue. As the sampling variance of every estimator converges to zero, our attention should focus on eliminating the bias, and this is best done by applying the most complex model. Asymptotically, we do not have to pursue model parsimony, because one or a few redundant parameters (degrees of freedom used unnecessarily) inflate the sampling variance only slightly when we have degrees of freedom in abundance. Asymptotically, maximum likelihood has no competitors, so long as the regularity conditions are satisfied, and model adequacy should be the only concern. In practice, asymptotics is usually far away and, if we are committed to working with a single model, parsimony is highly relevant. Synthesis frees us up from this constraint but does not offer a uniformly more efficient solution.

Suggested Reading

The classical text [27] contains a comprehensive treatment of the likelihood theory, including proofs of all the key properties of maximum likelihood estimators. For a more recent monograph on likelihood, see [143]. The original references to the two information criteria, AIC and BIC, are [3] and [175], respectively. Maximum likelihood estimators for many models are evaluated using methods for numerical optimisation. Useful references to such methods are [35, 58], and [94]. A more recent monograph [98] is addressed specifically to statisticians.

Problems and Exercises

2.1. Write down the likelihood for a random sample of size n from the binary distribution with unknown probability p. How does it differ from the likelihood for a single draw from the binomial distribution $\mathcal{B}(n, p)$ with known sample size n and unknown probability p? Find a linear sufficient statistic for p. Derive the maximum likelihood estimator of p. Check that its sampling variance agrees with the reciprocal of the expected information and explain why this agreement is not maintained for the parameter $v = p(1 - p)$.

2.2. Generate random samples from a binomial distribution of your choice and verify empirically the properties of the maximum likelihood estimator $\hat{p}$ of the proportion p. Experiment with estimators of the form $c\hat{p}$ for a range of values of c near unity. Show that if p were known the minimum MSE would be attained for $c = p/\{p + (1 - p)/n\}$.

2.3. Suppose the probability $p > 0$ is known but the sample size n in a random sample from a binary distribution is not. Derive the likelihood for this setting with parameter n. How does this likelihood differ from the likelihood(s) in Exercise 2.2? Compare the maximum likelihood estimator of n with $\hat{n} = k/p$, where k is the number of positive outcomes.

2.4. Consider the following experiment comprising binary outcomes (success and failure). We keep generating outcomes independently from a binary distribution with probability $p > 0$ and stop when we reach a given positive number M of successes. Derive the likelihood for p and the maximum likelihood estimator. Explain why you would expect it to have a positive bias, especially for small M. Check your conclusion by simulations.

2.5. Derive the results in Example 1 without the aid of matrix algebra and matrix differentiation.

2.6. The *beta distributions* are defined by the densities

$$f(y; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1}(1-y)^{b-1}$$

for positive constants (parameters) a and b; their support is the interval $(0, 1)$. (Note that the uniform distribution corresponds to $a = b = 1$.) Find a set of sufficient statistics for a and b. Suppose we know that a and b are integers. How would you go about maximising the likelihood? Derive the expectation and variance of the beta distribution and use them to derive a moment-matching estimator.
Hint: A *moment-matching estimator* is defined as a solution of an equation, or of a set of equations, that matches sample moments (expectation, variance, and the like) to the population moments expressed as functions of parameters. This method is called the *method of moments*.

2.7. The beta distributions with $b = 1$ are called *power distributions*. Show that they can be derived as powers of the continuous uniform distribution. Derive the Cramér–Rao inequality for the power distributions. Compare it with the sampling variance of the moment-matching estimator based on the expectation and variance. Explain why the moment-matching estimator is not efficient.
Hint: Look at the sufficient statistic for a.

2.8. Show that the χ^2 distributions are a subset of the gamma distributions.
Hint: Show this first for χ_1^2 distribution and then proceed by induction.
Using the densities of the gamma distributions, check that the expectation and variance of χ_n^2 are n and $2n$, respectively. Check these results using the definition of χ_n^2 by construction from a random sample from $\mathcal{N}(0, 1)$; see Exercise 1.13. Derive the expectation and variance of the reciprocal of χ^2, that is, of a variable X such that $1/X \sim \chi_n^2$.
Hint: Relate the integrals $\int y^{-k} f(y) \mathrm{d}y$, $k = 1, 2$ to the densities of some χ^2 distributions.

2.9. Either using Exercise 1.14 or independently, prove that the (unbiased) estimator of the population variance σ^2, $\hat{\sigma}^2 = (\mathbf{y} - \bar{y})^\top (\mathbf{y} - \bar{y})/(n-1)$, based on a random sample $\mathbf{y}$ of size n from $\mathcal{N}(\mu, \sigma^2)$ has χ^2 distribution with $n-1$ degrees of freedom. Compare the efficiencies of the estimators with denominators n (the maximum likelihood estimator), $n-1$, and $n+1$. Find the denominator that yields the efficient estimator. Extend this result to estimating the residual variance in ordinary regression.

2.10. Estimate the reciprocal of the variance σ^2 in the setting of the previous example. Find an estimator more efficient than $1/\hat{\sigma}^2$.
Hint: Use the results of Exercise 2.8.

2.11. Suppose $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ are the ordinary least squares estimators with respective models 1 and 2. Models 1 and 2 may be invalid. Let $\hat{\beta}_{\mathrm{x}}$, $\hat{\beta}_{\mathrm{x1}}$, and $\hat{\beta}_{\mathrm{x2}}$ be the slopes on a covariate included in both models. Show that $\mathrm{cov}(\hat{\beta}_{\mathrm{x1}}, \hat{\beta}_{\mathrm{x2}}) = \mathrm{var}(\hat{\beta}_{\mathrm{x2}})$. Generalise this result to a sequence of K nested models and describe the pattern of the variance matrix of the K estimators of the coefficient with respect to the same covariate.

2.12. Let $(v_1, \ldots, v_K)$ be a decreasing sequence of positive numbers and $\mathbf{V}$ the matrix defined by its elements $V_{kh} = v_{\max(k,h)}$. Find the determinant and inverse of $\mathbf{V}$.
Hint: Proceed by induction. Find in the literature on matrices formulae for the determinant and inverse of a partitioned matrix $\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^\top & \mathbf{C} \end{pmatrix}$; in our case, $\mathbf{B} = v_K \mathbf{1}_{K-1}$ and $\mathbf{C} = v_K$.

2.13. Compile a programme for simulating the selected-model-based and synthetic estimators of the expectation for a group in the setting of ANOVA with several (say, eight) groups of ten observations each. Set the differences among the groups in such a way that the target group has in one case an extreme expectation, in another case is close to the mean of the expectations, and in another has about the average deviation from the mean of the within-group expectations. Describe the results of the simulations and present them in a diagram.

2.14. The *F-distribution* with n_1 and n_2 degrees of freedom is defined as the ratio of two scaled independent χ^2-distributed variables with n_1 and n_2 degrees of freedom in the numerator and denominator, respectively:

$$X = \frac{X_1}{X_2} \frac{n_2}{n_1}$$

where $X_1 \sim \chi^2_{n_1}$ and $X_2 \sim \chi^2_{n_2}$. Derive the expectation and variance of these distributions.

2.15. Consider the ordinary regression model

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \varepsilon_j$$

for two covariates, X_1 and X_2, with $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, independently. Derive the likelihood ratio test statistic for the hypothesis that $\beta_2 = 0$.

Hint: Replace the variable X_2 with $X_2^* = X_2 - b_1 X_1 - b_0$ with b_1 and b_0 set so that X_2 is orthogonal to both X_1 and the intercept $\mathbf{1}$.

Show that the test is equivalent to rejecting the hypothesis for large values of $\hat{\beta}_2^* / \sqrt{\widehat{\mathrm{var}}\left(\hat{\beta}_2^*\right)}$, where $\hat{\beta}_2^*$ is the least squares estimator of the slope on X_2.

2.16. Compare (analytically) the estimators $(1 + 1/n) \max_j X_j$ and $2\bar{X}$ for the parameter θ of the uniform distribution on $(0, \theta)$. Do you think the estimator $\min_j X_j + \max_j X_j$ is more efficient than either of these? Check your view by simulations.

Explore the analogous problem with the uniform distribution replaced by the distribution that is formed as the mean of a random sample of size K from $\mathcal{U}(0, \theta)$. Show that this distribution is beta. Check by simulations that this distribution converges to the normal as $K \to \infty$. Presumably, as K increases, twice the sample mean becomes a relatively more efficient estimator of θ than $(1 + 1/n) \max_j X_j$. Can you confirm this? For which K are the two estimators about equally efficient? (You may consider all positive numbers for K, not only integers.)

2.17. The *Poisson distributions* are defined by the probabilities

$$P(Y = k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

for $k = 0, 1, \dots$ and parameter $\lambda > 0$. Show that $E(Y) = \mathrm{var}(Y) = \lambda$ and that the sum of two independent variables, each with a Poisson distribution, also has a Poisson distribution. For a simple random sample from a Poisson distribution, find the maximum likelihood estimator of λ and its sampling variance. Explain why and verify by simulations that λ is estimated more efficiently as the sample mean than as the sample variance.

2.18. A local authority conducted an experiment in which all men below the age of 21 were encouraged to be at home by 11 pm every night during the months April to October 2004. To evaluate its success, they compared the numbers of reported public-order offences that involved young men in this period of 30 weeks with the same period in 2003 (dataset `EX2a.dat` on `www.sntl.co.uk/BookA/Data`). Test by the likelihood ratio the hypothesis that the average weekly numbers of offences are the same in the two years, assuming that each sequence of 30 outcomes is a random sample from a Poisson distribution. Assess how valuable it is to know that both samples are from Poisson distributions. Check and discuss how realistic such an assumption is. Discuss the problems with interpreting the result of the test given that it is

not possible to implement all the principles of experimental design.

Hint: Generate many pairs of random samples from the Poisson distributions with the same means as the two years have in the data, calculate their variances as features, and plot the pairs of these variances, together with the realised pair of the within-year means which, according to the adopted model, are unbiased estimators of the variances.

2.19. *Permutation test.* As an alternative to the solution in the previous exercise consider the following. Generate a *permutation dataset* by assigning the two observations for a week to the years 2003 and 2004 at random, with these assignments being independent across the weeks. Evaluate the difference of the within-year sample means. Replicate this process many times and compare the (realised) version of this difference with the distribution of its permutation counterparts. Reject the null-hypothesis of equal means if the realised difference is in the tail of the distribution of simulated (permutation) differences. Discuss the relevance of the assumptions of the Poisson distributions and of independence among the weeks for this test.