Chapter 3

Maximum likelihood estimation

3.1 Theory

Consider the problem of estimating the unknown parameter from a population when the population distribution is known (up to the unknown parameter or unknown vector of parameters), and when a random sample of n observations has been drawn from that population. Common examples are sampling from a Bernoulli distribution with unknown parameter π, sampling from a Poisson distribution with unknown parameter λ, or sampling from a normal distribution with unknown parameters μ and σ. Generically, we can write the probability or density function of y_i, $i = 1, \ldots, n$ as $f(y_i; \theta)$, where y_i is the i-th drawing from the distribution and θ is the unknown parameter. Because of the assumption of independent sampling, we can write the joint probability or density function of the sample as

$$f(y_1, \ldots, y_n; \theta) = \prod_{i=1}^{n} f(y_i; \theta) \tag{3.1}$$

For notational simplicity, we introduce maximum likelihood estimation in the context of such marginal models. As will be shown later, the extension of the maximum likelihood method to conditional models $f(y_i|x_i; \theta)$, as those described in the previous chapter, does not change any of the basic results, although some of the implementation issues become more complex. In this case the assumption of random sampling extends to pairs of observations (y_i, x_i), requiring that they are independent $\forall i$.

3.1.1 Likelihood Function

In the case of a discrete random variable, equation (3.1) is the joint probability of the sample given the parameters. In the case of a continuous random variable, it is the joint density. Alternatively, we can interpret (3.1) not as a function of the random sample $y_1, \ldots, y_n$ given the parameter θ, but rather as a function of θ for a given random sample $y_1, \ldots, y_n$. When we do this, we call (3.1) the *likelihood function*, and we denote it by capital L:

$$L(\theta; y) = \prod_{i=1}^{n} L(\theta; y_i) = \prod_{i=1}^{n} f(y_i; \theta) \tag{3.2}$$

where $y = (y_1, \ldots, y_n)$, $L(\theta; y) = L(\theta; y_1, \ldots, y_n)$ is the likelihood function of the sample, and $L(\theta; y_i)$ is the likelihood contribution of one individual observation. Despite the apparent equality in the expression underlying the two equations (3.1) and (3.2), it is worth emphasizing the subtle conceptual difference between the two. The likelihood function $L(\theta; y)$ is not a probability function since the argument is θ, not y, and θ is a parameter, not a random variable. On the other hand, for each point θ_0, $L(\theta_0)$ is a random variable since it depends on the random sample that will be drawn. Similarly, all functions of $L(\theta; y)$, such as $\log L(\theta; y)$ or $\partial \log L(\theta; y)/\partial\theta$ are random variables.

Actually, the definition of the likelihood function in (3.2) is somewhat more restrictive than necessary. Any function that is proportional to (3.1) can serve as likelihood function, that is, we require that $L(\theta; y_1, \ldots, y_n) \propto \prod_{i=1}^{n} f(y_i; \theta)$. This is the definition that was given by Fisher (1922, p. 310) in his original development of the method:

> “The likelihood that any parameter (or set of parameters) should have any assigned value (or set of values) is proportional to the probability that if this were so, the totality of observations should be that observed.”

By taking logarithms of equation (3.2) we obtain the *log-likelihood function*

$$\log L(\theta; y_1, \ldots, y_n) = \sum_{i=1}^{n} \log f(y_i; \theta) \tag{3.3}$$

The log-likelihood function will be central in the following analysis.

Example: Sampling from a Bernoulli Distribution

Assume that a random sample of size 5 is drawn from a Bernoulli distribution with parameter π. Then

$$L(p; y_1, \ldots, y_5) = \prod_{i=1}^{5} (1-p)^{1-y_i} p^{y_i}$$

and

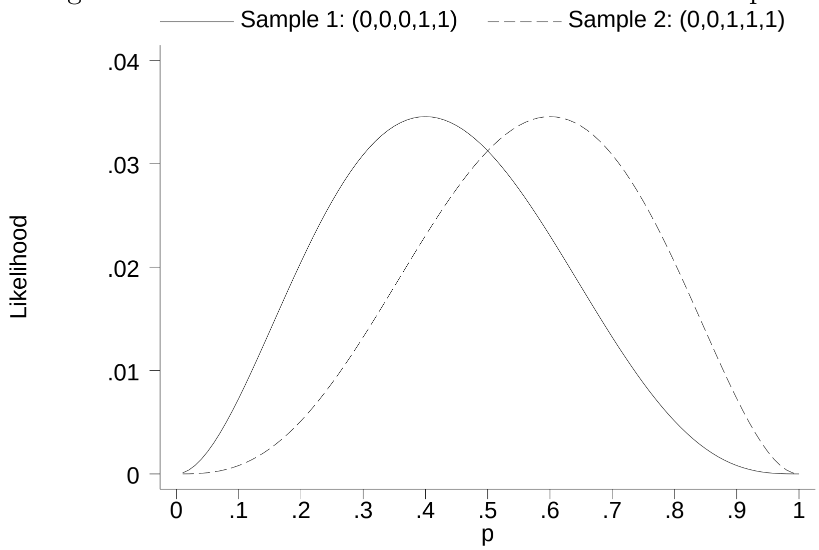
$$\log L(p; y_1, \ldots, y_5) = \sum_{i=1}^{5} (1-y_i)\log(1-p) + y_i \log p$$

To illustrate that the likelihood function is a random variable, Figure 3.1 plots the likelihood function for two different samples, one being the sequence $\{0,0,0,1,1\}$, the other being the sequence $\{0,0,1,1,1\}$. Under random sampling, the order of observations does not matter, so the difference between the two samples is the number of ones and zeros. The probability distribution of the log-likelihood function depends on the unknown true population parameter π. If for example $\pi = 0.4$, the likelihood function for the first sample has probability 0.035, whereas the likelihood function of the second sample has probability 0.023.

3.1.2 Expected Score and Information Matrix Equality

Consider the functions $\log L(\theta; y)$, $\partial \log L(\theta; y)/\partial\theta$, and $\partial^2 \log L(\theta; y)/\partial\theta\partial\theta'$. The first and second derivatives may be scalars or vector valued expressions, depending on the dimension of θ. The first derivative of the log-likelihood function, $\partial \log L(\theta; y)/\partial\theta$, is called the *score function*, or simply *score*. When appropriate, we will simply denote it as $s(\theta; y)$. The second derivative of the log-likelihood

Figure 3.1: Likelihood Function for the Bernoulli Example



function, $\partial^2 \log L(\theta; y)/\partial\theta\partial\theta'$ – a matrix if θ is a vector – is commonly refered to as the *Hessian matrix*. We will denote it as $H(\theta; y)$.

Because of the additivity of the log-likelihood function, the first and second derivatives are additive functions as well. For example,

$$\frac{\partial \log L(\theta; y)}{\partial\theta} = \sum_{i=1}^{n} \frac{\partial \log L(\theta; y_i)}{\partial\theta} = \sum_{i=1}^{n} \frac{\partial \log f(y_i; \theta)}{\partial\theta}$$

The first and second derivatives depend on the sample y_i and hence are random variables (they differ in repeated samples). In the following, we will derive the expectation of the score vector and of the Hessian matrix. Minus the expectation of the Hessian is also known as *information matrix* $I(\theta)$.

Expected Score

Consider first a single element of the score vector, $\partial \log L(\theta; y_i)/\partial\theta = \partial \log f(y_i; \theta)/\partial\theta$ and assume for simplicity the case where y_i is a continuous random variable such that

$$\int_{-\infty}^{\infty} f(y_i; \theta) dy_i = 1$$

In this expression, θ is understood to be the true, albeit unknown, parameter of the model. We can take the derivative of this identity with respect to θ.

$$\frac{\partial}{\partial\theta} \int_{-\infty}^{\infty} f(y_i; \theta) dy_i = 0$$

Under mild regularity conditions that allow us to interchange the order of differentiation and integration on the left side of the equation, we obtain

$$\frac{\partial}{\partial\theta} \int_{-\infty}^{\infty} f(y_i; \theta) dy_i = \int \frac{\partial f(y_i; \theta)}{\partial\theta}\, dy_i = \int \frac{\partial \log f(y_i; \theta)}{\partial\theta} f(y_i; \theta)\, dy_i = \mathrm{E}\left[\frac{\partial \log f(y_i; \theta)}{\partial\theta}\right] = 0$$

In words, the expected value of each element of the score vector, if evaluated at the true parameter, is zero. As a consequence, since

$$\mathrm{E}\left[\frac{\partial \log L(\theta; y)}{\partial \theta}\right] = \sum_{i=1}^{n} \mathrm{E}\left[\frac{\partial \log f(y_i; \theta)}{\partial \theta}\right]$$

the expected value of the vector of first derivatives of the log-likelihood function, or the score vector of the sample, is zero at the true parameter.

Example: Sampling from a Bernoulli Distribution

Assume as before that a random sample of size 5 is drawn from a Bernoulli distribution with parameter π. Then, the score function is given by

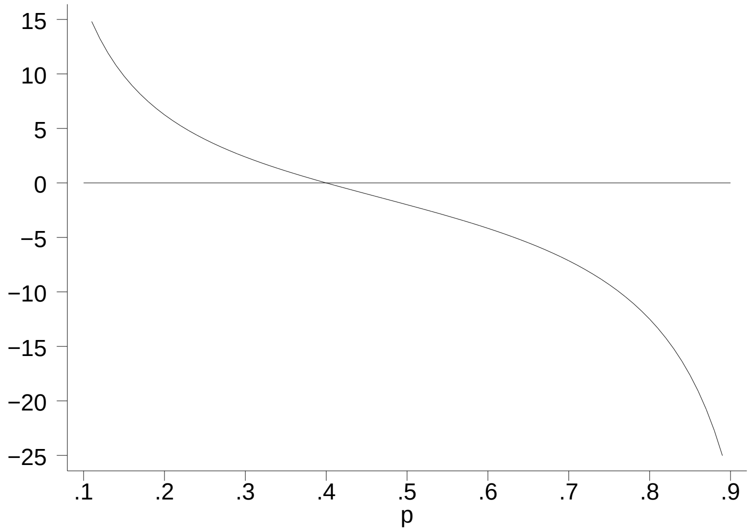
$$\frac{d \log L(p; y_1, \ldots, y_5)}{dp} = \sum_{i=1}^{5} -\frac{1-y_i}{1-p} + \frac{y_i}{p} = \sum_{i=1}^{5} \frac{(y_i - p)}{p(1-p)}$$

Taking expectations, we obtain

$$\mathrm{E}\left[\frac{d \log L(p; y_1, \ldots, y_5)}{dp}\right] = \sum_{i=1}^{5} \frac{(\mathrm{E}[y_i] - p)}{p(1-p)} = \frac{5(\pi - p)}{p(1-p)}$$

Evaluating this function of p at the true parameter vector, i.e., at the point $p = \pi$, we see that $\mathrm{E}(d \log L(p; y_1, \ldots, y_5)/dp)|_{p=\pi} = 0$ as required. Figure 3.2 plots the expected score function for $\pi = 0.4$ and various values of p.

Figure 3.2: Expected Score Function for the Bernoulli Example



Information Matrix

Now consider the second derivative of the log-likelihood function – a matrix if θ is a vector:

$$H(\theta; y) = \frac{\partial^2 \log L(\theta; y)}{\partial\theta\partial\theta'} = \sum_{i=1}^{n} \frac{\partial^2 \log f(y_i; \theta)}{\partial\theta\partial\theta'}$$

Moreover, define the *information matrix* of a sample as

$$I(\theta) = -\mathrm{E}[H(\theta; y)] \tag{3.4}$$

The information matrix is important in a number of ways in the development of maximum likelihood methodology. First, the information matrix can be used to assess whether the likelihood function is "well behaved". This relates to the issue of identification that is discussed below in chapter xxx. A lack of identification means that no matter how large the sample size, the information provided by it is insufficient to estimate the parameters of interest.

Second, the information matrix is important, because it is the inverse of the variance of the maximum likelihood estimator, a result we will derive in chapter xxx. And third, it links results for maximum likelihood estimation to an important result on the precision of estimators from general estimation theory, the so-called Cramér Rao lower bound. This result states that, under certain regularity conditions, the variance of an unbiased (efficient?) estimator of a parameter θ will always be at least as large as $I(\theta)^{-1}$.

One result pertaining to $I(\theta)$ is the so-called *information matrix equality.* This equality establishes that the information matrix can be derived in two ways, either as minus the expected Hessian, as in (3.4), or alternatively as the variance of the score function. In other words,

$$\mathrm{Var}[s(\theta; y)] = -\mathrm{E}[H(\theta; y)] \tag{3.5}$$

The derivation of the variance of the score function is based on the following considerations. As before, $f(y_i; \theta)$ denotes the probability model for the i-th observation. The Hessian matrix for this observation can be written as

$$\begin{aligned} H(\theta; y_i) &= \frac{\partial s(\theta; y_i)}{\partial\theta'} \\ &= \frac{\partial}{\partial\theta'} \frac{\partial f(y_i; \theta)/\partial\theta}{f(y_i; \theta)} \\ &= \frac{\partial^2 f(y_i; \theta)/\partial\theta\partial\theta'}{f(y_i; \theta)} - \frac{\partial f(y_i; \theta)/\partial\theta}{f(y_i; \theta)^2} \frac{\partial f(y_i; \theta)}{\partial\theta'} \\ &= \frac{\partial^2 f(y_i; \theta)/\partial\theta\partial\theta'}{f(y_i; \theta)} - s(\theta; y_i)s(\theta; y_i)' \end{aligned}$$

Upon taking expectations the first term on the right disappears since

$$\mathrm{E}\left[\frac{\partial^2 f(y_i; \theta)/\partial\theta\partial\theta'}{f(y_i; \theta)}\right] = \int \frac{\partial^2 f(y_i; \theta)}{\partial\theta\partial\theta'} \mathrm{d}y_i = 0$$

Therefore,

$$\mathrm{E}[H(\theta; y_i)] = -\mathrm{E}[s(\theta; y_i)s(\theta; y_i)'] = -\mathrm{Var}[s(\theta; y_i)]$$

as stated. This is the information matrix equality. Its extension to the full sample score and Hessian functions is straightforward due to the additivity of the log-likelihood function.

Example: Sampling from a Bernoulli Distribution

Assume as before that a random sample of size 5 is drawn from a Bernoulli distribution with parameter π. The score function is given by

$$s(p; y_1, \ldots, y_5) = \sum_{i=1}^{5} \frac{y_i - p}{p(1-p)}$$

The variance of the score is then

$$\mathrm{Var}[s(p; y_1, \ldots, y_5)] = \sum_{i=1}^{5} \frac{\mathrm{Var}(y_i - p)}{p^2(1-p)^2} = \frac{5\pi(1-\pi)}{p^2(1-p)^2}$$

Evaluated at the true parameter value (for $p = \pi$) this expression simplifies to

$$\mathrm{Var}[s(\pi; y_1, \ldots, y_5)] = \frac{5}{\pi(1-\pi)}$$

The Hessian matrix (here a scalar) is

$$H(p; y_1, \ldots, y_5) = \sum_{i=1}^{5} -\frac{1-y_i}{(1-p)^2} - \frac{y_i}{p^2}$$

with expectation

$$\mathrm{E}[H(p; y_1, \ldots, y_5)] = \sum_{i=1}^{5} -\frac{1-\pi}{(1-p)^2} - \frac{\pi}{p^2}$$

Evaluating the expected Hessian matrix at the true parameter value, we obtain

$$\mathrm{E}[H(\pi; y_1, \ldots, y_5)] = -\frac{5}{\pi(1-\pi)}$$

3.1.3 Conditional Models

All results mentioned so far require only minor modifications if conditional rather than marginal probability models are considered. Recall the examples for conditional probability models from chapter xxx.

- $y_i|x_i$ is Bernoulli distributed with parameter $\pi_i = \exp(x_i'\beta)/(1 + \exp(x_i'\beta)$.
- $y_i|x_i$ is Poisson distributed with parameter $\lambda_i = \exp(x_i'\beta)$
- $y_i|x_i$ is normal distributed with parameters $\mu_i = x_i'\beta$ and σ^2.

In order to accommodate such models within the previous framework, all we need to do is to replace the marginal probability or density function $f(y_i; \theta)$ by the conditional probability or density function $f(y_i|x_i; \theta)$ implied by the model. θ now comprises the β's plus any other parameters of the model, σ^2 in the case of the normal linear model. If β is a $(k \times 1)$ vector, then the score function has dimension $(k \times 1)$ as well, if there are no other parameters, and the Hessian matrix has dimension $(k \times k)$. Finally, we need to assume that pairs of observations (y_i, x_i) are independent for $i \neq j$.

3.1.4 Maximization

The value of θ that maximizes $L(\theta; y)$ is called the *maximum likelihood estimator* (MLE). We will use the symbol $\hat{\theta}$, or, if the type of estimation procedure used is not implicitly clear, $\hat{\theta}_{ML}$. Since the logarithm is a monotonic transformation, any value $\hat{\theta}$ that maximizes $L(\theta; y)$ also maximizes $\log L(\theta; y)$. There are two reasons why we focus on the maximization of the log-likelihood function in the following. First, maximization is much simpler if one takes logarithms, since this converts products to sums and the derivative needed for establishing the first-order condition for a maximum is a linear operator. Second, reliance on the log-likelihood function allows us to use results for the score vector and Hessian matrix developed above in order to establish properties of the maximum likelihood estimator including its asymptotic distribution.

As a starting point consider the problem of determining $\hat{\theta}$ such that

$$\arg\max_{\theta \in \Theta} \log L(\theta; y)$$

As usual, a necessary condition for a maximum is that the first derivative of the log-likelihood function, that is, the score vector, is equal to zero

$$\left[\frac{\partial \log L(\theta; y)}{\partial \theta}\right]_{\theta=\hat{\theta}} = s(\hat{\theta}; y) = 0$$

For a necessary and sufficient condition, we require in addition that the Hessian matrix

$$\left[\frac{\partial^2 \log L(\theta; y)}{\partial \theta \, \partial \theta'}\right]_{\theta=\hat{\theta}} = H(\hat{\theta})$$

is negative definite provided there is a solution at an inner point of the parameter space Θ. This maximum could be local or global. In "well-behaved" cases – and all problems considered in this book are well-behaved in this sense – the log-likelihood function is globally concave, from which it follows that the solution to the first-order condition gives the unique and global maximum of the log-likelihood function.

Example: Sampling from a Bernoulli Distribution

Assume as before that a random sample of size 5 is drawn from a Bernoulli distribution. The score and Hessian were derived before. The score is

$$s(p; y_1, \ldots, y_5) = \sum_{i=1}^{5} \frac{y_i - p}{p(1-p)}$$

Solving the first-order condition

$$\sum_{i=1}^{5} \frac{y_i - p}{p(1-p)} = 0$$

we find that

$$\hat{p} = \frac{1}{5} \sum_{i=1}^{5} y_i = \bar{y}$$

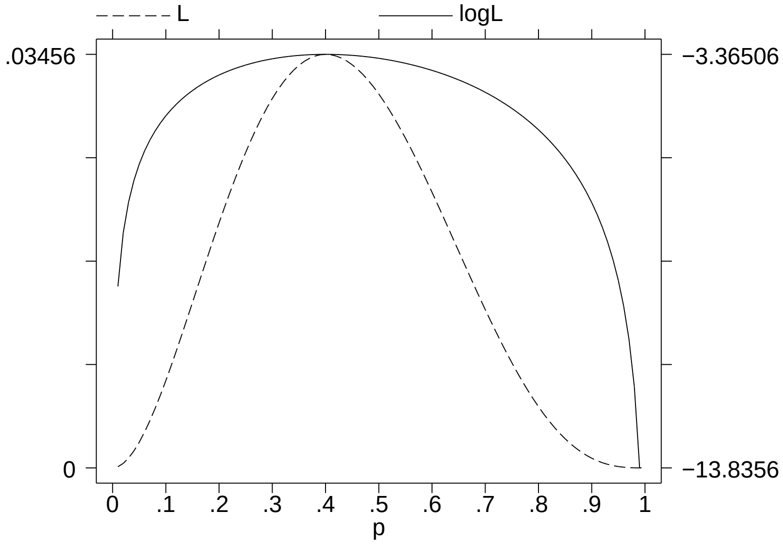
In the Bernoulli case, the maximum likelihood estimator for the probability of a one is equal to the sample mean, i.e., the proportion of ones in the sample. The second derivative of the log-likelihood function is equal to

$$H(p; y_1, \dots, y_5) = \sum_{i=1}^{5} -\frac{y_i}{p^2} - \frac{1-y_i}{(1-p)^2}$$

This term is negative for all possible samples $(y_1, \dots, y_5)$ as long as there is variation in y. Only then will there be an inner solution with well-defined score function and Hessian. Otherwise, if all observed values are either one or zero, the maximum likelihood estimator for p is at the boundary of the parameter space, either one or zero, respectively. We say that the model predicts the outcome perfectly. In the conditional Bernoulli model, perfect prediction causes a breakdown of the maximum likelihood method. For instance, if $p_i = \exp(x_i'\theta)/(1 + \exp(x_i'\theta)$, p_i cannot become zero or one for finite values of θ.

Now assume that a particular sample has been realized, for example $(0, 0, 0, 1, 1)$. Accordingly, the realization of the maximum likelihood estimator $\hat{\theta}$ will be the estimate $\hat{\theta} = 0.4$. This situation is depicted in Figure 3.3, where the likelihood and log-likelihood functions for this particular sample are plotted. The maximum is reached at $p = 0.4$. The corresponding value of the likelihood function is 0.035, and for the log-likelihood function it is -3.37.

Figure 3.3: Likelihood and Log-Likelihood for the Bernoulli Sample



3.1.5 Properties of the Maximum Likelihood Estimator

The maximum likelihood estimator $\hat{\theta} = \text{argmax } \log L(\theta; y)$ is in general a non-linear function of the dependent variable. Therefore, analytical small sample results are unavailable. Provided a number of regularity conditions are satisfied, it can be shown that the maximum likelihood estimator is:

- consistent
- asymptotically normal

- asymptotically efficient

The most important assumption is that the model is correctly specified, with true density given by $f(y_i;\theta)$. For further regularity conditions, see, for instance, Cramer (1986).

The approximate distribution of $\hat{\theta}$ is given by

$$\hat{\theta} \overset{\text{app}}{\sim} Normal(\theta, [-\mathrm{E}H(\theta)]^{-1}) \tag{3.6}$$

where $-\mathrm{E}H(\theta;y) = I(\theta)$ is the Fisher information matrix of the sample.

The main steps of a proof are as follows. Consider a first-order Taylor series approximation of $s(\hat{\theta})$ around the true parameter vector θ:

$$s(\hat{\theta};y) \approx s(\theta;y) + H(\theta;y)(\hat{\theta}-\theta)$$

Since $s(\hat{\theta};y)$ is zero by definition of a maximum likelihood estimator, we have

$$\hat{\theta}-\theta \approx -H(\theta;y)^{-1}s(\theta;y)$$

or

$$\sqrt{n}(\hat{\theta}-\theta) \approx \left(-\frac{1}{n}H(\theta;y)\right)^{-1}\frac{1}{\sqrt{n}}s(\theta;y)$$

Score function and Hessian are sums of independent components and law of large numbers and central limit theorems can be invoked. For increasing n, both $s(\theta;y)$ and $H(\theta;y)$ converge to their first moments $\mathrm{E}[s(\theta;y)] = 0$ and $\mathrm{E}[H(\theta;y)] = -I(\theta)$. The first convergence, that of $s(\theta;y)$ to $\mathrm{E}[s(\theta;y)]$ ensures consistency of the maximum likelihood estimator. In fact, this convergence allows for a re-interpretation of the maximum likelihood estimator as a method-of-moments estimator: the estimator is the value that solves the sample equivalent to the population moment restriction $\mathrm{E}[s(\theta;y)] = 0$.

The asymptotic distribution follows from a central limit theorem, whereby

$$\frac{1}{\sqrt{n}}s(\theta;y) \xrightarrow{d} Normal(0, n^{-1}I_n) \tag{3.7}$$

(where $\xrightarrow{d}$ stands for "converges in distribution") and therefore

$$\sqrt{n}(\hat{\theta}-\theta) \approx \left(-\frac{1}{n}H(\theta;y)\right)^{-1}\frac{1}{\sqrt{n}}s(\theta;y) \xrightarrow{d} Normal(0, nI(\theta)^{-1})$$

or

$$\hat{\theta} \overset{\text{app}}{\sim} Normal(\theta, I(\theta)^{-1})$$

(where $\overset{\text{app}}{\sim}$ stands for "approximately distributed as"). In practice, θ and thus $I(\theta)$ are not known. For the purpose of inference, the true variance covariance matrix of $\hat{\theta}$ can be replaced by a consistent estimator.

To summarize, the maximum likelihood estimator has the following properties. The asymptotic distribution is centered at the true parameter θ and its variance goes to zero. Hence we have mean squared convergence to zero and thus consistency. As mentioned earlier, $I(\theta)^{-1}$ is the smallest variance for any consistent estimator (linear or not). It is the so-called *Cramer-Rao lower bound*. Hence, the maximum likelihood estimator is asymptotically efficient. Also, the asymptotic distribution is normal which generates simple (asymptotic) procedures for inference. Of course, these properties in general only hold if the model is *correctly specified*. Both the population distribution must be correct, and the assumption of random sampling must be valid.

3.1.6 Estimating the Covariance Matrix

θ, $I(\theta)$ and thus $\text{Var}(\theta) = I(\theta)^{-1}$ are not known. For purposes of inference, the true covariance matrix of $\hat{\theta}$ is replaced by a consistent estimator. An obvious candidate is minus the expected Hessian evaluated at the parameter estimate (instead of the true value θ):

$$\widehat{\text{Var}}(\hat{\theta})_1 = [-\text{E}H(\hat{\theta}; y)]^{-1}$$

It is frequently the case that the Hessian matrix is a highly nonlinear function of y, making it impossible to obtain an exact expression for the expected value. A second alternative estimator for the covariance matrix takes the inverse of minus the *actual* Hessian matrix evaluated at the maximum likelihood estimator

$$\widehat{\text{Var}}(\hat{\theta})_2 = [-H(\hat{\theta}; y)]^{-1}$$

This is the standard procedure incorporated in most software routines for microeconometric models. Sometimes, even the computation of the Hessian is complicated. Then, it follows from the information matrix equality (see chapter xxx) that another estimator of the covariance matrix is the *outer product of the score*

$$\widehat{\text{Var}}(\hat{\theta})_3 = \left[\sum_{i=1}^{n} s(\hat{\theta}; y_i) s(\hat{\theta}; y_i)'\right]^{-1}$$

The practical relevance of this result is that the three estimators are asymptotically equivalent, and hence one can use whatever is most convenient.

It is also possible to combine these estimators. Indeed, it can be shown that the estimator

$$\widehat{\text{Var}}(\hat{\theta})_4 = \hat{I}_2^{-1} \hat{I}_3 \hat{I}_2^{-1}$$

has desirable properties. In particular, it is a consistent estimator of the covariance matrix of $\hat{\theta}$ even if the model is misspecified (the distributional assumption is violated). In this case, $\hat{\theta}$ may or may not be a consistent estimator of θ. If it remains consistent even if some aspects of the model are misspecified, one also talks about *pseudo-maximum likelihood estimation.*

Example: Sampling from a Bernoulli Distribution

The score and Hessian are

$$s(p; y_1, \ldots, y_n) = \sum_{i=1}^{n} \frac{y_i - p}{p(1-p)}$$

and

$$H(p; y_1, \ldots, y_n) = \sum_{i=1}^{n} -\frac{y_i}{p^2} - \frac{1-y_i}{(1-p)^2}$$

respectively. Therefore

$$\widehat{\mathrm{Var}}(\hat{p})_1 = \frac{\hat{p}(1-\hat{p})}{n}$$

$$\widehat{\mathrm{Var}}(\hat{p})_2 = \left(\sum_{i=1}^{n} \frac{y_i}{\hat{p}^2} + \frac{1-y_i}{(1-\hat{p})^2}\right)^{-1} = \left(\frac{n\bar{y}}{\hat{p}^2} + \frac{(n-n\bar{y})}{(1-\hat{p})^2}\right)^{-1} = \frac{\hat{p}(1-\hat{p})}{n}$$

$$\widehat{\mathrm{Var}}(\hat{p})_3 = \left(\sum_{i=1}^{n} \frac{(y_i-\hat{p})^2}{\hat{p}^2(1-\hat{p})^2}\right)^{-1} = \left(\frac{(n\bar{y} - 2n\hat{p}\bar{y} + n\hat{p}^2}{\hat{p}^2(1-\hat{p})^2}\right)^{-1} = \frac{\hat{p}(1-\hat{p})}{n}$$

since $\hat{p} = \bar{y}$. In this simple case, all three estimators are identical.

3.2 Applications

In this section, we illustrate the maximum likelihood methodology with the help of two further examples. The first example is an unconditional model, the maximum likelihood estimation of the Poisson parameter λ in a model without further covariates. The second example is a conditional model, the normal linear regression model.

3.2.1 Poisson model without covariates

We assume a random sample of size n from a population that has the following Poisson probability function:

$$f(y_i;\lambda) = \frac{e^{-\lambda}\lambda^{y_i}}{y_i!}$$

where $y = 0, 1, 2, \ldots$ can take on any non-negative integer values and λ must be non-negative. One property of the Poisson distribution is that its mean and variance are equal, because $\mathrm{E}(y_i) = \mathrm{Var}(y_i) = \lambda$.

To find the maximum likelihood estimator for λ, we start out with the log-likelihood function

$$\log L = \sum_{i=1}^{n} \log f(y_i;\lambda) = \sum_{i=1}^{n} (-\lambda + y_i \log\lambda - \log y_i!)$$

Notice that in this case, the last terms $\log y_i!$ do not depend on λ and are thus constants with respect to the maximization and can be dropped from the analysis. The score function is simply

$$s(\lambda;y) = \sum_{i=1}^{n} \left(-1 + \frac{y_i}{\lambda}\right) = -n + \frac{n\bar{y}}{\lambda}$$

from which it follows that the estimator $\hat{\lambda}$ that solves the first order condition $s(\hat{\lambda};y) = 0$ is the sample mean, $\hat{\lambda}_{ml} = \bar{y}$. In order to verify that this candidate estimator indeed maximizes the log-likelihood function, we need to verify the second order condition that the Hessian must be negative. Since

$$H(\lambda) = \sum_{i=1}^{n} -\frac{y_i}{\lambda^2}$$

for all interior solutions this is the case indeed. Moreover, we immediately obtain the asymptotic distribution of $\hat{\lambda}$ since

$$\hat{\lambda} \overset{\text{app}}{\sim} Normal(\lambda, \lambda/n)$$

Of course, in small samples, $\hat{\lambda}$, the mean of a variable with discrete support, is not normally distributed. Indeed, it has not even a continuous distribution. But the approximation holds for 'sufficiently large' n.

The asymptotic variance of the maximum likelihood estimated in the Poisson example can be estimated with any of the three methods mentioned in Section 3.1.6.

Method I.

$$\widehat{\text{Var}}(\hat{\lambda}) = \left[-\text{E}\left[\sum_{i=1}^{n} -\frac{y_i}{\lambda^2} \right]_{\hat{\lambda}} \right]^{-1} = \frac{\hat{\lambda}}{n}$$

Method II.

$$\widehat{\text{Var}}(\hat{\lambda}) = -\left[\sum_{i=1}^{n} -\frac{y_i}{\lambda^2} \right]_{\hat{\lambda}}^{-1} = \left[\frac{1}{\lambda^2} \sum_{i=1}^{n} y_i \right]_{\hat{\lambda}}^{-1} = \frac{\hat{\lambda}}{n}$$

Method III.

$$\widehat{\text{Var}}(\hat{\lambda}) = \left[\sum_{i=1}^{n} \left(\frac{y_i}{\lambda} - 1 \right)^2 \right]_{\hat{\lambda}}^{-1} = \left[\frac{\sum_{i=1}^{n} (y_i - \hat{\lambda})^2}{\hat{\lambda}^2} \right]^{-1}$$

If the model is correct, $\text{Var}(y) = \lambda$, and the numerator in III is a consistent estimator of $n\text{Var}(y) = n\lambda$. Therefore, as in the Bernoulli example derived earlier, all three methods to compute the covariance matrix give the same result.

3.2.2 The Normal Linear Model

The classical linear regression model is usually written as

$$y_i = \beta_1 + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + u_i = x_i'\beta + u_i$$

where $x_i = (1, x_{2i}, \ldots, x_{ki})'$ is a $(k \times 1)$ vector of explanatory variables and β is a conformable vector of parameters. Under the assuption that $\text{E}(u_i x_{ji}) = 0$ for all $j = 2 \ldots, k$ and that $\text{Var}(u_i) = \sigma^2$, the regression parameters of the model can be estimated by ordinary least squares and the resulting estimator is the best linear unbiased estimator. Now, we will in addition assume that u_i is normal distributed with mean 0 and variance σ^2. The resulting *normal linear model* can be written in the format of a conditional probability model because it follows from the above assumptions that

$$y_i | x_i \sim Normal(x_i'\beta, \sigma^2)$$

We will now show how the parameters of this model, β and σ^2, can be estimated by the maximum likelihood method. The density function for each observation can then be written explicitly as

$$f(y_i | x_i; \beta, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2} \left(\frac{y_i - x_i'\beta}{\sigma} \right)^2 \right]$$

Assuming a random sample of size n of pairs of observations (y_i, x_i), the likelihood function is

$$L(\beta, \sigma^2; y, x) = \prod_{i=1}^{n} f(y_i|x_i; \beta, \sigma^2)$$

and the log-likelihood function is

$$\begin{aligned}\log L(\beta, \sigma^2; y, x) &= \sum_{i=1}^{n} \log f(y_i|x_i; \beta, \sigma^2) \\ &= \sum_{i=1}^{n} \left[-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\sigma^2 - \frac{1}{2}\left(\frac{y_i - x_i'\beta}{\sigma}\right)^2 \right] \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i'\beta)^2\end{aligned}$$

The first-order conditions for maximizing this log-likelihood are:

$$\frac{\partial \log L}{\partial \beta} = \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i(y_i - x_i'\beta) \overset{!}{=} 0 \tag{3.8}$$

$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - x_i'\beta)^2 \overset{!}{=} 0 \tag{3.9}$$

The MLE for β, $\hat{\beta}$, can be determined from the first equation which can be rewritten as:

$$\sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} x_i x_i' \hat{\beta}$$

such that

$$\hat{\beta}_{ML} = \left(\sum_{i=1}^{n} x_i x_i'\right)^{-1} \left(\sum_{i=1}^{n} x_i y_i\right) \tag{3.10}$$

Here we solved the system of linear equations by using the concept of an inverse matrix: A^{-1} is the inverse of A if $A^{-1}A = I$ where I is a diagonal matrix with ones on the main diagonal and zeros elsewhere. See also chapter xxx for further details on this notation.

An inspection of the maximum likelihood estimator $\hat{\beta}_{ML}$ in (3.10) shows that it is the same as the ordinary least squares estimator. Hence, under the assumptions of the normal linear model, and as far as the slope vector β is concerned, there is no difference at all between maximum likelihood estimation and least squares. The reason is that the score vector is essentially equal to the normal equations of the least squares problem.

The second parameter of the model, σ^2, can be estimated by maximum likelihood as well. In order to do so, replace β in (3.9) by its maximum likelihood estimator $\hat{\beta}$, and define residuals $\hat{e}_i = y_i - x_i'\hat{\beta}$. Then

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{e}_i^2$$

This differs from the familiar variance estimator since the divisor is n and not $n-k$. In other words, the maximum likelihood estimator makes no adjustment for the degrees of freedom. Hence it is biased in small samples. However, since maximum likelihood estimation is anyway an approach relying on asymptotic properties, and since the bias disappears quickly for large n, this is not really something to worry about.

So far, we have taken for granted that the maximum likelihood estimators obtained from the first-order conditions are indeed maximizing the log-likelihood. Of course, we also should check the second order conditions to see whether this is the case in the normal linear model. We need to evaluate the three second derivatives

$$\frac{\partial^2 \log L}{\partial\beta\partial\beta'} = -\frac{1}{\sigma^2}\sum_{i=1}^n x_i x_i'$$

$$\frac{\partial^2 \log L}{\partial(\sigma^2)^2} = \frac{n}{2\sigma^4} + \frac{1}{\sigma^6}\sum_{i=1}^n (y_i - x_i'\beta)^2$$

$$\frac{\partial^2 \log L}{\partial\beta\partial\sigma^2} = -\frac{1}{\sigma^4}\sum_{i=1}^n x_i(y_i - x_i'\beta)$$

Because of symmetry, $\partial^2 \log L/\partial\beta\partial\sigma^2 = \partial^2 \log L/\partial\sigma^2\partial\beta$. Collecting terms, we get the Hessian matrix

$$H(\beta,\sigma^2;y,x) = \left[\begin{array}{cc} -\frac{1}{\sigma^2}\sum_{i=1}^n x_i x_i' & -\frac{1}{\sigma^4}\sum_{i=1}^n x_i(y_i - x_i'\beta) \\ -\frac{1}{\sigma^4}\sum_{i=1}^n x_i(y_i - x_i'\beta) & \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^n (y_i - x_i'\beta)^2 \end{array}\right]$$

The dimension of this matrix is $(k+1)\times(k+1)$. It can be shown that this matrix is a negative definite matrix, provided the x_i's are well-behaved and not collinear, and that therefore, the log-likelihood function of the normal linear model is globally concave, and the $\hat{\beta}$ and $\hat{\sigma}^2$ are indeed the maximizing values.

The information matrix contains the negatives of the expected values of the Hessian matrix:

$$I(\beta,\sigma^2) = \left[\begin{array}{cc} \frac{1}{\sigma^2}\sum_{i=1}^n x_i x_i' & 0 \\ 0 & n/2\sigma^4 \end{array}\right]$$

Its inverse provides the asymptotic covariance matrix of the maximum likelihood estimator in the normal linear regression model.

$$I(\beta,\sigma^2)^{-1} = \left[\begin{array}{cc} \sigma^2(\sum_{i=1}^n x_i x_i')^{-1} & 0 \\ 0 & 2\sigma^4/n \end{array}\right]$$

Finally, it follows that

$$\left(\begin{array}{c} \hat{\beta} \\ \hat{\sigma}^2 \end{array}\right) \overset{\text{app}}{\sim} N\left[\left(\begin{array}{c} \beta \\ \sigma^2 \end{array}\right), \left(\begin{array}{ll} \sigma^2(\sum_{i=1}^n x_i x_i')^{-1} & 0 \\ 0 & 2\sigma^4/n \end{array}\right)\right]$$

where, as before, the covariance matrix can be evaluated at the maximum likelihood estimators $\hat{\beta}$ and $\hat{\sigma}^2$.

3.3 Further Aspects of Maximum Likelihood Estimation

In this chapter, we will pick up some selected further issues. We start with a discussion of the invariance property of maximum likelihood estimation. Second, we point out that the examples given so far, Bernoulli and Poisson without regressors or the normal linear model, are somewhat untypical, in that in each case, it was possible to obtain the maximum likelihood estimator as a closed form solution to the first-order condition. In most micro data models, however, such a closed form solution is not available due to non-linearities in the score function. Therefore, estimation has to rely on numerical optimization routines. There are a number of such algorithms available but in this chapter we will focus on the Newton-Raphson method. Finally, there are situations when a maximum likelihood estimator does not exist. This can be due to data deficiencies or to modelling deficiencies, in particular a lack of identification.

3.3.1 Invariance

Invariance is a useful property of maximum likelihood estimation. Let $\hat{\theta}_{ml}$ denote the maximum likelihood estimator of θ. Assume that instead of θ, you are interested in $g(\theta)$, where g is an arbitrary function. What the invariance property says is that the maximum likelihood estimator of $g(\theta)$ will be simply $g(\hat{\theta}_{ml})$. The intuition for this result is simply that maximum likelihood yields, under the assumptions of the model, a consistent estimator and that, for large enough samples, we can treat the estimator as if it were a constant.

Example 1:
Let $\vartheta = \ln\theta$ with inverse function $\theta = \exp(\vartheta)$. Assume one wants to impose the constraint that θ is nonnegative. This is achieved by parameterizing the model in terms of ϑ (since $\exp(\vartheta)$ is nonnegative for all $\vartheta \in \mathbb{R}$) and obtain the MLE for ϑ. The MLE for θ is then $\hat{\theta}_{ml} = \exp(\hat{\vartheta}_{ml})$.

Example 2:
Let $\hat{\theta}$ be the maximum likelihood estimator for the mean μ of a homoskedastic normal population and a sample of size n. We know that $\mathrm{E}(\hat{\theta}) = \mu$ and $\mathrm{E}[\exp(\hat{\theta})] = \exp(\mu + 0.5\sigma^2/n)$. Hence, $\exp(\hat{\theta})$ is a biased but consistent estimator of $\exp(\mu)$. The invariance property holds approximately for large enough samples.

To obtain asymptotic standard errors for transformed estimators, we can use the *Delta method.* For a scalar parameter θ,

$$\widehat{\mathrm{Var}}[g(\hat{\theta}_{ml})] = g'(\hat{\theta}_{ml})^2\widehat{\mathrm{Var}}[\hat{\theta}_{ml}]$$

For example, assume that $g(\hat{\theta}_{ml}) = \exp(\hat{\theta}_{ml})$. It follows that $\widehat{\mathrm{Var}}[g(\hat{\theta}_{ml})] = \exp(2\hat{\theta}_{ml})\widehat{\mathrm{Var}}[\hat{\theta}_{ml}]$.

If θ is a vector, the delta method requires computation of a quadratic form involving the vector of derivatives of the g-function with respect to the elements of θ and the covariance matrix of θ.

3.3.2 Optimization Routines

In this chapter we will introduce some numerical methods for maximizing the log-likelihood function when the first-order condition has no closed form solution. Consider the following example.

Example: Normal distribution with log-linear conditional expectation function

In this example for a non-linear regression model, the conditional model of y_i given x_i is a normal distribution with conditional expectation $\mu(x_i) = \exp(x_i'\beta)$ and constant variance σ^2. The conditional density function can therefore be written as

$$f(y_i|x_i;\beta,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{y_i-\exp(x_i'\beta)}{\sigma}\right)^2\right]$$

Given an independent sample of n observations on y_i and x_i, the log-likelihood of this model is

$$\log L(\beta,\sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\exp(x_i'\beta))^2$$

and the score vector is

$$s(\beta) = \frac{\partial \log L}{\partial \beta} = \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i(y_i-\exp(x_i'\beta))\exp(x_i'\beta)$$

The first-order condition for a maximum of the log-likelihood function with respect to β, $s(\beta)=0$, has no analytical solution.

A frequently used method of quadratic approximation is the *Newton-Raphson method.* It can be motivated as follows. Given any initial parameter estimate, say $\hat\theta^0$, we can obtain a second-order approximation of $\log L(\theta)$ around $\hat\theta^0$:

$$\log L^*(\theta) = \log L(\hat\theta^0) + s(\hat\theta^0)'(\theta-\hat\theta^0) + \frac{1}{2}(\theta-\hat\theta^0)'H(\hat\theta^0)(\theta-\hat\theta^0) \approx \log L(\theta)$$

Now, we can maximize $\log L^*(\theta)$ (rather than $\log L(\theta)$) with respect to θ, yielding a new parameter value which we call $\hat\theta^1$. The first order condition of this simpler problem is

$$s(\hat\theta^0) + H(\hat\theta^0)(\hat\theta^1-\hat\theta^0) = 0$$

or

$$\hat\theta^1 = \hat\theta^0 - [H(\hat\theta^0)]^{-1}s(\hat\theta^0)$$

Thus, for arbitrary starting value $\hat\theta^0$, the Newton-Raphson updating rule is given by

$$\hat\theta^{t+1} = \hat\theta^t - [H(\hat\theta^t)]^{-1}s(\hat\theta^t) \qquad t=0,1,\ldots \tag{3.11}$$

where $s(\cdot)$ denotes the score and $H(\cdot)$ the Hessian of the log-likelihood function.

The iterative procedure ends when a predefined convergence criterion is satisfied. Possible criteria are the change in the value of the estimate $\hat\theta^{t+1}-\hat\theta^t$, the change in the log-likelihood $\log L(\hat\theta^{t+1})-\log L(\hat\theta^t)$, or the value of the gradient at the estimate $s(\hat\theta^t)$. Convergence occurs when any of these values, or a combination of them, are close to zero (say, smaller than 10^{-5} in absolute value).

Example: Normal distribution with log-linear conditional expectation function

In order to implement the Newton-Raphson algorithm, we need to compute the score function and the Hessian matrix and evaluate them iteratively at the current parameter values:

$$s(\beta_j) = \left.\frac{\partial \log L}{\partial \beta}\right|_{\beta=\beta_j} = \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i(y_i - \exp(x_i'\beta_j))\exp(x_i'\beta_j)$$

$$H(\beta_j) = \left.\frac{\partial^2 \log L}{\partial \beta\,\partial \beta'}\right|_{\beta=\beta_j} = \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i x_i'(y_i - 2\exp(x_i'\beta_j))\exp(x_i'\beta_j)$$

A numerical approximation to the solution can be found by selecting starting values β_0, for example from a linear regression of $\log y$ on x, and then updating according to the Newton-Raphson formula, using the update $\hat{\beta}_{j+1} = \hat{\beta}_j - H(\beta_j)^{-1}s(\beta_j)$, until convergence is reached.

3.3.3 When Maximum Likelihood Estimation Fails

There are two distinct type of problems that may cause maximum likelihood estimation to fail. The first one is a deficiency of the sample at hand, such as no variation in the dependent variable or no variation in an explanatory variable, or a combination of the two. The most common incarnation of such a combination arises when binary dependent variables and binary explanatory variables are mixed and there is no variation in the dependent variable for a given value of the binary explanatory variable. This leads to a perfect prediction problem, and an associated unboundedness of the log-likelihood function. See also chapter 3.1.4. To remedy such a problem, one can collect more data and hope that some useful variation arises in the additional observations.

A fundamentally different problem is that of identification, or the lack thereof. Identification is the study of what conclusions can and cannot be drawn in infinite samples, given specified combinations of assumptions and types of data.

> "Identification problems cannot be solved by gathering more of the same kind of data. These inferential difficulties can be alleviated only by invoking stronger assumptions or by initiating new sampling processes that yield different kinds of data." (Manski, 1995, p.3/4)

Identification is a general issue in any econometric model and for any estimation procedure. It certainly is not restricted to maximum likelihood estimation. Some examples follow:

- Sampling of y_i conditional on $x_i = 1, 2$. This can identify $\mathrm{E}(y_i|1)$ and $\mathrm{E}(y_i|2)$ but not, without further assumption, $\mathrm{E}(y_i|1.5)$. Identification could be achieved in two ways:
  - Change the sampling scheme and sample conditional on $x_i = 1.5$ as well.
  - Add the assumption that the conditional expectation function has a particular functional form, such as

    $$\mathrm{E}(y_i|x_i) = \beta_1 + \beta_2 x_i$$

- In other cases, identification is achieved by imposing an exclusion restriction. Here is an example:

  $$\mathrm{E}(y_i|x_i) = \beta_1 + \beta_2\ \mathrm{Male}_i + \beta_3\ \mathrm{Female}_i$$

In this conditional expectation function all parameters $\beta_{1,2,3}$ are not identified at the same time. One can either exclude the constant, or one of the dummy variables. Exclusion restrictions are also key when estimating the structural parameters from a reduced form model in the context of simultaneous equations, but this is not further discussed here.

- In a probability framework, two parameters θ_1 and θ_2 are said to be *observationally equivalent* if $f(y, \theta_1) = f(y, \theta_2)$ for all y. A parameter point θ_0 is then identifiable if there is no other $\theta \in \Theta$ which is observationally equivalent (Rothenberg, 1971).
- A sufficient condition for identification is that the information matrix $-\mathrm{E}[H(\theta)]$ is nonsingular.
- An example for a non-identified model, a mixture of two normal distributions, can be found in Maddala (1983, page xxx).

3.3.4 When Assumptions Are Violated

The single most important drawback of maximum likelihood estimation is that it requires the specification of a true probability model. If the specified model is incorrect, the resulting maximum likelihood estimator is inconsistent in general. This situation reflects a classical trade-off in empirical modelling: to be able to obtain strong results, including consistency, asymptotic normal distribution, and the ability to make inferences on conditional probability effects, one must also be prepared to make strong assumptions. The weaker the assumptions of a model, the more robust (to model misspecification) but also the weaker the inferences one can draw. A practical approach should not rely on a single estimate based on a single specification, but rather employ various models. To assess the usefulness of the evidence thus obtained, one can then ask whether the conclusions across the different specifications are qualitatively similar.

Whether or not a maximum likelihood estimator retains some of its desirable properties even when the underlying probability model is misspecified needs to be checked in each single case. The study of the behaviour of maximum likelihood estimators under misspecification is the topic of so called *quasi-likelihood estimation* (see, for instance, Gourieroux, Monfort and Trognon, 1984). Many frequently used maximum likelihood estimators remain consistent even if the underlying distributional assumption is violated. The leading example is the estimation of θ by maximum likelihood in the normal linear model (chapter 3.2.2). Since the maximum likelihood estimator is identical to the ordinary least squares estimator, it inherits all its desirable properties, that is, best linear unbiadness, that do not depend on the distribution $y_i|x_i$ at all. Of course, we still must require that the conditional expectation function is correct. In fact, all of the univariate examples given in this chapter, as well as the logit and Poisson regression models to be introduced in detail later, are robust to distributional misspecification in the sense that the (conditional) expectation can be estimated consistently if the chosen distribution is the wrong one.

3.4 Testing

Having specified a model and estimated its parameters, the next step is to conduct inferences, i.e., generalize from the estimates obtained for the sample to the true population parameters. In conditional probability models considered in this book, the main attention belongs to the regression parameters that capture the effect of explanatory variables on the conditional probabilities and conditional expectation function, although the models frequently include other parameters as well (such as σ^2 in the normal linear model).

The ultimate goal of any empirical analysis is to learn something about the values of these parameters in the population. The MLE gives us point estimates. In order to draw inferences on the population parameters, we need to account for the sampling variability. The standard way to do this is by formulating a hypothesis and assessing whether, from a statistical point of view, the evidence found in the data leads to a rejection of the hypothesis or not. Hypotheses for parameter values can also be interpreted as *restrictions* on the parameter space. In the next subsections, we will consider restrictions common in the analysis of micro data, and show how to impose them into the maximum likelihood estimation.

While we discuss the handling of restrictions with maximum likelihood applications in mind, the concepts are obviously much more general and can equally applied to least squares estimation (remember restricted least squares) or other estimation techniques. What is specific to maximum likelihood estimation, however, are two of the three methods that are available to test the validity of the restrictions, namely the likelihood ratio test and the score test, with the Wald test being a third possibility known from least squares analysis.

3.4.1 Restrictions

Assume that we have formulated a model

$$\log \text{wage} = \beta_1 + \beta_2 \,\text{age}_i + \beta_3 \,\text{age}_i^2 + \beta_4 \,\text{years of education}_i + e_i$$

$\beta_4 = 0$ if and only if education does not affect wages. This is a restriction and we can test wether it holds. If we reject it, then we say that education has a "significant effect" on wages, or simply that "education is statistically significant" (not to be confused with "economically significant", see McCloskey, xxx). An example for another restriction is $\beta_4 = 0.06$. Whether particular restrictions are interesting depends on the economic context.

Linear restrictions

All restrictions considered in this book, and almost all restrictions considered in practice, are linear. They come in three basic varieties.

1. The simplest conceivable restriction involves a single parameter:

   $$\beta_j = c$$

   From a practical point of view most important is the case where $c = 0$.

2. Sometimes, one is interested in linear combinations of two or more parameters. For example

   $$\beta_2 = \beta_3 \text{ or } 2\beta_3 = -\beta_4$$

   In the wage equation above, assume that we want to find out whether earnings peak is at the age of 55. Taking derivatives of the relevant polynomial part (as a necessary condition for a maximum), we find

   $$\beta_2 + 2\beta_3 \text{ age}_i = 0$$

   The restriction implies that $\beta_2 = -110\beta_3$.

3. A third case is that of multiple restrictions. For example, the restriction that none of the socio-economic variables matter in the wage equation can be expressed as

   $$\beta_2 = 0 \text{ and } \beta_3 = 0 \text{ and } \beta_4 = 0$$

   or $\beta_2 = \beta_3 = \beta_4 = 0$.

3.4.2 Restricted maximum likelihood

Restrictions can be imposed in the model. Estimating a model that guarantees the observance of restrictions (by maximum likelihood) is called *restricted maximum likelihood.*

Examples:

1. To enforce the restriction $\beta_4 = 0$, one simply estimates the wage equation without the education variable among the regressors.
2. Consider a Cobb-Douglas production function $Y_i = AL_i^\alpha K_i^\beta e^{u_i}$, or, after taking logarithms

   $$\ln Y_i = \ln A + \alpha \ln L_i + \beta \ln K_i + u_i$$

   (note: the Cobb-Douglas production function does not allow for zero inputs.) A production function is linear homogeneous if $f(\lambda K, \lambda L) = \lambda f(K, L)$. If all inputs are doubled (or multiplied by any factor λ), then output doubles as well (or is multiplied by the same factor λ). In the Cobb-Douglas case, linear homogeneity requires that $\alpha + \beta = 1$. To impose this restriction, rewrite the model as

   $$\ln Y_i = \ln A + \alpha \ln L_i + (1 - \alpha) \ln K_i + u_i$$

   i.e.

   $$\ln(Y_i/K_i) = \ln A + \alpha \ln(L_i/K_i) + u_i$$

In both examples a single restriction was imposed. That means that the number of parameters to be estimated is reduced by one, relative to the number of parameters of the unrestricted model.

3.4.3 Testing restrictions

All tests of restrictions involving maximum likelihood estimators are asymptotic by nature, i.e. the small sample distributions of the test statistics are unknown (although they can be, and have been on occasion, evaluated through Monte Carlo experimentation). The validity of the inference depends on the availability of a sample of sufficient size.

The tests distinguish between a null hypothesis

$$H_0: \text{ the restriction is true}$$

and an alternative hypothesis

$$H_1: \text{ the restriction is false}$$

Three tests are available in the context of maximum likelihood estimation:

- Wald test
- Likelihood ratio test
- Score test

The basic approach for all tests is to construct a test statistic, derive its distribution under H_0 (i.e., under the assumption that the restriction is correct), and then compare the observed value of the test statistic for the sample to the distribution under H_0: if what we see is unlikely to occur given this distribution (where "unlikely" could for example mean that values such as the observed test value or greater occur in less than 5 percent of all cases in repeated samples), H_0 is rejected. The tests are asymptotically equivalent but not so in finite samples. We will discuss them in their order.

3.4.4 Wald test

The basic variant of the Wald test, the so-called "t"- or z test is the most important test in empirical practice. Starting point of this test is the approximate distribution of the maximum likelihood estimator:

$$\hat{\theta} \overset{\text{app}}{\sim} Normal(\theta, \text{Var}(\theta)) \tag{3.12}$$

where $\text{Var}(\theta) = -\text{E}[H(\theta)]^{-1}$. Since we consider inference for large samples, it does not matter whether $\text{Var}(\theta)$ is known or we replace it by a consistent estimator $\widehat{\text{Var}(\theta)}$ (three such estimators were proposed before, the most common being minus the inverse of the Hessian of the log-likelihood function, evaluated at the maximum likelihood estimator).

In general, $\hat{\theta}$ is a vector of parameters. Denote the j-th element of $\hat{\theta}$ by $\hat{\theta}_j$. It follows from (3.12), that $\hat{\theta}_j$ is approximately univariate normal distributed with

$$\hat{\theta}_j \overset{\text{app}}{\sim} Normal(\theta_j, \hat{\sigma}_{jj})$$

where the estimated variance $\hat{\sigma}_{jj}$ is the j-th diagonal element of the covariance matrix. Of course, θ_j, and thus the distribution of $\hat{\theta}_j$, are unknown. However, we can impose a restriction. Under the restriction, the sampling distribution of $\hat{\theta}_j$ is known. We can then use the known sampling distribution to assess whether the restriction appears valid or invalid, given the evidence (the observed estimate).

For example, assume that $\theta_j = 0$. It follows under H_0 that $\hat{\theta}_j \sim Normal(0, \hat{\sigma}_{jj})$ and the z-statistic

$$\hat{z} = \frac{\hat{\theta}_j}{\sqrt{\hat{\sigma}_{jj}}} = \frac{\hat{\theta}_j}{s.e.(\hat{\theta}_j)}$$

is (approximately) standard normal distributed. On occasion, the z-statistic is referred to as t-statistic, although this is an abuse of language, since the t-distribution is a small sample distribution and all inference under maximum likelihood rests on asymptotic arguments. $s.e.$ is an often used acronym to denote the estimated standard error.

There are two ways to present the results from a hypothesis test, either through p-values or through z-values. The p-value of the standardized coefficient is defined by

$$P(|z| > |\hat{z}|)$$

If the p-value is smaller than a pre-determined level of significance (5% for example) the restriction (null hypothesis) is rejected. Alternatively, we can compare the z-score to the critical value of the standard normal distribution. For example, the critical value for a two-sided test at the 5% significance level is ± 1.96.

Such z scores are routinely reported in published research papers and also by statistical and econometric software packages. It is important to keep in mind that they assume a particular type of restriction, namely that the true parameter is zero. This restriction may not be relevant in some situations. For different restrictions, different z-scores need to be computed. Some researchers hence prefer to report standard errors rather than z scores and let the reader decide which restriction to be considered.

This methodology easily extends to the case of a single restriction involving more than one parameter. Consider the linear restriction $a\theta_1 + b\theta_2 = 0$ (which is the same as saying that $\theta_1 = -b/a\,\theta_2$). If the restriction is true, it follows that

$$a\hat{\theta_1} + b\hat{\theta_2} \sim Normal(0, a^2\sigma_{11} + 2ab\sigma_{12} + b^2\sigma_{22})$$

The z-score to test this restriction is therefore

$$\hat{z} = \frac{a\hat{\theta_1} + b\hat{\theta_2}}{\sqrt{a^2\hat{\sigma}_{11} + 2ab\hat{\sigma}_{12} + b^2\hat{\sigma}_{22}}}$$

where $\hat{\sigma}_{12}$ is the estimated covariance of the two parameters.

This is conceptionally straightforward. In practice, computation of the z- score requires knowledge of the covariance matrix of the MLE (or at least the relevant elements thereof). Not in all econometric software programs is it easy to display this matrix.

The Wald test is usually introduced in a more general form that allows for multiple restrictions (with a single restriction as special case). The test statistic is a quadratic form of a restriction vector and a covariance matrix that has a χ^2 distribution when the restriction is valid. We do not present this test here, because it is not used very frequently in practice, the reason being that alternative tests are easier to compute. For single restrictions, the general Wald test is equivalent to the simple z-tests and offers no extra benefit. For multiple restrictions, it is much easier to use the likelihood ratio test. This test is discussed next.

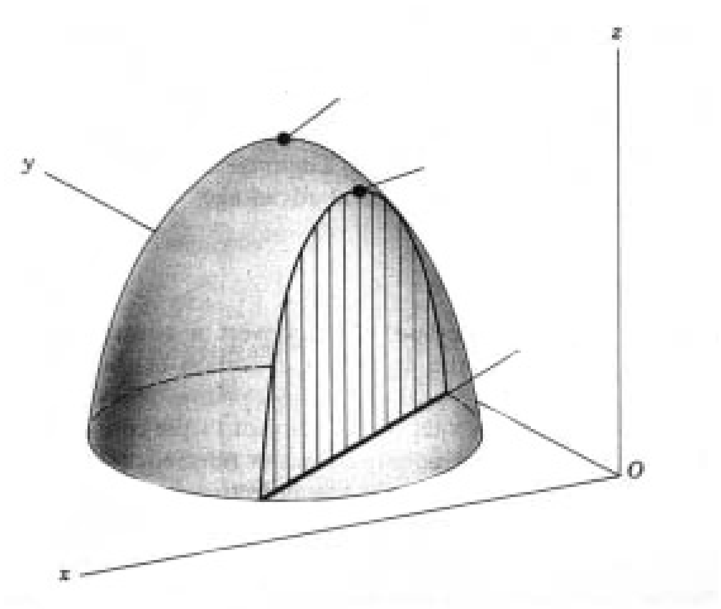
3.4.5 Likelihood ratio test

We have seen before that it is possible to impose the restriction and estimate the remaining parameters of the model by restricted maximum likelihood. Denote the value of the log-likelihood function at the restricted maximum by $\log L(\hat{\theta}_r)$ where the subscript 'r' reminds us that the restriction has been imposed. Denote the value of the log-likelihood function at the unrestricted maximum by $\log L(\hat{\theta}_u)$ where the subscript 'u' stands for 'unrestricted'. We know that $\log L(\hat{\theta}_r) \leq \log L(\hat{\theta}_u)$. Consult Figure 3.4 for an illustration that the restricted maximum cannot exceed the unrestricted maximum, and that it will fall short of it unless the unrestricted maximum happens to satisfy the restriction.

The drop in the log-likelihood associated with imposing restrictions is an indicator for the disagreement between the restrictions and the data. If the drop is sufficiently large, the restrictions are rejected. Let q denote the number of restrictions. The critical value for rejection comes from a $\chi^2(q)$ distribution since it can be shown that if the restrictions are true

$$LR = 2(\log L(\hat{\theta}_u) - \log L(\hat{\theta}_r)) \sim \chi^2_q$$

Figure 3.4: Unrestricted and Restricted Maximization



The test is referred to as likelihood *ratio* test, although it in fact involves the computation of a *difference* of *log*-likelihoods. The main advantage of this test is that it is easy to perform. It always works when a restricted and an unrestricted version of a model can be compared. The minor disadvantage is that two models need to be computed separately. This might have been a greater obstacle at times when computing was expensive but it should no longer be one nowadays.

Example: The Probability of Default on Student Loans

Knapp and Seaks (1992) study a sample of two thousand guaranteed student loans in order to determine the factors that increase or decrease the probability of default. They report that while individual characteristics (such as parent's income, presence of two parents at home, student's graduation and student's race) have a significant impact on default rates, institutional characteristics such a 2-year or 4-year college or private versus public college have little effect.

Their outcome variable is binary (default yes/no). As their conditional probability model, Knapp and Seaks use the probit model (see Chapter xxx) and estimate the parameters by maximum likelihood. All models include a constant. In one equation, they include as regressors the variables `Family Status`, `Graduation`, `Parent's income`, and `Race:Black`. The reported log-likelihood value is -436.00. In a second model, they add to the aforementioned variables `Gender:Female`, `Loan Amount`, and `Total College Cost`. The log-likelihood value increases now to -432.62.

To test whether the three additional variable are jointly significant, we consider the hypothesis H_0 : $\beta_{female} = \beta_{amount} = \beta_{cost} = 0$. All we need to do is to compute twice the log-likelihood difference between the two models. Therefore, the LR-test statistic is $2 \times (-432.62 - (-436.00)) = 6.76$. Under H_0, this statistic is a realization from a χ^2 distribution with 3 degrees of freedom. The 0.95 quantile for such a distribution is 7.82. Since the observed test statistic is greater than the critical value, we cannot reject H_0 at the 5-percent level of significance.

In order to test the significance of single parameters, it is easier to perform a z-test. For example, Knapp and Seaks report in the restricted model that the variable `Race:Black` has a coefficient of 0.902 with z-statistic 7.865. Since a value of 7.865 or beyond is highly unlikely to occur under $H_0 : \beta_{black} = 0$, in which case the z-statistik is approximately standard normal distributed, the null hypothesis of no effect of race is clearly rejected.

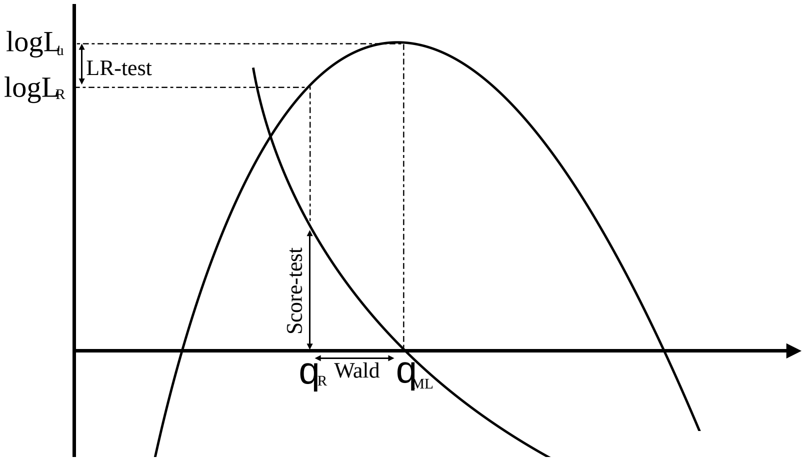
3.4.6 Score test

A third, asymptotically equivalent, test for a single restriction or a number of restrictions is the *score test*. Recall from Chapter 3.1.2 the property of the score, i.e., the first derivative of the log-likelihood function, that its expectation, evaluated at the true parameter, equals zero:

$$\mathrm{E}[s(\theta; y)] = 0$$

Moreover, it was shown that the score $s(\theta; y)$ is asymptotically normal distributed with mean zero and variance $\mathrm{Var}[s(\theta; y)] = -E[H(\theta; y)]$ (recall the information matrix equality in Chapter 3.1.2).

Figure 3.5: Wald, Likelihood ratio and Score Test



The crucial idea of the score test is that if the restriction was true, then it should be the case that the score, evaluated at the restricted parameter estimates $\hat{\theta}_r$, should not be "too" far away from its theoretical mean zero, where the distance is measured relative to its variance, by computing the quadratic form

$$S = s(\theta_r; y)' \mathrm{Var}[s(\theta_r; y)]^{-1} s(\theta_r; y)$$

Under H_0, S has a $\chi^2(q)$ distribution, where q is the number of restrictions.

The score test is also sometimes referred to as *Lagrange multiplier test* (Greene, 2000), although the presentation here makes the name "score test" seem more appropriate. In contrast to the two other test methods, the score test requires only estimation of the restricted model, in order to obtain the restricted parameter estimates $\hat{\theta}_r$. This may be an advantage. On the downside, the score test requires a considerable amount of algebra that is specific to the unrestricted model under consideration. Figure 3.5 compares the three test approaches.

3.4.7 Discriminating Between Nonnested Models

The Wald, likelihood-ratio and score tests compare two models among which one model is a restricted version of the other. The restrictions we mean here are those on the possible range of values that parameters might take, as detailed in Chapter 3.4.1. Consider instead a comparison of the two models

Model 1 : $y_i|x_{1i} \sim Normal(\beta_0 + \beta_1 x_{1i}, \sigma^2)$

Model 2 : $y_i|x_{2i} \sim Normal(\beta_0 + \beta_2 x_{2i}, \sigma^2)$

We call such two models nonnested, because none of them can be obtained as a restricted version of the other. If we still are forced to decide which of the two models, Model 1 or Model 2, we prefer, we have to choose different methods. We cannot simply test. A similar situation would arise for the following two models:

Model 1' : $y_i|x_{1i} \sim Normal(\beta_0 + \beta_1 x_{1i}, \sigma^2)$

Model 2' : $y_i|x_{1i} \sim Normal(\exp(\beta_0 + \beta_1 x_{1i}), \sigma^2)$

Again, the two models are nonnested. There is no parametric restriction that would transform (reduce) one model to the other. A simple rule to select between two nonnested model is to estimate both models by maximum likelihood and choose the one with the higher value of the log-likelihood. However, since one can increase the log-likelihood of any model arbitrarily by adding more and more parameters, the number of parameters should be taken into account as well when comparing the models. The model with the larger number of parameters should be penalized for this proliferation, and the two standard model silection criteria functions differ in their choice of penalty function:

- *Akaike information criterion*
  $AIC = -2 \log L(\hat{\theta}) + 2k$
- *Schwarz information criterion*
  $SIC = -2 \log L(\hat{\theta}) + k \log n$

Here, k is the number of parameters and n is the number of observations. Clearly, as in our two examples above, if the number of parameters is the same in the two models that are to be compared, the penalty function does not matter and the choice between the two models comes indeed simply down to choosing the model with the higher log-likelihood function.

3.4.8 Goodness-of-fit

Finally we consider the issue of goodness-of-fit in maximum likelihood estimation. In the standard linear models, the goodness of fit is usually assessed by the R^2, the proportion of the total variation in the dependent variable that is explained by the model.

In many non-linear micro data models, the underlying variance decomposition does not work and a standard R^2 measure is not available. One possible substitute is a log-likelihood comparison of the full model (with all regressors) and a constant-only model. Clearly, $\log L(\hat{\theta}_u) \geq \log L(\hat{\theta}_r)$. Since $\log L(\hat{\theta}_r) < 0$ it follows that

$$\frac{\log L(\hat{\theta}_u)}{\log L(\hat{\theta}_r)} \leq 1$$

Moreover define $R^2_{pseudo} = 1 - \log L(\hat{\theta}_u)/\log L(\hat{\theta}_r)$. Then

$$0 \leq R^2_{pseudo} \leq 1$$

This likelihood based pseudo R^2 was advocated by McFadden (1974).

3.4.9 Empirical strategies

When interpreting test results it is important to remember that tests of restrictions are asymmetric by construction. If a restriction is rejected then we know that the data provide strong evidence against it. Only in 5 (or 1, or 10) percent of all cases a restriction is rejected although it is correct.

If a restriction is not rejected, then we *cannot* say that the data provide strong evidence for the validity of the restriction. The *power* of the test, i.e. the probability of rejecting the restriction when it is false, can be quite low. But with low power the probability of not rejecting the restriction when it is false (the 'Type-II' error) is high.

These basic facts of hypothesis testing are sometimes ignored in practice. For example, Spencer (1985) modelled the relationship between money demand and the price level in the following way

$$m_t = \theta_0 + \theta_1 Y_t + \theta_2 R_t + \delta P_t + u_t$$

where m_t are real money balances and P_t is the log of the price level. According to some theories, the price level elasticity of the demand for real money balances should be zero. Hence, it is interesting to test the hypothesis $\delta = 0$. A rejection of this hypothesis would cast doubt on the aforementioned theory, which predicts δ to be zero. Rejecting this hypothesis therefore means that the theory cannot be quite true.

Spencer (1985), however, wants to use the evidence of not rejecting H_0 as proof that the theory is valid. He for instance reports that "for the period 1952:2 - 1972:4, zero price elasticity is strongly supported"; "...the hypothesis of zero price level elasticity is rejected for the full period but accepted for each of the two subperiods"; zero price elasticity "... receives strong support", andsoforth. From a statistical point of view, such statements are doubtful at best. We learn from rejecting restrictions, not from accepting them.

3.5 Summary: advantages and disadvantages of MLE

Maximum likelihood is not the only method for estimating models for micro data. Alternative methods include

- ordinary least squares
- non-linear least squares
- (generalized) method of moments

The first method is encountered in all introductory econometrics courses, the second and third methods, while also important, are not yet so standard. One justification for concentration on maximum likelihood is that it is widely applicable and, in practice, the most frequently used method for this type of data, with the exception of OLS. Still, one should be aware that the method has limitations and disadvantages.

Disadvantages

- Specific assumptions on parametric probability distribution are needed.
- Often theory provides little guidance on functional form and probability model and one has to make somewhat arbitrary assumptions.
- In general, MLE is not robust with respect to misspecification.
- The MLE does not always exist, since the likelihood function may become unbounded.

Advantages

- Convenience: all that one needs to do is to write down the likelihood function. First and second derivatives can be obtained numerically.
- The computer produces a numerical answer (as long as the problem is well defined, parameters are identified etc.)
- There is a well established large sample theory (asymptotic normality, consistency, and efficiency)
- The invariance property generates flexibility in reformulating the model.
- The inference is simple.
- The formulation of the likelihood function forces one to think carefully of the problem, the way the sample was generated, etc.

3.6 Exercises

Exercise 3.1 Consider a random sample $y_1, y_2, \ldots, y_n$ drawn from a Poisson distribution with parameter λ. Show that the information matrix can also be calculated as

$$I(\lambda) = \mathrm{E}\left[\left(\frac{\partial \log L(\lambda)}{\partial \lambda}\right)^2\right]$$

Exercise 3.2 Suppose $y_i|x_i$ is Poisson distributed with parameter $\lambda_i = \alpha + \theta x_i$.

(a) Does this specification make sense?

(b) Show that the first-order conditions can be written as

$$\sum_{i=1}^{n} \frac{(y_i - \mathrm{E}(y_i|x_i))}{\mathrm{Var}(y_i|x_i)} = 0$$

$$\sum_{i=1}^{n} \frac{(y_i - \mathrm{E}(y_i|x_i))x_i}{\mathrm{Var}(y_i|x_i)} = 0$$

Exercise 3.3 Suppose you sample y_i's independently from an exponential density

$$f(y_i) = \lambda e^{-\lambda y_i}, \qquad y_i \geq 0,\ \lambda \geq 0$$

where $\mathrm{E}(y_i) = 1/\lambda$ and $\mathrm{Var}(y_i) = 1/\lambda^2$.

(a) Find the maximum likelihood estimator for λ. Is this estimator consistent? What is the asymptotic distribution of $\hat{\lambda}$?

(b) Now, assume that you observe explanatory variables as well, and that the conditional density of $y_i|x_i$ is of exponential form with parameter

$$\lambda_i = \frac{1}{\alpha + \theta x_i}$$

Does this specification make sense?

(c) Show that the first-order conditions can be written as

$$\sum_{i=1}^{n} \frac{(y_i - \mathrm{E}(y_i|x_i))}{\mathrm{Var}(y_i|x_i)} = 0$$

$$\sum_{i=1}^{n} \frac{(y_i - \mathrm{E}(y_i|x_i))x_i}{\mathrm{Var}(y_i|x_i)} = 0$$

Exercise 3.4 Assume that you have a sample of n independent observations from a *geometric* distribution with

$$f(y_i) = \theta(1-\theta)^{y_i}, \qquad y_i = 0, 1, 2, \ldots$$

(a) Write down the log-likelihood function.

(b) Derive the maximum likelihood estimator.

(c) Calculate the Fisher information matrix of the sample.

(d) What is the asymptotic variance of the maximum likelihood estimator?

Exercise 3.5 Assume that you have a sample of n independent observations from a distribution with p.d.f.

$$f(y_i) = \theta y_i^{\theta-1} \text{ for } 0 < y_i < 1$$

(a) Find the maximum likelihood estimator of θ.

(b) Find the asymptotic distribution of the maximum likelihood estimator.

Exercise 3.6 In the literature on the economics of happiness, one often finds statements such as 'happiness is u-shaped in age'. What does this statement mean? What type of regression specification is needed to have such a conclusion?

Exercise 3.7 Depending on your answer in 3.6, can you give a different regression specification that would allow for u-shaped age effect as well?

Exercise 3.8 Consider the model

$$\ln y_i \sim Normal(\theta_0 + \theta_1 \text{schooling}_i + \theta_2 \text{age}_i + \theta_3 \text{age}_i^2, \sigma^2)$$

where $\ln y_i$ are log-earnings. How can you test whether age has no influence on log-earnings?

Exercise 3.9 You observe a random sample with population model

$$f(y_i|x_{i1}, x_{i2}, T_i) = f(y_i|\theta_1 + \theta_2 T_i + \theta_3 x_{i1} + \theta_4 x_{i1} T_i + \theta_5 x_{i2} + \theta_6 x_{i2} T_i)$$

where f is a known probability (density) function, x_{i1} and x_{i2} are continuous variables and T_i is a treatment indicator:

$$T_i = \begin{cases} 1 \text{ if individual receives treatment} \\ 0 \text{ else} \end{cases}$$

How could you test whether the conditional model $f(y_i|x_{i1}, x_{i2}, T_i)$ is the same for the treated and the non-treated, i.e.,

$$f(y_i|x_{i1}, x_{i2}, T_i) = f(y_i|x_{i1}, x_{i2}) \quad ?$$

Exercise 3.10 Show that in 3.9, the estimation of the unrestricted model is equivalent to estimating the model separately for the two groups.

Exercise 3.11 The following earnings equation has been estimated (estimated standard errors in parentheses):

$$\begin{array}{lcllll} \widehat{\ln y_i} & = & 8.5633\ + & 0.3733\,\text{voc}_i\ + & 0.6658\,\text{uni}_i\ + & 0.1023\,\text{age}_i\ - \quad 0.0010\,\text{age}_i^2 \\ & & (0.1421) & (0.0409) & (0.0510) & (0.0072) \qquad\quad (0.0001) \end{array}$$

(a) Determine the age at which earnings are maximized.

(b) What (if anything) can you say about the standard error of this age?

References

Cramer, J.S. (1986): *Econometric Applications of Maximum Likelihood Methods*, Cambridge University Press

Gourieroux C., A. Monfort, and A. Trognon (1984): Pseudo-Maximum Likelihood Methods: Theory, *Econometrica*, 52, 681700.

Greene, W.H. (2000): *Econometric Analysis*, fourth edition, Prentice Hall.

Knapp, L.G. and T.G. Seaks (1992): An Analysis of the Probability of Default on Federally Guaranteed Student Loans, *Review of Economics and Statistics*, 74(3), 404-11.)

Maddala, G.S. (1983) *Limited Dependent and Qualitative Variables in Econometrics*, Cambridge University Press.

Manski, C.F. (1995): *Identification Problems in the Social Sciences*, Harvard University Press.

McCloskey, D.N. (1985): The Loss Function Has Been Mislaid: The Rhetoric of Significance Tests, *American Economic Review*, 75, 201-205.

McFadden, D. (1974): Conditional logit analysis of qualitative choice behavior. In: P. Zarembka (ed.), *Frontiers in Econometrics*, Academic Press, New York.

Rothenberg, T.J. (1971): Identification in parametric models, *Econometrica*, 39, 577-591.

Spencer, D.E. (1985): Money demand and the price level, *Review of Economics and Statistics*, 67(3), 490-496.