University College London

Department of Economics

M.Sc. in Economics

MC3: Econometric Theory and Methods

Course Notes 4

Notes on maximum likelihood methods

Andrew Chesher

25/10/2005

1. Introduction

These notes introduce the concept of a likelihood function and the maximum likelihood estimator. The estimator's properties are studied using the tools developed for M-estimators. Important microeconometric examples of the application of maximum likelihood estimation are introduced.

2. Maximum likelihood estimation

Some of the models used in econometrics specify the complete probability distribution of the outcomes of interest are specified rather than just a regression function. Sometimes this is because of special features of the outcomes under study - for example because they are discrete or censored, or because there is serial dependence of a complex form, a situation which arises when event histories (e.g. of labour market transitions or fertility) are studied.[1]

When the complete probability distribution of outcomes given covariates is specified we can develop an expression for the probability of observation of the responses we see as a function of the unknown parameters embedded in the specification. With this to hand we can ask what values of these parameters maximise this probability for the data we have. The resulting statistics, functions of the observed data, are called *maximum likelihood estimators*. They possess important optimality properties and have the advantage that they can be produced in a rule directed fashion.

We start with a very simple problem, the estimation of the probability that an event occurs (e.g. finding a job, getting married) in a situation in which this probability is the same for all agents that are observed. Later we extend the model to cover situations in which the probability may vary across agents, that is when we are interested in the conditional probability of the event occurring given characteristics of the agents and their environment. Some elements of the theory of maximum likelihood estimators are outlined.

2.1. Estimating a probability

Suppose $Y_1, \ldots Y_n$ are binary independently and identically distributed random variables with $P[Y_i = 1] = p$, $P[Y_i = 0] = 1 - p$ for all i. We might use such a model for data recording the occurrence or otherwise of an event for n individuals, for example being in work or not, buying a good or service or not, etc. Shortly we will consider how to proceed when p depends upon characteristics of individuals and of their environment, but for the moment we stay with the very simple model in which p is the same for all individuals. Let $y_1, \ldots, y_n$ indicate the data

[1] The fine detail of this sort of specification rarely flows from economic theory so we must be on the alert for misspecification and not place too much trust on information flowing from a fitted model that is sensitive to minor changes in the detail of the model specification. Analysis of this sort is particularly fragile if the identifiability of interesting features of structures rests on restrictions only arising in the detail of the specification of the probability law.

values obtained and note that in this model

$$\begin{aligned} P[Y_1 = y_1 \cap \cdots \cap Y_n = y_n] &= \prod_{i=1}^{n} p^{y_i}(1-p)^{(1-y_i)} \\ &= p^{\sum_{i=1}^{n} y_i}(1-p)^{\sum_{i=1}^{n}(1-y_i)} \\ &= L(p;y). \end{aligned}$$

With any set of data $L(p;y)$ can be calculated for any value of p between 0 and 1. The result is the probability of observing the data to hand for each chosen value of p. One strategy for estimating p is to use that value that maximises this probability. The resulting estimator is called the *maximum likelihood estimator* (MLE) and the maximand, $L(p;y)$, is called the *likelihood function.*

The maximum of the *log likelihood function*, $l(p;y) = \log L(p,y)$, is at the same value of p as is the maximum of the likelihood function (because the log function is monotonic). It is often easier to maximise the log likelihood function (LLF) - further, because in many cases this is a sum of terms, one for each data point, central limit theorems will apply to suitably scaled versions of the LLF and statistics derived from it. For the problem considered here the LLF is

$$l(p;y) = \left(\sum_{i=1}^{n} y_i\right)\log p + \sum_{i=1}^{n}(1-y_i)\log(1-p).$$

Let[2]

$$\hat{p} = \arg\max_{p} L(p;y) = \arg\max_{p} l(p;y).$$

On differentiating we have the following.

$$\begin{aligned} l_p(p;y) &= \frac{1}{p}\sum_{i=1}^{n} y_i - \frac{1}{1-p}\sum_{i=1}^{n}(1-y_i) \\ l_{pp}(p;y) &= -\frac{1}{p^2}\sum_{i=1}^{n} y_i - \frac{1}{(1-p)^2}\sum_{i=1}^{n}(1-y_i). \end{aligned}$$

Note that $l_{pp}(p;y)$ is always negative for admissable p so the optimisation problem has a unique solution corresponding to a maximum. The solution to $l_p(\hat{p};y) = 0$ is

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

just the mean of the observed values of the binary indicators, equivalently the proportion of 1's observed in the data.

[2] By $\arg\max_{p} l(p,y)$ I mean the value of the argument p at which $l(p,y)$ achieves a maximum.

2.2. Likelihood functions and estimation in general

Let Y_i, $i = 1, \ldots, n$ be continuously distributed random variables with joint probability density function $f(y_1, \ldots, y_n, \theta)$. The probability that Y falls in infinitesimal intervals of width $dy_1, \ldots dy_n$ centred on values $y_1, \ldots, y_n$ is

$$A = f(y_1, \ldots, y_n, \theta)dy_1 dy_2 \ldots dy_n$$

Here only the joint density function depends upon θ and the value of θ that maximises $f(y_1, \ldots, y_n, \theta)$ also maximises A. In this case the likelihood function is defined to be the joint *density* function of the Y_i's.

When the Y_i's are discrete random variables the likelihood function is the joint probability mass function of the Y_i's, and in cases in which there are discrete and continuous elements the likelihood function is a combination of probability density elements and probability mass elements. In all cases the likelihood function is a function of the observed data values that is equal to, or proportional to, the probability of observing these particular values, where the constant of proportionality does not depend upon the parameters which are to be estimated.

When Y_i, $i = 1, \ldots, n$ are *independently* distributed the joint density (mass) function is the *product* of the marginal density (mass) functions of each Y_i, the likelihood function is

$$L(y; \theta) = \prod_{i=1}^{n} f_i(y_i; \theta),$$

and the log likelihood function is the *sum*:

$$l(y; \theta) = \sum_{i=1}^{n} \log f_i(y_i; \theta).$$

There is a subscript on f to allow for the possibility that each Y_i has a distinct probability distribution. This situation arises when modelling conditional distributions of Y given some covariates x. Then f varies depending upon the covariate values. In particular, $f_i(y_i; \theta) = f_i(y_i|x_i; \theta)$, the conditional density (mass) function of Y given x. Often in this context we will define the conditional distribution of Y given x to be the same function for all i, i.e. $f_i(y_i|x_i; \theta) = f(y_i|x_i; \theta)$. In setting out the properties of maximum likelihood estimators below we will suppress any dependence on covariates in the notation except where this would cause confusion.

In time series and panel data problems there is often dependence among the Y_i's, indeed this is sometimes of central interest. For any list of random variables $Y = \{Y_1, \ldots, Y_n\}$ define the $i-1$ element list $Y_{i-} = \{Y_1, \ldots, Y_{i-1}\}$. Since the joint density (mass) function of Y can be written as the product of conditional density (mass) functions and the marginal density of Y_1 as follows,

$$f(y) = \prod_{i=2}^{n} f_{y_i|y_{i-}}(y_i|y_{i-}) f_{y_1}(y_1),$$

we can always write the log likelihood function as the sum

$$f(y) = \sum_{i=1}^{n} \log f_{y_i|y_{i-}}(y_i|y_{i-}) + \log f_{y_1}(y_1).$$

2.2.1. Invariance

Note that (parameter free) monotonic transformations of the Y_i's (for example, a change of units of measurement, or use of logs rather than the original y data) usually leads to a change in the value of the maximised likelihood function when we work with continuous distributions. For example if we transform from y to z where $y = h(z)$ and the joint density function of y is $f_y(y;\theta)$ then the joint density function of z is

$$f_z(z;\theta) = \left|\frac{\partial h(z)}{\partial z}\right| f_y(h(z);\theta).$$

For any given set of values, y^*, the value of θ that maximises the likelihood function $f_y(y^*,\theta)$ also maximises the likelihood function $f_z(z^*;\theta)$ where $y^* = h(z^*)$, so the maximum likelihood estimator is invariant with respect to such changes in the way the data are presented. However the maximised likelihood functions will differ by a factor equal to $\left|\frac{\partial h(z)}{\partial z}\right|_{z=z^*}$. The reason for this is that we omit the infinitesimals $dy_1, \ldots dy_n$ from the likelihood function for continuous variates and these change when we move from y to z because they are denominated in the units in which y or z are measured. The implication of this is that two researchers estimating the same model but using different transformations of the data will produce the *same* MLE but *different* values for the maximised likelihood function, so these values cannot be directly compared, though they can clearly be adjusted to make them comparable, by multiplying by the factor identified above.

Maximum likelihood estimators possess another important *invariance property*. Suppose two researchers choose different ways in which to parameterise the same model. One uses θ, and the other uses $\lambda = h(\theta)$ where this function is one-to-one. Then faced with the same data and producing estimators $\hat{\theta}$ and $\hat{\lambda}$, it will always be the case that $\hat{\lambda} = h(\hat{\theta})$. There are a number of important consequences of this.

One arises in the following situation. There are some cases in which an economically interesting magnitude is a function of the parameters in which a model is perhaps naturally parameterised. For example in travel demand studies we find models in which the indirect utility gained using alternative travel modes is a function of an index $x'\theta$ where one element of x records travel time and another records travel cost. The ratio of the coefficients on these two covariates can be interpreted as the value of time because this ratio shows the rate at which cost has to be adjusted to compensate for travel time increases, keeping utility unchanged. The invariance property just described implies that the MLE of the value of time is just the ratio of the MLEs of the two coefficients in the index.

Another important consequence of the invariance property arises because sometimes a re-parameterisation can improve the numerical properties of the likelihood function. Newton's method and its variants may in practice work better if parameters are rescaled. An example of this often arises when, in index models, elements of x involve squares, cubes, etc., of some covariate, say x_1. Then maximisation of the likelihood function may be easier if instead of x_1^2, x_1^3, etc., you use $x_1^2/10$, $x_1^3/100$, etc., with consequent rescaling of the coefficients on these covariates. You can always recover the MLEs you would have obtained without the rescaling by rescaling the estimates.

There are some cases in which a re-parameterisation can produce a globally concave likelihood function where in the original parameterisation there was not global concavity - recall Newton's

method works best with concave maximisation problems (convex minimisation problems). An example of this arises in the "Tobit" model. This is a model in which each Y_i is $N(x_i'\beta, \sigma^2)$ with negative realisations replaced by zeros. The model is sometimes used to model expenditures and hours worked, which are necessarily non-negative. In this model the likelihood as parameterised here is not globally concave, but re-parameterising to $\lambda = \beta/\sigma$, and $\gamma = 1/\sigma$, produces a globally concave likelihood function. The invariance property tells us that having maximised the "easy" likelihood function and obtained estimates $\hat{\lambda}$ and $\hat{\gamma}$, we can recover the maximum likelihood estimates we might have had difficulty finding in the original parameterisation by calculating $\hat{\beta} = \hat{\lambda}/\hat{\gamma}$ and $\hat{\sigma} = 1/\hat{\gamma}$.

2.3. Properties of maximum likelihood estimators

Here we just sketch the main results which we will use later. Let $l(\theta; Y)$ be the log likelihood function now regarded as a random variable, a function of a set of (possibly vector) random variables $Y = \{Y_1, \ldots, Y_n\}$. Let $l_\theta(\theta; Y)$ be the gradient of this function, itself a vector of random variables (scalar if θ is scalar) and let $l_{\theta\theta}(\theta; Y)$ be the matrix of second derivatives of this function (also a scalar if θ is a scalar). Let

$$\hat{\theta} = \arg\max_{\theta} l(\theta; Y).$$

This is a function of Y as long as the maximum is defined. In order to make inferences about θ using $\hat{\theta}$ we need to determine the distribution of $\hat{\theta}$. This is hard to obtain exactly except in simple problems[3] so instead we consider developing a large sample approximation.

The limiting distribution for a quite wide class of maximum likelihood problems is as follows.

$$n^{1/2}(\hat{\theta} - \theta) \xrightarrow{d} N(0, V_0)$$

where

$$V_0 = -\operatorname*{plim}_{n\to\infty}(n^{-1} l_{\theta\theta}(\theta_0; Y))^{-1}$$

and θ_0 is the unknown parameter value. To get an approximate distribution that can be used in practice we use $(n^{-1} l_{\theta\theta}(\hat{\theta}; Y))^{-1}$ or some other consistent estimator of V_0 in place of V_0.

[3] Actually we can obtain it exactly in the case considered in Section 2.1. In that model

$$P[\sum_{i=1}^{n} Y_i = m] = \frac{n!}{m!(n-m)!} p^m (1-p)^{n-m}$$

which is the probability mass associated with a binomial distribution. Therefore

$$\begin{aligned} P[\hat{p} = a] &= P[\sum_{i=1}^{n} Y_i = na] \\ &= \frac{n!}{(na)!(n-na)!} p^{na} (1-p)^{n-na} \end{aligned}$$

where $a \in \{0, 1/n, 2/n, \ldots, 1\}$. Further

$$\begin{aligned} E[\hat{p}] &= p \\ Var[\hat{p}] &= p(1-p)/n. \end{aligned}$$

Clearly $\hat{p}$ converges in mean square to p and so $\hat{p}$ is a consistent estimator.

The argument that leads to this comes by applying our method for dealing with M-estimators set out earlier.

Suppose $\hat{\theta}$ is uniquely determined as the solution to the first order condition

$$l_\theta(\hat{\theta}; Y) = 0$$

and that $\hat{\theta}$ is a consistent estimator of the unknown value of the parameter, θ_0. Weak conditions required for consistency are quite complicated and will not be given here - they can be found in the intermediate textbooks. They require independence or at most weak dependence across log likelihood function contributions, existence and boundedness of low order derivatives of the log likelihood function, existence of certain moments involving these derivatives, and, when there are covariates, consideration of a benign evolution of covariate values as the sample size increases. An important condition is that the probability limit of n^{-1} times the log likelihood function have a unique maximum, located at the unknown parameter value.

Taking a Taylor series expansion around $\theta = \theta_0$ and then evaluating this at $\theta = \hat{\theta}$ gives

$$0 \simeq l_\theta(\theta_0; Y) + l_{\theta\theta'}(\theta_0; Y)(\hat{\theta} - \theta_0)$$

and rearranging and scaling by powers of the sample size n

$$n^{1/2}(\hat{\theta} - \theta_0) \simeq -\left(n^{-1} l_{\theta\theta'}(\theta; Y)\right)^{-1} n^{-1/2} l_\theta(\theta; Y).$$

As in our general treatment of M-estimators if we can show that

$$n^{-1} l_{\theta\theta'}(\theta_0; Y) \xrightarrow{p} A(\theta_0)$$

and

$$n^{-1/2} l_\theta(\theta_0; Y) \xrightarrow{d} N(0, B(\theta_0))$$

then

$$n^{1/2}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, A(\theta_0)^{-1} B(\theta_0) A(\theta_0)^{-1\prime}).$$

Clearly,

$$A(\theta_0) = \operatorname*{plim}_{n\to\infty} n^{-1} l_{\theta\theta'}(\theta_0; Y),$$

but what is the limiting distribution of $n^{-1/2} l_\theta(\theta_0; Y)$?

First note that in problems for which the Y_i's are independently distributed, $n^{-1/2} l_\theta(\theta_0; Y)$ is a scaled (by $n^{1/2}$) mean of random variables and we may be able to find conditions under which a central limit theorem applies, indicating a limiting *normal* distribution. We must now find the mean and variance of this distribution.

Since $L(\theta; Y)$ is a joint probability density function (we just consider the continuous distribution case here),

$$\int L(\theta; y) dy = 1$$

where multiple integration is over the support of Y. If this support *does not depend upon* θ, then

$$\frac{\partial}{\partial \theta} \int L(\theta; y) dy = \int L_\theta(\theta; y) dy = 0.$$

But, because $l(\theta; y) = \log L(\theta; y)$, and $l_\theta(\theta; y) = L_\theta(\theta; y)/L(\theta; y)$, we have

$$\int L_\theta(\theta; y)dy = \int l_\theta(\theta; y)L(\theta; y)dy = E\left[l_\theta(\theta; Y)\right]$$

and so $E\left[l_\theta(\theta; Y)\right] = 0$. This holds for any value of θ, in particular for θ_0 above. If the variance of $l_\theta(\theta_0; Y)$ converges to zero as n becomes large then $l_\theta(\theta_0; Y)$ will converge in probability to zero and the mean of the limiting distribution of $n^{-1/2}l_\theta(\theta_0; Y)$ will be zero.

We turn now to the variance of the limiting distribution. We have just shown that

$$\int l_\theta(\theta; y)L(\theta; y)dy = 0.$$

Differentiating again

$$\begin{aligned}
\frac{\partial}{\partial \theta'} \int l_\theta(\theta; y)L(\theta; y)dy &= \int \left(l_{\theta\theta'}(\theta; y)L(\theta; y) + l_\theta(\theta; y)L_{\theta'}(\theta; y)\right) dy \\
&= \int \left(l_{\theta\theta'}(\theta; y) + l_\theta(\theta; y)l_\theta(\theta; y)'\right) L(\theta; y)dy \\
&= E\left[l_{\theta\theta'}(\theta; Y) + l_\theta(\theta; Y)l_\theta(\theta; Y)'\right] \\
&= 0.
\end{aligned}$$

Separating the two terms in the penultimate line,

$$E\left[l_\theta(\theta; Y)l_\theta(\theta; Y)'\right] = -E\left[l_{\theta\theta'}(\theta; Y)\right] \tag{2.1}$$

and note that, since $E\left[l_\theta(\theta; Y)\right] = 0$,

$$Var[l_\theta(\theta; Y)] = E\left[l_\theta(\theta; Y)l_\theta(\theta; Y)'\right]$$

and so

$$\begin{aligned}
Var[l_\theta(\theta; Y)] &= -E\left[l_{\theta\theta'}(\theta; Y)\right] \\
\Rightarrow Var[n^{-1/2}l_\theta(\theta; Y)] &= -E\left[n^{-1}l_{\theta\theta'}(\theta; Y)\right]
\end{aligned}$$

giving

$$B(\theta_0) = -\operatorname*{plim}_{n\to\infty} n^{-1}l_{\theta\theta'}(\theta_0; Y).$$

The matrix

$$I(\theta) = -E\left[l_{\theta\theta}(\theta; Y)\right]$$

plays a central role in likelihood theory - it is called the *Information Matrix*.

Finally, because $B(\theta_0) = -A(\theta_0)$

$$A(\theta)^{-1}B(\theta)A(\theta)^{-1\prime} = -\left(\operatorname*{plim}_{n\to\infty} n^{-1}l_{\theta\theta'}(\theta; Y)\right)^{-1}.$$

Of course a number of conditions are required to hold for the results above to hold. These include the boundedness of third order derivatives of the log likelihood function, independence

or at most weak dependence of the Y_i's, existence of moments of derivatives of the log likelihood, or at least of probability limits of suitably scaled versions of them, and lack of dependence of the support of the Y_i's on θ.

The result in equation (2.1) above leads, under suitable conditions concerning convergence, to

$$\operatorname*{plim}_{n\to\infty}\left(n^{-1}l_\theta(\theta;Y)l_\theta(\theta;Y)'\right) = -\operatorname*{plim}_{n\to\infty}\left(n^{-1}l_{\theta\theta'}(\theta;Y)\right).$$

This gives an alternative way of "estimating " V_0, namely

$$\hat{V}_0^o = \left\{n^{-1}l_\theta(\hat{\theta};Y)l_\theta(\hat{\theta};Y)'\right\}^{-1}$$

which compared with

$$\tilde{V}_0^o = \left\{-n^{-1}l_{\theta\theta'}(\hat{\theta};Y)\right\}^{-1}$$

has the advantage that only first derivatives of the log likelihood function need to be calculated. Sometimes $\hat{V}_0^o$ is referred to as the "outer product of gradient" (OPG) estimator. Both these estimators use the "observed" values of functions of derivatives of the LLF and. It may be possible to derive explicit expressions for the expected values of these functions. Then one can estimate V_0 by

$$\begin{aligned}\hat{V}_0^e &= \left\{E[n^{-1}l_\theta(\theta;Y)l_\theta(\theta;Y)']|_{\theta=\hat{\theta}}\right\}^{-1}\\ &= \left\{-E[n^{-1}l_{\theta\theta'}(\theta;Y)]|_{\theta=\hat{\theta}}\right\}^{-1}.\end{aligned}$$

These two sorts of estimators are sometimes referred to as "observed information" ($\hat{V}_0^o$, $\tilde{V}_0^o$) and "expected information" ($\hat{V}_0^e$) estimators.

Maximum likelihood estimators possess optimality property, namely that, among the class of consistent and asymptotically normally distributed estimators, the variance matrix of their limiting distribution is the smallest that can be achieved in the sense that other estimators in the class have limiting distributions with variance matrices exceeding the MLE's by a positive semidefinite matrix.

We will shortly consider how to conduct inference in a likelihood framework. First, here are a couple of applications.

2.4. Estimating a conditional probability

Suppose $Y_1, \ldots Y_n$ are binary independently and identically distributed random variables with

$$\begin{aligned}P[Y_i &= 1|X = x_i] = p(x_i,\theta)\\ P[Y_i &= 0|X = x_i] = 1 - p(x,\theta).\end{aligned}$$

This is an obvious extension of the model in the previous section which we would use if we wanted to understand how characteristics of agents and their environment affect choices or the occurrence of events.

The likelihood function for this problem is

$$\begin{aligned} P[Y_1 = y_1 \cap \cdots \cap Y_1 = y_1 | x] &= \prod_{i=1}^{n} p(x_i, \theta)^{y_i} (1 - p(x_i, \theta))^{(1-y_i)} \\ &= L(\theta; y). \end{aligned}$$

where y denotes the complete set of values of y_i and dependence on x is suppressed in the notation. The log likelihood function is

$$l(\theta; y) = \sum_{i=1}^{n} y_i \log p(x_i, \theta) + \sum_{i=1}^{n} (1 - y_i) \log(1 - p(x_i, \theta))$$

and the maximum likelihood estimator of θ is

$$\hat{\theta} = \arg\max_{\theta} l(\theta; y).$$

So far this is an obvious generalisation of the simple problem met in the last section.

To implement the model we choose a form for the function $p(x, \theta)$, which must of course lie between zero and one. One common choice is

$$p(x, \theta) = \frac{\exp(x'\theta)}{1 + \exp(x'\theta)}$$

which produces what is commonly called a *logit model*. Another common choice is

$$\begin{aligned} p(x, \theta) &= \Phi(x'\theta) = \int_{-\infty}^{x'\theta} \phi(w) dw \\ \phi(w) &= (2\pi)^{-1/2} \exp(-w^2/2) \end{aligned}$$

in which Φ is the standard normal distribution function. This produces what is known as a *probit model*. Both models are widely used, and often rather uncritically. Note that in both cases a single index model is specified, the probability functions are monotonic increasing, probabilities arbitrarily close to zero or one are obtained when $x'\theta$ is sufficiently large or small, and there is a symmetry in both of the models in the sense that $p(-x, \theta) = 1 - p(x, \theta)$. Any or all of these properties might be inappropriate in a particular application but there is rarely discussion of this in the applied econometrics literature.

2.4.1. Single index models

We can cover both cases by considering general single index models, so for the moment rewrite $p(x, \theta)$ as $g(w)$ where $w = x'\theta$. Then the first derivative of the log likelihood function is as follows.

$$\begin{aligned} l_\theta(\theta; y) &= \sum_{i=1}^{n} \frac{g_w(x_i'\theta) x_i}{g(x_i'\theta)} y_i - \frac{g_w(x_i'\theta) x_i}{1 - g(x_i'\theta)} (1 - y_i) \\ &= \sum_{i=1}^{n} (y_i - g(x_i'\theta)) \frac{g_w(x_i'\theta)}{g(x_i'\theta)\,(1 - g(x_i'\theta))} x_i \end{aligned}$$

Here $g_w(w)$ is the derivative of $g(w)$ with respect to w. The expression for the second derivative is rather messy. Here we just note that its expected value given x is quite simple, namely

$$E[l_{\theta\theta}(\theta; y)|x] = -\sum_{i=1}^{n} \frac{g_w(x_i'\theta)^2}{g(x_i'\theta)\left(1 - g(x_i'\theta)\right)} x_i x_i',$$

the negative of which is the Information Matrix for general single index binary data models.

2.4.2. The logit model

For the logit model there is major simplification

$$\begin{aligned} g(w) &= \frac{\exp(w)}{1+\exp(w)} \\ g_w(w) &= \frac{\exp(w)}{(1+\exp(w))^2} \\ \Rightarrow \frac{g_w(w)}{g(w)\left(1-g(w)\right)} &= 1. \end{aligned}$$

Therefore in the logit model the MLE satisfies

$$\sum_{i=1}^{n} \left(y_i - \frac{\exp(x_i'\hat{\theta})}{1+\exp(x_i'\hat{\theta})} \right) x_i = 0,$$

the Information Matrix is

$$I(\theta) = \sum_{i=1}^{n} \frac{\exp(x_i'\theta)}{(1+\exp(x_i'\theta))^2} x_i x_i',$$

the MLE has the limiting distribution

$$\begin{aligned} & n^{1/2}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, V_0) \\ V_0 &= \left(\operatorname*{plim}_{n\to\infty} n^{-1} \sum_{i=1}^{n} \frac{\exp(x_i'\theta)}{(1+\exp(x_i'\theta))^2} x_i x_i' \right)^{-1}, \end{aligned}$$

and we can conduct approximate inference using the following approximation

$$n^{1/2}(\hat{\theta}_n - \theta) \simeq N(0, V_0)$$

using the estimator

$$\hat{V_0} = \left(n^{-1} \sum_{i=1}^{n} \frac{\exp(x_i'\hat{\theta})}{\left(1+\exp(x_i'\hat{\theta})\right)^2} x_i x_i' \right)^{-1}$$

when producing approximate hypothesis tests and confidence intervals.

2.4.3. The probit model

In the probit model

$$\begin{aligned} g(w) &= \Phi(w) \\ g_w(w) &= \phi(w) \\ \Rightarrow \frac{g_w(w)}{g(w)\,(1-g(w))} &= \frac{\phi(w)}{\Phi(w)(1-\Phi(w))}. \end{aligned}$$

Therefore in the probit model the MLE satisfies

$$\sum_{i=1}^{n} \left(y_i - \Phi(x_i'\hat{\theta})\right) \frac{\phi(x_i'\hat{\theta})}{\Phi(x_i'\hat{\theta})(1-\Phi(x_i'\hat{\theta}))} x_i = 0,$$

the Information Matrix is

$$I(\theta) = \sum_{i=1}^{n} \frac{\phi(x_i'\theta)^2}{\Phi(x_i'\theta)(1-\Phi(x_i'\theta))} x_i x_i',$$

the MLE has the limiting distribution

$$\begin{aligned} & n^{1/2}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, V_0) \\ V_0 &= \left(\operatorname*{plim}_{n\to\infty} n^{-1} \sum_{i=1}^{n} \frac{\phi(x_i'\theta)^2}{\Phi(x_i'\theta)(1-\Phi(x_i'\theta))} x_i x_i' \right)^{-1}, \end{aligned}$$

and we can conduct approximate inference using the following approximation

$$n^{1/2}(\hat{\theta}_n - \theta) \simeq N(0, V_0)$$

using the estimator

$$\hat{V}_0 = \left(n^{-1} \sum_{i=1}^{n} \frac{\phi(x_i'\hat{\theta})^2}{\Phi(x_i'\hat{\theta})(1-\Phi(x_i'\hat{\theta}))} x_i x_i' \right)^{-1}$$

when producing approximate tests and confidence intervals.

2.5. Models for count data

The methods developed above are useful when we want to model the occurrence or otherwise of an event. Sometimes we want to model the number of times an event occurs - above this was zero or one. In general it might be any nonnegative integer. Count data are being used increasingly in econometrics. An interesting application is to the modelling of the returns to R&D investment in which data on numbers of patents filed in a series of years by a sample of companies is studied and related to data on R&D investments.

Binomial and Poisson probability models provide common starting points in the development of count data models. If $Z_1, \ldots, Z_m$ are identically and independently distributed binary random

variables with $P[Z_i = 1] = p$, $P[Z_i = 0] = 1 - p$, then the sum of the Z_i's has a Binomial distribution,

$$Y = \sum_{i=1}^{m} Z_i \sim Bi(m, p)$$

and

$$P[Y = j] = \frac{m!}{j!(m-j)!} p^j (1-p)^{m-j}, \qquad j \in \{0, 1, 2, \ldots, m\}$$

As m becomes large, $m^{1/2}(m^{-1}Y - p)$ becomes approximately normally distributed, $N(0, p(1-p))$, and as m becomes large while $mp = \lambda$ remains constant, Y comes to have a Poisson distribution,

$$Y \sim Po(\lambda)$$

and

$$P[Y = j] = \frac{\lambda^j}{j!} \exp(-\lambda), \qquad j \in \{0, 1, 2, \ldots\}.$$

In each case letting p or λ be functions of covariates creates a model for the conditional distribution of a count of events given covariate values. The Poisson model is much more widely used, in part because there is no need to specify or estimate the parameter m. In the application to R&D investment one might imagine that a firm seeds a large number of research projects in a period of time, each of which has only a small probability of producing a patent. This is consonant with the Poisson probability model but note that one might be concerned about the underlying assumption of independence across projects built into the Poisson model.

With a model specified, maximum likelihood estimation proceeds as set out above. The Poisson model is used as an example. Suppose that we specify a single index model:

$$P[Y_i = y_i | x_i] = \frac{\lambda(x_i'\theta)^{y_i}}{y_i!} \exp(-\lambda(x_i'\theta)), \qquad j \in \{0, 1, 2, \ldots\}.$$

The log likelihood function is

$$l(\theta, y) = \sum_{i=1}^{n} y_i \log \lambda(x_i'\theta) - \lambda(x_i'\theta) - \log y_i!$$

with first derivative

$$\begin{aligned} l_\theta(\theta, y) &= \sum_{i=1}^{n} \left(y_i \frac{\lambda_w(x_i'\theta)}{\lambda(x_i'\theta)} - \lambda_w(x_i'\theta) \right) x_i \\ &= \sum_{i=1}^{n} \left(y_i - \lambda(x_i'\theta) \right) \frac{\lambda_w(x_i'\theta)}{\lambda(x_i'\theta)} x_i \end{aligned}$$

where $\lambda_w(w)$ is the derivative of $\lambda(w)$ with respect to w.

The MLE satisfies

$$\sum_{i=1}^{n} \left(y_i - \lambda(x_i'\hat{\theta}) \right) \frac{\lambda_w(x_i'\hat{\theta})}{\lambda(x_i'\hat{\theta})} x_i.$$

The second derivative matrix is

$$l_{\theta\theta}(\theta, y) = \sum_{i=1}^{n} (y_i - \lambda(x_i'\theta)) \left(\frac{\lambda_{ww}(x_i'\theta)}{\lambda(x_i'\theta)} - \left(\frac{\lambda_w(x_i'\theta)}{\lambda(x_i'\theta)} \right)^2 \right) x_i x_i' - \sum_{i=1}^{n} \frac{\lambda_w(x_i'\theta)^2}{\lambda(x_i'\theta)} x_i x_i'$$

where, note, the first term has expected value zero. Therefore the Information Matrix for this conditional Poisson model is

$$I(\theta) = \sum_{i=1}^{n} \frac{\lambda_w(x_i'\theta)^2}{\lambda(x_i'\theta)} x_i x_i'.$$

The limiting distribution of the MLE is (under suitable conditions)

$$\begin{aligned} n^{1/2}(\hat{\theta} - \theta_0) \quad & \xrightarrow{d} N(0, V_0) \\ V_0 \quad &= \quad \left(\operatorname*{plim}_{n\to\infty} n^{-1} \sum_{i=1}^{n} \frac{\lambda_w(x_i'\theta)^2}{\lambda(x_i'\theta)} x_i x_i' \right)^{-1} \end{aligned}$$

and we can make approximate inference about θ_0 using

$$(\hat{\theta} - \theta_0) \simeq N\left(0, n^{-1} V_0\right)$$

with V_0 estimated by

$$\hat{V}_0 = \left(n^{-1} \sum_{i=1}^{n} \frac{\lambda_w(x_i'\hat{\theta})^2}{\lambda(x_i'\hat{\theta})} x_i x_i' \right)^{-1}.$$

In applied work a common choice is $\lambda(w) = \exp(w)$ for which

$$\frac{\lambda_w(w)}{\lambda(w)} = 1 \qquad \frac{\lambda_w(w)^2}{\lambda(w)} = \exp(w).$$