University College London

Department of Economics

M.Sc. in Economics

MC3: Econometric Theory and Methods

Notes on likelihood based hypothesis testing and specification tests

Andrew Chesher

6/11/2005

1. Introduction

These notes start with a discussion of hypothesis testing in a likelihood context and we introduce the notions of Wald, Likelihood ratio and score tests. The latter are widely used as the basis for specification tests which is the second topic covered here.

The notes continue with an example of a commonly occurring problem in which maximum likelihood methods provide one of the few easily implemented solutions - the modelling of outcomes when they are not fully revealed.

The notes end with a brief discussion of Bayesian inferential methods.

2. Tests of hypotheses in a likelihood framework

We now consider test of hypotheses in econometric models in which the complete probability distribution of outcomes given conditioning variables is specified. In this situation the maximum likelihood estimator can be computed and it possesses optimality properties.

There are three natural ways to develop tests of hypotheses when a likelihood function is available.

1. Is the unrestricted ML estimator significantly far from the hypothesised value? This leads to what is known as the Wald test.

2. If the ML estimator is restricted to satisfy the hypothesis, is the value of the maximised likelihood function significantly smaller than the value obtained when the restrictions of the hypothesis are not imposed? This leads to what is known as the likelihood ratio test.

3. If the ML estimator is restricted to satisfy the hypothesis, are the Lagrange multipliers associated with the restrictions of the hypothesis significantly far from zero? This leads to what is known as the Lagrange multiplier or score test.

In the normal linear regression model all three approaches, after minor adjustments, lead to the same statistic which has an $F^{(j)}_{(n-k)}$ distribution when the null hypothesis is true and there are j restrictions.

Outside that special case, in general the three methods lead to different statistics, but in large samples the differences tend to be small. All three statistics have, under certain weak conditions, $\chi^2_{(j)}$ limiting distributions when the null hypothesis is true and there are j restrictions. The exact distributional result in the normal linear regression model fits into this large sample theory on noting that $\text{plim}_{n\to\infty}\left(jF^{(j)}_{(n-k)}\right) = \chi^2_{(j)}$.

We now consider tests of a hypothesis $H_0 : \theta_2 = 0$ where the full parameter vector is partitioned into $\theta' = [\theta_1' \vdots \theta_2']$ and θ_2 contains j elements. Recall that the MLE has the approximate distribution

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

where

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

and $\overline{I}(\theta_0)$ is the asymptotic information matrix per observation.

2.1. The Wald test

This test is obtained by making a direct comparison of $\hat{\theta}_2$ with the hypothesised value of θ_2, zero. Using the approximate distributional result given above leads to the following test statistic.

$$S_W = n\hat{\theta}_2' \widehat{W}_{22}^{-1} \hat{\theta}_2'$$

Here $\widehat{W}_{22}$ is a consistent estimator of the lower right hand $j \times j$ block of V_0. Recall that a variety of ways of estimating this matrix were given in the previous notes. Any of these can be used here. Under the null hypothesis $S_W \xrightarrow{d} \chi^2_{(j)}$ and we reject the null hypothesis for large values of S_W.

Using one of the formulas for the inverse of a partitioned matrix the Wald statistic can also be written as

$$S_W = n\hat{\theta}_2' \left(\widehat{\overline{I}}(\hat{\theta})_{22} - \widehat{\overline{I}}(\hat{\theta})_{21}' \widehat{\overline{I}}(\hat{\theta})_{11}^{-1} \widehat{\overline{I}}(\hat{\theta})_{12} \right) \hat{\theta}_2'$$

where the elements $\widehat{\overline{I}}(\hat{\theta})_{ij}$ are consistent estimators of the appropriate blocks of the asymptotic Information Matrix per observation evaluated at the (unrestricted) MLE.

2.2. The Score - or Lagrange Multiplier - test

To conduct a Wald test we have to estimate θ_2. Sometimes we are in a situation where a model has been estimated with $\theta_2 = 0$, and we would like to see whether the model should be extended by adding additional parameters and perhaps associated conditioning variables or functions of ones already present. In such a situation it is convenient to have a method of conducting a test of the hypothesis that the additional parameters are zero (in which case we might decide not to extend the model) without having to estimate the additional parameters. The *score test* provides such a method.

The score test considers the gradient of the log likelihood function evaluated at the point

$$\hat{\theta}^R = \begin{bmatrix} \hat{\theta}_1^R \\ 0 \end{bmatrix}$$

and examines the departure from zero of that part of the gradient of the log likelihood function that is associated with θ_2. Here $\hat{\theta}_1^R$ is the MLE of θ_1 when θ_2 is restricted to be zero. If the unknown value of θ_2 is in fact zero then this part of the gradient should be close to zero - recall that the expected value of the gradient evaluated at the true parameter values is zero. The score test statistic is

$$S_S = n^{-1} l_\theta(\hat{\theta}^R; Y)' \widehat{\overline{I}}(\hat{\theta}^R)^{-1} l_\theta(\hat{\theta}^R; Y)$$

and $S_S \xrightarrow{d} \chi^2_{(j)}$ under the null hypothesis. Again there are a variety of ways, as set out earlier, of estimating $\widehat{\overline{I}}(\theta_0)$ and hence its inverse.

Note that the complete score (gradient) vector appears in this formula. Of course the part of that associated with θ_1 is zero because we are evaluating at the restricted MLE. That means the score statistic can also be written, using the formula for the inverse of a partitioned matrix, as the algebraically identical

$$S_S = n^{-1} l_{\theta_2}(\hat{\theta}^R; Y)' \left(\widehat{\overline{I}}(\hat{\theta}^R)_{22} - \widehat{\overline{I}}(\hat{\theta}^R)_{21}' \widehat{\overline{I}}(\hat{\theta}^R)_{11}^{-1} \widehat{\overline{I}}(\hat{\theta}^R)_{12} \right)^{-1} l_{\theta_2}(\hat{\theta}^R; Y).$$

When the information matrix is block diagonal[1], which means that the MLEs of θ_1 and θ_2 are asymptotically uncorrelated, the second term in the inverse above vanishes.

2.3. Likelihood ratio tests

The final method for constructing hypothesis tests that we will consider involves comparing the value of the maximised likelihood function at the restricted MLE ($\hat{\theta}^R$) and the unrestricted MLE (now written as $\hat{\theta}^U$). This likelihood ratio test statistic takes the form

$$S_L = 2\left(l(\hat{\theta}^U; Y) - l(\hat{\theta}^R; Y)\right)$$

and it can be shown that under H_0, $S_L \xrightarrow{d} \chi^2_{(j)}$.

2.4. Discussion

All three statistics have the same limiting distribution under the null hypothesis but they have different exact distributions and generally produce different, though frequently similar, values in any application. Which of the tests to use in any particular application is partly a matter of convenience. The Wald test does not require estimation of the restricted model and when that estimation would be difficult one might choose to use a Wald test. Conversely the score test does not require estimation of the unrestricted model and when that is difficult one might choose the score test. The likelihood ratio test requires calculation of both estimates but sometimes that is simple enough and the subsequent calculation is very simple.

An issue not given enough attention in practice is that in finite samples the Wald test is not parameterisation invariant. This means that, for example, testing $H_0 : \theta^\alpha = 0$ produces different answers depending on the value of $\alpha > 0$ that we use. Indeed by choosing a mad enough value of α it is possible in some cases to always reject the null hypothesis. This is rather bizarre as for any value of α the only value of θ consonant with the null hypothesis is zero, so tests using any positive value of α are essentially testing the same hypothesis. The score and likelihood ratio tests are parameterisation invariant in the sense that whatever one-to-one reparameterisation is adopted the same numerical values of the test statistics are produced. In this respect they are to be preferred.

3. Specification testing

Maximum likelihood estimation requires a complete specification of the probability distribution of the random variables whose realisations we observe[2]. In practice we do not *know* this distribution though we may be able to make a good guess. If our guess is badly wrong then we may produce poor quality estimates, for example badly biased estimates, and the inferences we draw using the properties of the likelihood function may be incorrect. In regression models the same sorts of problems occur. If there is homoskedasticity or serial correlation then, though we may produce reasonable[3] point estimates of regression coefficients if we ignore these features of the

[1] In this case there is said to be "parameter orthogonality".

[2] Note that the distribution of conditioning variables does not have to be specified.

[3] Though maybe not efficient.

data generating process, our inferences will usually be incorrect if these features are not allowed for, because we will use incorrect formulae for standard errors and so forth.

It is important then to seek for evidence of departure from a model specification, that is to conduct *specification tests*. In a likelihood context the score test provides an easy way of generating specification tests. One produces a generalisation of the specified model which does capture potential elements not picked up in the model as originally specified and which is a special case of that model when a subset of the parameters are set to zero. One then conducts a score test of the hypothesis that these additional parameters are zero. Note that we never need to estimate the more general model when conducting the score test. Further it turns out that many classes of generalisation of any given model lead to identical (score) specification tests. This is good because it means that we sometimes do not have to be absolutely specific about the potential failure of the originally specified model. On the other hand it is bad in the sense that on detecting model misspecification the (score) specification test does not tell us exactly how the model should be extended.

3.1. Detecting heteroskedasticity

We consider one example here, namely detecting heteroskedasticity in a normal linear regression model. In the model considered, $Y_1, \ldots, Y_n$ are independently distributed with Y_i given x_i being $N(x_i'\beta, \sigma^2 h(z_i'\alpha))$ where $h(0) = 1$ and $h'(0) = 1$, both achievable by suitable scaling of $h(\cdot)$. Let $\theta^U = [\beta, \sigma^2, \alpha]$ and let $\theta^R = [\beta, \sigma^2, 0]$. A score test of $H_0 : \alpha = 0$ will provide a specification test to detect heteroskedasticity.

The log likelihood function when $\alpha = 0$, in which case there is homoskedasticity, is as follows.

$$l(\theta^R; y|x) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i'\beta)^2$$

whose gradients with respect to β and σ^2 are

$$\begin{aligned} l_\beta(\theta^R; y|x) &= -\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - x_i'\beta)\,x_i \\ l_{\sigma^2}(\theta^R; y|x) &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - x_i'\beta)^2 \end{aligned}$$

which lead to the restricted MLEs under homoskedasticity, as follows.

$$\begin{aligned} \hat\beta &= (X'X)^{-1}X'y \\ \hat\sigma^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(y_i - x_i'\hat\beta\right)^2 \end{aligned}$$

The log likelihood function for the unrestricted model is

$$l(\theta^U; y|x) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2}\sum_{i=1}^{n}\log h(z_i'\alpha) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\frac{(y_i - x_i'\beta)^2}{h(z_i'\alpha)}$$

whose gradient with respect to α is

$$l_\alpha(\theta^U; y|x) = -\frac{1}{2}\sum_{i=1}^{n}\frac{h'(z_i'\alpha)}{h(z_i'\alpha)}z_i + \frac{1}{2\sigma^2}\sum_{i=1}^{n}\frac{(y_i - x_i'\beta)^2\, h'(z_i'\alpha)}{h(z_i'\alpha)^2}z_i$$

which evaluated at the restricted MLE (for which $\alpha = 0$) is

$$\begin{aligned} l_\alpha(\hat{\theta}^R; y|x) &= -\frac{1}{2}\sum_{i=1}^{n} z_i + \frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}\left(y_i - x_i'\hat{\beta}\right)^2 z_i \\ &= \frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}\left(\hat{\varepsilon}_i^2 - \hat{\sigma}^2\right) z_i. \end{aligned}$$

The specification test examines the correlation between the squared OLS residuals and z_i. The score test will lead to rejection when this correlation is large. Details of calculation of this test are given in the intermediate textbooks and the test (Breusch-Pagan-Godfrey) is built into many of the econometric software packages. Note that the form of the function $h(\cdot)$ does not figure in the score test. This would not be the case had we developed either a Wald test or a Likelihood Ratio test.

3.2. Information Matrix tests

When the complete probability distribution of outcomes given conditioning variables is specified maximum likelihood estimation is usually feasible. We have seen that the results on the limiting distribution of the MLE rest at one point on the Information Matrix Equality

$$E[l_\theta(\theta_0, Y)l_\theta(\theta_0, Y)'] = -E[l_{\theta\theta'}(\theta_0, Y)]$$

where $Y = (Y_1, \ldots, Y_n)$ are n random variables whose realisations constitute our data.

In the case relevant to much microeconometric work the log likelihood function is a sum of independently distributed random variables, e.g. in the continuous Y case

$$l(\theta, Y) = \sum_{i=1}^{n}\log f(Y_i, \theta),$$

where $f(Y_i, \theta)$ is the probability density function of Y_i. Here the Information Matrix Equality derives from the result

$$E[\frac{\partial}{\partial\theta}\log f(Y,\theta)\frac{\partial}{\partial\theta'}\log f(Y,\theta) + \frac{\partial^2}{\partial\theta\partial\theta'}\log f(Y,\theta)] = 0.$$

Given a value $\hat{\theta}$ of the MLE we can calculate a sample analogue of the left hand side of this equation

$$IM = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial}{\partial\theta}\log f(Y_i,\theta)\frac{\partial}{\partial\theta'}\log f(Y_i,\theta) + \frac{\partial^2}{\partial\theta\partial\theta'}\log f(Y_i,\theta)|_{\theta=\hat{\theta}}\right)$$

and, if the likelihood function is a correct specification for the data generating process, then we expect the resulting statistic (which is a matrix of values unless θi is scalar) to be close to (a matrix of zeros). A general purpose statistic for detecting incorrect specification of a likelihood function is produced by considering a quadratic form in a vectorised version of all or part of $n^{1/2}IM$. This Information Matrix Test statistic was introduced by Halbert White[4] in 1982.

4. (*) An application of maximum likelihood methods: censored data

Censored data is data which is incomplete in the sense that some values are not revealed. We always consider cases in which there is an unambiguous revelation rule. Here are some examples.

Consider an investigation of the influences on the speed of return to work of unemployed workers. We might study this by developing a model, based in job search theory for example, and then try to estimate parameters of the model using information from a survey of unemployed workers. One way in which such a survey might be conducted involves sampling entrants to unemployment, observing how long it takes them to return to work, also recording features of the workers (e.g. marital status, educational attainment, wage in their previous job) and of the labour market in which they operate. Since unemployment durations can be very long it is possible that, by the time such a survey is terminated, some workers will not have returned to work. For these workers we will not know their unemployment duration, all we will know is that it is longer than the duration observed at the time the study was terminated. This is an example of *censored* data. In this case we talk of "*right censored*" data because high (to the right on a horizontal axis) values are not revealed to us.

Left censoring arises when low values of a response are not revealed to us. Many years ago James Tobin developed a simple model for household expenditure on cars. He proposed that the expenditure data revealed in surveys, which contains many zero expenditures over typical recording periods (e.g. a year), could be modelled as a left censored value of a normally distributed variable with a mean which depends upon household characteristics such as income. Specifically he modelled observed expenditures as if they were realisations of a random variable Z, where conditional on covariate values, x, there is a latent variable $Z^* \sim N(x'\beta, \sigma^2)$, with $Z = Z^*$ when $Z^* > 0$, $Z = 0$ otherwise. This model is often referred to as a *Tobit model*.

Here the censoring occurs as an essential part of the modelling process rather than as a part of the observation process. We might question the appropriateness of the Tobin model, perhaps because the appearance or otherwise of zero expenditures may be a consequence of the necessary discreteness of purchasing.

Another example arises in labour supply studies in which, we observe wages only for people who are working. In the context of a simple model of the decision to work such people will be those for whom the wage rate exceeds the value of leisure time at zero hours of work. Here we will have left censored wage rate data. If we were to study hours worked data we would find these only available for working people. In this case we can think of the hours data as censored according to a revelation rule that involves variables other than hours.

[4] See "Maximum Likelihood Estimation in Misspecified Models", Halbert White, *Econometrica* 1982. In "Testing for Neglected Heterogeneity", Andrew Chesher, *Econometrica*, 1984, I show that in a wide class of problems the Information Matrix test is in fact a test of the hypothesis that the "parameter" vector, θ, is constant across observations (individuals).

Frequently in surveys we find responses which tell us a range in which some value lies (e.g. that household annual income is in the range £20,000 - £30,000, etc.). There were discussions recently involving a major Government survey in which it was argued that there should be a move towards recording income data in this way and away from asking for an "exact" income value, on the grounds that non-response to the sensitive income question would be reduced. Data like this are usually called "*grouped*" data but sometimes they are called "*interval censored*".

4.1. Maximum likelihood methods for censored data

Here we tackle estimation in censored data models using maximum likelihood methods. Recall that the likelihood function is a function of parameters and data values which is proportional to[5] the probability of observing the data values at the values specified for the parameters. The maximum likelihood estimator is a value for the parameters, a function of the data, that maximises the likelihood function. We now derive this probability for some censored data cases.

First consider *left censoring* and suppose that in the absence of censoring we observe realisations of random variables Y_i which, conditional on covariates, x_i have probability density functions $f(y_i|x_i;\theta)$. Suppose left censoring occurs at a value c_i so that if a realisation of Y_i is less than c_i then its value is not observed.

Define binary random variables D_i such that $D_i = 0$ if $Y_i \leq c_i$ and $D_i = 1$ if $Y_i > c_i$. The probability mass function for these binary variables is

$$P[D_i = d_i|x_i] = F(c_i|x_i;\theta)^{1-d_i}\left(1 - F(c_i|x_i;\theta)\right)^{d_i} \tag{4.1}$$

where

$$F(c_i|x_i;\theta) = P[Y_i \leq c_i|x_i] = \int_{-\infty}^{c_i} f(y_i|x_i;\theta)dy_i \tag{4.2}$$

is the distribution function of Y_i evaluated at c_i. This would be the basis for constructing a likelihood function which only employed information on whether or not data were censored.

Values of Y_i are only observed when $D_i = 1$. The conditional probability density function of Y_i given $D_i = 1$ (and x_i) is the *truncated* density function

$$g(y_i|x_i, D_i = 1) = \frac{f(y_i|x_i;\theta)}{\left(1 - F(c_i|x_i;\theta)\right)}, \qquad y_i > c_i.$$

Note that this is a proper density function, integrating to one over the range of Y_i. This would be the basis for constructing a likelihood function which only employed information contained in non-censored data. Such data is sometimes called *truncated data*.

The joint probability density - probability mass function for the censoring indicators and the revealed values, y_i^r of Y_i is the product of (4.1) and (4.2), namely

$$\begin{aligned} h(d_i, y_i^r|x_i;\theta) &= F(c_i|x_i;\theta)^{1-d_i}\left(1 - F(c_i|x_i;\theta)\right)^{d_i}\left(\frac{f(y_i^r|x_i;\theta)}{\left(1 - F(c_i|x_i;\theta)\right)}\right)^{d_i} \\ &= F(c_i|x_i;\theta)^{1-d_i} f(y_i^r|x_i;\theta)^{d_i}. \end{aligned}$$

[5] Equal to, in the case of discrete data.

The log likelihood function is the sum across i (from 1 to n, the sample size) of the logarithm of this expression, as follows.

$$l(\theta; d, y^r|x) = \sum_{i=1}^{n} d_i \log f(y_i^r|x_i;\theta) + (1-d_i)\log F(c_i|x_i;\theta)$$

Check for yourself that, with right censoring and with $D_i = 1$ when $Y_i \leq c_i$ in which case y_i^r is revealed, and equal to zero otherwise, the log likelihood function is

$$l(\theta; d, y^r|x) = \sum_{i=1}^{n} d_i \log f(y_i^r|x_i;\theta) + (1-d_i)\log\left(1 - F(c_i|x_i;\theta)\right).$$

A case that arises frequently in econometric practice has the continuous random variables Y_i *normally* distributed with mean $x_i'\beta$ and variance σ^2. With left censoring and with $c_i = 0$ for all realisations this gives the classical Tobit model for which the log likelihood function is as follows.

$$\begin{aligned} l(\theta; d, y^r|x) \quad = \quad & \sum_{i=1}^{n} d_i \left(-\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left(y_i^r - x_i'\beta\right)^2\right) \\ & + \sum_{i=1}^{n}(1-d_i)\log\Phi\left(-\frac{x_i'\beta}{\sigma}\right) \end{aligned}$$

Here $\Phi(\cdot)$ is the standard normal distribution function. Maximisation of the log likelihood function is done using numerical methods and is more straightforward if done with respect to $\gamma = \beta/\sigma$ and $\lambda = 1/\sigma$, in which parameterisation the log likelihood function is globally concave. By the parameterisation invariance of ML, the MLEs of β and σ are $\hat{\beta} = \hat{\gamma}/\hat{\lambda}$ and $\hat{\sigma} = 1/\hat{\lambda}$. Most modern econometrics software programmes have estimation of this and related models built in.

Finally consider the case in which there is grouping, or interval censoring. Suppose that all we know is which of M intervals realisations of Y_i falls in: $(c_i, c_{i+1}]$, $i = 1, \ldots M$, $c_{i+1} > c_i$. For random variables with unbounded support we will have one or both of $c_1 = -\infty$, $c_{M+1} = \infty$ holding.

Define M binary indicators, $D_i^m = 1$ if $c_i < Y_i \leq c_{i+1}$, equal to zero otherwise, $m = 1, \ldots, M$. The data consist of realisations of these M indicators. The probability that D_i^m equals one is

$$P\left[D_i = 1|x_i\right] = F(c_{i+1}|x_i;\theta) - F(c_i|x_i;\theta)$$

so the probability mass function for $D_i^1, \ldots D_i^M$ is

$$P\left[\bigcap_{m=1}^{M}\left(D_i^m = d_i^m\right)|x_i\right] = \prod_{m=1}^{M}\left(F(c_{i+1}|x_i;\theta) - F(c_i|x_i;\theta)\right)^{d_i^m}$$

and the log likelihood function for n realisations of these M binary indicators, i.e. for the interval censored data is

$$l(\theta; d|x) = \sum_{i=1}^{n}\sum_{m=1}^{M} d_i^m \log\left(F(c_{i+1}|x_i;\theta) - F(c_i|x_i;\theta)\right).$$

5. Bayesian methods[6]

With outcomes $Y_1, \ldots, Y_n$ the likelihood function, $L(\theta, y_{[n]})$, gives the probability[7] of the realised outcomes, $y_{[n]} = \{y_1, \ldots, y_n\}$, as a function of θ, the value of the parameter (vector) that determines the data generating process. The subscript on $y_{[n]}$ is now helpful in keeping track of the sample size.

In a Bayesian approach the value of this parameter vector is regarded as a *random variable* with a prior distribution, say $p(\theta)$. This is "prior" in the sense of prior to observing the realisations of the outcomes. A Bayesian analysis uses the information about θ contained in the *posterior* distribution of θ given the realised values of outcomes. The posterior distribution is obtained by applying Bayes Theorem.

5.1. Bayes Theorem

For events A and B, since:

$$P[A \cap B] = P[B|A]P[A] = P[A|B]P[B]$$

there is

$$P[A|B] = \frac{P[B|A]P[A]}{P[B]}$$

which is known as Bayes Theorem.

For continuous random variables U and V there is the similar relationship amongst conditional and marginal *density* functions

$$f_{U|V}(u|v) = \frac{f_{V|U}(v|u)f_U(u)}{f_V(v)} \tag{5.1}$$

which arises because there are the alternative iterated decompositions of the joint density function

$$f_{UV}(u, v) = f_{V|U}(v|u)f_U(u) = f_{U|V}(u|v)f_V(v).$$

The denominator in (5.1) is the marginal density of V and is just the definite integral of the numerator across the support of U.

$$f_V(v) = \int f_{V|U}(v|u)f_U(u)du$$

There are similar expressions when U and/or V is discrete. U and V can be vector random variables.

[6] There is a fine exposition in Tony Lancaster's recently published *An Introduction to Modern Bayesian Econometrics*, Basil Blackwell, Oxford, 2004.

[7] The density function in the continuous case.

5.2. Posterior distributions

Let $p(y_{[n]}|\theta)$ denote the likelihood function, that is $p(y_{[n]}|\theta) \equiv L(\theta, y_{[n]})$. The discussion proceeds for the moment in terms of continuous random variables. The Bayesian posterior distribution is the conditional density of the parameter θ given the realised values of the outcomes, that is $p(\theta|y_{[n]})$ where

$$p(\theta|y_{[n]}) = \frac{p(y_{[n]}|\theta)p(\theta)}{p(y_{[n]})}$$

and

$$p(y_{[n]}) = \int p(y_{[n]}|\theta)p(\theta)d\theta. \tag{5.2}$$

In the case of discrete outcomes the function $p(y_{[n]}|\theta)$ is the probability mass function:

$$p(y_{[n]}|\theta) = P[Y_1 = y_1 \cap \cdots \cap Y_n = y_n].$$

In a Bayesian analysis the posterior distribution may be used to make explicit probability statements about the value of θ conditional on the realised outcomes. the posterior probability that θ falls in some set A is

$$P[\theta \in A|y_{[n]}] = \int_{\theta \in A} \frac{p(y_{[n]}|\theta)p(\theta)}{p(y_{[n]})}d\theta \tag{5.3}$$

Of course the prior distribution can have significant influence on these statements. That can be good in a policy and decision making contexts. In academic discourse it is common to work with prior distributions which do not have great influence - so called uninformative prior distributions.

It may be difficult to derive an exact expression for the integral in (5.3) and also for the integral in (5.2) which defines the function $p(y_{[n]})$. In modern Bayesian inference these difficulties are overcome by using sampling approximations of various sorts. For example when $p(\theta|y_{[n]})$ can be computed (i.e. there is no great difficulty in calculating $p(y_{[n]})$) one way to proceed is to draw a sample of values of θ from the posterior distribution, $p(\theta|y_{[n]})$ and calculate the proportion of sampled values that fall in the set A.

Moments of the posterior distributions, for example expected values and variances, can provide useful summaries of posterior distributions. Again the integrals may be difficult to calculate but means and variances of sampled values can provide good approximations. In cases in which the prior distribution is very slowly varying with θ the *mode* of the posterior distribution will be close to the maximum likelihood estimator. In many cases it is possible to show that in large samples Bayesian posterior distributions are approximately multivariate normal providing a link between large sample Bayesian and ML theory. In large samples with relatively uninformative prior distributions Bayesian estimators calculated as expected values of posterior distributions will be close to ML estimators. Cases in which this correspondence does not prevail are interesting and may be cases in which Bayesian methods produce.

5.3. Bayesian updating

The discussion proceeds in terms of densities, as if the outcomes are continuous. An additional observation is available, a realisation y_{n+1} of Y_{n+1}. There is

$$p(y_{[n+1]}|\theta) = p(y_{n+1}|y_{[n]}, \theta) \times p(y_{[n]}|\theta)$$

and so

$$p(\theta|y_{[n+1]}) \propto p(y_{n+1}|y_{[n]},\theta)p(\theta|y_{[n]}).$$

As each new realisation arrives a new posterior distribution, $p(\theta|y_{[n+1]})$, is obtained by updating the previous posterior distribution, $p(\theta|y_{[n]})$, using the probability law for the random variable whose realisation has arrived, allowing for any dependence on previously observed random variables. If outcomes are independent (given θ) then $p(y_{n+1}|y_{[n]},\theta) = p(y_{n+1}|\theta)$. For the purpose of processing the $(n+1)$th realisation, y_{n+1}, the prior distribution for θ is the posterior distribution for θafter n realisations, $p(\theta|y_{[n]})$.

5.4. Predictive distributions

We have n realisations and want to make probability statements about the values that the $(n+1)$th realisation, y_{n+1}, may take. Its density given the random variables for which there are already the realisations $y_{[n]}$, and θ is $p(y_{n+1}|y_{[n]},\theta)$. If outcomes are independent given θ this simplifies to $p(y_{n+1}|\theta)$.

For this purpose one can use the predictive distribution defined as:

$$p(y_{n+1}|y_{[n]}) = \int p(y_{n+1},\theta|y_{[n]})d\theta = \int p(y_{n+1}|y_{[n]},\theta)p(\theta|y_{[n]})d\theta \propto \int p(y_{n+1}|y_{[n]},\theta)p(y_{[n]}|\theta)p(\theta)d\theta.$$

This captures prior uncertainty about θ and the updates on this provided by the information in $y_{[n]}$. The predictive distribution can be summarised in a variety of ways. One could report its mean, median or mode, or report quantiles in the style of the charts (not Bayesian) issued recently by the UK Monetary Policy Committee.

5.5. Examples

5.5.1. Identically distributed binary outcomes.

$Y_1,\ldots,Y_n$ are identically and independently distributed binary random variables and $y = \{y_1,\ldots,y_n\}$ are realisations. For all i, $P[Y_i = y_i] = \theta^{y_i}(1-\theta)^{1-y_i}$ and of course $y_i \in \{0,1\}$. The likelihood function is

$$p(y|\theta) = \theta^{S_n}(1-\theta)^{n-S_n}$$

where $S_n^0 \equiv \Sigma_{i=1}^n y_i$ and $S_n^1 \equiv n - \Sigma_{i=1}^n y_i$ are the numbers of 0's and 1's amongst the realised outcomes. The posterior distribution of θ is

$$p(\theta|y) = \theta^{S_n^0}(1-\theta)^{S_n^1}p(\theta)$$

where $p(\theta)$ is the prior distribution one chooses to employ. A common and convenient choice is the Beta distribution in which[8]

$$p(\theta) \propto \theta^a(1-\theta)^b$$

with $\theta \in [0,1]$, for which the posterior distribution is

$$p(\theta|y) \propto \theta^{a+S_n^0}(1-\theta)^{b+S_n^1}p(\theta).$$

Choosing $a = b = 0$ gives a uniform prior for θ.[9] Then the likelihood function and posterior density function are very similar functions but with very different interpretations.

[8] "$\propto$" indicates "is proportionate to".

[9] Is that uninformative about θ?

5.5.2. The probit model

Now for each independently distributed Y_i there is a vector of covariates x_i, and the data generating process is supposed to be such that

$$P[Y_i = y_i|x_i] = \Phi(x_i'\theta)^{y_i}(1 - \Phi(x_i'\theta))^{1-y_i}.$$

The posterior density of the vector of parameters θ is

$$p(\theta|y, x) \propto \prod_{i=1}^{n} \Phi(x_i'\theta)^{y_i}(1 - \Phi(x_i'\theta))^{1-y_i} p(\theta)$$

where $p(\theta)$ is the prior distribution one chooses to employ. Now there are conditioning variables x the posterior density is conditional on their values as well as on the values of the realised outcomes. Clearly whatever the form of the prior distribution integration with respect to θ cannot be done explicitly, for example to find the value of $p(y|x)$ or the value of a probability as in (5.3). Here as in most cases arising in microeconometric practice sampling based approximations are the route to practical Bayesian inference.

5.5.3. Exponentially distributed outcomes

$Y_1, \ldots, Y_n$ are identically and independently distributed *continuous* positive random variables and $y = \{y_1, \ldots, y_n\}$ are realisations. For all i the density function of Y_i is

$$f(y|\theta) = \theta \exp(-\theta y), \qquad \theta, y > 0.$$

The posterior density of θ is

$$p(\theta|y) \propto \theta^n \exp(-\theta \Sigma_{i=1}^{n} y_i) p(\theta)$$

and if $p(\theta)$ is chosen to be a gamma distribution with

$$p(\theta) \propto \theta^a \exp(-b\theta)$$

there is obviously simplification and the posterior density is in the gamma distribution family too. This is an example of a so called *natural conjugate prior* which are such that posterior densities are in the same family of distributions as the prior density.

5.6. Issues and Challenges

Some of the difficulties that arise in classical (Fisherian) inference do not arise in a Bayesian approach. Given a prior distribution and a parametric specification of the data generating process one can proceed to exact inference based on the posterior distribution which carries a wealth of information and can be summarised in many ways honed to the purpose at hand. The predictive distribution is a powerful tool for forecasting and for detecting misspecification of the probability law of the data generating process. Posterior and predictive distributions are just what is required in a decision theoretic attack on policy determination.

Computational issues posed a huge barrier to use of Bayesian methods until high speed simulation and sampling became feasible. Now Bayesian methods can be almost routinely employed in problems of realistic scale but one rarely sees them used. Why?

1. Some people have a philosophical difficulty with the notion that parameters can be regarded as random variables. Some see an "objective" random variation in outcomes and accept the idea of a "subjective" random variation in parameters but have difficulty with the blending of these two types of randomness that a Bayesian analysis entails. When there is a real decision to be made and money to be gained or lives to be lost these philosophical difficulties can sometimes be put to one side.

2. There is the problem of constructing a prior distribution. This is difficult when there is a high dimensional parameter. It is easier if one is prepared to use an uninformative prior but even then some care must be taken to ensure that specious information is not inadvertently introduced. For academic discourse one will often not want to bring any prior information to the problem.

3. One must be quite specific about details of the data generating process. In these brief introductory notes I have concentrated on cases in which there is a *parametric* specification of the data generating process - that is a parametric specification of the likelihood function. Much modern econometrics tries to get by with less than this although in professional practice parametric likelihood based analysis is quite common. Identification is frequently obtained through moment conditions without specification of the precise families within which distributions of outcomes lie. Inference requires one to be more specific about the distributions involved but not very specific. Some of the most exciting challenges in Bayesian research lie in trying to find ways of doing Bayesian analysis without strong parametric restrictions. Empirical likelihood seems to be a good place to start.[10]

[10] See Art Owen's *Empirical Likelihood*, Chapman and Hall and CRC Press, 2001.