

# LIKELIHOOD METHODS IN FINANCIAL ECONOMETRICS

Three-Day Course

Economic Research Southern Africa  
University of Stellenbosch, April 2009

## COURSE OVERVIEW

This is a 3-day course on likelihood methods in financial econometrics at the intermediate level. The main objective of the course is to provide a thorough grounding in the maximum likelihood principle, encompassing specification, estimation and testing. On each day the morning session will focus on the theory and the afternoon session will be a practical illustration of the methods in action. A full set of lecture notes and exercises are available. The computer software used is MATLAB. It is not assumed that participants are familiar with MATLAB as all the relevant code will be provided in an attempt to promote learning-by-doing.

### **Stan Hurn**

School of Economics and Finance, Queensland University of Technology,  
<http://www.bus.qut.edu.au/stanhurn/>

and

National Centre for Econometric Research  
<http://www.ncer.edu.au/>

**Email:** [s.hurn@qut.edu.au](mailto:s.hurn@qut.edu.au)

# Contents

<b>1</b>	<b>THE MAXIMUM LIKELIHOOD PRINCIPLE</b>	<b>4</b>
1.1	Introduction . . . . .	4
1.2	Motivating Examples . . . . .	4
1.3	Joint Probability Distributions . . . . .	8
1.4	Maximum Likelihood Framework . . . . .	9
1.4.1	The Likelihood and Log-likelihood Functions . . . . .	9
1.4.2	Gradient . . . . .	11
1.4.3	Hessian . . . . .	11
1.5	Moments of the Gradient . . . . .	12
1.6	Asymptotic Properties . . . . .	16
1.6.1	Consistency . . . . .	16
1.6.2	Asymptotic Normality . . . . .	18
1.6.3	Asymptotic Efficiency . . . . .	20
1.7	A Very Brief Introduction to Testing . . . . .	21
1.7.1	Likelihood Ratio Test . . . . .	23
1.7.2	Wald Test . . . . .	24
1.7.3	Lagrange Multiplier Test . . . . .	24
1.8	Computer Applications . . . . .	25
1.8.1	Consistency . . . . .	25
1.8.2	Exponential Distribution . . . . .	26
1.8.3	The Classical Regression Model . . . . .	29
<b>2</b>	<b>ESTIMATING NONLINEAR MODELS</b>	<b>33</b>
2.1	Motivating Examples . . . . .	33
2.2	Newton Methods . . . . .	36
2.2.1	Newton-Raphson . . . . .	37
2.2.2	Method of Scoring . . . . .	38
2.2.3	BHHH Algorithm . . . . .	38
2.3	Quasi-Newton Methods . . . . .	40
2.4	Line Searching . . . . .	41
2.5	Simplex algorithm . . . . .	41
2.6	Choice of Algorithm . . . . .	43
2.7	Computing Standard Errors . . . . .	44
2.8	Parameter Constraints . . . . .	47
2.9	Maximum Likelihood Estimation of Nonlinear Regression Models . . . . .	49

2.10	Computer Applications . . . . .	57
2.10.1	Robust Estimation of the CAPM . . . . .	57
2.10.2	Nonnested Test US Money Demand . . . . .	60
2.10.3	A GARCH(1,1) Model of US Yields . . . . .	62
<b>3</b>	<b>QUASI MAXIMUM LIKELIHOOD ESTIMATION</b>	<b>65</b>
3.1	Introduction . . . . .	65
3.2	Motivating Examples . . . . .	65
3.3	The Information Equality . . . . .	71
3.4	Variance of the Quasi-Maximum Likelihood Estimator . . . . .	75
3.5	Misspecification of Regression Models . . . . .	80
3.5.1	White Variance Estimator . . . . .	81
3.5.2	Newey-West Variance Estimator . . . . .	85
3.6	Testing . . . . .	88
3.7	Computer Applications . . . . .	92
3.7.1	The Effects of Misspecifying a Distribution . . . . .	92
3.7.2	A Model of US Investment . . . . .	93
3.7.3	Conditional Nonnormality and QMLE in Volatility Models . . . . .	94

# 1 THE MAXIMUM LIKELIHOOD PRINCIPLE

## 1.1 Introduction

Maximum likelihood estimation is a general method for estimating the parameters of an econometric model. Intuitively, the principle of maximum likelihood estimation may be understood as follows. Consider an observed random variable  $y_t$  with probability density function  $f(y_1, \dots, y_T; \theta)$  where the form of  $f$  is known, but the parameter vector  $\theta$  is not known. The principle of maximum likelihood advocates choosing the values of the parameters that yield the greatest probability of giving rise to the observed sample of data.

The principle of maximum likelihood plays a central role in the exposition of this book, as a number of estimators used in econometrics can be derived within the maximum likelihood framework. In deriving the maximum likelihood estimator of a particular model, the following conditions must be satisfied.

1. The distribution of the observed random variable,  $y_t$ , must be known.
2. The likelihood function must be tractable in the sense that it can be evaluated for all admissible values  $\theta$ .

## 1.2 Motivating Examples

Maximum likelihood estimation is perhaps best illustrated with the following series of simple examples.

### **EXAMPLE: One observation from a Normal Distribution**

Suppose that you have a continuous random variable,  $x$ , with unknown mean  $\mu$  and standard deviation of unity. Suppose also that you are able to assume that it has a normal distribution. Suppose finally that you have two alternative hypotheses,  $\mu = \mu_0$  and  $\mu = \mu_1$  and one observation  $x_1$ . Which hypothesis do you select? The maximum likelihood principle says that you should choose the hypothesis that gives  $x_1$  the highest probability of occurring. Since  $x$  is continuous the probability of any particular value is infinitesimal. Instead we compare the probability density at  $x_1$  under the two hypotheses, as given by the height of the probability density function.

Now consider all possible values for  $\mu$  and choose that giving  $x_1$  the highest probability density. The density function of  $x$  given  $\mu$  is

$$f(x|\mu) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x-\mu)^2\right)$$

What value of  $\mu$  maximises the density given the observation  $x_1$ ? The answer is obvious... $\mu = x_1$ , because then the distribution will be centered over  $x_1$ , and the density is highest at the centre of the distribution.

This can be shown slightly more formally as follows. Note that  $x_1$  is given and  $\mu$  is being treated as variable, hence we can view the density function as a function of  $\mu$  with  $x_1$  given. Hence

$$L(\mu|x_1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_1 - \mu)^2\right)$$

where  $L(\mu|x_1)$  is now called the Likelihood Function.

Recall the likelihood function

$$L(\mu|x_1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_1 - \mu)^2\right)$$

It transpires that it is more convenient to work with the log-likelihood function:

$$\log L(\mu|x_1) = -\log \sqrt{2\pi} - \frac{1}{2}(x_1 - \mu)^2$$

Note also that  $\log L$  will have a maximum for the same value of  $\mu$  that maximises  $L$ . Differentiating with respect to  $\mu$  and setting the result equal to zero gives us the maximum likelihood estimator for  $\mu$

$$\frac{\partial \log L}{\partial \mu} = 0 \implies x_1 - \mu = 0$$

### **EXAMPLE: Two observations from a Normal Distribution**

Suppose now that we have two independent observations  $x_1$  and  $x_2$  and we wish to estimate  $\mu$ . The ML procedure is to find the value of  $\mu$  that maximises their joint probability density. The assumption of independence is crucial as we can then write

$$P(x_1, x_2|\mu) = P(x_1|\mu) \cdot P(x_2|\mu)$$

ie. the joint probability density is given by the product of the individual densities

$$f(x_1, x_2 | \mu) = \left[ \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_1 - \mu)^2\right) \right] \left[ \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_2 - \mu)^2\right) \right]$$

This can again be reinterpreted as the likelihood function for  $\mu$  given  $x_1$  and  $x_2$  and we can maximise it indirectly by maximising the log-likelihood function

$$\begin{aligned} \log L(\mu) &= -2 \log \sqrt{2\pi} - \frac{1}{2}(x_1 - \mu)^2 - \frac{1}{2}(x_2 - \mu)^2 \\ \frac{\partial \log L}{\partial \mu} &= 0 \\ &\implies (x_1 - \mu) + (x_2 - \mu) = 0 \\ \mu &= \frac{1}{2}(x_1 + x_2) \end{aligned}$$

The ML estimator of the mean is the sample mean!

### EXAMPLE: The normal distribution

Let us return to the normal distribution and estimate the mean,  $\mu$ , and variance  $\sigma^2$ . Recall the probability density function

$$f(Y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(Y - \mu)^2\right\}$$

Since  $\mu$  and  $\sigma^2$  specify the normal distribution in full we have two parameters to estimate.

$$\log L(\mu, \sigma^2) = -\frac{1}{2} \log 2\pi - \frac{1}{2} \log \sigma^2 - \frac{1}{2\sigma^2}(Y - \mu)^2$$

which for a sample of size  $n$  becomes

$$\log L(\mu, \sigma^2 | Y_1 \dots Y_n) = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - \mu)^2$$

As before we need to maximise  $\log L(\mu, \sigma^2)$  we differentiate partially with respect to  $\mu$  and  $\sigma^2$  and set the resultant derivatives to zero.

$$\begin{aligned} \frac{\partial \log L}{\partial \mu} &= \frac{1}{\sigma^2} \sum_{i=1}^n (Y_i - \mu) = 0 \\ \frac{\partial \log L}{\partial (\sigma^2)} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (Y_i - \mu)^2 = 0 \end{aligned}$$

For non-zero  $\sigma^2$  we have from the first condition

$$\sum_{i=1}^n (Y_i - \mu) = 0 \implies \sum_{i=1}^n Y_i - n\mu = 0 \implies \hat{\mu} = \sum_{i=1}^n \frac{Y_i}{n}$$

To obtain the MLE of  $\sigma^2$  multiply the second first-order condition by  $2\sigma^4$  to yield

$$-n\sigma^2 + \sum_{i=1}^n (Y_i - \mu)^2 = 0 \implies \hat{\sigma}^2 = \sum_{i=1}^n \frac{(Y_i - \mu)^2}{n}$$

Clearly  $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \mu)^2$  is a biased estimator of the variance  $\sigma^2$ . Hence here is a simple example of an MLE that does not possess the small sample property of unbiasedness.

Note that we are not limited to the assumption of normality. The major requirement for maximum likelihood estimation is that the distribution of the observed random variable is known. We can therefore estimate the parameters of any known distribution by maximum likelihood from a sample of independent observations.

### Example: The Poisson distribution

Consider

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

such a distribution might describe the frequency of calls per hour at a telephone exchange where the single parameter  $\lambda$  is to be interpreted as the average call rate per hour.

Suppose we have a sample  $Y_1 \dots Y_n$ , which is drawn from a population characterised by the single parameter  $\lambda$ . We require the MLE of  $\lambda$  (the average call rate per hour)

$$\begin{aligned} L(\lambda) &= P(\lambda|Y_1) \cdot P(\lambda|Y_2) \dots P(\lambda|Y_n) \\ &= \left[ \frac{\lambda^{Y_1} e^{-\lambda}}{Y_1!} \right] \left[ \frac{\lambda^{Y_2} e^{-\lambda}}{Y_2!} \right] \dots \left[ \frac{\lambda^{Y_n} e^{-\lambda}}{Y_n!} \right] \end{aligned}$$

The log-likelihood function is

$$\begin{aligned} \log L(\lambda) &= [Y_1 \log \lambda - \lambda - \log Y_1!] + [Y_2 \log \lambda - \lambda - \log Y_2!] + \dots + [Y_n \log \lambda - \lambda - \log Y_n!] \\ &= (\log \lambda) \sum_{i=1}^n Y_i - n\lambda - \sum_{i=1}^n \log Y_i! \end{aligned}$$

now

$$\begin{aligned} \frac{\partial \log L(\lambda)}{\partial \lambda} &= \frac{1}{\lambda} \sum_{i=1}^n Y_i - n = 0 \\ \implies \hat{\lambda} &= \frac{\sum_{i=1}^n Y_i}{n} \end{aligned}$$

Thus the MLE of  $\lambda$ , the population average call rate is simply the average of the sample values.

### 1.3 Joint Probability Distributions

The previous set of examples emphasized that observed time series can be viewed as draws from probability distributions. These concepts are now formalized in more detail. Let  $\{y_1, y_2, \dots, y_T\}$ , represent a set of  $T$  random variables conditional on  $\{\theta_1, \theta_2, \dots, \theta_T\}$  parameters. The joint probability density function (pdf) is given by

$$f(y_1, y_2, \dots, y_T; \theta_1, \theta_2, \dots, \theta_T). \quad (1)$$

The parameters at each point in time,  $\theta_t$ , in general represent a vector of parameters. The time subscript on  $\theta$  arises when the distribution at each point in time is a function of variables that are time-varying. For example, in regression examples, the mean is represented as  $\theta_t = \beta x_t$  which is time-varying as  $x_t$  is time-varying.

The joint pdf can be written as a sequence of conditional distributions by assuming that  $y_1$  is the initial value with marginal density  $f(y_1)$ , and the joint distributions can be constructed using standard rules of probability as

$$\begin{aligned} f(y_1, y_2; \theta_1, \theta_2) &= f(y_2|y_1; \theta_2)f(y_1; \theta_1) \\ f(y_1, y_2, y_3; \theta_1, \theta_2, \theta_3) &= f(y_3|y_2, y_1; \theta_3)f(y_2|y_1; \theta_2)f(y_1; \theta_1), \end{aligned}$$

and hence for the full sample

$$f(y_1, y_2, \dots, y_T; \theta_1, \theta_2, \dots, \theta_T) = f(y_1; \theta_1) \prod_{t=2}^T f(y_t|y_{t-1}, y_{t-2}, \dots, y_1; \theta_t).$$

Three important special cases which simplify the pdf are as follows.

1. Independent and non-identically distributed

$$f(y_1, y_2, \dots, y_T; \theta_1, \theta_2, \dots, \theta_T) = \prod_{t=1}^T f(y_t; \theta_t).$$

2. Dependent and identically distributed

$$f(y_1, y_2, \dots, y_T; \theta_1, \theta_2, \dots, \theta_T) = \prod_{t=1}^T f(y_t|y_{t-1}, y_{t-2}, \dots, y_1; \theta).$$

3. Independent and identically distributed (*iid*)

$$f(y_1, y_2, \dots, y_T; \theta_1, \theta_2, \dots, \theta_T) = \prod_{t=1}^T f(y_t; \theta),$$

where  $\theta_1 = \theta_2 = \dots = \theta_T = \theta$ , is a vector of parameters that is constant over time.

A time series of data is then simply the observed realization of repeated draws from the joint pdf. For example, in the case where the probability distribution is assumed to be *iid*, all the realized draws (observations), come from the same distribution with the same parameters  $\theta$ . The maximum likelihood principle makes use of this result by providing a general framework for estimating the parameters  $\theta$  from observed time-series data.

## 1.4 Maximum Likelihood Framework

In this section the likelihood and log-likelihood functions are introduced and the maximum likelihood estimators derived. It is also necessary to discuss the first and second derivatives of the log-likelihood function.

### 1.4.1 The Likelihood and Log-likelihood Functions

The starting point for deriving the maximum likelihood estimator is the joint pdf in equation (1). The standard interpretation of the pdf in (1) is that  $f$  is interpreted as a function of  $y_t$  for given parameters,  $\theta$ . In defining the maximum likelihood estimators this interpretation is reversed, so that  $f$  is taken as a function of  $\theta$  for given  $y_t$ . The motivation behind this change in the interpretation of the arguments of the pdf is to regard  $\{y_1, y_2, \dots, y_T\}$  as a realized data set which is no longer random. The maximum likelihood estimator is then obtained by finding the value of  $\theta$  which is “most likely” to have generated the observed data. Here the phrase “most likely” is loosely interpreted in a probability sense. To highlight this change in the status of the arguments of the pdf, the likelihood is expressed as

$$L(\theta) = f(y_1, y_2, \dots, y_T; \theta_1, \theta_2, \dots, \theta_T). \quad (2)$$

It is important to remember that the likelihood function is simply a redefinition of the joint pdf. The maximum likelihood estimate of  $\theta$  is then defined as that value of  $\theta$ , denoted  $\hat{\theta}$ , that maximizes the likelihood function in equation (2). In a large number of cases, this may be achieved using standard calculus. The next session discusses numerical approaches to the problem of finding maximum likelihood estimates when no analytical solutions exist, or are difficult to derive.

For many of the problems encountered in econometrics, it is often simpler to work with

the log-likelihood function

$$\ln L(\theta) = \ln f(y_1, y_2, \dots, y_T; \theta_1, \theta_2, \dots, \theta_T), \quad (3)$$

or the average log-likelihood function

$$A(\theta) = \frac{\ln L(\theta)}{T} = \frac{1}{T} \ln f(y_1, y_2, \dots, y_T; \theta_1, \theta_2, \dots, \theta_T). \quad (4)$$

In both cases, the maxima are the same as the maxima of (2), as the natural logarithm function is monotonic in  $\theta$ .

The most general form for the likelihood function is where the variable is dependent and non-identically distributed. The log-likelihood function for this case is

$$\ln L(\theta) = \sum_{t=1}^T \ln f(y_t | y_{t-1}, y_{t-2}, \dots, y_1; \theta_t) + \ln f(y_0; \theta_0), \quad (5)$$

whereas the average log-likelihood function is

$$A(\theta) = \frac{\sum_{t=1}^T \ln f(y_t | y_{t-1}, y_{t-2}, \dots, y_1; \theta_t)}{T} + \frac{\ln f(y_0; \theta_0)}{T}. \quad (6)$$

For large  $T$ , the contribution of the last term to the average log-likelihood is insignificant and may be disregarded.

The log-likelihood functions of the three special cases discussed in the previous section are:

1. Independent and non-identically distributed

$$\ln L(\theta) = \sum_{t=1}^T \ln f(y_t; \theta_t). \quad (7)$$

2. Dependent and identically distributed

$$\ln L(\theta) = \sum_{t=1}^T \ln f(y_t | y_{t-1}, y_{t-2}, \dots, y_1; \theta). \quad (8)$$

3. Independent and identically distributed (*iid*)

$$\ln L(\theta) = \sum_{t=1}^T \ln f(y_t; \theta). \quad (9)$$

As the aim of maximum likelihood estimation is to find the value of  $\theta$  that maximizes either the log-likelihood function or the average log-likelihood function, a natural way to do this is to use the rules of calculus. This involves computing the first derivatives (gradient) and second derivatives (Hessian) of the log-likelihood function with respect to the parameter vector  $\theta$ .

### 1.4.2 Gradient

The first derivative of the log-likelihood function with respect to the parameter vector  $\theta$ ,

$$G(\theta) = \frac{\partial \ln L(\theta)}{\partial \theta}, \quad (10)$$

is known as the gradient or the score. In the *iid* case, where  $\theta$  is a fixed  $(K \times 1)$  vector of parameters, the gradient is

$$\frac{\partial \ln L(\theta)}{\partial \theta} = \begin{bmatrix} \frac{\partial \ln L(\theta)}{\partial \theta_1} \\ \frac{\partial \ln L(\theta)}{\partial \theta_2} \\ \vdots \\ \frac{\partial \ln L(\theta)}{\partial \theta_K} \end{bmatrix}. \quad (11)$$

The maximum likelihood estimate of  $\theta$ , namely  $\hat{\theta}$ , is obtained by solving the set of first-order conditions for a maximum obtained by setting the gradient equal to zero. In other words,  $\hat{\theta}$  satisfies

$$G(\hat{\theta}) = \left. \frac{\partial \ln L(\theta)}{\partial \theta} \right|_{\theta=\hat{\theta}} = 0. \quad (12)$$

### 1.4.3 Hessian

The second derivative of the log-likelihood function with respect to the parameter vector  $\theta$  is known as the Hessian

$$H(\theta) = \frac{\partial^2 \ln L(\theta)}{\partial \theta \partial \theta'}. \quad (13)$$

Consider the first element of the gradient in equation (11),  $\partial \ln L(\theta)/\partial \theta_1$ . This entry is a function of all the elements in the parameter vector  $\theta$  and may therefore be differentiated with respect to each element. This gives

$$\frac{\partial^2 \ln L(\theta)/\partial \theta_1}{\partial \theta'} = \left[ \frac{\partial^2 \ln L(\theta)/\partial \theta_1}{\partial \theta_1} \quad \frac{\partial^2 \ln L(\theta)/\partial \theta_1}{\partial \theta_2} \quad \cdots \quad \frac{\partial^2 \ln L(\theta)/\partial \theta_1}{\partial \theta_K} \right],$$

or, in slightly simpler notation

$$\frac{\partial^2 \ln L(\theta)/\partial \theta_1}{\partial \theta'} = \left[ \frac{\partial^2 \ln L(\theta)}{\partial \theta_1 \partial \theta_1} \quad \frac{\partial^2 \ln L(\theta)}{\partial \theta_1 \partial \theta_2} \quad \cdots \quad \frac{\partial^2 \ln L(\theta)}{\partial \theta_1 \partial \theta_K} \right].$$

Repeating this operation for all elements in the gradient vector gives the symmetric, square matrix

$$\frac{\partial^2 \ln L(\theta)}{\partial \theta \partial \theta'} = \begin{bmatrix} \frac{\partial^2 \ln L(\theta)}{\partial \theta_1 \partial \theta_1} & \frac{\partial^2 \ln L(\theta)}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 \ln L(\theta)}{\partial \theta_1 \partial \theta_K} \\ \frac{\partial^2 \ln L(\theta)}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 \ln L(\theta)}{\partial \theta_2 \partial \theta_2} & \cdots & \frac{\partial^2 \ln L(\theta)}{\partial \theta_2 \partial \theta_K} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial^2 \ln L(\theta)}{\partial \theta_k \partial \theta_1} & \cdots & \cdots & \frac{\partial^2 \ln L(\theta)}{\partial \theta_k \partial \theta_K} \end{bmatrix}.$$

The Hessian plays two important roles in the maximum likelihood framework.

1. *Second-order condition for a maximum*

The Hessian is used to establish that a maximum for the log-likelihood function has been achieved. In a univariate optimization problem, the maximum of a function is obtained by solving the first order condition obtained by setting the gradient to zero and checking to see if the second derivative of the function at the optimum is negative. In the case of the maximum likelihood estimate, the requirement is that the Hessian matrix evaluated at  $\hat{\theta}$

$$H(\hat{\theta}) = \left. \frac{\partial^2 \ln L(\theta)}{\partial \theta \partial \theta'} \right|_{\theta=\hat{\theta}}, \quad (14)$$

is a negative definite matrix.<sup>1</sup>

2. *Computing the covariance matrix of the estimator*

The Hessian also plays a role in determining the precision of the maximum likelihood estimator. In the single parameter case, if the likelihood function is relatively flat (peaked) in the vicinity of the maximum likelihood estimate,  $\hat{\theta}$ , corresponding to a small (large) absolute value of the second derivative, then one would interpret the estimate of  $\theta$  to be relatively imprecise (precise).

## 1.5 Moments of the Gradient

The gradient is given by

$$G(\theta) = \frac{\partial \ln L}{\partial \theta}. \quad (15)$$

---

<sup>1</sup>A matrix is negative definite if and only if  $x'Hx < 0$  for all non-zero vectors  $x$ .

This function has two properties which are important in establishing the properties of maximum likelihood estimators.

### Mean of the Gradient

The first property is

$$E[G(\theta)] = 0. \quad (16)$$

Consider the likelihood function

$$L(\theta) = f(y_1, y_2, \dots, y_T; \theta).$$

It follows from the properties of the joint density function  $f(y_1, y_2, \dots, y_T; \theta)$  that

$$\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} L(\theta) dy_1 dy_2 \dots dy_T = 1.$$

Now differentiating both sides with respect to  $\theta$  gives

$$\frac{\partial}{\partial \theta} \left( \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} L(\theta) dy_1 dy_2 \dots dy_T \right) = 0.$$

Using the interchangeability of differentiation and integration regularity condition and the property of natural logarithms that

$$\frac{\partial L(\theta)}{\partial \theta} = \frac{\partial \ln L(\theta)}{\partial \theta} L(\theta)$$

gives

$$\begin{aligned} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \frac{\partial L(\theta)}{\partial \theta} dy_1 dy_2 \dots dy_T &= 0 \\ \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \frac{\partial \ln L(\theta)}{\partial \theta} L(\theta) dy_1 dy_2 \dots dy_T &= 0 \\ E \left[ \frac{\partial \ln L(\theta)}{\partial \theta} \right] &= 0 \\ E[G(\theta)] &= 0, \end{aligned}$$

which proves the result.

### Variance of the Gradient

The second property is

$$\text{var}[G(\theta)] = E[G(\theta)G(\theta)'] = -E[H(\theta)]. \quad (17)$$

This expression links the first and second derivatives of the likelihood function and establishes that the expectation of the square of the gradient is equal to the expectation of the negative of the Hessian.

Differentiating

$$\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} L(\theta) dy_1 dy_2 \dots dy_T = 1,$$

twice with respect to  $\theta$  and using the same regularity conditions to establish the first property of the gradient, gives

$$\begin{aligned} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \left[ \frac{\partial \ln L(\theta)}{\partial \theta} \frac{\partial L(\theta)}{\partial \theta'} + \frac{\partial^2 \ln L(\theta)}{\partial \theta \partial \theta'} L(\theta) \right] dy_1 dy_2 \dots dy_T &= 0 \\ \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \left[ \frac{\partial \ln L(\theta)}{\partial \theta} \frac{\partial \ln L(\theta)}{\partial \theta'} L(\theta) + \frac{\partial^2 \ln L(\theta)}{\partial \theta \partial \theta'} L(\theta) \right] dy_1 dy_2 \dots dy_T &= 0 \\ \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \left[ \frac{\partial \ln L(\theta)}{\partial \theta} \frac{\partial \ln L(\theta)}{\partial \theta'} + \frac{\partial^2 \ln L(\theta)}{\partial \theta \partial \theta'} \right] L(\theta) dy_1 dy_2 \dots dy_T &= 0 \\ E \left[ \frac{\partial \ln L(\theta)}{\partial \theta} \frac{\partial \ln L(\theta)}{\partial \theta'} \right] + E \left[ \frac{\partial^2 \ln L(\theta)}{\partial \theta \partial \theta'} \right] &= 0 \\ E [G(\theta)G(\theta)'] + E [H(\theta)] &= 0, \end{aligned}$$

which proves the result.

These results are completely general, as they hold for any arbitrary distribution, as the following example based on the Poisson distribution shows.

### Poisson Distribution

Let  $\{y_1, y_2, \dots, y_T\}$  be *iid* observations from a Poisson distribution

$$f(y) = \frac{\theta^y \exp(-\theta)}{y!},$$

where  $\theta > 0$ . The log-likelihood function for the sample is

$$\ln L(\theta) = \sum_{t=1}^T y_t \ln \theta - T\theta - \ln(y_1! y_2! \dots y_T!).$$

Taking first and second derivatives yields respectively the gradient and Hessian

$$G(\theta) = \frac{1}{\theta} \sum_{t=1}^T y_t - T, \quad H(\theta) = -\frac{1}{\theta^2} \sum_{t=1}^T y_t.$$

To establish the first property of the gradient consider

$$E[G(\theta)] = E\left[\frac{1}{\theta} \sum_{t=1}^T y_t - T\right] = \frac{1}{\theta} \sum_{t=1}^T E[y_t] - T = \frac{1}{\theta} \sum_{t=1}^T \theta - T,$$

as  $E[y_t] = \theta$ , for the Poisson distribution. Hence

$$E[G(\theta)] = \frac{T\theta}{\theta} - T = 0.$$

To establish the second property of the gradient consider

$$\text{var}[G(\theta)] = \text{var}\left[\frac{1}{\theta} \sum_{t=1}^T y_t - T\right] = \text{var}\left[\frac{1}{\theta} \sum_{t=1}^T y_t\right] = \frac{1}{\theta^2} \sum_{t=1}^T \text{var}[y_t],$$

the last step following from the assumed independence of  $y_t$ . Since  $\text{var}[y_t] = \theta$  for the Poisson distribution, it follows that

$$\text{var}[G(\theta)] = \frac{1}{\theta^2} \sum_{t=1}^T \theta = \frac{1}{\theta^2} T\theta = \frac{T}{\theta}.$$

Alternatively

$$E[H(\theta)] = E\left[-\frac{1}{\theta^2} \sum_{t=1}^T y_t\right] = -\frac{1}{\theta^2} \sum_{t=1}^T E[y_t] = -\frac{T}{\theta},$$

again using the result that  $E[y_t] = \theta$ , for the Poisson distribution. Thus,

$$E[G(\theta)^2] + E[H(\theta)] = \frac{T}{\theta} - \frac{T}{\theta} = 0,$$

which establishes the result for the Poisson distribution.

## The Information Matrix

An important concept used to establish the asymptotic distribution of the maximum likelihood estimator is the information matrix, defined to be the negative expectation of the Hessian:

$$I(\theta) = -E[H(\theta)].$$

To derive the information matrix rewrite equation (17)

$$\begin{aligned} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} & \left[ \frac{\partial \ln L(\theta)}{\partial \theta} \frac{\partial \ln L(\theta)}{\partial \theta'} L(\theta) \right] dy_1 dy_2 \cdots dy_T \\ & = - \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left[ \frac{\partial^2 \ln L(\theta)}{\partial \theta \partial \theta'} L(\theta) \right] dy_1 dy_2 \cdots dy_T. \end{aligned}$$

The information matrix may now be computed either as

$$I(\theta) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{\partial \ln L(\theta)}{\partial \theta} \frac{\partial \ln L(\theta)}{\partial \theta'} L(\theta) dy_1 dy_2 \cdots dy_T. \quad (18)$$

or

$$I(\theta) = - \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{\partial^2 \ln L(\theta)}{\partial \theta \partial \theta'} L(\theta) dy_1 dy_2 \cdots dy_T. \quad (19)$$

In other words

$$I(\theta) = E [G(\theta)G(\theta)'] = -E [H(\theta)]. \quad (20)$$

This result is the well-known information equality which is used to motivate alternative numerical estimation algorithms, the derivation of alternative test statistics based on the likelihood principle and in establishing the basis of quasi-maximum likelihood estimation.

## 1.6 Asymptotic Properties

To derive the asymptotic properties of maximum likelihood estimators, let the maximum likelihood estimator of the parameter vector  $\theta$  be denoted as  $\hat{\theta}$  and the true value by  $\theta_0$ .

### 1.6.1 Consistency

A minimum requirement of an estimator is that as the sample size increases, the estimate approaches the true population parameter value,  $\theta_0$ . This is formally represented as

$$\text{plim}(\hat{\theta}) = \theta_0, \quad (21)$$

a result which requires that any finite-sample bias and the variance of the estimator both tend to zero as  $T \rightarrow \infty$ . Given the regularity conditions, all maximum likelihood estimators are consistent.

#### **EXAMPLE: Mean of the Normal Distribution**

As established previously, the sample mean is the maximum likelihood estimator of the population mean of a normal distribution. The means for samples of increasing size  $T = 1, 2, \dots, 500$ , from a  $N(1, 2)$  distribution are presented in Figure 1. For small samples the sample means exhibit large deviations from the true population mean  $\mu = 1$ . As the sample size increases, the size of the deviations decrease, thereby demonstrating the consistency property of the sample mean for this example. While the sample means are not necessarily identically equal to the true

population mean in larger samples as would be the case in a deterministic limit, the convergence in probability is apparent.

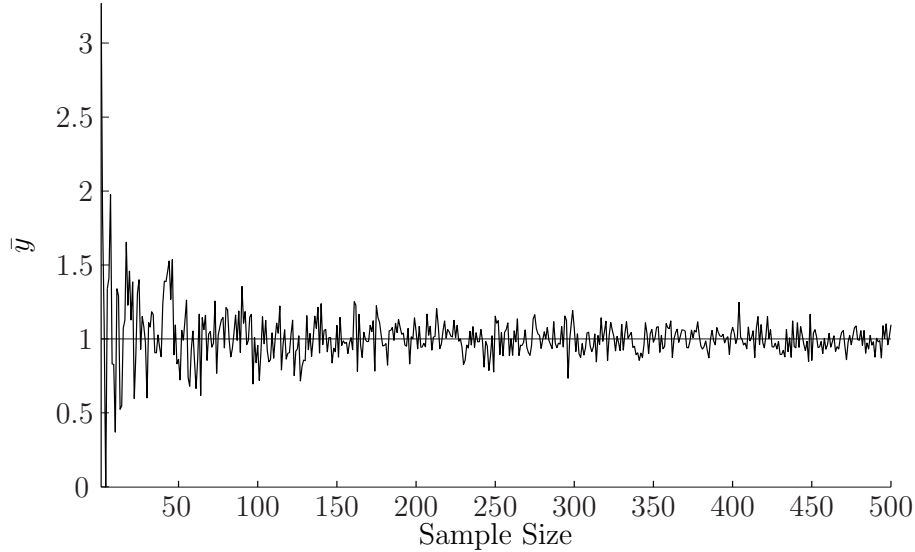


Figure 1: Demonstration of the consistency properties of the sample mean for the mean of a normal distribution for samples of increasing size  $T = 1, 2, \dots, 500$ .

For a large number of distributions the sample mean represents a consistent estimator of the population mean. A counter example is given next where the sample mean is shown to be an inconsistent estimator of the location parameter of the Cauchy distribution.

**EXAMPLE: Location parameter of the Cauchy Distribution**

This example is similar to the previous one, except that the distribution of  $y$  is now a Cauchy distribution given by

$$f(y) = \frac{1}{\pi} \frac{1}{1 + (y - \theta)^2},$$

where the location parameter of this distribution is  $\theta = 1$ . Panel (a) of Figure 2 demonstrates that the sample mean is not a consistent estimator of  $\theta$ . By contrast with the normal distribution used in the previous example, the sample means now exhibit large deviations from  $\theta = 1$  as the sample size increases and do not settle down. The sampling distribution of the sample mean as an estimate of  $\theta$  is itself a Cauchy distribution and there is no convergence in probability of this estimate to  $\theta$  as  $T$  increases. For the Cauchy distribution a consistent estimator of  $\theta$  is the median. The consistency property of the median is demonstrated in panel (b) of

Figure 2, where for progressively larger samples the median is seen to converge in probability to  $\theta = 1$ .

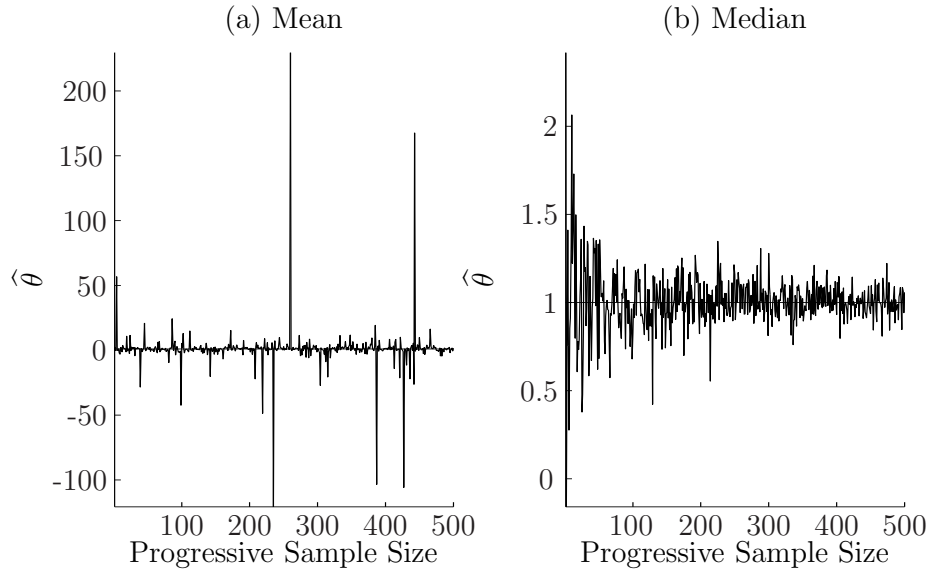


Figure 2: Demonstration of the inconsistency of the sample mean and the consistency of the sample median as estimators of the location parameter of a Cauchy distribution with  $\theta = 1$ , for samples of increasing size  $T = 1, 2, \dots, 500$ .

### 1.6.2 Asymptotic Normality

The sampling distribution of the maximum likelihood estimator is

$$\sqrt{T}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, I_t(\theta_0)^{-1}) \quad (22)$$

or, alternatively,

$$\hat{\theta} \stackrel{a}{\sim} N(\theta_0, I(\theta_0)^{-1}). \quad (23)$$

The square roots of the diagonal elements of  $I(\theta_0)^{-1}$  represent the standard errors. A full derivation is this result beyond the scope of this intermediate course.

#### EXAMPLE: Illustrating Asymptotic Normality

Figure 3 gives the results of sampling *iid* random variables from an exponential distribution with  $\theta_0 = 1$  for samples of size  $T = 5$  and  $T = 100$ . The number of replications is  $R = 5000$ . For each replication the maximum likelihood estimator is computed as

$$\hat{\theta}_i = \bar{X}_i, \quad i = 1, 2, \dots, 5000.$$

The sample means are then standardized using the population mean ( $\theta_0$ ) and the population variance ( $\theta_0^2/T$ ) as

$$Z_i = \frac{\bar{X}_i - 1}{\sqrt{1^2/T}}, \quad i = 1, 2, \dots, 5000.$$

(a) Exponential distribution

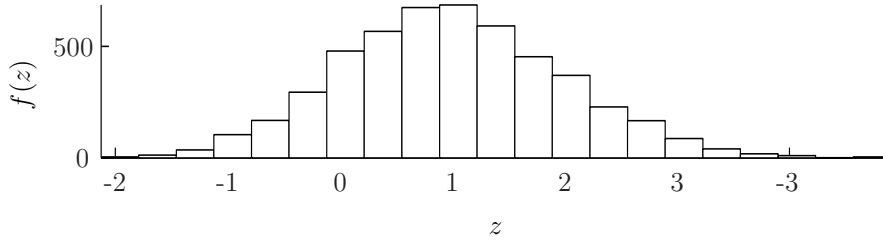
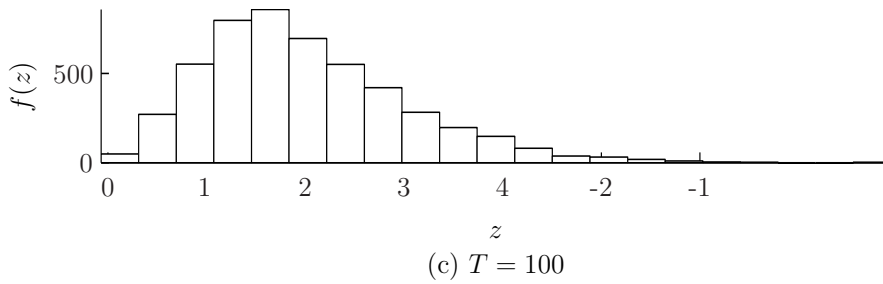
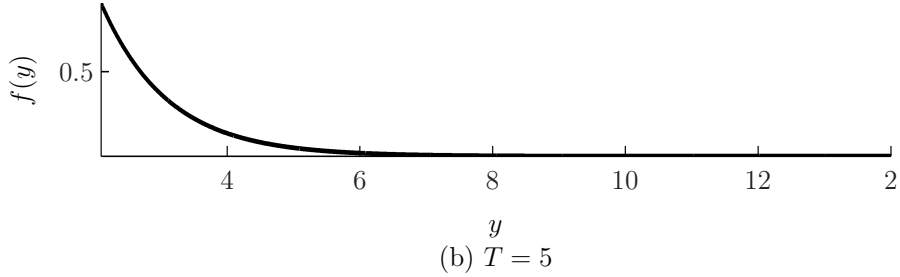


Figure 3: Demonstration of asymptotic normality of maximum likelihood estimators based on samples of size  $T = 5$  and  $T = 100$  from an exponential distribution.

The sampling distributions, as represented by the histograms, are presented in Figure 3 for samples of size  $T = 5$  (panel (b)) and  $T = 100$  (panel (c)). For samples of size  $T = 5$ , the sampling distribution is humped but still slightly skewed to the right, thereby mimicking the positive skewness characteristics of the population distribution given in Figure 3 panel (a). As the sample size is increased to  $T = 100$ , the distribution is now less skewed and nearly normal.

### 1.6.3 Asymptotic Efficiency

Consider an estimator  $\tilde{\theta}$  that is a consistent estimator of the true value  $\theta_0$  and has asymptotic distribution

$$\sqrt{T}(\tilde{\theta} - \theta_0) \xrightarrow{d} N(0, V), \quad (24)$$

for some variance matrix  $V$ , called the asymptotic variance of  $\tilde{\theta}$ . If  $\tilde{\theta}$  is the maximum likelihood estimator  $\hat{\theta}$  then  $V = I_t(\theta_0)^{-1}$ . However  $\tilde{\theta}$  may be derived from some other principle, such as the generalized methods of moments.

Suppose there are two consistent estimators,  $\tilde{\theta}_1$  and  $\tilde{\theta}_2$ , that satisfy (24) with asymptotic variances  $V_1$  and  $V_2$  respectively. Then  $\tilde{\theta}_1$  is defined to be *asymptotically more efficient* than  $\tilde{\theta}_2$  if the difference  $V_2 - V_1$  is a positive semi-definite matrix. If the parameter is a scalar, this simply means that  $\tilde{\theta}_1$  has a smaller asymptotic variance than  $\tilde{\theta}_2$ .

More generally, a consistent estimator  $\tilde{\theta}$  that satisfies (24) is defined to be *asymptotically efficient* if it is asymptotically more efficient than all other consistent estimators that satisfy (24). That is, an asymptotically efficient estimator has the smallest possible asymptotic variance.

The *Cramer-Rao Lower Bound* provides a bound on the possible efficiency of an estimator. It states that for any consistent estimator that satisfies (24),  $V - I_t(\theta_0)$  is a positive semi-definite matrix. That is, the information matrix  $I_t(\theta_0)$  is the smallest possible asymptotic variance. Since the maximum likelihood estimator has been shown in equation (22) to have precisely this asymptotic variance, we can conclude that it is asymptotically efficient.

Asymptotic efficiency is a desirable property for an estimator insofar as it predicts that the estimator will also have good finite sample efficiency properties. That is, we might hope that if two consistent estimators  $\tilde{\theta}_1$  and  $\tilde{\theta}_2$  have asymptotic variances  $V_1$  and  $V_2$  with  $V_1 < V_2$ , then  $\text{var}(\tilde{\theta}_1) < \text{var}(\tilde{\theta}_2)$  for finite  $T$  as well. In practice this holds often, but not always, and finite sample efficiency is, by necessity, most often evaluated using simulation.

#### **EXAMPLE: Relative Efficiency of the Mean and the Median**

The log-likelihood function of a normal distribution with unknown population mean

$\mu$  and with known variance  $\sigma^2$ , is

$$\ln L(\mu) = -\frac{T}{2} \ln 2\pi - \frac{T}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{t=1}^T (y_t - \mu)^2.$$

As established previously, the maximum likelihood estimator of  $\mu$  is the sample mean  $\bar{y}$ . The second derivative is

$$\frac{\partial^2 \ln L(\mu)}{\partial \mu^2} = -\frac{T}{\sigma^2}.$$

Taking expectations, changing the sign and inverting the result gives the variance of the maximum likelihood estimator

$$\text{var}(\bar{y}) = \frac{\sigma^2}{T},$$

which is the usual formula for the variance of the sample mean based on the assumption of a normal distribution.

The variance of the sample median,  $\bar{m}$ , is approximately

$$\text{var}(\bar{m}) = \frac{1}{4Tf^2},$$

where  $f$  is the value of the pdf at the population median. For the normal distribution, the population median is zero so  $f$  is given by

$$f = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(0-\theta)^2}{2\sigma^2}\right] = \frac{1}{\sqrt{2\pi\sigma^2}}.$$

Thus the variance of the sample median is

$$\text{var}(\bar{m}) = \frac{1}{4Tf^2} = \frac{\pi\sigma^2}{2T}.$$

In which case

$$\text{var}(\bar{y}) < \text{var}(\bar{m}),$$

as  $\pi/2 > 1$ , so the maximum likelihood estimator,  $\bar{y}$ , has a smaller variance than another consistent estimator, as given by the sample median.

## 1.7 A Very Brief Introduction to Testing

Although it is not possible to consider the very important area of testing in any great detail, it is useful to illustrate the general idea underlying tests based on the maximum

likelihood principal for the case of a single parameter. In constructing the test statistics, the following definitions are adopted

$$\begin{aligned}\widehat{\theta}_0 &= \text{restricted maximum likelihood estimator,} \\ \widehat{\theta}_1 &= \text{unrestricted maximum likelihood estimator,}\end{aligned}$$

which are obtained by estimating the model under the null and alternative hypotheses respectively. As there is just one parameter in this case, the restricted maximum likelihood estimator is just  $\widehat{\theta}_0 = \theta_0$ .

A “simple” hypothesis test is to compare the unrestricted parameter estimator  $\widehat{\theta}_1$ , with the value under the null,  $\theta_0$ . If the two values are considered to be “too far apart” from each other, the null is rejected. If the values are considered to be “close”, then the null hypothesis is not rejected. The three asymptotically equivalent testing procedures are now illustrated with reference to Figures 4 and 5.

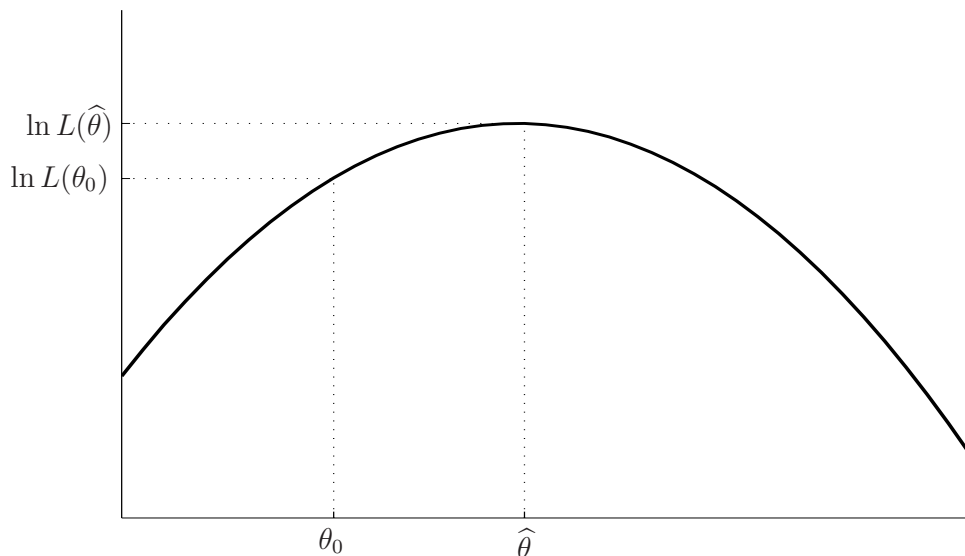


Figure 4: Comparison of the value of the log-likelihood function under the null hypothesis,  $\theta_0$ , and under the alternative hypothesis,  $\widehat{\theta}$ .

One measure of how close  $\widehat{\theta}_1$  and  $\theta_0$  are is to consider the distance between the values of the restricted and unrestricted log-likelihood functions,  $(\ln L(\widehat{\theta}_1) - \ln L(\theta_0))$ . This distance is measured on the vertical axis of Figure 4 and the test based on this measure is the LR test.

The distance  $(\widehat{\theta}_1 - \theta_0)$ , illustrated on the horizontal axis of Figure 4, is an alternative measure of the difference between  $\widehat{\theta}_1$  and  $\theta_0$ . A test based on this measure is known as a Wald test. Note that in contrast to the LR test, which requires estimation of the

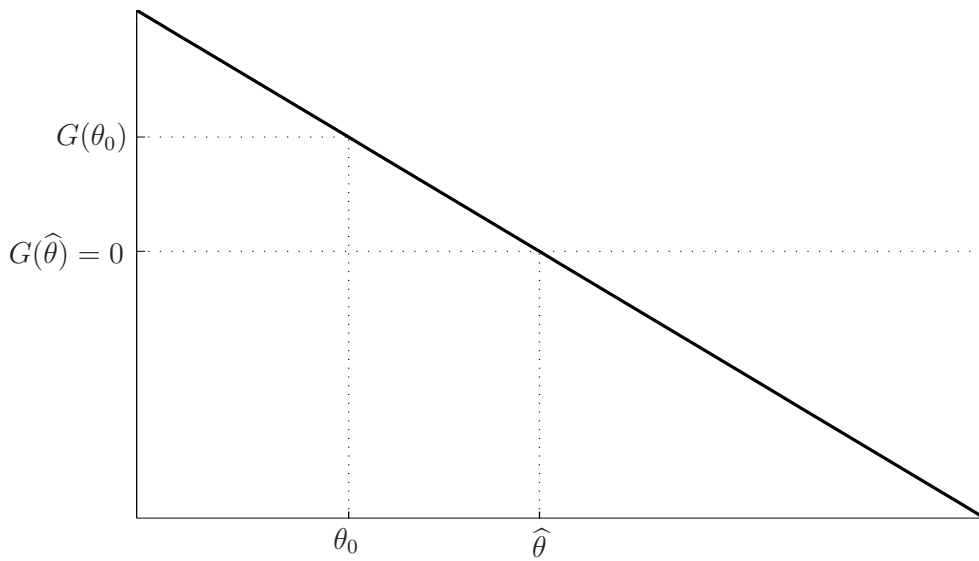


Figure 5: Comparison of the value of the gradient of the log-likelihood function under the null hypothesis,  $\theta_0$ , and under the alternative hypothesis,  $\hat{\theta}$ .

log-likelihood for both the restricted and unrestricted models, the Wald test only requires the estimation of the unrestricted model, as  $\theta_0$  is directly determined by the specification of the null hypothesis.

A hypothesis test constructed as the distance on the vertical axis between the value of the gradient evaluated at the restricted and unrestricted parameter values respectively, illustrated in Figure 5, is an LM test. As the gradient of the log-likelihood function at the unrestricted maximum likelihood estimator is zero by definition, only the restricted model needs to be estimated in order to implement the LM test.

The construction of each of these tests is now discussed in a little more detail.

### 1.7.1 Likelihood Ratio Test

The classical test procedure is based on the *likelihood ratio*, which is the ratio of the maximum value of the likelihood function under the null hypothesis divided by its maximum value when no restrictions are imposed denoted

$$\lambda = \frac{L(\theta_0)}{L(\hat{\theta})}$$

Because the denominator is based on the unrestricted model, the likelihood function value must be greater than that of the numerator, therefore providing the following bounds on  $\lambda$

$$0 \leq \lambda \leq 1$$

If the null hypothesis is correct we expect  $\lambda \simeq 1$ . What we need is a way in which  $\lambda$  can be transformed into another statistic whose distribution is known. It may be shown that for large sample sizes the statistic

$$LR = -2[L(\theta_0) - L(\hat{\theta})] = 2[L(\hat{\theta}) - L(\theta_0)]$$

is asymptotically distributed as a  $\chi^2$  random variable, with degrees of freedom equal to the number of hypotheses (in this case one). Formally, the null hypothesis is rejected if the value of the test statistic LR is too large, that is if  $LR \geq \chi_{(\alpha, J)}^2$  where  $\chi_{(\alpha, J)}^2$  is the upper  $\alpha$ -percentile of a  $\chi_{(J)}^2$  distribution.

### 1.7.2 Wald Test

It is clear that  $LR/2$  will depend on the distance  $\hat{\theta} - \theta_0$  and the curvature of  $L(\theta)$ : the more curved the likelihood surface the distance  $\hat{\theta} - \theta_0$  translates into a larger value of  $LR/2$ . this fact is used by the Wald testing procedure.

We measure the curvature of the likelihood function using the *negative* of its second derivative evaluated at the unrestricted ML estimator  $\hat{\theta}$

$$-\left. \frac{\partial^2 L(\theta)}{\partial \theta^2} \right|_{\theta=\hat{\theta}}$$

The greater this magnitude, the more curved is the likelihood function. It would seem reasonable to construct a test statistic by weighting the squared distance between  $\hat{\theta}$  and  $\theta_0$  by the curvature of the likelihood function. This is the Wald Test

$$W = (\hat{\theta} - \theta_0)^2 \left[ -\left. \frac{\partial^2 L(\theta)}{\partial \theta^2} \right|_{\theta=\hat{\theta}} \right] = (\hat{\theta} - \theta_0)^2 I(\theta)$$

where  $I(\theta)$  is the information matrix,

$$I(\theta) = -E \left[ \frac{\partial^2 L(\theta)}{\partial \theta^2} \right] \text{ evaluated at } \hat{\theta}$$

These two forms are asymptotically equivalent. Once again  $W \geq \chi_{(\alpha, J)}^2$  will lead to the rejection of the null hypothesis.

### 1.7.3 Lagrange Multiplier Test

The Lagrange Multiplier test derives from a restricted maximum likelihood estimation using Lagrange multipliers. If we maximise  $L(\theta)$  subject to the condition  $\theta = \theta_0$  we

would formulate the Lagrangian function

$$\mathcal{L} = L(\theta) - \lambda(\theta - \theta_0)$$

Differentiating  $\mathcal{L}$  W.R.T.  $\theta$  and  $\lambda$  and setting to zero yields the restricted ML estimator  $\theta^* = \theta_0$ , and the Lagrange multiplier value  $\lambda^* = S(\theta^*) = S(\theta_0)$  where  $S(\theta)$  is the slope of the log-likelihood function.

$$S(\theta) = \frac{\partial L(\theta)}{\partial \theta}$$

Since  $\hat{\theta}$  maximizes  $L(\theta)$  it must be that  $S(\hat{\theta}) = 0$ . The more the data "agree with" the null hypothesis that  $\theta = \theta_0$  the closer  $\theta^*$  will be to  $\hat{\theta}$  and the smaller will be  $\lambda^* = S(\theta^*) = S(\theta_0)$ . Thus the magnitude of  $S(\theta_0)$  measures the distance between  $\theta^* = \theta_0$  and  $\hat{\theta}$ . However, equal values of  $S(\theta_0)$  may imply different values of LR depending on the curvature of  $L(\theta)$  at  $\theta_0$ . Thus we weight  $S(\theta_0)^2$  by the reciprocal of the curvature of  $L(\theta)$  since now more curvature implies smaller differences between  $\theta_0$  and  $\hat{\theta}$ .

$$LM = \frac{[S(\theta_0)]^2}{[-\partial^2 L(\theta) / \partial \theta^2 |_{\theta=\theta_0}]} = [S(\theta_0)]^2 [I(\theta_0)]^{-1}$$

Again  $LM \sim \chi_{(J)}^2$  and we reject the null if  $LM$  is too large.

## 1.8 Computer Applications

### 1.8.1 Consistency

The aim of this exercise is to demonstrate the property of consistency, formally represented as

$$\text{plim}(\hat{\theta}) = \theta_0, \tag{25}$$

a result which requires that any finite-sample bias and the variance of the estimator both tend to zero as  $T \rightarrow \infty$ . Given the regularity conditions, all maximum likelihood estimators are consistent.

#### Mean of the Normal Distribution

The aim is to demonstrate the behaviour of the sample mean for samples of increasing size  $T = 1, 2, \dots, 500$ , from a  $N(1, 2)$  distribution and hence reproduce Figure 1.

#### Location parameter of the Cauchy Distribution

This example is similar to the previous one, except that the distribution of  $y$  is now a Cauchy distribution given by

$$f(y) = \frac{1}{\pi} \frac{1}{1 + (y - \theta)^2},$$

where the location parameter of this distribution is  $\theta = 1$ . The aim is to reproduce Figure 2 which demonstrates that the sample mean is not a consistent estimator of  $\theta$ .

### 1.8.2 Exponential Distribution

Let  $\{y_1, y_2, \dots, y_T\}$  be *iid* observations drawn from an exponential distribution

$$f(y) = \theta \exp(-\theta y), \quad y \in [0, \infty),$$

where  $\theta > 0$ . As the variables are *iid*, the log-likelihood function is

$$\begin{aligned} \ln L(\theta) &= \sum_{t=1}^T \ln f(y_t; \theta) \\ &= \sum_{t=1}^T (\ln \theta - \theta y_t) \\ &= T \ln \theta - \theta \sum_{t=1}^T y_t. \end{aligned}$$

The first and second derivatives of  $\ln L(\theta)$  with respect to  $\theta$  are respectively

$$\begin{aligned} G(\theta) &= \frac{T}{\theta} - \sum_{t=1}^T y_t \\ H(\theta) &= -\frac{T}{\theta^2}. \end{aligned}$$

The maximum likelihood estimate of  $\theta$  is obtained by setting  $G(\hat{\theta}) = 0$  and solving the resultant first-order condition. This yields

$$\hat{\theta} = \frac{T}{\sum_{t=1}^T y_t} = \frac{1}{\bar{y}},$$

which is the reciprocal of the sample mean. To establish that a maximum has been attained the Hessian evaluated at the maximum likelihood estimate is

$$H(\hat{\theta}) = -\frac{T}{\hat{\theta}^2} < 0.$$

As this term is negative for any  $\hat{\theta}$ , a maximum is achieved.

## Demonstrating Asymptotic Normality

In this example, for convenience, the exponential distribution is defined as

$$f(y) = \frac{1}{\theta} \exp\left(-\frac{1}{\theta}y\right), \quad y \in [0, \infty),$$

so that the maximum likelihood estimator is in fact the sample mean rather than its inverse. The aim here is to reproduce Figure 3 that gives the results of sampling *iid* random variables from an exponential distribution with  $\theta_0 = 1$  for samples of size  $T = 5$  and  $T = 100$ . The number of replications is  $R = 5000$ . For each replication the maximum likelihood estimator is computed as

$$\hat{\theta}_i = \bar{X}_i, \quad i = 1, 2, \dots, 5000.$$

The sample means are then standardized using the population mean ( $\theta_0$ ) and the population variance ( $\theta_0^2/T$ ) as

$$Z_i = \frac{\bar{X}_i - 1}{1/\sqrt{T}}, \quad i = 1, 2, \dots, 5000.$$

## Demonstrating the Maximum Likelihood Estimators

Let  $\{y_1, y_2, \dots, y_T\}$  be *iid* observations drawn from the conventionally-defined exponential distribution

$$f(y) = \theta \exp(-\theta y), \quad y \in [0, \infty),$$

where  $\theta > 0$ . Consider the following  $T = 6$  observations

$$y_t = \{2.1, 2.2, 3.1, 1.6, 2.5, 0.5\}.$$

As

$$T = 6, \quad \sum_{t=1}^T y_t = 12,$$

the log-likelihood function is

$$\begin{aligned} \ln L(\theta) &= T \ln \theta - \theta \sum_{t=1}^T y_t \\ &= 6 \ln \theta - 12 \theta. \end{aligned}$$

Plots of the log-likelihood,  $\ln L(\theta)$ , and the likelihood  $L(\theta)$  functions are given in Figure 6, which show that a maximum occurs at 0.5. This agrees with the closed-form solution

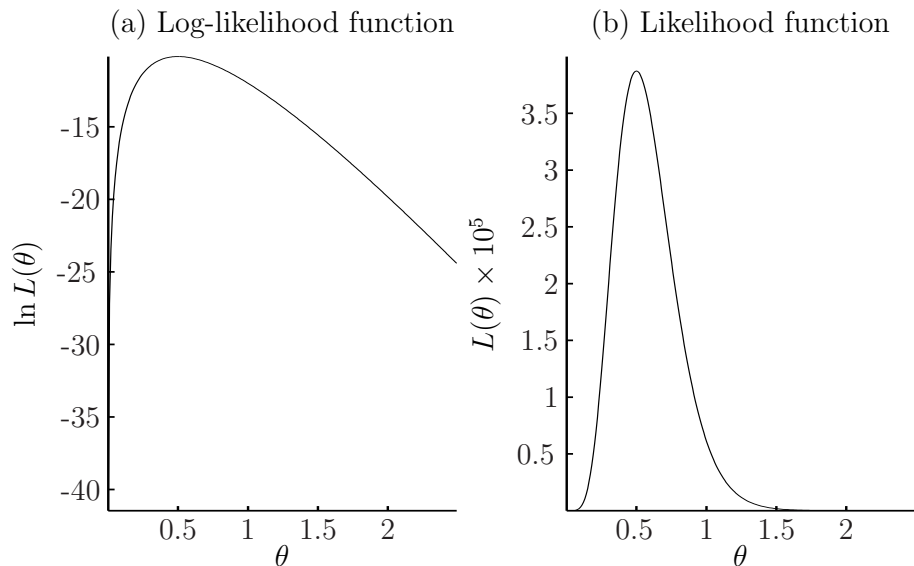


Figure 6: Plot of  $\ln L(\theta)$  for the exponential distribution example.

Table 1:

Maximum likelihood calculations for the exponential distribution problem. Maximum likelihood estimate is  $\hat{\theta} = 0.5$ .

$y_t$	$\ln L_t(0.5)$	$G_t(0.5)$	$H_t(0.5)$
2.1	-1.743	-0.100	-4.000
2.2	-1.793	-0.200	-4.000
3.1	-2.243	-1.100	-4.000
1.6	-1.493	0.400	-4.000
2.5	-1.943	-0.500	-4.000
0.5	-0.943	1.500	-4.000
$\ln L(0.5) = -10.159$		$G(0.5) = 0.000$	$H(0.5) = -24.000$

for the maximum likelihood estimate

$$\hat{\theta} = \frac{T}{\sum_{t=1}^T y_t} = \frac{6}{12} = 0.5.$$

Table 1 provides details of the calculations. Let the log-likelihood function at each observation evaluated at the maximum likelihood estimate be denoted  $\ln L_t(\theta)$ . The second column shows  $\ln L_t(\theta)$  evaluated at  $\hat{\theta} = 0.5$

$$\ln L_t(0.5) = \ln(0.5) - 0.5y_t.$$

The log-likelihood function evaluated at the maximum likelihood estimate is

$$\ln L(0.5) = \sum_{t=1}^6 \ln L_t(0.5) = -10.159,$$

which is also highlighted in Figure 6. The third column gives the gradients at each observation evaluated at the maximum likelihood estimate

$$G_t(0.5) = \frac{1}{0.5} - y_t.$$

Summing the gradients gives

$$G(0.5) = \sum_{t=1}^6 G_t(0.5) = 0.000,$$

which follows immediately from the properties of the maximum likelihood estimator. Finally, the last column of Table 1 gives the Hessian at each observation evaluated at the maximum likelihood estimate. As, in this particular case, the Hessian is not a function of the data,  $H_t$  is constant for all  $t$  and given by

$$H_t(0.5) = -\frac{1}{0.5^2} = -4.$$

The Hessian for the likelihood is then

$$H(0.5) = \sum_{t=1}^6 H_t(0.5) = -24.000,$$

which is negative confirming that a maximum has been reached.

### 1.8.3 The Classical Regression Model

Consider the regression model

$$y_t = \beta x_t + u_t,$$

where  $y_t$  is the dependent variable,  $x_t$  is the explanatory variable and the disturbance term is  $u_t \sim N(0, \sigma^2)$ . Problem with writing down the likelihood function is that the distribution of  $u_t$  is given, but it is the joint probability distribution of the observed sample, ie. the  $y_t$ 's that is needed.

---

## ASIDE: Change of Variable in Probability Density Functions

Let  $X$  be a continuous random variable with pdf  $f(x)$ . Define a new random variable  $Y$  by means of the relation  $Y = g(X)$  where the  $g$  is a monotonic one-to-one mapping, so that its inverse function exists. The pdf of the continuous random variable  $Y$ ,  $h(y)$  is given by

$$h(y) = f(g^{-1}(y)) \left| \frac{dg^{-1}(y)}{dy} \right| = f(x) \left| \frac{dx}{dy} \right|.$$

where  $dx/dy$  is known as the Jacobian of the transformation.

This result generalizes to functions of more than one variable. Let  $X_1, \dots, X_n$  be random variables whose joint pdf is given by  $f(x_1, \dots, x_n)$ . Once again define the functions  $g_i$  to be monotonic bijections so that

$$\begin{aligned} Y_i &= g_i(X_1, \dots, X_n) \\ X_i &= g_i^{-1}(Y_1, \dots, Y_n), \end{aligned}$$

and assume that the derivatives  $\partial g_i^{-1}(x_1, \dots, x_n)/\partial y_j$  exist for all  $i, j$ .

The Jacobian of the transformation is now defined as the determinant of the matrix of partial derivatives

$$J = \begin{vmatrix} \frac{\partial g_1^{-1}(x_1, \dots, x_n)}{\partial y_1} & \dots & \frac{\partial g_1^{-1}(x_1, \dots, x_n)}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_n^{-1}(x_1, \dots, x_n)}{\partial y_1} & \dots & \frac{\partial g_n^{-1}(x_1, \dots, x_n)}{\partial y_n} \end{vmatrix} = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \dots & \frac{\partial x_1}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial y_1} & \dots & \frac{\partial x_n}{\partial y_n} \end{vmatrix}$$

The joint pdf of  $Y_1, \dots, Y_n$  is now given by

$$\begin{aligned} h(y_1, \dots, y_n) &= f(g_1^{-1}(y_1, \dots, y_n), \dots, g_n^{-1}(y_1, \dots, y_n)) |J| \\ &= f(x_1, \dots, x_n) |J|. \end{aligned}$$

---

Using the change of variable method, the distribution of  $y_t$  is

$$f(y_t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{(y_t - \beta x_t)^2}{2\sigma^2} \right].$$

The log-likelihood function is

$$\begin{aligned}\ln L(\theta) &= \sum_{t=1}^T \left( -\frac{1}{2} \ln 2\pi - \frac{1}{2} \ln \sigma^2 - \frac{(y_t - \beta x_t)^2}{2\sigma^2} \right) \\ &= -\frac{T}{2} \ln 2\pi - \frac{T}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{t=1}^T (y_t - \beta x_t)^2,\end{aligned}$$

where  $\theta = \{\beta, \sigma^2\}$ .

The first derivatives of the log-likelihood function are

$$\begin{aligned}\frac{\partial \ln L}{\partial \beta} &= \frac{1}{\sigma^2} \sum_{t=1}^T (y_t - \beta x_t) x_t \\ \frac{\partial \ln L}{\partial (\sigma^2)} &= -\frac{T}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{t=1}^T (y_t - \beta x_t)^2,\end{aligned}$$

and the gradient is therefore

$$G(\theta) = \begin{bmatrix} \frac{1}{\sigma^2} \sum_{t=1}^T (y_t - \beta x_t) x_t \\ -\frac{T}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{t=1}^T (y_t - \beta x_t)^2 \end{bmatrix}.$$

Setting  $G(\hat{\theta}) = 0$  yields the maximum likelihood estimators

$$\begin{aligned}\hat{\beta} &= \frac{\sum_{t=1}^T x_t y_t}{\sum_{t=1}^T x_t^2} \\ \hat{\sigma}^2 &= \frac{\sum_{t=1}^T (y_t - \hat{\beta} x_t)^2}{T}.\end{aligned}$$

The second-order conditions are

$$\begin{aligned}\frac{\partial^2 \ln L}{\partial \beta^2} &= -\frac{1}{\sigma^2} \sum_{t=1}^T x_t^2 \\ \frac{\partial^2 \ln L}{\partial \beta \partial \sigma^2} &= -\frac{1}{\sigma^4} \sum_{t=1}^T (y_t - \beta x_t) x_t \\ \frac{\partial^2 \ln L}{\partial (\sigma^2)^2} &= \frac{T}{2\sigma^4} - \frac{1}{\sigma^6} \sum_{t=1}^T (y_t - \beta x_t)^2.\end{aligned}$$

Constructing the Hessian and evaluating it at  $\hat{\theta} = \{\hat{\beta}, \hat{\sigma}^2\}$ , gives

$$H(\hat{\theta}) = \begin{bmatrix} -\frac{1}{\hat{\sigma}^2} \sum_{t=1}^T x_t^2 & 0 \\ 0 & -\frac{T}{2\hat{\sigma}^4} \end{bmatrix},$$

given that the independence of  $x_t$  and  $u_t$  ensures that

$$\sum_{t=1}^T (y_t - \hat{\beta}x_t)x_t = 0.$$

The Hessian matrix,  $H(\hat{\theta})$ , is negative definite and the second-order condition for a maximum is therefore satisfied.

Assume now that the independent variable  $x_t$  is given by  $x = \{1, 2, 4, 5, 8\}$ .

1. Simulate the model for  $T = 5$  observations using the parameter values  $\beta = 1$  and  $\sigma^2 = 4$ .
2. Compute the average log-likelihood function  $A(\theta)$ .
3. Compute the gradient vector  $G(\theta)$ .
4. Compute the Hessian matrix  $H(\theta)$ .
5. Compute the maximum likelihood estimators based on the analytical expressions.

## 2 ESTIMATING NONLINEAR MODELS

The maximum likelihood estimator of the parameter vector  $\theta$  is obtained by solving the first-order conditions given by setting the gradient of the log-likelihood function to zero. For all the examples considered so far, analytical expressions for the maximum likelihood estimators are available. In many interesting cases, however, no closed-form expression exists, as the following example demonstrates.

To obtain the maximum likelihood estimator when no analytical solution is available, numerical optimization algorithms need to be used. These algorithms begin by assuming starting values for the unknown parameters and then proceeding iteratively. A general form for the  $k^{\text{th}}$  iteration is

$$\theta_{(k)} = F(\theta_{(k-1)}), \quad (26)$$

where the form of the function  $F(\cdot)$  is governed by the choice of the numerical algorithm. Convergence of the algorithm is achieved when the log-likelihood function cannot be further improved, a situation in which  $\theta_{(k)} \simeq \theta_{(k-1)}$ , resulting in  $\theta_{(k)}$  being the maximum likelihood estimator of  $\theta$ .

The aim of this session is to introduce the algorithms most commonly used in econometrics for computing maximum likelihood estimates. Broadly speaking, these may be divided into Newton methods (Newton-Raphson, method of scoring and BHHH algorithm) and quasi-Newton methods (BFGS algorithm). In addition to discussing these algorithms, alternative methods for computing the variance-covariance matrix of the maximum likelihood estimators are addressed.

### 2.1 Motivating Examples

#### Cauchy Distribution

Let  $\{y_1, y_2, \dots, y_T\}$  be  $T$  *iid* realized values from the Cauchy distribution

$$f(y) = \frac{1}{\pi} \frac{1}{1 + (y - \theta)^2},$$

where  $\theta$  is the unknown parameter. The log-likelihood function is

$$\ln L(\theta) = -T \ln \pi - \sum_{t=1}^T \ln [1 + (y_t - \theta)^2],$$

resulting in the gradient

$$\frac{d \ln L(\theta)}{d\theta} = G(\theta) = 2 \sum_{t=1}^T \frac{y_t - \theta}{1 + (y_t - \theta)^2}.$$

The first-order condition required to obtain the maximum likelihood estimator,  $\hat{\theta}$ , is

$$2 \sum_{t=1}^T \frac{y_t - \hat{\theta}}{1 + (y_t - \hat{\theta})^2} = 0.$$

This is a nonlinear function of  $\hat{\theta}$  for which no analytical solution exists.

### EXAMPLE: The GARCH Model

The ARCH( $q$ ) model, which was covered extensively in the Introduction to Financial Econometrics course earlier in the year, has the property that the memory in the variance stops at the lag  $q$ . This means that for processes that exhibit long memory in the variance, it would be necessary to specify and estimate a high dimensional model. A natural way to circumvent this problem is to specify the conditional variance as a function of its own lags. The equation for the conditional variance then becomes

$$h_t = \alpha_0 + \sum_{i=1}^q \alpha_i u_{t-i}^2 + \sum_{i=1}^p \beta_i h_{t-i}, \quad (27)$$

which is known as GARCH( $p, q$ ) where the  $p$  and the  $q$  identify the lags of the model and the “G” stands for Generalized ARCH. Once again without any loss of generality it is convenient to work with the GARCH(1,1) model, which is

$$\begin{aligned} y_t &= u_t \\ u_t &\sim N(0, h_t) \\ h_t &= \alpha_0 + \alpha_1 y_{t-1}^2 + \beta_1 h_{t-1}. \end{aligned} \quad (28)$$

To highlight the long memory properties of this model, rewrite the expression for the conditional variance,  $h_t$ , using the lag operator  $L^k y_t = y_{t-k}$  to yield

$$(1 - \beta_1 L) h_t = \alpha_0 + \alpha_1 y_{t-1}^2.$$

Assuming that  $|\beta_1| < 1$ , and using the properties of the lag operator, the conditional variance can be expressed as

$$\begin{aligned} h_t &= (1 - \beta_1 L)^{-1} \alpha_0 + \alpha_1 (1 - \beta_1 L)^{-1} y_{t-1}^2 \\ &= \frac{\alpha_0}{1 - \beta_1} + \alpha_1 y_{t-1}^2 + \alpha_1 \sum_{i=1}^{\infty} \beta_1^i y_{t-1-i}^2, \end{aligned} \quad (29)$$

which is instantly recognizable as an ARCH( $\infty$ ) model. The first two terms correspond to the ARCH(1) model with the second term determining the role of the first lag of  $y_t^2$  on the conditional variance, while the third term captures the effects of higher order lags  $\{y_{t-2}^2, y_{t-3}^2, y_{t-4}^2, \dots\}$  on the conditional variance which are now controlled by a sole parameter, namely  $\beta_1$ . It is this property that makes this model an attractive specification to use in modelling financial data as it provides a parsimonious representation of the memory characteristics commonly observed in the variance of financial returns.

Another way to highlight the memory characteristics of the GARCH conditional variance is to define the (forecast) error

$$v_t = y_t^2 - h_t, \quad (30)$$

which has the property  $E_{t-1}[v_t] = 0$ . Rearranging this expression and using (28) gives

$$\begin{aligned} y_t^2 &= h_t + v_t \\ y_t^2 &= \alpha_0 + \alpha_1 y_{t-1}^2 + \beta_1 h_{t-1} + v_t \\ y_t^2 &= \alpha_0 + \alpha_1 y_{t-1}^2 + \beta_1 (y_{t-1}^2 - v_{t-1}) + v_t \\ y_t^2 &= \alpha_0 + (\alpha_1 + \beta_1) y_{t-1}^2 - \beta_1 v_{t-1} + v_t, \end{aligned} \quad (31)$$

which is an ARMA(1,1) model in terms of  $y_t^2$ . The memory of this process is determined by the autoregressive parameter  $\alpha_1 + \beta_1$ . The closer is  $\alpha_1 + \beta_1$  to unity, the longer is the effect of a shock on volatility.

The effect of a shock in the long-run on volatility is obtained from the unconditional variance of  $y_t$ , defined as  $h = E[y_t^2]$ . Taking unconditional expectations of (31) and using the result  $E[v_t] = E[v_{t-1}] = 0$

$$\begin{aligned} E[y_t^2] &= E[\alpha_0 + (\alpha_1 + \beta_1) y_{t-1}^2 - \beta_1 v_{t-1} + v_t,] \\ h &= \alpha_0 + (\alpha_1 + \beta_1) y_{t-1}^2 h, \end{aligned}$$

upon rearranging gives an expression of the unconditional, or long-run, variance

$$h = \frac{\alpha_0}{1 - \alpha_1 - \beta_1}. \quad (32)$$

The GARCH model can be estimated by maximizing the log likelihood function using a gradient algorithm. For the GARCH(1,1) model the unknown parameters are

$$\theta = \{\alpha_0, \alpha_1, \beta_1\},$$

while the log of the conditional distribution is

$$\begin{aligned} \ln L_t &= \ln f(y_t|y_{t-1}) \\ &= -\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln(h_t) - \frac{1}{2} \frac{y_t^2}{h_t} \\ &= -\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln(\alpha_0 + \alpha_1 y_{t-1}^2 + \beta_1 h_{t-1}) \\ &\quad - \frac{1}{2} \frac{y_t^2}{\alpha_0 + \alpha_1 y_{t-1}^2 + \beta_1 h_{t-1}}. \end{aligned} \tag{33}$$

In estimating this model it may be necessary to restrict the parameters  $\{\alpha_0, \alpha_1, \beta_1\}$  to be positive to ensure that the conditional variance is positive for all  $t$ .

## 2.2 Newton Methods

Recall that the gradient and Hessian are defined to be

$$G(\theta) = \frac{\partial \ln L(\theta)}{\partial \theta} \tag{34}$$

$$H(\theta) = \frac{\partial^2 \ln L(\theta)}{\partial \theta \partial \theta'}, \tag{35}$$

respectively. A first-order Taylor series expansion of the gradient function around the true parameter vector  $\theta_0$  is

$$\frac{\partial \ln L(\theta)}{\partial \theta} \simeq \left. \frac{\partial \ln L(\theta)}{\partial \theta} \right|_{\theta=\theta_0} + \left. \frac{\partial^2 \ln L(\theta)}{\partial \theta^2} \right|_{\theta=\theta_0} (\theta - \theta_0), \tag{36}$$

where higher-order terms are excluded in the expansion. This expression is written more compactly as

$$G(\theta) \simeq G(\theta_0) + H(\theta_0)(\theta - \theta_0), \tag{37}$$

where  $G(\theta_0)$  and  $H(\theta_0)$  are the gradient and Hessian evaluated at the true parameter value,  $\theta_0$ .

As the maximum likelihood estimator,  $\hat{\theta}$ , is the solution to the equation

$$G(\hat{\theta}) = 0, \tag{38}$$

then from equation (37) the maximum likelihood estimator satisfies

$$G(\hat{\theta}) = 0 \simeq G(\theta_0) + H(\theta_0)(\hat{\theta} - \theta_0), \quad (39)$$

where for convenience the equation is now written as an equality. This is a linear equation in  $\hat{\theta}$  with solution

$$\hat{\theta} \simeq \theta_0 - H(\theta_0)^{-1}G(\theta_0). \quad (40)$$

As it stands, this equation is of no practical use because it expresses the maximum likelihood estimator as a function of the unknown parameter that it is trying to estimate, namely  $\theta_0$ . It does suggest, however, that a natural way to proceed is to replace  $\theta_0$  with some starting value and use (40) as an updating scheme. This is indeed the basis of Newton methods. Three algorithms are discussed, differing only in the way that the Hessian  $H(\theta)$  is evaluated.

### 2.2.1 Newton-Raphson

Let  $\theta_{(k)}$  be the value of the unknown parameters at the  $k^{\text{th}}$  iteration. The Newton-Raphson algorithm (NR) is given by replacing  $\theta_0$  in (40) by  $\theta_{(k-1)}$  to yield the updated parameter  $\theta_{(k)}$

$$\theta_{(k)} = \theta_{(k-1)} - H_{(k-1)}^{-1}G_{(k-1)}, \quad (41)$$

where

$$G_{(k)} = \left. \frac{\partial \ln L(\theta)}{\partial \theta} \right|_{\theta=\theta_{(k)}} \quad (42)$$

$$H_{(k)} = \left. \frac{\partial^2 \ln L(\theta)}{\partial \theta^2} \right|_{\theta=\theta_{(k)}}. \quad (43)$$

The algorithm proceeds until  $\theta_{(k)} \simeq \theta_{(k-1)}$ , subject to some tolerance level which is discussed in more detail later. From (41) this occurs when

$$\theta_{(k)} - \theta_{(k-1)} = -H_{(k-1)}^{-1}G_{(k-1)} \simeq 0,$$

which can only be satisfied when

$$G_{(k)} \simeq G_{(k-1)} \simeq 0,$$

as  $H_{(k-1)}^{-1}$  and hence  $H_{(k)}^{-1}$  are negative definite. But this is exactly the condition that defines the maximum likelihood estimator,  $\hat{\theta}$ , and thus  $\theta_{(k)} \simeq \hat{\theta}$  at the final iteration.

To implement the Newton-Raphson algorithm both the first and second derivatives of the log-likelihood function,  $G(\cdot)$  and  $H(\cdot)$ , are needed at each iteration.

### 2.2.2 Method of Scoring

The method of scoring uses the fact that the information matrix is given by

$$I(\theta) = E \left[ -\frac{\partial^2 \ln L}{\partial \theta \partial \theta'} \right].$$

This suggests that another variation of (40) is to replace  $-H(\theta_0)$  by the information matrix evaluated at  $\theta^{(k)}$ . The iterative scheme of the method of scoring is then

$$\theta_{(k)} = \theta_{(k-1)} + I_{(k-1)}^{-1} G_{(k-1)}. \quad (44)$$

To implement the scoring algorithm, it is necessary to derive the information matrix. The main problem, however, is that for many of the models used in econometrics the calculation of the information matrix is potentially difficult.

### 2.2.3 BHHH Algorithm

The Berndt, Hall, Hall and Hausman or BHHH algorithm is based on the information matrix equality

$$E \left[ \frac{\partial \ln L(\theta)}{\partial \theta} \frac{\partial \ln L(\theta)}{\partial \theta'} \right] + E \left[ \frac{\partial^2 \ln L(\theta)}{\partial \theta \partial \theta'} \right] = 0, \quad (45)$$

discussed in the previous session. Upon re-arranging this expression becomes

$$E \left[ \frac{\partial \ln L(\theta)}{\partial \theta} \frac{\partial \ln L(\theta)}{\partial \theta'} \right] = -E \left[ \frac{\partial^2 \ln L(\theta)}{\partial \theta \partial \theta'} \right] = I(\theta),$$

which also holds for the log-likelihood function defined at the  $t^{\text{th}}$  observation,  $\ln L_t(\theta)$ , so that

$$E \left[ \frac{\partial \ln L_t(\theta)}{\partial \theta} \frac{\partial \ln L_t(\theta)}{\partial \theta'} \right] = -E \left[ \frac{\partial^2 \ln L_t(\theta)}{\partial \theta \partial \theta'} \right]. \quad (46)$$

The BHHH algorithm replaces the expectation on the left hand side of equation (46) by the sample average of the gradient of the log-likelihood function evaluated at each observation. Let the gradient vector at the  $t^{\text{th}}$  observation be

$$G_t = \frac{\partial \ln L_t(\theta)}{\partial \theta}.$$

The relevant expectation in equation (46) is now taken to be

$$E \left[ \frac{\partial \ln L_t(\theta)}{\partial \theta} \frac{\partial \ln L_t(\theta)}{\partial \theta'} \right] = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T G_t G_t'. \quad (47)$$

This implies that for the full log-likelihood function the expectation of the outer product of gradients (OPG) matrix is

$$E \left[ \frac{\partial \ln L(\theta)}{\partial \theta} \frac{\partial \ln L(\theta)}{\partial \theta'} \right] = \lim_{T \rightarrow \infty} \sum_{t=1}^T G_t G_t' . \quad (48)$$

In practice the sample size  $T$  is finite, in which case the OPG matrix is computed as

$$J(\theta) = \sum_{t=1}^T \frac{\partial \ln L_t(\theta)}{\partial \theta} \frac{\partial \ln L_t(\theta)}{\partial \theta'} . \quad (49)$$

The BHHH algorithm is thus obtained by replacing  $I(\theta)$  in (44) by  $J(\theta)$  to yield

$$\theta_{(k)} = \theta_{(k-1)} + J_{k-1}^{-1} G_{(k-1)} . \quad (50)$$

The BHHH algorithm just requires the gradients of the log-likelihood function and is therefore relatively easy to implement. A potential advantage of the BHHH algorithm, in econometric problems, is that the OPG approximation to the Information matrix and hence the Hessian matrix is guaranteed to be positive semi-definite. The cost of using this algorithm, however, is that it may require more iterations than the Newton-Raphson and scoring algorithms as there is a loss of information due to approximating the information matrix by the OPG matrix.

Another useful way to think about the structure of the BHHH algorithm is as follows. Consider the  $(T \times K)$  matrix,  $X$ , whose elements are the derivatives of the log-likelihood function at each observation with respect to each parameter in  $\theta$ , and the  $(T \times 1)$  vector of ones,  $Y$ ,

$$X = \begin{bmatrix} \frac{\partial \ln L_1(\theta)}{\partial \theta_1} & \frac{\partial \ln L_1(\theta)}{\partial \theta_2} & \dots & \frac{\partial \ln L_1(\theta)}{\partial \theta_K} \\ \frac{\partial \ln L_2(\theta)}{\partial \theta_1} & \frac{\partial \ln L_2(\theta)}{\partial \theta_2} & \dots & \frac{\partial \ln L_2(\theta)}{\partial \theta_K} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \ln L_T(\theta)}{\partial \theta_1} & \frac{\partial \ln L_T(\theta)}{\partial \theta_2} & \dots & \frac{\partial \ln L_T(\theta)}{\partial \theta_K} \end{bmatrix}, \quad Y = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} .$$

An iteration of the BHHH algorithm may now be written as

$$\theta_{(k)} = \theta_{(k-1)} + (X'_{(k-1)} X_{(k-1)})^{-1} X'_{(k-1)} Y , \quad (51)$$

where

$$J_{(k-1)}^{-1} = X'_{(k-1)}X_{(k-1)}^{-1} \quad G_{(k-1)} = X'_{(k-1)}Y.$$

The second term on the right hand side of equation (51) represents an ordinary least squares regression where the dependent variable  $Y$  is regressed on the explanatory variables given by the matrix of gradients,  $X_{(k-1)}$ , evaluated at  $\theta_{(k-1)}$ .

### 2.3 Quasi-Newton Methods

The distinguishing feature of the Newton algorithms is that they all attempt to compute or approximate the Hessian directly. An alternative approach is to build up an estimate of the Hessian at each iteration, starting from an initial estimate which is known to be negative definite, usually taken to be the negative of the identity matrix. This class of algorithm is known as Quasi-Newton. The general form for the updating sequence of the Hessian is

$$H_{(k)} = H_{(k-1)} + U_{(k-1)},$$

where  $H_{(k)}$  is the estimate of the Hessian at the  $k^{\text{th}}$  iteration and  $U_{(k)}$  is an update matrix. The quasi-Newton algorithms differ only in their choice of this update matrix. One of the more important variants is the BFGS algorithm.

In the BFGS algorithm, the update at each iteration is given by

$$U_{(k-1)} = -\frac{H_{(k-1)}\Delta_{\theta}\Delta'_G + \Delta_G\Delta'_\theta H_{(k-1)}}{\Delta'_G\Delta_{\theta}} + \left(1 + \frac{\Delta'_\theta H_{(k-1)}\Delta_{\theta}}{\Delta'_G\Delta_{\theta}}\right) \frac{\Delta_G\Delta'_G}{\Delta'_G\Delta_{\theta}},$$

where

$$\begin{aligned} \Delta_{\theta} &= \theta_{(k)} - \theta_{(k-1)}, \\ \Delta_G &= G_{(k)} - G_{(k-1)}, \end{aligned}$$

represent the changes in the parameter values and the gradients between iterations respectively. To highlight the properties of this scheme for updating the Hessian, consider the one parameter case where all terms are scalars. In this situation, the Hessian reduces to

$$H_{(k)} = \frac{\Delta_G}{\Delta_{\theta}} = \frac{G_{(k)} - G_{(k-1)}}{\theta_{(k)} - \theta_{(k-1)}},$$

which is simply the numerical first derivative of the gradient. Numerical derivatives are discussed in more detail below. For the initial iterations, the numerical approximation will

be crude as the size of this step,  $\Delta_\theta$ , is potentially large. As the iterations progress, the step interval will diminish as the algorithm approaches the maximum likelihood estimate resulting in an improvement in the accuracy of the numerical derivative.

## 2.4 Line Searching

In the optimization algorithms discussed so far, gradient,  $G_{(k-1)}$ , is multiplied by the inverse of the Hessian,  $H_{(k-1)}^{-1}$ , with the result added to  $\theta_{(k-1)}$  to generate the updated value  $\theta_{(k)}$ . One problem with this simple iterative scheme is that the updated parameter estimates are not guaranteed to improve the log-likelihood.

To ensure that the log-likelihood increases at each iteration, the algorithm is now augmented by a parameter,  $\lambda$ , that controls the size of updating at each step

$$\theta_{(k)} = \theta_{(k-1)} - \lambda H_{(k-1)}^{-1} G_{(k-1)}, \quad \lambda \in [0, 1]. \quad (52)$$

For  $\lambda = 1$ , the full step is taken so updating is as before, whereas for smaller values of  $\lambda$ , updating is not based on the full step. Determining the optimal value of  $\lambda$  at each iteration is a one-dimensional optimization problem, known as line searching.

The simplest way to choose  $\lambda$ , is to perform a coarse grid search over possible values for  $\lambda$ , known as squeezing. Potential choices of  $\lambda$  are as follows

$$\lambda = 1, \lambda = \frac{1}{2}, \lambda = \frac{1}{3}, \lambda = \frac{1}{4}, \dots$$

The strategy is to calculate  $\theta_{(k+1)}$  for  $\lambda = 1$  initially and check to see if

$$\ln L(\theta_{(k)}) > \ln L(\theta_{(k-1)}).$$

If this condition is not satisfied, choose  $\lambda = \frac{1}{2}$  and test to see if there is an improvement in the log-likelihood function. If there is still no improvement, then choose  $\lambda = \frac{1}{3}$  and repeat the function evaluation. Once a value of  $\lambda$  is chosen and an updated parameter value is computed, the procedure begins again at the next step with  $\lambda = 1$ .

## 2.5 Simplex algorithm

The most commonly-used algorithm for function optimization based on function evaluation is the downhill-simplex method Consider a function  $F$  in  $n$  dimensions which may be

evaluated at the following  $n + 1$  vertices,  $x_1, x_2 \dots x_{n+1}$  where each vertex has the required  $n$  dimensions. At each iteration of the algorithm  $n + 1$  points,  $x_1, x_2 \dots x_{n+1}$  are retained together with the value of the function at these points, which are ordered so that

$$F(x_{n+1}) \geq F(x_n) \geq \dots \geq F(x_1)$$

and the object is to replace the worst point  $x_{n+1}$ .

The basic iteration of the simplex algorithm consists of a few simple steps.

1. Define

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i.$$

as the centroid of the best  $n$  vertices. In a two dimensional problem this point would merely be the midpoint of the line joining the two best vertices of the current simplex.

2. Construct a new point

$$x_r = \bar{x} + \alpha(\bar{x} - x_{n+1}) \quad \alpha > 0.$$

This move is known as a *reflection*. This new trial point is generated by reflecting the worst vertex through the opposite face of the simplex.

3. If  $F(x_r) < F(x_{n+1})$  and the original reflection is successful, the reflected point,  $x_r$ , replaces the previous worst point,  $x_{n+1}$ , and the next iteration is started.
4. If  $F(x_r) < F(x_1)$  and  $x_r$  is in fact the best point, an additional reflection defined by

$$x_e = \bar{x} + \beta(x_r - \bar{x}) \quad \beta > 0$$

is tried. If  $F(x_e) < F(x_r)$ , then  $x_e$  replaces  $x_{n+1}$  and the next iteration is started.

5. If  $F(x_r) > F(x_n)$  and the original reflection is not successful, a contraction step is carried out. There are two forms of contraction which may be tried. The first of these computes

$$x_c = \begin{cases} \bar{x} + \gamma(x_{n+1} - \bar{x}) & \text{if } F(x_r) \geq F(x_{n+1}) \\ \bar{x} + \gamma(x_r - \bar{x}) & \text{if } F(x_r) < F(x_{n+1}) \end{cases}.$$

The second contraction method uses the current best point as the focal point of the contraction and  $x_c$  is computed as follows

$$x_c = \begin{cases} x_1 + \gamma(x_{n+1} - x_1) & \text{if } F(x_r) \geq F(x_{n+1}) \\ x_1 + \gamma(x_r - x_1) & \text{if } F(x_r) < F(x_{n+1}) \end{cases} .$$

If the contraction is successful  $x_c$  replaces  $x_{n+1}$  and the next iteration is started.

6. If the contraction is not successful the simplex algorithm performs a final adjustment before starting a new iteration. To try and prevent the algorithm from getting stuck and keeping the best  $n$  vertices until the number of function evaluations set by the user is reached, a simple but effective move is to shrink the vertices of the simplex half-way toward the current best point. At this stage as all the vertices of the simplex have been altered, the next iteration is started.

This simple sequence of moves proves to be remarkably good at isolating the minimum of a given function. Upon reaching its destination the simplex will shrink around the best vertex until the termination criterion is triggered.

## 2.6 Choice of Algorithm

In theory there is little to choose between the gradient algorithms discussed in this chapter, because in the vicinity of a minimum each should enjoy quadratic convergence, which means that

$$\|\theta_{(k+1)} - \theta\| < \kappa \|\theta_{(k)} - \theta\|^2 \quad \kappa > 0 .$$

In other words, if  $\theta_{(k)}$  is accurate to 2 decimal places, then it is anticipated that  $\theta_{(k+1)}$  will be accurate to 4 decimal places and that  $\theta_{(k+2)}$  will be accurate to 8 decimal places and so on. In choosing an algorithm, however, there are a few practical considerations to bear in mind.

- (1) The Newton-Raphson and the Method of Scoring require the first two derivatives of the log-likelihood function. Because the Information matrix is the expected value of the negative Hessian matrix, it is problem specific and typically is not easy to compute. Consequently, the Method of Scoring is largely of theoretical interest.

- (2) Close to the minimum Newton-Raphson converges quadratically but further away from the minimum, the Hessian matrix may not be positive definite and this may cause the algorithm to become unstable.
- (3) BHHH provides an estimate of the Hessian that is guaranteed to be positive definite and is therefore a popular choice in econometrics.
- (4) The current consensus is that quasi-Newton algorithms that construct approximations to the Hessian matrix over successive iterations are the preferred choice. Specifically, the Hessian update of the BFGS algorithm is particularly robust and therefore most optimization toolboxes will provide BFGS as an option, very often as the default choice.

Practical optimization problems frequently throw up irregular surfaces to the target function being minimized. In particular, if the gradient is nearly flat in several dimensions, numerical errors can cause a gradient algorithm to misbehave. As a consequence, there are now very sophisticated optimization routines available which are based solely on function evaluation. These algorithms are fairly robust, but, in general they are more inefficient than gradient-based algorithms, requiring many more function evaluations to isolate the optimum.

A popular practical strategy is to use the simplex method to start the numerical optimization process. After a few iterations, BFGS can be employed to speed up convergence. Note, however, that if the BFGS algorithm terminates too soon after taking over from the simplex, the approximation to the inverse of the Hessian is potentially unreliable. This can have serious consequences for the quality of the estimated standard errors of the maximum likelihood estimators.

## 2.7 Computing Standard Errors

The asymptotic distribution of the maximum likelihood estimator is given by

$$\hat{\theta} \xrightarrow{a} N(\theta_0, I(\theta_0)^{-1}).$$

In practice, the covariance matrix of the maximum likelihood estimator is computed by replacing  $\theta_0$  by  $\hat{\theta}$  and inverting the information matrix

$$\text{var}(\hat{\theta}) = I^{-1}(\hat{\theta}) = \left( - E \left[ \frac{\partial^2 \ln L(\theta)}{\partial \theta \partial \theta'} \right]_{\theta=\hat{\theta}} \right)^{-1}. \quad (53)$$

The standard error of each element of  $\hat{\theta}$  is therefore given by the square root of the main-diagonal entries of this matrix.

In most practical situations, the expected value of the second-derivatives of the log-likelihood function cannot easily be evaluated. A more common approach, therefore, is simply to use the negative of the inverse Hessian, rather than its expectation, that is,

$$\text{var}(\hat{\theta}) = -H^{-1}(\hat{\theta}) = - \left( \frac{\partial^2 \ln L(\theta)}{\partial \theta \partial \theta'} \right)_{\theta=\hat{\theta}}^{-1}. \quad (54)$$

If the Hessian is not negative-definite at the maximum likelihood estimator, computation of the standard errors from equation (54) is not possible. A popular alternative is to use the OPG matrix,  $J(\hat{\theta})$  from equation (49), instead of the negative of the Hessian, is

$$\text{var}(\hat{\theta}) = J^{-1}(\hat{\theta}). \quad (55)$$

A situation which arises often in practice is the need to estimate the covariance matrix of a non-linear function of the maximum likelihood estimators, say  $C(\theta)$ . There are two approaches to this problem. The first approach simply imposes the nonlinearity and then uses the constrained likelihood to compute standard errors based on the OPG approximation. The second approach is the so-called delta method. The delta method uses a Taylor series expansion of  $C(\theta)$  around the maximum likelihood estimator,  $\hat{\theta}$ , which is truncated at the first-order term

$$C(\theta) \simeq C(\hat{\theta}) + \frac{\partial C}{\partial \theta}(\theta - \hat{\theta}), \quad (56)$$

where it is understood that the first derivative is evaluated at  $\hat{\theta}$ . After re-arranging, squaring and taking expectations, it follows that

$$E \left[ (C(\theta) - C(\hat{\theta}))^2 \right] = E \left[ \frac{\partial C}{\partial \theta}(\theta - \hat{\theta})(\theta - \hat{\theta})' \left( \frac{\partial C}{\partial \theta} \right)' \right]. \quad (57)$$

**EXAMPLE: Standard Error of a Nonlinear Function of the Maximum Likelihood Estimator**

This example demonstrates the equivalence of the outer product of the gradient and delta methods for computing standard errors. Consider the problem of finding the standard error for  $\bar{y}^2$ , where  $\hat{\theta} = \bar{y}$  and the observations are drawn from  $N(\theta, \sigma^2)$ .

### 1. Outer Product of Gradient

Consider the log-likelihood at time  $t$  for the unconstrained problem

$$\ln L_t = -\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} (y_t - \theta)^2.$$

Now define the parameter

$$\psi = \theta^2.$$

Rewriting this expression in terms of  $\psi$  gives

$$\ln L_t = -\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} (y_t - \psi^{1/2})^2.$$

The first derivative is

$$\frac{d \ln L_t}{d\psi} = \frac{1}{\sigma^2} (y_t - \psi^{1/2}) \left( \frac{1}{2} \psi^{-1/2} \right).$$

Squaring and taking expectations yields

$$\begin{aligned} E \left[ \left( \frac{d \ln L_t}{d\psi} \right)^2 \right] &= E \left[ \left( \frac{1}{\sigma^2} (y_t - \psi^{1/2}) \left( \frac{1}{2} \psi^{-1/2} \right) \right)^2 \right] \\ &= \frac{1}{4\sigma^4\psi} E \left[ (y_t - \psi^{1/2})^2 \right] \\ &= \frac{1}{4\sigma^4\psi} \sigma^2 = \frac{1}{4\sigma^2\psi} = \frac{1}{4\sigma^2\theta^2}. \end{aligned}$$

The variance is then

$$\text{var}(\hat{\psi}) = \left( E \left[ \left( \frac{d \ln L_t}{d\psi} \right)^2 \right] \right)^{-1} = 4\sigma^2\theta^2.$$

### 2. Delta Method

Defining

$$C(\theta) = \theta^2,$$

then

$$\text{var}(\hat{\psi}) = \left( \frac{dC}{d\theta} \right)^2 \text{Var}(\hat{\theta}) = (2\theta)^2 \sigma^2 = 4\sigma^2\theta^2.$$

This agrees with the expression of the variance obtained previously in (i).

## 2.8 Parameter Constraints

In some econometric applications, the values of the parameters need to be constrained to lie within certain intervals. Some examples are as follows.

1. The variance is required to be positive:

$$\sigma^2 > 0.$$

2. The marginal propensity to consume is constrained to be in the unit interval:

$$0 < MPC < 1.$$

3. For an MA(1) process to be invertible, the moving average parameter must lie within the unit interval:

$$|\theta| < 1.$$

4. The degrees of freedom parameter in the Student t distribution must be greater than 2, to ensure that the variance of the distribution exists:

$$\gamma > 2.$$

5. In a model based on the Poisson distribution, the mean parameter must be positive:

$$\theta > 0.$$

Consider the case of estimating a single parameter  $\theta$  where  $\theta \in (a, b)$ . The key idea is to transform the parameter  $\theta$  by means of a non-linear bijective mapping,  $\phi = g(\theta)$ , between the constrained interval  $(a, b)$  and the real line. Thus each and every value of  $\phi$  corresponds to a unique value of  $\theta$  satisfying the desired constraint, and obtained by applying the inverse transform  $\theta = g^{-1}(\phi)$ . In other words, when the minimization algorithm returns  $\hat{\phi}$ , the value of  $\phi$  that minimizes the objective function, the associated estimate of  $\theta$  is given by  $\hat{\theta} = g^{-1}(\hat{\phi})$ . Some useful one-dimensional transformations, their associated inverse functions and the gradients of the transformations are presented in Table 2. Note, however, that the application of the bijective mapping alters the scale of the variable being transformed and therefore its gradient. This has important implications for the computation of standard errors.

Table 2:

Some useful transformations for imposing constraints on  $\theta$ .

Constraint	Transform $\phi = g(\theta)$	Inverse Transform $\theta = g^{-1}(\phi)$	Jacobian $d\phi/d\theta$
$(0, \infty)$	$\phi = \log \theta$	$\theta = e^\phi$	$\frac{d\phi}{d\theta} = \frac{1}{\theta}$
$(-\infty, 0)$	$\phi = \log(-\theta)$	$\theta = -e^\phi$	$\frac{d\phi}{d\theta} = \frac{1}{\theta}$
$(0, 1)$	$\phi = \log\left(\frac{\theta}{1-\theta}\right)$	$\theta = \frac{1}{1+e^{-\phi}}$	$\frac{d\phi}{d\theta} = \frac{1}{\theta(1-\theta)}$
$(0, \pi)$	$\phi = \log\left(\frac{\theta}{\pi-\theta}\right)$	$\theta = \frac{\pi}{1+e^{-\phi}}$	$\frac{d\phi}{d\theta} = \frac{\pi}{\theta(\pi-\theta)}$
$(a, b)$	$\phi = \log\left(\frac{\theta-a}{b-\theta}\right)$	$\theta = \frac{b+ae^{-\phi}}{1+e^{-\phi}}$	$\frac{d\phi}{d\theta} = \frac{b-a}{(\theta-a)(b-\theta)}$
$(-1, 1)$	$\phi = \operatorname{atanh}(\theta)$	$\theta = \tanh(\phi)$	$\frac{d\phi}{d\theta} = \frac{1}{1-\theta^2}$
$(-1, 1)$	$\phi = \frac{\theta}{1- \theta }$	$\theta = \frac{\phi}{1+ \phi }$	$\frac{d\phi}{d\theta} = \frac{1}{(1- \theta )^2}$
$(-1, 1)$	$\phi = \tan\left(\frac{\pi\theta}{2}\right)$	$\theta = \frac{2}{\pi} \tan^{-1} \phi$	$\frac{d\phi}{d\theta} = \frac{\pi}{2} \sec^2\left(\frac{\pi\theta}{2}\right)$

Self-evidently the choice of mapping used will be problem specific, but it is useful to note the following.

- (1) The log transform is particularly useful in problems where the parameters are required to be positive. Note that in many cases not only  $d\phi/d\theta$  is required but also  $d\theta/d\phi$  and in the case of the log transform, the latter is simply the parameter  $\theta$  itself.
- (2) A popular mapping used to constrain a parameter to the interval  $(-1, 1)$  is

$$\phi = \frac{\theta}{1-|\theta|}.$$

This is fundamentally different to all of the other mappings listed in Table 2 in the respect that its second derivative is not defined at  $\theta = 0$ .

The price to be paid for the convenience of using an unconstrained minimisation algorithm on what is essentially a constrained problem is that the Hessian matrix returned by the algorithm cannot be used immediately to obtain the standard errors of the model parameters simply by taking the square roots of the diagonal elements of the inverse

Hessian matrix. In this situation, there are two ways of obtaining standard errors for the parameters.

- (1) A straightforward way to compute standard errors is to express the objective function in terms of the true parameters  $\theta$ . The gradient vector and Hessian matrix can then be computed numerically at the minimum using the estimated values of the parameters. Because no searching is involved and the perturbation of the optimal parameters used in the computation of the finite differences is small, the probability of numerical instability is much reduced.
- (2) Alternatively, the Delta method can be used. The Jacobian matrix of the transformation from  $\theta$  to  $\phi$ , in this instance evaluated at the minimum of the objective function is

$$J = \begin{bmatrix} \frac{\partial \phi_1}{\partial \theta_1} & \frac{\partial \phi_1}{\partial \theta_2} & \frac{\partial \phi_1}{\partial \theta_3} & \cdots & \frac{\partial \phi_1}{\partial \theta_K} \\ \frac{\partial \phi_2}{\partial \theta_1} & \frac{\partial \phi_2}{\partial \theta_2} & \frac{\partial \phi_2}{\partial \theta_3} & \cdots & \frac{\partial \phi_2}{\partial \theta_K} \\ \frac{\partial \phi_3}{\partial \theta_1} & \frac{\partial \phi_3}{\partial \theta_2} & \frac{\partial \phi_3}{\partial \theta_3} & \cdots & \frac{\partial \phi_3}{\partial \theta_K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \phi_n}{\partial \theta_1} & \frac{\partial \phi_n}{\partial \theta_2} & \frac{\partial \phi_n}{\partial \theta_3} & \cdots & \frac{\partial \phi_n}{\partial \theta_K} \end{bmatrix}.$$

The relationship between  $H_\theta$ , the Hessian matrix with respect to the parameters of the model and  $H_\phi$ , the Hessian matrix evaluated at the mapped parameters, is therefore given by

$$H_\theta = J' H_\phi J,$$

and the square of the diagonal elements of the inverse of this matrix provide estimates of the standard errors of the model parameters,  $\theta$ .

## 2.9 Maximum Likelihood Estimation of Nonlinear Regression Models

A typical form for the nonlinear regression model is

$$g(y_t; \alpha) = \mu(x_t; \beta) + u_t, \tag{58}$$

where  $y_t$  is the dependent variable,  $x_t$  is the independent variable and  $u_t$  is a disturbance term distributed as  $N(0, \sigma^2)$ . The nonlinear functions  $g(\cdot)$  and  $\mu(\cdot)$  of  $y_t$  and  $x_t$ , respec-

tively have parameter vectors  $\alpha = \{\alpha_0, \dots, \alpha_m\}$  and  $\beta = \{\beta_0, \dots, \beta_k\}$ . Let the unknown parameters to be estimated be given by the  $(m + k + 1)$  vector  $\theta = \{\alpha, \beta, \sigma^2\}$ .

An example of a nonlinear regression model is the Zellner-Revankar production function

$$\ln y_t + \alpha y_t = \beta_0 + \beta_1 \ln k_t + \beta_2 \ln l_t + u_t,$$

where  $y_t$  is output,  $k_t$  is capital,  $l_t$  is labour, and the relevant functions in (58) are

$$\begin{aligned} g(y_t; \alpha) &= \ln y_t + \alpha y_t, \\ \mu(x_t; \beta) &= \beta_0 + \beta_1 \ln k_t + \beta_2 \ln l_t. \end{aligned}$$

Another example is the following exponential regression model

$$y_t = \beta_0 \exp [\beta_1 x_t] + u_t, \tag{59}$$

where

$$\begin{aligned} g(y_t; \alpha) &= y_t, \\ \mu(x_t; \beta) &= \beta_0 \exp [\beta_1 x_t]. \end{aligned}$$

An alternative exponential regression model is

$$y_t = \beta_0 \exp [\beta_1 x_t + u_t]. \tag{60}$$

In contrast to equation (59) this model is not intrinsically nonlinear, as it can be transformed as

$$\ln y_t = \ln \beta_0 + \beta_1 x_t + u_t,$$

which is linear in the transformed variable while  $u_t$  is normally distributed. Despite having similar specifications, they nonetheless have very different time-series properties as the following example demonstrates.

The iterative algorithms discussed previously can be utilized to find the maximum likelihood estimates of the parameters of the nonlinear regression model in equation (58), together with their standard errors. Typically, the specification assumes that the nonlinear regression disturbance,  $u_t$ , is normally distributed

$$f(u_t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{u_t^2}{2\sigma^2} \right]. \tag{61}$$

In order to derive the corresponding density of  $y_t$ , it is necessary to use the transformation of variable technique. The density of  $y_t$  is given by

$$f(y_t) = f(u_t) \left| \frac{du_t}{dy_t} \right|. \quad (62)$$

Taking the derivative with respect to  $y_t$  on both sides of equation (58) gives

$$\frac{du_t}{dy_t} = \frac{dg(y_t; \alpha)}{dy_t},$$

so the density of  $y_t$  is

$$f(y_t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{(g(y_t; \alpha) - \mu(x_t; \beta))^2}{2\sigma^2} \right] \left| \frac{dg(y_t; \alpha)}{dy_t} \right|.$$

The log-likelihood function for observation  $t$  is

$$\ln L_t = -\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln(\sigma^2) - \frac{(g(y_t; \alpha) - \mu(x_t; \beta))^2}{2\sigma^2} + \ln \left| \frac{dg(y_t; \alpha)}{dy_t} \right|,$$

and the log-likelihood function for  $t = 1, 2, \dots, T$  observations, is

$$\begin{aligned} \ln L &= \sum_{t=1}^T \ln L_t \\ &= -\frac{T}{2} \ln(2\pi) - \frac{T}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{t=1}^T (g(y_t; \alpha) - \mu(x_t; \beta))^2 \\ &\quad + \sum_{t=1}^T \ln \left| \frac{dg(y_t; \alpha)}{dy_t} \right|. \end{aligned}$$

This function is maximized with respect to the unknown parameters  $\theta = \{\alpha, \beta, \sigma^2\}$ . The first-order conditions for a maximum are

$$\frac{\partial \ln L}{\partial \alpha} = -\frac{1}{\sigma^2} \sum_{t=1}^T (g(y_t; \alpha) - \mu(x_t; \beta)) \frac{\partial g(y_t; \alpha)}{\partial \alpha} + \sum_{t=1}^T \frac{\partial}{\partial \alpha} \ln \left| \frac{dg(y_t; \alpha)}{dy_t} \right|$$

$$\frac{\partial \ln L}{\partial \beta} = \frac{1}{\sigma^2} \sum_{t=1}^T (g(y_t; \alpha) - \mu(x_t; \beta)) \frac{\partial \mu(x_t; \beta)}{\partial \beta}$$

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{T}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{t=1}^T (g(y_t; \alpha) - \mu(x_t; \beta))^2.$$

Setting

$$\frac{\partial \ln L}{\partial \sigma^2} \Big|_{\sigma^2 = \hat{\sigma}^2} = 0,$$

and solving for  $\hat{\sigma}^2$  yields

$$\hat{\sigma}^2 = \frac{1}{T} \sum_{t=1}^T (g(y_t; \alpha) - \mu(x_t; \hat{\beta}))^2, \quad (63)$$

which is used to concentrate  $\hat{\sigma}^2$  out of the log-likelihood function, which in turn, is then maximized with respect to  $\alpha$  and  $\beta$  only.

---

### ASIDE: Block Diagonality and the Information Matrix

The asymptotic distribution of the maximum likelihood estimator is given by

$$\hat{\theta} \xrightarrow{a} N\left(\theta_0, \frac{1}{I(\theta_0)}\right) \quad \text{univariate}$$

$$\hat{\theta} \xrightarrow{a} N\left(\theta_0, I(\theta_0)^{-1}\right) \quad \text{multivariate,}$$

where  $I(\theta_0)^{-1}$  is the inverse of the information matrix and the standard errors of the parameters are obtained by taking the square root of the diagonal elements. In practice, this requires that all the elements of the information matrix need to be computed and this matrix inverted in order to obtain a standard error for any given parameter. In other words, asymptotic efficiency requires that all the parameters are estimated simultaneously.

There is an important instance where asymptotic efficiency does not require the simultaneous estimation of all the parameters of a the model. This instance will be illustrated below with reference to some simple rules for matrix inversion.

Consider a two parameter case  $\theta = \{\theta_1, \theta_2\}$ , the information matrix is

$$I(\theta) = \text{E} \begin{bmatrix} \frac{\partial \ln L(\theta)}{\partial \theta_1 \partial \theta_1} & \frac{\partial \ln L(\theta)}{\partial \theta_1 \partial \theta_2} \\ \frac{\partial \ln L(\theta)}{\partial \theta_2 \partial \theta_1} & \frac{\partial \ln L(\theta)}{\partial \theta_2 \partial \theta_2} \end{bmatrix} = \begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix},$$

where, purely for notational convenience, the fact that the matrix will be symmetric has been ignored. The inverse of the matrix is given by

$$I(\theta)^{-1} = \frac{1}{|I(\theta)|} \begin{bmatrix} I_{22} & -I_{12} \\ -I_{21} & I_{11} \end{bmatrix}$$

where the relevant determinant is

$$|I(\theta)| = I_{11}I_{22} - I_{12}I_{21}.$$

The asymptotic distribution of  $\theta$  requires the inversion of the full matrix and hence estimates of all of the parameters in  $\theta$ .

The operation of taking the inverse rapidly increases in difficulty. For the three parameter case  $\theta = \{\theta_1, \theta_2, \theta_3\}$ , the information matrix is

$$I(\theta) = E \begin{bmatrix} \frac{\partial \ln L(\theta)}{\partial \theta_1 \partial \theta_1} & \frac{\partial \ln L(\theta)}{\partial \theta_1 \partial \theta_2} & \frac{\partial \ln L(\theta)}{\partial \theta_1 \partial \theta_3} \\ \frac{\partial \ln L(\theta)}{\partial \theta_2 \partial \theta_1} & \frac{\partial \ln L(\theta)}{\partial \theta_2 \partial \theta_2} & \frac{\partial \ln L(\theta)}{\partial \theta_2 \partial \theta_3} \\ \frac{\partial \ln L(\theta)}{\partial \theta_3 \partial \theta_1} & \frac{\partial \ln L(\theta)}{\partial \theta_3 \partial \theta_2} & \frac{\partial \ln L(\theta)}{\partial \theta_3 \partial \theta_3} \end{bmatrix} = \begin{bmatrix} I_{11} & I_{12} & I_{13} \\ I_{21} & I_{22} & I_{23} \\ I_{31} & I_{32} & I_{33} \end{bmatrix}$$

and the inverse is obtained by computing

$$I(\theta)^{-1} = \frac{1}{|I(\theta)|} \begin{bmatrix} \left| \begin{array}{cc|cc} I_{22} & I_{23} & | & I_{13} & I_{12} \\ I_{32} & I_{33} & | & I_{33} & I_{32} \end{array} \right| & \left| \begin{array}{cc|cc} I_{12} & I_{13} \\ I_{22} & I_{23} \end{array} \right| \\ \left| \begin{array}{cc|cc} I_{23} & I_{21} \\ I_{33} & I_{31} \end{array} \right| & \left| \begin{array}{cc|cc} I_{11} & I_{13} \\ I_{31} & I_{33} \end{array} \right| & \left| \begin{array}{cc|cc} I_{13} & I_{11} \\ I_{23} & I_{21} \end{array} \right| \\ \left| \begin{array}{cc|cc} I_{21} & I_{22} \\ I_{31} & I_{32} \end{array} \right| & \left| \begin{array}{cc|cc} I_{12} & I_{11} \\ I_{32} & I_{31} \end{array} \right| & \left| \begin{array}{cc|cc} I_{11} & I_{12} \\ I_{21} & I_{22} \end{array} \right| \end{bmatrix}$$

where the determinant,  $|I(\theta)|$ , is

$$|I(\theta)| = I_{11}I_{22}I_{33} - I_{11}I_{23}I_{32} - I_{12}I_{21}I_{33} + I_{12}I_{31}I_{23} + I_{21}I_{13}I_{32} - I_{13}I_{22}I_{31}.$$

Completing the computation is a tedious task, but the general idea is once again that computing and inverting the information matrix will require knowledge of all the parameters.

The situation is eased slightly if the information matrix is a block diagonal matrix

$$I(\theta) = E \begin{bmatrix} \frac{\partial \ln L(\theta)}{\partial \theta_1 \partial \theta_1} & \frac{\partial \ln L(\theta)}{\partial \theta_1 \partial \theta_2} & 0 \\ \frac{\partial \ln L(\theta)}{\partial \theta_2 \partial \theta_1} & \frac{\partial \ln L(\theta)}{\partial \theta_2 \partial \theta_2} & 0 \\ 0 & 0 & \frac{\partial \ln L(\theta)}{\partial \theta_3 \partial \theta_3} \end{bmatrix} = \begin{bmatrix} I_{11} & I_{12} & 0 \\ I_{21} & I_{22} & 0 \\ 0 & 0 & I_{33} \end{bmatrix}$$

In this case the inverse is

$$I(\theta)^{-1} = \begin{bmatrix} \frac{I_{22}}{I_{11}I_{22} - I_{12}I_{21}} & -\frac{I_{12}}{I_{11}I_{22} - I_{12}I_{21}} & 0 \\ -\frac{I_{12}}{I_{11}I_{22} - I_{12}I_{21}} & \frac{I_{11}}{I_{11}I_{22} - I_{12}I_{21}} & 0 \\ 0 & 0 & \frac{1}{I_{33}} \end{bmatrix}.$$

In other words, the inverse of the block diagonal matrix may be obtained by inverting the  $(2 \times 2)$  matrix in the upper left corner of  $I(\theta)$  and inverting  $I_{33}$ . Now note that the standard error of the parameter  $\theta_3$  is not influenced by either  $\theta_1$  or  $\theta_2$ . In

other words the asymptotic distribution of  $\theta_3$  is not dependent on knowing  $\theta_1$  and  $\theta_2$ . Similarly, the asymptotic distributions of  $\theta_1$  and  $\theta_2$  do not depend on  $\theta_3$ . The block diagonal information matrix implies that the parameter estimation problem is separable:  $\theta_1$  and  $\theta_2$  may be estimated separately from  $\theta_3$  without any loss of asymptotic efficiency.

---

To estimate the model by Newton-Raphson the second derivatives of the log-likelihood are needed

$$\begin{aligned} \frac{\partial^2 \ln L}{\partial \alpha \partial \alpha'} &= -\frac{1}{\sigma^2} \sum_{t=1}^T (g(y_t; \alpha) - \mu(x_t; \beta)) \frac{\partial g(y_t; \alpha)}{\partial \alpha \partial \alpha'} - \frac{1}{\sigma^2} \sum_{t=1}^T \left( \frac{\partial g(y_t; \alpha)}{\partial \alpha \partial \alpha'} \right)^2 \\ &\quad + \sum_{t=1}^T \frac{\partial^2}{\partial \alpha \partial \alpha'} \ln \left| \frac{dg(y_t; \alpha)}{dy_t} \right| \\ \frac{\partial^2 \ln L}{\partial \alpha \partial \beta'} &= \frac{1}{\sigma^2} \sum_{t=1}^T (g(y_t; \alpha) - \mu(x_t; \beta)) \frac{\partial g(y_t; \alpha)}{\partial \alpha} \frac{\partial \mu(x_t; \beta)}{\partial \beta'} \\ \frac{\partial^2 \ln L}{\partial \beta \partial \beta'} &= \frac{1}{\sigma^2} \sum_{t=1}^T (g(y_t; \alpha) - \mu(x_t; \beta)) \frac{\partial^2 \mu(x_t; \beta)}{\partial \beta \partial \beta'} - \frac{1}{\sigma^2} \sum_{t=1}^T \frac{\partial^2 \mu(x_t; \beta)}{\partial \beta \partial \beta'} \\ \frac{\partial^2 \ln L}{\partial (\sigma^2)^2} &= -\frac{T}{2\sigma^4} + \frac{1}{\sigma^6} \sum_{t=1}^T (g(y_t; \alpha) - \mu(x_t; \beta))^2 \\ \frac{\partial^2 \ln L}{\partial \alpha \partial \sigma^2} &= \frac{1}{\sigma^4} \sum_{t=1}^T (g(y_t; \alpha) - \mu(x_t; \beta)) \frac{\partial g(y_t; \alpha)}{\partial \alpha} \\ \frac{\partial^2 \ln L}{\partial \beta \partial \sigma^2} &= -\frac{1}{\sigma^4} \sum_{t=1}^T (g(y_t; \alpha) - \mu(x_t; \beta)) \frac{\partial \mu(x_t; \beta)}{\partial \beta}. \end{aligned}$$

The Newton-Raphson algorithm is

$$\theta_{(k)} = \theta_{(k-1)} - H^{-1}(\theta_{(k-1)}) G(\theta_{(k-1)}), \quad (64)$$

where  $G$  and  $H$  are respectively the gradient vector and Hessian.

Alternatively, the method of scoring replaces  $-H(\theta)$  in (64), by the information matrix  $I(\theta)$ . The updated parameter vector is calculated as

$$\theta_{(k)} = \theta_{(k-1)} + I^{-1}(\theta_{(k-1)}) G(\theta_{(k-1)}), \quad (65)$$

where

$$I(\theta) = -E \begin{bmatrix} \frac{\partial^2 \ln L}{\partial \alpha \partial \alpha'} & \frac{\partial^2 \ln L}{\partial \alpha \partial \beta'} & \frac{\partial^2 \ln L}{\partial \alpha \partial \sigma^2} \\ \frac{\partial^2 \ln L}{\partial \beta \partial \alpha'} & \frac{\partial^2 \ln L}{\partial \beta \partial \beta'} & \frac{\partial^2 \ln L}{\partial \beta \partial \sigma^2} \\ \frac{\partial^2 \ln L}{\partial \sigma^2 \partial \alpha'} & \frac{\partial^2 \ln L}{\partial \sigma^2 \partial \beta'} & \frac{\partial^2 \ln L}{\partial (\sigma^2)^2} \end{bmatrix}.$$

An advantage of this algorithm is that estimation is simplified as a result of  $I(\theta)$  being block-diagonal for this class of models. To see this, note that from equation (58)

$$E[g(y_t; \alpha)] = E[\mu(x_t; \beta) + u_t] = \mu(x_t; \beta),$$

so that

$$\begin{aligned} E \left[ \frac{\partial^2 \ln L}{\partial \alpha \partial \sigma^2} \right] &= E \left[ \frac{1}{\sigma^4} \sum_{t=1}^T (g(y_t; \alpha) - \mu(x_t; \beta)) \frac{\partial g(y_t; \alpha)}{\partial \alpha} \right] = 0 \\ E \left[ \frac{\partial^2 \ln L}{\partial \beta \partial \sigma^2} \right] &= -E \left[ \frac{1}{\sigma^4} \sum_{t=1}^T (g(y_t; \alpha) - \mu(x_t; \beta)) \frac{\partial \mu(x_t; \beta)}{\partial \beta} \right] = 0. \end{aligned}$$

In which case  $I(\theta)$  reduces to

$$I(\theta) = -E \begin{bmatrix} \frac{\partial^2 \ln L}{\partial \alpha \partial \alpha'} & \frac{\partial^2 \ln L}{\partial \alpha \partial \beta'} & 0 \\ \frac{\partial^2 \ln L}{\partial \beta \partial \alpha'} & \frac{\partial^2 \ln L}{\partial \beta \partial \beta'} & 0 \\ 0 & 0 & \frac{\partial^2 \ln L}{\partial (\sigma^2)^2} \end{bmatrix} = \begin{bmatrix} I_{1,1} & 0 \\ 0 & I_{2,2} \end{bmatrix}, \quad (66)$$

where

$$I_{1,1} = -E \begin{bmatrix} \frac{\partial^2 \ln L}{\partial \alpha \partial \alpha'} & \frac{\partial^2 \ln L}{\partial \alpha \partial \beta'} \\ \frac{\partial^2 \ln L}{\partial \beta \partial \alpha'} & \frac{\partial^2 \ln L}{\partial \beta \partial \beta'} \end{bmatrix}, \quad I_{2,2} = -E \left[ \frac{\partial^2 \ln L}{\partial (\sigma^2)^2} \right].$$

The scoring algorithm now proceeds in two parts

$$\begin{bmatrix} \alpha^{(k)} \\ \beta^{(k)} \end{bmatrix} = \begin{bmatrix} \alpha^{(k-1)} \\ \beta^{(k-1)} \end{bmatrix} + I_{1,1}^{-1}(\theta_{(k-1)}) G_1(\theta_{(k-1)}) \quad (67)$$

$$\begin{bmatrix} \sigma_{(k)}^2 \end{bmatrix} = \begin{bmatrix} \sigma_{(k-1)}^2 \end{bmatrix} + I_{2,2}^{-1}(\theta_{(k-1)}) G_2(\theta_{(k-1)}), \quad (68)$$

where

$$G_1 = \begin{bmatrix} \frac{\partial \ln L}{\partial \alpha} & \frac{\partial \ln L}{\partial \beta} \end{bmatrix}', \quad G_2 = \begin{bmatrix} \frac{\partial \ln L}{\partial \sigma^2} \end{bmatrix}.$$

The variance-covariance matrix of the parameter estimators is obtained by inverting the relevant blocks of the information matrix at the last iteration. For example, the standard error of  $\hat{\sigma}^2$  is simply given by

$$se(\hat{\sigma}^2) = \sqrt{\frac{2\hat{\sigma}^4}{T}}.$$

## Zellner-Revankar Production Function

Consider the production function

$$\begin{aligned}\ln y_t + \alpha y_t &= \beta_0 + \beta_1 \ln k_t + \beta_2 \ln l_t + u_t \\ u_t &\sim N(0, \sigma^2),\end{aligned}$$

where  $y_t$  is output,  $k_t$  is capital,  $l_t$  is labour,  $u_t$  is a disturbance term and the unknown parameters are  $\theta = \{\alpha, \beta_0, \beta_1, \beta_2, \sigma^2\}$ . The probability density function of  $u_t$  is

$$f(u_t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{u_t^2}{2\sigma^2}\right].$$

using equation (62) with

$$\frac{du_t}{dy_t} = \frac{1}{y_t} + \alpha,$$

the density for  $y_t$  is

$$f(y_t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(\ln y_t + \alpha y_t - \beta_0 - \beta_1 \ln k_t - \beta_2 \ln l_t)^2}{2\sigma^2}\right] \left|\frac{1}{y_t} + \alpha\right|.$$

The log-likelihood function for observation  $t$  is

$$\begin{aligned}\ln L_t(\theta) &= \ln f(y_t) \\ &= -\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln(\sigma^2) + \ln \left|\frac{1}{y_t} + \alpha\right| \\ &\quad - \frac{(\ln y_t + \alpha y_t - \beta_0 - \beta_1 \ln k_t - \beta_2 \ln l_t)^2}{2\sigma^2},\end{aligned}$$

and the log-likelihood function for a sample of  $t = 1, 2, \dots, T$  observations, is then

$$\begin{aligned}\ln L(\theta) &= \sum_{t=1}^T \ln L_t(\theta) \\ &= -\frac{T}{2} \ln(2\pi) - \frac{T}{2} \ln(\sigma^2) + \sum_{t=1}^T \ln \left|\frac{1}{y_t} + \alpha\right| \\ &\quad - \frac{1}{2\sigma^2} \sum_{t=1}^T (\ln y_t + \alpha y_t - \beta_0 - \beta_1 \ln k_t - \beta_2 \ln l_t)^2.\end{aligned}$$

This function is then maximized with respect to the unknown parameters  $\theta = \{\alpha, \beta_0, \beta_1, \beta_2, \sigma^2\}$ .

The problem can be simplified by concentrating the log-likelihood function with respect to  $\hat{\sigma}^2$  which is given by the variance of the residuals

$$\hat{\sigma}^2 = \frac{1}{T} \sum_{t=1}^T (\ln y_t + \hat{\alpha} y_t - \hat{\beta}_0 - \hat{\beta}_1 \ln k_t - \hat{\beta}_2 \ln l_t)^2.$$

## 2.10 Computer Applications

### 2.10.1 Robust Estimation of the CAPM

One way to ensure that parameter estimates of the nonlinear regression model are robust to the presence of outliers, is to model the outliers using a heavy-tailed distribution such as the Student t distribution. This is a natural approach to modelling outliers as by definition an outlier represents an extreme draw from the tails of the distribution. If the presence of significant outliers is believed to be inconsistent with assuming normality, then a fat-tailed distribution is needed, such as a Student t distribution. The general idea is that the additional parameters of the heavy-tailed distribution capture the effects of the outliers which help to quarantine the estimates of the regression parameters from the outliers.

To demonstrate the approach consider the capital asset pricing model

$$\begin{aligned}r_t &= \beta_0 + \beta_1 m_t + u_t \\ u_t &\sim N(0, \sigma^2),\end{aligned}$$

where  $r_t$  is the return on the  $i^{th}$  asset relative to a risk free rate,  $m_t$  is the return on the market portfolio relative to a risk free rate, and  $u_t$  is a normal disturbance term. The parameter  $\beta_1$  is of importance in finance as it provides a measure of the risk of the asset. To control for outliers in the data, the model is rewritten as

$$r_t = \beta_0 + \beta_1 m_t + \sigma \sqrt{\frac{\beta_2 - 2}{\beta_2}} v_t,$$

where the disturbance term  $v_t$  now has a Student-t distribution given by

$$f(v_t) = \frac{\Gamma\left(\frac{\beta_2 + 1}{2}\right)}{\sqrt{\pi\beta_2} \Gamma\left(\frac{\beta_2}{2}\right)} \left(1 + \frac{v_t^2}{\beta_2}\right)^{-(\beta_2+1)/2},$$

where  $\beta_2$  is the degrees of freedom parameter and  $\Gamma(\cdot)$  is the Gamma function. The term  $\sigma\sqrt{(\beta_2 - 2)/\beta_2}$  next to  $v_t$  ensures that the variance of  $r_t$  is  $\sigma^2$ , as the variance of a Student t distribution is  $\beta_2/(\beta_2 - 2)$ .

Using the transformation of variable technique, the distribution of  $r_t$  is

$$\begin{aligned} f(r_t) &= f(v_t) \left| \frac{dv_t}{dr_t} \right| \\ &= \frac{\Gamma\left(\frac{\beta_2 + 1}{2}\right)}{\sqrt{\pi\beta_2} \Gamma\left(\frac{\beta_2}{2}\right)} \left(1 + \frac{v_t^2}{\beta_2}\right)^{-(\beta_2+1)/2} \left| \frac{1}{\sigma} \sqrt{\frac{\beta_2}{\beta_2 - 2}} \right|, \end{aligned}$$

yielding the log-likelihood function at observation  $t$

$$\begin{aligned} \ln L_t(\theta) &= \ln f(r_t) \\ &= \ln \left( \frac{\Gamma\left(\frac{\beta_2 + 1}{2}\right)}{\sqrt{\pi\beta_2} \Gamma\left(\frac{\beta_2}{2}\right)} \right) - \frac{\beta_2 + 1}{2} \ln \left( 1 + \frac{v_t^2}{\beta_2} \right) - \ln \sigma + \ln \sqrt{\frac{\beta_2}{\beta_2 - 2}}. \end{aligned}$$

The parameters  $\theta = \{\beta_0, \beta_1, \sigma^2, \beta_2\}$  are estimated by maximum likelihood using an iterative algorithm.

As an illustration, consider the monthly returns on the company Martin Marietta, over the period January 1982 to December 1986, taken from Butler et. al., (?), pp.321-327. A scatter plot of the data in Figure 7 suggests that estimation of the CAPM by least squares may yield an estimate of  $\beta_1$  that is biased upwards as a result of the outlier in  $r_t$  where the monthly excess return of the asset in one month is 0.688.

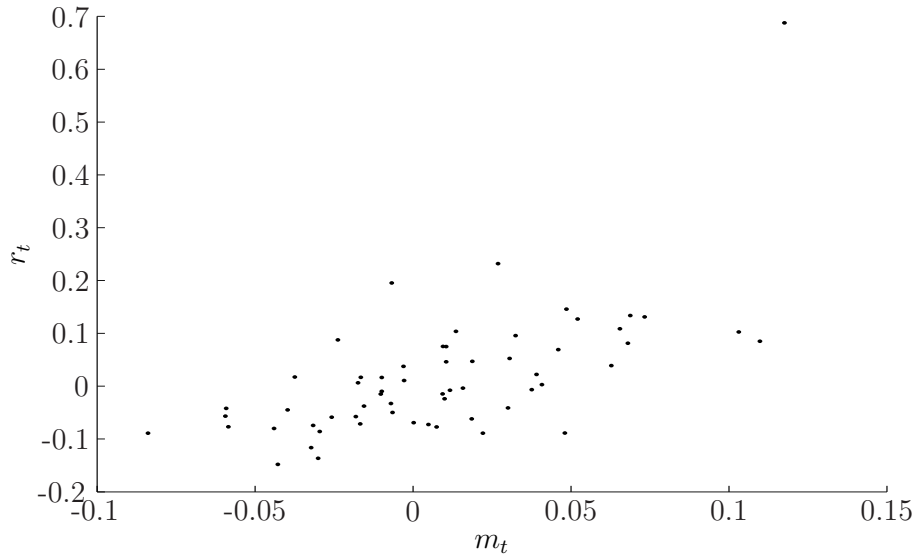


Figure 7: Scatter plot of the CAPM variables.

The results of estimating the the CAPM by maximum likelihood assuming normal dis-

Table 3:

Maximum likelihood estimates of the robust capital asset pricing model.  
Standard errors based on the inverse of the Hessian.

Parameter	Estimate	Std error	t-stat.
$\beta_0$	-0.007	0.008	-0.887
$\beta_1$	1.263	0.190	6.665
$\sigma^2$	0.008	0.006	1.338
$\beta_2$	2.837	1.021	2.779

turbances, are

$$\hat{r}_t = 0.001 + 1.803 m_t,$$

where the estimates are obtained by simply regressing  $r_t$  on a constant and  $m_t$ . The estimate of 1.803 suggests that this asset is very risky relative to the market portfolio as the asset moves in excess of the movements in the market excess returns  $m_t$ , on average. A test of the hypothesis that  $\beta_1 = 1$ , provides a test that movements in the returns on the asset mirror the market one-to-one. The Wald statistic is

$$W = \left( \frac{1.803 - 1}{0.285} \right)^2 = 7.930.$$

The p-value is 0.000, showing strong rejection of the null hypothesis.

The maximum likelihood estimates of the robust version of the CAPM model are given in Table 3. The estimate of  $\beta_1$  is now 1.263 which is much less than the OLS estimate of 1.803. A Wald test of the hypothesis that  $\beta_1 = 1$ , now yields

$$W = \left( \frac{1.263 - 1}{0.190} \right)^2 = 1.930.$$

The p-value is 0.164 showing that the null hypothesis that the asset tracks the market one-for-one, fails to be rejected.

The use of the Student-t distribution to model the outlier has helped to reduce the effect of the outlier on the estimate of  $\beta_1$ . The degrees of freedom parameter estimate of 2.837 shows that the tails of the distribution are indeed very fat, with just the first two moments of the distribution existing.

## 2.10.2 Nonnested Test US Money Demand

Two models are nonnested if it is not possible to express one of the models as a subset of the other model. Whilst a number of procedures have been developed to test nonnested models, in this application a maximum likelihood approach is discussed. The basic approach is to convert the likelihoods of two competing models into a common likelihood using the transformation of variable technique and perform a likelihood ratio test. The details of this approach are best demonstrated using the following example.

Consider the following two alternative money demand equations

$$\begin{aligned} \text{Model 1: } m_t &= \beta_0 + \beta_1 r_t + \beta_2 y_t + u_{1,t} \\ u_{1,t} &\sim N(0, \sigma_1^2), \end{aligned}$$

$$\begin{aligned} \text{Model 2: } \ln m_t &= \alpha_0 + \alpha_1 \ln r_t + \alpha_2 \ln y_t + u_{2,t} \\ u_{2,t} &\sim N(0, \sigma_2^2), \end{aligned}$$

where  $m_t$  is real money,  $y_t$  is real income,  $r_t$  is the nominal interest rate and  $\theta_1 = \{\beta_0, \beta_1, \beta_2, \sigma_1^2\}$  and  $\theta_2 = \{\alpha_0, \alpha_1, \alpha_2, \sigma_2^2\}$  are the unknown parameters of two models. The models are not nested as it is not possible to express one model as a subset of the other. Another way to view this problem is to observe that Model 1 is based on the distribution of  $m_t$  whereas Model 2 is based on the distribution of  $\ln m_t$ . Specifically, the respective distributions of the two models are

$$\begin{aligned} f_1(m_t) &= \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp \left[ -\frac{(m_t - \beta_0 - \beta_1 r_t - \beta_2 y_t)^2}{2\sigma_1^2} \right] \\ f_2(\ln m_t) &= \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp \left[ -\frac{(\ln m_t - \alpha_0 - \alpha_1 \ln r_t - \alpha_2 \ln y_t)^2}{2\sigma_2^2} \right]. \end{aligned}$$

To make the comparison between the two models commensurate, the transformation of variable technique is used to convert the distribution  $f_2$  into a distribution of the level of  $m_t$ . Formally this link between the two distributions is given by

$$\begin{aligned} f_1(m_t) &= f_2(\ln m_t) \left| \frac{d \ln m_t}{dm_t} \right| \\ &= f_2(\ln m_t) \left| \frac{1}{m_t} \right|, \end{aligned}$$

which allows the log-likelihoods of the two models to be compared using a likelihood ratio test. Formally, the steps to perform the test are as follows:

1. Estimate Model 1 by regressing  $m_t$  on  $\{c, r_t, y_t\}$  and construct the log-likelihood

function at each observation:

$$\ln L_{1,t}(\hat{\theta}_1) = -\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln(\hat{\sigma}_1^2) - \frac{(m_t - \hat{\beta}_0 - \hat{\beta}_1 r_t - \hat{\beta}_2 y_t)^2}{2\hat{\sigma}_1^2}.$$

2. Estimate Model 2 by regressing  $\ln m_t$  on  $\{c, \ln r_t, \ln y_t\}$  and construct the log-likelihood function at each observation for  $m_t$  by using:

$$\begin{aligned} \ln L_{2,t}(\hat{\theta}_2) &= -\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln(\hat{\sigma}_2^2) - \frac{(\ln m_t - \hat{\alpha}_0 - \hat{\alpha}_1 \ln r_t - \hat{\alpha}_2 \ln y_t)^2}{2\hat{\sigma}_2^2} \\ &\quad - \ln m_t. \end{aligned}$$

3. Compute the difference in the log likelihoods of the two models at each observation

$$d_t = \ln L_{1,t}(\hat{\theta}_1) - \ln L_{2,t}(\hat{\theta}_2).$$

4. Construct the test statistic

$$V = \sqrt{T} \frac{\bar{d}}{s},$$

where

$$\bar{d} = \frac{1}{T} \sum_{t=1}^T d_t, \quad s^2 = \frac{1}{T} \sum_{t=1}^T (d_t - \bar{d})^2,$$

are the mean and the variance of  $d_t$  respectively.

5. Using the result in Vuong (1989), the statistic  $V$  is asymptotically normally distributed under the null hypothesis that the two models are equivalent

$$V \xrightarrow{a} N(0, 1).$$

The models are estimated using annual data for the U.S. on real money,  $m_t$ , the nominal interest rate,  $r_t$ , and real income,  $y_t$ , for the period 1966 to 1985. The data are taken from Greene (? , Table 11.4, p.342). The estimates of Model 1 are

$$\hat{m}_t = -3169.418 + -14.922 r_t + 1.588 y_t.$$

The estimates of Model 2 are

$$\widehat{\ln m}_t = -21.992 + -0.032 \ln r_t + 3.656 \ln y_t.$$

The mean and variance of  $d_t$  are respectively

$$\bar{d} = -0.603$$

$$s^2 = 0.459,$$

yielding the value of the test statistic

$$V = \sqrt{T} \frac{\bar{d}}{s} = \sqrt{20} \frac{-0.603}{\sqrt{0.459}} = -3.980.$$

The p-value is 0.000, showing that the null hypothesis that the models are equivalent representations of the money demand, is rejected at conventional significance levels. As the statistic is negative this suggests that Model 2 has the higher log-likelihood, suggesting that it is to be preferred.

### 2.10.3 A GARCH(1,1) Model of US Yields

This application is based on the file `garch_estimate.m`. The data are in the Matlab `USYields.mat` and consist of daily yields ( $r_t$ ) on US zero coupon bonds, expressed as a percentage, over the period October 10th 1988 to December 28th 2001, a total of 3307 observations. The maturities of the bonds are 3 months, 1 year, 3 year, 5 year, 7 year and 10 year.

The following GARCH(1,1) model of yields ( $r_t$ ) is estimated for various US zero coupon bonds

$$\begin{aligned} dr_t &= \gamma_0 + u_t \\ h_t &= \alpha_0 + \alpha_1 u_{t-1}^2 + \alpha_2 h_{t-1} \\ u_t &\sim N(0, h_t), \end{aligned}$$

with  $dr_t = 100(r_t - r_{t-1})$ .

The parameter estimates are presented in Table 4 with standard errors based on the outer product of the gradients. The initial value value of  $h_1$  is taken as the sample variance of  $dr_t$ . The average log likelihood values are reported in the last column. An important feature of the empirical results is that the parameter estimates are consistent across all maturities greater than or equal to one year, with the estimates of  $\alpha_2$  being around 0.9 and the estimates of  $\alpha_1$  being around 0.04. Moreover, the sum of the  $\alpha_1$  and  $\alpha_2$  parameter estimates across all maturities, without any exceptions, are

$$\hat{\alpha}_1 + \hat{\beta}_1 \simeq 1.$$

From (31) this suggests that volatility in yields exhibits very long memory as there is a near unit root, a property that is returned to below. A plot of the conditional variance

of the 3-month yield is given in Figure ???. The conditional variance for observations  $t = 1, 2, 3$  are estimated as

$$\begin{aligned}\widehat{h}_1 &= \frac{1}{T-1} \sum (dr_t - \overline{dr_t})^2 \\ &= 30.330,\end{aligned}$$

$$\begin{aligned}\widehat{h}_2 &= \widehat{\alpha}_0 + \widehat{\alpha}_1(dr_1 - \widehat{\gamma}_0)^2 + \widehat{\beta}_1\widehat{\sigma}_1^2 \\ &= 1.788 + 0.211 \times (-1.000 - 0.081) + 0.762 \times 30.330 \\ &= 25.235,\end{aligned}$$

$$\begin{aligned}\widehat{h}_3 &= \widehat{\alpha}_0 + \widehat{\alpha}_1(dr_2 - \widehat{\gamma}_0)^2 + \widehat{\beta}_1\widehat{\sigma}_2^2 \\ &= 1.788 + 0.211 \times (9.000 - 0.081) + 0.762 \times 30.330 \\ &= 37.724,\end{aligned}$$

where the  $\widehat{h}_1$  represents the unconditional variance of the data, and the first and second observations of the change in the 3-month yields multiplied by 100, are respectively  $dr_1 = -1$  and  $dr_2 = 9$ .

Table 4:

Maximum likelihood estimates of GARCH(1,1) models of U.S. yields:  
 standard errors based on the outer product of gradients in parentheses.

Yield	$\gamma_0$	$\alpha_0$	$\alpha_1$	$\beta_1$	$\ln L/T$
3 month	0.081 (0.076)	1.788 (0.126)	0.211 (0.008)	0.762 (0.009)	-3.016
1 year	-0.072 (0.087)	0.452 (0.053)	0.056 (0.003)	0.933 0.003	-3.098
3 year	-0.196 (0.105)	0.688 (0.123)	0.037 (0.003)	0.946 (0.006)	-3.244
5 year	-0.184 (0.107)	0.835 (0.172)	0.035 (0.004)	0.945 (0.007)	-3.249
7 year	-0.194 (0.105)	0.946 (0.221)	0.035 (0.005)	0.941 (0.009)	-3.224
10 year	-0.186 (0.100)	1.022 (0.236)	0.034 (0.005)	0.937 (0.010)	-3.179

# 3 QUASI MAXIMUM LIKELIHOOD ESTIMATION

## 3.1 Introduction

An important assumption made thus far is that the specification of the likelihood function, in terms of the joint probability distribution of the variables, is correct. Under these conditions the maximum likelihood estimator has the desirable properties of consistency, achieving the Cramér-Rao lower bound given by the inverse of the Information matrix and asymptotically normality.

In this session, the problem of misspecification of the likelihood function is investigated. In general, the maximum likelihood estimator in the presence of misspecification does not display the usual properties. However, there are certain cases in which the maximum likelihood estimator of a misspecified model is still a consistent estimator of the population parameter of the true model. In these cases, this estimator is referred to as the quasi-maximum likelihood estimator.

One important difference between the maximum likelihood estimator based on the true probability distribution and the quasi-maximum likelihood estimator is that the usual variance formulae based on either the information matrix or the outer product of the gradient matrix are no longer appropriate. Nonetheless an estimate of the variance can be still be computed in this situation by combining both these matrices. In the context of regression models, misspecification may take the form of incorrect functional forms of heteroskedasticity and/or autocorrelation. In these cases, a quasi-maximum likelihood estimator exists where the correct estimate of variance is based on either White or Newey-West estimators respectively. The strategy of correcting for potential heteroskedasticity and autocorrelation by using a quasi-maximum likelihood estimator with a robust estimator of the covariance matrix, is commonly referred to as nonparametric because the solution does not require the explicit specification of the functional form of the heteroskedasticity and autocorrelation.

## 3.2 Motivating Examples

Consider the situation where the true probability distribution of  $y_t$  is  $f_0(y_t; \theta_0)$  but an incorrect probability distribution given by  $f_1(y_t; \theta_1)$  is used to construct the likelihood

function. Some specific examples of misspecified models are:

(i) **Duration Analysis**

The true model of duration times (between trades say) between significant events follows a exponential distribution with parameter  $\theta_0$  and the misspecified model posits that the durations are normally distributed with mean  $\theta_1$  and unit variance:

$$\begin{aligned} f_0(y_t; \theta_0) &= \frac{1}{\theta_0} \exp \left[ -\frac{y_t}{\theta_0} \right] \\ f_1(y_t; \theta_1) &= \frac{1}{\sqrt{2\pi}} \exp \left[ -\frac{(y_t - \theta_1)^2}{2} \right]. \end{aligned}$$

(ii) **Financial Returns**

In the true model financial returns follow a standardized Student t distribution with  $\gamma_0$  degrees of freedom and in the misspecified model they are assumed to be normally distributed:

$$\begin{aligned} f_0(y_t; \theta_0) &= \frac{\Gamma \left( \frac{\gamma_0 + 1}{2} \right)}{\sqrt{\pi \sigma_0^2 (\gamma_0 - 2)} \Gamma \left( \frac{\gamma_0}{2} \right)} \left( 1 + \frac{(y_t - \mu_0)^2}{\sigma_0 (\gamma_0 - 2)} \right)^{-(\gamma_0 + 1)/2} \\ f_1(y_t; \theta_1) &= \frac{1}{\sqrt{2\pi \sigma_1^2}} \exp \left[ -\frac{(y_t - \mu_1)^2}{2\sigma_1^2} \right], \end{aligned}$$

where  $\theta_0 = \{\mu_0, \sigma_0^2, \gamma_0\}$  and  $\theta_1 = \{\mu_1, \sigma_1^2\}$ .

(iii) **Income Distribution**

Per capita income is log-normally distributed with parameters  $\theta_0 = \{\mu_0, \sigma_0^2\}$  and the misspecified model is normal with parameters  $\theta_1 = \{\mu_1, \sigma_1^2\}$ :

$$\begin{aligned} f_0(y_t; \theta_0) &= \frac{1}{y_t \sqrt{2\pi \sigma_0^2}} \exp \left[ -\frac{(\ln y_t - \mu_0)^2}{2\sigma_0^2} \right] \\ f_1(y_t; \theta_1) &= \frac{1}{\sqrt{2\pi \sigma_1^2}} \exp \left[ -\frac{(y_t - \mu_1)^2}{2\sigma_1^2} \right]. \end{aligned}$$

(iv) **Heteroskedasticity**

The true model is the heteroskedastic regression model and the misspecified model assumes the disturbance variance is homoskedastic:

$$\begin{aligned} f_0(y_t; \theta_0) &= \frac{1}{\sqrt{2\pi \sigma_t^2}} \exp \left[ -\frac{(y_t - \alpha_0 - \beta_0 x_t)^2}{2\sigma_t^2} \right] \\ f_1(y_t; \theta_1) &= \frac{1}{\sqrt{2\pi \sigma^2}} \exp \left[ -\frac{(y_t - \alpha_1 - \beta_1 x_t)^2}{2\sigma^2} \right]. \end{aligned}$$

(v) **Autocorrelation**

The true model is the regression model with autocorrelation and the misspecified model assumes an independent disturbance:

$$\begin{aligned} f_0(y_t; \theta_0) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{(y_t - \alpha_0(1 - \rho_0) - \beta_0(x_t - \rho_0 x_{t-1}))^2}{2\sigma^2} \right] \\ f_1(y_t; \theta_1) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{(y_t - \alpha_1 - \beta_1 x_t)^2}{2\sigma^2} \right]. \end{aligned}$$

(vi) **Time Series Models**

The true model is a moving average model

$$y_t = u_t - \rho_0 u_{t-1}, \quad u_t \sim N(0, \sigma_0^2),$$

and the misspecified model is the autoregressive model

$$y_t = \rho_1 y_{t-1} + v_t, \quad v_t \sim N(0, \sigma_1^2),$$

so the true and misspecified distributions are respectively:

$$\begin{aligned} f_0(y_t; \rho_0) &= \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp \left[ -\frac{(y_t + \rho_0 u_{t-1})^2}{2\sigma_0^2} \right] \\ f_1(y_t; \rho_1) &= \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp \left[ -\frac{(y_t - \rho_1 y_{t-1})^2}{2\sigma_1^2} \right], \end{aligned}$$

where  $\theta_0 = \{\rho_0, \sigma_0^2\}$  and  $\theta_1 = \{\rho_1, \sigma_1^2\}$ .

In the examples above, the maximum likelihood estimators of the parameters of the misspecified distribution are obtained by maximizing the log-likelihood function

$$\ln L = \sum_{t=1}^T \ln f_1(y_t; \theta_1). \quad (69)$$

The estimator of  $\hat{\theta}_1$  is obtained by setting the first-order conditions given by

$$G(\theta_1) = \frac{\partial \ln L}{\partial \theta_1} = \sum_{t=1}^T \frac{\partial \ln f_1(y_t; \theta_1)}{\partial \theta_1}, \quad (70)$$

to zero. It turns  $\hat{\theta}_1$  may still provide useful information on  $\theta_0$  or some elements of  $\theta_0$ . To highlight this point, take expectations of the gradient vector in equation (70) with

respect to the true probability distribution  $f_0(y_t; \theta_0)$

$$\begin{aligned}
E_0 [G(\theta_1)] &= \int_{-\infty}^{\infty} G(\theta_1) f_0(y_t; \theta_0) dy_t \\
&= \int_{-\infty}^{\infty} \sum_{t=1}^T \frac{\partial \ln f_1(y_t; \theta_1)}{\partial \theta_1} f_0(y_t; \theta_0) dy_t \\
&= \int_{-\infty}^{\infty} \sum_{t=1}^T \frac{\partial f_1(y_t; \theta_1)}{\partial \theta_1} \frac{f_0(y_t; \theta_0)}{f_1(y_t; \theta_1)} dy_t, \tag{71}
\end{aligned}$$

where  $E_0 [\cdot]$  signifies that the expectation taken with respect to the true distribution. This expression is not guaranteed to equal zero except in the case where the distribution is specified correctly

$$f_1(y_t; \theta_1) = f_0(y_t; \theta_0).$$

In this case, (71) may be simplified by exchanging the integration and differentiation operators and using the property of a probability distribution to give

$$E_0 [G(\theta_1)] = \sum_{t=1}^T \frac{\partial}{\partial \theta_1} \int_{-\infty}^{\infty} f_1(y_t; \theta_1) dy_t = \sum_{t=1}^T \frac{\partial}{\partial \theta_1} 1 = 0. \tag{72}$$

Thus sufficient condition for (72) to hold is that the model is specified correctly. There are, however, some important cases where  $E_0 [G(\theta_1)] = 0$  even when the distribution is misspecified, as the following example demonstrates.

### Exponential Distribution

Suppose that in the true model  $y_t$  follows an exponential distribution with parameter  $\theta_0$  and that in the misspecified model  $y_t$  is normally distributed with mean  $\theta_1$  and unit variance, that is

$$\begin{aligned}
f_0(y_t) &= \frac{1}{\theta_0} \exp \left[ -\frac{y_t}{\theta_0} \right] \\
f_1(y_t) &= \frac{1}{\sqrt{2\pi}} \exp \left[ -\frac{(y_t - \theta_1)^2}{2} \right].
\end{aligned}$$

The gradient of the misspecified log-likelihood function is

$$G(\theta_1) = \sum_{t=1}^T (y_t - \theta_1),$$

which yields the sample mean as the maximum likelihood estimator  $\hat{\theta}_1 = \bar{y}$ . Taking expectations under the true model

$$E_0 [G(\hat{\theta}_1)] = \sum_{t=1}^T (E_0 [y_t] - E_0 [\hat{\theta}_1]) = \sum_{t=1}^T (\theta_0 - \theta_0) = 0,$$

because  $\theta_0 = E_0[y_t]$  from the properties of the exponential distribution and therefore

$$E_0[\widehat{\theta}_1] = E_0 \left[ T^{-1} \sum y_t \right] = T^{-1} \sum E_0[y_t] = \theta_0 .$$

This example highlights an important result. If the mean of the misspecified distribution happens to be the same as the mean of true distribution, then the expectation of the gradient function  $E_0 [G(\theta_1)]$  will be zero, despite the shape of the distribution being misspecified. In these circumstances, the implication is that the estimator of the parameter of the misspecified model provides a consistent estimator of the true parameter

$$\text{plim}(\widehat{\theta}_1) = \theta_0 ,$$

and is known as the quasi-maximum likelihood estimator. In the case of the normal distribution in the previous example, the sample mean is the maximum likelihood estimator of  $\theta_1$ , which is also the quasi-maximum likelihood estimator of the parameter of the exponential distribution.

Although the examples thus far have emphasized the first moment being specified correctly, in general this result extends to higher moments. If the first  $K$  moments of the misspecified model are specified correctly, the parameters of these moments of the true model can still be estimated consistently despite the shape of the distribution being misspecified.

## ARCH

Suppose that  $y_t$  has mean  $\mu_t$  and variance  $\sigma_t^2$ :

$$\begin{aligned} y_t &= \mu_t + u_t \\ u_t &= \sigma_t z_t \\ \mu_t &= \alpha + \beta x_t \\ \sigma_t^2 &= \gamma + \delta u_{t-1}^2 . \end{aligned}$$

If the true distribution of  $z_t$  is not normal but estimation is based on the normal distribution, the maximum likelihood estimator of the misspecified model  $\widehat{\theta}_1 = \{\widehat{\alpha}, \widehat{\beta}, \widehat{\gamma}, \widehat{\delta}\}$  provides a consistent estimator of the corresponding population parameters. This result was first established by Bollerslev and Wooldridge in the context of GARCH models.

An example where  $E_0[G(\hat{\theta}_1)] \neq 0$ , is as follows.

### Moving Average

Suppose that the true model is a moving average  $y_t = u_t - \theta_0 u_{t-1}$ , and the misspecified model is autoregressive  $y_t = \theta_1 y_{t-1} + v_t$  where both  $u_t$  and  $v_t$  are  $N(0, 1)$ . The gradient function of the misspecified (conditional) log-likelihood function is

$$G(\theta_1) = \sum_{t=1}^T (y_t - \theta_1 y_{t-1}) y_{t-1},$$

which yields the maximum likelihood estimator

$$\hat{\theta}_1 = \frac{\frac{1}{T} \sum_{t=1}^T y_t y_{t-1}}{\frac{1}{T} \sum_{t=1}^T y_{t-1}^2} = \frac{\text{var}(y_t)}{\text{cov}(y_t, y_{t-1})}.$$

Evaluating the gradient at  $\hat{\theta}_1$  and taking probability limits with respect to the true model gives<sup>2</sup>

$$\text{plim}_0[G(\hat{\theta}_1)] = T \text{plim}_0 \left[ \frac{1}{T} \sum_{t=1}^T y_t y_{t-1} \right] - \text{plim}_0[\hat{\theta}_1] T \text{plim}_0 \left[ \frac{1}{T} \sum_{t=1}^T y_{t-1}^2 \right].$$

From the properties of an MA(1) (which you may remember from the Introductory Financial Econometrics course!)

$$\text{plim}_0 \left[ \frac{1}{T} \sum_{t=1}^T y_t y_{t-1} \right] = -\theta_0, \quad \text{plim}_0 \left[ \frac{1}{T} \sum_{t=1}^T y_{t-1}^2 \right] = 1 + \theta_0^2,$$

then

$$\text{plim}_0[G(\hat{\theta}_1)] = -T \theta_0 - \theta_1 T (1 + \theta_0^2),$$

which in general is not equal to zero, except for the special case of

$$\theta_1 = -\frac{\theta_0}{1 + \theta_0^2}.$$

By contrast with the previous example,  $\hat{\theta}_1$  is not a consistent estimator of  $\theta_0$  because the mean of the distribution is misspecified. However, it is worth noting at this point that there is a monotonic relationship between  $\theta_0$  and  $\theta_1$ , subject to certain invertibility conditions being satisfied, which suggests that given an estimate

---

<sup>2</sup>It is more convenient to use probability limits than expectations here as a result of the nonlinear structure of the gradient function.

of  $\theta_1$  an estimate of  $\theta_0$  can be found. **This point may be further developed in an ingenious way to deliver an estimator based on estimating the incorrect model! This will be the subject of hopefully a third course in the financial econometrics sometime later in the year or possibly in 2010 on Computationally Intensive Methods, namely, Estimation by Simulation, Nonparametric Estimation and Bootstrapping. Watch Johannes carefully for details!!**

### 3.3 The Information Equality

The analysis of the previous section shows that when the model is misspecified it is still possible to obtain a consistent estimator of the true parameter provided that the relevant moment is specified correctly. The following result shows that even in the special case where a parameter is estimated consistently, the usual variance formula of the maximum likelihood estimator based on the information matrix is not consistent. To highlight the implication of misspecifying the distribution on the variance, differentiate (70) again

$$\begin{aligned} \frac{\partial^2 \ln L}{\partial \theta_1 \partial \theta'_1} &= \frac{\partial}{\partial \theta_1} \sum_{t=1}^T \left[ \frac{\partial \ln f_1(y_t; \theta_1)}{\partial \theta'_1} \right] = \frac{\partial}{\partial \theta_1} \sum_{t=1}^T \left[ \frac{1}{f_1(y_t; \theta_1)} \frac{\partial f_1(y_t; \theta_1)}{\partial \theta'_1} \right] \\ &= \sum_{t=1}^T \frac{1}{f_1(y_t; \theta_1)^2} \left[ \frac{\partial^2 f_1(y_t; \theta_1)}{\partial \theta_1 \partial \theta'_1} f_1(y_t; \theta_1) - \frac{\partial f_1(y_t; \theta_1)}{\partial \theta_1} \frac{\partial f_1(y_t; \theta_1)}{\partial \theta'_1} \right]. \end{aligned}$$

Taking expectations with respect to the true probability distribution and rearranging gives

$$\begin{aligned} E_0 \left[ \frac{\partial^2 \ln L}{\partial \theta_1 \partial \theta'_1} \right] &= \int_{-\infty}^{\infty} \sum_{t=1}^T \left[ \frac{1}{f_1(y_t; \theta_1)} \frac{\partial^2 f_1(y_t; \theta_1)}{\partial \theta_1 \partial \theta'_1} \right] f_0(y_t; \theta_0) dy_t \\ &\quad - \int_{-\infty}^{\infty} \sum_{t=1}^T \left[ \frac{\partial \ln f_1(y_t; \theta_1)}{\partial \theta_1} \frac{\partial \ln f_1(y_t; \theta_1)}{\partial \theta'_1} \right] f_0(y_t; \theta_0) dy_t, \end{aligned} \tag{73}$$

where the second term on the right-hand side of equation (73) follows from the properties of differentiation of the logarithmic function. If the model is correctly specified,  $f_1(y_t; \theta_1) = f_0(y_t; \theta_0)$ , the first term on the right-hand side of equation (73) becomes

$$\begin{aligned} \sum_{t=1}^T \int_{-\infty}^{\infty} \frac{1}{f_1(y_t; \theta_1)} \frac{\partial^2 f_1(y_t; \theta_1)}{\partial \theta_1 \partial \theta'_1} f_0(y_t; \theta_0) dy_t &= \sum_{t=1}^T \int_{-\infty}^{\infty} \frac{\partial^2 f_1(y_t; \theta_1)}{\partial \theta_1 \partial \theta'_1} dy_t \\ &= \sum_{t=1}^T \frac{\partial^2}{\partial \theta_1 \partial \theta'_1} \int_{-\infty}^{\infty} f_1(y_t; \theta_1) dy_t \\ &= 0. \end{aligned}$$

In this case, equation (73) simplifies to

$$E_0 \left[ \frac{\partial^2 \ln L}{\partial \theta_1 \partial \theta_1'} \right] = -E_0 \left[ \sum_{t=1}^T \frac{\partial \ln f_1(y_t; \theta_1)}{\partial \theta_1} \frac{\partial \ln f_1(y_t; \theta_1)}{\partial \theta_1'} \right], \quad (74)$$

which is the information matrix equality derived on Day 1 of this course. However, if the model is misspecified then

$$\int_{-\infty}^{\infty} \frac{1}{f_1(y_t; \theta_1)} \frac{\partial^2 f_1(y_t; \theta_1)}{\partial \theta_1 \partial \theta_1'} f_0(y_t; \theta_0) dy_t \neq 0, \quad (75)$$

so the equality in equation (74) no longer holds. The following examples use simulation methods to show the magnitude of the violation of the information equality in (74) when the model is misspecified.

### $f_0$ is Exponential and $f_1$ is Normal

Consider the true model where  $y_t$  is *iid* and is distributed as exponential with parameter  $\theta_0$

$$f_0(y_t; \theta_0) = \frac{1}{\theta_0} \exp \left[ -\frac{y_t}{\theta_0} \right],$$

while in the misspecified model  $y_t$  is assumed to be normally distributed with mean  $\theta_1$  and unit variance

$$f_1(y_t; \theta_1) = \frac{1}{\sqrt{2\pi}} \exp \left[ -\frac{(y_t - \theta_1)^2}{2} \right].$$

The gradient and Hessian of the misspecified log-likelihood are respectively

$$G(\theta_1) = \frac{d \ln L}{d \theta_1} = \sum_{t=1}^T (y_t - \theta_1), \quad H(\theta_1) = \frac{d^2 \ln L}{d \theta_1^2} = -T.$$

Setting the gradient vector to zero yields the sample mean as the maximum likelihood estimator  $\hat{\theta}_1 = \bar{y}$ . To compute the outer product of gradients and the information matrix of the misspecified log-likelihood under the true model,  $T = 500,000$  observations on  $y_t$  are drawn from the exponential distribution with parameter  $\theta_0 = 0.5$ . The (average) outer product of the gradients evaluated at  $\hat{\theta}_1 = \bar{y}$ , is computed as

$$E_0 \left[ \frac{\partial \ln L_t}{\partial \theta_1} \frac{\partial \ln L_t}{\partial \theta_1'} \right] \simeq \frac{1}{T} \sum_{t=1}^T \frac{\partial \ln L_t}{\partial \theta_1} \frac{\partial \ln L_t}{\partial \theta_1'} = \frac{1}{T} \sum_{t=1}^T (y_t - \bar{y})^2 = 0.249.$$

The (average) information matrix is

$$-E_0 \left[ \frac{\partial^2 \ln L_t}{\partial \theta_1 \partial \theta_1'} \right] \simeq \frac{1}{T} \sum_{t=1}^T \frac{\partial^2 \ln L_t}{\partial \theta_1 \partial \theta_1'} = \frac{1}{T} \sum_{t=1}^T 1 = 1.$$

The effect of misspecifying the log-likelihood function has resulted in a violation of the information matrix equality, equation (74).

### $f_0$ is Student t and $f_1$ is Normal

Consider the true model where  $y_t$  is *iid* and distributed as standardized Student t with parameters  $\theta_0 = \{\mu_0, \sigma_0^2, \gamma_0\}$

$$f_0(y_t; \theta_0) = \frac{\Gamma\left(\frac{\gamma_0 + 1}{2}\right)}{\sqrt{\pi\sigma_0^2(\gamma_0 - 2)}\Gamma\left(\frac{\gamma_0}{2}\right)} \left(1 + \frac{(y_t - \mu_0)^2}{\sigma_0(\gamma_0 - 2)}\right)^{-(\gamma_0+1)/2},$$

while the in the misspecified model  $y_t$  is assumed to be normally distributed with parameters  $\theta_1 = \{\mu_1, \sigma_1^2\}$

$$f_1(y_t; \theta_1) = \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left[-\frac{(y_t - \mu_1)^2}{2\sigma_1^2}\right].$$

The gradient and Hessian of the misspecified log-likelihood are respectively

$$G(\theta_1) = \frac{\partial \ln L}{\partial \theta_1} = \sum_{t=1}^T \begin{bmatrix} \frac{y_t - \mu_1}{\sigma_1^2} \\ -\frac{1}{2\sigma_1^2} + \frac{(y_t - \mu_1)^2}{2\sigma_1^4} \end{bmatrix},$$

$$H(\theta_1) = \frac{\partial^2 \ln L}{\partial \theta_1 \partial \theta_1'} = \sum_{t=1}^T \begin{bmatrix} \frac{1}{\sigma_1^2} & \frac{y_t - \mu_1}{\sigma_1^4} \\ \frac{y_t - \mu_1}{\sigma_1^4} & -\frac{1}{2\sigma_1^4} + \frac{(y_t - \mu_1)^2}{\sigma_1^6} \end{bmatrix}.$$

Setting the gradient vector to zero yields the sample mean and the sample variance as the maximum likelihood estimators  $\hat{\theta}_1 = \{\bar{y}, s^2\}$ . To compute the outer product of gradients and the information matrix of the misspecified log-likelihood under the true model,  $T = 500,000$  observations on  $y_t$  are drawn from the Student t distribution with parameters  $\{\mu_0 = 1, \sigma_0^2 = 1\}$  for different values of  $\gamma_0$ . The (average) outer product of the gradients evaluated at  $\hat{\theta}_1$  is computed as

$$\frac{1}{T} \sum_{t=1}^T \begin{bmatrix} \left(\frac{y_t - \bar{y}}{s^2}\right)^2 & \left(-\frac{y_t - \bar{y}}{2s^4} + \frac{(y_t - \bar{y})^3}{2s^4}\right) \\ \left(-\frac{y_t - \bar{y}}{2s^4} + \frac{(y_t - \bar{y})^3}{2s^4}\right) & \left(-\frac{1}{2s^2} + \frac{(y_t - \bar{y})^2}{2s^4}\right)^2 \end{bmatrix},$$

Table 5:

The effect of misspecifying the log-likelihood on the relationship between the information matrix and outer product of the gradients matrix, with a simulation size of  $T = 500000$ . The true model is standardized Student t with degrees of freedom given by  $\gamma_0$ , and the misspecified model is based on the normal distribution.

$\gamma_0$	$-E_0 \left[ \frac{\partial^2 \ln L_t}{\partial \theta_1 \partial \theta_1'} \right]$		$E_0 \left[ \frac{\partial \ln L_t}{\partial \theta_1} \frac{\partial \ln L_t}{\partial \theta_1'} \right]$		Difference	
5	1.003	0.000	1.003	0.007	0.000	-0.007
	0.000	0.503	0.007	1.640	-0.007	-1.137
10	1.002	0.000	1.002	0.001	0.000	-0.001
	0.000	0.502	0.001	0.750	-0.001	-0.248
20	1.002	0.000	1.002	0.000	0.000	0.000
	0.000	0.502	0.000	0.597	0.000	-0.095
30	1.002	0.000	1.002	0.000	0.000	0.000
	0.000	0.502	0.000	0.561	0.000	-0.059
100	1.002	0.000	1.002	0.000	0.000	0.000
	0.000	0.502	0.000	0.519	0.000	-0.017
500	1.002	0.000	1.002	0.000	0.000	0.000
	0.000	0.502	0.000	0.507	0.000	-0.005

and the (average) information matrix is computed as

$$\frac{1}{T} \sum_{t=1}^T \begin{bmatrix} \frac{1}{s^2} & \frac{y_t - \bar{y}}{s^4} \\ \frac{y_t - \bar{y}}{s^4} & -\frac{1}{2s^4} + \frac{(y_t - \bar{y})^2}{s^6} \end{bmatrix}.$$

The results are summarized in Table 5. It is clear that low values of the degrees of freedom parameter  $\gamma_0$  result in large differences in the estimates of the variance of  $\sigma_1^2$ . As  $\gamma_0$  increases, this difference reduces as the true and misspecified models converge. Notice that in the case of the mean,  $\mu_1$ , the corresponding elements of the two matrices are equal. To see why this is the case, note that the relevant entry in the outer product of gradients matrix is identical to the corresponding element

of the information matrix, that is

$$T^{-1} \sum_{t=1}^T (y_t - \bar{y})^2 / s^4 = 1/s^2.$$

### 3.4 Variance of the Quasi-Maximum Likelihood Estimator

It has been argued so far that either the information matrix or the outer product of gradients can be used to estimate the variance of the maximum likelihood estimator as they are asymptotically equivalent. However, a key result of the previous section is that if the true model is misspecified, the condition that establishes the equality of these variance measures is no longer applicable. It turns out that it is possible to combine the two variance estimators to yield an estimator that is robust to certain types of misspecification.

To derive the robust variance of the quasi-maximum likelihood estimator, consider a first-order Taylor series expansion of the misspecified gradient function  $G(\theta)$  around the true parameter of the misspecified model  $\theta_1$

$$G(\theta) \simeq G(\theta_1) + H(\theta_1)(\theta - \theta_1),$$

where  $G(\theta_1)$  and  $H(\theta_1)$  are respectively the gradient and the Hessian of the misspecified model. Evaluating this expression at the quasi-maximum likelihood estimator of the misspecified model  $\hat{\theta}_1$ , gives

$$G(\hat{\theta}_1) \simeq G(\theta_1) + H(\theta_1)(\hat{\theta}_1 - \theta_1),$$

which, by construction reduces to

$$0 \simeq G(\theta_1) + H(\theta_1)(\hat{\theta}_1 - \theta_1).$$

Rearranging this expression and ignoring the contribution of higher-order terms yields

$$\hat{\theta}_1 - \theta_1 = -H(\theta_1)^{-1}G(\theta_1),$$

so that the variance of the quasi-maximum likelihood estimator is given by

$$\text{E}_0 \left[ (\hat{\theta}_1 - \theta_1)(\hat{\theta}_1 - \theta_1)' \right] = \text{E}_0 \left[ H(\theta_1)^{-1}G(\theta_1)G(\theta_1)'H(\theta_1)^{-1} \right],$$

which uses the fact that the Hessian is symmetric,  $H(\theta_1) = H(\theta_1)'$ . An alternative asymptotic equivalent formulation is to replace  $-H(\theta_1)$  by the information matrix  $I(\theta_1)$

$$\begin{aligned} \text{E}_0 \left[ (\hat{\theta}_1 - \theta_1)(\hat{\theta}_1 - \theta_1)' \right] &= \text{E}_0 \left[ I(\theta_1)^{-1}G(\theta_1)G(\theta_1)'I(\theta_1)^{-1} \right] \\ &= I(\theta_1)^{-1}\text{E}_0 \left[ G(\theta_1)G(\theta_1)' \right] I(\theta_1)^{-1}. \end{aligned}$$

In practice, the unknown parameter  $\theta_1$  is replaced by a consistent estimator  $\hat{\theta}_1$  to yield a robust estimator of the covariance matrix of the quasi-maximum likelihood estimator

$$\text{var}(\hat{\theta}_1) = I(\hat{\theta}_1)^{-1} E_0 \left[ G(\hat{\theta}_1) G(\hat{\theta}_1)' \right] I(\hat{\theta}_1)^{-1}. \quad (76)$$

In order to implement the robust estimate of the variance of the quasi-maximum likelihood estimator in equation (76), it is necessary to evaluate the expectations under the true model of the information matrix,  $I(\hat{\theta}_1 = -E_0[H(\theta_1)])$ , and the outer-product of the gradient matrix,  $E_0 [G(\theta_1)G(\theta_1)']$ . If no analytical expression can be derived for the expectation of the negative Hessian or information matrix, then it is common practice to approximate the information matrix with the actual negative of the Hessian matrix, namely  $H(\hat{\theta}_1)$ .

In the case of the outer product of gradients, the situation is slightly more complex. Consider

$$\begin{aligned} E_0 \left[ \frac{\partial \ln L}{\partial \theta_1} \frac{\partial \ln L}{\partial \theta_1'} \right] &= E_0 \left[ \left( \sum_{t=1}^T G_t \right) \left( \sum_{t=1}^T G_t \right)' \right] \\ &= E_0 \left[ \sum_{t=1}^T G_t G_t' + \left( \sum_{t=1}^T G_t G_{t-1}' + \sum_{t=1}^T G_{t-1} G_t' \right) \right. \\ &\quad \left. + \left( \sum_{t=2}^T G_t G_{t-2}' + \sum_{t=2}^T G_{t-2} G_t' \right) + \dots \right] \\ &= E_0 \left[ \sum_{t=1}^T G_t G_t' \right] + E_0 \left[ \sum_{t=1}^T G_t G_{t-1}' + \sum_{t=1}^T G_{t-1} G_t' \right] \\ &\quad + E_0 \left[ \sum_{t=2}^T G_t G_{t-2}' + \sum_{t=2}^T G_{t-2} G_t' \right] + \dots, \end{aligned} \quad (77)$$

where  $G_t$  is the  $(N \times 1)$  vector of gradients at  $t$  corresponding to the  $N$  parameters in  $\theta_1$ . The first term represents the contemporaneous covariances of the gradient vector, the second term represents the first order auto-covariances of gradient vector, the third term represents the second order auto-covariances and so on.

In some cases equation (77) may be simplified to yield an analytical expression for  $E_0 [G(\theta_1)G(\theta_1)']$ . The following example shows one such instance where an analytical expectation can be derived and how the robust covariance matrix of the quasi-maximum likelihood estimator in (76) is able to correct for the misspecification of the distribution.

**$f_0$  is Exponential and  $f_1$  is Normal revisited**

Reconsider the case where where the  $y_t$  are *iid* drawings from the exponential distribution with parameter  $\theta_0$ ,

$$f_0(y_t) = \frac{1}{\theta_0} \exp \left[ -\frac{y_t}{\theta_0} \right],$$

while in the misspecified model  $y_t$  is assumed to be normally distributed with mean  $\theta_1$  and unit variance

$$f_1(y_t) = \frac{1}{\sqrt{2\pi}} \exp \left[ -\frac{(y_t - \theta_1)^2}{2} \right].$$

Performing maximum likelihood estimation based on the probability density function of the true model,  $f_0(y_t)$ , yields the log-likelihood

$$\ln L = \sum_{t=1}^T \ln f_0(y_t) = -T \ln(\theta_0) - \frac{1}{\theta_0} \sum_{t=1}^T y_t,$$

with respective first and second derivatives

$$G(\theta_0) = \frac{d \ln L}{d\theta_0} = -\frac{T}{\theta_0} + \frac{1}{\theta_0^2} \sum_{t=1}^T y_t, \quad H(\theta_0) = \frac{d^2 \ln L}{d\theta_0^2} = \frac{T}{\theta_0^2} - \frac{2}{\theta_0^3} \sum_{t=1}^T y_t.$$

Setting the gradient to zero and solving yields the maximum likelihood estimator for  $\theta_0$ , namely, the sample mean  $\hat{\theta}_0 = \bar{y}$ . The variance of this estimator is

$$\text{var}(\hat{\theta}_0) = \left( -\text{E}_0 \left[ \frac{d^2 \ln L}{d\theta_0^2} \right] \right)^{-1} = \left( -\text{E}_0 \left[ \frac{T}{\theta_0^2} - \frac{2}{\theta_0^3} \sum_{t=1}^T y_t \right] \right)^{-1} = \frac{\theta_0^2}{T}.$$

The log-likelihood function of the misspecified model is based on the distribution  $f_1(y_t)$  and is given by

$$\ln L = \sum_{t=1}^T \ln f_1(y_t) = -\frac{1}{2} \ln(2\pi) - \frac{1}{2} \sum_{t=1}^T (y_t - \theta_1)^2,$$

with respective first and second derivatives

$$G(\theta_1) = \frac{d \ln L}{d\theta_1} = \sum_{t=1}^T (y_t - \theta_1), \quad H(\theta_1) = \frac{d^2 \ln L}{d\theta_1^2} = -T.$$

Setting the gradient to zero also yields the sample mean as the maximum likelihood estimator  $\hat{\theta}_1 = \bar{y}$ . To derive the robust variance, the information and outer product of gradient matrices are needed. The information matrix evaluated at  $\hat{\theta}_1$ , is

$$I(\hat{\theta}_1) = -\text{E}_0 \left[ \frac{d^2 \ln L}{d\theta_1^2} \right]_{\theta_1=\hat{\theta}_1} = T.$$

The outer product of gradient matrix evaluated at  $\widehat{\theta}_1$  is

$$\begin{aligned}
\mathbb{E}_0 \left[ G(\widehat{\theta}_1)G(\widehat{\theta}_1)' \right] &= \mathbb{E}_0 \left[ \sum_{t=1}^T (y_t - \widehat{\theta}_1) \sum_{t=1}^T (y_t - \widehat{\theta}_1) \right] \\
&= \mathbb{E}_0 \left[ \sum_{t=1}^T (y_t - \widehat{\theta}_1)^2 + 2(y_1 - \widehat{\theta}_1)(y_2 - \widehat{\theta}_1) + \dots \right] \\
&= \mathbb{E}_0 \left[ \sum_{t=1}^T (y_t - \theta_0)^2 + 2(y_1 - \theta_0)(y_2 - \theta_0) + \dots \right] \\
&= \mathbb{E}_0 \left[ \sum_{t=1}^T (y_t - \theta_0)^2 \right] \\
&= T\theta_0^2,
\end{aligned}$$

where the following results are used

$$\text{plim}(\widehat{\theta}_1) = \theta_0 \quad \theta_0^2 = \mathbb{E}_0 [(y_t - \theta_0)^2].$$

Using (76), the robust estimator of the variance of the quasi-maximum likelihood estimator is then

$$\text{var}(\widehat{\theta}_1) = I(\widehat{\theta}_1)^{-1} \mathbb{E}_0 [G(\widehat{\theta}_1)G(\widehat{\theta}_1)'] I(\widehat{\theta}_1)^{-1} = \frac{1}{T} \times T\theta_0^2 \times \frac{1}{T} = \frac{\theta_0^2}{T},$$

which matches the variance based on the true distribution  $f_0(y_t)$ .

The following simulation exercise illustrates these results. Let the true distribution be the standardized Student t distribution with parameters  $\theta_0 = \{\mu_0 = 10, \sigma_0^2 = 1, \gamma_0 = 10\}$  and the misspecified model be normal with parameters  $\theta_1 = \{\mu_1, \sigma_1^2\}$ . A simulated sample of size  $T = 500,000$  is now used to approximate the variance-covariance matrices of the parameters of the true model based on the Hessian (H), outer product of gradients (OPG) and robust estimator (QMLE). After scaling by  $T$  these are respectively

$$\begin{aligned}
\text{H} &: \begin{bmatrix} 0.943 & 0.000 & 0.088 \\ 0.000 & 2.983 & -54.824 \\ 0.088 & -54.824 & 7788.228 \end{bmatrix} \\
\text{OPG} &: \begin{bmatrix} 0.943 & -0.001 & -0.038 \\ -0.001 & 2.991 & -55.877 \\ -0.038 & -55.877 & 7921.335 \end{bmatrix} \\
\text{QMLE} &: \begin{bmatrix} 0.943 & 0.000 & 0.212 \\ 0.000 & 2.975 & -53.788 \\ 0.212 & -53.788 & 7657.380 \end{bmatrix}.
\end{aligned}$$

All three variance-covariance matrices are very similar which is not surprising given that the variance estimates are based on the true model. The corresponding variance-covariance matrices of the parameters of the misspecified model (scaled by  $T$ ) are respectively

$$\begin{aligned} \text{H} & : T \text{var}(\hat{\theta}_1) = \begin{bmatrix} 0.998 & 0.000 \\ 0.000 & 1.992 \end{bmatrix} \\ \text{OPG} & : T \text{var}(\hat{\theta}_1) = \begin{bmatrix} 0.998 & -0.001 \\ -0.001 & 1.334 \end{bmatrix} \\ \text{QMLE} & : T \text{var}(\hat{\theta}_1) = \begin{bmatrix} 0.998 & 0.001 \\ 0.001 & 2.975 \end{bmatrix}. \end{aligned}$$

The alternative variance-covariance matrices are now different as the estimates are based on the misspecified model. Moreover, comparing the last variance-covariance matrix based on the robust estimate of the variance of the quasi-maximum likelihood estimator, the top  $(2 \times 2)$  block of any of the three variance-covariance estimates based on the true model, shows that the robust estimator has helped to correct the misspecification.

In practise, an analytical expression for  $E_0 [G(\theta_1)G(\theta_1)']$  in equation (77) is rarely available and must be approximated by the pertinent sample moment evaluated at  $\hat{\theta}_1$ . There are two important cases to consider.

### Independent case:

All the auto-covariances of the gradient vector in equation (77) are zero,  $E_0[G_t G'_{t-j}] = 0 \forall j$ , a case which arises when the observations  $y_t$  are independently distributed. In this instance, the robust covariance matrix is computed as

$$\text{var}(\hat{\theta}_1) = H(\hat{\theta}_1)^{-1} J_0(\hat{\theta}_1) H(\hat{\theta}_1)^{-1}, \quad (78)$$

where

$$\begin{aligned} J_0(\hat{\theta}_1) & = E_0 \left[ \sum_{t=1}^T G_t G'_t \right] = \sum_{t=1}^T E_0 \left[ \sum_{t=1}^T G_t G'_t \right] \\ & = \sum_{t=1}^T \left[ \frac{1}{T} \sum_{t=1}^T G_t G'_t \right] = \sum_{t=1}^T G_t G'_t, \end{aligned}$$

where the expectation has been approximated by the relevant sample moment. The notation  $J_0(\hat{\theta}_1)$  indicates that the approximation is evaluated at the consistent estimator  $\hat{\theta}_1$

and the subscript 0 indicates that no auto-covariances of the gradient vector are included in the approximation. This is recognizable as the approximation given in the BHHH algorithm.

**Dependent case:**

If the observations  $y_t$  are not independently distributed, then  $E_0[G_t G'_{t-j}] \neq 0$  and significantly non-zero auto-covariances must be included in the approximation. The expectation is simply computed as the actual outer product of the gradients  $G_t G'_{t-j}$  evaluated at  $\hat{\theta}_1$ . The robust covariance matrix in the dependent case is computed as

$$\text{var}(\hat{\theta}_1) = H(\hat{\theta}_1)^{-1} J_P(\hat{\theta}_1) H(\hat{\theta}_1)^{-1}, \quad (79)$$

where

$$J_P(\hat{\theta}_1) = \sum_{t=1}^T G_t G'_t + \sum_{i=1}^P w_i \left( \sum_{t=1+i}^T G_t G'_{t-i} + \sum_{t=1+i}^T G_{t-i} G'_t \right), \quad (80)$$

and  $w_i$  are weights chosen in such a way as to ensure that  $J_P$ , evaluated at the consistent estimator  $\hat{\theta}_1$ , is positive semi-definite. Once again, the relevant expectation has been approximated by sample moments and the subscript  $P$  indicates that the approximation is curtailed at  $P$  lags of the auto-covariances of the gradient vector. The number of auto-covariances used to model the dependence is given by

$$P = \text{Floor} \left( 4 \left( \frac{T}{100} \right)^{2/9} \right), \quad (81)$$

where the function  $\text{Floor}(\cdot)$  rounds the term in brackets down to the nearest integer. The weights attached to the auto-covariances are given by

$$w_i = 1 - \frac{i}{P+1}, \quad (82)$$

with properties  $w_i > 0$  and  $\sum_{i=1}^P w_i = 1$ .

### 3.5 Misspecification of Regression Models

The quasi-maximum likelihood estimator for misspecified models is now applied to regression models. Two examples of misspecified regression models are considered. The first is where the true regression model exhibits heteroskedasticity, but the variance is misspecified to be homoskedastic. In this case the variance is computed using (78) which results in what is commonly referred to as the White estimator. The second is where the

true regression model exhibits autocorrelation, but the disturbance is misspecified to be independent. In this case the variance estimator is computed using (79) which results in what is commonly referred to as the Newey-West estimator. In deriving the White and Newey-West variance estimators the focus is on the mean parameters of the regression model, although the analysis is easily extended to include the parameters of higher-order moments which are specified correctly even though the shape of the distribution is not.

### 3.5.1 White Variance Estimator

Consider the linear heteroskedastic regression model

$$\begin{aligned} y_t &= \alpha_0 + \beta_0 x_t + u_t \\ u_t &\sim N(0, \sigma_t^2), \end{aligned} \tag{83}$$

where the specification of  $\sigma_t^2$  is unknown, so the true distribution is

$$f_0(y_t) = \frac{1}{\sqrt{2\pi\sigma_t^2}} \exp \left[ -\frac{(y_t - \alpha_0 - \beta_0 x_t)^2}{2\sigma_t^2} \right]. \tag{84}$$

Now suppose that the model is misspecified by assuming that the disturbance term is homoskedastic

$$\begin{aligned} y_t &= \alpha_1 + \beta_1 x_t + u_t \\ u_t &\sim N(0, \sigma_1^2), \end{aligned} \tag{85}$$

with parameters  $\theta_1 = \{\alpha_1, \beta_1, \sigma_1\}$ . The misspecified distribution is

$$f_1(y_t) = \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp \left[ -\frac{(y_t - \alpha_1 - \beta_1 x_t)^2}{2\sigma_1^2} \right], \tag{86}$$

resulting in the log-likelihood

$$\ln L = -\frac{T}{2} \ln 2\pi - \frac{T}{2} \ln \sigma_1^2 - \frac{1}{2\sigma_1^2} \sum_{t=1}^T (y_t - \alpha_1 - \beta_1 x_t)^2. \tag{87}$$

The first derivatives are

$$\begin{aligned} \frac{\partial \ln L}{\partial \alpha_1} &= \frac{1}{\sigma_1^2} \sum_{t=1}^T (y_t - \alpha_1 - \beta_1 x_t) = \frac{1}{\sigma_1^2} \sum_{t=1}^T u_t \\ \frac{\partial \ln L}{\partial \beta_1} &= \frac{1}{\sigma_1^2} \sum_{t=1}^T (y_t - \alpha_1 - \beta_1 x_t) x_t = \frac{1}{\sigma_1^2} \sum_{t=1}^T u_t x_t \\ \frac{\partial \ln L}{\partial \sigma_1^2} &= -\frac{T}{2\sigma_1^2} + \frac{1}{2\sigma_1^4} \sum_{t=1}^T (y_t - \alpha_1 - \beta_1 x_t)^2, \end{aligned} \tag{88}$$

and the second derivatives are

$$\begin{aligned}
\frac{\partial^2 \ln L}{\partial \alpha_1^2} &= -\frac{T}{\sigma_1^2} \\
\frac{\partial^2 \ln L}{\partial \alpha_1 \partial \beta_1} &= -\frac{1}{\sigma_1^2} \sum_{t=1}^T x_t \\
\frac{\partial^2 \ln L}{\partial \alpha_1 \partial \sigma_1^2} &= -\frac{1}{\sigma_1^4} \sum_{t=1}^T (y_t - \alpha_1 - \beta_1 x_t) \\
\frac{\partial^2 \ln L}{\partial \beta_1^2} &= -\frac{1}{\sigma_1^2} \sum_{t=1}^T x_t^2 \\
\frac{\partial^2 \ln L}{\partial \beta_1 \partial \sigma_1^2} &= -\frac{1}{\sigma_1^4} \sum_{t=1}^T (y_t - \alpha_1 - \beta_1 x_t) x_t \\
\frac{\partial^2 \ln L}{\partial (\sigma_1^2)^2} &= \frac{T}{2\sigma_1^4} - \frac{1}{\sigma_1^6} \sum_{t=1}^T (y_t - \alpha_1 - \beta_1 x_t)^2.
\end{aligned} \tag{89}$$

Setting the first derivatives in (88) to zero and solving yields the maximum likelihood estimators of the parameters of the misspecified model,  $\hat{\theta}_1$ .

Consider computing the quasi-maximum likelihood covariance matrix of the mean parameters of the misspecified model  $\alpha_1$  and  $\beta_1$  in equation (85), which requires the Hessian and the outer product of gradients. From equation (89) the Hessian evaluated at  $\hat{\theta}_1$  is

$$H(\hat{\theta}_1) = \frac{1}{\hat{\sigma}_1^2} \sum_{t=1}^T \begin{bmatrix} 1 & x_t \\ x_t & x_t^2 \end{bmatrix}. \tag{90}$$

To compute the outer product of the gradients, from (88) define

$$G_{1,t} = \frac{u_t}{\sigma_1^2}, \quad G_{2,t} = \frac{u_t x_t}{\sigma_1^2}. \tag{91}$$

From the independence property of  $u_t$  under the true model it follows that

$$\begin{aligned}
E_0 [G_{1,t} G_{1,s}] &= E_0 \left[ \frac{u_t}{\sigma_1^2} \times \frac{u_s}{\sigma_1^2} \right] = \frac{1}{\sigma_1^4} E [u_t u_s] = 0 \quad t \neq s \\
E_0 [G_{2,t} G_{2,s}] &= E_0 \left[ \frac{u_t x_t}{\sigma_1^2} \times \frac{u_s x_s}{\sigma_1^2} \right] = \frac{1}{\sigma_1^4} E [u_t u_s] x_t x_s = 0 \quad t \neq s \\
E_0 [G_{1,t} G_{2,s}] &= E_0 \left[ \frac{u_t}{\sigma_1^2} \times \frac{u_s x_s}{\sigma_1^2} \right] = \frac{1}{\sigma_1^4} E [u_t u_s x_s] = 0 \quad t \neq s.
\end{aligned}$$

The robust quasi-maximum likelihood covariance matrix is now given by equation (78) with the outer product of gradients matrix computed as

$$J_0(\hat{\theta}_1) = \sum_{t=1}^T \begin{bmatrix} G_{1,t}^2 & G_{1,t} G_{2,t} \\ G_{2,t} G_{1,t} & G_{2,t}^2 \end{bmatrix} = \frac{1}{\hat{\sigma}_1^4} \sum_{t=1}^T \hat{u}_t^2 \begin{bmatrix} 1 & x_t \\ x_t & x_t^2 \end{bmatrix},$$

where  $G_{1,t}$  and  $G_{2,t}$  are replaced by their sample estimates. Substituting the expressions for  $H(\hat{\theta}_1)$  in (90) and using  $J_0(\hat{\theta}_1)$  in equation (78) gives

$$\begin{aligned} \text{var}(\hat{\theta}_1) &= H(\hat{\theta}_1)^{-1} J_0(\hat{\theta}_1) H(\hat{\theta}_1)^{-1} \\ &= \left( \frac{1}{\hat{\sigma}^2} \sum_{t=1}^T \begin{bmatrix} 1 & x_t \\ x_t & x_t^2 \end{bmatrix} \right)^{-1} \left( \frac{1}{\hat{\sigma}^4} \sum_{t=1}^T \hat{u}_t^2 \begin{bmatrix} 1 & x_t \\ x_t & x_t^2 \end{bmatrix} \right) \left( \frac{1}{\hat{\sigma}^2} \sum_{t=1}^T \begin{bmatrix} 1 & x_t \\ x_t & x_t^2 \end{bmatrix} \right)^{-1} \\ &= \left( \sum_{t=1}^T \begin{bmatrix} 1 & x_t \\ x_t & x_t^2 \end{bmatrix} \right)^{-1} \left( \sum_{t=1}^T \hat{u}_t^2 \begin{bmatrix} 1 & x_t \\ x_t & x_t^2 \end{bmatrix} \right) \left( \sum_{t=1}^T \begin{bmatrix} 1 & x_t \\ x_t & x_t^2 \end{bmatrix} \right)^{-1}. \end{aligned} \quad (92)$$

This expression is also known as the White variance estimator which is robust to unknown heteroskedasticity. The last step also shows that the White estimator of the mean parameters is independent of  $\hat{\sigma}_1^2$ .

More generally, letting  $X$  be a  $(T \times K)$  matrix of regressors, equation (92) is written more compactly as

$$\text{var}(\hat{\theta}_1) = (X'X)^{-1} X' \begin{bmatrix} \hat{u}_1^2 & & \\ & \ddots & \\ & & \hat{u}_T^2 \end{bmatrix} X(X'X)^{-1}.$$

A special case of equation (92) is if the true model is indeed homoskedastic

$$E_0 [\hat{u}_t^2] = \sigma_0^2 = \sigma_1^2 = \sigma^2,$$

so that  $\hat{u}_t^2$  is replaced by  $\hat{\sigma}^2 = T^{-1} \sum \hat{u}_t^2$  and the robust variance estimator reduces to the standard least squares estimator

$$\text{var}(\hat{\theta}_1) = \hat{\sigma}^2 \left( \sum_{t=1}^T \begin{bmatrix} 1 & x_t \\ x_t & x_t^2 \end{bmatrix} \right)^{-1}.$$

### Calculation of the White Estimator

Compute the White estimator of the mean parameters  $\alpha_1$  and  $\beta_1$ , in the linear regression model

$$y_t = \alpha_1 + \beta_1 x_t + u_t$$

$$u_t \sim N(0, \sigma_1^2),$$

based on the following data:

Year	:	2004	2005	2006	2007	2008
$y_t$	:	3	2	4	6	10
$x_t$	:	1	2	3	4	5

The estimated model is

$$y_t = -0.4 + 1.8 x_t + \hat{u}_t$$

$$\hat{\sigma}_1^2 = \frac{(1.6)^2 + (-1.2)^2 + (-1.0)^2 + (-0.8)^2 + (1.4)^2}{5} = 1.52.$$

The least squares residuals and gradients are given in Table 6.

To compute the White estimator the outer product of gradients is

$$J_0(\hat{\theta}_1) = \left[ \sum_{t=1}^T G_t G_t' \right] = \sum_{t=1}^T \begin{bmatrix} G_{1,t}^2 & G_{1,t}G_{2,t} \\ G_{1,t}G_{2,t} & G_{2,t}^2 \end{bmatrix} = \begin{bmatrix} 3.289 & 9.003 \\ 9.003 & 33.137 \end{bmatrix},$$

while the Hessian is

$$H(\hat{\theta}_1) = -\frac{1}{\hat{\sigma}_1^2} \sum_{t=1}^T \begin{bmatrix} 1 & x_t \\ x_t & x_t^2 \end{bmatrix} = -\frac{1}{1.52} \begin{bmatrix} 5 & 15 \\ 15 & 55 \end{bmatrix}$$

$$= \begin{bmatrix} -3.289 & -9.868 \\ -9.868 & -36.184 \end{bmatrix}.$$

The White estimate of variance is therefore

$$\text{var}(\hat{\theta}_1) = H(\hat{\theta}_1)^{-1} J_0(\hat{\theta}_1) H(\hat{\theta}_1)^{-1}$$

$$= \begin{bmatrix} 2.358 & -0.645 \\ -0.645 & 0.202 \end{bmatrix}.$$

Table 6:

Computing the White and Newey-West estimates of variance for the linear regression model in the numerical examples.

Year	$y_t$	$x_t$	$\hat{u}_t$	$G_{1,t} = \hat{u}_t/\hat{\sigma}_1^2$	$G_{2,t} = \hat{u}_t x_t/\hat{\sigma}_1^2$
2004	3	1	1.6	1.053	1.053
2005	2	2	-1.2	-0.789	-1.579
2006	4	3	-1.0	-0.658	-1.974
2007	6	4	-0.8	-0.526	-2.105
2008	10	5	1.4	0.921	4.605

### 3.5.2 Newey-West Variance Estimator

Consider the linear autocorrelated regression model

$$\begin{aligned} y_t &= \alpha_0 + \beta_0 x_t + u_t \\ u_t &= \rho_0 u_{t-1} + v_t \\ v_t &\sim N(0, \sigma_0^2). \end{aligned} \tag{93}$$

The true distribution is

$$f_0(y_t | y_{t-1}) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left[-\frac{v_t^2}{2\sigma_0^2}\right], \tag{94}$$

where  $v_t = y_t - \alpha_0(1 - \rho_0) - \beta_1(x_t - \rho_0 x_{t-1}) - \rho_0 y_{t-1}$ . Now suppose that the model is misspecified by assuming that the disturbance term is independent ( $\rho_0 = 0$ ) so that the misspecified model is

$$\begin{aligned} y_t &= \alpha_1 + \beta_1 x_t + u_t \\ u_t &= v_t \\ v_t &\sim N(0, \sigma_1^2), \end{aligned} \tag{95}$$

and associated misspecified distribution

$$f_1(y_t) = \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left[-\frac{1}{2\sigma_1^2} (y_t - \alpha_1 - \beta_1 x_t)^2\right]. \tag{96}$$

Computing the quasi-maximum likelihood covariance matrix of the mean parameters of the misspecified model  $\alpha_1$  and  $\beta_1$  requires the Hessian and the outer product of gradients matrices. As equation (95) is the same as (85), the Hessian is given by equation (88). From the gradients in equation (88),

$$\begin{aligned} E_0 [G_{1,t} G_{1,s}] &= E_0 \left[ \frac{u_t}{\sigma^2} \times \frac{u_s}{\sigma^2} \right] = \frac{1}{\sigma_1^4} E [u_t u_s] \neq 0 \quad t \neq s \\ E_0 [G_{2,t} G_{2,s}] &= E_0 \left[ \frac{u_t x_t}{\sigma^2} \times \frac{u_s x_s}{\sigma^2} \right] = \frac{1}{\sigma_1^4} E [u_t u_s] x_t x_s \neq 0 \quad t \neq s \\ E_0 [G_{1,t} G_{2,s}] &= E_0 \left[ \frac{u_t}{\sigma^2} \times \frac{u_s x_s}{\sigma^2} \right] = \frac{1}{\sigma_1^4} E [u_t u_s x_s] \neq 0 \quad t \neq s, \end{aligned}$$

which follows immediately from the autocorrelation properties of  $u_t$  under the true model. This is in contrast to the situation of the White estimator where the disturbance term under the true model is independent. In which case the quasi-maximum likelihood variance-covariance matrix is now (79) where the outer product of gradients matrix is computed

as

$$\begin{aligned}
J_P(\hat{\theta}_1) &= \sum_{t=1}^T \begin{bmatrix} G_{1,t}^2 & G_{1,t}G_{2,t} \\ G_{2,t}G_{1,t} & G_{2,t}^2 \end{bmatrix} + w_1 \left( \sum_{t=2}^T \begin{bmatrix} G_{1,t}G_{1,t-1} & G_{1,t}G_{2,t-1} \\ G_{2,t}G_{1,t-1} & G_{2,t}G_{2,t-1} \end{bmatrix} \right) \\
&+ \sum_{t=2}^T \begin{bmatrix} G_{1,t-1}G_{1,t} & G_{1,t-1}G_{2,t} \\ G_{2,t-1}G_{1,t} & G_{2,t-1}G_{2,t} \end{bmatrix} + w_2 \left( \sum_{t=3}^T \begin{bmatrix} G_{1,t}G_{1,t-2} & G_{1,t}G_{2,t-2} \\ G_{2,t}G_{1,t-2} & G_{2,t}G_{2,t-2} \end{bmatrix} \right) \\
&+ \sum_{t=3}^T \begin{bmatrix} G_{1,t-2}G_{1,t} & G_{1,t-2}G_{2,t} \\ G_{2,t-2}G_{1,t} & G_{2,t-2}G_{2,t} \end{bmatrix} + \dots \\
&= \frac{1}{\hat{\sigma}^4} \sum_{t=1}^T \hat{u}_t^2 \begin{bmatrix} 1 & x_t \\ x_t & x_t^2 \end{bmatrix} + w_1 \frac{1}{\hat{\sigma}^4} \sum_{t=2}^T \hat{u}_t \hat{u}_{t-1} \left( \begin{bmatrix} 1 & x_{t-1} \\ x_t & x_t x_{t-1} \end{bmatrix} \right) \\
&+ \begin{bmatrix} 1 & x_t \\ x_{t-1} & x_t x_{t-1} \end{bmatrix} + \dots,
\end{aligned}$$

where  $G_{1,t}$  and  $G_{2,t}$  in equation (91) are replaced by their sample estimates, the lag length  $P$  is determined by (81) and the weight  $w_i$  is given by equation (82). Substituting the expressions for  $H(\hat{\theta}_1)$  in (90) and using  $J_P(\hat{\theta}_1)$  for  $J(\hat{\theta}_1)$  in (79) gives the required estimate of the variance, which is known as the Newey-West estimator that is robust to autocorrelation of an unknown functional form. Moreover, as seen previously, the White estimator in equation (92) is a special case of the estimator obtained when  $P = 0$ , it follows that the Newey-West estimator is robust to both autocorrelation and heteroskedasticity and standard errors based on equation (79) are therefore known as heteroskedastic-autoregressive consistent (HAC) standard errors. As with the White variance estimator in equation (92), the Newey-West estimator is independent of  $\hat{\sigma}_1^2$ .

### Calculation of the Newey-West Estimator

Compute the Newey-West estimator of the mean parameters of  $\alpha_1$  and  $\beta_1$ , in the linear regression model

$$\begin{aligned}
y_t &= \alpha_1 + \beta_1 x_t + u_t \\
u_t &\sim N(0, \sigma^2).
\end{aligned}$$

The data and key calculations are given in Table 6. To compute the Newey-West estimator, the maximum lagged auto-covariance from (81) is

$$P = \text{Floor} \left( 4 \left( \frac{5}{100} \right)^{2/9} \right) = \text{Floor} (2.0556) = 2,$$

and from (82) the weight function for  $P = 2$  is

$$w_1 = 1 - \frac{1}{2+1} = \frac{2}{3}, \quad w_2 = 1 - \frac{2}{2+1} = \frac{1}{3}.$$

Thus the outer product of gradient matrix is

$$\begin{aligned} J_2(\hat{\theta}_1) &= \left[ \sum_{t=1}^T G_t G_t' \right] + w_1 \left[ \sum_{t=2}^T G_t G_{t-1}' + \sum_{t=2}^T G_{t-1} G_t' \right] \\ &\quad + w_2 \left[ \sum_{t=3}^T G_t G_{t-2}' + \sum_{t=3}^T G_{t-2} G_t' \right], \end{aligned}$$

where the first term is

$$\left[ \sum_{t=1}^T G_t G_t' \right] = \sum_{t=1}^T \begin{bmatrix} G_{1,t}^2 & G_{1,t} G_{2,t} \\ G_{1,t} G_{2,t} & G_{2,t}^2 \end{bmatrix} = \begin{bmatrix} 3.289 & 9.003 \\ 9.003 & 33.137 \end{bmatrix},$$

the second term is

$$\begin{aligned} \sum_{t=2}^T G_t G_{t-1}' &= \sum_{t=2}^T \begin{bmatrix} G_{1,t} G_{1,t-1} & G_{1,t} G_{2,t-1} \\ G_{2,t} G_{1,t-1} & G_{2,t} G_{2,t-1} \end{bmatrix} = \begin{bmatrix} -0.450 & -0.692 \\ -1.143 & -4.086 \end{bmatrix} \\ \sum_{t=2}^T G_{t-1} G_t' &= \sum_{t=2}^T \begin{bmatrix} G_{1,t-1} G_{1,t} & G_{1,t-1} G_{2,t} \\ G_{2,t-1} G_{1,t} & G_{2,t-1} G_{2,t} \end{bmatrix} = \begin{bmatrix} -0.450 & -1.143 \\ -0.692 & -4.086 \end{bmatrix}, \end{aligned}$$

and the third term is

$$\begin{aligned} \sum_{t=3}^T G_t G_{t-2}' &= \sum_{t=3}^T \begin{bmatrix} G_{1,t} G_{1,t-2} & G_{1,t} G_{2,t-2} \\ G_{2,t} G_{1,t-2} & G_{2,t} G_{2,t-2} \end{bmatrix} = \begin{bmatrix} -0.883 & -1.679 \\ 3.445 & -7.843 \end{bmatrix} \\ \sum_{t=3}^T G_{t-2} G_t' &= \sum_{t=3}^T \begin{bmatrix} G_{1,t-2} G_{1,t} & G_{1,t-2} G_{2,t} \\ G_{2,t-2} G_{1,t} & G_{2,t-2} G_{2,t} \end{bmatrix} = \begin{bmatrix} -0.883 & 3.445 \\ -1.679 & -7.843 \end{bmatrix}. \end{aligned}$$

Combining all terms

$$\begin{aligned} J_P(\hat{\theta}_1) &= \begin{bmatrix} 3.289 & 9.003 \\ 9.003 & 33.137 \end{bmatrix} \\ &\quad + \frac{2}{3} \left( \begin{bmatrix} -0.450 & -0.692 \\ -1.143 & -4.086 \end{bmatrix} + \begin{bmatrix} -0.450 & -1.143 \\ -0.692 & -4.086 \end{bmatrix} \right) \\ &\quad + \frac{1}{3} \left( \begin{bmatrix} -0.883 & -1.679 \\ 3.445 & -7.843 \end{bmatrix} + \begin{bmatrix} -0.883 & 3.445 \\ -1.679 & -7.843 \end{bmatrix} \right) \\ &= \begin{bmatrix} 2.101 & 6.071 \\ 6.071 & 22.461 \end{bmatrix}. \end{aligned}$$

Also, from the calculations of the White estimator

$$H(\hat{\theta}_1) = -\frac{1}{\hat{\sigma}^2} \sum_{t=1}^T \begin{bmatrix} 1 & x_t \\ x_t & x_t^2 \end{bmatrix} = -\frac{1}{1.52} \begin{bmatrix} 5 & 15 \\ 15 & 55 \end{bmatrix} = \begin{bmatrix} -3.289 & -9.868 \\ -9.868 & -36.184 \end{bmatrix}.$$

Hence

$$\begin{aligned} \text{var}(\hat{\theta}_1) &= H(\hat{\theta}_1)^{-1} J_2(\hat{\theta}_1) H(\hat{\theta}_1)^{-1} \\ &= \begin{bmatrix} 1.285 & -0.353 \\ -0.353 & 0.114 \end{bmatrix}. \end{aligned}$$

### 3.6 Testing

The calculation of standard errors in the linear regression model that are robust to misspecification of heteroskedasticity and autocorrelation suggests that the test statistics based on the likelihood principle can be made robust to these types of misspecification by using the quasi-maximum likelihood variance-covariance matrix. For example, to test the hypotheses

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_1 &: \theta \neq \theta_0, \end{aligned}$$

robust versions of the Wald and Lagrange Multiplier tests are respectively

$$\begin{aligned} W &= [\hat{\theta} - \theta_0]' [H^{-1}(\hat{\theta}) J_P(\hat{\theta}) H^{-1}(\hat{\theta})]^{-1} [\hat{\theta} - \theta_0] \\ LM &= G(\theta_0)' [H^{-1}(\theta_0) J_P(\theta_0) H^{-1}(\theta_0)] G(\theta_0). \end{aligned} \tag{97}$$

with  $J_P$  given by equation (80). By contrast, there is no natural adjustment of the likelihood ratio statistic to render it robust to misspecification of heteroskedasticity and autocorrelation. An implication of this result is that the robust analogues of the Wald and Lagrange multiplier tests have correct size under misspecification, whereas the likelihood ratio test does not.

An important property discussed in this chapter is that the information equality in (74) does not hold if the model is misspecified. This property suggests that a general test of misspecification can be based on comparing the elements of the Hessian  $H(\hat{\theta}_1)$  and the outer product of gradients  $J(\hat{\theta}_1)$ . The hypotheses tested are

$$\begin{aligned} H_0 &: \text{vech}(J(\theta)) + \text{vech}(H(\theta)) = 0 \\ H_1 &: \text{vech}(J(\theta)) + \text{vech}(H(\theta)) \neq 0. \end{aligned}$$

To demonstrate the form of this test statistic, consider the case where a normal likelihood is used as the misspecified model with unknown parameters  $\theta_1 = \{\mu_1, \sigma_1^2\}$ . The misspecified log-likelihood function is

$$\ln L = \sum_{t=1}^T \ln L_t = \sum_{t=1}^T \left( -\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln \sigma_1^2 - \frac{(y_t - \mu_1)^2}{2\sigma_1^2} \right),$$

resulting in the gradient vector

$$G(\theta_1) = \sum_{t=1}^T \begin{bmatrix} \frac{\partial \ln L_t}{\partial \mu_1} \\ \frac{\partial \ln L_t}{\partial \sigma_1^2} \end{bmatrix} = \frac{1}{\sigma_1^2} \sum_{t=1}^T \begin{bmatrix} (y_t - \mu_1) \\ -\frac{1}{2} + \frac{1}{2\sigma_1^2} (y_t - \mu_1)^2 \end{bmatrix}.$$

Setting the gradient vector to zero and rearranging gives the maximum likelihood estimators

$$\hat{\mu}_1 = \frac{1}{T} \sum_{t=1}^T y_t, \quad \hat{\sigma}_1^2 = \frac{1}{T} \sum_{t=1}^T (y_t - \hat{\mu}_1)^2.$$

To construct the test statistics, the Hessian and the outer product of the gradients, evaluated at the maximum likelihood estimator,  $\hat{\theta}_1$ , are respectively

$$\begin{aligned} H(\hat{\theta}_1) &= \sum_{t=1}^T \begin{bmatrix} \frac{\partial^2 \ln L_t}{\partial \mu_1^2} & \frac{\partial^2 \ln L_t}{\partial \mu_1 \partial \sigma_1^2} \\ \frac{\partial^2 \ln L_t}{\partial \sigma_1^2 \partial \mu_1} & \frac{\partial^2 \ln L_t}{\partial (\sigma_1^2)^2} \end{bmatrix}_{\theta_1 = \hat{\theta}_1} \\ &= \frac{1}{\hat{\sigma}_1^4} \sum_{t=1}^T \begin{bmatrix} -\hat{\sigma}_1^2 & -(y_t - \hat{\mu}_1) \\ -(y_t - \hat{\mu}_1) & \frac{1}{2} - \frac{(y_t - \hat{\mu}_1)^2}{\hat{\sigma}_1^2} \end{bmatrix}, \end{aligned} \tag{98}$$

and

$$\begin{aligned}
J(\hat{\theta}_1) &= \sum_{t=1}^T \begin{bmatrix} \left(\frac{\partial \ln L_t}{\partial \mu_1}\right)^2 & \frac{\partial \ln L_t}{\partial \mu_1} \frac{\partial \ln L_t}{\partial \sigma_1^2} \\ \frac{\partial \ln L_t}{\partial \sigma_1^2} \frac{\partial \ln L_t}{\partial \mu_1} & \left(\frac{\partial \ln L_t}{\partial \sigma_1^2}\right)^2 \end{bmatrix}_{\theta_1=\hat{\theta}_1} \\
&= \frac{1}{\hat{\sigma}_1^4} \sum_{t=1}^T \begin{bmatrix} (y_t - \hat{\mu}_1)^2 & -\frac{y_t - \hat{\mu}_1}{2} + \frac{(y_t - \hat{\mu}_1)^3}{2\hat{\sigma}_1^2} \\ -\frac{y_t - \hat{\mu}_1}{2} + \frac{(y_t - \hat{\mu}_1)^3}{2\hat{\sigma}_1^2} & \left(-\frac{1}{2} + \frac{(y_t - \hat{\mu}_1)^2}{2\hat{\sigma}_1^2}\right)^2 \end{bmatrix} \\
&= \frac{1}{\hat{\sigma}_1^4} \sum_{t=1}^T \begin{bmatrix} (y_t - \hat{\mu}_1)^2 & -\frac{(y_t - \hat{\mu}_1)}{2} + \frac{(y_t - \hat{\mu}_1)^3}{2\hat{\sigma}_1^2} \\ -\frac{(y_t - \hat{\mu}_1)}{2} + \frac{(y_t - \hat{\mu}_1)^3}{2\hat{\sigma}_1^2} & \frac{1}{4} + \frac{(y_t - \hat{\mu}_1)^4}{4\hat{\sigma}_1^4} - \frac{(y_t - \hat{\mu}_1)^2}{2\hat{\sigma}_1^2} \end{bmatrix}. \tag{99}
\end{aligned}$$

Three examples of test statistics are now derived depending upon which elements of  $H$  and  $J$  are used.

### Testing for Skewness

Consider the element

$$\begin{aligned}
J_{1,2}(\hat{\theta}_1) + H_{1,2}(\hat{\theta}_1) &= \sum_{t=1}^T \left[ \frac{\partial \ln L_t}{\partial \mu_1} \frac{\partial \ln L_t}{\partial \sigma_1^2} + \frac{\partial^2 \ln L_t}{\partial \mu_1 \partial \sigma_1^2} \right] \\
&= \frac{1}{\hat{\sigma}_1^4} \sum_{t=1}^T \left[ \left( -\frac{(y_t - \hat{\mu}_1)}{2} + \frac{(y_t - \hat{\mu}_1)^3}{2\hat{\sigma}_1^2} \right) - (y_t - \hat{\mu}_1) \right] \\
&= \frac{T}{2\hat{\sigma}_1^3} \left[ \frac{1}{T} \sum_{t=1}^T \frac{(y_t - \hat{\mu}_1)^3}{\hat{\sigma}_1^3} - 0 \right],
\end{aligned}$$

where the last step uses  $T\hat{\mu}_1 = \sum_{t=1}^T y_t$ , and  $T\hat{\sigma}_1^2 = \sum_{t=1}^T (y_t - \hat{\mu}_1)^2$ . The term in brackets is commonly used to test for skewness as

$$\mathbb{E}_0 \left[ \frac{(y_t - \hat{\mu}_1)^3}{\hat{\sigma}_1^3} \right] = 0,$$

if the true model is normal. In which case there is no misspecification and

$$\mathbb{E}_0 \left[ J_{1,2}(\hat{\theta}_1) + H_{1,2}(\hat{\theta}_1) \right] = 0.$$

The usual form of the skewness statistic is

$$Sk = 6 \left[ \frac{1}{T} \sum_{t=1}^T \left( \frac{y_t - \hat{\mu}_1}{\hat{\sigma}_1} \right)^3 - 0 \right]^2,$$

which is distributed asymptotically as  $\chi^2$  with one degree of freedom under the null hypothesis of no skewness.

### Testing for Kurtosis

Consider the element

$$\begin{aligned}
J_{2,2}(\hat{\theta}_1) + H_{2,2}(\hat{\theta}_1) &= \sum_{t=1}^T \left[ \left( \frac{\partial \ln L_t}{\partial \sigma_1^2} \right)^2 + \frac{\partial^2 \ln L_t}{\partial (\sigma_1^2)^2} \right] \\
&= \frac{1}{\hat{\sigma}_1^4} \sum_{t=1}^T \left[ \left( \frac{1}{4} + \frac{(y_t - \hat{\mu}_1)^4}{4\hat{\sigma}_1^4} - \frac{(y_t - \hat{\mu}_1)^2}{2\hat{\sigma}_1^2} \right) + \left( \frac{1}{2} - \frac{(y_t - \hat{\mu}_1)^2}{\hat{\sigma}_1^2} \right) \right] \\
&= \frac{1}{\hat{\sigma}_1^4} \sum_{t=1}^T \left[ \frac{3}{4} + \frac{(y_t - \hat{\mu}_1)^4}{4\hat{\sigma}_1^4} - \frac{3(y_t - \hat{\mu}_1)^2}{2\hat{\sigma}_1^2} \right] \\
&= \frac{T}{4\hat{\sigma}_1^4} \left[ \frac{1}{T} \sum_{t=1}^T \frac{(y_t - \hat{\mu}_1)^4}{\hat{\sigma}_1^4} - 3 \right],
\end{aligned}$$

where the last step uses  $T\hat{\sigma}_1^2 = \sum_{t=1}^T (y_t - \hat{\mu}_1)^2$ . The term in brackets is commonly used to test for kurtosis as

$$E_0 \left[ \frac{(y_t - \hat{\mu}_1)^4}{\hat{\sigma}_1^4} \right] = 3,$$

if the true model is normal. In which case there is no misspecification and

$$E_0 \left[ J_{2,2}(\hat{\theta}_1) + H_{2,2}(\hat{\theta}_1) \right] = 0.$$

The usual form of the kurtosis statistic is

$$Kt = 24 \left[ \frac{1}{T} \sum_{t=1}^T \left( \frac{y_t - \hat{\mu}_1}{\hat{\sigma}_1} \right)^4 - 3 \right]^2,$$

which is distributed asymptotically as  $\chi^2$  with one degree of freedom under the null hypothesis of no kurtosis.

### Testing for Normality

A test of normality is based on a joint test of skewness and kurtosis

$$\begin{aligned}
JB &= Sk + Kt \\
&= 6 \left[ \frac{1}{T} \sum_{t=1}^T \left( \frac{y_t - \hat{\mu}_1}{\hat{\sigma}_1} \right)^3 - 0 \right]^2 + 24 \left[ \frac{1}{T} \sum_{t=1}^T \left( \frac{y_t - \hat{\mu}_1}{\hat{\sigma}_1} \right)^4 - 3 \right]^2,
\end{aligned}$$

which is distributed asymptotically as  $\chi^2$  with two degrees of freedom under the null hypothesis of normality.

Constructing a test statistic using the elements of  $H(\hat{\theta}_1)$  in (98) and  $J(\hat{\theta}_1)$  in (99) corresponding to  $\mu_1$ , yields

$$J_{1,1}(\hat{\theta}_1) + H_{1,1}(\hat{\theta}_1) = 0.$$

This result was also found in an earlier example where the true distribution is Student  $t$  and the misspecified model is normal. The implication of this result from a testing point of view is that this statistic has zero power in testing for misspecification.

### 3.7 Computer Applications

The following examples illustrates the computation of the robust covariance of the quasi-maximum likelihood estimator.

#### 3.7.1 The Effects of Misspecifying a Distribution

Consider again the example where  $y_t$  is *iid* and distributed as standardized Student  $t$  with parameters  $\theta_0 = \{\mu_0, \sigma_0^2, \gamma_0\}$

$$f_0(y_t; \theta_0) = \frac{\Gamma\left(\frac{\gamma_0 + 1}{2}\right)}{\sqrt{\pi\sigma_0^2(\gamma_0 - 2)}\Gamma\left(\frac{\gamma_0}{2}\right)} \left(1 + \frac{(y_t - \mu_0)^2}{\sigma_0(\gamma_0 - 2)}\right)^{-(\gamma_0+1)/2},$$

while the in the misspecified model  $y_t$  is assumed to be normally distributed with parameters  $\theta_1 = \{\mu_1, \sigma_1^2\}$

$$f_1(y_t; \theta_1) = \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left[-\frac{(y_t - \mu_1)^2}{2\sigma_1^2}\right].$$

Recall that the gradient and Hessian of the misspecified log-likelihood are respectively

$$G(\theta_1) = \frac{\partial \ln L}{\partial \theta_1} = \sum_{t=1}^T \begin{bmatrix} \frac{y_t - \mu_1}{\sigma_1^2} \\ -\frac{1}{2\sigma_1^2} + \frac{(y_t - \mu_1)^2}{2\sigma_1^4} \end{bmatrix},$$

$$H(\theta_1) = \frac{\partial^2 \ln L}{\partial \theta_1 \partial \theta_1'} = \sum_{t=1}^T \begin{bmatrix} \frac{1}{\sigma_1^2} & \frac{y_t - \mu_1}{\sigma_1^4} \\ \frac{y_t - \mu_1}{\sigma_1^4} & -\frac{1}{2\sigma_1^4} + \frac{(y_t - \mu_1)^2}{\sigma_1^6} \end{bmatrix}.$$

Setting the gradient vector to zero yields the sample mean and the sample variance as the maximum likelihood estimators  $\hat{\theta}_1 = \{\bar{y}, s^2\}$ .

The aim of this exercise is to reproduce Table 5. To compute the outer product of gradients and the information matrix of the misspecified log-likelihood under the true model,  $T = 500,000$  observations on  $y_t$  are drawn from the Student t distribution with parameters  $\{\mu_0 = 1, \sigma_0^2 = 1\}$  for different values of  $\gamma_0$ . The (average) outer product of the gradients evaluated at  $\hat{\theta}_1$  is computed as

$$\frac{1}{T} \sum_{t=1}^T \begin{bmatrix} \left( \frac{y_t - \bar{y}}{s^2} \right)^2 & \left( -\frac{y_t - \bar{y}}{2s^4} + \frac{(y_t - \bar{y})^3}{2s^4} \right) \\ \left( -\frac{y_t - \bar{y}}{2s^4} + \frac{(y_t - \bar{y})^3}{2s^4} \right) & \left( -\frac{1}{2s^2} + \frac{(y_t - \bar{y})^2}{2s^4} \right)^2 \end{bmatrix},$$

and the (average) information matrix is computed as

$$\frac{1}{T} \sum_{t=1}^T \begin{bmatrix} \frac{1}{s^2} & \frac{y_t - \bar{y}}{s^4} \\ \frac{y_t - \bar{y}}{s^4} & -\frac{1}{2s^4} + \frac{(y_t - \bar{y})^2}{s^6} \end{bmatrix}.$$

### 3.7.2 A Model of US Investment

Consider the simple static investment model

$$\begin{aligned} \ln ri_t &= \beta_0 + \beta_1 \ln ry_t + \beta_2 rint_t + \beta_3 trend_t + u_t \\ u_t &\sim N(0, \sigma^2), \end{aligned}$$

where  $ri_t$  is real investment,  $ry_t$  is real income,  $rint_t$  is the real interest rate expressed as a percentage and  $trend_t$  is a time trend. The model is estimated for U.S. quarterly data from June 1950 to December 2000 ( $T = 203$ ) provided by Greene in his text *Econometric Analysis* (see Table F5.1, p 1083 of the latest edition). The estimated parameters are reported in Table 7 where the standard errors are based respectively on the Hessian, the outer product of the gradients and the robust quasi-maximum likelihood variance estimator. In computing the latter two variances, both the independent ( $P = 0$ ) case and dependent ( $P > 0$ ) case with  $P$  determined as

$$P = \text{Floor} \left( 4 \left( \frac{203}{100} \right)^{2/9} \right) = \text{Floor} (4.682) = 4,$$

are evaluated. The results show large differences in the standard errors based on the Hessian and the outer product of gradients, suggesting some misspecification and the

need to compute the robust standard errors. A comparison of the robust standard errors with  $P = 0$  and  $P = 4$  also shows some large differences suggesting that  $ri_t$  exhibits some dependence and that the robust standard errors based on  $P = 4$  are more appropriate. A result of using the robust standard errors for the investment model is that the semi-elasticity parameter estimate on the real interest rate is not statistically significant at the 5% level, whereas it is statistically significant if the standard errors based on the Hessian are used.

Table 7:

Investment model estimated using U.S. quarterly data from June 1950 to December 2000. Standard errors based on the square root of the diagonal elements of the variance-covariance matrix.

Parameter	Estimate	Standard errors				
		Hessian	OPG (P=0)	OPG (P=4)	QMLE (P=0)	QMLE (P=4)
$\beta_0$	-7.900	1.224	1.654	0.931	1.140	2.044
$\beta_1$	1.763	0.164	0.222	0.125	0.152	0.273
$\beta_2$	-0.443	0.225	0.207	0.157	0.307	0.495
$\beta_3$	-0.439	0.136	0.184	0.102	0.127	0.230
$\sigma^2$	0.007	0.001	0.001	0.001	0.001	0.002

### 3.7.3 Conditional Nonnormality and QMLE in Volatility Models

An important feature of the volatility models discussed so far is the role of conditional normality. Whilst the combination of conditional normality and GARCH conditional variances yields unconditional financial returns distributions which are leptokurtotic, fat-tails and a sharp peak relative to the normal distribution, in practice this class of models is not able to capture all of the leptokurtosis in the data. There are three approaches to generalise the GARCH model to account for leptokurtosis.

1. The parametric solution is to replace the conditional normality specification by a nonnormal parametric distribution that displays the correct behaviour. We have already done this.

2. A semi-parametric approach is to specify the conditional mean and conditional variance equations parametrically, but to use a nonparametric density estimator for the distribution of the disturbance term. Estimation would proceed as follows:

**Step 1** Choose starting values for the parameters  $\{\gamma_0, \gamma_1, \alpha_0, \alpha_1, \beta_1\}$ .

**Step 2** For each  $t$ , construct the conditional mean  $\gamma_0 + \gamma_1 x_t$ , the disturbance  $u_t = y_t - \gamma_0 - \gamma_1 x_t$ , and the conditional variance  $h_t = \alpha_0 + \alpha_1 u_{t-1}^2 + \beta_1 \cdot h_{t-1}$ .

**Step 3** For each  $t$ , construct the standardised disturbance

$$z_t = \frac{\gamma_0 + \gamma_1 x_t}{h_t}.$$

**Step 4** Use all  $z_t$  to estimate  $f(z)$  using a nonparametric density estimator, say  $f_{np}(z)$ .

**Step 5** Evaluate  $f_{np}(z)$  at each observation and hence compute  $f_{np}(z_t)$ .

**Step 6** For each  $t$ , compute the log of the likelihood  $\ln L_t = \ln f_{np}(z_t) - 0.5 \ln h_t$ , and maximize  $\ln L = \sum_t \ln L_t$  using a gradient algorithm.

3. A simpler approach to the above two methods is to compute quasi maximum likelihood standard errors (QMLE), also known as Bollerslev-Wooldridge standard errors in this context. The QMLE standard errors are computed as the square root of the following matrix

$$H^{-1} \left( \hat{\theta} \right) J \left( \hat{\theta} \right) H^{-1} \left( \hat{\theta} \right), \quad (100)$$

where

$$H = \frac{\partial^2 \ln L}{\partial \theta \partial \theta'} \quad (101)$$

$$J = G'G + \sum_{I=1}^L w_i (G'G_{-i} + (G'G_{-i})'), \quad (102)$$

where  $G$  is the matrix of gradients

$$G \left( \hat{\theta} \right) = \begin{bmatrix} \frac{\partial \ln L_1}{\partial \theta} \\ \vdots \\ \frac{\partial \ln L_T}{\partial \theta} \end{bmatrix}_{\theta=\hat{\theta}},$$

and  $w_i = 1 - \frac{i}{L+1}$  are the Newey-West weights. The vector  $\hat{\theta}$ , represents the maximum likelihood parameter estimates at the final iteration. Notice that the only

difference between the MLE and QMLE estimates are the standard errors, in which case the point estimates of  $\theta$  for the two estimators are exactly the same.