Maximum Likelihood and Structural Models in Microeconometrics

Lecture and class notes for the M.Sc in Development Economics
(Quantitative Methods course)

Simon Quinn*

Michaelmas 2012 and Hilary 2013

Outline of these classes

Overview

These notes provide some background for a series of classes and lectures that I will be giving this year for the Quantitative Methods course. The classes and lectures focus on the themes of structural models and estimation with maximum likelihood; we will discuss (i) maximum likelihood estimation (in Michaelmas), (ii) estimation of a structural model using maximum likelihood (in Hilary), and (iii) some more general discussion on structural modelling (also in Hilary).

*Post-doctoral researcher, All Souls College (Oxford): simon.quinn@economics.ox.ac.uk. Without implicating them in any shortcomings, I would like to thank Marcel Fafchamps, Avi Feller, Matthew Polisson, Victoria Prowse, Francis Teal, Nicolas Van de Sijpe and Andrew Zeitlin for their help and suggestions on these notes.

Schedule of classes and lectures

The schedule of lectures and classes will be as follows.

MICHAELMAS		
Week 3	Class 3	**Introducing Maximum Likelihood in Stata**
Week 3	Lecture 5	**Maximum Likelihood Estimation**
HILARY		
Week 3	Lecture 21	**Microeconometric Structural Models I**
Week 3	Lecture 22	**Microeconometric Structural Models II**
Week 3	Class 11	**Using Structural Models I**
Week 4	Class 12	**Using Structural Models II**

Notes, slides and readings

You should read all of the starred references ('$\star$'). You are not required or expected to read the others.

You should also read these notes; they provide the background to our lectures and classes. I will use slides for the lectures and classes themselves. I will make these slides available online *after* each class and lecture. It is likely that there will be some things in these notes that we do not have time to cover in class, and we *may* cover some things in class that are not covered in these notes. Though we will focus in class on the most important issues, please consider all of the lectures, all of the notes and all of the starred readings to be potentially relevant for the exam.

1 Class and Lecture: Maximum Likelihood Estimation

References:

- [★] HENDRY, D., AND B. NIELSEN (2007): *Econometric modeling: A likelihood approach*. Princeton University Press, chapters 1–3 and chapter 5. This is a very useful introduction to maximum likelihood. Even if you are not interested in understanding the details of maximum likelihood in Stata, *you should read* chapters 1–3 and chapter 5 of this book.
- GOULD, W., J. PITBLADO, AND W. SRIBNEY (2006): *Maximum likelihood estimation with Stata*. Stata Press. This is the 'official' guide to maximum likelihood estimation in Stata; the authors of the book also wrote the relevant Stata commands. If you are interested in learning more about the details of coding your own likelihood estimator in Stata — for example, for your extended essay — then you will find this book very helpful. However, you do *not* need to consult this book as a general text on maximum likelihood.

In this lecture, we will look at a class of estimators known as *maximum likelihood estimators*. There are two main reasons that this will be important. First, maximum likelihood estimators are *very common* in empirical work. If you have done any empirical research before, it is likely that you have used a maximum likelihood estimator — even if you didn't realise you were doing so! Understanding the concept of a maximum likelihood estimator — and the relative strengths and weaknesses of those estimators — is therefore very important for doing empirical work, and for interpreting others' empirical results. Second, maximum likelihood estimation is *very flexible*. It is therefore very important if we want to tie a theoretical model directly to an empirical estimation. This is known as *structural modelling*, and we will discuss it in two lectures and two classes in Hilary.

1.1 The concept of maximum likelihood

I would like to outline the key concepts of maximum likelihood by considering a simple model and a simple empirical problem. Suppose that we have a cross-section of N individuals, indexed $i = 0, \ldots, N$, and that we observe a single outcome variable y_i (for example, log earnings) and a single explanatory variable x_i (for example, years of education). Suppose we believe that there is a very simple relationship between x_i and y_i, with an additive error term, ε_i:

$$y_i = \beta x_i + \varepsilon_i. \tag{1.1}$$

For the moment, we don't assume anything in particular about the distribution of ε, but we will see later that distributional assumptions about ε play a fundamental role for estimating models by maximum likelihood. We are interested simply in estimating β. (Note that, for simplicity, we are not even allowing here for an intercept term — in reality, this would be too simple, but hopefully it will help us explore the concepts of maximum likelihood.) You will know that we could estimate this very easily by OLS, using Stata's `reg` command. However, I would like to consider this problem in more detail, in order to outline the principles of maximum likelihood.

There are many different methods that we could choose to estimate β. We could think of some very sensible ways (for example, running OLS), or — for the sake of argument — we could think of some very silly ways (for example, guessing a number between 1 and 10). Therefore, we need some *criterion* that tells us what constitutes a 'good' estimate of β. You have already seen a very simple and logical criterion: "choose a value for $\hat{\beta}$ that minimises the sum of squared differences between y_i and $\hat{\beta}x_i$". This is the criterion used to define the OLS estimator. If we wanted to be formal, we could write:

$$\hat{\beta}_{OLS} = \arg\min_{\beta} \sum_{i=1}^{N} (y_i - \beta x_i)^2 . \tag{1.2}$$

We could define $S(\beta; y_1, \ldots, y_N, x_1, \ldots, x_N) = \sum_{i=1}^{N} (y_i - \beta x_i)^2$, and call $S(\beta; y_1, \ldots, y_N, x_1, \ldots, x_N)$ the *objective function* for our problem. That is, $S(\cdot)$ is *the function which tells us, for some given data, what makes a 'good' estimate*.

It would be tempting to think of this function — the sum of squares — as the *only* good objective function for this problem. However, this is not the case. For example, we might choose instead to minimise the *sum of absolute deviations*, giving us an objective function $D(\beta; y_1, \ldots, y_N, x_1, \ldots, x_N) = \sum_{i=1}^{N} |y_i - \beta x_i|$. In that case, we would get an estimator known as a LAD ('least absolute deviation') estimator:[1]

$$\hat{\beta}_{LAD} = \arg\min_{\beta} \sum_{i=1}^{N} |y_i - \beta x_i| . \tag{1.3}$$

These concepts are easily sufficient for dealing with simple models like this one. However, what if we want a more flexible estimator, so that we can deal with more complicated models? In that case, we might want to use a *maximum likelihood* estimator. A maximum likelihood estimator of β is *an estimator that tells us* ***the most likely value*** *for β, given the data that we have observed*. This is a *different* criterion to either of the criteria that we have discussed so far. That is, it may be *sensible* to estimate β by minimising the sum of squared deviations, or the sum of absolute deviations, but there is no necessary reason that either $\hat{\beta}_{OLS}$ or $\hat{\beta}_{LAD}$ estimates the *most likely* value for β.

Hopefully, this sounds like a good idea. However, the devil lies in the details — in order to find the 'most likely value for β, given the data that we have observed', we need to know (or assume) precisely *how* the likelihood of observing a given value for β depends upon the observed data. Sometimes, it can be difficult to figure this out — and, in many cases, it is impossible! Fortunately, it is relatively straightforward in this simple case. However, to think carefully about this, we need to draw a fundamental distinction between the concept of a *sample* and the concept of a *population*.

[1] We could run this estimator using Stata's `qreg` command.

1.2 The concept of population

The word 'population' is a common one in society. Usually, we use the term to talk about *a complete set of things* — as in, "the population of the UK is about 61 million people". Consider our data on the relationship between education and earnings in South Africa. It would be tempting, in that context, to discuss 'population' as referring to *all of the income-earners in South Africa*. However, for our purposes, I would like to consider population slightly differently: I would like to think about a population *not* as a finite collection of *things* (workers, firms, *etc*), but as an '*infinite hypothetical population*' from which finite samples are drawn.

This characterisation has a long history in statistics and econometrics. For example, R.A. Fisher said this in his seminal text *Statistical Methods for Research Workers* (10th edition, 1948, page 41, emphasis in original):

> The idea of an infinite **population** distributed in a **frequency distribution** in respect of one or more characters is fundamental to all statistical work. From a limited experience, for example, of individuals of a species, or of the weather of a locality, we may obtain some idea of the infinite hypothetical population from which our sample is drawn, and so of the probable nature of future samples to which our conclusions are to be applied.

In our starred reading, Hendry and Nielsen (2007, p.3) consider the example of estimating the probability of a newborn child in the UK being a boy. They describe the sample/population distinction in that context as follows:

> We will think of a sample distribution as a random realization from a population distribution. In the above example, the sample is all newborn children in the UK in 2004, whereas the population distribution is thought of as representing the biological causal mechanism that determines the sex of children. Thus, although the sample here is actually the population of all newborn children in the UK in 2004, the population from which that sample is drawn is a hypothetical one.

We distinguish between *population parameters* and *sample parameters* by the use of the 'hat' — which you will have seen already in these notes. Therefore, when we assumed earlier that

$$y_i = \beta x_i + \varepsilon_i, \tag{1.1}$$

we were assuming an *underlying causal mechanism* explaining how x causes y. That is, we were *specifying a population relationship*. When we consider different estimators — whether OLS, LAD, maximum likelihood or so on — we are considering different possible ways of *estimating* β using a *finite sample from the true population*. This is why we use the 'hat'; to denote that $\hat{\beta}_{OLS}$, $\hat{\beta}_{LAD}$, $\hat{\beta}_{ML}$ and so forth are *statistics estimating* β *from a finite sample*, in contrast to the 'true' β itself.

1.3 Distributional assumptions and the log-likelihood function

THE
NORMAL
LAW OF ERROR
STANDS OUT IN THE
EXPERIENCE OF MAN
AS ONE OF THE BROADEST
GENERALIZATIONS OF NATURAL
PHILOSOPHY • IT SERVES AS THE
GUIDING INSTRUMENT IN RESEARCHES
IN THE PHYSICAL AND SOCIAL SCIENCES AND
IN MEDICINE AGRICULTURE AND ENGINEERING •
IT IS AN INDISPENSABLE TOOL FOR THE ANALYSIS AND THE
INTERPRETATION OF THE BASIC DATA OBTAINED BY OBSERVATION AND EXPERIMENT

W.J. Youden

To develop a maximum-likelihood estimator, we need to *assume a probability distribution* for the unobservable, ε. (Note that this is *not* something that we had to do for $\hat{\beta}_{OLS}$ (or, for that matter, $\hat{\beta}_{LAD}$).) Given the preceding discussion, we can now be very clear: what we are assuming is the *true population probability distribution* for ε.

For simplicity, we will assume that ε has a normal distribution with mean 0 and variance σ^2, and that ε has this distribution for *all* observations.

Assumption 1.1 *The unobservable, ε, has an identical normal distribution with mean 0 and variance σ^2:*

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2). \tag{1.4}$$

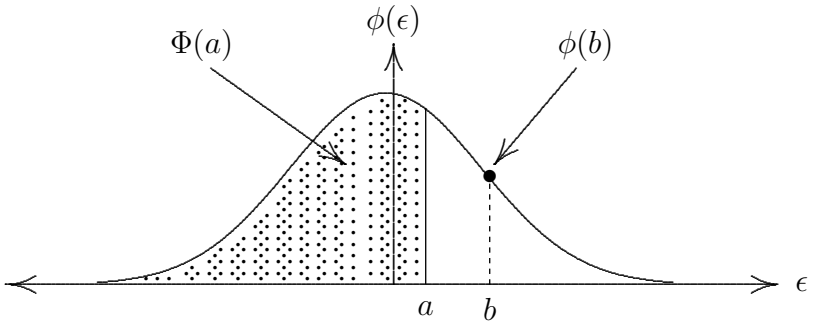
From this assumption, we will — with the help of a few more simplifying assumptions — be able to derive the estimator $\hat{\beta}_{ML}$.

I would like to build incrementally to this result. Let's start, then, by considering the more general case where $\varepsilon \sim \mathcal{N}(\mu, \sigma^2)$ (that is, the case where ε has mean μ, rather than mean 0). In that case, we know that the *probability density function* for ε is:

$$f(\varepsilon_i; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{(\varepsilon_i - \mu)^2}{2\sigma^2}\right). \tag{1.5}$$

This function simply describes the shape of the normal distribution. When $\mu = 0$ and $\sigma^2 = 1$, we have the *standard normal distribution*, often referred to simply by the function $\phi(\cdot)$. Figure 1.1 shows this.

Figure 1.1: **The standard normal: probability density ($\phi(\cdot)$) and cumulative density ($\Phi(\cdot)$)**



Therefore, we know that, if $\mu = 0$,

$$f(\varepsilon_i; \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{\varepsilon_i^2}{2\sigma^2}\right). \tag{1.6}$$

This is interesting, but not very useful. What we *really* want is a statement about the probability of observing y_i (rather than ε_i). For this, we need another assumption.

Assumption 1.2 *The variables ε_i and x_i are independent. That is, knowing x_i tells us nothing about the probability of obtaining any particular value of ε_i:*

$$\varepsilon_i \,|\, x_i \sim \mathcal{N}(0, \sigma^2). \tag{1.7}$$

This is a *very strong assumption* — it is saying that *everything affecting income apart from education is independent of the amount of education someone obtains*. So, in the South African example, it is saying that the amount of education someone undertakes is independent of their intelligence, ability, location, gender, family background, *etc*. Indeed, if we believed this assumption generally, we could probably do away with most econometric techniques (for example, there would be no need to use instrumental variables, no need to use panel data, *etc*). However, we will make this strong assumption now to illustrate the concept of maximum likelihood in a simple context.

From equation 1.1, it follows that $\varepsilon_i = y_i - \beta x_i$. Equation 1.7 tells us that we can rewrite our probability density as a function of y_i and x_i, rather than ε_i. Therefore, we can rewrite equation 1.6 as:

$$f(y_i \,|\, x_i; \beta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{(y_i - \beta x_i)^2}{2\sigma^2}\right). \tag{1.8}$$

Equation 1.8 tells us *the probability density of observing y_i, conditional on x_i, given some true parameters β and σ^2*. This is useful — but, fortunately, we have more than one observation on y_i and x_i! Therefore, we need to write the probability for observing a *collection* of x_i and y_i for multiple respondents. To do this, we need one more assumption.

Assumption 1.3 *The unobservable, ε, is independent between different observations. That is, for any $i \neq j$, we would learn nothing about the probability of observing any particular value for ε_i if we could observe ε_j:*

$$\varepsilon_i \,|\, \varepsilon_j \sim \mathcal{N}(0, \sigma^2). \tag{1.9}$$

This is quite a strong assumption, too. It will be violated, for example, if two individuals in the sample are hit by a *common shock*, or other *common unobservable* — for example, if individuals i and j are in the same household and therefore subject to the same 'household effect'.

Nonetheless, if we make this assumption, we can write the *joint probability density* across multiple observations. Remember that *the joint probability of two independent events equals the product of their individual probabilities* — so, for example, if the weather in Oxford is independent of the weather in Brisbane,

$$\begin{aligned}&\Pr(\text{rain today in Oxford and rain today in Brisbane})\\ &\quad = \Pr(\text{rain today in Oxford}) \times \Pr(\text{rain today in Brisbane}).\end{aligned}$$

Using this rule, we can write the *joint probability density* for our sample — it is simply the product of the probability densities for the separate individuals:[2]

$$f(y_1, \ldots, y_N \,|\, x_1, \ldots, x_N; \beta, \sigma^2) = f(y_1; x_1, \beta, \sigma^2) \times f(y_2; x_2, \beta, \sigma^2) \times \ldots \times f(y_N; x_N, \beta, \sigma^2) \tag{1.10}$$

$$\begin{aligned}&= \prod_{i=1}^{N} f(y_i; x_i, \beta, \sigma^2)\\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \cdot \prod_{i=1}^{N} \exp\left(-\frac{(y_i - \beta x_i)^2}{2\sigma^2}\right)\\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \cdot \exp\left(-\frac{\sum_{i=1}^{N}(y_i - \beta x_i)^2}{2\sigma^2}\right)\\ &= \left(2\pi\sigma^2\right)^{-\frac{1}{2}N} \cdot \exp\left(-\frac{\sum_{i=1}^{N}(y_i - \beta x_i)^2}{2\sigma^2}\right).\end{aligned} \tag{1.11}$$

So we are almost done. However, what we have written is the *probability of observing a combination of $y_1, \ldots, y_N$, conditional on $x_1, \ldots, x_N$, **given the true parameters** β **and** σ^2*. To write the likelihood function we write the *same thing*, but interpret it as stating *probability of the true parameters being some β and σ^2, **given the data we have observed***. That is, we write the likelihood function as:

$$L(\beta, \sigma^2; y_1, \ldots, y_N, x_1, \ldots, x_N) = \left(2\pi\sigma^2\right)^{-\frac{1}{2}N} \cdot \exp\left(-\frac{\sum_{i=1}^{N}(y_i - \beta x_i)^2}{2\sigma^2}\right). \tag{1.12}$$

[2] I am overlooking here the distinction between the concepts of *probability* and *probability density*. This is something that we can discuss if you wish, but I would prefer to keep things simpler in these notes.

1.4 Maximising the (log-)likelihood

We can now define the maximum likelihood estimator, β_{ML}, in this context:

$$\hat{\beta}_{ML} = \arg\max_{\beta} L(\beta, \sigma^2; y_1, \ldots, y_N, x_1, \ldots, x_N). \tag{1.13}$$

That is, β_{ML} is *the value of β that maximises the likelihood function*. This can sometimes be a messy and difficult problem to solve, either analytically (that is, using algebra) or numerically (that is, using a sophisticated 'guess and check' algorithm on a computer). However, fortunately, we don't *actually* have to maximise $L(\cdot)$; we can, alternatively, maximise *any* monotonically increasing function of $L(\cdot)$. One extremely convenient function is the log function — we can write:

$$\hat{\beta}_{ML} = \arg\max_{\beta} \ \ln L(\beta, \sigma^2; y_1, \ldots, y_N, x_1, \ldots, x_N), \tag{1.14}$$

because $\ln(\cdot)$ is a monotone increasing transformation (and given that $L(\beta, \sigma^2; y_1, \ldots, y_N, x_1, \ldots, x_N) > 0$).

Therefore, I would like to define $\ell(\beta, \sigma^2)$ as the *log of the likelihood function* (or 'log-likelihood' for short):

$$\ell(\beta, \sigma^2) \equiv \ln L(\beta, \sigma^2; y_1, \ldots, y_N, x_1, \ldots, x_N) \tag{1.15}$$

$$= -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{\sum_{i=1}^{N}(y_i - \beta x_i)^2}{2\sigma^2}. \tag{1.16}$$

In many cases, we cannot write a 'closed form' solution for the estimator that maximises the log-likelihood function; in that case, we can optimise it numerically (which we will do shortly in Stata). However, in this case, we *can* find an analytical solution. We do this simply by differentiating the log-likelihood function with respect to β to find its global maximum:

$$\left.\frac{\partial \ell(\beta, \sigma^2)}{\partial \beta}\right|_{\beta=\hat{\beta}_{ML}} = \frac{1}{\sigma^2}\sum_{i=1}^{N}(y_i - \beta x_i)\cdot x_i \tag{1.17}$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{N}\left(x_i y_i - \beta x_i^2\right) = 0 \tag{1.18}$$

This implies, then, that

$$\sum_{i=1}^{N}\left(x_i y_i - \hat{\beta}_{ML}\cdot x_i^2\right) = 0 \tag{1.19}$$

$$\Leftrightarrow \sum_{i=1}^{N} x_i y_i = \hat{\beta}_{ML}\sum_{i=1}^{N} x_i^2 \tag{1.20}$$

$$\therefore \hat{\beta}_{ML} = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2}. \tag{1.21}$$

Hopefully, this formula looks familiar! This, of course, is the same formula for $\hat{\beta}_{OLS}$. And this is no coincidence: if you look again at the log-likelihood function, you will see that maximising with respect to β *requires* us to minimise the sum of squared deviations. Therefore, we can say that OLS *is* a maximum likelihood estimator, *if* we believe that ε is drawn from a normal distribution.

1.5 Maximum likelihood in Stata

Stata has a command designed specifically for doing maximum likelihood estimation with customised likelihood functions: **the `ml` command**. Aside from some general discussion on principles of maximum likelihood, the purpose of the first computer class is to discuss the use of that command.

What does `ml` do? The `ml` command does several useful things.

(i) It accepts a user-specified likelihood function.

(ii) It can do some *limited* checking on the way this function has been entered (though it *cannot* check that the function makes sense in the context of the economic model, of course!).

(iii) It chooses values of the relevant parameters to *maximise* this function. It does this *numerically*, rather than *analytically* — that is, `ml` uses a clever 'guess and check' algorithm to try to find a maximum (rather than taking derivatives and solving algebraically, as we just did).[3]

(iv) It estimates confidence intervals and performs hypothesis testing.

How does `ml` work? The best way to learn how to use `ml` is to work through some examples. We will do this in a moment. However, first, we should see some general overview of how `ml` works.[4] To maximise a likelihood function in `ml`, we should do the following.

(i) Write an economic or statistical model and derive the log-likelihood function.

(ii) Write a short program in Stata to calculate that log-likehood function. Our program — coded using the '`program`' command — will accept data as *inputs* and will provide the resulting log-likelihood value as *output*.

(iii) Tell Stata that the program estimates a log-likelihood function, and tell Stata what data will enter the model and what algorithms we would like to use to maximise the function. This will involve a command like:
`. ml model lf` {*name of program*} {*name of parameters and variables*}, `tech`({*optimisation techniques*}) `constraints`({*numbered list of constraints imposed*})

[3] The term 'guess and check' doesn't do the algorithms justice, of course — numerical optimisation methods form a massive distinct field of study, but we will leave it as 'guess and check' for our purposes!

[4] This overview very closely follows Gould et al (2006, pp.48–49).

(iv) Perform some basic checks on the program:
```
. ml check
```

(v) Tell Stata to search for useful starting values for the parameters:
```
. ml search
```

(vi) Tell Stata to maximise the log-likelihood function:
```
. ml max
```
(or, for more troublesome log-likelihood functions, '`ml max, difficult`').

(vii) Graph the convergence to the maximum:
```
. ml graph
```

Clearly, we need an example...

1.6 An empirical example: OLS as a maximum likelihood estimator in Stata

Let's return to an example from an earlier computer class: the South African wage curve data. You will remember that, using OLS to regress log earnings on years of education, you obtained the following.

```
. clear

. use sa_wage_curve_1

. reg logwphy edyrs

      Source |       SS       df       MS              Number of obs =    6980
-------------+------------------------------           F(  1,  6978) = 2486.11
       Model |  2242.29713     1  2242.29713           Prob > F      =  0.0000
    Residual |  6293.66923  6978  .901930242           R-squared     =  0.2627
-------------+------------------------------           Adj R-squared =  0.2626
       Total |  8535.96636  6979  1.22309304           Root MSE      =   .9497

------------------------------------------------------------------------------
     logwphy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       edyrs |    .138002   .0027677    49.86   0.000     .1325764    .1434276
       _cons |   .4611801    .024596    18.75   0.000     .4129645    .5093957
------------------------------------------------------------------------------
```

Let's now use the `ml` command to estimate this *same* model using maximum likelihood. As before, let's fit a model of the form:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \tag{1.22}$$

where y_i is log earnings and x_i is years of education. To estimate this model as a maximum likelihood estimator, we need to make an assumption about the distribution of ε. Let's assume that

ε_i is *independently and identically distributed* as a normal distribution with mean zero and variance σ^2:

$$\varepsilon_i \,|\, x_i \sim \text{i.i.d. } \mathcal{N}(0, \sigma^2). \tag{1.23}$$

The file `OLS_ML.do` implements this model as a maximum likelihood estimator. Let's run it...

First, I want to clear the memory, load the data and then tell Stata to run the file.

```
. clear

. use sa_wage_curve_1

. do ols_ml
```

The file will start by clearing the program 'MyOLS' from memory, if it's already there.

```
. capture program drop MyOLS
```

Next, it will define 'MyOLS' to calculate the log-likelihood function. It uses Stata's built-in function '`normalden`' to calculate the density of a normal distribution with mean 'beta' and standard deviation 'sigma'.[5]

```
. program MyOLS
    args lfn beta sigma
    quietly {
                replace `lfn' = ln(normalden($ML_y1, `beta', `sigma'))
            }
  end
```

Having defined the log-likelihood function, we should tell Stata what data will enter the model and what algorithms should be used for maximisation.

```
. ml model lf MyOLS (logwphy = edyrs) /sigma, tech(bfgs 5 dfp 5 nr 5 bhhh 5)
```

Let's check that the function and model pass some basic tests.

```
. ml check
```

We may be able to improve the maximisation by asking Stata to search for sensible starting values.

```
. ml search
initial:       log likelihood =  -12438.44
improve:       log likelihood =  -12438.44
rescale:       log likelihood = -12403.364
rescale eq:    log likelihood = -11375.412
```

[5] You can learn more about `normalden` by typing '`help normalden`'.

At last, we're ready to maximise our log-likelihood function and obtain some results!

```
. ml max

initial:       log likelihood = -11375.412
rescale:       log likelihood = -11375.412
rescale eq:    log likelihood = -11375.412
(setting optimization to BFGS)
Iteration 0:   log likelihood = -11375.412
Iteration 1:   log likelihood = -11179.421  (backed up)
Iteration 2:   log likelihood = -10634.969  (backed up)
Iteration 3:   log likelihood = -10562.157  (backed up)
Iteration 4:   log likelihood = -10208.066
(switching optimization to DFP)
Iteration 5:   log likelihood = -9819.9101
Iteration 6:   log likelihood = -9552.0497
Iteration 7:   log likelihood = -9543.6422
Iteration 8:   log likelihood = -9543.0248
Iteration 9:   log likelihood = -9542.9599
(switching optimization to Newton-Raphson)
Iteration 10:  log likelihood = -9542.9597

                                                  Number of obs   =       6980
                                                  Wald chi2(1)    =    2486.81
Log likelihood = -9542.9597                       Prob > chi2     =     0.0000

------------------------------------------------------------------------------
     logwphy |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
eq1          |
       edyrs |   .1380017   .0027673    49.87   0.000     .1325778    .1434256
       _cons |   .4611829   .0245925    18.75   0.000     .4129826    .5093833
-------------+----------------------------------------------------------------
sigma        |
       _cons |   .9495643   .0080368   118.15   0.000     .9338126    .9653161
------------------------------------------------------------------------------
```

Finally, we should graph the convergence.

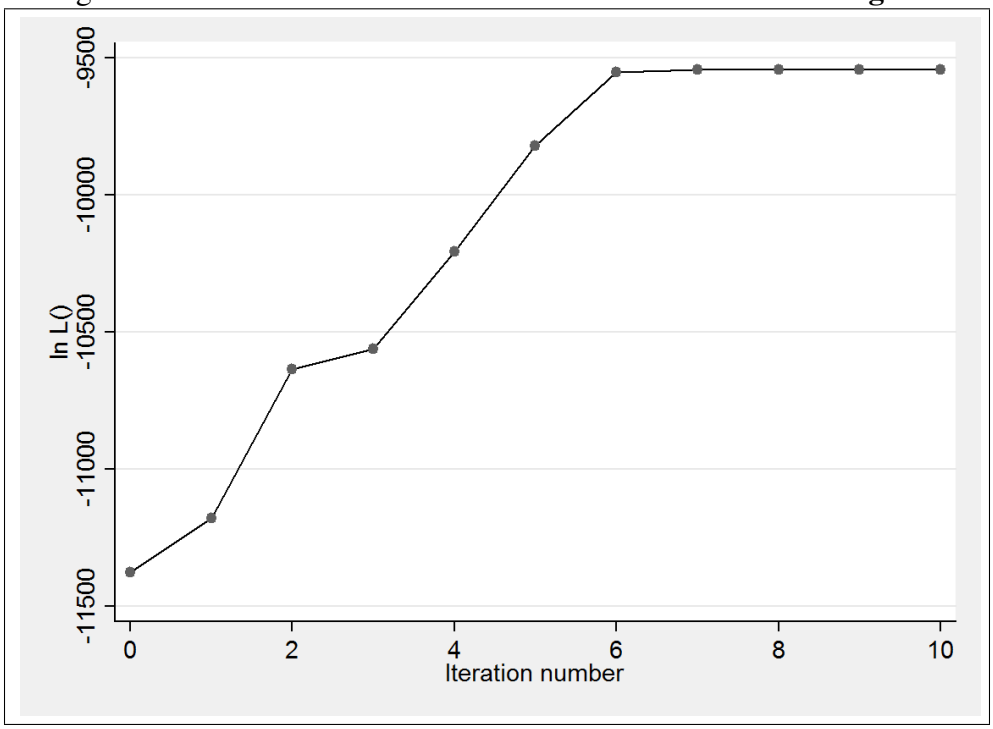
```
. ml graph

. exit

end of do-file
```

Figure 1.2 shows the graph of convergence that results.

Figure 1.2: **OLS as a maximum likelihood estimator: Convergence**



As we expected, the maximum likelihood estimation provides the same results as the OLS regression (there's some disagreement about the last few decimal places, but let's not worry about that!). Of course, we would never use `ml` to fit an OLS regression in practice — it's much faster, much simpler and much easier just to use `reg` — but this simple example is a useful illustration of the way that `ml` works.

Notice that `ml` produces two results that `reg` didn't report: the maximised value of the log-likelihood function (that is, -9542.9597) and an estimate of σ (that is, 0.9495643). You might wonder whether it's possible to recover these values from the original `reg` command; it *is* possible, using the `ereturn` command. Try entering the following.

```
. reg logwphy edyrs
. ereturn list
. display e(rmse)
. display e(ll)
```

1.7 Problems and warnings...

All econometric methods come with fine print, and maximum likelihood is no exception. Before going further, we should note some of the potential problems associated with this approach. There are several.

1.7.1 Maximum likelihood and endogeneity

It is vital to remember the purpose that maximum likelihood serves: it provides a useful method of *estimating a particular econometric model* — it does *not*, of course, provide a general solution to the problem of endogeneity. That is, *any maximum likelihood estimation is only as good as the assumptions that justify it*. In the previous example, we were required to assume that everything affecting income apart from education is independent of the amount of education someone obtains. This is a very strong assumption, and it is *not* an assumption that the maximum likelihood method can directly address. Imagine, for example, that you presented these estimation results at a seminar, and imagine that someone said, "We really shouldn't believe this as a causal relationship, because there are unobserved factors that probably affect *both* education *and* earnings — for example, workers' differing intelligence." This may be a valid criticism, and it would obviously *not* be valid to respond, "But I've estimated by maximum likelihood". Maximum likelihood may be a useful method of estimation in many contexts, but there is nothing inherent in maximum likelihood that deals generally with the problem of endogeneity: again, *maximum likelihood is only as good as the assumptions that justify it*.

1.7.2 Maximum likelihood and convergence

In *theory*, the `ml` command stops running when it has found the *global maximum* to our likelihood function. However, in *practice*, this is not always the case — `ml` can sometimes (i) find a *local maximum* rather than the *global maximum*, and/or (ii) make bold and unjustified claims about having converged upon a maximum, or (iii) throw its proverbial arms in the air and give up. We need to have some sense of when this might be happening!

Unfortunately, short of finding an analytical expression for the maximum likelihood estimator (in which case we don't even need to use the `ml` command), we can *never* be *absolutely* sure that `ml` has reached the global maximum.[6] However, I would like to suggest several rules of thumb for using `ml` sensibly. They are as follows.

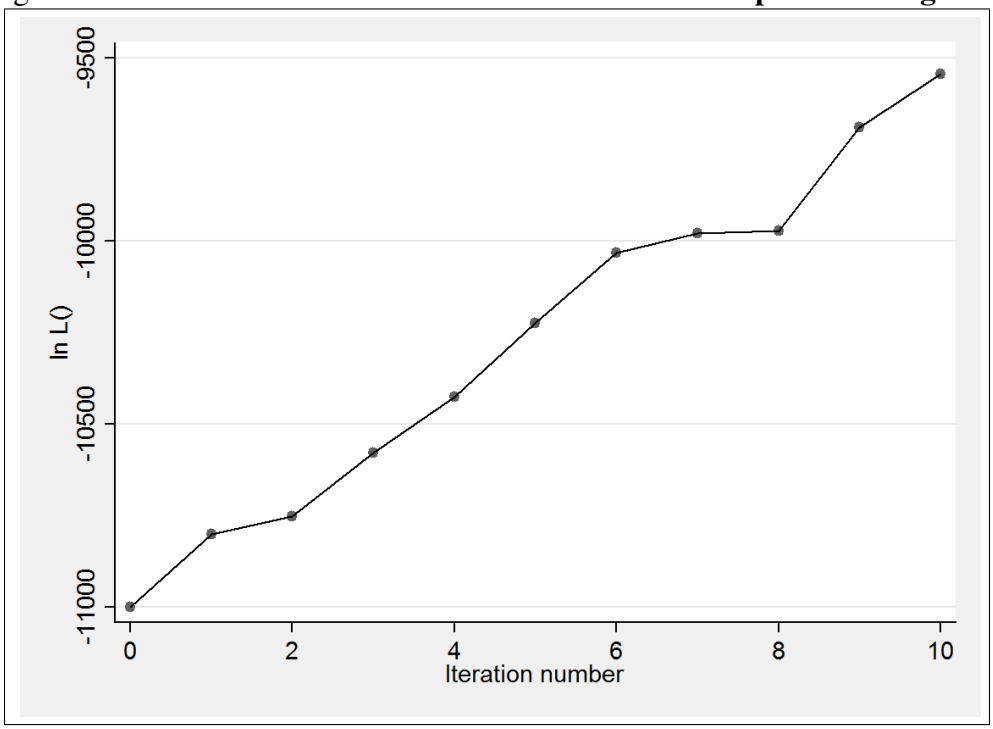
[6] Indeed, we may not know for sure that a unique global maximum even exists!

(i) ***Always run the same estimation several times.*** We should be more suspicious of results in which:

- Stata stops at different values of the log-likelihood on different estimations of the same problem; or
- Stata stops at the *same* value of the log-likelihood but those estimations produce substantially different estimates.[7]

(ii) ***Always check that Stata has converged 'slowly' to the final log-likelihood, rather than jumped suddenly.*** Figure 1.3 is an example of where Stata has jumped quite suddenly to the maximum; you should compare it to Figure 1.2. You can check this by the `ml graph` command, or just by looking at the different log-likelihoods reported at different steps of convergence. The 'convergence' in Figure 1.3 is not necessarily *wrong*, but it is certainly *suspect* — and would justify even more care in running and re-running the estimation.

Figure 1.3: **OLS as a maximum likelihood estimator: Suspect 'convergence'**



(iii) ***Never accept results for which the final iteration is not 'clean'.*** Usually, `ml` will report a series of iterations as it searches for the maximum of the log-likelihood function. For example, `ml` may say something like:

[7] If this second case arises, it is likely that your model is 'under-identified', so that Stata has climbed to a 'ridge' on the log-likelihood function, rather than a single 'point'. We will talk more about this in the next lecture.

```
Iteration 2:  log likelihood = -4859.2522.
```

In doing this, it is common for `ml` to complain that it has entered a region in the parameter space in which the log-likelihood is not locally concave; for example, `ml` may say something like this:

```
Iteration 3:  log likelihood = -4849.7574 (not concave).
```

You might get other messages, too. For example, you will commonly see messages like '`(backed up)`', '`numerical derivatives are approximate`', and/or '`flat or discontinuous region encountered`'.

This is generally no problem — ***unless such a message occurs on the final line of the iteration***. In that case, you should *always* estimate again — Stata is trying to tell you (albeit quite obliquely) that it couldn't really find a local maximum.

(iv) ***Never accept results for which, for* any *estimated coefficient, Stata was unable to estimate the standard error.*** If you choose to *restrict* a coefficient — something that we will discuss in Class 2 of next term — then you will simply get '`.`' instead of a standard error. This is absolutely correct — after all, Stata cannot estimate a confidence interval on a parameter that you claim to know with certainty. However, this is the *only* time that Stata should return '`.`' instead of a standard error. If an estimated parameter returns '`.`' as an estimated standard error, you should discard *all* the parameter estimates and estimate again. (That is, the '`.`' is suggesting that the log-likelihood is not strictly concave at the estimated maximum — and this means that we should treat *all* the parameter estimates as suspect.)

1.7.3 Properties of maximum likelihood estimates

All maximum likelihood estimators share two important properties:

- Estimates are *consistent*, and
- Estimates are *asymptotically efficient*.

These are important and valuable properties. We should revise what they mean — however, we should also emphasise what they do *not* mean.

Consistency: In very general terms, an estimator is *consistent* if, as the number of observations becomes very large, the probability of the estimator missing the true parameter value goes to zero. Suppose, for example, that we are trying to estimate some true parameter β, and that we are using a maximum likelihood estimator $\hat{\beta}_{ML}$, with N observations in our sample. Then consistency means that, for *any* $\varepsilon > 0$,

$$\lim_{N\to\infty} \Pr(|\hat{\beta}_{ML} - \beta| > \varepsilon) = 0. \tag{1.24}$$

We can describe this by saying "$\hat{\beta}_{ML}$ converges in probability to the true value β", and we can write

$$\operatorname*{plim}_{N\to\infty} \hat{\beta}_{ML} = \beta. \tag{1.25}$$

Consistency, of course, is a Good Thing — it is hard to see much use for an estimator that is not consistent. However, consistency is also quite a *mild* claim. For example, suppose that *you* have some 'good estimator', $\hat{\beta}_{GOOD}$, which is consistent. Then I can propose the following estimator that is *also* consistent:

$$\hat{\beta}_{BAD} = \hat{\beta}_{GOOD} + \frac{1000000}{N}. \tag{1.26}$$

This, of course, is not so much a 'bad' estimator as a *really stupid* one — for a sample of $N = 1000$, for example, my estimator will be $\hat{\beta}_{BAD} = \hat{\beta}_{GOOD} + 1000$. But *both* of our estimators are clearly still *consistent*, since $\text{plim}_{N\to\infty} \hat{\beta}_{BAD} = \text{plim}_{N\to\infty} \hat{\beta}_{GOOD} = \beta$.

Efficiency: Informally, we can say that maximum likelihood makes the *best possible use* of the data available, to give us the *smallest possible standard errors on our estimates*. More formally, we can say that maximum likelihood estimators *achieve the Cramér-Rao Lower Bound*. We don't need to worry for now about what this means — however, the important point is that we can generally say only that maximum likelihood estimators achieve this *asymptotically* — that is, again, we can only draw general conclusions for the case where $N \to \infty$.

So what? Consistency and efficiency are valuable, but *they are properties that are only relevant to 'large' samples (really, to 'infinite' samples)*. There are at least two important implications for this.

(i) As Gould et al (2006, p.8) put it, "**Everyone knows this and knows never to fit a maximum likelihood model with only a handful of observations.**" Unfortunately, nobody can tell us what a 'handful' of observations means, but there is an important general principle here — you should beware of fitting maximum likelihood models (even simple models, like probit) on relatively small samples (for example, $N \leq 100$ — though, again, there is no clear guidance on this).

(ii) Because nobody can tell us how large a sample is 'large enough' for these purposes, there is a *strong case* for simulating a model to see how it performs in the sample size we are using. We will talk more about this when we come to discuss structural models later.

1.8 Hypothesis testing under maximum likelihood

1.8.1 Two main types of test

There are two main ways of testing a hypothesis after doing a maximum likelihood estimation. It is important to understand the different approaches taken by these tests — because the test that is easier and more commonly used is also generally considered to be the inferior of the two. Let's understand why.

For simplicity, suppose that we have a model involving only one parameter to be estimated, β, and that we are estimating using the log-likelihood function, $\ell(\beta)$. Suppose that we are testing the hypothesis:

$$H_0 : \beta = \beta_0. \tag{1.27}$$

One method of testing this hypothesis is to use a *likelihood ratio* test. This requires estimating our model *twice* — once when we allow β to take on any value (the 'unrestricted model') and once where we force $\beta = \beta_0$ (the 'restricted model'). If the hypothesis is *true*, this restriction should not matter much — *it should not make much difference to the log-likelihood*. On the other hand, if the hypothesis is *false*, we should expect that imposing the restriction $\beta = \beta_0$ should matter a *lot*. The likelihood ratio test simply formalises this intuition; formally, we have:

$$2\left[\ell(\hat{\beta}_{ML}) - \ell(\beta_0)\right] \sim \chi^2(1) \tag{1.28}$$

This implies a simple procedure:

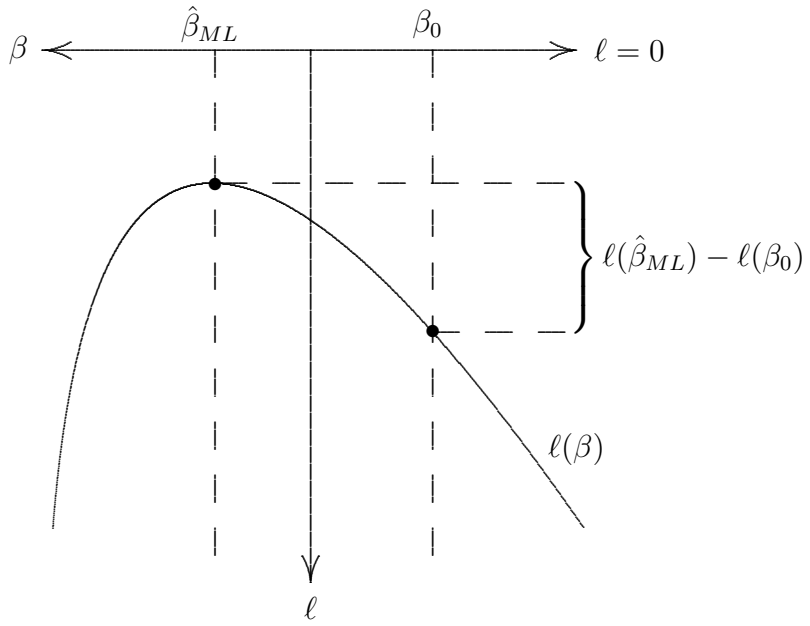
(i) Estimate the model without restrictions: the resulting log-likelihood is $\ell(\hat{\beta}_{ML})$;

(ii) Estimate the model restricting $\beta = \beta_0$ (we can do this using `ml`'s '`constraints`' option): the resulting log-likelihood is $\ell(\beta_0)$;

(iii) Calculate the difference, double it, and compare it to a χ^2 distribution with one degree of freedom.

Figure 1.4 illustrates the key idea.

Figure 1.4: **The Wald test and the Likelihood Ratio test**



Of course, it is possible that we might have more than one parameter and that we might want to test *multiple* restrictions. In that case, Stata will allow us to impose multiple constraints. The only catch, then, is to compare to a χ^2 distribution with multiple degrees of freedom; if we are imposing k independent restrictions, we compare to the $\chi^2(k)$ distribution.

The likelihood ratio test is clearly a very *simple* one — however, it is not necessarily *easy*. In particular, we need to run every estimation at least twice (and possibly many more times — if, for example, we want to *separately* test the significance of multiple coefficients). This is where the Wald test is useful. The Wald test requires estimating the model only *once* — to obtain $\hat{\beta}_{ML}$. The test then uses the *curvature* of the log-likelihood function at $\hat{\beta}_{ML}$ to obtain an estimate of the variance of $\hat{\beta}_{ML}$, *taken at* $\hat{\beta}_{ML}$. Let's call this $\mathrm{Var}(\hat{\beta}_{ML})$. The Wald test then involves calculating

$$W = \frac{\left(\hat{\beta}_{ML} - \beta_0\right)^2}{\mathrm{Var}\left(\hat{\beta}_{ML}\right)}, \tag{1.29}$$

which is compared to a $\chi^2(1)$ distribution. (As with the likelihood ratio test, the Wald test can also be used to test k restrictions jointly.)

We can implement a Wald test using Stata's `test` command. However, importantly, `ml` *performs Wald tests for us anyway, when it reports the standard errors,* z *values and* p*-values for each parameter estimate*. It is very important, then, to remember that, *the Wald test is only* ***approximating***

the job it is really supposed to be doing.[8] Rather than actually ***calculating*** the difference between $\ell(\hat{\beta}_{ML})$ and $\ell(\beta_0)$, the Wald test is using the *curvature of* $\ell(\beta)$ *at* $\beta = \hat{\beta}_{ML}$ *to* ***approximate*** *it*.

In short, it is *fine* to rely on `ml`'s reported p-values for getting a general sense of the significance of paricular estimates. However, if we are particularly interested in testing a hypothesis about a coefficient or group of coefficients, *we should use the likelihood ratio test*.[9]

1.8.2 Hypothesis testing using `ml` in Stata

We should implement the Wald and Likelihood Ratio tests in Stata using the South African wage curve data. First, let's try the Wald test.

```
. do ols_ml
```

⋮

```
                                                Number of obs   =       6980
                                                Wald chi2(1)    =    2486.93
Log likelihood = -9542.9597                     Prob > chi2     =     0.0000

------------------------------------------------------------------------------
     logwphy |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
eq1          |
       edyrs |   .1380054   .0027674    49.87   0.000     .1325814    .1434293
       _cons |   .4611689   .0245925    18.75   0.000     .4129684    .5093694
-------------+----------------------------------------------------------------
sigma        |
       _cons |   .9495667   .0080368   118.15   0.000     .9338148    .9653185
------------------------------------------------------------------------------
```

Let's practice performing a Wald test in Stata using the `test` command. When we have output from `ml`, the easiest way to refer to coefficients is by '`[{equation name}]{coefficient name}`' — for example, we can refer to the estimate on `edyrs` as '`[eq1]edyrs`'.

Let's test whether the coefficient on `edyrs` is significantly different from zero...

[8] If you are interested in this, you should read Gould et al (2006, pp.9-10). In short, the problem is this: the concept of hypothesis testing really requires us to use the distribution of $\hat{\beta}_{ML}$ when $\beta = \beta_0$; by using $\text{Var}(\hat{\beta}_{ML})$, the Wald test is approximating the estimated variance (except where the log-likelihood is quadratic, as it is — for example — in the case of linear regression). Gould et al show how we can think of this as the Wald test taking a *second-order Taylor approximation* around $\hat{\beta}_{ML}$ to approximate the likelihood ratio; I think this is a really nice way of thinking about it, but it's not something that we need to worry about in these lectures.

[9] Gould et al (2006, p.10) caution against using the likelihood ratio test if the number of independent restrictions is greater than about 100... hopefully this won't concern us!

```
. test [eq1]edyrs = 0

 ( 1)  [eq1]edyrs = 0

           chi2(  1) = 2486.93
         Prob > chi2 =    0.0000
```

We find a tiny p-value (less than 0.00005), meaning that the coefficient is statistically significant at the 99.99% confidence level. Of course, we could also have got this result directly from the estimation itself (you should calculate $\sqrt{2486.93}$; do you notice anything?).

However, we can also perform a Wald test to test two parameters *jointly*. Suppose that, for some reason, we have a theory that predicts that $\beta_0 = 0.4$ and $\beta_1 = 0.14$. Clearly, the coefficients seem very *close* to this, but we should test whether they are statistically different...

```
test ([eq1]_cons = 0.4) ([eq1]edyrs = .14)

 ( 1)  [eq1]_cons = .4
 ( 2)  [eq1]edyrs = .14

           chi2(  2) =   16.51
         Prob > chi2 =    0.0003
```

The test *rejects* a joint null hypothesis $H_0 : \beta_0 = 0.4; \beta_1 = 0.14$; we obtain $p = 0.0003$, meaning that we can reject at the 99.9% confidence level.

Before we proceed to the Likelihood Ratio test, notice that we can use the `test` command to perform the same kind of test after `reg`:

```
. reg logwphy edyrs

      Source |       SS       df       MS              Number of obs =    6980
-------------+------------------------------           F(  1,  6978) = 2486.11
       Model |  2242.29713     1  2242.29713           Prob > F      =  0.0000
    Residual |  6293.66923  6978  .901930242           R-squared     =  0.2627
-------------+------------------------------           Adj R-squared =  0.2626
       Total |  8535.96636  6979  1.22309304           Root MSE      =   .9497

------------------------------------------------------------------------------
     logwphy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       edyrs |    .138002   .0027677    49.86   0.000     .1325764    .1434276
       _cons |   .4611801    .024596    18.75   0.000     .4129645    .5093957
------------------------------------------------------------------------------

. test (_cons = .4) (edyrs = .14)

 ( 1)  _cons = .4
 ( 2)  edyrs = .14

       F(  2,  6978) =    8.25
            Prob > F =    0.0003
```

The likelihood ratio test is a bit more complicated to implement, because we need to estimate *twice*: once under the unrestricted case and once under the restricted case. Suppose we are testing the initial hypothesis, $H_0 : \beta_1 = 0$. We need to go back and constrain the `ml` command to ensure that $\hat{\beta}_1 = 0$. We do this as follows.

First, at the start of the file `OLS_ML.do`, we define the constraint, using a number:

```
constraint 1 [eq1]edyrs = 0
```

Second, we add to the end of the line defining the model, to tell `ml` to use this constraint:

```
ml model ...constraints(1)
```

Make these changes and save the file. Now we can use `ml` to estimate *under the restriction*:

```
. do ols_ml

⋮
⋮

                                                  Number of obs   =       6980
                                                  Wald chi2(0)    =          .
Log likelihood = -10606.517                       Prob > chi2     =          .

 ( 1)  [eq1]edyrs = 0
------------------------------------------------------------------------------
     logwphy |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
eq1          |
       edyrs |  (dropped)
       _cons |   1.548722   .0132364   117.00   0.000     1.522779    1.574665
-------------+----------------------------------------------------------------
sigma        |
       _cons |   1.105857   .0093596   118.15   0.000     1.087512    1.124201
------------------------------------------------------------------------------
```

We now have *two* values for the log-likelihood: the unrestricted value was $\ell(\hat{\beta}_{ML}) = -9542.9597$ and the restricted value is now $\ell(\beta_0) = -10606.517$. Equation 1.28 tells us that we should take the difference, double it, and compare to a $\chi^2(1)$ distribution. **Let's do this in Stata…**

```
. display 10606.517 -9542.9597
1063.5573

. display 2*(10606.517 -9542.9597)
2127.1146

. display chi2tail(1, 2127.1146)
0
```

This last value is the p-value; as with the Wald test, we obtained an extremely small p-value (approximately zero).

Exercise: Go back to the joint null hypothesis considered earlier, $H_0 : \beta_0 = 0.4; \beta_1 = 0.14$. Implement this with `ml` using *two* constraints and obtain the p-value on the Likelihood Ratio test. How does this compare to the p-value from the Wald test? (I *think* you should end up with something very similar to this...)

```
. display chi2tail(2, 2*(9551.2006 -9542.9597))
.00026365
```

1.9 Conclusions

In this lecture, we have considered concept of maximum likelihood estimation, and we have estimated a simple maximum-likelihood model using the `ml` command in Stata. Hopefully, there are at least two reasons that this will be valuable. First, maximum likelihood is one powerful tool for estimating structural models; we will look at this in more detail in Week 1 of next term.[10] Second, even if you are not interested in ever maximising a customised log-likelihood model, the underlying concepts of maximum likelihood will be important for understanding several of the 'standard' estimators that you will come across (including the probit model, the logit model, the tobit model and models of multinomial choice). Indeed, if you *really* want to understand such models, it may be a very useful exercise to code your own version of each model using the `ml` command, in order to check that it produces the same output as the coded estimators in Stata. For the probit and logit models, I have provided examples on Weblearn which may be a useful starting point: see the files `probit_ml.do` and `logit_ml.do`.

1.10 A warning: maximum likelihood and instrumental variables

Let me flag briefly one important practical issue that you may encounter when writing your extended essay.[11] Suppose that you have a binary outcome variable, y, a continuous endogenous explanatory variable, x, and an instrumental variable, z. Intuitively, you might:

(i) Run a regression of $x_i = \gamma_0 + \gamma_1 z_i + \mu$ and form the 'predicted values' $\hat{x}_i = \hat{\gamma}_0 + \hat{\gamma}_1 z_i$; then

(ii) Run a probit estimation using $\hat{x}_i$ instead of x_i.

This may seem like a *perfectly normal and intuitive* method to use, particularly once you have learned the principles of 'two-stage least squares' estimation. However, it is ***wrong***! Indeed, in general (though not in every case) you should *not* be running maximum likelihood estimations with predicted values, no matter how intuitive it may seem. The details — and alternative approaches — are beyond the scope of my teaching. However, I mention the problem as a warning.[12]

[10] You should note, however, that there are *many* other tools for estimating structural models. Structural models do not require maximum likelihood estimation, and maximum likelihood estimators do not require a structural model!

[11] Of course, this section will probably make more sense *after* you have studied instrumental variable methods later in the term.

[12] You could, for example, look at Chesher (2010, *Econometrica*) for a technical discussion of the problem. The simplest alternative method in this case would be to run the second stage as a Linear Probability Model; in that case, you could use $\hat{x}_i$ instead of x_i. You could also use 'ivprobit' in Stata, which estimates both first and second stages jointly (*i.e.* rather than using predicted values). But I will leave the details for you to discuss with your essay supervisor if the issue arises.

2 Lecture: Microeconometric Structural Models I

Plato's *Allegory of the Cave* (c.380 BCE)

SOCRATES And now, I said, let me show in a figure how far our nature is enlightened or unenlightened: – Behold! human beings living in a underground cave, which has a mouth open towards the light and reaching all along the cave; here they have been from their childhood, and have their legs and necks chained so that they cannot move, and can only see before them, being prevented by the chains from turning round their heads. Above and behind them a fire is blazing at a distance, and between the fire and the prisoners there is a raised way; and you will see, if you look, a low wall built along the way, like the screen which marionette players have in front of them, over which they show the puppets.

GLAUCON I see.

SOCRATES And do you see, I said, men passing along the wall carrying all sorts of vessels, and statues and figures of animals made of wood and stone and various materials, which appear over the wall? Some of them are talking, others silent.

GLAUCON You have shown me a strange image, and they are strange prisoners.

SOCRATES Like ourselves, I replied; and they see only their own shadows, or the shadows of one another, which the fire throws on the opposite wall of the cave?

GLAUCON True, he said; how could they see anything but the shadows if they were never allowed to move their heads?

SOCRATES And of the objects which are being carried in like manner they would only see the shadows?

GLAUCON Yes, he said.

SOCRATES And if they were able to converse with one another, would they not suppose that they were naming what was actually before them?

GLAUCON Very true.

SOCRATES And suppose further that the prison had an echo which came from the other side, would they not be sure to fancy when one of the passers-by spoke that the voice which they heard came from the passing shadow?

GLAUCON No question, he replied.

SOCRATES To them, I said, the truth would be literally nothing but the shadows of the images.

GLAUCON That is certain.

SOCRATES And now look again, and see what will naturally follow if the prisoners are released and disabused of their error. At first, when any of them is liberated and compelled suddenly to stand up and turn his neck round and walk and look towards the light, he will suffer sharp pains; the glare will distress him, and he will be unable to see the realities of which in his former state he had seen the shadows; and then conceive some one saying to him, that what he saw before was an illusion, but that now, when he is approaching nearer to being and his eye is turned towards more real existence, he has a clearer vision, — what will be his reply? And you may further imagine that his instructor is pointing to the objects as they pass and requiring him to name them, — will he not be perplexed? Will he not fancy that the shadows which he formerly saw are truer than the objects which are now shown to him?

Koopmans and Reiersøl (1950, p.165)

In many fields the objective of the investigator's inquisitiveness is not just a "population" in the sense of a distribution of observable variables, but a physical structure projected behind this distribution, by which the latter is thought to be generated. The word "physical" is used merely to convey that the structure concept is basd on the investigator's ideas as to the "explanation" or "formation" of the phenomena studied, briefly, on his [or her] theory of these phenomena, whether they are classified as physical in the literal sense, biological psychological, sociological, economic or otherwise.

2.1 An introduction to structural models in microeconometrics

References:

- ⋆ ATTANASIO, O., MEGHIR, C., AND SANTIAGO, A. (2012): "Education Choices in Mexico: Using a Structural Model and a Randomised Experiment to Evaluate PROGRESA," *The Review of Economic Studies*, 79(1), 37-66.
- ⋆ BELZIL, C., AND HANSEN, J. (2002): "Unobserved Ability and the Return to Schooling," *Econometrica*, 70(5), 2075–2091.
- BROWNING, M. (2009): "Two Examples of Structural Modelling," *Lecure notes for the M.Phil in Economics*; see `http://www.nuffield.ox.ac.uk/Teaching/Economics/Browning/Structural/`.
- JENSEN, R. (2010): "The (Perceived) Returns to Education and the Demand for Schooling," *The Quarterly Journal of Economics*, 125(2), 515.
- KEANE, M. (2010): "Structural vs. Atheoretic Approaches to Econometrics," *Journal of Econometrics*, 156, 3–20.
- KOOPMANS, T.C. AND REIERSØL, O. (1950): "The Identification of Structural Characteristics," *The Annals of Mathematical Statistics*, 21(2), 165–181.
- MANSKI, C. (1993): "Adolescent Econometricians: How do Youth Infer the Returns to Schooling?", in *Studies of Supply and Demand in Higher Education*, by C.Clotfelter and M.Rothschild.
- SÖDERBOM, M., TEAL, F., WAMBUGU, A., AND KAHYARARA, G. (2006): "The Dynamics of Returns to Education in Kenyan and Tanzanian Manufacturing," *Oxford Bulletin of Economics and Statistics*, 68(3), 261–288.

Data is always an imperfect shadow of the real world. As development economists, however, it is the real world that primarily concerns us. *Structural modelling* is a methodology that allows a direct dialogue between economic theory and econometric estimation; it allow us to think in rigorous theoretical terms about the real incentive structures that agents face and the implications of those structures for empirical observation — the 'shadows cast' by agents' incentives upon the variables that we observe.

There is no single, simple definition of what constitutes a structural model. However, for our purposes, we will consider a structural model to be *any economic model that specifies a theory about how agents optimise and then uses the implications of that theory* ***literally*** *as a basis for empirical estimation*. In doing so, structural models often try to estimate 'deep parameters' governing human behaviour (for example, the return to capital, or preferences over risk) rather than merely identifying particular causal relationships in a particular context (for example, the 'average treatment

effect' of a given policy on some outcome).

Figure 2.1 gives a very simple illustration — perhaps offensively simple! — of what structural models do: they add a 'front end' to the estimation procedure. In doing so, they provide a formal theoretical context for interpreting the resulting estimates that is lacking in more 'traditional' microeconometric methodologies.

Figure 2.1: **A simplistic distinction between 'traditional' and 'structural' methodologies**

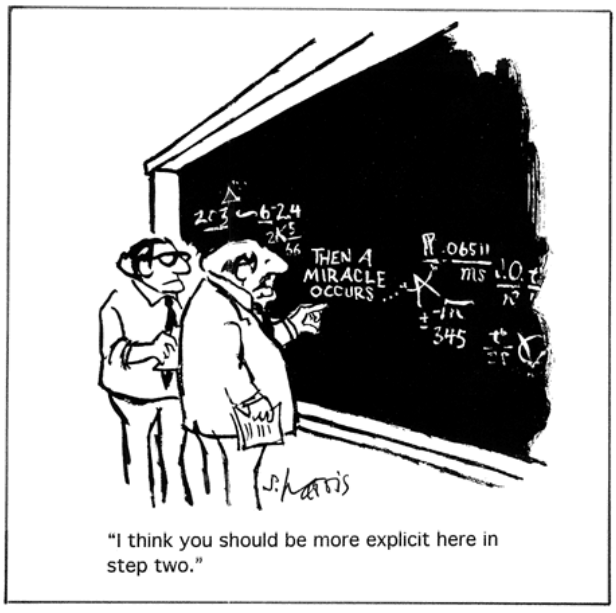
'Traditional' microeconometric methodologies...

EQUATION(S) TO BE ESTIMATED ⇒ ESTIMATION

The structural microeconometric methodology...

THEORETICAL MODEL ⇒ EQUATION(S) TO BE ESTIMATED ⇒ ESTIMATION

Or, in slightly less formal terms...



Because theoretical models often do not produce 'neat' equations to be estimated (for example, they often do not imply relationships as neat as $y_i = \boldsymbol{\beta}\boldsymbol{x}_i + \varepsilon_i$), structural modelling often requires more complicated estimation techniques than 'traditional' microeconometric methods — as we will see in this lecture. However, this is *not* a necessary part of structural modelling — it is entirely possible to have a structural model that demands only a garden-variety estimator (for example, OLS). We will see an example of this in the next lecture.

What do we mean, then, by having a 'theoretical model'? In short, we mean exactly what we would mean when discussing this in microeconomic theory. That is, we will generally specify a model that has:

- Clear assumptions about who the agents ('players', perhaps) are, and what choices they make;
- Clear assumptions about what the agents maximise, and the constraints that they face in doing so;
- Clear assumptions about how agents do or do not interact.

As you know, *no* model ever fits data perfectly (if we run OLS and get $R^2 = 1$, for example, we know we have made a mistake!). The same must be true in structural modelling — we can *never* come close to explaining any data perfectly. In addition to the elements above, therefore, we also need to think about *heterogeneity across agents*. That is, we will generally also add:

- Clear assumptions about how agents differ between each other, and/or
- Clear assumptions about how our observed data differs from the 'true underlying model' that we posit.

That is — as in traditional microeconometric methods — *we need to make clear assumptions about the role played by unobservables*.

Structural modelling has many advantages over traditional microeconometric methods — but it also has many disadvantages. My goal here is not to proselytise; rather, I would like to discuss the key ideas of structural modelling, and work through two examples of structural models in the context of developing economies. Hopefully, this will be useful both for understanding structural approaches in the literature and — if you are feeling adventurous — it may encourage you to consider a structural approach for your extended essay or for subsequent research. Either way, I propose that we leave a discussion about the relative merits and demerits of the methodology until later!

Instead, I would like to build a simple structural model to illustrate the basic ideas of the structural methodology. We will discuss and build the model in this lecture and estimate the model using Stata in the following class.

2.2 Example: *Belzil and Hansen Go To Africa*

2.2.1 The question

How does investment in education affect earnings in developing economies? This is the basic question that concerns us in this lecture and in the next class. It is a question that we are all quite familiar with; for example, it is the same question that we considered last term when we used the South African wage curve data.

We did not worry much about endogeneity in the earlier lecture or class on maximum likelihood. You should be familiar with the standard endogeneity issue in the education-earnings literature: if more *able* individuals choose to undertake *more education*, an OLS estimation of earnings on education will produce a biased estimate of the true causal value of education. There are several ways that we could try to address this endogeneity concern:

- We could run OLS with a number of *controls*, arguing that those controls together adequately capture individuals' 'unobserved ability';
- We could find an *instrument* for education (for example, respondents' distance to the nearest school when they were young);
- We could find a *natural experiment* changing educational attainment (for example, a change in compulsory school attendance laws);
- We could run a *randomised experiment* (for example, by randomly choosing some households to receive free education for their children).

However, it may be that we cannot use *any* of these methods. First, we may be concerned that even an extensive set of controls do not adequately capture unobserved ability. Second, we may struggle to find a valid instrument or natural experiment (for example, we may be concerned that a family's decision to live close to a school itself reflects some important unobservable characteristics of that family).[13] Third, we may be concerned that randomisation is impractical (for example, because of the long lag time between allocation of education vouchers and observing the recipients in the labour force) and not likely to be representative of any population of interest (for example, because any randomisation is likely to affect a relatively small group of participants who are not necessarily representative of any particular region or nation). Finally — and perhaps most importantly — we may feel that, *even if we could randomise*, we may still want to be able to interpret our estimates in the context of a clear theoretical model.[14]

[13] Indeed, we may struggle to justify the validity of an instrument without an accompanying theoretical model. If you are interested, you should read Keane's (2010) short article (see the reference list for this lecture), which discusses this in some detail. Professor Browning makes a similar point in Section 5 of his lecture notes on 'identification', which are a required reading for our next lecture.

[14] The recent paper by Attanasio, Meghir and Santiago (2012) — which we will consider shortly through a set of exercises — is an interesting illustration of this last point.

In this lecture and in the next class, we will consider an alternative approach. We will build a simple *structural model* to capture directly the relationship between unobserved ability and choice of schooling. This model is a simple version of the model in Belzil and Hansen (2002). Belzil and Hansen estimated their model using data from the National Longitudinal Survey of Youth in the US; we will estimate using a random sample of 2000 income-earners from the Tanzania's 2006 Integrated Labour Force Survey (a nationally-representative survey of Tanzanian households).

In doing so, we will be interested both in the *magnitude* of the return to education (that is, whether education has a substantial effect upon earnings) and upon the *shape* of that return (that is, whether the marginal value of education *increases* or *decreases* with the level of education). Both questions have attracted interest in the literature: for example, Söderbom *et al* (2006) find (using OLS and IV estimation) that education has a *significant and convex* effect on earnings in Tanzania.

2.2.2 Specifying the model

> Essentially, all models are wrong, but some are useful.
>
> *George Box and Norman Draper (1987)*

Purpose of the model: In this lecture, we will build a microeconomic model to consider the *supply* of educated labour. That is, we will assume that the *demand* for educated labour is given exogenously, and we will worry about how the *supply* of that educated labour is determined. Specifically, we will build a model in which students (and their families) *balance the costs and benefits of education*, and do so on the basis of unobservable 'abilities' for schooling and for work. In short, *the purpose of this model is to estimate the returns to education while allowing for education to correlate with unobserved ability*.

Outline of the model: We will:

(i) Make some strong assumptions about the utility that a student gets from schooling and from work;

(ii) Make some strong assumptions about how the student weighs utility from different time periods;

(iii) On the basis of these assumptions, solve for the student's decision rule as to how many years of education to obtain;

(iv) Make a strong assumption about the distribution of unobservables; and

(v) On the basis of the decision rule and the distributional assumption, write an expression for the log-likelihood for the model.

Assumption 2.1 ***Instantaneous utility****: Every student has preferences of an identical form. A given student i receives utility v_i^s from a year of schooling, and this is the same for that student for* all *years of schooling.*[15] *Following Belzil and Hansen, we can refer to v_i^s as 'school ability'. Upon entering the workforce, the student i receives a wage determined by his or her schooling achievement (S_i) and the student's idiosyncratic 'market ability', v_i^w. The student gains utility from the wage according to the log transformation. Specifically, the student receives the following in-period utilities for schooling and work:*

$$U_{it}^S(\cdot) = v_i^s; \tag{2.1}$$

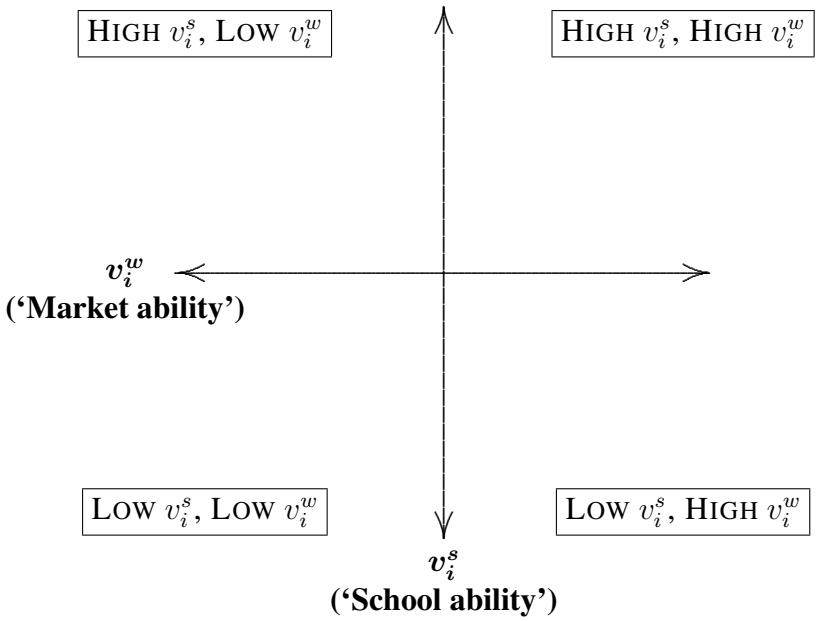
$$U_{it}^W(\cdot) = \ln w_{it} = \psi(S_i) + v_i^w. \tag{2.2}$$

Thus, right from the outset, we have a sense of the key aspects of the model:

(i) Students *choose* their investment in education on the basis of their respective abilities for schooling and work;

(ii) We immediately see the main *object of interest* for our estimation: the function $\psi(S)$, which determines the causal *returns to education*.

Figure 2.2: **'School ability' (v_i^s) and 'market ability' (v_i^w)**



[15] That is, you should never listen to bold claims around the Manor Road café along the lines of "wait until you get onto the D.Phil — then you'll *really* come to appreciate the value of daytime television"; the Belzil and Hansen model strictly rules out this kind of attitude.

Figure 2.2 shows the key idea: each student is endowed with some combination of 'market ability' and 'school ability', so that the combination of the two will — in due course — determine how much education a student chooses to obtain *and* how much that student consequently earns.

This, of course, is a very strong assumption indeed. In particular, note that v_i^s and v_i^w cannot change over time — so that, for example, we are ruling out the possibility that age or experience cause higher earnings. Similarly, by using the log transformation, we are *completely* abstracting away from the question of *whether* someone earns income (that is, the question of employment and unemployment) — we are trying instead only to explain the level of income that income-earners receive. For these reasons, the assumption is probably not really satisfactory — but it is simple, so we will stick with it for the illustrative purposes of this example.

Assumption 2.2 *Students may only choose zero, four, seven or 11 years of education.*

There is nothing in the subsequent reasoning that rules out students choosing other investments (for example, they may choose $S = 5$). However, this assumption will keep the model simpler and more manageable. In the data, we observe 'clumping' around these choices, so we will build the model as though these are the *only* choices available. That is, I create dummy variables for whether workers have no education, four years of education, seven years of education or 11 years; if someone has five years of education, for example, I code this as $S = 4$.

In doing this, we can define $\psi(S)$ as providing the return to education *relative to zero years of schooling* — that is, we can define $\psi(0) = 0$ and restrict attention to estimating $\psi(4)$, $\psi(7)$ and $\psi(11)$.

Assumption 2.3 *Every student has time-separable preferences using exponential discounting with discount factor β. That is, if the instantaneous utility from period t is u_t, the student's utility from time $t = 0$ to $t = T$ is:*

$$U = \sum_{t=0}^{T} \beta^t u_t. \tag{2.3}$$

This assumption, too, is quite strong — but very standard. We will accept it and move on.

Assumption 2.4 *Once a student leaves school, the student cannot return.*

My sense is that this is *not* a very strong assumption, given the other assumptions we have made. After all, if it pays to get S years of education at *any* point in your life, it will surely pay to get that education *as soon as possible*, to maximise lifetime utility.[16] However, to keep things simple, we will state this as an assumption and not bother proving it!

Assumption 2.5 *Conditional on knowing v_i^s and v_i^w, everything is deterministic. That is, no student faces any uncertainty — including, for example, about (i) whether or not the student will get a job, or (ii) how much the student will subsequently earn.*

[16] That is, in the context of this simple model with no random shocks and a very simple income process.

Of course, this assumption borders on the ridiculous — it completely abstracts away from important aspects of both the job market and life generally (for example, we will discuss the literature on risk and mutual insurance later in this term). However, it makes the model manageable, so we will stick with it.

2.2.3 Solving the model

On the basis of these assumptions, we can write the *value functions* for the lifetime utility of a student choosing $S = 0$, $S = 4$, $S = 7$ and $S = 11$ respectively. They are:

$$V_0(0) = \sum_{t=0}^{T} \beta^t \cdot v_i^w \tag{2.4}$$

$$V_0(4) = \sum_{t=0}^{3} \beta^t \cdot v_i^s + \sum_{t=4}^{T} \beta^t \cdot (\psi(4) + v_i^w) \tag{2.5}$$

$$V_0(7) = \sum_{t=0}^{6} \beta^t \cdot v_i^s + \sum_{t=7}^{T} \beta^t \cdot (\psi(7) + v_i^w) \tag{2.6}$$

$$V_0(11) = \sum_{t=0}^{10} \beta^t \cdot v_i^s + \sum_{t=11}^{T} \beta^t \cdot (\psi(11) + v_i^w)\,. \tag{2.7}$$

From these functions, we can solve the model: we simply find the conditions, in terms of v_i^s and v_i^w, for the student to choose $S = 0$, $S = 4$, $S = 7$ or $S = 11$. First, consider the choice between $S = 0$ and $S = 4$; clearly, the student will prefer $S = 4$ to $S = 0$ if:

$$V_0(4) \geq V_0(0) \tag{2.8}$$

$$\sum_{t=0}^{3} \beta^t \cdot v_i^s + \sum_{t=4}^{T} \beta^t \cdot (\psi(4) + v_i^w) \geq \sum_{t=0}^{T} \beta^t \cdot v_i^w \tag{2.9}$$

$$\sum_{t=0}^{3} \beta^t \cdot v_i^s + \sum_{t=4}^{T} \beta^t \cdot \psi(4) \geq \sum_{t=0}^{3} \beta^t \cdot v_i^w \tag{2.10}$$

$$\sum_{t=0}^{3} \beta^t \cdot (v_i^s - v_i^w) \geq -\sum_{t=4}^{T} \beta^t \cdot \psi(4) \tag{2.11}$$

$$(v_i^s - v_i^w) \cdot \left(\frac{1-\beta^4}{1-\beta}\right) \geq -\left(\frac{\beta^4 - \beta^{T+1}}{1-\beta}\right) \cdot \psi(4) \tag{2.12}$$

$$v_i^s \geq v_i^w - \left(\frac{\beta^4 - \beta^{T+1}}{1-\beta^4}\right) \cdot \psi(4) \tag{2.13}$$

Note here that, to get to the last two lines, we need to know that, for any $|\beta| < 1$,

$$\sum_{t=m}^{n} \beta^t = \frac{\beta^m - \beta^{n+1}}{1-\beta}. \tag{2.14}$$

We can use the same logic, then, to think about comparing $S = 7$ to $S = 4$. The student prefers $S = 7$ to $S = 4$ if:

$$V_0(7) \geq V_0(4) \tag{2.15}$$

$$\sum_{t=0}^{6} \beta^t \cdot v_i^s + \sum_{t=7}^{T} \beta^t \cdot (\psi(7) + v_i^w) \geq \sum_{t=0}^{3} \beta^t \cdot v_i^s + \sum_{t=4}^{T} \beta^t \cdot (\psi(4) + v_i^w) \tag{2.16}$$

$$\sum_{t=4}^{6} \beta^t \cdot v_i^s + \sum_{t=7}^{T} \beta^t \cdot (\psi(7) - \psi(4)) \geq \sum_{t=4}^{6} \beta^t \cdot (\psi(4) + v_i^w) \tag{2.17}$$

$$\sum_{t=4}^{6} \beta^t \cdot (v_i^s - v_i^w - \psi(4)) \geq -\sum_{t=7}^{T} \beta^t \cdot (\psi(7) - \psi(4)) \tag{2.18}$$

$$\therefore \sum_{t=0}^{2} \beta^t \cdot (v_i^s - v_i^w - \psi(4)) \geq -\sum_{t=3}^{T-4} \beta^t \cdot (\psi(7) - \psi(4)) \tag{2.19}$$

$$\left(\frac{1-\beta^3}{1-\beta}\right) \cdot (v_i^s - v_i^w - \psi(4)) \geq -\left(\frac{\beta^3 - \beta^{T-3}}{1-\beta}\right) \cdot (\psi(7) - \psi(4)) \tag{2.20}$$

$$\therefore v_i^s \geq v_i^w + \psi(4) - \left(\frac{\beta^3 - \beta^{T-3}}{1-\beta^3}\right) \cdot (\psi(7) - \psi(4)). \tag{2.21}$$

Finally, consider the decision to choose $S = 11$ rather than $S = 7$:

$$V_0(11) \geq V_0(7) \tag{2.22}$$

$$\sum_{t=0}^{10} \beta^t \cdot v_i^s + \sum_{t=11}^{T} \beta^t \cdot (\psi(11) + v_i^w) \geq \sum_{t=0}^{6} \beta^t \cdot v_i^s + \sum_{t=7}^{T} \beta^t \cdot (\psi(7) + v_i^w) \tag{2.23}$$

$$\sum_{t=7}^{10} \beta^t \cdot v_i^s + \sum_{t=11}^{T} \beta^t \cdot (\psi(11) - \psi(7)) \geq \sum_{t=7}^{10} \beta^t \cdot (\psi(7) + v_i^w) \tag{2.24}$$

$$\sum_{t=7}^{10} \beta^t \cdot (v_i^s - v_i^w - \psi(7)) \geq -\sum_{t=11}^{T} \beta^t \cdot (\psi(11) - \psi(7)) \tag{2.25}$$

$$\therefore \sum_{t=0}^{3} \beta^t \cdot (v_i^s - v_i^w - \psi(7)) \geq -\sum_{t=4}^{T-7} \beta^t \cdot (\psi(11) - \psi(7)) \tag{2.26}$$

$$\left(\frac{1-\beta^4}{1-\beta}\right) \cdot (v_i^s - v_i^w - \psi(7)) \geq -\left(\frac{\beta^4 - \beta^{T-6}}{1-\beta}\right) \cdot (\psi(11) - \psi(7)) \tag{2.27}$$

$$\therefore v_i^s \geq v_i^w + \psi(7) - \left(\frac{\beta^4 - \beta^{T-6}}{1-\beta^4}\right) \cdot (\psi(11) - \psi(7)) \tag{2.28}$$

To be very clear, we can rewrite these results in terms of different cut-offs for different choices of schooling:

$$S = \begin{cases} 11 & \text{if} \quad v_i^s - v_i^w \geq \gamma_{11} \\ 7 & \text{if} \quad v_i^s - v_i^w \in [\gamma_7, \gamma_{11}) \\ 4 & \text{if} \quad v_i^s - v_i^w \in [\gamma_4, \gamma_7) \\ 0 & \text{if} \quad v_i^s - v_i^w < \gamma_4, \end{cases} \tag{2.29}$$

where (2.30)

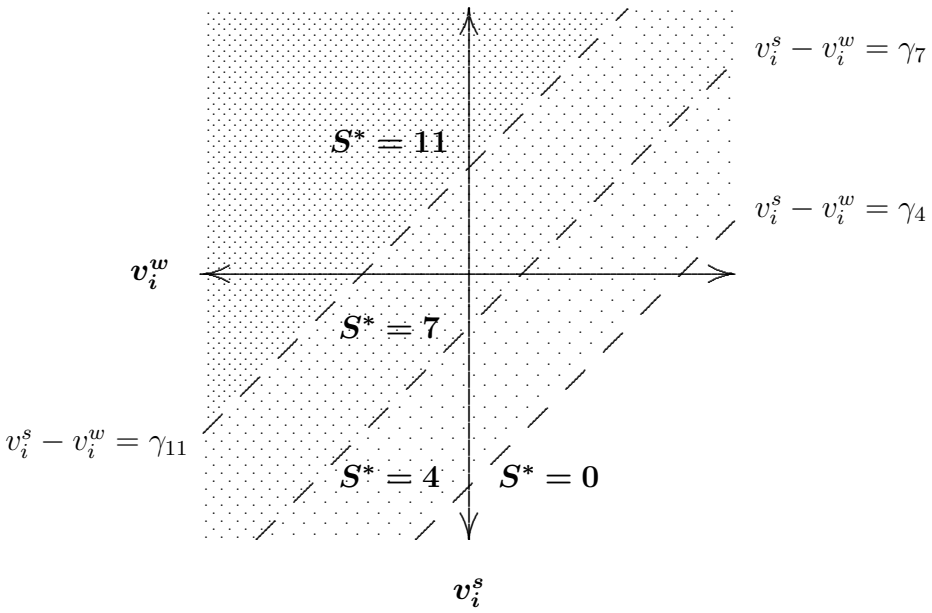
$$\gamma_{11} = \psi(7) - \left(\frac{\beta^4 - \beta^{T-6}}{1 - \beta^4} \right) \cdot (\psi(11) - \psi(7)) \tag{2.31}$$

$$\gamma_7 = \psi(4) - \left(\frac{\beta^3 - \beta^{T-3}}{1 - \beta^3} \right) \cdot (\psi(7) - \psi(4)) \tag{2.32}$$

and $$\gamma_4 = - \left(\frac{\beta^4 - \beta^{T+1}}{1 - \beta^4} \right) \cdot \psi(4). \tag{2.33}$$

Figure 2.3 shows these cutoffs — and the resulting choice of S — in terms of segments of the (v_i^w, v_i^s) space.

Figure 2.3: **Choosing between** $S = 0$**,** $S = 4$**,** $S = 7$ **and** $S = 11$



2.2.4 The distribution of unobservables

We have, then, solved the model — if we know (or posit) values of v_i^s and v_i^w for a student, we now know how much education the student will obtain and how much the student will earn. At least, we know this in terms of the *parameters of the model*; the next step is to use the data to find maximum likelihood estimates of those parameters.

But not so fast. As we discussed in the introductory class, we need to make *distributional assumptions* about the unobservables in the model. In this model, we have two unobservables: v_i^s and v_i^w. The simplest approach would be to assume that each has a normal distribution. This is *almost* good enough, but not quite — we need to make an assumption not merely about their separate *marginal* distributions, but also about their *joint* distribution. The simplest joint distribution for our purposes is the *bivariate normal distribution with zero covariance*. This distribution implies that (i) v_i^s and v_i^w *each* has a normal distribution, and (ii) v_i^s and v_i^w are independent of each other. Formally, we can write the following.

Assumption 2.6 *The joint distribution of v_i^s and v_i^w is bivariate normal with zero covariance:*

$$\begin{pmatrix} v_i^s \\ v_i^w \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_s^2 & 0 \\ 0 & \sigma_w^2 \end{pmatrix} \right). \tag{2.34}$$

This assumption, too, seems very strange — intuitively, we would probably imagine that people who have higher 'market ability' also have higher 'school ability'. Indeed, when we usually speak of 'unobserved ability' (for example, when discussing the standard endogeneity problem of regressing earnings on education) we are essentially implying that there is a *single type* of 'ability' — in effect, that $v_i^w = v_i^s$.

What, then, is the *object of interest*? Clearly, our primary interest is in estimating the returns to education — that is, the shape of the function $\psi(S)$. Since S can only take on the values $S = 0$, $S = 4$, $S = 7$ and $S = 11$, and since we assume that $\psi(0) = 0$, we are really just trying to estimate $\psi(4)$, $\psi(7)$ and $\psi(11)$. For clarity, let's refer to these as parameters ψ_4, ψ_7 and ψ_{11}. We have to do a bit more than just estimate these three parameters, however; our model also requires us to estimate (or assume) the discount parameter, β. Further, our distributional assumption provides two more parameters to estimate: σ_s^2 and σ_w^2. Therefore, the *object of interest* is a *collection* (vector) of parameters that we can refer to, for simplicity, as $\mathbf{\Theta}$:

$$\mathbf{\Theta} = \begin{pmatrix} \psi_4 \\ \psi_7 \\ \psi_{11} \\ \beta \\ \sigma_s^2 \\ \sigma_w^2 \end{pmatrix}. \tag{2.35}$$

Having made the distributional assumption, the trick now is to use this assumption to derive the log-likelihood function in terms of $\mathbf{\Theta}$. Having made quite strong simplifying assumptions throughout, this now transpires to be relatively straightforward. The fundamental point is this: our model

makes predictions *both* about how v_i^s and v_i^w determine S *and* about how v_i^w and S then determine income; we therefore want a function that explains not merely the likelihood of observing a particular wage, *or* the likelihood of observing a given number of years of schooling, but a function that describes *both together*.

For this we can use Bayes' Rule:

$$\Pr(A \text{ and } B) = \Pr(A|B) \times \Pr(B), \tag{2.36}$$

where we read $\Pr(A|B)$ as 'the probability of A *conditional on* B'. We can therefore say that:

the likelihood of observing wage w_{it} *and* schooling S_i
$=$ the likelihood of 'observing' market ability v_i^w *and* schooling S_i
$=$ the likelihood of observing schooling S_{it} *conditional* on market ability v_i^w
$\times$ the likelihood of observing market ability v_i^w,

or, more formally,

$$L\left(\boldsymbol{\Theta}; S_i, v_i^w\right) = L\left(\boldsymbol{\Theta}; S_i \,|\, v_i^w\right) \times L\left(\boldsymbol{\Theta}; v_i^w\right). \tag{2.37}$$

(Though they don't explain it in as many words — or with very formal notation — this principle also justifies Belzil and Hansen's approach (for example, see the first paragraph on page 2080).)

Of course, what we really need is the *log*-likelihood; for this, we simply take logs of equation 2.37:

$$\ell\left(\boldsymbol{\Theta}; S_i, v_i^w\right) = \ell\left(\boldsymbol{\Theta}; S_i \,|\, v_i^w\right) + \ell\left(\boldsymbol{\Theta}; v_i^w\right). \tag{2.38}$$

Equation 2.38 tells us that we will be done if only we can get expressions for $\ell\left(\boldsymbol{\Theta}; S_i \,|\, v_i^w\right)$ and $\ell\left(\boldsymbol{\Theta}; v_i^w\right)$. The second term is straightforward; this is equivalent to writing the log-likelihood for OLS, as we did in the previous lecture. Specifically, we know that:

$$\ln w_{it} = \psi_4 \cdot D_{4i} + \psi_7 \cdot D_{7i} + \psi_{11} \cdot D_{11i} + v_i^w, \tag{2.39}$$

where we define D_4 as a dummy variable for $S = 4$, D_7 as a dummy for $S = 7$ and D_{11} as a dummy for $S = 11$. Therefore, just as we did in the earlier lecture, we can write:

$$\ell\left(\boldsymbol{\Theta}; v_i^w\right) = \ln\left[\sigma_w^{-1} \cdot \phi\left(\frac{\ln w_{it} - \psi_4 \cdot D_{4i} - \psi_7 \cdot D_{7i} - \psi_{11} \cdot D_{11i}}{\sigma_w}\right)\right], \tag{2.40}$$

where $\phi(\cdot)$ refers (as before) to the *pdf* of the normal distribution.

Finding an expression for $\ell(\boldsymbol{\Theta}; S_i \,|\, v_i^w)$ is a bit more difficult. We learned earlier that S_i is determined by whether $v_i^s - v_i^w$ falls between any two cut-offs. Let's consider the case of $S = 4$; we know that $S = 4$ if $v_i^s - v_i^w \in [\gamma_4, \gamma_7)$. For simplicity, we will use this to write the probability that $S = 4$.

First, recall that we know that:

$$v_i^s \,|\, v_i^w \sim \mathcal{N}\left(0, \sigma_s^2\right), \tag{2.41}$$

so that

$$\left.\frac{v_i^s}{\sigma_s}\right| \, v_i^w \sim \mathcal{N}(0,1). \tag{2.42}$$

The previous result is useful, but not useful enough — we need to be able to write the probability that v_i^s is less than $v_i^w + \gamma_7$, *conditional* on observing v_i^w. That is:

$$\Pr\left(v_i^s \leq v_i^w + \gamma_7 \,|\, v_i^w\right) = \Pr\left(\left.\frac{v_i^s}{\sigma_s} \leq \frac{v_i^w + \gamma_7}{\sigma_s}\right| \, v_i^w\right) \tag{2.43}$$

$$= \Phi\left(\frac{v_i^w + \gamma_7}{\sigma_s}\right), \tag{2.44}$$

where $\Phi(\cdot)$ refers (as before) to the *cdf* of the normal distribution.

Of course, this reasoning extends to other cutoffs as well. And, once we have the conditional probability of v_i^s lying below a given cutoff, we can use this to write the probability of $v_i^s - v_i^w$ lying between, say, γ_4 and γ_7, conditional on knowing v_i^w:

$$\Pr\left(v_i^s - v_i^w \in [\gamma_4, \gamma_7) \,|\, v_i^w\right) = \Pr\left(v_i^s \in [v_i^w + \gamma_4, v_i^w + \gamma_7) \,|\, v_i^w\right) \tag{2.45}$$

$$= \Phi\left(\frac{v_i^w + \gamma_7}{\sigma_s}\right) - \Phi\left(\frac{v_i^w + \gamma_4}{\sigma_s}\right). \tag{2.46}$$

(This will look more familiar after you study the 'ordered probit' model later this term.)

Therefore, by the same logic and with some abuse of notation (we're not really supposed to use ∞ like this!), we can write $\ell(\boldsymbol{\Theta}; S_i \,|\, v_i^w)$ for $S = 0$, $S = 4$, $S = 7$ and $S = 11$:

$$\ell(\boldsymbol{\Theta}; S_i \,|\, v_i^w) = \begin{cases} \ln\left\{\Phi\left(\frac{v_i^w+\gamma_4}{\sigma_s}\right) - \Phi\left(\frac{v_i^w-\infty}{\sigma_s}\right)\right\} & \text{if} \quad S = 0 \\ \ln\left\{\Phi\left(\frac{v_i^w+\gamma_7}{\sigma_s}\right) - \Phi\left(\frac{v_i^w+\gamma_4}{\sigma_s}\right)\right\} & \text{if} \quad S = 4 \\ \ln\left\{\Phi\left(\frac{v_i^w+\gamma_{11}}{\sigma_s}\right) - \Phi\left(\frac{v_i^w+\gamma_7}{\sigma_s}\right)\right\} & \text{if} \quad S = 7 \\ \ln\left\{\Phi\left(\frac{v_i^w+\infty}{\sigma_s}\right) - \Phi\left(\frac{v_i^w+\gamma_{11}}{\sigma_s}\right)\right\} & \text{if} \quad S = 11. \end{cases} \tag{2.47}$$

So we're done, at last!

2.2.5 The story so far…

We started with a simple question: *how does education affect earnings?* The key difficulty in answering that question is the problem of unobservables: the concern that those with more education may not be comparable to those with less. In this lecture, we have built a structural model in order to try to deal with that problem. (Of course, structural models can be used for many more purposes than merely dealing with endogeneity — as we will consider in more detail in the next lecture.)

Specifically, we built a *microeconomic model* in which students decide their optimal investment in education depending on their own (unobservable) abilities. Then, by adding an assumption about the distribution of those unobservables, we were able to write down an *econometric model*, capable estimation by the maximum-likelihood methods that we considered last term. We will simulate and estimate the model in the next class.

2.2.6 'Adolescent econometricians'?

Our structural model rests upon a number of important assumptions that we have stated explicity. However, one of our most fundamental assumptions has remained implicit: namely, that Tanzanian households *know* the true return to education, and act rationally in deciding their educational investments. This is not necessarily an unreasonable assumption — it is surely a much better approach, for example, than assuming that expectations are formed arbitrarily or uniformly — but there is no reason necessarily to believe that Tanzanians' subjective expectations about the returns to education match the true market return. Manski (1993) summarised the issue in his paper on 'Adolescent Econometricians':

> Economists analyzing schooling decisions assume that youth, having compared the expected outcomes from schooling and other activities, choose the best feasible option. Viewing education as an investment in human capital, we use the term *returns to schooling* to refer to the outcomes from schooling relative to nonschooling.
>
> Given the centrality of the expected returns to schooling in economic thinking on educational behavior, it might be anticipated that economists would make substantial efforts to learn how youth form their expectations. But the profession has traditionally been skeptical of subjective data; so much so that we have generally been unwilling to collect data on expectations. Instead, the norm has been to make assumptions about expectations formation.

Recent empirical evidence suggests that subjective expectations about the returns to education may not match earnings patterns. For example, Jensen (2010) found in the Dominican Republic that students' perceived returns to schooling were substantially lower than what earnings data would suggest are the true returns.[17] Further, Jensen found that students who were randomly selected to

[17] Of course, this is merely indicative evidence: it does not necessarily mean that the perceived returns are wrong and that the naïve interpretation of the earnings data is correct.

have the higher measured returns explained to them completed (on average) an additional 0.20 to 0.35 years of school over the next four years as a result.

2.2.7 Structural modelling: Hubris or humility?

The social anthropologist Harri Englund has said this:[18]

> A key procedure by which human rights discourse in Malawi and elsewhere has depoliticized the exercise of power is the denial that human rights acquire significance situationally. The procedure is familiar from a wide range of contemporary contexts and is one that, according to the French philosopher Alain Badiou, posits a universal human subject who is split into two modalities. *On the one hand, the subject is passive and pathetic, the one who suffers. On the other, the subject is active, the one who identifies suffering and knows how to act*... The education of the poor and the ignorant has long been an aspect of liberal democracy, pregnant with historical parallels with missionary and colonial projects in many African settings.
>
> *(Englund, 2006, p.32; quotes omitted; emphasis added)*

Englund was writing about human rights in Malawi, but the distinction that he identifies might also be relevant to us as researchers in development economics: how often, I wonder, do *we* fall into the trap of framing ourselves as the 'active subjects', identifying the suffering of 'passive and pathetic' communities and presuming that we know how they (and policymakers) should act?

The emphasis of this lecture has been upon the problem of *endogeneity*: we are using a structural model in order to identify the causal effect of education on earnings. But perhaps there is a deeper, larger advantage to using structural models: by demanding that we write down very explicitly our assumptions about how individuals optimise, the structural methodology encourages us to think about individuals in developing economies as *active, intelligent agents*, trying to make the best decisions possible given a range of constraints. Of course, there is nothing about alternative methods (for example, randomised control trials) that *precludes* thinking about poor communities in this way — though I think it *is* much easier, under traditional econometric methodologies, to decide to "first, find out what works" without much worrying about *why* it works and *how it might work in different circumstances*.[19] In effect, experimental and quasi-experimental methods can sometimes encourage researchers to study the *consequences* of a policy without thinking very carefully about the *incentives and constraints* faced by those affected by the policy. In contrast, for the structural methodology, questions of 'what works' and 'why it works' are largely inseparable.

On the other hand, a critic might argue that the assumptions of structural models are far too strong to describe adequately the process by which poor households make decisions. In this lecture, for example, we have assumed that Tanzanian households choose their children's education perfectly

[18] Englund (2006): *Prisoners of Freedom: Human Rights and the African Poor*, University of California Press.
[19] See, for example, Andrew Leigh, *First, Find Out What Works*, Australian Financial Review, 4 October 2007.

rationally, using time-separable exponential discounting, and that their only objective is to maximise future earnings. We know that this cannot literally be *true*, but should we consider it as *reasonable*? And, reasonable or otherwise, do we *really* treat Tanzanian households as active, intelligent agents if we make such strong assumptions about their behaviour? Is it intellectual hubris to reduce the complexity and subtlety of African education to four equations? Or does the structural methodology demand intellectual humility, by forcing empirical researchers to state explicitly their underlying assumptions? I don't know the answer to any of these questions, but I think they present interesting issues for thinking about the value of structural models for development economics. They are questions that I leave for you to consider...

3 Class: Using Structural Models

I haven't tried it because I don't like it.

Guinness Beer poster

Note: We will consider Exercise 1 and Exercise 2 in class (and you do not need to submit any work from these exercises). Exercise 3 is a take-home exercise that we will *not* have time to discuss in class. However, you are encouraged to submit answers to this question; I will be happy to provide feedback or answer questions on any exercises that you submit (including Exercise 1 and Exercise 2, if you wish).

3.1 Exercise 1: Estimating from simulated data

In this exercise, we simulate data according to the model that we have created. We then use the `ml` command to specify the customised log-likelihood for the model and use it to estimate the data.

First, look at the file to simulate the data:

```
. doedit SimulateData
```

Question 1 *What is the discount factor, β, in the simulated data?*

Question 2 *What is the duration of the student's schooling/work life, T, in the simulated data?*

Question 3 *What values are taken by ψ_4, ψ_7 and ψ_{11} in the simulated data?*

Now simulate the data:

```
. do SimulateData
```

Question 4 *How many students complete 11 years of education in the simulated data? What about if $T = 80$? How does this number change if $T = 20$? Explain the inuition for the effect of T on the number of students choosing $S = 11$.*

Second, look at the file to estimate the model:

```
. doedit SchoolingML
```

Question 5 *What is the name of the program used to define the log-likelihood function for `ml`?*

Question 6 *How does the program deal with a cut-off being ∞ or $-\infty$?*

Question 7 *What numerical **tech**niques is `ml` instructed to use, and in what order? (*Hint*: `help ml` might be useful...)*

Third, run the estimation program to define the log-likelihood:

```
. do SchoolingML
```

(Notice that there is a premature `exit` command in `SchoolingML.do`. This is so that we can run the additional commands separately.)

Question 8 *Run the next line of code (`ml model lf...`). What does it do?*

Question 9 *`Check` the estimator. Are there any problems? If we had solved our model wrongly, would Stata realise this?*

Question 10 *`Search` for a set of starting values, and `report` on what those values are. What is the starting value of the log-likelihood function? What is the starting value for ψ_4?*

Question 11 *`Maximise` the log-likelihood function and `graph` the convergence. What was the maximum value of the log-likelihood function achieved? What is $\hat{\psi}_4$? Did `ml` converge 'slowly', or did it jump to the final value?*

Question 12 *`Maximise` again (after issuing the `ml model` and `ml search` commands again). Did you get the same results?*

Find the estimate for β:

```
. nlcom exp(-exp(-[EstBeta]_cons/10))
```

Question 13 *What is the estimated value for β, and what is its 95% confidence interval? Any thoughts on why we enter β in such a strange way? (*Hint*: What happens if `[EstBeta]_cons` goes towards ∞? What if it goes towards $-\infty$?)*

3.2 Exercise 2: Estimating and testing on real data

We are now going to take our structural model to real data. Notice that, if we hadn't simulated first, we would have *no way of knowing* whether our estimator was performing correctly — if you are intending to build a customised log-likelihood function, you really should simulate first!

First, clear the memory and load the file `SimpleILFS`:

```
. clear
. use SimpleILFS
```

(`SimpleILFS` is a sample of 2000 observations from the ILFS data. For present purposes, we need only (log) income and education (recorded as dummy variables).)

Question 14 *Run an OLS regression of income on the schooling dummies (allow a constant term). What are the OLS estimates of ψ_4, ψ_7 and ψ_{11}?*

Second, `max`imise the log-likelihood using the ILFS data (you may want to delete the premature `exit` and the `ml check` in the file `SchoolingML.do`):

```
. do SchoolingML
```

Question 15 *What are the estimates from our structural model for ψ_4, ψ_7 and ψ_{11}? Why are these estimates so much higher than the OLS estimates?*

Question 16 *In many respects, the estimates for ψ_4, ψ_7 and ψ_{11} are ridiculously large. Perhaps this should not surprise us; after all, our model is very simplistic. What, if anything, would* you *suggest changing about the model in order to produce more reasonable estimates?*

Question 17 *What is the estimated value of β from the ILFS data? Does this seem like a reasonable value to you?*

We often treat education as having a linear effect — that is, we often assume that each year of education increases log earnings by the same amount. Therefore, we might be interested in testing the hypothesis:

$$H_0 : \psi_4 = \frac{4}{7} \cdot \psi_7 = \frac{4}{11} \cdot \psi_{11}. \tag{3.1}$$

Question 18 *What is the value of the maximum of the log-likelihood function when we do* not *impose this restriction?*

Now let's impose the restriction implied by H_0. I have already coded `constraint 6` and `constraint 7` in `SchoolingML.do`; now we need to apply these constraints to the estimation: do this by adding '`constraints(6 7)`' at the end of the `ml model` command (and save the file, of course!).

Question 19 *Run `SchoolingML.do` again with the constraints imposed. What is the value of the maximum of the log-likelihood under the restrictions? What is the value of the Likelihood Ratio test statistic? What is the p-value for the test? (*Hint*: Decide how many degrees of freedom the test implies, and use the `chi2tail` function.)*

Finally, let's consider the *shape* of $\psi(S)$, so far as we can estimate it.

Question 20 *What does the (unrestricted) model imply to be the annual return for each of four years of education? What does the model imply is the annual return for the next three years? And the next four years after that?*

Suppose that I hypothesised, on the basis of the answer to the previous question, that the annual return to education for years 1–4 is the same as the annual return to education for years 8–11.

Question 21 *Write this claim as a hypothesis (*i.e. *algebraically). Implement it as a constraint and test the hypothesis using a Likelihood Ratio test. (Note: You may need to use `ml max, difficult` instead of just `ml max` to maximise the restricted model.) Find the p-value for the test. What do you conclude? Perform a Wald test on the same hypothesis. Which test is easier to implement? Which p-value should we prefer?*

3.3 Exercise 3: Attanasio, Meghir and Santiago (2012)

This take-home exercise requires reading a recent working paper in which a structural model is used to analyse the results of a randomised control trial. You should read the paper and submit answers to the following questions. You are *not* expected to follow the mathematical derivations closely; however,you *are* expected to understand the principles behind the structural estimation technique, and the motivation for using such a structural method in this context. (The estimation technique, of course, is very similar — albeit a bit more complicated — than what we have covered in our simple version of the Belzil and Hansen model.)

(i) Attanasio *et al* discuss the possibility that 'income earned by the child (in school as a scholaraship or in work as a wage) to be non-separable from the activity that generated it' (page 47). What do the authors mean by this? Suggest some reasons for this non-separability. Why is this particularly relevant to the use of a structural model (rather than, say, a simple difference-in-differences estimator)?

(ii) Attanasio *et al* use their structural model to consider possible 'general equilibrium effects' of the PROGRESA program. Explain how such effects might operate in this context.

(iii) The estimation is limited to boys. There are surely many good reasons *against* such a limitation. What might *justify* such a choice?

(iv) Attanasio *at al* follow other authors — including Belzil and Hansen — in modelling the choice of schooling with a forward-looking *dynamic* model. But why not just simplify things into a *static* model?

(v) What variables do the authors use as 'taste shifter variables'? We usually use a different name for such variables; what is it?

(vi) What type of estimator do Attanasio *et al* use? (For example, is it OLS, 2SLS, GMM, . . . ?)

(vii) A standard difference-in-difference estimator would compare treatment villages with control villages. But, by using a structural model, Attanasio *et al* are able to draw information simultaneously from *three* comparisons. What are they?

(viii) Attanasio *et al* 'obtain an estimate of the average return to education of 5% per year. . . ' (page 59); however, they obtain this estimate *without* using data on adult earnings! How do they manage this?

(ix) Duflo, Hanna and Ryan (2012) say this: "A primary benefit of estimating structural model of behavior is the ability to calculate outcomes under economic environments not observed in the data" (page 1265). Explain what they mean, using Attanasio *et al*'s analysis to illustrate.

4 Lecture: Microeconometric Structural Models II

References:

- ⋆ BROWNING, M. (2008): "Identification," *Lecure notes for the M.Phil in Economics*; see http://www.nuffield.ox.ac.uk/Teaching/Economics/Browning/AM/. (I encourage you to read all of this; however, only pages 1–18 are required reading.)
- ⋆ HARRISON, G. (2011): "Experimental Methods in Developing Countries", video of a presentation to the CSAE Annual Conference panel session "*Methodology update: Randomised Controlled Trials or Structural Models (or both... or neither...)?*"; see http://www.csae.ox.ac.uk/conferences/2011-EdiA/video.html. (I encourage you to watch the other videos in this panel session — from David McKenzie and Leonard Wantchekon — but only Glenn Harrison's video is required 'reading'.)
- ⋆ MACE, B. (1991): "Full Insurance in the Presence of Aggregate Uncertainty," *The Journal of Political Economy*, 99(5), 928–956.
- BROWNING, M. (2009): "Estimation," *Lecure notes for the M.Phil in Economics*; see http://www.nuffield.ox.ac.uk/Teaching/Economics/Browning/AM/.
- DERCON, S., DE WEERDT, J., BOLD, T., AND PANKHURST, A. (2006): "Group-based Funeral Insurance in Ethiopia and Tanzania," *World Development*, 34(4), 685–703.
- DERCON, S., AND P.KRISHNAN (2000): "In Sickness and in Health: Risk Sharing Within Households in Rural Ethiopia," *The Journal of Political Economy*, 108(4), 688–727.
- KOOPMANS, T.C. AND REIERSØL, O. (1950): "The Identification of Structural Characteristics," *The Annals of Mathematical Statistics*, 21(2), 165–181.
- KINNAN, C. (2010): "Distinguishing Barriers to Insurance in Thai Villages," *Working paper*; see http://faculty.wcas.northwestern.edu/~cgk281/.
- PORTER, C. (2008): "Examining the Impact of Idiosyncratic and Covariate Shocks on Ethiopian Households' Consumption and Income Sources," *Working paper*; see http://oxford.academia.edu/CatherinePorter/Papers.

In the previous lecture and class on structural modelling, we considered one specific structural model — a simplified version of Belzil and Hansen (2002) — in order to consider the basic concept of structural modelling and to explore the use of maximum likelihood as a method for estimating such a model. In this lecture, I would like to consider some broader issues about structural modelling in microeconometrics; we will consider some motivations for the methodology, examine the concepts of identification and estimation, and weigh costs and benefits from using the approach. To explore these issues, we will use a motivating example about community insurance in Ethiopia.

4.1 Mace (1991): "Full Insurance in the Presence of Aggregate Uncertainty"

What role does exposure to risk play in determining vulnerability to poverty? This is a very important question in development economics. If risk is an important determinant or incident of poverty then the implementation of institutions to manage risk (for example, microinsurance schemes) would be a valuable policy for the poor.

In this lecture, we pose a question fundamental to thinking about the effects of risk: *how well do households manage to insure shocks between themselves?* It is entirely possible that many poor communities in developing economies are so tightly-knit (or, indeed, have such efficient economic institutions) that they manage to insure individuals fully against risk — for example, through the benevolence of the group in individuals' times of hardship, or through the operation of institutions for formal risk management (such as funeral societies in Ethiopia and Tanzania: see Dercon *et al* (2006)). The extent to which a community manages to insure its members against idiosyncratic risk should be a very important issue in thinking about how best to assist that community. For example, if a community is *very good* at insuring idiosyncratic risk, it may be that offering individual microinsurance contracts is not very helpful — the money may be better spent through insuring aggregate risk, or even on completely different kinds of support programs. If a community *does not* manage to insure idiosyncratic risk well, risk is likely a more important determinant of individual poverty, and the case for policy intervention to insure idiosyncratic risk is stronger.[20]

In this lecture, will explore the way that a structural model may help us to answer that question. Specifically, we will explore the methodology developed by Mace (1991) (for the US context) and subsequently implemented using data from a number of developing economies; we will use the methodology to consider vulnerability to risk in rural Ethiopia.

Mace's method is an interesting and important one. However, as discussed above, it is *not* the ultimate focus of this lecture. Instead, I would like to *use* Mace's method to provide a general overview of how we might use a structural model to think about an important issue in development economics.

4.2 Starting with an intuition...

We have our research question: *how well do rural Ethiopians manage to insure shocks between themselves?* We also have our data: the Ethiopian Rural Household Survey (discussed further shortly). We will soon develop a formal model to help us with this question. However, formal models really only ever implement and tighten researchers' *intuition* about a problem — because formal models can be developed to emphasise one aspect or another of a particular circumstance,

[20] Note immediately that *exposure to risk* has nothing directly to do with whether or not households have purchased *formal insurance products*. As this paragraph has noted, households may be well insured against risk despite not having any formal insurance, or may be poorly insured even with such products. Sadly, this is not the impression sometimes given by leading researchers in development economics; for example, click the 'Risk' tab on the 'data' page of the website for Banerjee and Duflo's 2011 book *Poor Economics*...

there is little point in trying to build a formal model without having thought carefully about the problem and what we might expect to observe.

In this case, we are thinking about people's exposure to — and management of — idiosyncratic risk. Intuitively, our sense might go something like this...

(i) Everyone faces shocks in their life — good shocks (*e.g.* finding a new job, harvesting a good crop, *etc*) and bad shocks (*e.g.* death of a family member, requirement to pay a dowry, *etc*).

(ii) An individual 'in autarky' — having no access to economic interactions with others — would have to bear the full cost (or benefit) of those shocks. That is, an individual in autarky would have no way of insuring the shocks.[21]

(iii) However, an individual living in a society with risk insurance would bear *less than the full cost (or benefit)* of a given shock. The 'bad' shocks would not produce an impact as costly as under no-insurance, and the 'good' shocks would not be as beneficial.

(iv) Consider the most extreme case — a collective society where everyone always receives an equal share of the group wealth. In this case, the individual would feel *no* individual impact from a shock — 'good' and 'bad' shocks at the level of the individual would be spread completely across the group, so that *everyone* feels the effect of *every* shock, but *no individual feels the effect of his or her own idiosyncratic shocks more than does the group*.

From this intuition, we might think of a useful reduced-form testing strategy: *test whether individual outcomes vary when individuals face measurable 'good' and 'bad' shocks* — for example, test whether households consume more when harvesting a good crop, or consume less after the funeral of a family member. This is a good intuition, and a useful starting point. However, I think there are several important reasons why *merely* proceeding to estimation at this point would be missing a valuable opportunity.

First, **we haven't really formulated a test**! We have an *intuitive sense* that it would be *interesting* to know how individual outcomes vary with individual shocks — but we have no sense of what a particular coefficient would mean. Suppose, for example, that we regress individual consumption on whether a household has paid for a funeral in the past month. Suppose we then find that having to pay for a funeral reduces consumption by 10%, and suppose we find that this is significantly different from zero. This is interesting and important — but what does it *mean*? Is "10%" a 'big number', or a 'small number'? What does this tell us about poor communities and their ability to deal with risk? It is likely that policymakers — and academic researchers generally — will be much *less* interested about simply the effect of funeral payments and much *more* interested in general insights into the economic problems that households face; yet a result like this, without an accompanying theoretic framework, gives very little general insight. The economist Arthur Goldberger is famously credited with declaring, "Every estimator is a consistent estimator of *something*"; our

[21] The individual *could*, perhaps, 'self-insure' — by saving in order to smooth consumption over time; however, we will be ignoring this kind of behaviour.

simple reduced-form approach would give us consistent estimates of *something* — but of *what*?

Second, by leaping straight to a reduced-form approach, **we have closed our eyes to the potential insights of economic theory**. Even if we do not think that rural Ethiopians behave as perfectly rational economic agents, we may be concerned to ensure that our testing strategy is at least consistent with economic theory. In some sense, we may think of economic theory as bringing some *information* to assist with the problem; just as we would never just throw away observations from survey data, we should not throw away the information that theory can provide.[22] Further, we should be concerned that, without economic theory, we have *no clear sense* about *how* to interpret our estimates. For example, the intuitive discussion earlier suggested that, in a completely collective society, individual outcomes would vary with *group* outcomes, but *not* with individual shocks. This might be a useful limiting example, but we know that very few completely collective societies actually exist. Is it reasonable to apply the same test to a free-market context in which people own private property and can trade as they wish? Only a formal model can really help us think about this.

Third, we have no real guidance on the **choice and form of outcome variable**. It is fine to talk in general terms about the effect of a shock upon 'individual outcomes' — but *which variable should we use as the outcome variable?* And, once we have selected it, *what functional form should it take (for example, should we take it in levels, logs or something else)?*

Fourth, we have very little **clarity about unobservables** — we will have some intuitive sense that 'other factors' may complicate our analysis (for example, in the way that 'ability' complicates an analysis of the causal effect of education on earnings), but our intuition itself will provide very little guidance on what these 'other factors' may be. Inevitably, we will need to make some kind of assumption about the role of unobservables; if we proceed merely on the basis of intuition, we may struggle to understand precisely what that assumption implies.

In short, our intuitive approach is a *useful starting point*. However, if we can go further — by building a formal microeconononomic model of how people manage risk — this will likely help a *lot* in addressing these issues. That is what we will do in a moment. First, however, we should understand formally the role of such modelling in econometric analysis.

[22] For example, in section 4 of his "Two examples of structural modelling" (a listed reference for the previous lecture), Professor Browning writes that structural models "impose a discipline on explanations. It is easy enough to think up models and assumptions that rationalise any particular empirical regularity. But these assumptions have implications for other observables and these also have to be rationalised with the same parameters."

4.3 Identification and Estimation: Two separate concepts and two separate steps

The process of econometric analysis always involves moving back and forth between empirical specification and empirical results. If a variable is not significant in my regression, we may consider dropping it; if a variable is highly significant, we might see how it performs in a quadratic specification, or see how it interacts with some other variable.

This is well and good as a matter of *actual empirical practice*. However, when we are trying to think very clearly about the process of econometric analysis — and, in particular, when we are trying to build and estimate structural models — it is generally very useful to *separate* the concepts of (i) deciding on the specification of our model and the appropriate estimator, and (ii) taking that model and estimator to the data. We refer to the first step as 'identification' and to the second as 'estimation'.

More specifically, this is how Professor Browning describes the distinction, in his lecture notes on "Identification" (see the reading list earlier):

> The first [step] is the *identification step*. Very broadly, this asks: what assumptions are needed to answer the question I have given the data I have. More specifically, identification is concerned with whether we can find estimates of our objects of interest if we know what the *population* distributions of the data were. This step does not involve any statistics. The idea is that if we cannot answer the question of interest with the best possible data situation (knowing population values) then we cannot hope to answer it with imperfect information (samples from the population). The second step is the *estimation* (or *inference*) *step*. In this stage we are concerned with finding estimates of the population distributions given the finite sample to hand. This involves the usual econometrics and statistics that you have been learning for a while now.

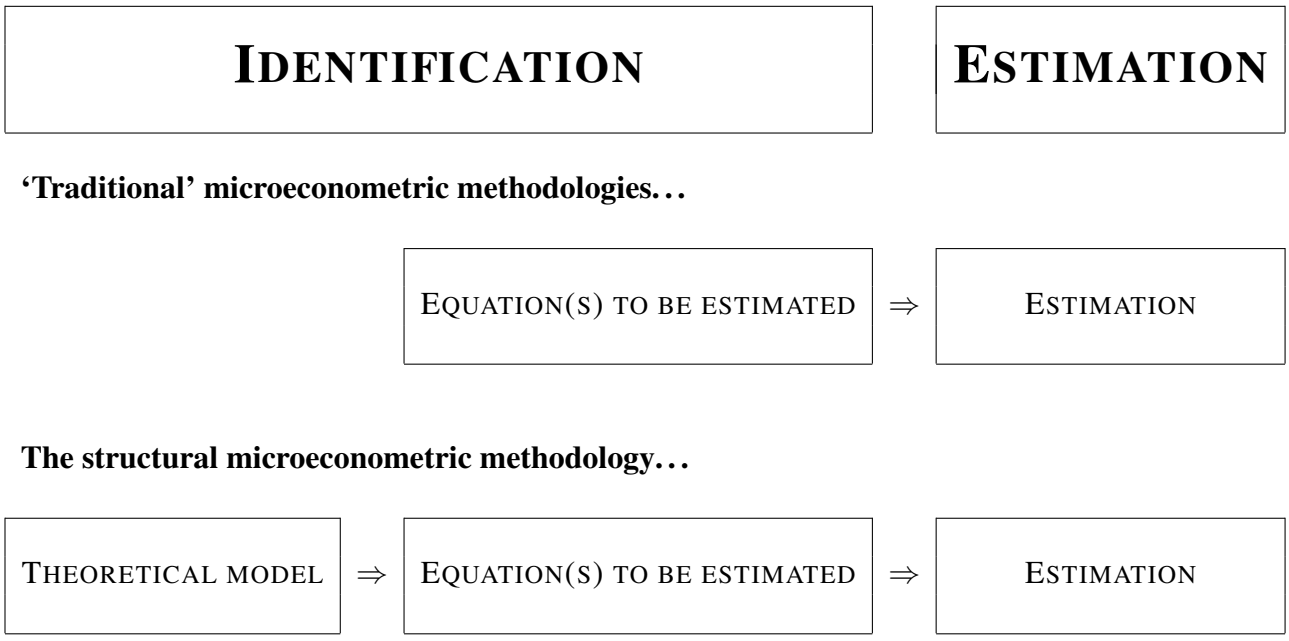
In a seminal article on identification, Koopmans and Reiersøl (1950, pp.169–170) described the issue as follows (emphasis in original):

> [By definition,] a given structure S generates one and only probability distribution $H(y \mid S)$ of the apparent variables. However, statistical inference from any number of observations can relate only to charcateristics of the distribution of the observed variables. The limit of statistical inference is an exact knowledge of this distribution function, a limit not attainable but approachable if very large samples are taken. Anything not implied in this distribution is not a possible object of statistical inference.
> ⋮
> It is therefore a question of great practical importance whether a statement converse to the one just made is valid: can the distribution H of apparent variables, generated by a given structure S contained in a model, be generated by only one structure in that model? This is by no means implied in the definitions given, and it is not generally true.

> ⋮
> Identification problems are not problems of *statistical* inference in a strict sense, since the study of identifiability proceeds from a hypothetical exact knowledge of the probability distribution of observed variables rather than from a finite sample of observations.

In Figure 2.1 (page 26) we considered a very simplistic distinction between 'traditional' microeconometric methodologies and structural methods; Figure 4.1 now adds an equally-simplistic distinction between 'identification' and 'estimation'.

Figure 4.1: **A simplistic distinction between 'identification' and 'estimation'**



Why bother with the distinction? I think there are several good reasons.

(i) The distinction encourages us to think about 'identification' as a distinct and important step; it encourages us to think clearly about the *underlying processes* that we expect our data to reflect. This has obvious relevance to structural modelling, but I think the point is broader than that.

(ii) The distinction encourages us to think carefully about the assumptions that we make about *unobservables*. It is sometimes tempting to think, for example, that we can completely assess the importance of unobservables by looking at our *data*, but this is not true; by thinking about identification as a distinction step to estimation, we remind ourselves that it is, ultimately, *the*

assumptions that we bring to the data, rather than the data itself, that ultimately determines the reasonableness of our approach.

(iii) The distinction forces us to abstract away from problems with data — small samples, measurement error, *etc* — to think about what information we *might possibly* learn from our data. We will consider this last issue again soon, when we talk about under-identification, just-identification and over-identification.

We will get to estimation in due course. But first — before we have even *looked* at our data — we will complete the identification step. We will do this by (i) building an economic model (following Mace (1991)) and (ii) making assumptions about unobservables. To be very clear, I would never *encourage* anyone to theorise without even having looked at the data in question — in fact, that would generally be a bad idea — but my point here is that we don't *need* data for identification. *Estimation is about data; identification is about underlying behaviours and processes.*

4.4 Identification: Building a model

4.4.1 Specifying the model

In this section, we make a series of assumptions in order to build a microeconomic model. We then solve the model by taking the first-order conditions; these conditions will make a specific testable prediction about household behaviour. The discussion here follows Mace (1991) very closely.

Let's start by imaginging that each household chooses to maximise expected utility over an infinite horizon. Imagine that preferences are time separable, so that households discount by the parameter β. Imagine that there are S possible states of the world, where the probability of state τ occurring in period t is $\pi(s_{\tau t})$ (such that the sum of probabilities in each period is always one). Suppose that each household has an in-period utility of $U(c, b)$, where c and b are consumption and shocks respectively. We allow that the shock b and the choice of consumption c may be functions of the state of the world.

Assumption 4.1 *Information*

All households share common information with all other households. Information at time t is represented by one of S possible states of the world $s_{\tau t}$, where $\tau \in \{1, \dots S\}$. The probability of state $s_{\tau t}$ occurring is $\pi(s_{\tau t})$, such that the sum of probabilities for each time period equals 1.

Assumption 4.2 *Preferences*

There are J households which, for simplicity, are treated as living forever. Each household gains utility ($U(\cdot)$) from (i) consumption ($C_t^j(s_{\tau t})$) and (ii) shocks ($b(s_{\tau t})$). As the notation shows, consumption and shocks can change across time (t) and states of the world ($s_{\tau t}$). Each household

is assumed to have preferences that are time-separable and state-separable, with a discount factor $\beta \in (0, 1)$. We can therefore write household j's expected net present utility as:

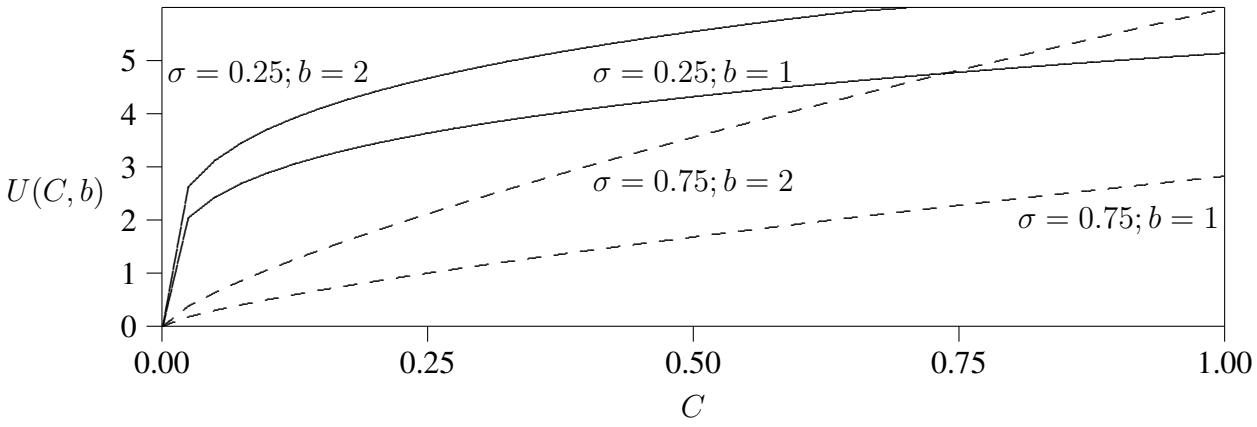
$$\sum_{t=0}^{\infty} \beta^t \sum_{\tau=1}^{S} \pi(s_{\tau t}) \cdot U\left(c_t(s_{\tau t}), b_t(s_{\tau t})\right). \tag{4.1}$$

This, of course, does not give us a functional form for $U(\cdot)$. However, in due course, we will consider Mace's second case (see p.934 of her paper), using 'power utility with multiplicative preference shocks':

$$U(C, b) = \exp(\sigma b) \cdot \frac{1}{\sigma} \cdot C^{\sigma}. \tag{4.2}$$

Figure 4.2 illustrates this functional form, with several different parameters.

Figure 4.2: **Power utility with multiplicative preference shocks**



Assumption 4.3 *Endowments*

In each time period and each state of the world, each household j receives an endowment of the consumption good, $y_t^j(s_{\tau t})$ and this is assumed to be *exogenous* — that is, the household does not, for example, trade off between consumption and production. Similarly, the household cannot save between periods.

Assumption 4.4 *Complete markets*

We assume that 'markets are complete' — that is, in the context of the present model, we assume that households can trade the consumption good *contingent* on any state of the world occurring (*e.g.* I can make a contract with the market such that I will receive x extra units of the consumption good if a family member dies). In effect, this is where the model embeds the critical assumption of full insurance: we are *assuming* complete markets (with full insurance) and then testing whether

that assumption survives the subsequent empirical analysis.

Of course, in making this assumption, Mace does not *literally* ask us to believe that consumers can agree contingent contracts to trade every good in any state of the world. Rather, the point is to test whether that assumption is *reasonable and cannot be falsified by the data*. As Mace herself says:

> The object [of this paper] is to determine how much mileage can be obtained from a model with complete markets, with such features as private information or liquidity constraints omitted. The goal of this research is not to provide evidence that all markets are perfect, but rather to determine whether market imperfections or lack of completeness if an essential feature in explaining consumption allocations. Hence a complete-markets model provides a useful benchmark without requiring researchers to literally accept market perfection.
>
> *(Mace, 1991, pp. 928–929)*

Assumption 4.5 *Price-taking behaviour*

We assume that no individual household can affect the price of the consumption good through its trades; that is, we assume that every household is a price-taker.

Assumption 4.6 *Every household is risk-averse.*

We assume that every household is risk-averse. This will be important shortly because it implies that we can solve the model just by finding the 'interior solution'; if even one household is allowed to be risk-neutral (or even risk-loving), we will obtain a 'corner solution' in which a single household insures all of the other households. Given our assumption about preferences, this assumption simply reduces to a claim that $U(\cdot)$ is concave in consumption.

4.4.2 Solving the model

At first glance, it seems that we have a very difficult problem! Specifically, we have J households, each of whom is assumed to be *separately optimising*, and we are trying to solve for the outcome — to me, at least, it initially seems difficult to know even where to start! Fortunately, Mace's assumptions allow us to simplify the problem substantially, by using a very important and well-established result.

Theorem 1 *The first fundamental theorem of welfare economics*

We have assumed (i) that markets are complete, (ii) that every consumer is a price-taker, and (iii) that more consumption is always preferred to less consumption. Therefore, we know by the first fundamental theorem of welfare economics that the outcome — whatever it may be — *must* be Pareto efficient.

As you will recall, for an outcome to be 'Pareto efficient' means that *it is impossible to make any individual better off without making another individual worse off*. More importantly for our purposes, it *also* means that the outcome can be obtained *as though* produced by a single 'social planner' using a set of welfare weights for the individual members of the community. (This should seem familiar from looking at the Edgeworth Box — remember that, in the Edgeworth Box, the Pareto efficient outcomes are the outcomes along the 'contract curve', and the contract curve maps out outcomes by varying the weight of the different consumers' welfare.)

Therefore, instead of having to solve a very complicated problem for J separate consumers, we can solve the problem *as if* there is a single social planner with a set of J welfare weights. Specifically, the social planner's problem, therefore, is:

$$\max_{\{C_t^j(s_{\tau t})\}} EU = \sum_{j=1}^{J} \omega^j \sum_{t=0}^{\infty} \beta^t \sum_{\tau=1}^{S} \pi(s_{\tau t}) \cdot U\left[C_t^j(s_{\tau t}), b_t^j(s_{\tau t})\right] \tag{4.3}$$

subject to, for *all* t and *all* $s_{\tau t}$,

$$\sum_{j=1}^{J} C_t^j(s_{\tau t}) = \sum_{j=1}^{J} y_t^j(s_{\tau t}). \tag{4.4}$$

Therefore, to simplify for present purposes, we can reduce this to an optimisation problem *only* for time t and state of the world $s_{\tau t}$, using a Lagrangian:

$$\begin{aligned}\mathcal{L}\left(C_t^1(s_{\tau t}), \ldots, C_t^J(s_{\tau t}), \mu_{t,s_{\tau t}}\right) = \beta^t \pi(s_{\tau t}) \sum_{j=1}^{J} \omega^j \cdot U\left[C_t^j(s_{\tau t}), b_t^j(s_{\tau t})\right] \\ + \mu_{t,s_{\tau t}} \cdot \sum_{j=1}^{J} \left[y_t^j(s_{\tau t}) - C_t^j(s_{\tau t})\right]\end{aligned} \tag{4.5}$$

This gives us J first-order conditions in terms of consumption for each period and each state of the world — that is, one for each household.[23] They are as follows:

$$\begin{aligned}\frac{\partial \mathcal{L}\left(C_t^1(s_{\tau t}), \ldots, C_t^J(s_{\tau t}), \mu_{t,s_{\tau t}}\right)}{\partial C_t^1(s_{\tau t})} &= \beta^t \pi(s_{\tau t}) \cdot \omega^1 \cdot U'\left[C_t^1(s_{\tau t}), b_t^1(s_{\tau t})\right] - \mu_{t,s_{\tau t}} = 0; \\ &\vdots \\ \frac{\partial \mathcal{L}\left(C_t^1(s_{\tau t}), \ldots, C_t^J(s_{\tau t}), \mu_{t,s_{\tau t}}\right)}{\partial C_t^j(s_{\tau t})} &= \beta^t \pi(s_{\tau t}) \cdot \omega^j \cdot U'\left[C_t^j(s_{\tau t}), b_t^j(s_{\tau t})\right] - \mu_{t,s_{\tau t}} = 0; \\ &\vdots \\ \frac{\partial \mathcal{L}\left(C_t^1(s_{\tau t}), \ldots, C_t^J(s_{\tau t}), \mu_{t,s_{\tau t}}\right)}{\partial C_t^J(s_{\tau t})} &= \beta^t \pi(s_{\tau t}) \cdot \omega^J \cdot U'\left[C_t^J(s_{\tau t}), b_t^J(s_{\tau t})\right] - \mu_{t,s_{\tau t}} = 0.\end{aligned} \tag{4.6}$$

[23] Of course, if we were doing this explicitly, we might exploit the fact that we only have to solve for $J-1$ consumers, because the Jth consumer's outcome will be precisely implied by the consumption of the other $J-1$. But we don't need to worry about that in simply setting out the principles, as here.

We should pause to consider one of the implications of equation 4.6. It tells us that, for *every* possible state of the world $s_{\tau t}$ in period t, the 'shadow price of income' must be:

$$\mu_{t,s_{\tau t}} = \beta^t \pi(s_{\tau t}) \cdot \omega^j \cdot U' \left[C_t^j(s_{\tau t}), b_t^j(s_{\tau t}) \right], \tag{4.7}$$

across *all* individuals $j \in \{1, \ldots, J\}$. Consider, then, any two individuals, denoted i and j. Equation 4.7 tells us that the consumption of i and j must be determined such that the ratio of marginal utilities equals the inverse ratio of their social weights:

$$\frac{\omega^j}{\omega^i} = \frac{U_i'(\cdot)}{U_j'(\cdot)}. \tag{4.8}$$

But equation 4.7 is not merely of theoretical interest. If we define $\hat{\mu}_t \equiv \mu_t / (\beta^t \cdot \pi(s_{s_{\tau t}}))$, we can take the (natural) log of equation 4.7 to obtain:

$$\ln \hat{\mu}_t = \ln \omega^j + \ln U' \left[C_t^j(s_{\tau t}), b_t^j(s_{\tau t}) \right]. \tag{4.9}$$

Now, returning to the functional form assumed in equation 4.2, we can be more specific about the form of $U'(\cdot)$:

$$U \left[C_t^j(s_{\tau t}), b_t^j(s_{\tau t}) \right] = \exp(\sigma \cdot b_t^j(s_{\tau t})) \cdot \frac{1}{\sigma} \cdot C_t^j(s_{\tau t})^\sigma \tag{4.2}$$

$$\therefore U' \left[C_t^j(s_{\tau t}), b_t^j(s_{\tau t}) \right] = \exp(\sigma \cdot b_t^j(s_{\tau t})) \cdot C_t^j(s_{\tau t})^{\sigma - 1}. \tag{4.10}$$

Substituting back into equation 4.9, we obtain:

$$\ln \hat{\mu}_t = \ln \omega^j + \sigma \cdot b_t^j(s_{\tau t}) + (\sigma - 1) \cdot \ln C_t^j(s_{\tau t}). \tag{4.11}$$

So far, so good! But this doesn't give us anything to estimate yet. In order to do that, we need to think about how to compare the consumption of the *individual* with the consumption of the *group*. To do this, we need to sum equation 4.11 across all households:

$$\sum_{j=1}^{J} \ln \hat{\mu}_t = J \cdot \ln \hat{\mu}_t = \sum_{j=1}^{J} \ln \omega^j + \sigma \cdot \sum_{j=1}^{J} b_t^j(s_{\tau t}) + (\sigma - 1) \cdot \sum_{j=1}^{J} \ln C_t^j(s_{\tau t}) \tag{4.12}$$

$$\therefore \ln \hat{\mu}_t = \frac{1}{J} \sum_{j=1}^{J} \ln \omega^j + \sigma \cdot \frac{1}{J} \sum_{j=1}^{J} b_t^j(s_{\tau t}) + (\sigma - 1) \cdot \frac{1}{J} \sum_{j=1}^{J} \ln C_t^j(s_{\tau t}) \tag{4.13}$$

$$= \omega^a + \sigma \cdot b_t^a + (\sigma - 1) \cdot \ln C_t^a, \tag{4.14}$$

where the last line simply implements some definitions for 'aggregate' terms (following Mace (p.934)). We can, therefore, substitute this expression back into equation 4.11:

$$\ln \hat{\mu}_t = \ln \omega^j + \sigma \cdot b_t^j(s_{\tau t}) + (\sigma - 1) \cdot \ln C_t^j(s_{\tau t}) \tag{4.11}$$

$$\therefore \ln C_t^j(s_{\tau t}) = -\left(\frac{\ln \hat{\mu}_t}{1 - \sigma} \right) + \frac{1}{1 - \sigma} \cdot \ln \omega^j + \frac{\sigma}{1 - \sigma} \cdot b_t^j(s_{\tau t}) \tag{4.15}$$

$$\therefore \ln C_t^j(s_{\tau t}) = \ln C_t^a + \frac{1}{1 - \sigma} \cdot (\ln \omega_j - \omega^a) + \frac{\sigma}{1 - \sigma} \cdot (b_t^j(s_{\tau t}) - b_t^a). \tag{4.16}$$

This, then, is the solution to our model. Equation 4.16, therefore, makes a *very specific* claim, on the basis of our assumed model, about the way that a household's consumption varies with (i) the consumption of the group as a whole, (ii) the household's 'welfare weight' in the community and (iii) idiosyncratic shocks. What we need to know now, then, is whether we can sensibly take equation 4.16 to our data — and, if so, how.

4.5 Under-identification, just identification and over-identification

Earlier, we considered identification and estimation as 'two separate concepts and two separate steps'. By working through the Mace (1991) model, we have gone most of the way towards 'identifying' our model. However, we're not done yet — as attractive as equation 4.16 is, we are not yet able to take it to our data. The reason, of course, is that we cannot observe everything in equation 4.16 — and we do not yet know what estimator we need to use and what assumptions we need to make about the unobservables.

4.5.1 A brief recap: Back to the education-earnings model...

To think carefully about the assumptions we need, I would like to take a brief detour back to a very simple linear model. Let's assume that we are trying to identify the effect of education x on earnings y, and let's presume that we have a theoretical model that predicts the following very simple relationship:

$$y = \beta x + \varepsilon. \tag{4.17}$$

Now imagine, for a moment, that we can observe the *true population distributions* for x and y. That is, let's imagine that we can abstract away from the world of mere mortals and their finite datasets and see the *true underlying distribution* of x and y. (Alternatively, imagine that we can have a *huge* dataset that's *free of measurement error* — this is another way of thinking about the problem.) Will this be enough to correctly estimate β? The answer depends upon what we *believe* about the unobserable, ε. We can see this easily if we (i) multiply everything by x and (ii) take the expectation of both sides:

$$\begin{aligned} y &= \beta x + \varepsilon \\ \therefore xy &= \beta x^2 + x\varepsilon \qquad (4.18) \\ \therefore \mathbb{E}(xy) &= \beta \cdot \mathbb{E}(x^2) + \mathbb{E}(x\varepsilon). \qquad (4.19) \end{aligned}$$

Now, let's assume that the unobservable, ε, is *linearly independent* of years of education, x: that is, $\mathbb{E}(\varepsilon \mid x) = 0$, implying $\mathbb{E}(\varepsilon x) = 0$. In that case, we can simplify our earlier equation further:

$$\begin{aligned} \mathbb{E}(xy) &= \beta \cdot \mathbb{E}(x^2) + \mathbb{E}(x\varepsilon) \\ \mathbb{E}(x\varepsilon) &= 0 \\ \therefore \mathbb{E}(xy) &= \beta \cdot \mathbb{E}(x^2) \qquad (4.20) \\ \therefore \beta &= \frac{\mathbb{E}(xy)}{\mathbb{E}(x^2)}. \qquad (4.21) \end{aligned}$$

This tells us that, if we knew the *true distributions of* x *and* y, we would therefore also know the *true value of* β, ***if*** we *believed* that unobservables (for example, 'ability') were linearly independent of schooling x.

Of course, we mere mortals never know true population distributions — we have to make do with finite datasets. However, we can always do our best, by replacing true population moments by finite estimates — so, for example, we can replace $\mathbb{E}(xy)$ with $\frac{1}{N}\sum_{i=1}^{N} x_i y_i$ and replace $\mathbb{E}(x^2)$ with $\frac{1}{N}\sum_{i=1}^{N} x_i^2$ (this is an example of what is sometimes called the 'analogy principle'). In that case, we get an estimator like this:

$$\hat{\beta} = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2}. \tag{4.22}$$

Seem familiar? This, of course, is the OLS estimator for β! What we have just learned — by using expectations to think about the true population distributions — is that $\hat{\beta}_{OLS}$ estimates the *true value of* β if, *in the true underlying population*, ε is linearly independent of x. What's more, we cannot *test* whether ε is linearly independent of x in the underlying population because we cannot observe ε! That is, whether we we accept $\hat{\beta}_{OLS}$ as an estimator of the true value of β depends on *whether we believe* that $\mathbb{E}(x\varepsilon) = 0$.[24]

Let's press on and accept that we *do* believe that $\mathbb{E}(x\varepsilon) = 0$. In that case, *we have solved for our unknown parameter,* β*, in terms of the population distribution of observable variables*. We can say that 'β is **identified**'. In fact, given that we have no *other* way of knowing or estimating β — for example, no specific guidance from our model, no instrumental variable, *etc* — we can say that 'β is **just-identified**'.

But what if we *did* have some other information on β? Suppose, for example, that our theoretical model predicted instead that:

$$y = 42x + \varepsilon. \tag{4.23}$$

In that case — assuming that we still believe that $\mathbb{E}(x\varepsilon) = 0$ — we can say that β is identified from *two* sources. That is, (i) we can run OLS and obtain $\hat{\beta}$ as we did before, *and/or* (ii) we can trust the model, and take β to equal 42. In that case, we can say not only that 'β is identified' but that 'β is **over-identified**'. This is very interesting and very important — in this case, we can use $\hat{\beta}$ to test whether $\beta = 42$; *if we reject that* $\beta = 42$*, then we must reject our model*.

Finally, consider a slightly alternative model. Suppose we think that education affects earnings through a 'human capital effect' — which we will call β — and through a 'signalling effect' — which we will call α. In that case, we might say:

$$y = (\alpha + \beta) \cdot x + \varepsilon. \tag{4.24}$$

[24] Of course, we *can* test this if we have an instrument. But then, of course, we need to *believe* that the instrument is valid...

Suppose we still believe that $\mathbb{E}(x\varepsilon) = 0$. In that case, for the reasons just set out, we can say that $\alpha + \beta$ is **just-identified** but that neither α nor β are *separately identified*; instead, we can say that α and β are each **under-identified**.[25] That is, *even if we could see the true population distributions of the variables, we would still have no way of inferring the 'human capital effect' distinct from the 'signalling effect'*. By the same logic, if we do *not* believe that $\mathbb{E}(x\varepsilon) = 0$ in the earlier model, we would then have to say there that β is under-identified.

In short...

- If a parameter is **just-identified**, we can estimate it.
- If a parameter is **under-identified**, we cannot estimate it. (In some cases, we can still push the button in Stata and *run* an estimation — but we cannot then give our estimated parameter a causal interpretation. This would be the case, for example, if we were to run OLS on $y = \beta x + \varepsilon$ *without* assuming $\mathbb{E}(x\varepsilon) = 0$.) Another way of thinking about this is that *different values of the parameters could produce identical observed data*.
- If a parameter is **over-identified**, we have *multiple sources* of information from which to estimate it — for example, from the data and from a theoretical model, or from multiple different aspects of the data.[26] In that case, we can *test* whether the multiple possible estimates of the parameter are mutually consistent. If they are *not* mutually consistent, we need to reject the model in its current form; if they *are* mutually consistent, we can continue to believe our model.

4.5.2 Identifiation in the Ethiopian example

So what, then, of our Ethiopian example? Recall the key implication of our model:

$$\ln C_t^j(s_{\tau t}) = \ln C_t^a + \frac{1}{1-\sigma} \cdot (\ln \omega_j - \omega^a) + \frac{\sigma}{1-\sigma} \cdot (b_t^j(s_{\tau t}) - b_t^a). \tag{4.16}$$

If we substitute μ for the last two terms, and include individual j's log income ($\ln y_t^j$), we obtain:

$$\ln C_t^j = \beta_1 \cdot \ln C_t^a + \beta_2 \cdot \ln y_t^j + \mu_{tj}, \tag{4.25}$$

where equation 4.16 implies that $\beta_1 = 1$ and $\beta_2 = 0$.

What can we say, then, about β_1 and β_2? Clearly, they *are* identified — they are identified by our model. But this is not much use — after all, they are identified by the model irrespecive of whether we bother looking at the data! We really need to know whether they are *over-identified* — that is, whether they are also identified by equation 4.25 alone. This requires the standard condition

[25] A question for possible class discussion: *What would the objective function (in this case, the sum of squares) look like in the case of under-identification?*

[26] For example, if we have one or more valid instruments.

for unbiased estimation of OLS, that the unobservable be linearly independent of the explanatory variables:

$$\mathbb{E}(\mu_{tj} \mid \ln C_t^a, \ln y_t^j) = 0. \tag{4.26}$$

Of course, there is nothing to *stop* us making this assumption — but is it *reasonable*? Almost certainly, the answer is *no*. To see why, recall the definition of μ — and note that it includes the household's (log) welfare weight, ω^j. In order to feel comfortable estimating on the basis of equation 4.25, we therefore need to believe that a household's welfare weight (that is, its 'importance' in obtaining consumption) is unrelated (linearly, at least) to that person's income ($\ln y_t^j$). This, frankly, is just not plausible! More importantly, this is an example of a structural model giving *very clear guidance* on what we need to assume in order to believe a particular estimation.

We need a way, therefore, of removing the welfare weight from equation 4.16. Fortunately, we have already assumed that the welfare weight is time-invariant. Therefore, we can remove it by using a *fixed-effect specification* or — more straightforwardly — a *first-difference specification*. Mace uses the latter, and so will we; let's assume that we have access to panel data and rewrite equation 4.16 in differences:

$$\Delta \ln C_t^j(s_{\tau t}) = \Delta \ln C_t^a + \frac{\sigma}{1-\sigma} \cdot (\Delta b_t^j(s_{\tau t}) - \Delta b_t^a). \tag{4.27}$$

As Mace puts it (and I substitute our equation numbers),

> For econometric reasons, the implication for differenced consumption of equation 4.27 is exploited in the empirical work rather than the implication for level consumption from equation 4.16. Remember that equation 4.27 is the first difference of equation 4.16. Individual j's additive fixed effect, $\ln \omega^j - \omega^a$ in equation 4.16, is removed by first-differencing. Using the first-difference specification avoids problems from an omitted-variables bias when the fixed effect is not obseved by the econometrician.
>
> *Mace (1991, p.936)*

Now, if we substitute ε for the last term, and include the *change* in individual j's log income, we obtain:

$$\Delta \ln C_t^j = \beta_1 \cdot \Delta \ln C_t^a + \beta_2 \cdot \Delta \ln y_t^j + \varepsilon_{tj}, \tag{4.28}$$

where equation 4.16 still implies that $\beta_1 = 1$ and $\beta_2 = 0$.

Of course, this does not *remove* the need for an assumption on the unobservable (now denoted ε), but the assumption is now much weaker: we now assume that the change in a household's 'shock' is linearly independent of the change in the household's income (and group consumption). We therefore make one final assumption.

Assumption 4.7 *Preference shocks are linearly independent of log group consumption and log individual income:*

$$\mathbb{E}(\varepsilon_{tj} \mid \ln C_t^a, \ln y_t^j) = 0. \tag{4.29}$$

Of course, it is unlikely that this assumption is literally *true*. As Mace puts it,

> Unbiased estimation of the coefficients requires a zero correlation between the disturbance term and the right-hand-side variables of aggregate consumption and household income. The disturbance term may include both preference shocks and measurement errors from the data. For example, correlations between preference shocks to consumption and income might arise, such as an illness resulting in no employment and reduced consumpton. Correlations between the disturbance and right-hand-side variables might also arise because of measurement errors in the data.
>
> *Mace (1991, p.949)*

However, notwithstanding this concern, it is probably nonetheless *reasonable* to impose Assumption 4.7. With this assumption — and the model that preceded it — we are done at last! We can now say that both β_1 and β_2 are over-identified:

- β_1 is identified by an OLS regression on equation 4.28, *and* is identified by the model such that $\beta_1 = 1$;
- β_2 is identified by an OLS regression on equation 4.28, *and* is identified by the model such that $\beta_2 = 0$.

After a relatively intricate *identification* step, the *estimation* step is — in this case, at least — relatively straightforward. In particular, the over-identification of β_1 and β_2 suggests the following estimation approach.

(i) Run OLS on equation 4.27.

(ii) Perform a *joint test* of the null hypothesis $H_0 : \beta_1 = 1; \beta_2 = 0$. (Since we are doing OLS, there is no problem using Stata's `test` command to perform a Wald test for this.)

(iii) *If the test rejects* H_0, then we *reject the model*. This does *not necessarily* mean that that there is not full risk-sharing — it could be, for example, that there *is* full risk-sharing but that households don't use the utility function outlined in equation 4.2, or that the other assumptions of the model do not hold. However, the usual approach — implicitly, at least — is to maintain the assumption that the functional form and other assumptions are correct, so interpret rejecting H_0 as a *rejection of full insurance*. Of course, this then does *not* tell us *anything* about how *far* Ethiopian communities are from perfect insurance: formulated in this way, the model is capable only of telling us *whether or not* risk is perfectly insured.

(iv) *If the test does not reject* H_0, then we *do not reject the model*. As with any 'do not reject' result, this does *not* mean that we will have 'proved the model correct' — rather, it merely means that we have not *rejected* the model. Our conclusion can then be something like, "We do not reject that agents fully insure, as if having access to perfect and complete markets."

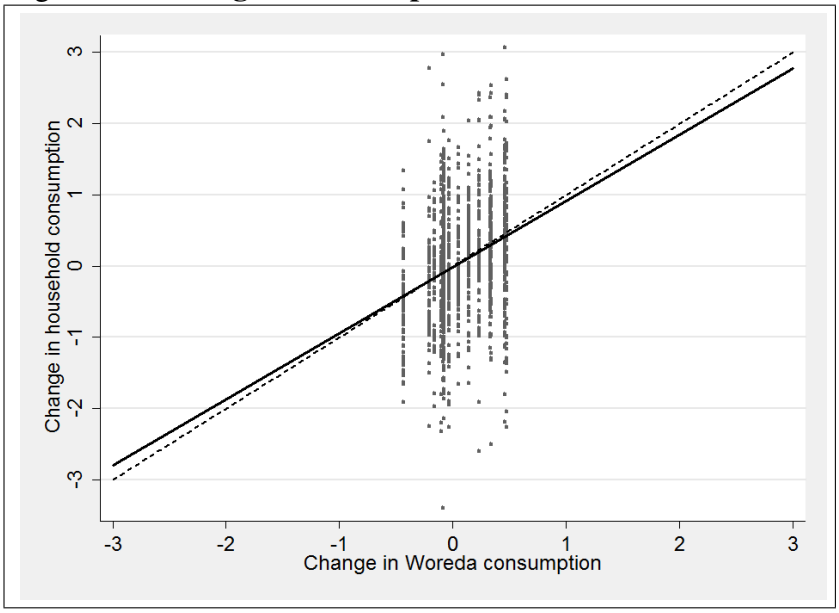
4.6 Estimation

4.6.1 The data

I use data from the Ethiopian Rural Household Survey. This is a panel survey collected by the University of Addis Ababa, the Centre for the Study of African Economies and the International Food Policy Research Institute. The survey covers fifteen districts (called 'Woredas'), and has been conducted in 1989, 1994, 1995, 1997, 1999 and 2004.[27] For this analysis, I use just the data from 1994 and 1995. I use data at the level of the household; there are 1370 households with income and consumption data from the two periods.

Figure 4.3 shows the relationship between change in household consumption and change in Woreda consumption (without controlling for anything). The figure shows that a fitted line between the variables is very close to a 45^o line through the origin (as the model predicts). Figure 4.4 shows the relationship between change in household consumption and change in household income (again, without controlling for anything). The figure shows a slight positive relationship; this suggests that changes in household income cause changes in household consumption.

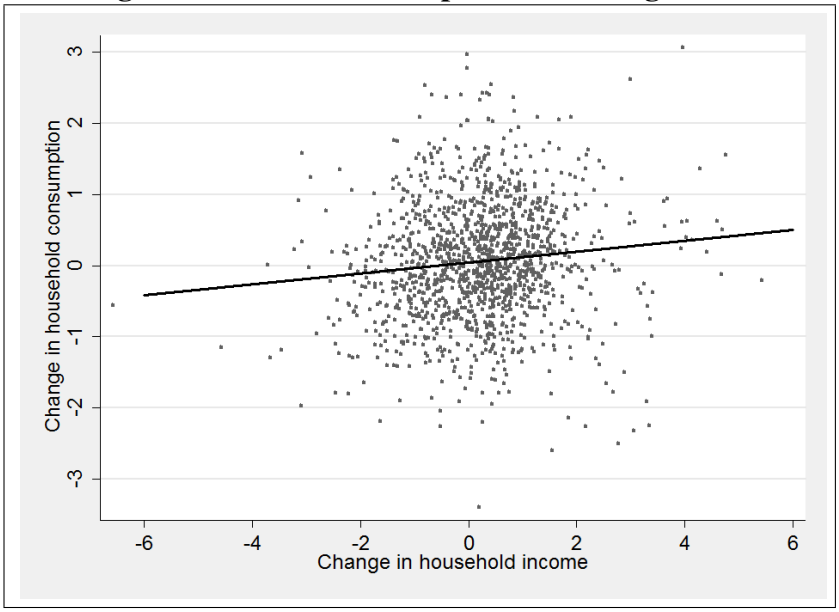
Figure 4.3: **Change in consumption: Households and Woredas**



The solid line shows a linear fit on the data; the dotted line is a 45^o line through the origin.

[27] For more information on the survey, see `http://www.csae.ox.ac.uk/datasets/Ethiopia-ERHS/ERHS-main.html`.

Figure 4.4: **Change in household consumption and change in household income**



The solid line shows a linear fit on the data.

4.6.2 Estimation results

Of course, to test the full-insurance hypothesis properly, we need to run a regression. Table 1 (on the following page) reports two regressions of the form in equation 4.28 (they differ only in that the second specification controls for change in household size; we could, of course, control for other factors too).

The last row in the table reports the p-values from a joint test of the full-insurance null hypothesis.[28] Under both specifications, the test rejects at the 99% confidence level. Therefore, we conclude that rural Ethiopian households do *not* achieve full insurance. Broadly speaking, this result has been shown before using this dataset: see Dercon and Krishnan (2000) and Porter (2008). It is also a result that has been found in a number of other developing countries: see Kinnan (2010). Finally, note that the same result was found in the original Mace (1991) paper (for the power-utility specification presented here) using data from the Consumer Expenditure Survey in the US.

Of course, this is an extremely brief empirical analysis; in reality, we would want to run things much more carefully and with much more attention to the robustness of the result. However, the analysis has helped — hopefully! — to illustrate the principles and the methodology that we have discussed in this lecture.

[28] This is a Wald test, producing an F-statistic having a $\chi^2(2)$ distribution.

Table 1: **A Mace (1991) test on the Ethiopian Rural Household Survey (rounds 2 and 3)**

	OLS.1	OLS.2
	(1)	(2)
Dependent variable: $\Delta \ln(\text{consumption})$ ($\Delta \ln C_t^j$)		
β_1: ΔWoreda consumption ($\Delta \ln C_t^a$)	0.903 (0.037)***	0.902 (0.039)***
β_2: $\Delta \ln(\text{income})$ ($\Delta \ln y_t^j$)	0.06 (0.018)***	0.059 (0.018)***
Δ household size		0.036 (0.019)*
Const.	-.021 (0.01)**	-.018 (0.01)*
Obs.	1370	1370
R^2	0.087	0.089
$H_0 : \beta_1 = 1; \beta_2 = 0$ (p-value)	**0.0049*****	**0.0054*****

Confidence: ***** ↔ 99%, ** ↔ 95%, * ↔ 90%**.

Parentheses show robust standard errors, clustered by Woreda.

4.7 Structural modelling '*Lite*'®

Structural modelling is an important part of econometric analysis, and understanding the key principles of such models can be very valuable in understanding modern empirical research, including in development economics. However, this does *not* mean that every empirical model should use a structural approach. As M.Sc students with a relatively short time available for research, you should be aware of some of the practical *risks* and *dangers* of a structural approach...

- Structural modelling can be very time-consuming; you need to find time to build a microeconomic model *and* to estimate — and, depending on the estimation results, you may need to go back to change the model.
- There are more 'moving parts' — in developing the theory, in identification and in estimation. For this reason, there is more opportunity for things to go wrong! This is particularly true if, as in our Belzil and Hansen example, you use a custom-built estimator.

For this reason, it would probably *not* be a good idea to try to build a full structural model for your extended essay! However, I think there is still a lot that we can learn from the structural approach, even if we are not building a structural model; this is what I mean by structural modelling '*Lite*'.

First, structural models emphasise the importance of *identification*, in addition to *estimation*. This distinction holds in traditional reduced-form contexts, too. Even if you are not building a structural model, it is still very useful to think clearly about the 'identifying assumptions'. What are you forced to assume in order to give your estimation results a causal interpretation? How reasonable are these assumptions? Is it possible to test the robustness of the assumptions? Is it possible that a different estimator (for example, a fixed-effects estimator in the context of panel data) may allow for less restrictive assumptions?

Second, formal microeconomic modelling is often very insightful, even if you can't take the implications of that modelling directly to the data in a structural approach. If you have time — and the approval of your supervisor, of course! — it might be a great idea to try to use a *formal microeconomic model* to capture the key ideas motivating your estimation. This can be valuable for formalising your intuition about a problem and for suggesting additional testable relationships. It may also suggest a potential structural methodology for further research later (for example, "In my extended essay, I use OLS — but, if I have more time to work on this in future, my microeconomic model suggests that I really should use this slightly different estimator to capture this slightly different aspect of the model…"). For example…

- If you are interested in health outcomes (for example, the effects of smoking, or of risky sexual behaviour), you might want to consider a model of optimising *individuals*, in which individuals choose more or less risky behaviour on the basis of different anticipated outcomes (for example, their beliefs about risks of contracting HIV/AIDS, their general life expectancy, *etc*);
- If you are interested in the monitoring and delivery of foreign aid, you could think of a 'principal-agent' model in which *donor governments* and *recipient governments* each optimise (one in the delivery and monitoring of aid, the other in its spending), subject to different incentives and constraints;
- If you are interested in the efficiency of bank credit markets, you could build a model in which *firms seek credit* and *banks respond*, with the firm optimising in the choice of information to reveal and the bank optimising its credit provision;
- If you are interested in political stability in developing nations, you could model the state's incentive to provide public goods and extract natural resource rents, subject to the effect of such policies on the government's probability of remaining in power.

In short, structural modelling is important because it requires us to think about the underlying incentives faced by the people whose behaviour we measure: *it reminds us not to forget the economics in econometrics*.

The last word, however, belongs not to an economist but to a mathematician. This is what Professor Ian Stewart (from the University of Warwick) said on the BBC's 2011 physics documentary *Everything and Nothing*, when explaining why mathematical models are useful in physics:

> When you're trying to understand the universe, it's easy to think what you do is you make lots and lots of observations — you see what's there — and then you fit it all together into your grand picture. But the problem is ***unless you have some sort of idea what the picture should be, you don't know what observations to make — you don't know what's significant***.

5 Class: Evaluation methods

`Leonard Nimoy`: Well, my work is done here.
`Barney Gumble`: What do you mean your work is done? You didn't do anything!
`Leonard Nimoy`: Didn't I?

Marge vs. The Monorail (1993)

We will do two things in this class. In the first hour, we will run some regressions like those in the table on page 63. We will use a modified version (that is, a ***fake*** version) of the Ethiopia Rural Household Survey.[29] The exercises are set out in `EthiopiaAnalysis.do`, and the data is in `EthiopiaClassData.dta`. In the second hour, we will discuss the set exercise on Attanasio et al (2010) (on page 44).

Question 1 *How is 'aggregate log consumption' created in the Mace model? (Is it **log mean consumption**, or **mean log consumption**?)*

Question 2 *What's wrong with the `tsset` command as implemented?*

Question 3 *What are the two levels of 'aggregate' consumption used? Can you suggest some other levels?*

Question 4 *Does it seem to matter that we ran the estimation in first differences? Compare the random-effect and fixed-effect estimations; do they seem different? How could we formally test for their equivalence? What would such a test mean in terms of the Mace model?*

Question 5 *How does the empirical specification change if we assume 'exponential utility' rather than 'power utility with multiplicative shocks'?*

Question 6 *How do these results compare to the results in Mace (1991)?*

Question 7 *What are we overlooking in running the regressions like this? Try some alternative specifications to check the robustness of the earlier conclusions.*

[29] Of course, you are welcome to apply through CSAE to use the actual ERHS; however, note that the version that I am providing for this class is a version that has been further anonymised and modified from that data.