

# Limited Dependent Variables

## (M.Sc in Economics for Development)

Simon Quinn\*

Michaelmas 2022



These notes provide some background for four lectures that I will be giving this year for the M.Sc Quantitative Methods course. I will use slides for the lectures themselves. I will make the slides available online *after* our last 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 and all of these notes to be potentially relevant for the exam (except where noted).

For each lecture, I have ‘starred’ (‘★’) references to Cameron and Trivedi (2005) and to Wooldridge (2002 and 2010). You are required to read at least *one* of these, but you do not need to read more. I have also provided other references; you are not required or expected to read these.

---

\*Department of Economics, Centre for the Study of African Economies and St Antony’s College, University of Oxford: [simon.quinn@economics.ox.ac.uk](mailto:simon.quinn@economics.ox.ac.uk). I must particularly thank Victoria Prowse for her assistance in preparing these lectures. I must also thank Cameron Chisholm, Cosmina Dorobantu, Sebastian Königs, Julien Labonne and Jeremy Magruder for very useful comments. All errors remain my own.

---

# 1 Lecture 1: Binary Choice I

## Required readings (for Lectures 1 and 2):

- ★ CAMERON, A.C. AND TRIVEDI, P.K. (2005): *Microeconometrics: Methods and Applications*. Cambridge University Press, pages 463 – 478 (*i.e.* sections 14.1 to 14.4, inclusive)  
*or*
- ★ WOOLDRIDGE, J. (2002): *Econometric Analysis of Cross Section and Panel Data*. The MIT Press, pages 453 – 461 (*i.e.* sections 15.1 to 15.4, inclusive)  
*or*
- ★ WOOLDRIDGE, J. (2010): *Econometric Analysis of Cross Section and Panel Data* (2nd ed.). The MIT Press, pages 561 – 569 (*i.e.* sections 15.1 to 15.4, inclusive).

## Other references:

- TRAIN, K. (2009): *Discrete Choice Methods with Simulation*. Cambridge University Press.
- GOULD, W., PITBLADO, J. AND SRIBNEY, W. (2006): *Maximum Likelihood Estimation with Stata*. Stata Press.

## 1.1 An illustrative empirical question

Our first two lectures consider models for binary dependent variables; that is, models for contexts in which our outcome of interest takes just two values. We will focus on a simple illustrative question: *how has primary school attendance changed over time in Tanzania?* There are many reasons that this question may be important for empirical researchers — for example, it may be of historical interest in understanding Tanzania’s long-run economic development, or it may be important for considering present-day earnings differences across Tanzanian age cohorts.

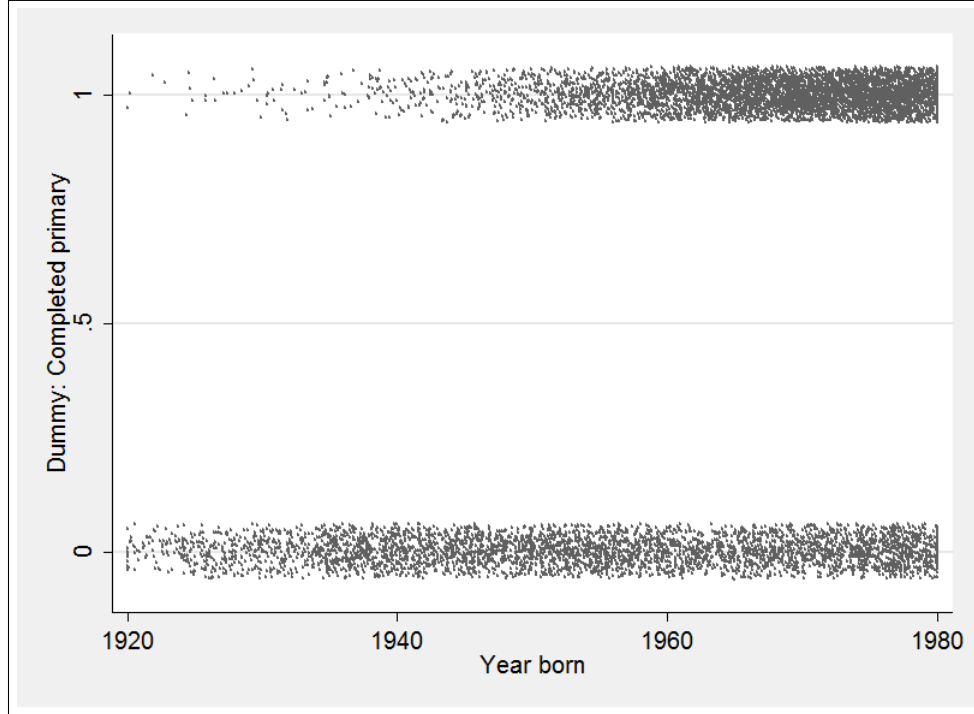
We shall consider this question using data from Tanzania’s 2005/2006 Integrated Labour Force Survey (‘ILFS’). For simplicity, we will consider a single explanatory variable: the year in which a respondent was born. We index individuals by  $i$ , and denote the  $i$ th individual’s year of birth as  $x_i$ . We record educational attainment by a dummy variable,  $y_i$ , defined such that:

$$y_i = \begin{cases} 0 & \text{if the } i\text{th individual } \textit{did not} \text{ complete primary education;} \\ 1 & \text{if the } i\text{th individual } \textit{did} \text{ complete primary education.} \end{cases} \quad (1.1)$$

(Note immediately that — as with all binary outcome models — this denomination is arbitrary: for example, we could just as easily reverse the assignment of 0 and 1 without changing anything of the structure of the problem.)

Figure 1.1 illustrates the data: it shows the education dummy variable on the  $y$  axis (with data points ‘jittered’, for illustrative clarity), and the age variable on the  $x$  axis. Note that we will limit consideration to individuals born between 1920 and 1980 (inclusive).

Figure 1.1: Primary school attainment in Tanzania across age cohorts



## 1.2 A simple model of binary choice

We began with a somewhat imprecise question: *how has primary school attendance changed over time in Tanzania?* More formally, we will be interested in estimating the following ‘object of interest’:

$$\Pr(y_i = 1 \mid x_i). \quad (1.2)$$

That is, we will build and estimate a model of the *probability of attaining a primary school education, conditional upon year of birth*.

As with most econometric outcome variables, investment in primary school education is a matter of choice. We therefore begin by specifying a (very simple) microeconomic model of investment in education. Denote the  $i$ th household’s utility of attending primary school as  $U_i^S(x_i)$  and the utility of not attending school (*i.e.* ‘staying home’) as  $U_i^H(x_i)$ . For simplicity, we will assume that each utility function is additive in the year in which a child was born:<sup>1</sup>

$$U_i^S(x_i) = \alpha_0^S + \alpha_1^S x_i + \mu_i^S \quad (1.3)$$

$$U_i^H(x_i) = \alpha_0^H + \alpha_1^H x_i + \mu_i^H. \quad (1.4)$$

<sup>1</sup> This would be the case, for example, if we think that the ‘utility cost’ of primary education has changed linearly over time — or, indeed, the ‘utility benefit’ from a primary education.

This is a very simple example of an ‘*additive random utility model*’ (‘ARUM’).

Define  $\beta_0 \equiv \alpha_0^S - \alpha_0^H$ ,  $\beta_1 \equiv \alpha_1^S - \alpha_1^H$  and  $\varepsilon_i \equiv \mu_i^S - \mu_i^H$ . Then, trivially, we model a household as having invested in primary education if:

$$\beta_0 + \beta_1 x_i + \varepsilon_i \geq 0. \quad (1.5)$$

We can therefore define a ‘*latent variable*’,  $y_i^*$ :

$$y_i^*(x_i; \beta_0, \beta_1) \equiv \beta_0 + \beta_1 x_i + \varepsilon_i. \quad (1.6)$$

We can express this latent variable as determining our outcome variable for the  $i$ th individual:

$$y_i = \begin{cases} 0 & \text{if } y_i^* < 0 \\ 1 & \text{if } y_i^* \geq 0. \end{cases} \quad (1.7)$$

So far, so good — but we’re still not in a position to estimate the object of interest. To do this, we need to make a distributional assumption.

### 1.3 The probit model

**Assumption 1.1 (DISTRIBUTION OF  $\varepsilon_i$ )**  $\varepsilon_i$  is i.i.d. with a standard normal distribution, independent of  $x_i$ :

$$\varepsilon_i | x_i \sim \mathcal{N}(0, 1). \quad (1.8)$$

You will be familiar with the normal distribution, and with the concepts of the probability density and the cumulative density. Recall that the *probability density function*  $\phi(x)$  is:

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{x^2}{2}\right), \quad (1.9)$$

and that the *cumulative density function* ( $\Phi(x)$ ) has no closed-form expression:

$$\Phi(x) = \int_{-\infty}^x \phi(z) dz. \quad (1.10)$$

Figure 1.2 illustrates. With our distributional assumption, we can now write the conditional probability of primary education:<sup>2</sup>

$$\Pr(y_i = 1 | x_i; \beta_0, \beta_1) = \Pr(\beta_0 + \beta_1 x_i + \varepsilon_i \geq 0 | x_i) \quad (1.11)$$

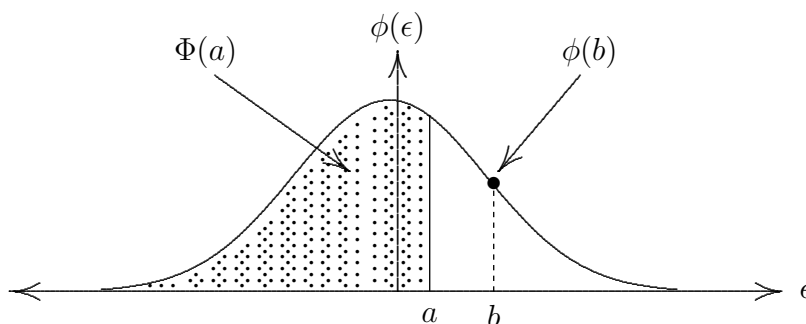
$$= \Pr(-\varepsilon_i \leq \beta_0 + \beta_1 x_i | x_i) \quad (1.12)$$

$$= \Pr(\varepsilon_i \leq \beta_0 + \beta_1 x_i | x_i) \quad (1.13)$$

$$= \Phi(\beta_0 + \beta_1 x_i) \quad (1.14)$$

$$\therefore \Pr(y_i = 0 | x_i; \beta_0, \beta_1) = 1 - \Phi(\beta_0 + \beta_1 x_i). \quad (1.15)$$

<sup>2</sup> Note that equation 1.13 follows from equation 1.12 because, under the assumption of normality, the distribution of  $\varepsilon$  is symmetric.

Figure 1.2: **The standard normal: probability density ( $\phi(\cdot)$ ) and cumulative density ( $\Phi(\cdot)$ )**

Equation 1.14 (and, equivalently, equation 1.15) defines the ‘*probit model*’. There is certainly no need to motivate the probit model with an additive random utility approach, as we have done; indeed, the vast majority of empirical papers (and econometrics textbooks) start merely with some equivalent to equation 1.14. But the additive random utility approach is useful (i) for thinking about how microeconomic models may motivate econometric analysis, (ii) for explaining the ‘latent variable interpretation’ of the probit model, and (iii) to lay the conceptual groundwork for other limited dependent variable models that we will discuss later (in particular, models of multinomial choice). Train (2009, page 14) explains the utility-based approach in discrete choice models as follows:

Discrete choice models are usually derived under an assumption of utility-maximising behaviour by the decision maker. . . It is important to note, however, that models derived from utility maximisation can also be used to represent decision making that does not entail utility maximisation. The derivation assumes that the model is consistent with utility maximisation; it does not preclude the model from being consistent with other forms of behaviour. The models can also be seen as simply describing the relation of explanatory variables to the outcome of a choice, without reference to exactly how the choice is made.

## 1.4 Estimation by maximum likelihood

### 1.4.1 The log-likelihood

Equation 1.14 defines the probit model. But this still requires a method of estimation. The method used for the probit model is *maximum likelihood*.<sup>3</sup> For the  $i$ th individual, the likelihood can be written as:

$$L_i(\beta_0, \beta_1; y_i | x_i) = \Pr(y_i = 1 | x_i; \beta_0, \beta_1)^{y_i} \cdot \Pr(y_i = 0 | x_i; \beta_0, \beta_1)^{1-y_i} \quad (1.16)$$

$$= \Phi(\beta_0 + \beta_1 x_i)^{y_i} \cdot [1 - \Phi(\beta_0 + \beta_1 x_i)]^{1-y_i}. \quad (1.17)$$

The log-likelihood, therefore, is:

$$\ell_i(\beta_0, \beta_1; y_i | x_i) = y_i \cdot \ln \Phi(\beta_0 + \beta_1 x_i) + (1 - y_i) \cdot \ln [1 - \Phi(\beta_0 + \beta_1 x_i)]. \quad (1.18)$$

Denoting the stacked values of  $y_i$  and  $x_i$  as  $\mathbf{y}$  and  $\mathbf{x}$  respectively, we can write the likelihood for a sample of  $N$  individuals as:

$$\ell(\beta_0, \beta_1; \mathbf{y} | \mathbf{x}) = \sum_{i=1}^N \{y_i \cdot \ln \Phi(\beta_0 + \beta_1 x_i) + (1 - y_i) \cdot \ln [1 - \Phi(\beta_0 + \beta_1 x_i)]\}. \quad (1.19)$$

You will be aware that several numerical algorithms may be used to find the values  $(\hat{\beta}_0, \hat{\beta}_1)$  that jointly maximise this log-likelihood — for example, the Newton-Raphson method, the Berndt-Hall-Hausman algorithm, the Davidson-Fletcher-Powell algorithm, the Broyden-Fletcher-Goldfarb-Shanno algorithm, *etc.* Happily, Stata (and other statistical packages) has these algorithms built in, so we can use these algorithms without having to code them ourselves.

### 1.4.2 Properties of the maximum likelihood estimator

Before we go on with our probit example, we should briefly revise several important properties of maximum likelihood estimators. Suppose that we have an outcome vector,  $\mathbf{y}$ , and a matrix of explanatory variables,  $\mathbf{X}$ ; further, suppose that we are interested in fitting a parameter vector  $\beta$ . You will recall that we can generally specify the log-likelihood as:

$$\ell(\beta; \mathbf{y} | \mathbf{X}) = \ln f(\mathbf{y} | \mathbf{X}; \beta); \quad (1.20)$$

that is, the log-likelihood is the log of the conditional probability density (or probability mass) of  $\mathbf{y}$ , given  $\mathbf{X}$ , for some parameter value  $\beta$ . This can formally be described as the *conditional* log-likelihood function, but we usually just term it ‘the log-likelihood’.<sup>4</sup> Further, you will recall that, if

<sup>3</sup> However, this is certainly not the only way we could estimate the probit model. For example, equation 1.14 implies that  $\mathbb{E}(y_i | x_i) = \Phi(\beta_0 + \beta_1 x_i)$ ; the model could therefore also be estimated by Nonlinear Least Squares (*i.e.* a method-of-moments estimator).

<sup>4</sup> Cameron and Trivedi (page 139) note that ignoring the marginal likelihood of  $\mathbf{X}$  is not a problem “if  $f(\mathbf{y} | \mathbf{X})$  and  $f(\mathbf{X})$  depend on mutually exclusive sets of parameters”; that is, if there is no endogeneity problem.

we assume observations are independent across individuals, we can decompose the log-likelihood as:

$$\ell(\boldsymbol{\beta}; \mathbf{y} | \mathbf{X}) = \sum_{i=1}^N \ell_i(\boldsymbol{\beta}; y_i | \mathbf{x}_i) = \sum_{i=1}^N \ln f(y_i | \mathbf{x}_i; \boldsymbol{\beta}). \quad (1.21)$$

The maximum likelihood estimate  $\hat{\boldsymbol{\beta}}_{ML}$  therefore solves:

$$\left. \frac{\partial \ell(\boldsymbol{\beta}; \mathbf{y} | \mathbf{X})}{\partial \boldsymbol{\beta}} \right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_{ML}} = \mathbf{0}, \quad (1.22)$$

where the lefthand side of this expression is called the ‘*score vector*’.

You will recall further that all maximum likelihood estimators share at least four important properties:

- (i) **Consistency:** In 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 that we are trying to estimate some true scalar parameter  $\beta$ , and that we are using a maximum likelihood estimator  $\hat{\beta}_{ML}$ , with  $N$  observations in our sample. Then consistency means that, for *any*  $\varepsilon > 0$ ,

$$\lim_{N \rightarrow \infty} \Pr(|\hat{\beta}_{ML} - \beta| > \varepsilon) = 0. \quad (1.23)$$

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

$$\text{plim}_{N \rightarrow \infty} \hat{\beta}_{ML} = \beta. \quad (1.24)$$

- (ii) **Asymptotic normality:** Assuming some regularity conditions, the asymptotic distribution of a maximum likelihood estimator is normal:<sup>5</sup>

$$\sqrt{N} \cdot (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, I(\boldsymbol{\beta})^{-1}), \quad (1.25)$$

$$\text{where } I(\boldsymbol{\beta}) = -\mathbb{E} \left( \frac{\partial^2 \ell_i(\boldsymbol{\beta}; y_i | \mathbf{x}_i)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \right). \quad (1.26)$$

We generally estimate  $I(\boldsymbol{\beta})$  using:

$$\hat{I}(\hat{\boldsymbol{\beta}}_{ML}) = -\frac{1}{N} \sum_{i=1}^N \left. \frac{\partial^2 \ell_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_{ML}}. \quad (1.27)$$

<sup>5</sup> See, for example, page 392 of Wooldridge (2002) for those conditions. In this context, the conditions are: (i) that  $\boldsymbol{\beta}$  is interior to the set of possible values for  $\boldsymbol{\beta}$ , and (ii) that the log-likelihood is twice continuously differentiable on the interior of that set. We will not need to worry about these conditions in these lectures.

Suppose, then, that we want to test a hypothesis  $H_0: \beta = \beta_0$ . The estimated covariance matrix  $\hat{I}(\hat{\beta}_{ML})^{-1}$  can be used to perform a Wald test. Alternatively, we can perform a Likelihood Ratio test:

$$2 \cdot [\ell(\hat{\beta}_{ML}) - \ell(\beta_0)] \sim \chi^2(k), \quad (1.28)$$

where  $k$  is the number of restricted parameters in  $\beta_0$ .

- (iii) **Efficiency:** Equations 1.25 and 1.26 show that, asymptotically, the variance of maximum likelihood estimators is the ‘*Cramér-Rao lower bound*’ (i.e. the inverse of the Fisher information matrix). That is, maximum likelihood estimators are *efficient*: the asymptotic variance of the maximum likelihood estimator is at least as small as the variance of any other consistent estimator of the parameter.
- (iv) **Invariance:** If  $\gamma = f(\beta)$  is a one-to-one, continuous and continuously differentiable function,  $\hat{\gamma}_{ML} = f(\hat{\beta}_{ML})$ .

### 1.4.3 Goodness of fit in the probit model

For simplicity, let’s return to our earlier example of a probit model with a single explanatory variable. You will be familiar with the  $R^2$  statistic from linear regression models; this statistic reports the proportion of variation in the outcome variable that is explained by variation in the regressors. Unfortunately, this statistic does not generalise naturally to the maximum likelihood context. Instead, the standard goodness-of-fit statistic for maximum likelihood estimates is ‘*McFadden’s Pseudo- $R^2$* ’:

$$R_p^2 \equiv 1 - \frac{\ell(\hat{\beta})}{\ell_0}, \quad (1.29)$$

where  $\ell(\hat{\beta})$  is the value of the maximised log-likelihood, and  $\ell_0$  is the log-likelihood for a model without explanatory variables (so, in the context of our probit model,  $\ell_0$  is the log-likelihood for a probit estimation using  $\Pr(y_i = 1 | x) = \Phi(\beta_0)$ ). You should confirm that we will always obtain  $R_p^2 \in (0, 1)$ .

Additionally, in a binary outcome model, we may wish to report the ‘*percent correctly predicted*’. Wooldridge (2002, page 465) explains:

For each  $i$ , we compute the predicted probability that  $y_i = 1$ , given the explanatory variables,  $x_i$ . If  $[\Phi(\hat{\beta}_0 + \hat{\beta}_1 x_i) > 0.5]$ , we predict  $y_i$  to be unity; if  $[\Phi(\hat{\beta}_0 + \hat{\beta}_1 x_i) \leq 0.5]$ ,  $y_i$  is predicted to be zero. The percentage of times the predicted  $y_i$  matches the actual  $y_i$  is the percent correctly predicted. In many cases it is easy to predict one of the outcomes and much harder to predict another outcome, in which case the percent correctly predicted can be misleading as a goodness-of-fit statistic. More informative is to compute the percent correctly predicted for each outcome,  $y = 0$  and  $y = 1$ .



### 1.4.4 Back to Tanzania...

Table 1.1 reports the results of the probit estimation for Tanzania (see column (1)). We estimate  $\hat{\beta}_0 = -90.395$  and  $\hat{\beta}_1 = 0.046$ ; both estimates are highly significant. Columns (2) and (3) show respectively the estimated mean marginal effect and the estimated marginal effect at the mean (that is, the estimated marginal effect for  $x_i = 1962.627$ ). (We will discuss the concept of marginal effects shortly.) Figure 1.3 shows the predicted probability of primary school attainment:  $\Phi(\hat{\beta}_0 + \hat{\beta}_1 \cdot x_i)$ . (Appendix 1 provides the basic Stata commands for producing these estimates.)

Table 1.1: Probit estimates for primary school attainment in Tanzania

	Estimates	Marginal Effects	
	(1)	Mean effect	Effect at mean
	(1)	(2)	(3)
Year born	0.046 (0.001)***	0.015 (0.0002)***	0.018 (0.0004)***
Const.	-90.395 (2.075)***		
Obs.	10000		
Log-likelihood	-5684.679		
Pseudo- $R^2$	0.165		
Successes correctly predicted (%)	84.7		
Failures correctly predicted (%)	57.2		
Mean of "year born"			1962.627
<i>Confidence:</i> *** $\leftrightarrow$ 99%, ** $\leftrightarrow$ 95%, * $\leftrightarrow$ 90%.			

## 1.5 Normalisations in the probit model

We assumed earlier that  $\varepsilon_i \sim \mathcal{N}(0, 1)$ . But suppose instead that we had assumed more generally that  $\varepsilon_i \sim \mathcal{N}(\mu, \sigma^2)$ . In that case, we would write the conditional probability as:

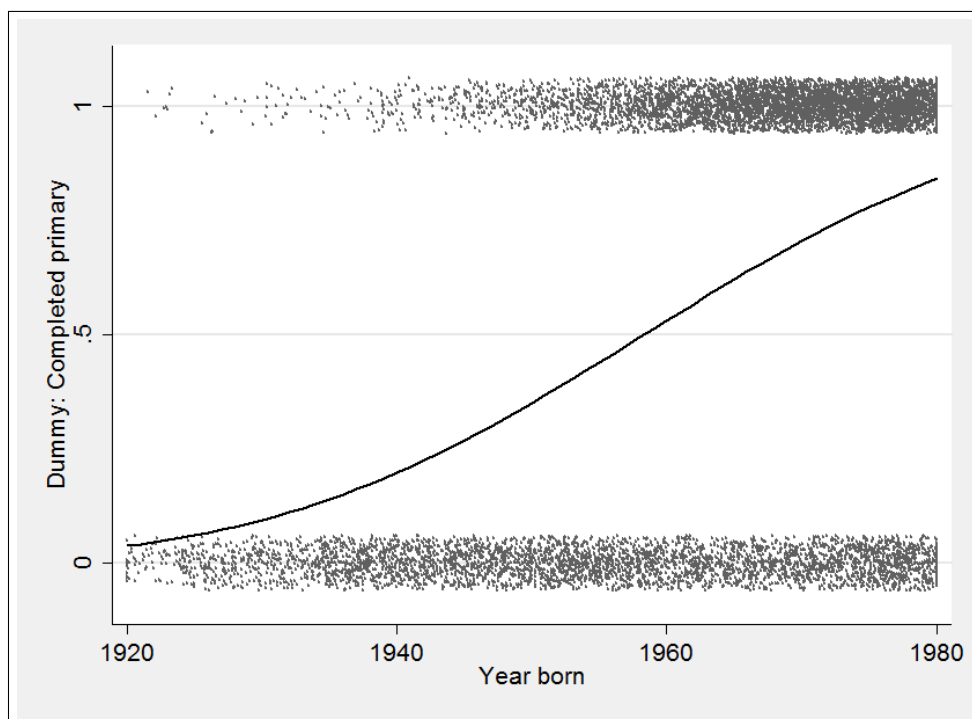
$$\Pr(y_i = 1 \mid x_i; \beta_0, \beta_1) = \Pr(\beta_0 + \beta_1 x_i + \varepsilon_i \geq 0 \mid x_i) \quad (1.30)$$

$$= \Pr(\varepsilon_i \leq \beta_0 + \beta_1 x_i \mid x_i) \quad (1.31)$$

$$= \Pr\left(\frac{\varepsilon_i - \mu}{\sigma} \leq \frac{\beta_0 - \mu}{\sigma} + \frac{\beta_1}{\sigma} \cdot x_i \mid x_i\right) \quad (1.32)$$

$$= \Phi\left(\frac{\beta_0 - \mu}{\sigma} + \frac{\beta_1}{\sigma} \cdot x_i\right) \quad (1.33)$$

$$\therefore \Pr(y_i = 0 \mid x_i; \beta_0, \beta_1) = 1 - \Phi\left(\frac{\beta_0 - \mu}{\sigma} + \frac{\beta_1}{\sigma} \cdot x_i\right). \quad (1.34)$$

Figure 1.3: **Probit estimates for primary school attainment in Tanzania**

This clearly presents a problem: the best that we can now do is to identify the objects  $(\beta_0 - \mu) \cdot \sigma^{-1}$  and  $\beta_1 \cdot \sigma^{-1}$ . That is, the earlier assumptions that  $\mu = 0$  and  $\sigma = 1$  are *identifying assumptions*: they are normalisations without which we cannot identify either  $\beta_0$  or  $\beta_1$ .

This should not come as a surprise: remember that we can always take a monotone increasing transformation of a utility function without changing any of the observed choices. This has an important implication for the way that we interpret the magnitude of parameter estimates from discrete choice models; as Train explains (2009, page 24, emphasis in original):

The [estimated coefficients in a probit model] reflect, therefore, the effect of the observed variables *relative* to the standard deviation of the unobserved factors.

## 1.6 Interpreting the results: Marginal effects

The parameter estimates from a probit model are often difficult to interpret in any intuitive sense; a policymaker, for example, is hardly likely to be impressed if told that “the estimated effect of age on primary school completion in Tanzania is  $\hat{\beta}_1 = 0.048$ ”! Instead, the interpretation of probit estimates tends to focus upon (i) the predicted probabilities of success and, consequently, (ii) the *estimated marginal effects*.

In a binary outcome model, a given marginal effect is the *ceteris paribus* effect of changing one individual characteristic upon an individual’s probability of ‘success’. In the context of the Tanzanian education data, the marginal effects measure the *predicted difference in the probability of primary school attainment between individuals born one year apart*.

Having estimated the parameters  $\hat{\beta}_0$  and  $\hat{\beta}_1$ , the estimated marginal effects follow straightforwardly. For an individual  $i$  born in year  $x_i$ , the predicted probability of completing primary education is:

$$\Pr(y_i = 1 \mid x_i; \hat{\beta}_0, \hat{\beta}_1) = \Phi(\hat{\beta}_0 + \hat{\beta}_1 \cdot x_i). \quad (1.35)$$

Had the  $i$ th individual been born a year later, (s)he would have a predicted probability of completing primary education of:

$$\Phi(\hat{\beta}_0 + \hat{\beta}_1 \cdot (x_i + 1)). \quad (1.36)$$

For the  $i$ th individual, the estimated marginal effect of the variable  $x$  is therefore:

$$M_d(x_i; \hat{\beta}_0, \hat{\beta}_1) = \Phi(\hat{\beta}_0 + \hat{\beta}_1 \cdot (x_i + 1)) - \Phi(\hat{\beta}_0 + \hat{\beta}_1 \cdot x_i). \quad (1.37)$$

$M_d(x_i; \hat{\beta}_0, \hat{\beta}_1)$  provides the marginal effect for the *discrete* variable  $x_i$ . If  $x_i$  were continuous — or treated as being continuous for simplicity (an approximation that often works well) — we would instead use:

$$M_c(x_i; \hat{\beta}_0, \hat{\beta}_1) = \frac{\partial \Pr(y_i = 1 \mid x_i; \hat{\beta}_0, \hat{\beta}_1)}{\partial x_i} \quad (1.38)$$

$$= \hat{\beta}_1 \cdot \phi(\hat{\beta}_0 + \hat{\beta}_1 \cdot x_i). \quad (1.39)$$

Figure 1.4 shows the predicted probabilities of success for the Tanzanian data — as in Figure 1.3 — but superimposes the estimated marginal effects (where, for simplicity, I have treated year of birth as a continuous variable). You will see the the estimated marginal effect is greatest for individuals born in 1959 — and that this is the year for which the predicted probability of primary attainment is (approximately) 0.5. This is clearly no coincidence: the function  $\phi(x)$  is maximised at  $x = 0$ , and  $\Phi(0) = 0.5$ .

Figure 1.4: Probit estimates and marginal effects for primary school attainment in Tanzania

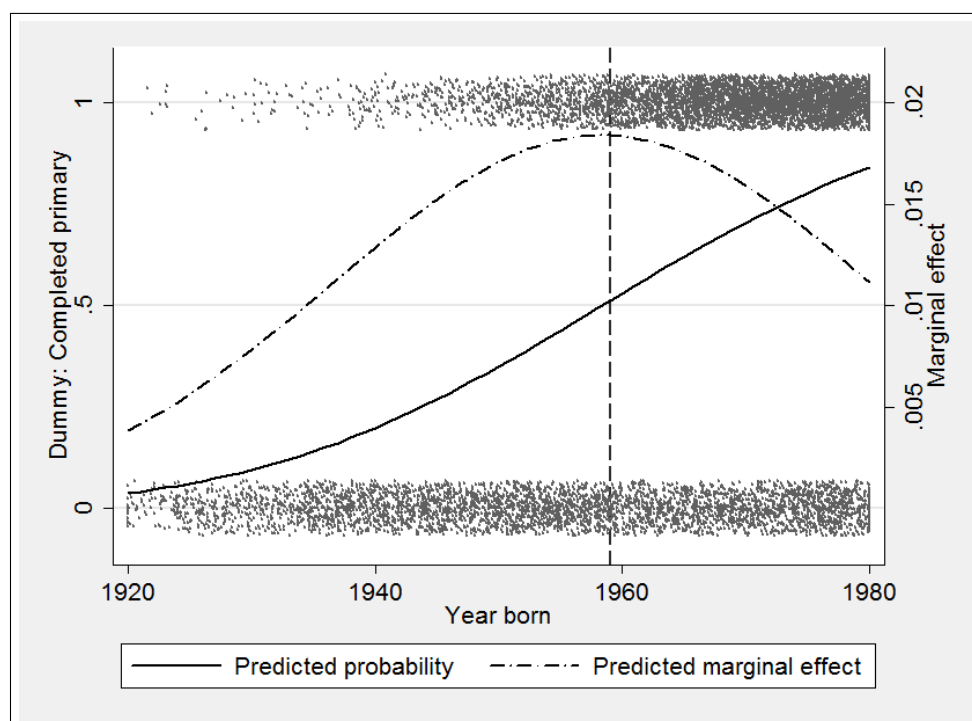


Figure 1.4 shows one way of reporting the marginal effects — *i.e.* by calculating the marginal effects separately for all individuals in the sample, and then graphing. But there are several alternative approaches: for example, statistical packages generally report either the average of the marginal effects across the sample, or the marginal effect at the mean of the regressors.<sup>6</sup> Alternatively, you may wish to take some weighted average of the sample marginal effects, if the sample is unrepresentative of the population of interest. Table 1.1 earlier reported both the mean marginal effect and the marginal effect at the mean.

In short, the marginal effects are particularly important for binary outcome variables, and it is generally a very good idea to report marginal effects alongside estimates of the parameters (or, indeed, instead of them). Standard statistical packages can compute estimated marginal effects straightforwardly — and, similarly, can use the delta method to calculate corresponding standard errors.

<sup>6</sup> Cameron and Trivedi (page 467) prefer the former; they say, “it is best to use... the sample average of the marginal effects. Some programs instead evaluate at the sample average of the regressors...”.

## 1.7 Appendix to Lecture 1: Stata code

First, let's clear Stata's memory and load our dataset.

```
clear
```

```
use WorkingSample
```

We can then tabulate `primaryplus`, the dummy variable that records whether or not a respondent has primary education (or higher):

```
tab primaryplus
```

Similarly, let's summarise the variable `yborn`, which records the year in which each respondent was born:

```
summarize yborn
```




Time to estimate. We begin with our probit model; try `help probit`.

Now let's calculate marginal effects — first as the mean across the sample, and then as the marginal effect at the mean of `yborn`. To do this, try `help margins`.

---

## 2 Lecture 2: Binary Choice II

### Required readings (for Lectures 1 and 2):

-  CAMERON, A.C. AND TRIVEDI, P.K. (2005): *Microeconometrics: Methods and Applications*. Cambridge University Press, pages 463 – 478 (*i.e.* sections 14.1 to 14.4, inclusive)  
*or*
-  WOOLDRIDGE, J. (2002): *Econometric Analysis of Cross Section and Panel Data*. The MIT Press, pages 453 – 461 (*i.e.* sections 15.1 to 15.4, inclusive)  
*or*
-  WOOLDRIDGE, J. (2010): *Econometric Analysis of Cross Section and Panel Data* (2nd ed.). The MIT Press, pages 561 – 569 (*i.e.* sections 15.1 to 15.4, inclusive).

### Other references:

- HARRISON, G. (2011): “Randomisation and Its Discontents,” *Journal of African Economies*, 20(4), 626–652.
- ANGRIST, J.D. AND PISCHKE, J.S. (2008): *Mostly Harmless Econometrics: An Empiricist’s Companion*. Princeton University Press.

### 2.1 The logit model

Lecture 1 considered the probit model: a model of binary choice in which the latent error variable is assumed to have a standard normal distribution. You will recall that, in the context of a single explanatory variable ( $x_i$ ), this model can be summarised succinctly by our earlier equation 1.14:

$$\Pr(y_i = 1 \mid x_i; \beta_0, \beta_1) = \Phi(\beta_0 + \beta_1 x_i). \quad (1.14)$$

An alternative approach is to assume that the latent unobservable has a ‘logistic distribution’:

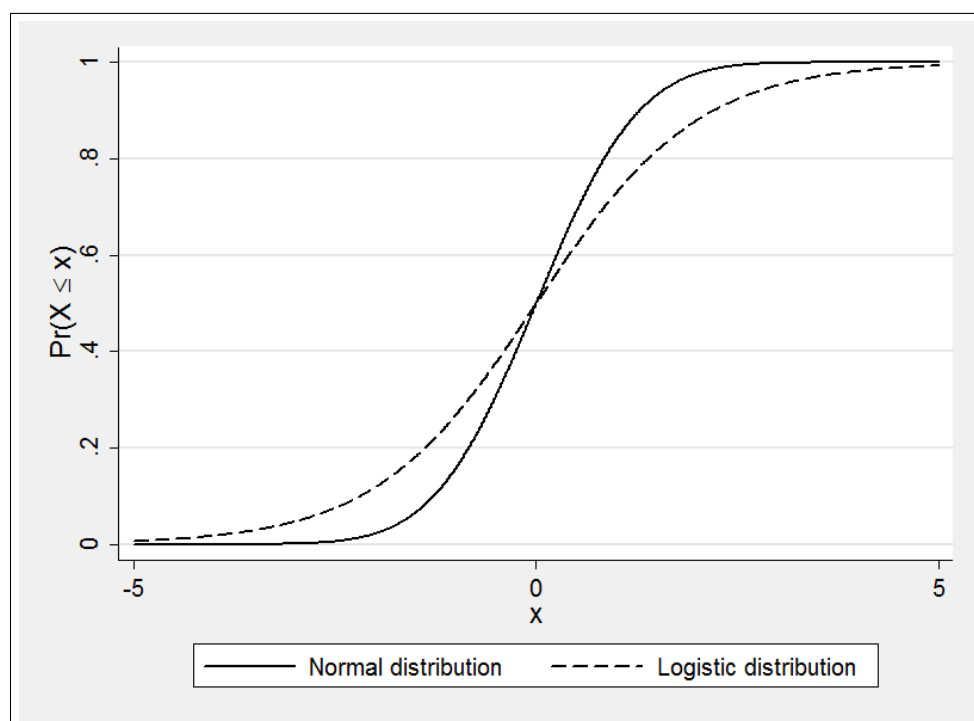
**Assumption 2.1 (DISTRIBUTION OF  $\varepsilon_i$ )**  $\varepsilon$  is *i.i.d.* with a logistic distribution, independent of  $x$ :

$$\Pr(\varepsilon \leq Z \mid x) = \Lambda(Z) \quad (2.1)$$

$$= \frac{\exp(Z)}{1 + \exp(Z)}. \quad (2.2)$$

Figure 2.1 shows the *cdf* of the logistic distribution, compared to the normal.

Figure 2.1: Cumulative density functions: normal and logistic distributions



Symmetric to our derivation of the probit specification, we can write:

$$\Pr(y_i = 1 \mid x_i; \beta_0, \beta_1) = \Pr(\varepsilon_i \leq \beta_0 + \beta_1 x_i \mid x_i) \quad (2.3)$$

$$= \Lambda(\beta_0 + \beta_1 x_i). \quad (2.4)$$

Equation 2.4 is directly analogous to equation 1.14; it defines the ‘*logit model*’. The logit model is an alternative to the probit model for estimating the conditional probability of a binary outcome. For any given dataset, the predicted probabilities from a logit model are generally almost identical to those from a probit model, as we will see later in this lecture.

All of the reasoning from Lecture 1 extends by analogy to the case of the logit model: we can follow the same principles to (i) form the log-likelihood, (ii) maximise the log-likelihood, (iii) measure the goodness-of-fit, (iv) normalise our parameter estimates and (v) interpret the marginal effects. We will not rehearse these principles in this lecture, but you should understand the way that they extend from the probit case to the logit case.

## 2.2 The log-odds ratio in the logit model

The probit model and the logit model are almost identical in their implications. However, when we use the logit model, we sometimes speak about the ‘*odds ratio*’, because this ratio has a natural relationship to the estimated parameters from a logit specification.<sup>7</sup>

Generally, the odds ratio is defined as:

$$\text{odds ratio} = \frac{\text{probability of success}}{\text{probability of failure}}. \quad (2.5)$$

In the context of our Tanzanian problem, and using the logit specification, we can write:

$$\text{odds ratio}_i = \frac{\Pr(y_i = 1 \mid x_i; \beta_0, \beta_1)}{\Pr(y_i = 0 \mid x_i; \beta_0, \beta_1)} \quad (2.6)$$

$$= \frac{\Pr(y_i = 1 \mid x_i; \beta_0, \beta_1)}{1 - \Pr(y_i = 1 \mid x_i; \beta_0, \beta_1)} \quad (2.7)$$

$$= \left( \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \right) \cdot \left( \frac{1}{1 + \exp(\beta_0 + \beta_1 x_i)} \right)^{-1} \quad (2.8)$$

$$= \exp(\beta_0 + \beta_1 x_i). \quad (2.9)$$

In the logit model, we can therefore interpret the index  $\beta_0 + \beta_1 x_i$  as providing the ‘*log odds ratio*’, so that the parameter  $\beta_1$  shows the effect of  $x_i$  on this log ratio:

$$\beta_0 + \beta_1 x_i = \ln(\text{odds ratio}_i) \quad (2.10)$$

$$\beta_1 = \frac{\partial \ln(\text{odds ratio}_i)}{\partial x_i}. \quad (2.11)$$

This implies that, for a small change in  $x_i$ , the value  $100\beta_1 \cdot \Delta x_i$  is approximately the *percentage change in the odds ratio*.

## 2.3 Probit or logit?

Cameron and Trivedi have an extensive discussion of the theoretical and empirical considerations in choosing between the probit or logit model: see pages 471–473. In these notes, I would like simply to emphasise their comment about empirical considerations:

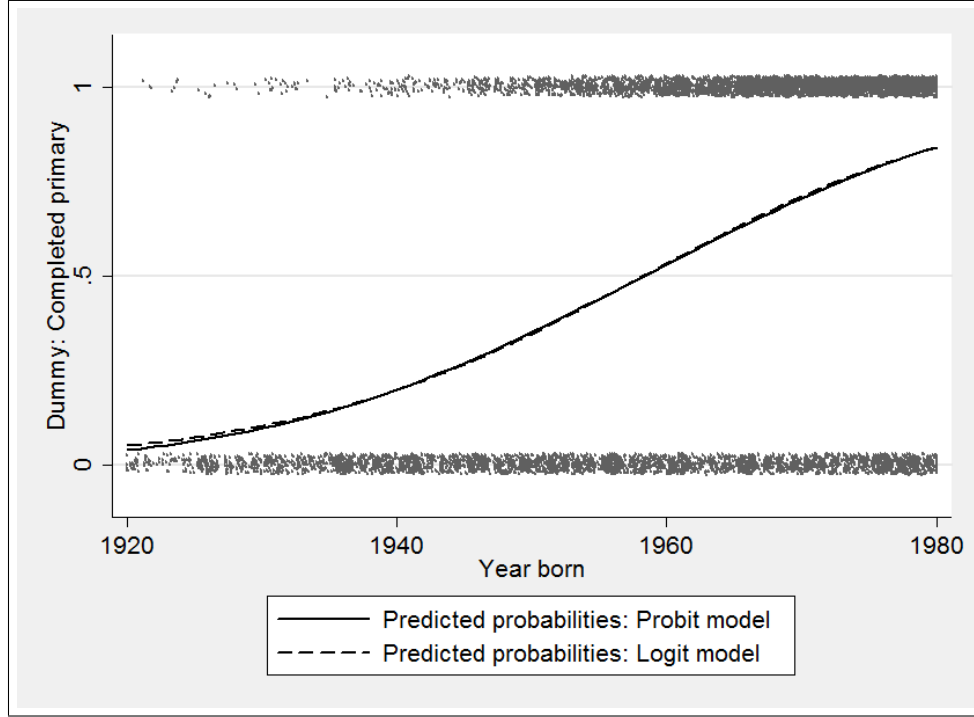
Empirically, either logit and probit can be used. There is often little difference between the predicted probabilities from probit and logit models. The difference is greatest in the tails where probabilities are close to 0 or 1. The difference is much less if interest lies only in marginal effects averaged over the sample rather than for each individual.

Figure 2.2 illustrates this point by comparing estimates from the Tanzanian data.

<sup>7</sup> Of course, this doesn’t mean that we *can’t* talk about the odds ratio when discussing other models — just that the ratio has a more intuitive relationship to the parameters of interest in the logit model.



Figure 2.2: Probit estimates and logit estimates for primary school attainment in Tanzania



## 2.4 The Linear Probability Model

To this point, we have considered two models: probit and logit. We have specified these models in terms of a *conditional probability of success*, but we could equally specify them in terms of the *conditional expectation of the outcome variable*:

$$\mathbb{E}(y_i | x_i; \beta_0, \beta_1) = 1 \times \Pr(y_i = 1 | x_i; \beta_0, \beta_1) + 0 \times \Pr(y_i = 0 | x_i; \beta_0, \beta_1) \quad (2.12)$$

$$= \Pr(y_i = 1 | x_i; \beta_0, \beta_1). \quad (2.13)$$

Thus, for the probit model, we used:

$$\mathbb{E}(y_i | x_i; \beta_0, \beta_1) = \Phi(\beta_0 + \beta_1 x_i); \quad (2.14)$$

for the logit model, we used:

$$\mathbb{E}(y_i | x_i; \beta_0, \beta_1) = \Lambda(\beta_0 + \beta_1 x_i). \quad (2.15)$$

A simpler approach is to assume that the conditional probability of success — and, therefore, the conditional expectation of the outcome — is linear in the explanatory variable(s):

$$\mathbb{E}(y_i | x_i; \beta_0, \beta_1) = \beta_0 + \beta_1 x_i \quad (2.16)$$

$$\Leftrightarrow y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad (2.17)$$

where  $\mathbb{E}(\varepsilon_i | x_i) = 0$ .

This is known as the ‘*Linear Probability Model*’, or ‘LPM’ for short. As equation 2.17 implies, the parameters of interest for the LPM can be obtained very simply: *just use OLS*. We will not rehearse here the principles involved in OLS estimation.

### 2.4.1 Predicted probabilities and marginal effects in the Linear Probability Model

Predicted probabilities in the LPM are trivial:

$$\widehat{\Pr}(y_i = 1 | x_i; \hat{\beta}_0, \hat{\beta}_1) = \hat{\beta}_0 + \hat{\beta}_1 x_i. \quad (2.18)$$

Note that nothing constrains this predicted probability to lie in the unit interval. We will return to this point shortly.

Marginal effects in the LPM are similarly trivial: whether  $x_i$  is discrete or continuous, its estimated marginal effect is  $\hat{\beta}_1$ . Note that this marginal effect is identical across all values of  $x$ .

### 2.4.2 Heteroskedasticity in the Linear Probability Model

The Linear Probability Model generally produces heteroskedastic errors. We can illustrate this straightforwardly using our simple example; for a given  $x_i$ , we have:

$$\varepsilon_i = \begin{cases} 1 - \beta_0 - \beta_1 x_i & \text{with conditional probability } \beta_0 + \beta_1 x_i \\ -\beta_0 - \beta_1 x_i & \text{with conditional probability } 1 - \beta_0 - \beta_1 x_i. \end{cases} \quad (2.19)$$

Figure 2.3 illustrates. We know that  $\text{Var}(\varepsilon_i | x_i) = \mathbb{E}(\varepsilon_i^2 | x_i) - [\mathbb{E}(\varepsilon_i | x_i)]^2$ , and — as we noted earlier —  $\mathbb{E}(\varepsilon_i | x_i) = 0$ . We therefore have:

$$\text{Var}(\varepsilon_i | x_i) = \mathbb{E}(\varepsilon_i^2 | x_i) \quad (2.20)$$

$$= \Pr(y_i = 1 | x_i) \cdot (1 - \beta_0 - \beta_1 x_i)^2 + \Pr(y_i = 0 | x_i) \cdot (-\beta_0 - \beta_1 x_i)^2 \quad (2.21)$$

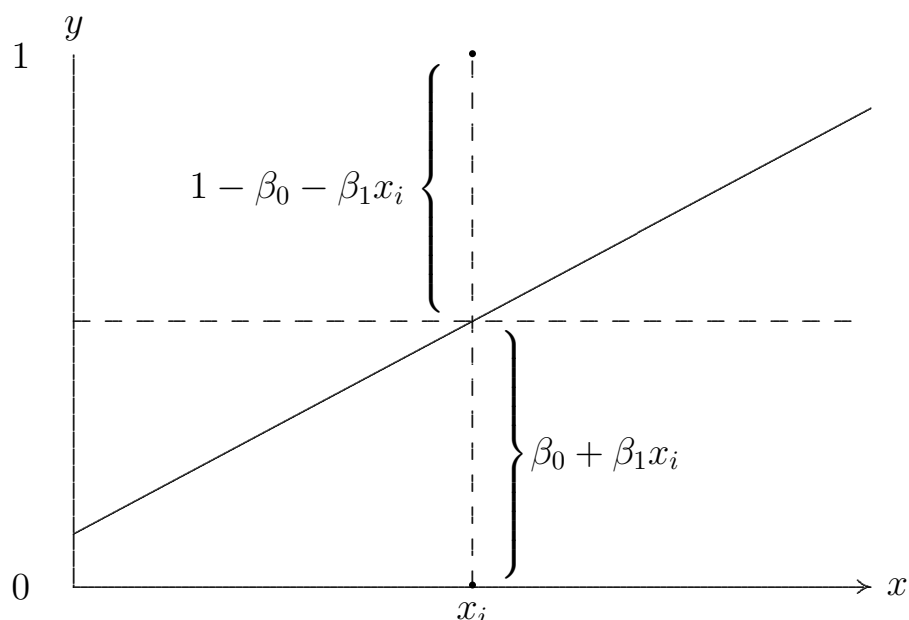
$$= (\beta_0 + \beta_1 x_i) \cdot (1 - \beta_0 - \beta_1 x_i)^2 + (1 - \beta_0 - \beta_1 x_i) \cdot (-\beta_0 - \beta_1 x_i)^2 \quad (2.22)$$

$$= (\beta_0 + \beta_1 x_i) \cdot (1 - \beta_0 - \beta_1 x_i). \quad (2.23)$$

Therefore,  $\text{Var}(\varepsilon_i | x_i)$  depends upon  $x_i$  so long as  $\beta_1 \neq 0$ .<sup>8</sup> The simplest way of dealing with this problem is to use White’s heteroskedasticity-robust standard errors (which can be implemented straightforwardly in Stata using the ‘robust’ option). Alternatively, we could use Weighted Least Squares; this produces more efficient estimates, but requires predicted probabilities to lie between 0 and 1 (which, as we discussed, is not guaranteed in the LPM).

<sup>8</sup> Note, then, that if we are testing a null hypothesis  $H_0 : \beta_1 = 0$  in this model, we do not need to worry about heteroskedasticity.

Figure 2.3: Heteroskedasticity in the Linear Probability Model



## 2.5 LPM or MLE?

### 2.5.1 Relative advantages and disadvantages

We noted earlier that the difference between probit and logit is very small — both in terms of the estimates that they provide and in terms of their underlying structure. The Linear Probability Model, however, is clearly quite different — for example, as Cameron and Trivedi note (page 466), the Linear Probability Model, unlike probit and logit, “does not use a cdf”. So which approach should be preferred — probit/logit on the one hand, or the Linear Probability Model on the other?

This can be quite a controversial issue in applied research! On the one hand, many researchers prefer the probit or logit models, on the basis that they constrain predicted probabilities to the unit interval, and that they therefore imply sensible marginal effects across the entire range of explanatory variables. Cameron and Trivedi, for example, say (page 471):

Although OLS estimation with heteroskedastic standard errors can be a useful exploratory data analysis tool, it is best to use the logit or probit MLE for final data analysis.

In a 2011 article about Randomised Controlled Trials (‘RCT’) in the *Journal of African Economies*, Harrison said this about the use of OLS for limited dependent variable models (footnote omitted):<sup>9</sup>

<sup>9</sup> Harrison also discussed the issue in his presentation at the 2011 CSAE Annual Conference, available at <http://www.csae.ox.ac.uk/conferences/2011-EdiA/video.html>.

One side-effect of the popularity of RCT is the increasing use of Ordinary Least Squares estimators when dependent variables are binary, count or otherwise truncated in some manner. One is tempted to call this the *OLS Gone Wild* reality show, akin to the *Girls Gone Wild* reality TV show, but it is much more sober and demeaning stuff. I have long given up asking researchers in seminars why they do not just report the marginal effects for the right econometric specification. Instead I ask if we should just sack those faculty in the room who seem to waste our time teaching things like logit, count models or hurdle models.

In their book *Mostly Harmless Econometrics*, Angrist and Pischke (2009, page 94) take a different approach:

Should the fact that a dependent variable is limited affect empirical practice? Many econometrics textbooks argue that, while OLS is fine for continuous dependent variables, when the outcome of interest is a limited dependent variable (LDV), linear regression models are inappropriate and nonlinear models such as probit and Tobit are preferred. In contrast, our view of regression as inheriting its legitimacy from the [Conditional Expectation Function] makes LDVness less central.

That is, the Linear Probability Model can still be used to estimate the average marginal effect. Cameron and Trivedi acknowledge (page 471) that:

The OLS estimator [that is, the Linear Probability Model] is nonetheless useful as an exploratory tool. In practice it provides a reasonable direct estimate of the sample-average marginal effect on the probability that  $y = 1$  as  $x$  changes, even though it provides a poor model for individual probabilities. In practice it provides a good guide to which variables are statistically significant.

Further, the Linear Probability Model is sometimes preferred for computational reasons; maximum likelihood models can prove much more difficult to estimate where, for example, there is a very large number of observations or a large number of explanatory variables.

### 2.5.2 Estimates from Tanzania

Table 2.1 reports estimates from the probit, logit and LPM models for the Tanzanian education example; Figure 2.4 shows the predicted probabilities. Together, the table and figure illustrate several important features of the three models. First, all three models predict very similar mean marginal effects. Second, the mean marginal effect for the Linear Probability Model is identical to the parameter estimate.<sup>10</sup> Third, the probit and logit models predict conditional probabilities in the unit interval; in contrast, the LPM implies nonsensical predicted probabilities for people born before about 1925.

<sup>10</sup> For this reason, we would never report the estimate and the marginal effect separately for the LPM; I have done so here simply to emphasise their equivalence.

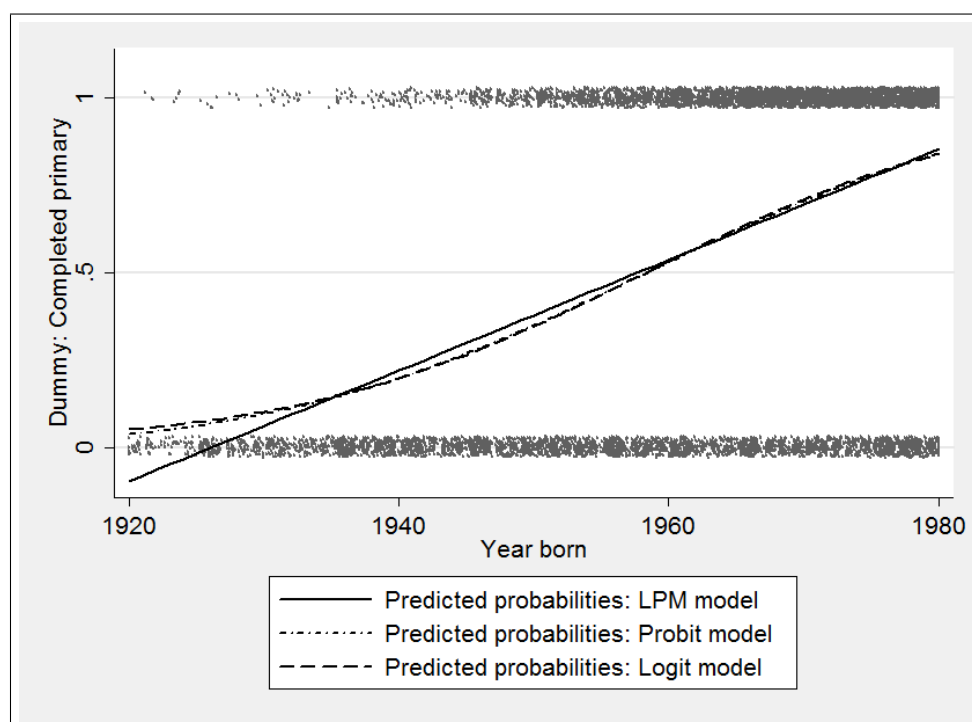
Table 2.1: Probit, logit and LPM results from Tanzania

	Probit		Logit		LPM	
	<i>Estimate</i>	<i>Marginal</i>	<i>Estimate</i>	<i>Marginal</i>	<i>Estimate</i>	<i>Marginal</i>
	(1)	(2)	(3)	(4)	(5)	(6)
Year born	0.046 (0.001)***	0.015 (0.0002)***	0.077 (0.002)***	0.015 (0.0002)***	0.016 (0.0003)***	0.016 (0.0003)***
Const.	-90.395 (2.075)***		-149.997 (3.651)***		-30.510 (0.514)***	
Obs.	10000		10000		10000	
Log-likelihood	-5684.679		-5680.831			
Pseudo- $R^2$	0.165		0.165			
$R^2$					0.210	
<i>Correctly predicted:</i>						
Successes (%)	84.7		84.7		86.3	
Failures (%)	57.2		57.2		55.2	

**Confidence:** \*\*\*  $\leftrightarrow$  99%, \*\*  $\leftrightarrow$  95%, \*  $\leftrightarrow$  90%.

'Marginal' refers to the mean marginal effect. The Linear Probability Model was run using White's heteroskedasticity-robust standard errors.

Figure 2.4: LPM, probit and logit estimates for primary school attainment in Tanzania



## 2.6 The single-index assumption

Albert Einstein once famously declared that “the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience”. (This is often misquoted simply as: “everything should be made as simple as possible, but no simpler”.) In this spirit, we have considered the probit, logit and LPM models solely in the context of a single (scalar) explanatory variable,  $x_i$ . All of the basic principles of these estimators can be understood in this way, so we have not yet considered the multivariate context.

In most empirical applications, however, we have more than one explanatory variable. It is straightforward to take all of our previous reasoning on  $x_i$  and generalise it to a vector  $\mathbf{x}_i$ , by replacing  $\beta_0 + \beta_1 x_i$  with the linear index  $\boldsymbol{\beta} \cdot \mathbf{x}_i$  (where, generally,  $\mathbf{x}_i$  is understood as including an element ‘1’, to allow an intercept). This is how we deal with multiple explanatory variables in the context of the probit, logit and LPM models; thus, in the multivariate case, we specify either:

$$\Pr(y_i = 1 | \mathbf{x}_i) = \Phi(\boldsymbol{\beta} \cdot \mathbf{x}_i) \text{ for probit,} \quad (2.24)$$

$$\text{or } \Pr(y_i = 1 | \mathbf{x}_i) = \Lambda(\boldsymbol{\beta} \cdot \mathbf{x}_i) \text{ for logit,} \quad (2.25)$$

$$\text{or } \Pr(y_i = 1 | \mathbf{x}_i) = \boldsymbol{\beta} \cdot \mathbf{x}_i \text{ for LPM.} \quad (2.26)$$

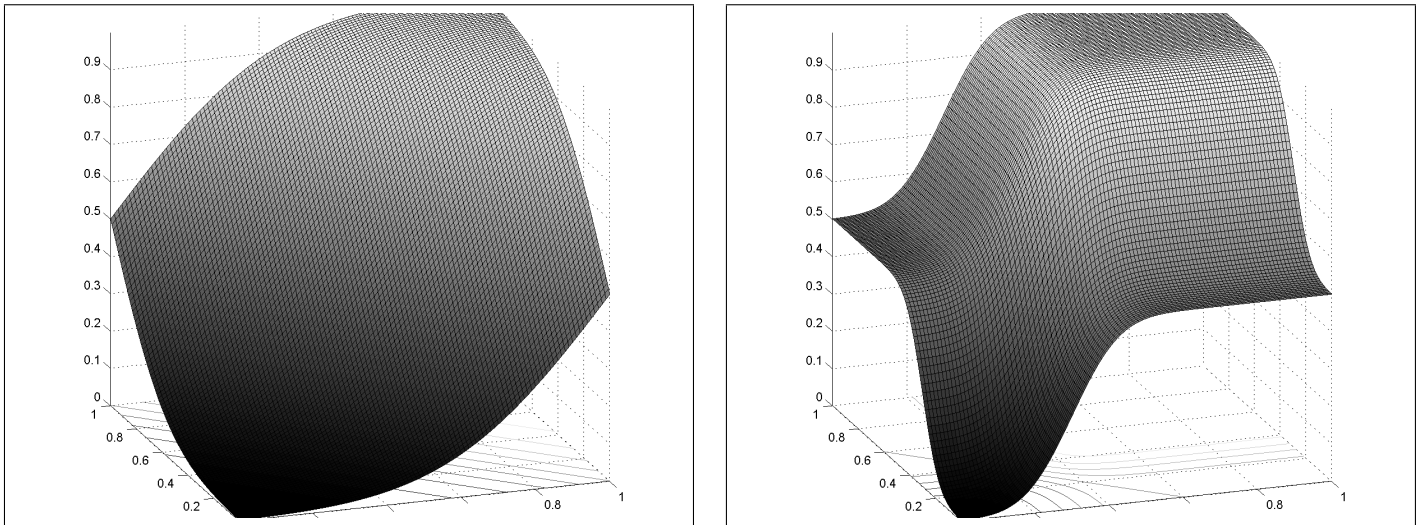
This is a very general structure; it permits quite flexible estimation of binary outcome models with a large number of explanatory variables. However, note that the explanatory variables enter linearly through a ‘single index’,  $\boldsymbol{\beta} \cdot \mathbf{x}_i$ . It is easy to think of functions that violate this assumption. For example, the left surface of Figure 2.5 shows the function  $y = \Phi(3(x_1 + x_2 - 1))$ ; this satisfies the single index assumption because we can write  $y = F(\boldsymbol{\beta} \cdot \mathbf{x})$ . But the right surface in Figure 2.5 shows the function  $y = 0.5(\Phi(10x_1 - 3) + \Phi(10x_2 - 3))$ ; this cannot be expressed as  $y = F(\boldsymbol{\beta} \cdot \mathbf{x})$ , so it violates the single-index assumption.

## 2.7 A general framework

If we are willing to impose the single-index assumption, we can write the probit, logit and LPM models as special cases of a very general structure:

$$\Pr(y_i = 1 | \mathbf{x}_i) = F(\boldsymbol{\beta} \cdot \mathbf{x}_i), \quad (2.27)$$

where  $F(z) = \Phi(z)$  for probit,  $F(z) = \Lambda(z)$  for logit and  $F(z) = z$  for the LPM.  $F$  can be referred to as a ‘link function’. If  $F(z)$  is a *cdf* — as it is, for example, for probit and logit — then we can rewrite all of our earlier maximum likelihood results more generally in terms of  $F(z)$ . This is the way, for example, that Cameron and Trivedi introduce probit and logit (see pages 465 to 469 of their text); Wooldridge takes the same approach (see pages 457 – 458 of his 2002 text).

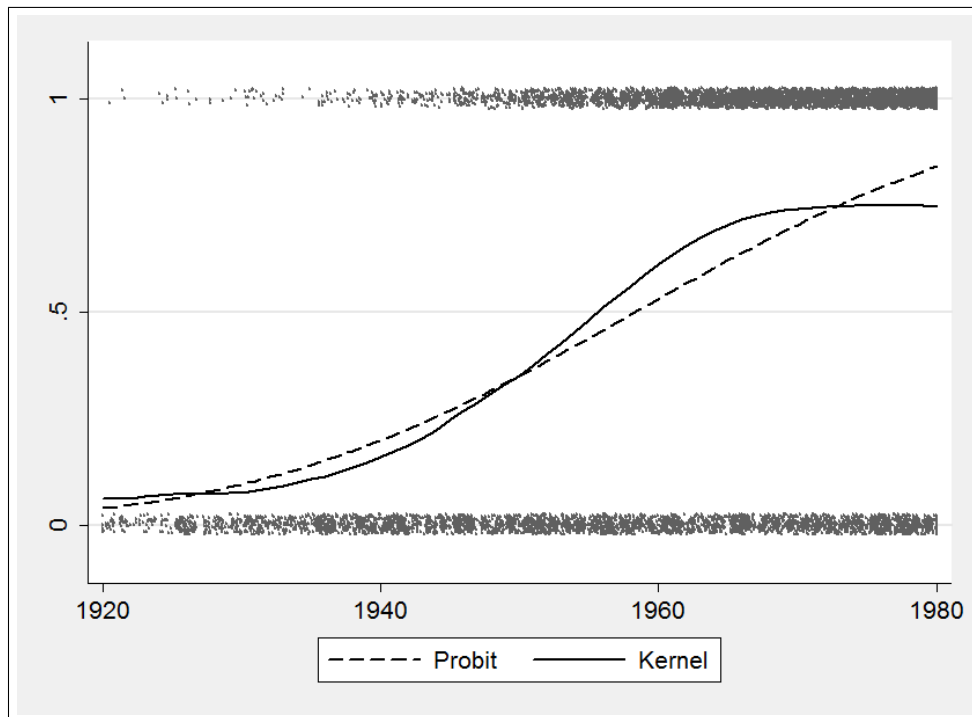
Figure 2.5: The single index restriction: an example (*left*) and a violation (*right*)

*The surface on the left shows the function  $y = \Phi(3(x_1 + x_2 - 1))$ ; the graph on the right shows the function  $y = 0.5(\Phi(10x_1 - 3) + \Phi(10x_2 - 3))$ . Thus, the surface on the left may be expressed as  $y = f(\beta \cdot \mathbf{x})$ , but the surface on the right requires a bivariate function  $y = g(x_1, x_2)$ .*

More generally, equation 2.27 permits semiparametric estimation, in which a researcher can jointly estimate the parameter vector  $\beta$  and the link function  $F$ . In our simple application with one explanatory variable, this kind of approach just implies fitting a univariate nonparametric function,  $y = f(x_i) + \varepsilon_i$ , where  $\mathbb{E}(\varepsilon_i | x_i) = 0$ . Figure 2.6 illustrates, where the function  $F$  is fitted with a kernel.

You do not need to understand any nonparametric or semiparametric methods for these lectures. However, the underlying point is worth remembering: the probit, logit and LPM models all impose particular assumptions on the data, and we can sometimes relax these assumptions using more flexible estimation techniques.

Figure 2.6: Probit and kernel estimates for primary school attainment in Tanzania



## 2.8 Appendix to Lecture 2: Stata code

Let's again clear Stata's memory and load our dataset.

```
clear
```

```
use WorkingSample
```

We can run a logit estimation with the `logit` command ('help logit'). We can then use the same `margins` command as for `probit`.

Finally, we can run the Linear Probability Model using the command `regress`, for an OLS regression. Remember to think about the standard errors!



---

## 3 Lecture 3: Discrete multinomial choice

### Required reading:

- ★ CAMERON, A.C. AND TRIVEDI, P.K. (2005): *Microeconometrics: Methods and Applications*. Cambridge University Press, pages 490 – 506 (*i.e.* sections 15.1 to 15.5.3, inclusive)  
*or*
- ★ WOOLDRIDGE, J. (2002): *Econometric Analysis of Cross Section and Panel Data*. The MIT Press, pages 497 – 502 (*i.e.* section 15.9.1 and part of section 15.9.2)  
*or*
- ★ WOOLDRIDGE, J. (2010): *Econometric Analysis of Cross Section and Panel Data* (2nd ed.). The MIT Press, pages 643 – 649 (*i.e.* sections 16.1, 16.2.1 and part of 16.2.2).

### Other references:

- LEWIS, W.A. (1954): “Economic Development with Unlimited Supplies of Labour,” *The Manchester School*, 22(2), 139–191.
- MCFADDEN, D. (1974): “The Measurement of Urban Travel Demand,” *Journal of Public Economics*, 3(4), 303–328.
- MCFADDEN, D. (2000): “Economic Choices”, Nobel Prize Lecture, 8 December 2000.

### 3.1 Occupational choice in Tanzania

Travel demand forecasting has long been the province of transportation engineers, who have built up over the years considerable empirical wisdom and a repertory of largely ad hoc models which have proved successful in various applications. . . [but] there still does not exist a solid foundation in behavioral theory for demand forecasting practices. Because travel behavior is complex and multifaceted, and involves ‘non-marginal’ choices, the task of bringing economic consumer theory to bear is a challenging one. *Particularly difficult is the integration of a satisfactory behavioural theory with practical statistical procedures for calibration and forecasting.*

McFadden (1974, emphasis added)

The main sources from which workers come as economic development proceeds are subsistence agriculture, casual labour, petty trade, domestic service, wives and daughters in the household, and the increase of population. In most but not all of these sectors, if the country is overpopulated relatively to its natural resources, the marginal productivity of labour is negligible, zero, or even negative.

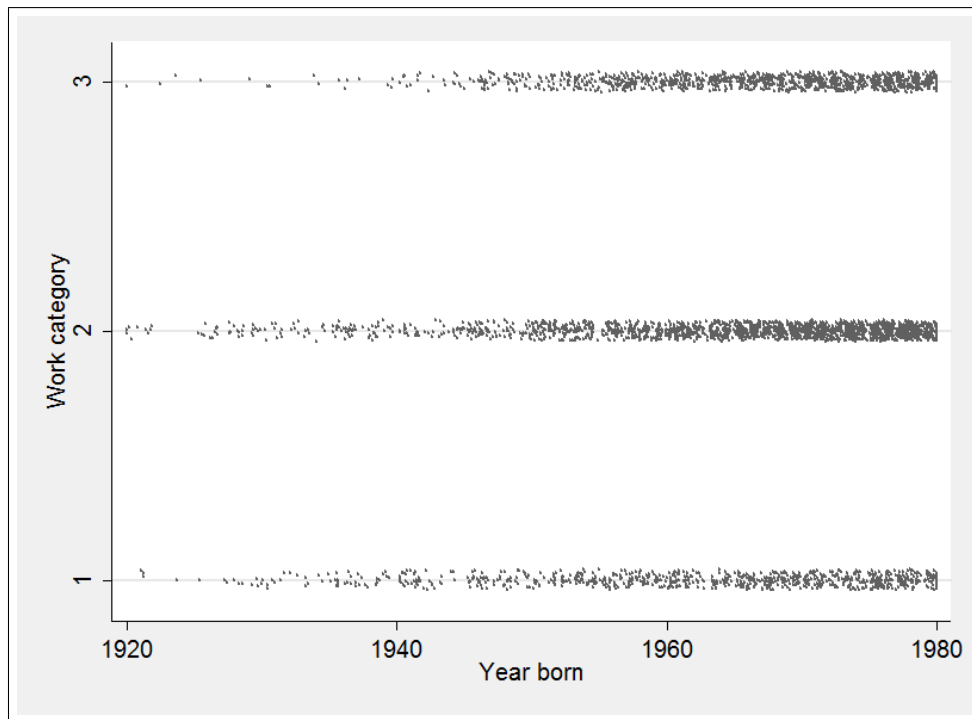
Lewis (1954)

In this lecture, we consider occupational choice in Tanzania. Both geographically and conceptually, the Tanzanian labour market is a long way from the San Francisco Bay Area Rapid Transit system (the ‘BART’). Nonetheless, we will analyse occupational choice using some of the econometric methods that Daniel McFadden famously developed to predict demand for the new BART. Like McFadden, our concern shall be to estimate the conditional probability of various *discrete and unordered choices*, and to do so with — as McFadden termed it — “a solid foundation in behavioral theory”.

Let’s begin in Tanzania. Figure 3.1 describes occupational choice among employed Tanzanians of different ages. The figure — and our subsequent analysis — uses a ternary outcome variable, covering three mutually exclusive categories:

$$y_i = \begin{cases} 1 & \text{if the } i\text{th respondent works in \textbf{agriculture};} \\ 2 & \text{if the } i\text{th respondent is \textbf{self-employed} (outside of agriculture);} \\ 3 & \text{if the } i\text{th respondent is \textbf{wage employed} (outside of agriculture).} \end{cases} \quad (3.1)$$

Figure 3.1: **Occupational categories and age in Tanzania**



We will not worry too much today about why occupational choice in Tanzania might matter; as in earlier lectures, we will use the Tanzanian data as an illustrative vehicle for our econometric techniques. However, it is not difficult to see why this kind of occupational choice might be important for understanding Tanzania’s development; the quote from Lewis’s famous 1954 paper,

for example, highlights sectoral shifts as an important mechanism for long-run development, and the data in Figure 3.1 might provide insights into the flexibility with which workers can achieve such shifts. (For example, if older workers are more likely to choose employment in agriculture than younger workers, this may suggest some ‘switching costs’ between sectors.)

## 3.2 An Additive Random Utility Model

As in earlier lectures, we will motivate our econometric methods by a simple underlying microeconomic model. Typically, this kind of choice-theoretic foundation is more common in the analysis of discrete unordered choice than in the models that we have studied earlier. For example, the latent variable interpretation is a useful approach for thinking about the probit and logit models, but is not generally a starting point for analysis; similarly, an optimal stopping model is just one possible foundation for models of discrete ordered choice. But in the analysis of discrete unordered choice, an additive random utility model is a common starting point. As Cameron and Trivedi (page 506) explain:

The econometrics literature has placed great emphasis in restricting attention to multinomial models that are consistent with maximisation of a random utility function. This is similar to restricting analysis to demand functions that are consistent with consumer choice theory.

Suppose, therefore, that we again have data on  $N$  individuals, indexed  $i \in \{1, \dots, N\}$ . Assume that each individual makes a choice  $y_i = j$ , where there are a finite number  $J$  options available. Critically, suppose that we observe information at the level of each *individual*, including his/her choices (that is, we observe  $x_i$  and  $y_i$ ). That is, we do *not* observe information at the level of each *available option*; we will consider this alternative kind of data structure later in this lecture. You may query how reasonable it may be to model occupational outcomes purely as a matter of choice — after all, could an agricultural employee simply *choose* to take a wage job? — but we will leave this concern aside for this lecture.

As in Lecture 1, we will assume that the  $i$ th individual’s utility from the  $j$ th choice is determined by an additive random utility model:

$$U_{ij}(x_i) = \alpha_0^{(j)} + \alpha_1^{(j)}x_i + \varepsilon_{ij}. \quad (3.2)$$

Thus, for example, for choices  $j \in \{1, 2, 3\}$ , the individual obtains the following utilities:

$$U_{i1}(x_i) = \alpha_0^{(1)} + \alpha_1^{(1)}x_i + \varepsilon_{i1} \quad (3.3)$$

$$U_{i2}(x_i) = \alpha_0^{(2)} + \alpha_1^{(2)}x_i + \varepsilon_{i2} \quad (3.4)$$

$$U_{i3}(x_i) = \alpha_0^{(3)} + \alpha_1^{(3)}x_i + \varepsilon_{i3}. \quad (3.5)$$

Together, these three utilities determine the choice of an optimising agent. Figure 3.2 illustrates preferences between the three options in two-dimensional space; in each box, the bold outcome represents the agent’s choice.

Figure 3.2: **Multinomial choice among three options**

We can, therefore, express the conditional probability of the  $i$ th individual choosing, say, option 1:

$$\Pr(y_i = 1 \mid x_i) = \Pr \left[ U^{(1)}(x_i) > U^{(2)}(x_i) \text{ and } U^{(1)}(x_i) > U^{(3)}(x_i) \mid x_i \right] \quad (3.6)$$

$$= \Pr \left[ \alpha_0^{(1)} + \alpha_1^{(1)}x_i + \varepsilon_{i1} > \alpha_0^{(2)} + \alpha_1^{(2)}x_i + \varepsilon_{i2} \text{ and } \right. \\ \left. \alpha_0^{(1)} + \alpha_1^{(1)}x_i + \varepsilon_{i1} > \alpha_0^{(3)} + \alpha_1^{(3)}x_i + \varepsilon_{i3} \mid x_i \right]. \quad (3.7)$$

More generally, if the  $i$ th individual were to choose  $y_i = j$  out of  $J$  choices, we could write:

$$\Pr(y_i = j \mid x_i) = \Pr \left[ U^{(j)}(x_i) > \max_{k \neq j} (U^{(k)}(x_i)) \mid x_i \right] \quad (3.8)$$

$$= \Pr \left[ \alpha_0^j + \alpha_1^j x_i + \varepsilon_{ij} > \max_{k \neq j} (\alpha_0^k + \alpha_1^k x_i + \varepsilon_{ik}) \mid x_i \right]. \quad (3.9)$$

In order to estimate using equation 3.9, we again need to make a distributional assumption.

### 3.3 The Multinomial Logit model

**Assumption 3.1 (DISTRIBUTION OF  $\varepsilon_{ij}$ )**  $\varepsilon_{ij}$  is i.i.d. with a Type I Extreme Value distribution, independent of  $x_i$ :

$$\Pr(\varepsilon_{ij} < z | x_i) = \Pr(\varepsilon_{ij} < z) = \exp(-\exp(-z)). \quad (3.10)$$

Equation 3.10, of course, defines the *cumulative density function*  $F(z)$ ; this implies a *probability density function* of:

$$f(z) = \frac{d}{dz} \exp(-\exp(-z)) = \exp(-z) \cdot \exp(-\exp(-z)) \quad (3.11)$$

$$= \exp(-z) \cdot F(z). \quad (3.12)$$

Figure 3.3 shows the cumulative density function for the Type I Extreme Value function, compared to the *cdf* of the normal.

Figure 3.3: Cumulative density functions: Normal and Type I Extreme Value distributions

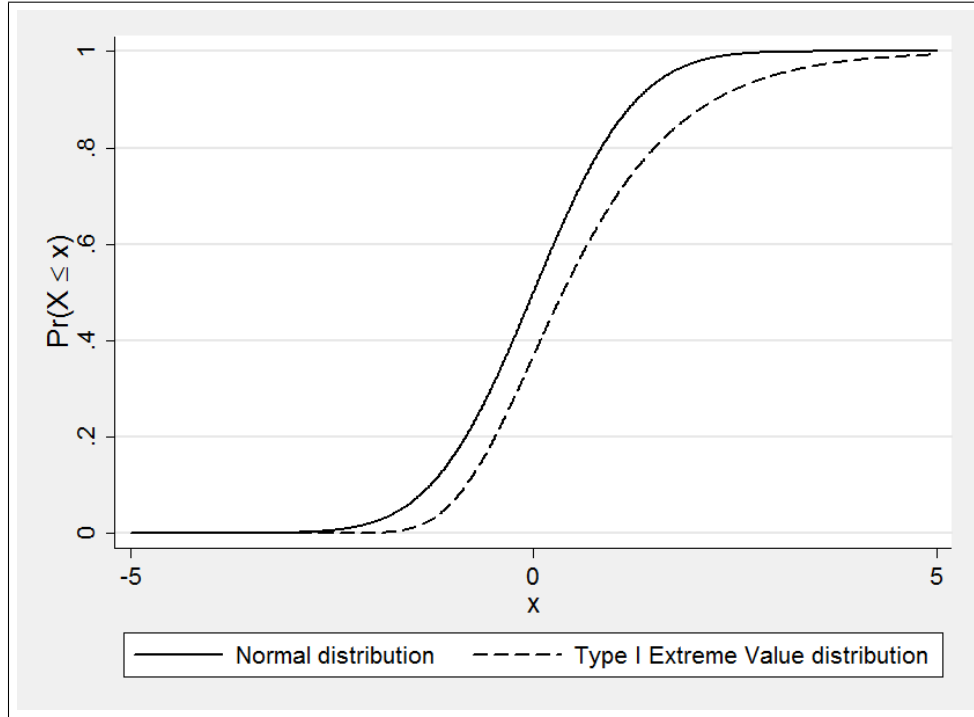
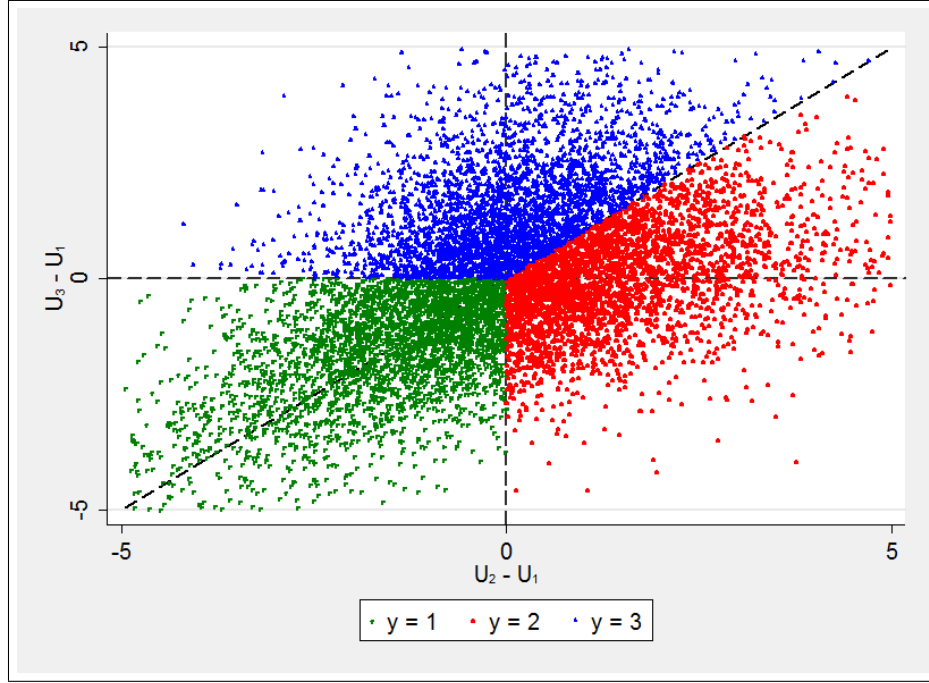


Figure 3.4 shows how this distributional assumption might imply the different outcomes  $y_i = 1$ ,  $y_i = 2$  and  $y_i = 3$ ; the figure shows the same two-dimensional space as Figure 3.2, but with simulated values for  $U_{i1}$ ,  $U_{i2}$  and  $U_{i3}$ . (For simplicity, I have generated the graph by setting all of the parameters  $\alpha_0^{(j)}$  and  $\alpha_1^{(j)}$  to zero; that is, the graph shows variation generated solely by the Type I Extreme Value distribution on  $\varepsilon_{ij}$ .)

Figure 3.4: **Multinomial choice among three options: Simulated data**

With this distributional assumption, we can now find an expression for the conditional probability that the  $i$ th individual chooses outcome  $j$  from  $J$  choices.<sup>11</sup> *Note that you do not need to memorise this derivation for the exam; however, I think the derivation is useful for understanding the underlying structure required by multinomial choice models.*

First, suppose that the error term for the chosen option,  $\varepsilon_{ij}$ , were known. Then we could write:

$$\Pr(y_i = j | x_i, \varepsilon_{ij}) = \Pr \left[ U^{(j)}(x_i) > \max_{k \neq j} (U^{(k)}(x_i)) \mid x_i, \varepsilon_{ij} \right] \quad (3.13)$$

$$= \Pr [\alpha_0^k + \alpha_1^k x_i + \varepsilon_{ik} < \alpha_0^j + \alpha_1^j x_i + \varepsilon_{ij} \mid x_i, \varepsilon_{ij} \forall k \neq j] \quad (3.14)$$

$$= \Pr [\varepsilon_{ik} < \varepsilon_{ij} + \alpha_0^j + \alpha_1^j x_i - \alpha_0^k - \alpha_1^k x_i \mid x_i, \varepsilon_{ij} \forall k \neq j] \quad (3.15)$$

$$= \prod_{k \neq j} \exp \left\{ - \exp \left[ - (\varepsilon_{ij} + \alpha_0^j + \alpha_1^j x_i - \alpha_0^k - \alpha_1^k x_i) \right] \right\}. \quad (3.16)$$

<sup>11</sup> This derivation is taken from Train (2009, pages 74-75).

Of course,  $\varepsilon_{ij}$  is not known; we therefore need to integrate across its possible values:

$$\Pr(y_i = j | x_i) = \int_{-\infty}^{\infty} f(\varepsilon_{ij}) \cdot \Pr(y_i = j | x_i, \varepsilon_{ij}) d\varepsilon_{ij} \quad (3.17)$$

$$= \int_{-\infty}^{\infty} \exp(-\varepsilon_{ij}) \cdot \exp[-\exp(-\varepsilon_{ij})] \cdot \prod_{k \neq j} \exp\{-\exp[-(\varepsilon_{ij} + \alpha_0^j + \alpha_1^j x_i - \alpha_0^k - \alpha_1^k x_i)]\} d\varepsilon_{ij} \quad (3.18)$$

$$= \int_{-\infty}^{\infty} \exp(-\varepsilon_{ij}) \cdot \prod_k \exp\{-\exp[-(\varepsilon_{ij} + \alpha_0^j + \alpha_1^j x_i - \alpha_0^k - \alpha_1^k x_i)]\} d\varepsilon_{ij} \quad (3.19)$$

$$= \int_{-\infty}^{\infty} \exp(-\varepsilon_{ij}) \cdot \exp\left\{-\sum_k \exp[-(\varepsilon_{ij} + \alpha_0^j + \alpha_1^j x_i - \alpha_0^k - \alpha_1^k x_i)]\right\} d\varepsilon_{ij} \quad (3.20)$$

$$= \int_{-\infty}^{\infty} \exp(-\varepsilon_{ij}) \cdot \exp\left\{-\exp(-\varepsilon_{ij}) \cdot \sum_k \exp[-(\alpha_0^j + \alpha_1^j x_i - \alpha_0^k - \alpha_1^k x_i)]\right\} d\varepsilon_{ij}. \quad (3.21)$$

We can now integrate by substitution. Define  $t = \exp(-\varepsilon_{ij})$ , so that  $dt = -\exp(-\varepsilon_{ij}) \cdot d\varepsilon_{ij}$ . Note that  $\lim_{\varepsilon_{ij} \rightarrow -\infty} t = \infty$  and  $\lim_{\varepsilon_{ij} \rightarrow \infty} t = 0$ . Then we can rewrite our integral as:

$$\Pr(y_i = j | x_i) = \int_0^{\infty} \exp\left\{-t \cdot \sum_k \exp[-(\alpha_0^j + \alpha_1^j x_i - \alpha_0^k - \alpha_1^k x_i)]\right\} dt \quad (3.22)$$

$$= \left[ \frac{\exp\{-t \cdot \sum_k \exp[-(\alpha_0^j + \alpha_1^j x_i - \alpha_0^k - \alpha_1^k x_i)]\}}{-\sum_k \exp[-(\alpha_0^j + \alpha_1^j x_i - \alpha_0^k - \alpha_1^k x_i)]} \right]_0^{\infty} \quad (3.23)$$

$$= \frac{1}{\sum_k \exp[-(\alpha_0^j + \alpha_1^j x_i - \alpha_0^k - \alpha_1^k x_i)]} \quad (3.24)$$

$$= \frac{\exp[\alpha_0^j + \alpha_1^j x_i]}{\sum_k \exp[\alpha_0^k + \alpha_1^k x_i]}. \quad (3.25)$$

Let's return to our example with three outcomes,  $y \in \{1, 2, 3\}$ . The last derivation implies that we

can write the conditional probabilities of the outcomes as:

$$\Pr(y_i = 1 | x_i) = \frac{\exp(\alpha_0^{(1)} + \alpha_1^{(1)}x_i)}{\exp(\alpha_0^{(1)} + \alpha_1^{(1)}x_i) + \exp(\alpha_0^{(2)} + \alpha_1^{(2)}x_i) + \exp(\alpha_0^{(3)} + \alpha_1^{(3)}x_i)}; \quad (3.26)$$

$$\Pr(y_i = 2 | x_i) = \frac{\exp(\alpha_0^{(2)} + \alpha_1^{(2)}x_i)}{\exp(\alpha_0^{(1)} + \alpha_1^{(1)}x_i) + \exp(\alpha_0^{(2)} + \alpha_1^{(2)}x_i) + \exp(\alpha_0^{(3)} + \alpha_1^{(3)}x_i)}; \quad (3.27)$$

$$\Pr(y_i = 3 | x_i) = \frac{\exp(\alpha_0^{(3)} + \alpha_1^{(3)}x_i)}{\exp(\alpha_0^{(1)} + \alpha_1^{(1)}x_i) + \exp(\alpha_0^{(2)} + \alpha_1^{(2)}x_i) + \exp(\alpha_0^{(3)} + \alpha_1^{(3)}x_i)}. \quad (3.28)$$

It would be tempting to take these three conditional probabilities and write the log-likelihood; for our three outcomes, we could therefore try to maximise the log-likelihood across six unknown parameters (that is, the parameters  $\alpha_0^{(1)}$ ,  $\alpha_0^{(2)}$ ,  $\alpha_0^{(3)}$ ,  $\alpha_1^{(1)}$ ,  $\alpha_1^{(2)}$  and  $\alpha_1^{(3)}$ ). However, this would be a mistake, because we would be unable to find a unique set of values that would maximise the function. (That is, the model would be *underidentified*.) The reason, of course, is that we can only ever express utility in *relative* terms: we have seen this principle already in deriving both the probit and the Ordered Probit models. We therefore need to choose a ‘base category’, and estimate relative to the utility from that category. We shall choose 1 as the base category, and define  $\beta_0^{(2)} \equiv \alpha_0^{(2)} - \alpha_0^{(1)}$  and  $\beta_1^{(2)} \equiv \alpha_1^{(2)} - \alpha_1^{(1)}$  (and symmetrically for  $\beta_0^{(3)}$  and  $\beta_1^{(3)}$ ). Note the emphasis here that the choice of base category is *arbitrary*: our predicted probabilities would be identical if we were to choose a different base category. Then we can multiply numerator and denominator of each conditional probability by  $\exp(-\alpha_0^{(1)} - \alpha_1^{(1)}x_i)$ , to obtain:

$$\Pr(y_i = 1 | x_i) = \frac{1}{1 + \exp(\beta_0^{(2)} + \beta_1^{(2)}x_i) + \exp(\beta_0^{(3)} + \beta_1^{(3)}x_i)}; \quad (3.29)$$

$$\Pr(y_i = 2 | x_i) = \frac{\exp(\beta_0^{(2)} + \beta_1^{(2)}x_i)}{1 + \exp(\beta_0^{(2)} + \beta_1^{(2)}x_i) + \exp(\beta_0^{(3)} + \beta_1^{(3)}x_i)}; \quad (3.30)$$

$$\Pr(y_i = 3 | x_i) = \frac{\exp(\beta_0^{(3)} + \beta_1^{(3)}x_i)}{1 + \exp(\beta_0^{(2)} + \beta_1^{(2)}x_i) + \exp(\beta_0^{(3)} + \beta_1^{(3)}x_i)}. \quad (3.31)$$

These conditional probabilities can then be used to define the log-likelihood; it is now straightfor-



ward that, for the  $i$ th individual, the log-likelihood is:<sup>12</sup>

$$\begin{aligned} \ell_i \left( \beta_0^{(1)}, \beta_1^{(1)}, \beta_0^{(2)}, \beta_1^{(2)}; y_i | x_i \right) \\ = \mathbf{1}(y_i = 1) \cdot \ln [\Pr(y_i = 1 | x_i)] + \mathbf{1}(y_i = 2) \cdot \ln [\Pr(y_i = 2 | x_i)] \\ + \mathbf{1}(y_i = 3) \cdot \ln [\Pr(y_i = 3 | x_i)]. \end{aligned} \quad (3.32)$$

This log-likelihood function defines the ‘*Multinomial Logit*’ model. The Multinomial Logit is the simplest model for unordered choice. (Note that, if  $J = 2$ , the Multinomial Logit collapses to the logit model that we considered in Lecture 2.) Marginal effects in the Multinomial Logit model are directly analogous to marginal effects in the earlier models.

### 3.4 Estimates from Tanzania

Table 3.1 shows the estimates from the Tanzanian data. Note that, given the foundation of our additive random utility model, we can express the estimates in terms of ‘relative utility’ from self-employment and wage employment; we estimate  $\hat{\beta}_0^{(1)} = -36.383$ ,  $\hat{\beta}_1^{(1)} = 0.019$ ,  $\hat{\beta}_0^{(2)} = -48.562$  and  $\hat{\beta}_1^{(2)} = 0.025$ . All four estimates are highly significant. Figure 3.5 shows the consequent predicted probabilities of all three work categories (conditional upon having employment); this shows that older employed Tanzanians are significantly more likely to be working in agriculture than are younger employed Tanzanians, and that the converse applies for wage employment and self-employment.

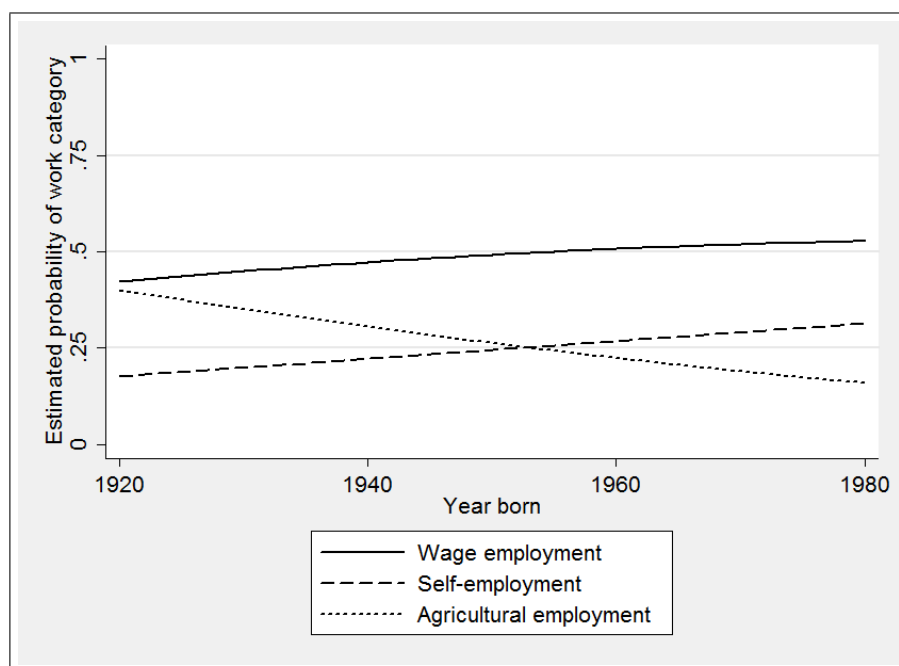
<sup>12</sup> I denote the indicator function by  $\mathbf{1}(\cdot)$ . Note that, for simplicity, I have not explicitly written each conditional probability as depending upon the parameters of interest, though they clearly do.

Table 3.1: Estimates from Tanzania: Multinomial Logit

	Estimates	Mean Marginal Effects	
	(1)	$y = 1$ (2)	$y = 2$ (3)
Year born		0.001 (0.0007)***	0.002 (0.0006)***
<i>Relative utility of self-employment:</i>			
Year born	0.019 (0.003)***		
Const.	-36.383 (6.308)***		
<i>Relative utility of wage employment:</i>			
Year born	0.025 (0.004)***		
Const.	-48.562 (7.289)***		
Obs.	4136		
Log-likelihood	-4218.448		
Pseudo- $R^2$	0.006		

Confidence: \*\*\*  $\leftrightarrow$  99%, \*\*  $\leftrightarrow$  95%, \*  $\leftrightarrow$  90%.

Figure 3.5: Occupational categories and age in Tanzania: Multinomial Logit estimates



### 3.5 From Multinomial Logit to Conditional Logit

We assumed earlier that the  $i$ th individual's utility from the  $j$ th choice depends linearly upon (i) the *observable* characteristics of the *individual*,  $x_i$  and (ii) *unobservable* characteristics of the *individual's taste for that choice*,  $\varepsilon_{ij}$ :

$$U_{ij}(x_i) = \alpha_0^{(j)} + \alpha_1^{(j)} x_i + \varepsilon_{ij}. \quad (3.2)$$

Critically, this structure does not allow for *observable characteristics of different options*. However, we can allow straightforwardly for that possibility, by allowing the variable  $x$  to be indexed by both individual and potential choice:  $x_{ij}$ . That is, we now assume that the researcher observes information on characteristics of options that the  $i$ th individual did not choose — for example, a researcher might know what the  $i$ th individual would have paid to take the train, even though she actually chose the bus.<sup>13</sup> We can therefore write  $U_{ij}$  as a linear function of characteristics — and, for the sake of generality, we now use vector notation to allow for multiple characteristics:<sup>14</sup>

$$U_{ij}(\mathbf{x}_{ij}) = \boldsymbol{\alpha} \cdot \mathbf{x}_{ij} + \varepsilon_{ij}. \quad (3.33)$$

Note that  $\mathbf{x}_{ij}$  is indexed at the level of the option-individual combination; the vector includes characteristics that vary at the level of the individual and the choice — for example, respondents may face different relative costs of using different types of transport. Wooldridge cautions (2002, page 500) that “ $\mathbf{x}_{ij}$  cannot contain elements that vary only across  $i$  and not  $j$ ; in particular,  $\mathbf{x}_{ij}$  does not contain unity”. This means, therefore, that  $\mathbf{x}_{ij}$  cannot include personal characteristics (for example, age, sex, income, *etc*); we will see an intuitive reason for this shortly (in equation 3.39).

In the Multinomial Logit, the outcome  $y$  was indexed by individuals, and took the value of the particular choice made; for example, we might write  $y_i = j$ . But in the Conditional Logit model, we need to express the outcome at the level of the individual-choice combination. Therefore, we redefine our outcome as a binary variable:

$$y_{ij} = \begin{cases} 1 & \text{if the } i\text{th individual chooses option } j, \text{ and;} \\ 0 & \text{if the } i\text{th individual chooses some other option } k \neq j. \end{cases} \quad (3.34)$$

All of our reasoning from the Multinomial Logit extends to the Conditional Logit. The conditional probability of individual  $i$  choosing outcome  $j$  is:

$$\Pr(y_{ij} = 1 \mid \mathbf{x}_{i1}, \dots, \mathbf{x}_{iJ}) = \frac{\exp(\boldsymbol{\beta} \cdot \mathbf{x}_{ij})}{\sum_{k=1}^J \exp(\boldsymbol{\beta} \cdot \mathbf{x}_{ik})}. \quad (3.35)$$

<sup>13</sup> Cameron and Trivedi provide an example of this kind of data structure in their Table 15.1 on page 492; in that application, a researcher observes the price that different respondents faced for each of four types of fishing, even though each respondent chose only one type.

<sup>14</sup> Of course, we could also have used vector notation for all of our earlier reasoning in the Multinomial Logit model; this would have implied estimating two vectors  $\beta_1^{(1)}$  and  $\beta_1^{(2)}$ . But the scalar case captured all of the important aspects of Multinomial Logit in a simpler context, and matched directly our illustrative empirical application.

The log-likelihood follows straightforwardly:

$$\ell_i(\beta; y_{i1}, \dots, y_{iJ} | \mathbf{x}_{i1}, \dots, \mathbf{x}_{iJ}) = \sum_{j=1}^J y_{ij} \cdot \ln \Pr(y_{ij} = 1 | \mathbf{x}_{i1}, \dots, \mathbf{x}_{iJ}). \quad (3.36)$$

### 3.6 Independence assumptions

The Multinomial Logit and Conditional Logit are very tractable models. As we have discussed, they provide an analytical expression for the log-likelihood; this function can therefore be evaluated and maximised easily. But this analytical tractability comes at a cost: the Multinomial Logit and Conditional Logit both require that the unobservable terms,  $\varepsilon_{ij}$ , have a Type I Extreme Value distribution, *and* that these terms are distributed independently of each other. This has serious implications for a structure of individual choice. Consider, for example, the Multinomial Logit. Using equation 3.25, we can write the ratio of the conditional probability that  $y_i = A$  and that  $y_i = B$ :

$$\frac{\Pr(y_i = A | x_i)}{\Pr(y_i = B | x_i)} = \frac{\exp[\alpha_0^{(A)} + \alpha_1^{(A)} x_i]}{\exp[\alpha_0^{(B)} + \alpha_1^{(B)} x_i]} \quad (3.37)$$

$$= \exp\left[\alpha_0^{(A)} - \alpha_0^{(B)} + \left(\alpha_1^{(A)} - \alpha_1^{(B)}\right) \cdot x_i\right]. \quad (3.38)$$

That is, the ratio of probabilities for any two alternatives *cannot* depend upon how much the respondents like any of the *other* alternatives on offer. Similarly, consider the Conditional Logit. Using equation 3.35, the ratio of conditional probabilities for two choices is:

$$\frac{\Pr(y_{iA} = 1 | \mathbf{x}_{i1}, \dots, \mathbf{x}_{iJ})}{\Pr(y_{iB} = 1 | \mathbf{x}_{i1}, \dots, \mathbf{x}_{iJ})} = \exp[\beta \cdot (\mathbf{x}_{iA} - \mathbf{x}_{iB})]. \quad (3.39)$$

Thus, in the Conditional Logit model, the ratio of probabilities for two alternatives cannot depend upon the *characteristics* of any other alternative (or, as noted, on any characteristics that do not vary across  $j$ ).

Cameron and Trivedi (page 503, emphasis in original) describe why these results are so concerning:

A limitation of the [Conditional Logit] and [Multinomial Logit] models is that discrimination among the [ $J$ ] alternatives reduces to a series of pairwise comparisons that are unaffected by the characteristics of alternatives other than the pair under consideration...

As an extreme example, the conditional probability of commute by car given commute by car or red bus is assumed in an MNL or CL model to be independent of whether commuting by blue bus is an option. However, in practice we would expect introduction of a blue bus, which is the same as a red bus in every aspect except colour, to have

little impact on car use and to halve use of the red bus, leading to an increase in the conditional probability of car use given commute by car or red bus.

This weakness of MNL is known in the literature as the red bus – blue bus problem, or more formally as the assumption of [Luce] **independence of irrelevant alternatives**.

This is clearly a serious limitation of the conditional logit and multinomial logit. Indeed, in his Nobel Prize Lecture in 2000, Daniel McFadden even went so far as to say (page 339):

The MNL model has proven to have wide empirical applicability, but as a theoretical model of choice behaviour its IIA property is unsatisfactorily restrictive.

A variety of alternative models have been developed in order to relax these independence assumptions while still maintaining a clear basis in individual utility maximisation. For example, the Alternative-Specific Multinomial Probit assumes that the values of  $\varepsilon_{ij}$  are drawn from a multivariate normal distribution with a flexible covariance matrix; this would allow, for example, that the unobservable utility from taking a blue bus is very close to the unobservable utility from taking a red bus. However, this model — like most other alternative models — does not admit a closed form expression for the log-likelihood. The log-likelihood is therefore evaluated using simulation-based methods (*e.g.* ‘Maximum Simulated Likelihood’). These models are beyond the scope of our lectures — though they build directly upon the principles that we have discussed.

### 3.7 Appendix to Lecture 4: Stata code

First, clear the memory and load the data, as in previous lectures. We can then tabulate the variable `WorkCat`:




```
tab work_cat
```

We can estimate a multinomial logit — with a base category of 1 (agricultural employment) — using the ‘`mlogit`’ command. The ‘`margins`’ command will again provide the mean marginal effects (assuming that you can correctly specify the outcomes of interest!).

---

## 4 Lecture 4: Count models

### Required readings:

-  CAMERON, A.C. AND TRIVEDI, P.K. (2005): *Microeconometrics: Methods and Applications*. Cambridge University Press, pages 665 – 677 (*i.e.* sections 20.1 to 20.4.1, inclusive)  
*or*
-  WOOLDRIDGE, J. (2002): *Econometric Analysis of Cross Section and Panel Data*. The MIT Press, pages 645 – 657 (*i.e.* sections 19.1 to 19.3.1, inclusive)  
*or*
-  WOOLDRIDGE, J. (2010): *Econometric Analysis of Cross Section and Panel Data* (2nd ed.). The MIT Press, pages 723 – 738 (*i.e.* sections 18.1 to 18.3.1, inclusive).

### 4.1 Introduction: The concept of a count model

There are many ways in which a dependent variable may be ‘limited’. In our first and second lectures, we considered the simplest way: an outcome may be binary. In our third lecture, we generalised this to the case of multinomial choice – in which the dependent variable takes a *finite* set of values. In this final lecture, we consider models for the case of a dependent variable that can take any *non-negative integer values*: that is, *count models*.

### 4.2 Motivating example: Fertility trends in Tanzania

As in our three previous lectures on Limited Dependent Variables, we will illustrate using data from Tanzania. In this lecture, we will use data from the Tanzanian DHS surveys of 1992 and 2017.<sup>15</sup>

Figure 4.1 shows the relationship between (i) the age of respondent women (in 1992) and (ii) the number of children ever born to each woman. Figure 4.2 shows the equivalent figure for 2017. There appears to be an interesting ‘demographic shift’ evident in these graphs: in 1992, for example, 40-year old Tanzanian women had, on average, given birth to about six children each; in 2017, the equivalent figure was about five.

Suppose that we want to formalise this insight – by estimating an econometric model that allows directly for the ‘count’ nature of the outcome variable. In this lecture, we will consider two models for doing so: the Poisson and the Negative Binomial.

---

<sup>15</sup> Unlike the previous lectures, this is not data that is publicly available; if you would like to work with this data, you need to apply at <https://dhsprogram.com/>.

Figure 4.1: Age and children, Tanzania, 1992 (DHS data)

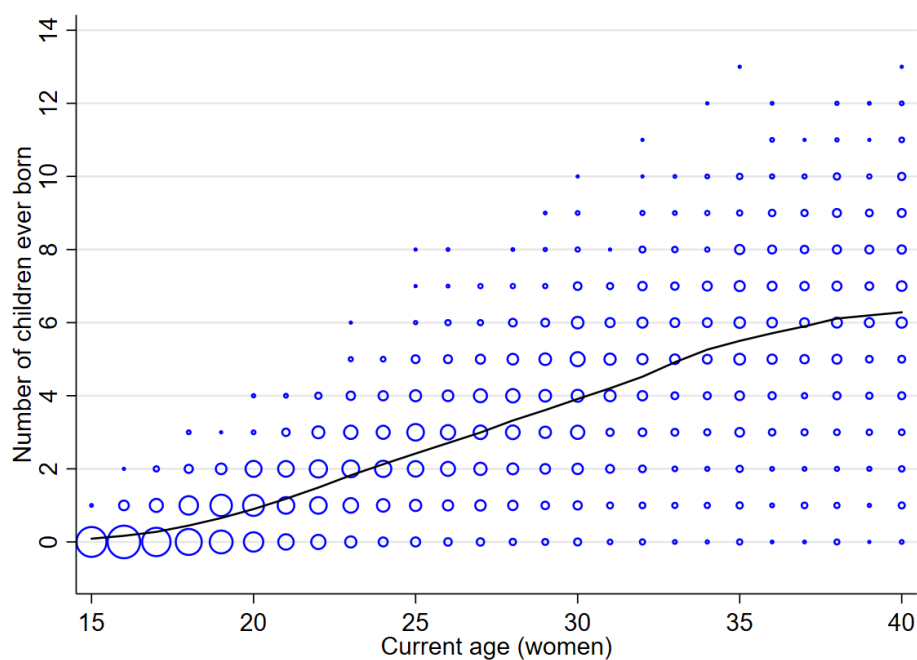
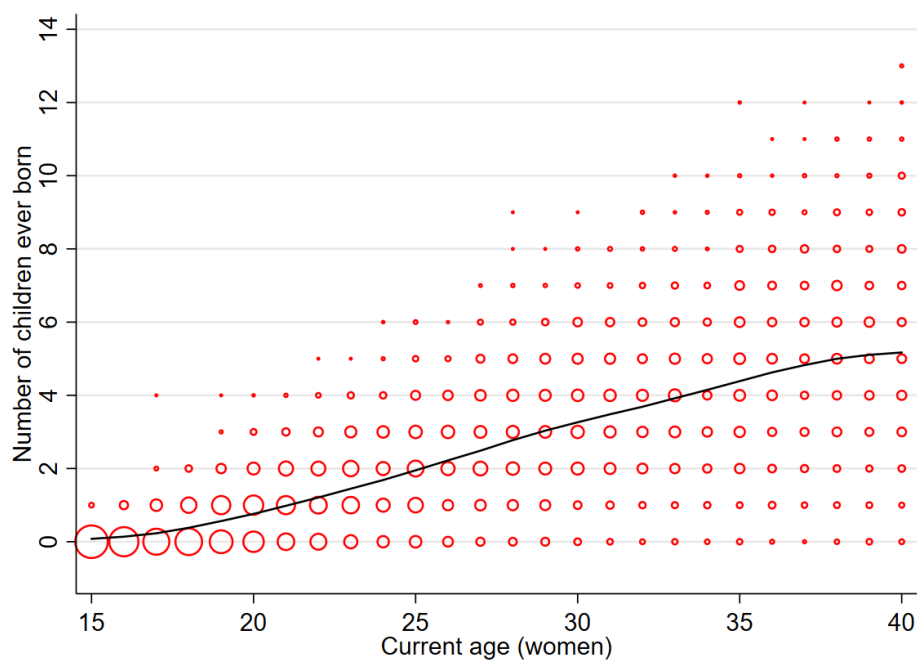


Figure 4.2: Age and children, Tanzania, 2017 (DHS data)



### 4.3 Introducing the Poisson model

Prussian cavalry soldiers died in many different ways. Sadly, this included being killed by being kicked by horses: from 1875 to 1894, a total of 196 Prussian cavalry members died in this way.<sup>16</sup> Famously, the Russian statistician Ladislaus Bortkiewicz used this data – disaggregated by year and by cavalry corps – to introduce the Poisson distribution.<sup>17</sup> You have already met the Poisson distribution on this course: as an application of the general principles of Maximum Likelihood, in a class exercise. In this lecture, we will discuss the model in more detail, and apply it – along with a model generalisation – to the issue of declining fertility in Tanzania.

In each of our three previous lectures, we started with random choice frameworks: you will recall that we respectively discussed the ways in which Additive Random Utility Models could be used as the conceptual foundation for the Probit model, the Logit model and the Multinomial Logit model. In the case of the Poisson model, the typical foundation is more ‘statistical’: specifically, we can think of the Poisson as a limiting case of a repeated draw from the Binomial distribution.

Specifically, start by considering a Binomial distribution, in which we have  $n$  independent draws, each having a probability of success of  $p$ . In that case, the distribution of the total number of successes,  $Y \in \{0, \dots, n\}$ , is:

$$\Pr(Y = y) = \frac{n!}{y! \cdot (n - y)!} \cdot p^y \cdot (1 - p)^{n-y} \quad (4.1)$$

$$= \frac{n \times (n - 1) \times \dots \times (n - y + 1)}{y!} \cdot p^y \cdot (1 - p)^{n-y}. \quad (4.2)$$

We know that this has an expectation of  $np$ ; we will denote this as  $\lambda \equiv np$ . We can therefore substitute  $p = \lambda/n$  into the expression:

$$\Pr(Y = y) = \frac{n \times (n - 1) \times \dots \times (n - y + 1)}{y!} \cdot \left(\frac{\lambda}{n}\right)^y \cdot \left(1 - \frac{\lambda}{n}\right)^{n-y} \quad (4.3)$$

$$= \frac{n \times (n - 1) \times \dots \times (n - y + 1)}{y!} \cdot \left(\frac{\lambda}{n}\right)^y \cdot \left(1 - \frac{\lambda}{n}\right)^n \cdot \left(1 - \frac{\lambda}{n}\right)^{-y} \quad (4.4)$$

$$= \frac{n}{n} \times \frac{n-1}{n} \times \dots \times \frac{n-y+1}{n} \cdot \frac{\lambda^y}{y!} \cdot \left(1 - \frac{\lambda}{n}\right)^n \cdot \left(1 - \frac{\lambda}{n}\right)^{-y} \quad (4.5)$$

Let’s think about what happens if we (i) hold fixed this expectation,  $\lambda$ , but (ii) increase to infinity the number of draws,  $n$ . (If we think of the  $n$  draws as occurring within some fixed total time period, this limit corresponds to the limit as the duration of each draw goes to zero.) From the previous expression – and remembering the rule that  $\lim_{n \rightarrow \infty} (1 - \lambda/n)^n = \exp(-\lambda)$  – we can say:

$$\lim_{n \rightarrow \infty} \Pr(Y = y) = \frac{\lambda^y \cdot \exp(-\lambda)}{y!}. \quad (4.6)$$

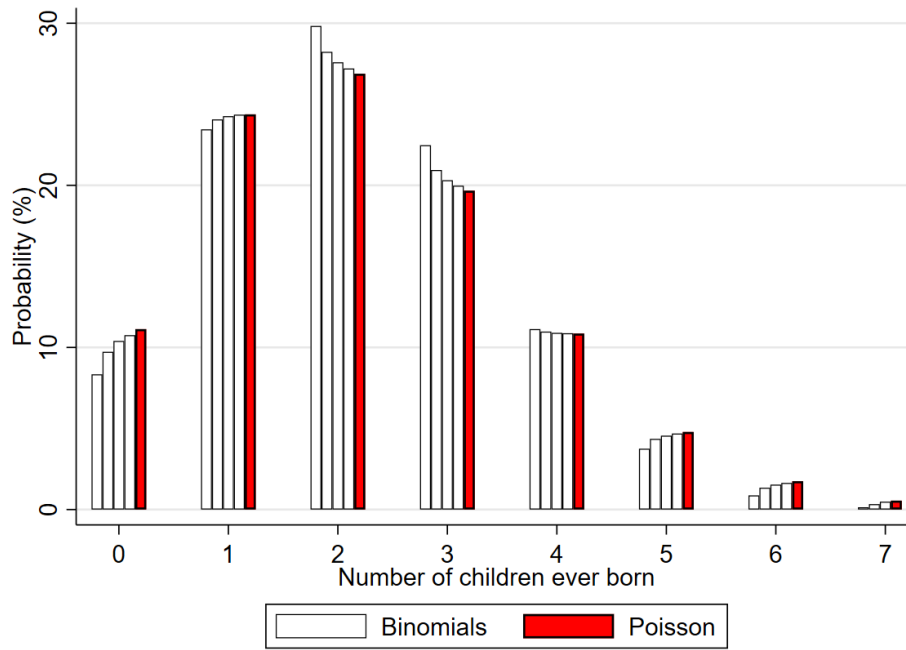
<sup>16</sup> See, for example, <https://www.randomservices.org/random/data/HorseKicks.html>.

<sup>17</sup> See [https://en.wikipedia.org/wiki/Ladislaus\\_Bortkiewicz](https://en.wikipedia.org/wiki/Ladislaus_Bortkiewicz).



This expression defines the probability mass function for the Poisson. Figure 4.3 shows the probability mass function for  $y \in \{0, \dots, 7\}$  children, for  $\lambda = 2.2$ ; the figure shows the *pmf* for four binomial distributions (corresponding to  $n = 10, n = 20, n = 40$  and  $n = 80$ ), as well as for the Poisson.

Figure 4.3: The Poisson as limiting case of the Binomial



## 4.4 Fitting the Poisson model

Let's now take our Poisson model to the data. Before we start introducing a role for age, we will think about how to estimate the Poisson model on the unconditional distribution of the number of children born to each woman. Specifically, we will think about how, for a given set of observations,  $y_i$  (for respondents indexed  $i \in \{1, \dots, N\}$ ), we can find the Maximum Likelihood estimate for the Poisson parameter,  $\lambda$ . We will do this using the 2017 DHS data described earlier.

Returning to equation 4.6, we can proceed as follows:

$$\Pr(Y = y; \lambda) = \frac{\lambda^y \cdot \exp(-\lambda)}{y!} \quad (4.7)$$

$$\therefore L(\lambda; y_1, \dots, y_N) = \prod_{i=1}^N \frac{\lambda^{y_i} \cdot \exp(-\lambda)}{y_i!} \quad (4.8)$$

$$\therefore \ell(\lambda; y_1, \dots, y_N) = \sum_{i=1}^N (y_i \cdot \ln \lambda - \lambda - \ln y_i!). \quad (4.9)$$

Taking the first-order condition with respect to  $\lambda$ , it follows that:

$$\sum_{i=1}^N \frac{y_i}{\hat{\lambda}} = N \quad (4.10)$$

$$\therefore \hat{\lambda} = \frac{1}{N} \cdot \sum_{i=1}^N y_i. \quad (4.11)$$

So – as one might intuitively expect – the Maximum Likelihood estimate of  $\lambda$  is simply the sample average number of children. (This derivation approach should be familiar from the class exercise on the Poisson model that you did for our ‘introduction to Maximum Likelihood’ lectures.)

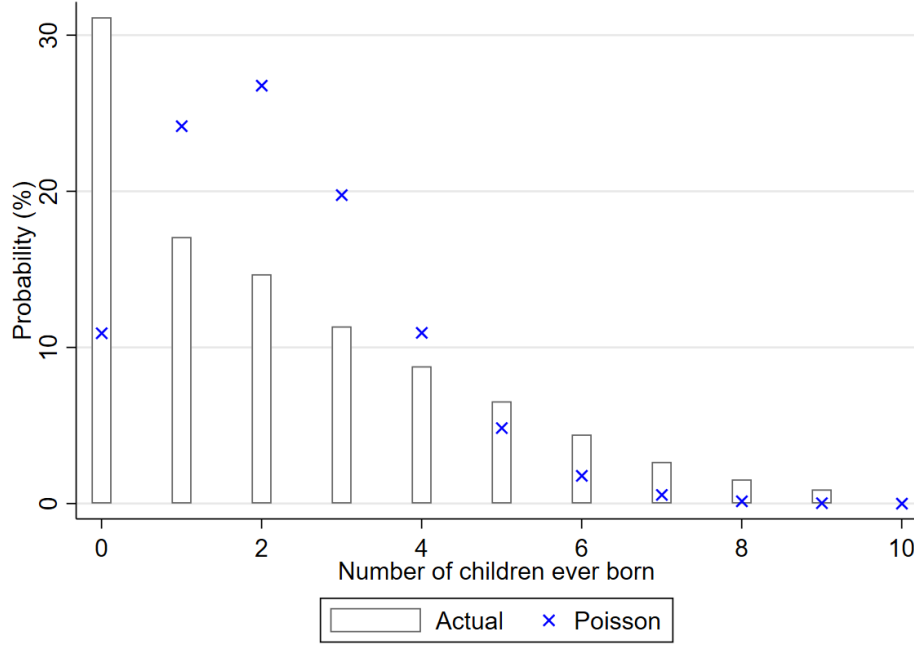
In the 2017 Tanzanian DHS data, the average number of children per woman is 2.2; therefore, this is our Maximum Likelihood estimate:  $\hat{\lambda} = 2.2$ . Figure 4.4 shows the implications of this estimate: the figure shows (i) a histogram of the distribution of children per woman, and (ii) the predicted probabilities from the Poisson model, using  $\lambda = 2.2$ .

Clearly, we have a problem. Even though  $\hat{\lambda} = 2.2$  is the Maximum Likelihood estimate, the Poisson model clearly fits the data very poorly. Note, for example, that the Poisson model (with  $\hat{\lambda} = 2.2$ ) implies that the modal number of children per mother should be 2; instead, the true mode is zero. Indeed, the model predicts that about 12% of Tanzanian women in 2017 had no children; the true number is just over 30%.

## 4.5 Overdispersion and the Negative Binomial model

### 4.5.1 Introducing the Negative Binomial

The problem illustrated in Figure 4.4 is often referred to as ‘*overdispersion*’: that is, the data shows greater heterogeneity than the underlying model allows. The fundamental problem here is that the Poisson distribution – for all its elegance and its relative statistical simplicity – is simply not flexible enough to model count outcomes in many real-world scenarios. This is because the Poisson has the ‘equidispersion’ property that  $\mathbb{E}(Y) = \text{Var}(Y) = \lambda$ . (We consider  $\mathbb{E}(Y)$  in the appendix to this lecture; you will have the opportunity to consider  $\text{Var}(Y)$  in the class exercises.)

Figure 4.4: **Poisson estimates: Distribution of children per woman in Tanzania (2017)**

There are several possible ways of proceeding. In this lecture, we will consider the Negative Binomial model: a common and flexible generalisation of the Poisson. The Negative Binomial deals with the overdispersion problem in a very elegant way: it assumes a Poisson model for each individual, but – even before we add covariates – it allows each individual to have a *different* value of  $\lambda$ . We can therefore think of the Negative Binomial as a ‘random parameter’ extension of the Poisson model.

Specifically, instead of treating  $\lambda$  as a fixed parameter, common to all respondents, we assume that  $\lambda$  itself is a random parameter, drawn such that:<sup>18</sup>

$$\lambda_i \equiv \mu \cdot \nu_i \quad (4.12)$$

$$\nu_i \sim \text{Gamma}\left(\frac{1}{\alpha}, \alpha\right). \quad (4.13)$$

We will not discuss the details of the Gamma distribution (and you do not need to know such details for any exam question on my part of the course). Note, though, that a variable with a distribution  $\text{Gamma}(\alpha^{-1}, \alpha)$  has a mean 1 and a variance  $\alpha$ .

This implies that the *pdf* for  $\nu_i$  is:

$$\frac{\nu^{(1-\alpha)/\alpha} \cdot \exp(-\nu/\alpha)}{\alpha^{1/\alpha} \cdot \Gamma(1/\alpha)}. \quad (4.14)$$

<sup>18</sup> I follow here the explanation in the Stata manual for the command `nbreg`.

We have just introduced the Gamma function,  $\Gamma(z)$ . We will not discuss this function in any detail<sup>19</sup> – however, note that, for any positive integer  $z$ ,  $\Gamma(z) = (z - 1)!$ .

We can therefore say:<sup>20</sup>

$$\Pr(Y = y; \alpha, \mu) = \int_0^\infty \underbrace{\frac{(\mu\nu)^y \cdot \exp(-\mu\nu)}{y!}}_{\text{conditional density for } Y \text{ given } \nu} \cdot \underbrace{\frac{\nu^{(1-\alpha)/\alpha} \cdot \exp(-\nu/\alpha)}{\alpha^{1/\alpha} \cdot \Gamma(1/\alpha)}}_{\text{marginal density for } \nu} d\nu \quad (4.15)$$

*= a bunch of steps that we won't worry about, and then...*

$$= \frac{\Gamma(\alpha^{-1} + y)}{\Gamma(\alpha^{-1}) \cdot \Gamma(y + 1)} \cdot \left( \frac{1}{1 + \alpha\mu} \right)^{\alpha^{-1}} \cdot \left( 1 - \frac{1}{1 + \alpha\mu} \right)^y \quad (4.16)$$

$$= \frac{\mu^y}{y!} \cdot \frac{\Gamma(\alpha^{-1} + y)}{\Gamma(\alpha^{-1}) \cdot (\alpha^{-1} + \mu)^y} \cdot \frac{1}{(1 + \alpha\mu)^{\alpha^{-1}}}. \quad (4.17)$$

This is an interesting function, for several reasons:

(i) Consider the limit as  $\alpha \rightarrow 0$ . We can say:

$$\lim_{\alpha \rightarrow 0} \Pr(Y = y; \alpha, \mu) = \frac{\mu^y}{y!} \cdot \lim_{\alpha \rightarrow 0} \frac{\Gamma(\alpha^{-1} + y)}{\Gamma(\alpha^{-1}) \cdot (\alpha^{-1} + \mu)^y} \cdot \lim_{\alpha \rightarrow 0} \frac{1}{(1 + \alpha\mu)^{\alpha^{-1}}} \quad (4.18)$$

$$= \frac{\mu^y}{y!} \cdot 1 \cdot \frac{1}{\exp(\mu)} \quad (4.19)$$

$$= \frac{\mu^y}{y!} \cdot \exp(-\mu), \quad (4.20)$$

which, of course, is the probability mass function for the Poisson with parameter  $\mu$ . That is, as we should expect, we recover the Poisson model in the limiting case where the variance of  $\nu$  goes to zero.

(ii) Suppose that  $\alpha^{-1}$  is some integer  $r$ . Then we can rewrite the probability as:

$$\Pr(Y = y) = \frac{(y + r - 1)!}{(r - 1)! \cdot y!} \cdot (1 - p)^r \cdot p^y \quad (4.21)$$

$$= \binom{r + y - 1}{y} \cdot (1 - p)^r \cdot p^y, \quad (4.22)$$

<sup>19</sup> Formally,  $\Gamma(z) = \int_0^\infty x^{z-1} \cdot \exp(-x) dx$ .

<sup>20</sup> To get from equation 4.16 to 4.17, note that

$$\left( 1 - \frac{1}{1 + \alpha\mu} \right)^y = \left( \frac{\alpha\mu}{1 + \alpha\mu} \right)^y = \mu^y \cdot \left( \frac{\alpha}{1 + \alpha\mu} \right)^y = \mu^y \cdot \left( \frac{1}{\alpha^{-1} + \mu} \right)^y.$$

where  $p = \mu / (r + \mu)$ . With this formulation, the probability is equivalent to *the probability, in a sequence of independent binary events, of  $y$  successes before  $r$  failures* (where  $p$  is the probability of a ‘success’). This is an alternative (and common) application for the Negative Binomial model.

### 4.5.2 Estimating the Negative Binomial

When we estimate the Negative Binomial we obtain  $\hat{\mu} = 2.2$  and  $\hat{\alpha} = 0.82$ . Figure 4.5 shows the implication of these estimates for the distribution of  $\lambda_i$ .

Figure 4.5: Estimated distribution of  $\lambda_i$

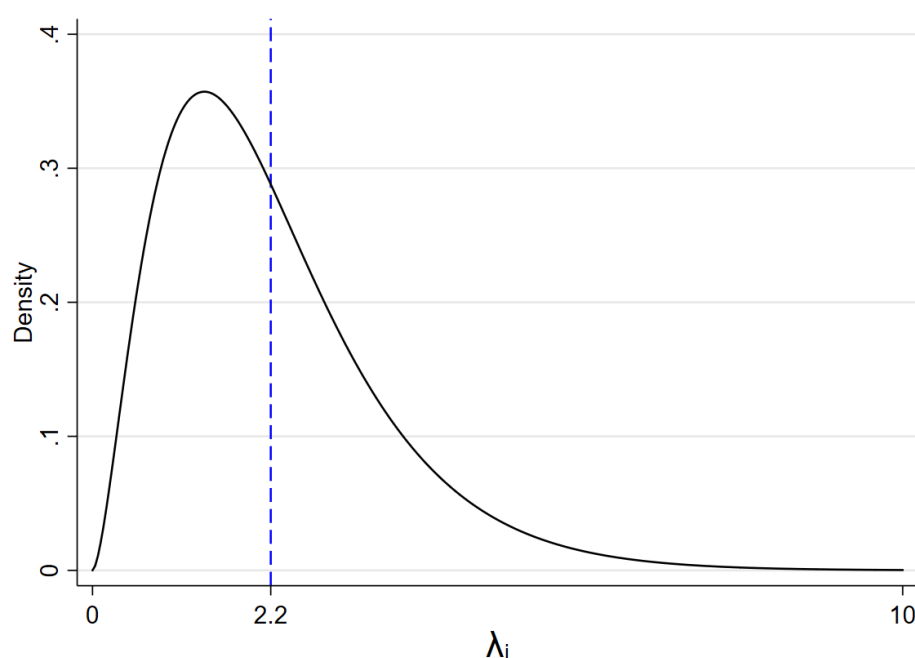


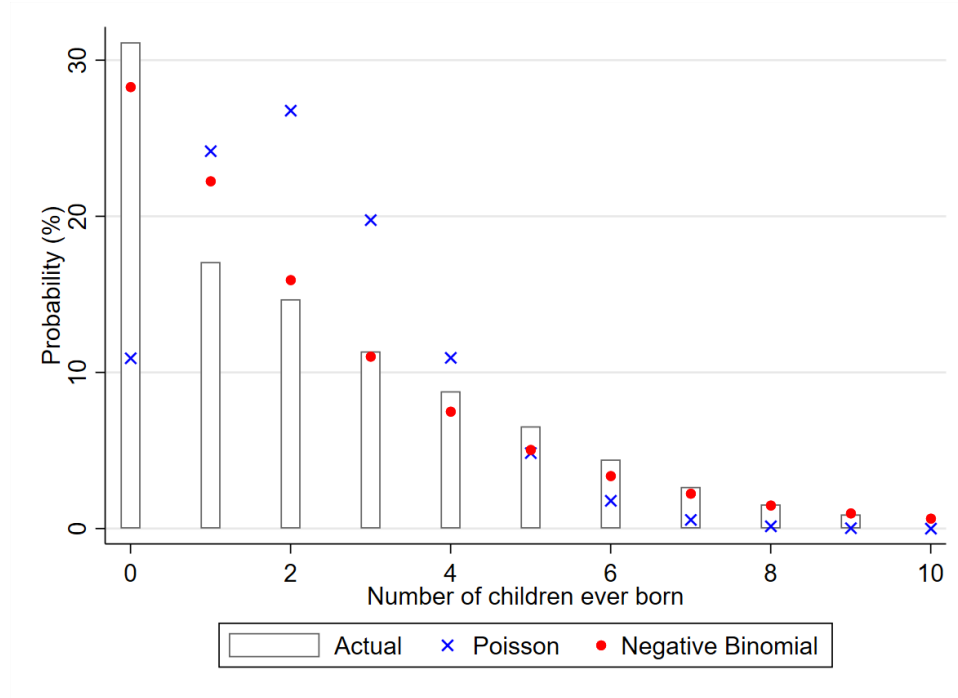
Figure 4.6 shows the predicted value for the number of children born.

### 4.5.3 Testing the restriction that $\alpha = 0$

As we just discussed, the Poisson is a special case of the Negative Binomial. We can, therefore, test the null hypothesis that the data is generated by a Poisson model, against the less restrictive hypothesis that the data is generated by a Negative Binomial. We can do this by constructing the standard Likelihood Ratio test statistic – where the log-likelihood for the Negative Binomial is the ‘unrestricted’ log-likelihood, and the Poisson provides the relevant ‘restricted’ log-likelihood.

However, there is one important twist to the Likelihood Ratio test in this case. It is true that the Poisson is a special case of the Negative Binomial – but, specifically, it is a special *limiting* case. That is, the Poisson corresponds to the Negative Binomial when  $\alpha = 0$ , and  $\alpha = 0$  lies on the

Figure 4.6: **Negative Binomial estimates: Distribution of children per woman in Tanzania (2017)**



boundary of the admissible parameter set for  $\alpha$ . (Put simply, we want to test whether the variance of  $\lambda_i$  is zero – and negative variances are impossible.)

For this reason, we should not compare the Likelihood Ratio statistic to the usual  $\chi^2(1)$  distribution, but to a distribution known as the  $\bar{\chi}^2(0, 1)$ . In practice, this amounts to calculating the  $p$ -value from the usual  $\chi^2(1)$  distribution and then halving it. (Intuitively, this is very similar to the distinction between a two-tailed test and a one-tailed test; for a single parameter restriction, we can think of the usual LR test as a two-tailed test, whereas we are interested here in testing  $H_0 : \alpha = 0$  against the one-sided alternative hypothesis,  $H_1 : \alpha > 0$ .)

## 4.6 Introducing the covariate

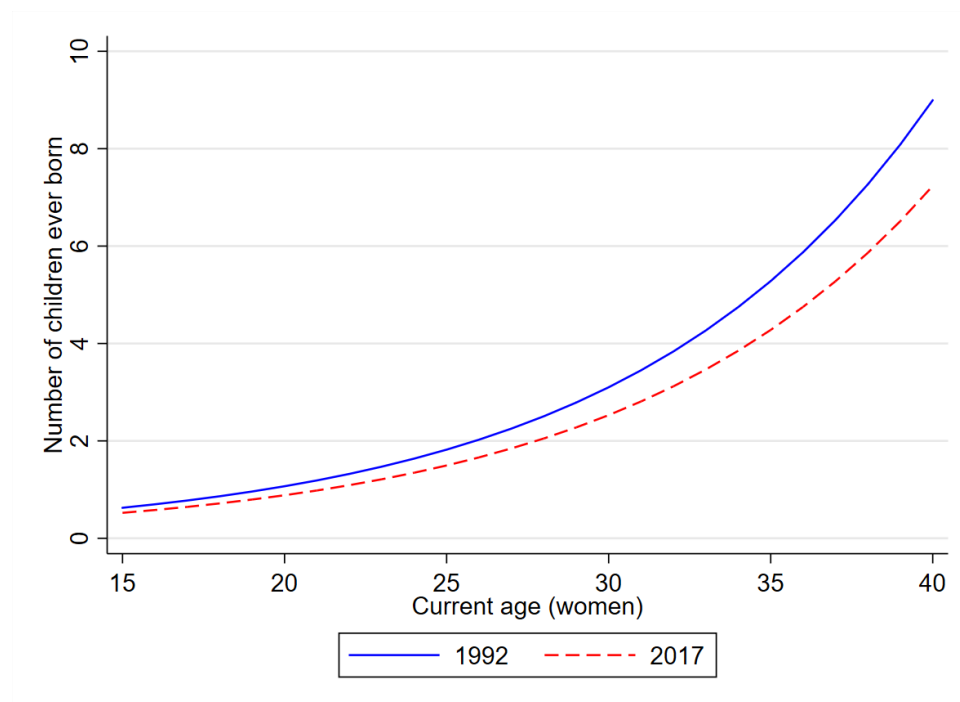
Time, at last, to introduce women's age to the analysis. Having discussed the construction of the log-likelihood for both the Poisson and the Negative Binomial models, it is relatively straightforward to include covariates: we simply need to choose some functional form that links the Poisson mean to the covariates of interest. There are any number of ways in which you might choose to do this; the only restriction is that, for any values of the covariates,  $\lambda$  must remain positive.

One very standard parameterisation is to use an exponential link function:<sup>21</sup>

$$\lambda_i = \exp(\beta \cdot x_i). \quad (4.23)$$

Figure 4.7 shows the predicted values from this estimation for our Tanzanian data (that is, using Maximum Likelihood to estimate with  $\lambda_i = \beta_0 + \beta_1 \cdot x_i$ , where  $x_i$  is the age of female respondent).

Figure 4.7: **Negative Binomial estimates: Distribution of children per woman in Tanzania (2017)**



<sup>21</sup> In Stata, the `Poisson` command imposes this link as standard. If you would like to use a different link function to estimate the Poisson in Stata, you can explore the `glm` command.

## 4.7 Other comments on Poisson

To conclude our lecture, let me make three brief points:

- (i) When using the Poisson model for your own research, you need to take a stance on whether you want to use ‘robust’ standard errors. In this lecture, we have discussed Poisson as a Maximum Likelihood estimator. In most applications, ‘robust’ estimation is probably a good idea; this then involves ‘pseudo-Maximum Likelihood’ (sometimes also called ‘quasi-Maximum Likelihood’). This is discussed, for example, in section 5.2.3 of Cameron and Trivedi.
- (ii) When estimating the Negative Binomial model, there is an important decision to be made about the structure of the variance. One approach – which we discussed in this lecture – is ‘mean dispersion’. This is the default in Stata, and is sometimes referred to as the ‘NB2’ option. There is also a ‘constant-dispersion’ option, sometimes referred to as ‘NB1’.
- (iii) Note that there are panel estimator versions available for both the Poisson and the Negative Binomial model (in Stata, these are ‘xtpoisson’ and ‘xtnbreg’).
- (iv) The Poisson – with robust or clustered errors – can be an excellent estimator for any situation where you have a skewed non-negative outcome variable (for example, wages). We will discuss this briefly in the lecture; it may be useful (for example) for your dissertation work.



## 4.8 Appendix: The mean of the Poisson

In this appendix, we prove that the expectation of a Poisson-distributed random variable is  $\lambda$ :

$$\Pr(Y = y) = \frac{\lambda^y \cdot \exp(-\lambda)}{y!} \quad (4.24)$$

$$\therefore \mathbb{E}(Y) = \sum_{y=0}^{\infty} y \cdot \Pr(Y = y) \quad (4.25)$$

$$= \sum_{y=1}^{\infty} y \cdot \Pr(Y = y) \quad (4.26)$$

$$= \sum_{y=1}^{\infty} y \cdot \frac{\lambda^y \cdot \exp(-\lambda)}{y!} \quad (4.27)$$

$$= \sum_{y=1}^{\infty} \frac{\lambda^y \cdot \exp(-\lambda)}{(y-1)!} \quad (4.28)$$

$$= \lambda \cdot \sum_{y=1}^{\infty} \frac{\lambda^{y-1} \cdot \exp(-\lambda)}{(y-1)!} \quad (4.29)$$

$$= \lambda \cdot \sum_{y=0}^{\infty} \frac{\lambda^y \cdot \exp(-\lambda)}{y!} \quad (4.30)$$

$$= \lambda \cdot \sum_{y=0}^{\infty} \Pr(Y = y) \quad (4.31)$$




$$= \lambda. \quad (4.32)$$

You will have the opportunity, in your class exercises, to adapt this proof to find the variance of a Poisson-distributed random variable.

---

## 5 Lecture 5 (NOT EXAMINABLE): Discrete ordered choice

### Required reading:

-  CAMERON, A.C. AND TRIVEDI, P.K. (2005): *Microeconometrics: Methods and Applications*. Cambridge University Press, pages 519 – 520 (*i.e.* section 15.9.1)  
*or*
-  WOOLDRIDGE, J. (2002): *Econometric Analysis of Cross Section and Panel Data*. The MIT Press, pages 504 – 507 (*i.e.* section 15.10)  
*or*
-  WOOLDRIDGE, J. (2010): *Econometric Analysis of Cross Section and Panel Data* (2nd ed.). The MIT Press, pages 655 – 659 (*i.e.* sections 16.3.1 and 16.3.2).

### Other references:

- CUNHA, F., HECKMAN, J. AND NAVARRO, S. (2007): “The Identification and Economic Content of Ordered Choice Models with Stochastic Thresholds,” *International Economic Review*, 48(4), 1273–1309.

### 5.1 The concept of ordered choice

In Lectures 1 and 2, we considered the problem of binary outcome variables; we did so by considering Tanzanians’ decision whether or not to complete primary school education. In this lecture, we extend our earlier reasoning to consider the problem of *discrete ordered choice*. To do so, we will continue to work with the Tanzanian ILFS dataset; we will now consider Tanzanians’ decision between three choices: (i) not completing primary education, (ii) completing primary education but not secondary education, and (iii) completing secondary education.

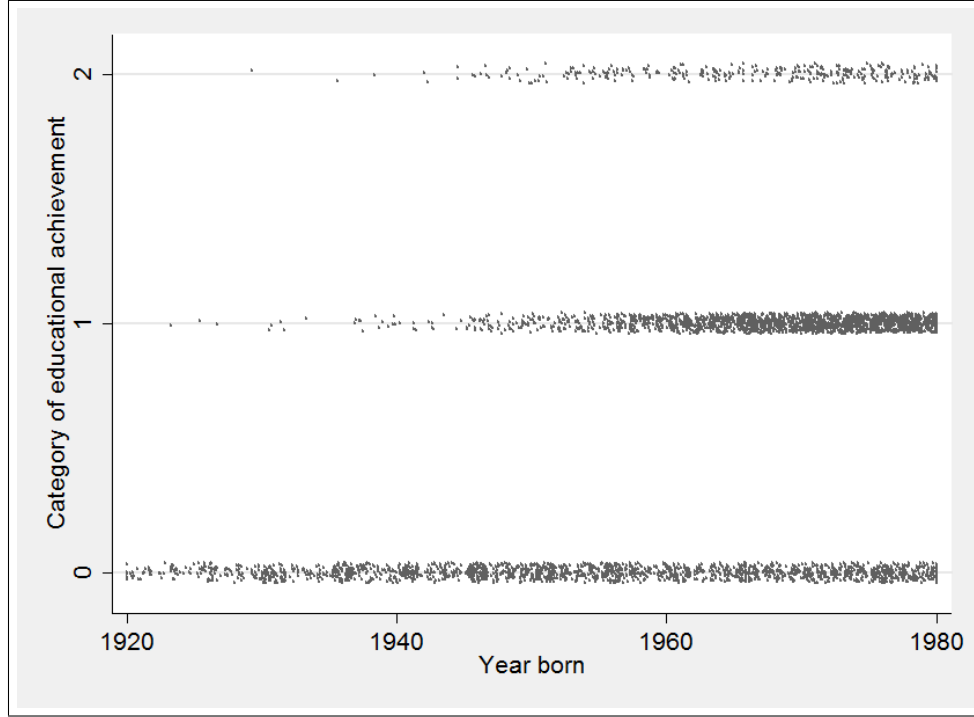
We will denote our outcome variable as follows:

$$y_i = \begin{cases} 0 & \text{if the } i\text{th individual did not complete primary education;} \\ 1 & \text{if the } i\text{th individual completed primary education but not secondary education;} \\ 2 & \text{if the } i\text{th individual completed secondary education.} \end{cases} \quad (5.1)$$

Figure 5.1 shows how attainment of primary and secondary education has changed over time in Tanzania; it plots our new variable  $y_i$  against respondents’ year of birth ( $x_i$ ). Note again one of the key characteristics of many limited dependent variable models: the outcome variable is *categorical*, so the numerical values taken by  $y_i$  have no cardinal meaning. There is no sense, for example, in which completing secondary education ( $y_i = 2$ ) is ‘twice as good’, or ‘twice as useful’, or ‘twice as *anything*’ as completing primary education ( $y_i = 1$ ).

We will model this education decision as an *ordered choice*. It is clear that, in some *intuitive* sense, the categories ‘no education’ – ‘primary education’ – ‘secondary education’ are ordered; for

Figure 5.1: Primary and secondary school attainment in Tanzania across age cohorts



example, secondary education requires more time than primary education, which itself (obviously) requires more time than no education. This kind of intuitive reasoning often justifies the description of a choice as a ‘*discrete ordered choice*’. Ideally, though, we should be able to go further: we should be able to describe the outcome as a *monotone step function* of some continuous latent variable. To illustrate what this might mean, we will consider a simple microeconomic model of Tanzanians’ investment in education.

## 5.2 A simple optimal stopping model

Assume that a student obtains some utility from attending school (or, equivalently, pays some utility cost), and that this utility changes with (i) the student’s year of birth ( $x_i$ ), and (ii) the student’s unobserved taste for education ( $\varepsilon_i$ ):

$$u_{it}^s(x_i) = \beta_0 + \beta_1 x_i + \varepsilon_i. \quad (5.2)$$

Additionally, assume that the student may work and receive in-period utility determined by the student’s level of education ( $s_i$ ):

$$u_{it}(s_i) = \alpha s_i. \quad (5.3)$$

We assume that  $\varepsilon_i$  is known to the student, but unobservable to a researcher.

For simplicity, let's assume that the student faces only three choices: (i) do not attend school ( $s = 0$ ), (ii) finish primary school ( $s = 7$ ), and (iii) finish secondary school ( $s = 12$ ). We will assume that students have a lifetime of known finite duration  $T > 12$  years and, for simplicity, we will make the extreme assumption that students assign equal utility weight to each period.<sup>22</sup> Given these assumptions, we can write three value functions, one corresponding to each choice:

$$V_0(0, x_i, \varepsilon_i) = 0 \quad (5.4)$$

$$V_0(7, x_i, \varepsilon_i) = 7 \cdot (\beta_0 + \beta_1 x_i + \varepsilon_i) + (T - 7) \cdot 7\alpha \quad (5.5)$$

$$V_0(12, x_i, \varepsilon_i) = 12 \cdot (\beta_0 + \beta_1 x_i + \varepsilon_i) + (T - 12) \cdot 12\alpha. \quad (5.6)$$

Therefore, the student prefers  $s = 7$  to  $s = 0$  if and only if:<sup>23</sup>

$$\beta_0 + \beta_1 x_i + \varepsilon_i \geq (7 - T) \cdot \alpha. \quad (5.7)$$

Similarly, the student prefers  $s = 12$  to  $s = 7$  if and only if:

$$12 \cdot (\beta_0 + \beta_1 x_i + \varepsilon_i) + (T - 12) \cdot 12\alpha \geq 7 \cdot (\beta_0 + \beta_1 x_i + \varepsilon_i) + (T - 7) \cdot 7\alpha \quad (5.8)$$

$$\Leftrightarrow \beta_0 + \beta_1 x_i + \varepsilon_i \geq (19 - T) \cdot \alpha. \quad (5.9)$$

We can therefore define two ‘cutpoints’,

$$\kappa_1 = \alpha \cdot (7 - T) - \beta_0 \quad (5.10)$$

$$\kappa_2 = \alpha \cdot (19 - T) - \beta_0, \quad (5.11)$$

and express the  $i$ th student's decision ( $y_i$ ) as an ordered choice in the latent variable  $\beta_1 x_i + \varepsilon_i$ :

$$y(x_i, \varepsilon_i; \beta_1, \kappa_1, \kappa_2) = \begin{cases} 0 & \text{if } \beta_1 x_i + \varepsilon_i < \kappa_1; \\ 1 & \text{if } \beta_1 x_i + \varepsilon_i \in [\kappa_1, \kappa_2); \\ 2 & \text{if } \beta_1 x_i + \varepsilon_i \geq \kappa_2. \end{cases} \quad (5.12)$$

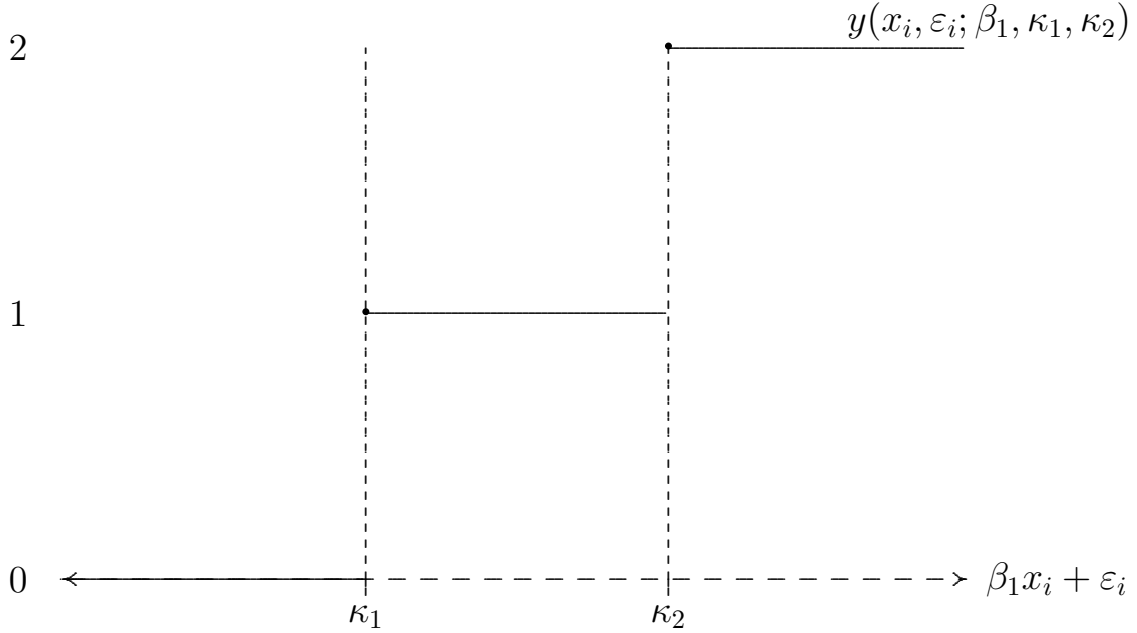
Our simple optimal stopping model therefore implies that  $y(x, \varepsilon; \beta_0, \beta_1)$  is a *monotone step function* in  $\beta_1 x_i + \varepsilon_i$ . Cunha, Heckman and Navarro (2007) discuss several classes of models — including a dynamic schooling choice model — that imply this kind of monotone step function solution. Figure 5.2 illustrates.

Notice that, in this model, the observable covariate  $x$  affects the *latent index* rather than the *cutpoints*; for this reason, we can describe the Ordered Probit as a model of ‘*index shift*’. There are two reasons that this result matters for empirical analysis:

- (i) We may wish to exploit the ordered nature of the outcome variable for more efficient estimation;
- (ii) We may wish to test the microeconomic model by testing the implication that  $x$  affects educational choice through ‘index shift’.

<sup>22</sup> That is, we will use a subjective discount factor of 1.

<sup>23</sup> We will assume that the indifferent student chooses the higher level of education.

Figure 5.2: Optimal schooling as a monotone step function in  $\beta_1 x_i + \varepsilon_i$ 

### 5.3 The Ordered Probit

The implications of our simple optimal stopping model are important. However, we need more before we can take these implications to data: once again, we need a distributional assumption about  $\varepsilon$ . We will make the same assumption that we made in Lecture 1.

**Assumption 5.1 (DISTRIBUTION OF  $\varepsilon_i$ )**  $\varepsilon_i$  is i.i.d. with a standard normal distribution, independent of  $x_i$ :

$$\varepsilon_i | x_i \sim \mathcal{N}(0, 1). \quad (5.13)$$

Armed with this assumption, it is straightforward to write the log-likelihood for the  $i$ th individual:

$$\ell_i(\beta_1, \kappa_1, \kappa_2; y_i | x_i) = \begin{cases} \ln \Phi(\kappa_1 - \beta_1 x_i) & \text{if } y_i = 0; \\ \ln [\Phi(\kappa_2 - \beta_1 x_i) - \Phi(\kappa_1 - \beta_1 x_i)] & \text{if } y_i = 1; \\ \ln [1 - \Phi(\kappa_2 - \beta_1 x_i)] & \text{if } y_i = 2. \end{cases} \quad (5.14)$$

### 5.4 Marginal effects in the Ordered Probit model

Marginal effects in the Ordered Probit model are directly analogous to marginal effects in the probit model. For simplicity, we will consider only the case in which  $x_i$  is continuous. Consider first the marginal effects for the extreme categories,  $y_i = 2$  and  $y_i = 0$ . Following the reasoning in

subsection 1.6, we have:

$$M_0(x_i; \hat{\beta}_1, \hat{\kappa}_1) = \frac{\partial \Pr(y_i = 0 | x_i; \hat{\beta}_1, \hat{\kappa}_1)}{\partial x_i} = -\hat{\beta}_1 \cdot \phi(\hat{\kappa}_1 - \hat{\beta}_1 \cdot x_i), \text{ and} \quad (5.15)$$

$$M_2(x_i; \hat{\beta}_1, \hat{\kappa}_2) = \frac{\partial \Pr(y_i = 2 | x_i; \hat{\beta}_1, \hat{\kappa}_2)}{\partial x_i} = \hat{\beta}_1 \cdot \phi(\hat{\kappa}_2 - \hat{\beta}_1 \cdot x_i). \quad (5.16)$$

For the intermediate category, we can find the marginal effect simply by considering the effect of  $x_i$  at both cutoffs:

$$M_1(x_i; \hat{\beta}_1, \hat{\kappa}_1, \hat{\kappa}_2) = \frac{\partial \Pr(y_i = 1 | x_i; \hat{\beta}_1, \hat{\kappa}_1, \hat{\kappa}_2)}{\partial x_i} = \hat{\beta}_1 \cdot [\phi(\hat{\kappa}_1 - \hat{\beta}_1 \cdot x_i) - \phi(\hat{\kappa}_2 - \hat{\beta}_1 \cdot x_i)]. \quad (5.17)$$

These principles generalise naturally to the case where  $x_i$  is discrete, and to the case in which there are more than three categories.

## 5.5 The Ordered Probit in Tanzania

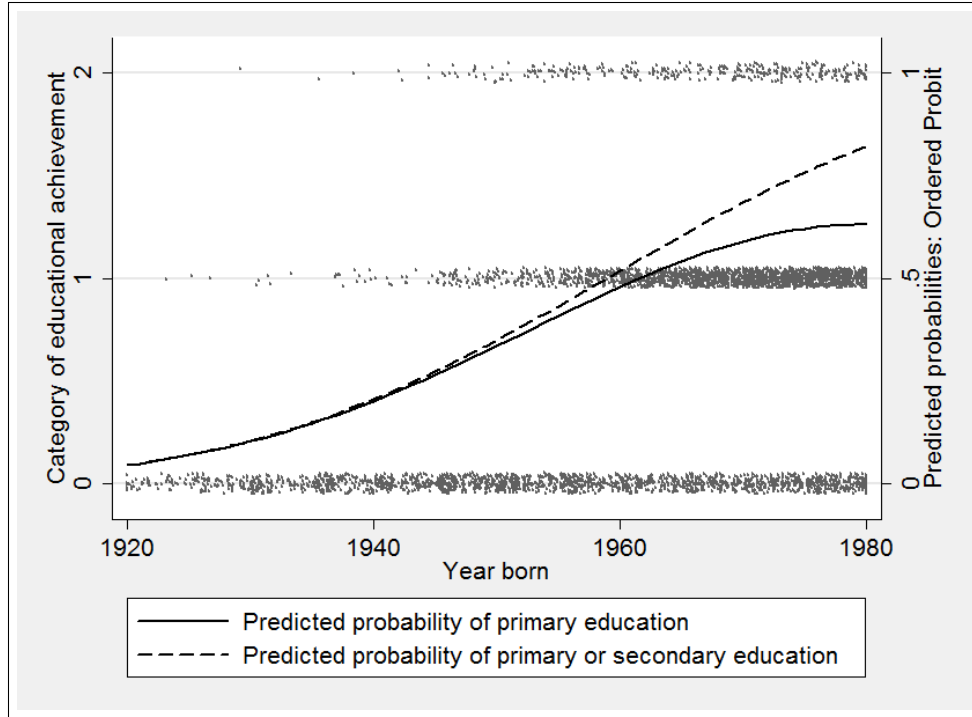
Table 5.1 shows the estimates from the Ordered Probit model for our Tanzanian data: we estimate  $\hat{\beta}_1 = 0.039$ ,  $\hat{\kappa}_1 = 76.672$  and  $\hat{\kappa}_2 = 78.517$ . Columns 2 and 3 respectively show the mean marginal effects for the outcomes  $y = 1$  and  $y = 2$  (that is, I omit the mean marginal effect for outcome  $y = 0$ ; you should be able to calculate this, however). Figure 5.3 shows the consequent predicted probabilities.

Table 5.1: Estimates from Tanzania: Ordered Probit

	Estimates	Mean Marginal Effects	
		$y = 1$	$y = 2$
	(1)	(2)	(3)
Year born	0.039 (0.001)***	0.008 (0.002)***	0.005 (0.002)***
Cutoff 1 ( $\hat{\kappa}_1$ )	76.672 (1.886)***		
Cutoff 2 ( $\hat{\kappa}_2$ )	78.517 (1.890)***		
Obs.	10000		
Log-likelihood	-7972.275		
Pseudo- $R^2$	0.105		

**Confidence:** \*\*\*  $\leftrightarrow$  99%, \*\*  $\leftrightarrow$  95%, \*  $\leftrightarrow$  90%.

Figure 5.3: Ordered Probit estimates for primary and secondary school attainment in Tanzania across age cohorts



## 5.6 The Generalised Ordered Probit

We just noted that the Ordered Probit model is a model of *index shift*: the observable variable  $x_i$  affects the latent index  $\beta_1 x_i + \varepsilon_i$ . This was justified by our simple optimal stopping model, in which year of birth ( $x_i$ ) directly affected each student's utility from attending school. This model — *i.e.* both the optimal stopping model and the Ordered Probit — therefore implied that we could summarise the effect of age on *both* primary and secondary education with a single parameter:  $\hat{\beta}_1$ . This implies, for example, that if we estimate that, over time, students are more likely to complete primary school (*i.e.*  $\Pr(y_i = 0 | x_i)$  is decreasing), we must also estimate that students are more likely to complete secondary school (*i.e.* that  $\Pr(y_i = 2 | x_i)$  is increasing). This is implied in equations 5.15 and 5.16; the marginal effects on the largest and smallest outcomes must have opposite signs.

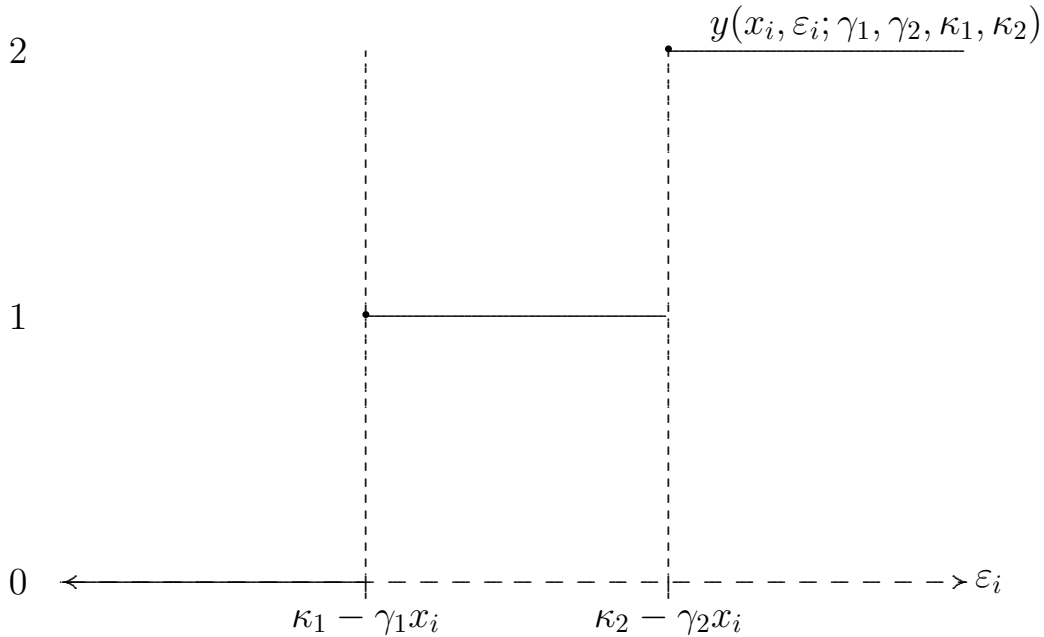
However, we might be concerned that this structure is too restrictive. After all, there might be many good reasons that, over time, students have become less likely to complete secondary education *and* less likely to complete no education at all (*i.e.* with more students stopping after primary school). This might be the case if, for example, the cost of primary education has fallen over time but the cost of secondary education has increased. In that case, we may still believe that educational choice is a monotone step function in the student's unobserved taste for education ( $\varepsilon_i$ ), but we may

want to allow the explanatory variable to affect each cutpoint differently. That is, we may want to use a model of ‘*cutpoint shift*’, rather than of ‘*index shift*’:

$$y(x_i, \varepsilon_i; \beta_1, \kappa_1, \kappa_2) = \begin{cases} 0 & \text{if } \varepsilon_i < \kappa_1 - \gamma_1 x_i; \\ 1 & \text{if } \varepsilon_i \in [\kappa_1 - \gamma_1 x_i, \kappa_2 - \gamma_2 x_i); \\ 2 & \text{if } \varepsilon_i \geq \kappa_2 - \gamma_2 x_i. \end{cases} \quad (5.18)$$

This model is identical to our earlier model in the special case  $\gamma_1 = \gamma_2 = \beta_1$ . But, by allowing  $\gamma_1$  and  $\gamma_2$  to vary separately, we can allow for a more flexible model while still exploiting the ordered structure of the decision. (Note, of course, that we haven’t gone back to modify our simple optimal stopping model to reflect this change; however, we could certainly do so — for example, by allowing  $x_i$  to affect the cost of each schooling level differently.) Figure 5.4 illustrates this more general model.

Figure 5.4: **Optimal schooling as a monotone step function in  $\varepsilon_i$**



If we maintain the assumption that  $\varepsilon_i$  has a standard normal distribution, we can describe this new model as a ‘*Generalised Ordered Probit*’. We will not write the log-likelihood for this model, but it is straightforward and directly analogous to the log-likelihood for the Ordered Probit. Table 5.2 shows the estimation results, with estimated mean marginal effect. Compared to the Ordered Probit, the Generalised Ordered Probit implies a slightly *higher* mean marginal effect upon the probability of primary education ( $y_i = 1$ ), but a slightly *lower* effect upon the probability of secondary ( $y_i = 2$ ).



Table 5.2: Estimates from Tanzania: Generalised Ordered Probit

	Estimates	Mean Marginal Effects	
		$y = 1$	$y = 2$
Year born		0.013 (0.0002)***	0.001 (0.0002)***
<b>Cutoff 1:</b>			
Year born	0.045 (0.001)***		
Const.	88.725 (2.017)***		
<b>Cutoff 2:</b>			
Year born	0.01 (0.001)***		
Const.	21.794 (2.882)***		
Obs.	10000		
Log-likelihood	-7779.772		
Pseudo- $R^2$	0.126		
<b>Confidence:</b> *** $\leftrightarrow$ 99%, ** $\leftrightarrow$ 95%, * $\leftrightarrow$ 90%.			

Figure 5.5 shows the consequent predicted probabilities. Figure 5.6 shows the estimated cutoff functions for the Generalised Ordered Probit (that is,  $\hat{\kappa}_1 - \hat{\gamma}_1 x_i$  and  $\hat{\kappa}_2 - \hat{\gamma}_2 x_i$ ) — along with the cutoffs for the Ordered Probit ( $\hat{\kappa}_1 - \hat{\beta}_1 x_i$  and  $\hat{\kappa}_2 - \hat{\beta}_1 x_i$ ). The diagram shows the fundamental difference between the Ordered Probit and Generalised Ordered Probit: the Ordered Probit restricts the cutoff functions to be parallel. Of course, this may or may not be a valid (or useful) restriction, depending on our particular empirical context. We can test the restriction straightforwardly: you should verify that, using a Likelihood Ratio test, we can compare the results in Tables 5.2 and 5.1 and obtain  $LR = 2 \times (7972.275 - 7779.772) = 385.006$ , which implies a tiny  $p$ -value when compared to a  $\chi^2(1)$  distribution.

Figure 5.5: Generalised Ordered Probit estimates for primary and secondary school attainment in Tanzania across age cohorts

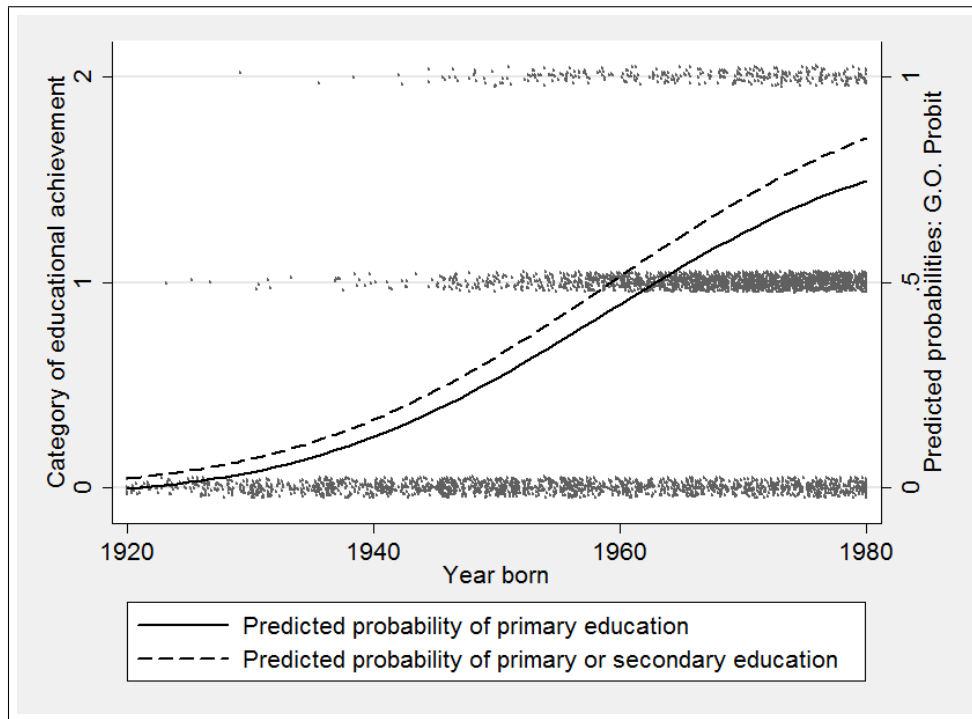
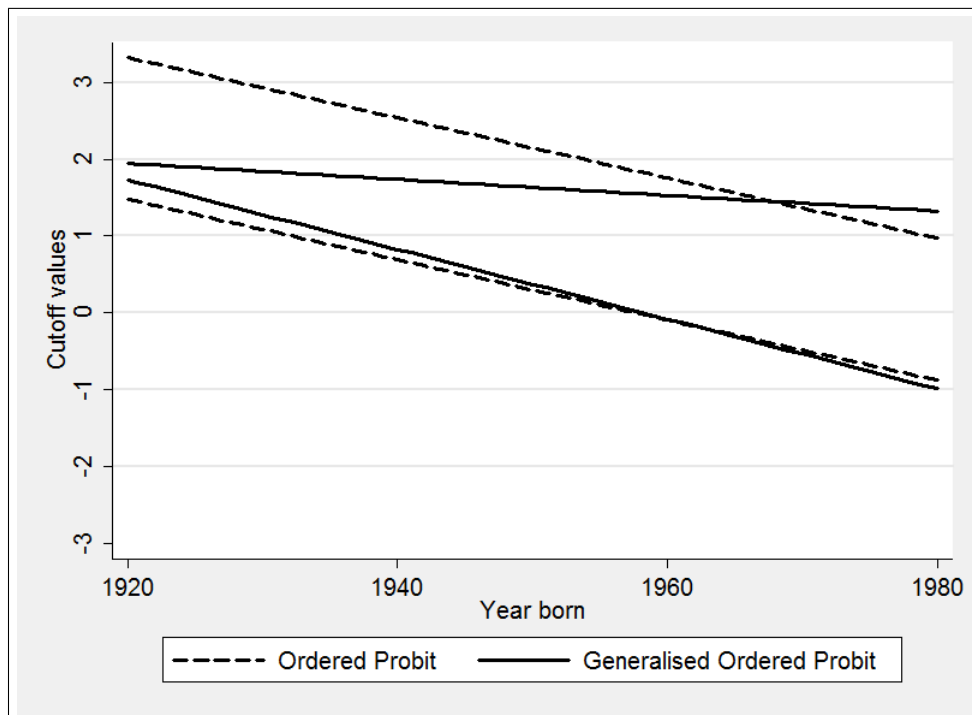


Figure 5.6: Estimated cutoff functions: Ordered Probit and Generalised Ordered Probit



## 5.7 The Ordered Logit and Generalised Ordered Logit

Recall that, in the binary outcome case, the *probit* model is motivated by the assumption that the latent error term has a *normal* distribution, and the *logit* model is motivated by the assumption that the error has a *logistic distribution*. In this lecture, we have considered the Ordered Probit and the Generalised Ordered Probit. Both specifications have relied upon the assumption that  $\varepsilon$  has a normal distribution. However, as in the binary outcome case, we could assume instead that  $\varepsilon$  has a logistic distribution. By analogy to the binary outcome case, we would then call our estimators the Ordered Logit and the Generalised Ordered Logit.

## 5.8 The Linear Probability Model and discrete ordered choice

In Lecture 2, we considered the Linear Probability Model as an alternative to the probit or logit model. We can also use a Linear Probability Model as an alternative to the Generalised Ordered Probit (or Generalised Ordered Logit). It would be *tempting* to write such an alternative like this:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad (5.19)$$

where  $y_i$  again refers to our three-outcome measure of educational achievement and  $x_i$  is again year of birth. That is, we could simply run an OLS regression of  $y_i$  on  $x_i$ . However, it is very difficult — *if not impossible* — to justify this approach. The reason for the difficulty is simple: as we noted earlier,  $y_i$  is a *categorical outcome*, where the values ‘0’, ‘1’ and ‘2’ have no cardinal meaning. It is therefore not meaningful to talk about a ‘one unit increase in the outcome variable’ (for example, we cannot interpret our estimate of  $\beta_1$  in terms of a marginal effect on a conditional probability). Unfortunately, it is not uncommon to see researchers using specifications like equation 5.19 for studying discrete outcomes.

Instead, we ought to estimate in a way that respects the categorical nature of the dependent variable. If we want to use a linear probability structure, we can do this by using *multiple* LPM estimates. In our Tanzanian example, we can do this by defining two new binary outcomes:

$$p_i = \begin{cases} 1 & \text{if } y_i = 1. \\ 0 & \text{if } y_i \neq 1; \end{cases} \quad (5.20)$$

$$s_i = \begin{cases} 1 & \text{if } y_i = 2. \\ 0 & \text{if } y_i \neq 2; \end{cases} \quad (5.21)$$

We can then run *two separate* Linear Probability Models to estimate the effect of  $x_i$  on the probability of primary completion and the probability of secondary completion:

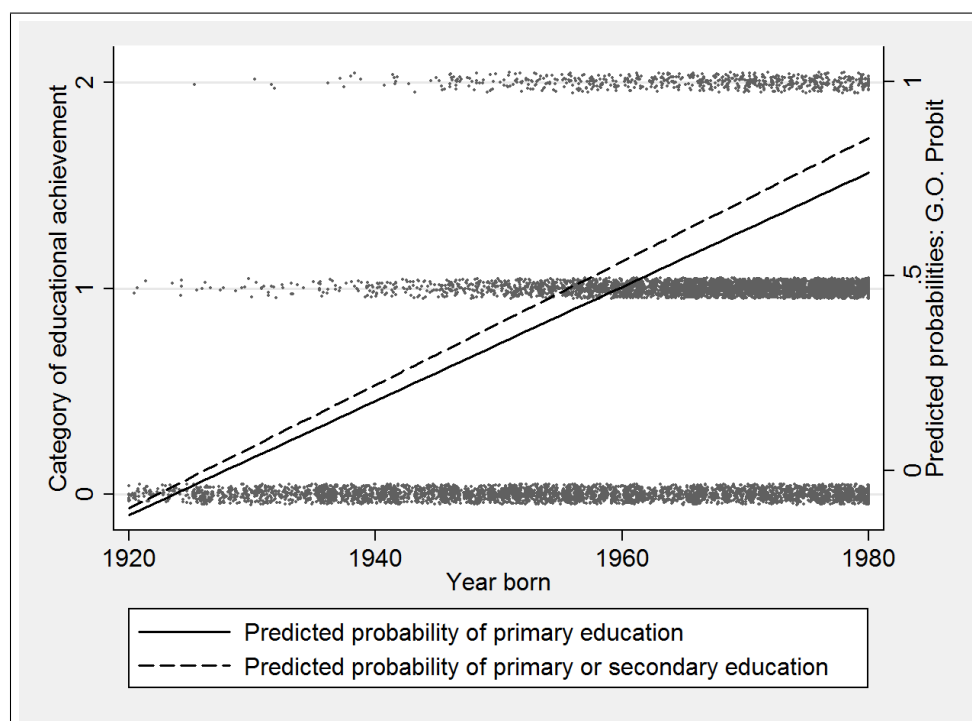
$$p_i = \gamma_0 + \gamma_1 x_i + \varepsilon_i \quad (5.22)$$

$$s_i = \delta_0 + \delta_1 x_i + \mu_i. \quad (5.23)$$

Figure 5.7 shows the resulting estimates. We can compare this graph directly to Figure 5.5. Of course, the estimates illustrated in Figure 5.7 are not necessarily *good* estimates: all of the objections to the Linear Probability Model that we discussed in Lecture 2 still apply. Arguably, these

objections apply with even more force where the dependent variable has multiple outcomes: we may think that it is even more important, in this case, to use an estimator that can be rationalised in terms of an underlying economic structure. But these estimates can at least be defended as providing reasonable estimates of the marginal effect of  $x_i$  on the probability of choosing  $y_i = 1$  and the probability of choosing  $y_i = 2$ . Unfortunately, this is not something that we can say about equation 5.19.<sup>24</sup>

Figure 5.7: **Linear Probability Model estimates for primary and secondary school attainment in Tanzania across cohorts**



<sup>24</sup> I have introduced this ‘multiple LPM’ approach as an alternative to the Generalised Ordered Probit. We could also use it as an alternative to models of discrete multinomial choice (*i.e.* ‘unordered’ choice), which we will discuss in Lecture 4. However, as we will see in that lecture, models of unordered choice have traditionally placed particular emphasis upon having choice-theoretic foundations — which, as we saw in Lecture 2, the Linear Probability Model does not provide.

## 5.9 Appendix to Lecture 5: Stata code

We can start by clearing the memory and loading the data — as we did in the exercises for Lectures 1 and 2. Then, we can tabulate our categorical education variable:

```
tab educ_cat
```

We can run an Ordered Probit with the `oprobit` command (`'help oprobit'`). We can then calculate mean marginal effects for the outcomes  $y = 1$  and  $y = 2$ . (Notice that the `margins` command works slightly differently where we need to specify which outcome we are thinking about; try the option `'predict(outcome(1))'`).

We can fit the Generalised Ordered Probit using the `goprobit` command. (Note that `goprobit` actually estimates, in our terminology,  $-\kappa_1$  and  $-\kappa_2$ , rather than  $\kappa_1$  and  $\kappa_2$ . Note also that this command may not be installed on your computer; the command is not currently built in to Stata, so you may have to download it separately.)

To use multiple Linear Probability Models instead, we can generate new dummy variables, then run OLS regressions. For example, you might want to start with something like `'gen p = (educ_cat == 1)'`.