Econometrics II

Lecture 2: Discrete Choice Models

Måns Söderbom*

4 April 2011

*University of Gothenburg. mans.soderbom@economics.gu.se. www.economics.gu.se/soderbom. www.soderbom.net

1. Introduction

Linear regression is primarily designed for modelling a **continuous**, **quantitative** variable - e.g. economic growth, the log of value-added or output, the log of earnings etc.

Many economic phenomena of interest, however, concern variables that are not continuous or perhaps not even quantitative.

- What characteristics (e.g. parental) affect the likelihood that an individual obtains a higher degree?
- What determines labour force participation (employed vs not employed)?
- What factors drive the incidence of civil war?

Today we will discuss **binary choice models**. These are central models in applied econometrics. Obviously binary choice models are useful when our outcome variable of interest is binary - a common situation in applied work. Moreover, the binary choice model is often used as an **ingredient** in other models. For example:

- In **propensity score matching** models (to be covered in lecture 3), we identify the average treatment effect by comparing outcomes of treated and non-treated individuals who, a priori, have similar probabilities of being treated. The probability of being treated is typically modelled using probit.
- In **Heckman's selection model**, we use probit in the first stage to predict the likelihood that someone is included (selected) in the sample. We then control for the likelihood of being selected when estimating our equation of interest (e.g. a wage equation)

The binary choice model is also a good starting point if we want to study more complicated models. Later on in the course we will thus cover **extensions** of the binary choice model, such as models for multinomial or ordered response, and models combining continuous and discrete outcomes (e.g. corner response models).

Useful references for this lecture:

Greene, W (2008). *Econometric Analysis*, 6th edition.

Angrist, Joshua and Jörn-Stefen Pischke (2009). *Mostly Harmless Econometrics. An Empiricist's Companion.* Chapter 3.4.2

In addition, for my empirical examples I will draw on material presented in the following paper:.

Kingdon, G. (1996) 'The quality and efficiency of private and public education: a case-study of urban India,' *Oxford Bulletin of Economics and Statistics* 58: 57-81

2. Binary Response

Whenever the variable that we want to model is binary, it is natural to think in terms of **probabilities**, e.g.

- 'What is the probability that an individual with such and such characteristics owns a car?'
- 'If some variable X changes by one unit, what is the effect on the probability of owning a car?'

When the dependent variable y is binary, it is typically equal to one for all observations in the data for which the event of interest has happened ('success') and zero for the remaining observations ('failure').

Provided we have a random sample, the sample mean of this binary variable is an unbiased estimate of the unconditional probability that the event happens. That is, letting y denote our binary dependent variable, we have

$$\Pr(y = 1) = E(y) = \frac{\sum_i y_i}{N},$$

where N is the number of observations in the sample.

Estimating the unconditional probability is trivial, but usually not the most interesting thing we can do with the data. Suppose we want to analyze what factors 'determine' changes in the probability that y equals one. Can we use the classical linear regression framework to this end?

3. The Regression Approach

Consider the linear regression model

$$\begin{aligned} y &= \beta_1 + \beta_2 x_2 + ... + \beta_K x_K + u \\ &= \boldsymbol{x}\boldsymbol{\beta} + u, \end{aligned} \tag{3.1}$$

where $\boldsymbol{\beta}$ is a $K \times 1$ vector of parameters, $\boldsymbol{x}$ is a $N \times K$ matrix of explanatory variables, and u is a residual. Assume that the residual is uncorrelated with the regressors, i.e. endogeneity is not a problem. This allows us to use OLS to estimate the parameters of interest.

- To interpret the results, note that if we take expectations on both sides of the equation above we obtain

$$E\left(y|\boldsymbol{x};\boldsymbol{\beta}\right) = \boldsymbol{x}\boldsymbol{\beta}.$$

- Now, just like the unconditional probability that y equals one is equal to the unconditional expected value of y, i.e. $E\left(y\right) = \Pr\left(y = 1\right)$, the conditional probability that y equals one is equal to the conditional expected value of y:

$$\begin{aligned} \Pr\left(y = 1|\boldsymbol{x}\right) &= E\left(y|\boldsymbol{x};\boldsymbol{\beta}\right), \\ \Pr\left(y = 1|\boldsymbol{x}\right) &= \boldsymbol{x}\boldsymbol{\beta}. \end{aligned} \tag{3.2}$$

  Because probabilities must sum to one, it must also be that

$$\Pr\left(y = 0|x\right) = 1 - \boldsymbol{x}\boldsymbol{\beta}.$$

- Equation (3.2) is a **binary response model**. In this particular model the probability of success (i.e. $y = 1$) is a **linear** function of the explanatory variables in the vector $\boldsymbol{x}$. This is why using OLS with a binary dependent variable is called the **linear probability model** (LPM).

Notice that in the LPM the parameter β_j measures the change in the probability of 'success', resulting from a change in the variable x_j, holding other factors fixed:

$$\Delta \Pr(y = 1|\boldsymbol{x}) = \beta_j \Delta x_j.$$

This can be interpreted as a partial effect on the probability of 'success'.

EXAMPLE: Modelling the probability of going to a private, unaided school (PUA) in India.[1] See appendix, Table 1a.

3.1. Shortcomings of the Linear Probability Model

Clearly the LPM is straightforward to estimate, however there are some important shortcomings.

- One undesirable property of the LPM is that, if we plug in certain combinations of values for the independent variables into (3.2), we can get predictions either less than zero or greater than one. Of course a probability by definition falls within the (0,1) interval, so predictions outside this range are hard to interpret. This is not an unusual result; for instance, based on the above LPM results, there are 61 observations for which the predicted probability is larger than one and 81 observations for which the predicted probability is less than zero. That is, 16 per cent of the predictions fall outside the (0,1) interval in this application (see Figure 1 in the appendix, and the summary statistics for the predictions reported below the table).

- Angrist and Pischke (p.103): "...[linear regression] may generate fitted values outside the LDV boundaries. This fact bothers some researchers and has generated a lot of bad press for the linear probability model."

- A related problem is that, conceptually, it does not make sense to say that a probability is **linearly** related to a continuous independent variable for all possible values. If it were, then continually increasing this explanatory variable would eventually drive $P(y = 1|x)$ above one or below zero.

[1] The data for this example are taken from the study by Kingdon (1996).

For example, the model above predicts that an increase in parental wealth by 1 unit increases the probability of going to a PUA school by about 1 percentage point. This may seem reasonable for families with average levels of wealth, however in very rich or very poor families the wealth effect is probably smaller. In fact, when taken to the extreme our model implies that a hundred-fold increase in wealth increases the probability of going to a PUA by more than 1 which, of course, is impossible (the wealth variable ranges from 0.072 to 82 in the data, so such an comparison is not unrealistic).

- A third problem with the LPM - arguably less serious than those above - is that the residual is heteroskedastic by definition. Why is this? Because y takes the value of 1 or 0, the residuals in equation (3.1) can take only two values, conditional on x: $1-\beta x$ and $-\beta x$. Further, the respective probabilities of these events are βx and $1-\beta x$. Hence,

$$\begin{aligned} var\left(u|\boldsymbol{x}\right) &= \Pr\left(y=1|\boldsymbol{x}\right)\left[1-\boldsymbol{x\beta}\right]^2 \\ &\quad + \Pr\left(y=0|\boldsymbol{x}\right)\left[-\boldsymbol{x\beta}\right]^2 \\ &= \boldsymbol{x\beta}\left[1-\boldsymbol{x\beta}\right]^2 + \left(1-\boldsymbol{x\beta}\right)\left[-\boldsymbol{x\beta}\right]^2 \\ &= \boldsymbol{x\beta}\left[1-\boldsymbol{x\beta}\right], \end{aligned}$$

  which clearly varies with the explanatory variables $\boldsymbol{x}$. The OLS estimator is still unbiased, but the conventional formula for estimating the standard errors, and hence the t-values, will be wrong. The easiest way of solving this problem is to obtain estimates of the standard errors that are robust to heteroskedasticity.

- EXAMPLE continued: Appendix - LPM with robust standard errors, Table 1b; compare to LPM with non-robust standard errors (Table 1a).

- A fourth and related problem is that, because the residual can only take two values, it cannot be normally distributed. The problem of non-normality means that OLS point estimates are unbiased

but its violation does mean that inference in small samples cannot be based on the usual suite of normality-based distributions such as the t test.

Summarizing:

- The LPM can be useful as a first step in the analysis of binary choices, but awkward issues arise if we want to argue that we are modelling a probability.
- As we shall see next, probit and logit solve these particular problems. Nowadays, these are just as easy to implement as LPM/OLS - but they are less straightforward to interpret.
- However, LPM remains a reasonably popular modelling framework because certain econometric problems are easier to address within the LPM framework than with probits and logits.
- If, for whatever reason, we use the LPM, it is important to recognize that it tends to give better estimates of the partial effects on the response probability near the centre of the distribution of $\boldsymbol{x\beta}$ than at extreme values (i.e. close to 0 and 1). The LPM graph in the appendix illustrates this (Figure 1).

3.2. Logit and Probit Models for Binary Response

The two main problems with the LPM were: nonsense predictions are possible (there is nothing to bind the value of Y to the (0,1) range); and linearity doesn't make much sense conceptually.

To address these problems we abandon the LPM and thus the OLS approach to estimating binary response models. Consider instead a class of binary response models of the form

$$\begin{aligned} \Pr(y=1|\boldsymbol{x}) &= F(\beta_1 + \beta_2 x_2 + ... + \beta_K x_K) \\ \Pr(y=1|\boldsymbol{x}) &= F(\boldsymbol{x\beta}), \end{aligned} \tag{3.3}$$

where F is a function taking on values strictly between zero and one: $0 < F(z) < 1$, for all real numbers z. The model (3.3) is often referred to in general terms as an **index model**, because $\Pr(y=1|x)$ is a

function of the vector x only through the **index**

$$\boldsymbol{x\beta} = \beta_1 + \beta_2 x_2 + ... + \beta_k x_k,$$

which is simply a scalar. Notice that $0 < F(\boldsymbol{x\beta}) < 1$ ensures that the estimated response probabilities are strictly between zero and one, which thus addresses the main worries of using LPM. F is usually a **cumulative density function** (cdf), monotonically increasing in the index z (i.e. $\boldsymbol{x\beta}$), with

$$\begin{aligned} \Pr(y = 1|\boldsymbol{x}) &\rightarrow 1 \text{ as } \boldsymbol{x\beta} \rightarrow \infty \\ \Pr(y = 1|\boldsymbol{x}) &\rightarrow 0 \text{ as } \boldsymbol{x\beta} \rightarrow -\infty. \end{aligned}$$

It follows that F must be a non-linear function, and hence we cannot use OLS.

Various non-linear functions for F have been suggested in the literature. By far the most common ones are the logistic distribution, yielding the **logit** model, and the standard normal distribution, yielding the **probit** model.

The logit model:

$$\Pr(y = 1|\boldsymbol{x}) = \frac{\exp(\boldsymbol{x\beta})}{1 + \exp(\boldsymbol{x\beta})} = \Lambda(\boldsymbol{x\beta}),$$

which is between zero and one for all values of $\boldsymbol{x\beta}$ (recall that $\boldsymbol{x\beta}$ is a scalar). This is the cumulative distribution function (CDF) for a logistic variable.

The probit model:

$$F(\boldsymbol{x\beta}) = \Phi(\boldsymbol{x\beta}) \equiv \int_{-\infty}^{\boldsymbol{x\beta}} \phi(v)\, dv,$$

where

$$\phi(v) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{v^2}{2}\right),$$

is the standard normal density. This choice of F also ensures that the probability of 'success' is strictly between zero and one for all values of the parameters and the explanatory variables.

EXAMPLE: See graphs in Figure 2, appendix.

The logit and probit functions are both increasing in $\boldsymbol{x\beta}$. Both functions increase relatively quickly at $\boldsymbol{x\beta} = 0$, while the effect on F at extreme values of $\boldsymbol{x\beta}$ tends to zero. The latter result ensures that the **partial effects** of changes in explanatory variables are **not constant**, a concern we had with the LPM. Also notice that the standard normal CDF has a shape very similar to of the logistic CDF, suggesting that it doesn't much matter which one of the two we choose to use in our analysis. I will come back to this point later.

Interpretation of probit and logit estimates is less straightforward than what we are used to for linear regression. Note in particular that the **marginal effects -** the effects on the response probability $\Pr(y = 1|x)$ resulting from a change in one of the explanatory variables - cannot be read off the parameter vector $\boldsymbol{\beta}$ directly. Let's look at how to compute marginal effects in a few cases.

3.2.1. Case I: The explanatory variable is continuous.

- In linear models the marginal effect of a unit change in some explanatory variable on the dependent variable is simply the associated coefficient on the relevant explanatory variable.

- However, for logit and probit models obtaining measures of the marginal effect is more complicated (which should come as no surprise, as these models are non-linear). When x_j is a continuous variable, its partial effect on $\Pr(y = 1|x)$ is obtained from the partial derivative:

$$\begin{aligned} \frac{\partial \Pr(y = 1|x)}{\partial x_j} &= \frac{\partial F(\boldsymbol{x\beta})}{\partial x_j} \\ &= f(\boldsymbol{x\beta}) \cdot \beta_j, \end{aligned}$$

  where

$$f(z) \equiv \frac{dF(z)}{dz}$$

  is the **probability density function** associated with F.

- Because the density function is non-negative, the partial effect of x_j will always have **the same**

sign as β_j.

- Notice that the partial effect depends on $f(\boldsymbol{x\beta})$; i.e. for different values of $x_1, x_2, ..., x_k$ the partial effect will be different. Hence, one has to take a stand on how to evaluate the marginal effects.
  - One possibility is to evaluate marginal effects at the **sample mean** values of $x_1, x_2, ..., x_k$. This is what Stata command 'mfx compute' does (unless you tell it otherwise). This command also provides standard errors for the marginal effects - more on this below.
  - Alternatively, you could compute marginal effects for each observation in the sample and average them - this gives you the average marginal effect.
  - Or, you could evaluate them anywhere you like, depending on what kind of argument you want to make (e.g. suppose income is an explanatory variable, and suppose you want to say something about the effect among low-income people - then it makes sense to evaluate the marginal effect at a low income level.
- Can you see at what values of $\boldsymbol{x\beta}$ the partial (or marginal) effect will be relatively small/large? See graphs of the standard normal and the logistic CDFs in handout.

EXAMPLE: Suppose we use the Indian data introduced above to estimate a probit modelling the probability that a child goes to a private unaided school as a function of the child's ability, measured by the score on the Raven's test. For simplicity, abstract from other explanatory variables. Our model is thus:

$$\Pr(pua = 1|sraven) = \Phi(\beta_0 + \beta_1 sraven).$$

The probit results are

	coef.	t-value
β_0	-1.82	12.84
β_1	0.050	11.76

Since the coefficient on *sraven* is positive, we know that the marginal effect must be positive. Treating

sraven as a continuous variable, it follows that the marginal effect is equal to

$$\begin{aligned}\frac{\partial \Pr\left(pua = 1|sraven\right)}{\partial sraven} &= \phi\left(\beta_0 + \beta_1 \cdot sraven\right)\beta_1 \\ &= \phi\left(-1.82 + 0.05 \cdot sraven\right)0.05,\end{aligned}$$

where $\phi\left(.\right)$ is the standard normal density function:

$$\phi\left(z\right) = \frac{1}{\sqrt{2\pi}}\exp\left(-z^2/2\right).$$

We see straight away that the marginal effect depends on the level of *sraven*. We see from the summary statistics that the mean value of *sraven* is about 31, so let's evaluate the marginal effect at $sraven = 31$:

$$\begin{aligned}&\frac{\partial \Pr\left(pua = 1|sraven = 31\right)}{\partial sraven} \\ = \;& \frac{1}{\sqrt{2\pi}}\exp\left(-\left(-1.82 + 0.05 \cdot 31\right)^2/2\right)0.05 \\ = \;& 0.019,\end{aligned}$$

Evaluated at the mean of *sraven*, we see that the results imply that an increase in *sraven* by one unit raises the probability of going to a private school by about two percentage points. At lower levels of *sraven*, the marginal effect is smaller:

$$\begin{aligned}&\frac{\partial \Pr\left(pua = 1|sraven = 15\right)}{\partial sraven} \\ = \;& \frac{1}{\sqrt{2\pi}}\exp\left(-\left(-1.82 + 0.05 \cdot 20\right)^2/2\right)0.05 \\ = \;& 0.011.\end{aligned}$$

Of course, the fact that the marginal effect is smaller at lower levels reflects the non-linearity of the probit model (again: see graphs in Figure 2 in handout).

STUDENT EXERCISE: Now consider **logit**:

$$\Pr(pua = 1|sraven) = \Lambda(\beta_0 + \beta_1 sraven),$$

$$\Lambda(z) = \frac{\exp(z)}{1 + \exp(z)}.$$

The logit results are

	coef.	t-value
β_0	-3.07	12.00
β_1	0.084	11.20

Task: Calculate and interpret the marginal effect. Compare the result to the probit marginal effect.

3.2.2. Case II: The explanatory variable is discrete.

If x_j is a discrete variable then we should not rely on calculus in evaluating the effect on the response probability. To keep things simple, suppose x_2 is binary. In this case the partial effect from changing x_2 from zero to one, holding all other variables fixed, is

$$F(\beta_1 + \beta_2 \cdot 1 + ... + \beta_K x_K) - F(\beta_1 + \beta_2 \cdot 0 + ... + \beta_K x_K).$$

Again this depends on all the values of the other explanatory variables and the values of all the other coefficients.

Again, knowing the **sign** of β_2 is sufficient for determining whether the effect is positive or not, but to find the **magnitude** of the effect we have to use the formula above.

The Stata command 'mfx compute' can spot dummy explanatory variables. In such a case it will use the above formula for estimating the partial effect.

3.2.3. Case III: Non-linear explanatory variables.

Suppose the model is

$$\Pr(y = 1|x) = F\left(\beta_1 + \beta_2 x_2 + \beta_2 x_3 + \beta_{22} x_2^2\right),$$

where x_2^2 is a continuous variable. What is the marginal effect of x_2 on the response probability?

4. Latent Regression - Index Function Models

As we have seen, the probit and logit models resolve some of the problems with the LPM model. The key, really, is the specification

$$\Pr(y = 1|x) = F(\boldsymbol{x\beta}),$$

where F is the cdf for either the standard normal or the logistic distribution, because with any of these models we have a functional form that is easier to defend than the linear model.

The traditional way of introducing probits and logits in econometrics, however, is not as a response to a functional form problem. Instead, probits and logits are traditionally viewed as models suitable for estimating parameters of interest when the dependent variable is not fully observed. Let's have a look at this perspective.

Let y^* be a continuous variable that we do not observe - a **latent variable** - and assume y^* is determined by the model

$$\begin{aligned} y^* &= \beta_1 + \beta_2 x_2 + ... + \beta_K x_K + e \\ &= \boldsymbol{x\beta} + e, \end{aligned} \tag{4.1}$$

where e is a residual, assumed uncorrelated with $\boldsymbol{x}$ (i.e. $\boldsymbol{x}$ is not endogenous). While we do not observe y^*, we do observe the discrete choice made by the individual, according to the following choice rule:

$$\begin{aligned} y &= 1 \text{ if } y^* > 0 \\ y &= 0 \text{ if } y^* \leq 0. \end{aligned}$$

Why is y^* unobserved? Think about y^* as representing net utility of, say, buying a car. The individual

undertakes a cost-benefit analysis and decides to purchase the car if the net utility is positive. We do not observe (because we cannot measure) the 'amount' of net utility; all we observe is the actual outcome of whether or not the individual does buy a car. (If we had data on y^* we could estimate the model (5.4) with OLS as usual.)

Now, we want to model the probability that a 'positive' choice is made (e.g. buying, as distinct from not buying, a car). Assuming that e follows a logistic distribution[2],

$$\begin{aligned}\lambda(e) &= \frac{\exp(-e)}{(1+\exp(-e))^2} \text{ (density)},\\ \Lambda(e) &= \frac{\exp(e)}{1+\exp(e)} \text{ (CDF)},\end{aligned}$$

it follows that

$$\begin{aligned}\Pr(y=1|x) &= \Pr(y^*>0|x)\\ &= \Pr(\boldsymbol{x\beta}+e>0|x)\\ &= \Pr(e>-\boldsymbol{x\beta})\\ &= 1-\Lambda(-\boldsymbol{x\beta}) \text{ (integrate)}\\ &= \Lambda(\boldsymbol{x\beta}) \text{ (exploit symmetry)}.\end{aligned}$$

Notice that the last step here exploits the fact that the logistic distribution is symmetric, so that $F(z) = 1-F(-z)$ for all z. This equation is exactly the binary response model (3.3) for the logit model. This is how the binary response model can be derived from an underlying latent variable model.

We can follow the same route to derive the probit model. Assume e follows a standard normal distribution

[2] Note that symmetry of the probability density function implies $\lambda(e) = \frac{\exp(e)}{(1+\exp(e))^2} = \frac{\exp(-e)}{(1+\exp(-e))^2}$. In some expositions you see $\frac{\exp(e)}{(1+\exp(e))^2}$, in others $\frac{\exp(-e)}{(1+\exp(-e))^2}$. Don't let this confuse you.

$$
\begin{aligned}
\Pr(y=1|\boldsymbol{x}) &= \Pr(y^*>0|\boldsymbol{x}) \\
&= \Pr(\boldsymbol{x\beta}+e>0|\boldsymbol{x}) \\
&= \Pr(e>-\boldsymbol{x\beta}) \\
&= 1-N\left(-\frac{\boldsymbol{x\beta}}{\sigma}\right) \text{ (integrate)} \\
&= \Phi(\boldsymbol{x\beta}),
\end{aligned}
$$

where again we exploit symmetry and use $\sigma = 1$ implied by the standard normal distribution. This is the binary response model (3.3) for the probit model.[3]

5. Estimation and Inference in Binary Choice Models

To estimate the LPM we can use OLS. Because of the non-linear nature of the probit and logit models (see graphs), however, linear estimators are not applicable for these. Instead we rely on **Maximum Likelihood (ML)** estimation. The principle of ML is very general and not confined to probit and logit models. Before turning the details of how ML is used to estimate probits and logits, here is an informal recap of ML.

5.1. Maximum Likelihood: Recap

- Suppose that, in the population, there is a variable w which is distributed according to some distribution $f(w;\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of unknown parameters.
- Suppose we have a random sample $\{w_1, w_2, ..., w_N\}$ drawn from the population distribution $f(w;\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is unknown.
- Our objective is to estimate $\boldsymbol{\theta}$. Our sample is more likely to have come from a population charac-

[3] The assumption that $\sigma = 1$ may appear restrictive. In fact, this is a necessary normalisation, because we cannot estimate σ by means of a binary response model.

terized by one particular set of parameter values, say $\tilde{\boldsymbol{\theta}}$, than from another set of parameter values, say $\breve{\boldsymbol{\theta}}$.

- The maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$ is simply the particular vector $\hat{\boldsymbol{\theta}}^{ML}$ that gives the **greatest likelihood** (or, if you prefer, probability) of observing the sample $\{w_1, w_2, ..., w_N\}$.

- **Random sampling** (an assumption) implies that $w_1, w_2, ..., w_N$ are independent of each other, hence the likelihood of observing $\{w_1, w_2, ..., w_N\}$ (i.e. the sample) is simply

$$L(\boldsymbol{\theta}; w_1, w_2, ..., w_N) = f(w_1; \boldsymbol{\theta}) f(w_2; \boldsymbol{\theta}) \cdot ... \cdot f(w_N; \boldsymbol{\theta}),$$

or, in more compact notation,

$$L(\boldsymbol{\theta}; w_1, w_2, ..., w_N) = \prod_{i=1}^{N} f(w_i; \boldsymbol{\theta}).$$

i.e. the **product** of the individual likelihoods. The equation just defined is a function of $\boldsymbol{\theta}$: for some values of $\boldsymbol{\theta}$ the resulting L will be relatively high while for other values of $\boldsymbol{\theta}$ it will be low. This is why we refer to equations of this form as **likelihood functions**.

- The value of $\boldsymbol{\theta}$ that gives the maximum value of the likelihood function is the maximum likelihood estimate of θ.

- For computational reasons it is much more convenient to work with the **log-likelihood function**:

$$\ln L(\boldsymbol{\theta}; w_1, w_2, ..., w_N) = \sum_{i=1}^{N} \ln f(w_i; \boldsymbol{\theta}).$$

The value of $\boldsymbol{\theta}$ that gives the maximum value of the log likelihood function is the $\hat{\boldsymbol{\theta}}^{ML}$.

5.2. Maximum likelihood estimation of logit and probit models

We now return to the logit and probit models. How can ML be used to estimate the parameters of interest in these models, i.e. $\boldsymbol{\beta}$? Assume that we have random sample of size N. The ML estimate of $\boldsymbol{\beta}$ is the particular vector $\hat{\boldsymbol{\beta}}^{ML}$ that gives the greatest likelihood of observing the sample $\{y_1, y_2, ..., y_N\}$, conditional on the explanatory variables $\boldsymbol{x}$.

By assumption, the probability of observing $y_i = 1$ is $F(\boldsymbol{x}\boldsymbol{\beta})$ while the probability of observing $y_i = 0$ is $1 - F(\boldsymbol{x}\boldsymbol{\beta})$. It follows that the probability of observing the entire sample is

$$L(y|\boldsymbol{x};\boldsymbol{\beta}) = \prod_{y_i=0} [1 - F(\boldsymbol{x}_i\boldsymbol{\beta})] \prod_{y_i=1} F(\boldsymbol{x}_i\boldsymbol{\beta}),$$

We can rewrite this as

$$L(y|\boldsymbol{x};\boldsymbol{\beta}) = \prod_{i=1}^{N} F(\boldsymbol{x}_i\boldsymbol{\beta})^{y_i} [1 - F(\boldsymbol{x}_i\boldsymbol{\beta})]^{(1-y_i)},$$

because when $y = 1$ we get $F(\boldsymbol{x}_i\boldsymbol{\beta})$ and when $y = 0$ we get $[1 - F(\boldsymbol{x}_i\boldsymbol{\beta})]$.

- The **log** likelihood for the sample is

$$\ln L(y|\boldsymbol{x};\boldsymbol{\beta}) = \sum_{i=1}^{N} \{y_i \ln F(\boldsymbol{x}_i\boldsymbol{\beta}) + (1 - y_i) \ln [1 - F(\boldsymbol{x}_i\boldsymbol{\beta})]\}.$$

  The MLE of β **maximizes** this log likelihood function.

- If F is the logistic CDF then we obtain the logit likelihood:

$$\begin{aligned}
\ln L(y|\boldsymbol{x};\boldsymbol{\beta}) &= \sum_{i=1}^{N} \{y_i \ln \Lambda(\boldsymbol{x}_i\boldsymbol{\beta}) + (1 - y_i) \ln [1 - \Lambda(\boldsymbol{x}_i\boldsymbol{\beta})]\} \\
\ln L(y|\boldsymbol{x};\boldsymbol{\beta}) &= \sum_{i=1}^{N} \left\{ y_i \ln \left(\frac{\exp(\boldsymbol{x}_i\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i\boldsymbol{\beta})} \right) + (1 - y_i) \ln \left(\frac{1}{1 + \exp(\boldsymbol{x}_i\boldsymbol{\beta})} \right) \right\},
\end{aligned}$$

which simplifies to

$$\ln L\left(y|\boldsymbol{x};\boldsymbol{\beta}\right)=\sum_{i=1}^{N}\left\{y_i\left[\boldsymbol{x}_i\boldsymbol{\beta}-\ln\left(1+\exp\left(\boldsymbol{x}_i\boldsymbol{\beta}\right)\right)\right]-\left(1-y_i\right)\ln\left(1+\exp\left(\boldsymbol{x}_i\boldsymbol{\beta}\right)\right)\right\}.$$

- If F is the standard normal CDF we get the probit estimator:

$$\ln L\left(y|\boldsymbol{x};\boldsymbol{\beta}\right)=\sum_{i=1}^{N}\left\{y_i\ln\Phi\left(\boldsymbol{x}_i\boldsymbol{\beta}\right)+\left(1-y_i\right)\ln\left[1-\Phi\left(\boldsymbol{x}_i\boldsymbol{\beta}\right)\right]\right\}.$$

How maximize the log likelihood function?

- Sample log likelihood:

$$\ln L\left(y|\boldsymbol{x};\boldsymbol{\beta}\right)=\sum_{i=1}^{N}\left\{y_i\ln F\left(\boldsymbol{x}_i\boldsymbol{\beta}\right)+\left(1-y_i\right)\ln\left[1-F\left(\boldsymbol{x}_i\boldsymbol{\beta}\right)\right]\right\}.$$

- Because the objective is to maximize the log likelihood function with respect to the parameters in the vector $\boldsymbol{\beta}$, it must be that, at the maximum, the following K first order conditions will hold:

$$\sum_{i=1}^{N}\left\{y_i\frac{f\left(\boldsymbol{x}_i\boldsymbol{\beta}\right)}{F\left(\boldsymbol{x}_i\boldsymbol{\beta}\right)}+\left(1-y_i\right)\frac{f\left(\boldsymbol{x}_i\boldsymbol{\beta}\right)}{\left[1-F\left(\boldsymbol{x}_i\boldsymbol{\beta}\right)\right]}\right\}\boldsymbol{x}_i=0.$$

1 x 1 1 x K

Greene refers to these as **likelihood equations**.

- It is typically not possible to solve analytically for $\boldsymbol{\beta}$ here. Instead, to obtain parameter estimates, we rely on some sophisticated iterative 'trial and error' technique. There are lots of **algorithms** that can be used, but we will not study these here. The most common ones are based on first and sometimes second derivatives of the log likelihood function. Think of a blind man walking up a hill, and whose only knowledge of the hill comes from what passes under his feet. Provided the hill is strictly concave, the man should have no trouble finding the top. Fortunately the log likelihood

functions for logit and probit are concave, but this is not always the case for other models.

EXAMPLE: Appendix, Tables 2-3.

5.3. Hypothesis Tests

5.3.1. Inference based on the log likelihood function

- We have already discussed how the ML estimates of the parameters are those that maximize the likelihood of observing the sample. It must then be that **all other parameter values** - which, by definition, are not the ML estimates - will result in a **lower** (worse) log likelihood value.

- Now let's revisit our Indian dataset and investigate what happens to the log likelihood value if we change the value of the coefficient on *sraven*. See Figure 3 in the handout.

- As expected, values of b_raven not equal to 0.03 produce lower log likelihood values.

- Is it important **how much** the log L falls as a result of moving b_sraven a given distance away from the ML estimate of 0.03? Yes, very important, because this, essentially, **is** the general basis for our inference. Think about the log likelihood ratio test.

- The **log likelihood ratio** test is defined as two times the difference in two log likelihood values:

$$LR = -2\left(\ln L_R - \ln L_U\right),$$

where $\ln L_U$ is the log likelihood value for the **unrestricted** model and $\ln L_R$ is the log likelihood value for the **restricted** model. LR follows a chi-squared distribution with q degrees of freedom under H_0, where q is the number of restrictions.

- Suppose now I want to test the following null hypothesis:

$$H_0 : b\_sraven = 0.$$

Looking at Figure 3, I see that

$$\ln L_R \simeq -356,$$

(i.e. this is the log likelihood value associated with $b\_sraven = 0$) and I know from the regression output (or from the graph) that

$$\ln L_U = -340.4.$$

Hence

$$LR = 2\left(-340.4 + 356\right) = 31.2.$$

To test H_0 at the 5% level we use as our critical value the 95^{th} percentile in the χ^2_q distribution. With $q = 1$ (because there is only one restriction here) the critical value is 3.84, so I firmly reject the null hypothesis at the 5% level. If you want a specific p-value, we can type

$$chiprob(1, 31.2)$$

in Stata which is equal to 0.00000002. We can thus reject the null at any conventional level of significance.

- Key point: It is the **large fall** in the log L resulting from imposing b_sraven = 0 that enables us to reject the null hypothesis. Had the log likelihood function been flatter in b_sraven, we might not have been able to reject the null hypothesis.

- The log likelihood ratio is often used to test whether a sub-set of the explanatory variables can be omitted from the model. Again the idea is that, because ML maximizes the log likelihood function, dropping variables will lead to a lower log likelihood value (this is similar to the result that the R-squared falls when variables are dropped from an OLS regression). The question is whether the fall in the log likelihood is large enough to conclude that the dropped variables are important. The

likelihood ratio statistic:

$$LR = -2\left(\ln L_R - \ln L_U\right),$$

where $\ln L_U$ is the log likelihood value for the unrestricted model, e.g.

$$F\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4\right),$$

and $\ln L_R$ is the value for the restricted model, e.g.

$$F\left(\beta_0 + \beta_1 x_1\right).$$

So estimate these two models and compare the two log likelihood values. We can obtain p-values directly in Stata by using the command

$$chiprob(q, LR).$$

What's q in this case?

- Example in appendix, Table 8.
- In Table 2 in the appendix, how should we interpret the information in 'LR chi2(9)'?

5.3.2. Standard errors for parameters

In linear models the conventional covariance matrix is given by $\sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$ which is straightforward to estimate. In non-linear models, however, such as the probit and the logit, deriving formulas for the covariance matrix, and hence the standard errors, is more complicated. The conventional estimator for the covariance matrix is based on the inverse of the negative Hessian:

$$Var\left(\hat{\boldsymbol{\beta}}^{ML}\right) = -H^{-1}, \tag{5.1}$$

where the Hessian is the matrix of second order derivatives of the log likelihood function:

$$H(\boldsymbol{\beta}) = \frac{\partial^2 \ln L\left(y|\boldsymbol{x}_i; \hat{\boldsymbol{\beta}}^{ML}\right)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}.$$

Note that provided the log likelihood function is concave, the second derivative is negative which ensures that the variance is positive.

- This is (somewhat) intuitive. Note that the second derivative of the log likelihood function with respect to β (evaluated at $\hat{\boldsymbol{\beta}}^{ML}$) measures the **curvature** of the log likelihood function - and so variance formula (5.1) says that **the more curvature, the lower is the variance.**

- Recall that how big is the quantitative fall in the log L as a result of imposing other parameter values than the ML estimate, is central for our inference if we use the log likelihood ratio test. Clearly curvature plays a central role here: with little curvature you have to move the parameter value for b_sraven a long way away from its ML estimate before the LR test rejects, but with a lot of curvature you will not have to move far. So you see the variance and the LR test are very closely related.

- Sometimes you see 'robust' standard errors reported - these are obtained from the 'sandwich' formula:

$$Var\left(\hat{\boldsymbol{\beta}}^{ML}\right) = \left[-H\left(\hat{\boldsymbol{\beta}}^{ML}\right)\right]^{-1} Var\left[\boldsymbol{s}\left(\hat{\boldsymbol{\beta}}^{ML}\right)\right]\left[-H\left(\hat{\boldsymbol{\beta}}^{ML}\right)\right]^{-1},$$

  where

$$\boldsymbol{s}\left(\hat{\boldsymbol{\beta}}^{ML}\right) \equiv \frac{\partial \ln L\left(y|\boldsymbol{x}; \hat{\boldsymbol{\beta}}^{ML}\right)}{\partial \hat{\boldsymbol{\beta}}^{ML}}$$

  is the **gradient** vector, or **score** vector.

5.3.3. Standard errors for marginal effects

- Once we have estimated the variance matrix, we can calculate standard errors by taking the square root of the diagonal elements of the covariance matrix, and subsequently obtain t-values and con-

fidence intervals in the usual ways.

- We can also calculate standard errors for the **marginal effects** (recall these are non-linear functions of the parameters). The Stata mfx command does this for us using the **delta method**, which involves transforming the standard errors of $\hat{\boldsymbol{\beta}}^{ML}$ into standard errors of $\frac{\partial \Pr(y=1|\boldsymbol{x})}{\partial x_j}$ by means of a Taylor series approximation. Here is how it works:

  - Our goal is to estimate the standard error of the marginal effect $\frac{\partial \Pr(y=1|\boldsymbol{x})}{\partial x_j}$. Define

$$\frac{\partial \Pr(y=1|\boldsymbol{x})}{\partial x_j} \equiv h(\boldsymbol{\beta}),$$

    making it explicit that the marginal effect is a function of the parameters $\boldsymbol{\beta}$. We have obtained an estimate of $\boldsymbol{\beta}$, denoted $\hat{\boldsymbol{\beta}}^{ML}$. We have also estimated the covariance matrix $Var\left(\hat{\boldsymbol{\beta}}^{ML}\right)$. We now need to obtain $Var\left(\hat{\boldsymbol{\gamma}}_j^{ML}\right)$.

  - Now define $\hat{\gamma}_j^{MLE} = h\left(\hat{\boldsymbol{\beta}}^{MLE}\right)$, and then take a Taylor series approximation of $\hat{\gamma}_j^{MLE} = h\left(\hat{\boldsymbol{\beta}}^{MLE}\right)$ around the true value $\boldsymbol{\beta}$:

$$\hat{\gamma}_j^{MLE} \simeq \gamma_j + \sum_{i=1}^{K} \frac{\partial h}{\partial \beta_i}\left(\hat{\beta}_i^{MLE} - \beta_i\right).$$

    In matrix notation

$$\hat{\boldsymbol{\gamma}}^{MLE} - \boldsymbol{\gamma} \simeq \Psi\left(\hat{\beta}^{MLE} - \beta\right), \tag{5.2}$$

    where

$$\Psi = \begin{bmatrix} \frac{\partial h_1}{\partial \beta_1} & \frac{\partial h_1}{\partial \beta_2} & \cdots & \frac{\partial h_1}{\partial \beta_K} \\ \frac{\partial h_2}{\partial \beta_1} & \frac{\partial h_2}{\partial \beta_2} & \cdots & \frac{\partial h_2}{\partial \beta_K} \\ .. & & & \\ \frac{\partial h_J}{\partial \beta_1} & \cdots & & \frac{\partial h_J}{\partial \beta_K} \end{bmatrix}$$

    is a $J \times K$ matrix of derivatives. Post-multiply (5.2) by the transpose of (5.2), and take

expectations, and you get the variance matrix for the marginal effects:

$$Var\left(\hat{\boldsymbol{\gamma}}^{MLE}\right) = \Psi Var\left(\hat{\boldsymbol{\beta}}^{ML}\right)\Psi'.$$

5.4. Specification Tests for Binary Choice Models

A lot has been written about the problems posed by heteroskedasticity for the probit and logit models. You often hear statements to the effect that probit and logit estimates are **inconsistent** in the presence of heteroskedasticity. Greene (2006, p. 787) argues that this is a serious problem "because the probit model is most often used with microeconomic data, which are frequently heteroscedastic". What is the nature of the problem? Consider the following illustration:[4]

Start from a latent variable model with one explanatory variable x_{i1}:

$$y_i^* = \psi_0 + \psi_1 x_{i1} + u_i. \tag{5.3}$$

Suppose the residual u_i is heteroskedastic. Consider the following - admittedly very special and arguably peculiar - form of heteroskedasticity:

$$u_i \sim Normal\left(0, x_{i1}^2\right).$$

Recall that we do not observe y_i^* - all we observe is the binary dependent variable:

$$\begin{aligned} y_i &= 1 \text{ if } y_i^* > 0 \\ y_i &= 0 \text{ if } y_i^* \leq 0. \end{aligned}$$

Thus,

$$y_i = 1 \text{ if } \psi_0 + \psi_1 x_{i1} + u_i > 0.$$

[4] This is taken from Section 15.7.4 in Wooldridge (2002) "Econometric Analysis of Cross Section and Panel Data".

What is the probability that $y = 1$? We have

$$\begin{aligned}
\Pr(y_i = 1|x_i) &= \Pr(y_i^* > 0|x_i) \\
&= \Pr(\psi_0 + \psi_1 x_{i1} + u_i > 0|x_i) \\
&= \Pr\left(\psi_0 + \psi_1 x_{i1} + \sqrt{x_{i1}^2} e_i > 0|x_i\right),
\end{aligned}$$

where e_i follows a standard normal distribution (i.e. with mean zero and variance equal to one). Hence,

$$\begin{aligned}
\Pr(y_i = 1|x_{i1}) &= \Pr\left(e_i > -\frac{1}{x_{i1}}(\psi_0 + \psi_1 x_{i1})\right) \\
&= 1 - \Phi\left(-\frac{1}{x_{i1}}(\psi_0 + \psi_1 x_{i1})\right) \text{ (integrate)} \\
&= \Phi\left(\frac{1}{x_{i1}}(\psi_0 + \psi_1 x_{i1})\right), \text{ (symmetry)} \\
&= \Phi\left(\psi_0 \frac{1}{x_{i1}} + \psi_1\right).
\end{aligned}$$

We now see how the presence of heteroskedasticity radically has **altered the functional form** of the probit model. Given that the underlying latent model is

$$y_i^* = \psi_0 + \psi_1 x_{i1} + u_i,$$

we might be tempted to specify the probit model as

$$\Pr(y_i = 1|x_i) = \Phi(\psi_0 + \psi_1 x_{i1}),$$

but this would **not** be the correct specification. This is quite important. Think about the partial effect of x_{i1}. The correct specification is

$$\Pr(y_i = 1|x_i) = \Phi\left(\psi_0 \frac{1}{x_{i1}} + \psi_1\right),$$

hence the correct marginal effect is

$$\frac{\partial \Pr\left(y_i = 1|x_{i1}\right)}{\partial x_{i1}} = \phi\left(\psi_0 \frac{1}{x_{i1}} + \psi_1\right)\left(-\psi_0\left(\frac{1}{x_{i1}}\right)^2\right).$$

Remarkably, the sign of the marginal effect is the opposite of that of ψ_0 - i.e. the constant in the latent variable model - and does not depend on the sign of ψ_1 - the slope coefficient on x_{i1} in the latent variable model. It follows that if ψ_0 and ψ_1 are both positive, the marginal effect of x_{i1} on the probability of 'success' has the **opposite sign** to the marginal effect of x_{i1} on the latent dependent variable y_i^*.

Of course the latter result is driven by the specific form of heteroskedasticity considered here, and should not be viewed as a general result. The main point is that if the residual in the latent variable model is heteroskedastic this alters the functional form. Exactly how depends on the form of heteroskedasticity.

Now, suppose you were to specify (incorrectly) the probit as

$$\Pr\left(y_i = 1|x_i\right) = \Phi\left(\eta_0 + \eta_1 x_{i1}\right).$$

Do you think your coefficient on x_1 would be a good estimate of the coefficient ψ_1 in the latent variable model

$$y_i^* = \psi_0 + \psi_1 x_{i1} + u_i \ ?$$

Answer: no. And this is an example of how the presence of heteroskedasticity leads to "inconsistent estimates" of the parameters in the latent variable model.

How can we proceed if we believe heteroskedasticity is a problem? One possibility is to use Stata's **hetprob** command, which estimates a generalized probit model:

$$\begin{aligned} y^* &= \beta_1 + \beta_2 x_2 + ... + \beta_K x_K + e \\ y^* &= \boldsymbol{x\beta} + e, \end{aligned} \tag{5.4}$$

where

$$\sigma_e^2 = [\exp(\boldsymbol{z\gamma})]^2,$$

where $\boldsymbol{z}$ is a vector of variables (not including a constant - since not identified) thought to affect the variance of e, and $\boldsymbol{\gamma}$ is the corresponding vector of coefficients. We obtain

$$\begin{aligned}
\Pr(y=1|\boldsymbol{x},\boldsymbol{z}) &= \Pr(y^*>0|\boldsymbol{x},\boldsymbol{z}) \\
&= \Pr(\boldsymbol{x\beta}+e>0|\boldsymbol{x},\boldsymbol{z}) \\
&= \Pr(\boldsymbol{x\beta}+\exp(\boldsymbol{z\gamma})\,u>0|\boldsymbol{x},\boldsymbol{z}),
\end{aligned}$$

where u follows a standard normal distribution (a normalization). Hence

$$\begin{aligned}
\Pr(y=1|\boldsymbol{x},\boldsymbol{z}) &= \Pr\left(u>\frac{-\boldsymbol{x\beta}}{\exp(\boldsymbol{z\gamma})}\right) \\
\Pr(y=1|\boldsymbol{x},\boldsymbol{z}) &= 1-N\left(-\frac{\boldsymbol{x\beta}}{\exp(\boldsymbol{z\gamma})}\right) \text{ (integrate)} \\
\Pr(y=1|\boldsymbol{x},\boldsymbol{z}) &= \Phi\left(\frac{\boldsymbol{x\beta}}{\exp(\boldsymbol{z\gamma})}\right).
\end{aligned}$$

Of course, if a variable x_k is included in both $\boldsymbol{x}$ and $\boldsymbol{z}$, the marginal effect is somewhat more involved:

$$\frac{\partial \Pr(y=1|\boldsymbol{x},\boldsymbol{z})}{\partial x_k} = \phi\left(\frac{\boldsymbol{x\beta}}{\exp(\boldsymbol{z\gamma})}\right)\left(\frac{\beta_k-(\boldsymbol{x\beta})\,\gamma_k}{\exp(\boldsymbol{z\gamma})}\right).$$

This shows that the sign of the marginal effect is not necessarily the same as the sign of β_k.

EXAMPLE: Heteroskedasticity in school choice probit. Appendix.

5.5. Measuring Goodness of Fit

In linear models where the dependent variable is continuous, we often rely on the R-squared as a measure of the goodness of fit of the model. If for some reason we use linear regression in a binary choice setting (i.e. LPM here), you will obviously get an estimate of the R-squared. However, you should probably not

pay too much attention to this statistic. Why?

Recall:

$$R^2 = \frac{var\left(\hat{y}\right)}{var\left(y\right)},$$

where $\hat{y}$ denotes the predictions from the regression. But remember the main problem with LPM is that linearity is an unattractive feature of the model - both conceptually and in the sense that nonsense probability predictions may result. Consequently, we should not take the predictions of the LPM too seriously and so any measures of how 'good' these predictions are, is of limited interest.

The two most common alternative measures of goodness of fit for binary choice models are the **percent correctly predicted**, and the **pseudo R-squared**.

Percent correctly predicted. To obtain the percent correctly predicted we begin by computing the estimated probability that y_i equals one for each observation in the sample. For the probit model, for instance, this is given by

$$Est.\Pr\left(y=1|x\right) = \Phi\left(\boldsymbol{x}\hat{\boldsymbol{\beta}}^{MLE}\right),$$

where $Est.$ denotes 'estimated'. We then say that the **predicted outcome** of y_i is one if $\Phi\left(\boldsymbol{x}\hat{\boldsymbol{\beta}}^{MLE}\right) > 0.50$ and zero otherwise. The percentages of times the predicted y_i matches the actual y_i is the per cent correctly specified. Note the difference between predicted outcome (which is binary, 0 or 1) and predicted probability (any number between 0 and 1).

The per cent correctly predicted is a useful measure in this context, but we need to be careful. Consider a case where out of 200 observations, 180 have $y_i = 0$. If, say, 150 of these are predicted to be zero we obtain 75% correct predictions, even if our model fails to predict any of the observations for which $y = 1$ correctly. This is not an uncommon outcome in practice. For this reason, it is a good idea to report the percentages (or frequencies) correctly predicted **for each of the two outcomes**.

- Appendix, Table 7.
- Note: Hard to say a priori what makes up a 'satisfactory' percentage of correct predictions.

Pseudo R-squared. Various pseudo R-squared measures for binary response models have been developed. The most common one is

$$\tilde{R}^2 = 1 - \frac{\ln L_{ur}}{\ln L_r},$$

where $\ln L_{ur}$ is the value of the log likelihood at the ML estimates (the 'unrestricted' model) and $\ln L_r$ is the log likelihood value for a 'restricted' model in which the only 'explanatory' variable is a constant.

What is the logic of using this formula? Notice that if our explanatory variables have no explanatory power at all, then $\ln L_r = \ln L_{ur}$ (why?). In this case we get $\tilde{R}^2 = 0$.

In contrast, if our model is doing very well indeed in predicting the actual observations of y, then the log likelihood value (of the unrestricted model) will approach zero from below, and hence $\tilde{R}^2$ will tend to one. Why?

Recall that the log likelihood function is

$$\ln L\left(y|x_i;\beta\right) = \sum_{i=1}^{N} \left\{ y_i \ln F\left(\boldsymbol{x}_i\boldsymbol{\beta}\right) + \left(1-y_i\right) \ln\left[1 - F\left(\boldsymbol{x}_i\boldsymbol{\beta}\right)\right] \right\}.$$

A very good model will be such that $F\left(\boldsymbol{x}_i\boldsymbol{\beta}\right)$ will be very close to one for all observations for which $y_i = 1$ and very close to zero for all observations for which $y_i = 0$. To illustrate the point, suppose $F\left(\boldsymbol{x}_i\boldsymbol{\beta}\right)$ is exactly one for all observations for which $y_i = 1$ and exactly zero for all observations for which $y_i = 0$ - i.e. the model predicts the dependent variable perfectly. In that extreme case, we have

$$\begin{aligned}
\ln L\left(y|\boldsymbol{x}_i;\boldsymbol{\beta}\right) &= \sum_{i=1}^{N} \left\{ y_i \ln 1 + \left(1-y_i\right) \ln\left[1-0\right] \right\} \\
&= \sum_{i=1}^{N} \left\{ y_i \cdot 0 + \left(1-y_i\right) \cdot 0 \right\} \\
&= 0,
\end{aligned}$$

and so

$$\begin{aligned} \tilde{R}^2 &= 1 - \frac{0}{\ln L_r} \\ &= 1. \end{aligned}$$

Notice that $\tilde{R}^2$ uses the same information as that underlying the log likelihood ratio test.

- See Table 2 in the appendix. Verify that the reported $\tilde{R}^2$ is consistent with the LR test.

PhD Programme: Econometrics II
Department of Economics, University of Gothenburg
Appendix Lecture 2
Måns Söderbom

Binary Choice Models

Application: School Choice in India

The data used below were kindly provided by Dr Geeta Kingdon. These data have been used in

Kingdon, G. (1996) 'The quality and efficiency of private and public education: a case-study of urban India,' *Oxford Bulletin of Economics and Statistics* 58: 57-81.

See Table 1 in the paper for details on how variables are defined.

Key variables and summary statistics:

```
Contains data from kingdon96.dta
  obs:           902
 vars:             9
 size:        36,080 (99.7% of memory free)
------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------
numsib          float  %9.0g                  Number of siblings
sraven          float  %9.0g                  Raven ability score
wealth          float  %9.0g                  Index of household asset value
male            float  %9.0g                  Gender dummy: male=1, female=0
lowcaste        float  %9.0g                  Low caste? yes=1,no=0
muslim          float  %9.0g                  Muslim? yes=1,no=0
medyrs          float  %9.0g                  Mother's education in years
sikhchr         float  %9.0g                  Sikh or Christian? yes=1,no=0
stype           float  %9.0g                  School type: 0=govt, 1=private
                                                aided, 2=private unaided
------------------------------------------------------------------------------
```

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      puaind |       902    .3991131    .4899878          0          1
      numsib |       902    3.988914    1.705215          1         11
      sraven |       902    30.52661    11.22551          3         57
      wealth |       902    24.25723    21.08854       .072         82
        male |       902    .5321508    .4992421          0          1
-------------+--------------------------------------------------------
    lowcaste |       902    .1330377    .3398039          0          1
      muslim |       902    .2184035    .4133919          0          1
      medyrs |       902    8.665188    4.954049          0         20
     medyrsq |       902    99.60089    79.36289          0        400
     sikhchr |       902    .0310421    .1735278          0          1
```

Now consider results from OLS, probit and logit using the Stata code in Box 1.

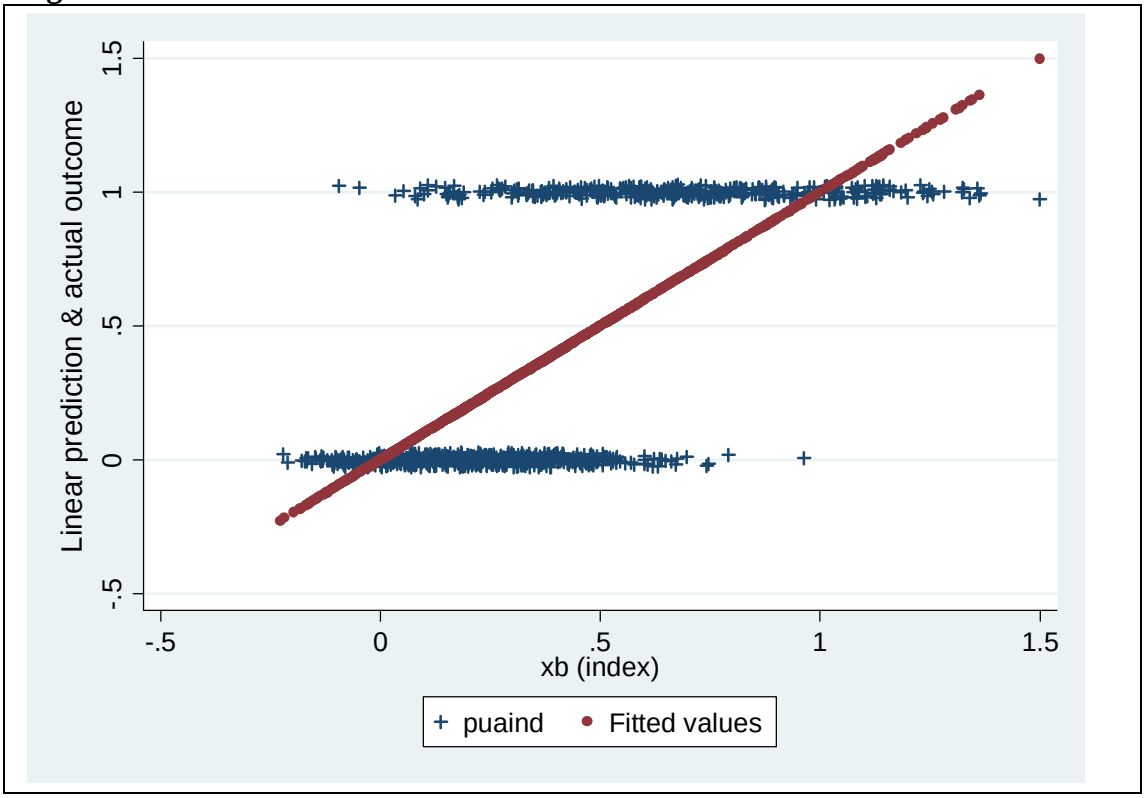
Box 1: Stata code for estimation of binary choice models

```
#delimit;

use kingdon96.dta, clear;

describe;
summarize;

tabstat numsib sraven wealth male ,
by(stype) s(mean p50 sd);

tabstat lowcaste muslim medyrs  sikhchr,
by(stype) s(mean p50 sd);

ge puaind=stype==2;
replace puaind=. if stype==.;

ge medyrsq = medyrs^2;

sum puaind numsib sraven wealth male lowcaste muslim medyrs medyrsq
sikhchr;

/*** LPM, probit, logit ***/

reg puaind numsib sraven wealth male lowcaste muslim medyrs medyrsq
sikhchr;
reg puaind numsib sraven wealth male lowcaste muslim medyrs medyrsq
sikhchr, robust;
predict yhat;           /* obtain predicted probability */
predict xb, xb;
label var xb "xb (index)";
scatter puaind yhat xb, symbol(+ o) jitter(2) l1title("Linear prediction &
actual outcome");

probit puaind numsib sraven wealth male lowcaste muslim medyrs medyrsq
sikhchr;
predict phat, p;        /* obtain predicted probability */

ge phat_d=phat>.5;      /* predicted outcome */
tab phat_d puaind;      /* compare predicted & acutal outcomes */

test sraven wealth;     /* Wald test, joint significance */

mfx compute;

logit puaind numsib sraven wealth male lowcaste muslim medyrs medyrsq
sikhchr;
predict lhat, p;        /* obtain predicted probability */

sum yhat phat lhat;
count if yhat<0;        /* number of negative predicted probabilities, LPM
*/
count if yhat>1;        /* number of predicted probabilities in excess of
one, LPM */

exit;
```

Table 1a. LINEAR PROBABILITY MODEL

```
> regress puaind numsib sraven wealth male lowcaste muslim medyrs medyrsq sikhchr;

      Source |       SS       df       MS              Number of obs =     902
-------------+------------------------------           F(  9,   892) =   85.25
       Model |  100.026088     9  11.1140097           Prob > F      =  0.0000
    Residual |  116.293203   892  .130373546           R-squared     =  0.4624
-------------+------------------------------           Adj R-squared =  0.4570
       Total |   216.31929   901  .240088003           Root MSE      =  .36107

------------------------------------------------------------------------------
      puaind |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      numsib |  -.0223168   .0082608    -2.70   0.007    -.0385297   -.0061038
      sraven |   .0075825   .0012103     6.27   0.000     .0052072    .0099578
      wealth |   .0101314   .0007259    13.96   0.000     .0087067    .0115561
        male |   .1732116   .0245567     7.05   0.000     .1250159    .2214072
    lowcaste |  -.1412188   .0392124    -3.60   0.000    -.2181782   -.0642594
      muslim |  -.1387535   .0321586    -4.31   0.000    -.2018689   -.0756381
      medyrs |  -.0245589   .0078772    -3.12   0.002    -.0400188   -.0090989
     medyrsq |   .0016972   .0005077     3.34   0.001     .0007008    .0026937
     sikhchr |    .220197   .0702409     3.13   0.002     .0823403    .3580538
       _cons |   .0047471   .0624763     0.08   0.939    -.1178706    .1273648
------------------------------------------------------------------------------
```

Figure 1: Predictions based on LPM shown in Table 1a



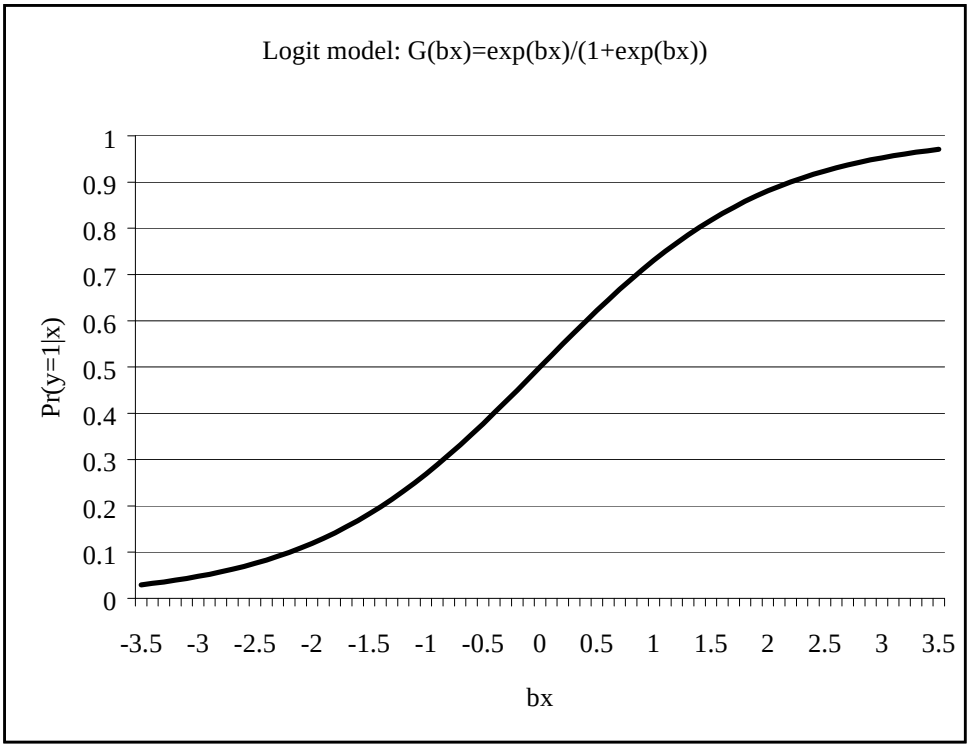
Note: The linear prediction is denoted yhat, and puaind is the actual binary dependent variable. The puaind variable has been "jittered" to facilitate interpretation.

```
. sum yhat;

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        yhat |       902    .3991131    .3331918  -.2283615   1.499489

. count if yhat>1;
   61

. count if yhat<0;
   81
```

Table 1b. LINEAR PROBABILITY MODEL WITH STANDARD ERRORS ROBUST TO HETEROSKEDASTICITY

```
> regress puaind numsib sraven wealth male lowcaste muslim medyrs medyrsq sikhchr,
> robust ;

Regression with robust standard errors                Number of obs =     902
                                                      F(  9,   892) =  189.05
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.4624
                                                      Root MSE      =  .36107

------------------------------------------------------------------------------
             |               Robust
      puaind |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      numsib |  -.0223168   .0084851    -2.63   0.009    -.0389699   -.0056636
      sraven |   .0075825     .00126     6.02   0.000     .0051096    .0100554
      wealth |   .0101314   .0006368    15.91   0.000     .0088817    .0113811
        male |   .1732116   .0241983     7.16   0.000     .1257193    .2207039
    lowcaste |  -.1412188   .0374195    -3.77   0.000    -.2146593   -.0677782
      muslim |  -.1387535   .0317488    -4.37   0.000    -.2010645   -.0764425
      medyrs |  -.0245589    .007757    -3.17   0.002     -.039783   -.0093347
     medyrsq |   .0016972    .000503     3.37   0.001       .00071    .0026845
     sikhchr |    .220197   .0775071     2.84   0.005     .0680796    .3723145
       _cons |   .0047471   .0618058     0.08   0.939    -.1165547    .1260488
------------------------------------------------------------------------------
```

Figure 2: The Logit and Probit CDFs



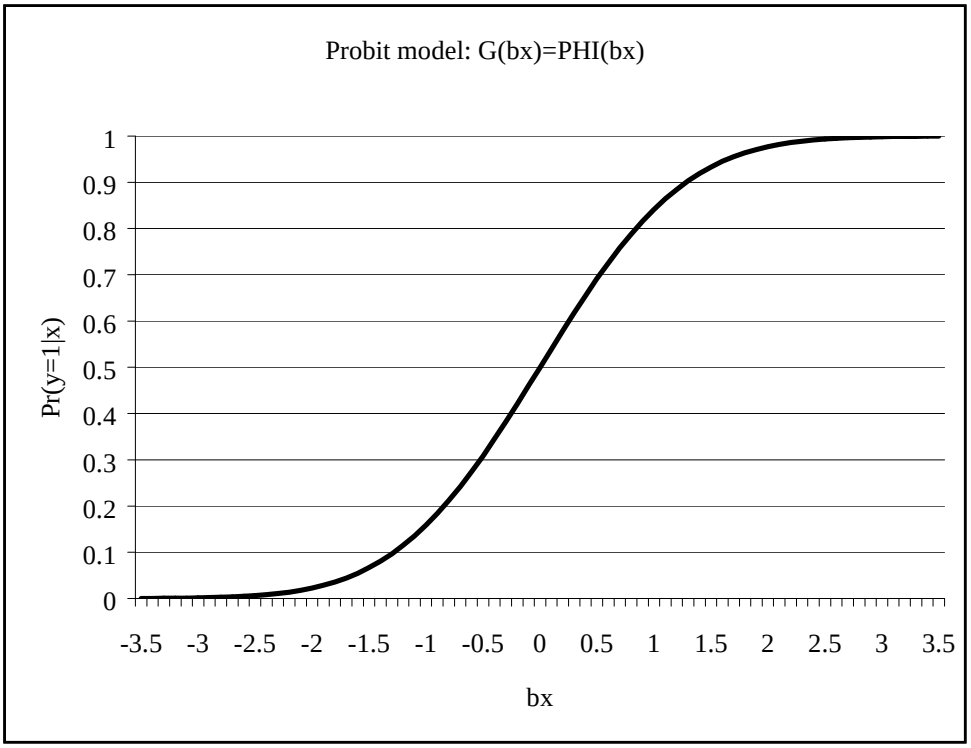


Table 2. PROBIT MODEL

```
> probit puaind numsib sraven wealth male lowcaste muslim medyrs medyrsq sikhchr;

Iteration 0:   log likelihood = -606.73067
Iteration 1:   log likelihood = -373.10677
Iteration 2:   log likelihood = -343.27331
Iteration 3:   log likelihood = -340.47774
Iteration 4:   log likelihood = -340.43889
Iteration 5:   log likelihood = -340.43888

Probit estimates                                  Number of obs   =        902
                                                  LR chi2(9)      =     532.58
                                                  Prob > chi2     =     0.0000
Log likelihood = -340.43888                       Pseudo R2       =     0.4389

------------------------------------------------------------------------------
      puaind |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      numsib |  -.0998298   .0382835    -2.61   0.009    -.1748641   -.0247956
      sraven |   .0301986   .0054653     5.53   0.000     .0194869    .0409103
      wealth |   .0461453   .0043108    10.70   0.000     .0376963    .0545943
        male |   .8575159   .1199153     7.15   0.000     .6224862    1.092546
    lowcaste |  -.5496526   .1865875    -2.95   0.003    -.9153575   -.1839478
      muslim |  -.7229197   .1530685    -4.72   0.000    -1.022929   -.4229109
      medyrs |  -.1260082   .0373075    -3.38   0.001    -.1991296   -.0528868
     medyrsq |   .0079365   .0024278     3.27   0.001     .0031781    .0126948
     sikhchr |   .8875504   .3272338     2.71   0.007      .246184    1.528917
       _cons |  -1.882662    .287822    -6.54   0.000    -2.446783   -1.318541
------------------------------------------------------------------------------

. /* marginal effects using mfx compute */
> mfx compute;

Marginal effects after probit
      y  = Pr(puaind) (predict)
         =  .38659838
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.     z    P>|z|  [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
  numsib |  -.0382063      .01465   -2.61   0.009  -.066911 -.009502   3.98891
  sraven |   .0115574      .00208    5.54   0.000   .007472  .015643   30.5266
  wealth |   .0176605      .00172   10.27   0.000   .014289  .021032   24.2572
    male*|   .3167018      .04152    7.63   0.000   .235328  .398076   .532151
lowcaste*|  -.1926379      .05762   -3.34   0.001  -.305563 -.079713   .133038
  muslim*|  -.2513949      .04612   -5.45   0.000  -.341793 -.160997   .218404
  medyrs |  -.0482251      .01433   -3.37   0.001  -.076308 -.020142   8.66519
 medyrsq |   .0030374      .00093    3.26   0.001   .001211  .004864   99.6009
 sikhchr*|   .3401752      .11123    3.06   0.002   .122159  .558191   .031042
------------------------------------------------------------------------------
(*) dy/dx is for discrete change of dummy variable from 0 to 1
```

Table 3. LOGIT MODEL

```
> logit puaind numsib sraven wealth male lowcaste muslim medyrs medyrsq sikhchr;

Iteration 0:   log likelihood = -606.73067
Iteration 1:   log likelihood = -373.59867
Iteration 2:   log likelihood = -342.99987
Iteration 3:   log likelihood =  -338.7387
Iteration 4:   log likelihood = -338.60417
Iteration 5:   log likelihood =   -338.604

Logit estimates                                   Number of obs   =        902
                                                  LR chi2(9)      =     536.25
                                                  Prob > chi2     =     0.0000
Log likelihood =   -338.604                       Pseudo R2       =     0.4419

------------------------------------------------------------------------------
      puaind |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      numsib |  -.1763513   .0702654    -2.51   0.012    -.3140689   -.0386336
      sraven |   .0566268   .0099437     5.69   0.000     .0371376    .0761161
      wealth |   .0812388   .0079356    10.24   0.000     .0656852    .0967923
        male |   1.573975   .2206061     7.13   0.000     1.141595    2.006355
    lowcaste |  -1.097276     .35503    -3.09   0.002    -1.793122   -.4014304
      muslim |   -1.30201   .2766134    -4.71   0.000    -1.844162   -.7598576
      medyrs |  -.2326064   .0666913    -3.49   0.000     -.363319   -.1018938
     medyrsq |   .0142063   .0043246     3.28   0.001     .0057302    .0226824
     sikhchr |   1.672074   .5791094     2.89   0.004     .5370403    2.807107
       _cons |  -3.388287   .5269294    -6.43   0.000    -4.421049   -2.355524
------------------------------------------------------------------------------

. /* marginal effects using mfx compute */
> mfx compute;

Marginal effects after logit
      y  = Pr(puaind) (predict)
         =  .36959733
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.     z    P>|z|  [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
  numsib |   -.041089      .01632   -2.52   0.012  -.073077 -.009101   3.98891
  sraven |   .0131938      .00229    5.76   0.000   .008701  .017686   30.5266
  wealth |   .0189282      .00196    9.65   0.000   .015082  .022774   24.2572
    male*|   .3480658      .04426    7.86   0.000   .261309  .434823   .532151
lowcaste*|  -.2195702      .05702   -3.85   0.000  -.331332 -.107809   .133038
  muslim*|    -.26307      .04577   -5.75   0.000  -.352781 -.173359   .218404
  medyrs |  -.0541962      .01563   -3.47   0.001  -.084826 -.023567   8.66519
 medyrsq |     .00331      .00101    3.27   0.001   .001324  .005296   99.6009
 sikhchr*|   .3900823      .10997    3.55   0.000   .174542  .605622   .031042
------------------------------------------------------------------------------
(*) dy/dx is for discrete change of dummy variable from 0 to 1
```

Table 4. PREDICTED PROBABILITIES: PROBIT AND LOGIT

```
. sum phat lhat;

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        phat |       902    .4023303    .3436492   .0009482    .999997
        lhat |       902    .3991131    .3473132   .0037879   .9996895
```

Note: phat = predicted probability based on probit model; lhat = predicted probability based on logit model.

Table **5. SIMPLE PROBIT MODEL: PUA = PHI(SRAVEN)**

```
Probit estimates                                  Number of obs   =        902
                                                  LR chi2(1)      =     152.66
                                                  Prob > chi2     =     0.0000
Log likelihood = -530.40283                       Pseudo R2       =     0.1258

------------------------------------------------------------------------------
      puaind |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      sraven |   .0499438   .0042474    11.76   0.000     .0416191    .0582685
       _cons |  -1.816083   .1414252   -12.84   0.000    -2.093271   -1.538894
------------------------------------------------------------------------------
```

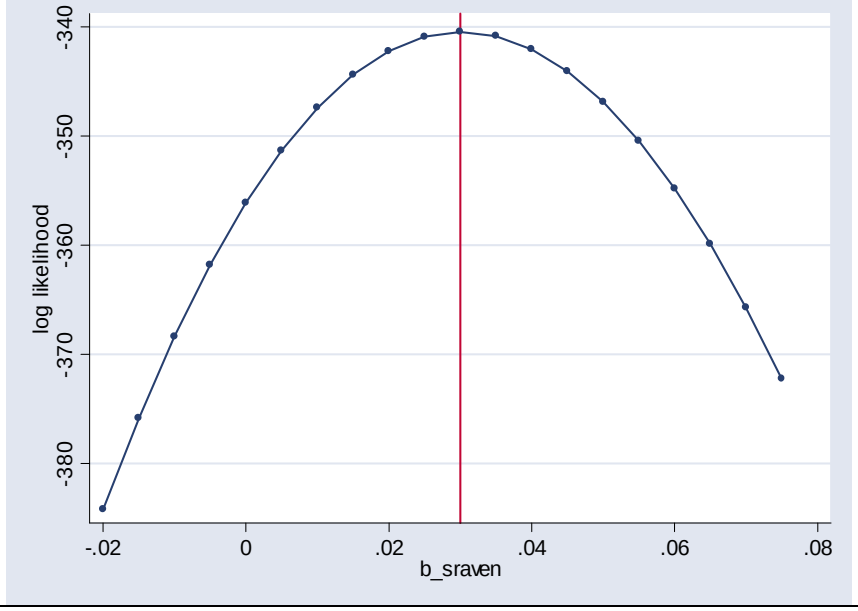
Table 6: The baseline probit model (same as Table 2)

```
Probit estimates                                  Number of obs   =        902
                                                  LR chi2(9)      =     532.58
                                                  Prob > chi2     =     0.0000
Log likelihood = -340.43888                       Pseudo R2       =     0.4389

------------------------------------------------------------------------------
      puaind |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      numsib |  -.0998298   .0382835    -2.61   0.009    -.1748641   -.0247956
      sraven |   .0301986   .0054653     5.53   0.000     .0194869    .0409103
      wealth |   .0461453   .0043108    10.70   0.000     .0376963    .0545943
        male |   .8575159   .1199153     7.15   0.000     .6224862    1.092546
    lowcaste |  -.5496526   .1865875    -2.95   0.003    -.9153575   -.1839478
      muslim |  -.7229197   .1530685    -4.72   0.000    -1.022929   -.4229109
      medyrs |  -.1260082   .0373075    -3.38   0.001    -.1991296   -.0528868
     medyrsq |   .0079365   .0024278     3.27   0.001     .0031781    .0126948
     sikhchr |   .8875504   .3272338     2.71   0.007      .246184    1.528917
       _cons |  -1.882662    .287822    -6.54   0.000    -2.446783   -1.318541
------------------------------------------------------------------------------
```

```
. test sraven wealth;

 ( 1)  sraven = 0
 ( 2)  wealth = 0

           chi2(  2) =  150.74
         Prob > chi2 =    0.0000
```

Now vary the coefficient on *sraven* around the ML estimate of 0.03 – see Figure 1.

Figure 3: The log-likelihood as function of b_sraven



As expected, values of b_raven not equal to 0.03 produce a lower log likelihood value. Is it important how much the log L falls as a result of moving b_sraven away from the ML estimate of 0.03?

Predictions:

```
. predict phat, p;
. ge phat_d=phat>.5;
. table phat_d pua;
```

Table 7: Frequencies of correct predictions

phat_d | puaind 0	puaind 1
0	491	87
1	51	273

Illustration of LR test

Estimate restricted model without sraven and wealth. Compare the resulting log likelihood value to that obtained in the unrestricted model (Table 2):

Table 8: Restricted probit: sraven and wealth omitted

```
Probit estimates                                  Number of obs   =        902
                                                  LR chi2(7)      =     311.29
                                                  Prob > chi2     =     0.0000
Log likelihood = -451.08324                       Pseudo R2       =     0.2565

------------------------------------------------------------------------------
      puaind |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      numsib |  -.1092096   .0343527    -3.18   0.001    -.1765395   -.0418796
        male |   .6587308   .1010961     6.52   0.000     .4605861    .8568756
    lowcaste |  -.5847738   .1689565    -3.46   0.001    -.9159225    -.253625
      muslim |  -.6309888   .1307152    -4.83   0.000    -.8871859   -.3747918
      medyrs |  -.0969399   .0336861    -2.88   0.004    -.1629634   -.0309163
     medyrsq |   .0127564   .0021721     5.87   0.000     .0084992    .0170135
     sikhchr |   .8240591   .3017656     2.73   0.006     .2326093    1.415509
       _cons |  -.4723225   .2159322    -2.19   0.029    -.8955419   -.0491031
------------------------------------------------------------------------------

. display 2*(-340.43888 - -451.08324 )
221.28872

. disp chiprob(2,221.29)
8.861e-49

(=0.0000000000…)
```

Heteroskedasticity in the school choice probit

Consider the benchmark probit model reported in Table 2 above:

```
Probit estimates                                  Number of obs   =        902
                                                  LR chi2(9)      =     532.58
                                                  Prob > chi2     =     0.0000
Log likelihood = -340.43888                       Pseudo R2       =     0.4389

------------------------------------------------------------------------------
      puaind |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      numsib |  -.0998298   .0382835    -2.61   0.009    -.1748641   -.0247956
      sraven |   .0301986   .0054653     5.53   0.000     .0194869    .0409103
      wealth |   .0461453   .0043108    10.70   0.000     .0376963    .0545943
        male |   .8575159   .1199153     7.15   0.000     .6224862    1.092546
    lowcaste |  -.5496526   .1865875    -2.95   0.003    -.9153575   -.1839478
      muslim |  -.7229197   .1530685    -4.72   0.000    -1.022929   -.4229109
      medyrs |  -.1260082   .0373075    -3.38   0.001    -.1991296   -.0528868
     medyrsq |   .0079365   .0024278     3.27   0.001     .0031781    .0126948
     sikhchr |   .8875504   .3272338     2.71   0.007      .246184    1.528917
       _cons |  -1.882662    .287822    -6.54   0.000    -2.446783   -1.318541
------------------------------------------------------------------------------
```

Now relax the assumption that the error term is homoskedastic, by writing the variance of the error term as [exp(g*sraven)]^2, where g is a parameter to be estimated (note: if g=0 we're back to the homoskedastic model). I can obtain results for this generalized model by using the hetprob command in Stata:

```
hetprob puaind numsib sraven wealth male lowcaste muslim medyrs medyrsq
sikhchr, het(sraven);
```

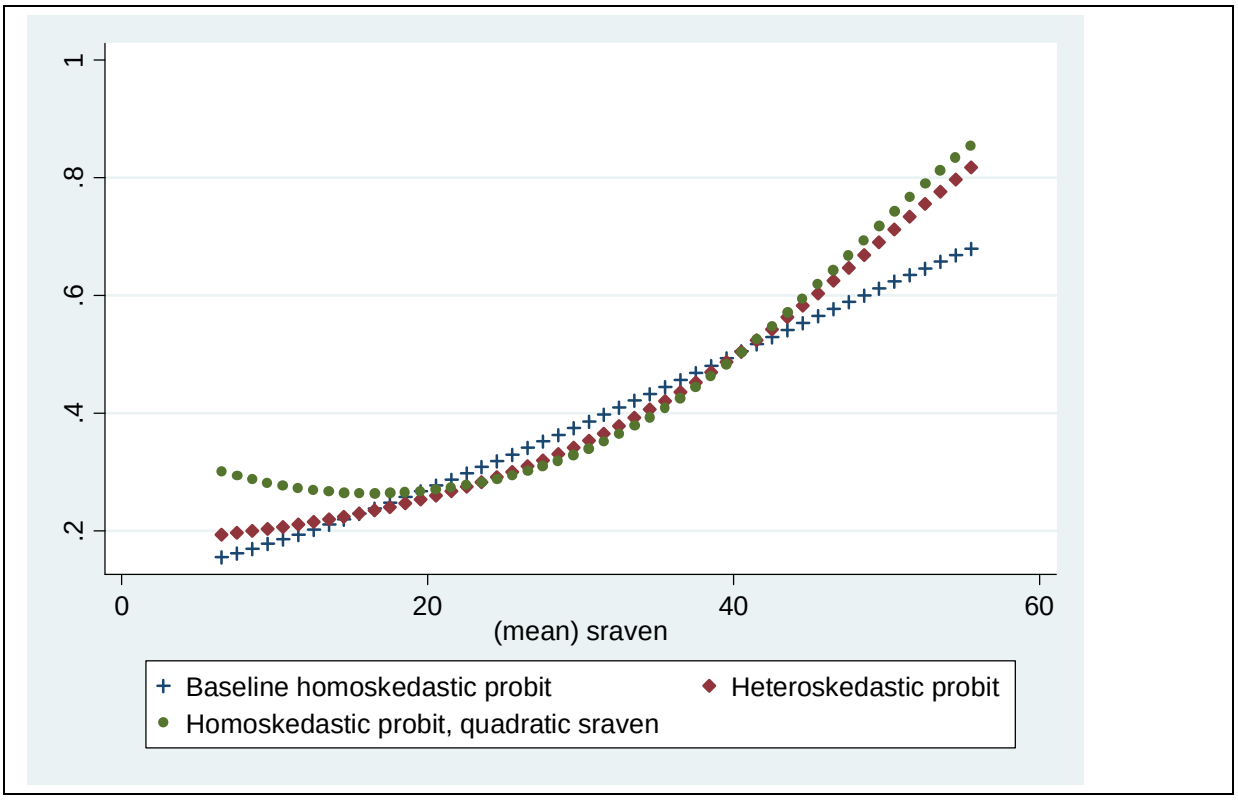
```
Heteroskedastic probit model                      Number of obs     =        902
                                                  Zero outcomes     =        542
                                                  Nonzero outcomes  =        360

                                                  Wald chi2(9)      =      36.56
Log likelihood = -336.3521                        Prob > chi2       =     0.0000

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
puaind       |
      numsib |  -.0672515   .0246799    -2.72   0.006    -.1156232   -.0188798
      sraven |   .0229242   .0041345     5.54   0.000     .0148208    .0310276
      wealth |   .0270485   .0057435     4.71   0.000     .0157914    .0383056
        male |   .5055253   .1190219     4.25   0.000     .2722466    .7388039
    lowcaste |  -.3304765    .127884    -2.58   0.010    -.5811245   -.0798284
      muslim |   -.410586   .1214911    -3.38   0.001    -.6487041   -.1724678
      medyrs |  -.0691932   .0268416    -2.58   0.010    -.1218016   -.0165847
     medyrsq |   .0041276   .0017081     2.42   0.016     .0007798    .0074754
     sikhchr |   .4802327   .2136518     2.25   0.025     .0614829    .8989825
       _cons |  -1.272155   .2532967    -5.02   0.000    -1.768607   -.7757022
-------------+----------------------------------------------------------------
lnsigma2     |
      sraven |  -.0171179   .0059309    -2.89   0.004    -.0287422   -.0054935
------------------------------------------------------------------------------
Likelihood-ratio test of lnsigma2=0: chi2(1) =     8.17   Prob > chi2 = 0.0043
```

Clearly there is evidence here that the variance of the error term falls with sraven. Alternatively, this can be interpreted as indicating that the functional form of the baseline probit is wrong. Now consider adding **sraven squared** to the baseline model, on the grounds that this is a generalization of the baseline probit. Results:

```
. ge sraven2=sraven^2;

. probit puaind numsib sraven wealth male lowcaste muslim medyrs medyrsq sikhchr
sraven2;

Probit regression                                 Number of obs   =        902
                                                  LR chi2(10)     =     539.44
                                                  Prob > chi2     =     0.0000
Log likelihood = -337.00859                       Pseudo R2       =     0.4445

------------------------------------------------------------------------------
      puaind |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      numsib |  -.0976174   .0383671    -2.54   0.011    -.1728155   -.0224192
      sraven |  -.0364237   .0257712    -1.41   0.158    -.0869344    .0140869
      wealth |   .0467619   .0044055    10.61   0.000     .0381273    .0553964
        male |   .8598127   .1210329     7.10   0.000     .6225926    1.097033
    lowcaste |  -.5200105   .1866679    -2.79   0.005    -.8858729   -.1541481
      muslim |  -.7285526   .1536653    -4.74   0.000    -1.029731   -.4273742
      medyrs |  -.1232646   .0372906    -3.31   0.001    -.1963527   -.0501765
     medyrsq |     .00768   .0024345     3.15   0.002     .0029084    .0124516
     sikhchr |   .9275063   .3282048     2.83   0.005     .2842368    1.570776
     sraven2 |   .0011059   .0004214     2.62   0.009       .00028    .0019317
       _cons |    -1.0303   .4258538    -2.42   0.016    -1.864958    -.195642
------------------------------------------------------------------------------
```

Clearly the squared term is quite significant.

Based on the three models shown above, the following graph illustrates how the predicted probability of going to a private unaided school varies with sraven, holding all other explanatory factors constant.



It seems the heteroskedastic probit and the homoscedastic probit with sraven squared included tell a similar story: the likelihood that y=1 is relatively insensitive to changes to sraven at low levels, but more sensitive to changes to sraven at high levels than what is implied by the benchmark model.

Box 2: Stata code generating the graph on the previous page

```
probit puaind numsib sraven wealth male lowcaste muslim medyrs medyrsq
sikhchr;
estimates store base;
hetprob puaind numsib sraven wealth male lowcaste muslim medyrs medyrsq
sikhchr, het(sraven);
estimates store het;

ge sraven2=sraven^2;
probit puaind numsib sraven wealth male lowcaste muslim medyrs medyrsq
sikhchr sraven2;
estimates store quad;

collapse _all;
ge junk=50;
expand junk;
replace sraven=sraven+(_n-25);
replace sraven2=sraven^2;

estimates restore base;
predict p1, p;
estimates restore het;
predict p2, p;
estimates restore quad;
predict p3, p;

label var p1 "Baseline homoskedastic probit";
label var p2 "Heteroskedastic probit";
label var p3 "Homoskedastic probit, quadratic sraven";



scatter p1 p2 p3 sraven, s(+ d o);
exit;
```