Econometrics II
Tutorial No. 2

Lennart Hoogerheide & Agnieszka Borowska

22.02.2017

Outline

Summary

Key terms

- **Multinomial data:** dependent variable can attain m possible outcomes ($y_i \in \{0, 1, \ldots, m-1\}$).
- **Ordered and unordered variables:** variables with or without a natural ordering.
  [ordered: e.g. education level, job category; unordered: e.g. means of transport]

Key terms – cont'd

- **Ordered response model:** a model where the categorical outcome y_i is related to the latent variable

$$y_i^* = x_i'\beta + e_i, \qquad e_i \sim IID(0,1)$$

by means of $m-1$ unknown threshold values $\tau_1 < \cdots < \tau_{m-1}$ as follows

$$y_i = \begin{cases} 0 & \text{if } -\infty < y_i^* \leq \tau_1, \\ j & \text{if } \tau_j < y_i^* \leq \tau_{j+1},\ j = 1, \ldots, m-2, \\ m-1 & \text{if } \tau_{m-1} < y_i^* < \infty \end{cases}$$

($k+m-2$ parameters, no constant term in β which has $k-1$ elements).

Key terms – cont'd

- **Ordered response model: (cont'd)** The probability of choosing the alternative j:

$$\begin{aligned} p_{ij} &= \mathbb{P}[y_i = j] \\ &= \mathbb{P}[\tau_j < y_i^* \leq \tau_{j+1}] \\ &= \mathbb{P}[y_i^* \leq \tau_{j+1}] - \mathbb{P}[y_i^* \leq \tau_j] \\ &= G(\tau_{j+1}) - G(\tau_j), \end{aligned}$$

where $\tau_0 = -\infty$ and $\tau_m = \infty$.

Depending on $G(\cdot)$, the distribution of e_i, we have the ordered probit $(G(\cdot) = \Phi(\cdot))$ or logit $(G(\cdot) = \Lambda(\cdot))$ model.

Key terms – cont'd

- **Multinomial logit:**

$$p_{ij} = \frac{\exp(x_i'\beta_j)}{\sum_{h=1}^{m} \exp(x_i'\beta_h)} = \frac{\exp(x_i'\beta_j)}{1 + \sum_{h=2}^{m} \exp(x_i'\beta_h)}$$

$\Rightarrow$ individual-specific data.

- **Conditional logit:**

$$p_{ij} = \frac{\exp(x_i'\beta))}{\sum_{h=1}^{m} \exp(x_i'\beta)}$$

$\Rightarrow$ alternative-specific data.

Key terms – cont'd

- **Marginal effects of explanatory variables**: (in multinomial logit model) all the parameters $\beta_1, \ldots, \beta_{m-1}$ **together** determine the marginal effect of x_i on the probability to choose the jth alternative. So the sign of the parameter $\beta_l^{(j)}$ cannot always be interpreted **directly** as the sign of the effect of the x_l on the probability to choose the jth alternative.
- **Odds ratio:** the relative odds to choose between the alternatives j and h, given by (in multinomial logit):

$$\frac{\mathbb{P}(y_i = j|x_i)}{\mathbb{P}(y_i = h|x_i)} = \exp\left(x_i'(\beta^{(j)} - \beta^{(h)})\right).$$

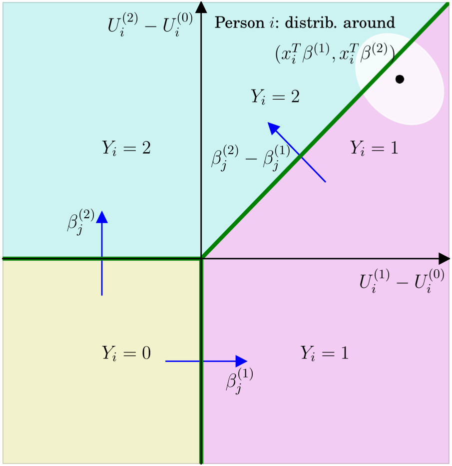
Then: $(\beta_l^{(j)} - \beta_l^{(h)}) > 0$ indicates a positive effect of x_{li} on $\mathbb{P}(y_i = j|x_i)$ **relative to** $\mathbb{P}(y_i = h|x_i)$.

Key terms – cont'd

- **Utilities Model:** A model where the observed dependent variable is assumed to be a function of utilities experienced from alternative choices, $U_i^{(j)}$, $j = 0, 1, \ldots, m$. The observed choice depends on the difference in the utilities. [interpretation of multinomial/binary logit/probit model alternative to the latent variables model]
- (Cf. from last week) **Latent Variable Model:** A model where the observed dependent variable is assumed to be a function of an underlying latent, or unobserved, variable. [interpretation of binary logit/probit model]

Key terms – cont'd

- **Multinomial logit: 3 categories case** (the jth variable)



Key terms – cont'd

- **Standard extreme value distribution:**

$$G(x) = \exp(-\exp(-x)), \qquad \text{(CDF)}$$
$$p(x) = \exp(-\exp(-x) - x). \qquad \text{(PDF)}$$

The difference between two independent variables with (standard) extreme value distribution has (standard) logistic distribution
$\Rightarrow$ used in defining of binary logit model in terms of utilities.

Extra Topics – from last week

The Perfect Classifier Problem

Recall – the loglikelihood:

$$\begin{aligned} \ln L(\beta) &= \ln p(y_1, \ldots, y_n | x_1, \ldots, x_n) \\ &= \sum_{i=1}^{n} \Big\{ \underbrace{y_i \ln[G(x_i'\beta)]}_{(*)} + \underbrace{(1 - y_i) \ln[1 - G(x_i'\beta)]}_{(**)} \Big\}. \qquad (1) \end{aligned}$$

We have

$$0 < G(x_i'\beta) < 1,$$

hence

$$-\infty < \ln[G(x_i'\beta)] < 0.$$

Notice that

$$y_i = 1 \Rightarrow (*) < 0 \ \& \ (**) = 0,$$
$$y_i = 0 \Rightarrow (*) = 0 \ \& \ (**) < 0.$$

Perfect fit:

$$y_i = 1 \iff G(x_i'\beta) = 1,$$
$$y_i = 0 \iff G(x_i'\beta) = 0.$$

This could happen only when

$$y_i = 1 \iff x_i'\beta = \infty, \tag{2}$$
$$y_i = 0 \iff x_i'\beta = -\infty. \tag{3}$$

We say that the loglikelihood (1) is *bounded above by* 0, and it achieves this bound if (2) and (3) hold.

Now, suppose that there is some linear combination of the independent variables, say $x_i'\beta^\bullet$, such that

$$y_i = 1 \iff x_i'\beta^\bullet > 0, \tag{4}$$
$$y_i = 0 \iff x_i'\beta^\bullet < 0. \tag{5}$$

In other words, there is some range of the regressor(s) for which y_i is always 1 or 0.

Then, we say that $x_i'\beta^\bullet$ describes a **separating hyperplane** (see Figure 2.1) and there is **complete separation** of the data.

$x_i'\beta^\bullet$ is said to be a **perfect classifier**, since it allows us to predict y_i with perfect accuracy for every observation.

Problem?

Yes, for ML estimation!

Then, it is possible to make the value of $\ln L$ arbitrarily close to 0 (the upper bound) by choosing β arbitrarily large (in an absolute sense)[1].

Hence, no finite ML estimator exists.

[1] Formally: by setting $\beta = \gamma\beta^{\bullet}$ and letting $\gamma \to \infty$.

Computer arithmetic

This is exactly what any nonlinear maximization algorithm will attempt to do if there exists a vector $\beta^{\bullet}$ for which conditions (4) and (5) are satisfied.

Because of the numerical limitations, the algorithm will eventually terminate (with some numerical error) at a value of $\ln L$ slightly less than 0.

This is likely to occur in practice when the sample is very small, when almost all of the y_i are equal to 0 or almost all of them are equal to 1, or when the model fits extremely well.

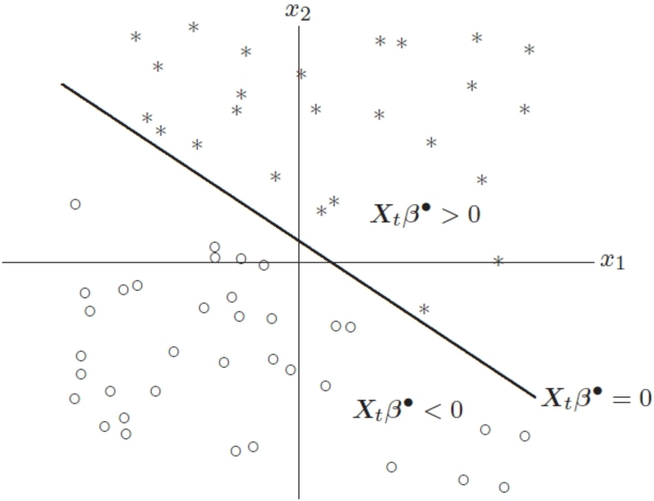


Figure 2.1: Figure 11.2 from Davidson and MacKinnon (1999), "Econometric Theory and Methods": A perfect classifier yields a separating hyperplane.

Simulation from the latent variable model

Consider the latent variable model

$$
\begin{aligned}
y_i^* &= \beta_0 + \beta_1 x_i + e_i, \\
e_i &\sim \mathcal{N}(0,1), \\
y_i &= \begin{cases} 1, & \text{if } y_i^* > 0, \\ 0, & \text{if } y_i^* \leq 0 \end{cases}
\end{aligned}
$$

Suppose that $x_i \sim \mathcal{N}(0,1)$. *We will generate* $5,000$ *samples of* 20 *observations on* (x_i, y_i) *pairs in the following way:*

- $1,000$ *assuming that* $\beta_0 = 0$ *and* $\beta_1 = 1$*;*
- $1,000$ *assuming that* $\beta_0 = 1$ *and* $\beta_1 = 1$*;*
- $1,000$ *assuming that* $\beta_0 = -1$ *and* $\beta_1 = 1$*;*
- $1,000$ *assuming that* $\beta_0 = 0$ *and* $\beta_1 = 2$*;*
- $1,000$ *assuming that* $\beta_0 = 0$ *and* $\beta_1 = 3$*.*

For each of the $5,000$ *samples, we will attempt to estimate a probit model.*

We are interested in the following question:

In each of the five cases, what proportion of the time does the estimation fail because of perfect classifiers?

We also want to explain why there will be more failures in some cases than in others.

Next, we will repeat this exercise for five sets of $1,000$ *samples of size* 40*, with the same parameter values.*

This will allow us to draw a conclusion about the effect of sample size on the perfect classifier problem.

EViews code for the first case ($N = 20$ with β_0 and β_1).

Program: LATENTVARIABLE_DM17_5 - (h:\desktop\econometric2\latentvariable_dm1...

Run | Print | Save | SaveAs | Cut | Copy | Paste | InsertTxt | Find | Replace | Wrap+/- | LineNum+/- | Encrypt

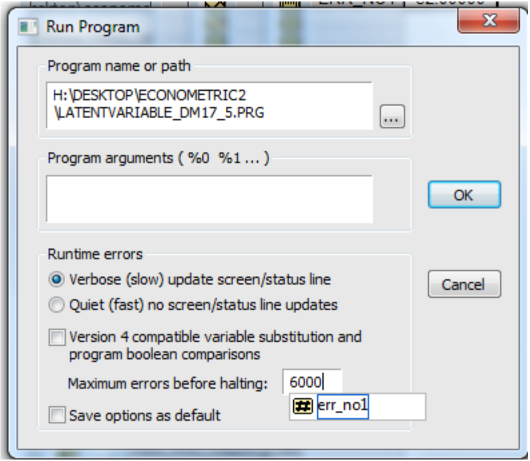
```
wfcreate(wf=latentvariable_dm17_5_0_1) u 20
'Control Variables
!N = 20
!M =1000
setmaxerrs 6*!M '6 because if the estimation fails no coefs, stderrs and loglik are created, and
assigning of these creates next errors

'Parameters
!beta0 = 0
!beta1 = 1

matrix(!N,!M) xs
matrix(!N,!M) us
matrix(!N,!M) ys
matrix(!N,!M) y_stars

matrix(2,!M) eq_coeff
matrix(2,!M) eq_stderrs
matrix(1,!M) eq_loglik

for !i=1 to !M
    series u = nrnd
    matplace(us,u,1,!i)
    series x = nrnd
    matplace(xs,x,1,!i)
    series y_star = !beta0 + !beta1*x + u
    matplace(y_stars,y_star,1,!i)
    series y = @recode(y_star>0, 1, 0)
    matplace(ys,y,1,!i)

    equation eq.binary(d="n") y c x
    eq_coeff(1,!i) = eq.@coefs(1)
    eq_coeff(2,!i) = eq.@coefs(2)
    eq_stderrs(1,!i) = eq.@stderrs(1)
    eq_stderrs(2,!i) = eq.@stderrs(2)

    eq_loglik = eq.@logl
next

scalar err_no1 = @errorcount/6

wfsave "H:\Desktop\Econometric2\DM17_5_N20_betas_0_1"
```

If you are interested, you can check the results of each probit estimation: the coefficients estimates, their standard errors and the loglikelihood values are stored in matrices `eq_coeff`, `eq_stderrs` and `eq_logl`, respectively.

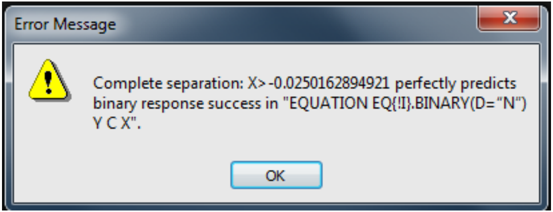
But what we are truly after, is the error count variable, `err_no1`, which reports how many times an estimation error occurred.

Notice, that we used the command `setmaxerr` to set the maximum number of error that the program may encounter before execution is halted.

Alternatively, you can specify it in the box showing up after clicking on the `run` button.



Without changing the value of maximum error allowed, the program would shortly break with the error message reporting the perfect separation problem.



The proportion of the time that perfect classifiers were encountered for each of the five cases and each of the two sample sizes:

Parameters	$n = 20$	$n = 40$
$\beta_0 = 0,\ \beta_1 = 1$	0.012	0.000
$\beta_0 = 1,\ \beta_1 = 1$	0.074	0.001
$\beta_0 = -1,\ \beta_1 = 1$	0.056	0.002
$\beta_0 = 0,\ \beta_1 = 2$	0.143	0.008
$\beta_0 = 0,\ \beta_1 = 3$	0.286	0.052

The proportion of samples with perfect classifiers increases as both β_0 and β_1 increase in absolute value. When $\beta_0 = 0$, the unconditional expectation of y_i is 0.5.

As β_0 increases in absolute value, this expectation becomes larger, and the proportion of 1s in the sample increases.

As β_1 becomes larger in absolute value, the model fits better on average, which obviously increases the chance that it fits perfectly.

The results for parameters $(1, 1)$ are almost identical to those for parameters $(-1, 1)$ because, with x_i having mean 0, the fraction of 1s in the samples with parameters $(1, 1)$ is the same, on average, as the fraction of 0s in the samples with parameters $(-1, 1)$.

Comparing the results for $n = 20$ and $n = 40$, it is clear that the probability of encountering a perfect classifier falls very rapidly as the sample size increases.

Lecture Problems

Lecture Problems: Exercise 3

Show that the ordered probit model (with two explanatory variables x_{i1} and x_{i2}) with $m = 2$ alternatives is the binary probit model with constant term $\beta_0 = -\tau_1$, by showing that $\mathbb{P}(y_i = 1|x_i)$ is the same in both models.

Ordered probit model in case of 2 categories $y_i \in \{0, 1\}$ and two explanatory variables x_{i1} and x_{i2}.

Latent variable y_i^*:

$$y_i^* = \beta_1 x_{i1} + \beta_2 x_{i2} + e_i,$$

where $e_i \sim N(0, 1)$, i.i.d. We have observed variable y_i:

$$y_i = \begin{cases} 0 & \text{if } -\infty < y_i^* \leq \tau_1, \\ 1 & \text{if } \tau_1 < y_i^* \leq \infty, \end{cases}$$

with threshold value τ_1.

We have:

$$\begin{aligned}
\mathbb{P}(y_i = 1|x_i) &= \mathbb{P}(y_i^* > \tau_1|x_i) \\
&= \mathbb{P}(\beta_1 x_{i1} + \beta_2 x_{i2} + e_i > \tau_1|x_i) \\
&= \mathbb{P}(e_i > \tau_1 - \beta_1 x_{i1} - \beta_2 x_{i2}|x_i) \\
&\overset{(*)}{=} \mathbb{P}(e_i < -\tau_1 + \beta_1 x_{i1} + \beta_2 x_{i2}|x_i) \\
&\overset{(**)}{=} \mathbb{P}(e_i \leq -\tau_1 + \beta_1 x_{i1} + \beta_2 x_{i2}|x_i) \\
&\overset{(***)}{=} \mathbb{P}(e_i \leq -\tau_1 + \beta_1 x_{i1} + \beta_2 x_{i2}) \\
&= \Phi(-\tau_1 + \beta_1 x_{i1} + \beta_2 x_{i2}),
\end{aligned}$$

the standard normal distribution of e_i is $(*)$ symmetric around 0, $(**)$ continuous and $(***)$ independent of x_i; $\Phi(.)$ is the CDF of the standard normal distribution.

Further, since $y_i = 0$ or $y_i = 1$ we have

$$\mathbb{P}(y_i = 0|x_i) + \mathbb{P}(y_i = 1|x_i) = 1,$$

so that

$$\mathbb{P}(y_i = 0|x_i) = 1 - \mathbb{P}(y_i = 1|x_i) = 1 - \Phi(-\tau_1 + \beta_1 x_{i1} + \beta_2 x_{i2}).$$

In the **binary probit** model we have

$$\mathbb{P}(y_i = 1|x_i) = \Phi(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}),$$
$$\mathbb{P}(y_i = 0|x_i) = 1 - \mathbb{P}(y_i = 1|x_i) = 1 - \Phi(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}).$$

So, indeed $\mathbb{P}(y_i = 1|x_i)$ is the same in the binary probit model and in the ordered probit model with $m = 2$ alternatives (with $-\tau_1 = \beta_0$).

Therefore: the ordered probit model reduces to the binary probit model if we have only $m = 2$ alternatives.
I.e.: they have the same Bernoulli distribution for y_i (conditionally upon x_i).

In a similar way it holds that the ordered logit model reduces to the binary logit model if we have only $m = 2$ alternatives.

Lecture Problems: Exercise 4

The EViews file `bank_employees_exercise13.wf1` *contains the data, where also two variables have been added:*

- `admin0_manage1_cust2` *(where 0 = administrative, 1 = management, 2 = custodial), where the ordering is done based on average value of* $male_i$ *per category;*
- `admin0_cust1_manage2` *(where 0 = administrative, 1 = custodial, 2 = management), where the ordering is done based on average value of salary per category.*

Estimate two ordered logit models using these series as dependent variable (and `education` *and* `male` *as explanatory variables). Compare the AIC, SC and prediction quality with the model where the categories are ordered with education (with dependent variable* `ORDERED_JOB_CATEGORY`*, which is used on the slides). Can you explain the differences?*

We have:

dependent variable	AIC	SC	percentage correctly predicted
ORDERED_JOB_CATEGORY	0.829509	0.864624	85.232%
admin0_manage1_cust2	1.151120	1.186236	77.215%
admin0_cust1_manage2	1.051928	1.087044	84.810%

Note:

- `ORDERED_JOB_CATEGORY` could be called '`cust0_admin1_manage2`' (where 0 = custodial, 1 = administrative, 2 = management).
- The model with (0 = custodial, 1 = administrative, 2 = management) is the best: the lowest (best) AIC and SC, and the highest (best) percentage correctly predicted.

Reason: *education* is the most important explanatory variable (more important than *male*), so it is best to order the categories with *education*. A higher education **increases** the probability of going from category 0=custodial to 1=administrative, and it **increases** the probability of going from category 1=administrative to 2=management.

dependent variable	AIC	SC	percentage correctly predicted
`ORDERED_JOB_CATEGORY`	0.829509	0.864624	85.232%
`admin0_manage1_cust2`	1.151120	1.186236	77.215%
`admin0_cust1_manage2`	1.051928	1.087044	84.810%

Note:

- `ORDERED_JOB_CATEGORY` could be called '`cust0_admin1_manage2`' (where 0 = custodial, 1 = administrative, 2 = management).
- The model with (where 0 = administrative, 1 = management, 2 = custodial) is the worst: the highest (worst) AIC and SC, and the lowest (worst) percentage correctly predicted.

Reason: *male* is a relatively unimportant explanatory variable (less important than *education*), so it is not good to order the categories with *male*. Here the estimated coefficient of education is 'damaged', because education **increases** the probability of going from category 0=administrative to 1=management, but it **decreases** the probability of going from category 1=management to 2=custodial.

dependent variable	AIC	SC	percentage correctly predicted
`ORDERED_JOB_CATEGORY`	0.829509	0.864624	85.232%
`admin0_manage1_cust2`	1.151120	1.186236	77.215%
`admin0_cust1_manage2`	1.051928	1.087044	84.810%

Note:

- `ORDERED_JOB_CATEGORY` could be called '`cust0_admin1_manage2`' (where 0 = custodial, 1 = administrative, 2 = management).
- The model with (0 = administrative, 1 = custodial, 2 = management) is also bad: the AIC, SC and percentage correctly predicted are bad (close to the worst model and much worse than the best model).

Reason: Here the estimated coefficient of *education* is again 'damaged', because *education* **decreases** the probability of going from category 0=administrative to 1=custodial, but it **increases** the probability of going from category 1=management to 2=custodial.

dependent variable	AIC	SC	percentage correctly predicted
`ORDERED_JOB_CATEGORY`	0.829509	0.864624	85.232%
`admin0_manage1_cust2`	1.151120	1.186236	77.215%
`admin0_cust1_manage2`	1.051928	1.087044	84.810%

Note:

- `ORDERED_JOB_CATEGORY` could be called '`cust0_admin1_manage2`' (where 0 = custodial, 1 = administrative, 2 = management).
- Beforehand we could **not** say whether the model with (0 = administrative, 1 = management, 2 = custodial) or the model with (0 = administrative, 1 = custodial, 2 = management) would be the worst.

Both of these models have a poor ordering of the categories (when looking at the effect of *education* on the probabilities of being in the categories).

Problem on binary, ordered & multinomial logit models

Problem on binary, ordered & multinomial logit models

Consider the binary logit model where

$$y_i^* = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i, \qquad y_i = \begin{cases} 1 & \textit{if } y_i^* > 0, \\ 0 & \textit{if } y_i^* \leq 0, \end{cases}$$

where the e_i *(*$i = 1, 2, \ldots, n$*) are i.i.d. errors that have the (standard) logistic distribution with cumulative distribution function (CDF) given by*

$$G(a) = \mathbb{P}(e_i \leq a) = \frac{1}{1 + \exp(-a)} = \frac{\exp(a)}{1 + \exp(a)},$$

and where the e_i *(*$i = 1, 2, \ldots, n$*) are independent of* x_{j1} *and* x_{j2} *(*$j = 1, 2, \ldots, n$*).*

Problem (a)

Derive the probability $\mathbb{P}(y_i = 1|x_{i1}, x_{i2})$ *and the probability* $\mathbb{P}(y_i = 0|x_{i1}, x_{i2})$.

We have:

$$
\begin{aligned}
\mathbb{P}(y_i = 1|x_i) &= \mathbb{P}(y_i^* > 0|x_i) \\
&= \mathbb{P}(x_i'\beta + e_i > 0|x_i) \\
&= \mathbb{P}(e_i > -x_i'\beta|x_i) \\
&\overset{(*)}{=} \mathbb{P}(e_i < x_i'\beta|x_i) \\
&\overset{(**)}{=} \mathbb{P}(e_i \leq x_i'\beta|x_i) \\
&\overset{(***)}{=} \mathbb{P}(e_i \leq x_i'\beta) \\
&= G(x_i'\beta) = \frac{1}{1 + \exp(-x_i'\beta)} \\
&= \frac{\exp(x_i'\beta)}{1 + \exp(x_i'\beta)},
\end{aligned}
$$

the standard logistic distribution of e_i is $(*)$ symmetric around 0, $(**)$ continuous and $(***)$ independent of x_i.

Further, y_i is either 0 or 1, so that

$$\mathbb{P}(y_i = 0|x_{i1}, x_{i2}) + \mathbb{P}(y_i = 1|x_{i1}, x_{i2}) = 1,$$

so we have:

$$\mathbb{P}(y_i = 0|x_{i1}, x_{i2}) = 1 - G(x_i'\beta) = \frac{1}{1 + \exp(x_i'\beta)}.$$

Problem (b)

Derive the loglikelihood in this model.

The likelihood per observation i is the probability function of y_i, conditionally upon x_i:

$$p(y_i|x_i) = [G(x_i'\beta)]^{y_i}[1 - G(x_i'\beta)]^{1-y_i} = \begin{cases} G(x_i'\beta) & \text{if } y_i = 1, \\ 1 - G(x_i'\beta) & \text{if } y_i = 0. \end{cases}$$

The likelihood is the joint probability function of the y_i $(i = 1, 2, \dots, n)$, conditionally upon the x_i $(i = 1, 2, \dots, n)$:

$$\begin{aligned} L(\beta) &= p(y_1, \dots, y_n | x_1, \dots, x_n) \\ &\overset{(*)}{=} \prod_{i=1}^{n} p(y_i | x_i) \\ &= \prod_{i=1}^{n} [G(x_i'\beta)]^{y_i} [1 - G(x_i'\beta)]^{1-y_i}, \end{aligned}$$

where in $(*)$ we used the assumption that the y_i are independent (conditionally upon the x_i). In other words, we assume that the e_i are independent.

The loglikelihood is simply the (natural) logarithm of the likelihood:

$$\begin{aligned}\ln L(\beta) &= \ln p(y_1, \ldots, y_n | x_1, \ldots, x_n) \\ &= \sum_{i=1}^{n} \left\{ y_i \ln[G(x_i'\beta)] + (1 - y_i) \ln[1 - G(x_i'\beta)] \right\}.\end{aligned}$$

Problem (c)

Suppose that we analyse data on a presidential election, where there are two candidates, say C and T. We observe $n = 1000$ *observations. We have:*

$$y_i = \begin{cases} 1 & \textit{if person } i \textit{ votes for candidate } C, \\ 0 & \textit{if person } i \textit{ votes for candidate } T, \end{cases}$$

$$x_{2i} = \textit{number of years of education of person } i, x_{2i} \in [12, 20],$$

and

$$x_{3i} = \begin{cases} 1 & \textit{if person } i \textit{ is a female}, \\ 0 & \textit{if person } i \textit{ is a male}. \end{cases}$$

Figure 4.1 contains ML estimation output and graphs of the estimated probability $\hat{\mathbb{P}}(y_i = 1|x_{i1}, x_{i2})$.
Explain why the estimates of β_1 *and* β_2 *match with the graphs of the estimated probability* $\hat{\mathbb{P}}(y_i = 1|x_{i1}, x_{i2})$.

Dependent Variable: Y
Method: ML - Binary Logit (Quadratic hill climbing)
Sample: 1 1000
Included observations: 1000
Convergence achieved after 4 iterations
Covariance matrix computed using second derivatives

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	-2.472179	0.493827	-5.006160	0.0000
EDUCATION	0.171900	0.030299	5.673508	0.0000
FEMALE	0.914104	0.141307	6.468922	0.0000

McFadden R-squared	0.059675	Mean dependent var	0.671000
S.D. dependent var	0.470085	S.E. of regression	0.453222
Akaike info criterion	1.197331	Sum squared resid	204.7936
Schwarz criterion	1.212055	Log likelihood	-595.6657
Hannan-Quinn criter.	1.202927	Deviance	1191.331
Restr. deviance	1266.936	Restr. log likelihood	-633.4682
LR statistic	75.60489	Avg. log likelihood	-0.595666
Prob(LR statistic)	0.000000		

Obs with Dep=0	329	Total obs	1000
Obs with Dep=1	671		

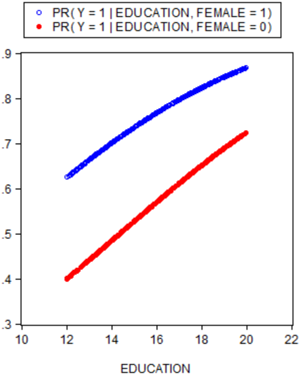


Figure 4.1: Binary logit model: estimation output and graphs of the estimated probability $\widehat{\mathbb{P}}(y_i = 1|x_{i1}, x_{i2})$.

The estimated coefficients $\hat{\beta}_1$ (at education x_{i1}) and $\hat{\beta}_2$ (at female x_{i2}) are significantly positive, which matches with the fact that $\hat{\mathbb{P}}(y_i = 1|x_{i1}, x_{i2})$ is increasing with education and is higher for females than for males.

The graph for males is the graph for females shifted 0.91/0.17=5.35 to the right.

Problem (d)

Now suppose there are three candidates, say C, T and B. We have:

$$y_i = \begin{cases} 0 & \textit{if person } i \textit{ votes for candidate } C, \\ 1 & \textit{if person } i \textit{ votes for candidate } T, \\ 2 & \textit{if person } i \textit{ votes for candidate } B. \end{cases}$$

Figures 4.2 and 4.3 contain ML estimation output and graphs of the estimated probabilities $\hat{\mathbb{P}}(y_i = 0|x_{i1}, x_{i2})$, $\hat{\mathbb{P}}(y_i = 1|x_{i1}, x_{i2})$ *and* $\hat{\mathbb{P}}(y_i = 2|x_{i1}, x_{i2})$ *in the multinomial logit model (with reference category 0). Explain why the estimates of the coefficients match with the graphs of the estimated probabilities* $\widehat{\mathbb{P}}(y_i = 0|x_{i1}, x_{i2})$, $\widehat{\mathbb{P}}(y_i = 1|x_{i1}, x_{i2})$ *and* $\widehat{\mathbb{P}}(y_i = 2|x_{i1}, x_{i2})$.

LogL: ML_MULTINOMIAL_LOGIT
Method: Maximum Likelihood (Marquardt)
Sample: 1 1000
Included observations: 1000
Evaluation order: By equation
Convergence achieved after 8 iterations

	Coefficient	Std. Error	z-Statistic	Prob.
FOR PROBABILITY OF VOTING T (VERSUS REFERENCE CATEGORY OF VOTING C)				
C(1) (CONSTANT)	1.779986	0.519733	3.424805	0.0006
C(2) (EDUCATION)	-0.124617	0.032334	-3.854041	0.0001
C(3) (FEMALE)	-0.863168	0.145194	-5.944921	0.0000
FOR PROBABILITY OF VOTING B (VERSUS REFERENCE CATEGORY OF VOTING C)				
C(4) (CONSTANT)	-16.29794	1.901439	-8.571372	0.0000
C(5) (EDUCATION)	0.820380	0.102006	8.042430	0.0000
C(6) (FEMALE)	0.191768	0.232677	0.824182	0.4098

Log likelihood	-805.0163	Akaike info criterion	1.622033
Avg. log likelihood	-0.805016	Schwarz criterion	1.651479
Number of Coefs.	6	Hannan-Quinn criter.	1.633224

Figure 4.2: Multinomial logit model: estimation output.

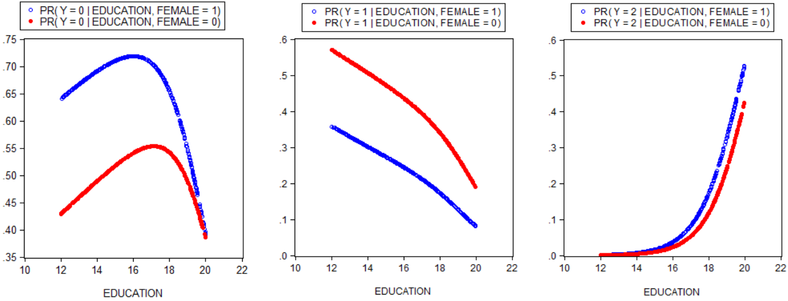


Figure 4.3: Multinomial logit model: graphs of the estimated probabilities $\hat{\mathbb{P}}(y_i = 0|x_{i1}, x_{i2})$, $\hat{\mathbb{P}}(y_i = 1|x_{i1}, x_{i2})$ and $\hat{\mathbb{P}}(y_i = 2|x_{i1}, x_{i2})$.

In this multinomial logit model we have probabilities:

$$\mathbb{P}(y_i = 0|x_i)$$
$$= \frac{1}{1 + \exp(\beta_0^{(1)} + \beta_1^{(1)} x_{i1} + \beta_2^{(1)} x_{i2}) + \exp(\beta_0^{(2)} + \beta_1^{(2)} x_{i1} + \beta_2^{(2)} x_{i2})},$$
$$\mathbb{P}(y_i = 1|x_i)$$
$$= \frac{\exp(\beta_0^{(1)} + \beta_1^{(1)} x_{i1} + \beta_2^{(1)} x_{i2})}{1 + \exp(\beta_0^{(1)} + \beta_1^{(1)} x_{i1} + \beta_2^{(1)} x_{i2}) + \exp(\beta_0^{(2)} + \beta_1^{(2)} x_{i1} + \beta_2^{(2)} x_{i2})},$$
$$\mathbb{P}(y_i = 2|x_i)$$
$$= \frac{\exp(\beta_0^{(2)} + \beta_1^{(2)} x_{i1} + \beta_2^{(2)} x_{i2})}{1 + \exp(\beta_0^{(1)} + \beta_1^{(1)} x_{i1} + \beta_2^{(1)} x_{i2}) + \exp(\beta_0^{(2)} + \beta_1^{(2)} x_{i1} + \beta_2^{(2)} x_{i2})}.$$

Note: we have **odds ratio**

$$\frac{\mathbb{P}(y_i = 1|x_i)}{\mathbb{P}(y_i = 0|x_i)} = \exp(\beta_0^{(1)} + \beta_1^{(1)} x_{i1} + \beta_2^{(1)} x_{i2}),$$

so that

$$\ln\left(\frac{\mathbb{P}(y_i = 1|x_i)}{\mathbb{P}(y_i = 0|x_i)}\right) = \beta_0^{(1)} + \beta_1^{(1)} x_{i1} + \beta_2^{(1)} x_{i2}.$$

Looking at the effect of *education*:

- Reference category 0 (voting C) has coefficient 0 (by definition).
- Category 1 (voting T) has significantly negative estimated coefficient `C(2) = -0.12`: an increase in education decreases
$$\frac{\widehat{\mathbb{P}}(y_i = 1|x_{i1}, x_{i2})}{\widehat{\mathbb{P}}(y_i = 0|x_{i1}, x_{i2})}.$$
- Category 2 (voting B) has significantly positive estimated coefficient `C(5) = 0.82`: an increase in education increases
$$\frac{\hat{\mathbb{P}}(y_i = 2|x_{i1}, x_{i2})}{\hat{\mathbb{P}}(y_i = 0|x_{i1}, x_{i2})}.$$

Hence, an increase in education increases $\hat{\mathbb{P}}(y_i = 2|x_{i1}, x_{i2})$ (B) and decreases $\hat{\mathbb{P}}(y_i = 1|x_{i1}, x_{i2})$ (T).

Looking at the effect of $female$ ($x_{i2} = 1$ for female, $x_{i2} = 0$ for male):

- Reference category 0 (voting C) has coefficient 0 (by definition).
- Category 1 (voting T) has significantly negative estimated coefficient `C(3) = -0.86`:
$$\frac{\hat{\mathbb{P}}(y_i = 1|x_{i1}, x_{i2} = 1)}{\widehat{\mathbb{P}}(y_i = 0|x_{i1}, x_{i2} = 0)} < 1.$$
- Category 2 (voting B) has **insignificant** estimated coefficient `C(6) = 0.19`: we can not reject that
$$\frac{\hat{\mathbb{P}}(y_i = 2|x_{i1}, x_{i2} = 1)}{\hat{\mathbb{P}}(y_i = 0|x_{i1}, x_{i2} = 0)} = 1.$$

Hence,

- $\hat{\mathbb{P}}(y_i = 1|x_{i1}, x_{i2})$ (T) is lower for females than for males.
- $\hat{\mathbb{P}}(y_i = 0|x_{i1}, x_{i2})$ (C) and $\widehat{\mathbb{P}}(y_i = 2|x_{i1}, x_{i2})$ (B) are higher for females than for males.

Problem (e)

Could an ordered logit model be appropriate in this case? Motivate your answer.

No: the alternatives can not be ordered in such a way that the explanatory variables 'push' someone from the first to the second alternative and from the second to the third alternative.

- *education* 'pushes' from T to C and from C to B;
- *female* 'pushes' from T to C **but not (significantly) from C to B**.

However, if we ignore the fact that the positive estimated effect of *female* on

$$\frac{\widehat{\mathbb{P}}(y_i = 2|x_{i1}, x_{i2} = 1)}{\widehat{\mathbb{P}}(y_i = 0|x_{i1}, x_{i2} = 0)}$$

is **not** significant, then yes: we can order the alternatives T, C, B, where both the variables *education* and $female$ ‘push’ persons from T to C and from C to B. In that case the ordered logit model **could** be appropriate.

Computer Exercise

W17/C2 (i)

Use the data in `loanapp.wf1` *for this exercise.*

(i) Estimate a probit model of approve on white. Find the estimated probability of loan approval for both whites and nonwhites. How do these compare with the linear probability estimates?

Equation: EQ_PROBIT Workfile: LOANAPP::Loanapp\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: APPROVE
Method: ML - Binary Probit (Newton-Raphson / Marquardt steps)
Sample (adjusted): 1 1988
Included observations: 1988 after adjustments
Convergence achieved after 3 iterations
Coefficient covariance computed using observed Hessian

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	0.546946	0.075435	7.250563	0.0000
WHITE	0.783615	0.086714	9.036738	0.0000

McFadden R-squared	0.053274	Mean dependent var	0.877264
S.D. dependent var	0.328217	S.E. of regression	0.320172
Akaike info criterion	0.707023	Sum squared resid	203.5846
Schwarz criterion	0.712652	Log likelihood	-700.7813
Hannan-Quinn criter.	0.709091	Deviance	1401.563
Restr. deviance	1480.431	Restr. log likelihood	-740.2157
LR statistic	78.86870	Avg. log likelihood	-0.352506
Prob(LR statistic)	0.000000		

Obs with Dep=0	244	Total obs	1988
Obs with Dep=1	1744		

As there is only one explanatory variable that takes on just two values, there are only two different predicted values: the estimated probabilities of loan approval for white and nonwhite applicants. Rounded to three decimal places these are:

$$\begin{aligned}
\mathbb{P}(approve = 1 | white = 0) &= \Phi(\beta_0 + \beta_1 \cdot 0) \\
&= \Phi(0.547) = 0.708, \\
\mathbb{P}(approve = 1 | white = 1) &= \Phi(\beta_0 + \beta_1 \cdot 1) \\
&= \Phi(0.547 + 0.784) = 0.908,
\end{aligned}$$

for nonwhites and whites, respectively.

(In other words, 0.708 is the proportion of loans approved for nonwhites and 0.908 is the proportion approved for whites.)

W17/C2 (ii)

*(ii) Now, add the variables hrat, obrat, loanprc, unem, male, married, dep, sch, cosign, chist, pubrec, mortlat*1*, mortlat*2*, and vr to the probit model. Is there statistically significant evidence of discrimination against nonwhites?*

Equation: EQ_PROBIT2 Workfile: LOANAPP::Loanapp\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: APPROVE
Method: ML - Binary Probit (Newton-Raphson / Marquardt steps)
Sample (adjusted): 1 1988
Included observations: 1971 after adjustments
Convergence achieved after 3 iterations
Coefficient covariance computed using observed Hessian

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	2.062327	0.313176	6.585194	0.0000
WHITE	0.520253	0.096959	5.365707	0.0000
HRAT	0.007876	0.006962	1.131394	0.2579
OBRAT	-0.027692	0.006049	-4.577783	0.0000
LOANPRC	-1.011969	0.237240	-4.265600	0.0000
UNEM	-0.036685	0.017481	-2.098594	0.0359
MALE	-0.037001	0.109927	-0.336599	0.7364
MARRIED	0.265747	0.094252	2.819528	0.0048
DEP	-0.049576	0.039057	-1.269304	0.2043
SCH	0.014650	0.095842	0.152851	0.8785
COSIGN	0.086071	0.245751	0.350238	0.7262
CHIST	0.585281	0.095971	6.098491	0.0000
PUBREC	-0.778741	0.126320	-6.164823	0.0000
MORTLAT1	-0.187624	0.253113	-0.741265	0.4585
MORTLAT2	-0.494356	0.326556	-1.513847	0.1301
VR	-0.201062	0.081493	-2.467220	0.0136

McFadden R-squared	0.186602	Mean dependent var	0.876205
S.D. dependent var	0.329431	S.E. of regression	0.299475
Akaike info criterion	0.625338	Sum squared resid	175.3347
Schwarz criterion	0.670686	Log likelihood	-600.2710
Hannan-Quinn criter.	0.642002	Deviance	1200.542
Restr. deviance	1475.959	Restr. log likelihood	-737.9793
LR statistic	275.4167	Avg. log likelihood	-0.304551
Prob(LR statistic)	0.000000		

Obs with Dep=0	244	Total obs	1971
Obs with Dep=1	1727		

W17/C2 (iii)

(iii) Estimate the model from part (ii) by logit. Compare the coefficient on white to the probit estimate.

Equation: EQ_LOGIT2 Workfile: LOANAPP::Loanapp\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: APPROVE
Method: ML - Binary Logit (Newton-Raphson / Marquardt steps)
Sample (adjusted): 1 1988
Included observations: 1971 after adjustments
Convergence achieved after 4 iterations
Coefficient covariance computed using observed Hessian

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	3.801710	0.594707	6.392572	0.0000
WHITE	0.937764	0.172904	5.423603	0.0000
HRAT	0.013263	0.012880	1.029730	0.3031
OBRAT	-0.053034	0.011280	-4.701462	0.0000
LOANPRC	-1.904951	0.460443	-4.137212	0.0000
UNEM	-0.066579	0.032809	-2.029310	0.0424
MALE	-0.066385	0.206429	-0.321588	0.7478
MARRIED	0.503282	0.177998	2.827452	0.0047
DEP	-0.090734	0.073334	-1.237261	0.2160
SCH	0.041229	0.178404	0.231098	0.8172
COSIGN	0.132059	0.446094	0.296034	0.7672
CHIST	1.066577	0.171212	6.229570	0.0000
PUBREC	-1.340665	0.217366	-6.167781	0.0000
MORTLAT1	-0.309882	0.463520	-0.668541	0.5038
MORTLAT2	-0.894675	0.568581	-1.573522	0.1156
VR	-0.349828	0.153725	-2.275671	0.0229

McFadden R-squared	0.186297	Mean dependent var	0.876205
S.D. dependent var	0.329431	S.E. of regression	0.299487
Akaike info criterion	0.625567	Sum squared resid	175.3487
Schwarz criterion	0.670915	Log likelihood	-600.4962
Hannan-Quinn criter.	0.642230	Deviance	1200.992
Restr. deviance	1475.959	Restr. log likelihood	-737.9793
LR statistic	274.9664	Avg. log likelihood	-0.304666
Prob(LR statistic)	0.000000		

Obs with Dep=0	244	Total obs	1971
Obs with Dep=1	1727		

W17/C1 (iv)

(iv) Use equation

$$n^{-1}\sum_{i=1}^{n}\Big\{G\big[\hat{\beta}_0+\hat{\beta}_1 x_{i1}+\cdots+\hat{\beta}_{k-1}x_{ik-1}+\hat{\beta}_k(c_k+1)\big] \\ -G\big[\hat{\beta}_0+\hat{\beta}_1 x_{i1}+\cdots+\hat{\beta}_{k-1}x_{ik-1}+\hat{\beta}_k c_k\big]\Big\} \quad (17.17)$$

to estimate the sizes of the discrimination effects for probit and logit.

Note that (17.17) is the average partial effect for a discrete explanatory variable.

We consider all the variables but *white*. Instead, for each individual we consider two counterfactual scenarios: as if he or she was white and otherwise (new generated variables `white1` and `white0`), which we use to create two `groups` (`variables_white1` and `variables_white0`).

Then, we use the coefficients from two estimations (`coef_probit` and `coef_logit`) to sum all the variables multiplied by their respective coefficient.

This gives us the arguments inside $G(\cdot)$ in (17.17).

To evaluate $G(\cdot)$ we need to apply the appropriate function for each model.

For probit, it is $\Phi(z)$, the cdf of the standard normal distribution;
for logit, it is $\frac{1}{1+\exp(-z)}$.

Finally, we subtract the vector with $G(\cdot)$ applied to the sum under the "nonwhites scenario" from that under the "whites scenario" and average out.

The obtained values are $APE_{probit} = 0.1042$ and $APE_{logit} = 0.1009$, hence quite similar.

Program: LOANAPP - (h:\desktop\loanapp.prg)

Run Print Save SaveAs Cut Copy Paste InsertTxt Find Replace Wrap+/- LineNum+/- Encrypt

```
equation eq_probit.binary(d=n) approve c white hrat obrat loanprc unem male married dep sch cosign chist pubrec mortlat1 mortlat2 vr
equation eq_logit.binary(d=l) approve c white hrat obrat loanprc unem male married dep sch cosign chist pubrec mortlat1 mortlat2 vr

vector(16) coef_probit
coef_probit = eq_probit.@coefs

vector(16) coef_logit
coef_logit = eq_logit.@coefs

' counterfactual scenarios
genr white1 = 1
genr white0 = 0

' all variables under counterfactual scenarios
group variables_white1 white1 hrat obrat loanprc unem male married dep sch cosign chist pubrec mortlat1 mortlat2 vr
group variables_white0 white0 hrat obrat loanprc unem male married dep sch cosign chist pubrec mortlat1 mortlat2 vr

' sum inside the G functions
series sum_white0_probit
series  sum_white1_probit
series  sum_white0_logit
series  sum_white1_logit

' start summing with the intercept (beta0)
sum_white0_probit = coef_probit(1)
sum_white1_probit = coef_probit(1)
sum_white0_logit = coef_logit(1)
sum_white1_logit = coef_logit(1)

' add subsequent variables multiplied by their coefficients
' (there are more coefs because the one for the constant term is also there - hence !i-1 for the grouped variables)
for !i = 2 to 16
    series temp =  coef_probit(!i)* variables_white0(!i-1)
    sum_white0_probit = sum_white0_probit + temp

    series temp =  coef_probit(!i)* variables_white1(!i-1)
    sum_white1_probit = sum_white1_probit + temp

    series temp =  coef_logit(!i)* variables_white0(!i-1)
    sum_white0_logit = sum_white0_logit + temp

    series temp =  coef_logit(!i)* variables_white1(!i-1)
    sum_white1_logit = sum_white1_logit + temp
next

' for probit: compute G as the cdf of the standard normal distribution
series G_white0_probit  = @cnorm(sum_white0_probit )
series G_white1_probit  = @cnorm(sum_white1_probit )
series diff_probit = G_white1_probit  - G_white0_probit
scalar apf_probit = @mean(diff_probit )

' for logit: compute G as the logistic function
series G_white0_logit  = 1/(1+@exp(-sum_white0_logit))
series G_white1_logit  = 1/(1+@exp(-sum_white1_logit))
series diff_logit = G_white1_logit  - G_white0_logit
scalar apf_logit = @mean(diff_logit )
```