DEPARTMENT OF ECONOMICS
UNIVERSITY OF COPENHAGEN

Econometrics 2, Class 1

Problem Set #6
October 24, 2005

DEPARTMENT OF ECONOMICS
UNIVERSITY OF COPENHAGEN

Practical information

- You need to register for the exam by the end of this week at the latest.
  - This is done through www.punkt.ku.dk.
  - If you do not register, you will not be allowed to participate in the exam.
- You were supposed to evaluate the course by yesterday at the latest. Unfortunately, I was only told this after our last class.
  - I have written to those of you who have sent me their e-mail addresses.
  - If you did not get a chance to evaluate the course, you are very welcome to send me an e-mail with your comments.

Problem Set #6

- We start by working through an empirical exercise with a binary choice model.
- We then look at some properties of ARMA models.

6.1 Binary Choice Model

- We want to analyze who passes the first year of the BA in Economics at Copenhagen (after a year).
- We do this first "manually" using the general maximum likelihood routine in PcGive…
- …and secondly in more detail using the built-in procedures.
- We have a data set of 1105 students enrolled from Summer 1997 to Winter 2000 (when I started!).

DEPARTMENT OF ECONOMICS
UNIVERSITY OF COPENHAGEN

The variables

	bestd12	1 if the student passes first year within 12 month.
	kvotient	Average score in the qualifying high school exam (adgangsgivende eksamen). The variable is truncated at 6.9 and at 10.
	halvaar	1 for summer-intake, 0 for winter-intake.
	eksald	Years since high school exam.
	kvinde	1 if woman.
*	matB	1 if level B in mathematics in high school (mathematics programme).
*	matM	1 if unspecified level in mathematics in high school (mathematics programme).
*	sproglig	1 if language programme in high school.
*	hf	1 if attended HF (high school that takes 2 rather than 3 years).
*	hhx	1 if attended HHX (high school with focus on business and trade).
*	htx	1 if attended HTX (high school with focus on technical issues and engineering).
*	udland	1 if the qualifying high school exam is taken abroad.
*	GIF	1 in case of other qualifying high school exam.

Table 6.1: Variables in the data set `polit.in7`.

Note: The benchmark qualifying exam is Danish "level A" in mathematics, i.e. all the *-dummies are equal to 0.

DEPARTMENT OF ECONOMICS
UNIVERSITY OF COPENHAGEN

Binary choice model

We want to analyze the effects of the characteristics in the data set on the probability of passing the first year exams, i.e. the binary choice model

$$\text{Prob}(\text{bestd}12_i = 1 \mid x_i) = F(x_i'\beta),$$

where x_i contains explanatory variables and $F(\cdot)$ is some distribution function.

DEPARTMENT OF ECONOMICS
UNIVERSITY OF COPENHAGEN

(1) Logit model

$$F(w) = \frac{\exp(w)}{1+\exp(w)}$$

$$w = \beta_0 + \beta_1 \cdot kvotient + \beta_2 \cdot halvaar + ... + \beta_{12} \cdot GIF$$

$$L_i(\beta) = \Pr(y_i = 1 \mid w)^{y_i} \cdot \Pr(y_i = 0 \mid w)^{1-y_i}$$

$$\Rightarrow \log L_i(\beta) = y_i \cdot \log F(w) + (1-y_i) \cdot \log(1-F(w))$$

$$\Rightarrow \log L(\beta) = \sum_{i=1}^{N} \left[y_i \cdot \log F(w) + (1-y_i) \cdot \log(1-F(w)) \right]$$

DEPARTMENT OF ECONOMICS
UNIVERSITY OF COPENHAGEN

(2) PcGive – Package – Descriptive Statistics; Model – Formulate – (Select all variables)

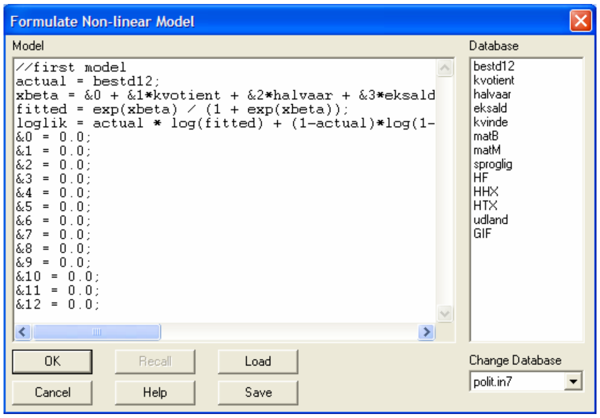
```
Descriptive Statistics package version 1.0, object created on 21-10-2005

Means, standard deviations and correlations (using polit.in7)
The sample is 1 - 1105
Means
      bestd12     kvotient      halvaar       eksald       kvinde         matB
      0.46878       8.5314      0.83258       1.9928      0.28597      0.10136
         matM     sproglig           HF          HHX          HTX       udland
      0.11041     0.051584     0.084163      0.12670     0.019005     0.027149
          GIF
     0.014480
Standard deviations (using T-1)
      bestd12     kvotient      halvaar       eksald       kvinde         matB
      0.49925      0.83501      0.37352       2.7103      0.45208      0.30194
         matM     sproglig           HF          HHX          HTX       udland
      0.31354      0.22129      0.27776      0.33278      0.13660      0.16259
          GIF
      0.11951
Correlation matrix:
```

I haven't inserted the correlation matrix!

47% pass, the average grade is 8.5, most people start in the Summer and are men, and the average years since taking the high school exam is 2. Most have level "A" mathematics.

(3) Package – Econometric Modelling; Model – Non-Linear Modelling



You will need to enter some code in this window – I will explain it on the next slide.

DEPARTMENT OF ECONOMICS
UNIVERSITY OF COPENHAGEN

How to formulate the model

```
//first model
actual = bestd12;
xbeta = &0 + &1*kvotient + &2*halvaar + &3*eksald + &4*kvinde +
    &5*matB + &6*matM + &7*sproglig + &8*HF + &9*HHX +
    &10*HTX + &11*udland + &12*GIF;
fitted = exp(xbeta) / (1 + exp(xbeta));
loglik = actual * log(fitted) + (1-actual)*log(1-fitted);
&0 = 0.0;
&1 = 0.0;
&2 = 0.0;
&3 = 0.0;
&4 = 0.0;
&5 = 0.0;
&6 = 0.0;
&7 = 0.0;
&8 = 0.0;
&9 = 0.0;
&10 = 0.0;
&11 = 0.0;
&12 = 0.0;
```

Actual value of dependent variable.

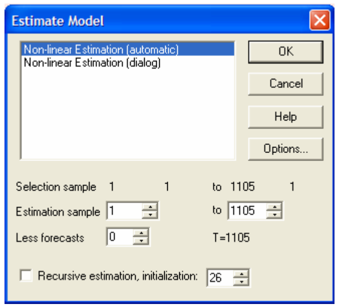
Fitted value of dependent variable. (Probability of passing.)

An auxiliary variable.

Specifies the log-likelihood contribution for one observation.

The parameters have to be given starting values.

Estimate model



PcGive maximizes the log-likelihood function

$\sum_{i=1}^{n} \texttt{loglik}_i$

using an iterative numerical procedure, i.e. it finds the scores, puts them equal to zero and then solves the system of equations for the 13 parameters (&0 to &12) numerically.

(Here it is not possible to solve for the ML-estimators analytically.)

DEPARTMENT OF ECONOMICS
UNIVERSITY OF COPENHAGEN

Results

```
EQ( 1) Modelling actual by ML (using polit.in7)
       The estimation sample is: 1 to 1105

                  Coefficient  Std.Error  t-value  t-prob
&0                  -9.12756     0.8284    -11.0    0.000
&1                   1.07331     0.09302    11.5    0.000
&2                   0.418142    0.1909     2.19    0.029
&3                  -0.0770681   0.03030   -2.54    0.011
&4                   0.176896    0.1535     1.15    0.249
&5                  -0.811773    0.2390    -3.40    0.001
&6                  -0.354658    0.2263    -1.57    0.117
&7                  -0.845276    0.3219    -2.63    0.009
&8                  -1.48581     0.2981    -4.98    0.000
&9                  -0.862620    0.2139    -4.03    0.000
&10                 -0.260739    0.4978    -0.524   0.601
&11                 -0.735209    0.4004    -1.84    0.067
&12                 -1.94947     0.7089    -2.75    0.006
loglik = -632.6335672 for 13 parameters and 1105 observations

Standard errors based on numerical second derivatives
BFGS/warm-up using numerical derivatives (eps1=0.0001; eps2=0.005):
Strong convergence
```

This is to do with the iteration procedure. (If there was no convergence, we wouldn't get any estimation results!)

&0 – intercept

&1 – *kvotient*: Higher high school grade gives better chance of passing.

&2 – *halvaar*: Better to start in the Summer!

&3 – *eksald*: Best to start sooner rather than later!

&4 – *kvinde*: Not significant.

&5-&12 – Any exam other than level "A" mathematics lowers the chance of passing!

Interpretation of the magnitudes of the coefficients

- The signs are interpretable, but not the magnitudes.
- To interpret the model we can look at the marginal effects:

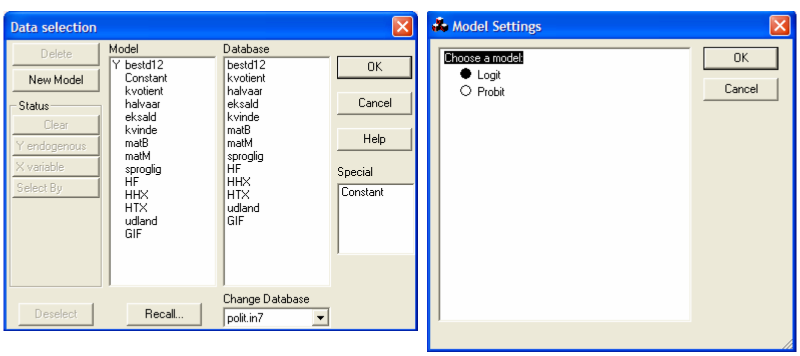
$$\frac{\partial \text{Prob}(y_i = 1 \mid x_i)}{\partial x_{ik}} = \frac{\partial F(x_i'\beta)}{\partial x_{ik}} = F'(x_i'\beta)\,\beta_k$$

- But these depend on the values of *x*.
- Two possible ways around this problem:
  1. Evaluate for the sample mean. Here: the average student. But this often makes little sense, for example here we would be looking at someone who is 29% woman.
  2. Evaluate for a standard individual. But maybe there is a difficulty defining this.

> Something extra: Try saving the fitted values (Test – Store Residuals etc. in Database – Fitted values) – these are the probabilities of passing. Look at observation 377 in the dataset – this is me. I had a 62% probability of passing!

DEPARTMENT OF ECONOMICS
UNIVERSITY OF COPENHAGEN

(4) Package – Limited Dependent Models; Model – Binary Discrete Choice



DEPARTMENT OF ECONOMICS
UNIVERSITY OF COPENHAGEN

Results

CS(1) Modelling bestd12 by Logit
The estimation sample is 1 - 1105

	Coefficient	Std.Error	t-value	t-prob
Constant	-9.12756	0.8284	-11.0	0.000
kvotient	1.07331	0.09302	11.5	0.000
halvaar	0.418137	0.1909	2.19	0.029
eksald	-0.0770580	0.03030	-2.54	0.011
kvinde	0.176897	0.1535	1.15	0.249
matB	-0.811774	0.2390	-3.40	0.001
matM	-0.354646	0.2263	-1.57	0.117
sproglig	-0.845282	0.3219	-2.63	0.009
HF	-1.48571	0.2981	-4.98	0.000
HHX	-0.862616	0.2139	-4.03	0.000
HTX	-0.260759	0.4978	-0.524	0.600
udland	-0.735218	0.4004	-1.84	0.067
GIF	-1.94973	0.7089	-2.75	0.006

log-likelihood	-632.633567	no. of states	2
no. of observations	1105	no. of parameters	13
baseline log-lik	-763.7719	Test: Chi^2(12)	262.28 [0.0000]**
AIC	1291.26713	AIC/n	1.16856754
mean(bestd12)	0.468778	var(bestd12)	0.249025

Newton estimation (eps1=0.0001; eps2=0.005): Strong convergence

	Count	Frequency	Probability	loglik
State 0	587	0.53122	0.53122	-309.2
State 1	518	0.46878	0.46878	-323.4
Total	1105	1.00000	1.00000	-632.6

Identical to the results from (3).

DEPARTMENT OF ECONOMICS
UNIVERSITY OF COPENHAGEN

(5) Goodness-of-fit

First we check whether the (unrestricted) model performs significantly better than a model with only a constant (the restricted model). The null-hypothesis is that there is no significant difference.

The maximized log-likelihood function in the restricted model - *baseline log-lik*

The maximized log-likelihood function in the unrestricted model - *log-likelihood*

$$LR = -2\cdot(\log L_R - \log L_{UR}) = -2\cdot(-763.77-(-632.63)) = 262.28 \sim \chi^2(12)$$

We have imposed 12 restrictions, i.e. that &1 to &12 equal 0. Critical value at 5% level is 21.03. The null is clearly rejected!

Here are some goodness-of-fit measures:

$$McFadden - R^2 = 1 - \frac{\log L_{UR}}{\log L_R} = 1 - \frac{632.63}{763.77} = 0.17$$

$$Pseudo - R^2 = 1 - \frac{1}{1+2(\log L_{UR} - \log L_R)/N} = \frac{1}{1+2(632.63-(-763.77))/1105} = 0.28$$

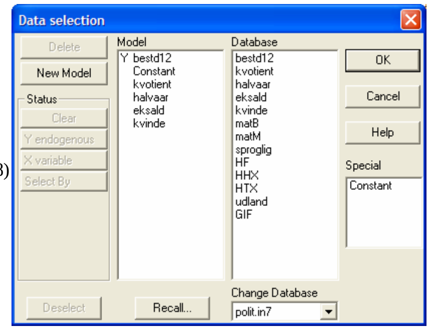
DEPARTMENT OF ECONOMICS
UNIVERSITY OF COPENHAGEN

(6) Test for joint significance of education dummies

The restricted model is the model without the education dummies.

Estimate this model to find the restricted maximum of the log-likelihood function (=-659.12).

$$LR=-2\cdot(\log L_R-\log L_{UR})=-2\cdot(-659.12-(-632.63))=52.98\sim\chi^2(8)$$



This is a clear rejection of the restricted model.

N.B. It is also possible to get PcGive to calculate this automatically by choosing Model – Progress...

Progress to date

Model	T	p		log-likelihood	SC	HQ	AIC
CS(1)	1105	13	Logit	-632.63357	1.2275	1.1908	1.1686
CS(2)	1105	5	Logit	-659.12184	1.2247	1.2106	1.2020

Tests of model reduction (please ensure models are nested for test validity)
CS(1) --> CS(2): Chi^2(8) = 52.977 [0.0000] **

DEPARTMENT OF ECONOMICS
UNIVERSITY OF COPENHAGEN

(7) Test – Further Output – Derivatives of probabilities at regressor means

This calculates $\frac{\partial \text{Prob}(y_i=1\mid x_i)}{\partial x_{ik}}=\frac{\partial F(x_i'\beta)}{\partial x_{ik}}=F'(x_i'\beta)\beta_k$

evaluated for the sample means. Only valid locally!

Derivatives of probabilities at regressor means

Probabilities:

State 0	0.54440 (Fail)
State 1	0.45560 (Pass)

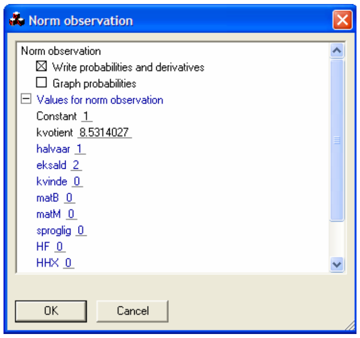
Derivatives:

	mean	State 0	State 1
Constant	1.0000	2.2639	-2.2639
kvotient	8.5314	-0.26621	0.26621
halvaar	0.83258	-0.10371	0.10371
eksald	1.9928	0.019113	-0.019113
kvinde	0.28597	-0.043875	0.043875
matB	0.10136	0.20134	-0.20134
matM	0.11041	0.087962	-0.087962
sproglig	0.051584	0.20965	-0.20965
HF	0.084163	0.36850	-0.36850
HHX	0.12670	0.21395	-0.21395
HTX	0.019005	0.064676	-0.064676
udland	0.027149	0.18235	-0.18235
GIF	0.014480	0.48359	-0.48359

An extra grade increases the probability of passing by 27%-points!

This is the first option discussed in (3).

(8) Test – Norm observation… - Write probabilities and derivatives



This is the second option suggested in (3). Choose some appropriate values. Here I choose someone who has the average qualifying grade, starts in the summer, has had two years between high school and university, is male and has level "A" mathematics.

```
Derivatives of probabilities at specified values
Probabilities:
State 0          0.42720
State 1          0.57280
Derivatives:
```

Has a 57% probability of passing!

	at values	State 0	State 1
Constant	1.0000	2.2335	-2.2335
kvotient	8.5314	-0.26264	0.26264
halvaar	1.0000	-0.10232	0.10232
eksald	2.0000	0.018856	-0.018856
kvinde	0.00000	-0.043287	0.043287
matB	0.00000	0.19864	-0.19864
matM	0.00000	0.086782	-0.086782
sproglig	0.00000	0.20684	-0.20684
HF	0.00000	0.36355	-0.36355
HHX	0.00000	0.21108	-0.21108
HTX	0.00000	0.063808	-0.063808
udland	0.00000	0.17991	-0.17991
GIF	0.00000	0.47710	-0.47710

Now we can see the marginal effects of changing the dummy variables

(9) How to increase the number of students who pass the first year

- Increase the required grade for starting the course.
- Encourage students to start soon after finishing high school.
- Require level "A" in mathematics.
- Make the first year easier! 

(10) Test – Further Output... – Table of actual and predicted

Table of actual and predicted

	State 0	State 1	Sum actual
State 0	434	153	587
State 1	169	349	518
Sum pred	603	502	1105

The model predicted (434+349)/1105 = 70.9% correctly.

If we had a simple model that stated that everybody failed we would predict 587/1105 = 53.1% correctly. This is the relevant benchmark.

(11) Estimate using Probit

CS(2) Modelling bestd12 by Probit
The estimation sample is 1 - 1105

	Coefficient	Std.Error	t-value	t-prob
Constant	-5.48832	0.4798	-11.4	0.000
kvotient	0.643779	0.05366	12.0	0.000
halvaar	0.258930	0.1136	2.28	0.023
eksald	-0.0436182	0.01732	-2.52	0.012
kvinde	0.114272	0.09182	1.24	0.214
matB	-0.487327	0.1428	-3.41	0.001
matM	-0.212651	0.1361	-1.56	0.118
sproglig	-0.497021	0.1917	-2.59	0.010
HF	-0.883615	0.1706	-5.18	0.000
HHX	-0.512194	0.1283	-3.99	0.000
HTX	-0.156295	0.2981	-0.524	0.600
udland	-0.444038	0.2444	-1.82	0.070
GIF	-1.10235	0.3997	-2.76	0.006

log-likelihood	-633.075545	no. of states	2
no. of observations	1105	no. of parameters	13
zeroline log-lik	-765.9276	Test: Chi^2(13)	265.7 [0.0000]**
AIC	1292.15109	AIC/n	1.1693675
mean(bestd12)	0.468778	var(bestd12)	0.249025

Newton estimation (eps1=0.0001; eps2=0.005): Strong convergence

	Count	Frequency	Probability	loglik
State 0	587	0.53122	0.53068	-310.2
State 1	518	0.46878	0.46932	-322.9
Total	1105	1.00000	1.00000	-633.1

The signs are the same. *log-likelihood* is almost identical.

This will usually be the case, since the models are almost identical.

DEPARTMENT OF ECONOMICS
UNIVERSITY OF COPENHAGEN

6.2 Autocorrelation of the ARMA(1,1) model

I will go through this on the blackboard.

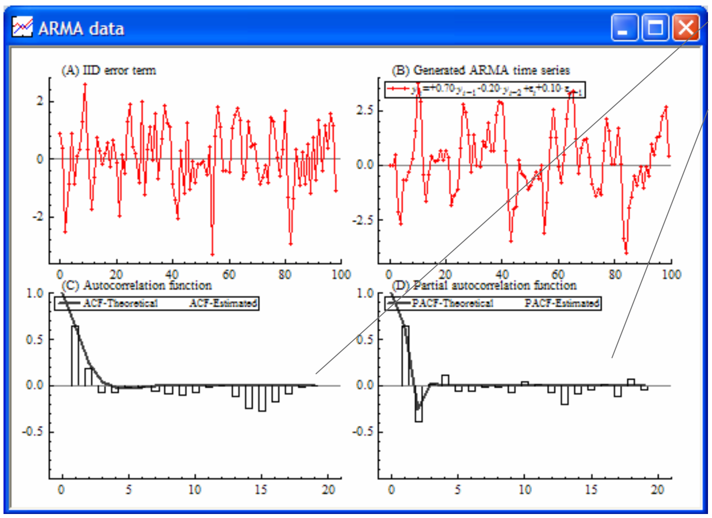
DEPARTMENT OF ECONOMICS
UNIVERSITY OF COPENHAGEN

6.3 Generated ARMA time series

- This exercise involves generating ARMA time series using some Ox code. Try it yourselves!
- You don't need to learn to program, but it is useful to be able to use Ox code.
- This model with T=100:
  $y_t = 0.5 \cdot y_{t-1} - 0.2 \cdot y_{t-2} + \epsilon_t + 0.1 \cdot \epsilon_{t-1}$
- is represented by changing the relevant code to this:

```
obs = 100;
ar = < 0.5 , -0.2 >;
ma = < 0.1 >;
```

(1) Modules – Run Ox



$$\rho_k = \frac{\text{cov}(Y_t, Y_{t-k})}{\text{var}(Y_t)} = \frac{\gamma_k}{\gamma_0}$$

$$\theta_k$$

(2) AR(1) models

- The theoretical ACF is poor when close to a unit root, improves for larger numbers of observations.
- The theoretical ACF is undefined when there is a unit root.

MA(1) models

- Note that for the MA(1) process the autocorrelation coefficients are

$$\rho_1 = \frac{\alpha}{1+\alpha^2}$$

$$\rho_k = 0$$

- so a shock to an MA(1) process affects Y_t in two periods only.

(4) ARMA processes

ARMA(1,1) model with $\theta_1 = -\alpha_1$
e.g.

$$Y_t = 0.7Y_{t-1} + \varepsilon_t - 0.7\varepsilon_{t-1}$$
$$\Leftrightarrow (1-0.7L)Y_t = (1-0.7L)\varepsilon_t$$
$$\Leftrightarrow Y_t = \varepsilon_t$$

a white noise process.

DEPARTMENT OF ECONOMICS
UNIVERSITY OF COPENHAGEN

Next week

- We will have another multiple choice test, so please remember your notes etc.
- We will use an ARIMA model in PcGive.

FINALLY:

- Don’t forget to register for the exam!