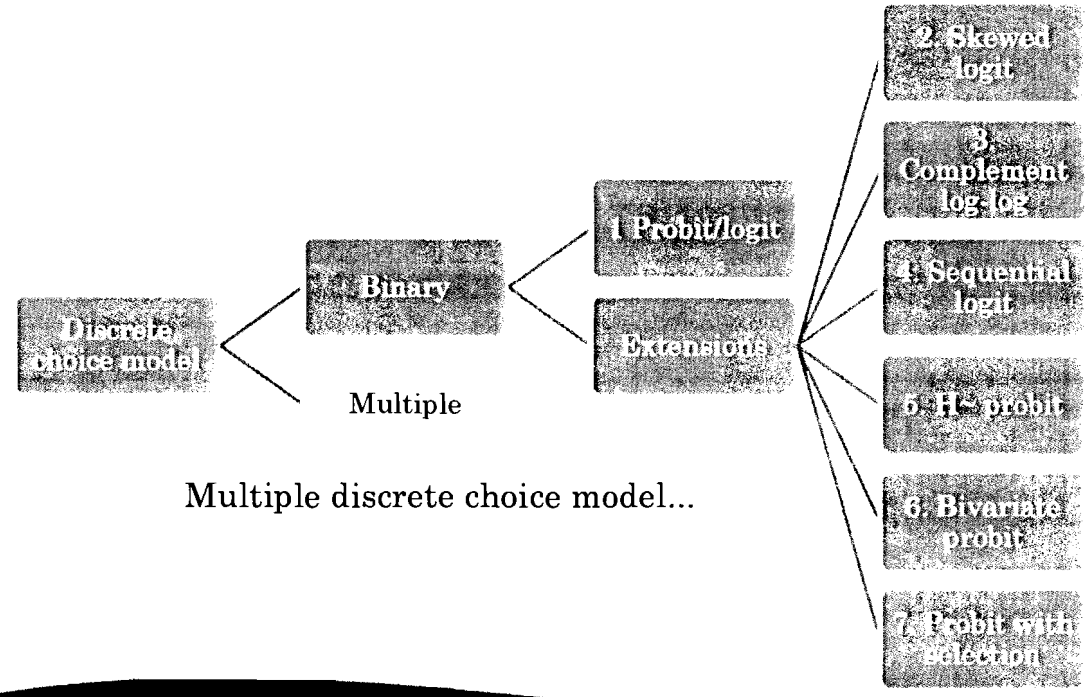
SESSION I: WHY OLS IS NOT SUITABLE? (CONT.)



Multiple discrete choice model...

UUM Universiti Utara Malaysia
The Eminent Management University

SESSION I: WHY OLS IS NOT SUITABLE? (CONT.)

- When the dependent variable takes on just a few discrete values.
- For example, we may be interested in the effect of a number of children on married women's labour force participation decisions, a brand's advertising on consumers' decisions to buy that brand.
- LPM (OLS)
- So, what are the problems?

SESSION I: WHY OLS IS NOT SUITABLE?

- Simple case with one explanatory variable X
- The BIG problem:

  Y values can only be 0 or 1 so a straight line fit through the points will result in predicted Y values outside the range 0-1

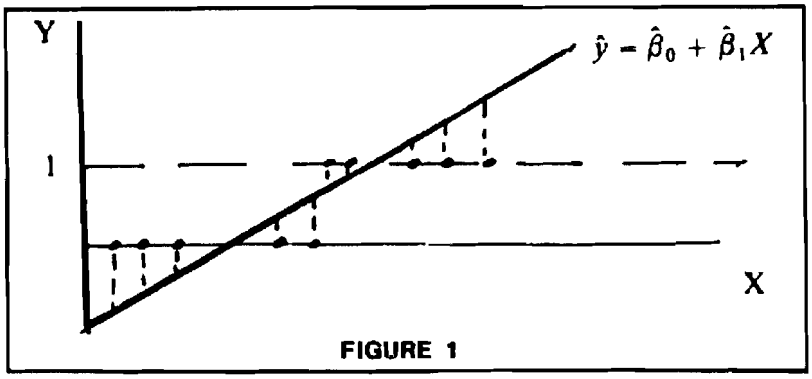


FIGURE 1

UUM Universiti Utara Malaysia
The Eminent Management University

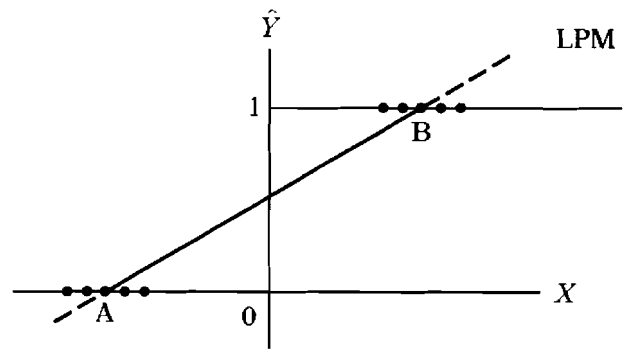
SESSION I: WHY OLS IS NOT SUITABLE?

- Other not so big problem:

  - Residuals will also be heteroskedastic – so if we do use OLS we should use robust standard errors to calculate t values

  - R squared has no meaning here

- Linear probability model:



PREDICTING A BOND RATING

Based on a pooled time series and cross-sectional data of 200 Aa (high-quality) and Baa (medium-quality) bonds over the period 1961–1966, Joseph Cappelleri estimated the following bond rating prediction model:[10]

$$Y_i = \beta_1 + \beta_2 X_{2i}^2 + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + u_i$$

where Y_i = 1 if the bond rating is Aa (Moody's rating)
= 0 if the bond rating is Baa (Moody's rating)
X_2 = debt capitalization ratio, a measure of leverage
$$= \frac{\text{dollar value of long-term debt}}{\text{dollar value of total capitalization}} \cdot 100$$
X_3 = profit rate
$$= \frac{\text{dollar value of after-tax income}}{\text{dollar value of net total assets}} \cdot 100$$
X_4 = standard deviation of the profit rate, a measure of profit rate variability
X_5 = net total assets (thousands of dollars), a measure of size

A priori, β_2 and β_4 are expected to be negative (why?) and β_3 and β_5 are expected to be positive.

After correcting for heteroscedasticity and first-order autocorrelation, Cappelleri obtained the following results[11]:

$$\hat{Y}_i = \underset{(0.1775)}{0.6860} - \underset{(0.0024)}{0.0179 X_{2i}^2} + \underset{(0.0486)}{0.0486 X_{3i}} + \underset{(0.0178)}{0.0572 X_{4i}} + \underset{(0.039)(E\text{-}8)}{0.378(E\text{-}7) X_5} \qquad \textbf{(15.3.1)}$$

$$\bar{R}^2 = 0.6933$$

Thomas F. Pogue and Robert M. Soldofsky, "What Is in a Bond Rating?" *Journal of Financial and Quantitative Analysis*, June 1969, pp. 201–228.

UUM Universiti Utara Malaysia
The Eminent Management University

USING STATA

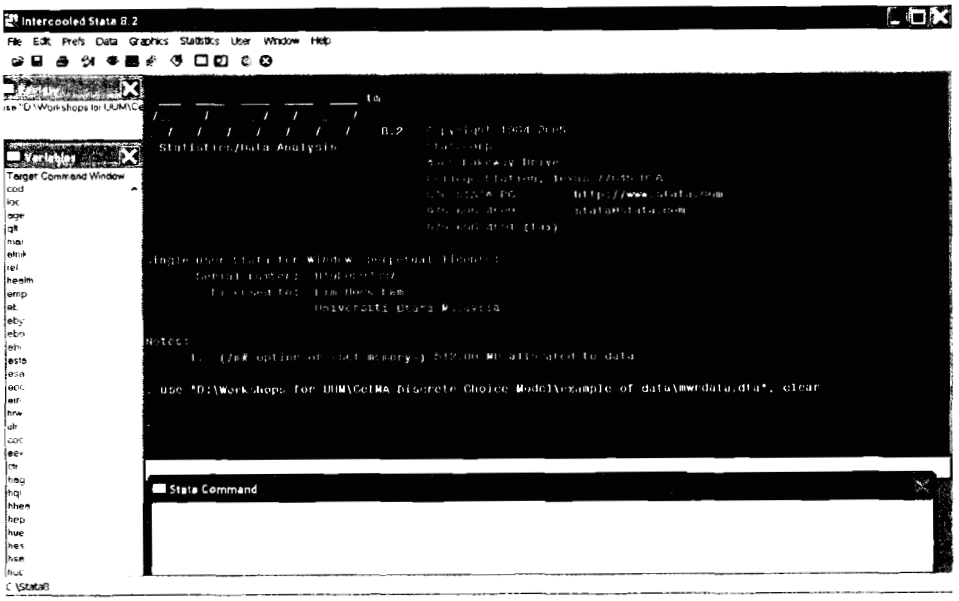
- Level of user: Interactive, do files and ado files
- STATA version 8
- Data

  - mwrdata.dta (2002 – old data)

  Lim, H.E., Zalina Mohd Mohaideen & Norehan Abdullah. (2003). Penyertaan tenaga buruh wanita berkahwin di Kedah: kesan institusi agama, faktor anak dan pendidikan. *Jurnal Ekonomi Malaysia*, 37, 49-79.

- Let us have a look at the data and estimate a LPM using STATA (a brief and hands on introduction to STATA included)

LET US ENJOY STATA...



UUM Universiti Utara Malaysia
The Eminent Management University

SESSION II: THEORY AND APPLICATION

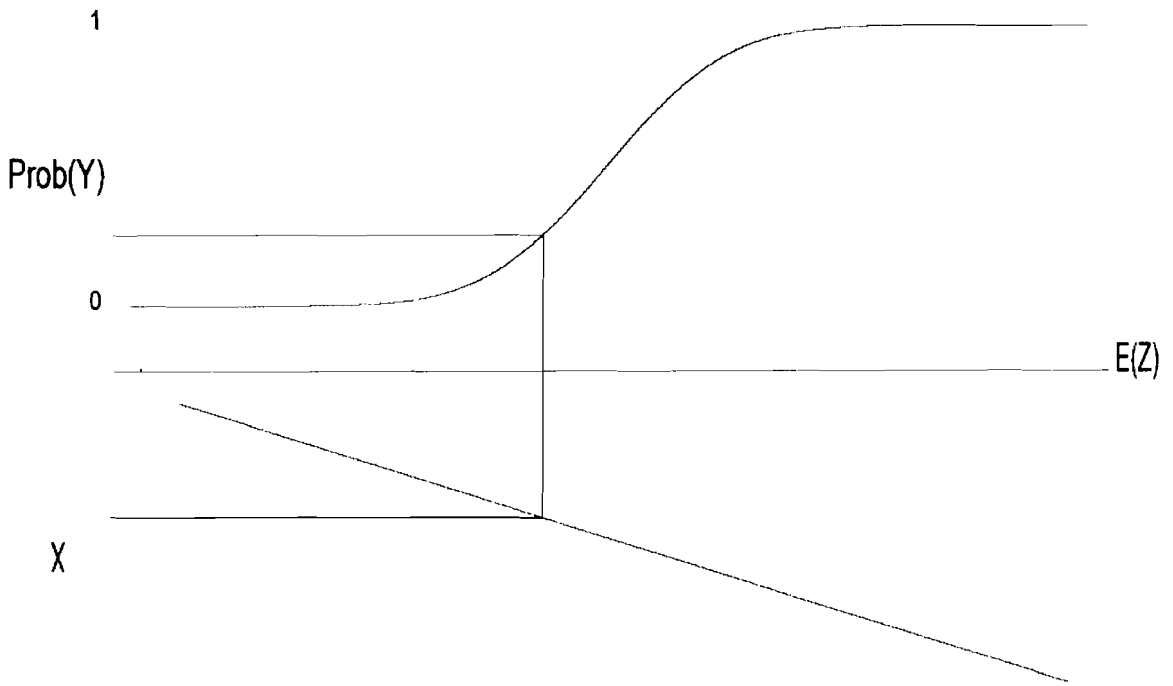
- RECALL: BIG problem of OLS – interpretation!
- Linear regression methods predict values between $-\infty$ and $+\infty$.
- Probabilities must fall between 0 and 1.
- The linear probability model cannot guarantee sensible predictions.
- We need a TRANSFORMER!

SESSION II: THEORY AND APPLICATION (CONT.)

- The transformer - a "squash function", G, to ensure that the fitted values lie strictly between 0 and 1
  - Logit Model: G follows a logistic distribution
  - Probit Model: G follows a cumulative normal distribution

- The differences between the two models are subtle.
- There is almost no practical difference between the two models.

UUM Universiti Utara Malaysia
The Eminent Management University



Methodology:

Lim, H.E. (2013). Overeducation and happiness in the Malaysian graduate labour market. *International Journal of Business and Society*, 14(1), forthcoming

Assume that for each employed graduates, there is a latent variable that represent his or her tendency to be overeducated. This overeducated tendency is associated with individual characteristics of the graduate (x_i.) Let y^* represent this latent variable and assume that y^* is a linear function of x_i, then,

$$y_i^* = \sum_{i=1}^{n} \beta x_i + u_i \quad ...(1)$$

where

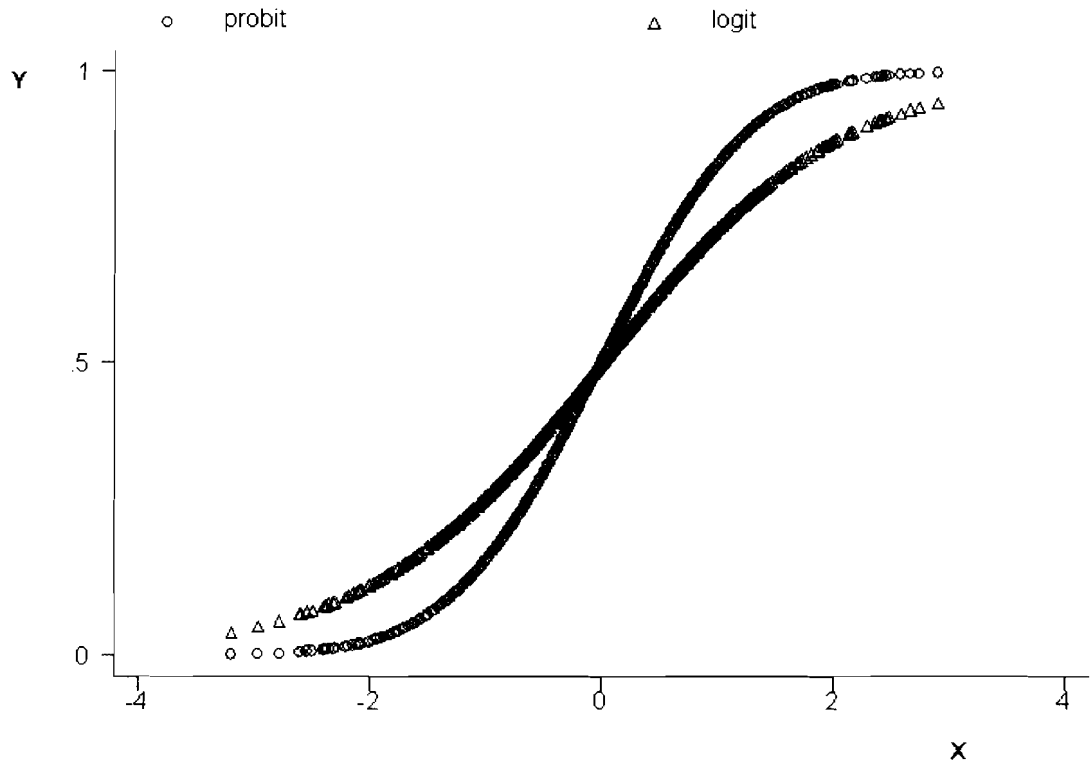
y^* = the unobserved tendency to be employed
x = the individual characteristics
u = the error term

UUM Universiti Utara Malaysia
The Eminent Management University

If y is the random variable that represent the observed outcomes, j, of the graduate, where j=1 if overeducated, j=0 if otherwise. Assume that the error term follows a normal distribution, we have the probit model. The probability of overeducated can be specified as below:

$$Prob(y = 1 \mid x) = Prob(y^* > 0) = Prob(\beta x + u > 0) = Prob(u > -\beta x)$$
$$= Prob(u < \beta x) = F(\beta x)$$

- The Prob(y=0) = 1- Prob(y=1) = 1-F(Bx)
- The F(Bx) represents the CDF, i.e., the transformer!
- F(.): CDF of standard normal function – Probit
- F(.): CDF of logistic function - Logit



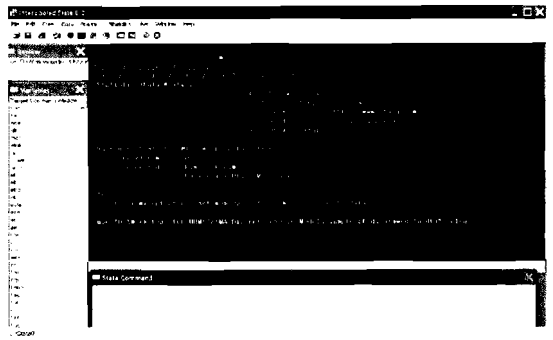
UUM Universiti Utara Malaysia
The Eminent Management University

SESSION II: THEORY AND APPLICATION (CONT.)

- Notice that the slope varies dramatically.
- When there is very very likely or very very unlikely to success (y=1 or y=0), a small change in the X has very little impact.
- When the team's chance of success is 50/50, a small change in the X can lead to a large change in probabilities.

SESSION II: THEORY AND APPLICATION (CONT.)

- Let us try in STATA
- Use the mwrdata.dta again



STATA is back...

UUM Universiti Utara Malaysia
The Eminent Management University

SESSION III: APPLICATION EXAMPLE 1

- Application example 1
- Lim, H.E., Zalina Mohd Mohaideen & Norehan Abdullah. (2003). Penyertaan tenaga buruh wanita berkahwin di Kedah: kesan institusi agama, faktor anak dan pendidikan. *Jurnal Ekonomi Malaysia,* 37, 49-79.
- Gap in literature: married women labour force participation decision and religion

SESSION III: APPLICATION EXAMPLE 1 (CONT.)

- Collect the data
- Screen the data to minimize typing errors
- Descriptive statistics: describe your dependent variables, focus independent variables, and related both (if you wish to)
- Estimating your probit/logit model:
  - goodness of fit statistics

	%	Bil
Penyertaan Guna Tenaga:		
Bekerja	63.87	449
Tidak bekerja	36.13	254
Agama:		
Islam	88.48	622
Bukan Islam	11.52	81
Pendidikan:		
Tidak pernah bersekolah	13.41	94
Tidak tamat Sekolah Rendah	12.84	90
Tamat Sekolah Rendah	25.39	178
Tidak tamat sekolah menengah	13.84	97
Tamat sekolah menengah	28.82	202
Tamat kolej/maktab/STPM	4.42	31
Tamat universiti	1.28	9
Umur:		
18 sehingga 30	21.34	150
31 sehingga 42	37.41	263
43 sehingga 54	25.60	180

Bilangan anak < 6 tahun	%	Bil
0	45.51	299
1	29.38	193
2	18.87	124
3 sehingga 5	6.24	41
Bilangan anak 6 - 19 tahun		
0	32.88	216
1	22.07	145
2	20.4	134
3	12.79	84
4	7.76	51
5 sehingga 9	4.1	27
Bilangan anak 20 dan atas		
0	53.58	352
1	9.59	63
2	9.74	64
3	10.2	67
4	7.31	48
5 sehingga 9	9.59	63

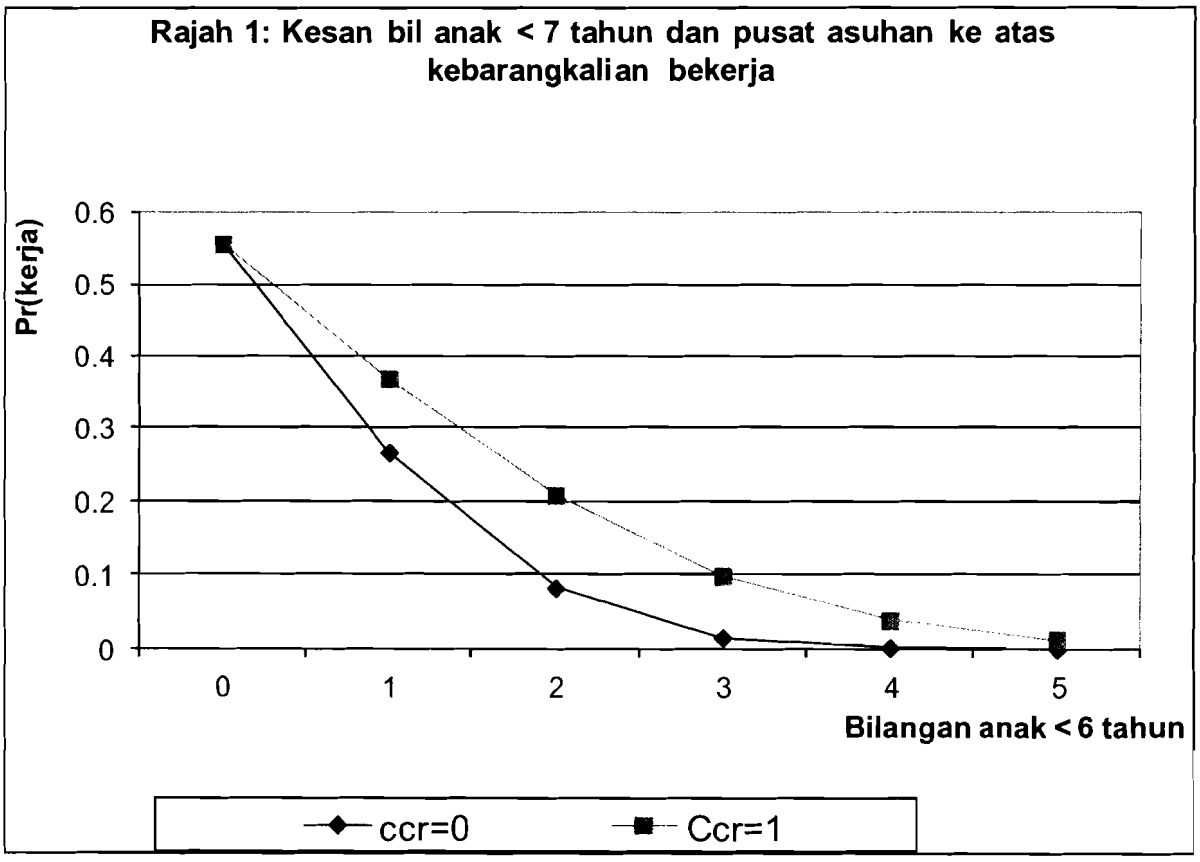
	Model asal		Model baru	
Pemboleh ubah	Pekali Dianggarkan	Nilai-p	Pekali Dianggarkan	Nilai-p
Dbdr	-0.6106	0.004***	-0.5687	0.003***
Age	0.3030	0.008***	0.2351	0.017**
age2	-0.0041	0.003***	-0.0033	0.005***
Dqlf1	-0.5363	0.234	-0.5063	0.213
Dqlf2	-0.5422	0.162	-0.4078	0.24
Dqlf3	-0.4862	0.268	-0.3590	0.365
Dqlf4	-0.7948	0.075*	-0.7931	0.044**
Dqlf5	-0.3936	0.565	-0.1460	0.836
Dqlf6	0.3416	0.771	-0.3223	0.749
EexDqlf1	0.0102	0.768	0.0008	0.978
EexDqlf2	0.0548	0.074*	0.0387	0.167
EexDqlf3	0.0433	0.293	0.0301	0.422
EexDqlf4	0.1575	0.000***	0.1301	0.001***
EexDqlf5	0.3843	0.001***	0.3573	0.009***
EexDqlf6	0.2495	0.126	0.3026	0.054*
Eex	0.0617	0.009***	0.0741	0.001***

Cco	-0.0004	0.301		
Ocodorn	0.0006	0.682		
Ccr	-0.2019	0.503		
Dhom1	-0.0005	0.999	0.1277	0.658
Dhom2	-1.0635	0.014**	-0.9919	0.014**
Dhom3	-0.4823	0.057*	-0.4704	0.051*
Cage06	-0.7957	0.000***	-0.7647	0.000***
Cage619	0.0577	0.483		
Cage20	0.0130	0.875		
Chp	5.4142	0.019**	4.8710	0.031**
Health	-0.3446	0.429		
Oin	-0.0555	0.769		
Agechp	-0.1708	0.005***	-0.1538	0.01***
Ccrc06	0.3530	0.084*	0.2893	0.025**
Dagama	-0.4104	0.138		
_cons	-0.9919	0.623	-1.8849	0.271
Faktor-faktor Suami[1]	Ujian Kekangan	0.0078***	Ujian Kekangan	0.0751*
Ujian Kebagusan (nilai-p):		0.0000	0.0000	
% correctly predicted:		0.8100	0.8000	
Pseudo R2:		0.4287	0.4000	

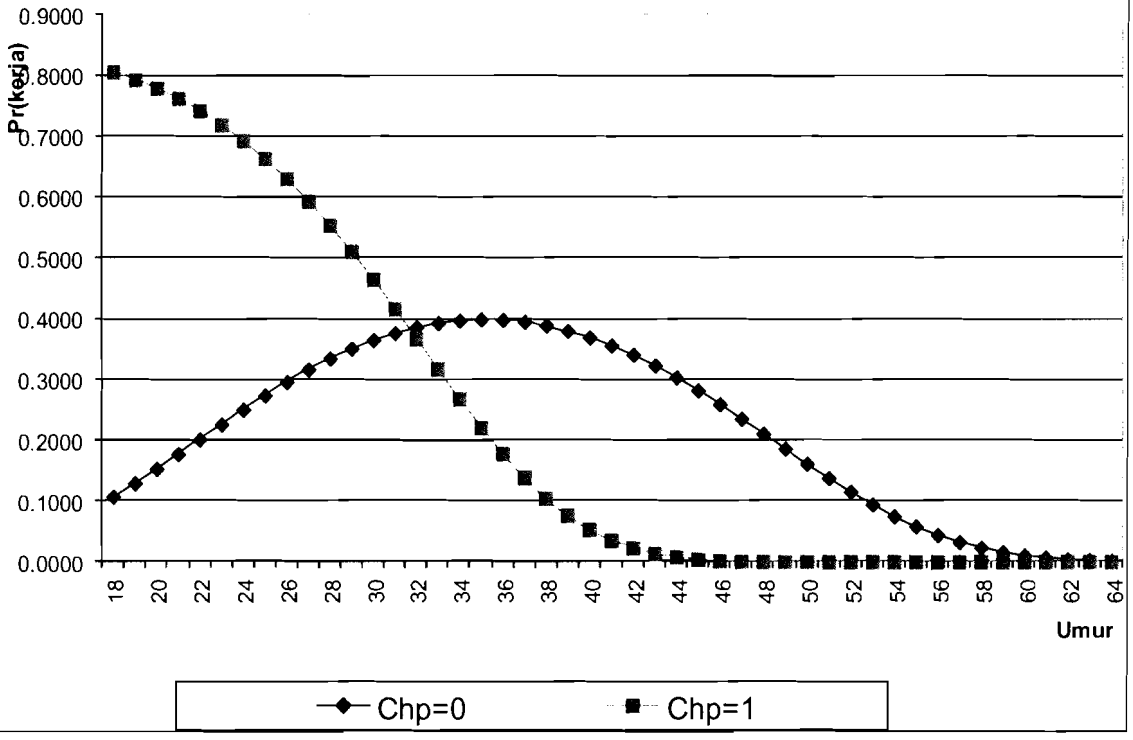
Pemboleh ubah dikekangkan kosong	Nilai-p
Model asal: cco ocodorn ccr cage619 cage20 health Dagama oin	0.7844
Model baru: Dqlf1-6	0.4982

SESSION III: APPLICATION EXAMPLE 1 (CONT.)

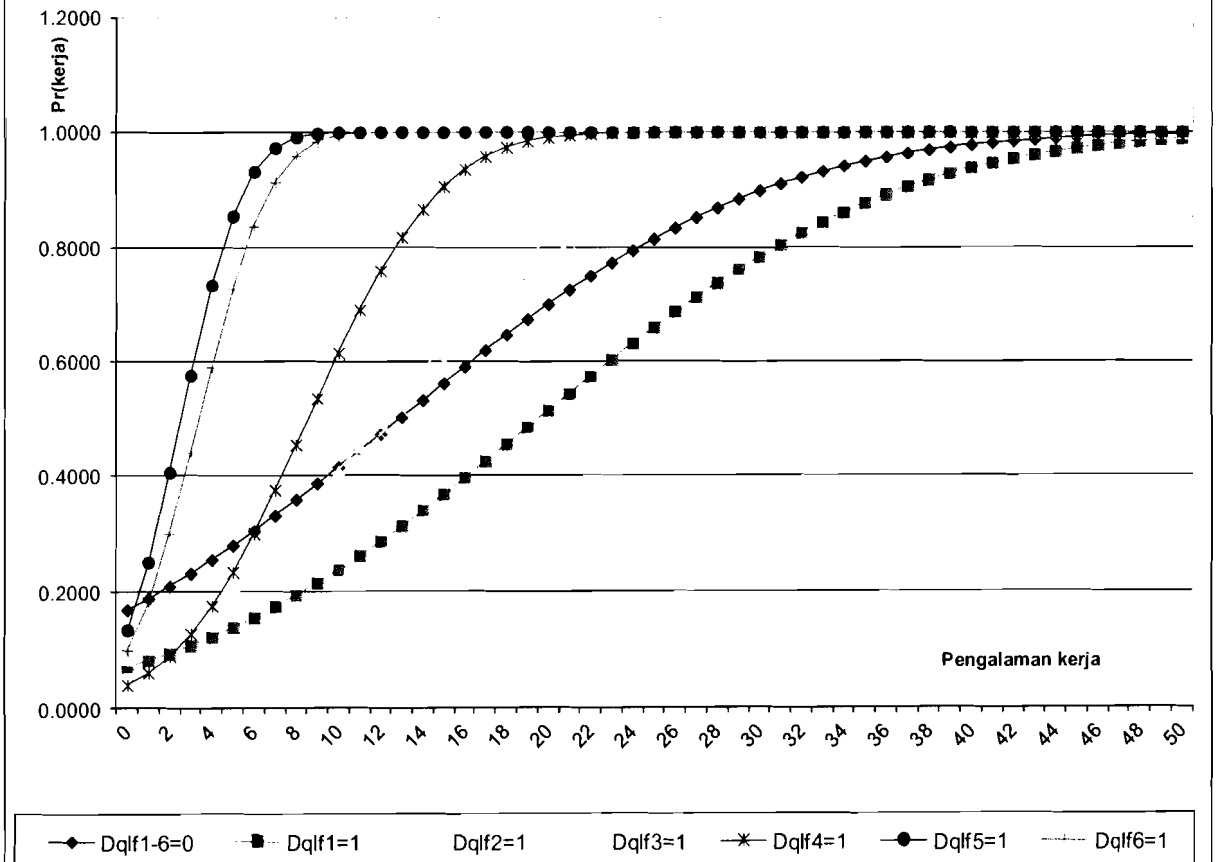
- Sufficient?
- Reviewers might ask for more...
- Prob(Y=1) vs Xs
- Physical calculation using the CDF (old days)
- Now, use the prvalue



Rajah 2: Kesan Bantuan Jagaan Anak dan Umur atas kebarangkalian kerja



Rajah 3: Kesan tahap pendidikan dan pengalaman kerja ke atas kebarangkalian bekerja



SESSION IV: APPLICATION EXAMPLE 2

- Application example 2
- **Lim, H.E.** (2013). Overeducation and happiness in the Malaysian graduate labour market. *International Journal of Business and Society*, 14(1), forthcoming.
- Gap in literature: overeducation – how it relates to happiness?

Categorical variable	Category	%
Gender:	Female	70.13
	Male	29.87
Ethnic group:	Non-Malay	77.92
	Malay	22.08
University:	UUM	68.83
	UTAR	31.17
Home town (rural):	No	43.51
	Yes	56.49
Car driving licence:	No	13.64
	Yes	86.36
Father economically inactive:	No	90.67
	Yes	9.33
Mother economically inactive:	No	40.41
	Yes	59.59
Work during uni vacation:	No	34.64
	Yes	65.36
Practicum/ind training:	No	51.03
	Yes	48.97

Types of degree:	
UUM: Economics	9.09
Public Mgt	3.25
Business Admin	11.04
Accounting	9.74
Communication	4.55
Info Technology	6.49
Others[1]	6.49
HumanRes/SocW	5.19
International Bus	5.84
Finance/banking	7.14
UTAR: Business Admin	9.09
Accounting	11.04
IT/Comp Sciences	5.84
Others[2]	5.19

Continuous variables	Mean	Std Deviation
Age	23.46	1.71
Health	4.42	0.98
Father's education level[3]	4.21	1.79
Mother's education level[3]	3.91	1.75
Family size	6.04	1.66
Self-perceived marketability	4.52	1.17
Happiness (predetermined)	4.96	1.09

Categorical variables	Category	%
Training for job interview/search:	No	79.22
	Yes	20.78
Sharing labour market information:	No	1.99
	Yes	87.01
Overeducation	No	59.09
	Yes	40.91

Non-categorical variables	Mean	Std Deviation
CGPA	3.08	0.28
Happiness (current)	4.56	1.57
Unemployment duration	56.83	53.05
Job application submitted	15.16	20.45
Financial difficulties	2.90	1.17

Variable		%
Overeducation	Yes	40.91
	No	59.09

	Happiness Overall life(predetermined)	
	Mean	Std Deviation
Overeducated: Yes	4.81	0.94
No	5.14	1.10

	Overall life Happiness (current)	
	Mean	Std Deviation
Overeducated: Yes	4.29	1.67
No	5.10	1.33

Variable	Model I Coefficient	Model II Coefficient
HAPPINESS		
Happiness (predetermined)	-0.3020 (0.1495)**	-0.1989 (0.0955)**
OTHER CONTROL VARIABLES:		
Types of degree:		
UUM Economics	2.1941 (0.9635)**	-
UUM Public Mgt	2.3317 (0.9453)**	-
UUM Business Admin	2.5106 (0.9442)***	-
:		
:		
Constant	-2.3737 (5.0493)	0.7500 (0.4906)

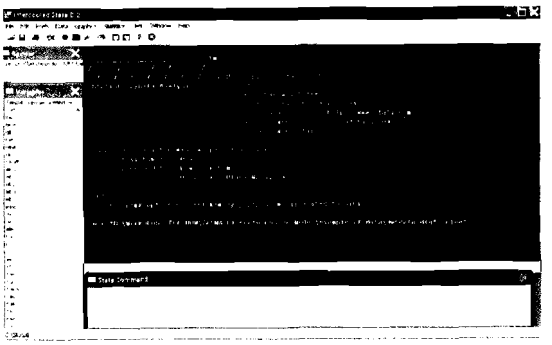
SESSION V: APPLICATION EXAMPLE 3

- Application example 3
- Extension of Probit/Logit model: skewed logit model
- Effects of the regressors on the probability of success are not constrained to be the largest when the probability is 0.5.
- Relax the assumption – skewed logit model:

$$\Pr(y_j \neq 0 \mid \mathbf{x}_j) = 1 - 1 \Big/ \left\{1 + \exp(\mathbf{x}_j \beta)\right\}^{\alpha}$$

SESSION V: APPLICATION EXAMPLE 3 (CONT.)

- Using the overedu1.dta
- Should I use the skewed logit model?
- STATA command: scobit



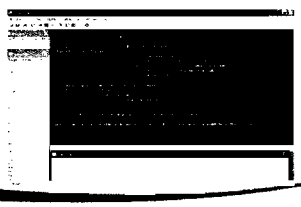
UUM Universiti Utara Malaysia
The Eminent Management University

SESSION V: APPLICATION EXAMPLE 3 (CONT.)

- Data problem: it might occur – % of y=1 (or y=0) is too low
- Would you like to have a increase in your salary?
- How satisfy are you with the compensation that you receive?
- Abandon the data? Collect the data again?

SESSION V: APPLICATION EXAMPLE 3 (CONT.)

- The error distribution of the latent variable follows a standard extreme value distribution
- STATA command: cloglog (complementary log-log model)
- Use HS_beyond1.dta
- Math_scholar: math score more than 70% (tabu this variable and see, what is the % of math scholar?)



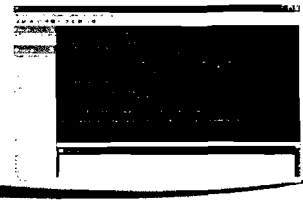
UUM Universiti Utara Malaysia
The Eminent Management University

SESSION V: APPLICATION EXAMPLE 3 (CONT.)

- Problem: reviewer – two stage process
- Example: prob(FS in Malaysia)
  First stage: FS or not
  Second stage: Malaysia or not
- What should we do?

SESSION V: APPLICATION EXAMPLE 3 (CONT.)

- Use nlsw88.dta
- Ed: *highest achieved level of education*: less than high school (1), high school (2), some college (3), and college (4).
- Sequential stages: 1. HS or not 2. HS, then College or not. 3. College, finished or not
- Pr(y=1, y=2,3,4). Pr(y=2, y=3,4). Pr(y=3, y=4)



UUM Universiti Utara Malaysia
The Eminent Management University

SESSION V: APPLICATION EXAMPLE 3 (CONT.)

- If you have STATA 10 and above, you can use the command: seqlogit10

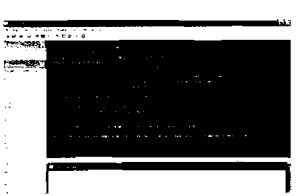
SESSION V: APPLICATION EXAMPLE 3 (CONT.)

- Problem: reviewer – instead of sequential decision, it is a simultaneous decision
- Prob(vote state gov); Prob(vote federal gov)
- Stata command: biprobit
- Use mwrdata.dta

UUM Universiti Utara Malaysia
The Eminent Management University

SESSION V: APPLICATION EXAMPLE 3 (CONT.)

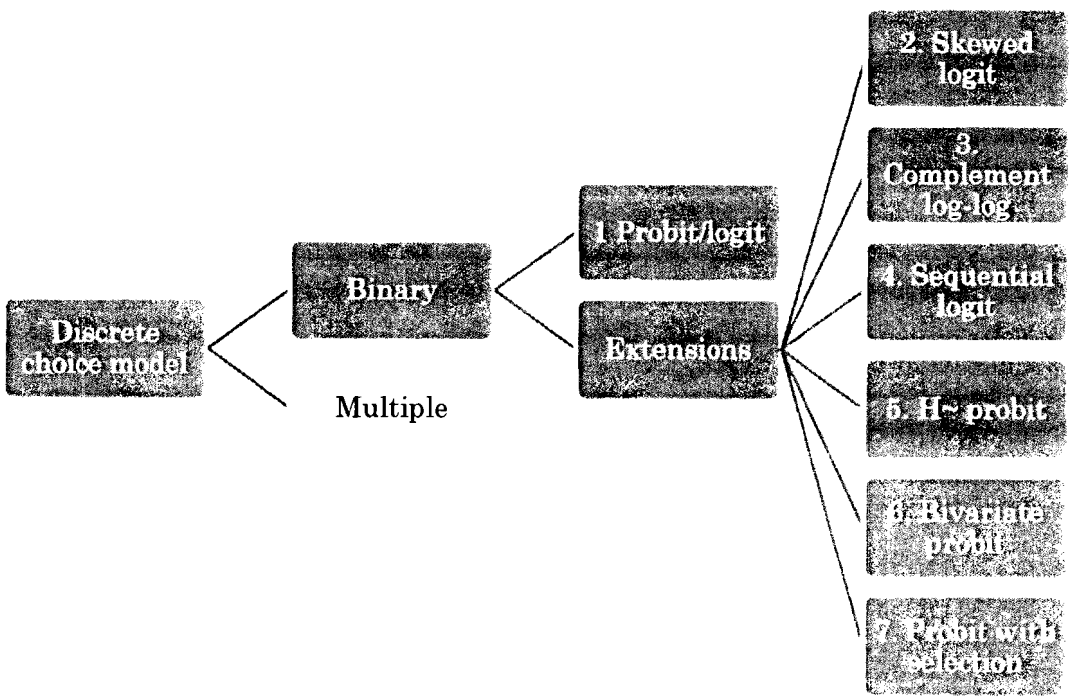
- Problem: reviewer – problem of heteroscedasticiy
- Cross section data – very likely to have
- Test: hettest; imtest, white
- Reject H0 of no hetero. What should we do?
- Stata command: hetprob
- Use mwrdata.dta



SESSION V: APPLICATION EXAMPLE 3 (CONT.)

- Problem: reviewer – problem of sample selection
- High income or not – only those who work (selected), then we can observe.
- Observe whether children attend private school only if the family votes for increasing the property taxes
- Use school_v8.dta

RECAP: what you have been learned?



Thank you