

BSM946: Topic 3: Discrete Choice Models

David Morris

Economics, Finance and Entrepreneurship Group

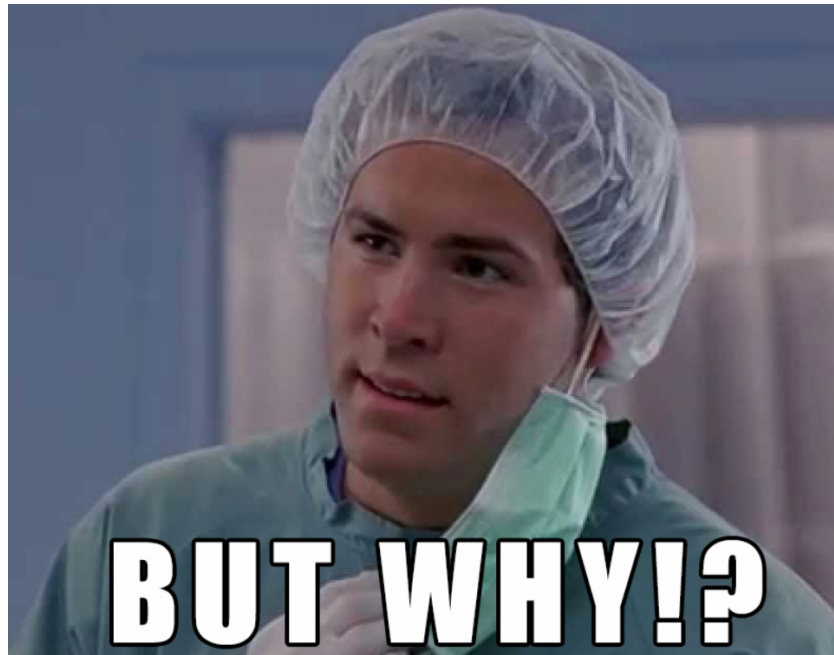
d.morris5@aston.ac.uk

House Keeping

Coursework:

- ① Read the marking framework
- ② Check your work back against the marking framework
- ③ Think about the key words you are using
- ④ Think about the learning outcomes!
 - be able to review the limitations of simple linear regression analysis and the situations where it is not the most appropriate method to use
 - have developed an awareness of which models are appropriate to use when an OLS regression may not be robust, and be able to apply these models correctly
 - command an experienced understanding of the empirical approach to economics, and be able to undertake a robust piece of econometric analysis

Explain....



Today's Plan

- ① What are discrete choice models?
- ② When are they used?
- ③ Linear Probability Model
 - And its flaws
- ④ Probit Model
- ⑤ Logit Model
- ⑥ Nuts and Bolts of Probit and Logit models
 - Applying Probit and Logit models in STATA

What are Discrete Choice Models?

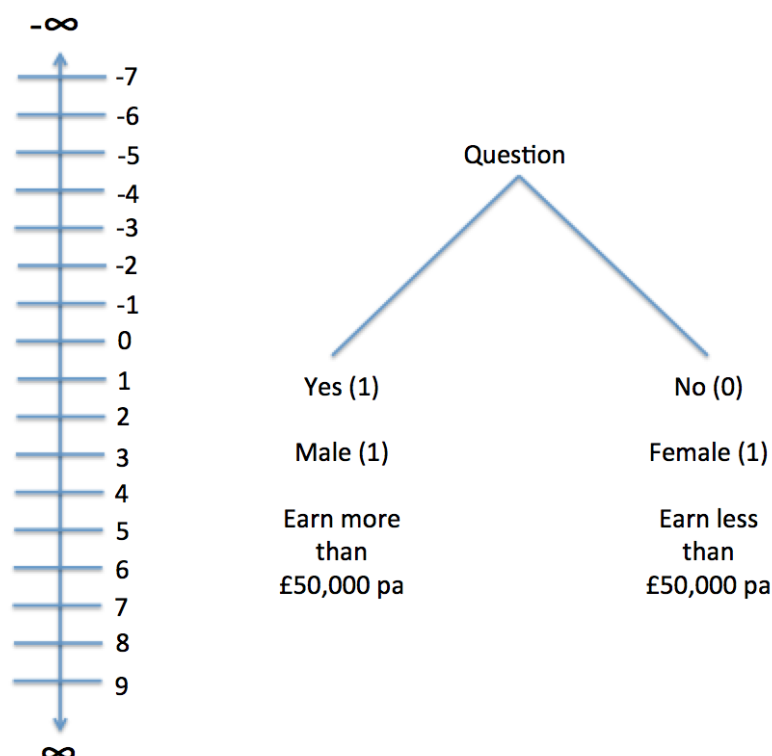
Discrete choice models describe decision maker's choices among alternatives

- The alternatives might represent competing products, courses of action, or any other decision with limited outcomes
- We will look at simple binary options today, can expand to multiple options

Examples include:

- Firm's choice of opening a second store or not (binary choice)
- Different ways of travelling to university: by car, bus, tram or on foot (multiple choice)
- Firm's choice of foreign direct investment locations
- Winner of the best picture award in the Oscars

Discrete Choice



Why is it useful?

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 if a business fails?

Not possible to estimate these properly through linear models!

Today we will discuss binary choice models

Why start here?

- ① Useful for when our outcome variable of interest is binary
- ② The binary choice model is also a good starting point if we want to study more complicated models
 - Can easily extend these models to things such multinomial or ordered response models
- ③ These are central models in applied econometrics
 - Used in various other more complex models: Propensity Score Matching, Heckman Selection Models, etc.

Getting Acquainted with Binary Models



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?

Probabilities are obviously bounded by 0 and 1 so we need to translate our data to this!

So how does a binary variable translate to 0 to 1?

When the dependent variable (y) is binary, we typically outline it as:

- Equal to one for all observations in the data for which the event of interest has happened (success)
- Equal to zero for the remaining observations (failure)

Let's try this with a simple regression approach (called the Linear Probability Model)

Linear Probability Model

Thinking of a simple model:

$$y = \alpha + \beta_1 X + \varepsilon$$

Where: ε is i.i.d with $(0, \sigma^2)$

We know how to deal with this

What if y is **binary** though.... i.e. :

$$y = \begin{cases} 0, & \text{if do not go to University;} \\ 1, & \text{if go to University} \end{cases}$$

Important thing is the conditional expectation of the dependent variable (y), given, the independent variable x

$$\mathbb{E} [y \mid x] = \beta_0 + \beta_1 X + \varepsilon$$

$\mathbb{E}$ of ε is going to be 0 (due to its mean being zero) thus giving...

$$\mathbb{E} [y \mid x] = \beta_0 + \beta_1 X$$

So we can interpret these coefficients as:

- β_0 = Probability that $y=1$, if $x=0$
- β_1 = Change in the probability of $y=1$ due to a 1 unit change in x

Today's Example...



Affairs!!!

The data come from a 1969 survey, filled out by the readers of Psychology Today (results published in the July 1970 issue)

- Sample includes employed men & women, married for the first time (601 observations)

$$\text{Affair} = \beta_0 + \beta_1 \text{Male} + \beta_2 \text{Age} + \beta_3 \text{YearsMarried} + \beta_4 \text{WithKids} \\ + \beta_5 \text{Religiosity} + \beta_6 \text{YearsEducation} + \beta_7 \text{HappinessinMarriage} + u_i$$

where:

- Affair is =1 if the person had at least one affair, 0 otherwise
- Religiosity is a categorical variable from 1 to 5, where 5 is very religious
- Happiness in Marriage is a categorical variable from 1 to 5, where 5 is very happy

bcuse affairs

reg affair male age yrsmarr kids relig educ ratemarr, r

Linear regression

Number of obs = 601
F(7, 593) = 9.61
Prob > F = 0.0000
R-squared = 0.1062
Root MSE = .41189

affair	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
male	.0517169	.0389454	1.33	0.185	-.0247707	.1282046
age	-.0073149	.0031354	-2.33	0.020	-.0134728	-.0011571
yrsmarr	.0160812	.00563	2.86	0.004	.005024	.0271384
kids	.050184	.0449966	1.12	0.265	-.0381881	.1385561
relig	-.0539739	.0153308	-3.52	0.000	-.0840832	-.0238647
educ	.004867	.0078468	0.62	0.535	-.0105439	.020278
ratemarr	-.087728	.0170375	-5.15	0.000	-.1211892	-.0542668
_cons	.729658	.161796	4.51	0.000	.4118952	1.047421

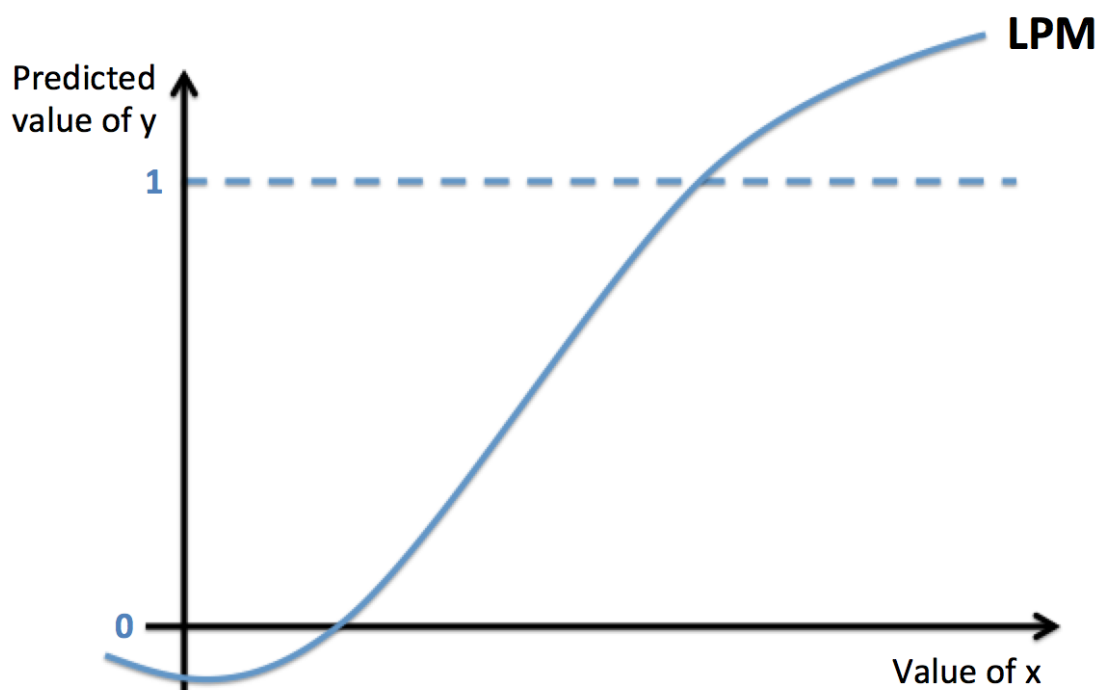
Why would we not just use this model then?

- Simple to interpret the coefficients
- Can easily calculate the probability of a individual with certain characteristics having an affair

The Linear Probability Model has a few problems

- 1 Linear Probability Models can produce fitted values that fall outside of the acceptable range of probabilities at either end of the scale
 - Probabilities are bounded by $0 \leq Prob \leq 1$ but the LPM is bounded by $-\infty \leq \widehat{LPM} \leq \infty$
- 2 Less of a problem, but the model has inbuilt heteroskedasticity
 - Can use robust standard errors to correct this, which has been done in the above example

Fitted values that fall outside of the acceptable range



Example of bad LPM fit

Now, what is the expected probability of having an affair for a very religious (relig=5) 25-year old woman, without kids, who is very happily married (ratemarr=5) for a year and has 9 years of education?

$$P(\widehat{Affair} = 1) = \hat{\beta}_0 + \hat{\beta}_1 Male + \hat{\beta}_2 Age + \hat{\beta}_3 YearsMarried + \\ \hat{\beta}_4 WithKids + \hat{\beta}_5 Religiosity \\ + \hat{\beta}_6 YearsEducation + \hat{\beta}_7 HappinessinMarriage$$

$$P(\widehat{A} = 1) = \hat{\beta}_0 + \hat{\beta}_2(25) + \hat{\beta}_3(1) + \hat{\beta}_5(5) + \hat{\beta}_6(9) + \hat{\beta}_7(5) = -0.1$$

$$P(\widehat{A} = 1) = \\ 0.729 - 0.007(25) + 0.016 - 0.054(5) + 0.005(9) - 0.0877(5) = -0.1$$

How can a probability be negative? It cannot

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

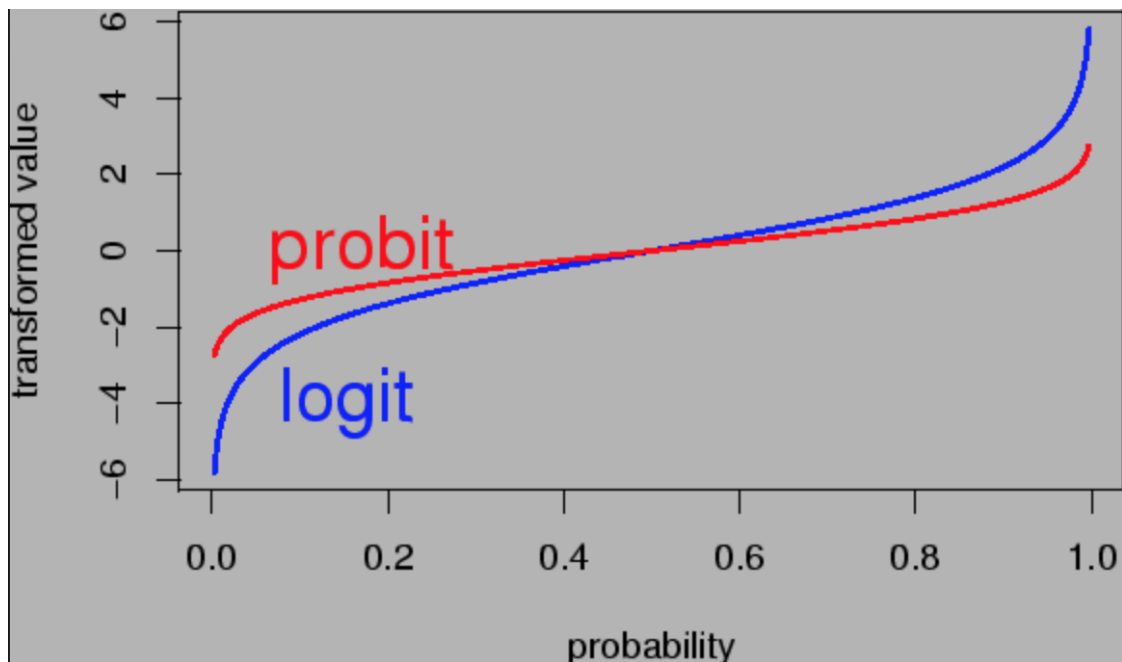
As we shall see next, **Probit** and **Logit** models solve this problem

- Just as easy to implement as LPM/OLS, **but less straightforward to interpret**

LPM tends to predict the mean pretty well, but the partial effects (i.e. what is the probability of an individual with certain characteristics having an affair) is not as accurate

- **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 than at extreme values**

Probit and Logit



Probit and Logit

With the LPM we were looking at:

$$Prob(y = 1 | X) = \beta_0 + \beta_1 x$$

But it was not bounded by 0 and 1

The idea with these two models is to transform this model but a given function (F), so it is bounded

$$Prob(y = 1 | X) = F(\beta_0 + \beta_1 x)$$

Where $F(-\infty) = 0$ and $F(\infty) = 1$

Logit Model

The transformations ensure both the Logit and the Probit lie $[0, 1]$

The two models are just different transformations of the standard model we have looked at

Logit is given by:

$$F(Z) = \frac{1}{1 + \exp^{-Z}} = \text{often referred to as } \Lambda(Z)$$

Why use this?

- $Z \rightarrow \infty$: $\exp^{-Z}$ tends to 0 giving $\frac{1}{1}$ equal to 1
- $Z \rightarrow -\infty$: $\exp^{-Z}$ tends to infinity leaving $\frac{1}{\infty}$, ≈ 0

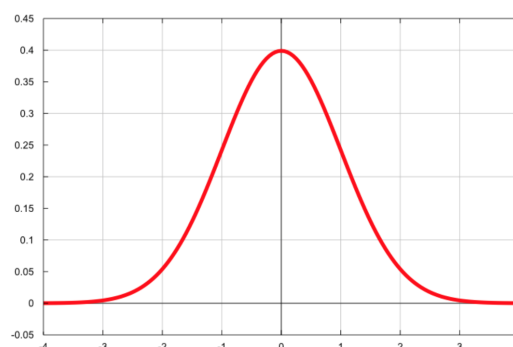
Probit Model

Transformation with the Probit model is:

$$F(Z) = \int_{-\infty}^Z \Phi(u) du$$

Integral of normal probability density function between $-\infty$ and Z

- Area under the normal probability density function is equal to one



STATA Implementation: Probit

probit affair male age yrsmarr kids relig educ ratemarr, r

```

Probit regression                               Number of obs   =       601
                                                Wald chi2(7)    =       54.93
                                                Prob > chi2     =       0.0000
Log pseudolikelihood = -305.2525              Pseudo R2      =       0.0961

```

affair	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
male	.1888175	.1319271	1.43	0.152	-.0697548	.4473898
age	-.0243996	.0111237	-2.19	0.028	-.0462017	-.0025975
yrsmarr	.054608	.0189634	2.88	0.004	.0174405	.0917755
kids	.2080754	.1662426	1.25	0.211	-.117754	.5339049
relig	-.1860854	.0532405	-3.50	0.000	-.2904348	-.081736
educ	.0155052	.0263552	0.59	0.556	-.0361501	.0671605
ratemarr	-.2727101	.0533916	-5.11	0.000	-.3773558	-.1680644
_cons	.76416	.534335	1.43	0.153	-.2831173	1.811437

STATA Implementation: Logit

logit affair male age yrsmarr kids relig educ ratemarr, r

```

Logistic regression                               Number of obs   =       601
                                                Wald chi2(7)    =       53.00
                                                Prob > chi2     =       0.0000
Log pseudolikelihood = -304.84818              Pseudo R2      =       0.0973

```

affair	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
male	.3164296	.2290344	1.38	0.167	-.1324697	.7653288
age	-.0437373	.0190217	-2.30	0.021	-.0810191	-.0064555
yrsmarr	.0952295	.0323227	2.95	0.003	.0318782	.1585808
kids	.3791557	.2931454	1.29	0.196	-.1953987	.95371
relig	-.3254421	.0924565	-3.52	0.000	-.5066535	-.1442307
educ	.0305846	.0451685	0.68	0.498	-.0579441	.1191132
ratemarr	-.4700074	.0915244	-5.14	0.000	-.649392	-.2906228
_cons	1.336704	.9258706	1.44	0.149	-.4779688	3.151377

Logit and Probit Coefficients

- So far we have only been discussing fit.... we need to be able to interpret coefficients too
- How to interpret Logit results?
- How to interpret Probit results?
- STATA Example

Previously....

- 1 Use logit or probit models when ever your dependent variable is binary (also called dummy) which takes values 0 or 1
- 2 Logit and probit regression are non-linear regression models that forces the output (predicted values) to be either 0 or 1
- 3 Logit and probit models estimate the probability of your dependent variable to be 1 ($Y=1$). This is the probability that some event happens.
- 4 Cannot read off the coefficients as in OLS

Logit Coefficients

Logit coefficients are in log-odds units and cannot be read as regular OLS coefficients

They represent the log of the odds of $Y=1$ when X increases by 1 unit

If we convert them to odds ratios.....

- ① If the OR > 1 then the odds of $Y=1$ increases
- ② If the OR < 1 then the odds of $Y=1$ decreases

Look at the sign of the logit coefficient

Dummy Variables:

$$[\exp(\text{coefficient}) - 1] \times 100$$

Continuous Variables:

$$[\exp(\text{coefficient}) - 1] \times 100$$

So we can get the odds ratio by exponentiating the coefficients

STATA Example

Use an inbuilt STATA dataset `sysuse nlsw88.dta`

Look at whether an individual is married or not (married =1 if married and 0 if not)

Dichotomous variable so will look at logits and probits...

Independent variables:

- age - Continuous variable from 34 to 46 in this sample
- race - Dummy variable =0 if white, 1 otherwise
- grade - Continuous from 0 to 18
- south - Dummy variable =1 if in the south, 0 otherwise
- wage - Continuous variable from 1 to 40

STATA Example

```
. logit married age race grade south wage
```

```
Iteration 0:  log likelihood = -1463.463
Iteration 1:  log likelihood = -1408.9033
Iteration 2:  log likelihood = -1408.7176
Iteration 3:  log likelihood = -1408.7176
```

Logistic regression

```
Number of obs   =      2244
LR chi2(5)      =     109.49
Prob > chi2     =     0.0000
Pseudo R2      =     0.0374
```

Log likelihood = -1408.7176

married	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
age	-.0230983	.0148391	-1.56	0.120	-.0521824	.0059859
race	-1.041891	.1033673	-10.08	0.000	-1.244488	-.8392952
grade	.0001168	.0189896	0.01	0.995	-.0371022	.0373357
south	.1969104	.0959908	2.05	0.040	.0087719	.3850489
wage	-.0209626	.0081078	-2.59	0.010	-.0368536	-.0050716
_cons	1.87414	.6453517	2.90	0.004	.6092736	3.139006

Interpreting Coefficients

Continuous Variables:

How do we interpret the coefficient for wage?

We can translate the log odds to odds ratio using the exponentiation

Given the coefficient on wages: -0.02096

$$\begin{aligned}\text{Odds ratio} &= (\exp(-0.02096) - 1) \times 100 \\ &= (0.979 - 1) \times 100 \\ &= -0.021 \times 100 = -2.1\%\end{aligned}$$

So we can say for a one-unit increase in wage, we expect to see about a 2.1% decrease in the odds of being married

Interpreting Coefficients

Dummy Variables:

Again, can translate the log odds to odds ratio using the exponentiation

Given the coefficient on race: -1.041

$$\begin{aligned}\text{Odds ratio} &= (\exp(-1.041) - 1) \times 100 \\ &= (0.353 - 1) \times 100 \\ &= -0.647 \times 100 = -64.5\%\end{aligned}$$

The odds for individuals coded as non-white in the dataset being married is 64.5% less than the odds of an individual coded as white

Odds Ratio

Can do this in STATA using **,or**

```
. logit married age race grade south wage, or
```

```
Iteration 0:    log likelihood =  -1463.463
Iteration 1:    log likelihood = -1408.9033
Iteration 2:    log likelihood = -1408.7176
Iteration 3:    log likelihood = -1408.7176
```

```
Logistic regression                                Number of obs   =      2244
                                                    LR chi2(5)      =      109.49
                                                    Prob > chi2     =      0.0000
Log likelihood = -1408.7176                      Pseudo R2      =      0.0374
```

married	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
age	.9771665	.0145003	-1.56	0.120	.9491557	1.006004
race	.3527868	.0364666	-10.08	0.000	.2880885	.4320149
grade	1.000117	.0189918	0.01	0.995	.9635777	1.038041
south	1.217635	.1168818	2.05	0.040	1.00881	1.469686
wage	.9792556	.0079396	-2.59	0.010	.9638172	.9949412
_cons	6.515212	4.204603	2.90	0.004	1.839095	23.08091

Probit Coefficients

The probit regression coefficients give the change in the z-score or probit index for a one unit change in the predictor

For a one unit increase in X, the z-score increases by β

Not easy to manually find an interpretation for these coefficients....

Probit Coefficients

```
. probit married age race grade south wage, r
```

```
Iteration 0: log pseudolikelihood = -1463.463
Iteration 1: log pseudolikelihood = -1417.28
Iteration 2: log pseudolikelihood = -1417.2668
Iteration 3: log pseudolikelihood = -1417.2668
```

```
Probit regression                                Number of obs   =      2244
                                                Wald chi2(5)    =      82.48
                                                Prob > chi2     =      0.0000
Log pseudolikelihood = -1417.2668              Pseudo R2      =      0.0316
```

married	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
age	-.0126476	.0090273	-1.40	0.161	-.0303409	.0050456
race	-.5414551	.0617276	-8.77	0.000	-.662439	-.4204712
grade	.0022606	.0116873	0.19	0.847	-.020646	.0251672
south	.0888138	.0577044	1.54	0.124	-.0242848	.2019124
wage	-.0125216	.0050513	-2.48	0.013	-.022422	-.0026213
_cons	1.594881	.4079203	3.91	0.000	.7953724	2.394391

In linear regression, if the coefficient on x is β , then a 1-unit increase in x increases Y by β

But what exactly does it mean that the probit coefficient on wage is -0.0125 and significant?

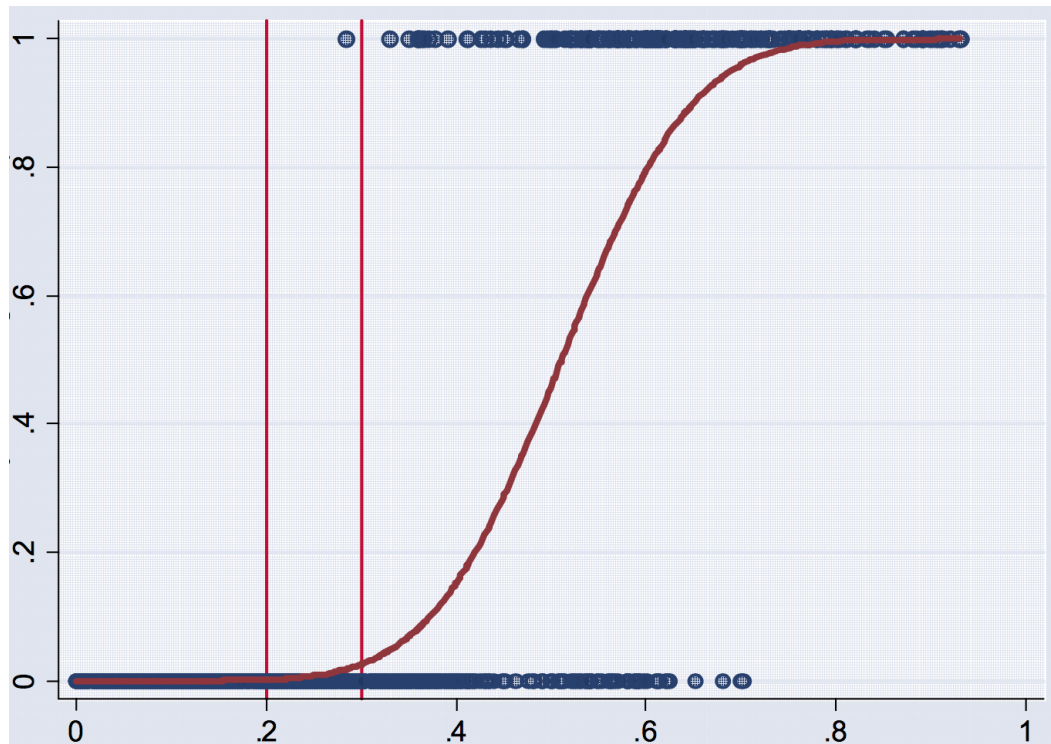
- It means that a 1% increase in wage will decrease the z-score of $\Pr(\text{Married}=1)$ by -0.0125
- And this coefficient is different from 0 at the 5% significance level

So raising wage has a constant effect on the z-score...

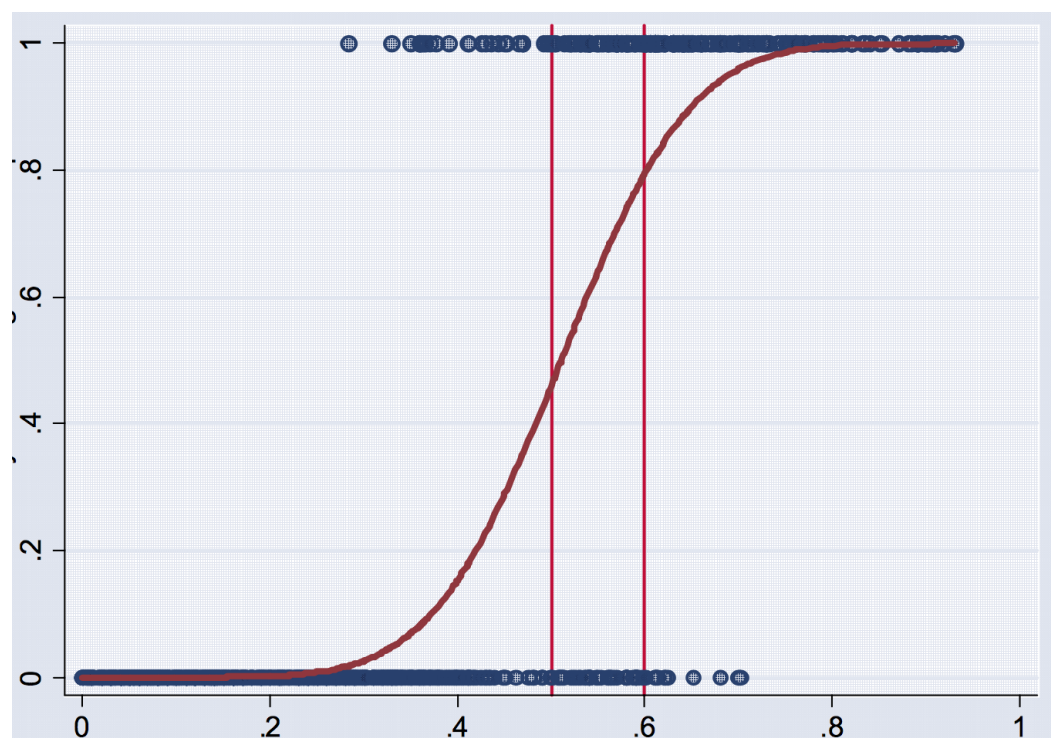
- But this doesn't translate into a constant effect on the original Y
 - This depends on your starting point.....

Similar to what we saw last week when we needed to look at marginal effects

For instance, raising wage from 0.2 to 0.3 has little appreciable impact on $\Pr(\text{Married})$



But increasing wage from 0.5 to 0.6 does have a big impact on the probability of being married



So if you want to calculate the impact of X_i on Y you have to choose values for all other variables X_j

Numerous different ways you can do this....

- One way is to set all variables at their mean
- Can also let the variables vary from their min to their max and look at the change across the observed range

We will use the average method above today... this gives average marginal effects

margins, dydx(*)

Now we can see that increasing the wage by 1, on average, decreases the probability of being married by 0.4%

- Only an average!

```
. margins, dydx(*)
```

```
Average marginal effects          Number of obs   =      2244
Model VCE      : OIM
```

```
Expression      : Pr(married), predict()
dy/dx w.r.t.    : age race grade south wage
```

	Delta-method				[95% Conf. Interval]	
	dy/dx	Std. Err.	z	P> z		
age	-.0045606	.0032419	-1.41	0.159	-.0109146	.0017933
race	-.195245	.019801	-9.86	0.000	-.2340542	-.1564359
grade	.0008152	.0041611	0.20	0.845	-.0073405	.0089708
south	.0320257	.0207406	1.54	0.123	-.0086251	.0726764
wage	-.0045152	.0017831	-2.53	0.011	-.0080101	-.0010204

Important to tell STATA which variables are dummies if using margins! So should have run....

```
. probit married age i.race grade i.south wage
```

```
Iteration 0:  log likelihood = -1463.463
Iteration 1:  log likelihood = -1408.6493
Iteration 2:  log likelihood = -1408.6324
Iteration 3:  log likelihood = -1408.6324
```

```
Probit regression                               Number of obs   =       2244
                                                LR chi2(5)      =       109.66
                                                Prob > chi2     =       0.0000
Log likelihood = -1408.6324                    Pseudo R2      =       0.0375
```

married	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
age	-.0139531	.0090329	-1.54	0.122	-.0316572	.003751
race						
white	-.6438341	.0635291	-10.13	0.000	-.7683489	-.5193192
grade	5.00e-06	.0115858	0.00	1.000	-.0227028	.0227128
1.south	.121001	.0583221	2.07	0.038	.0066919	.2353102
wage	-.0128143	.004973	-2.58	0.010	-.0225612	-.0030674
_cons	1.145929	.3917018	2.93	0.003	.3782079	1.913651

```
. margins, dydx(*)
```

```
Average marginal effects                               Number of obs   =       2244
Model VCE      : OIM
```

```
Expression   : Pr(married), predict()
dy/dx w.r.t. : age 1.race grade 1.south wage
```

	Delta-method					
	dy/dx	Std. Err.	z	P> z	[95% Conf. Interval]	
age	-.004998	.0032308	-1.55	0.122	-.0113302	.0013342
race						
white	-.2448406	.0239766	-10.21	0.000	-.2918339	-.1978474
grade	1.79e-06	.0041501	0.00	1.000	-.0081322	.0081358
1.south	.0430785	.0205918	2.09	0.036	.0027193	.0834376
wage	-.0045901	.0017736	-2.59	0.010	-.0080664	-.0011138

Note: dy/dx for factor levels is the discrete change from the base level.

Predicted Probabilities

So far we have only been discussing coefficients, these models are useful for predicted probabilities as well....

- Let's go back to the first example we looked at...

Predicted Probabilities

Let's start with a simple linear model:

```
. regress affair male age yrsmarr kids relig educ ratemarr, r
```

Linear regression

Number of obs = 601
F(7, 593) = 9.61
Prob > F = 0.0000
R-squared = 0.1062
Root MSE = .41189

affair	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
male	.0517169	.0389454	1.33	0.185	-.0247707	.1282046
age	-.0073149	.0031354	-2.33	0.020	-.0134728	-.0011571
yrsmarr	.0160812	.00563	2.86	0.004	.005024	.0271384
kids	.050184	.0449966	1.12	0.265	-.0381881	.1385561
relig	-.0539739	.0153308	-3.52	0.000	-.0840832	-.0238647
educ	.004867	.0078468	0.62	0.535	-.0105439	.020278
ratemarr	-.087728	.0170375	-5.15	0.000	-.1211892	-.0542668
_cons	.729658	.161796	4.51	0.000	.4118952	1.047421

What is the expected probability of having an affair for a 25 year-old woman, high school graduate (i.e., 12 years of education), without kids, who is married for 3 years, estimates her religiousness level as average (=3) and is very happy (=5) in her marriage?

$$\begin{aligned} P(\widehat{Affair} = 1) = & \hat{\beta}_0 + \hat{\beta}_1 Male + \hat{\beta}_2 Age + \hat{\beta}_3 YearsMarried + \\ & \hat{\beta}_4 WithKids + \hat{\beta}_5 Religiosity \\ & + \hat{\beta}_6 YearsEducation + \hat{\beta}_7 HappinessinMarriage \end{aligned}$$

$$P(\widehat{A} = 1) = \beta_0 + \beta_2(25) + \beta_3(3) + \beta_5(3) + \beta_6(12) + \beta_7(5)$$

$$\begin{aligned} P(\widehat{A} = 1) = \\ 0.730 - 0.007(25) + 0.016(3) - 0.054(3) + 0.005(12) - 0.088(5) \end{aligned}$$

$$P(\widehat{A} = 1) = 0.05$$

Can do this using STATA... Given that we want:

$$P(\widehat{A} = 1) = \beta_0 + \beta_2(25) + \beta_3(3) + \beta_5(3) + \beta_6(12) + \beta_7(5)$$

Using STATA:

```
sca z =_b[_cons]+_b[age]*25+_b[yrsmarr]*3+_b[relig]*3+
_b[educ]*12+_b[ratemarr]*5
```

```
display z
```

Gives the answer: 0.0529

How will the probability of an affair change if she goes from being very happy (ratemarr=5) to very unhappy (ratemarr=1) in the marriage (all else constant)?

$$\Delta P(\widehat{A} = 1) = \text{Final } P(\widehat{A} = 1) - \text{Initial } P(\widehat{A} = 1)$$

$$P(\widehat{A} = 1) = \beta_0 + \beta_2(25) + \beta_3(3) + \beta_5(3) + \beta_6(12) + \beta_7(1) -$$

$$[\beta_0 + \beta_2(25) + \beta_3(3) + \beta_5(3) + \beta_6(12) + \beta_7(5)]$$

$$= \hat{\beta}_7(1) - \hat{\beta}_7(5)$$

$$= \hat{\beta}_7(-4)$$

$$= 0.35$$

$$= \hat{\beta}_7(-4) = 0.35$$

The probability of an affair will increase by 35 percentage points

How can we interpret $\beta_{\text{HappinessinMarriage}}$?

It is the change in the probability of a person having at least one affair, for a one-unit change in the rating of how happy the person is in the marriage, holding everything else constant

Be really careful with your interpretations throughout this course!!!

Predicted Probabilities from a Probit

```

Probit regression                               Number of obs   =       601
                                                Wald chi2(7)    =       54.93
                                                Prob > chi2     =       0.0000
Log pseudolikelihood = -305.2525              Pseudo R2      =       0.0961
  
```

affair	Robust		z	P> z	[95% Conf. Interval]	
	Coef.	Std. Err.				
male	.1888175	.1319271	1.43	0.152	-.0697548	.4473898
age	-.0243996	.0111237	-2.19	0.028	-.0462017	-.0025975
yrsmarr	.054608	.0189634	2.88	0.004	.0174405	.0917755
kids	.2080754	.1662426	1.25	0.211	-.117754	.5339049
relig	-.1860854	.0532405	-3.50	0.000	-.2904348	-.081736
educ	.0155052	.0263552	0.59	0.556	-.0361501	.0671605
ratemarr	-.2727101	.0533916	-5.11	0.000	-.3773558	-.1680644
_cons	.76416	.534335	1.43	0.153	-.2831173	1.811437

What is the expected probability of having an affair for a 25 year-old woman, high school graduate (i.e., 12 years of education), without kids, who is married for 3 years, estimates her religiousness level as average (=3) and is very happy (=5) in her marriage?

$$\begin{aligned}
 P(\widehat{Affair} = 1) = & \Phi(\hat{\beta}_0 + \hat{\beta}_1 Male + \hat{\beta}_2 Age + \hat{\beta}_3 YearsMarried + \\
 & \hat{\beta}_4 WithKids + \hat{\beta}_5 Religiosity \\
 & + \hat{\beta}_6 YearsEducation + \hat{\beta}_7 HappinessinMarriage)
 \end{aligned}$$

$$P(\widehat{A} = 1) = \Phi(\hat{\beta}_0 + \hat{\beta}_2(25) + \hat{\beta}_3(3) + \hat{\beta}_5(3) + \hat{\beta}_6(12) + \hat{\beta}_7(5))$$

$$P(\widehat{A=1}) = \Phi[0.764 + (-0.024 \times 25) + (0.055 \times 3) + (-0.186 \times 3) + (0.015 \times 12) + (-0.273 \times 5)]$$

$$P(\widehat{A=1}) = \Phi[0.764 - 0.600 + 0.165 - 0.558 + 0.180 - 1.365]$$

$$P(\widehat{A=1}) = \Phi(-1.414)$$

This is not the end however.... need to find this value in the normal distribution tables:

-1.41 in the normal distribution table gives: 0.07927

- Statistical tables are available on BB

Thus there is a 7% chance that a person with these characteristics will have an affair

How will the probability of an affair change if she goes from being very happy (ratemarr=5) to very unhappy (ratemarr=1) in the marriage (all else constant)?

$$\Delta P(\widehat{A=1}) = \text{Final}P(\widehat{A=1}) - \text{Initial}P(\widehat{A=1})$$

$$= \phi[P(\widehat{A=1}) = \Phi(\hat{\beta}_0 + \hat{\beta}_2(25) + \hat{\beta}_3(3) + \hat{\beta}_5(3) + \hat{\beta}_6(12) + \hat{\beta}_7(1))] - (\phi[P(\widehat{A=1}) = \Phi(\hat{\beta}_0 + \hat{\beta}_2(25) + \hat{\beta}_3(3) + \hat{\beta}_5(3) + \hat{\beta}_6(12) + \hat{\beta}_7(5))])$$

$$\Delta P(\widehat{A=1}) = \hat{\beta}_7(1) - \hat{\beta}_7(5)$$

$$\Delta P(\widehat{A=1}) = \Phi(-0.273) - \Phi(-1.365)$$

$$\Delta P(\widehat{A=1}) = (0.39358) - (0.08534) = 0.3936 - 0.0869 = 0.3067$$

The probability of an affair will increase by 31 percentage points

Why couldn't we simply calculate the estimated effect as:

$$\Phi[\beta_{HappinessinMarriage} \times (4)]?$$

Because the estimated effect of Happiness in Marriage on the probability of having an affair depends on the initial values of all the regressors

- Probit is a non-linear regression model

How shall we interpret $\beta_{HappinessinMarriage}$?

It is the change in the z (where the estimated probability of having at least one affair is $\Phi(z)$) corresponding to a one-unit change in the rating of how happy the person is in the marriage, holding everything else constant.

Predicted Probabilities from a Logit

Logistic regression

Number of obs = 601

Wald chi2(7) = 53.00

Prob > chi2 = 0.0000

Log pseudolikelihood = -304.84818

Pseudo R2 = 0.0973

affair	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
male	.3164296	.2290344	1.38	0.167	-.1324697	.7653288
age	-.0437373	.0190217	-2.30	0.021	-.0810191	-.0064555
yrsmarr	.0952295	.0323227	2.95	0.003	.0318782	.1585808
kids	.3791557	.2931454	1.29	0.196	-.1953987	.95371
relig	-.3254421	.0924565	-3.52	0.000	-.5066535	-.1442307
educ	.0305846	.0451685	0.68	0.498	-.0579441	.1191132
ratemarr	-.4700074	.0915244	-5.14	0.000	-.649392	-.2906228
_cons	1.336704	.9258706	1.44	0.149	-.4779688	3.151377

What is the expected probability of having an affair for a 25 year-old woman, high school graduate (i.e., 12 years of education), without kids, who is married for 3 years, estimates her religiousness level as average (=3) and is very happy (=5) in her marriage?

$$P(\widehat{Affair} = 1) = \frac{1}{1 + \exp^{-z}}$$

where:

$$z = \hat{\beta}_0 + \hat{\beta}_1 Male + \hat{\beta}_2 Age + \hat{\beta}_3 YearsMarried + \hat{\beta}_4 WithKids + \hat{\beta}_5 Religiosity + \hat{\beta}_6 YearsEducation + \hat{\beta}_7 HappinessinMarriage$$

$$z = (\hat{\beta}_0 + \hat{\beta}_2(25) + \hat{\beta}_3(3) + \hat{\beta}_5(3) + \hat{\beta}_6(12) + \hat{\beta}_7(5))$$

$$z = (1.3367 + (-0.043 \times 25) + (0.095 \times 3) + (-0.325 \times 3) + (0.031 \times 12) + (-0.470 \times 5))$$

$$z = 1.3367 - 1.075 + 0.285 - 0.975 + 0.372 - 2.35 \quad z = -2.41$$

Predicted Probabilities from a Logit

Again, this is not the end....

$$P(\widehat{Affair} = 1) = \frac{1}{1 + \exp^{-z}} = \frac{1}{1 + \exp^{-(-2.41)}}$$

$$P(\widehat{Affair} = 1) = \frac{1}{1 + 11.13} = 0.08$$

Thus there is a 8% chance that a person with these characteristics will have an affair

For a change in ratemarr of 5 to 1...

$$z = (\hat{\beta}_0 + \hat{\beta}_2(25) + \hat{\beta}_3(3) + \hat{\beta}_5(3) + \hat{\beta}_6(12) + \hat{\beta}_7(1))$$

$$z = (1.3367 + (-0.044 \times 25) + (0.095 \times 3) + (-0.325 \times 3) + (0.03 \times 12) + (-0.47 \times 1))$$

$$z = -0.550$$

$$\text{Difference} = Z(-0.550) - Z(-2.41)$$

$$\text{Difference} = \frac{1}{1 + \exp(-0.550)} - 0.08 = \frac{1}{1 + 1.73} - 0.08$$

$$\text{Difference} = 0.36 - 0.08 = 0.28$$

The person would be 28% more likely to have an affair if their happiness in marriage went from 5 to 1

Again changes in probabilities cannot be calculated as in the LPM

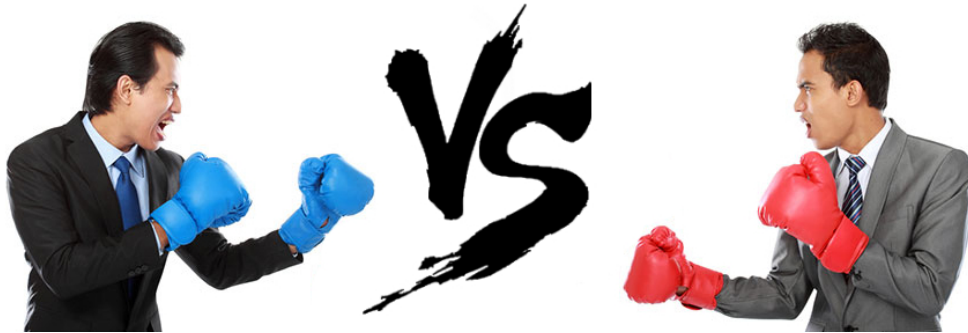
- Need to follow similar steps to Probit

How shall we interpret $\beta_{\text{Happiness in Marriage}}$?

Two models give similar but slightly different results.....

How do we choose between them?!

Probit vs Logit vs LPM



Probit vs Logit

How do the coefficients calculated by probit compare to those calculated by logit?

- For example, look at the coefficients on **ratemarr**

The estimated coefficients are very different:

- β ratemarr estimated by probit: **-0.27**
- β ratemarr estimated by logit: **-0.47**

Why is that?

They mean very different things (See interpretations)

How do the predicted values calculated by probit compare to those calculated by logit

- For example, look at our example of the predicted change in the probability of having an affair for the woman who went from being very happy to very unhappy in her marriage)?

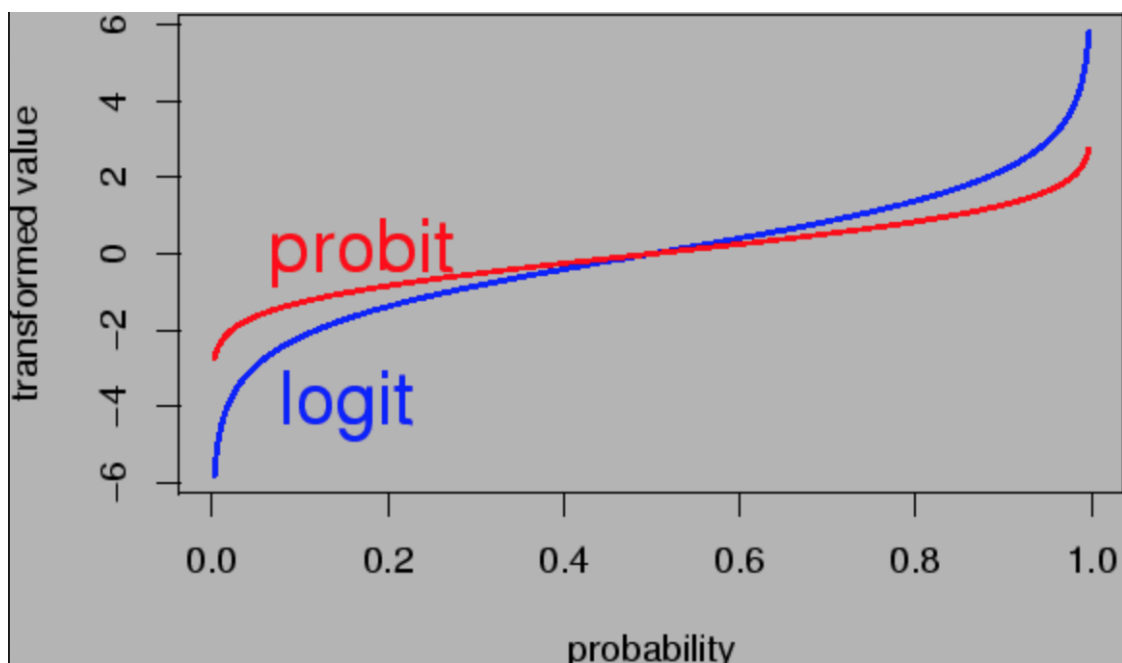
The estimated change in the predicted values are very similar

The estimated change in $\hat{Y}$ calculated:

- by probit was of 31 percentage points
- by logit was of 28 percentage points

Why is that?

The shape of the cumulative probability distribution function of the standard normal and the shape of the cumulative probability distribution function of the standard logistic are very similar



STATA Test

quietly regress affair male age yrsmarr kids relig educ ratemarr, r
predict ols

quietly probit affair male age yrsmarr kids relig educ ratemarr, r
predict probit

quietly logit affair male age yrsmarr kids relig educ ratemarr, r
predict logit

sum affair ols probit logit

Variable	Obs	Mean	Std. Dev.	Min	Max
affair	601	.249584	.4331328	0	1
ols	601	.249584	.141164	-.1128995	.6256906
probit	601	.249917	.1412915	.0242042	.6901009
logit	601	.249584	.1441054	.0320915	.7124752

STATA Test

estat gof, group(10) table

Probit model for affair, goodness-of-fit test

(Table collapsed on quantiles of estimated probabilities)

Group	Prob	Obs_1	Exp_1	Obs_0	Exp_0	Total
1	0.0902	5	4.2	56	56.8	61
2	0.1210	5	6.4	55	53.6	60
3	0.1583	12	8.6	49	52.4	61
4	0.1823	11	10.3	49	49.7	60
5	0.2190	10	12.0	50	48.0	60
6	0.2594	13	14.0	46	45.0	59
7	0.3165	13	17.1	47	42.9	60
8	0.3773	21	20.6	39	39.4	60
9	0.4576	24	24.9	36	35.1	60
10	0.6901	36	32.2	24	27.8	60

STATA Test

estat gof, group(10) table

Logistic model for affair, goodness-of-fit test

(Table collapsed on quantiles of estimated probabilities)

Group	Prob	Obs_1	Exp_1	Obs_0	Exp_0	Total
1	0.0928	5	4.5	57	57.5	62
2	0.1198	5	6.4	54	52.6	59
3	0.1545	11	8.3	49	51.7	60
4	0.1773	13	10.1	47	49.9	60
5	0.2135	9	11.7	51	48.3	60
6	0.2529	13	13.9	47	46.1	60
7	0.3113	12	16.7	48	43.3	60
8	0.3761	24	20.4	36	39.6	60
9	0.4656	22	25.0	38	35.0	60
10	0.7125	36	32.9	24	27.1	60

Summary

What are the advantages and disadvantages of the linear probability model?

- **Advantage:** The coefficients are easy to interpret
- **Disadvantage:** Because it assumes linearity, it predicts probabilities outside [0,1]

What about probit and logit?

- **Advantage:** The predicted values (predicted probabilities) are bounded between 0 and 1
- **Disadvantage:** The coefficients are more difficult to interpret
 - For interpretations, should always calculate differences in predicted probabilities: $Y^{final} - Y^{initial}$

Learning Outcomes

- ① Be able to review the limitations of simple liner regression analysis and the situations where it is not the most appropriate method to use;
 - Do you know why a LPM may not work?
- ② Have developed an awareness of which models are appropriate to use when an OLS regression may not be robust, and be able to apply these models correctly;
 - Do you know why a Probit or logit may work better?
- ③ Command an experienced understanding of the empirical approach to economics, and be able to undertake a robust piece of econometric analysis;
 - Estimating these models (chance in the coursework and labs)
 - Do you know what the results mean?
- ④ Be able to explain the covered econometric concepts clearly in concise non-technical language

READING

Check the notes on non-continuous data and the simulation work on the LPM

Your questions please....

Glossary

Discrete Choice Model - Models that describe the decision maker's choices among a finite number of alternative options. Binary choice models are a form of discrete choice model where there are only two options.

Linear Probability Model (LPM) - A simple OLS regression using the binary outcome measure as the dependent variable.

Predicted Probabilities - The probability of an individual being in the outcome group equal to one given a set of independent variables.