

ECO 372: Introduction to Econometrics

Spring 2025

Discrete and Limited Dependent Variable Models

Sakib Bin Amin, Ph.D.

Associate Professor in Economics

Director, Accreditation Project Team (APT)



Outline

Our objectives for this lecture will be to learn:

- What We have Learnt So Far?
- Binary Dependent Variables
- Issues with LPM
- LMP Example
- Data Visualisation
- Estimation Results
- Fitted Plot of LPM Model
- STATA Commands
- The Emergence of Probit and Logit Models
- General Framework of Probit and Logit Models
- Statistical Inference of the Parameters
- Empirical Example from Bangladesh
- Probit and Logit with STATA
- STATA Commands

A Quick Recap

➤ So far, we know how to deal with linear estimation models. That is:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i$$

➤ Sometimes, we have to transform or add variables to make the equation linear.

- ❑ Taking logs of Y and/or the X's

- ❑ Adding squared terms

- ❑ Using dummies as regressors

- ❑ Adding interactions

- ❑ Using quantile/decile

➤ Then, we can run the estimation, check model consistency, interpret results, etc.

➤ One thing is common in all the above situations: **the outcome variable Y_i was always continuous.**

Binary Dependent Variables

- Many dependent variables of interest in economics and other social sciences can only take two values.
- The two possible outcomes are usually denoted by 0 and 1.
- As we know, such variables are called dummy variables or dichotomous variables.
- Some Examples:
 - ❑ The labor market status of a person. The variable takes the value 1 if a person is employed and 0 if he is unemployed.
 - ❑ Access to Solar Home Systems (SHS) in off-grid households. Again this variable takes 1 if the off-grid household has a solar home system device and 0 otherwise.
 - ❑ Cooking choice preference of women in households. It takes 1 if the cooking technology is clean/modern (stove, induction, LPG, clean cooking stove, etc.) or 0 if traditional cooking technology (firewood, inefficient fire stove, cow dung, etc.).

Binary Dependent Variables

- What would be the estimation strategy for unknown parameters if the dependent variable is binary?
- We can consider the Linear Probability Model (LPM).
- The LPM is an altered version of OLS.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i$$

- When Y_i is binary/dummy then the expected value can be written as follows :

$$E(Y_i|X_{ij}) = [0 \times \text{Pr}(Y_i = 0)] + [1 \times \text{Pr}(Y_i = 1)] = \text{Pr}(Y_i = 1)$$

- Extending this to the binary dependent variable, we can rewrite the LPM as:

$$\text{Pr}(Y_i = 1|X_{1i}, \dots, X_{ki}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}$$

- The regression coefficient β_1 is the **probability** that $Y = 1$ outcome will be achieved when X_1 changes by 1 unit, holding all other things constant.

Issues with LPM

- Although the LPM framework gives a very straightforward way of estimating models with binary-dependent variables, it is not an adequate statistical model.
- When empirically tested, it has been found that the expected value can lie outside $[0,1]$ and does not represent a probability.
- In addition, estimated variance is always heteroscedastic.
- Given the issues, a new class of estimators was developed.
- These estimators are popularly known as Discrete Estimators or Binary Estimators.

LPM Example

➤ Let us take an example. We will attempt to find out what factors influence women labour force participation.

$$FLP_i = \beta_0 + \beta_1 Fam_Income_i + \beta_2 Edu_i + \beta_3 Exp_i + \beta_4 Exp_i^2 + \beta_5 Age_i + \beta_6 kids_lt6_i + \beta_7 kidsage6_i + u_i$$

FLP_i = Female labour force participation. 1= participate, 0= otherwise

Fam_Income_i = Family income

Edu_i = Education level

EXP_i = Experience in labour market

Exp_i^2 = Squared of experience

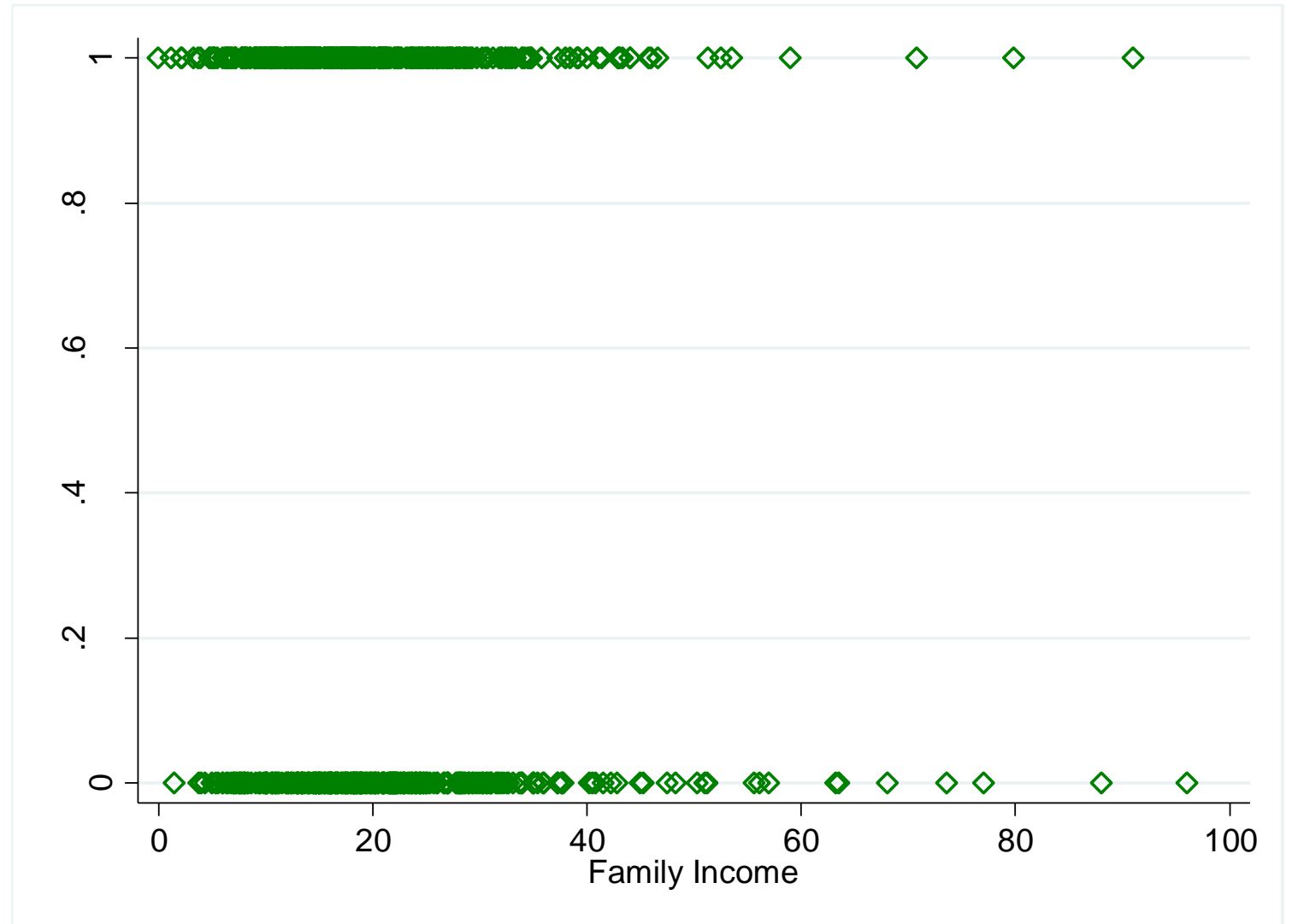
Age_i = Age of the respondent

$kids_lt6_i$ = Has kids with age less than 6

$kidsage6_i$ = Has kids with age 6-18

Data Visualisation

- Scatter plot of FLP and Family income.
- The only value FLP takes is 0 or 1.

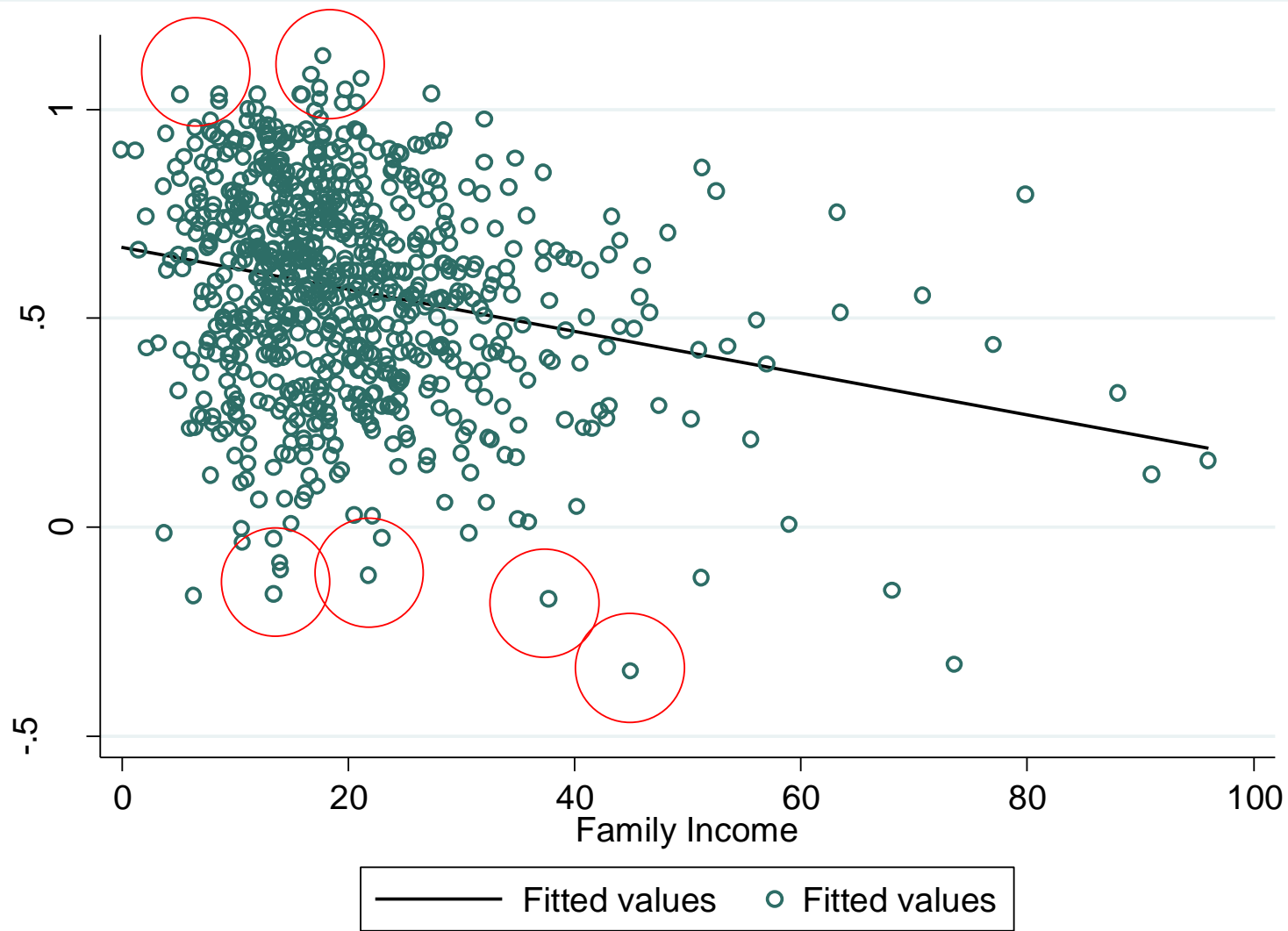


Estimation Results

Linear regression	Number of obs	=	753
	F(7, 745)	=	62.48
	Prob > F	=	0.0000
	R-squared	=	0.2642
	Root MSE	=	.42713

inlf	Robust					
	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
nwifeinc	-.0034052	.0015249	-2.23	0.026	-.0063988	-.0004115
educ	.0379953	.007266	5.23	0.000	.023731	.0522596
exper	.0394924	.00581	6.80	0.000	.0280864	.0508983
expersq	-.0005963	.00019	-3.14	0.002	-.0009693	-.0002233
age	-.0160908	.002399	-6.71	0.000	-.0208004	-.0113812
kidslt6	-.2618105	.0317832	-8.24	0.000	-.3242058	-.1994152
kidsge6	.0130122	.0135329	0.96	0.337	-.013555	.0395795
_cons	.5855192	.1522599	3.85	0.000	.2866098	.8844287

Fitted Plot



- Focus on the circles.
- Do the values fall between 0 and 1?
- Not At All!
- That means LPM fails to estimate the expected values.

STATA Commands

```
//Linear Probability Model: (page.455 of Woolridge)
```

```
//open MROZ.dta
```

```
describe
```

```
sum
```

```
histogram inlf
```

```
histogram nwifeinc
```

```
//linear probability model (simple ols)
```

```
tab inlf
```

```
reg inlf nwifeinc educ exper expersq age kidslt6 kidsge6
```

```
predict flp_hat if e(sample)
```

```
sum flp_hat
```

```
twoway lfit flp_hat nwifeinc || scatter flp_hat nwifeinc
```

The Emergence of Probit and Logit Models

- Recall the LPM model and its problems. We have seen that the LPM model does not predict the binary-dependent variable well and does not represent probability.
- Additionally, estimated variance is always heteroscedastic.
- Therefore, we need different types of estimators to address these issues.
- The two most widely used models in this case are the **Probit** and **Logit** models.
- Based on **Probit** and **Logit** models, further models have been developed to provide robust estimates and predictions when a dependent variable takes values beyond binary. For instance, categorical, un-categorical, ordered, etc.

General Framework of Probit and Logit Models

- We will use the concept of the latent variable approach.
- By definition, a latent variable means a variable that is unobserved in the realm of a modelling framework (we refer to it as Y_i^*).
- The condition Probit and Logit models impose is that the probability of the latent variable having a value greater than zero will ensure the probability of the intended outcome having a value of 1, given regressors.
- Given this condition, transforming the right-hand side of the specification can predict the dependent variable properly.
- Suppose, we have the following model to express the latent variable:

$$Y_i^* = \beta_0 + \beta_1 X_i + u_i ; \quad E(u_i | X_i) = 0$$

- The latent variable can be interpreted as the utility difference between choosing $Y_i = 1$ and 0.

General Framework of Probit and Logit Models

- A researcher observes only Y_i , which is a binary variable.
- Following the condition, the researcher chooses $Y_i = 1$ if the latent variable is positive and 0 otherwise.

$$Y_i = \begin{cases} 1 & \text{if } Y_i^* > 0 \\ 0 & \text{if } Y_i^* < 0 \end{cases}$$

- Furthermore, assume that the individual observations are i.i.d, that the explanatory variables are exogenous, and that the error term is normally distributed and homoscedastic

$$u_i | X_i \sim N(0, \sigma^2)$$

- So:

$$\Pr(Y_i = 1 | X_i) = \Pr(Y_i^* > 0 | X_i) = \Phi(\beta_0 + \beta_1 X_i)$$

- Φ is the transformation function that keeps the distribution of the parameters between 0 and 1.

General Framework of Probit and Logit Models

- The model is known as the Probit Model when Φ follows the cumulative distribution function of the standard normal distribution. The integral of PDF is CDF.

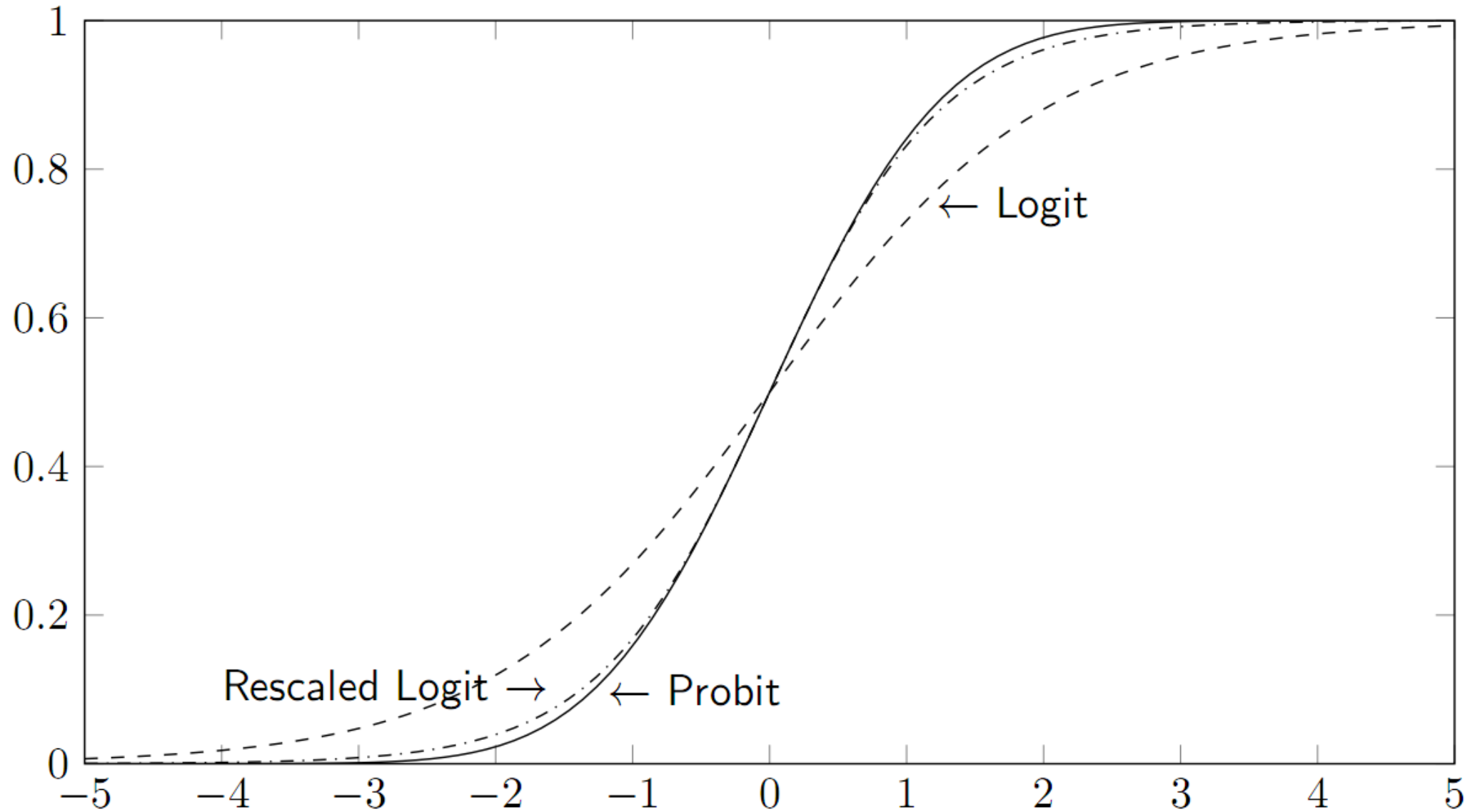
$$CDF = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

- On the other hand, when Φ follows the Standard Logistic distribution function, the model is known as the Logit Model.

$$LDF = \frac{1}{1 + e^{-x}}$$

- The Probit and Logit models are estimated by Maximum Likelihood (ML) method, which maximises the log-likelihood function (i.e., value function) to get the unknown parameters.

General Framework of Probit and Logit Models



Statistical Inference of the Parameters

- Unlike the linear models, the parameters cannot directly be interpreted as marginal effects (i.e., how a change in predictor leads to a change in the expected value of the dependent variable) on the dependent variable.
- Because both Probit and Logit models are non-linear due to the addition of distinct distribution functions that help to transform the data.
- The marginal effects for these models are as follows:

$$\frac{\partial E(Y_i|X_i)}{\partial X_{ki}} = \frac{\partial \Pr(Y_i = 1|X_i)}{\partial X_{ki}} = \phi(\beta_k)$$

$$\frac{\partial E(Y_i|X_i)}{\partial X_{ki}} = \frac{\partial \Pr(Y_i = 1|X_i)}{\partial X_{ki}} = \psi(\beta_k)$$

- ϕ and ψ are partial differentiations of CDF and LDF, respectively.

Goodness of Fit

- The goodness of fit is primarily judged by Pseudo-R². Recall that Pseudo-R² is used when components of the R² are not available.
- Since the Probit and Logit model is estimated with the ML method, derivation of SSR and TSR is not possible.
- Pseudo-R² in this regard is as follows:

$$1 - \frac{\ln \hat{L}(M_{unrestricted})}{\ln \hat{L}(M_{restricted})}$$

- $\ln \hat{L}(M_{unrestricted})$ is the value of the objective function of the intended model and $\ln \hat{L}(M_{restricted})$ is the value of the objective function of the restricted model.
- Similar to the R² of the linear regression model, it holds that the higher the value of R², the better the model. Adjusted Pseudo-R² can also be derived.

Goodness of Fit

- Apart from Pseudo-R², the Likelihood Ratio test is another way to check the goodness of fit.
- The likelihood ratio test is the ratio of the log-likelihood functions of the unrestricted and restricted model. It follows a Chi-square distribution with DF equal to the number of restrictions. The higher the value, the lower the model's predictability

$$LR = -2\ln[\ln\hat{L}(M_{restricted}) - \ln\hat{L}(M_{unrestricted})]$$

$$0 \leq LR \leq 1$$

- Another test is the classification test. It tests how much the model has predicted the data. The higher the degree of correct or positive prediction across observations [i.e. $\Pr(Y_i = 1|X_i)$], the better the model.
- Finally, the Receiver Operating Characteristic (ROC) curve is used. It plots Sensitivity versus (1-specificity). Sensitivity is the fraction of observed positive outcomes. (1-specificity) is the fraction of observed negative outcomes.

□ The greater the predictive power, the more bowed the curve will be.

Which One to Choose? Probit or Logit?

- When it comes to choosing one from Probit and Logit, according to many scholars, it is largely a matter of taste.
- The Logit and Probit distributions are very close to each other and differ appreciably only in the case where the distribution is very spread.
- Many researchers used to choose the Logit over the Probit because of its comparative mathematical simplicity. However, given the advancements in computational software, It takes the same effort to run Probit and Logit models these days.
- Some of the investigations reveal that estimated parameters with the Logit Model are approximately equal to 1.6 times that parameter obtained by the Probit model.
- When the independent variable is found to be an extreme independent variable (i.e., when most of the observations are skewed toward the extreme side), the Logit Model seems to work better. But then again, recent computational updates make Probit models better in this scenario as well.

Empirical Example from Bangladesh

Table 5 Results of participation equation (without and with wage)

Column 1	Column 2	Column 3
Variable	Participation (without wage)	Participation (with wage)
Imputed wage		0.084 ^a (0.004)
Age	0.016 ^a (0.001)	0.036 ^a (0.001)
Age ²	-0.0003 ^a (0.000)	-0.0006 ^a (0.000)
Primary and secondary passed	0.057 ^a (0.006)	0.122 ^a (0.007)
SSC and HSC passed	0.124 ^a (0.013)	0.163 ^a (0.013)
University passed	0.577 ^a (0.018)	0.542 ^a (0.021)
Marital status	-0.130 ^a (0.010)	-0.130 ^a (0.010)
Child	0.004 ^c (0.002)	0.005 ^c (0.002)
Child under 6 years	-0.036 ^a (0.003)	-0.036 ^a (0.003)
Net family income (natural log)	-0.069 ^a (0.001)	-0.069 ^a (0.001)
Household land	7.26E-05 ^a (0.000)	7.26E-05 ^a (0.000)
Urban	0.119 ^a (0.006)	
Living with in-laws	-0.036 ^a (0.007)	-0.036 ^a (0.007)
Head primary and secondary passed	0.001 (0.001)	0.001 (0.006)
Head SSC and HSC passed	0.005 (0.010)	0.005 (0.010)
Head university passed	-0.037 ^b (0.015)	-0.037 ^b (0.015)
Head employed in agriculture	0.009 (0.005)	0.009 (0.005)
Head self employed	0.476 ^a (0.004)	0.476 ^a (0.005)
Chittagong division	0.051 ^a (0.009)	0.051 ^a (0.009)
Dhaka division	0.087 ^a (0.009)	0.087 ^a (0.009)
Khulna division	0.038 ^a (0.010)	0.038 ^a (0.010)
Rajshahi division	0.042 ^a (0.009)	0.042 ^a (0.009)
Sylhet division	0.010 (0.011)	0.010 ^a (0.011)
Constant	1.525	-0.729
Pseudo R ²	0.293	0.293
Chi ²	0.000	0.000
N	42,646	42,646

- The imputed wage has a significant impact on female labour force participation in Bangladesh. An increase in imputed wage by 1 unit increases the likelihood of joining the labour force by 8.4% for females.
- Female university education pushes up the probability of joining the labour force by 57.7%-54.2% in Bangladesh.
- Having 1 more kid with an age less than 6 reduces the propensity of labour force participation by 3.6% for females in Bangladesh.
- Living with in-laws actually diminishes the probability of joining the labour force by 3.6% in Bangladesh for females.

Source: Mahmud, S., Bidisha, S.H. (2018). Female labor market participation in Bangladesh: Structural changes and determinants of labor supply. In: Raihan, S. (eds) *Structural Change and Dynamics of Labor Markets in Bangladesh. South Asia Economic and Policy Studies*. Springer, Singapore.

Probit & Logit with STATA

➤ Let us take the previous model:

$$FLP_i = \beta_0 + \beta_1 Fam_Income_i + \beta_2 Edu_i + \beta_3 Exp_i + \beta_4 Exp_i^2 + \beta_5 Age_i + \beta_6 kids_lt6_i + \beta_7 kidsage6_i + u_i$$

FLP_i = Female labour force participation. 1= participate, 0= otherwise

Fam_Income_i = Family income

Edu_i = Education level

EXP_i = Experience in labour market

Exp_i^2 = Squared of experience

Age_i = Age of the respondent

$kids_lt6_i$ = Has kids with age less than 6

$kidsage6_i$ = Has kids with age 6-18

Our objective is to check what determines female labour force participation.

Probit and Logit with STATA

Probit regression

Number of obs = 753

LR chi2(7) = 227.14

Prob > chi2 = 0.0000


Log likelihood = -401.30219

Pseudo R2 = 0.2206


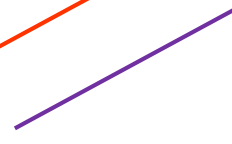



inlf	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
nwifeinc	-.0120237	.0048398	-2.48	0.013	-.0215096	-.0025378
educ	.1309047	.0252542	5.18	0.000	.0814074	.180402
exper	.1233476	.0187164	6.59	0.000	.0866641	.1600311
expersq	-.0018871	.00006	-3.15	0.002	-.003063	-.0007111
age	-.0528527	.0084772	-6.23	0.000	-.0694678	-.0362376
kidslt6	-.8683285	.1185223	-7.33	0.000	-1.100628	-.636029
kidsge6	.036005	.0434768	0.83	0.408	-.049208	.1212179
_cons	.2700768	.508593	0.53	0.595	-.7267473	1.266901

Probit and Logit with STATA

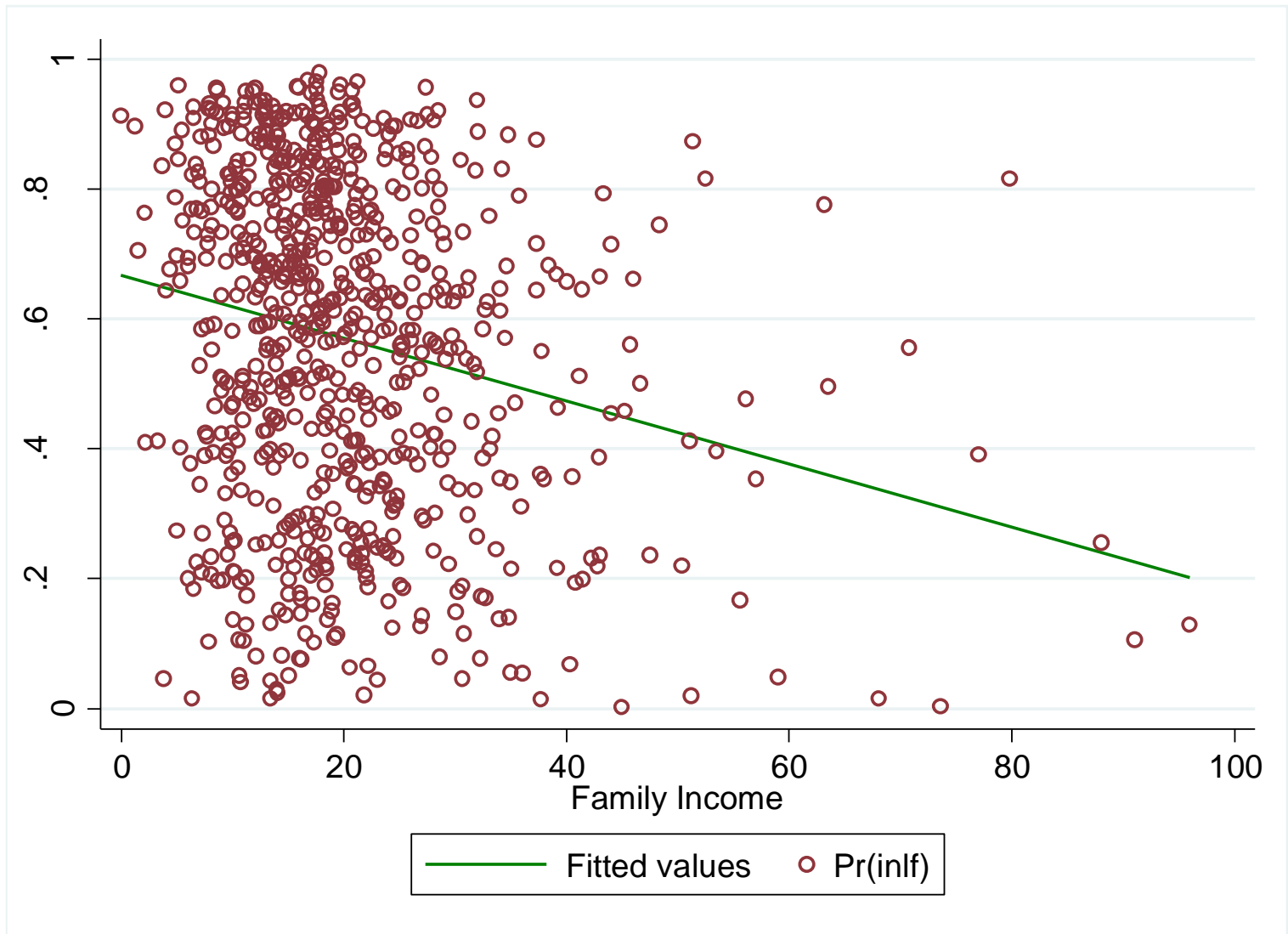


	Delta-method					
	dy/dx	Std. Err.	z	P> z	[95% Conf. Interval]	
nwifeinc	-.0046962	.0018903	-2.48	0.013	-.0084012	-.0009913
educ	.0511287	.0098592	5.19	0.000	.0318051	.0704523
exper	.0481771	.0073278	6.57	0.000	.0338149	.0625392
expersq	-.0007371	.0002347	-3.14	0.002	-.001197	-.0002771
age	-.0206432	.0033079	-6.24	0.000	-.0271265	-.0141598
kidslt6	-.3391514	.0463581	-7.32	0.000	-.4300117	-.2482911
kidsge6	.0140628	.0169852	0.83	0.408	-.0192275	.0473531



- Increase in family income by 1 unit reduces the likelihood of female labour force participation by 0.47%.
- Increase in education by 1 year leads to 5.11% higher probability to join labour force for females.
- Having 1 more kid with age less than 6 reduces the propensity of labour force participation by 33.92% for females.

Probit and Logit with STATA



➤ Predicted probabilities are within the range of 0 and 1.

Probit and Logit with STATA

Classified	True		Total
	D	~D	
+	348	120	468
-	80	205	285
Total	428	325	753

Classified + if predicted $\Pr(D) \geq .5$
 True D defined as `inlf != 0`

Sensitivity	$\Pr(+ D)$	81.31%
Specificity	$\Pr(- \sim D)$	63.08%
Positive predictive value	$\Pr(D +)$	74.36%
Negative predictive value	$\Pr(\sim D -)$	71.93%

False + rate for true ~D	$\Pr(+ \sim D)$	36.92%
False - rate for true D	$\Pr(- D)$	18.69%
False + rate for classified +	$\Pr(\sim D +)$	25.64%
False - rate for classified -	$\Pr(D -)$	28.07%

Correctly classified	73.44%
----------------------	--------

- Positive outcome is when $\Pr(\text{outcome}) \geq 0.5$.
- 73.44% is Correctly specified.

Probit and Logit with STATA

Likelihood-ratio test

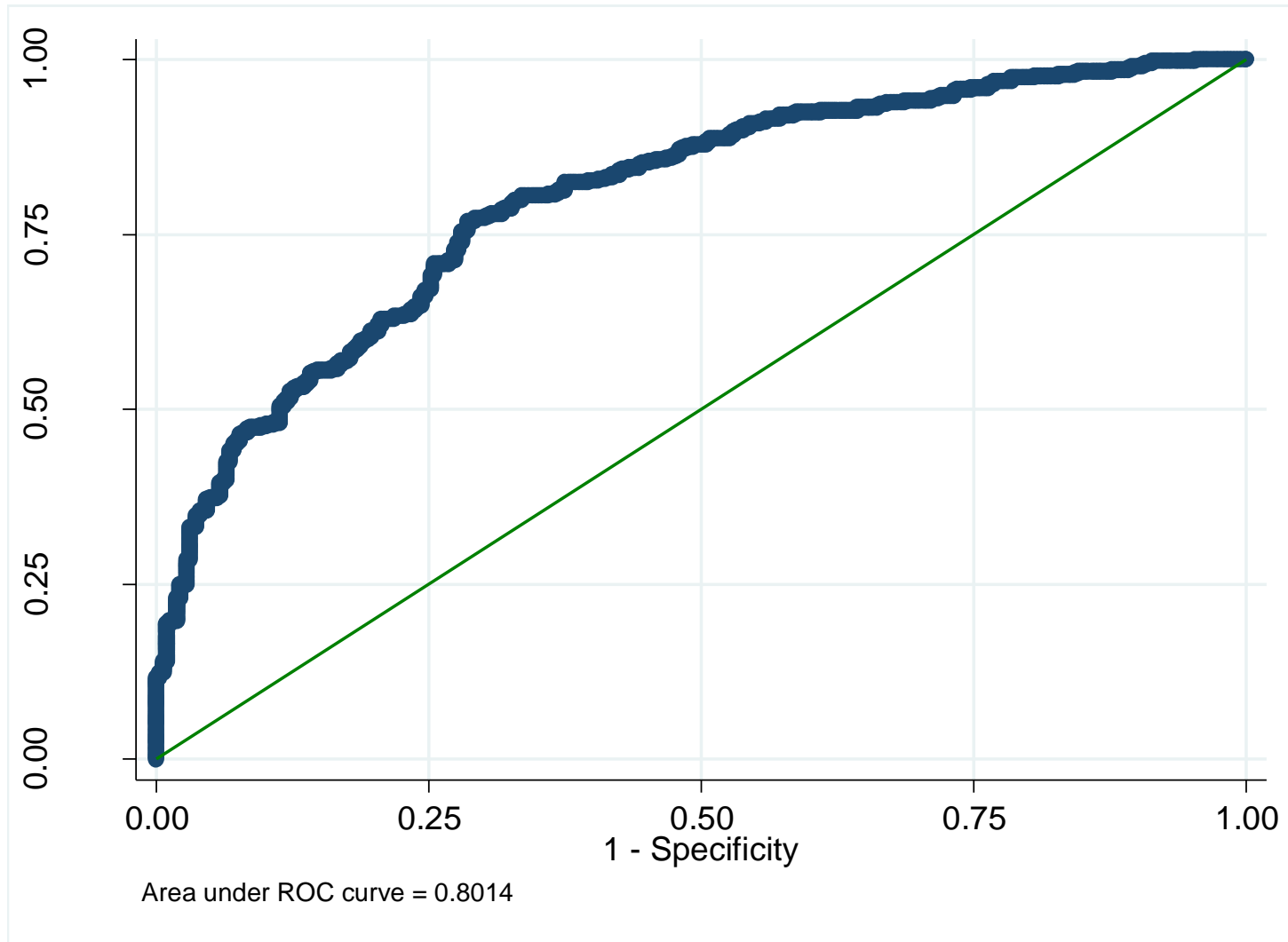
(Assumption: m1 nested in m2)

LR chi2(2) = 0.29

Prob > chi2 = 0.8647

- Model 1 is just depended variable is a function of constant.
- Model 2 is the main model showed earlier.
- Probability is 0.86, indicating we can not reject the assumption that model 1 is a part of model 2. Means that model 2 is a better model .

Probit and Logit with STATA



➤ The curve is bowed, meaning the model is well fit.

STATA Commands

//probit model

```
probit inlf nwifeinc educ exper expersq age kidslt6 kidsge6
      margins, dydx(*) atmeans
      predict flp_hat1 if e(sample)
      sum flp_hat1
      twoway lfit flp_hat1 nwifeinc || scatter flp_hat1 nwifeinc
```

//correct specified

```
estat classification
```

// roc curve

```
lroc
```

// likelihood ratio

```
quietly probit inlf
      estimates store m1_restricted
quietly probit inlf nwifeinc educ exper expersq age kidslt6 kidsge6
      estimates store m2_unrestricted
lrtest m1 m2
```

*for logit model, just substitute “probit” with “logit”

```
logit inlf nwifeinc educ exper expersq age kidslt6 kidsge6
```