Elements of Econometrics. Lecture 14. Linear Probability Model. Binary Choice Models

FCS, 2022-2023

Binary model
for pregnancy prediction

TECH

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

Kashmir Hill Former Staff

Welcome to The Not-So Private Parts where technology & privacy collide

Feb 16, 2012, 11:02am EST

This article is more than 10 years old.

Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy. Target , for example, has figured out how to data-mine its way into your womb, to figure out whether you have a baby on the way long before you need to start buying diapers.

TARGET

LINEAR PROBABILITY MODEL

The linear probability model is the linear multiple regression model

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

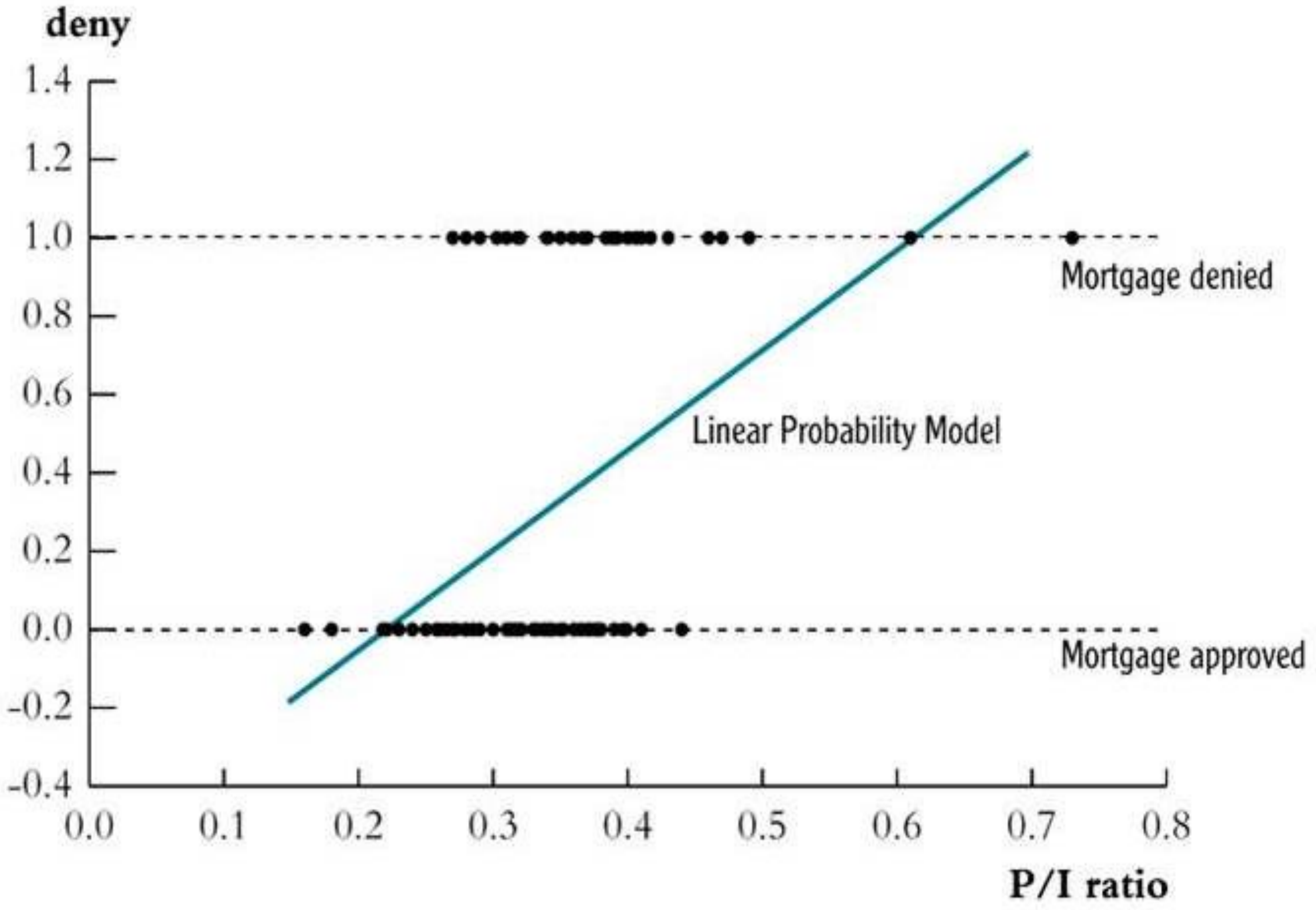
where Y_i is binary, so that

$$\Pr(Y = 1\,|X_1, X_2, \dots, X_k) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

The regression coefficient β_1 is the change in the probability that $Y = 1$ associated with a unit change in X_1, holding constant other regressors, and so forth for β_2, etc. The regression coefficients can be estimated by OLS, and the usual (heteroscedasticity-robust) OLS standard errors can be used for confidence intervals and hypotheses testing.

LINEAR PROBABILITY MODEL: Scatterplot of Mortgage Application Denial and the Payment-to-Income Ratio

Mortgage applicants with a high ratio of debt payments to income (*P/I ratio*) are more likely to have their application denied (*deny* = 1 if denied, *deny* = 0 if approved). The linear probability model uses a straight line to model the probability of denial, conditional on the *P/I ratio*.



LINEAR PROBABILITY MODEL

$$p_i = p(Y_i = 1) = \beta_1 + \beta_2 X_i$$

$$Y_i = \mathrm{E}(Y_i) + u_i$$

$$\mathrm{E}(Y_i) = 1 \times p_i + 0 \times (1 - p_i) = p_i = \beta_1 + \beta_2 X_i$$

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

$$Y_i = 1 \quad \Rightarrow \quad u_i = 1 - \beta_1 - \beta_2 X_i$$

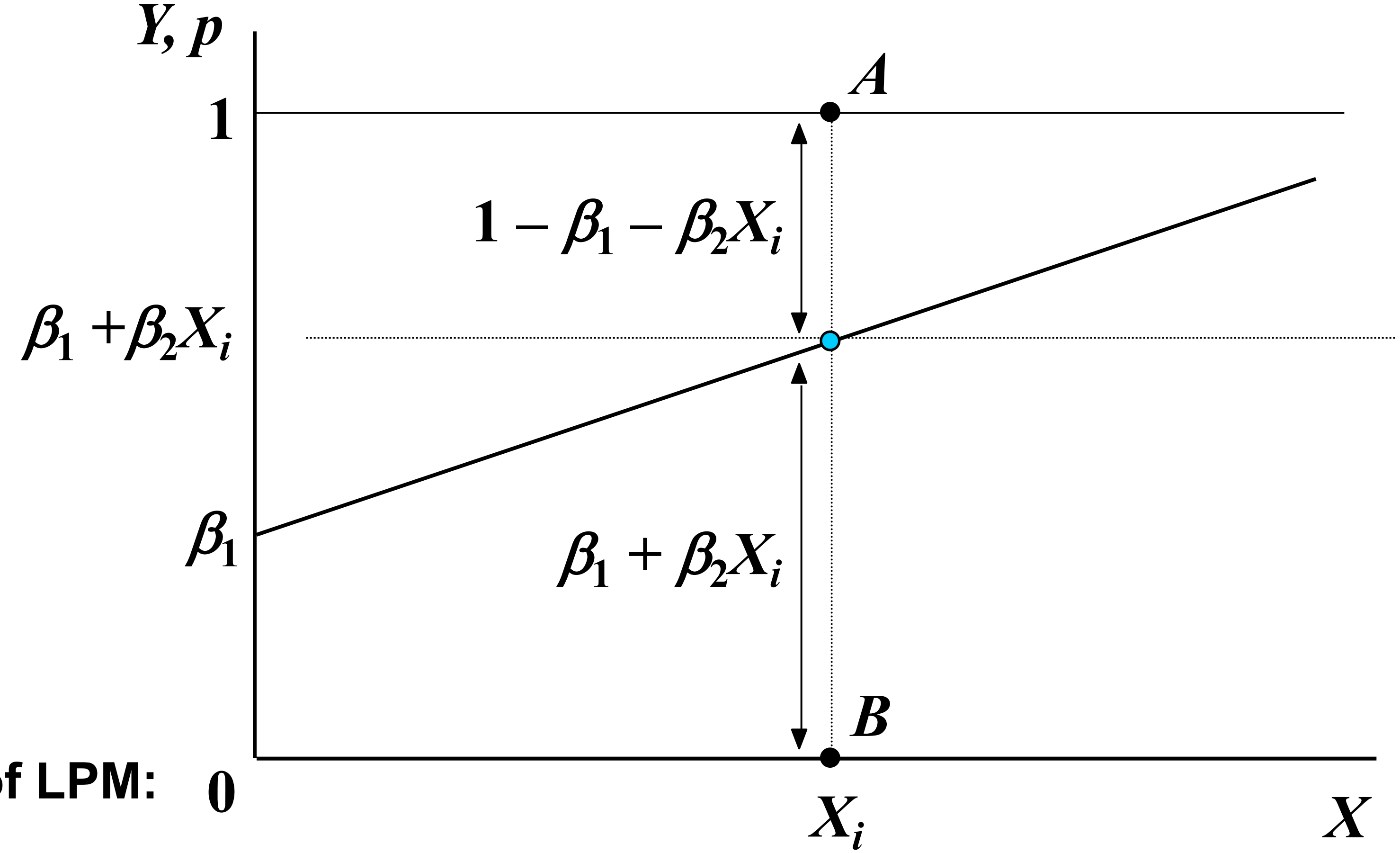
$$Y_i = 0 \quad \Rightarrow \quad u_i = -\beta_1 - \beta_2 X_i$$

The value of dependent variable Y_i in observation *i* has a nonstochastic component and a random component. The nonstochastic component depends on X_i and the parameters. The random component is the disturbance term u_i.

In observation *i*, for Y_i to be 1, u_i is $(1 - \beta_1 - \beta_2 X_i)$. For Y_i to be 0, u_i is $(- \beta_1 - \beta_2 X_i)$.

LINEAR PROBABILITY MODEL

$$p_i = p(Y_i = 1) = \beta_1 + \beta_2 X_i$$



Shortcomings of LPM:

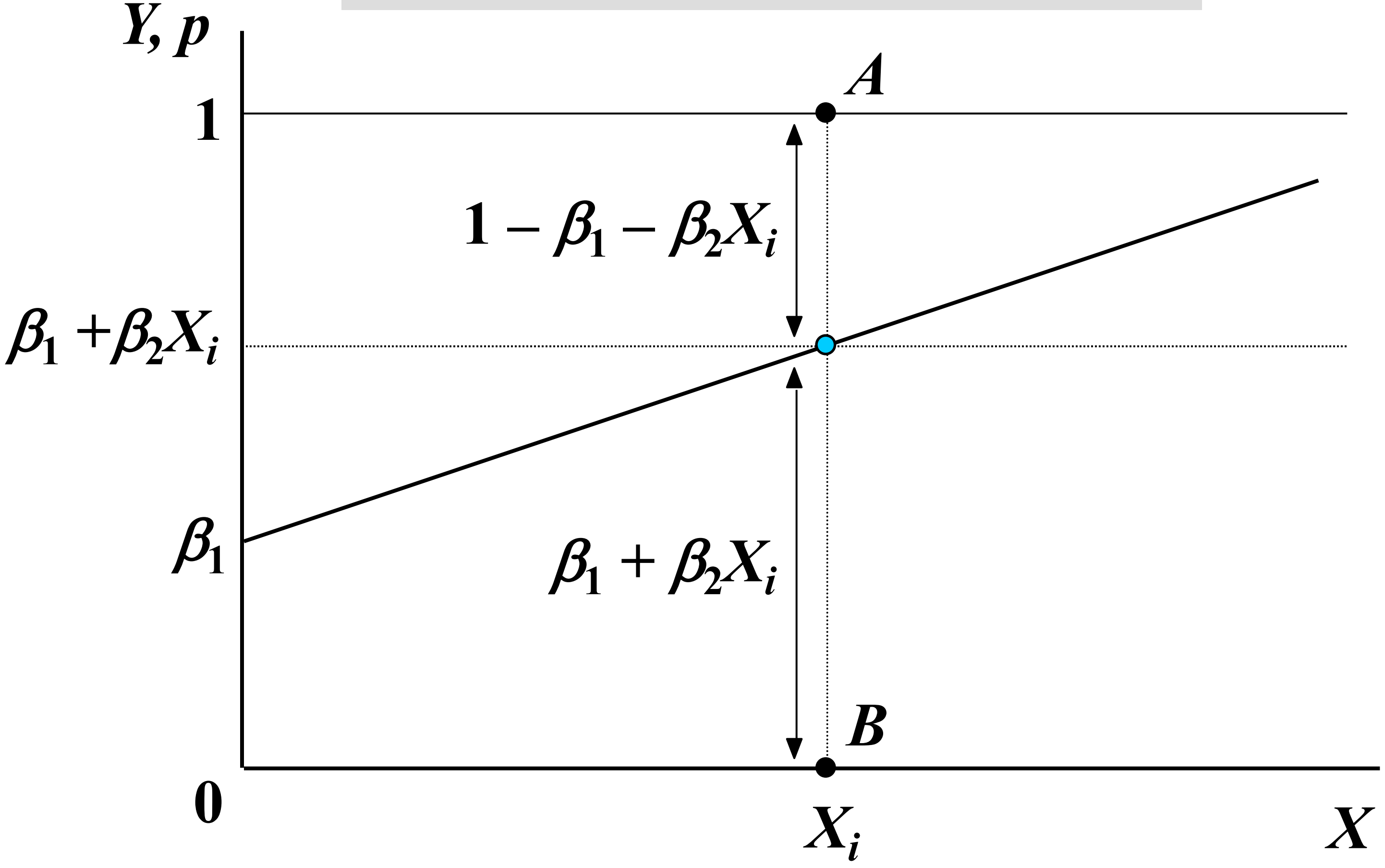
Since *u* does not have a normal distribution, the standard errors and test statistics are invalid. Its distribution is not continuous.

It may predict probabilities of more than 1 or less than 0.

Marginal effect of each factor is constant.

LINEAR PROBABILITY MODEL

$$\sigma_{u_i}^2 = (\beta_1 + \beta_2 X_i)(1 - \beta_1 - \beta_2 X_i)$$



Further, it can be shown that the population variance of the disturbance term in observation *i* is given by (β_1 + $\beta_2 X_i$)(1 – β_1 – $\beta_2 X_i$). This changes with X_i, and so the distribution is heteroscedastic. It is discrete and has only 2 possible values for each X. Test statistics are calculated wrongly, tests are invalid.

LINEAR PROBABILITY MODEL: EXAMPLE.

ICEF STUDENTS UoL First Class Honours Degrees

Depending on their Econometrics Performance

In the pre-covid year, 66 ICEF BSc graduates got the First Class Honours UoL Degrees.

Dependent Variable: First (binary).

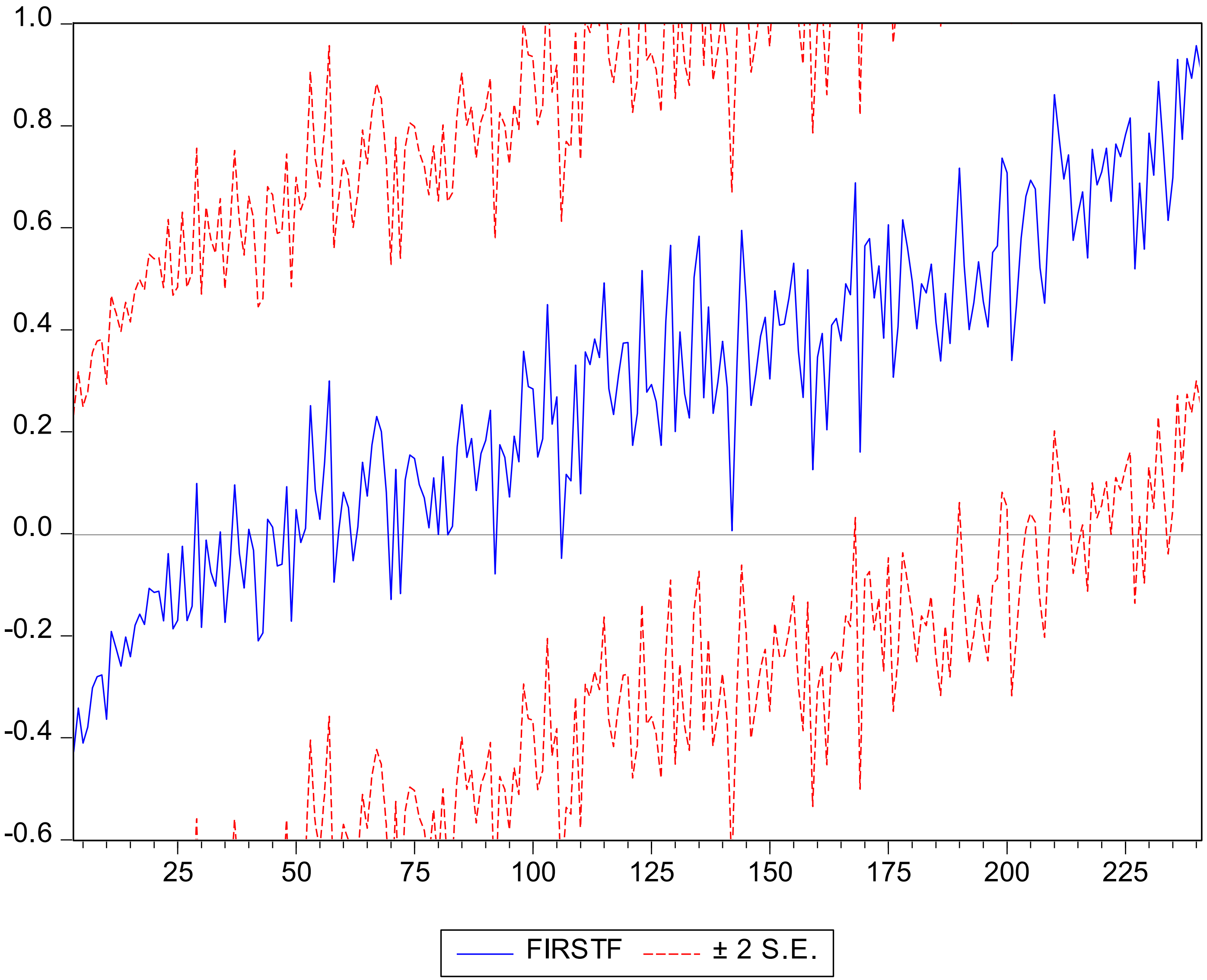
Explanatory variables: UOL – Elements of Econometrics UoL grade; SEM1 – grade (out of 100) in Econometrics for Semester 1 (proxy for regular studying).

239 observations in the sample, sorted by UOL.

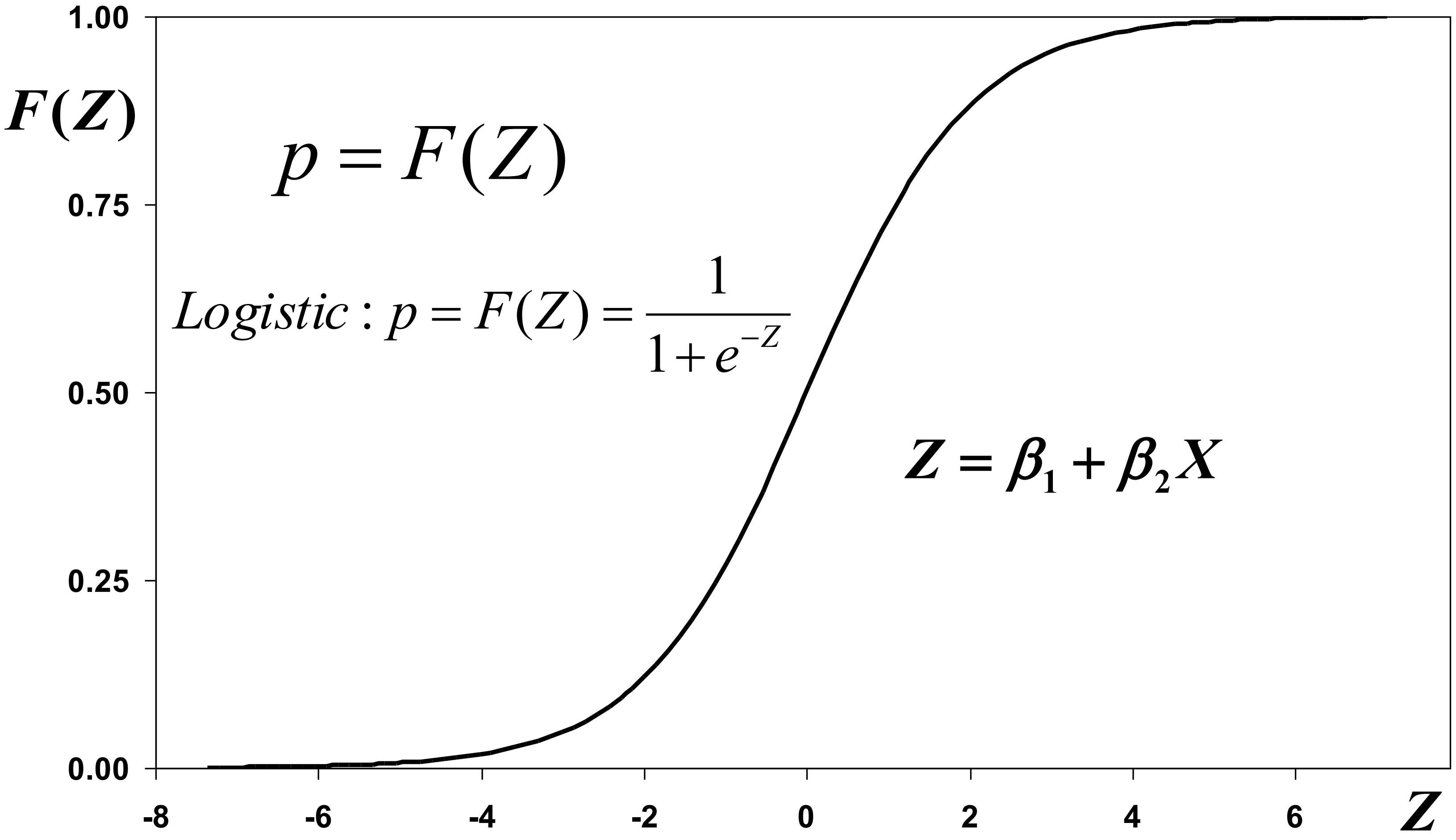
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.440911	0.059930	-7.357061	0.0000
UOL	0.007704	0.001674	4.601300	0.0000
SEM1	0.010440	0.001948	5.359645	0.0000
R-squared	0.479206	Mean dependent var		0.276151

LINEAR PROBABILITY MODEL: EXAMPLE.

ICEF STUDENTS UoL First Class Honours Degrees (pre-covid year), Depending on their Econometrics Performance: LPM Forecast



BINARY CHOICE MODELS: LOGIT ANALYSIS



To avoid the specification problem, sigmoid (*S*-shaped) function of *Z*, *F*(*Z*), where *Z* is a linear function of the explanatory variables, can be used.

Logistic function: $F(z) = \frac{1}{1+e^{-z}}$

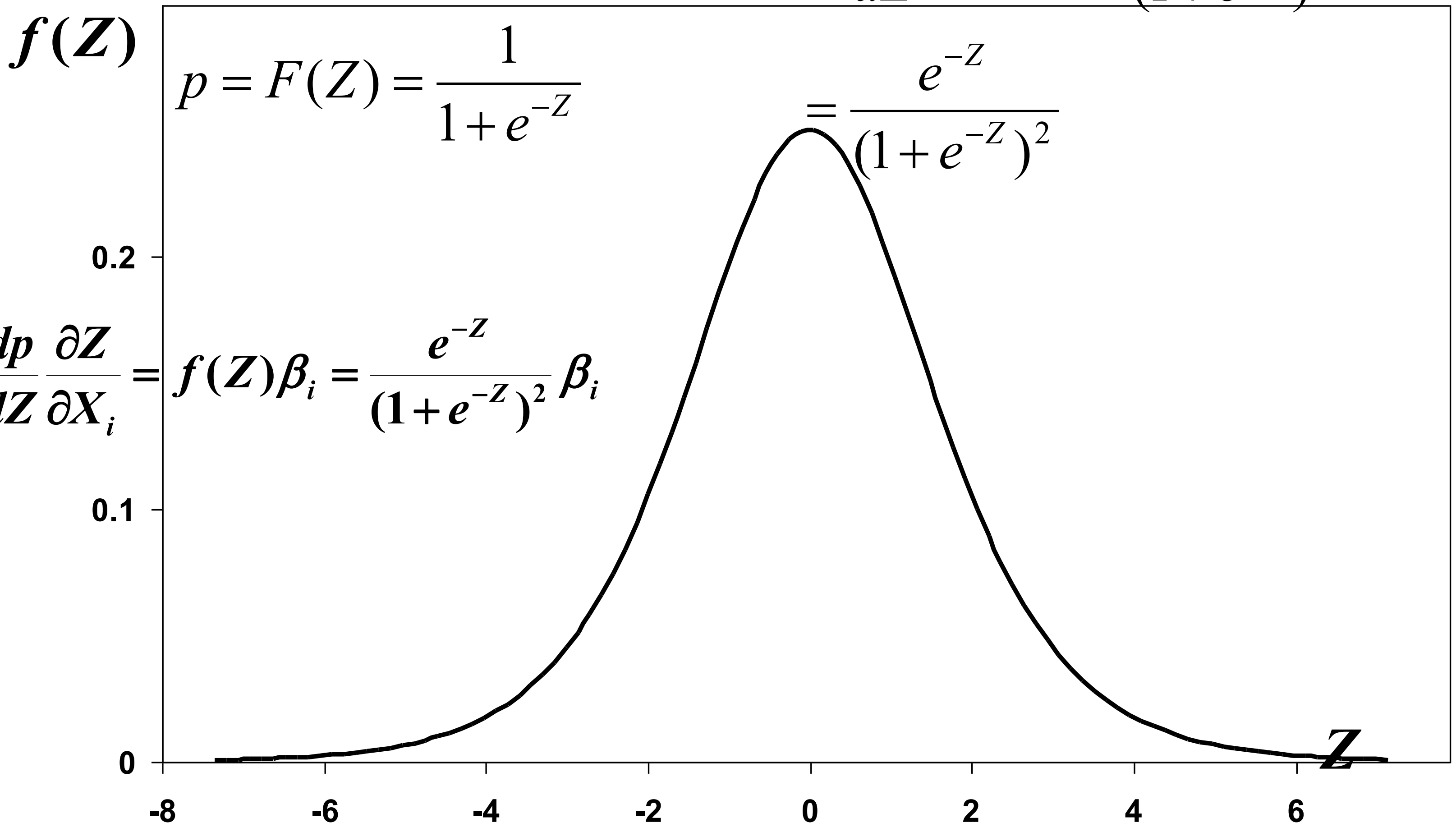
BINARY CHOICE MODELS: LOGIT ANALYSIS

$$f(Z) = \frac{dp}{dZ} = \frac{(1+e^{-Z}) \times 0 - 1 \times (-e^{-Z})}{(1+e^{-Z})^2}$$

$$= \frac{e^{-Z}}{(1+e^{-Z})^2}$$

$$p = F(Z) = \frac{1}{1+e^{-Z}}$$

$$\frac{\partial p}{\partial X_i} = \frac{dp}{dZ}\frac{\partial Z}{\partial X_i} = f(Z)\beta_i = \frac{e^{-Z}}{(1+e^{-Z})^2}\beta_i$$



The marginal partial effect is measured by the slope, gets maximum if *Z*=0. The marginal function, *f*(*Z*), reaches a maximum at this point.

ICEF, EGE_SUM and UL_PASS
LOGIT MODEL

Dependent Variable: UL_PASS   Method: ML - Binary Logit   Included observations: 238

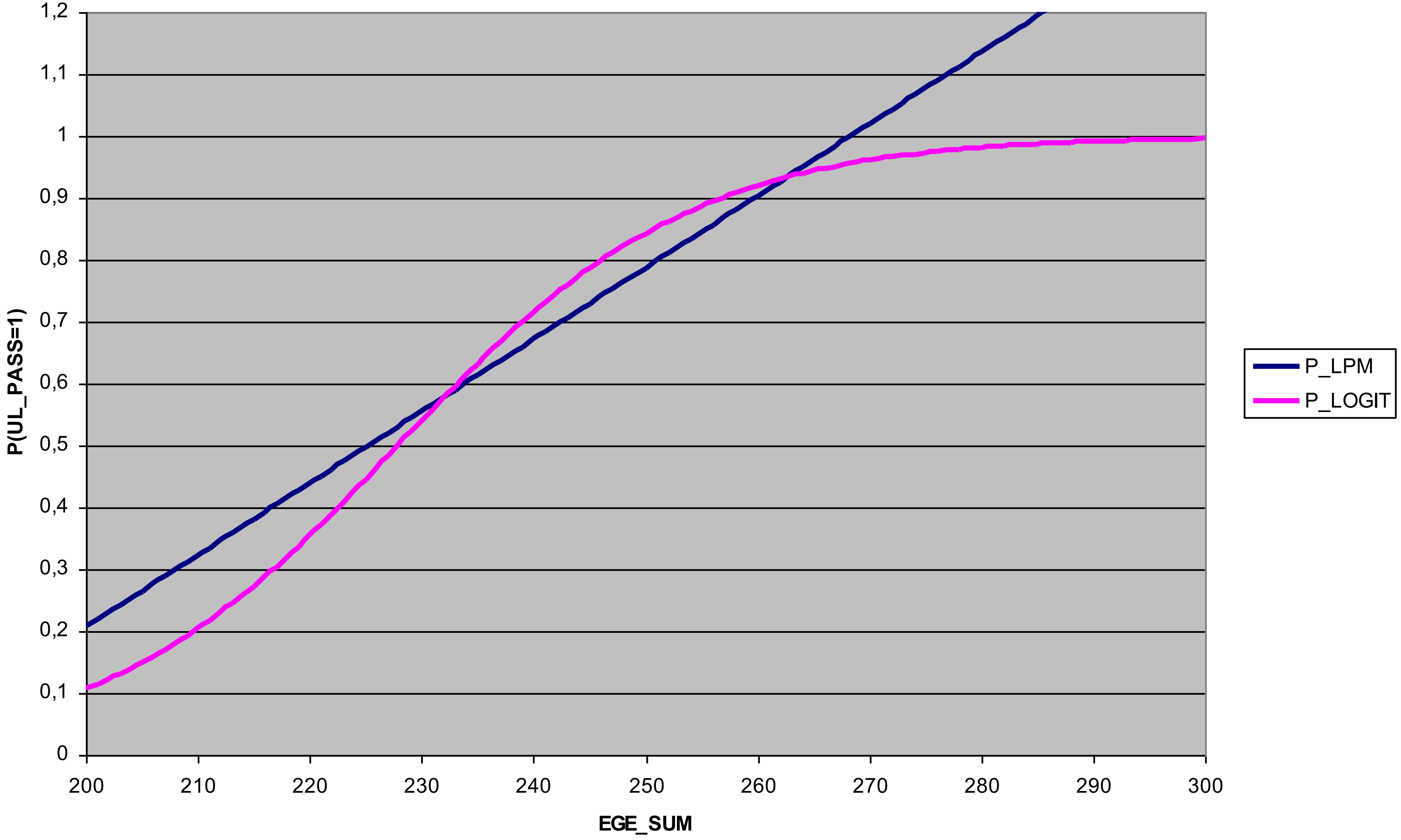
Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	-17.57812	3.853663	-4.561405	0.0000
EGE_SUM	0.075565	0.015478	4.882173	0.0000

McFadden R-squared	0.132192	Mean dependent var	0.815126
S.E. of regression	0.364	S.D. dependent var	0.4317
Sum squared resid	31.23	Log likelihood	-98.87
LR statistic (1 df)	30.12	Probability(LR stat)	0.0000

$$Z = -17.58 + 0.0756 * EGE\_SUM$$

$$p(UL\_PASS = 1) = F(Z) = \frac{1}{1+e^{-Z}} = \frac{1}{1+e^{17.58-0.0756*EGE\_SUM}}$$

Probability to pass to the UoL: LPM and Logit



LOGIT MODEL, ICEF, EGE_SUM and UL_PASS:

MARGINAL EFFECT

$$Mean\ EGE\_SUM = \bar{X} = 256.4$$

$$Z = \beta_1 + \beta_2 \bar{X} = -17.58 + 0.076 \times 256.4 = 1.906$$

$$e^{-Z} = e^{-1.906} = 0.149$$

$$f(Z) = \frac{dp}{dZ} = \frac{e^{-Z}}{(1 + e^{-Z})^2} = \frac{0.149}{(1 + 0.149)^2} = 0.113$$

$$\frac{\partial p}{\partial X} = \frac{dp}{dZ} \frac{\partial Z}{\partial X} = f(Z)\beta_2 = 0.113 \times 0.076 = 0.0085$$

The marginal effect, evaluated at the mean, equals 0.0085. This implies that a one point increase in EGE_SUM would increase the probability of admission to the University of London by 0.85 percent points. It is slightly less than the LPM model slope 0.0096.

LOGIT MODEL, ICEF, EGE_SUM and UL_PASS:

MARGINAL EFFECT

$$Z = -17.578 + 0.0756X = 0 \quad \Rightarrow \quad X \approx 232.5$$

$$e^{-Z} = e^{0} = 1$$

$$f(Z) = \frac{dp}{dZ} = \frac{e^{-Z}}{(1+e^{-Z})^2} = \frac{1}{(1+1)^2} = 0.25$$

$$\frac{\partial p}{\partial X} = \frac{dp}{dZ}\frac{\partial Z}{\partial X} = f(Z)\beta = 0.25 \times 0.0756 = 0.019$$

The biggest marginal effect is in the point where Z=0, or EGE_SUM≈232.5, it equals 0.019. One point increase in EGE_SUM would increase the probability of admission to the University of London by 1.9 percent points. It is bigger than at EGE_SUM=256.4, or than that of the LPM.

Marginal partial effects of explanatory variables

The marginal partial effects are not constant in the Logit model.

– Partial effects at the average (the notation *g(z) = f(z) = F`(z) = G`(z))*:

$$\widehat{PEA}_j = g(\bar{\boldsymbol{x}}\hat{\boldsymbol{\beta}})\hat{\beta}_j$$

The partial effect of explanatory variable x_j is considered for an "average" individual (this is problematic in the case of explanatory variables such as gender)

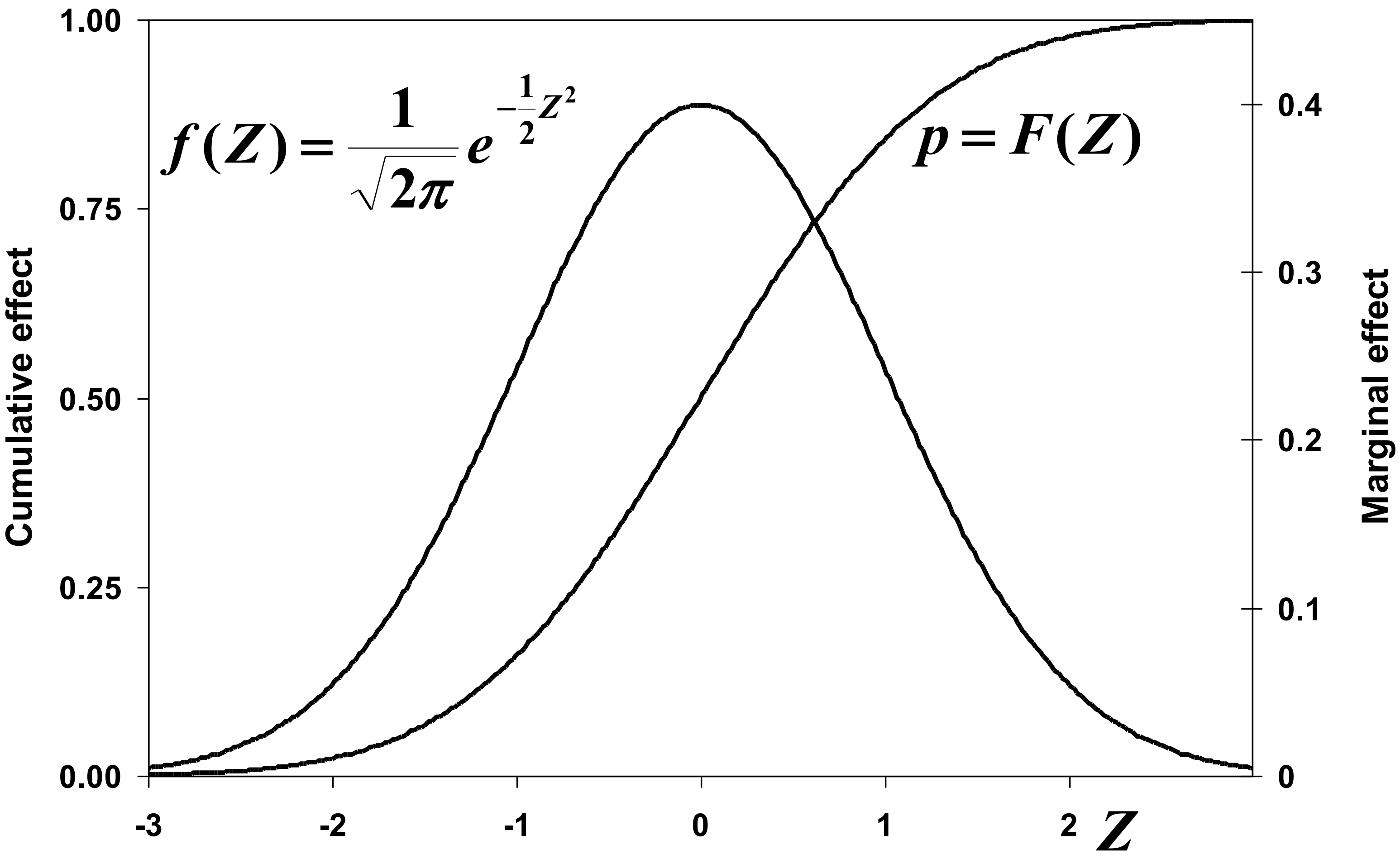
– Average partial effects:

$$\widehat{APE}_j = n^{-1}\sum_{i=1}^{n} g(\mathbf{x}_i\hat{\boldsymbol{\beta}})\hat{\beta}_j$$

The partial effect of explanatory variable x_j is computed for each individual in the sample and then averaged across all sample members (makes more sense)

– Analogous formulas hold for discrete explanatory variables.

BINARY CHOICE MODELS: PROBIT ANALYSIS



$$Z = \beta_1 + \beta_2 X_2 + ... + \beta_k X_k$$

Probit model: sigmoid function F is the cumulative standardized normal distribution. *f(Z)* – probability density function.

PROBIT MODEL: MARGINAL PARTIAL EFFECT

$$\boldsymbol{p = F(Z)} \qquad \boldsymbol{Z = \beta_1 + \beta_2 X_2 + \ldots \beta_k X_k}$$

$$f(Z) = \frac{dp}{dZ} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}Z^2}$$

$$\frac{\partial p}{\partial X_i} = \frac{dp}{dZ}\frac{\partial Z}{\partial X_i} = f(Z)\beta_i = \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}Z^2}\right)\beta_i$$

EViews: Quick – Estimate Equation – Equation Specification (type) –

Method: Binary - Probit

ICEF, EGE_SUM and UL_PASS
PROBIT MODEL

Dependent Variable: UL_PASS   Method: ML - Binary Probit   Included observations: 238

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	-9.762781	2.093597	-4.663161	0.0000
EGE_SUM	0.042141	0.008332	5.057498	0.0000

McFadden R-squared	0.132392	Mean dependent var	0.815126
S.E. of regression	0.364	S.D. dependent var	0.4317
Sum squared resid	31.29	Log likelihood	-98.85
LR statistic (1 df)	30.17	Probability(LR stat)	0.0000

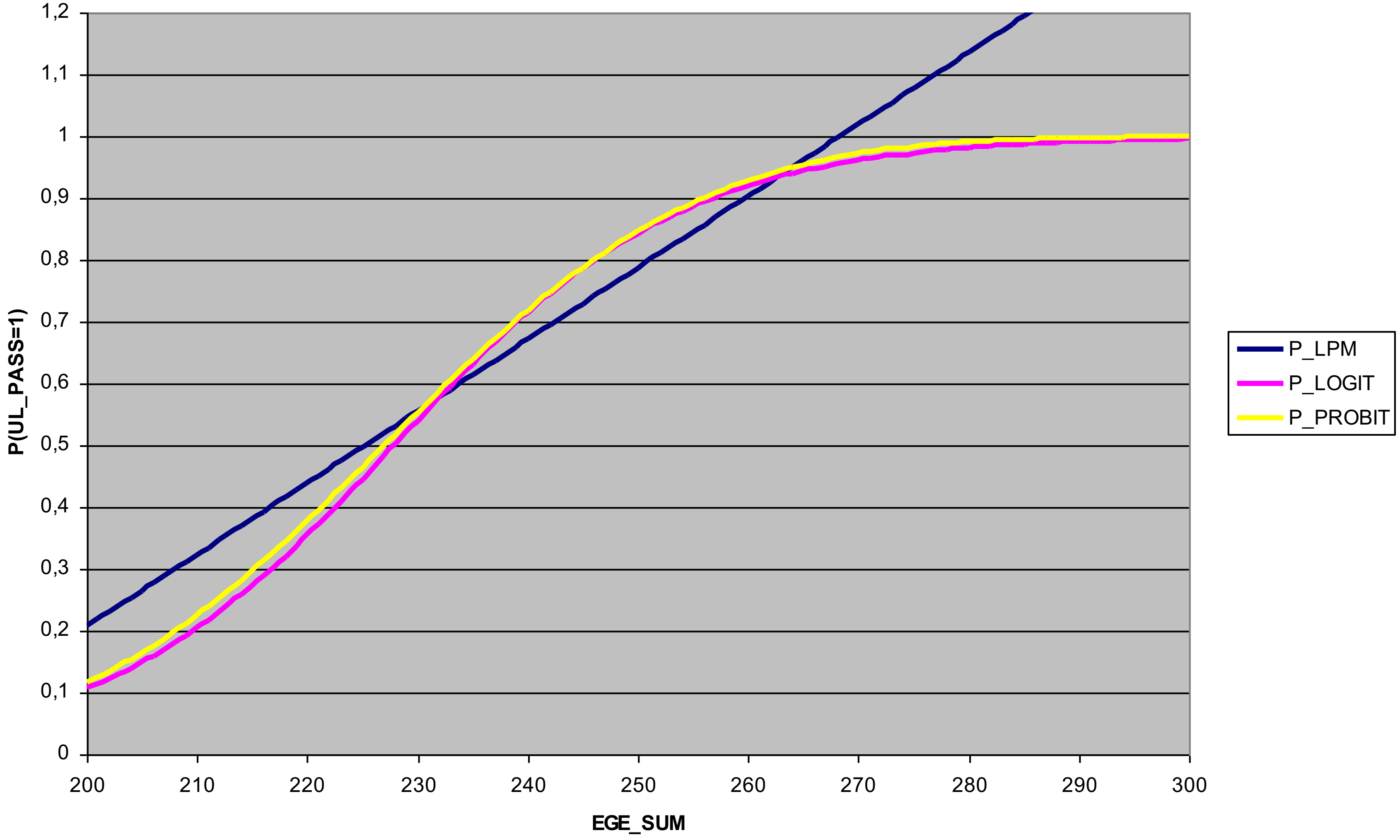
$$Z = -9.763 + 0.04214 * EGE\_SUM$$

$$p(UL\_PASS = 1) = F(Z)$$

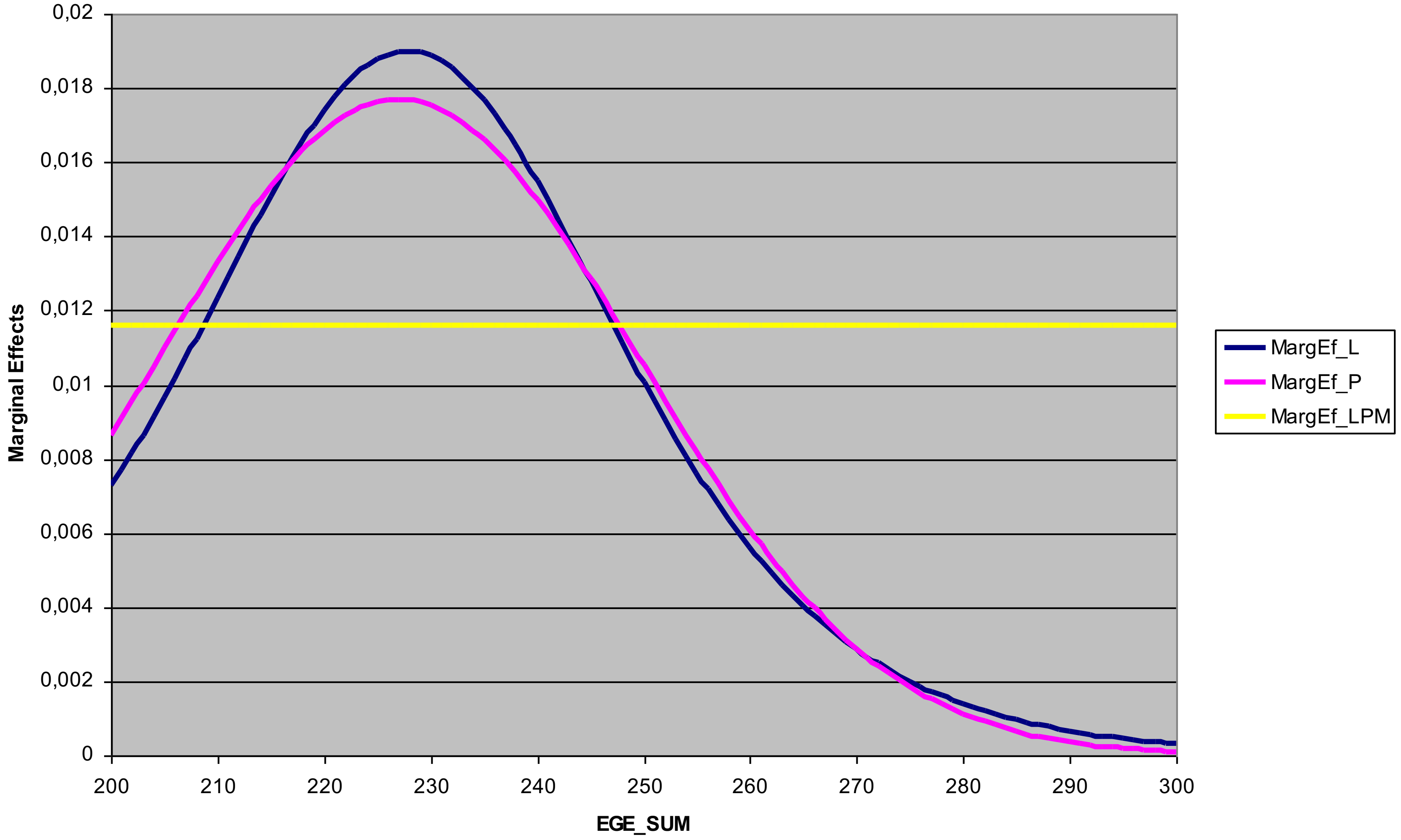
$$Z = -9.763 + 0.04214X = 0 \Rightarrow X \approx 227.2$$

$$\frac{\partial p}{\partial X} = \frac{dp}{dZ}\frac{\partial Z}{\partial X} = f(Z)\beta_2 = \frac{1}{\sqrt{2\pi}}0.04214 = 0.0168$$

Probability to pass to the UoL: LPM, Logit and Probit



Marginal Effects: LPM, Logit, Probit



How to fit Probit or Logit?
How to choose between them?

- How to fit logit and probit models? Why the Maximum Likelihood estimation is applied?
- What are the statistics and tests?
- How to choose between logit and probit?