Elements of Econometrics.
Lecture 30.
Revision 4

FCS, 2022-2023

Maximum likelihood estimation

MLE is widely used in estimating various models, and for some of them it is the principal estimation method.

It provides the estimates of parameters θ which maximise joint probability (or probability density) of the sample available:

$f(y_1,y_2,\ldots,y_n;\theta)\rightarrow max$, or

$f(y_1;\theta)f(y_2;\theta)\ldots f(y_n;\theta)\rightarrow max$ for the case of independent y_i.

MLE is **usually consistent** and often unbiased.

Can be used for special types of models where Least Squares is not applicable (like Limited Dependent Variable Models).

MLE is generally the most asymptotically efficient estimator when the population model $f(y;\theta)$ is correctly specified.

MLE is sometimes the **minimum variance unbiased estimator.**

Maximum Likelihood Estimation

- Suppose we have a regression

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

Where ε is iid (independent & identically distributed) random variable and all G-M conditions are satisfied.

How to estimate its parameters?
OLS

RSS->min

MLE
•We have 3 pieces of info:

1. A sample of observations - $(Y_1, X_1), (Y_2, X_2), \cdots, (Y_N, X_N)$
2. The model - $Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$
3. A specified pdf for the error - $\varepsilon_i \sim N(0, \sigma^2)$

Maximum Likelihood Estimation

We want to estimate: β_1, β_2, σ on the basis of N sample values of Y&X

Let us construct Likelihood function (assume u_i independent (G-M)). Remember that Likelihood function is a product of density (or probability) functions and stays for probability of the sample. So ML estimates give the highest probability (likelihood) of the sample

$$L(\beta_1, \beta_2, \sigma \mid Y_1, \ldots, Y_n) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{Y_1-\beta_1-\beta_2 X_1}{\sigma}\right)^2} \times \ldots \times \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{Y_n-\beta_1-\beta_2 X_n}{\sigma}\right)^2} = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{Y_i-\beta_1-\beta_2 X_i}{\sigma}\right)^2}$$

We then choose β_1, β_2 which maximizes the likelihood

$$L = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N e^{-\frac{1}{2}\left(\frac{Y_1-\beta_1-\beta_2 X_1}{\sigma}\right)^2} e^{-\frac{1}{2}\left(\frac{Y_2-\beta_1-\beta_2 X_2}{\sigma}\right)^2} \cdots e^{-\frac{1}{2}\left(\frac{Y_N-\beta_1-\beta_2 X_N}{\sigma}\right)^2}$$

Alternatively we can maximize the log likelihood which is easier mathematically

$$log(L) = log\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(Y_i - \beta_1 - \beta_2 X_i)^2$$

Compare with the OLS? The same, but estimate of σ is slightly different

Maximum Likelihood Estimation

How to find β_1, β_2, σ

Log likelihood becomes:

$$\log L = l(\beta_1, \beta_2, \sigma \mid Y_1, \ldots, Y_n) = \sum_{i=1}^{n} \log\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{Y_i - \beta_1 - \beta_2 X_i}{\sigma}\right)^2} \right) = n \log\left(\frac{1}{\sigma\sqrt{2\pi}} \right) - \frac{1}{2} \sum_{i=1}^{n} \left(\frac{Y_i - \beta_1 - \beta_2 X_i}{\sigma} \right)^2$$

$$= n \log\left(\frac{1}{\sigma\sqrt{2\pi}} \right) - \frac{\sigma^{-2}}{2} Z \quad where\ Z = RSS = \sum_{i=1}^{n} (Y_i - \beta_1 - \beta_2 X_i)^2$$

From this obtained expression, maximization of the log likelihood implies minimization of Z. Therefore, it is identical to the OLS procedure ($RSS \to \min_{\beta_1, \beta_2}$) for choosing estimators of β_1 and β_2. Hence, ML estimators of β_1 and β_2 coincide with the OLS ones.

Let's obtain the expression for $\hat{\sigma}_{ML}$ .FOC: $\frac{\partial l}{\partial \sigma} = -\frac{n}{\sigma} + \sigma^{-3} Z = \sigma^{-3}\left(Z - n\sigma^2\right)$

Hence, $\hat{\sigma}^2_{ML} = \frac{Z}{n} = \frac{\sum_{i=1}^{n} e_i^2}{n}$

Properties of ML estimators

- Consistent
- Asymptotically Normally distributed
- Asymptotically efficient in a class of consistent normally distributed estimators

- Thus, in general ML estimates are biased (but they can be unbiased in special cases such as Linear regression)
- Moreover, you should use *z*-test instead of *t*-test, because ML estimators are asymptotically normally distributed

OLS vs ML

- OLS in general gives unbiased estimators (ML estimators are generally biased)
- ML estimators are asymptotically normally distributed which is better for testing, because we know the exact distribution for large samples
- ML can be used for estimation of nonlinear in parameters models and for some types of models where the Least Squares is not applicable.

Binary Choice Models

Qualitative response models (discrete choice: binary or multinominal)

Binary Choice (binary dependent variable):

1. Linear Probability Model
2. Logit
3. Probit

$$Y_i = \begin{cases} 1 \\ 0 \end{cases}$$

- What characteristics affect the likelihood that an individual obtains a higher degree?
- Why do some people buy houses while others rent?
- What determines labour force participation?

LINEAR PROBABILITY MODEL

The linear probability model is the linear multiple regression model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i, \tag{9.2}$$

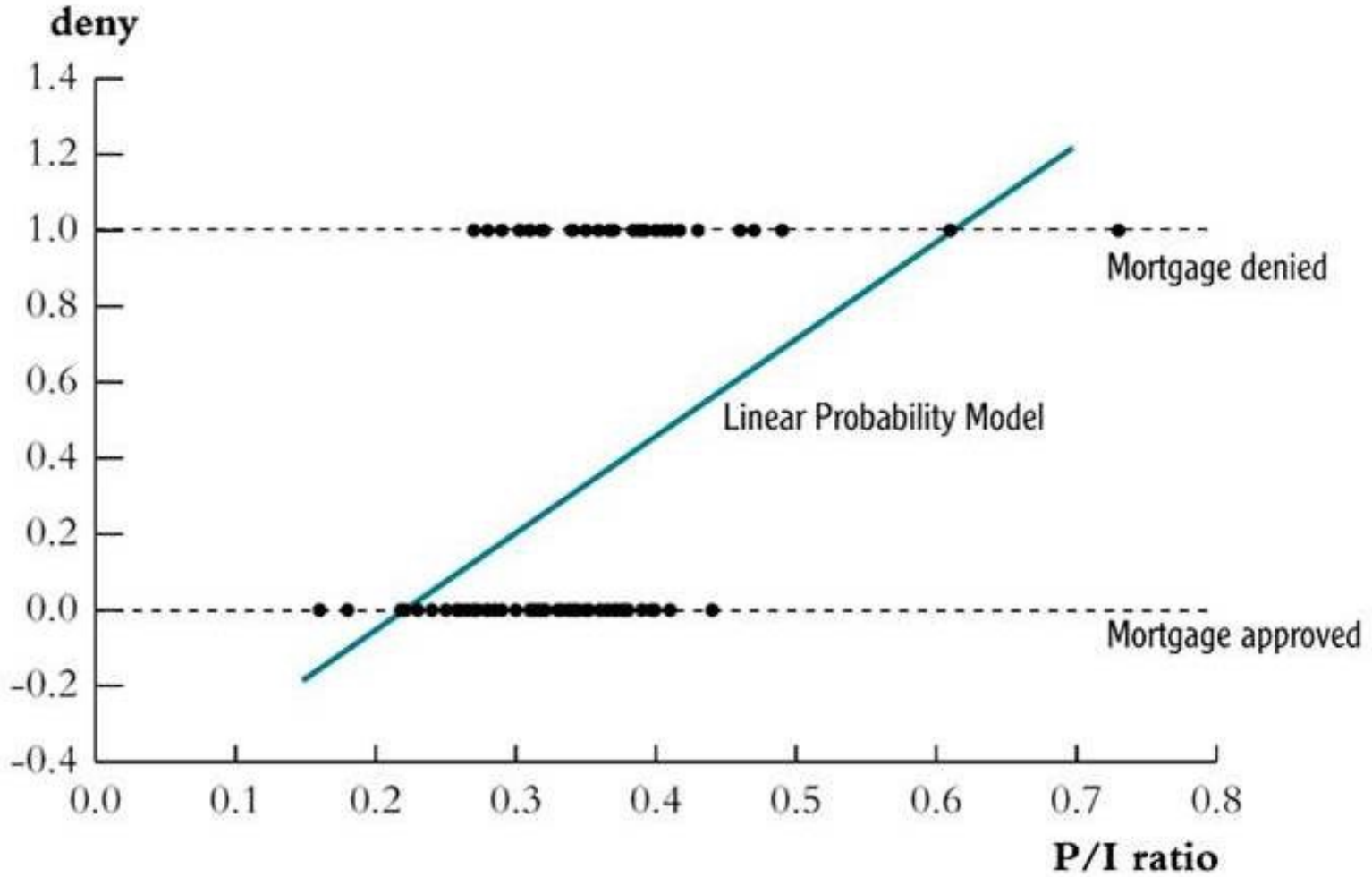
where Y_i is binary, so that

$$\Pr(Y = 1 \mid X_1, X_2, \ldots, X_k) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k.$$

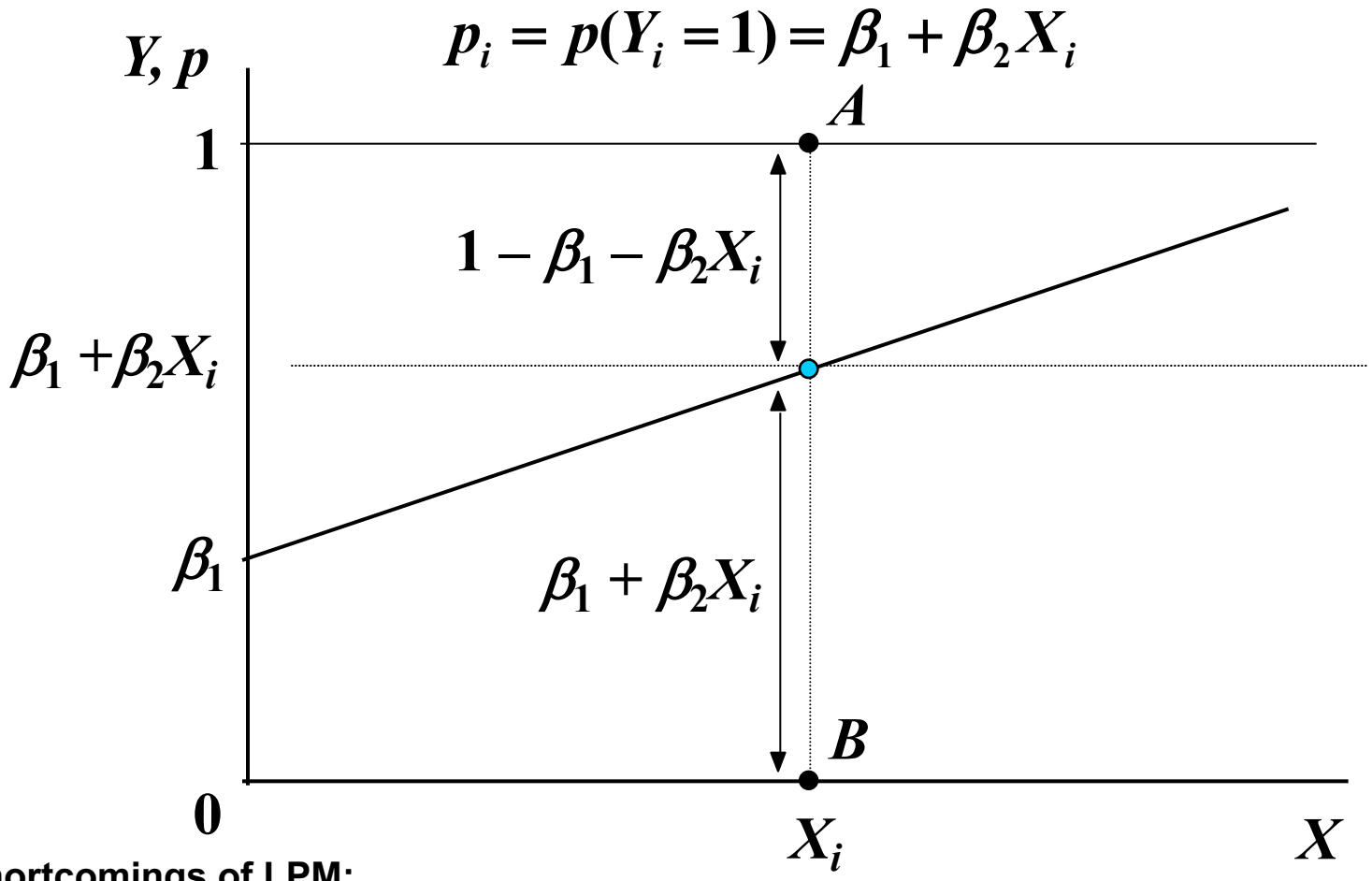
The regression coefficient β_1 is the change in the probability that $Y = 1$ associated with a unit change in X_1, holding constant the other regressors, and so forth for β_2, etc. The regression coefficients can be estimated by OLS, and the usual (heteroskedasticity-robust) OLS standard errors can be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

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



Shortcomings of LPM:

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

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

Marginal effect of each factor is constant.

Logit Model

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

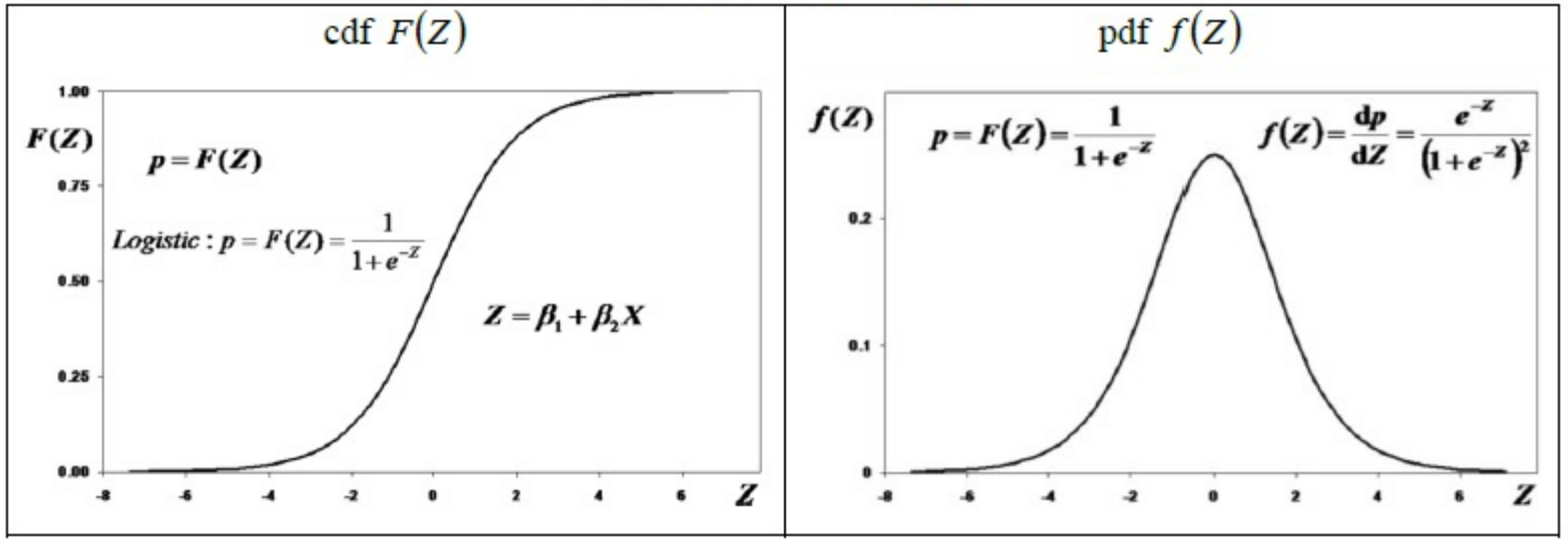
Logit function: $$Pr(Y_i = 1|X_i) = F(Z) = \frac{1}{1 + e^{-Z_i}}$$

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

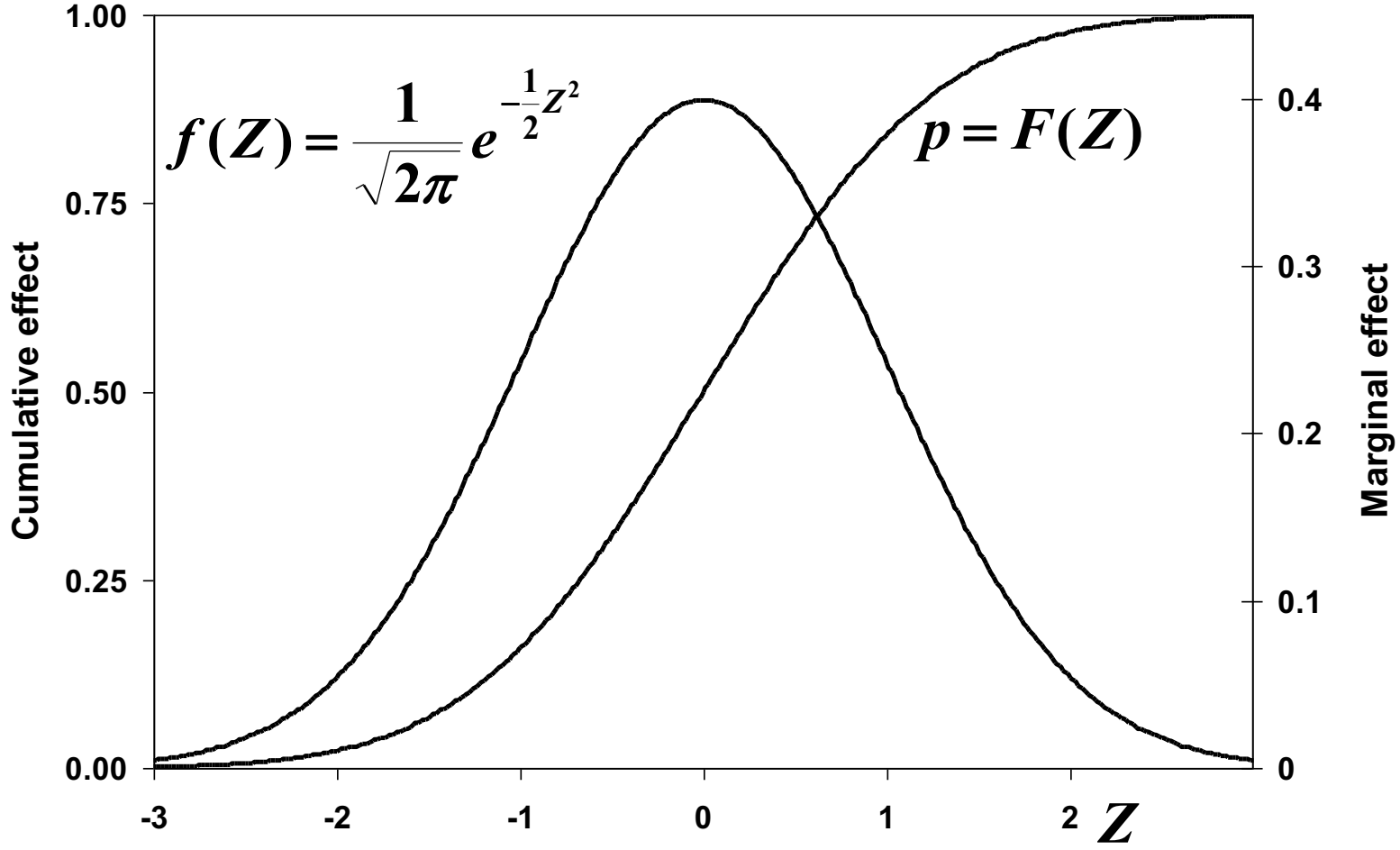
where $Z_i = \beta_1 + \beta_2 X_i$

- $\beta_2 f(Z)$ is now the change in the probability of choosing 1 given a unit change in X
- f(Z) is the derivative of F(Z) i.e. $f(Z) = \frac{e^{-Z}}{(1+e^{-Z})^2}$
- f(Z) is usually evaluated at the mean of all the independent variables

Logit model: cdf and pdf



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 EFFECT

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

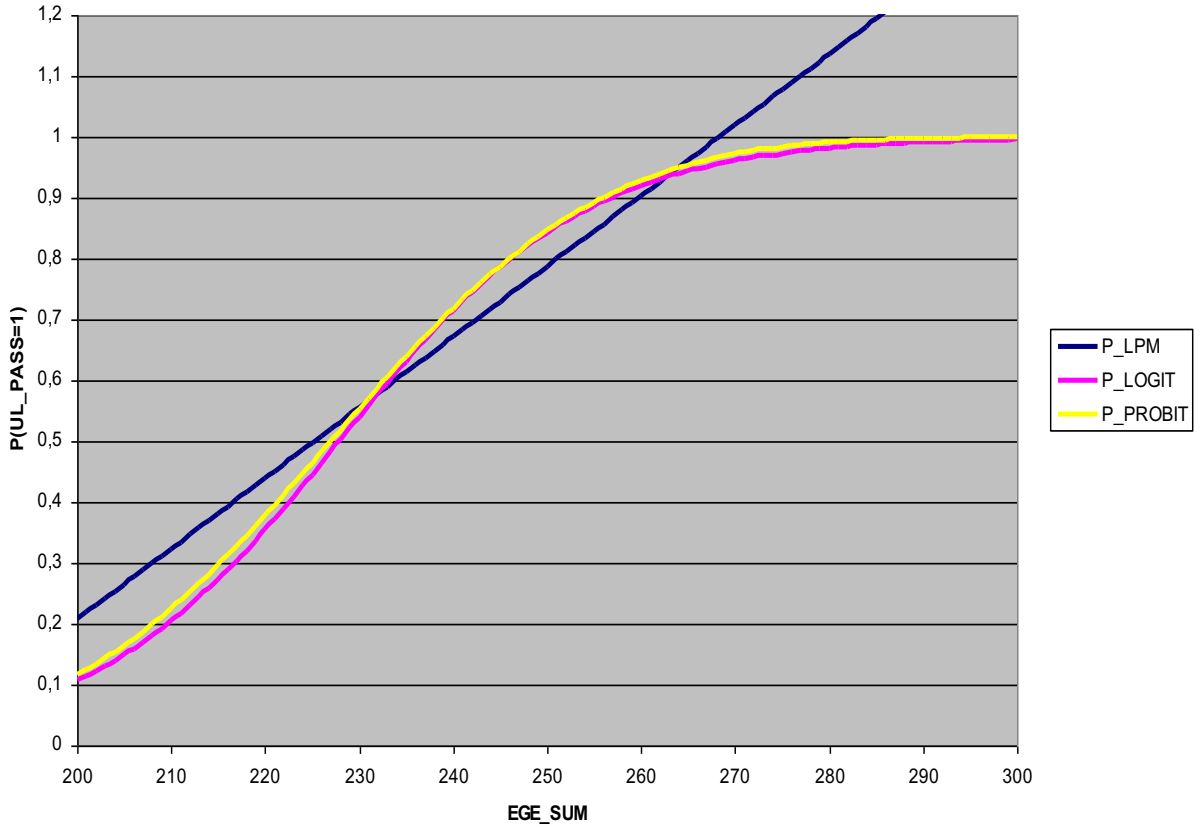
$$f(Z) = \frac{dp}{dZ} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}Z^2}$$

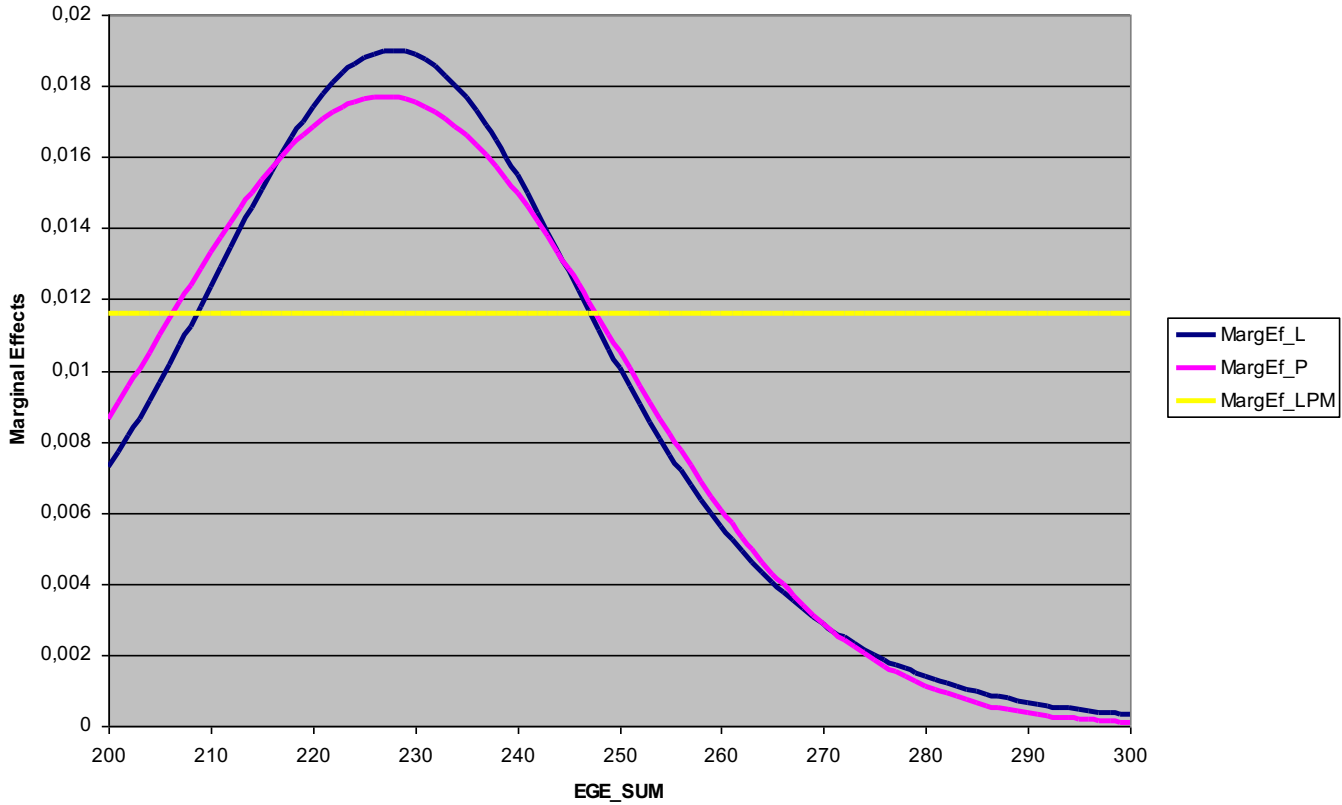
$$\frac{\partial p}{\partial X_i} = \frac{dp}{dZ}\frac{\partial Z}{\partial X_i} = f(Z)\beta_i = \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}Z^2}\right)\beta_i$$

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

Method: Binary - Probit

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





Maximum Likelihood Estimation of the Logit and Probit Models

$$L = \prod_i p(Y = Y_i \,|X_i, \beta) = \prod_{i:Y_i=1} F(\beta_1 + \beta_2 \cdot X_i) \cdot \prod_{i:Y_i=0} (1 - F(\beta_1 + \beta_2 \cdot X_i)) \to \max_\beta$$

$$l(\beta) = \log L = \sum_i (\log p(Y = Y_i \,|X_i, \beta)) =$$

$$= \sum_{i:Y_i=1} \log F(\beta_1 + \beta_2 \cdot X_i) + \sum_{i:Y_i=0} \log(1 - F(\beta_1 + \beta_2 \cdot X_i)) =$$

$$= \sum_i Y_i (\log F(\beta_1 + \beta_2 \cdot X_i)) + \sum_i (1 - Y_i)(\log(1 - F(\beta_1 + \beta_2 \cdot X_i))) \to \max_\beta$$

For logit model $F(\beta_1 + \beta_2 \cdot X_i) = \dfrac{1}{1 + e^{-(\beta_1 + \beta_2 \cdot X_i)}}$

For probit model $F(\beta_1 + \beta_2 \cdot X_i) -$

cumulative function of standardized normal distribution

The goodness of fit in maximum likelihood estimation

$$\text{"Pseudo-}R^2\text{"}\ (\text{or } McFadden\ R^2) = 1 - \frac{\log L}{\log L_0}$$

where $log\ L_0$ is the natural logarithm of the value the likelihood function would take with only the intercept in the regression.

$\log L < 0$, since $0 < L < 1$.

The values of Pseudo- R^2 range from 0 to 1; the closer this coefficient is to 1, the better the fit.

The Goodness of Fit in Maximum Likelihood Estimation

***The likelihood ratio*:**

$$LR = 2\log(\frac{L}{L_0}) = 2(\log L - \log L_0)$$

The likelihood ratio is used to test the following hypothesis:
H_0: the coefficients of all explanatory variables are equal to zero
H_1: the coefficient of at least one explanatory variable is not equal to zero.

Under the null hypothesis, the statistic *LR* has a χ^2-distribution with (*k*-1) degrees of freedom, where *k* is the number of parameters estimated, and, accordingly, (*k*-1) is the number of explanatory variables.

The significance of individual coefficients is tested via z-statistics,
whose distribution approaches the standard normal in large samples.

The Likelihood Ratio Test for Variables Exclusion Restrictions

Maximum likelihood estimation (MLE), provides with a log-likelihood, L

As in F test, we estimate the restricted and unrestricted models, then form

$LR = 2(L_{ur} - L_r) \sim \chi^2_q$ where q is the number of restrictions (excluded explanatory variables).

Probit or Logit?

- Both the probit and logit are nonlinear and require maximum likelihood estimation
- The functions *F*, pseudo-R^2, *LR* statistics and *p*-values are usually close to each other
- No real reason to prefer one over the other
- Traditionally logit was wider used because the logistic function leads to a more easily computed model
- Now, probit is easy to compute with standard packages, and is used widely