SOS3003
Applied data analysis for social science
Seminar note 03-2009

Erling Berge
Department of sociology and political science
NTNU

Why logistic regression?

- Hamilton Ch 7 p217-219

LOGIT REGRESSION

- **Should be used if the dependent variable (Y) is a nominal scale**
- Here it is assumed that Y has the values 0 or 1
- The model of the conditional probability of Y, E[Y | X], is based on the logistic function

  (E[Y | X] is read “the expected value of Y given the value of X”)
- But

  Why cannot E[Y | X] be a linear function also in this case?

The linear probability model: LPM

- The linear probability model (LPM) of Y_i when Y_i can take only two values (0, 1) assumes that we can interpret $E[Y_i \mid \mathbf{X}]$ as a probability
- $E[Y_i \mid \mathbf{X}] = b_0 + \Sigma_j \, b_j \, x_{ji} = Pr[Y_i = 1]$
- This leads to severe problems:

Are the assumptions of a linear regression model satisfied for the LPM?

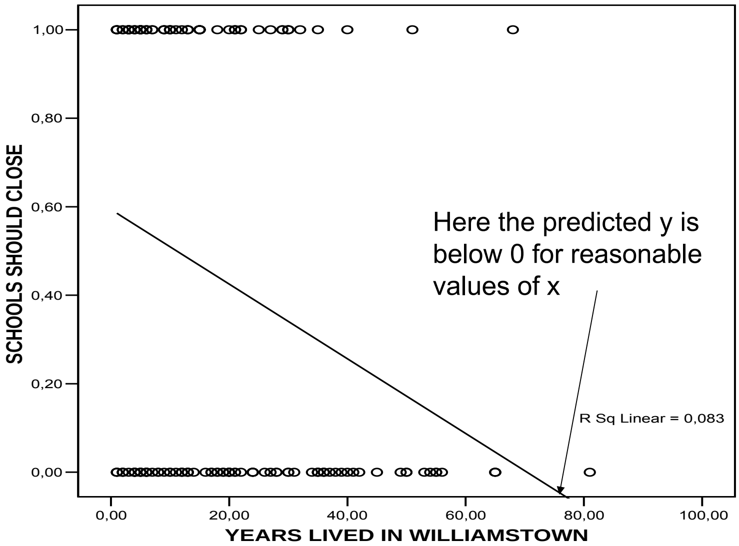
- One assumptions of the LPM is that the residual, e_i satisfies the requirements of OLS
- The the residual must be either $e_i = 1 – (b_0 + \Sigma_j b_j x_{ji})$ or $e_i = 0 – (b_0 + \Sigma_j b_j x_{ji})$
- This means that there is heteroscedasticity (the residual varies with the size of the values on the x-variables)
- There are estimation methods that can get around this problem (such as 2-stage weighted least squares method)
- One example of LPM:

OLS regression of a binary dependent variable on the independent variable ”years lived in town”

ANOVA tabell	Sum of Squares	df	Mean Square	F	Sig.
Regression	3,111	1	3,111	13,648	,000(a)
Residual	34,418	151	,228		
Total	37,529	152			

Dependent Variable: SCHOOLS SHOULD CLOSE	B	Std. Error	t	Sig.
(Constant)	,594	,059	10,147	,000
YEARS LIVED IN TOWN	-,008	,002	-3,694	,000

The regression looks OK in these tables



Scatter plot with line of regression. Figure 7.1 Hamilton

Conclusion: LPM model is wrong

- The example shows that for reasonable values of the x variable we can get values of the predicted y where

  $E[Y_i \mid \mathbf{X}] > 1$ or $E[Y_i \mid \mathbf{X}] < 0$,
- For this there is no remedy
- <u>LPM is for substantial reasons a wrong model</u>
- We need a model where we always will have

  $0 \leq E[Y_i \mid \mathbf{X}] \leq 1$
- The logistic function can provide such a model

The logistic function

The general logistic function is written

- $Y_i = \alpha/(1+\gamma^*\exp[-\beta X_i]) + \varepsilon_i$

$\alpha>0$ provides an upper limit for Y

this means that $0<Y<\alpha$

γ determines the horizontal point for rapid growth

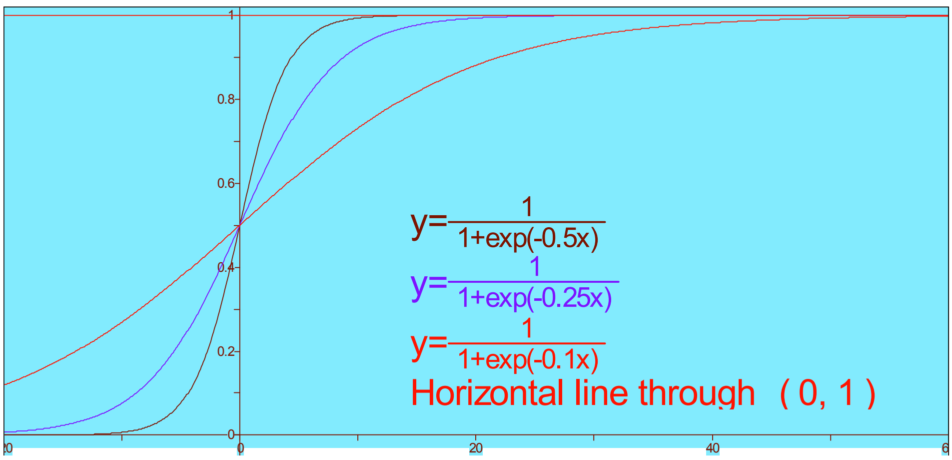
If we determines that $\alpha = 1$ and $\gamma = 1$

One will always find that

- $0 < 1/(1+\exp[-\beta X_i]) < 1$

The logistic function will for all values of x lie between 0 and 1

Logistic curves for different β



β determines how rapidly the curve grows

MODELL (1)

Definisjonar

- Sannsynet for at person i skal ha verdien 1 på variabelen Y skriv vi $Pr(Y_i=1)$. Da er $Pr(Y_i \neq 1) = 1 - Pr(Y_i=1)$
- Oddsen for at person i skal ha verdien 1 på variabelen Y_i, her kalla O_i , er tilhøvet mellom to sannsyn:

$$O_i\left(y_i = 1\right) = \frac{\Pr\left(y_i = 1\right)}{1 - \Pr\left(y_i = 1\right)} = \frac{p_i}{1 - p_i}$$