SOS3003
Applied data analysis for social science
Lecture note 07-2009

Erling Berge
Department of sociology and political science
NTNU

Literature

- Logistic regression I
  Hamilton Ch 7 p217-234

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 | **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?

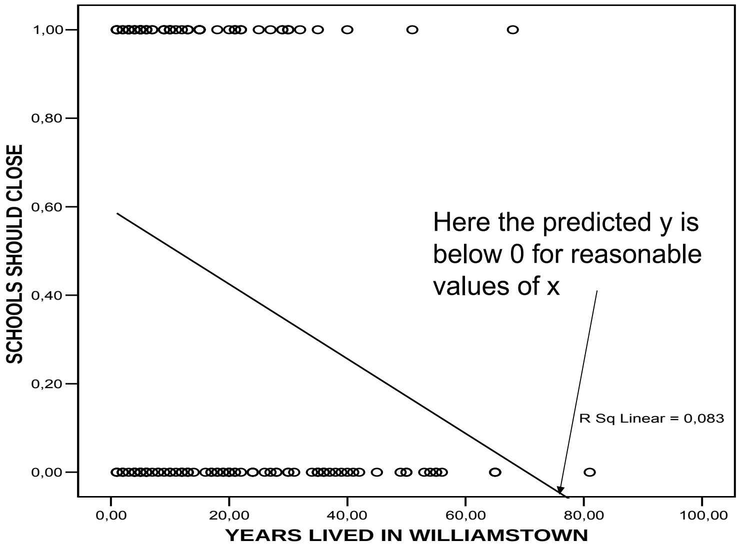
- One assumptions of the LPM is that the residual, e_i satisfies the requirements of OLS
- The the residual must be either
  - $e_i = 1 - (b_0 + \Sigma_j b_j x_{ji})$ or
  - $e_i = 0 - (b_0 + \Sigma_j b_j x_{ji})$
- This means that there is heteroscedasticity (the residual varies with the size of the 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

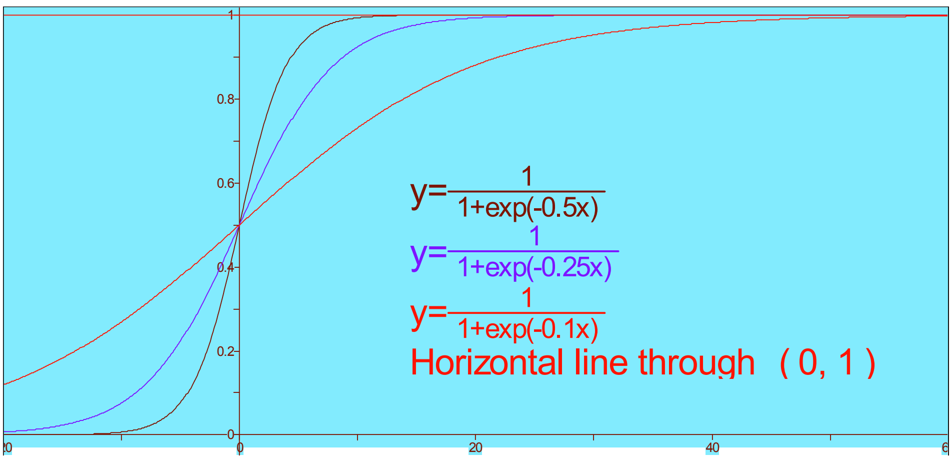
If we determines that $\alpha = 1$ and $\gamma = 1$

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

MODEL (1)

Definitions:

- The probability that person no i shall have the value 1 on the variable Y will be written $Pr(Y_i = 1)$. Then $Pr(Y_i \neq 1) = 1 - Pr(Y_i = 1)$
- The odds that person no i shall have the value 1 on the variable Y, here called O_i, is the ratio between two probabilities

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

MODEL (2)

Definitions:

- The LOGIT, L_i , is the natural logarithm of the odds, O_i , for person no i:

  $L_i = \ln(O_i)$
- The model assumes that L_i is a linear function of the explanatory variables x_j ,
- i.e.:
- $L_i = \beta_0 + \Sigma_j \beta_j x_{ji}$, where j=1,..,K-1, and i=1,..,n

MODEL (3)

- Let X = (the collection of all x_j), then the probability of Y_i = 1 for person no i

$$\Pr(y_i = 1) = E\left[y_i \mid \mathrm{X}\right] = \frac{1}{1+\exp\left(-L_i\right)} = \frac{\exp(L_i)}{1+\exp(L_i)}$$

$$\text{where } L_i = \beta_0 + \sum_{j=1}^{K-1} \beta_j X_{ji}$$

The graph of this relationship is useful for the interpretation what a change in x means

MODEL (4)

In the model $Y_i = E[Y_i \mid \mathrm{X}] + \varepsilon_i$ the error is either

- $\varepsilon_i = 1 - E[Y_i \mid \mathrm{X}]$ with probability $E[Y_i \mid \mathrm{X}]$
  (since $\Pr(Y_i = 1) = E[Y_i \mid \mathrm{X}]$),

or the error is

- $\varepsilon_i = - E[Y_i \mid \mathrm{X}]$ with probability $1 - E[Y_i \mid \mathrm{X}]$

- Meaning that the error has a distribution known as the binomial distribution with
  $p_i = E[Y_i \mid \mathrm{X}]$

Estimation

- The method used to estimate the parameters in the model is Maximum Likelihood
- The ML-method gives us the parameters that maximize the Likelihood of finding just the observations we have got
- This likelihood we call $\mathcal{L}$
- The criterion for choosing regression parameters is that the likelihood becomes as large as possible

Maximum Likelihood (1)

- The Likelihood equals the product of the probability of each observation. For a dichotomous variable where $\Pr(Y_i = 1)=P_i$ this can be written

$$\mathrm{L} = \prod_{i=1}^{n} \left\{ P_i^{Y_i} \left(1 - P_i\right)^{\left(1-Y_i\right)} \right\}$$

Maximum Likelihood (2)

- It is easier to maximize the likelihood $\mathcal{L}$

  if one uses the natural logarithm of $\mathcal{L}$:

$$\ln\left(\mathrm{L}\right)=\sum_{i=1}^{n}\left\{y_i \ln P_i+\left(1-y_i\right)\ln\left(1-P_i\right)\right\}$$

- The natural logarithm of $\mathcal{L}$ is called the LogLikelihood, It may be called $\mathcal{LL}$.
- $\mathcal{LL}$ has a central role in logistic regression.

Logistic model instead of LPM

Iteration	-2 Log Likelihood	Coefficients	
		Constant	Lived in town
Step 0	209,212	-,275	0
1	195,684	,376	-,034
2	195,269	,455	-,041
3	195,267	,460	-,041
4	195,267	,460	-,041

Dependent: Schools should close	B	S.E.	Wald	df	Sig.	Exp(B)
Lived in town	-,041	,012	11,399	1	,001	,960
Constant	,460	,263	3,069	1	,080	1,584

Footnotes to the table

- Step 0: Point of departure is a model with a constant and no variables
- **Iterative estimation**
  - Estimation ends at iteration no 4 since the parameter estimates changed less than 0.001
- The Wald statistic that SPSS provides equals the square of the "t" that Hamilton (and STATA) provides (Wald = t^2)

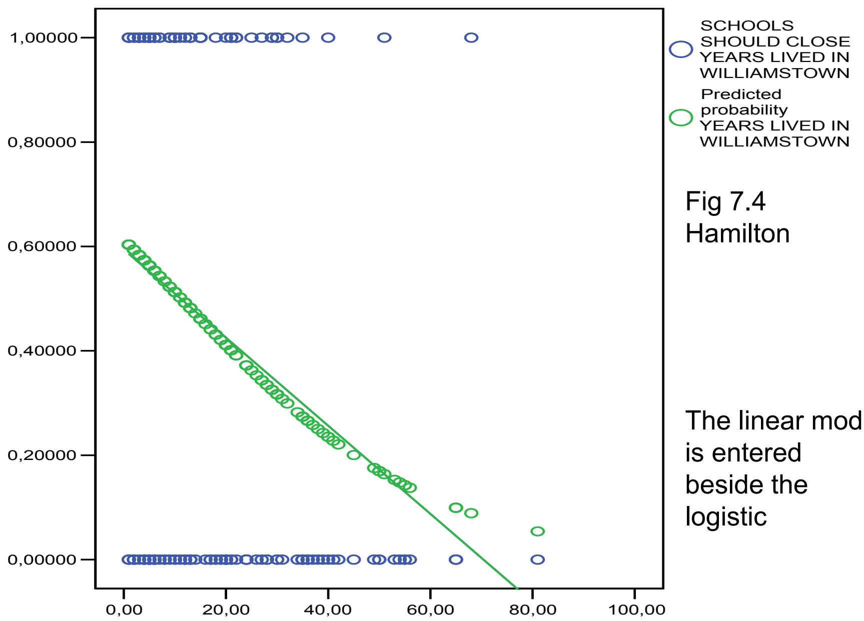


Fig 7.4 Hamilton

The linear model is entered beside the logistic

TESTING

Two tests are useful

- (1) The Likelihood ratio test
  - This can be used analogous to the F-test
- (2) Wald test
  - The square root of this can be used analogous to the t-test

Interpretation (1)

- The difference between the linear model and the logistic is large in the neighbourhood of 0 and 1
- LPM is easy to interpret: $Y_i = \beta_0$ when x_{1i}=0, and when x_{1i} increases with one unit Y_i increases with β_1 units
- The logistic model is more difficult to interpret. It is non-linear both in relation to the odds and the probability

ODDS and ODDS RATIOS

- The Logit, L_i, ($L_i = \beta_0 + \Sigma_j \beta_j x_{ji}$) is defined as the natural logarithm of the odds

This means that

- odds = $O_i(Y_i=1) = \exp(L_i) = e^{L_i}$

and

- **Odds ratio**= $O_i(Y_i=1| L_i') / O_i(Y_i=1| L_i)$
  - where L_i' and L_i have different values on only one variable $x_{.j.}$

Interpretation (2)

- When all x equals 0 then $L_i = \beta_0$ This means that the odds for $y_i = 1$ in this case is $\exp\{\beta_0\}$
- If all x-variables are kept fixed (they sum up to a constant) while x_1 increases with 1, the odds for $y_i = 1$ will be multiplied by $\exp\{\beta_1\}$
- This means that it will change with

  $100(\exp\{\beta_1\} – 1)$ %
- The probability $\Pr\{y_i = 1\}$ will change with a factor affect by all elements in the logit

Logistic regression: assumptions

- The model is correctly specified
  - The logit is linear in its parameteres
  - All relevant variables is included
  - No irrelevant variables are included
- x-variables are measured without error
- Observations are independent
- No perfect multicollinearity
- No perfect discrimination
- Sufficiently large sample

Assumptions that cannot be tested

- Model specification
  - All relevant variables are included
- x-variables are measured without error
- Observations are independent

Two will be tested automatically.

If the model can be estimated there is

- No perfect multicollinearity and
- No perfect discrimination

LOGISTIC REGRESSION
Statistical problems may be due to

- Too small a sample
- High degree of **multicollinearity**
  - Leading to large standard errors (imprecise estimates)
  - M is discovered and treated in the same way as in OLS regression
- High degree of **discrimination** (or separation)
  - Leading to large standard errors (imprecise estimates)
  - Will be discovered automatically by SPSS

Discrimination/ separation

- Problems with discrimination appear when we for a given x-value get almost perfect prediction of the y-value (nearly all with a given x-value have the same y-value)
- In SPSS it may produce the following message:

Warnings

• There is possibly a quasi-complete separation in the data. Either the maximum likelihood estimates do not exist or some parameter estimates are infinite.
• The NOMREG procedure continues despite the above warning(s). Subsequent results shown are based on the last iteration. Validity of the model fit is uncertain.

Discrimination in Hamilton table 7.5

- Odds for weaker requirements is 44/202 = 0,218 among women without small children
- Odds for weaker requirement is 0/79 = 0 among women with small children
- Odds rate is 0/0,218 = 0 hence $\exp\{b_{woman}\}=0$
- This means that b_{woman} = minus infinity

	Women without small children	Women with small children
No weaker require ments	202	79
Weaker require ments OK	44	0

Logistic regression

- If the assumptions are satisfied logistic regression will provide normally distributed, unbiased and efficient (minimal variance) estimates of the parameters