# SOS3003
# **Applied data analysis for social science**
# Lecture note 04-2010

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 $E[y_i \mid \mathbf{X}_i]$ as a probability
- $\mathbf{X}_i = \{x_{1i}, x_{2i}, x_{3i}, \ldots, x_{(K-1)i}\}$
- $E[y_i \mid \mathbf{X}_i] = b_0 + \Sigma_j\, b_j\, x_{ji} = Pr[y_i = 1]$
- 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:

Spring 2010 © Erling Berge 2010 5

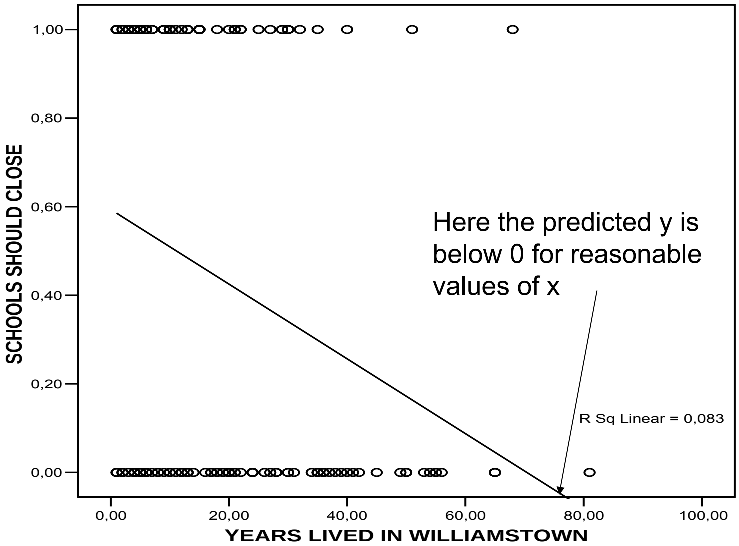
## OLS regression of a binary dependent variable on the independent variable "years lived in town"

| **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

Spring 2010 © Erling Berge 2010 6




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

this means that $0< y_i < \alpha$

$\gamma$ determines the horizontal point for rapid growth

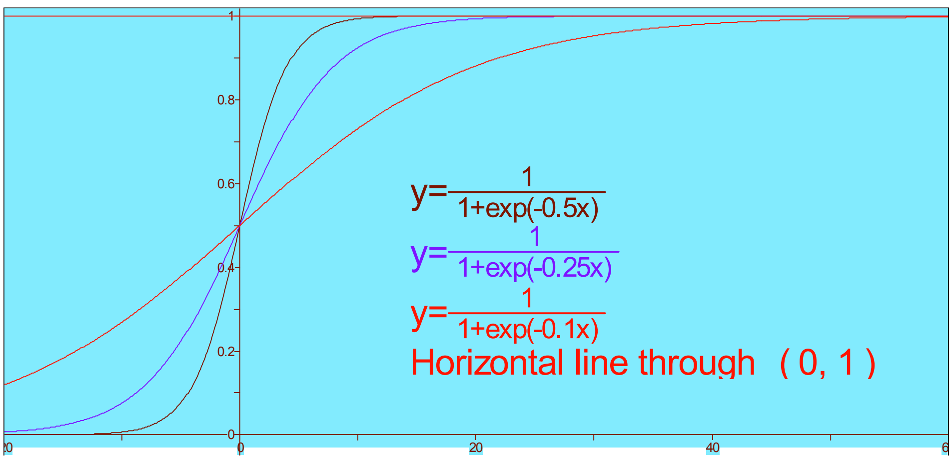
If we determine that $\alpha = 1$ and $\gamma = 1$ one will always find that

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

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



# Logistic curves for different $\beta$




$\beta$ determines how rapidly the curve grows

## MODEL (1)

Definitions:

- The probability that person no i shall have the value 1 on the variable $y_i$ 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 $y_i$ , here called $O_i$, is the ratio between two probabilities

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



## MODEL (2)

Definitions:

- The LOGIT , $L_i$ , for person no i (corresponding to $Pr(y_i=1)$) is the natural logarithm of the odds, $O_i$ , that person no i has the value 1 on variable $y_i$, is written:

  $L_i = \ln(O_i) = \ln\{p_i/(1-p_i)\}$
- 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

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

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

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



## MODEL (4)

In the model $Y_i = E[y_i \mid X_i] + \varepsilon_i$ the error is either

- $\varepsilon_i = 1 - E[y_i \mid X_i]$ with probability $E[y_i \mid X_i]$
  (since $\Pr(y_i = 1) = E[y_i \mid X_i]$ ),

or the error is

- $\varepsilon_i = - E[y_i \mid X_i]$ with probability $1 - E[y_i \mid X_i]$

- Meaning that the error has a distribution known as the binomial distribution with $p_i = E[y_i \mid X_i]$

# Estimation by the ML method

- 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

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

# Maximum Likelihood (2)

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

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

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

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



# Maximum Likelihood (3)

- The LogLikelihood $\mathcal{LL}$ will always be negative
- Maximizing $\mathcal{LL}$ is the same as minimizing the **positive LogLikelihood**; i.e. minimizing **-$\mathcal{LL}$**
- Finding parameter values that minimizes - $\mathcal{LL}$ can be done only by ”trial and error”, i.e. using an iterative procedure

# Iterative estimation

| From Hamilton Tabell 7.1 | Iteration | **-2 Log Likelihood** | Coefficients | |
|---|---|---|---|---|
| | | | Constant | lived |
| Initial | 0 | 209,212 | -,276 | |
| Step | 1 | 195,684 | ,376 | -,034 |
| | 2 | 195,269 | ,455 | -,041 |
| | 3 | 195,267 | ,460 | -,041 |
| | 4 | 195,267 | ,460 | -,041 |

Note the column titled *-2 LogLikelihood*



# Footnotes to the tables

- Step 0: Point of departure is a model with only a constant and no variables
- **Iterative estimation**
  - Estimation ends at iteration no 4 since the parameter estimates changed less than 0.001

For the next slide:

- The Wald statistic that SPSS provides equals the square of the "t" that Hamilton (and STATA) provides (Wald = $t^2$)

# Logistic model instead of LPM

## OLS regression (slide 6 above)

| 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 |

## Logistic regression

| 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 |



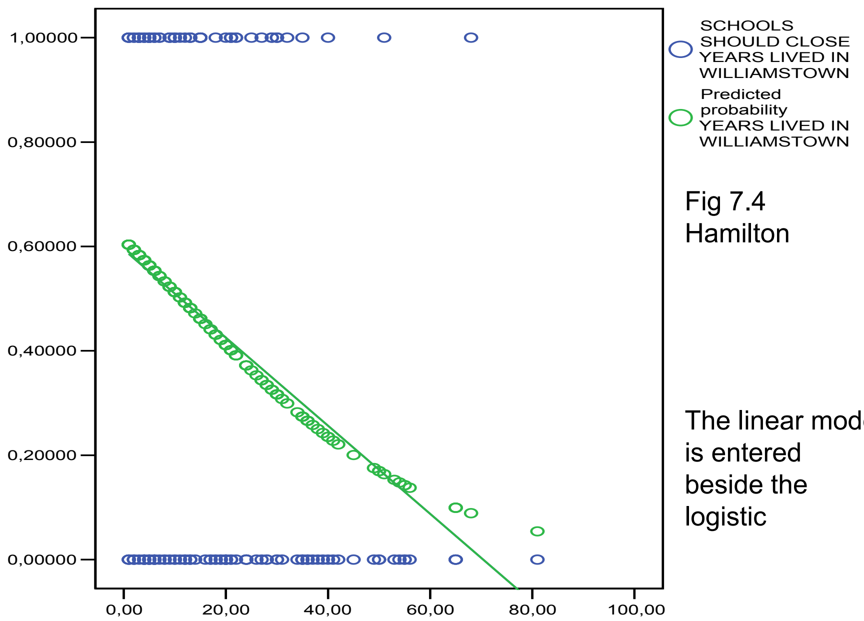



Fig 7.4
Hamilton

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 (e.g. comparing two NESTED models)
- (2) Wald test
  - The square root of this can be used analogous to the t-test but is normally distributed



# 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 $\beta_1$ 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 parameters
  - All relevant variables are 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 by SPSS there is
  - No perfect multicollinearity and
  - No perfect discrimination

## Assumptions that can be tested

- Model specification
  - logit is linear in the parameters
  - no irrelevant variables are included
- Sufficiently large sample
  - What is “sufficiently large” depends on the number of different patterns in the sample and how cases are distributed across these
- Testing implies an assessment of whether statistical problems leads to departure from the assumptions



## LOGISTIC REGRESSION
## Statistical problems may be due to

- Too small a sample
- High degree of **multicollinearity**
  - Leading to large standard errors (imprecise estimates)
  - Multicollinearity 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 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

| Y = Strength of water quality standards | Women without small children | Women with small children |
|---|---|---|
| Not weaker | 202 | 79 |
| Weaker OK | 44 | 0 |



# 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. |

# Logistic regression

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



# The LikeLihood Ratio test (1)

- The ratio between two Likelihoods equals the difference between two **LogLikelihoods**
- The difference between the **LogLikelihood** ($\mathcal{LL}$) of two **nested** models, estimated on **the same data**, can be used to test which of two models fits the data best, just like the F-statistic is used in OLS regression
- The test can also be used for singe regression coefficients (single variables). In small samples it has better properties than the Wald statistic

# The LikeLihood Ratio test (2)

The LikeLihood Ratio test statistic

- $\chi^2_H$ **= -2[$\mathcal{LL}$(model1) - $\mathcal{LL}$(model2)]**

will, if the null hypothesis of no difference between the two models is correct, be distributed approximately (for large n) as the chi-square distribution with number of degrees of freedom equal to the difference in number of parameters in the two models (H)



# Example of a Likelihood Ratio test

- Model 1: just constant
- Model 2: constant plus one variable

- $\chi^2_H$ **= -2[$\mathcal{LL}$(model1) - $\mathcal{LL}$(model2)]**
  **= -2$\mathcal{LL}$(model1) + 2$\mathcal{LL}$(model2)**
- Find the value of the ChiSquare and the number of degrees of freedom
- e.g.: LogLikelihood (mod1) = 209,212/(-2)
- LogLikelihood (mod2) = 195,267/(-2)

| From Tab 7.1: **-2 Log Likelihood** |
|---|
| 209,212 |
| 195,684 |
| 195,269 |
| 195,267 |
| 195,267 |

# The Wald test (1)

- The Wald (or chisquare) test statistic provided by SPSS = $t^2 = (b_k / SE(b_k))^2$ (where t is the normally distributed t used by Hamilton) can be used for testing single parameters similarly to the t-statistic of the OLS regression
- If the null hypothesis is correct, t will (for large n) in logistic regression be approximately normally distributed
- If the null hypothesis is correct, the Wald statistic will (for large n) in logistic regression be approximately chisquare distributed with df=1



## Excerpt from Hamilton Table 7.2

| **Iterasjon** | **-2 Log likelihood** | | | | | |
|---|---|---|---|---|---|---|
| 0 | 209,212 | | | | | |
| 1 | 152,534 | | | | | |
| 2 | 149,466 | | | | | |
| 3 | 149,382 | | | | | |
| 4 | 149,382 | | | | | |
| 5 | 149,382 | | | | | |
| | | | | | | |
| **Variables** | **B** | **S.E.** | **Wald** | **df** | **Sig.** | **Exp(B)** |
| Lived | -,046 | ,015 | 9,698 | 1 | ,002 | ,955 |
| Educ | -,166 | ,090 | 3,404 | 1 | ,065 | ,847 |
| Contam | 1,208 | ,465 | 6,739 | 1 | ,009 | 3,347 |
| Hsc | 2,173 | ,464 | 21,919 | 1 | ,000 | 8,784 |
| Constant | 1,731 | 1,302 | 1,768 | 1 | ,184 | 5,649 |

## Confidence interval for parameter estimates

- Can be constructed based on the fact that the square root of the Wald statistic approximately follows a normal distribution with 1 degree of freedom
- $b_k - t_\alpha * SE(b_k) < \beta_k < b_k + t_\alpha * SE(b_k)$
  where $t_\alpha$ is a value taken from the table of the **normal distribution** with level of significance equal to $\alpha$



## Can be constructed based on the t-distribution (1)

- If a table of the normal distribution is missing one may use the **t-distribution** since the t-distribution is approximately normally distributed for large n-K (e.g. for n-K > 120)

## Excerpt from Hamilton Table 7.3 (from SPSS)

| STATA<br>SPSS | | B | S.E. | $t^2$<br>Wald | df | Prob>t<br>Sig. | Exp(B) |
|---|---|---|---|---|---|---|---|
| Step 1 | lived | -,047 | ,017 | 7,550 | 1 | ,006 | ,954 |
| | educ | -,206 | ,093 | 4,887 | 1 | ,027 | ,814 |
| | contam | 1,282 | ,481 | 7,094 | 1 | ,008 | 3,604 |
| | hsc | 2,418 | ,510 | 22,508 | 1 | ,000 | 11,223 |
| | female | -,052 | ,557 | ,009 | 1 | ,926 | ,950 |
| | kids | -,671 | ,566 | 1,406 | 1 | ,236 | ,511 |
| | nodad | -2,226 | ,999 | 4,964 | 1 | ,026 | ,108 |
| | Constant | 2,894 | 1,603 | 3,259 | 1 | ,071 | 18,060 |



## More from Hamilton Table 7.3

| Iteration | | -2 Log likelihood | Coefficients | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Const | lived | educ | contam | hsc | female | kids | nodad |
| Step0 | | 209,212 | -0,276 | | | | | | | |
| Step1 | 1 | 147,028 | 1,565 | -,027 | -,130 | ,782 | 1,764 | -,015 | -,365 | -1,074 |
| | 2 | 141,482 | 2,538 | -,041 | -,187 | 1,147 | 2,239 | -,037 | -,580 | -1,844 |
| | 3 | 141,054 | 2,859 | -,046 | -,204 | 1,269 | 2,401 | -,050 | -,662 | -2,184 |
| | 4 | 141,049 | 2,893 | -,047 | -,206 | 1,282 | 2,418 | -,052 | -,671 | -2,225 |
| | 5 | 141,049 | 2,894 | -,047 | -,206 | 1,282 | 2,418 | -,052 | -,671 | -2,226 |

## Is the model in table 7.3 better than the model in table 7.2 ?

- **$\mathcal{LL}$(model in 7.3) = 141,049/(-2)**
- **$\mathcal{LL}$(model in 7.2) = 149,382/(-2)**

- $\chi^2_H$ = -2[$\mathcal{LL}$(model 7.2) - $\mathcal{LL}$(model 7.3)]
- Find $\chi^2_H$ value
- Find H
- Look up the table of the chisquare distribution



## The model of the probability of observing y=1 for person i

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

where the logit $L_i = \beta_0 + \sum_{j=1}^{K-1} \beta_j X_{ji}$ is a linear function of the explanatory variables

It is not easy to interpret the meaning of the $\beta$ coefficients just based on this formula

## The odds ratio

- The odds ratio, **O**, can be interpreted as the relative effect of having one variable value rather than another
- e.g. if $x_{ki}$ = t+1 in $L_i$' and $x_{ki}$ = t in $L_i$
- **O** = $O_i$ ($Y_i$=1| $L_i$')/ $O_i$ ($Y_i$=1| $L_i$)
  = exp[$L_i$' ]/ exp[$L_i$]
  = exp[$\beta_k$]
- Why $\beta_k$ ?



## The odds ratio : example I

- The Odds for answering yes =
  $e^{b_0+b_1*\text{Alder}+b_2*\text{Kvinne}+b_3*\text{E.utd}+b_4*\text{Barn i HH}}$
- The odds ratio for answering yes between women and men =

$$\frac{e^{b_0+b_1*Alder+b_2*1+b_3*E.utd+b_4*Barn\_i\_HH}}{e^{b_0+b_1*Alder+b_2*0+b_3*E.utd+b_4*Barn\_i\_HH}} = e^{b_2}$$

Remember the rules of power exponents

## The odds ratio : example II

- The Odds for answering yes given one year of extra education

$$\frac{e^{b_0+b_1*Alder+b_2*Kvinne+b_3*(E.utd+1)+b_4*Barn\_i\_HH}}{e^{b_0+b_1*Alder+b_2*Kvinne+b_3*E.utd+b_4*Barn\_i\_HH}}=e^{b_3}$$

Remember the rules of power exponents



## Example from Hamilton table 7.2

- What is the odds ratio for yes to closing the school from one year extra education?
- The odds ratio is the ratio of two odds where one odds is the odds for a person with one year extra education

$$\frac{e^{b_0+b_1*\text{ÅrBuddIByen}+b_2*(Utdanning+1)+b_3*UreiningEigEigedom+b_4*\text{MangeHSCmøter}}}{e^{b_0+b_1*\text{ÅrBuddIByen}+b_2*Utdanning+b_3*UreiningEigEigedom+b_4*\text{MangeHSCmøter}}}$$

$$=\frac{e^{b_2*(Utdanning+1)}}{e^{b_2*Utdanning}}=e^{b_2}$$

## Example from Hamilton table 7.2 cont.

- Odds ratio = Exp{$b_2$} = exp(-0,166) = 0,847
- One extra year of education implies that the odds is reduced with a factor of 0.847
- One may also say that the odds has increased with a factor of

  100(0,847-1)% = -15,3%
- Meaning that it has declined with 15,3%



## Concluding on logistic regression

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