

Logistic regression

Morten Frydenberg ©  
Department of Biostatistics, Aarhus Univ, Denmark

When one might use logistic regression.

Some examples:

- One **binary** independent variable. (**one odds ratio**).
- Probabilities, odds and the logit function
- One **continuous** independent variable.
- One **categorical** independent variable. (The **Wald test**)
- One **binary** independent variable and **continuous** independent variable no interaction.
- One **binary** independent variable and **continuous** independent variable with interaction.

Morten Frydenberg

Linear and Logistic regression - Note 4

1

Watch out for 'small' reference groups

The **likelihood ratio test**: comparing two nested models.

The **logistic regression model in general**

- The model and the **assumptions**.
- The **data** and the assumption of **independence**.

**Estimation and inference**

Morten Frydenberg

Linear and Logistic regression - Note 4

2

Logistic regression models: Introduction

A logistic regression is a **possible** model if the **dependent** variable (the response) is **dichotomous** dead/alive obese/not obese etc.

Contrary to what many believe there are **no assumptions** about the **independent** variables.

They can be categorical or continuous.

When working with binary response it is **custom** to **code** the "positive" event (eg. dead) as **1** and a "negative" event (alive) as **0**.

A logistic regression models the **probability** of a "positive event" via odds.

And the associations via **odds ratio**.

If the **event is rare** then **odds ratios** estimate the **relative risk**.

Morten Frydenberg

Linear and Logistic regression - Note 4

3

Logistic regression models: Introduction

A logistic regression can also be used to estimate the odds ratios in a **unmatched case-control** study.

For such data the **constant terms** have **no meaning**.

And the odds ratios comparable odds ratio from a **follow-up study**.

Many **other epidemiological design** are analyzed by logistic regression models.

Morten Frydenberg

Linear and Logistic regression - Note 4

4

Estimating one odds ratio using logistic regression

We are now considering a larger part of the Frammingham data set, consisting of 4690 person with **known BMI** at the start.

We will focus on the risk obesity ( $BMI \geq 30$  kg/m<sup>2</sup>).

Out of the 4690 persons 601 = 12.8% were **obese**.

Divided into gender

|       | Obese       | Not-Obese |
|-------|-------------|-----------|
| Women | 375 (14.2%) | 2268      |
| Men   | 226 (11.0%) | 1821      |

We see a higher prevalence among women: OR: **1.33 (1.12;1.59)**.

That is **the odds** of being obese is between **12** and **59** percent higher for women. ( $\chi^2=10.2$  p-value=0.001)

Morten Frydenberg

Linear and Logistic regression - Note 4

5

Finding an odds ratio using logistic regression

The odds ratio is defined as:  $OR = \frac{odds_{Women}}{odds_{Men}}$

So applying the logarithm we get:

$$\ln(OR) = \ln\left(\frac{odds_{Women}}{odds_{Men}}\right) = \ln(odds_{Women}) - \ln(odds_{Men})$$

And rearranging terms :

$$\ln(odds_{Women}) = \ln(odds_{Men}) + \ln(OR)$$

That is the log-odds obesity for the women can be written as the sum of two terms:

- The log-odds in **reference** group (men)
- The log of the odds ratio

Morten Frydenberg

Linear and Logistic regression - Note 4

6

Finding an odds ratio using logistic regresion

$$\ln(odds_{Women}) = \ln(odds_{Men}) + \ln(OR)$$

If we again let *women* be a indicator/dummy variable, then we can consider the model:

$$\ln(odds) = \beta_0 + \beta_1 \cdot woman$$

For **men** we get:  $\ln(odds) = \beta_0$

And for **women**:  $\ln(odds) = \beta_0 + \beta_1$

Comparing with the equation on top we get:

$$\beta_0 = \ln(odds_{Men})$$

and

$$\beta_1 = \ln(OR)$$

Morten FrydenbergLinear and Logistic regression - Note 47

Finding an odds ratio using logistic regresion

$$\ln(odds) = \beta_0 + \beta_1 \cdot woman$$

$\ln(odds_{Men})$  $\ln(OR)$

Or to be more precise:  $\beta_1 = \ln(OR_{Women\ vs\ Men})$

So, if we can fit the model above to the data, then we can get an estimate of the  $\ln(OR)$  and hence of *OR*!

Morten FrydenbergLinear and Logistic regression - Note 48

Probabilities and odds

If *p* denote the probability of an event (the **risk**, the **prevalence** proportion, or **cumulated incidence** proportion) then the odds is given by :

$$odds = \frac{p}{1-p}$$

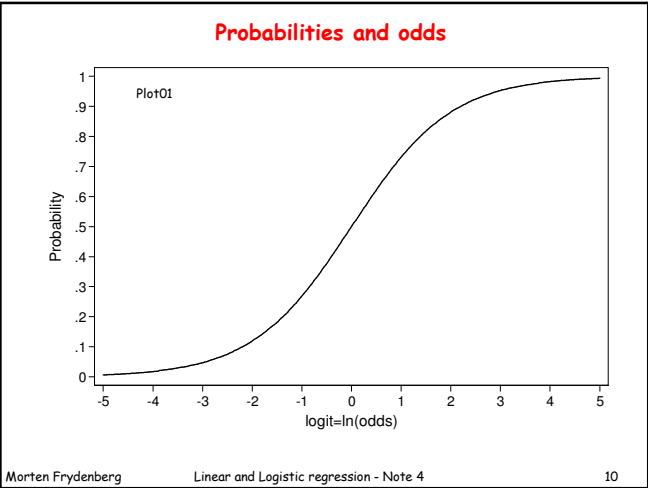
Note:  $odds=1 \Leftrightarrow p=0.5 \Leftrightarrow \ln(odds)=0$

$$\ln(odds) = \ln\left(\frac{p}{1-p}\right)$$

In mathematics the last function of *p* is called the "logit" function.

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right)$$

Morten FrydenbergLinear and Logistic regression - Note 49



Probabilities and odds

$$\ln(odds) = \beta_0 + \beta_1 \cdot woman$$

So modelling the **log-odds** is the same as modelling  $\text{logit}(p)$  and model from before could be written.

$$\text{logit}(p) = \beta_0 + \beta_1 \cdot woman$$

Going from odds to probabilities:  $p = \frac{odds}{1+odds}$

The model on **probability scale** is :

$$p = \frac{\exp(\beta_0 + \beta_1 \cdot woman)}{1 + \exp(\beta_0 + \beta_1 \cdot woman)}$$

Morten FrydenbergLinear and Logistic regression - Note 411

Finding an odds ratio using logistic regresion

$$\text{logit}(p) = \ln(odds) = \beta_0 + \beta_1 \cdot woman$$

Back to finding the estimates.

In Stata:

```
char sex[omit]1
xi: logit obese i.sex
```

i.sex

\_Isex\_1-2

(naturally coded; \_Isex\_1 omitted)

Iteration 0: log likelihood = -1795.5437

Iteration 3: log likelihood = -1790.3703

Logit estimates

Number of obs = 4690

LR chi2(1) = 10.35

Prob > chi2 = 0.0013

Pseudo R2 = 0.0029

Log likelihood = -1790.3703

|         | obese | Coef.     | Std. Err. | z      | P> z  | [95% Conf. Interval] |
|---------|-------|-----------|-----------|--------|-------|----------------------|
| _Isex_2 |       | .2868784  | .0898972  | 3.19   | 0.001 | .1106831 .4630738    |
| _cons   |       | -2.086606 | .070526   | -29.59 | 0.000 | -2.224835 -1.948378  |

Morten FrydenbergLinear and Logistic regression - Note 412

Finding an odds ratio using logistic regression

$$\text{logit}(p) = \ln(\text{odds}) = \beta_0 + \beta_1 \cdot \text{woman}$$

$$\hat{\beta}_1 = \ln(\widehat{OR})$$

$$95\% \text{ CI for } \ln(OR)$$

|         | Coef.     | Std. Err. | z      | P> z  | [95% Conf. Interval] |
|---------|-----------|-----------|--------|-------|----------------------|
| obese   |           |           |        |       |                      |
| _Isex_2 | .2868784  | .0898972  | 3.19   | 0.001 | .1106831 .4630738    |
| _cons   | -2.086606 | .070526   | -29.59 | 0.000 | -2.224835 -1.948378  |

$$\widehat{OR} = \exp(0.2868784) = 1.33$$

$$95\% \text{ CI: } (1.12; 1.59).$$

Test for the hypothesis :  $\ln(OR)=0 \Leftrightarrow OR=1$

Odds in reference group (men) =  $\exp(-2.086606)=0.1241$

$$95\% \text{ CI: } (0.1081; 0.1425).$$

Prevalence among men: 0.1104 (0.0975;0.1247).

Morten Frydenberg      Linear and Logistic regression - Note 4      13

Finding an odds ratio using logistic regression

$$\text{logit}(p) = \ln(\text{odds}) = \beta_0 + \beta_1 \cdot \text{woman}$$

An easier way to obtain the odds ratio.

`xi: logit obese i.sex, or`

```
i.sex      _Isex_1-2      (naturally coded; _Isex_1 omitted)
Iteration 0:  log likelihood = -1795.5437
Iteration 3:  log likelihood = -1790.3703
Logit estimates
Number of obs      =    4690
LR chi2(1)         =    10.35
Prob > chi2        =    0.0013
Pseudo R2          =    0.0029
```

Log likelihood = -1790.3703

|         | Odds Ratio | Std. Err. | z    | P> z  | [95% Conf. Interval] |
|---------|------------|-----------|------|-------|----------------------|
| obese   |            |           |      |       |                      |
| _Isex_2 | 1.332262   | .1197667  | 3.19 | 0.001 | 1.117041 1.588951    |

Note, we cannot find any information about the risk in the reference group , i.e. the odds and prevalence among men!

Morten Frydenberg      Linear and Logistic regression - Note 4      14

The obesity and age: version 1

In the previous section we saw that the prevalence of obesity was different between men and women.

Is it also associated with age?

The simplest model on the logit scale would be:

$$\text{logit}(p) = \ln(\text{odds}) = \beta_0 + \beta_1 \cdot \text{age}$$

That is a linear relation on the log-odds scale.

As we have seen before using age implies that  $\beta_0$  references to a newborn (age=0).

So we will chose age=45 reference instead:

$$\text{logit}(p) = \ln(\text{odds}) = \beta_0 + \beta_1 \cdot (\text{age} - 45)$$

Morten Frydenberg      Linear and Logistic regression - Note 4      15

The obesity and age: version 1

$$\text{logit}(p) = \ln(\text{odds}) = \beta_0 + \beta_1 \cdot (\text{age} - 45)$$

The interpretation of the parameters:

$\beta_0$  : the log odds for 45 year old person.

$\beta_1$  : the log odds ratio, when comparing two persons who differ 1 year in age.

$\exp(\beta_1)$  : the odds ratio, when comparing two persons who differ 1 year in age.

Note, that this odds ratio is assumed to be the same no matter what age the two persons have, as long as they differ by one year!

The log odds ratio is proportional to the age differences, e.g. OR increases exponentially with the age differences.

Morten Frydenberg      Linear and Logistic regression - Note 4      16

The obesity and age: version 1

$$\text{logit}(p) = \ln(\text{odds}) = \beta_0 + \beta_1 \cdot (\text{age} - 45)$$

Obtaining the estimates in Stata:

`generate age45=age-45`

`logit obese age45`

|       | Coef.     | Std. Err. | z      | P> z  | [95% Conf. Interval] |
|-------|-----------|-----------|--------|-------|----------------------|
| obese |           |           |        |       |                      |
| age45 | .0348023  | .0051296  | 6.78   | 0.000 | .0247484 .0448561    |
| _cons | -1.985922 | .0463594  | -42.84 | 0.000 | -2.076785 -1.895059  |

Test for no association with age

`logit obese age45, or`

|       | Odds Ratio | Std. Err. | z    | P> z  | [95% Conf. Interval] |
|-------|------------|-----------|------|-------|----------------------|
| obese |            |           |      |       |                      |
| age45 | 1.035415   | .0053113  | 6.78 | 0.000 | 1.025057 1.045877    |

Morten Frydenberg      Linear and Logistic regression - Note 4      17

The obesity and age: version 1

$$\text{logit}(p) = \ln(\text{odds}) = \beta_0 + \beta_1 \cdot (\text{age} - 45)$$

Estimate:  $\beta_0$  : -1.985 (-2.0767;-1.8951)

The odds for obesity for among 45 year old:

$$0.1373 \text{ (0.1253;0.1503)}$$

The prevalence of obesity for among 45 year old:

$$0.1207 \text{ (0.1114;0.1307)}$$

$$\text{odds} = \exp(\log(\text{odds}))$$

$$\text{Prob} = \frac{\text{odds}}{1 + \text{odds}}$$

Morten Frydenberg      Linear and Logistic regression - Note 4      18

The obesity and age: version 1

$$\text{logit}(p) = \ln(\text{odds}) = \beta_0 + \beta_1 \cdot (\text{age} - 45)$$

Estimates:  $\beta_1 : 0.0348 (0.0247; 0.0449)$

The **odds ratio** for being obese is **1.0354 (1.0251; 1.0459)** when comparing the old person to the young person, if they differ with **one year in age**.

If they differ with **4.5 years** then the odds ratio is **1.0354<sup>4.5</sup> (1.0251<sup>4.5</sup>; 1.0459<sup>4.5</sup>) = 1.17 (1.12; 1.22)**

In Stata: `lincom age45*4.5, OR`

```
( 1) 4.5 age45 = 0
```

|     | obese | Odds Ratio | Std. Err.            | z    | P> z  | [95% Conf. Interval] |
|-----|-------|------------|----------------------|------|-------|----------------------|
| (1) |       | 1.16954    | <del>0.0269568</del> | 6.78 | 0.000 | 1.117806 1.223668    |

Morten Frydenberg

Linear and Logistic regression - Note 4

19

The obesity and age: version 1

Estimated relationship:  $\ln(\text{odds}) = -1.986 + 0.0348 \cdot (\text{age} - 45)$

Morten Frydenberg

Linear and Logistic regression - Note 4

20

The obesity and age: version 1

Estimated relationship:

$$\text{prevalence} = \frac{\exp(-1.986 + 0.0348 \cdot (\text{age} - 45))}{1 + \exp(-1.986 + 0.0348 \cdot (\text{age} - 45))}$$

Morten Frydenberg

Linear and Logistic regression - Note 4

21

The obesity and age: version 2

$$\ln(\text{odds}) = \beta_0 + \beta_1 \cdot (\text{age} - 45)$$

This model assumes that one year of age difference is associated with the same odds ratio irrespective of the age.

An other way to model the prevalence could be to assume a step function that is to categorize age.

We will here look at age divided in seven five-years groups:  
`egen agegrp7=cut(age), at(0,35,40,45,50,55,60,120) label 1`

With this command the **youngest** age group will be number **0** the **second youngest**: **1** and the **oldest**: **6**

Morten Frydenberg

Linear and Logistic regression - Note 4

22

The obesity and age: version 2

```
table agegrp7, c(min age max age count obese sum obese) row
```

| agegrp7 | min(age) | max(age) | N(obese) | sum(obese) |
|---------|----------|----------|----------|------------|
| 0-      | 30       | 34       | 352      | 23         |
| 35-     | 35       | 39       | 973      | 105        |
| 40-     | 40       | 44       | 885      | 93         |
| 45-     | 45       | 49       | 799      | 95         |
| 50-     | 50       | 54       | 733      | 115        |
| 55-     | 55       | 59       | 613      | 95         |
| 60-     | 60       | 66       | 335      | 75         |
| Total   | 30       | 66       | 4,690    | 601        |

A model that have different odds in each age group :

$$\ln(\text{odds}) = \alpha_0 + \sum_{i=1}^6 \alpha_i \cdot \text{age}_i$$

Where *age<sub>i</sub>* is an indicator for being in the *i*th age group

Morten Frydenberg

Linear and Logistic regression - Note 4

23

The obesity and age: version 2

$$\ln(\text{odds}) = \alpha_0 + \sum_{i=1}^6 \alpha_i \cdot \text{age}_i$$

The interpretation of the parameters:

$\alpha_0$  : the **log odds** in **reference** group=the youngest.

$\alpha_i$  : the **log odds ratio**, when comparing one person in age group *i* with one in the reference group=the youngest.

```
char agegrp7[omit]0
xi: logit obese i.agegrp7
```

Not all output

|             | obese | Coef.    | Std. Err. | z      | P> z  | [95% Conf. Interval] |
|-------------|-------|----------|-----------|--------|-------|----------------------|
| _Iagegrp7_1 |       | .54833   | .23915    | 2.29   | 0.022 | .079603 1.017061     |
| _Iagegrp7_2 |       | .51860   | .24193    | 2.14   | 0.032 | .0444155 .992787     |
| _Iagegrp7_3 |       | .65766   | .24179    | 2.72   | 0.007 | .1837537 1.13157     |
| _Iagegrp7_4 |       | .97900   | .23839    | 4.11   | 0.000 | .5117642 1.44625     |
| _Iagegrp7_5 |       | .96446   | .24284    | 3.97   | 0.000 | .4884941 1.440436    |
| _Iagegrp7_6 |       | 1.41737  | .25238    | 5.62   | 0.000 | .9227081 1.912032    |
| _cons       |       | -2.66056 | .21567    | -12.34 | 0.000 | -3.083288 -2.237839  |

Morten Frydenberg

Linear and Logistic regression - Note 4

24

The obesity and age: version 2

$$\ln(odds) = \alpha_0 + \sum_{i=1}^6 \beta_i \cdot age_i$$

xi: logit obese i.agegrp7,or

Not all output

|             | obese | Odds Ratio | Std. Err. | z    | P> z  | [95% Conf. Interval] |
|-------------|-------|------------|-----------|------|-------|----------------------|
| _Iagegrp7_1 |       | 1.730365   | .4138201  | 2.29 | 0.022 | 1.082857 2.765057    |
| _Iagegrp7_2 |       | 1.679677   | .4063746  | 2.14 | 0.032 | 1.045417 2.698747    |
| _Iagegrp7_3 |       | 1.930274   | .4667295  | 2.72 | 0.007 | 1.20172 3.100522     |
| _Iagegrp7_4 |       | 2.661812   | .634592   | 4.11 | 0.000 | 1.668232 4.247159    |
| _Iagegrp7_5 |       | 2.623384   | .6370806  | 3.97 | 0.000 | 1.62986 4.222538     |
| _Iagegrp7_6 |       | 4.126254   | 1.041397  | 5.62 | 0.000 | 2.516095 6.766825    |

The OR between the **second oldest** and the **youngest**:  
2.62 (1.63;4.22)

Between a **63** and **322** percent **increase** in odds.

Small prevalence: **63** and **322** percent **increase** in prevalence.

A statistical significant difference in prevalence!

Morten Frydenberg

Linear and Logistic regression - Note 4

25

The obesity and age: version 2

$$\ln(odds) = \alpha_0 + \sum_{i=1}^6 \alpha_i \cdot age_i$$

The output contains **six tests** of no difference in risk - comparing each of the six groups with the **reference** (the youngest) group.

The command: `testparm _Iagegrp*` will give a **"Wald test"** of no difference between the **seven** groups .

( 1) \_Iagegrp7\_1 = 0  
( 2) \_Iagegrp7\_2 = 0  
( 3) \_Iagegrp7\_3 = 0  
( 4) \_Iagegrp7\_4 = 0  
( 5) \_Iagegrp7\_5 = 0  
( 6) \_Iagegrp7\_6 = 0

chi2( 6) = 55.26  
Prob > chi2 = 0.0000

Highly significant differences

Morten Frydenberg

Linear and Logistic regression - Note 4

26

The obesity and age: version 2

Using the age group 45-49 as **reference**

char agegrp7[omit]3

xi: logit obese i.agegrp7,or

Not all output

|             | obese | Odds Ratio | Std. Err. | z     | P> z  | [95% Conf. Interval] |
|-------------|-------|------------|-----------|-------|-------|----------------------|
| _Iagegrp7_0 |       | .518061    | .1252643  | -2.72 | 0.007 | .3225264 .8321407    |
| _Iagegrp7_1 |       | .896434    | .1348312  | -0.73 | 0.467 | .6675609 1.203778    |
| _Iagegrp7_2 |       | .870175    | .1347005  | -0.90 | 0.369 | .6424561 1.17861     |
| _Iagegrp7_4 |       | 1.378981   | .2057436  | 2.15  | 0.031 | 1.029341 1.847385    |
| _Iagegrp7_5 |       | 1.359073   | .2123007  | 1.96  | 0.050 | 1.000625 1.845927    |
| _Iagegrp7_6 |       | 2.137652   | .3648206  | 4.45  | 0.000 | 1.529915 2.986803    |

The OR between the **second oldest** and the **45-49 old**:  
1.36 (1.00;1.85)

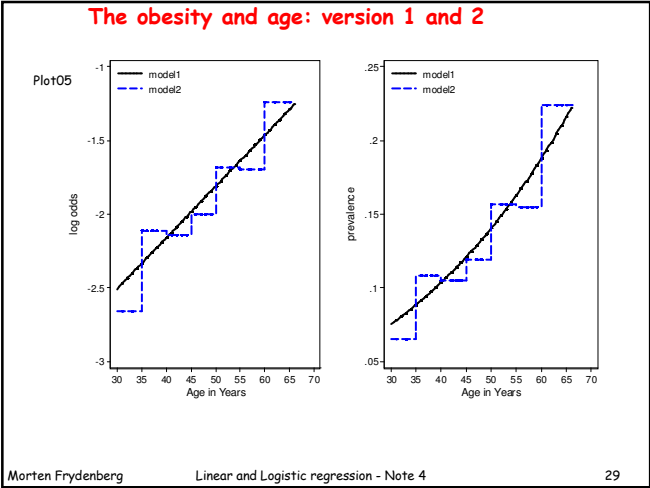
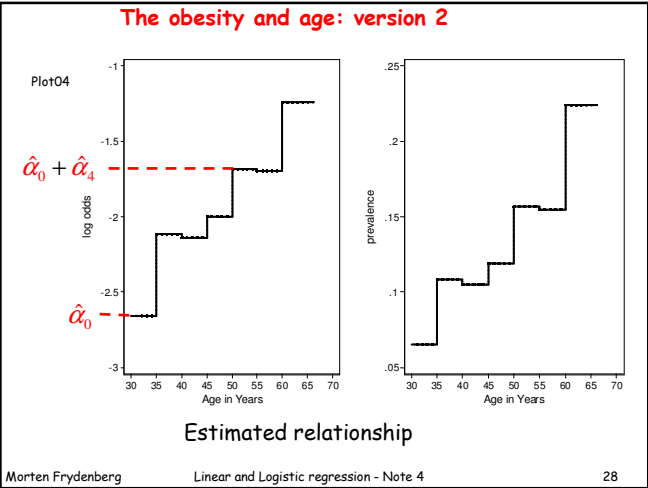
Between a **no** and **85** percent **increase** in (odds) prevalence.

A borderline significant different in prevalence!

Morten Frydenberg

Linear and Logistic regression - Note 4

27



The obesity, sex and age: version 1

The first analysis only looked at sex and the second only at age.

Let us try to look at those two at the same time

The simplest model **on the logit scale** would be:

$$\ln(odds) = \beta_0 + \beta_1 \cdot woman + \beta_2 \cdot (age - 45)$$

This is based on three **assumptions**:

**Additivity on logit scale:** The contribution from sex and age are **added**.

**Proportionality on logit scale:** The contribution from age is **proportional** to it is value.

**No effectmodification on logit scale:** The contribution from one independent variable is **the same** whatever the value is for the other.

Morten Frydenberg

Linear and Logistic regression - Note 4

30

The obesity, sex and age : version 1

$\ln(odds) = \beta_0 + \beta_1 \cdot woman + \beta_2 \cdot (age - 45)$

The interpretation of the parameters:

$\beta_0$  : the log odds for 45 year old man.

$\beta_1$  : the log odds ratio, when comparing a woman to a man of the same age.

$\beta_2$  : the log odds ratio, when comparing two persons of the same sex, where the first is one year older than the other.

$\beta_2 \cdot \Delta age$ : the log odds ratio, when comparing two persons of the same sex, where the first is  $\Delta age$  years older than the other.

Morten FrydenbergLinear and Logistic regression - Note 431

The obesity, sex and age : version 1

$\ln(odds) = \beta_0 + \beta_1 \cdot woman + \beta_2 \cdot (age - 45)$

Obtaining the estimates in Stata:

`xi:logit obese i.sex age45`

i.sex

\_Isex\_1-2

(naturally coded; \_Isex\_1 omitted)

Iteration 0: log likelihood = -1795.5437

Iteration 3: log likelihood = -1767.7019

Logit estimates

Number of obs = 4690

LR chi2(2) = 55.68

Prob > chi2 = 0.0000

Pseudo R2 = 0.0155

Log Likelihood = -1767.7019

|         | obese | Coef.     | Std. Err. | z      | P> z  | [95% Conf. Interval] |
|---------|-------|-----------|-----------|--------|-------|----------------------|
| _Isex_2 |       | .2743977  | .0903385  | 3.04   | 0.002 | .0973375 .451458     |
| age45   |       | .0344723  | .0051354  | 6.71   | 0.000 | .0244072 .0445374    |
| _cons   |       | -2.147056 | .0721981  | -29.74 | 0.000 | -2.288561 -2.00555   |

Tests: 

No association with sex

No association with age

Prevalence is 50% among 45 year old men

Morten FrydenbergLinear and Logistic regression - Note 432

The obesity, sex and age : version 1

$\ln(odds) = \beta_0 + \beta_1 \cdot woman + \beta_2 \cdot (age - 45)$

`xi:logit obese i.sex age45, or`

|         | obese | Odds Ratio | Std. Err. | z    | P> z  | [95% Conf. Interval] |
|---------|-------|------------|-----------|------|-------|----------------------|
| _Isex_2 |       | 1.315738   | .1188618  | 3.04 | 0.002 | 1.102232 1.5706      |
| age45   |       | 1.035073   | .0053155  | 6.71 | 0.000 | 1.024707 1.045544    |

OR for women compared to men "adjusted for age" :  
1.32 (1.10;1.57)

The unadjusted was  
1.33 (1.12;1.59).

OR for one year age difference "adjusted for sex" :  
1.04 (1.02;1.05)

The unadjusted was  
1.04 (1.03;1.05)

Not much has changed!

Morten FrydenbergLinear and Logistic regression - Note 433

The obesity, sex and age : version 1

$\ln(odds) = \beta_0 + \beta_1 \cdot woman + \beta_2 \cdot (age - 45)$

Plot06

The estimated relationship

Morten FrydenbergLinear and Logistic regression - Note 434

The obesity, sex and age: version 2

A more complicated model on the logit scale would be:

men:  $\ln(odds) = \alpha_0 + \alpha_1 \cdot (age - 45)$

women:  $\ln(odds) = \gamma_0 + \gamma_1 \cdot (age - 45)$

This is based on one assumptions:

Proportionality on logit scale: The contribution age is proportional to it is value.

It can be written in just one formula (with interaction):

$\ln(odds) = \beta_0 + \beta_1 \cdot woman + \beta_2 \cdot (age - 45) + \beta_3 \cdot woman \cdot (age - 45)$

Where:  $\alpha_0 = \beta_0$   $\alpha_1 = \beta_2$   
 $\gamma_0 = \beta_0 + \beta_1$   $\gamma_1 = \beta_2 + \beta_3$

That is:  $\beta_1 = \gamma_0 - \alpha_0$   $\beta_3 = \gamma_1 - \alpha_1$

Morten FrydenbergLinear and Logistic regression - Note 435

The obesity, sex and age: version 2

$\ln(odds) = \beta_0 + \beta_1 \cdot woman + \beta_2 \cdot (age - 45) + \beta_3 \cdot woman \cdot (age - 45)$

Estimates log odds:

`xi: logit obese i.sex*age45`

|              | obese | Coef.     | Std. Err. | z      | P> z  | [95% Conf. Interval] |
|--------------|-------|-----------|-----------|--------|-------|----------------------|
| _Isex_2      |       | -.116797  | .095034   | 1.23   | 0.219 | -.0694672 -.303061   |
| age45        |       | -.0056849 | .008372   | -0.68  | 0.497 | -.022095 -.010725    |
| _IsexXage4-2 |       | -.065803  | .01074    | 6.13   | 0.000 | -.044747 -.0868588   |
| _cons        |       | -2.083041 | .070643   | -29.49 | 0.000 | -2.22149 -1.944583   |

Men 

Difference between women and men

Estimates odds ratios:

|              | obese | Odds Ratio | Std. Err | z     | P> z  | [95% Conf. Interval] |
|--------------|-------|------------|----------|-------|-------|----------------------|
| _Isex_2      |       | 1.123891   | .10660   | 1.23  | 0.219 | .9328908 1.353997    |
| age45        |       | .994331    | .00882   | -0.68 | 0.497 | .978147 1.010783     |
| _IsexXage4-2 |       | 1.068016   | .01147   | 6.13  | 0.000 | 1.045763 1.090743    |

Morten FrydenbergLinear and Logistic regression - Note 436

Linear and Logistic Regression: Note 4

6



**A small case-control example**

tabodds cancer age

| age   | cases | controls | odds    | [95% Conf. Interval] |         |  |
|-------|-------|----------|---------|----------------------|---------|--|
| 25-34 | 2     | 116      | 0.01724 | 0.00426              | 0.06976 |  |
| 35-44 | 9     | 190      | 0.04737 | 0.02427              | 0.09244 |  |
| 45-54 | 46    | 167      | 0.27545 | 0.19875              | 0.38175 |  |
| 55-64 | 76    | 166      | 0.45783 | 0.34899              | 0.60061 |  |
| 65-74 | 55    | 106      | 0.51887 | 0.37463              | 0.71864 |  |
| >=75  | 13    | 31       | 0.41935 | 0.21944              | 0.80138 |  |

Few events in reference group= wide CI's

tabodds cancer age, or

| age   | Odds Ratio | chi2  | P>chi2 | [95% Conf. Interval] |            |
|-------|------------|-------|--------|----------------------|------------|
| 25-34 | 1.000000   | .     |        |                      |            |
| 35-44 | 2.747368   | 1.76  | 0.1843 | 0.579474             | 13.025660  |
| 45-54 | 15.976048  | 24.18 | 0.0000 | 3.588609             | 71.123412  |
| 55-64 | 26.554217  | 41.14 | 0.0000 | 5.834718             | 120.850133 |
| 65-74 | 30.094340  | 43.99 | 0.0000 | 6.278745             | 144.243682 |
| >=75  | 24.322581  | 29.40 | 0.0000 | 4.402342             | 134.380270 |

Morten Frydenberg      Linear and Logistic regression - Note 4      38

**A small case-control example**

tabodds cancer age

| age   | cases | controls | odds    | [95% Conf. Interval] |         |  |
|-------|-------|----------|---------|----------------------|---------|--|
| 25-34 | 2     | 116      | 0.01724 | 0.00426              | 0.06976 |  |
| 35-44 | 9     | 190      | 0.04737 | 0.02427              | 0.09244 |  |
| 45-54 | 46    | 167      | 0.27545 | 0.19875              | 0.38175 |  |
| 55-64 | 76    | 166      | 0.45783 | 0.34899              | 0.60061 |  |
| 65-74 | 55    | 106      | 0.51887 | 0.37463              | 0.71864 |  |
| >=75  | 13    | 31       | 0.41935 | 0.21944              | 0.80138 |  |

Many events in reference group= narrow CI's

tabodds cancer age, or **base(3)**

| age   | Odds Ratio | chi2  | P>chi2 | [95% Conf. Interval] |          |
|-------|------------|-------|--------|----------------------|----------|
| 25-34 | 0.062594   | 24.18 | 0.0000 | 0.014060             | 0.278660 |
| 35-44 | 0.171968   | 25.86 | 0.0000 | 0.079661             | 0.371235 |
| 45-54 | 1.000000   | .     |        |                      |          |
| 55-64 | 1.662127   | 5.54  | 0.0186 | 1.083844             | 2.548952 |
| 65-74 | 1.883716   | 7.32  | 0.0068 | 1.181689             | 3.002809 |
| >=75  | 1.522440   | 1.30  | 0.2546 | 0.734799             | 3.154365 |

Morten Frydenberg      Linear and Logistic regression - Note 4      39

**A small case-control example**

```
char age [omit]1
xi:logit cancer i.smoker i.age,or
i.smoker      _smoker_0-1      (naturally coded; _smoker_0 omitted)
i.age         _age_1-6         (naturally coded; _age_1 omitted)
```

Iteration 0: log likelihood = -496.55682  
Iteration 1: log likelihood = -437.55133  
Iteration 2: log likelihood = -429.86007  
Iteration 3: log likelihood = -428.99383  
Iteration 4: log likelihood = -428.94473  
Iteration 5: log likelihood = -428.94432  
Iteration 6: log likelihood = -428.94432

Logit estimates

Number of obs = 977  
LR chi2(6) = 135.23  
Prob > chi2 = 0.0000  
Pseudo R2 = 0.1362

Log likelihood = -428.94432

| cancer    | Odds Ratio | Std. Err. | z    | P> z  | [95% Conf. Interval] |          |
|-----------|------------|-----------|------|-------|----------------------|----------|
| _smoker_1 | 2.350      | .4513038  | 4.45 | 0.000 | 1.613342             | 3.424472 |
| _age_2    | 2.832      | 2.24368   | 1.31 | 0.189 | .5995103             | 13.3798  |
| _age_3    | 16.58      | 12.17378  | 3.82 | 0.000 | 3.932286             | 69.91422 |
| _age_4    | 27.89      | 20.32374  | 4.57 | 0.000 | 6.691356             | 116.3235 |
| _age_5    | 34.79      | 25.59029  | 4.83 | 0.000 | 8.231516             | 147.0764 |
| _age_6    | 27.71      | 21.89267  | 4.21 | 0.000 | 5.891878             | 130.3509 |

Many iterations

Morten Frydenberg      Linear and Logistic regression - Note 4      40

**A small case-control example**

```
char age [omit]3
xi:logit cancer i.smoker i.age,or
i.smoker      _smoker_0-1      (naturally coded; _smoker_0 omitted)
i.age         _age_1-6         (naturally coded; _age_3 omitted)
```

Iteration 0: log likelihood = -496.55682  
Iteration 1: log likelihood = -437.55133  
Iteration 2: log likelihood = -429.86007  
Iteration 3: log likelihood = -428.99383  
Iteration 4: log likelihood = -428.94473  
Iteration 5: log likelihood = -428.94432

Logit estimates

Number of obs = 977  
LR chi2(6) = 135.23  
Prob > chi2 = 0.0000  
Pseudo R2 = 0.1362

Log likelihood = -428.94432

| cancer    | Odds Ratio | Std. Err. | z     | P> z  | [95% Conf. Interval] |          |
|-----------|------------|-----------|-------|-------|----------------------|----------|
| _smoker_1 | 2.3504     | .451303   | 4.45  | 0.000 | 1.613343             | 3.424469 |
| _age_1    | .0603      | .0442767  | -3.83 | 0.000 | .0143051             | .2542718 |
| _age_2    | .1708      | .0652397  | -4.63 | 0.000 | .0807999             | .3610977 |
| _age_4    | 1.6826     | .3701188  | 2.37  | 0.018 | 1.093327             | 2.58953  |
| _age_5    | 2.0984     | .5042862  | 3.08  | 0.002 | 1.31025              | 3.360918 |
| _age_6    | 1.6713     | .6277714  | 1.37  | 0.171 | .8005146             | 3.489699 |

Morten Frydenberg      Linear and Logistic regression - Note 4      41

**Things to look out for in the output**

In general:

**Wide CI's or large standard errors** in a logistic regression indicates that at least one group has **few events**!

**Many iterations** in a logistic regression indicates that some of the **parameters are hard to estimate**.

Morten Frydenberg      Linear and Logistic regression - Note 4      42

### Comparing two models: the likelihood ratio test

Earlier we saw how one could use a **Wald** to test if several coefficients could be zero.

An other way to "compare" two models is by a **likelihood ratio test**.

In the logistic regression output from Stata we find a likelihood ratio test comparing the **fitted model** with the model with no dependent variables the **constant odds model**:

```
LR chi2(6)      =    135.23
Prob > chi2     =    0.0000
```

**The conclusion:** The model with smoker and age is **statistical significant** better, than a model assuming the same odds, risk for everybody.

Morten Frydenberg

Linear and Logistic regression - Note 4

43

### Comparing two models: the likelihood ratio test

One can compare two models with a likelihood ratio test if:

- The two models are fitted on exactly the **same data set**.
- The two models are **nested**, i.e. one can go from one model to the other by setting some coefficients to zero.

In Stata the test is found in this way:

```
xi:logit cancer i.smoker i.age
estimates store model1
xi:logit cancer i.smoker
estimates store model2
lrtest model1 model2
```

```
Output:
likelihood-ratio test      LR chi2(5) =    120.82
(Assumption: model2 nested in model1)  Prob > chi2 =    0.0000
```

i.age adds **statistical significant** information to the model only containing smoking!

Morten Frydenberg

Linear and Logistic regression - Note 4

44

### Logistic regression model in general

$$\ln(\text{odds}) = \beta_0 + \sum_{p=1}^k \beta_p \cdot x_p$$

This is based on three assumptions:

- Additivity on log-odds scale:** The contribution from each of the independent variables are **added**.
- Proportionality:** The contribution from independent variables is **proportional** to it is value (with a factor  $\beta$ )
- No effectmodification:** The contribution from one independent variables is **the same** whatever the values are for the other.

Note a. can also be formulate as **multiplicativity on odds scale**

$$\text{odds} = \text{odds}_0 \cdot \text{OR}_1^{x_1} \cdot \text{OR}_2^{x_2} \dots \text{OR}_k^{x_k}$$

Morten Frydenberg

Linear and Logistic regression - Note 4

45

### Logistic regression model in general

$$\ln(\text{odds}) = \beta_0 + \sum_{p=1}^k \beta_p \cdot x_p$$

If one consider two persons who differ with

$$\Delta x_1 \text{ in } x_1, \Delta x_2 \text{ in } x_2 \dots \text{ and } \Delta x_k \text{ in } x_k$$

then difference in the **log odds** is :

$$\sum_{p=1}^k \beta_p \cdot \Delta x_p$$

Again we see that the contribution for each of the explanatory variables:

- are **added**,
- are **proportional** to the difference
- and **does not dependent** of the difference in the other

**on the log odds scale.**

Morten Frydenberg

Linear and Logistic regression - Note 4

46

### Logistic regression model in general

$$\ln(\text{odds}) = \beta_0 + \sum_{p=1}^k \beta_p \cdot x_p$$

If one consider two persons who differ with

$$\Delta x_1 \text{ in } x_1, \Delta x_2 \text{ in } x_2 \dots \text{ and } \Delta x_k \text{ in } x_k$$

then odds ratio :

$$\text{OR} = \text{OR}_1^{\Delta x_1} \cdot \text{OR}_2^{\Delta x_2} \dots \text{OR}_k^{\Delta x_k}$$

**Note** the model might also be formulated:

$$p = \Pr[Y=1] = \frac{\exp\left(\beta_0 + \sum_{p=1}^k \beta_p \cdot x_p\right)}{1 + \exp\left(\beta_0 + \sum_{p=1}^k \beta_p \cdot x_p\right)}$$

Morten Frydenberg

Linear and Logistic regression - Note 4

47

### Logistic regression model in general

$$\ln(\text{odds}) = \beta_0 + \sum_{p=1}^k \beta_p \cdot x_p$$

**The data:**  $Y=1/0$  dichotomous dependent variable

$x_1, x_2 \dots x_k$  independent/explanatory variables

Like in the normal regression models it is assumed that the Y's are **independent** given the explanatory variables.

This assumption can, in general, only be checked by **scrutinising** the design.

Look out for data sampled in **clusters**:

Patients within the **same GP**

Children within the **same family**

**Twins.**

Morten Frydenberg

Linear and Logistic regression - Note 4

48

**Logistic regression model in general****Estimation:**

Excepting the two by two tables, there are **no closed form** for the estimates.

The **distribution** of the estimates **are not known**.

Estimates are found by the method of **maximum likelihood**.

Estimates are using **iterative methods**.

Standard errors, confidence intervals and all tests are based on **asymptotics**.

That is, all statistical **inference** are **approximate**.

The **more data** - the more events -the **better** the approximations.