Logistic regression
Morten Frydenberg ©
Section of Biostatistics, Aarhus Univ, Denmark
Stata 11+

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 one **continuous** independent variable without interaction.

One **binary** independent variable and one **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, i.e.** 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 ratios**.

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

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 an **unmatched case-control** study.

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

And the odds ratios are comparable to the odds ratios based on 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 persons with **known BMI** at the start.

We will focus on the risk obesity (BMI≥30 kg/m^2) .

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.(χ^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: $OR = \dfrac{odds_{Women}}{odds_{Men}}$

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 regression

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

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

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

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

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

Comparing with the equation above we get:

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

and

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

Finding an odds ratio using logistic regression

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

$$\ln\left(odds_{Men}\right) \qquad\qquad \ln\left(OR\right)$$

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

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

Probabilities and odds

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

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

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

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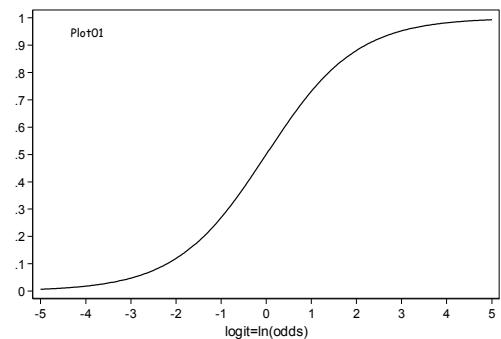
Probabilities and odds

Probabilities and odds

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

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

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

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

The model on **probability scale** is :

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

Finding an odds ratio using logistic regression

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

Back to finding the estimates.

In Stata: *logit obese b1.sex, baselevel*

```
Iteration 0:   log likelihood = -1795.5437
Iteration 1:   log likelihood = -1790.3856
Iteration 2:   log likelihood = -1790.3703
Iteration 3:   log likelihood = -1790.3703
Logistic regression                              Number of obs   =     4690
                                                 LR chi2(1)      =    10.35
                                                 Prob > chi2     =   0.0013
Log likelihood = -1790.3703                      Pseudo R2       =   0.0029
------------------------------------------------------------------------------
  obese |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------+---------------------------------------------------------------------
    sex |
      1 |   (base)
      2 |   .2868784   .0898972     3.19   0.001     .1106831    .4630738
        |
  _cons |  -2.086606   .0705261   -29.59   0.000    -2.224835   -1.948378
------------------------------------------------------------------------------
```

Finding an odds ratio using logistic regression

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

$$\hat{\beta}_1 = \ln(OR) \qquad\qquad 95\% \text{ CI for } \ln(OR)$$

```
obese |    Coef.   Std. Err.      z    P>|z|    [95% Conf. Interval]
------+---------------------------------------------------------------
    2 |  .2868784   .0898972     3.19   0.001    .1106831    .4630738
 _cons | -2.086606   .070526   -29.59   0.000   -2.224835   -1.948378
```

$$OR = \exp(0.2868784) = 1.33 \qquad 95\% \text{ CI}: (1.12;1.59).$$

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

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

95% 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

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

An easier way to obtain the odds ratio.
`logit obese b1.sex ,or baselevel cformat(%6.4f)`

```
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
Log likelihood = -1790.3703              Pseudo R2       =   0.0029

   obese | Odds Ratio  Std. Err.     z    P>|z|    [95% Conf. Interval]
---------+---------------------------------------------------------------
   sex   |
     1   |   1.0000   (base)
     2   |   1.3323    0.1198     3.19   0.001    1.1170     1.5890
         |
   _cons |   0.1241    0.0088   -29.59   0.000    0.1081     0.1425
```

Odds not odds ratio

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 for men and women.

Is it also associated with age?

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

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

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

As we have seen before using *age* implies that β_0 references to a newborn (*age*=0).

So we will choose *age*=45 reference instead:

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

Morten Frydenberg   Linear and Logistic regression - Note 4   15

The obesity and age: version 1

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

The interpretation of the parameters:

β_0 : the **log odds** for a 45-year-old person.

β_1 : the **log odds ratio**, when comparing two persons who differ 1 year in age.

$\exp(\beta_1)$: the **odds ratio** 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

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

Obtaining the estimates in Stata:
generate *age45=age-45*
`logit obese age45`

```
obese |    Coef.   Std. Err.      z    P>|z|    [95% Conf. Interval]
------+---------------------------------------------------------------
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`
```
obese | Odds Ratio  Std. Err.      z    P>|z|    [95% Conf.Interval]
------+---------------------------------------------------------------
age45 |   1.0354     0.0053     6.78   0.000    1.0251     1.0459
_cons |   0.1373     0.0064   -42.84   0.000    0.1253     0.1503
```

Morten Frydenberg   Linear and Logistic regression - Note 4   17

The obesity and age: version 1

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

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

The **odds** for obesity **among 45-year**-olds:

$$0.1373\ (0.1253;0.1503)$$

The **prevalence** of obesity **among 45-year**-olds:

$$0.1207\ (0.1114;0.1307)$$

$$odds = \exp(\log(odds)) \qquad \text{Prob} = \frac{odds}{1+odds}$$

Morten Frydenberg   Linear and Logistic regression - Note 4   18

The obesity and age: version 1

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

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

$1.0354^{4.5}\ (1.0251^{4.5}; 1.0459^{4.5}) = 1.17\ (1.12; 1.22)$

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   .0269968    6.78   0.000   1.117806   1.223668
------------------------------------------------------------
```

The obesity and age: version 1

Estimated relationship: $\ln(odds) = -1.986 + 0.0348 \cdot (age - 45)$
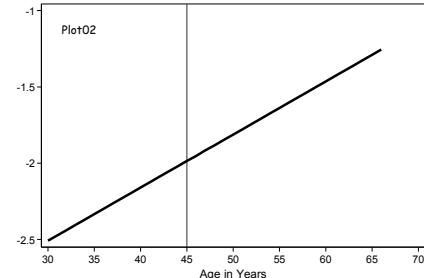
Plot02

The obesity and age: version 1

Estimated relationship:

$$prevalence = \frac{\exp(-1.986 + 0.0348 \cdot (age - 45))}{1 + \exp(-1.986 + 0.0348 \cdot (age - 45))}$$
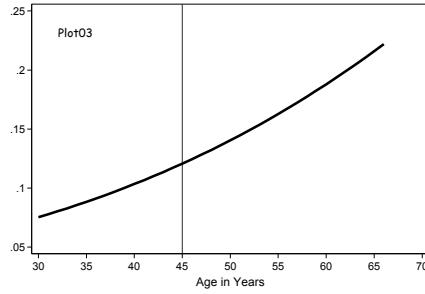
Plot03

The obesity and age: version 2

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

This model assumes that one year of age difference is associated with the same odds ratio irrespectively 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 into seven five-year groups:
`egen agegrp7=cut(age), at(0,35,40,45,50,55,60,120) label`

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

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 :

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

Where *agei* is an indicator for being in the *i*th age group

---

The obesity and age: version 2

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

The interpretation of the parameters:

α_0 : the **log odds** in the **reference** group=the youngest.

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

`logit obese i.agegrp7,baselevel`   **Not all output**

```
------------------------------------------------------------------
   obese |   Coef.   Std. Err.     z    P>|z|   [95% Conf. Interval]
---------+--------------------------------------------------------
 agegrp7 |
      0  |  (base)
      1  | .5483322  .239152    2.29   0.022   .0796029   1.017061
      2  | .5186016  .2419361   2.14   0.032   .0444155   .9927877
      3  | .6576621  .2417944   2.72   0.007   .1837537   1.13157
      4  | .9790072  .2383937   4.11   0.000   .5117642   1.44625
      5  | .9644652  .2428468   3.97   0.000   .4884941   1.440436
      6  | 1.41737   .2523832   5.62   0.000   .9227081   1.912032
   _cons | -2.660564 .2156798 -12.34   0.000  -3.083288  -2.237839
------------------------------------------------------------------
```

The obesity and age: version 2

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

`logit obese i.agegrp7,or baselevel`   **Not all output**

obese	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf. Interval]
0	1.0000	(base)				
1	1.7304	0.4138	2.29	0.022	1.0829	2.7651
2	1.6797	0.4064	2.14	0.032	1.0454	2.6987
3	1.9303	0.4667	2.72	0.007	1.2017	3.1005
4	2.6618	0.6346	4.11	0.000	1.6682	4.2472
5	2.6234	0.6371	3.97	0.000	1.6299	4.2225
6	4.1263	1.0414	5.62	0.000	2.5161	6.7668

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!

The obesity and age: version 2

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

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 i.agegrp7
will give a **"Wald test"** of no difference between the **seven** groups .

```
( 1)  [obese]1.agegrp7 = 0
( 2)  [obese]2.agegrp7 = 0
( 3)  [obese]3.agegrp7 = 0
( 4)  [obese]4.agegrp7 = 0
( 5)  [obese]5.agegrp7 = 0
( 6)  [obese]6.agegrp7 =
      chi2(  6) =   55.26
      Prob > chi2 =   0.0000
```

Highly significant differences

The obesity and age: version 2

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

`logit obese b3.agegrp7,or baselevel`   **Not all output**

obese	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf. Interval]
agegrp7						
0	0.5181	0.1253	-2.72	0.007	0.3225	0.8321
1	0.8964	0.1348	-0.73	0.467	0.6676	1.2038
2	0.8702	0.1347	-0.90	0.369	0.6425	1.1786
3	1.0000	(base)				
4	1.3790	0.2057	2.15	0.031	1.0293	1.8474
5	1.3591	0.2123	1.96	0.050	1.0006	1.8459
6	2.1377	0.3648	4.45	0.000	1.5299	2.9868

The OR between the **second oldest** and the **45-49-year-old**:

1.36 (1.00;1.85)

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

A borderline significant different in prevalence!

The obesity and age: version 2

$\hat{\alpha}_0 + \hat{\alpha}_4$

$\hat{\alpha}_0$

Estimated relationship

The obesity and age: version 1 and 2

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 model is based on three **assumptions**:

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

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

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

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:

β_0 : the **log odds** for a 45-year-old **man**.

β_1 : the **log odds ratio**, when comparing a woman to a man of the same age.

β_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 Δage years older than the other.

Morten Frydenberg   Linear and Logistic regression - Note 4   31

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:

`logit obese b1.sex age45`

```
Iteration 0:   log likelihood = -1795.5437
Iteration 3:   log likelihood = -1767.7019
Logistic regression                          Number of obs  =     4690
                                             LR chi2(2)     =    55.68
                                             Prob > chi2    =   0.0000
Log likelihood = -1767.7019                  Pseudo R2      =   0.0155

------------------------------------------------------------------------
 obese |    Coef.   Std. Err.      z    P>|z|    [95% Conf. Interval]
-------+----------------------------------------------------------------
  sex  |
    1  |  (base)
    2  | .2743976  .0903385     3.04   0.002    .0973374    .4514579
age45  | .0344723  .0053154     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 Frydenberg   Linear and Logistic regression - Note 4   32

The obesity, sex and age : version 1

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

`logit obese b1.sex age45, or`

```
 obese | Odds Ratio  Std. Err.     z    P>|z|    [95% Conf. Interval]
-------+----------------------------------------------------------------
  sex  |
    1  |   1.0000    (base)
    2  |   1.3157    0.1189     3.04   0.002    1.1022      1.5706
age45  |   1.0351    0.0053     6.71   0.000    1.0247      1.0455
 _cons |   0.1168    0.0084   -29.74   0.000    0.1014      0.1346
```

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 a **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 Frydenberg   Linear and Logistic regression - Note 4   33

The obesity, sex and age : version 1

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

The estimated relationship
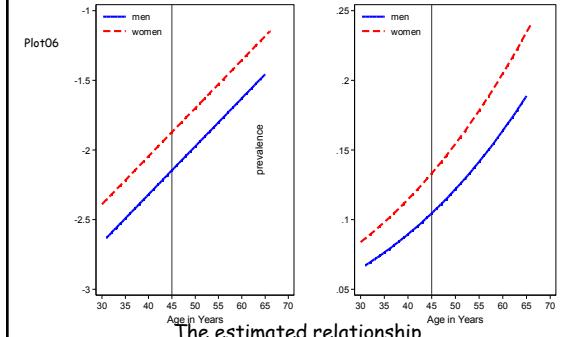
Morten Frydenberg   Linear and Logistic regression - Note 4   34

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 **assumption**:

Proportionality on the **logit scale**: The contribution age is **proportional** to its 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 Frydenberg   Linear and Logistic regression - Note 4   35

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:

`logit obese b1.sex##c.age45`

```
------------------------------------------------------------------------
  obese |    Coef.   Std. Err.     z    P>|z|    [95% Conf. Interval]
--------+---------------------------------------------------------------
  2.sex |  .116797  .0950345    1.23   0.219   -.0694672    .3030611
  age45 | -.005684  .0083728   -0.68   0.497   -.0220953    .0107255
sex#c.age45 |
      2 | .065803   .010743     6.13   0.000    .0447472    .0868588
  _cons |-2.083041 .0706433   -29.49   0.000   -2.221499   -1.944583
```

| Men | | Difference between women and men |

Estimates odds ratios:

```
  obese | Odds Ratio  Std. Err     z    P>|z|    [95% Conf. Interval]
--------+---------------------------------------------------------------
  2.sex |  1.1239               1.23   0.219    0.9329      1.3540
  age45 |  0.9943              -0.68   0.497    0.9781      1.0108
sex#c.age45 |
      2 |  1.0680               6.13   0.000    1.0458      1.0907
  _cons |  0.1246             -29.49   0.000    0.1084      0.1430
```
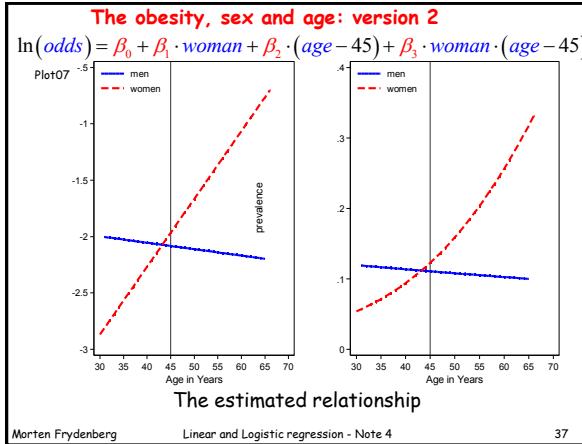
Slide 37

The obesity, sex and age: version 2

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

The estimated relationship

Morten Frydenberg   Linear and Logistic regression - Note 4   37

Slide 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

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

Slide 39

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

Slide 40

A small case-control example

```
logit cancer b0.smoker b1.age,or
Iteration 0:   log likelihood = -496.55682
Iteration 1:   log likelihood = -437.36405
Iteration 2:   log likelihood = -429.36499
Iteration 3:   log likelihood = -428.94718
Iteration 4:   log likelihood = -428.94432
Iteration 5:   log likelihood = -428.94432
Logistic regression
```

"Many" iterations

```
                                    Number of obs  =     977
                                    LR chi2(6)     =  135.23
                                    Prob > chi2    =  0.0000
Log likelihood = -428.94432         Pseudo R2      =  0.1362
```

cancer	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf. Interval]	
smoker						
0	(base)					
1	2.350498	.4513038	4.45	0.000	1.613342	3.424472
age						
1	(base)					
2	2.832192	2.243677	1.31	0.189	.5995101	13.37978
3	16.58078	12.17376	3.82	0.000	3.932284	69.91412
4	27.89911	20.32372	4.57	0.000	6.691354	116.3233
5	34.79453	25.59025	4.83	0.000	8.231513	147.1061
6	27.713	21.89264	4.21	0.000	5.891876	130.3507

Morten Frydenberg   Linear and Logistic regression - Note 4   40

Slide 41

A small case-control example

```
logit cancer b0.smoker b3.age,or baselev
Iteration 0:   log likelihood = -496.55682
Iteration 1:   log likelihood = -437.36405
Iteration 2:   log likelihood = -429.36499
Iteration 3:   log likelihood = -428.94718
Iteration 4:   log likelihood = -428.94432
Iteration 5:   log likelihood = -428.94432
Logistic regression
```

```
                                    Number of obs  =     977
                                    LR chi2(6)     =  135.23
                                    Prob > chi2    =  0.0000
Log likelihood = -428.94432         Pseudo R2      =  0.1362
```

cancer	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf. Interval]	
smoker						
0	(base)					
1	2.350498	.4513038	4.45	0.000	1.613342	3.424472
age						
1	.0603108	.0442807	-3.82	0.000	.014303	.254305
2	.1708118	.0652397	-4.63	0.000	.080800	.361098
3	(base)					
4	1.682618	.3701188	2.37	0.018	1.093327	2.58953
5	2.098486	.5042862	3.08	0.002	1.31025	3.360918
6	1.671393	.6277714	1.37	0.171	.800514	3.489699

Morten Frydenberg   Linear and Logistic regression - Note 4   41

Slide 42

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** test 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:
```
logit cancer i.smoker i.age
estimates store model1
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

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

This is based on three assumptions:

a. **Additivity on log-odds scale**: The contribution from each of the independent variables are **added**.

b. **Proportionality**: The contribution from independent variables is **proportional** to its value (with a factor β)

c. **No effectmodification**: The contribution from one independent variable **is the same** whatever the values of the other.

Note **a.** can also be formulated as **multiplicitivity** on the **odds scale**

$$odds = odds_0 \cdot OR_1^{x_1} \cdot OR_2^{x_2} \cdots$$

Morten Frydenberg   Linear and Logistic regression - Note 4   45

Logistic regression model in general

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

If one consider two persons who differ with

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

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

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

Again we see that the contribution from each of the explanatory variables:
  are **added**,
  are **proportional** to the difference
  and **does not depend** on the difference in the other explanatory variables

On the log odds scale!

Morten Frydenberg   Linear and Logistic regression - Note 4   46

Logistic regression model in general

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

If one consider two persons who differ with

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

then the odds ratio is:

$$OR = OR_1^{\Delta x_1} \cdot OR_2^{\Delta x_2} \cdots$$

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

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

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

x_1 , x_2 ... 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.

Morten Frydenberg   Linear and Logistic regression - Note 4   49