Working with logistics regression models

Morten Frydenberg ©
Department of Biostatistics, Aarhus Univ, Denmark

The `lincom` command for logistic regression

Further remarks on logistic regression

- **Diagnostics**: residuals and leverages
- **Enough data?**
- **Test of fit**: The Hosmer-Lemeshow test
- **ROC-curves** and the **area under** the ROC-curve

Extensions to the ordinary logistic regression:

Conditional logistic regression
- **When?**
- **What?**
- **How?**

Other methods for analyzing binary data
- Models for **relative risks**
- Models for **risk differences**

Data with **several random components: Binary outcome**

Clustered binary data with **one random components**

The `lincom` command after `logit` or `regress`

Consider the model:

$$\text{logit}\left(\Pr\left(\text{obese}\right)\right) = \beta_0 + \beta_1 \cdot woman + \beta_2 \cdot \left(age - 45\right)$$

```
   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
--------------------------------------------------------------------------
```

Here men are reference.

If we want to find the log odds for a 45 year old women we can calculate by hand $-2.147+0.274=-1.873$

But what about confidence interval?

We could change the reference to women and fit the model once more.
But.......

The `lincom` command after `logit` or `regress`

$$\text{logit}\left(\Pr\left(\text{obese}\right)\right) = \beta_0 + \beta_1 \cdot woman + \beta_2 \cdot \left(age - 45\right)$$

Stata has a command that can be used for this: "lincom"

```
lincom _cons+_Isex

 ( 1)  _Isex_2 + _cons = 0
------------------------------------------------------------------
obese |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
------+-----------------------------------------------------------
  (1) |   -1.8726     .05813   -32.21   0.000    -1.986602   -1.758714
------------------------------------------------------------------
```

You can add "`,or`" to get odds/odds ratios.

```
lincom _cons+_Isex,or
( 1)  _Isex_2 + _cons = 0
--------------------------------------------------------------------------
   obese | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------+----------------------------------------------------------------
     (1) |   .1537145   .0089363   -32.21   0.000     .1371606     .172266
--------------------------------------------------------------------------
```


The `lincom` command after `logit` or `regress`

$$\text{logit}\left(\Pr\left(\text{obese}\right)\right) = \beta_0 + \beta_1 \cdot woman + \beta_2 \cdot \left(age - 45\right)$$

Some examples:

Odds for a 42 year old woman:

```
lincom _cons+_Isex-age45*3,or
 ( 1)  _Isex_2 - 3 age45 + _cons = 0
--------------------------------------------------------------------------
   obese | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------+----------------------------------------------------------------
     (1) |   .1386122   .0088378   -30.89   0.000     .1222772    .1571295
--------------------------------------------------------------------------
```

Odds ratio for 4.5 age difference:

```
lincom age45*4.5,or
 ( 1)  4.5 age45 = 0
--------------------------------------------------------------------------
   obese | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------+----------------------------------------------------------------
     (1) |   1.167804   .0269869     6.71   0.000     1.116091    1.221914
--------------------------------------------------------------------------
```


Logistic regression models: Diagnostics

In the linear regression we saw some example of statistics:

residuals, **standardized residuals** and **leverage**

which can be used in the **model checking** and search for strange or **influential** data points.

Such statistics can also be defined for the logistic regression model.

But they are much more **difficult to interpret** and **cannot** in general be **recommended**.

Checking the validity of a logistic regression model will mainly be based on **comparing** it with other **models**.

Logistic regression models: Test of fit

A common, and to some extend informative, test of fit is the **Hosmer-Lemeshow** test.

Consider the model for obesity from Monday

$$\text{logit}\left(\Pr(\text{obese})\right) = \beta_0 + \beta_1 \cdot woman + \beta_2 \cdot (age - 45)$$

```
Logit estimates                                   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]
-------------+----------------------------------------------------------------
      _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
------------------------------------------------------------------------------
```

Significantly better than nothing – but is it good?

Logistic regression models: Do you have enough data?

All inference in logistic regression models are based on asymptotics , i.e. **assuming that you have a lot of data** !

Rule of thumb:
You should have at least **10 events** per variable (parameter) in the model.

A large standard error typical indicates that you have to little information concerning the variable and that the **estimate and standard error are not valid.**

Lower your **ambitions** or get **more data** !

A exact methods exists, but only one (**expensive**) program can do it.

And it will give also wide confidence intervals.

Logistic regression models: Test of fit

What about comparing the **estimated prevalence** with the **observed prevalence**?

In the Hosmer-Lemeshow test the data is **divided** into groups (traditionally 10) according to the **estimated** probabilities

and the **observed** and **expected** counts are compared in these groups by a chi-square test.

Most programs, that can fit a logistic regression model, can calculate this test.

In Stata it is done by (**after fitting the model**):

lfit, group(10) *table*

The data is divided into **deciles** after the estimated probabilities.

Logistic regression models: Test of fit

OUTPUT

```
Logistic model for obese, goodness-of-fit test
  (Table collapsed on quantiles of estimated probabilities)
  +--------------------------------------------------------+
  | Group |   Prob | Obs_1 | Exp_1 | Obs_0 | Exp_0 | Total |
  |-------+--------+-------+-------+-------+-------+-------|
  |     1 | 0.0841 |    64 |  40.9 |   462 | 485.1 |   526 |
  |     2 | 0.0953 |    43 |  45.5 |   453 | 450.5 |   496 |
  |     3 | 0.1045 |    44 |  44.6 |   398 | 397.4 |   442 |
  |     4 | 0.1112 |    42 |  50.3 |   422 | 413.7 |   464 |
  |     5 | 0.1217 |    44 |  51.4 |   394 | 386.6 |   438 |
  |     6 | 0.1332 |    52 |  63.0 |   441 | 430.0 |   493 |
  |     7 | 0.1456 |    53 |  61.7 |   389 | 380.3 |   442 |
  |     8 | 0.1592 |    62 |  69.8 |   392 | 384.2 |   454 |
  |     9 | 0.1834 |    98 |  89.9 |   424 | 432.1 |   522 |
  |    10 | 0.2407 |    99 |  83.8 |   314 | 329.2 |   413 |
  +--------------------------------------------------------+
       number of observations =      4690
             number of groups =        10
      Hosmer-Lemeshow chi2(8) =        26.01
                  Prob > chi2 =         0.0010
```

One problem: Too many in the tails

Significant difference between observed and expected!

Logistic regression models: Test of fit

xi: logit obese i.sex*age45

```
lfit, group(10) table
Logistic model for obese, goodness-of-fit test
  (Table collapsed on quantiles of estimated probabilities)
  +--------------------------------------------------------+
  | Group |   Prob | Obs_1 | Exp_1 | Obs_0 | Exp_0 | Total |
  |-------+--------+-------+-------+-------+-------+-------|
  |     1 | 0.0796 |    36 |  35.9 |   466 | 466.1 |   502 |
  |     2 | 0.1011 |    42 |  41.1 |   406 | 406.9 |   448 |
  |     3 | 0.1053 |    49 |  49.6 |   429 | 428.4 |   478 |
  |     4 | 0.1096 |    50 |  54.8 |   458 | 453.2 |   508 |
  |     5 | 0.1124 |    52 |  54.2 |   436 | 433.8 |   488 |
  |     6 | 0.1153 |    51 |  46.4 |   355 | 359.6 |   406 |
  |     7 | 0.1182 |    52 |  53.9 |   410 | 408.1 |   462 |
  |     8 | 0.1590 |    76 |  70.3 |   428 | 433.7 |   504 |
  |     9 | 0.2133 |    96 |  91.8 |   391 | 395.2 |   487 |
  |    10 | 0.3310 |    97 | 103.0 |   310 | 304.0 |   407 |
  +--------------------------------------------------------+
       number of observations =      4690
             number of groups =        10
      Hosmer-Lemeshow chi2(8) =         2.43
                  Prob > chi2 =         0.9650
```

The model 'fits' – when we look at in this way !!!!!!

ROC curves – sensitivity and specificity

```
generate over45=(age>45) if age!=.
diagt obese over45
```

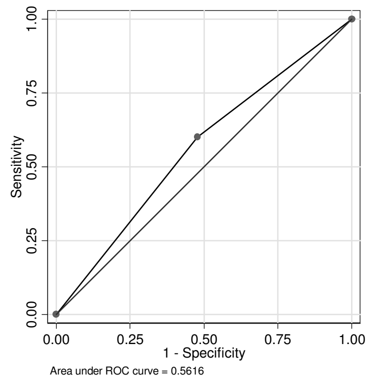
```
     obese |      Pos.       Neg. |     Total
-----------+----------------------+----------
  Abnormal |       361        240 |       601
    Normal |     1,952      2,137 |     4,089
-----------+----------------------+----------
     Total |     2,313      2,377 |     4,690
True abnormal diagnosis defined as obese = 1
                                                     [95% Confidence Interval]
---------------------------------------------------------------------------
Prevalence                          Pr(A)       13%        12%      13.8%
---------------------------------------------------------------------------
Sensitivity                        Pr(+|A)     60.1%        56%        64%
Specificity                        Pr(-|N)     52.3%      50.7%      53.8%
ROC area                     (Sens. + Spec.)/2  .562       .541       .583
---------------------------------------------------------------------------
Likelihood ratio (+)          Pr(+|A)/Pr(+|N)   1.26       1.17       1.35
Likelihood ratio (-)          Pr(-|A)/Pr(-|N)   .764        .69       .846
Odds ratio                         LR(+)/LR(-)  1.65       1.38       1.96
Positive predictive value          Pr(A|+)     15.6%      14.2%      17.2%
Negative predictive value          Pr(N|-)     89.9%      88.6%      91.1%
---------------------------------------------------------------------------
```

ROC curves – sensitivity and specificity

```
roctab obese over45,graph tab de

           |        over45
     obese |         0          1 |     Total
-----------+----------------------+----------
         0 |     2,137      1,952 |     4,089
         1 |       240        361 |       601
-----------+----------------------+----------
     Total |     2,377      2,313 |     4,690

Detailed report of Sensitivity and Specificity
------------------------------------------------------------------------------
                                           Correctly
Cutpoint      Sensitivity   Specificity   Classified          LR+          LR-
------------------------------------------------------------------------------
( >= 0 )         100.00%         0.00%       12.81%       1.0000
( >= 1 )          60.07%        52.26%       53.26%       1.2583       0.7641
( >  1 )           0.00%       100.00%       87.19%                    1.0000
------------------------------------------------------------------------------

                      ROC                    -Asymptotic Normal--
           Obs       Area     Std. Err.      [95% Conf. Interval]
     ------------------------------------------------------------
          4690     0.5616       0.0107        0.54061     0.58268
```


ROC curves – sensitivity and specificity

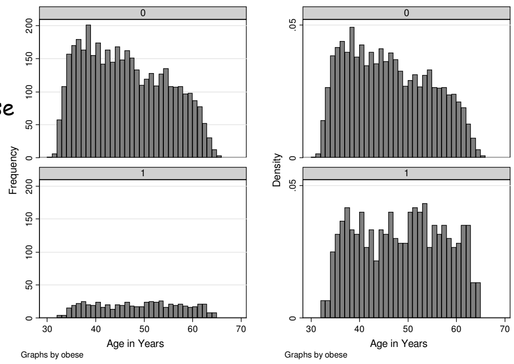
roctab obese over45,graph tab de



ROC curves – sensitivity and specificity

Non-obese

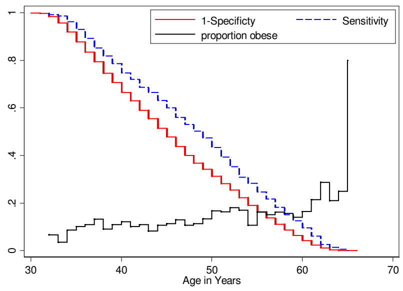
Obese



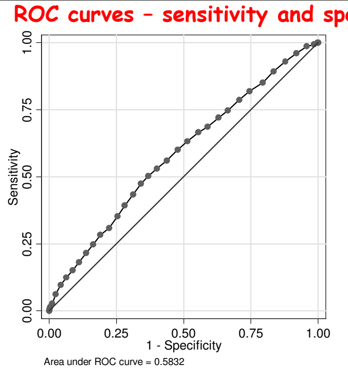
In population Within group

ROC curves – sensitivity and specificity

Cutpoint	Sensitivity	Specificity
(>= 30)	100.00%	0.00%
(>= 31)	100.00%	0.02%
(>= 32)	100.00%	0.17%
(>= 33)	99.33%	1.57%
(>= 34)	98.67%	4.21%
(>= 35)	96.17%	8.05%
(>= 36)	93.01%	12.20%
(>= 37)	89.35%	16.58%
(>= 38)	85.19%	20.57%
(>= 39)	81.86%	25.48%
(>= 40)	78.70%	29.27%
(>= 41)	74.71%	33.53%
(>= 42)	72.05%	37.00%
(>= 43)	68.72%	40.99%
(>= 44)	66.56%	44.53%
(>= 45)	63.23%	48.64%
(>= 46)	60.07%	52.26%
(>= 47)	56.07%	56.22%
(>= 48)	53.08%	59.92%
(>= 49)	50.25%	63.17%
(>= 50)	47.42%	65.86%
(>= 51)	43.43%	68.77%
(>= 52)	39.27%	71.90%
(>= 53)	35.27%	74.57%
(>= 54)	30.95%	77.67%
(>= 55)	28.29%	80.97%
(>= 56)	24.79%	83.61%
(>= 57)	21.63%	86.23%
(>= 58)	18.14%	88.87%
(>= 59)	15.14%	91.24%
(>= 60)	12.48%	93.64%
(>= 61)	9.65%	95.74%
(>= 62)	6.16%	97.63%
(>= 63)	2.66%	98.90%
(>= 64)	1.33%	99.63%
(>= 65)	0.67%	99.93%
(>= 66)	0.00%	99.95%
(> 66)	0.00%	100.00%



ROC curves – sensitivity and specificity




```
roctab obese age, graph tab de
           ROC      -Asymptotic Normal--
Obs        Area     [95% Conf. Interval]
 -------------------------------------------
 4690     0.5832    0.55866      0.60779
```


ROC curves – the area under the curve

The area under the ROC curve – what is it?

Note, it only depends on the sensitivity and the specificity , but not on the prevalence!

The mathematical definition of the are under the ROC-curve is:

Suppose we take one **random obese** person and one **random non-obese person** then :

$$\boxed{\begin{aligned}&\Pr(\text{age obese} > \text{age non-obese})\\ &\qquad +\tfrac{1}{2}\Pr(\text{age obese} = \text{age non-obese})\end{aligned}}$$

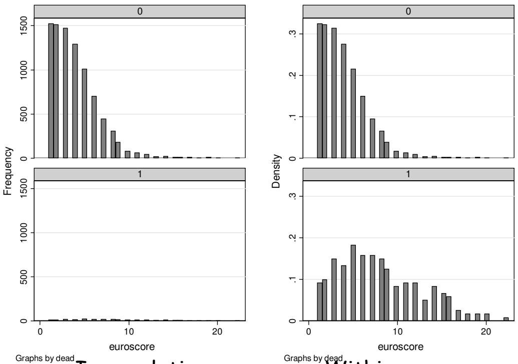
Note, this is not related to the predictive values!

Predicting dead after operation - the (additive) euroscore

Data on 8949 operations 223 deaths (2.42%)

Survived

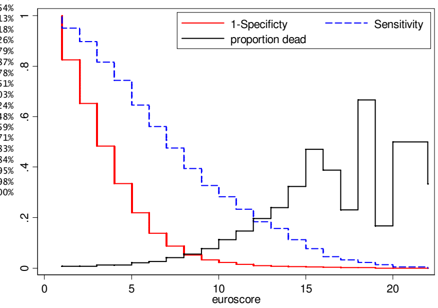
Dead



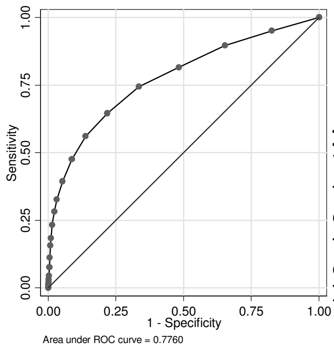
In population Within group

ROC curves - sensitivity and specificity

Cutpoint	Sensitivity	Specificity
(>= 1)	100.00%	0.00%
(>= 2)	95.07%	17.45%
(>= 3)	89.69%	34.79%
(>= 4)	81.61%	51.72%
(>= 5)	74.44%	66.54%
(>= 6)	64.57%	78.13%
(>= 7)	56.05%	86.18%
(>= 8)	47.53%	91.26%
(>= 9)	39.46%	94.79%
(>= 10)	32.74%	96.87%
(>= 11)	28.25%	97.78%
(>= 12)	23.32%	98.51%
(>= 13)	18.39%	99.03%
(>= 14)	15.70%	99.24%
(>= 15)	11.21%	99.48%
(>= 16)	7.62%	99.59%
(>= 17)	4.48%	99.71%
(>= 18)	3.14%	99.83%
(>= 19)	2.24%	99.84%
(>= 20)	1.35%	99.95%
(>= 22)	0.45%	99.98%
(> 22)	0.00%	100.00%



ROC curves - sensitivity and specificity

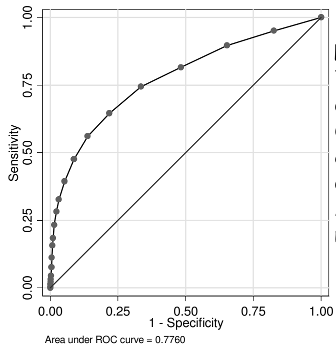



```
roctab obese age, graph tab de
                 ROC      -Asymptotic Normal--
Obs              Area     [95% Conf. Interval]
-----------------------------------------------
8949            0.7760     0.74090     0.81105
```

Interpretation:
There is 78 (74;81)% chance that a random person, who died, had a higher Euroscore, than a random person, who died not die.
Is this relevant ?

What about the predictive value?
If the Euroscore is 15 then 40% will die.
What is the consequence of this information?

ROC curves - sensitivity and specificity

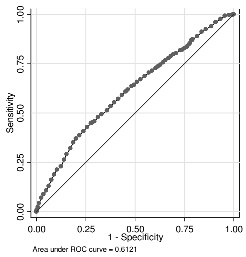


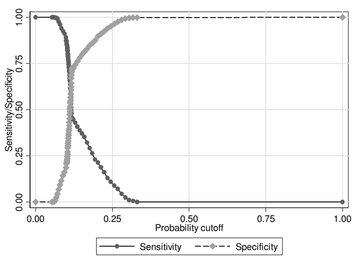
Looking at the ROC-curve:
The different cut-points have different consequences.
Often one have to choose a cut-point and use that in the decision making in the future.
All other cut-point will have no relevance.

ROC curves - sensitivity and specificity

In Stata and other programs you can plot roc curves using on the predicted probabilities based on a logistic regression:

```
xi:logit obese i.sex*age45
lroc
lsens
```





Conditional logistic regression When

Used in two situations:

1. **Matched** studies (binary response).
2. **Unmatched** studies with a **confounder** with **many distinct values.**

In **1.** the models correspond to **the way data was collected.**

In **2.** the method adjust for a **'mathematical' flaw** in the unconditional method.

An example of situation **2.** the confounder is "*kommune*" having 275 distinct values.

Conditional logistic regression
What

The logistic regression model (outcome disease yes/no):

$$\ln(odds) = \alpha + \sum_{i=1}^{k} (\beta_i \cdot x_i)$$

ln(odds) in reference ln(odds ratios)

Suppose the model above hold in each strata:

$$\ln(odds) = \alpha_s + \sum_{i=1}^{k} (\beta_i \cdot x_i)$$

ln(odds) in reference **different in each strata**

ln(odds ratios) **the same in each strata**

Conditional logistic regression
What

$$\ln(odds) = \alpha_s + \sum_{i=1}^{k} (\beta_i \cdot x_i)$$

ln(odds) different in each strata

We are not interested in these !

In a **matched** study these are 'controlled'.

In a **conditional** logistic regression one **'condition on the odds in each strata'** , i.e. these case/control ratio.

In the conditional model the α's **disappear** !

The β's , the log OR's, are still in and **can be estimated**.

Conditional logistic regression
How

It is easy !

You need a statistical software package.

A package made for **research in epidemiology**

Not in social science

Not *SPSS*

But ***Stata***, *EPICURE, EPILOG, EGRET, EPIINFO(2000) and SAS can do it.*

Conditional logistic regression
How

An example using *Stata*

A study of cancer in the oral cavity

Matched on **gender** and **10 years age groups**

Ten strata (*genage*)

Here we focus on

textile-worker and

life time consumption of alcohol (three groups)

Conditional logistic regression
How

logistic regression in *Stata*

xi:logit cancer textile i.alkcon i.**genage**

Part of the output:

cancer	Coef.	Std. Err.	z	P>|z|	CI	
textile	**.5022**	**.4141**	**1.213**	**0.225**	**-.3094**	**1.3139**
_Ialkcon_1	**.4628**	**.2823**	**1.639**	**0.101**	**-.0905**	**1.0163**
_Ialkcon_2	**2.7165**	**.3232**	**8.404**	**0.000**	**2.0829**	**3.3501**
_Igenage_2	.2450	1.2514	0.196	0.845	-2.2075	2.6977
_Igenage_3	-.4940	.5503	-0.898	0.369	-1.5726	.5846
_Igenage_4	.1798	.6406	0.281	0.779	-1.0758	1.4353
_Igenage_5	-.2899	.5482	-0.529	0.597	-1.3644	.7844
_Igenage_6	.2127	.6262	0.340	0.734	-1.0147	1.4401
_Igenage_7	-.2305	.5355	-0.431	0.667	-1.2802	.8190
_Igenage_8	.5507	.5263	1.046	0.295	-.4809	1.5825
_Igenage_9	.0315	.5884	0.054	0.957	-1.1217	1.1847
_Igenage_10	.5572	.5595	0.996	0.319	-.53954	1.6539
_const	-1.4692	.4762	-3.085	0.002	-2.4027	-.5356

Conditional logistic regression in *Stata*

The syntax:

xi:clogit cancer textile i.alkcon,***group*(genage)**

Part of the output:

cancer	Coef.	Std. Err.	z	P>|z|	CI	
textile	**.4929**	**.4103**	**1.201**	**0.230**	**-.3112**	**1.2971**
_Ialkcon_1	**.452**	**.27923**	**1.621**	**0.105**	**-.094**	**.9999**
_Ialkcon_2	**2.660**	**.31936**	**8.332**	**0.000**	**2.034**	**3.2868**

xi:clogit cancer textile i.alkcon, ***group*(genage)** *or*

cases	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
textile	1.63708	.6717022	1.20	0.230	.732517	3.658661
_Ialkcon_1	1.572508	.4390957	1.62	0.105	.909724	2.718168
_Ialkcon_2	14.30908	4.569879	8.33	0.000	7.651811	26.75835

Other methods to analysis of binary response data
Relative Risk models

Logistic regression model focus on the **Odds Ratios**

This is the correct thing to do in **case-control** studies.

In **follow-up** studies **Relative Risk** is often the appropriate measure of association, (personal risk).

I.e. a model like this might be more relevant:

$$\Pr(\text{event}) = p_0 \times RR_1 \times RR_2 \times RR_3$$

$$\ln\{\Pr(\text{event})\} = \ln(p_0) + \ln(RR_1) + \ln(RR_2) + \ln(RR_3)$$

$$\ln\{\Pr(\text{event given the covariates})\} = \alpha + \sum_{i=1}^{p}(\beta_i \cdot x_i)$$

That is linear on **log-probability** scale

Other methods to analysis of binary response data
Relative Risk models

$$\ln\{\Pr(\text{event given the covariates})\} = \alpha + \sum_{i=1}^{p}(\beta_i \cdot x_i)$$

Such a model **modelling the relative risk** can easily be fitted by many programs (not SPSS).

Logistic regression in Stata:

```
xi: logit obese age i.sex
```

or

```
xi: glm obese age i.sex, fam(bin) link(logit)
```

Relative risk model:

```
xi: glm obese age i.sex, fam(bin) link(log)
```

The *link* is **log** instead of **logit**

Other methods to analysis of binary response data
Risk difference models

Logistic regression model focus on the **Odds Ratios**

This is the correct thing to do in **case-control** studies.

In **follow-up** studies **Risk Difference** is often the appropriate measure of association, (community effect).

I.e. a model like this might be more relevant:

$$\Pr(\text{event}) = p_0 + RD_1 + RD_2 + RD_3$$

$$\Pr(\text{event given the covariates}) = \alpha + \sum_{i=1}^{p}(\beta_i \cdot x_i)$$

That is linear on **probability** scale

Other methods to analysis of binary response data
Risk difference models

$$\Pr(\text{event given the covariates}) = \alpha + \sum_{i=1}^{p}(\beta_i \cdot x_i)$$

Such a model **modelling the risk difference** can easily be fitted by many programs (not SPSS).

Logistic regression in Stata:

```
xi: logit obese age i.sex
```

or

```
xi: glm obese age i.sex, fam(bin) link(logit)
```

Risk difference model:

```
xi: glm obese age i.sex, fam(bin) link(id)
```

The *link* is **identity** instead of **logit**

Other methods to analysis of binary response data

Three different links for Obese "=" sex "+" age

Other methods to analysis of binary response data
Problems

$$\Pr(\text{event}) = p_0 \times RR_1 \times RR_2 \times RR_3$$

As the relative risk can be **larger** than one the product might be **larger than one** !

$$\Pr(\text{event}) = p_0 + RD_1 + RD_2 + RD_3$$

The sum might **negative** and be **larger than one** !

Note: In Stata you can also use the `binreg` command

Clustered data / data with several random components
Dichotomous outcome

A different outcome:

$$H_{fpd} = \begin{cases} 1 & \text{if the person has hayfewer} \\ 0 & \text{else} \end{cases}$$

A statistical model:

Systematic part

$$\text{logit}\left(H_{fpd}=1\right) = \boxed{\beta_0 + \beta_I \cdot I + \beta_U \cdot U + \beta_A \cdot A + \beta_S \cdot S + \beta_G \cdot G}$$
$$\boxed{+F_f + P_{fp} + \cancel{E_{fpd}}}$$

Random part

This is not needed due to the binomial error

Clustered data / data with several random components
Dichotomous outcome

$$\text{logit}\left(H_{fpd}=1\right) = \beta_0 + \beta_I \cdot I + \beta_U \cdot U + \beta_A \cdot A + \beta_S \cdot S + \beta_G \cdot G$$
$$+F_f + P_{fp}$$

That is, an ordinary logistic regression + **random components.**

- A **generalized linear mixed model**
- A **multilevel model for dichotomous outcome**

Comments 1:

- It is **important** to include the **relevant random** components in the model.
- 'Multilevel models' is **essential** in medical/epidemiological research.

Clustered data / data with several random components
Dichotomous outcome

Comments 2:

- The theory and insight into the models for non-normal data are **not yet fully developed**.
- The main problem being that it is very difficult find **valid (unbiased) estimates**.
- Several software programs **falsely claim** to estimate the models.
- Some programs like Stata and NLwin can give you valid estimates if you take care and have **a lot of data**.

Advice:
Do not try to estimate this kind of models without consulting a specialist.

Clustered data / data with one random components
Dichotomous outcome

If the models only involves **one random components**, e.g. **variation between families** or between **GP's**,

then methods exists which can **adjust the standards errors**.

Remember that if the **data contains clusters**, then the precision of the estimates overestimated, that is the reported **standard errors is too small**.

So called **robust methods** or **sandwich estimates** of the standard errors will (try) adjust for this problem.

Only a **few** programs have this option – Stata does!