Regression models for binary data

Morten Frydenberg ©
Section for Biostatistics, Aarhus Univ, Denmark

When to use binary regressions models.

The three measures of association:
- RR: Risk Ratio (Relative Risk)
- OR: Odds Ratio
- RD: Risk Difference

Switching the outcome
Changing the reference

One (three) examples: RD, RR and OR -models:
Interpretation, estimation, lincom

Plotting the "response curves"

The connection between OR and RR

1

The limitations of the RD and RR models
- The bounds for RD and RR
- Invalid "probabilities"
- Problem with estimation/fitting

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

Binary regression models: **The assumptions**

Checking the models
- No valid "residuals" → No diagnostic plots

General comments to estimation
- Subtle details with standard errors.
- Watch out for **'small' reference** groups

Why the logistic regression model is so popular.

Conditoinal logistic regression

2

Binary regression models: Introduction

A binary 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.

3

Binary regression models: Introduction

A **OR-regression** model, **logistic regression**, models the **probability** of a "positive event" via odds and associations via **odds ratios**.

A **RR-regression** model, **relative risk (risk ratio) regression**, models the **probability** of a "positive" event and associations via **risk ratios, i.e. relative risks**.

A **RD-regression** model, **risk difference regression** models the **probability** of a "positive" event and associations via **risk differences.**

There other types of models that can be used for binary outcome.
In psychometrics one often used **Probit-models.**
These model are **not covered** in this course.

4

Risk ratios, odds ratios, risk differences

Risk (chance) of event - comparing two groups

Let π_1 be the risk of the event in group 1 and
π_2 be the risk of the event in group 2

The odds in group i is defined as : $odds_i = \pi_i/(1-\pi_i)$

$$RR_{1vs2} = \frac{\pi_1}{\pi_2}$$

$$OR_{1vs2} = \frac{odds_1}{odds_2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = \frac{\pi_1 \cdot (1-\pi_2)}{\pi_2 \cdot (1-\pi_1)}$$

$$RD_{1vs2} = \pi_1 - \pi_2$$

5

Risk ratios, odds ratios, risk differences

Parity of the mother	Post term delivery No	Yes	Total	Risk of post term delivery(%) Estimate	Lower	Upper
At least one previous	4,696	1,677	6,373	26.3%	25.2%	27.4%
No previous deliveries	4,216	1,722	5,938	29.0%	27.8%	30.2%
Total	8,912	3,399	12,311			

At least one versus first	Risk of post term delivery(%) Estimate	Lower	Upper
RR	0.91	0.86	0.96
OR	0.87	0.81	0.95
RD	-2.7%	-4.3%	-1.1%

$$\widehat{RR}(yes)_{1vs2} = \frac{26.3\%}{29.0\%} = 0.91$$

$$\widehat{OR}(yes)_{1vs2} = \frac{26.3\% \cdot (100\% - 29.0\%)}{29.0\% \cdot (100\% - 26.3\%)} = \frac{1677 \cdot 4216}{1722 \cdot 4696} = 0.87$$

$$\widehat{RD}(yes)_{1vs2} = 26.3\% - 29.0\% = -2.7\%$$

6

Risk ratios, odds ratios, risk differences

Parity of the mother	Post term delivery No	Yes	Total	Risk of **not** post term delivery(%) Estimate	Lower	Upper
At least one previous	4,696	1,677	6,373	73.7%	72.6%	74.8%
No previous deliveries	4,216	1,722	5,938	71.0%	69.8%	72.2%
Total	8,912	3,399	12,311			

At least one versus first	Risk of **not** post term delivery(%) Estimate	Lower	Upper
RR	1.04	1.02	1.06
OR	1.14	1.06	1.24
RD	-2.7%	-4.3%	-1.1%

$$\widehat{RR}(no)_{1vs2} = \frac{73.7\%}{71.0\%} = 1.04$$

$$\widehat{OR}(no)_{1vs2} = \frac{73.7\% \cdot (100\% - 71.0\%)}{71.0\% \cdot (100\% - 73.7\%)} = \frac{73.7\% \cdot 29.0\%}{71.0\% \cdot 26.3\%} = 1.14$$

$$\widehat{RD}(no)_{1vs2} = 73.7\% - 71.0\% = 2.7\%$$

7

Risk ratios, odds ratios, risk differences
Switching outcome yes→no

$$OR(yes)_{1vs2} = \frac{\pi_1 \cdot (1-\pi_2)}{\pi_2 \cdot (1-\pi_1)} = \frac{1}{OR(no)_{1vs2}}$$

$$RD(yes)_{1vs2} = \pi_1 - \pi_2 = -\left[(1-\pi_1) - (1-\pi_2)\right] = -RD(no)_{1vs2}$$

$$\widehat{OR}(yes)_{1vs2} = 0.87 = \frac{1}{1.14}$$

$$\widehat{RD}(yes)_{1vs2} = -2.7\% = -[2.7\%]$$

No nice relationship between $RR(yes)_{1vs2}$ and $RR(no)_{1vs2}$

8

Risk ratios, odds ratios, risk differences
Changing "reference"

	Post term delivery			Risk of post term delivery(%)		
Parity of the mother	No	Yes	Total	Estimate	Lower	Upper
At least one previous	4,696	1,677	6,373	26.3%	25.2%	27.4%
No previous deliveries	4,216	1,722	5,938	29.0%	27.8%	30.2%
Total	8,912	3,399	12,311			

	Risk of post term delivery(%)		
First versus at least one	Estimate	Lower	Upper
RR	0.91	0.86	0.96
OR	1.14	1.06	1.24
RD	2.7%	1.1%	4.3%

9

Risk ratios, odds ratios, risk differences
Changing "reference"

$$RR(yes)_{1vs2} = \frac{\pi_1}{\pi_2} = \frac{1}{\pi_2/\pi_1} = \frac{1}{RR(yes)_{2vs1}}$$

$$OR(yes)_{1vs2} = \frac{\pi_1 \cdot (1-\pi_2)}{\pi_2 \cdot (1-\pi_1)} = \frac{1}{OR(yes)_{2vs1}}$$

$$RD(yes)_{1vs2} = \pi_1 - \pi_2 = -[\pi_2 - \pi_1] = -RD(yes)_{2vs1}$$

$$\widehat{RR}(yes)_{1vs2} = 0.91 = \frac{1}{1.10}$$

$$\widehat{OR}(yes)_{1vs2} = 0.87 = \frac{1}{1.14}$$

$$\widehat{RD}(yes)_{1vs2} = -2.7\% = -[2.7\%]$$

10

The example

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/m²) .

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

Divided into gender

	Obese	Not-Obese
Women	375 (14.2%)	2268 (85.8%)
Men	226 (11.0%)	1821 (89.0%)

We will also look at age divided in three group and serum cholesterol.

11

A risk difference model

$$\Pr(obese) = \beta_0 + \beta_1 \cdot (scl - 200) + \beta_2 \cdot woman + \beta_3 \cdot (40 \le age < 50) + \beta_4 \cdot (50 \le age)$$

β_0: **Risk** among men, age<40, with scl=200

β_1: **Risk Difference** comparing two persons, where the first has one unit higher serum cholesterol, adjusted for sex and age

β_2: **Risk Difference** comparing two persons, where the first is a woman and the second a man, adjusted for serum cholesterol and age

β_3: **Risk Difference** comparing two persons, where the first is in the age group 40≤age<50 and the second in age<40, adjusted for serum cholesterol and sex

β_4: **Risk Difference** comparing two persons, where the first is in the age group 50≤age and the second in age<40, adjusted for serum cholesterol and sex

12

A risk difference model

```
binreg obese  b1.sex b0.agegrp3 scl200,rd

Iteration 1:   deviance =  3496.151
Iteration 2:   deviance =  3494.521
Iteration 3:   deviance =  3494.449
Iteration 4:   deviance =  3494.445
Iteration 5:   deviance =  3494.445
Iteration 6:   deviance =  3494.445
Iteration 7:   deviance =  3494.445

Generalized linear models                          No. of obs      =      4,658
Optimization     : MQL Fisher scoring              Residual df     =      4,653
                   (IRLS EIM)                      Scale parameter =          1
Deviance         =  3494.444982                    (1/df) Deviance =    .751009
Pearson          =  4657.969064                    (1/df) Pearson  =   1.001068

Variance function: V(u) = u*(1-u)                  [Bernoulli]
Link function    : g(u) = u                        [Identity]

                                                   BIC             =  -35806.38

Output omitted
```

Not much of interest - we will return to this later!

13

A risk difference model

```
binreg obese  b1.sex b0.agegrp3 scl200,rd
Output omitted
```

This is not a RD. It is a risk!

```
------------------------------------------------------------------------------
             |                 EIM
       obese | Risk Diff.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |
         Men |          0  (base)
       Women |   .0200744   .0093267     2.15   0.031     .0017943    .0383545
             |
     agegrp3 |
          0- |          0  (base)
         40- |   .0049258   .0110113     0.45   0.655    -.0166559    .0265076
         50- |   .0559626   .0126235     4.43   0.000      .031221    .0807042
             |
      scl200 |   .0005806   .0001144     5.08   0.000     .0003564    .0008048
       _cons |   .0782201   .0092233     8.48   0.000     .0601428    .0962973
------------------------------------------------------------------------------
```

You can used `lincom`, `regeq` and `testparm`
You can get estimated probabilities/risk by
`predict….,mu`
Residuals and leverage does not make any sense

14

```
------------------------------------------------------------------------------
             |                 EIM
       obese | Risk Diff.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |
         Men |          0  (base)
       Women |   .0200744   .0093267     2.15   0.031     .0017943    .0383545
             |
     agegrp3 |
          0- |          0  (base)
         40- |   .0049258   .0110113     0.45   0.655    -.0166559    .0265076
         50- |   .0559626   .0126235     4.43   0.000      .031221    .0807042
             |
      scl200 |   .0005806   .0001144     5.08   0.000     .0003564    .0008048
       _cons |   .0782201   .0092233     8.48   0.000     .0601428    .0962973
------------------------------------------------------------------------------
```

Risk, man, age<40 scl=200: **7.8 (6.0;9.6)%**

Women **2.0 (0.2;3.8)%-point** higher risk than men
adjusted for age and serum cholesterol level

100 units difference in serum cholesterol level corresponds to a 5.8(3.6;8.0) %-point increase in risk
adjusted for age and sex

15

A risk ratio model

A "usual" additive model on log-probability scale

$$\ln\left[\Pr(obese)\right] = \beta_0 + \beta_1 \cdot (scl - 200) + \beta_2 \cdot woman + \beta_3 \cdot (40 \le age < 50) + \beta_4 \cdot (50 \le age)$$

$$\gamma_i = \exp\left[\beta_i\right]$$

A multiplicative model on probability scale

$$\Pr(obese) = \exp\left[\beta_0 + \beta_1 \cdot (scl - 200) + \beta_2 \cdot woman + \beta_3 \cdot (40 \le age < 50) + \beta_4 \cdot (50 \le age)\right]$$

$$= \gamma_0 \cdot \gamma_1^{(scl-200)} \cdot \gamma_2^{woman} \cdot \gamma_3^{(40 \le age < 50)} \cdot \gamma_4^{(50 \le age)}$$

16

A risk ratio model

$$\ln\left[\Pr(obese)\right] = \beta_0 + \beta_1 \cdot (scl - 200) + \beta_2 \cdot woman$$
$$+\beta_3 \cdot (40 \le age < 50) + \beta_4 \cdot (50 \le age)$$

$$\gamma_i = \exp[\beta_i]$$

γ_0: **Risk** among men, age<40, with scl=200

γ_1: **Risk Ratio** comparing two persons, where the first has one unit higher serum cholesterol, adjusted for sex and age

γ_2: **Risk Ratio** comparing two persons, where the first is a woman and the second a man, adjusted for serum cholesterol and age

γ_3: **Risk Ratio** comparing two persons, where the first is in the age group 40≤age<50 and the second in age<40, adjusted for serum cholesterol and sex

γ_4: **Risk Ratio** comparing two persons, where the first is in the age group 50≤age and the second in age<40, adjusted for serum cholesterol and sex

17

A risk ratio model

```
binreg obese  b1.sex b0.agegrp3 scl200,rr

Iteration 1:   deviance =   4849.97
Iteration 2:   deviance =  3631.229
Iteration 3:   deviance =  3503.426
Iteration 4:   deviance =  3499.733
Iteration 5:   deviance =  3499.727
Iteration 6:   deviance =  3499.727

Generalized linear models                          No. of obs      =       4658
Optimization     : MQL Fisher scoring              Residual df     =       4653
                   (IRLS EIM)                      Scale parameter =          1
Deviance         =  3499.727216                    (1/df) Deviance =   .7521443
Pearson          =  4644.143131                    (1/df) Pearson  =   .9980965

Variance function: V(u) = u*(1-u)                  [Bernoulli]
Link function    : g(u) = ln(u)                    [Log]

                                                   BIC             =   -35801.1

Output omitted
```

Not much of interest - we will return to this later!

18

A risk ratio model

```
. binreg obese  b1.sex b0.agegrp3 scl200,rr
Output omitted
```

This not a RR. It is a risk!

```
------------------------------------------------------------------------------
             |                 EIM
       obese | Risk Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |
         Men |          1  (base)
       Women |   1.250198   .0988399     2.82   0.005     1.070738    1.459736
             |
     agegrp3 |
          0- |          1  (base)
         40- |   1.074009   .1176699     0.65   0.515      .866461    1.331272
         50- |   1.549492   .1609833     4.21   0.000     1.264014    1.899447
             |
      scl200 |   1.003053    .000818     3.74   0.000      1.00145    1.004659
       _cons |   .0825146   .0079357   -25.94   0.000      .068339    .0996307
------------------------------------------------------------------------------
```

You can get estimated probabilities/risk by `predict….,mu`

Residuals and leverage does not make any sense

You can used `lincom`, `regeq` and `testparm`,

but the estimates, se and CIs are found on log scale

19

```
------------------------------------------------------------------------------
             |                 EIM
       obese | Risk Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |
         Men |          1  (base)
       Women |   1.250198   .0988399     2.82   0.005     1.070738    1.459736
             |
     agegrp3 |
          0- |          1  (base)
         40- |   1.074009   .1176699     0.65   0.515      .866461    1.331272
         50- |   1.549492   .1609833     4.21   0.000     1.264014    1.899447
             |
      scl200 |   1.003053    .000818     3.74   0.000      1.00145    1.004659
       _cons |   .0825146   .0079357   -25.94   0.000      .068339    .0996307
------------------------------------------------------------------------------
```

Risk, man, age<40 scl=200: **8.3 (6.8;10.0)%**

Women **25 (7;46)%** higher risk than men
adjusted for age and serum cholesterol level

100 units difference in scl corresponds to a
36(16;59) % increase in risk adjusted for age and sex

$1.003053^{100}\ (1.00145^{100};1.004659^{100})=1.36(1.16;1.59)$

20

A risk ratio model

```
. regeq
estimated equation
-2.4948 +0.2233 * 2.sex  +0.0714 * 1.agegrp3  +0.4379 * 2.agegrp3  ///
+0.0030 * scl200

equation
b0 + b1 * 2.sex + b2 * 1.agegrp3 + b3 * 2.agegrp3 + ///
b4 * scl200

. lincom scl200*100

 ( 1)  100*scl200 = 0
------------------------------------------------------------------------------
       obese |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .3048838   .0816075     3.74   0.000     .1449361    .4648316
------------------------------------------------------------------------------

. lincom scl200*100, eform
 ( 1)  100*scl200 = 0
------------------------------------------------------------------------------
       obese |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   1.356467   .1106979     3.74   0.000     1.155966    1.591746
------------------------------------------------------------------------------
```

Log RR

exp

RR

21

A risk ratio model

RR woman 40≤age<50 versus man age<40, same scl

```
lincom (2.sex+1.agegrp3)-(1.sex+0.agegrp3)
 ( 1)  - 1b.sex + 2.sex - 0b.agegrp3 + 1.agegrp3 = 0
------------------------------------------------------------------------------
       obese |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .2947002   .1337717     2.20   0.028     .0325124    .5568879
------------------------------------------------------------------------------

lincom (2.sex+1.agegrp3)-(1.sex+0.agegrp3), eform
 ( 1)  - 1b.sex + 2.sex - 0b.agegrp3 + 1.agegrp3 = 0
------------------------------------------------------------------------------
       obese |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   1.342724   .1796185     2.20   0.028     1.033047    1.745233
------------------------------------------------------------------------------
```

Log RR

exp

RR

22

A risk ratio model

Risk for women 40≤age<50 with scl=150scl

```
. lincom _cons+2.sex+1.agegrp3+scl200*(-50)
 ( 1)  2.sex + 1.agegrp3 - 50*scl200 + _cons = 0
------------------------------------------------------------------------------
       obese |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -2.352522   .1028584   -22.87   0.000     -2.55412   -2.150923
------------------------------------------------------------------------------

. lincom _cons+2.sex+1.agegrp3+scl200*(-50), eform
 ( 1)  2.sex + 1.agegrp3 - 50*scl200 + _cons = 0
------------------------------------------------------------------------------
       obese |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |    .095129   .0097848   -22.87   0.000     .0777606    .1163767
------------------------------------------------------------------------------
```

Log Risk

exp

Risk

23

A risk ratio model

The estimates, se, CI, tests and p-values are found/calculated on log scale

```
. binreg obese  b1.sex b0.agegrp3 scl200,rr coef
Output omitted
------------------------------------------------------------------------------
             |                 EIM
       obese |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |
        Men  |          0  (base)
      Women  |   .2233019   .0790594     2.82   0.005     .0683483    .3782555
             |
     agegrp3 |
         0-  |          0  (base)
        40-  |   .0713982   .1095614     0.65   0.515    -.1433382    .2861347
        50-  |   .4379273   .1038975     4.21   0.000      .234292    .6415626
             |
      scl200 |   .0030488   .0008161     3.74   0.000     .0014494    .0046483
       _cons |   -2.49478   .0961727   -25.94   0.000    -2.683275   -2.306285
------------------------------------------------------------------------------
```

ok

24

A odds ratio model

A "usual" additiv model on log-odds scale

$$\ln\left[Odds(obese)\right] = \beta_0 + \beta_1 \cdot (scl - 200) + \beta_2 \cdot woman$$
$$+ \beta_3 \cdot (40 \le age < 50) + \beta_4 \cdot (50 \le age)$$

$$\gamma_i = \exp[\beta_i]$$

A multiplicative model on odds scale

$$Odds(obese) = \exp\left[\beta_0 + \beta_1 \cdot (scl - 200) + \beta_2 \cdot woman + \beta_3 \cdot (40 \le age < 50) + \beta_4 \cdot (50 \le age)\right]$$
$$= \gamma_0 \cdot \gamma_1^{(scl-200)} \cdot \gamma_2^{woman} \cdot \gamma_3^{(40 \le age < 50)} \cdot \gamma_4^{(50 \le age)}$$

A complicated model on probabilty scale

$$\Pr(obese) = \frac{Odds(obese)}{1 + Odds(obese)}$$

25

A odds ratio model

$$\ln\left[Odds(obese)\right] = \beta_0 + \beta_1 \cdot (scl - 200) + \beta_2 \cdot woman + \beta_3 \cdot (40 \le age < 50) + \beta_4 \cdot (50 \le age)$$

$$\gamma_i = \exp[\beta_i]$$

γ_0: **Odds** among men, age<40, with scl=200

γ_1: **Odds Ratio** comparing two persons, where the first has one unit higher serum cholesterol, adjusted for sex and age

γ_2: **Odds Ratio** comparing two persons, where the first is a woman and the second a man, adjusted for serum cholesterol and age

γ_3: **Odds Ratio** comparing two persons, where the first is in the age group 40≤age<50 and the second in age<40, adjusted for serum cholesterol and sex

γ_4: **Odds Ratio** comparing two persons, where the first is in the age group 50≤age and the second in age<40, adjusted for serum cholesterol and sex

26

A odds ratio model

```
binreg obese  b1.sex b0.agegrp3 scl200,or

Iteration 1:   deviance =  3519.095
Iteration 2:   deviance =  3498.984
Iteration 3:   deviance =  3498.815
Iteration 4:   deviance =  3498.815
Iteration 5:   deviance =  3498.815

Generalized linear models                          No. of obs      =       4658
Optimization     : MQL Fisher scoring              Residual df     =       4653
                   (IRLS EIM)                      Scale parameter =          1
Deviance         =  3498.815069                    (1/df) Deviance =   .7519482
Pearson          =   4643.16574                    (1/df) Pearson  =   .9978865

Variance function: V(u) = u*(1-u)                  [Bernoulli]
Link function    : g(u) = ln(u/(1-u))              [Logit]

                                                   BIC             =  -35802.01

Output omitted
```

Not much of interest - we will return to this later!

27

A odds ratio model

```
. binreg obese  b1.sex b0.agegrp3 scl200,or
Output omitted

------------------------------------------------------------------------------
             |                 EIM
       obese | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |
        Men  |          1  (base)
      Women  |   1.282514   .1164761     2.74   0.006     1.073389    1.532382
             |
     agegrp3 |
         0-  |          1  (base)
        40-  |   1.075132   .1318739     0.59   0.555     .8453855    1.367315
        50-  |    1.64556   .1947525     4.21   0.000      1.30489     2.07517
             |
         scl |   1.003923   .0009843     3.99   0.000     1.001996    1.005854
       _cons |   .0890348   .0095747   -22.49   0.000     .0721145    .1099252
------------------------------------------------------------------------------
```

This not an OR. It is an odds

You can get estimated probabilities/risk by `predict….,mu`

Residuals and leverage does not make any sense

You can used `lincom`, `regeq` and `testparm`,

but the estimates, se and CIs are found on log scale

28

```
------------------------------------------------------------------------------
             |                 EIM
       obese | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |
         Men |          1  (base)
       Women |   1.282514   .1164761     2.74   0.006      1.073389    1.532382
             |
     agegrp3 |
          0- |          1  (base)
         40- |   1.075132   .1313739     0.59   0.555      .8453855    1.367315
         50- |    1.64556   .1947525     4.21   0.000       1.30489     2.07517
             |
         scl |   1.003923   .0009843     3.99   0.000      1.001996    1.005854
       _cons |   .0890348   .0095747   -22.49   0.000      .0721145    .1099252
------------------------------------------------------------------------------
```

Odds, man, age<40 scl=200: **0.089 (0.072;0.110)**

Women **28 (7;53)%** higher **odds** than men
adjusted for age and serum cholesterol level

100 units difference in scl corresponds to a
48(22;79) % increase in **odds** adjusted for age and sex
1.003923^{100} $(1.001996^{100};1.005854^{100})$=1.48(1.22;1.79)

29

A odds ratio model

```
. regeq
estimated equation
-2.4187 +0.2488 * 2.sex  +0.0724 * 1.agegrp3  +0.4981 * 2.agegrp3  ///
+0.0039 * scl200

equation
b0 + b1 * 2.sex + b2 * 1.agegrp3 + b3 * 2.agegrp3 + ///
b4 * scl200
```

Log OR

```
lincom scl200*100
 ( 1)  100*scl200 = 0
------------------------------------------------------------------------------
       obese |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .3915803   .0980411     3.99   0.000     .1994233    .5837373
------------------------------------------------------------------------------

. lincom scl200*100,eform
 ( 1)  100*scl200 = 0
------------------------------------------------------------------------------
       obese |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   1.479317   .1450338     3.99   0.000     1.220699    1.792726
------------------------------------------------------------------------------
```

exp

OR

30

A odds ratio model

OR woman 40≤age<50 versus man age<40, same scl

Log OR

```
lincom (2.sex+1.agegrp3)-(1.sex+0.agegrp3)
 ( 1)  - 1b.sex + 2.sex - 0b.agegrp3 + 1.agegrp3 = 0
------------------------------------------------------------------------------
       obese |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .3212652   .1510523     2.13   0.033     .0252081    .6173223
------------------------------------------------------------------------------

lincom (2.sex+1.agegrp3)-(1.sex+0.agegrp3),eform
 ( 1)  - 1b.sex + 2.sex - 0b.agegrp3 + 1.agegrp3 = 0
------------------------------------------------------------------------------
       obese |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   1.378871   .2082817     2.13   0.033     1.025529    1.853957
------------------------------------------------------------------------------
```

exp

OR

31

A odds ratio model

Risk for women 40≤age<50 with scl=150scl

Log Risk

```
lincom _cons+2.sex+1.agegrp3+scl200*(-50)
 ( 1)  2.sex + 1.agegrp3 - 50*scl200 + _cons = 0
------------------------------------------------------------------------------
       obese |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -2.293253   .1188768   -19.29   0.000     -2.526247   -2.060258
------------------------------------------------------------------------------

lincom _cons+2.sex+1.agegrp3+scl200*(-50), eform
 ( 1)  2.sex + 1.agegrp3 - 50*scl200 + _cons = 0
------------------------------------------------------------------------------
       obese |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .1009376   .0119991   -19.29   0.000      .0799586     .127421
------------------------------------------------------------------------------
```

exp

Odds

```
disp %12.6f .1009376/1.1009376 %12.6f .0799586/1.0799586 %12.6f .127421/1.127421

    0.091683     0.074039     0.113020
```

Risk = Odds/(1+Odds) by hand

32

A odds ratio model

Risk for women 40≤age<50 with scl=150scl

$$Probability = \frac{odds}{1+odds} = \frac{\exp(logodds)}{1+\exp(logodds)} = \text{invlogit}(logodds)$$

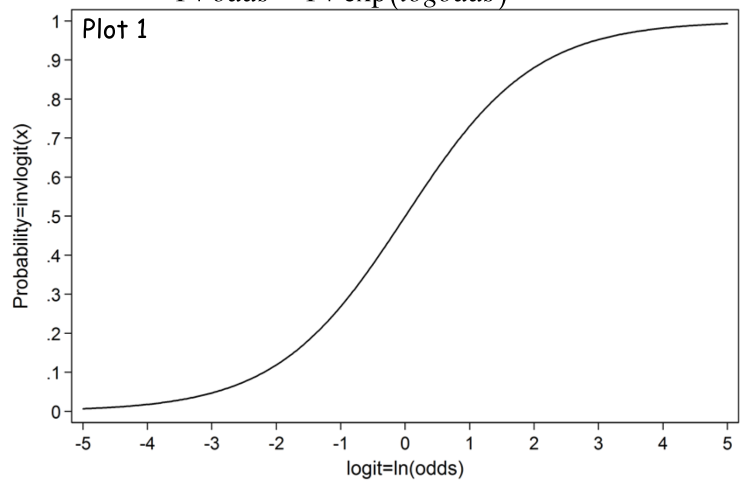
No , eform

```
lincom _cons+2.sex+1.agegrp3+scl200*(-50)
 ( 1)  2.sex + 1.agegrp3 - 50*scl200 + _cons = 0
------------------------------------------------------------------------------
       obese |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -2.293253   .1188768   -19.29   0.000    -2.526247   -2.060258
------------------------------------------------------------------------------
disp  %12.6f invlogit( r(estimate)             ) ///
      %12.6f invlogit( r(estimate)-1.96*r(se) ) ///
      %12.6f invlogit( r(estimate)+1.96*r(se) )

    0.091683     0.074038     0.113020
```

33

$$Probability = \frac{odds}{1+odds} = \frac{\exp(logodds)}{1+\exp(logodds)} = \text{invlogit}(logodds)$$



34

binreg: The tests in the output

binreg…, **rd**:
- Risk difference =0
- _cons: Risk=0 ?????

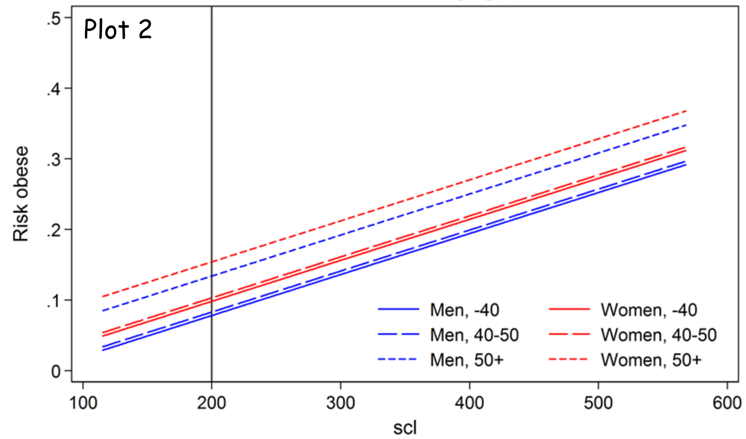
binreg…, **rr**:
- Risk ratio = 1
- _cons: log(risk)=0 that is Risk=1 ?????

binreg…, **or**:
- Odds ratio = 1
- _cons: log(odds)=0 that is Risk=0.5 ?????

35

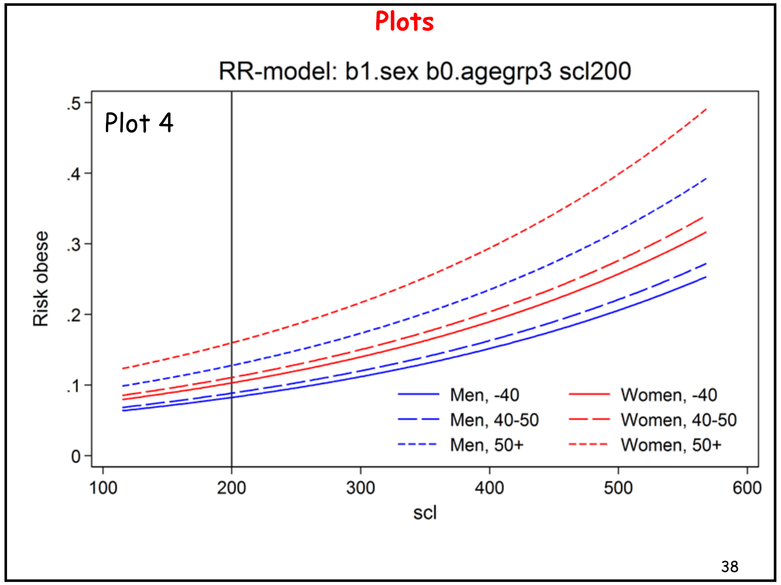
Plots

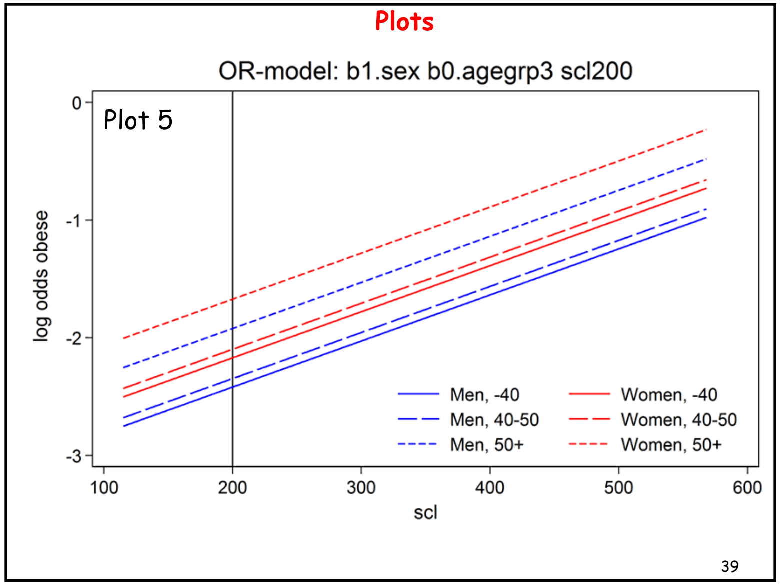
RD-model: b1.sex b0.agegrp3 scl200

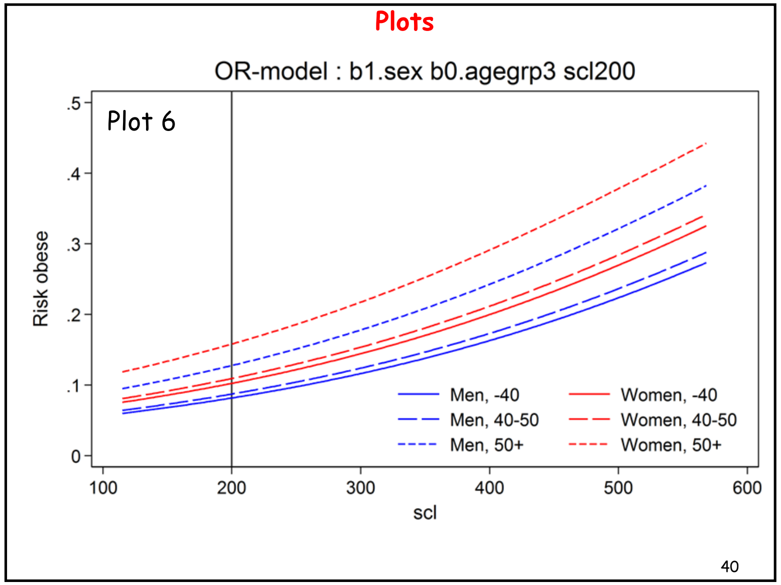


36

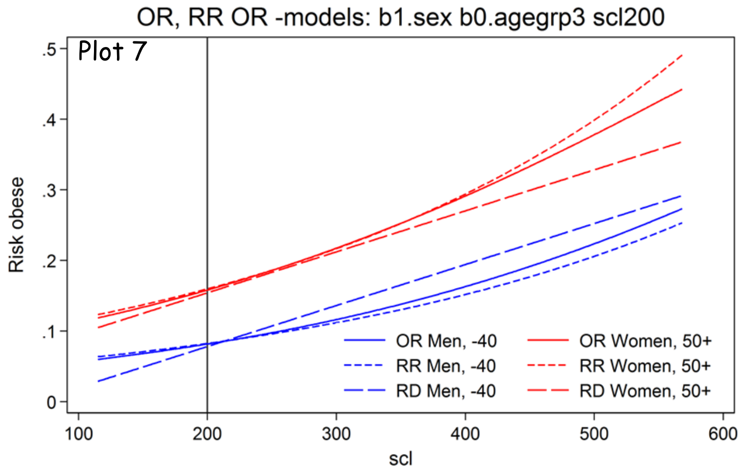

Plots
RR-model: b1.sex b0.agegrp3 scl200
Plot 3
log Risk obese
-.5
-1
-1.5
-2
-2.5
-3
100
200
300
400
500
600
scl
Men, -40
Men, 40-50
Men, 50+
Women, -40
Women, 40-50
Women, 50+
37


Plots
RR-model: b1.sex b0.agegrp3 scl200
Plot 4
Risk obese
.5
.4
.3
.2
.1
0
100
200
300
400
500
600
scl
Men, -40
Men, 40-50
Men, 50+
Women, -40
Women, 40-50
Women, 50+
38


Plots
OR-model: b1.sex b0.agegrp3 scl200
Plot 5
log odds obese
0
-1
-2
-3
100
200
300
400
500
600
scl
Men, -40
Men, 40-50
Men, 50+
Women, -40
Women, 40-50
Women, 50+
39


Plots
OR-model : b1.sex b0.agegrp3 scl200
Plot 6
Risk obese
.5
.4
.3
.2
.1
0
100
200
300
400
500
600
scl
Men, -40
Men, 40-50
Men, 50+
Women, -40
Women, 40-50
Women, 50+
40

Plots



41

When is the risk ratio and the odds ratio close?

$$OR_{1vs2} = \frac{\pi_1 \cdot (1-\pi_2)}{\pi_2 \cdot (1-\pi_1)} = RR_{1vs2} \cdot \frac{(1-\pi_2)}{(1-\pi_1)}$$

$$= RR_{1vs2} \cdot \frac{(1-\pi_2)}{(1-\pi_2 \cdot RR_{1vs2})}$$

We see that:

$$\pi_1 > \pi_2 \Rightarrow 1 < RR_{1vs2} < OR_{1vs2}$$
$$\pi_1 < \pi_2 \Rightarrow 1 > RR_{1vs2} > OR_{1vs2}$$

If π_2 is small or RR_{1vs2} is close to 1,
then the last term is close to 1,
so OR_{1vs2} will be close to RR_{1vs2}

42

The limitations of the RD and the RR

$RD_{1vs2} = \pi_1 - \pi_2$ $\qquad RR_{1vs2} = \frac{\pi_1}{\pi_2}$

The Risk Difference cannot obtain all values:

For fixed value of π_2 we have: $-\pi_2 \le RD_{1vs2} \le 1-\pi_2$

if $\pi_2 = 5\%$ then the risk difference must be larger than -5%

if $\pi_2 = 95\%$ then the risk difference must be smaller than 5%

The Risk Ratio cannot obtain all values:

For fixed value of π_2 we have: $0 \le RR_{1vs2} \le \frac{1}{\pi_2}$

if $\pi_2 = 95\% = 0.95$
then the risk ratio must be smaller than $1.05 = 1/0.95$

i.e. **RR can be very restricted for frequent events.**

43

The limitations of the RD and the RR models

Per definition the risk (probality) is limited to the interval zero to one.

A RD model can violate this:

$$\Pr(obese) = \beta_0 + \beta_1 \cdot (scl - 200) + \beta_2 \cdot woman + \beta_3 \cdot (40 \le age < 50) + \beta_4 \cdot (50 \le age)$$

If β_1 is positive then
very large values of `scl` will give risks higher than one
very small values of `scl` will give negative risks

A RR model can also violate this:

$$\ln\left[\Pr(obese)\right] = \beta_0 + \beta_1 \cdot (scl - 200) + \beta_2 \cdot woman + \beta_3 \cdot (40 \le age < 50) + \beta_4 \cdot (50 \le age)$$

If β_1 is positive then very large values of `scl`
will give positive log-risks, i.e. risk larger than 1.

44

The limitations of the RD and the RR models

RD and RR models can be difficult or impossible to fit to a data set:

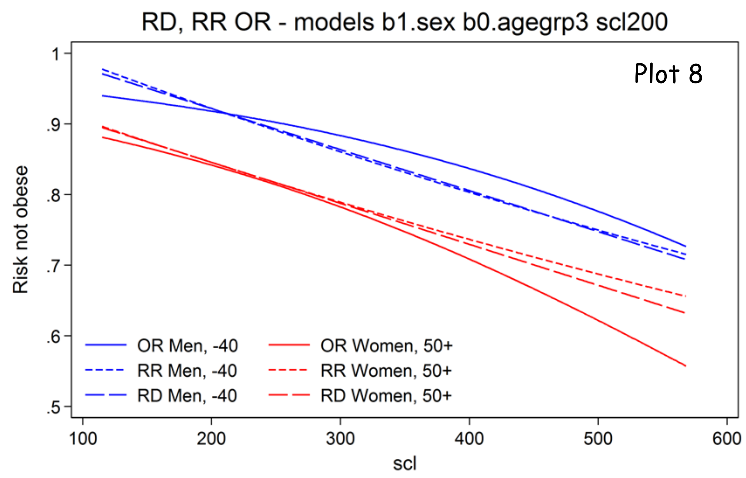
```
generate notobese=1-obese
binreg notobese  b1.sex b0.agegrp3 scl200,rr

Iteration 1:   deviance =  14758.67
Iteration 2:   deviance =  3530.169
.
.
.
Iteration 154: deviance =  3509.491
Iteration 155: deviance =   3509.49
--Break--
```

This trick might solve the "convergence" problem:

```
binreg notobese  b1.sex b0.agegrp3 scl200,or
predict pr, mu
binreg notobese  b1.sex b0.agegrp3 scl200,rr mu(pr)
```

45



46

Comparing two models: the likelihood ratio test

Until now we have used `testparm` to test if several coefficients could be zero.

In a "normal" regression model `testparm` will give a **exact F-test**.

In all other models, including binary regression models, `testparm` will results in what is so-called a **Wald-test**, which is approximative/"asymptotic" test.

```
binreg obese  b1.sex b0.agegrp3 scl200,rr
```

```
testparm i.agegrp3
 ( 1)  1.agegrp3 = 0
 ( 2)  2.agegrp3 = 0
           chi2(  2) =    25.46
         Prob > chi2 =    0.0000
```

```
testparm i.agegrp3 scl200
( 1)  1.agegrp3 = 0
( 2)  2.agegrp3 = 0
( 3)  scl200 = 0
       chi2(  3) =    53.76
     Prob > chi2 =    0.0000
```

An often used asymptotic test is the likelihood ratio test.

47

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:

```
binreg obese  b1.sex b0.agegrp3 scl200,rr ml
estimates store Modelrr1
binreg obese  b1.sex            scl200,rr ml
estimates store Modelrr2
lrtest Modelrr1 Modelrr2
```

Output:

```
Likelihood-ratio test                          LR chi2(2)  =     24.92
(Assumption: Modelrr2 nested in Modelrr1)      Prob > chi2 =    0.0000
```

i.agegrp3 adds **statistical significant** information to the model containing sex and scl smoking!

48

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:

```
binreg obese  b1.sex b0.agegrp3 scl200,rr ml
estimates store Modelrr1

binreg obese  b1.sex                  ,rr ml
estimates store Modelrr3

lrtest Modelrr1 Modelrr3
```

Output:

```
observations differ: 4658 vs. 4690
```

49

Comparing two models: the likelihood ratio test

```
estimates table Modelrr*,stats(N ll)
```

Variable	Modelrr1	Modelrr2	Modelrr3
sex			
Men	(base)	(base)	(base)
Women	.22330342	.2309925	.2508517
agegrp3			
0	(base)		
1	.07139944		
2	.43793034		
scl200	.00304885	.00416867	
_cons	-2.4947825	-2.3300943	-2.2035956
N	4658	4658	4690
ll	-1749.8636	-1762.3225	-1790.3703

The model without scl is fitted to a larger data set.
The results cannot be compared!!!

50

Comparing two models: the likelihood ratio test

Likelihood ratio test safe method

(All models fitted to the same data):

```
quietly: binreg obese  b1.sex b0.agegrp3 scl200,rr ml
estimates store Modelrr1

generate inmodel1=e(sample)

quietly: binreg obese  b1.sex  scl200 if inmodel1 ,rr ml
estimates store Modelrr2

quietly: binreg obese  b1.sex         if inmodel1, rr ml
estimates store Modelrr3

lrtest Modelrr1 Modelrr2
Likelihood-ratio test                        LR chi2(2)  = 24.92
(Assumption: Modelrr2 nested in Modelrr1)    Prob > chi2 = 0.0000

lrtest Modelrr1 Modelrr3
Likelihood-ratio test                        LR chi2(3)  =  53.72
(Assumption: Modelrr3 nested in Modelrr1)    Prob > chi2 = 0.0000
```

51

Comparing two models: the likelihood ratio test

```
estimates table Modelrr*,stats(N ll)
```

Variable	Modelrr1	Modelrr2	Modelrr3
sex			
Men	(base)	(base)	(base)
Women	.22330342	.2309925	.24350715
agegrp3			
0	(base)		
1	.07139944		
2	.43793034		
scl200	.00304885	.00416867	
_cons	-2.4947825	-2.3300943	-2.2001701
N	4658	4658	4658
ll	-1749.8636	-1762.3225	-1776.7225

52

The assumptions:

The model:

$$f\left(risk\right) = \beta_0 + \beta_1 \cdot \left(scl - 200\right) + \beta_2 \cdot woman + \beta_3 \cdot (40 \le age < 50) + \beta_4 \cdot (50 \le age)$$

Note:
We model the probability, some no room for additional random variation, - **no unexplained deviations**.

Two assumptions:
1. Linearity

2. Independency

Checking independency :
Just like in the normal case

53

The assumptions: Linearity

$$f\left(risk\right) = \beta_0 + \beta_1 \cdot \left(scl - 200\right) + \beta_2 \cdot woman + \beta_3 \cdot (40 \le age < 50) + \beta_4 \cdot (50 \le age)$$

The linearity can be decomposed in the sub-**assumptions**:

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

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

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

54

The assumptions ratio models: Linearity→multiplicativity

$$\Pr\left(obese\right) = \gamma_0 \cdot \gamma_1^{(scl-200)} \cdot \gamma_2^{woman} \cdot \gamma_3^{(40\le age<50)} \cdot \gamma_4^{(50\le age)}$$

$$Odds\left(obese\right) = \gamma_0 \cdot \gamma_1^{(scl-200)} \cdot \gamma_2^{woman} \cdot \gamma_3^{(40\le age<50)} \cdot \gamma_4^{(50\le age)}$$

The linearity can be decomposed in the sub-**assumptions**:

Multiplicativity on **risk /odds-scale**:
The contributions from sex and age are **multiplied**.

Exponential on **risk /odds-scale** :
The contribution from age is **raised to its value**.

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

55

Model checking

As there are no additional "random variation" there are no residuals, so you **cannot make any of the diagnostic plots** known from the normal regression models.

Model checking are typically done by **expanding** the model with interactions, cubic splines etc.

or looking at alternative way to introduce central variables.

In large data sets you can get some insight to the fit of the model by plotting observed frequencies against estimated probabilities in subgroups.

There exist many "statistics", like generalised r-squared, AUC-roc and Brier score, that measured the quality of an estimated model. But they will not give any insight what could be wrong with the model.

56

Binary regression models 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.

`binreg` can also be run with the option `ml`, this will give slightly different standard errors for RD and RR models

57

Things to look out for in the output

In general:

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

As a rule of thumb there should be at **least 15 events** per parameters in the model.

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

(for RD and RR it might help to using starting values from a OR – model).

58

OR regression = logistic regression

A OR regression model in usually called a logistic regression.

It can be fitted in Stata by `logit` or `logistics` command

```
binreg   obese  b1.sex b0.agegrp3 scl200,or
logit    obese  b1.sex b0.agegrp3 scl200,or
logistic obese  b1.sex b0.agegrp3 scl200
```

59

Logistic regression - why

Logistic regression is the most used binary regression model:

- It is always valid as it the probability always is between 0 and one.
- It was the first ever programmed.
  The option of RD or RR models in standard software is relative new.
- It can be used to used to analyse data from many types of case-control designs.
- It have done the job for many years!!????

60

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

61

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

Δx_1 in x_1 , Δx_2 in x_2 ... and Δx_k 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!

62

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

Δx_1 in x_1 , Δx_2 in x_2 ... and Δx_k in x_k

then the odds ratio is:

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

Note, the model might also be formulated:

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

63