Working with logistic regression models

Morten Frydenberg ©
Section 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

Extensions to the ordinary logistic regression:

- **Conditional logistic regression**
- **Other methods for analysing binary data**
  - Models for **relative risks**
  - Models for **risk differences**

Missing data

- A small example – non completely random sample
- Complete data analysis – bias
- Missing at random vs missing **completely** at random
- Introduction to techniques
  - Sampling weights
  - Imputation
  - Full modelling
- Sensitivity analyses

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

Clustered binary data with **one random component**

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

Consider the model:

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

```
       obese |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |
          1  |  (base)
          2  |   .2743976   .0903385     3.04   0.002      .0973374    .4514579
             |
       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`

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

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

```
 lincom _cons+2.sex

 ( 1)  [obese]2.sex + [obese]_cons = 0

------------------------------------------------------------------------------
       obese |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -1.872658    .058136   -32.21   0.000     -1.986602   -1.758714
------------------------------------------------------------------------------
```

To get to the risk/probability with confidence interval:

```
disp invlogit(r(estimate))
.13323448

disp invlogit(r(estimate)-1.96*r(se)) ";“ ///
     invlogit(r(estimate)+1.96*r(se))
.12061656 ;  .1469518
```

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

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

Some examples:

Log Odds for a 42 year old woman:

```
lincom _cons+2.sex-age45*3
 ( 1)  [obese]2.sex - 3*[obese]age45 + [obese]_cons = 0
------------------------------------------------------------------------------
       obese |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -1.976075   .0639755   -30.89   0.000     -2.101465   -1.850685
------------------------------------------------------------------------------
```

Odds ratio for 4.5 age difference:

```
 lincom age45*4.5,or
 ( 1)  4.5*[obese]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: 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 **15 events** per variable (parameter) in the model.

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

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

An exact method exists.

But it will also give wide confidence intervals.

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 more complicated **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 Day 4

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

```
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]
---------+--------------------------------------------------------------------
     sex |
      1  |  (base)
      2  |   .2743976   .0903385     3.04   0.002      .0973374    .4514579
         |
   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: 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**):

estat gof, *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

```
logit obese i.sex##age45
 estat gof, 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 it this way !!!!!!!

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. the 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

A study of cancer in the oral cavity

Matched on **gender** and **10-year 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*

`logit cancer textile i.alkcon i.genage`

Part of the output:

```
------------------------------------------------------------------------------
      cancer |      Coef.   Std. Err.      z    P>|z|            CI
-------------+----------------------------------------------------------------
     textile |   .5022796   .4141317     1.21   0.225    -.3094036    1.313963
      alkcon |
          0  |  (base)
          1  |   .4628618   .2823836     1.64   0.101    -.0905998    1.016323
          2  |   2.716577    .323265     8.40   0.000     2.082989    3.350165
      genage |
          1  |  (base)
          2  |    .245086   1.251388     0.20   0.845     -2.20759    2.697762
          3  |  -.4940138   .5503273    -0.90   0.369    -1.572635    .5846079
          4  |    .179786   .6406249     0.28   0.779    -1.075816    1.435388
          5  |  -.2899853   .5482076    -0.53   0.597    -1.364452    .7844818
          6  |   .2127169   .6262462     0.34   0.734    -1.014703    1.440137
          7  |  -.2305881   .5355411    -0.43   0.667    -1.280229    .8190532
          8  |   .5507988   .5263922     1.05   0.295    -.4809109    1.582509
          9  |   .0315165   .5884123     0.05   0.957     -1.12175    1.184783
         10  |   .5572024   .5595749     1.00   0.319    -.5395442    1.653949
       _cons |  -1.469219    .476301    -3.08   0.002    -2.402752   -.5356865
------------------------------------------------------------------------------
```

Conditional logistic regression in *Stata*

The syntax:

`clogit cancer textile i.alkcon,group(genage)`

Part of the output:

```
------------------------------------------------------------------------------
      cancer |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     textile |   .4929143    .410305     1.20   0.230    -.3112687    1.297097
      alkcon |
          0  |  (base)
          1  |    .452672   .2792327     1.62   0.105     -.094614     .999958
          2  |   2.660894   .3193692     8.33   0.000     2.034942    3.286846
------------------------------------------------------------------------------
```

`clogit cancer textile i.alkcon, group(genage) or`

```
------------------------------------------------------------------------------
      cancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     textile |    1.63708   .6717022     1.20   0.230      .732517    3.658661
      alkcon |
          0  |  (base)
          1  |   1.572508   .4390957     1.62   0.105      .909724    2.718168
          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
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
Estimatong RR or RD models

The Relative Risk models and the Risk Difference models can be estimated in many programs using what is called **Generalized** (not general) **Linear Models**.

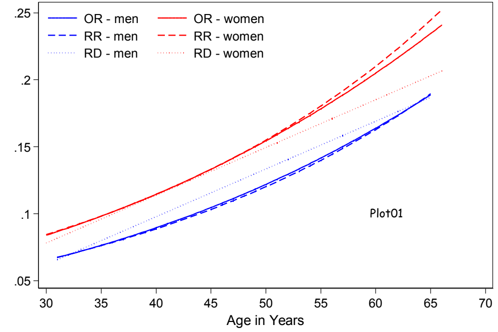
In Stata this is most easily done by the `binreg` command with the option `rr` og `rd`.

But be careful – estimation procedure might not work/converges, as

- the risk of the event in a RR-model is **not restricted** to be below one.
- the risk of the event in a RD-model is **not restricted** to be positive or below one.

Other methods to analysis of binary response data

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



Missing data – example 1

Consider the Frammingham study and imagine, that (due to a limited budget) only 500 measurements of SBP were allowed.

It was decided to take SBP measurements on **100** random participants in each of the age groups -40 and 60+ and **150** in each of the age groups 40-50 and 50-60.

That is we have missing SBP on 4190 of the 4,690 participants!

A short description of the design and the data:

```
----------------------------------------------------------
   agegrp |      Freq.     N(sbp)  mean(sbp)     sd(sbp)
----------+-----------------------------------------------
       0- |      1,325        100     122.18     15.4327
      40- |      1,684        150     130.85     22.2366
      50- |      1,346        150     140.93     22.4819
      60- |        335        100     149.51     26.9251
          |
    Total |      4,690        500     135.87     24.0783
----------------------------------------------------------
```

Missing data – example 1

```
----------------------------------------------------------
   agegrp |      Freq.     N(sbp)  mean(sbp)     sd(sbp)
----------+-----------------------------------------------
       0- |      1,325        100     122.18     15.4327
      40- |      1,684        150     130.85     22.2366
      50- |      1,346        150     140.93     22.4819
      60- |        335        100     149.51     26.9251
          |
    Total |      4,690        500     135.87     24.0783
----------------------------------------------------------
```

We note:

This is not a **completely** random sample
– the chance of being sample depends on age group!

The overall (total) average SBP is a biased estimate of the mean SBP among participants in the Frammingham study!

I.e. an analysis of the 500 participants (a complete data analysis) will be biased.

Missing data – example 1

```
----------------------------------------------------------
   agegrp |      Freq.     N(sbp)  mean(sbp)     sd(sbp)
----------+-----------------------------------------------
       0- |      1,325        100     122.18     15.4327
      40- |      1,684        150     130.85     22.2366
      50- |      1,346        150     140.93     22.4819
      60- |        335        100     149.51     26.9251
          |
    Total |      4,690        500     135.87     24.0783
----------------------------------------------------------
```

We also note:

Within each age group the sample is **completely** random.

Within each age group the average SBP is an **unbiased** estimate of the mean SBP in the age group.

We know the size of each age group.

We can **calculate an unbiased** estimate of the total mean by weighing the group averages.

Missing data – example 1

```
------------------------------------------------------------
    agegrp |      Freq.      N(sbp)   mean(sbp)      sd(sbp)
-----------+------------------------------------------------
        0- |      1,325         100      122.18      15.4327
       40- |      1,684         150      130.85      22.2366
       50- |      1,346         150      140.93      22.4819
       60- |        335         100      149.51      26.9251
           |
     Total |      4,690         500      135.87      24.0783
------------------------------------------------------------
```

An unbiased estimate can be found as the **weighted average** of the group averages using the group sizes as weights:

$$\frac{122.18\cdot 1325+130.85\cdot 1684+140.93\cdot 1346+149.51\cdot 335}{4690}=132.62$$

Conclusion: Although this is not a completely random sample, we have enough information in the data to find an unbiased estimate!!!!
(Assuming completely random sample **within** age group!)

Assuming that SBP is related to age:

Being missing is **not independent** of the **unobserved** SBP.

but

Being missing is **independent** of the unobserved SBP,
when we know the age group of the individual.

The first statement means that the data is not **missing completely at random (MCAR)**.

The second statement corresponds to **missing at random (MAR)**, i.e. that given **all what we have observed** (including age group), then the missingness is (completely) random, i.e. independent of the unobserved data.

Mathematically Missing At Random implies that one (in theory) has enough information in the **observed data** to correct for the missing data – in principle.

Missing data: Standard terminology

Missing completely at random (MCAR).
The observed data is a (completely) random sample:
A complete data analysis will be unbiased

Missing at random (MAR)
Given **all what we have observed**, then the missingness is (completely) random (independent of the unobserved data):
The biased sampling **might be adjusted for**.

Missing not at random (MNAR)
Non of the two above apply:
We will need further assumptions in order to analyse the data.

Missing at random

When the data is **missing at random**, then one can, in theory, make unbiased inference based on the observed data.

In the SBP example such an analysis could be to use the **weighted average** SBP instead of the biased unweighted average.

In general

If the sampled persons are not a completely random sample, but the *i*th person is sampled with a **known** probability, p_i , then we can obtain unbiased estimates by weighing the *i*th person with $1/p_i$.

The method is called Inverse Probability Weighing.

Inverse probability weighting

The SBP data:
Four different sampling probabilities and weights:

$$p_0 = 100/1325 = 0.0755 \qquad w_0 = 1/p_0 = 13.25$$

$$p_1 = 150/1684 = 0.0891 \qquad w_1 = 1/p_1 = 11.23$$

$$p_2 = 150/1346 = 0.1114 \qquad w_2 = 1/p_2 = 8.97$$

$$p_3 = 100/335 = 0.2985 \qquad w_3 = 1/p_3 = 3.35$$

That is information from each of the youngest should weight by 13.25 and information from the each of the oldest should weight by 3.35.
Sampling weights can be used in many Stata commands:

```
 mean sbp [pw= sampw]
Mean estimation                     Number of obs    =     500
--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
         sbp |   132.6242   1.032943      130.5947    134.6536
--------------------------------------------------------------
```

Missing values – not by design

Most often the missing is **not per design**
and both in the **outcome** and in the **covariates**:

id	*y*	x_1	x_2	x_3
1	o	o	o	o
2	o	m	o	o
3	m	o	o	o
4	m	m	o	o
5	o	o	o	o
6	o	m	m	o

o observed
m observed

Here we have only **complete data** on 2 persons, but partial information on 4 additoinal persons.

Missing values – not by design

If the missing is **completely at random**, then the analysis of the complete cases will be unbiased.

If this is not the case, then complete data analysis can give biased estimates.

If the data is **missing at random**, then it is **in theory** possible to make an unbiased analysis of all the data.

id	*y*	x_1	x_2	x_3
1	o	o	o	o
2	o	m	o	o
3	m	o	o	o
4	m	m	o	o
5	o	o	o	o
6	o	m	m	o

Imputation

One way to try solve the problem with missing is to **fill in** the data for the missing values and then make the analysis on the whole data set with the **'imputed'** values.

The imputation can be done in many ways.

One way is to fill in an "average" value.

id	*y*	x_1	x_2	x_3
1	o	o	o	o
2	o	a_1	o	o
3	a_y	o	o	o
4	a_y	a_1	o	o
5	o	o	o	o
6	o	a_1	a_2	o

This could be the total average of the observed values for the specific variable or the average in a **relevant subgroup**.

This method will not in general solve the bias problem.

And of course the **standard error** stated in the output, when you analyse the imputed data set, is **wrong**.

The missing SBP example

Imputation by **observed mean** in age group:

```
bysort agegrp: egen msbp=mean(sbp)
generate isbp=sbp
replace isbp=msbp if missing(sbp)

mean isbp
Mean estimation                     Number of obs    =    4690
--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
        isbp |   132.6242   .1627486      132.3051    132.9432
--------------------------------------------------------------
```

Correct mean, but a much too small standard error – incorrectly **assuming 4690 independent observations**.

Correct analysis using sampling weights:

```
mean sbp [pw=sampw]
Mean estimation                     Number of obs    =     500
--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
         sbp |   132.6242   1.032943      130.5947    134.6536
--------------------------------------------------------------
```

Imputation – random multiple

A fixed imputation will not take into account the random variation of the unobserved observation or the uncertainty of the parameters.

id	*y*	x_1	x_2	x_3
1	o	o	o	o
2	o	m	o	o
3	m	o	o	o
4	m	m	o	o
5	o	o	o	o
6	o	m	m	o

Imputation methods should add some random variation to the imputed data.

For that we need a **statistical model** for the missing data.

In **multiple imputations** one generates **several imputed** data sets.

For each imputed data set one fit the model of interest.

The point estimate, then the average across the imputed data sets.

One tricky thing is **calculation of the standard errors**.

Multiple imputations

Questions:

How to find **the models** from which to generate the missing data?

How should you handle missing data in this process?

How to find the uncertainty (**standard errors**) of the estimates?

Bookkeeping.

id	*y*	x_1	x_2	x_3
1	o	o	o	o
2	o	m	o	o
3	m	o	o	o
4	m	m	o	o
5	o	o	o	o
6	o	m	m	o

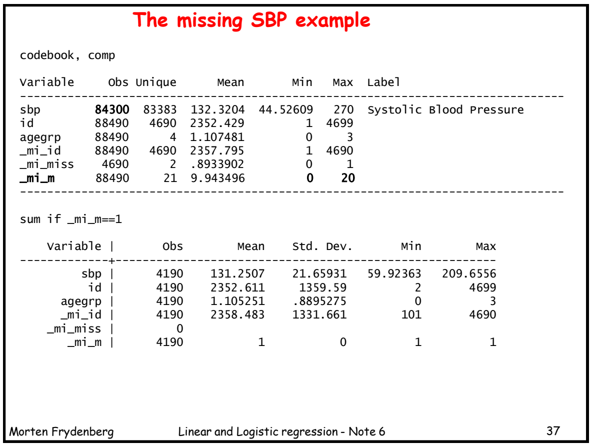
Most important: **Missing at random is required**!

The missing SBP example

```
use sbpdata,clear
mi set mlong
mi register imputed sbp
(4190 m=0 obs. now marked as incomplete)

 mi impute regress sbp i.agegrp, add(20)

Univariate imputation                      Imputations =       20
Linear regression                                added =       20
Imputed: m=1 through m=20                      updated =        0
              |               Observations per m
              |----------------------------------------------
     Variable |   complete   incomplete   imputed |     total
--------------+-----------------------------------+----------
          sbp |        500         4190      4190 |      4690
-------------------------------------------------------------
(complete + incomplete = total; imputed is the minimum across m
 of the number of filled in observations.)
```



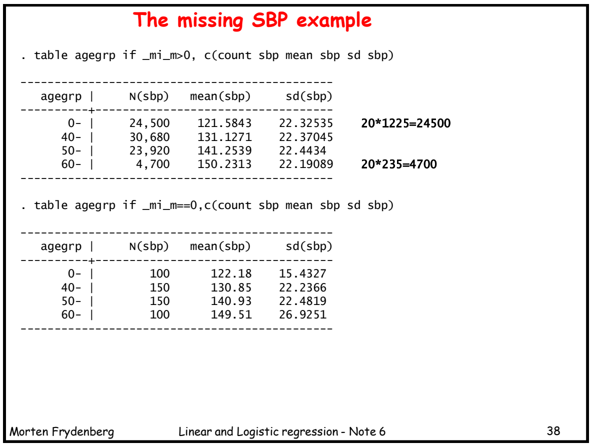
The missing SBP example

```
codebook, comp

Variable      Obs Unique      Mean       Min    Max  Label
--------------------------------------------------------------------------------
sbp         84300  83383  132.3204  44.52609    270  Systolic Blood Pressure
id          88490   4690  2352.429         1   4699
agegrp      88490      4  1.107481         0      3
_mi_id      88490   4690  2357.795         1   4690
_mi_miss     4690      2  .8933902         0      1
_mi_m       88490     21  9.943496         0     20
--------------------------------------------------------------------------------

sum if _mi_m==1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         sbp |      4190    131.2507    21.65931   59.92363   209.6556
          id |      4190    2352.611     1359.59          2       4699
      agegrp |      4190    1.105251    .8895275          0          3
      _mi_id |      4190    2358.483    1331.661        101       4690
    _mi_miss |         0
       _mi_m |      4190           1           0          1          1
```



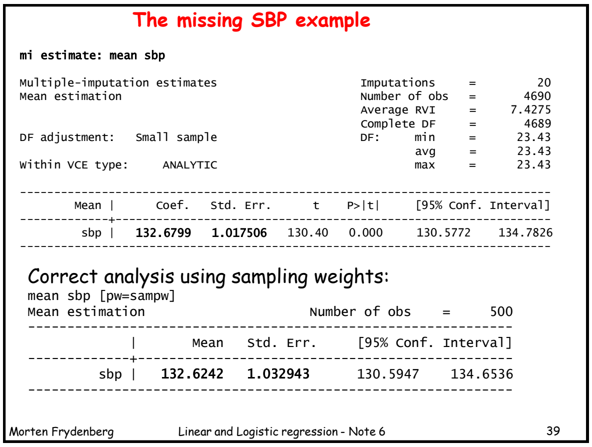
The missing SBP example

```
. table agegrp if _mi_m>0, c(count sbp mean sbp sd sbp)

----------------------------------------------
   agegrp |     N(sbp)   mean(sbp)     sd(sbp)
----------+-----------------------------------
       0- |     24,500    121.5843    22.32535     20*1225=24500
      40- |     30,680    131.1271    22.37045
      50- |     23,920    141.2539    22.4434
      60- |      4,700    150.2313    22.19089     20*235=4700
----------------------------------------------

. table agegrp if _mi_m==0,c(count sbp mean sbp sd sbp)

----------------------------------------------
   agegrp |     N(sbp)   mean(sbp)     sd(sbp)
----------+-----------------------------------
       0- |        100      122.18    15.4327
      40- |        150      130.85    22.2366
      50- |        150      140.93    22.4819
      60- |        100      149.51    26.9251
----------------------------------------------
```



The missing SBP example

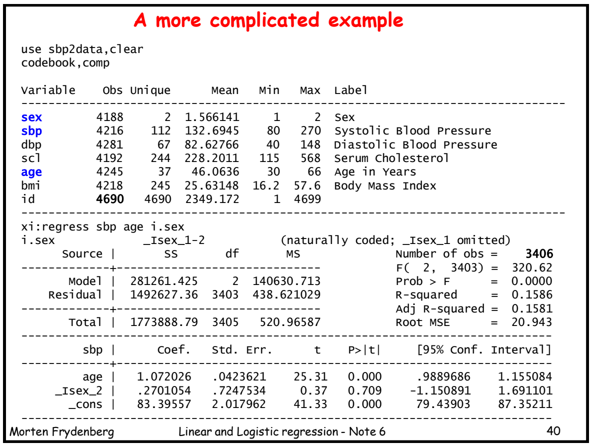
```
mi estimate: mean sbp

Multiple-imputation estimates                     Imputations     =         20
Mean estimation                                   Number of obs   =       4690
                                                  Average RVI     =     7.4275
                                                  Complete DF     =       4689
DF adjustment:   Small sample                     DF:      min    =      23.43
                                                           avg    =      23.43
Within VCE type:     ANALYTIC                              max    =      23.43

------------------------------------------------------------------------------
        Mean |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sbp |   132.6799   1.017506   130.40   0.000     130.5772    134.7826
------------------------------------------------------------------------------
```

Correct analysis using sampling weights:

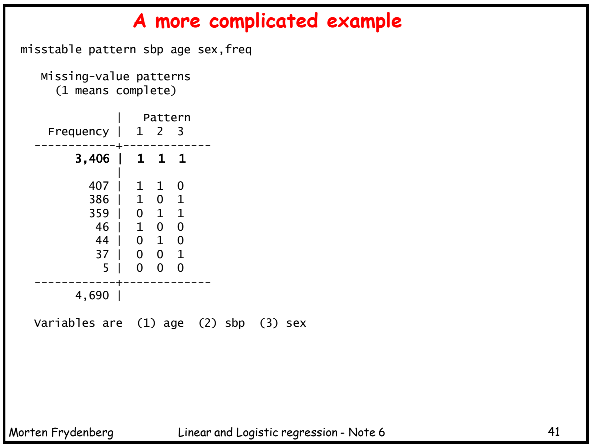
```
mean sbp [pw=sampw]
Mean estimation                     Number of obs    =     500
--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
         sbp |   132.6242   1.032943      130.5947    134.6536
--------------------------------------------------------------
```



A more complicated example

```
use sbp2data,clear
codebook,comp

Variable     Obs Unique      Mean   Min   Max  Label
--------------------------------------------------------------------------------
sex         4188      2  1.566141     1     2  Sex
sbp         4216    112  132.6945    80   270  Systolic Blood Pressure
dbp         4281     67  82.62766    40   148  Diastolic Blood Pressure
scl         4192    244  228.2011   115   568  Serum Cholesterol
age         4245     37   46.0636    30    66  Age in Years
bmi         4218    245  25.63148  16.2  57.6  Body Mass Index
id          4690   4690  2349.172     1  4699
--------------------------------------------------------------------------------
xi:regress sbp age i.sex
i.sex             _Isex_1-2          (naturally coded; _Isex_1 omitted)
      Source |       SS       df       MS              Number of obs =    3406
-------------+------------------------------           F(  2,  3403) =  320.62
       Model |  281261.425     2  140630.713           Prob > F      =  0.0000
    Residual |  1492627.36  3403  438.621029           R-squared     =  0.1586
-------------+------------------------------           Adj R-squared =  0.1581
       Total |  1773888.79  3405   520.96587           Root MSE      =  20.943
------------------------------------------------------------------------------
         sbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.072026   .0423621    25.31   0.000     .9889686    1.155084
     _Isex_2 |   .2701054   .7247534     0.37   0.709    -1.150891    1.691101
       _cons |   83.39557   2.017962    41.33   0.000     79.43903    87.35211
------------------------------------------------------------------------------
```



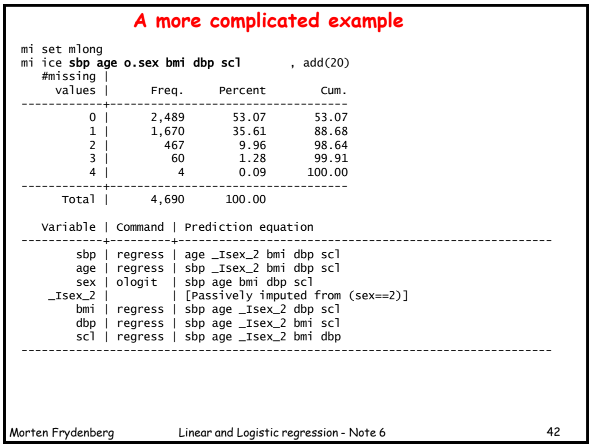
A more complicated example

```
misstable pattern sbp age sex,freq

   Missing-value patterns
     (1 means complete)

              |   Pattern
    Frequency |  1  2  3
  ------------+-------------
        3,406 |  1  1  1
              |
          407 |  1  1  0
          386 |  1  0  1
          359 |  0  1  1
           46 |  1  0  0
           44 |  0  1  0
           37 |  0  0  1
            5 |  0  0  0
  ------------+-------------
        4,690 |

  Variables are  (1) age  (2) sbp  (3) sex
```



A more complicated example

```
mi set mlong
mi ice sbp age o.sex bmi dbp scl       , add(20)
   #missing |
     values |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      2,489       53.07       53.07
          1 |      1,670       35.61       88.68
          2 |        467        9.96       98.64
          3 |         60        1.28       99.91
          4 |          4        0.09      100.00
------------+-----------------------------------
      Total |      4,690      100.00

   Variable | Command | Prediction equation
------------+---------+-------------------------------------------------------
        sbp | regress | age _Isex_2 bmi dbp scl
        age | regress | sbp _Isex_2 bmi dbp scl
        sex | ologit  | sbp age bmi dbp scl
    _Isex_2 |         | [Passively imputed from (sex==2)]
        bmi | regress | sbp age _Isex_2 dbp scl
        dbp | regress | sbp age _Isex_2 bmi scl
        scl | regress | sbp age _Isex_2 bmi dbp
------------------------------------------------------------------------------
```

A more complicated example

```
codebook,comp

Variable      Obs Unique      Mean       Min       Max  Label
-------------------------------------------------------------------------------
sex         48208      2  1.568682         1         2  Sex
sbp         48236   9585  132.3171  55.04445       270  Systolic Blood Pressure
dbp         48301   8239  82.44462  39.00607       148  Diastolic Blood Pressure
scl         48212  10200  227.2202  71.84563       568  Serum Cholesterol
age         48265   8932  45.94714  14.28921  83.50232  Age in Years
bmi         48238   9679  25.52701  10.58046      57.6  Body Mass Index
id          48710   4690  2348.166         1      4699
_mi_id      48710   4690  2330.321         1      4690
_mi_miss     4690      2  .4692964         0         1
_mi_m       48710     21  9.489017         0        20
-------------------------------------------------------------------------------
```

A more complicated example

```
 mi estimate: regress sbp age sex

Multiple-imputation estimates                   Imputations       =         20
Linear regression                               Number of obs     =       4690
                                                Average RVI       =     0.1115
                                                Complete DF       =       4687
DF adjustment:   Small sample                   DF:     min       =     784.98
                                                        avg       =     982.49
                                                        max       =    1366.36
Model F test:       Equal FMI                   F(   2, 1480.0) =     397.31
Within VCE type:          OLS                   Prob > F          =     0.0000

------------------------------------------------------------------------------
         sbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.074694   .0376721    28.53   0.000     1.000792    1.148595
         sex |   .2725589   .6618376     0.41   0.681    -1.026622     1.57174
       _cons |    82.8989   2.061978    40.20   0.000     78.85135    86.94646
------------------------------------------------------------------------------
```

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:

$$\operatorname{logit}\left(H_{fpd}=1\right)=\overbrace{\boxed{\beta_0+\beta_I\cdot I+\beta_U\cdot U+\beta_A\cdot A+\beta_S\cdot S+\beta_G\cdot G}}^{\textbf{Systematic}\text{ part}}$$
$$\underbrace{\boxed{+F_f+P_{fp}+\cancel{E_{fpd}}}}_{\textbf{Random}\text{ part}}$$

This is not needed due to the binomial error

Clustered data / data with several random components
Dichotomous outcome

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

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 to 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 involve **one random component**, e.g. **variation between families** or between **GP's**,

then methods exist which can **adjust the standard errors**.

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

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

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