

Working with logistic 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

Extensions to the ordinary logistic regression:
Conditional logistic regression

Other methods for analysing binary data
Models for relative risks
Models for risk differences

Morten Frydenberg

Linear and Logistic regression - Note 6

1

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

Morten Frydenberg

Linear and Logistic regression - Note 6

2

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

Clustered binary data with **one** random component

ROC-curves and the area under the ROC-curve

Morten Frydenberg

Linear and Logistic regression - Note 6

3

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

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

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.....

Morten Frydenberg

Linear and Logistic regression - Note 6

4

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

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

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
-----+-----
```

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
```

Morten Frydenberg

Linear and Logistic regression - Note 6

5

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

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

Some examples:

Log Odds for a 42 year old woman:

```
lincom _cons+_Isex-age45*3
( 1)  _Isex_2 - 3 age45 + _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 age45 = 0
-----
obese | Odds Ratio   Std. Err.      z    P>|z|    [95% Conf. Interval]
-----+-----
(1) |  1.167804   .0209869    6.71   0.000    1.116091    1.221914
-----+-----
```

Morten Frydenberg

Linear and Logistic regression - Note 6

6

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 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 only one (**expensive**) program can do it.

And it will also give wide confidence intervals.

Morten Frydenberg Linear and Logistic regression - Note 6 7

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.**

Morten Frydenberg Linear and Logistic regression - Note 6 8

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

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

Logit estimates

Log likelihood = -1767.7019						Number of obs = 4690	
						LR chi2(2) = 55.68	
						Prob > chi2 = 0.0000	
						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?

Morten Frydenberg Linear and Logistic regression - Note 6 9

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.

Morten Frydenberg Linear and Logistic regression - Note 6 10

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!

Morten Frydenberg Linear and Logistic regression - Note 6 11

Logistic regression models: Test of fit

xi: 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 !!!!!

Morten Frydenberg Linear and Logistic regression - Note 6 12

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.

Morten FrydenbergLinear and Logistic regression - Note 613

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 $\ln(odds)$ ratios
different in each strata the same in each strata

Morten FrydenbergLinear and Logistic regression - Note 614

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**.

Morten FrydenbergLinear and Logistic regression - Note 615

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.

Morten FrydenbergLinear and Logistic regression - Note 616

Conditional logistic regression

How

An example using *Stata*

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)

Morten FrydenbergLinear and Logistic regression - Note 617

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
_1alkcon_1	.4628	.2823	1.639	0.101	-.0905 1.0163
_1alkcon_2	2.7165	.3232	8.404	0.000	2.0829 3.3501
_1genage_2	.2450	1.2514	0.196	0.845	-2.2075 2.6977
_1genage_3	-.4940	.5503	-0.898	0.369	-1.5726 .5846
_1genage_4	-.1798	.6406	0.281	0.779	-1.0758 1.4353
_1genage_5	-.2899	.5482	-0.529	0.597	-1.3644 .7844
_1genage_6	.2127	.6262	0.340	0.734	-1.0147 1.4401
_1genage_7	-.2305	.5355	-0.431	0.667	-1.2802 .8190
_1genage_8	.5507	.5263	1.046	0.295	-.4809 1.5825
_1genage_9	.0315	.5884	0.054	0.957	-1.1217 1.1847
_1genage_10	-.5572	.5595	0.996	0.319	-.53954 1.6539
_const	-1.4692	.4762	-3.085	0.002	-2.4027 .5356

Morten FrydenbergLinear and Logistic regression - Note 618

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
_alkcon_1	.452	.27923	1.621	0.105	-.094 .9999
_alkcon_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
_alkcon_1	1.572508	.4390957	1.62	0.105	.909724 2.718168
_alkcon_2	14.30908	4.569879	8.33	0.000	7.651811 26.75835

Morten FrydenbergLinear and Logistic regression - Note 619

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

Morten FrydenbergLinear and Logistic regression - Note 620

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

Morten FrydenbergLinear and Logistic regression - Note 621

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

Morten FrydenbergLinear and Logistic regression - Note 622

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

Morten FrydenbergLinear and Logistic regression - Note 623

Other methods to analysis of binary response data

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

Morten FrydenbergLinear and Logistic regression - Note 624

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** !

Here/in Stata glm is an acronym for **generalized** linear model

not
general linear model

Note: In Stata you can also use the **binreg** command with option **rr** or **rd**

Morten FrydenbergLinear and Logistic regression - Note 625

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 4690 participants!

A short description of the data:

agegrp	Freq.	N(sbp)	mean(sbp)	sd(sbp)
0-	1,325	100	122.18	15.43273
40-	1,684	150	130.8467	22.2366
50-	1,346	150	140.9267	22.48194
60-	335	100	149.51	26.92507
Total	4,690	500	135.87	24.0783

Morten FrydenbergLinear and Logistic regression - Note 626

Missing data - example 1

agegrp	Freq.	N(sbp)	mean(sbp)	sd(sbp)
0-	1,325	100	122.18	15.43273
40-	1,684	150	130.8467	22.2366
50-	1,346	150	140.9267	22.48194
60-	335	100	149.51	26.92507
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.

Morten FrydenbergLinear and Logistic regression - Note 627

Missing data - example 1

agegrp	Freq.	N(sbp)	mean(sbp)	sd(sbp)
0-	1,325	100	122.18	15.43273
40-	1,684	150	130.8467	22.2366
50-	1,346	150	140.9267	22.48194
60-	335	100	149.51	26.92507
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.

Morten FrydenbergLinear and Logistic regression - Note 628

Missing data - example 1

agegrp	Freq.	N(sbp)	mean(sbp)	sd(sbp)
0-	1,325	100	122.18	15.43273
40-	1,684	150	130.8467	22.2366
50-	1,346	150	140.9267	22.48194
60-	335	100	149.51	26.92507
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!)

Morten FrydenbergLinear and Logistic regression - Note 629

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.

Morten FrydenbergLinear and Logistic regression - Note 630

Linear and Logistic Regression: Note 6

5

The missing SBP example

mi estimate: mean sbp

Multiple-imputation estimates

Mean estimation

DF adjustment: Small sample

Within VCE type: ANALYTIC

Imputations = 20

Number of obs = 4690

Average RVI = 7.4275

Complete DF = 4689

DF: min = 23.43

avg = 23.43

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

Morten FrydenbergLinear and Logistic regression - Note 643

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
sc1	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

Source | SS df MS

Model | 281261.425 2 140630.713

Residual | 1492627.36 3403 438.621029

Total | 1773888.79 3405 520.96587

(naturally coded; _Isex_1 omitted)

Number of obs = 3406

F(2, 3403) = 320.62

Prob > F = 0.0000

R-squared = 0.1586

Adj R-squared = 0.1581

Root MSE = 20.943

	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

Morten FrydenbergLinear and Logistic regression - Note 644

A more complicated example

misstable pattern sbp age sex,freq

Missing-value patterns

(1 means complete)

Frequency	Pattern	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

Morten FrydenbergLinear and Logistic regression - Note 645

A more complicated example

mi set mlong

mi ice sbp age o.sex bmi dbp sc1 , 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 sc1

age | regress | sbp _Isex_2 bmi dbp sc1

sex | ologit | sbp age bmi dbp sc1

_Isex_2 | [Passively imputed from (sex==2)]

bmi | regress | sbp age _Isex_2 dbp sc1

dbp | regress | sbp age _Isex_2 bmi sc1

sc1 | regress | sbp age _Isex_2 bmi dbp

Morten FrydenbergLinear and Logistic regression - Note 646

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
sc1	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	

Morten FrydenbergLinear and Logistic regression - Note 647

A more complicated example

mi estimate: regress sbp age sex

Multiple-imputation estimates

Linear regression

DF adjustment: Small sample

Model F test: Equal FMI

within VCE type: OLS

Imputations = 20

Number of obs = 4690

Average RVI = 0.1115

Complete DF = 4687

DF: min = 784.98

avg = 982.49

max = 1366.36

F(2, 1480.0) = 397.31

Prob > F = 0.0000

	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

Morten FrydenbergLinear and Logistic regression - Note 648

Clustered data / data with several random components
Dichotomous outcome

A different outcome:

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

A statistical model:

$$\text{logit}(H_{jpd} = 1) = \beta_0 + \beta_I \cdot I + \beta_U \cdot U + \beta_A \cdot A + \beta_S \cdot S + \beta_G \cdot G + F_j + P_{jp} + X_{jm}$$

Systematic part

Random part

This is not needed due to the binomial error

Clustered data / data with several random components
Dichotomous outcome

$$\text{logit}(H_{jpd} = 1) = \beta_0 + \beta_I \cdot I + \beta_U \cdot U + \beta_A \cdot A + \beta_S \cdot S + \beta_G \cdot G + F_j + P_{jp}$$

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!

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

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

obese	over45	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

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

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



