

Extensions to linear and logistic regressions

Morten Frydenberg ©  
Institut for Biostatistik

Conditional logistic regression

- When?
- What?
- How?

Other methods for analyzing binary data

Models for relative risks

Models for risk differences

Clustered data / data with several random components

Continuous outcome

Dichotomous outcome

Clustered binary data with one random components

Nonlinear regression models

Morten Frydenberg      Linear and Logistic regression - Note 4.2      1

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 Frydenberg      Linear and Logistic regression - Note 4.2      2

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

Morten Frydenberg      Linear and Logistic regression - Note 4.2      3

Conditional logistic regression

What

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

↗

ln(odds) different in each strata

We are not interested in these !

In a matched study these are 'controlled'.

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

In the conditional model the  $\alpha$ 's disappear !

The  $\beta$ 's, the log OR's, are still in and can be estimated.

Morten Frydenberg      Linear and Logistic regression - Note 4.2      4

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 Frydenberg      Linear and Logistic regression - Note 4.2      5

Conditional logistic regression

How

An example using STATA

A study of cancer in the oral cavity

Matched on gender and 10 years age groups

Ten strata (genage)

Here we focus on

textile-worker and

life time consumption of alcohol (three groups)

Morten Frydenberg      Linear and Logistic regression - Note 4.2      6

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
_alkcon_1	.4628	.2823	1.639	0.101	-.0905 1.0163
_alkcon_2	2.7165	.3232	8.404	0.000	2.0829 3.3501
_igenage_2	.2450	1.2514	0.196	0.845	-2.2075 2.6977
_igenage_3	-.4940	.5503	-0.898	0.369	-1.5726 .5846
_igenage_4	.1798	.6406	0.281	0.779	-1.0758 1.4353
_igenage_5	-.2899	.5482	-0.529	0.597	-1.3644 .7844
_igenage_6	.2127	.6262	0.340	0.734	-1.0147 1.4401
_igenage_7	-.2305	.5355	-0.431	0.667	-1.2802 .8190
_igenage_8	.5507	.5283	1.046	0.295	-.4809 1.5825
_igenage_9	.0315	.5884	0.054	0.957	-1.1217 1.1847
_igenage_10	-.5572	.5595	-0.996	0.319	-1.6339 .5195
_const	-1.4692	.4762	-3.085	0.002	-2.4027 -.5356

Morten FrydenbergLinear and Logistic regression - Note 4.27

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 4.28

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 4.29

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 4.210

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 4.211

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

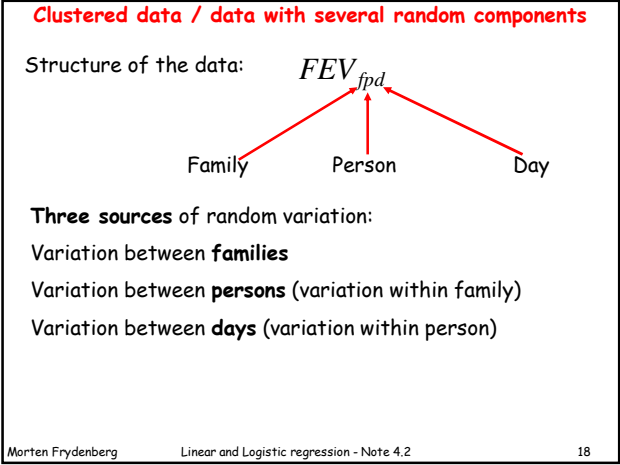
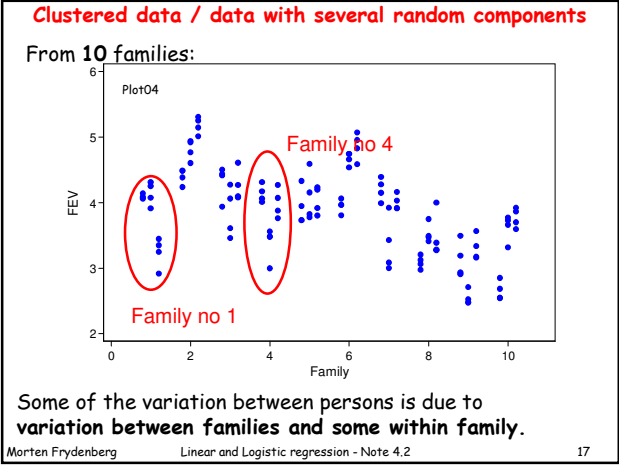
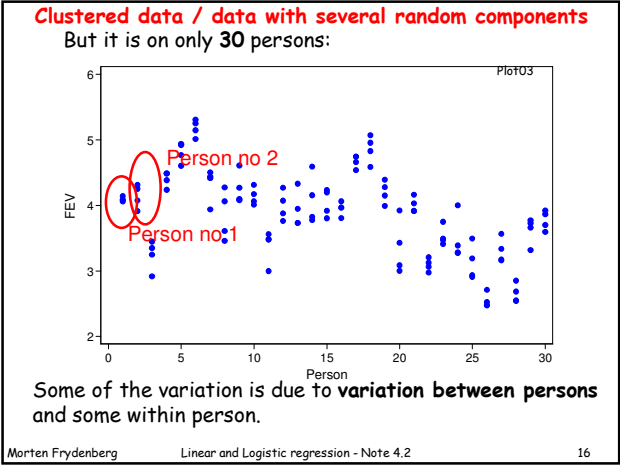
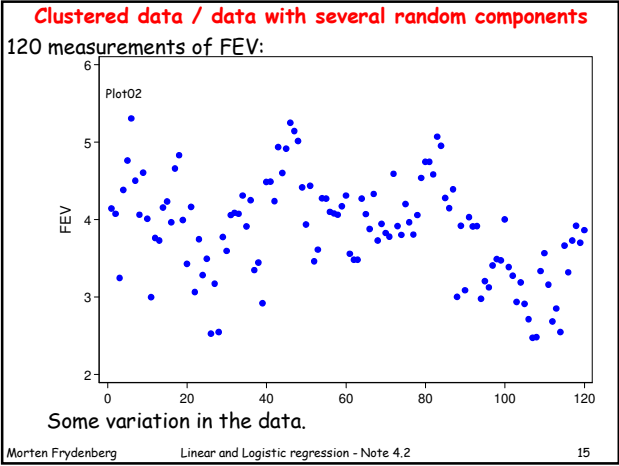
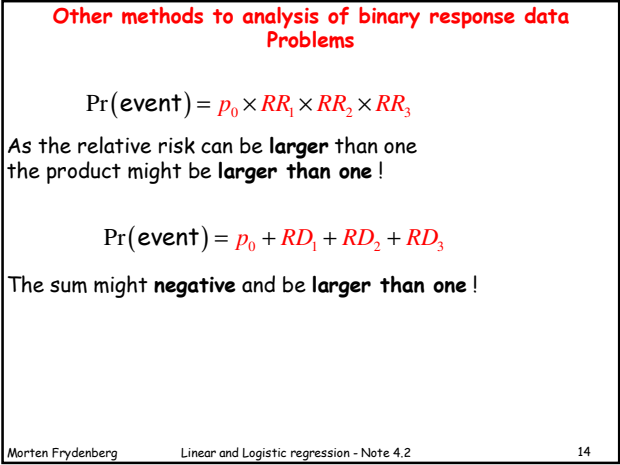
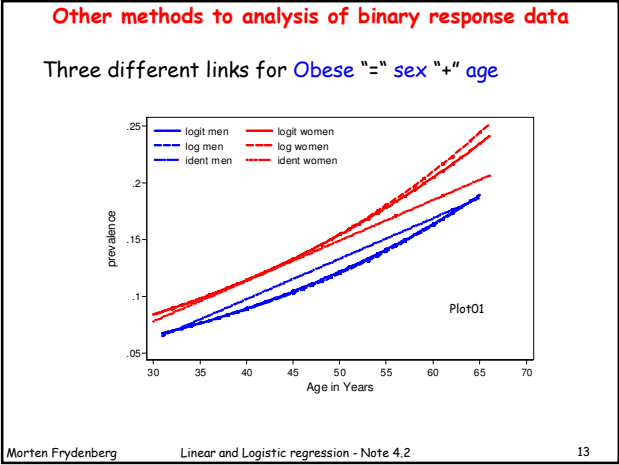
xi: glm obese age i.sex, fam(bin) link(logit)

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 4.212



Clustered data / data with several random components

Factors of interest:

household Income

Constant within family

Urbanization

Constant within family

Age

Constant within person; varies within family

Sex

Constant within person; varies within family

Grass pollen

Constant within day; varies within person

A model:

$$FEV = \beta_0 + \beta_I \cdot I + \beta_U \cdot U + \beta_A \cdot A + \beta_S \cdot S + \beta_G \cdot G$$

+random variation

Morten Frydenberg      Linear and Logistic regression - Note 4.2      19

Clustered data / data with several random components

$$FEV = \beta_0 + \beta_I \cdot I + \beta_U \cdot U + \beta_A \cdot A + \beta_S \cdot S + \beta_G \cdot G$$

+random variation

If the **three** levels/sources of **random** variation are **not** taken into account :

- The **precision** of the  $\beta_I$  and  $\beta_U$  are **highly overestimated**
- The **precision** of the  $\beta_A$  and  $\beta_S$  are **overestimated**
- The **estimates** of the  $\beta_I$  and  $\beta_U$  will be **biased** if the not all families are represented by the **same number of persons** and each person is measured the **same number of times**.
- The **estimates** of the  $\beta_A$  and  $\beta_S$  will be **biased** if the not all persons are measured the **same number of times**.

Morten Frydenberg      Linear and Logistic regression - Note 4.2      20

Clustered data / data with several random components

$$FEV = \beta_0 + \beta_I \cdot I + \beta_U \cdot U + \beta_A \cdot A + \beta_S \cdot S + \beta_G \cdot G$$

$+F_f + P_{fp} + E_{fpd}$

$F_f$

$P_{fp}$

$E_{fpd}$

: Random family contribution

: Random person contribution

: Random day contribution

$\sigma_F^2$

$\sigma_P^2$

$\sigma_E^2$

$$\text{var}(FEV_{fpd}) = \sigma_F^2 + \sigma_P^2 + \sigma_E^2$$

Variance components

Assumed to be normal distributed

Morten Frydenberg      Linear and Logistic regression - Note 4.2      21

Clustered data / data with several random components

Systematic part

$$FEV = \beta_0 + \beta_I \cdot I + \beta_U \cdot U + \beta_A \cdot A + \beta_S \cdot S + \beta_G \cdot G$$

Random part

$+F_f + P_{fp} + E_{fpd}$

$\beta_0, \beta_I, \beta_U, \beta_A, \beta_S$  and  $\beta_G$       Quantify the **systematic** variation

$\sigma_F^2, \sigma_P^2$  and  $\sigma_E^2$       Quantify the **random** variation

This is a:

- **Variance component** model
- **Mixed** model (both systematic and random variation)
- **Multilevel** model

The theory behind and the understanding of such models is well **established!!!**

Morten Frydenberg      Linear and Logistic regression - Note 4.2      22

Clustered data / data with several random components

Dichotomous outcome

A different outcome:

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

A statistical model:

Systematic part

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

Random part

$+F_f + P_{fp} + E_{fpd}$

This is not needed due to the binomial error

Morten Frydenberg      Linear and Logistic regression - Note 4.2      23

Clustered data / data with several random components

Dichotomous outcome

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

$+F_f + P_{fp}$

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.

Morten Frydenberg      Linear and Logistic regression - Note 4.2      24

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 not known how to obtain **valid (unbiased) estimates**.
- Several software programs **falsely claim** to estimate the models. (SAS, STATA, SUDAAN, NLwin)
- The programs/algorithms are not able give 'the correct' estimates.

Advice:

At the moment, do not trust results based on multilevel models.

Wait and see, the statisticians might solve the problems.

Morten Frydenberg      Linear and Logistic regression - Note 4.2      25

Clustered data / data with one random components  
Dichotomous outcome

If the models only involves **one random components**, e.g. **variation between families** or between GP's, then methods exists which can **adjust the standards errors**.

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

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

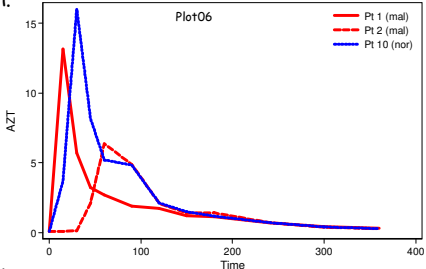
Only a **few** programs have this option - STATA does!

Morten Frydenberg      Linear and Logistic regression - Note 4.2      26

Nonlinear regression models

Concentration in the blood of zidovudine (AZT) after administration of the drug.

One person with normal fat absorption and two with malabsorption.



Clearly non linear.

Morten Frydenberg      Linear and Logistic regression - Note 4.2      27

Nonlinear regression models

The is **no way** that the above data can be described by a linear regression.

Furthermore **pharmaco kinetic** theory specify a simple model for the **expected concentration** as a function of time:

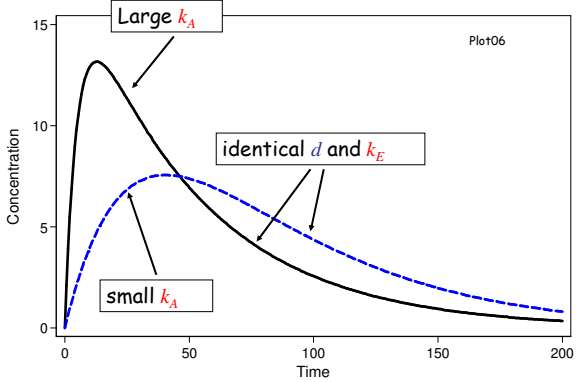
$$\text{concentration}(t) = d \frac{k_A}{k_A - k_E} (\exp(-k_E \cdot t) - \exp(-k_A \cdot t))$$

Where:

- $d$  : dose (per kg bodyweight)
- $k_A$  : absorptionsrate
- $k_E$  : eliminationsrate

Morten Frydenberg      Linear and Logistic regression - Note 4.2      28

Nonlinear regression models



Morten Frydenberg      Linear and Logistic regression - Note 4.2      29

Nonlinear regression models

How do persons with malabsorption differ from normal in  $k_E$  and  $k_A$  ?

One type of analysis:

1. Fit this pharmaco-kinetic model to the data for each person.
2. Extract the estimates of  $k_E$  and  $k_A$  from each analysis.
3. Compare the distributions of  $k_E$  and  $k_A$  from the two types for persons.

Morten Frydenberg      Linear and Logistic regression - Note 4.2      30