

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

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

Conditional logistic regression
What

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

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

ln(odds) in reference ln(odds ratios)

Suppose the model above hold in each strata:

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

ln(odds) in reference ln(odds ratios)
different in each strata the same in each strata

Morten Frydenberg Linear and Logistic regression - Note 4.2

Conditional logistic regression
What

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

ln(odds) different in each strata

We are not interested in these !

In a matched study these are 'controlled'.

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

In the conditional model the α 's disappear !

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

Morten Frydenberg Linear and Logistic regression - Note 4.2

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

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

Conditional logistic regression How																																																																																										
logistic regression in STATA																																																																																										
<code>xi: logit cancer textile i.alkcon i.genage</code>																																																																																										
Part of the output:																																																																																										
<table border="1"> <thead> <tr> <th>cancer</th> <th>Coef.</th> <th>Std. Err.</th> <th>z</th> <th>P> z </th> <th>CI</th> </tr> </thead> <tbody> <tr> <td>textile</td> <td>.5022</td> <td>.4141</td> <td>1.213</td> <td>0.225</td> <td>-.3094 1.3139</td> </tr> <tr> <td><u>alkcon_1</u></td> <td><u>.4628</u></td> <td><u>.2823</u></td> <td><u>1.639</u></td> <td><u>0.101</u></td> <td><u>-.0905 1.0163</u></td> </tr> <tr> <td><u>alkcon_2</u></td> <td><u>2.7165</u></td> <td><u>.3232</u></td> <td><u>8.404</u></td> <td><u>0.000</u></td> <td><u>2.0829 3.3501</u></td> </tr> <tr> <td><u>genage_2</u></td> <td><u>.2450</u></td> <td><u>1.2514</u></td> <td><u>0.196</u></td> <td><u>0.845</u></td> <td><u>-2.2075 2.6977</u></td> </tr> <tr> <td><u>genage_3</u></td> <td><u>-.4940</u></td> <td><u>.5503</u></td> <td><u>-0.898</u></td> <td><u>0.369</u></td> <td><u>-1.5726 0.5846</u></td> </tr> <tr> <td><u>genage_4</u></td> <td><u>1.798</u></td> <td><u>.6406</u></td> <td><u>0.281</u></td> <td><u>0.779</u></td> <td><u>-1.0758 1.4353</u></td> </tr> <tr> <td><u>genage_5</u></td> <td><u>-.2899</u></td> <td><u>.5482</u></td> <td><u>-0.529</u></td> <td><u>0.597</u></td> <td><u>-1.3644 0.7844</u></td> </tr> <tr> <td><u>genage_6</u></td> <td><u>.2127</u></td> <td><u>.6262</u></td> <td><u>0.340</u></td> <td><u>0.734</u></td> <td><u>-1.0147 1.4401</u></td> </tr> <tr> <td><u>genage_7</u></td> <td><u>-.2305</u></td> <td><u>.5355</u></td> <td><u>-0.431</u></td> <td><u>0.667</u></td> <td><u>-1.2802 0.8190</u></td> </tr> <tr> <td><u>genage_8</u></td> <td><u>.5507</u></td> <td><u>.5263</u></td> <td><u>1.046</u></td> <td><u>0.295</u></td> <td><u>-.4809 1.5825</u></td> </tr> <tr> <td><u>genage_9</u></td> <td><u>.0315</u></td> <td><u>.5884</u></td> <td><u>0.054</u></td> <td><u>0.957</u></td> <td><u>-1.1217 1.1847</u></td> </tr> <tr> <td><u>genage_10</u></td> <td><u>.5572</u></td> <td><u>.5595</u></td> <td><u>0.996</u></td> <td><u>0.319</u></td> <td><u>-.53954 1.6539</u></td> </tr> <tr> <td><u>const</u></td> <td><u>-1.4692</u></td> <td><u>.4762</u></td> <td><u>-3.085</u></td> <td><u>0.002</u></td> <td><u>-2.4027 0.5356</u></td> </tr> </tbody> </table>							cancer	Coef.	Std. Err.	z	P> z	CI	textile	.5022	.4141	1.213	0.225	-.3094 1.3139	<u>alkcon_1</u>	<u>.4628</u>	<u>.2823</u>	<u>1.639</u>	<u>0.101</u>	<u>-.0905 1.0163</u>	<u>alkcon_2</u>	<u>2.7165</u>	<u>.3232</u>	<u>8.404</u>	<u>0.000</u>	<u>2.0829 3.3501</u>	<u>genage_2</u>	<u>.2450</u>	<u>1.2514</u>	<u>0.196</u>	<u>0.845</u>	<u>-2.2075 2.6977</u>	<u>genage_3</u>	<u>-.4940</u>	<u>.5503</u>	<u>-0.898</u>	<u>0.369</u>	<u>-1.5726 0.5846</u>	<u>genage_4</u>	<u>1.798</u>	<u>.6406</u>	<u>0.281</u>	<u>0.779</u>	<u>-1.0758 1.4353</u>	<u>genage_5</u>	<u>-.2899</u>	<u>.5482</u>	<u>-0.529</u>	<u>0.597</u>	<u>-1.3644 0.7844</u>	<u>genage_6</u>	<u>.2127</u>	<u>.6262</u>	<u>0.340</u>	<u>0.734</u>	<u>-1.0147 1.4401</u>	<u>genage_7</u>	<u>-.2305</u>	<u>.5355</u>	<u>-0.431</u>	<u>0.667</u>	<u>-1.2802 0.8190</u>	<u>genage_8</u>	<u>.5507</u>	<u>.5263</u>	<u>1.046</u>	<u>0.295</u>	<u>-.4809 1.5825</u>	<u>genage_9</u>	<u>.0315</u>	<u>.5884</u>	<u>0.054</u>	<u>0.957</u>	<u>-1.1217 1.1847</u>	<u>genage_10</u>	<u>.5572</u>	<u>.5595</u>	<u>0.996</u>	<u>0.319</u>	<u>-.53954 1.6539</u>	<u>const</u>	<u>-1.4692</u>	<u>.4762</u>	<u>-3.085</u>	<u>0.002</u>	<u>-2.4027 0.5356</u>
cancer	Coef.	Std. Err.	z	P> z	CI																																																																																					
textile	.5022	.4141	1.213	0.225	-.3094 1.3139																																																																																					
<u>alkcon_1</u>	<u>.4628</u>	<u>.2823</u>	<u>1.639</u>	<u>0.101</u>	<u>-.0905 1.0163</u>																																																																																					
<u>alkcon_2</u>	<u>2.7165</u>	<u>.3232</u>	<u>8.404</u>	<u>0.000</u>	<u>2.0829 3.3501</u>																																																																																					
<u>genage_2</u>	<u>.2450</u>	<u>1.2514</u>	<u>0.196</u>	<u>0.845</u>	<u>-2.2075 2.6977</u>																																																																																					
<u>genage_3</u>	<u>-.4940</u>	<u>.5503</u>	<u>-0.898</u>	<u>0.369</u>	<u>-1.5726 0.5846</u>																																																																																					
<u>genage_4</u>	<u>1.798</u>	<u>.6406</u>	<u>0.281</u>	<u>0.779</u>	<u>-1.0758 1.4353</u>																																																																																					
<u>genage_5</u>	<u>-.2899</u>	<u>.5482</u>	<u>-0.529</u>	<u>0.597</u>	<u>-1.3644 0.7844</u>																																																																																					
<u>genage_6</u>	<u>.2127</u>	<u>.6262</u>	<u>0.340</u>	<u>0.734</u>	<u>-1.0147 1.4401</u>																																																																																					
<u>genage_7</u>	<u>-.2305</u>	<u>.5355</u>	<u>-0.431</u>	<u>0.667</u>	<u>-1.2802 0.8190</u>																																																																																					
<u>genage_8</u>	<u>.5507</u>	<u>.5263</u>	<u>1.046</u>	<u>0.295</u>	<u>-.4809 1.5825</u>																																																																																					
<u>genage_9</u>	<u>.0315</u>	<u>.5884</u>	<u>0.054</u>	<u>0.957</u>	<u>-1.1217 1.1847</u>																																																																																					
<u>genage_10</u>	<u>.5572</u>	<u>.5595</u>	<u>0.996</u>	<u>0.319</u>	<u>-.53954 1.6539</u>																																																																																					
<u>const</u>	<u>-1.4692</u>	<u>.4762</u>	<u>-3.085</u>	<u>0.002</u>	<u>-2.4027 0.5356</u>																																																																																					

Morten Frydenberg

Linear and Logistic regression - Note 4.2

7

Conditional logistic regression in STATA																														
The syntax:																														
<code>xi: clogit cancer textile i.alkcon, group(genage)</code>																														
Part of the output:																														
<table border="1"> <thead> <tr> <th>cancer</th> <th>Coef.</th> <th>Std. Err.</th> <th>z</th> <th>P> z </th> <th>CI</th> </tr> </thead> <tbody> <tr> <td>textile</td> <td>.4929</td> <td>.4103</td> <td>1.201</td> <td>0.230</td> <td>-.3112 1.2971</td> </tr> <tr> <td><u>alkcon_1</u></td> <td><u>.452</u></td> <td><u>.27923</u></td> <td><u>1.621</u></td> <td><u>0.105</u></td> <td><u>-.094 .9999</u></td> </tr> <tr> <td><u>alkcon_2</u></td> <td><u>2.660</u></td> <td><u>.31936</u></td> <td><u>8.332</u></td> <td><u>0.000</u></td> <td><u>2.034 3.2868</u></td> </tr> </tbody> </table>							cancer	Coef.	Std. Err.	z	P> z	CI	textile	.4929	.4103	1.201	0.230	-.3112 1.2971	<u>alkcon_1</u>	<u>.452</u>	<u>.27923</u>	<u>1.621</u>	<u>0.105</u>	<u>-.094 .9999</u>	<u>alkcon_2</u>	<u>2.660</u>	<u>.31936</u>	<u>8.332</u>	<u>0.000</u>	<u>2.034 3.2868</u>
cancer	Coef.	Std. Err.	z	P> z	CI																									
textile	.4929	.4103	1.201	0.230	-.3112 1.2971																									
<u>alkcon_1</u>	<u>.452</u>	<u>.27923</u>	<u>1.621</u>	<u>0.105</u>	<u>-.094 .9999</u>																									
<u>alkcon_2</u>	<u>2.660</u>	<u>.31936</u>	<u>8.332</u>	<u>0.000</u>	<u>2.034 3.2868</u>																									

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
<u>alkcon_1</u>	<u>1.572508</u>	<u>.4390957</u>	<u>1.62</u>	<u>0.105</u>	<u>.909724 2.718168</u>
<u>alkcon_2</u>	<u>14.30908</u>	<u>4.569879</u>	<u>8.33</u>	<u>0.000</u>	<u>7.651811 26.75835</u>

Morten Frydenberg Linear and Logistic regression - Note 4.2

8

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 Frydenberg

Linear and Logistic regression - Note 4.2

9

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 Frydenberg

Linear and Logistic regression - Note 4.2

11

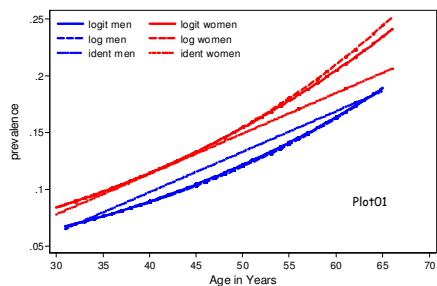
Other methods to analysis of binary response data Risk difference models						
$\Pr(\text{event given the covariates}) = \alpha + \sum_{i=1}^p (\beta_i \cdot x_i)$						
Such a model modelling the risk difference can easily be fitted by many programs (not SPSS).						
Logistic regression in STATA :						
<code>xi: logit obese age i.sex</code>						
or						
<code>xi: g1m obese age i.sex, fam(bin) link(logit)</code>						
Relative risk model:						
<code>xi: g1m obese age i.sex, fam(bin) link(log)</code>						
The link is log instead of logit						

Morten Frydenberg Linear and Logistic regression - Note 4.2

10

Other methods to analysis of binary response data

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



Morten Frydenberg

Linear and Logistic regression - Note 4.2

13

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**!

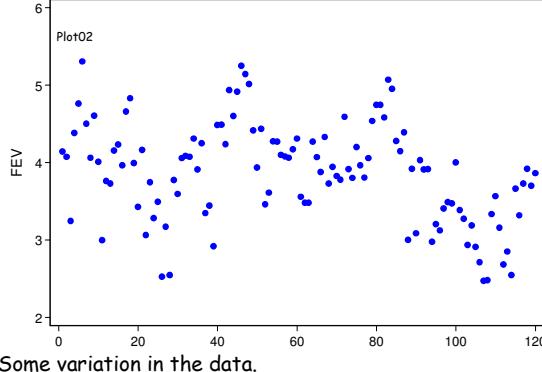
Morten Frydenberg

Linear and Logistic regression - Note 4.2

14

Clustered data / data with several random components

120 measurements of FEV:



Some variation in the data.

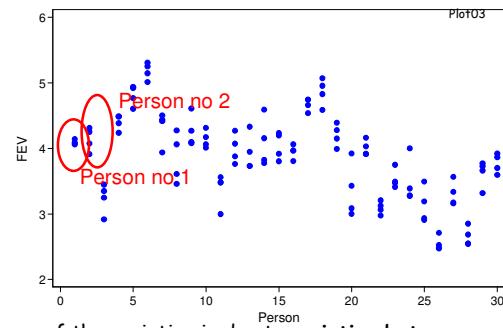
Morten Frydenberg

Linear and Logistic regression - Note 4.2

15

Clustered data / data with several random components

But it is on only 30 persons:



Some of the variation is due to **variation between persons** and some within person.

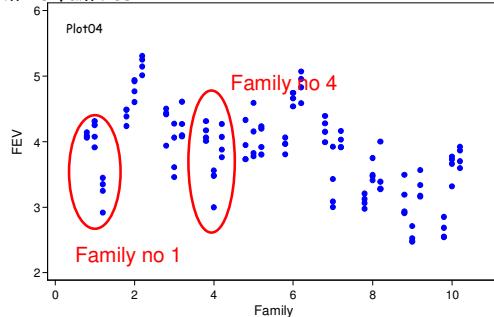
Morten Frydenberg

Linear and Logistic regression - Note 4.2

16

Clustered data / data with several random components

From 10 families:



Some of the variation between persons is due to **variation between families and some within family**.

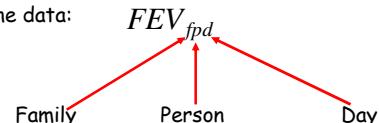
Morten Frydenberg

Linear and Logistic regression - Note 4.2

17

Clustered data / data with several random components

Structure of the data:



Three sources of random variation:

Variation between **families**

Variation between **persons** (variation within family)

Variation between **days** (variation within person)

Morten Frydenberg

Linear and Logistic regression - Note 4.2

18

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 β_I and β_U are highly overestimated
- The precision of the β_A and β_S are overestimated
- The estimates of the β_I and β_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 β_A and β_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

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

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

variance

F_f	: Random family contribution	σ_F^2
P_{fp}	: Random person contribution	σ_P^2
E_{fpd}	: Random day contribution	σ_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

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

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

Random part

 $\beta_0, \beta_I, \beta_U, \beta_A, \beta_S$ and β_G Quantify the systematic variation σ_F^2, σ_P^2 and σ_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

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

$$+ \boxed{F_f + P_{fp} + X_{bs}}$$

Random part
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

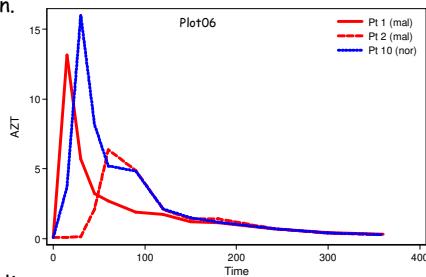
Linear and Logistic regression - Note 4.2

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

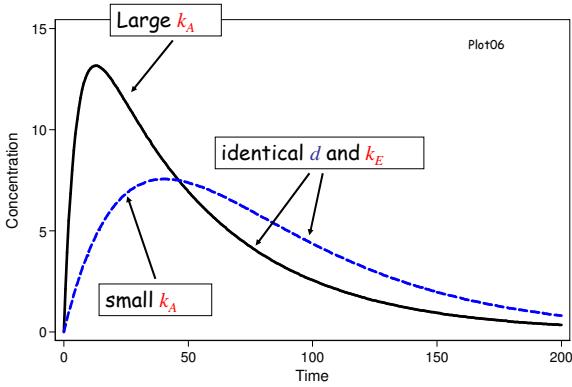
k_E : eliminationsrate

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