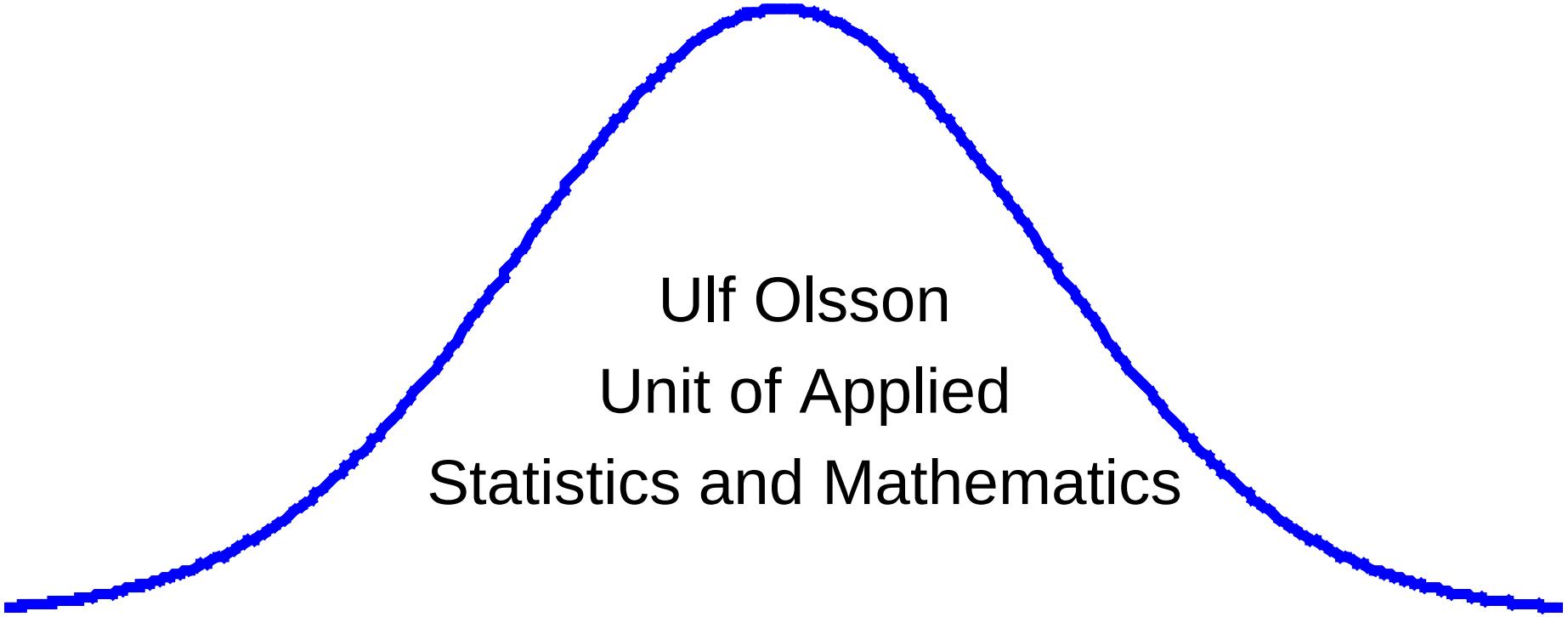
Workshop in mixed models

Umeå, August 27-28, 2015

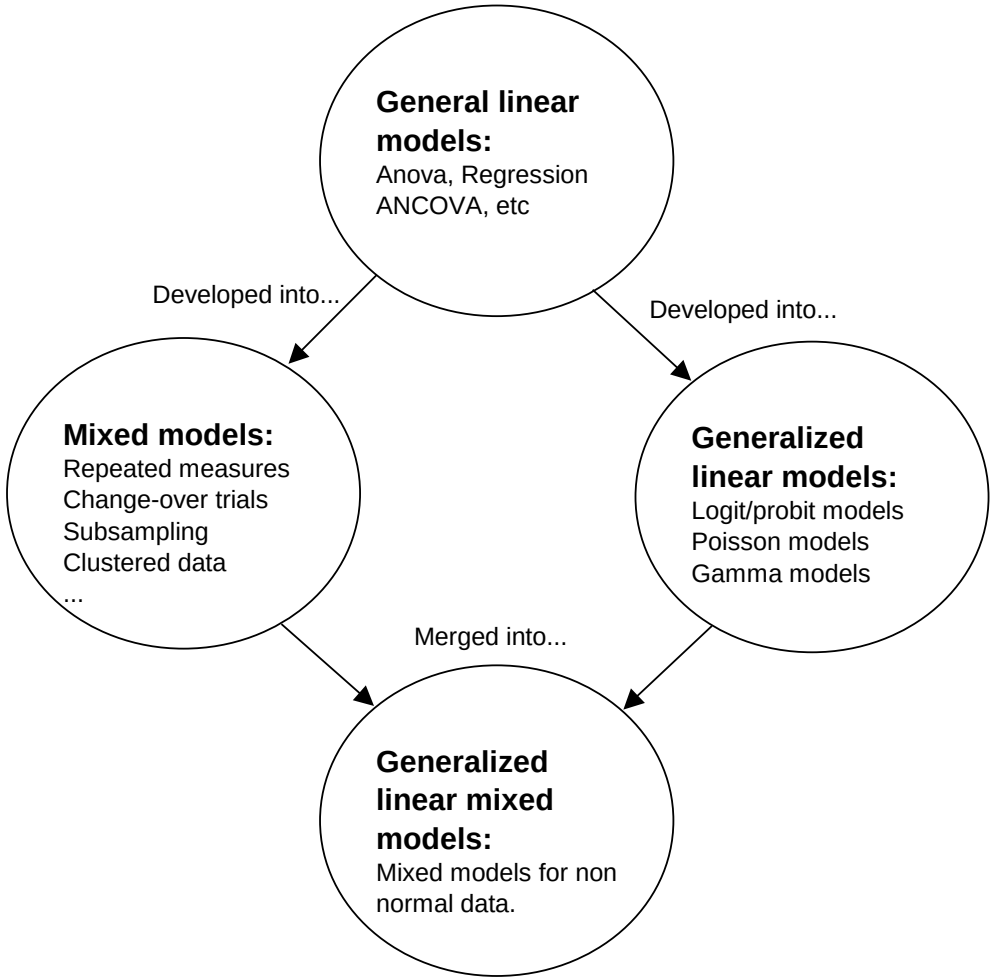


Ulf Olsson

Unit of Applied

Statistics and Mathematics

1. Introduction



2. General linear models (GLM)

GLM (cont.)

Data: Response variable y for n ”individuals”

Some type of design (+ possibly covariates)

Linear model:

$$y = f(\text{design, covariates}) + e$$

$$\mathbf{y} = \mathbf{XB} + e$$

GLM (cont.)

Examples of GLM:

(Multiple) linear regression

Analysis of Variance (ANOVA, including t test)

Analysis of covariance (ANCOVA)

GLM (cont.)

- Parameters are estimated using either the Least squares, or Maximum Likelihood methods
- Possible to test statistical hypotheses, for example to test if different treatments give the same mean values
- Assumption: The residuals e_i are independent, normally distributed and have constant variance.

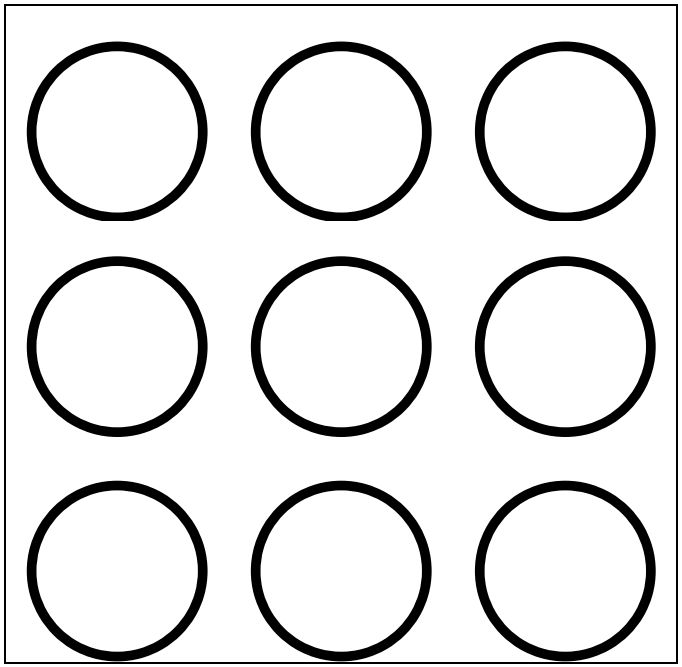
GLM (cont): some definitions

- **Factor**: e.g. treatments, or properties such as sex
  - Levels

    Example : Facor: type of fertilizer

    Levels: Low Medium High level of N
- **Experimental unit**: The smallest unit that is given an individual treatment
- **Replication**: To repeat the same treatments on new experimental units

Experimetal unit



Pupils	**Class**
Chicken	**Box**
Plants	**Bench**
Trees	**Plot**

3. “Mixed models”: Fixed and random factors

Fixed factor: those who planned the experiment decided which levels to use

Random factor: The levels are (or may be regarded as) a sample from a population of levels

Fixed and random factors

Example: 40 forest stands. In each stand, one plot fertilized with A and one with B.

Response variable: e.g. diameter of 5 trees on each plot

Fixed factor: fertilizer, 2 levels (A and B)

Experimental unit: the plot (NOT the tree!)

Replication on 40 stands

”Stand” may be regarded as a random factor

Mixed models (cont.)

Examples of random factors

- ”Block” in some designs
- ”Individual” (when several measurements are made on each individual)
- ”School class” (in experiments with teaching methods: then exp. unit is the class)
- …i.e. in situations when many measurements are made on the same experimental unit.

Mixed models (cont.)

Mixed models are models that include both fixed and random factors.

Programs for mixed models can also analyze models with only fixed, or only random, factors.

Mixed models: formally

$$\mathbf{y} = \mathbf{XB} + \mathbf{Zu} + \mathbf{e}$$

y is a vector of responses
XB is the fixed part of the model
    X: design matrix
    B: parameter matrix
Zu is the random part of the model
e is a vector of residuals

y = f(fixed part) + g(random part) + e

Parameters to estimate

- **Fixed effects:** the parameters in **B**
- **Ramdom effects:**
  - the variances and covariances of the random effects in **u**: **Var(u)=G**

    ”G-side random effects”
  - The variances and covariances of the residual effects: **Var(e)=R**

    ”R-side random effects”

To formulate a mixed model you might

Decide the design matrix **X** for fixed effects

Decide the design matrix **Z** for random effefcts

In some types of models:

Decide the structure of the covariance matrices **G** or, more commonly, **R**.

Example 1

Two-factor model with one random factor

Treatments: two mosquito repellants A_1 and A_2

(Schwartz, 2005)

24 volonteeers divided into three groups

4 in each group apply A_1, 4 apply A_2

Each group visits one of three different areas

y=number of bites after 2 hours

Ex 1: data

Bites	Brand	Site
21	A1	1
19	A1	1
20	A1	1
22	A1	1
14	A2	1
15	A2	1
13	A2	1
16	A2	1
14	A1	2
17	A1	2
15	A1	2
17	A1	2
12	A2	2
11	A2	2
12	A2	2
14	A2	2
16	A1	3
20	A1	3
18	A1	3
19	A1	3
14	A2	3
14	A2	3
14	A2	3
12	A2	3

Ex 1: Model

$$y_{ijk}=\mu+\alpha_i+b_j+ab_{ij}+e_{ijk}$$

Where

μ is a general mean value,

α_i is the effect of brand *i*

b_j is the random effect of site j

ab_{ij} is the interaction between factors a and b

e_{ijk} is a random residual

$b_j \sim N(o, \sigma^2_b)$ $\quad$ $e_{ijk} \sim N(o, \sigma^2_e)$

Note: Fixed effects and parameters – Greek letters

Random effects – Latin letters

Ex 1: Program

SAS code

```
PROC MIXED
DATA=Bites;
CLASS brand site;
MODEL bites=brand;
RANDOM site brand*site;
RUN;
```

R code

```
lme(
data=bites,

fixed=bites~brand,
random=~1|site/brand)
```

Ex 1, results

The Mixed Procedure

Covariance Parameter Estimates

Cov Parm	Estimate
Site	2.6771
Brand*Site	0.3194
Residual	1.8472

Ex 1, results

Fit Statistics

-2 Res Log Likelihood	87.1
AIC (smaller is better)	93.1
AICC (smaller is better)	94.5
BIC (smaller is better)	90.4

Type 3 Tests of Fixed Effects

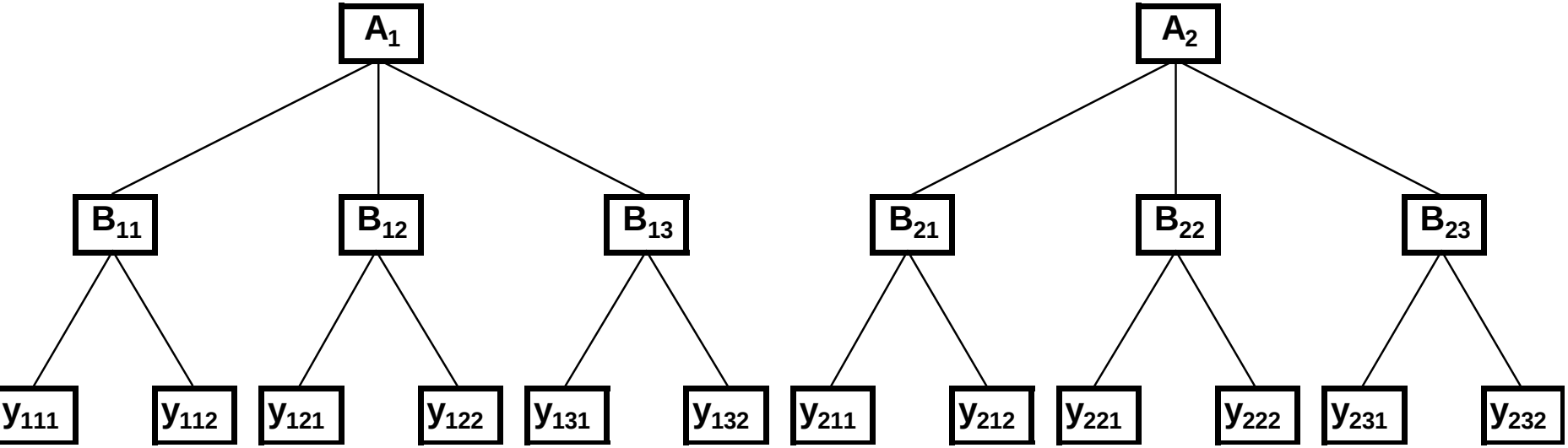
Effect	Num DF	Den DF	F Value	Pr > F
Brand	1	2	43.32	0.0223

Example 2: Subsampling

Two treatments

Three experimental units per treatment

Two measurements on each experimental unit



Ex 2

An example of this type:

3 different fertilizers

4 plots with each fertilizer

2 mangold plants harvested from each plot

y = iron content

Ex 2: data

Treat	Plot	Plant	Iron
1	1	1	102.4
1	1	2	98.3
1	2	1	99.7
1	2	2	99.3
1	3	1	100.1
1	3	2	100.4
1	4	1	97.0
1	4	2	99.2
2	1	1	96.4
2	1	2	98.8
2	2	1	100.7
2	2	2	98.1
2	3	1	101.2
2	3	2	101.5
2	4	1	97.5
2	4	2	97.6
3	1	1	103.8
3	1	2	104.1
3	2	1	105.6
3	2	2	104.7
3	3	1	109.1
3	3	2	108.4
3	4	1	101.4
3	4	2	102.6

Ex 2: model

$$y_{ij}=\mu+\alpha_i + b_{ij} + e_{ijk}$$

μ	General mean value
α_i	Fixed effect of treatment i
b_{ij}	Random effect of plot j within treatment i
e_{ijk}	Random residual

Ex 2: results

Covariance Parameter Estimates	
Cov Parm	**Estimate**
Treat*Plot	3.3507
Residual	1.5563

Type 3 Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Treat	2	9	10.57	0.0043

Example 3: ”Split-plot models” Models with several error terms

y=The dry weight yield of grass

Cultivar, levels A and B.

Bacterial inoculation, levels, C, L, D

Four replications in blocks.

Ex 3: design

Repl. 1	Repl. 2	Repl. 3	Repl. 4
C 29.4	D 28.7	D 29.7	C 26.7
L 34.4	L 33.4	C 28.6	L 31.8
D 32.5	C 28.9	L 32.9	D 28.9
C 27.4	L 36.4	C 27.2	D 28.6
L 34.5	D 32.4	L 32.6	L 30.7
D 29.7	C 28.7	D 29.1	C 26.8

Legend: C=Control L=Live D=Dead

Cultivar A

Cultivar B

Ex 3

Block and Block*cult used as random factors.

Results for random factors:

Covariance Parameter Estimates	
Cov Parm	**Estimate**
Block	0.8800
Cult*Block	0.8182
Residual	0.7054

Ex 3

Results for fixed factiors

Type 3 Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Cult	1	3	0.76	0.4471
Inoc	2	12	83.76	<.0001
Cult*Inoc	2	12	1.29	0.3098

Example 4: repeated measures

4 treatments

9 dogs per treatment

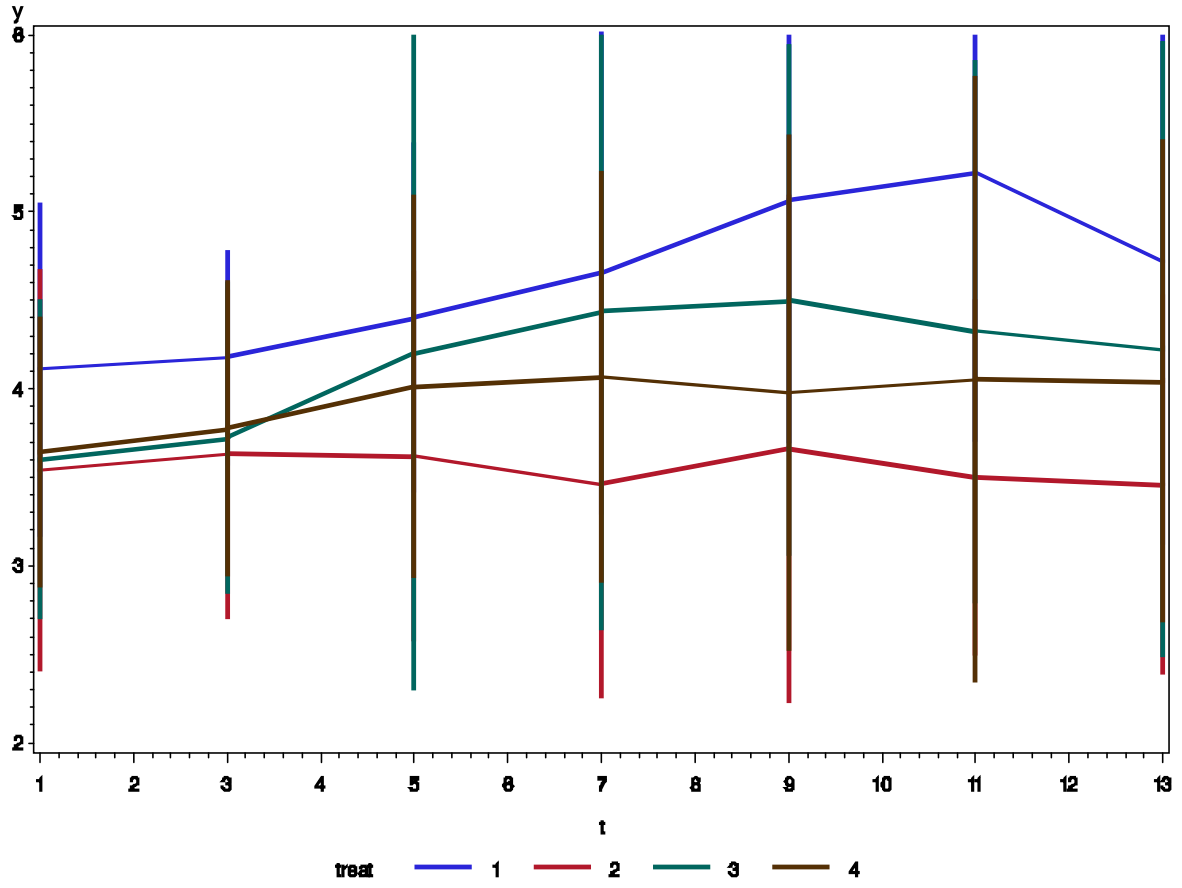
Each dog measured at several time points

y: some measurement, for example level of stress hormone

Ex 4: data structure

treat	dog	t	y
1	1	1	4.0
1	1	3	4.0
1	1	5	4.1
1	1	7	3.6
1	1	9	3.6
1	1	11	3.8
1	1	13	3.1

Ex 4: plot



Ex 4: program

SAS code	R code
PROC MIXED	lme(
DATA=dogs;	data=dogs,
CLASS treat t dog;	
MODEL y = treat t treat*t;	fixed=y~treat*t,
REPEATED /subject=dog*treat TYPE=UN;	random=~1|treat/dog, weights=varIdent(form=~1|t), correlation= corSymm(form=~1|treat/dog)
RUN;	)

Ex 4, results

Type 3 Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
treat	3	32	6.91	0.0010
t	6	27	6.75	0.0002
treat*t	18	48.5	1.93	0.0356

Covariance structures for repeated-measurses data

Model: **y** = **XB** + **Zu** + **e**

The residuals **e** (”R-side random effects”) are correlateded over time, correlation matrix **R**.

If **R** is left free (unstructured) this gives tx(t-1)/2 parameters to estimate (t=# of time points).

If n is small and t is large, we might run into problems (non-convergence, negative definite Hessian matrix).

Covariance structure

One solution: Apply some structure on **R** to reduce the number of parameters.

Covariance structure

Name	Structure	Example
Compound symmetry	CS	$\begin{bmatrix} \sigma^2+\sigma_1 & \sigma_1 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma^2+\sigma_1 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma^2+\sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma_1 & \sigma^2+\sigma_1 \end{bmatrix}$
Unctructured	UN	$\begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 \end{bmatrix}$
First order AR	AR(1)	$\begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$
Spetial power	SP(POW)(c)	$\sigma^2 \begin{bmatrix} 1 & \rho^{d_{12}} & \rho^{d_{13}} & \rho^{d_{14}} \\ \rho^{d_{21}} & 1 & \rho^{d_{23}} & \rho^{d_{24}} \\ \rho^{d_{31}} & \rho^{d_{32}} & 1 & \rho^{d_{34}} \\ \rho^{d_{41}} & \rho^{d_{42}} & \rho^{d_{43}} & 1 \end{bmatrix}$

Analysis strategy

Baseline model: Time as a ”class” variable

MODEL treatment time treatment*time;

”Repeated” part: First try UN. Simplify if needed:

AR(1) for equidistant time points, else SP(POW)

CS is only a last resort!

To simplify the fixed part: Polynomials in time can be used. Or other known functions.

Other tricks

Comparisons between models:

Akaike’s Information Criterion (AIC)

Denominator degrees of freedom for tests:

Use the method by Kenward and Roger (1997)

Normal distribution?

Make diagnostic plots! Transformations?

Robust (”sandwich”) estimators can be used

-or Generalized Linear Mixed Models…

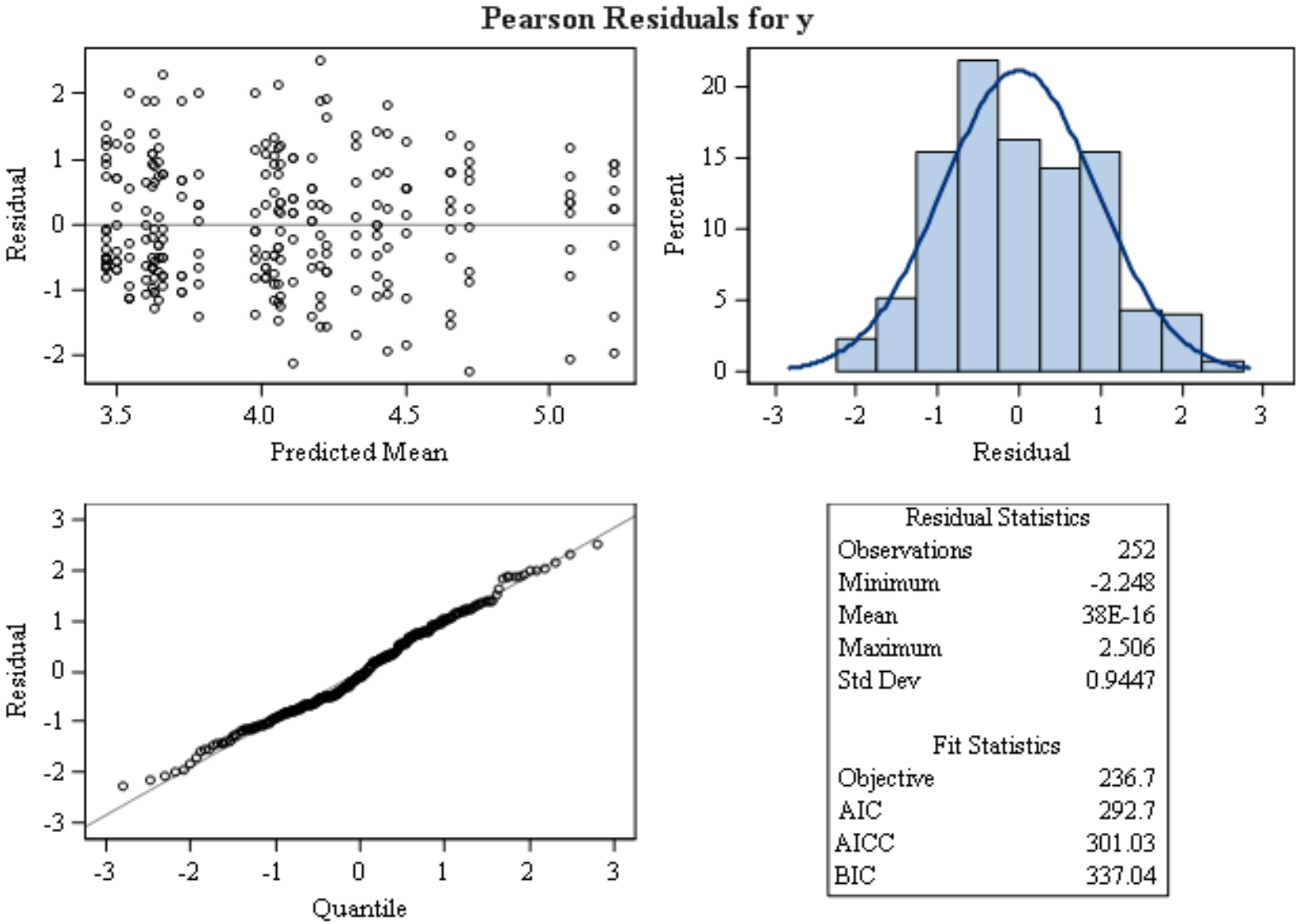
For the ”dog” data:

Unstructured: AIC=292.7

AR(1): AIC=296.3

Denominator d.f. according to Kenward and Rogers were used on page 35.

Residual plot



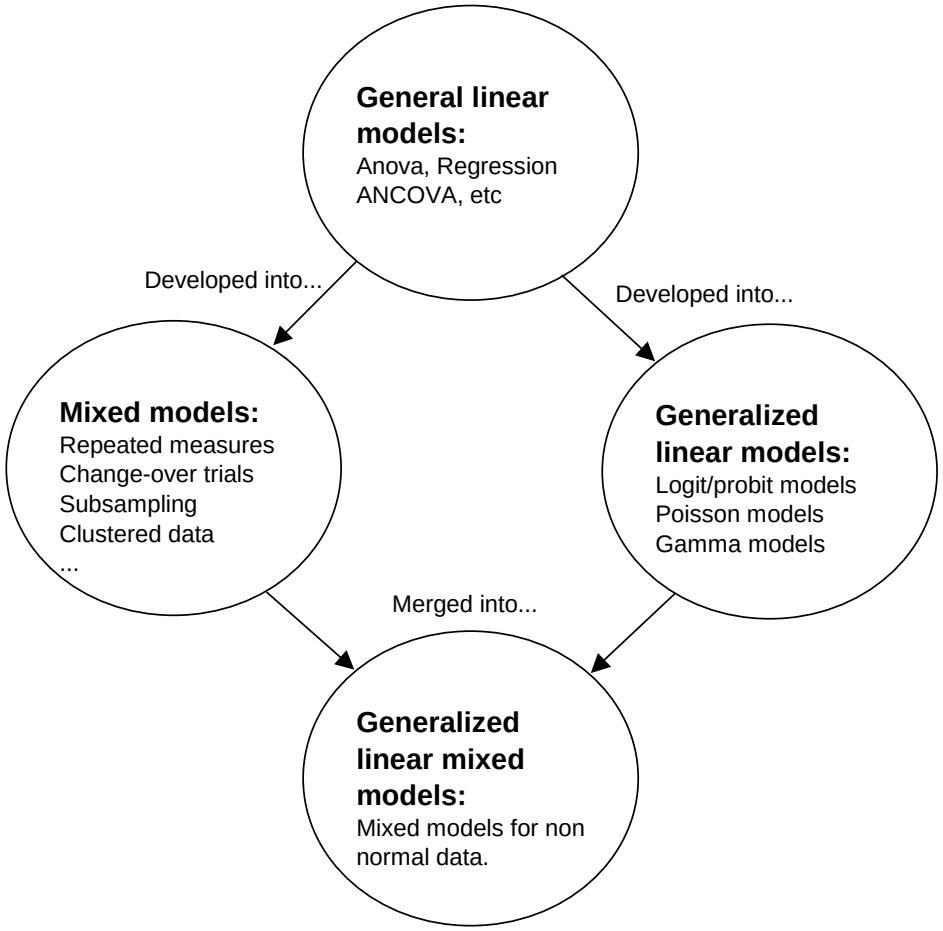
Residual Statistics	
Observations	252
Minimum	-2.248
Mean	38E-16
Maximum	2.506
Std Dev	0.9447

Fit Statistics	
Objective	236.7
AIC	292.7
AICC	301.03
BIC	337.04

Not covered…

- Models with spatial variation
  - Lecture by Johannes Forkman
- Models with non-normal responses
  - (Generalized Linear Mixed Models)
- …and much more

Summary


General linear models:
Anova, Regression
ANCOVA, etc
Developed into...
Developed into...
Mixed models:
Repeated measures
Change-over trials
Subsampling
Clustered data
...
Generalized linear models:
Logit/probit models
Poisson models
Gamma models
Merged into...
Generalized linear mixed models:
Mixed models for non normal data.

”All models are wrong…

…but some are useful.” (G. E. P. Box)

References

Fitzmaurice, G. M., Laird, N. M. and Ware, J. H. (2004): A*pplied longitudinal analysis*. New York, Wiley

Littell, R., Milliken, G., Stroup, W. Wolfinger, R. and and Schabenberger O. (2006): SAS for mixed models, second ed. Cary, N. C., SAS Institute Inc.

(R solutions to this can be found on the net)

Ulf Olsson: *Generalized linear models: an applied approach*. Lund, Studentlitteratur, 2002

Ulf Olsson (2011):Statistics for Life Science 1. Lund, Studentlitteratur

Ulf Olsson (2011):Statistics for Life Science 2. Lund, Studentlitteratur