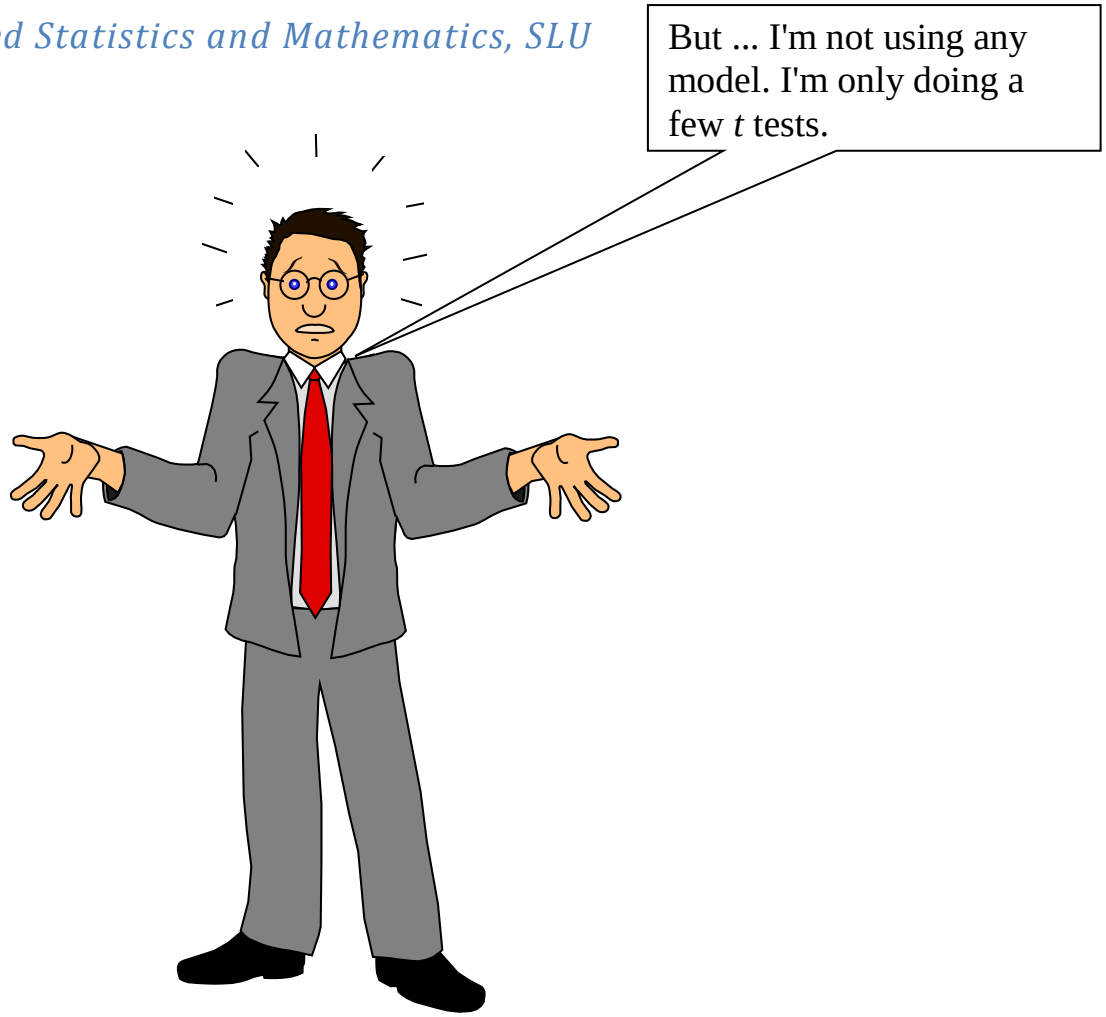
Generalized Linear Models

Ulf Olsson, Unit of Applied Statistics and Mathematics, SLU



Model-based statistical methods

Many statistical methods, like t tests, are model based:

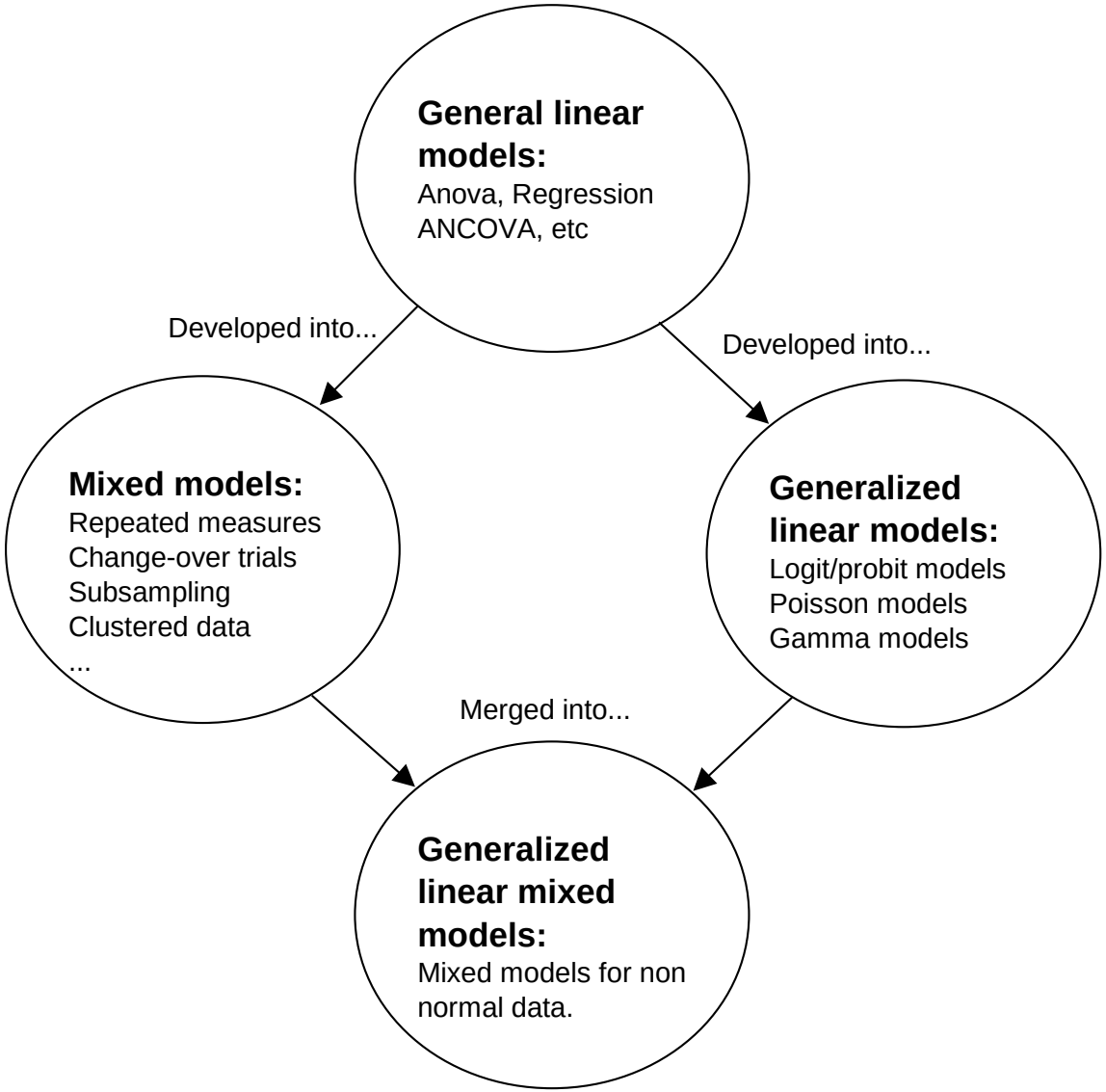
$$y = f(x) + e$$

y Response variable

x Covariates and factors

f some function

e Residuals (i.e. differences between the model and the data)


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.

General linear models

If the function f is linear, we are dealing with *General linear models*

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + e$$

or in matrix terms

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

X is a design matrix that contains data for the model, similar to a data spreadsheet

Assumptions:

1. The residuals e are i*ndependent*
2. The residuals have *constant variance*
3. The residuals (approximately) follow a *normal distribution*

It is possible to

- Estimate the parameters $\beta_0 \ldots \beta_p$ and σ_e^2 (the *residual variance*) of the model. (ML, LS)
- Test if parameters are significantly different from zero
- Assess the fit of the model
- Make predictions based on the model

Tests of type 1, 2, 3 and 4

The order in which factors are added to a model may affect the significance of the factor.

Example

y = yield of a crop

x_1 = soil humidity

x_2 = rain during the growing season

Possible result: x_1 has a significant effect on yield

After x_1 is in the model, x_2 has no significant effect

Does that mean that amount of rain has no significant effect on yield?

Tests of type 1, 2, 3 and 4 (cont.)

Type I	SS(A), SS(B|A) and SS(AB|A;B).	(Sequential tests)
Type II	SS(A|B); SS(B|A) and SS(AB|A;B).	"As if the factor was added last".
Type III	Computes SS "as if the experiment had been balanced".	
Type IV	As Type III but different handling of empty cells	

Examples of General Linear Models

Simple regression	$y = \beta_0 + \beta_1 x + e$
Multiple regression	$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + e$
t test	$y = \beta_0 + \beta_1 x + e$
	where x=1 for one group, x=0 for the other group
	"dummy variable"
Analysis of Variance	$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + e$
	Where $x_1, x_2, \ldots, x_p$ are dummy variables

Example 1: A t test is a regression model!

Number of gill movements per minute was recorded for water louse (*Asellus*) in either stagnant water or in oxygen-rich water. Data were organized as follows:

Site	Movements	Dummy
Stagnant	44	0
Stagnant	53	0
Stagnant	54	0
Stagnant	43	0
Stagnant	48	0
Stagnant	49	0
Stagnant	53	0
Oxygen-rich	42	1
Oxygen-rich	48	1
Oxygen-rich	46	1
Oxygen-rich	43	1
Oxygen-rich	49	1
Oxygen-rich	42	1
Oxygen-rich	41	1
Oxygen-rich	40	1
Oxygen-rich	44	1
Oxygen-rich	48	1

Two analyses are given on the next page

Two-sample T for Movements

Site	N	Mean	StDev	SE Mean
Oxygen-rich	10	44,30	3,23	1,0
Stagnant	7	49,14	4,45	1,7

T-Test of difference = 0 (vs not =): T-Value = -2,61
P-Value = 0,020 DF = 15

Regression Analysis: Movements versus Dummy

The regression equation is
Movements = 49,1 - 4,84 Dummy

Predictor	Coef	SE Coef	T	P
Constant	49,143	1,424	34,51	0,000
Dummy	-4,843	1,857	-2,61	0,020

S = 3,76791   R-Sq = 31,2%   R-Sq(adj) = 26,6%

The dummy variable idea can be used when there are more than two groups (Analysis of Variance). This is done automatically in computer programs.

Variables in GLM models

1. Numeric variables (covariates)
2. Non-numeric ("class") variables (factors)

   (Translated to dummy variables by the program)

Term: Linear predictor

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

Generalized linear models, GLIM

GLIM is a class of statistical models that are based on the following building blocks:

1. The response variable is assumed to follow some distribution in the *exponential family*
2. The mean value μ of y is assumed to be related to covariates and factors through

$$g(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

Link function | **Linear predictor**

The link function g(μ) is often chosen as the *canonical link* for the chosen distribution

To analyze data, you have to specify

1. The distribution
2. The link function
3. The linear predictor ("model").

Examples of distributions and their canonical links

Distribution	Canonical link	Use: type of data
Normal	Identity	Continuous, Normal
Binomial	$logit(p) = \log(\frac{p}{1-p})$	Proportion
Poisson	log: log(μ)	Count
Gamma	Inverse: $-\frac{1}{\mu}$	Time duration, "lifetime"
Multinomial	(Cumulative logit)[1]	Ordinal data

[1]: Not the canonical link but a link e.g. Glimmix can handle for multinomial data

Example 2: Models for binary data

(Bliss, 1934)

Carbon disulphide, in different concentrations, was applied on groups of beetles.

Response: y = number of dead beetles in a group of n.

```
DATA beetles;
INPUT x n y;
p=y/n;
CARDS;
1.6907 59 6
1.7242 60 13
1.7552 62 18
1.7842 56 28
1.8113 63 52
1.8369 59 53
1.8610 62 61
1.8839 60 60
;
```

Note: x=log(dose) is often used in dose-response models.

A logistic model for this type of data can be written

$$\text{logit}(p) = \log(p/(1-p)) = \beta_0 + \beta_1 x$$

This model can be fitted in SAS:

```
PROC GLIMMIX data=beetles plots=PearsonPanel;
   MODEL y/n=x
   /dist=bin link=logit;
RUN;
```

Parts of the output:

Fit Statistics	
-2 Log Likelihood	37.43
AIC (smaller is better)	41.43
AICC (smaller is better)	43.83
BIC (smaller is better)	41.59
CAIC (smaller is better)	43.59
HQIC (smaller is better)	40.36
Pearson Chi-Square	10.03
Pearson Chi-Square / DF	1.67

Deviance/df should be “close to 1”. Is 1.67 “too large”?

$X^2 = 10.03$ on 6 d.f., $p = 0.12$

Type III Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
x	1	6	138.49	<.0001

Parameter Estimates					
Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	-60.7175	5.1807	6	-11.72	<.0001
x	34.2703	2.9121	6	11.77	<.0001

The fitted model is $\log(p/(1-p)))=-60.71 + 34.27x$

LD_{50}

If p (the probability of being killed) is 0.5 then

$$\log(p/(1-p))=0$$

$$b_0 + b_1x = 0$$

$$x = - b_0/b_1$$

For our example data,

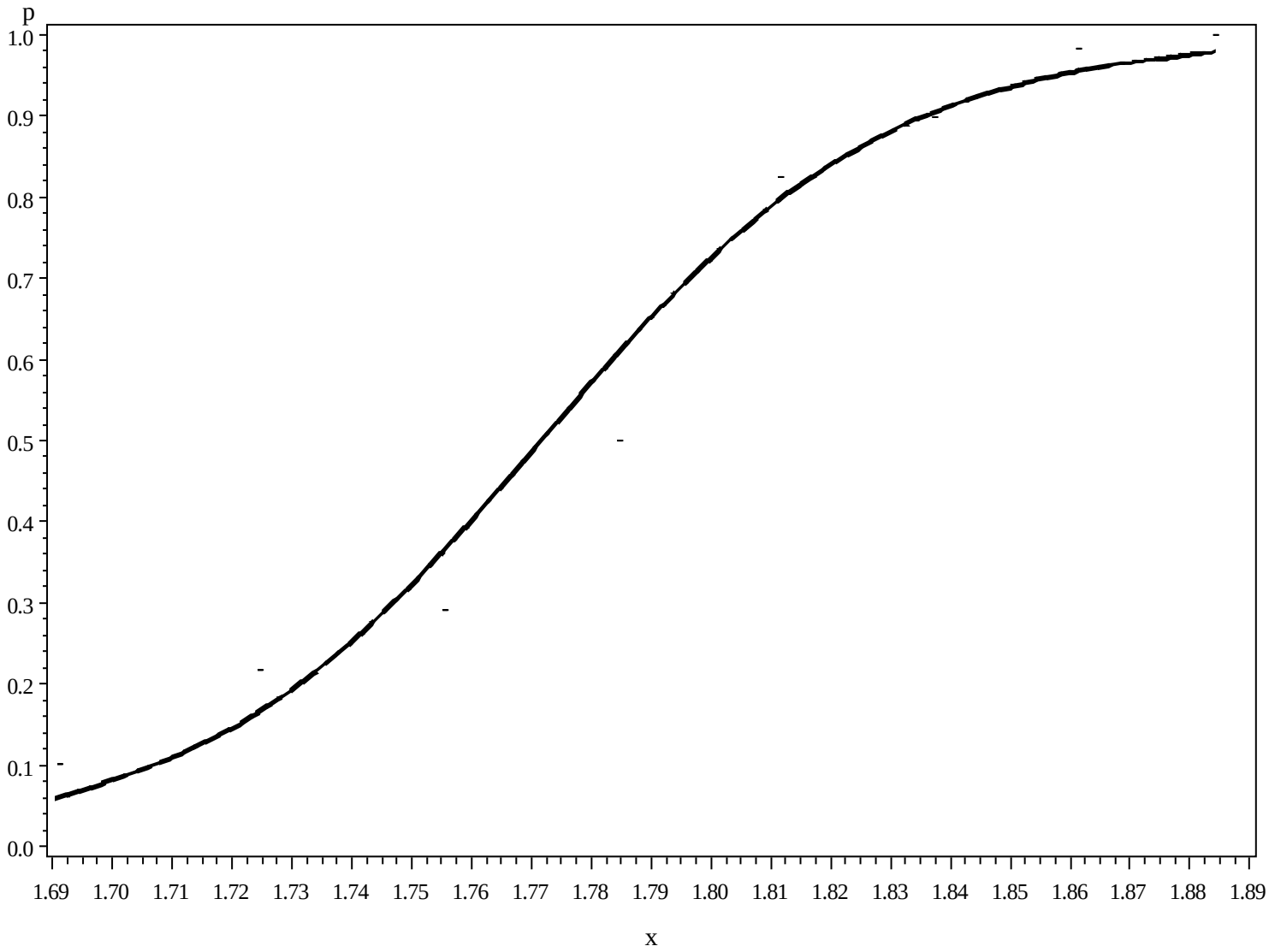
b_0 = -60.7175 (“intercept” in the printout)

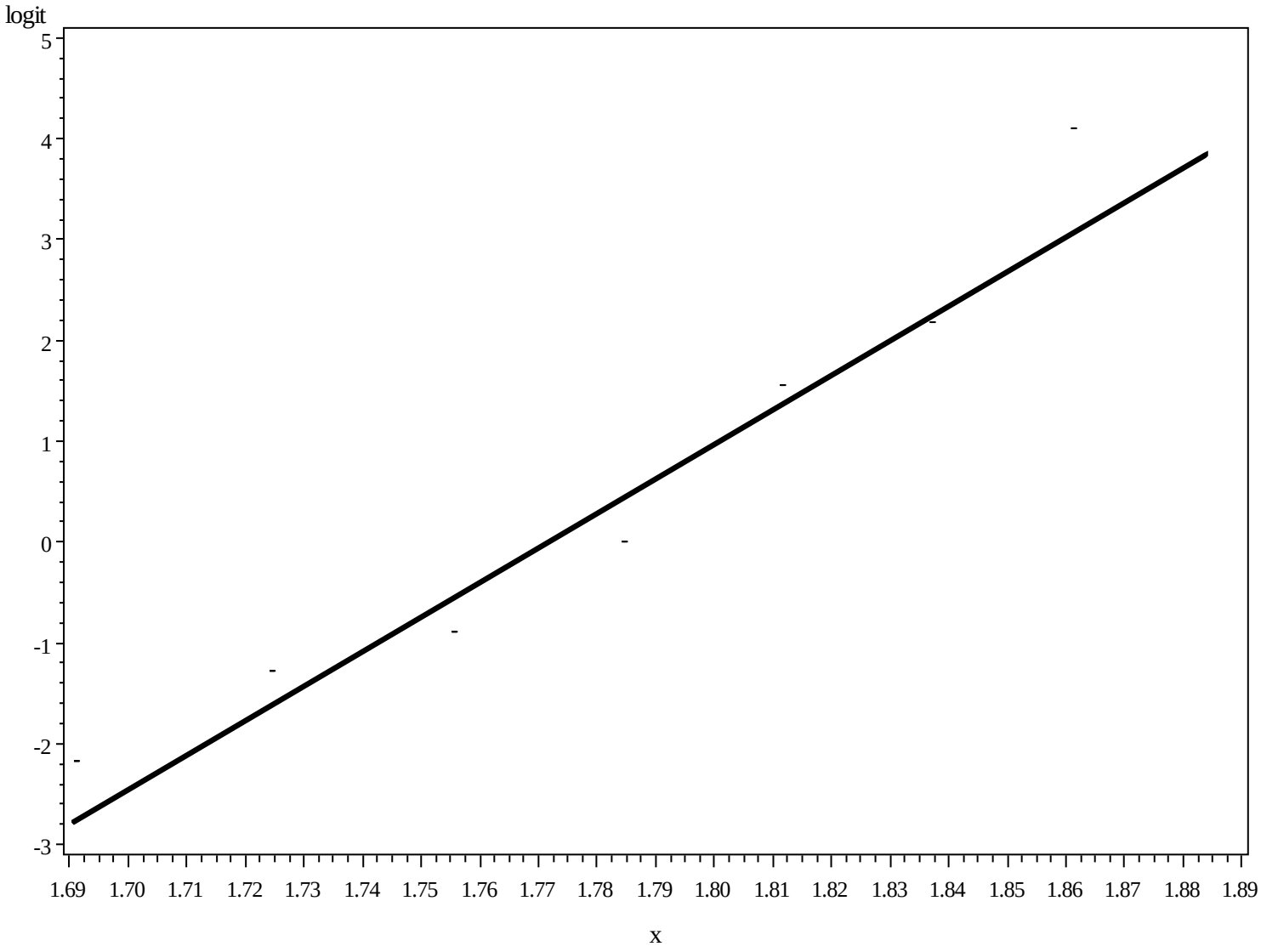
b_1 = 34.2703 (“x” in the printout) so

LD_{50} = - (-60.7175/34.2703) =1.77

Notation: b_0 denotes the sample estimate of β_0

b_1 denotes the sample estimate of β_1


p
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
1.69
1.70
1.71
1.72
1.73
1.74
1.75
1.76
1.77
1.78
1.79
1.80
1.81
1.82
1.83
1.84
1.85
1.86
1.87
1.88
1.89
x


logit
5
4
3
2
1
0
-1
-2
-3
1.69
1.70
1.71
1.72
1.73
1.74
1.75
1.76
1.77
1.78
1.79
1.80
1.81
1.82
1.83
1.84
1.85
1.86
1.87
1.88
1.89
x

Example 3: Binomial "Anova-like" model

Do blood stains on egg shells depend on hen hybrid and/or on diet?

Data: 18 cages with about 100 hens in each: 3 diets x 2 hybrids x 3 replicates

10 eggs randomly selected from each cage. y=blood stains/no blood stains

Cage	hybrid	food	FREQ	nblood
1	LB	fiber	10	5
2	LSL	pellets	10	0
3	LB	Control	10	3
4	LSL	fiber	10	1
5	LB	pellets	10	5
6	LSL	Control	10	0
7	LB	fiber	10	4
8	LSL	pellets	10	1
9	LB	Control	10	5
10	LSL	fiber	10	0
11	LB	pellets	10	3
12	LSL	Control	10	1
13	LB	fiber	10	3
14	LSL	Control	10	1
15	LB	pellets	10	1
16	LSL	fiber	10	0
17	LB	Control	10	5
18	LSL	pellets	10	1

SAS program

```
PROC GLIMMIX data=blood ;
   CLASS hybrid food;
   MODEL nblood/freq = hybrid food hybrid*food/
   DIST=bin LINK=logit ;
   LSMEANS hybrid/pdiff ilink;
RUN;
```

Output:

Fit Statistics	
-2 Log Likelihood	44.96
AIC (smaller is better)	56.96
AICC (smaller is better)	64.60
BIC (smaller is better)	62.30
CAIC (smaller is better)	68.30
HQIC (smaller is better)	57.70
Pearson Chi-Square	9.94
Pearson Chi-Square / DF	0.55

Type III Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
hybrid	1	12	19.99	0.0008
food	2	12	0.23	0.7948
hybrid* food	2	12	0.38	0.6945

hybrid Least Squares Means							
hybrid	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**	**Mean**	**Standard Error Mean**
LB	-0.5070	0.2194	12	-2.31	0.0394	0.3759	0.05148
LSL	-2.8818	0.4837	12	-5.96	<.0001	0.05306	0.02430

Differences of hybrid Least Squares Means						
hybrid	_hybrid	Estimate	Standard Error	DF	t Value	Pr > |t|
LB	LSL	2.3748	0.5312	12	4.47	0.0008

Odds ratios

Odds are defined as

$$Odds = \frac{p}{1-p}$$

Example: 20 out of 100 in group 1 are cured after treatment:

$$Odds_1 = \frac{0.20}{0.80} = 0.25$$

...and 10 out of 100 in group 2

$$Odds_2 = \frac{0.10}{0.90} \approx 0.111$$

To compare the two groups the odds ratio (OR) is often used:

$$OR = \frac{Odds_1}{Odds_2} = \frac{0.25}{0.111} \approx 2.52$$

The Odds of being cured are 2.52 times higher in group 1.

A logistic regression model is

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

If x = 0 then

$$\log\left(\frac{p}{1-p}\right) = \beta_0$$

If x = 1, then

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1$$

To compare a group with x = 0 with a group with x = 1, the Odds ratio is.

$$OR = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

Thus, the regression parameter, when exponentiated, can be interpreted as an odds ratio.

Example 4: Poisson model ("Poisson regression")

The Poisson distribution is often used to model "count data"

Number of wireworms/plot in a Latin Square experiment with 5 treatments (Snedecor and Cochran,1960)

Row	Column				
	1	2	3	4	5
1	P 3	O 2	N 5	K 1	M 4
2	M 6	K 0	O 6	N 4	P 4
3	0 4	M 9	K 1	P 6	N 5
4	N 17	P 8	M 8	O 9	K 0
5	K 4	N 4	P 2	M 4	O 8

(K M N O P are treatments, the numbers are number of wireworms)

SAS program

```
PROC GLIMMIX data=Poisson PLOTS=PearsonPanel;
CLASS row col treat ;
MODEL count = row col treat/
DIST=poisson LINK=log ;
LSMEANS treat/adjust=Tukey ilink;
RUN;
```

Output

Fit Statistics	
-2 Log Likelihood	97.12
AIC (smaller is better)	123.12
AICC (smaller is better)	156.21
BIC (smaller is better)	138.97
CAIC (smaller is better)	151.97
HQIC (smaller is better)	127.52
Pearson Chi-Square	18.01
Pearson Chi-Square / DF	1.50

(p=0.12)

Type III Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
row	4	12	3.63	0.0367
col	4	12	0.74	0.5847
treat	4	12	4.06	0.0263

Differences of treat Least Squares Means Adjustment for Multiple Comparisons: Tukey-Kramer							
treat	**_treat**	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**	**Adj P**
K	M	-1.6707	0.4504	12	-3.71	0.0030	0.0204
K	N	-1.7121	0.4445	12	-3.85	0.0023	0.0160
K	O	-1.5801	0.4523	12	-3.49	0.0044	0.0296

…and so on

Some model fitting issues

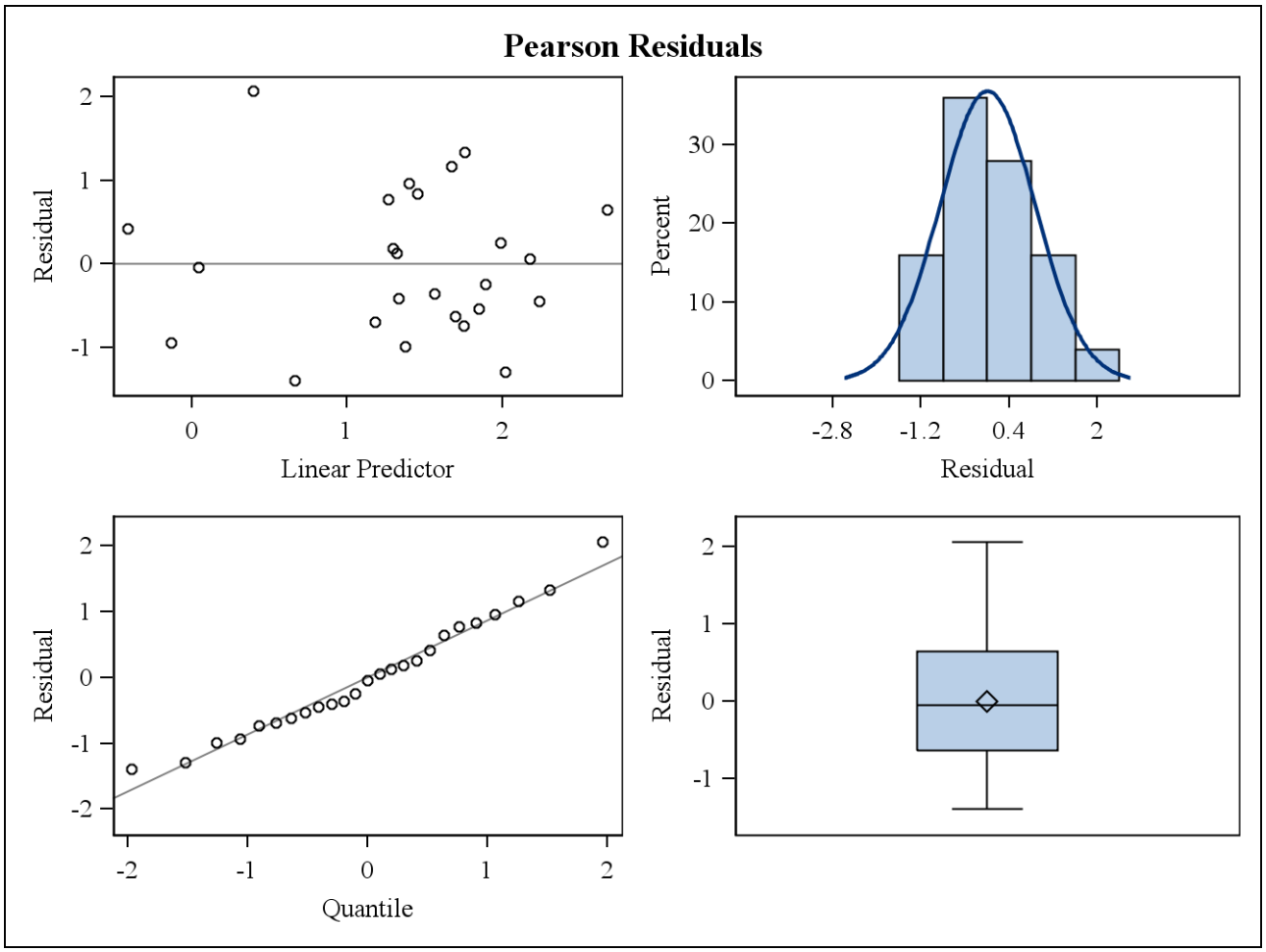
Models may be assessed using:

Deviance	For most models, Deviance/d.f. should be close to 1
	(or at least non-significant, interpreted as Chi-square)

Over-dispersion: When Deviance/df is “large”. Makes p-values “too small”.

	The deviance can be used to compare models (χ^2 tests), but
AIC	(Akaike Information Criterion) is better for that purpose
Residuals	(Pearson residuals) residuals should be approximately Normal; see example on next page
R-square	In general: not available. Some types of GLIM models have “Pseudo R^2”
Stepwise	For example, stepwise logistic regression

Example of residual plot (Poisson regression data)



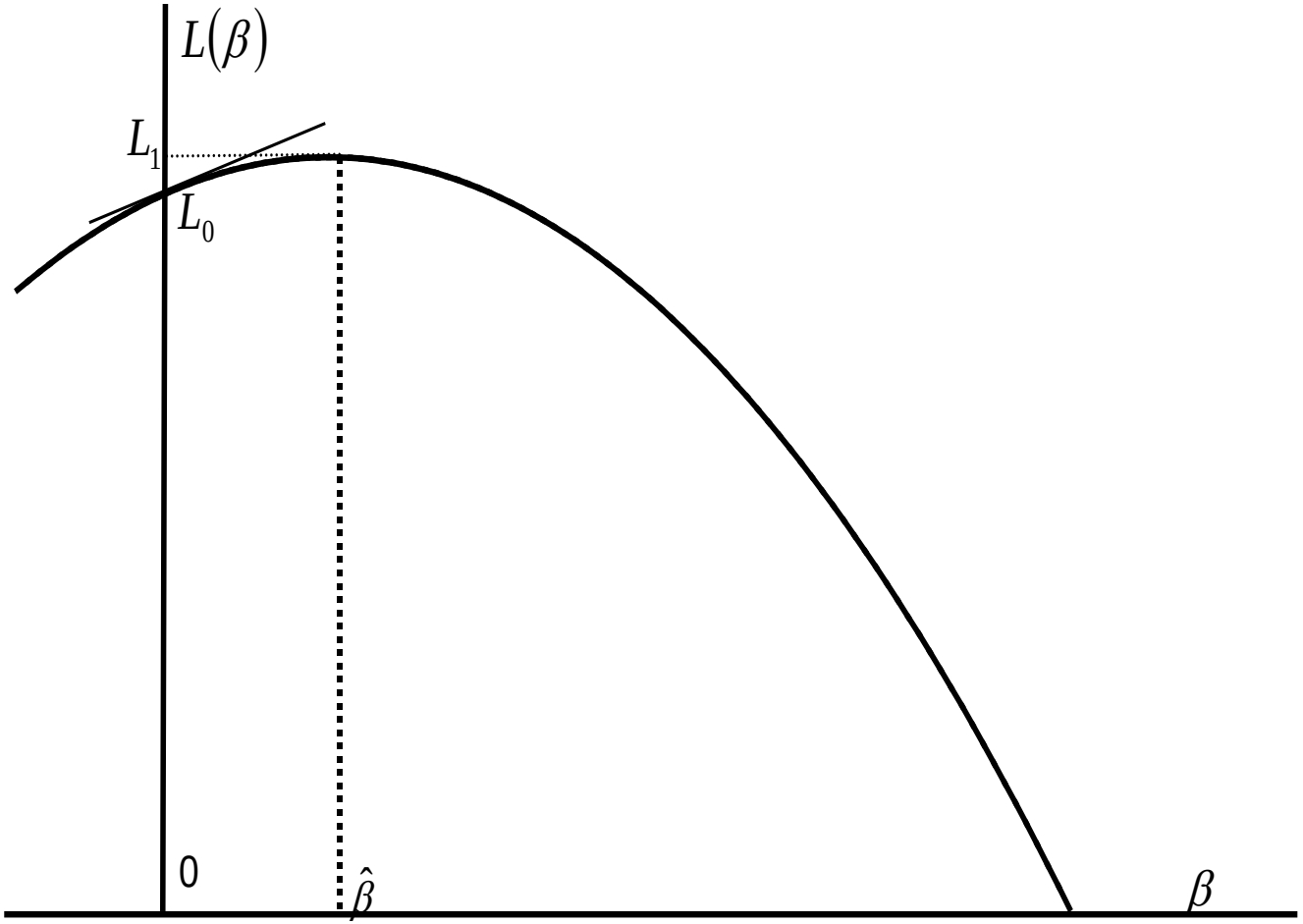
Wald, LR and Score tests

Programs for GLIM may use different methods for test construction:

Wald tests	$z = \frac{\hat{\beta}}{s.e.(\hat{\beta})}$ (or $\chi^2=z^2$)
Likelihood ratio test	Based on difference in the log likelihood function.
	X^2 approximation or F approximation
Score tests	Based on the slope of the log likelihood at H_o

All these tests are Large-sample tests

The variety may cause confusion: Why are different p values given in different parts of the output?


$L(\beta)$
L_1
L_0
0
$\hat{\beta}$
β

Over-dispersion (More in a separate lecture)

May be present when deviance/df is “large”.

(But: wrong choice of model may also affect the deviance)

Symptom: The variance in the data is larger than expected, for the chosen distribution

Example: In a Poisson distribution, the mean value is μ and the variance is also μ. If the observed variance is larger than the mean, we may have over-dispersion

Causes: Often some form of clustering in the data.

Remedies:

1. Choose some other distribution
2. Force Deviance/df to be exactly 1
3. Use robust (“sandwich”) estimators

Ordinal data

Ordinal data: e.g. school marks, assessment of symptom severity on a scale 1 2 3 4 5

One approach:

Assume that the data were generated from an unknown distribution as

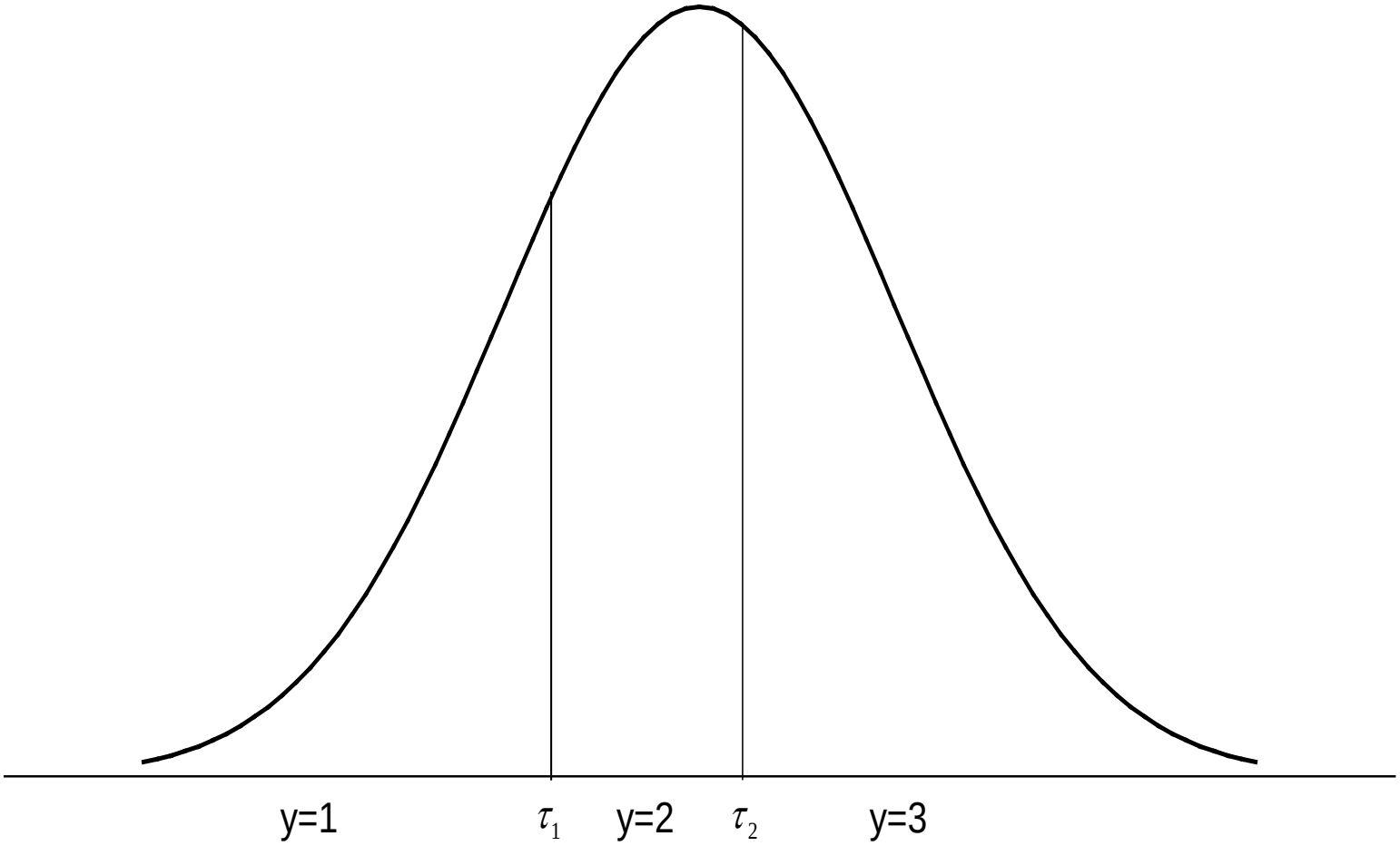
$y=1$ if $\eta < \tau_1$

$y=2$ if $\tau_1 \leq \eta < \tau_2$

⋮

$y=s$ if $\tau_{s-1} \leq \eta$

$\tau_1, \tau_2 \ldots$ are called thresholds


y=1
τ_1
y=2
τ_2
y=3

Another approach ("proportional odds model")

Make one logistic regression for each threshold.

Illustrated for a simple regression model with three ordered categories 1 2 3:

$$\text{logit}(P(y \leq 1) = \alpha_1 + \beta x$$

$$\text{logit}(P(y \leq 2) = \alpha_2 + \beta x$$

Assume that the intercepts are different but the slopes equal

It turns out that these two approaches are mathematically identical.

Example 6: Ordinal regression

Treatment for arthritis pain. Response:

2 = “Marked improvement”

1 = “Some improvement”

0 = “No improvement”

Data:

Gender	Treatment	Marked	Some	None
Female	Active	16	5	6
Female	Placebo	6	7	19
Male	Active	5	2	7
Male	Placebo	1	0	10

SAS program:

```
PROC GLIMMIX data=a;
CLASS gender treatment;
MODEL y = gender treatment gender*treatment
   /dist=mult link=cumlogit;
   FREQ f;
RUN;
```

Output:

Type III Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Gender	1	79	5.23	0.0248
Treatment	1	79	9.76	0.0025
Gender* Treatment	1	79	0.29	0.5945

Mixed models

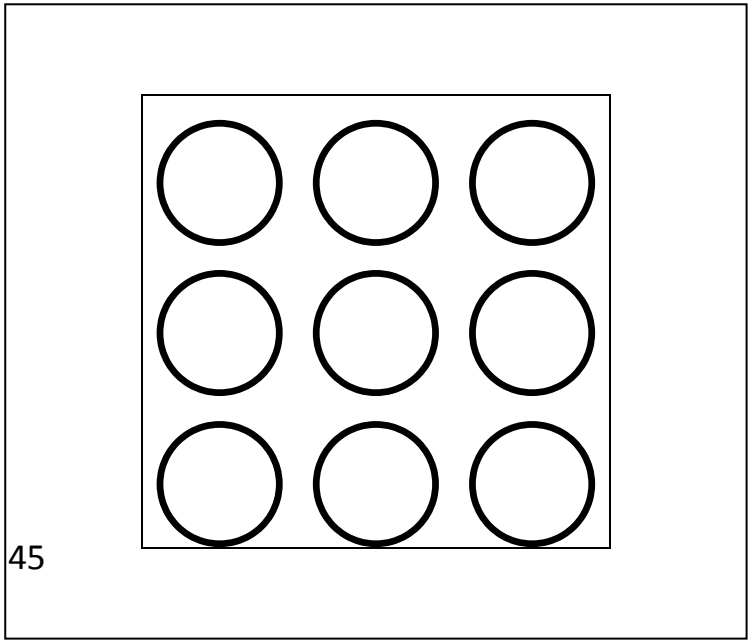
Are used when we make several measurements on the same

Experimental unit

The smallest unit that gets an individual treatment

Examples:

- Subsampling
- Clustered data
- Repeated measurements



Example 7: A Mixed Generalized Linear Model

Purpose: investigate whether different treatments have different attraction on ladybirds and whether this changes with time.

3 treatments

3 replicates of each treatment

6 time points (5, 6, 9, 13, 16 and 21 days). Response y: Number of ladybirds

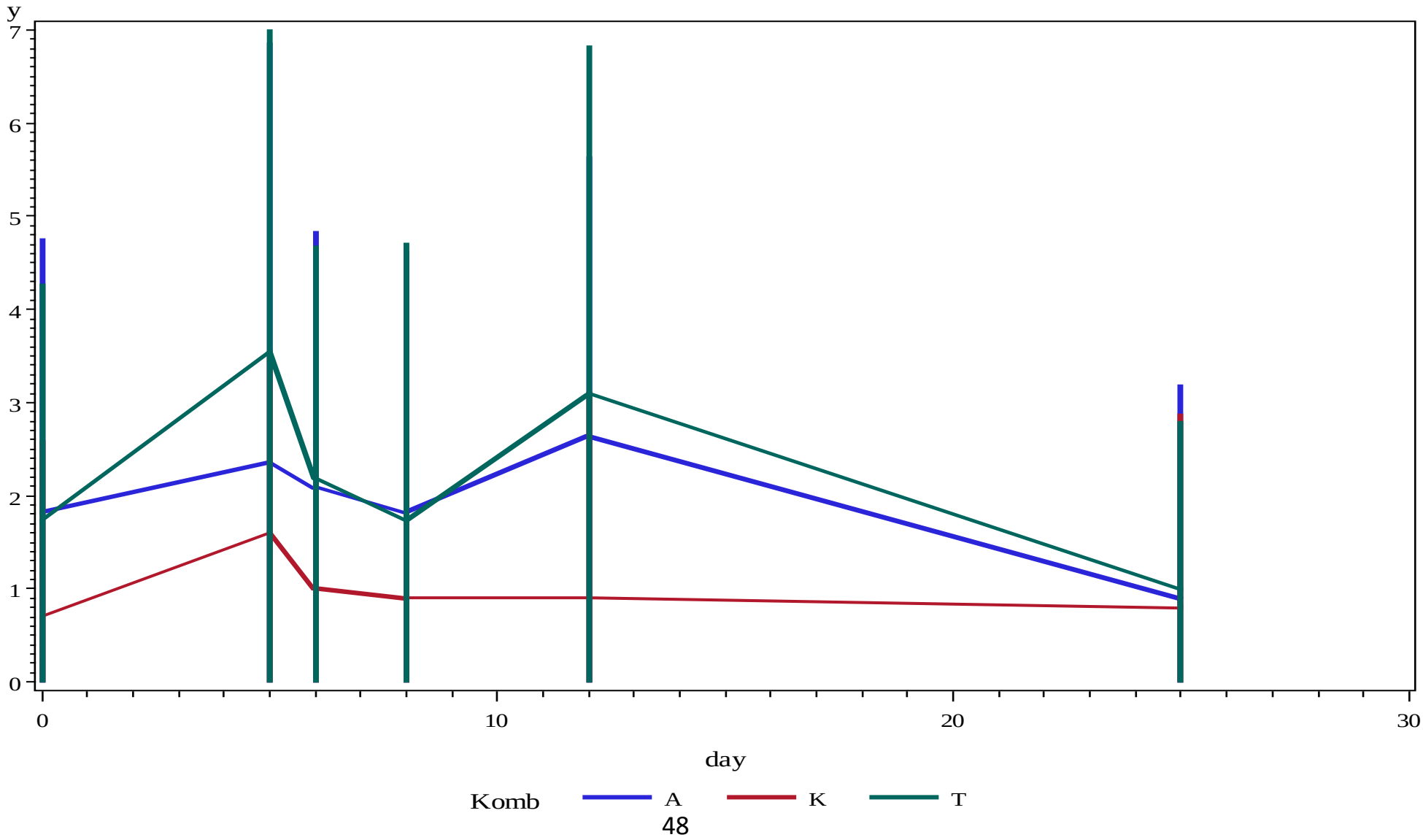
```
PROC GLIMMIX data=ladybird ;
CLASS parcell komb day ;
MODEL y = komb day komb*day /dist=Poisson link=log ;
RANDOM residual / type=sp(pow)(day) subject=parcell*komb ;
lsmeans komb /pdiff;
RUN;
```

Output:

Fit Statistics	
-2 Res Log Pseudo-Likelihood	475.72
Generalized Chi-Square	190.69
Gener. Chi-Square / DF	1.10

Type III Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Komb	2	29	10.91	0.0003
day	5	145	4.45	0.0008
Komb* day	10	145	0.50	0.8853

("komb" is the treatment)


y
7
6
5
4
3
2
1
0
0
10
20
30
day
Komb
A
K
T

Model building

- Include all relevant main effects (even non-significant ones)
- If the A*B interaction is in the model, it should also include A and B
- In polynomial models, include ALL terms lower than the chosen degree
- Use tools such as Akaike Information Criterion (AIC) to choose between models (Do not care too much about p values when building models!)

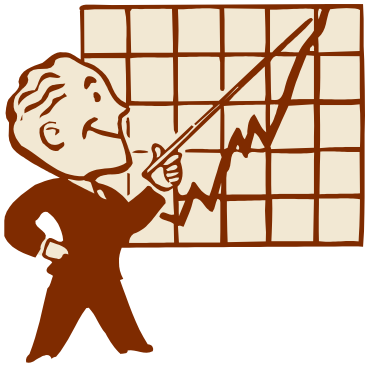
Deviance/df is affected

By the choice of distribution and link

By the linear model used

So “overdispersion” may be caused by using a bad MODEL!

”All models are wrong…



…but some are useful.” (G. E. P. Box)

References

Bliss, C. I. (1934). The Method of Probits. Science 79: 38-39.

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

McCullagh, P. and Nelder, J.A. (1989): Generalized linear models. London, Chapman and Hall.

Ninkovic, V., Sate Al Abassi, Jan Pettersson, The Influence of Aphid-Induced Plant Volatiles on Ladybird Beetle Searching Behavior, Biological Control, Volume 21, Issue 2, June 2001,

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

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

The R system: http://ftp.sunet.se/pub/lang/CRAN/

In particular the Glm and lme4 packages

SAS Institute Inc. (2011): SAS/Stat 9.3 user's guide. Cary, N.C., SAS Institute Inc.

(In particular: the Glimmix procedure)

Snedecor, G, W. and Cochran, W. G. (1960): Statistical methods. Ames, Iowa State University Press.

R scripts

Example 2

```
ex2 <- glm(cbind(y,n-y)~x, data=beetles,
family=binomial(link=”logit”))

summary(ex2)
```

Example 3

```
ex3 <- glm(cbind(nblood,freq-nblood)~hybrid*food, data=blood,
family=binomial(link=”logit”))

summary(ex3)
```

Example 4

```
ex4 <- glm(count~method, data=o, family=poisson(link=”log”))

summary(ex4)
```

Example 5

```
library(MASS)

ex5 <- glm.nb(count~method, data=o)

summary(ex5)
```

Example 6

```
library(VGAM)

ex6 <- vglm(cbind(Marked,Some,None)~Gender*Treatment, data=a,
family=propodds)

summary(ex6)
```

Example 7

```
ex7 <- glmmPQL(y~komb*day, random=~1|komb/parcell,
correlation=corCAR1(form=~day|komb/parcell), data=ladybird,
family=poisson(link="log"))

summary(ex7)
```