Faculty of Health Sciences

Correlated data

Count variables

Lene Theil Skovgaard

December 8, 2015

Non-normal outcomes

- The Poisson distribution for counts
- Poisson models, log-linear models
- Overdispersion
- Generalized linear mixed models
  - Population average models (PA)
  - Subject specific models (SS)
- Examples:
  - Leprosy
  - Seizures (briefly)

Study on epilepsy

Controlled clinical trial, with 58 epileptic patients:

- 28 treated with placebo
- 31 treated with pragabide=active

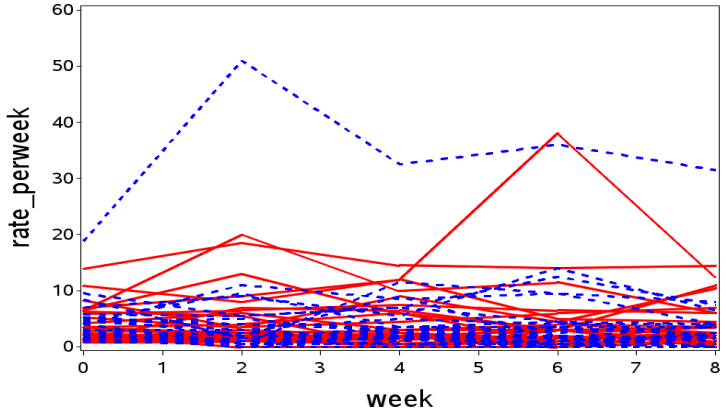
Recording of the number of epileptic seizures during

- an 8-week interval before treatment
- 4 2-weeks intervals after treatment

Reference: Thall, P.F. and Vail, S.C. (1990). Some covariance models for longitudinal count data with overdispersion. Biometrics.

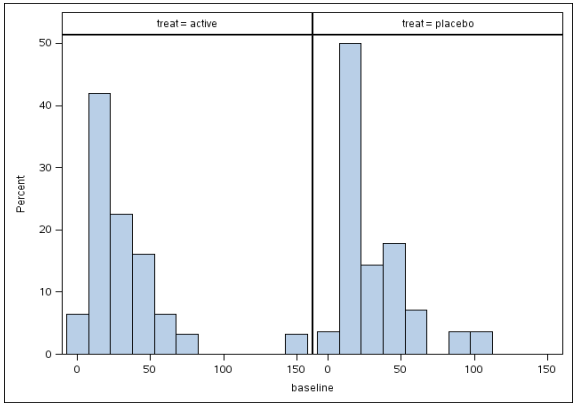
Spaghettiplot - the epilepsy example

Number af seizures per week (rates):



Legends: Progabide Placebo

Baseline measurements, total number of seizures



Note the skew distributions

Do we see a difference between the groups at baseline?

Number of seizures at baseline

The MEANS Procedure

Analysis Variable : baseline

treat	N Obs	N	Mean	Std Dev	Minimum	Maximum
active	31	31	31.6451613	27.9935092	7.0000000	151.0000000
placebo	28	28	30.7857143	26.1042882	6.0000000	111.0000000

Note:

- The variance is obviously bigger than the average (overdispersion, to be discussed later)
- We obviously do *not* have normal distributions here
- Maybe on log-scale? But we have zeroes at later times.....

Binary data

Examples of binary outcomes:

- infection after surgery
- smoking among school children
- seizures on a single day (or hour)

A binary variable X has a Bernoulli distribution, meaning that

- $P(U = 1) = p$
- $P(U = 0) = 1 - p$

For such an outcome, the mean value is $\mathrm{E}(U) = p$,
and the variance is $\mathrm{Var}(U) = p(1 - p)$

Binomial data

If we sum up n binary observations,

$$Y = \sum_{i=1}^{n} U_i = U_1 + \cdots + U_n$$

e.g.

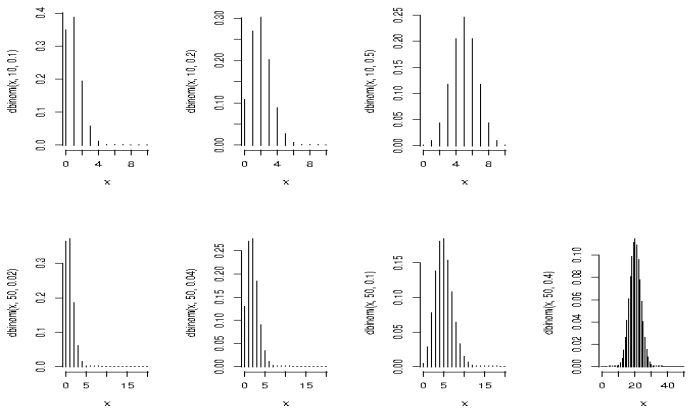
- number of infections for each hospital
- number of smokers in each school class
- number of seizures in a specific interval

we get a Binomial distribution, $Y \sim \mathrm{Bin}(n, p)$, with

$\mathrm{E}(Y) = np, \quad \mathrm{Var}(Y) = np(1 - p)$

Examples of Binomial distributions

n=10, 50; np=1, 2, 5 or 20 (mean value)



Approximations to the Binomial distribution

The Binomial variable Y has point probabilities

$$P(Y = m) = \binom{n}{m} p^m (1-p)^{n-m}$$

Its mean is $\mu = np$ and its variance $V = np(1-p)$

When n is large, this distribution is very intractable, so we use approximations

- p moderate (not too close to 0 or 1) and $np > 5$: Normal distribution
- p close to 0: Poisson distribution

Poisson distribution

Counts with no well-defined upper limit:

- the number of cancer cases in a specific community during a specific year
- the number of positive swabs over a certain period of time

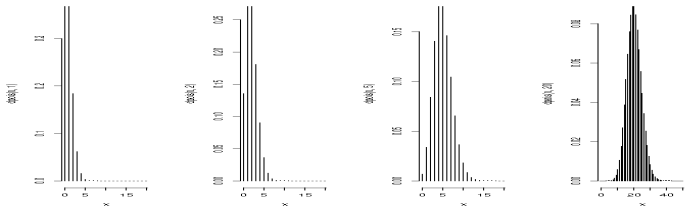
Law of rare events:
As the count parameter n in a Binomial distribution gets larger and the parameter p gets close to 0, the Binomial distribution is approximately equal to the Poisson distribution

$$P(Y = m) = \frac{\mu^m}{m!} \exp(-\mu)$$

where $\mu = np$ is the mean value,
as well as the variance, i.e. $V = np$ as well.

Poisson distribution

Poisson distribution with mean value: μ =1,2,5 and 20



Important note:
In a Poisson distribution, the mean and variance are equal:

$$\mathrm{E}(Y) = \mu = V = \mathrm{Var}(Y)$$

This fact is unfortunately often overlooked....

Models for non-normal data

Generalized linear models

are just like Multiple regression models, but
on a scale that *corresponds* to the data:

- Normal (link=`identity`),
  traditional linear models
- Binomial (link=`logit`),
  logistic regression (next lecture)
- Poisson (link=`log`),
  Log-linear models, Poisson regression

Generalized linear models, for count data

Outcome variable Y_i, following a Poisson distribution, with

- Mean value: $\mathrm{E}(Y_i) = \mu_i$
- Link funktion: $\log$, the *natural* logarithm.
  On this scale, we assume linearity in the covariates, i.e.

$$\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} (= X_i^T \beta)$$

where $x_{i1}, \ldots, x_{ik}$ denote the covariate values for individual i.

The $\log$-link ensures that $\mu_i = E(Y_i)$ will always be positive

Comparing two distributions of counts

Comparison of distributions from p. 5:
Do we see a difference in mean number of seizures at baseline in the two groups?

- This problem corresponds to a T-test
  (in case of Normal distributions)
- We do not expect any difference, since we are dealing with a randomized study

Model: $Y_i \sim \text{Poisson}(\mu_i), \quad \log(\mu_i) = \beta_t$
where the subscript t denotes **t**reatment, which can be either active (=progabide) or placebo.

Poisson analysis in SAS

```
proc genmod data=seizures; where week=0;
     class treat;
     model baseline = treat / dist=poisson link=log;
estimate "active vs. placebo" treat 1 -1;
run;
```

with output

```
The GENMOD Procedure

        Model Information

Data Set              WORK.SEIZURES
Distribution               Poisson
Link Function                  Log
Dependent Variable        baseline

Number of Observations Read          59
Number of Observations Used          59
```

Output, II

```
      Class Level Information
 Class      Levels    Values
treat            2    active placebo

            Criteria For Assessing Goodness Of Fit

Criterion                     DF           Value          Value/DF
Deviance                      57       1059.5228           18.5881
Scaled Deviance               57       1059.5228           18.5881
Pearson Chi-Square            57       1340.5348           23.5182
Scaled Pearson X2             57       1340.5348           23.5182

Algorithm converged.
```

The values above 1 in the last column indicates a misfit, see p. 28-29 and p. 37

Output, III

```
                Analysis Of Maximum Likelihood Parameter Estimates

                                                Standard       Wald 95%            Wald
Parameter            DF  Estimate     Error  Confidence Limits  Chi-Square
Intercept             1    3.4271    0.0341    3.3603    3.4938     10123.9
treat      active     1    0.0275    0.0467   -0.0640    0.1190        0.35
treat      placebo    0    0.0000    0.0000    0.0000    0.0000           .
Scale                 0    1.0000    0.0000    1.0000    1.0000


Parameter            Pr > ChiSq
Intercept                <.0001
treat      active         0.5553
treat      placebo             .
Scale

NOTE: The scale parameter was held fixed.
```

Estimate of $\beta_{\text{active}} - \beta_{\text{placebo}} = 0.0275$

This is on log-scale

Output, IV

Output from `Estimate` statement:

```
                    Contrast Estimate Results

                        Mean          Mean          L'Beta  Standard
Label                Estimate    Confidence Limits Estimate     Error
active vs. placebo     1.0279     0.9380    1.1264    0.0275    0.0467

                                L'Beta          Chi-
Label                  Alpha    Confidence Limits  Square  Pr > ChiSq
active vs. placebo      0.05    -0.0640    0.1190    0.35      0.5553
```

Estimated ratio:

$$\exp(\beta_{\text{active}} - \beta_{\text{placebo}}) = \frac{\exp(\beta_{\text{active}})}{\exp(\beta_{\text{placebo}})} = 1.028$$

with CI=(0.938, 1.126)

The active group starts off a little worse than the placebo group.

New example: Counts of leprosy bacilli

Controlled clinical trial:

- 10 patients treated with placebo P
- 10 patients treated with antibiotic A
- 10 patients treated with antibiotic B

Recording of the number of bacilli at six sites of the body, i.e. a count variable

- before treatment (baseline, `time=0`)
- several months after treatment, (`time=1`)

Reference: Snedecor, G.W. and Cochran, W.G. (1967). Statistical Methods, (6th edn). Iowa State University Press

Averages for the leprosy example

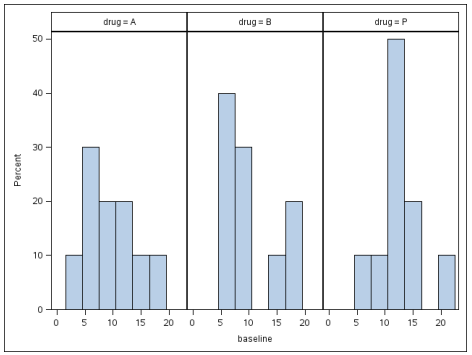
Analysis Variable : bacilli

drug	time	N Obs	N	Median	Mean	Variance
A	0	10	10	9.0000000	9.3000000	22.6777778
	1	10	10	5.0000000	5.3000000	21.5666667
B	0	10	10	8.0000000	10.0000000	27.5555556
	1	10	10	3.5000000	6.1000000	37.8777778
P	0	10	10	12.0000000	12.9000000	15.6555556
	1	10	10	12.5000000	12.3000000	51.1222222

	time	N Obs	N	Median	Mean	Variance
All	0	30	30	10.5000000	10.7333333	22.9609195

Note: The variance is obviously bigger than the average.....overdispersion

Baseline measurements



Do we see a difference between the baseline counts?

Poisson analysis in SAS

Code just like as on p. 16
except for the option `type3`,
which compares all 3 drugs simultaneously, like an ANOVA

```
proc genmod data=leprosy; where time=0;
     class drug;
     model baseline = drug /
                    d=poisson link=log type3;
     estimate 'Effect B minus A' drug -1 1 0;
     estimate 'Antibiotic effect' drug 0.5 0.5 -1;
     contrast 'Antibiotic effect?' drug 1 1 -2;
run;
```

The estimate `Antibiotic effect` evaluates $\frac{A+B}{2} - P$,
i.e. the average effect of the two active drugs.

Output

The GENMOD Procedure

Model Information

Data Set	WORK.LEPROSY
Distribution	Poisson
Link Function	Log
Dependent Variable	baseline
Number of Observations Read	30
Number of Observations Used	30

Class Level Information

Class	Levels	Values
drug	3	A B P

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	27	55.0366	2.0384
Scaled Deviance	27	55.0366	2.0384
Pearson Chi-Square	27	57.6687	2.1359
Scaled Pearson X2	27	57.6687	2.1359

Output, II

Analysis Of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square
Intercept		1	2.5572	0.0880	2.3847	2.7298	843.58
drug	A	1	-0.3272	0.1360	-0.5938	-0.0606	5.79
drug	B	1	-0.2546	0.1332	-0.5158	0.0065	3.65
drug	P	0	0.0000	0.0000	0.0000	0.0000	.
Scale		0	1.0000	0.0000	1.0000	1.0000	

Parameter		Pr > ChiSq
Intercept		<.0001
drug	A	0.0162
drug	B	0.0560
drug	P	.
Scale		

NOTE: The scale parameter was held fixed.

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
drug	2	6.62	0.0364

Unfortunately, we have a significant difference already at baseline

Output, III

Contrast Estimate Results

Label	Mean Estimate	Mean Confidence Limits		L'Beta Estimate	Standard Error	Alpha
Effect B minus A	1.0753	0.8108	1.4261	0.0726	0.1441	0.05
Antibiotic effect	0.7476	0.5982	0.9343	-0.2909	0.1138	0.05

Label	L'Beta Confidence Limits		Chi-Square	Pr > ChiSq
Effect B minus A	-0.2098	0.3549	0.25	0.6144
Antibiotic effect	-0.5139	-0.0680	6.54	0.0105

Contrast Results

Contrast	DF	Chi-Square	Pr > ChiSq	Type
Antibiotic effect?	1	6.40	0.0114	LR

The two active drugs have a somewhat lower level of bacilli, estimated to be 75% of that of placebo patients, with CI=60-93% and P=0.01.
This is a nuisance.... but is it real?

Overdispersion

Overdispersion:
The variance has been noted to be larger than expected for a Poisson distribution.
This may be caused by

- omitted covariates (isn't that always the case?)
- unrecognized clusters
- heterogeneity, e.g. a "zero"-group (non-susceptibles)

Traditional solution:
An over-dispersion parameter ϕ can be estimated and multiplied onto the standard errors

Overdispersion, II

When overdispersion is disregarded

- The standard errors are erroneously small
- The P-values are erroneously small
- We get type I errors

Multiplying with the over-dispersion parameter $\hat{\phi}$ yields

- Larger standard errors
- Larger P-values

ϕ is estimated from either `Pearson Chi-Square Value/DF` or `Pearson Chi-Square Value/DF` (see p. 24), by taking square roots.
Use options `scale=p` or `scale=d`.

Overdispersion in PROC GENMOD, scale=pearson

```
proc genmod data=leprosy; where time=0;
     class drug;
     model baseline = drug /
                    d=poisson link=log type3 scale=pearson;
     estimate 'Effect B minus A' drug -1 1 0;
     estimate 'Antibiotic effect' drug 0.5 0.5 -1;
     contrast 'Antibiotic effect?' drug 1 1 -2;
run;
```

Analysis Of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr>ChiSq
Intercept		1	2.5572	0.1287	2.3050	2.8094	394.96	<.0001
drug	A	1	-0.3272	0.1988	-0.7169	0.0624	2.71	0.0998
drug	B	1	-0.2546	0.1947	-0.6363	0.1270	1.71	0.1910
drug	P	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.4615	0.0000	1.4615	1.4615		

NOTE: The scale parameter was estimated by the square root of Pearson's Chi-Square/DOF.

LR Statistics For Type 3 Analysis

Source	Num DF	Den DF	F Value	Pr > F	Chi-Square	Pr > ChiSq
drug	2	27	1.55	0.2304	3.10	0.2121

Output, continued

Contrast Estimate Results

Label	Mean Estimate	Mean Confidence Limits		L'Beta Estimate	Standard Error	Alpha
Effect B minus A	1.0753	0.7117	1.6245	0.0726	0.2105	0.05
Antibiotic effect	0.7476	0.5397	1.0355	-0.2909	0.1662	0.05

Label	L'Beta Confidence Limits		Chi-Square	Pr > ChiSq
Effect B minus A	-0.3401	0.4852	0.12	0.7303
Antibiotic effect	-0.6168	0.0349	3.06	0.0801

Contrast Results

Contrast	Num DF	Den DF	F Value	Pr > F	Chi-Square	Pr > ChiSq	Type
Antibiotic effect?	1	27	2.99	0.0950	2.99	0.0835	LR

The two active drugs are still estimated to be only 75% of that of placebo patients, but now with CI=54-104% and P=0.08, i.e. not significant any more.

Problems with overdispersion

Overdispersion is *unnatural*:

- The distribution does not exist
- The multiplication factor $\hat{\phi}$ is defined ad-hoc

If we believe the over-dispersion to be caused by omitted covariates, it would be more natural to include an extra random variation, e.g.

$$\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + b_i$$

with some assumption on the distribution of the b_i's, i.e. with $\exp(b_i)$ multiplied on the mean value

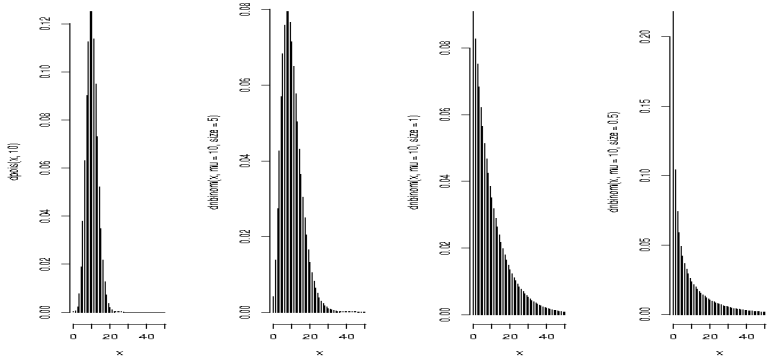
Additional random variation

Possible models for b_i:

- $b_i \sim N(0, \omega_b^2)$: leads to a complicated model, which changes the level of the mean, since $E(\exp(b_i)) = \exp(\omega_b^2/2) > 1$
- $b_i \sim$ log Gamma: leads to Y_i being distributed as a Negative binomial distribution:
  - Overdispersed counts
  - Unbounded positive range
  - Variance varies independent of mean
  - $\mathrm{E}(Y_i) = \mu_i$
  - $\mathrm{Var}(Y_i) = \mu_i + \theta\mu_i^2$

Negative binomial distributions, with mean 10

Poisson distribution, followed by 3 negative binomial distributions, with variance 30, 110 and 210:



Negative binomial analysis in SAS

The code is identical to that on p. 23, except for `dist=negbin`.

Analysis Of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square
Intercept		1	2.5572	0.1250	2.3122	2.8022	418.43
drug	A	1	-0.3272	0.1851	-0.6900	0.0356	3.13
drug	B	1	-0.2546	0.1830	-0.6134	0.1041	1.94
drug	P	0	0.0000	0.0000	0.0000	0.0000	.
Dispersion		1	0.0788	0.0450	0.0257	0.2412	

Parameter		Pr > ChiSq
Intercept		<.0001
drug	A	0.0771
drug	B	0.1642
drug	P	.
Dispersion		

NOTE: The negative binomial dispersion parameter was estimated by maximum likelihood.

Output, continued

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
drug	2	3.35	0.1874

Contrast Estimate Results

Label	Mean Estimate	Mean Confidence Limits		L'Beta Estimate	Standard Error	Alpha
Effect B minus A	1.0753	0.7394	1.5637	0.0726	0.1911	0.05
Antibiotic effect	0.7476	0.5492	1.0176	-0.2909	0.1573	0.05

Label	L'Beta Confidence Limits		Chi-Square	Pr > ChiSq
Effect B minus A	-0.3019	0.4470	0.14	0.7041
Antibiotic effect	-0.5993	0.0174	3.42	0.0644

Contrast Results

Contrast	DF	Chi-Square	Pr > ChiSq	Type
Antibiotic effect?	1	3.23	0.0722	LR

Effect of overdispersion

Leprosy example, baseline comparison:

LR test, Drug difference?	$\chi^2(2)$	P-value
Poisson	6.62	0.036
- with overdispersion	3.10	0.21
Negative Binomial	3.35	0.19

Effect (A and B) vs. P	Ratio (CI)	P-value Wald
Poisson	0.75 (0.60, 0.93)	0.011
- with overdispersion	0.75 (0.54, 1.04)	0.080
Negative Binomial	0.75 (0.55, 1.02)	0.064

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

∗Effect of overdispersion, II

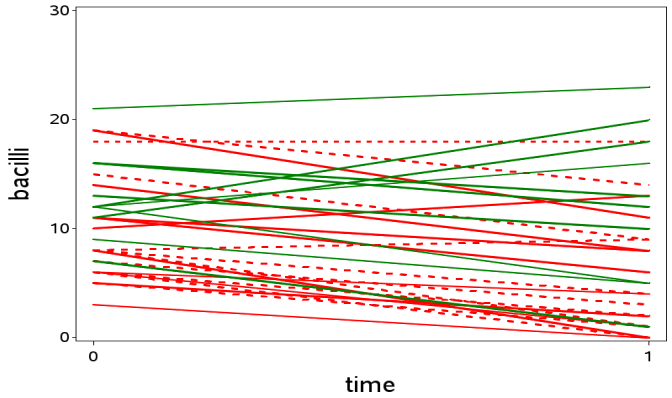
Recall the Seizure example, from p. 3:

Baseline comparison:

Effect: Active vs. Placebo	Ratio (CI)	P-value LR
"as usual"	1.028 (0.938, 1.126)	0.56
with overdispersion	1.028 (0.660, 1.602)	0.90
Negative Binomial	1.028 (0.708, 1.493)	0.89

Spaghettiplot - the leprosy example

Now including both time points (0 and 1):

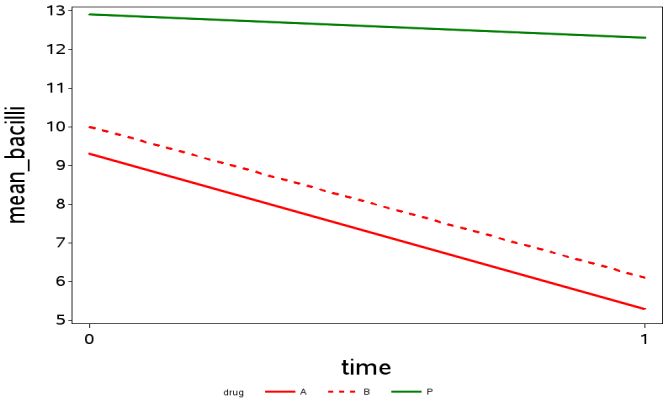


Legends:
A —— B ..... P ——

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

Average plot - the leprosy example

Note: New scaling, different from p. 38



Legends:
A —— B ..... P ——

Possible purposes of the investigation

1. Evaluate the efficiency of antibiotics:
   red lines vs green line
2. Compare the two drugs, A and B:
   solid vs dotted red line
3. Quantify the effects of each of the two antibiotic drugs separately

Randomization:
At baseline, all patients have the same expected mean count (mean value), but *by chance*, the placebo individuals have somewhat larger values than the remaining groups.

Model reflections

This is just a before-after study....*but*

- We are dealing with counts, so it is natural to consider a Poisson distribution, with `log`-link (natural $\log$)
- Because it is a randomized study, the mean values at baseline should be identical for the three groups
- We are prepared to see 3 different changes over time - but some of these may be identical (this is actually the main scientific question)
- Baseline and follow-up measurement are correlated within individuals (just like a random effect of individual)

Model reflections, II

- Can't we just take logarithms?
  No, because we have zeroes
- Some other transformation then?
  Yes, square roots, or arcsine,
  but the interpretation would suffer a lot
- Could we just condition on the baseline value?
  Yes, we could do that.....*but*
  it becomes more tricky when we have multiple time points
- Could we analyze differences? Or rather, ratios? Hmm....
- Could we build a Constrained Model,
  forcing mean values to be equal at baseline?

Model reflections, III

Parametrization of mean values (on the `log`-scale):

Treatment	Period	Mean (on log scale)
P	Baseline	β_1
P	Follow-up	$\beta_1 + \beta_2$
A	Baseline	β_1
A	Follow-up	$\beta_1 + \beta_2 + \beta_3$
B	Baseline	β_1
B	Follow-up	$\beta_1 + \beta_2 + \beta_4$

β_3 resp. β_4 denote additional effects of A and B, when compared to placebo

Generalized linear MIXED models

Outcome variable Y_{ij}, e.g. j'th measurement time for individual i:

Mean value: μ_{ij}
Link funktion g: $g(\mu_{ij})$ is assumed linear in covariate *vector* X_{ij}.
Two kinds of models:

- Population average models (PA):
  $g(\mu_{ij}) = \beta_0 + \beta_1 x_{ij1} + \cdots + \beta_k x_{ijk} = X_{ij}^T \beta$
  and (Y_{ij_1}, Y_{ij_2}) are associated (correlated),
  with som (patterned) covariance
- Subject-specific models (SS):
  $g(\mu_{ij}) = \beta_0 + \beta_1 x_{ij1} + \cdots + \beta_k x_{ijk} + b_i$
  $b_i \sim N(0, \omega_b^2)$, random intercepts (levels)
  may be generalized to other random effects: slopes,...

The two model types

Marginal models: or Population average (PA):
Describe covariate effects on the population mean, e.g. expected difference between the effects of two treatments
(corresponds to the `repeated` statement)

Mixed effects model: or Subject specific (SS):
Describe covariate effects on specific individuals (or clusters), e.g. expected change over time, or differences between boys and girls in the same school class
(corresponds to the `random` statement)

For traditional linear models (Normality)

with *identity* link:

Subject-specific model with random intercept/level

is equal to

Marginal model with compound symmetry covariance structure (`type=CS`)

More generally:
The interpretation of the parameters β does not depend on the way that we model the covariance/correlation
(although the estimate may change somewhat depending on the assumed structure)

For non-normal outcomes

The above is no longer true *in general*,
due to *non-linearity* of the link-function

This means:
The interpretation of the parameters β does depend on the way that we model the covariance/correlation.
...with some important exceptions (in this lecture)

This implies that effects may either be interpreted cross-sectionally (marginally, for comparison of different populations, say, of different age) or subject-specific (effect of ageing for a single individual)

Marginal models = Population Average (PA)

We specify only

- Marginal mean, $E(Y_{ij}|X_{ij}) = \mu_{ij}$, where $\log(\mu_{ij}) = X_{ij}^T\beta$, i.e. covariate effects as usual
- Distribution..... Poisson (in a way), but...
- Marginal variance, $\phi V(\mu_{ij}) = \phi\mu_{ij}$ (overdispersion)
- Some measure of association for Y's belonging to the same individual/unit, $\mathrm{V}_i = \mathrm{Cov}(Y_i)$

This creates problems:

- Multivariate Poisson distribution does not exist
- It is more of an estimation procedure rather than a model

∗ Marginal models, technicalities

Since we do not actually have a model,
we cannot use a maximum likelihood approach.
This has implications for the handling of missing values (lecture 4).

Instead, we use a socalled
GEE: Generalized estimating equation,
(written in vector notation)

$$\sum D^T V_i^{-1}(y_i - \mu_i) = 0$$

where V_i is the (*working*) covariance matrix for Y_i, and D_i is the matrix of derivatives of the mean value μ_i with respect to β

∗Marginal models, technicalities II

The GEE-method

- requires an iterative procedure,
- gives consistent estimates of β
  (they have the correct mean when the sample size is large),
  even if $V_i \neq \text{Cov}(Y_i)$ is incorrectly specified
- standard error of $\hat{\beta}$ should be based on the
  empirical sandwich estimator,
  to allow for possible overdispersion and general
  misspecification of Cov(Y_i)
  (only possible for balanced designs)
- the estimates are asymptotically Normal
  (i.e for large samle size, we can construct confidence intervals
  with plus/minus 2 standard errors)

Marginal model (PA) for leprosy

With the variable definitions:

```
A_effect=(drug='A')*time;
B_effect=(drug='B')*time;
```

we write the code, including a `repeated`-statement:

```
proc genmod data=leprosy;
     class id;
     model bacilli= time A_effect B_effect /
                    d=poisson link=log type3;
     repeated subject=id / type=un modelse corrw;
contrast 'Antibiotic effect' A_effect 1, B_effect 1;
contrast 'Antibiotic effect?' A_effect 1, B_effect 1 / wald;
estimate 'Effect B minus A' A_effect 1 B_effect -1;
estimate "changes for A" time 1 A_effect 1 / exp;
estimate "changes for B" time 1 B_effect 1 / exp;
estimate 'changes (A,B) vs. P' A_effect 0.5 B_effect 0.5;
output out=pa pred=pred_pa;
run;
```

Comments to code

- `time` indicates the change over time for the placebo group
  (the parameter β_2)
- `A_effect` indicates the additional change over time for drug A
  (the parameter β_3)
- `B_effect` indicates the additional change over time for drug B
  (the parameter β_4)
- `d=poisson`: specifies the `link`-function as `log`, and the
  working correlation matrix as (proportional to) the mean
- `link=log`: may overrule the `link`-function from
  `dist=poisson`, if so needed
- `repeated`: specifies an unstructured (`type=un`) association
  between measurements on the same `id` (`corrw` requests
  printing)

Comments to code, II

- `estimate` statements:
  Estimate combinations of the β's, here
  - Effect of B-A: $\beta_4 - \beta_3$
  - Changes for A: $\beta_2 + \beta_3$
  - Changes for B: $\beta_2 + \beta_4$
- `contrast` statements:
  Useful for testing several parameters simultaneously, here the tests
  - $\beta_3 = \beta_4 = 0$: No (extra) effect of either A nor B
- Output data set, with predicted values `pred_pa`, for illustration purposes

Output

```
The GENMOD Procedure

        Model Information

Data Set              WORK.LEPROSY
Distribution               Poisson
Link Function                  Log
Dependent Variable         bacilli

Number of Observations Read          60
Number of Observations Used          60

                    Class Level Information

Class      Levels    Values
id             30    1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
                     21 22 23 24 25 26 27 28 29 30

  Parameter Information

Parameter       Effect
Prm1            Intercept
Prm2            time
Prm3            A_effect
Prm4            B_effect
```

Output, II

```
          GEE Model Information

Correlation Structure             Unstructured
Subject Effect                  id (30 levels)
Number of Clusters                         30
Correlation Matrix Dimension                2
Maximum Cluster Size                        2
Minimum Cluster Size                        2


Algorithm converged.

  Working Correlation Matrix

             Col1         Col2
Row1       1.0000       0.7966
Row2       0.7966       1.0000
```

Output, III: Estimation

```
The GENMOD Procedure

            Analysis Of GEE Parameter Estimates
             Empirical Standard Error Estimates

                      Standard   95% Confidence
Parameter Estimate       Error        Limits            Z Pr > |Z|
Intercept   2.3734     0.0801    2.2163    2.5304    29.62    <.0001
time       -0.0138     0.1573   -0.3222    0.2946    -0.09    0.9300
A_effect   -0.5406     0.2186   -0.9690   -0.1122    -2.47    0.0134
B_effect   -0.4791     0.2279   -0.9257   -0.0325    -2.10    0.0355
```

Here, the covariance between repeated measurements on the same subject is accounted for, even if it is misspecified (`Empirical Standard Error`)

*Output, IV: Estimation results without overdispersion

Not reasonable:, from option `modelse`:

Analysis Of GEE Parameter Estimates
Model-Based Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	2.3734	0.1035	2.1704	2.5763	22.92	<.0001
time	-0.0138	0.1111	-0.2315	0.2039	-0.12	0.9010
A_effect	-0.5406	0.1818	-0.8969	-0.1843	-2.97	0.0029
B_effect	-0.4791	0.1779	-0.8278	-0.1303	-2.69	0.0071
Scale	1.8578	.	.	.	.	.

Here, the covariance between repeated measurements on the same subject is only accounted for if it is correctly specified (`Model-Based Standard Error`, with no additional overdispersion)

Output, V (additional statements)

Contrast Estimate Results

Label	Mean Estimate	Mean Confidence Limits		L'Beta Estimate	Standard Error
Effect B minus A	0.9403	0.6148	1.4381	-0.0615	0.2168
changes for A	0.5744	0.4281	0.7707	-0.5544	0.1499
changes for B	0.6109	0.4478	0.8333	-0.4929	0.1585
changes (A,B) vs. P	0.6006	0.4097	0.8805	-0.5098	0.1952

Label	Alpha	L'Beta Confidence Limits		Chi-Square	Pr > ChiSq
Effect B minus A	0.05	-0.4864	0.3633	0.08	0.7765
changes for A	0.05	-0.8483	-0.2605	13.67	0.0002
changes for B	0.05	-0.8035	-0.1823	9.68	0.0019
changes (A,B) vs. P	0.05	-0.8924	-0.1273	6.82	0.0090

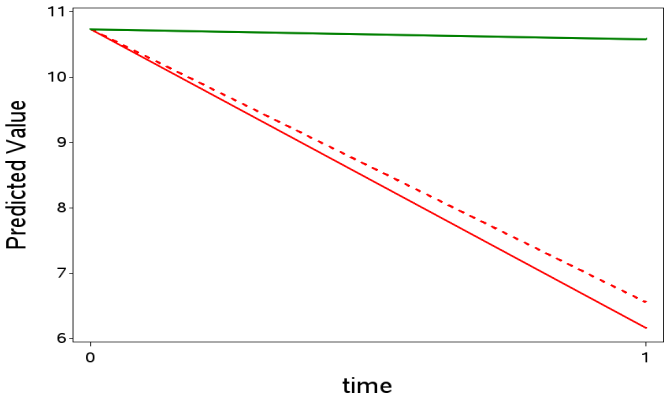
Contrast Results for GEE Analysis

Contrast	DF	Chi-Square	Pr > ChiSq	Type
Antibiotic effect	2	4.56	0.1024	Score
Antibiotic effect?	2	6.99	0.0303	Wald

Interpretations

- There is a significant effect of antibiotics:
  Score test: $4.56 \sim \chi^2(2) \Rightarrow P = 0.10$
  Walds test: $6.99 \sim \chi^2(2) \Rightarrow P = 0.03$
- The *effect* of placebo is estimated to
  $\exp(\hat{\beta}_2) = \exp(-0.0138) = 0.986$, i.e a decrease of 1.4%
- The additional effect of drug A is estimated to
  $\exp(\hat{\beta}_3) = 0.58$, and the total effect to
  $\exp(\hat{\beta}_2 + \hat{\beta}_3) = \exp(-0.5544) = 0.574$,
  i.e a decrease of 42.6%
- The two antibiotics are not significantly different:
  $0.08 \sim \chi^2(1) \Rightarrow P = 0.78$
  (although the estimated effect is a tiny bit larger for drug A)

Predicted means from Population Average model (PA)



Legends:
A —— B ..... P ——

Comments to estimates time profiles

in comparison to the simple averages (p. 39):

- Treatment B starts off at a higher level
- Due to *Regression to the mean*, we therefore expect this group to have the steepest decline
- Since they are close to parallel in the averages (so that B is not steeper than A), this leads us to conclude that B is not as effective as A, and therefore, we see a difference in slope in the predicted means
- Same type of argument concerning P, which would decrease the most *if it was equally effective*

Mixed effects models = Subject Specific models (SS)

New type of model, see p. 45:

Observations: Y_{ij}, covariate vector X_{ij}

We specify Y_{ij} to be Poisson distributed, with

- Mean, $E(Y_{ij}|X_{ij}, b_i) = \mu_{ij}$, where $\log(\mu_{ij}) = X_{ij}^T\beta + Z_{ij}^T b_i$
- Conditional variance, $\text{Var}(Y_{ij}|X_{ij}, b_i) = V(\mu_{ij})$
  Additional overdispersion? Then $\text{Var}(Y_{ij}|X_{ij}, b_i) = \phi V(\mu_{ij})$
- Distribution of random effects, e.g. $b_i \sim N_p(0, G)$,
  where G is the matrix (and software) notation for ω_b^2
- Conditional indepence,
  given the covariates and the random effects

Interpretation of SS

This is a real model (except for possible overdispersion), but

- Inference is *conditional* on random effects and therefore specific to the subject
- The effect of a covariate is interpreted as being for "fixed value of all other covariates", including for fixed *value* of the individual

For models with a log-link, however, the interpretation of covariate effects are still "as usual", except for

- The intercept
- Covariates that also enter as random effects
  e.g. random slope = random effect of time

A very simple example of random slope

A population consisting of two individuals
(e.g. mean number of bacilli:)

Individual	Baseline	Follow up	Ratio
1	12	8	0.667
2	8	7	0.875
Average	10	7.5	0.771

but for the *population*, the ratio is

$$\frac{7.5}{10} = 0.75 \neq 0.771$$

The "average" of individual ratios is not equal to the ratio of the averages

Mixed effects model (SS)

We now assume random intercepts, $b_i \sim N(0, \omega_b^2)$ by specifying a random level for each individual:

```
proc glimmix data=leprosy method=quad(qpoints=50);
     class id;
     model bacilli = time A_effect B_effect /
       dist=poisson link=log solution;
     random intercept / subject=id type=vc g;
contrast 'Antibiotic effect' A_effect 1, B_effect 1;
estimate 'Effect B minus A' A_effect 1 B_effect -1 /exp cl;
estimate "changes for A" time 1 A_effect 1 /exp cl;
estimate "changes for B" time 1 B_effect 1 /exp cl;
estimate 'changes (A,B) vs. P' A_effect 0.5 B_effect 0.5 / exp cl;
output out=ss pred=Pred pred(noblup)=PredPA
              pred(ilink)=PredMu pred(ilink noblup)=PredMuPA;
run;
```

Comments to glimmix code

- `method=quad(qpoints=50)`: maximizes the likelihood function
- `qpoints=50`: the more quadrature points, the better accuracy
- `random`: here we have only one random intercept, so `type=...` is unimportant
- `g`: prints the estimate of ω_b^2 (In `glimmix`, the parameter ω_b^2 is generally denoted G)
- The output data set contains ... see p. 69

Output from glimmix analysis

Estimated G Matrix

Effect	Row	Col1
Intercept	1	0.2814

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error
Intercept	id	0.2814	0.09557

Solutions for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	2.2412	0.1148	29	19.53	<.0001
time	0.003088	0.1235	27	0.03	0.9802
A_effect	-0.6055	0.2036	27	-2.97	0.0061
B_effect	-0.5228	0.1963	27	-2.66	0.0129

Note: Somewhat steeper lines than for PA-model

Output from glimmix analysis, II

Label	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha
Effect B minus A	-0.08271	0.2242	27	-0.37	0.7151	0.05
changes for A	-0.6024	0.1657	27	-3.64	0.0012	0.05
changes for B	-0.5197	0.1567	27	-3.32	0.0026	0.05
changes (A,B) vs. P	-0.5641	0.1656	27	-3.41	0.0021	0.05

Label	Lower	Upper	Exponentiated Estimate
Effect B minus A	-0.5427	0.3773	0.9206
changes for A	-0.9425	-0.2624	0.5475
changes for B	-0.8412	-0.1982	0.5947
changes (A,B) vs. P	-0.9039	-0.2244	0.5688

Label	Exponentiated Lower	Exponentiated Upper
Effect B minus A	0.5812	1.4583
changes for A	0.3897	0.7692
changes for B	0.4312	0.8202
changes (A,B) vs. P	0.4050	0.7990

Label	Num DF	Den DF	F Value	Pr > F
Antibiotic effect	2	27	5.83	0.0079

Note again:
Some differences to PA-analysis, but overall same conclusion

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

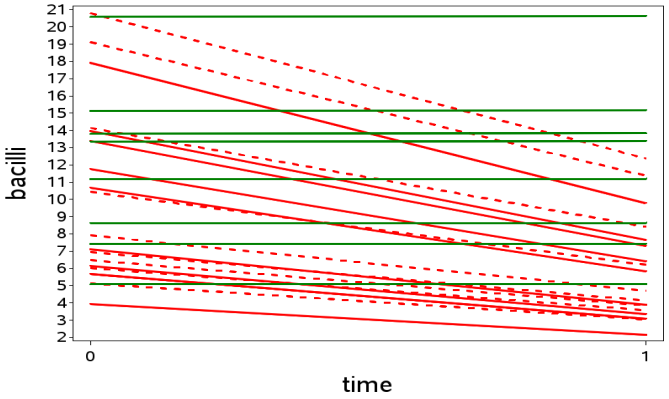
Output dataset from GLIMMIX analysis

The data set `ss`, created p. 65, contains 4 different predicted values:
output out=ss pred=Pred pred(noblup)=PredPA
pred(ilink)=PredMu pred(ilink noblup)=PredMuPA;

- Predicions on log-scale:
  - Pred: Individual predictions (`pred=`)
  - PredPA: Predictions, averaged over population (`pred(noblup)=`)
- Predictions on original scale:
  - PredMu: Individual predictions (`pred(ilink)=`)
  - PredMuPA: Back-transformed average predictions (`pred(ilink noblup)=`)

Individual predicted curves, SS

`pred(ilink)=PredMu`

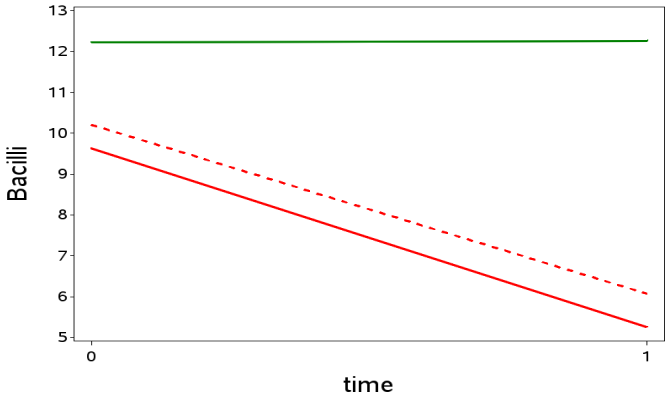


Legends:
A —— B ..... P ——

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

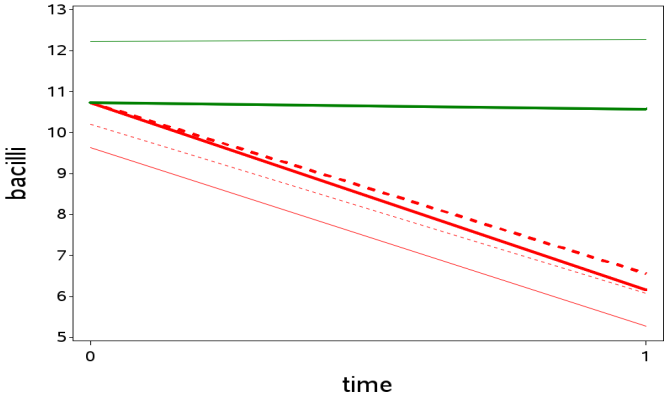
Average individual predictions, SS

Averages from p. 70



Legends:
A —— B ..... P ——

Predicted means from PA and SS



Legends:
A —— B ..... P ——
Thick lines: PA, thin lines: SS

Additional overdispersion in GLIMMIX

Add the line `random _residual_;` to the code from p. 65. Output:

Cov Parm	Subject	Estimate	Standard Error
Intercept	id	0.2522	0.09656
Residual (VC)		1.3292	0.3738

Solutions for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	2.2842	0.1136	29	20.11	<.0001
time	0.01617	0.1402	27	0.12	0.9090
A_effect	-0.6285	0.2293	27	-2.74	0.0107
B_effect	-0.5420	0.2213	27	-2.45	0.0211

Estimates

Label	Estimate	Standard Error	DF	t Value	Pr > |t|
Effect B minus A	-0.08657	0.2540	27	-0.34	0.7359
changes for A	-0.6124	0.1885	27	-3.25	0.0031
changes for B	-0.5258	0.1784	27	-2.95	0.0065

Output, continued

Contrasts

Label	Num DF	Den DF	F Value	Pr > F
Antibiotic effect	2	27	4.97	0.0146

Estimates

Label	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha
Effect B minus A	-0.08657	0.2540	27	-0.34	0.7359	0.05
changes for A	-0.6124	0.1885	27	-3.25	0.0031	0.05
changes for B	-0.5258	0.1784	27	-2.95	0.0065	0.05
changes (A,B) vs. P	-0.5853	0.1862	27	-3.14	0.0040	0.05

Label	Lower	Upper	Exponentiated Estimate	Exp Lower	Exp Upper
Effect B minus A	-0.6078	0.4346	0.9171	0.5446	1.5444
changes for A	-0.9992	-0.2255	0.5421	0.3682	0.7981
changes for B	-0.8919	-0.1597	0.5911	0.4099	0.8524
changes (A,B) vs. P	-0.9672	-0.2033	0.5570	0.3801	0.8160

Overdispersion using Negative Binomial, SS

Change to `dist=negbin` on p.65, and get output

Estimates

Label	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha
Effect B minus A	-0.08199	0.2261	27	-0.36	0.7196	0.05
changes for A	-0.6040	0.1719	27	-3.51	0.0016	0.05
changes for B	-0.5220	0.1694	27	-3.08	0.0047	0.05
changes (A,B) vs. P	-0.5654	0.1703	27	-3.32	0.0026	0.05

Label	Lower	Upper	Exponentiated Estimate
Effect B minus A	-0.5458	0.3818	0.9213
changes for A	-0.9567	-0.2512	0.5466
changes for B	-0.8695	-0.1744	0.5933
changes (A,B) vs. P	-0.9149	-0.2160	0.5681

Label	Exponentiated Lower	Exponentiated Upper
Effect B minus A	0.5794	1.4650
changes for A	0.3841	0.7778
changes for B	0.4191	0.8399
changes (A,B) vs. P	0.4006	0.8058

Contrasts

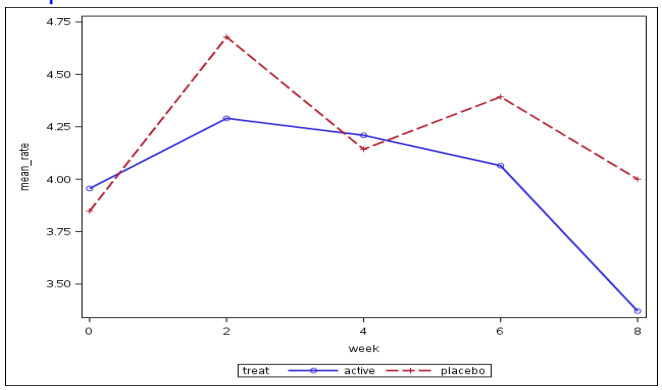
Label	Num DF	Den DF	F Value	Pr > F
Antibiotic effect	2	27	5.55	0.0095

Overview of results for Leprosy

Decrease (A,B) vs. P	Ratio (CI)	P-value
No correlation	0.46 (0.36, 0.60)	< 0.0001
No corr., overdispersion	0.46 (0.29, 0.74)	0.0014
No corr., Negative Binomial	0.46 (0.28, 0.75)	0.0020
PA		
PA, *Poisson*	0.60 (0.41, 0.88)	0.0090
PA, *Negative Binomial*	0.58 (0.39, 0.86)	0.0073
SS, *Poisson*	0.57 (0.41, 0.80)	0.0021
SS, *Poisson, overdispersion*	0.56 (0.38, 0.82)	0.0040
SS, *Negative Binomial*	0.57 (0.40, 0.81)	0.0026

Back to: the Seizure example

Mean value plot



Legends: Progabide Placebo

Not linear...but for illustration....

Purpose of investigation

1. Investigate what happens over time,
   does the number of seizures decrease?
2. Compare the decrease for a patient treated with pragabide to the decrease for a *similar* patient in the placebo group
3. Compare the decrease for a population treated with pragabide to the decrease for a population treated with placebo

Notation:

- T_{ij} denotes the time span
  corresponding to the number of seizures, Y_{ij},
  so T_{ij} is either 2 or 8 weeks

Model building

Model (in principle, not reasonable here)
for the number of seizures:

- Poisson outcome
- Random regression, i.e.
  linear effect of `week`, with individual intercepts and slopes
- Mean value proportional to length of period (8 or 2 weeks)
  `log(8)` and `log(2)` used as `offsets`
  This ensures that we model the ratio $\frac{Y_{ij}}{T_{ij}}$, on `log`-scale, i.e.

$$\log\left(\frac{\mathsf{E}(Y_{ij})}{T_{ij}}\right) = \alpha_{treat} + \beta_{time} + \gamma_{treat*time}$$

or

$$\log(\mathsf{E}(Y_{ij})) = \alpha_{treat} + \beta_{time} + \gamma_{treat*time} + \log(T_{ij})$$

Random regression, SS model in glimmix

Important: The model is not reasonable here (see figure on p. 77), and is *only* showed to hint at possible extensions....

```
proc glimmix data=seizures method=quad(qpoints=50);
     class id trt visit;
     model seizures = weeks trt trt*weeks /
       dist=poisson offset=lweeks link=log solution;
     random intercept weeks / subject=id type=un g;
     estimate 'weekly decline trt=0' weeks 1 weeks*trt 1 0;
     estimate 'weekly decline trt=1' weeks 1 weeks*trt 0 1;
     estimate 'slope, active vs. placebo??' week*trt -1 1 / exp cl;
output out=ss pred=Pred pred(noblup)=PredPA
              pred(ilink)=PredMu pred(ilink noblup)=PredMuPA;
run;
```

Since time (`weeks`) here enter as a random effect, the interpreation of time effects have to be conditional on the specific subject.

Output from random regression

Remember: All the results to follow are only to show possible extensions, model is not reasonable, and results should not be trusted

The GLIMMIX Procedure

Class Level Information

Class	Levels	Values
id	59	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59
trt	2	0 1
visit	5	0 1 2 3 4

Number of Observations Read	295
Number of Observations Used	295

Dimensions

G-side Cov. Parameters	3
Columns in X	6
Columns in Z per Subject	2
Subjects (Blocks in V)	59
Max Obs per Subject	5

Output II

Estimated G Matrix

Effect	Row	Col1	Col2
Intercept	1	0.5277	0.01120
week	2	0.01120	0.005060

Solutions for Fixed Effects

Effect	trt	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		1.1214	0.1356	57	8.27	<.0001
week		-0.05809	0.01697	57	-3.42	0.0012
trt	0	-0.01754	0.1966	177	-0.09	0.9290
trt	1	0	.	.	.	.
week*trt	0	0.04676	0.02351	177	1.99	0.0482
week*trt	1	0	.	.	.	.

The slopes of the two lines are both negative, and the difference in slopes is borderline significant (P=0.048).

Output III

Estimates

Label	Estimate	Standard Error	DF	t Value	Pr > |t|
weekly decline, placebo	-0.01133	0.01692	177	-0.67	0.5040
weekly decline, active	-0.05809	0.01697	177	-3.42	0.0008
slope, active vs. placebo??	-0.04676	0.02351	177	-1.99	0.0482

Label	Alpha	Lower	Upper	Exponentiated Estimate
weekly decline, placebo	0.05	-0.04472	0.02206	0.9887
weekly decline, active	0.05	-0.09157	-0.02461	0.9436
slope, active vs. placebo??	0.05	-0.09314	-0.00037	0.9543

Label	Exponentiated Lower	Exponentiated Upper
weekly decline, placebo	0.9563	1.0223
weekly decline, active	0.9125	0.9757
slope, active vs. placebo??	0.9111	0.9996

Results from random regression, Seizures

- Expected decline for a specific patient (taken from p. 83):
  - Placebo: 0.9887 (0.9563, 1.0223), i.a a weekly decline of 1.1%
  - Active: 0.9436 (0.9125, 0.9757), i.a a weekly decline of 5.6%
- The difference between the slopes, which can be translated to the expected ratio of decline for active vs. placebo patients: This should not be estimated from this model, since it refers to a Population Average comparison.

PA model for random regression

In the sense that the working covariance is specified by a random regression structure, Use

```
proc glimmix data=seizures method=rmpl;
```

Estimates

Label	Estimate	Standard Error	DF	t Value	Pr > |t|
weekly decline, placebo??	0.009604	0.02276	177	0.42	0.6736
weekly decline, active??	-0.00953	0.02172	177	-0.44	0.6614
slope, active vs. placebo	-0.01913	0.03146	177	-0.61	0.5438

Label	Alpha	Lower	Upper	Exponentiated Estimate
weekly decline, placebo??	0.05	-0.03531	0.05452	1.0096
weekly decline, active??	0.05	-0.05239	0.03333	0.9905
slope, active vs. placebo	0.05	-0.08122	0.04295	0.9810

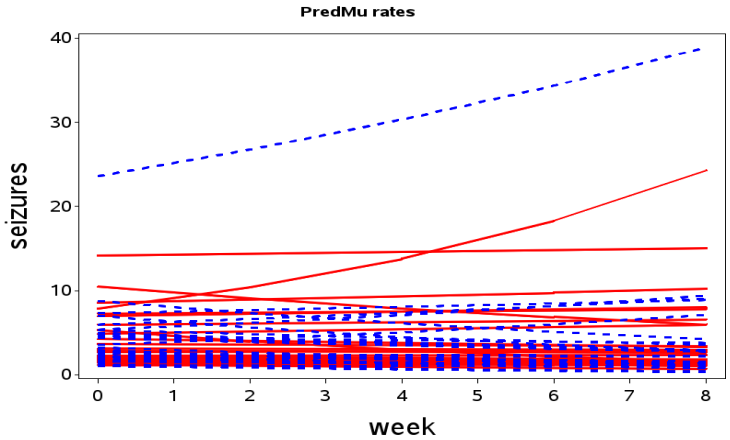
Label	Exponentiated Lower	Exponentiated Upper
weekly decline, placebo??	0.9653	1.0560
weekly decline, active??	0.9490	1.0339
slope, active vs. placebo	0.9220	1.0439

Results from random regression, Seizures, II

- Expected decline for a specific patient:
  This should not be estimated from this model,
  since it refers to a Subject Specific comparison.
- Expected ratio of decline for active vs. placebo patients
  (taken from p. 85):
  0.9810 (0.9220, 1.0439), corresponding to the phrase:
  Patients treated with Pragabide are expected to decrease in seizures with approximately 2% more than if treated with Placebo.
  However, the difference is not significant, and might as well turn out to be an increased rate of approximately 4% compared to Placebo treatment.

Predictions from SS, random regression

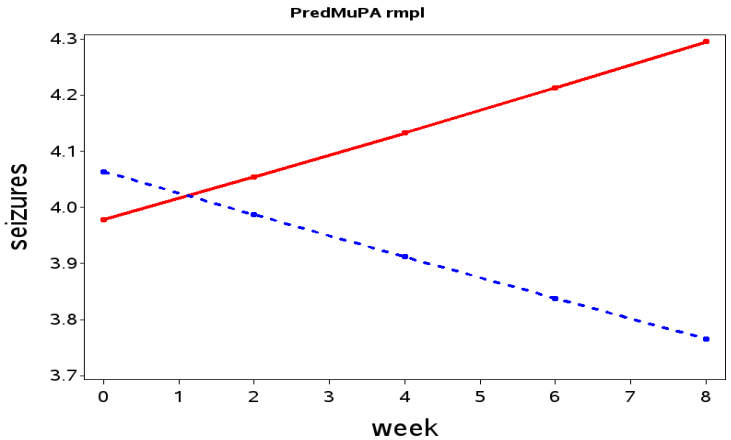
Individual predictions of rates



Legends: Progabide Placebo

Predictions from PA, (random regression)

Population predictions of rates (Note the scale!!)



Legends: Progabide Placebo