Models for longitudinal data
Analysis of repeated measurements, 2016

Julie Lyng Forman & Lene Theil Skovgaard
Department of Biostatistics, University of Copenhagen

Contents
▶ Longitudinal designs
▶ Models for the mean
▶ Models for the covariance
▶ Linear growth model (aka random regression)

Suggested reading: FLW chapters 6, 7, and 8.

Outline
Longitudinal designs
Models for the mean
Covariance pattern models
Analysis of summary statistics
The linear growth model (random regression)
Unbalanced data

Typical set-up for longitudinal measurements
Two or more groups of subjects
▶ E.g. two different treatments, possibly randomized.

Repeated measurements over time for each subject.
▶ calendar time / age / duration of treatment
▶ planned or ad hoc times of measurement.

ATT: statistical results may be biased unless we account for correlation between measurements on the same subject.
Characterizing your longitudinal study

Type of study:
▶ Randomized or observational?

Time schedule:
▶ Fixed or ad hoc observations?

Type of data:
▶ Continuous, binary, count, ordinal, or categorical?

Data structure:
▶ Wide or long format?

Sample size:
▶ How many subjects? How many time points?

Study type

Randomized: One homogeneous population is studied.
▶ Randomization to two (or more) treatment groups
▶ One or more follow-up measurements + usually a measurement at baseline.

Observational: Two (or more) populations are studied.
▶ E.g. men and women or two different groups of patients
▶ A well defined starting point (e.g. diagnosis).

Observation schedules

Fixed time points (balanced design):
▶ Measurements collected at prespecified time points.
▶ Equidistant: 5, 10, 15, 20 minutes, or every month.
▶ Non-equidistant: 5, 10, 20, and 60 minutes.

Ad hoc time points (unbalanced design):
▶ Subjects were seen e.g. when the doctor decided or when the patient felt the need.
▶ Beware of selection bias. What are the time points representative of?

BUT: In practice, designs planned to be balanced often turn out more or less unbalanced....

Case: Calcium supplements

Randomized study: including 112 girls at age 11.

Treatment: calcium supplement or placebo.

Outcome: BMD=bone mineral density, in mg/cm³

Planned follow-up: every 6 months in two years
▶ 5 visits in total including baseline.

Does calcium increase bone gain in adolescent women?
Time points in the calcium study

**visit**: Number 1, 2, 3, 4, or 5.

**time**: Scheduled follow-up 0, 1/2, 1, 1 1/2, 2 years.

**obstime**: Actual time of follow-up (individual).

Analysis Variable : obstime

<table>
<thead>
<tr>
<th>time</th>
<th>NObs</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>112</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.5</td>
<td>112</td>
<td>0.5192530</td>
<td>0.0418429</td>
<td>0.4161533</td>
<td>0.7227926</td>
</tr>
<tr>
<td>1</td>
<td>112</td>
<td>0.9606814</td>
<td>0.0617612</td>
<td>0.8487337</td>
<td>1.1882272</td>
</tr>
<tr>
<td>1.5</td>
<td>112</td>
<td>1.4997597</td>
<td>0.0829828</td>
<td>1.3360712</td>
<td>1.7768652</td>
</tr>
<tr>
<td>2</td>
<td>112</td>
<td>1.9760965</td>
<td>0.0807148</td>
<td>1.7960301</td>
<td>2.2340862</td>
</tr>
</tbody>
</table>

Planned time line (ignore deviations from time schedule)

Repetition: Analysis of response profiles

Comparison of change over \( n \) time points (visits) within \( g \) groups (treatment) of subjects.

- Similar to two-way ANOVA, only with correlated data.
- Use a constrained model (cLMM) for baseline adjustment, if treatment groups are randomized.
- An **unstructured covariance** is assumed.

**Drawbacks:**

- Can only handle balanced designs.
- Not good with many groups or time points, due to having too many model parameters.
- Do not make use of a priori known data patterns, e.g.
  - correlation decreasing with time.
  - monotone growth.

Results

Mean gain in BMD (mg/cm\(^3\)) since baseline with calcium supplement or placebo

<table>
<thead>
<tr>
<th>years</th>
<th>Calcium group</th>
<th>Placebo group</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/2</td>
<td>27 (20;33)</td>
<td>20 (14;27)</td>
<td>6 (0;13)</td>
</tr>
<tr>
<td>1</td>
<td>56 (47;65)</td>
<td>45 (35;54)</td>
<td>12 (2;21)</td>
</tr>
<tr>
<td>1 1/2</td>
<td>83 (72;94)</td>
<td>71 (60;82)</td>
<td>12 (1;23)</td>
</tr>
<tr>
<td>2</td>
<td>106 (94;118)</td>
<td>87 (75;99)</td>
<td>19 (6;31)</td>
</tr>
</tbody>
</table>

- We see a significantly higher gain in BMD with calcium at last follow-up (\( P=0.0032 \))
- Estimated mean at baseline 875 mg/cm\(^3\) (95% CI 863 to 887)
Models for the mean

Changes over time usually appear gradually and often following a distinct pattern.

- We gain power by incorporating this in our models.
- We only need to report few parameters such as growth rates.

Model the mean as a **continuous function of time**:

- Linear
- Exponential (log-linear)
- Piecewise linear (linear spline)
- Nonlinear (cyclic, dose-response curve, etc)
- Nonparametric (loess, smoothing spline, etc)

(many possibilities, not all treated in this course).

Examples of mean curves

Calcium: cLMM estimated means

- Looks as mean BMD increases linearly with time.
Calcium: linear model

To fit a linear model for the mean, use the continuous variable time (scheduled visit in years since baseline).

title1 'Linear trend';

proc mixed data=calcium method=ml plots=all;
class grp (ref='P') girl visit;
model bmd = time grp*time / ddfm=kr solution cl;
repeated visit / type=un subject=girl r rcorr;
run;

▶ Main effect of grp omitted (no difference at baseline).
▶ Use the categorical variable visit to specify the unstructured model for the covariance.
▶ Use the method=ml option for model comparisons.

Calcium: Estimates

Fit Statistics

<table>
<thead>
<tr>
<th>Effect</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2 Log Likelihood</td>
<td>-2430.4</td>
</tr>
<tr>
<td>AIC (Smaller is Better)</td>
<td>-2394.4</td>
</tr>
<tr>
<td>AICC (Smaller is Better)</td>
<td>-2393.0</td>
</tr>
<tr>
<td>BIC (Smaller is Better)</td>
<td>-2346.5</td>
</tr>
</tbody>
</table>

Solution for Fixed Effects

| Effect | Estimate | Error | DF  | t Value | Pr > |t| Alpha | Lower | Upper |
|--------|----------|-------|-----|---------|-------|-------|-------|-------|
| Intercept | 872.0 | 5.712 | 111 | 152.66 | <.0001 | 0.05 | 860.7 | 883.4 |
| time | 44.10 | 2.185 | 98.3 | 20.18 | <.0001 | 0.05 | 39.76 | 48.43 |
| time*grp C | 8.828 | 3.141 | 98.9 | 2.81 | 0.0060 | 0.05 | 2.596 | 15.06 |
| time*grp P | 0.00 | 0.00 | - | - | - | - | - | - |

Type 3 Tests of Fixed Effects

<table>
<thead>
<tr>
<th>Effect</th>
<th>Run</th>
<th>Den</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>time</td>
<td>199.1</td>
<td>969.59</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td>time*grp</td>
<td>98.9</td>
<td>7.90</td>
<td>0.0060</td>
<td></td>
</tr>
</tbody>
</table>

Extra increase in BMD with calcium of 8.8 mg/cm^3 per year, (95%CI: 2.6 to 15.1, P=0.006).

Calcium: Estimated response profiles

Comparison of models for the mean*

Goodness of fit is measured by likelihood.

▶ Better fitting models have large values of likelihood and therefore small values of deviance: $-2 \log \text{Likelihood}$.
▶ ATT: Use the method=ml-option in proc mixed.
▶ Compute the difference in deviances (called $-2 \log Q$) and compare to a $\chi^2$-distribution with df = Δ no. params.
▶ Only nested models can be compared this way.

Example: linear trend vs unrestricted response profiles.

$-2 \log Q = 2444.1 - 2430.4 = 13.7$

$\sim \chi^2(9 - 3) = \chi^2(6) \Rightarrow P = 0.0332$

BUT: Is linear evolution at all plausible?
Technical note on likelihood-types*

When testing a submodel for the mean, the deviances of the compared models must be computed in proc mixed using

- the full (or conventional) likelihood method=ml
- not residual likelihood (method=rem1) which is default.

But don’t forget: Most hypothesis about the mean can be tested using just the default F-tests (optimal choice).

proc mixed data=calcium;
class grp girl visit treat;
model bmd = time grp*time treat*visit/ ddfm=kr solution;
repeated visit / type=un subject=girl;
run;

<table>
<thead>
<tr>
<th>Effect</th>
<th>NumDF</th>
<th>DenDF</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>time</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>time*grp</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>visit*treat</td>
<td>6</td>
<td>126</td>
<td>2.29</td>
<td>0.0392</td>
</tr>
</tbody>
</table>

Residuals for linear trend model

Deviations from linearity are not that pronounced.

The unstructured covariance

So far we have made no assumptions about the covariance.

Advantages

- We make no wrong assumptions about the covariance of our observations. No need to think more about them.
- We gain insight in the actual structure of the covariance by looking at the estimates.

Drawbacks

- We use quite a lot of parameters to describe the covariance structure. Thus our analysis becomes less powerful.
- No good with small data sets; The results may be unstable.
- It can only be used in case of balanced data, i.e. all subjects have to be measured at identical times.
Estimated covariance

Using the cLMM model for the mean (we are sure this is correct).

Estimated R Matrix for girl 101

<table>
<thead>
<tr>
<th>Row</th>
<th>Col1</th>
<th>Col2</th>
<th>Col3</th>
<th>Col4</th>
<th>Col5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3942.77</td>
<td>4185.68</td>
<td>4163.02</td>
<td>4238.18</td>
<td>3946.77</td>
</tr>
<tr>
<td>2</td>
<td>4185.68</td>
<td>4724.53</td>
<td>4709.06</td>
<td>4807.88</td>
<td>4518.94</td>
</tr>
<tr>
<td>3</td>
<td>4163.02</td>
<td>4709.06</td>
<td>4961.24</td>
<td>5042.13</td>
<td>4726.14</td>
</tr>
<tr>
<td>4</td>
<td>4238.18</td>
<td>4807.88</td>
<td>5042.13</td>
<td>5326.11</td>
<td>4980.45</td>
</tr>
<tr>
<td>5</td>
<td>3946.77</td>
<td>4518.94</td>
<td>4726.14</td>
<td>4980.45</td>
<td>4894.13</td>
</tr>
</tbody>
</table>

Estimated R Correlation Matrix for girl 101

<table>
<thead>
<tr>
<th>Row</th>
<th>Col1</th>
<th>Col2</th>
<th>Col3</th>
<th>Col4</th>
<th>Col5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.0000</td>
<td>0.9698</td>
<td>0.9413</td>
<td>0.9249</td>
<td>0.8985</td>
</tr>
<tr>
<td>2</td>
<td>0.9698</td>
<td>1.0000</td>
<td>0.9727</td>
<td>0.9585</td>
<td>0.9398</td>
</tr>
<tr>
<td>3</td>
<td>0.9413</td>
<td>0.9727</td>
<td>1.0000</td>
<td>0.9809</td>
<td>0.9591</td>
</tr>
<tr>
<td>4</td>
<td>0.9249</td>
<td>0.9585</td>
<td>0.9809</td>
<td>1.0000</td>
<td>0.9755</td>
</tr>
<tr>
<td>5</td>
<td>0.8985</td>
<td>0.9398</td>
<td>0.9591</td>
<td>0.9755</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

Models for the covariance

Most often covariance display distinct features.

E.g. decreasing correlation with increasing time span between observations as in the calcium data.

- We gain power by incorporating these features in our model.

Possibilities:

- Unstructured covariance (lecture 1)
- Covariance pattern models (lecture 2)
- Variance components / random effects (lecture 3)

A huge selection is available with proc mixed.

Stationary covariance patterns

Large selection of models for equidistant observations.

Assumption: variances and correlations are stationary:

- The variance is constant over time.
- Correlation depend only on the time-distance between the observations not the specific times of measurements.

Examples: compound symmetry, autoregressive, autoregressive moving average, and the Toeplitz models.

<table>
<thead>
<tr>
<th>proc mixed type</th>
<th>Cov(Y_{ij}, Y_{ik})</th>
<th>parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS</td>
<td>\sigma^2[I{j = k} + \rho \cdot I{j \neq k}]</td>
<td>2</td>
</tr>
<tr>
<td>AR(1)</td>
<td>\sigma^2 \rho^{</td>
<td>k-j</td>
</tr>
<tr>
<td>ARMA(1,1)</td>
<td>\sigma^2[I{j = k} + \gamma \cdot \rho^{</td>
<td>k-j</td>
</tr>
<tr>
<td>TOEP</td>
<td>\sigma^2[I{j = k} + \rho</td>
<td>k-j</td>
</tr>
</tbody>
</table>

Examples of stationary correlations

- Autoregressive
- Gaussian
- Linear
- ARMA
Example: Compound symmetry (type=cs)

Also called exchangeable covariance (more on this in lecture 3)

- The variance \( \sigma^2 \) is the same at all time points.
- The correlation \( \rho \) is the same between all time points.

Calcium data: \( \hat{\sigma}^2 = 4660 \) and \( \hat{\rho} = 0.9496 \).

Heterogeneous covariance patterns

The assumption that the variance does not change with time can be dropped when assuming a heterogeneous covariance pattern.

- No restrictions on the variances \( \sigma_1^2, \ldots, \sigma_n^2 \).
- Correlations are assumed stationary; They depend only on the time-distance between observations.

Examples: the heterogeneous compound symmetry, heterogeneous autoregressive, the heterogeneous Toeplitz, and the antedependence covariance structures.

```
<table>
<thead>
<tr>
<th>proc mixed type</th>
<th>Cov((Y_{ij}, Y_{ik}))</th>
<th>parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSH</td>
<td>(\sigma_j \sigma_k [I(j = k)] + \rho \cdot I[j \neq k])</td>
<td>(n + 1)</td>
</tr>
<tr>
<td>ARH(1)</td>
<td>(\sigma_j \sigma_k [I(j = k)] + \rho</td>
<td>k-j</td>
</tr>
<tr>
<td>TOEPH</td>
<td>(\sigma_j \sigma_k [I(j = k)] + \rho</td>
<td>k-j</td>
</tr>
<tr>
<td>ANTE(1)</td>
<td>(\sigma_j \sigma_k \prod_{l=j}^{k-1} \rho_l)</td>
<td>(2n - 1)</td>
</tr>
</tbody>
</table>
```

Example: Autoregressive pattern (type=ar(1))

The so-called autoregressive covariance structure has

- Constant variance \( \sigma^2 \) over time.
- Correlation decreasing exponentially with the distance between the observations, \( \text{Cor}(Y_{ij}, Y_{ik}) = \rho^{\mid k-j \mid} \)

Calcium data: \( \hat{\sigma}^2 = 4401 \) and \( \hat{\rho} = 0.9708 \).

Example: Heterogeneous AR (type=arh(1))

Correlations are similar to AR(1), but variances differ.

- Correlation decreasing exponentially with the distance between the observations, \( \text{Cor}(Y_{ij}, Y_{ik}) = \rho^{\mid k-j \mid} \)
- Variances \( \sigma_1^2, \ldots, \sigma_n^2 \) are specific to each time point.

Calcium data: \( \hat{\sigma}_1^2 = 4013, \ldots, \hat{\sigma}_5^2 = 4758, \) and \( \hat{\rho} = 0.9744 \).
Comparison of covariance structures

<table>
<thead>
<tr>
<th>Model</th>
<th>$-2 \log L$</th>
<th>par.</th>
<th>$-2 \log Q$</th>
<th>$\Delta df$</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>UN</td>
<td>-2352.6</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARH(1)</td>
<td>-2343.4</td>
<td>6</td>
<td>9.2</td>
<td>9</td>
<td>0.42</td>
</tr>
<tr>
<td>AR(1)</td>
<td>-2324.8</td>
<td>2</td>
<td>27.8</td>
<td>13</td>
<td>0.0096</td>
</tr>
<tr>
<td>CS</td>
<td>-2195.1</td>
<td>2</td>
<td>157.5</td>
<td>13</td>
<td>$&lt; 0.0001$</td>
</tr>
</tbody>
</table>

* For comparison of covariance patterns use either of the two likelihood-types (ml or reml), just don’t compare one to the other.

Predicted means

The estimated response profiles are almost identical for all our choices of covariance patterns.

Tests of treatment effect

**BUT:** Confidence intervals and tests depend on the covariance.

Tests of treatment effect at last follow-up

<table>
<thead>
<tr>
<th>Covariance pattern</th>
<th>Estimate (95% CI)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compound symmetry</td>
<td>19.6 (11.0-28.3)</td>
<td>$&lt; 0.0001$</td>
</tr>
<tr>
<td>Autoregressive</td>
<td>20.1 (7.8-32.4)</td>
<td>0.0014</td>
</tr>
<tr>
<td>Heterogeneous autoregr.</td>
<td>19.4 (7.3-31.5)</td>
<td>0.0018</td>
</tr>
<tr>
<td>Unstructured</td>
<td>19.0 (6.5-31.4)</td>
<td>0.034</td>
</tr>
</tbody>
</table>

Modeling strategy

As suggested by Fitzmaurice et al. (2011):

1. Put up a plausible (e.g. saturated) model for the mean
2. Fit the data so far ignoring correlation (GLM).
3. Check the residuals for assessing the adequacy of the model for the mean and in order to get an impression of the error covariance.
4. Pick a reasonable model for the covariance (if possible test against the unstructured model).
5. Re-check the model fit.
6. Do the analysis.

**Note:** Lecture 4 on model diagnostics among others.
Alternative: Robust standard errors

Goodness of fit test for the covariance might approve an incorrect model due to lack of power.

- If data is balanced and complete estimates are unbiased
- but the standard errors may be biased.

Instead one could use the robust sandwich covariance estimator:

* proc mixed empirical data=calcium;

However: The sandwich covariance estimator . . .

- is useless in small samples because it is anti-conservative.
- needs additional weighting to get unbiased estimates when there are missing data.

(More about sandwich covariance estimators in lectures 5 and 6.)

Outline

- Longitudinal designs
- Models for the mean
- Covariance pattern models
- Analysis of summary statistics
- The linear growth model (random regression)
- Unbalanced data

Many time points and few subjects

In this situation choosing a reliable model for the covariance is just about impossible.

- Unstructured covariance has too many parameters.
- Compound symmetry underestimates correlation between observations close in time and overestimates correlation between observations far apart in time.
- We have no crystal ball for choosing a simple yet correct covariance pattern.
- Robust standard errors are anti-conservative.

So what can we do?

Reduction to independent data

By analyzing carefully chosen characteristics for each individual we can resolve to simple analyses which have no repeated measurements issues.

- Maybe not optimal, but feasible – and easy!

Examples of useful summary statistics:

- The changes from baseline to endpoint
- The slopes for individual time effects
- The area under the curve (AUC)
- The time to peak or peak value

Note: Comparing measurements for each time point in turn is not recommended without adjustment for multiple testing.
Case: Calcium supplements

The overall time course looks reasonably linear, but maybe the girls have individual growth rates?

Individual regression

Fit an ordinary linear regression for each girl:

\[ Y_{ij} = a_i + b_i t_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2) \]

Are the slopes of the Calcium group systematically bigger?

Analysis of summary statistics

Comparison of individual intercepts and slopes:

- Use the two-sample t-test for independent data.

<table>
<thead>
<tr>
<th>Group</th>
<th>Level at baseline (mg/cm(^3))</th>
<th>Slope (mg/cm(^3) per year)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>869.7 (851.5;888.0)</td>
<td>41.1 (36.1;46.2)</td>
</tr>
<tr>
<td>C</td>
<td>881.7 (865.0;898.5)</td>
<td>49.7 (43.8;55.6)</td>
</tr>
<tr>
<td>Dif</td>
<td>12.0 (-12.5;36.5)</td>
<td>8.54 (0.9;16.2)</td>
</tr>
<tr>
<td>P-value</td>
<td>0.33</td>
<td>0.029</td>
</tr>
</tbody>
</table>

Limitations of summary statistics

No baseline adjustment, hence less power.

Some of the girls in the calcium study dropped out:

- We get less accurate slope-estimates from girls with few observations.
- No slope at all if drop out was right after baseline.
- And maybe those with low BMD are more likely to drop out; Parents think the girl needs supplement and won’t risk placebo.

Can we make better use of the full data?
Outline

- Longitudinal designs
- Models for the mean
- Covariance pattern models
- Analysis of summary statistics
- The linear growth model (random regression)
- Unbalanced data

Random regression

We let each girl have her own level $a_i$ and her own slope $b_i$

We **assume** these individual 'parameters' ($a_i$ and $b_i$) follow a bivariate normal distribution in the population

$$\begin{pmatrix} a_i \\ b_i \end{pmatrix} \sim N_2 \left( \begin{pmatrix} \alpha_{g(i)} \\ \beta_{g(i)} \end{pmatrix}, \begin{pmatrix} \tau_a^2 & \omega_{ab} \\ \omega_{ab} & \tau_b^2 \end{pmatrix} \right)$$

The covariance is the so-called **G-matrix**.

- $G$ describes the **population variance** of the regression parameters, the **inter-individual variation**.
- Note the correlation $\rho_{ab} = \frac{\omega_{ab}}{\tau_a \cdot \tau_b}$.

Subjects with lower baseline might tend to also have smaller (or larger) growth rates.

**PROC MIXED: random regression**

```r
proc mixed data=calcium plots=all;
class grp girl;
model bmd = time grp*time / ddfm=kr solution cl;
random intercept time / type=un subject=girl g v vcorr;
run;
```

Individual intercepts and slopes are so-called **random effects**. They must be specified in a **random-statement**.

- **Note:** The intercept refers to $time=0$ (baseline).
- **Note:** Here `type=un` refers to an unstructured specification of the G-matrix. If it is omitted, we may experience convergence problems and sometimes totally incomprehensible results.
- **Option g** prints the estimated G-matrix.

(For more about random effects in lecture 3).

**Output from random regression**

<table>
<thead>
<tr>
<th>Estimated G Matrix</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Row</td>
<td>Effect</td>
</tr>
<tr>
<td>-----</td>
<td>--------</td>
</tr>
<tr>
<td>1</td>
<td>Intercept</td>
</tr>
<tr>
<td>2</td>
<td>time</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Covariance Parameter Estimates</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cov Parm</td>
<td>Subject</td>
</tr>
<tr>
<td>----------</td>
<td>---------</td>
</tr>
<tr>
<td>UN(1,1)</td>
<td>girl</td>
</tr>
<tr>
<td>UN(2,1)</td>
<td>girl</td>
</tr>
<tr>
<td>UN(2,2)</td>
<td>girl</td>
</tr>
<tr>
<td>Residual</td>
<td></td>
</tr>
</tbody>
</table>

**Fit Statistics**

- $-2$ Res Log Likelihood: $-2350.4$
- AIC (smaller is better): $-2342.4$
- AICC (smaller is better): $-2342.3$
- BIC (smaller is better): $-2331.5$

**Note:** $\rho_{ab} = \frac{103}{\sqrt{4155 \cdot 191}} = 0.12$. 
Output from random regression

### Solution for Fixed Effects

| Effect      | grp | Estimate | Standard Error | DF  | t Value | Pr > |t| | Alpha | Lower  | Upper |
|-------------|-----|----------|----------------|-----|---------|------|---|-------|--------|--------|
| Intercept   |     | 875.2    | 6.149          | 111 | 142.32  | <.0001 | 0.05 | 863.0 | 887.4  |
| time        |     | 44.90    | 2.208          | 96  | 20.33   | <.0001 | 0.05 | 40.51 | 48.28  |
| time*grp C  |     | 8.859    | 3.174          | 96.5| 2.79    | 0.0063 | 0.05 | 2.169 | 15.16  |
| time*grp P  |     | 0        |                |     |         |       |     |       |        |

### Type 3 Tests of Fixed Effects

<table>
<thead>
<tr>
<th>Effect</th>
<th>Num</th>
<th>Den</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>time</td>
<td>1</td>
<td>96.4</td>
<td>982.47</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>time*grp C</td>
<td>1</td>
<td>96.5</td>
<td>7.79</td>
<td>0.0063</td>
</tr>
</tbody>
</table>

We find an extra increase in BMD of 8.9 mg/cm³ per year, with calcium supplement, (95% CI: 2.6 to 15.2, P=0.0063).

Implied covariance

The random regression implies the covariance pattern:

$$\text{Cov}(Y_{ij}, Y_{ik}) = \tau_a^2 + (t_j + t_k)\omega_{ab} + t_j t_k \tau_b^2$$

as a function of the G-matrix parameters and the time points.

### Does this fit the data well?

Random regression vs UN (linear trend for the mean)

$$-2 \log Q = 2400.9 - 2350.4 = 50.5 \sim \chi^2(15-3) = \chi^2(12) \Rightarrow P = 0.00001$$

No, apparently not.

Implied covariance

➤ Options v and vcorr prints the covariance and correlation.

#### Estimated V Matrix for girl 101

<table>
<thead>
<tr>
<th>Row</th>
<th>Col1</th>
<th>Col2</th>
<th>Col3</th>
<th>Col4</th>
<th>Col5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4280</td>
<td>4207</td>
<td>4258</td>
<td>4309</td>
<td>4360</td>
</tr>
<tr>
<td>2</td>
<td>4207</td>
<td>4430</td>
<td>4405</td>
<td>4503</td>
<td>4602</td>
</tr>
<tr>
<td>3</td>
<td>4258</td>
<td>4405</td>
<td>4676</td>
<td>4698</td>
<td>4844</td>
</tr>
<tr>
<td>4</td>
<td>4309</td>
<td>4503</td>
<td>4698</td>
<td>5017</td>
<td>5086</td>
</tr>
<tr>
<td>5</td>
<td>4360</td>
<td>4602</td>
<td>4844</td>
<td>5086</td>
<td>5453</td>
</tr>
</tbody>
</table>

#### Estimated V Correlation Matrix for girl 101

<table>
<thead>
<tr>
<th>Row</th>
<th>Col1</th>
<th>Col2</th>
<th>Col3</th>
<th>Col4</th>
<th>Col5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.0000</td>
<td>0.9660</td>
<td>0.9518</td>
<td>0.9299</td>
<td>0.9026</td>
</tr>
<tr>
<td>2</td>
<td>0.9660</td>
<td>1.0000</td>
<td>0.9677</td>
<td>0.9552</td>
<td>0.9364</td>
</tr>
<tr>
<td>3</td>
<td>0.9518</td>
<td>0.9677</td>
<td>1.0000</td>
<td>0.9699</td>
<td>0.9594</td>
</tr>
<tr>
<td>4</td>
<td>0.9299</td>
<td>0.9552</td>
<td>0.9699</td>
<td>1.0000</td>
<td>0.9724</td>
</tr>
<tr>
<td>5</td>
<td>0.9026</td>
<td>0.9364</td>
<td>0.9594</td>
<td>0.9724</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

ATT: These are specific to the individual time points.

Outline

- Longitudinal designs
- Models for the mean
- Covariance pattern models
- Analysis of summary statistics
- The linear growth model (random regression)
- Unbalanced data
Nonequidistant time points

In the calcium study the girls are seen approximately twice a year.

- Perhaps we get better estimates of the slopes when replacing planned time of visit with the actual individual times?
- But we loose the option of an unstructured covariance.

Other covariance patterns can still be fitted, e.g.
- the autoregressive pattern,
- the random regression model,
- the compound symmetry pattern.

Non-equidistant observations

Only a limited number of covariance patterns are available in case time points are individual or non-equidistant. All are stationary:

- The variance is constant over time.
- The correlation depend only on the time-distance between the observations.

The ctime-variable must be a numerical variable in SAS

<table>
<thead>
<tr>
<th>proc mixed type</th>
<th>Cov(Y_{ij}, Y_{ik})</th>
<th>parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS</td>
<td>\sigma^2 I{j = k} + \rho \cdot I{j \neq k}</td>
<td>2</td>
</tr>
<tr>
<td>SP(POW)(ctime)</td>
<td>\sigma^2 \rho</td>
<td>t_j - t_k</td>
</tr>
<tr>
<td>SP(GAU)(ctime)</td>
<td>\sigma^2 e^{-</td>
<td>t_j - t_k</td>
</tr>
<tr>
<td>SP(LIN)(ctime)</td>
<td>\sigma^2(1 - \rho</td>
<td>t_k - t_j</td>
</tr>
</tbody>
</table>

Random regression in observed time

Syntax is exactly the same as with scheduled time points, only the name of the time-variable changes.

```sas
title1 'random regression (actual time)';
proc mixed data=calcium plots=all;
class grp girl;
model bmd = obstime grp*obstime / ddfm=kr solution cl;
random intercept obstime / type=un subject=girl g;
run;
```
Random regression, using actual duration

**Estimated G Matrix**

<table>
<thead>
<tr>
<th>Row</th>
<th>Effect</th>
<th>girl</th>
<th>Col1</th>
<th>Col2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Intercept</td>
<td>101</td>
<td>4174</td>
<td>103</td>
</tr>
<tr>
<td>2</td>
<td>obstime</td>
<td>101</td>
<td>103</td>
<td>179</td>
</tr>
</tbody>
</table>

**Solution for Fixed Effects**

<table>
<thead>
<tr>
<th>Effect</th>
<th>grp</th>
<th>Estimate</th>
<th>Error</th>
<th>DF</th>
<th>t Value</th>
<th>Pr &gt;</th>
<th>Alpha</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td></td>
<td>875.1</td>
<td>6.163</td>
<td>111</td>
<td>141.99</td>
<td>&lt;.0001</td>
<td>0.05</td>
<td>862.9</td>
<td>887.3</td>
</tr>
<tr>
<td>obstime</td>
<td></td>
<td>45.38</td>
<td>2.162</td>
<td>96.3</td>
<td>20.99</td>
<td>&lt;.0001</td>
<td>0.05</td>
<td>41.09</td>
<td>49.68</td>
</tr>
<tr>
<td>obstime*grp</td>
<td>C</td>
<td>8.758</td>
<td>3.104</td>
<td>96.9</td>
<td>2.82</td>
<td>0.0058</td>
<td>0.05</td>
<td>2.596</td>
<td>14.92</td>
</tr>
</tbody>
</table>

**Type 3 Tests of Fixed Effects**

<table>
<thead>
<tr>
<th>Effect</th>
<th>Num</th>
<th>Den</th>
<th>F Value</th>
<th>Pr &gt;</th>
<th>F Value</th>
<th>Pr &gt;</th>
<th>Alpha</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>obstime*grp</td>
<td>1</td>
<td>96.6</td>
<td>1046.11</td>
<td>&lt;.0001</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>obstime*grp</td>
<td>1</td>
<td>96.9</td>
<td>7.96</td>
<td>0.0058</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

We find an extra increase in BMD of 8.8 mg/cm³ per year with calcium supplement, 95% CI: 2.6 to 14.9, P=0.0058.

Tests of treatment effect

**Comparison of estimates for different covariance structures:**

<table>
<thead>
<tr>
<th>Covariance pattern</th>
<th>Estimate (95% CI)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compound symmetry</td>
<td>9.2 (5.3;13.8)</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Autoregressive</td>
<td>10.0 (3.8;16.2)</td>
<td>0.0017</td>
</tr>
<tr>
<td>Random regression</td>
<td>8.8 (2.6;14.9)</td>
<td>0.0058</td>
</tr>
</tbody>
</table>

▶ Estimates and tests depend on the covariance!

Scheduled vs observed times

**Estimated slopes from the two random regressions:**

<table>
<thead>
<tr>
<th>Group</th>
<th>Scheduled time</th>
<th>Actual time</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>44.9 (40.5;49.3)</td>
<td>45.4 (41.1;49.7)</td>
</tr>
<tr>
<td>C</td>
<td>53.7 (49.3;58.2)</td>
<td>54.1 (49.8;58.5)</td>
</tr>
</tbody>
</table>

| Difference | 8.9 (2.6;15.2) | 8.8 (2.6;14.9) |
| P-value    | 0.0063         | 0.0058         |

▶ Slightly steeper mean slopes in both groups with actual times.
▶ Slightly smaller standard errors with actual times.

The latter would have been more pronounced if actual times of visit deviated more from the scheduled times.

Concluding remarks

Results depend on choice of covariance pattern

▶ Obvious bias for unrealistic models (independence, CS).
▶ More similar results for the more complex models.

Not much impact of using exact times of measurement instead of planned times (visit) - WHY?

▶ Gain in modeling flexibility by rounding the times.
▶ There are sophisticated statistical arguments implying that rounding to the nearest scheduled time do not cause bias.
▶ But it increases variance.

Choosing an appropriate model is a compromise between practical feasibility, realistic model assumptions, and interpretable results.