Analyzing Longitudinal Quality of Life Outcome Data

Stephen Walters
Professor of Medical Statistics and Clinical Trials
School of Health and Related Research
University of Sheffield, United Kingdom
• To download full colour versions of the powerpoint slides go to....

• [http://www.shef.ac.uk/scharr/sections/dts/Statistics/staff/walters](http://www.shef.ac.uk/scharr/sections/dts/Statistics/staff/walters)
Aims

• The aim of the workshop is to be a practical guide to the analysis of longitudinal QoL data.

• At the end of the workshop participants should know about:
  – Summarising, tabulating and graphically displaying repeated QoL assessments;
  – Response feature analysis – the use of summary measures;
  – Modelling longitudinal QoL data using marginal and random-effect general linear models (GLMs).
  – How to analyse and present longitudinal data from a two group comparative study.
Introduction

• Some studies using QoL outcomes have repeated assessments over time and are longitudinal in nature.
• In a RCT and other longitudinal studies there may be a baseline QoL assessment and several follow-up assessments over time.
• This session will describe how the QoL data from such studies can be summarised, tabulated and graphically displayed.
• These repeated QoL measurements, on the same individual subject, are likely to be related or correlated.
• This means that the usual statistical methods for analysing such data which assume independent outcomes may not be appropriate.
• This session will show how repeated QoL measures for each subject can be reduced to a single summary measure for statistical analysis and how standard statistical methods of analysis can then be used.
• Finally, the session will describe a more complex modelling approach, based on an extension of the linear regression model which allows for the fact that successive QoL assessments by a particular patient are likely to be correlated.
Three broad approaches to analysing repeated QoL assessments

• With one QoL observation on each subject or experimental unit, we are confined to modelling the population average QoL, called the *marginal* mean response; there is no other choice.

• However, with repeated QoL measurements, there are several different approaches that can be adopted. Three broad approaches are (Everitt, 2002):

  1. Time by time analysis;
  2. Response feature analysis – the use of summary measures;
  3. Modelling of longitudinal data.
Summarising repeated QoL assessments

• An important initial step, prior to analysing the repeated QoL assessments is to tabulate the data and/or graphically display it.

• This will give us an idea of how the QoL outcomes change over time.
Example Dataset

• To illustrate some of the methods we will use data from an RCT.
• OBJECTIVE: To determine whether a short course of traditional acupuncture improves longer term outcomes for patients with persistent non-specific low back pain in primary care.
• DESIGN: Pragmatic, open, RCT with 24 months follow-up.
• SETTING: Three private acupuncture clinics and 18 general practices in York, England
• SUBJECTS: 241 adults aged 18-65 with non-specific low back pain of 4-52 weeks’ duration allocated at random to Acupuncture group (160) or Usual Care (81) group.
• INTERVENTIONS: 10 individualised acupuncture treatments from one of six qualified acupuncturists or usual care only.
Main outcome measures

- Patient quality of life (QoL) as measured by the SF-36 at baseline (0), 3, 12 and 24 months.
- The primary outcome was SF-36 bodily pain, measured at 12 months.
- Other outcomes included reported use of analgesics, scores on the Oswestry pain disability index, safety, and patient satisfaction.
Flow of patients through trial

Eligible patients identified by general practitioners (n=289)

- Refused (n=28): On speaking to general practitioner (n=13) On speaking to researcher (n=15)
- Exclusions (n=20): Back pain resolved (n=12) Outside age range (n=5) Back pain > 12 months (n=1) Pending litigation (n=1) Currently receiving acupuncture (n=1)

Randomised (n=241)

- Allocated to offer of acupuncture (n=160) Chose acupuncture (n=160) Did not receive acupuncture care: patient withdrew due to intercurrent illness (n=1) Received acupuncture appointment (n=159) Did not receive acupuncture treatment (n=9) Withdrew early from acupuncture treatment (n=16)
- Allocated to usual management (n=81) Received usual management (n=81) Patient withdrew consent to participate immediately after randomisation (n=1)

Follow-up

- Lost to follow-up: Non-responders at 3 months (n=13) Non-responders at 12 months (n=12) General practitioner notes found (12/12) Non-responders at 24 months (n=36) General practitioner notes found (24/36)
- Lost to follow-up: Non-responders at 3 months (n=9) Non-responders at 12 months (n=12) General practitioner notes found (12/12) Non-responders at 24 months (n=21) General practitioner notes found (19/21)

Analysis

- Outcomes analysed: 3 months (n=146, 92%) 12 months (n=147, 93%) 24 months (n=123, 77%) All included in primary and secondary analyses
- Outcomes analysed: 3 months (n=71, 92%) 12 months (n=69, 85%) 24 months (n=59, 73%) All included in primary and secondary analyses

Thomas, K J et al. BMJ 2006;333:623
Table of results: what do you think?
What is good (and bad) about the table below?

**Table:** Mean SF-36 Pain scores over time by treatment group for 239 patients (data from Thomas et al 2006)

<table>
<thead>
<tr>
<th>SF-36 Pain Outcome Time (months)</th>
<th>Treatment group</th>
<th>Independent Samples Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Usual care Mean</td>
<td>Acupuncture Mean</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>SE</td>
</tr>
<tr>
<td>0</td>
<td>30.417</td>
<td>2.0132</td>
</tr>
<tr>
<td>3</td>
<td>55.399</td>
<td>3.0193</td>
</tr>
<tr>
<td>12</td>
<td>58.333</td>
<td>2.6885</td>
</tr>
<tr>
<td>24</td>
<td>59.51</td>
<td>3.0481</td>
</tr>
</tbody>
</table>
Results: what do you think?

- Hard to identify structure
- Title uninformative
- Not clear what numbers mean: is a high QoL score good or bad?
- Spurious numerical precision and decimal places add to clutter.
- Hard to identify size and direction of effect – no mean difference or confidence intervals.
- Not clear what abbreviations e.g. SE and df mean
- No sample size.
- Repeated/multiple hypothesis/significance testing.
Table of results: now what do you think?
Tabulating repeated QoL assessments (I)

Table: Mean SF-36 Pain scores over time by treatment group with all valid patients at each time-point (data from Thomas et al 2006)

<table>
<thead>
<tr>
<th>SF-36 Pain Outcome‡</th>
<th>Treatment group</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Time (months)</td>
<td>Usual care n</td>
<td>Mean</td>
<td>SD</td>
<td>Acupuncture n</td>
<td>Mean</td>
<td>SD</td>
<td>Mean Difference†</td>
<td>95% CI Lower</td>
</tr>
<tr>
<td>0</td>
<td>80</td>
<td>30.4</td>
<td>(18.0)</td>
<td>159</td>
<td>30.8</td>
<td>(16.2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>71</td>
<td>55.4</td>
<td>(25.4)</td>
<td>146</td>
<td>60.9</td>
<td>(23.0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>68</td>
<td>58.3</td>
<td>(22.2)</td>
<td>147</td>
<td>64.0</td>
<td>(25.6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>59</td>
<td>59.5</td>
<td>(23.4)</td>
<td>123</td>
<td>67.8</td>
<td>(24.1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean follow-up SF-36 pain score</td>
<td>76</td>
<td>57.2</td>
<td>(19.8)</td>
<td>153</td>
<td>63.4</td>
<td>(20.9)</td>
<td>6.3</td>
<td>0.6</td>
</tr>
<tr>
<td>Pain Area under the curve (AUC)</td>
<td>55</td>
<td>112.7</td>
<td>(36.7)</td>
<td>118</td>
<td>125.2</td>
<td>(39.4)</td>
<td>12.6</td>
<td>0.1</td>
</tr>
</tbody>
</table>

‡The SF-36 pain dimension is scored on a 0 (poor) to 100 (good health) scale.
*P-value from two independent samples t-test.
†A positive mean difference indicates the Acupuncture group has the better QoL.
CI Confidence Interval AUC Area Under the Curve.
Tabulating repeated QoL assessments (II)

**Table:** Mean SF-36 Pain scores over time by treatment group with patients who completed all four QoL assessments (data from Thomas *et al* 2006)

<table>
<thead>
<tr>
<th>SF-36 Pain Outcome,‡</th>
<th>Treatment group</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Usual care</td>
<td>Acupuncture</td>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>95% CI</td>
</tr>
<tr>
<td>Time (months)</td>
<td>n n Mean SD N</td>
<td>n Mean SD N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lower</td>
</tr>
<tr>
<td>0</td>
<td>55 29.9 (18.5) 118 31.5 (16.6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>55 57.4 (26.9) 118 62.3 (22.4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>55 57.8 (21.8) 118 64.1 (25.4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>55 59.4 (23.7) 118 68.1 (23.8)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean follow-up</td>
<td>55 58.2 (19.5) 118 64.8 (20.1)</td>
<td>6.7</td>
<td>0.3</td>
<td>13.1</td>
<td>0.042</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SF-36 pain score</td>
<td>Pain Area under the curve (AUC)</td>
<td>12.6</td>
<td>0.1</td>
<td>25.0</td>
<td>0.048</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

‡The SF-36 Pain dimension is scored on a 0 (poor) to 100 (good health) scale.
*P-value from two independent samples t-test.
†A positive mean difference indicates the Acupuncture group has the better QoL.
Recommendations when presenting data and results in tables

- The amount of information should be maximised for the minimum amount of ink.
- Numerical precision should be consistent throughout a paper or presentation, as far as possible.
- Avoid spurious accuracy. Numbers should be rounded to two effective digits.
- Quantitative data should be summarised using either the mean and standard deviation (for symmetrically distributed data) or the median and interquartile range or range (for skewed data). The number of observations on which these summary measures are based should be included.
- Categorical data should be summarised as frequencies and percentages. As with quantitative data, the number of observations should be included.
- Each table should have a title explaining what is being displayed and columns and rows should be clearly labelled.
- Gridlines in tables should be kept to a minimum.
- Where variables have no natural ordering, rows and columns should be ordered by size.
Guidelines for presenting numerical information

- Test statistics such as values of $t$ or $c^2$ and correlation coefficients should be given to no more than two decimal places.
- P-values should be given to one or two significant figures, even for non-significant results as these may conceal important information.
- Confidence intervals are better presented as 12.4 to 52.9 because the format 12.4-52.9 is confusing when one or both numbers are negative.
Graphically displaying repeated QoL assessments

Figure: Profile of individual SF-36 Pain scores over time for the first 20 participants
Graphically displaying repeated QoL assessments

Figure: Profile of individual SF-36 Pain scores over time (n=239)
Graphically displaying repeated QoL assessments

Figure: Profile of Mean SF-36 Pain scores over time by treatment group all available data (data from Thomas et al 2006)

<table>
<thead>
<tr>
<th>Group</th>
<th>0</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>25</th>
</tr>
</thead>
<tbody>
<tr>
<td>Usual care</td>
<td>80</td>
<td>71</td>
<td>68</td>
<td>59</td>
<td></td>
</tr>
<tr>
<td>Acupuncture</td>
<td>159</td>
<td>146</td>
<td>147</td>
<td>123</td>
<td></td>
</tr>
</tbody>
</table>
Figure: Profile of Mean SF-36 Pain scores over time by treatment group complete case analysis (data from Thomas et al 2006)
Further examples of patterns of QoL scores over time
Mean EORTC QLC-30 Global Health scores over time by treatment group (AIM-High) in malignant melanoma patients

- Interferon-alpha (n=39)
- Control (n=37)
Mean EQ-5D scores over time by hip fracture status in elderly women (aged 75 or more)
Mean Health Assessment Questionnaire (HAQ) Scores scores over time by gender in patients with RA

![Graph showing mean HAQ scores over time by gender in patients with RA.](image)
Mean SF-6D scores over time in new mothers by treatment group (PONDER RCT)

<table>
<thead>
<tr>
<th>Time (years)</th>
<th>Control (n=246)</th>
<th>Intervention (n=533)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Mean Chronic Respiratory Disease Questionnaire (CRQ) scores over time by treatment group in patients with COPD

Mean CRQ Mastery score

Time (years)

- Community rehabilitation (n=49)
- Hospital rehabilitation (n=63)
• Each graph should have a title explaining what is being displayed.
• Axes should be clearly labelled.
• Gridlines should be kept to a minimum (and drawn in a faded shade).
• With a small sample size (<20) plot the individual QoL scores over time.
• For larger sample sizes (>20) summarise the data with the mean or median QoL score and plot these over time.
• Preferable to join the summary points by a dotted line if different number of subjects at each time point.
• The number of observations (at each time point) should be included.
Time by time analysis

• A series of two independent samples $t$-tests (or the non-parametric equivalent), could be used to test for differences, in QoL, between the two groups at each time point.
• The procedure is straightforward but has a number of serious flaws and weaknesses (Everitt, 2001).

1. The QoL measurements in a subject from one time point to the next are not independent, so interpretation of the results is difficult.
2. The large number of hypothesis tests carried out implies that we are likely to obtain significant results purely by chance.
3. We lose information about the within-subject changes in QoL over time.

• Consequently, it will not be pursued further here.
Response feature analysis

(1)

• Here the repeated QoL measures for each participant are transformed into a single number considered to capture some important aspect of the participant’s response (Mathew’s et al 1990).

• A simple and often effective strategy (Diggle et al 2002) is to:

1. Reduce the repeated QoL values into one or two summaries.

2. Analyse each summary as a function of covariates or explanatory variables, $x_1, x_2, \ldots, x_p$. 
• Diggle *et al* (2002) call this strategy a *two-stage* or *derived variable* analysis, and mention that it works well when $x_{ij} = x_i$ for all $i$ and $j$ (i.e. the important explanatory variables do not change over time, such as baseline QoL), since the summary value which results from stage (1) can only be regressed on $x_i$ in stage (2).

• Examples of summary measures include the Area Under the Curve (AUC) or the overall mean of post-randomisation measures.
## Summary measures

<table>
<thead>
<tr>
<th>Type of data</th>
<th>Property to be compared between groups</th>
<th>Summary measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peaked</td>
<td>Overall value of response</td>
<td>Mean or Area Under the Curve</td>
</tr>
<tr>
<td>Peaked</td>
<td>Value of most extreme response</td>
<td>Maximum (minimum)</td>
</tr>
<tr>
<td>Peaked</td>
<td>Delay in response</td>
<td>Time to maximum or minimum</td>
</tr>
<tr>
<td>Growth</td>
<td>Rate of change of response</td>
<td>Linear regression coefficient</td>
</tr>
<tr>
<td>Growth</td>
<td>Final level of response</td>
<td>Final value or (relative) difference between first and last</td>
</tr>
<tr>
<td>Growth</td>
<td>Delay in response</td>
<td>Time to reach a particular value</td>
</tr>
</tbody>
</table>
Area Under the Curve (AUC)

Utility

0.0
("death")

1.0
("full health")

Time (years)

0.25
0.5
0.75
1.0
Calculation of the AUC

- The area can be split into a series of shapes called trapeziums.
- The areas of the separate individual trapeziums are calculated and then summed for each patient.
- Let $Y_{ij}$ represent the QoL response variable observed at time $t_{ij}$, for observation $j = 1, \ldots, n_i$ on subject $i = 1, \ldots, m$.
- The AUC for the $i^{th}$ subject is calculated by

$$AUC_i = \frac{1}{2} \sum_{j=1}^{n_i} (t_{j+1} - t_j)(Y_j + Y_{j+1})$$

- The units of AUC are the product of the units used for $Y_{ij}$ and $t_{ij}$, and may not be easy to understand, since QoL outcomes have no natural units.
- We can calculate the AUC even when there are missing data, except when the first and final observations are missing.
Calculation of a AUC for an individual patient

- In the Acupuncture study, the patients’ QoL was assessed four times; at baseline (0), 3, 12 and 24 months using the SF-36.
- If the time $t_{ij}$ for each QoL assessment is represented as a fraction of a year then the AUCs represent the weighted average level of QoL over the two year period.
- An AUC of 200, corresponds to “good health” over the year, conversely an AUC of 0, corresponds to “poor health” over the period.
- If we divide by the total time (of 2 years) then we get back to the 0 to 100 scale of the original SF-36 measurement which may make interpretation of the results easier.
Calculation of a AUC for an individual patient

- Consider a patient in the Acupuncture study, with SF-36 pain scores of 33.3, 44.4, 55.6 and 77.8 at baseline (0), 3, 12 and 24 months.
- The AUC for this patient is calculated as:
  \[
  0.5 \times \left\{ [0.25 \times (33.3 + 44.4)] + [0.75 \times (44.4 + 55.6)] + [1 \times (55.6 + 77.7)] \right\} = 113.9.
  \]
Comparison of AUCs

• The Area Under the Curve (AUC) is a useful way of summarising the information from a series of measurements on one individual.

• Parametric CIs for the mean difference in AUC between groups can also be calculated as again the AUCs are more likely to be a fairly good fit to the Normal.

• Multiple linear regression methods can be used to adjust AUCs for other covariates (e.g. age, sex, centre).
The Figures show the histograms of the distribution of the AUC summary measure for the SF-36 Pain dimensions separately for the Acupuncture and Usual care groups. Although the distributions are not symmetric, the histograms are not as skewed as the raw data at each time point.

**Figure**: Histograms of SF-36 Pain AUC summary from Acupuncture data by group (data from Thomas et al 2006)
Example: Acupuncture trial AUC analysis

Table: Mean SF-36 Pain scores over time by treatment group with all valid patients at each time-point (data from Thomas et al 2006)

<table>
<thead>
<tr>
<th>SF-36 Pain Outcome‡</th>
<th>Treatment group</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
<th>Mean Difference†</th>
<th>95% CI Lower</th>
<th>Upper</th>
<th>P-value*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Usual care n</td>
<td></td>
<td></td>
<td>Acupuncture n</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time (months)</td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>80</td>
<td>30.4 (18.0)</td>
<td>159</td>
<td>30.8 (16.2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>71</td>
<td>55.4 (25.4)</td>
<td>146</td>
<td>60.9 (23.0)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>68</td>
<td>58.3 (22.2)</td>
<td>147</td>
<td>64.0 (25.6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>59</td>
<td>59.5 (23.4)</td>
<td>123</td>
<td>67.8 (24.1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean follow-up</td>
<td>76</td>
<td>57.2 (19.8)</td>
<td>153</td>
<td>63.4 (20.9)</td>
<td>6.3</td>
<td>0.6</td>
<td>12.0</td>
<td></td>
<td>0.030</td>
</tr>
<tr>
<td>Pain Area under the</td>
<td>SF-36 pain score</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>curve (AUC)</td>
<td>55</td>
<td>112.7 (36.7)</td>
<td>118</td>
<td>125.2 (39.4)</td>
<td>12.6</td>
<td>0.1</td>
<td>25.0</td>
<td></td>
<td>0.048</td>
</tr>
</tbody>
</table>

‡The SF-36 pain dimension is scored on a 0 (poor) to 100 (good health) scale.
P-value from two independent samples t-test.
†A positive mean difference indicates the Acupuncture group has the better QoL.
CI Confidence Interval AUC Area Under the Curve.
AUCs: a cautionary note (1)

- Consider two patients, A and B, whose QoL was assessed at four time points, 0, 6, 12 and 24 months with a utility or preference based QoL measure.
- Patient A, had utility scores of 0.80, 0.70, 0.60 and 0.60 at 0, 6, 12 and 24 months respectively.
- Patient B, had utility scores of 0.50, 0.50, 0.70 and 0.80 at 0, 6, 12 and 24 months respectively.
AUCs: a cautionary note

(2)

Patient A’s QoL declines over time whereas Patient B’s increases over time.

Patients A and B have the same AUC but different utility scores at the four assessment points.
Other summary measures

• The Figures and Tables from the Acupuncture study suggest that SF-36 pain scores at 3, 12, and 24 months follow-up are fairly similar (the lines in the graph appear to be almost horizontal at these time points).

• Therefore another sensible summary measure would be the mean follow-up SF-36 pain score.

• For this summary measure patients need only to have one valid follow-up Pain score.
ANCOVA
(Analysis of Covariance)

• A simple analysis would be to use the two-independent sample $t$-test to compare mean follow-up Pain scores between the two groups.
• The correlation between baseline & mean follow-up pain scores is 0.30.
• Despite this low correlation a more powerful statistical analysis, is an ANCOVA or multiple regression.
• This involves a multiple regression analysis with the average follow-up QoL (the mean of the 3-, 12-, 24-month assessments) as the dependent variable, $\bar{Y}_i$, and the baseline QoL and treatment group (coded Usual care = 0, Acupuncture = 1) as covariates.
• The linear regression model for the $i^{th}$ subject is:

$$\bar{Y}_i = \beta_0 + \beta_{Base}x_{Base_i} + \beta_{Group}x_{Group_i} + \varepsilon_i$$

• where $\varepsilon_i$ is a random error term with $\varepsilon_i \sim N(0, \sigma^2)$ and $\beta_0$ is a constant.
## Results of ANCOVA

Table: Unadjusted and adjusted differences in mean follow-up SF-36 pain outcome scores between Acupuncture and Usual care groups (data from Thomas et al 2006)

<table>
<thead>
<tr>
<th>SF-36 dimension‡</th>
<th>Treatment group</th>
<th>Acupuncture</th>
<th>Unadjusted Difference* 95% CI</th>
<th>P-value</th>
<th>Adjusted† Difference* (95% CI)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean follow-up Pain score</td>
<td><strong>Usual care</strong></td>
<td>N 76</td>
<td>Mean 57.2 (19.8)</td>
<td>n 153</td>
<td>Mean 63.4 (20.9)</td>
<td>P-value 0.030</td>
</tr>
</tbody>
</table>

‡The SF-36 pain dimension is scored on a 0 to 100 (no pain) scale.
† N=229 difference adjusted for baseline pain score.
* Improvement is indicated by a positive difference on the SF-36 pain dimension
Above all - graph the longitudinal QoL data

• Both Diggle (*et al* 2002) and Fayers and Machin (2007) emphasise the importance of graphical presentation of longitudinal data prior to modelling.

• The graphs from the Acupuncture Trial show the mean levels of QoL in patients with low-back pain, before and during treatment, for the Pain dimension of the SF-36.

• The curves do not overlap and there is some evidence to suggest that for later QoL measurements the curves are parallel and that the mean difference between treatments is now fairly constant.
In longitudinal studies, multiple assessments of QoL on the same subject at different time points are used and the within subject responses are then correlated.

This correlation must be accounted for by analysis methods appropriate to the data.

Several models have been proposed for the analysis of such data.

They are usually classified into *marginal* or *random-effects* models.

Random-effects models are also called *generalized linear mixed models* or *multilevel models* or *conditional models*.

The choice of one or the other depends on the objectives of the study.
Modelling longitudinal QoL data

- Figures a) to e) show some simple example profiles of the possible treatment effects on the QoL outcome over time.
- These five graphs lead to the specification of five possible statistical models for the QoL outcome:
  a) QoL Outcome = constant
  b) QoL Outcome = baseline + time
  c) QoL Outcome = baseline + group
  d) QoL Outcome = baseline + time + group
  e) QoL Outcome = baseline + time + group + group*time interaction
Model a)

No time effect and no group effect
(flat horizontal lines that coincide)
Model b)

Time effect but no group effect
(one coincident line with a non-zero gradient)
Model c)

Group effect but no time effect
(two flat parallel horizontal lines)
**Model d)**

Group effect and time effect (two parallel lines with same gradient)

- **Outcome**
- **Time**

![Graph showing parallel lines with treatment and control groups.](image-url)
Group x time interaction effect and (two lines with different gradient)
Order of model fitting

- Ideally these models should be investigated in reverse order i.e. Model e) first.
- If there is no significant group*time interaction then we can a fit a simpler Model d) to the QoL outcome data to see if there is both a significant group and time effect in this model.
- If only the group or time effect was statistically significant, but not both, we would then go on to fit Model b) or c).
- Depending on the results of Model d) we would either go on to fit Model b) if there was no group effect but a significant time effect (see Figure b) or Model c) (See Figure c) if there was no significant time effect but a group effect.
- In the event of no significant group or time effect then model a) and Figure a) is most appropriate for the outcome data.
If $y_{i1}$ and $y_{i2}$ represent the values of two successive QoL assessments by the same ($i^{th}$) patient and $m$ represents the total number of patients completing both assessments in the sample.

Then the equation below measures the strength of association or *auto-correlation* between successive longitudinal measurements of QoL on the same patient,

$$r_T(1,2) = \frac{\sum_{i=1}^{m} (y_{i1} - \bar{Y}_1)(y_{i2} - \bar{Y}_2)}{\sqrt{\sum_{i=1}^{m} (y_{i1} - \bar{Y}_1)^2 \sum_{i=1}^{m} (y_{i2} - \bar{Y}_2)^2}},$$

where $\bar{Y}_1$ and $\bar{Y}_2$ are the sample mean QoL scores at times $t_1$ and $t_2$ respectively. (This is equivalent to Pearson’s product moment correlation coefficient.)
Patterns of autocorrelation

• Several underlying patterns of the auto-correlation matrix $R$ are used in the modelling of QoL data.

1. Independent (sometimes termed random)
2. Unstructured
3. Exchangeable, uniform or compound symmetric.
4. Autoregressive structure (sometimes called multiplicative or time series).

• The auto-correlation pattern affects the way in which the computer packages estimate the regression coefficients in the corresponding statistical model, and so it should be chosen with care.
Patterns of autocorrelation: independent

• Several underlying patterns of the auto-correlation matrix $R$ are used in the modelling of QoL data.

• The error structure is independent (sometimes termed random) if the off diagonal terms of the auto-correlation matrix $R$ are zero.
  – The repeated QoL observations on the same subject are then independent of each other, and can be regarded as though they were observations from different individuals.
**Independent structure**

\[ R_{t,s} = 1 \text{ if } t = s \]
\[ 0 \text{ otherwise} \]

<table>
<thead>
<tr>
<th>Time</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Patterns of autocorrelation: exchangeable

- If all the correlations are approximately equal or uniform then the matrix of correlation coefficients is termed exchangeable, or compound symmetric.
  - This means that we can re-order (exchange) the successive observations in any way we choose in our data file without affecting the pattern in the correlation matrix.
Exchangeable structure

- \( R_{t,s} = 1 \) if \( t = s \)
- \( \rho \) otherwise

<table>
<thead>
<tr>
<th>Time</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>( \rho )</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>( \rho )</td>
<td>( \rho )</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>( \rho )</td>
<td>( \rho )</td>
<td>( \rho )</td>
<td>1</td>
</tr>
</tbody>
</table>
Patterns of autocorrelation: unstructured

- Unstructured imposes only the constraint that the diagonal elements of the working correlation matrix be 1.
- $R_{ts} = 1$ if $t = s$
- $r_{ts}$ otherwise, $r_{ts} = r_{st}$
Example of unstructured autocorrelation

Unstructured imposes only the constraint that the diagonal elements of the working correlation matrix be 1.
\[ R_{ts} = 1 \text{ if } t = s \]
\[ r_{ts} \text{ otherwise, } r_{ts} = r_{st} \]

<table>
<thead>
<tr>
<th>Time</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>( \rho_{10} )</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>( \rho_{20} )</td>
<td>( \rho_{21} )</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>( \rho_{30} )</td>
<td>( \rho_{31} )</td>
<td>( \rho_{32} )</td>
<td>1</td>
</tr>
</tbody>
</table>
Patterns of autocorrelation: autoregressive

- Frequently, as the time between successive observations increases, the auto-correlation between the observations decreases.
- Thus, we would expect a higher auto-correlation between QoL assessments made only two days apart than between two QoL assessments made one month apart.
- A correlation matrix of this form is said to have an autoregressive structure (sometimes called multiplicative or time series).
Autoregressive

\[ R_{t,s} = 1 \text{ if } t = s \]
\[ \rho^{|t-s|} \text{ otherwise} \]

<table>
<thead>
<tr>
<th>Time</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>\rho_{</td>
<td>1-0</td>
<td>}=\rho_1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>\rho_{</td>
<td>2-0</td>
<td>}=\rho_2</td>
<td>\rho_{</td>
</tr>
<tr>
<td>3</td>
<td>\rho_{</td>
<td>3-0</td>
<td>}=\rho_3</td>
<td>\rho_{</td>
</tr>
</tbody>
</table>
Why not always use unstructured correlation?

• It might appear that using this option would be the most sensible one to choose for all longitudinal data sets.
• This is not the case since it necessitates the estimation of many nuisance parameters.
• This can sometimes cause problems in the estimation of the parameters of interest particularly when the sample size is small and the number of time points is large (Rabe-Hesketh and Everitt, 2007).
Observed correlations for the acupuncture QoL data

Table: Auto-correlation matrices for the Pain dimension of the SF-36 from back pain patients in the Acupuncture study assessed at four time points

<table>
<thead>
<tr>
<th>Bodily Pain (n = 173)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
</tr>
<tr>
<td>(months)  0  3  12  24</td>
</tr>
<tr>
<td>0        1.00</td>
</tr>
<tr>
<td>3        0.24        1.00</td>
</tr>
<tr>
<td>12       0.27        0.56 1.00</td>
</tr>
<tr>
<td>24       0.19        0.47 0.57 1.00</td>
</tr>
</tbody>
</table>

- The correlations in the Table clearly show the off-diagonal terms are non-zero and that the assumption of an independent auto-correlation matrix for the marginal model is unrealistic.
- The correlations between the 3 post-baseline QoL assessments at 3, 12 and 24 months are of similar magnitude and range between 0.47 and 0.57.
- This suggests the assumption of an exchangeable correlation structure for the repeated QoL assessment for this data in not unrealistic.
Repeated measures ANOVA

- In some situations QoL assessments may be made over a limited period rather than over an extended time span.
- In this case it may be reasonable to assume that all the subjects complete all the assessments.
- Thus instead of having a fragmented data file with the number of observations for each subject varying from subject to subject, the file has a regular or rectangular shape.
- This enables the repeated measures ANOVA approach to be considered.
- Diggle et al (2002) say that ANOVA has limitations that prevent its recommendation as a general approach for longitudinal data.
  1. It fails to exploit the potential gains in efficiency from modelling the covariance among repeated observations.
  2. ANOVA methods usually require a complete balanced array of data.
  3. The use of repeated measures ANOVA implies an exchangeable auto-correlation between any two observations on the same patient. This may not always be appropriate for QoL assessments.
- It is therefore better to use a regression modelling approach rather than repeated measures ANOVA for analysing longitudinal QoL data.
Simple (independence) model

- Let $y_{ij}$ be the QoL outcome for the $i^{th}$ subject for observation $j$ measured at time $t_{ij}$, for observation $j = 1$ to $n_i$ on subject $i = 1$ to $m$.

- Simple model for the data, assuming independent outcomes is:

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \varepsilon_{ij}$$

- where $t_{ij}$ is the time variable
  - $\varepsilon_{ij}$ is a random error term with $\varepsilon_{ij} \sim N(0, \sigma^2_e)$ and $\text{Corr}(\varepsilon_{ij}, \varepsilon_{ik}) = 0$
  - $\beta_0$ is the mean outcome at baseline
  - $\beta_1$ is the time effect
The basic marginal model takes the same form as the simple (independence) model

\[ y_{ij} = \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij} \]

But the residuals, \( \varepsilon_{ij} \), are correlated i.e. \( \text{Corr}(\varepsilon_{ij}, \varepsilon_{ik}; \mathcal{R}) = \rho(x_{ij}, x_{ik}; \mathcal{R}) \).

The correlation matrix, \( \mathcal{R} \), is usually estimated by an exchangeable correlation matrix, \( R \), that assumes the outcomes for a subject at observation \( j \) are equally correlated with the outcomes at observation \( k \).

But we can also assume an unstructured matrix.

This common correlation, \( \rho \), is the intracluster correlation coefficient (ICC).
In statistics, a generalized estimating equation (GEE) is used to estimate the parameters of a generalized linear model with a possible unknown correlation between outcomes.

Parameter estimates from the GEE are consistent even when the correlation/covariance structure is misspecified.

The focus of the GEE is on estimating the average response over the population ("population-averaged" effects) rather than the regression parameters that would enable prediction of the effect of changing one or more covariates on a given individual.

GEEs are usually used in conjunction with Huber–White standard error estimates, also known as "robust standard error" or "sandwich variance" estimates.

GEEs belong to a class of semiparametric regression techniques because they rely on specification of only the first two moments.

They are a popular alternative to the likelihood–based generalized linear mixed model which is more sensitive to variance structure specification.
Generalized Estimating Equations (GEEs)

• The marginal generalised linear modelling approach uses Generalized Estimating Equations (GEEs) to estimate the regression coefficients (Liang and Zeger, 1986).

• Using GEE any required covariance structure and link function may be assumed and the parameters estimated without specifying the joint distribution of the repeated observations.

• Estimation is via a multivariate analogue of a quasi-likelihood approach (Wedderburn, 1974).
Marginal models

• In the marginal modelling approach, we only need to specify the first two moments of the responses for each person (i.e. the mean and variance).

• With continuous Normally distributed data, the first two moments fully determine the likelihood, but this is not the case for other Generalized Linear Models.

• Since the parameters specifying the structure of the correlation matrix are rarely of great practical interest (they are what is known as nuisance parameters), simple structures (e.g. exchangeable or 1st order autoregressive) are used for the within subject correlations giving rise to the so-called working correlation matrix.

• Liang and Zeger (1986) show that the estimates of the parameters of most interest, i.e. those that determine the mean profiles over time, are still valid even when the correlation structure is incorrectly specified.
Treatment x time interactions

• The non-overlapping lines in graphs imply there is unlikely to be a ‘Treatment x Time’ interaction.
• However, it is still important to test for any such interaction in any regression model.
• Fortunately, with the marginal model approach this is relatively easy to do and simply involves the addition of an extra regression coefficient to the model.
• If treatment is coded as a 0/1 variable (i.e. 0 = Usual Care and 1 = Acupuncture) and assessment time as a continuous variable, then the additional interaction term is simply the product of these two variables (which will be 0 for all the Usual Care group patients and equal to the QoL assessment time in the Acupuncture group patients).
Coding Treatment $\times$ time interactions

- If treatment is coded as a 0/1 variable (i.e. 0 = Usual Care and 1 = Acupuncture) and assessment time as a continuous variable, then the additional interaction term is simply the product of these two variables (which will be 0 for all the Usual Care group patients and equal to the QoL assessment time in the Acupuncture group patients).
Baseline measurements

- For RCTs, with a baseline measurement of the outcome variable, at time 0, since it is not an “outcome” it seems sensible to fit this variable as a covariate in the model and treat it like other baseline covariates such as age, gender and treatment group.
The marginal model

• Let $Pain_{ij}$ be the QoL outcome for the $i^{th}$ subject for observation $j$ measured at time $t_{ij}$, for observation $j = 1$ to $n_i$ on subject $i = 1$ to $m$.

\[
\hat{Pain}_{ij} = \hat{\beta}_0 + \hat{\beta}_1 \text{pain}_i \text{ base}_i + \hat{\beta}_2 \text{time}_{ij} + \hat{\beta}_3 \text{group}_i + \hat{\beta}_4 \text{group}_i \times \text{time}_{ij} + \hat{\epsilon}_{ij}
\]
Estimated regression coefficients from a marginal regression model, in STATA 13 using the *xtgee* procedure with coefficients estimated by GEE with robust standard errors to show the effect of group on outcome, SF-36 pain score, from the Acupuncture RCT (Thomas *et al* 2006) n=229.

The interaction term is not statistically significant. Thus there was no reliable evidence of a ‘Treatment x Time’ interaction. Therefore we can now use a simpler model without the interaction term to test for a group and time effect on QoL.
Estimated regression coefficients from a marginal regression model, in STATA 13 using the `xtgee` procedure with coefficients estimated by GEE with robust standard errors to show the effect of group on outcome, SF-36 pain score, from the Acupuncture RCT (Thomas et al 2006) n=229

```
. xtgee pain pain_base time group, family(gaussian) link(identity) corr(exchangeable) vce(robust)

GEE population-averaged model
Number of obs = 614
Group variable: studyid Number of groups = 229
Link: identity Obs per group: min = 1
Family: Gaussian avg = 2.7
Correlation: exchangeable max = 3
Scale parameter: 541.1862 Wald chi2(3) = 35.74
(Std. Err. adjusted for clustering on studyid)

| Robust | Coef. | Std. Err. | z     | P>|z|  | [95% Conf. Interval] |
|--------|-------|-----------|-------|-------|-------------------|
| pain   |       |           |       |       |                   |
| pain_base | 0.36  | 0.07      | 4.86  | 0.000 | 0.22              | 0.51             |
| time   | 0.24  | 0.09      | 2.79  | 0.005 | 0.07              | 0.41             |
| group  | 6.10  | 2.66      | 2.29  | 0.022 | 0.88              | 11.31            |
| _cons  | 43.46 | 3.57      | 12.18 | 0.000 | 36.46             | 50.45            |
```

```
. estat wcorrelation, compact
Error structure: exchangeable
Estimated within-studyid correlation: .49484305
```

The estimated exchangeable correlation between the outcomes
Analysis of Acupuncture Trial data using a marginal model

Table: Estimated regression coefficients from a marginal model (Model d) to show the effect of treatment (Acupuncture or Usual care) on outcome (SF-36 pain score) over time after adjustment for baseline pain assuming a exchangeable correlation (n=229)

<table>
<thead>
<tr>
<th>Pain*</th>
<th>b</th>
<th>Semi-robust SE(b)</th>
<th>z</th>
<th>P-value</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pain (baseline)</td>
<td>0.4</td>
<td>0.07</td>
<td>4.86</td>
<td>0.001</td>
<td>0.2</td>
<td>0.5</td>
</tr>
<tr>
<td>Time (months)</td>
<td>0.2</td>
<td>0.09</td>
<td>2.79</td>
<td>0.005</td>
<td>0.1</td>
<td>0.4</td>
</tr>
<tr>
<td>Group</td>
<td>6.1</td>
<td>2.66</td>
<td>2.29</td>
<td>0.022</td>
<td>0.9</td>
<td>11.3</td>
</tr>
<tr>
<td>Constant</td>
<td>43.5</td>
<td>3.57</td>
<td>12.18</td>
<td>0.001</td>
<td>36.5</td>
<td>50.4</td>
</tr>
</tbody>
</table>

*The outcome variable is SF-36 pain score with a higher score indicating less pain.*

There is some evidence that SF-36 Pain scores increase over time. The P-value, of 0.022, for the treatment group regression coefficient suggests a significant difference in Pain scores between the Usual care and Acupuncture treated groups.
Checking the assumptions

- The table below shows the estimated within subject correlation matrices for the SF-36 pain outcome if we assume a compound symmetric or exchangeable correlation structure for the repeated QoL assessments.
- The upper diagonal gives the observed matrix before the model fitting. The fitted autocorrelation was 0.49.
- The observed deviation between the fitted model and observed autocorrelations are not too great, suggesting that the assumption of compound symmetry is not unreasonable.

**Table:** Observed and estimated within-patient auto-correlation matrices (exchangeable model) from the low back pain patients in the Acupuncture Trial. The upper diagonal gives the observed matrix before model-fitting whilst the lower gives the exchangeable form after model-fitting

<table>
<thead>
<tr>
<th>Time (months)</th>
<th>3</th>
<th>12</th>
<th>24</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1.00</td>
<td>0.56</td>
<td>0.47</td>
</tr>
<tr>
<td>12</td>
<td>0.49</td>
<td>1.00</td>
<td>0.57</td>
</tr>
<tr>
<td>24</td>
<td>0.49</td>
<td>0.49</td>
<td>1.00</td>
</tr>
</tbody>
</table>

*a) The model contains time, baseline QoL and group as covariates.*
The random effects model, assumes that the correlation arises among repeated responses because the regression coefficients vary across individuals.

Random effects models are particularly useful when inferences are to be made about individuals, rather than the population average.

Thus a random effects approach will allow us to estimate the QoL status of an individual patient.

The regression coefficients, $\beta$, represent the effect of the explanatory variables on an individual patient’s QoL.

This is in contrast to the marginal model coefficients, which describe the effect of the explanatory variables on the population average.

It is based on the assumption that the subjects in the study are chosen at random from some wider patient population.
Random Effects Model

• Let $y_{ij}$ be the QoL outcome for the $i^{th}$ subject for observation $j$ measured at time $t_{ij}$, for observation $j = 1$ to $n_i$ on subject $i = 1$ to $m$ is:

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \omega_i + \epsilon_{ij}$$

where $t_{ij}$ is the time variable and $\epsilon_{ij}$ is a random error term with $\epsilon_{ij} \sim N(0, \sigma^2_e)$;

$\beta_0$ is the mean outcome at time 0 and $\beta_1$ is the time effect.

• $\omega_i$ is the random effect of subject $i$ across all time points with $\omega_i \sim N(0, \sigma^2_\omega)$
Random Effects Model

\[ y_{ij} = \beta_0 + \beta_1 t_{ij} + \omega_i + \epsilon_{ij} \]

- Variation in \( \omega_i \) induces variation in the mean outcome across all subjects.
- Assumes the treatment effect is homogenous across the subjects.
- Sometimes known as the “random intercept” model.
- The fixed portion of the model states that we want one overall regression line representing the population average QoL over time.
- The random effect serves to shift this regression line up or down according to each individual subject.
Random intercepts model

The **bold** line represents the mean regression line for all subjects and the each of the lighter lines represents the regression line for a different subject. The intercept of the $i^{th}$ subject specific regression line differs from that of the mean line by a residual $= \alpha_i - \alpha$ where these residuals are Normally distributed with zero mean and variance $\sigma_\omega^2$. Every line has a common slope $\beta$. 
The random intercept model

- Let \( \text{Pain}_{ij} \) be the QoL outcome for the \( i^{th} \) subject (\( i = 1 \) to \( m \)) for observation \( j \) (\( j = 1 \) to \( n_i \)) measured at time \( t_{ij} \)

\[
\hat{\text{Pain}}_{ij} = \hat{\beta}_0 + \hat{\beta}_1 \text{pain } \text{base}_i + \hat{\beta}_2 \text{time}_{ij} + \hat{\beta}_3 \text{group}_i \\
+ \hat{\omega}_i + \hat{\varepsilon}_{ij}
\]
Estimated regression coefficients from random/mixed effects model, in STATA 13 using the mixed procedure with coefficients estimated by ML to show the effect of group on outcome, SF-36 pain score, from the Acupuncture RCT (Thomas et al 2006) n=229

```
mixed pain pain_base time group || studyid:
Mixed-effects ML regression Number of obs = 614
Group variable: studyid Number of groups = 229
Obs per group: min = 1
avg = 2.7
max = 3
Wald chi2(3) = 36.69
Log likelihood = -2740.2075 Prob > chi2 = 0.0000
------------------------------------------------------------------------------
pain |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
pain_base |   .3613839   .0753737     4.79   0.000     .2136542    .5091137
  time |   .2390793   .0814212     2.94   0.003     .0794967    .3986618
  group |   6.095832   2.719466     2.24   0.025     .7657756    11.42589
  _cons |   43.45541   3.360395    12.93   0.000     36.86916    50.04167
------------------------------------------------------------------------------
Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
studyid: Identity            |
  var(_cons) |   264.0637   35.38377      203.0721    343.3739
-----------------------------+------------------------------------------------
  var(Residual) |   275.8723   19.80373      239.6646    317.5502
------------------------------------------------------------------------------
LR test vs. linear regression: chibar2(01) = 126.03 Prob >= chibar2 = 0.0000
```

The ICC is 264.06/(264.06 + 275.87) = 0.49
Estimated regression coefficients from random/mixed effects model, in STATA 13 using the mixed procedure with coefficients estimated by REML to show the effect of group on outcome, SF-36 pain score, from the Acupuncture RCT (Thomas et al 2006) n=229

```
.mixed pain pain_base time group || studyid:, reml
Mixed-effects REML regression
Group variable: studyid
Number of obs = 614
Number of groups = 229

Obs per group: min = 1
avg = 2.7
max = 3

Log restricted-likelihood = -2740.3706
Wald chi2(3) = 36.27
Prob > chi2 = 0.0000

------------------------------------------------------------------------------
pain |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]  
------------------------------------------------------------------------------
pain_base |   12.850   0.000    36.82845    50.08307
------------------------------------------------------------------------------
Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
studyid: Identity            |                   var(_cons) |   268.7707   36.07768      206.5967    349.6557
                          |                   var(Residual) |   276.5753    19.8798      240.2317    318.4171
------------------------------------------------------------------------------
LR test vs. linear regression: chibar2(01) = 127.26 Prob > chibar2 = 0.0000
```

The ICC is $268.77/(268.77 + 276.57) = 0.49$
Random slopes

• To allow for a random slope over time for each subject we need a random slopes model.
The **bold** line represents the mean regression line for all subjects and the each of the lighter lines represents the regression line for a different subject. The intercept of the \(i^{th}\) subject specific regression line differs from that of the mean line by a residual \(\alpha_i - \alpha\); the slope of the \(i^{th}\) specific regression line differs from that of the mean line by a residual \(\beta_i - \beta\) where these residuals are Normally distributed with zero mean and variance \(\sigma_{o0}^2\) and \(\sigma_{o1}^2\) respectively.
Random slopes model

- Let $y_{ij}$ be the QoL outcome for the $i^{th}$ subject for observation $j$ measured at time $t_{ij}$, for observation $j = 1$ to $n_i$ on subject $i = 1$ to $m$ is:

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \omega_{0i} + \omega_{1i} t_{ij} + \varepsilon_{ij}$$

where $t_{ij}$ is the time variable; $\varepsilon_{ij}$ is a random error term with $\varepsilon_{ij} \sim N(0, \sigma_e^2)$;

$\beta_0$ is the mean (baseline) outcome and $\beta_1$ is the time effect.

- $\omega_{0i}$ is the random (intercept) effect of subject $i$ across all time points with $\omega_{0i} \sim N(0, \sigma_{\omega_0}^2)$

- $\omega_{1i}$ is the random (slope) effect of subject $i$ over time with $\omega_{1i} \sim N(0, \sigma_{\omega_1}^2)$
The random intercept and slopes model

• Let $Pain_{ij}$ be the QoL outcome for the $i^{th}$ subject ($i = 1$ to $m$) for observation $j$ ($j = 1$ to $n_i$) measured at time $t_{ij}$

$$
\hat{Pain}_{ij} = \hat{\beta}_0 + \hat{\beta}_1pain\_\text{base}_i + \hat{\beta}_2time_{ij} + \hat{\beta}_3\text{group}_i \\
+ \hat{\omega}_0i + \hat{\omega}_1time_{ij} + \hat{\epsilon}_{ij}
$$
Estimated regression coefficients from random intercept and slopes effects model, in STATA 13 using the mixed procedure with coefficients estimated by ML to show the effect of group on outcome, SF-36 pain score, from the Acupuncture RCT (Thomas et al. 2006) n=229

![](Mixed-effects ML regression)

| pain | Coef. | Std. Err. | z    | P>|z| | [95% Conf. Interval] |
|------|-------|-----------|------|------|------------------------|
| pain_base | 0.36  | 0.08      | 4.82 | 0.000 | 0.22       0.51     |
| time  | 0.24  | 0.08      | 2.87 | 0.004 | 0.08       0.40     |
| group | 5.98  | 2.72      | 2.20 | 0.028 | 0.64       11.32    |
| _cons | 43.45 | 3.35      | 12.96| 0.000 | 36.88      50.02    |

Random-effects Parameters

<table>
<thead>
<tr>
<th>studyid: Independent</th>
<th>Estimate</th>
<th>Std. Err.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>var(time)</td>
<td>0.08</td>
<td>0.10</td>
<td>0.01     0.98</td>
</tr>
<tr>
<td>var(_cons)</td>
<td>256.24</td>
<td>36.52</td>
<td>193.79   338.82</td>
</tr>
<tr>
<td>var(Residual)</td>
<td>266.66</td>
<td>22.23</td>
<td>226.47   313.98</td>
</tr>
</tbody>
</table>

The random slope effect for time is non-significant from the Likelihood-ratio test.
Missing data

• Different assumptions are required for the marginal and R-E models regarding missing data.
Why does missing data matter?

• Bias

• If the proportion of missing data is small then little bias will result.

• If the proportion of data missing is not small then
  
  – Are the characteristics of patients with missing data different from those for whom complete data are available?
Patterns of missing data

• Missing Completely at Random (MCAR)
  – When the probability of response at time $t$ is independent of both the previously observed values and the unobserved values at time $t$.

• Missing At Random (MAR)
  – When the probability of response at time $t$ depends on the previously observed values but not the unobserved values at time $t$.

• Not Missing At Random (NMAR)
  – When the probability of response at time $t$ depends on the unobserved values at time $t$. 
Describing the extent and patterns of missing data

• Graph QoL outcome vs. time stratified by drop out time.
• Explore missing data mechanisms by comparing those who dropout vs. those who do not via
  – $t$-tests
  – Cross-tabulations
  – Logistic regression
  – Survival analysis to look at predictors of time to dropout
Graphical investigation of patterns of missing data

Mean QoL scores over time by treatment group

- Complete data
- Died after 0.5 year
- Died after 1 year
- Died after 1.5 years

Mean QoL score vs Time (years)
Acupuncture RCT patterns of missing data

Mean SF-36 Pain scores over time by treatment group and completion status

- Control group: complete data to 3 months only (n=8)
- Acupuncture group: complete data to 3 months only (n=6)
- Control group: complete data to 12 months only (n=8)
- Acupuncture group: complete data to 12 months only (n=22)
- Control group: complete data (n=55)

Trajectories over time are similar, so data are likely to MCAR.
Analyses which assume the outcome data are MCAR

- Complete case analysis
- Repeated univariate (time-by-time) analysis
- Marginal models with coefficients estimated by “standard” GEE
- Summary measures, such as AUC
Analyses which assume the outcome data are MAR.

• Random effects models (maximum likelihood methods)
• Multiple imputation (MI)
• Marginal models
  – with coefficients estimated by extensions to GEE with inverse probability weights (IPW), MI or both (doubly robust estimation- DR-GEE)
Missing data

• Different assumptions are required for the two models regarding missing data.
• The marginal model using the GEE requires a missing data process completely at random (MCAR).
  – Under this assumption, missingness does not depend on individual characteristics (observed or not).

• In contrast, random effects models only need the less stringent assumption of missing at random (MAR).
  – In this process, the probability of missingness depends only on observed variables (previous covariates or outcomes).
GEE extensions

• Estimates obtained by marginal models with unweighted GEE as biased for data which are not MCAR.

• We can use marginal models with GEE either with weights, Multiple Imputation (MI) or both.

• Marginal Modelling with weighed GEEs (WGEE) is a two step process.
Marginal Modelling with weighed GEEs (WGEE) is a two step process:

1. Model the probability of the data being observed to obtain predicted probabilities for each subject.

2. Fit a marginal model with GEE using the inverse of the probabilities as weights (IPW).
   - Only observed data is used but this data is weighed to account for those who drop out.

• Doubly robust GEE (DR-GEE) combine inverse probability weighting with Multiple Imputation (MI).
Inverse Probability Weighting (IPW)

1. Generate a response variable $R_{ij} = 1$ if the outcome is observed and 0 otherwise.
2. Use logistic regression to find predictions for $R_{ij}$
3. Obtain fitted or predicted probabilities for being observed/valid data
4. Generate $IPW = 1/\text{Predicted probability of seeing the outcome/data}$

- Subjects with missing outcome data contribute more information to the analysis model.
- Necessary to assume an independent “working covariance” in models.
What if we think the data is MNAR?

• Unfortunately cannot test whether or not the data are MNAR.
• We have not observed the missing QoL score, so it is not possible to formally test the hypothesis that the missingness does not depend on the QoL at the time the assessment is missing.
• The data we need to test the hypothesis are missing!
• More complex models are required.
Analyses which assume the outcome data are MNAR.

- Mixture models (MM)
  - Pattern mixture models are a special case of MM
- Shared random effects parameter models
- Joint multivariate models
- Selection models
- See Fairclough (2002, 2010) and Diggle et al 2002 for more details
In practice both R-E and marginal models provide valid methods for the analysis of longitudinal QoL data. The two approaches lead to different interpretations of between subject effects (particularly for binary outcomes).

Marginal model
- Treatment group coefficients from model represent the average difference between the intervention or control treatments.

Random effects model
- Treatment group coefficients from model represent the difference in effect of offering either the intervention or control treatment on an individual subject.

But for continuous outcomes,
- using a linear regression model, the coefficients from a R-E model can have a marginal interpretation!
Choice of model?

• Choose the model which best answers the scientific research question being asked in the study.
• In RCTs we are clearly interested in the average difference in the treatment effect between the intervention and control groups.
• For this a marginal model appears to be appropriate as the treatment effect of a marginal model represents the average difference between the treatment and control groups across the whole population without being specific to the individuals used in the trial.
• However, in an RCT we may also be interested in the effect of the intervention or control treatment on an individual subject.
• In these circumstances, the R-E model would give the effect of either the intervention or control treatment on an individual subject.
• There is a continuing debate on this subject!
By now you should know about

• How QoL data from longitudinal studies can be summarised, tabulated and graphically displayed.
• How repeated QoL measures for each subject can be reduced to a single summary measure for statistical analysis and how standard statistical methods of analysis can then be used.
• Two extensions of the linear regression model, marginal and random effects models which allows for the fact that successive QoL assessments by a particular patient are likely to be correlated.
Summary

• This session has described how QoL data from longitudinal studies can be summarised, tabulated and graphically displayed.
• This session has shown how repeated QoL measures for each subject can be reduced to a single summary measure for statistical analysis and how standard statistical methods of analysis can then be used.
• Finally, the session has described two extensions of the linear regression model, marginal and random effects models which allows for the fact that successive QoL assessments by a particular patient are likely to be correlated.
Questions?
Exercises

• Now have a go at the exercises on analysing longitudinal QoL data.
Recommended reading

Recommended reading

References (1)


References (3)

- Bell ML and Fairclough DL. Practical and statistical issues in missing data for longitudinal patient reported outcomes. Stat Methods Med Res October 2014 vol. 23 no. 5 440-459
Appendices
## Data format – long form

### Table Format

<table>
<thead>
<tr>
<th>Patient ID</th>
<th>Age</th>
<th>Duration (yrs)</th>
<th>Expectations</th>
<th>IMM Park</th>
<th>Group</th>
<th>comparator</th>
<th>SEX</th>
<th>MRC Park</th>
<th>Inclusion</th>
<th>Time (months)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>87</td>
<td>12</td>
<td>Yes</td>
<td>Yes</td>
<td>Male</td>
<td>Yes</td>
<td>1</td>
<td>12</td>
<td>Flu_200</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>87</td>
<td>12</td>
<td>Yes</td>
<td>Yes</td>
<td>Male</td>
<td>Yes</td>
<td>1</td>
<td>12</td>
<td>Flu_200</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>87</td>
<td>12</td>
<td>Yes</td>
<td>Yes</td>
<td>Male</td>
<td>Yes</td>
<td>1</td>
<td>12</td>
<td>Flu_200</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>87</td>
<td>12</td>
<td>Yes</td>
<td>Yes</td>
<td>Male</td>
<td>Yes</td>
<td>1</td>
<td>12</td>
<td>Flu_200</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>87</td>
<td>12</td>
<td>Yes</td>
<td>Yes</td>
<td>Male</td>
<td>Yes</td>
<td>1</td>
<td>12</td>
<td>Flu_200</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>87</td>
<td>12</td>
<td>Yes</td>
<td>Yes</td>
<td>Male</td>
<td>Yes</td>
<td>1</td>
<td>12</td>
<td>Flu_200</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>87</td>
<td>12</td>
<td>Yes</td>
<td>Yes</td>
<td>Male</td>
<td>Yes</td>
<td>1</td>
<td>12</td>
<td>Flu_200</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>87</td>
<td>12</td>
<td>Yes</td>
<td>Yes</td>
<td>Male</td>
<td>Yes</td>
<td>1</td>
<td>12</td>
<td>Flu_200</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>87</td>
<td>12</td>
<td>Yes</td>
<td>Yes</td>
<td>Male</td>
<td>Yes</td>
<td>1</td>
<td>12</td>
<td>Flu_200</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>87</td>
<td>12</td>
<td>Yes</td>
<td>Yes</td>
<td>Male</td>
<td>Yes</td>
<td>1</td>
<td>12</td>
<td>Flu_200</td>
<td>0</td>
</tr>
</tbody>
</table>

### Variables

- **Cohort ID**: Unique identifier for each subject
- **Age**: Age of subject at start of trial
- **Duration (yrs)**: Duration of study participation
- **Expectations**: Whether the subject expected to benefit from the study
- **IMM Park**: Presence of Immune System involvement
- **Group**: Assignment to treatment group
- **Comparator**: Control group assigned
- **Sex**: Gender of participant
- **MRC Park**: Motor Function Assessment
- **Inclusion**: Details of inclusion criteria
- **Time (months)**: Length of follow-up
• Marginal models can be fitted in R using the `gee` function in the package `gee`.
• Linear mixed “random” effects models are fitted in R by using the `lmer` function contained in `lme4` package.
• Data needs to be in `long` format.
• Simple generalized estimating equations (GEEs) and weighted generalized estimating equations (WGEEs) in longitudinal studies with dropouts: guidelines and implementation in R. *Statistics in Medicine.*
• Alejandro Salazar, Begoña Ojeda, María Dueñas, Fernando Fernández and Inmaculada Failde Version of Record online: 5 APR 2016 | DOI: 10.1002/sim.6947
Software - SAS

• Marginal models can be fitted in SAS using the **PROC GENMOD** function.
• Linear mixed “random” effects models are fitted in SAS by using the **PROC MIXED** function.
• Data needs to be in **long** format.
Robust standard errors

The formula for the robust estimator of variance, for $\beta$, for the OLS model is:

$$
\hat{V}(\beta)_{Robust} = \hat{V}_{OLS}(\sum_{i=1}^{N} u_i^T u_i) \hat{V}_{OLS}.
$$

(1)

Where $\hat{V}_{OLS}$ is the conventional estimator of variance and $u_i$ is a row vector (and $u_i^T$ is its transpose) and the contribution from the $i^{th}$ observation to $\frac{\partial \ln L}{\partial \beta}$ (the first derivative of the log-likelihood for the parameters given the data with respect to $\beta$).

For the simple linear model, $\hat{V}(\beta)_{OLS} = \frac{s^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ where $s^2 = \frac{\sum_{i=1}^{n} e_i^2}{(n - 2)}$ and $e_i$ is the residual for the $i^{th}$ observation i.e. $(e_i = y_i - (\alpha + \beta x_i))$. Then equation (2) becomes:

$$
\hat{V}(\beta)_{Robust} = \frac{s^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \left( \frac{\sum_{i=1}^{N} (e_i x_i)^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \right) \frac{s^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.
$$

(2)
Robust standard errors

In equations 1 and 2, the observations are assumed to be independent. If the observations denoted by $i$ are not independent but can be divided into $M$ clusters or groups, $G^1, G^2, \ldots, G^M$ which are independent, then the robust clustered estimate of variance, is:

$$
\hat{\mathbf{V}}(\beta)_{\text{cluster}} = \hat{\mathbf{V}}_{\text{OLS}} \left( \sum_{j=1}^{M} u_j^{(G)^T} u_j^{(G)} \right) \hat{\mathbf{V}}_{\text{OLS}} . (3)
$$

For the simple linear model $\hat{u}_j^{(G)} = \sum_{j=1}^{(G)} e_{ij} x_{ij}$. The formula for the clustered estimator of the variance is simply that of the robust (unclustered) estimator with the individual $e_i x_i$'s replaced by their sums over each cluster.
Marginal models

• In a marginal model, the regression of the response on the explanatory variables is modelled separately from the within–person correlation.
• We model the marginal expectation, $E(y_{ij})$, as a function of explanatory variables.
• By marginal expectation we mean the average response over the sub-population that shares a common value of $x$.
• Marginal expectation is what we model in a cross sectional study.
Assumptions for Marginal models (Diggle et al 2002)

1) The marginal expectation of the response $E(y_{ij}) = m_{ij}$, depends on explanatory variables, $x_{ij}$, by

$$h(m_{ij}) = b_0 + b_1 x_{ij}$$

where $h$ is a known link function such as the logit for binary responses or log for counts.

2) The marginal variance depends on the marginal means according to

$$\text{Var}(y_{ij}) = n(m_{ij})f$$

where $n$ is a known variance function and $f$ is a scale parameter which may need to be estimated.

3) The correlation between $y_{ij}$ and $y_{ik}$ is a function of the marginal means and additional parameters $a$, i.e.

$$\text{Corr}(y_{ij}, y_{ik}) = r(m_{ij}, m_{ik}; a)$$

where $r ()$ is a known function.

4) The Marginal regression coefficients, $b$, have the same interpretation as coefficients from a cross-sectional analysis.
Fitting a marginal model

1. Fit the standard (naïve) regression model assuming all observations to be independent.
2. Take the residuals from the regression and use these to estimate the parameters which quantify the correlations between observations in the same individual (cluster/therapist).
3. Refit the regression model using a modified algorithm incorporating a matrix, “working correlation matrix R” which
   • Reflects the magnitude of the correlation estimated in step (2).
4. Keep alternating between steps 2 and 3 until the estimates all stabilise.
Marginal vs random effects

• In the linear case it is possible to formulate the two regression approaches to have coefficients with the same interpretation.

• That is coefficients from random effects models can have marginal interpretations as well.
Consider the marginal model

\[ y_{ij} = \beta_0 + \beta_1 t_{ij} + \epsilon_{ij} \]

The marginal modelling approach is to assume:
1. \( E(y_{ij}) = \beta_0 + \beta_1 t_{ij} \);
2. \( \text{Corr}(\epsilon_{ij}, \epsilon_{ik}) = \rho(t_{ij}, t_{ik}; \alpha) \)

Assumption 1 is that the average outcome/QoL for all participants in the population at any time \( t \) is \( \beta_0 + \beta_1 t \).

The parameter \( \beta_1 \) is the change per unit time in the population average QoL.

Assumption 2 specifies the nature of the autocorrelation.
Random slopes model

• Let $y_{ij}$ be the QoL outcome for the $i^{th}$ subject for observation $j$ measured at time $t_{ij}$, for observation $j = 1$ to $n_i$ on subject $i = 1$ to $m$ is:

$$y_{ij} = \beta_0^* + \beta_1^* t_{ij} + \omega_{0i} + \omega_{1i} t_{ij} + \varepsilon_{ij}$$

where $t_{ij}$ is the time variable; $\varepsilon_{ij}$ is a random error term with $\varepsilon_{ij} \sim N(0, \sigma_e^2)$;

$\beta_0$ is the mean (baseline) outcome and $\beta_1$ is the time effect.

• $\omega_{0i}$ is the random (intercept) effect of subject $i$ across all time points with $\omega_{0i} \sim N(0, \sigma_{\omega_0}^2)$

• $\omega_{1i}$ is the random (slope) effect of subject $i$ over time with $\omega_{1i} \sim N(0, \sigma_{\omega_1}^2)$
Random slopes model

In a linear random effects model the regression coefficients also have a marginal interpretation since

\[ E(y_{ij}) = \beta_0^* + \beta_1^* t_{ij} \]

- This is because the average change in QoL for individuals is the same as the change in the population-average over time in a linear model.