M459: Longitudinal Data Analysis

Contents

1 Introduction

1.1 Administration

Books:

Brown, H and Prescott, R (1999). Applied Mixed Models in Medicine. Wiley.

Davidian M and Giltinan DM (1995). Nonlinear Models for Repeated Measurement Data.

Diggle, PJ, Liang, K-Y and Zeger, SL (1994). Analysis of Longitudinal Data. Oxford University Press.

Hand, DJ and Crowder, MJ (1996). Practical Longitudinal Data Analysis. Chapman and Hall.

Jones, RH (1993). Longitudinal Data with Serial Correlation: A State-space Approach. Chapman and Hall.

Lindsey, JK (1993). Models for Repeated Measurements. Oxford University Press.

Verbeke G and Molenberghs G (2000). Linear Mixed Models for Longitudinal Data. Springer.

Assessment:

20%	assignments
30%	project
50%	examination

1.2 Longitudinal data

Longitudinal data occur when responses are obtained on individuals through time. Each subject in a study then provides a vector of time ordered measurements, which, because they are made on the same individual, usually cannot be assumed to be independent. A simple example is a crossover design for a clinical trial, in which two or more treatments are applied sequentially to a subject, allowing the *change* in response to be studied.

Longitudinal data analysis combines ideas from multivariate and time series methodologies: the difference from standard multivariate methods is the time ordering of responses; the difference from time series techniques is that usually the data consist of a relatively large number of relatively short sequences, rather than one (or a small number of) long series.

Example 1. PCA data. Appendix A provides data from a clinical trial into the use of patient controlled analgesia, which allows patients to control their own pain relief following surgery. On request, a machine infuses a bolus of drug provided a sufficient period has elapsed since the previous delivery. The tables shows the numbers of requests in successive four-hourly intervals for two days following abdominal surgery for 65 patients in two groups. In group 1 the bolus was 2mg of morphine and the PCA machine was locked out for 8 minutes after each delivery, and in group 2 the bolus was 1mg of morphine with 4 minute lockout.

This an example of a *balanced* study, in the sense that measurements are made at common times for all patients.

□

Example 2. PANSS data. A clinical trial was carried out into the treatment of schizophrenia. Mental health was assessed using the PANSS score (positive and negative symptom rating scale, high values implying worse condition) with measurements scheduled for 0, 1, 2, 4, 6 and 8 weeks, measured from the start of treatment. Three treatments were considered: placebo, haloperidol and risperidone.

Not all patients completed the trial. We will return to this point later but for now consider only patients who provided all six measurements and inference will be conditional upon completion. 269 subjects completed the trial, of whom 29 were in the placebo group, 41 were treated with haloperidol and 199 were treated with risperidone. The risperidone group is larger mainly because this treatment was subdivided into five different doseage levels. There was little difference in response between these doseage groups, which are therefore pooled for our analyses.

This is another balanced design. □

Example 3. Liver data.

Liver cirrhosis patients were randomly allocated at diagnosis to prednisone treatment or placebo. Data are available for 86 prednisone and 80 placebo patients who survived at least five years after recruitment and all inference is conditional upon this survival. Prothrombin index measurements, an indicator of liver function (high values good), were obtained at entry and then scheduled for 3 months, 6 months, 12 months and annually thereafter, though the achieved times varied considerably between patients, which means this is an unbalanced design. Numbers of measurements per patient ranged from 2 to 12.

□

1.3 Plots

As with any statistical investigation, we need to undertake some exploratory work before performing any sophisticated analysis, to get a feel for the data and to screen for unusual observations. Inspection of appropriate plots is essential.

Individual traces. Figure 1 shows the PCA count data. The left column shows the profiles for the 2mg group, randomly split into three subgroups for presentational clarity, the right column shows the same for the 1mg group. Note that the plots have common scale. What do we learn from this figure?

Dotplots. When the sample sizes are large usually it is not sensible to try to show all data as lines in a figure: all we will see is a mess. Using points allows all the data to be shown but loses the longitudinal structure. Figure 2 illustrates for the liver data. What do we learn from this plot?

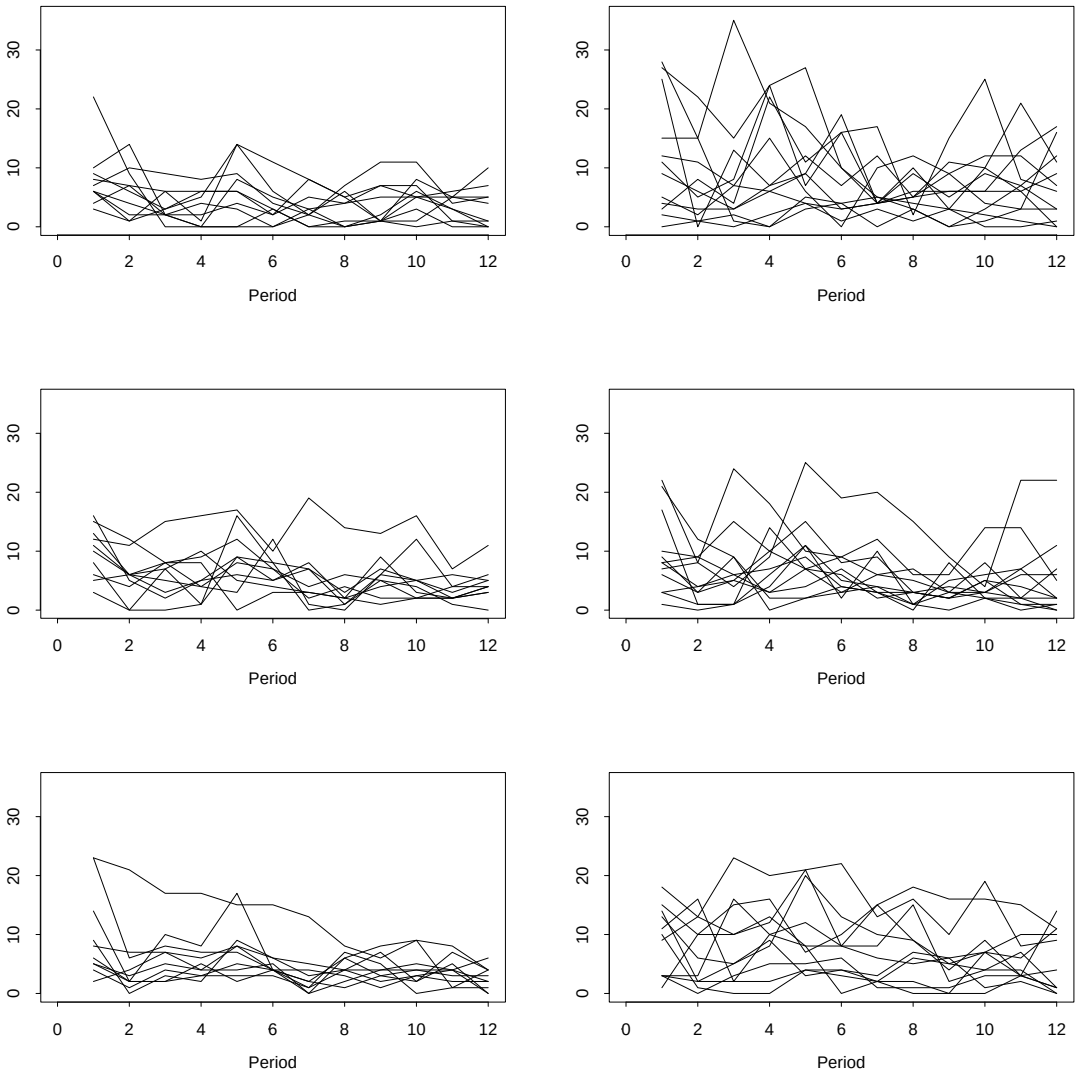


Figure 1: PCA data

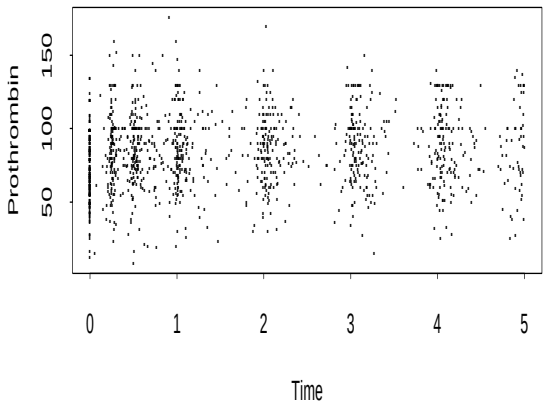


Figure 2: Liver data

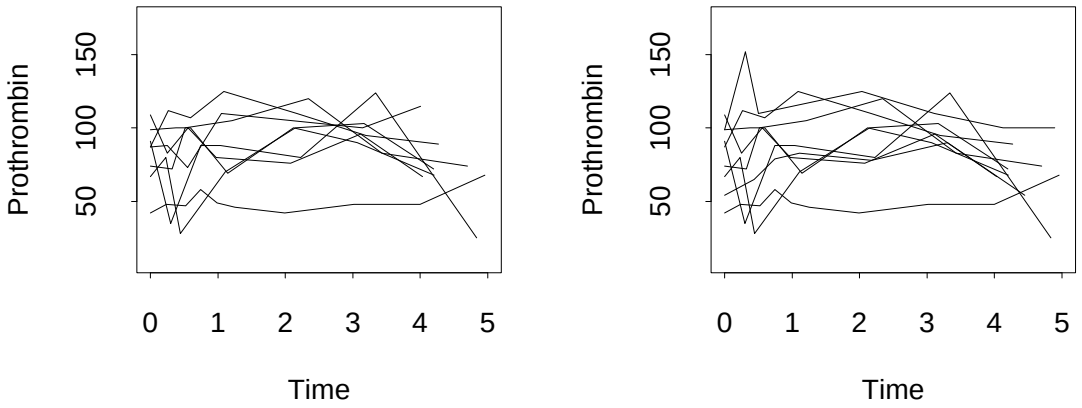


Figure 3: 10% samples of liver data: placebo (left) and prednisone (right)

Selected traces. It is easy to inspect longitudinal traces of all subjects on a computer screen, looking at a few at a time. For a final written report a small selection might be plotted to illustrate the types of pattern which can occur. The selection might be random or judgemental. Figure 3 illustrates with a 10% random sample of the liver data. What do we learn from this plot?

Mean and variance profiles. Inspection of the mean and standard deviation or variance in response against time for each treatment group is always useful. These are easy to calculate for balanced designs: Figures 4 and 5 illustrate. For unbalanced situations some smoothing will be needed (Appendix B) as illustrated in Figure 6 for the liver data means. Note that the vertical scales in Figures 4 and 6 are reduced in comparison with plots of the original data.

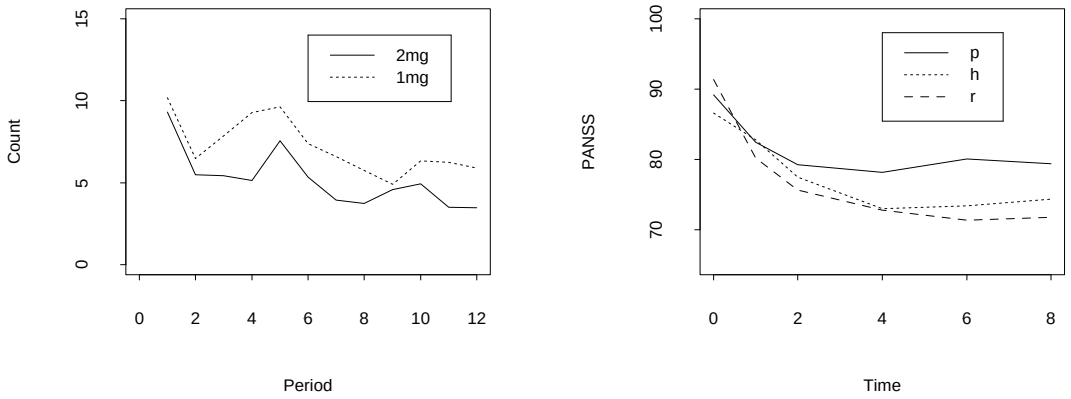


Figure 4: Mean profiles for PCA (left) and PANSS (right) data

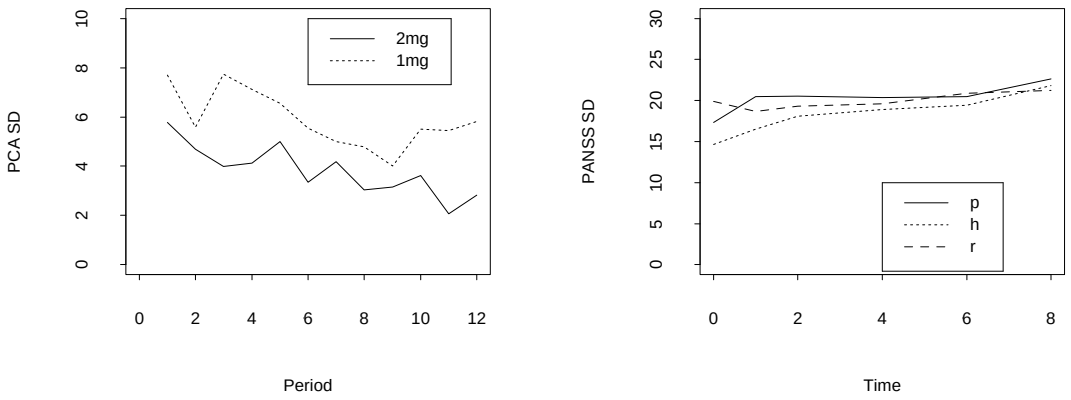


Figure 5: Standard deviation profiles for PCA (left) and PANSS (right) data

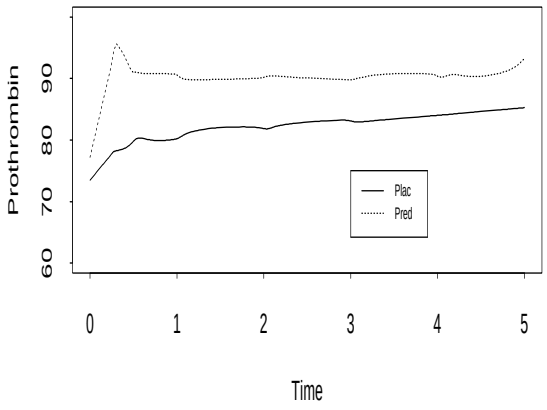


Figure 6: Smoothed means for liver data

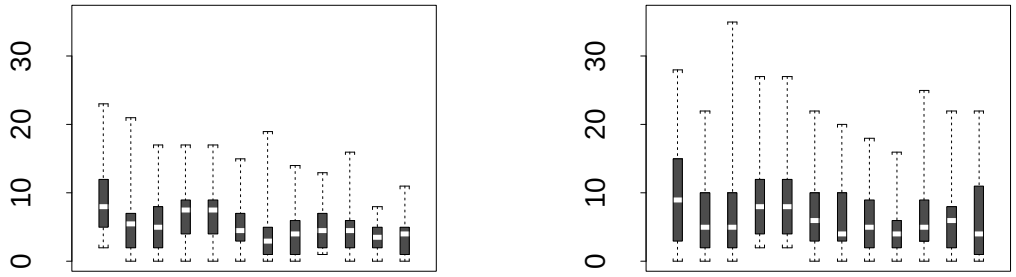


Figure 7: Boxplots of PCA data: 2mg group (left) and 1mg group (right)

Boxplots. Use of the mean profile shows the overall trend but by ignoring between-subject variability can exaggerate time trends and treatment effects. Side-by-side boxplots can be used to show trends and distributions, though at the expense of losing the linkage between observations on the same individual. Figure 7 illustrates.

1.4 Notation

We do not use special notation to distinguish scalars, vectors and matrices - it should be obvious from the context. Nor do we use upper and lower cases to distinguish random variables from observations.

$f(y;\theta)$ - generic for any probability distribution or density and parameter vector.

m - number of subjects.

n_i - number of observations on subject i.

$t_{i1}, t_{i2}, \ldots, t_{in_i}$ - observation times for subject i.

$Y_{i1}, Y_{i2}, \ldots, Y_{in_i}$ - the observations.

$x_{i1}, x_{i2}, \ldots, x_{i,n_i}$ - corresponding p-vectors of covariates (including an intercept term $x_{ij1} = 1$). Covariates may be time-fixed ($x_{ij} = x_{ik}$), time varying, or involve a mixture. Time itself can be a covariate.

Y_i - n_i-vector of responses for subject i.

μ_i, V_i - mean vector and variance matrix of Y_i.

$N = \sum_i n_i$ - total number of observations.

$Y = (Y_1, Y_2, \ldots, Y_m)^T$ - combined N-vector of observations.

μ, V - mean and variance matrix of Y.

Throughout we will assume independence between subjects so that V is *block diagonal.*

2 Simple analysis strategies

2.1 Derived variables

Sometimes we do not need any complicated analysis technique which allows for within-subject dependence. If interest is only in some summary of the individual's profile, we can reduce the n_j vector Y_j to a single summary statistic or *derived variable*, which can be analysed using standard univariate methods. If the summary is based on unequal numbers of observations then care must be taken to allow for the possibility of unequal variances.

Reduced data. We may be interested only in the response at one time period, probably the last, and may decide to ignore the others. Alternatively we may be interested only in the change between first and last and may ignore the intermediate observations. Analysis of such reduced data is valid but inefficient, and also raises the question of why all repeated measurements are taken if they are not to be considered in the analysis.

Example. Analyses of the week 0 and week 8 PANSS data:

```
ANALYSIS OF VARIANCE ON PANSS WEEK 0
SOURCE     DF        SS        MS        F        p
TMT         2       814       407     1.14    0.323
ERROR     266     95216       358
TOTAL     268     96030

                                    INDIVIDUAL 95% CI'S FOR MEAN
                                    BASED ON POOLED STDEV
 LEVEL      N      MEAN     STDEV
 ---------+---------+---------+-------
     1     29     89.14     17.35     (-------------*-------------)
     2     41     86.61     14.65   (----------*-----------)
     3    199     91.35     19.87                  (-----*----)
                                    ---------+---------+---------+-------
POOLED STDEV =    18.92                85.0      90.0      95.0



ANALYSIS OF VARIANCE ON PANSS WEEK 8
SOURCE     DF        SS        MS        F        p
TMT         2      1543       772     1.68    0.189
ERROR     266    122433       460
TOTAL     268    123976



                                    INDIVIDUAL 95% CI'S FOR MEAN
                                    BASED ON POOLED STDEV
 LEVEL      N      MEAN     STDEV  --------+---------+---------+--------
     1     29     79.38     22.64         (------------*------------)
     2     41     74.32     21.82   (----------*----------)
     3    199     71.79     21.21     (----*----)
                                    --------+---------+---------+--------
POOLED STDEV =    21.45             72.0      78.0      84.0
```

Conclusions?

□

Means or totals. Self explanatory. Eg for for the PCA data we may be interested only in the total number of requests made over the 48 hours. After a log transformation to give approximate Normality:

```
T-TEST FOR LOG(PCA COUNT) DATA

TWOSAMPLE T FOR C17
C1    N      MEAN     STDEV   SE MEAN
0    30     4.037     0.437     0.080
1    35     4.302     0.581     0.098

95 PCT CI FOR MU 0 - MU 1: ( -0.517,  -0.011)

TTEST MU 0 = MU 1 (VS NE): T= -2.09  P=0.041  DF=  62
```

What other test might we be interested in performing?

□

Regression coefficients. We might be able to summarise individual profiles through a small number of regression coefficients, which can then be considered as responses in their own right. To illustrate, for the liver data we fit to profile i the linear model

$$E[Y_{ij}] = \alpha_i + \beta_i t_{ij}.$$

A summary of the estimates is

		alpha		beta	
	m	mean	sd	mean	sd
Placebo	80	77.33	18.99	2.355	5.718
Prednisone	86	85.76	19.70	1.926	6.828

Comments?

□

Area under the curve. This is popular for unbalanced designs. As the name implies, the response profile is summarised by the total area under the line segments joining the measurements. Clearly analysis of this summary is closely related to that of the mean response. A related summary is the total area under (or over) some threshold. For instance, a level of 100 in prothrombin index is considered normal. The total area under this level for each patient might be calculated and used for analysis.

2.2 Independence working assumption

If there is no real interest in the correlation structure a *marginal* analysis might be appropriate. The idea is to write down a model for the marginal distribution of each observation (ie ignoring the others on the same subject) and then form a *pseudo-likelihood* as if the joint distribution of all measurements on a subject is given by the product of the marginals, which of course is only really true if the observations are independent. In other words we make an *independence working assumption*. Maximisation of this pseudo-likelihood leads to consistent estimates of the parameters involved in the marginals, though the standard errors obtained from the observed information will not be correct and have to be adjusted.

In more detail, for balanced designs assume that the marginal density of observations at time t_j is $f_j(y, \theta_j)$. Denote the combined parameter vector made up of all distinct terms in $\theta_1, \theta_2, \ldots$ by θ (some of the θ_j may have elements in common with others). Then we obtain a consistent estimate of θ_j by maximising the marginal likelihood

$$L_j(\theta_j) = \prod_i f_j(y_{ij}; \theta_j)$$

or equivalently the marginal log-likelihood

$$l_j(\theta_j) = \sum_i \log\{f_j(y_{ij}; \theta_j)\}.$$

If the observations were independent then the full likelihood would be

$$L(\theta) = \prod_j L_j(\theta_j),$$

the full log-likelihood would be

$$l(\theta) = \sum_j l_j(\theta_j), \tag{1}$$

and maximisation would lead to consistent estimation as usual. Because (1) is additive this result also holds when the data are *not* independent: provided each $l_j(\theta_j)$ is correctly specified we still get consistent estimation by solving

$$u(\theta) = \partial l(\theta)/\partial\theta = 0.$$

Thus we can estimate many parameters without making any assumptions whatsoever about the correlation structure. Efficiency may be lost, but not validity of the estimates, $\tilde{\theta}$ say. These ideas also work for unbalanced data.

This method has the advantages of simplicity and robustness against misspecified correlation structures, though we do need to be sure the marginal distributions $f_j(.)$ are correctly specified. Disadvantages are:

- it does not lead to estimates of parameters which do not appear in the marginals, which usually means the association/correlation parameters;
- standard errors for $\tilde{\theta}$ obtained from the observed pseudo-information are not correct, even asyptotically, unless the observations are genuinely mutually independent.

The second of these difficulties can be overcome by using a *robust* or *sandwich* variance estimator for $\tilde{\theta}$, as shown in an Econometrics paper by White in 1982. First, let $A(\theta)$ be the observed information obtained from the pseudo-likelihood:

$$A(\theta) = -\partial^2 l(\theta)/\partial\theta^2,$$

remembering that usually θ will be a vector and so $A(\theta)$ will be a matrix. Now let $u_i(\theta)$ be the score contribution from subject i, ie

$$u_i(\theta) = \sum_j \partial \log\{f_j(y_{ij};\theta_j\}/\partial\theta$$

so that $u(\theta) = \sum_i u_i(\theta)$. Next form a matrix

$$B(\theta) = \sum_i u_i(\theta)u_i(\theta)^T.$$

Then it turns out that (subject to the usual regularity conditions) $\tilde{\theta}$ is asmptotically unbiased and has variance (matrix) which can be consistently estimated by

$$C(\theta) = A(\tilde{\theta})^{-1}B(\tilde{\theta})A(\tilde{\theta})^{-1}.$$

Appendix C outlines why in the case of a correctly specified likelihood both $A(\theta)$ and $B(\theta)$ are estimates of the same thing, so that $C(\theta)$ and $A(\theta)^{-1}$ are asymptotically equivalent at the true value of θ.

Example. We will see later that a reasonable assumption for the PCA data is that the count Y_{ij} for person i in period j has a negative binomial distribution

$$P(Y_{ij} = y_{ij}) = \frac{\lambda_{ij}^{y_{ij}}}{y_{ij}!(1+\xi\lambda_{ij})^{(y_{ij}+1/\xi)}} \prod_{k=0}^{y_{ij}-1}(1+k\xi).$$

Here $\xi > 0$ is a parameter which measures overdispersion in comparison with Poisson counts with mean $\lambda_{ij} = \alpha_j e^{\beta x_i}$, where $x_i = 0$ for the 2mg group and $x_i = 1$ for the 1mg group. Note that some parameters appear in one period only (ie α_j) and others affect the counts in all periods (β, ξ). Maximum pseudo-likelihood estimates under an independence working assumption are given in the following table together with naive (from $A(\tilde{\theta})$) and robust (from $C(\tilde{\theta})$) standard errors and t-statistics.

		Naive		Robust	
	Est	SE	T	SE	T
α_1	8.16	0.82	10.00	0.85	9.64
α_2	5.00	0.53	9.50	0.64	7.77
α_3	5.52	0.58	9.56	0.68	8.13
α_4	5.95	0.62	9.57	0.69	8.60
α_5	7.17	0.73	9.84	0.76	9.45
α_6	5.28	0.55	9.52	0.54	9.75
α_7	4.36	0.47	9.26	0.59	7.39
α_8	3.93	0.43	9.16	0.49	8.07
α_9	3.98	0.43	9.23	0.45	8.92
α_{10}	4.70	0.50	9.41	0.54	8.68
α_{11}	4.04	0.44	9.16	0.42	9.64
α_{12}	3.87	0.42	9.12	0.48	8.08
β	0.34	0.06	5.81	0.13	2.60
ξ	0.49	0.03	14.09	0.05	9.00

Why are the robust standard errors generally larger than the naive ones?

□

3 Normal linear model with correlated errors

3.1 Model and assumptions

In this section we assume that the n_i-vector of responses for subject i has multivariate Normal distribution and that the mean follows a linear model. So we have a linear regression problem with the complication that the errors may not be independent. In the following we shall concentrate on the special features of and requirements for regression analysis of longitudinal data. We omit discussion of standard methods in model building and diagnostic assessment, taking the importance of these as given, noting in particular that model building is an iterative process. Some revision on methods for multiple linear regression is provided in Appendix D.

The model for subject i is

$$Y_i = X_i\beta + \epsilon_i \tag{2}$$

where

X_i is $n_i \times p$ with j'th row x_{ij}^T

β is a p-vector of coefficients to be estimated

$\epsilon_i \sim \mathrm{N}(0, V_i)$ is an n_i-vector of errors.

When the errors have common variance σ^2 we will write $V_i = \sigma^2 W_i$, where W_i is the correlation matrix.

Note that we can combine the equations (2) column-wise to produce the single model

$$Y = X\beta + \epsilon \tag{3}$$

with $\epsilon \sim \mathrm{N}(0, V)$ (and $V = \sigma^2 W$ in the homoscedastic case). Given independence between subjects, V is block-diagonal.

Recall that the ordinary least squares (OLS) procedure yields an unbiased estimate

$$\hat{\beta}_{OLS} = (X^T X)^{-1} X^T Y$$

whatever the correlation structure. This is also the maximum pseudo-likelihood estimate under an independence working assumption (as previous section) and can be used with the robust variance estimator to provide valid inference, at least asymptotically. However, efficiency can be lost. Nonetheless an OLS analysis usually makes a natural starting point.

3.2 Exploring the correlation structure

Let the vector of OLS residuals for subject i be

$$R_i = Y_i - X_i \hat{\beta}_{OLS}$$

For large samples R_i *should* exhibit similar properties to ϵ_i and can be used to assess the variance and correlation structure.

For balanced data scatterplot and correlation matrices of residuals at the n time points can both be informative.

Example. For the PANSS data we fit a saturated model which allows a different mean for each of the 18 treatment×time combinations. This means for each observation Y_{ij} there is an 18-vector x_{ij} whose elements are all zero except

$$x_{ij,j} = 1 \text{ to measure the time effect in the placebo group,}$$

$$x_{ij,j+6} = 1 \text{ if haloperidol, to measure the difference from placebo at time } j,$$

$$x_{ij,j+12} = 1 \text{ if risperidone, to measure the difference from placebo at time } j.$$

Estimated coefficients are

	0	1	2	4	6	8
p	89.138	82.414	79.241	78.138	80.034	79.379
h-p	-2.528	0.391	-1.778	-5.138	-6.669	-5.062
r-p	2.209	-2.248	-3.588	-5.359	-8.683	-7.590

Residual variances are

Week	0	1	2	4	6	8
Var	355.3	342.6	368.3	380.5	422.7	456.8

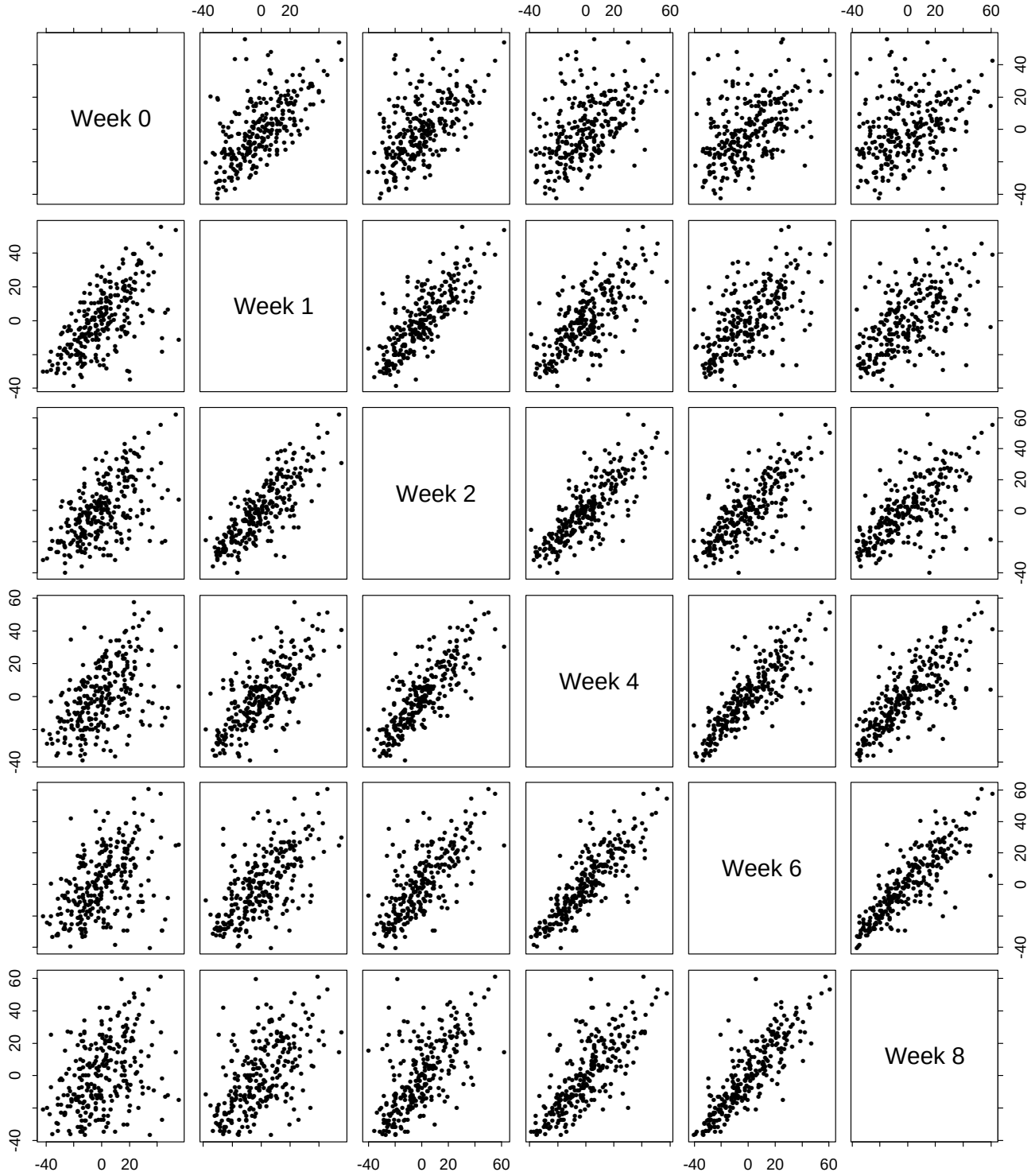


Figure 8: PANSS OLS residuals

and the residual correlation matrix is

Week	0	1	2	4	6	8
0	1.000	0.636	0.568	0.507	0.445	0.378
1	0.636	1.000	0.806	0.718	0.639	0.573
2	0.568	0.806	1.000	0.815	0.728	0.638
4	0.507	0.718	0.815	1.000	0.845	0.765
6	0.445	0.639	0.728	0.845	1.000	0.855
8	0.378	0.573	0.638	0.765	0.855	1.000

A scatterplot matrix of the residuals is in Figure 8. What do these results tell us?

□

For unbalanced data the *sample variogram* is more useful. With a slight notation change, the variogram (sometimes called the semi-variogram) is defined by

$$\gamma(u) = \frac{1}{2}E[\{R(t) - R(t+u)\}^2].$$

Note that if $R(t)$ is stationary (so the distribution doesn't depend on t) the variogram is equivalent to $\text{Var}(R)\{1 - \text{Corr}(R(t), R(t+u))\}$. The *sample variogram* is a smooth trace through a scatterplot of

$$\frac{1}{2}(R_{ij} - R_{ik})^2$$

plotted against the separation $|t_{ij} - t_{ik}|$. Sometimes extreme distances are omitted, and often only the smooth curve is given - the points usually just look like a mess.

Example. For the liver data we fit (by OLS) a model which includes 6 covariates:

$x_{ij,1} = 1$ to measure the intercept in the placebo group

$x_{ij,2} = t_{ij}$ to measure slope in the placebo group

$x_{ij,3} = I(\text{tmt} = \text{pred})$ to measure the difference in intercept in the prednisone group from placebo

$x_{ij,4} = t_{ij}I(\text{tmt} = \text{pred})$ to measure the difference in slope in the prednisone group from placebo

$x_{ij,5} = I(t_{ij} = 0)$ to account for any sudden change after time 0 in the placebo group

$x_{ij,6} = I(t_{ij} = 0)I(\text{tmt} = \text{pred})$ to allow that change to be different in the prednisone group

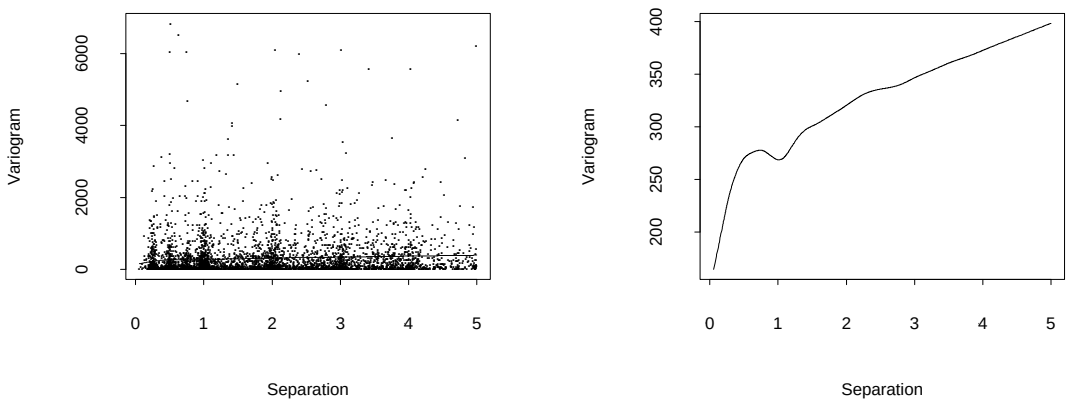


Figure 9: Sample variogram for liver data with and without points. Note different vertical scales.

Ordinary least squares estimates of the corresponding coefficients β are

```
78.065   1.594  12.848  -1.746  -5.103 -10.997
```

and the sample variogram is in Figure 9. Comments?

□

3.3 Generalised least squares for known variance structure

Now suppose (temporarily) that the variance matrix V is known up to a scale factor, ie $V = \sigma^2 W$ with $N \times N$ block diagonal W fully known. The log-likelihood based on the combined vector Y is

$$-\frac{N}{2}\log\{2\pi\sigma^2\} - \frac{1}{2}\log\{|W|\} - \frac{1}{2\sigma^2}(Y - X\beta)^T W^{-1}(Y - X\beta)$$

so that the maximum likelihood estimator of β minimises the *generalised* or *weighted* sum of squares (and products)

$$GSS = (Y - X\beta)^T W^{-1}(Y - X\beta).$$

This leads to the estimator

$$\hat{\beta}_{GSS} = (X^T W^{-1} X)^{-1} X^T W^{-1} Y \tag{4}$$

Proof:

□

The scale parameter σ^2 (which is the variance when errors are homoscedastic and W is a correlation matrix) is estimated by

$$\hat{\sigma}^2 = \frac{1}{N}(Y - X\beta_{GSS})^T W^{-1}(Y - X\beta_{GSS})$$

Proof: exercise.

Notes:

1. Some simplification of $\hat{\sigma}^2$ is achieved by writing

   $$Y - X\beta_{GSS} = \{I - X(X^T W^{-1} X)^{-1} X^T W^{-1}\}Y = DY$$

   say, and noting $D^T W^{-1} D = W^{-1} D$ so that

   $$\hat{\sigma^2} = \frac{1}{N} Y^T W^{-1} DY.$$

2. $E[\beta_{GSS}] = \beta \qquad \mathrm{Var}(\beta_{GSS}) = \sigma^2 (X^T W^{-1} X)^{-1} \qquad E[\hat{\sigma}^2] = (N-p)\sigma^2/N$

3. Because of the block-diagonal structure we can avoid the need to invert an $N \times N$ matrix W by noting

   $$W = \mathrm{diag}(W_1, W_2, \ldots, W_m)$$

   and so

   $$W^{-1} = \mathrm{diag}(W_1^{-1}, W_2^{-1}, \ldots, W_m^{-1}).$$

   This means

   $$X^T W^{-1} X = \sum_i X_i^T W_i^{-1} X_i$$

   and

   $$\hat{\beta}_{GSS} = (\sum_i X_i^T W_i^{-1} X_i)^{-1} \sum_i X_i^T W_i^{-1} Y_i.$$

   So to obtain $\hat{\beta}_{GSS}$ we need to invert m matrices W_i (dimension $n_i \times n_i$) and the final $p \times p$ matrix $X^T W^{-1} X$.

4. We can also used *generalised* or *weighted* least squares to estimate β by minimising

   $$(Y - X\beta)^T G (Y - X\beta)$$

   for arbitrary weight matrix G. The estimate is not then maximum likelihood but is always unbiased, has properties which depend on X, V, G and β, and often will be quite efficient. Moreover this method can be applied to linear models with non-Normal errors.

Example. The variogram in Figure 9 suggests that the correlation in the liver data decreases with separation between measurements. As a first approximation we shall assume an *exponential correlation structure*

$$\text{Corr}(Y_{ij}, Y_{ik}) = \rho^{|t_{ij} - t_{ik}|}.$$

For this illustrative analysis we assume $\rho = 0.25$, returning to this point later. Estimates are

```
beta   78.401   1.751  13.381  -2.097  -3.138 -15.242
se      2.424   0.864   3.379   1.200   2.235   3.163
sigma  24.294
```

Corresponding estimates under an assumption of independence (using OLS) are

```
beta   78.065   1.594  12.848  -1.746  -5.103 -10.997
se      1.804   0.727   2.514   1.011   3.294   4.580

sigma  24.647
```

3.4 Parametric variance structures

Now suppose that the variance matrix V depends upon an unknown parameter vector α, where the dimensionality of α usually is small. For instance in the previous example with exponential correlation matrix, V depended upon σ^2 and ρ. Another common correlation structure (called *compound symmetry*) has all correlations equal

$$\text{Corr}(Y_{ij}, Y_{ik}) = \rho$$

and again V depends upon σ^2 and ρ. Other correlation structures will be discussed later.

Write $V = V(\alpha)$ to remind us of this dependence. Then β and α can be estimated jointly by maximum likelihood. First note that for fixed α the results of the previous section apply and we have a closed form expression for the estimate (dropping the GSS subscript)

$$\hat{\beta}(\alpha) = (X^T V^{-1}(\alpha) X)^{-1} X^T V^{-1}(\alpha) Y.$$

This can be substituted back into the log-likelihood to give, apart from a constant, the *profile* log-likelihood for α

$$\begin{aligned} l(\alpha) &= -\frac{1}{2}\log\{|V(\alpha)|\} - \frac{1}{2}\{Y - X\beta(\alpha)\}^T V^{-1}(\alpha)\{Y - X\beta(\alpha)\} \\ &= -\frac{1}{2}\log\{|V(\alpha)|\} - \frac{1}{2} RSS(\alpha) \quad \text{(say)}. \end{aligned}$$

Given the assumed independence between units, for calculation we have

$$\log\{|V(\alpha)|\} = \sum_{i=1}^{m} \log\{|V_i(\alpha)|\}$$

and

$$RSS(\alpha) = \sum_{i=1}^{m} \{Y_i - X_i\beta(\alpha)\}^T V_i^{-1}(\alpha)\{Y_i - X_i\beta(\alpha)\}.$$

Usually we maximise $l(\alpha)$ by a numerical search over α, which avoids the need to calculate derivatives (difficult for determinants). This means we can't use the observed information to calculate standard errors for α, which must be obtained by other methods. Often α is treated as a nuisance parameter and we do not consider its standard errors, interest mainly being in β. We then simply plug in $\hat{\alpha}$ as if it was fixed into the results of the previous section.

Notes:

1. In the homoscedastic case we can extract the common variance σ^2 from α and estimate it directly by
$$\hat{\sigma}^2(\alpha^*) = RSS_1(\alpha^*)/N$$
where $\alpha = (\alpha^*, \sigma^2)$ and
$$RSS_1(\alpha^*) = \{Y - X\hat{\beta}(\alpha^*)\}^T W^{-1}(\alpha^*)\{Y - X\hat{\beta}(\alpha^*)\}.$$
We then choose α^* to maximise
$$l_1(\alpha^*) = -\frac{1}{2}\log\{|W(\alpha^*)|\} - \frac{N}{2}\log\{RSS_1(\alpha^*)\}.$$

2. Maximum likelihood estimates can have finite sample bias: eg for a single univariate sample $\hat{\sigma}^2 = \sum(y_i - \bar{y})^2/n$ rather than the unbiased $s^2 = \sum(y_i - \bar{y})^2/(n-1)$. If N is large in comparison with the number of unknown parameters usually the bias can be ignored. Otherwise *restricted maximum likelihood* can be used, under which
$$\hat{\sigma}^2(\alpha^*) = RSS_1(\alpha^*)/(N-p)$$
and α^* maximises
$$l_1^R(\alpha^*) = -\frac{1}{2}\log\{|W(\alpha^*)|\} - \frac{N}{2}\log\{RSS_1(\alpha^*)\} - \frac{1}{2}\log|X^T W^{-1}(\alpha^*)X|$$
rather than $l_1(\alpha^*)$. The idea behind this method is to estimate α from linearly transformed data HY, where H is a matrix selected so that the distribution of HY does not depend on β. The effect is to introduce into the likelihood based on HY an additional bias correction term (of order p) in comparison with the likelihood based on Y, namely the final term in $l_1^R(\alpha^*)$.

Example. Returning to the liver data with exponential correlation structure, a profile likelihood plot for ρ is shown in Figure 10. The maximum likelihood estimate is $\hat{\rho} = 0.12$ (with likelihood based 95% CI 0.09 to 0.15) and at this value the other estimates are

```
beta   78.701   1.633  12.879  -1.913  -3.515 -14.596
se      2.247   0.816   3.134   1.134   2.386   3.370

sigma  23.686
```

Alternative correlation structures are discussed later. □.

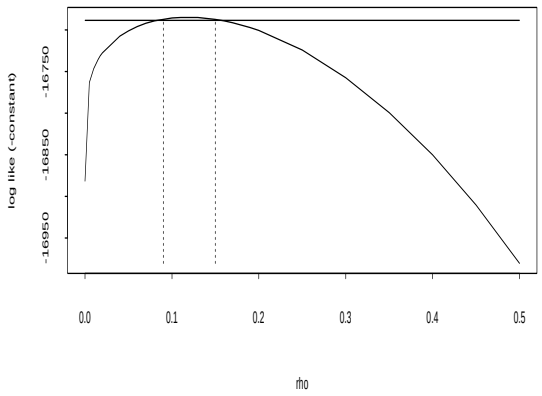


Figure 10: Profile log-likelihood for liver data parameter ρ

3.5 General variance structure for balanced designs

Now suppose that we do not wish to make any assumptions about the form of the variance structure. This is realistic only for balanced data, which will be assumed for this section. Thus all subjects have common measurement times and are assumed to have common $n \times n$ variance matrix V_0. Again we can estimate β directly for given V_0 and can substitute back into the log-likelihood to give

$$l(V_0) = -\frac{m}{2}\log\{|V_0|\} - \frac{1}{2}\sum_{i=1}^{m}\{Y_i - X_i\beta(V_0)\}^T V_0^{-1}\{Y_i - X_i\beta(V_0)\}.$$

Or in the homoscedastic case $V_0 = \sigma^2 W_0$

$$\hat{\sigma}^2(W_0) = \frac{1}{N}RSS(W_0) = \frac{1}{N}\sum_{i=1}^{m}\{Y_i - X_i\beta(W_0)\}^T W_0^{-1}\{Y_i - X_i\beta(W_0)\}$$

and

$$l_1(W_0) = -\frac{m}{2}\log\{|W_0|\} - \frac{N}{2}\log\{RSS(W_0)\}.$$

In principle we can choose the $n(n+1)/2$ elements of V_0 (or the $n(n-1)/2$ elements of W_0) to maximise the likelihood. In practice this is feasible only for small n. In practice an iterative reweighting procedure is reasonably simple to apply and usually converges in a few cycles.

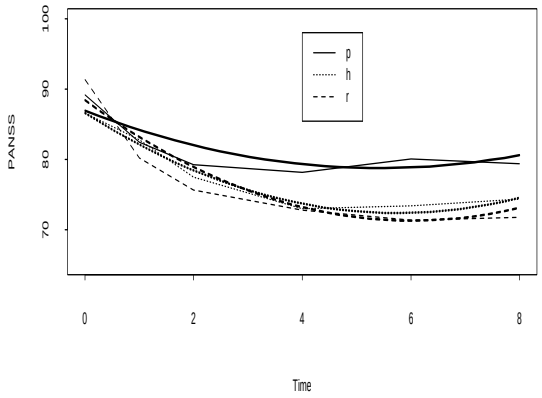
1. Begin with the OLS estimate $\hat{\beta}$ of β.
2. Find the residuals $R_i = Y_i - X_i\beta$ (each $n \times 1$).
3. Estimate V_0 by
$$\frac{1}{m}\sum_{i=1}^{m} R_i R_i^T$$
(the average estimated residual correlation matrix).

4. Update $\hat{\beta}$ using generalised least squares and return to 2.

Example. For the PANSS data we fit quadratic time curves to each treatment profile using this method, hence using 9 covariates

$$x_{ij1} = 1,\ x_{ij2} = t_j,\ x_{ij3} = (t_j - 4)^2$$
$$x_{ij4} = I(\text{tmt} = h),\ x_{ij5} = t_j I(\text{tmt} = h),\ x_{ij6} = (t_j - 4)^2 I(\text{tmt} = h)$$
$$x_{ij7} = I(\text{tmt} = r),\ x_{ij8} = t_j I(\text{tmt} = r),\ x_{ij9} = (t_j - 4)^2 I(\text{tmt} = r).$$

Figure 11 shows the elements of β for the first 10 iterations of this procedure: it seems that only three or four are required. The final variance matrix is close to that to be shown in the next section. Fitted curves are



□

3.6 ANOVA methods

Longitudinal data analysis has its roots in *repeated measures* analysis, in turn derived from early work on *split plots* experiments. Without computer facilities for matrix manipulation, regression methods were infeasible and most early work was based on ANOVA methods. Regression methods are in general preferable for longitudinal data, though ANOVA can be useful in certain limited circumstances. First, we can use standard univariate ANOVA to compare treatment groups for any of the derived variables discussed earlier. (Revision of basic ANOVA is in Appendix E). Second, split-plots ANOVA is appropriate for balanced longitudinal data without covariates provided the correlation matrix is *spherical*. This means that the variance of all within-subject differences $Y_{ij} - Y_{ik}$ is a constant. For instance with three observations per subject, the variance matrix

$$v = \begin{pmatrix} 1 & 0.5 & 1.5 \\ 0.5 & 3 & 2.5 \\ 1.5 & 2.5 & 5 \end{pmatrix}$$

is spherical. Probably the only realistic form of sphericity is compound symmetry, under which all variances are σ^2 and all within-subject correlations are ρ.

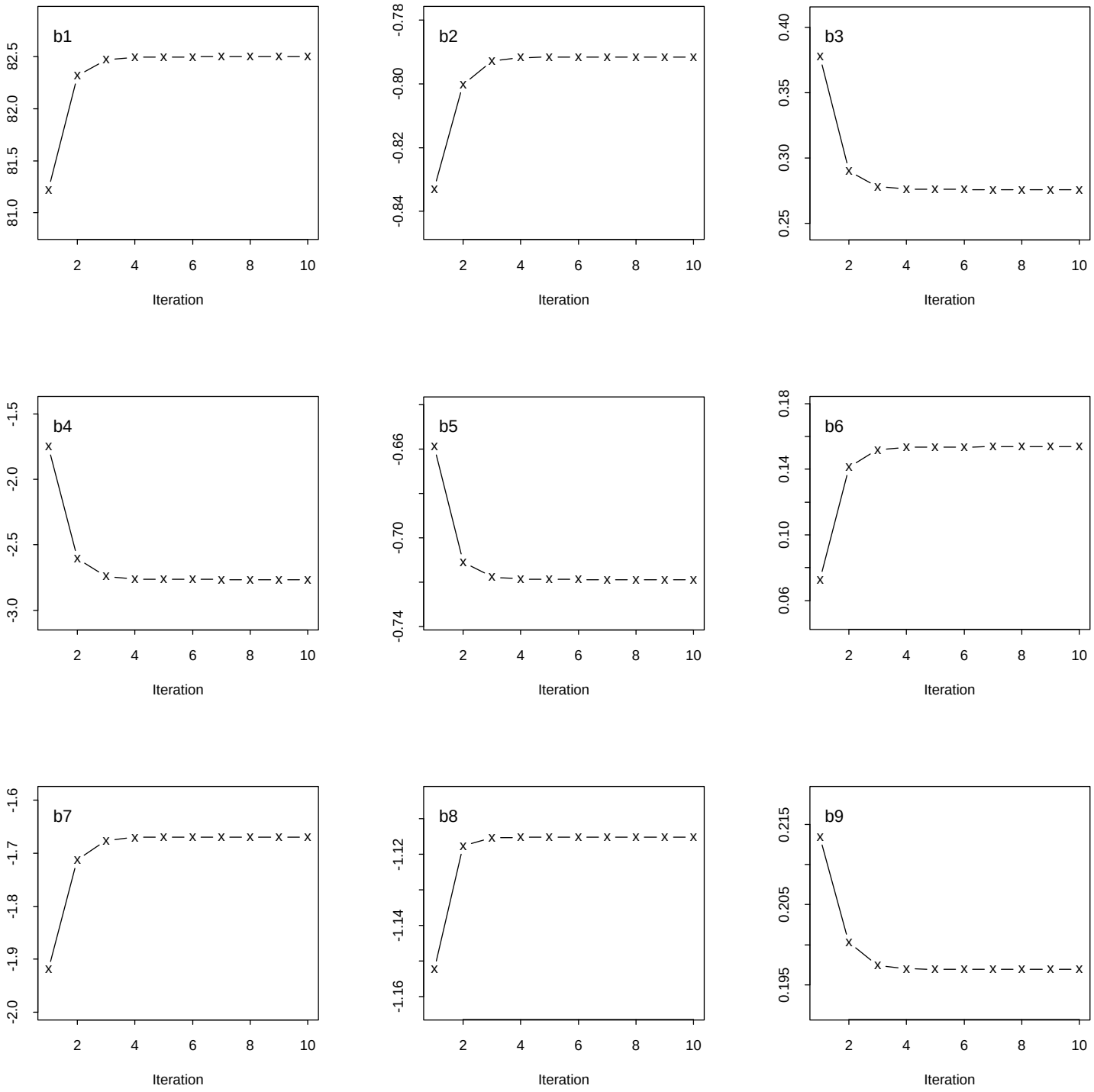


Figure 11: Iteratively reweighted least squares for PANSS data with quadratic trends

Source	SS	df	EMS
Groups	$BGSS$	$g-1$	$\sigma^2\{1+(n-1)\rho\}+q_1$
Error 1	RSS_1	$m-g$	$\sigma^2\{1+(n-1)\rho\}$
Times	$BTSS$	$n-1$	$\sigma^2(1-\rho)+q_2$
Groups $\times$ times	$GTSS$	$(n-1)(g-1)$	$\sigma^2(1-\rho)+q_3$
Error 2	RSS_2	$(m-g)(n-1)$	$\sigma^2(1-\rho)$
Total	TSS	$mn-1$	

Table 1: Split plots ANOVA

For the majority of this section we will assume that we have a balanced structure, there are no covariates, and the variance matrix has compound symmetry. Changing notation slightly we assume g treatment groups, m_h subjects in group h, and n (common) measurement times. Response j on person i in group h will be denoted Y_{hij} and our assumptions are

$$E[Y_{hij}]=\mu_{hj} \qquad \mathrm{Var}(Y_{hij})=\sigma^2 \qquad \mathrm{Cov}(Y_{hij},Y_{hik})=\rho\sigma^2 \qquad (j\neq k)$$

together with between-subject independence. An ANOVA table can be set up with two sections. The first section is used to test for differences between groups in mean response, is simply based on the subject totals, and does not rely on compound symmetry for validity. The second section tests for differences between time points in mean levels, and for treatment by time interaction, ie whether the profiles for treatments are parallel. The second section uses a different error term than the first.

The split plots ANOVA can be drawn up as in Table 1 (with EMS indicating expected mean square and $m=\sum m_h$).

In this table q_1, q_2 and q_3 are non-negative functions which are zero when the appropriate hypothesis is true. Using + notation to indicate totalling over the appropriate subscript and . to indicate averaging, formulae for the sums of squares are

$$BGSS=\sum_h\left(\frac{Y_{h++}^2}{m_h n}\right)-\frac{Y_{+++}^2}{N}=n\sum_h m_h(Y_{h..}-Y_{...})^2$$

$$RSS_1=\sum_h\sum_i\left(\frac{Y_{hi+}^2}{n}\right)-\sum_h\left(\frac{Y_{h++}^2}{m_h n}\right)=n\sum_h\sum_i(Y_{hi.}-Y_{h..})^2$$

$$BTSS=\sum_j\left(\frac{Y_{++j}^2}{m}\right)-\frac{Y_{+++}^2}{N}=m\sum_j(Y_{..j}-Y_{...})^2$$

$$GTSS=\sum_h\sum_j\left(\frac{Y_{h+j}^2}{m_h}\right)-\sum_h\left(\frac{Y_{h++}^2}{m_h n}\right)-\sum_j\left(\frac{Y_{++j}^2}{m}\right)+\frac{Y_{+++}^2}{N}=\sum_h\sum_j m_h(Y_{h.j}-Y_{h..}-Y_{..j}+Y_{...})^2$$

$$\begin{aligned}RSS_2 &= \sum_h\sum_i\sum_j Y_{hij}^2-\sum_h\sum_j\left(\frac{Y_{h+j}^2}{m_h}\right)-\sum_h\sum_i\left(\frac{Y_{hi+}^2}{n}\right)+\sum_h\left(\frac{Y_{h++}^2}{m_h n}\right)\\ &= \sum_h\sum_i\sum_j(Y_{hij}-Y_{h.j}-Y_{hi.}+Y_{h..})^2\end{aligned}$$

$$TSS = \sum_h \sum_i \sum_j Y_{hij}^2 - \frac{Y_{+++}^2}{N} = \sum_h \sum_i \sum_j (Y_{hij} - Y_{...})^2$$

(Don't try to remember these!)

The lower test statistics in the ANOVA only have the nominal F distributions if the within-subject correlation matrices are spherical. If not, the tests are too liberal and lead to false rejection of the null hypotheses too often. Thus non-significant results can always be believed, but borderline significant ones should be questioned. We can do this by adjusting downwards the degrees of freedom: it turns out that the F-tests are approximately valid if all the degrees of freedom are multiplied by a factor ϵ, where $(n-1)^{-1} \leq \epsilon \leq 1$ is an adjustment which depends upon the correlation matrix, and $\epsilon = 1$ if this is spherical. A conservative approach sets ϵ to the minimum value: if the result stays significant we can be confident it is genuine. Otherwise we can estimate ϵ using

$$\hat{\epsilon} = \frac{n^2(\bar{v}_{jj} - \bar{v}_{..})^2}{(n-1)(\sum_j \sum_k v_{jk}^2 - 2n \sum_j \bar{v}_{j.}^2 + n^2 \bar{v}_{..}^2)}$$

where

v_{jk} is element jk of the estimated common covariance matrix V_0 (obtained from residuals),

$\bar{v}_{jj}$ is the mean diagonal element,

$\bar{v}_{j.}$ is the mean of row j,

$\bar{v}_{..}$ is the overall mean.

Of course we should remember not to over-interpret borderline results.

Example. The ANOVA table for the PANSS data is

Source	SS	df	MS	F	p-val	adj-p
Groups	2700.8	2	1350.4	0.803	0.449	
Error 1	447131.0	266	1680.9			
Times	64377.8	5	12875.6	97.118	0.000	0.000
Groups*times	3037.0	10	303.7	2.291	0.011	0.033
Error 2	176327.2	1330	132.6			
Total	693573.9	1613				

The estimated variance matrix is

Week	0	1	2	4	6	8
0	354.0	221.2	204.7	185.8	172.0	151.7
1	221.2	341.3	285.3	258.3	242.3	226.0
2	204.7	285.3	367.0	304.1	286.1	260.9
4	185.8	258.3	304.1	379.1	337.5	317.6
6	172.0	242.3	286.1	337.5	421.2	374.4
8	151.7	226.0	260.9	317.6	374.4	455.1

from which the adjustment factor is $\hat{\epsilon} = 0.62$. Comments?

□

4 Random effects models

The parameters we have considered so far have *population average* interpretations - the overall mean value of a response at time t for instance - with an assumption of independence between subjects. Now suppose (hypothetically) that we could obtain two sets of measurements, under identical circumstances, from the same subject. Would these behave like observations from two separate subjects? Unlikely, because in practice one of the sources of variability is *heterogeneity* between subjects. *Random effects models* try to allow for this property.

The basic assumption is that at least some parameters take different values on different subjects. Differences from population averages can be assumed to be unobservable random variables, say U_i for subject i, and usually we are interested in the probability distribution of the U_i rather than the specific values taken for the subjects in our study, so that we can generalise our conclusions. The most common strategy is to assume a marginal distribution $g(u)$ for U, a conditional distribution $f(y|u)$ for the observable data given the random effects, and then attempt to draw inferences about population parameters from the marginal distribution of y

$$f(y) = \int_u f(y|u)g(u)du.$$

Note that $f(y)$ will depend on parameters from both $f(y|u)$ and $g(u)$.

4.1 Normal linear models

We will assume that at least some of the regression coefficients vary between subjects and that the deviations are zero-mean Normal variables. We can write this as a *linear mixed model*

$$Y_i = X_i\beta + Z_iU_i + \epsilon_i$$

where

X_i is an $n_i \times p$ matrix of covariates (possibly including time t)

β is a p-vector of parameters

Z_i is a $n_i \times q$ matrix of covariates (again possibly including time t)

U_i is a q-vector of unobservable subject-specific random effects, assumed to be $\mathrm{N}(0, \Sigma)$

$\epsilon_i \sim \mathrm{N}(0, V_i)$ is an n_i-vector of errors. Often we take $V_i = \sigma^2 I$ to represent pure measurement error.

Note that U_i varies between subjects but is fixed for each subject, so that if it were possible to obtain further observations for subject i then the *same* U_i would apply, but we would get a *new* ϵ_i. In Crowder and Hand's words, β is an immutable constant of the universe, U_i is a lasting characteristic of the individual, and ϵ_i is but a fleeting aberration of the moment.

The conditional distribution of $Y_i|U_i$ is $\mathrm{N}(X_i\beta + Z_iU_i, V_i)$ and the marginal distribution is $\mathrm{N}(X_i\beta, Z_i\Sigma Z_i^T + V_i)$. Inference can be based on this marginal distribution, using maximum likelihood for structured covariance matrices.

Example (a). Random intercept only. Take $V_i = \sigma^2 I$ and allow each subject to have their own mean level through

$$Y_i = X_i\beta + U_i 1_{n_i} + \epsilon_i$$

with $U_i \sim \mathrm{N}(0, \sigma_1^2)$ a scalar random variable and 1_{n_i} an n_i-vector of ones. This induces the compound symmetry correlation structure since

$$\mathrm{Var}(Y_{ij}) = \sigma_1^2 + \sigma^2 \qquad \mathrm{Cov}(Y_{ij}, Y_{ik}) = \sigma_1^2 \qquad \mathrm{Corr}(Y_{ij}, Y_{ik}) = \frac{\sigma_1^2}{\sigma_1^2 + \sigma^2} \qquad (j \neq k).$$

Example (b). Random intercept and random slope. This assumes one of the covariates in X_i is time, to allow a general linear trend. Subject-specific trends are modelled via two random effects, again taking $V_i = \sigma^2 I$:

$$Y_i = X_i\beta + U_{i1} 1_{n_i} + U_{i2} t_i + \epsilon_i$$

with $U_{i1} \sim \mathrm{N}(0, \sigma_1^2)$, $U_{i2} \sim \mathrm{N}(0, \sigma_2^2)$, $\mathrm{Cov}(U_{i1}, U_{i2}) = \sigma_{12}$ and $t_i = (t_{i1}, t_{i2}, \ldots, t_{in_i})^T$. So

$$\mathrm{Var}(Y_{ij}) = \sigma_1^2 + \sigma_2^2 t_{ij}^2 + 2\sigma_{12} t_{ij} + \sigma^2 \qquad \mathrm{Cov}(Y_{ij}, Y_{ik}) = \sigma_1^2 + \sigma_2^2 t_{ij} t_{ik} + \sigma_{12}(t_{ij} + t_{ik}).$$

Note that the variances and covariances change with time. Using the Splus routine "lme" to fit this model to the PANSS data with quadratic underlying trends produced the following output summary (using x1 to denote time)

```
   Standard Deviation(s) of Random Effect(s)
 (Intercept)       x1
    16.13578 2.168015
 Correlation of Random Effects
   (Intercept)
x1 -0.2161575

 Cluster Residual Variance: 91.10771
Fixed Effects Estimate(s):
  (Intercept)         x1       x2        x3         x4         x5        x6
    81.22395 -0.8331269 0.377485 -1.752372 -0.6589619 0.07282809 -1.919499

        x7        x8
 -1.152382 0.2134208
```

The variance matrix can be constructed from these estimates for comparison with the unstructured version obtained earlier

Week	0	1	2	4	6	8
0	351.5	252.8	245.3	230.1	215.0	199.9
1	252.8	341.1	247.1	241.4	235.7	230.0
2	245.3	247.1	340.1	252.6	256.3	260.0
4	230.1	241.4	252.6	366.2	297.6	320.1
6	215.0	235.7	256.3	297.6	430.0	380.2
8	199.9	230.0	260.0	320.1	380.2	531.4

Comments?

□

Notes

1. These examples are cases of so-called Laird-Ware models.

2. An extension adds a serially correlated term to give a Diggle-Laird-Ware model

$$Y_{ij} = x_{ij}^T \beta + U_{i1} + U_{i2} t_{ij} + S_i(t_{ij}) + \epsilon_{ij}$$

   where $S_i(t_{ij}) \sim \mathrm{N}(0, \sigma_s^2)$, $\mathrm{Cov}\{S_i(t_{ij}), S_i(t_{ik})\} = \rho^{|t_{ij} - t_{ik}|}$ and the ϵ_{ij} are mutually independent.

3. The likelihood can be used to compare models in an informal way. Formal likelihood ratio tests are usually invalid here because we have *non-regular* problems. Consider a comparison between the first two rows for instance. Testing for significant improvement in log-likelihood is equivalent to testing $H_0 : \sigma_1^2 = 0$. But as variances can never be negative the null hypothesis is that σ_1^2 is on the boundary of its parameter space, a condition excluded from regular asymptotic likelihood theory.

   When using log-likelihoods to compare models informally, *information criteria* can be used to adjust for the number of parameters. Probabably the best known is the Akaike information criterion

$$\mathrm{AIC} = -2l + p$$

   where l is the log-likelihood and p is the number of free parameters. (Sometimes this is divided by two, and sometimes there is a constant subtracted). Small values are preferred.

   Continuing the PANSS example:

Variance terms	log-likelihood	AIC
σ^2	-5620.5	11251.0
σ_1^2, σ^2	-5102.7	10216.4
σ_2^2, σ_1^2, σ^2	-5000.6	10013.2
σ_{12}, σ_2^2, σ_1^2, σ^2	-4996.6	10006.2
Unstructured (21 terms)	-4917.4	9864.8

   Comments?

4. If required, a best linear unbiased predictor (BLUP) of the random effects U_i can be obtained by treating U_i as a fixed parameter, or an empirical Bayes estimate can be obtained as the mean of a posterior distribution: see exercises.

5. These models can be applied without a Normality assumption for Y, using generalised least squares for estimation.

6. *Multilevel* extensions are sometimes used when we have nested clusters. For instance in education there can be child, school and district random effects.

4.2 Binary data

Now suppose each Y_{ij} is binary. One simple approach to describe heterogeneity is to assume constant within-subject probabilities

$$P(Y_{ij} = 1|U_i) = U_i$$

together with conditional independence and a beta distribution for between-subject differences

$$U_i \sim g(.) \qquad g(u) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} u^{a-1}(1-u)^{b-1} \qquad (0 < u < 1)$$

with parameters $a, b > 0$. The marginal properties are then

$$P(Y_{ij} = 1) = a/(a+b) \qquad E[Y_{ij}] = a/(a+b) \qquad \mathrm{Var}(Y_{ij}) = ab/(a+b)^2,$$

the covariances are

$$\mathrm{Cov}(Y_{ij}, Y_{ik}) = \frac{ab}{(a+b)^2(a+b+1)}$$

and the likelihood contribution is

$$\begin{aligned} P(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, \ldots, Y_{in_i} = y_{in_i}) &= E[U^{y_{i+}}(1-U)^{n_i - y_{i+}}] \\ &= \frac{\left(\prod_{j=0}^{y_{i+}-1}(a+j)\right)\left(\prod_{j=0}^{n_i-y_{i+}-1}(b+j)\right)}{\prod_{j=0}^{n_i-1}(a+b+j)} \end{aligned}$$

where we use the convention that $\prod_{j=0}^{-1} = 1$ if necessary. (Exercise: prove these results, but don't try to remember them).

Other approaches to allow covariate (including time) effects are possible: see Diggle, Liang and Zeger.

4.3 Count data

The standard model for count data is Poisson, under which the mean and variance should be equal. Heterogeneity between subjects leads to *extra-Poisson* variability. For the PCA data for instance:

```
2mg mean  9.3  5.5  5.4  5.1  7.6  5.3  3.9  3.7  4.6   4.9   3.5   3.5
    var  33.5 22.1 16.0 17.0 25.1 11.3 17.6  9.2 10.0  13.2   4.3   8.0

1mg mean 10.2  6.5  7.9  9.3  9.6  7.4  6.6  5.7  4.9   6.3   6.3   5.9
    var  59.2 31.1 60.0 50.7 43.1 30.7 25.0 22.8 16.1  30.3  29.7  33.8
```

A random effects interpretation can be useful for modelling longitudinal data with these characteristics. We assume that the counts Y_{ij} are conditionally independent Poisson given the value of a gamma distributed random effect. If covariates are to be included or the baseline mean allowed to change with time the gamma distribution should be scaled to have mean one for identifiabilty. This leaves one parameter to measure the heterogeneity effect, conveniently measured as the gamma variance, ξ say. So the assumptions are

$$Y_{ij}|U_i \sim \mathrm{P}(U_i \alpha_j e^{\beta x_i})$$

(with conditional independence between times) and

$$U_i \sim \Gamma(1/\xi, 1/\xi) \qquad E[U_i] = 1 \qquad \mathrm{Var}(U_i) = \xi$$

Writing $\lambda_{ij} = \alpha_j e^{\beta x_i}$, the counts have negative binomial marginals

$$P(Y_{ij} = y_{ij}) = \frac{\lambda_{ij}^{y_{ij}}}{y_{ij}!(1+\xi\lambda_{ij})^{(y_{ij}+1/\xi)}} \prod_{k=0}^{y_{ij}-1}(1+k\xi)$$

with

$$E[Y_{ij}] = \lambda_{ij} \qquad \mathrm{Var}(Y_{ij}) = \xi\lambda_{ij}^2 + \lambda_{ij} \qquad \mathrm{Corr}(Y_{ij}, Y_{ik}) = \frac{1}{\sqrt{\{1+1/(\xi\lambda_{ij})\}\{1+1/(\xi\lambda_{ik})\}}}$$

and likelihood contribution

$$P(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, \ldots, Y_{in_i} = y_{in_i}) = \left(\prod_{j=0}^{n_i} \frac{\lambda_{ij}^{y_{ij}}}{y_{ij}!}\right)\left(\prod_{j=0}^{n_i-1}(1+j\xi)\right)\left(\frac{1}{1+\xi\sum_j \lambda_{ij}}\right)^{y_{i+}+1/\xi}$$

Example. Maximum likelihood estimates for the PCA data:

	Est	SE	T
α_1	8.09	0.80	10.09
α_2	4.99	0.59	9.62
α_3	5.58	0.57	9.74
α_4	6.10	0.62	9.83
α_5	7.18	0.72	9.98
α_6	5.32	0.55	9.69
α_7	4.45	0.47	9.48
α_8	3.99	0.43	9.35
α_9	3.95	0.42	9.33
α_{10}	4.71	0.49	9.55
α_{11}	4.13	0.44	9.39
α_{12}	3.95	0.42	9.33
β	0.33	0.12	2.66
ξ	0.24	0.04	5.59

Apart from ξ, these are comparable with those obtained earlier under an independence working assumption. Why is the estimate of ξ smaller?

□

5 Generalised estimating equations

There are at least two difficulties with the random effects approach to longitudinal data analysis. First, the assumptions in the previous sections were made at least in part in order that a closed form expression is available for $f(y)$ after integrating out the random effects, so that full likelihood inference based on the observable y is tractable. Usually however we will not be able to write down an explicit expression for $f(y)$, which means that approximate numerical methods will be needed. The second difficulty is the necessity to make assumptions about the distribution of the unobservable random effects. How can we be confident in these assumptions when there are no data to provide direct support?

Liang and Zeger proposed a *generalised estimating equation* approach to overcome these difficulties. Inference is based on the marginal distribution $f(y)$ only, but without writing down an explicit form for it. Instead we make assumptions only about the *mean* $\mu_i = E[Y_i]$ and the variance $V_i = \mathrm{Var}(Y_i)$. We assume that the mean depends on parameters β and the variance may depend on β and additional parameters α. The motivation comes from the treatment of the Normal linear model $Y_i \sim \mathrm{N}(X_i\beta, V_i)$ where for given V_i the maximum likelihood estimator of β solves

$$\sum_i X_i^T V_i^{-1}(Y_i - X_i^T\beta) = 0.$$

Noting $\mu_i = X_i\beta$ this is equivalent to

$$S_\beta(\beta, \alpha) = \sum_i \left(\frac{\partial \mu_i}{\partial \beta}\right)^T V_i^{-1}(Y_i - \mu_i) = 0.$$

Provided μ_i is the correctly specified function of β this equation will lead to asymptotically unbiased estimates of β even without the assumption of a Normal linear model. Moreover we can replace V_i^{-1} by any weight matrix and have the same result, though a less efficient estimator. Similarly if we define W_i to be a vector of all second-order products, ie

$$W_i = (Y_{i1}^2, Y_{i1}Y_{i2}, \ldots, Y_{in_i}^2)^T$$

with expectation η_i depending on α and β, we get asymptotically unbiased estimates by solving

$$S_\alpha(\beta, \alpha) = \sum_i \left(\frac{\partial \eta_i}{\partial \alpha}\right)^T H_i^{-1}(W_i - \eta_i) = 0.$$

for any weight matrix H_i.

Simultaneously solving $S_\beta(\beta, \alpha) = 0$ and $S_\alpha(\beta, \alpha) = 0$ thus provides the generalised estimating equation approach to estimation relying only on specification of the first two moments of Y_i, ie μ_i and η_i.

Notes:

1. The optimal choice of weight H_i requires knowledge of $\mathrm{Var}(W_i)$ and thus third and fourth order moments. Hence sub-optimal selections are usually made, to trade simplicity for efficiency. Exact choice depends upon the form of data - continuous, binary, count etc.

2. The solution $(\hat{\alpha}, \hat{\beta})$ is asymptotically Normal, with mean (α, β) and variance which can be estimated through an extended form of sandwich estimator - details omitted.

Example

The following output shows use of the function `gee` (from the R library of the same name) to fit a model of the form

$$\log\{E[Y_{ij}]\} = x'_{ij}\beta \qquad \mathrm{Var}(Y_{ij}) = \phi E[Y_{ij}]$$

to the PCA data. The design matrix X is set up to allow a different mean for each measurement time but a common treatment effect. Use of the `family=poisson` term implicitly tells the program to form the mean and variance structure above. Other families include 'gaussian', 'binomial', 'Gamma', and 'quasi': use `help(gee,package="gee")` for further information.

Variables (in column form) in the data frame `pca` are

`y`	Response counts
`tmt`	Indicator for treatment group
`time.1, time.2, ...`	Indicators for measurement times
`id`	Patient identifier

```
>    gfit_gee(y~-1+time.1+time.2+time.3+time.4+time.5+time.6+time.7+time.8+
 + time.9+time.10+time.11+time.12+tmt,id=id,data=pca,family=poisson)

[1] "Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27"
[1] "running glm to get initial regression estimate"
 [1] 2.0903256 1.6079636 1.7189204 1.8084016 1.9717548 1.6721820 1.4917731
 [8] 1.3829042 1.3732731 1.5502038 1.4174443 1.3732781 0.3265408
>
>  names(gfit)
 [1] "title"               "version"             "model"
 [4] "call"                "terms"               "nobs"
 [7] "iterations"          "coefficients"        "nas"
[10] "linear.predictors"   "fitted.values"       "residuals"
[13] "family"              "y"                   "id"
[16] "max.id"              "working.correlation" "scale"
[19] "robust.variance"     "naive.variance"      "xnames"
[22] "error"
>
>  rbse_diag(gfit$robust.variance)^.5
>  nvse_diag(gfit$naive.variance)^.5
>  outsum_cbind(gfit$coefficients,rbse,nvse)
>  print(round(outsum,3))
```

```
                rbse  nvse
time.1  2.090 0.110 0.087
time.2  1.608 0.130 0.107
time.3  1.719 0.124 0.102
time.4  1.808 0.112 0.098
time.5  1.972 0.106 0.091
time.6  1.672 0.103 0.104
time.7  1.492 0.130 0.113
time.8  1.383 0.123 0.118
time.9  1.373 0.116 0.119
time.10 1.550 0.116 0.110
time.11 1.417 0.103 0.117
time.12 1.373 0.123 0.119
tmt     0.327 0.129 0.059
> print(round(gfit$scale,3))
[1] 3.973
>
```

Comments?

6 Dealing with dropout

Usually in longitudinal trials there are protocols which states that a sequence of measurements should be taken on each subject, usually at pre-specified times although in practice there can be some variation, just as for the liver data. Minor variation in timing of measurements is unlikely to lead to any problems, but potentially more serious is the occurrence of missing data, where one or more scheduled observations are simply not obtained at all. *Intermittent missingness* occurs when occasional observations are not recorded. Potentially more serious for longitudinal data analysis is *dropout*, where a patient withdraws (or is withdrawn) early. This is the focus of the current section.

Ignoring dropout can lead to biased results. As a simple example, if dropout is a result of an inadequate response to treatment, then the observed mean response amongst subjects who complete the study will give an over-optimistic estimate of the mean response for the target population.

To simplify notation we assume a balanced design with n scheduled measurements per subject, at common times $t_1, t_2, \ldots, t_n$. Dropping the subject subscript, we introduce a scalar random variable D to represent the time at which a subject drops out. Then, a subject will provide a complete set of data $Y = (Y_1, ..., Y_n)$ if and only if $D > t_n$. Otherwise, the subject will provide an incomplete set of observations $Y_{obs} = (Y_1, ..., Y_d)$ where $d < n$ is such that $t_d < D < t_{d+1}$. In this situation, let $Y_{miss} = (Y_{d+1}, ..., Y_n)$ denote the missing data from the subject in question. A model for the data must now specify the joint distribution of Y and D, which we shall write as $[Y_{obs}, Y_{miss}, D]$ (the square brackets to be read as "the distribution of"). The likelihood is then

$$L(\theta) \quad = \quad \int [Y_{obs}, Y_{miss}, D] dY_{miss}$$

$$= \int [Y_{obs}, Y_{miss}][D|Y_{obs}, Y_{miss}]dY_{miss}. \quad (5)$$

Following Rubin's missing data terminology we can distinguish three cases:

Completely random dropout : the conditional distribution of dropout time depends neither on Y_{obs} nor on Y_{miss}

Random dropout : the conditional distribution of dropout time may depend on Y_{obs}, but does not depend on Y_{miss}

Informative dropout : the conditional distribution of dropout time depends on Y_{miss}

Suppose that dropout is completely random. Then, in (5) we have that $[D|Y_{obs}, Y_{miss}] = [D]$ and it follows that

$$L(\theta) = [D] \int [Y_{obs}, Y_{miss}]dY_{miss} = [D][Y_{obs}]$$

Similarly, if dropout is random, then $[D|Y_{obs}, Y_{miss}] = [D|Y_{obs}]$ and (5) reduces to

$$L(\theta) = [D|Y_{obs}] \int [Y_{obs}, Y_{miss}]dY_{miss} = [D|Y_{obs}][Y_{obs}]$$

In either case, the likelihood factorises into two terms, one for the contribution from the observed values of Y_{obs}, and one for the contribution from the observed value of D conditional on Y_{obs} (the conditioning being irrelevant in the case of completely random dropout). Hence, provided these two components of the likelihood are separately parameterised, a likelihood-based analysis requires only *separate* likelihood-based analyses of the data Y_{obs}, and of the data D treating Y_{obs} as covariate information. Furthermore, if the questions of interest concern only the parameters of the measurement process Y, that part of the overall analysis which involves D contributes nothing of interest, and can be ignored. Indeed, the case of completely random or random dropout with separate parameterisation is usually given the name *ignorable dropout.*

Example. We use simulated data to illustrate the effect of dropout. Measurements are scheduled at times $j = 1, 2, ..., 10$ and are drawn from the model

$$Y_{ij} = 10 + U_i + \epsilon_{ij}$$

where $U_i \sim \mathrm{N}(0, \sigma_1^2)$ are independent between subjects, and $\epsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$ are independent between times within subjects. Note that $\mathrm{Corr}(Y_{ij}, Y_{ik}) = \sigma_1^2/(\sigma_1^2 + \sigma^2) = \rho$ say. At each time $j \geq 2$, any subject who has not dropped out at or before time $j-1$ will drop out at time j with probability $p(Y_{i,j-1})$, where

$$\log\left(\frac{p(y)}{1-p(y)}\right) = \alpha + \beta y$$

Simulations were made with $\sigma_1^2 + \sigma^2 = 2$, $\alpha = -17$ and $\beta = 1.4$ so that the marginal distribution of each measurement is $\mathrm{N}(10, 2)$ and the dropout probabilities when the most recently observed measurement is at the lower and upper quartiles of this distribution are 0.013 and 0.158, respectively. The two panels of Figure 12 each show the results of a simulation in the following format: the solid line shows the observed mean at time j amongst all subjects who have not yet dropped out; the dashed lines show, for $k = 2, .., 10$, the observed means at times 1 to k amongst subjects who drop out at time $k+1$. The horizontal dotted line shows the marginal population mean of each Y_{ij}.

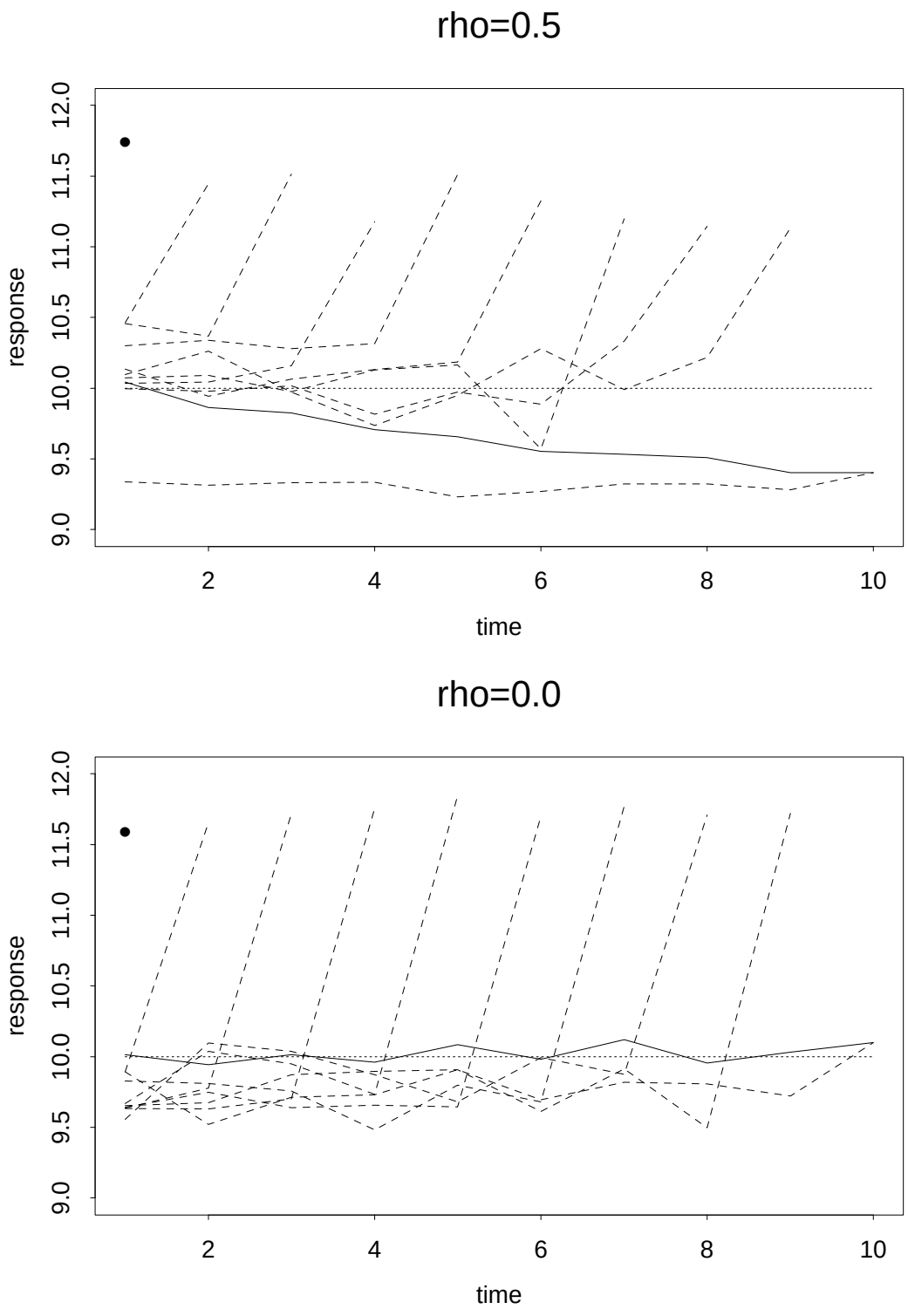


Figure 12: Dropout effects

In the upper panel $\sigma_1^2 = 1$, hence $\rho = 0.5$, and the solid line shows the selective effect of the dropout process. In the lower panel $\sigma_1^2 = 0$ so $\rho = 0$. Comments?

□.

In order to allow for dropout in analysis we need to specify a joint distribution $[Y, D]$. This can be factorised as either $[Y|D][D]$ or $[D|Y][Y]$, which are mathematically equivalent but in practice lead to different simplifying assumptions and therefore to different models and analyses of the data. The factorisation $[D|Y][Y]$ is a *selection model* for dropout, the connotation being that dropouts are "selected" on the basis of their evolving sequence of measurements over time. The factorisation $[Y|D][D]$ is a *pattern mixture model*, the understanding now being that the population is a mixture of sub-populations identified by their differing *a priori* propensities to dropout, and that potentially, subjects with different propensities to dropout may also differ with regard to their measurement profiles over time.

Random effects U may also be included in the model. In many ways a conditional independence assumption and factorisation

$$[Y, D] = \int [Y|U][D|U][U]dU$$

is the most natural, in that we make the biologically plausible assertion that both Y and D are affected by some internal unobservable characteristic U of the subject. Methods for fitting this type of model are becoming available but are computationally intensive and not yet incorporated in commercial software.

Example. The PANSS data considered throughout these notes is a subset only of the results of the trial, consisting of data on the 269 patients who completed the 8 weeks of treatment. Initially there were 518 patients (88 placebo, 85 haloperidol and 345 risperidone) but 249 dropped out before week 8. The upper part of Figure 13 shows the mean PANSS score for each treatment group against time of measurement, using all available data at each time point. The lower part gives survival curves for time to dropout for the three groups. The fall in mean for each group suggests, superficially, that each treatment is beneficial because the average score improves over time. Such a conclusion is not as yet justified however, because of the possible selective effect of dropout: if patients with the worst health (and so highest scores) are more likely to dropout, then the mean score amongst those remaining under study would fall over time even in the absence of any treatment effect. An apparent linkage between dropout and response is evident in Figure 14, the mean PANSS scores for each dropout cohort.

For completeness we present results of a joint analysis of Y and D for these data. We assumed a random intercept plus stochastic process model

$$Y_{ij} = x_{ij}^T\beta + U_{i1} + S_i(t_{ij}) + \epsilon_{ij}$$

with a different mean for each treatment × time combination, $U_{i1} \sim \mathrm{N}(0, \sigma_1^2)$, $S_i(t_{ij}) \sim \mathrm{N}(0, \sigma_s^2)$, $\mathrm{Cov}\{S_i(t_{ij}), S_i(t_{ik})\} = \rho^{|t_{ij}-t_{ik}|}$ and the ϵ_{ij} mutually independent. A proportional hazards model

$$\alpha_i(t) = \alpha_0(t)\exp\{x_i(t)^T\gamma + W_i^*(t)\}$$

was taken for the dropout mechanism, where $\alpha_0(t)$ is unspecified and $W_i^*(t)$ is allowed to depend on U_{1i} and $S_i(t)$. Estimation was by EM and results included:

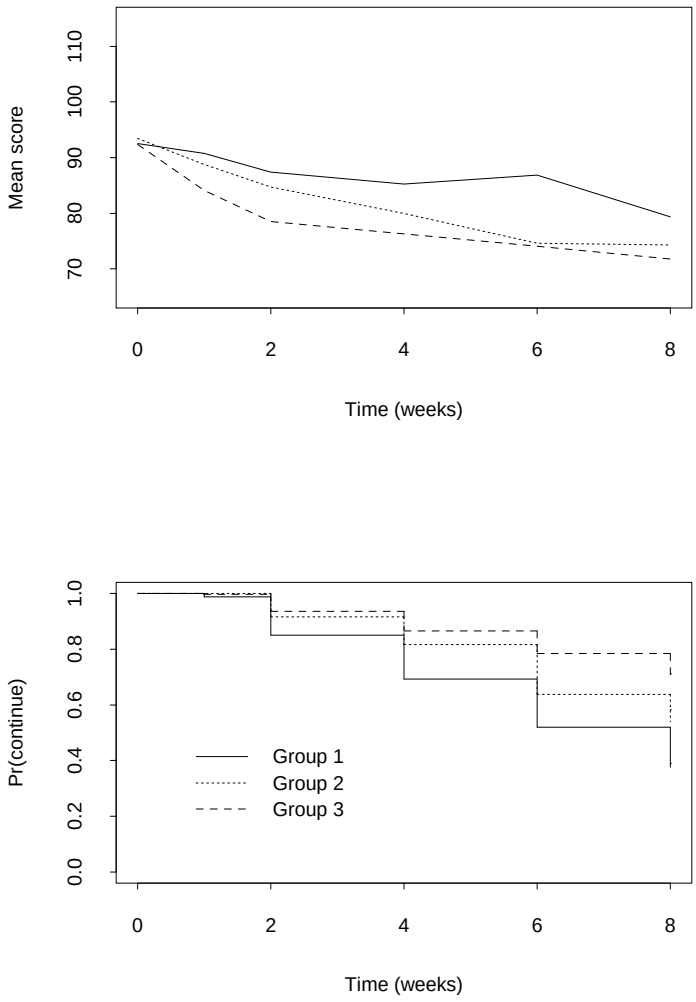


Figure 13: PANSS observed means within each treatment group (upper plot) and survival curves for time to dropout (lower plot)

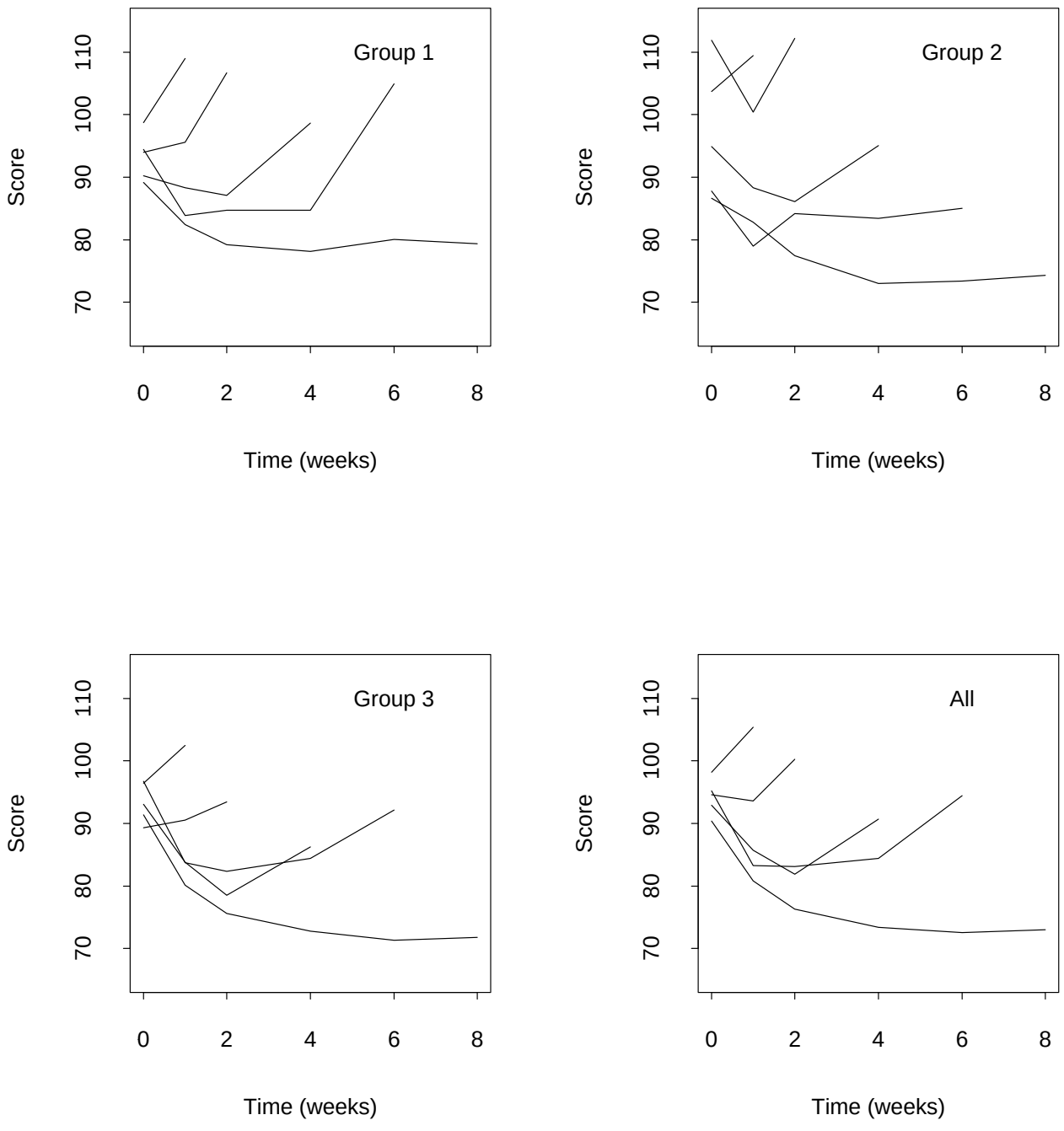


Figure 14: PANSS mean profiles by dropout cohort

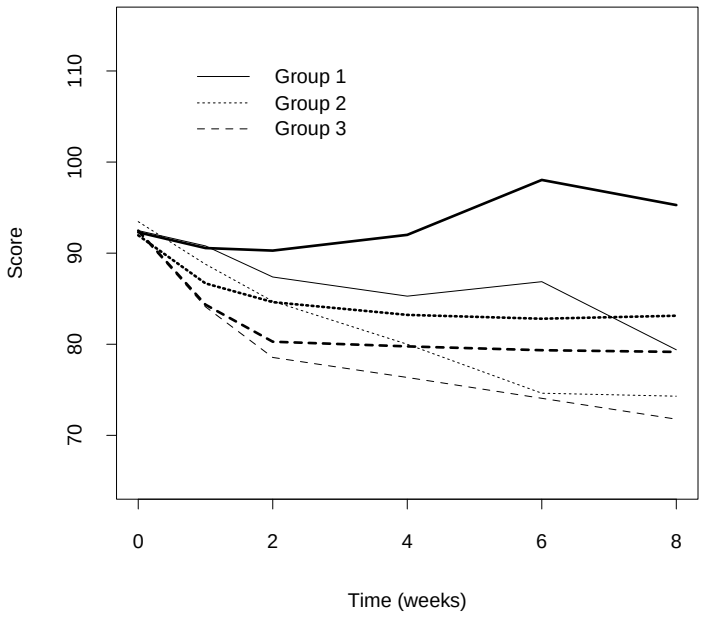


Figure 15: PANSS observed means and estimated droput-free profiles

		log-likelihood		
		Measurements	Dropout	Combined
Model	Estimate	$[Y]$	$[D\|Y]$	$[Y, D]$
Completely random	$W_i^*(t) = 0$	-10126.7	-602.5	-10729.2
Proportional	$W_i^*(t) = 0.016(U_i + S_i(t))$	-10131.5	-584.3	-10715.8
Separate	$W_i^*(t) = 0.063U_i + 0.006S_i(t)$	-10131.5	-568.7	-10700.2

Fitted profiles are in Figure 15. Comments?

□.

Final remark: of course the best way to deal with dropouts is not to have them in the first place.

7 Appendices

Appendix A. PCA data

Group 1: 2mg bolus

Patient	Z_1	Z_2	Z_3	Z_3	Z_5	Z_6	Z_7	Z_8	Z_9	Z_{10}	Z_{11}	Z_{12}
1	5	2	2	5	2	4	0	2	4	4	4	2
2	5	3	5	4	4	5	0	6	3	2	7	4
3	6	4	8	4	3	12	1	0	6	5	3	5
4	4	1	4	3	3	3	1	7	5	0	1	1
5	8	7	7	6	8	6	5	4	7	2	5	0
6	10	6	5	4	8	7	4	6	5	2	4	4
7	2	4	7	4	8	4	4	3	1	3	4	6
8	14	2	10	8	17	4	2	6	8	9	1	4
9	7	10	9	8	9	4	2	6	1	8	5	5
10	3	0	7	1	9	8	0	1	5	4	1	0
11	6	2	2	3	5	4	2	1	3	4	3	3
12	10	14	2	2	4	2	8	5	1	1	5	10
13	23	6	8	7	7	4	1	4	2	3	2	2
14	15	12	8	9	12	7	2	4	2	2	2	3
15	11	6	8	8	0	3	3	2	1	2	2	4
16	9	0	3	2	9	6	3	4	4	5	4	0
17	4	7	2	0	0	3	0	1	1	3	0	0
18	3	1	6	6	6	2	5	4	5	5	5	4
19	16	5	2	5	6	5	8	2	9	3	2	4
20	6	1	3	6	6	3	0	0	2	6	3	1
21	13	6	7	10	5	4	3	2	4	5	2	3
22	12	11	15	16	17	10	19	14	13	16	7	11
23	6	2	2	0	8	5	3	4	7	5	6	7
24	8	0	0	1	16	8	7	3	7	5	6	5
25	9	6	3	5	14	11	8	5	7	7	1	1
26	22	9	0	0	0	0	3	0	1	0	1	0
27	8	7	6	1	14	6	2	7	11	11	4	5
28	6	4	2	4	3	0	2	0	1	5	3	0
29	5	6	3	5	9	5	7	1	6	12	4	6
30	23	21	17	17	15	15	13	8	6	9	8	4

Group 2: 1mg bolus

Patient	Z_1	Z_2	Z_3	Z_3	Z_5	Z_6	Z_7	Z_8	Z_9	Z_{10}	Z_{11}	Z_{12}
31	9	6	3	6	4	1	3	1	3	2	1	0
32	8	4	5	3	4	7	3	3	2	5	1	1
33	9	3	9	2	2	4	3	1	3	2	0	1
34	2	1	2	0	3	4	5	3	0	3	7	3
35	9	13	23	20	21	22	13	16	10	19	8	9
36	15	10	15	16	7	10	15	18	16	16	15	11
37	11	16	2	10	8	0	2	2	0	7	3	1
38	3	3	16	10	12	8	6	5	6	7	6	11
39	18	13	10	12	21	8	8	15	2	4	7	1
40	10	1	0	0	4	3	2	6	5	4	4	0
41	4	3	3	7	12	7	12	5	6	6	13	17
42	27	22	15	24	27	10	4	9	5	9	7	12
43	10	9	15	10	15	8	9	3	4	3	6	6
44	6	3	5	10	7	9	6	7	3	8	2	7
45	15	15	1	0	5	4	0	3	0	1	3	3
46	8	9	6	7	9	4	3	0	8	2	2	0
47	22	8	4	9	25	19	20	15	9	4	22	22
48	1	0	1	6	11	5	4	3	2	3	7	2
49	11	5	8	24	7	16	17	2	15	25	8	6
50	21	12	9	0	2	3	6	5	3	3	2	2
51	1	10	10	13	8	8	15	9	4	9	3	14
52	28	15	35	21	17	10	5	4	3	10	21	11
53	0	1	0	2	4	3	4	10	3	0	0	1
54	3	0	3	5	5	6	1	1	1	3	3	1
55	5	2	7	15	6	0	10	12	9	4	3	16
56	13	6	5	8	20	13	10	9	5	7	10	10
57	3	1	1	14	7	6	2	3	2	5	4	2
58	7	8	24	18	10	9	12	6	6	14	14	5
59	17	1	1	4	11	2	10	1	5	6	7	11
60	3	4	6	3	7	3	4	1	0	2	1	0
61	3	2	5	9	3	4	2	0	0	0	3	4
62	25	0	13	7	9	19	4	5	11	10	6	9
63	12	11	7	6	9	3	4	6	6	6	6	0
64	3	8	4	22	11	16	4	5	9	12	12	7
65	14	2	2	2	4	4	3	7	6	1	2	0

Appendix B. Scatterplot smoothing

Assume we have data pairs (t_i, y_i). Scatterplot smoothers aim to estimate the underlying trend $E[Y(t)] = \mu(t)$ without making parametric assumptions about the shape of $\mu(t)$. Various methods are available, but probably the two most used are kernel smoothers and spline smoothers.

Kernel smoothers estimate $\mu(t)$ at each point t by a weighted average of the y's, with weight on y_i depending on the distance $|t - t_i|$. Various weight functions or *kernels* have been suggested. Some have finite range and so only use y_i if t_i is within a certain distance of t. Others use all the points but decrease the weight with distance. In general if the kernel function is $K(.)$ then

$$\hat{\mu}(t) = \sum_i K\{(t - t_i)/h\} y_i / \sum_i K\{(t - t_i)/h\}$$

where h is the bandwidth. This controls the smoothing, with large h evening out the weights more and so producing smoother curves. Accepted wisdom is that the shape of the kernel function is less important than the choice of bandwidth parameter.

Cubic spline smoothers estimate $\mu(t)$ by minimising

$$\sum_i \{y_i - \mu(t_i)\}^2 + \lambda \int \{\mu''(t)\}^2 dt.$$

The second term is a *roughness penalty*: if $\mu(t)$ varies rapidly then its second derivative will be large and the penalty is high. The parameter λ plays the same role as h, ie controls the degree of smoothing.

Adaptive methods allow the amount of smoothing to depend on the local intensity of data points, to have less smoothing when there are lots of points near t, more when the data are more sparse near t.

Cross-validation is sometimes used to find the smoothing parameter, choosing the value which minimises the mean square error between the data points y_i and their estimates $\hat{\mu}_{(i)}(t_i)$ based on all the data *except* i.

Appendix C. Likelihood revision for robust estimation

Assume a regular problem with iid continuous random variables having density $f(y; \theta_0)$. For simplicity assume θ is scalar (though everything works for vector θ too). Likelihood contributions are $l(\theta) = \log f(y; \theta)$, the score contribution is $u(\theta) = \partial l(\theta)/\partial\theta$ and the information contribution is $i(\theta) = -\partial^2 l(\theta)/\partial\theta^2$. Of course asymptotically the maximum likelihood estimator $\hat{\theta}$ is $\mathrm{N}(\theta_0, 1/\{nE_Y[i(\theta); \theta_0]\})$ (using Taylor series and the central limit theorem). Usually we estimate $E_Y[i(\theta); \theta_0]$ by the sample average of $i(\hat{\theta})$ (ie the observed information divided by n). But

$$E_Y[u(\theta); \theta_0] = 0$$

$$\begin{aligned} \mathrm{Var}_Y\{u(\theta); \theta_0\} &= E_Y[u(\theta)^2; \theta_0] \\ &= E_Y[i(\theta); \theta_0] \end{aligned}$$

(using $(f'/f)^2 = f''/f - \partial^2 \log(f)/\partial\theta^2$). This means we can also estimate $E_Y[i(\theta); \theta_0]$ by the sample average of the $u(\hat{\theta})^2$ terms.

Appendix D. Multivariate revision

Moments

Suppose A is an $n \times m$ matrix of constants, a is an n-vector of constants and Y is an n-vector of univariate random variables $Y_1, Y_2, \ldots, Y_n$. Vectors are always columns.

$E[Y]$ should be taken as the vector of means, ie $E[Y] = (E[Y_1], E[Y_2], \ldots, E[Y_n])^T$. $\mathrm{Var}(Y)$ is a positive semi-definite symmetric matrix with variances down the diagonal and covariances elsewhere, ie

$$\{\mathrm{Var}(Y)\}_{ij} = \begin{cases} \mathrm{Var}(Y_i) & i = j \\ \mathrm{Cov}(Y_i, Y_j) & i \neq j \end{cases}$$

Then

$$E[a^T Y] = a^T E[Y] \qquad \mathrm{Var}(a^T Y) = a^T \mathrm{Var}(Y) a$$

$$E[AY] = AE[Y] \qquad \mathrm{Var}(AY) = A\mathrm{Var}(Y)A^T.$$

Multivariate Normal

If Y is multivariate Normal with mean (vector) μ and variance (matrix) V, its probabilty density is

$$f(y) = \frac{1}{(\sqrt{2\pi})^{n/2}|V|^{1/2}} \exp\{-\frac{1}{2}(y-\mu)^T V^{-1}(y-\mu)\}$$

Partitioning Y as

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \qquad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \qquad V = \begin{pmatrix} V_{11} & V_{12} \\ V_{12}^T & V_{22} \end{pmatrix}$$

then $Y_2|Y_1$ is multivariate Normal with

$$E[Y_2|Y_1] = \mu_2 + V_{12}^T V_{11}^{-1}(Y_1 - \mu_1) \qquad \mathrm{Var}(Y_2|Y_1) = V_{22} - V_{12}^T V_{11}^{-1} V_{12}.$$

On differentiation

Suppose θ is a p-vector of parameters and $F(\theta)$ is a scalar function of the elements of θ. Then we write the vector of derivatives as

$$\frac{\partial F(\theta)}{\partial \theta} = (\frac{\partial F(\theta)}{\partial \theta_1}, \frac{\partial F(\theta)}{\partial \theta_2}, \ldots, \frac{\partial F(\theta)}{\partial \theta_p})^T.$$

If b is a p-vector of constants and B a $p \times p$ symmetric matrix of constants

$$\frac{\partial (b^T\theta)}{\partial \theta} = \frac{\partial (\theta^T b)}{\partial \theta} = b \qquad \frac{\partial (\theta^T B \theta)}{\partial \theta} = 2B\theta$$

OLS estimates for linear model

Assume $E[Y] = X\beta$ and define

$$SS = (Y - X\beta)^T(Y - X\beta) = Y^TY - 2Y^TX\beta + \beta^TX^TX\beta$$

So

$$\frac{\partial SS}{\partial \beta} = -2X^TY + 2X^TX\beta$$

leading to

$$\hat{\beta} = (X^TX)^{-1}X^TY$$

provided the inverse exists, and

$$E[\hat{\beta}] = \beta \qquad \text{Var}(\hat{\beta}) = (X^TX)^{-1}X^T\text{Var}(Y)X(X^TX)^{-1}.$$

Appendix E. One-way Analysis of Variance revision

Assumptions and notation

Assume g groups, m_h (scalar) observations in group h, Y_{ih} is observation i in group h, $m = \sum m_h$ observations in total. Usually the groups correspond to different forms of *treatment*.

Assume mutually independent Normal data, common variance σ^2, group means $\mu_1, \mu_2, \mu_3, \ldots, \mu_g$. So $Y_{ih} \sim N(\mu_h, \sigma^2)$.

Denote

$$Y_{+h} = \sum_{i=1}^{m_h} Y_{ih} \qquad Y_{i+} = \sum_{h=1}^{g} Y_{ih} \qquad Y_{++} = \sum_{i=1}^{m_h}\sum_{h=1}^{g} Y_{ih}$$

and

$$Y_{.h} = Y_{+h}/m_h \qquad Y_{i.} = Y_{i+}/g \qquad Y_{..} = Y_{ih}/m$$

Hypothesis to test $H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_g$

Rationale

We *partition* the variability in the data into two components:

1. Variablity *between* group means
2. Variability *within* groups

With appopriate scaling we derive two estimates of the variance σ^2. One, from the between group variability (called *between group mean square BGMS*) is unbiased only if the null hypothesis is true. Otherwise it tends to be larger than σ^2. The second, from the within group

variability (called the *within group mean square* WMS or *residual mean square* or *error mean square*) is always an unbiased estimate of σ^2.

Hence we make a decision about the hypothesis by comparing $BGMS$ with WMS. If $BGSS >> WMS$ then the hypothesis is unlikely to be true.

The estimates are constructed so that $BGSS/WMS$ has an $F-$distribution if the hypothesis of equal means is correct. We use this for the formal comparison of $BGMS$ and WMS: does $BGMS/WMS$ look as though it could have come from an $F-$distribution (with the appropriate df)? If yes, we conclude the hypothesis could be true. If no, we conclude it is likely to be false.

Partitioning the total sum of squares

Let TSS denote the total sum of squares about the overall mean:

$$TSS = \sum_g \sum_h \{Y_{ih} - Y_{..}\}^2$$

To partition we use the common trick of adding in something and taking it away again. So

$$\begin{aligned}
TSS &= \sum_h \sum_i \{(Y_{ih} - Y_{.h}) + (Y_{.h} - Y_{..})\}^2 \\
&= \sum_h \sum_i (Y_{ih} - Y_{.h})^2 + 2\sum_h \sum_i (Y_{ih} - Y_{.h})(Y_{.h} - Y_{..}) + \sum_h \sum_i (Y_{.h} - Y_{..})^2 \\
&= \sum_h \sum_i (Y_{ih} - Y_{.h})^2 + 2\sum_h \left\{ (Y_{.h} - Y_{..}) \sum_i (Y_{ih} - Y_{.h}) \right\} + \sum_h m_h (Y_{.h} - Y_{..})^2 \\
&= \sum_h \sum_i (Y_{ih} - Y_{.h})^2 + 0 + \sum_h m_h (Y_{.h} - Y_{..})^2 \\
&= WSS + BGSS
\end{aligned}$$

say.

Two estimates of variance

Note that WSS is a function of the within-group sample variances $s_h^2 = \sum_i (Y_{ih} - Y_{.h})^2/(m_h - 1)$, and

$$WSS = \sum_h (m_h - 1) s_h^2$$

Each s_h^2 is an unbiased estimate of σ^2 which means

$$E[WSS] = (m - g)\sigma^2 \qquad (m = \sum_h m_h \text{ remember})$$

So the *within mean square*

$$WMS = WSS/(m - g)$$

is unbiased for σ^2.

Now note

$$\begin{aligned}
BGSS &= \sum_h m_h (Y_{.h} - Y_{..})^2 \\
&= \sum_h m_h Y_{.h}^2 - 2Y_{..} \sum_h m_h Y_{.h} + mY_{..}^2 \\
&= \sum_h m_h Y_{.h}^2 - mY_{..}^2
\end{aligned}$$

Under the null hypothesis of common means, say μ, $Y_{.h} \sim \mathrm{N}(\mu, \sigma^2/m_h)$ and so

$$E[Y_{.h}^2] = \frac{\sigma^2}{m_h} + \mu^2$$

Also $Y_{..} \sim \mathrm{N}(\mu, \sigma^2/m)$ and

$$E[Y_{..}^2] = \frac{\sigma^2}{m} + \mu^2$$

Hence

$$E[BGSS] = (g-1)\sigma^2$$

and the *between group mean square*

$$BGMS = BGSS/(g-1)$$

is unbiased for σ^2 if the null hypothesis is true. If the group means are not all equal it is possible (and not too difficult - try it) to show that $E[BGMS] > \sigma^2$, with the excess becoming larger as the differences between means increases.

Another point is that WMS is a function only of the sample variances s_h^2 and $BGMS$ is a function only of the sample means $Y_{.h}$. For Normal data the sample mean is independent of the sample variance, which means $WMSS$ is independent of $BGMS$. From this we can show that the ratio $BGMS/WMS$ has an F-distribution with $g-1$ and $m-g$ degrees of freedom under the assumption of equal means.

Layout

Calculations are usually laid out in the form of a table giving the sums of squares, degrees of freedom, mean squares and final F ratio.

Example:

```
ANALYSIS OF VARIANCE ON PANSS WEEK 0
SOURCE      DF         SS         MS          F         p
TMT          2        814        407       1.14     0.323
ERROR      266      95216        358
TOTAL      268      96030
```

Extensions

This is the simplest case of ANOVA. Extensions allow the data to have a more complicated structure, such as structured treatments, say different types of drugs and different administration methods. We can use ANOVA to compare both drugs and administration methods simultaneously, provided certain balance conditions are satisfied.

There are lots more extensions too, not least on how to proceed if the ANOVA leads to the conclusion that the group means are not equal.