MATH5885 Longitudinal Data Analysis

Week 1, Lecture 2

Spiridon Penev, William Dunsmuir, Sally Galbraith, Andrew Hayen, Marc Donoghoe, Gordana Popovic

30 May 2024

Outline

Terminology and Notation

Terminology

▶️正在研究的参与者或单位通常被称为个人或受试者。

▶️主题在不同场合或时间进行测量。

▶️如果在常见的场合对所有受试者进行测量，则该研究据说随着时间的推移是平衡的，否则它是不平衡的。

- Participants or items being studied are usually called individuals or subjects.
- Subjects are measured at different occasions or times.
- If measurements are obtained at a common set of occasions[1] on all subjects, the study is said to be balanced over time, otherwise it is unbalanced.

[1] Note that this depends on the definition of the time origin.

Missing data

数据缺失是纵向研究中的一个常见问题——带有缺失数据的纵向数据集是“不完整的”，以将其与其他形式的不平衡数据区分开来。 这强调了没有获得计划的测量结果。

请注意，平均响应时间的图可能会误导缺失的数据——这些图可能反映缺失的模式，而不是个人内部的变化。

缺少数据可能会导致效率下降，并可能导致估计偏差。 失踪的原因需要调查。

在对人类受试者的纵向研究中，平衡和完整的数据很少见。在长时间随访的研究中尤其如此。

Missing data are a common problem in longitudinal studies—longitudinal datasets with missing data are “incomplete” to distinguish them from other forms of unbalanced data. This emphasises that a planned measurement was not obtained.

Note that plots of mean response against time may be misleading with missing data—the plots may reflect patterns of missingness rather than change within individuals.

Missing data can cause loss of efficiency and can bias estimates. Causes of missingness need to be investigated.

Data that are balanced and complete are rare in longitudinal studies on human subjects. This is particularly true in studies with long periods of follow-up.

Notation

We will use the following notation throughout the course. Let Y_{ij} denote the response for the ith individual ($i = 1, \ldots, N$) at the jth occasion ($j = 1, \ldots, n$).

The mean of a response Y_{ij} is denoted

$$\mathbb{E}(Y_{ij}) = \mu_{ij}$$

The variance of Y_{ij} is defined as

$$\text{var}(Y_{ij}) = \mathbb{E}(Y_{ij} - \mu_{ij})^2 = \sigma_j^2$$

(Note that we are allowing the variance to be different from occasion to occasion, but we are not at this stage allowing it differ across subjects)

Notation

The covariance between responses at two different occasions j and k is

$$\text{cov}(Y_{ij}, Y_{ik}) = \mathbb{E}((Y_{ij} - \mu_{ij})(Y_{ik} - \mu_{ik})) = \sigma_{jk}$$

and the correlation is

$$\text{corr}(Y_{ij}, Y_{ik}) = \frac{\sigma_{jk}}{\sigma_j \sigma_k} = \rho_{jk}$$

Covariance matrix

We can collect the n repeated measurements on subject i into a vector $Y_i = (Y_{i1}, \ldots, Y_{in})$. The covariance of Y_i is then defined to be the matrix:

$$\text{cov}(Y_i) = \begin{pmatrix} \text{var}(Y_{i1}) & \text{cov}(Y_{i1}, Y_{i2}) & \cdots & \text{cov}(Y_{i1}, Y_{in}) \\ \text{cov}(Y_{i2}, Y_{i1}) & \text{var}(Y_{i2}) & \cdots & \text{cov}(Y_{i2}, Y_{in}) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(Y_{in}, Y_{i1}) & \text{cov}(Y_{in}, Y_{i2}) & \cdots & \text{var}(Y_{in}) \end{pmatrix}$$

This matrix is symmetric because $\text{cov}(Y_{ij}, Y_{ik}) = \text{cov}(Y_{ik}, Y_{ij})$.

Note: we won't use boldface notation to distinguish vectors from scalars in this course: the context will make this clear.

Covariance matrix...

More compactly, this can be written as:

$$\text{cov}(Y_i) = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{pmatrix}$$

The correlation matrix (also symmetric) is defined as

$$\text{corr}(Y_i) = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & 1 & \cdots & \rho_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \cdots & 1 \end{pmatrix}$$

Correlation in Longitudinal Data

Factors affecting correlation in longitudinal data

Recall from Tuesday's lecture that there are three main sources of variation that impact on correlation in longitudinal data:

1. Between-individual heterogeneity
2. Within-individual biological variation
3. Measurement error

Nature of correlation in longitudinal data

- Repeated observations from the same individual are unlikely to be independent.
- The correlation between measurements on the same individual is usually positive.
- Commonly, measurements that are taken close together in time are usually more correlated than measurements taken further apart.
- FLW (p 36) note that correlations between measurements from the same individual are rarely close to 0 (even when taken many years apart) or 1 (even when very close together in time).

- In addition, the variance of measurements may not be constant over time. For example, baseline measurements are often less variable than post-baseline measurements.

Effect of measurement error

- Recall that measurement error occurs if two measurements taken under identical conditions do not agree.
- The effect of measurement error is to shrink the correlation between repeated measures towards zero.
- The use of a less reliable measurement tool will result in repeated measurements with smaller correlations than if a more reliable instrument is used.

Exercise 1

Effect of between-individual heterogeneity

Consider the following simple model for longitudinal data:

$$Y_{ij} = \alpha + u_i + \varepsilon_{ij} \tag{1}$$

where:

- Y_{ij} is the response for individual i at time j,
- α is the overall mean response,
- u_i represents deviation from the overall mean for individual i, and
- ε_{ij} represents measurement error

Assume that $u_i \sim$ i.i.d. $N(0, \nu^2)$, $\varepsilon_{ij} \sim$ i.i.d. $N(0, \sigma^2)$, and the u_i and ε_{ij} are independent.

Exercise 1

Effect of between-individual heterogeneity

In Tuesday's lecture, I claimed that for longitudinal data:

> *Two responses taken on the same individual are expected to be more similar than responses from two different individuals. We would thus expect to see positive correlation between the repeated measurements.*

Use (1) to justify the "thus" in this claim: i.e. show how the second sentence follows from the first if responses are "more similar" in the sense of Model (1).

Exercise 1

Effect of between-individual heterogeneity

Exercise 1

Effect of between-individual heterogeneity

Exercise 2

Effect of measurement error

A claim made in today's lecture was:

> *The effect of measurement error is to shrink the correlation among the repeated measures towards 0.*

Show that this claim also follows if Model (1) holds.

Exercise 2

Effect of measurement error

Ignoring Correlation in Longitudinal Data

TLC trial

For the TLC data (using just the succimer group), suppose we wanted to see if there is a change from baseline to week 1. An estimate of the change is

$$\hat{\delta} = \hat{\mu}_2 - \hat{\mu}_1$$

where

$$\hat{\mu}_j = \frac{1}{N} \sum_{i=1}^{N} Y_{ij}$$

The variance of our estimate is

$$\text{var}(\hat{\delta}) = \frac{1}{N}(\sigma_1^2 + \sigma_2^2 - 2\sigma_{12})$$

Estimated variance of $\hat{\delta}$

If we use the sample variances and covariances we obtain the following estimate of the variance:

$$\widehat{\text{var}}(\hat{\delta}) = \frac{1}{50}(25.2 + 58.9 - 2 \times 15.5) = 1.06$$

If we ignore the correlation between the repeated measurements, then our estimate of the variance would be:

$$\frac{1}{50}(25.2 + 58.9) = 1.68$$

Ignoring correlation

If we were to ignore correlation in the repeated measurements, then we would obtain standard errors that are too large, and confidence intervals that are too wide. Note that in this case we are making a comparison within individuals.

R analysis

The previous example is easily done using R:

```
> vnames <- c("id","group",
+             paste("wk",c(0,1,4,6),sep=""))
> tlc <- read.table("../../data/TLC100.txt",
+                   header = FALSE, col.names = vnames)
> succimer <- subset(tlc, group == "A")
> lead1 <- succimer[,3:6]
> mean1 <- apply(lead1, 2, mean)
> varmat1 <- var(lead1)
> var1 <- diag(varmat1)
```

R analysis

```
> mean1

   wk0     wk1     wk4     wk6
26.540 13.522 15.514 20.762

> varmat1

          wk0      wk1      wk4      wk6
wk0 25.20980 15.46543 15.13800 22.98543
wk1 15.46543 58.86706 44.02907 35.96596
wk4 15.13800 44.02907 61.65715 33.02197
wk6 22.98543 35.96596 33.02197 85.49465

> var1

     wk0      wk1      wk4      wk6
25.20980 58.86706 61.65715 85.49465
```

R analysis

Accounting for correlation:

```
> N <- nrow(lead1)
> delta.hat <- diff(mean1[1:2])
> var.delta.hat <- (sum(var1[1:2])-2*varmat1[1,2])/N
> se.delta.hat <- sqrt(var.delta.hat)
> tstat <- delta.hat/se.delta.hat
```

Ignoring correlation:

```
> var.delta.hat0 <- sum(var1[1:2])/N
> se.delta.hat0 <- sqrt(var.delta.hat0)
> tstat0 <- delta.hat/se.delta.hat0
```

R analysis

Get the results together

```
> res <- cbind(
+ c(delta.hat,var.delta.hat,se.delta.hat,tstat),
+ c(delta.hat,var.delta.hat0,se.delta.hat0,tstat0))
> rownames(res) <- c("estimate","var","se","tstat")
> colnames(res) <- c("accounting","ignoring")
```

R analysis

```
> res
```

	accounting	ignoring
estimate	-13.01800	-13.018000
var	1.06292	1.681537
se	1.03098	1.296741
tstat	-12.62682	-10.039014

Using SAS for this analysis

Please see the text web site for data and SAS programs.

https://content.sph.harvard.edu/fitzmaur/ala2e/

Exploratory Data Analysis

Exploratory data analysis

An analysis should begin with the use of some graphical displays of data.

The aim of exploratory data analysis (EDA) is to assess the relationship between the mean response and covariates (especially time).

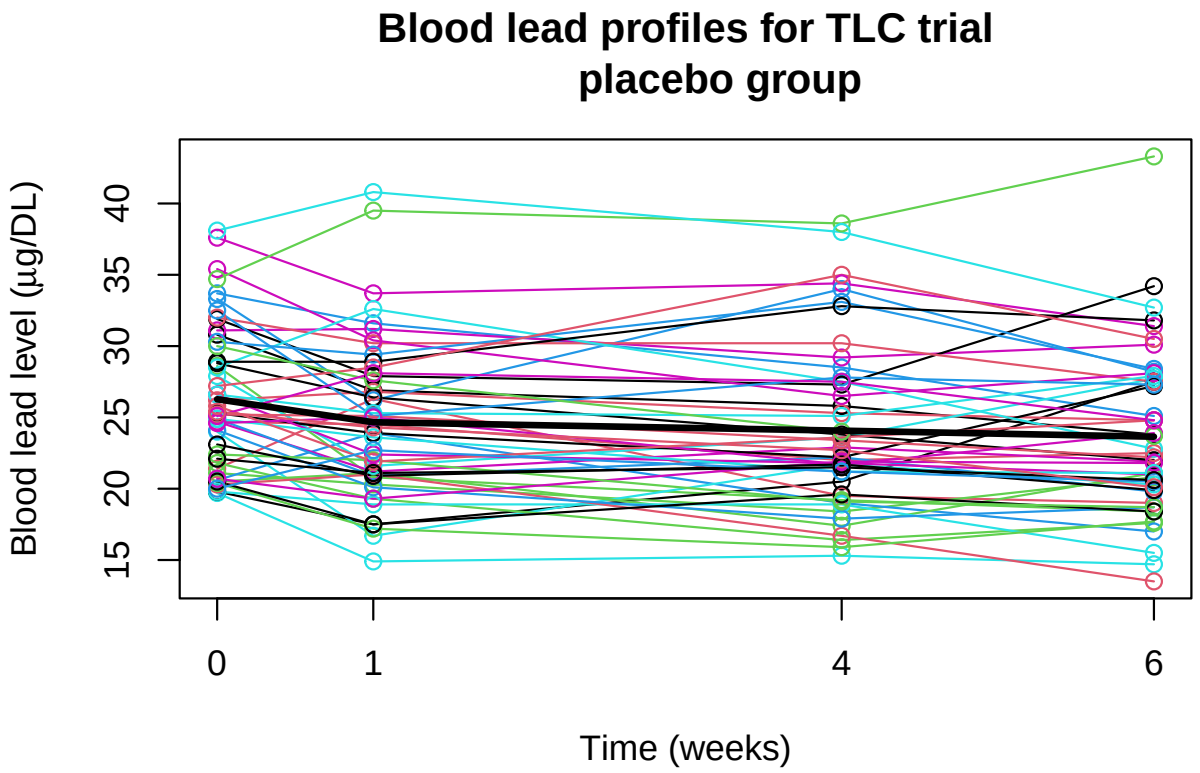
Another aim is to get a feel for the the variance and correlation structure in the data.

Individual profiles: "spaghetti plots"

- It is useful to plot each individual's measurements over time, joining the points corresponding to the same individual. This gives a plot showing a series of piecewise linear curves, each curve representing one individual.
- Connecting the repeated measurements helps to display changes over time for individuals.
- However, for large datasets, these can be too cluttered to allow us to clearly discern patterns.
- One remedy is to show individual profiles using a pale colour, with a typical profile, and perhaps profiles for a small subset of individuals, shown in bold.

"Spaghetti plots" for TLC data

We have already seen an example of a spaghetti plot for the placebo group in the TLC trial



R function to make "spaghetti plots"

We can write an R function to produce a spaghetti plot for any dataset in which there are complete measurements at the same times for all individuals:

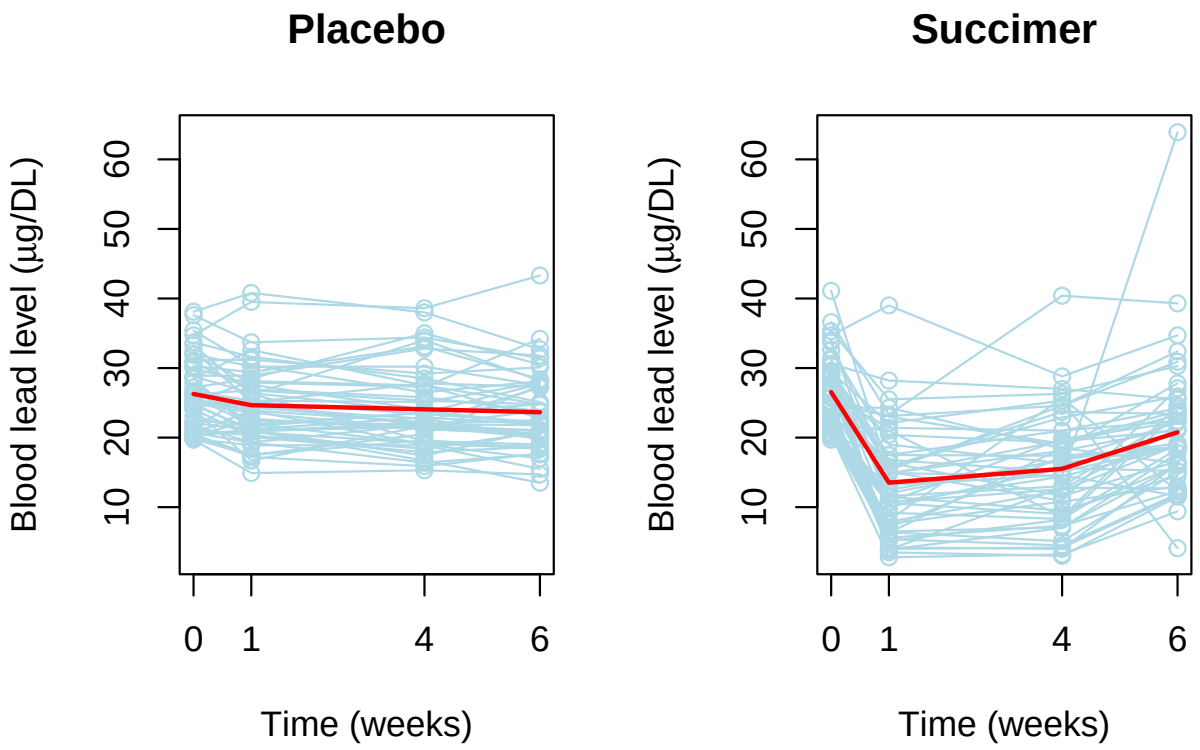
```
> spagplot <- function(x,y,ylimit,xlabel,ylabel,heading) {
+   meany <- apply(y,1,mean)
+   matplot(x,y,type="l",lty=1,xlab=xlabel,
+           ylab=ylabel,ylim=ylimit,xaxt="n",
+           col="lightblue")
+   axis(side=1,at=x)
+   matpoints(x,y,type="p",pch=1,col="lightblue")
+   title(main=heading)
+   lines(x,meany,type="l",lwd=2,col="red")
+ }
```

"Spaghetti plots" for TLC data

Try out the function on the TLC data:

```
> x <- c(0,1,4,6)
> lead0 <- placebo[,3:6]
> lead1 <- succimer[,3:6]
> ylab0 <- expression(paste("Blood lead level (",
+                           mu,"g/DL)"))
> xlab0 <- "Time (weeks)"
> ylim0 <- c(min(lead0,lead1),max(lead0,lead1))
> title0 <- "Placebo"
> title1 <- "Succimer"
> par(mfrow=c(1,2))
> spagplot(x,t(lead0),ylim0,xlab0,ylab0,title0)
> spagplot(x,t(lead1),ylim0,xlab0,ylab0,title1)
```

"Spaghetti plots" for TLC data



Mean profiles

当以所有个人的一组共同时间进行测量时，可以对时间计算和绘制每个时间点的平均值。为协变量的不同值（例如，不同的治疗组）显示单独的图通常很有用。

在上一张图表中，这些平均轮廓被覆盖在意大利面条图上。我们还可以编写一个R函数，在单个图上显示每个组的平均配置文件。

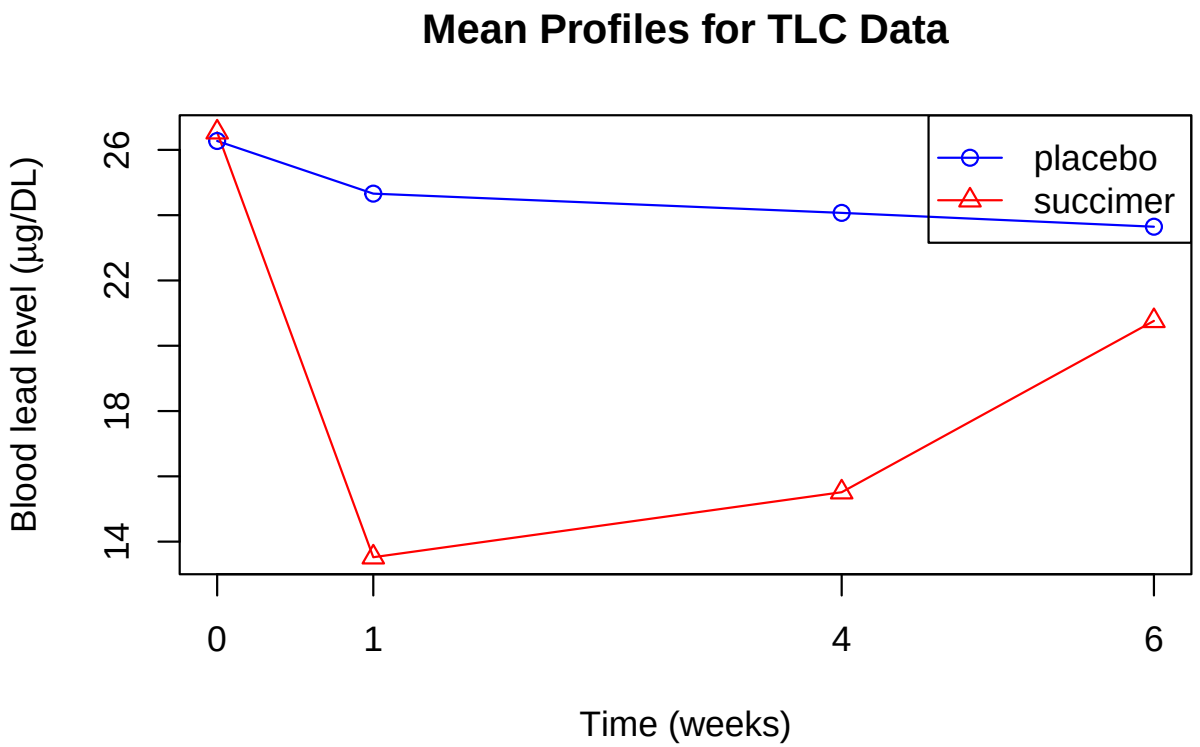
When measurements are taken at a common set of times for all individuals, the mean at each timepoint can be calculated and plotted against time. It is often useful to display separate plots for different values of covariates (for example, different treatment groups).

In the previous graph, these mean profiles were overlaid on the spaghetti plots. We can also write an R function to display the mean profiles for each group on a single plot.

Mean profiles function

```
> meanplot <- function(x,means,xlabel,ylabel,heading,
+                      legnames,legloc) {
+   matplot(x,means,type="l",lty=1,xlab=xlabel,ylab=ylabel,
+           xaxt="n",col=c("blue","red"))
+   axis(side=1,at=x)
+   matpoints(x,means,type="p",pch=1:2,col=c("blue","red"))
+   title(main=heading)
+   legend(legloc, legnames, lty=1, col=c("blue","red"),
+          pch=1:2)
+ }
> means1 <- cbind(apply(lead0,2,mean),apply(lead1,2,mean))
> meanplot(x, means1, xlab0, ylab0,
+          "Mean Profiles for TLC Data",
+          c("placebo","succimer"),"topright")
```

Mean profiles for TLC data



Mean profiles for TLC data

从之前的情节来看，琥珀的效果在第一周似乎更大，然后在后几周反弹到基线水平。

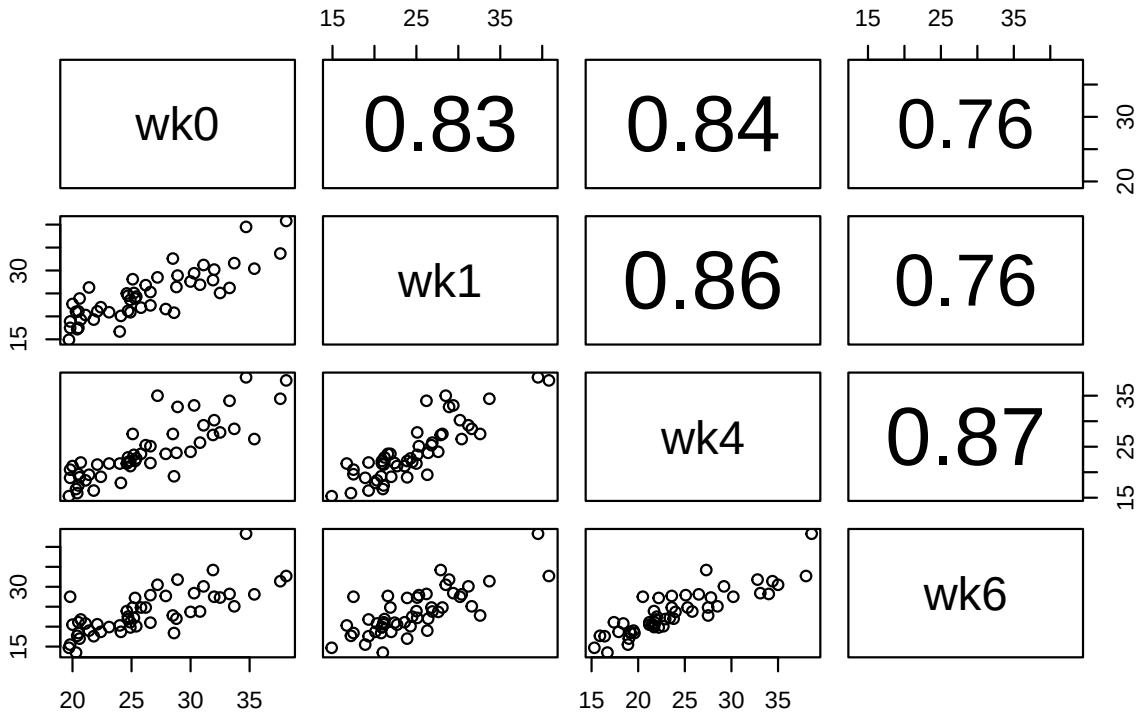
此外，这个群体似乎存在非线性趋势。

From the previous plot, it appears that the effect of succimer is greater at week one, before rebounding towards the baseline level in later weeks.

Additionally, it appears that there is a non-linear trend for this group.

Exploring correlation

Recall the correlation plot for the placebo group in the TLC trial:



Exploring correlation

In general, to remove the effects of explanatory variables, the first step is to regress the response against any covariates and obtain residuals:

$$r_{ij} = Y_{ij} - x_{ij}^{\mathsf{T}} \hat{\beta}$$

If the measurements are equally spaced, we can plot r_{ij} against r_{ik} for all $j < k$.

Exploring correlation: TLC data

First we need to convert the data from 'wide' into 'long' format:

```
> tlc.long <- reshape(tlc, varying = list(3:6),
+                     timevar = "week", v.names = "lead",
+                     direction = "long",
+                     times = factor(paste0("wk", c(0,1,4,6))
> head(tlc.long)
```

```
      id group week lead
1.wk0  1     P  wk0 30.8
2.wk0  2     A  wk0 26.5
3.wk0  3     A  wk0 25.8
4.wk0  4     P  wk0 24.7
5.wk0  5     A  wk0 20.4
6.wk0  6     A  wk0 20.4
```

Exploring correlation: TLC data

Regress the lead level on time and treatment, and save the residuals:

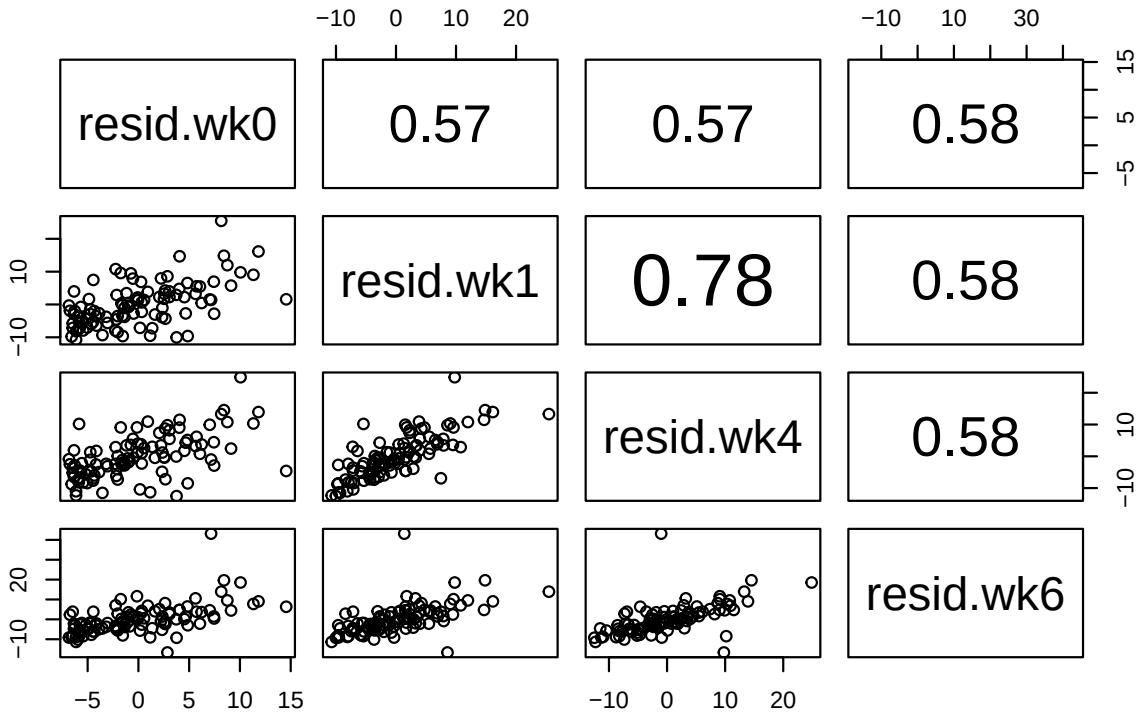
在时间和处理上回归铅水平，并保存残留物：

```
> tlc.lm <- lm(lead ~ group*week, data = tlc.long)
> tlc.long$resid <- resid(tlc.lm)
> tlc.wide <- reshape(tlc.long[,c("id", "week", "resid")],
+                     timevar = "week", v.names = "resid",
+                     idvar = "id", direction = "wide")
> head(tlc.wide)
```

	id	resid.wk0	resid.wk1	resid.wk4	resid.wk6
1.wk0	1	4.528	2.240	1.730	0.154
2.wk0	2	-0.040	1.278	3.986	0.238
3.wk0	3	-0.740	9.478	3.586	2.438
4.wk0	4	-1.572	-0.160	-2.070	-1.146
5.wk0	5	-6.140	-10.722	-12.314	-11.362
6.wk0	6	-6.140	-8.122	-11.014	-8.862

Exploring correlation: TLC data

```
> pairs(tlc.wide[,-1], upper.panel = panel.cor)
```



Exercise 3

HRT trial

The comma-delimited text file `hrt.txt` on the course web page gives data from a randomised, placebo controlled, double-blind clinical trial of hormone replacement therapy (HRT). The outcome of interest is a measure of depression, the Hamilton Depression Score (higher values corresponding to a greater level of depression). Women taking part in the trial were rated on this scale on two occasions prior to randomization, and then monthly for four consecutive months after treatment commenced.

课程网页上的逗号分隔文本文件hrt.txt提供了随机、安慰剂对照、双盲激素替代疗法（HRT）临床试验的数据。兴趣的结果是衡量抑郁症的指标，汉密尔顿抑郁症评分（更高的值对应于更高的抑郁症水平）。参加试验的妇女在随机化前两次按此比例进行评级，然后在治疗开始后连续四个月每月进行评级。

出于我们的目的，假设所有测量，包括基线，都是每月一次进行的。

For our purposes, assume that all measurements, including the baselines, took place at monthly intervals.

Exercise 3

HRT trial data

数据集中的变量是两个基线测量，基数1和基数2，以及四个基线后测量月1、月2、月3和4月。前20名患者（20行）属于安慰剂组，其余20名患者被分配到HRT。

请注意，缺少一些测量值（表示为NA）

The variables in the dataset are the two baseline measurements, `base1` and `base2`, and the four post-baseline measurements `month1`, `month2`, `month3` and `month4`. The first 20 patients (20 rows) belong to the placebo group, with the remaining 20 patients allocated to HRT.

Note that there are some missing measurements (denoted `NA`)

Read in the data:

```
> hrt <- read.csv(file = "../../data/hrt.txt")
```

Exercise 3

HRT trial exercises

1. Produce "spaghetti plots" of the individual depression score profiles, adding the mean profiles, separately for each group. (You may wish to write a function).
2. Plot the mean profiles on the same graph, with a legend.
3. Comment on the graphs in (a) and (b).
4. Calculate the changes from the first baseline measurement, and repeat parts (a) and (b). Comment.
5. Calculate the covariance and correlation matrices of the six measurements, separately for each treatment group. To deal with the missing measurements, use the `use="pairwise.complete.obs"` option with the `var` and `cor` functions. Also calculate the pooled covariance matrix, and use it to derive the pooled correlation matrix. (Hint: the function `cov2cor` calculates the correlation matrix from a given covariance matrix).
6. Plot the variances as a function of time for each group. Also produce pairwise scatterplot matrices for each group, using `pairs`, and add the correlations on the upper diagonal.

1. 生成单个抑郁评分配置文件的"意大利面条图"，为每个组分别添加平均配置文件。（您可能希望编写一个函数）。
2. 用图例在同一图上绘制平均剖面图。
3. 评论（a）和（b）中的图表。
4. 计算第一次基线测量的变化，并重复（a）和（b）部分。评论。
5. 为每个治疗组分别计算六种测量值的协方差和相关矩阵。要处理缺失的测量值，请使用带有var和cor函数的use="pairwise.complete.obs"选项。还要计算集合协方差矩阵，并使用它来推导集合相关矩阵。（提示：函数cov2cor从给定的协方差矩阵计算相关矩阵）。