

Longitudinal data analysis (day 1)

Advanced modelling with R

Juan R Gonzalez

juanr.gonzalez@isglobal.org

BRGE - Bioinformatics Research Group in Epidemiology

ISGlobal - Barcelona Institute for Global Health

<http://brge.isglobal.org>

Data format

Data can be available in width or long format

```
data.width <- read.table("../data/hypothetical_anchors.txt",  
                          header=TRUE)  
head(data.width)
```

	id	time1	time2	time3	time4	group
1	1	31	29	15	26	A
2	2	24	28	20	32	A
3	3	14	20	28	30	A
4	4	38	34	30	34	B
5	5	25	29	25	29	B
6	6	30	28	16	34	B

Data format

```
data.long <- read.table("../data/hypothetical_largo.txt",  
                        header=TRUE)  
head(data.long)
```

	id	time	score	group
1	1	1	31	A
2	1	2	29	A
3	1	3	15	A
4	1	4	26	A
5	2	1	24	A
6	2	2	28	A

Data format

Data are normally collected in width format since they can be easily collected in a database. You can use the function `reshape` to transform data into long format as following

```
data.long2 <- reshape(data.width, direction = "long",  
                      varying = list(paste0("time", 1:4)),  
                      idvar = c("id", "group"),  
                      v.names="score",  
                      timevar = "time", times=1:4)
```

Data format

```
subset(data.long, id==1)
```

	id	time	score	group
1	1	1	31	A
2	1	2	29	A
3	1	3	15	A
4	1	4	26	A

```
subset(data.long2, id==1)
```

	id	group	time	score
1.A.1	1	A	1	31
1.A.2	1	A	2	29
1.A.3	1	A	3	15
1.A.4	1	A	4	26

Data format

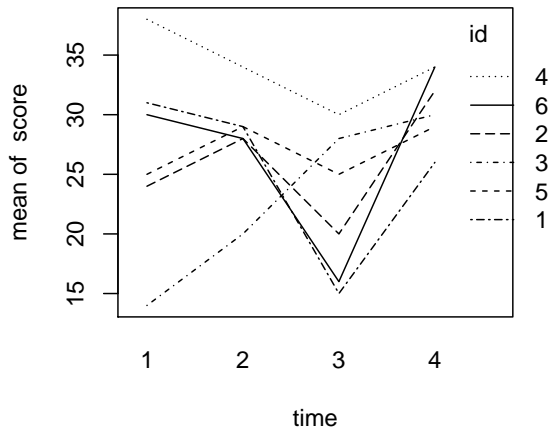
Another option with reshape2

```
data.long3 <- reshape2::melt(data.width,  
                             id.vars = c("id", "group"))  
subset(data.long3, id==1)
```

	id	group	variable	value
1	1	A	time1	31
7	1	A	time2	29
13	1	A	time3	15
19	1	A	time4	26

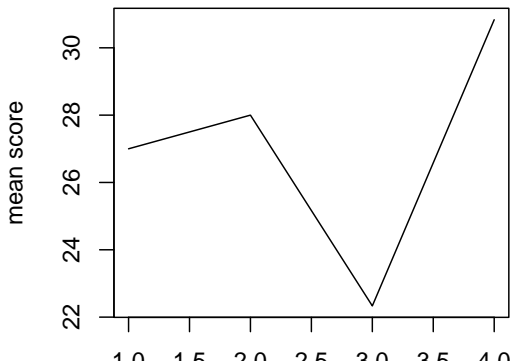
Data visualization - Individual profiles

```
with(data.long, interaction.plot(time, id, score))
```



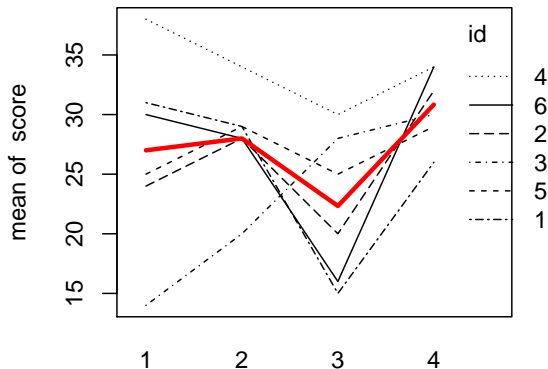
Data visualization - Overall profile

```
datos.agg <- aggregate(score ~ time,  
                        data = data.long, mean)  
with(datos.agg, plot(time, score,  
                     ylab="mean score", type="l"))
```



Data visualization - Both

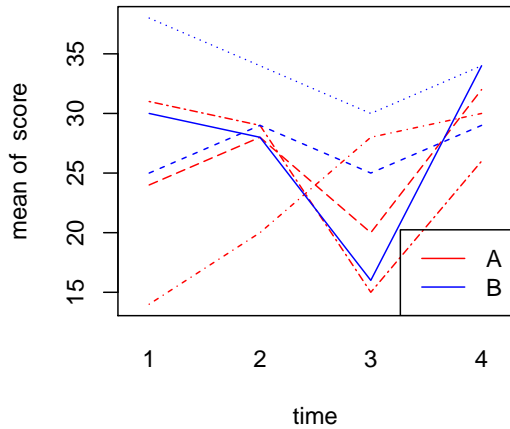
```
with(data.long, interaction.plot(time, id, score))  
with(datos.agg, lines(time, score, ylab="mean score",  
                      col="red", lwd=3))
```



Data visualization - Grouping

```
with(data.long, interaction.plot(time, id, score,  
  col=c(rep("red",3), rep("blue",3)),  
  legend=FALSE))  
legend("bottomright", lty=1, c("A","B"),  
  col=c("red", "blue"))
```

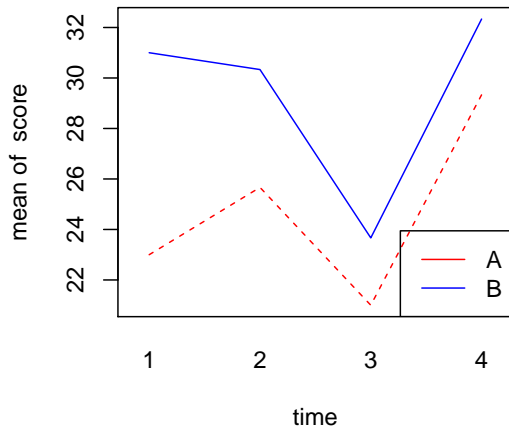
Data visualization - grouping



Data visualization - overall profile by group

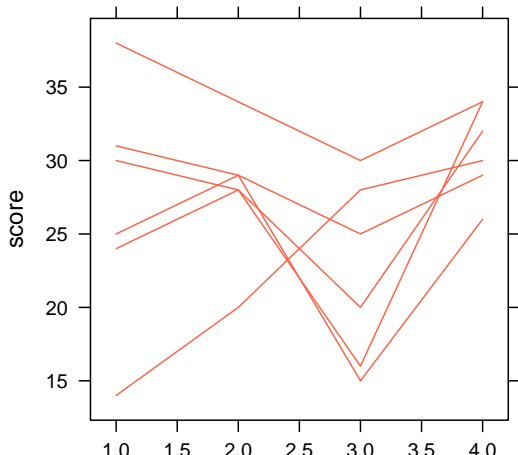
```
datos.agg <- aggregate(score ~ time + group,  
                        data = data.long, mean)  
with(datos.agg, interaction.plot(time, group, score,  
                                col=c("red", "blue"), legend=FALSE))  
legend("bottomright", lty=1, c("A","B"),  
      col=c("red", "blue"))
```

Data visualization - overall profile by group



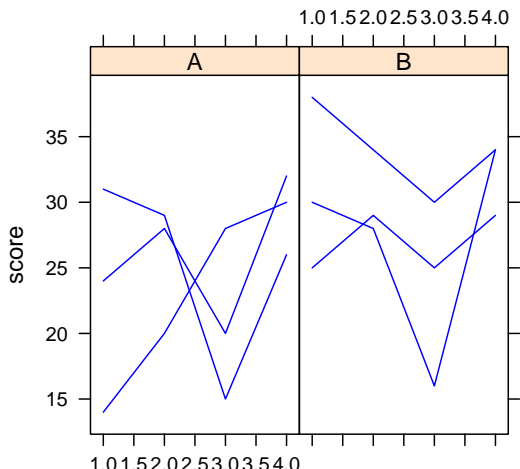
Data visualization - lattice

```
library(lattice)  
xyplot(score ~ time, group=id,  
        data = data.long, type="l", col="tomato")
```



Data visualization - lattice

```
library(lattice)  
xyplot(score ~ time | group, group=id,  
       data = data.long, type="l", col="blue")
```



Data Analysis: End-point analysis

```
mod <- aov(time4 ~ time1 + group, data.width)  
summary(mod)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
time1	1	3.95	3.951	0.335	0.603
group	1	9.55	9.549	0.811	0.434
Residuals	3	35.33	11.778		

Data Analysis: ANOVA repeated measurements

The naive version is obtained by

```
mod <- aov(score ~ as.factor(time)*group, data.long)
summary(mod)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
as.factor(time)	3	224.8	74.93	2.291	0.1173
group	1	126.0	126.04	3.854	0.0673 .
as.factor(time):group	3	26.8	8.93	0.273	0.8439
Residuals	16	523.3	32.71		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Data Analysis: ANOVA repeated measurements

The proper analysis has to be performed by

```
mod <- aov(score ~ as.factor(time)*group + Error(id), data = data)
summary(mod)
```

Error: id

	Df	Sum Sq	Mean Sq
group	1	29.57	29.57

Error: Within

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
as.factor(time)	3	224.8	74.93	2.567	0.0932 .
group	1	182.0	182.03	6.237	0.0246 *
as.factor(time):group	3	26.8	8.93	0.306	0.8207
Residuals	15	437.8	29.18		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Data Analysis: MANOVA

```
mod <- manova(cbind(time1, time2, time3, time4) ~ group, data = da)
summary(mod)
```

	Df	Pillai	approx F	num Df	den Df	Pr(>F)
group	1	0.49226	0.24238	4	1	0.8879
Residuals	4					

Data Analysis: MANOVA

Other tests can be applied

```
summary(mod, test = "Wilks")
```

	Df	Wilks	approx F	num Df	den Df	Pr(>F)
group	1	0.50774	0.24238	4	1	0.8879
Residuals	4					

```
summary(mod, test = "Hotelling-Lawley")
```

	Df	Hotelling-Lawley	approx F	num Df	den Df	Pr(>F)
group	1	0.96951	0.24238	4	1	0.8879
Residuals	4					

Exercise 1

Para estudiar las diferencias entre dos procedimientos diferentes de recuperación de pacientes de un infarto, se consideraron dos grupos experimentales en sendos hospitales, de 8 pacientes cada uno. La variable respuesta es el índice de Bartel, que varía entre 0 y 100, y que constituye una medida de la habilidad funcional con la que se valoran diferentes capacidades, de forma que valores más altos se corresponden con una mejor situación del paciente. De cada uno de los 16 pacientes se dispone de su respuesta cada semana a lo largo de 5 semanas consecutivas. Los datos se pueden encontrar en el archivo *recuperainfarto.txt*.

1. Representa gráficamente las trayectorias de las mediones de forma separada para cada hospital que conforma cada uno de los dos procedimientos.
2. ¿Cuál de los dos procedimientos presenta una mejor recuperación de los pacientes? ¿Es esta diferencia estadísticamente significativa? [NOTA: contesta a estas preguntas planteando la pregunta estadística que creas conveniente]

Exercise 2

En un estudio sobre la agudeza visual se dispone de la respuesta de 7 individuos. La respuesta en cada ojo es el retraso en milisegundos entre la emisión de una luz y la respuesta en a la misma por el cortex. Cada ojo se somete a cuatro mediciones correspondientes a cuatro instantes consecutivos. Se tienen mediciones en el ojo izquierdo y derecho. Los datos se pueden encontrar en el archivo *agudezavisual.txt*

1. Crea una nueva base de datos agregando la información para cada una de la medida repetida (ojo) [NOTA: toma la media - usa la función `aggregate` - esto es el procedimiento habitual cuando no se conocen modelos más avanzados]
2. Representa de forma gráfica la respuesta en cada tiempo.
3. ¿Existen diferencias entre la medición final y la basal?
4. ¿Existe un efecto temporal en la respuesta?