BIOS6643 Longitudinal

L2 Graphs

EJC

Department of Biostatistics & Informatics

Learning objectives:

Introduction

PCA (Principal components analysis)

Learning objectives:

1. Become familiar with multiple ways of representing longitudinal and cluster data.
2. Understand the basic ideas of principal component analysis

Introduction

- Line graph
  - Visual staple for longitudinal data.
  - Generalization of a scatterplot in which points are connected either within subjects or the 'correlated unit'.
  - Intuitive and indicates nested responses (e.g., repeated measures within subjects).
- Scatterplot
  - Can use different symbols for subjects/objects on which repeated measures are taken (avoids criss-cross and tangle of lines in a line graph).
  - Can use scatterplot for time 'x' versus time 'y'.
- Panels
  - Can be used for multiple line graphs or scatterplots, e.g., if a longitudinal study has multiple groups with many subjects. (Also see the growth curve graphs for boys and girls, presented in the Introduction Chapter.)

Graphs for repeated measures data with one sample

Data: The Ramus data come from a prospective study that has existed for over 40 years and was used by dentists to establish a growth curve for the ramus (part of the lower jaw bone) for young boys. Four measurements were made on 20 boys, at ages 8 (h1), 8.5 (h2), 9 (h3) and 9.5 (h4) in mm.

Table 1: The first 3 samples

boy	h1	h2	h3	h4
1	47.8	48.8	49.0	49.7
2	46.4	47.3	47.7	48.4
3	46.3	46.8	47.8	48.5

Table 2: The last 3 samples

	boy	h1	h2	h3	h4
18	18	53.3	54.6	55.1	55.3
19	19	46.2	47.5	48.1	48.4
20	20	46.3	47.6	51.3	51.8

Table 3: Mean

h1	h2	h3	h4
48.655	49.625	50.57	51.45

Table 4: Standard deviation

h1	h2	h3	h4
2.5159	2.5396	2.6302	2.7322

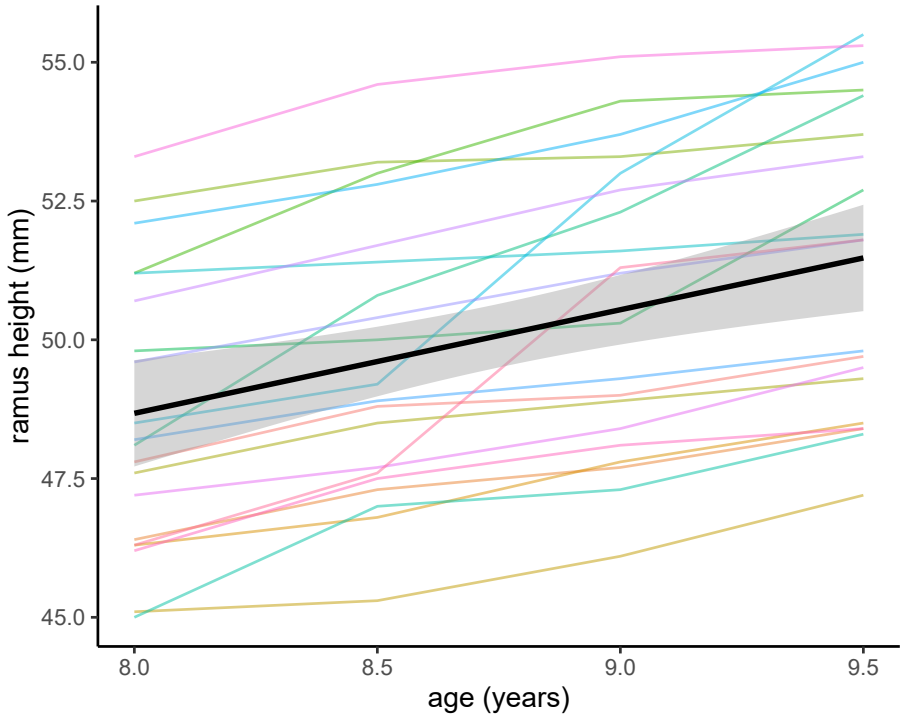
Table 5: Standard error

h1	h2	h3	h4
0.5626	0.5679	0.5881	0.6109

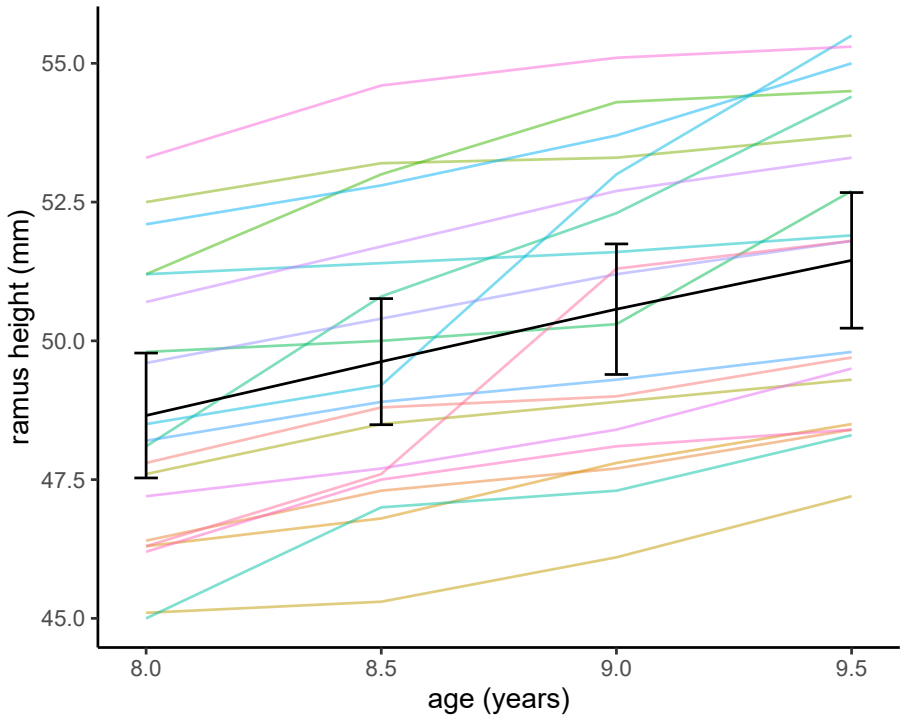
In the following graph, subject lines are in grey and the group mean function is in black. Error bars indicate +/- 2 standard errors from the mean. The grey lines comprise what is sometimes referred to as a spaghetti plot.

```
ramus <- here::here("data", "ramus.dat") %>%
  read.table(header = T, row.names = 1,
             sep = ",",  skip = 0) %>%
  rename("8" = 2, "8.5" = 3, "9" = 4, "9.5" = 5) %>%
  pivot_longer(cols = c("8", "8.5", "9", "9.5"),
               names_to = "time",
               values_to = "ramus height (mm)") %>%
  mutate(`age (years)` = as.numeric(time),
         boy = as.factor(boy))

plot1 <- ggplot() +
  geom_line(data = ramus,
            aes(group = boy,
                x = `age (years)`,
                y = `ramus height (mm)`,
                color = boy),
            alpha = 0.5)
```

```
plot1 +  geom_smooth(data = ramus,
                     aes(x = `age (years)`,
                         y = `ramus height (mm)`),
                     method = "lm", se = TRUE,
                     level = 0.95, color = "black") +
  theme_classic() +
  theme(legend.position = "none")
```

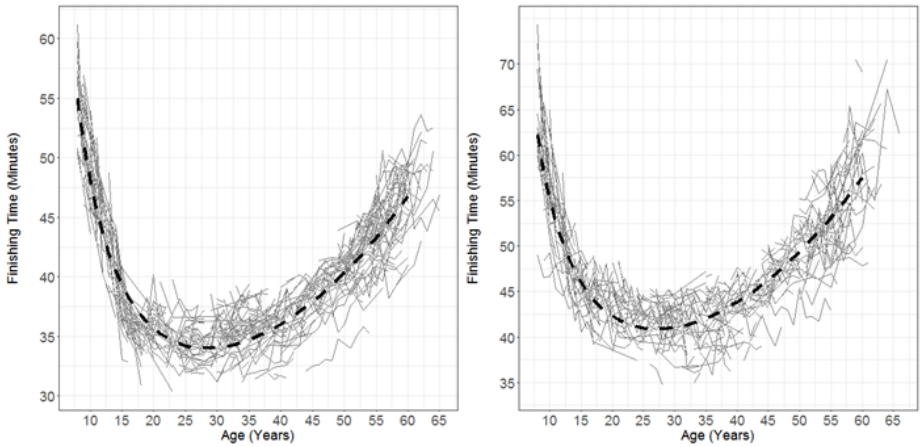


```
ramus_s <- cbind(ramus_mean, ramus_sd,
                 ramus_se, `age (years)`) %>%
  as.data.frame() %>%
  mutate_all(as.numeric) %>%
  round(4)

plot1 + geom_line(data = ramus_s,
                  aes(x = `age (years)`,
                      y = ramus_mean)) +
  geom_errorbar(data = ramus_s,
                aes(x = `age (years)`, y = ramus_mean,
                    ymin = ramus_mean-2*ramus_se,
                    ymax = ramus_mean+2*ramus_se),
                width = 0.05,  position = position_dodge(0.05)) +
  theme_classic() +
  theme(legend.position = "none")
```



Using GGPLOT in R

The package `ggplot2::ggplot` is a more current graphing package; example code and plots shown below. Data are from Strand et al., 2018; these are fastest times by age in the Bolder Boulder 10K road race for men (left) and women (right); individual runners are shown in multiple years by using spaghetti noodles (those with only 1 point were modeled but not shown in these graphs); the dashed curve is the group average. For more detail, see Strand et al., 2018 (Journal of Quantitative Analysis in Sports).

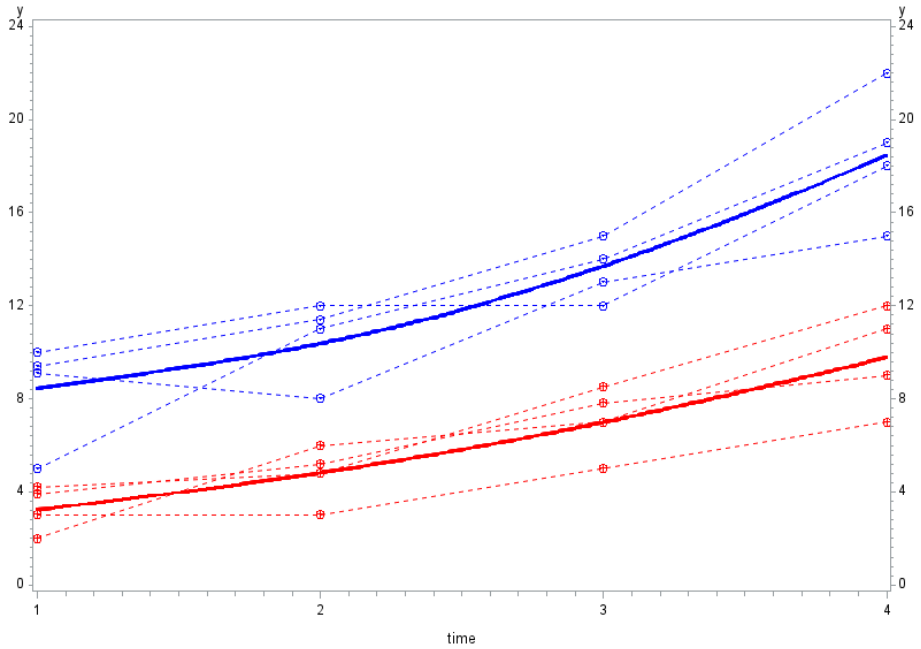


need data and not sure about the code

Graphs for repeated measures data with multiple samples

- Multiple samples present a whole new set of issues when constructing graphs. Consider a simple generic data set with 2 groups (e.g., men, women), where individuals are monitored over time.
- The curves are obtained from PROC MIXED, a procedure that we'll learn more about later. For now, it is enough to understand that it yields predicted values based on the function in the MODEL statement.

Plotting longitudinal data at the group level, with fitted curves, using SAS



SAS code used to obtain the preceding graph

```
data tday; input group id time y @@;
datalines;
1 1 1 3.0 1 1 2 3.0 1 1 3 5.0 1 1 4  7 1 2 1 3.9 1 2 2 5.2  1 2 3 7.8 1 2 4 9
1 3 1 2.0 1 3 2 6.0 1 3 3 7.0 1 3 4 11 1 4 1 4.2 1 4 2 4.8  1 4 3 8.5 1 4 4 12
2 5 1 10  2 5 2 12  2 5 3 12  2 5 4 18 2 6 1 9.1 2 6 2 8.0  2 6 3 13  2 6 4 15
2 7 1 5.0 2 7 2 11  2 7 3 14  2 7 4 19 2 8 1 9.4 2 8 2 11.4 2 8 3 15  2 8 4 22
;
proc mixed data=tday;
class id group;
model y=time time*time group group*time
 group*time*time / outpm=out solution;
random intercept /subject=id solution; run;

symbol c=red i=join value="+" r=4 line=2;
symbol2 c=blue i=join value="-" r=4 line=2;
symbol3 c=red i=spline r=4 width=2;
symbol4 c=blue i=spline r=4 width=2;
axis1 label=("y") order=(0 to 24 by 4);
axis2 label=("time");
proc gplot data=out;
plot y*time=id / vaxis=axis1 haxis=axis2 nolegend;
plot2 pred*time=id / vaxis=axis1 nolegend; run;
```

The outpm option outputs predicted means (not including random effect deviations). Since all subjects were observed at the same time points and there are no subject-variant covariates, the predicted means will be the same for subjects within groups – i.e. one curve per group.

The 'r=4' option tells SAS that we need 4 noodles, one per subject. The r=4 on symbols 3 and 4 are the predicted means, which are the same for each subject. The 'line=2' option makes dashed lines for subjects; a thicker line for the mean was obtained using 'width=2'.

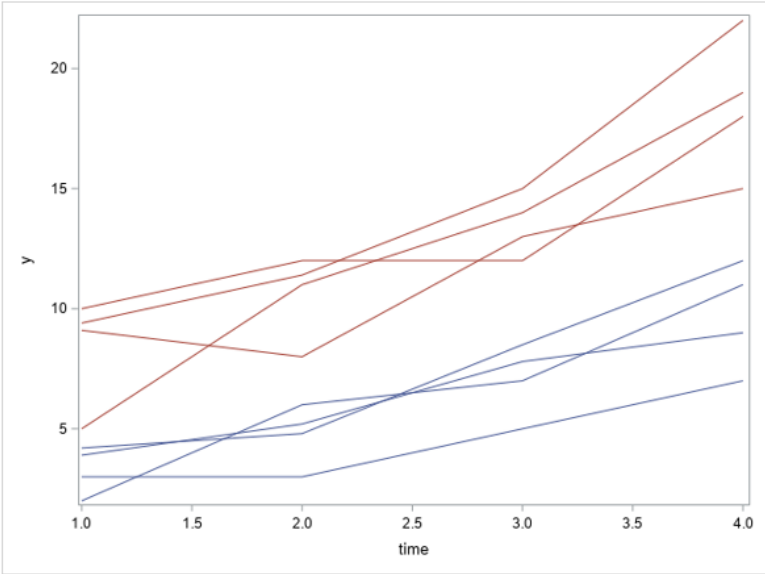
The plot2 option plots a 'second' y-axis, indicated on the right side. In order to get the numbers from plot and plot2 to align, I used the vaxis option, forcing the scale to be the same.

Graphing in SAS has become somewhat easier via the SGPLOT, which mimics some of the features that R graphing has. Below is a spaghetti plot of the same data. Note the minimal amount of coding required to get the plot. The 'reg' statement would allow for plotting of group means, and the 'degree' option can be added to get polynomial curves. See the SAS Help Documentation for more detail.

SAS code used to obtain the following graph:

```
proc sgplot data=out
 noautolegend;
 series x=time y=y
 / group=id grouplc=group;run;
```

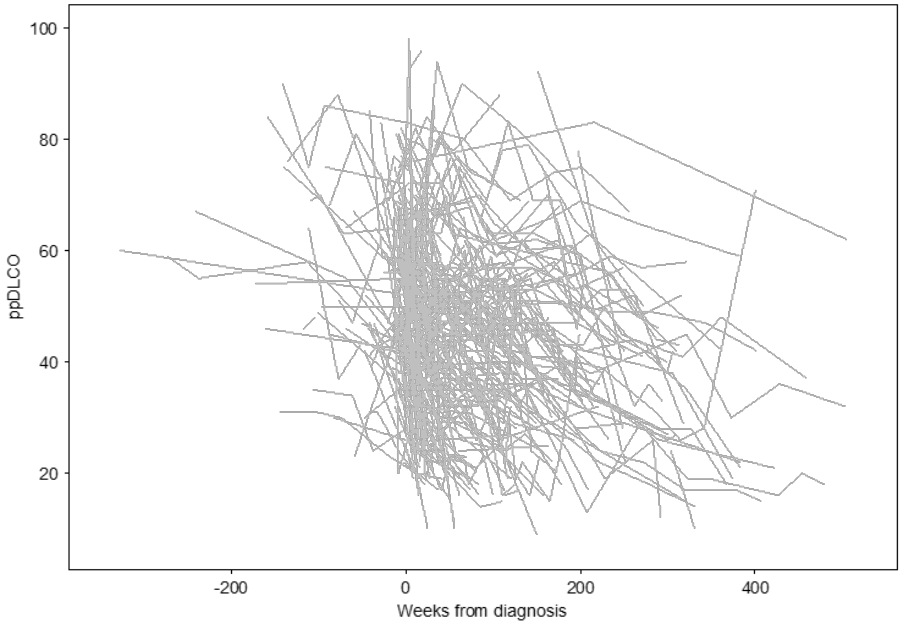
The 'group=id' option allows for the spaghetti noodles for subjects, while the 'grouplc=group' option tells SAS to allow for different colors by group.



Graphs for large amounts of data

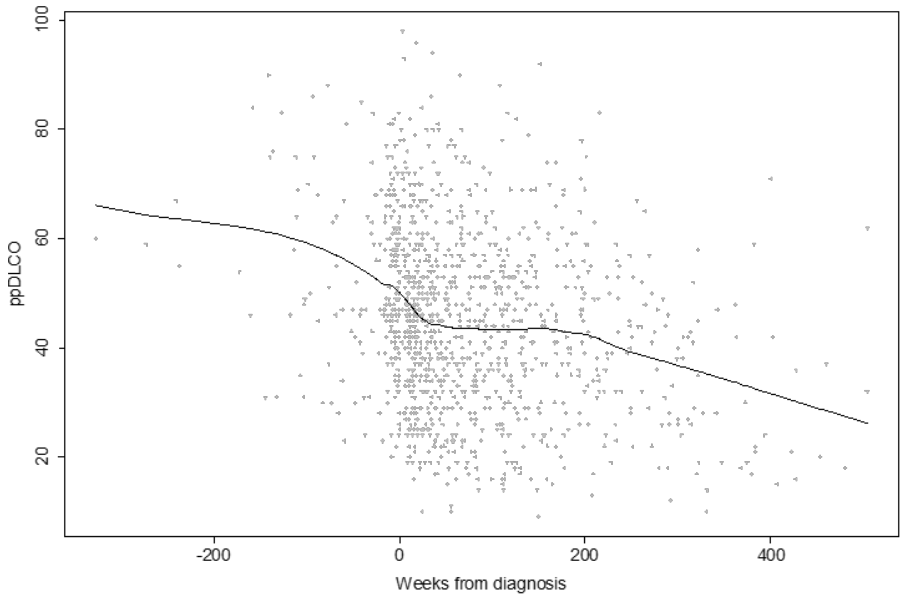
- With large amounts of longitudinal data, a question arises as to the best way to present the data for visual appeal and to best allow for interpretations.
- Diggle, et al., (Analysis of Longitudinal Data; 1994, 1996) discuss approaches to create graphs for a large data set from the Multicenter AIDS Cohort Study (MACS).
- Some of Diggle et al.'s graphing concepts are used here, for data involving subjects with idiopathic pulmonary fibrosis (IPF) that I analyzed at NJH.
- The outcome '% predicted diffusing capacity of the lung' was measured on 321 IPF subjects, both before and after diagnosis. This measure tends to decrease as subjects progress in their illness (see Strand et al., 2014).

Typical spaghetti plot for the data.



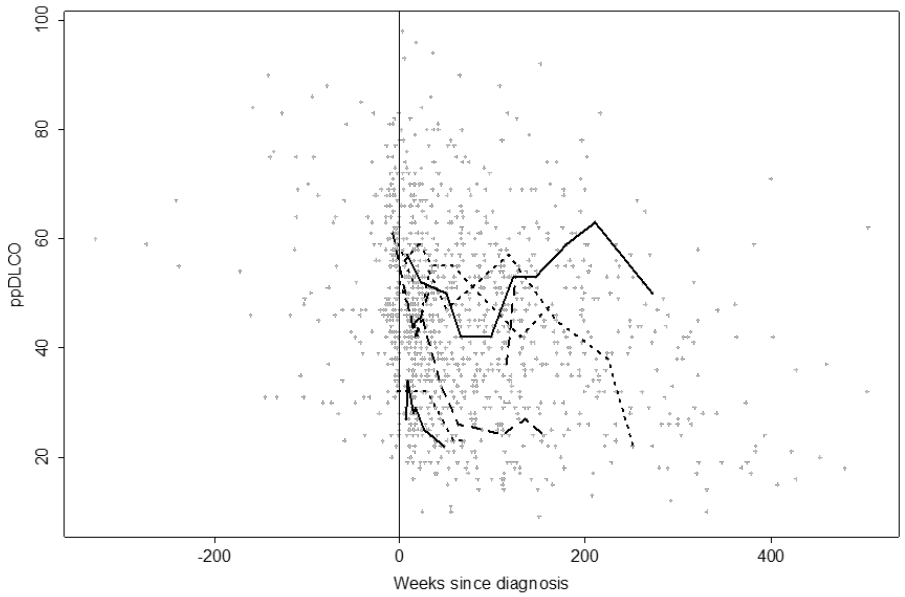
Alternative local polynomial regression

An alternative to the spaghetti plot is to use symbols for subject-day values rather than connecting them, and then overlaying the mean function. Here, local polynomial regression was used to get the fitted function, using order 1 and a span parameter value of 0.5.



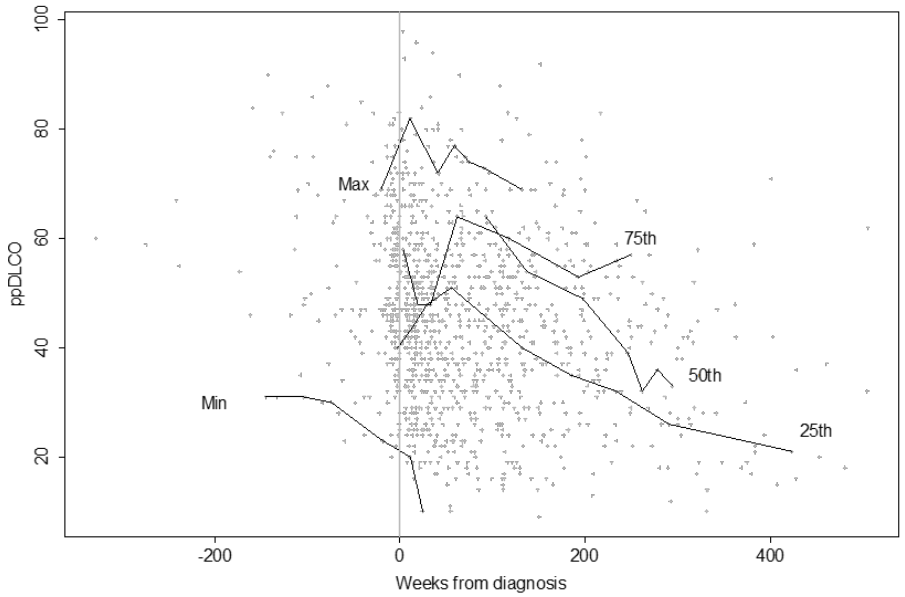
Scatterplot with selective subject trajectory

Scatterplot of IPF data with line graph of 9 randomly selected subjects.



systematically selected subjects

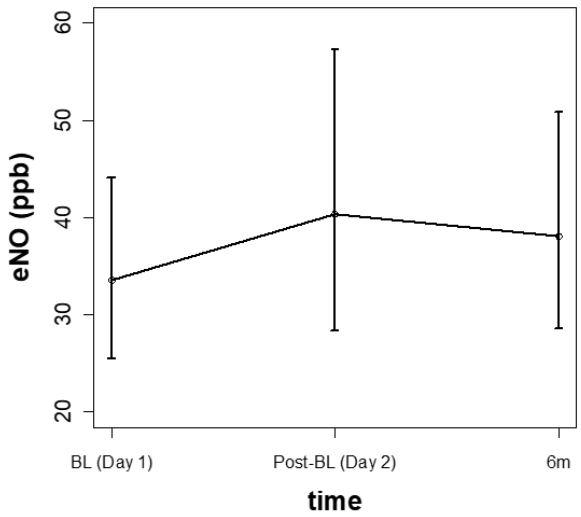
Scatterplot of IPF data with line graph of systematically selected subjects.



Graphs that demonstrate between-subject or within-subject variability

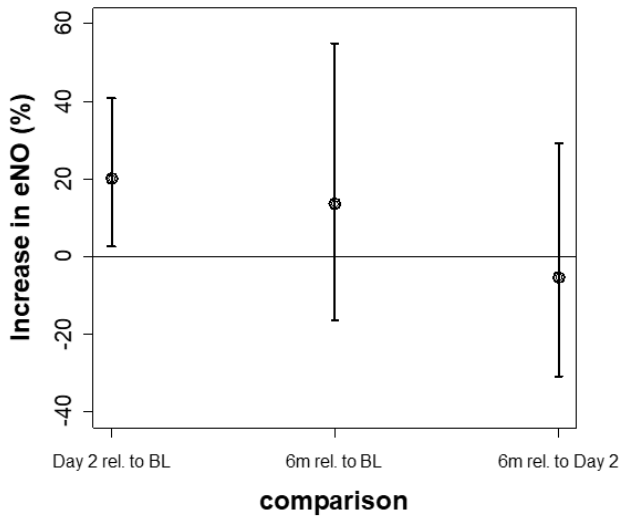
- Mean estimates at individual time points are often graphed including 'error bars' to indicate variability of data.
- Consider the following data from a clinical trial (Katial et al., 2010). Subjects allergic to aspirin were given an aspirin desensitization test over 1 day period. Several measures were taken immediately before and after the desensitization, one being exhaled nitric oxide (eNO). In addition, measures were taken again at 6 months. [Three measures were taken on each subject for several subjects: BL, post-BL, 6m.]
- The following graph displays estimates at individual time points (time as class variable), with confidence intervals, based on a linear mixed model fit. If we perform tests for differences between time points, we find that $p = 0.025$ for the difference between the first and second time points, and $p = 0.40$ between the first and third.

Line graph of means with CI's, eNO data



Question: are the means at BL and Post-BL time points significantly different?

- The previous graph does not demonstrate variability of within-subject changes over time that may be quite different than the SDs of the individual time points.
- This graph shows the variability of the difference estimates. Graphed are relative change estimates, which result since analysis of eNO was on the natural log scale; also plotted in the graph are 95% CI's for these relative estimates.



- In the first position of the graph we have the Day 2 estimate relative to BL, and the CI does not contain 0, which is consistent with $p = 0.025$ (since we constructed a 95% CI).
- In this graph I do not join the difference estimates with a line since the x-axis is not time, but rather it is the comparison of pairs of time points (one relative to another). A reference line at $y = 0$ is included.

Lasagna plots

While we're on the topic of Italian food, a former student of mine got creative and developed the lasagna plot as an alternative to the spaghetti plot (Also see Lasagna plots: a saucy alternative to spaghetti plots, Bruce Swihart et al., 2010, Epidemiology 21: 621-625.)

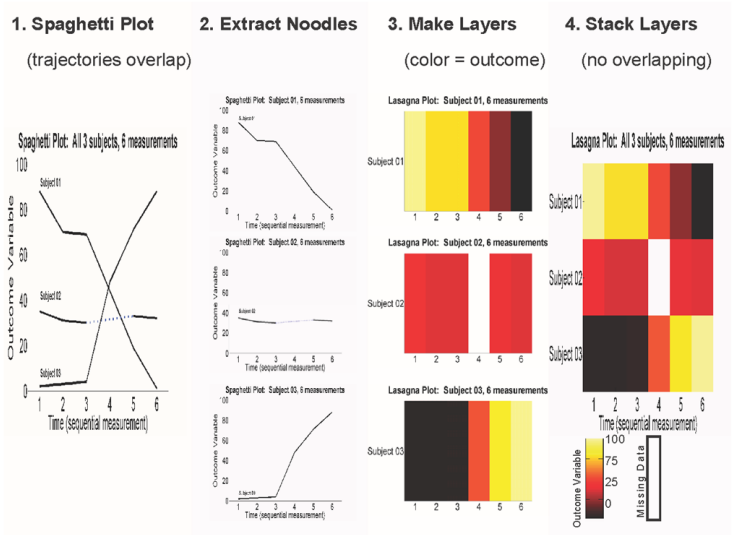
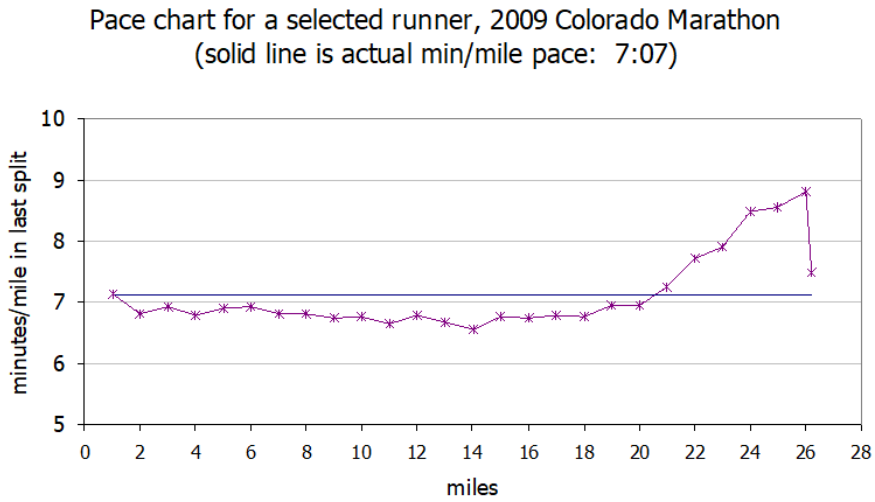


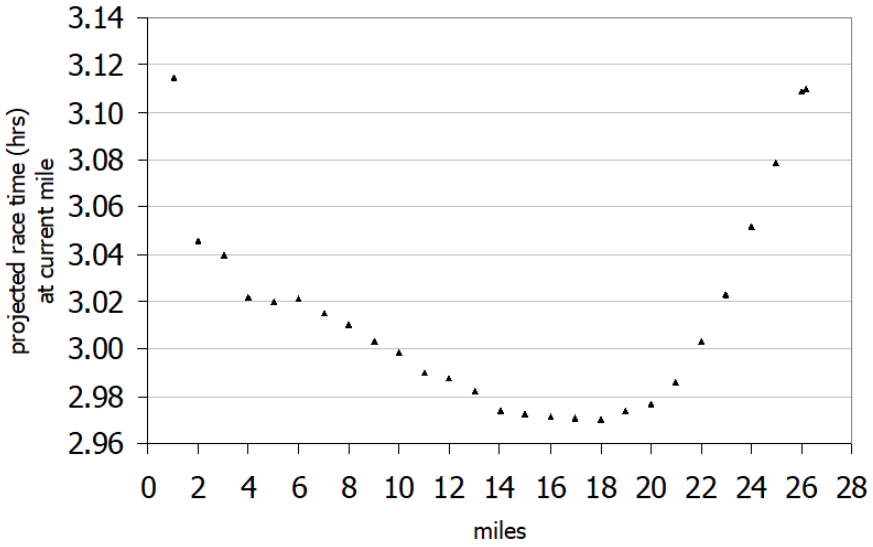
Figure 1: Lasagna plots as derived from spaghetti plots involve making noodles into layers. From left to right, a spaghetti plot with three noodles where trajectories overlap. Extracting each noodle representing repeated measures on a subject, a layer is made by letting color represent the outcome. Individual layers are then stacked to make a lasagna plot, with no overlapping of subject information.

Pace charts

The last "0.2" miles of the 26.2 mile race was adjusted per mile distance, and shows that although the runner 'hit the wall' in the last 3 to 5 miles, he was able to finish strong. The finish time was 3 hours and 6 minutes.



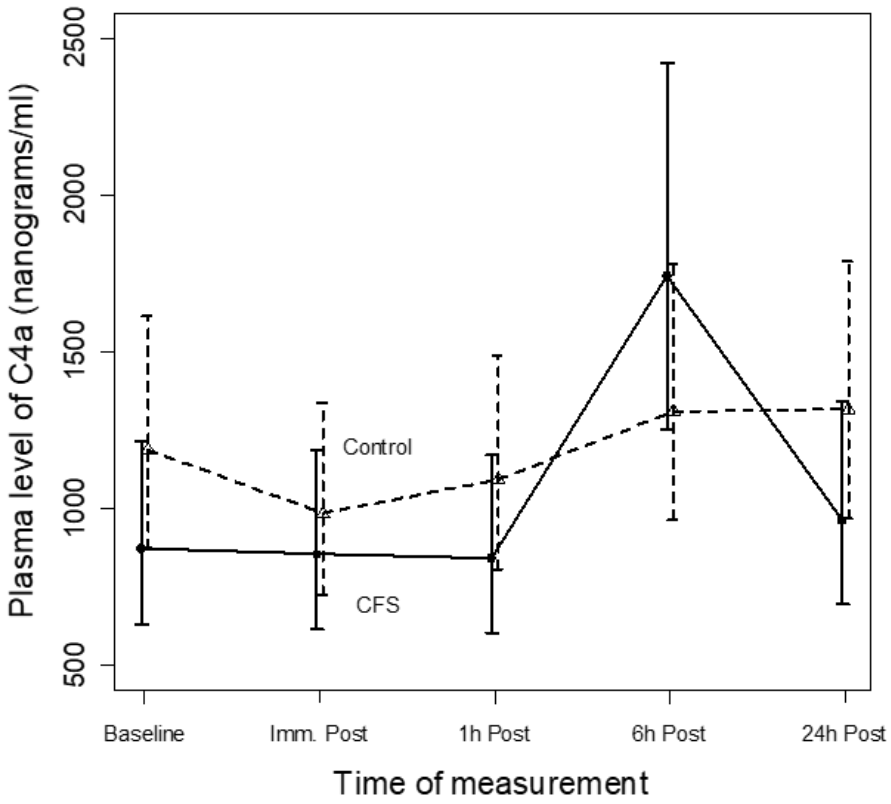
Projected finish time by time at current distance, for the selected runner in the 2009 Colorado Marathon.



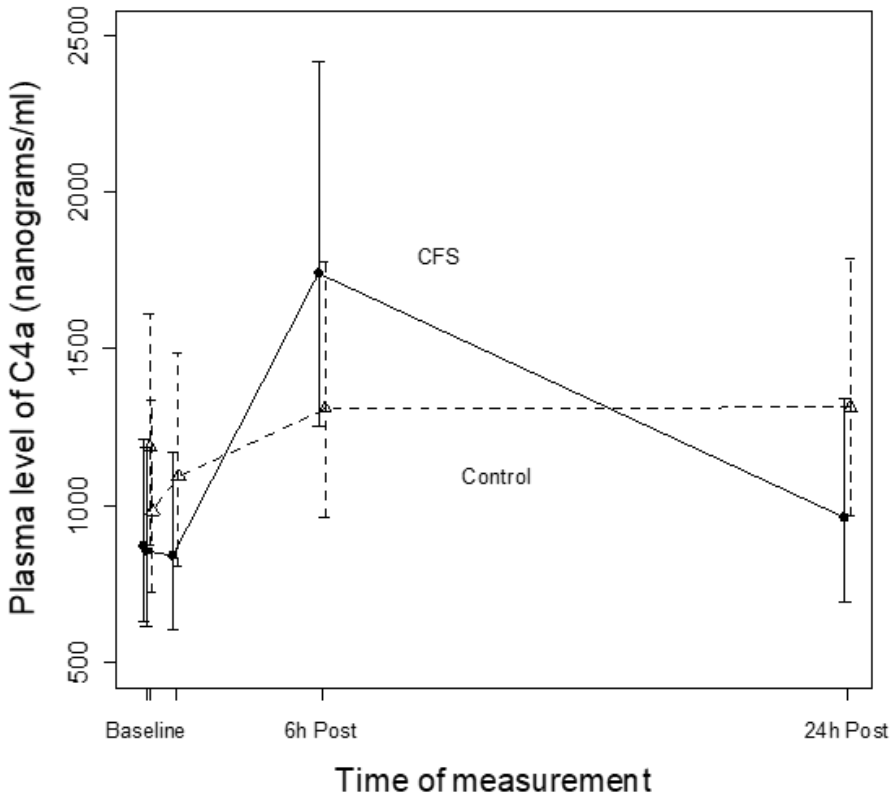
Graphs for unequally spaced data with common time points

- A longitudinal experiment was conducted by Sorensen et al. (2003, JACI) where measurements were taken at unequally spaced times.
- This experiment involved complement split products, which are biological markers measured in the body that may be related to symptoms of chronic fatigue syndrome.
- This research aimed at determining which complements correlated with symptoms induced with exercise and allergen challenges. One such complement was "C4a".
- Estimates of geometric mean C4a levels before and after exercise challenge for chronic fatigue syndrome (CFS) and Control populations, are presented in the following graphs, with 95% confidence intervals. Data were analyzed on the log scale and then inverted back for presentation, resulting in a longer upper bars than lower.

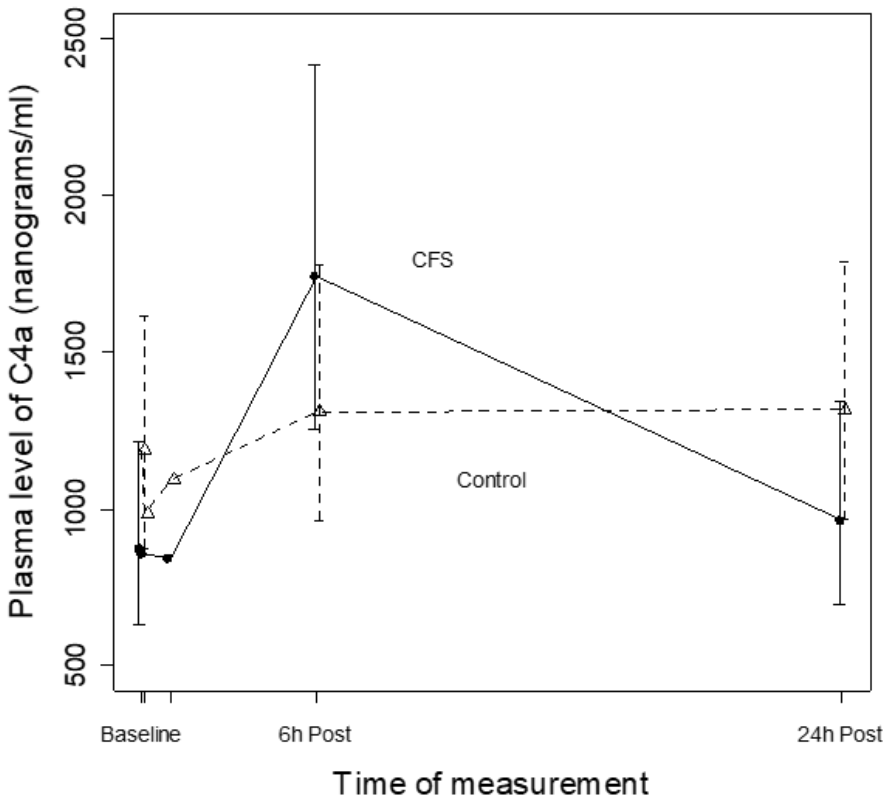
CFS data, time points presented as equally-spaced categories



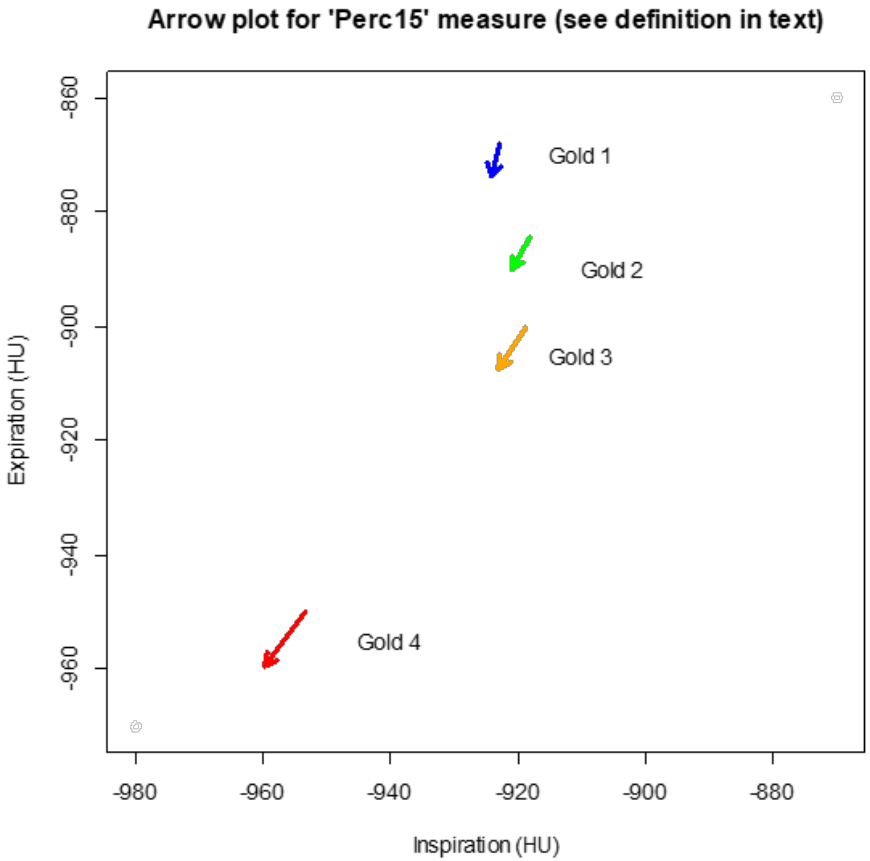
CFS data, time as metric variable. This is a time-metric sensitive graph with the same data. Clearly the concentration of data on the left side makes it difficult to see what is going on.



This is the same basic display, but suppressing CI's for the 2nd and 3rd time points. Which of the 3 graphs is best?



Arrow plots


Arrow plot for 'Perc15' measure (see definition in text)
Gold 1
Gold 2
Gold 3
Gold 4
Expiration (HU)
-860
-880
-900
-920
-940
-960
-980
-960
-940
-920
-900
-880
Inspiration (HU)

Fundamentals of PCA

- Principal components analysis as a descriptive tool for longitudinal data
- PCA becomes particularly useful for very large data sets, as a data reduction technique or to find important patterns.
- Used in genetic data analyses, pattern recognition data, growth curve analysis, and even with meteorological data to identify important climate change patterns.
- Related to factor analysis (FA). In FA, the primary goal is to determine latent 'factors' in the data. While PCA tends to be more of a descriptive technique, FA uses factor rotations of create a reduced set of factors that typically have even stronger patterns than PCs; the remaining unexplained variation is attributed to error.

Eigendecomposition

- Let $\boldsymbol{Y} = (Y_1,\ Y_2,\ ...,\ Y_r)^\top \sim \mathcal{N}ormal_r(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$$\sum_{i=1}^{r} Var[Y_i] = trace(\boldsymbol{\Sigma}) = trace(\Lambda) = \sum_{i=1}^{r} Var[PC_i]$$

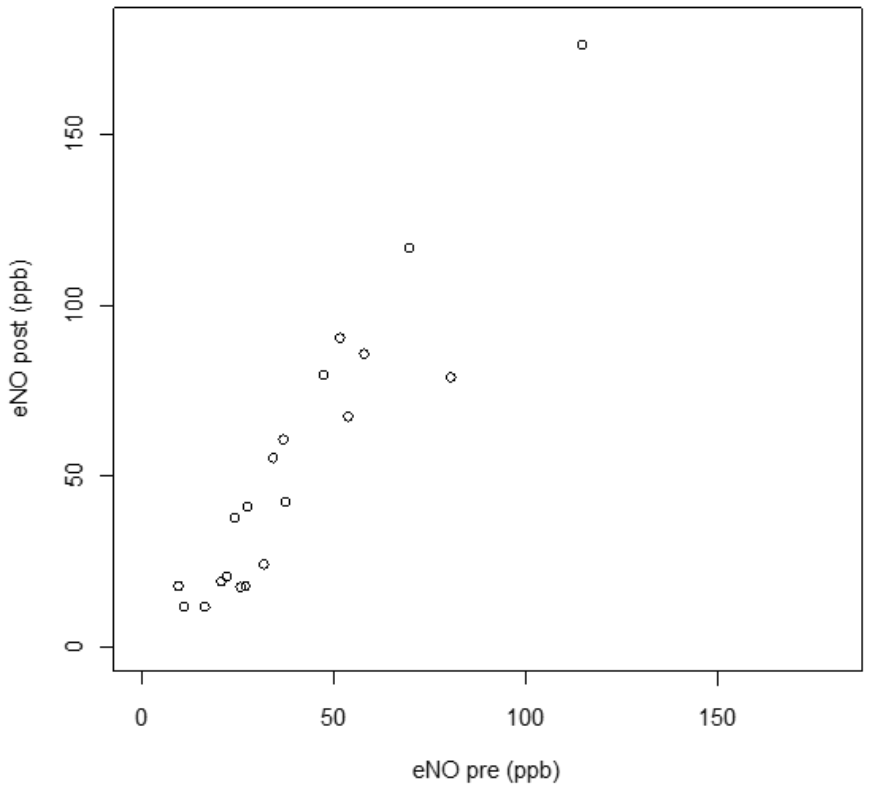
- where $\boldsymbol{\Sigma}_{r\times r} = \boldsymbol{P}_{r\times r}\boldsymbol{\Lambda}_{r\times r}\boldsymbol{P}^\top_{r\times r}$, $\boldsymbol{P} = (\boldsymbol{e}_1,\ ...,\ \boldsymbol{e}_r)$; $\boldsymbol{\Lambda}$ is the diagonal matrix of eigenvalues, $\boldsymbol{e}_i$s are eigenvectors. **(check the eigen-decomposition section in the Matrix notes)**
- The quantity $\frac{\lambda_i}{\sum \lambda_j}$ indicates the proportion of variability in the data accounted for by PC_i.
- In principal components analysis:
  - Eigenvalues indicate magnitude of variances of the principle components (PC's)
  - Eigenvectors indicate direction of the PC's.

Applications: Aspirin/eNO data

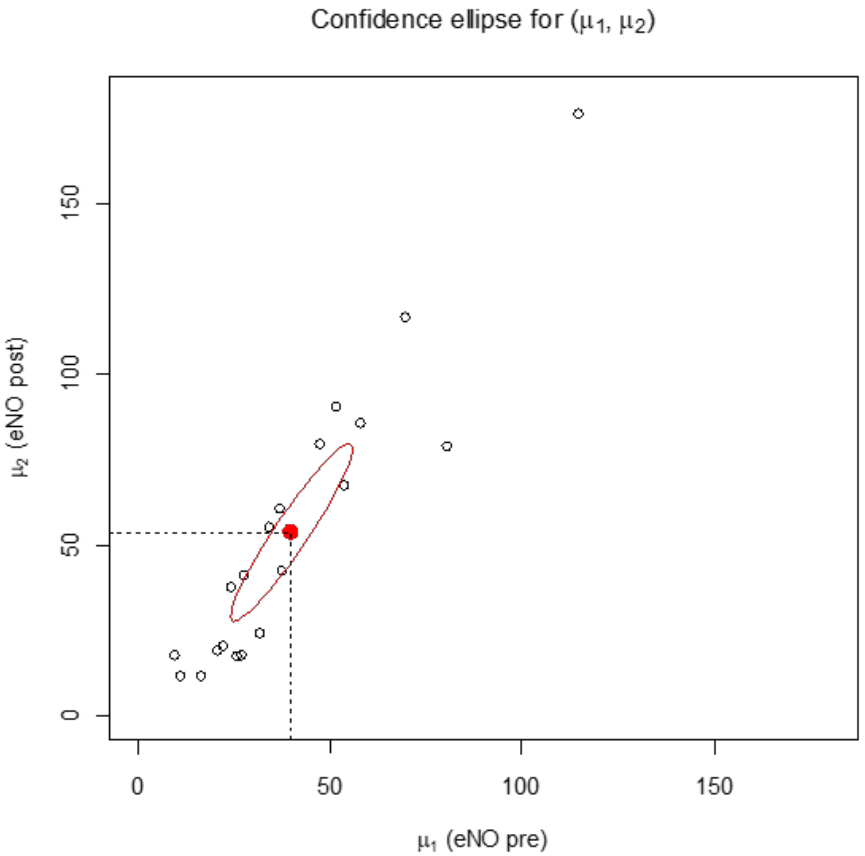
- Aspirin/eNO data, pre and post-aspirin challenge variables
  - Only 2 variables, hence only 2 principle components
  - Somewhat unusual to perform a PCA on only 2 variables
  - Done here primarily for pedagogical purposes, although even a PCA for descriptive analysis purposes that uses only 2 variables can be helpful!
- $PC_1 = \boldsymbol{e}_1^\top \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$, $PC_2 = \boldsymbol{e}_2^\top \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$, where $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ are the eigenvectors associated with λ_1 and λ_2, respectively, where $\lambda_1 \leq \lambda_2$, and Y_1 and Y_2 are the original variables.

$$\begin{cases} PC_1 = 0.51Y_1 + 0.86Y_2 \\ PC_2 = -0.86Y_1 + 0.51Y_2 \end{cases}$$

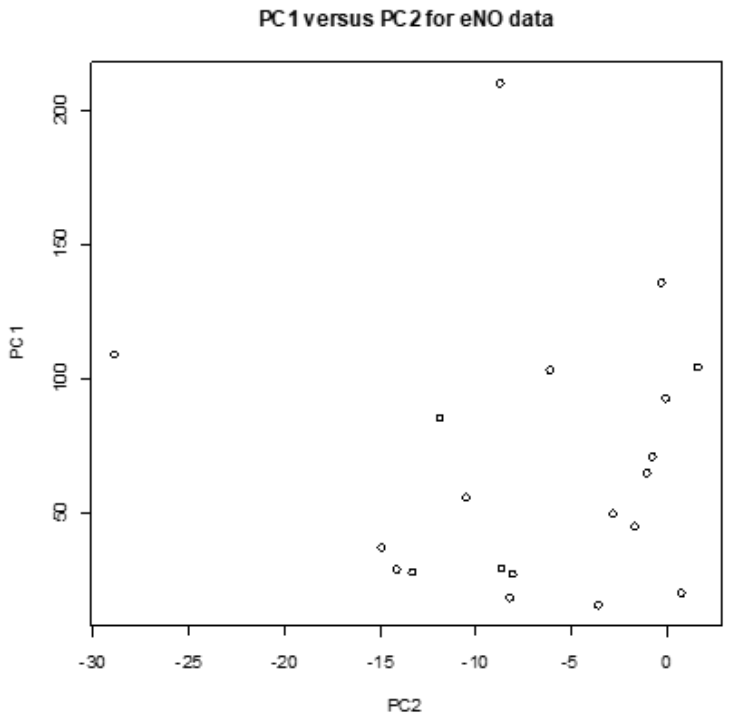
Aspirin challenge data. These are graphs of pre and post challenge eNO values for subjects with aspirin allergies. Here is the same graph, with a confidence ellipse superimposed. Note that the long side of the ellipse extends in the direction of PC1, while 90 degrees to that is PC2.



- $PC1$ is a weighted average eNO value (with a higher weight given to the post measurement, since it contained more variability); well call $PC1$ the 'average' component.
- $PC2$ differentiates pre $(Y1)$ and post $(Y2)$ eNO values; subjects with relatively low values did not react as strongly to the aspirin challenge, while subjects with higher values had post $(Y2)$ measurements that were much higher than pre measurements. For this reason, we can call $PC2$ the 'reactivity' component.



This scatterplot is really the same as the previous one (with the confidence ellipse); it is just tilted and stretched. However, it allows us to see some patterns that we wouldn't otherwise see so easily. In particular, the subject to the far left could be considered an outlier on the reactivity component ($PC2$). If we go back to the original values, we see that their pre eNO value was 80.5, and post value was 79.1, which is unusual because after the aspirin challenge, we would expect most subjects to increase in eNO, particularly those with higher starting values. (Some other subjects also had drops in eNO, but they were ones that had smaller pre eNO values – the lower middle scores on the plot.



The data shows that in fact it was unusual. Those on the far right were more common. Another point that stands out is the high point on $PC1$ – the subject had a very 'average' eNO value and did in fact increase from pre to post.

Note that there are several different ways that PC's can be standardized. For example, in SAS, PC's are mean corrected. In the previous plot, no standardization was done.

The Ramus data

The Ramus data (presented earlier) involves measurements of the Ramus bone in the jaw for boys. Each boy was measured at 8, 8.5, 9, and 9.5 years. Here, we have 4 variables, which are the measurements at each of the ages. The following SAS code can be used to carry out a standard PCA. The output and relevant graphs follow.

```
proc princomp data=ramus out=ramus_out; var h1 h2 h3 h4; run;
proc gplot data=ramus_out; plot prin1*prin2; run;
proc gplot data=ramus_out; plot prin2*prin3; run;
```

The PRINCOMP Procedure

Observations 20
Variables 4

Simple Statistics

	h1	h2	h3	h4
Mean	48.6550000	49.6250000	50.5700000	51.4500000
StD	2.5159439	2.5395555	2.6302091	2.7321671

Correlation Matrix

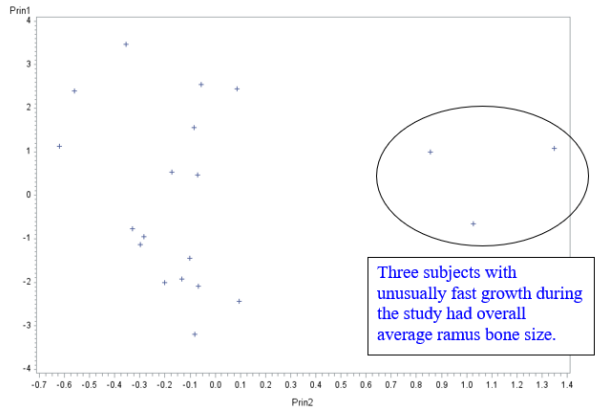
	h1	h2	h3	h4
h1	1.0000	0.9687	0.8730	0.8071
h2	0.9687	1.0000	0.9212	0.8537
h3	0.8730	0.9212	1.0000	0.9666
h4	0.8071	0.8537	0.9666	1.0000

Eigenvalues of the Correlation Matrix

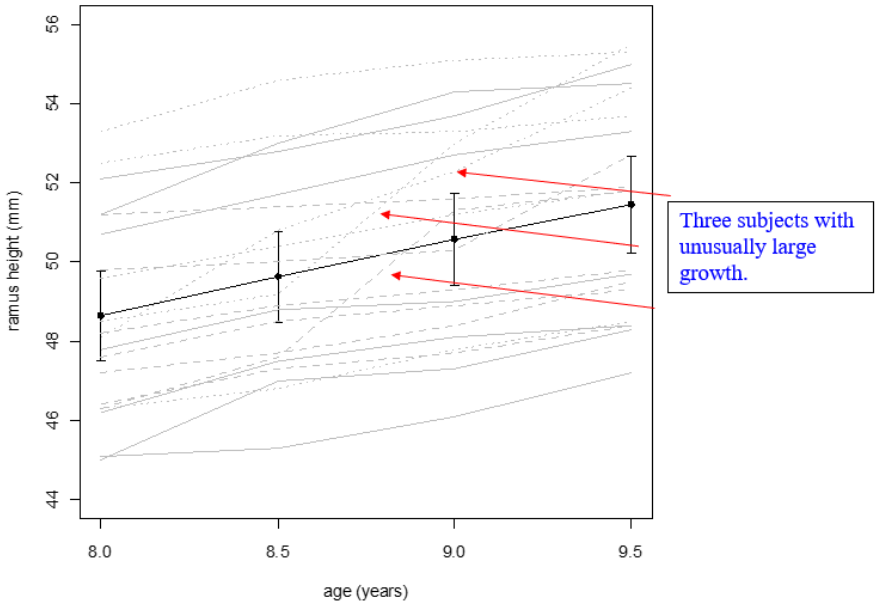
	Eigenvalue	Difference	Proportion	Cumulative
1	3.69607865	3.44097766	0.9240	0.9240
2	0.25510100	0.22304846	0.0638	0.9878
3	0.03205254	0.01528473	0.0080	0.9958
4	0.01676781		0.0042	1.0000

Eigenvectors

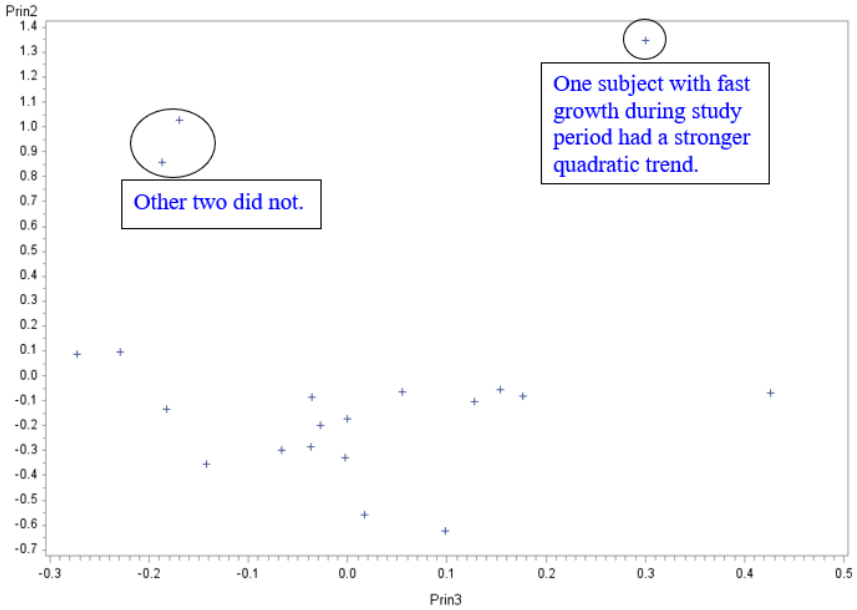
	Prin1	Prin2	Prin3	Prin4
h1	0.493661	-.585366	0.567059	-.303462
h2	0.506595	-.380976	-.535784	0.557813
h3	0.508856	0.339895	-.442819	-.655323
h4	0.490639	0.629822	0.441917	0.409033



Original line graph of data, with markers to 3 subjects with unusually large growth. These subjects are the same as those 3 on the right side of the previous graph.



Plot of PC2 versus PC3. This graph further breaks down the kids with unusual large growth into those with a quadratic trend (1 subject) and those without stronger quadratic trend (2 on the left). Also see the previous graph.



We can also identify subjects with stronger values of PC3 and PC4 on the original line graph. (Other subjects are removed in order to see patterns more clearly.

This PCA allowed us to see quickly subjects with more unusual trends. It also showed us that the variability in the data is captured through orthogonal polynomial trends, with decreasing variability as the order increases (from 'intercept' to cubic); nearly 99% of the between-subject variability could be captured by the 'intercept' and 'linear' components.