

NCGS (Fitzmaurice exercise 5.1)

We consider data from the National Cooperative Gallstone Study (NCGS). In this study patients were randomly assigned to high-dose (750 mg/day) or low-dose (375 mg/day) of the drug chenondiol or to a placebo. We focus on a subset of data on patients who had floating gallstones and who were assigned to either the high-dose or the placebo group.

In the NCGS it was suggested that chenondiol would dissolve gallstones but in doing so might increase levels of serum cholesterol. As a result serum cholesterol (mg/dL) was measured at baseline and at 6, 12, 20, and 24 months of follow-up. Note that many cholesterol measurements are missing due to missed visits, drop out, or missing or inadequate laboratory specimens.

Note the groups: 1=high dose, 2=placebo.

1. Read in the data. In SAS by opening and running the program file `ncgs.sas`. In R similarly by running the program `ncgs.R`.
 - How many variables does the data contain? What are they called?
 - Is the data in the *long* or in the *wide* format?
 - How many observations in total does the dataset contain?
2. Make summary statistics and scatterplots for each treatment group as exemplified in the lecture.
 - Does it seem reasonable to assume that the repeated serum cholesterol measurements follows a multivariate normal distribution?
 - Is there a time-trend in the mean-cholesterol levels within the two groups?
 - Is there a time-trend in the variances of cholesterol within the two groups?
 - Is there a time-trend in the correlations between measurements at different time points?
3. Before conducting any further analyses, data must be transformed to the *long format*. Check the SAS- or R-notes to see how to do this.
4. Make two spaghettiplots showing the data in each group.

5. Construct a plot of the response profiles for the two groups showing the sample means for each occasion. Describe the time trends in each group.
6. Make an analysis to compare the response profiles between the groups as in a **parallel group study**. Note that this is just for the exercise, we **pretend** that treatment was not randomised (as in an observational study).
 - What are the estimated mean changes from baseline to each follow-up in the placebo group? And in the high dose group? Provide estimates for the differences between these with 95% confidence intervals.
 - Does the overall pattern of change over time differ significantly between the groups? I.e. are the response profiles parallel?
 - What is the estimated difference in means between the groups at baseline? Is this an interesting difference?
 - Save the predicted values from the model in an output dataset and use these data to construct a plot of the predicted response profiles. Compare this to the plot of response profiles based on the sample means. Can you guess why these are almost but not exactly the same?
7. Since the NCGS study was actually randomized, re-analyze the data using the constrained model from the lectures. Hint: Start by adding a new variable `treatment` to the data. This should be a copy of the `group` variable except that all records at baseline should have `treatment=2` because the de facto treatment at this time is similar to placebo for all groups.
 - What are the estimated mean changes from baseline to each follow-up in the placebo group? And in the high dose group? Provide estimates for the differences between these with 95% confidence intervals.
 - Does the overall pattern of change over time differ significantly between the groups? I.e. are the response profiles identical?
 - Save the predicted values from the model in an output dataset and use these data to construct a plot of the predicted response profiles. Compare this to the plot of response profiles in question 5.