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. The data is contained within the sas-program file `ncgs.sas`.

In the NCGS it was suggested that chenodiol 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. Open the program file `ncgs.sas` in SAS and run it by either pressing "Run"(enterprise guide) or the button with the running man (SAS 9.4 or earlier versions). This will generate the sas-dataset `ncgs`.

   - How many variables does the data contain? What are they called?
   - Is data in the *long* or in the *wide* format?
   - Switch to the log-window. Are there any error messages or warnings?
   - How many observations in total does the dataset contain?

   Scroll to the buttom of the program file to start writing your own sas-code. Don't forget to save the program every time you have added a new part.

2. Use `proc corr` to construct 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?

To conduct further analyses we will have to transform data to the *long format*. This can be done using the following code:

```
data ncgslong (drop = y1-y4); set ncgs;
  month = 0; chol = y0; output;
  month = 6; chol = y1; output;
  month = 12; chol = y2; output;
  month = 20; chol = y3; output;
  month = 24; chol = y4; output;
  run;
```

3. Make two spaghettiplots showing the data in each group.

4. Construct a plot of the response profiles for the two groups showing the sample means for each occation. Describe the time trends in each group.

   We next turn to the analysis of response profiles. The NCGS study was a randomised study so we ought to do baseline adjustment. However, to exercise the general analysis of response profiles we will first conduct an analysis pretending that treatment was not randomised (as in an observational study).

5. Conduct an analysis of response profiles using `proc mixed` as was done in the lectures.

   - 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?
   - What is the estimated mean change from baseline to final follow-up in the placebo group? And in the high dose group? Provide an estimate for the difference between these with a 95% confidence interval.
   - Save the predicted group means from the model in an output dataset (`outpm=ncgsfit`). 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?

The last two questions about baseline adjustment are optional. **If your own data are from a randomized study you should do them!**

Since the NCGS study was indeed randomized, do an analysis of response profiles based on the constrained model from the lectures. Hint: Start by adding the variable `treat_adj` to the data:

```
data ncgsajdust;
set ncgslong;
treat_adj = treat;
if month = 0 then treat_adj = 2;
run;
```

6. Run `proc mixed` to conduct the analysis of response profiles with baseline adjustment. I.e. with the model defined by

   ```
   model chol = month treat_adj*month / ....
   ```

   - Does the overall pattern of change over time differ significantly between the groups? I.e. are the response profiles identical?
   - What is the estimated mean change from baseline to final follow-up in the placebo group? And in the high dose group? Provide an estimate for the difference between these with a 95% confidence interval.
   - Save the predicted group means from the model in an output dataset (`outpm=ncgsfit`). Use these data to construct a plot of the predicted response profiles. Compare this to the plot of response profiles in question 5.

7. Finally, define a new dataset consisting only of the post baseline measurements. Run `proc mixed` on these data to do the analysis of response profiles using the baseline measurement (`y0`) as a covariate.

   - What is the expected difference between a person on high dose and a person on placebo at final follow-up given that they had the same baseline value. Find a 95% confidence interval for this estimate and compare to question 6.
   - Save the predicted values from the model in an output dataset. Construct a plot of these against time for each of the treatment groups. Why is this plot different from the ones in questions 5 and 6?