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 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. 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?

     The data contains the treatment variable `group`, with value `1` for the high-dose and value `2` for placebo treatment; the variable `id` which is a numbering of the patients in the study; the measurement `y0` at baseline; and the measurements `y1`-`y4` at 6, 12, 20 and 24 months of follow-up. Below is the output from a `proc contents` in `SAS`.

     ```
     # Variable Type Len
     2 id       Num  8
     1 group    Num  8
     3 y0       Num  8
     4 y1       Num  8
     5 y2       Num  8
     6 y3       Num  8
     7 y4       Num  8
     ```

   - Is data in the *long* or in the *wide* format?

     The data is in the wide format. Measurements from each of the time points are included in the data as a separate column.

- How many observations in total does the dataset contain?

  When reading in the data, we get the following output in the **log-window**:

  ```
   data ncgs;
   input group id y0-y4;
   datalines;

  NOTE: The data set WORK.NCGS has 103 observations and 7 variables.
  NOTE: DATA statement used (Total process time):
        real time           0.00 seconds
        cpu time            0.00 seconds
    ;
   run;
  ```

  telling that the dataset ncgs in the work-library was created sucessfully and that it contains 103 observations in total.

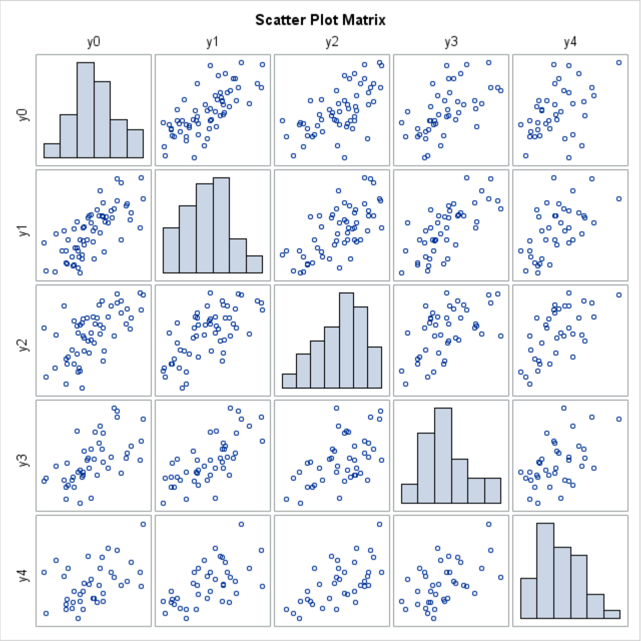
Now it's time to scroll to the buttom of the program file to start writing your own sas-code. Don't forget to save the program every time you have added a new part.

2. We can use `proc corr` to construct summary statistics and scatterplots for each treatment group as exemplified in the lecture.

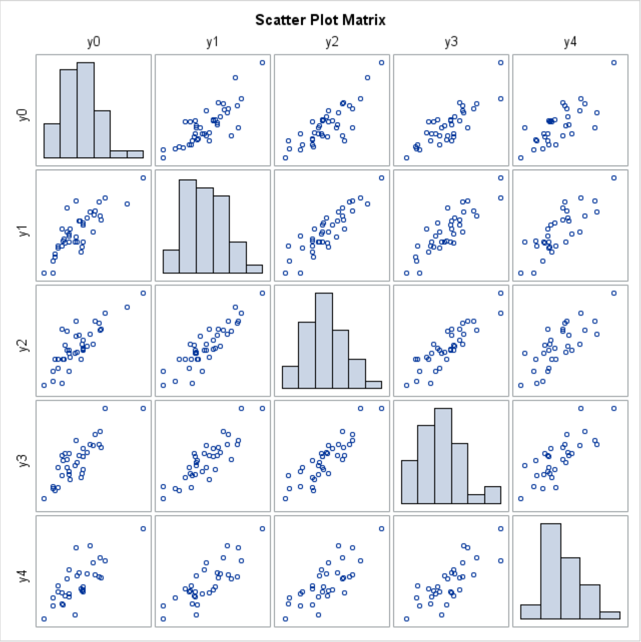
   ```
   proc sort data = ncgs; by group; run;

   ods graphics on;
   proc corr data = ncgs plots=matrix(histogram) noprob;
   by group;
   var y0-y4;
   run;
   ```

   - Does it seem reasonable to assume that the repeated serum cholesterol measurements follow a multivariate normal distribution?

     We look at the `Scatter Plot Matrix`, cf. figure 1 and figure 2. Within each of the two treatment groups the histograms at each time-point do not deviate substantially from the normal distribution. Further, all of the scatterplots look reasonably elliptical. All in all we find no appearant reason to doubt the multivariate normal distribution.



Figur 1: Treatment with high-dose.



Figur 2: Placebo treatment.

- Is there a time-trend in the mean-cholesterol levels within the two groups?

  To compare the mean-cholesterol levels within the two groups, we look at the tables of summary statistics.

  For `group=1`:

```
Simple Statistics
Variable N  Mean   Std Dev Sum   Minimum   Maximum
y0       62 226.02 39.66   14013 144.00000 313.00000
y1       62 245.53 39.45   15223 177.00000 334.00000
y2       55 252.02 38.33   13861 167.00000 316.00000
y3       44 256.80 34.49   11299 194.00000 334.00000
y4       38 254.55 49.96   9673  172.00000 397.00000
```

  For `group=2`:

```
Simple Statistics
Variable N  Mean   Std Dev Sum  Minimum Maximum
y0       41 235.93 55.875  9673 141.00  418.00
y1       41 243.17 49.240  9970 142.00  371.00
y2       38 244.76 46.111  9301 157.00  363.00
y3       35 257.60 51.142  9016 162.00  384.00
y4       31 257.48 49.388  7982 169.00  387.00
```

  In both groups mean levels of cholesterol tend to increase with time.

- Is there a time-trend in the variances of cholesterol within the two groups?

  We do not see any systematic trend in the standard deviations in either group, but the overall variability appears to be somewhat higher in the placebo group. This seems to be due to the (random) assignment of the most outlying patients to the placebo group.

- Is there a time-trend in the correlations between measurements at different time points?

  We look at the `Pearson Correlations` in the output:

  For `group=1`:

```
Pearson Correlation Coefficients
Number of Observations
```

```
          y0        y1        y2        y3        y4
y0   1.00000   0.72034   0.62267   0.59078   0.45819
          62        62        55        44        38
y1   0.72034   1.00000   0.66953   0.71531   0.58330
          62        62        55        44        38
y2   0.62267   0.66953   1.00000   0.53743   0.63632
          55        55        55        43        36
y3   0.59078   0.71531   0.53743   1.00000   0.51410
          44        44        43        44        37
y4   0.45819   0.58330   0.63632   0.51410   1.00000
          38        38        36        37        38
```

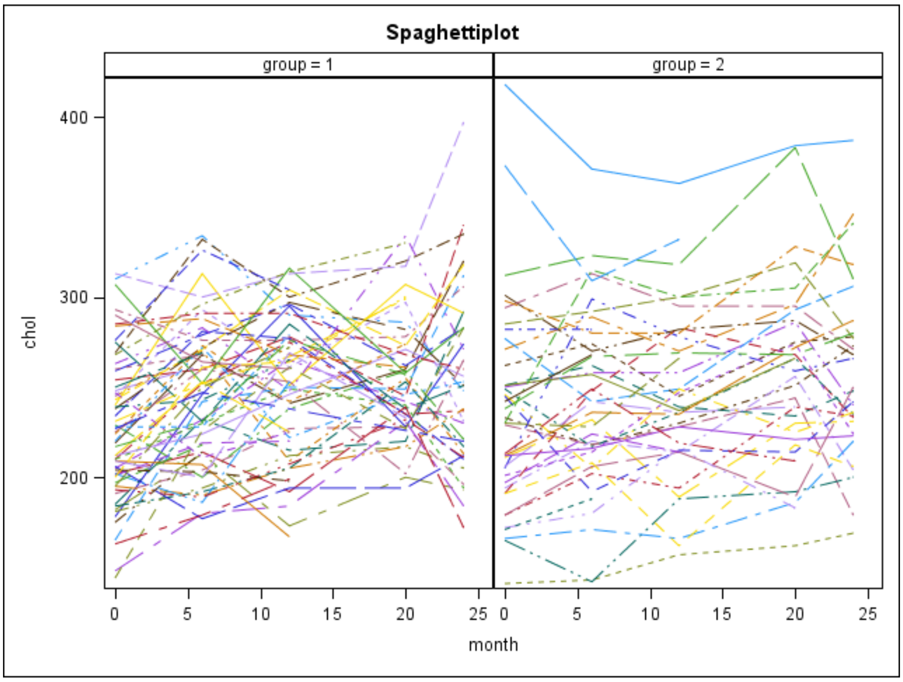
For `group=2`:

```
Pearson Correlation Coefficients
Number of Observations
          y0        y1        y2        y3        y4
y0   1.00000   0.81613   0.83232   0.84425   0.76128
          41        41        38        35        31
y1   0.81613   1.00000   0.88740   0.86885   0.81910
          41        41        38        35        31
y2   0.83232   0.88740   1.00000   0.87795   0.77765
          38        38        38        35        31
y3   0.84425   0.86885   0.87795   1.00000   0.78924
          35        35        35        35        31
y4   0.76128   0.81910   0.77765   0.78924   1.00000
          31        31        31        31        31
```

Correlations close to the diagonal tend to be the stronger and the smallest correlation is the one between the baseline measurement and the final follow-up measurement which are furthest apart in time. However, the time-trend is not all that strong since the ordering of the correlations is not completely monotone and the correlation between the first and the last measurement is still modestly high. We note that the correlations in the placebo group are overall higher than in the treatment group. Looking at the scatterplot matrices, this seems to be mainly due to the outliers which have randomly ended up in this group.

3. Before we conduct further analyses we 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;
```



Figur 3: Spaghettiplots showing the data in each group.

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

   See figure 3 which is the output from running the following code.

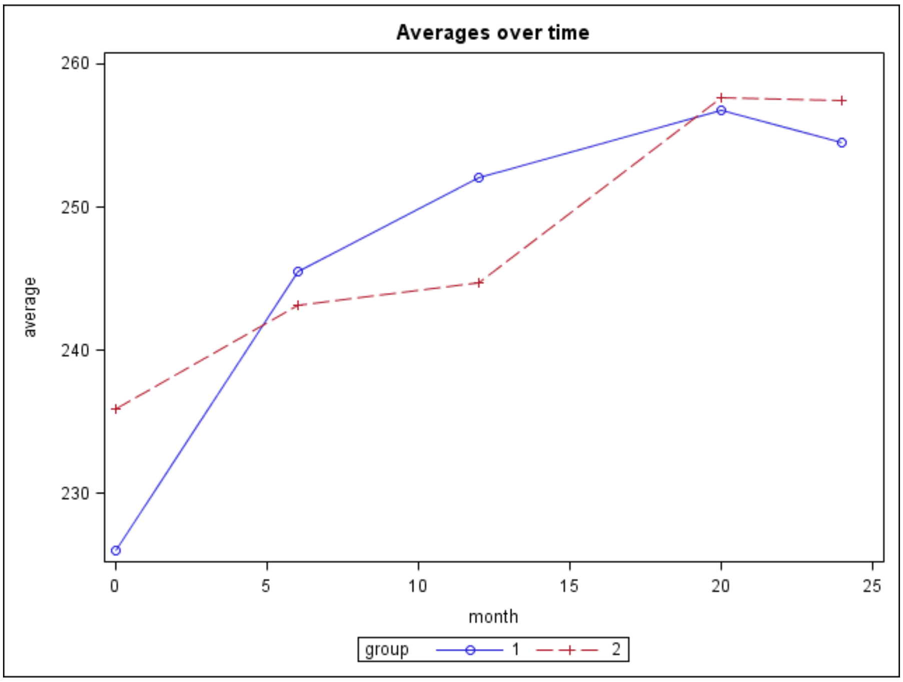
```
proc sgpanel data = ncgslong;
panelby group;
series x = month y = chol / group = id;
run;
```

   Again we see an overall higher variation between subjects in the placebo group. One could worry whether the difference in variance could be caused by the treatment, but since the difference was there already at baseline and persists throughout the study it is most likely caused by the inclusion of a few more outlying subjects in the placebo group. The increasing trend in mean cholesterol levels can hardly be seen from the spaghettiplots. This suggests that a possible increase in serum cholesterol due to treatment will be small compared to the inter-individual variation in cholesterol levels.

5. 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 run the code:

```
proc sort data = ncgslong;
by group month;
run;

proc means nway data = ncgslong noprint;
by group month;
var chol;
output out = ncgsmeans mean = average;
run;

proc sgplot data = ncgsmeans;
series x = month y = average / group = group markers;
run;
```



Figur 4: Plot of the sample means from the two groups.

On figure 4 we see how the average changes across the time-points for each of the two treatment groups. For both groups we see an increasing trend, with the biggest change for the high-dose group. We note that the two groups intersect in each end of the time interval, such that the placebo group has the highest average in the beginning and in the end of the study, whereas the high-dose group has higher measurements in between. An appealing interpretation is that treatment temporarily increases serum cholesterol but that the effect has vanished again after 20-24 months. However, we need to do a formal statitical analysis to make sure that this is not just a chance finding.

Note that, due to the many persons who drop out during the study the plot of averages against time may not give an accurate picture of the potential treatment effect. We will return to this problem later in the lectures.

6. The NCGS study was a randomised study so we ought to do baseline adjustment. However, to start out more gently on the exercise, we will first conduct an analysis **pretending** that treatment was not randomised (as in an parallel group study). To do so we run the code from the *Introduction to SAS proc mixed*. Note that the natural reference points are `month=0` and `group=2` (the placebo group).

```
proc mixed data = ncgslong;
class month (ref='0') id group (ref='2');
model chol = month group month*group / solution cl ddfm = kr
                                       outpm=ncgsfit0;
repeated month / type = un R Rcorr subject = id ;
run;
```

```
(Less interesting output omitted)

    Solution for Fixed Effects


                                            Standard
Effect         month   group    Estimate       Error      DF   t Value    Pr > |t|    Alpha       Lower       Upper
Intercept                         235.93      7.3029     101     32.31      <.0001     0.05      221.44      250.41
month              6              7.2439      4.8054     101      1.51      0.1348     0.05     -2.2888     16.7766
month             12              8.8483      5.2118    99.6      1.70      0.0927     0.05     -1.4923     19.1889
month             20             23.1028      5.3114    88.8      4.35      <.0001     0.05     12.5487     33.6569
month             24             21.1238      7.4251    80.5      2.84      0.0056     0.05      6.3487     35.8989
month              0                   0           .       .         .           .        .           .           .
group                      1     -9.9107      9.4128     101     -1.05      0.2949     0.05    -28.5831      8.7617
group                      2           0           .       .         .           .        .           .           .
month*group        6       1     12.2722      6.1938     101      1.98      0.0503     0.05    -0.01457     24.5590
month*group        6       2           0           .       .         .           .        .           .           .
month*group       12       1     16.4175      6.7516     100      2.43      0.0168     0.05      3.0230     29.8121
month*group       12       2           0           .       .         .           .        .           .           .
month*group       20       1      4.9770      7.0113    91.4      0.71      0.4796     0.05     -8.9492     18.9032
month*group       20       2           0           .       .         .           .        .           .           .
month*group       24       1      6.9031      9.8868    82.2      0.70      0.4870     0.05    -12.7642     26.5704
month*group       24       2           0           .       .         .           .        .           .           .
month*group        0       1           0           .       .         .           .        .           .           .
month*group        0       2           0           .       .         .           .        .           .           .
```

```
         Type 3 Tests of Fixed Effects

                Num     Den
Effect           DF      DF    F Value    Pr > F

month             4    85.2      14.36    <.0001
group             1     101       0.05    0.8205
month*group       4    85.2       1.89    0.1195
```

- 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 difference between these with 95% confidence intervals.
  For instance, the estimated mean change from baseline to final follow-up in the placebo group is 21.12. In the high-dose group this figure should be 6.90 higher, i.e. 28.02. The estimated difference in changes between the groups is 6.90 with a confidence interval of (-12.76;26.57). In particular, there is no significance difference in change between the two groups at final follow-up. Note however, that after 12 months the increase in cholesterol levels appears to be significantly higher in the high dose group than in the placebo group (P=0.0168).

- Does the overall pattern of change over time differ significantly between the groups? I.e. are the response profiles parallel?

  To test the null hypothesis that the two response profiles are parallel, we look at the type 3 test of the interaction term `month*group` which has P=0.12. From this we conclude that there is no evidence indicating that the two groups evolve differently over time. Note that we would have concluded differently if 12 months follow-up had been the primary end point.

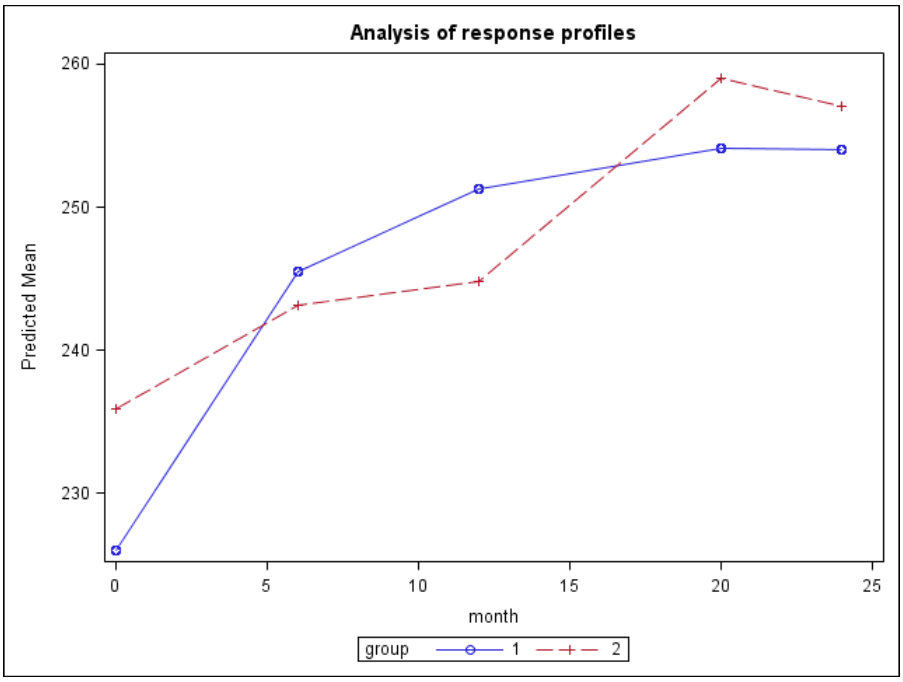
- What is the estimated difference in means between the groups at baseline? Is this an interesting difference?
  Pretending that data are from a non-randomised study, it is plausible that there is a diffence between the baseline means, which is represented by the main effect of `group` in the model. From the output we see that the estimated difference between the groups is -9.91 with a confidence interval of (-28.58;8.76). The difference is not significant (P=0.49). This is of course no surprise since the study was in fact randomised; We **know** that the true difference is zero!

- 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?

  The output from the model is saved in the data `ncgsfit0` (see code above). The plot of the estimated response profiles is made with:

  ```
  proc sort data=ncgsfit0; by group month id; run;
  proc sgplot data = ncgsfit0;
  series x = month y = pred / group = group markers;
  run;
  ```

  The resulting plot and the plot on the sample means are almost but not quite the same. The largest deviation is seen at the last two time points. This is due to the persons dropping out of the study. Drop outs do not contribute to the time-group averages. However, due to the correlation in the data the initial measurements from the drop outs are partially predictive of their future cholesterol levels. This information is taken into account by the mixed model are therefore the predicted means differ from the sample means.



Figur 5: Estimated response profiles for the two groups based on a mixed model.

7. Since the NCGS study was in fact randomized, we finally make the suggested analysis based on the constrained model from the lectures. The first step is to add a new variable `treat` to the data:

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

   Next we run `proc mixed` to conduct the analysis of response profiles with baseline adjustment. I.e. with the model defined by:

   ```
   proc mixed data = ncgsadjust;
   class id month (ref='0') treat (ref='2');
   model chol = month treat*month / solution ddfm = kr outpm=ncgsfit2;
   repeated month / subject = id type = un R Rcorr;
   run;
   ```

   This produces the following parameterestimates and tests.

   ```
                                Solution for Fixed Effects

                                                     Standard
   Effect       month  treat      Estimate      Error      DF    t Value     Pr > |t|      Alpha       Lower       Upper
   Intercept                        229.96     4.6100     102      49.88       <.0001       0.05      220.82      239.11
   month            6               8.9261     4.5570     111       1.96       0.0526       0.05     -0.1039     17.9560
   month           12              10.9739     4.8301     111       2.27       0.0250       0.05      1.4032     20.5447
   month           20              25.1274     4.9793    96.5       5.05       <.0001       0.05     15.2441     35.0106
   month           24              23.4712     7.1223    82.5       3.30       0.0014       0.05      9.3039     37.6385
   month            0                    0          .       .          .            .          .           .           .
   month*treat      6      1        9.4777     5.6515     101       1.68       0.0966       0.05     -1.7334     20.6887
   month*treat      6      2             0          .       .          .            .          .           .           .
   month*treat     12      1       12.8863     5.9160    97.2       2.18       0.0318       0.05      1.1451     24.6274
   month*treat     12      2             0          .       .          .            .          .           .           .
   month*treat     20      1        1.6136     6.3023    87.1       0.26       0.7985       0.05    -10.9127     14.1398
   month*treat     20      2             0          .       .          .            .          .           .           .
   month*treat     24      1        3.0033     9.2511    76.4       0.32       0.7463       0.05    -15.4202     21.4268
   month*treat     24      2             0          .       .          .            .          .           .           .
   month*treat      0      2             0          .       .          .            .          .           .           .
   ```

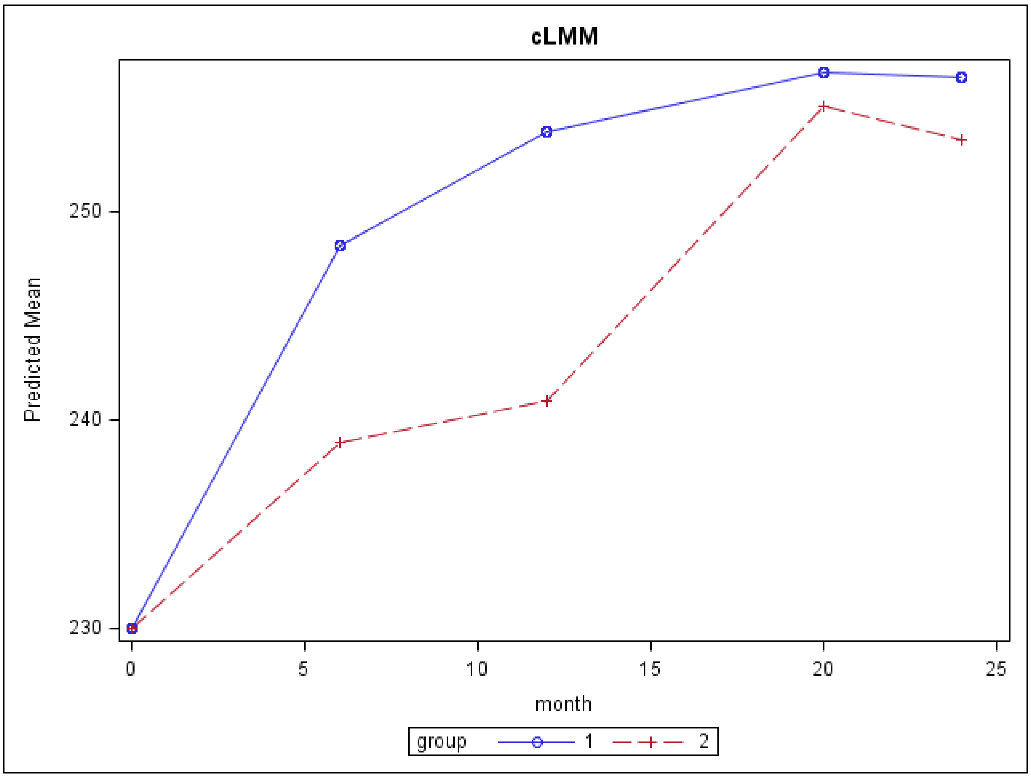
   ```
               Type 3 Tests of Fixed Effects

                         Num      Den
   Effect                 DF       DF    F Value    Pr > F
   month                   4     85.5      14.91    <.0001
   month*treat             4     84.2       1.59    0.1840
   ```

   - 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 difference between these with 95% confidence intervals.

     For instance, in the placebo group the estimated mean change from baseline to final follow-up is 23.47. In the high-dose group the estimated mean change is 23.47+3.00=26.47. The estimated difference in mean change between the groups is 3.00 (95% CI -15.42 to 21.43).

- Does the overall pattern of change over time differ significantly between the groups? I.e. are the response profiles identical.
  Since the model assumes that mean cholesterol is the same in the two groups at baseline, by testing the interaction term `month*treat` we are testing the null hypothesis that the response profiles are identical in the two groups. The P-value is 0.18. Hence, there is no evidence that treatment affects the serum cholesterol levels.
- 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.

```
proc sort data=ncgsfit2; by group month id; run;

proc sgplot data = ncgsfit2;
series x = month y = pred / group = group markers;
run;
```



Figur 6: Predicted response profiles from the constrained mixed model.

The predicted baseline means for the two groups are now identical reflecting the fact that treatment was randomized.