

Stat 402A, HW 11 Answers

My SAS code for almost all parts:

```
data serum;
  infile 'serum.txt' firstobs = 2;
  input diet subject glucose time;

/* calculate then plot the means */
proc means noprint data = serum;
  by diet time;
  var glucose;
  output out = means mean = glucose;

proc plot;
  plot glucose*time = diet;
  title 'Means for each food and time';
run;

proc print;
  run;

/* fit split model */
proc mixed data = serum;
  class diet subject time;
  model glucose = diet time diet*time /ddfm = kr;
  random subject(diet);
  lsmeans diet*time / slice = time;

/* fit ar(1) model, other correlation models fit */
/* by changing the type and/or adding a random */
/* stmt or having neither (indep. model) */
proc mixed data = serum;
  class diet subject time;
  model glucose = diet time diet*time /ddfm = kr;
  repeated time /subject = subject type = ar(1);
```

1. 3 pts. Plot not included. The pattern I see is:
  - Diet 1: mean Glucose levels increase from 15 min to 30 min, then little change
  - Diet 2: mean Glucose levels decrease over time slightly
  - Diet 3: mean Glucose levels increase slightly over time
2. 4 pts. The split plot model.  
The fit statistics for each model are:

model	# param	AIC	AICc	BIC
indep	1	167.7	167.8	169.0
split plot	2	152.0	152.5	152.9
ar(1)	2	154.4	154.9	155.4
ar(1)+RE	3	153.9	154.9	155.3
un	6	155.2	159.4	158.1

The model with the smallest value is the most appropriate. By any criterion, the split plot model is the most appropriate.

Notes: If the three criteria are not the same, I usually rely on AIC. There are folks who prefer AICc and folks who prefer BIC. I go with AIC because the other two tend to select simpler models (in general). In this application (choosing a correlation structure), the 'cost' of a wrong model that is too simple can be much higher than the 'cost' of the wrong too complex model. BIC and to a lesser extent AICc tend to select simpler models, hence my preference for AIC in this application.

There are other applications of model selection statistics for which I much prefer BIC.

3. 2 pts. The observations, means and residuals for the 6 obs from subjects 1 and 2 are:

Subject	Time	Glucose	mean	Residual
1	15min	22	17.5	4.5
	30min	34	35.0	-1.0
	45min	32	31.5	0.5
2	15min	15	17.5	-2.5
	30min	29	35.0	-6.0
	45min	27	31.5	-4.5

Yes. Subject 2 has all negative residuals. Subject 4 has all positive residuals. If the observation at one time is above average (or below average), the other observations on that subject are above (or below) average. That's the meaning of correlation over time within a subject.

If you look at all subjects, subjects 2, 3, 7 and 12 have all negative residuals; subjects 4, 8, and 9 have all positive residuals.

4. 3 pts. Using the split plot model and K-R d.f. adjustment, I get:

Source	F	p-value
diet	11.41	0.0034
time	25.96	<0.0001
diet*time	59.34	<0.0001

Note: The Satterthwaite d.f. adjustment should give the same results, since this analysis is a balanced split-plot.

Antonio: If a student choses a different model, in the previous part, they will get slightly different answers. Don't mark off here if that happens. Do mark off a point or so if the answers are very different, which suggests a major misunderstanding.

5. 3 pts. No, the interaction of diet\*time is highly significant. The differences between diets are quite different at 15 min, 30 min and 45 min.

6. 1 pt. Using the split plot model, I get

Time	F	p-value
15 min	4.24	0.039
30 min	13.20	0.0008
45 min	39.33	<0.0001

There is an effect of diet at all three times. It is more significant at 30min and 45 min.

7. 4 pts. The skeleton ANOVA table for each time separately is:

Source	d.f.
Diet	2
Error	9
total	11

The ANOVA table for the split-plot repeated measures is:

Source	d.f.
Diet	2
Subject(Diet)	9
Time	2
Diet*Time	4
Error	18
total	35

There are two differences between the two ANOVA tables:

1) The repeated measures allows you answer questions about differences in time (either averaged over diets = TIME) or consistency of differences between diets (= DIET\*TIME)

2) More information (3 obs per subject) are being used to estimate the variability between subjects (=SUBJECT(DIET)) and the variability between observations (=ERROR).

I.e. The rep. meas. analysis pools error variances for 15 min, 30min, and 45 min.

Point 1 is a good thing in favor of the rep. meas. analysis. The 2nd point is a good thing if it is reasonable to pool, i.e. if the observations have similar variance at all three times.

The one disadvantage to the repeated measures analysis is that you have to determine a reasonable correlation structure.