**BIOS6643 HW3 Due Oct. 7, 2016**

For practice:

1. (GLM question) Determine the X matrix for the model that imposes ‘sum-to-0’ restrictions up front, for the Myostatin data, using the 2-way effects model. Note that the solution can be found in the Appendix of the GLM notes. As a starting point, for Factor A that has 3 levels, the sum-to-0 restriction means that , and hence . But recall that for the Myostatin data, there are 2 groups and 3 times; also see page 13 of the GLM notes for the sum-to-0 restrictions definition (including interaction). (If you do a search on “sum to 0” in the course notes, it should get you there.)

x={ 1 1 1 0 1 0, 1 1 1 0 1 0, 1 1 1 0 1 0, 1 1 1 0 1 0, 1 1 0 1 0 1,

1 1 0 1 0 1, 1 1 0 1 0 1, 1 1 0 1 0 1, 1 1 0 0 0 0, 1 1 0 0 0 0,

1 1 0 0 0 0, 1 1 0 0 0 0, 1 0 1 0 0 0, 1 0 1 0 0 0, 1 0 1 0 0 0,

1 0 1 0 0 0, 1 0 0 1 0 0, 1 0 0 1 0 0, 1 0 0 1 0 0, 1 0 0 1 0 0,

1 0 0 0 0 0, 1 0 0 0 0 0, 1 0 0 0 0 0, 1 0 0 0 0 0};

1. Write observation-, subject- and complete-data linear mixed models for the following
   1. Model with time as class, random intercept for subjects.
   2. Model with time as continuous, no random terms, AR(1) covariance structure for errors.
   3. Model with time as continuous, random intercept for subjects, AR(1) covariance structure for errors.
   4. Determine *Var*(**Y***i*) for each of the models above (a through c).

Turn in:

Note: For simplicity, you can write your answers for (1) and (2) below for *ri*=*r*=4. I.e., each answer should be a 4x4 matrix.

1. Consider the linear mixed model , where  and , where  and  are independent; *μ* and *τ* are fixed effects; *i* denotes subject, *i*=1,…,*n* and *j* =1,…,*ri* denotes time. In class we determined that *Var*(**Y***i*) had a compound symmetric structure. [This is actually the covariance matrix for **Y***i*, or *Cov*(**Y***i*), but it is also often referred to as *Var*(**Y***i*).] Note that .
   1. Determine the correlation matrix, *Corr*(**Y***i*).
   2. Write the form of , where .
2. Consider the model , where  (no random intercept).
   1. Write the form of **R***i* that yields the compound symmetric structure for *Var*(**Y***i*).

Since there is no random effect (slope or intercept) component, the Ri will look identical to the above 1b.

* 1. Write the form of **R***i* that yields the AR(1) structure for *Var*(**Y***i*). (Hint: refer to HW1.)

The AR(1) structure will have included.

1. Re: slides 17 and 18 in the LMM II slides: on slide 18, tests for linear and quadratic trend are included in **C**. Show that the test H0: **Cτ**=**0** versus HA: H0C (considering rows simultaneously in the same test) is just the main effect test for time. I.e., it is the same test as when using the form of **C** as given on slide 17. Does it make sense that this would be right? Explain.

Main effect test of time:

Polynomial trend lack-of-fit test:

.

For the polynomial trend test since the first row is indicating, which is testing the same contrast as the row 2 of main effect test of time, the equation can be substituted in the second row which would yield which would then become , then whiich would yield the exact same constraints (and contrasts) as the Main effect test of time. Since the orthogonal polynomial contrasts of (r-1) degrees of freedom (in this case 3 time points so r=2) is supposed to be partitioned without overlap, in this case the contrasts for the linear and the quadratic seems to be partitioned to test whether = and thus , which is exactly equal to the main effect test of time given some algebraic manipulation. Also, the contrasts in row 1 and row 2 of the polynomial trend test are orthogonal (no linearly dependent columns or rows) thus the setup of the contrasts initially has no overlap for testing thus the sums of squares essentially add up to the sums of squares of the main effect test of time.

1. With the dog data in the LMM II slides, group and time were modeled as class variables, plus group\*time, and a random intercept for individual dogs. The following questions involve this same model. some questions you can answer by just reviewing the notes, and some you will need to fit yourself. Note that 2-way effects model formulation is shown on Slide 3 and the means model formulation is shown on Slide 16 (in the LMM II slides).
2. List 2 reasons why adding a random intercept for dogs (relative to the same model but without the random term) might help the model.
3. Adding a random intercept allows us to use the clustered data (within, repeated measures of each individual dog) to identify separate intercepts for each unit of each level (of group and time thus multiple parallel slopes) compared to only identifying the overall population effect of group and time (one population slope) – between subject variability.
4. We can also identify the variance in gallbladder that is due to the group (between variance) and also the variance in gallbladder that is due to subject (within variance). Additionally, we can identify/induce complex correlation structure – correlation structure for responses.
5. By adding a random intercept component, we can somewhat control for the various unmeasured aspects of dogs that may affect the repeated measurements.
6. The random intercept model empirically (see below table) has a better fit (with regards to AIC).

Table A.1. Fit statistic of model without random intercept.

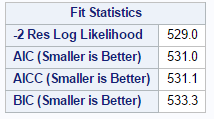


Table A.2. Fit statistic of model with random intercept.



1. Graph the mean response by time for each group.

Figure 1A. Mean response by time for each group of dogs using SAS.

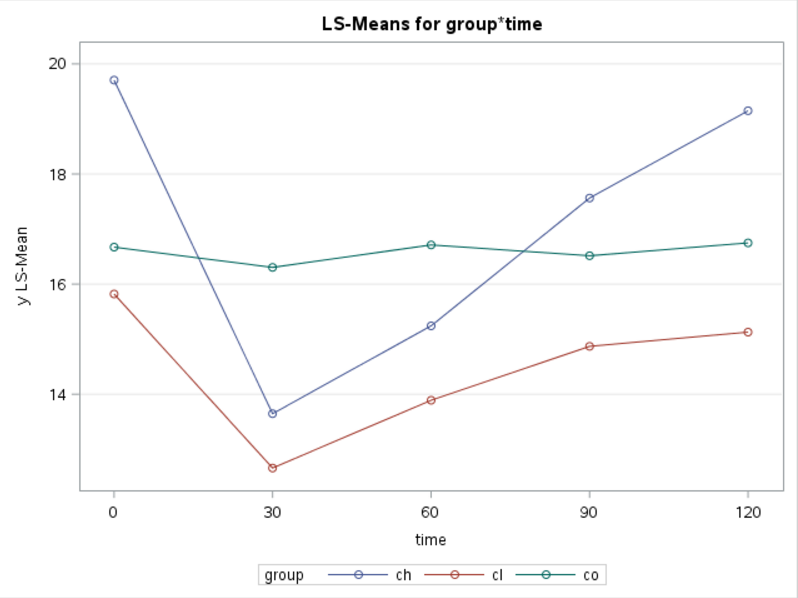
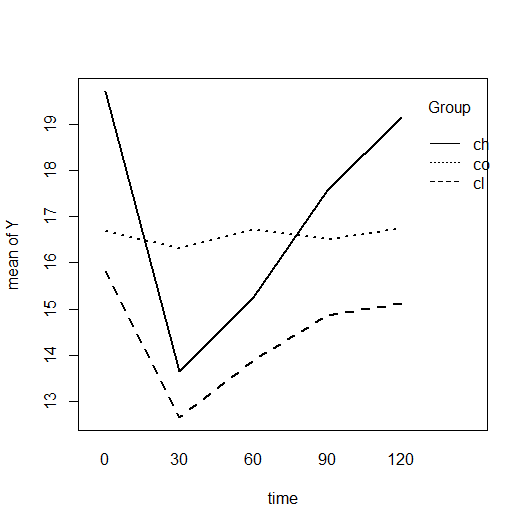
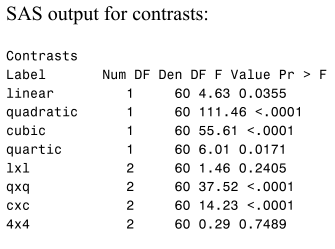


Figure 1B. Mean response by time for each group of dogs using R.



1. Refer to class notes and slides for polynomial test output. Which polynomial trends for time and group\*time most apparent in the graph? Which significant ones are not as apparent in the graph? (There is no correct or incorrect here, it’s just based on effort.)

Table 1.



Based on the output for polynomial trends for time (Table 1), the quadratic and the cubic have p-values < 0.0001 while the same goes for the group\*time where qxq and cxc have the highest test statistic (F) value and lowest p-value. But based on the Figure 1 (A and B), the mean GBV by timer per group indicates that the relationship is perhaps closer to a quadratic relationship and/or linear rather than cubic or quartic. The cubic order term isn’t as apparent in the Figure 1 compared to the linear and quadratic relationship visible in Figure1. And for the sake of parsimony (of the model), the linear component should be reasonable enough to make inferences with the data. (but if another research insisted using polynomial terms then a check on AIC would be in order to identify if polynomial’s would be appropriate for the model).

1. Consider a CONTRAST to test for differences over time between the CH and CL groups. Using the means model, the test can be written as H0:  for all *j*, *j*׳, where subscript ‘1’ denotes CH group, ‘2’ denotes CL. Show that the same test can also be written as H0:  for all *j*, *j*׳ using the effects model. (See LMM II slide 3 for the two-way effects model and 16 for the means model.)

Since for the means model

, for group h=0,1,and 2 where 0 is control, 1 is CH and 2 is CL and for time j=1…5.

Thus

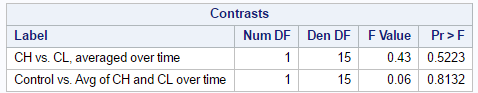
H0: 

H0:

1. Write an ESTIMATE or CONTRAST statement for each question below and carry out the analysis. Summarize your results. (Note that these involve all 5 time points.)

Does the difference between the 2 drug groups change over time?

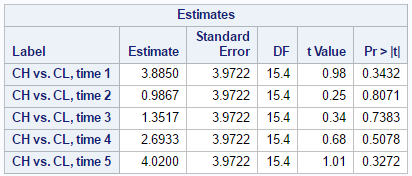
Table 2. Proc Mixed contrast statement for comparison of mean difference between GBV among dogs between CH vs. CL and Control vs. Average of CH and CL.



Based on the p-value (Table 2) from the contrast statement that compared the difference in GBV between CH vs. CL averaged over time, which is 0.5223, there is no statistically significant difference between the 2 drugs groups over time.

Is there a mean difference between drug groups for at least one time point?

Table 3. Mean difference between CH vs. CL at different time points



Based on the p-values (Table 3), there is no mean difference in CH vs. CL during time 1-5. (P-values all greater than 0.05).

Does control differ from the average of the CH and CL drug groups over time? (For the drug group means, you can use the straight average of drug groups means for each time.)

Based on the p-value indicated in Table 2 regarding the difference between control and the average of CH and CL, there is no statistically significant difference; p-value = 0.8132.

1. Get estimates and 95% confidence intervals for each of the following.

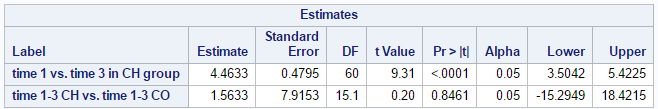
The mean change in scores (from BL to 60 minutes after) for the CH group.

The mean change from baseline to 60 minutes after (time 1 to 3) for CH group is 4.4633 units and the 95% CI is from 3.5042 to 5.4225 (Table 4).

The difference in mean change of scores (BL to 60 min) for the CH group relative to the Control group.

The difference in mean from baseline to 60 minutes after (time 1 to 3) for CH group vs CO group is 1.5633 units and the 95% CI is from -15.2949 to 18.4215 (Table 4).

Table 4. Estimates and 95% CI for 4fi and ii; time 1 vs. 3 in CH group and time 1-3 CH vs. time 1-3 CO.

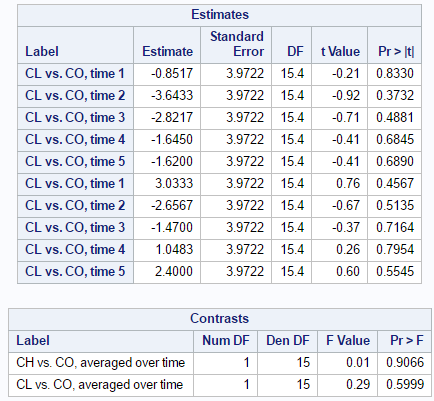


1. Based on what you have done or seen up to this point or your own analysis (if applicable), write a one paragraph (min. 5 sentences) summary of results and discussion for the dog and gallbladder volume data. Include what you believe is important and include statistical results in your write up.

Based on the random intercept model for the dog data, the mean GBV over time by group (Figure 1) indicated a linear relationship. Although the data is probably truncated and with the full dataset there could be a possibility of identifying a quadratic or a cubic relationship, the polynomial trend for time indicated that the linear trend was still significant (p=0.0355) and by justification of parsimony perhaps not including other polynomial terms will be wise. Thus using the linear random intercept model, the difference between CH vs. CL was not statistically significantly different (Table 2, p=0.5223) over time and there was no time point (time 1-5) where the mean difference between CH vs. CL was statistically significantly different (Table 3). Additionally, the difference in GBV values of CL vs. CO was not statistically significantly different over time (Table 5, contrast table) where none of the time points had difference in GBV (Table 5, estimates table). There was a statistically significant difference in time 1 vs. time 3 in the CH group GBV (p-value<0.0001, 95% CI: 3.5042-5.4225) but this doesn’t answer the question of whether or not there is a difference in GBV between treatment groups/control group over time. The difference in GBV change from time 1 and time 3 between CH vs. CO indicated no statistically significant difference (p value = 0.8461). Overall, the hypothesis test regarding whether or not there is a difference in GBV between CH vs. CO, CL vs. CO, and average of CH and CL vs. CO indicated that there were no statistically significant difference amongst the different treatment groups over time regarding GBV in our data.

1. Say that your client requests that time be modeled as a continuous variable rather than class. In particular, they want to be able to estimate GBV values that might be in between the 30 minute intervals (e.g., 45 minutes) in addition to at the observed time points. Any cautions or things to consider when setting up the model? How would you proceed? Answer in 2 to 3 sentences.

Specifically in our data, if we decide to treat time as continuous, the assumption of linear time effect may not be appropriate due to the fact that our polynomial test has indicated a quadratic or cubic trend. Also it may be wise to construct the model in both ways (time as continuous and class) and do a LR or a wald test to identify which model would best fit the data. If the purpose of the suggestion was to be able to estimate GBV values that are between the time marks, it may be wise to identify the best fitting model (through LR or wald) and also potentially fit higher order terms while at it.

Table 5. Proc Mixed contrast statement for comparison of mean difference between GBV among dogs between CH vs. CL and Control vs. Average of CH and CL.  


**Appendix**

**R code**

|  |
| --- |
| #1. Remove Previous Data  rm(list=ls()) |
|  |
| #2. Import dog data |
| library(sas7bdat) |
| dat <- read.sas7bdat("C:/Users/ck/Dropbox/Academic Coursework/FS2016/Long/uni\_dogs.sas7bdat") |
|  |
| library(ggplot2) |
| #3. set p object |
| p <- ggplot(data = dat, aes(x = time, y = y, group = group)) |
| ##4. plot using ggplot2 (a.k.a., scatterplot, mean graph) |
| p + geom\_smooth(aes(group = group, size = 2, method = "lm", se = FALSE) , ) |
| ) |
|  |
| library(nlme) |
| #5. Plot using nlme package |
| interaction.plot (dat$time, factor(dat$group), dat$y, lty=c(1:3),lwd=2,ylab="mean of Y", xlab="time", trace.label="Group") |

|  |
| --- |
| SAS code |
| \*Import data from folder; |
| libname onpc "/folders/myfolders"; |
|  |
| \*set temporary data; |
| data dog; |
| set onpc.uni\_dogs; |
| run; |
|  |
| \*Q4.e.i. Does the difference between the 2 drug groups change over time? |
| ii. Is there a mean difference between drug groups for at least one time point? |
| iii. Control vs. average of CH and CL over time? |
| ; |
|  |
| proc mixed data=dog; class id group time; |
| model y = group\*time / noint ddfm=sat solution; random id(group); |
| contrast 'CH vs. CL, averaged over time' group\*time 1 1 1 1 1 -1 -1 -1 -1 -1 0 0 0 0 0; |
| estimate 'CH vs. CL, time 1' group\*time 1 0 0 0 0 -1 0 0 0 0 0 0 0 0 0; |
| estimate 'CH vs. CL, time 2' group\*time 0 1 0 0 0 0 -1 0 0 0 0 0 0 0 0; |
| estimate 'CH vs. CL, time 3' group\*time 0 0 1 0 0 0 0 -1 0 0 0 0 0 0 0; |
| estimate 'CH vs. CL, time 4' group\*time 0 0 0 1 0 0 0 0 -1 0 0 0 0 0 0; |
| estimate 'CH vs. CL, time 5' group\*time 0 0 0 0 1 0 0 0 0 -1 0 0 0 0 0; |
| contrast 'Control vs. Avg of CH and CL over time' group\*time 1 1 1 1 1 1 1 1 1 1 -2 -2 -2 -2 -2; |
|  |
| run; |
|  |
| \*Q4f. Get estimates and 95% confidence intervals for each of the following. |
| i. The mean change in scores (from BL to 60 minutes after) for the CH group. |
| ii. The difference in mean change of scores (BL to 60 min) for the CH group relative to the Control group. |
| ; |
|  |
| proc mixed data=dog; class id group time; |
| model y = group\*time / noint ddfm=sat solution; random id(group); |
| estimate 'time 1 vs. time 3 in CH group' group\*time 1 0 -1 0 0 0 0 0 0 0 0 0 0 0 0 / CL; |
| estimate 'time 1-3 CH vs. time 1-3 CO' group\*time 1 0 1 0 0 0 0 0 0 0 -1 0 -1 0 0 / CL; |
| run; |

|  |
| --- |
| \*Difference in CH vs. CO and CL vs. CO;  proc mixed data=dog; class id group time; |
| model y = group\*time / noint ddfm=sat solution; random id(group); |
| contrast 'CH vs. CO, averaged over time' group\*time 1 1 1 1 1 0 0 0 0 0 -1 -1 -1 -1 -1; |
| contrast 'CL vs. CO, averaged over time' group\*time 0 0 0 0 0 1 1 1 1 1 -1 -1 -1 -1 -1; |
| contrast 'average of CH and CL vs. CO, averaged over time' group\*time 1 1 1 1 1 1 1 1 1 1 -2 -2 -2 -2 -2; |
|  |
| estimate 'CL vs. CO, time 1' group\*time 0 0 0 0 0 1 0 0 0 0 -1 0 0 0 0; |
| estimate 'CL vs. CO, time 2' group\*time 0 0 0 0 0 0 1 0 0 0 0 -1 0 0 0; |
| estimate 'CL vs. CO, time 3' group\*time 0 0 0 0 0 0 0 1 0 0 0 0 -1 0 0; |
| estimate 'CL vs. CO, time 4' group\*time 0 0 0 0 0 0 0 0 1 0 0 0 0 -1 0; |
| estimate 'CL vs. CO, time 5' group\*time 0 0 0 0 0 0 0 0 0 1 0 0 0 0 -1; |
| estimate 'CL vs. CO, time 1' group\*time 1 0 0 0 0 0 0 0 0 0 -1 0 0 0 0; |
| estimate 'CL vs. CO, time 2' group\*time 0 1 0 0 0 0 0 0 0 0 0 -1 0 0 0; |
| estimate 'CL vs. CO, time 3' group\*time 0 0 1 0 0 0 0 0 0 0 0 0 -1 0 0; |
| estimate 'CL vs. CO, time 4' group\*time 0 0 0 1 0 0 0 0 0 0 0 0 0 -1 0; |
| estimate 'CL vs. CO, time 5' group\*time 0 0 0 0 1 0 0 0 0 0 0 0 0 0 -1; |
|  |
| run; |

**Cal for 3.**

SSlinear = [(-1) () + (0)()+(1)()]^2/(2\*r)

SSquadratic = [(1)()+(-2)()+(1)()]^2/(6\*r)

SSlinear + SSquadratic =

SSrow1 = [(1) () + (-1)()+(0)()]^2 /(2\*r)

SSrow2 = [(1) () + (0)()+(-1)()]^2 /(2\*r)

SStime =

Set equal equations to see if they equal each other

Because there are 3 time points and thus the maximum amount of polynomial contrast would be the (r-1=2) quadratic form. Adding up the (linear) and the (quadratic) contrasts, which are non-overlapping partitions of the SS of the main effect for time due to properties of orthogonality, would yield the SS(time).

Using the equations above:

For the linear and quadratic contrast the first row of the linear contrast is testing the same contrast as the second row of the main effect hypothesis. The second row of the polynomial contrast is

SSlinear = (-1) () + (0)()+(1)()/(2\*r)

SSquadratic = (1)()+(-2)()+(1)()/(6\*r)

SSlinear + SSquadratic = SStime