# Homework 2 Solutions

Alexander McLain

1. The longitudinal data from an insulin study contain 36 rabbits where 12 rabbits were randomly assigned to each of 3 groups: group 1 rabbits received the standard insulin mixture, group 2 rabbits received a mixture containing 1% less protamine than the standard, and group 3 rabbits received a mixture containing 5% less protamine. Rabbits were injected with the assigned mixture at time 0, and blood sugar measurements taken on each rabbit at the time of injection (time 0) and 0.5, 1.0, 1.5, 2.0, 2.5, and 3.0 hours post-injection.

The data file *"insulin"* is on the course website. The variables appearing in columns are: (1) rabbit id, (2) insulin group, and (3-9) response (blood sugar level) at 7 time points.

```
library(tidyverse)
insulin <- read.csv("insulin.csv",header = TRUE, na.strings = "",
                    stringsAsFactors = FALSE)
str(insulin)
```

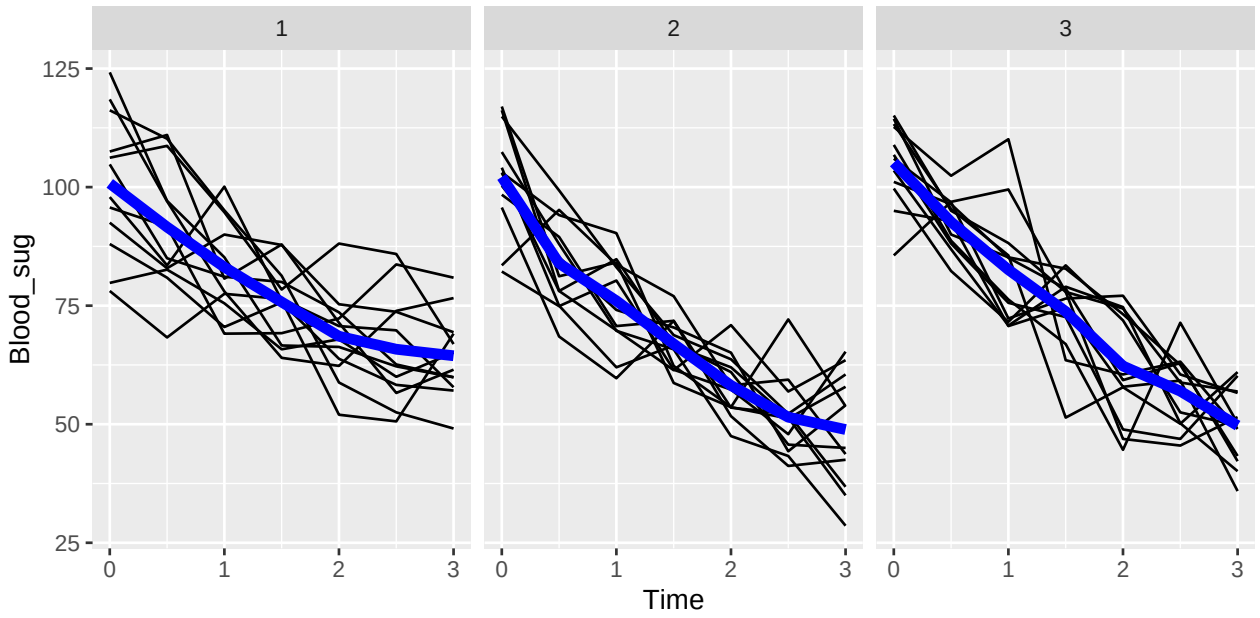
```
## 'data.frame':    252 obs. of  4 variables:
##  $ id       : int  1 1 1 1 1 1 1 2 2 2 ...
##  $ group    : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Time     : num  0 0.5 1 1.5 2 2.5 3 0 0.5 1 ...
##  $ Blood_sug: num  78.1 68.3 77.5 76.4 70.7 ...
```

```
head(insulin)
```

```
##   id group Time Blood_sug
## 1  1     1  0.0      78.1
## 2  1     1  0.5      68.3
## 3  1     1  1.0      77.5
## 4  1     1  1.5      76.4
## 5  1     1  2.0      70.7
## 6  1     1  2.5      69.8
```

a. **(10 points) Create a spagetti-plot of the data with separate panels for each group. Comment on the heterogeneity in the data.**

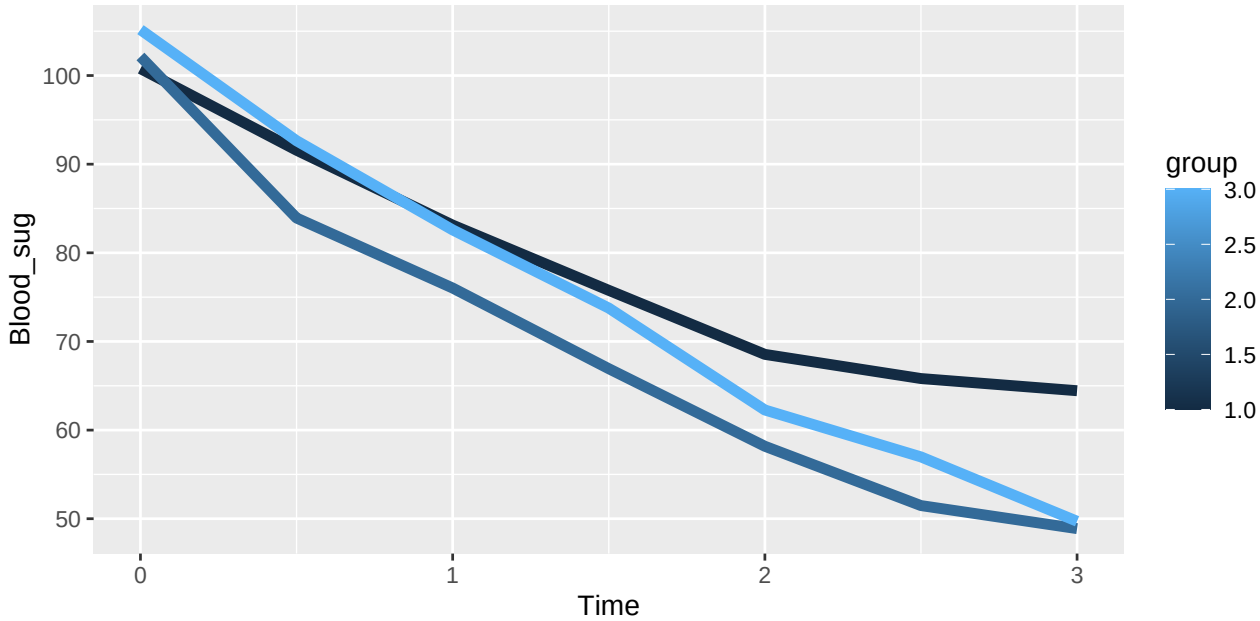
```
p <- ggplot(data = insulin, aes(x = Time, y = Blood_sug, group = id))
p + geom_line() +
stat_summary(aes(group = 1), geom = "line", fun = mean,
color = "blue", size = 2) +
facet_grid(. ~ group)
```




The within subject variability appears to be fairly constant over time and across groups. The mean of the outcome for the three groups is similar at hour=0, but shows different rates of change as hour increases. The change in the outcome over time appears to be approximately linear.

b. **(10 points) Create a plot that has the groups means over time on the same plot with different colors (if you'll print/submit in color) or line-types (if you print/submit in black and white). What trends might be appropriate (e.g., profile, linear, quadratic, etc.)?**

```
p + stat_summary(aes(group = group, color = group), geom = "line", fun = mean,
                 size = 2)
```




Groups 2 and 3 appear to be more linear than group 1, which has a more quadratic shape.

c. **(20 points) Fit a full interaction model using a profile analysis with the following covariance matrices. Hand in the estimates of the estimated covariance matrices for each. No other output.**

Since this is going to be a profile analysis, I want `Time` and `group` to be a factor variables not a numeric variables. I'll make that change first.

```
insulin <- insulin %>% mutate( group = as.factor(group), Time = as.factor(Time) )
# Set up the model
library(nlme)
```

```
##
## Attaching package: 'nlme'
```

```
## The following object is masked from 'package:dplyr':
##
##     collapse
```

```
formula_inter <- Blood_sug ~ Time + group + group*Time
```

Now we'll fit it to the various covariance matrix assumptions

i. **Unstructured (heterogeneous and symmetric)**

```
cor_fun <- corSymm(form = ~ 1|id)
var_fun <- varIdent(form = ~ 1|Time)
# Run the model
lm_UN <- gls(model = formula_inter, data = insulin, correlation = cor_fun,
               weights = var_fun)
getVarCov(lm_UN)
```

```
## Marginal variance covariance matrix
##          [,1]     [,2]      [,3]     [,4]     [,5]      [,6]     [,7]
## [1,] 142.1000  55.3720  27.39900 -20.4940   8.1449  11.60500  -9.8088
## [2,]  55.3720 100.8800  49.96100   4.0967   1.8378   3.57170 -16.8320
## [3,]  27.3990  49.9610 116.31000  -4.6387 -15.3240  -0.14531   3.7046
## [4,] -20.4940   4.0967  -4.63870  59.1380  21.6110 -10.95700 -27.5570
## [5,]   8.1449   1.8378 -15.32400  21.6110  88.2290  29.73300   6.4098
## [6,]  11.6050   3.5717  -0.14531 -10.9570  29.7330  89.16500  23.7740
## [7,]  -9.8088 -16.8320   3.70460 -27.5570   6.4098  23.77400  95.7140
##   Standard Deviations: 11.921 10.044 10.785 7.6901 9.393 9.4427 9.7834
```

ii. **Compound Symmetry**

```
cor_fun <- corCompSymm(form = ~ 1|id)
var_fun <- varIdent(form = ~ 1)
# Run the model
lm_CS <- gls(model = formula_inter, data = insulin, correlation = cor_fun,
               weights = var_fun)
getVarCov(lm_CS)
```

```
## Marginal variance covariance matrix
##         [,1]    [,2]    [,3]    [,4]    [,5]    [,6]    [,7]
## [1,] 98.7920  6.7364  6.7364  6.7364  6.7364  6.7364  6.7364
## [2,]  6.7364 98.7920  6.7364  6.7364  6.7364  6.7364  6.7364
## [3,]  6.7364  6.7364 98.7920  6.7364  6.7364  6.7364  6.7364
## [4,]  6.7364  6.7364  6.7364 98.7920  6.7364  6.7364  6.7364
## [5,]  6.7364  6.7364  6.7364  6.7364 98.7920  6.7364  6.7364
## [6,]  6.7364  6.7364  6.7364  6.7364  6.7364 98.7920  6.7364
## [7,]  6.7364  6.7364  6.7364  6.7364  6.7364  6.7364 98.7920
##   Standard Deviations: 9.9394 9.9394 9.9394 9.9394 9.9394 9.9394 9.9394
```

iii. **Heterogeneous Compound Symmetry.**

```
cor_fun <- corCompSymm(form = ~ 1|id)
var_fun <- varIdent(form = ~ 1|Time)
# Run the model
lm_HetCS <- gls(model = formula_inter, data = insulin, correlation = cor_fun,
                weights = var_fun)
getVarCov(lm_HetCS)
```

```
## Marginal variance covariance matrix
##          [,1]    [,2]     [,3]    [,4]    [,5]    [,6]    [,7]
## [1,] 140.7900  7.0003   7.5798  5.5241  6.5812  6.6247  6.9977
## [2,]   7.0003 98.5240   6.3407  4.6211  5.5053  5.5417  5.8537
## [3,]   7.5798  6.3407 115.5100  5.0036  5.9611  6.0005  6.3383
## [4,]   5.5241  4.6211   5.0036 61.3530  4.3444  4.3731  4.6193
## [5,]   6.5812  5.5053   5.9611  4.3444 87.0800  5.2099  5.5033
## [6,]   6.6247  5.5417   6.0005  4.3731  5.2099 88.2350  5.5397
## [7,]   6.9977  5.8537   6.3383  4.6193  5.5033  5.5397 98.4500
##   Standard Deviations: 11.866 9.9259 10.748 7.8328 9.3317 9.3934 9.9222
```

iv. **Exponential structure with a nugget**

```
insulin <- insulin %>% mutate( Time_c = as.numeric(Time) )
cor_fun <- corExp(form = ~ Time_c|id, nugget = TRUE)
var_fun <- varIdent(form = ~ 1)
# Run the model
lm_Exp <- gls(model = formula_inter, data = insulin, correlation = cor_fun,
              weights = var_fun)
getVarCov(lm_Exp)
```

```
## Marginal variance covariance matrix
##          [,1]     [,2]    [,3]    [,4]    [,5]     [,6]     [,7]
## [1,] 99.41800 31.59900 10.0440  3.1924  1.0147  0.32251  0.10251
## [2,] 31.59900 99.41800 31.5990 10.0440  3.1924  1.01470  0.32251
## [3,] 10.04400 31.59900 99.4180 31.5990 10.0440  3.19240  1.01470
## [4,]  3.19240 10.04400 31.5990 99.4180 31.5990 10.04400  3.19240
## [5,]  1.01470  3.19240 10.0440 31.5990 99.4180 31.59900 10.04400
## [6,]  0.32251  1.01470  3.1924 10.0440 31.5990 99.41800 31.59900
## [7,]  0.10251  0.32251  1.0147  3.1924 10.0440 31.59900 99.41800
##   Standard Deviations: 9.9708 9.9708 9.9708 9.9708 9.9708 9.9708 9.9708
```

v. **Exponential structure without a nugget.**

```
insulin <- insulin %>% mutate( Time_c = as.numeric(Time) )
cor_fun <- corExp(form = ~ Time_c|id)
var_fun <- varIdent(form = ~ 1)
# Run the model
lm_ExpNoNug <- gls(model = formula_inter, data = insulin, correlation = cor_fun,
                   weights = var_fun)
getVarCov(lm_ExpNoNug)
```

```
## Marginal variance covariance matrix
##          [,1]     [,2]    [,3]    [,4]    [,5]     [,6]     [,7]
## [1,] 99.41800 31.59900 10.0440  3.1924  1.0147  0.32251  0.10251
## [2,] 31.59900 99.41800 31.5990 10.0440  3.1924  1.01470  0.32251
## [3,] 10.04400 31.59900 99.4180 31.5990 10.0440  3.19240  1.01470
## [4,]  3.19240 10.04400 31.5990 99.4180 31.5990 10.04400  3.19240
## [5,]  1.01470  3.19240 10.0440 31.5990 99.4180 31.59900 10.04400
```

```
## [6,]  0.32251  1.01470  3.1924 10.0440 31.5990 99.41800 31.59900
## [7,]  0.10251  0.32251  1.0147  3.1924 10.0440 31.59900 99.41800
##   Standard Deviations: 9.9708 9.9708 9.9708 9.9708 9.9708 9.9708 9.9708
```

d. **(5 points) Based on the estimated covariance matrices what do you think is best and why?**

- Based on the unstructured matrix there does not appear to be much heterogeneity in the model. So, we can cross off the two heterogeneous options.
- For the exponential structures, the estimates are basically the same with a nugget and without nugget. As a result, without a nugget would be preferred since it has one less parameter.
- So it comes down to the homogeneous commpound symmetric and the homogeneous exponential structure without a nugget. I think the exponential structure without a nugget will better here. If the compound symmetric were the best choice, I would expect the "with nugget" and "without nugget" to be more similar.

e. **(10 points) For the models that were fit in (c), which model has the best fit according to AIC and BIC?**

We can get all the fit statistics using the `anova` function. Ignore the likelihood ratio tests for now.

```
anova(lm_UN,lm_CS,lm_HetCS,lm_Exp,lm_ExpNoNug)
```

```
##             Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## lm_UN           1 49 1819.619 1988.297 -860.8095
## lm_CS           2 23 1812.054 1891.229 -883.0268 1 vs 2 44.43451  0.0136
## lm_HetCS        3 29 1817.684 1917.514 -879.8422 2 vs 3  6.36922  0.3831
## lm_Exp          4 24 1797.091 1879.709 -874.5454 3 vs 4 10.59362  0.0601
## lm_ExpNoNug     5 23 1795.091 1874.266 -874.5454 4 vs 5  0.00000  0.9996
```

According to AIC and BIC and homogeneous exponential structure without a nugget has the best fit.

e. **(10 points) Complete a likelihood ratio test between the following structures. To each of the results coincide with the results from AIC and BIC?**

f. Unstructured and Heterogeneous Compound Symmetry.

```
anova(lm_UN,lm_HetCS)
```

```
##          Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## lm_UN        1 49 1819.619 1988.297 -860.8095
## lm_HetCS     2 29 1817.684 1917.514 -879.8422 1 vs 2 38.06529  0.0087
```

Here, the likelihood ratio test prefers the **unstructured** model. This is opposite of what the AIC and BIC criteria, which prefer the heterogeneous compound symmetric.

ii. Unstructured and Homogeneous Exponential without a nugget.

```
anova(lm_UN,lm_ExpNoNug)
```

```
##             Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## lm_UN           1 49 1819.619 1988.297 -860.8095
## lm_ExpNoNug     2 23 1795.091 1874.266 -874.5454 1 vs 2 27.47167  0.3849
```

Here, the likelihood ratio test prefers the **exponential** model. This is the same the AIC and BIC criteria.

g. **(10 points) Using the model that fit best from (c), test whether the time profiles of means are different in the groups.**

```
anova(lm_ExpNoNug)
```

```
## Denom. DF: 231
##             numDF  F-value p-value
## (Intercept)     1 8321.666  <.0001
## Time            6   85.347  <.0001
## group           2    9.422  0.0001
## Time:group     12    1.420  0.1575
```

Based on the above model there does not appear to be an interaction in the data. As a result, we cannot say that the time profiles of the groups are different.

h. **(10 points) Does time have a significant impact on the response (this *may* require you to fit another model).**

To test this we'll fit another model without an interaction term.

```
formula_no_inter <- Blood_sug ~ Time + group
cor_fun <- corExp(form = ~ Time_c|id)
# Run the model
lm_no_inter <- gls(model = formula_no_inter, data = insulin, correlation = cor_fun)
# First look at the type III tests
anova(lm_no_inter)
```

```
## Denom. DF: 243
##             numDF  F-value p-value
## (Intercept)     1 7513.324  <.0001
## Time            6   80.075  <.0001
## group           2    8.423   3e-04
```

It's clear that time (and group) have a significant effect on the outcome.

i. **(5 points) Using the model that fit best from (c), test and give an estimate of the difference in the mean response level at 0.5 hour (0 hour) from baseline in group 2.**

There are a number of ways to do this. I'll show two of them. First, we can use the `emmeans` package. Notice here that we use `specs = Time|group`, which tells `emeans` that we want to look at time **by** group. Here, I did not use any multiple comparisons corrections because I am only interested in one comparison.

```
library(emmeans)
est_means <- emmeans(lm_ExpNoNug, specs = ~Time|group, mode = "df.error")
contrast(est_means, interaction = c(group = "trt.vs.ctrl"))
```

```
## group = 1:
##  Time_trt.vs.ctrl estimate   SE  df t.ratio p.value
##  0.5 - 0             -9.21 3.36 229  -2.739 0.0066
##  1 - 0              -17.66 3.86 229  -4.575 <.0001
##  1.5 - 0            -25.02 4.00 229  -6.247 <.0001
##  2 - 0              -32.24 4.05 229  -7.961 <.0001
##  2.5 - 0            -34.97 4.06 229  -8.604 <.0001
##  3 - 0              -36.34 4.07 229  -8.932 <.0001
##
## group = 2:
##  Time_trt.vs.ctrl estimate   SE  df t.ratio p.value
##  0.5 - 0            -18.23 3.36 229  -5.421 <.0001
##  1 - 0              -26.12 3.86 229  -6.769 <.0001
##  1.5 - 0            -35.23 4.00 229  -8.796 <.0001
##  2 - 0              -43.98 4.05 229 -10.859 <.0001
```

```
##  2.5 - 0            -50.67 4.06 229 -12.467 <.0001
##  3 - 0              -53.27 4.07 229 -13.093 <.0001
##
## group = 3:
##  Time_trt.vs.ctrl estimate   SE  df t.ratio p.value
##  0.5 - 0            -12.53 3.36 229  -3.728 0.0002
##  1 - 0              -22.59 3.86 229  -5.854 <.0001
##  1.5 - 0            -31.41 4.00 229  -7.843 <.0001
##  2 - 0              -42.91 4.05 229 -10.595 <.0001
##  2.5 - 0            -48.20 4.06 229 -11.860 <.0001
##  3 - 0              -55.46 4.07 229 -13.631 <.0001
##
## Degrees-of-freedom method: df.error
```

The second way is to refit the above interaction model using group 2 as the referent group. When this is done, the main effects of time are the estimated difference between the time of the coefficient and the baseline time for group 2. For example, `time0.5` is the estimated difference between time 0.5 and time 0 (baseline the referent time) for group 2.

```
insulin <- insulin %>% mutate( group = relevel(as.factor(group), ref = "2") )
cor_fun <- corExp(form = ~ Time_c|id, nugget = TRUE)
# Run the model
lm_inter <- gls(model = formula_inter, data = insulin, correlation = cor_fun)
```

Now we can look at the coefficients to get the answer.

```
summary(lm_inter)
```

```
## Generalized least squares fit by REML
##   Model: formula_inter
##   Data: insulin
##        AIC      BIC    logLik
##   1797.091 1879.709 -874.5454
##
## Correlation Structure: Exponential spatial correlation
##  Formula: ~Time_c | id
##  Parameter estimate(s):
##        range       nugget
## 8.724553e-01 8.278504e-08
##
## Coefficients:
##                     Value Std.Error   t-value p-value
## (Intercept)    102.15000  2.878334  35.48928  0.0000
## Time0.5        -18.22500  3.361999  -5.42088  0.0000
## Time1          -26.12500  3.859489  -6.76903  0.0000
## Time1.5        -35.22500  4.004692  -8.79593  0.0000
## Time2          -43.97500  4.049753 -10.85869  0.0000
## Time2.5        -50.66667  4.063971 -12.46728  0.0000
## Time3          -53.26667  4.068480 -13.09252  0.0000
## group1          -1.36667  4.070579  -0.33574  0.7374
## group3           3.02500  4.070579   0.74314  0.4582
## Time0.5:group1   9.01667  4.754585   1.89641  0.0592
## Time1:group1     8.46667  5.458142   1.55120  0.1222
## Time1.5:group1  10.20833  5.663489   1.80248  0.0728
## Time2:group1    11.73333  5.727216   2.04870  0.0416
## Time2.5:group1  15.70000  5.747323   2.73171  0.0068
```

```
## Time3:group1    16.92500  5.753700   2.94159  0.0036
## Time0.5:group3   5.69167  4.754585   1.19709  0.2325
## Time1:group3     3.53333  5.458142   0.64735  0.5180
## Time1.5:group3   3.81667  5.663489   0.67391  0.5010
## Time2:group3     1.06667  5.727216   0.18625  0.8524
## Time2.5:group3   2.46667  5.747323   0.42919  0.6682
## Time3:group3    -2.19167  5.753700  -0.38091  0.7036
##
##  Correlation:
##                (Intr) Tim0.5 Time1  Tim1.5 Time2  Tim2.5 Time3  group1 group3
## Time0.5        -0.584
## Time1          -0.670  0.574
## Time1.5        -0.696  0.462  0.635
## Time2          -0.703  0.428  0.525  0.652
## Time2.5        -0.706  0.418  0.490  0.542  0.657
## Time3          -0.707  0.415  0.479  0.508  0.548  0.658
## group1         -0.707  0.413  0.474  0.492  0.497  0.499  0.500
## group3         -0.707  0.413  0.474  0.492  0.497  0.499  0.500  0.500
## Time0.5:group1  0.413 -0.707 -0.406 -0.327 -0.303 -0.295 -0.293 -0.584 -0.292
## Time1:group1    0.474 -0.406 -0.707 -0.449 -0.371 -0.347 -0.339 -0.670 -0.335
## Time1.5:group1  0.492 -0.327 -0.449 -0.707 -0.461 -0.384 -0.359 -0.696 -0.348
## Time2:group1    0.497 -0.303 -0.371 -0.461 -0.707 -0.464 -0.387 -0.703 -0.352
## Time2.5:group1  0.499 -0.295 -0.347 -0.384 -0.464 -0.707 -0.465 -0.706 -0.353
## Time3:group1    0.500 -0.293 -0.339 -0.359 -0.387 -0.465 -0.707 -0.707 -0.353
## Time0.5:group3  0.413 -0.707 -0.406 -0.327 -0.303 -0.295 -0.293 -0.292 -0.584
## Time1:group3    0.474 -0.406 -0.707 -0.449 -0.371 -0.347 -0.339 -0.335 -0.670
## Time1.5:group3  0.492 -0.327 -0.449 -0.707 -0.461 -0.384 -0.359 -0.348 -0.696
## Time2:group3    0.497 -0.303 -0.371 -0.461 -0.707 -0.464 -0.387 -0.352 -0.703
## Time2.5:group3  0.499 -0.295 -0.347 -0.384 -0.464 -0.707 -0.465 -0.353 -0.706
## Time3:group3    0.500 -0.293 -0.339 -0.359 -0.387 -0.465 -0.707 -0.353 -0.707
##                T0.5:1 Tm1:g1 T1.5:1 Tm2:g1 T2.5:1 Tm3:g1 T0.5:3 Tm1:g3 T1.5:3
## Time0.5
## Time1
## Time1.5
## Time2
## Time2.5
## Time3
## group1
## group3
## Time0.5:group1
## Time1:group1    0.574
## Time1.5:group1  0.462  0.635
## Time2:group1    0.428  0.525  0.652
## Time2.5:group1  0.418  0.490  0.542  0.657
## Time3:group1    0.415  0.479  0.508  0.548  0.658
## Time0.5:group3  0.500  0.287  0.231  0.214  0.209  0.207
## Time1:group3    0.287  0.500  0.318  0.262  0.245  0.240  0.574
## Time1.5:group3  0.231  0.318  0.500  0.326  0.271  0.254  0.462  0.635
## Time2:group3    0.214  0.262  0.326  0.500  0.328  0.274  0.428  0.525  0.652
## Time2.5:group3  0.209  0.245  0.271  0.328  0.500  0.329  0.418  0.490  0.542
## Time3:group3    0.207  0.240  0.254  0.274  0.329  0.500  0.415  0.479  0.508
##                Tm2:g3 T2.5:3
## Time0.5
## Time1
```

```
## Time1.5
## Time2
## Time2.5
## Time3
## group1
## group3
## Time0.5:group1
## Time1:group1
## Time1.5:group1
## Time2:group1
## Time2.5:group1
## Time3:group1
## Time0.5:group3
## Time1:group3
## Time1.5:group3
## Time2:group3
## Time2.5:group3  0.657
## Time3:group3    0.548  0.658
##
## Standardized residuals:
##        Min         Q1        Med         Q3        Max
## -2.3343064 -0.6475130  0.0142081  0.5601750  2.7597135
##
## Residual standard error: 9.970842
## Degrees of freedom: 252 total; 231 residual
```

Both of these give the same answer. In group 2, the estimated difference between hour 0.5 and baseline is $-18.23$ (p.value $< 0.001$)

j. **(10 points) Using the model that fit best from (c), interpret at least two of the parameters in context of the problem. Have one of the parameters you interpret be from an interaction.**

- The average difference in insulin between baseline and hour 0.5 was $-9.21$ (baseline was higher, p.value $= 0.007$) for rabbits that received the standard mixture.
- The average change between baseline to hour 0.5 for rabbits in the 1% less protamine group was 9.02 units *less* than the same difference for rabbits in the standard mixture group (p.value $= 0.059$).

Putting these results together it shows that the average insulin decreased between baseline and hour 0.5 in the standard mixture group. In the 1% less protamine group insulin decreased more, but not *significantly more* at the $\alpha = 0.05$ level.