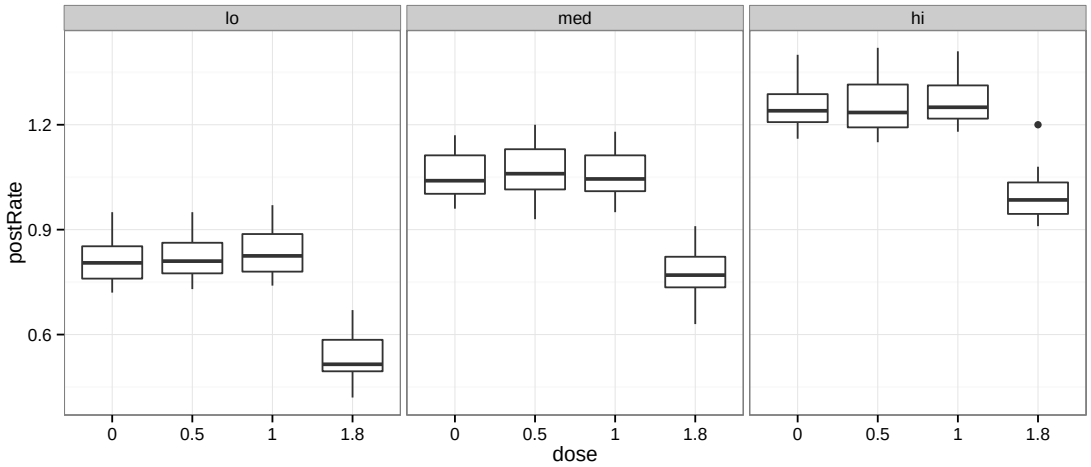
Stat 505 Assignment 5

October 3, 2014

1. The experimental units are 12 thirsty albino rats who are trained to press a lever to get water prior to the experiment. Their pre-experiment pressing rate is recorded as low (1), medium (2), or high (3). They are then injected with one of four levels of a drug where 0 is a control saline solution, the other values are mg per kg of the rat's weight. This was a cross-over design replicated twice, so each rat has 8 measurements of postRate (number of lever presses per second), two at each of four drug levels. Make classification variables (including both predictors) into factors.



(a)

(b) Write out a model for these data using preRate as a three-level factor and drug as a four-level factor. Include distributions for all random components (assuming normality throughout).

$$\begin{aligned} \boldsymbol{y}_i &= \boldsymbol{X}_i\boldsymbol{\beta} + b_i\mathbf{1} + \epsilon_\mathbf{i} \\ b_i &\sim N(0, \sigma_b^2) \\ \epsilon_i &\sim N(0, \sigma^2) \end{aligned}$$

where $i \in \{1, 2, 3, ..., 12\}$ and

$$\boldsymbol{\beta} = \begin{bmatrix} \mu \\ \tau_L \\ \tau_M \\ \tau_H \\ s\alpha_0 \\ \alpha_{0.5} \\ \alpha_1 \\ \alpha_{1.8} \end{bmatrix} .$$

(c) What is the variance-covariance matrix for the 96 observations? (Use Greek letters, not estimated values.)

$$Var(\boldsymbol{y}) = \begin{bmatrix} \boldsymbol{\Sigma_1} & 0 & \dots & 0 \\ 0 & \boldsymbol{\Sigma_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \boldsymbol{\Sigma_{12}} \end{bmatrix} \quad \text{where}$$

$$\boldsymbol{\Sigma_i} = \begin{bmatrix} \sigma^2 + \sigma_b^2 & \sigma_b^2 & \dots & \sigma_b^2 \\ \sigma_b^2 & \sigma^2 + \sigma_b^2 & \dots & \sigma_b^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_b^2 & \sigma_b^2 & \dots & \sigma^2 + \sigma_b^2 \end{bmatrix}$$

i. Explain the structure as in the class notes.
The variance-covariance matrix of y is block diagonal with Σ_i on the diagonal where each Σ_i is an $8x8$ compound symmetric matrix as shown above.

ii. What are the variances of all responses?
The variance of all the responses is $\sigma^2 + \sigma_b^2$.

iii. What are the covariances between each possible pair of responses?
The covariance between each pair of responses is σ_b^2.

(d) Is a 'split plot' analysis appropriate for these data? Explain.
Yes, a split plot analysis would be appropriate as it is a split plot design. There are four rats within each level of pre-experiment pressing rate (lo, med, and hi), and each rat is exposed to all four drug treatments.

(e) Fit the above model to the data using `aov` in R using the 'whole plot' factor within the Error function. Explain the output including the results for each F test shown.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
preRate	2	3.18	1.59	34.46	0.0001
Residuals	9	0.41	0.05		
dose	3	1.37	0.46	251.68	0.0000
dose:preRate	6	0.01	0.00	0.59	0.7378
Residuals	75	0.14	0.00		

There is strong evidence that the mean post experiment pressing rate depends on the pre-experiment pressing rate after accounting for rat and dose (p-value$<$ 0.0001 from F-stat= 34.46 on 2 and 9 df). There is also strong evidence that the mean post experiment pressing rate depends on drug dose after accounting for rat and pre-experiment rate (p=value$<$ 0.0001 from F-stat= 251.68 on 3 and 75 df). There is no evidence that the difference in the mean post experiment pressing rates among doses changes across pre-experiment pressing rates, after accounting for rat (p-value= 0.7378 from F-stat= 0.59 on 6 and 75 df).

(f) Fit the model using `gls` and the correct correlation specification in R. Compare F tests with those from aov.
*The F-tests for dose, preRate, and dose*preRate are the same as above.*

(g) Run `anova(lm(postRate ~ preRate * dose * ratID,data=rats))` and explain where each line of output (its df and its Sum Sq) shows up in the table produced by `aov`. Which lines provide the proper F test to test for effects of preRate?

	numDF	F-value	p-value
(Intercept)	1	1996.94	0.00
dose	3	251.68	0.00
preRate	2	34.46	0.00
dose:preRate	6	0.59	0.74

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
preRate	2	3.18	1.59	627.67	0.0000
dose	3	1.37	0.46	180.90	0.0000
ratID	9	0.41	0.05	18.21	0.0000
preRate:dose	6	0.01	0.00	0.42	0.8596
dose:ratID	27	0.01	0.00	0.22	1.0000
Residuals	48	0.12	0.00		

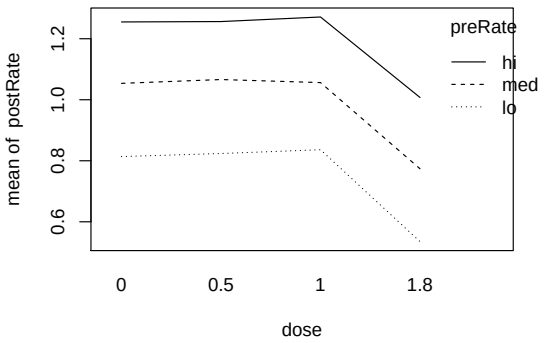
*The degrees of freedom and sums of squares for preRate, dose, and preRate*dose terms are the same as the aov() output above. The df and sums of square for 'ratID' is the same as the wholeplot residual df and SS in the aov() output. If you sum the df and SS of the dose:ratID and residual rows in the lm output, you get the splitplot residual df and SS in aov() above. To get the proper F-test to test for the effects of pre-Rate, you would have to divide $MS_{PreRate}/MS_{RatID}$. This would give you the F-statistic for the preRate row in the aov() output.*

(h) Return to (c) and provide estimates for all variances and covariances.

```
## Marginal variance covariance matrix
##         [,1]    [,2]    [,3]    [,4]    [,5]    [,6]    [,7]    [,8]
## [1,] 0.00735 0.00553 0.00553 0.00553 0.00553 0.00553 0.00553 0.00553
## [2,] 0.00553 0.00735 0.00553 0.00553 0.00553 0.00553 0.00553 0.00553
## [3,] 0.00553 0.00553 0.00735 0.00553 0.00553 0.00553 0.00553 0.00553
## [4,] 0.00553 0.00553 0.00553 0.00735 0.00553 0.00553 0.00553 0.00553
## [5,] 0.00553 0.00553 0.00553 0.00553 0.00735 0.00553 0.00553 0.00553
## [6,] 0.00553 0.00553 0.00553 0.00553 0.00553 0.00735 0.00553 0.00553
## [7,] 0.00553 0.00553 0.00553 0.00553 0.00553 0.00553 0.00735 0.00553
## [8,] 0.00553 0.00553 0.00553 0.00553 0.00553 0.00553 0.00553 0.00735
##   Standard Deviations: 0.0857 0.0857 0.0857 0.0857 0.0857 0.0857 0.0857 0.0857
```

Above is the $8x8$ *variance covariance matrix. Note that each of the twelve rats has the same variance/covariance matrix. The estimates of the variances are on the diagonal, and the estimates of the covariances are on the off-diagonal. Essentially, the variance of any of the* 8 *observations is estimated to be* 0.00735*, and the covariance between any two observations is estimated to be* 0.00553*.*

(i) Explain what we've learned about the drug effects from these data. Does dose interact with pretreatment press rate? How does the identification of rats (as opposed to random assignment) as low/medium/high rates of bar pushing effect the scope of inference?

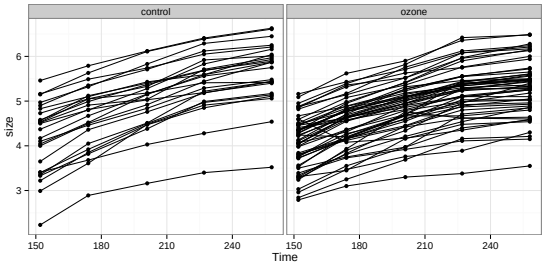


	Df	Sum Sq	Mean Sq	F value	Pr(>F)
preRate	2	3.18	1.59	34.46	0.0001
Residuals	9	0.41	0.05		
dose	3	1.37	0.46	259.57	0.0000
Residuals1	81	0.14	0.00		

As we saw above, there is no evidence of an interaction between drug and pretreatment press rate (p-value$= 0.7378$ *from F-stat*$= 0.59$ *on* 6 *and* 75 *df). I went ahead and fit an additive model using the aov() function to evaluate the effect of drug dose and preRate. After controlling for rat and pretreatment pressing rate, there is strong evidence that the mean posttreatment pressing rate depends on drug dose (p-value*< 0.00001 *from F-stat*$= 259.57$ *on* 3 *and* 81 *df). There is also strong evidence that the mean posttreatment pressing rate depends on pretreatment pressing rate (p-value*< 0.0001 *from F-stat*$= 34.46$ *on* 2 *and* 9 *df). Since rats were not randomly assigned to pretreatment pressing rates, we cannot infer that higher pretreatment pressing rates causes rats to have higher posttreatment pressing rates.*

2. Load the Sitka data (from the MASS library in R) on the growth of 79 sitka spruce trees.

   (a) Plot size over time, separating the two groups, and using a different line for each individual tree. Does it appear that rate of growth is constant? (a linear relationship) or does it change for at least some trees?



   The rate of growth does appear to change for some trees. The rate of growth appears to slow down for some trees as they get older.

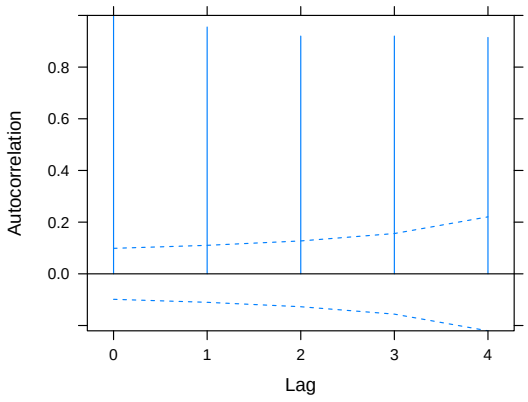
   (b) Use `gls` to fit a quadratic model across all the data. Update the model adding treatment effects which allow the intercept, slope, or quadratic coefficients to depend on

treatment.

	numDF	F-value	p-value
(Intercept)	1	22988.45	0.00
Time	1	222.44	0.00
time2	1	9.02	0.00

	numDF	F-value	p-value
(Intercept)	1	23461.50	0.00
Time	1	227.02	0.00
time2	1	9.21	0.00
treat	1	9.66	0.00
Time:treat	1	1.40	0.24
time2:treat	1	0.01	0.90

(c) The times at which the data were gathered are not quite equally spaced, but assume that they are close enough to equal, and check for serial correlation with an appropriate plot. Conclusions?



There is clearly serial correlation among observations within a tree. The plot shows that the correlation between times 0 *and* 1 *for a given tree is about* 0.95*. The correlation between times* 0 *and* 2 *is about* 0.90*.*

(d) Update the above model to obtain three other models:

- add AR1 correlation structure (within a tree as `corAR1(form = 1|tree)`)
- add compound symmetric correlation (within a tree).
- add symmetric correlation (within a tree). We are having some troubles getting convergence, you might need to specify some starting values for the correlations. I got it to work using `corSymm(rep(.9,10), ...)` (where ... is the same formula used with the other correlation models).

Compare the four models with the `anova` function. Which of the four models has smallest AIC? Which does the F test favor? (There is no nesting in either direction between AR1 and CompSymm models, but each is intermediate between no correlation and the full correlation fit, so one anova could compare AR1 to null and full, and a second could compare CompSymm to null and full models.)

```
##            Model df AIC BIC logLik   Test L.Ratio p-value
## sitka.gls2     1  7 830 858   -408
## sitka.gls3     2  8 -43 -11     29 1 vs 2     875  <.0001
## sitka.gls5     3 17 -63   4     49 2 vs 3      39  <.0001
##            Model df AIC BIC logLik   Test L.Ratio p-value
## sitka.gls2     1  7 830 858   -408
## sitka.gls4     2  8 106 138    -45 1 vs 2     726  <.0001
## sitka.gls5     3 17 -63   4     49 2 vs 3     188  <.0001
```

The F-test favors the model with symmetric correlation structure. We can see in the above anova that sitka.gls5, which is the model with symmetric correlation structure, has the largest log likelihood and the smallest AIC of all the models.

(e) Now we should have a model with lots of treatment terms and a reasonable variance-covariance structure, since you've looked at several correlation models and allowed for increasing variance as trees get bigger. Now examine the treatment effects. Are all of the terms in the model formula needed? Reduce the model one term at a time until you can justify all remaining terms.

The standardized residual plot didn't show a clear trend, but I went ahead and included VarPower() to allow for increasing variance as trees get bigger because biologically, this makes sense. Below is the anova for this model.

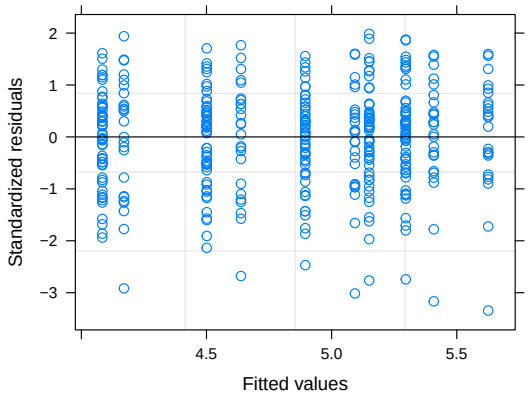
	Value	Std.Error	t-value	p-value
(Intercept)	-1.20	0.48	-2.50	0.01
Time	0.05	0.00	11.34	0.00
time2	-0.00	0.00	-8.92	0.00
treatozone	-0.07	0.58	-0.12	0.90
Time:treatozone	0.00	0.01	0.17	0.87
time2:treatozone	-0.00	0.00	-0.61	0.54

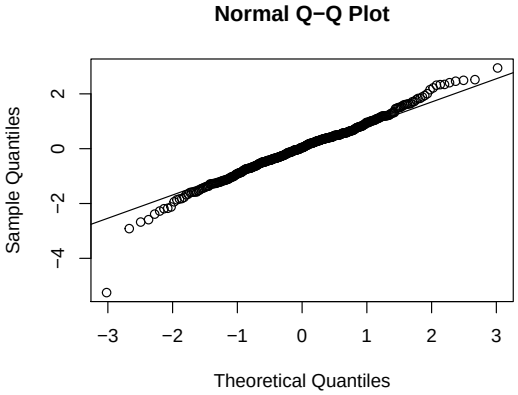
	Value	Std.Error	t-value	p-value
(Intercept)	-1.43	0.30	-4.71	0.00
Time	0.05	0.00	20.78	0.00
time2	-0.00	0.00	-16.77	0.00
treatozone	0.26	0.20	1.30	0.20
Time:treatozone	-0.00	0.00	-3.95	0.00

	Value	Std.Error	t-value	p-value
(Intercept)	-1.17	0.28	-4.21	0.00
Time	0.05	0.00	20.05	0.00
time2	-0.00	0.00	-16.77	0.00
treat2control	-0.26	0.20	-1.30	0.20
Time:treat2control	0.00	0.00	3.95	0.00

Initially, there is no evidence of a time^2 by treatment interaction. I removed this term from the model. All other terms are important except for the treatment term. We do, however, find evidence of a treatment by time interaction, so I will leave the treatment main effect in the model to allow for different intercepts between the ozone and control groups.

(f) Plot the residuals versus fitted and normal quantile plots. Discuss any problems.





There is no apparent trend in the standardized residuals plot, but nevertheless we have allowed for increasing variance as trees get bigger. The residuals do have longer tails than the normal distribution, but with a sample size of 395 *the central limit theorem kicks in and I don't worry much about the normality assumption.*

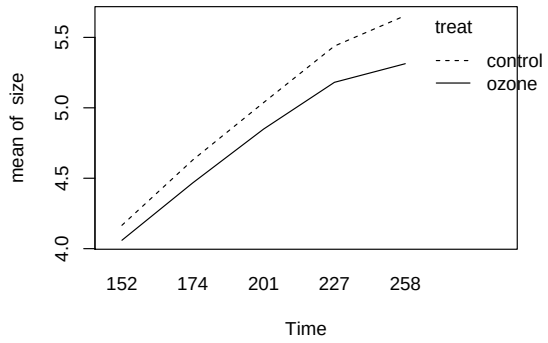
(g) Check your favorite model with the `intervals` function to be sure that we don't have an over-specified model.

```
## Approximate 95% confidence intervals
##
##  Coefficients:
##                     lower      est.     upper
## (Intercept)     -2.02742 -1.43e+00 -8.34e-01
## Time             0.04569  5.05e-02  5.52e-02
## time2           -0.00010 -8.96e-05 -7.91e-05
## treatozone      -0.13581  2.62e-01  6.60e-01
## Time:treatozone -0.00343 -2.29e-03 -1.15e-03
## attr(,"label")
## [1] "Coefficients:"
##
##  Correlation structure:
##          lower  est. upper
## cor(1,2) 0.945 0.964 0.977
## cor(1,3) 0.884 0.925 0.951
## cor(1,4) 0.828 0.884 0.922
## cor(1,5) 0.818 0.876 0.916
## cor(2,3) 0.961 0.974 0.983
## cor(2,4) 0.907 0.939 0.961
## cor(2,5) 0.895 0.932 0.956
## cor(3,4) 0.938 0.960 0.974
```

```
## cor(3,5) 0.918 0.948 0.967
## cor(4,5) 0.981 0.988 0.992
## attr(,"label")
## [1] "Correlation structure:"
##
##  Variance function:
##        lower   est. upper
## power -0.598 -0.183 0.231
## attr(,"label")
## [1] "Variance function:"
##
##  Residual standard error:
## lower  est. upper
## 0.445 0.864 1.679
```

The variances of the estimates are attainable, so we do not seem to have an over-specified model.

(h) How does the ozone treatment affect growth of these trees?



There is strong evidence that the relationship between time and mean size (on the log scale) depends on the treatment group (p-value<0.0001 *from t-stat*$=3.953$ *on* 390 *df). We conclude that ozone treatment does have an affect on the growth rate of these trees.*

For trees in the control group, there is strong evidence of curvature in the relationship between time and and size on the log scale (p-value<0.0001 *from t-stat*$=-16.77$ *on* 390 *df). At a time of* 152 *days, a one day increase in time is estimated to be associated with a* 1.023 *factor change in the median size of trees. At a time of* 201 *days, a one day increase in time is estimated to be associated with a* 1.014 *factor change in the median size of trees. At a time of* 258 *days, a one day increase in time is estimated to be associated with a* 1.004 *factor change in the median size of trees. The rate of growth is estimated to decrease within the time span studied.*

For trees in the ozone group, there is also strong evidence of curvature in the relationship between time and size on the log scale (p-value<0.0001 *from t-stat*$=-16.77$ *on* 390 *df.) At a time of* 152 *days, a one day increase in time is estimated to be associated with a* 1.021 *factor increase in the median size of trees. At a time of* 201 *days, a one day increase in time is estimated to be associated with a* 1.012 *factor increase in the median size of trees. At a time of* 258 *days, a one day increase in*

time is estimated to be associated with a 1.002 *factor change in the median size of trees. Overall, we see that the rate of growth is estimated to be slower when trees are grown in an ozone enriched chamber.*

R Code

```
require(ggplot2)
rats <-  read.csv("http://www.math.montana.edu/~jimrc/classes/stat505/data/drugResponse.csv")
rats$ratID <-  factor(rats$ratID)
rats$preRate <- factor(rats$preRate, labels=c("lo","med","hi"))
rats$dose <- factor(rats$dose)
```

```
qplot(x=dose, y = postRate, data = rats, geom="boxplot", facets=~preRate)+theme_bw()
```

```
split.fit <- aov(postRate~dose*preRate+Error(ratID), data=rats)
require(xtable)
xtable(summary(split.fit))
```

```
require(nlme)
rat.gls <- gls(postRate~dose*preRate, data=rats, correlation=corCompSymm(form=~1|ratID))
require(xtable)
xtable(anova(rat.gls))
```

```
rat.wrong <- lm(postRate ~  preRate * dose * ratID,data=rats)
require(xtable)
xtable(anova(rat.wrong))
```

```
vcov <- getVarCov(rat.gls)
vcov
```

```
with(rats,interaction.plot(dose,preRate,postRate))
```

```
split.fitadd <- aov(postRate~preRate+dose+Error(ratID), data=rats)
require(xtable)
xtable(summary(split.fitadd))
```

```
data(Sitka, package="MASS")
qplot(x=Time, y = size, data = Sitka, group=tree, geom=c("point","line")) + theme_bw() +facet_grid(. ~ treat)
## or
#require(lattice)
#print(xyplot(size ~ Time | treat, data = Sitka, group = tree, type = "l"))
```

```
require(MASS)
time2 <- Sitka$Time^2
sitka.gls <- gls(size~Time+time2, data=Sitka)
require(xtable)
xtable(anova(sitka.gls))
sitka.gls2 <- with(Sitka,update(sitka.gls, .~.*treat))
xtable(anova(sitka.gls2))
```

```
plot(ACF(sitka.gls2, form=~1|tree),alpha=0.05)
```

```
sitka.gls3 <- update(sitka.gls2, correlation=corAR1(form=~1|tree))
```

```
sitka.gls4 <- update(sitka.gls2, correlation=corCompSymm(form=~1|tree))
```

```
sitka.gls5 <- update(sitka.gls2, correlation=corSymm(rep(.9,10),form=~1|tree))
```

```
require(xtable)
anova(sitka.gls2, sitka.gls3, sitka.gls5)
anova(sitka.gls2, sitka.gls4, sitka.gls5)
```

```
require(xtable)
require(MASS)
sitka.gls5 <- update(sitka.gls5, weights=varPower())
xtable(summary(sitka.gls5)$tTable)
sitka.gls51 <- gls(size~Time+time2+treat+Time*treat, data=Sitka, correlation=corSymm(rep(.9,10),form=~1|tree), weights=varPower(
treat2 <- factor(Sitka$treat,levels=c("ozone","control"))
sitka.gls52 <- gls(size~Time+time2+treat2+Time*treat2, data=Sitka, correlation=corSymm(rep(.9,10),form=~1|tree), weights=varPowe
xtable(summary(sitka.gls51)$tTable)
xtable(summary(sitka.gls52)$tTable)
```

```
plot(sitka.gls51)
qqnorm(resid(sitka.gls51, type="n"))
qqline(resid(sitka.gls51, type="n"))
```

```
intervals(sitka.gls51)
```

```
with(Sitka, interaction.plot(Time,treat,size))
```