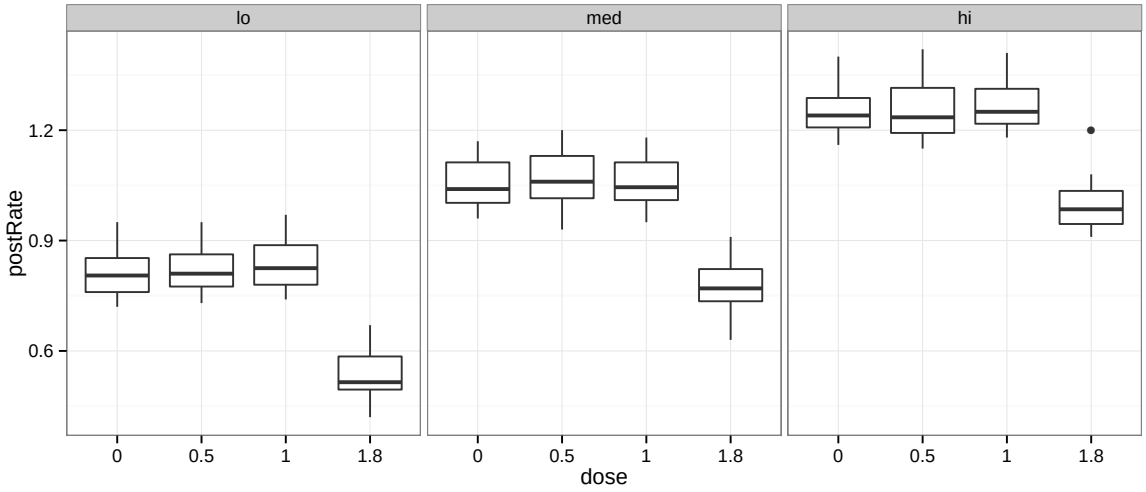
Stat 505 Assignment 5 Fall 2014

Solutions

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, and the order of treatments was randomized for each rat. Time intervals between treatments were long enough to remove any carryover effects.



There is a strong difference in postRate between the three preRate groups, and a strong difference in postRate between the highest doses of the drug and the three lower doses. It appears to be a threshold effect. The drug doesn't kick in til you get over 1.5 mg/Kg, and then it has the same negative effect within each preRate group.

(a) 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).

$$y_i = \mu + \alpha_{j[i]} + \tau_{k[i]} + \delta_{j[i]k[i]} + b_{m[i]} + \epsilon_i$$

Where i goes from 1 to 96 to specify row number, $j[i] = 1, 2, 3$ tells us which prerate group that rat belongs to, $k[i] = 1, 2, 3, 4$ is the dose given this rat in row i. The α's are preRate effects, τ's are dose effects, and δ's are interactions betweem prerate and dose. $b_{m[i]}$ is the random effect for the mth (1 to 12) rat which has $N(0, \sigma_b^2)$ distribution. Finally the residual, $\epsilon_i \sim N(0, \sigma^2)$. The ϵ_i's are iid as are the b_m, and the two levels of variation are independent.

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

i. Explain the structure of the matrix.
Block diagonal with a block $\mathbf{\Sigma}_m$ (8 by 8) for each of the 12 rats. Within $\mathbf{\Sigma}_m$ we have $\sigma^2 + \sigma_b^2$ on the diagonal and σ_b^2 off diagonal.

ii. What are the variances of all responses?

$$\sigma^2_{total} = \sigma^2 + \sigma^2_b$$

iii. What are the covariances between each possible pair of responses?

σ^2_b for two observations on the same rat, 0 for observations on different rats.

(c) Is a 'split plot' analysis appropriate for these data? Explain.

A split plot agricultural experiment uses whole plots as blocks for the split plot treatment. Here we do not assign a treatment at the rat level, we just observe the rats' pressing speed before any intervention. Like a split plot treatment, dose is assigned at random within each rat, so rats are like the 'whole plot' factor.

(d) Fit the above model to the data using `aov` in R using RatID 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
preRate:dose	6	0.01	0.00	0.59	0.7378
Residuals	75	0.14	0.00		

We can test for differences in mean postRate response depending on the preRate level in the "rat" or "whole plot" stratum, but it seems a bit silly, since preRate was observed, not assigned. The F test on 2 and 9 df has a value of 34 and a small p-value. The effect of Dose is strong ($F = 251$ on 3, 75 df with p-value $< .0001$), and the interaction between Dose and preRate is weak ($F = 0.59$ on 6, 75 df, p-value $> .5$).

(e) Fit the model using `gls` and the correct correlation specification in R. Compare F tests with those from aov.

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

The correct structure is "compound symmetric", and the F statistics and p-values agree exactly. It does not say which denominator is used, but it must have used the same ones as in aov output.

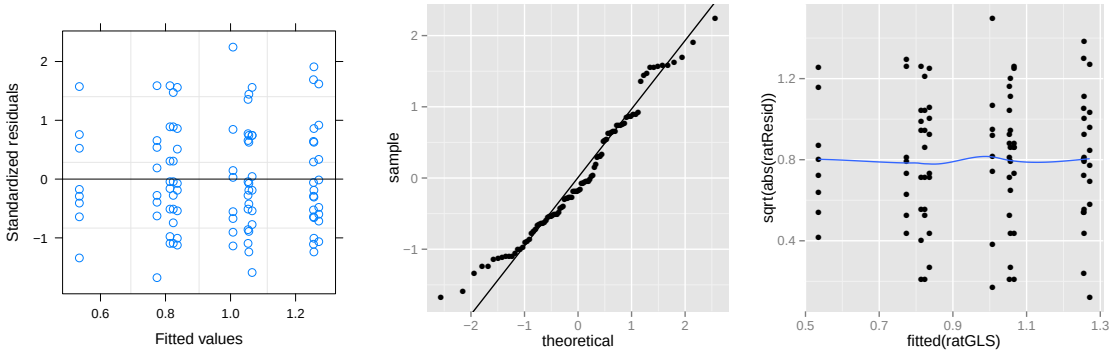
(f) Run `anova(lm(postRate ~ preRate * dose * ratID,data=rats))` and explain where each line of output (its df and its Sum Sq) shows up in the table produced by `aov`. Which lines provide the proper F test to test for effects of preRate?

I'll refer to the `aov` *output as table 1 and the* `anova(lm())` *output as table 2. The preRate lines are the same in both tables. The first Residuals line in table 1 is the ratID line in table 2. The dose line agrees in both tables, as does preRate:dose. Table 2 splits the df and SSq for Residuals from Table 1 into a piece labeled dose:ratID ($3 \times 9 = 27$ df SSq = 0.01) and a residuals line with 48 df and ssq=0.12 which is the replication error. If we use higher accuracy than I've shown the SSq's also add up to give 0.1364 in the aov Residuals row.*

	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		

(g) Explain what we've learned about the drug effects from these data. Does dose interact with initial 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?

```
## geom_smooth:  method="auto" and size of largest group is <1000, so using loess.  Use 'method = x' to
change the smoothing method.
```



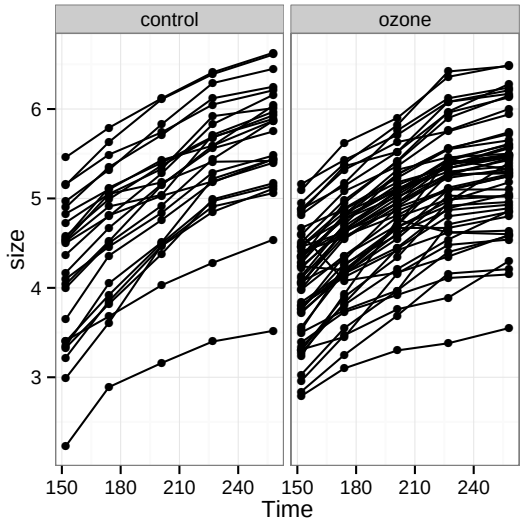
From the plot of the raw data we can see that there are large differences between the three groups of preRate rats, and that the drug has little to no effect until we get to the highest dose. Then it kicks in and decreases postRate by about 0.3 units. No preRate by Dose is visible and the large p-value confirms the fact that preRate and Dose effects seem to be simply additive. I'm not saying we should drop the interaction because this is not a model building exercise. We needed to see if there was an interaction, and we found that there is not one.

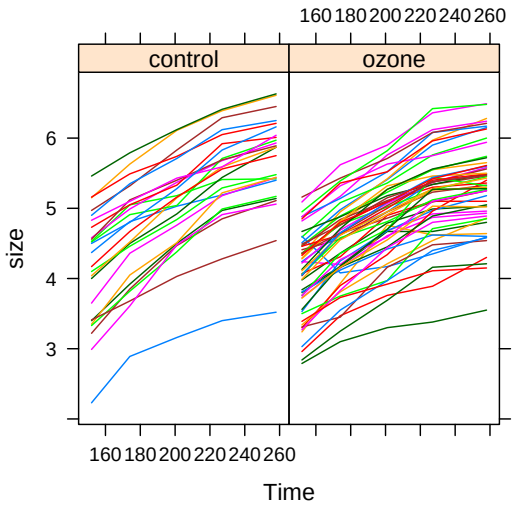
Diagnostic plots show no problem with the usual assumptions of constant variance and normality.

Because dose was randomly assigned, we can infer that high dose does cause a decrease in pressing rate. We know nothing about how these 12 rats were selected. I assume they are a convenience sample from available lab rats. Our statistical inference is limited to the sample unless someone can argue that these are a genetically pure strain of rats, and the inference extends to all rats of this type.

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.





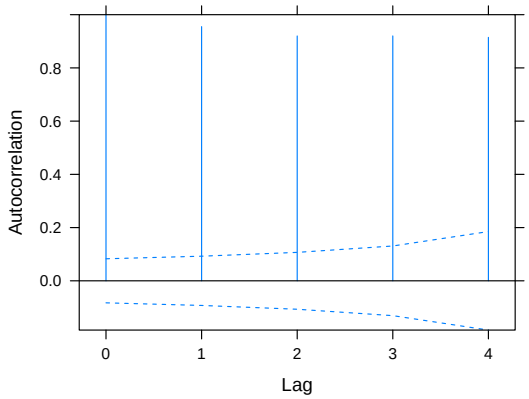
The rate of growth seems to be decreasing. The curves are getting flatter. There is wide variation in starting sizes at the first observation, with control group seeming to have a larger mean size. It's difficult to tell if slopes between the two groups vary, but we do see lots of variation between trees.

(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.

	Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
update(sitka.gls1, method = "ML")	1	4	766.62	782.53	-379.31			
update(sitka.gls2, method = "ML")	2	7	761.54	789.39	-373.77	1 vs 2	11.08	0.01

There is strong evidence ($\chi^2_3 = 11.08$, p-value $= 0.0113$) that a linear model with treatment effects is preferred over the model with just a single quadratic curve.

(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?



Strong correlations are evident at lags 1, 2, 3, and 4 within the same tree. It looks like a compound symmetric structure.

(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).

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	830.20	857.94	-408.10			
sitka.gls3	2	8	-42.88	-11.17	29.44	1 vs 2	875.07	0.00
sitka.gls5	3	17	-63.43	3.95	48.72	2 vs 3	38.56	0.00

	Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
sitka.gls2	1	7	830.20	857.94	-408.10			
sitka.gls4	2	8	106.09	137.80	-45.05	1 vs 2	726.10	0.00
sitka.gls5	3	17	-63.43	3.95	48.72	2 vs 3	187.53	0.00

Comparing AIC's, we note that each correlation structure is a big improvement over no correlation, decreasing AIC by at least 700 units. General symmetric structure has the smallest AIC by about 20 units over the nearest competitor: AR1. Compound symmetric does not fit as well as the AR1, though both use one parameter to model all correlations. I'm surprised that general symmetric is this much better than AR1 or CS, because if you print the intervals, they all overlap substantially.

(e) As trees get bigger, sizes might get more variable. Check to see if that is the case, and if so, adjust your model.

	Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
sitka.gls5	1	17	-63.43	3.95	48.72			
sitka.gls6	2	18	-62.16	9.19	49.08	1 vs 2	0.72	0.39

Adding varPower weights did not improve the fit, giving a $\chi^2_1 = 0.72$ *and a p–value of 0.40. We can omit weights.*

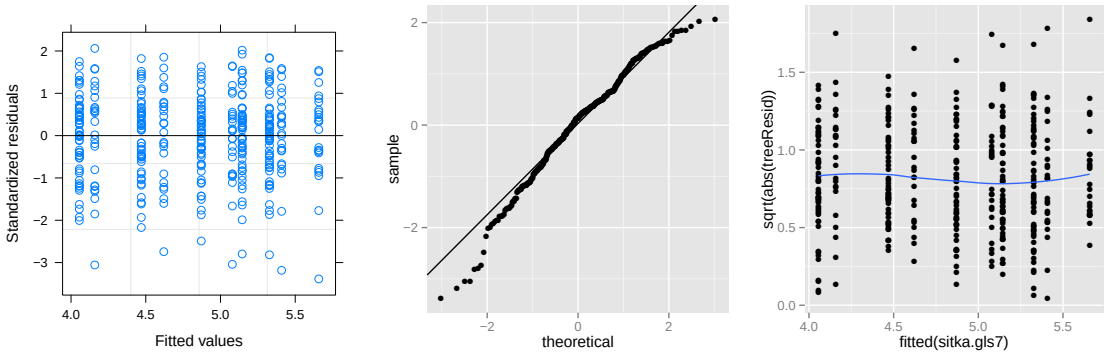
(f) Now we should have a model with lots of treatment terms and a reasonable variance-covariance structure, since you've looked at several correlation models and allowed for increasing variance as trees get bigger. Now examine the treatment effects. Are all of the terms in the model formula needed? Reduce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

The interaction between treatment and time^2 *is not needed, so I removed it. I would not remove the treatment effect, even though it has a large p-value because it's needed for the interaction with time. Conclusion: there is a treatment effect on the growth rate (slope), see below for more discussion.*

(g) Diagnostic plots.

	numDF	F-value	p-value
(Intercept)	1	4880.57	0.00
Time	1	3363.29	0.00
I((Time - 202)^2)	1	136.39	0.00
treat	1	2.01	0.16
Time:treat	1	20.68	0.00
I((Time - 202)^2):treat	1	0.22	0.64

	numDF	F-value	p-value
(Intercept)	1	4905.69	0.00
Time	1	3378.23	0.00
I((Time - 202)^2)	1	137.00	0.00
treat	1	2.02	0.16
Time:treat	1	20.77	0.00



There are a few small residuals, and slight evidence of a long left tail, but I think the data is close to normally distributed.

The intervals command works without generating an error, so we do not have an over-specified model.

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

The estimated decreases in growth rate .00229 (log volume units)/day, but I'd rather report a 95% confidence interval estimate of (-0.0012, -0.0034) for the ozone effect on growth rate. We could rescale that to a change in the annual growth rate of (-0.422, -1.25). Growth rates decrease as the trees get bigger, but the curvature (rate of change in the slope) is the same for both groups of trees.

I assume that the trees used are a convenience sample of available trees, so our inference applies only to this sample (unless experts can argue otherwise). The treatment was assigned at random, so we can say that ozone treatment causes a decrease in growth rate.

I do wonder how many growth chambers were used in the study. I fear that all controls were in one chamber, and all ozone trees in another. Then treatments were really assigned to chambers, and we have only two independent units with pseudo–replication inside each.

R Code

```
require(xtable, quietly = TRUE)
require(nlme, quietly = TRUE)
require(lattice, quietly = TRUE)
require(ggplot2, quietly = TRUE)
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()
```

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

```
ratGLS <- gls(postRate ~ preRate * dose, rats, cor = corCompSymm(form = ~1 | ratID))
xtable(anova(ratGLS))
```

```
xtable(anova(lm(postRate ~ preRate * dose * ratID, rats)))
```

```
data(Sitka, package = "MASS")
qp1 <- qplot(x = Time, y = size, data = Sitka, group = tree, geom = c("point", "line")) + theme_bw() +
    facet_grid(. ~ treat)
pp2 <- xyplot(size ~ Time | treat, data = Sitka, group = tree, type = "l")
grid.arrange(qp1, pp2, ncol = 2)
```

```
sitka.gls1 <- gls(size ~ Time + I((Time - 202)^2), data = Sitka)
sitka.gls2 <- update(sitka.gls1, . ~ . * treat)
xtable(anova(update(sitka.gls1, method = "ML"), update(sitka.gls2, method = "ML"))[, 2:9])
```

```
plot(ACF(sitka.gls2, form = ~1 | tree), alpha = 0.1)
```

```
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(form = ~1 | tree))
xtable(anova(sitka.gls2, sitka.gls3, sitka.gls5)[, 2:9])
xtable(anova(sitka.gls2, sitka.gls4, sitka.gls5)[, 2:9])
```

```
xtable(anova(update(sitka.gls4, method = "ML")))
sitka.gls7 <- update(sitka.gls4, . ~ . - I((Time - 202)^2):treat)
xtable(anova(update(sitka.gls7, method = "ML")))
```

```
treeResid <- resid(sitka.gls7, type = "pearson")
defaultPlot <- plot(sitka.gls7)
qqplotR <- qqGG2(treeResid)
locScalePlot <- qplot(x = fitted(sitka.gls7), y = sqrt(abs(treeResid))) + geom_smooth(level = 0)
grid.arrange(defaultPlot, qqplotR, locScalePlot, ncol = 3)

## geom_smooth:  method="auto" and size of largest group is <1000, so using loess.  Use 'method = x' to change the
smoothing method.
```

```
intervals(sitka.gls7)
```

```
## Approximate 95% confidence intervals
##
##  Coefficients:
##                       lower      est.     upper
## (Intercept)        1.85e+00  2.14e+00  2.43e+00
## Time               1.39e-02  1.46e-02  1.54e-02
## I((Time - 202)^2) -9.51e-05 -8.15e-05 -6.79e-05
## treatozone        -1.24e-01  2.22e-01  5.68e-01
## Time:treatozone   -3.05e-03 -2.14e-03 -1.22e-03
## attr(,"label")
## [1] "Coefficients:"
##
##  Correlation structure:
##     lower  est. upper
## Rho 0.909 0.934 0.953
## attr(,"label")
## [1] "Correlation structure:"
##
##  Residual standard error:
## lower  est. upper
## 0.543 0.631 0.733
```