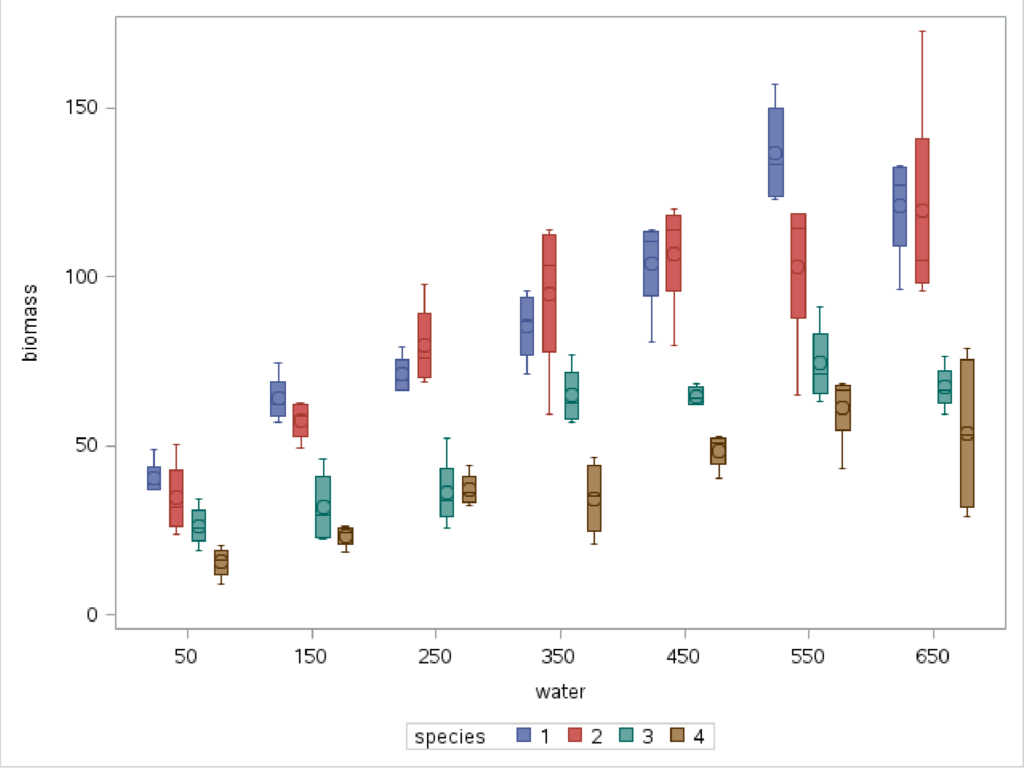
Stat 506 Assignment 2

Leslie Gains-Germain

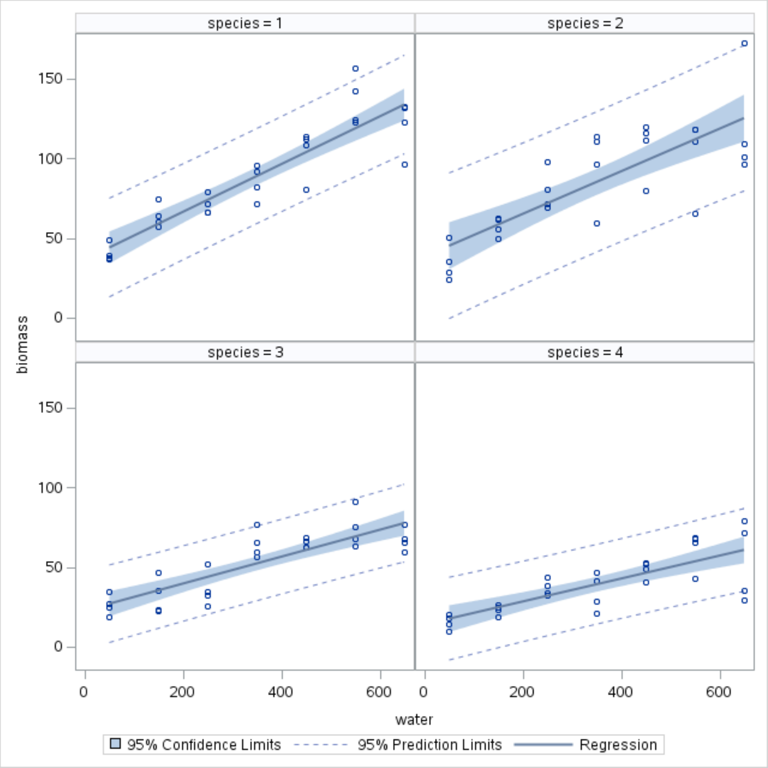
January 30, 2015

1. Researchers in Jordan are interested in plants useable for animal fodder which require little moisture. They tested four plant species in a greenhouse experiment varying the daily watering from 50 to 650 mm in 100 mm increments. Within each species, water amounts were allocated at random. The response is dry biomass. Read in the data from here.

   (a) Use SAS to plot the data in a manner which allows us to easily compare mean biomass for each species as a function of water. *Note the dot at the mean for each combination of species and water in the boxplots below.*

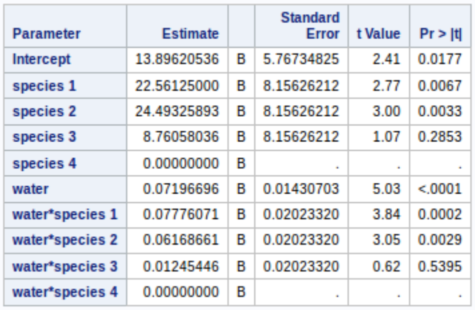


   (b) Also plot biomass by water separating them into panels by species and adding a regression line to each panel.



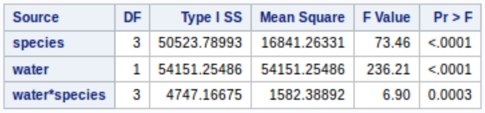
(c) Fit a model with an intercept and a slope for each species. $y_i = \beta_0 + \beta_1 x_i + \alpha_{0j[i]} + \alpha_{1j[i]} x_i + \epsilon_i$. In this notation, i is the row number, j[i] is the plant species of the plant in row i (j = 1 to 4), β's are overall effects, and α's are adjustments for each plant species.

i. Fit this model in SAS and show the estimated coefficients.



Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	13.89620536	B	5.76734825	2.41	0.0177
species 1	22.56125000	B	8.15626212	2.77	0.0067
species 2	24.49325893	B	8.15626212	3.00	0.0033
species 3	8.76058036	B	8.15626212	1.07	0.2853
species 4	0.00000000	B	.	.	.
water	0.07196696	B	0.01430703	5.03	<.0001
water*species 1	0.07776071	B	0.02023320	3.84	0.0002
water*species 2	0.06168661	B	0.02023320	3.05	0.0029
water*species 3	0.01245446	B	0.02023320	0.62	0.5395
water*species 4	0.00000000	B	.	.	.

ii. Provide either the Type I or Type III output table, and explain why you think this table is preferred. Is the interaction needed?



Source	DF	Type I SS	Mean Square	F Value	Pr > F
species	3	50523.78993	16841.26331	73.46	<.0001
water	1	54151.25486	54151.25486	236.21	<.0001
water*species	3	4747.16675	1582.38892	6.90	0.0003

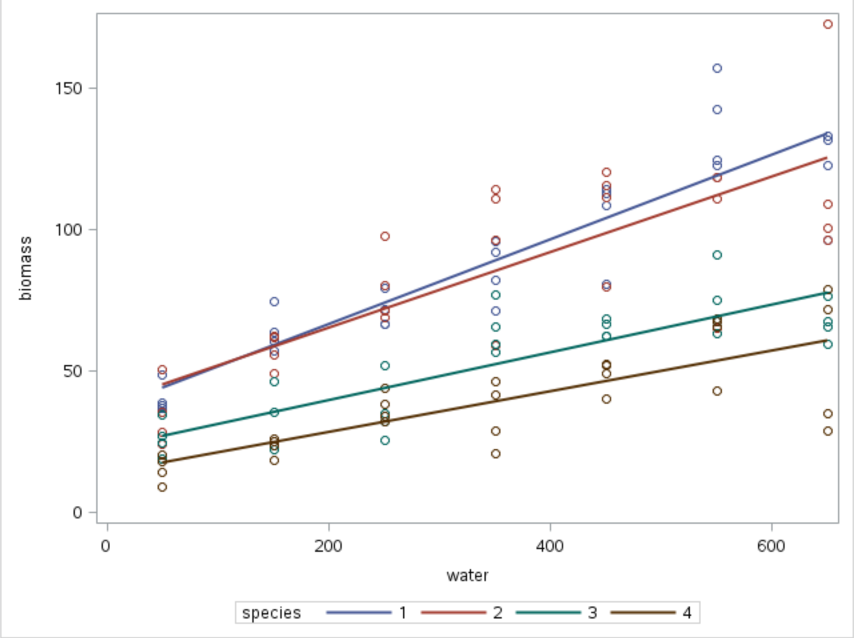
I prefer the type I sums of squares because I think the sequential sums of squares provide useful information. Consider the water row in the above table. The p-value in this row compares a model with species and water and predictors to a model with just species as a predictor. I think this p-value could be meaningful in some situations. The p-value for the water row in the type III sums of squares table is not meaningful because we would almost never consider a model with a *water*species interaction but no water main effect.*

Yes, the interaction is needed. There is strong evidence that the relationship between mean biomass and water depends on species (p-value= 0.0003*).*

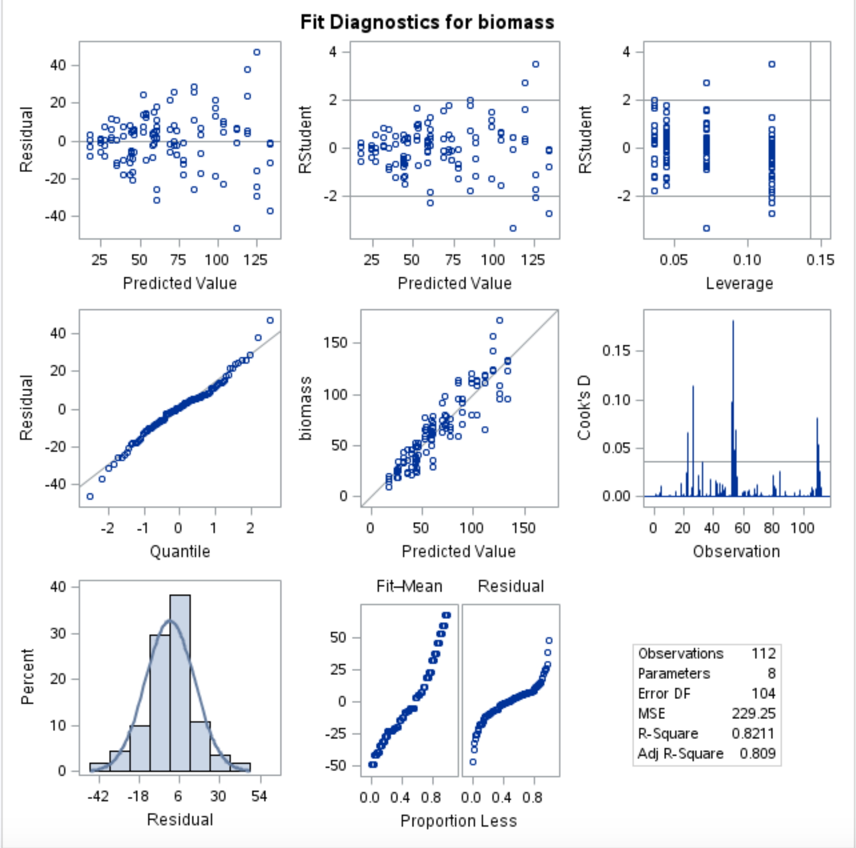
iii. What combination of Greek letters is estimated by each coefficient shown?

Intercept	$\beta_0 + \alpha_{04}$
Species 1	$\alpha_{01} - \alpha_{04}$
Species 2	$\alpha_{02} - \alpha_{04}$
Species 3	$\alpha_{03} - \alpha_{04}$
water	$\beta_1 + \alpha_{14}$
water*species 1	$\alpha_{11} - \alpha_{14}$
water*species 2	$\alpha_{12} - \alpha_{14}$
water*species 3	$\alpha_{13} - \alpha_{14}$

iv. Plot the four lines from the interaction model.



v. Provide the default diagnostic plots and comment on how well the assumptions are met. No random effects on this one, but do comment on your plots and models.

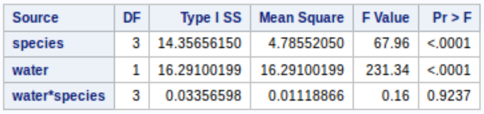


Observations	112
Parameters	8
Error DF	104
MSE	229.25
R-Square	0.8211
Adj R-Square	0.809

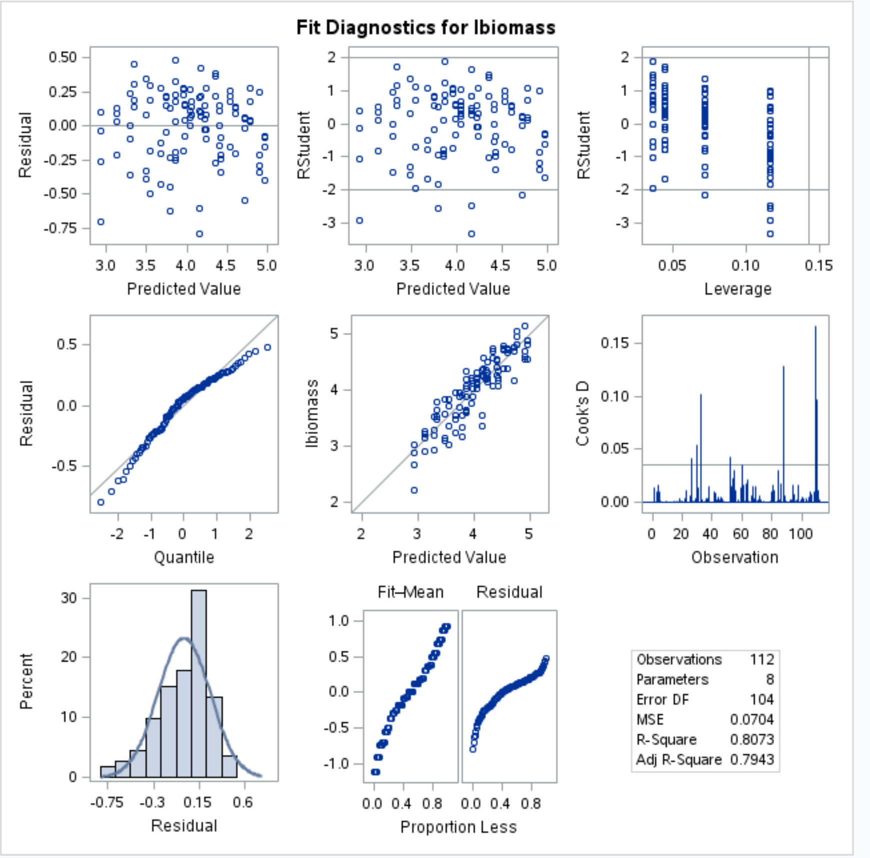
The residual vs. predicted values plot shows a funnel shape, which indicates that the constant variance assumption is violated. I'm hesitant to trust the p-values with these severe of a violation. We also see that the distribution of responses is heavy tailed, which is a violation that can interfere with our analysis. I did a log transformation to see how that would change our results.

The funnel pattern is no longer evident in the residuals vs. predicted values plot, but there are more large negative residuals than we would expect under normality. This is confirmed by the long left tail we see in the histogram of log(responses). It looks like the log transformation overdid it slightly.

The most notable change, however, is that the p-value for the interaction term is now 0.9237*! So perhaps the interaction between water and species isn't needed. I think we should take more time to think about what model is appropriate for this situation.*



Source	DF	Type I SS	Mean Square	F Value	Pr > F
species	3	14.35656150	4.78552050	67.96	<.0001
water	1	16.29100199	16.29100199	231.34	<.0001
water*species	3	0.03356598	0.01118866	0.16	0.9237



Observations	112
Parameters	8
Error DF	104
MSE	0.0704
R-Square	0.8073
Adj R-Square	0.7943

(d) Fit the same model in R and compare your conclusions from both fits.

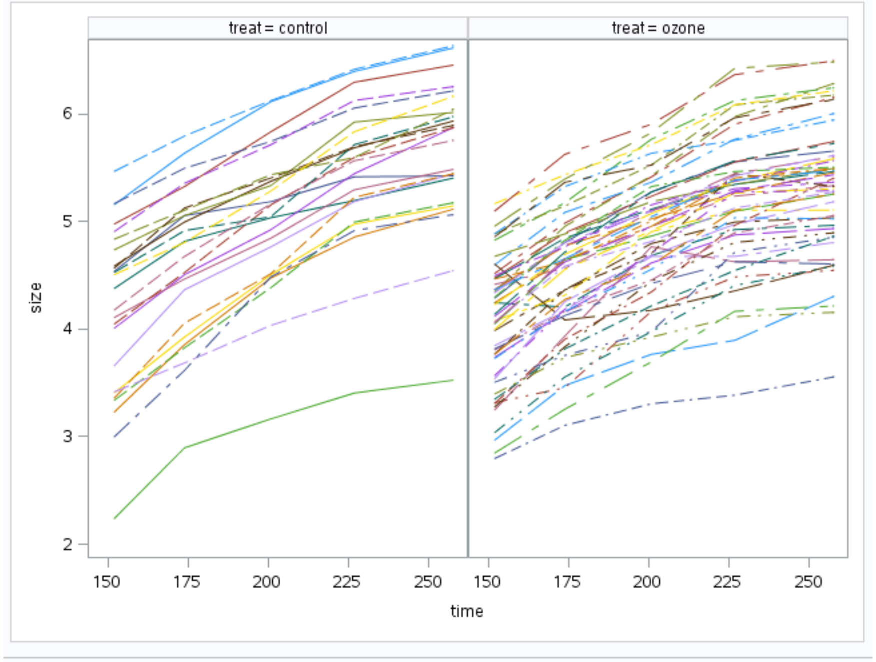
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	13.8962	5.7673	2.41	0.0177
species1	22.5612	8.1563	2.77	0.0067
species2	24.4933	8.1563	3.00	0.0033
species3	8.7606	8.1563	1.07	0.2853
water	0.0720	0.0143	5.03	0.0000
species1:water	0.0778	0.0202	3.84	0.0002
species2:water	0.0617	0.0202	3.05	0.0029
species3:water	0.0125	0.0202	0.62	0.5395

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
species	3	50523.79	16841.26	73.46	0.0000
water	1	54151.25	54151.25	236.21	0.0000
species:water	3	4747.17	1582.39	6.90	0.0003
Residuals	104	23842.42	229.25		

The ANOVA type I SS table is exactly the same when we fit the same model in R. I set species four as the reference level in R, and after doing so the table of estimates is exactly the same as the table of estimates given by SAS. Our conclusions are the same no matter which software package we choose.

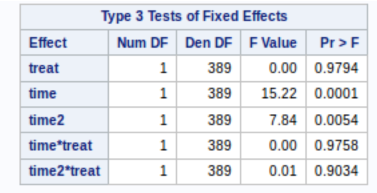
2. In Stat 505 HW5 we analyzed the Sitka data from the MASS package. Pull it into SAS, and rerun the analysis. Specifically:

   (a) Plot size over time, separating the two groups, and using a different line for each individual tree.



   (b) Use PROC MIXED to fit a quadratic model across all the data. Do we need intercept, slope, and/or quadratic coefficients to depend on treatment? As in R, use REML (the default) when comparing random effects, ML for fixed effects.

   There is weak evidence of curvature in the relationship between time and biomass (p-value= 0.0054*.) According to this model, there is no evidence that the slope or quadratic coefficients depend on treatment (p-values=* 0.9758 *and* 0.9034 *respectively).*



Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
treat	1	389	0.00	0.9794
time	1	389	15.22	0.0001
time2	1	389	7.84	0.0054
time*treat	1	389	0.00	0.9758
time2*treat	1	389	0.01	0.9034

   (c) The only ACF plot I'm finding in SAS is in PROC MI (multiple imputation), and I think it refers to MCMC sequences. That's not helpful. So we may have to do it blindly. Fit the same model(s) for correlation as you did in R by using PROC MIXED. Compare with the models in (a).

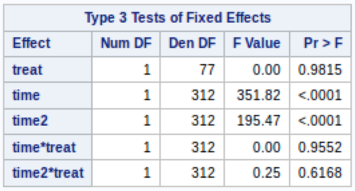
   I fit three different models with compound symmetric, AR(1), and symmetric correlation structures. I compared the AIC's of all four models I fit. The table below shows the AIC for each model. Note that each model included a time, time^2*, treatment, time*treatment, and* $time^2 * treatment$ *term.*

Model	AIC
Independent Errors	761.5
Compound Symmetric	29.7
AR-1	-118.4
Symmetric	-138.7

We see that the symmetric correlation structure has the lowest AIC. The AR-1 model also has a low AIC, so I will choose this model for inference because it is simpler(and has only one covariance parameter to estimate).

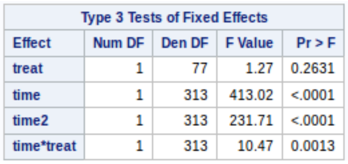
(d) After finding a model with reasonable correlation structure, see if you can prune back the fixed effects. Provide diagnostic plots and explain your model.

Below on the left is the table of coefficients for the model with AR-1 correlation structure. I chose to remove the time$^2 *$ *treatment term (p-value*$= 0.6168$*). After doing so, there is strong evidence that the relationship between time and mean biomass depends on treatment (p-value*< 0.0013*).*



Type 3 Tests of Fixed Effects

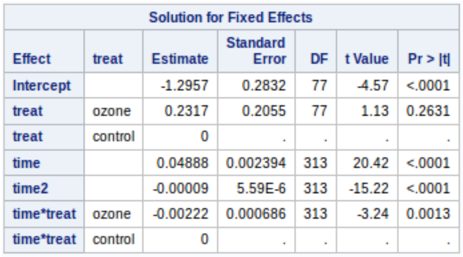
Effect	Num DF	Den DF	F Value	Pr > F
treat	1	77	0.00	0.9815
time	1	312	351.82	<.0001
time2	1	312	195.47	<.0001
time*treat	1	312	0.00	0.9552
time2*treat	1	312	0.25	0.6168



Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
treat	1	77	1.27	0.2631
time	1	313	413.02	<.0001
time2	1	313	231.71	<.0001
time*treat	1	313	10.47	0.0013

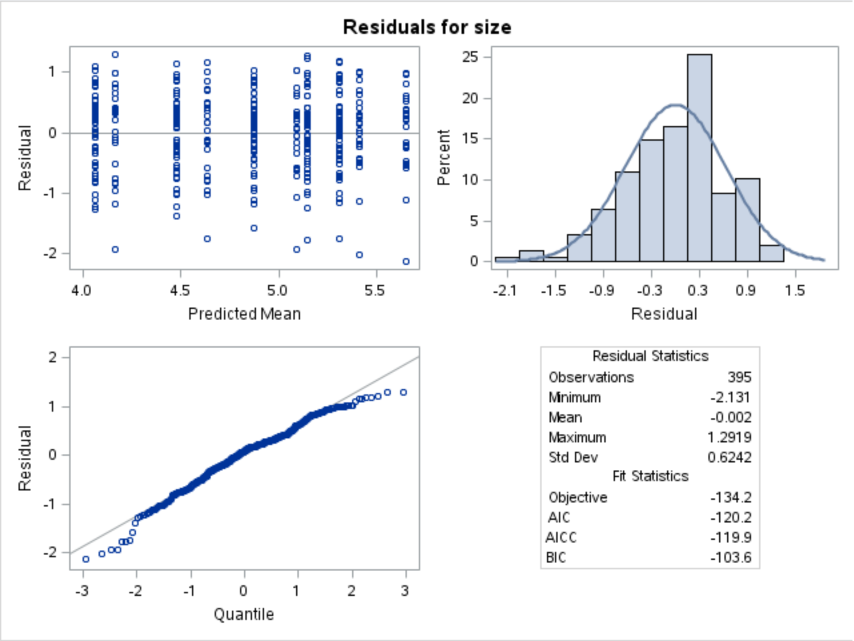
This means that there is evidence that the rate of tree growth depends on the treatment. Further examination of the coefficient estimates shows us that the rate of growth in an ozone environment is estimated to be 0.0022 *units lower than the rate of growth in the control environment.*



Solution for Fixed Effects

Effect	treat	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		-1.2957	0.2832	77	-4.57	<.0001
treat	ozone	0.2317	0.2055	77	1.13	0.2631
treat	control	0	.	.	.	.
time		0.04888	0.002394	313	20.42	<.0001
time2		-0.00009	5.59E-6	313	-15.22	<.0001
time*treat	ozone	-0.00222	0.000686	313	-3.24	0.0013
time*treat	control	0	.	.	.	.

The diagnostic plots do not show any severe violations of the model assumptions.



Residual Statistics	
Observations	395
Minimum	-2.131
Mean	-0.002
Maximum	1.2919
Std Dev	0.6242
Fit Statistics	
Objective	-134.2
AIC	-120.2
AICC	-119.9
BIC	-103.6

(e) Compare to the results from HW5 last fall.

I printed the results from the model I fit in the fall. The ANOVA table looks different because R gives us type I SS by default and SAS gives us type III SS by default in the PROC mixed procedure. Additionally, the coefficient estimates are slightly different. I chose to use a symmetric correlation structure in the fall, and I used varPower() to allow for increasing variance as the mean biomass increases. Biologically, this makes sense because as trees grow and age we would expect to see larger variation in biomass.

Overall, however, our conclusions are the same in both models. The ozone enriched environment does appear to be associated with slower growth rates of sitka trees.

	numDF	F-value	p-value
(Intercept)	1	5226.90	0.00
Time	1	1105.96	0.00
time2	1	281.03	0.00
treat	1	3.29	0.07
Time:treat	1	15.63	0.00

	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