STA 723 HW06 Report **Zach White**

Executive Summary

We analyze tree diameter data for a trees in Coweeta Hydrologic Laboratory. We are interested in inferring the pattern of tree growth over time and given certain covariates. We compare the results of a hierarchical model with a model whose parameters we estimate using generalized estimating equations (GEE), and we assess the advantages and disadvantages of using each of the models. When comparing the results for our strategies, we find some both harmonious and conflicting results. However, there are clear limitations to both of our models.

Introduction

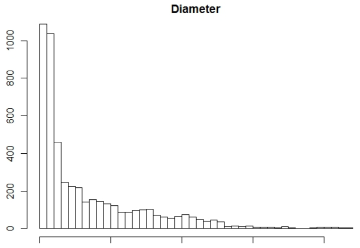
Tree growth provides important information about forest ecology. In general, we can gauge certain aspect of the health of a system based on tree growth. We have repeated tape measurements of the diameter of 848 trees from a mapped stand in the Coweeta Hydrologic Laboratory. They performed the census every one to four years. In the time between the censuses, trees died and were removed from the study, and trees were planted and added. We have 5 possible explanatory variables in total: ID, year (of measurement), annualprec (annual precipitation), summerpdsi (Summer Palmer Drought Severity Index, a standardized scale between -10(dry) and 10(wet)), and wintertemp (average winter temperature). The response variable of interest is cm (diameter of tree in centimeter).

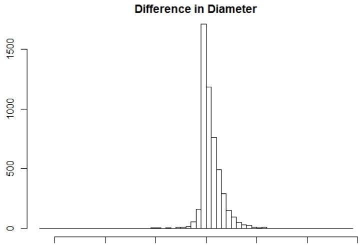
The overall goal of this analysis is to infer the pattern of tree growth. We are interested in both population (stand) and individual (each tree) level growth. These data present unique challenges both because of the way it is collected and because of the covariates of interest. We compare the results of a hierarchical model with generalized estimating equations (GEE). Each of these techniques has distinct advantages and disadvantages, and we discuss these.

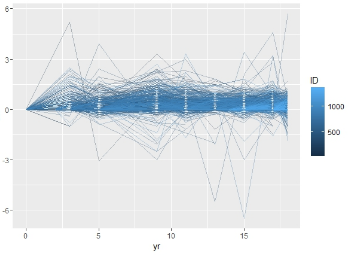
Analysis

Exploratory Data Analysis

The overall goal of this analysis is to infer patterns of tree growth. We are interested in the distribution of the response. The first two plots are histograms of diameter of tree and the change in diameter of the trees. The last plot shows the change in diameter by year and tree ID. As we analyze the first plot, it is clear that the raw diameter data is not normally distributed. It is clearly right skewed, and a large proportion of the values are very small. However, when we analyze the distribution of the change of cm between time intervals, it does look more normal, and they seem to centered around zero. Because of these differences in distribution, we choose to model the differences. Since we are both interested in tree growth for both site and stand levels, we want to visualize the change in diameter by both the covariate of interest and tree ID. As we do that, there appears to be a slight positive trend. Much of the data is positive, but there is not a clear trend. There is a lot of noise in this plot, which will pose some issues. When we visualize the







change in diameter by the other covariates and site ID, there aren't any patterns that are clearly identifiable. This is largely due to some issues with the data we have.

There are some clear issues with this data. One is that the time intervals are not consistent. They range from one to four years, which makes it difficult to understand what change we are modeling. Under the construct of our model that we will discuss, we are not modeling yearly change, but rather, we are modeling change over inconsistent time periods. For this reason, we will include time as a covariate. Another issue with the data is that all the covariates except tree ID are stand level. This means that there isn't variation in the covariates between the trees for the years of the census. This means that we are assuming to some degree that each tree responds to each of the covariates in the same way, and the entire stand will respond in the same way. This is an overly simplistic assumption because there are interactions between trees which will influence the growth of each of them. Due to the fact that the covariates are stand level, we are most likely missing some more tree specific variables like some spatial component for example. A spatial component could complete the data.

Model Choice

We propose and compare two different models. The first of which is a hierarchical model with random effects on each of the trees and fixed effects for the covariates years, annualprec, summerpdsi, wintertemp, and an interaction between annualprec and wintertemp. For our purposes, a hierarchical model is good because it not only allows us to analyze the conditional effects but also the marginal effects. However, hierarchical models are not robust for model misspecification. Thus, if our model isn't correct, inference will suffer. However, GEE is robust under model misspecification. Thus, if our model is incorrect, our inference can still be valid. Under GEE, though, we cannot perform inference on conditional effects, and random effects do not work within this framework.

Hierarchical Model

$$Y_{ij} - Y_{ij-1} = \alpha_i + \beta_1 \text{yr\_from} + \beta_2 \text{prec} + \beta_3 \text{pdsi} + \beta_5 \text{w.temp} + \beta_6 \text{prec} \times \text{w.temp} + \epsilon_{ij}$$
$$\epsilon_{ij} \sim N(0, \sigma^2)$$

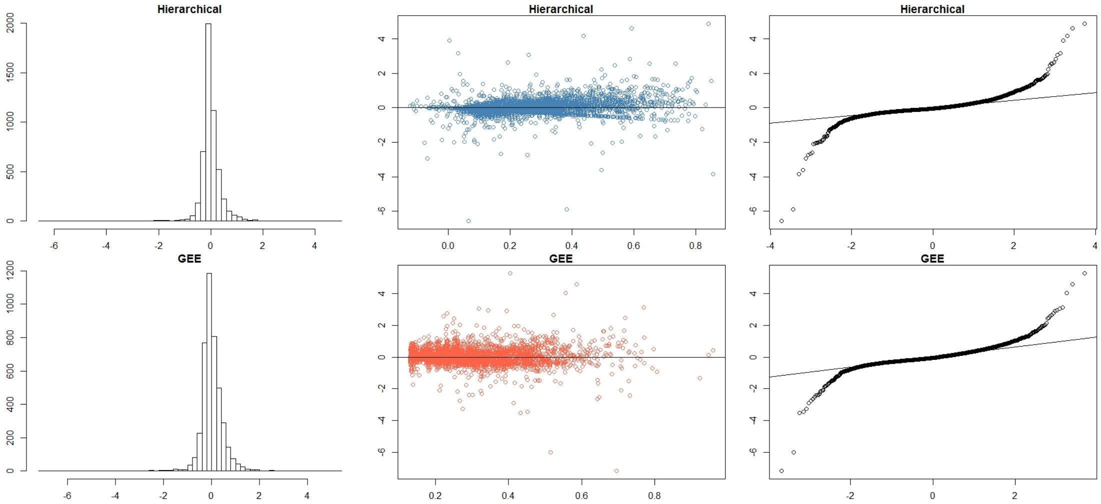
Model for GEE

$$Y_{ij} - Y_{ij-1} = \beta_0 + \beta_1 \text{yr\_from} + \beta_2 \text{prec} + \beta_3 \text{pdsi} + \beta_5 \text{w.temp} + \beta_6 \text{prec} \times \text{w.temp} + \epsilon_{ij}$$
$$\epsilon_{ij} \sim N(0, \sigma^2)$$

Under this model, each i represents a different tree, and j represents a different time frame. Thus, we are modeling the difference between the tree$_i$ at times j and $j-1$. We choose to include this interaction term because the other interaction terms include covariates that are highly correlated. Also, this interaction term makes intuitive sense, and we include in both our models.

Results

We will analyze the fit of the model using the following plots. The first column of plots shows the histogram of residuals. Both of these plots are centered about zero, and they seem to be at least somewhat normally distributed. However, there are also residuals significantly larger than would be expected under the normal distribution, which could indicate that the errors of this model might be distributed with a distribution with heavier tails, like a t-distribution. The second column shows fitted values versus residuals. If our model is correct, we look for random scatter about zero. There should not be a pattern. If there is a clear pattern, then it means there is some trend or relationship that



our model is missing. Both of these seem to be reasonable. The plot in the first row represents the hierarchical model and represents a possibly better fit of the data. The qq-plot, which is perhaps the most important diagnostic plot in our situation is the third column. Ideally, we want the points to follow the line. However, in both of our plots, there are clear deviations on the edges of the plot. This could indicate again that the model has heavier tails than we are modeling. Also, there might be some curve in the middle of the plot. This might indicate that this model is not appropriate for the data. Overall, it seems that these models that we have fit may not be appropriate.

	N.L	N.U	R.L	R.U	HM.L	HM.U
Intercept					-1.239	-0.242
Intercept	-0632	0.65	-0.745	0.76		
Year	-0.021	-0.011	-0.002	-0.011	-0.004	0.004
prec	-0.066	0.71	-0.143	0.786	0.295	0.899
pdsi	-0.085	0.03	-0.084	0.03	-0.145	-0.038
w.temp	-0.095	0.15	-.112	0.1635	-0.043	0.142
tree.int	0.005	0.007	0.004	0.008		
temp*prec	-0.11	0.036	-0.125	0.049	-0.125	0.049

As we analyze the confidence intervals above, year and some measure of intercept are significant under both models. There are multiple intercepts under GEE because under the model specification, we choose to include an intercept, which is the first measurement of each tree. The tree.int in the table above represents this. Under the hierarchical model, the intercept, prec, and pdsi are all significant. This is conflicting because these are not significant under the GEE model. From these intervals, we conclude that some measure of intercept and year from the beginning are significant.

Conclusion

These data present multiple challenges, which we don't fully address. For our goal of individual inference, we cannot use GEE directly, but our modified GEE formula with a specific intercept value can capture this. We can conclude that each individual is different and warrants its own intercept. As far as population level, we have mixed results because our hierarchical model captures significance of both precipitation and pdsi. Overall, the model diagnostics show these model are not great for the data. If we were to repeat this analysis, we would choose a different, more creative hierarchy to model the mean change. or possibly a ratio of the change.