Comparing GEE and Hierarchical Models in Tree Diameter

Zach White

STA 723 Case Studies

February 28, 2017

Overview

Introduction and EDA

Model Choice

Model Comparison

Conclusion

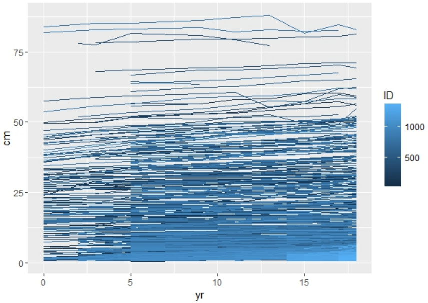
Introduction

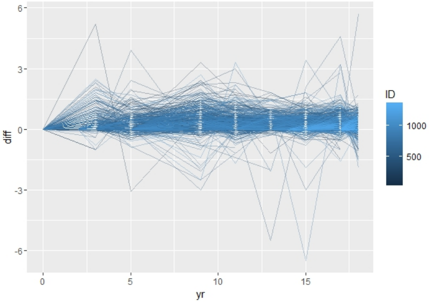
- Tree growth provides essential information about forest ecology.
- We can estimate tree growth based on repeated tape measurements of the diameter of the same tree.
- We want to analyze the data to infer the pattern of tree growth.
  - Population level
  - Individual level
- We compare two possible methods to achieve these goals:
  - Hierarchical Models
  - Generalized Estimating Equations (GEE)

Data

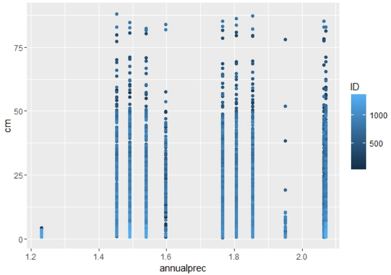
- We have diameter measurements for a large number of trees obtained from a mapped stand in Coweeta Hydrologic Laboratory
- The data span from 1993-2010
  - Censuses were conducted at intervals of one to four years
  - Each year, some trees died and were removed from the census, and some trees were planted and added to the census.]
- We have the following variables:
  - cm: Diameter of tree
  - ID: A unique ID for each tree
  - year: The year of the measurement
  - annualprec: Annual precipitation
  - summerpdsi: Average summer Palmer Drought Severity Index (-10 (dry) to 10 (wet))
  - wintertemp: Average winter temperature

EDA

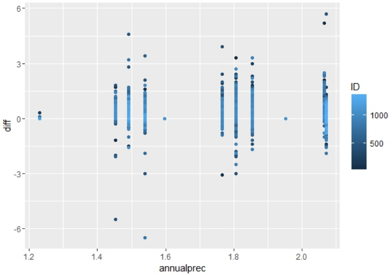

75
50
25
0
cm
0
5
10
15
yr
ID
1000
500


6
3
0
-3
-6
diff
0
5
10
15
yr
ID
1000
500

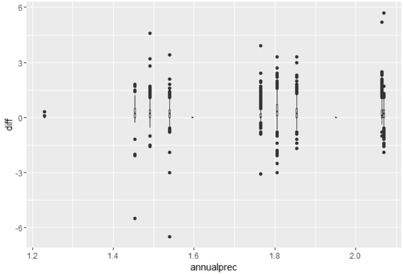
EDA



cm
0
25
50
75
annualprec
1.2
1.4
1.6
1.8
2.0
ID
1000
500

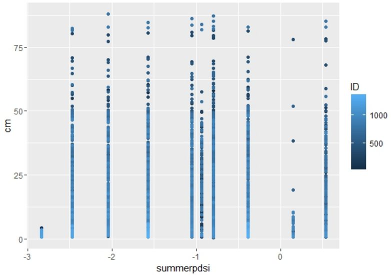


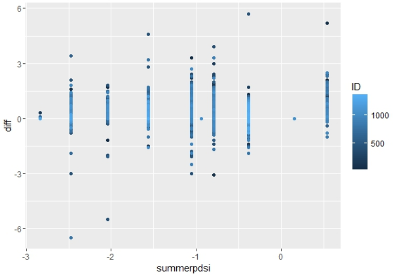
diff
6
3
0
-3
-6
annualprec
1.2
1.4
1.6
1.8
2.0
ID
1000
500

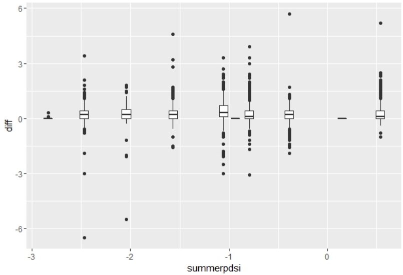


diff
6
3
0
-3
-6
annualprec
1.2
1.4
1.6
1.8
2.0

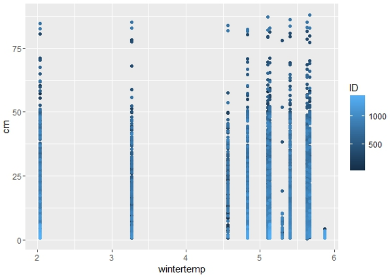
EDA

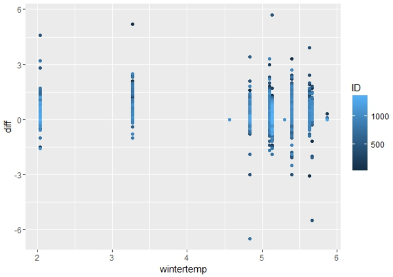

cm
75
50
25
0
-3
-2
-1
0
summerpdsi
ID
1000
500

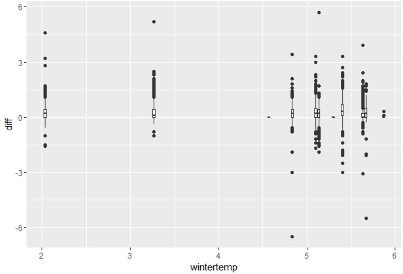

diff
6
3
0
-3
-6
-3
-2
-1
0
summerpdsi
ID
1000
500


diff
6
3
0
-3
-6
-3
-2
-1
0
summerpdsi

EDA


cm
75
50
25
0
2
3
4
5
6
wintertemp
ID
1000
500


diff
6
3
0
-3
-6
2
3
4
5
6
wintertemp
ID
1000
500


diff
6
3
0
-3
-6
2
3
4
5
6
wintertemp

EDA Conclusions

- We are missing a lot of years.
- The measurements for the covariates are perfectly aligned with the year, and we are probably missing parts of the story. The inconsistency of measurements may make it difficult to identify trends.
    - For example, a very wet year, like 1996 might be followed by growth that is observed in the following year, but we just have data in 1998.
    - Overly simplistic covariates
- Certain species may respond to changes in different ways, which might lead to a better nesting structure.
- There are some trees with fewer measurements i.e. ID multiple trees with just two measurements.

GEE vs. Hierarchical

Hierarchical Model

Advantages

- Inference on conditional and marginal effects

Disadvantages

- Not robust to Model misspecification

GEE

Advantages

- Robust to model specification
- Population Effects

Disadvantages

- Inference on conditional effects
- Dependent on number of groups

Hierarchical model

For a given tree diameter Y_ij, where $i =$ site ID, and $j =$ index of measurement.

$$Y_{ij} = \alpha_i + \beta_1 \text{prec}_{ij} + \beta_2 \text{pdsi}_{ij} + \beta_3 \text{wtemp} + \beta_4 \text{year} + \epsilon_{ij}$$

$$\epsilon_{ij} \sim N(0, \sigma^2)$$

GEE

Goal: We are trying to minimize

$$U(\beta) = \sum_{i=1}^{N} \boldsymbol{D_i^T V_i^{-1} (y_i - \mu_i)}$$

where

$$\boldsymbol{D_i = \frac{\delta \mu_i}{\delta \beta_i}} \quad V_i = (A_i^{1/2} R_i A_i^{1/2})\phi \tag{1}$$

where V_i is the working covariance matrix of Y_i.
$A_i = \text{diag}(\text{var}(Y_{ij}))$ and R_i is the correlation matrix for Y_i and ϕ is an overdispersion parameter we can estimate.
We find $\hat{\beta}$ iteratively by proposing V_i, finding $\hat{\boldsymbol{\beta}}$ and then estimating a new V_i based on the residuals.

Model Validation

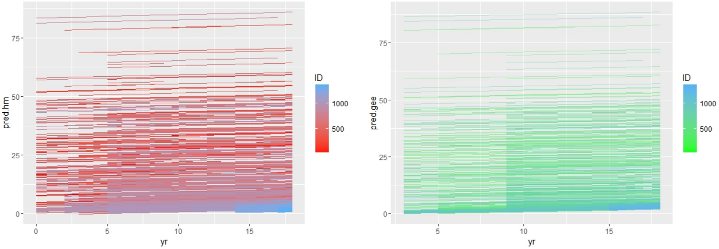


Figure 1: HM and GEE

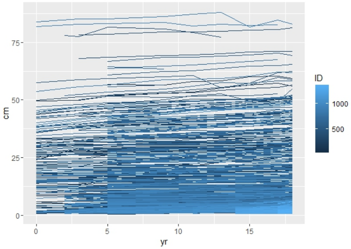
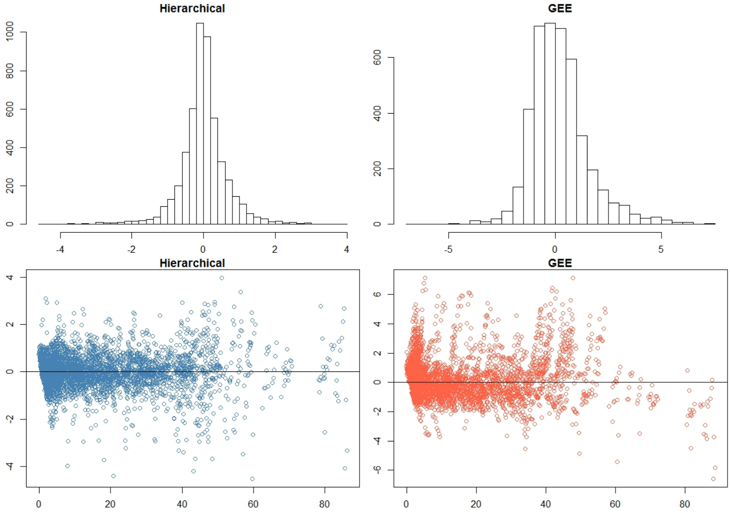
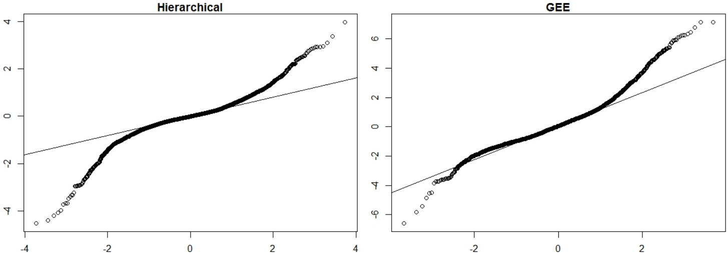


Figure 2: Observed Data

Model Validation



Model Validation



Meaningful Results and conclusions

	N.L	N.U	R.L	R.U	HM.L	HM.U
Intercept	-1.3963	-0.93568	-1.6058	-1.022	8.88	11.074
Year	0.138	0.148	0.131	0.155	0.142	0.155
prec	-0.1838	0.24683	-0.109	0.172	-0.444	0.164
pdsi	-0.26	.098	-0.003	0.075	-0.0067	0.174
w.temp	-0.020	0.162	-.020	0.0160	-.0055	.044
site	1.036	1.048	1.034	1.05		

Table 1: GEE and HM Confidence intervals

Conclusion

- Neither of these models find any of the meaningful covariates to be significant. They both find intercepts to be significant.
- Both of these models have clear downfalls. They both to overfit the data with r^2 on the order of ≈ 0.99.
- We are most likely missing key components of the data.
- Both have strong assumptions. For example
  - Hierarchical Model: Each site is independent of each other
  - GEE: There is a shared intercept
- Without modification, prediction for GEE suffers