Factors Affecting Leaf Mass Analysis

ID:6677[1]
Department of Statistics, University of California, Santa Cruz[1]

Abstract

In this report, we explored the relationship between the leaf mass per area, the distance from top of tree to where leaf sample was taken and the species of the tree. When treating the distance as a categorical variable and considering the clustering properties of the variable, the model simply became an ANOVA model and not the mean of all the groups are same. When treating it as a continuous variable and considering the species, there are significant difference among species LMA mean. In a Bayesian setting, no matter whether we consider the tree-level effects on the slope, the result are not so different from each other. And by WAIC and looic criteria, the model excluding tree-level slope effects performs better.

KEY WORDS:

ANOVA, Linear Regression, Random Effect Model, WAIC

1. Background and Data Overview

1.1 Data Set Description

The dataset contains data from leaves of two pine species (20 trees in total) that were sampled throughout their canopy. Eight samples were taken at various heights in each tree, with the objective of investigating whether or not there is a pattern of higher leaf thickness (higher LMA) toward the top of the trees.

In total, there are 160 rows and the following variables: **ID**: an ID of the individual tree, **species**: a categorical variable with two levels, Pinus ponderosa and Pinus monticola, **dfromtop**: a numerical variable corresponding to the distance from top of tree to where leaf sample was taken in meters, **height**: a numerical variable corresponding to the height from the ground where sample was taken in meters, and **LMA**: a numerical value corresponding to the leaf mass per area in g/m^2. Noticing that the data could be divided according to the tree ID, the assumptions of independent data sampling process are not satisfied, which will be discussed later in the regression part.

Table 1: Summary of Numeric Variables

	Min	0.25Q	Median	0.75Q	Max
Dftop	0.1	2.3	4.3	10.7	23.2
Height	0.9	10.1	19.7	27.1	43.0
LMA	125.1	152.4	179.2	267.7	383.3

1.2 Variable Properties and EDA

1.2.1 Variable Summary

Since the **ID** includes both the letter and the number, I will rearrange the ID from 1 to 20. Noticing that in each ID group, there are all actually eight samples, we can also deal with the variable **species**, in which 0 means "Pinus ponderosa" and 1 means "Pinus monticola".(To simplify the notation, I will use "P" and "M" instead for description). A basic summary of the variable **dfromtop**, **height** and **LMA** is shown in table 1. The species variable is a categorical one, and the height, LMA and dfromtop are numerical continuous variables. The response variable is LMA, and others are treated as covariates.

1.2.2 Exploratory Data Analysis

We are pretty interested in whether the species will affect the LMA value, which is a relationship between a categorical and continuous numerical variable. The overall LME comparison among different species are shown in figure 1. As we can see there are very significant LMA difference among different species, it seems that Ponderosa has an overall higher LMA comparing with that of Monticola. Also, the scatterplot of the other variables are also as follows in figure 2, the red and blue line are describing the estimated trend **in each group**. Also, the scatterplot also shows that the LMA have significant separation between these two species. Also, the positive correlation between the distance from the top and LMA are more significant in Ponderosa species than Monticola. Furthermore, the height is also positively correlated with LMA in both of the groups and tends to have a different slope for different species. Noticing that we did not consider the correlation between samples who belong to the same tree, the correlation is considered later.

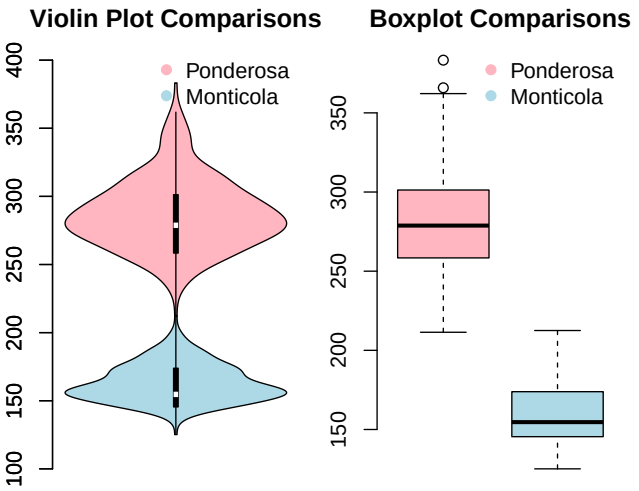

Figure 1: Comparisons Among Species

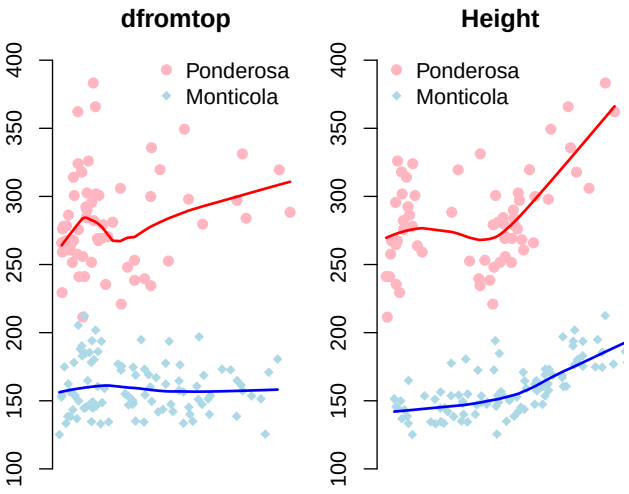

Figure 2: Scatterplot of Numeric Variables

2.Frequentist Model

2.1 Variable Transformation Model (Numerical to Categorical)

2.1.1 Variable Transformation Creteria

According to the past documentation of the variable **dfromtop**, this variable is always studied as a categorical variable. Therefore, to construct a new variable from **dfromtop**, I will create the variable **catdfromtop** in the way of adding two cut point to the data. We know the sample size is 120, one intuitive way to do the separation is using the quantile, which indicates that the first 40 are group 1, then 41-80 is group 2, the rest is group three. However, this method didn't consider the clustering properties of the data, thereofore, using histogram and checking whether the data has a clustering properties is a better choice. As shown in figure 3, a good cut point for the first category is 4, and the second is hard to decide. However, to avoid a small sample in the last group, I will choose the second cut point to be 13. Therefore, for those whose **dfromtop** values are between zero and 4, I will set the **catdfromtop** to be 0, between 4 and 14 to be 1, and the rest to be 2.

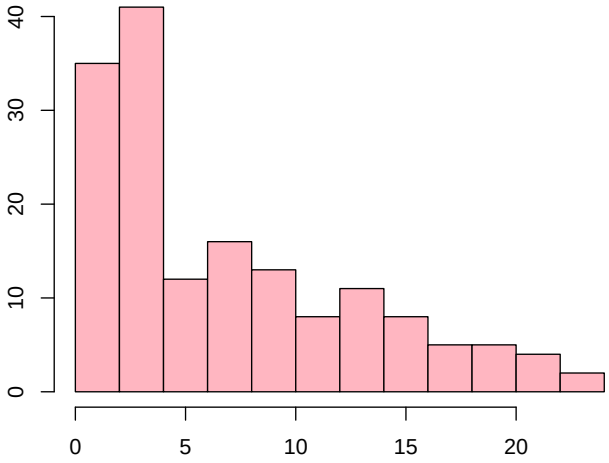

Figure 3: Histogram of dfromtop

2.1.2 ANOVA Model

After transferring the variable **dfromtop** to **catdfromtop**, one proposed model is as follows:

$$Y_{i,j,k} = \mu + \beta_1 I_{\{c_1 < x_{i,j,k} < c_2\}} + \beta_2 I_{\{c_2 < x_{i,j,k}\}} + \epsilon_{i,j,k}$$

with $\epsilon_{i,j,k} \sim^{i.i.d} N(0, \sigma^2)$ and I to be indicator function indicating whether $x_{i,j,k}$ satisfies the condition. The re-

Table 2: Regression Result

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	224.656	7.420	30.276	0.000
Beta1	-28.171	11.452	-2.460	0.015
Beta2	-33.691	14.119	-2.386	0.018

gression result is shown in table 2, from the table we can see that the β_1 and β_2 are significant. Our next goal is to compare the group-wise difference and check whether the group-wise difference is significant for all the three groups. ANOVA is a good way to achieve this goal. After using the ANOVA approach to analysis the data to check whether the group mean are same for all the groups, we got a p-value of 0.0151, smaller than 0.05, indicating that not all of the group means are same. By Tukey's HSD interval plot, we can exactly get the group-wise difference and check whether there are significant difference among the mean of different groups in figure 4. According to the plot, both the difference between the group 0 (the smallest group) and the group 1, whose distance from the top ranges from 4 to 13 and the difference between group 0 and group 2 are significant at 0.05 level, but the difference between group 1 and group 2 are not significant. Therefore, a more proper way to divide the data could be dividing them into just two groups, with the cut point equals to 4.

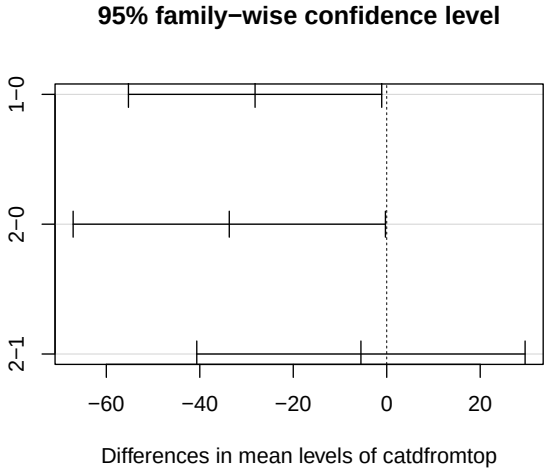

Figure 4: Tukey's HSD Interval

Finally, checking the residual assumptions are necessary for linear regression models. We strongly suspect there should be a species based difference based on our EDA part. Usually, we use the scatter plot between fitted values and the residuals as shown in figure 5, in which species type is also indicated. It seems that for different groups the residuals are not constant but not differ from each other too much. Also, it increases as the value of fitted values increase. Furthermore, the residuals have a pattern of clustering to different species. First, my change should be include species as a covariate, that changes the model to a two-way ANOVA, which exactly will be better for our data in this case. For now, I don't think our model is a good model to explain the response variable because the adjusted R^2 is only 0.0408 according to our regression result. This is really small and could not explain the factors affecting LMA well.

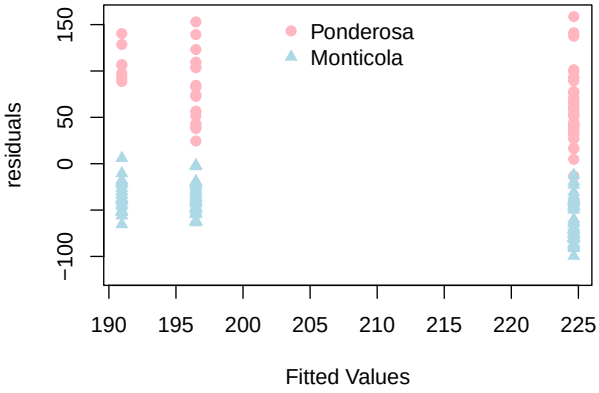

Figure 5: Residual Plot

2.2 Group Separated Linear Regression

Instead of using a ANOVA model, in this section, a linear regression model which separates the group is included. However, the design matrix should be a little different. The model can be expressed as follows:

$$y_{i,j,k} = \mu_0 I_{\{i=0\}} + \mu_1 I_{\{i=1\}} + \beta_0 I_{\{i=0\}} x_{i,j,k} + \beta_1 I_{\{i=1\}} x_{i,j,k} + \epsilon_{i,j,k}$$

and the regression results are shown in table 3. μ_{dif} here simply means the difference between μ_1 and μ_0 above, i.e., $\mu_1 - \mu_0$. From the table we can see the difference of intercept between species are significant. However, the distance from the top became not significant in both of the species. Finally, the significance of the affect of distance from the top are more significant in Ponderosa than Monticola. This model has an adjusted R^2 of 0.8351, which increases a lot compared with our previous model. The μ_0 can be expressed as the expected LMA of individuals who belongs to Ponderosa species and have zero distance from the top. $\mu_0 + \mu_{dif} = \mu_1$ could be explained as the expected LMA of individuals who belongs to Monticola species and have zero distance from the top. β_0 can be interpreted in this way: the average increment of LMA will be 1.2052 if the distance

Table 3: Regression Coefficents

	Estimate	Std. Error	t value	Pr(>|t|)
Mu0	276.078	4.689	58.873	0.000
Mu Dif	-114.247	6.583	-17.356	0.000
Beta0	1.205	0.615	1.961	0.052
Beta1	-0.258	0.467	-0.554	0.581

from top increases 1 unit for an individual belonging to Ponderosa group, with all other variables remaining the same. Finally, β_1 can be interpreted in this way: the average reduction of LMA will be 0.2585 if the distance from top increases 1 unit for an individual belonging to Monticola group, with all other variables remaining the same. However, the β's are not significant.

The assumptions of this model is obvious linearity, constant variance, the residuals are normally distributed, and the observations are independent from each other. However, in this case, data from the same tree is correlated with each other.

To further make an evaluation of the performance of this model, a residual analysis is needed. As shown in the figure 6, the residuals of the Monticola species seems to have an increasing pattern as the fitted values increase. However, the Ponderosa species tends to have decreasing variance as the increase of the fitted values. To improve this case, we can try to do data transformation, a square root transformation of LMA is considered since the log transformation is more fit for dealing with severer case of non-constant variance. The interpretation of the coefficient will be changed correspondingly to the mean of the square root of LMA instead of LMA. Also, another better way to analyze the data is to divide them into two groups based on species, then analyze them differently because the residuals tends to have clustering properties.

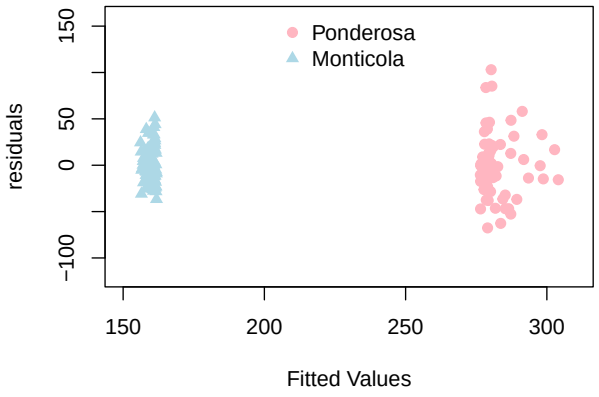

Figure 6: Residual Plot(Group Separated Model)

2.3 Model Comparisons

We have multiple creteria for model selection like AIC, adjusted R-squared and so on. The second model simply uses a group separated way to make the data analysis. However, the residuals seems to violate the assumptions of constant variance. For the first model, it is a one way ANOVA approach, that one is easier to carry out since fewer parameters are included in the model, but the strength of interpretation is so weak. The AIC for the first model is 1793.296, and 1512.506 for the second model. The BIC for the first model is 1805.597, and 1527.882 for the second model. Obviously, I will choose the second model because the AIC is smaller, BIC is smaller and adjusted R-squared is much larger. To further improve the model, another way could be using the weighted least squares or do some square root transformation of the response variable.

3.Bayesian Model

3.1 Speciese-wise Slope Model

In this model, a Bayesian approach is applied. And we are still interested in the relationship between the distance from the top and the LMA. We consider observations from a typical tree of a certain species will have the same intercept, but all the observations from the same species share the same slope, in which we ignore the tree-level effect on the slopes.

3.1.1 Model Assumption

In this Bayesian model, different trees from different species have different intercepts. However, the slope of the trees in the same species remains the same as each other. The model could be described as follows:

$$y_{i,j,k} = \mu_{ij} + \beta_i x_{i,j,k} + \epsilon_{i,j,k}$$
$$\epsilon_{i,j,k} \sim N(0, \sigma^2), \quad \mu_{ij} \sim N(\mu_{i0}, \tau^2)$$
$$\beta_i \sim N(0, \phi^2), \quad p(\mu_{i0}, \sigma^2, \tau^2) \propto \frac{1}{\sigma^2 \tau^2}$$

With this model, we can get the posterior full conditional distribution for all the parameter, which are needed for Gibbs sampler and posterior inference. First, denote n_i the number of trees in species i, and m_{ij} is the number of leaves of species i and the j^{th} individual in this species. Same as before, $i = 0$ indicates the species Ponderosa, $i = 1$ for Monticola. I set the prior for τ^2 to be $\frac{1}{\tau^2}$ since it's both non-informative and the posterior could be written in a closed form, same for σ^2.

Furthermore, I prefer to set a more flexible and weak prior, so I chose ϕ to be 10 because based on our fre-

quentist linear regression result, $|\beta|$ should be single-digit order of magnitude, so if I set ϕ to be 10, according to the rule of thumb or empirical rule, it is not a very strong prior but also can shrink β towards zero.

$$\mu_{ij}|\cdot \sim N(\frac{\frac{\sum_{k=1}^{m_{ij}}(y_{ijk}-\beta_i x_{ijk})^2}{\sigma^2}+\frac{\mu_{i0}}{\tau^2}}{\frac{m_{ij}}{\sigma^2}+\frac{1}{\tau^2}}, (\frac{m_{ij}}{\sigma^2}+\frac{1}{\tau^2})^{-1})$$

$$\mu_{i0}|\cdot \sim N(\frac{\sum_{j=1}^{n_i}\mu_{ij}}{n_i}, \frac{\tau^2}{n_i})$$

$$\beta_i|\cdot \sim N(\frac{\frac{\sum_{j=1}^{n_i}\sum_{k=1}^{m_{ij}} x_{ijk}(y_{ijk}-\mu_{ij})}{\sigma^2}}{\frac{\sum_{j=1}^{n_i}\sum_{k=1}^{m_{ij}} x_{ijk}^2}{\sigma^2}+\frac{1}{\phi^2}}, \frac{1}{\frac{\sum_{j=1}^{n_i}\sum_{k=1}^{m_{ij}} x_{ijk}^2}{\sigma^2}+\frac{1}{\phi^2}})$$

$$\sigma^2|\cdot \sim IG(\frac{N}{2}, \frac{SSR_{temp}}{2})$$

with $N=\sum_{i=0}^{1}\sum_{j=1}^{n_i} m_{ij}$ i.e., the sample size, and

$$SSR_{temp}=\sum_{i=0}^{1}\sum_{j=1}^{n_i}\sum_{k=1}^{m_{ij}}(y_{ijk}-\mu_{ij}-\beta_i x_{ijk})^2$$

$$\tau^2|\cdot \sim IG(\frac{\sum_{i=0}^{1} n_i}{2}, \frac{\sum_{i=0}^{1}\sum_{j=1}^{n_i}(\mu_{ij}-\mu_{i0})^2}{2})$$

3.1.2 Posterior Summary

The posterior MCMC trace plot are in figure 7, all chains mix well and converge to a similar distribution. And the posterior density is shown in figure 8. There are very significant difference in the mean LMA of these two species, but the slope seems similar to each other. The Ponderosa group has an overall mean LMA of 295.35, and the Monticola group has an overall mean LMA of 174.93. And the posterior mean slope of distance from top for species Ponderosa is -2.41, and -1.89 for Monticola.

3.1.3 Posterior Predictive Distribution

First, I generated the "Tree", like the $\mu_{0,new}$ because this new leave does not belong to any existing tree, after generating this new tree, then $LMA_{new} \sim N(\mu_{0,new}, \sigma^2)$. Using our posterior samples of the overall mean μ_{i0},β and σ^2, we can get a posterior predictive sample for each of the trees.

The predictive distribution for two samples from Ponderosa group with distance from top being 2.330 and 10.742 are shown in the figure 9, and the 95% credible interval is shown in figure 10.

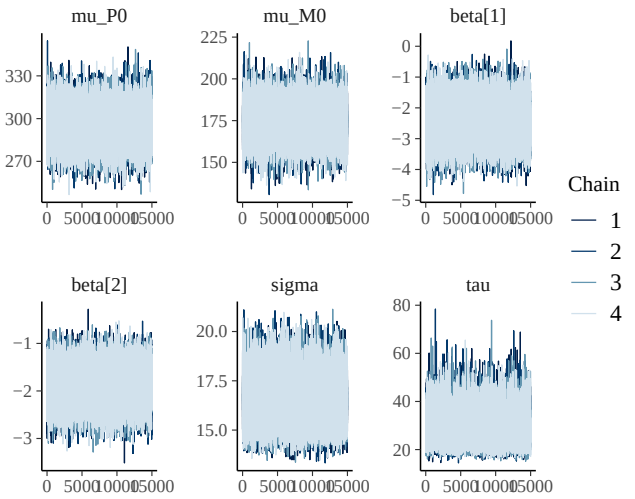

Figure 7: Posterior Mixing Performance

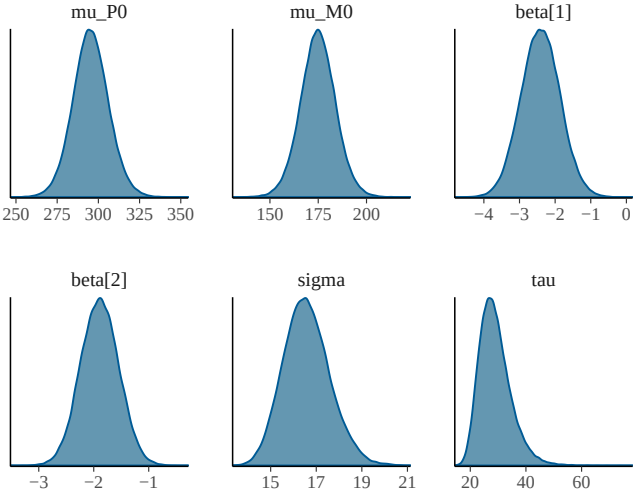

Figure 8: Posterior Density

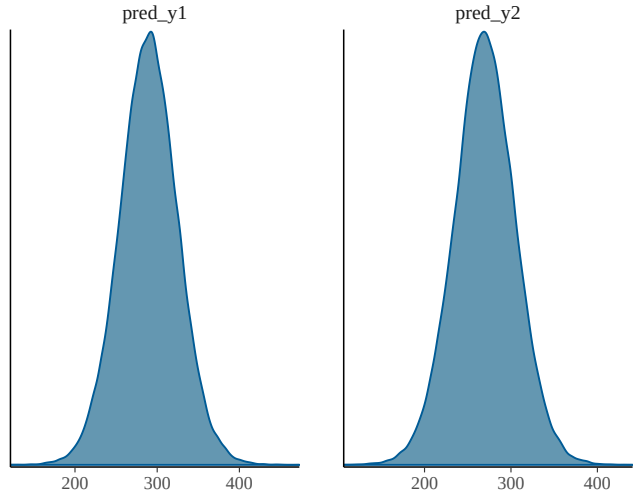

Figure 9: Posterior Predictive Distribution

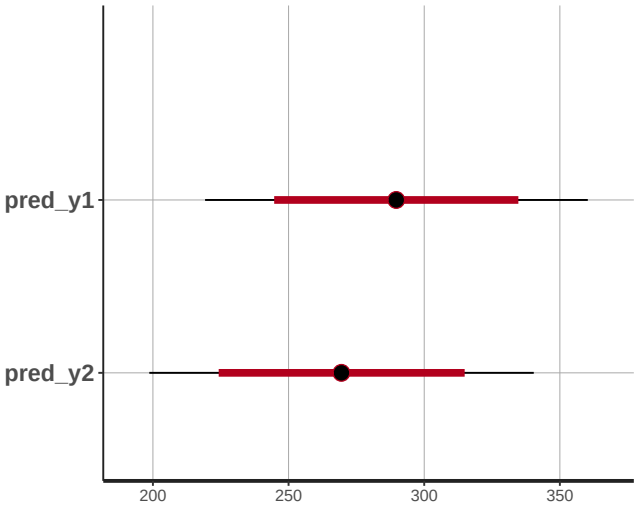

Figure 10: Posterior Credible Interval

3.2 Speciese-wise and Tree-level Slope Model

Based on the model mentioned in 3.1, we introduce the tree-level effect on slope to the model, indicating that observations from different individual trees of different species share the same slope and intercept. As long as the observations are from different individuals or different species, the slope and the intercept will both be different from each other.

3.2.1 Model Assumption

In this model, we not only consider the tree-level intercept, but also the tree-level slopes for all the trees. Therefore, I consider for each species, the slope shares the same mean, which means in each species group, the β_{ij} are iid normal distributed with mean β_{i0} for $i = 0, 1$. Also, we should put a prior distribution on ϕ since it is the variance parameter of the β_{i0}. Similar to the previous model, I put an non-informative prior on it. Therefore, the model can be express as follows:

$$y_{i,j,k} = \mu_{ij} + \beta_{ij} x_{i,j,k} + \epsilon_{i,j,k}$$

$$\epsilon_{i,j,k} \sim N(0, \sigma^2), \quad \mu_{ij} \sim N(\mu_{i0}, \tau^2)$$

$$\beta_{ij} \sim N(\beta_{i0}, \phi^2), \quad p(\mu_{i0}, \beta_{i0}, \sigma^2, \tau^2, \phi^2) \propto \frac{1}{\sigma^2 \tau^2 \phi^2}$$

3.2.2 Posterior Summary

I will simply plot the trace plot and posterior density for $\mu_{i0}, \beta_{i0}, \tau^2, \sigma^2, \phi^2$ in figure 11 and figure 12. The model mixed well because the Rhat of each variables are 1. And the posterior density still shows that the overall mean intercept of Ponderosa group still higher than the Monticola group. A big difference is that the slope of each individual tree are not so close to each other according to our posterior summary. But the overall mean slope of Ponderosa group is around -2.7, which differed from the Monticola group, -1.95.

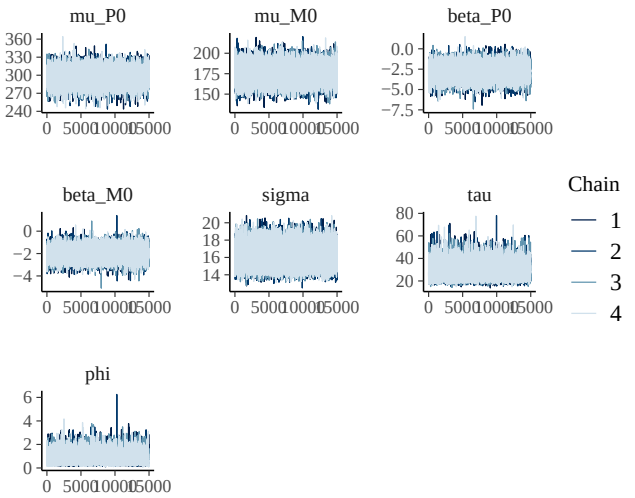

Figure 11: Trace Plot (Tree-level Slope)

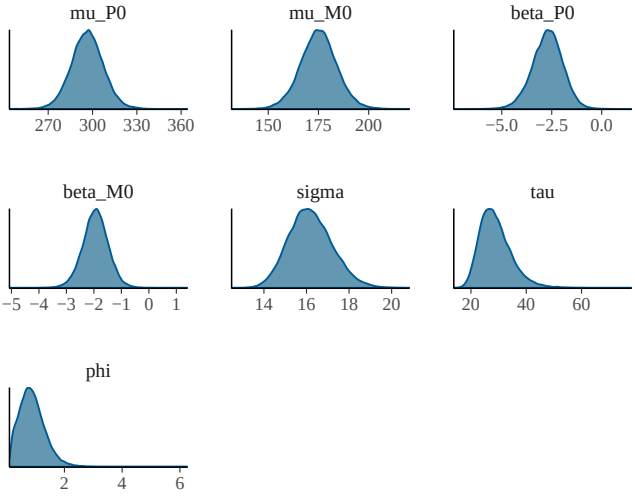

Figure 12: Posterior Density (Tree-level Slope)

3.3 Model Comparisons

3.3.1 Predictive Performance

One important criteria to judge the goodness of fit for a model is how well it can predict the samples. In this section, I simply use the posterior mean which is the bayes decision estimator under the quadratic loss function to make the predictive distribution. And then plot the true LMA value together with the posterior predictive mean

and check how similar they are. The model excluding tree level slopes effect predicts the LMA in figure 13. And the model including the tree level slopes effect prediction is shown in figure 14. They are pretty similar to each other. And the predictive sum squares of the first model is 3.7422339×10^4, and 3.4439193×10^4 for the second model. Therefore, the second model actually behaves better than the one that excludes the tree-level slope effects. However, the second model includes more parameters than the first one, which makes it more flexible obviously. Therefore, more numerical ways to carry out model comparisons are needed.

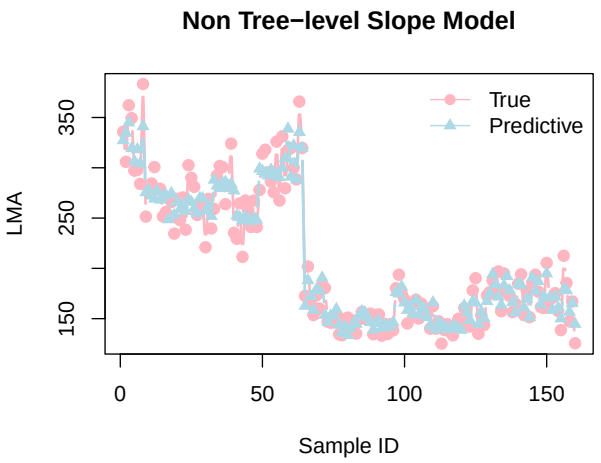

Figure 13: Prediction Excluding Tree-level Slope

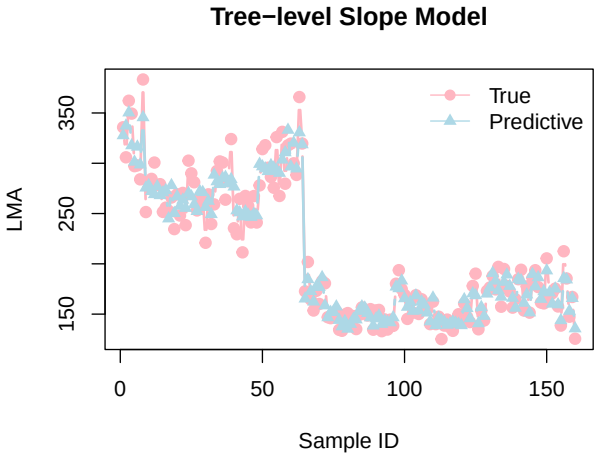

Figure 14: Prediction Including Tree-level Slope

3.3.2 WAIC and looic

The WAIC for the first model is 69590, and 80996 for the second model, which indicates the first model is better. Furthermore, I also used the looic, and it is 41719 for the first model and 45219 for the tree-level slopes included model. Furthermore, although we includes more parameters in our model, the results shows that we still prefer the first one. We can also see the distribution of residuals, i.e, σ^2. They both center around 16, therefore, introducing extra β's will not make the residuals reduce significantly. Therefore, due to the penalty of parameters, we prefer the first model, which does not include the tree-level slope effect.

4. Results

As shown in the Bayesian regression model which excludes the the tree-level effect, the effect of distance from top on LMA is more significant in species ponderosa than monticola. And also, the posterior mean effect are similar to each other. Furthermore, the effect of the distance from top is both negative for two species. Also, the overall intercept mean μ_{00} for spices ponderosa is much bigger than that of monticola. Furthermore, if we transfer the distance from the top from a numerical variable to an categorical by setting up thresholds, there are significant difference in mean among all the groups if we set the cut point to be 4 and 13. In our case, we did not see significant difference between the highest distance group and the second highest one, but we see both of these two groups have a significant difference of there means between the lowsest distance group. Also, as both shown in the results of Bayesian model and frequentist model, an important property of this dataset is that we should consider the interaction between species and distance from the top. If we don't consider the species as a covarate, both the adjusted R^2 and AIC will prefer the model that includes the species and the interaction between species and distance from the top.

5. Discussion

In this model, we didn't include the height as a explanatory variable. From the EDA part, we can see that the height is a more significant factor that affects the LMA. Also, the slope for these two species are not the same, if in the future I have the chance going back to this data, I will add the height as a covariate and make sure difference species have different slopes. Since we have checked that the model including tree-level slope effect actually performed not so good, a new probable model can be with this form:

$$y_{i,j,k} = \mu_{ij} + \beta_i x_{i,j,k} + \gamma_i z_{i,j,k} + \epsilon_{i,j,k}$$

$$\epsilon_{i,j,k} \sim N(0, \sigma^2), \quad \mu_{ij} \sim N(\mu_{i0}, \tau^2)$$

$$\beta_i \sim N(0, \phi^2), \quad p(\mu_{i0}, \sigma^2, \tau^2) \propto \frac{1}{\sigma^2 \tau^2}$$

$$\gamma_i \sim N(0, \nu^2)$$

in which, $z_{i,j,k}$ is the height variable of the sample from species i, the tree j and the k^{th} record. ν and ϕ are still hyperparameters, and i will not put a prior on them. This will propose a new model including the effects of the height. Noticing that the height is a measure of the distance from the ground, and seems to have a significant effect on the LMA based on the EDA part, and also the sum of the distance from the top and distance from the ground is actually the height of the tree, cases could also be that difference species have different overall height. Therefore, the distance from the ground could be correlated to the species, which is another interesting result if verified.