Plant Pals (Group 4)
Dr. Amy T. Kim
STA 101
June 3, 2024

STA 101 Final Project

Paper: Liu M, Feng Z, Zhang Z, Ma C, Wang M, et al. (2017) Development and evaluation of height diameter at breast models for native Chinese Metasequoia. PLOS ONE 12(8): e0182170. https://doi.org/10.1371/journal.pone.0182170

Introduction

The paper we selected analyzes the relationship between tree height and diameter at breast height (referred to as dbh) in Chinese Metasequoia trees. The Metasequoia plant is very old, dating back to the Mesozoic Cretaceous period, and is currently considered an endangered species in China. This study aimed to study the allometry equation of the trees, by using height and dbh data of 5503 total trees. 53 models were considered in the study in order to select the best fit, wherein 7 were linear and 46 were non-linear. Of those models, 37 only included height and diameter (dbh), while the other 16 had more variables. These other variables included: basal area in cm^2 (BA), meters above sea level (ASL), age of the stand in years (T), dominant height of the stand in meters (H_0), and dominant dbh of the stand in centimeters (D_0).

Dataset

The dataset provided for us included the height and dbh of 500 trees selected from the full sample used in the paper. It was not communicated on how these 500 observations were selected. The data also did not include information about the age of the trees or any other variable. Because of this, we were unable to replicate models that included any explanatory variable outside of diameter at breast height.

Before we began to fit models, we first wanted to compare our data to the larger dataset so we could understand any differences in the fit of our models and the models found in the paper using the large dataset. The data used in the paper was separated into two categories: Fitting Data and Validation Data. When comparing the data, we found the following information[1]:

> Mean Height: Fitting Data = 27.61 m, Validation Data = 27.73 m, Our Data = 26.69130 m
> Max Height: Fitting Data = 46.41 m, Validation Data = 40.09 m, Our Data = 45.62 m
> Min Height: Fitting Data = 16.69 m, Validation Data = 18.8 m, Our Data = 17.35 m
> Standard Deviation of Height: Fitting Data = 3.78 m, Validation Data = 3.55 m, Our Data = 4.44614 m
> Mean Diameter: Fitting Data = 57.03 cm, Validation Data = 57.3 cm, Our Data = 53.36036 cm
> Max Diameter: Fitting Data = 134.68 cm, Validation Data = 109.8 cm, Our Data = 134.68 cm
> Min Diameter: Fitting Data = 26.35 cm, Validation Data = 28.47 cm, Our Data = 26.35 cm

[1] Table 1: Comparison of Paper's Sample Statistics to Our Sample Statistics

Standard Deviation Diameter: Fitting Data = 12.43 cm,Validation Data = 11.75 cm,Our Data=13.33801 cm

The mean height and diameter found in our data are slightly lower than the averages found in the larger dataset, and the standard deviation in our data is slightly higher in both as well. While it was not mentioned how the observations in the sample dataset were selected, we assumed that since our data and their data were very similar, that it was a simple random sample (SRS).

Replication Process

Firstly, we attempted to replicate all of the single variable linear models proposed in the paper, in order to determine if we would find the same model from these choices to be the model of best fit.

Model Replication

There were 5 proposed single variable linear models in our paper (Models 1 through 5)[2]:

$$h = \beta_0 + \beta_1 d$$
$$h = \beta_0 + \beta_1 log(d)$$
$$h = \beta_0 + \beta_1 d + \beta_2 d^2$$
$$h = \beta_0 + \beta_1 d^2 + \beta_2 d^3$$
$$h = \beta_0 + \beta_1 d^{-1} + \beta_2 d^2$$

Our replication using R produced the following models:

Model 1: $h = 10.1942 + 0.3092d$

Model 2: $h = -39.31 + 16.72log(d)$

Model 3: $h = 7.4610696 + 0.4066729d - 0.0008166d^2$

Model 4: $h = 15.75 + 0.005308d^2 - 0.00002802d^3$

Model 5: $h = 30.73 - 411.7d^{-1} - 0.001373d^2$

Model Selection

After obtaining these models, we used R to run the series of tests used in the paper on our models to compare them[3]. The package used was ModelMetrics. First, we found the root mean squared error (RMSE) of each model:

Model 1: 1.66083

Model 2: 1.739887

Model 3: 1.635669

[2] Table 2: Replicate Models
[3] Table 3: Model Selection Values

Model 4: 1.631005

Model 5: 1.70729

Next, we found the R^2_{adj} :

Model 1: 0.8599043

Model 2: 0.8462496

Model 3: 0.8638436

Model 4: 0.8646189

Model 5: 0.8516588

Next, we found the AIC:

Model 1: 1932.256

Model 2: 1978.759

Model 3: 1918.99

Model 4: 1916.135

Model 5: 1961.846

Lastly, we calculated the bias of each model since this was used in the paper to judge their models. We found extremely small values for each model, however, and concluded that since each model had a bias of close to zero, the bias values were not significant.

After computing all of these values, we determined that Model 4 was the best selection from the 5 models, as it had the lowest RMSE, the highest R^2_{adj}, and the lowest AIC.

Additionally, we created a Residual Plot[4], a Q-Q Plot[5], and an Observed vs Predicted Plot[6] for each model in order to visualize the fit of each model. While there does not appear to be major noticeable differences across the plots, the Q-Q Plot of Model 4 appears to be one of the better fit Q-Q Plots, and the residual plots seem to show that Model 4 is one of the models with the most constant variance.

Comparison to Paper:

Our findings aligned with the findings of the paper, which had also concluded that Model 4 was

[4] Figure 1: All Residual Plots
[5] Figure 2: All Q-Q Plots
[6] Figure 3: All Observed vs Predicted PlotsFigure 4: QQ-Plots of 3 Best Models

the best fit out of the 5 models. In their findings, Model 4 had the lowest RMSE = 1.8277, and the highest $R^2_{adj} = 0.7583$.

Exploratory Analysis

Hypothetical Model

The original article has two points that can be criticized: 1. It doesn't contain all the data in the dataset, but it is still understandable. 2. The model fit part doesn't explain how the analysts conclude these models are appropriate for fitting. We aim to find our best model based on multiple linear regression for the exploratory analysis; however, the incomplete dataset might cause a problem.

First, we need to transform the variables in the five linear regression models we replicated into separate variables. Consequently, we created four extra columns in the dataset including d^2, d^3, log(d), and d^{-1}. We can use X's to substitute them.

Thus, we come up with a multiple linear regression model consisting of all the transformed X's mentioned in the current five models:

$$h = \beta_0 + \beta_1 d + \beta_2 d^2 + \beta_3 d^3 + \beta_4 log(d) + \beta_5 d^{-1}$$

We set this model as the full model, then perform stepwise model selection to select the best model from incomplete data, and compare with the current best model-model 4 in the model replication part.

Model Selection

Overall, we performed model selection using forward model, backward model, forward-backward model, and backward-forward model to examine both AIC and BIC. This was done using R packages, ModelMetrics and MASS.

We found that the backward AIC model stated below has the smallest AIC = 1912.894, which is even smaller than that of model 4's AIC = 1916.135.

$$h = \beta_0 + \beta_1 d + \beta_2 d^2 + \beta_4 log(d)$$

The forward BIC model stated below has the smallest BIC = 1932.985, which is also smaller than that of model 4's BIC = 1932.994.

$$h = \beta_0 + \beta_1 d + \beta_3 d^3$$

These results prove that model 4 may not be the best model under the given data. Therefore, we decided to test and record the statistics using the backward AIC model as Model A, the forward BIC model as Model B, and model 4 as Model C. However, before we proceed to the next step, we operated test statistics for linear assumptions to support

Linear Assumptions

1. **Normally distributed data**
   To assess the normality of the data, we use the Shapiro-Wilks Tests. We found that p-values of all 3 models are too small, showing that the data is not normally distributed. Thus, we need to remove outliers.
2. **Constance variance of error**
   To assess the homoscedasticity, we use the Fligner-Killeen Tests. The p-values are all over common significant levels, showing that errors have constant variances.
3. **Linearity**
   To access the linearity test, we find that the Q-Q plots[7] of each model. The pattern shows the linearity of each model.
4. **Random Sample**
   We don't have enough information about whether the selected values are randomly selected; however, we can assume that a random sample was taken.

Identifying Outliers

We use standardized residuals to set 2 as the cutoff to identify any potential outliers that are influential to the model. The three models we have included in this process are the best model based on 1) AIC, 2) BIC, and 3) the article's claim. Any datapoint that has a standard residual that is greater than +2 or less than -2 will be removed from the model. After removing potential outliers, we performed SW test again and found out this time p-values are big enough. After removing outliers, we have 477 observations for Model A, 474 for Model B, and 477 for Model C.

Conclusion

We compare the AIC, BIC, RMSE, and adjusted R^2 of each model, and make a table to show which one is the best model.

Model	AIC	BIC	RMSE	R^2_{adj}
Model A	1629.31	1650.148	1.321083	0.9077727
Model B	1608.176	1624.821	1.30867	0.9093569

[7]Figure 4: QQ-Plots of 3 Best Models

Model C	1632.427	1649.098	1.328188	0.9032784

Since Model B has the smallest AIC, BIC, and RMSE; plus, it has the biggest adjusted R^2, it is the best model.

Further Analysis

The full equation of Model B is $h = 8.659 + 0.3496\ d\ -3.763x10^{-6}\ d^3$. The 95% confidence interval for (0.3299, 0.3693) for β_1 and ($-5.276*10^{-6}$, $-2.249*10^{-6}$) for β_2.
Since the CIs don't contain 0, the explanatory variables have a significant influence in the model and should not be dropped.

Findings

Replication Result

(Part 1) Our replication of the original paper resulted in the same best model, Model 4. Even though we had different dataset sizes, which caused different RMSE, AIC, and R^2_{adj} values, the same trend was found between the paper and our data.

Hypothetical Model

(Part 2) For further analysis, we created new models to compare to the models from the paper. This resulted in the best model being Model B, with diameter and diameter cubed as the variables.

Model Selection

In the first part, we did not create any new models. Models were replicated from the original paper based on what variables and data we had. This resulted in five models to replicate, Models 1-5. The original paper does not elaborate on how they chose/fitted the models.

In the second part, we used more tests to create new models. Using the R package, MASS, we used forward selection, backward selection, forward backward selection, and backward forward selection for both AIC and BIC to create the best model. The best two models with the smallest AICs or BICs were selected and renamed Model A and Model B, while we kept Model 4 from the first part and renamed it Model C.

Identifying Outliers

Any observation that had a standard residual greater than 2 or less than -2 was removed from the data. After removing potential outliers, we performed SW test again and found out this time p-values were greater than $\alpha = 0.05$, so there were no more outliers.

Linear Assumptions

1. **Normally distributed data**
   In all of the models we ran, we used Q-Q plots to visually determine the normality of the dataset. All models had some variance, but generally satisfied the normality assumption of linear regression.
   We also used the Shapiro Wilks test to determine the normality of the dataset. We used an $\alpha = 0.05$ to determine if the dataset was normal.
2. **Constance variance of error**
   For constant variance, we used residual plots to visually determine the variance of the dataset. This included finding a model that had no visible patterns in the plot.
3. **Linearity**
   For linearity, we used the combined results of the three tests above to come to a conclusion. If both normality and constancy were satisfied, then a linear model is appropriate.
   We also followed the assumption that linear models were appropriate, since the paper labeled Models 1-5 as linear.
4. **Random Sample**
   We don't have enough information about whether the selected values are randomly selected; however, we can assume that a random sample was taken.

Comparison to Paper Findings

In the first part (Replication Process), we found that of the models we performed, the paper concluded the best model was Model 4 using RMSE, AIC, R^2_{adj}, and Bias. While we had a smaller dataset, we used the same test (See Table 3) and also found that Model 4 was the best model at predicting the height based on diameter at breast height.

In the second part (Exploratory Analysis), we created our own models to compare to Model 4 from the paper. This time, we took all the variables from Models 1 through 5 to create a very large model with diameter, log transformed diameter, diameter squared, diameter cubed, and inverse diameter. After running several selection techniques, we resulted in a model that was better than Model 4 that was not the same as any of the other models from the paper. This was Model B which had diameter and diameter cubed as its explanatory variables. It had the lowest AIC, BIC, RMSE and highest R^2_{adj} values.

Interpretation

Interpretation of Model 4 Betas

β_0: We cannot interpret the intercept of this model, as a tree cannot have a dbh of 0.

β_1: As the $(dbh)^2$ of a Chinese Metasequoia tree increases by 1 cm, we expect the height of the tree to increase by 0.005308 meters on average.

β_2: As the $(dbh)^3$ of a Chinese Metasequoia tree increases by 1 cm, we expect the height of the tree to decrease by 0.00002802 meters on average.

Interpretation of Model 4 95% Confidence Intervals

β_0: We cannot interpret the confidence of β_0, as a tree cannot have a dbh of 0.

β_1: We are 95% confident that as the $(dbh)^2$ of a Chinese Metasequoia tree increases by 1 cm^2 the height of the tree will increase by between 0.004946444 meters and 0.005670146 meters.

β_2: We are 95% confident that as the $(dbh)^3$ of a Chinese Metasequoia tree increases by 1 cm^3 the height of the tree will decrease by between 0.00003143211 meters and 0.00002460142 meters.

Interpretation of Model B Betas

β_0: We cannot interpret the intercept of this model, as a tree cannot have a dbh of 0.

β_1: As the dbh of a Chinese Metasequoia tree increases by 1 cm, we expect the height of the tree to increase by 0.3496 meters on average.

β_2: As the $(dbh)^3$ of a Chinese Metasequoia tree increases by 1 cm, we expect the height of the tree to decrease by 0.000003763 meters on average.

Interpretation of Model B 95% Confidence Intervals

β_0: We cannot interpret the confidence of β_0, as a tree cannot have a dbh of 0.

β_1: We are 95% confident that as the dbh of a Chinese Metasequoia tree increases by 1 cm, the height of the tree will increase by between 0.3299 meters and 0.3693 meters.

β_2: We are 95% confident that as the $(dbh)^3$ of a Chinese Metasequoia tree increases by 1 cm^3 the height of the tree will decrease by between $-5.276*10^{-6}$ meters and $-2.249*10^{-6}$ meters.

Conclusion

Summary

We were able to replicate the 5 single variable linear models analyzed in the paper, and found the same model as the paper found to be the best fit model for our data. This model was Model 4: $h = 15.75 + 0.005308d^2 - 0.00002802d^3$. Although Model 4 was the best fit out of the 5 proposed models, we found a different model, Model B, to be a better fit for our data. We found Model B using Forward BIC selection, using every transformation of dbh used in Models 1-5 as a potential variable. Due to the limitation of our incomplete dataset, Model B might not be the best model under the full dataset, so we can only declare it the best fit model for our smaller dataset.

Challenges

We found our greatest challenge to be our limited dataset, especially having only one explanatory variable. This greatly limited the number of models we were able to create and test.

Appendix

[1]Table 1: Comparison of Paper's Sample Statistics to Our Sample Statistics

	Paper's Data (Fitting Data) n = 4401	**Paper's Data (Validation Data) n = 1102**	**Our Data n = 500**
Mean Height (m)	27.61	27.73	26.69130
Max Height	46.41	40.09	45.62
Min Height	16.69	18.8	17.35
SD Height	3.78	3.55	4.44614
Mean Diameter (cm)	57.03	57.3	53.36036
Max Diameter	134.68	109.8	134.68
Min Diameter	26.35	28.47	26.35
SD Diameter	12.43	11.75	13.33801

[2]Table 2: Replicated Models

	General Model	**Our Model**
Model 1	$h = \beta_0 + \beta_1 d$	$h = 10.1942 + 0.3092d$
Model 2	$h = \beta_0 + \beta_1 log(d)$	$h = -39.31 + 16.72log(d)$
Model 3	$h = \beta_0 + \beta_1 d + \beta_2 d^2$	$h = 7.4610696 + 0.4066729d - 0.0008166d^2$
Model 4	$h = \beta_0 + \beta_1 d^2 + \beta_2 d^3$	$h = 15.75 + 0.005308d^2 - 0.00002802d^3$

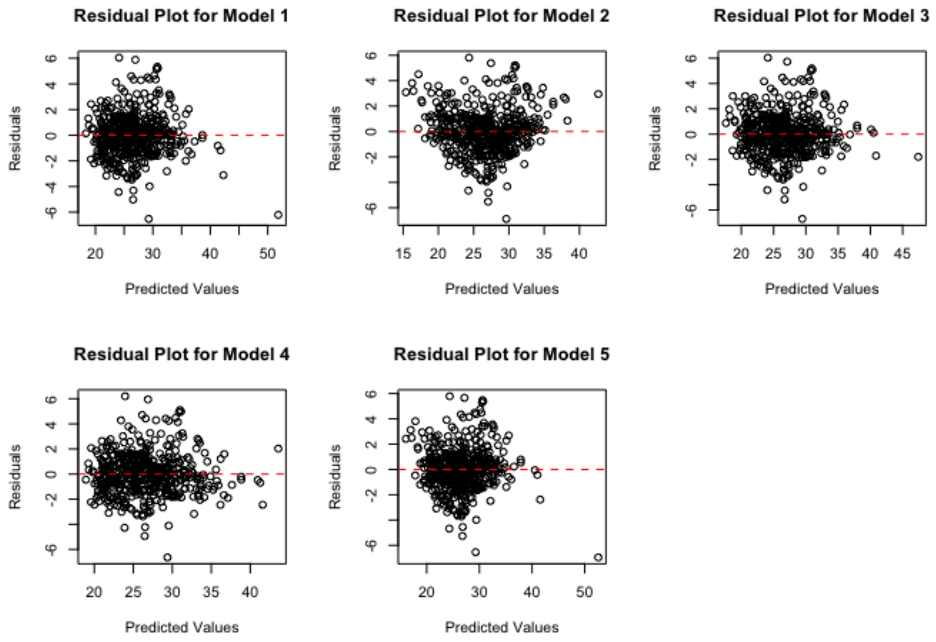
Model 5	$h = \beta_0 + \beta_1 d^{-1} + \beta_2 d^2$	$h = 30.73 - 411.7d^{-1} - 0.001373d^2$

[3]Table 3: Model Selection Values

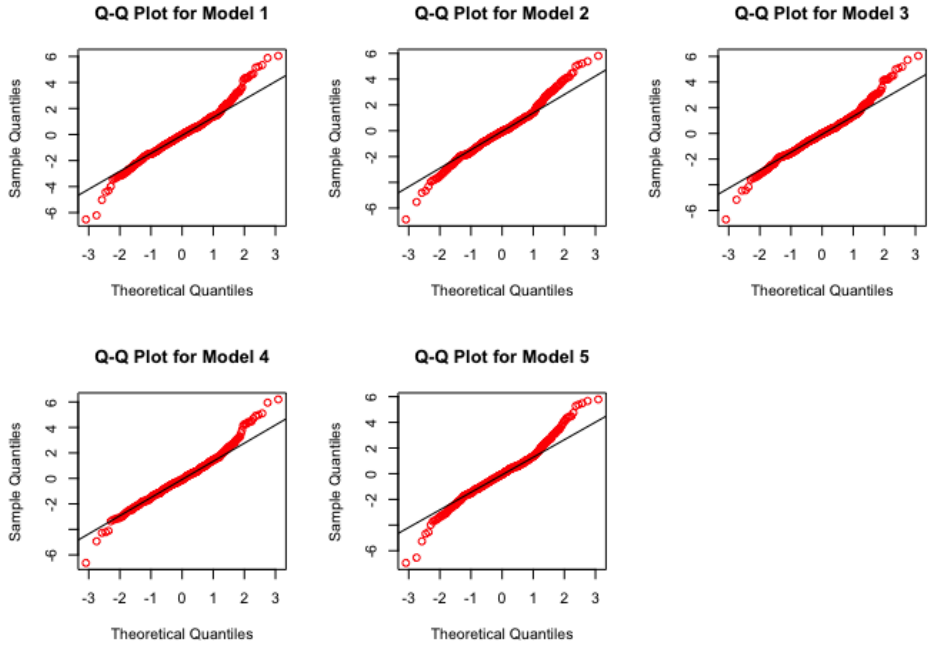
	RMSE	**R^2_{adj}**	**AIC**	**Bias**
Model 1	1.66083	0.8599043	1932.256	0.0000
Model 2	1.739887	0.8462496	1978.759	0.0000
Model 3	1.635669	0.8638436	1918.99	0.0000
Model 4*	1.631005	0.8646189	1916.135	0.0000
Model 5	1.70729	0.8516588	1961.846	0.0000

*Red text indicates best values that result in which model we picked was the best

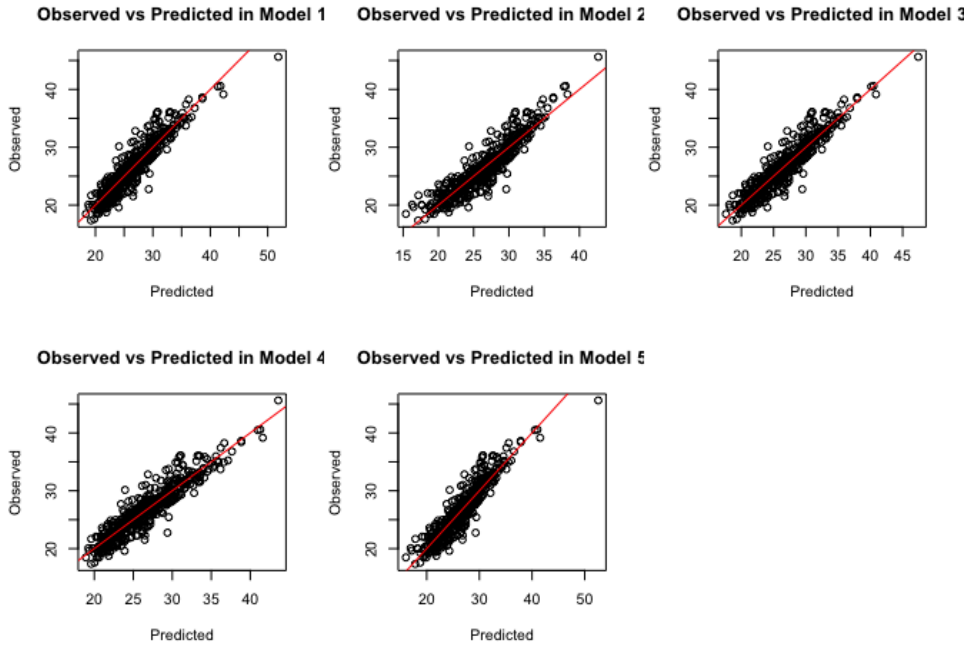
[4]Figure 1: All Residual Plots



[5]Figure 2: All QQ-Plots



[6]Figure 3: All Observed vs Predicted Plots



[7]Figure 4: QQ-Plots of 3 Best Models