STA 101 Final Project

Group 4 - Plant Pals

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

Development and evaluation of height diameter at breast models for native Chinese Metasequoia

About the Paper

- Trying to predict the growth of Metasequoia trees through the relationship between height and diameter
- Examination of several different models and variables to see which model or variable has the most predictive power regarding growth
  - 53 total models: 7 linear and 46 non-linear
  - 2 model groups: group 1 has one variable (dbh) group 2 had multiple variables
- Reason: measuring height of individuals trees can be tricky and expensive

Variables

- **h**: height of the tree in meters
- **dbh**: diameter at breast height in centimeters
- **BA**: basal area in cm^2
- **ASL**: above sea level in meters
- **T**: age of the stand in years
- **H_0**: dominant height of the stand in meters
- **D_0**: dominant dbh of the stand in centimeters

Data Observation

Fitting data - used to make and fit models
Validation data - used to check predictive power of models

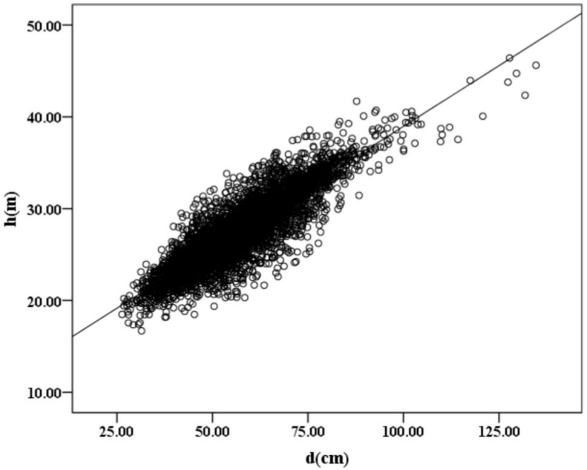


Table 1. Regional Metasequoia sample statistics.

		h(m)	d(cm)	BA(cm^2)	ASL(m)	T(y)	H_0(m)	D_0(cm)
Fitting data	Mean	27.61	57.03	2675.36	1187.16	95	43.04	122.59
N = 4401	Max	46.41	134.68	14246.35	1590	485	46.41	134.68
	Min	16.69	26.35	545.21	750	50	40.5	110.14
	Standard deviation	3.78	12.43	1221.36	111.38	32.58	3.26	8.79
Validation data	Mean	27.73	57.3	2686.72	1185.33	95.92	38.11	95.68
N = 1102	Max	40.09	109.8	9468.78	1605	325	40.09	109.8
	Min	18.8	28.47	636.69	856	50	37.12	95.68
	Standard deviation	3.55	11.75	1124.55	11.93	32.91	0.89	6.39

h: height; d: diameter at breast height (dbh); BA: basal area; ASL: Above Sea Level; T: age of the stand; H_0: dominant height of the stand, m; D_0: dominant dbh of the stand, cm; N: number of trees. doi: 10.6084/m9.figshare.4956284.t001

Data Comparison

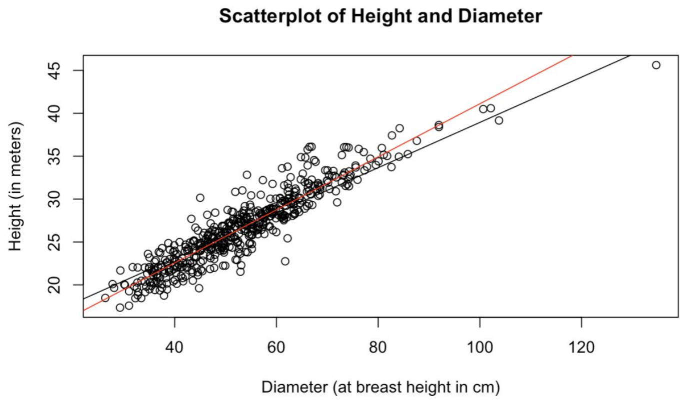
Paper's Data

- 5746 observations (reduced to 5503)
  - divided into fitting data (4401 observations) and validation data (1102 observations)
- Many variables:
  - height (h)
  - diameter (dbh)
  - basal area (BA)
  - above sea level (ASL)
  - age of the stand (T)
  - dominant height of the stand (H_0)
  - dominant dbh of the stand (D_0)

Our Data

- 500 observations
- Two variables:
  - height
  - diameter

Data Comparison



Red Line: Trendline for Our Data (Model 1)
Black Line: Trendline for Paper's Data

	Paper's Data (Fitting Data) n = 4401	**Paper's Data (Validation Data) n = 1102**	**Our Data n = 500**
Mean Height	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	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

Replication Process

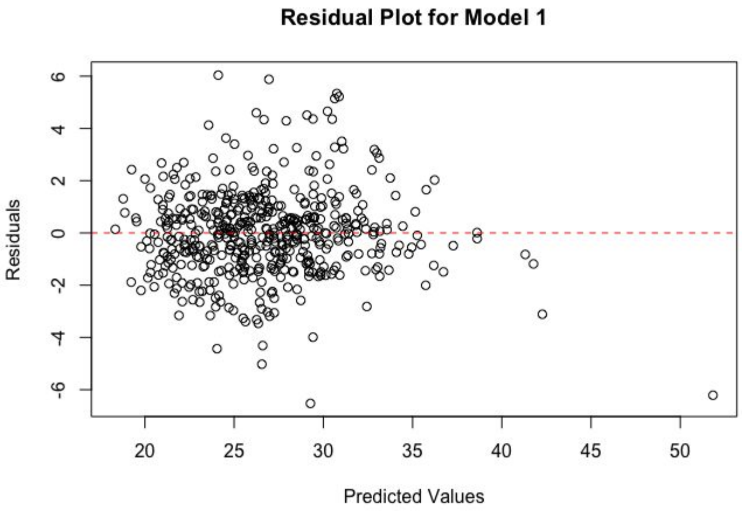
Testing Group 1 Linear Models:

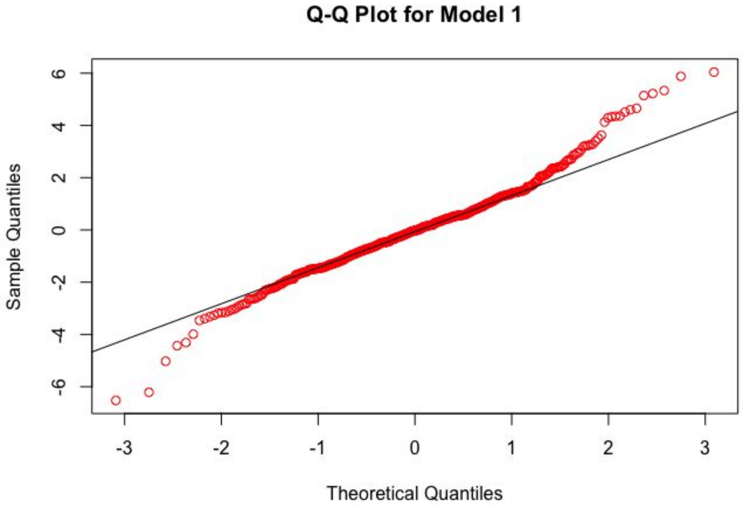
- Group 1: single variable models

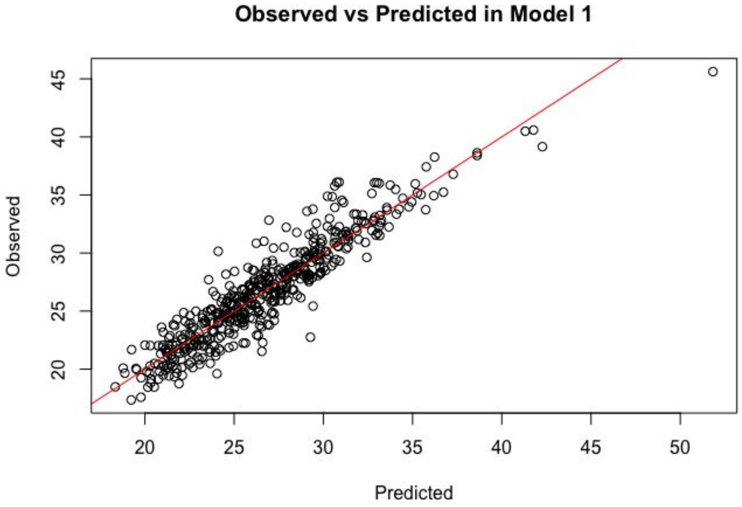
	General Model	Our Model
Model 1	$h = \beta_0 + \beta_1 d$	$h = 10.1942 + 0.3092d$
Model 2	$h = \beta_0 + \beta_1 \log(d)$	$h = -39.31 + 16.72\log(d)$
Model 3	$h = \beta_0 + \beta_1 d + \beta_2 d^2$	$h = 7.4610696 + 0.4066729d - 0.0008166d^2$
Model 4	$h = \beta_0 + \beta_1 d^2 + \beta_2 d^3$	$h = 15.75 + 0.005308d^2 - 0.00002802d^3$
Model 5	$h = \beta_0 + \beta_1 d^{-1} + \beta_2 d^2$	$h = 30.73 - 411.7d^{-1} - 0.001373d^2$

Model 1

$$Y = 10.1942 + 0.3092x$$

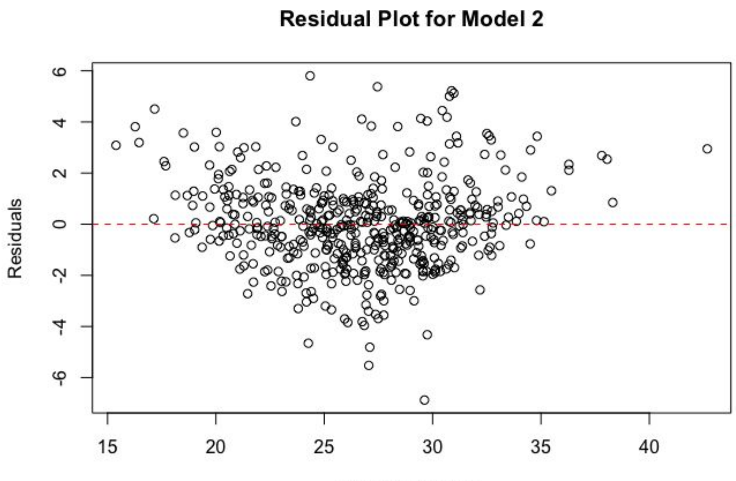


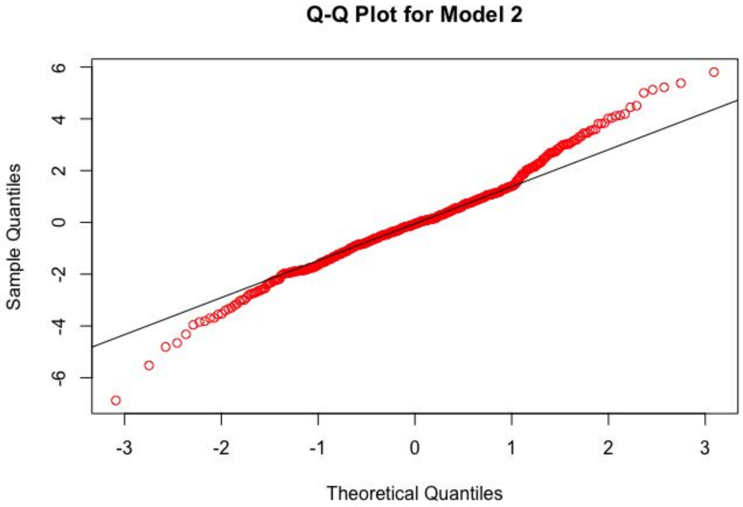


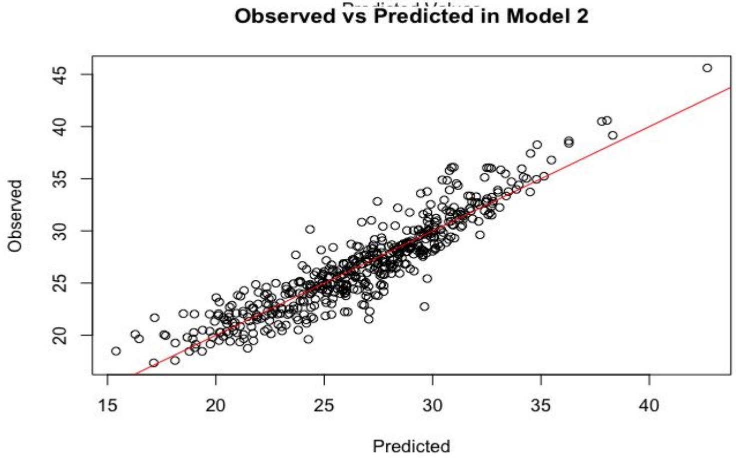


Model 2

$$Y = -39.31 + 16.72log(x)$$

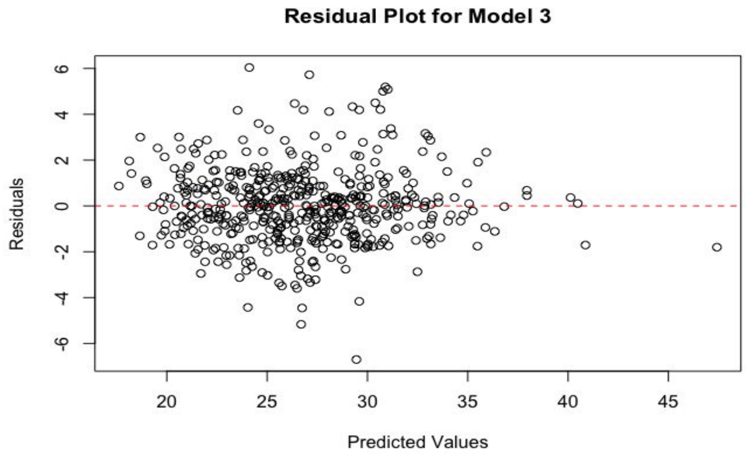


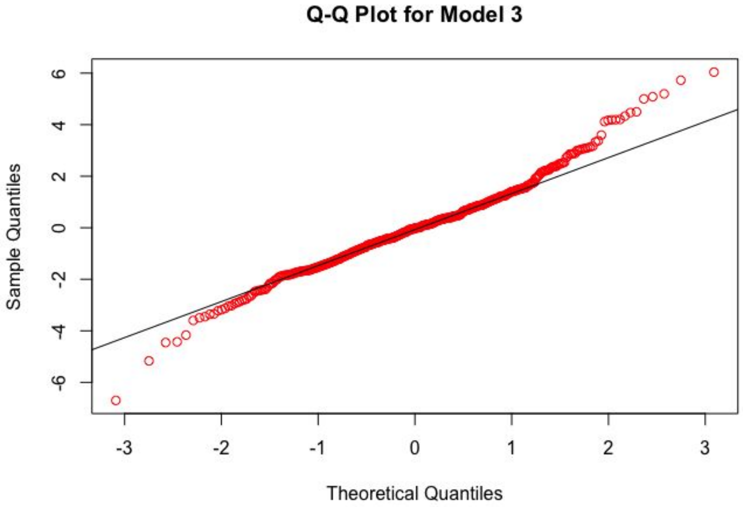


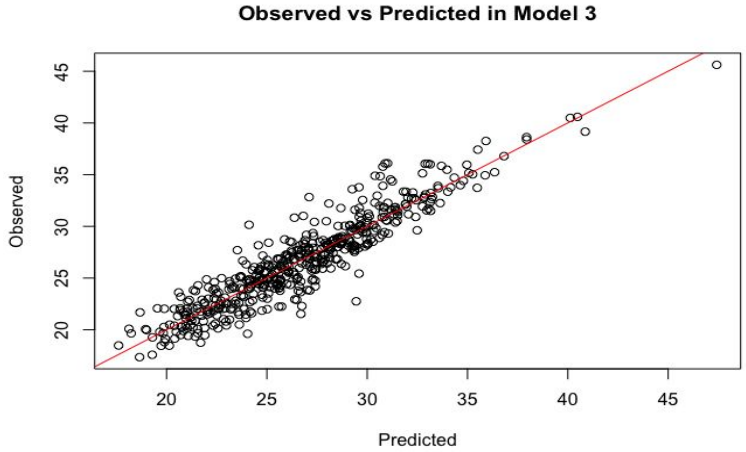


Model 3

$$Y = 7.4610696 + 0.4066729x - 0.0008166x^2$$

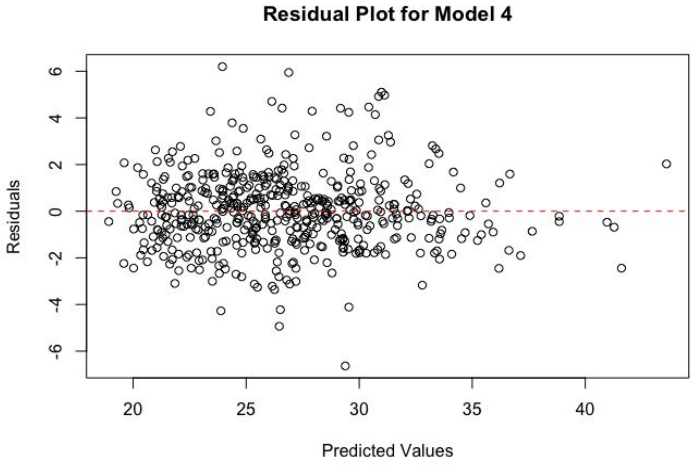


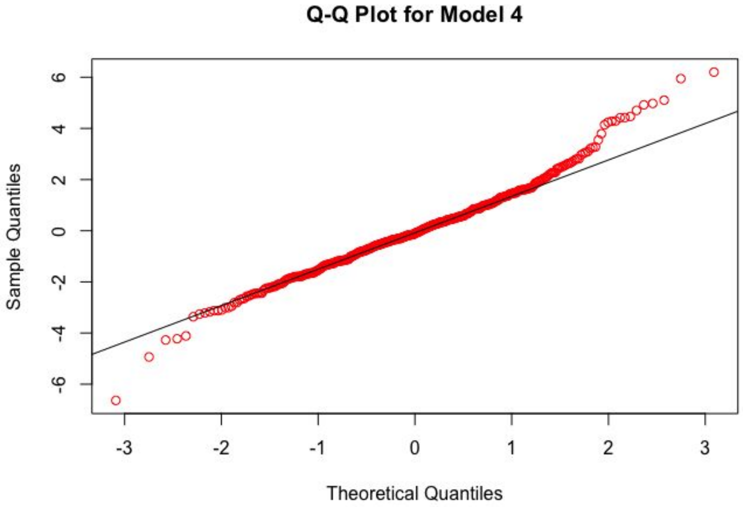


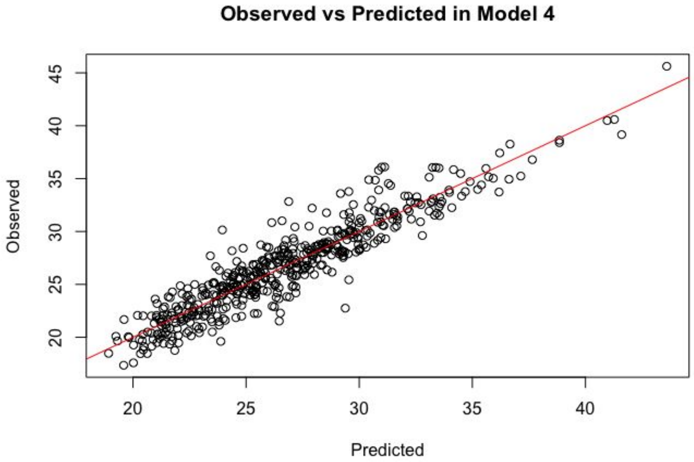


Model 4

$$Y = 15.75 + 0.005308x^2 - 0.00002802x^3$$

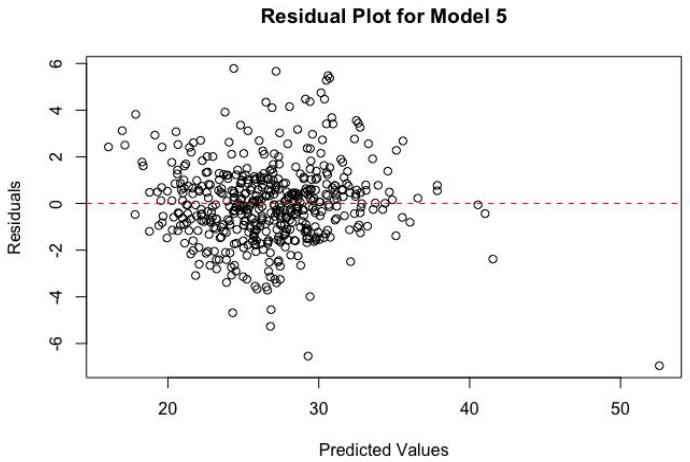


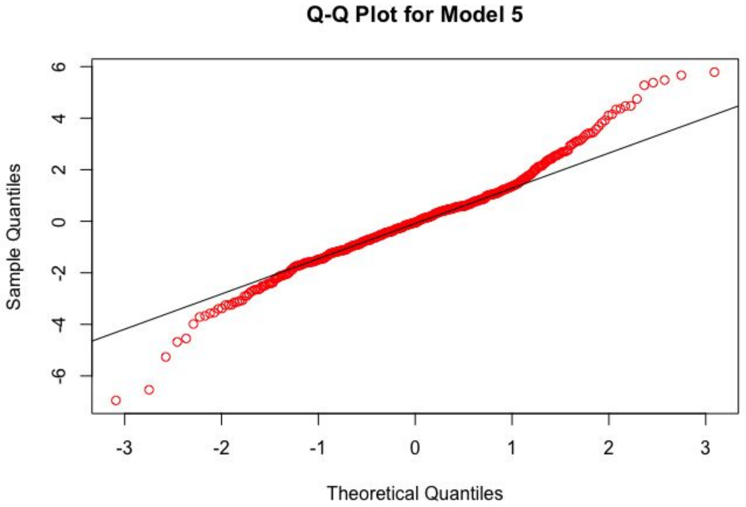


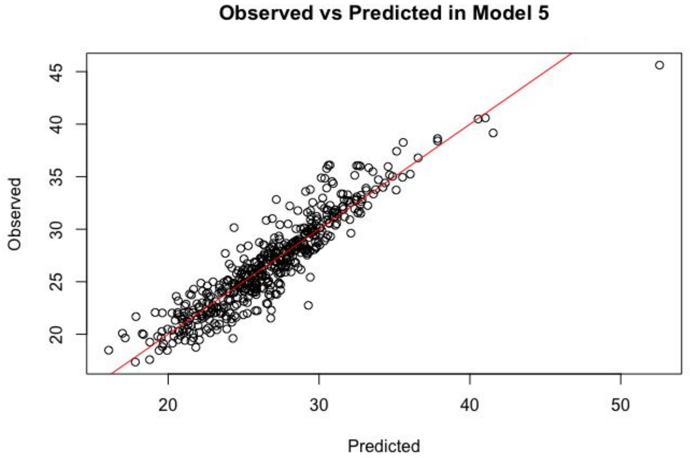


Model 5

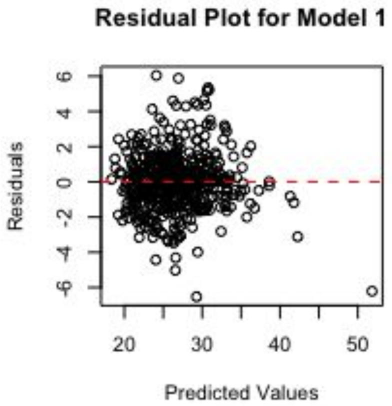
$$Y = 30.73 - 411.7x^{-1} + 0.001373x^{2}$$

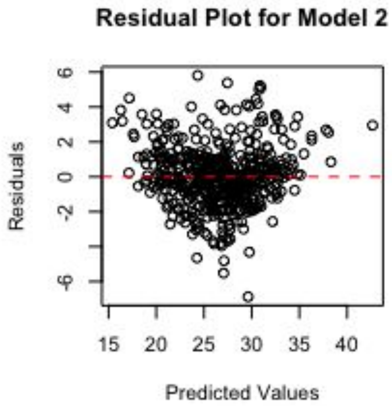


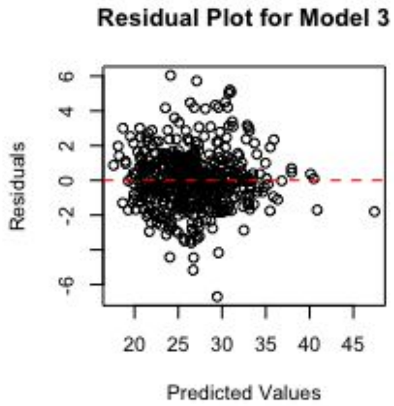


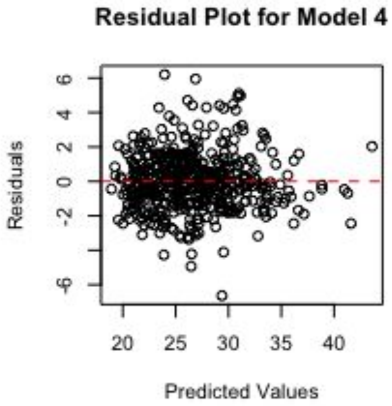


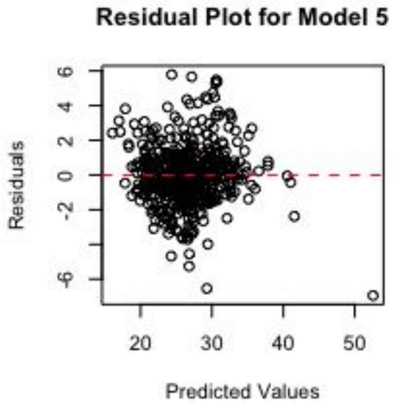
Residual Plot Comparison



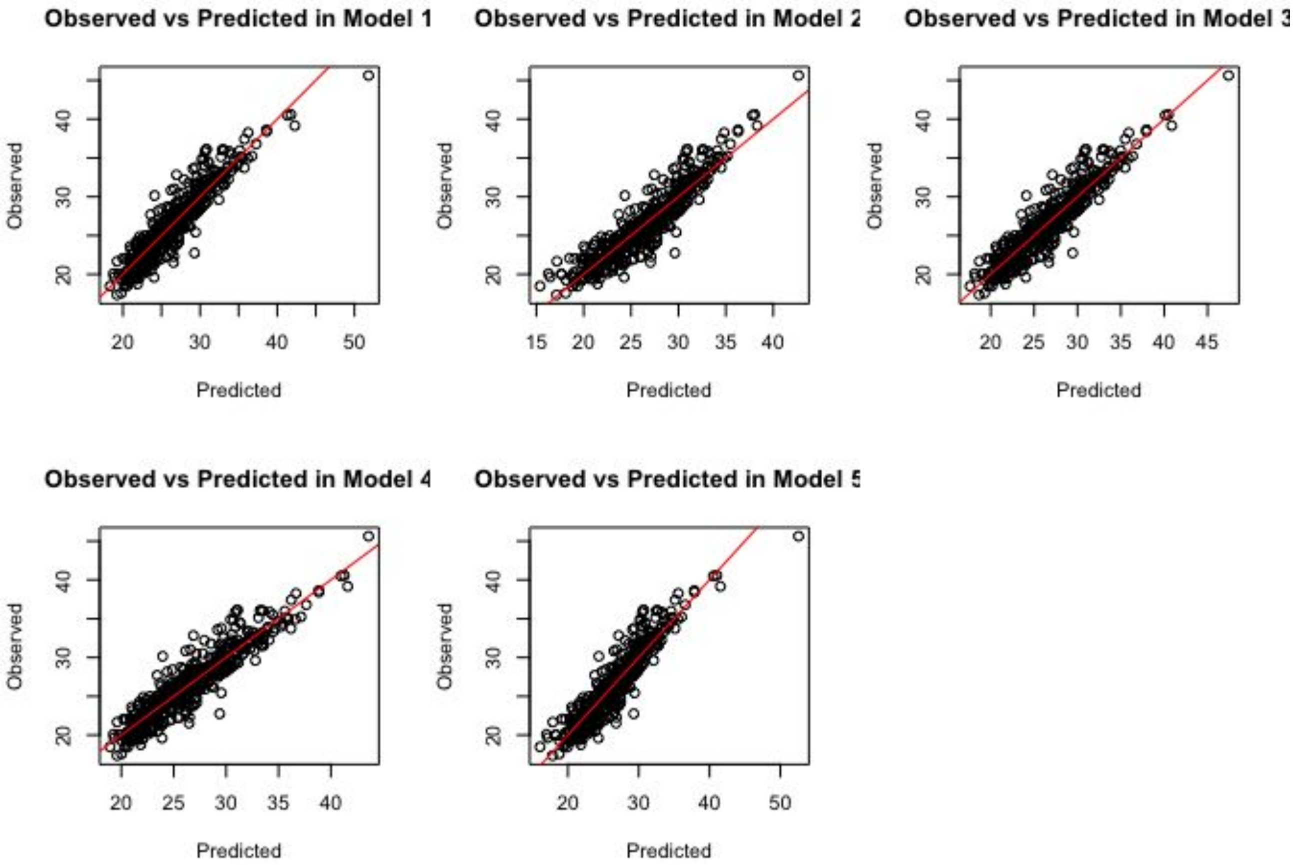




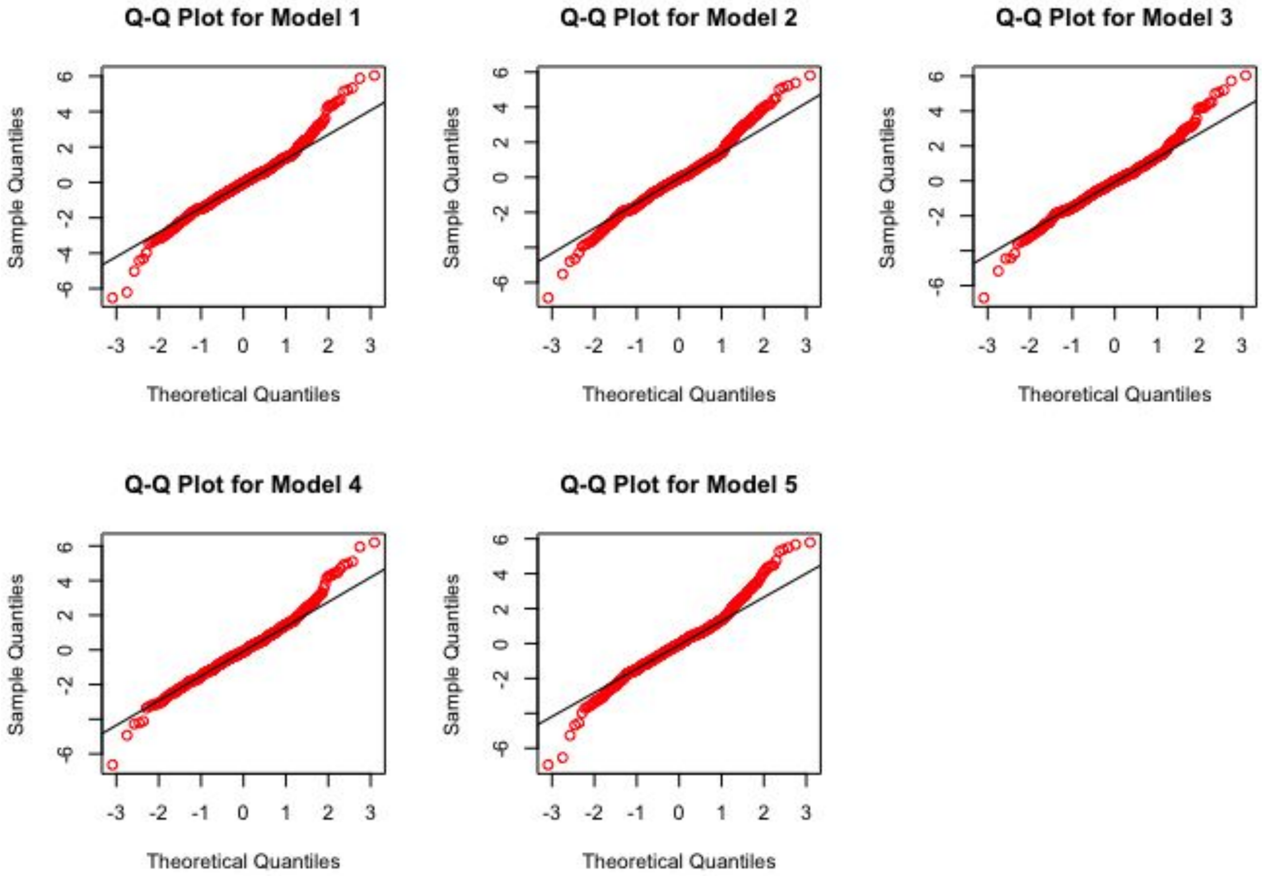




Observed vs Predicted Plot Comparison



Q-Q Plot Comparison



Model Diagnostics Comparison

	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

*Bias values were found to be extremely small, and therefore not significant

Final Selection: Model 4

CI for Betas:

B0 (Intercept): (15.21801e+01, 16.28303)

B1 (diameter^2): (0.004946444, 0.005670146)

B2(diameter^3): (-0.00003143211, -0.00002460142)

```
Call:
lm(formula = height ~ I(diameter^2) + I(diameter^3), data = metasequoia)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.6346 -1.0426 -0.1073  0.8784  6.2001 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)    1.575e+01  2.710e-01   58.11   <2e-16 ***
I(diameter^2)  5.308e-03  1.842e-04   28.82   <2e-16 ***
I(diameter^3) -2.802e-05  1.738e-06  -16.12   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.636 on 497 degrees of freedom
Multiple R-squared:  0.8652,	Adjusted R-squared:  0.8646 
F-statistic:  1594 on 2 and 497 DF,  p-value: < 2.2e-16
```

Model 4 Interpretation

$\beta 0$: We cannot interpret the intercept, as a tree cannot have a dbh of 0.

$\beta 1$: 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.

$\beta 2$: 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.

Model 4 Interpretation

β1: We are 95% confident that as the $(dbh)^2$ of a Chinese Metasequoia tree increases by 1 cm the height of the tree will increase by between 0.004946444 meters and 0.005670146 meters.

β2: We are 95% confident that as the $(dbh)^3$ of a Chinese Metasequoia tree increases by 1 cm the height of the tree will decrease by between 0.00003143211 meters and 0.00002460142 meters.

Findings: Comparison with Paper

From the 5 linear models we replicated, we found Model 4 to be the best fit.

Our findings align with the findings of the paper, where Model 4 was their best fit linear model.

	RMSE	R^2_{adj}	Bias
Our Model 4	1.631005	0.8646189	0.0000
Paper's Model 4	1.8277	0.7583	0.0000

Exploratory Analysis

1	$h = a_0 + a_1 d$
2	$h = a_0 + a_1 \log d$
3	$h = a_0 + a_1 d + a_2 d^2$
4	$h = a_0 + a_1 d^2 + a_2 d^3$
5	$h = a_0 + a_1 d^{-1} + a_2 d^2$

Didn't have a complete dataset.

Didn't fully explain why these explanatory variables are included in these models

Data Processing

tree_number	diameter	height	log.diameter	squared.diameter	cubic.diameter	diameter.to.the.power.of.negativeone
1	134.68	45.62	2.129303	18138.7024	2442920.44	0.007425007
2	64.38	30.10	1.808751	4144.7844	266841.22	0.015532774
3	47.96	28.42	1.680879	2300.1616	110315.75	0.020850709
4	38.75	24.87	1.588272	1501.5625	58185.55	0.025806452
5	40.95	24.99	1.612254	1676.9025	68669.16	0.024420024
6	61.66	28.08	1.790004	3801.9556	234428.58	0.016217970
7	52.88	23.64	1.723291	2796.2944	147868.05	0.018910741
8	59.24	31.77	1.772615	3509.3776	207895.53	0.016880486
9	45.01	30.15	1.653309	2025.9001	91185.76	0.022217285
10	74.40	32.63	1.871573	5535.3600	411830.78	0.013440860

X_1 -> d; X_2 ->d^2; X_3 -> d^3; X_4 -> log(d); X_5 -> d^{-1}

$h = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$ (Multiple linear regression)

Stepwise Selection

Forward/Backward/Forward-Backward/Backward-Forward

AIC/BIC

AIC: Backward AIC Model: 1912.894; Model4: 1916.135

BIC: Forward BIC Model: 1932.985; Model4: 1932.994

Backward Model = Model A; Forward Model = Model B; Model 4 = Model C

Removing Outliers

Model	P-values of SW Test (Before)	P-values of SW Test (After)
Model A	2.051e-06	0.1147
Model B	1.835e-06	0.06006
Model C	2.871e-06	0.2097

Evaluation

Model	AIC	BIC	RMSE	Adjusted R^2
Model A*	1629.31	1650.148	1.321083	0.9077727
Model B*	1608.176	1624.821	1.30867	0.9093569
Model C*	1632.427	1649.098	1.328188	0.9032784

Best Model Based On the 500 Data

Model B:

$Y = 8.659 + 0.3496X_1 - 3.763x10^{-6} X_3$

$Y = 8.659 + 0.3496x - 3.763x10^{-6} x^3$

βs	95% CI	t-statistics	p-values
β_0	(7.8401, 9.4784)	20.771728	1.648354e-68
β_1	(0.3299, 0.3693)	34.862165	1.530932e-132
β_3	($-5.276*10^{-6}$, $-2.249*10^{-6}$)	-4.885436	1.416719e-06

Model B Interpretation

β_0: We cannot interpret the intercept, 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.

Model B Interpretation

β1: 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.

β2: We are 95% confident that as the $(dbh)^3$ of a Chinese Metasequoia tree increases by 1 cm the height of the tree will decrease by between $-5.276*10^{-6}$ meters and $-2.249*10^{-6}$ meters.

Overall Findings

- Overall, we found that out of the 5 linear models proposed in the paper Model 4 was the best fit for the data
- This is in alignment with what the paper found, as they found Model 4 to be the best fit as well
- Although Model 4 was the best fit out of the 5 proposed models, we found Model B (our Forward BIC Model) to be a better fit for the our data
- Due to the limitation of the incomplete dataset, Model B might not be the best model under the full dataset

Challenges

- Limited dataset
    - Only one explanatory variable, so we had very limited number of potential model
    - We were unable to determine if our final model is the best fit for the full dataset, or just our smaller dataset as the model was not tested in the paper