

Final Project Report

Trevor Drees and Emily Howerton

December 15, 2019

Abstract

Introduction and Data

As the world has become significantly more interconnected, the increased movement of people across the world has also facilitated an increase in the spread of species outside of their native ranges. When species that are introduced outside of their native range become invasive, they can have a detrimental impact on local biodiversity, quantity and quality of valuable ecosystem services, and other aspects of the local ecosystem. Because of this, the study of invasive species has been a large part of the ecology literature in recent decades, ranging from theoretical models of how they spread to investigating possible management practices that keep them at bay.

One particular invasive species that has greatly expanded its range due to human activity is *Carduus nutans*, also known as “musk thistle” or “nodding thistle”. This thistle is native to Europe and Central Asia, but has expanded its range into North America, Australia, and New Zealand, among other parts of the world¹. Within the U.S., this thistle has been reported in all U.S. states except for Alaska, Florida, Hawaii, Maine, and Vermont²; the thistle may even be present in these states, but has not yet been reported. It is also been reported all Canadian provinces except Nunavut, Northwest Territories, and Yukon Territories².

C. nutans is considered to be a noxious weed in many U.S. states for several reasons. Because it can occur in very large numbers and grow to be quite large, this thistle may form dense and often impenetrable stands. The plant is also covered in numerous large spines, making it painful when touched as well as unpalatable to grazing animals. The adverse impacts of this weed on grazing can also lead to substantial economic losses.

Another reason why there is concern over *C. nutans* is because it has a high potential to spread locally when introduced, as it is wind dispersed and large pappi on the seeds allow them to be transported great distances. Models have been proposed to model such wind-driven seed dispersal³, and such models have been applied to *C. nutans*⁴, showing significant potential for long-range dispersal events. While there are abiotic and biotic factors that affect how far a seed like those in *C. nutans* can be dispersed, a noteworthy predictor of dispersal distance is seed terminal velocity. For seeds, a higher terminal velocity generally means a decreased dispersal distance; this is because a higher terminal velocity means the seed falls faster and thus spends less time in the air, which means less of an opportunity for wind to carry it further from its source.

However, it is not entirely clear what affects terminal velocity in *C. nutans* seeds, though the most obvious candidates would be physical properties of the seed such as shape and mass. In general, seeds with a larger area perpendicular to the direction of motion will have higher drag and a lower terminal velocity. Seeds with a higher mass will have a higher downward force (mg) from gravity and thus a higher air resistance force that must equal it to achieve terminal velocity, which leads to a higher terminal velocity since said resistance force is proportional to that velocity. However, the physical properties of the seed may be affected by the morphology and physiology of the parent plant; abiotic and biotic factors can affect the parent plant in such a way that may ultimately influence the terminal velocity of its seeds.

Given that there may be a link between abiotic influences on *C. nutans* and the terminal velocity of its seeds, we wish to investigate whether certain treatments applied to the plant before it flowers have any effect on seed dispersal capabilities. Any treatment effects that can reduce the dispersal capability of these thistles may then be used to inform management decisions. By using mowing treatments as well as warming treatments (and combinations of the two), we will examine the effects of said treatments on seed terminal velocity and thus on dispersal capability.

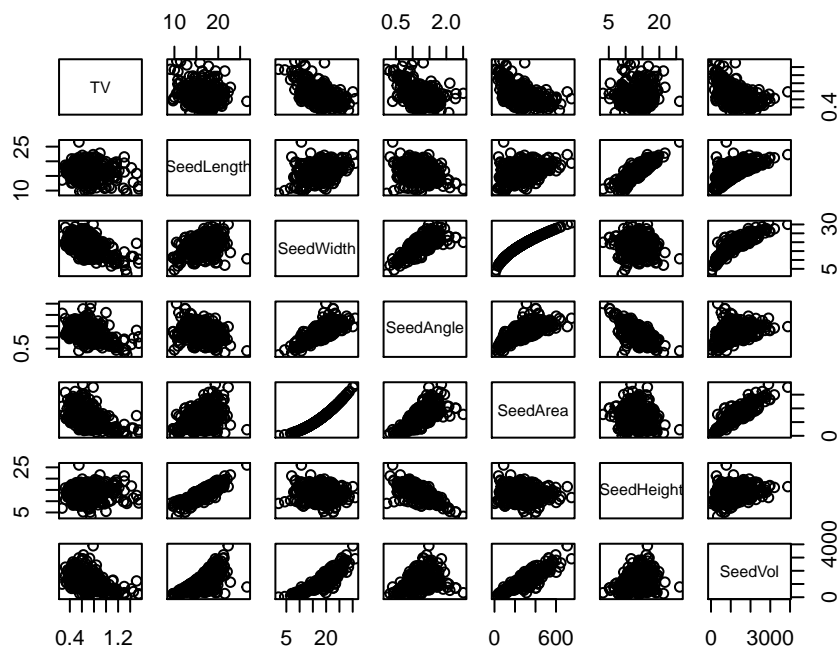
The data used to assess the effects of mowing and warming on seed terminal velocity were collected during a field experiment that involved applying these treatments to parent plants and collecting their flower heads after they had set seed. The experimental setup involves ten blocks, with each block containing two plots: one plot with a warming treatment and one without. Within each plot there are three positions: one with an early mowing treatment, one with a late mowing treatment, and one with no mowing at all. Overall, this yields six unique combinations of warming and mowing, with 10 replicates for each combination. Ten seeds were planted at each position, and one flower head was harvested from all individuals that survived to harvest date. Seeds were collected from individual flower heads and their terminal velocities were determined by timing the amount of time it took them to fall through a 1.25-m drop chamber. For each seed, these drop tests were repeated until two consecutive drop times were recorded within 0.1 seconds of each other.

Using the data from the experiment described above, we seek answers to three research questions of interest:

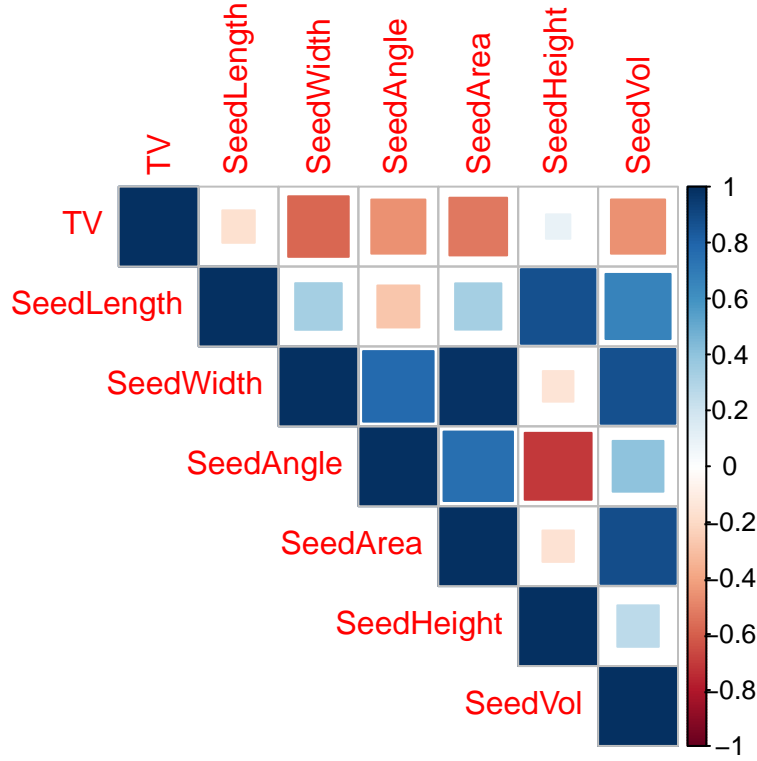
1. Is seed terminal velocity predicted by seed shape parameters?
2. To what extent do warming and mowing treatments change seed terminal velocity?
3. Can changes in terminal velocity by treatment be explained by changes in seed shape parameters?

Research Question 1

Before any models were fit to the data, we first examined correlations between the various seed physical properties. The `pairs` plot for terminal velocity and the various physical properties is shown below:



Linear correlations between the variables can also be shown in a heatmap of the correlation matrix:



Between both of these plots, it is clear that multicollinearity will be an issue since several of the variables are highly correlated with each other (e.g. **SeedWidth** and **SeedAngle**, **SeedWidth** and **SeedArea**, etc.).

Our first model used the terminal velocity **TV** as a response and each of the physical properties as predictors. The resulting coefficients and their significance are listed below:

| | Estimate | Std. Error | t value | Pr(> t) | VIF | Significance |
|-------------|----------|------------|---------|----------|-----------|--------------|
| (Intercept) | 1.7939 | 0.2872 | 6.2465 | 0 | | *** |
| SeedWidth | -0.0902 | 0.0294 | -3.0629 | 0.0024 | 231.7271 | ** |
| SeedLength | 0.2323 | 0.1394 | 1.6664 | 0.0965 | 1672.8695 | . |
| SeedArea | -0.0012 | 0.0021 | -0.5918 | 0.5544 | 918.6289 | |
| SeedAngle | -0.2099 | 0.4397 | -0.4774 | 0.6334 | 291.9196 | |
| SeedHeight | -0.2336 | 0.1536 | -1.5214 | 0.129 | 2709.2157 | |
| SeedVol | 4e-04 | 2e-04 | 2.2275 | 0.0265 | 124.7234 | * |

It is clear that this model is a poor fit for two main reasons. First, many of the terms are not significant to $\alpha = 0.05$ and add no significant additional predictive power to the model. Second, the VIF for all of the coefficients is quite high, as was predicted.

Given that some of these variables must be removed to improve the fit of the model, we then performed variable selection on the full model to reduce clutter. The first method of selection was backwards selection from the full model; this involved removing subsequent variables based on their p -value; the variable with the highest p -value was removed, the resulting model examined, and the process repeated until all predictors were significant. The second method of selection involved backwards selection from the full model using AIC, where terms were removed until the AIC of the model was minimised. The third method of selection started from the full model and used **step** selection in both directions, while the fourth method started from the null model and added terms; again, in both cases, the selection continued until the AIC of the model was minimised.

The results from the model selection can be seen below, along with the VIF for each term as well as both the AIC and R^2 for each resulting model. All coefficients are significant to $\alpha < 0.001$.

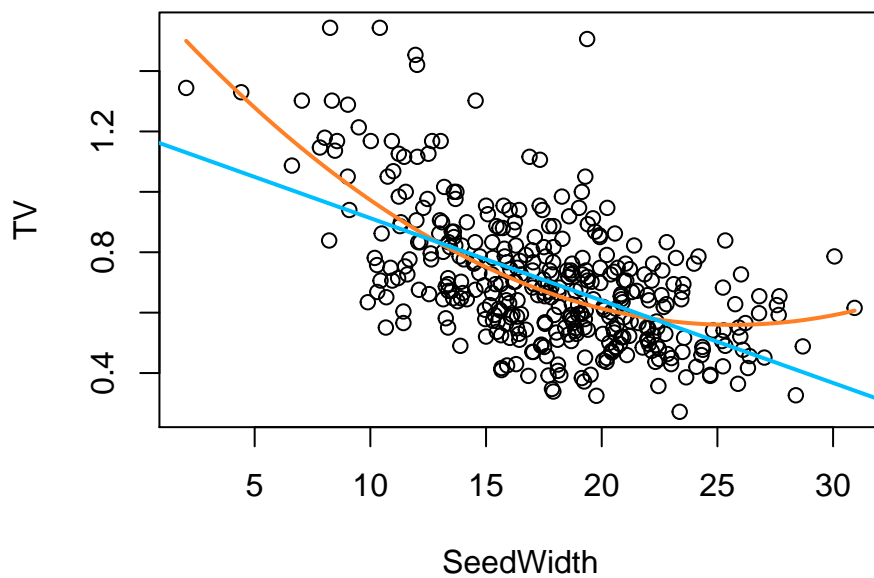
| | Method 1 | Method 2 | Method 3 | Method 4 | Method 1,2,3 VIF | Method 4 VIF |
|-------------|---------------|---------------|---------------|---------------|------------------|---------------|
| (Intercept) | 1.6698759 | 1.6698759 | 1.6698759 | 1.6694531 | | |
| SeedWidth | -0.1027270 | -0.1027270 | -0.1027270 | -0.0865305 | 37.3815562 | 29.5080384 |
| SeedLength | 0.1536113 | 0.1536113 | 0.1536113 | | 69.9871228 | |
| SeedArea | | | | -0.1470827 | | 29.5080384 |
| SeedAngle | | | | | | |
| SeedHeight | -0.1470827 | -0.1470827 | -0.1470827 | | 72.0248940 | |
| SeedVol | 0.0002749 | 0.0002749 | 0.0002749 | | 14.8502466 | |
| AIC | -1293.6000000 | -1293.6000000 | -1293.6000000 | -1288.4000000 | -1293.6000000 | -1293.6000000 |
| R^2 | 0.3921194 | 0.3921194 | 0.3921194 | 0.3803212 | 0.3921194 | 0.3921194 |

As we can see in the table, the first three selection methods produced the exact same result, while the fourth method produced a result with fewer terms. While the model derived from the first three methods has a higher R^2 and lower AIC than the model derived from the fourth method, it suffers greatly from high VIF. Because the VIF on the second model is much lower than that of the first and the difference in R^2 is very small, we will proceed with the second model, effectively trading a small increase in the amount of unexplained variance for a large decrease in multicollinearity.

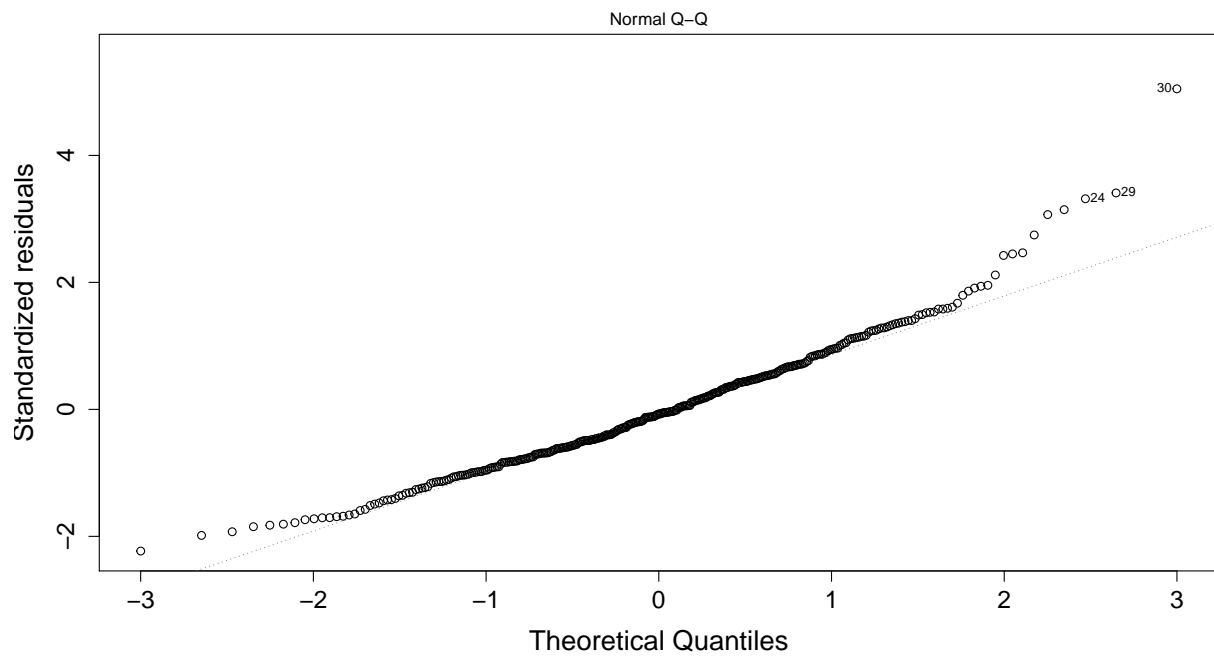
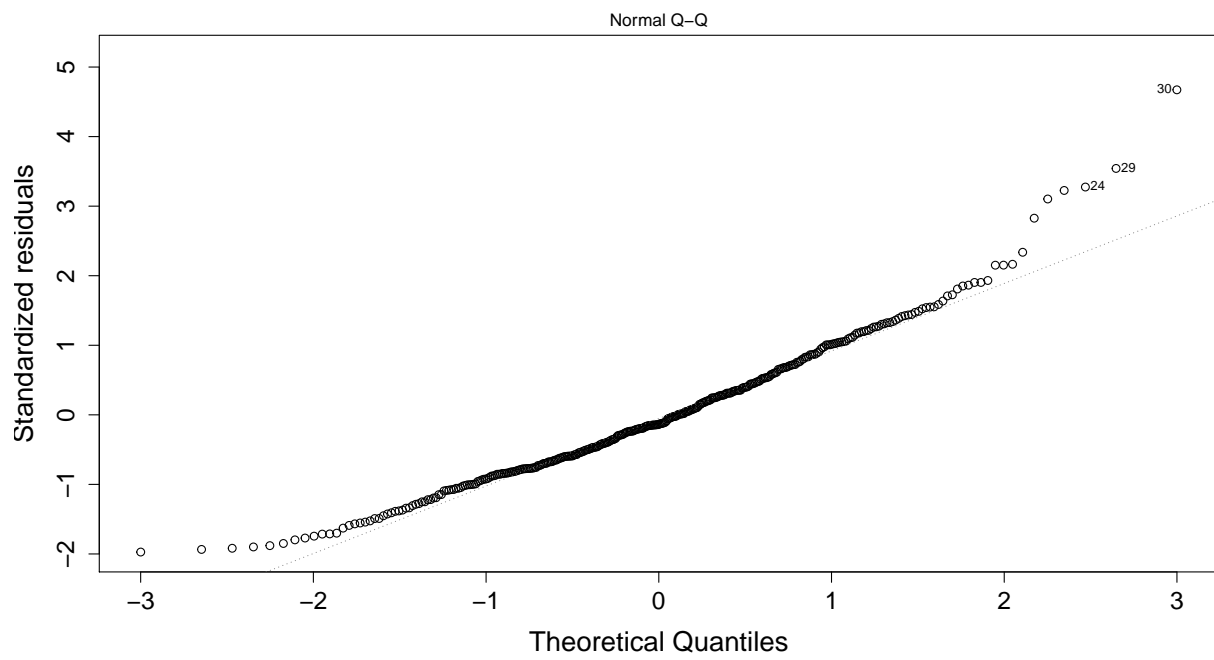
However, we still see a rather high VIF for the **SeedWidth** and **SeedArea** terms. While the **SeedArea** term may be inflating estimates when added to **SeedWidth**, a partial F -test confirms that its addition is statistically significant ($F = 32.43$, $p < 0.001$). Since **SeedArea** is proportional to the square of **SeedWidth**, we can simply express it as such in a polynomial regression instead of using a separate **SeedArea** variable for it.

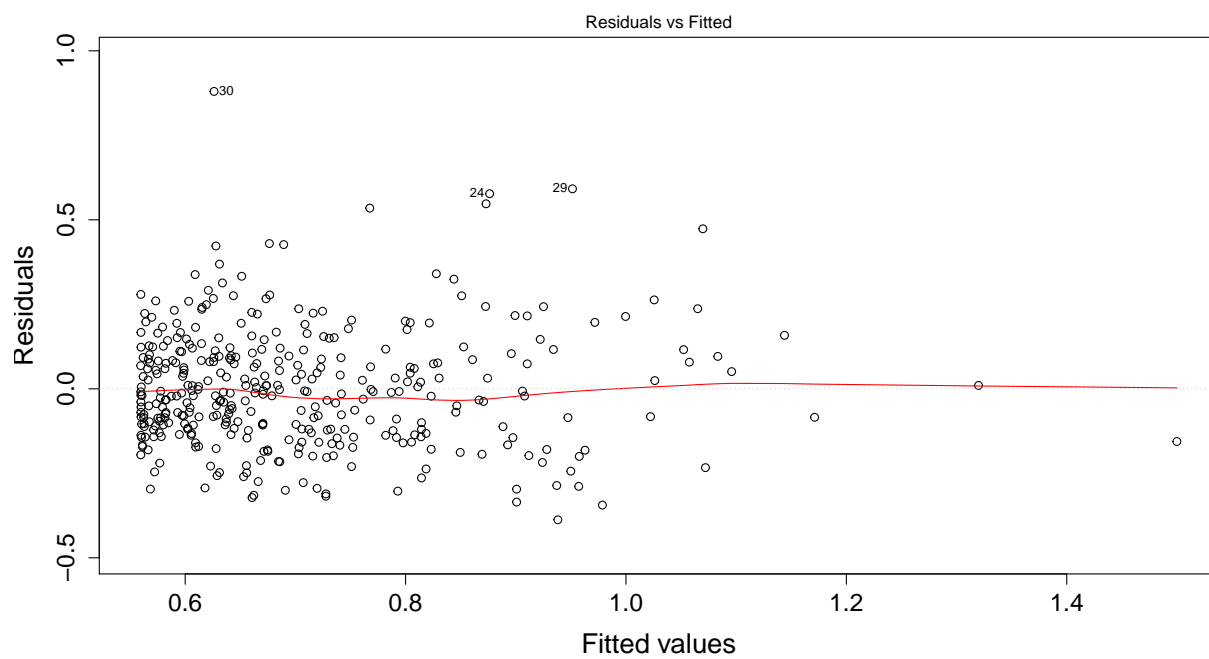
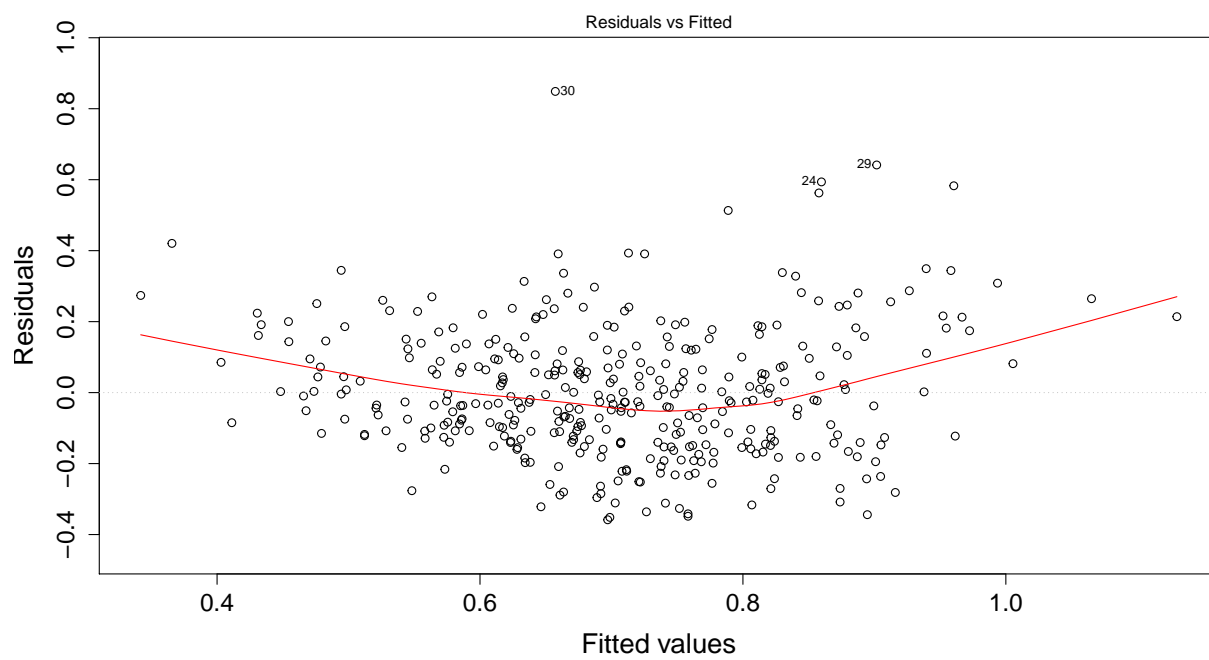
Comparing this to a single-term model where VIF is not an issue, we see that a model with only **SeedWidth** is not as good of a fit, with an R^2 of 0.3274 compared to 0.3804 for the quadratic model. This is also clear when plotting the models against the data, as can be seen below:

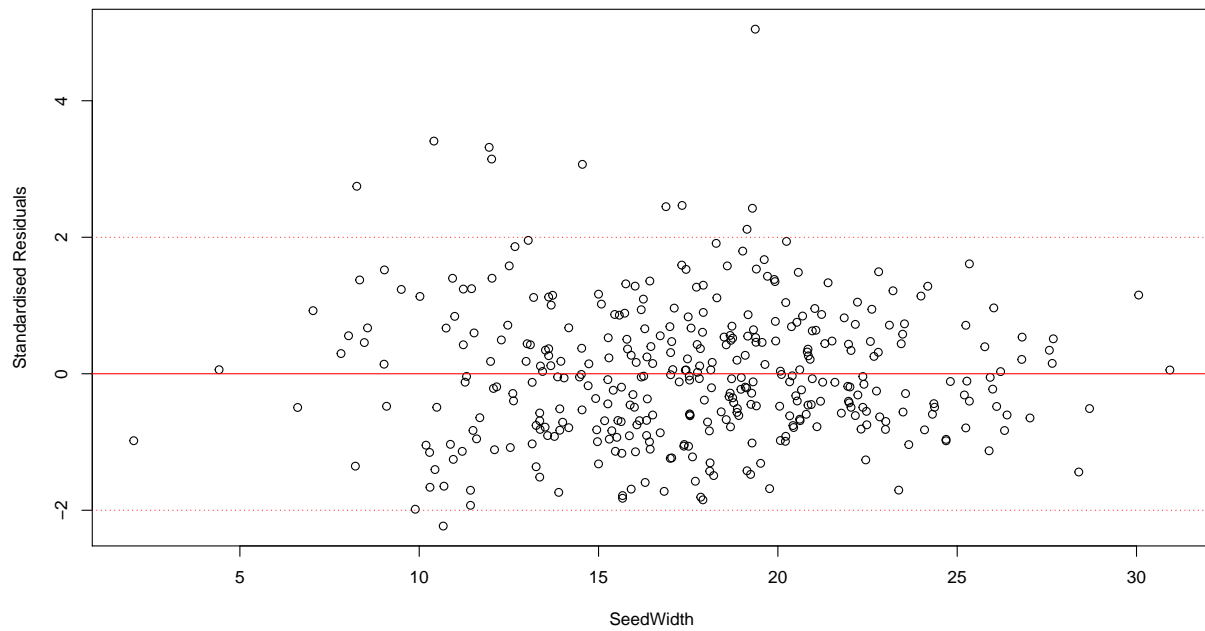
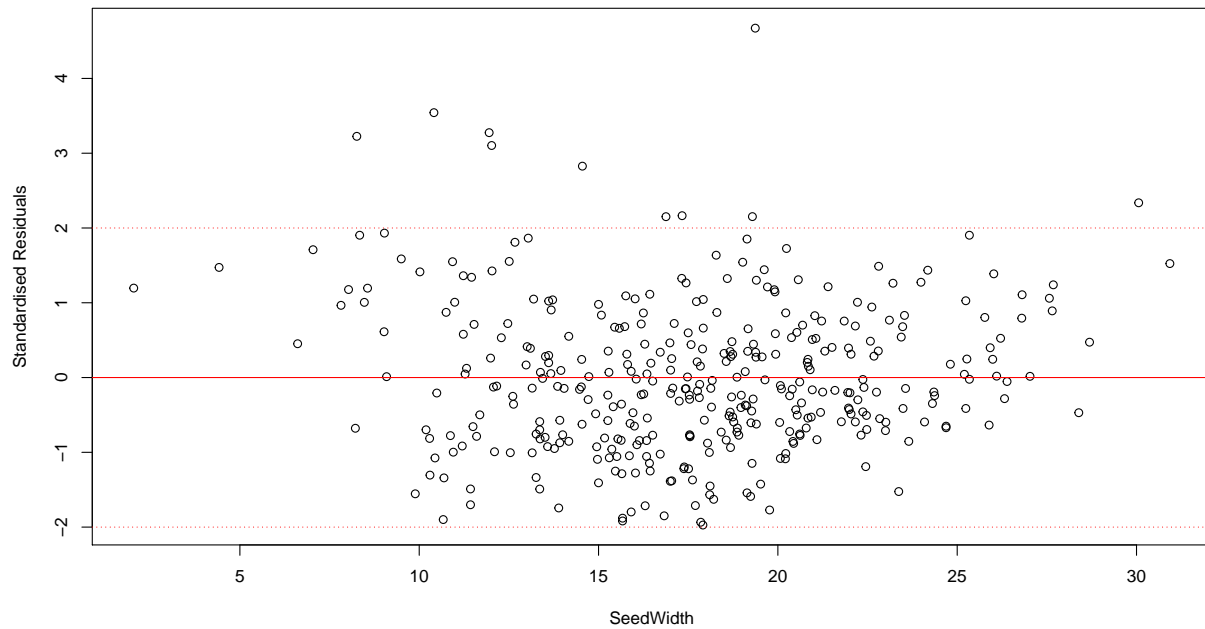
We can also compare the distribution of residuals between the two models to check their validity. As can be seen below, the residuals are approximately normally-distributed for both models, with a bit of deviation from normality at the tails of the distribution.



However, there are clearly differences regarding patterns of distribution in the residuals. For the model with only `SeedWidth`, there appears to be some curvature in the distribution of residuals, suggesting that an invalid model has been fit to the data. However, in the model with both linear and quadratic `SeedWidth` terms, the curved pattern in the residuals disappears, suggesting that this model is more valid.







Research Question 2

Not only are we interested in predicting seed terminal velocity based on the physical properties of the seed, we are also interested in quantifying the extent to which two environmental variables, warming and mowing, affect seed terminal velocity. Figure 1 shows variation in terminal velocity across all six treatments. Using this plot, we can generate a preliminary hypothesis that warming increases terminal velocity, while mowing decreases it. The following analyses will determine whether such differences are statistically significant. As we

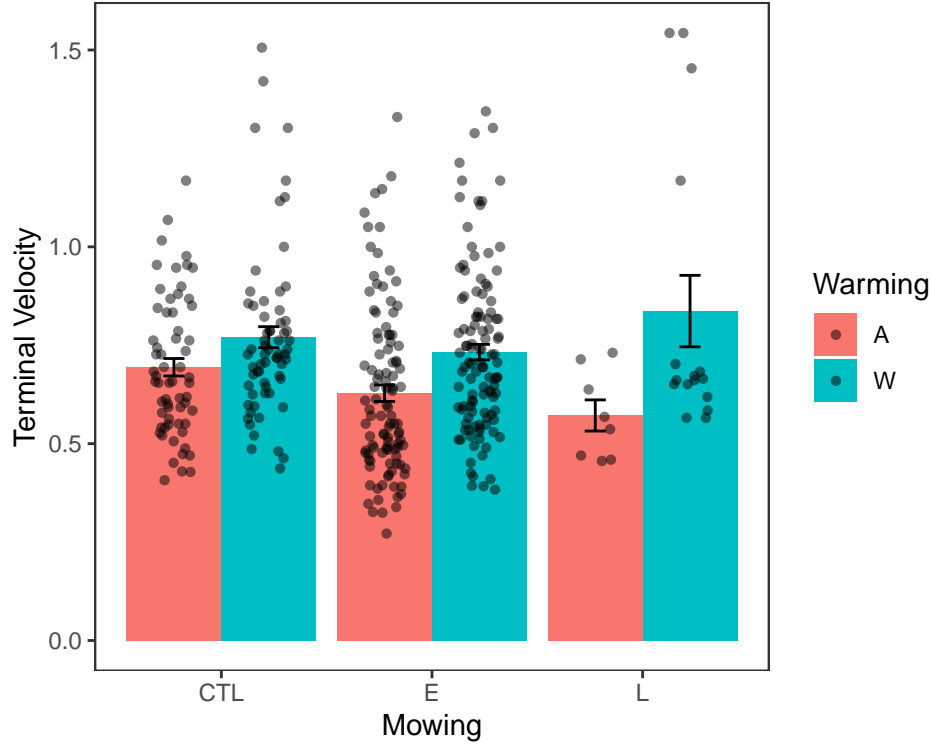


Figure 1: Terminal velocity for plants grown under warming and mowing treatments. Height of bar shows group mean, and error bars show one standard error. Actual data shown with black dots.

can also see in Figure 1, there are fewer points for late mowed plants, as many of them died due to mowing late in the life cycle of the plant.

Because we are fitting a continuous response variable to categorical predictor variables, we will use an ANOVA model to characterize changes in terminal velocity by treatment. We log-transformed terminal velocity to compensate for skewed right distribution of the terminal velocity data and significant deviations from error normality, as seen in Figure 2.

Using log-transformed terminal velocity, we use two methods to identify the most appropriate model. First, we fit the following four models:

- Mowing Only: $TerminalVelocity = \beta_0 + \beta_1 Mow$
- Warming Only: $TerminalVelocity = \beta_0 + \beta_1 Warm$
- Warming and Mowing: $TerminalVelocity = \beta_0 + \beta_1 Mow + \beta_2 Warm$
- Interaction: $TerminalVelocity = \beta_0 + \beta_1 Mow + \beta_2 Warm + \beta_3 Warm \cdot Mow$

We used model p-values for the Mowing and Warming only models to establish which provided a better fit for the data. Then, using this better fitting model as a base, we performed a partial F-test to identify whether adding additional variables to the model significantly improved model fit.

Here, the warming only model has a p-value of 0 whereas the mowing only model has a p-value of 0.0248. Therefore, we use the warming only model as the base model in our partial F-test. Results, shown in the table below, suggest that the model with both mowing and warming, but no interaction term is the best fit for the data.

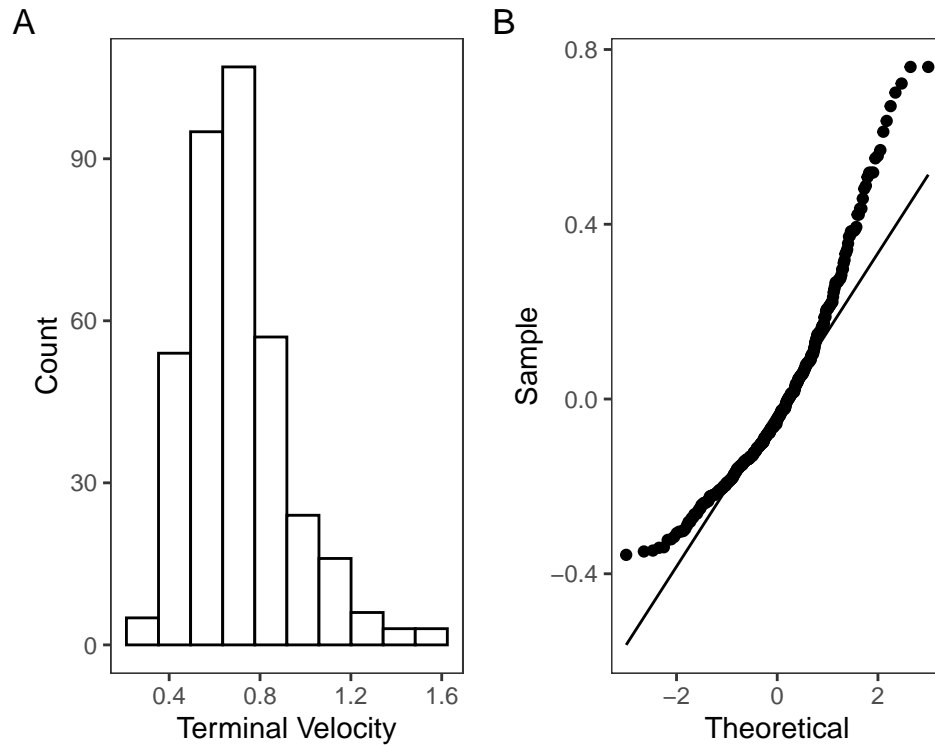


Figure 2: Justification for transformation of terminal velocity. Histogram of terminal velocity (A) and Normal-QQ plot of linear model with both warming and mowing predictors (B).

| Model | P.Value |
|--------------------|---------|
| Warming | |
| Mowing and Warming | 0.0217 |
| Interaction | 0.2134 |

To validate these model selection results, we also performed model selection with both AIC and BIC; the results are shown in the table below.

| Variable | AIC: Forward | AIC: Backward | AIC: Both | BIC: Forward | BIC: Backward | BIC: Both |
|----------------|--------------|---------------|-----------|--------------|---------------|-----------|
| Intercept | 1 | 1 | 1 | 1 | 1 | 1 |
| Mowing | 0 | 1 | 1 | 0 | 1 | 1 |
| Warming | 0 | 1 | 1 | 0 | 1 | 1 |
| Warming:Mowing | 0 | 0 | 0 | 0 | 0 | 0 |

Backward and bi-directional selection for AIC and BIC support the warmign and mowing model, though forward selection suggests an intercept only model in both cases of AIC and BIC. The AIC and BIC for the warming and mowing model is 138.44 and 147.13, respectively, whereas the intercept only model has 165.61 and 169.09 respectively.

Thus, based on the results of the partial F-test and AIC/BIC model selection, we conclude that $TerminalVelocity = \beta_0 + \beta_1 Mowing + \beta_2 Warming$ is the best model for our data. The residuals for this model are shown in Figure 3. The residuals show approximately random distribution, and though there are a few points with high leverage, they do are not high includence accroding to Cook's distance. Lastly, we can see that our log-transformation significantly improved the normality of the errors, as shown in the Normal Q-Q plot.

Model results are shown in the table below and treatment level predictions are shown in Figure 4. From these results, we conclude that early mowing significnatly decreases terminal velocity, whereas warming increases terminal velocity.

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|----------|------------|----------|----------|
| (Intercept) | -0.4232 | 0.0300 | -14.1166 | 0.0000 |
| MowE | -0.0891 | 0.0323 | -2.7618 | 0.0060 |
| MowL | -0.0382 | 0.0646 | -0.5908 | 0.5551 |
| WarmingW | 0.1551 | 0.0302 | 5.1323 | 0.0000 |

Research Question 3

Next, we will explore the extent to which the results of Research Question 2, or the changes in terminal velocity by treatment, can be explained by seed width, as the results of Research Question 1 suggest that seed width is a good predictor of terminal velocity. Figure 5 shows the variation in seed width across treatments.

Our data fitting methods are similar to those of Research Question 2. We use partial F-tests and AIC/BIC analyses for model selection, but did not transform our response variable. First, we analyze the fit of the four basic models - mowing only, warming only, warming and mowing, and interaction - to identify which model is the best fit.

Here, the warming only model has a p-value of 0.0152 whereas the mowing only model has a p-value of 0.0031. Therefore, we use the mowing only model as the base model in our partial F-test. Results, shown in the table below, suggest that the model with both mowing and warming, but no interaction term is again the best fit for the data.

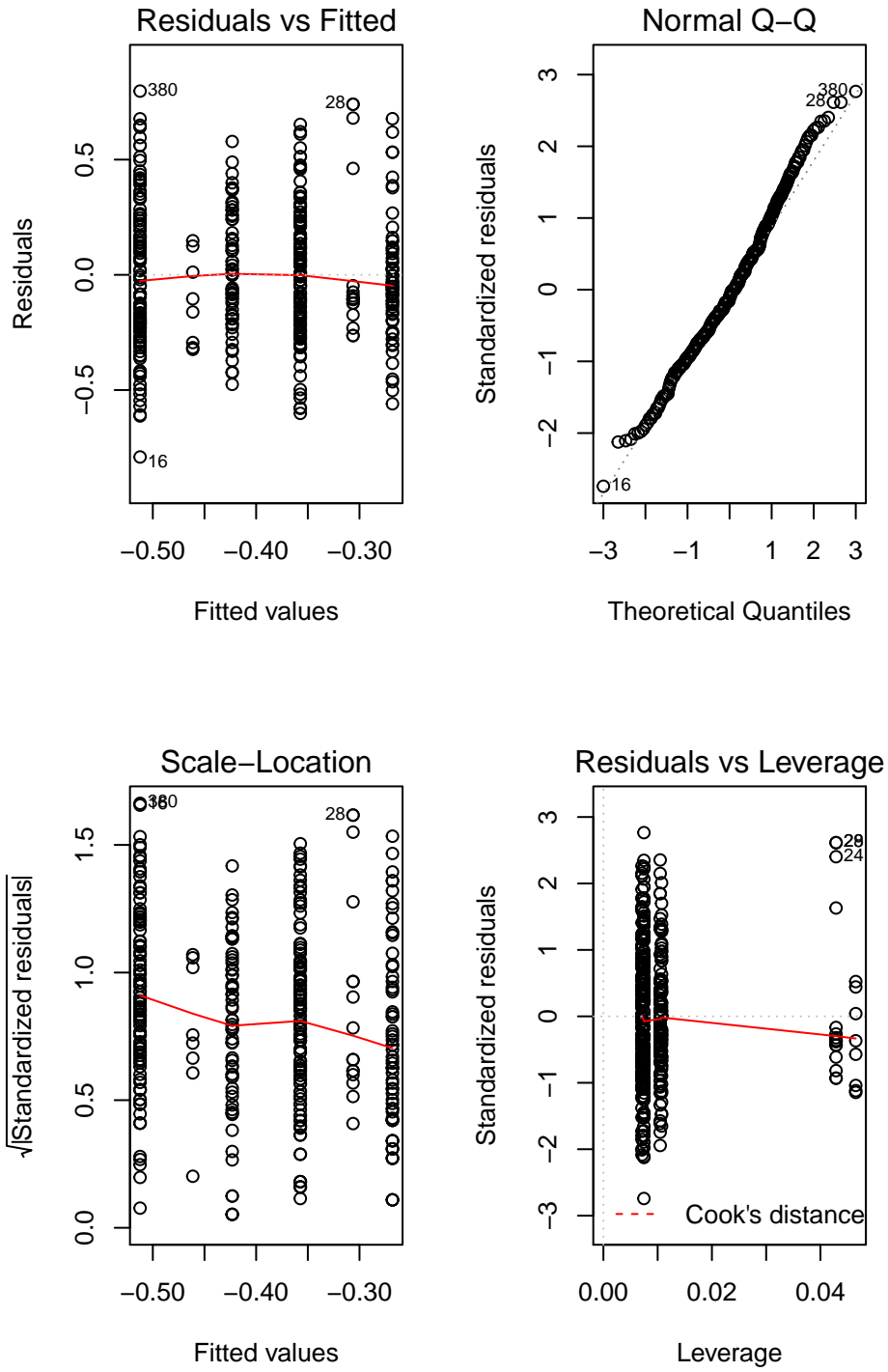


Figure 3: Diagnostic plots for warming and mowing model.

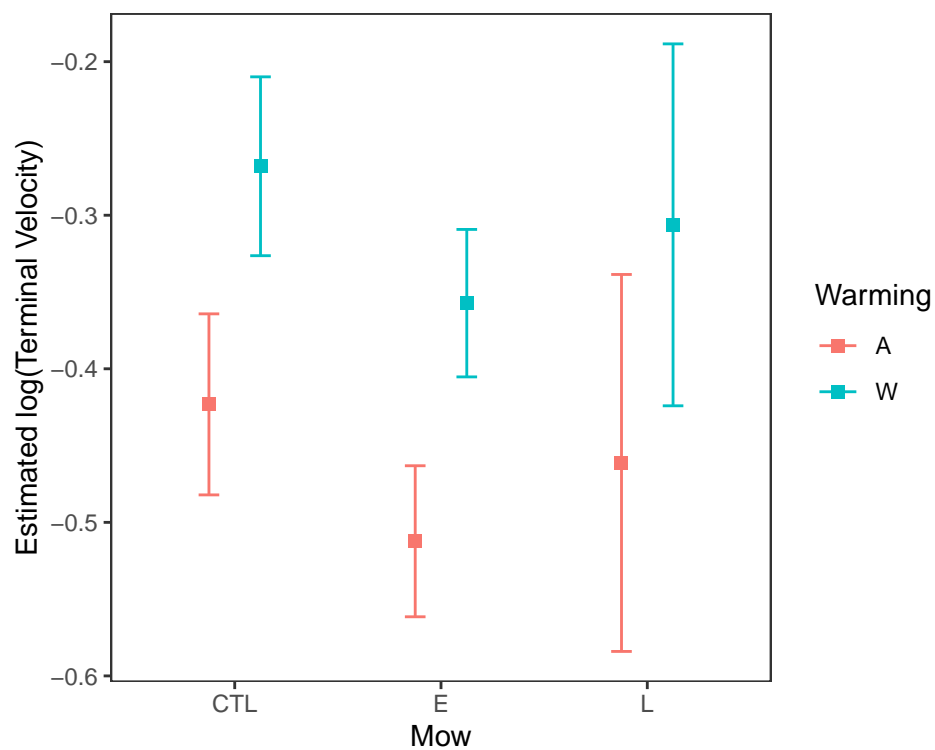


Figure 4: Results of best fitting ANOVA model, which included both warming and mowing as predictors. Square dots show mean predicted value and error bars show 95% confidence intervals.

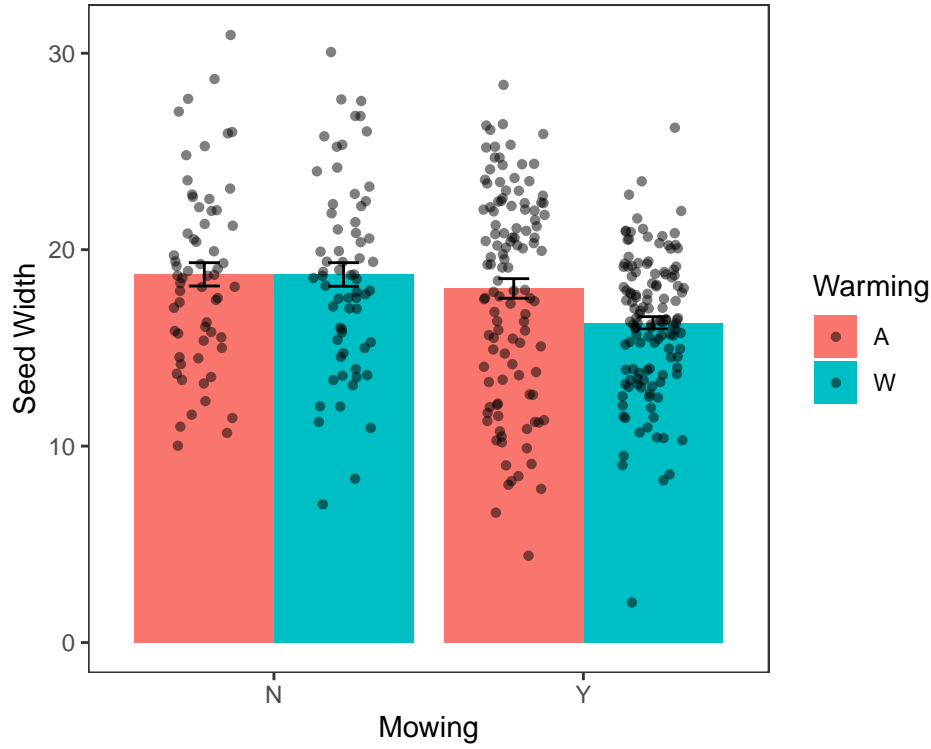


Figure 5: Seed width by warming and mowing treatments. Height of bar shows group mean, and error bars show one standard error. Actual data shown with black dots.

| Model | P.Value |
|--------------------|---------|
| Warming | |
| Mowing and Warming | 0.0198 |
| Interaction | 0.2272 |

Again, we use AIC/BIC model selection to validate these results.

| Variable | AIC: Forward | AIC: Backward | AIC: Both | BIC: Forward | BIC: Backward | BIC: Both |
|----------------|--------------|---------------|-----------|--------------|---------------|-----------|
| Intercept | 1 | 1 | 1 | 1 | 1 | 1 |
| Mowing | 0 | 1 | 1 | 0 | 1 | 1 |
| Warming | 0 | 1 | 1 | 0 | 1 | 1 |
| Warming:Mowing | 0 | 0 | 0 | 0 | 0 | 0 |

Again, we see similar results to those in Research Question 2. The results of the AIC/BIC backward and bi-directional suggest the warming and mowing model is the best fit (AIC: 2181.59 and BIC: 2190.28), whereas forward selection indicates an intercept only model (AIC:2192.72 and BIC:2196.19). Based on the AIC and BIC values of the warming and mowing vs. intercept only model, we conclude that the warming and mowing model is the best fit.

Thus, both the partial F-test and the AIC/BIC model selection support the mowing and warming model, whose model diagnostic plots are shown in Figure 6. These residuals suggest our errors are approximately normally distributed. Again, there may be some leverage points, though they do not have high influence.

The table below shows the results of the mowing and warming model for predicting seed width. These results suggest that both early mowing, late mowing, and warming significantly decrease seed width.

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|----------|------------|---------|----------|
| (Intercept) | 19.3124 | 0.4741 | 40.7347 | 0.0000 |
| MowE | -1.5332 | 0.5105 | -3.0032 | 0.0029 |

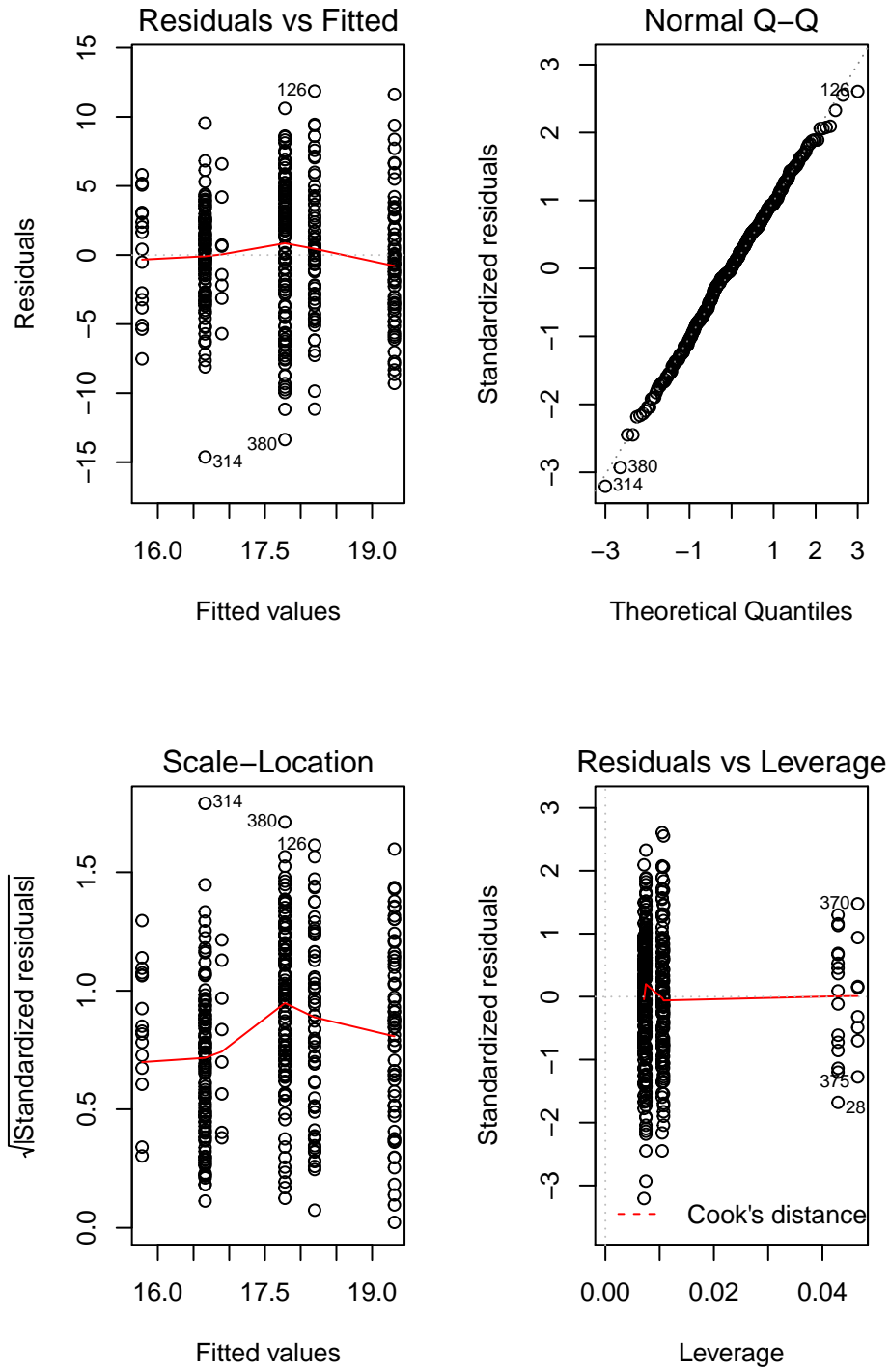


Figure 6: Diagnostic plots for predicting seed width using warming and mowing model.

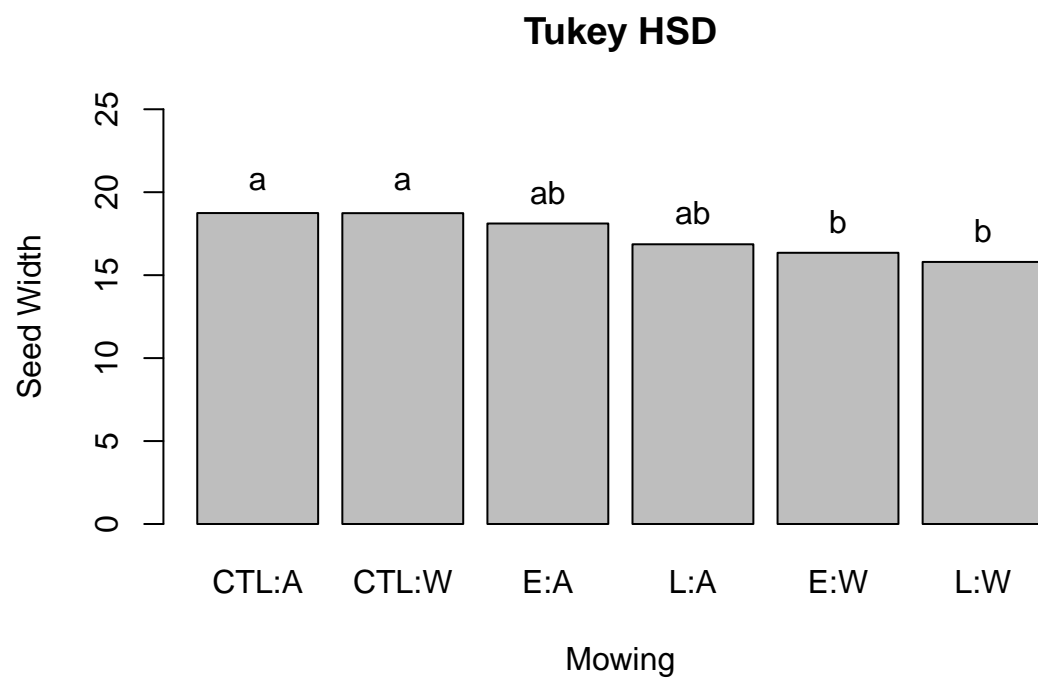


Figure 7: Results of post-hoc Tukey HSD test. Letters indicate groups which are significantly different.