Maple Trees Dataset Regression Analysis

Ian McLeod

October 28, 2017

Introduction and Exploratory Analysis

A tree scientist measured the speed or **velocity** with which the maple keys (or samara) fell to the ground
when released by the tree during autumn. As well the aerodynamic properties of the samara were summarized
in the variable **loading**. Measurements were collected for three different trees in the forest. The tree scientist
would like to summarize this data using regression and find out if the speeds vary between trees after adjusting
for the differing aerodynamic properties of the keys.

An interesting 3D visualization of the samara fruit is available from Wikipedia.

The Maple Dataset may be briefly summarized using R's str() function.

```
## 'data.frame':    35 obs. of  3 variables:
##  $ tree    : Factor w/ 3 levels "Tree 1","Tree 2",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ loading : num  0.239 0.208 0.223 0.224 0.246 0.213 0.198 0.219 0.241 0.21 ...
##  $ velocity: num  1.34 1.06 1.14 1.13 1.35 1.23 1.23 1.15 1.25 1.24 ...
```

The notched boxplot shown in Figure 1 indicates that Tree 3 **velocity** is less than that of Trees 1 and 2 while
Trees 1 and 2 don't look much difference. Using regression analysis we can investigate whether this holds
true after adjusting for **loading**. Sometimes this particular application of regression is called *analysis of
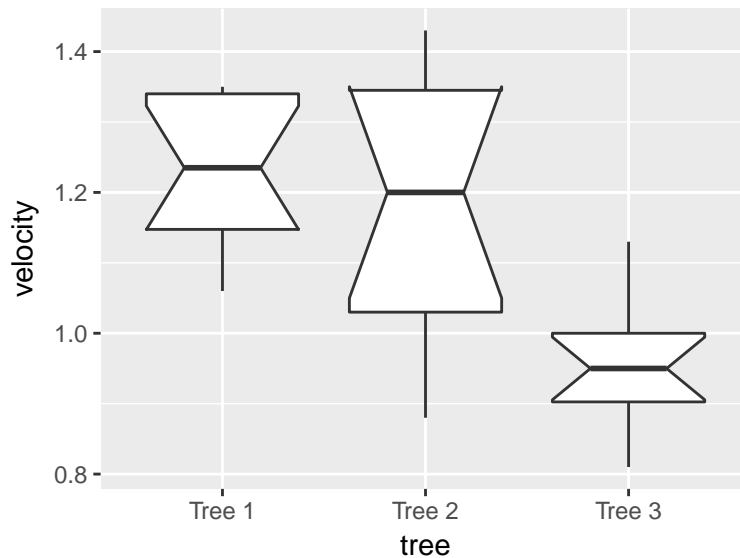covariance*.

Figure 1: Notched boxplot

As pointed out by Anscombe it is important visualize the data before fitting it. We look for unusual features such as systematic nonlinear, changes in variability, discontinuities and major outliers.
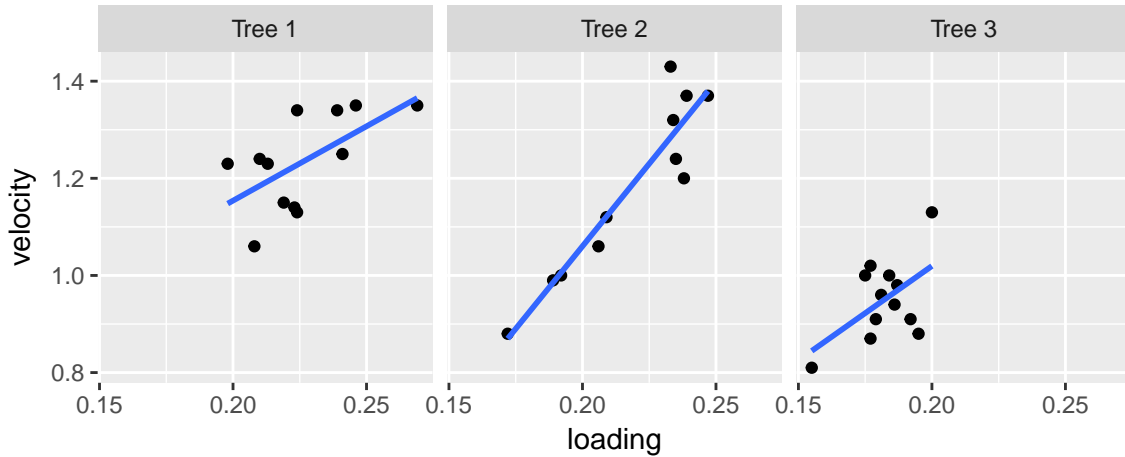

Figure 2: Maple Trees dataset with fitted least squares regression

Simple Linear Regression Model

From the initial plot we see that the slope parameters are certainly non-zero. They appear to be different but it is not clear whether or not this difference is statistically signficant. It is reasonable to begin by fitting a simple linear regression assuming the slope and intercept parameter are the same for each tree. The fitted model is summarized in Table 1.

Table 1: Model Summary

	Dependent variable:
	velocity
loading	5.820***
	(0.511)
Constant	−0.093
	(0.107)
Observations	35
R^2	0.797
Adjusted R^2	0.791
Residual Std. Error	0.081 (df = 33)
F Statistic	129.631*** (df = 1; 33)
Note:	*p<0.1; **p<0.05; ***p<0.01

We see that $R^2 = 79.7\%$ so the model accounts for a lot of variation and the overall regression is signficant at less than 0.1% which is agrees with the visualization in Figure 2.

Common Intercept Model

Before doing further diagnostic checks, lets see if a slightly more complex model gives a better fit. So next we consider the model where the intercept term is the same for all trees but the slopes are different. The fitted model is summarized in Table 2.

Table 2: Model Summary: Different slopes

	Dependent variable:
	velocity
treeTree 1:loading	5.169***
	(0.707)
treeTree 2:loading	5.158***
	(0.732)
treeTree 3:loading	4.873***
	(0.880)
Constant	0.062
	(0.159)
Observations	35
R^2	0.808
Adjusted R^2	0.790
Residual Std. Error	0.081 (df = 31)
F Statistic	43.531*** (df = 3; 31)
Note:	*p<0.1; **p<0.05; ***p<0.01

Comparing the slope estimates with their standard errors, it would seem that the slopes are not signficantly different. We can use the extra-sum-of-squares ANOVA test to check this. The P-value for testing the null hypothesis that the slopes and intercepts are equal (Table 1) against the model with equal slopes and intercepts not equal (Table 2) is 41.9%. So there is no evidence to support the claim the slopes are different.

Common Slope Model

So the model allowing different slopes, does not improve the fit of the simple model. Another simple model would be to allow the intercepts to vary but assume the slopes are the same.

Table 3: Model Summary: Different intercepts

	Dependent variable:
	velocity
treeTree 2	−0.010
	(0.034)
treeTree 3	−0.059
	(0.046)
loading	5.123***
	(0.739)
Constant	0.076
	(0.169)
Observations	35
R^2	0.808
Adjusted R^2	0.789
Residual Std. Error	0.081 (df = 31)
F Statistic	43.426*** (df = 3; 31)
Note:	*p<0.1; **p<0.05; ***p<0.01

The P-value for testing the null hypothesis that the intercepts are equal against the model with intercepts not equal (but intercepts are assumed equal - see Table 2) is 43.2%. So there is no evidence that the intercepts are different as we would have expected from Table 3.

Full Model: Slopes and Intercepts are Different

Finally let's see if fitting the full model is an improvement over the simple linear regression model in Table 1.

Table 4: Model Summary: Full model, different slopes and intercepts

	Dependent variable:
	velocity
treeTree 2	−0.841**
	(0.336)
treeTree 3	−0.299
	(0.445)
loading	3.063**
	(1.160)
treeTree 2:loading	3.734**
	(1.500)
treeTree 3:loading	0.820
	(2.284)
Constant	0.541**
	(0.263)
Observations	35
R^2	0.844
Adjusted R^2	0.817
Residual Std. Error	0.076 (df = 29)
F Statistic	31.294*** (df = 5; 29)
Note:	*p<0.1; **p<0.05; ***p<0.01

The P-value for testing the null hypothesis that the simple model (Table 1) where the slopes and intercepts are equal against the model with slopes different but intercepts equal (Table 3) is 9.9%. So there is no evidence to support the claim the slopes are different.

Diagnostic Checks for the Simple Model

The basic diagnostic checks do not indicate any problem with the Simple Linear Regression Model.
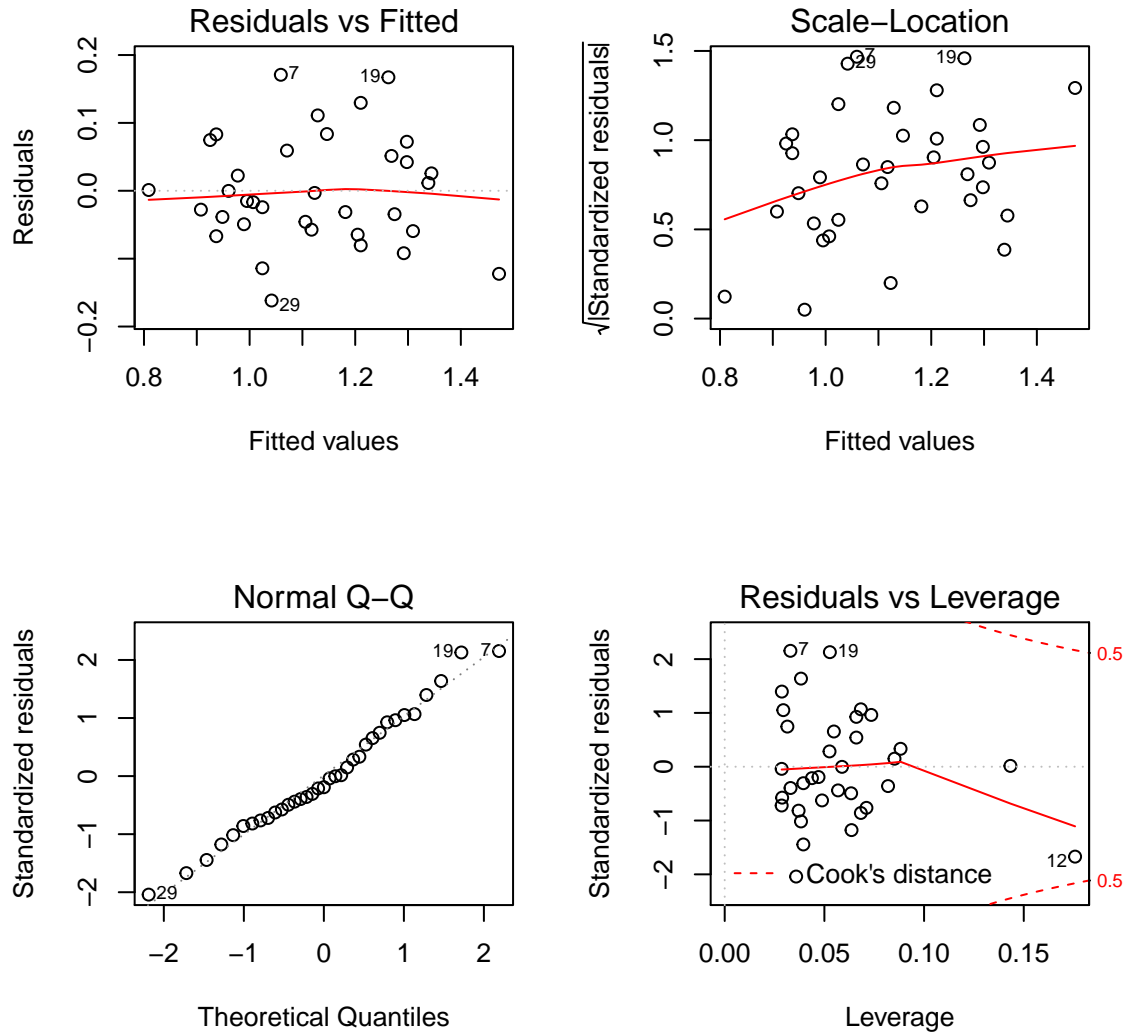

Figure 3: Basic Diagnostic Checks for Full Model

Since there is only one input in the Simple Model, the residual dependency plot does not add beyond what is revealed in the Residuals vs Fit Plot.

The p-value for the Jarque-Bera normality test is approximately 68.3% agrees the conclusion from the normal probablity plot which indicates there is no evidence against the normal distribution assumption.

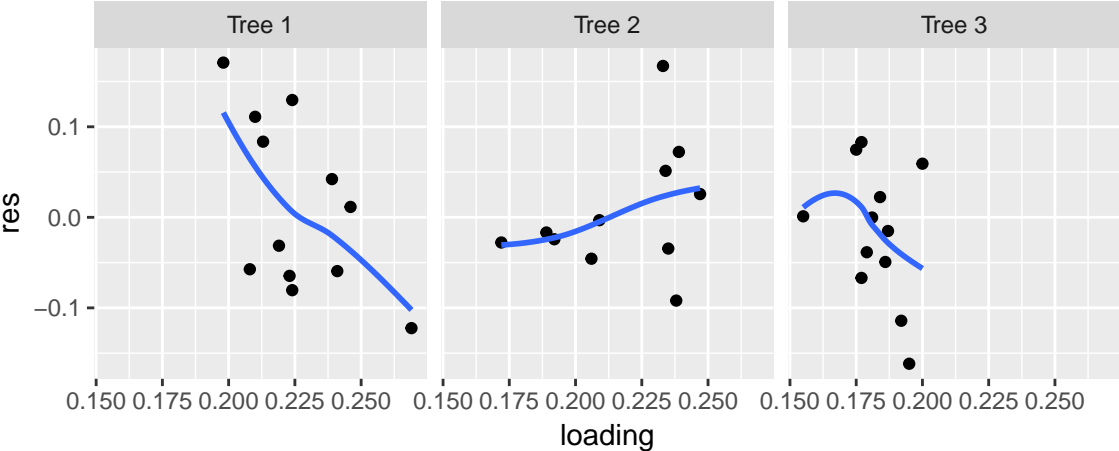Figure 4 shows the residual dependency plot conditional on Tree. Visually this suggests lack-of-fit but our statistical analysis did not support this.

Figure 4: Residual dependency plot conditional on Tree