Homework 1 STAT 505 Fall 2014

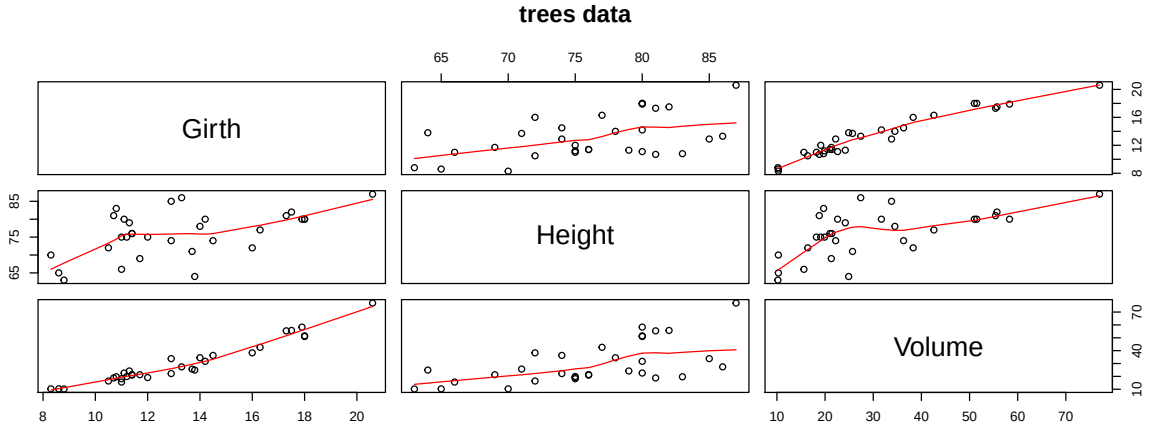
Leslie Gains-Germain

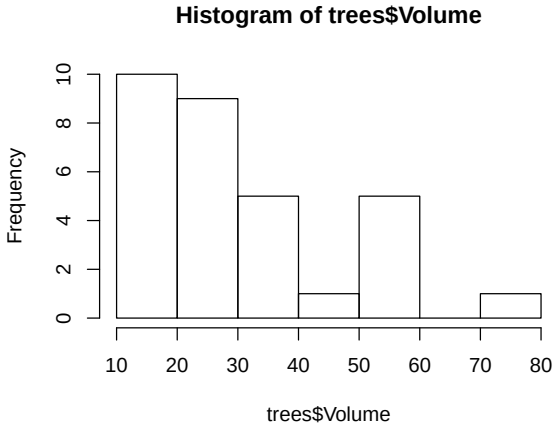
Introduction

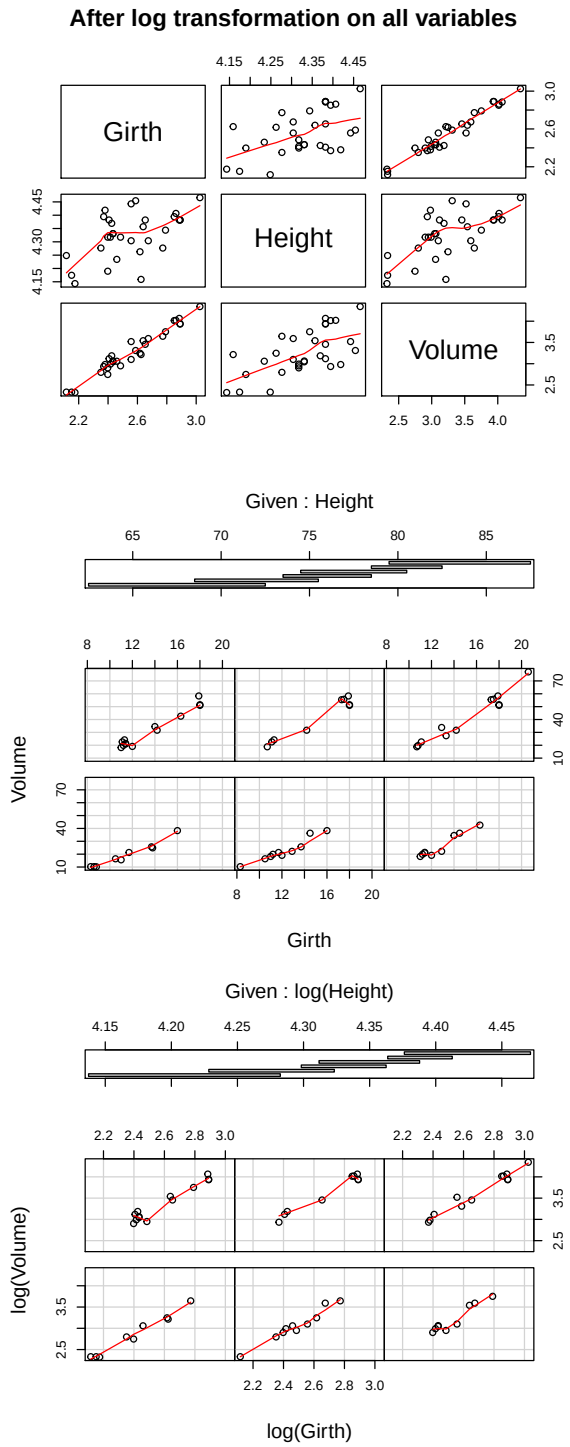
The girth (in), height (ft), and volume (ft^3) of 31 *felled black cherry trees was recorded. No additional information is given about the method of data collection or the location of the trees.*

Exploratory Analysis

In general, trees with larger girth tend to be taller and have larger volume. More specifically, volume $= \pi r^2 h$, *or* $V = \frac{1}{4\pi} C^2 h$ *where C=circumference or girth. Therefore, I would expect the relationship between girth and volume to be curved and this is what we see in the pairs plot below. A log transformation on girth may be appropriate before fitting a linear model to these data. Second, we see increasing variability in the sample volumes as the mean volume of trees increases (see the volume vs. height plot below). Lastly, the histogram of tree volumes shows that the data are right skewed. A log transformation on volume will help equalize the standard deviations and make the data more symmetric.*





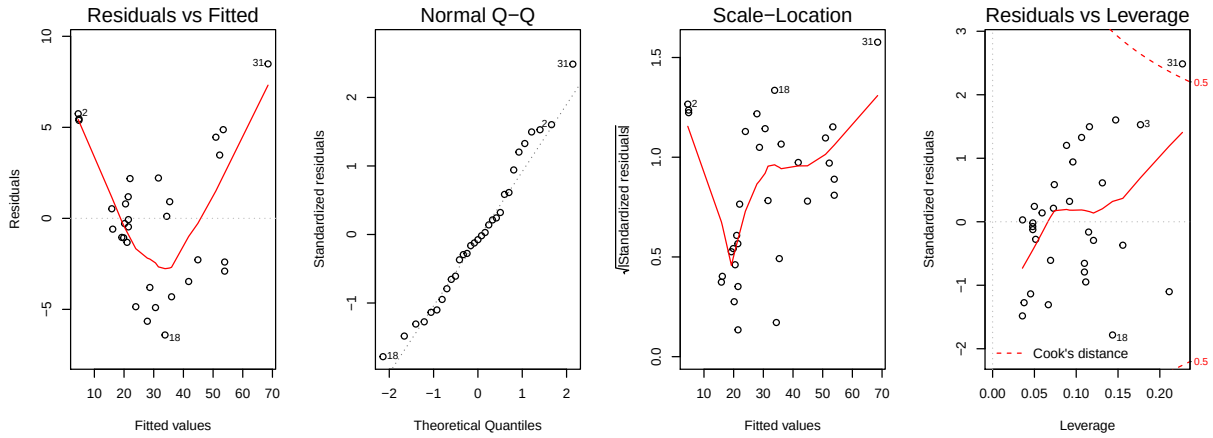


After the log transformation of all the variables, we reexamine the pairs plot. We can see that the relationship between girth and volume looks more linear. The relationship between height and volume also looks more linear with more constant spread. The coplot function shows the relationship between girth and volume for tree heights within specified intervals. If we switch the two predictors in coplot(), we are shown the relationship between height and volume for tree girths in specified intervals. We can see in the above plots that the relationship between girth and volume at specified heights becomes more linear after the transformation.

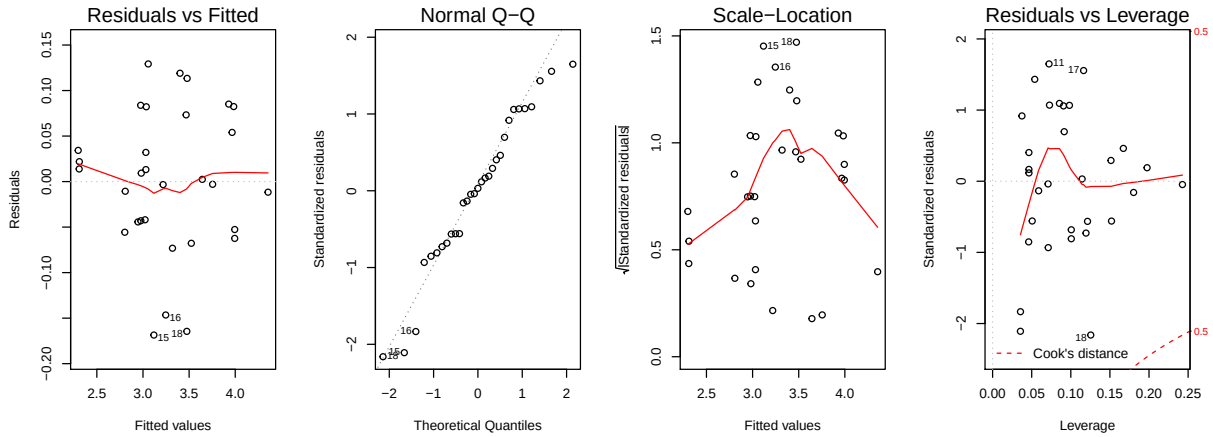
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-57.9877	8.6382	-6.71	0.0000
Height	0.3393	0.1302	2.61	0.0145
Girth	4.7082	0.2643	17.82	0.0000

I first fit a linear model on the original scale. According to this model, there is moderate evidence of a relationship between the mean volume of trees and tree height (p-value= 0.0145*). There is strong evidence of a relationship between the mean volume of trees and tree girth (p-value<* 0.0001*). The mean volume of trees is estimated to increase by* 0.339 *cubic feet for a one foot increase in tree height with a* 95% *confidence interval from a* 0.073 *to a* 0.606 *cubic foot increase. The model also estimates that a one inch increase in tree girth is associated with a* 4.708 *cubic foot increase in the mean volume of trees with a* 95% *confidence interval from a* 4.167 *to a* 5.249 *cubic foot increase.*

Discussion of Assumptions



The diagnostic plots do show some problems with the "usual assumptions". The Residuals vs. Fitted Values plot shows that the residuals are not centered at 0 *for all fitted values. For some fitted values, all the residuals are positive and for others all are negative. This indicates that a line is not appropriate to model these data, or in other words the linearity assumption is not met. The constant variance assumption also does not appear to be adequately met because the spread of the residuals does not seem to be constant across fitted values. The normal Q-Q plot shows that the points do lie along a straight line, which indicates that the normality assumption is met. There is one outlier, however, which is flagged in the last diagnostic plot. We will also assume the independence assumption is met although we have no information to verify this assumption.*



After the log transformation, we see that the residuals are more centered at 0*. Although the spread appears to change somewhat across fitted values, I wouldn't expect the spread to appear perfectly constant with a relatively small sample size of* 31*. The assumptions of linearity and constant variance are adequately met. We do see some departures from normality in the tails, but I am not too worried about this because our parametric tools are robust to departures from normality. Additionally, there are no outliers flagged after the transformation.*

Summary of Findings

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-6.6316	0.7998	-8.29	0.0000
log(Girth)	1.9826	0.0750	26.43	0.0000
log(Height)	1.1171	0.2044	5.46	0.0000

There is strong evidence of a relationship between tree girth and the median tree volume on the log scale (p-value< 0.0001*). The median tree volume is estimated to change by* 3.95 *times for every doubling of tree girth with an associated* 95% *confidence interval from* 3.55 *to* 4.40 *times. There is also strong evidence of a relationship between between tree height and the median tree volume on the log scale (p-value<* 0.0001*). The median tree volume is estimated to change by* 2.17 *times for every doubling of tree height with an associated* 95% *confidence interval from* 1.62 *to* 2.90 *times. Note also in the model summary that the coefficient for girth is close to* 2 *and the coefficient for height is close to* 1 *which is what we would expect mathematically (*$V = \frac{1}{4\pi} G^2 h$ *and* $V = e^{\beta_0} G^{\beta_1} h^{\beta_2}$ *after backtransforming).*

Scope of Inference

Since it is not known whether the trees in the study were randomly selected, inference extends only to the 31 *trees in the study. Since this was a purely observational study and no treatments were imposed, we cannot establish whether there is a causal relationship between [height, girth] and volume of trees.*

R Code Appendix

```
require(datasets)
data(trees)
  ## also get a description with:
help(trees)
  ## and examine structure with
#str(trees)
par(mfrow=c(1,2))
pairs(trees, panel = panel.smooth, main = "trees data")
#par(mfrow=c(1,3))
#plot(Volume ~ Girth, data = trees)
#plot(Volume ~ Height, data = trees)
#plot(Height ~ Girth, data = trees)
#library(lattice)
```

```
par(mfrow=c(1,2))
coplot(Volume ~ Girth | Height, data = trees,
      panel = panel.smooth)
coplot(log(Volume) ~ log(Girth) | log(Height), data = trees,
      panel = panel.smooth)
```

```
require(xtable)
fit1 <- lm(Volume ~ Height + Girth, trees)
xtable(summary(fit1))
fm2 <- lm(log(Volume) ~ log(Girth) + log(Height), data = trees)
```

```
par(mfrow=c(1,4))
plot(fit1)
```

```
require(stats); require(graphics); require(xtable)
fm2 <- lm(log(Volume) ~ log(Girth) + log(Height), data = trees)
xtable(summary(fm2))
#step(fm2)
```

```
coef(fm2)
2^coef(fm2)
confint(fm2)
2^confint(fm2)
```