**Joseph Crockett**

**February 23, 2016**

**ES 207: Environmental Data Analysis**

**Homework 3: Pre-regression**

Objective Statement:  
Aboveground carbon stocks, as a function of biomass, cannot be directly measured without destroying the sample. An alternative is to estimate volume using measurments of tree height and diameter at breast height (DBH). Our objective is to 1) develop a linear model relating height to DBH and 2) determine whether project site or genus affects the model.

Methods:  
Following preliminary data analysis, we developed a global linear model using the R function "lm(ripdata[,htcm] ~ ripdata[,dbh]) relating height to DBH. Outliers were then found and removed in order to improve model fit. To further improve fit, we natural log transformed sample values for a second global model. Tertiary linear models for each project site and genus were also created.

Data:  
Measurements of five most frequent tree genera at four project sites in Northern California.

Code:

#Examining subset of riparian data  
str(ripdata\_s)

## 'data.frame': 3240 obs. of 17 variables:  
## $ SurveyID : int 6 6 6 6 6 6 6 6 7 7 ...  
## $ ProjectID : chr "Heritage Oak Winery" "Heritage Oak Winery" "Heritage Oak Winery" "Heritage Oak Winery" ...  
## $ LocationName : chr "RIP01" "RIP01" "RIP01" "RIP01" ...  
## $ Date : chr "3/20/2012" "3/20/2012" "3/20/2012" "3/20/2012" ...  
## $ Collectors : chr "G. Venicx, K. Swan, M. Vaghti, C. Guillen" "G. Venicx, K. Swan, M. Vaghti, C. Guillen" "G. Venicx, K. Swan, M. Vaghti, C. Guillen" "G. Venicx, K. Swan, M. Vaghti, C. Guillen" ...  
## $ Longitude : num -121 -121 -121 -121 -121 ...  
## $ Latitude : num 38.1 38.1 38.1 38.1 38.1 ...  
## $ SurveyTypeID : chr "Plant" "Plant" "Plant" "Plant" ...  
## $ Plot.Name : chr "RIP01" "RIP01" "RIP01" "RIP01" ...  
## $ SpeciesVarietalCode: chr "ACNE" "POFR" "POFR" "POFR" ...  
## $ SpeciesVarietalName: chr "Acer negundo" "Populus fremontii" "Populus fremontii" "Populus fremontii" ...  
## $ Measurement : int 1 2 3 4 5 6 9 21 1 2 ...  
## $ CanopyID : chr "" "" "" "" ...  
## $ Woody\_DBH\_cm : num 6.5 9.3 6.5 7.6 6.3 9.9 7 18.9 6.5 24 ...  
## $ Woody\_Height\_m : num 2.57 5.08 3.74 2.68 3.03 4.53 4.4 8.1 4.61 8 ...  
## $ ProjCode : chr "HOWY" "HOWY" "HOWY" "HOWY" ...  
## $ Genus : chr "Acer" "Populus" "Populus" "Populus" ...

head(ripdata\_s)

## SurveyID ProjectID LocationName Date  
## 1 6 Heritage Oak Winery RIP01 3/20/2012  
## 2 6 Heritage Oak Winery RIP01 3/20/2012  
## 3 6 Heritage Oak Winery RIP01 3/20/2012  
## 4 6 Heritage Oak Winery RIP01 3/20/2012  
## 5 6 Heritage Oak Winery RIP01 3/20/2012  
## 6 6 Heritage Oak Winery RIP01 3/20/2012  
## Collectors Longitude Latitude  
## 1 G. Venicx, K. Swan, M. Vaghti, C. Guillen -121.2 38.15  
## 2 G. Venicx, K. Swan, M. Vaghti, C. Guillen -121.2 38.15  
## 3 G. Venicx, K. Swan, M. Vaghti, C. Guillen -121.2 38.15  
## 4 G. Venicx, K. Swan, M. Vaghti, C. Guillen -121.2 38.15  
## 5 G. Venicx, K. Swan, M. Vaghti, C. Guillen -121.2 38.15  
## 6 G. Venicx, K. Swan, M. Vaghti, C. Guillen -121.2 38.15  
## SurveyTypeID Plot.Name SpeciesVarietalCode SpeciesVarietalName  
## 1 Plant RIP01 ACNE Acer negundo  
## 2 Plant RIP01 POFR Populus fremontii  
## 3 Plant RIP01 POFR Populus fremontii  
## 4 Plant RIP01 POFR Populus fremontii  
## 5 Plant RIP01 POFR Populus fremontii  
## 6 Plant RIP01 POFR Populus fremontii  
## Measurement CanopyID Woody\_DBH\_cm Woody\_Height\_m ProjCode Genus  
## 1 1 6.5 2.57 HOWY Acer  
## 2 2 9.3 5.08 HOWY Populus  
## 3 3 6.5 3.74 HOWY Populus  
## 4 4 7.6 2.68 HOWY Populus  
## 5 5 6.3 3.03 HOWY Populus  
## 6 6 9.9 4.53 HOWY Populus

tail(ripdata\_s)

## SurveyID ProjectID LocationName Date Collectors  
## 4627 857 Cosumnes River Preserve Denier 9/1/2013 RH,DB,AS,CK  
## 4628 857 Cosumnes River Preserve Denier 9/1/2013 RH,DB,AS,CK  
## 4629 857 Cosumnes River Preserve Denier 9/1/2013 RH,DB,AS,CK  
## 4630 857 Cosumnes River Preserve Denier 9/1/2013 RH,DB,AS,CK  
## 4631 857 Cosumnes River Preserve Denier 9/1/2013 RH,DB,AS,CK  
## 4632 857 Cosumnes River Preserve Denier 9/1/2013 RH,DB,AS,CK  
## Longitude Latitude SurveyTypeID Plot.Name SpeciesVarietalCode  
## 4627 -121.38 38.31 Plant 6 FRLA  
## 4628 -121.38 38.31 Plant 6 FRLA  
## 4629 -121.38 38.31 Plant 6 FRLA  
## 4630 -121.38 38.31 Plant 6 FRLA  
## 4631 -121.38 38.31 Plant 6 FRLA  
## 4632 -121.38 38.31 Plant 6 FRLA  
## SpeciesVarietalName Measurement CanopyID Woody\_DBH\_cm Woody\_Height\_m  
## 4627 Fraxinus latifolia 112 111 1.8 1.20  
## 4628 Fraxinus latifolia 113 112 2.7 1.65  
## 4629 Fraxinus latifolia 114 113 1.6 1.41  
## 4630 Fraxinus latifolia 115 114 1.2 0.82  
## 4631 Fraxinus latifolia 116 115 2.4 1.05  
## 4632 Fraxinus latifolia 117 116 3.8 1.35  
## ProjCode Genus  
## 4627 CORP Fraxinus  
## 4628 CORP Fraxinus  
## 4629 CORP Fraxinus  
## 4630 CORP Fraxinus  
## 4631 CORP Fraxinus  
## 4632 CORP Fraxinus

unique(ripdata\_s[,"Genus"])

## [1] "Acer" "Populus" "Fraxinus" "Quercus" "Salix"

unique(ripdata\_s[,"ProjCode"])

## [1] "HOWY" "CORP" "SRRB" "NASO"

#plotting height by project site and genus  
g <- ggplot(data = ripdata\_s)   
  
#Histograms and densities of ht and DBH  
p1 <- g + geom\_boxplot(aes(x = ProjCode, y = Woody\_Height\_m, fill = Genus)) +ggtitle("Distribution of values, Genus by ProjCode")  
p2 <- g + geom\_histogram(aes(x = Woody\_Height\_m))  
p3 <- g + geom\_density(aes(x = Woody\_Height\_m))

#Change height units to DBH units (m to cm)  
ripdata\_s[,"Woody\_Height\_cm"] <- ripdata\_s[,"Woody\_Height\_m"] \*100  
  
g2 <- ggplot(data = ripdata\_s)   
p4 <- g2 + geom\_histogram(aes(x = Woody\_DBH\_cm))  
p5 <-g2 + geom\_density(aes(x = Woody\_DBH\_cm))  
#Force a zero intercept  
incpt <- 0.0  
ripdata\_s\_lm <- lm(I(Woody\_Height\_cm - incpt) ~ 0 + Woody\_DBH\_cm, ripdata\_s)  
summary(ripdata\_s\_lm)

##   
## Call:  
## lm(formula = I(Woody\_Height\_cm - incpt) ~ 0 + Woody\_DBH\_cm, data = ripdata\_s)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6013.6 118.0 310.0 541.4 2106.2   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## Woody\_DBH\_cm 34.482 0.372 92.7 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 576.8 on 3239 degrees of freedom  
## Multiple R-squared: 0.7263, Adjusted R-squared: 0.7262   
## F-statistic: 8594 on 1 and 3239 DF, p-value: < 2.2e-16

#Plot line & observations, sep by genus  
p6 <- g2 + geom\_point(aes(x = Woody\_DBH\_cm, y = Woody\_Height\_cm, color = Genus)) + geom\_abline(intercept = 0, slope = coef(ripdata\_s\_lm)) + ggtitle("Base Model, WH vs DBH")

#Find outliers  
rip\_ol <- outlierTest(ripdata\_s\_lm)  
rip\_ol\_id <- as.numeric(names(rip\_ol$rstudent))  
  
p7 <- g2 + geom\_point(aes(x = Woody\_DBH\_cm, y = Woody\_Height\_cm, color = Genus)) + geom\_abline(intercept = 0, slope = coef(ripdata\_s\_lm)) + geom\_point(data = ripdata\_s[rip\_ol\_id,],aes(x = Woody\_DBH\_cm, y = Woody\_Height\_cm), color = "black", shape = 17) +ggtitle("WH vs DBH, outliers identified")  
  
  
ripdata\_ss <- ripdata\_s[!(rownames(ripdata\_s) %in% rip\_ol\_id) ,]  
  
ripdata\_ss\_lm <- lm(I(Woody\_Height\_cm - incpt) ~ 0 + Woody\_DBH\_cm, ripdata\_ss)  
  
p8 <- ggplot(data = ripdata\_ss) + geom\_point(aes(x = Woody\_DBH\_cm, y = Woody\_Height\_cm, color = Genus))+ geom\_abline(intercept = 0, slope = coef(ripdata\_ss\_lm)) + ggtitle("WH vs. DBH, outliers removed")

ripdata\_s[,"WH\_cm\_log"] <- log10(ripdata\_s[,"Woody\_Height\_cm"])  
ripdata\_s[,"DBH\_cm\_log"] <- log10(ripdata\_s[,"Woody\_DBH\_cm"])  
  
  
g3 <- ggplot(data = ripdata\_s)  
p9 <- g3 + geom\_histogram(aes(x = WH\_cm\_log))  
p10 <- g3 + geom\_density(aes(x = WH\_cm\_log))  
p11 <- g3 + geom\_histogram(aes(x = DBH\_cm\_log))  
p12 <- g3 + geom\_density(aes(x = DBH\_cm\_log))  
  
ripdata\_s\_lm\_log <- lm(I(WH\_cm\_log - incpt) ~ 0 + DBH\_cm\_log, ripdata\_s)  
  
summary(ripdata\_s\_lm\_log)

##   
## Call:  
## lm(formula = I(WH\_cm\_log - incpt) ~ 0 + DBH\_cm\_log, data = ripdata\_s)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.4586 -0.1998 0.2893 0.6598 2.3542   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## DBH\_cm\_log 2.4303 0.0104 233.6 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.6857 on 3239 degrees of freedom  
## Multiple R-squared: 0.944, Adjusted R-squared: 0.944   
## F-statistic: 5.458e+04 on 1 and 3239 DF, p-value: < 2.2e-16

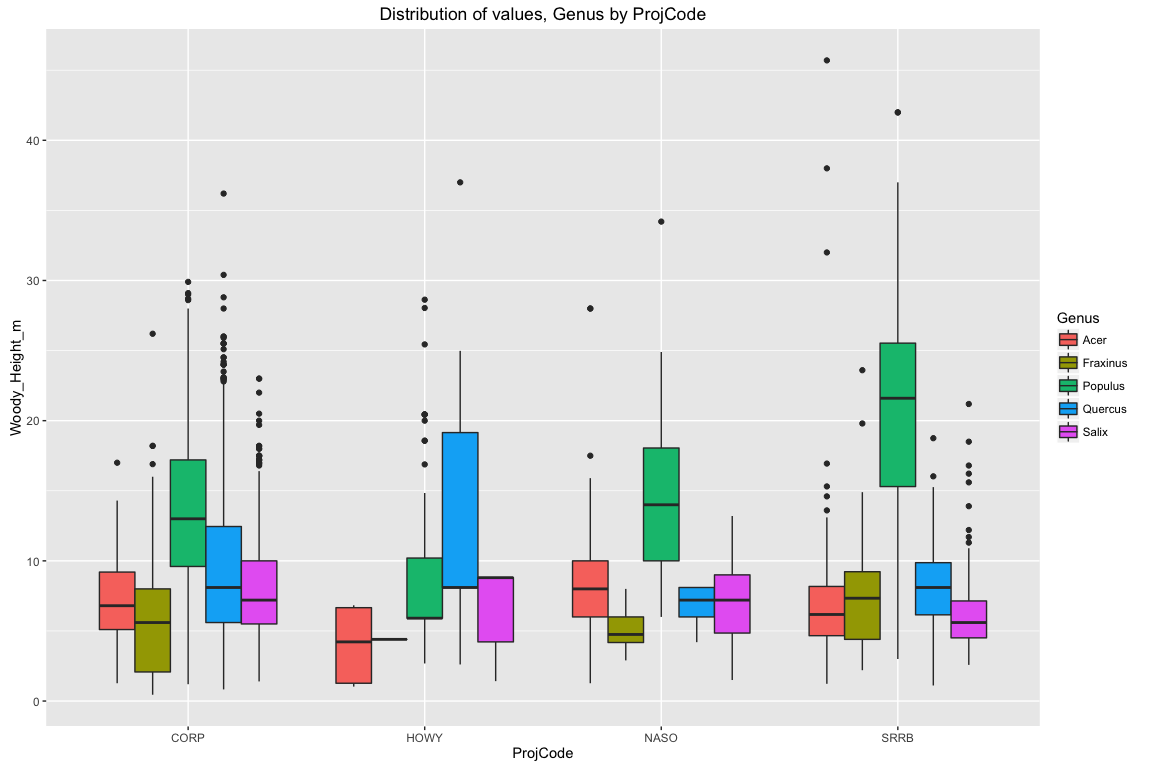
p13 <- g3 + geom\_point(aes(x = DBH\_cm\_log, y = WH\_cm\_log, color = Genus)) + stat\_smooth(formula = I(y - incpt) ~ (0 + x), method = "lm", se = FALSE) + ggtitle("Log10 Transformed WH and DBH")  
  
#Making labels for the graph  
lm\_prj\_gen <- function(xx){  
 lmod <- lm(I(Woody\_Height\_cm - incpt) ~ 0 + Woody\_DBH\_cm, data = xx)  
 r2 <- summary(lmod)$r.squared   
 data.frame(r2 = r2, stringsAsFactors = F)  
}  
  
lm\_labs <- ddply(ripdata\_ss, .(Genus,ProjCode), lm\_prj\_gen)  
  
#graphing genus vs projcode, with lm line and r2 labeled  
p14 <- ggplot(data = ripdata\_ss, aes(x = Woody\_DBH\_cm, y = Woody\_Height\_cm)) + geom\_point() + stat\_smooth(formula = I(y - incpt) ~ (0 + x), method = "lm", se = FALSE) + facet\_grid(ProjCode ~ Genus) + geom\_text(data = lm\_labs, aes(x = 50, y = 7000, label = paste("r2 = ", round(r2, digits = 2), sep = " "))) + ggtitle("Competing Linear models, Genus by Project Site")

p15 <- ggplot(data = ripdata\_ss, aes(x = Woody\_DBH\_cm, y = Woody\_Height\_cm)) + geom\_point(aes(shape = Genus, color = ProjCode)) + stat\_smooth(formula = I(y - incpt) ~ (0 + x), method = "lm", se = FALSE) + coord\_trans(x = "log10", y = "log10") + geom\_text(aes( x = 10, y = 4000, label = paste0("f(x)=", round(coef(lm(I(Woody\_Height\_cm - incpt) ~ (0 + Woody\_DBH\_cm))), digits = 4),"x,R2 =",round(summary(lm(I(Woody\_Height\_cm - incpt) ~ (0 + Woody\_DBH\_cm), data = ripdata\_s))$r.squared, digits = 4)))) + ggtitle("Global fit model, WH vs DBH")

Results:

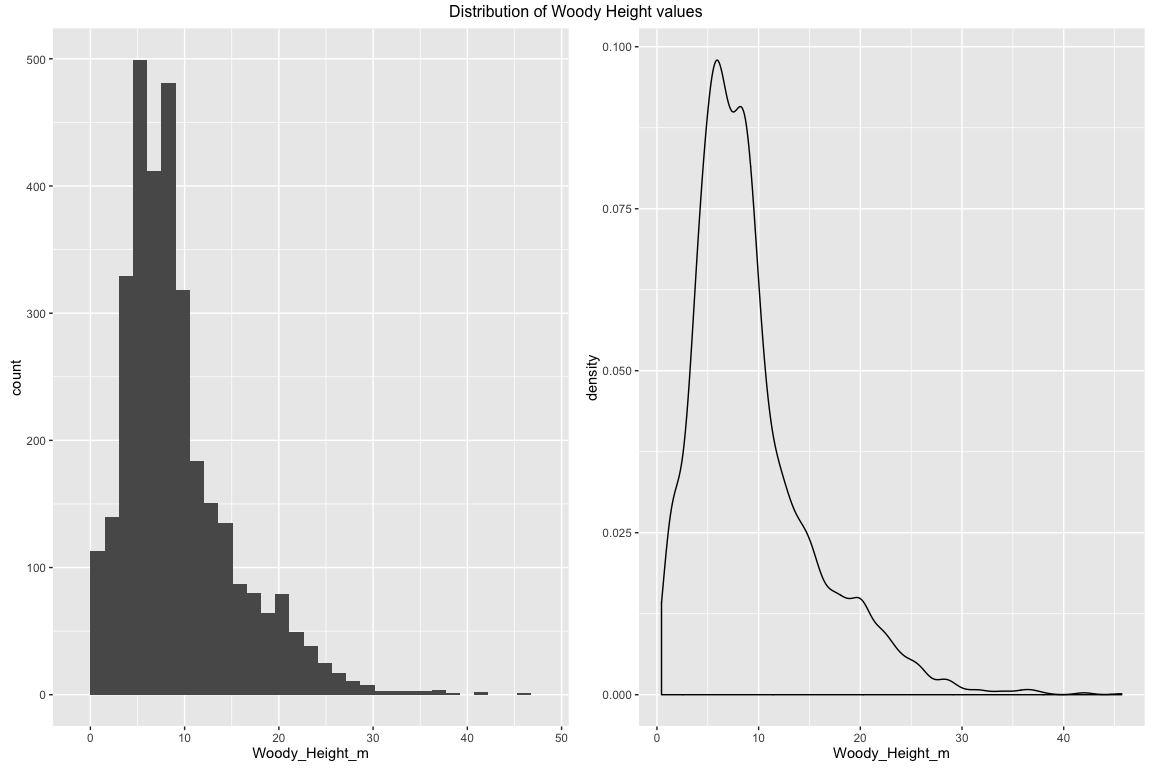
Step 1

p1



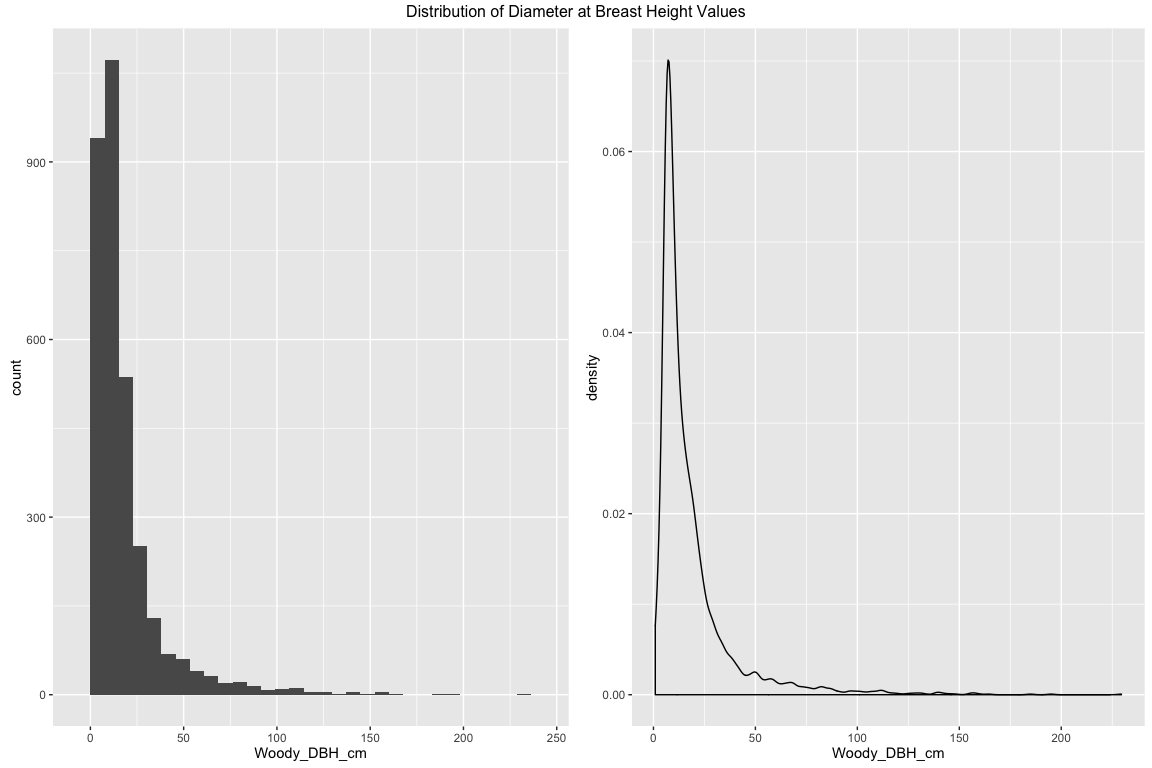
Of the most frequently occuring genera, Populus exhibits both the greatest range of heights and the greatest mean heights in CORP, NASO and SRRB. Genera in CORP have the most outliers and in general the closes means between genera.

grid.arrange(p2,p3, ncol = 2, nrow =1, top = "Distribution of Woody Height values")



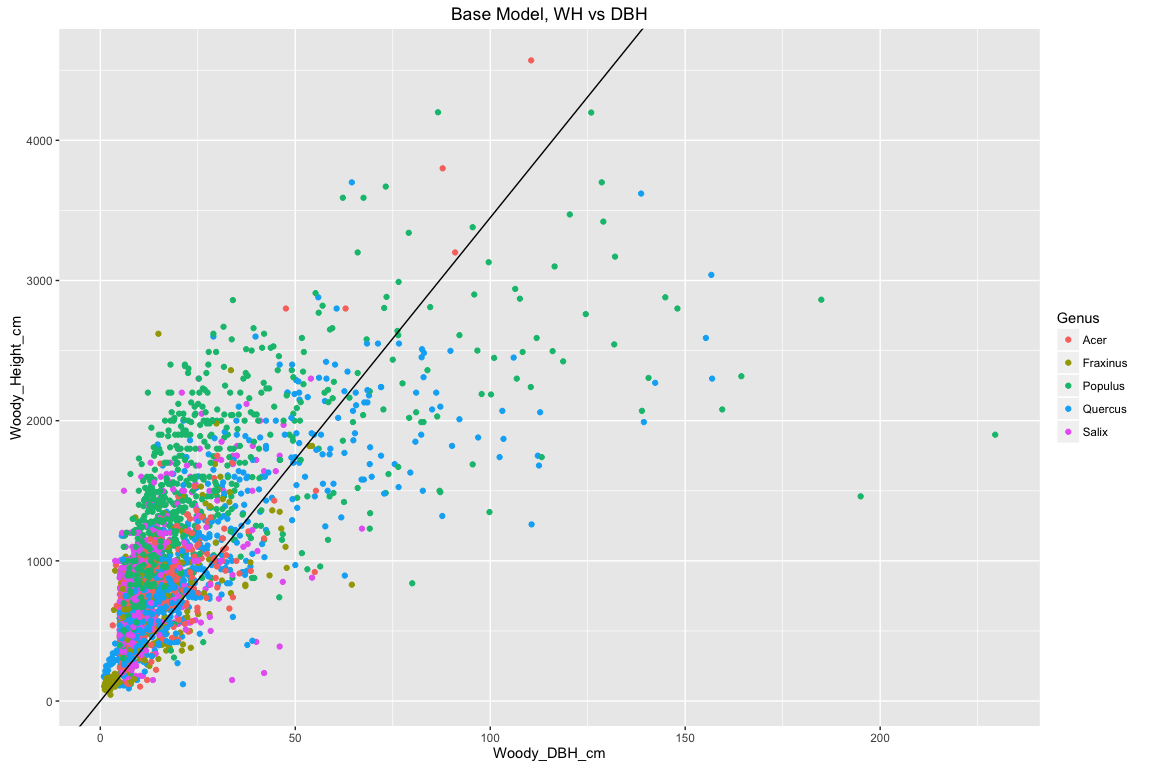
The heights and diameters are not normally distributed and may require transformation.

grid.arrange(p4,p5, ncol = 2, nrow = 1, top = "Distribution of Diameter at Breast Height Values")



Linear Model

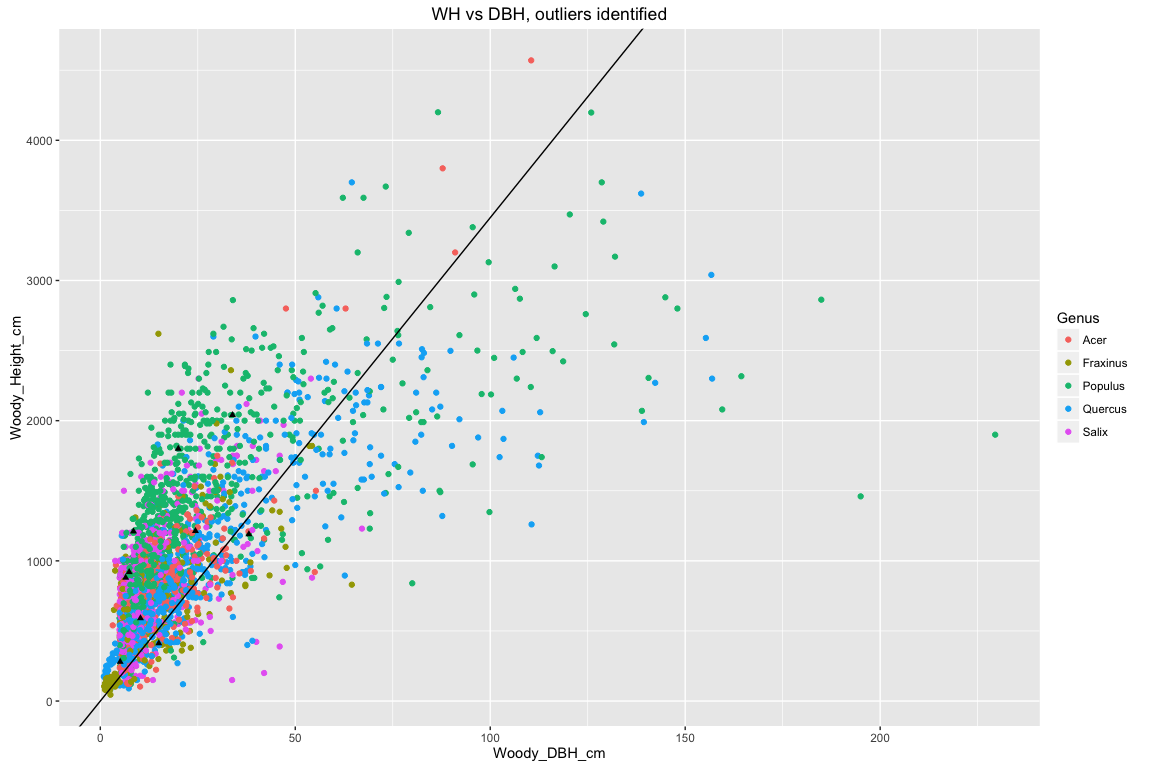
p6



A basic linear model of the relationship between tree height and diameter at breast height explaines approximately 72% of the variation, with a p-value less than 2e-16. This fits the given data, but normalized data may increase the percent variation explained.

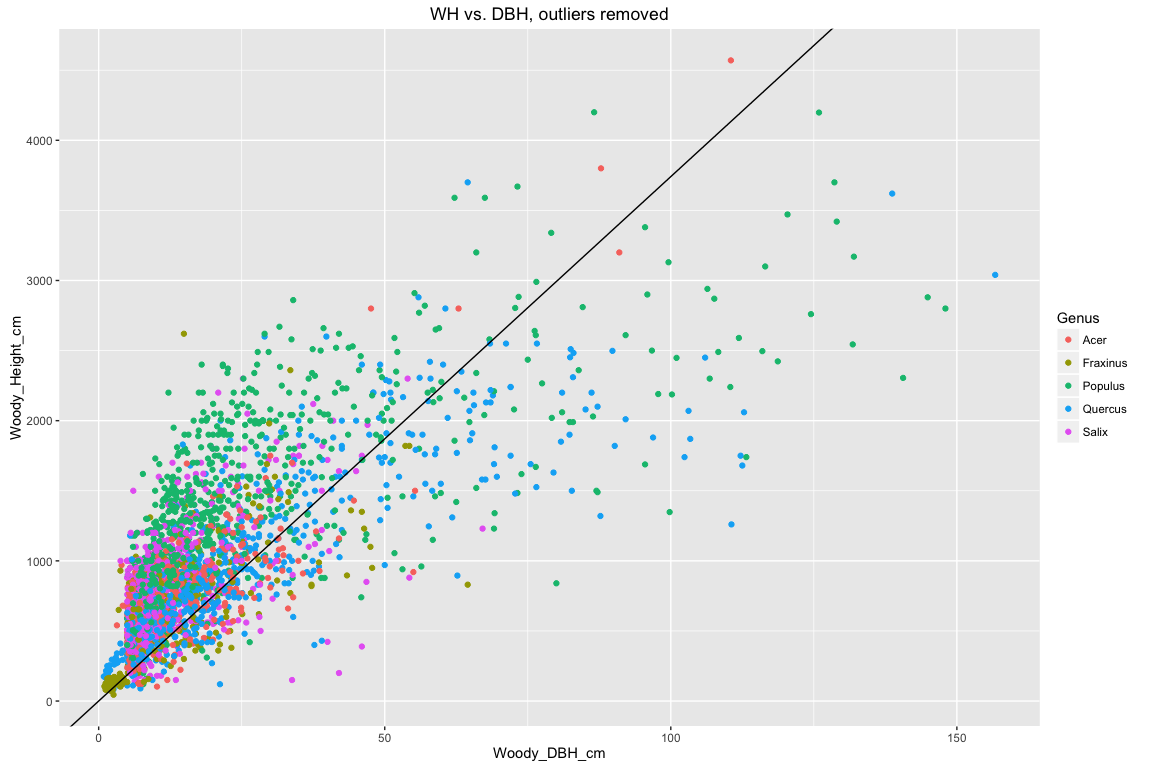
Outliers

p7



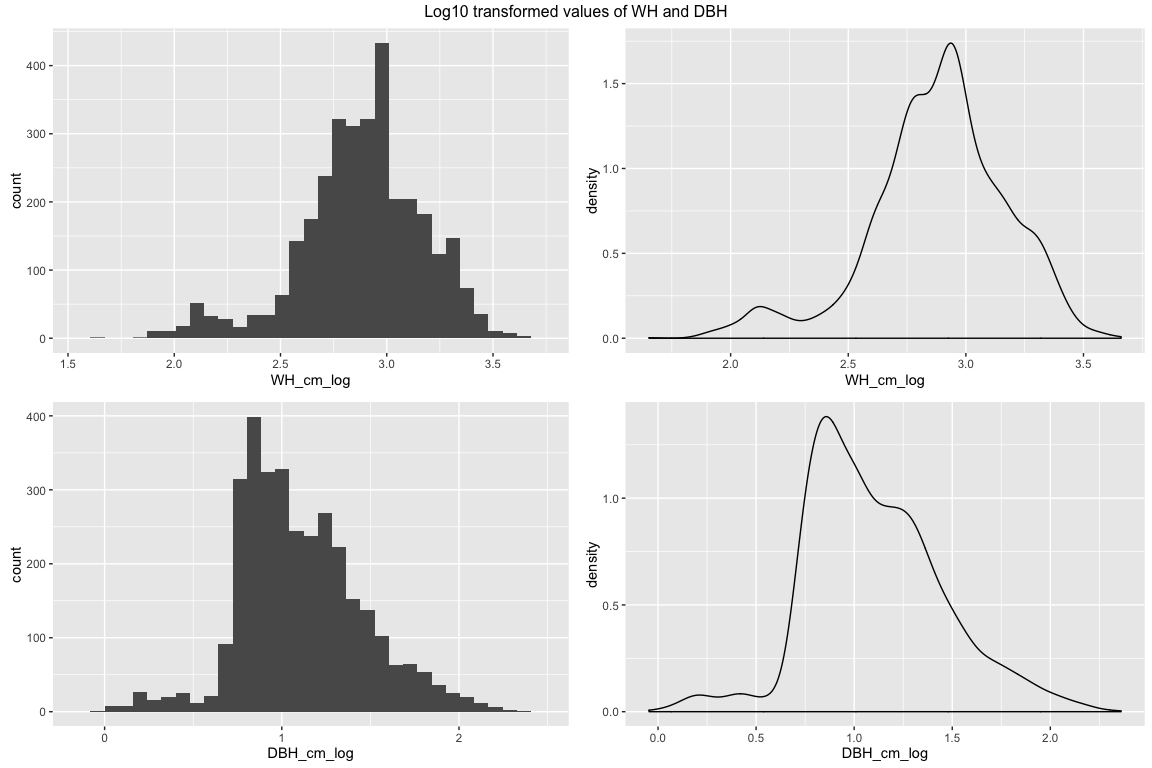
We identified 10 points as outliers. Their removal results in a 4% increase in R2.

p8



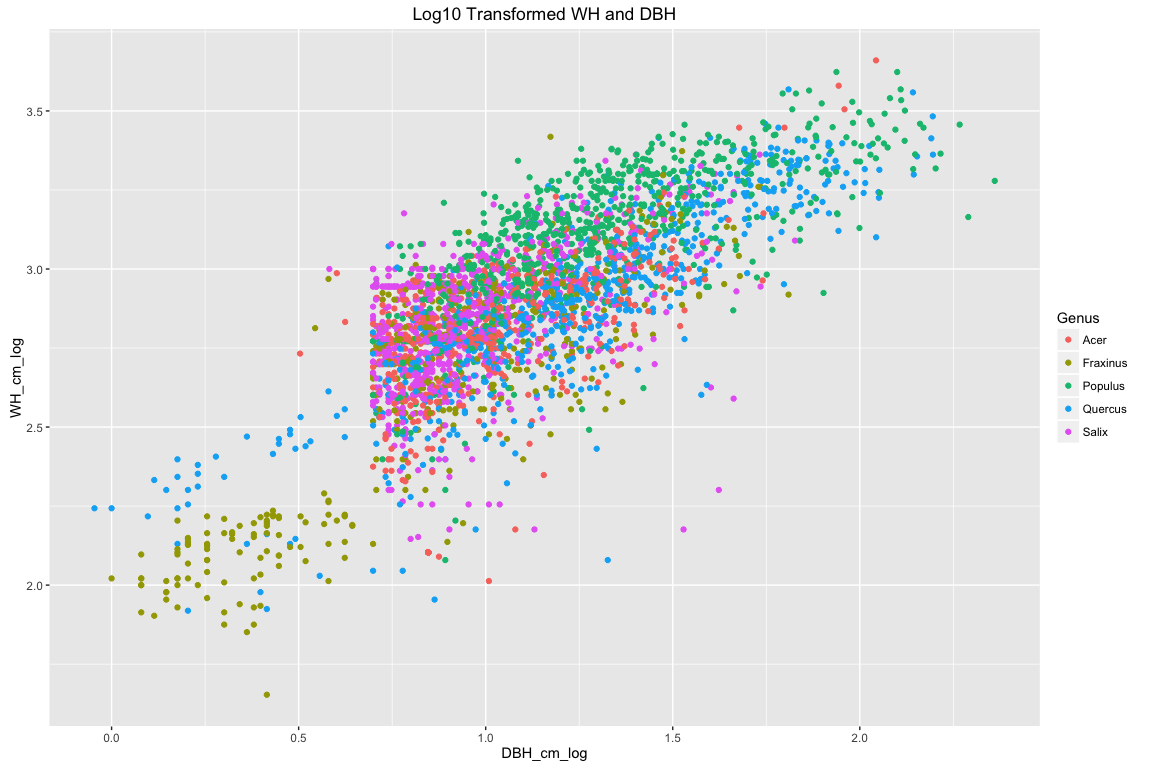
Competing models

grid.arrange(p9,p10,p11,p12, ncol=2,nrow=2, top = "Log10 transformed values of WH and DBH")



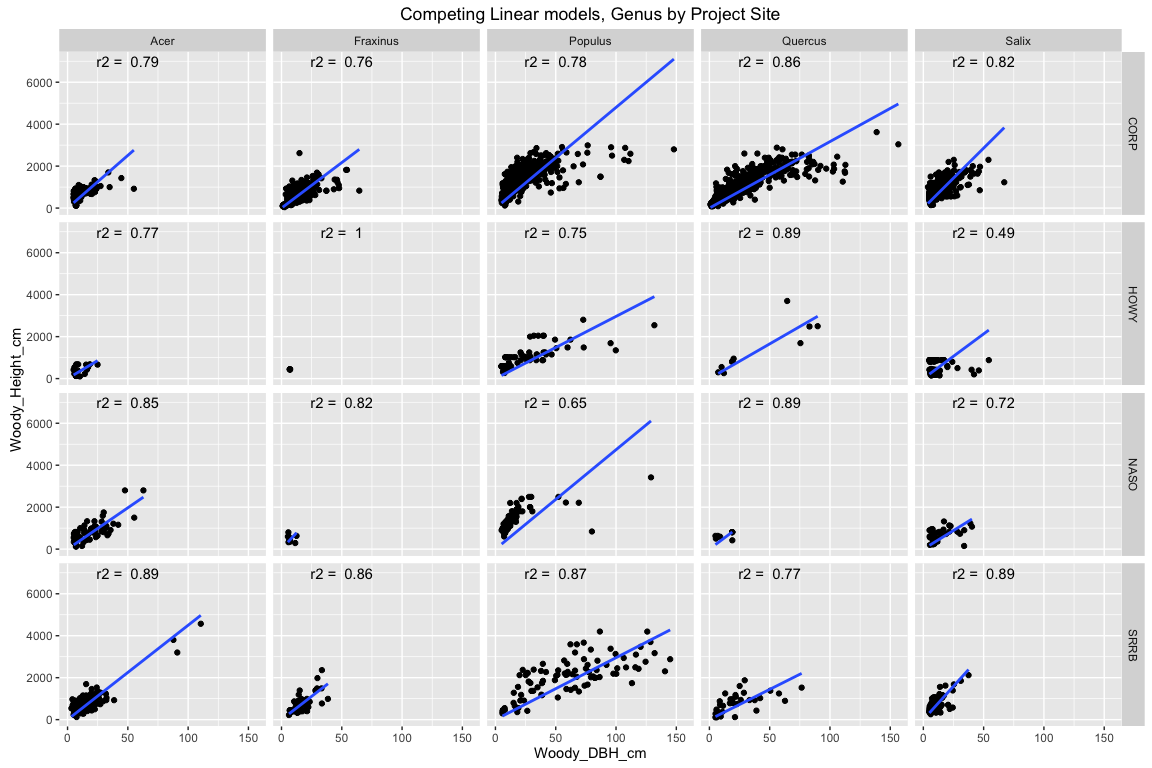
The log tranformed values have approximately normal distributions.

p13



When outliers are removed and values transformed, our explained variance is approximately 94%. We can begin to separate out what combinations of variables is influencing this with the following table:

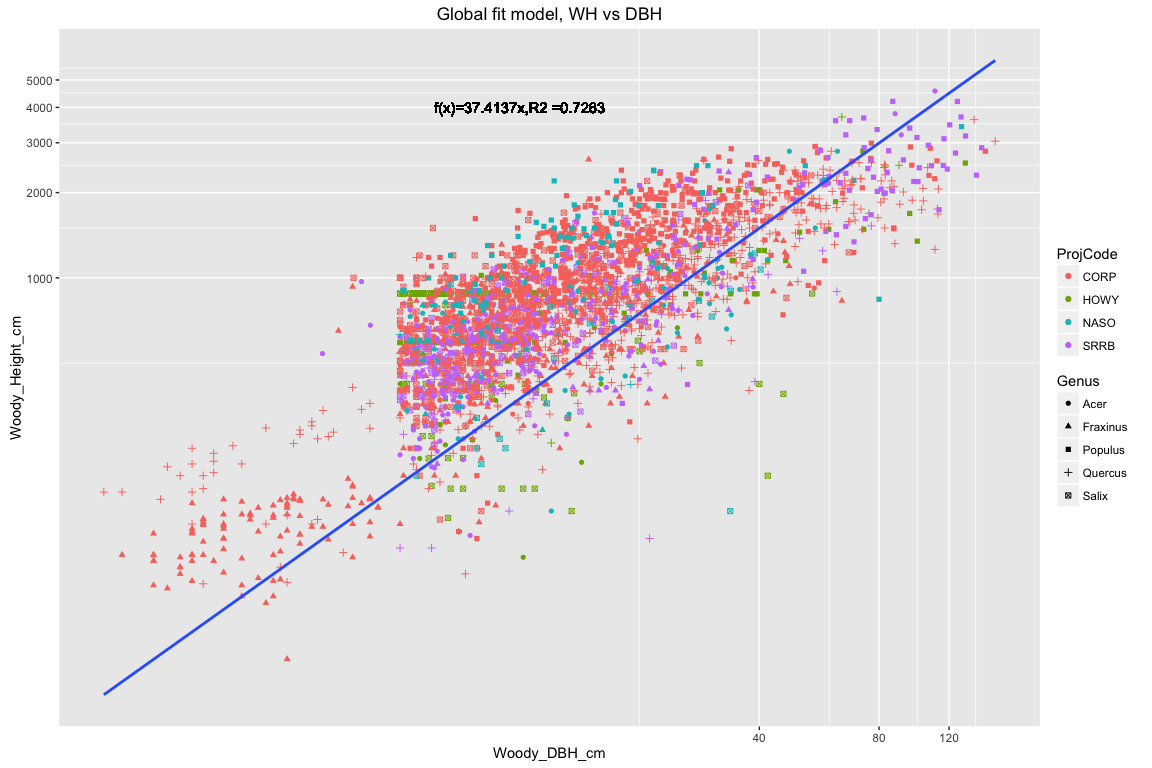
p14



It appears that a model based only on the Acer genus in the SRRB site would be close to our total explained variance; however, I prefer one of the models that shows a greater spread of values, such as Quercus x CORP.

Master Plot

p15

 The master plot includes the best fit model, with log transformed values, separated by genus and project code.

Discussion:  
There appears to be a fairly strong relationship between DBH and height based upon our subset of most frequent genera. In previous data cleansing, we notice data errors that were summarily removed before analysis. It was not noticed then, but can be seen now that there is some amount of bias at the HOWY, NASO and SRRB sites for samples greater than ~ 10cm DBH. This is unfortunate, because though our model explains a high amount of variation in the relationship, we cannot be certain that our model accurately portrays the population.

Limitations:  
Our linear models obscure the relationships between DBH and WH at low values of each. A better fitting model would utilize an exponential function to describe the relationship.