Modelling of DBH and height of Maple and Linden

Oyedayo Oyelowo, student number: 014717208

1 November 2017

Task is to create regression models for diameter and height of linden (lehmus) and maple (vaahtera). Laser scanning-based features are used as explanatory variables. Data: Modeling data (koepuut.xlsx): The file contains 1138 trees from the test area. DBH and species (puulaji) have been determined from every tree. In addition, height is known for about half of the trees. 10 ALS metrics have been calculated for each tree. The data contains several tree species. Trees with no measured DBH and height (mallinnettavat.xlsx): 10 ALS metrics have been calculated for each tree. DBH and height are predicted with models. The data consists of linden and maples.

Fork flow: Creating the models with modelling data - Separate linden and maples from the data (there are also other species in the modelling data).

Choose (by testing) best predictors from ALS features for each model (best features may vary between species).

Predicting variables

- Form your own functions that utilize the created models models' explanatory variables are used as parameters for the functions
- Form a loop structure that runs through all trees in mallinnettavat.txt and calls for the correct functions according to tree species save the predicted diameter (mm) and height (dm) values for example in dbh and h vectors
- Create a result matrix that contains the following columns: tree number (Puunro), tree species (puulaji), dbh in cm and height in m
- Export the matrix into csv file

In this mini-project, I will be performing a multiple regression analysis to predict the dbh and height. I will be using the lm function in R. However, there are other more efficient models that can deal with e.g multucollinearity and homoscedacity better. Such include generalized linear model(GLM), generalised additive model(GAM) and general boosting model(GBM)/BRT. These models give the options to deal with binomial distribution and count data with poisson distribution. This is because, in many cases, we deal with data with non-normally distributed error. Packaages that can be used in r include "mgcv", "gbm", "glm", "dismo" etc. It is also possitble to use higher order polynomial and also look at interactions between variables. GAM also gives the opportunity to see the response curves.

However, for simpplicity, I will be using the lm(linear model) function in R and just the first order polynomials. It is also possible to test the prediction of te model by dividing the data into 70:30 training and testing data or using the leave one out method. AFterwards, correlation can be used to see how related the predictd is to the observed. AUC curves can also be compared by using the wilcox test. To make it simple, I will be doin the prediction alone as requested in this exercise.

```
rm(list = ls())

setwd("C:/Users/oyeda/Desktop/R_COURSE/modelling")

#load the data
data1 <- read.table("koepuut.txt", header = T, sep = "\t")
```

```r
data2 <- read.table("mallinnettavat.txt", header = T, sep = "\t")
## NOTE: linden (lehmus) and maple (vaahtera)
```

The dimension and structure of both dataets

```r
str(data1)
```

```
## 'data.frame':    1138 obs. of  16 variables:
##  $ Puunro : int  356 357 358 359 360 361 362 363 364 365 ...
##  $ X      : num  50844 50837 50832 50827 50824 ...
##  $ Y      : num  22598 22591 22585 22576 22564 ...
##  $ Hmax   : num  8.81 9.25 8.61 7.65 7.88 ...
##  $ Hmean  : num  5.07 5 5.02 4.9 4.76 ...
##  $ h30    : num  4.31 4.13 4.12 4.08 4.02 ...
##  $ h50    : num  4.92 4.88 4.85 4.67 4.57 ...
##  $ h70    : num  5.57 5.64 5.82 5.49 5.28 ...
##  $ h90    : num  6.72 6.64 7.06 6.72 6.37 ...
##  $ p30    : num  0.12 0.3224 0.2273 0.0788 0.1673 ...
##  $ p50    : num  0.528 0.69 0.553 0.436 0.585 ...
##  $ p70    : num  0.825 0.932 0.842 0.695 0.815 ...
##  $ p90    : num  0.937 0.987 0.963 0.877 0.951 ...
##  $ puulaji: Factor w/ 15 levels "jalava","kirsikka",..: 5 5 5 5 5 5 11 11 11 11 ...
##  $ dbh_mm : int  142 138 139 131 134 149 133 176 185 141 ...
##  $ h_dm   : int  86 NA 87 NA 79 NA 78 NA 65 NA ...
```

```r
dim(data1)
```

```
## [1] 1138   16
```

```r
str(data2)
```

```
## 'data.frame':    6181 obs. of  14 variables:
##  $ Puunro : int  134 143 173 175 176 177 178 179 180 181 ...
##  $ X      : num  46872 47008 47515 49787 49790 ...
##  $ Y      : num  22728 22692 22662 22697 22689 ...
##  $ Hmax   : num  12.52 9.55 2.83 9.23 9.12 ...
##  $ Hmean  : num  12.36 8.27 2.6 6.96 6.55 ...
##  $ h30    : num  12.33 8.15 2.55 6.43 6 ...
##  $ h50    : num  12.4 8.2 2.73 7.1 6.51 ...
##  $ h70    : num  12.43 8.27 2.76 7.69 7 ...
##  $ h90    : num  12.49 8.38 2.79 8.21 8.01 ...
##  $ p30    : num  0 0 0 0.00396 0.00756 ...
##  $ p50    : num  0 0 0.25 0.0839 0.1371 ...
##  $ p70    : num  0 0 0.25 0.338 0.674 ...
##  $ p90    : num  0 0 0.375 0.764 0.954 ...
##  $ puulaji: Factor w/ 2 levels "lehmus","vaahtera": 2 2 2 1 1 1 1 1 1 1 ...
```

```r
dim(data2)
```

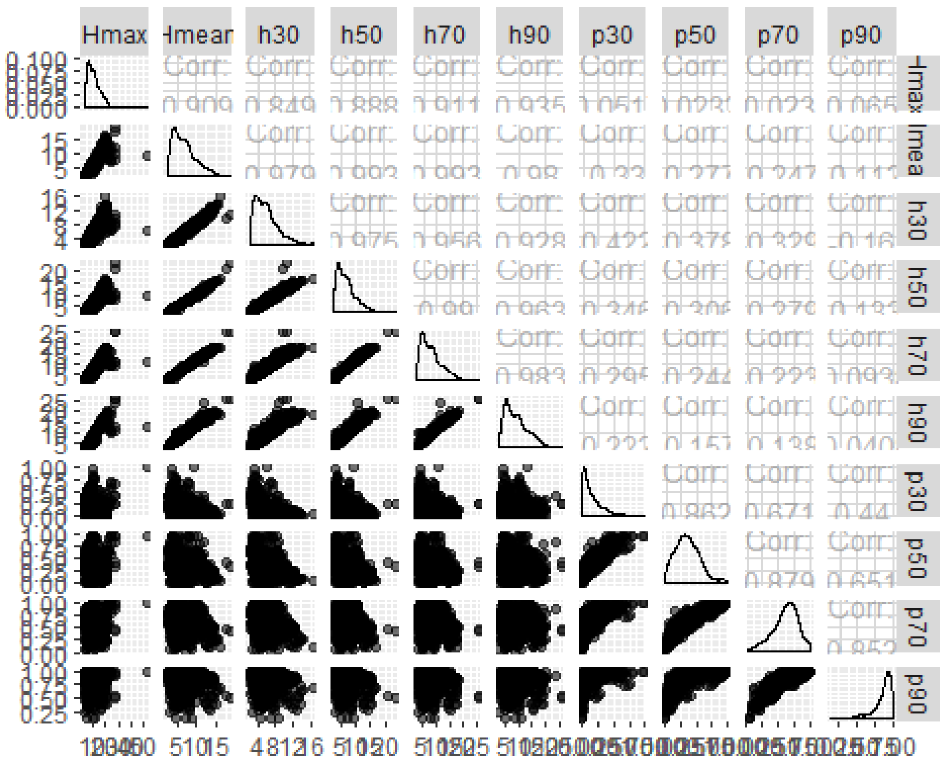
```
## [1] 6181   14
```

Let's see how the predictors are distributed

```r
library(GGally)
library(ggplot2)
d1<-data1[, c("Hmax", "Hmean", "h30", "h50", "h70", "h90", "p30"
              , "p50", "p70", "p90")]
# create a more advanced plot matrix with ggpairs()
```

```r
p <- ggpairs(d1, mapping = aes(alpha=0.3), lower = list(combo = wrap("facethist", bins =
20)))

# draw the plot
p
```



firstly, I have to use the data with some known heights to create the model for linden

```r
linden1 <- data1[data1$puulaji == "lehmus",]
#linear model for dbh of linden
linfit_dbh <- lm(dbh_mm ~ Hmax + Hmean + h30 + h50 + h70 + h90 + p30
                 + p50 + p70 + p90, linden1)
summary(linfit_dbh)

##
## Call:
## lm(formula = dbh_mm ~ Hmax + Hmean + h30 + h50 + h70 + h90 +
##     p30 + p50 + p70 + p90, data = linden1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -292.65  -29.34   -3.79   25.34  305.10
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -70.199     21.831  -3.216 0.001378 **
## Hmax            6.818      3.268   2.086 0.037429 *
## Hmean          64.526     27.326   2.361 0.018555 *
## h30           -31.825     12.282  -2.591 0.009817 **
## h50            -3.033     11.354  -0.267 0.789448
## h70           -23.176     10.823  -2.141 0.032687 *
## h90            12.948      7.800   1.660 0.097488 .
## p30            46.208     39.959   1.156 0.248025
```

```
## p50         -195.585     46.578  -4.199 3.12e-05 ***
## p70           18.229     55.552   0.328 0.742926    
## p90          158.242     46.969   3.369 0.000807 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 58.55 on 552 degrees of freedom
## Multiple R-squared:  0.8024, Adjusted R-squared:  0.7988 
## F-statistic: 224.2 on 10 and 552 DF,  p-value: < 2.2e-16
```

from the above, we can take away h50, p30, and p70, because, *they all have p values above 0.05* thus, i'm left with *dbh_mm ~ Hmax + Hmean + h30 + h70 + p50 + h90 + p90*

To further corroborate this, I used a stepwise regression next

```
# Stepwise Regression
library(MASS)
linfit_dbh <- lm(dbh_mm ~ Hmax + Hmean + h30 + h50 + h70 + h90 + p30
                 + p50 + p70 + p90, linden1)
#?stepAIC
step <- stepAIC(linfit_dbh, direction = "both")
```

```
## Start:  AIC=4593.59
## dbh_mm ~ Hmax + Hmean + h30 + h50 + h70 + h90 + p30 + p50 + p70 + 
##     p90
## 
##         Df Sum of Sq     RSS    AIC
## - h50    1       245 1892613 4591.7
## - p70    1       369 1892738 4591.7
## - p30    1      4584 1896953 4593.0
## <none>               1892368 4593.6
## - h90    1      9446 1901815 4594.4
## - Hmax   1     14919 1907287 4596.0
## - h70    1     15719 1908087 4596.3
## - Hmean  1     19115 1911484 4597.3
## - h30    1     23018 1915387 4598.4
## - p90    1     38912 1931281 4603.1
## - p50    1     60447 1952815 4609.3
## 
## Step:  AIC=4591.67
## dbh_mm ~ Hmax + Hmean + h30 + h70 + h90 + p30 + p50 + p70 + p90
## 
##         Df Sum of Sq     RSS    AIC
## - p70    1       436 1893050 4589.8
## - p30    1      4364 1896977 4591.0
## <none>               1892613 4591.7
## - h90    1     11043 1903656 4592.9
## + h50    1       245 1892368 4593.6
## - Hmax   1     15483 1908096 4594.3
## - h70    1     17339 1909952 4594.8
## - Hmean  1     21290 1913903 4596.0
## - h30    1     23304 1915917 4596.6
## - p90    1     38770 1931383 4601.1
## - p50    1     60302 1952915 4607.3
## 
## Step:  AIC=4589.8
```

```
## dbh_mm ~ Hmax + Hmean + h30 + h70 + h90 + p30 + p50 + p90
##
##          Df Sum of Sq     RSS    AIC
## - p30    1      3995 1897044 4589.0
## <none>               1893050 4589.8
## - h90    1     12371 1905421 4591.5
## + p70    1       436 1892613 4591.7
## + h50    1       312 1892738 4591.7
## - h70    1     19327 1912377 4593.5
## - Hmax   1     19588 1912638 4593.6
## - Hmean  1     20913 1913962 4594.0
## - h30    1     22937 1915987 4594.6
## - p50    1     78604 1971654 4610.7
## - p90    1    122382 2015432 4623.1
##
## Step:  AIC=4588.98
## dbh_mm ~ Hmax + Hmean + h30 + h70 + h90 + p50 + p90
##
##          Df Sum of Sq     RSS    AIC
## <none>               1897044 4589.0
## + p30    1      3995 1893050 4589.8
## - h90    1     12304 1909348 4590.6
## + p70    1        67 1896977 4591.0
## + h50    1        36 1897008 4591.0
## - h70    1     16364 1913408 4591.8
## - Hmax   1     17685 1914729 4592.2
## - Hmean  1     19105 1916149 4592.6
## - h30    1     22130 1919175 4593.5
## - p90    1    118980 2016024 4621.2
## - p50    1    119950 2016994 4621.5
```

```
step$anova # display results
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## dbh_mm ~ Hmax + Hmean + h30 + h50 + h70 + h90 + p30 + p50 + p70 +
##     p90
##
## Final Model:
## dbh_mm ~ Hmax + Hmean + h30 + h70 + h90 + p50 + p90
##
##
##    Step Df  Deviance Resid. Df Resid. Dev      AIC
## 1                          552    1892368 4593.594
## 2 - h50  1  244.6803       553    1892613 4591.667
## 3 - p70  1  436.4771       554    1893050 4589.796
## 4 - p30  1 3994.8186       555    1897044 4588.983
```

The result of the analaysis confirms earlier the choice made earlier. Thus, my final model for dbh for linden would be: $dbh_m m = a + b_1 Hmax + b_2 Hmean + b_4 h30 + b_5 h70 + b_6 p50 + b_7 h90 + b_8 p90$

Next is for the height

```
linden1 <- data1[data1$puulaji == "lehmus",]
#linear model for dbh of linden
```

```r
linfit_h <- lm(h_dm ~ Hmax + Hmean + h30 + h50 + h70 + h90 + p30
               + p50 + p70 + p90, linden1)
summary(linfit_h)
```

```
## 
## Call:
## lm(formula = h_dm ~ Hmax + Hmean + h30 + h50 + h70 + h90 + p30 + 
##     p50 + p70 + p90, data = linden1)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -88.831  -5.673   1.032   7.379  27.274 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)   
## (Intercept)    5.192      8.298   0.626  0.53223   
## Hmax           3.693      1.233   2.994  0.00307 **
## Hmean        -12.727     12.139  -1.048  0.29559   
## h30            9.596      5.316   1.805  0.07242 . 
## h50           -9.270      4.706  -1.970  0.05012 . 
## h70           13.175      4.720   2.791  0.00572 **
## h90            4.222      3.102   1.361  0.17489   
## p30            4.994     14.336   0.348  0.72790   
## p50          -29.220     17.123  -1.707  0.08935 . 
## p70           22.127     21.603   1.024  0.30686   
## p90           14.675     18.186   0.807  0.42060   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 14.51 on 216 degrees of freedom
##   (336 observations deleted due to missingness)
## Multiple R-squared:  0.8759, Adjusted R-squared:  0.8702 
## F-statistic: 152.5 on 10 and 216 DF,  p-value: < 2.2e-16
```

from the above, I can eliminate Hmean, h50, h90, p30, p50, p70 and p90. Model can then be: $h_d m = a + b_1 Hmax + b_2 h30 + b_3 h70$

*Stepwise regression for height of linden

```r
#next, use stepwise regression to eliminate the redundant variables:
step_h <- stepAIC(linfit_h, direction = ("both"))
```

```
## Start:  AIC=1225.26
## h_dm ~ Hmax + Hmean + h30 + h50 + h70 + h90 + p30 + p50 + p70 + 
##     p90
## 
##         Df Sum of Sq   RSS    AIC
## - p30    1     25.57 45533 1223.4
## - p90    1    137.18 45645 1223.9
## - p70    1    221.03 45729 1224.4
## - Hmean  1    231.61 45739 1224.4
## - h90    1    390.34 45898 1225.2
## <none>               45508 1225.3
## - p50    1    613.55 46121 1226.3
## - h30    1    686.63 46194 1226.7
## - h50    1    817.58 46325 1227.3
```

```
## - h70    1   1641.67 47150 1231.3
## - Hmax   1   1888.66 47397 1232.5
##
## Step:  AIC=1223.38
## h_dm ~ Hmax + Hmean + h30 + h50 + h70 + h90 + p50 + p70 + p90
##
##         Df Sum of Sq   RSS    AIC
## - p90    1    138.33 45672 1222.1
## - p70    1    201.39 45735 1222.4
## - Hmean  1    265.40 45799 1222.7
## - h90    1    401.54 45935 1223.4
## <none>               45533 1223.4
## - h30    1    707.02 46240 1224.9
## + p30    1     25.57 45508 1225.3
## - h50    1    794.42 46328 1225.3
## - p50    1    864.75 46398 1225.7
## - h70    1   1730.34 47264 1229.8
## - Hmax   1   1899.14 47433 1230.7
##
## Step:  AIC=1222.07
## h_dm ~ Hmax + Hmean + h30 + h50 + h70 + h90 + p50 + p70
##
##         Df Sum of Sq   RSS    AIC
## - Hmean  1    283.46 45955 1221.5
## - h90    1    342.99 46015 1221.8
## <none>               45672 1222.1
## + p90    1    138.33 45533 1223.4
## - h30    1    693.68 46365 1223.5
## - h50    1    718.62 46390 1223.6
## + p30    1     26.72 45645 1223.9
## - p50    1   1124.71 46796 1225.6
## - Hmax   1   1760.83 47433 1228.7
## - p70    1   1821.60 47493 1229.0
## - h70    1   2056.69 47728 1230.1
##
## Step:  AIC=1221.48
## h_dm ~ Hmax + h30 + h50 + h70 + h90 + p50 + p70
##
##         Df Sum of Sq   RSS    AIC
## - h90    1    107.23 46062 1220.0
## <none>               45955 1221.5
## - h30    1    469.95 46425 1221.8
## + Hmean  1    283.46 45672 1222.1
## + p90    1    156.38 45799 1222.7
## + p30    1     62.65 45893 1223.2
## - p50    1    954.70 46910 1224.1
## - h50    1   1463.41 47419 1226.6
## - Hmax   1   1479.51 47435 1226.7
## - p70    1   1765.41 47721 1228.0
## - h70    1   1855.17 47810 1228.5
##
## Step:  AIC=1220.01
## h_dm ~ Hmax + h30 + h50 + h70 + p50 + p70
##
##         Df Sum of Sq   RSS    AIC
```

```
## <none>               46062 1220.0
## - h30    1     438.9 46501 1220.2
## + h90    1     107.2 45955 1221.5
## + p90    1      99.3 45963 1221.5
## + p30    1      51.7 46011 1221.8
## + Hmean  1      47.7 46015 1221.8
## - p50    1    1010.2 47073 1222.9
## - h50    1    1472.8 47535 1225.2
## - p70    1    1942.9 48005 1227.4
## - Hmax   1    2207.3 48270 1228.6
## - h70    1    4046.5 50109 1237.1
```

```
step_h$anova #display results
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## h_dm ~ Hmax + Hmean + h30 + h50 + h70 + h90 + p30 + p50 + p70 +
##     p90
##
## Final Model:
## h_dm ~ Hmax + h30 + h50 + h70 + p50 + p70
##
##
##      Step Df  Deviance Resid. Df Resid. Dev      AIC
## 1                            216   45507.84 1225.257
## 2   - p30  1  25.56837       217   45533.41 1223.384
## 3   - p90  1 138.32822       218   45671.74 1222.073
## 4 - Hmean  1 283.46075       219   45955.20 1221.477
## 5   - h90  1 107.22846       220   46062.43 1220.006
```

final model: $h_dm = Hmax + h30 + h50 + h70 + p50 + p70$

multiple regression analysis and stepwise regression for Diameter At Breast Height of *Maple*

```
#subset the dataframe to vaahtera species
maple1 <- data1[data1$puulaji == "vaahtera", ]

#multiple linear regression, using all the variables
mapfit_dbh <- lm(dbh_mm ~ Hmax + Hmean + h30 + h50 + h70 + h90 + p30
                 + p50 + p70 + p90, maple1)
#get the summary details
summary(mapfit_dbh)
```

```
##
## Call:
## lm(formula = dbh_mm ~ Hmax + Hmean + h30 + h50 + h70 + h90 +
##     p30 + p50 + p70 + p90, data = maple1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -109.381  -22.123   -8.636   19.235  257.898
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -75.3943    44.0312  -1.712   0.0920 .
## Hmax           1.6715     5.8207   0.287   0.7750
## Hmean        -45.3215    62.3283  -0.727   0.4700
## h30           54.2880    22.3993   2.424   0.0184 *
## h50           10.4815    21.7662   0.482   0.6319
## h70           11.6399    21.3723   0.545   0.5880
## h90           -0.6115    10.1094  -0.060   0.9520
## p30          -94.5943   135.6017  -0.698   0.4881
## p50          116.8221   116.4486   1.003   0.3198
## p70         -166.7410   119.9211  -1.390   0.1695
## p90          192.9563    74.3820   2.594   0.0119 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 55.9 on 60 degrees of freedom
## Multiple R-squared:  0.6733, Adjusted R-squared:  0.6188
## F-statistic: 12.36 on 10 and 60 DF,  p-value: 2.707e-11
```

```
#Perform a stepwise regression to remove the redundant variables
m_step_dbh <- stepAIC(mapfit_dbh, direction = ("both"))
```

```
## Start:  AIC=581.39
## dbh_mm ~ Hmax + Hmean + h30 + h50 + h70 + h90 + p30 + p50 + p70 +
##     p90
##
##         Df Sum of Sq    RSS    AIC
## - h90    1      11.4 187484 579.39
## - Hmax   1     257.7 187730 579.49
## - h50    1     724.6 188197 579.66
## - h70    1     926.8 188399 579.74
## - p30    1    1520.5 188993 579.96
## - Hmean  1    1652.1 189125 580.01
## - p50    1    3144.6 190617 580.57
## <none>                187473 581.39
## - p70    1    6040.6 193513 581.64
## - h30    1   18353.8 205826 586.02
## - p90    1   21026.6 208499 586.94
##
## Step:  AIC=579.39
## dbh_mm ~ Hmax + Hmean + h30 + h50 + h70 + p30 + p50 + p70 + p90
##
##         Df Sum of Sq    RSS    AIC
## - Hmax   1       302 187786 577.51
## - h50    1      1088 188572 577.80
## - h70    1      1231 188715 577.86
## - p30    1      1738 189222 578.05
## - p50    1      3174 190658 578.58
## <none>                187484 579.39
## - p70    1      6070 193555 579.66
## - Hmean  1      6766 194250 579.91
## + h90    1        11 187473 581.39
## - p90    1     21034 208518 584.94
## - h30    1     50811 238295 594.42
##
## Step:  AIC=577.51
```

```
## dbh_mm ~ Hmean + h30 + h50 + h70 + p30 + p50 + p70 + p90
##
##         Df Sum of Sq    RSS    AIC
## - h50    1       869 188655 575.83
## - h70    1      1077 188863 575.91
## - p30    1      1490 189276 576.07
## - p50    1      4088 191874 577.04
## <none>               187786 577.51
## - p70    1      5871 193657 577.69
## - Hmean  1      7891 195677 578.43
## + Hmax   1       302 187484 579.39
## + h90    1        56 187730 579.49
## - p90    1     21209 208996 583.10
## - h30    1     50931 238717 592.55
##
## Step:  AIC=575.83
## dbh_mm ~ Hmean + h30 + h70 + p30 + p50 + p70 + p90
##
##         Df Sum of Sq    RSS    AIC
## - p30    1      1238 189893 574.30
## - p50    1      4045 192700 575.34
## - h70    1      5135 193790 575.74
## <none>               188655 575.83
## - p70    1      7121 195777 576.47
## - Hmean  1      7383 196039 576.56
## + h50    1       869 187786 577.51
## + h90    1       404 188251 577.68
## + Hmax   1        83 188572 577.80
## - p90    1     22211 210866 581.74
## - h30    1     52506 241162 591.27
##
## Step:  AIC=574.3
## dbh_mm ~ Hmean + h30 + h70 + p50 + p70 + p90
##
##         Df Sum of Sq    RSS    AIC
## - p50    1      3089 192982 573.44
## - h70    1      4171 194065 573.84
## <none>               189893 574.30
## - Hmean  1      6301 196195 574.62
## - p70    1      6459 196353 574.67
## + p30    1      1238 188655 575.83
## + h50    1       617 189276 576.07
## + h90    1       564 189329 576.09
## + Hmax   1       122 189771 576.25
## - p90    1     24353 214247 580.87
## - h30    1     51795 241689 589.42
##
## Step:  AIC=573.44
## dbh_mm ~ Hmean + h30 + h70 + p70 + p90
##
##         Df Sum of Sq    RSS    AIC
## - h70    1      3379 196362 572.68
## - p70    1      4344 197327 573.03
## - Hmean  1      5255 198237 573.35
## <none>               192982 573.44
```

```
## + p50    1      3089 189893 574.30
## + h50    1       972 192010 575.09
## + Hmax   1       893 192089 575.12
## + h90    1       675 192307 575.20
## + p30    1       282 192700 575.34
## - p90    1     21724 214706 579.02
## - h30    1     50677 243659 588.00
##
## Step:  AIC=572.68
## dbh_mm ~ Hmean + h30 + p70 + p90
##
##          Df Sum of Sq    RSS    AIC
## - Hmean  1      4655 201017 572.34
## <none>                196362 572.68
## + h50    1      3948 192414 573.24
## - p70    1      7703 204064 573.41
## + h90    1      3424 192938 573.43
## + h70    1      3379 192982 573.44
## + p50    1      2297 194065 573.84
## + p30    1       504 195858 574.49
## + Hmax   1       202 196160 574.60
## - p90    1     25231 221593 579.26
## - h30    1     87926 284287 596.95
##
## Step:  AIC=572.34
## dbh_mm ~ h30 + p70 + p90
##
##          Df Sum of Sq    RSS    AIC
## + h90    1      7600 193417 571.60
## <none>                201017 572.34
## + Hmean  1      4655 196362 572.68
## + h70    1      2780 198237 573.35
## + h50    1      2159 198858 573.57
## + p50    1      1874 199143 573.68
## + Hmax   1       686 200330 574.10
## + p30    1       441 200576 574.18
## - p70    1     15680 216696 575.67
## - p90    1     36596 237612 582.22
## - h30    1    336023 537040 640.11
##
## Step:  AIC=571.6
## dbh_mm ~ h30 + p70 + p90 + h90
##
##          Df Sum of Sq    RSS    AIC
## - p70    1      4738 198155 571.32
## <none>                193417 571.60
## - h90    1      7600 201017 572.34
## + p50    1      2578 190839 572.65
## + h50    1       939 192477 573.26
## + h70    1       793 192624 573.31
## + Hmax   1       686 192731 573.35
## + Hmean  1       479 192938 573.43
## + p30    1       452 192964 573.44
## - p90    1     22894 216311 577.55
## - h30    1    157000 350417 611.80
```

```
##
## Step:  AIC=571.32
## dbh_mm ~ h30 + p90 + h90
##
##         Df Sum of Sq    RSS    AIC
## <none>               198155 571.32
## + p70    1      4738 193417 571.60
## + h50    1      1778 196377 572.68
## + Hmax   1      1505 196650 572.78
## + p30    1      1396 196759 572.82
## + h70    1      1370 196785 572.83
## + p50    1      1333 196822 572.84
## + Hmean  1       809 197346 573.03
## - h90    1     18541 216696 575.67
## - p90    1     28738 226893 578.94
## - h30    1    228507 426662 623.78
```

```
m_step_dbh$anova #display results
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## dbh_mm ~ Hmax + Hmean + h30 + h50 + h70 + h90 + p30 + p50 + p70 +
##     p90
##
## Final Model:
## dbh_mm ~ h30 + p90 + h90
##
##
##       Step Df   Deviance Resid. Df Resid. Dev      AIC
## 1                                 60   187472.6 581.3883
## 2    - h90  1   11.43085        61   187484.1 579.3926
## 3   - Hmax  1  302.20328        62   187786.3 577.5070
## 4    - h50  1  869.04424        63   188655.3 575.8348
## 5    - p30  1 1238.14149        64   189893.4 574.2992
## 6    - p50  1 3088.65800        65   192982.1 573.4448
## 7    - h70  1 3379.43011        66   196361.5 572.6773
## 8  - Hmean  1 4655.13389        67   201016.7 572.3409
## 9    + h90  1 7599.78245        66   193416.9 571.6046
## 10   - p70  1 4738.08029        67   198155.0 571.3229
```

Final Model based on the chosen variables by stepwise regression: $dbh_m m = h30 + p90 + h90$

linear model and Stepwise regression for height of maple trees

```
#Linear model using all the variables
mapfit_h <- lm(h_dm ~ Hmax + Hmean + h30 + h50 + h70 + h90 + p30
               + p50 + p70 + p90, maple1)
summary(mapfit_h)
```

```
##
## Call:
## lm(formula = h_dm ~ Hmax + Hmean + h30 + h50 + h70 + h90 + p30 +
##     p50 + p70 + p90, data = maple1)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -30.644  -8.948   1.248   8.017  24.733 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)  
## (Intercept)   32.953     18.310   1.800   0.0845 .
## Hmax           1.531      2.820   0.543   0.5922  
## Hmean        -34.497     34.361  -1.004   0.3254  
## h30           26.208     11.809   2.219   0.0362 *
## h50           -9.848      8.264  -1.192   0.2450  
## h70           21.393     15.222   1.405   0.1727  
## h90            5.025      5.211   0.964   0.3446  
## p30            9.811     67.419   0.146   0.8855  
## p50          -60.450     52.854  -1.144   0.2640  
## p70           33.589     51.193   0.656   0.5180  
## p90          -10.524     28.321  -0.372   0.7134  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 16.23 on 24 degrees of freedom
##   (36 observations deleted due to missingness)
## Multiple R-squared:  0.841,  Adjusted R-squared:  0.7747 
## F-statistic: 12.69 on 10 and 24 DF,  p-value: 2.527e-07
```

```
#Stepwise regression
m_step_h <- stepAIC(mapfit_h, direction = ("both"))
```

```
## Start:  AIC=203.86
## h_dm ~ Hmax + Hmean + h30 + h50 + h70 + h90 + p30 + p50 + p70 + 
##     p90
## 
##         Df Sum of Sq    RSS    AIC
## - p30    1      5.58 6324.2 201.89
## - p90    1     36.36 6355.0 202.06
## - Hmax   1     77.62 6396.3 202.28
## - p70    1    113.34 6432.0 202.48
## - h90    1    244.74 6563.4 203.19
## - Hmean  1    265.36 6584.0 203.30
## - p50    1    344.39 6663.0 203.71
## <none>               6318.6 203.86
## - h50    1    373.93 6692.6 203.87
## - h70    1    519.96 6838.6 204.62
## - h30    1   1296.64 7615.3 208.39
## 
## Step:  AIC=201.89
## h_dm ~ Hmax + Hmean + h30 + h50 + h70 + h90 + p50 + p70 + p90
## 
##         Df Sum of Sq    RSS    AIC
## - p90    1     31.67 6355.9 200.06
## - p70    1    110.95 6435.2 200.50
## - Hmax   1    229.71 6553.9 201.14
## - h90    1    285.49 6609.7 201.43
## <none>               6324.2 201.89
## - h50    1    376.00 6700.2 201.91
## - p50    1    399.23 6723.4 202.03
```

```
## - Hmean  1    502.47 6826.7 202.56
## - h70    1    753.28 7077.5 203.83
## + p30    1      5.58 6318.6 203.86
## - h30    1   1867.73 8191.9 208.94
##
## Step:  AIC=200.06
## h_dm ~ Hmax + Hmean + h30 + h50 + h70 + h90 + p50 + p70
##
##         Df Sum of Sq    RSS    AIC
## - p70    1     82.99 6438.9 198.52
## - h90    1    276.67 6632.6 199.55
## - Hmax   1    290.35 6646.2 199.63
## - h50    1    359.96 6715.9 199.99
## <none>               6355.9 200.06
## - p50    1    378.31 6734.2 200.09
## - Hmean  1    481.24 6837.1 200.62
## - h70    1    724.02 7079.9 201.84
## + p90    1     31.67 6324.2 201.89
## + p30    1      0.89 6355.0 202.06
## - h30    1   1855.55 8211.4 207.03
##
## Step:  AIC=198.52
## h_dm ~ Hmax + Hmean + h30 + h50 + h70 + h90 + p50
##
##         Df Sum of Sq    RSS    AIC
## - Hmax   1    222.97 6661.8 197.71
## - h90    1    306.24 6745.1 198.14
## <none>               6438.9 198.52
## - h50    1    430.79 6869.7 198.78
## - p50    1    437.12 6876.0 198.82
## - Hmean  1    529.50 6968.4 199.28
## + p70    1     82.99 6355.9 200.06
## + p90    1      3.71 6435.2 200.50
## + p30    1      3.40 6435.5 200.50
## - h70    1    858.06 7296.9 200.90
## - h30    1   1982.91 8421.8 205.91
##
## Step:  AIC=197.71
## h_dm ~ Hmean + h30 + h50 + h70 + h90 + p50
##
##         Df Sum of Sq    RSS    AIC
## - p50    1    245.08 6906.9 196.97
## - h90    1    309.97 6971.8 197.30
## <none>               6661.8 197.71
## - Hmean  1    422.11 7084.0 197.86
## - h50    1    505.19 7167.0 198.27
## + Hmax   1    222.97 6438.9 198.52
## + p30    1     94.55 6567.3 199.21
## + p90    1     15.88 6646.0 199.62
## + p70    1     15.61 6646.2 199.63
## - h70    1    806.35 7468.2 199.71
## - h30    1   1918.83 8580.7 204.57
##
## Step:  AIC=196.97
## h_dm ~ Hmean + h30 + h50 + h70 + h90
```

```
##
##          Df Sum of Sq    RSS    AIC
## - h90    1    246.97 7153.9 196.20
## <none>               6906.9 196.97
## - h50    1    407.89 7314.8 196.98
## - Hmean  1    531.06 7438.0 197.56
## + p50    1    245.08 6661.8 197.71
## + p90    1    164.41 6742.5 198.13
## + p70    1    153.96 6753.0 198.18
## + p30    1     78.54 6828.4 198.57
## + Hmax   1     30.93 6876.0 198.82
## - h70    1    922.91 7829.8 199.36
## - h30    1   2293.52 9200.5 205.01
##
## Step:  AIC=196.2
## h_dm ~ Hmean + h30 + h50 + h70
##
##          Df Sum of Sq    RSS    AIC
## - Hmean  1    299.56 7453.5 195.64
## <none>               7153.9 196.20
## + h90    1    246.97 6906.9 196.97
## + p50    1    182.09 6971.8 197.30
## + p90    1    107.42 7046.5 197.67
## + p70    1     97.81 7056.1 197.72
## - h70    1    758.68 7912.6 197.73
## + p30    1     31.81 7122.1 198.05
## + Hmax   1     13.93 7140.0 198.13
## - h50    1   1287.55 8441.4 200.00
## - h30    1   2727.59 9881.5 205.51
##
## Step:  AIC=195.64
## h_dm ~ h30 + h50 + h70
##
##          Df Sum of Sq     RSS    AIC
## <none>                 7453.5 195.64
## + p50    1     359.1  7094.4 195.91
## + Hmean  1     299.6  7153.9 196.20
## + p70    1     205.1  7248.4 196.66
## + p90    1     128.5  7325.0 197.03
## + Hmax   1     125.3  7328.1 197.04
## + p30    1      72.0  7381.5 197.30
## - h70    1     863.1  8316.6 197.47
## + h90    1      15.5  7438.0 197.56
## - h50    1     995.2  8448.7 198.03
## - h30    1   12258.4 19711.8 227.68
```

```
m_step_h$anova #display results
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## h_dm ~ Hmax + Hmean + h30 + h50 + h70 + h90 + p30 + p50 + p70 +
##     p90
##
## Final Model:
```

```
## h_dm ~ h30 + h50 + h70
##
##
##      Step Df   Deviance Resid. Df Resid. Dev      AIC
## 1                              24   6318.642 203.8569
## 2   - p30  1   5.575097        25   6324.217 201.8878
## 3   - p90  1  31.669329        26   6355.887 200.0626
## 4   - p70  1  82.989887        27   6438.877 198.5166
## 5  - Hmax  1 222.971013        28   6661.848 197.7081
## 6   - p50  1 245.078713        29   6906.926 196.9726
## 7   - h90  1 246.972425        30   7153.899 196.2023
## 8 - Hmean  1 299.563677        31   7453.462 195.6380
```

Final Model chosen after using the pvalues and stepwise regression: h_dm ~ h30 + h50 + h70

The final models created are:

FOR LINDEN $dbh_m m = Hmax + Hmean + h30 + h70 + p50 + h90 + p90$ $h_d m = Hmax + h30 + h50 + h70 + p50 + p70$

FOR MAPLE $dbh_m m = h30 + p90 + h90$ $h_d m$ $h30 + h50 + h70$

Final creation of models based on the chosen predictors

LINDEN

```
#for diameter at breast height(DBH)
lindbh_model <- lm(dbh_mm ~ Hmax + Hmean + h30 + h70
                   + p50 + h90 + p90, linden1)
lindbh_coef <- coef(lindbh_model) #extract the coefficients


#model for height for linden species
linh_model <-
  lm(h_dm ~ Hmax + h30 + h50 + h70 + p50 + p70, linden1)
linh_coef <- coef(linh_model) #extract the coefficients


#MAPLES
#model for diameter at breast height(DBH) of maple
mapdbh_model <- lm(dbh_mm ~ Hmax + Hmean + h30 + h70 + p50
                   + h90 + p90, maple1)
mapdbh_coef <- coef(mapdbh_model) #extract the coefficients

#model for height of maple
maph_model <- lm(h_dm ~ h30 + h50 + h70, maple1)
maph_coef <- coef(maph_model) #extract the coefficients
```

Creating functions to calculate the parameters

```
#create function for calculating the DBH of linden, by using the model
lin_DBH_fun <-
  function(Hmax , Hmean , h30 , h70 , p50 , h90 , p90) {
    lindbh_mod <-
      round((
        lindbh_coef[1] + (lindbh_coef[2] * Hmax) + (lindbh_coef[3] * Hmean) +
```

```
          (lindbh_coef[4] * h30) + lindbh_coef[5] * h70 + lindbh_coef[6] * p50
        + (lindbh_coef[7] * h90) + (lindbh_coef[8] * p90)
      ),
      2)
    return(lindbh_mod)
  }

#function for calculating the Height of linden, by using the model
lin_H_fun <- function(Hmax,  h30 , h50, h70, p50, p70) {
  linh_mod <- round((
    linh_coef[1] + (linh_coef[2] * Hmax) +
      (linh_coef[3] * h30) + linh_coef[4] * h50 + linh_coef[5] *
      h70 + linh_coef[6] * p50
    + (linh_coef[7] * p70)
  ),
  2)
  return(linh_mod)
}

#create function for calculating the DBH of maple, by using the model
map_DBH_fun <-
  function(Hmax , Hmean , h30 , h70 , p50 , h90 , p90) {
    mapdbh_mod <-
      round((
        mapdbh_coef[1] + (mapdbh_coef[2] * Hmax) + (mapdbh_coef[3] * Hmean) +
          (mapdbh_coef[4] * h30) + mapdbh_coef[5] * h70 + mapdbh_coef[6] * p50
        + (mapdbh_coef[7] * h90) + (mapdbh_coef[8] * p90)
      ),
      2)
    return(mapdbh_mod)
  }

#create function for calculating the Height of Maple, by using the model
map_H_fun <- function(h30 , h50, h70) {
  maph_mod <- round((
    maph_coef[1] + (maph_coef[2] * h30)
    + maph_coef[3] * h50 + maph_coef[4] * h70
  ), 2)
  return(maph_mod)
}
```

Loop to calculate the predictions into a dataframe

```
#Create a loop to predict the heigt and dbh of maple and linden
#by using the created models
{
  dbh_ln <- h_ln <- dbh_map <- h_map <- puunro <- puulaji <- c()
  for (i in 1:nrow(data2)) {
    if (data2$puulaji[i] == "lehmus") {
      dbh_ln <- append(
        dbh_ln,
        lin_DBH_fun(
          data2$Hmax[i] ,
          data2$Hmean[i]
          ,
          data2$h30[i] ,
```

```r
          data2$h70[i] ,
          data2$p50[i]
          ,
          data2$h90[i] ,
          data2$p90[i]
        )
      )

      h_ln <-
        append(
          h_ln,
          lin_H_fun(
            data2$Hmax[i] ,
            data2$h30[i] ,
            data2$h50[i] ,
            data2$h70[i]
            ,
            data2$p50[i] ,
            data2$p70[i]
          )
        )
      puunro <- append(puunro, data2$Puunro[i])
      puulaji <- append(puulaji, as.character(data2$puulaji[i]))

    }

#here, I can also use  else alone instead of if (data2$puulaji[i] == "vaahtera")
    if (data2$puulaji[i] == "vaahtera") {
      dbh_map <-
        append(dbh_map, (
          map_DBH_fun(
            data2$Hmax[i] ,
            data2$Hmean[i],
            data2$h30[i]
            ,
            data2$h70[i] ,
            data2$p50[i],
            data2$h90[i] ,
            data2$p90[i]
          )
        ))

      h_map <-
        append(h_map, (map_H_fun(data2$h30[i] , data2$h50[i] , data2$h70[i])))
      puunro <- append(puunro, data2$Puunro[i])
      puulaji <- append(puulaji, as.character(data2$puulaji[i]))
    }
  }
  #combine the  dbh and height vectors created for linden column-wise
  a <- cbind(dbh_ln, h_ln)
  #combine the  dbh and height vectors created for maple column-wise
  b <- cbind(dbh_map, h_map)

  #now, combine both data but row_wise since we want them to be merged
  c <- rbind(a, b)
```

```r
#finally, add the plot number and names of the species which are
  #vectors created earlier for the entire data
  maple_linden <- cbind.data.frame(puunro, puulaji, c)

  #reset the index/rownames to default index
  rownames(maple_linden) <- NULL

  #rownames(maple_linden)<-rownames(maple_linden, do.NULL=T, prefix = "Obs.")

  #rename columns one and two
 colnames(maple_linden)[colnames(maple_linden)=="dbh_ln"] <- "DBH"
 colnames(maple_linden)[colnames(maple_linden)=="h_ln"] <- "height"

 #names(maple_linden)[1]<-"DBH"  #wont use this cos column number might change
  #names(maple_linden) = c("DBH", "height")
}
```

let's see the summary and the head part of the modelled data

```r
summary(maple_linden)

##      puunro          puulaji           DBH                height
##  Min.   :  134   lehmus   :5225   Min.   : -67.19   Min.   :   6.53
##  1st Qu.: 3600   vaahtera: 956   1st Qu.: 192.96   1st Qu.:  88.57
##  Median : 5893                    Median : 261.12   Median : 113.44
##  Mean   :13296                    Mean   : 278.56   Mean   : 118.28
##  3rd Qu.: 9506                    3rd Qu.: 359.10   3rd Qu.: 144.27
##  Max.   :51921                    Max.   :2476.28   Max.   :1171.89

head(maple_linden, n=15)

##    puunro  puulaji    DBH height
## 1     134 vaahtera 253.38  98.27
## 2     143 vaahtera 277.50 103.96
## 3     173 vaahtera 285.79 109.77
## 4     175   lehmus 278.35 103.70
## 5     176   lehmus 336.67 123.63
## 6     177   lehmus 304.81 115.54
## 7     178   lehmus 357.86 136.43
## 8     179   lehmus 376.53 147.17
## 9     180   lehmus 414.90 160.16
## 10    181   lehmus 445.74 165.78
## 11    182   lehmus 435.29 165.75
## 12    183   lehmus 444.21 161.59
## 13    184   lehmus 394.79 162.45
## 14    185   lehmus 451.77 171.20
## 15    186   lehmus 467.53 173.25
```

create a more advanced plot matrix with ggpairs()

```r
p2 <- ggpairs(maple_linden, mapping = aes(col=puulaji, alpha=0.3), lower = list(combo =
wrap("facethist", bins = 20)))

#draw the plot
p2
```

As we can see from the distribution above, lehmus is much more than vaahtera. We can also see that the diameter is highly correlated with the DBH. There also seems to be some outliers in the predicted dbh and height for both species. The predicted height seems to be mostly around <=300dm. The predicted man height is about 118.28dm while the mean dbh is 278.56mm.

Regression models are useful ways to make predictions for extrapolating and interpolating because it is pratically impossible to capture the entire reality. In this exercise, I adopted the principle of parsimony by using as less predictors as possible. I also tried my hands on creating functions for making te predictions. However, it can be simply done by using a function in R called "predict.lm()"

```
#write the data into csv format
write.csv(maple_linden, file = "C:/Users/oyeda/Desktop/R_COURSE/modelling/Trees_DBH_H"
          , row.names = TRUE)
```

NOTE: THERE ARE SIMPLER APPROACHES TO CALCULATING THE HEIGHT AND DBH INTO NEW COLUMNS E.G DATA$DBH<- FORMULA(USING NECESSARY COLUMNS). BUT I CHOSE TO TRY OUT LOOPING, BINDING AND APPENDING. THE PREDICTION CAN ALSO BE DONE BY USING THE PREDICT.LM FUNCTION IN R