Stat 400 Project

28 February 2022

GLM

```
rm(list=ls())
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v tibble  3.1.6     v purrr   0.3.4
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(psych)
```

```
##
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
library(lme4)
```

```
## Loading required package: Matrix

##
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
library(lmerTest)
```

```
##
## Attaching package: 'lmerTest'

## The following object is masked from 'package:lme4':
##
##     lmer

## The following object is masked from 'package:stats':
##
##     step
```

Source

The following data source was acquired from Kaggle. It lists multiple types of halloween candies and their traits. Data was collected from a game where participants chose between 2 candies. The win percentage of candies was the main focus of the study. (https://www.kaggle.com/fivethirtyeight/the-ultimate-halloween-candy-power-ranking)

```
candy = read.csv(file='C:/users/omerc/Downloads/candy-data.csv', header=T)
```

EDA

```
head(candy)
```

```
##   competitorname chocolate fruity caramel peanutyalmondy nougat
## 1      100 Grand         1      0       1              0      0
## 2   3 Musketeers         1      0       0              0      1
## 3       One dime         0      0       0              0      0
## 4    One quarter         0      0       0              0      0
## 5      Air Heads         0      1       0              0      0
## 6     Almond Joy         1      0       0              1      0
##   crispedricewafer hard bar pluribus sugarpercent pricepercent winpercent
## 1                1    0   1        0        0.732        0.860   66.97173
## 2                0    0   1        0        0.604        0.511   67.60294
## 3                0    0   0        0        0.011        0.116   32.26109
## 4                0    0   0        0        0.011        0.511   46.11650
## 5                0    0   0        0        0.906        0.511   52.34146
## 6                0    0   1        0        0.465        0.767   50.34755
```

```
str(candy)
```

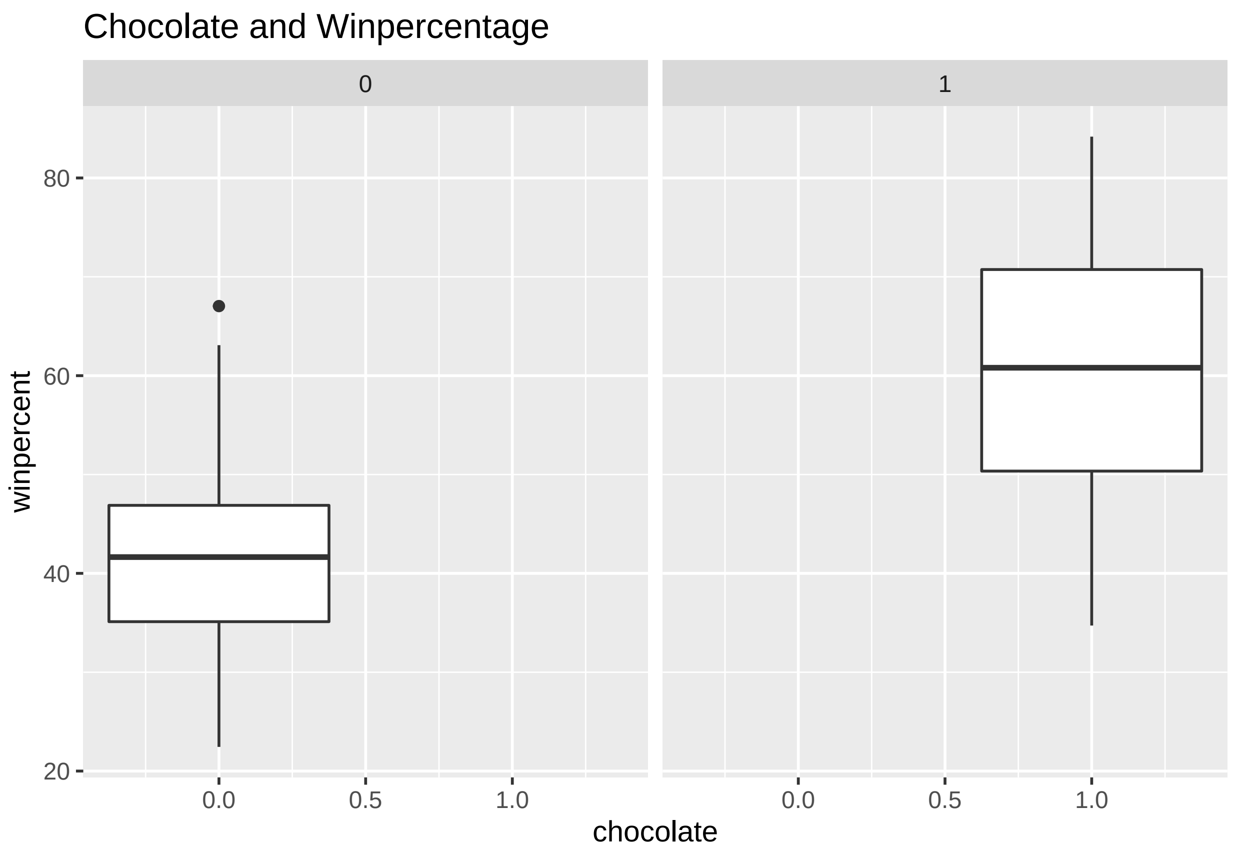
```
## 'data.frame':    85 obs. of  13 variables:
##  $ competitorname  : chr  "100 Grand" "3 Musketeers" "One dime" "One quarter" ...
##  $ chocolate       : int  1 1 0 0 0 1 1 0 0 0 ...
##  $ fruity          : int  0 0 0 0 1 0 0 0 0 1 ...
##  $ caramel         : int  1 0 0 0 0 0 1 0 0 1 ...
##  $ peanutyalmondy  : int  0 0 0 0 0 1 1 1 0 0 ...
##  $ nougat          : int  0 1 0 0 0 0 1 0 0 0 ...
##  $ crispedricewafer: int  1 0 0 0 0 0 0 0 0 0 ...
##  $ hard            : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ bar             : int  1 1 0 0 0 1 1 0 0 0 ...
##  $ pluribus        : int  0 0 0 0 0 0 0 1 1 0 ...
##  $ sugarpercent    : num  0.732 0.604 0.011 0.011 0.906 ...
##  $ pricepercent    : num  0.86 0.511 0.116 0.511 0.511 ...
##  $ winpercent      : num  67 67.6 32.3 46.1 52.3 ...
```

We would like to make a variable that tells us if the candy won the majority of its matchups. If the value is 1, then it won the majority. If it's 0, it did not. We also add an ID column for possible future reference. We also notice that some columns are characters columns, so we switch the relevant ones to integer columns.

```
candy <- mutate(candy, majority = ifelse(winpercent > 50, 1, "0"))
candy <- tibble::rowid_to_column(candy, "ID")
candy$majority <- as.numeric(candy$majority)
```

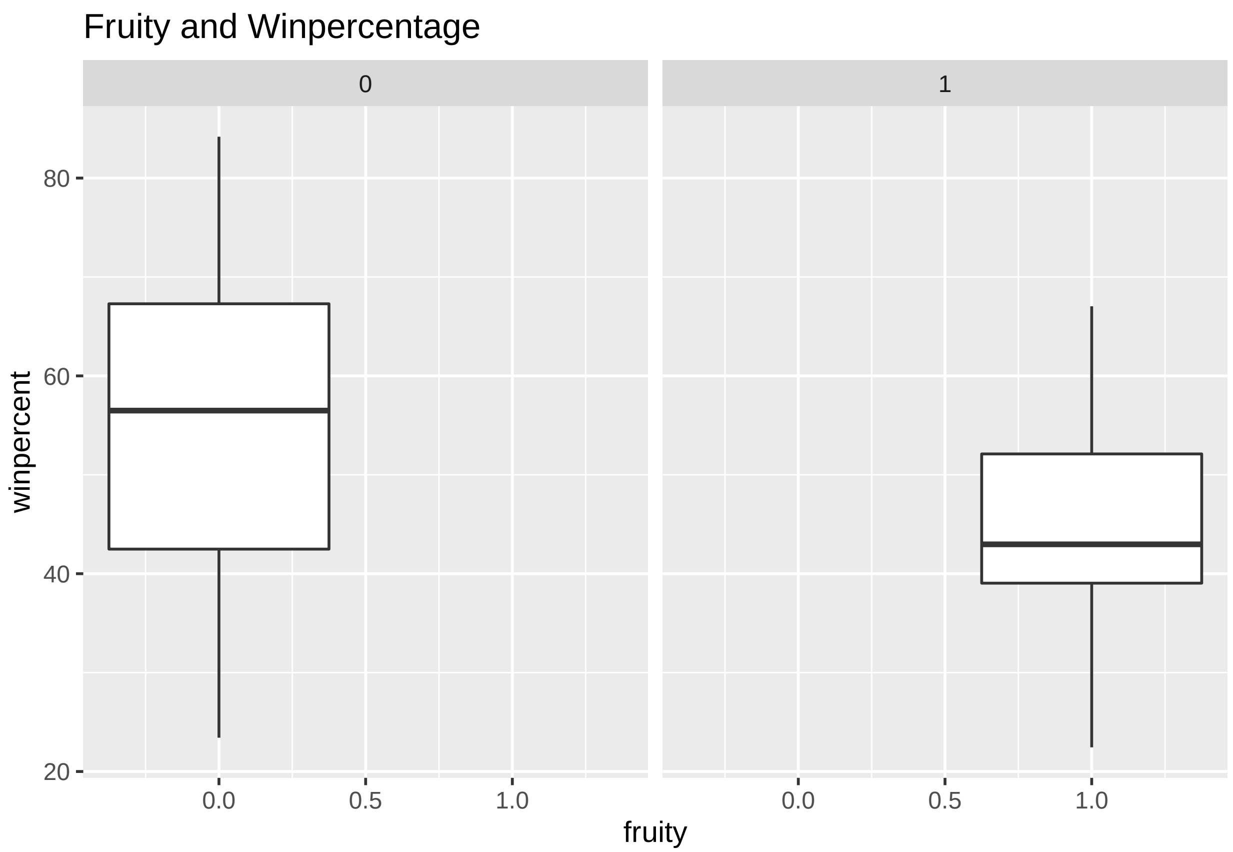
```
candy %>%
ggplot(aes(x=chocolate,y=winpercent)) +
  geom_boxplot()+ facet_wrap( ~ chocolate) +  labs(title = 'Chocolate and Winpercentage')
```

```
## Warning: Continuous x aesthetic -- did you forget aes(group=...)?
```



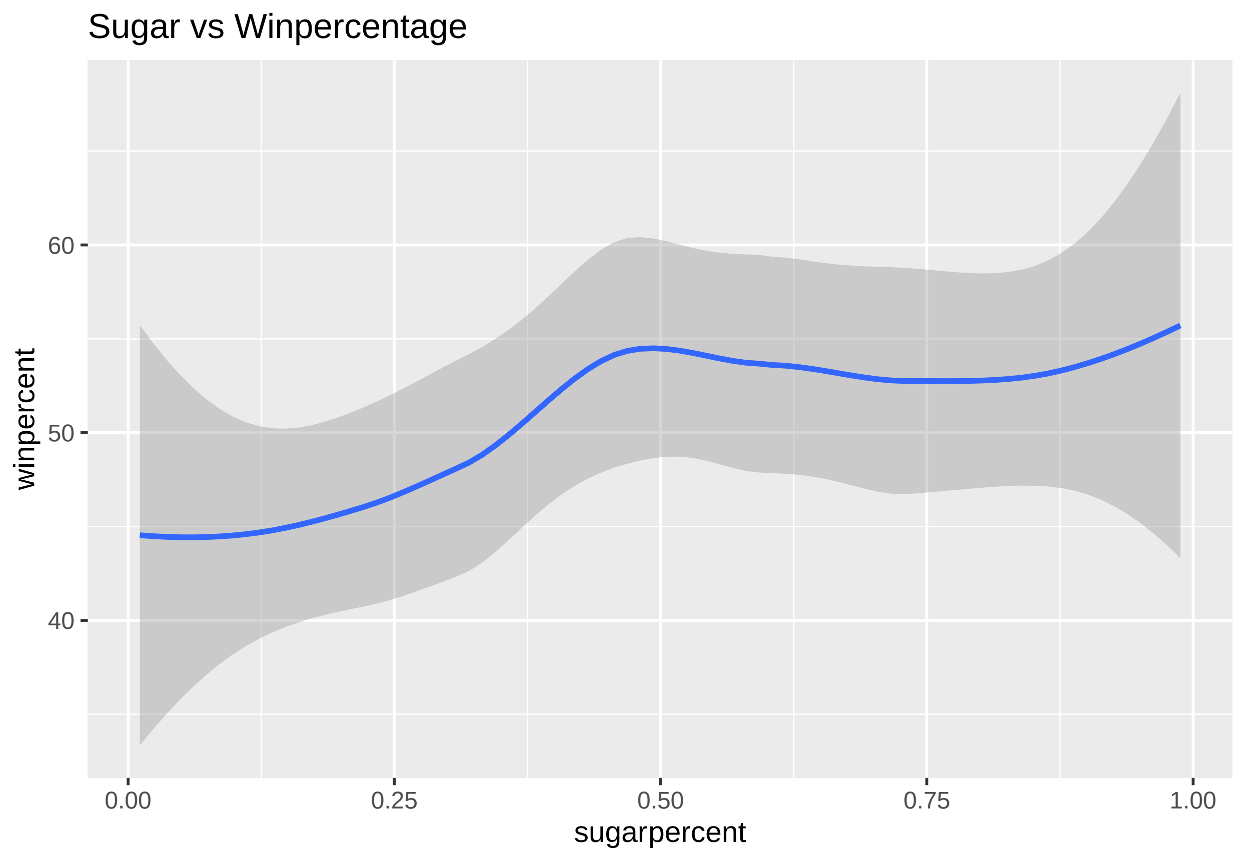
```
candy %>%
ggplot(aes(x=fruity,y=winpercent)) +
  geom_boxplot()+facet_wrap( ~ fruity)+  labs(title = 'Fruity and Winpercentage')
```

```
## Warning: Continuous x aesthetic -- did you forget aes(group=...)?
```



```
candy %>%
ggplot(aes(x=sugarpercent,y=winpercent)) + geom_smooth()+  labs(title = 'Sugar vs Winpercentage')
```

```
## 'geom_smooth()' using method = 'loess' and formula 'y ~ x'
```



We notice that if a candy contains chocolate, it usually wins the majority of its match-ups. The graphs also tell us that fruity candy's do not usually the majority of their match-ups. Lastly, we see that as sugar percentile rises, the win percentage tends to rise.

Statistical Analysis

Model 1

We begin with a model containing solely chocolate.

```
model1 <- glm(majority ~ chocolate, family = binomial, data = candy)
summary(model1)
```

```
## 
## Call:
## glm(formula = majority ~ chocolate, family = binomial, data = candy)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.6815  -0.7215  -0.7215   0.7466   1.7166  
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept)  -1.2130     0.3434  -3.532 0.000412 ***
## chocolate     2.3480     0.5145   4.563 5.04e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 117.258  on 84  degrees of freedom
## Residual deviance:  92.728  on 83  degrees of freedom
## AIC: 96.728
##
## Number of Fisher Scoring iterations: 4
```

Our model tells us that that chocolate is a significant variable. Our estimated binomial regression model is:

$$\log\left(\frac{p_i}{1-p_i}\right) = -1.2130 + 2.348 chocolate$$

where p is the estimated proportion of candies who win the majority of their matchups. We can interpret the coefficient on chocolate as

$$e^2 = 10.8$$

indicating that the odds of a chocolate winning the majority of their matchups is 10 times the odds of of a non-chocolate winning the majority of their matchups

Model 2

We notice that chocolate is significant, but our EDA told us that fruity also seemed to have correlation with the response. We therefore create a model with these two variables.

```
model2 <- glm(majority ~ chocolate + fruity, family = binomial, data = candy)
summary(model2)
```

```
##
## Call:
## glm(formula = majority ~ chocolate + fruity, family = binomial,
##     data = candy)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2770  -0.7973  -0.4192   0.7549   1.6130
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -2.388      1.045  -2.284  0.02235 *
## chocolate      3.498      1.104   3.169  0.00153 **
## fruity         1.405      1.099   1.278  0.20112
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 117.258  on 84  degrees of freedom
## Residual deviance:  90.535  on 82  degrees of freedom
## AIC: 96.535
##
## Number of Fisher Scoring iterations: 5
```

We see that there is some correlation with fruity, but it is not significant. However, our AIC has decreased so we will keep it in the model.

Model 3

We consider the hard variable

```
model3 <- glm(majority ~ chocolate + hard + fruity , family = binomial, data = candy)
summary(model3)
```

```
##
## Call:
## glm(formula = majority ~ chocolate + hard + fruity, family = binomial,
##     data = candy)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6920  -1.0112  -0.2573   0.7390   2.6172
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -2.648      1.185  -2.235  0.02542 *
## chocolate      3.806      1.221   3.118  0.00182 **
## hard          -2.987      1.403  -2.129  0.03328 *
## fruity         2.244      1.251   1.793  0.07295 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 117.258  on 84  degrees of freedom
## Residual deviance:  82.513  on 81  degrees of freedom
## AIC: 90.513
##
## Number of Fisher Scoring iterations: 6
```

Hard is added and is also a significant variable. We also see a decrease in AIC.

Model 4

We consider the sugarpercent and peanutyalmondy variable.

```
model4 <- glm(majority ~ chocolate + hard + fruity + sugarpercent + peanutyalmondy, family = binomial,
summary(model4)
```

```
##
## Call:
## glm(formula = majority ~ chocolate + hard + fruity + sugarpercent +
##     peanutyalmondy, family = binomial, data = candy)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -1.8019  -0.8191  -0.1470   0.8235   2.5165
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -4.466      1.517  -2.944  0.00324 **
## chocolate         4.254      1.387   3.067  0.00216 **
## hard             -3.474      1.542  -2.253  0.02423 *
## fruity            3.246      1.465   2.216  0.02669 *
## sugarpercent      1.852      1.109   1.670  0.09489 .
## peanutyalmondy    2.508      1.339   1.873  0.06104 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 117.258  on 84  degrees of freedom
## Residual deviance:  73.792  on 79  degrees of freedom
## AIC: 85.792
##
## Number of Fisher Scoring iterations: 6
```

Sugarpercent is a little farther from being significant. However, because it lowered our AIC even farther, we will keep it as our final model.

Model 5 (Interactions)

Most fruity candies are hard, so I wanted to test if there was any interaction bertween fruity and hard

```
model5 = glm(majority ~ chocolate + hard* fruity + sugarpercent + peanutyalmondy, family = binomial, da
summary(model5)
```

```
##
## Call:
## glm(formula = majority ~ chocolate + hard * fruity + sugarpercent +
##     peanutyalmondy, family = binomial, data = candy)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8019  -0.8190  -0.1474   0.8234   2.5145
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -4.461      1.520  -2.935  0.00334 **
## chocolate         4.249      1.390   3.057  0.00223 **
## hard            -13.032   1658.131  -0.008  0.99373
## fruity            3.240      1.468   2.207  0.02731 *
## sugarpercent      1.852      1.109   1.670  0.09493 .
## peanutyalmondy    2.506      1.339   1.872  0.06125 .
## hard:fruity       9.564   1658.132   0.006  0.99540
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 117.258  on 84  degrees of freedom
## Residual deviance:  73.788  on 78  degrees of freedom
## AIC: 87.788
##
## Number of Fisher Scoring iterations: 15
```

The interaction term is far from significant and AIC increases.

Model 6 (More Interactions)

Most chocolate candies have nuts, so I wanted to test if there was any interaction between chocolate and peanutyalmondy

```
model6 = glm(majority ~ chocolate * peanutyalmondy + hard + fruity + sugarpercent, family = binomial, da
summary(model6)
```

```
##
## Call:
## glm(formula = majority ~ chocolate * peanutyalmondy + hard +
##     fruity + sugarpercent, family = binomial, data = candy)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7553  -0.8307  -0.1023   0.8497   2.3770
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -3.453      1.420  -2.431   0.0151 *
## chocolate                   3.235      1.308   2.472   0.0134 *
## peanutyalmondy            -15.796   4604.684  -0.003   0.9973
## hard                       -3.071      1.355  -2.266   0.0234 *
## fruity                      2.284      1.341   1.703   0.0885 .
## sugarpercent                1.740      1.092   1.594   0.1110
## chocolate:peanutyalmondy   33.664   4963.668   0.007   0.9946
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 117.258  on 84  degrees of freedom
## Residual deviance:  71.332  on 78  degrees of freedom
## AIC: 85.332
##
## Number of Fisher Scoring iterations: 17
```

The interaction term is far from significant but AIC decreases

```
two.way <- aov(majority ~  hard + fruity + sugarpercent + peanutyalmondy * chocolate, data = candy)
summary(two.way)
```

```
##                           Df Sum Sq Mean Sq F value   Pr(>F)    
## hard                       1  2.801  2.8011  17.581 7.21e-05 ***
## fruity                     1  0.664  0.6639   4.167   0.0446 *  
## sugarpercent               1  0.876  0.8756   5.496   0.0216 *  
## peanutyalmondy             1  1.038  1.0376   6.512   0.0127 *  
## chocolate                  1  2.990  2.9904  18.769 4.35e-05 ***
## peanutyalmondy:chocolate  1  0.310  0.3097   1.944   0.1672    
## Residuals                 78 12.428  0.1593                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

PeanutAlmondy is not significant so we will use the model without the interaction (model4).

Overdispersion

We deny the need for an overdispersion parameter because our residual deviance is close enough to a 1:1 ratio with our degrees of freedom.

Interpreting Parameters

$$e^4 = 70.386$$

so the odds of a chocolate winning the majority of their matchups is 70.386 times the odds of of a non-chocolate winning the majority of their matchups

$$e^-3 = 0.03099$$

so the odds of a hard candy winning the majority of their matchups is 0.03099 times the odds of of a non-hard winning the majority of their matchups

$$e^3 = 25.687$$

so the odds of a fruity winning the majority of their matchups is 25.687 times the odds of of a non-fruity winning the majority of their matchups

$$e^1 = 6.37$$

this is log odds increase of winning the majority of matchups for every rise in sugar percentile

$$e^2 = 12.28$$

so the odds of a peanutyalnmondy winning the majority of their matchups is 12.28 times the odds of of a non-peanutalmondy winning the majority of their matchups

Conclusion

Our final model to predict whether certain traits give a candy a higher chance of winning the majority of its matchups is

$$\log\left(\frac{p_i}{1-p_i}\right) = -4.466 + 4.254 chocolate - 3.474 hard + 3.246 fruity + 1.852 sugarpercent + 2.508 peanutyalmondy$$

We can be sure of this for a couple of reasons. This model gave us the most significant terms. Although some terms were not significant, it also gave us the lowest AIC. This tells us that this model gives us the least prediction error while also keeping the most predictors.

LMM

Source

The following data source was acquired from uvm.edu. It shows various tree statistics grouped by multiple variables. The data will be analyzed at 2 levels, including Plot and Tree At each level, we will analyze a variable (e.g. height at a certain year). The data was collected by the state of Vermont to capture broad temporal changes in the condition of the national forest resource. (https://www.uvm.edu/femc/data/archive/project/federal-forest-inventory-analysis-data-for/dataset/tree-data-for-intensive-sampling-forest)

```
data = read.csv(file = "C:/users/omerc/Downloads/tree.csv")
```

Summarized results and Variable Renaming

Very few data cleaning was required. Sample of the data set using is shown below as well as renaming one variable (YEAR) for convenience when model building. Also added an ID column

```
head(data)
```

```
##   PLOT SUBP TREE INVYR      LAT       LON ELEV SPCD     COMMON_NAME GENUS
## 1 1129    1    1  2003 44.57789 -72.06068 1079  261 eastern hemlock Tsuga
## 2 1129    1    2  2003 44.57789 -72.06068 1079  261 eastern hemlock Tsuga
## 3 1129    1    3  2003 44.57789 -72.06068 1079  261 eastern hemlock Tsuga
## 4 1129    1    4  2003 44.57789 -72.06068 1079  261 eastern hemlock Tsuga
## 5 1129    2    1  2003 44.57789 -72.06068 1079  318     sugar maple  Acer
## 6 1129    2    2  2003 44.57789 -72.06068 1079  318     sugar maple  Acer
##      SPECIES  DIA HT ACTUALHT MORTCFAL TPAMORT_UNADJ GROWCFAL TPAGROW_UNADJ
## 1 canadensis 10.2 66       66       NA            NA       NA            NA
## 2 canadensis 11.0 64       64       NA            NA       NA            NA
## 3 canadensis  6.9 30       30       NA            NA       NA            NA
## 4 canadensis 12.0 67       67       NA            NA       NA            NA
## 5  saccharum  2.8 32       32       NA            NA       NA            NA
## 6  saccharum  3.9 36       36       NA            NA       NA            NA
##   CDIEBKCD TRANSCD DRYBIO_STUMP DRYBIO_BOLE DRYBIO_TOP  DRYBIO_BG CARBON_AG
## 1        0      20    20.894415    346.2560   60.75819  98.122302 213.95429
## 2        5      25    23.566141    406.9286   70.40636 114.552869 250.45055
## 3        5      40     8.627827    114.7891   22.14140  33.979989  72.77918
## 4        5      25    28.006881    506.8650   86.41649 141.682737 310.64416
## 5       NA      NA           NA          NA         NA   6.072238  14.11971
## 6       NA      NA           NA          NA         NA  13.170183  31.89602
##   CARBON_BG P3PANEL          CN      PLT_CN
## 1 49.061151       5 5.59493e+13 5.59492e+13
## 2 57.276435       5 5.59493e+13 5.59492e+13
## 3 16.989995       5 5.59493e+13 5.59492e+13
## 4 70.841369       5 5.59493e+13 5.59492e+13
## 5  3.036119       5 5.59493e+13 5.59492e+13
## 6  6.585092       5 5.59493e+13 5.59492e+13
```

```
str(data)
```

```
## ’data.frame':    4236 obs. of  29 variables:
##  $ PLOT         : int  1129 1129 1129 1129 1129 1129 1129 1129 1129 1129 ...
##  $ SUBP         : int  1 1 1 1 2 2 2 2 2 2 ...
##  $ TREE         : int  1 2 3 4 1 2 3 4 5 6 ...
##  $ INVYR        : int  2003 2003 2003 2003 2003 2003 2003 2003 2003 2003 ...
##  $ LAT          : num  44.6 44.6 44.6 44.6 44.6 ...
##  $ LON          : num  -72.1 -72.1 -72.1 -72.1 -72.1 ...
##  $ ELEV         : int  1079 1079 1079 1079 1079 1079 1079 1079 1079 1079 ...
##  $ SPCD         : int  261 261 261 261 318 318 318 531 375 318 ...
##  $ COMMON_NAME  : chr  "eastern hemlock" "eastern hemlock" "eastern hemlock" "eastern hemlock" ...
##  $ GENUS        : chr  "Tsuga" "Tsuga" "Tsuga" "Tsuga" ...
##  $ SPECIES      : chr  "canadensis" "canadensis" "canadensis" "canadensis" ...
##  $ DIA          : num  10.2 11 6.9 12 2.8 3.9 3.1 8.9 7.7 5 ...
##  $ HT           : int  66 64 30 67 32 36 36 67 68 52 ...
##  $ ACTUALHT     : int  66 64 30 67 32 36 36 9 68 52 ...
##  $ MORTCFAL     : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ TPAMORT_UNADJ: num  NA NA NA NA NA NA NA NA NA NA ...
##  $ GROWCFAL     : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ TPAGROW_UNADJ: num  NA NA NA NA NA NA NA NA NA NA ...
##  $ CDIEBKCD     : int  0 5 5 5 NA NA NA NA 5 5 ...
##  $ TRANSCD      : int  20 25 40 25 NA NA NA NA 30 15 ...
##  $ DRYBIO_STUMP : num  20.89 23.57 8.63 28.01 NA ...
##  $ DRYBIO_BOLE  : num  346 407 115 507 NA ...
##  $ DRYBIO_TOP   : num  60.8 70.4 22.1 86.4 NA ...
##  $ DRYBIO_BG    : num  98.12 114.55 33.98 141.68 6.07 ...
##  $ CARBON_AG    : num  214 250.5 72.8 310.6 14.1 ...
##  $ CARBON_BG    : num  49.06 57.28 16.99 70.84 3.04 ...
##  $ P3PANEL      : int  5 5 5 5 5 5 5 5 5 5 ...
##  $ CN           : num  5.59e+13 5.59e+13 5.59e+13 5.59e+13 5.59e+13 ...
##  $ PLT_CN       : num  5.59e+13 5.59e+13 5.59e+13 5.59e+13 5.59e+13 ...
```

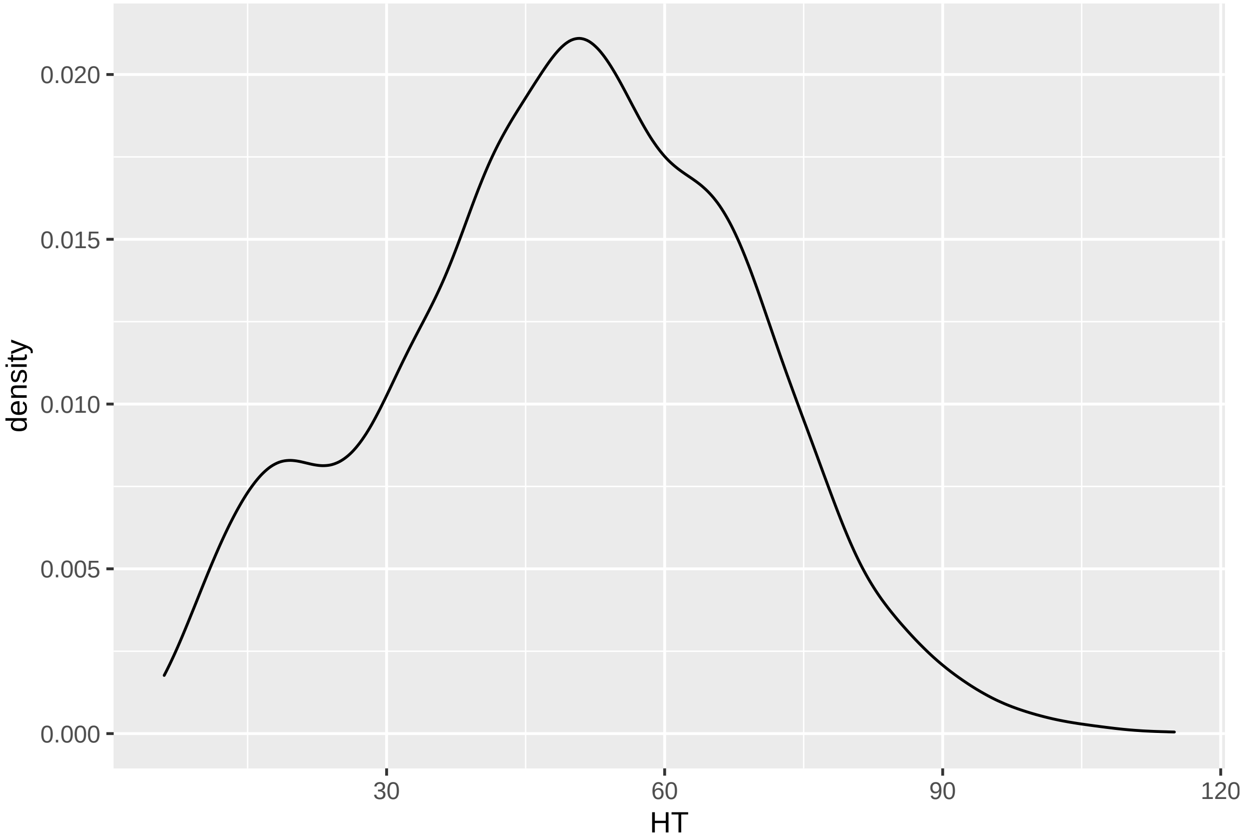
```
names(data)[names(data) == 'INVYR'] <- 'YEAR'
data$ID <- seq.int(nrow(data))
```

EDA

At level 2 we have have variables that have to do with the plot: Subplot, At level 1 we have variables that have to do with the trees: Height, SpeciesID, Diameter, Year

```
ggplot(data=data, aes(HT)) + geom_density()
```

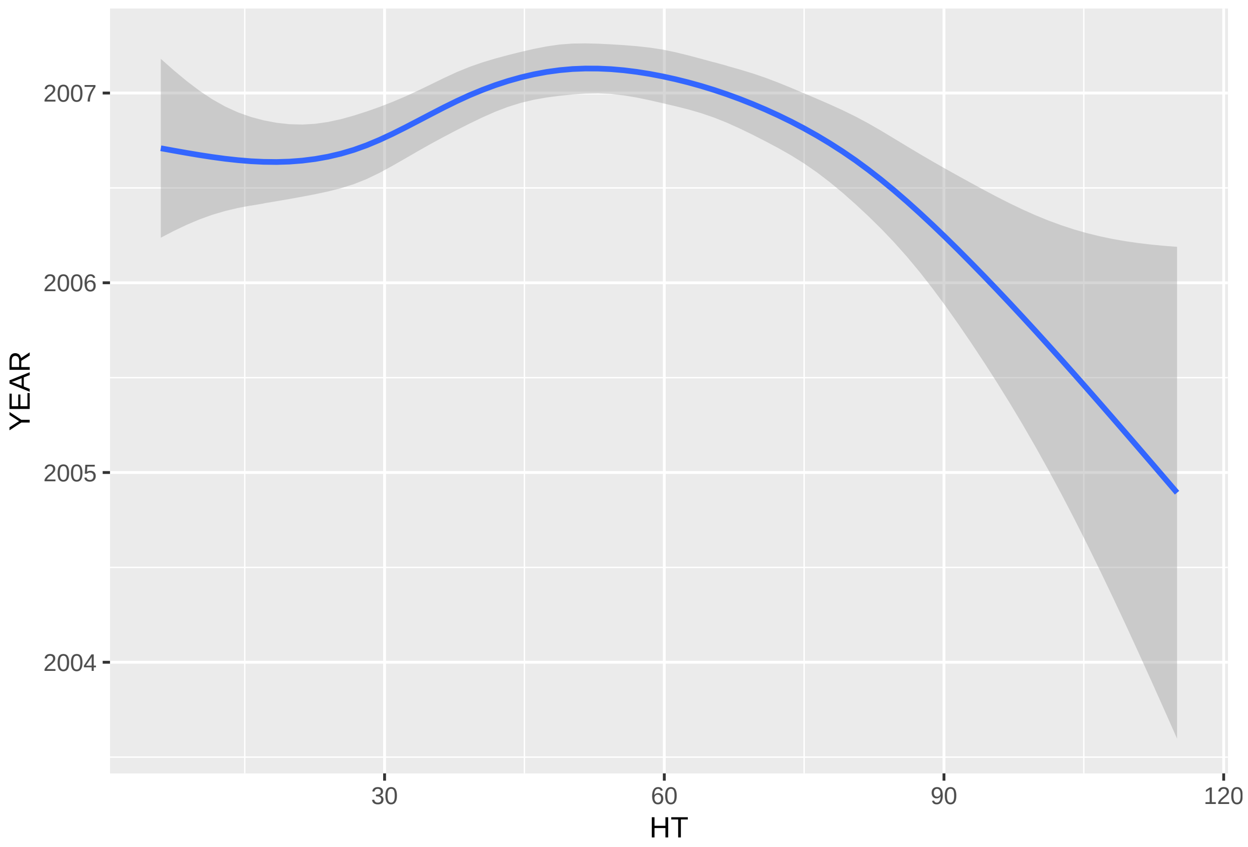
```
## Warning: Removed 725 rows containing non-finite values (stat_density).
```



```
ggplot(data=data, aes(x=HT,y=YEAR)) + geom_smooth()
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 725 rows containing non-finite values (stat_smooth).
```



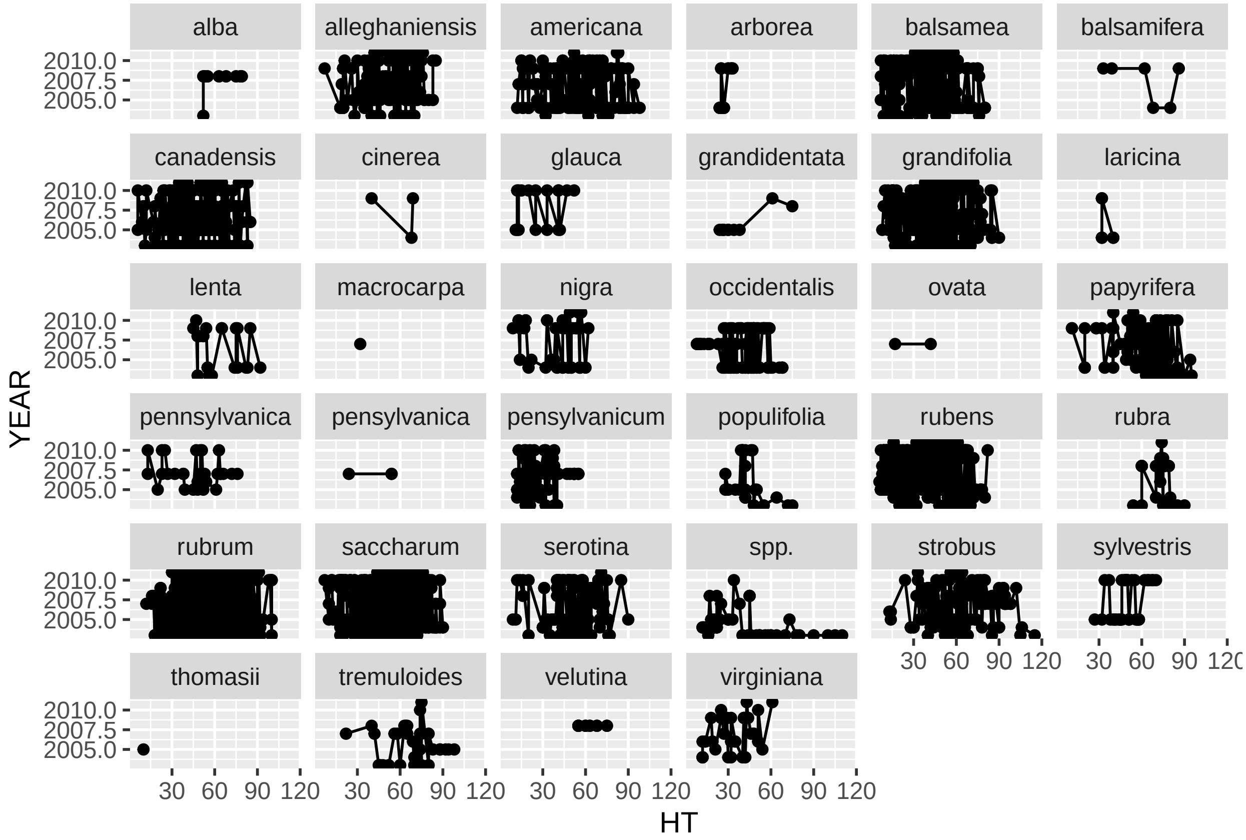
```r
ggplot(data, aes(x=HT,y=YEAR)) +
geom_point() + geom_line() + facet_wrap(~SPECIES,ncol=6)
```

```
## Warning: Removed 725 rows containing missing values (geom_point).
```

```
## Warning: Removed 9 row(s) containing missing values (geom_path).
```

```
## geom_path: Each group consists of only one observation. Do you need to adjust
## the group aesthetic?
```

```
## geom_path: Each group consists of only one observation. Do you need to adjust
## the group aesthetic?
```



Our first graph shows the density of height. It's distribution is approximately normal.

Our second graph shows a relationship between year and height. We see that trees in earlier years seemed to have been taller. Perhaps this means that trees were planted late into the data collection.

Our final graph shows HT vs Year for each species. There does not seem to be a significant enough relationship here. However, the model may tell us otherwise.

Model A: Unconditional Means Model

```
model.a <- lmer(HT ~ 1 + (1 | TREE), REML = T, data = data)
summary(model.a)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: HT ~ 1 + (1 | TREE)
##    Data: data
##
## REML criterion at convergence: 30629.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.4097 -0.6410  0.0466  0.6949  3.6359
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  TREE     (Intercept)  16.56    4.07
##  Residual             355.36   18.85
## Number of obs: 3511, groups:  TREE, 36
##
## Fixed effects:
##             Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)  47.5683     0.8713 25.7179    54.6   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Model B

Add a level 1 covariate (DIA)

```
model.b <- lmer(HT ~ DIA +  (DIA | TREE), data = data)
summary(model.b)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite’s method [
## lmerModLmerTest]
## Formula: HT ~ DIA + (DIA | TREE)
##    Data: data
##
## REML criterion at convergence: 27295.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.6125 -0.6986 -0.0115  0.6770  4.0950
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  TREE     (Intercept)   9.1205  3.020
##           DIA           0.1798  0.424   -0.84
##  Residual             136.7933 11.696
## Number of obs: 3511, groups:  TREE, 36
##
## Fixed effects:
##             Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)  21.0331     0.7866 16.8544   26.74 3.06e-15 ***
## DIA           3.5438     0.1058 13.1893   33.50 3.76e-14 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Correlation of Fixed Effects:
##     (Intr)
## DIA -0.859
```

21.03 mean ht before DIA
0.1798 mean increase in HT for increase in DIA 136.79 variance in tree deviations DIA is significant

Model C

Added a level 2 covariate

```
model.c <- lmer(HT ~ ELEV + DIA + ELEV:DIA +
  (1|TREE), data = data)
```

```
## Warning: Some predictor variables are on very different scales: consider
## rescaling

## Warning: Some predictor variables are on very different scales: consider
## rescaling
```

```
summary(model.c)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: HT ~ ELEV + DIA + ELEV:DIA + (1 | TREE)
##    Data: data
##
## REML criterion at convergence: 27250.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.9605 -0.6696 -0.0106  0.6523  4.0084
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  TREE     (Intercept)   1.993   1.412
##  Residual             135.251  11.630
## Number of obs: 3511, groups:  TREE, 36
##
## Fixed effects:
##                Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)  2.145e+01  1.071e+00  9.481e+02  20.023  < 2e-16 ***
## ELEV        -1.775e-04  7.137e-04  3.506e+03  -0.249    0.804
## DIA          4.047e+00  1.177e-01  3.505e+03  34.373  < 2e-16 ***
## ELEV:DIA    -3.726e-04  7.873e-05  3.504e+03  -4.733  2.3e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##          (Intr) ELEV   DIA
## ELEV     -0.881
## DIA      -0.830  0.782
## ELEV:DIA  0.781 -0.875 -0.919
## fit warnings:
## Some predictor variables are on very different scales: consider rescaling
```

We see that DIA and the interaction is significant

Model D

Added more level 1 covariates (SpeciesID, Year)

```
model.d <- lmer(HT ~ ELEV + DIA + SPCD + YEAR + YEAR:DIA + SPCD:DIA + ELEV:DIA +
  (DIA|TREE), data = data)
```

```
## Warning: Some predictor variables are on very different scales: consider
## rescaling

## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.020683 (tol = 0.002, component 1)

## Warning: Some predictor variables are on very different scales: consider
## rescaling
```

```
summary(model.d)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: HT ~ ELEV + DIA + SPCD + YEAR + YEAR:DIA + SPCD:DIA + ELEV:DIA +
##     (DIA | TREE)
##    Data: data
##
## REML criterion at convergence: 27149.9
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.4530 -0.6922 -0.0036  0.6548  3.8454
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  TREE     (Intercept)   7.7834  2.7899
##           DIA           0.1329  0.3646  -0.90
##  Residual             129.3755 11.3743
## Number of obs: 3511, groups:  TREE, 36
##
## Fixed effects:
##               Estimate Std. Error         df t value Pr(>|t|)
## (Intercept) -4.330e+02  3.355e+02  3.471e+03  -1.291 0.196869
## ELEV         7.636e-04  7.173e-04  3.485e+03   1.065 0.287152
## DIA          1.047e+02  3.652e+01  3.480e+03   2.867 0.004162 **
## SPCD         7.072e-03  2.113e-03  3.477e+03   3.348 0.000824 ***
## YEAR         2.244e-01  1.672e-01  3.472e+03   1.342 0.179642
## DIA:YEAR    -5.014e-02  1.820e-02  3.480e+03  -2.756 0.005890 **
## DIA:SPCD     3.565e-04  2.416e-04  3.405e+03   1.476 0.140105
## ELEV:DIA    -4.015e-04  7.889e-05  3.334e+03  -5.089 3.79e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##          (Intr) ELEV   DIA    SPCD   YEAR   DIA:YE DIA:SP
## ELEV      0.107
## DIA      -0.888 -0.075
## SPCD     -0.017  0.085  0.016
## YEAR     -1.000 -0.110  0.888  0.015
## DIA:YEAR  0.888  0.078 -1.000 -0.014 -0.888
## DIA:SPCD  0.018 -0.027 -0.021 -0.883 -0.016  0.019
## ELEV:DIA -0.074 -0.874  0.062 -0.032  0.077 -0.065  0.013
## fit warnings:
## Some predictor variables are on very different scales: consider rescaling
## optimizer (nloptwrap) convergence code: 0 (OK)
## Model failed to converge with max|grad| = 0.020683 (tol = 0.002, component 1)
```

We see that the SPCD variable is significant. Its added interaction is nearly significant at the 90% confidence level.

We do not see the year as a significant variable. This may be because the trees heights were taken after the trees had reached their full growth. We will run anova to see if it is a significant enough variable

```
model.e <- lmer(HT ~ ELEV + DIA + SPCD +  SPCD:DIA + ELEV:DIA +
  (DIA|TREE), data = data)
```

```
## Warning: Some predictor variables are on very different scales: consider
## rescaling
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.00641402 (tol = 0.002, component 1)
```

```
## Warning: Some predictor variables are on very different scales: consider
## rescaling
```

```
anova(model.d, model.e)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: data
## Models:
## model.e: HT ~ ELEV + DIA + SPCD + SPCD:DIA + ELEV:DIA + (DIA | TREE)
## model.d: HT ~ ELEV + DIA + SPCD + YEAR + YEAR:DIA + SPCD:DIA + ELEV:DIA + (DIA | TREE)
##         npar   AIC   BIC logLik deviance  Chisq Df Pr(>Chisq)
## model.e   10 27112 27174 -13546    27092
## model.d   12 27103 27177 -13540    27079 13.375  2   0.001247 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

The interaction and year varialbe is significant. We will keep it.

Final Model

Below is our final model.

```
model.d
```

```
## Linear mixed model fit by REML [’lmerModLmerTest’]
## Formula: HT ~ ELEV + DIA + SPCD + YEAR + YEAR:DIA + SPCD:DIA + ELEV:DIA +
##     (DIA | TREE)
##    Data: data
## REML criterion at convergence: 27149.86
## Random effects:
##  Groups   Name        Std.Dev. Corr 
##  TREE     (Intercept)  2.7899       
##           DIA          0.3646  -0.90
##  Residual             11.3743       
## Number of obs: 3511, groups:  TREE, 36
## Fixed Effects:
## (Intercept)         ELEV          DIA         SPCD         YEAR     DIA:YEAR  
##  -4.330e+02    7.636e-04    1.047e+02    7.072e-03    2.244e-01   -5.014e-02  
##    DIA:SPCD     ELEV:DIA  
##   3.565e-04   -4.015e-04  
## fit warnings:
## Some predictor variables are on very different scales: consider rescaling
## optimizer (nloptwrap) convergence code: 0 (OK) ; 0 optimizer warnings; 1 lme4 warnings
```

Conclusion

Our final model to predict whether the heights of trees based on level 2(traits of the plot) and level 1(traits of the tree) variables is as see in model.d. We choose this model for various reasons. This model gave us the most significant terms. Although some terms were not significant, after running the anova function, we see that the Year variable is significant. Also, we see that that DIA, SPCD, and a few interactions are also significant enough to keep in the model. This means that they are valid predictors for height/

Relevant Items

Presentation Link

Variables Link