Solution to Kim Iles Challenge
WESTERN MENSURATIONISTS CONFERENCE 2023

Mauricio Zapata
mzapatacuartas@finitecarbon.com

May 31, 2023

Announcement

0.1 Theoretical question

State the Basal Area Factor for Variable Plot Sampling (in Imperial units) for the following three angles: 90 degrees, 180 degrees, and 270 degrees.

0.2 Practical question

A series of diameter measurements along the stem was executed on a set of Ponderosa Pines located on the east sides of the Cascades. The data, which contains the tree number (i.e., Tree), age at the time of the measurement (i.e., Age-in years), the total height of the tree (i.e., TotalHeight-in feet), the height of the measured diameter (i.e., DiamHeight-in feet), and the diameter at the specified height (i.e., Diameter-in inches), can be downloaded HERE. You are tasked to develop a taper equation with the following requirements:

- The model can have at most four (4) parameters (Note: values estimated from the data and used in the final form of the model are considered parameters)
- Achievement of a pseudo coefficient of determination of at least 95%
- Fulfillment of all modeling requirements, with tests included

1 Solution to the theoretical question

The Basal Area Factor (BAF) depends on the measurement instrument used, and it allows us to determine the stand basal area at the point sampling by only knowing the number of trees counted or included in the horizontal sampling. Let's derive a general

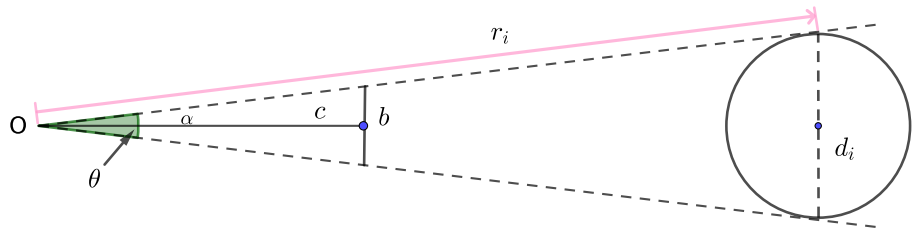

Figure 1: Simplified illustration of the Bitterlich Angle Gauge where the visual match exactly the tree dbh d (border tree). In this case, a virtual circular sub-plot of radio r is associated with this tree with diameter d. If the sampling plot (O) is located at a short distance than r, then the tree width at dbh should not covered by the visual angle defined with the horizontal target.

formula to compute the BAF given any angle, and then I'll calculate the BAFs for the requested angles.

To derive the BAF, we need to consider the size of each possible tree-specific virtual sub-plot. Lets first consider the situation presented in Figure 1 where there is an exact covering of the opening angle at the dbh (actually, it is an approximate approach that the tree is viewed at its true diameter, but for this exercise, lets ignore the error produced by this assumption), and assume that the distance from the sampling point (O) to the tree center is the radius of the virtual sub-plot (r_i). These assumptions let us state the following triangle proportion:

$$d_i : r_i = b : c \tag{1}$$

It implies that any tree counted lies within a marginal circle whose radius is

$$r_i = \frac{cd_i}{b} \tag{2}$$

The respective sub-plot area F is

$$F_i = \pi r_i^2 = \pi \frac{c^2 d_i^2}{b^2} \tag{3}$$

Now, let's expand this per-sub-plot observation to a per-acre dimension. Remember that the virtual sub-plot of a tree has to include the sampling point to be counted, then this particular tree with diameter d_i and sub-plot with area F_i at the per-acre base has a probability of selection of $F_i/43560$. The basal area of the ith tree is

$$g_i = \frac{\pi}{4} d_i^2 \tag{4}$$

From the relationship of ratios:

$$\frac{\text{Basal area of the stand per acre}\,(G)}{g_i} = \frac{\text{Area of one acre}}{F_i} \tag{5}$$

$$\frac{G}{\frac{\pi}{4}d_i^2} = \frac{43560\,\text{sq.ft}}{\pi\frac{c^2 d_i^2}{b^2}} \tag{6}$$

$$G = 43560\left(\frac{b^2}{4c^2}\right) \tag{7}$$

From Figure 1, $tan(\alpha) = \frac{b/2}{c}$, then

$$G = 43560\,(tan(\alpha))^2 \tag{8}$$

G is the definition of BAF for each tree count.

1.1 BAF for $\theta = 90$

When the angle gauge is $\theta = 90°$, then $\alpha = 45°$ or $\pi/4$ rad. From the unitary circle (see Figure 2), we know that $sin(45°) = cos(45°) = 1/\sqrt{2}$. And also, by trigonometric identities, $tan(45°) = sin(45°)/cos(45°) = 1$ then

$$G = 43560(tan(45°))^2 = 43560 \tag{9}$$

Using an angle gauge of 90°, the BAF = 43560 sq.ft. That means that each tree counts represent 43560 sq ft. or an acre of basal area.

1.2 BAF for $\theta = 180°$

When the angle gauge is $\theta = 180°$, then $\alpha = 90°$ or $\pi/2$ rad. From the unitary circle (see Figure 3), we know that $sin(90°) = 1, and cos(90°) = 0$. Then by the trigonometric identities, $tan(90°) = sin(90°)/cos(90°) = \frac{1}{0}$. Therefore when half of the gauge angle is exactly 90°, $tan(90°)$ is undefined, and so is the BAF. As the gauge angle approaches 180°, the BAF becomes infinite (inf).

1.3 BAF for $\theta = 270°$

Since the tangent function is periodic, we can represent $tan(270°)$ as, $tan(270°) = tan(90° + n180°), n \in Z$. Then $tan(90°) = tan(270°) =$ undefined(inf).

2 Solution to the practical question

Given the ponderosa data with information by sections, I derived a taper equation from the volume information by sections. That is, Instead of modeling cross-sectional stem areas, I propose modeling the relative accumulative volume. To do that, I assumed that

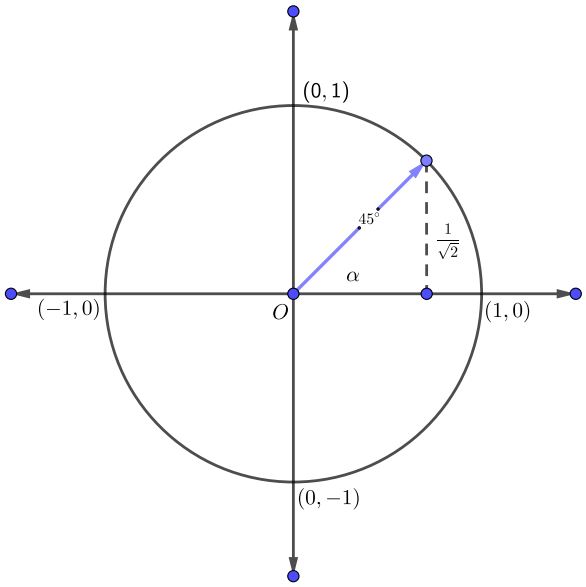

Figure 2: Illustration of the unit circle and an angle of 45°. The angle $\alpha = 45°$ is half of the gauge angle $\theta = 90°$.

the volume from the stump to the DBH follows a cylindrical form, and the section from the top last measurement to the tip follows a conical section.

The model considers the relationship between the accumulative volume from the stump and relative height. It is known that this relationship follows a convex curve.

Trees with missing total hight were excluded from the analysis.

Let's define $R_{z(p)} = v/Vt$, where v (cu.ft) is the cumulative stem volume from the stump to the height h (ft) above ground, Vt (cu.ft) is the total stem volume, and $p = h/H$, where H is total tree height with $0 \leq h \leq H$. A simple initial model for $R_{z(p)}$ could be a cubic polynomial equation with constraints to force it to take the value 1 when $p = 1$ and the value of 0 when $p = 0$. The initial model is:

$$R_{z(p)} = a + b * p + c * p^2 + d * p^3 \tag{10}$$

Considering $R_{z(p)} = 1$ when $p = 1$. Then $1 = a + b + c + d$, or $a = 1 - b - c - d$. Substituting in (Eq. 10):

$$R_{z(p)} = 1 - b - c - d + bp + cp^2 + dp^3 \tag{11}$$

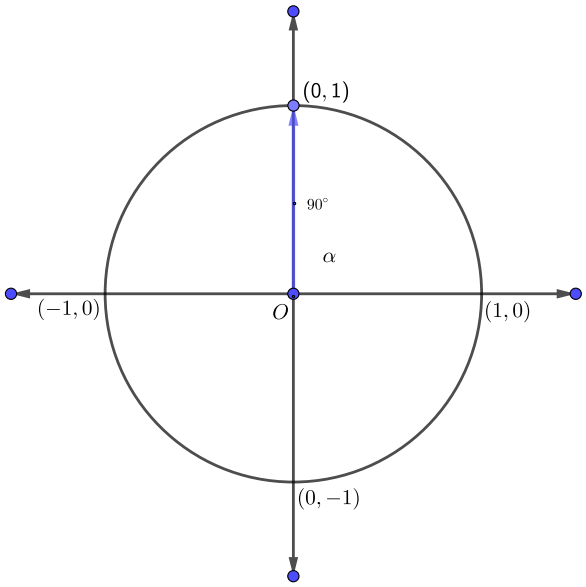

Figure 3: Illustration of the unit circle and a an angle of 90°. The angle $\alpha = 90°$ is half of the gauge angle $\theta = 180°$.

$$R_{z(p)} = 1 + b(p-1) + c(p^2-1) + d(p^3-1) \tag{12}$$

Now I need to constraint $R_{z(p)} = 0$ when $p = 0$, or $0 = 1 - b - c - d$, or $b = 1 - c - d$. Substituting in (Eq. 12) I got:

$$R_{z(p)} = 1 + (1-c-d) * (p-1) + c(p^2-1) + d(p^3-1) \tag{13}$$

$$R_{z(p)} = 1 + (p-1) - c(p-1) - d(p-1) + c(p^2-1) + d(p^3-1) \tag{14}$$

simplifying

$$R_z(p) = p + c(p^2-p) + d(p^3-p) \tag{15}$$

Now I need to guarantee that $\frac{dR_{z(p)}}{dh} \geq 0$, that is, the relationship always is convex. Lets check the firs derivative $\frac{dR_{z(p)}}{dh}$:

$$\frac{dR_{z(p)}}{dh} = \frac{1}{H} + 2c(p)\frac{1}{H} - \frac{c}{H} + 3d(p^2)\frac{1}{H} - \frac{d}{H} \tag{16}$$

$$\frac{dR_{z(p)}}{dh} = \frac{1}{H}\left(1 + 2cp - c + 3dp^2 - d\right) \tag{17}$$

$$\frac{dR_{z(p)}}{dh} = \frac{1}{H}\left(1 + c(2p-1) + d(3p^2 - 1)\right) \tag{18}$$

The constraint holds when the parameter $d \geq 0$ (for this model, d is always positive). An additional constrain comes from equaling the derivative to 0 and doing $p = 1$. Then $c = -1 - 2d$. Substituting it again in $R_{z(p)}$ I got:

$$R_{z(p)} = p + (-1 - 2d)(p^2 - p) + d(p^3 - p) \tag{19}$$

$$R_{z(p)} = p - (p^2 - p) - 2d(p^2 - p) + d(p^3 - p) \tag{20}$$

$$R_{z(p)} = p - p^2 + p + dp(p^2 - 2p + 1) \tag{21}$$

$$R_{z(p)} = 2p - p^2 + dp(p-1)^2 \tag{22}$$

Eq. 22 is our ratio-volume equation. This equation only has one parameter d ($d > 0$).

I will now develop an expression for the **taper model**. Let's use the definition for merchantable volume from a volume ratio function and equaling with the integral on the taper $D(h)$:

$$V_t \times R_{z(p)} = \int_0^h k * [D(h)]^2 dh \tag{23}$$

where $D(h)$ is the taper function for outside bark diameter (inches) at upper-stem height h (ft) from ground line (or stump height), k is a constant ($k = \pi/576$). Now differentiate both sides with respect to upper-stem height:

$$V_t \frac{dR_{z(p)}}{dh} = k * [D(h)]^2 \tag{24}$$

Solving for $D(h)$ yields the **taper function**:

$$D(h) = \sqrt{\frac{V_t}{k} * \frac{dR_{z(p)}}{dh}} \tag{25}$$

Using (Eq. 22) and changing p by h/H, I got:

$$\frac{dR_{z(p)}}{dh} = \frac{2}{H} - 2\left(\frac{h}{H}\right)\left(\frac{1}{H}\right) + \frac{d}{H}\left(\frac{h}{H} - 1\right)^2 + 2d\left(\frac{h}{H}\right)\left(\frac{h}{H} - 1\right)\left(\frac{1}{H}\right) \tag{26}$$

$$\frac{dR(p)}{dh} = \frac{1}{H}\left[2 - 2\left(\frac{h}{H}\right) + d\left(\frac{h}{H} - 1\right)^2 + 2d\left(\frac{h}{H}\right)\left(\frac{h}{H} - 1\right)\right] \tag{27}$$

According to equation (25), the **taper equation** then becomes:

$$d(h) = \left[\frac{V_t}{k} * \frac{1}{H}\left[2 - 2\left(\frac{h}{H}\right) + d\left(\frac{h}{H} - 1\right)^2 + 2d\left(\frac{h}{H}\right)\left(\frac{h}{H} - 1\right)\right]\right]^{0.5} \quad (28)$$

The taper is deducted as an implied equation from the volumen×ratio-equation, and the final model only uses **two** parameter: d and total tree volume V_t.

2.1 Model fitting

I used a linear regression model to fit equation 12.

I defined the following variables:

- $R_{z(p)} = v/V_t - 1$,
- $x1 = p - 1$,
- $x2 = p^2 - 1$,
- $x3 = p^3 - 1$

The following are the results of fitting this model using R (see the Appendix for the code used):

```
Call:
lm(formula = Rz ~ x1 + x2 + x3 - 1, data = ponderosa3)

Residuals:
Min        1Q    Median        3Q       Max
-0.282852 -0.006754  0.000000  0.004982  0.251911

Coefficients:
Estimate Std. Error t value Pr(>|t|)
x1  2.355964   0.002504  941.04   <2e-16 ***
x2 -1.599410   0.005494 -291.10   <2e-16 ***
x3  0.247951   0.003258   76.11   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0208 on 77766 degrees of freedom
Multiple R-squared:  0.998,Adjusted R-squared:  0.998
F-statistic: 1.314e+07 on 3 and 77766 DF,  p-value: < 2.2e-16
```

```
> extractAIC(pond_model)
[1]       3.0 -602372.3

# mean square error
> mean((ponderosa3$Rz - predict(pond_model))^2)
[1] 0.0004325823
```

Our median residual value is centered around zero, as this would tell us our residuals were somewhat symmetrical and that the model is predicting evenly at both the high and low ends of volume ratios. Our residual distribution looks symmetric but does not follow a normal distribution. We can visualize this with a quantile-quantile plot (see Fig 4). Looking at the chart below, you can see few outliers on both ends of the chart. And the distribution has heavy tails compared with the normal distribution. Overall this is a violation of the linear regression assumptions. One possible explanation is the restriction on our model to pass by 1 when $p = 1$ and take the value of 0 when $p = 0$. However, the fitted line looks to capture the theoretical convex pattern needed for the model construction.

The estimated value for the d parameter is 0.247951 with a standard error of 0.003258. the coefficient is large in comparison to its standard error, then statistically, the coefficient will most likely not be zero.

The residual standard error measures how well the model fits the data. In this case, it is 0.0208. The average value in R_z is -0.3108418. In average, this value says that the model predictions is low and be off on average by 0.02.

The Adjusted R-squared value shows that the combinations of predictors explain 99.8% of the variation within our dependent variable (volume ratios). The plot of observed Vs predicted in Fig 5 can confirm this high value.

The AIC ($-2 * log(L) + 2 * edf$, $edf = 3$) is -602372.3

A few observations have standardized residuals greater than 3 in absolute value (see Fig 4). Additionally, there is no high leverage point in the data. That is, all data points have a leverage statistic below $2(p + 1)/n = 8/77769 = 0.0001028688$

The ANOVA analysis confirms that all the predictors are important in the model.

```
Analysis of Variance Table

Response: Rz
Df  Sum Sq Mean Sq    F value    Pr(>F)
x1            1 15710.1 15710.1 36315652.8 < 2.2e-16 ***
x2            1  1335.7  1335.7  3087585.4 < 2.2e-16 ***
x3            1     2.5     2.5     5792.7 < 2.2e-16 ***
Residuals 77766    33.6     0.0
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

In conclusion, the taper equation is

$$d(h) = \left[\frac{V_t}{k} * \frac{1}{H}\left[2 - 2\left(\frac{h}{H}\right) + d\left(\frac{h}{H} - 1\right)^2 + 0.495902\left(\frac{h}{H}\right)\left(\frac{h}{H} - 1\right)\right]\right]^{0.5} \quad (29)$$

The equation uses two parameters: $2 \times d = 0.495902$, and total tree volumen V_t. The coefficient of determination is 99.9%. $AIC = -602372.3$, and mean square error of 0.0004325823.

3 Appendix—

```
# load some libraries
library(tidyverse)
library(ggpubr)

# read the data
ponderosa <- read.table("ponderosa_taper.txt", header = T, sep = ",")
head(ponderosa, 10)

# create a unique tree-age id:
ponderosa %<>% mutate(
ID = paste(Tree, Age, TotalHeight, sep = "_")
) %>%
arrange(
Tree, Age, TotalHeight, DiamHeight
) %>%
drop_na()

# fix repeated heights
ponderosa2 <- ponderosa %>% group_by(Tree, Age, TotalHeight, DiamHeight) %>%
summarise(
d = mean(Diameter),
nd = n()
) %>%
mutate(
ID = paste(Tree, Age, TotalHeight, sep = "_")
)

## compute volume
ponderosa3 <- ponderosa2 %>% split(f= ~ID) %>%
map_dfr(function(x){
v0 <- NULL
```

```
v0 <- (x[1,"d"]^2)*0.005454 * x[1,"DiamHeight"]
if(nrow(x)>1){
for(i in 2:nrow(x)){
vp <- (((x[i,"d"] + x[i-1,"d"])/2)^2)*0.005454 *(x[i,"DiamHeight"] -
 x[i-1,"DiamHeight"])
v0 <- c(v0,vp)
}
}
x$vp <- unlist(v0)

## Add the volume of the tip
last <- x[nrow(x),]
last$DiamHeight <- last$TotalHeight
last$d <- 0
last$vp <- unlist(
(1/3)*(x[nrow(x),"d"]^2)*0.004545*(last$TotalHeight - x[nrow(x),"DiamHeight"])
)
x <- rbind(x, last)
x$V <- sum(x$vp)
x$v <- cumsum(x$vp)
x
})

# Create variables for regression
ponderosa3 <- ponderosa3 %>% mutate(
Rz = (v/V) - 1,
p = DiamHeight/TotalHeight,
x1 = p - 1,
x2 = (p^2) - 1,
x3 = (p^3) - 1
) %>%
dplyr::filter(
v >= 0
)

## Regression (eq(12))
pond_model <- lm(Rz~x1+x2+x3-1, data = ponderosa3)
summary(pond_model)
anova(pond_model)
par(mfrow = c(2, 2))
plot(pond_model)

par(mfrow = c(1, 1))
plot(pond_model$fitted, ponderosa3$Rz,
```

```
xlab = "Predicted", ylab= "Observed")
abline(0,1, col = "Red")

## AIC
extractAIC(pond_model)

# mean square error
mean((ponderosa3$Rz - predict(pond_model))^2)
```

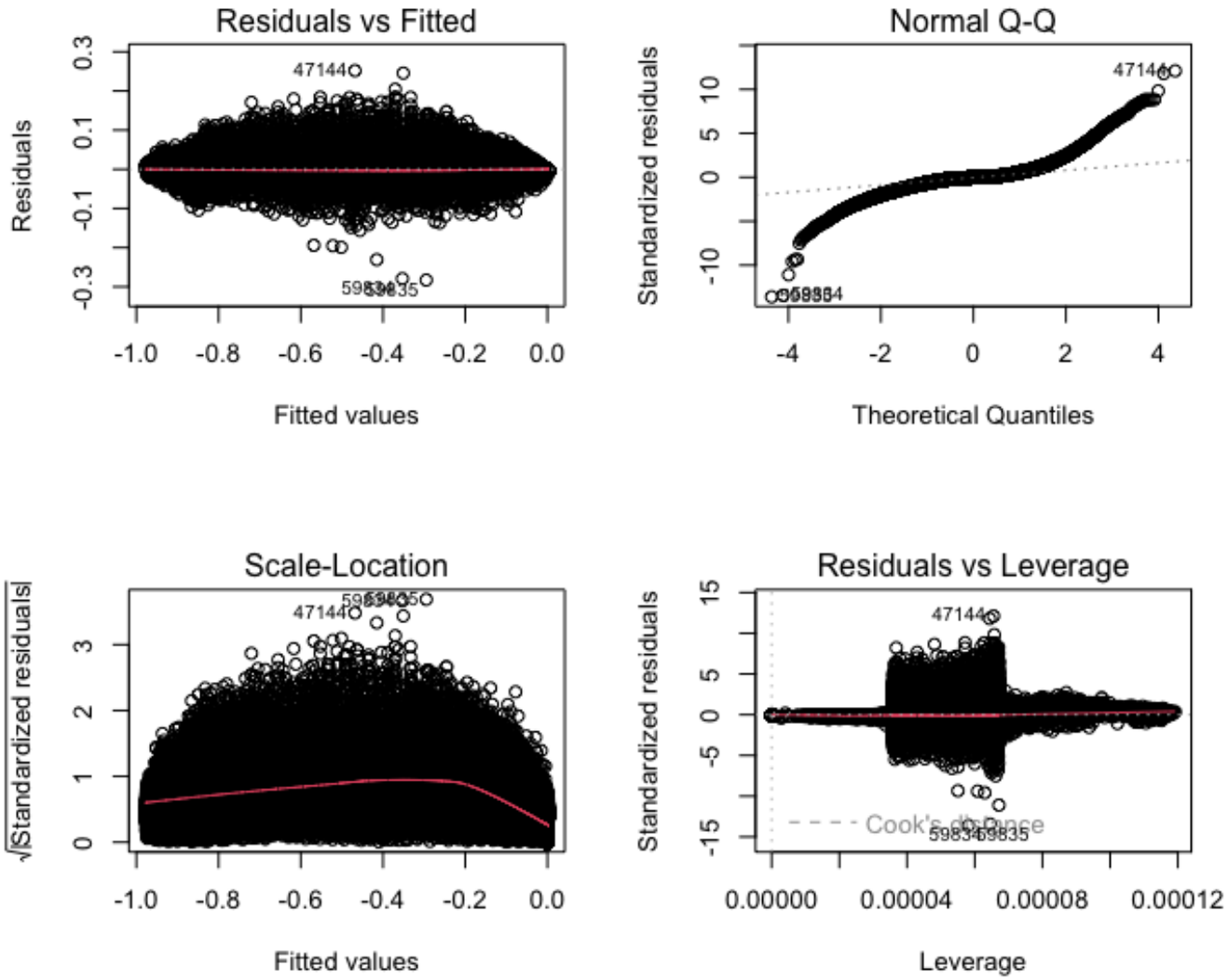

Figure 4: Regression diagnostics plots for the linear model in Eq 12 .

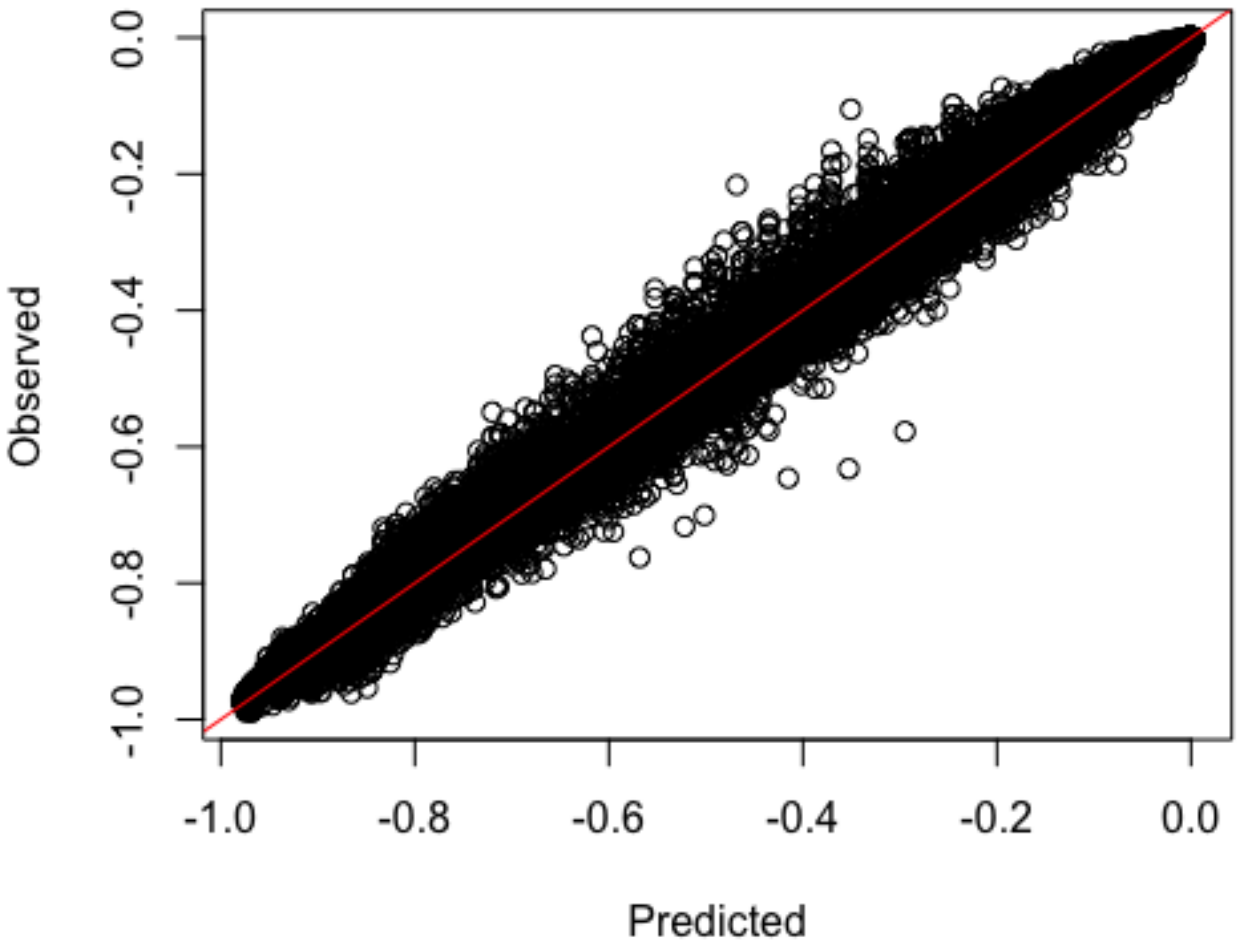

Figure 5: Observed $R_z = v/V_t - 1$ versus predicted by the model in equation 12. The red line indicate the 1:1 relation.