ST552 Homework 8

Brian Cervantes Alvarez

March 12, 2024

Problem 3

Given all these transformed models, the log-transformed model appears to be the best model for predicting volume from girth and height in the trees dataset. It has the highest explanatory power (highest R-squared) and the predictions are closest to the actual data (lowest RSE). Hence, compared to the original model, I would favor the log-transformed model. However, I would proceed with caution when using a transformed model, as it may require me to adjust my original research question, especially if that question was based on a different assumption or form of analysis.

```r
library(faraway)
data(trees)

# Linear model with original variables
model1 <- lm(Volume ~ Girth + Height, data = trees)
# Linear model with logarithmic transformation
logTrees <- log(trees)
model2 <- lm(Volume ~ Girth + Height, data = logTrees)
# Linear model with square root transformation
sqrtTrees <- sqrt(trees)
model3 <- lm(Volume ~ Girth + Height, data = sqrtTrees)
# Linear model with cube root transformation
cubeRootTrees <- (trees)^(1/3)
model4 <- lm(Volume ~ Girth + Height, data = cubeRootTrees)
# Compare model summaries
summary(model1)
```

```

Call:
lm(formula = Volume ~ Girth + Height, data = trees)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.4065 -2.6493 -0.2876  2.2003  8.4847 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
Girth         4.7082     0.2643  17.816  < 2e-16 ***
```

```
Height       0.3393     0.1302   2.607   0.0145 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-squared:  0.948, Adjusted R-squared:  0.9442
F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

```
summary(model2)
```

```

Call:
lm(formula = Volume ~ Girth + Height, data = logTrees)

Residuals:
      Min        1Q    Median        3Q       Max
-0.168561 -0.048488  0.002431  0.063637  0.129223

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.63162    0.79979  -8.292 5.06e-09 ***
Girth        1.98265    0.07501  26.432  < 2e-16 ***
Height       1.11712    0.20444   5.464 7.81e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08139 on 28 degrees of freedom
Multiple R-squared:  0.9777,    Adjusted R-squared:  0.9761
F-statistic: 613.2 on 2 and 28 DF,  p-value: < 2.2e-16
```

```
summary(model3)
```

```

Call:
lm(formula = Volume ~ Girth + Height, data = sqrtTrees)

Residuals:
     Min       1Q   Median       3Q      Max
-0.50788 -0.14043 -0.01882  0.25518  0.34851

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.6839     1.1625  -9.191 5.99e-10 ***
```

```
Girth         2.9826     0.1335   22.337   < 2e-16 ***
Height        0.5984     0.1538    3.891 0.000563 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2648 on 28 degrees of freedom
Multiple R-squared:  0.9675,    Adjusted R-squared:  0.9652
F-statistic: 417.4 on 2 and 28 DF,  p-value: < 2.2e-16
```

```r
summary(model4)
```

```
Call:
lm(formula = Volume ~ Girth + Height, data = cubeRootTrees)

Residuals:
      Min        1Q    Median        3Q       Max
-0.184031 -0.051048 -0.003254  0.081786  0.117567

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6.1677     0.6053 -10.189 6.35e-11 ***
Girth         2.5882     0.1077  24.029  < 2e-16 ***
Height        0.7327     0.1651   4.438 0.000128 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0923 on 28 degrees of freedom
Multiple R-squared:  0.9722,    Adjusted R-squared:  0.9703
F-statistic: 490.5 on 2 and 28 DF,  p-value: < 2.2e-16
```

Problem 4

Part A

In the additive model, H2S and Lactic were significant predictors with an R-squared value of 0.6518, indicating that about 65.18% of taste variability was explained. I applied data transformations logarithmic, square root, and cube root: the logarithmic transformation decreased explanatory power (R-squared = 0.4924), while square and cube root transformations maintained significance for H2S and Lactic with R-squared values of 0.6327 and 0.6015, respectively. The square root model appeared most effective, balancing high explanatory power with reduced residual variance, thus potentially improving model fit while preserving the significance of key predictors.

```r
data(cheddar)
model <- lm(taste ~ Acetic + H2S + Lactic, data = cheddar)
summary(model)
```

```
Call:
lm(formula = taste ~ Acetic + H2S + Lactic, data = cheddar)

Residuals:
    Min      1Q  Median      3Q     Max 
-17.390  -6.612  -1.009   4.908  25.449 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept) -28.8768    19.7354  -1.463  0.15540   
Acetic        0.3277     4.4598   0.073  0.94198   
H2S           3.9118     1.2484   3.133  0.00425 **
Lactic       19.6705     8.6291   2.280  0.03108 * 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.13 on 26 degrees of freedom
Multiple R-squared:  0.6518,	Adjusted R-squared:  0.6116 
F-statistic: 16.22 on 3 and 26 DF,  p-value: 3.81e-06
```

```r
# Log-Transformed Model
logCheddar <- log(cheddar)
model2 <- lm(taste ~ Acetic + H2S + Lactic, data = logCheddar)
summary(model2)
```

```
Call:
lm(formula = taste ~ Acetic + H2S + Lactic, data = logCheddar)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.4924 -0.2247  0.1769  0.5082  1.1364 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)   0.1368     3.0091   0.045   0.9641  
Acetic       -0.2856     2.0311  -0.141   0.8893  
H2S           1.6191     0.5908   2.740   0.0109 *
Lactic        1.2023     0.9749   1.233   0.2285  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8138 on 26 degrees of freedom
Multiple R-squared:  0.4924,	Adjusted R-squared:  0.4339 
F-statistic: 8.408 on 3 and 26 DF,  p-value: 0.0004513
```

```r
# Square Root Model
sqrtCheddar <- sqrt(cheddar)
model3 <- lm(taste ~ Acetic + H2S + Lactic, data = sqrtCheddar)
summary(model3)
```

```
Call:
lm(formula = taste ~ Acetic + H2S + Lactic, data = sqrtCheddar)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.47172 -0.60881  0.06132  0.82278  2.19358 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)  -5.6434     4.4767  -1.261  0.21865   
Acetic       -0.3743     2.4141  -0.155  0.87799   
H2S           2.3083     0.6966   3.314  0.00271 **
Lactic        4.6958     2.3329   2.013  0.05459 . 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.149 on 26 degrees of freedom
Multiple R-squared:  0.6327,	Adjusted R-squared:  0.5903 
```

```
F-statistic: 14.93 on 3 and 26 DF,  p-value: 7.52e-06
```

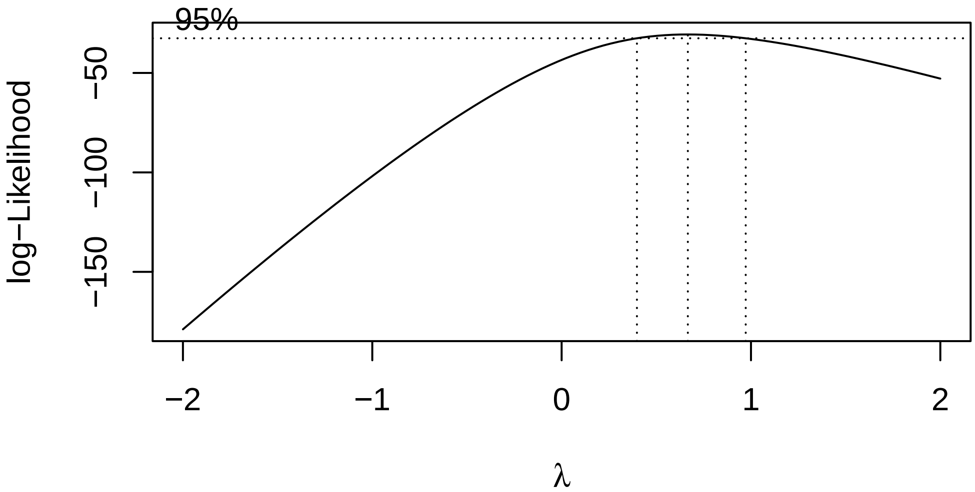
```r
# Linear model with cube root transformation
cubeCheddar <- (cheddar)^(1/3)
model4 <- lm(taste ~ Acetic + H2S + Lactic, data = cubeCheddar)
summary(model4)
```

```

Call:
lm(formula = taste ~ Acetic + H2S + Lactic, data = cubeCheddar)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.28089 -0.22503  0.05775  0.37511  0.82227 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)  -3.4865     2.9884  -1.167  0.25393   
Acetic       -0.3927     2.1512  -0.183  0.85657   
H2S           1.9979     0.6239   3.202  0.00358 **
Lactic        2.9642     1.6472   1.800  0.08355 . 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5111 on 26 degrees of freedom
Multiple R-squared:  0.6015,	Adjusted R-squared:  0.5556 
F-statistic: 13.08 on 3 and 26 DF,  p-value: 2.119e-05
```

Part B

Based on the Box-Cox plot, which indicates a clear peak around $\lambda = (0.6, 0.7)$, I recommend transforming the response variable using the Box-Cox method with the optimal λ in this range. This transformation is advised because the optimal λ is significantly different from 1, suggesting that the original data do not follow a normal distribution closely enough. This could also address our issue with our wide spread of residuals.

```
library(MASS)
boxcoxResults <- boxcox(model)
```



Part C

The optimized model demonstrates a better fit with smaller residuals and a lower residual standard error compared to the original model, indicating more accurate predictions. Both models have similar R-squared values, suggesting they explain a comparable amount of variability in taste, though significant predictors have larger effect sizes in the original model. The direction of the `Acetic` acid coefficient differs between models, highlighting potential differences in underlying processes between the two types of cheese.

```
maxLogLik <- max(boxcoxResults$y)
optimalLambdaIndex <- which(boxcoxResults$y == maxLogLik)
optimalLambda <- boxcoxResults$x[optimalLambdaIndex]
print(paste0("Optimal Lambda = ", optimalLambda))
```

```
[1] "Optimal Lambda = 0.666666666666667"
```

```
# Linear model with optimal root transformation
optimalCheddar <- (cheddar)^(optimalLambda)
model5 <- lm(taste ~ Acetic + H2S + Lactic, data = optimalCheddar)
summary(model5)
```

```

Call:
lm(formula = taste ~ Acetic + H2S + Lactic, data = optimalCheddar)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.3395 -1.4806 -0.0576  1.5319  5.2293 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)  -9.4775     7.0637  -1.342  0.19129   
Acetic       -0.2770     2.8520  -0.097  0.92337   
H2S           2.7169     0.8166   3.327  0.00263 **
Lactic        7.5017     3.4759   2.158  0.04033 * 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.417 on 26 degrees of freedom
Multiple R-squared:  0.6491,	Adjusted R-squared:  0.6086 
F-statistic: 16.03 on 3 and 26 DF,  p-value: 4.197e-06
```