Does Size Matter? (Estimation of Banana Weight with a Regression Modeling Approach)

Scott Graham, Kaisa Roggeveen

February 13, 2018

Summary

Introduction

The purpose of this study was to determine the most effective regression model to predict the weight of a banana using external measurements. This study also demonstrated multiple techniques for developing regression models. These models were then examined to demonstrate their effectiveness at creating regression models.

For the purpose of this study, it is assumed $\alpha = 0.05$ for all tests done.

Data Collection

First a small sample set bananas were purchased from the Real Canadian Superstore. The weight, length, diameter and circumference were then calculated using a scale and a ruler.

Table 1: Summary Statisitcs for Simulated R-Squared

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0004513	0.4005	0.6352	0.5719	0.7617	0.9812

In order to determine the minimum sample size needed, random sample sizes of 10 were generated using radius and length as the predictors. The correlation of the random sample sizes were calculated and a matrix of the correlations were generated. The value of the squared population multiple correlation coefficients with two predictor variables was then calculated and determined to be approximately 0.5719. From this the minimum sample size required was then determined from the table from Gregory T. Knofcznski's Sample Size When Using Multiple Linear Regression for Prediction, the minimum sample size was determined to be between 15 and 35, therefore the minimum number of bananas required was finalized at 24 bananas.

Analysis

Preliminary Analysis

Table 2: Banana Summary Statisitcs

Statistic	Weight	Radius	Length	Circumference
Min.	157.0000	18.30000	128.0000	115.0000
1st Qu.	179.5000	18.90000	243.2500	119.0000

Statistic	Weight	Radius	Length	Circumference
Median	193.5000	19.30000	256.5000	121.0000
Mean	193.7917	19.50417	250.2083	122.5417
3rd Qu.	203.2500	20.20000	268.5000	127.0000
Max	231.0000	20.50000	283.0000	129.0000

Figure 01: Sample Distributions of Banana Data

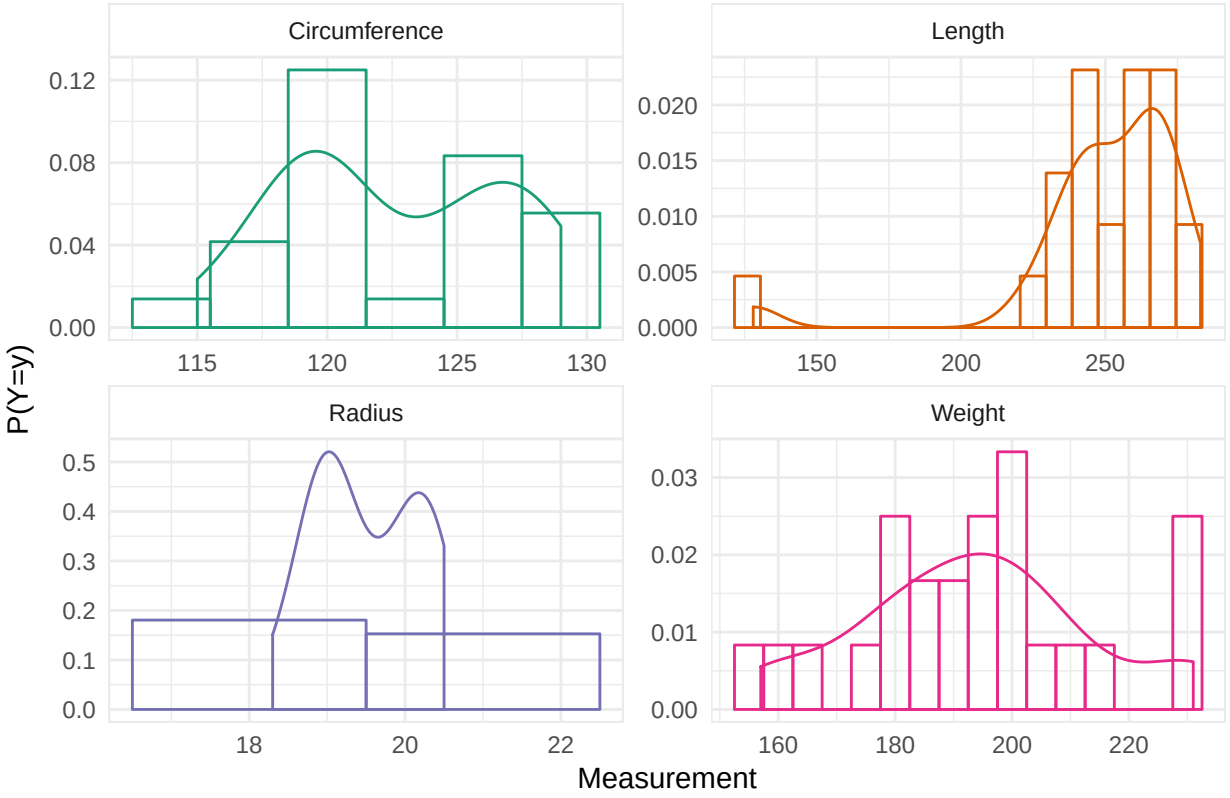
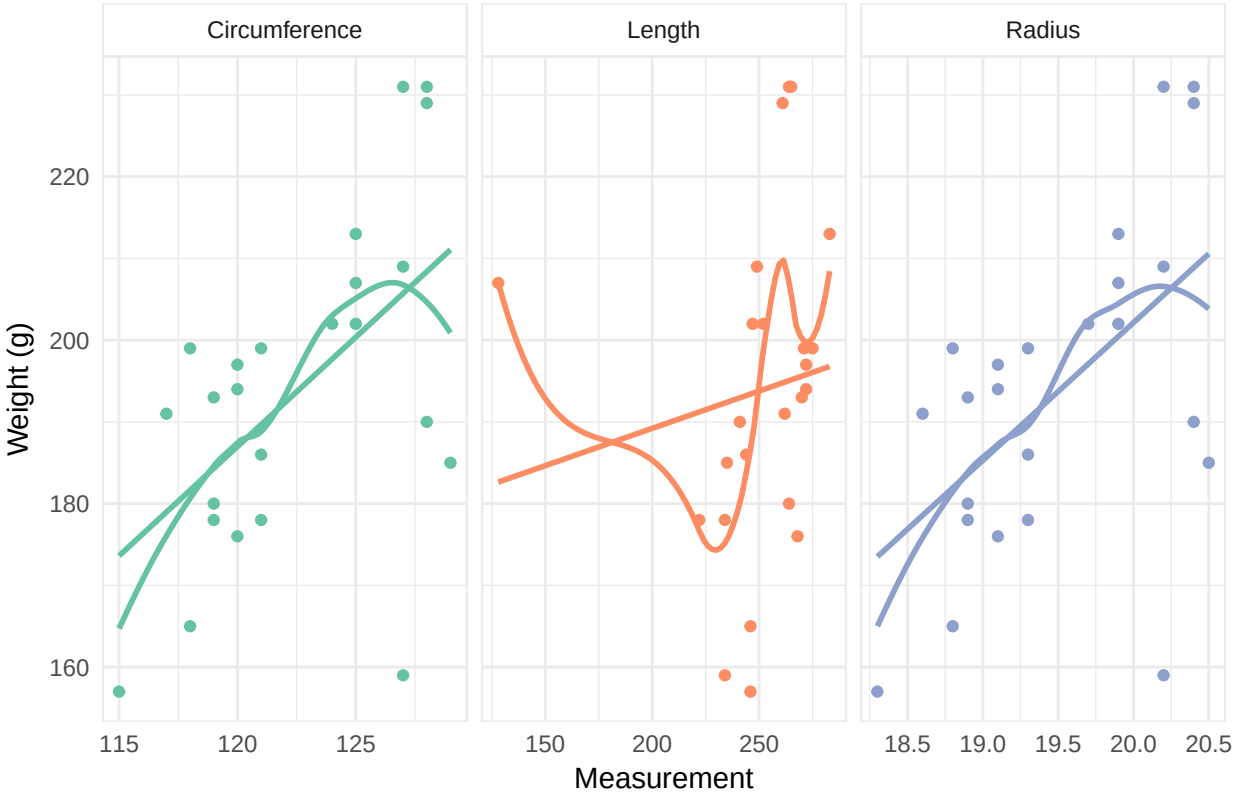


Figure 02: Weight vs. Predictors



Initial Regression Models

To begin analysis, a model using all predictor variables was created. In this case the density of the banana is assumed to be a constant. In the following models all measured bananas were considered.

Let:

$$W = \text{Weight (g)},\ L = \text{Length (mm)},\ R = \text{Radius (mm)},\ C = \text{Circumference (mm)}$$

Then:

$$\ln(W) = \beta_0 + \beta_1 \ln(L) + \beta_2 \ln(R) + \beta_3 \ln(C) \implies W = e^{\beta_0} \times L^{\beta_1} \times R^{\beta_2} \times C^{\beta_3} \tag{1}$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	10.06	26.32	0.3822	0.7063
Length_log	0.123	0.1275	0.9652	0.346
Radius_log	7.526	14.09	0.5341	0.5992
Circumference_log	-5.788	14.16	-0.4088	0.687

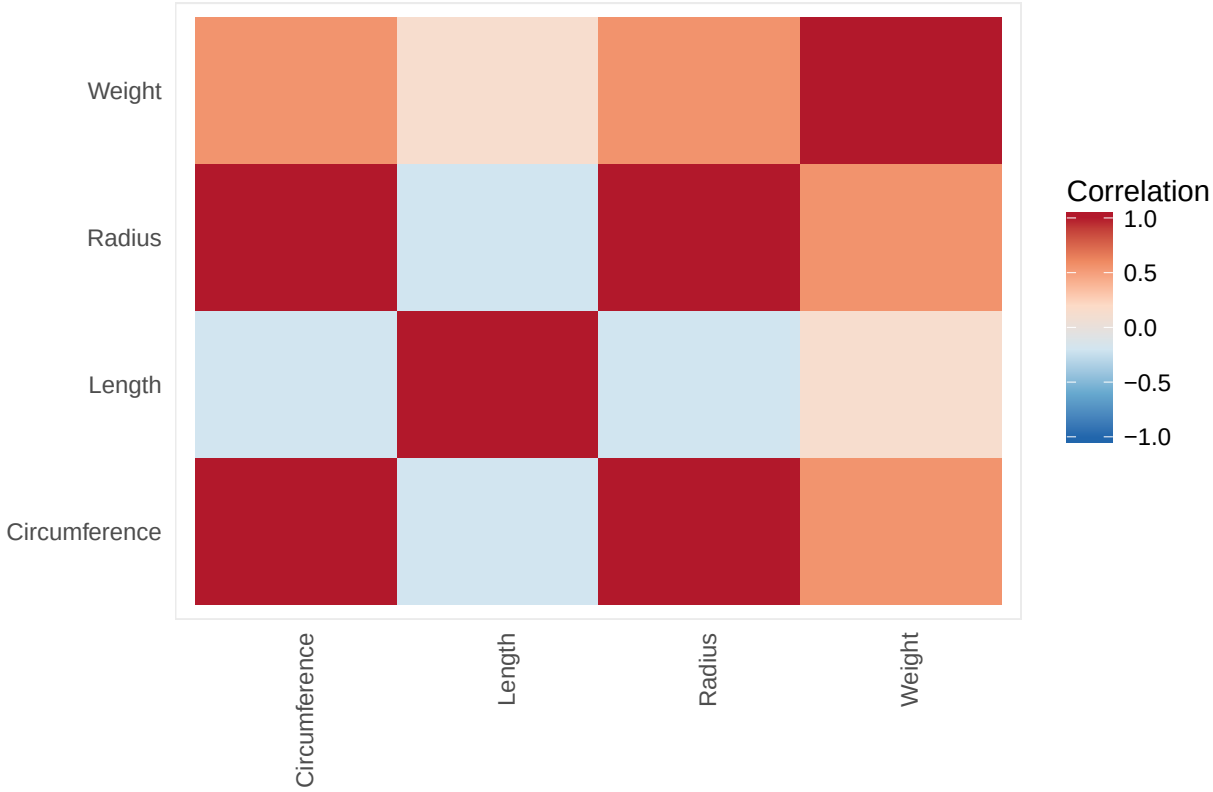
Table 4: Fitting linear model: Weight_log ~ Length_log + Radius_log + Circumference_log

Observations	Residual Std. Error	R^2	Adjusted R^2
24	0.09248	0.3318	0.2316

Observations	Residual Std. Error	R^2	Adjusted R^2

None of the predictor variables were found to be statistically significant. This is due to the high degree of collinearity exhibited between Radius and Circumference. As such, in the second model, the predictor variable, circumference, was removed. This is because $C = 2\pi R$, leading to collinearity. This can also be seen by examining the correlation plot produced by the variables:

Figure 03: Correlation Plot



The second model considered only the predictor variables length and radius:

$$\ln(W) = \beta_0 + \beta_1 \ln(L) + \beta_2 \ln(R) \tag{2}$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.6702	1.913	-0.3503	0.7296
Length_log	0.1223	0.1249	0.9787	0.3389
Radius_log	1.77	0.5596	3.163	0.004684

Table 6: Fitting linear model: Weight_log ~ Length_log + Radius_log

Observations	Residual Std. Error	R^2	Adjusted R^2
24	0.09062	0.3262	0.2621

The third model considered the predictor, length.

$$\ln(W) = \beta_0 + \beta_1 \ln(L) \tag{3}$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.99	0.8037	6.209	3e-06
Length_log	0.04917	0.1457	0.3374	0.739

Table 8: Fitting linear model: Weight_log ~ Length_log

Observations	Residual Std. Error	R^2	Adjusted R^2
24	0.1076	0.005146	-0.04007

The fourth model considered only one predictor, radius.

$$\ln(W) = \beta_0 + \beta_2 \ln(R) \tag{4}$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.3046	1.632	0.1867	0.8536
Radius_log	1.669	0.5494	3.038	0.006043

Table 10: Fitting linear model: Weight_log ~ Radius_log

Observations	Residual Std. Error	R^2	Adjusted R^2
24	0.09054	0.2955	0.2635

ANOVAs were then done to determine if a statistically significant difference in levels of explained variation existed between the 4 models. 1st, 01 and 02 were compared, testing the effect of Circumference:

Table 11: Analysis of Variance Table: Model 01 vs. Model 02

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
21	0.1725	NA	NA	NA	NA
20	0.171	1	0.001429	0.1671	0.687

Based on the data, we failed to reject the null hypothesis of equal explained variance between the two models. As such, we believe the model 02 is more appropriate, as it is smaller and more parsimonious.

The following test compares models 02 and 03, testing the effect of Radius:

Table 12: Analysis of Variance Table: Model 02 vs. Model 03

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
22	0.2547	NA	NA	NA	NA

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
21	0.1725	1	0.08219	10.01	0.004684

Based on the data, we rejected the null hypothesis of equal explained variance between the two models. As such, we believe the model 02 is more appropriate, as it does a better job at explaining the underlying variance in the data.

The following test compares models 02 and 03, testing the effect of Length:

Table 13: Analysis of Variance Table: Model 02 vs. Model 04

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
22	0.1803	NA	NA	NA	NA
21	0.1725	1	0.007867	0.9578	0.3389

Based on the data, we failed to reject the null hypothesis of equal explained variance between the two models. As such, we believe the model 04 is more appropriate, as it is smaller and more parsimonious.

From this, we found the most appropriate model was model 04, as it is the smallest model that explains an adequate amount of variance. This model being:

$$\ln(W) = 0.3046 + 1.6690 \ln(R) \implies W = e^{0.3046} R^{1.6690}$$

This is an unexpected result, as it doesn't include the Length predictor, implying that the weight of a banana is purely a function of its radius. Due to this, further analysis was required.

Removal of Outliers

Due to the surprising results obtained above, the choice to check for potential outliers was made. For this, model 02 was chosen, as we wished to examine the potential presence of outlier with respect to both Length and Radius, which model 04 would fail to accomplish.

Figure 04: Standardized Residuals vs. Predicted for Model 02

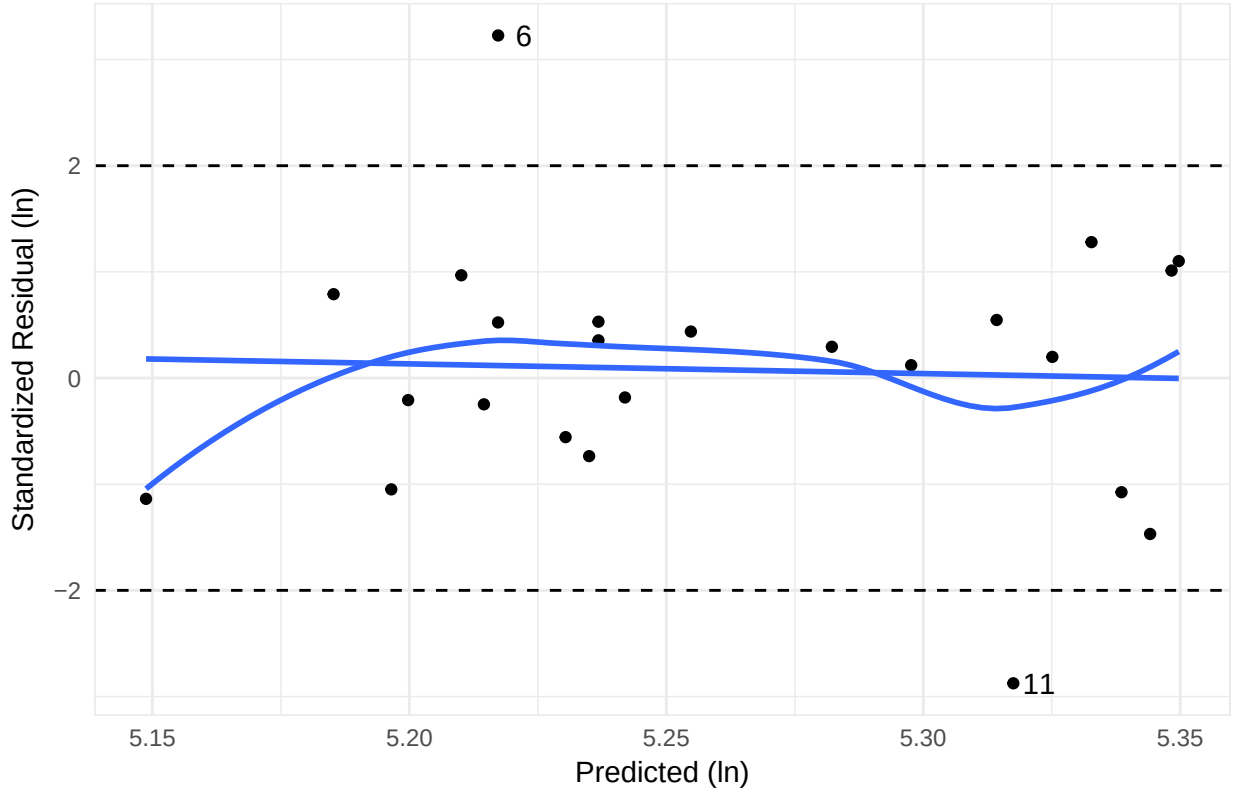


Figure 05: Standardized Residuals vs. Leverage for Model 02

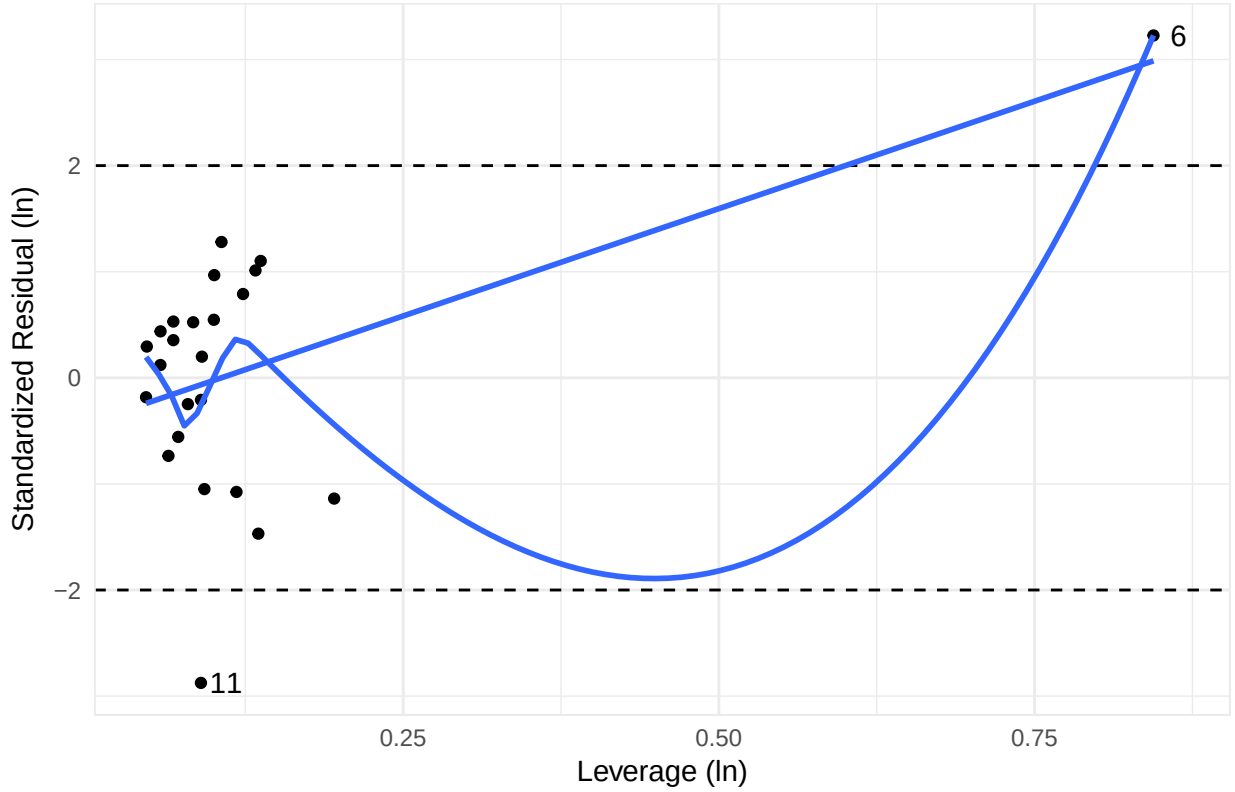


Table 14: Entries with a |Standardized Residual| >2

ID	Weight	Radius	Length	Circumference	Std Residuals	Leverage
6	207	19.9	128	125	3.226762	0.8441502
11	159	20.2	234	127	-2.875567	0.0898602

By looking at values with |Std. Residual| > 2, we can identify potential outliers. Then, by examining the leverage of those outliers we can make a determination of whether or not they are severely distorting the regression line from the true slope. Although both 6 and 11 have a large standardized residual, only 6 has a large leverage associated with it, and as such is excluded from the dataset.

Cross Validation

```
## Analysis of Variance Table
##
## Response: Weight_log
##             Df Sum Sq Mean Sq F value  Pr(>F)
## Length_log   1 0.0643  0.0643    14.8 0.00101 **
## Radius_log   1 0.0994  0.0994    22.9 0.00011 ***
## Residuals   20 0.0870  0.0043
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
```

```
## fold 1
## Observations in test set: 7
##                 [,1]    [,2]  [,3]  [,4]   [,5]   [,6]  [,7]
## Predicted    5.2865  5.2865 5.111 5.098 5.3693  5.235 5.268
## cvpred       5.2964  5.2964 5.059 5.026 5.3739  5.232 5.242
## Weight_log   5.2832  5.2679 5.182 5.182 5.4424  5.193 5.308
## CV residual -0.0132 -0.0285 0.123 0.155 0.0685 -0.039 0.066
##
## Sum of squares = 0.05    Mean square = 0.01    n = 7
##
## fold 2
## Observations in test set: 8
##                  [,1]  [,2]    [,3]   [,4]    [,5]   [,6]   [,7]    [,8]
## Predicted    5.19520  5.24  5.0995  5.275 5.30308  5.291 5.2668  5.3053
## cvpred       5.23493  5.34  5.0927  5.381 5.28996  5.381 5.2254  5.3711
## Weight_log   5.22575  5.07  5.0562  5.220 5.29330  5.247 5.2933  5.3423
## CV residual -0.00918 -0.27 -0.0364 -0.161 0.00334 -0.134 0.0679 -0.0287
##
## Sum of squares = 0.12    Mean square = 0.02    n = 8
##
## fold 3
## Observations in test set: 8
##                [,1]    [,2]   [,3]   [,4]   [,5]   [,6]   [,7]     [,8]
## Predicted    5.152  5.4075 5.3729 5.3847  5.271 5.2685 5.1961  5.25829
## cvpred       5.163  5.3975 5.3550 5.3667  5.276 5.2629 5.2099  5.26643
## Weight_log   5.106  5.3613 5.4337 5.4424  5.170 5.3083 5.2523  5.26269
## CV residual -0.057 -0.0363 0.0787 0.0757 -0.105 0.0453 0.0423 -0.00374
##
## Sum of squares = 0.03    Mean square = 0    n = 8
##
## Overall (Sum over all 8 folds)
##      ms
## 0.00894
```

Table 15: CV Model (continued below)

ID	Weight	Radius	Length	Circumference	Weight_log	Radius_log
1	197	19.1	272	120	5.283	2.95
2	194	19.1	272	120	5.268	2.95
3	165	18.8	246	118	5.106	2.934
4	186	19.3	244	121	5.226	2.96
5	178	18.9	234	119	5.182	2.939
7	213	19.9	283	125	5.361	2.991
8	178	19.3	222	121	5.182	2.96
9	229	20.4	261	128	5.434	3.016
10	231	20.2	265	127	5.442	3.006
11	159	20.2	234	127	5.069	3.006
12	157	18.3	246	115	5.056	2.907
13	231	20.4	264	128	5.442	3.016
14	185	20.5	235	129	5.22	3.02
15	176	19.1	268	120	5.17	2.95
16	202	19.7	252	124	5.308	2.981
17	199	19.3	271	121	5.293	2.96
18	190	20.4	241	128	5.247	3.016

ID	Weight	Radius	Length	Circumference	Weight_log	Radius_log
19	180	18.9	264	119	5.193	2.939
20	191	18.6	262	117	5.252	2.923
21	202	19.9	247	125	5.308	2.991
22	193	18.9	270	119	5.263	2.939
23	199	18.8	275	118	5.293	2.934
24	209	20.2	249	127	5.342	3.006

Length_log	Circumference_log	Predicted	cvpred
5.606	4.787	5.286	5.296
5.606	4.787	5.286	5.296
5.505	4.771	5.152	5.163
5.497	4.796	5.195	5.235
5.455	4.779	5.111	5.059
5.645	4.828	5.408	5.398
5.403	4.796	5.098	5.026
5.565	4.852	5.373	5.355
5.58	4.844	5.369	5.374
5.455	4.844	5.241	5.338
5.505	4.745	5.099	5.093
5.576	4.852	5.385	5.367
5.46	4.86	5.275	5.381
5.591	4.787	5.271	5.276
5.529	4.82	5.269	5.263
5.602	4.796	5.303	5.29
5.485	4.852	5.291	5.381
5.576	4.779	5.235	5.232
5.568	4.762	5.196	5.21
5.509	4.828	5.268	5.242
5.598	4.779	5.258	5.266
5.617	4.771	5.267	5.225
5.517	4.844	5.305	5.371

MAE

Table 17: Calculated Error Terms for CV Model

MSE	MAE	MPAE
0.009	0.072	0.014

Recommendations

Using the first set of data before the outlier was removed, it can be determined that the best way to predict the weight of a banana is by measuring the radius of the banana. The model that is then used for banana weight prediction is the following:

$$\ln(W) = \beta_0 + \beta_1 \ln(R)$$

$$\ln(W) = 0.3046 + 1.669 \ln(R)$$

After the removal of the outlier, the model that was determined to be the best predictor for banana weight was the following:

Appendix

Appendix A: Code

```
library(pander, warn.conflicts = FALSE, quietly = TRUE)
library(MAAS, warn.conflicts = FALSE, quietly = TRUE)
library(DAAG, warn.conflicts = FALSE, quietly = TRUE)
library(tidyverse, warn.conflicts = FALSE, quietly = TRUE)
library(magrittr, warn.conflicts = FALSE, quietly = TRUE)
library(ggfortify, warn.conflicts = FALSE, quietly = TRUE)
library(knitr, warn.conflicts = FALSE, quietly = TRUE)

set.seed(5609)

theme_minimal2 <- theme_minimal() %>%  theme_set()
theme_minimal2 <-
  theme_update(
    panel.border = element_rect(
      linetype = "solid"
      ,colour = "grey92"
      ,fill = NA
    )
    ,strip.background = element_rect(
      linetype = "solid"
      ,colour = "grey92"
        ,fill = NA
      )
    )

banana_data <-
  "mybanana.txt" %>%
  read_tsv()

banana_data <-
  banana_data %>%
  mutate_at(
    .vars = vars(Weight:Circumference)
    ,.funs = funs(log = log)
  )

banana_tidy <-
  banana_data %>%
  select(
    -c(
      Weight_log
      ,Radius_log
      ,Length_log
      ,Circumference_log
    )
```

```
) %>%
gather(
  key = "Type"
  ,value = "Measurement"
  ,-ID
)
```