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

Scott Graham, Kaisa Roggeveen

February 13, 2018

Summary

This study developed linear regression models to effectively predict the weight of bananas based on external measurements. In order to develop the different models the techniques used were, determining significance of variables, elimination of variables and elimination of outliers. The models were then compared by ANOVA. This study concludes that the weight of the banana is dependent on the length and the radius of the banana.

Introduction

The weight of a banana is difficult to predict since it depends on density and volume, which are not easily measurable. In this study it was assumed that the density of the bananas remained consistent throughout the entire banana and that volume was approximately $V = \pi R^2 L$, where R was the radius of the body and L was the curved length. As such we have $W = VD = \pi R^2 LD$.

This report states the data collection methods, the different methods and techniques used for creating regression models, the analysis of the regression models and the recommendations for the best method to predict banana weight. For the purpose of this study, it is assumed $\alpha = 0.05$ for all tests done.

Data Collection

First a small sample set of 24 bananas were purchased from the Real Canadian Superstore. The weight, length, diameter and circumference were then calculated using a scale and a ruler. This data was recorded using an Excel spreadsheet for further analysis.

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 the table presented in Appendix B, the minimum sample size was determined to be between 15 and 35, therefore the minimum number of bananas required was finalized at 24 bananas (Knofczynski, 2007).

Analysis

Preliminary Analysis

To begin the analysis the summary statistics of all predictor variables were calculated and recorded in Table 2. The summary statistics were used to highlight the data to obtain a general understanding of the spread for each predictor variable.

Next graphs of the sampling distributions of the predictor variables were used to visualize the trends of the data in Figure 1. The graphs were examined for skewness and normality where, length appears to be highly skewed to the left, weight appears to be more normally distributed circumference and radius are very similarly distributed with potential bimodal distribution.

In Figure 2, correlation between weight and the predictor variables were produced. From the analysis of the graphs, circumference and radius appear to be much more highly correlated to weight than length.

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
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 1: Sample Distributions of Banana Data

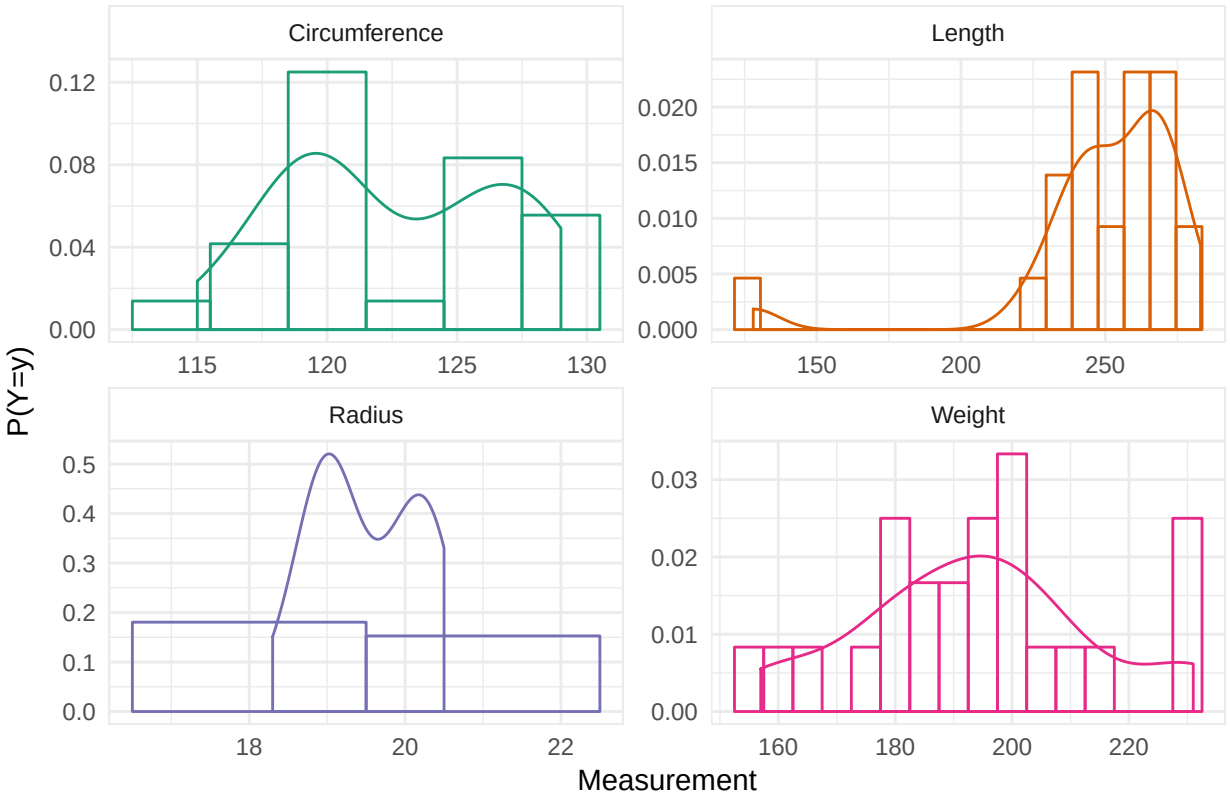
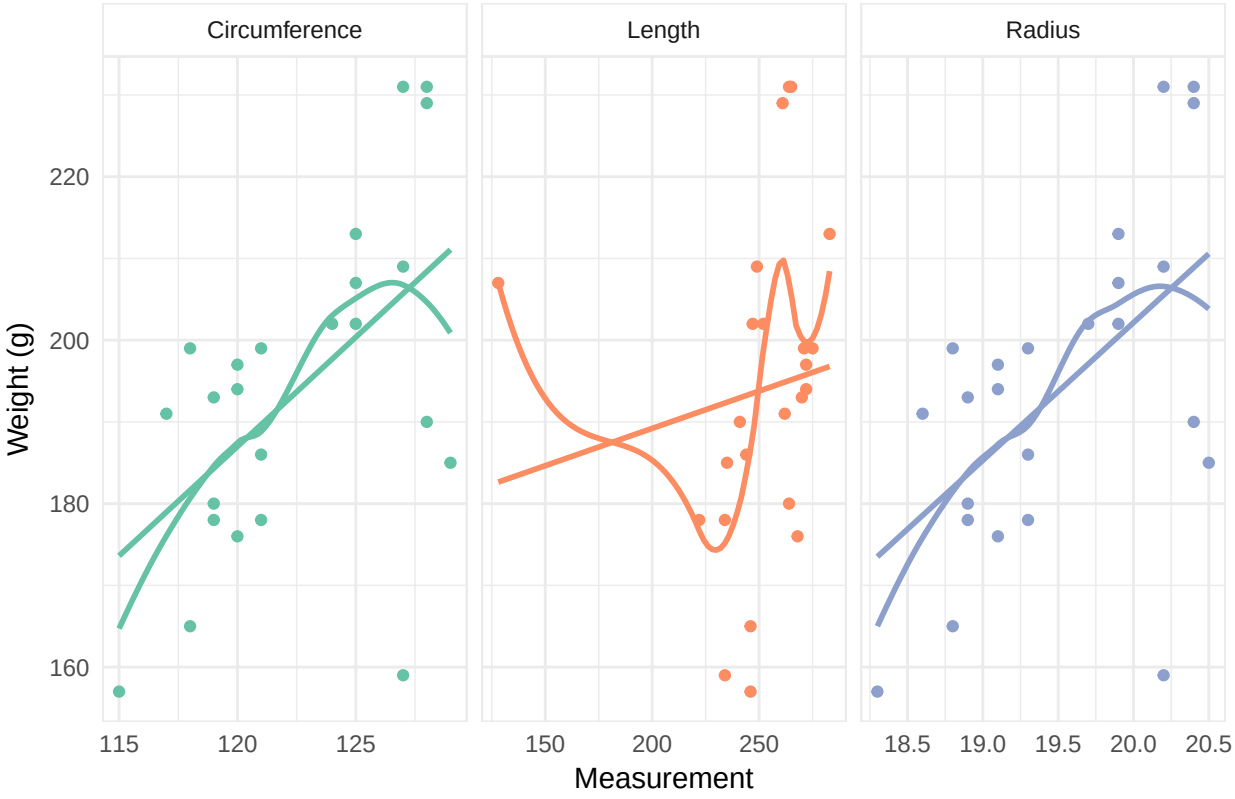


Figure 2: Weight vs. Predictors



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

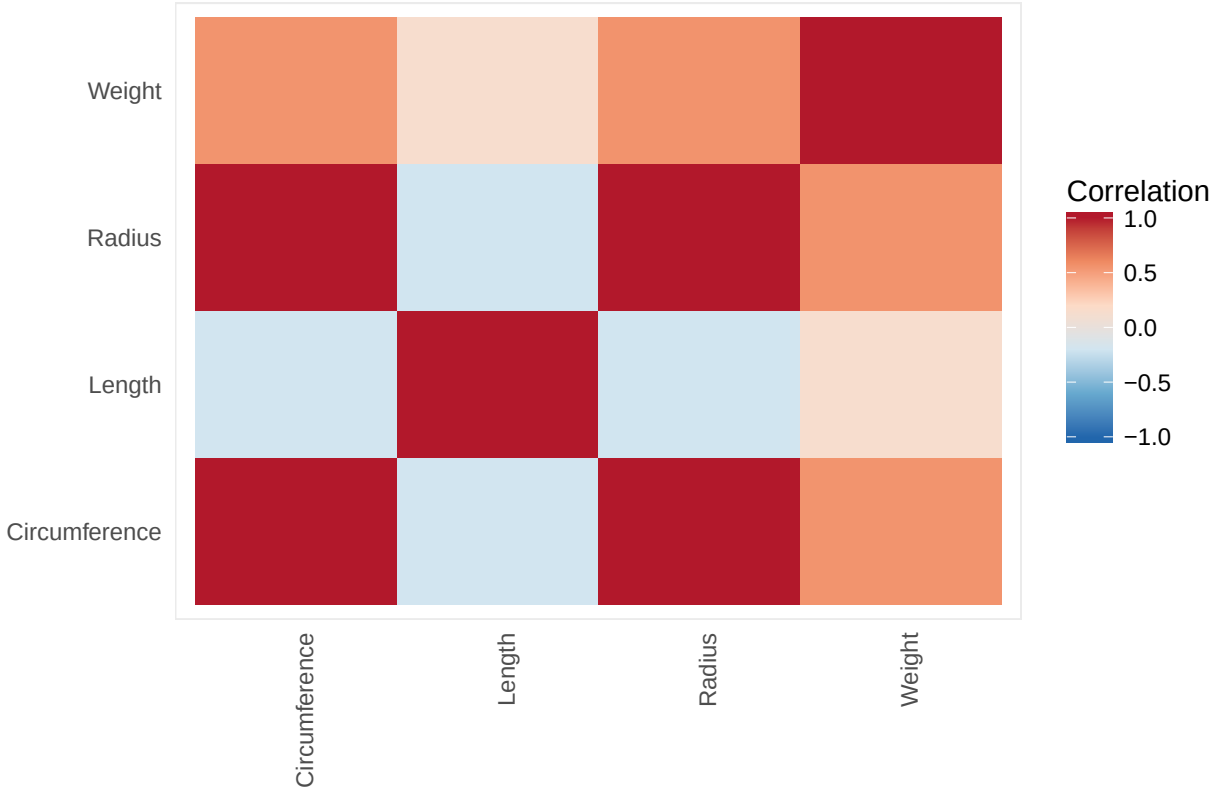
Table 4: Fitting linear model: Weight_log ~ Length_log + Radius_log + Circumference_log

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 in Figure 3:

Figure 3: Correlation Plot



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

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

This model is similar to the one described in the Introduction, with $e^{\beta_0} \approx \pi D$. as such, it'd be reasonable to expect it to perform reasonably well.

	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
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 4: Standardized Residuals vs. Predicted for Model 02

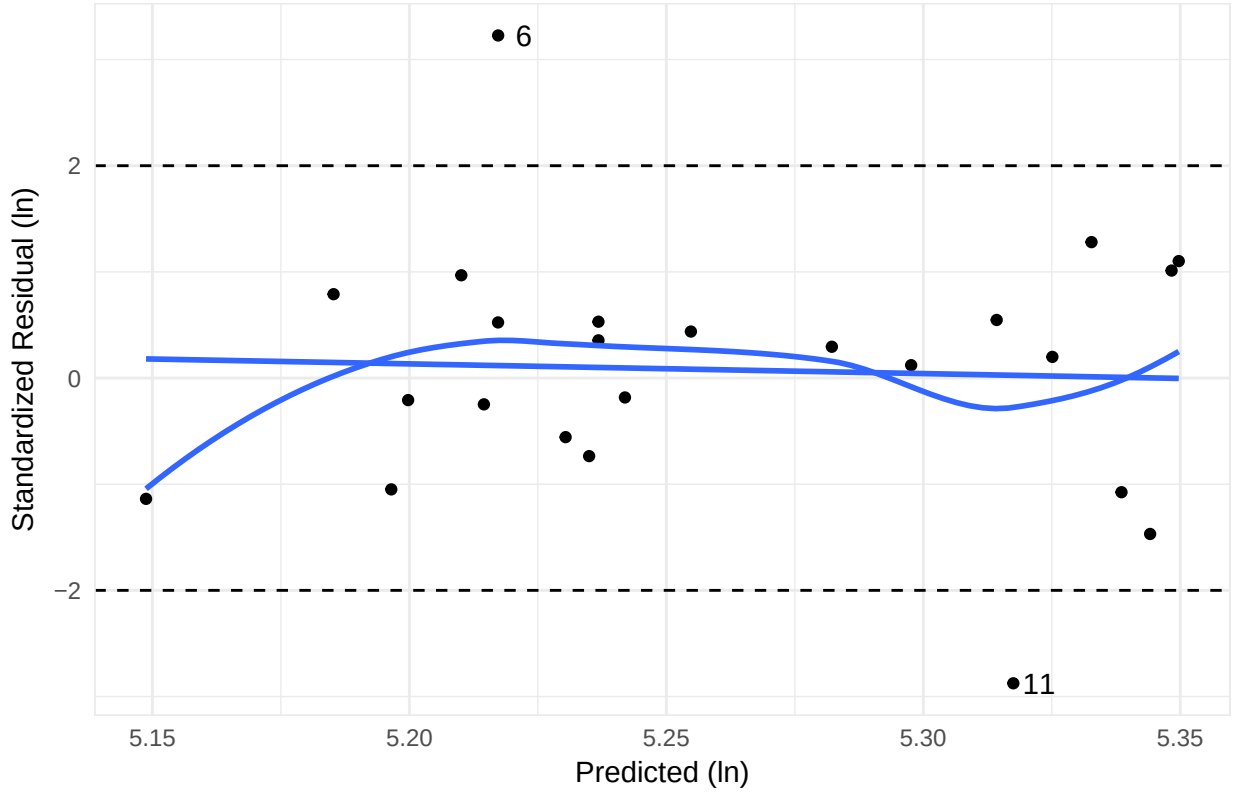


Figure 5: Standardized Residuals vs. Leverage for Model 02

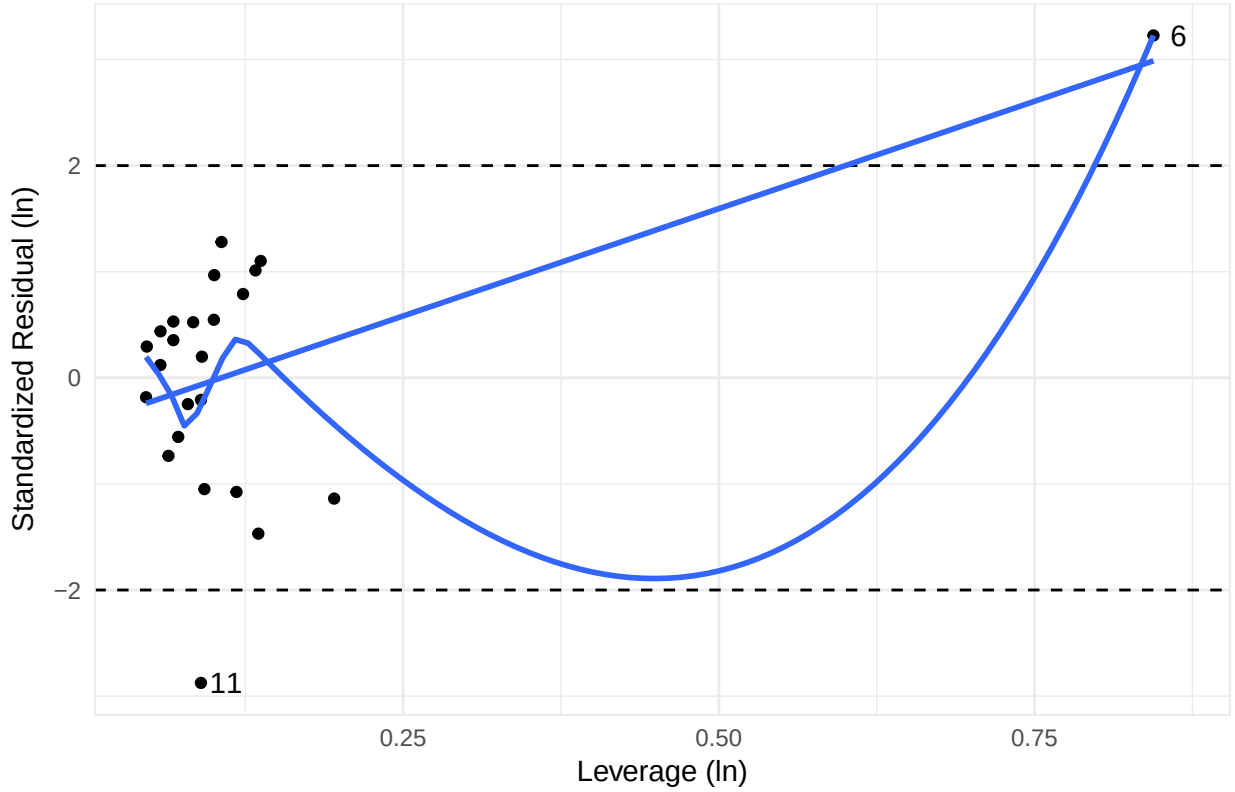


Table 14: Entries with a |Standardized Residual| >2

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 data set.

Cross Validation

We then performed cross validation as described in the `caret` package on Equation (2). For this 10-fold cross validation was used, and the following regression output was produced, using the post-outlier data.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-6.249	1.876	-3.331	0.003334
Length_log	1.028	0.2235	4.599	0.0001739
Radius_log	1.957	0.4093	4.781	0.0001138

Table 16: Fitting linear model: .outcome ~ .

Observations	Residual Std. Error	R^2	Adjusted R^2
23	0.06594	0.6531	0.6184

This gives the model:

$$W = e^{-0.6249} L^{1.028} R^{1.957} = 0.00193 L^{1.028} R^{1.957} \implies D = \frac{0.00193}{\pi} = 0.000615 g/cm$$

Which is very similar to that described in the Introduction section.

Mean Error

The error was measured using the cross validated predictions, in both terms of the $\ln(W)$, and the W. Firstly, we found the Mean Squared Error (MSE), Mean Squared Error (MAE), and Mean Percent Absolute Error (MPAE). The residual was calculated as:

$$y_i - \hat{y}_i$$

Table 17: Calculated Error Terms for Log CV Model

MSE	MAE	MPAE	RMSE
0.009	0.072	0.014	0.095

For this, the residual term was put back into grams, and was calculated as following:

$$e^{y_i} - e^{\hat{y}_i}$$

Table 18: Calculated Error Terms for CV Model

MSE	MAE	MPAE	RMSE
317	13.7	0.073	17.8

In both cases the errors are quite small, especially considering the sample size, and lend to the conclusion that the model is an acceptable model.

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:

$$\ln(W) = \beta_0 + \beta_1 \ln(R) = 0.3046 + 1.669 \ln(R)$$

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

$$\ln(W) = \beta_0 + \beta_1 \ln(L) + \beta_2 \ln(R) \implies W = 0.00193 L^{1.028} R^{1.957} \approx 0.000615 \pi L R^2$$

References

Knofczynski, G. T., & Mundfrom, D. (2007). Sample Sizes When Using Multiple Linear Regression for Prediction. Educational and Psychological Measurement, 68(3), 431-442. doi:10.1177/0013164407310131

Appendix

Appendix A: Code

```
# Setup
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
  )

# Sample Size
rsquare_sim <- c()
for(i in 1:1000){
  banana_cor <-
    banana_data %>%
    sample_n(10) %>%
    select(
      Weight
      ,Radius
      ,Length
      # ,Circumference
    ) %>%
    cor()

  xy_vec <- banana_cor[2:3, 1]
  C_mat <- banana_cor[2:3, 2:3]

  rsquare_sim[i] <- t(xy_vec) %*% solve(C_mat) %*% xy_vec
}
rsquare_sim %>%
  summary() %>%
  pander(caption = "Summary Statisitcs for Simulated R-Squared")

# Summary Stats
banana_summary <-
  cbind(
    Statistic =
      c(
        "Min."
        ,"1st Qu."
        ,"Median"
        ,"Mean"
        ,"3rd Qu."
        ,"Max"
      )
    ,banana_data %>%
      select(
        -c(
          ID
          ,Weight_log
          ,Radius_log
          ,Length_log
```

```
          ,Circumference_log
        )
      ) %>%
      map_df(summary)
  ) %>%
  as.tibble()
kable(banana_summary, caption = "Banana Summary Statisitcs")


banana_tidy %>%
  ggplot(aes(x = Measurement, colour = Type)) +
  geom_histogram(
    aes(y = ..density..)
    ,alpha = 0
    ,binwidth = function(x) nclass.FD(x)
  ) +
  geom_density() +
  facet_wrap(
    ~ Type
    ,scales = "free"
  ) +
  scale_colour_brewer(
    palette = "Dark2"
    ,type = "qual"
  ) +
  labs(
    title = "Figure 1: Sample Distributions of Banana Data"
    ,y = "P(Y=y)"
  ) +
  theme(legend.position = "none")

# Visualizations
banana_data %>%
  select(
    -c(
      Weight_log
      ,Radius_log
      ,Length_log
      ,Circumference_log
    )
  ) %>%
  gather(
    key = "Type"
    ,value = "Measurement"
    ,-ID
    ,-Weight
  ) %>%
  ggplot(
    aes(
      x = Measurement
      ,y = Weight
      ,colour = Type
    )
```

```
  ) +
  geom_smooth(
    method = "loess"
    ,se = FALSE
  ) +
  geom_smooth(
    method = "lm"
    ,se = FALSE
  ) +
  geom_point() +
  facet_wrap(
    ~ Type
    ,scales = "free_x"
  ) +
  scale_colour_brewer(
    palette = "Set2"
    ,type = "qual"
  ) +
  labs(
    title = "Figure 2: Weight vs. Predictors"
    ,y = "Weight (g)"
  ) +
  theme(legend.position = "none")

# Reg 01
banana_reg_01 <-
  banana_data %>%
  lm(
    Weight_log ~ Length_log + Radius_log + Circumference_log
    ,data = .
  )
pander(summary(banana_reg_01))

# Cor Plot
banana_data %>%
  select(
    -c(
      ID
      ,Weight_log
      ,Radius_log
      ,Length_log
      ,Circumference_log
    )
  ) %>%
  cor() %>%
  as.data.frame() %>%
  rownames_to_column() %>%
  as.tibble() %>%
  gather(
    key = Column
    ,value = Correlation
    ,-rowname
  ) %>%
```

```
  rename(Row = rowname) %>%
  ggplot(
    aes(
      x = Column
      ,y = Row
      ,fill = Correlation
    )
  ) +
  geom_raster() +
  scale_fill_distiller(
    type = "div"
    ,palette = "RdBu"
    ,limits = c(-1, 1)
  ) +
  labs(title= "Figure 3: Correlation Plot") +
  theme(
    axis.text.x = element_text(angle = 90, hjust = 1)
    ,axis.title.x = element_blank()
    ,axis.title.y = element_blank()
    ,panel.grid = element_blank()
  )

# Reg 02
banana_reg_02 <-
  banana_data %>%
  lm(
    Weight_log ~ Length_log + Radius_log
    ,data = .
  )
pander(summary(banana_reg_02))

# Reg 03
banana_reg_03 <-
  banana_data %>%
  lm(
    Weight_log ~ Length_log
    ,data = .
  )
pander(summary(banana_reg_03))

# Reg 04
banana_reg_04 <-
  banana_data %>%
  lm(
    Weight_log ~ Radius_log
    ,data = .
  )
pander(summary(banana_reg_04))

# ANOVA 01 vs 02
anova(banana_reg_02, banana_reg_01) %>%
  pander(caption = "Analysis of Variance Table: Model 01 vs. Model 02")
```

```r
# ANOVA 01 vs 02
anova(banana_reg_03, banana_reg_02) %>%
  pander(caption = "Analysis of Variance Table: Model 02 vs. Model 03")

# ANOVA 01 vs 02
anova(banana_reg_04, banana_reg_02) %>%
  pander(caption = "Analysis of Variance Table: Model 02 vs. Model 04")

# Outlier Check
banana_resid_data <-
  tibble(
    Predicted = predict(banana_reg_02)
    ,Actual = banana_data$Weight_log
    ,ID = banana_data$ID
    ,`Std Residuals` = stdres(banana_reg_02)
    ,Leverage = hatvalues(banana_reg_02)
  ) %>%
  mutate(Residual = Actual - Predicted)

banana_resid_data %>%
  ggplot(aes(x = Predicted, y = `Std Residuals`)) +
  geom_hline(
    aes(yintercept = -2)
    ,linetype = "dashed"
  ) +
  geom_hline(
    aes(yintercept = 2)
    ,linetype = "dashed"
  ) +
  geom_point() +
  geom_text(
    data =
      banana_resid_data %>%
      filter(abs(`Std Residuals`) >= 2)
    ,aes(label = ID)
    ,nudge_x = 0.005
  ) +
  geom_smooth(
    method = "loess"
    ,se = FALSE
  ) +
  geom_smooth(
    method = "lm"
    ,se = FALSE
  ) +
  labs(
    title = "Figure 4: Standardized Residuals vs. Predicted for Model 02"
    ,x = "Predicted (ln)"
    ,y = "Standardized Residual (ln)"
  )

banana_resid_data %>%
  ggplot(aes(x = Leverage, y = `Std Residuals`)) +
```

```
  geom_hline(
    aes(yintercept = -2)
    ,linetype = "dashed"
  ) +
  geom_hline(
    aes(yintercept = 2)
    ,linetype = "dashed"
  ) +
  geom_point() +
  geom_text(
    data =
      banana_resid_data %>%
      filter(abs(`Std Residuals`) >= 2)
    ,aes(label = ID)
    ,nudge_x = 0.02
  ) +
  geom_smooth(
    method = "loess"
    ,se = FALSE
  ) +
  geom_smooth(
    method = "lm"
    ,se = FALSE
  ) +
  labs(
    title = "Figure 5: Standardized Residuals vs. Leverage for Model 02"
    ,x = "Leverage (ln)"
    ,y = "Standardized Residual (ln)"
  )

banana_data %>%
  inner_join(
    banana_resid_data %>%
      filter(abs(`Std Residuals`) > 2)
    ,by = "ID"
  ) %>%
  select(
    ID
    ,Weight
    ,Radius
    ,Length
    ,Circumference
    ,`Std Residuals`
    ,Leverage
  ) %>%
  kable(caption = "Entries with a |Standardized Residual| >2")

# CV DAAG
banana_data_post <-
  banana_data %>%
    inner_join(
    banana_resid_data
    ,by = "ID"
```

```
  ) %>%
  filter(
    !(abs(`Std Residuals`) > 2 & Leverage > 0.2)
  ) %>%
  select(
    -c(
      Predicted
      ,Actual
      ,`Std Residuals`
      ,Leverage
      ,Residual
    )
  )

banana_reg_cv <-
  banana_data_post %>%
  cv.lm(
    Weight_log ~ Length_log + Radius_log
    ,plotit = FALSE
  )

# CV caret
banana_train_control <- trainControl(method = "cv", number = 10)
banana_caret_cv <-
  train(
    Weight_log~.
    ,data =
      banana_data_post %>%
      select(Weight_log, Length_log, Radius_log)
    ,trControl = banana_train_control
    ,method = "lm"
  )
banana_caret_cv %>%
  summary() %>%
  pander()

# MAE Log
mae_log <-
  tibble(
    MSE =
      (
        banana_reg_cv %>%
          transmute((Weight_log - cvpred)^2) %>%
          sum()
      )/(as.numeric(count(banana_reg_cv)))
    ,MAE =
      (
        banana_reg_cv %>%
          transmute(abs(Weight_log - cvpred)) %>%
          sum()
      )/(as.numeric(count(banana_reg_cv)))
    ,MPAE =
      (
```

```
        banana_reg_cv %>%
          transmute(abs((Weight_log - cvpred) / Weight_log)) %>%
          sum()
      )/(as.numeric(count(banana_reg_cv)))
  ) %>%
  mutate(RMSE = sqrt(MSE))
kable(mae_log, caption = "Calculated Error Terms for Log CV Model")

# MAE
mae_regular <-
  tibble(
    MSE =
      (
        banana_reg_cv %>%
          transmute((Weight - exp(cvpred))^2) %>%
          sum()
      )/(as.numeric(count(banana_reg_cv)))
    ,MAE =
      (
        banana_reg_cv %>%
          transmute(abs(Weight - exp(cvpred))) %>%
          sum()
      )/(as.numeric(count(banana_reg_cv)))
    ,MPAE =
      (
        banana_reg_cv %>%
          transmute(abs((Weight - exp(cvpred)) / Weight)) %>%
          sum()
      )/(as.numeric(count(banana_reg_cv)))
  ) %>%
  mutate(RMSE = sqrt(MSE))
kable(mae_regular, caption = "Calculated Error Terms for CV Model")
```

Appendix B: Sample Size Table

Sample Size Recommendations at Selected Levels of Squared Population Multiple Correlation Coefficients for Varying Numbers of Predictor Variables

	Number of Predictor Variables					
ρ^2	2	3	4	5	7	9
			Good prediction level			
.10	240	380	440	550	700	900
.15	160	220	280	340	440	550
.20	110	170	200	260	320	400
.25	85	120	150	180	240	300
.30	65	95	130	150	190	240
.40	45	65	80	95	120	150
.50	35	45	55	65	85	100
.70	15	21	25	35	40	50
.90	7	9	10	11	14	16
			Excellent prediction level			
.10	950	1,500	1,800	2,200	2,800	—
.15	600	850	1,200	1,400	1,800	2,200
.20	420	650	800	950	1,300	1,500
.25	320	460	600	750	950	1,200
.30	260	360	480	600	800	1,000
.40	160	260	300	380	480	600
.50	110	130	220	230	320	400
.70	50	70	95	110	140	170
.90	15	21	29	35	40	50

Figure 1: