Does Size Matter? (Estimation of Banana Weight with a regression modeling appraoch)

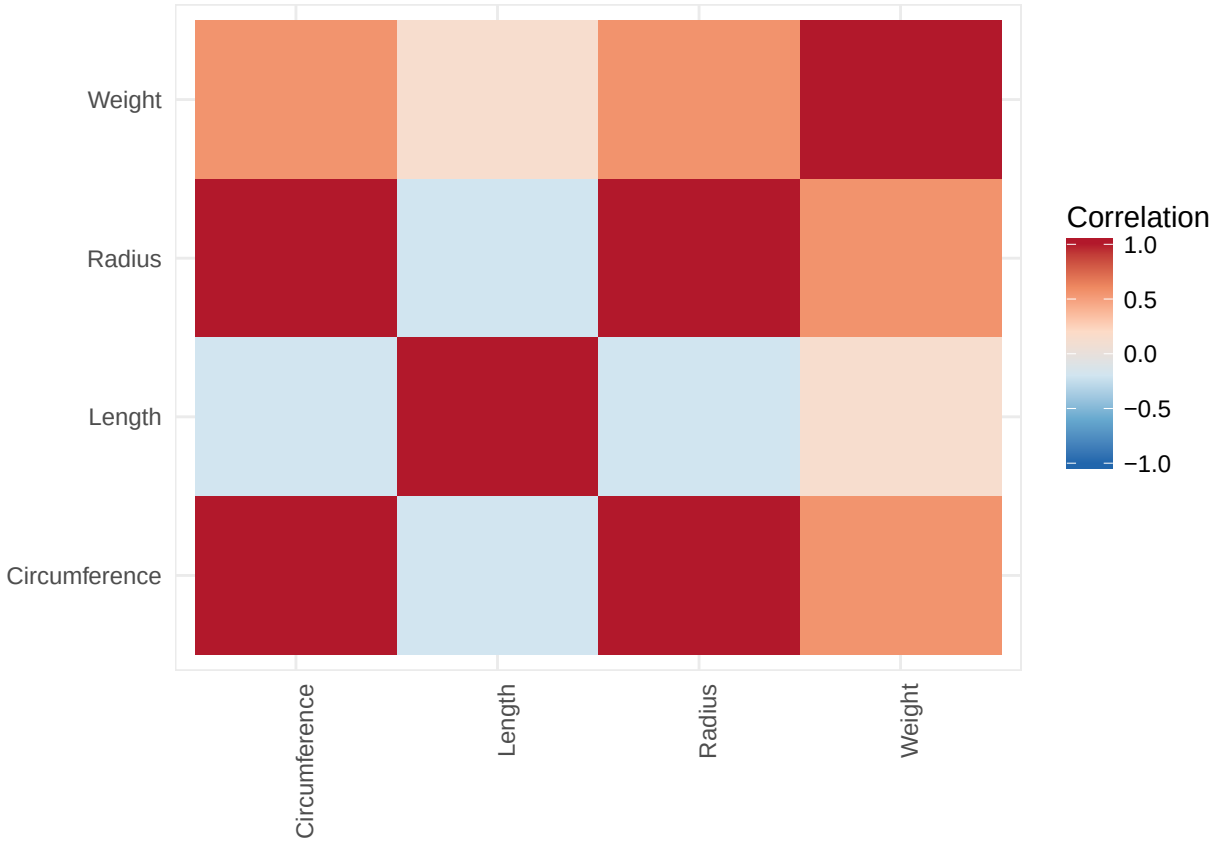
Scott Graham, Kaisa Roggeveen

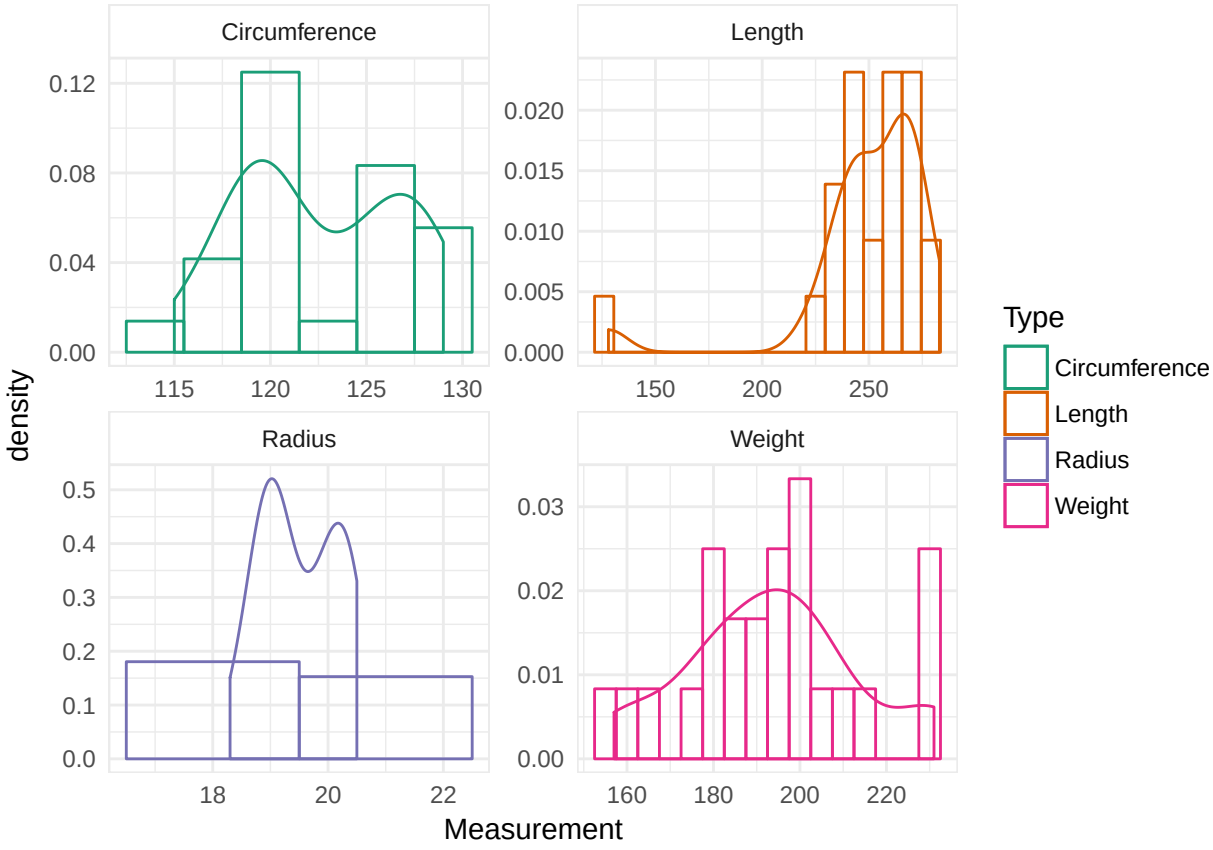
February 13, 2018

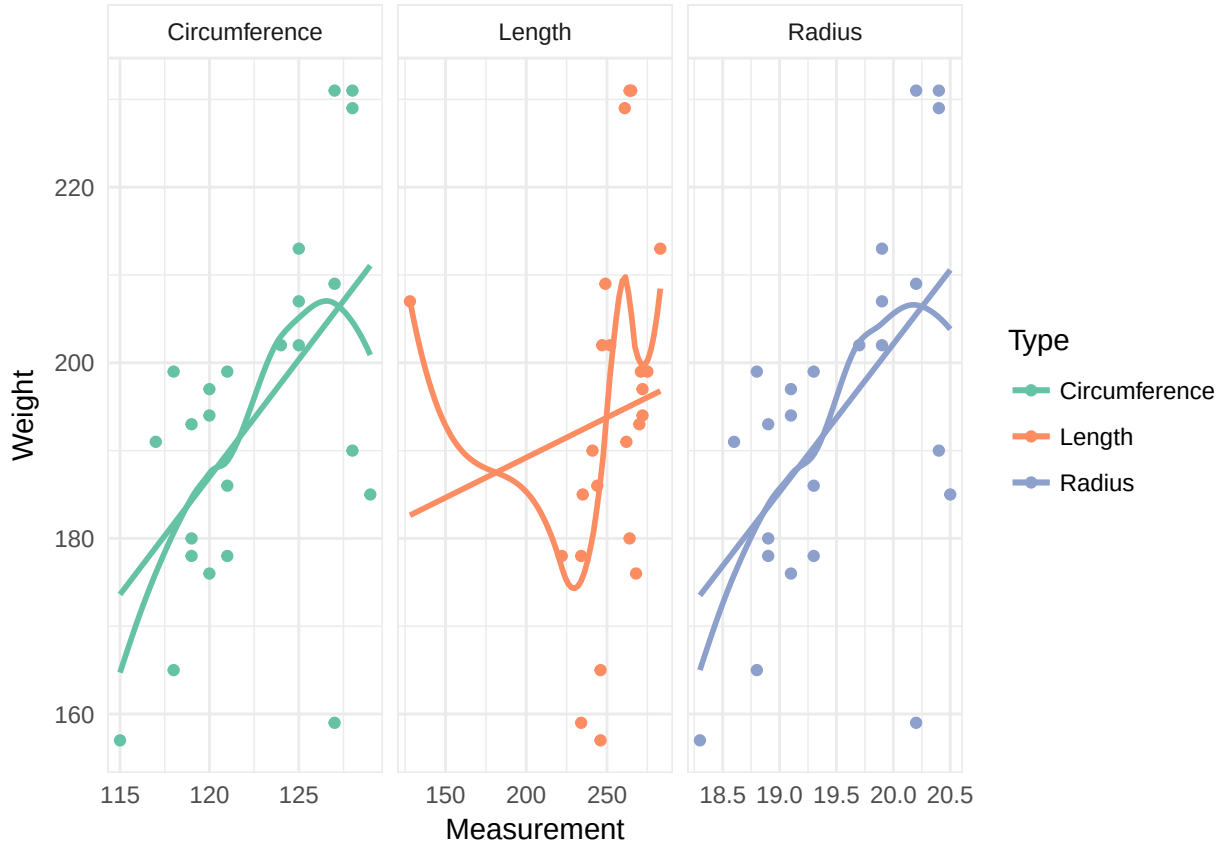
Summary

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.




Circumference
Length
Radius
Weight
density
Measurement
Type
Circumference
Length
Radius
Weight




```
##
## Call:
## lm(formula = Weight_log ~ Length_log + Radius_log + Circumference_log,
##     data = .)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.24351 -0.06228  0.02400  0.06062  0.11528
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        10.0594    26.3199   0.382    0.706
## Length_log          0.1230     0.1275   0.965    0.346
## Radius_log          7.5264    14.0928   0.534    0.599
## Circumference_log  -5.7885    14.1601  -0.409    0.687
##
## Residual standard error: 0.09248 on 20 degrees of freedom
## Multiple R-squared:  0.3318, Adjusted R-squared:  0.2316
## F-statistic:  3.31 on 3 and 20 DF,  p-value: 0.04104
```

```
##
## Call:
## lm(formula = Weight_log ~ Radius_log + Length_log, data = .)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.24861 -0.05259  0.02164  0.05202  0.11544
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.6702     1.9132  -0.350  0.72961
## Radius_log    1.7702     0.5596   3.163  0.00468 **
## Length_log    0.1223     0.1249   0.979  0.33888
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09062 on 21 degrees of freedom
## Multiple R-squared:  0.3262, Adjusted R-squared:  0.2621
## F-statistic: 5.084 on 2 and 21 DF,  p-value: 0.01583
```

```
##
## Call:
## lm(formula = Weight_log ~ Radius_log, data = .)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.25201 -0.05851  0.02530  0.05814  0.12150
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.3046     1.6319   0.187  0.85363
## Radius_log    1.6689     0.5494   3.038  0.00604 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09054 on 22 degrees of freedom
## Multiple R-squared:  0.2955, Adjusted R-squared:  0.2635
## F-statistic: 9.227 on 1 and 22 DF,  p-value: 0.006043
```

```
##
## Call:
## lm(formula = Weight_log ~ Length_log, data = .)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.204865 -0.072290 -0.000595  0.055375  0.177835
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.99043    0.80371   6.209    3e-06 ***
## Length_log   0.04917    0.14574   0.337    0.739
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1076 on 22 degrees of freedom
## Multiple R-squared:  0.005146,   Adjusted R-squared:  -0.04007
## F-statistic: 0.1138 on 1 and 22 DF,  p-value: 0.739
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
21	0.1724705	NA	NA	NA	NA
20	0.1710414	1	0.0014291	0.1671057	0.687041

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
22	0.1803370	NA	NA	NA	NA
20	0.1710414	2	0.0092956	0.5434713	0.5890658

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
22	0.1803370	NA	NA	NA	NA
21	0.1724705	1	0.0078665	0.9578258	0.3388762

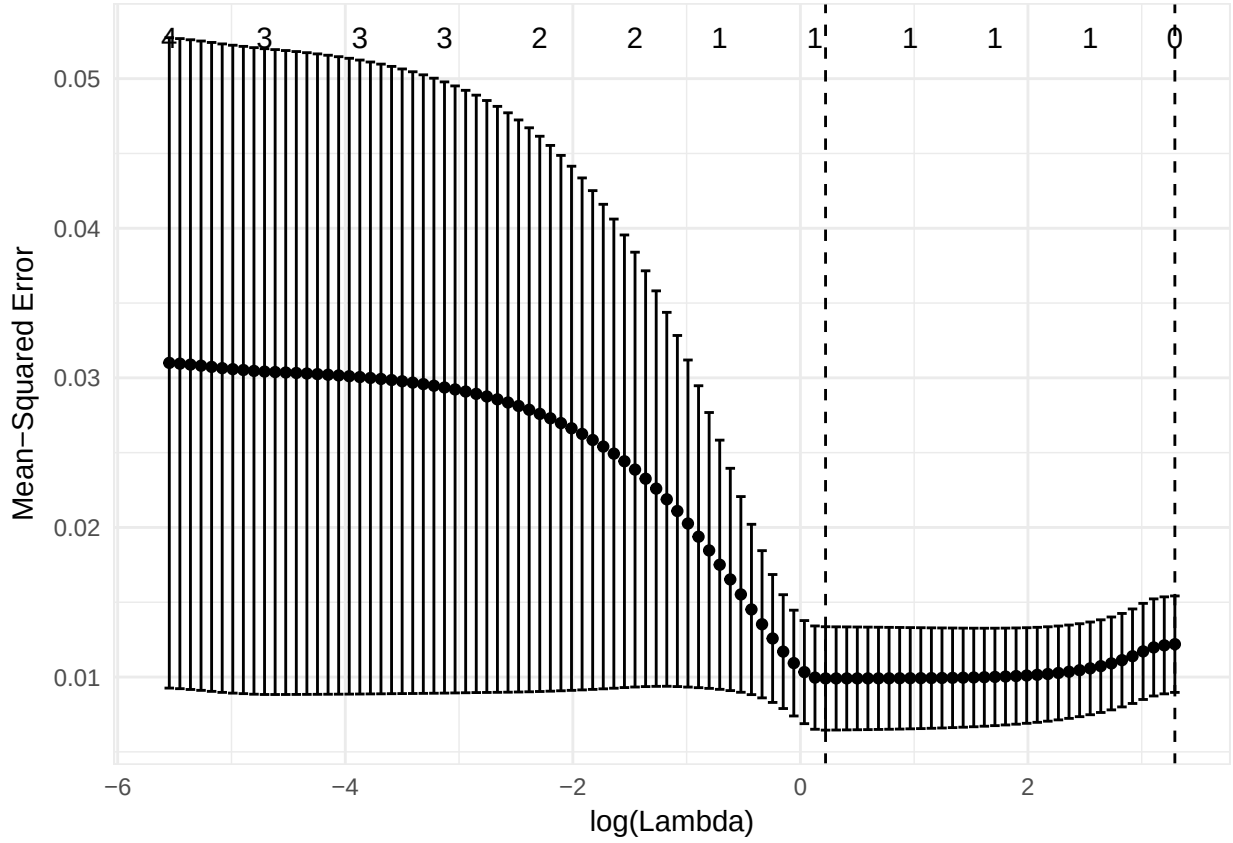
```
## Analysis of Variance Table
##
## Response: Weight_log
##            Df Sum Sq Mean Sq F value Pr(>F)
## Length_log  1 0.0013  0.0013    0.16 0.6928
## Radius_log  1 0.0822  0.0822   10.01 0.0047 **
## Residuals  21 0.1725  0.0082
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## fold 1
## Observations in test set: 8
##                [,1]   [,2]   [,3]    [,4]   [,5]    [,6]   [,7]   [,8]
## Predicted   5.2368 5.2368  5.196  5.1998 5.3143  5.2304  5.344  5.215
## cvpred      5.2428 5.2428  5.207  5.2131 5.3184  5.2464  5.358  5.222
## Weight_log  5.2832 5.2679  5.106  5.1818 5.3613  5.1818  5.220  5.193
## CV residual 0.0404 0.0251 -0.101 -0.0314 0.0429 -0.0646 -0.137 -0.029
##
## Sum of squares = 0.04    Mean square = 0    n = 8
##
## fold 2
## Observations in test set: 8
##                 [,1]  [,2]  [,3]    [,4]   [,5]    [,6]   [,7]   [,8]
## Predicted    5.24195 5.348 5.333  5.2350 5.2548  5.3386 5.1853 5.3251
## cvpred       5.23389 5.324 5.310  5.2248 5.2417  5.3184 5.1822 5.3056
## Weight_log   5.22575 5.434 5.442  5.1705 5.2933  5.2470 5.2523 5.3423
## CV residual -0.00814 0.109 0.132 -0.0543 0.0516 -0.0714 0.0701 0.0367
##
## Sum of squares = 0.05    Mean square = 0.01    n = 8
##
## fold 3
## Observations in test set: 8
##             [,1]   [,2]    [,3]   [,4]   [,5]   [,6]   [,7]   [,8]
## Predicted   5.22  5.318  5.1488 5.3497 5.2822 5.2976 5.2173 5.2101
## cvpred      4.78  5.267  5.1137 5.3771 5.2744 5.2788 5.2464 5.2500
## Weight_log  5.33  5.069  5.0562 5.4424 5.3083 5.3083 5.2627 5.2933
## CV residual 0.55 -0.198 -0.0574 0.0654 0.0339 0.0295 0.0163 0.0433
##
## Sum of squares = 0.35    Mean square = 0.04    n = 8
##
## Overall (Sum over all 8 folds)
##     ms
```

```
## 0.0183
```

```
## # A tibble: 24 x 13
##       ID Weight Radius Length Circumference Weight_log Radius_log
##    <int>  <int>  <dbl>  <int>         <int>      <dbl>      <dbl>
##  1     1    197   19.1    272           120       5.28       2.95
##  2     2    194   19.1    272           120       5.27       2.95
##  3     3    165   18.8    246           118       5.11       2.93
##  4     4    186   19.3    244           121       5.23       2.96
##  5     5    178   18.9    234           119       5.18       2.94
##  6     6    207   19.9    128           125       5.33       2.99
##  7     7    213   19.9    283           125       5.36       2.99
##  8     8    178   19.3    222           121       5.18       2.96
##  9     9    229   20.4    261           128       5.43       3.02
## 10    10    231   20.2    265           127       5.44       3.01
## # ... with 14 more rows, and 6 more variables: Length_log <dbl>,
## #   Circumference_log <dbl>, Predicted <dbl>, cvpred <dbl>, `CV Residual`
## #   <dbl>, Residual <dbl>
```

```
## Warning: Option grouped=FALSE enforced in cv.glmnet, since < 3 observations
## per fold
```

```
## Warning: Option grouped=FALSE enforced in cv.glmnet, since < 3 observations
## per fold
```



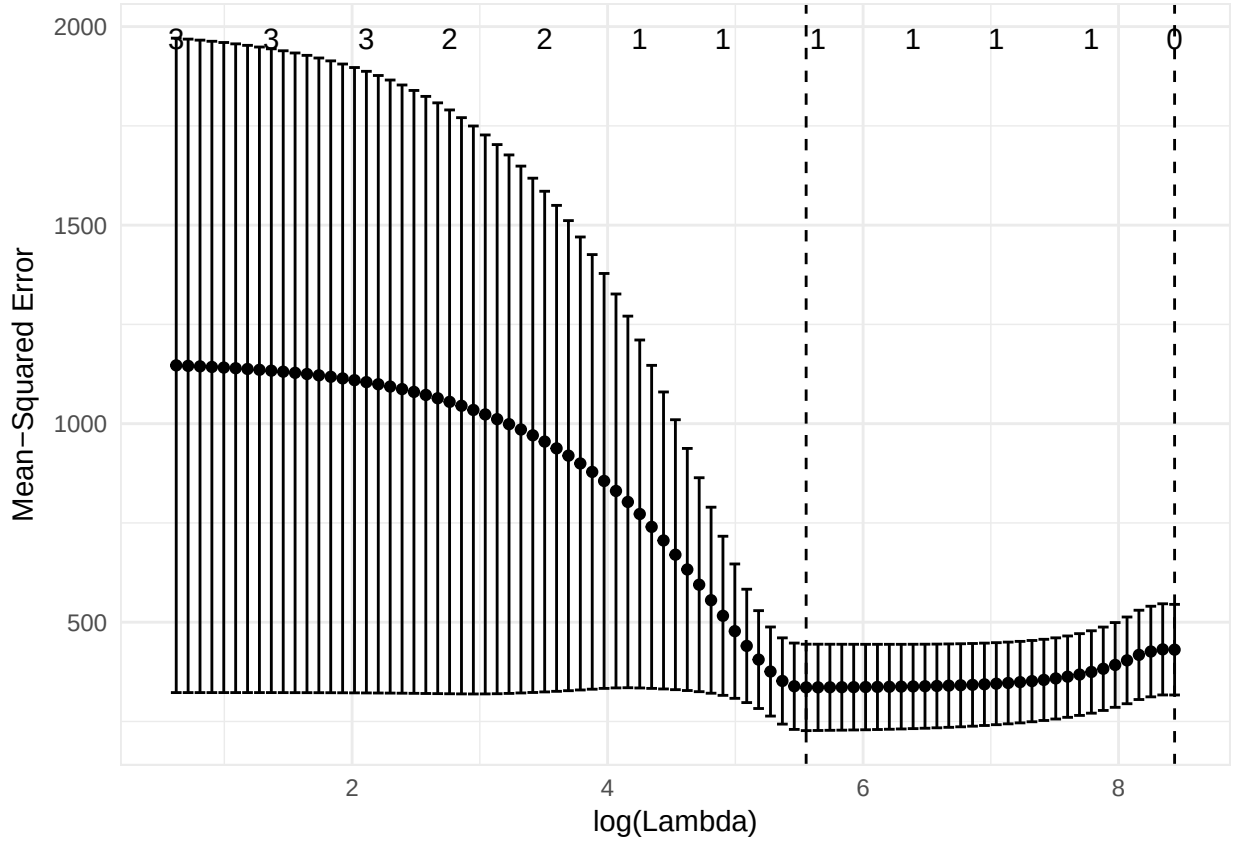
	1
(Intercept)	0.535

	1
Length	0.000
Radius	0.000
Circumference	0.000
Length_log	0.000
Radius_log	1.591
Circumference_log	0.000

	1
(Intercept)	5.26
Length	0.00
Radius	0.00
Circumference	0.00
Length_log	0.00
Radius_log	0.00
Circumference_log	0.00

```
## Warning: Option grouped=FALSE enforced in cv.glmnet, since < 3 observations
## per fold

## Warning: Option grouped=FALSE enforced in cv.glmnet, since < 3 observations
## per fold
```



	1
(Intercept)	-731
Length	0
Radius	0
Circumference	0
Length_log	0
Radius_log	311
Circumference_log	0

	1
(Intercept)	194
Length	0
Radius	0
Circumference	0
Length_log	0
Radius_log	0
Circumference_log	0

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.003   0.418   0.621   0.572   0.757   0.979
```

Cross Validation

MAE

```
## # A tibble: 1 x 2
##     MAE   MPAE
##   <dbl>  <dbl>
## 1  13.7 0.0722
```