Haiyang Zhou B00849335
Han Cao B00770581
Xinyan Chen B00810311
Pengcheng Liang B00783818

Abstract

This article uses data from the Fish Market Dataset from LionBridge which shows the relationship of body length, width, height, and weight in Bream, Roach, Whitefish, Parkki, Perche, Smelt, seven species in the first alphabetic order. We hypothesized that weight in these species was positively related to length, width, and height. We use multiple linear regression and multiple regression model to analyze the linear and non-linear relation between the data in the dataset, through the analysis of the correlation coefficient and the backward of the selection of the final model, the length2 and length3 did not produce relationship with weight, so our final model is the weight and length of 1, species, height between the regression model. Weight and length1, species and height have relationship, which confirmed our hypothesis.

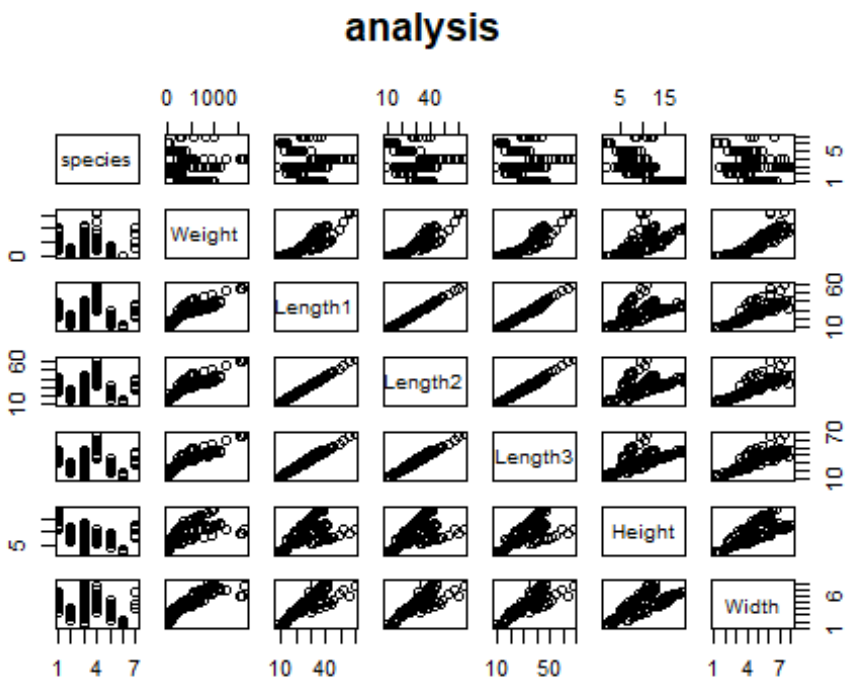
Introduction.

Are you curious about different kinds of fish, whether there is a certain relationship between their weight and body length, width, and height? We collected our data to explore what other factors are related to fish weight.

From our dataset collected from fisher market, we got the length, width, height, and weight of seven kinds of fish. We assumed that the weight of the fish has relationship with the body length, species and height. We will establish multiple regression to analyze the relationship between weight and body length, species and height.

To get this model, we did the following. We added a new data of Bream based on original data.

```
#1.Data visualization
fishdata <- read.csv("C:/Users/jerry/Desktop/project.csv", header=TRUE)
pairs(fishdata, main="analysis")
```




```
typeof(fishdata)
```

```
## [1] "list"
```

```
table(fishdata$Weight)
```

```
##
##    0  5.9  6.7    7  7.5  8.7  9.7  9.8  9.9   10 12.2 13.4 19.7 19.
9   32   40
##    1    1    1    1    1    1    1    2    1    1    2    1    1
1    1    2
## 51.5   55   60   69   70   78   80   85   87   90  100  110  115  12
0  125  130
##    1    1    1    1    1    2    1    2    1    1    1    3    1
5    1    3
##  135  140  145  150  160  161  169  170  180  188  197  200  218  22
5  230  242
##    1    2    4    4    2    1    1    2    2    1    1    3    1
1    1    1
##  250  260  265  270  272  273  290  300  306  320  340  345  363  39
0  430  450
##    2    1    1    2    1    1    2    6    1    1    2    1    1
2    2    2
##  456  475  500  510  514  540  556  567  575  600  610  620  650  68
0  685  690
##    1    1    5    1    1    2    1    1    1    2    1    1    2
1    2    1
##  700  714  720  725  770  800  820  840  850  900  920  925  950  95
5  975 1000
##    5    1    1    1    1    1    2    1    2    2    1    1    2
1    1    5
## 1015 1100 1250 1550 1600 1650
##    1    2    1    1    1    1
```

```
a1=sub("Bream","0",fishdata$species)
a2=sub('Parkki','1',a1)
a3=sub('Perch','2',a2)
a4=sub('Pike','3',a3)
a5=sub('Roach','4',a4)
a6=sub("Smelt",'5',a5)
a7=sub("Whitefish",'6',a6)
fishdata$species=a7
fishdata$species<-as.numeric(fishdata$species)
#this is the analysis of two variable.
#we can saw from graph, the lengh 1, length 2, length 3 has strongest l
inear relationship with each other, which is a linear in the graph
#We use 0,1,2,3,4,5,6 to represent Bream, Parkki, Perch, Pike, Roach, S
melt, Whitefish respectively.

#2.correlation matrix
cov(fishdata)
```

```
##             species      Weight     Length1     Length2     Length3
     Height
## species    2.918239   -187.3489   -4.469025   -5.098585   -6.565566
  -5.096774
## Weight  -187.348899 127519.5960 3259.689624 3505.443125 3814.583483
1102.288683
## Length1   -4.469025   3259.6896   99.366254  106.469335  114.457318
  26.585754
## Length2   -5.098585   3505.4431  106.469335  114.190346  122.955106
  29.188659
## Length3   -6.565566   3814.5835  114.457318  122.955106  133.981022
  34.755994
## Height    -5.096774   1102.2887   26.585754   29.188659   34.755994
  18.281762
## Width     -1.122085    532.5906   14.538280   15.701503   17.099548
   5.681458
##              Width
## species  -1.122085
## Weight  532.590612
## Length1  14.538280
## Length2  15.701503
## Length3  17.099548
## Height    5.681458
## Width     2.829322

#3.corr
cor(fishdata)

##             species     Weight    Length1    Length2    Length3     H
eight
## species  1.0000000 -0.3071159 -0.2624417 -0.2793023 -0.3320398 -0.69
77918
## Weight  -0.3071159  1.0000000  0.9157319  0.9186282  0.9228625  0.72
19346
## Length1 -0.2624417  0.9157319  1.0000000  0.9995175  0.9919790  0.62
37643
## Length2 -0.2793023  0.9186282  0.9995175  1.0000000  0.9940546  0.63
88375
## Length3 -0.3320398  0.9228625  0.9919790  0.9940546  1.0000000  0.70
22618
## Height  -0.6977918  0.7219346  0.6237643  0.6388375  0.7022618  1.00
00000
## Width   -0.3905029  0.8866738  0.8670664  0.8735451  0.8782573  0.78
99681
##              Width
## species -0.3905029
## Weight   0.8866738
## Length1  0.8670664
## Length2  0.8735451
## Length3  0.8782573
```

```
## Height   0.7899681
## Width    1.0000000
```

```
#4. model selection
library(olsrr)
```

```
## Warning: package 'olsrr' was built under R version 3.6.3
```

```
##
## Attaching package: 'olsrr'
```

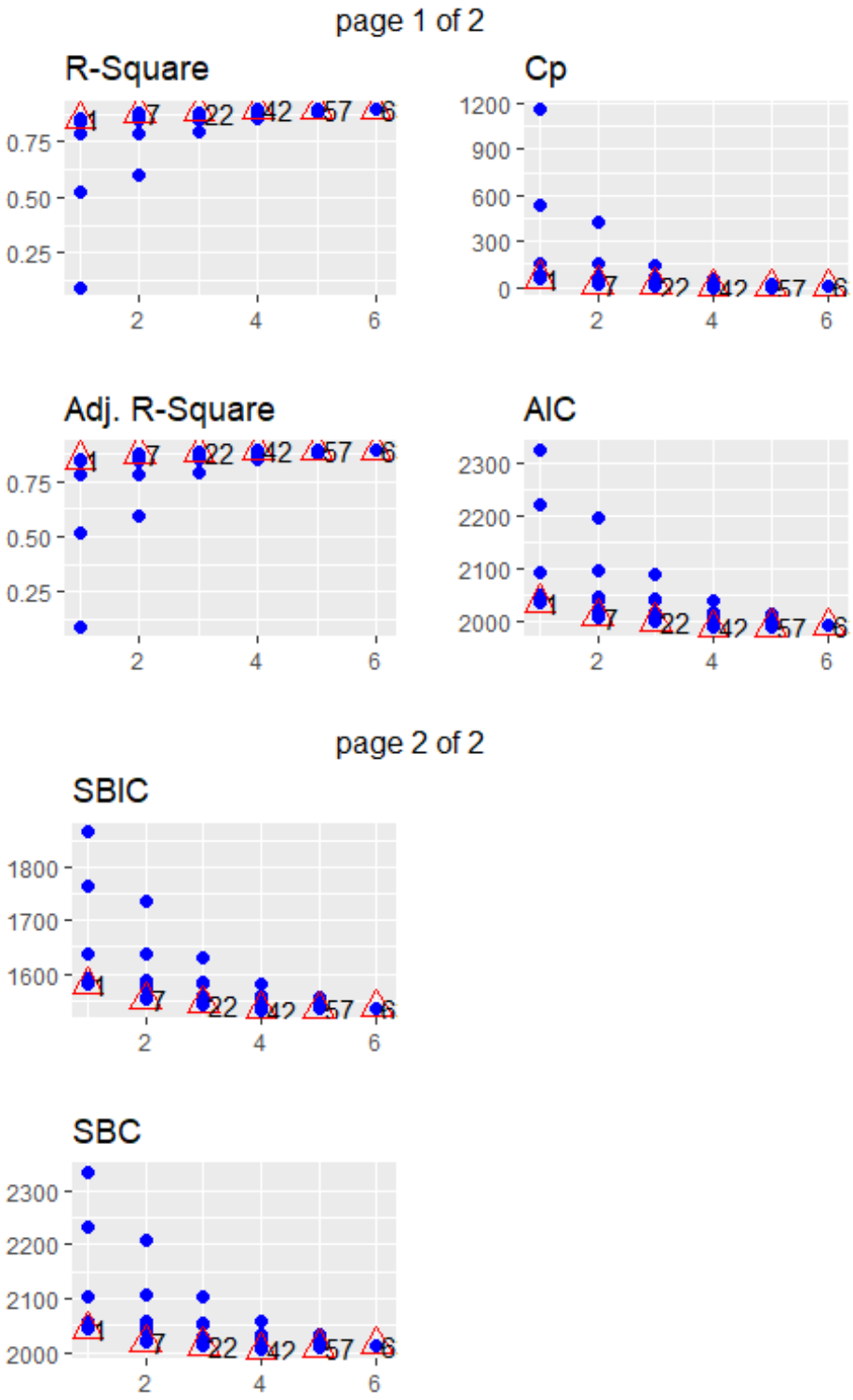
```
## The following object is masked from 'package:datasets':
##
##     rivers
```

```
model=lm(Weight~species+Length1+Length2+Length3+Height+Width,data=fishd
ata)
allpossible=ols_step_all_possible(model)
allpossible
```

```
##    Index N                                    Predictors   R-Square
## 4      1 1                                       Length3 0.85167527
## 3      2 1                                       Length2 0.84387775
## 2      3 1                                       Length1 0.83856485
## 6      4 1                                         Width 0.78619035
## 5      5 1                                        Height 0.52118958
## 1      6 1                                       species 0.09432016
## 20     7 2                                 Length3 Width 0.87704347
## 14     8 2                                Length1 Height 0.87575588
## 17     9 2                                Length2 Height 0.87470582
## 18    10 2                                 Length2 Width 0.87380960
## 15    11 2                                 Length1 Width 0.87316801
## 19    12 2                                Length3 Height 0.86243406
## 16    13 2                               Length2 Length3 0.85180758
## 13    14 2                               Length1 Length3 0.85167988
## 9     15 2                               species Length3 0.85167580
## 12    16 2                               Length1 Length2 0.85011449
## 8     17 2                               species Length2 0.84664826
## 7     18 2                               species Length1 0.84335567
## 11    19 2                                 species Width 0.78799726
## 21    20 2                                  Height Width 0.78741883
## 10    21 2                                species Height 0.59655494
## 24    22 3                         species Length1 Height 0.88485423
## 27    23 3                         species Length2 Height 0.88461950
## 35    24 3                        Length1 Length3 Height 0.88397873
## 37    25 3                          Length1 Height Width 0.88239842
## 38    26 3                        Length2 Length3 Height 0.88225695
## 40    27 3                          Length2 Height Width 0.88140647
## 30    28 3                         species Length3 Width 0.87827128
## 41    29 3                          Length3 Height Width 0.87773815
## 36    30 3                         Length1 Length3 Width 0.87721523
```

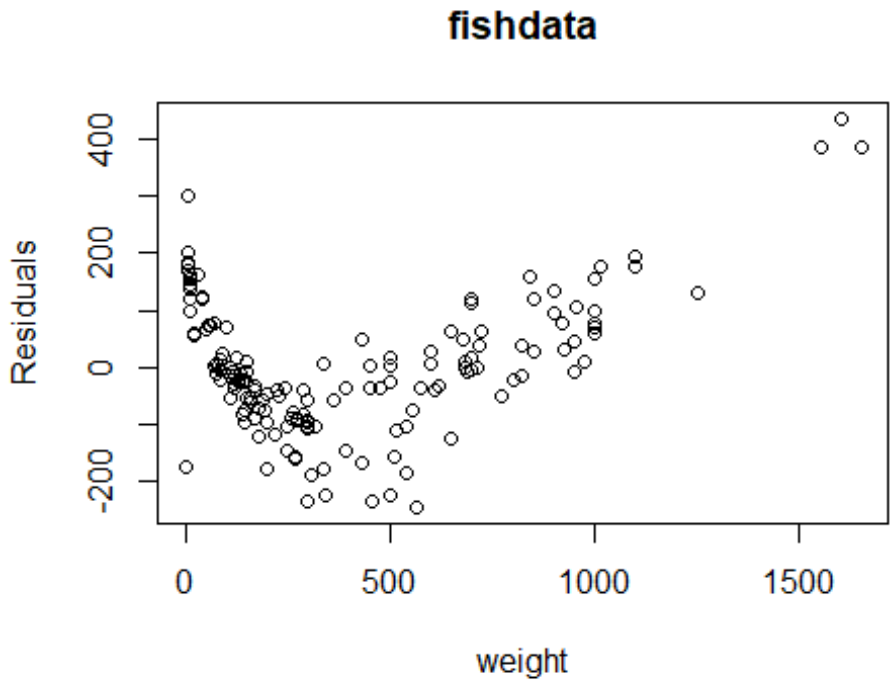
```
## 39    31 3                         Length2 Length3 Width 0.87714237
## 33    32 3                        Length1 Length2 Height 0.87611432
## 34    33 3                         Length1 Length2 Width 0.87392078
## 28    34 3                         species Length2 Width 0.87381245
## 25    35 3                         species Length1 Width 0.87321092
## 29    36 3                        species Length3 Height 0.87211322
## 32    37 3                       Length1 Length2 Length3 0.85356420
## 26    38 3                       species Length2 Length3 0.85186204
## 23    39 3                       species Length1 Length3 0.85168536
## 22    40 3                       species Length1 Length2 0.85021617
## 31    41 3                          species Height Width 0.79589906
## 48    42 4                 species Length2 Length3 Height 0.89416827
## 45    43 4                 species Length1 Length3 Height 0.89357856
## 47    44 4                   species Length1 Height Width 0.88880473
## 50    45 4                   species Length2 Height Width 0.88841452
## 55    46 4                   Length1 Length3 Height Width 0.88504484
## 43    47 4                 species Length1 Length2 Height 0.88485734
## 52    48 4                 Length1 Length2 Length3 Height 0.88402250
## 56    49 4                   Length2 Length3 Height Width 0.88331143
## 51    50 4                   species Length3 Height Width 0.88316363
## 54    51 4                   Length1 Length2 Height Width 0.88300100
## 49    52 4                  species Length2 Length3 Width 0.87835514
## 46    53 4                  species Length1 Length3 Width 0.87833414
## 53    54 4                  Length1 Length2 Length3 Width 0.87741869
## 44    55 4                  species Length1 Length2 Width 0.87398540
## 42    56 4                species Length1 Length2 Length3 0.85368960
## 57    57 5        species Length1 Length2 Length3 Height 0.89450015
## 61    58 5          species Length2 Length3 Height Width 0.89427437
## 60    59 5          species Length1 Length3 Height Width 0.89359940
## 59    60 5          species Length1 Length2 Height Width 0.88883800
## 62    61 5          Length1 Length2 Length3 Height Width 0.88506017
## 58    62 5         species Length1 Length2 Length3 Width 0.87837387
## 63    63 6 species Length1 Length2 Length3 Height Width 0.89457635
##    Adj. R-Square Mallow's Cp
## 4      0.8507365   59.261797
## 3      0.8428896   70.578236
## 2      0.8375431   78.288781
## 6      0.7848371  154.299226
## 5      0.5181591  538.891461
## 1      0.0885880 1158.401623
## 20     0.8754771   24.445252
## 14     0.8741732   26.313908
## 17     0.8731097   27.837845
## 18     0.8722021   29.138517
## 15     0.8715523   30.069654
## 19     0.8606816   45.647698
## 16     0.8499198   61.069773
## 13     0.8497905   61.255095
## 9      0.8497863   61.261024
```

```
## 12     0.8482051   63.526926
## 8      0.8446947   68.557433
## 7      0.8413602   73.335922
## 11     0.7852966  153.676873
## 21     0.7847108  154.516350
## 10     0.5914155  431.514681
## 24     0.8826399   15.109584
## 27     0.8824006   15.450247
## 35     0.8817476   16.380185
## 37     0.8801368   18.673679
## 38     0.8799927   18.878984
## 40     0.8791258   20.113287
## 30     0.8759303   24.663341
## 41     0.8753870   25.437069
## 36     0.8748540   26.195976
## 39     0.8747797   26.301715
## 33     0.8737319   27.793704
## 34     0.8714962   30.977167
## 28     0.8713858   31.134387
## 25     0.8707727   32.007383
## 29     0.8696539   33.600456
## 32     0.8507481   60.520406
## 26     0.8490132   62.990728
## 23     0.8488332   63.247147
## 22     0.8473357   65.379364
## 31     0.7919740  144.209100
## 48     0.8914371    3.592244
## 45     0.8908322    4.448075
## 47     0.8859352   11.376276
## 50     0.8855349   11.942590
## 55     0.8820783   16.832959
## 43     0.8818859   17.105079
## 52     0.8810295   18.316663
## 56     0.8803001   19.348641
## 51     0.8801485   19.563136
## 54     0.8799817   19.799162
## 49     0.8752159   26.541631
## 46     0.8751944   26.572109
## 53     0.8742553   27.900695
## 44     0.8707334   32.883382
## 42     0.8499139   62.338415
## 57     0.8910748    5.110593
## 61     0.8908417    5.438259
## 60     0.8901448    6.417835
## 59     0.8852288   13.328002
## 62     0.8813284   18.810707
## 58     0.8744250   28.514453
## 63     0.8904421    7.000000

plot(allpossible)
```




```
#1. more independent variable means larger R square. R^2=SSR/SST

#2. the lower the cp the better variable, which 16, 26, 31 is better.
```

```
#3. adj.R^2 is not influence by the number of variables.

#4. the smaller AIC, the better variables, which is 16, 26

fish.res=resid(model)
plot(fishdata$Weight,fish.res, ylab="Residuals", xlab="weight", main="f
ishdata")
```



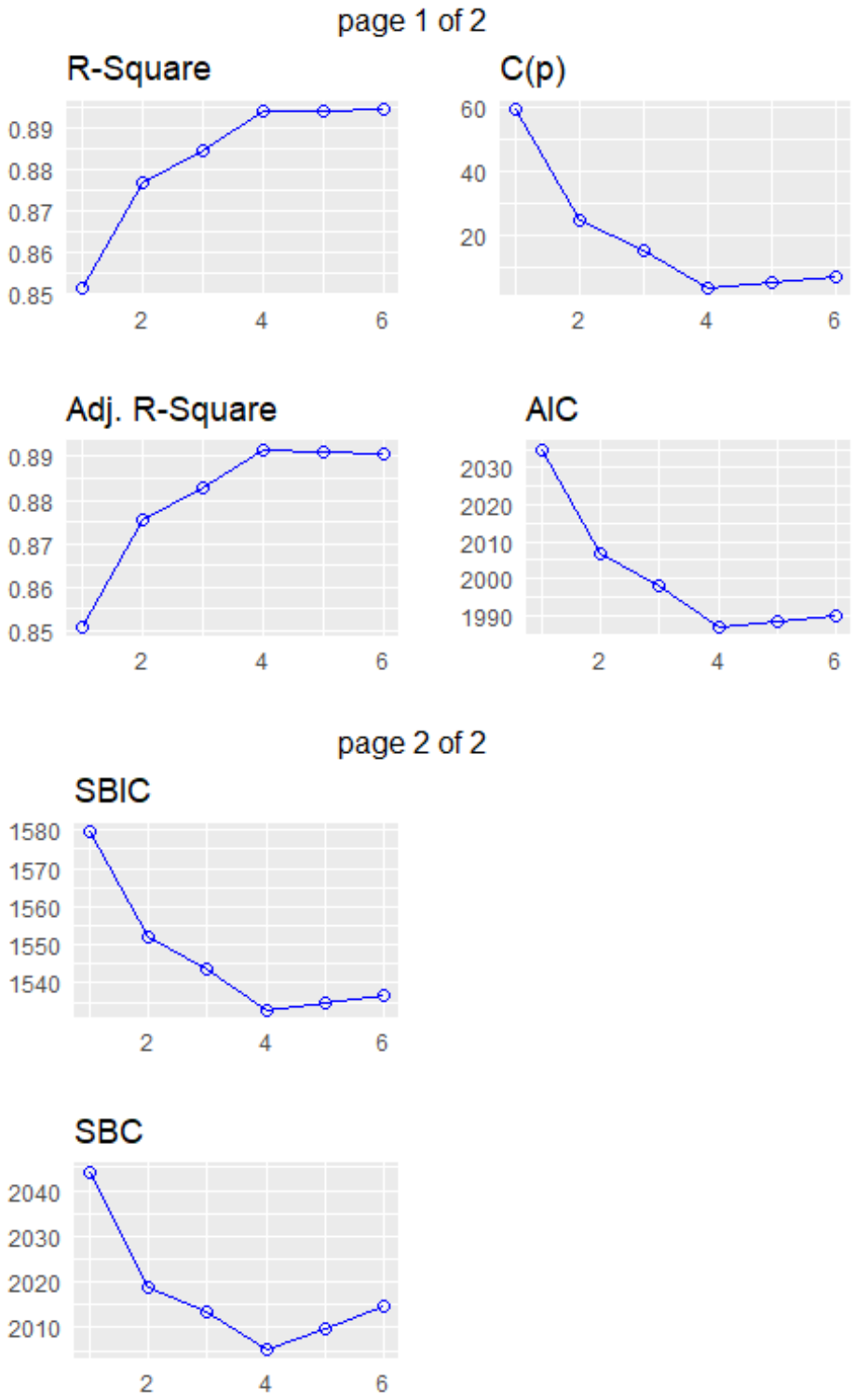

```
#it obeys the regression assumption, the residual variance should be a
normal number. the graph shows that the residual variance is not
constant. besides, the residual should be random, however in this
graph, it seems like an x^2 relationship. We can use log(y) to deal
with this problem. However after we used it, the result is not much
better.
```

Maybe it is because we still do not have enough data and it's not clear that there is a quadratic relationship from the graph. I think if we have enough data, it should follow a linear relationship otherwise there is no way that our transformation won't make it better. So I do not do any transformation.

```
model=lm(Weight~species+Length1+Length2+Length3+Height+Width,data=fishd
ata)
best=ols_step_best_subset(model)
best
```

```
##                    Best Subsets Regression
## ---------------------------------------------------------
## Model Index    Predictors
## ---------------------------------------------------------
##      1         Length3
##      2         Length3 Width
##      3         species Length1 Height
##      4         species Length2 Length3 Height
##      5         species Length1 Length2 Length3 Height
##      6         species Length1 Length2 Length3 Height Width
## ---------------------------------------------------------
##
##                                                          Subsets Re
gression Summary

## ------------------------------------------------------------------
-------------------------------------------------------------------
----
##                        Adj.        Pred


## Model    R-Square    R-Square    R-Square     C(p)          AIC
   SBIC           SBC            MSEP            FPE           HSP        A
PC
## ------------------------------------------------------------------
-------------------------------------------------------------------
----
##   1        0.8517      0.8507      0.8466    59.2618    2034.6850
 1579.4660    2043.9106    3045446.3900    19271.9463    121.2358    0.
1521
##   2        0.8770      0.8755       0.866    24.4453    2006.6733
 1551.9714    2018.9740    2540762.3596    16176.8379    101.7891    0.
1277
##   3        0.8849      0.8826      0.8748    15.1096    1998.1722
 1543.7679    2013.5481    2394712.1051    15339.8564     96.5530    0.
1211
##   4        0.8942      0.8914      0.8814     3.5922    1986.6765
 1533.0298    2005.1275    2215298.2914    14276.5153     89.8954    0.
1127
##   5        0.8945      0.8911      0.8803     5.1106    1988.1739
 1534.6495    2009.7001    2222784.9916    14410.9725     90.7849    0.
1137
##   6        0.8946      0.8904      0.8782     7.0000    1990.0583
```

```
 1536.6343    2014.6597    2235792.4843    14582.0021     91.9130    0.
1151
## ------------------------------------------------------------------
------------------------------------------------------------------
----
## AIC: Akaike Information Criteria
##  SBIC: Sawa's Bayesian Information Criteria
##  SBC: Schwarz Bayesian Criteria
##  MSEP: Estimated error of prediction, assuming multivariate normalit
y
##  FPE: Final Prediction Error
##  HSP: Hocking's Sp
##  APC: Amemiya Prediction Criteria

plot(best)
```




```
#so the best number of variable should be 4,because C(p) and AIC are th
e smallest.
```

```r
#backwards analysis

#we set up alpha=0.01

summary(lm(Weight~species+Length1+Length2+Length3+Height+Width,data=fis
hdata))
```

```
##
## Call:
## lm(formula = Weight ~ species + Length1 + Length2 + Length3 +
##     Height + Width, data = fishdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -246.17  -72.35  -13.36   64.85  436.66
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -634.233     45.714 -13.874  < 2e-16 ***
## species       33.491      9.012   3.716 0.000283 ***
## Length1       26.279     39.695   0.662 0.508960
## Length2       51.076     42.894   1.191 0.235604
## Length3      -51.051     17.690  -2.886 0.004468 **
## Height        49.111     10.128   4.849 3.02e-06 ***
## Width         -7.036     21.157  -0.333 0.739926
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 118.2 on 153 degrees of freedom
## Multiple R-squared:  0.8946, Adjusted R-squared:  0.8904
## F-statistic: 216.4 on 6 and 153 DF,  p-value: < 2.2e-16
```

```r
#find the largest p-value which is width that 0.72197 and larger than t
he alpha, and we delete it,then we have a new model and do regression a
nalysis again

summary(lm(Weight~species+Length1+Length3+Height+Length2,data=fishdat
a))
```

```
##
## Call:
## lm(formula = Weight ~ species + Length1 + Length3 + Height +
##     Length2, data = fishdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -240.50  -71.73  -13.54   62.45  442.22
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

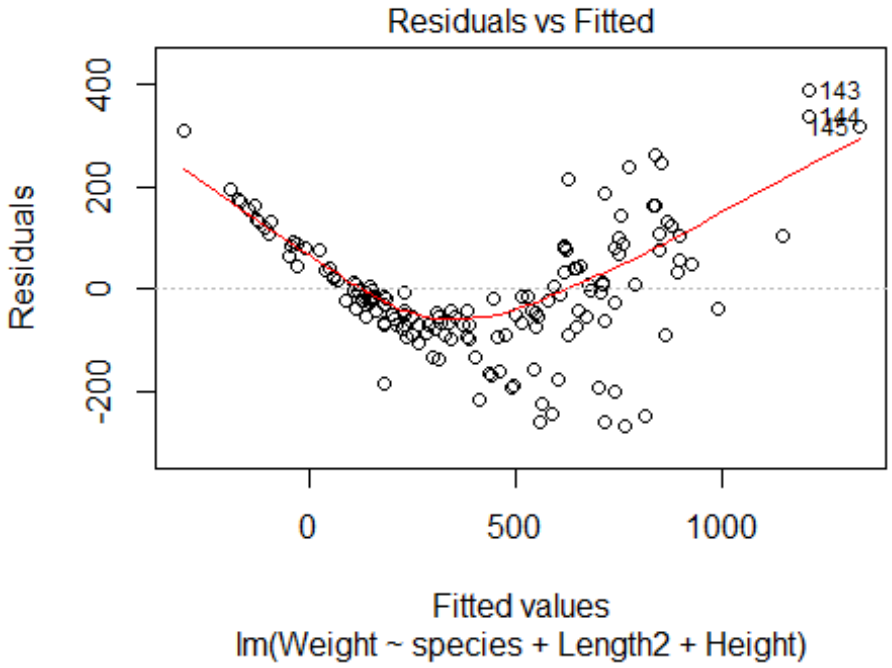
```
## (Intercept) -631.354     44.757 -14.106  < 2e-16 ***
## species       32.313      8.263   3.911 0.000138 ***
## Length1       27.442     39.427   0.696 0.487463
## Length3      -46.901     12.501  -3.752 0.000248 ***
## Height        46.405      6.012   7.718 1.39e-12 ***
## Length2       45.194     38.965   1.160 0.247905
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 117.9 on 154 degrees of freedom
## Multiple R-squared:  0.8945, Adjusted R-squared:  0.8911
## F-statistic: 261.1 on 5 and 154 DF,  p-value: < 2.2e-16

#find the largest p-value which is Length1 which is 0.474379, and large
r than the alpha, and we delete it,then we have a new model and do regr
ession analysis again

summary(lm(Weight~species+Length2+Length3+Height,data=fishdata))

##
## Call:
## lm(formula = Weight ~ species + Length2 + Length3 + Height, data = f
ishdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -233.75  -73.44  -14.32   64.80  438.99
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -640.692     42.628 -15.030  < 2e-16 ***
## species       33.591      8.042   4.177 4.92e-05 ***
## Length2       70.883     12.472   5.683 6.37e-08 ***
## Length3      -46.653     12.475  -3.740 0.000259 ***
## Height        45.181      5.740   7.872 5.66e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 117.7 on 155 degrees of freedom
## Multiple R-squared:  0.8942, Adjusted R-squared:  0.8914
## F-statistic: 327.4 on 4 and 155 DF,  p-value: < 2.2e-16

#at this time we find that p-value of Length3>alpha, so we delete lengt
h 3. Our final model is Weight~species+Length1+Height

#5. modify mnodel

#by checking if that suitable gauss-markov assumption

summary(lm(Weight~species+Length2+Height,data=fishdata))
```

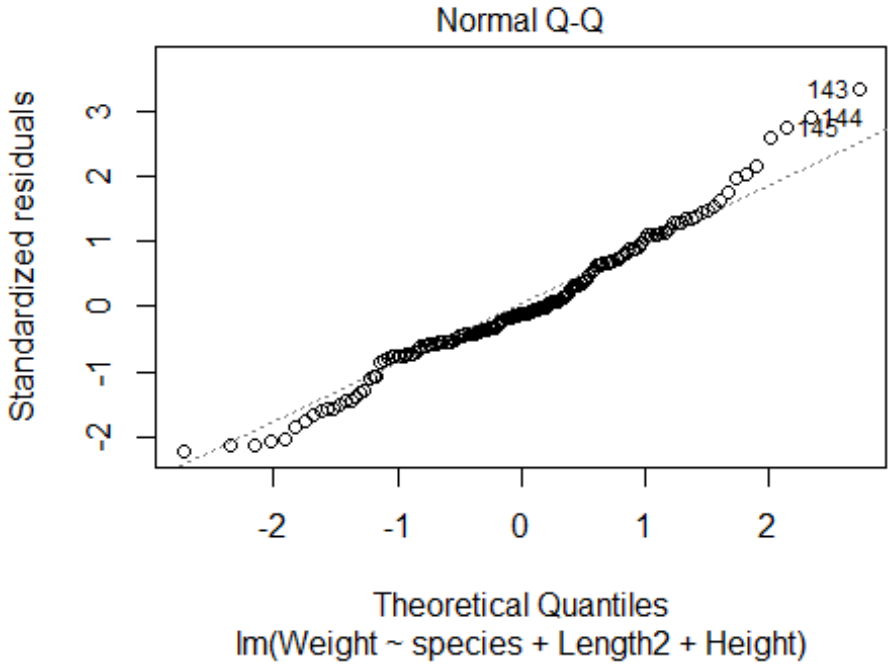
```
##
## Call:
## lm(formula = Weight ~ species + Length2 + Height, data = fishdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -266.04  -68.57  -15.21   80.37  388.60
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -632.942     44.315 -14.283  < 2e-16 ***
## species       30.481      8.326   3.661 0.000343 ***
## Length2       24.456      1.239  19.735  < 2e-16 ***
## Height        29.746      4.152   7.165  2.9e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 122.5 on 156 degrees of freedom
## Multiple R-squared:  0.8846, Adjusted R-squared:  0.8824
## F-statistic: 398.7 on 3 and 156 DF,  p-value: < 2.2e-16
```

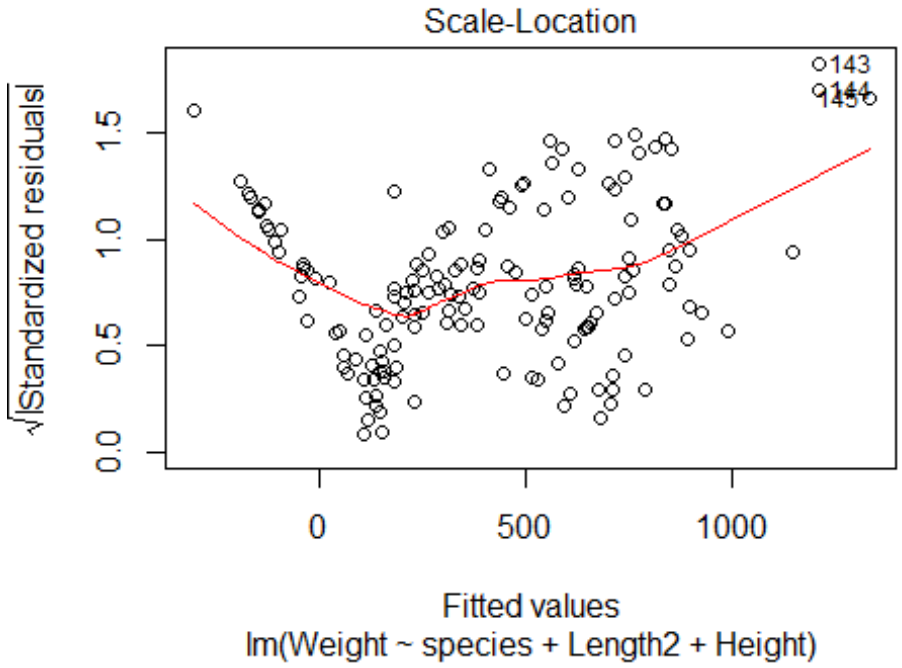
```
#we can see the bestmodel is Weight~species+Length1+Height
anova(lm(Weight~species+Length2+Height,data=fishdata))
```

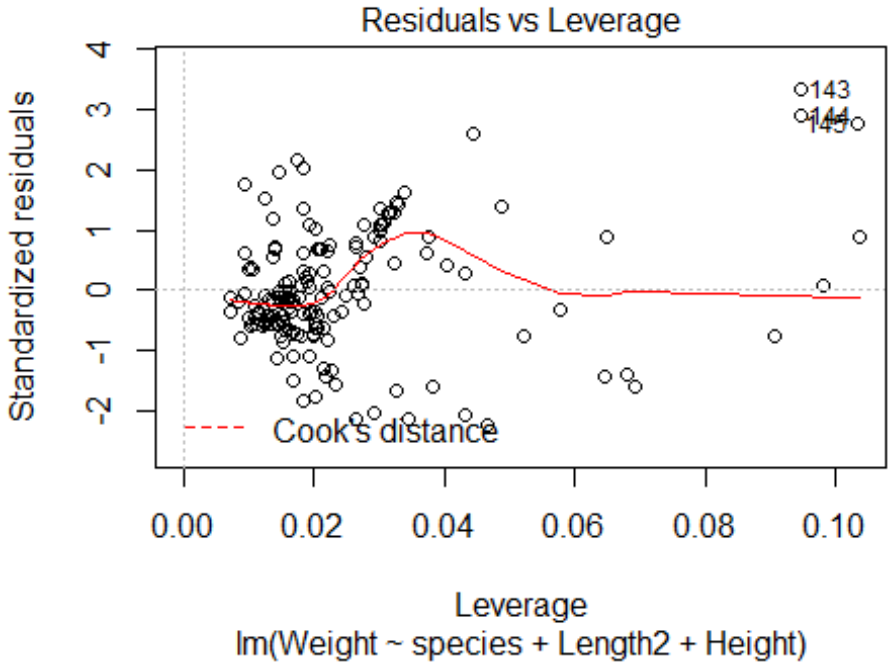
```
## Analysis of Variance Table
##
## Response: Weight
##            Df   Sum Sq  Mean Sq  F value  Pr(>F)
## species     1  1912399  1912399  127.525 < 2e-16 ***
## Length2     1 15253915 15253915 1017.184 < 2e-16 ***
## Height      1   769890   769890   51.339 2.9e-11 ***
## Residuals 156  2339411    14996
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#The anova table of this model
plot(lm(Weight~species+Length2+Height,data=fishdata))
```


Residuals vs Fitted
Residuals
400
200
0
-200
143
144
145
0
500
1000
Fitted values
lm(Weight ~ species + Length2 + Height)


Normal Q-Q
Standardized residuals
3
2
1
0
-1
-2
143
144
145
-2
-1
0
1
2
Theoretical Quantiles
lm(Weight ~ species + Length2 + Height)






```
#P1. residual vs fitted

#the model should be modified, the residual should be randomly distribu
```

```
ted

#P2. normal Q-Q plot

#the model suitable for the normal distribution, however there is few u
nsual point.

#P3. scale-location plot

#We can see that there is relationship between FItted values and standa
rd variance.

#P4. residual vs leverage

#There are few leverage points. In reality, many fish can grow very lar
ge or small in the same growth environment, possibly due to genetic mut
ations or other factors. And we can see from the figure that there are
only few leverage points, which are in line with the actual situation.

library(MPV)

## Warning: package 'MPV' was built under R version 3.6.3

## Loading required package: KernSmooth

## KernSmooth 2.23 loaded
## Copyright M. P. Wand 1997-2009

##
## Attaching package: 'MPV'

## The following object is masked from 'package:olsrr':
##
##     cement

library(car)

## Warning: package 'car' was built under R version 3.6.3

## Loading required package: carData

## Warning: package 'carData' was built under R version 3.6.3

length2=fishdata$Length2
length3=fishdata$Length3
height=fishdata$Height
weight=fishdata$Weight
species=fishdata$species

s2<-sqrt(sum((length2-mean(length2))^2)/(159-1))
s3<-sqrt(sum((height-mean(height))^2)/(159-1))
```

```
s1<-sqrt(sum((species-mean(species))^2)/(159-1))
sy<-sqrt(sum((weight-mean(weight))^2)/(159-1))



z2<-(length2-mean(length2))/s1
z3<-(height-mean(height))/s3
z1<-(species-mean(species))/s3
ys<-(weight-mean(weight))/sy
fishmodel1=lm(ys~z1+z2)
fishmodel2=lm(ys~z1+z3)
fishmodel3=lm(ys~z3+z2)
vif(fishmodel1)

##      z1      z2
## 1.08461 1.08461

vif(fishmodel2)

##       z1       z3
## 1.948989 1.948989

vif(fishmodel3)

##       z3       z2
## 1.689513 1.689513
```

1.08641<10, 1.948989<10 and 1.689513<10. Therefore, we cannot see Species, Length2 and Height have serious molticolinearity. So our final model should be Weight~species+Length2+Height (yhat= -632.942+ 30.481x1+ 24.456x2+ 29.746x3).

According to our initial hypothesis and final model, it can be estimated that the length 1 and height are the same. Bream, Parkki, Perch, Pike, Roach, Smelt, Whitefish increase in weight. When the species of fish is the same, the weight and length of the fish also show a positive correlation with the height.

During the analysis, we encountered extreme points, but we thought it was normal. The relationship between the length and weight of fish may be influenced by many factors, such as season, habitat, gonadal maturity, sex, diet and satiety, health and preservation techniques, and differences in capture sample length. (Kuriakose,2020). There are often various other factors that cause a fish to be very heavy or very light. We also need to pay attention to this when cultivating or fishing fish.

Reference

Fish Market (data), retrieved from: https://lionbridge.ai/datasets/10-open-datasets-for-linear-regression/

Kuriakose, S., n.d. ESTIMATION OF LENGTH WEIGHT RELATIONSHIP IN FISHES. [online] Core.ac.uk. Available at:<https://.ac.uk/download/pdf/95776221.pdf> [Accessed 10 December 2020].