Model_Selection_Of_Abalones

Brook Tarekegn Assefa

2022-11-14

Model Selection Of Abalones

Import the necessary packages

Import the data and look at the first six rows

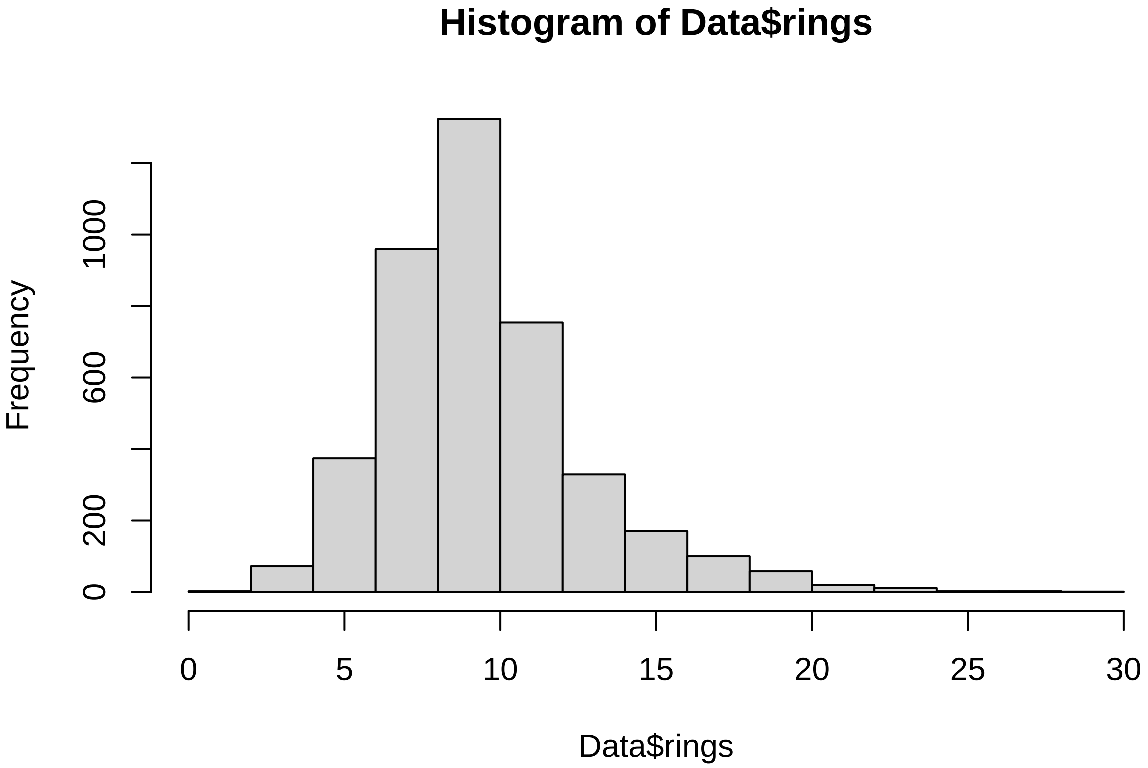
```
Data <- read.csv("Data_Set--Abalone_Marine_Snails--With_Column_Names.csv")
head(Data)
```

```
##   sex length diameter height whole_weight shucked_weight viscera_weight
## 1   M  0.455    0.365  0.095       0.5140         0.2245         0.1010
## 2   M  0.350    0.265  0.090       0.2255         0.0995         0.0485
## 3   F  0.530    0.420  0.135       0.6770         0.2565         0.1415
## 4   M  0.440    0.365  0.125       0.5160         0.2155         0.1140
## 5   I  0.330    0.255  0.080       0.2050         0.0895         0.0395
## 6   I  0.425    0.300  0.095       0.3515         0.1410         0.0775
##   shell_weight rings
## 1        0.150    15
## 2        0.070     7
## 3        0.210     9
## 4        0.155    10
## 5        0.055     7
## 6        0.120     8
```

Let us first look at the Histogram of the rings Dataset, it seems to be slightly

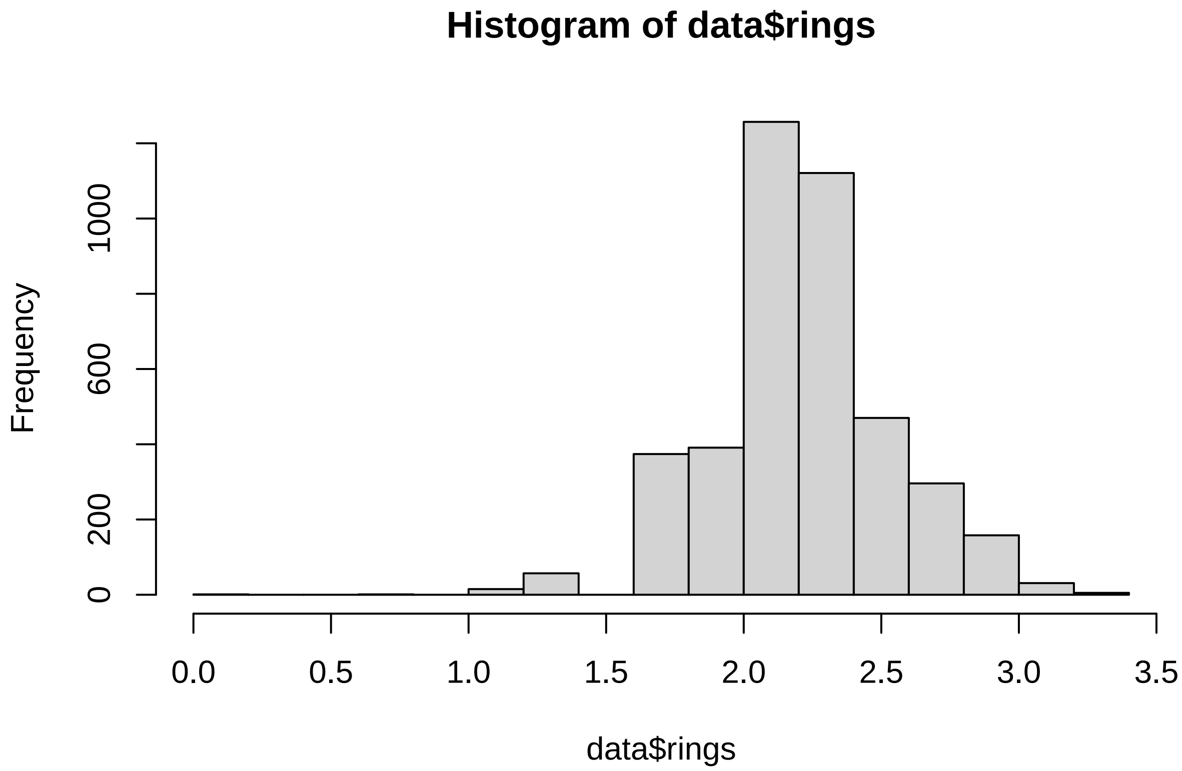
right skewed

```
hist(Data$rings)
```



We might have to consider the Log transformation to flatten out the data and create a fairly Normal distribution ***Waiting for comment from other teammates***

```
data<- Data %>%
  select(-c(sex))
data<- log(data)
hist(data$rings)
```



### First, perform all possible regressions

The age and number of rings of a blacklip abalone has a moderate correlation with all variables other than sex and shucked weight of the abalone, for which correlation is low.

The following order was suggested in our proposal: shell weight, shucked weight, diameter, whole weight, sex, viscera weight, and height. As we can observer the if we consider one predictor model, what is the best one predictor? This turns out to be the shell_weight as stated in our proposal. If I consider two predictive model what turned out to be the best two predictive model are shell weight and shucked weight respectively; and the process goes on as per the suggested order.

```
allregressions <- regsubsets(rings~., force.out=NULL, data=Data, nbest=1)
summary(allregressions)
```

```
## Subset selection object
## Call: regsubsets.formula(rings ~ ., force.out = NULL, data = Data,
##     nbest = 1)
## 9 Variables  (and intercept)
##                 Forced in Forced out
## sexI                FALSE      FALSE
## sexM                FALSE      FALSE
## length              FALSE      FALSE
## diameter            FALSE      FALSE
## height              FALSE      FALSE
## whole_weight        FALSE      FALSE
## shucked_weight      FALSE      FALSE
## viscera_weight      FALSE      FALSE
## shell_weight        FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          sexI sexM length diameter height whole_weight shucked_weight
## 1  ( 1 ) " "  " "  " "    " "      " "    " "          " "
## 2  ( 1 ) " "  " "  " "    " "      " "    " "          "*"
## 3  ( 1 ) " "  " "  " "    "*"      " "    " "          "*"
## 4  ( 1 ) "*"  " "  " "    "*"      " "    " "          "*"
## 5  ( 1 ) "*"  " "  " "    "*"      " "    "*"          "*"
## 6  ( 1 ) "*"  " "  " "    "*"      " "    "*"          "*"
## 7  ( 1 ) "*"  " "  " "    "*"      "*"    "*"          "*"
## 8  ( 1 ) "*"  "*"  " "    "*"      "*"    "*"          "*"
##          viscera_weight shell_weight
## 1  ( 1 ) " "            "*"
## 2  ( 1 ) " "            "*"
## 3  ( 1 ) " "            "*"
## 4  ( 1 ) " "            "*"
## 5  ( 1 ) " "            "*"
## 6  ( 1 ) "*"            "*"
## 7  ( 1 ) "*"            "*"
## 8  ( 1 ) "*"            "*"
```

From this we would like to extract the best model based on the following criteria:- **criteria 1**:- Having High/Maximum Adjusted R Squared value **criteria 2**:- Having Low Mallow's CP **criteria 3**:- Having Low BIC The Mallow's Cp and BIC are usually similar.

### These can now be extracted from the summary

```
names(summary(allregressions))
```

```
## [1] "which"  "rsq"    "rss"    "adjr2"  "cp"     "bic"    "outmat" "obj"
```

adjusted R 2 ?

```
which.max(summary(allregressions)$adjr2)
```

```
## [1] 7
```

Mallow's Cp ?

```
which.min(summary(allregressions)$cp)
```

```
## [1] 7
```

BIC ?

```
which.min(summary(allregressions)$bic)
```

```
## [1] 7
```

Based on the above response of the *allregressions object* we can observe that Model 7 has the best adjusted R2 , Mallow's Cp and BIC. Therefore we can get the corresponding coefficients and predictors of these models, and they all have shell_weight, sex, diameter, height, whole_weight, shucked_weight and viscera_weight.

Since Model 7 is the best model selected we can get the coefficient as follows:-

```
coef(allregressions, which.max(summary(allregressions)$adjr2))
```

```
##    (Intercept)           sexI       diameter         height   whole_weight
##       3.915734      -0.860655      10.538302      10.725118       8.974326
## shucked_weight viscera_weight   shell_weight
##     -19.769037     -10.648149       8.749681
```

We will then use the **Forward selection** , **Backward Selection** and the **Bidirectional selection** to find the best model according to AIC?

Intercept only model

```
regnull <- lm(rings~1, data=Data)
```

Model with all predictors

```
regfull <- lm(rings~., data=Data)
```

Let us first carry out the Forward selection.

```
step(regnull, scope=list(lower=regnull, upper=regfull), direction="forward")
```

```
## Start:  AIC=9780.82
## rings ~ 1
##
##                  Df Sum of Sq   RSS    AIC
## + shell_weight    1   17097.2 26313 7691.7
## + diameter        1   14335.7 29075 8108.6
## + height          1   13490.7 29920 8228.2
## + length          1   13454.5 29956 8233.3
## + whole_weight    1   12676.8 30734 8340.3
## + viscera_weight  1   11019.1 32392 8559.8
## + sex             2    8381.1 35030 8888.8
## + shucked_weight  1    7689.9 35721 8968.4
## <none>                        43411 9780.8
##
## Step:  AIC=7691.7
## rings ~ shell_weight
##
##                  Df Sum of Sq   RSS    AIC
## + shucked_weight  1    3476.1 22837 7101.9
## + whole_weight    1    1740.8 24573 7407.8
## + viscera_weight  1    1067.0 25246 7520.8
## + sex             2     553.1 25760 7607.0
## + height          1     259.3 26054 7652.3
## <none>                        26313 7691.7
## + diameter        1      10.2 26303 7692.1
## + length          1       9.9 26304 7692.1
##
## Step:  AIC=7101.9
## rings ~ shell_weight + shucked_weight
##
##                  Df Sum of Sq   RSS    AIC
## + diameter        1   1233.62 21604 6871.9
## + length          1    978.17 21859 6921.0
## + sex             2    847.75 21990 6947.9
## + whole_weight    1    803.77 22034 6954.2
## + height          1    802.26 22035 6954.5
## + viscera_weight  1     73.91 22763 7090.4
## <none>                        22837 7101.9
##
## Step:  AIC=6871.94
## rings ~ shell_weight + shucked_weight + diameter
##
##                  Df Sum of Sq   RSS    AIC
## + whole_weight    1    549.30 21054 6766.4
## + sex             2    554.46 21049 6767.3
## + height          1    308.62 21295 6813.8
## <none>                        21604 6871.9
## + viscera_weight  1      3.93 21600 6873.2
## + length          1      2.58 21601 6873.4
##
## Step:  AIC=6766.36
## rings ~ shell_weight + shucked_weight + diameter + whole_weight
##
##                  Df Sum of Sq   RSS    AIC
## + sex             2    455.23 20599 6679.1
## + viscera_weight  1    259.38 20795 6716.6
## + height          1    257.21 20797 6717.0
## <none>                        21054 6766.4
## + length          1      8.56 21046 6766.7
##
## Step:  AIC=6679.06
## rings ~ shell_weight + shucked_weight + diameter + whole_weight +
##     sex
##
##                  Df Sum of Sq   RSS    AIC
## + viscera_weight  1   302.180 20297 6619.3
## + height          1   211.258 20388 6638.0
## <none>                        20599 6679.1
## + length          1     3.004 20596 6680.5
##
## Step:  AIC=6619.33
## rings ~ shell_weight + shucked_weight + diameter + whole_weight +
##     sex + viscera_weight
##
##          Df Sum of Sq   RSS    AIC
## + height  1   235.947 20061 6572.5
## <none>                20297 6619.3
## + length  1     0.004 20297 6621.3
##
## Step:  AIC=6572.49
## rings ~ shell_weight + shucked_weight + diameter + whole_weight +
##     sex + viscera_weight + height
##
##          Df Sum of Sq   RSS    AIC
## <none>                20061 6572.5
## + length  1     0.309 20061 6574.4
```

```
##
```

```
## Call:
## lm(formula = rings ~ shell_weight + shucked_weight + diameter +
##     whole_weight + sex + viscera_weight + height, data = Data)
##
## Coefficients:
##    (Intercept)    shell_weight  shucked_weight        diameter    whole_weight
##        3.87038         8.75078       -19.80258        10.56951         8.97751
##           sexI            sexM  viscera_weight          height
##       -0.82644         0.05755       -10.61279        10.74911
```

Therefore we can understand from the **Forward selection** all predictors are included except the **length predictor**. Indicating that the seven predictors listed below are significant predictors.

Therefore the regression equation selected is as the following for y hat:-

$$\hat{y} = 3.87038 + 8.75078x_1 - 19.80258x_2 + 10.56951x_3 + 8.97751x_4 - 0.76889x_5 - 10.61279x_6 + 10.74911x_7$$

Let us now carry out the Backward selection.

```
step(regfull, scope=list(lower=regnull, upper=regfull), direction="backward")
```

```
## Start:  AIC=6574.43
## rings ~ sex + length + diameter + height + whole_weight + shucked_weight +
##     viscera_weight + shell_weight
##
##                  Df Sum of Sq   RSS    AIC
## - length          1      0.31 20061 6572.5
## <none>                        20061 6574.4
## - diameter        1    119.03 20180 6597.1
## - height          1    236.25 20297 6621.3
## - shell_weight    1    290.82 20352 6632.5
## - viscera_weight  1    322.07 20383 6639.0
## - sex             2    445.15 20506 6662.1
## - whole_weight    1    737.01 20798 6723.1
## - shucked_weight  1   2821.38 22882 7122.1
##
## Step:  AIC=6572.49
## rings ~ sex + diameter + height + whole_weight + shucked_weight +
##     viscera_weight + shell_weight
##
##                  Df Sum of Sq   RSS    AIC
## <none>                        20061 6572.5
## - height          1    235.95 20297 6619.3
## - shell_weight    1    291.71 20353 6630.8
## - viscera_weight  1    326.87 20388 6638.0
## - sex             2    448.49 20510 6660.8
## - diameter        1    549.86 20611 6683.4
## - whole_weight    1    737.45 20798 6721.3
## - shucked_weight  1   2842.24 22903 7123.9

##
## Call:
## lm(formula = rings ~ sex + diameter + height + whole_weight +
##     shucked_weight + viscera_weight + shell_weight, data = Data)
##
```

```
## Coefficients:
##    (Intercept)             sexI             sexM         diameter           height
##        3.87038         -0.82644          0.05755         10.56951         10.74911
##   whole_weight   shucked_weight  viscera_weight     shell_weight
##        8.97751        -19.80258        -10.61279          8.75078
```

Therefore we can understand from the **Backward selection** also includes all predictors except the **length predictor**. Indicating that the seven predictors listed below are significant predictors. T ### Therefore the regression equation selected is as the following for y hat:-

$$\hat{y} = 3.87038 + 8.75078x_1 - 19.80258x_2 + 10.56951x_3 + 8.97751x_4 - 0.76889x_5 - 10.61279x_6 + 10.74911x_7$$

This doesn't help much as it is similar to the forward selection.

Since in our proposal we have mentioned that the observed correlation among the rings has a moderate correlation with all variables other than sex and shucked weight of the abalone. We would like to force out these values and perform the **Backward selection**.

```
data_set_with_numeric_sex <-
    Data %>%
    mutate(sex = replace(sex, sex == 'M', 0)) %>%
    mutate(sex = replace(sex, sex == 'F', 1)) %>%
    mutate(sex = replace(sex, sex == 'I', 2))
data_set_with_numeric_sex$sex <- as.numeric(data_set_with_numeric_sex$sex)
correlation_matrix <- cor(data_set_with_numeric_sex)
correlation_matrix
```

```
##                       sex     length   diameter     height whole_weight
## sex             1.0000000 -0.4487653 -0.4582451 -0.4179278   -0.4612384
## length         -0.4487653  1.0000000  0.9868116  0.8275536    0.9252612
## diameter       -0.4582451  0.9868116  1.0000000  0.8336837    0.9254521
## height         -0.4179278  0.8275536  0.8336837  1.0000000    0.8192208
## whole_weight   -0.4612384  0.9252612  0.9254521  0.8192208    1.0000000
## shucked_weight -0.4409269  0.8979137  0.8931625  0.7749723    0.9694055
## viscera_weight -0.4546577  0.9030177  0.8997244  0.7983193    0.9663751
## shell_weight   -0.4455492  0.8977056  0.9053298  0.8173380    0.9553554
## rings          -0.3518216  0.5567196  0.5746599  0.5574673    0.5403897
##                shucked_weight viscera_weight shell_weight      rings
## sex                -0.4409269     -0.4546577   -0.4455492 -0.3518216
## length              0.8979137      0.9030177    0.8977056  0.5567196
## diameter            0.8931625      0.8997244    0.9053298  0.5746599
## height              0.7749723      0.7983193    0.8173380  0.5574673
## whole_weight        0.9694055      0.9663751    0.9553554  0.5403897
## shucked_weight      1.0000000      0.9319613    0.8826171  0.4208837
## viscera_weight      0.9319613      1.0000000    0.9076563  0.5038192
## shell_weight        0.8826171      0.9076563    1.0000000  0.6275740
## rings               0.4208837      0.5038192    0.6275740  1.0000000
```

Model with all predictors except excluding the sex and shucked weight of the abalone.

```
regfull_excluding_sex_shucked_weight <- lm(rings~ diameter+height+whole_weight+viscera_weight+shell_wei
```

Perform the Backward elimination from the regfull

```
step(regfull_excluding_sex_shucked_weight, scope=list(lower=regnull, upper=regfull_excluding_sex_shucked
```

```
## Start:  AIC=7224.2
## rings ~ diameter + height + whole_weight + viscera_weight + shell_weight +
##     length
##
##                  Df Sum of Sq   RSS    AIC
## - viscera_weight  1       4.5 23475 7223.0
## <none>                        23471 7224.2
## - length          1      39.9 23511 7229.3
## - diameter        1     201.5 23672 7257.9
## - height          1     356.7 23828 7285.2
## - whole_weight    1     943.8 24415 7386.9
## - shell_weight    1    4227.5 27698 7914.0
##
## Step:  AIC=7223
## rings ~ diameter + height + whole_weight + shell_weight + length
##
##                Df Sum of Sq   RSS    AIC
## <none>                      23475 7223.0
## - length        1      42.1 23518 7228.5
## - diameter      1     203.7 23679 7257.1
## - height        1     353.2 23828 7283.4
## - whole_weight  1    2268.4 25744 7606.3
## - shell_weight  1    4500.5 27976 7953.6

##
## Call:
## lm(formula = rings ~ diameter + height + whole_weight + shell_weight +
##     length, data = Data)
##
## Coefficients:
##  (Intercept)      diameter        height  whole_weight  shell_weight
##        3.702        14.387        13.085        -5.959        26.202
##       length
##       -5.308
```

After excluding the sex and shucked_weight we can observe that this change has rendered the **viscera weight** as insignificant predictor. ### Therefore our adjusted regression equation for y hat is :-

$$\hat{y} = 3.702 + 14.387x_1 + 13.085x_2 - 5.959x_3 + 26.202x_4 - 5.308x_5$$

If We would like to force out the sex and shucked weight of the abalone and perform the **Bidirectional selection**, we would have found the same result as the one stated above.

But if we do perform the **Forward selection** we can see that the *viscera_weight* weight is considered as a significant predictor , which was dropped by **Backward and Bidirectional selections.**

```
step(regfull_excluding_sex_shucked_weight, scope=list(lower=regnull, upper=regfull_excluding_sex_shucked
```

```
## Start:  AIC=7224.2
## rings ~ diameter + height + whole_weight + viscera_weight + shell_weight +
##     length
```

```
##
## Call:
## lm(formula = rings ~ diameter + height + whole_weight + viscera_weight +
##     shell_weight + length, data = Data)
##
## Coefficients:
##    (Intercept)        diameter          height    whole_weight  viscera_weight
##          3.674          14.318          13.176          -5.671          -1.203
##   shell_weight          length
##         26.017          -5.183
```

### Therefore our Forward selection regression equation for y hat is :-

$$\hat{y} = 3.674 + 14.318x_1 + 13.176x_2 - 5.671x_3 - 1.203x_4 + 26.017x_5 - 5.183x_6$$

Therefore **Forward selection**:-

$$x_1, x_2, x_3, x_4, x_5, x_6$$

Backward elimination and **Stepwise regression**

$$x_1, x_2, x_3, x_4, x_5$$

"***Berk [1978]*** has noted that **forward selection** tends to agree with all possible regressions for **small subset sizes** but not for large ones, while b**ackward elimination** tends to agree with all possible regressions for large subset sizes but not for small ones." Based on this our final regression equation and the model selection based on the AIC is excluding the **viscera weight** (which is of a blacklip abalone is the gut weight after bleeding.) When thinking logically about it , it seems like it does not have any relation with the age of an abalone at all.

$$\hat{y} = 3.702 + 14.387x_1 + 13.085x_2 - 5.959x_3 + 26.202x_4 - 5.308x_5$$