Project 2 Linear regression

Math 6375 - 2022 Spring

Achraf cherkaoui

For this project, we use a real data on abalone fishing in Australia (taken from http://archive.ics.uci.edu/ml/machine-learning-databases/abalone). Abalone is a rich nutritious food resource in the many parts of the world. The economic value of abalone is positively correlated with its age. The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope – a boring and time-consuming task. Other measurements, which are easier to obtain, are used to predict the age.

From the original data examples with missing values were removed (the majority having the predicted value missing), and the ranges of the continuous values have been scaled for use with an ANN (by dividing by 200). The provided dataset has 4177 rows and 9 columns. Each row corresponds to a particular caught abalone, and the columns correspond to the following attributes:

`Name / Data Type / Measurement Unit / Description`

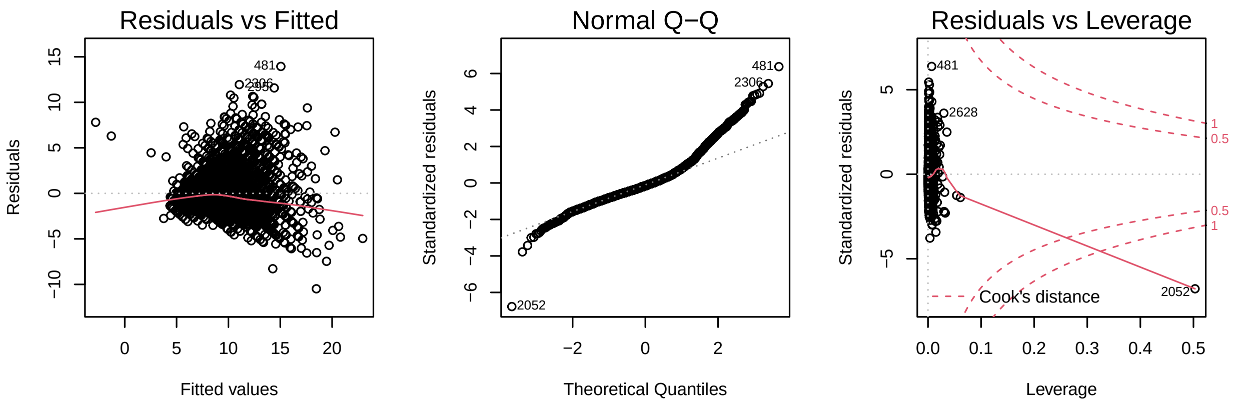
Sex / nominal / – / M, F, and I (infant)
Length / continuous / mm / Longest shell measurement
Diameter / continuous / mm / perpendicular to length
Height / continuous / mm / with meat in shell
Whole weight / continuous / grams / whole abalone
Shucked weight / continuous / grams / weight of meat
Viscera weight / continuous / grams / gut weight (after bleeding)
Shell weight / continuous / grams / after being dried
Rings / integer / – / +1.5 gives the age in years

By checking the correlation matrix of all of the numeric measurements, or based on common sense, these measurements are highly correlated. I fit the linear regression including all features, and made the diagnosis plots and found a couple of outliers and abnormality. Hence I removed these outliers and transformed `rings` to be `log2(rings)`. The diagnosis plots improved a little bit. All the following is based on the transformed data.

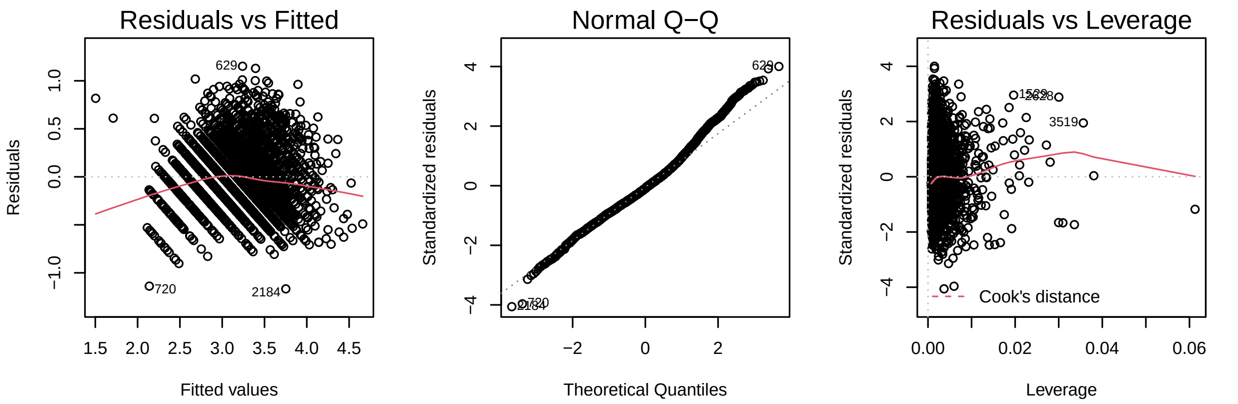
```
library(ggplot2)

abalone=read.table("http://archive.ics.uci.edu/ml/machine-learning-databases/abalone/abalone.data", sep=

abnames = c('sex','length','diameter','height','wt.w','wt.s', 'wt.v','wt.sh','rings')
names(abalone)=abnames

View(abalone)
# check correlation matrix
round(cor(x=abalone[,2:ncol(abalone)]) ,3)
```

```
##          length diameter height  wt.w  wt.s  wt.v wt.sh rings
## length    1.000    0.987  0.828 0.925 0.898 0.903 0.898 0.557
## diameter  0.987    1.000  0.834 0.925 0.893 0.900 0.905 0.575
## height    0.828    0.834  1.000 0.819 0.775 0.798 0.817 0.557
## wt.w      0.925    0.925  0.819 1.000 0.969 0.966 0.955 0.540
## wt.s      0.898    0.893  0.775 0.969 1.000 0.932 0.883 0.421
## wt.v      0.903    0.900  0.798 0.966 0.932 1.000 0.908 0.504
## wt.sh     0.898    0.905  0.817 0.955 0.883 0.908 1.000 0.628
## rings     0.557    0.575  0.557 0.540 0.421 0.504 0.628 1.000
```

```
# fit linear model
model = lm(rings~.,data=abalone)
#summary(model)
par(mfrow=c(1,3))
plot(model, which=c(1,2,5))
```



```
# remove outliers and tranform the response
outliers =c(2052,1418,237)
abalone = abalone[!row.names(abalone)%in%outliers,]
abalone$rings = log2(abalone$rings)

# fit linear model again
model = lm(rings~.,data=abalone)
#summary(model)
par(mfrow=c(1,3))
plot(model, which=c(1,2,5))
```

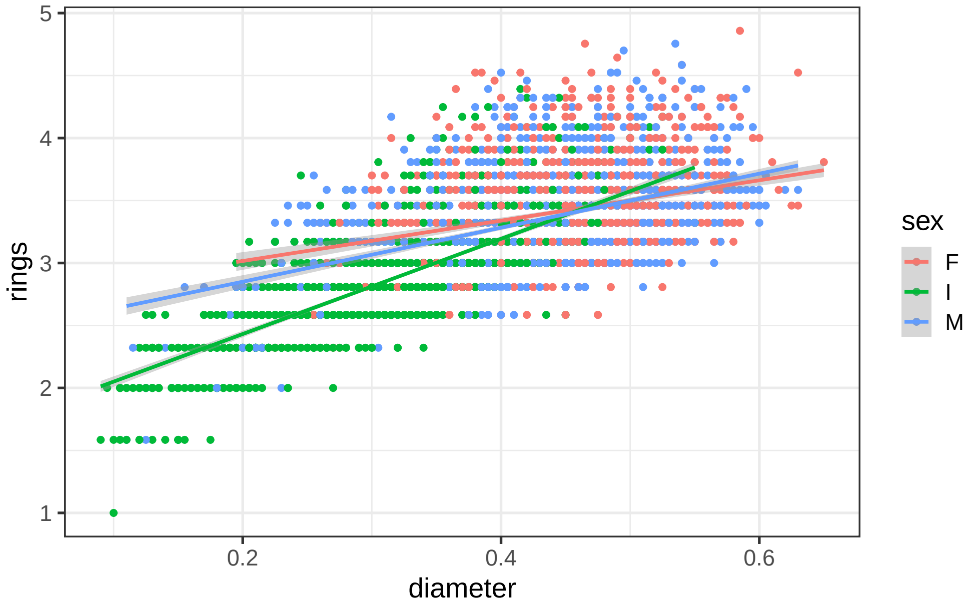


Objective: now, we would like to know if there is a significantly different relation between

diameter and the transformed rings (related with age) among groups of sex and what is the relation for each group. To help answer the question, I first made the scatterplot to visualize the data.

```
# make the plot to help visualize the relation
ggplot(abalone,aes(x=diameter,y=rings,color=sex) )+
  geom_point()+
  geom_smooth(method=lm)+
  theme_bw(base_size=16)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



We will need a few pieces of information or take a couple of steps to investigate the objective.

- We will need to fit linear models with interaction terms to investigate the difference of effects:

Question 1: Fit the model, and show the summary.

Answer1:

```
model_1 <- lm(rings ~ sex*diameter, data = abalone)
summary(model_1)
```

```
##
## Call:
## lm(formula = rings ~ sex * diameter, data = abalone)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1572 -0.2177 -0.0573  0.1651  1.3108
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.69307    0.05876  45.830  < 2e-16 ***
## sexI          -1.02332    0.06810 -15.027  < 2e-16 ***
## sexM          -0.27520    0.07364  -3.737 0.000189 ***
## diameter       1.61502    0.12766  12.651  < 2e-16 ***
## sexI:diameter  2.19450    0.16324  13.443  < 2e-16 ***
## sexM:diameter  0.54676    0.16170   3.381 0.000728 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3271 on 4168 degrees of freedom
## Multiple R-squared:  0.4914, Adjusted R-squared:  0.4908
## F-statistic: 805.4 on 5 and 4168 DF,  p-value: < 2.2e-16
```

- Since `sex` has 3 levels, we will need to use the function `relevel` to **change the baseline and fit the linear model again**, so that we can compare the difference of effects of diameter for each of the 3 pairs: “Female” vs “Male”, “Female” vs “Infant”, “Male” vs “Infant”.

Question 2: Use the function ‘relevel‘ to change the baseline category for the variable ‘sex‘ and fit the linear model again.

Answer2:

```
abalone$sex <- as.factor(abalone$sex)
abalone$sex <- relevel(abalone$sex, ref = 'I')

summary(lm(rings ~ sex*diameter, data = abalone))
```

```
##
## Call:
## lm(formula = rings ~ sex * diameter, data = abalone)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1572 -0.2177 -0.0573  0.1651  1.3108
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.66975    0.03442   48.52   <2e-16 ***
## sexF           1.02332    0.06810   15.03   <2e-16 ***
## sexM           0.74812    0.05617   13.32   <2e-16 ***
## diameter       3.80952    0.10174   37.44   <2e-16 ***
## sexF:diameter -2.19450    0.16324  -13.44   <2e-16 ***
## sexM:diameter -1.64773    0.14213  -11.59   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3271 on 4168 degrees of freedom
## Multiple R-squared:  0.4914, Adjusted R-squared:  0.4908
## F-statistic: 805.4 on 5 and 4168 DF,  p-value: < 2.2e-16
```

- Use the above results, please answer the following questions to discuss the different effects of `diameter` to the transformed `rings` among the 3 `sex` groups by **checking the corresponding coefficient estimates and the *p*-values:**
  - **Question 3:Are the effects significantly different among groups?** For example, are the effects significantly different between `female` and `male`, between `female` and `infant`, and between `infant` and `male`?

Answer3: The effects of diameters on rings in the three groups of Sex is significant based on the p-values (< 0.05) corresponding to the coefficients of appropriate interaction terms in the two models.Also the effect of diameters on rings in each of the 3 sex groups is significant.(“Female” vs “Male”, “Female” vs “Infant”, “Male” vs “Infant”)

- **Question 4:How different?** For example, if the effects of diameter to rings are significantly different between `female` group and `infant` group, then tell if the effects of female group is significantly larger or smaller than that of the infant group.

Answer4:

The additional effect of diameter on rings in infants in comparison to females is the largest. Then we have the additional effect diameters on rings in infants with respect to males.Finally, The additional effect of diameters on rings in males with respect to females is the smallest.

- **Question 5: What are the effects of ‘diameter‘ to ‘rings‘ for each group?** The problem basically asks you to write the mathematical expression for the fitted model for each sex group, separately. You should have 3 equations in total. And explain and interpret the coefficient estimate of diameter for each sex group.

Answer5: The mathematical expression for the fitted model for each sex group.

$$E(Y|X) = \begin{cases} \hat{\beta}_0 + \hat{\beta}_1 X_1, & \text{female} \\ \hat{\beta}_0 + \hat{\beta}_2 + (\hat{\beta}_1 + \hat{\beta}_4) X_1, & \text{male} \\ \hat{\beta}_0 + \hat{\beta}_3 + (\hat{\beta}_1 + \hat{\beta}_5) X_1, & \text{infant} \end{cases} \tag{1}$$

Group	Slope
Female	$\hat{\beta}_1 = 1.61502$
Male	$\hat{\beta}_1 + \hat{\beta}_4 = 2.16178$
Infant	$\hat{\beta}_1 + \hat{\beta}_5 = 3.80952$

For a fixed value of the diameter X1, the female abalone rings will grow by $\hat{\beta}_1 = 1.61502$ on average. The growing rings rate for male will be $\hat{\beta}_1 + \hat{\beta}_4 = 2.16178$.Infants will have the most growing rings rate by $\hat{\beta}_1 + \hat{\beta}_5 = 3.80952$.