36401-B Data Analysis Exam 2

Joong Ho Choi

TOTAL POINTS

89 / 100

QUESTION 1

1 Introduction 20 / 20

✓ **- 0 pts Correct**

- 20 pts Missing

- 10 pts Insufficient work

QUESTION 2

2 EDA 12 / 20

- 0 pts Correct

✓ **- 4 pts Also investigate univariate relationships (summary statistics, analysis of missing values etc.)**

- 4 pts Also investigate bivariate relationships such as scatterplots between independent and dependent variables

✓ **- 4 pts Hard to see you analysis visually -- figures / tables not legible**

- 10 pts Look at analysis of variables other than age only.

QUESTION 3

3 Modeling 20 / 20

✓ **- 0 pts Correct**

- 2 pts Did not mention strengths/weaknesses of linear regression

- 2 pts Did not mention strengths/weaknesses of additive model

- 2 pts Did not mention strengths/weaknesses of regression forest

- 5 pts Method is not adequate to assess predictive accuracy of model on new data. (Ex., running models on all data and not using test/train split).

- 2 pts Did not consider transformations for linear model

- 1 pts Not clear what model(s) are being tested

- 5 pts Incorrect method for GAM

- 5 pts Lacks comments or interpretation

- 2 pts Incorrect interpretation/diagnostics of RF model

1 Mention your method: using a train/test split. You didn't mention this, so I had to look at your code.

QUESTION 4

4 Diagnostics and Model Selection 17 / 20

- 0 pts Correct

- 2 pts labeling of plots makes narrative unclear

- 2 pts No QQ plots

- 2 pts Residual analysis is insufficient

- 10 pts Diagnostic modeling only described at surface level; no in-depth narrative/graph/code present

- 0.5 pts model plots are cut off between pages or cannot read formatting

- 0.1 pts Need to label sections properly

✓ **- 3 pts insufficient analysis of model error and justification of model selection**

QUESTION 5

5 Final Model 20 / 20

✓ **- 0 pts Correct**

- 5 pts More than seven pages.

- 2 pts Only used tree, not random forest.

- 2 pts Prediction error of the linear model seems unrealistically good.

- 2 pts Forest makes predictions by voting from all the trees, not taking the result from the best tree. This is a reason why random forest doesn't overfit much.

- 1 pts GAM has too few covariates, and the performance of GAM should be improved.

- 1 pts MSEs of the three models seem to be too close.

- 2 pts To fairly compare all three models, you should calculate prediction error or R square with the same data. An outlier tends to be hard to predict and increases the mean square error a lot. It's ok you keep outlier out in the training state, but it's not fair for only model 1b to "avoid" those difficult data points when you compare the three models. This explains why your model 1b has way smaller MSE ---- model 1b took an easier exam.

- 4 pts Didn't explain enough about how to choose the final models from the three models.

36401 test2

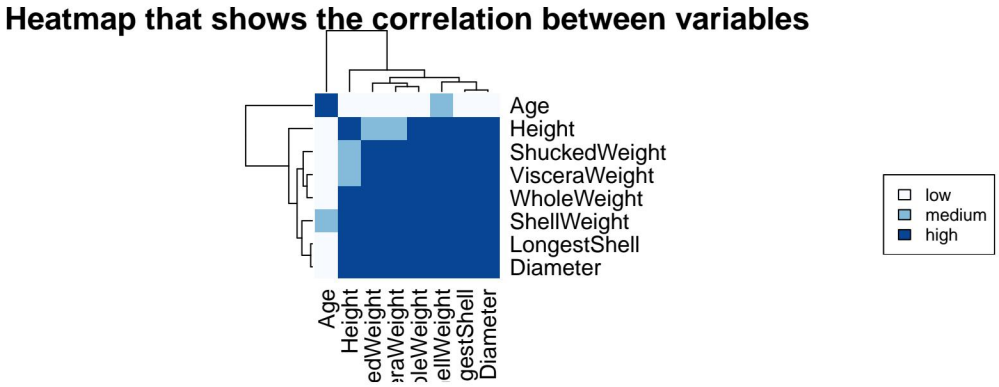
Introduction

```
##   Type LongestShell Diameter Height WholeWeight ShuckedWeight VisceraWeight
## 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
##   ShellWeight       Age
## 1        0.15 16.056874
## 2        0.07  8.151516
## 3        0.21 10.255053
```

The data comes from a file called abalone.csv, and the data's sample size is 4177. As it will not be used in the analysis, variable ID was removed. Excluding Age(our variable of interest), the data has the following 8 variables: -Measurements of the type (male, female and infant) -The longest shell measurement -The diameter -Height -Four weights (whole, shucked, viscera and shell).

While type has been changed from character type to factor type, the rest of variables remain as numeric types. The main research problem is to predict Age from the other eight variables. To do so, we will build and compare predictive accuracy between three types of models: linear, additive and forest. We will also examine which variables are most important.

Exploratory Data Analysis



From the correlation heatmap above, it seems only ShellWeight has medium correlation with Age; every other variable has low correlation with Age. Interestingly, while height has medium correlation with ShuckedWeight and VisceraWeight, the rest of other variables have high correlations with each other.

```
pairs(df[,2:9],pch=19,main="Pairwise plot of quantitative variables in data for EDA")
```

1 Introduction **20 / 20**

✓ **- 0 pts** **Correct**

- 20 pts Missing

- 10 pts Insufficient work

36401 test2

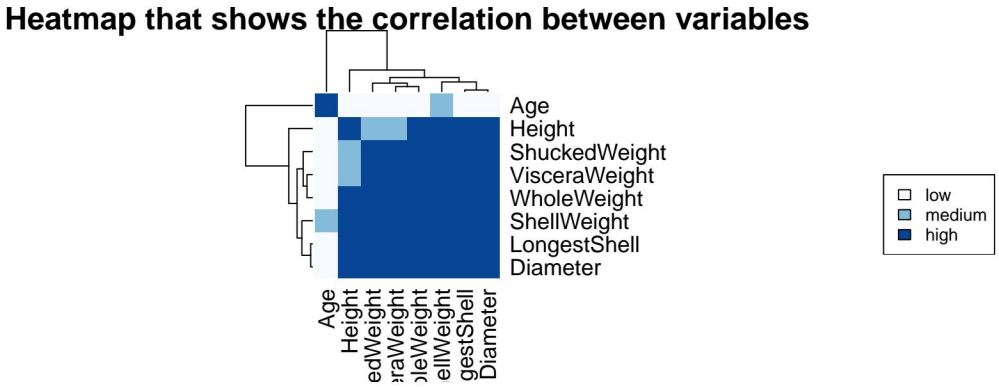
Introduction

```
##   Type LongestShell Diameter Height WholeWeight ShuckedWeight VisceraWeight
## 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
##   ShellWeight       Age
## 1        0.15 16.056874
## 2        0.07  8.151516
## 3        0.21 10.255053
```

The data comes from a file called abalone.csv, and the data's sample size is 4177. As it will not be used in the analysis, variable ID was removed. Excluding Age(our variable of interest), the data has the following 8 variables: -Measurements of the type (male, female and infant) -The longest shell measurement -The diameter -Height -Four weights (whole, shucked, viscera and shell).

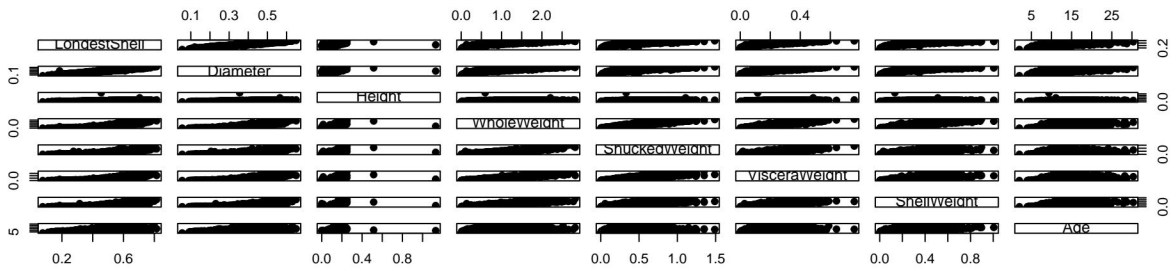
While type has been changed from character type to factor type, the rest of variables remain as numeric types. The main research problem is to predict Age from the other eight variables. To do so, we will build and compare predictive accuracy between three types of models: linear, additive and forest. We will also examine which variables are most important.

Exploratory Data Analysis



From the correlation heatmap above, it seems only ShellWeight has medium correlation with Age; every other variable has low correlation with Age. Interestingly, while height has medium correlation with ShuckedWeight and VisceraWeight, the rest of other variables have high correlations with each other.

```
pairs(df[,2:9],pch=19,main="Pairwise plot of quantitative variables in data for EDA")
```



From the pairwise plot above, it seems age shares logarithm relationship with all of the variables. Thus, when we are considering variable transformation in the future, logarithmic transformation would be the first choice. We will also partition the dataset into training (70%) and testing (30%) to compare three models' predictive accuracy. However, before partitioning, we will remove two data points(1258,3997) as we will likely perform log transformation and these two values can become infinite after log transformation.

Modeling

As the data does not have high number of features and is subsequently not high dimensional, I felt that there was no need for penalization. Thus, I created one linear model and one linear model with logarithm transformation. I decided to use variable transformation, specifically Log, because log transformation can make highly skewed distribution less skewed; this can be used to make data patterns more interpretable and for helping to meet the assumptions of inferential statistics.

```
##
## Call:
## lm(formula = Age ~ ., data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.8772 -1.3338 -0.3237  0.8601 13.8365
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     5.65148    0.35253  16.031  < 2e-16 ***
## TypeI          -0.96383    0.12299  -7.837 6.43e-15 ***
## TypeM           0.01076    0.10041   0.107    0.915
## LongestShell   -1.02573    2.17218  -0.472    0.637
## Diameter       12.05985    2.67500   4.508 6.79e-06 ***
## Height          7.86392    1.67905   4.684 2.95e-06 ***
## WholeWeight     8.14950    0.86900   9.378  < 2e-16 ***
## ShuckedWeight -18.88699    0.98581 -19.159  < 2e-16 ***
## VisceraWeight  -9.96966    1.56553  -6.368 2.21e-10 ***
## ShellWeight    10.24263    1.33077   7.697 1.90e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.204 on 2912 degrees of freedom
## Multiple R-squared:  0.5439, Adjusted R-squared:  0.5425
## F-statistic: 385.9 on 9 and 2912 DF,  p-value: < 2.2e-16

##
## Call:
```

2 EDA 12 / 20

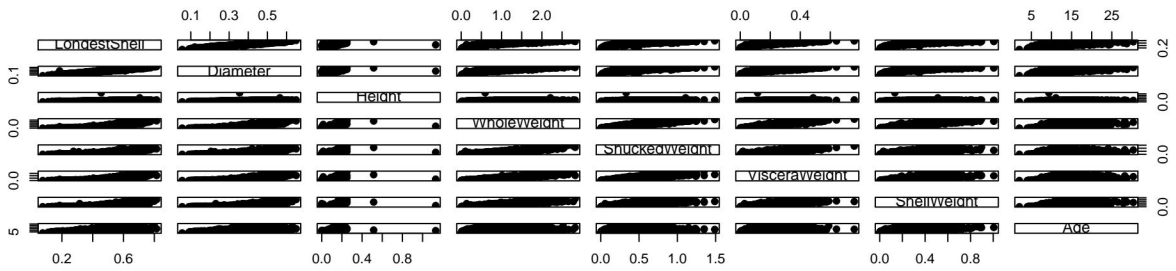
- **0 pts** Correct

✓ **- 4 pts** **Also investigate univariate relationships (summary statistics, analysis of missing values etc.)**

- **4 pts** Also investigate bivariate relationships such as scatterplots between independent and dependent variables

✓ **- 4 pts** **Hard to see you analysis visually -- figures / tables not legible**

- **10 pts** Look at analysis of variables other than age only.



From the pairwise plot above, it seems age shares logarithm relationship with all of the variables. Thus, when we are considering variable transformation in the future, logarithmic transformation would be the first choice. We will also partition the dataset into training (70%) and testing (30%) to compare three models' predictive accuracy. However, before partitioning, we will remove two data points(1258,3997) as we will likely perform log transformation and these two values can become infinite after log transformation.

Modeling

As the data does not have high number of features and is subsequently not high dimensional, I felt that there was no need for penalization. Thus, I created one linear model and one linear model with logarithm transformation. I decided to use variable transformation, specifically Log, because log transformation can make highly skewed distribution less skewed; this can be used to make data patterns more interpretable and for helping to meet the assumptions of inferential statistics.

```
##
## Call:
## lm(formula = Age ~ ., data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.8772 -1.3338 -0.3237  0.8601 13.8365
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     5.65148    0.35253  16.031  < 2e-16 ***
## TypeI          -0.96383    0.12299  -7.837 6.43e-15 ***
## TypeM           0.01076    0.10041   0.107    0.915
## LongestShell   -1.02573    2.17218  -0.472    0.637
## Diameter       12.05985    2.67500   4.508 6.79e-06 ***
## Height          7.86392    1.67905   4.684 2.95e-06 ***
## WholeWeight     8.14950    0.86900   9.378  < 2e-16 ***
## ShuckedWeight -18.88699    0.98581 -19.159  < 2e-16 ***
## VisceraWeight  -9.96966    1.56553  -6.368 2.21e-10 ***
## ShellWeight    10.24263    1.33077   7.697 1.90e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.204 on 2912 degrees of freedom
## Multiple R-squared:  0.5439, Adjusted R-squared:  0.5425
## F-statistic: 385.9 on 9 and 2912 DF,  p-value: < 2.2e-16

##
## Call:
```

```
## lm(formula = Age ~ Type + I(log(LongestShell)) + I(log(Diameter)) +
##     +I(log(Height)) + I(log(WholeWeight) + I(log(ShuckedWeight))) +
##     I(log(VisceraWeight)) + I(log(ShellWeight)), data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.4785 -1.4465 -0.3599  0.8784 16.7423
##
## Coefficients:
##                                           Estimate Std. Error t value
## (Intercept)                                21.0322     0.7312  28.764
## TypeI                                      -0.9521     0.1266  -7.520
## TypeM                                      -0.1386     0.1032  -1.342
## I(log(LongestShell))                       -3.5320     1.1464  -3.081
## I(log(Diameter))                            0.2900     1.0665   0.272
## I(log(Height))                              1.3909     0.3748   3.711
## I(log(WholeWeight) + I(log(ShuckedWeight)))  -2.3223     0.1655 -14.035
## I(log(VisceraWeight))                       0.1075     0.2412   0.446
## I(log(ShellWeight))                         7.3881     0.2816  26.232
##                                           Pr(>|t|)
## (Intercept)                                < 2e-16 ***
## TypeI                                     7.26e-14 ***
## TypeM                                      0.17962
## I(log(LongestShell))                       0.00208 **
## I(log(Diameter))                           0.78568
## I(log(Height))                             0.00021 ***
## I(log(WholeWeight) + I(log(ShuckedWeight)))  < 2e-16 ***
## I(log(VisceraWeight))                      0.65589
## I(log(ShellWeight))                        < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.275 on 2913 degrees of freedom
## Multiple R-squared:  0.514,  Adjusted R-squared:  0.5127
## F-statistic: 385.2 on 8 and 2913 DF,  p-value: < 2.2e-16
```

I created one additive model. As additive model already involves smoothing, I did not apply any transformation on the variables. Since Type is a discrete variable, no smoothing was applied.

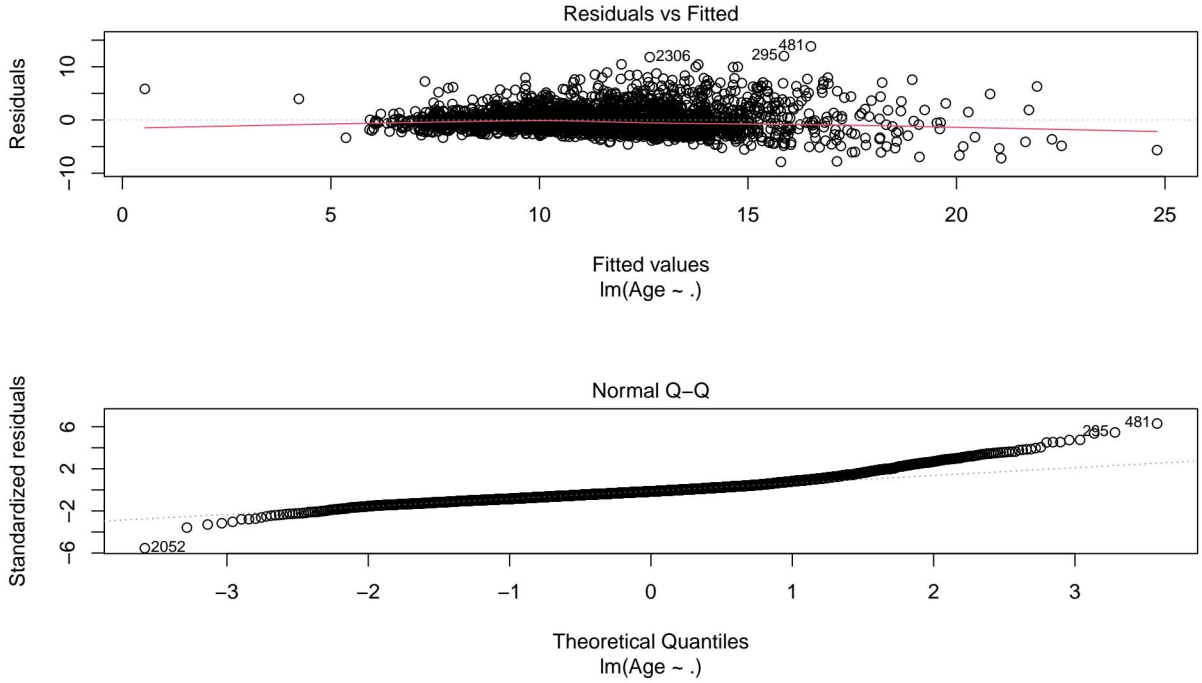
```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Age ~ Type + s(LongestShell) + s(Diameter) + s(Height) + s(WholeWeight) +
##     s(ShuckedWeight) + s(VisceraWeight) + s(ShellWeight)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.70399    0.07595 154.095  < 2e-16 ***
## TypeI       -0.73132    0.12474  -5.863 5.07e-09 ***
## TypeM       -0.02504    0.09732  -0.257    0.797
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Approximate significance of smooth terms:
##                     edf Ref.df      F p-value
## s(LongestShell)   4.982  6.120  2.883 0.00815 **
## s(Diameter)       6.482  7.467  1.786 0.06103 .
## s(Height)         2.922  3.659  4.819 0.00129 **
## s(WholeWeight)    5.977  7.174 26.368 < 2e-16 ***
## s(ShuckedWeight) 4.528  5.727 78.073 < 2e-16 ***
## s(VisceraWeight) 4.032  5.226 10.677 < 2e-16 ***
## s(ShellWeight)    1.000  1.000 61.785 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.576   Deviance explained = 58.1%
## GCV = 4.5516  Scale est. = 4.5003    n = 2922
```

Then, I created one random forest model.

Diagnostics and model selection

The followings are the diagnostic plots for the linear model that used all of the variables without any transformation



From the summary of a linear model above, it seems all the variables, except TypeM and LongestShell, are statistically significant variables as their p-values are less than 0.05.

The followings are the diagnostic plots for the linear model that had all of the variables(except Age) have log transformations applied.

3 Modeling 20 / 20

✓ - 0 pts Correct

- 2 pts Did not mention strengths/weaknesses of linear regression
- 2 pts Did not mention strengths/weaknesses of additive model
- 2 pts Did not mention strengths/weaknesses of regression forest
- 5 pts Method is not adequate to assess predictive accuracy of model on new data. (Ex., running models on all data and not using test/train split).
- 2 pts Did not consider transformations for linear model
- 1 pts Not clear what model(s) are being tested
- 5 pts Incorrect method for GAM
- 5 pts Lacks comments or interpretation
- 2 pts Incorrect interpretation/diagnostics of RF model

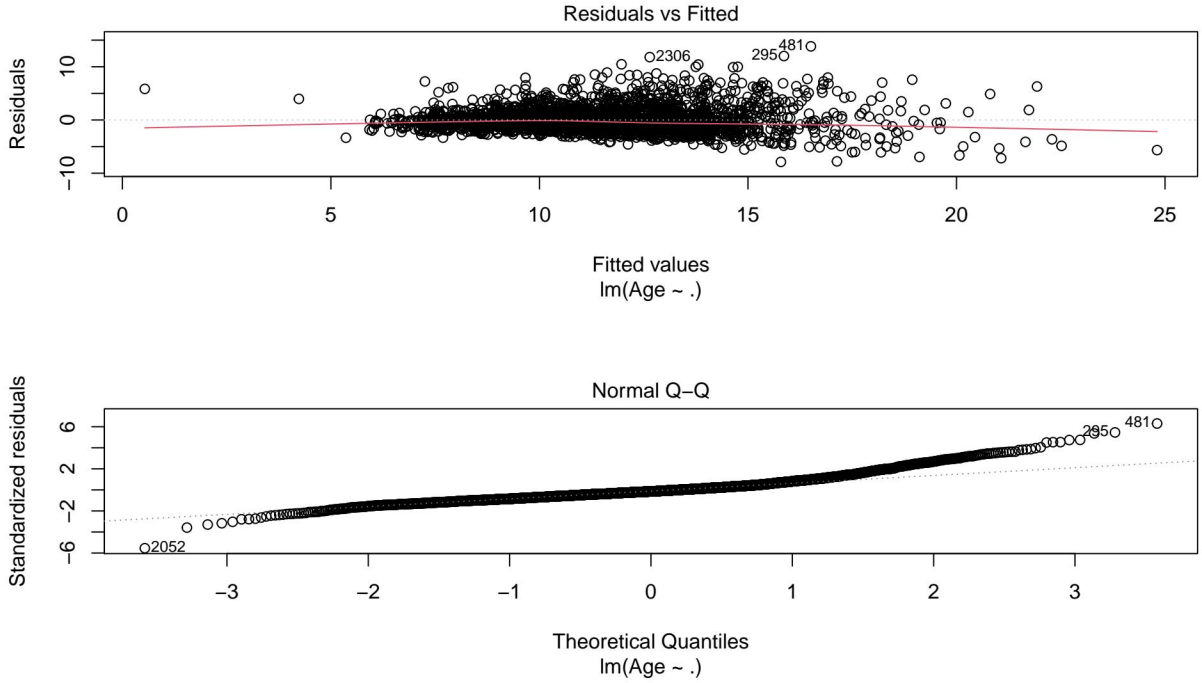
1 Mention your method: using a train/test split. You didn't mention this, so I had to look at your code.

```
##
## Approximate significance of smooth terms:
##                     edf Ref.df      F p-value
## s(LongestShell)   4.982  6.120  2.883 0.00815 **
## s(Diameter)       6.482  7.467  1.786 0.06103 .
## s(Height)         2.922  3.659  4.819 0.00129 **
## s(WholeWeight)    5.977  7.174 26.368 < 2e-16 ***
## s(ShuckedWeight) 4.528  5.727 78.073 < 2e-16 ***
## s(VisceraWeight) 4.032  5.226 10.677 < 2e-16 ***
## s(ShellWeight)    1.000  1.000 61.785 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.576   Deviance explained = 58.1%
## GCV = 4.5516  Scale est. = 4.5003    n = 2922
```

Then, I created one random forest model.

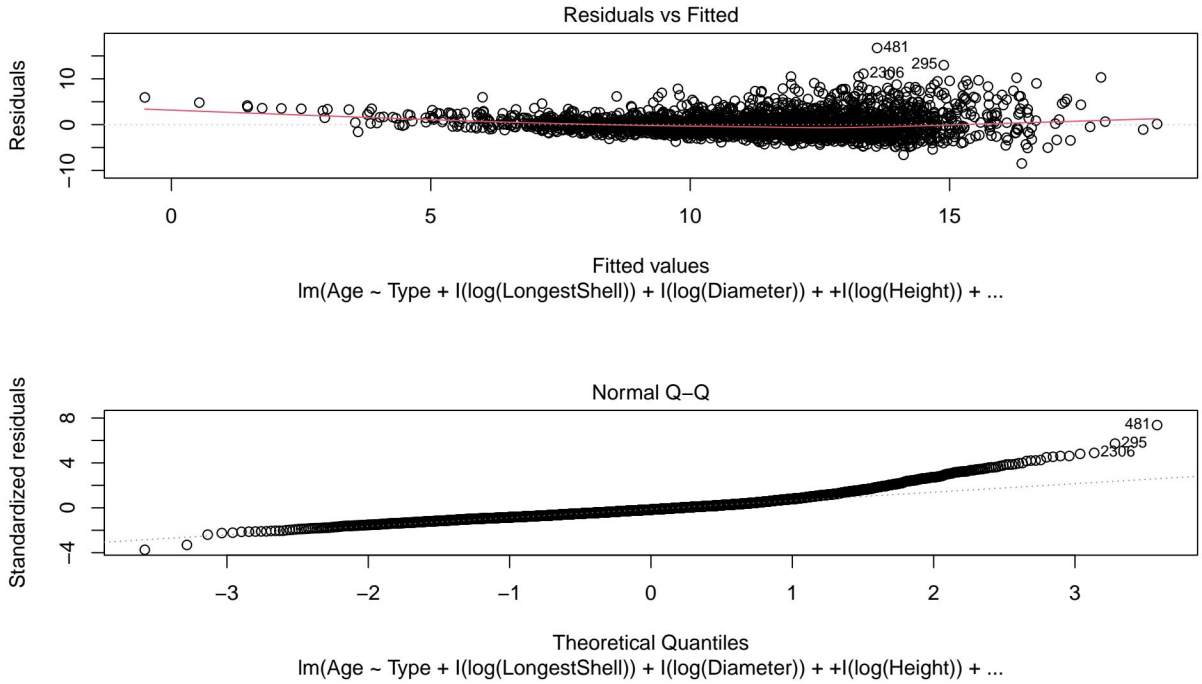
Diagnostics and model selection

The followings are the diagnostic plots for the linear model that used all of the variables without any transformation



From the summary of a linear model above, it seems all the variables, except TypeM and LongestShell, are statistically significant variables as their p-values are less than 0.05.

The followings are the diagnostic plots for the linear model that had all of the variables(except Age) have log transformations applied.



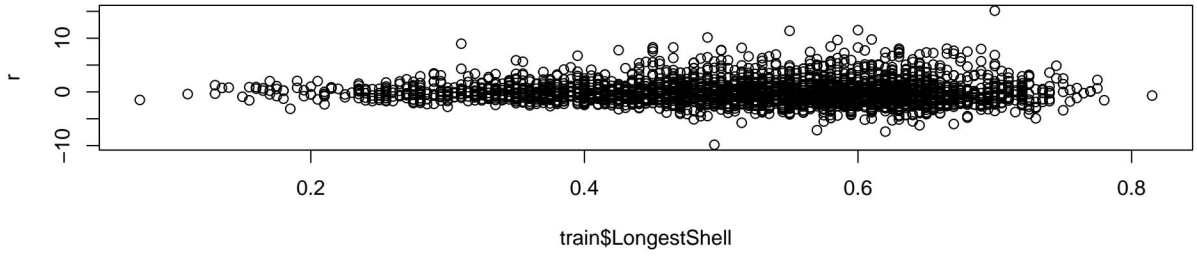
From the summary of a linear model with log transformation, it seems all the variables, except TypeM, log(LongestShell),log(Diameter),log(VisceraWeight), are statistically significant variables as their p-values are less than 0.05.

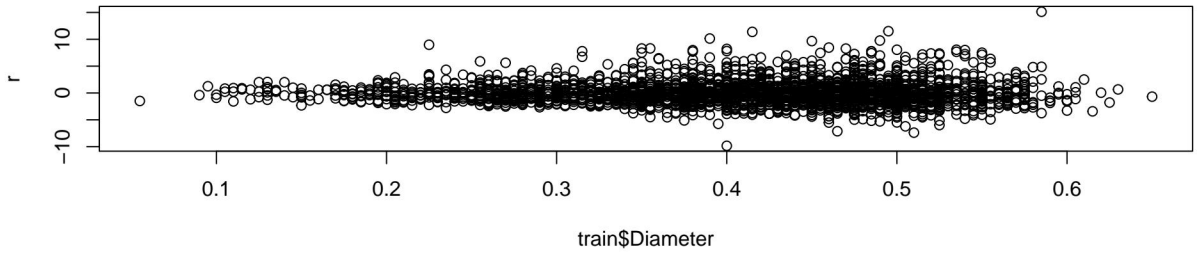
Also, from the diagnostics of the two models previously created, it seems they both violate the homoscedasticty as the residuals deviate from the dotted line by greatly different amounts. Also, the distribution of the data used to train the model does not seem normal, as the residuals in the front and later theoretical quantiles deviate a lot from the line. While both models have 3 outliers, it seems that points 295 and 481 are the two common outlier points.

Moreover, interestingly, as shown below, the model without any log transformation had higher adjusted R^2 value(0.5425) than the other's(0.5127), indicating the model without any log transformation fits the observed data better than the model with log transformation.

In addition, the model without any log transformation had lower prediction error (RMSE of 2.245197) than the other model's prediction error(RMSE of 2.302447). Also, we've tested to confirm that removing statistically insignificant variables for both models increased their prediction errors. Therefore, between the two models, we conclude that normal linear regression without variable transformation is better in terms of prediction.

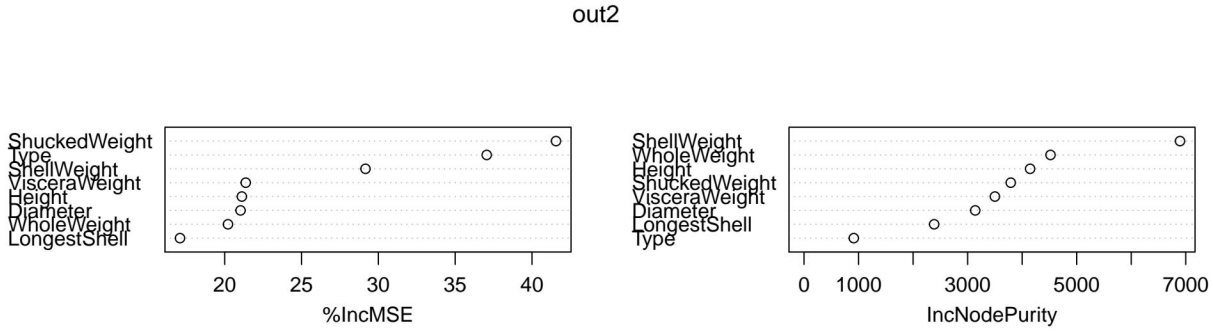
The followings are the diagnostic plots for the generalized additive model that used all of the variables.



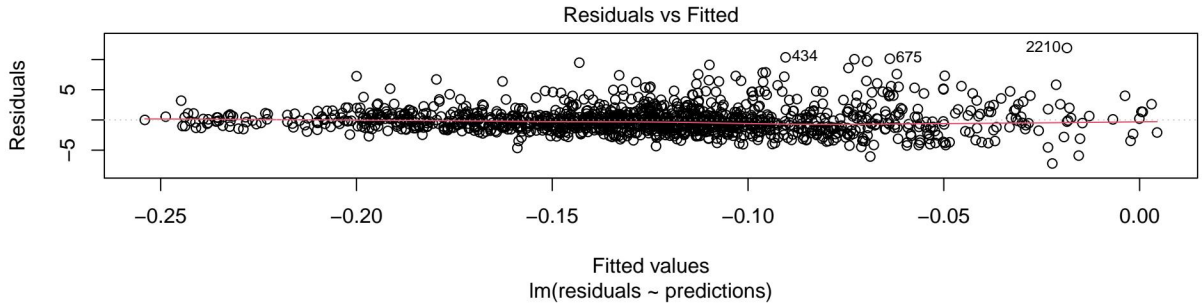


From the residual plots of two randomly chosen variables,it seems the assumption of homoscedasticty is violated as the residuals deviate from the dotted line by greatly different amounts. From the summary, it seems smoothed versions of height,wholeweight,shuckedweight,visceraweight,shellweight are most important for this model.The additive model explains 58.1% of deviance in training set and had the prediction error(RMSE of 2.145853), which was lower than those of two linear models.

Finally, the random forest.



%IncMSE indicates the increase in mse of predictions as a result of certain variable being permuted(values randomly shuffled); the higher number, more important the variable is. Thus, from the visualisation above, it seems there are 3 important variables(ShellWeight, Type and ShuckedWeight). Secondly, IncNodePurity is the total decrease in "node impurities" from splitting on the variable, averaged over all trees. Node impurity is just residual sums of squares.It is also worth nothing that from the IncNodePurity, we know ShellWeight split reduces the node impurity the most and Type split reduces the node impurity the least. Also, Random forest model had the prediction error(RMSE of 2.144028), which is the lowest error among all the models we have explored so far.



It is also worth noting that from the residual diagnosis of the random forest regression above,it seems the assumption of homoscedasticty is violated as the residuals deviate from the dotted line by greatly different amounts.

Final Models

Before deciding on a final model, I will first need to take the strengths and weaknesses of each machine learning model into account.

In general, linear regression models' strengths are that they are simple to understand can be regularized to avoid overfitting; however, they perform poorly when there are non-linear relationships. On the other hand, additive model is less general than the model but it is much more flexible than the linear model. It is often a good compromise between the linear model and a completely nonparametric model. However, model will lose predictability when the smoothed variables' values are outside of the range of training dataset. Lastly, with random forest, they are generally good at handling unbalanced and missing data. However, their weaknesses are that when used for regression they cannot predict beyond the range in the training data, and that they may over-fit data sets that are particularly noisy.

While considering strengths and weaknesses of each model, I chose random forest to be my final choice of the model for age prediction task because it had the lowest prediction error. Also, it's worth nothing that although additive model does not meet homoscedasticity assumption well, other two linear models seemed to violate homoscedasticity assumption even more.

While the 3 common important variables across all models were ShellWeight, Type and ShuckedWeight, the smoothed versions of height,wholeweight,shuckedweight,visceraweight,shellweight are most important for this additive model.

However, before using this model,there are several limitations that we should be aware of. This model is created using the data, which has 4176 rows. In the grand scheme, the dataset is not necessarily that large; I would be cautious when applying a model trained only on 2922 rows(of training set) to real application. Secondly, when used for regression, random forest models cannot predict beyond the range in the training data, so this is something we need to remember when using it to predict the age of abalones; we would want our training set to cover all possible ages of abalones. Also, this model seems to violate the assumption of homoscedasticity, one of the common assumptions that need to be satisfied if regression model is to be of any good. Lastly, random forests may over-fit data sets that are particularly noisy, so it is essential that we double-check that the dataset we used to train our model is reliable.T herefore, I would recommend still being cautious when applying the following model in real-life scenarios.

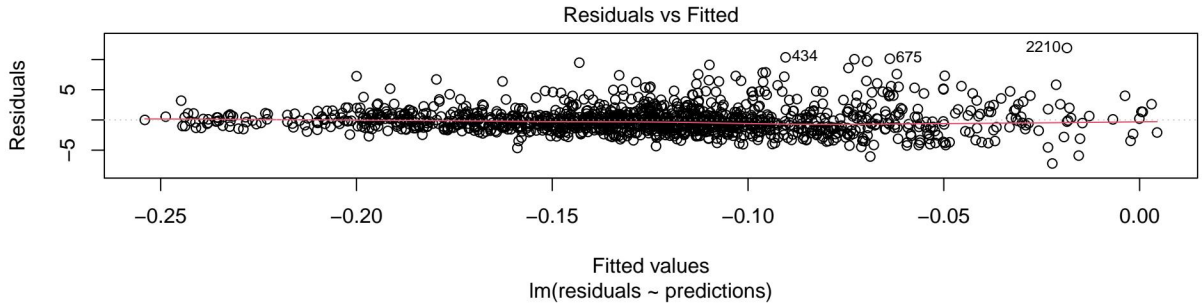
Code Appendix

```
## ----setup, include=FALSE-----------------------------------------------------
knitr::opts_chunk$set(fig.width=10, fig.height=3)
options(knitr.duplicate.label = "allow")
library(RColorBrewer)
library(mgcv) #Similar to example in lecture 14, age was not smoothed.
library(randomForest)
```

4 Diagnostics and Model Selection 17 / 20

- **0 pts** Correct
- **2 pts** labeling of plots makes narrative unclear
- **2 pts** No QQ plots
- **2 pts** Residual analysis is insufficient
- **10 pts** Diagnostic modeling only described at surface level; no in-depth narrative/graph/code present
- **0.5 pts** model plots are cut off between pages or cannot read formatting
- **0.1 pts** Need to label sections properly

✓ - **3 pts** **insufficient analysis of model error and justification of model selection**



It is also worth noting that from the residual diagnosis of the random forest regression above,it seems the assumption of homoscedasticty is violated as the residuals deviate from the dotted line by greatly different amounts.

Final Models

Before deciding on a final model, I will first need to take the strengths and weaknesses of each machine learning model into account.

In general, linear regression models' strengths are that they are simple to understand can be regularized to avoid overfitting; however, they perform poorly when there are non-linear relationships. On the other hand, additive model is less general than the model but it is much more flexible than the linear model. It is often a good compromise between the linear model and a completely nonparametric model. However, model will lose predictability when the smoothed variables' values are outside of the range of training dataset. Lastly, with random forest, they are generally good at handling unbalanced and missing data. However, their weaknesses are that when used for regression they cannot predict beyond the range in the training data, and that they may over-fit data sets that are particularly noisy.

While considering strengths and weaknesses of each model, I chose random forest to be my final choice of the model for age prediction task because it had the lowest prediction error. Also, it's worth nothing that although additive model does not meet homoscedasticity assumption well, other two linear models seemed to violate homoscedasticity assumption even more.

While the 3 common important variables across all models were ShellWeight, Type and ShuckedWeight, the smoothed versions of height,wholeweight,shuckedweight,visceraweight,shellweight are most important for this additive model.

However, before using this model,there are several limitations that we should be aware of. This model is created using the data, which has 4176 rows. In the grand scheme, the dataset is not necessarily that large; I would be cautious when applying a model trained only on 2922 rows(of training set) to real application. Secondly, when used for regression, random forest models cannot predict beyond the range in the training data, so this is something we need to remember when using it to predict the age of abalones; we would want our training set to cover all possible ages of abalones. Also, this model seems to violate the assumption of homoscedasticity, one of the common assumptions that need to be satisfied if regression model is to be of any good. Lastly, random forests may over-fit data sets that are particularly noisy, so it is essential that we double-check that the dataset we used to train our model is reliable.T herefore, I would recommend still being cautious when applying the following model in real-life scenarios.

Code Appendix

```
## ----setup, include=FALSE-----------------------------------------------------
knitr::opts_chunk$set(fig.width=10, fig.height=3)
options(knitr.duplicate.label = "allow")
library(RColorBrewer)
library(mgcv) #Similar to example in lecture 14, age was not smoothed.
library(randomForest)
```

```
library(tidyverse)


## ----echo=FALSE-----------------------------------------------------------
df<-read.csv("/Users/joonghochoi/Desktop/abalone.csv")
df<-df[ , -which(names(df) %in% c("ID"))]
df$Type=as.factor(df$Type)
head(df,3)


## ----echo=FALSE-----------------------------------------------------------
tmp<-df[ , -which(names(df) %in% c("Type"))]
res<-cor(tmp)
heatmap(x = res, col = colorRampPalette(brewer.pal(8,"Blues"))(3), symm = TRUE,main="Heatmap that shows
legend(x = "bottomright", legend = c("low", "medium", "high"),
       cex = 0.8,fill = colorRampPalette(brewer.pal(8, "Blues"))(3))


## -------------------------------------------------------------------------
pairs(df[,2:9],pch=19,main="Pairwise plot of quantitative variables in data for EDA")


## ----echo=FALSE-----------------------------------------------------------
meow<-df
meow$Height=log(meow$Height)
#a<-c(which(is.na(meow$Height)))
#b<-c(which(is.infinite(meow$Height)))
omit<-c(which(is.na(meow$Height) | is.infinite(meow$Height)))
df1<-df[-omit,]
#train<-train[-0,]
set.seed(42)
dt = sort(sample(nrow(df1), nrow(df1)*.7))
train<-df1[dt,]
test<-df1[-dt,]


## ----echo=FALSE-----------------------------------------------------------
out<-lm(Age~.,data=train)
summary(out)
out_log<-lm(Age~Type+I(log(LongestShell))+I(log(Diameter))++I(log(Height))+I(log(WholeWeight)+I(log(Shu
summary(out_log)


## ----echo=FALSE-----------------------------------------------------------
out1<-gam(Age~Type+s(LongestShell)+s(Diameter)+s(Height)+s(WholeWeight)+s(ShuckedWeight)+s(VisceraWeigh
summary(out1)

## ----echo=FALSE-----------------------------------------------------------
out2 = randomForest(Age ~ .,importance=TRUE,data=train)


## ----echo=FALSE-----------------------------------------------------------
plot(out,which=1)
```

```
plot(out,which=2)



## ----echo=FALSE----------------------------------------------------------
plot(out_log,which=1)
plot(out_log,which=2)


## ----echo=FALSE----------------------------------------------------------
predictions<-out%>%predict(test)
error <- test$Age - predictions
RMSE<-sqrt(mean(error^2))
#RMSE
predictions<-out_log%>%predict(test)
error <- test$Age - predictions
RMSE<-sqrt(mean(error^2))
#RMSE


## ----echo=FALSE----------------------------------------------------------
r=residuals(out1)
plot(train$LongestShell,r)
plot(train$Diameter,r)


## ----echo=FALSE----------------------------------------------------------
predictions<-out1%>%predict(test)
error <- test$Age - predictions
RMSE<-sqrt(mean(error^2))
#MSE


## ----echo=FALSE----------------------------------------------------------
varImpPlot(out2)


## ----echo=FALSE----------------------------------------------------------
predictions<-out2%>%predict(test)
error <- test$Age - predictions
RMSE<-sqrt(mean(error^2))
#RMSE


## ----echo=FALSE----------------------------------------------------------
predictions<-out2%>%predict(test)
residuals <- test$Age-predictions
plot(lm(residuals ~ predictions), which=1)


## ----code = readLines(knitr::purl(knitr::current_input(), documentation = 1)), echo = T, eval = F----
## NA
```

5 Final Model 20 / 20

✓ **- 0 pts** **Correct**

- 5 pts More than seven pages.

- 2 pts Only used tree, not random forest.

- 2 pts Prediction error of the linear model seems unrealistically good.

- 2 pts Forest makes predictions by voting from all the trees, not taking the result from the best tree. This is a reason why random forest doesn't overfit much.

- 1 pts GAM has too few covariates, and the performance of GAM should be improved.

- 1 pts MSEs of the three models seem to be too close.

- 2 pts To fairly compare all three models, you should calculate prediction error or R square with the same data. An outlier tends to be hard to predict and increases the mean square error a lot. It's ok you keep outlier out in the training state, but it's not fair for only model 1b to "avoid" those difficult data points when you compare the three models. This explains why your model 1b has way smaller MSE ---- model 1b took an easier exam.

- 4 pts Didn't explain enough about how to choose the final models from the three models.