BIOS 663 Final Project Report

Body Fat Data Set

Hantong Hu, Jie He, Yingnan Wu, Yirun Li

Abstract

In this report, we showed and interpreted the results of three selected statistical methods to model the relationship between percent of body fat (the response variable) and multiple body measurements (predictors) based on the body fat dataset. Specifically, we used data-splitting and employed three models to the training samples: (1) the linear regression model, (2) the LASSO and Ridge model, and (3) the general additive model (GAM). The selection criterion was **mean squared errors (MSE)** of the uniformly scaled response variable (**Siri_Y**: percent of body fat calculated by the Siri Equation based on corrected body densities) and **R-squared** values. The parametric estimates were reported to display the directions of different predictors' effects on the response variable, and the p-values were reported to single out the significance levels of the predictors. In addition, we also reported model specific statistics for thoughtfully evaluating the different models.

1. Introduction

1.1 Data Source

The data set was obtained from CMU StatLib (http://lib.stat.cmu.edu/datasets/bodyfat). The original dataset contains 252 observations aged 21- 81 and 15 continuous numeric variables with no missing value including body density determined from underwater weighing, biometrics such as age, height, and weight, and body circumference measurements of neck, chest, abdomen, hip, thigh, knee, ankle, biceps, forearm and wrist.

1.2 Objective

Some public health books suggested that their readers tend to use the percentage of body fat to assess their health conditions[1,2,3]. Therefore, the estimation of the percentage of body fat became of particular interest[4]. To accurately measure the body fat, however, is typically inconvenient and costly as it involves measurement under water. For this project, we aimed to find a simpler method of approximation using body measurements. To achieve this, different methodologies were used to fit the percentage of body fat data and a best-fit model was obtained.

2. Data Processing

2.1 Exploratory Data Analysis

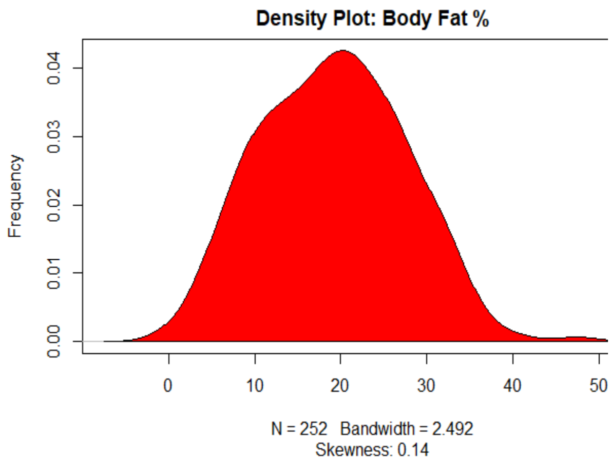


Fig 1. Density plot of percent body fat

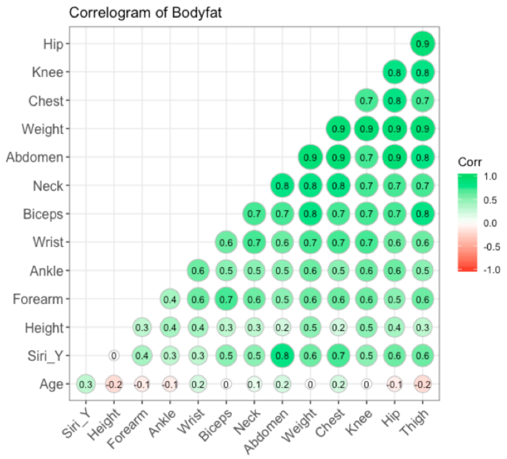


Fig 2. Correlogram of variables

Before analyzing the data, we visualized the data to identify potential outliers and to improve modeling efficiency. From the density plot, we observed that the distribution of the response variable, the percentage of the body fat, is approximately normal, indicating that the multiple linear regression is applicable. In addition, the correlation plot on the right demonstrates collinearity among covariates.

2.2 Data Cleaning

From ***Fig 3***., we observed several special values. Some of these values were apparent errors and were manually corrected according to their relationship with other variables. For example, in case 42, the individual had a weight of 205 pounds and a height of 29.5 inches, which was abnormal based on empirical knowledge. By inspecting the data set, the lean body weight was recorded as 104.1 pounds, which agrees with the result by Brozek's equation (Lean body weight= (1-proportion body fat) * weight). Therefore, the height value for this observation was erroneous. Based on the value of the adiposity index of 29.9 kg/meters^2, the corrected value of height was calculated to be 69.5 inches

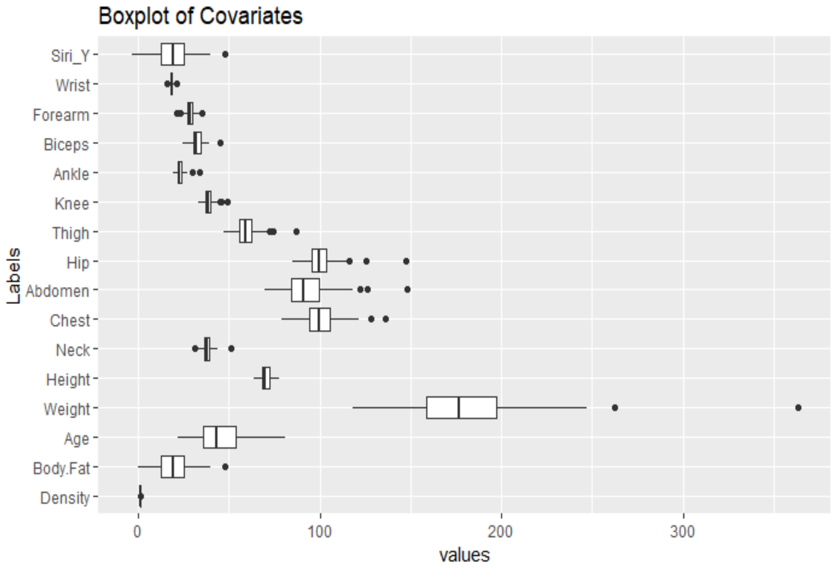


Fig 3. Boxplot of all covariates (from top- down: Siri_Y, Wrist, Forearm, Biceps, Ankle, Knee, Thigh, Hip, Abdomen, Chest, Neck, Height, Weight, Age, Percent body fat, and Density)

Similarly, three other erroneous cases were identified, as summarized below in ***Table 1***. Based on the corrected body densities, **Siri_Y** (percent of body fat) was calculated according to the Siri Equation to be the response variable.

Case	**48**	**76**	**96**
Variable	Body Density	Body Density	Body Density
Corrected Value	1.0865	1.0566	1.0591

Table 1. Summary of erroneous cases with the corrected variable values

3. Analysis Methods

As for data analysis, the original data set was split into a training set and a test set and MSE on the test set was employed as the validation criterion to choose the best fit model. Since the objective was to predict the **Siri_Y**, a quantitative variable, three supervised learning methods - multiple linear regression, regularization, and general additive modeling - were applied.

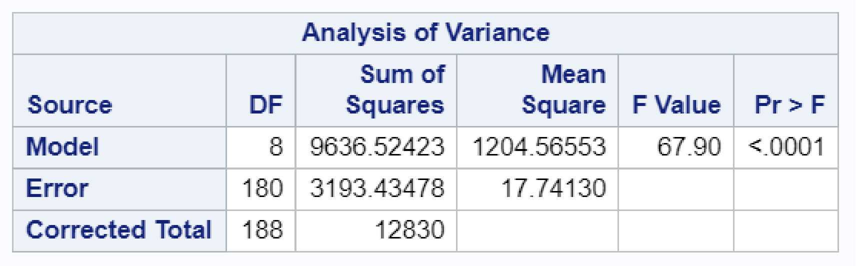
3.1 Multiple Linear Regression and Stepwise Selection

Multiple linear regression was first applied to fit the training set. From the exploratory data analysis, it was inferred that there might exist collinearity between variables. The stepwise selection was therefore used to narrow down the variables to make a relatively parsimonious model. A total of 7 predictors were selected from the list of 13 variables, which are weight, age, circumferences of wrist, thigh, neck, forearm, and abdomen. All of the variables left in the model were significant at the 0.1500 level. The analysis of the training set resulted in an adjusted R^2 value of 73.30%. By empirical knowledge and the literature review, we suspected that there existed interaction between weight and abdomen. Therefore, an interaction term was added to the linear regression model to account for the interaction between weight and abdomen circumference:

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \cdots + \beta_7 x_7 + \beta_8 x_2 x_7 + \varepsilon$, where the x denotes the circumference of wrist, weight, circumferences of thigh, neck, forearm, age, the circumference of abdomen, respectively. The null hypothesis is that the selection of predictors does not contribute to explaining any of the variability in the response variable, i.e.,

$$Ho: \beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 .$$

After fitting the equation to the data set, we obtained the test statistics of 67.90 with a significant p-value of <0.0001, meaning there is sufficient evidence against the null hypothesis to reject it. The interpretation is that the entire group of predictors has a significant relationship with the response variable.



Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	9636.52423	1204.56553	67.90	<.0001
Error	180	3193.43478	17.74130		
Corrected Total	188	12830			

Table 2. Test statistics of the hypothesis test

Also, the estimated linear regression equation was listed as follows,

$$y = -51.5 - 1.58 * \text{Wrist} + 0.025 * \text{Weight} + 0.31 * \text{Thigh} - 0.35 * \text{Neck} + 0.36 * \text{Forearm}$$
$$+0.056 * \text{Age} + \ 1.14 * \text{Abdomen} - 0.0012 * (\text{Weight} * \text{Abdomen})$$

Based on the equation, it can be inferred that an older male with higher weight, thicker thigh, forearm, and the abdomen would be more likely to have a higher percent body fat. The aforementioned estimated linear regression equation was then applied to the test set and the adjusted R^2 was 75.0% with the MSE of 19.30.

3.2 Ridge and LASSO Regression

While the stepwise selection used the subsets of the predictors to control the model complexity, an alternative method, the Ridge and Lasso regression, controls the model complexity by shrinking the regression coefficients.

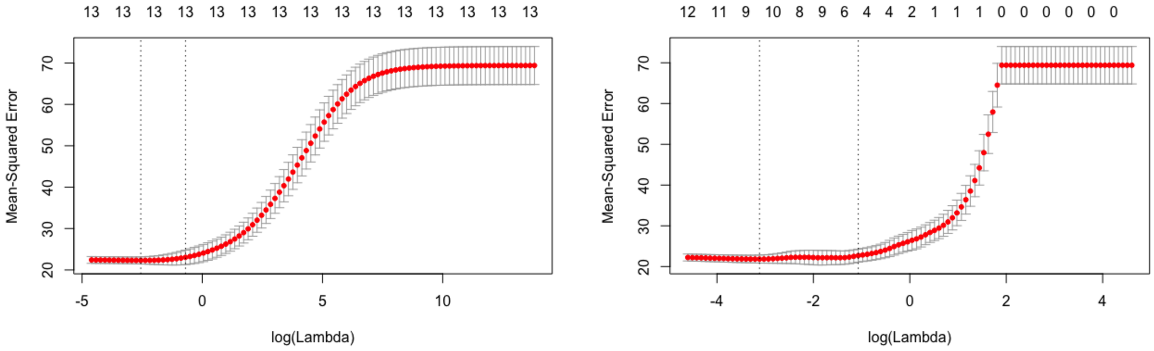


Fig 4. MSE plots of different values of log(λ) for Ridge [left] and Lasso [right].

For each regularization method, we used 5-fold cross-validation on the training set to select the optimal value of the tuning parameter λ among a grid of values and obtained the prediction MSE estimates using the test set data. ***Fig 4.*** demonstrates the values of MSE versus the different values of log(λ) for Ridge and Lasso respectively. It can be seen from ***Fig 5.*** that the coefficients in the Lasso model shrink to zero as λ increases.

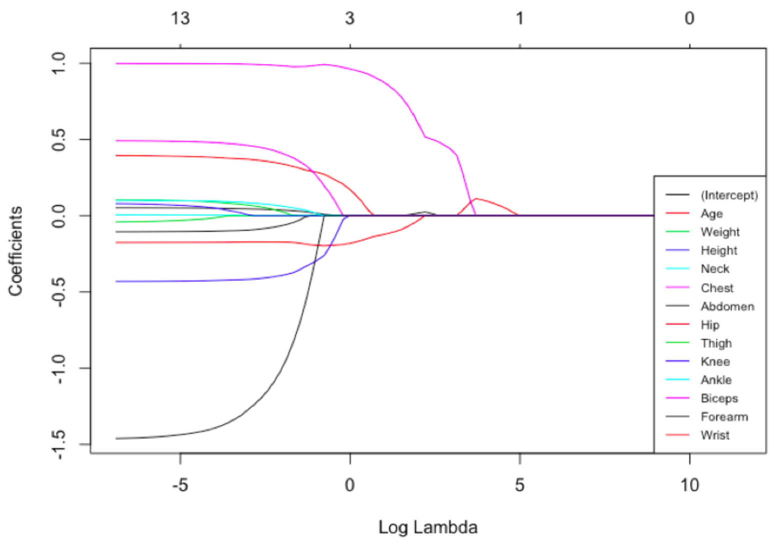


Fig 5. MSE plots of different values of log(λ) for Lasso

We chose λ_Ridge=0.07742637 and λ_Lasso=0.0443062. In particular, since the Lasso regression will force some coefficients to be zero, the Lasso model with the optimal tuning parameter consists of 11 predictors (Age, Weight, Height, Neck, Abdomen, Hip, Thigh, Biceps, Forearm, and Wrist) including the intercept. Test MSE for Ridge is 19.74 and 19.95 for Lasso. The resulting models are shown below:

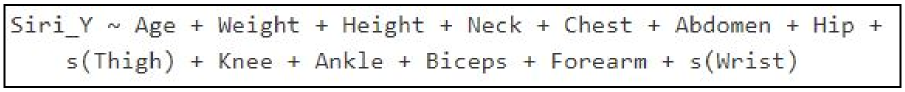
$$y_{Lasso} = -26.96 + 0.0496 * Age - 0.1285 * Weight - 0.0092 * Height + 0.92 * Abdomen \\ - 0.06 * Hip + 0.28 * Thigh + 0.03 * Biceps + 0.4495 * Forearm - 1.4707 * Wrist$$

$$y_{Ridge} = -18.76 + 0.0668 * Age - 0.1039 * Weight - 0.0580 * Height - 0.4827 * Neck \\ + 0.0168 * Chest + 0.8691 * Abdomen - 0.1182 * Hip + 0.36 * Thigh - 0.0432 * Knee \\ + 0.0128 * Ankle + 0.0454 * Biceps + 0.4696 * Forearm - 1.6056 * Wrist$$

3.3 General Additive Model

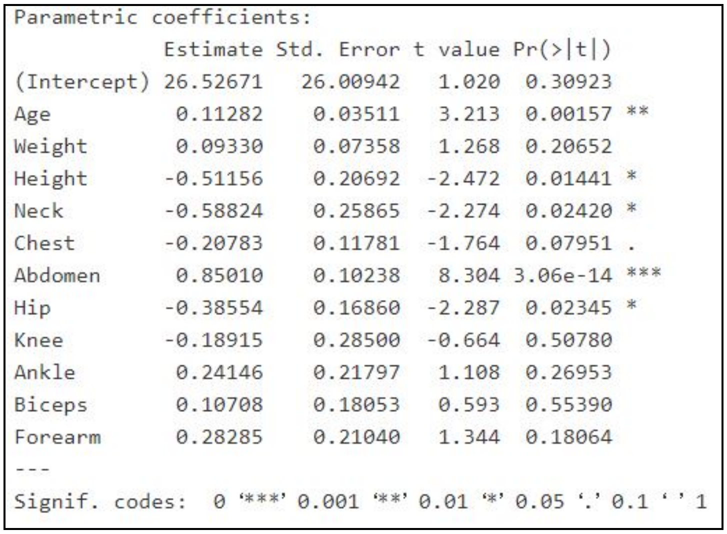
Beside the aforementioned linear regression methods, we also tried the general additive model (GAM) where relationships between the response variable and individual predictors could be non-linear.

3.3.1 Fitted Model:



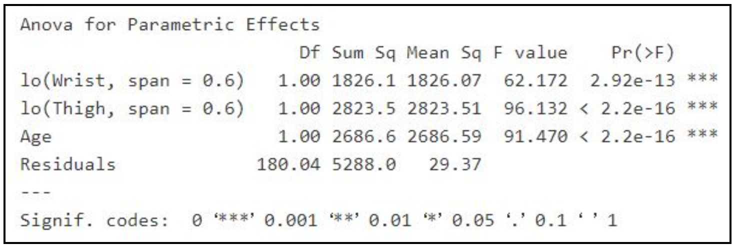
```
Siri_Y ~ Age + Weight + Height + Neck + Chest + Abdomen + Hip +
    s(Thigh) + Knee + Ankle + Biceps + Forearm + s(Wrist)
```

3.3.2 Parametric Estimates:



```
Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 26.52671   26.00942   1.020  0.30923
Age          0.11282    0.03511   3.213  0.00157 **
Weight       0.09330    0.07358   1.268  0.20652
Height      -0.51156    0.20692  -2.472  0.01441 *
Neck        -0.58824    0.25865  -2.274  0.02420 *
Chest       -0.20783    0.11781  -1.764  0.07951 .
Abdomen      0.85010    0.10238   8.304 3.06e-14 ***
Hip         -0.38554    0.16860  -2.287  0.02345 *
Knee        -0.18915    0.28500  -0.664  0.50780
Ankle        0.24146    0.21797   1.108  0.26953
Biceps       0.10708    0.18053   0.593  0.55390
Forearm      0.28285    0.21040   1.344  0.18064
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

3.3.3 ANOVA of Parametric Effects:



```
Anova for Parametric Effects
                        Df Sum Sq Mean Sq F value    Pr(>F)
lo(Wrist, span = 0.6)   1.00 1826.1 1826.07  62.172  2.92e-13 ***
lo(Thigh, span = 0.6)   1.00 2823.5 2823.51  96.132 < 2.2e-16 ***
Age                     1.00 2686.6 2686.59  91.470 < 2.2e-16 ***
Residuals             180.04 5288.0   29.37
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

4. Limitation

This analysis has some limitations. As for the data preparation stage, the outliers and measurement errors of the data set were identified by empirical knowledge, making the data cleaning criterion restricted to this particular data set. As for the model fitting stage, the multiple linear regression method was not exhaustive, so there may be a lack of consideration on interaction terms and polynomial terms in the model. Second, the GAM did not employ multiple-folded cross-validation so that its parametric estimates and test MSE can be unstable otherwise. Further, we would consider improving the power of the multiple linear regression model by adjusting VIFs in the future.

5. Results and Conclusion

Method	**Linear Model**	**Stepwise Selection**	**With Interaction**	**LASSO**	**Ridge Regression**	**GAM**
Test MSE	20.52	20.89	19.30	19.96	19.74	22.12
Test R^2	0.73	0.73	0.75	0.74	0.74	NA*

According to our selection criterion, the multiple linear regression model with an interaction term was chosen as our final model for estimating the percent of body fat by measurements because the model has the lowest test MSE **(19.30)** and the highest R squared value **(0.75).** In conclusion, one's percent of body fat would be additively associated with an individual’s weight, thigh, forearm, age, and abdomen measurements in the positive direction, while related in the negative direction to an individual’s wrist and neck measurements as well as the joint effect of weight and abdomen.

Bibliography

1. Brozek, J., Grande, F., Anderson, J., and Keys, A. (1963), "Densitometric Analysis of Body Composition: Revision of Some Quantitative Assumptions," Annals of the New York Academy of Sciences, 110, 113-140.

Appendix

[Multiple Linear Regression]

```
*****************************************
-{Final Project- BIOS 663}-
Group 9- 730052136
*****************************************;
ods rtf file='C:\Users\Yirun Li\Documents\Spring 2019\BIOS 663\project\project_output.rtf';

proc import datafile="C:\Users\Yirun Li\Documents\Spring 2019\BIOS 663\project\bodyfat_train.csv"
    out=bodyfat
    dbms=csv
    replace;
            datarow=2;
    getnames=yes;
run;

proc import datafile="C:\Users\Yirun Li\Documents\Spring 2019\BIOS 663\project\bodyfat_test.csv"
    out=test
    dbms=csv
    replace;
            datarow=2;
    getnames=yes;
run;

/*Proc contents data=bodyfat;*/
/*run;*/

Data bodyfat;
set bodyfat;
label siri_y="Body Fat Values based on Siri's Equation";
interaction=weight*abdomen;
interaction2=weight*neck;
run;

Data test;
set test;
label siri_y="Body Fat Values based on Siri's Equation";
interaction=weight*abdomen;
interaction2=weight*neck;
run;

/*run all possible models*/
title1 'all possible models';
Proc reg data=bodyfat outest=est;
model Siri_Y=wrist interaction interaction2 weight thigh neck knee hip height forearm chest biceps ankle age
abdomen/
```

```
selection=adjrsq sse aic;
output out=out p=p r=r;
run;
/**/
/*/*forward selection*/*/
/*title1 'forward selection';*/
/*Proc reg data=bodyfat outest=est1;*/
/*model Siri_Y=wrist weight thigh neck knee hip height forearm chest biceps ankle age abdomen/*/
/*slentry=0.15 selection=forward ss2 sse aic;*/
/*output out=out1 p=p r=r;*/
/*run;*/
/**/
/*/*backward selection*/*/
/*title1 'backward selection';*/
/*Proc reg data=bodyfat outest=est2;*/
/*model Siri_Y=wrist weight thigh neck knee hip height forearm chest biceps ankle age abdomen/*/
/*slstay=0.15 selection=backward ss2 sse aic;*/
/*output out=out1 p=p r=r;*/
/*run;*/

/*stepwise selection*/
title1 'stepwise selection';
Proc reg data=bodyfat outest=est3;
model Siri_Y=wrist weight thigh neck knee hip height forearm chest biceps ankle age abdomen interaction
interaction2/
slstay=0.15 slentry=0.15 selection=stepwise ss2 sse aic;
output out=out3 p=p r=r;
run;

/*Collinearity*/
title 'Check collinearity';
data bodyfat;
set bodyfat;
int=1;
run;

/*scaled sscp to check collinearity*/
proc princomp data=bodyfat noint;
var int Wrist Weight Thigh Neck Forearm Age Abdomen;
run;

proc princomp data=bodyfat;
var Wrist Weight Thigh Neck Forearm Age Abdomen;
run;

/*Obtain model equation*/
title 'Obtain model equation';
proc reg data=bodyfat;
```

```
model siri_y=wrist weight thigh neck forearm age abdomen interaction;
run;


/*two interaction terms*/
Data test3;
set test;
pred_y=-44.39555-1.57157*Wrist-0.06989*Weight+0.30988*Thigh-0.77549*Neck+Forearm*0.37016+Age*0.0552
4+Abdomen*1.24186-0.00167*interaction+interaction2*0.00234;
diff=pred_y-siri_y;
diffsq=diff*diff;
run;

/*no interaction*/
Data test1;
set test;
pred_y=-36.46942-1.39198*Wrist-0.16829*Weight+0.34139*Thigh-0.41358*Neck+Forearm*0.55703+Age*0.0519
6+Abdomen*0.95458;
diff=pred_y-siri_y;
diffsq=diff*diff;
run;

/*abdomenweight*/
Data test2;
set test;
pred_y=-51.52449-1.57615*Wrist-0.02503*Weight+0.31160*Thigh-0.35229*Neck+Forearm*0.35961+Age*0.0562
0+Abdomen*1.14166-0.00116*interaction;
diff=pred_y-siri_y;
diffsq=diff*diff;
run;

Proc means data=test1 sum;
var diffsq;
output out=testresult1 sum=/autoname;
run;

Proc means data=test2 sum;
var diffsq;
output out=testresult2 sum=/autoname;
run;

Proc means data=test3 sum;
var diffsq;
output out=testresult3 sum=/autoname;
run;

Data testresult1;
set testresult1;
```

```
mse=diffsq_sum/63;
run;

Data testresult2;
set testresult2;
mse=diffsq_sum/63;
run;

Data testresult3;
set testresult3;
mse=diffsq_sum/63;
run;

title 'MSE of model without interaction';
proc print data=testresult1;
var mse;
run;

title 'MSE of model with interaction';
proc print data=testresult2;
var mse;
run;

title 'MSE of model with two interaction';
proc print data=testresult3;
var mse;
run;

************************************************;
Proc means data=test var;
var siri_y;
run;

data testresult1;
set testresult1;
mv=(62/63)*77.6379731;
testrsq=1-mse/mv;
run;

data testresult2;
set testresult2;
mv=(62/63)*77.6379731;
testrsq=1-mse/mv;
run;

data testresult3;
set testresult3;
mv=(62/63)*77.6379731;
```

```
testrsq=1-mse/mv;
run;

title 'Test R-SQ of model without interaction';
proc print data=testresult1;
var testrsq;
run;

title 'Test R-SQ of model with interaction';
proc print data=testresult2;
var testrsq;
run;

title 'Test R-SQ of model with two interaction';
proc print data=testresult3;
var testrsq;
run;

***************************************************************;
/*Graphics*/
Proc print data=bodyfat (obs=10);
var wrist weight thigh neck forearm age abdomen;
run;

Proc univariate data=bodyfat;
var siri_y;
histogram/ normal;
run;

Proc means data=bodyfat min max mean median std;
var siri_y;
run;

proc univariate data=bodyfat;
qqplot siri_y / normal(mu=19.5686464 sigma=8.26) square ctext=blue;
run;
***************************************************************;

proc glm data=bodyfat;
model siri_y=wrist weight thigh neck forearm age abdomen interaction/solution;
output out=one predicted=y_hat rstudent=r_i;
run;

title1 "Gaussian Distribution for Studentized Residuals";
proc univariate plot normal data=one;
var r_i;
label r_i="Studentized Residuals";
run;
```

```
/*quit;*/

ods rtf close;
```

[Lasso and Ridge]

~~~~
---
title: "Group 9 Final Project"
date: "4/20/2019"
output: pdf_document
---



```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
if(!require(ISLR)) { install.packages("ISLR", repos = "http://cran.us.r-project.org"); library(ISLR) }
if(!require(leaps)) { install.packages("leaps", repos = "http://cran.us.r-project.org"); library(leaps) }
if(!require(glmnet)) { install.packages("glmnet", repos = "http://cran.us.r-project.org"); library(glmnet) }
if(!require(pls)) { install.packages("pls", repos = "http://cran.us.r-project.org"); library(pls) }
```



```{r}
# import data and rename the columns
data_corrected <- read.csv("data_corrected.csv", header = TRUE)
bodyfat <- data_corrected[ , 3:16]
head(bodyfat, 10)
```




##### Correlation
```{r warning=FALSE, message=FALSE}
# ++++++++++++++++++++++++++++
# flattenCorrMatrix
# ++++++++++++++++++++++++++++
# cormat : matrix of the correlation coefficients
# pmat : matrix of the correlation p-values
flattenCorrMatrix <- function(cormat, pmat) {
  ut <- upper.tri(cormat)
  data.frame(
            row = rownames(cormat)[row(cormat)[ut]],
            column = rownames(cormat)[col(cormat)[ut]],
            cor  =(cormat)[ut],
            p = pmat[ut]
            )
}
~~~~

```
x <- bodyfat[ ,1:13]
y <- bodyfat[ , 14]
head(x)
head(y)

# The default method of cor() is pearson correlation coefficient
# which measures the linear dependence between two variables.
# ----------------------------------------------------------------------
# The function rcorr() [in Hmisc package] can be used to
# compute the significance levels for pearson and spearman correlations.
# It returns both the correlation coefficients and
# the p-value of the correlation for
# all possible pairs of columns in the data table.

#library(Hmisc)
#res2<-rcorr(as.matrix(bodyfat), type = "pearson")
#flattenCorrMatrix(res2$r, res2$P)

# Use corrplot() function to draw a correlogram
library(corrplot)
res <- cor(bodyfat, method = "pearson")
corrplot(res, type = "upper", order = "hclust", tl.col = "black", tl.srt = 45)

# Use corrplot() function to draw a correlogram
# ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
# In the above plot:
# 1. The distribution of each variable is shown on the diagonal.
# 2. On the bottom of the diagonal:
#		the bivariate scatter plots with a fitted line are displayed
# 3. On the top of the diagonal:
#		the value of the correlation plus the significance level as stars
# 4. Each significance level is associated to a symbol:
#		p-values(0, 0.001, 0.01, 0.05, 0.1, 1) <=> symbols(“***”, “**”, “*”, “.”, " “)
# ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

library("PerformanceAnalytics")
bodyfat1 <- bodyfat[, c(1,2,3,4,5,14)]
bodyfat2 <- bodyfat[, c(6,7,8,9,10,14)]
bodyfat3 <- bodyfat[, c(11,12,13,14)]
bodyfat4 <- bodyfat[, c(13,2,8,4,12,1,6)]
chart.Correlation(bodyfat1, histogram=TRUE, pch=19)
chart.Correlation(bodyfat2, histogram=TRUE, pch=19)
chart.Correlation(bodyfat3, histogram=TRUE, pch=19)
chart.Correlation(bodyfat4, histogram=TRUE, pch=19)
```

##### Correlogram
```{r}
# devtools::install_github("kassambara/ggcorrplot")
library(ggplot2)
library(ggcorrplot)

# Correlation matrix
corr <- round(cor(bodyfat), 1)

# Plot
ggcorrplot(corr, hc.order = TRUE,
           type = "lower",
           lab = TRUE,
           lab_size = 3,
           method="circle",
           colors = c("tomato2", "white", "springgreen3"),
           title="Correlogram of Bodyfat",
           ggtheme=theme_bw)
```

##### Data Splitting
```{r include=FALSE}
set.seed(6639)
train <- sample(1:nrow(bodyfat),nrow(bodyfat)/4*3)
test <- (-train)
bodyfat.train <- bodyfat[train, ]
write.csv(bodyfat.train, "bodyfat_train.csv")
bodyfat.test <- bodyfat[test, ]
write.csv(bodyfat.test, "bodyfat_test.csv")
```

##### Best Subset Selection
```{r}
mod.best <- regsubsets(Siri_Y ~ ., data = bodyfat.train, nvmax = 13)
mod.best.sum <- summary(mod.best)
```

The best size of subsets identified by $C_p$, BIC and adjusted $R^2$ are `r which.min(mod.best.sum$cp)`, `r which.min(mod.best.sum$bic)` and `r which.max(mod.best.sum$adjr2)` respectively. The plots of $C_p$, BIC and adjusted $R^2$ with respect to the size of subsets are as follows.

```{r echo=FALSE}
opar <- par(mfrow = c(1,3), oma = c(0, 0, 1, 0))
plot(mod.best.sum$cp,
```

~~~~
        sub = "Cp",
        xlab = "Subset Size", ylab = "Cp",
        pch = 20, type = "l")
points(which.min(mod.best.sum$cp),
        min(mod.best.sum$cp),
        pch = 4, col = "red", lwd = 7)

plot(mod.best.sum$bic,
        sub = "BIC",
        xlab = "Subset Size", ylab = "BIC",
        pch = 20, type = "l")
points(which.min(mod.best.sum$bic),
        min(mod.best.sum$bic),
        pch = 4, col = "red", lwd = 7)

plot(mod.best.sum$adjr2,
        sub = "Adjusted R^2",
        xlab = "Subset Size", ylab = "Adjusted R^2",
        pch = 20, type = "l")
points(which.max(mod.best.sum$adjr2),
        max(mod.best.sum$adjr2),
        pch = 4, col = "red", lwd = 7)
title("Information Criteria ~ Subset Size", outer = TRUE)
par(opar)
```


The selected models are
```{r}
# by Cp
coefficients(mod.best, id = which.min(mod.best.sum$cp))

# by BIC
coefficients(mod.best, id = which.min(mod.best.sum$bic))

# by adjusted R^2
coefficients(mod.best, id = which.max(mod.best.sum$adjr2))
```




##### Linear Model Using Selected Best Subset

```{r}
mod.lm <- lm(Siri_Y ~ Age+Weight+Neck+Abdomen+Thigh+Forearm+Wrist, data = bodyfat.train)
mse.lm <- mean( ( bodyfat.test$Siri_Y - predict(mod.lm, bodyfat.test) )^2 )
mse.lm
~~~~

~~~
```
The prediction MSE estimate for linear model is `r formatC(mse.lm, format = "e", digits = 4)` when using the
selected best subset based on the Adjusted R-squared criteria.




##### Forward and Backward Stepwise Regression
Now we conduct forward and backward stepwise regressions.

```{r}
mod.fwd <- regsubsets(Siri_Y ~ ., data = bodyfat.train, nvmax = 13, method = "forward")
mod.bwd <- regsubsets(Siri_Y ~ ., data = bodyfat.train, nvmax = 13, method = "backward")
mod.fwd.sum <- summary(mod.fwd)
mod.bwd.sum <- summary(mod.bwd)
```
The selected models under three criteria are listed as follows.

|**Information Criteria**|**Forward Stepwise Regression** |**Backward Stepwise Regression**|
|:---------------------:|:-----------------------------:|:------------------------------:|
|$C_p$                  |`r which.min(mod.fwd.sum$cp)`   |`r which.min(mod.bwd.sum$cp)`   |
|BIC                    |`r which.min(mod.fwd.sum$bic)`  |`r which.min(mod.bwd.sum$bic)`  |
|Adjusted $R^2$                |`r which.max(mod.fwd.sum$adjr2)`|`r which.max(mod.bwd.sum$adjr2)`|

```{r echo=FALSE}
opar <- par(mfrow = c(2,3), oma = c(0, 0, 1, 0))
plot(mod.fwd.sum$cp,
        sub = "Forward Stepwise Regression",
        xlab = "Subset Size", ylab = "Cp",
        pch = 20, type = "l")
points(which.min(mod.fwd.sum$cp),
        min(mod.fwd.sum$cp),
        pch = 4, col = "red", lwd = 7)

plot(mod.fwd.sum$bic,
        sub = "Forward Stepwise Regression",
        xlab = "Subset Size", ylab = "BIC",
        pch = 20, type = "l")
points(which.min(mod.fwd.sum$bic),
        min(mod.fwd.sum$bic),
        pch = 4, col = "red", lwd = 7)

plot(mod.fwd.sum$adjr2,
        sub = "Forward Stepwise Regression",
        xlab = "Subset Size", ylab = "Adjusted R^2",
        pch = 20, type = "l")
points(which.max(mod.fwd.sum$adjr2),
        max(mod.fwd.sum$adjr2),
~~~

~~~
    pch = 4, col = "red", lwd = 7)

plot(mod.bwd.sum$cp,
    sub = "Backward Stepwise Regression",
    xlab = "Subset Size", ylab = "Cp",
    pch = 20, type = "l")
points(which.min(mod.bwd.sum$cp),
    min(mod.bwd.sum$cp),
    pch = 4, col = "red", lwd = 7)

plot(mod.bwd.sum$bic,
    sub = "Backward Stepwise Regression",
    xlab = "Subset Size", ylab = "BIC",
    pch = 20, type = "l")
points(which.min(mod.bwd.sum$bic),
    min(mod.bwd.sum$bic),
    pch = 4, col = "red", lwd = 7)

plot(mod.bwd.sum$adjr2,
    sub = "Backward Stepwise Regression",
    xlab = "Subset Size", ylab = "Adjusted R^2",
    pch = 20, type = "l")
points(which.max(mod.bwd.sum$adjr2),
    max(mod.bwd.sum$adjr2),
    pch = 4, col = "red", lwd = 7)
title("Information Criteria ~ Subset Size", outer = TRUE)
par(opar)
```
~~~

The selected models are

~~~
```{r}
# Forward Stepwise Regression with Cp
coefficients(mod.fwd, id = which.min(mod.fwd.sum$cp))

# Forward Stepwise Regression with BIC
coefficients(mod.fwd, id = which.min(mod.fwd.sum$bic))

# Forward Stepwise Regression with Adjusted R^2
coefficients(mod.fwd, id = which.max(mod.fwd.sum$adjr2))

# Backward Stepwise Regression with Cp
coefficients(mod.bwd, id = which.min(mod.bwd.sum$cp))

# Backward Stepwise Regression with BIC
coefficients(mod.bwd, id = which.min(mod.bwd.sum$bic))
~~~

```
# Backward Stepwise Regression with Adjusted R^2
coefficients(mod.bwd, id = which.max(mod.bwd.sum$adjr2))
```

Similar results are obtained as that in the best susbet selection, except that the backward selection under the adjusted $R^2$ criteria chooses a model of size 6 other than 5. Moreover, even under the case that both best subset selection and forward/backward stepwise regression choose the model of same subset size, as can be observed in the cases identified by $C_p$ and adjusted $R^2$.

\newpage

##### Linear Model

```{r}
mod2.lm <- lm(Siri_Y ~ ., data = bodyfat.train)
mse2.lm <- mean( ( bodyfat.test$Siri_Y - predict(mod2.lm, bodyfat.test) )^2 )
mse2.lm
```

The prediction MSE estimate for linear model is `r formatC(mse2.lm, format = "e", digits = 4)`.

##### LASSO

Now we train the LASSO model incorporating all 13 variables, using 5-fold cross-validation to tune the parameter $\lambda$.

```{r}
bodyfat.x <- model.matrix(Siri_Y ~ ., bodyfat.train)

set.seed(6639)
grid <- 10^seq (2, -2, length = 100)
mod.lasso <- cv.glmnet(x = bodyfat.x, y = bodyfat.train$Siri_Y,
                       lambda = grid, nfolds = 5, alpha = 1)
summary(mod.lasso)
plot(mod.lasso)
```

$\lambda$ is identified as `r mod.lasso$lambda.min`. Then we train the lasso regresssion on the training data with $\lambda =$ `r mod.lasso$lambda.min` and predict on the test set.

```{r}
mse.lasso <- mean( (bodyfat.test$Siri_Y -
```

~~~
                    predict(mod.lasso, s = "lambda.min",
                    newx = model.matrix(Siri_Y ~ ., bodyfat.test))
                    )^2 )
mod.lasso$lambda.min
mse.lasso
```

The prediction MSE estimate for lasso regression is `r formatC(mse.lasso, format = "e", digits = 4)`. The number of
predictive variables is `r sum(coef(mod.lasso, "lambda.min") != 0)`, with coefficients


```{r}
coef(mod.lasso, "lambda.min")
sum(coef(mod.lasso, s = "lambda.min") != 0)
```

$$y_{Lasso}=-26.96+0.0496*Age-0.1285*Weight-0.0092*Height+0.92*Abdomen\\-0.06*Hip+0.28*Thigh+0.03*B
iceps+0.4495*Forearm-1.4707*Wrist$$


```{r}
lambdas_to_try <- 10^seq(-3, 5, length.out = 100)
res <- glmnet(x=bodyfat.x, y=bodyfat.train$Siri_Y, alpha = 1, lambda = lambdas_to_try,
          standardize = FALSE)
plot(res, xvar = "lambda")
legend("bottomright", lwd = 1, col = 1:6, legend = colnames(bodyfat.x), cex = .7)
```



##### Ridge Regression
Now we use 5-fold cross-validation to tune $\lambda$ in the ridge regression of `Siri_Y` on other variables in
`bodyfat.train`.

```{r}
set.seed(6639)
grid = 10^seq(6, -2, length=100)
mod.ridge <- cv.glmnet(x = bodyfat.x, y = bodyfat.train$Siri_Y,
                    lambda = grid, nfolds = 5, alpha = 0)
plot(mod.ridge)
```


$\lambda$ is identified as `r mod.ridge$lambda.min`. Then we train the ridge regresssion on the training data with
$\lambda =$ `r mod.ridge$lambda.min` and predict on the test set.

```{r}
mse.ridge <- mean( (bodyfat.test$Siri_Y -
~~~

~~~~
                    predict(mod.ridge, s = "lambda.min",
                    newx = model.matrix(Siri_Y ~ ., bodyfat.test))
                    )^2 )
mod.ridge$lambda.min
mse.ridge
```
The prediction MSE estimate for ridge regression is `r formatC(mse.ridge, format = "e", digits = 4)`.


```{r}
coef(mod.ridge, "lambda.min")
sum(coef(mod.ridge, s = "lambda.min") != 0)
```


$$y_{Ridge}=-18.76+0.0668*Age-0.1039*Weight-0.0580*Height-0.4827*Neck\\+0.0168*Chest+0.8691*Abdomen-0.1182*Hip+0.36*Thigh-0.0432*Knee\\+0.0128*Ankle+0.0454*Biceps+0.4696*Forearm-1.6056*Wrist$$

##### Discussion
The predicted MSEs of linear model, ridge regression, and LASSO summarized as follows.


```{r echo=FALSE}
bodyfat.test.mv <- (nrow(bodyfat.test) - 1)/nrow(bodyfat.test) * var(bodyfat.test$Siri_Y)
```


|**Method**        |**Linear Model**|**Stepwise Selection**|**With Interaction**|**LASSO**|**Ridge Regression**|**GAM**|
|:------------:|:--------------:|:--------------------:|:------------------:|:-------:|:------------------:|:-------:|
|**Test MSE**  |`r formatC(mse2.lm, digits = 4)`|20.89|19.30|`r formatC(mse.lasso, digits = 4)`|`r formatC(mse.ridge, digits = 4)`|22.12|
|**Test $R^2$**|`r round(1-mse2.lm/bodyfat.test.mv, digits = 2)`|0.73|0.75|`r round(1-mse.lasso/bodyfat.test.mv, digits = 2)`|`r round(1-mse.ridge/bodyfat.test.mv, digits = 2)`|NA*|


As we can see, the ridge regression and LASSO have comparable prediction performance as the linear model, but both of them are relatively superior. Models accounts for a high proportion of variation in the test response `Siri_Y`.

```{r warning=FALSE, message=FALSE}
# GAM-normal
if(!require('gam')){install.packages('gam');library('gam')}
if(!require('data.table')){install.packages('data.table');library('data.table')}

gam.model=gam(Siri_Y~.,data = bodyfat.train,family = "gaussian")
gam.pred=predict.Gam(gam.model,newdata = bodyfat.test)
mean( (gam.pred-bodyfat.test$Siri_Y)^2 )
```
~~~~