Analysis of the Bodyfat Dataset using Linear Modeling: Prediction and Inference

Hee Won Ahn, hmahn@ucdavis.edu
Dae Hyeun Cheong, dcheong@ucdavis.edu

December 11, 2023

Table of Contents

I Abstract

We explored two key research questions using the body fat dataset, comprising 252 observations and fifteen distinct body characteristics: first, whether linear modeling could serve as a viable alternative in measuring body fat, and second, how individuals should manage their body fat levels. Our findings indicate that linear modeling is not an optimal alternative, primarily due to potential limitations. Notably, our analysis underscores that managing thigh circumference emerges as the most effective strategy for adjusting the level of body fat.

II Introduction

Obesity has emerged as a growing concern in the United States, leading many individuals to actively monitor their body fat percentage as a proactive approach to maintaining optimal health. However, in the absence of the body composition analyzer, gauging body fat percentage posed a formidable challenge. Historically, estimations relied on the Siri Equation (1956), which employed body density measured through labor-intensive methods like underwater weighing.

This report delves into the quest for a more convenient and cost-effective method to precisely measure body fat percentage. Utilizing linear modeling, we aim to identify key individual characteristics that predict body fat percentage while acknowledging the inherent limitations of such methods. Simultaneously, our investigation endeavors to pinpoint the most influential determinants of body fat and offers guidance to individuals on managing their body fat levels, promoting informed choices for individuals striving to maintain overall well-being.

III Exploratory Data Analysis

Our dataset comes from Department of Mathematics & Computer Science at South Dakota School of Mines & Technology, investigating 252 men and their fifteen different body characteristics.

Table 1: Data Dictionary

Body Characteristics	***Unit***
Density and Percent Body Fat	*N/A*
Age	*Years*
Weight	*Pounds*
Height	*Inches*
Neck, Chest, Abdomen, Hip, Thigh, Knee, Ankle, Biceps, Forearm, Wrist	*Circumference in centimeters*

Utilizing the provided body fat dataset, we conducted an exploration of summary statistics (refer to Code 1 in the appendix) and variable distributions using histograms and the summary function in R (refer to Figures 1 and 2 in the appendix). Additionally, we examined the relationship between the response variable (percent body fat) and other predictors using scatter plots (see Figures 3 and 4 in the appendix). The exploratory data analysis yielded valuable insights into the dataset.

1. The histogram depicting the response variable (percent body fat) revealed a distribution somewhat resembling the normal distribution. Consequently, it was determined that no further transformation of the response variables is necessary.

2. No predictor variables displayed significant data skewness or exhibited a non-linear relationship with the response variables within the dataset. This suggests that additional transformations for predictor variables are unnecessary, and the utilization of higher power terms is not entirely justified based on the observations from these scatter plots.

3. Observations indicate the presence of potential outliers in the dataset. Notably, in the height vs. percent body fat dataset (Figure 3 in the appendix), an individual with an almost 20-inch height is apparent, suggesting a measurement error. Both the scatter plot and histogram strongly signal the need for an outlier analysis.

Correlation Plot and Multicollinearity

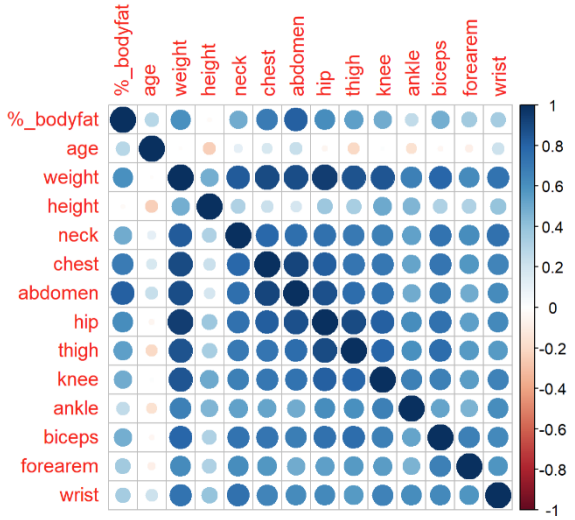


Figure 5: Correlation plot

We also generated a correlation plot to examine interactions between variables. Unsurprisingly, the correlation plot revealed pronounced multicollinearity issues, particularly concerning predictor variables, especially in the vicinity of weight and various circumferences. To identify which variables exhibited the highest correlation within the dataset, we calculated the Variance Inflation Factor (VIF) for each variable (refer to Code 2 in the appendix).

The VIF analysis indicated that weight and circumferences of the chest, abdomen, and hip had VIF values exceeding 10, signifying high correlation with other predictor variables. This outcome aligns with expectations, as weight could reasonably be considered a linear combination of other circumference variables in the dataset. Additionally, variables such as hip circumference likely exhibit linear associations with other predictors like thigh, knee, and ankle circumferences.

To mitigate multicollinearity within the dataset, we opted to eliminate predictors with a VIF greater than 10. This decision resulted in a significant reduction in VIF scores for the remaining variables and contributed to an improved correlation plot aesthetic (refer to Code 3 and Figure 6 in the appendix)

Exploring Outliers

Given the presence of clear outliers in the dataset, we computed Cook's distance for each observation, assuming the full model as the true model. While the full model will not serve as our final model, this process enables the identification of influential points, which may indicate outliers. According to the Cook's distance plot and residual vs. leverage plot (see Figure 7 and 8 in the appendix), observations 39, 86, and 175 emerged as the most influential points. Further investigation is warranted to determine whether to exclude these observations.

Observation 39: The high Cook's distance for this observation was attributed to the considerable weight of the individual, leading to extreme values in various circumferences, such as hip circumference. However, the percentage changes between the fitted value based

on all cases and those based on data without observation 39 fell within the range of 1.3% and 12.7% (see Code 4 in the appendix), indicating that observation 39 does not exert strong influence on our prediction. Additionally, upon reviewing the entire scatter plot (refer to Figure 3 and 4 in the appendix), observation 39 appears to align reasonably well with the association between predictor variables and response variables. As a result, we decided to retain observation 39.

Observation 86: The high Cook's distance for this observation was driven by the extreme size of ankle circumferences. However, upon comparison with other predictor variables (such as age, weight, height, and knee), we noted that the individual's ankle circumference was abnormally high, with a maximum percentage change in prediction power of 18% (refer to Code 5 in the appendix). Given the likelihood of a measurement error, we decided to eliminate observation 86.

Observation 175: The moderate Cook's distance for this observation did not reveal any unreasonable values, and its percentage change in prediction power was relatively low (maximum of 8.5%, refer to Code 6 in the appendix). Consequently, we opted to retain observation 175.

Additionally, we chose to exclude observations 42 and 182, as these two exhibited clear measurement errors (observation 42 had a height of about 20 inches, and observation 182 reported 0 percent body fat). In total, three observations are excluded from the analysis. After excluding these observations, we drew new histograms and scatter plot for the better view (refer to Figure 9, 10, 11, 12 in the appendix)

IV Model Selection and Model Diagnostics

Employing a dataset that has undergone outlier exclusion, we employed forward and backward stepwise selection, along with ridge regression, to discern the optimal model for our objectives. This model selection procedure was conducted for each first-order model, both with and without interactions, allowing us to evaluate the effectiveness of the model under different conditions. Since we also want to comment on inference and prediction power of the final model, we divided the dataset into train and test dataset and scaled the predictors to unify the unit (refer to Code 7 in the appendix). This is the summary of all model selection methods and criteria.

Table 2: Model Selection Method and Criteria

Method	Adjusted R	AIC	BIC	MSPE	Mallow Cp	p
Forward Step	0.5667031	1260.5226	1296.8590	44.21901	25.12780	10
Backward Step	0.6281853	1245.0272	1337.5198	42.77823	14.13630	27
Ridge	0.5125974	741.7820	994.0661	43.71722	79.16427	46
Stepwise w/o Interaction	0.5075199	1282.4261	1305.5492	39.10514	5.45108	6
Ridge w/o Interaction	0.4999856	716.8953	768.0195	38.63170	12.37304	10

(refer to Code 10 in the appendix)

1. Forward and Backward Stepwise Selection using AIC criteria:

   Initially, we applied the full model with all first-order interactions included. Subsequently, we executed forward and backward stepwise selection, yielding models with fewer

variables as presented in the table above. Excluding interaction terms, both forward and backward stepwise selections converged on the same model comprising five variables, including the intercept (refer to Code 8 in the appendix).

2. Ridge Regression:

   In ridge regression, we conducted two separate analyses—one with interaction terms and another without. Optimal lambdas were determined through cross-validation (lambda = 1.20120 for the model with interaction terms and lambda = 0.6506507 for the model without interaction terms), considering various candidates for lambda (refer to Code 9 and Figures 13 and 14 in the appendix). Given that we employed the full model without variable selection in ridge regression, assuming irrelevant coefficients would approach zero, the model complexity remained high.

The comparative table of different methods and criteria highlights crucial insights into model selection:

1. The ridge regression model with interaction terms exhibited elevated bias (Cp substantially exceeding p) and relatively high variance (model complexity is at the highest). Conversely, the ridge regression model without interaction terms demonstrated the most favorable outcome, displaying relatively low variance (low model complexity) and a notable reduction in bias (Cp is approximately equal to p). This emerges as a robust contender for the final model.

2. Overall, regardless of the inclusion of interaction terms, stepwise selection exhibited elevated AIC and BIC, indicative of high in-sample SSE. The model selected by forward and backward stepwise selection with interaction terms displayed considerable bias and variance. However, the model chosen by stepwise selection without interaction terms demonstrated improvement in both variance and bias (lower model complexity and Cp is approximately equal to p). This model also merits consideration as a strong candidate for the final model.

3. Notably, only the model without interaction terms was selected among our candidates. This underscores that numerous interactions may act as nuisances, potentially diverting our analysis of the dataset (e.g., overfitting).

Among the selected candidates, we opted for the model chosen by stepwise selection without interactions (the 4th row in the table above). Although the ridge model without interactions displayed significantly lower AIC and BIC values, a comparison of Cp and MSPE revealed similar predictive power and bias. Adhering to the Principle of Parsimony, we favored the model with fewer predictors.

An examination of the diagnostic plot for the chosen final model validates that the linearity, normality, and constant variance assumptions are well-maintained (refer to Figure 15 and 16 in the appendix)

V Prediction and Possible Limitation

Despite selecting a model with a relatively low MSPE, the final model's MSPE value itself remains suboptimal, suggesting that our model struggles to predict the percentage of body fat accurately. This conclusion is supported by the fitted vs. actual plot, revealing that our linear model lacks robust predictive capabilities. While the predictions are not entirely implausible, as evidenced by the clustered points around the line and a consistent positive slope, numerous points deviate significantly from the line, underscoring the model's limitations in prediction accuracy.

Interestingly, the fitted vs. actual plot not only underscores the model's predictive shortcomings but also highlights potential limitations in employing linear modeling for predicting the percentage of body fat. A noticeable pattern emerges: for lower actual values, our model tends to overpredict body fat, and conversely, for higher actual values, it tends to underestimate. This discrepancy may arise if our predictor variables exhibit inverse proportional relationships with the response variable in a multi-variable space[1]. The figure[2] below illustrates that in the presence of an inverse proportional relationship with predictor variables, the linear regression line may overestimate at lower x values and severely underestimate as x values increase—coinciding with the observed trend in the fitted vs. actual plot. If this limitation holds true, exploring non-linear models could enhance our prediction power. However, delving into this requires further investigation, which falls beyond the scope of this report.

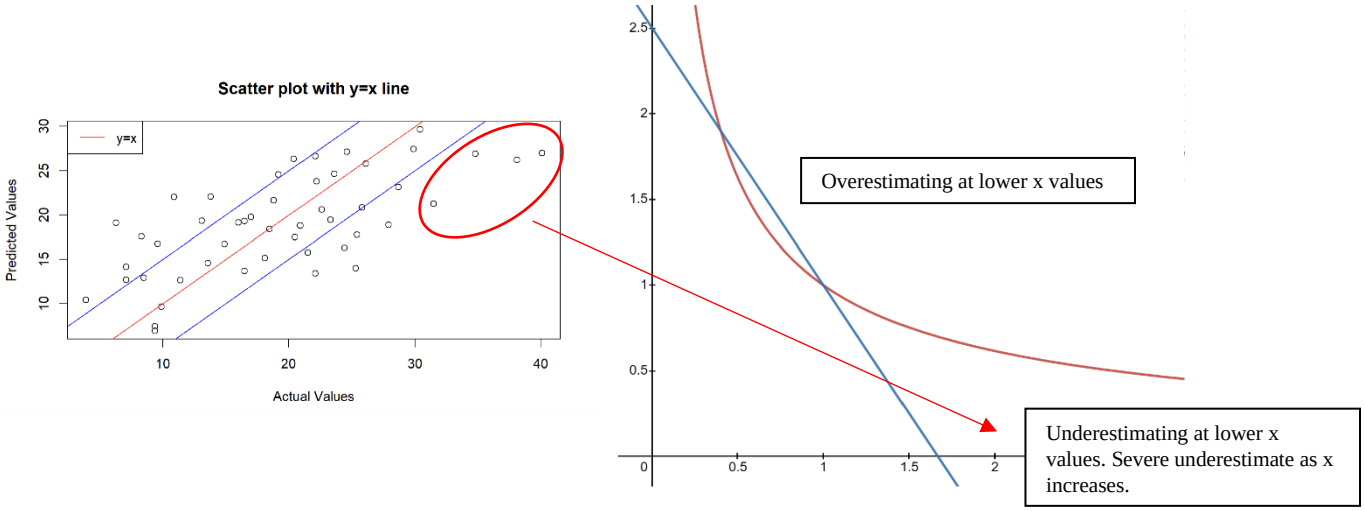


Figure 17: Fitted vs Actual Plot and Simplified Representation of Possible Limitation

VI Inference

Before delving into the model's inference, we amalgamated our training and test datasets and re-fitted the model to the combined data (see Code 11 in the appendix).

The stepwise forward/backward selection, utilizing the AIC criteria, identified a total of five predictors out of ten as relevant variables (thigh circumference, age, height, forearm

[1] The assumption is based on the fact that percent of body fat is inversely proportional to the body density by Siri Equation (1956)
[2] While the plot simplifies our model into a simple linear regression setting—an approximation rather than an accurate representation—it hints at the challenges linear regression may face.

circumference, and wrist circumference). The fitted coefficients of the model provide valuable insights into the percentage of body fat (refer to Code 11 in the appendix).

1. Height and wrist circumference exhibit negative regression coefficients, indicating that assuming all other factors remain constant, an increase in height or wrist circumference corresponds to a decrease in body fat percentage. This aligns with the notion that greater height and wrist circumference may serve as indicators of a larger body volume, potentially leading to a lower overall percentage of body fat.

2. Other variables (age, thigh circumference, and forearm circumference) show that, assuming all else remains constant, an increase in these values corresponds to an increase in the overall percentage of body fat. This is logical because:

    A) As individuals age, a higher percentage of body fat is expected due to slower metabolism and altered body composition.[3]

    B) Body fat tends to be stored in areas such as the thigh, biceps, and forearm.[4]

Among the predictors, thigh, age, and height have the largest coefficients. This suggests that, aside from age and height (which are beyond our control), managing thigh circumference could be a strategic approach to reducing body fat percentage. This conclusion also aligns with the ridge coefficient from the ridge regression model (refer to Code 12 in the appendix)

V Conclusion

As outlined in the introduction, our objective was to explore whether linear regression analysis could provide a more convenient and accurate alternative to measuring body fat compared to the Siri Equation. Additionally, we sought to offer guidance to individuals aiming to increase or decrease their body fat percentage for optimal health.

Regrettably, upon thorough analysis of the model, we arrived at the conclusion that employing a linear regression model may not serve as a viable alternative for accurately measuring body fat percentage, given its overall weak predictive power. However, the fitted vs. actual plot reveals a potential limitation of the linear regression model, suggesting an inverse proportional relationship between predictors and body fat percentage. This raises the possibility that employing a non-linear model may enhance overall prediction accuracy.

From an inferential standpoint, our results provide meaningful insights. The top three predictors with the most significant impact on determining body fat percentage are thigh circumference, age, and height. Individuals seeking to manage their body fat percentage should pay attention to thigh circumference, as it emerges as the most critical factor in determining body fat percentage according to our model. Since body fat percentage serves as a crucial indicator of individual health and our model clearly illustrates its increase with age, it is imperative for government health officials to monitor body fat percentage in older populations and guide them in maintaining healthy levels of body fat.

[3] See Reference 1
[4] See Reference 2

Appendix

Figure 1

```
par(mfrow = c(2,3), mar = c(2,1,2,1))
for(i in 1:6){
  hist(bodyfat[,i], main = paste("histogram of ", names(bodyfat)[i]))
}
```

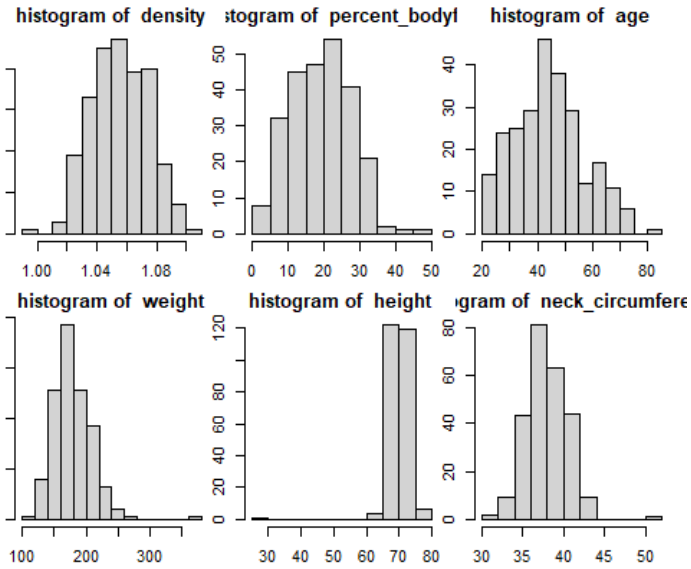


Figure 2

```
par(mfrow = c(3,3), mar = c(2,1,2,1))
for(i in 7:15){
  hist(bodyfat[,i], main = paste("histogram of ", names(bodyfat)[i]))
}
```

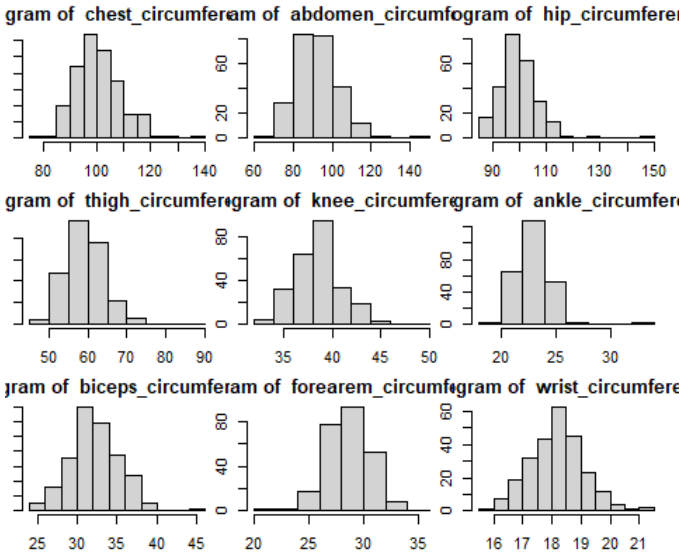


Figure 3

```
par(mfrow = c(3,3), mar = c(2,1,2,1))
for(i in 3:11){
  plot(bodyfat[,i], bodyfat$percent_bodyfat,
       main = paste(names(bodyfat)[i], "/ percent bodyfat"))
}
```

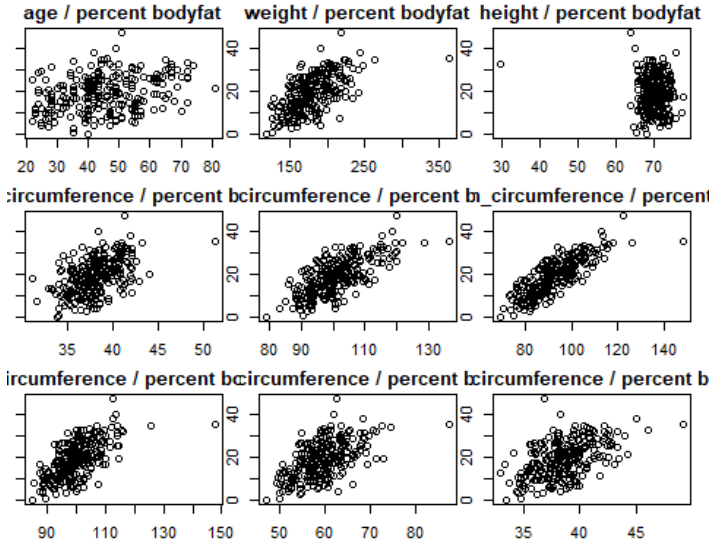


Figure 4

```
par(mfrow = c(2,2), mar = c(2,1,2,1))
for(i in 12:15){
  plot(bodyfat[,i], bodyfat$percent_bodyfat,
       main = paste(names(bodyfat)[i], "/percent bodyfat"))
}
```

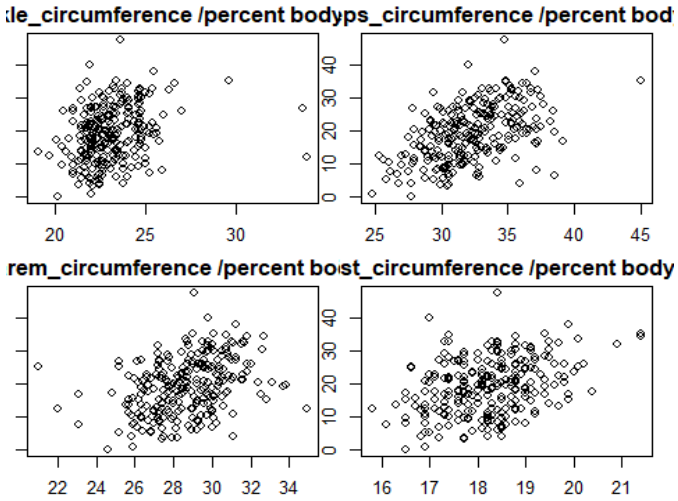


Figure 6

```
cleaned_bodyfat_2 <- cleaned_bodyfat[ ,c(-4,-7, -8,-9)]

M <- cor(cleaned_bodyfat_2[ ,2:11])
colnames(M) <- c("bodyfat", "age", "hght", "neck", "thigh", "knee", "ankle", "bicp", "arm", "w
rist")
rownames(M) <- c("bodyfat", "age", "hght", "neck", "thigh", "knee", "ankle", "bicp", "arm", "w
rist")
corrplot.mixed(M)
```

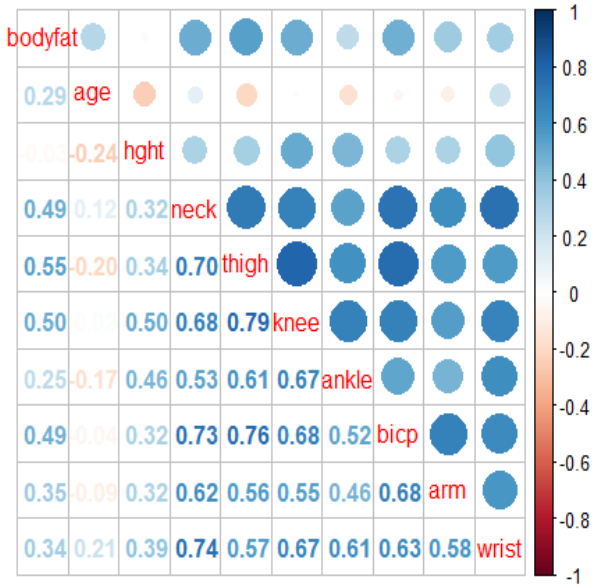
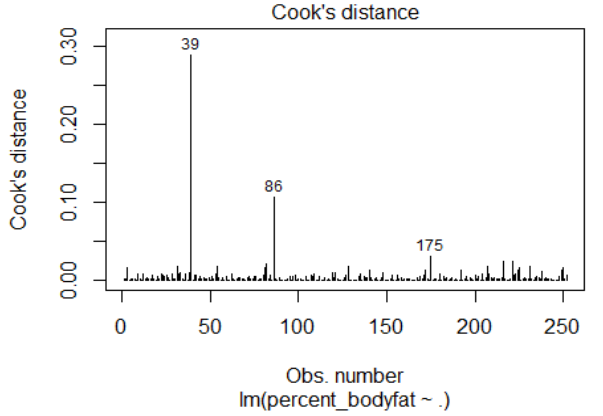


Figure 7 and 8

```
## Not using density as our predictors.
predictors <- bodyfat[ ,2:15]
sample_fit <- lm(percent_bodyfat ~ . , data = predictors)

plot(sample_fit, which = 4)
plot(sample_fit, which = 5
```



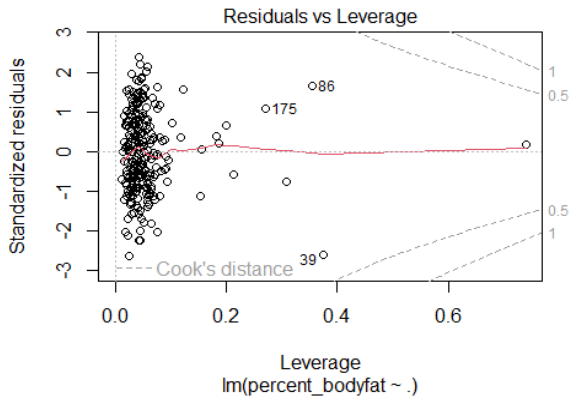
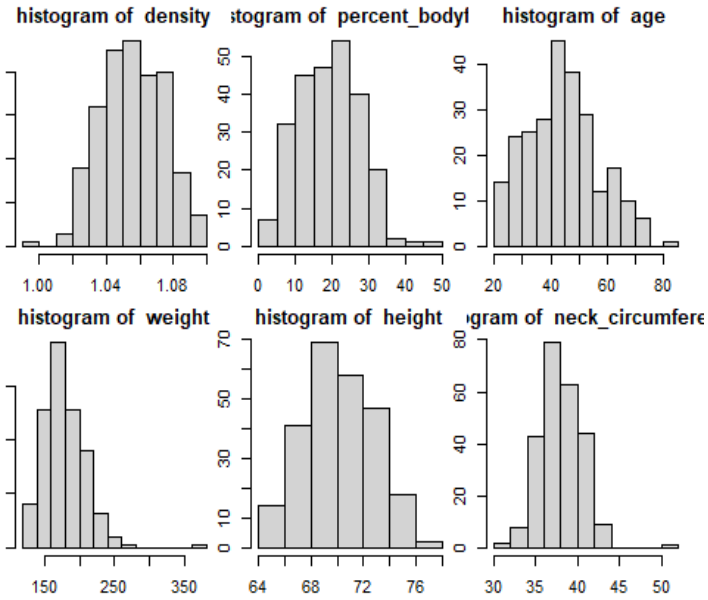
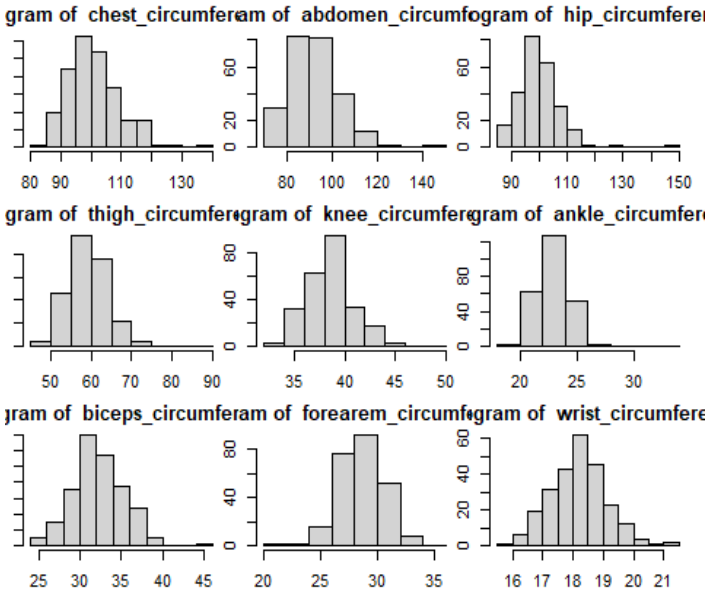


Figure 9, 10, 11, 12

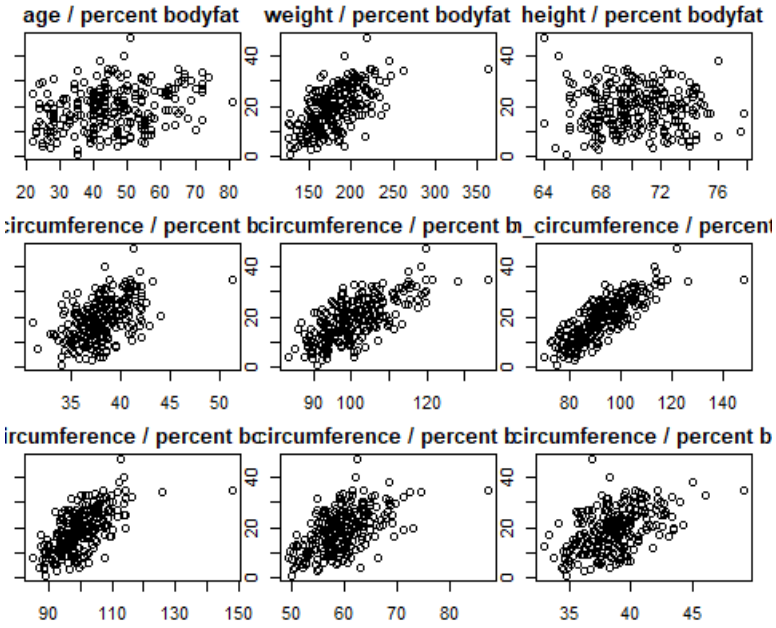
```
cleaned_bodyfat <- bodyfat[c(-86,-42, -182), ]

par(mfrow = c(2,3), mar = c(2,1,2,1))
for(i in 1:6){
  hist(cleaned_bodyfat[,i], main = paste("histogram of ", names(cleaned_bodyfat)[i]))
}
```




```
par(mfrow = c(3,3), mar = c(2,1,2,1))
for(i in 7:15){
  hist(cleaned_bodyfat[,i], main = paste("histogram of ", names(cleaned_bodyfat)[i]))
}
```



```
par(mfrow = c(3,3), mar = c(2,1,2,1))
for(i in 3:11){
  plot(cleaned_bodyfat[,i], cleaned_bodyfat$percent_bodyfat,
       main = paste(names(cleaned_bodyfat)[i], "/ percent bodyfat"))
}
```




```
par(mfrow = c(2,2), mar = c(2,1,2,1))
for(i in 12:15){
  plot(cleaned_bodyfat[,i], cleaned_bodyfat$percent_bodyfat,
       main = paste(names(cleaned_bodyfat)[i], "/ percent bodyfat"))
}
```

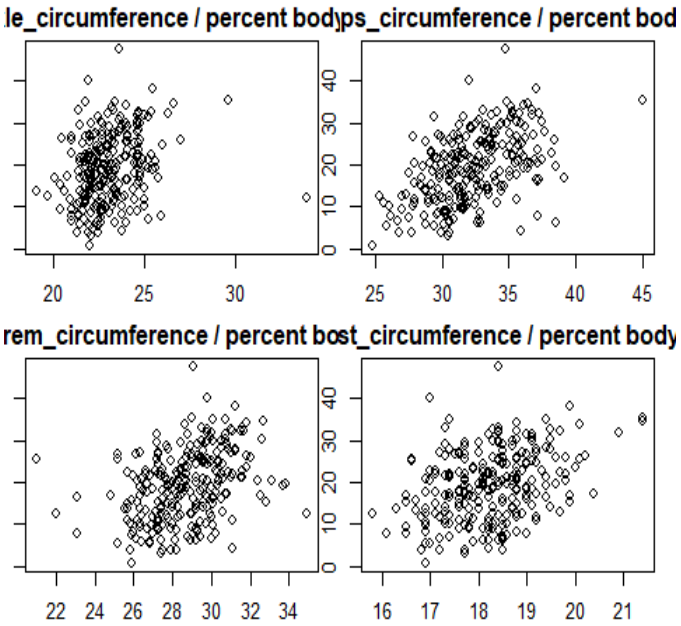


Figure 13

```
plot(cv_fit)
```

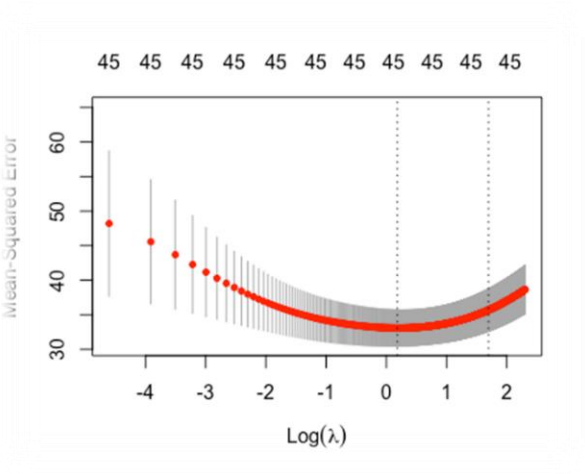


Figure 14

```
plot(cv_fit_no_interaction)
```

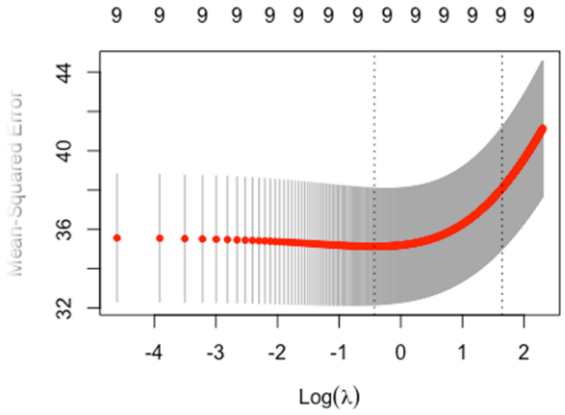
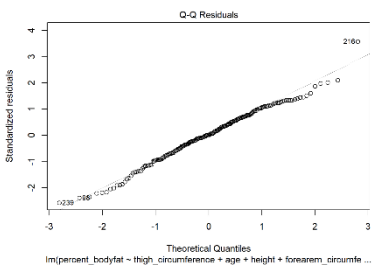


Figure 15, 16

```
plot(forward_aic_no_interation,which=1)
plot(forward_aic_no_interation,which=2)
```



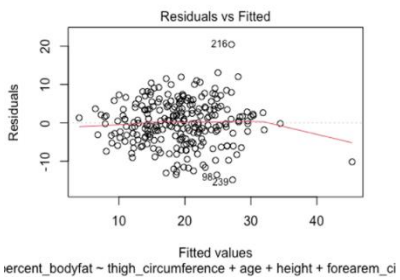
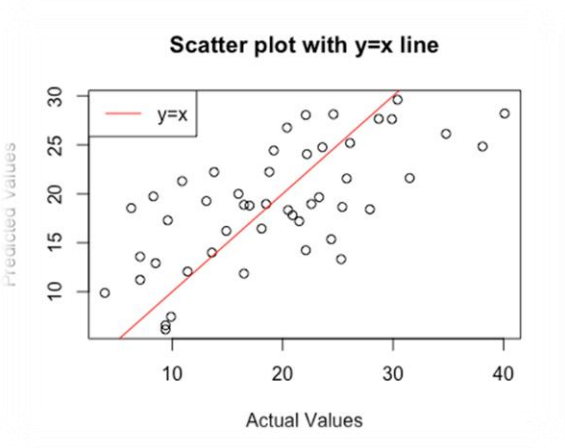


Figure 17

```
plot(test_Y, predict(forward_aic_no_interation, test_X),
     main = "Scatter plot with y=x line",
     xlab = "Actual Values",
     ylab = "Predicted Values")
abline(a = 0, b = 1, col = "red")
legend("topleft", legend = "y=x", col = "red", lty = 1)
```



Code 1

```
summary(bodyfat)
```

```
##     density      percent_bodyfat      age            weight
##  Min.   :0.995   Min.   : 0.00   Min.   :22.00   Min.   :118.5
##  1st Qu.:1.041   1st Qu.:12.47   1st Qu.:35.75   1st Qu.:159.0
##  Median :1.055   Median :19.20   Median :43.00   Median :176.5
##  Mean   :1.056   Mean   :19.15   Mean   :44.88   Mean   :178.9
##  3rd Qu.:1.070   3rd Qu.:25.30   3rd Qu.:54.00   3rd Qu.:197.0
##  Max.   :1.109   Max.   :47.50   Max.   :81.00   Max.   :363.1
##      height      neck_circumference chest_circumference abdomen_circumference
##  Min.   :29.50   Min.   :31.10      Min.   : 79.30      Min.   : 69.40
##  1st Qu.:68.25   1st Qu.:36.40      1st Qu.: 94.35      1st Qu.: 84.58
##  Median :70.00   Median :38.00      Median : 99.65      Median : 90.95
##  Mean   :70.15   Mean   :37.99      Mean   :100.82      Mean   : 92.56
##  3rd Qu.:72.25   3rd Qu.:39.42      3rd Qu.:105.38      3rd Qu.: 99.33
##  Max.   :77.75   Max.   :51.20      Max.   :136.20      Max.   :148.10
##  hip_circumference thigh_circumference knee_circumference ankle_circumference
##  Min.   : 85.0     Min.   :47.20       Min.   :33.00      Min.   :19.1
##  1st Qu.: 95.5     1st Qu.:56.00       1st Qu.:36.98      1st Qu.:22.0
##  Median : 99.3     Median :59.00       Median :38.50      Median :22.8
##  Mean   : 99.9     Mean   :59.41       Mean   :38.59      Mean   :23.1
##  3rd Qu.:103.5     3rd Qu.:62.35       3rd Qu.:39.92      3rd Qu.:24.0
##  Max.   :147.7     Max.   :87.30       Max.   :49.10      Max.   :33.9
##  biceps_circumference forearem_circumference wrist_circumference
##  Min.   :24.80        Min.   :21.00          Min.   :15.80
##  1st Qu.:30.20        1st Qu.:27.30          1st Qu.:17.60
##  Median :32.05        Median :28.70          Median :18.30
##  Mean   :32.27        Mean   :28.66          Mean   :18.23
##  3rd Qu.:34.33        3rd Qu.:30.00          3rd Qu.:18.80
##  Max.   :45.00        Max.   :34.90          Max.   :21.40
```

```r
sapply(bodyfat, class)
```

```
##                density        percent_bodyfat                    age
##              "numeric"              "numeric"              "integer"
##                 weight                 height     neck_circumference
##              "numeric"              "numeric"              "numeric"
##    chest_circumference  abdomen_circumference      hip_circumference
##              "numeric"              "numeric"              "numeric"
##    thigh_circumference     knee_circumference    ankle_circumference
##              "numeric"              "numeric"              "numeric"
##   biceps_circumference forearem_circumference    wrist_circumference
##              "numeric"              "numeric"              "numeric"
```

```r
any(is.na(bodyfat)) ## No NAs in the dataset. -> Imputation is not needed
```

```
## [1] FALSE
```

Code 2

```r
## Calculating VIF

X <- cleaned_bodyfat[ ,3:15]
r_inv <- solve(cor(X))
diag(r_inv)
```

```
##                    age                 weight                 height
##               2.307747              44.453361               2.942193
##     neck_circumference    chest_circumference  abdomen_circumference
##               4.388301              10.156506              12.529061
##      hip_circumference    thigh_circumference     knee_circumference
##              14.395300               7.782930               4.756334
##    ankle_circumference   biceps_circumference forearem_circumference
##               2.398978               3.680700               2.163727
##    wrist_circumference
##               3.435442
```

Code 3

```r
X <- cleaned_bodyfat[ ,3:15]
X_new <- X[ , c(-2,-5,-6,-7)]
r_inv <- solve(cor(X_new))
diag(r_inv)
```

```
##                    age                 height     neck_circumference
##               1.763273               1.659615               3.439267
##    thigh_circumference     knee_circumference    ankle_circumference
##               4.909824               4.450078               2.300685
##   biceps_circumference forearem_circumference    wrist_circumference
##               3.400674               2.105006               3.324212
```

Code 4, 5, 6

```r
## Due to very high weight. Kinda align with association. Included.
sample_fit2 <- lm(percent_bodyfat ~ . , data = predictors[-39, ])
per.change = abs((sample_fit$fitted.values - predict.lm(sample_fit2, predictors[, 2:14]))/samp
le_fit$fitted.values) * 100
summary(per.change)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  0.01305  0.64187  1.43711  1.88499  2.34386 12.71331

## Due to very abnormal ankle size
sample_fit2 <- lm(percent_bodyfat ~ . , data = predictors[-86, ])
per.change = abs((sample_fit$fitted.values - predict.lm(sample_fit2, predictors[, 2:14]))/samp
le_fit$fitted.values) * 100
summary(per.change)

##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  0.01845  0.33912  0.73447  1.15334  1.37163 18.22589

# Not very influencial in prediction <- included
sample_fit2 <- lm(percent_bodyfat ~ . , data = predictors[-175, ])
per.change = abs((sample_fit$fitted.values - predict.lm(sample_fit2, predictors[, 2:14]))/samp
le_fit$fitted.values) * 100
summary(per.change)

##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.008587 0.172810 0.411148 0.650410 0.786848 8.510426
```

Code 7

```
set.seed(1234)
train_index = createDataPartition(data$percent_bodyfat, p = 0.8, list = FALSE)
train_X = scale(data[train_index,-1],center=T,scale=T)
train_X=as.data.frame(train_X)
train_Y = data[train_index,1 ]
train_data=as.data.frame(cbind(train_Y,train_X))
colnames(train_data)[1] = 'percent_bodyfat'

test_X =scale(data[-train_index,-1],center=T,scale=T)
test_X= as.data.frame(test_X)
test_Y = data[-train_index, 1 ]
test_data=as.data.frame(cbind(test_Y,test_X))
colnames(test_data)[1] = 'percent_bodyfat'

none_mod = lm(percent_bodyfat~1, data=train_data) #model with only intercept
full_mod_interaction = lm(percent_bodyfat ~.^2, data = train_data)

#full model with interaction terms
full=lm(percent_bodyfat ~., data = train_data) # full model without interaction terms
```

Code 8

```
forward_aic=stepAIC(none_mod, scope=list(upper=full_mod_interaction, lower = ~1),
direction="both", k=2, trace = FALSE)
backward_aic=stepAIC(full_mod_interaction, scope=list(upper=full_mod_interaction, lower = ~1),
direction="both", k=2, trace = FALSE)


forward_aic_no_interation=stepAIC(none_mod, scope=list(upper=full, lower = ~1),
direction="both", k=2, trace = FALSE)
backward_aic_no_interation=stepAIC(full, scope=list(upper=full, lower = ~1), direction="both",
k=2, trace = FALSE)


# same model using full model without interaction term
```

```
all_variables = colnames(train_X)
for (i in 1:(length(all_variables) - 1)) {
  for (j in (i+1):length(all_variables)) {
    interaction_name <- paste(all_variables[i], all_variables[j], sep = ":")

train_X[,as.character(interaction_name)]=train_X[[all_variables[i]]]*train_X[[all_variables[j]
]]}}  # adding interaction terms to data matrix

all_variables = colnames(test_X)
for (i in 1:(length(all_variables) - 1)) {
  for (j in (i+1):length(all_variables)) {
    interaction_name <- paste(all_variables[i], all_variables[j], sep = ":")

test_X[,as.character(interaction_name)]=test_X[[all_variables[i]]]*test_X[[all_variables[j]]]}
} # adding interaction terms to data matrix
```

Code 9

```
X=train_X # interation terms included
y=train_Y

X_no_interaction=X[,1:9]

lambdas=seq(0,10,length=1000)
cv_fit = cv.glmnet(as.matrix(X), y, alpha = 0, lambda = lambdas)
plot(cv_fit)

lambda_min=cv_fit$lambda.min
ridge = glmnet(X, y, alpha = 0, lambda = lambda_min)

lambdas=seq(0,10,length=1000)
cv_fit_no_interaction = cv.glmnet(as.matrix(X_no_interaction), y, alpha = 0, lambda = lambdas)
plot(cv_fit_no_interaction)

lambda_min2=cv_fit_no_interaction$lambda.min
ridge_no_interation= glmnet(X_no_interaction, y, alpha = 0, lambda = lambda_min2)
```

Code 10

```
forward_sum=summary(forward_aic)
backward_sum=summary(backward_aic)
forward_sum_no_interaction=summary(forward_aic_no_interation)

# R square
ridge_y_pred = predict(ridge, newx = as.matrix(X), s = lambda_min)
ridge_SSE <- sum((y - ridge_y_pred)^2)  # Residual Sum of Squares
SST <- sum((y - mean(y))^2) # Total Sum of Squares
ridge_SSR=SST-ridge_SSE
ridge_r_squared <- 1 - ridge_SSE/SST

ridge_y_pred2 = predict(ridge_no_interation, newx = as.matrix(X_no_interaction), s =
lambda_min2)
ridge_SSE2<- sum((y - ridge_y_pred2)^2)  # Residual Sum of Squares
ridge_SSR2=SST-ridge_SSE2
```

```
ridge_r_squared2 <- 1 - ridge_SSE2/SST
R_square=c(forward_sum$r.squared,backward_sum$r.squared,ridge_r_squared,forward_sum_no_interac
tion$r.squared,ridge_r_squared2)
# adj_r_square
n=length(y)
p=length(coef(ridge))
p2=length(coef(ridge_no_interation))
ridge_adj.r=1-(ridge_SSE/(n-p))/(SST/(n-1))
ridge_adj.r2=1-(ridge_SSE2/(n-p2))/(SST/(n-1))

adj_R_square=c(forward_sum$adj.r.squared,backward_sum$adj.r.squared,ridge_adj.r,forward_sum_no
_interaction$adj.r.squared,ridge_adj.r2)
# AIC
aic_ridge=n*log(ridge_SSE/n)+2*p
aic_ridge2=n*log(ridge_SSE2/n)+2*p2

aic=c(AIC(forward_aic),AIC(backward_aic),aic_ridge,AIC(forward_aic_no_interation),aic_ridge2)
#BIC
bic_ridge=n*log(ridge_SSE/n)+log(n)*p
bic_ridge2=n*log(ridge_SSE2/n)+log(n)*p2

bic=c(BIC(forward_aic),BIC(backward_aic),bic_ridge,BIC(forward_aic_no_interation),bic_ridge2)
#Mean squared prediction error
forward_mspe=sum((test_Y-predict(forward_aic, test_X))^2)/length(test_Y)
backward_mspe=sum((test_Y-predict(backward_aic, test_X))^2)/length(test_Y)
ridge_mspe=sum((test_Y-predict(ridge, as.matrix(test_X)))^2)/length(test_Y)
forward2_mspe=sum((test_Y-predict(forward_aic_no_interation, test_X[,1:9]))^2)/length(test_Y)
ridge_mspe2=sum((test_Y-predict(ridge_no_interation,
as.matrix(test_X[,1:9])))^2)/length(test_Y)

MSPE=c(forward_mspe,backward_mspe,ridge_mspe,forward2_mspe,ridge_mspe2)

# Mallos' Cp criterion

forward_mallow=(anova(forward_aic)[10,2]/anova(full_mod_interaction)[46,3])-(length(train_Y)-
2*length(coef(forward_aic)))
backward_mallow=(anova(backward_aic)[27,2]/anova(full_mod_interaction)[46,3])-
(length(train_Y)-2*length(coef(backward_aic)))
ridge_mallow=(ridge_SSE/anova(full_mod_interaction)[46,3])-(length(train_Y)-
2*length(coef(ridge)))
forward_mallow2=(anova(forward_aic_no_interation)[6,2]/anova(full)[10,3])-(length(train_Y)-
2*length(coef(forward_aic_no_interation)))
ridge_mallow2=(ridge_SSE2/anova(full)[10,3])-(length(train_Y)-
2*length(coef(ridge_no_interation)))

mallow=c(forward_mallow,backward_mallow,ridge_mallow,forward_mallow2,ridge_mallow2)

critera= data.frame(
  R_square = R_square,
  adj_R_square = adj_R_square,
  AIC = aic,
  BIC = bic,
  MSPE = MSPE,
  Mallow = mallow,
```

```
number_of_variables=c(length(coef(forward_aic)),length(coef(backward_aic)),length(coef(ridge))
,length(coef(forward_aic_no_interation)),length(coef(ridge_no_interation)))
)

rownames(critera)=c("forward_stepwise","backward_stepwise","ridge","stepwise_without_interacti
on","ridge_without_interaction")

critera
```

```
##                                R_square adj_R_square       AIC       BIC
## forward_stepwise              0.5862014    0.5667031 1260.5226 1296.8590
## backward_stepwise             0.6765212    0.6281853 1245.0272 1337.5198
## ridge                         0.6222630    0.5125974  741.7820  893.7340
## stepwise_without_interaction 0.5198241    0.5075119 1282.4261 1305.5492
## ridge_without_interaction     0.5224862    0.4999856  716.8953  749.9284
##                                    MSPE   Mallow number_of_variables
## forward_stepwise              44.21901 25.12780                  10
## backward_stepwise             42.77823 14.13630                  27
## ridge                         43.71722 79.16427                  46
## stepwise_without_interaction 39.10514  5.45108                   6
## ridge_without_interaction     38.63170 12.37304                  10
```

Code 11

```
summary(forward_aic_no_interation)
```

```
Call:
lm(formula = percent_bodyfat ~ thigh_circumference + age + height +
    forearem_circumference + wrist_circumference, data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max
-14.6804  -3.8164   0.0556   4.0634  20.1126

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            19.0871     0.4065  46.954  < 2e-16 ***
thigh_circumference     5.6283     0.5716   9.846  < 2e-16 ***
age                     3.7640     0.4838   7.781 4.11e-13 ***
height                 -1.2788     0.4677  -2.734  0.00683 **
forearem_circumference  1.2235     0.5300   2.309  0.02202 *
wrist_circumference    -1.2020     0.6242  -1.926  0.05557 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.763 on 195 degrees of freedom
Multiple R-squared:  0.5198,    Adjusted R-squared:  0.5075
F-statistic: 42.22 on 5 and 195 DF,  p-value: < 2.2e-16
```

```
final_model=lm(percent_bodyfat~thigh_circumference+age+height+forearem_circumference+wrist_cir
cumference,data=data)
```

```
Call:
lm(formula = percent_bodyfat ~ thigh_circumference + age + height +
    forearem_circumference + wrist_circumference, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-14.8271  -3.7344   0.0518   3.8744  20.4427

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           19.1426     0.3705  51.666  < 2e-16 ***
thigh_circumference    5.8663     0.5115  11.468  < 2e-16 ***
age                    3.7455     0.4441   8.435 2.99e-15 ***
height                -1.0779     0.4339  -2.484  0.01365 *
forearem_circumference 1.1720     0.4920   2.382  0.01798 *
wrist_circumference   -1.5256     0.5738  -2.659  0.00837 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.847 on 243 degrees of freedom
Multiple R-squared:  0.5105,    Adjusted R-squared:  0.5004
F-statistic: 50.68 on 5 and 243 DF,  p-value: < 2.2e-16
```

Code 12

```
coef(ridge_no_interation)
```

```
## 10 x 1 sparse Matrix of class "dgCMatrix"
##                                s0
## (Intercept)            19.0870647
## age                     2.8267621
## height                 -1.5873122
## neck_circumference      0.6523462
## thigh_circumference     3.3134239
## knee_circumference      1.3283640
## ankle_circumference    -0.2178703
## biceps_circumference    1.0484538
## forearem_circumference  0.6941282
## wrist_circumference    -1.0847001
```

Reference

Murtaugh, Taysha. “The Real Reason You’re Gaining Weight as You Get Older .” Woman’s Day, 25 July 2017, www.womansday.com/health-fitness/a59692/why-we-gain-weight-as-we-age/.

Rail, Kevin. “The Most Common Places for the Body to Store Excess Fat | Livestrong.” LIVESTRONG.COM, Leaf Group, www.livestrong.com/article/446377-the-most-common-places-for-the-body-to-store-excess-fat/. Accessed 8 Dec. 2023.