Predicting Weight Based on Body Dimensions

Henry Cui

April 29, 2022

Contents

Project statement **2**

Data description **2**

Exploratory Data Analysis **2**

Data Analysis **3**

Summary and discussion **9**

References **10**

Code and data **10**

Project statement

The practice of predicting a person's weight based on height alone has a long history. For example, Body Mass Index (BMI) is the ratio of one's body weight (in kg) to height square (in m^2), which shows whether someone is overweight or not. However, the quality of such models has been found to be not very reliable.

Effects of many other body dimension measurements, which could be used to characterize one's body build, have been ignored. This report explores if we can use those measurements of the body, combined with one's height and age, to predict one's weight. In addition, we also try to find out if the model depends on gender or not.

Data description

The data this project uses were collected by two professors at San José State University and the U.S. Naval Postgraduate School in Monterey, California. The data was initially used for the paper "Exploring Relationships in Body Dimensions" in *Journal of Statistics Education.* Most measurements were taken at the two colleges. The additional measurements were performed in some fitness clubs from California by technicians. The data is from 507 physically active individuals, with 247 men and 260 women. Primarily, the individuals in this dataset are in their twenties and early thirties, with some older men and women. Therefore, the model we conclude at last may not be generalized to all age groups.

Exploratory Data Analysis

The original dataset contains 12 body girths and 9 skeletal diameters measurements of each individual. Age, weight, height, and gender information are also collected. To investigate the effect of gender on the model, we also add 23 interaction terms manually, which are binary variable $Gender\times$ 21 body dimension variables plus $Gender\times$ Age and $Gender\times$ $Height$.

- Basic information: Age(years), Weight(kg), Height(cm), Gender(1 - male, 0 - female)
- Skeletal Measurements (in cm):
  - Biacromial: its diameter
  - Biiliac: its diameter
  - Bitrochanteric: its diameter
  - ChestDepth: depth between spine and sternum at nipple level, mid-expiration
  - ChestDiameter: diameter at nipple level, mid-expiration
  - ElbowDiameter: sum of two elbows
  - Wrist: diameter sum of two wrists
  - Knee: diameter sum of two knees
  - Ankle: diameter sum of two ankles
- Girth Measurements (in cm):
  - ShoulderGirth, ChestGirth, WaistGirth, NavelGirth, HipGirth, ThighGirth
  - BicepGirth, ForearmGirth, KneeGirth, CalfMax, AnkleMin, WristMin

Data Analysis

Since we need to build a prediction model, we first randomly choose 456 data, which is about 90% of the whole dataset, as our train data. We use the trained statistical model to predict the remaining 10% data at the end.

Our training data contain 233 women and 223 men. First, we assume that the error terms are normally distributed and run a relatively simple MLR which predicts weight only using height, gender, and interaction term $Gender \times Height$. With the coefficients below and adjusted R^2 0.5554, we can conclude that it is not a good idea to predict weight only based on gender and height (The interaction terms with *Gender* are all marked by the variable name with a period at the end).

```
## (Intercept)      Height      Gender     Height.
## -41.2643857   0.6199079 -17.2741994   0.1491019
```

We then run MLR with all variables except weight as our explanatory variables. This time, we notice that the adjusted R^2 is 0.9763, which means most response variable data can be explained by this model. However, this model is too complex as it contains more than 40 predictors, and the absolute values of most coefficients are less than 0.5. Only values of the intercept, *Knee*, *Gender* and, *Gender* × *Ankle* are greater than 0.5.

```
## (Intercept)        Knee      Gender      Ankle.
##           1           9          25          34
```

In addition, we also find that variance inflation factors (VIF) for most variables are greater than 10. Only 7 variables have VIF less than 10. Therefore, multicollinearity exists in this model. Most variables are highly correlated with each other. The value of AIC for this model is 1990, and the value of BIC is 2192. Thus, though this model has a very high adjusted R^2, we need to do variable selection to have a more concise MLR model and avoid collinearity.

```
##         Biiliac Bitrochanteric      ChestDepth       KneeGirth        AnkleMin
##               2              3               4              18              20
##             Age         Height
##              22             23
```

```
## [1] 1990.246
```

```
## [1] 2192.248
```

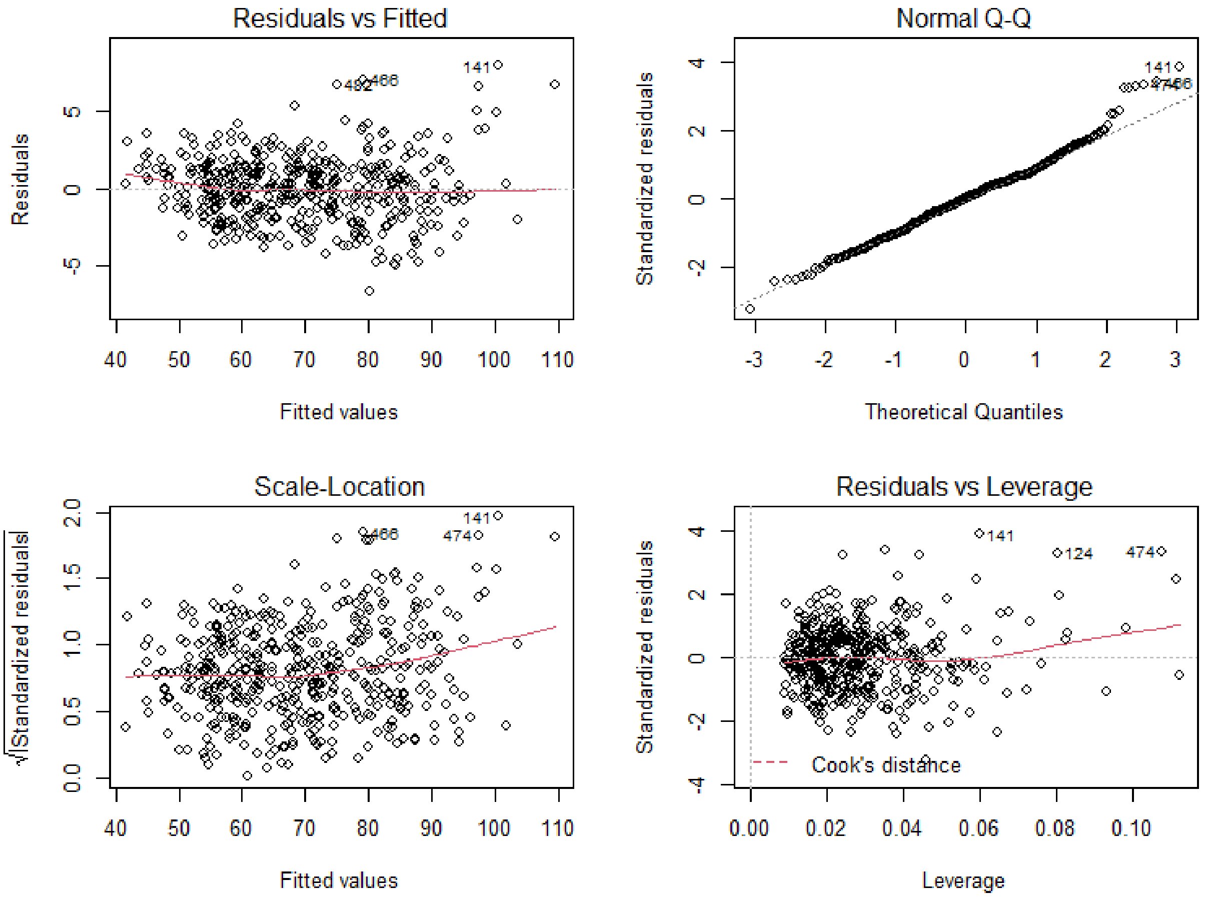
The variable selection methods we use are forward stepwise selection and backwards stepwise selection. We do not use Best Subset Selection method because it is typically only feasible when the number of predictors is fewer than 40. For the two methods we use, we apply both Akaike's Information Criterion and Bayesian Information Criterion. Therefore, we generate four different variable selection strategies.

With forward stepwise selection method and AIC as the criterion, the value of AIC after selection is 652.33 With BIC as our criterion, the value of BIC after selection is 718.68. With backward stepwise selection method and AIC as criterion, the value of AIC after selection is 652.33. While with BIC as our criterion, the final value of BIC is 718.31.

Among the 47 explanatory variables, forward stepwise selection with AIC method chooses 17 variables. Backward stepwise selection with AIC method chooses exactly the same 17 variables. If we use BIC criterion, the models tend to choose fewer variables and have higher BIC value. Forward selection chooses 14 variables, while backward selection chooses 13 variables.

Among the four variable selection strategies, all of them select the interaction term $Height \times Gender$. Therefore, the ideal model should contain interaction terms with gender. A smaller value of AIC or BIC does not mean we find the best model, because both forward and backward stepwise selection processes are greedy algorithms and do not search over all possible models. Checking the value of VIF for each variable selection, we notice that besides the interaction terms which are reasonable to have high VIF values, VIF of variable $ChestGirth$ has a value of 11.23 for the models with AIC criterion, which signals that multicollinearity still exists for such models. Therefore, we choose the variable selection strategy with the smaller BIC value and fewer variables, which contains 13 variables, and remove variable $ChestGirth$ because it has VIF 10.97. The model we now choose has 12 explanatory variables, and all non-interactive terms have VIF values lower than 10:

$Weight = \beta_0 + \beta_1(ChestDepth) + \beta_2(Knee) + \beta_3(WaistGirth) + \beta_4(HipGirth) + \beta_5(ThighGirth) + \beta_6(ForearmGirth) + \beta_7(CalfMax) + \beta_8(Age) + \beta_9(Height) + \beta_{10}(Gender) + \beta_{11}(ShoulderGirth \times Gender) + \beta_{12}(Height \times Gender)$



We then check the model diagnostic from the four plots above. The Residual vs. Fitted plot does not show any clear pattern, which means the linearity assumption holds for this MLR model. Points on the Q-Q plot follow roughly a diagonal line, so we can conclude that the errors are normally distributed. The Scale-Location plot has no pattern, which means variance is constant for the error term.

After checking the assumptions, we run this MLR with the chosen 12 variables. Adjusted R^2 is 0.9745, which is close to the value when we include all the 47 variables. In addition, the 12 variables we use for MLR are included in all the four variable selection method outputs. The predictive model for females is:

$Weight = -106.54 + 0.28(ChestDepth) + 0.76(Knee) + 0.46(WaistGirth) + 0.22(HipGirth) + 0.26(ThighGirth) + 0.83(ForearmGirth) + 0.40(CalfMax) - 0.05(Age) + 0.29(Height) -$

$32.60(Gender)$

The predictive model for males is:

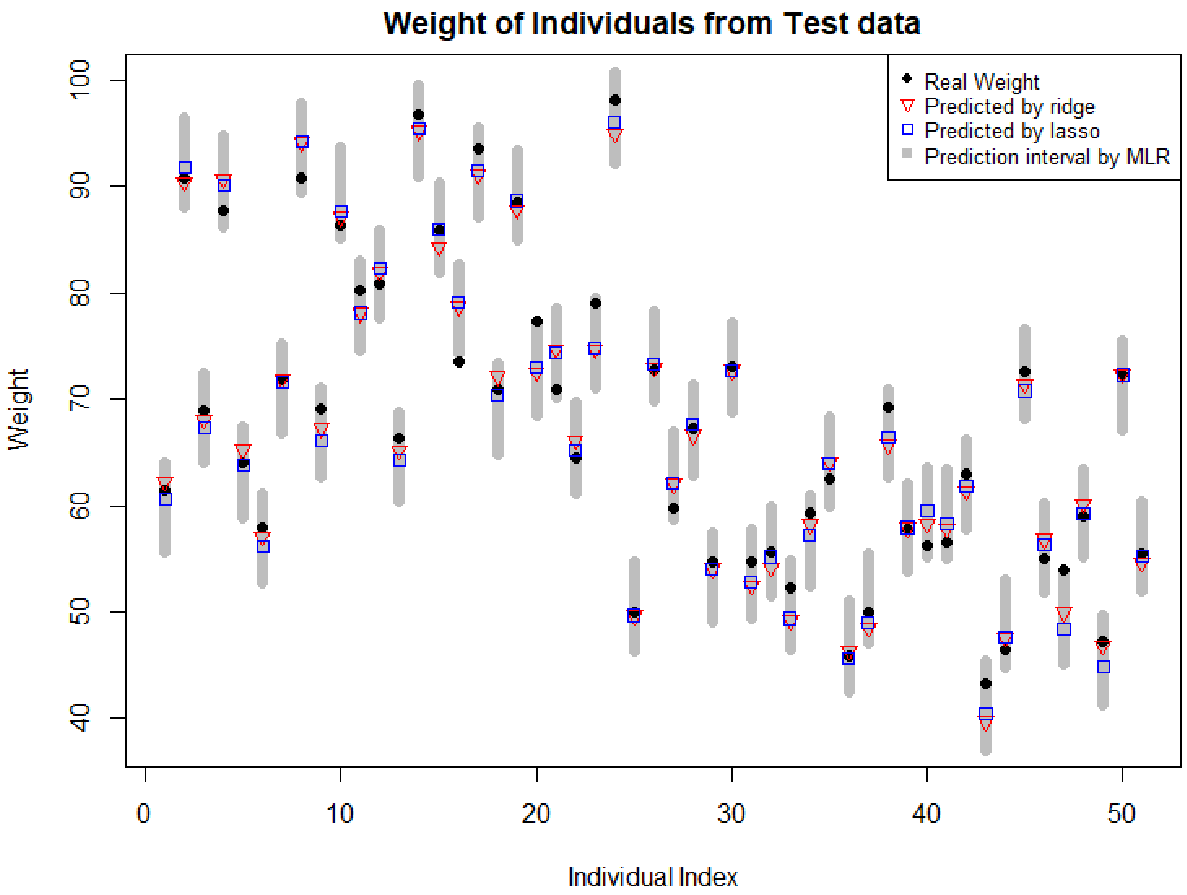
$Weight = -106.54 + 0.28(ChestDepth) + 0.76(Knee) + 0.46(WaistGirth) + 0.22(HipGirth) + 0.26(ThighGirth) + 0.83(ForearmGirth) + 0.40(CalfMax) - 0.05(Age) + 0.35(Height) - 32.60(Gender) + 0.19(ShoulderGirth)$

We then run both ridge regression and lasso regression, and compare different models' prediction for the test data. Lasso regression usually does well when there are a small number of significant parameters and others close to zero, which is the case we show after running MLR. However, it performs poorly when we have multicollinearity, which is also the case we have when we use all 47 predictors according to the value of VIF. Therefore, we need to test both lasso and ridge regression.

For both lasso and ridge regression, we use the package glmnet from R. We utilize cross-validation to find appropriate λ for each regression model. After running cross-validation, we choose the λ with the smallest value of MSE_λ respectively.

We first run ridge regression with all predictors and find that all 47 variables are included in the model as we expect. This result supports our claim that interaction terms with gender should be included in our predictive model, and the model should be different based on the individual's gender. Then we run lasso regression with 47 predictors. This time, lasso regression performs variable selection, and 21 variables are removed. Though lasso regression eliminates some predictors, it does not remove all the interaction terms. Therefore, we now have three different models: one from MLR after variable selection process, one from ridge regression, and one from lasso regression.

Our final step is to compare the three different models with our test data. In our test data, there are 27 women and 24 men. Since both ridge and lasso regression introduce bias in exchange for decreasing the variance of the estimates, we predict one single value of weight for each individual in the test data set when applying ridge and lasso model. Using the MLR model, we generate the 95% prediction interval of the weight for each individual.



We observe from the plot that weights predicted by ridge and lasso regression are usually very close to each other. However, among the 51 test individuals, there are 24 times that the difference between the real weight and the weight predicted by ridge regression is greater than 1.5kg. The gap is greater than 3kg for 11 individuals. The biggest gap is from the 16th individual, where the difference between the real value and the predicted value is 5.1kg. For the lasso regression model, the difference between the real weight and the predicted value is greater than 1.5kg for 25 individuals in the test dataset. The difference is greatest for the 16th subject, who has a gap of 5.48kg between the real value and the predicted value. From the plot, we notice that the 95% prediction interval from MLR captures the actual value of weights for 50 out of the 51 cases.

Therefore, compared with lasso and ridge regression, the MLR model provides prediction intervals that are more likely to capture the real weights. In addition, the result from MLR model is much more interpretable, as it has fewer predictors and the coefficients for the predictors can be used to explain their effects on individuals' weights. Hence, our final predictive model for predicting weights based on body dimensions, age, height, and gender is the MLR model with 12 chosen variables.

Summary and discussion

In conclusion, after variable selection with AIC and BIC as criterion, we remove collinearity for the MLR model. Our predictive model is:

$Weight = \beta_0 + \beta_1(ChestDepth) + \beta_2(Knee) + \beta_3(WaistGirth) + \beta_4(HipGirth) + \beta_5(ThighGirth) + \beta_6(ForearmGirth) + \beta_7(CalfMax) + \beta_8(Age) + \beta_9(Height) + \beta_{10}(Gender) + \beta_{11}(ShoulderGirth \times Gender) + \beta_{12}(Height \times Gender)$

It has a much higher adjusted R^2 compared with the predicting method that only uses height and gender. Meanwhile, its adjusted R^2 is very close to the model with all 47 variables that include all the body dimension measurements and interaction terms.

This final MLR predictive model is simple and interpretable, compared with both ridge and lasso regression model. It also performs better than the two penalized approaches when we apply the three models to the test dataset.

The model provides insight into the relationship between an individual's weight and his or her body dimensions. Our model shows that such a relationship should depend on gender. In addition, instead of measuring all 12 body girths and 9 skeletal diameters measurements, we only need to measure 7 body dimensions for our model besides height, age, and gender to predict one's weight.

One limitation of this study is that besides gender, race might also be one factor that influences the predictive model. However, we miss the race information for the individuals.

For future studies, we can test this model with race as one dummy variable and interaction terms *race*× body dimensions. We can also collect more data from individuals of all age groups to see if this predictive model can be generalized.

References

[1] Heinz, Grete, et al. "Exploring relationships in body dimensions." *Journal of Statistics Education* 11.2 (2003).

[2] Sheather, Simon. *A modern approach to regression with R.* Springer Science Business Media, 2009.

Code and data

```
[breaklines=true]
library(knitr)
library(RSQLite)
library(stringr)
library(dplyr)
library(ggplot2)
library(tidyverse)
library(ggmap)
library(car)
library(MASS)
library(leaps)
library(glmnet)

rawdata <- read.csv("bodydata.csv")
```

```
set.seed(1)

rand <- sample.int(507,456)

traindt <- rawdata[rand,]

testdt <- rawdata[-rand,]

table(traindt$Gender)

fit1 <- lm(Weight~Height+Gender+Height.,data=traindt) # first simple regression

summary(fit1)

print(fit1$coefficients)


fit2 <- lm(Weight~.-Weight,data=traindt) # run MLR with all predictors

summary(fit2)

which(abs(fit2$coefficients)>0.5)


y = vif(fit2) # check VIF

which(y<10)

AIC(fit2)

BIC(fit2)


#apply variable selection

fit_null <- lm(Weight~1, data=traindt)

fit_full <- lm(Weight ~.-Weight,data=traindt)


AICfit <- step(fit_null, scope= list(upper = fit_full), direction = c("forward") )

AICfit2 <- step(fit_full, scope= list(upper = fit_null), direction = c("backward") )


AICfit$coefficients #AIC = 652.33
```

```
AICfit2$coefficients #AIC = 652.33


n <-  length(traindt$Weight)
BICfit <- step(fit_null, scope= list(upper = fit_full)
, direction = c("forward")
, k = log(n))
BICfit2 <- step(fit_full, scope= list(upper = fit_null), direction = c("backward"),
k = log(n))


BICfit$coefficients #BIC =718.68
BICfit2$coefficients #BIC=718.31


vif(BICfit2)
vif(BICfit)
vif(AICfit)
vif(AICfit2)


# run MLR after variable selection
fit_mlr1 <-  lm(Weight~ChestDepth+Knee+ChestGirth+WaistGirth+HipGirth
+ThighGirth+ForearmGirth+CalfMax+Age+Height+Gender+
ShoulderGirth.+Height., data=traindt)
summary(fit_mlr1)
vif(fit_mlr1)


fit_mlr2 <-  lm(Weight~ChestDepth+Knee+WaistGirth+
HipGirth+ThighGirth+ForearmGirth+CalfMax+Age+
Height+Gender+ShoulderGirth.+Height., data=traindt)
```

```
summary(fit_mlr2)
vif(fit_mlr2)

plot(fitted(fit_mlr2),resid(fit_mlr2))
par(mfrow=c(2,2)) # the diagnostic plots
plot(fit_mlr2)

# Run lasso and ridge regression
set.seed(1)

newdt <- testdt[-23]

fit_mlr2 <-  lm(Weight~ChestDepth+Knee+WaistGirth+
HipGirth+ThighGirth+ForearmGirth+CalfMax+Age+Height+
Gender+ShoulderGirth.+Height., data=traindt)
set.seed(1)
y <- fit2$model[,1]
X = as.matrix(fit2$model[,-1])
fit.cv = cv.glmnet(X,  # Matrix of predictors (w/o intercept)
                   y,  # Response
                   alpha=0, # Corresponds to the penalty (0 for ridge, 1 for lasso)
)
lamb <- fit.cv$lambda.min
fit_ridge = glmnet(X,  # Matrix of predictors (w/o intercept)
              y,  # Response
              alpha=0, # Corresponds to the penalty (0 for ridge, 1 for lasso)
              lambda = lamb # lambda sequence
              )
y <- fit2$model[,1]
X = as.matrix(fit2$model[,-1])

fit.cv2 = cv.glmnet(X,  # Matrix of predictors (w/o intercept)
                   y,  # Response
                   alpha=1, # Corresponds to the penalty (0 for ridge, 1 for lasso)
)

lamb2 <- fit.cv2$lambda.min
fit_lasso = glmnet(X,  # Matrix of predictors (w/o intercept)
              y,  # Response
              alpha=1, # Corresponds to the penalty (0 for ridge, 1 for lasso)
              lambda = lamb2 # lambda sequence
              )
full_ridge_predict <- predict.glmnet(fit_ridge,
newx = as.matrix(newdt),interval = "prediction")
full_lasso_predict <- predict.glmnet(fit_lasso,
newx = as.matrix(newdt),interval="prediction")
real_weight <- testdt$Weight

# run the model with test data

mlr_predict <- predict(fit_mlr2,newdata = newdt,interval="prediction",level = 0.95)

plot(c(1:51),real_weight,xlab="Individual Index",ylab="Weight",
main="Weight of Individuals from Test data",pch=19,ylim=c(38,100))
```

```
points(c(1:51),full_ridge_predict,pch = 6,col="red")
points(c(1:51),full_lasso_predict,pch = 0,col="blue")

for (j in 1:51){
segments(x0 = j, y0 = mlr_predict[j,2], x1=j,y1=mlr_predict[j,3],col = "grey")
}

r <- matrix(nrow=51,ncol=3)
r[,1]=full_ridge_predict
r[,2]=full_lasso_predict
r[,3]=real_weight
testdiff1 <- r[,1]-r[,3]
which(testdiff1 == max(testdiff1))
max(testdiff1)
which(abs(testdiff1)>1.5)
which(abs(testdiff1)>3)

testdiff2 <- r[,2]-r[,3]
which(testdiff2 == max(testdiff2))
max(testdiff2)
which(abs(testdiff2)>1.5)
which(abs(testdiff2)>3)
```