Robert Dam

Identifying more accessible body fat percentage estimation procedures with regression analysis

Abstract

The current standard procedures for body fat percentage estimation, hydrostatic weighing and DXA scans, are very accurate but costly procedures that require specialized equipment. It is in the interests of the general public to know approximately their body fat percentage because high body fat percentage contributes to health risks like heart disease and more. So instead, we look for a way to predict body fat percentage on easily obtained body measurements like weight, age, height, and numerous body part circumferences obtained on a sample of 252 men. Questions of interest include which set of measurements provides an accurate and stable model for predicting body fat percentage? Can this method replace the standard methods or will it be used more as a screening measure? How effective is this model at out-of-sample prediction?

Comparing first-order linear regression models using an exhaustive best subset selection procedure and comparing interactive models with a forward stepwise procedure we arrive at a model with only two predictor variables: weight and abdomen circumference. In general, we found that it was hard to have models with many predictor variables since the body part circumferences had high pairwise correlations which made multicollinearity an issue.

This final model performs strongly to approximate body fat percentage, but it is not precise enough to give a definitive value. The adjusted coefficient of determination of this final model on all of the data is 0.7102 which means about 71% of the variability in body fat percentage is explained by variation in weight and abdominal circumference. Evaluating predictive performance on the validation set, we find that the MSPE is 22.69 compared to an SSE/n of 18.39 which indicates decent predictive performance. Do note that this model was only built on data from men so it should not be used on women and children.

We recommend this model for men to get an approximate idea of where their body fat percentage is at. If it is in a high range then one can take measures to reduce it. If one needs a very accurate measurement of body fat percentage we still recommend use of the standard methods of hydrostatic weighing and DXA scans.

Introduction

Excess weight contributes to a number of health risks including diabetes, high blood pressure, heart disease, stroke, cancer, and more (NIH). However, even at healthy weights, having high body fat percentages leads to similar health risks as being overweight or obese (Kim et al., 2013). This means that body fat percentage is another important measure to assess a person's health. However, body fat percentage is harder to measure than other metrics of health since it requires knowledge of total body composition. The current most accurate procedures to estimate body fat percentage are hydrostatic weighing and DXA scans. Hydrostatic weighing estimates body density by comparing dry and underwater body weights (Katch and McArdle,

1977). DXA scans conduct imaging of your body with x-rays to provide body composition measures (UC Davis). However, both methods are expensive and require specialized equipment.

So we would like to explore possible easier and cheaper methods that one could use to get a quick idea of approximately what their body fat percentage is. We will be investigating whether more accessible measurements can be used to estimate body fat percentage. Specifically, we will be working with a dataset that contains the age, weight, height, body density, body fat percentage estimate from Siri's (1956) equation, and numerous body part circumference measurements of 252 men. Note that these are only quantitative variables. Questions of interest include:

- Which set of measurements above provide the most stable and accurate estimation of body fat percentage?
- Is this estimate accurate enough for precise body fat estimation or is it more of a rough estimate?
- Does this model generalize well to out of sample cases i.e. how effective is this model for prediction?

This problem is of interest because if we are able to identify an accessible body fat percentage model, then a widespread amount of people will be able to get a cheap and convenient way to monitor an aspect of their health. This can help with identifying possible health risks and lead to preventative measures.

Methods and Results

We will be working with least-squares linear regression to examine the relationship between weight, age, height, and numerous body part circumferences with body fat. We first performed some exploratory analyses to get an understanding of what our data is like and to assess the suitability of a linear model. From marginal plots, we found that the response variable body fat percentage is slightly right skewed but mostly symmetric (Figure 1). Subjects in our sample have a mean body fat percentage of 19.1%. As for the other variables, most of them are similarly mostly symmetric with small skews, mostly right skews. From these plots we noticed some outliers which we were able to remove since they seemed to be due to input errors. For example, some body fat percentages were less than 1% and one person had a height of 29.5 inches which was less than his waistline. We additionally split the dataset by a 70%/30% split into a training set and a validation set. We initially performed modelling work with the training set and used the validation set to test predictive performance.

From scatterplot and correlation matrices (Figure 2) we see that body fat percentage has a pretty strong linear relationship with many body part circumferences. This justifies the use of a linear regression model. Unsurprisingly, the body part circumferences have strong correlations between each other. This leads to the concern of multicollinearity which we must take into account.

As an additional exploratory step, we fit a linear model with all predictor variables. This will help us get a rough idea of how the predictor variables explain body fat percentage. From an initial naive fit, we get a pretty strong R^2 value of 0.7404 and R_a^2 value of 0.7261. This suggests that this approach could be fruitful. However, the full model unsurprisingly overfits and has issues with strong multicollinearity. We can observe this from the fact that only 3 of the 13 coefficients are significant. We also see that 4 variables in this model have a VIF over 10, and body fat has a huge VIF of 47.09 (R Output 1).

So we would instead like to find a subset of variables that has high accuracy and predictive performance but avoids these issues. If we can find one with lower variance and multicollinearity we will have more stable predictions. Since in the first-order regression model we only have 13 predictor variables we are able to do an exhaustive search of every subset of models to compare their performance. For the first-order regression model, we performed a best subset selection procedure to select this set of variables. From this procedure, we arrived at a few candidate models by calculating several model selection criteria (Figure 3). By BIC, we had a candidate model (Candidate 1) with two variables: weight and abdomen circumference. By C_p and AIC we had a candidate model (Candidate 2) with five variables: age, height, chest, abdomen, and wrist circumference. By R_a^2 we had a candidate model (Candidate 3) with eight variables: age, height, chest, abdomen, wrist, ankle, hip, and forearm circumference.

We also considered possible interaction models since it could be plausible that different levels of weight, height, or age could make body part circumferences interact differently with body fat percentage. Since the number of total two-way interactions between all predictor variables is large, best subset selection becomes unfeasible so we instead performed a forward stepwise procedure by the AIC criterion. From this we arrived at a model, Candidate 4, with five variables: weight, height, abdomen-weight interaction, wrist and abdomen circumference.

By these various criteria these models perform similarly so we also calculated the VIF of each variable in each candidate model, examined regression summary output, and calculated the mean-squared prediction error of each model on the validation set. From the regression summary output, we see that in Candidate 2 chest and age are not significant at level 0.05. In Candidate 3 age, chest, hip, ankle, and forearm are not significant at level 0.05. In Candidate 4 weight and the abdomen-weight interaction term are not significant at level 0.05. From examining the VIF, we also see that in Candidate 3 abdomen circumference has a VIF of 10.53 which indicates multicollinearity. The lack of significance in a number of variables in these three models indicates possible high variability and high multicollinearity. Candidate 1 did not suffer from these problems, but did have a C_p value decently higher than its number of predictors which indicates the possibility of some underfitting. As for predictive performance, we can see that Candidate 1 and Candidate 4 have an MSPE that is closer to SSE/n than the other two models which indicates that they may have less of an issue with overfitting. Candidate 4 has the lowest MSPE. However, all four models still have pretty similar MSPE which suggests that they have

similar predictive performance. See Figure 4 for a table of SSE/n compared to MSPE for the three models.

Due to these considerations we decided to choose Candidate 1, the model with two predictors: weight and abdominal circumference. The other variables definitely do have explanatory value in predicting body fat percentage, but due high pairwise correlations between these circumference values that leads to high multicollinearity, including many of these values together creates high variance models. This high variance makes it hard to determine how stable our fitted coefficients are which then makes generalized out-of-sample prediction less precise. Just two variables: weight and abdominal circumference already capture much of the information needed to estimate body fat percentage. It is not too surprising that abdominal circumference has value for predicting body fat percentage since men hold much of their body fat in that area of the body. We can see this from Figure 3 where within the test set, Candidate 1 already has a R_a^2 value of 0.74 compared to values of 0.75 for more complex models.

For our final model, we fit a linear regression of body fat with weight and abdomen circumference using all of the data. This model has all the assumptions of the least-squares multiple linear regression model which are as follows:

1. Body fat percentage has a linear relationship with weight and abdomen circumference.
2. Errors from this relationship are normally distributed with 0 mean and constant variance.
3. These errors are uncorrelated with each other.

Referring to R Output 2, the regression summary output of this model, we see that the final model has a R^2 value of 0.7125 and a R_a^2 value of 0.7102. This can be interpreted as 71% of the variation in body fat can be explained by variation in weight and abdominal circumference. We can compare this to the full model and see that we lose little in R^2 value, but make up with a model with less variability. The fitted regression coefficients for this model are 0.9829 for abdomen circumference (in centimeters), -0.1504 for weight (in pounds), and -44.879 for the intercept. This gives a regression equation of :

$$B = -44.879 + 0.9829A - 0.1504W$$

B = Body fat percentage, A = Abdomen circumference in centimeters,
W = Weight in pounds

We can interpret the coefficients as changes in body fat percentage with a unit change in each variable holding the other variable constant. So for abdomen circumference, this would mean that holding weight constant, increasing abdomen circumference by 1 centimeter increases body fat percentage by 0.9829. Likewise for weight, this is interpreted as holding abdomen circumference constant, for every 1 pound increase in weight we decrease body fat percentage by 0.1504. A negative coefficient for weight initially looks counter-intuitive, but when we consider

that it is controlled by abdomen circumference it makes more sense. Having a larger waistline typically means one has a higher body fat percentage, but at a constant waistline level, having more weight then correlates with height and muscle mass which would lower body fat percentage.

The t-values testing for non-zero values for both slope coefficients are very significant at 17.45 for abdomen circumference and -7.28 for weight; both have near 0 p-values. The F-statistic which tests for general regression relation is also very significant with a value of 304.8 which also has a near 0 p-value.

Performing diagnostics (Figure 5) on this chosen model, we see from the residuals vs fitted values plot that there is no clear nonlinearity or heteroscedasticity in the residuals. This is consistent with the assumptions of linearity and constant variance. From the residuals normal QQ plot, we see that most of the data follows the line, but there the tails of the residuals distribution are light. This is not strong evidence of non-normality, but we should take caution with applying our predictions to more extreme values. Looking at the residuals vs leverage plot, we see that there is one influential case. To see if we should remove this case or not, we compute the percent change of the fitted values when we include vs exclude this case. On average, this percent change is 1.46% with a range of 0.0025% to 9.57%. This indicates that this data point does not heavily influence prediction so we do not need to remove it. Finally, looking at a Box-Cox procedure, we verify that body fat percentage varies with weight and abdomen circumference linearly so no further transformations are needed.

Conclusion

We are able to arrive at simple lower variance model that performs nearly as well or better than more complicated models in the metrics of R_a^2 and MSPE. Just two measurements, weight and abdominal circumference provide a reasonable linear model to predict body fat percentage. At an R_a^2 of 0.7102 on the full data, this model explains a large amount of the variability in body fat, but it is not a precise model. Instead, it should be used as an initial estimate to gauge around what level of body fat a person is at. From the exhaustive search, we found that the highest performing subsets of variables all ended up with a similar coefficient of determination. This suggests that using body part circumferences has its limits on how accurately it can estimate body fat percentage. This limitation along with the multicollinearity that comes with the high pairwise correlations between measurements lowers the accuracy of this analysis. We could possibly explore other accessible biometrics and see if they lead to a more accurate model if we want to try to further improve the accuracy of this method. We would also need more data to confidently prescribe more complicated models since more data would help with sampling variability.

As we saw in Figure 4, this model has a MSPE not much higher than SSE/n which means that it has reasonable predictive performance. It does have a C_p value decently higher than the number of predictors which could indicate underfitting, but with more variables we ran into

the issue of multicollinearity and non-significant coefficients so we decided on this compromise instead. Since this simpler model avoids the issues of multicollinearity and higher variance its predictions should be more stable. However, there are limitations with using this model for prediction. One limitation of this model is that there is less data in the extreme regions of body fat percentage. This makes prediction in these regions a less sure thing since we get into some extrapolation territory. We can address this by collecting more data points in the extreme body fat percentage regions. Another limitation is that this sample consists only of adult men which means that this model is not appropriate to predict the body fat percentage of women and children. We need to collect data from women and children, and would likely need a completely different model of body fat percentage for these two groups.

Overall, we recommend this model as a reasonable initial screening to get an idea of a person's body fat percentage. If one has body fat percentage in a high range from this screening they could take preventative measures to lower that value in order to avoid the health risks. In the cases when one needs a precise value, the methods of hydrostatic weighing and DXA scans are still needed.

Appendix 1: Figures and Tables

Figure 1: Histogram of body fat percentage

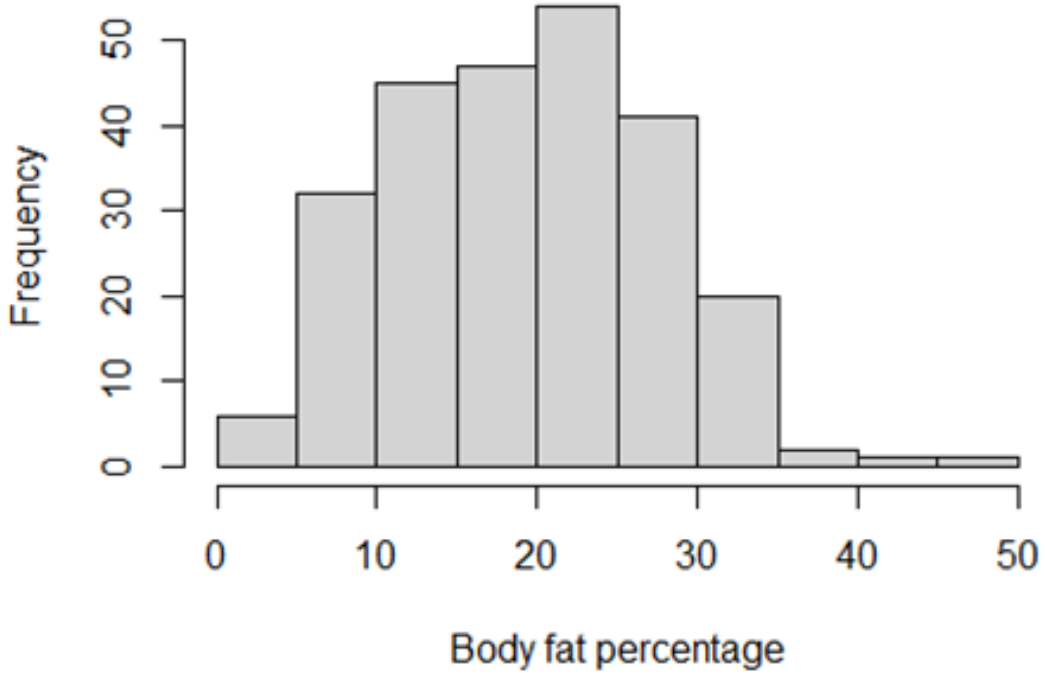


Figure 2: Scatterplot and correlation matrix

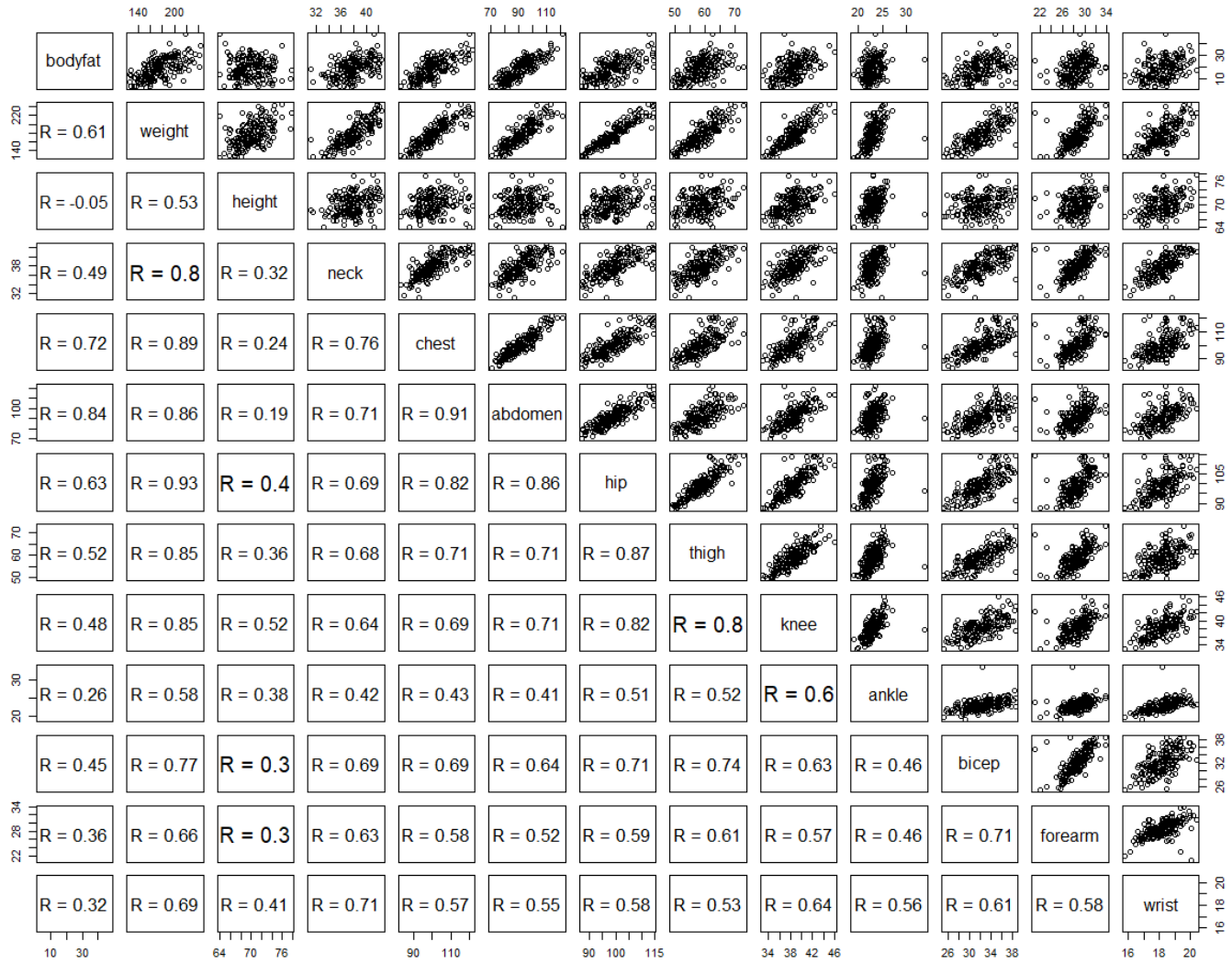


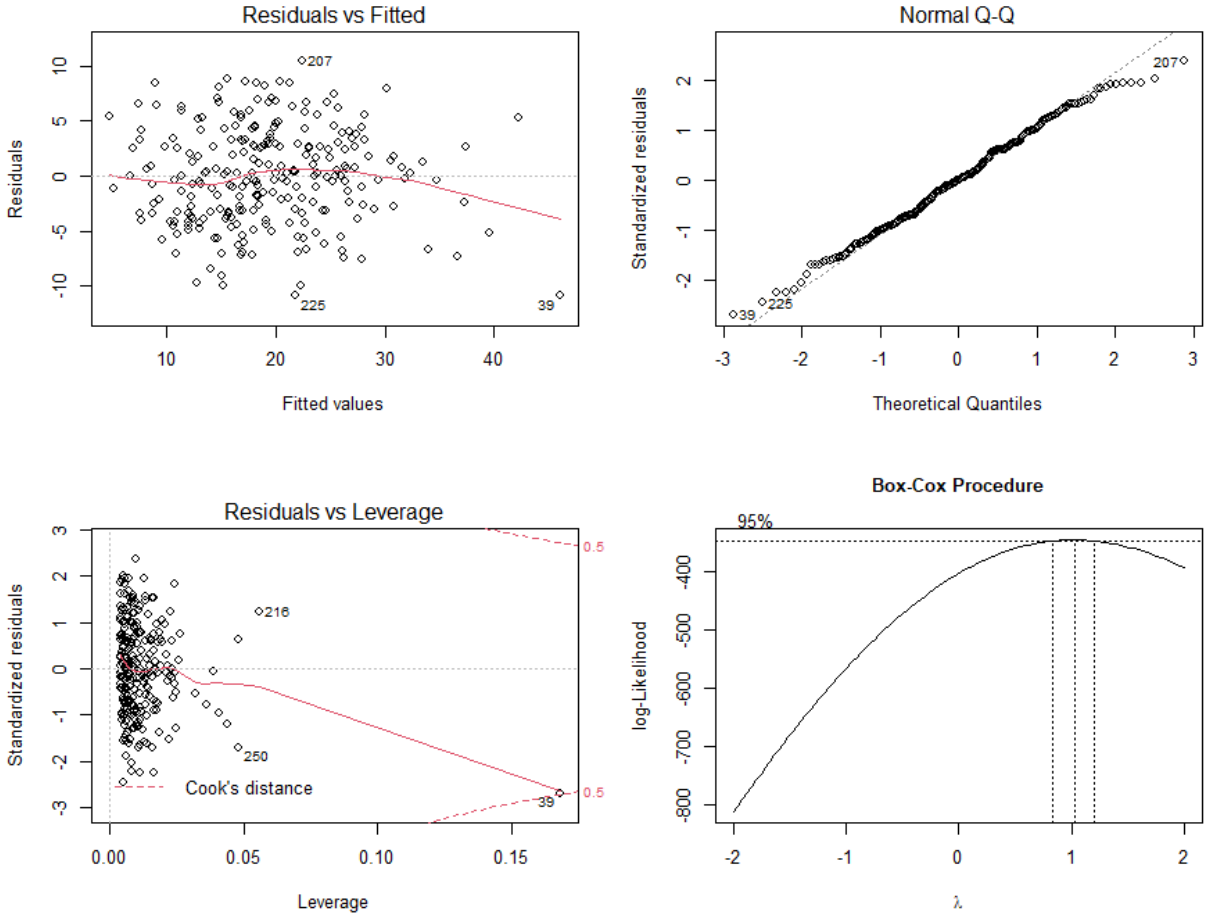
Figure 3: Candidate models from best subset selection and forward stepwise procedure

	Variables Selected	R_a^2	C_p	BIC	AIC
Candidate 1	Weight, abdomen	0.74	7.56	571.51	562.03
Candidate 2	age, height, chest, abdomen, wrist	0.75	3.58	576.80	557.85
Candidate 3	age, height, chest, abdomen, wrist, ankle, hip, forearm	0.75	6.26	588.75	560.32
Candidate 4	weight, height, abdomen-weight interaction, wrist, abdomen	0.75	9.97	577.20	558.25

Figure 4: MSPE vs SSE/n for the four candidate models

	SSE/n	**MSPE**
Candidate 1	22.69	18.39
Candidate 2	24.99	17.35
Candidate 3	23.24	17.00
Candidate 4	18.12	17.39

Figure 5: Model diagnostics for final model



Appendix 2: R Code and Output

R Output 1: Full model output summary

```
fit_all <- lm(bodyfat ~ . - density, data = bodyfat)

summary(fit_all)

##
## Call:
## lm(formula = bodyfat ~ . - density, data = bodyfat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.2788  -2.9475  -0.1618   3.1777   9.8986
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -15.32662   22.37419  -0.685  0.49401
## age           0.05625    0.03238   1.737  0.08366 .
## weight       -0.08619    0.06209  -1.388  0.16636
## height       -0.09211    0.17948  -0.513  0.60831
## neck         -0.46028    0.23522  -1.957  0.05155 .
## chest        -0.02214    0.10408  -0.213  0.83175
## abdomen       0.94972    0.09015  10.534  < 2e-16 ***
## hip          -0.19892    0.14630  -1.360  0.17523
## thigh         0.20623    0.14727   1.400  0.16273
## knee          0.02363    0.24717   0.096  0.92390
## ankle         0.17531    0.22202   0.790  0.43055
## bicep         0.15933    0.17326   0.920  0.35871
## forearm       0.44034    0.19894   2.213  0.02783 *
## wrist        -1.61265    0.53372  -3.021  0.00279 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.292 on 235 degrees of freedom
## Multiple R-squared:  0.7404, Adjusted R-squared:  0.7261
## F-statistic: 51.56 on 13 and 235 DF,  p-value: < 2.2e-16

car::vif(fit_all)

##       age    weight    height      neck     chest   abdomen       hip
thigh
##  2.261079 43.852021  2.937227  4.346421 10.139520 12.439730 14.290792
7.715646
##      knee     ankle     bicep   forearm     wrist
##  4.656591  1.902153  3.604014  2.149158  3.297124
```

R Output 2: Final model output summary

```
final_model <- lm(bodyfat ~ weight + abdomen, data=bodyfat)
summary(final_model)

##
## Call:
## lm(formula = bodyfat ~ weight + abdomen, data = bodyfat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.875  -3.252  -0.001   3.175  10.450
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -44.87931    2.61661 -17.152  < 2e-16 ***
## weight       -0.15037    0.02065  -7.281 4.47e-12 ***
## abdomen       0.98285    0.05631  17.455  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.415 on 246 degrees of freedom
## Multiple R-squared:  0.7125, Adjusted R-squared:  0.7102
## F-statistic: 304.8 on 2 and 246 DF,  p-value: < 2.2e-16
```

Raw R Code:

```
#Read in data
bodyfat <- read.table("bodyfat.txt", header=F)
colnames(bodyfat) <- c("density","bodyfat","age","weight","height","neck","chest","abdomen",
                       "hip", "thigh", "knee", "ankle", "bicep", "forearm","wrist")
#Remove height outlier (imputation error)
bodyfat <- bodyfat[-which(bodyfat$height == min(bodyfat$height)),]
bodyfat <- bodyfat[bodyfat$bodyfat > 1,]

#Train and test sets
set.seed(1239)
train_index <- sort(sample(1:nrow(bodyfat),as.integer(nrow(bodyfat)*0.7),replace=F))
train <- bodyfat[train_index,]
test <- bodyfat[-train_index,]

#Exploratory plots
hist(bodyfat$bodyfat, main="Figure 1: Histogram of body fat percentage", xlab = "Body fat percentage")
hist(bodyfat$age)
hist(bodyfat$height)
hist(bodyfat$weight)
hist(bodyfat$neck)
hist(bodyfat$chest)
hist(bodyfat$abdomen)
hist(bodyfat$hip)
hist(bodyfat$thigh)
hist(bodyfat$knee)
hist(bodyfat$ankle)
hist(bodyfat$bicep)
hist(bodyfat$forearm)
hist(bodyfat$wrist)

sapply(bodyfat, summary)

panel_cor <- function(x,y) {
  par(usr = c(0, 1, 0, 1))
  r <- round(cor(x, y, use = "complete.obs"), 2)
  txt <- paste0("R = ", r)
  cex_cor <- 0.8/strwidth(txt)
  text(0.5, 0.5, txt, cex = cex_cor)
}

pairs(train[,c("bodyfat", "weight","height","neck","chest","abdomen",
                       "hip", "thigh", "knee", "ankle", "bicep", "forearm","wrist")],
            lower.panel = panel_cor)

#Exploratory full model fit
fit_all <- lm(bodyfat ~ . - density, data = bodyfat)

summary(fit_all)

plot(fit_all)

car::vif(fit_all)
```

```
MASS::boxcox(fit_all)

#Best subset selection procedure
best_set <- leaps::regsubsets(bodyfat ~ . - density, data=train, nbest=1, nvmax=13, method="exhaustive")
bestsub <- summary(best_set)

n <- nrow(train)
p_m <- rowSums(bestsub$which)
ssto <- sum((bodyfat$bodyfat - mean(bodyfat$bodyfat))^2)
sse <- (1-bestsub$rsq)*ssto
aic <- n*log(sse/n)+2*p_m
bic <- n*log(sse/n)+log(n)*p_m

model_comp <- cbind(bestsub$which, sse, bestsub$rsq, bestsub$adjr2, bestsub$cp, bic, aic)
colnames(model_comp) <- c(colnames(bestsub$which), "sse","R^2","R^2_a","Cp","bic", "aic")

round(model_comp, 2)

#Candidate models
candidate1 <- lm(bodyfat ~ weight + abdomen, data=train)
candidate2 <- lm(bodyfat ~ age + height + chest + abdomen + wrist, data=train)
candidate3 <- lm(bodyfat ~ age + height + chest + abdomen + wrist + hip + ankle + forearm, data=train)

#Evaluating candidate models
summary(candidate1)
summary(candidate2)
summary(candidate3)
plot(candidate1)
plot(candidate2)
plot(candidate3)
anova(candidate1)
anova(candidate2)
anova(candidate3)
car::vif(candidate1)
car::vif(candidate2)
car::vif(candidate3)
MASS::boxcox(candidate1)
MASS::boxcox(candidate2)
MASS::boxcox(candidate3)

#MSPE calculation
test_fit1 <- lm(bodyfat ~ weight + abdomen, data=test)
summary(test_fit1)
mean((test$bodyfat - predict(candidate1, test[,-c(1,2)]))^2)
sum(candidate1$residuals^2)/nrow(train)

test_fit2 <- lm(bodyfat ~ age + height + chest + abdomen + wrist, data=test)
summary(test_fit2)
mean((test$bodyfat - predict(candidate2, test[,-c(1,2)]))^2)
sum(candidate2$residuals^2)/nrow(train)

test_fit3 <- lm(bodyfat ~ age + height + chest + abdomen + wrist + hip + ankle + forearm, data=test)
summary(test_fit3)
mean((test$bodyfat - predict(candidate3, test[,-c(1,2)]))^2)
```

```
sum(candidate3$residuals^2)/nrow(train)

#Interaction terms model
full_mod <- lm(bodyfat ~ (. - density)^2, data=train)
none_mod <- lm(bodyfat ~ 1, data=train)

candidate4 <- MASS::stepAIC(none_mod, scope=list(upper=full_mod, lower = ~1), direction="both", k=2, trace=F)
summary(candidate4)
car::vif(candidate4)
anova(candidate4)
plot(candidate4)

n*log((1-summary(candidate4)$r.squared)*ssto/n)+2*6
n*log((1-summary(candidate4)$r.squared)*ssto/n)+log(n)*6
olsrr::ols_mallows_cp(candidate4, full_mod)

#MSPE
test_int <- lm(bodyfat ~ weight + abdomen + height + wrist + weight:abdomen, data=train)
summary(test_int)
mean((test$bodyfat - predict(candidate4, test[,-c(1,2)]))^2)
sum(candidate4$residuals^2)/nrow(train)

#Final model summary and diagnostics
final_model <- lm(bodyfat ~ weight + abdomen, data=bodyfat)
summary(final_model)

par(mfrow=c(2,2))
plot(final_model, which=c(1,2,5))
MASS::boxcox(final_model)
title("Box-Cox Procedure", cex.main=1)
car::vif(final_model)

#Seeing effect of influential point
no_39 <- lm(bodyfat ~ weight + abdomen, data=bodyfat[-39,])
percent_change <- abs((final_model$fitted.values - predict(no_39, bodyfat[,-c(1,2)]))/final_model$fitted.values)*100
summary(percent_change)
plot(final_model$fitted.values, predict(no_39, bodyfat[,-c(1,2)]))
abline(0,1)
```

References

National Institute of Diabetes and Digestive and Kidney Diseases. “Health Risks of Overweight and Obesity”, U.S. Department of Health.

Kim, Ji Young, Han, Sang-Hwan and Yang, Bong-min (2013). “Implication of high-body-fat percentage on cardiometabolic risk in middle-aged, healthy, normal-weight adults”, Obesity, Silver Spring, MD.

UC Davis. “Dual X-ray Absorptiometry for Body Composition”, UC Regents, Sacramento, CA.

Katch, Frank and McArdle, William (1977). “Nutrition, Weight Control, and Exercise”, Houghton Mifflin Co., Boston.