DSA 8020 R Lab 3: Multiple Linear Regression II

Meredith Sliger

Contents

Percentage of Body Fat and Body Measurements **1**
- Load the dataset . . . 1
- Exploratory Data Analysis . . . 2
  - Numerical summary . . . 2
  - Graphical summary . . . 3
- General Linear F-Test . . . 5
- Prediction . . . 7
- Multicollinearity . . . 8

Percentage of Body Fat and Body Measurements

Age, weight, height, and 10 body circumference measurements are recorded for 252 men. Each man's percentage of body fat was accurately estimated by an underwater weighing technique.

Data Source: Johnson R. *Journal of Statistics Education* v.4, n.1 (1996)

Load the dataset

Code:

```
library(faraway)
```

```
## Warning: package 'faraway' was built under R version 4.4.2
```

```
data(fat)
head(fat)
```

```
##   brozek siri density age weight height adipos  free neck chest abdom   hip
## 1   12.6 12.3  1.0708  23 154.25  67.75   23.7 134.9 36.2  93.1  85.2  94.5
## 2    6.9  6.1  1.0853  22 173.25  72.25   23.4 161.3 38.5  93.6  83.0  98.7
## 3   24.6 25.3  1.0414  22 154.00  66.25   24.7 116.0 34.0  95.8  87.9  99.2
## 4   10.9 10.4  1.0751  26 184.75  72.25   24.9 164.7 37.4 101.8  86.4 101.2
## 5   27.8 28.7  1.0340  24 184.25  71.25   25.6 133.1 34.4  97.3 100.0 101.9
## 6   20.6 20.9  1.0502  24 210.25  74.75   26.5 167.0 39.0 104.5  94.4 107.8
##   thigh knee ankle biceps forearm wrist
```

```
## 1  59.0 37.3  21.9   32.0    27.4  17.1
## 2  58.7 37.3  23.4   30.5    28.9  18.2
## 3  59.6 38.9  24.0   28.8    25.2  16.6
## 4  60.1 37.3  22.8   32.4    29.4  18.2
## 5  63.2 42.2  24.0   32.2    27.7  17.7
## 6  66.0 42.0  25.6   35.7    30.6  18.8
```

For the purposes of this lab, we will use only the following variables for conducting data analysis:

1. y `brozek`: Percent body fat using Brozek's equation

$$\frac{457}{\text{Density}} - 414.2$$

2. x_1 `age`: Age (yrs);
3. x_2`weight`: Height (inches);
4. x_3 `height`: Height (inches);
5. x_4`chest`: Chest circumference (cm);
6. x_5 `abdom`: Abdomen circumference (cm) at the umbilicus and level with the iliac crest

Code:

You can use the code below to extract these variables:

```r
vars <- c("brozek", "age", "weight", "height", "chest", "abdom")
data <- fat[, vars]
```

Exploratory Data Analysis

Numerical summary

1. Use `summary` command to produce various numerical summaries of each of the 6 variables under consideration

Code:

```r
str(data)
```

```
## 'data.frame':    252 obs. of  6 variables:
##  $ brozek: num  12.6 6.9 24.6 10.9 27.8 20.6 19 12.8 5.1 12 ...
##  $ age   : int  23 22 22 26 24 24 26 25 25 23 ...
##  $ weight: num  154 173 154 185 184 ...
##  $ height: num  67.8 72.2 66.2 72.2 71.2 ...
##  $ chest : num  93.1 93.6 95.8 101.8 97.3 ...
##  $ abdom : num  85.2 83 87.9 86.4 100 94.4 90.7 88.5 82.5 88.6 ...
```

```r
summary(data)
```

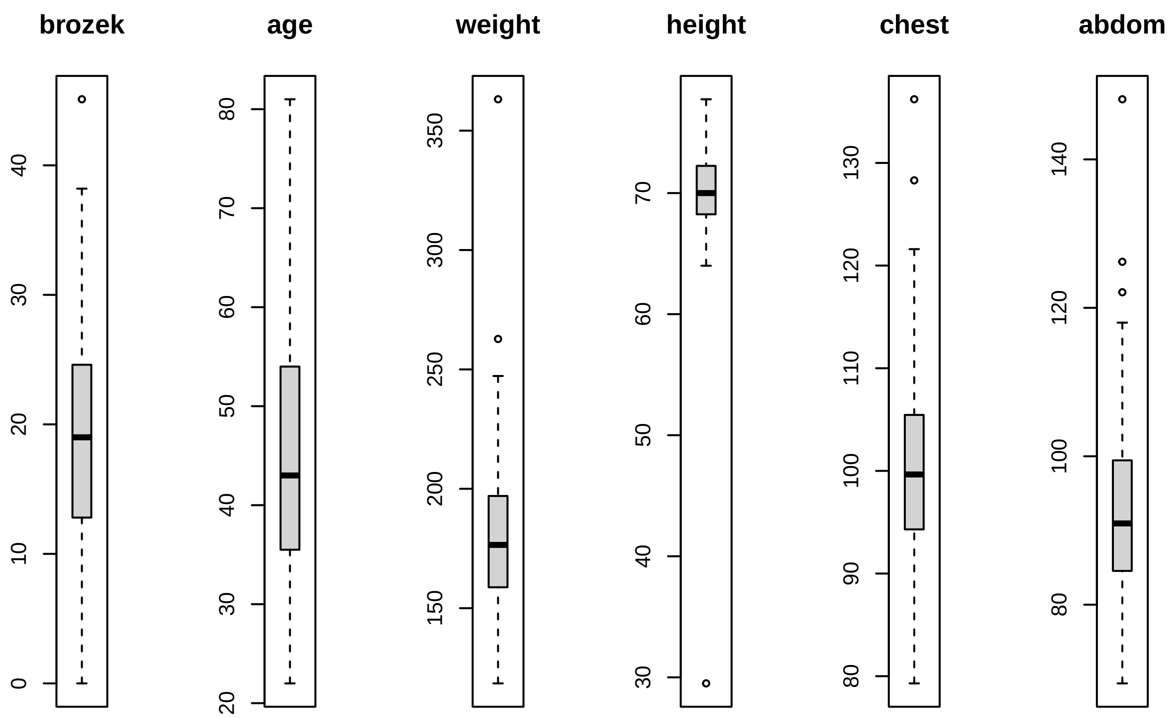
```
##      brozek           age            weight          height
##  Min.   : 0.00   Min.   :22.00   Min.   :118.5   Min.   :29.50
##  1st Qu.:12.80   1st Qu.:35.75   1st Qu.:159.0   1st Qu.:68.25
##  Median :19.00   Median :43.00   Median :176.5   Median :70.00
##  Mean   :18.94   Mean   :44.88   Mean   :178.9   Mean   :70.15
##  3rd Qu.:24.60   3rd Qu.:54.00   3rd Qu.:197.0   3rd Qu.:72.25
##  Max.   :45.10   Max.   :81.00   Max.   :363.1   Max.   :77.75
##      chest            abdom
##  Min.   : 79.30   Min.   : 69.40
##  1st Qu.: 94.35   1st Qu.: 84.58
##  Median : 99.65   Median : 90.95
##  Mean   :100.82   Mean   : 92.56
##  3rd Qu.:105.38   3rd Qu.: 99.33
##  Max.   :136.20   Max.   :148.10
```

Graphical summary

2. Make a boxplot for each variable

Code:

```r
par(mfrow=c(1,6)) # Arrange plots in a 1x6 grid
boxplot(data$brozek, main = vars[1])
boxplot(data$age, main = vars[2])
boxplot(data$weight, main = vars[3])
boxplot(data$height, main = vars[4])
boxplot(data$chest, main = vars[5])
boxplot(data$abdom, main = vars[6])
```



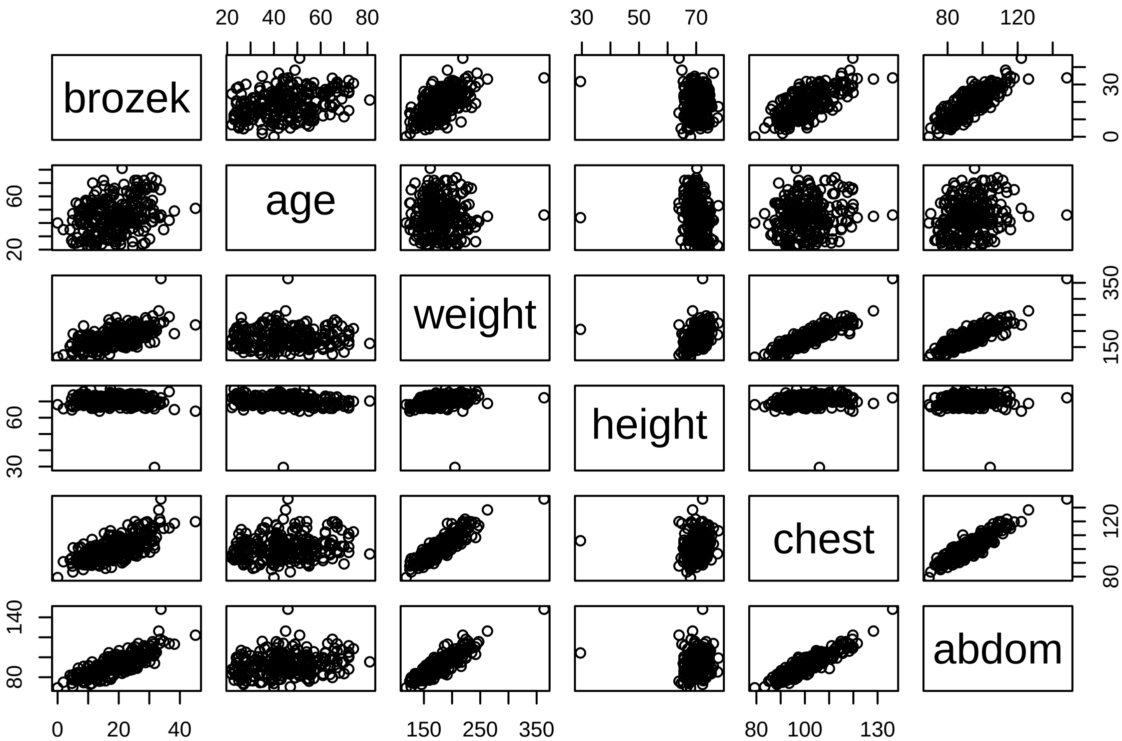
3. Briefly discuss the shape of the distribution of each variable

Answer: The boxplots show different distribution patterns for the variables. `Brozek` (percent body fat) appears slightly right-skewed, with the median positioned closer to the lower quartile and a few high outliers. `Age` looks fairly symmetric since the median is centered and the whiskers are about the same length. Weight is right-skewed, as shown by the longer upper whisker and a few high outliers. `Height`, however, is slightly left-skewed because the median is closer to the top of the box and the lower whisker is longer. `Chest` circumference seems roughly symmetric, with an even spread and some outliers on both ends. `Abdomen` circumference is right-skewed, with a longer upper whisker and several high outliers. Overall, `weight`, `brozek`, and `abdomen` show right skewness, `height` is slightly left-skewed, and `age` and `chest` circumference appear more symmetric.

4. Create a scatterplot matrix to explore the inter-dependence between these variables

Code:

```
pairs(data)
```



General Linear F-Test

Suppose a researcher would like to compare the "Full" model using all the 5 predictors and a "reduce" model where only x_1 (`age`) and x_5 (`abdom`) are used by performing a general linear F-test:

5. Write down the null and the alternative hypotheses.

Answer:

Null Hypothesis (H_0): The additional predictors of `weight`, `height`, `chest` do not improve the model (coefficients equal to zero).

Alternative Hypothesis (H_a): At least one of the additional predictors contributes to explaining `brozek`.

Mathematically, these can be demonstrated as:

$$H_0 : x_2 = x_3 = x_4 = 0$$

$$H_A : \text{At least one of } x_2, x_3, \text{ or } x_4 \neq 0$$

6. Fit the full model and write down the fitted linear regression equation.

Code:

```r
full_model <- lm(brozek ~ age + weight + height + chest + abdom, data = data)
summary(full_model)
```

```
##
## Call:
## lm(formula = brozek ~ age + weight + height + chest + abdom,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.6515  -2.9213   0.0552   2.9019   9.4269
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -32.153538   7.779978  -4.133 4.92e-05 ***
## age          -0.006447   0.024734  -0.261    0.795
## weight       -0.121843   0.028160  -4.327 2.20e-05 ***
## height       -0.118164   0.083492  -1.415    0.158
## chest        -0.012862   0.087484  -0.147    0.883
## abdom         0.894248   0.074150  12.060  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.134 on 246 degrees of freedom
## Multiple R-squared:  0.7212, Adjusted R-squared:  0.7155
## F-statistic: 127.2 on 5 and 246 DF,  p-value: < 2.2e-16
```

Answer:

Formula to follow:

$$\hat{y} = x_0 + x_1 \cdot \text{age} + x_2 \cdot \text{weight} + x_3 \cdot \text{height} + x_4 \cdot \text{chest} + x_5 \cdot \text{abdom}$$

With estimated coefficients:

$$\hat{y} = -32.15 - 0.0064 \cdot \text{age} - 0.1218 \cdot \text{weight} - 0.1182 \cdot \text{height} - 0.0129 \cdot \text{chest} + 0.8942 \cdot \text{abdom}$$

7. Fit the reduced model and write down the fitted linear regression equation.

Code:

```r
reduced_model <- lm(brozek ~ age + abdom, data = data)
summary(reduced_model)
```

```
##
## Call:
## lm(formula = brozek ~ age + abdom, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.7114  -3.2622   0.0285   3.2248  12.0577
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -36.51507    2.46972 -14.785  < 2e-16 ***
## age           0.06605    0.02290   2.884  0.00427 **
## abdom         0.56710    0.02677  21.187  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.45 on 249 degrees of freedom
## Multiple R-squared:  0.673,  Adjusted R-squared:  0.6704
## F-statistic: 256.3 on 2 and 249 DF,  p-value: < 2.2e-16
```

Answer:

Formula to follow:

$$\hat{y} = x_0 + x_1 \cdot \text{age} + x_5 \cdot \text{abdom}$$

With estimated coefficients:

$$\hat{y} = -36.52 + 0.0661 \cdot \text{age} + 0.5671 \cdot \text{abdom}$$

8. Perform a general linear F-test and state the conclusion at $\alpha = 0.05$

Code:

```r
anova(reduced_model, full_model)
```

```
## Analysis of Variance Table
##
## Model 1: brozek ~ age + abdom
## Model 2: brozek ~ age + weight + height + chest + abdom
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    249 4930.3
## 2    246 4204.7  3     725.6 14.151 1.543e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer:

The p-value is 1.543e-08 (or 0.00000001543), meaning it is much smaller than the 0.05 significance level. This means we reject the null hypothesis (H_0). The additional predictors of `weight, height, chest` significantly improve the prediction of body fat percentage.

Prediction

9. Predict a future response for an individual with `age` = 54, `weight` = 197, `height` = 72.25, `chest` = 105.375, and `abdom` = 99.325. Construct a 95% prediction interval.

Code:

```r
new_data <- data.frame(age = 54,
                       weight = 197,
                       height = 72.25,
                       chest = 105.375,
                       abdom = 99.325)

predict(full_model, newdata = new_data, interval = "prediction")
```

```
##        fit      lwr      upr
## 1 22.42373 14.24419 30.60327
```

Answer:

The prediction for future response with provided figures leads to a predicted value of 22.42373. The 95% prediction interval is [14.24419, 30.60327].

10. Construct a 95% confidence interval for the mean response of percent body fat with `age` = 54, `weight` = 197, `height` = 72.25, `chest` = 105.375, and `abdom` = 99.325.

Code:

```r
predict(full_model, newdata = new_data, interval = "confidence")
```

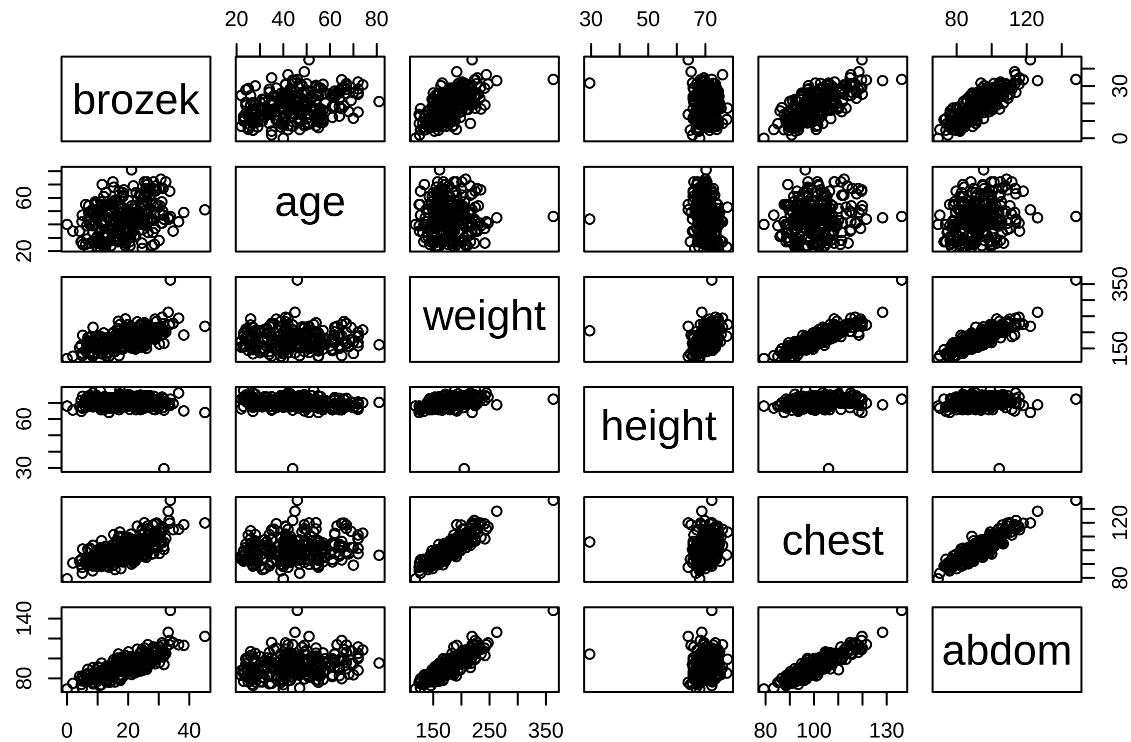
```
##        fit      lwr      upr
## 1 22.42373 21.65224 23.19523
```

Multicollinearity

11. Make the scatterplot matrix and compute the correlation matrix for all 6 variables (including the response).

Code:

```r
pairs(data)
```




```r
cor(data)
```

```
##             brozek         age      weight      height     chest      abdom
## brozek  1.00000000  0.28917352  0.61315611 -0.08910641 0.7028852 0.81370622
## age     0.28917352  1.00000000 -0.01274609 -0.17164514 0.1764497 0.23040942
## weight  0.61315611 -0.01274609  1.00000000  0.30827854 0.8941905 0.88799494
## height -0.08910641 -0.17164514  0.30827854  1.00000000 0.1348918 0.08781291
## chest   0.70288516  0.17644968  0.89419052  0.13489181 1.0000000 0.91582767
## abdom   0.81370622  0.23040942  0.88799494  0.08781291 0.9158277 1.00000000
```

12. Calculate VIF and briefly discuss your finding

Code:

```r
vif(full_model)
```

```
##       age    weight    height     chest     abdom
##  1.426799 10.058282  1.373446  7.987963  9.388374
```

Answer:

The Variance Inflation Factor (VIF) analysis reveals the presence of multicollinearity in the model, particularly for `weight`, `chest`, and `abdom`. The VIF value for `weight` is 10.06, which exceeds the threshold of 10, indicating severe multicollinearity and suggesting that `weight` is highly correlated with other predictors. Additionally, `chest` and `abdom` have VIF values of 7.99 and 9.39, respectively, which fall within the moderate-to-high range, further suggesting that these predictors share substantial information with one another. In contrast, `age` and `height` have low VIF values of 1.43 and 1.37, respectively, indicating minimal collinearity concerns for these variables. The presence of high multicollinearity, particularly for `weight`, may inflate standard errors and reduce the reliability of coefficient estimates, making it difficult to determine the true effect of individual predictors in the model.