Multiple Linear Regression (Model Selection and Model Counting) - Lab

Blake Pappas

December 17, 2023

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)
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
```

Only the following variables will be used 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:

Use the code below to extract these variables.

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

Exploratory Data Analysis

Numerical Summary

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

Code:

```
summary(data)
```

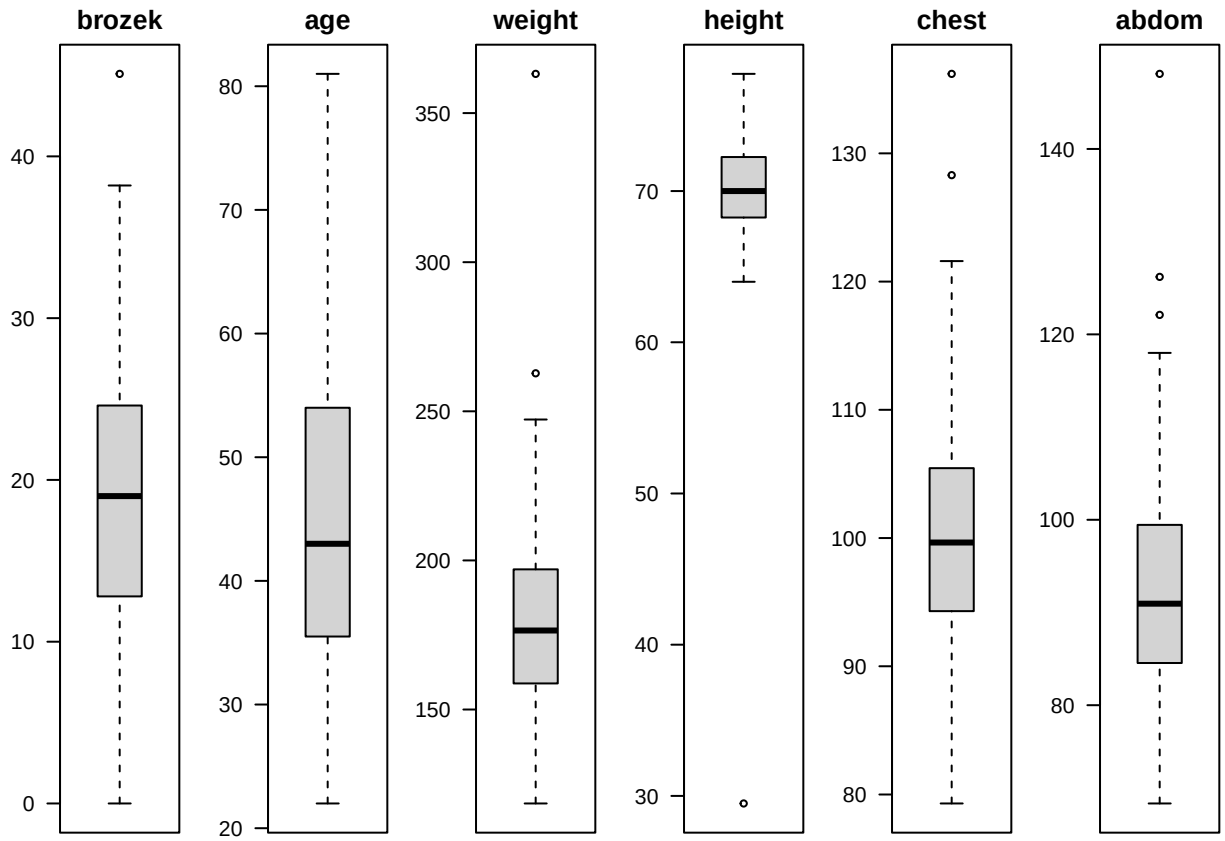
```
##      brozek           age            weight          height
##  Min.   : 0.00   Min.   :22.00   Min.   :118.5   Min.   :29.50
##  1st Qu.:12.80   1st Qu.:35.75   1st Qu.:159.0   1st Qu.:68.25
##  Median :19.00   Median :43.00   Median :176.5   Median :70.00
##  Mean   :18.94   Mean   :44.88   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:

```
p <- dim(data)[2]
par(mfrow = c(1, p), mar = c(1, 3, 2, 0.5))
for (i in 1:p)
boxplot(data[, i], main = colnames(data)[i], las = 1)
```



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

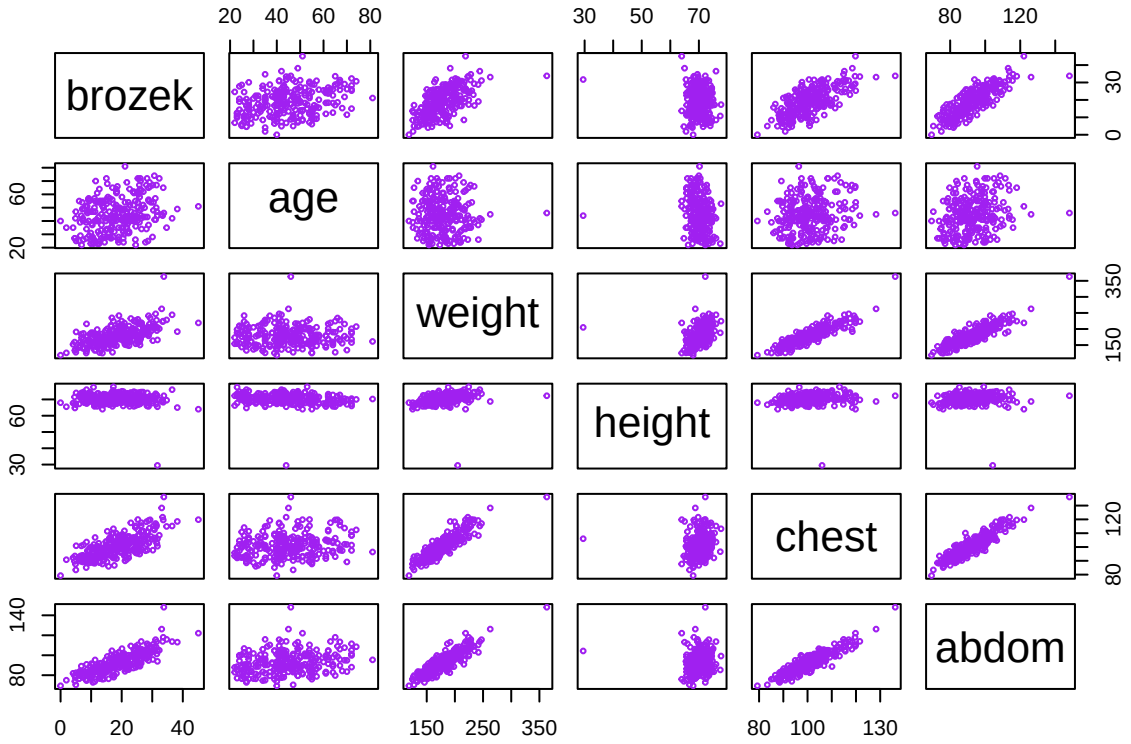
Answer:

`brozek` is close to symmetric with one upper outlier. Age is (slightly) right-skewed. `weight` is approximately symmetric with two upper outliers. `height` is approximately symmetric with a single lower outlier. `chest` is slightly right skewed with two upper outliers. `abdom` is right skewed with three upper outliers.

4. Create a scatterplot matrix to explore the interdependence between these variables.

Code:

```
pairs(data, cex = 0.5, col = "purple")
```




```
library(GGally)
```

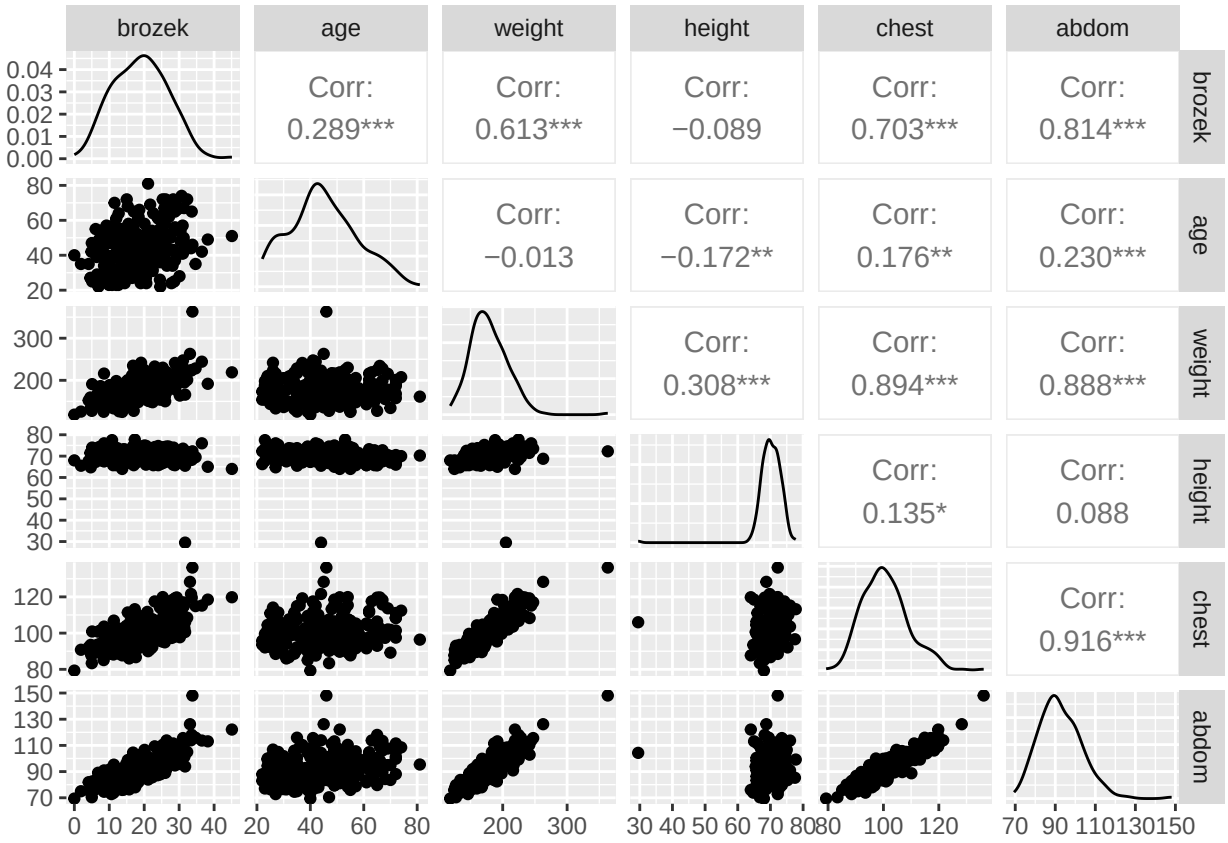
```
## Loading required package: ggplot2
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
##
## Attaching package: 'GGally'
```

```
## The following object is masked from 'package:faraway':
##
##     happy
```

```
ggpairs(data)
```



General Linear F-Test

Suppose a researcher would like to compare between the "full" model using all the 5 predictors and a "reduced" 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:

$H_0 : \beta_{\texttt{weight}} = \beta_{\texttt{height}} = \beta_{\texttt{chest}} = 0$ vs. H_a : At least one the above three coefficients $\neq 0$

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

Code:

```
full <- lm(brozek ~ ., data = data)
summary(full)
```

```
##
## Call:
## lm(formula = brozek ~ ., 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:

$\hat{\texttt{brozek}} = -32.153538 - 0.006447 \times \texttt{age} - 0.121843 \times \texttt{weight} - 0.118164 \times \texttt{height} - 0.012862 \times \texttt{chest} + 0.894248 \times \texttt{abdom}$

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

Code:

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

```
##
## 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:

$\hat{\texttt{brozek}} = -36.51507 + 0.06605 \times \texttt{age} + 0.56710 \times \texttt{abdom}$

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

Code:

```
anova(reduce, full)
```

```
## 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:

Since the p-value of this general linear F-test is less than α, we reject H_0 and conclude that we have sufficient evidence to support that at least one of the three regression coefficients is not equal to 0.

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:

```
new <- data.frame(age = 54, weight = 197, height = 72.25, chest = 105.375, abdom = 99.325)
predict(full, newdata = new, interval = "prediction")
```

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

Answer:

The predicted value is 22.4237316 and the 95% prediction interval is [14.2441941, 30.6032691].

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:

```
predict(full, newdata = new, interval = "confidence")
```

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

Answer:

The 95% prediction interval is [21.6522351, 23.195228].

Multicollinearity

11. Compute the correlation matrix for all 6 variables (including the response).

Code:

```
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 the VIF and briefly discuss the findings.

Code:

```
vif(full)
```

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

Answer:

Since `weight`, `chest`, and `abdom` all have "high" VIF values (i.e., > 5), we determine that we have multicollinearity between these predictors, meaning that these predictors are highly (positively) correlated. There does not appear to be multicollinearity with the other two predictors of `age` and `height`.