# 1 Initial Model

As stated previously, the initial model we are interested in fitting is of the form:
$\text{weight}_i = \beta_0 + \beta_1 \text{ chest.diam}_i + \beta_2 \text{ chest.dep}_i + \beta_3 \text{ bitro.diam}_i + \beta_4 \text{ wrist.min}_i + \beta_5 \text{ ankle.min}_i + \beta_6 \text{ height}_i$

```
## (Intercept)  chest.diam   chest.dep  bitro.diam   wrist.min
##    -109.890       1.340       1.537       1.196       1.113
##   ankle.min      height
##       1.152       0.177
```

This model indicates that the expected change in weight for a 1 unit change in chest.diam, holding all other variables constant, is 1.34 lbs. The expected change in weight for a 1 unit change in chest.dep, holding all other variables constant, is 1.54 lbs., etc. Note that chest depth has the largest impact on weight. In addition to the coefficients, the R-squared value of 0.8882 implies that our model explains 88.82% of the variation in weight and the $P$-values for each variable and for the model are significant. This model seems like a good tool to predict weight given these measurements, but are there better ones?

# 2 Model Selection

We are building a model to predict weight given various body measurements. Before running random models, we need to determine what predictors to use. The predictors needed in our models are age, height and gender. These variables contribute significantly to weight. The predictors we will allow in model selection are the initial predictors: chest diameter, chest depth and bitro diameter. In addition to these variables, pelvic bredth, shoulder, chest, waist, hip and thigh will be used. I chose to allow these predictors in my model since these are directly associated with weight (e.g. waist). However, for the other models we will fit, we will let "R" do it's work.

## 2.1 Criterion

The Information Criterions we will be using to evaluate our models are Akaike Information Crierion (AIC), Bayes Information Criterion (BIC), adjusted $R^2$ and Predictive Residual Sum of Squares (PReSS). In short, AIC and BIC measure goodness-of-fit through residual sum of squares (log likelihoods) and penalizes the model size; the smaller the AIC/BIC, the better. Adjusted $R^2$ adjusts $R^2$ so that the model is penalized for adding more predictors; the higher the value of the adjusted $R^2$ the better. Finally, PRESS is a summary measure focused on prediction; the lower the value of PRESS, the better.

$$\begin{aligned} \text{AIC} &= n \log\left(\frac{RSS}{n}\right) + 2(p+1) \\ \text{BIC} &= n \log\left(\frac{RSS}{n}\right) + (p+1)\log(n) \\ \text{adj}R^2 &= 1 - \frac{n-1}{n-p-1}(1-R^2) \end{aligned}$$

$$\text{PRESS} = \sum \left( \frac{\hat{\epsilon}_i}{1 - h_{ii}} \right)^2$$

## 2.2 Methods in R

There are multiple methods built into different packages in R for Model Selection. To illustrate these, we will use the variables: height, wrist.min, ankle.min and chest and call this model "MLRex".

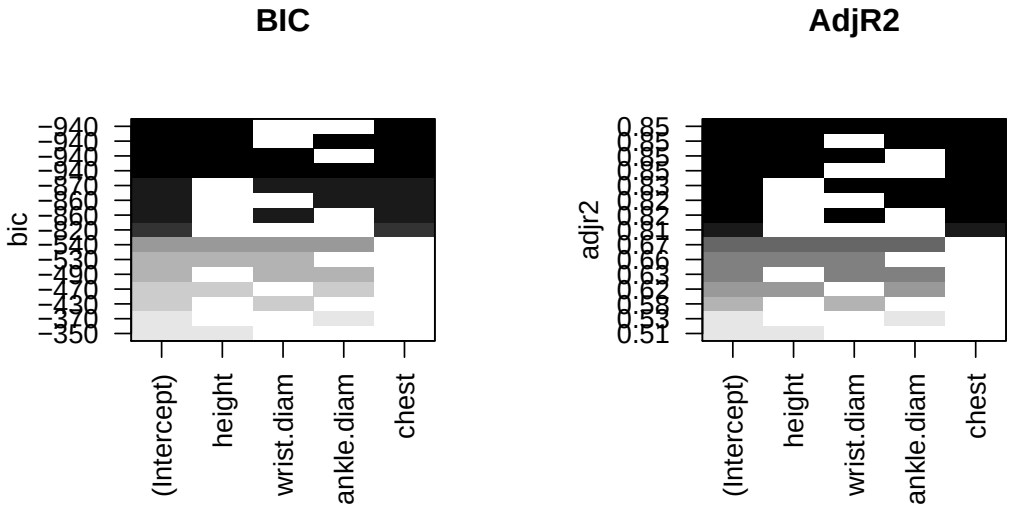
### 2.2.1 stepAIC()

The R function found in the package "MASS" called "stepAIC()" performs stepwise model selection by AIC. This will output the initial model and the final model (model of best fit determined by this method), and the steps taken. In the output below we can see that this method suggests using a different model that doesn't contain wrist.diam.

```
Initial Model:  weight   height + wrist.diam + ankle.diam + chest
Final Model:  weight   height + ankle.diam + chest
            Step  Df  Deviance  Resid. Df  Resid. Dev       AIC
 1                                    502     13294.14  1666.150
 2  - wrist.diam   1  47.90427         503     13342.05  1665.974
```

### 2.2.2 leaps()

The R package "leaps" contains a function "regsubsets()". This method performs an exhaustive search of models and plots the $R^2$ criterion by variables and subset size. The class "summary.regsubsets" outputs an object with multiple elements, including adjusted $R^2$ and BIC. Furthermore, the plots below plot the BIC and Adjusted $R^2$ values against each subset of variables.



In these plots, for example, the BIC plot is implying the best model is using height and chest as predictors. On the other hand, the AdjR2 plot is saying all variables give the best fit. Using both of these plots together, one might conflude height, ankle.diam and chest would be the best fit. This conclsion agrees with our analysis using stepAIC.

### 2.3 Model Selection in Action

The Model Selection Lab will explain how to obtain the models summarized in Table 1.
**Table 1**

| MLR# | AIC | BIC | PRESS | Adjusted $R^2$ | Method |
|---|---|---|---|---|---|
| all | 2216 | 2324 | 2383 | 0.9753 | all variables from dataset used |
| i | 2970 | 3004 | 10405 | 0.8869 | suggested by paper |
| 1 | 2256 | 2319 | 2560 | 0.9727 | suggested by paper |
| 2 | 2402 | 2441 | 3408 | 0.9632 | my model |
| 3 | 2206 | 2282 | 2329 | 0.9754 | stepAIC |
| 4 | 2195 | 2271 | 2281 | 0.9759 | stepAIC and adjustments |
| 5 | 2207 | 2292 | 2335 | 0.9755 | leaps (adj $R^2$) |
| 6 | 2189 | 2278 | 2255 | | leaps(adj $R^2$)and adjustments |
| 7 | 2213 | 2272 | 2353 | 0.9749 | leaps(BIC) |
| 8 | 2194 | 2257 | 2267 | 0.9759 | leaps(BIC) and adjustments |

Base on Table 1, we can see that each Criterion yields different results. It is up to our discretion to choose a model. Since AIC and BIC are lowest in MLR8, the adjusted $R^2$ is the larges, and the PRESS is close to 2300, I would choose MLR8 as the model of best fit. MLR8 is of the form:
$\text{weight}_i = \beta_0 + \beta_1\ \text{chest.dep}_i + \beta_2\ \text{knee.diam}_i + \beta_3\ \text{shoulder}_i + \beta_4\ \text{chest}_i + \beta_5\ \text{waist}_i + \beta_6\ \text{hip}_i + \beta_7\ \text{thigh}_i + \beta_7\ \text{forearm}_i + \beta_8\ \text{calf}_i + \beta_9\ \text{gender}_i + \beta_{10}\ \text{height}_i + \beta_{11}\ \text{height}_i^2 + \beta_{12}\ \text{age}_i^2$.

```
## (Intercept)   chest.dep   knee.diam    shoulder       chest
##  -1.512e+01   2.268e-01   6.364e-01   8.838e-02   1.673e-01
##       waist         hip       thigh     forearm        calf
##   3.779e-01   2.454e-01   2.469e-01   5.989e-01   4.210e-01
##      height      gender    I(age^2) I(height^2)
##  -9.414e-01  -1.480e+00  -7.361e-04   3.667e-03
```

## 3 Conclusion

Model Selection truely is an art form. R can mechanically run through steps, interactions, combinations, etc. However, R cannot subjectively look at the variables to determine the absolute best model. To acheive the model of best fit, a combination of methods and human adjustment is necessary.