Predicting Percent Body Fat

Arden Miller

Executive Summary

Measurements made on athletes at the Australian Institute of Sport were used to construct predictive models for percent body fat. It was found that a simple linear regression model that just uses skinfold thickness to predict percent body fat produces estimates that have a margin of error of $\pm 3.5\%$ at a 95% confidence level. In simple terms, this means that estimates of percent body fat made using this model should be within $\pm 3.4\%$ of the true value 95% of the time. A substantial improvement in the precision of predictions results from using separate models for female and male athletes. A slight further improvement can be obtained by using the height of the athlete in addition to skinfold as a regressor in these models. For these models percent body fat can be predicted with 95% confidence with a margin of error of $\pm 2.9\%$ for female athletes and $\pm 1.7\%$ for male athletes.

The Australian Institute of Sport Data

The data investigated in this report comes from the Australian Institute of Sport and consists of measurements taken on approximately 200 athletes. The data represents athletes from a number of different disciplines and observations are approximately evenly divided between female and male athletes. The measured variables are:

Sex	1=female 2=male
BMI	body mass index (weight/height^2)
SSF	sum of skinfolds
Bfat	percentage body fat
Ht	height (m)
Wt	weight (kg)

Figure 1 contains pairwise scatter plots of these variables.

Percent body fat (Bfat) is of particular interest as it is a key element of overall fitness but it is difficult to measure directly. The focus of this investigation was to develop regression models that could be used to predict percent body fat using other variables that are easier to measure. It is important to note that as this data consists of measurements made on elite athletes, the models discussed should only be used to predict percent body fat for other athletes.

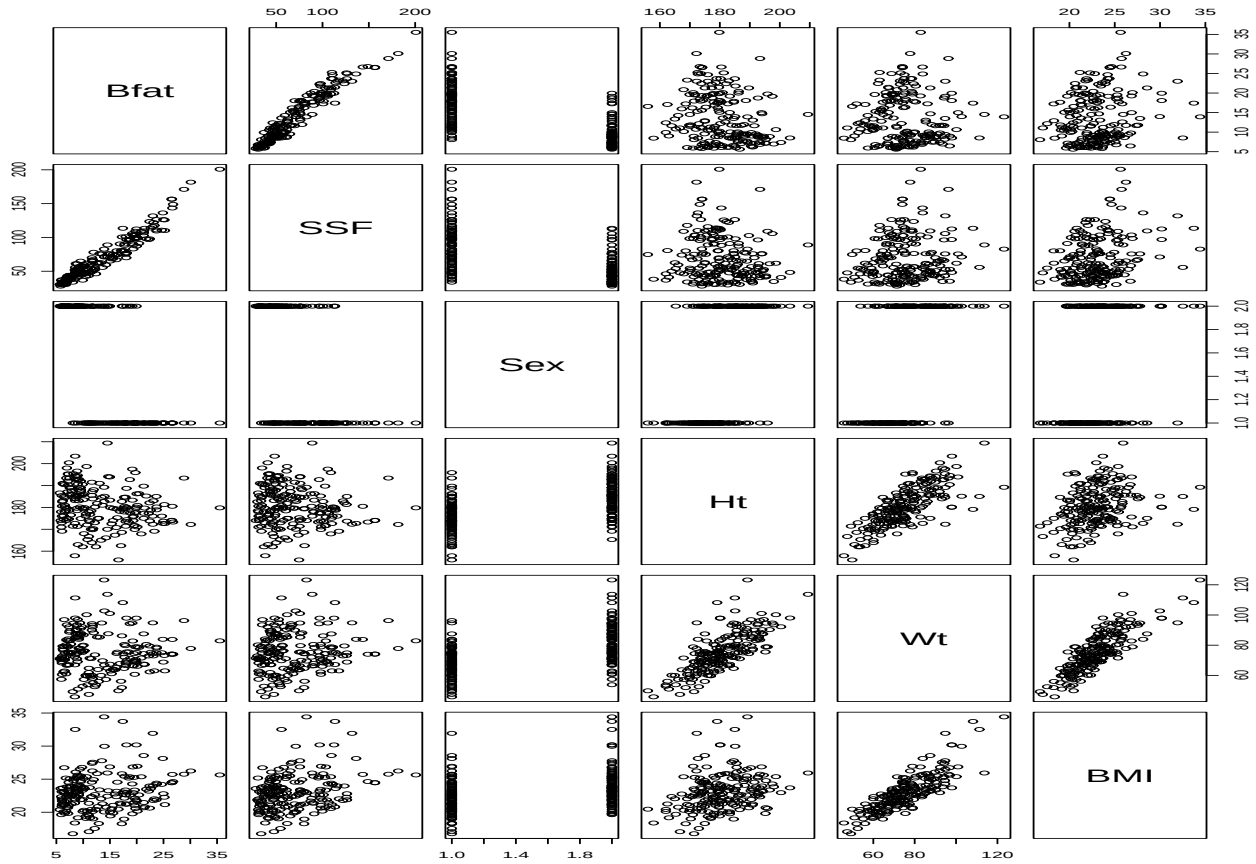

Figure 1: Scatterplots of variables from the AIS data.

Predicting Percent Body Fat using Regression Models

A number of different regression models for predicting percent body fat were explored. Skinfold thickness is often used as a predictor of percent body fat. Callipers are used to measure skinfold thickness at various body locations and the sum of these measurements is recorded. First, a very simple model that used a linear function of SSF to predict Bfat was considered. Then the data was split into two parts based on the sex of the athletes to see if this would improve the precision of predictions. Finally, the inclusion of additional variables (height, weight, and body mass index) was considered to see if this would improve predictions.

The scatter plot of Bfat versus SSF (skinfold thickness) in Figure 1 indicates that this relationship is particularly strong. As this relationship is reasonably linear a simple linear regression model that used SSF to predict Pfat was tried. Fitting the linear regression model produces:

$$\widehat{\text{Bfat}} = 0.820 + 0.183 \times \text{SSF}$$

where $\widehat{\text{Bfat}}$ represents the predicted values of Bfat. A statistical analysis of the precision of these predictions made with this equation indicates that at a 95% confidence level the predictions have a margin of error of $\pm 3.4\%$. That is we expect the predicted value should be within 3.4% of the true value 95% of the time.

Secondly, creating separate models for female athletes and for male athletes was considered. Figure 2 contains a scatterplot of Bfat versus SSF where the plotting character indicates the

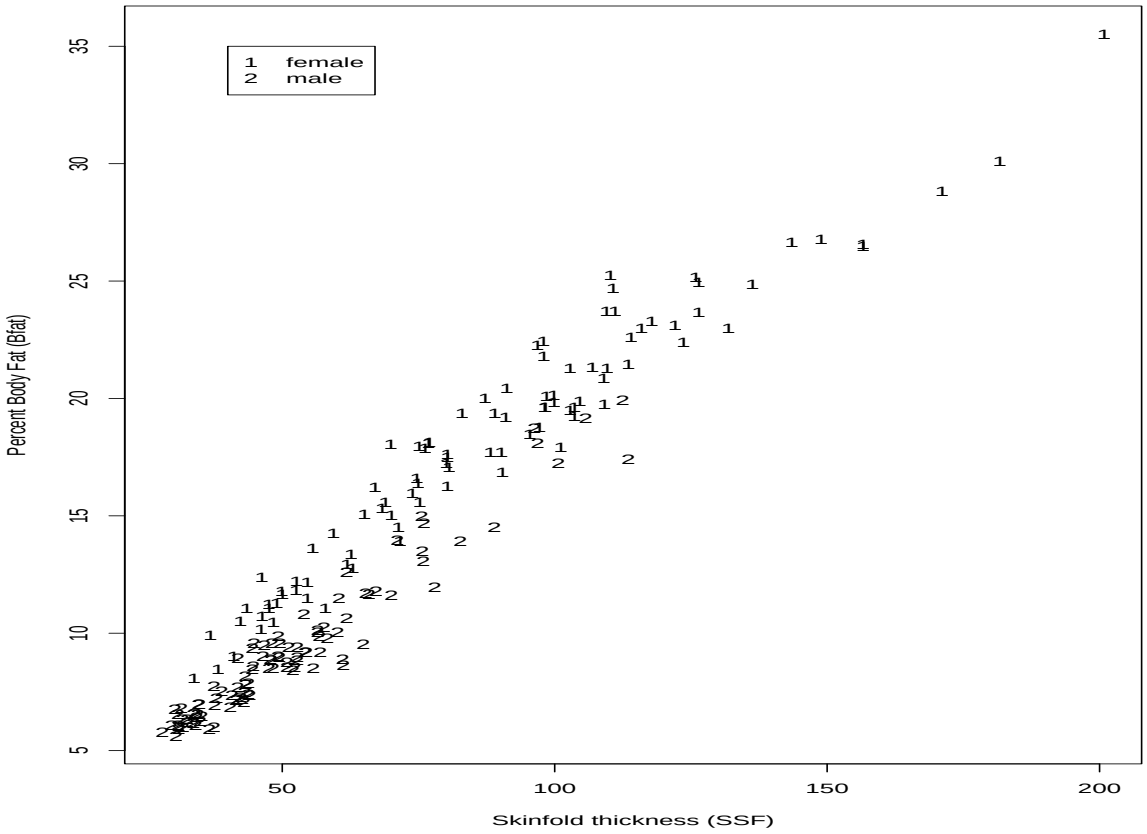

Figure 2: Scatterplot of Percent Body Fat versus Skinfold Thickness.

sex of the athlete. This plot clearly indicates that the relationship between Bfat and SSF is somewhat different for female athletes and male athletes. This suggests that separate models should be used. Fitting such models using this data set gives:

$$\begin{aligned} \text{Female athletes:} \quad & \widehat{\text{Bfat}} = 4.411 + 0.155 \times \text{SSF} \\ \text{Male athletes:} \quad & \widehat{\text{Bfat}} = 0.849 + 0.163 \times \text{SSF} \end{aligned}$$

For these models the margins of error for a 95% confidence level are $\pm 2.9\%$ for female athletes and $\pm 1.7\%$ for male athletes.

Finally, the possibility improving the prediction of Bfat by adding one or more of height, weight, and body mass index to the models was investigated. It was found that adding height to the existing models resulted in a slight improvement to each. The new models are:

$$\begin{aligned} \text{Female athletes:} \quad & \widehat{\text{Bfat}} = -1.961 + 0.152 \times \text{SSF} + 0.038 \times \text{Ht} \\ \text{Male athletes:} \quad & \widehat{\text{Bfat}} = 4.966 + 0.166 \times \text{SSF} - 0.023 \times \text{Ht} \end{aligned}$$

The margins of error for these models are essential the same as for the previous models: $\pm 2.9\%$ for female athletes and $\pm 1.7\%$ for male athletes.

Overall, I would recommend using separate regression models for female and male athletes and that both skinfold thickness and height be used as explanatory variables in these models.

However, height could be dropped from these models with only a very small reduction in the precision of the predictions.

Statistical Appendix

First a simple linear regression model that used SSF to predict Bfat was fitted using the entire data set. This model had a very small P-value for the test of the hypothesis that the coefficient of SSF was really 0. It also had a high value of R^2 (92.7%). To evaluate the precision of predictions made using this model, I considered 95% prediction intervals for SSF = 28, 69.4, and 200.8 (the minimum, mean, and maximum values of SSF for this data). The margins of error for these intervals were ± 3.3, ± 3.3, and ± 3.4. As these values are all very close to each other I decided to just quote $\pm 3.4\%$ as the margin of error.

I decided that it was a good idea to create separate models for female and male athletes since Figure 1 indicates that the relationship between Bfat and SSF is somewhat different for female and male athletes. As was stated in the report, fitting the separate models reduced the width of the 95% prediction intervals and thus improves the precision of predictions.

Finally, I tried adding all possible combinations of Ht, Wt, and BMI to these models. For both male and female athletes, the best model was the one that contained Ht in addition to SSF. For male athletes the P-value for Ht is significant (.03) and it is marginal for female athletes (.07). I calculated the widths of 95% prediction intervals using these new models and found the margins of error to be very nearly the same as for the models that don't contain height.