

## Homework #3 Solutions

Basic body measurements are taken on 86 incoming female cadets at one of the U.S. military academies. The variables measured are: 1. height (cm), 2. weight (kg), 3. hip circumference (cm), 4. forearm circumference (cm), 5. neck circumference (cm), 6. wrist circumference (cm), 7. tricep skinfold (mm), 8. scapular skinfold (mm), 9. suprailiac skinfold (mm), and 10. body fat as percent of body weight. We are interested in predicting body fat (which is the most difficult of these to

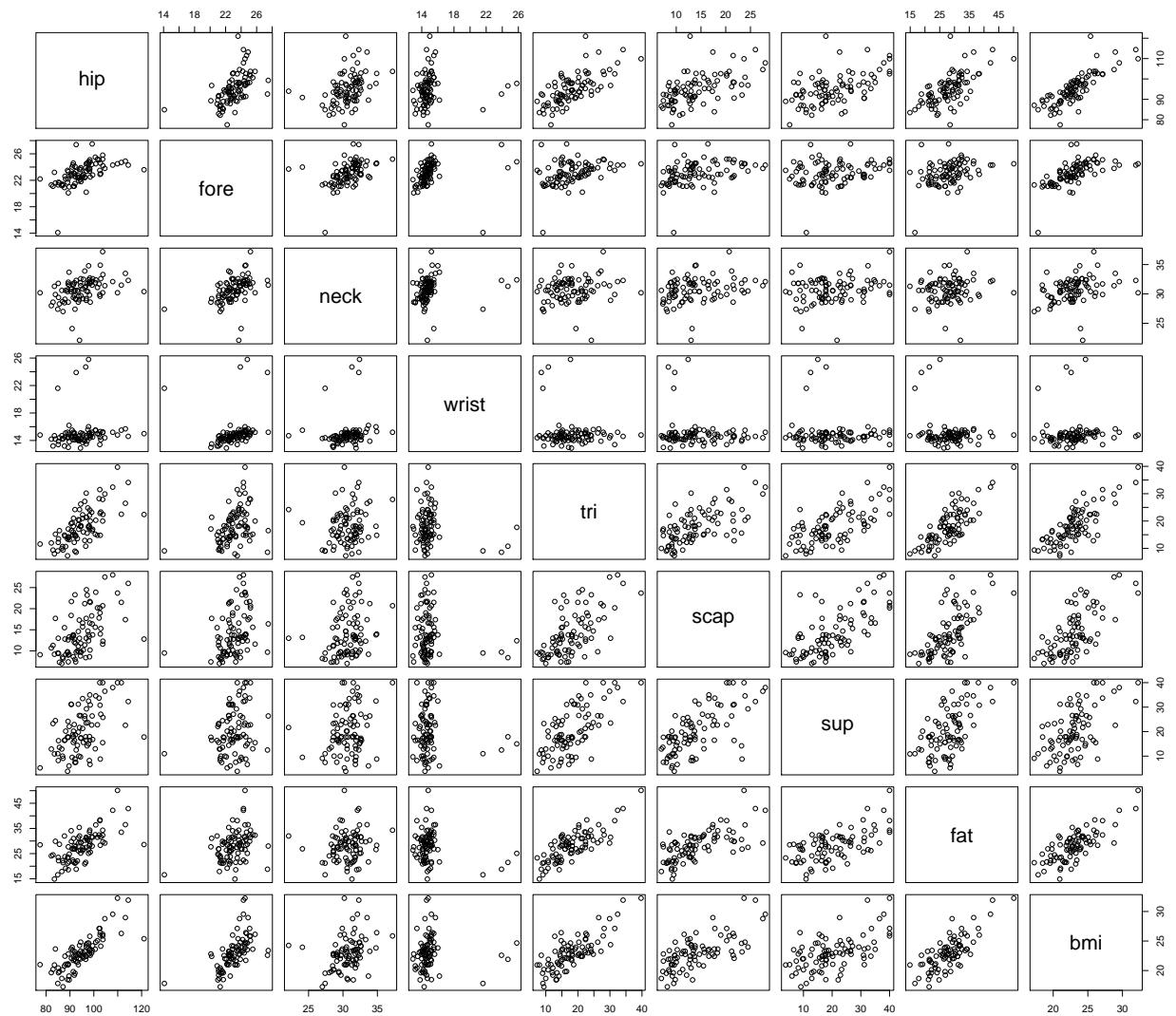


Figure 1: Scatterplot of entire dataset

A quick look at the correlation between the variables listed above shows that the weight variable is highly correlated with most other anthropomorphic measurements. Because the Body Mass Index ( $kg/m^2$ ) is widely used, I converted the height and weight variables to the BMI and therefore reduced some of the collinearity present in the dataset as well as the total number of variables to be analyzed.

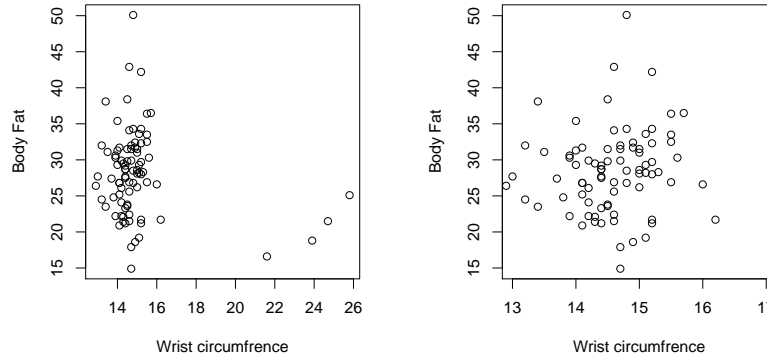


Figure 2: Scatterplot of the wrist covariate, (a) with and (b) without outliers

Next, to start the analysis, a scatterplot of all the potential covariates in the dataset, as well as the outcome, is drawn. This scatterplot is shown in Figure 1. It is immediately obvious that the forearm and neck circumference are not related to body fat. So, these are removed. Next, the plot of wrist circumference versus body fat shows very little association, as well as four very large outliers (see Figure 2a). A univariate analysis of wrist versus body fat is statistically significant, however. If the four outliers are removed, the relationship between wrist and body fat disappears. This is shown in Figure 2b. So, the wrist variable is also removed from the dataset. Hence, I proceeded to examine BMI, hip circumference, tricep skinfold, scapular skinfold, and suprailiac skinfold for their ability to predict body fat.

With five covariates remaining, I calculated the variance inflation factor (VIF) for each and fit the full linear regression model. The output is in Table 1.

The BMI variable and the hip circumference variable are highly correlated and therefore one variable

Table 1:  $\text{bodyfat} = \beta_0 + \beta_1\text{BMI} + \beta_2\text{hip} + \beta_3\text{tricep} + \beta_4\text{scapular} + \beta_5\text{suprailiac}$

Covariate	$\beta$	p-value	VIF	$R^2$
BMI	0.60	0.03	4.70	0.704
Hip	-0.004	0.96	3.03	
Tricep	0.42	0.00009	3.34	
Scapular	0.18	0.099	2.46	
Suprailiac	0.02	0.76	2.27	

eclipses the significance of the other. In univariate analysis, the BMI seems to perform better and hence we will keep this variable and eliminate the hip circumference variable. Also, the suprailiac skinfold is not statistically significant in models that include the BMI and tricep skinfold. So, this variable is also eliminated. Furthermore, a stepwise selection procedure that uses AIC as the selection criteria also eliminates hip circumference and suprailiac skinfold. Therefore, the final model is:

$$\text{bodyfat} = \beta_0 + \beta_1\text{BMI} + \beta_2\text{tricep} + \beta_3\text{scapular} \quad (1)$$

and the regression table for this model is in Table 2.

The VIFs for each covariate are now much less than five, and so the majority of the collinearity problem has been resolved. Also, there was no reduction in the  $R^2$  from the five to the three covariate model, so we have likely selected a minimally sufficient model for prediction. All of the regression coefficients are statistically significant and therefore different from zero. Further, they are also positive indicating a positive relationship between increasing covariate and increasing body fat. For example, when the tricep and scapular skinfold measurements are the same for two women but their BMI differs by one point, the increase in body fat is 0.587. Of course, because there is

Table 2:  $\text{bodyfat} = \beta_0 + \beta_1\text{BMI} + \beta_2\text{tricep} + \beta_3\text{scapular}$

	Full dataset				Outlier removed (obs. 74)			
Covariate	$\beta$	p-value	VIF	$R^2$	Covariate	$\beta$	p-value	$R^2$
Intercept	3.897	0.3		0.704	Intercept	2.65	0.45	0.74
BMI	0.587	0.01	3.09		BMI	0.614	0.004	
Tricep	0.435	0.00001	2.8		Tricep	0.434	0.000004	
Scapular	0.195	0.046	1.995		Scapular	0.248	0.009	

some correlation between these three variables, the probability that two covariates remain fixed while another changes is unlikely.

Figure 3 shows that added predictive advantage of the model given in equation (1) compared to other options. The first two plots in Figure 3 indicate that nothing is lost by eliminating hip and suprailiac from the model. The last two plots in Figure 3 show the fitted values compared to the outcome for univariate models containing tricep alone and BMI alone. Comparing these last two plots for the plot corresponding to equation (1) indicates a better relationship between fitted values and outcome for the model that we have shown in equation (1). This gives further confidence that the final model chosen for prediction is the best.

The major regression assumptions all appear to be met in this dataset. The 86 incoming cadets in the study are probably independent, and so we can generalize results to all incoming female cadets. Also, the relationships between the three covariates, BMI, tricep, and scapular skinfold, and body fat all appear to be linear. An examination of the residuals in Figure 4 indicates that residuals are evenly scattered and normally distributed. There does appear to be one outlying point and this point, indicated in the first panel of Figure 4, has a studentized residual less than  $-3$ . This point

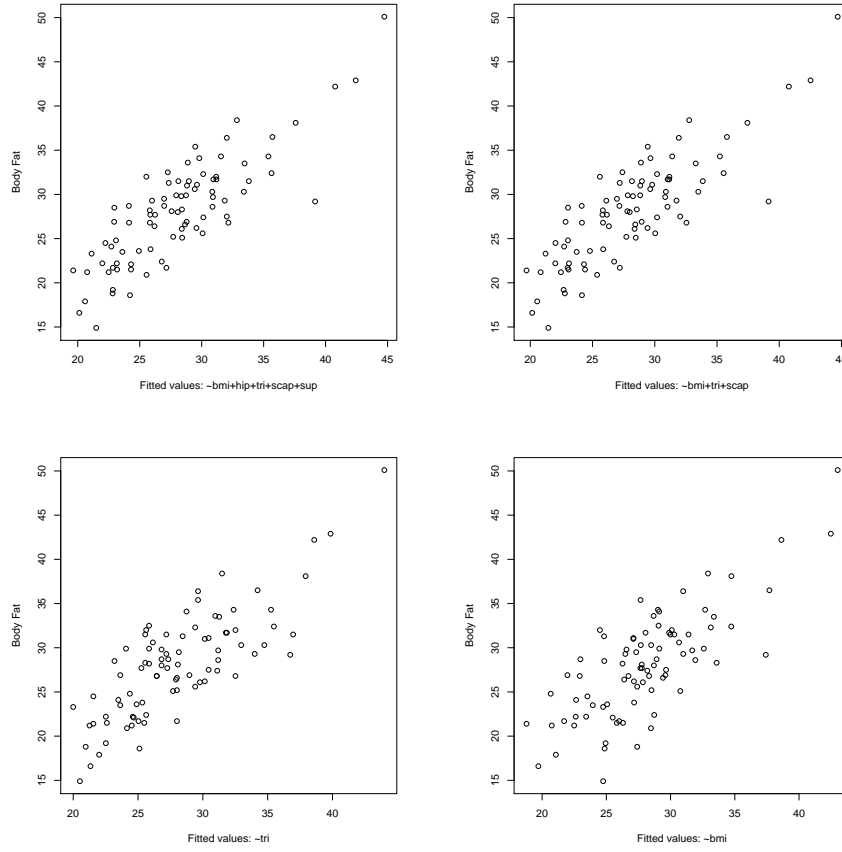


Figure 3: Plots comparing fitted values to outcome for four different models

also has the largest Cook's distance as indicated in the last panel of Figure 4.

Figure 5 gives the  $\Delta\beta$ 's for each of the linear regression coefficients. Given the variation in the influence diagnostics, it seemed best to fit the model again removing the outlying point. The results for this regression, eliminating the outlier, are also shown in Table 2. The regression coefficients are not altered too greatly and the major conclusions certainly do not change. Therefore, in the absence of further information about the outlier, it is best to leave it in the model and use the results from the complete dataset.

Here is the code in R used to analyze the dataset:

```
fatty <- read.table("bodyfat.dat",header=T)

fatty$bmi <- fatty$wt/((fatty$ht/100)^2)

pairs(fatty[,c(4:12)])

#Scatterplot indicates that fore and neck aren't very interesting,
#so we take them out, they also aren't sig. in the big full model
par(mfrow=c(1,2),pty='s')

plot(fatty$wrist,fatty$fat,xlab='Wrist circumfrence',ylab='Body Fat')

plot(fatty$wrist,fatty$fat,xlim=c(13,17),xlab='Wrist circumfrence',
ylab='Body Fat')

# The variable wrist has 4 outliers that are highly influential
# and so the variable appears sig. It really isn't. We'll take out
# the covariate.

pairs(fatty[,c(4,8:12)])

vifs <- rep(0,5)

out <- summary(lm(fatty[,4]~fatty[,8]+fatty[,9]+fatty[,10]+fatty[,12]))$r.sq
vifs[1] <- 1/(1-out)

out <- summary(lm(fatty[,8]~fatty[,4]+fatty[,9]+fatty[,10]+fatty[,12]))$r.sq
vifs[2] <- 1/(1-out)

out <- summary(lm(fatty[,9]~fatty[,4]+fatty[,8]+fatty[,10]+fatty[,12]))$r.sq
vifs[3] <- 1/(1-out)

out <- summary(lm(fatty[,10]~fatty[,4]+fatty[,8]+fatty[,9]+fatty[,12]))$r.sq
```

```

vifs[4] <- 1/(1-out)

out <- summary(lm(fatty[,12]~fatty[,4]+fatty[,8]+fatty[,9]+fatty[,10]))$r.sq

vifs[5] <- 1/(1-out)


modellout <- lm(fat~bmi+hip+tri+scap+sup,data=fatty)

step(modellout, list(upper = ~bmi+hip+tri+scap+sup, lower = ~tri))


out1 <- lm(fat~bmi+tri+scap,data=fatty)

out2 <- lm(fat~tri,data=fatty)

out3 <- lm(fat~bmi,data=fatty)


par(mfrow=c(2,2))

plot(modellout$fit,fatty$fat,

xlab='Fitted values: ~bmi+hip+tri+scap+sup',ylab='Body Fat')

plot(out1$fit,fatty$fat,xlab='Fitted values: ~bmi+tri+scap',ylab='Body Fat')

plot(out2$fit,fatty$fat,xlab='Fitted values: ~tri',ylab='Body Fat')

plot(out3$fit,fatty$fat,xlab='Fitted values: ~bmi',ylab='Body Fat')


plot(out1$fit,rstudent(out1),xlab='Fitted values',ylab='Studentized Residuals')

identify(out1$fit,rstudent(out1))

hist(rstudent(out1),xlab='Studentized Residuals',main="Histogram")

qqnorm(rstudent(out1))

abline(0,1)

plot(cooks.distance(out1),xlab='Observation',ylab='Cook\'s Distance')

```

```

plot(dfbetas(out1)[,1],out1$fit,xlab="Dfbetas for the intercept",
     ylab="Fitted values")
plot(dfbetas(out1)[,2],out1$fit,xlab="Dfbetas for the BMI",
     ylab="Fitted values")
plot(dfbetas(out1)[,3],out1$fit,xlab="Dfbetas for tricep skinfold",
     ylab="Fitted values")
plot(dfbetas(out1)[,4],out1$fit,xlab="Dfbetas for scapular skinfold",
     ylab="Fitted values")

summary(lm(fat ~ bmi + tri + scap, data = fatty[-74,]))

```



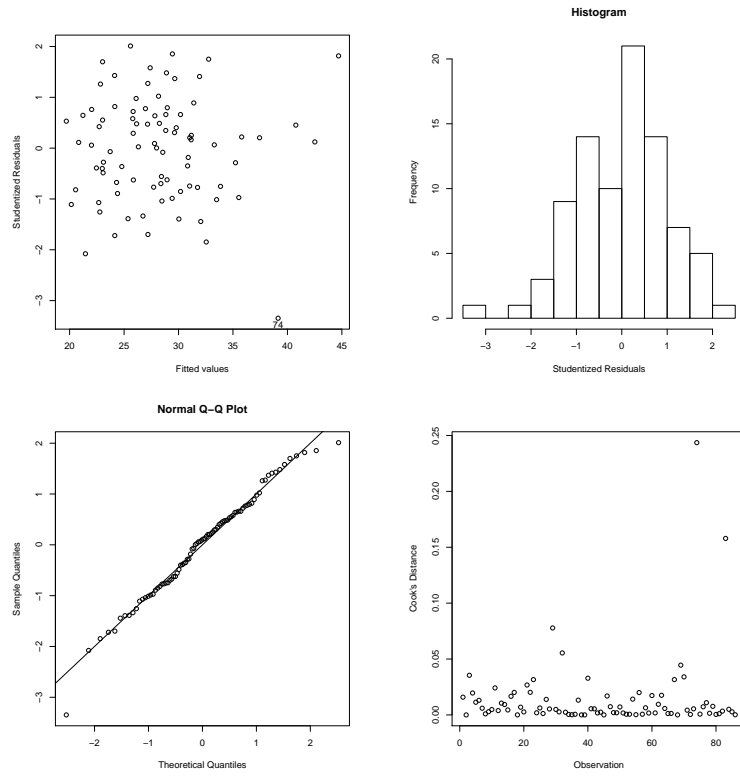


Figure 4: Residual diagnostic plots

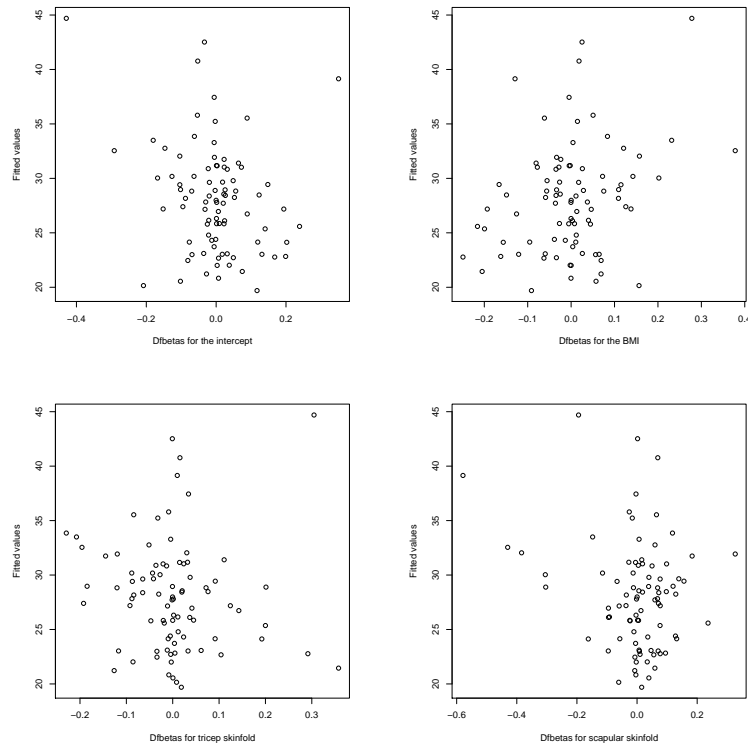


Figure 5: Influence diagnostic plots

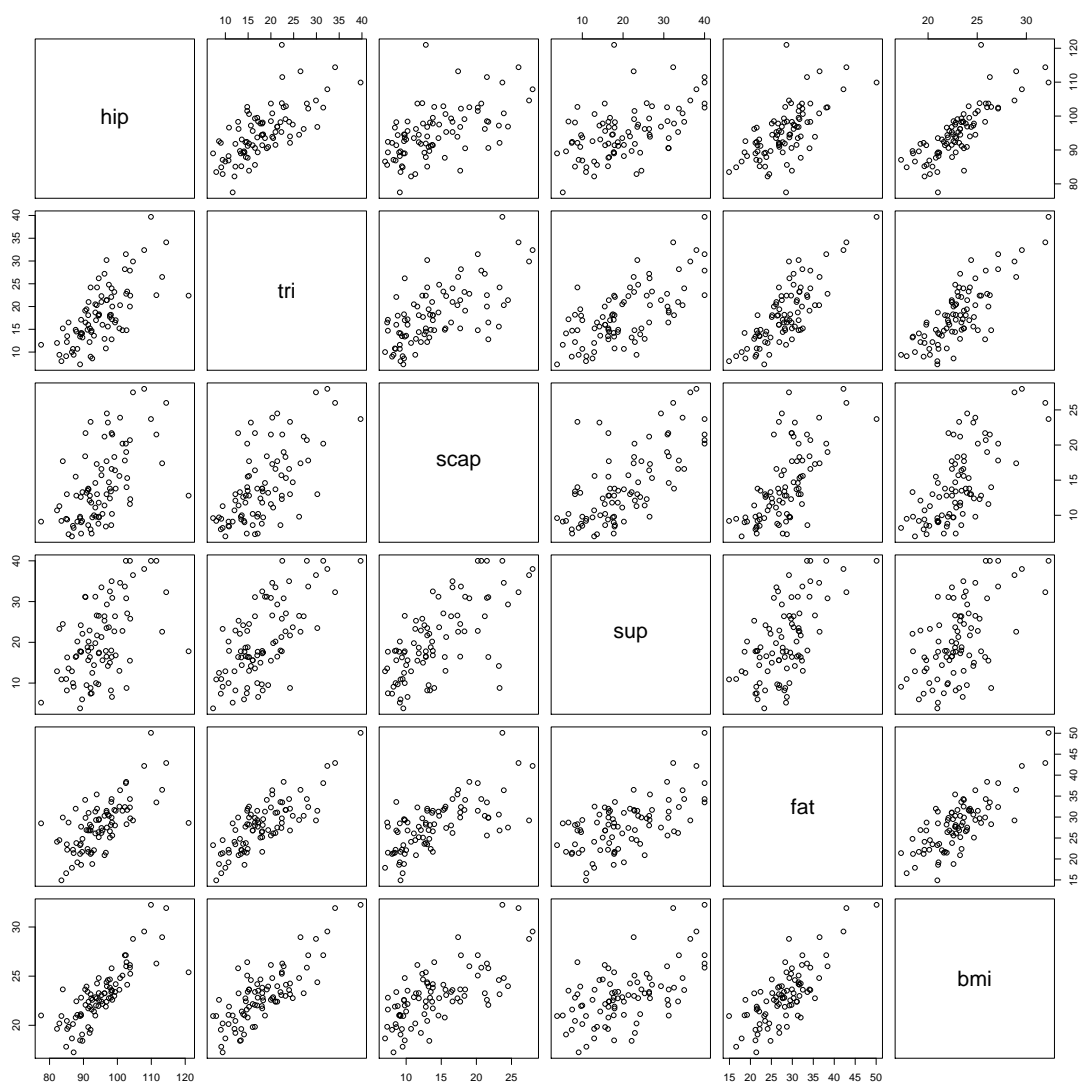


Figure 6: Scatterplot of covariates used in stepwise selection