Solutions to HW 7

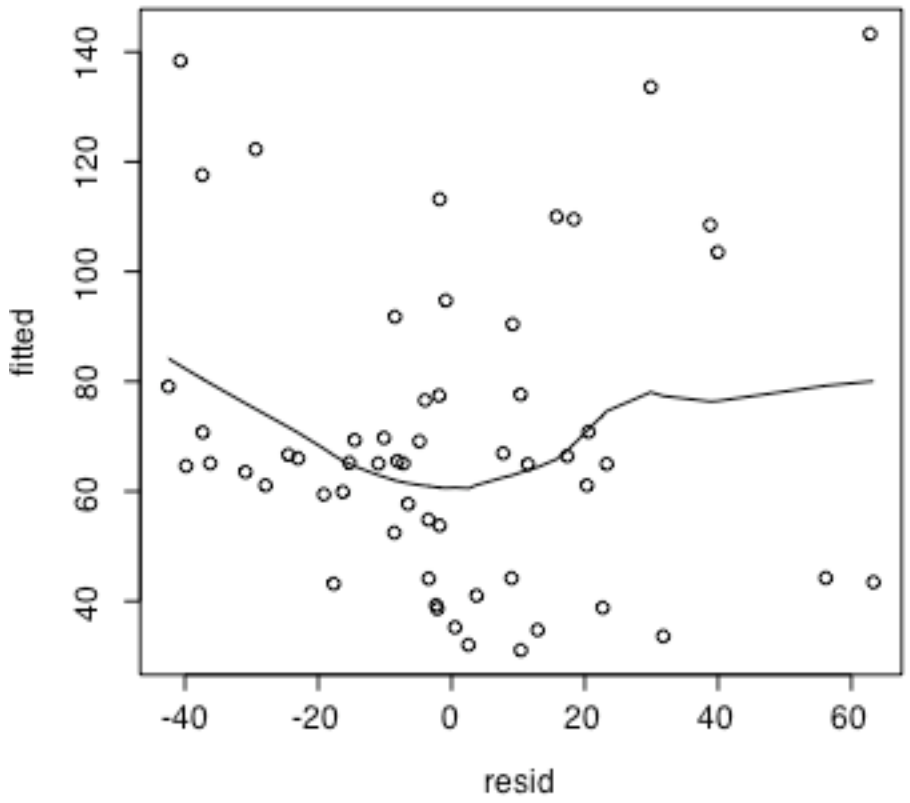
A.

i) Using the mussel data (mussels.short) , fit a linear model using food level to predict the thickness of the mussel beds. Comment on how well the data fit the model. No need to do adjustments; just check the diagnostics and comment. NOTE: to upload this data into R, you should type

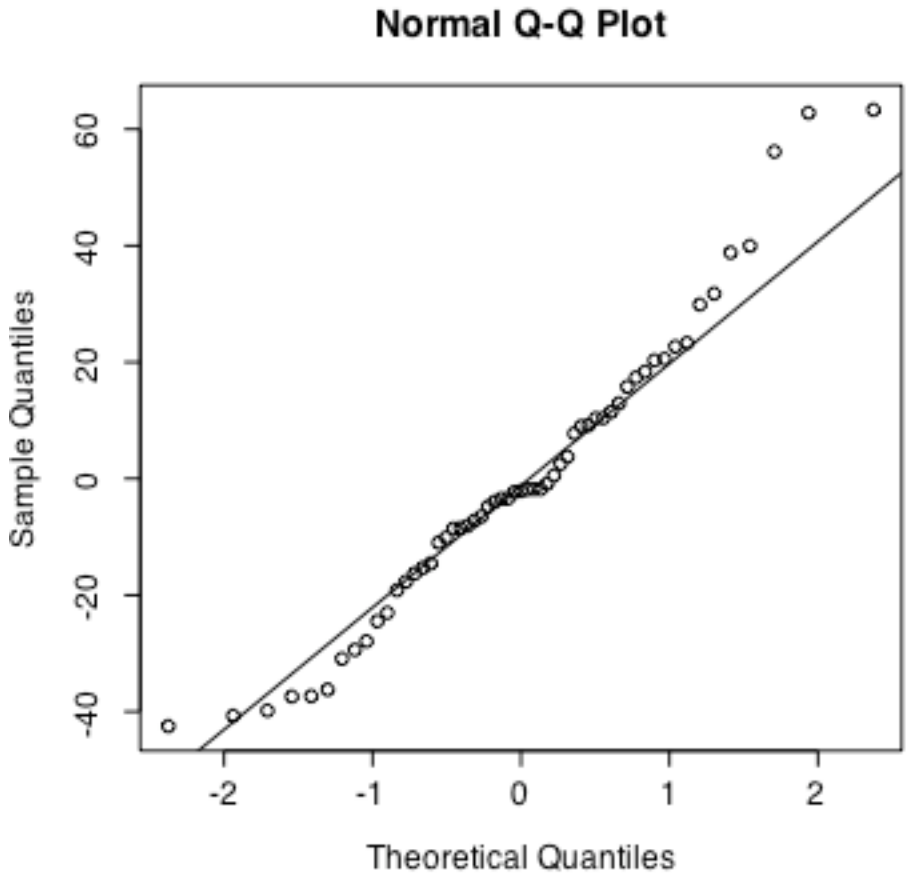
whatevernameyouwant <- read.table("mussels.short", header=T)

```
> fit <- lm(thickness~food+temp+waves+human.use)
> summary(fit)

Call:
lm(formula = thickness ~ food + temp + waves + human.use)

Residuals:
   Min     1Q Median     3Q    Max
-42.492 -15.320  -2.126  12.932  63.326

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -194.631     81.533  -2.387   0.0207 *
food           7.233      2.959   2.444   0.0179 *
temp          10.225      4.201   2.434   0.0184 *
waves         33.564      7.108   4.722 1.81e-05 ***
human.use     -4.080      3.146  -1.297   0.2005
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 25.77 on 52 degrees of freedom
Multiple R-Squared:  0.56,     Adjusted R-squared: 0.5262
F-statistic: 16.55 on 4 and 52 DF,  p-value: 8.332e-09
```

The residual plot shows a hint of a missing quadratic term, but there is also some hint of hetereskedasticity, since the variation seems to be much larger for fitted values above 60 than below.



The assumption that the residuals are normally distributed isn't perfectly sound, although this is not an important assumption. For example, it looks like there is some skew in the right-tail



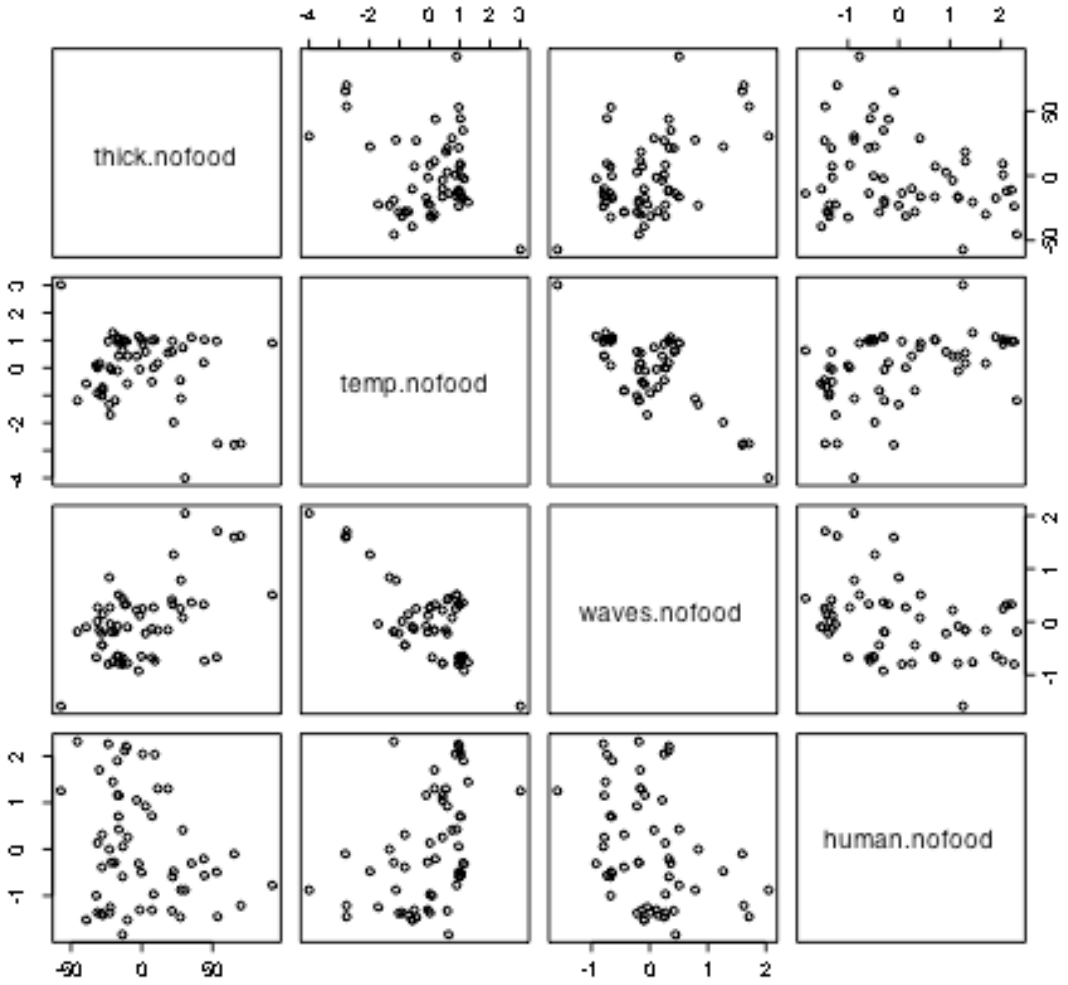
ii) Use an added variable plot to determine which of these varibles: temp, waves, human.use, will have a contribution to predicting thickness of the mussel bed if we already know the food level.

First, we find the residuals from the using food to explain thickness. Then, for each of the other predictors, we "remove" the variation due to food:

```
> fit.food <- lm(thickness~food)
> thick.nofood <- residuals(fit.food)
> temp.nofood <- residuals(lm(temp~food))
> waves.nofood <- residuals(lm(waves~food))
> human.nofood <- residuals(lm(human.use~food))
```

The added-variable plots are the first row of this matrix:

```
> pairs(cbind(thick.nofood, temp.nofood, waves.nofood, human.nofood))
```



The only clear trend seems to be with waves, suggesting that once we've included food, including waves is also useful, and suggesting that as wave levels increase, thickness increases. There might also be a negative dependence with temperature. The human use plot is hard to interpret.

We can check by seeing what the p-values and slopes actually are if we add these variables to the model with food:

```
> summary(lm(thickness~food+waves))

Call:
lm(formula = thickness ~ food + waves)

Residuals:
    Min      1Q  Median      3Q     Max
-41.909 -18.469  -3.137  14.652  80.662
```

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-9.589	13.247	-0.724	0.472	
food	3.606	2.681	1.345	0.184	
waves	22.770	5.028	4.528	3.32e-05	***

This shows that, given the food levels, knowing the wave activity is useful (because the p-value is significant.)

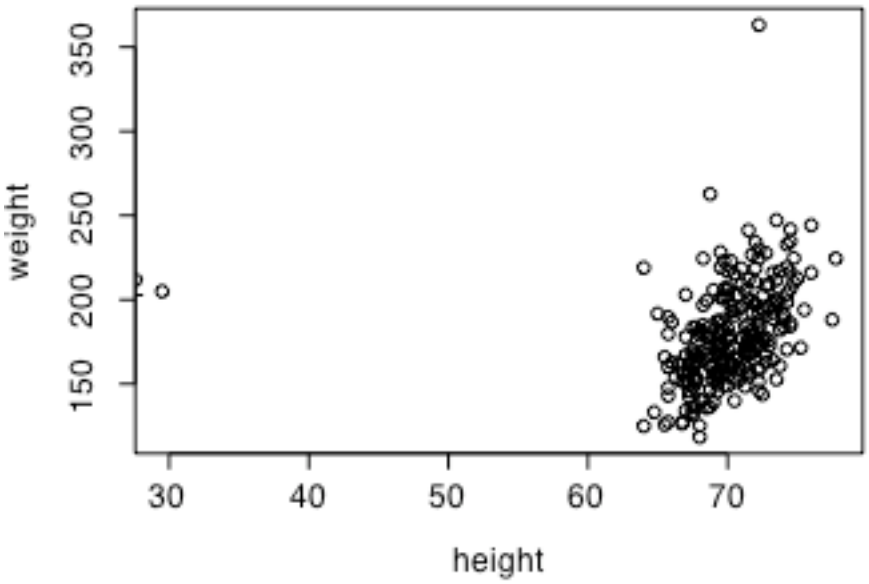
And the following shows that, given the food levels, knowing the temperature is not useful:

```
> summary(lm(thickness~food+temp))

Call:
lm(formula = thickness ~ food + temp)

Residuals:
    Min      1Q  Median      3Q     Max
-51.817 -22.456  -6.607  14.877  97.112

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  135.723     55.163   2.460   0.0171 *
food           7.469      3.444   2.169   0.0345 *
temp          -5.430      3.276  -1.657   0.1032
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 30.7 on 54 degrees of freedom
Multiple R-Squared: 0.3519,  Adjusted R-squared: 0.3279
F-statistic: 14.66 on 2 and 54 DF,  p-value: 8.207e-06
```

B

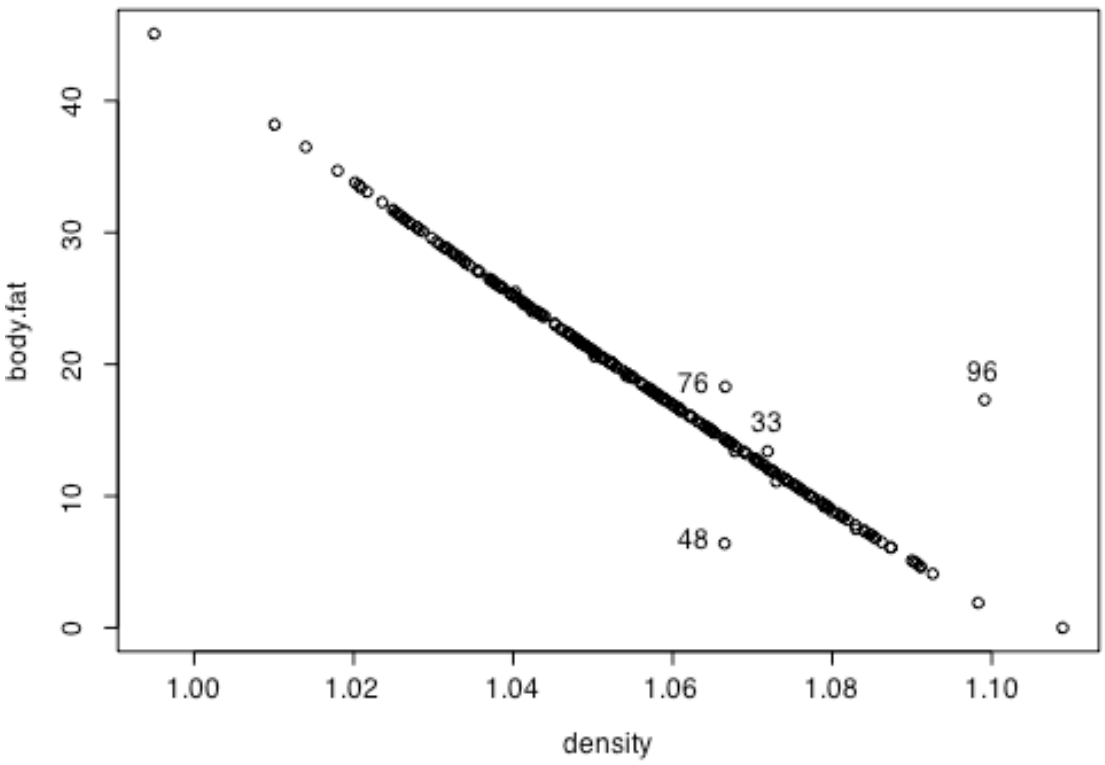
1. Examine the data and note any unusual cases. What should be done about the unusual cases? (The "about" file mentions one case in which is not hard to figure out how to correct one unusual case.)

I began with the "pairs" command to view all pairs of scatterplots. Several caught my attention:



Observations 39 and 42 are outliers. While it is not implausible that someone 70" tall could be 350 pounds, it is rather unlikely that someone 30 inches tall (2.5 feet) could weight over 200 pounds. Presumably his height was entered incorrectly. (It was entered as 29.5 inches and 69.5 would have been more in line with the data, but we will delete this point. This is observation 42.)

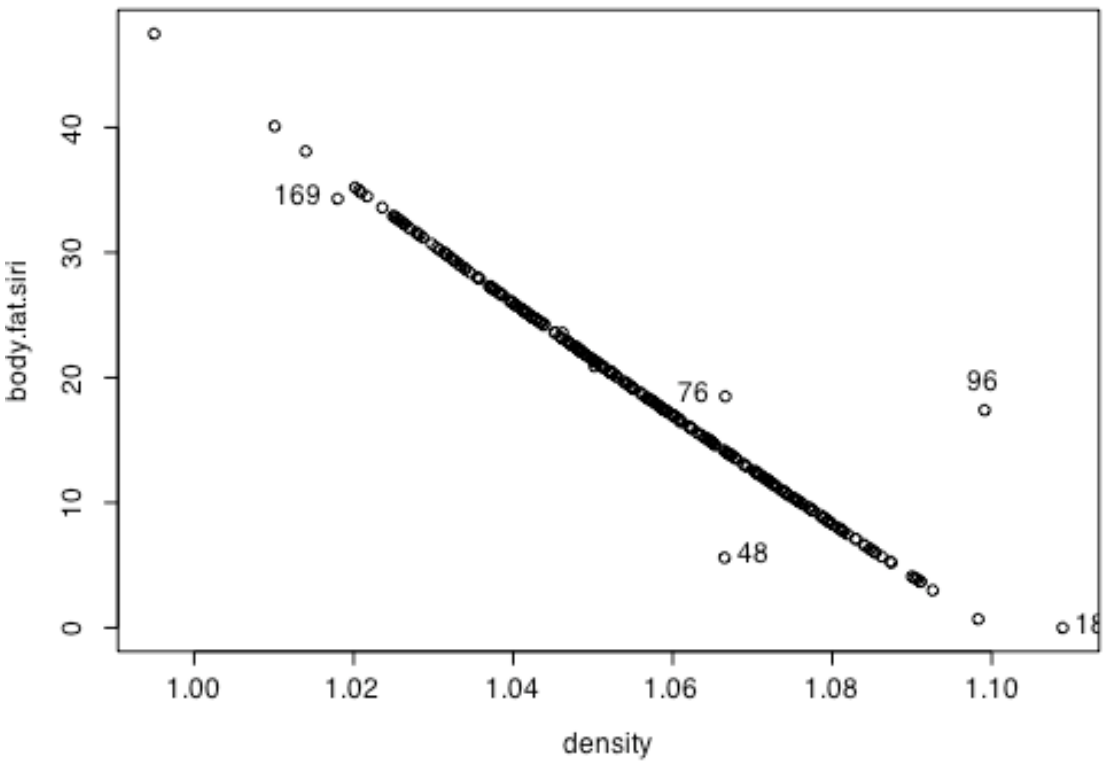
```
> fat <- fat[-42,]
```



Bodyfat should be a strict linear function of density, because it is computed from the density value. You see four observations, though, that don't fall on the line:

```
> body.fat.should <- 457/density - 414.2
> body.fat.should[c(33,48,76,96)]
[1] 12.145741 14.304454 14.264279  1.594741
> body.fat[c(33,48,76,96)]
[1] 13.4  6.4 18.3 17.3
```

These calculations show the discrepancies between what the bodyfat should be if the densities are correct and what they were actually reported as being. It's not clear, yet, whether the bodyfat or the density's are the incorrect values. However, a plot of densities against the other body fat measurement shows that it, too, has some funny values, and the same densities are to blame. This suggests that the body fats were calculated from the densities, and not the other way around. Which in turn suggests that the densities are wrong:



Presumably we could fix these, but let's take them out since I'm not sure how.

> fat <- fat[-c(48,76,96,169,18,33),]

. Choose one of the two percentage of body fat estimates (Brozek or Siri). Fit the percentage of body fat to some subset of the provided variables, but do not use density (which is too hard to measure). You need not describe every idea you try, but should justify your final choice using appropriate diagnostic tools and the like.

I chose variables that (a) seemed like they should be useful to me and (b) were not strongly correlated with each other. According to the scatterplots, many of these variables are so strongly correlated they are essentially measuring the same thing. Including variables that measure the same thing is not only redundant, but leads to instability in the calculation of parameters and makes the model difficult to interpret.

Call:
lm(formula = body.fat ~ age + weight + height + abdomen + thigh +
    bicep)

Residuals:

```
    Min      1Q  Median      3Q     Max
-10.1113  -3.0043  -0.0292   3.0601   9.2407

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -56.376650  13.257138  -4.253 3.04e-05 ***
age           0.009455   0.027517   0.344    0.731
weight       -0.194638   0.038987  -4.992 1.15e-06 ***
height        0.068069   0.152422   0.447    0.656
abdomen       0.939691   0.078931  11.905  < 2e-16 ***
thigh         0.183583   0.119381   1.538    0.125
bicep         0.218167   0.153752   1.419    0.157
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 4.11 on 238 degrees of freedom
Multiple R-Squared: 0.7259,  Adjusted R-squared: 0.719
F-statistic:   105 on 6 and 238 DF,  p-value: < 2.2e-16
```

Diagnostic plots showed that (a) the linear model seems to be a sound model (b) the normality assumption is not far off and (c) observation 37 is very influential.

A little detective work shows that 37 is the 350 pound person. Let's remove this person and see what happens:

```
Call:
lm(formula = body.fat ~ age + weight + height + abdomen + thigh +
    bicep)

Residuals:
     Min       1Q   Median       3Q      Max
-10.14607  -2.87479  -0.08961   2.97561   9.21722

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -46.07852   13.61159  -3.385 0.000832 ***
age           0.01394    0.02720   0.512 0.608786
weight       -0.15047    0.04172  -3.606 0.000378 ***
height       -0.07301    0.15899  -0.459 0.646486
abdomen       0.88038    0.08084  10.890  < 2e-16 ***
thigh         0.17404    0.11784   1.477 0.141027
bicep         0.14468    0.15406   0.939 0.348635
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 4.055 on 237 degrees of freedom
Multiple R-Squared: 0.7302,  Adjusted R-squared: 0.7234
F-statistic: 106.9 on 6 and 237 DF,  p-value: < 2.2e-16
```

Although the values of the coefficients change a bit, it doesn't seem to affect the statistical significance. It looks as if only weight and abdomen are important:

```
> fit3 <- lm(body.fat~weight+abdomen)
> summary(fit3)
```

```
Call:
lm(formula = body.fat ~ weight + abdomen)

Residuals:
    Min      1Q  Median      3Q     Max
-10.0770  -3.0034  -0.1184   2.9777   9.8712

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -42.97476    2.47704 -17.349  < 2e-16 ***
weight       -0.12576    0.01998  -6.295 1.44e-09 ***
abdomen       0.91245    0.05292  17.242  < 2e-16 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 4.065 on 241 degrees of freedom
Multiple R-Squared: 0.7243,  Adjusted R-squared: 0.722
F-statistic: 316.6 on 2 and 241 DF,  p-value: < 2.2e-16
```

A backwards, step-wise regression confirms that weight and abdomen are important, but also suggests including thigh.

```
Call:
lm(formula = body.fat ~ thigh + weight + abdomen)

Residuals:
   Min     1Q Median     3Q    Max
-9.6969 -2.8362 -0.0739  3.0158  9.5757

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -48.91999    4.06079 -12.047  < 2e-16 ***
thigh         0.18439    0.10009   1.842   0.0667 .
weight       -0.15504    0.02545  -6.091 4.42e-09 ***
abdomen       0.91503    0.05268  17.370  < 2e-16 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 4.045 on 240 degrees of freedom
Multiple R-Squared: 0.7282,  Adjusted R-squared: 0.7248
F-statistic: 214.3 on 3 and 240 DF,  p-value: < 2.2e-16
```

The diagnostic plots for this model look pretty good, so we'll stick with

body.fat = -49 + .18(thigh(-.16(weight) + .92(abdomen)

3. September 14, 1995 articles in The New England Journal of Medicine link high values of the adiposity index (weight/height^2), sometimes called the body mass index, to increased risk of premature death. See if this variable is useful in your model. Also try weight^1.2/height^3.3 as suggested in Abdel-Malek, et al. (1985

Neither of these terms seems to add anything of value to my model. Once thigh, abdomen, and weight are included, adding the BMI or this other variable doesn't seem to add useful predictive information.

```
Call:
lm(formula = body.fat ~ thigh + abdomen + weight + BMI)

Residuals:
    Min      1Q  Median      3Q     Max
-9.8840 -2.9096 -0.0945  2.9514  9.2621

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -47.93201    4.21972 -11.359  < 2e-16 ***
thigh         0.16093    0.10374   1.551    0.122
abdomen       0.87260    0.07192  12.132  < 2e-16 ***
weight       -0.15710    0.02558  -6.142 3.37e-09 ***
BMI           0.18504    0.21346   0.867    0.387


> predeath <- weight^(1.2)/height^(3.3)
lm(formula = body.fat ~ thigh + abdomen + weight + predeath)

Residuals:
    Min      1Q  Median      3Q     Max
-9.8885 -2.8781 -0.1073  2.9525  9.2042

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -47.17047    4.43312 -10.640  < 2e-16 ***
thigh         0.15864    0.10346   1.533    0.127
abdomen       0.86846    0.07081  12.265  < 2e-16 ***
weight       -0.14974    0.02602  -5.755 2.64e-08 ***
predeath    7754.51909 7879.56534   0.984    0.326
```

However, note that all of these variables are fairly strongly correlated with each other.

4. Comment on the predictive accuracy of your model. For example, how far off is an individual likely to be?

The residual standard error is 4.04, which means, roughly speaking, that individuals are likely to be within 4% of their true body-fat percentage, across all individuals. Given an

individuals actual measurements, we can calculate a prediction interval to get a better idea of how far off their true body-fat will be from the predicted body-fat. But as a rough guide, about 95% of the subjects will be within 8% of their true body-fat --- which isn't terribly good.

5. Estimate the percentage of US men whose bodyfat is less than 15% (which some experts say is the maximum for good health). What assumptions must you make?

Assuming the data represent a random sample of US men, then about 32% of the sample had body-fat less than 15%:

```
> length(body.fat[body.fat<15])/length(body.fat)
[1] 0.3155738
```

Our regression model tells us that for any given weight/thigh/abdomen measurements, this percentage will be different, but ignoring these terms, across the population we'd estimate that about 32% are under 15% bodyfat.