M5MS03 - Applied Statistics; Coursework 3 - Autumn 2018

Question 1

Provide a short description of the dataset mentioning briefly (a) what you believe the experimental procedure was, (b) any interesting features from a statistical perspective.

The `fat` dataset is summarised by `help(fat)` as

> Age, weight, height, and 10 body circumference measurements for 252 men. Each man's percentage body fat was accurately estimated by an underwater weighting technique.

Body fat is a variable of interest because it is correlated with other health measures. This particular dataset seems to have been collected on American men in Utah during the 1980s, with the intention of developing a predictive model for lean body weight[1]. Each of its features is described in Table 1 (see end of document). No dataset values were missing. The majority of the measurements are self-explanatory. `brozek` and `siri` are two methods of estimating % body fat based on the density of a person's solid matter (feature `density`).

The experimental method used to estimate density is likely to be hydrostatic weighting, wherein a person's density is estimated by considering four factors: their 'dry' mass m_d as measured on land, their 'wet' mass m_w as measured in water, the density of the water in which they are measured, ρ_w, and the volume occupied by cavities in the person's body, V_r. Essentially, Archimedes' principle can be applied to reason that the density of a person's solid matter can be estimated according to:

$$\rho = \frac{m_d}{\frac{m_a - m_w}{\rho_w} - V_r} \tag{1}$$

where ρ denotes solid-matter body density. The largest errors in this estimate are likely to be attributable to the value chosen for V_r, which is usually taken as a proportion of the 'maximum amount of air a person can expel from their lungs after a maximum inhalation'[2]. No indication is given in the dataset's description as to how this V_r was estimated, whether by experiment or by formula. For the purposes of analysis we will assume that the values for `density` are accurate estimates of the subjects' body densities.

The dataset has several interesting features of this dataset from a statistical perspective.

- Some features are strongly correlated with one another, while others are uncorrelated. This is unsurprising: everyday experience tells us that body dimensions, both mass and weight, are correlated. A correlation heatmap is provided in Figure 1 - it reveals that large body dimensions are negatively correlated with density and higher weight, and that all of the dimension measurements are strongly correlated.
- The dataset contains $n = 252$ examples and $p = 18$ features.
- The dataset contains one integer-valued feature, `age`. Otherwise its features are positive-valued and continuous (ignoring measurement resolution). The range of `brozek` and `siri` is restricted to $(0, 100)$.
- The feature distributions are unimodal without exception, as is shown in Figure 2. Many features contain extreme values on one side only.

Question 2

Identify which variables in the data you think could be used as the response.

[1]The specific paper is 'Generalized body composition prediction equation for men using simple measurement techniques', K.W. Penrose, A.G.Nelson, A.G. Fisher, FACSM, Human Performance Research Center, Brigham Young University, Provo, Utah 84602 as listed in Medicine and Science in Sports and Exercise, vol. 17, no. 2, April 1985, p. 189.

[2]This definition is courtesy of Wikipedia's 'Vital capacity' page.

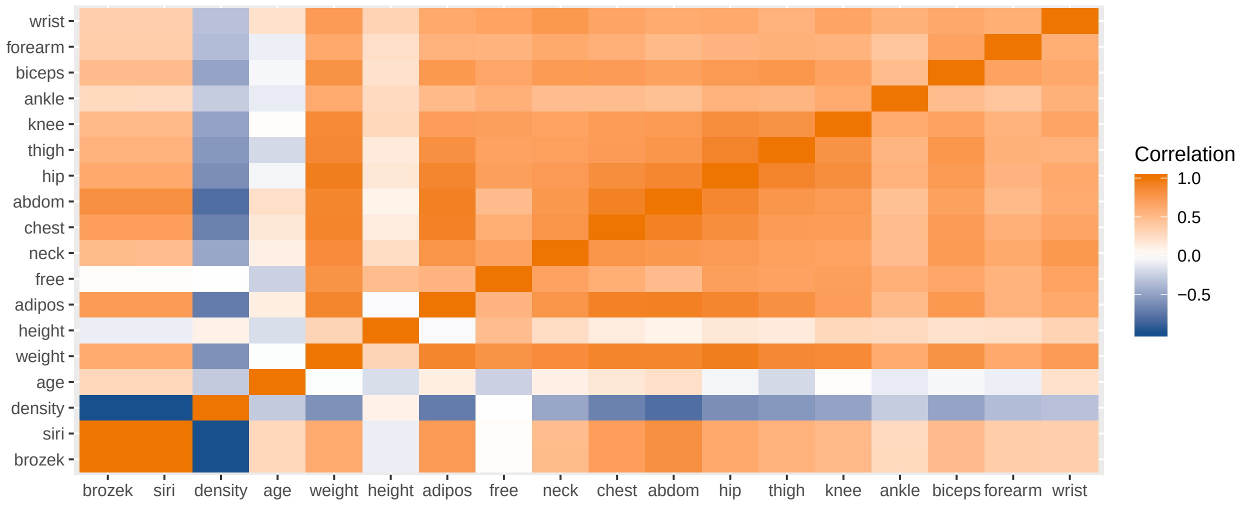

Figure 1: Correlation heatmap for the feature vectors in the `fat` dataset.

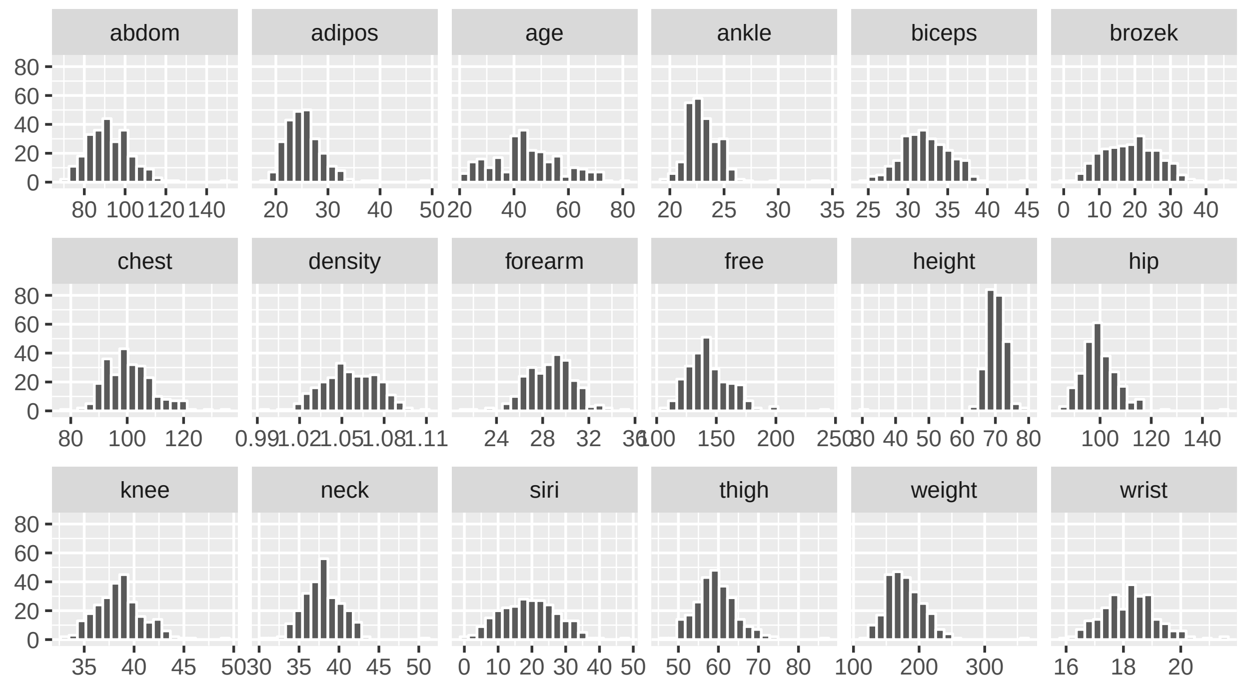

Figure 2: Frequency distributions for each feature in the dataset. These are unimodel, with each feature possessing at least one extreme value (e.g. the `height`<35in and `hip`>140cm).

In the original paper associated with this dataset, a predictive model was created for lean body weight by conducting a stepwise regression. The generated model had the form

$$w_l = 17.298 + .89946w - .2783a + .002617a^2 + 17.819h - .6798(\mathrm{Ab} - \mathrm{Wr}) \tag{2}$$

This model will be referred to as 'Penrose and Fisher's model'. w_l is lean body weight (body weight minus body fat weight), w is weight in kilograms, h is is height in meters, a is age in years, and Ab and Wr are the circumferences of the abdomen and wrist respectively ($R^2 = .924$).

Based on this context, lean body weight (a.k.a. fat-free body weight, `free`) seems like an appropriate response variable. It will be interesting to see whether the model developed by the authors of the dataset's paper is optimal for the purposes of prediction.

The original authors had access to `height`, `weight`, `age`, and the ten circumferences in the dataset. To enable a fair comparison, the other features were deleted. The units of `height` and `weight` were converted from imperial measurements to metric ones. To enable comparison of the practical significance of the regression coefficients, the sample distributions of the features and response were centered on zero and linearly scaled to have unit variance.

```
retained_features <- c('age', 'weight', 'height', 'free', 'neck', 'chest',
                       'abdom', 'hip', 'thigh', 'knee', 'ankle', 'biceps',
                       'forearm', 'wrist')
fat$weight <- fat$weight/2.204622 # lb to kg
fat$free   <- fat$free/2.204622 # lb to kg
fat$height <- fat$height*2.54/100 # into m
Fat <- data.frame(scale(fat))
brozek <- Fat$brozek # used later in report
Fat <- Fat[retained_features]
```

Question 3

For one suitable choice for the response, construct a model based on a single set of predictor variables (without interactions) that you select. Summarise the model output, and make an initial recommendation on the results.

In spite of the strong correlations between features, we consider a normal linear model that uses all of the predictor variables. This model has the form

$$y = X\beta + \epsilon : \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{3}$$

We fit this model in `R` on the full dataset and inspect a summary of the result.

```
Fat.lm1 <- lm(free ~ . - 1, data=Fat) # response has mean zero
summary(Fat.lm1)
```

```
##
## Call:
## lm(formula = free ~ . - 1, data = Fat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.21316 -0.27013  0.01804  0.25248  1.46100
##
## Coefficients:
##          Estimate Std. Error t value Pr(>|t|)
## age      -0.06767    0.03741  -1.809 0.071734 .
## weight    1.26011    0.14436   8.729  4.5e-16 ***
## height    0.07532    0.03227   2.334 0.020429 *
## neck      0.10276    0.05186   1.982 0.048672 *
## chest     0.07172    0.07670   0.935 0.350717
## abdom    -0.90857    0.08554 -10.621  < 2e-16 ***
## hip       0.12067    0.09592   1.258 0.209620
## thigh    -0.04527    0.06955  -0.651 0.515724
## knee      0.04348    0.05356   0.812 0.417662
## ankle    -0.01481    0.03445  -0.430 0.667710
## biceps   -0.02700    0.04744  -0.569 0.569894
## forearm  -0.04669    0.03692  -1.264 0.207313
```

```
## wrist    0.15638    0.04583   3.412 0.000757 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3951 on 239 degrees of freedom
## Multiple R-squared:  0.8514,Adjusted R-squared:  0.8433
## F-statistic: 105.3 on 13 and 239 DF,  p-value: < 2.2e-16
```

The F-statistic's observed value has probability $< 2.2e-16$ under the null hypothesis that the feature coefficients are all equal to zero. We can therefore safely reject the null and conclude that some of the features are explanatory. The t-statistics for each regression coefficient indicate that, in order of decreasing significance, `abdom`, `weight`, `wrist`, `height`, and `neck` have significantly non-zero coefficients at the 5% level. The t-statistic for the `age` coefficient is slightly below this significance threshold ($p(t) \approx 0.07$). The other features have coefficient values that are relatively non-significant (i.e. they satisfy $p(t) > 0.2$). Inspecting the ANOVA table for this model (omitted) also reveals that only these variables yield a statistically significant reduction in residual-sum-of-squares upon being included in a nested linear model.

The practical significance of each feature can be evaluated directly as a result of earlier standardisation. Restricting the analysis to statistically significant features, we see that `weight` and `abdom` have a large effect on a person's lean body weight, whereas the other features have relatively minor effects:

```
##  abdom    age height   neck  wrist weight
##  -0.91  -0.07   0.08   0.10   0.16   1.26
```

These coefficients can be interrogated further. `abdom` would be negatively correlated with fat-free body weight since a large abdomen circumference suggests that a person is fat. The fact that a large abdomen implies a larger overall weight (and therefore a higher fat-free weight) is controlled for because the multiple regression includes `weight` as a predictor (i.e. the `abdom` coefficient is the change in scaled free-body weight while holding `weight` fixed). In line with this, the large positive coefficient of `weight` higlights that fat-free weight increases with overall weight. By comparison, the values of the other coefficients - those for `age`, `height`, `neck`, `wrist` - are less easily explained.

Reassuringly, the coefficients of the variables deemed non-explanatory are generally quite small. This indicates that they would not dramatically affect a prediction were they to be included in the model.

My initial recommendation is that we drop the variables that are clearly non-significant and retain the ones that are significant. Consequently, we re-fit a linear model using only `age, weight, height, abdom, neck,` and `wrist` as predictors.

Question 4

Perform a set of model diagnostics, to assess the adequacy of the model.

The following model diagnostic plots were generated:

- Studentized[3] residuals versus fitted values - checking for independence, homoscedasticity, and nonlinearity.
- QQ-plot of the studentised residuals, to assess whether they are plausibly normal.
- Added variable plots for the set of significant features, again to assess linearity.
- Inspection of the leverage values, studentized residual magnitudes, and the Cook's distance of each datapoint, with the intention of identifying possible outliers.

The studentized residuals versus fitted values plot, shown in Figure 3, suggests that the residuals are not correlated with the fitted values, as would be expected under the model specification. Furthermore,

[3]'Studentised' in this document refers exclusively to externally studentised (i.e. jackknife) residuals.

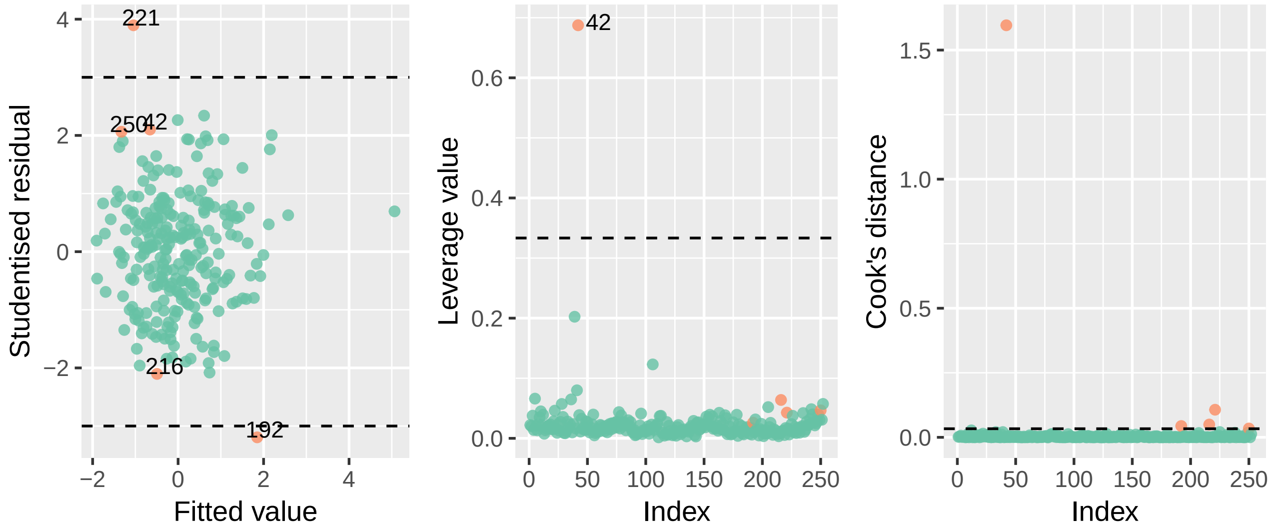

Figure 3: (Left) Studentized residuals verus fitted values. (Center) Leverage ('hat') values versus entry index. (Right) Cook's distance versus entry index. Thresholds are dashed lines at $\pm 3, 2/p$, and $\frac{8}{n-2p}$. Labels are omitted from the final plot for clarity - the 'high' point is `42`.

there is no evidence of nonlinearity or heteroscedasticity. The Cook's distance for each plot indicates a collection of points that might be considered outliers, those with a distance greater than $\frac{8}{n-2p}$, where $p = 6$. Foremost amongst these is entry `42`, which has both a large residual and a large leverage. Let's look at this entry in the original dataset.

```
print(fat[42, c(sig_coef_names, 'free')])

##    age   weight height neck abdom wrist     free
## 42  44 92.98646 0.7493 36.6 104.3  17.4 63.54831
```

This fourty-four year-old man is apparently 75cm tall and weighs 93kg. Either we are working with an obese dwarf or there is a data entry error. Either way, this point will be excluded and the model will be re-fit. But what of the other entries with high Cook's distances?

```
print(fat[192, c(sig_coef_names, 'free')])

##     age weight height neck abdom wrist     free
## 192  42 110.79 1.9304 41.8 113.7  19.9 70.39756

print(fat[221, c(sig_coef_names, 'free')])

##     age   weight height neck abdom wrist     free
## 221  54 69.51305 1.7907 38.5  91.8  18.9 68.62854
```

The `weight` median for the complete dataset is 81.2kg, with standard deviation 13.3kg. Entry `192` is therefore an outlier in terms of weight. Entry `221` is an outlier because their calculated free weight is remarkably low relative to their weight, as can also be seen from the added variable plots in Figure 4

The qq-plot of the studentised residuals (omitted) indicated that the model's residuals are approximately normally distributed, suggesting this part of the specification is correct. Furthermore, the linearity assumption also seems justified given the added variable plot - no distinct nonlinear trend exists. This seems to contravene including an order-2 (i.e. squared) age term as was in Penrose and Fisher's model.

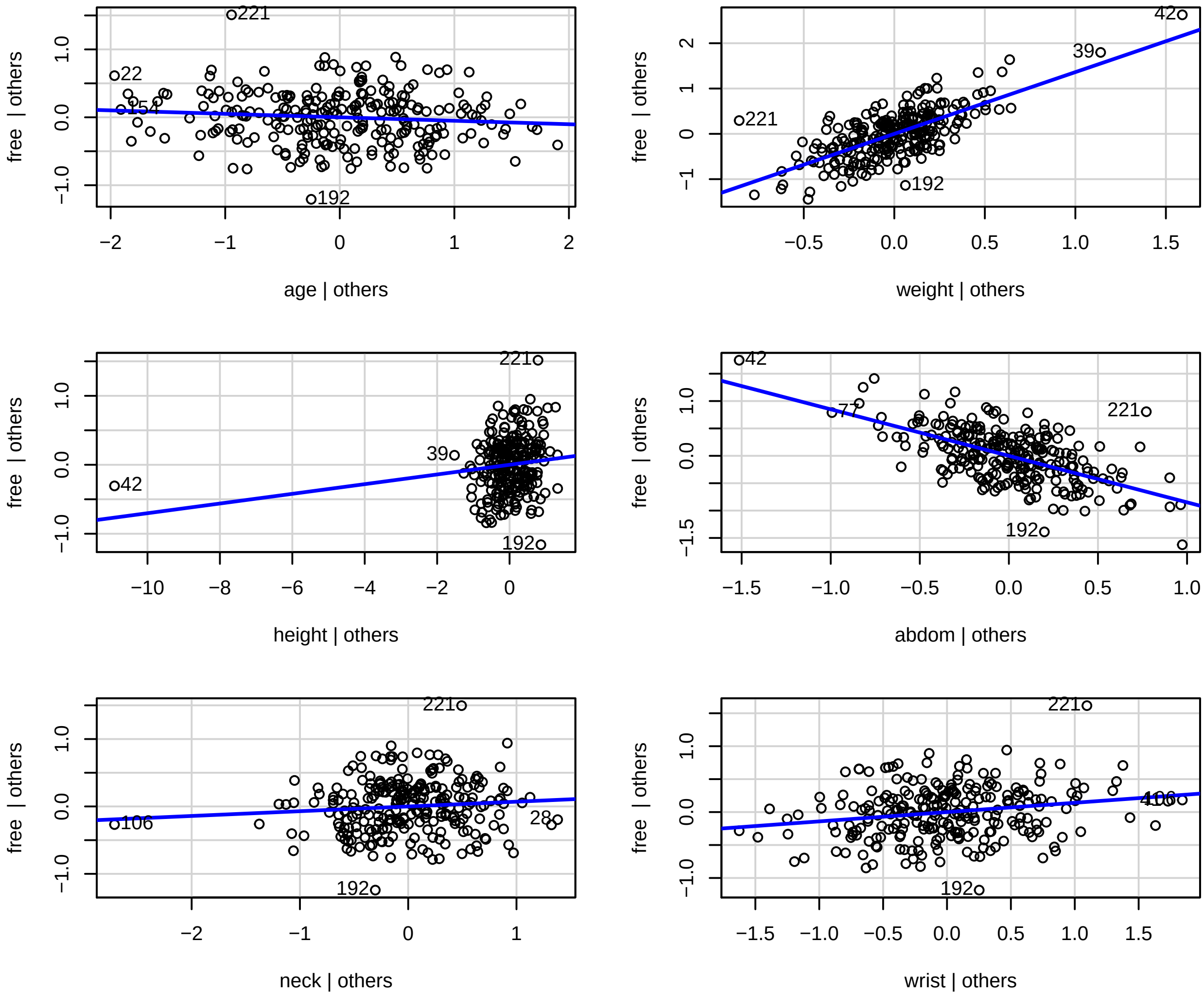

Figure 4: Added variable plots for the linear model fit using only variables deemed significant.

Question 5

Given the diagnostics above, conduct appropriate model refinement and recompute the model

Based on considerations listed in the previous section, particularly Cook's distance, the points `31, 42, 86, 192, 216,` and `221` were omitted from the dataset before refitting the model.

The change in model coefficients is provided below as a fraction of the outlier-included coefficient estimate. We can see that omitting the outliers has substantially increased the `height` coefficient estimate and has mildly decreased the `age` coefficient estimate. The change in `height` effect is likely to be attributable to omitting the 'dwarf', entry `42`.

```
##    age weight height  abdom   neck  wrist
##  -0.11  -0.02   0.97   0.04   0.06  -0.08
```

Question 6

Comment on potential advantages and disadvantages for this model.

Summary statistics for the outlier-omitted model are $R^2 = 0.87$ and a residual standard error of 0.37.

The key advantage of the proposed model is its simplicity and the fact that the inclusion of each explanatory variable is well-motivated by considerations of statistical and practical significance. The model can be explained verbally (with each coefficient corresponding to the change in fat-free weight per unit change in the predictor, all other predictors being held fixed). The R^2 value for the proposed model is marginally lower than that of Penrose and Fisher (they obtained $R^2 = 0.92$). It seems plausible that their model was overfit given their methodology (stepwise regression that included quadratic terms). It could be argued that `height` would be better replaced by its cube[4], however plotting fat-free weight against height provides little support for anything other than a linear relationship. Moreover, interaction terms have not been considered - this is on the basis that no physical motiviation for them could be identified, they would substantially increase the risk of falsely including a spuriously predictive term, and they would cause p (182 w/ interactions) to approach n (252), making overfit more likely.

Disadvantages of this model are that it is high-dimensional, which makes it difficult to visualize, and it does not account for the existence of the outliers listed earlier. It is also specific to American men - its validity outside of the sample population is unknown. The omission of interaction terms has already been discussed - future work could investigate whether these provide a meaningfully improved fit.

Fitting this model in the original feature space (i.e. that used by Fisher and Penrose), once more with outliers removed, produces

$$w_f = 4.93 + .825w - .037a + 12.4h - .623\text{Ab} + .256\text{Ne} + 1.147\text{Wr} \tag{4}$$

Compare this with Penrose and Fisher's model:

$$w_l = 17.298 + .89946w - .2783a + .002617a^2 + 17.819h - .6798(\text{Ab} - \text{Wr}) \tag{5}$$

Future analyses should investigate the differences in the coefficients above - there is not sufficient time now, but the conflict between the signs of a and a^2's coefficients may be cause for concern.

Question 7

Identify other possible choices for the response. Using a similar analysis to the one you did in parts 3.-4. check whether the explanatory variables in part 3. are still valid. Comment on the result.

[4] Given that weight scales with volume, and volume scales with the cube of length (modeling people as cuboids).

An alternative choice for the response would be percentage body fat, either `brozek` or `siri`. The dataset's documentation reveals that `free` was calculated according $(1 - \texttt{brozek})\texttt{weight}$. `brozek` is therefore a linear function of `free` but is non-linear with respect to `weight`. `free` and `brozek` also have near-zero correlation ($\rho = 0.02$). It is therefore difficult to anticipate whether the same features will be valid as predictors of `brozek`.

```
Fat.no$brozek <- brozek[-c(to_remove)]
Fat.brozlm <- lm(brozek ~ age + weight + height + neck + abdom + wrist - 1, data=Fat.no)
round(coef(summary(Fat.brozlm)), 3)
```

```
##        Estimate Std. Error t value Pr(>|t|)
## age       0.049      0.043   1.130    0.260
## weight   -0.294      0.143  -2.059    0.041
## height   -0.045      0.068  -0.653    0.514
## neck     -0.070      0.065  -1.063    0.289
## abdom     1.191      0.109  10.939    0.000
## wrist    -0.127      0.057  -2.219    0.027
```

```
anova(Fat.brozlm)
```

```
## Analysis of Variance Table
##
## Response: brozek
##            Df Sum Sq Mean Sq  F value    Pr(>F)
## age         1 19.706  19.706  73.7103 1.134e-15 ***
## weight      1 83.952  83.952 314.0188 < 2.2e-16 ***
## height      1 18.692  18.692  69.9158 4.991e-15 ***
## neck        1  3.969   3.969  14.8475 0.0001495 ***
## abdom       1 37.590  37.590 140.6060 < 2.2e-16 ***
## wrist       1  1.317   1.317   4.9249 0.0274041 *
## Residuals 241 64.430   0.267
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results in these tables highlight that the variables would still yield a reduction in variance according to the F-tests we applied when answering Question 3, but could not easily be said to be non-zero under the t-tests. This combination of diagnostics indicates multicollinearity - the variables are significant in combination, but not in isolation. Inspecting the correlation heatmap reveals that `weight` is strongly correlated with all predictors apart from `age`. Re-fitting the model with `weight` omitted yields t and F-statistics that are in agreement regarding the significance of the remaning variables (i.e., that all are statistically significant).

Question 8

> *Considering the model in 5., generate 3 new data instances (predictor variable and response) that have, respectively (i) high residual variance and low leverage, (ii) high leverage and low residual variance, and (iii) high influence. Explain your rational for generating the points. These can be derived from existing points, or otherwise.*

Under the specified model, the residual variance is $\sigma^2(I - H)$, where H denotes the hat matrix, $H = X(X^TX)^{-1}X^T$. The leverage of point i in this matrix is $h_i = H_{ii} = (X(X^TX)^{-1}X^T)_{ii}$. Note that it can be shown $0 \leq h_{ii} \leq 1$. It follows that a point with high leverage will necessarily have low residual variance. We can randomly generate an x vector, compute its corresponding h_{ii}, then accept or reject it according to its leverage value. We use an arbitrary threshold of $2/p = 1/3$ ($p = 6$).

```
library(MASS)
X0 <- model.matrix(Fat.lm4)
```

```r
h  <- 0
while(h < 1/3){ # could definitely be sped up
  x  <- matrix(mvrnorm(n=1, mu=colMeans(X0), Sigma=10*cov(X0)))
  X1 <- rbind(X0, t(x))
  H  <- X1 %*% ginv(t(X1) %*% X1) %*% t(X1)
  h  <- H[248, 248] # leverage
}
print(h) # print leverage
```

```
## [1] 0.3717456
```

```r
print(t(x)) # High leverage point
```

```
##           [,1]     [,2]       [,3]     [,4]      [,5]     [,6]
## [1,] -8.662107 2.368115 -0.4352294 2.051909 0.4298769 1.955905
```

```r
while(h > 1/12){ # low leverage, high residual variance point
  x  <- matrix(mvrnorm(n=1, mu=colMeans(X0), Sigma=10*cov(X0)))
  X1 <- rbind(X0, t(x))
  H  <- X1 %*% ginv(t(X1) %*% X1) %*% t(X1)
  h  <- H[248, 248] # leverage
}
print(h)
```

```
## [1] 0.05982127
```

```r
print(t(x)) # High residual variance point
```

```
##            [,1]       [,2]    [,3]      [,4]      [,5]     [,6]
## [1,] -0.4857404 -0.7812352 1.70157 -1.502337 -1.589377 -1.54464
```

There was insufficient time to generate a high-influence point, however Cook's distance would have been a suitable criterion for measuring influence (since it would be proportional to the change in the re-fitted coefficients when the point was omitted). It would have been implemented using similar code to that provided above.

Table 1: Descriptions of each feature in the `fat` dataset.

Feature	Description
`brozek`	Percent body fat using Brozek's equation, 457/Density - 414.2
`siri`	Percent body fat using Siri's equation, 495/Density - 450
`density`	Density (gm/cm^3)
`age`	Age (yrs)
`weight`	Weight (lbs)
`height`	Height (inches)
`adipos`	Adiposity index = Weight/Height^2 (kg/m^2)
`free`	Fat Free Weight = (1 - fraction of body fat) * Weight, using Brozek's formula (lbs)
`neck`	Neck circumference (cm)
`chest`	Chest circumference (cm)
`abdom`	Abdomen circumference (cm) at the umbilicus and level with the iliac crest
`hip`	Hip circumference (cm)
`thigh`	Thigh circumference (cm)
`knee`	Knee circumference (cm)
`ankle`	Ankle circumference (cm)
`biceps`	Extended biceps circumference (cm)
`forearm`	Forearm circumference (cm)
`wrist`	Wrist circumference (cm) distal to the styloid processes