RESEARCH SCHOOL OF
FINANCE, ACTUARIAL STUDIES AND STATISTICS
College of Business & Economics, The Australian National University

REGRESSION MODELLING

(STAT2008/STAT4038/STAT6038)

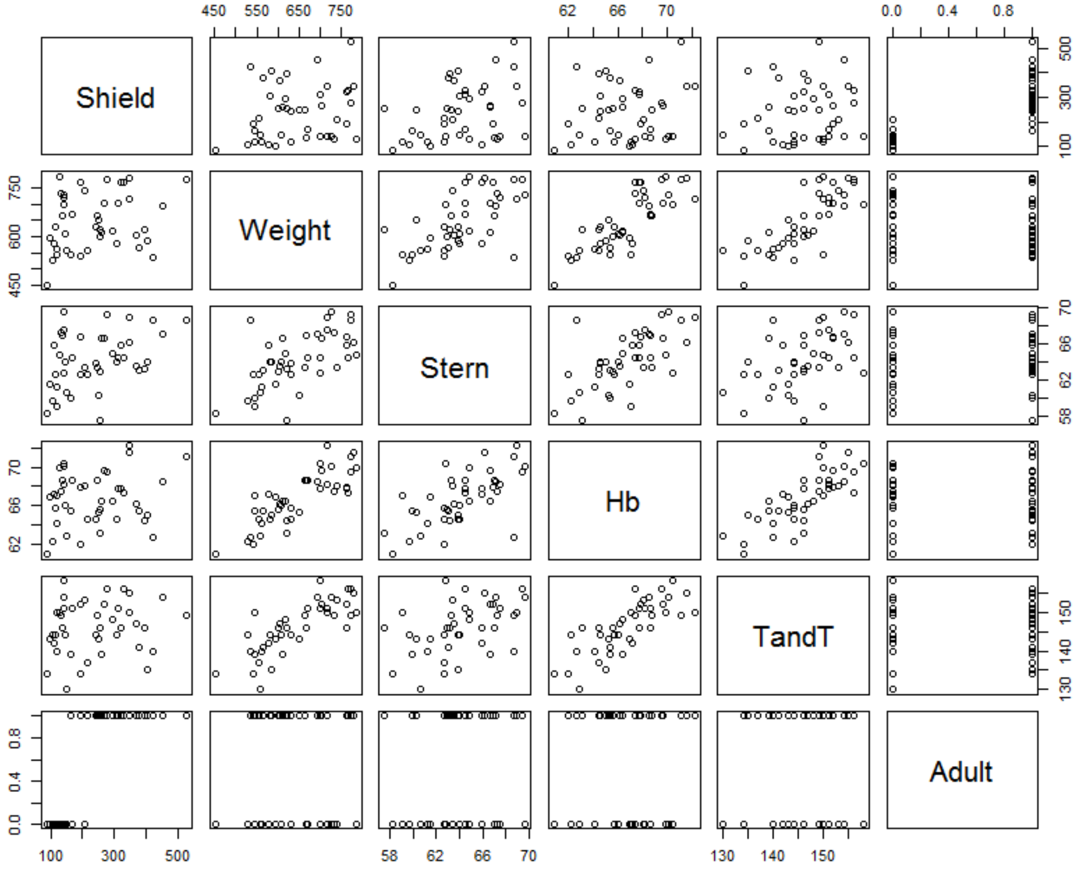
Solutions to Assignment 2 for 2017

Question 1 **(20 marks)**

Following on from the analysis of the moorhen data in Question 1 of Assignment 1, we would like to use all available variables to try and build a multiple regression model with Shield area as the response variable. The e-mail from the scientists that came with the data doesn't really describe the variables Stern, Hb and TandT, except to say that they are "three lineal measurements" taken on each bird. Adult is an indicator of whether the bird is a juvenile (0) or adult (1) bird.

Use R to further analyse the moorhen data and answer these questions:

(a) Produce both a scatterplot matrix and a correlation matrix for the variables in the moorhen data and comment on any important relationships between the variables (assuming you are planning to do a multiple regression analysis). **(3 marks)**



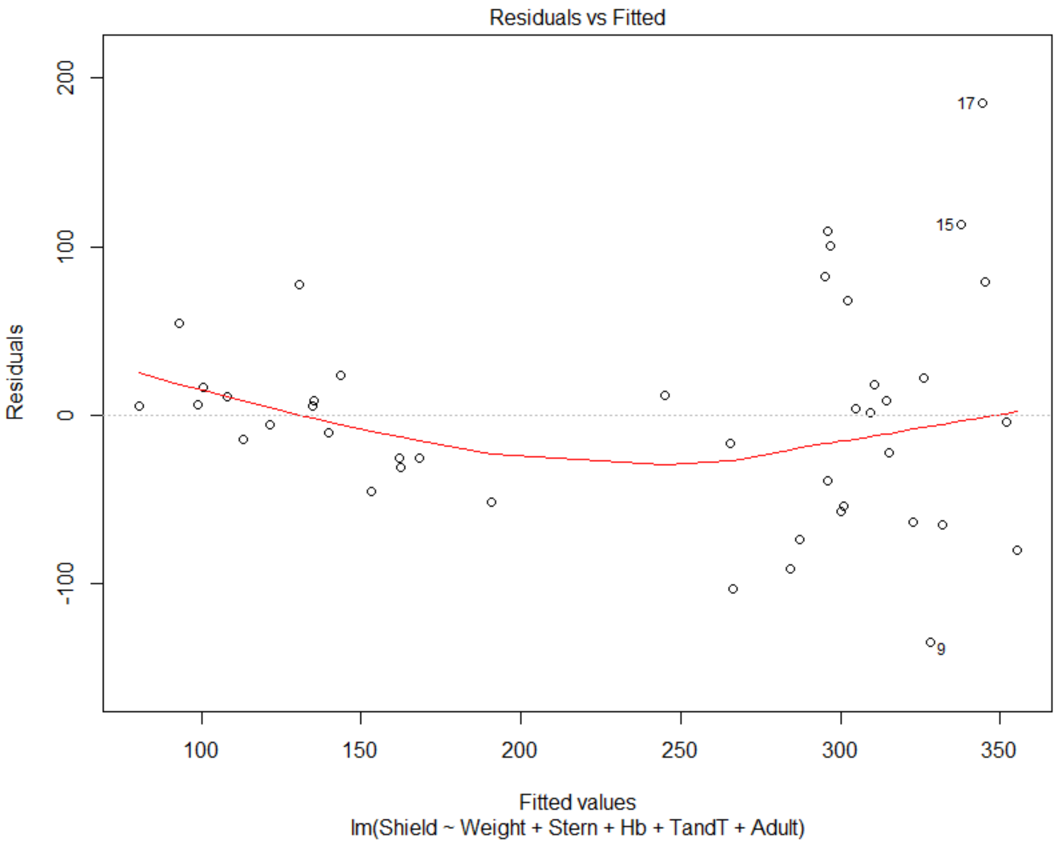
```
> cor(moorhen)
          Shield    Weight     Stern           Hb       TandT        Adult
Shield 1.0000000 0.2394694 0.3818278  0.171113116 0.144948682  0.782786730
Weight 0.2394694 1.0000000 0.6350777  0.826493514 0.793679060  0.100761751
Stern  0.3818278 0.6350777 1.0000000  0.644056172 0.461534419  0.176030285
Hb     0.1711131 0.8264935 0.6440562  1.000000000 0.782295402 -0.008168973
TandT  0.1449487 0.7936791 0.4615344  0.782295402 1.000000000  0.004246455
Adult  0.7827867 0.1007618 0.1760303 -0.008168973 0.004246455  1.000000000
```

Question 1, part (a) continued

The relationships between the response variable, Shield, and the potential explanatory variables do not look particularly promising. The strongest relationship is with Stern, but the correlation is only 0.38 and there appears to be signs of increasing variance in the plot.

On the other hand, there are some relatively strong relationships between some of the explanatory variables, notably Hb and TandT (correlation 0.78) and between Weight and all three of the other continuous predictors, suggesting that not all of these variables will be useful predictors in the same multiple regression model.

(b) Fit a multiple linear regression model with Shield as the response variable and with all the other variables in the data as explanatory variables. Present the main residual plot of the residuals against the fitted values for this model. Are there are any obvious problems with underlying assumptions? **(3 marks)**

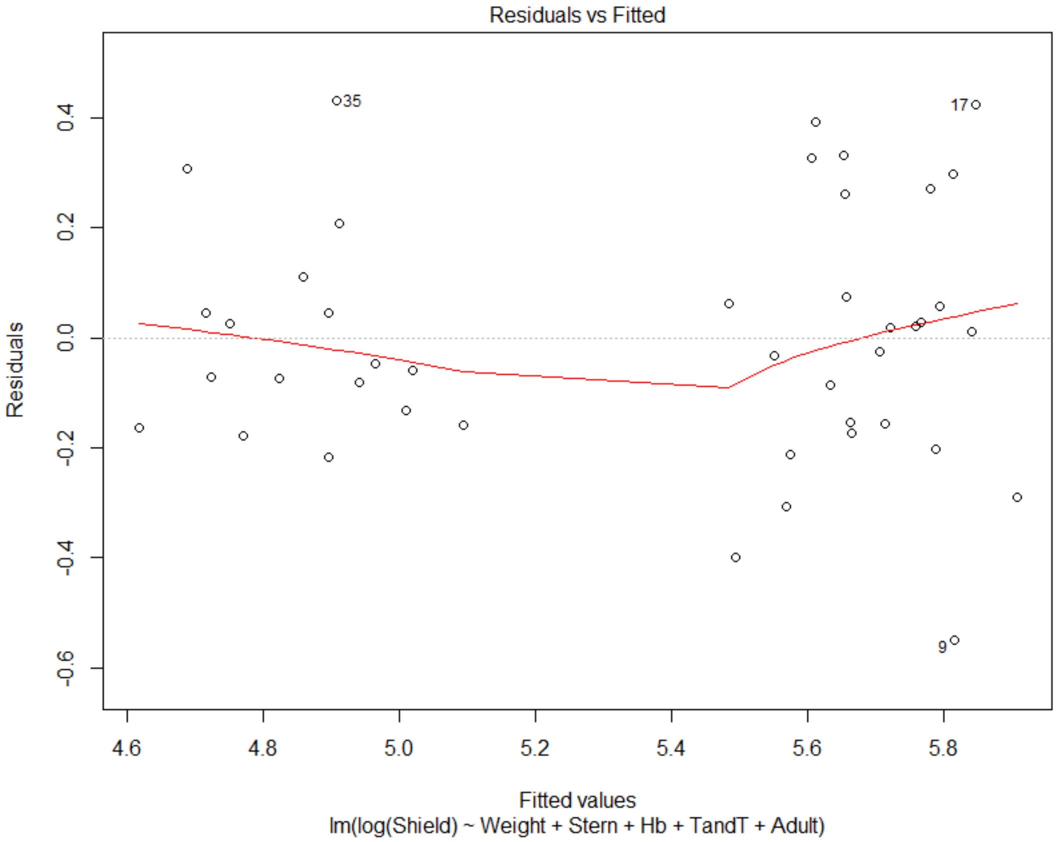


The additional median smoother (the curved red line), does suggest some slight curvature in the plot, which would be a violation of the assumption of independent errors. I am not particularly convinced by this suggestion, but there are definitely signs of increasing variance in the residuals as the fitted values increase. Any hint that observation 17 may be an outlier is probably just a result of the obvious non-constant variance.

A mid-range transformation, such as log, applied to the response variable, may help to remedy both the apparent problem with non-constant variance and possibly also the slight curvature.

Question 1 continued

(c) Now fit a multiple linear regression model with ln(`Shield`) as the response variable, still using all the other variables (not log transformed) as explanatory variables. Again present the main residual plot of the residuals against the fitted values for this new model. Does the transformation applied to the response variable appear to have corrected any problems you identified in part (b)? **(3 marks)**



The natural log transformation to the response variable does appear to have fixed most of the apparent problems There is still a hint of curvature (though this is now almost certainly an artefact of the median smoother), still possibly some increasing variance and/or a problem with an outlier (but now it is observation 9 which is the most likely culprit).

Note that there are two obvious clusters on the plot. All the juvenile birds have smaller fitted ln(`Shield`) values and are located on the left of the plot, whilst all the adult birds are located to the right of the plot. Any apparent increasing variance may simply be a result of the difference in the size of these two groups: 26 adult birds were observed, but only 17 juveniles.

Overall, this is much closer to being a plot where we can say the underlying assumptions are satisfied.

Question 1 continued

(d) Examine (but do not present) the ANOVA (Analysis of Variance) table and summary output for the model in part (c). Now adjust the order of the explanatory variables in the model in part (c), so that you can test the following "nested" hypotheses:

$$\mathrm{H}_0 : \beta_{\mathsf{Weight}} = \beta_{\mathsf{Hb}} = \beta_{\mathsf{TandT}} = 0$$

$$\mathrm{H}_0 : \beta_{\mathsf{Hb}} = \beta_{\mathsf{TandT}} = 0$$

$$\mathrm{H}_0 : \beta_{\mathsf{TandT}} = 0$$

Present the ANOVA table for the re-ordered model and discuss the result of the partial (nested) F-tests for the above hypotheses. Do your results suggest some possible modification(s) you could make to the model? **(3 marks)**

```
> model_1d <- lm(log(Shield) ~ Stern + Adult + Weight + Hb + TandT)
> anova(model_1d)
Analysis of Variance Table

Response: log(Shield)
          Df Sum Sq Mean Sq  F value    Pr(>F)    
Stern      1 1.4262  1.4262  24.8711 1.468e-05 ***
Adult      1 6.3003  6.3003 109.8661 1.253e-12 ***
Weight     1 0.0402  0.0402   0.7016    0.4076    
Hb         1 0.0001  0.0001   0.0014    0.9702    
TandT      1 0.0129  0.0129   0.2254    0.6377    
Residuals 37 2.1218  0.0573                       
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

$$\text{Model}: \mathsf{log(Shield)} = \beta_0 + \beta_{\mathrm{Stern}}\,\mathsf{Stern} + \beta_{\mathrm{Adult}}\,\mathsf{Adult} + \beta_{\mathrm{Weight}}\,\mathsf{Weight} + \beta_{\mathrm{Hb}}\,\mathsf{Hb} + \beta_{\mathrm{TandT}}\,\mathsf{TandT} + \varepsilon$$

$$\varepsilon \sim i.i.d.\ N\left(0, \sigma^2\right)$$

Taking the required tests in reverse, the last sequential F-statistic in the above ANOVA table, $F_{1,37} = 0.2254$ has an associated p-value which is greater than $\alpha = 0.05$, so we do not reject $\mathrm{H}_0 : \beta_{\mathsf{TandT}} = 0$ and conclude that TandT is not a significant addition to a model that already includes the four previous explanatory variables.

Similarly, for the other two nested F-tests:

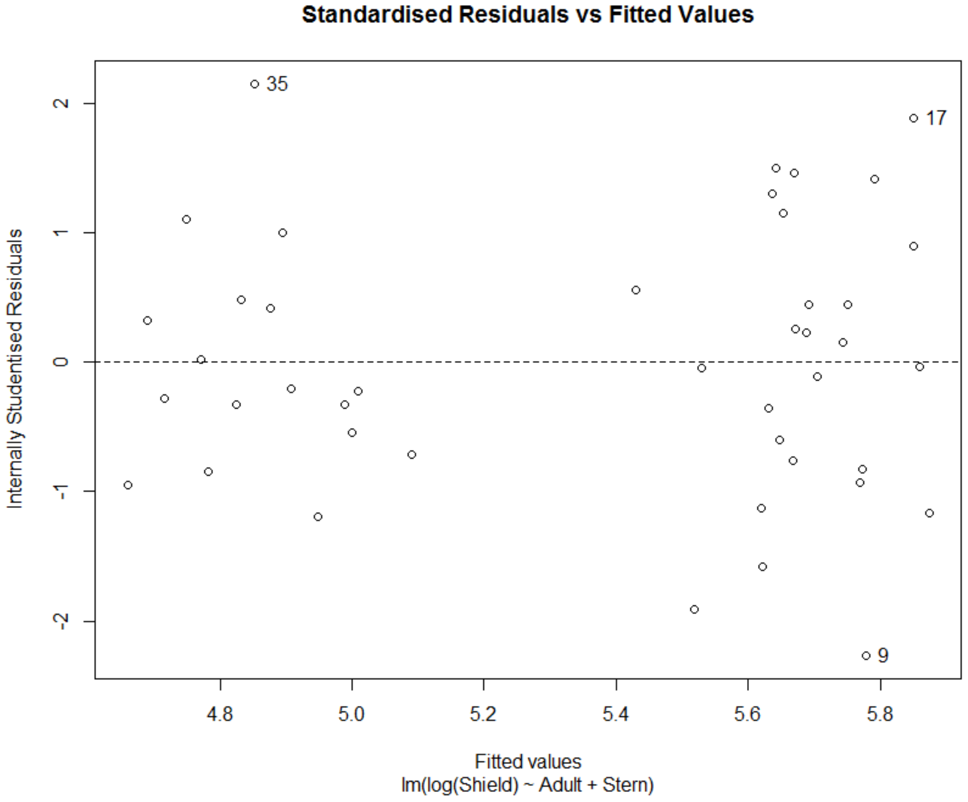
$F_{2,37} = \dfrac{(0.0001 + 0.0129)/(1+1)}{0.0573} = 0.1134$ is not greater than a critical value of $F_{2,37}(0.05) = 3.252$, so do not reject $\mathrm{H}_0 : \beta_{\mathsf{Hb}} = \beta_{\mathsf{TandT}} = 0$; and

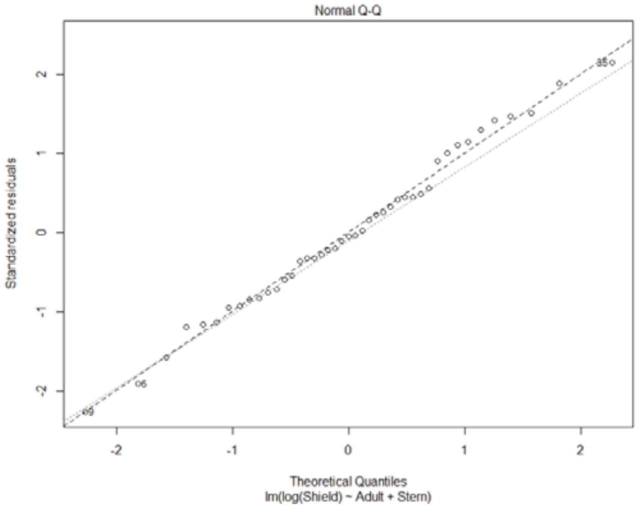
$F_{3,37} = \dfrac{(0.0402 + 0.0001 + 0.0129)/(1+1+1)}{0.0573} \approx 0.3095$ is not greater than a critical value of $F_{3,37}(0.05) = 2.859$, so do not reject $\mathrm{H}_0 : \beta_{\mathsf{Weight}} = \beta_{\mathsf{Hb}} = \beta_{\mathsf{TandT}} = 0.$

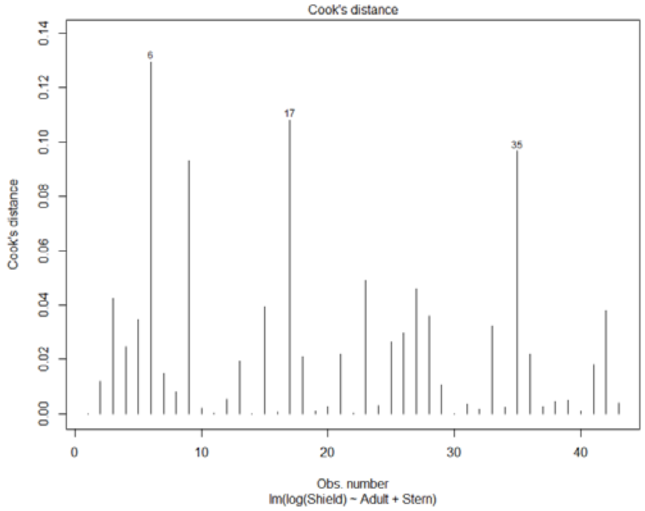
So none of Weight, Hb and TandT are significant additions to a model that already includes Stern and Adult, suggesting that we can refine the model by removing these optional observational covariates.

Question 1 continued

(e) Now fit the multiple linear regression model with ln(Shield) as the response variable and with Adult and Stern as explanatory variables. For this model, construct a plot of the internally Studentised residuals against the fitted values, a normal Q-Q plot of the residuals, and a bar plot of Cook's distances for each observation. Are there any obvious problems with the underlying assumptions? **(3 marks)**



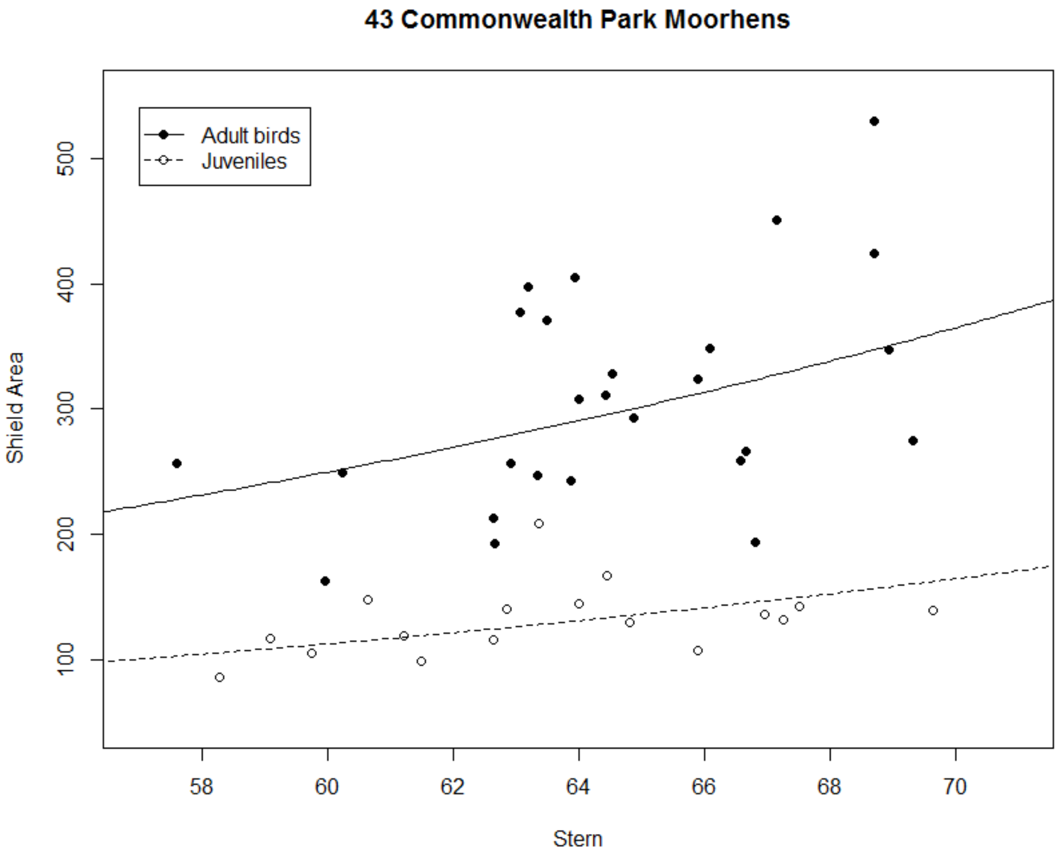




No obvious departures from the assumptions of either independence or constant variance in the main residual plot with only 2 of the 43 standardised residuals (4.7%) lying just outside ± 2 standard deviations. Similarly, no problems with the assumption of normality on the normal qq plot and none of the observations has a relatively large value of Cook's distance on the Cook's distance plot.

Question 1 continued

(f) Plot Shield against Stern, using different plotting symbols for juvenile and adult birds. Use the model from part (e) to predict the expected Shield area for both juvenile and adult birds over the full range of possible Stern measurements and include these on your plot as two different curves (using different line types). Include appropriate titles, axis labels, a legend and a brief discussion of your plot. **(3 marks)**



The plot suggests that Shield area increases as Stern increases (i.e. birds with bigger Stern measurements do indeed have larger shields) and also that adult birds typically have larger shields than juvenile birds (with the same Stern measurement).

(g) Consider two birds, one an adult bird and the other a juvenile, but who have the same Stern measurement. Present the table of coefficients for the model in part (e). Use this table to estimate the ratio of the expected Shield area of the adult bird to the expected Shield area of the juvenile bird with the same Stern measurement. Find a 95% confidence interval for this estimate. **(2 marks)**

```
> summary(model_1e)

Call:
lm(formula = log(Shield) ~ Adult + Stern)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.51460 -0.16352 -0.01033  0.11358  0.48673 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.45030    0.76743   3.193  0.00274 ** 
Adult        0.79532    0.07389  10.764 2.21e-13 ***
Stern        0.03791    0.01205   3.147  0.00311 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2332 on 40 degrees of freedom
Multiple R-squared:  0.7803,	Adjusted R-squared:  0.7694 
F-statistic: 71.05 on 2 and 40 DF,  p-value: 6.842e-14
```

As the ln (natural log) transformation is monotonically increasing, the positive and significant coefficient for Stern ($\hat{\beta}_{\text{Stern}} = 0.03791$, $t_{40} = 3.147$, $p = 0.00311$), confirms that both ln(Shield) and Shield area increase as Stern increases.

Similarly, the positive coefficient for Adult ($\hat{\beta}_{\text{Adult}} = 0.79532$, $t_{40} = 10.764$, $p = 0.00000$), suggests that adult birds (Adult = 1) have significantly larger shields than juvenile birds (Adult = 0). We can interpret the size of this coefficient as the expected increase in ln(Shield) as Adult increases by 1 from 0 (juvenile birds) to 1 (adult birds), with Stern remaining constant:

$$\hat{\beta}_{\text{Adult}} = \ln\left(\text{Shield}\right)_{\text{Adult=1}} - \ln\left(\text{Shield}\right)_{\text{Adult=0}} = \ln\left(\frac{\text{Shield}_{\text{Adult=1}}}{\text{Shield}_{\text{Adult=0}}}\right)$$

$$\frac{\text{Shield}_{\text{Adult=1}}}{\text{Shield}_{\text{Adult=0}}} = e^{\hat{\beta}_{\text{Adult}}} = \exp(0.79532) \approx 2.215$$

So the expected Shield area of an adult bird will be slightly more than twice the expected Shield area of a juvenile bird with the same Stern measurement.

A 95% confidence interval for this estimated ratio:

$$\hat{\beta}_{\text{Adult}} \pm t_{40}\left(0.975\right) \bullet se\left(\hat{\beta}_{\text{Adult}}\right) = 0.79532 \pm \left(2.021075\right)\left(0.07389\right) \approx 0.79532 \pm 0.14934$$

$$\left[\exp\left(0.64598\right),\ \exp\left(0.94466\right)\right] \approx \left(1.907,\ 2.572\right)$$

Question 2 **(20 marks)**

The dataset fat contains estimates of the percentage of adipose tissue (body.fat) and other related measurements taken on a sample of 252 adult men. The measurements are described in the help file for this dataset, help(fat), which also contains a link to another file, fat.txt, which contains yet more information about the data. Read both of these documents carefully to gain a better understanding of the contents of this dataset. You may also like to search the internet for a description of some of the terms used (e.g. “adipose tissue”).

For this assignment, we are interested in using all available information in the data to build a multiple linear regression model which can be used to estimate the percentage of body.fat, which is not easy to measure directly, as it has to be estimated using an underwater weighing technique.

(a) The file, fat.txt, suggests that there is an error in the height measurement for case 42, which should be 69.5 inches, rather than 29.5 inches. Apply this correction to the data. In fitting a multiple regression model with body.fat as the response variable, why should you not include case, body.fat.siri or density as possible explanatory variables? Is there a potential problem with including all three of weight, height and BMI as explanatory variables (hint: try looking up the definition of BMI)? What about including ffweight as a predictor in a model that already includes weight? **(4 marks)**

```
> c_height <- height
> c_height[42] <- 69.5
> c_height - height
  [1]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
 [22]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 40
 [43]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
 [64]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
 [85]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
[106]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
[127]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
[148]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
[169]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
[190]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
[211]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
[232]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
>
> cor(cbind(body.fat, body.fat.siri, density))
                body.fat body.fat.siri    density
body.fat       1.0000000     0.9997443 -0.9880867
body.fat.siri  0.9997443     1.0000000 -0.9877824
density       -0.9880867    -0.9877824  1.0000000
>
> cor(BMI, weight/(c_height^2))
[1] 0.9973828
>
> cor(ffweight, weight*(100-body.fat)/100)
[1] 0.9980995
>
```

Comments on the excluded variables:

- case is just a nominal identifier with numbers used as labels, there is nothing to indicate that there is any meaningful order or any other information included – as it is numerically coded, including this variable in a model would be very misleading, as it would be interpreted by R as a continuous covariate;
- reading the description of the data, both body.fat.siri and density are simply versions of the response variable measured on slightly different scales, so do not provide any new information. Both variables are very highly correlated with body.fat and it is only due to the mistakes in the density values and rounding in the calculation of the three variables, that they are not perfectly correlated; and
- similarly, both BMI and ffweight are functions of other explanatory variables which are likely to be included in the model and would cause extreme multicollinearity.

Question 2 continued

(b) Using body.fat (or some transformed version of body.fat) as your response variable and using just age, weight, height, neck, chest, abdomen, hip, thigh, knee, ankle, bicep, forearm and wrist as possible predictors, try to find a multiple regression model that is a "good" fit to these data. Note that weight is considered key and must be included, but BMI and any other variables not listed above should NOT be included. You are welcome to apply suitable scale transformations (e.g. log) to any or all of the explanatory variables, but do not use any interaction terms or higher order terms in any of the variables (except possibly for weight and height, as a way of including some proxy for BMI in the model). Do NOT use any "automatic" variable selection method in choosing your final model and do NOT present output from such a method as a justification for your choice of final model. Also, at this stage, do NOT delete any potential outliers.
Choose a promising candidate model and present a plot of the internally Studentised residuals against the fitted values, a Q-Q plot of the residuals and a bar plot of Cook's distances for each observation. Discuss these plots and comment on the model assumptions and on any unusual data points. **(4 marks)**

We are told in the question (presumably an instruction from the client) that "weight is considered key and must be included", so there is a prior assumption that this variable will be important and must be included in our model (so we can make the required predictions and test this underlying assumption). Now, weight is an important component of BMI and in Assignment 1 we saw that the simple linear regression model of body.fat regressed on log-transformed BMI was a reasonable fit to the data:

$$\begin{aligned} bod\hat{y}.fat &= \hat{\beta}_0 + \hat{\beta}_1 \ln\left(BMI\right) \ = \hat{\beta}_0 + \hat{\beta}_1 \ln\left(\frac{weight}{height^2}\right) \\ &= \hat{\beta}_0 + \hat{\beta}_1 \ln\left(weight\right) - 2\hat{\beta}_1 \ln\left(height\right) \\ &= \hat{\beta}_0 + \hat{\beta}_1^* \ln\left(weight\right) - \hat{\beta}_2^* \ln\left(height\right) \end{aligned}$$

This looks like a good starting point on which to build a multiple regression model and plots of body.fat against weight and body.fat against height (not shown, but code included in the *R* appendix), both look slightly better on a log-log scale, so we might also try a log transformation for the response variable body.fat as well. Note that there is one 0 value for body.fat (case 182), so we have to add 1 (or some other small positive constant) to all the body.fat values before taking logs.

If we now want to include the other biometric measurements, which are linear body measurements similar to height, it would seem sensible to also log those variables. The one exception is age, which is a little different to the other variables, but there is little to choose between plots of body.fat against age on the original and log-log scales, so I also chose to log age, for the sake of consistency.

A nested *F*-test for the addition of log(neck), log(chest), log(knee), log(ankle), log(bicep) and log(forearm), to a model which already includes log(weight), log(c_height), log(age), log(abdomen), log(hip), log(thigh) and log(wrist) gives $F_{6,238} = 1.8$, $p = 0.11$, suggesting that none of these 6 variables are significant additions to the base model.

Question 2, part (b) continued

This leaves log(abdomen), log(hip) and log(thigh) as variables which vary in significance in the ANOVA table and summary table of partial regression coefficients, depending on which order we include them in the model, which suggests we are dealing with multicollinearity. Reviewing the correlations between the variables on either the original or log scales, we find that most of the explanatory variables (except age and possibly height) are strongly correlated with weight. This is supported by the VIFs, which are relatively high for weight, abdomen, hip and thigh (the VIFs vary depending on which subset of variables you include and the scale used – see the code in the *R* appendix).

The "popular" literature on obesity, suggests that abdomen or belly fat is supposedly key in male obesity, whilst thigh and hip play a similar role for women, but may also be important for men. (Note, if we were preparing a report for publication, this sort of assertion would need the support of a good reference to the published literature). So, it would be nice to include these variables, but not at the expense of weight. Removing the variables with the highest VIFs (apart from weight), one at time, eventually leads us to exclude all three of log(abdomen), log(hip) and log(thigh), until we get a model in which all the remaining variables are significant in both the ANOVA and summary tables and for which the VIFs for the remaining explanatory variables are all reasonable.

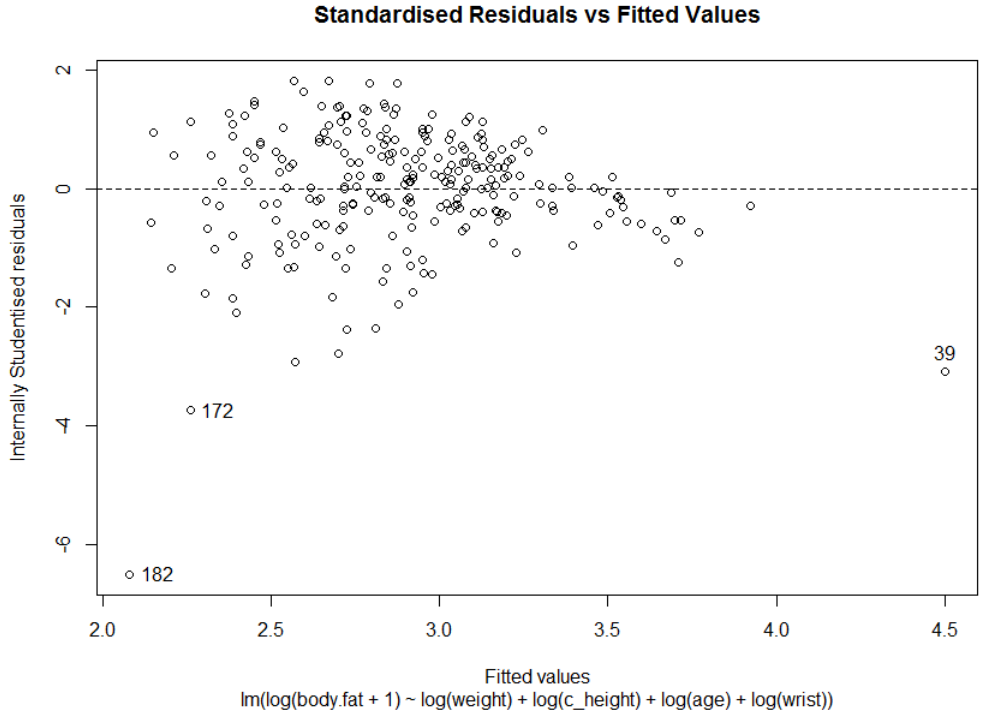
Note if we had started with a different prior assumption – say, that abdomen was key and weight was optional, we would have ended up with a different model in which abdomen was included and weight was excluded (again due to the collinearity between them). The final model (shown below) consists of just four explanatory variables: log(weight), log(c_height), log(age) and log(wrist):

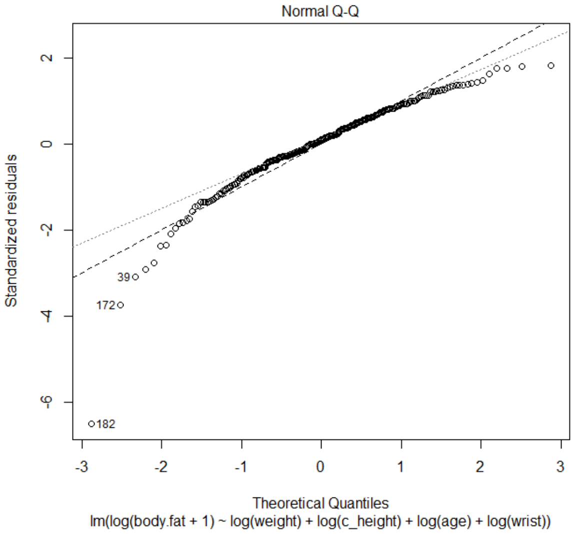
```
> fat.loglm <- lm(log(body.fat + 1) ~ log(weight) + log(c_height)
  + log(age) + log(wrist))
> fat.loglm

Call:
lm(formula = log(body.fat + 1) ~ log(weight) + log(c_height) +
    log(age) + log(wrist))

Coefficients:
  (Intercept)    log(weight)  log(c_height)       log(age)    log(wrist)
       9.6866         2.9980        -3.5888         0.4552       -3.0167
```

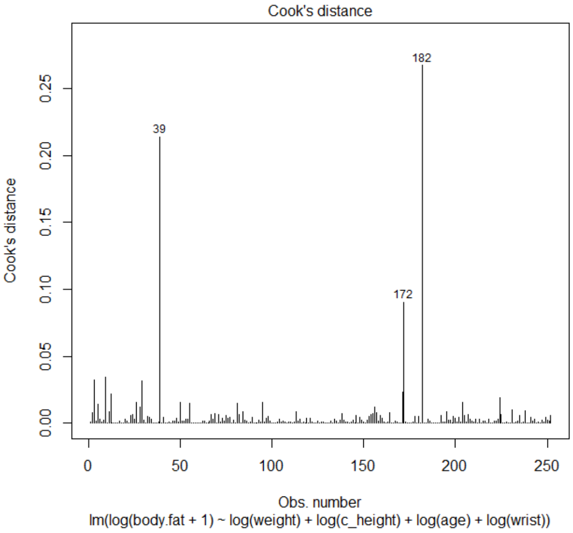
The partial regression coefficient for log(weight) is positive and the coefficient for log(c_height) is negative, which are both consistent with the assumptions made at the start. The inclusion of wrist is interesting and the interpretation of this model term should be checked with the clients. The partial regression coefficient for log(wrist) is negative, suggesting that like height, it belongs in the denominator of the calculation of BMI as an additional moderating factor (note: the Ponderal index, an alternative measure to BMI, uses weight over height cubed, on the grounds that weight is a volume measure and height is a linear measure).

Note that an extended explanation like the above is not a necessary part of a student answer to this part of the question, though a few comments (or even "dot points") on which variables you decided to include in the model and which you decided to exclude (and why), would be good. What is really required for this part of the question are the residual plots and discussion for some sensible candidate model (see the next page):

Question 2, part (b) continued





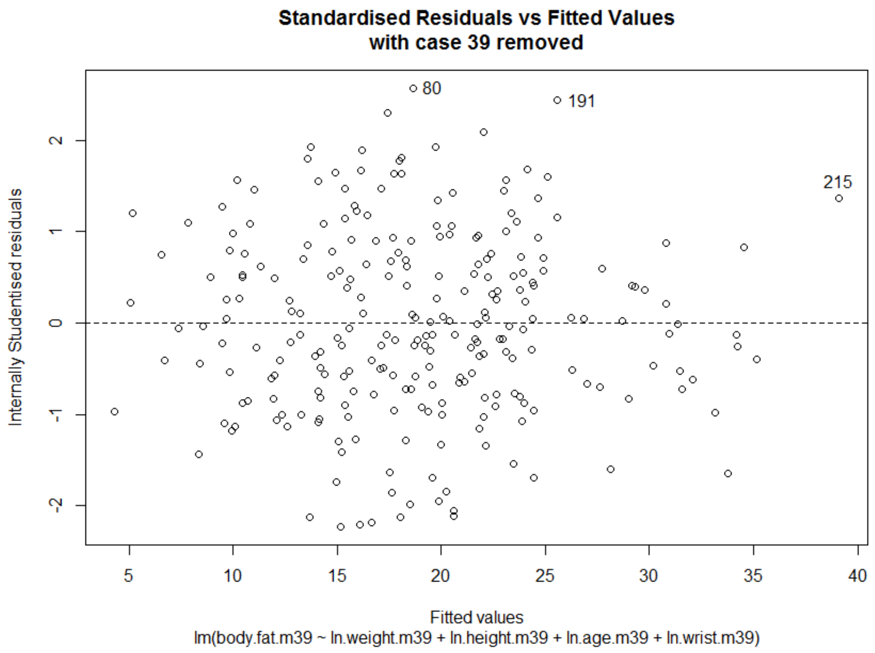


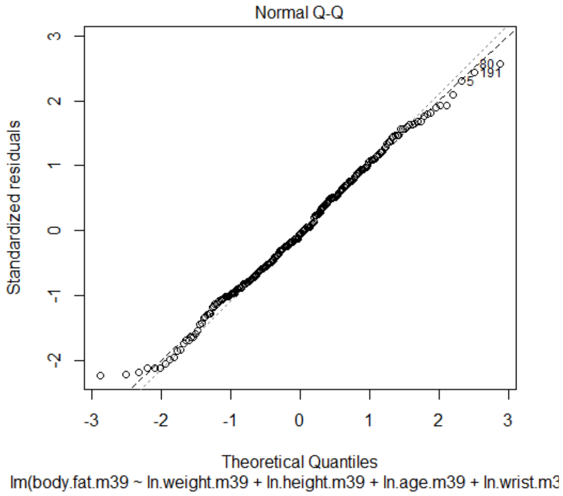
There are definitely problems with vertical outliers and highly influential points, with the most obvious ones (cases 39, 182 and 172) identified on the main residual plot ("Standardised Residuals vs Fitted Values"). The problems with these observations are also obvious on the bar plot of Cook's distances. The main residual plot also shows definite signs of decreasing variance, which is reinforced by the obvious departure from normality in the normal quantile plot (which is definitely of concern in this large a sample). These other "distributional" problems suggest that the log transformation may have been too strong.

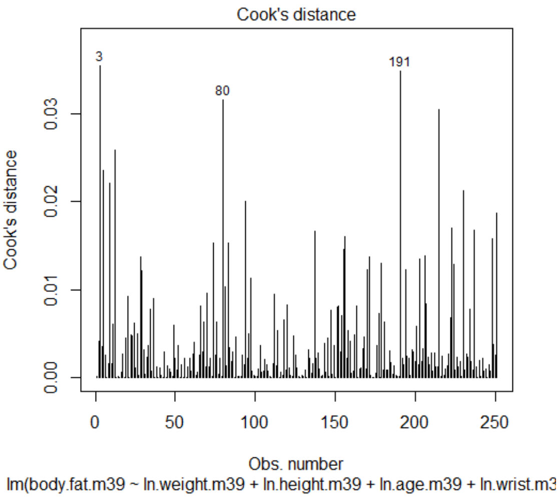
This is a good justification to go back and check the residual plots for the equivalent model with the response variable on the original untransformed scale, but keeping the explanatory variables on the log scale. The residual plots for this modified model (not shown, but relevant code is included in the R appendix) have none of the above "distributional" problems and less of a problem with outliers (observation 39 is still a potential problem), so this will be my chosen model in part (c).

Question 2 continued

(c) You may now delete no more than 3 potential outliers, but only if you can suggest a sensible justification for each exclusion. Refit the model to the reduced data set and again check residual plots. At this stage, you might decide to vary any transformations used and revisit the issue of which potential outliers to exclude. Choose just ONE final model and to justify your choice, present and discuss residual plots for your chosen model (with outliers removed). **(4 marks)**







On the residual plots that I didn't show for the modified model in part (b), case 39 was the observation with a relatively large value of Cook's distance plot and it was also the only observation that stood out on the main residual plot as both a possible outlier and highly influential point that also "passed" the test for being a mean shift outlier ($t_{39} = -3.03$ compared with $t_{246}(0.025) = -1.97$). Case 39 is the only observation with a BMI greater than 40 (BMI$_{40}$ = 48.9) and we cannot reasonably expect any fitted model to hold in a range where we have almost no data. Some of the other observations also "fail" the cut-offs on some of the other influence measures, but in my opinion, case 39 is the only observation that really warrants treatment as a highly influential outlier.

The residual plots for the modified look good with observation 39 removed. Maybe a slight deviation from normality in the lower tail of the normal quantile plot, but no obvious problems.

Question 2 continued

(d) Present the ANOVA table and the table of the estimated coefficients for your chosen model from part (c). Interpret the values of the estimated coefficients for this model and the results of the overall F test and the t-tests on the estimated coefficients. **(4 marks)**

```
> anova(fat.lm.m39)
Analysis of Variance Table

Response: body.fat.m39
               Df Sum Sq Mean Sq F value    Pr(>F)    
ln.weight.m39   1 5816.1  5816.1 281.055 < 2.2e-16 ***
ln.height.m39   1 2579.9  2579.9 124.668 < 2.2e-16 ***
ln.age.m39      1  558.0   558.0  26.966 4.341e-07 ***
ln.wrist.m39    1  812.5   812.5  39.264 1.644e-09 ***
Residuals     246 5090.7    20.7                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
> summary(fat.lm.m39)

Call:
lm(formula = body.fat.m39 ~ ln.weight.m39 + ln.height.m39 + ln.age.m39 + 
    ln.wrist.m39)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.0376  -3.2308  -0.2575   3.1743  11.5727 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    205.882     37.536   5.485 1.03e-07 ***
ln.weight.m39   55.525      2.979  18.639  < 2e-16 ***
ln.height.m39  -80.526      9.499  -8.478 2.14e-15 ***
ln.age.m39       7.403      1.073   6.897 4.47e-11 ***
ln.wrist.m39   -55.012      8.779  -6.266 1.64e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.549 on 246 degrees of freedom
Multiple R-squared:  0.6574,     Adjusted R-squared:  0.6518 
F-statistic:   118 on 4 and 246 DF,  p-value: < 2.2e-16
```

$$\text{Model}: \text{body.fat} = \beta_0 + \beta_1 \ln(\text{weight}) + \beta_2 \ln(\text{height}) + \beta_3 \ln(\text{age}) + \beta_4 \ln(\text{wrist}) + \varepsilon$$

$$\varepsilon \sim i.i.d.\ N\left(0, \sigma^2\right)$$

Overall F-test $H_0 : \dfrac{\sigma^2_{Model}}{\sigma^2_{Error}} = 1 \quad H_A : \dfrac{\sigma^2_{Model}}{\sigma^2_{Error}} > 1$ t-tests on β_j's $H_0 : \beta_j = 0 \quad H_A : \beta_j \neq 0$

$F_{4,246} = 118$, $p << 0.05$, so reject H_0 in favour of H_A and conclude the variance explained by the model is large compared to the error variance, i.e. the model is explaining a significant proportion of the variability in body.fat.

Similarly, the t-tests suggest that all of the partial regression coefficient are significant (for example, the test on the coefficient of weight: $t_{246} = 18.6$, $p << 0.05$, so reject H_0 in favour of H_A and conclude that the coefficient is significantly different from 0). This implies that each of the explanatory variables is having a significant effect on the response variable, body.fat, in the context of a multiple regression model that includes all four of the explanatory variables. All four explanatory variables are also significant additions to the model in the ANOVA table.

The size of the partial regression coefficients can be interpreted as the expected increase in body.fat as that predictor increases by 1, holding the other predictors constant. The intercept is the expected value of body.fat when all the predictors are zero, a situation which is well outside the range of the actual data.

Question 2 continued

(e) Assume you have four groups of new individuals categorised as “underweight”, “normal”, “overweight” and “obese”. Assuming that the four new groups have the same average for weight and the other measurements as the corresponding groups in the original sample, use your chosen model to repeat the predictions and confidence intervals made using the simple linear model in part (e) of Question 2 of Assignment 1. Is your chosen model a good model for making all of these predictions? **(4 marks)**

Here is a summary of the data:

			Variable means					
	BMI range	#cases	body.fat	BMI	weight	c_height	age	wrist
underweight	< 18.5	1	0.0	18.1	118.5	68.0	40.0	16.5
normal	[18.5, 25)	124	14.3	22.8	159.5	70.2	43.2	17.7
overweight	[25, 30)	103	22.2	27.0	191.5	70.6	46.2	18.6
obese:								
excl. 39	[30, 40)	23	29.5	32.1	221.8	69.8	48.4	19.0
incl. 39	≥ 30	24	29.6	32.8	227.7	69.9	48.3	19.1
case 39	≥ 40	1	33.8	48.9	363.2	72.3	46.0	21.4
Overall		252	18.9		178.9	70.3	44.9	18.2

Here are the predictions for body.fat from both the above multiple regression (MR) model and from the simple linear regression (SLR) model from Assignment 1

	MR model		SLR model from Assignment 1		
	body.fat	95% CI	BMI	body.fat	95% CI
underweight	4.3	(2.6, 6.0)	17.25	2.7	(0.7, 4.7)
normal	14.8	(14.1, 15.6)	21.75	12.6	(11.6, 13.6)
overweight	21.6	(20.9, 22.3)	27.50	22.7	(21.9, 23.5)
obese:			32.50	29.9	(28.4, 31.2)
excl. 39	30.8	(29.5, 32.8)			
incl. 39	31.9	(30.5, 33.2)			
case 39	48.3	(45.3, 51.4)			
Overall	19.8	(19.2, 20.5)			

I would not trust either of these models for the first of these predictions, as the models are at best based on only one observation (case 182) in the “underweight” end of the BMI range.

I think the multiple regression model predictions are reasonable for the two categories around the mean of the data (“normal” and “overweight”), where the residual plots for the model tended to look reasonable. The confidence intervals for the MR model predictions do at least include the observed data means of body.fat for both these categories.

If it is important to remove case 39 from the calculation of the model, it is also important to remove it from the calculation of the variable means used in the prediction for the “obese” category, however, I am not convinced that either model is doing a good job of making predictions in this “upper tail of the distribution”.