1. Question 1

a.

The scatterplot matrix and correlation matrix for the variables in the moorhen data are shown below.

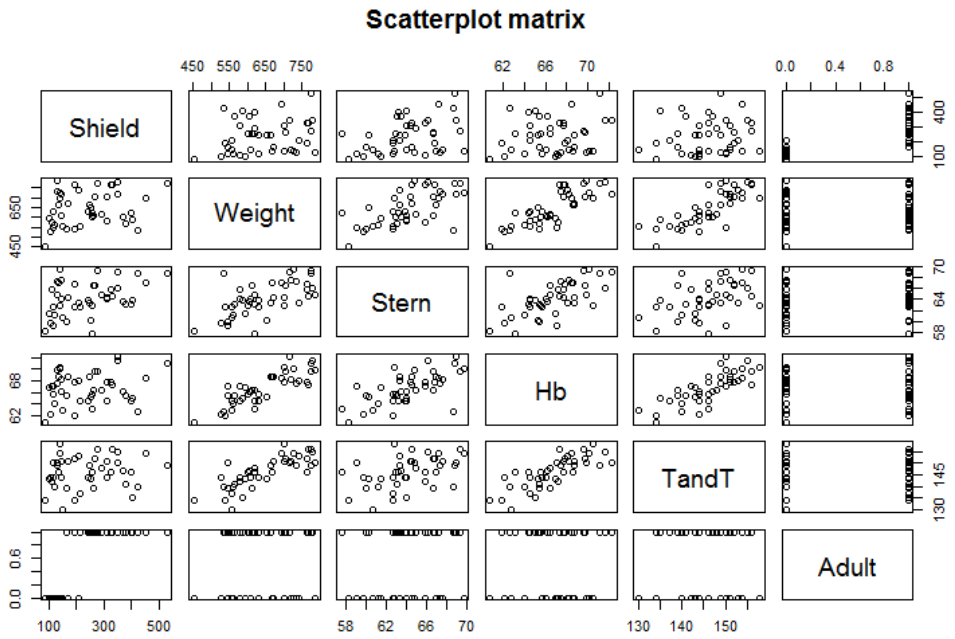

Figure 1: Scatterplot matrix for moorhen data

```
> 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
```

From scatterplot matrix, it is obvious that $Adult$ is a binary indicator which only takes on the values 0 and 1. From correlation matrix and correlation test, $Adult$ is significantly correlated with $Shield$, and NOT significantly correlated with $Weight, Stern, Hb$ and $TandT$. Therefore $Adult$ may be a proper explanatory variables for $Shield$.

Also from scatterplot matrix and correlation test, $Weight, Stern, Hb$ and $TandT$ show significant linear correlation with each other, also these four are all lineal measurements of each bird. For a multiple regression model, we should only include one of these four as explanatory variables to avoid multicollinearity. From correlation matrix and correlation test, we can find all $Weight, Stern, Hb$ and $TandT$ are NOT significantly correlated with $Shield$. Whether they can be explanatory variables for $Shield$ should be further tested.

b.

Fit a multiple linear regression model with Shield as response variable and with all the other variables in the data as explanatory variables:

```
> moorhen.lm <- lm(Shield ~ Weight + Stern + Hb + TandT + Adult)
> moorhen.lm

Call:
lm(formula = Shield ~ Weight + Stern + Hb + TandT + Adult)

Coefficients:
(Intercept)       Weight        Stern           Hb        TandT        Adult
 -583.78607     -0.06182      9.08199      0.69159      0.90056    168.66964
```

The main residual plot of the residuals against the fitted values is shown in Figure 2.

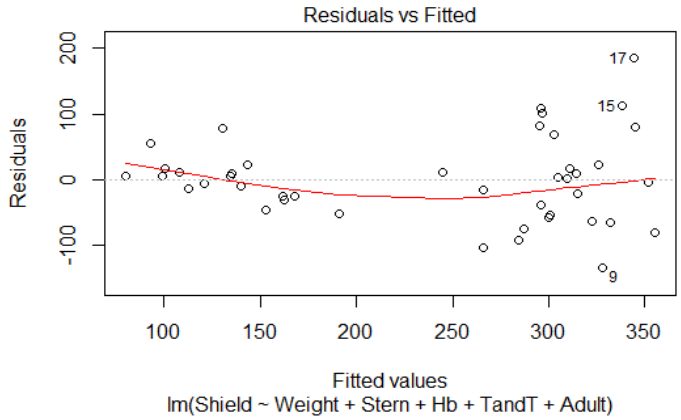

Figure 2: residuals vs fitted plot

The mean of residuals is close to zero. However, heteroscedasticity is an obvious problem of this model, that is, the residuals do not have constant variance. In detail, it can be found in the main residual plot (Figure 2) that the variance of residuals increases when fitted value increases.

c.

```
> moorhen.loglm <- lm(log(Shield) ~ Weight + Stern + Hb + TandT + Adult)
> moorhen.loglm

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

Coefficients:
(Intercept)       Weight        Stern           Hb        TandT        Adult
  2.3179549    0.0002523    0.0309669   -0.0043647    0.0048227    0.7979297
```

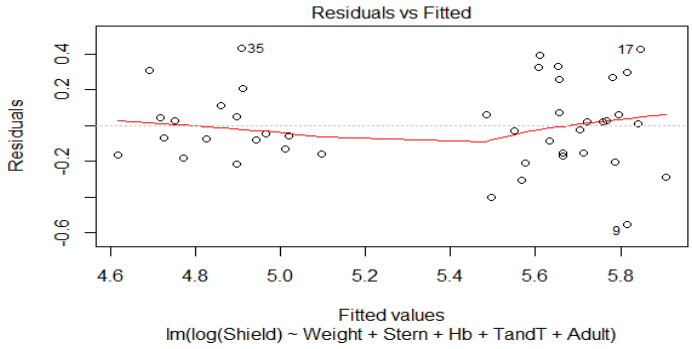

Figure 3: residuals vs fitted plot for log(Shield)

A second regression model is fitted with $\ln(Shield)$ as the response variable and all the other variables as explanatory variables. A main residual plot of that model is show in Figure 3.

Comparing with Figure 2, this plot shows the residuals have stable variance when fitted values changes. The log transformation of $Shield$ reduces heteroscedasticity of residuals.

d.

```
> moorhen.loglm_a <- lm(log(Shield) ~ Stern + Adult + cbind(Weight, Hb, Tand
T))
> anova(moorhen.loglm_a)
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 ***
cbind(Weight, Hb, TandT)  3 0.0532  0.0177   0.3095    0.8184
Residuals                37 2.1218  0.0573
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Model: $\log(Shield) = \beta_0 + \beta_1 Stern + \beta_2 Adult + \beta_i x_i + \varepsilon \quad \varepsilon \sim i.i.d. N(0, \sigma^2)$

And $x_i = [Weight, Hb, TandT], i = 3,4,5$

$H_0: \frac{\sigma^2_{x_i}}{\sigma^2_{Error}} = 1$ OR $H_0: \beta_{Weight} = \beta_{Hb} = \beta_{TandT} = 0$, equivalently in this case, $H_0: \beta_i = 0 \quad i = 3,4,5.$

$H_0: \frac{\sigma^2_{x_i}}{\sigma^2_{Error}} > 1$ OR H_A: at least one $\beta_i \neq 0$.

From the ANOVA table above, $F_{3,37} = 0.3095, p = 0.8184 > 0.05$, so do NOT reject H_0 in favour of H_A and conclude that the additional terms in the model $(Weight, Hb, TandT)$ do not significantly increase the proportion of the variance explained by the model and so are not significant additions to the model.

```
> moorhen.loglm_b <- lm(log(Shield) ~ Stern + Adult + Weight + cbind(Hb, Tand
T))
> anova(moorhen.loglm_b)
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
cbind(Hb, TandT)  2 0.0130  0.0065   0.1134    0.8931
Residuals        37 2.1218  0.0573
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Model: $\log(Shield) = \beta_0 + \beta_1 Stern + \beta_2 Adult + \beta_3 Weight + \beta_j x_j + \varepsilon \quad \varepsilon \sim i.i.d. N(0, \sigma^2)$

And $x_j = [Hb, TandT], j = 4{,}5$

$H_0: \frac{\sigma^2_{x_j}}{\sigma^2_{Error}} = 1$ OR $H_0: \beta_{Hb} = \beta_{TandT} = 0$, equivalently in this case, $H_0: \beta_j = 0 \quad j = 4{,}5$.

$H_0: \frac{\sigma^2_{x_j}}{\sigma^2_{Error}} > 1$ OR H_A: at least one $\beta_j \neq 0$.

From the ANOVA table above, $F_{2,37} = 0.1134, p = 0.8931 > 0.05$, so do NOT reject H_0 in favour of H_A and conclude that the additional terms in the model ($Hb, TandT$) do not significantly increase the proportion of the variance explained by the model and so are not significant additions to the model.

```
> moorhen.loglm_c <- lm(log(Shield) ~ Stern + Adult + Weight + Hb + TandT)
> anova(moorhen.loglm_c)
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
```

Model:

$$\log(Shield) = \beta_0 + \beta_1 Stern + \beta_2 Adult + \beta_3 Weight + \beta_4 Hb + \beta_5 TandT + \varepsilon \quad \varepsilon \sim i.i.d. N(0, \sigma^2)$$

$H_0: \frac{\sigma^2_{x_5}}{\sigma^2_{Error}} = 1$ OR $H_0: \beta_{TandT} = 0$, equivalently in this case, $H_0: \beta_5 = 0$.

$H_0: \frac{\sigma^2_{x_5}}{\sigma^2_{Error}} > 1$ OR $H_A: \beta_5 \neq 0$.

From the ANOVA table above, $F_{1,37} = 0.2254, p = 0.6377 > 0.05$, so do NOT reject H_0 in favour of H_A and conclude that the additional terms in the model ($TandT$) do not significantly increase the proportion of the variance explained by the model and so are not significant additions to the model.

After these three nested hypotheses tests, we can conclude that the variance of response variable $\log(Shield)$ is mainly explained by the variable $Stern$ and $Adult$. And $Weight, Hb$ and $TandT$ are not significant addition to this model. Therefore, a promising regression model for $\log(Shield)$ may only contains $Stern$ and $Adult$ as explanatory variables.

e.

A multiple linear regression model is fitted with $\ln(Shield)$ as the response variable and with $Adult$ and $Stern$ as explanatory variables. Three plots are generated.

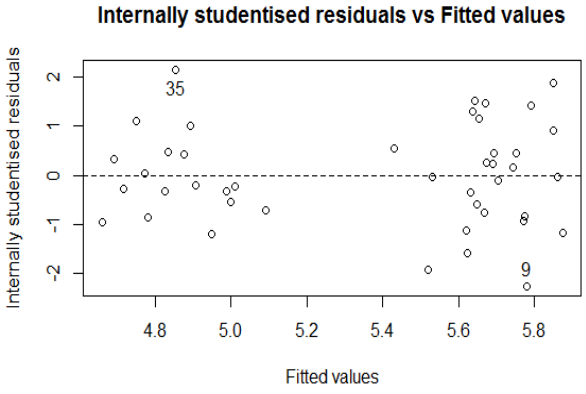

Figure 4: internally studentised residuals vs fitted values

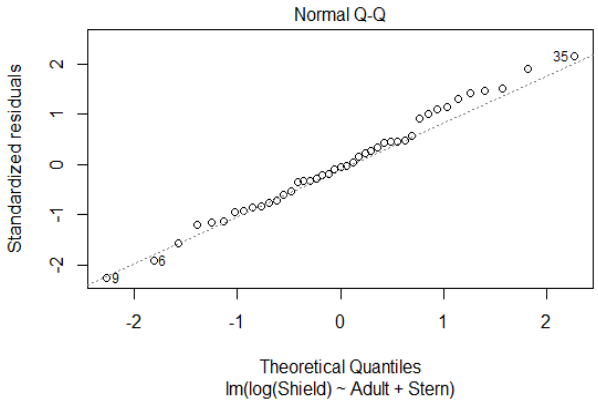

Figure 5: Normal Q-Q plot

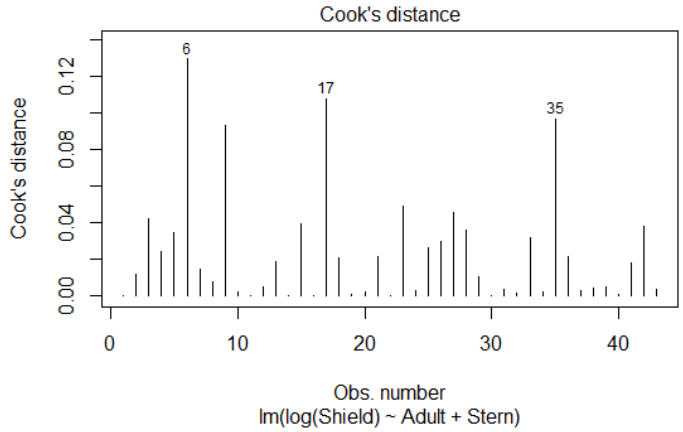

Figure 6: Bar plot of Cook's distance

Overall, the main residual plot (Figure 4) looks normal without obvious problems. The residuals do not show and patterns and the variance looks constant. The only possible outliers are point 9 and 35. However they are only slightly larger than 2 standard deviation away from 0, and that is totally acceptable. Therefore, the main residual plot does not show obvious problems. For normal Q-Q plot (Figure 5), the distribution of residuals is close enough to normal distribution considering the small sample size. Notably point 9 and 35 fit the normal distribution well. For Cook's distance (Figure 6), the maximum Cook's distance is about 0.13, which is very small. And point 35 and 9 are merely the third and fourth highest in Figure 6. Therefore, the Cook's distance plot does not show obvious outlier or highly influential points. In conclusion, these three plots show that residuals do not have obvious problems with the underlying assumptions, i.e. residuals do not show obvious pattern, the variance looks constant and no outliers are spotted.

f.

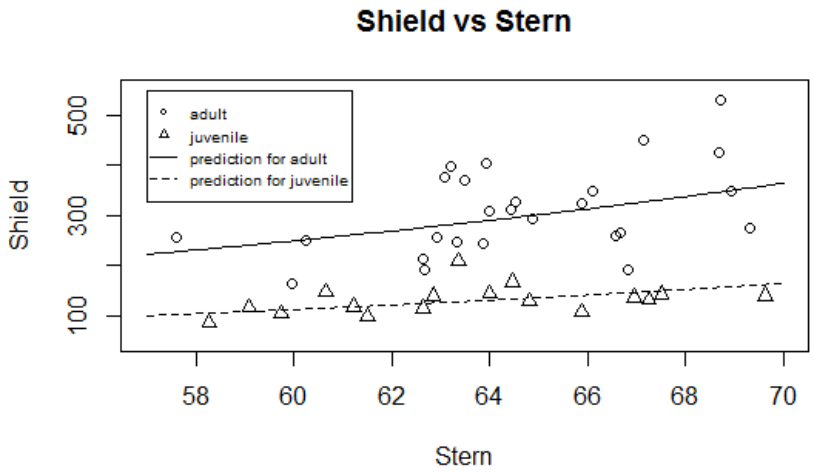

Figure 7: Shield vs Stern plot

From Figure 7, prediction line of juvenile moorhens fits the $Shield$ well, but for the prediction line of adult moorhens, when $Stern$ is larger, the variance of $Shield$ is also slightly larger. Besides, it can be found that given similar $Stern$, adult moorhens are expected to have larger Shield area than juvenile moorhens, which matches the common sense.

g.

```
> moorhen.e_lm <- lm(log(Shield) ~ Adult + Stern)
> summary(moorhen.e_lm)

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
```

Estimate the ratio of $Shield$ area of the adult bird to the juvenile bird with the same $Stern$ measurement:

For adult birds: $\log(Shield_a) = \beta_0 + \beta_1 Adult + \beta_2 Stern = \beta_0 + \beta_1 \times 1 + \beta_2 \times Stern_a$

For juvenile birds: $\log(Shield_j) = \beta_0 + \beta_1 Adult + \beta_2 Stern = \beta_0 + \beta_1 \times 0 + \beta_2 \times Stern_j$

Find that:

$$\log(Shield_a) - \log(Shield_j) = \log\left(\frac{Shield_a}{Shield_j}\right) = (\beta_0 - \beta_0) + \beta_1 \times (1 - 0) + \beta_2(Stern_a - Stern_j)$$
$$= \beta_1$$

The expected ratio:

$$\frac{Shield_a}{Shield_j} = e^{\beta_1} = e^{0.79531521} \approx 2.215139$$

Compute the 95% confidence interval using R:

```
> interval <- confint(moorhen.e_lm, "Adult")
> interval
          2.5 %    97.5 %
Adult 0.6459868 0.9446436
> exp(interval)
         2.5 %   97.5 %
Adult 1.907869 2.571897
```

Therefore the 95% confidence interval for estimated $\frac{Shield_a}{Shield_j}$ is $(1.907869,\ 2.571897)$.

2. Question 2

a.

Correcting $height$ measurement for case 42 is done in R:

```
> height[42]
[1] 29.5
> height[42] <- 69.5
> height[42]
[1] 69.5
```

Why should you not include $case$, $body.fat.siri$ or $density$ as possible explanatory variables?

From document, $case$ is just the index of sample (1, 2, 3, …), which doesn't have relationship with $body.fat$, so it should not be included. $body.fat.siri$ is the body fat percentage from another calculating method which has very close number as $body.fat$, therefore including $body.fat.siri$ is not meaningful. $density$ is the density of human body. From the document, $body.fat = \frac{457}{density} - 414.2$, which means the computation of $body.fat$ already contains $density$. So it's unnecessary to include $density$, otherwise multicollinearity may occur.

Is there a potential problem with including all three of $weight$, $height$ and BMI as explanatory variables?

The formula for BMI is $BMI = \frac{Weight(kg)}{(Height(m))^2}$, which relates with $height$ and $weight$. If we including $weight, height$ and BMI together, multicollinearity will appear.

What about including $ffweight$ as a predictor in a model that already includes $weight$?

From the document, $ffweight = (1 - fraction\ of\ body\ fat) \times weight$, which has linear relationship with $weight$. If we include $ffweight$ and $weight$ together, multicollinearity will appear.

b.

The promising candidate model is:

$$\log(body.fat + 1) = \beta_0 + \beta_1 \log(age) + \beta_2 \log(weight) + \beta_3 \log(height) + \beta_4 \log(wrist)$$

```
> fat.lm <- lm(log(body.fat+1) ~log(age)+log(weight)+log(height)+log(wrist))
> fat.lm
Call:
lm(formula = log(body.fat + 1) ~ log(age) + log(weight) + log(height) +
    log(wrist))
Coefficients:
(Intercept)     log(age)  log(weight)  log(height)   log(wrist)
     9.6866       0.4552       2.9980      -3.5888      -3.0167
```

Several aspects are considered when choosing this candidate model. Firstly, assignment 1 suggests that a promising simple linear regression model may be: $\log(body.fat) = \beta_0 + \beta_1 \log(BMI)$. By expanding BMI, we get $\log(body.fat) = \beta_0 + \beta_1 \log(BMI) = \beta_0 + \beta_1 \log\left(\frac{weight}{height^2}\right) = \beta_0 + \beta_1 \log(weight) - 2\beta_1 \log(height)$. Based on this expansion, I plan to include $height$ and $weight$ in this multiple regression model. Notably, $weight$ is a key factor and must be concluded as per requirement. Besides, due to data 182 has 0 $body.fat$ values, when do the log transformation I add 1 on all $body.fat$, like $\log(body.fat + 1)$.

Then from correlation test, I found age is significantly correlated with $body.fat$ and is not significantly correlated with $weight$. Hence it may be a proper explanatory variable and I plan to include it.

From correlation matrix and scatterplot matrix, $neck$, $chest$, $abdomen$, hip, $thigh$, $knee$, $ankle$, $bicep$, $forearm$ and $wrist$ are highly correlated. Also, considering they are all body measurements, they might have linear relationship with each other. To avoid multicollinearity, at most one of these explanatory variables can be included. Then I fit multiple models: $\log(body.fat + 1) = \beta_0 + \beta_1 \log(age) + \beta_2 \log(weight) + \beta_3 \log(height) + \beta_4 \log(X)$, where X is each body measurement. And then I check the variance inflation factor (VIF) and filter out 5 body measurements with least VIF, which are $neck$, $ankle$, $bicep$, $forearm$ and $wrist$. Then I check the ANOVA table of the same regression model for these five variables and find that, only $\log(wrist)$ and $\log(neck)$ are significant additions to the model. Finally, because $\log(wrist)$ has much lower p-value (3.771e-06) than $\log(neck)$ (0.0162975), I leave $\log(wrist)$ in the model. The same experiment can be repeated by the R code in Appendix.

For transformation, I apply log transformation for all variables to keep consistent, but the transformation will be adjusted in following steps. For $wrist$, applying log transformation is reasonable because $height$ and $wrist$ are both measurements of length. However, there might be problems to simply add log transformation on age because age is in different unit.

Three plots are generated for this model.

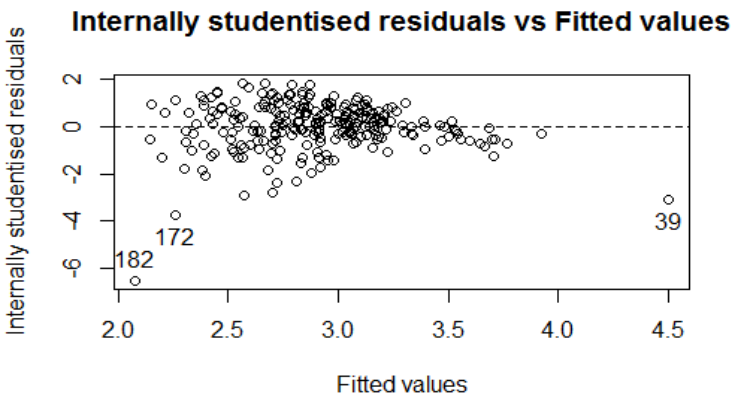

Figure 8: internally studentised residuals vs fitted values plot

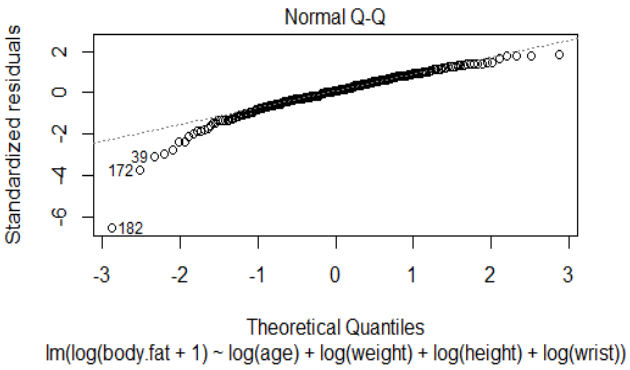

Figure 9: normal Q-Q plot

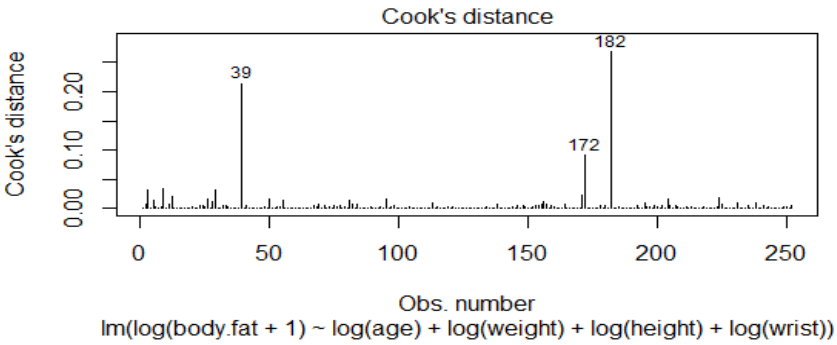

Figure 10: bar plot of Cook's distance

From main residual plot (Figure 8), there are several potential outliers identified on the plot. Point 39 is far from the most data points horizontally, so it may be a horizontal outlier. Point 182 is likely to be a vertical outlier because it has more than -6 standard deviation from 0. Point 172 has more than -3 standard deviation from 0, so it might be a vertical outlier. In normal Q-Q plot (Figure 9) and bar plot of Cook's distance (Figure 10), these three points (39, 172, 182) also appear as potential outliers.

Apart from that, the main residual plot (Figure 8) shows the variance of residuals reduce when fitted values increase, in other word, heteroscedasticity appears. Normal Q-Q plot (Figure 9) shows the distribution is slightly left skewed. These two observations indicate that the log transformation may be too strong.

c.

I would like to delete point 39, 172 and 182. As stated above, they are outliers in main residual plot (Figure 8), they do not fit normal distribution well (Figure 9) and they have high influence on the model (Figure 10). From raw data, we can find that point 39 is a case with extremely large BMI (48.9). Point 172 is a case with normal BMI (20.6) but having very low $body.fat$ (1.9). Point 182 is the case with zero $body.fat$. Therefore, these three cases have extreme measurements in the raw data, so excluding them is reasonable. However after excluding these three outliers, the residual plot still shows an obvious heteroscedasticity as stated in Section b.

Then I decide to adjust the model. I vary the log transformation on each item and check the residual plots, ANOVA table and coefficient summary table, finally I choose the best model:

$$body.fat = \beta_0 + \beta_1 age + \beta_2 \log(weight) + \beta_3 \log(height) + \beta_4 \log(wrist)$$

that is, I remove the log for $body.fat$ and age comparing with Section b. This is a reasonable justification because from Section b we know that the log transformation is too strong. Also this model ONLY shows one obvious outlier point (point 39).

After point 39 is removed, three new plots are generated for the final model.

```
> fat.lm_c1 <- lm(body.fat.r~age.r+log(weight.r)+log(height.r)+log(wrist.r))
> fat.lm_c1
Call:
lm(formula = body.fat.r ~ age.r + log(weight.r) + log(height.r) +
    log(wrist.r))
Coefficients:
  (Intercept)          age.r  log(weight.r)  log(height.r)   log(wrist.r)
     224.8876         0.1794         56.0871       -78.7867       -58.2890
```

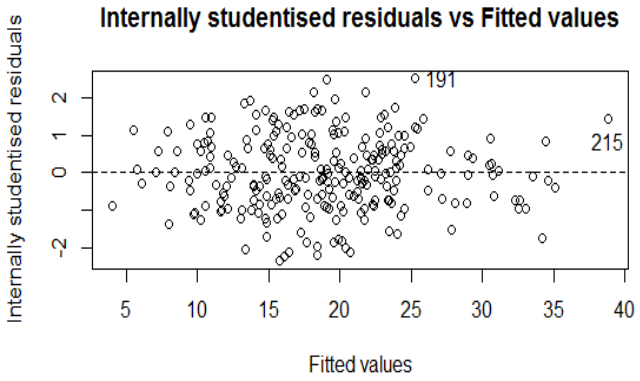

Figure 11: internally studentised residuals vs fitted values of final model after excluding point 39

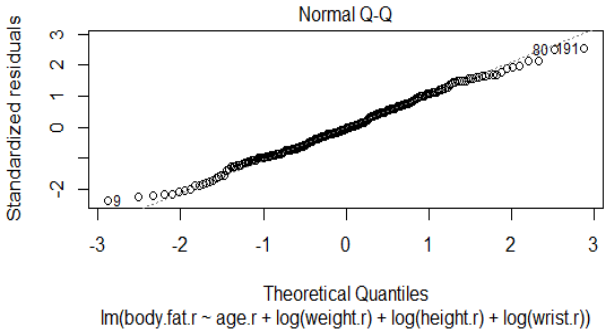

Figure 12: normal Q-Q plot of final model after excluding point 39

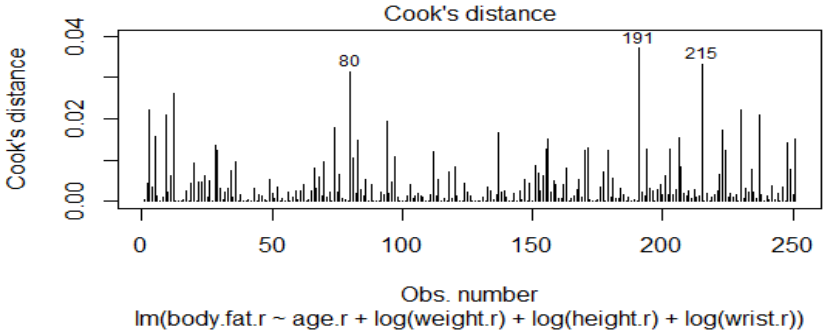

Figure 13: bar plot of Cook's distance of final model after excluding point 39

For the main residual plot (Figure 11), there is no obvious problems. The extreme residual points (like 191) are less than 3 standard deviation from 0, which is acceptable. From the normal Q-Q plot (Figure 12), the distribution of residuals fits closer to normal distribution comparing with Figure 9. From Cook's distance plot (Figure 13), there is no obvious problem identified. The maximum Cook's distance is point 191, but its value is not that large (less than 0.04) and not far from other points.

d.

```
> anova(fat.lm_c1)
Analysis of Variance Table

Response: body.fat.r
             Df Sum Sq Mean Sq F value    Pr(>F)
age.r         1 1255.1  1255.1  61.246 1.492e-13 ***
```

```
log(weight.r)   1 5887.6  5887.6 287.313 < 2.2e-16 ***
log(height.r)   1 1788.3  1788.3  87.266 < 2.2e-16 ***
log(wrist.r)    1  885.3   885.3  43.203 2.928e-10 ***
Residuals     246 5041.0    20.5
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(fat.lm_c1)

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

Residuals:
     Min       1Q   Median       3Q      Max
-10.6326  -3.2454  -0.1796   3.1381  11.2316

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   224.88759   36.71495   6.125 3.56e-09 ***
age.r           0.17936    0.02525   7.103 1.31e-11 ***
log(weight.r)  56.08715    2.97318  18.864  < 2e-16 ***
log(height.r) -78.78667    9.50571  -8.288 7.49e-15 ***
log(wrist.r)  -58.28904    8.86812  -6.573 2.93e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.527 on 246 degrees of freedom
Multiple R-squared:  0.6607,   Adjusted R-squared:  0.6552
F-statistic: 119.8 on 4 and 246 DF,  p-value: < 2.2e-16
```

Model:

$$body.fat = \beta_0 + \beta_1 age + \beta_2 \log(weight) + \beta_3 \log(height) + \beta_4 \log(wrist) + \varepsilon \quad \varepsilon \sim i.i.d.N(0, \sigma^2)$$

Interpret the values of the estimated coefficients:

Table 1: interpretation of all coefficients

Estimated values	Interpretation
$\beta_0 = 224.88759$	When all predictors are 0, $body.fat = 224.88759\%$, which is outside the range of given data.
$\beta_1 = 0.17936$	Holding the other predictors constant, an increase of 1 on age will cause an increase of 0.17936 percent in $body.fat$
$\beta_2 = 56.08715$	Holding the other predictors constant, an increase of 1 on log scale of $weight$ will cause an increase of 56.08715 percent in $body.fat$
$\beta_3 = -78.78667$	Holding the other predictors constant, an increase of 1 on log scale of $height$ will cause a decrease of 78.78667 percent in $body.fat$
$\beta_4 = -58.28904$	Holding the other predictors constant, an increase of 1 on log scale of $wrist$ will cause a decrease of 58.28904 percent in $body.fat$

Overall F test, $H_0: \frac{\sigma^2_{model}}{\sigma^2_{error}} = 1$. $H_A: \frac{\sigma^2_{model}}{\sigma^2_{error}} > 1$.

$F_{4,246} = 119.8$, $p \ll 0.05$, 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 involving $age, \log(weight), \log(height), \log(wrist)$ is explaining a significant proportion of the variability in $body.fat$.

Overall t-test, $H_0: t_j = 0, j = 0,1,2,3,4$. $H_A: any\ one\ of\ t_j \neq 0, j = 0,1,2,3,4$.

From summary table, $t_0 = 6.125; t_1 = 7.103; t_2 = 18.864; t_3 = -8.288; t_4 = -6.573$. All of them have p values much smaller than 0.05, so reject H_0 in favour of H_A and conclude that all the slope coefficients and the intercept are significantly different from 0.

e.

According to Assignment 1 question 2(e), I category these four groups by BMI and then compute the categorical average value of $age, weight, height$ and $wrist$.

Table 2: average values for four new categories

Category	BMI	Avg(age)	Avg(weight)	Avg(height)	Avg(wrist)
Underweight	(0,18.5]	40	118.5	68	16.5
Normal	(18.5,25]	43.256	159.5292	70.148	17.7496
Overweight	(25,30]	46.12745	191.8113	70.62745	18.64118
Obese	(30,+∞)	48.29167	227.6896	69.875	19.05417

Then I do the prediction based on these average values of these four categories, and find the 95% confidence interval as same as Assignment 1.

```
> predict(fat.lm_c1, newdata=data.frame(age.r=avg.age,
+                                       weight.r=avg.weight,
+                                       height.r=avg.height,
+                                       wrist.r=avg.wrist), interval="confide
nce")
        fit       lwr       upr
1  4.027138  2.406718  5.647557
2 14.581148 13.878436 15.283860
3 22.038788 21.381972 22.695603
4 31.610900 30.290830 32.930969
```

Then combine this prediction result with previous prediction in Assignment 1, get table ().

Table 3: prediction of body.fat and 95% confidence interval.

	Multiple Linear Regression			Simple Linear Regression (Assignment 1)		
Category	fit	lwr	upr	fit	lwr	upr
Underweight	4.027138	2.406718	5.647557	2.709611	0.793065	4.626157
Normal	14.58115	13.87844	15.28386	12.6353	11.68451	13.58609
Overweight	22.03879	21.38197	22.6956	22.67962	21.91453	23.44471
Obese	31.6109	30.29083	32.93097	29.83284	28.46111	31.20456

Firstly from table 3, the 95% confidence interval for prediction is narrow, and the fitted values are slightly shifted comparing with Simple Linear Regression from Assignment 1. It is reasonable as we have added more explanatory variables into the model. Then by checking four categories, I find the 'Underweight' category only contains one case (point 182). The group size is too small, so the prediction for 'Underweight' category is definitely not reliable. For 'Normal', 'Overweight' and 'Obese', the group size is large enough (125, 102, 24 cases respectively), and I think the model works well to make these predictions considering the good residual plots and the significant F-test and t-test in section d. Overall, the multiple linear regression model is NOT good at predicting 'Underweight' group, but it is a good model for predicting 'Normal', 'Overweight' and 'Obese' groups.

Appendix

```
# R code for assignment2 - Dingying Li
# Q1(a)

moorhen <- read.csv('moorhen.csv', header=T)
attach(moorhen)

pairs(moorhen, main='Scatterplot matrix')
cor(moorhen)

# Q1(b)
moorhen.lm <- lm(Shield ~ Weight + Stern + Hb + TandT + Adult)
moorhen.lm
plot(moorhen.lm, which=1) # residual vs fitted plot

# Q1(c)
moorhen.loglm <- lm(log(Shield) ~ Weight + Stern + Hb + TandT + Adult)
moorhen.loglm
plot(moorhen.loglm, which=1) # residual vs fitted plot


# Q1(d)
anova(moorhen.loglm)
summary(moorhen.loglm)

moorhen.loglm_a <- lm(log(Shield) ~ Stern + Adult + cbind(Weight, Hb, TandT))
anova(moorhen.loglm_a)
moorhen.loglm_b <- lm(log(Shield) ~ Stern + Adult + Weight + cbind(Hb, TandT))
anova(moorhen.loglm_b)
moorhen.loglm_c <- lm(log(Shield) ~ Stern + Adult + Weight + Hb + TandT)
anova(moorhen.loglm_c)


# Q1(e)
moorhen.e_lm <- lm(log(Shield) ~ Adult + Stern)
summary(moorhen.e_lm)
plot(fitted(moorhen.e_lm), rstandard(moorhen.e_lm), xlab="Fitted values", ylab="Internally studentised
residuals", main="Internally studentised residuals vs Fitted values")
identify(fitted(moorhen.e_lm), rstandard(moorhen.e_lm))
abline(0,0, lty=2)

plot(moorhen.e_lm, which=2) # normal Q-Q
plot(moorhen.e_lm, which=4) # cook's distance

# Q1(f)
plot(Stern[Adult==1], Shield[Adult==1], pch=1, xlim=c(57,70), ylim=c(50,550), xlab="Stern",
    ylab="Shield", main="Shield vs Stern")
points(Stern[Adult==0], Shield[Adult==0], pch=2)
pred_adult <- predict(moorhen.e_lm, data.frame(Adult=1, Stern=c(570:700)/10))
pred_juvenile <- predict(moorhen.e_lm, data.frame(Adult=0, Stern=c(570:700)/10))
```

```
lines(x=c(570:700)/10, y=exp(pred_adult), lty=1)
lines(x=c(570:700)/10, y=exp(pred_juvenile), lty=2)
legend(57,550,c('adult','juvenile', 'prediction for adult', 'prediction for
juvenile'),pch=c(1,2,NA,NA),lty=c(NA,NA,1,2),cex=0.6)

# Q1(g)
summary(moorhen.e_lm)
interval <- confint(moorhen.e_lm, "Adult")
interval
exp(interval)

# Q2(a)
fat <- read.csv('fat.csv', header=T)
attach(fat)
# correct case 42: replace height to 69.5, rather than 29.5
height[42]
height[42] <- 69.5
height[c(40:50)]

# Q2(b)
vif <- function(xmatrix) {
  if (class(xmatrix) == "matrix" | class(xmatrix) == "data.frame")
    diag(solve(cor(xmatrix)))
  else
    diag(solve(cor(model.matrix(xmatrix)[,-1])))
  # assuming a linear model object, if not a matrix
}

pairs(~body.fat+age+weight+height+neck+chest+abdomen+hip+thigh+knee+ankle+bicep+forearm+wrist)
# use external library corrplot to get better visualization
C <- cor(data.frame(body.fat,age,weight,height,neck,chest,abdomen,hip,thigh,knee,ankle,bicep,forearm,wrist))
C
library(corrplot)
corrplot(C, method='circle')

pairs(body.fat+age+weight+height+neck+chest+abdomen+hip+thigh+knee+ankle+bicep+forearm+wrist)
pairs(~log(body.fat+1)+log(age)+log(weight)+log(height)+log(neck)+log(chest)+log(abdomen)+log(hip)+log(thigh)+l
og(knee)+log(ankle)+log(bicep)+log(forearm)+log(wrist))

# Choose Model, change 'wrist' to any body measurement and run following 3 lines
test.lm <- lm(log(body.fat+1) ~log(age)+log(weight)+log(height)+log(wrist))
vif(test.lm)
anova(test.lm)

fat.lm <- lm(log(body.fat+1) ~log(age)+log(weight)+log(height)+log(wrist))
anova(fat.lm)
summary(fat.lm)
vif(fat.lm)

plot(fitted(fat.lm), rstandard(fat.lm),
    xlab="Fitted values", ylab="Internally studentised residuals",
    main="Internally studentised residuals vs Fitted values")
identify(fitted(fat.lm), rstandard(fat.lm))
```

```
abline(0, 0, lty=2)
plot(fat.lm, which=2)
plot(fat.lm, which=4)

# Q2(c)
body.fat.r <- body.fat[c(-39,-172,-182)]
age.r <- age[c(-39,-172,-182)]
weight.r <- weight[c(-39,-172,-182)]
height.r <- height[c(-39,-172,-182)]
wrist.r <- wrist[c(-39,-172,-182)]

fat.lm_c0 <- lm(log(body.fat.r+1)
            ~log(age.r)+log(weight.r)+
              log(height.r)+log(wrist.r))
anova(fat.lm_c0)
summary(fat.lm_c0)
vif(fat.lm_c0)
plot(fitted(fat.lm_c0), rstandard(fat.lm_c0))

body.fat.r <- body.fat[c(-39)]
age.r <- age[c(-39)]
weight.r <- weight[c(-39)]
height.r <- height[c(-39)]
wrist.r <- wrist[c(-39)]

fat.lm_c1 <- lm(body.fat.r
             ~age.r+log(weight.r)+
              log(height.r)+log(wrist.r))
anova(fat.lm_c1)
vif(fat.lm_c1)
plot(fitted(fat.lm_c1), rstandard(fat.lm_c1),
   xlab="Fitted values", ylab="Internally studentised residuals",
   main="Internally studentised residuals vs Fitted values")
identify(fitted(fat.lm_c1), rstandard(fat.lm_c1))
abline(0, 0, lty=2)
plot(fat.lm_c1, which=2)
plot(fat.lm_c1, which=4)

fat.lm_c2 <- lm(body.fat.r
             ~age.r+weight.r+
              height.r+wrist.r)
anova(fat.lm_c2)
vif(fat.lm_c2)
plot(fitted(fat.lm_c2), rstandard(fat.lm_c2))

# Q2(d)
anova(fat.lm_c1)
summary(fat.lm_c1)

# Q2(e)
length(weight[BMI<=18.5])
length(weight[BMI>18.5 & BMI<=25])
length(weight[BMI>25 & BMI<=30])
```

```
length(weight[BMI>30])

avg.weight.un <- mean(weight[BMI<=18.5]) # underweight
avg.weight.no <- mean(weight[BMI>18.5 & BMI<=25]) # normal
avg.weight.ov <- mean(weight[BMI>25 & BMI<=30]) # overweight
avg.weight.ob <- mean(weight[BMI>30]) # obese

avg.age.un <- mean(age[BMI<=18.5]) # underweight
avg.age.no <- mean(age[BMI>18.5 & BMI<=25]) # normal
avg.age.ov <- mean(age[BMI>25 & BMI<=30]) # overweight
avg.age.ob <- mean(age[BMI>30]) # obese

avg.height.un <- mean(height[BMI<=18.5]) # underweight
avg.height.no <- mean(height[BMI>18.5 & BMI<=25]) # normal
avg.height.ov <- mean(height[BMI>25 & BMI<=30]) # overweight
avg.height.ob <- mean(height[BMI>30]) # obese

avg.wrist.un <- mean(wrist[BMI<=18.5]) # underweight
avg.wrist.no <- mean(wrist[BMI>18.5 & BMI<=25]) # normal
avg.wrist.ov <- mean(wrist[BMI>25 & BMI<=30]) # overweight
avg.wrist.ob <- mean(wrist[BMI>30]) # obese

avg.weight <- c(avg.weight.un,
            avg.weight.no,
            avg.weight.ov,
            avg.weight.ob)

avg.age <- c(avg.age.un,
          avg.age.no,
          avg.age.ov,
          avg.age.ob)

avg.height <- c(avg.height.un,
            avg.height.no,
            avg.height.ov,
            avg.height.ob)

avg.wrist <- c(avg.wrist.un,
           avg.wrist.no,
           avg.wrist.ov,
           avg.wrist.ob)

predict(fat.lm_c1, newdata=data.frame(age.r=avg.age,
                           weight.r=avg.weight,
                           height.r=avg.height,
                           wrist.r=avg.wrist), interval="confidence")
```