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

REGRESSION MODELLING

(STAT2008/STAT4038/STAT6038)

Solutions to Assignment 1 for 2017

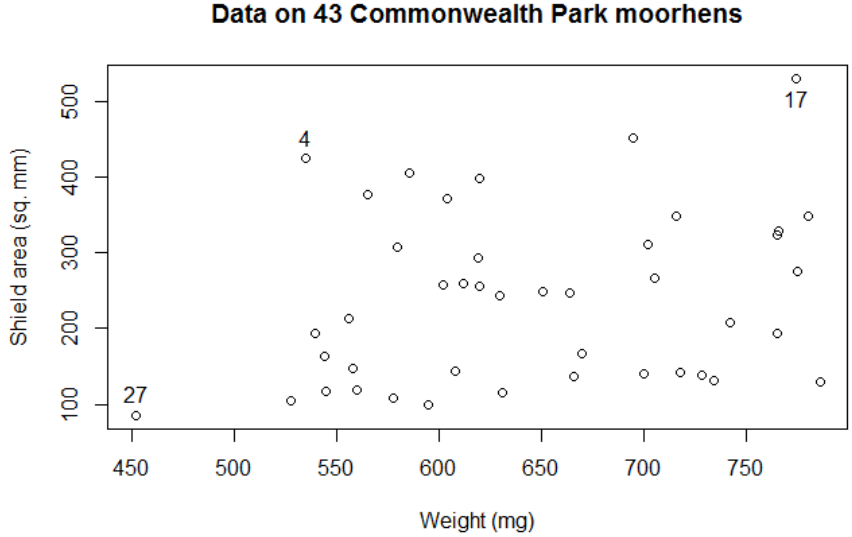
Question 1 **(20 marks)**

Moorhens are those blue-purple-red water birds often seen down near Lake Burley Griffin in Commonwealth Park. They are characterised by large, fleshy red shields that protrude from their heads. Some scientists have collected various measurements on a group of 43 moorhens in Commonwealth Park in the file moorhen.csv, which is available on Wattle. The scientists have sent the data to you for analysis.

The e-mail accompanying the data is a little light on the details, but there is a suggestion that moorhens form a fairly hierarchical society and that shield size is a relevant indicator of a bird's status within their group, so the variable of most interest (the response variable) is the area of each bird's Shield (units not specified, but presumably in mm^2). An alternative explanation might be that a bird's status is more strongly related to their overall size (which could be measured by the bird's Weight, presumably in mg) and that bigger birds simply have larger shields.

In this assignment, we will concentrate on the relationship between Weight and Shield (we will investigate the other available variables in Assignment 2). Read the data into R and conduct the following analyses:

(a) Plot Shield against Weight (which means that Shield should be the response variable on the Y or vertical axis and Weight should be the explanatory variable on the X or horizontal axis). Use the identify() function in R to identify any unusual data points on the plot. Discuss why you chose these observation(s) as being unusual. **(2 marks)**




```
> moorhen[c(4,17,27),]
   Shield Weight Stern    Hb TandT Adult
4   424.5    535 68.70 62.62   140     1
17  529.3    774 68.71 71.10   149     1
27   85.9    452 58.28 60.85   134     0
```

The plot shows a relationship that is generally positive, but not particularly strong. The three observations identified on the plot were chosen as standing out from the rest of the data: 17 is the bird with the largest shield; 4 is a relatively light-weight bird with a large shield; and 27 is a very light juvenile (non-adult) bird with the smallest shield.

Question 1 continued

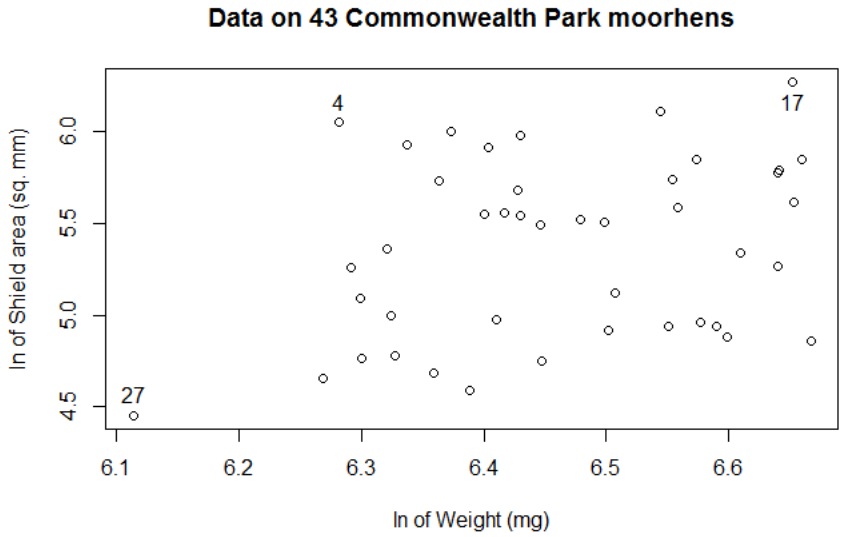
(b) Is there a significant correlation between Weight and Shield? Use the cor.test() function to conduct a suitable hypothesis test. Clearly specify the hypotheses you are testing and present and interpret the results. **(2 marks)**

```
> cor.test(Weight, Shield)

   Pearson's product-moment correlation

data:  Weight and Shield
t = 1.5793, df = 41, p-value = 0.122
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.06559203  0.50359325
sample estimates:
      cor
0.2394694
```

Test H_0: $\rho = 0$ vs H_A: $\rho \neq 0$, where ρ is the correlation between Weight and Shield (or the correlation between Shield and Weight as it doesn't really matter which way round we consider correlations, unlike variables in a regression model).

$t_{41} = 1.58$, $p = 0.122 > 0.05$, so do NOT reject H_0 in favour of H_A and conclude that ρ is NOT significantly different from 0. The observed sample correlation $r = 0.24$ suggests only a weak positive correlation between Weight and Shield.

(c) Experiment with applying natural log transformations (to the base e, which is the default for the log() function in R) and square root transformations to one or both of Weight and Shield, and repeat the analysis in parts (a) and (b). Do NOT show all of your results, just pick whichever one you think is the best choice of scale for the two variables and show and discuss the results for your chosen combination. **(4 marks)**

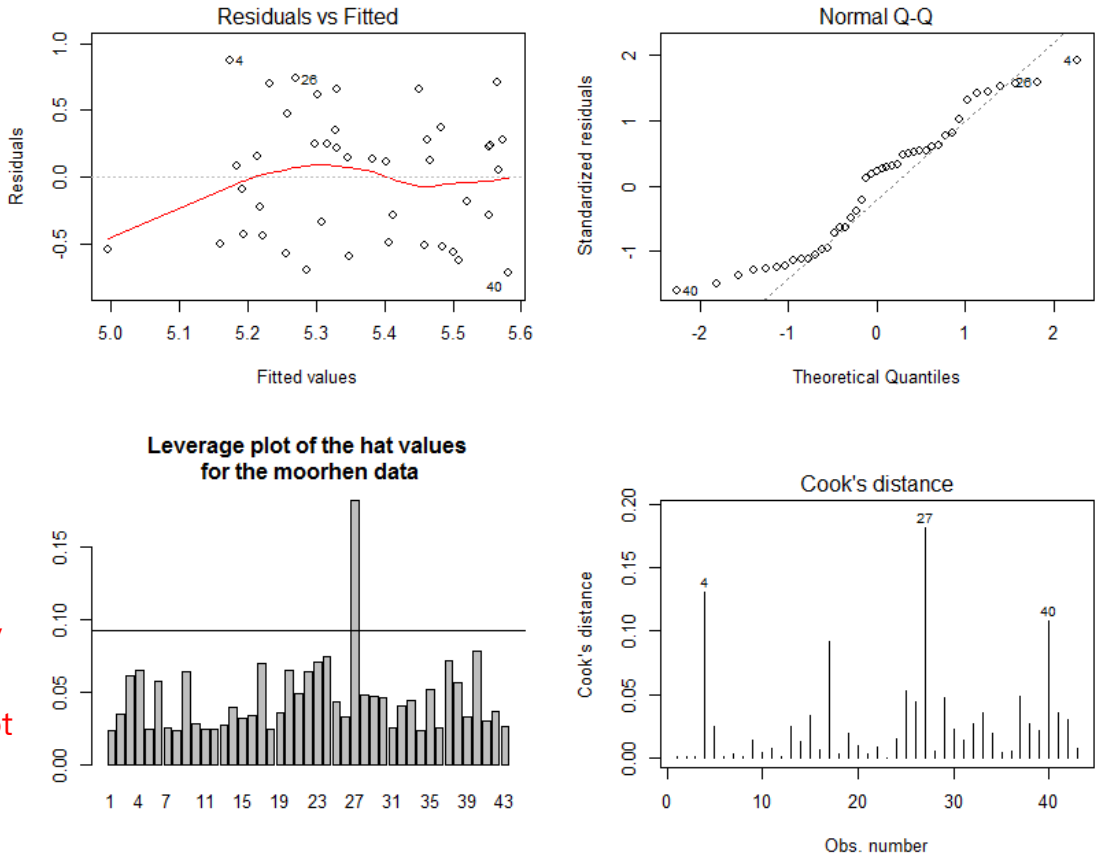



```
> cor.test(log(Weight), log(Shield))

   Pearson's product-moment correlation

data:  log(Weight) and log(Shield)
t = 1.9709, df = 41, p-value = 0.05552
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.006763403  0.546257547
sample estimates:
     cor
0.294178
```

After testing all nine possible combinations, the one shown above had the strongest sample correlation and the smallest *p*-value, but it is still not a significant correlation.

Question 1 continued

(d) Fit a simple linear regression (SLR) model with your chosen transformation of Shield as the response variable and your chosen transformation of Weight as the predictor. Construct a plot of the residuals against the fitted values, a normal Q-Q plot of the residuals, a bar plot of the leverages for each observation and a bar plot of Cook's distances for each observation. Use these plots to comment on the model assumptions and on any unusual data points. **(3 marks)**



automatically highlighted points are not necessarily problematic

The main (residuals vs fitted values) plot does not show any serious problems with either the assumption of independence or the assumption of constant variance of the errors. Similarly, the normal (Q-Q) quantile plot does not show any serious departure from the assumption of normally distributed errors, given the relatively small sample size (the residual distribution is possibly a little "fatter" than expected in the lower tail).

Despite the default labelling of the three observations with the largest residuals (in absolute value), the only observation that stands out is the one towards the left of the main residual plot, which is somewhat remote from the other observations in the horizontal direction (the direction of the fitted values).

This is observation 27, the juvenile bird identified in part (a), which does have relatively high leverage (as shown on the leverage bar plot) and also has the largest value for Cook's distance (as shown on the bar plot of Cook's distances). However, the Cook's distance for observation 27 is not notably extreme relative to the next largest Cook's distance, which suggests that this observation is still following the general trend of the other observations and has not been particularly influential in determining the estimated regression equation (or the overall "fit" of the model).

in relative terms…

Question 1 continued

(e) Produce the ANOVA (Analysis of Variance) table for the transformed SLR model in part (d) and interpret the results of the F test. What is the coefficient of determination for the model and how should you interpret this summary measure? **(3 marks)**

```
> anova(moorhen.lm)
Analysis of Variance Table
Response: log(Shield)
            Df Sum Sq Mean Sq F value  Pr(>F)
log(Weight)  1 0.8569 0.85689  3.8843 0.05552 .
Residuals   41 9.0447 0.22060
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

$$H_0 : \frac{\sigma^2_{Model}}{\sigma^2_{Error}} = 1 \quad H_A : \frac{\sigma^2_{Model}}{\sigma^2_{Error}} > 1 \quad ; \quad R^2 = \frac{0.8569}{0.8569 + 9.0447} \approx 0.08654$$

$F_{1,41} = 3.8843$, $p > 0.05$, so do NOT reject H_0 in favour of H_A and conclude the variance explained by the model is not large compared to the error variance, i.e. the model involving log(Weight) is not explaining a significant proportion of the variability in log(Shield).

Judging by the coefficient of determination (R^2), the model involving log(Weight) is explaining less than 9% of the variability in log(Shield).

(f) What are the estimated coefficients of the transformed SLR model in part (d) and the standard errors associated with these coefficients? Interpret the values of these estimated coefficients and perform t-tests to test whether or not these coefficients differ significantly from zero. What do you conclude as a result of these t-tests? **(3 marks)**

```
> summary(moorhen.lm)
Call:
lm(formula = log(Shield) ~ log(Weight))
Residuals:
    Min      1Q  Median      3Q     Max
-0.7196 -0.4674  0.1076  0.2787  0.8769
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.4844     3.4757  -0.427   0.6716
log(Weight)   1.0599     0.5378   1.971   0.0555 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.4697 on 41 degrees of freedom
Multiple R-squared:  0.08654,    Adjusted R-squared:  0.06426
F-statistic: 3.884 on 1 and 41 DF,  p-value: 0.05552
```

$$\text{Model}: \log(\text{Shield}) = \beta_0 + \beta_1 \log(\text{Weight}) + \varepsilon \quad \varepsilon \sim i.i.d.\ N\left(0, \sigma^2\right)$$

$$H_0 : \beta_1 = 0 \quad H_A : \beta_1 \neq 0$$

$t_{41} = 1.971$, $p > 0.05$, so do NOT reject H_0 in favour of H_A and conclude that the slope coefficient of log(Weight) is not significantly different from 0, implying there is not a strong (linear) relationship between log(Shield) and log(Weight). This is essentially the same test as the test on the correlation between log(Shield) and log(Weight) in part (c).

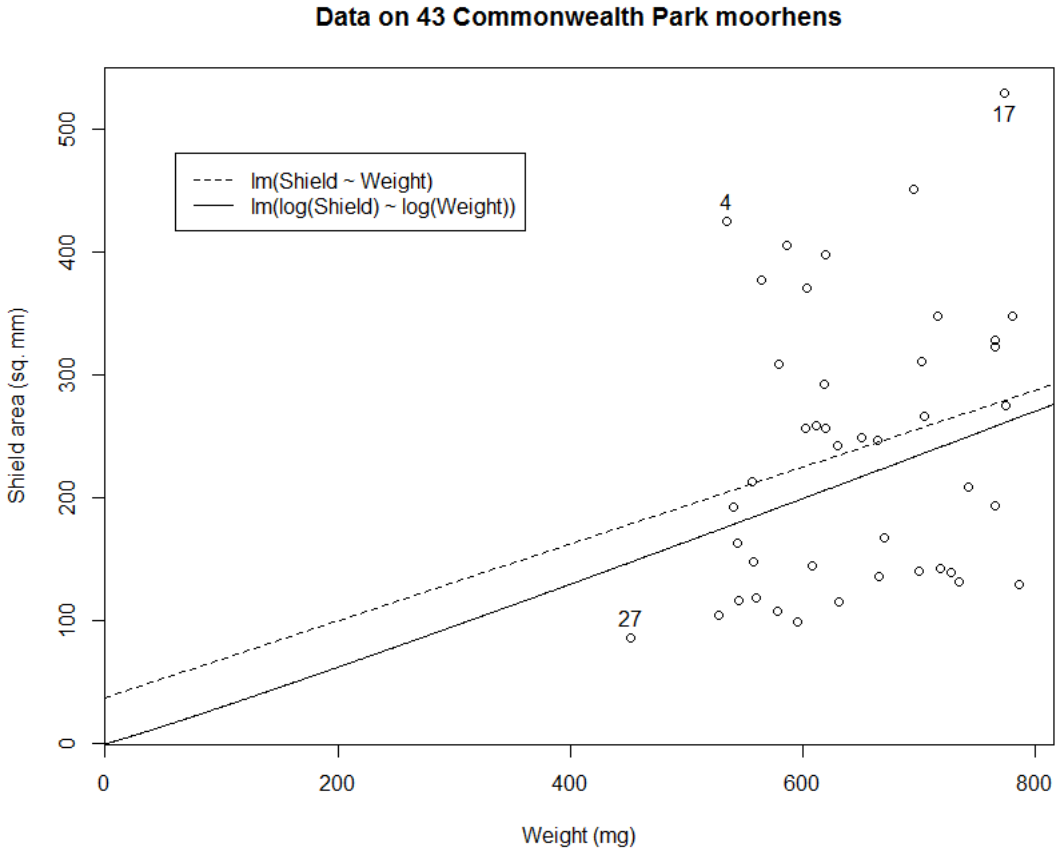
$$H_0 : \beta_0 = 0 \quad H_A : \beta_0 \neq 0$$

$t_{41} = -0.427$, $p > 0.05$, so do not reject H_0 in favour of H_A and conclude that the intercept is also not significantly different from 0. Note that the log(Weight) values range from 6.11 to 6.66, so the intercept is a considerable distance from the data and we probably shouldn't try to interpret the value of the intercept in this instance, it is fitted simply to allow maximum flexibility in the way that the model fits the data.

extrapolation

Question 1 continued

(g) Repeat part (a) and again plot Shield against Weight, but this time extend both X and Y axes to include the origin. Now include the transformed SLR model from part (d) as a curve on your plot and also include the untransformed SLR of Shield against Weight as a line on the plot. Use different line types for the two curves and also include an appropriate legend on the plot. What are you overall conclusions about the relationship between Weight and Shield, and the broader research questions discussed in the second paragraph of this question? **(3 marks)**



Neither model looks particularly convincing and we did try a number of modelling options. The fact that none of them resulted in a significant relationship leads us to suspect that there is not a significant relationship between Shield and Weight, so it appears that heavier birds do not necessarily have bigger shields.

This does not really address the question of whether bigger birds have bigger shields, as Weight is not the only possible measure of the size of the birds (there might be fat and heavy birds, but who are short and have relatively small shields). The dataset does include alternatives measures of the size of the birds and we will experiment with these in a multiple regression model in Assignment 2. Weight, though not significant in a simple linear regression model, may still be a useful addition to a multiple regression model with Shield as the response.

The deeper research question of which is the better indicator of social status amongst moorhens (shield size or overall size) cannot really be addressed using this dataset, as none of the included variables appears to be a measure of social status.

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 include a derived variable, BMI or body mass index, which is frequently used as a measure of obesity and is based on simple weight and height measurements.

For this assignment, we are interested in whether or not BMI, which is relatively easy to measure, can be used to predict the percentage of body.fat, which has to be estimated using an underwater weighing technique?

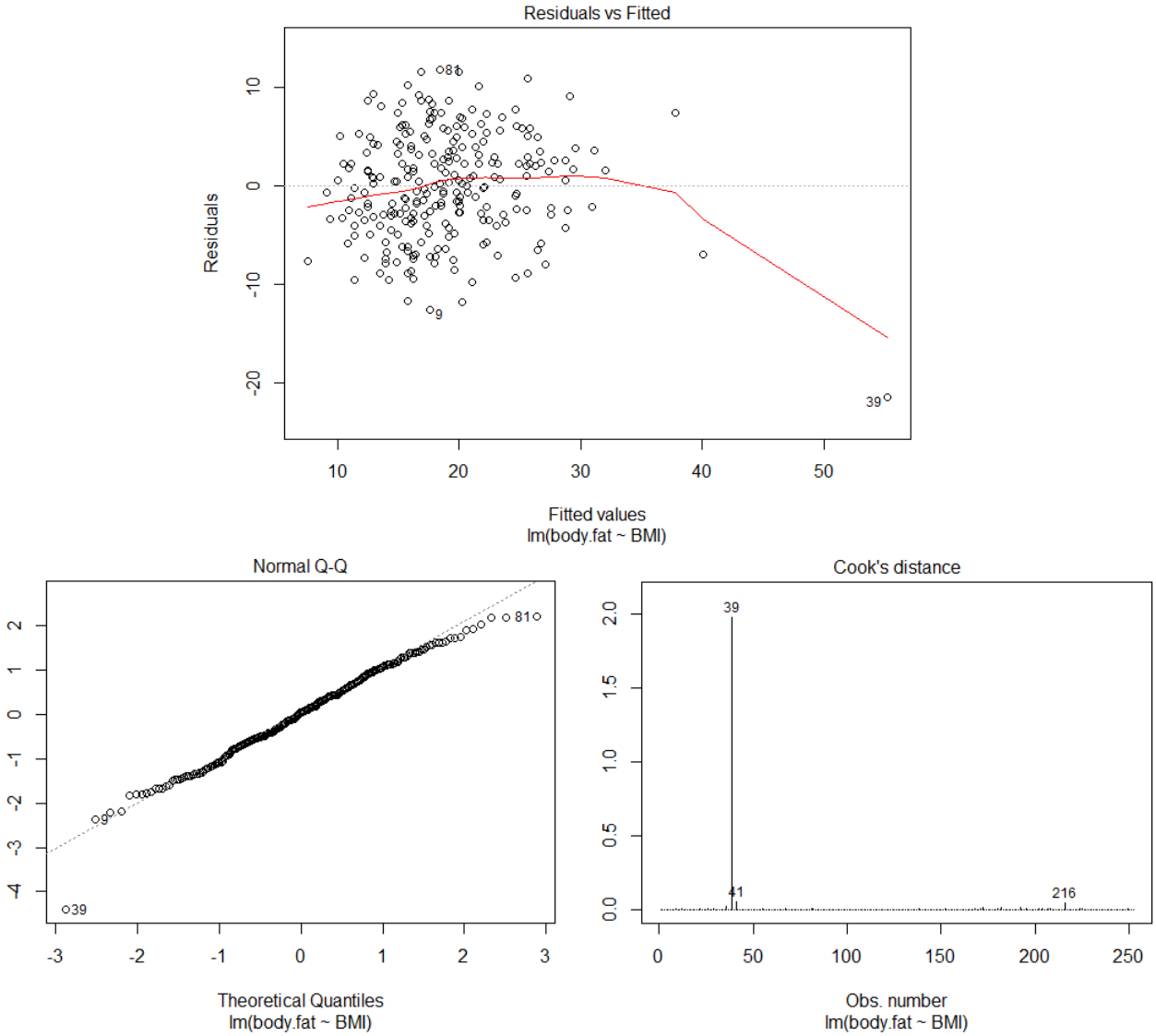
(a) Plot body.fat against BMI. Describe the correlation shown in the plot. Would you expect a simple linear regression model to be a reasonable model for the relationship shown in the plot? **(4 marks)**

The graph shows a reasonably strong moderate positive correlation, with the exception of the three observations with the highest BMI values, which turn out to be cases 39, 41 and 216. I would expect a simple linear regression model to be a reasonable first approximation, but I would also expect to have to deal with the above three cases as possible outliers and/or modify the model using some sort of transformation.

Question 2 continued

(b) Fit a simple linear regression (SLR) model with body.fat as the response variable and BMI as the predictor. Construct a plot of the residuals against the fitted values, a Q-Q plot of the residuals and a bar plot of Cook's Distances for each observation. Comment on the model assumptions and on any unusual data points. **(4 marks)**

```
> fat.lm <- lm(body.fat ~ BMI)
> fat.lm

Call:
lm(formula = body.fat ~ BMI)

Coefficients:
(Intercept)          BMI
    -20.405        1.547
```



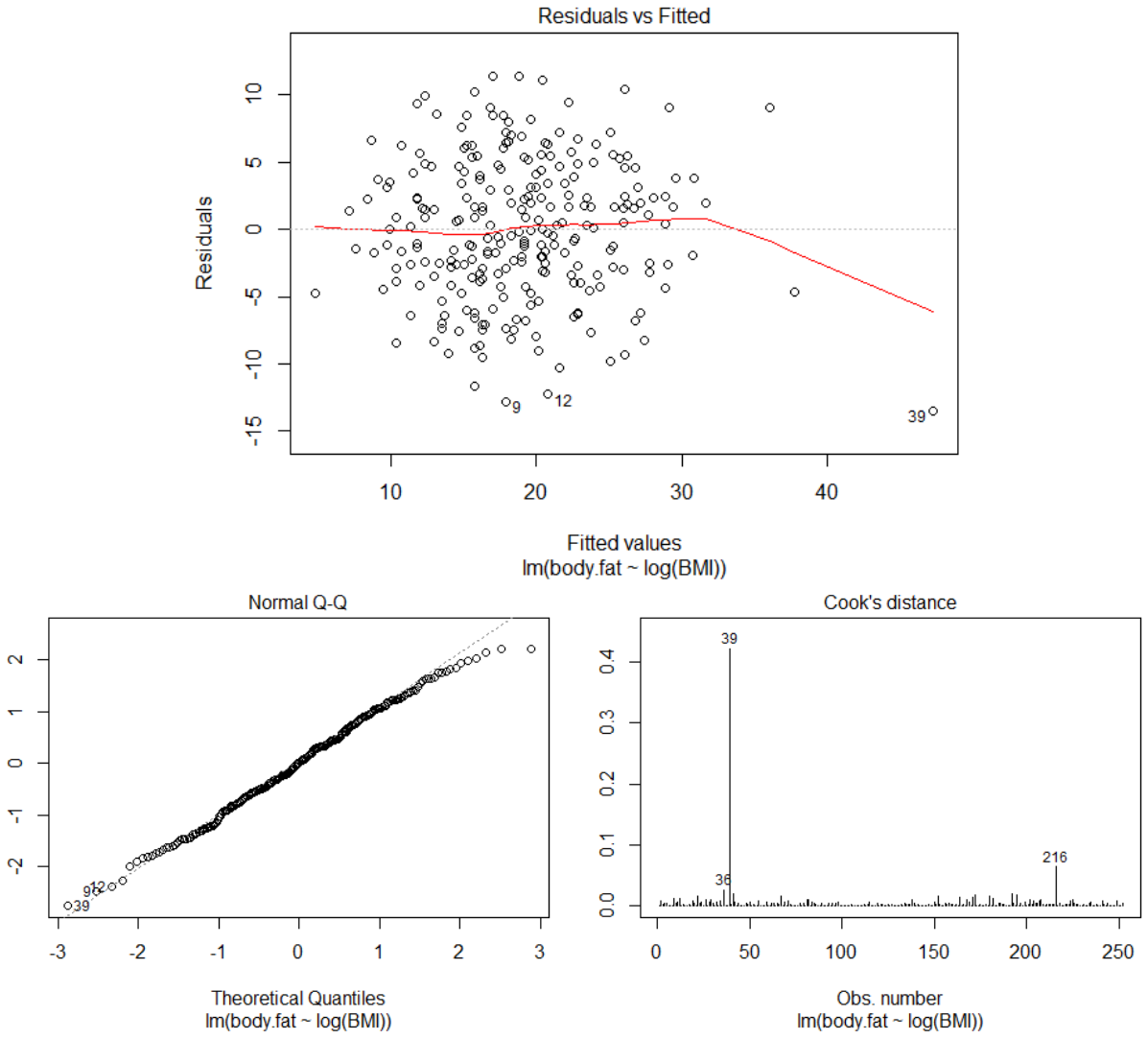
As expected, there are definitely problems with the three observations with large BMI, all three of which have been labelled on the "Cook's distance" plot, though observation 39 is currently the only one that really stands out in relative terms.

It is a little hard to tell if the assumptions are satisfied, given these obvious problems, but it may be better to try a transformation before we go down the route of declaring observations to be outliers and simply remove them from the data.

Question 2 continued

(c) A natural log (to the base *e*) transformation (to one or both of the response and predictor variables) is often used to adjust the scale of the variables prior to fitting an SLR model. Now fit another SLR model with body.fat as the response variable and log(BMI) as the predictor. What would be the problem with also applying a log transformation to the response variable? Check the same plots you produced for the earlier model in part (b). Are the same problems still apparent? **(4 marks)**

```
> fat.loglm <- lm(body.fat ~ log(BMI))
> fat.loglm

Call:
lm(formula = body.fat ~ log(BMI))

Coefficients:
(Intercept)     log(BMI)
    -119.23        42.82
```



The problem with also trying to log the body.fat measurements is that there is at least one observation (case 182) with a 0 value.

These are definitely better than the previous set of plots, though case 39 (and possibly also case 216) are still potential outliers/influential points and require further investigation and possible treatment (we will have another look at these problems in Assignment 2).

Question 2 continued

(d) Produce the ANOVA table and the table of the estimated coefficients for the transformed SLR model in part (c). Interpret the values of the estimated coefficients for this SLR model and the results of the overall F test and the t-tests on the estimated coefficients. **(4 marks)**

$$\text{Model}: \text{body.fat} = \beta_0 + \beta_1 \log(\text{BMI}) + \varepsilon \quad \varepsilon \sim i.i.d.\ N\left(0, \sigma^2\right)$$

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

Response: body.fat
           Df Sum Sq Mean Sq F value    Pr(>F)
log(BMI)    1 8386.5  8386.5  313.28 < 2.2e-16 ***
Residuals 250 6692.6    26.8
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

$$H_0 : \frac{\sigma^2_{Model}}{\sigma^2_{Error}} = 1 \quad H_A : \frac{\sigma^2_{Model}}{\sigma^2_{Error}} > 1$$

$F_{1,250} = 313.3$, $p << 0.05$, 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 involving log(BMI) is explaining a significant proportion of the variability in body.fat.

```
> summary(fat.loglm)

Call:
lm(formula = body.fat ~ log(BMI))

Residuals:
     Min       1Q   Median       3Q      Max
-13.5263  -3.3776   0.0751   3.8273  11.4306

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -119.233      7.813  -15.26   <2e-16 ***
log(BMI)      42.820      2.419   17.70   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.174 on 250 degrees of freedom
Multiple R-squared:  0.5562,     Adjusted R-squared:  0.5544
F-statistic: 313.3 on 1 and 250 DF,  p-value: < 2.2e-16
```

$$H_0 : \beta_1 = 0 \quad H_A : \beta_1 \neq 0$$

$t_{250} = 17.70$, $p << 0.05$, so reject H_0 in favour of H_A and conclude that the slope coefficient of log(BMI) is significantly different from 0, implying there is a significant linear relationship between body.fat and log(BMI). Note this test is directly equivalent to the above F test. The interpretation of this slope coefficient (42.82) is the expected increase in body.fat as log(BMI) increases by 1. An increase of 1 on the log scale is equivalent to multiplying BMI by a factor of e, which is not the easiest thing to interpret on the original scale, but the slope is positive, implying that body.fat is increasing as BMI increases (log is an order preserving transformation).

$$H_0 : \beta_0 = 0 \quad H_A : \beta_0 \neq 0$$

$t_{250} = -15.26$, $p << 0.05$, so reject H_0 in favour of H_A and conclude that the intercept is also significantly different from 0. Note that log(BMI) values range from 2.9 to 3.9, so as in question 1, we probably shouldn't try to interpret the actual value of the intercept in this instance, as a value of 0 would represent extrapolation a long way outside the range of our data.

Question 2 continued

(e) Body mass index values less than 18.5 are typically categorised as "underweight"; from 18.5 to 25 as "normal", 25 to 30 as "overweight" and over 30 as "obese". Use the transformed SLR model from part (c) to predict the body.fat percentage for groups of males with typical BMI values 17.25 ("moderately underweight"), 21.75 ("normal"), 27.5 ("overweight") and 32.5 ("moderately obese"), respectively. Find 95% confidence intervals for these predictions. Do you think this SLR model is a good model for making these predictions? **(4 marks)**

```
> logBMI <- log(BMI)
> fat.newloglm <- lm(body.fat ~ logBMI)
> newvalues <- log(c(17.25, 21.75, 27.5, 32.5))
> conf.ints <- predict(fat.newloglm, newdata=data.frame(logBMI=newvalu
es), interval="confidence")
> conf.ints
        fit        lwr       upr
1  2.709611  0.7930648  4.626157
2 12.635297 11.6845100 13.586085
3 22.679621 21.9145334 23.444709
4 29.832835 28.4611122 31.204557
```

The confidence intervals are reasonably narrow, but I am not yet convinced that this is a good model to use for prediction, at least not until we have resolved some of the issues with the residual plots that we identified in part (c).