STAT6038 Assignment 1 for 2017

Rui Qiu

2017-03-12

Question 1

Executive Summary

In this question, we are interested in the data of 43 observed moorhens. Specifically, we will investigate the relationship between `Weight`(unit in mg) and `Shield`(unit in mm^2) as two main variables of moorhens. Our goal is to determine which one of those two variables is more likely to be an indicator of the social status of a moorhen.

(a)

The plot is shown as below. We also use `identify()` command to manually select unusual data point on the left lower corner. The plot indicates such point is labeled with index 27 in our sample.

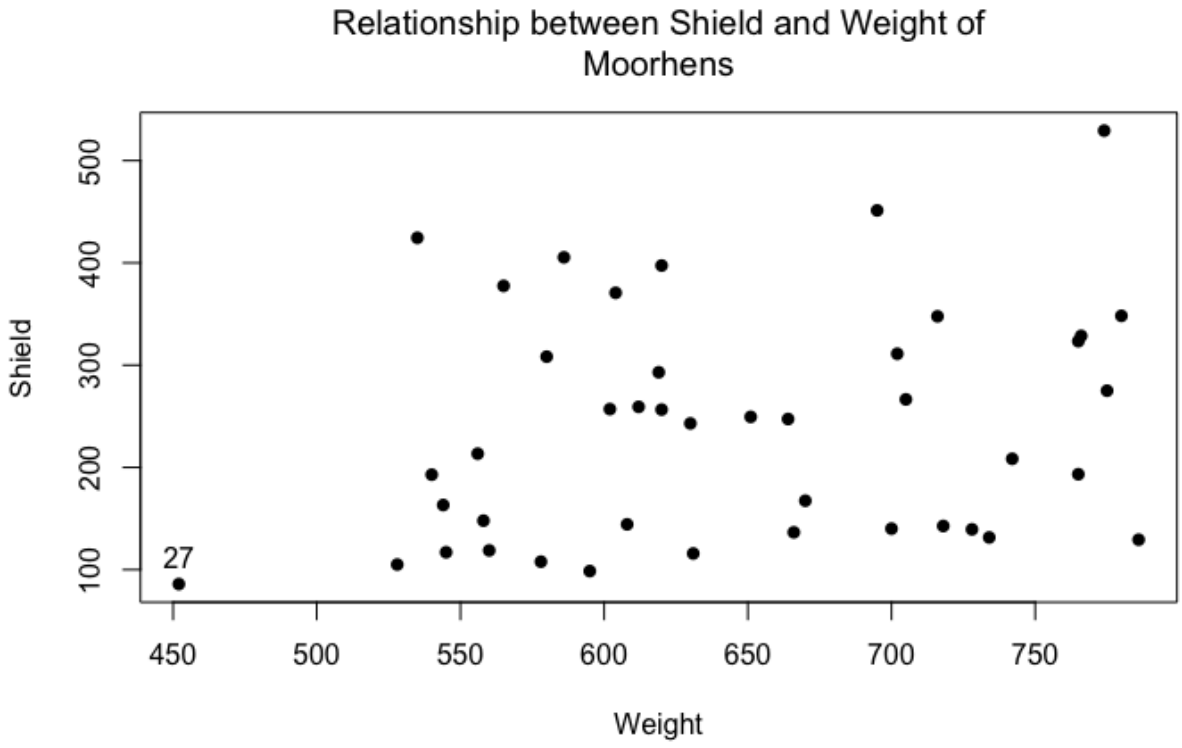


Figure 1:

The reason that we believe the moorhen with data with index 27 is an unusual is that, among all moorhens with similar `Shield` area, its has a lowest `Weight`. But the data points did not show an obvious pattern, so the evidence that it is unusual is not very strong.

(b)

```
##
##  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
```

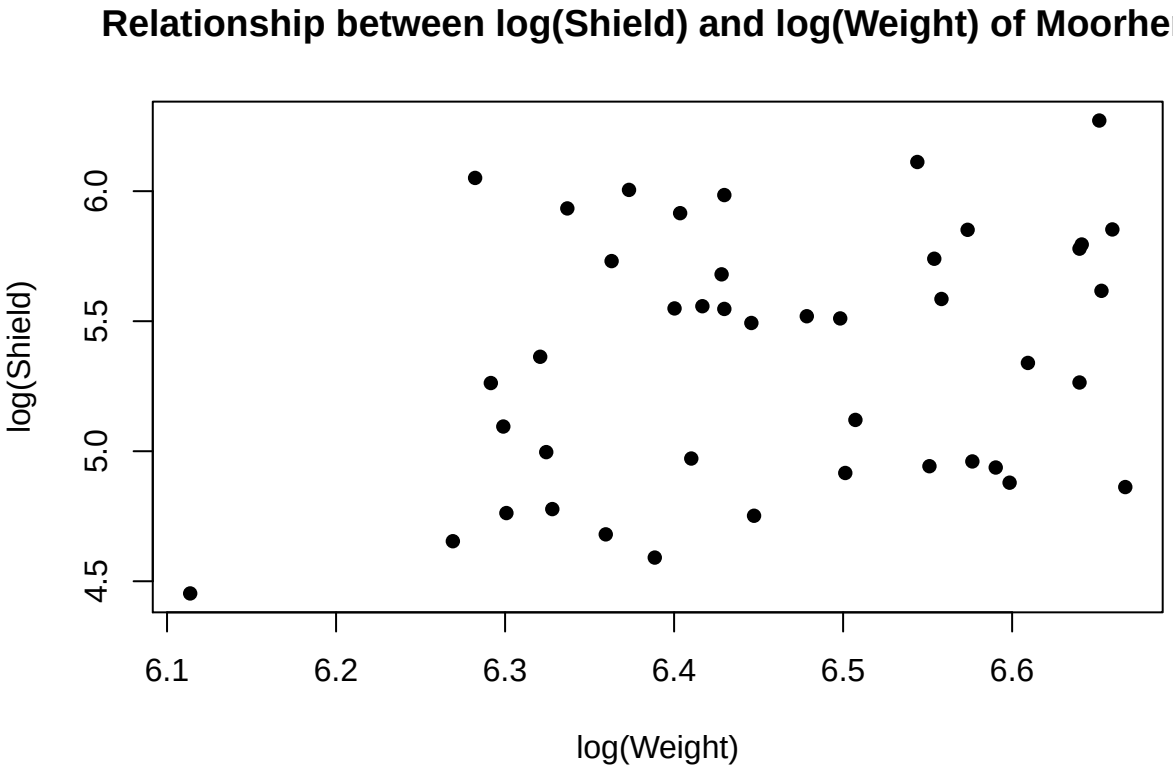
The hypotheses tested are:

- $H_0 : \rho = 0$, i.e. there is no corrleation between variable `Weight` and `Shield`.
- $H_A : \rho \neq 0$, i.e. there is some correlation.

Since $t_{95} = 1.5793,\ p = 0.122 > 0.05$, so we fail to reject H_0, so we cannot say there is a significant correlation between `Shield` and `Weight`.

(c)

After generating 8 different combinations of `log()` and `sqrt()` transformations on `Weight` and `Shield`, we pick




```
##
##  Pearson's product-moment correlation
##
## data:  log(Weight) and log(Shield)
```

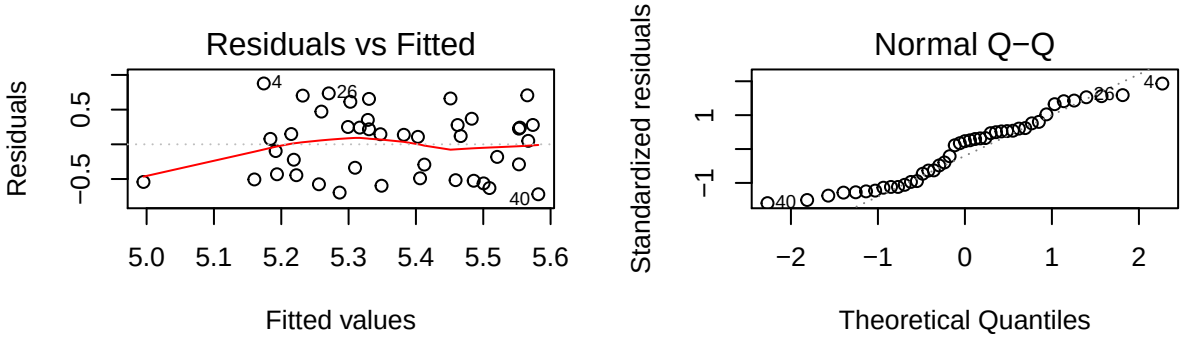
```
## t = 1.9709, df = 41, p-value = 0.05552
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.006763403  0.546257547
## sample estimates:
##      cor
## 0.294178
```

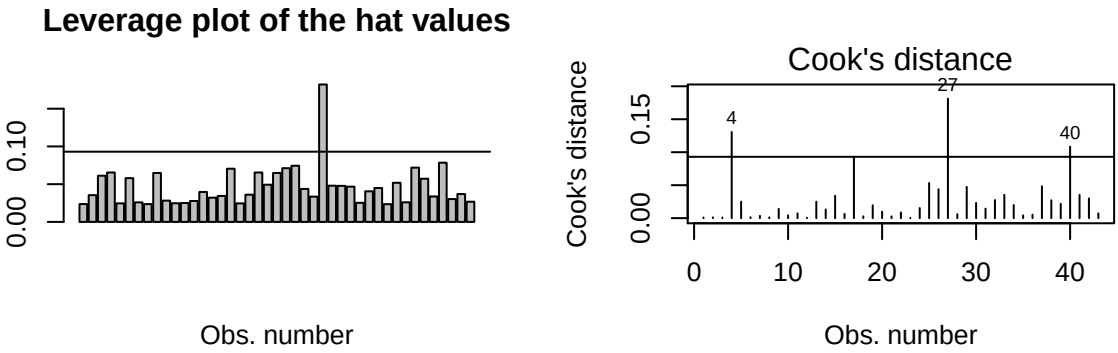
According to the results of t test, the case `log(Shield)` vs. `log(Weight)` has the lowest p-value, but still slightly above 0.05 (in fact it's 0.05552).

Still, for the same reason, we fail to reject null hypothesis. In other words, we can say that there isn't a significant evidence of correlation between `log(Shield)` and `log(Weight)`.

(d)

```
##
## Call:
## lm(formula = log(Shield) ~ log(Weight))
##
## Coefficients:
## (Intercept)  log(Weight)
##      -1.484        1.060
```





First we should have an impression that residuals, leverages, and Cook's distances are 3 measurements of influential statistics.

- The residual measure "outlierness", the distance between observations and its fitted value.
- The leverage measures how far away the independent variable values of an observation are from those of the other observations, it tells us about its potential to influence the fit.
- The Cook's distance measure influence of each data value on the fit.

Specifically,

- **The residuals vs. fitted plot** indicates the observations with index `4, 26, 40` are outliers. If we ignore these 3 data points, there are almost equally spread residuals around the horizontal line without distinct patterns.
- **The normal Q-Q plot** highlights that `4, 26, 40` again do not fall on the diagonal line. Generally speaking, the whole plot is slightly "S-shaped", which means the distribution is light-tailed comparing with normal distributions. And this contradicts the assumption that all errors are normally distributed.
- **The leverage barplot** shows that the observation `27` is an unusual data points whose leverage is high.
- **The Cook's distances barplot** shows that the observations `4, 27, 40` have high Cook's distances, which means they have great influence on our SLR.

Also different diagnostics outlines different unusual data points. Probably we have to reconsider the our model.

(e)

```
## 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
```

Here we have hypotheses:

- $H_0 : \frac{\sigma^2_{regression}}{\sigma^2_{Error}} = 1.$
- $H_A : \frac{\sigma^2_{regression}}{\sigma^2_{Error}} > 1.$

According to the results of F test:

$F_{1,95} = 3.8843,\ p = 0.05552 > 0.05$, so fail to reject null hypothesis that $\frac{\sigma^2_{regression}}{\sigma^2_{Error}} = 1.$

The result is consistent with the t-test we performed previously. In the end, for SLR, t-test and F-test are equivalent tests with:

$$\text{F-statistics} = 3.884 = 1.9709^2 = \text{t-statistics}^2$$

The coefficient of determination for this model (R^2) is calculated as $\frac{0.8569}{0.8569+9.0447}$:

```
## [1] 0.08654157
```

The coefficient of determination is 0.08654, which is the proportion of total variation of outcomes explained by our model. In this case, this proportion is not high at all, so the variation cannot be mainly explained by model. Variation explained by errors also plays an import role in our case, then our model could be problematic.

(f)

```
##
## 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
```

The estimated coefficients of the SLR model are $\hat{\beta}_0 = -1.4844, \hat{\beta}_1 = 1.0599$

The standard errors are $se(\beta_0) = 3.4757, se(\beta_1) = 0.5378$.

Our model is: `log(Shield)`$= \beta_0 + \beta_1$`log(Weight)`$+\epsilon, \epsilon \overset{iid}{\sim} N(0, \sigma^2)$.

First we do t-test on β_0:

- $H_0 : \beta_0 = 0$,
- $H_A : \beta_0 \neq 0$.

$t_{95} = -0.427,\ p = 0.6716 > 0.05$. Fail to reject H_0, we claim that β_0 (the intercept) is not significantly different from 0.

Then we do t-test on β_1:

- $H_0 : \beta_1 = 0$,
- $H_A : \beta_1 \neq 0$.

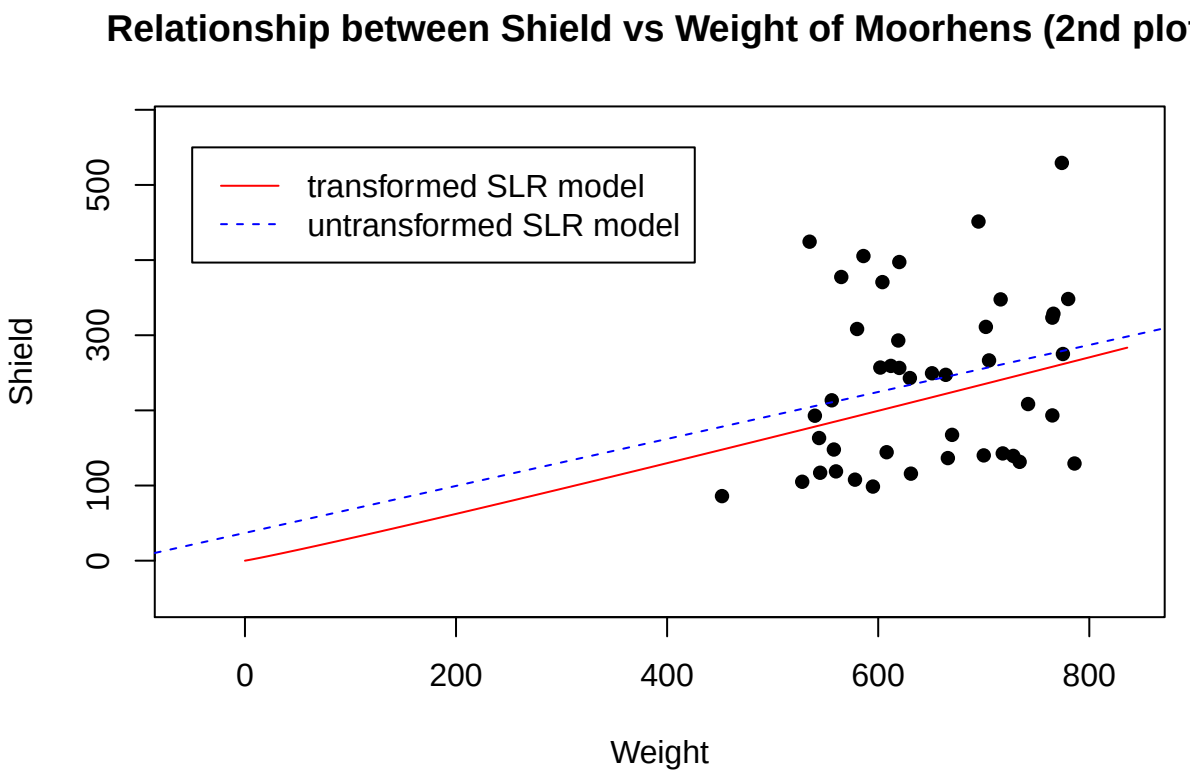
$t_{95} = 1.971,\ p = 0.0555 > 0.05$. Again fail to reject H_0, and we claim that β_0 (the coefficient of `Weight`) is not significantly different from 0.

Therefore, it looks like that the value of variable `log(Shield)` does not have a significant relationship with the variable `log(Weight)`. Our model definitely needs more improvement.

(g)

For the regression line between `log(Shield)` and `log(Weight)`, if we want to plot it in the original plot, we need to do the following 'backward' transformation.

$$\ln(Shield) = \beta_0 + \beta_1 \cdot \ln(Weight)$$
$$Shield = e^{\beta_0 + \beta_1 \cdot \ln(Weight)}$$



As we can see from the plot, the two lines standing for two different SLR models basically agree with each other. The untransformed model even looks better, with similar amounts of data points on two sides on the line.

But the problem is, the hypothesis tests' results for corrleation between `log(Weight)` vs `log(Shield)` and corrleation between `Weight` vs `Shield` are not good. In other words, there might be not very significant relationship between the two variables of interest. And back to the plot we have here, the regression lines seem to cut this flock of data points into halves, but there is not any obvious linear trend in these data points.

After all, neither of these two models is good enough to reflect the true relationship between `Weight` and `Shield`. And we can do the following to improve our model:

- We should include more data points (larger sample size) to increase the range of `Weight` (e.g. more data with `Weight` less than 400, to cover the blank space in the plot above).
- A more complex model (or to include more variables from the data set) would be introduced.

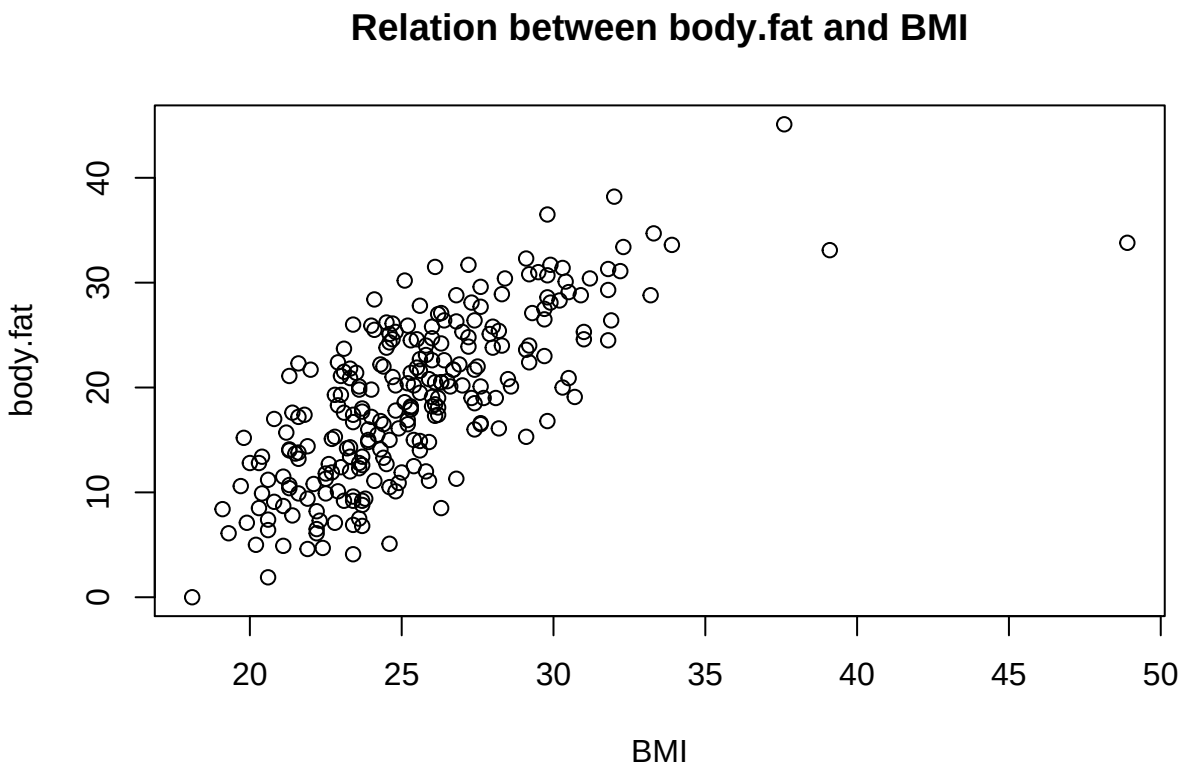
Question 2

Executive Summary

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?

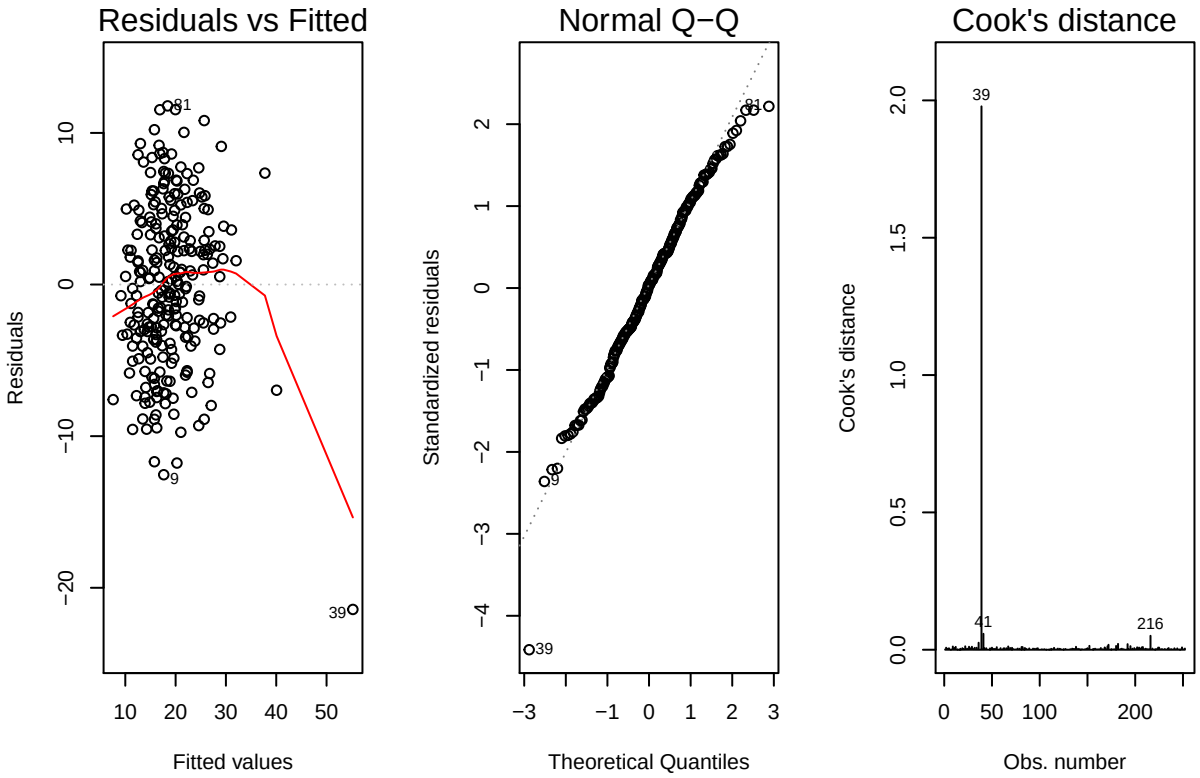
(a)



Looks like the two variables `body.fat` and `BMI` share a plausible SLR model as the two have a generally linear trend, i.e. when `BMI` increases, `body.fat` increases correspondingly and linearly.

(b)

```
##
## Call:
## lm(formula = body.fat ~ BMI)
##
## Coefficients:
## (Intercept)          BMI
##     -20.405        1.547
```

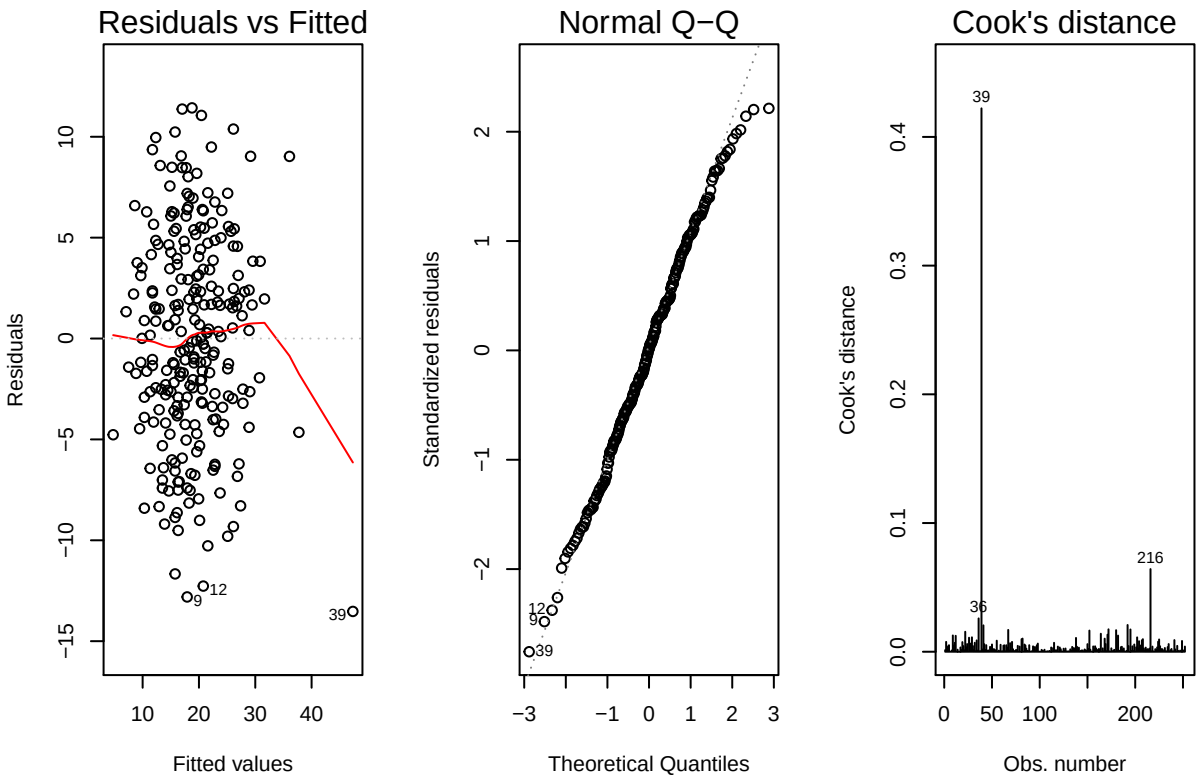


- In residual plot, the amounts of data points above and below the line are basically equal. But the 39th observation is way below the line, which is problematic.
- In normal Q-Q plot, some data points (e.g. 39th observation) don't lie on a diagonal line, which does not support the assumption that all errors are normally distributed.
- In Cook's distance plot, the value of the 39th observation still outnumbers other observations, which means the 39th observation has a great impact on the fitted model.

All three plots above indicate that data with index 39 is unusual, which has a great influence on our model.

(c)

```
##
## Call:
## lm(formula = body.fat ~ log_BMI)
##
## Coefficients:
## (Intercept)      log_BMI
##     -119.23        42.82
```



If we also apply a log transformation to `body.fat`, the program would not proceed successfully, since the minimum in `body.fat` is `0`, and `log(0)` is a singularity.

Comparing with previous plots, the 39th observation could still be an unusual data point in our SLR model between `body.fat` and `log(BMI)`.

(d)

```
## 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
```

```
##
## 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
```

The estimated coefficients of the SLR model are $\hat{\beta}_0 = -119.233, \hat{\beta}_1 = 42.820$.

In details, β_0 stands for the baseline of this SLR model. For example, it is the value of `body.fat` when `log(BMI)` is zero (although this might be impossible in reality). β_1 stands for the amount of increament in `body.fat` when `log(BMI)` increases by 1 unit.

For F test, we are testing the following hypotheses:

- $H_0 : \frac{\sigma^2_{regression}}{\sigma^2_{Error}} = 1.$
- $H_A : \frac{\sigma^2_{regression}}{\sigma^2_{Error}} > 1.$

$F_{1,95} = 313.28,\ p = 2.2 \times 10^{-16} \ll 0.05$, so reject H_0 in favour of H_A and conclude that the variance explained by the model is greater than the error variance, which means the model involving `log(BMI)` explains a significant proportion of the variability in `body.fat`.

We have the following model: `body.fat`$= \beta_0 + \beta_1$`log(BMI)`$+\epsilon, \epsilon \overset{iid}{\sim} N(0, \sigma^2)$.

For t test on β_0

- $H_0 : \beta_0 = 0,$
- $H_A : \beta_0 \neq 0.$

$t_{95} = -15.26,\ p \ll 0.05$, we reject H_0 in favour of H_A and conclude that β_0 (the intercept) is significantly different from 0.

Then we do t test on β_1:

- $H_0 : \beta_1 = 0,$
- $H_A : \beta_1 \neq 0.$

$t_{95} = 17.70,\ p = \ll 0.05,$, again we are to reject H_0 in favour of H_A, and we claim that β_0 (the slope of `log(BMO)`) is significantly different from 0.

So our model is pretty plausible.

(e)

```
##         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
```

Our SLR model from part (c) looks like a very good model for predictions, since all four 95% confidence intervals have no overlaps. Also, the span of each interval is not quite large, so we consider the predictions precise.

In this case, we believe around 95% of data with those critical `BMI` values should fall into our confidence intervals of `body.fat` correspondingly.

Appendix

```
# Prepare Data
moorhen <- read.csv('moorhen.csv', header = TRUE)
# As we only need Weight and Shield in this assignment, we ignore the other variables.
moorhen <- moorhen[,1:2]
attach(moorhen)

plot(Weight, Shield, pch=16, main="Relationship between Shield and Weight of Moorhens")

identify(Weight, Shield)

cor.test(Weight, Shield)

# loop for different combinations of transformations
sv <- cbind(Shield, log(Shield), sqrt(Shield))
wv <- cbind(Weight, log(Weight), sqrt(Weight))
varNames <- c("", "log", "sqrt")

i = 1; j = 2; count = 0
while (count != 3**2-1) {
        while (j <= 3) {
                plot(wv[,j], sv[,i], pch=16,
                     #xlim=c(0, max(wv[,j])), ylim=c(0,max(sv[,i])),
                     main=paste("Relationship between", varNames[i], "Shield and",
                                varNames[j], "Weight of Moorhens"),
                     xlab = paste(varNames[j], "Weight"),
                     ylab = paste(varNames[i], "Shield"))
                print(cor.test(wv[,j], sv[,i]))
                j <- j + 1
                count <- count + 1
        }
        i <- i + 1
        j <- 1
}

plot(log(Weight), log(Shield), pch=16,
     main="Relationship between log(Shield) and log(Weight) of Moorhens")
cor.test(log(Weight), log(Shield))

(moorhen.lm <- lm(log(Shield) ~ log(Weight)))

# Residuals vs Fitted, Normal Q-Q, Cook's Distances
par(mfrow=c(2,2))
plot(moorhen.lm, which=c(1,2,4))
# Here we use 4/n as the cut-off value for spotting highly influential points.
abline(h=4/length(log(Shield)))

# Leverage Barplot
barplot(hat(log(Shield)), main="Leverage plot of the hat values", xlab = "Obs. number")
abline(h=2*sum(hat(log(Shield)))/length(log(Shield)))

anova(moorhen.lm)
```

```r
(r2 <- 0.8569/(0.8569+9.0447))

summary(moorhen.lm)

plot(Weight, Shield, pch=16,
     xlim = range(-50, max(Weight)+50),
     ylim = range(-50, max(Shield)+50),
     main="Relationship between Shield vs Weight of Moorhens (2nd plot)")
moorhen.lm2 <- lm(Shield ~ Weight)
abline(moorhen.lm2$coefficients, lty=2, col="blue")

beta0 <- as.numeric(moorhen.lm$coefficients[1])
beta1 <- as.numeric(moorhen.lm$coefficients[2])
newXrange <- seq(0, max(Weight)+50, by=2)
lines(newXrange, exp(1)**(beta0+beta1*log(newXrange)), lty=1, col = "red")
legend(-50, 550, c("transformed SLR model", "untransformed SLR model"),
       lty=1:2, col=c("red","blue"))

bf <- read.csv('fat.csv')
attach(bf)
plot(BMI, body.fat, main="Relation between body.fat and BMI")

(bf.lm <- lm(body.fat ~ BMI))
par(mfrow=c(1,3))
plot(bf.lm, which=c(1,2,4))

log_BMI <- log(BMI)
(bflog.lm <- lm(body.fat ~ log_BMI))
par(mfrow=c(1,3))
plot(bflog.lm, which=c(1,2,4))

# min(body.fat)
# The following would cause an Error:
# loglog.lm <- lm(log(body.fat) ~ log_BMI)

anova(bflog.lm)
summary(bflog.lm)

crit_logbmi <- log(c(17.25, 21.75, 27.5, 32.5))
(predictions <- predict(bflog.lm, newdata = data.frame(log_BMI=crit_logbmi),
                        interval = "confidence"))
```