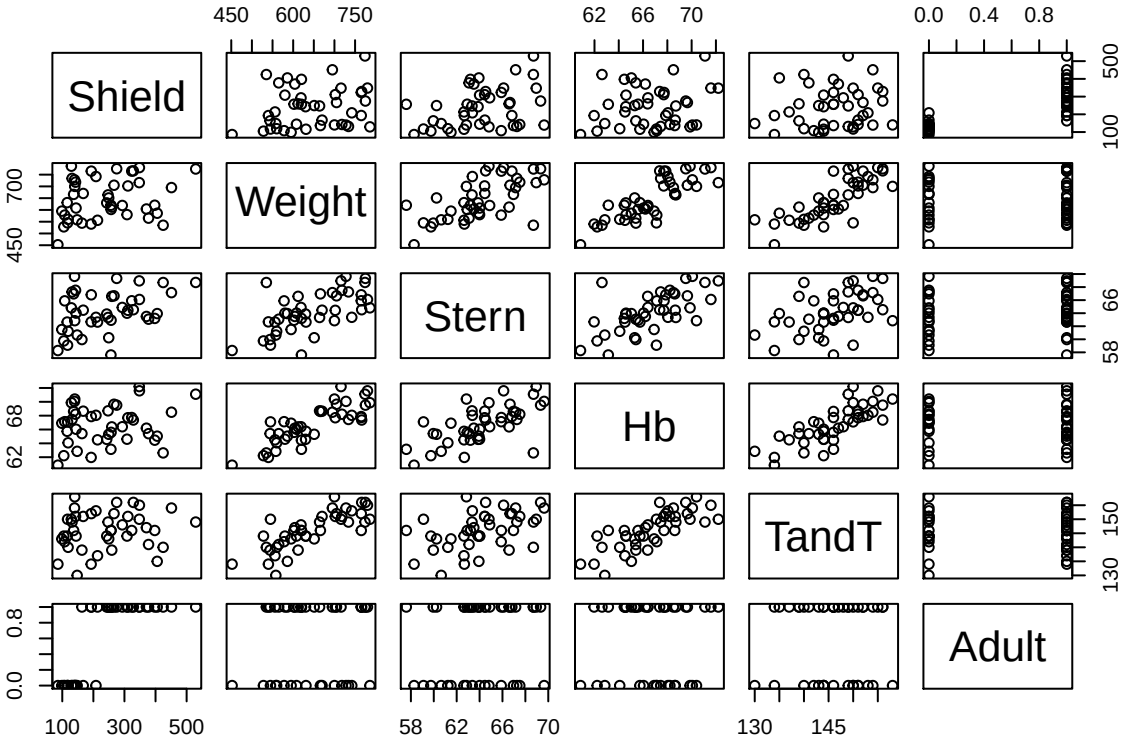
STAT6038 Assignment 2

Rui Qiu, Ming Zhang

2017-05-10

Question 1

(a)




```
##           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
```

Generally speaking, only the following pairs of predictors seem to have linear relations:

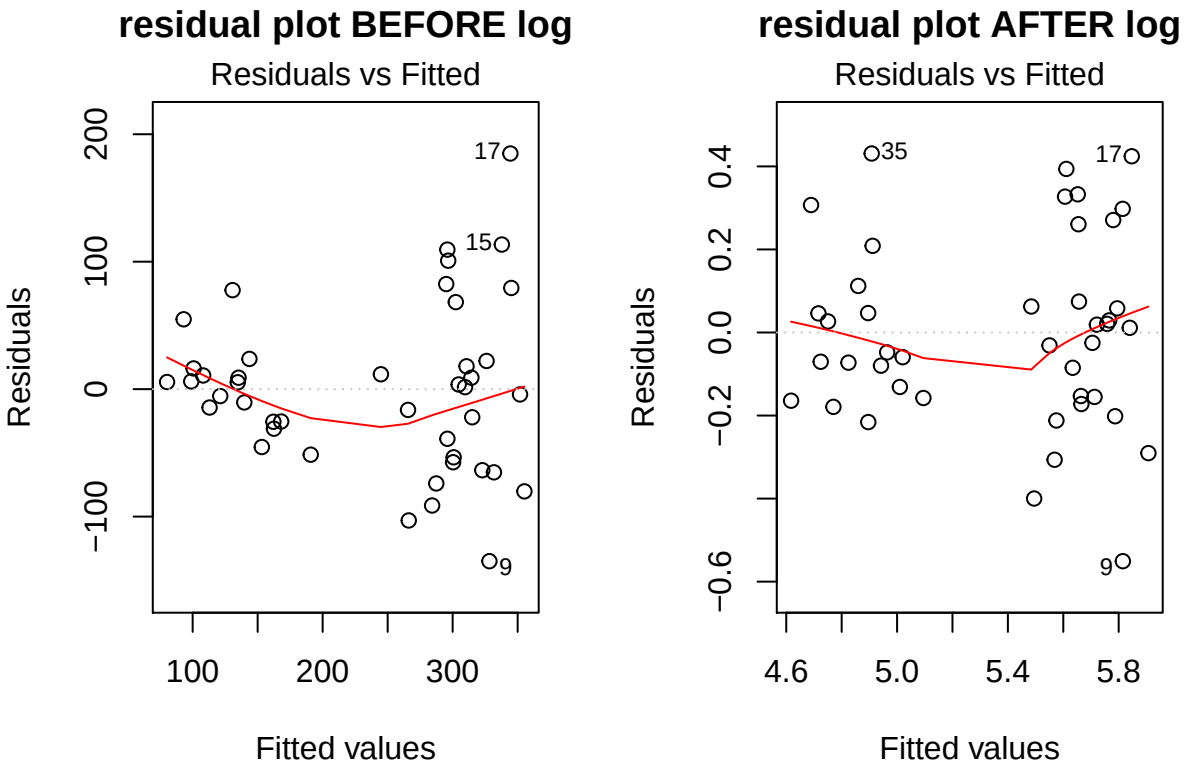
- Shield vs. Adult
- Weight vs. Stern
- Weight vs. Hb
- Weight vs. TandT
- Stern vs. Hb
- Hb vs. Tandt

Note: The variable `Adult` is an indicator variable, thus we only care about if different predictor inputs lead to different responses. But it looks like only the variable `Shield` differs dramatically. For `Stern` also looks

like adult moorhens have seemingly higher `Stern` values, but just not as obvious as `Shield`, i.e. generally adult moorhens have larger shield area than juvenile moorhens.

And the correlation matrix confirms our statement above, since only the mentioned 6 pairs of variables have correlations greater than 0.5 in the matrix.

(b)



The residuals vs. fitted values plot for our first multiple liner regression model seems to not satisfying the assumption of homoscedasticity, i.e. the variance of residuals is not constant. That's we can detect the existence of 'funnel'-shape in the residual plot. However, the assumption of zero-mean residual seems ok as no intense curvature observed.

Also, the sample data points are clustered basically on two different sides of the plot, thus leaving the centre of the plot quite blank. As we know, an ideal normally distributed residual plot should have a rectangular or ecliptical shape, while ours does not look anything like that. Hence we also check the normal qq-plot, and find out it is not fitting a straight line.

(c)

This time the assumption of homoscedasticity is fixed, since the scale of the 'funnel'-shape is smaller, i.e. not spreading far away as before. But still, the centre lacks of data points, leaving the overall shape like a 'dumbbell'. We will look into this issue later.

(d)

Examine the log-transformed model in part (c): The ANOVA table shows that only insignificant additional variable is `TandT` with p-value greater than 0.05, others are fine. But the summary table (t-test) infers that

every predictor variable is not significant, except the last one `Adult` which has the same p-value as it in ANOVA (F-test).

Nested hypotheses testing (with ANOVA tables):

```
## 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
```

```
(0.0402 + 0.0001 + 0.0129)/3/0.0573
```

```
## [1] 0.3094823
```

```
(0.0001 + 0.0129)/2/0.0573
```

```
## [1] 0.113438
```

```
(0.0129)/0.0573
```

```
## [1] 0.2251309
```

```
qf(0.95, 3, 37)
```

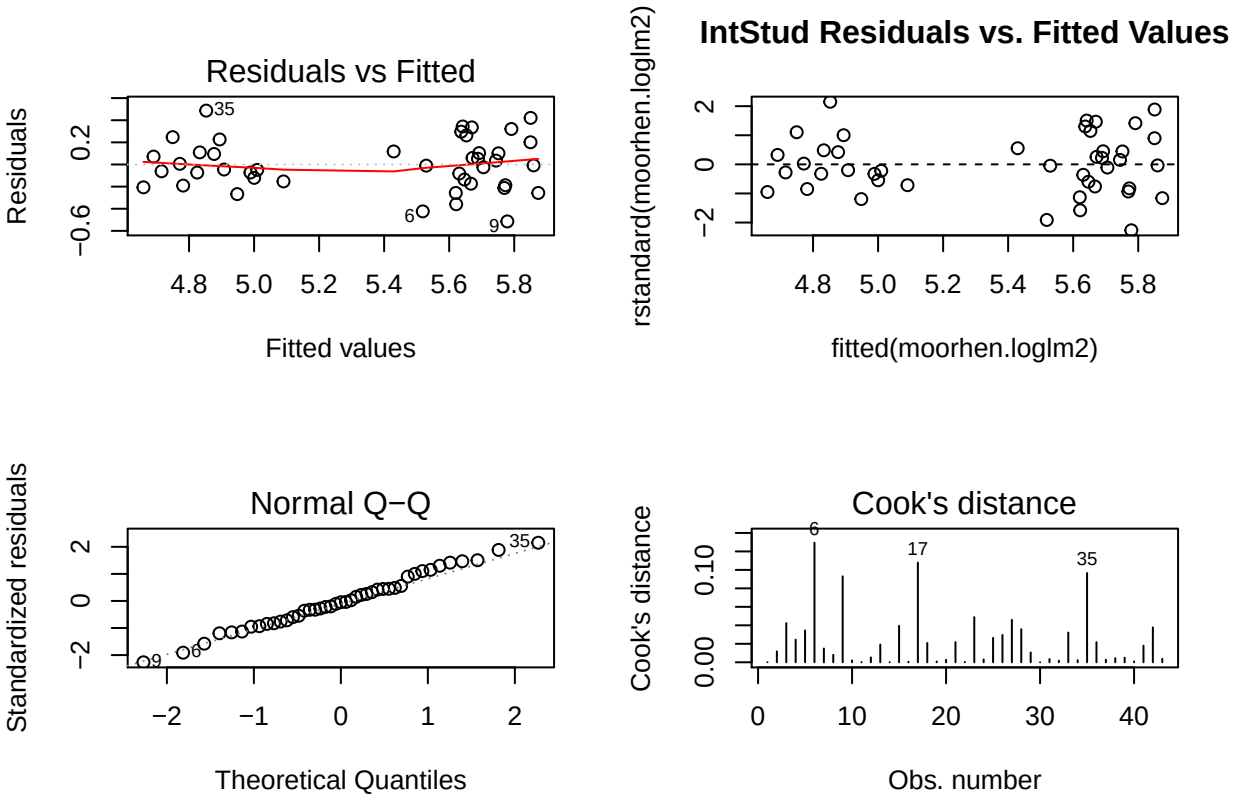
```
## [1] 2.858796
```

We calculated the theoretical 95% F-statistics, which is $F_{3,37} = 2.858796$.

- $H_0 : \beta_{\text{Weight}} = \beta_{\text{Hb}} = \beta_{\text{TandT}} = 0$
  - Given `Stern, Adult` included in the model, the inclusion of variables `Weight, Hb, TandT` has F-statistics $0.3094823 < 2.858796$, so they provide no significantly additional information.
- $H_0 : \beta_{\text{Hb}} = \beta_{\text{TandT}} = 0$
  - Given `Stern, Adult, Weight` included in the model, the inclusion of variables `Hb, TandT` has F-statistics $0.113438 < 2.858796$, so they provide no significantly additional information.
- $H_0 : \beta_{\text{TandT}} = 0$
  - Given `Stern, Adult, Weight, Hb` included in the model, the inclusion of variable `TandT` has F-statistics $0.2251309 < 2.858796$, so it provides no significantly additional information.

To conclude, we could possibly modify the model by removing predictors `Weight,Hb,TandT` in the following parts of this question.

(e)

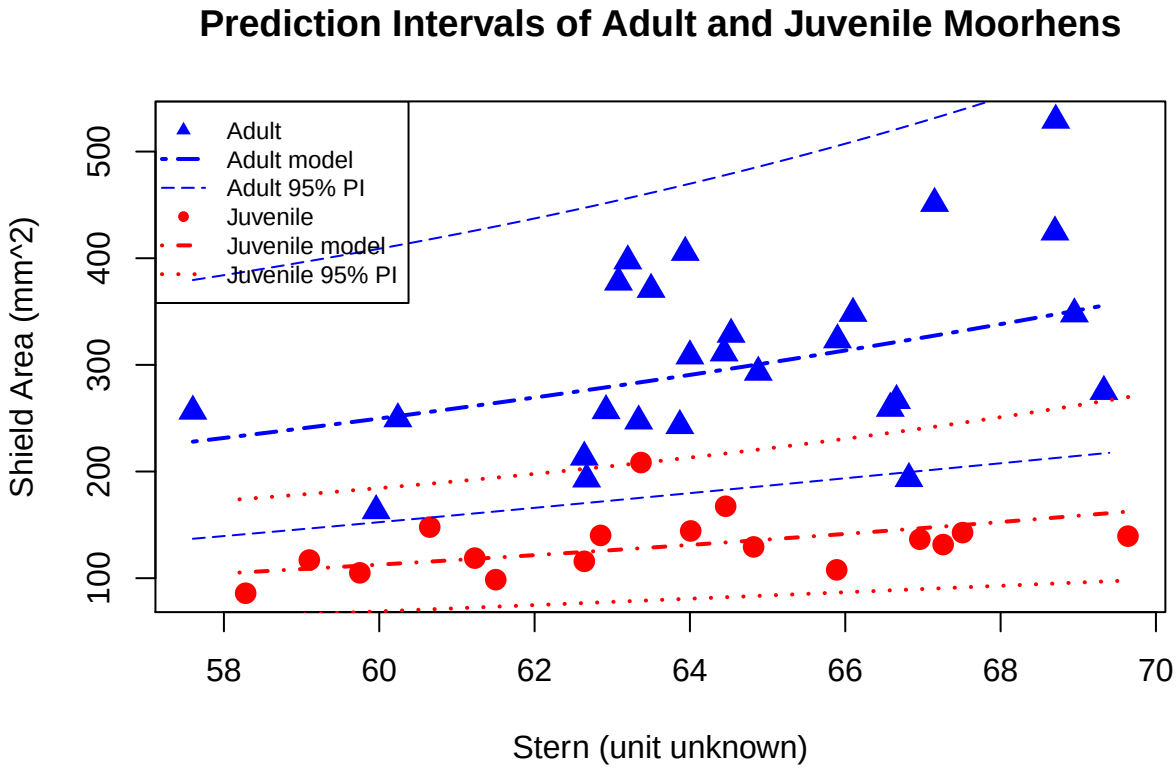


- Internally Studentised (standardised) residual plot: Data points generally spread into a rectangle, no obvious outliers observed. In fact, if we check the scale-location plot with default `plot()` function, we would find that then there is a general increasing trend from the left hand side to the right hand side. This might be explained by the fact that adult moorhens usually have larger `log(Shield)` value than juvenile ones.
- Normal qq-plot: Even less curved than the one we ploted in our first log-transformed model, and no obvious problem observed. Thus it satisfies the normal distribution assumption.
- Cook's Distance plot: The highest three Cook's distances are only around 0.12, and are not even relatively higher than any other points. So no influential points observed here. Also we could check Residuals vs Leverage plot to confirm our statement. (In fact, no data points are beyond the cut-off boundary.)

In addtion, we also include main residual plot besides the three requires plots above Based on the main residual plot, we believe our model satisfies the assumption of independent errors.

But back to our question itself, since we have indicator variables labeling every one of moorhens (either adult or juvenile), we probably would like to display this feature and try to investigate further if the dumbbell-shape clustering is related with this. Thinking about this from common sense, adult bird usually is larger in size than juvenile bird.

(f)



The plot surely captured the general trend of adult and juvenile moorhens' shield area, that is, adults usually have larger shield area. But note that, the two prediction intervals have overlaps.

(g)

$$\begin{aligned}\text{Ratio} &= \frac{e^{\beta_0+\beta_{\text{Adult}}\cdot 1+\beta_{\text{Stern}}\cdot\text{Stern}}}{e^{\beta_0+\beta_{\text{Adult}}\cdot 0+\beta_{\text{Stern}}\cdot\text{Stern}}} \\ &= e^{\beta_{\text{Adult}}}\end{aligned}$$

Since the ratio seems to be a fixed value, then the 95% confidence interval of this estimated ratio is in fact, an exponential of 95% confidence interval of parameter β_{Adult}. Therefore, we have

```
## (Intercept)       Adult       Stern
##  2.45029558  0.79531521  0.03791127
```

```
##           2.5 %    97.5 %
## Adult 0.6459868 0.9446436
```

```
##          2.5 %   97.5 %
## Adult 1.907869 2.571897
```

Question 2

(a)

- NOT including `case`, `body.fat.siri` or `density`:
  - `case`: This is just the index of each man, it has no real impact on the real body fat percentage.

  - `body.fat.siri`: This is just another expression of body fat. We have already included `body.fat` as the response, there is no need to include it as a predictor. If we do so, it would be affect the power of other predictors. (In other words, we are using body fat to fit body fat.)
  - `density`: According to Bozek's equation, body.fat $= \frac{457}{\text{Density}} - 414.2$, since we are trying to find a regression model for `body.fat`, we should not include a variable which is a direct transformation of that response in the predictors.
- Including all three `weight`, `height`, `BMI` at the same time: No need to include all of three variables, since `BMI` can be expressed as `weight/height^2`, so `BMI` is highly correlated with those two variables. We can keep `weight` and `height`, excluding `BMI`.
- Including `ffweight` when `weight` is included: No need to do this as well. `ffweight` is both related with `weight` and involved with `body.fat`. Pretty dangerous to include it.

(b)

Recall that `ln(BMI)` was a pretty good simple linear regression predictor in Assignment 1, and this problem itself addresses that `weight` is a considered as a key and should be included. By the formula

$$\begin{aligned}\log(\text{BMI}) &= \log\left(\frac{\text{weight}}{\text{height}^2}\right) \\ &= \log(\text{weight}) - 2\log(\text{height})\end{aligned}$$

So we put variable `ln(weight), ln(height)` into our draft model.

Next, we tend to add variables into this model. We add the variables left one by one, by checking the variance inflation factor (VIF), we only confirm the addition if no VIFs of current variables exceed 10 (conventionally we pick 10 as a cut-off value). In other words, we only add variables if we can avoid multicolinearity.

Repeat until we have a full model, in this case, which is `body.fat ~ log(weight) + log(height) + wrist + neck + age + forearm + bicep + ankle + knee`.

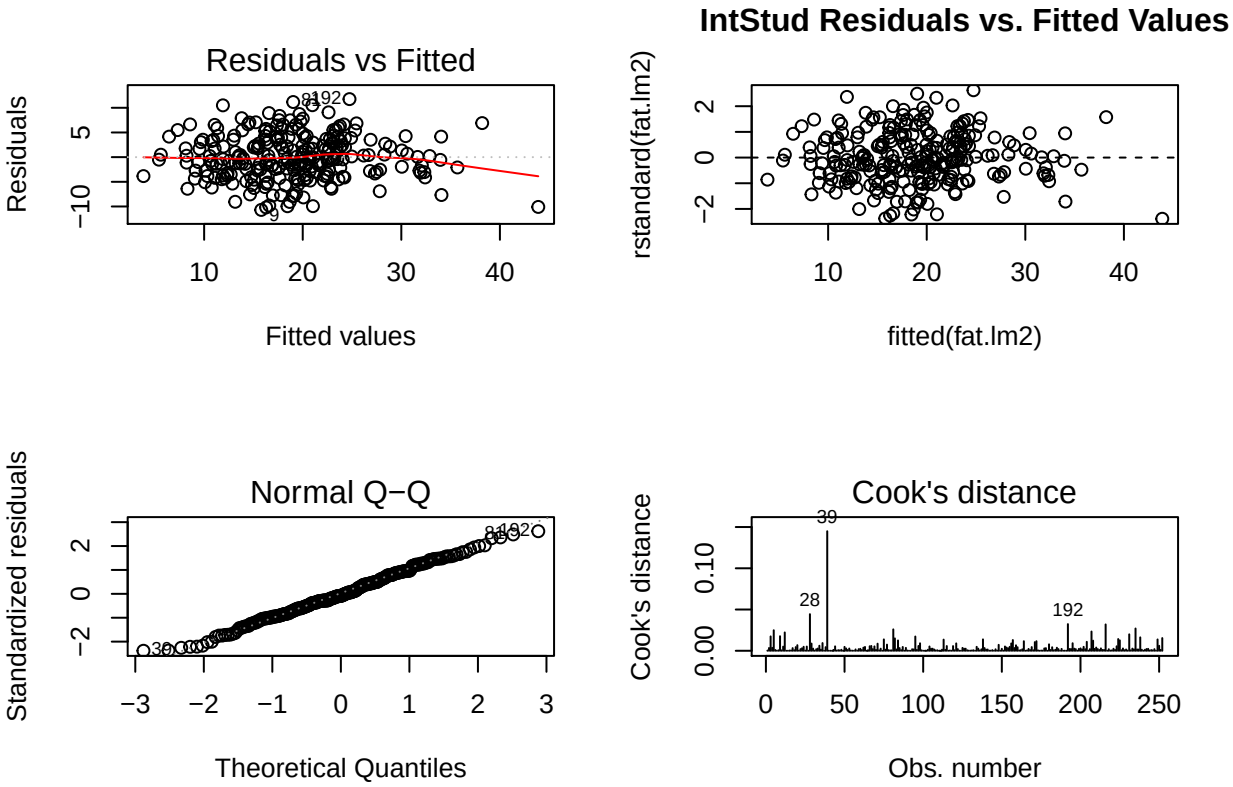
Then we check the ANOVA and `summary()` table, both suggests the last four variables are insignificant. And we think about it with common sense that, there is not much body fat in those parts of human body. So we would like to drop those variables.

Now the model is `body.fat ~ log(weight) + log(height) + wrist + neck + age`. Again we check the ANOVA and `summary()` table. Now every variable is significant.

```
## Analysis of Variance Table
##
## Response: body.fat
##              Df Sum Sq Mean Sq  F value    Pr(>F)
## log(weight)   1 5977.5  5977.5 288.7938 < 2.2e-16 ***
## log(height)   1 2447.4  2447.4 118.2406 < 2.2e-16 ***
## wrist         1  381.3   381.3  18.4225 2.547e-05 ***
## neck          1   99.0    99.0   4.7823    0.0297 *
## age           1 1082.0  1082.0  52.2754 6.073e-12 ***
## Residuals   246 5091.8    20.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Call:
## lm(formula = body.fat ~ log(weight) + log(height) + wrist + neck +
##     age)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.6935  -3.0179  -0.3545   3.3000  11.7562
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 115.74571   36.46230   3.174  0.00169 **
## log(weight)  60.05545    3.77094  15.926  < 2e-16 ***
## log(height) -80.91097    9.69494  -8.346 5.14e-15 ***
## wrist        -2.74117    0.51820  -5.290 2.71e-07 ***
## neck         -0.57301    0.22995  -2.492  0.01337 *
## age           0.18350    0.02538   7.230 6.07e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.55 on 246 degrees of freedom
## Multiple R-squared:  0.6623, Adjusted R-squared:  0.6555
## F-statistic:  96.5 on 5 and 246 DF,  p-value: < 2.2e-16
```

Afterwards, we plot and examine these three diagnostic plots:

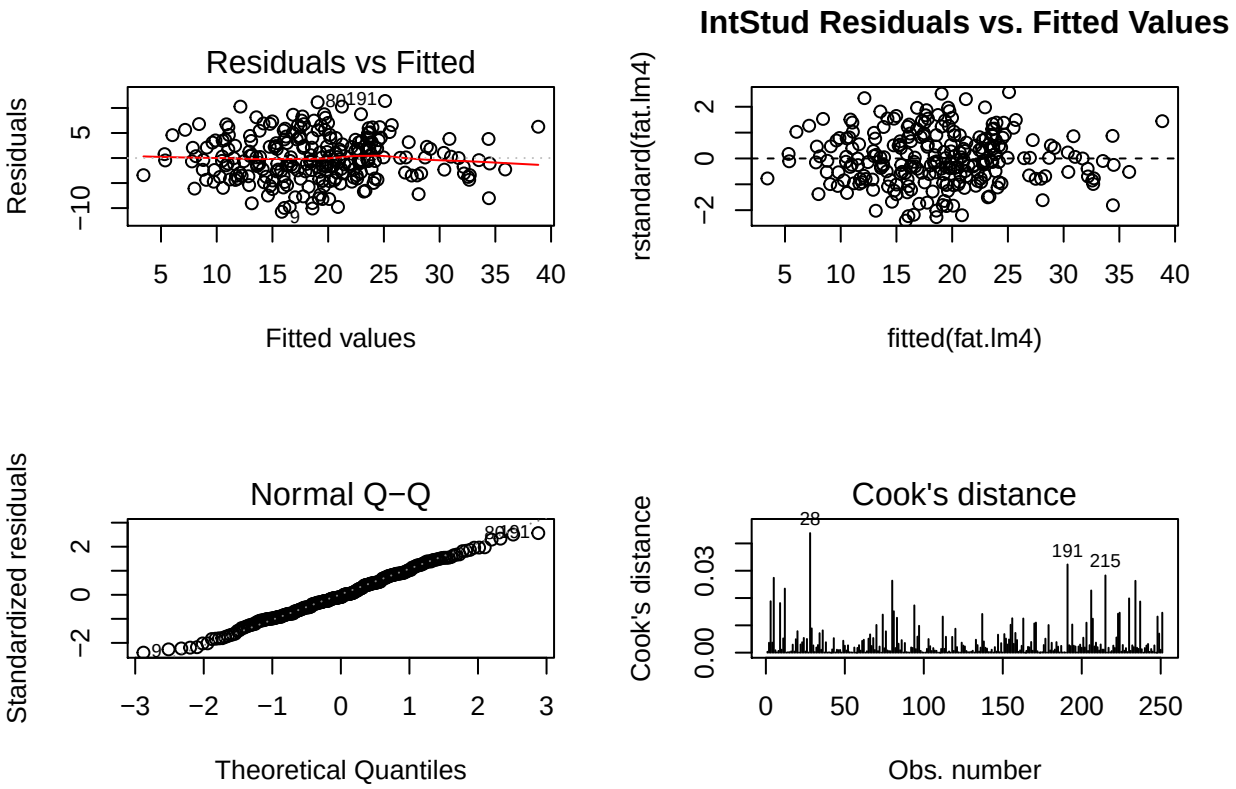


For here, we also include a main residual plot with some curvature on the right hand side, but generally the model satisfies the independent errors assumption.

- The internally studentized residuals plot looks fine, the data points cluster generally in a rectangular range. The only thing suspicious is the data point (case number 39) in the downright corner, which might be an potential outlier.
- The normal-qq looks a little bit light tailed, but generally we consider it satisfying the assumption.
- The Cook's distances plot suggests that the case number 39 stands out comparing with all other data. It probably is high leverage which affects the model dramatically.

(c)

```
## Analysis of Variance Table
##
## Response: body.fat[-39]
##                   Df Sum Sq Mean Sq  F value    Pr(>F)
## log(weight)[-39]   1 5816.1  5816.1 286.4627 < 2.2e-16 ***
## log(height)[-39]   1 2579.9  2579.9 127.0670 < 2.2e-16 ***
## wrist[-39]         1  377.5   377.5  18.5950 2.344e-05 ***
## neck[-39]          1   56.7    56.7   2.7923     0.096 .
## age[-39]           1 1052.7  1052.7  51.8502 7.326e-12 ***
## Residuals        245 4974.3    20.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Call:
## lm(formula = body.fat[-39] ~ log(weight)[-39] + log(height)[-39] +
##     wrist[-39] + neck[-39] + age[-39])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.7378  -3.0796  -0.4016   3.2972  11.3939
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      118.71664   36.13380   3.285  0.00117 **
## log(weight)[-39]  60.56758    3.74084  16.191  < 2e-16 ***
## log(height)[-39] -82.87330    9.63657  -8.600 9.64e-16 ***
## wrist[-39]        -2.80081    0.51383  -5.451 1.22e-07 ***
## neck[-39]         -0.46871    0.23184  -2.022  0.04429 *
## age[-39]           0.18113    0.02515   7.201 7.33e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.506 on 245 degrees of freedom
## Multiple R-squared:  0.6652, Adjusted R-squared:  0.6584
## F-statistic: 97.35 on 5 and 245 DF,  p-value: < 2.2e-16
```



According to the diagnostic plots, we tend to delete the case 39. After doing so, we refit the model and find out that in ANOVA table, the predictor `neck` becomes insignificant with p-value $0.096 > 0.05$. This might be because the removed data has rather high `neck` value, so it increased the influence of `neck`, but in fact, neck is not that important. So our modification is to switch the position of `neck` and `age`. Now our final model looks like: `body.fat ~ log(weight) + log(height) + wrist + age + neck`.

The new diagnostic plots indicate that no outliers or influential points exist any more.

(d)

The ANOVA and `summary()` table for our final model.

```
# ---------------(d)---------------
anova(fat.lm4)
```

```
## Analysis of Variance Table
##
## Response: body.fat[-39]
##                    Df Sum Sq Mean Sq  F value    Pr(>F)
## log(weight)[-39]    1 5816.1  5816.1 286.4627 < 2.2e-16 ***
## log(height)[-39]    1 2579.9  2579.9 127.0670 < 2.2e-16 ***
## wrist[-39]          1  377.5   377.5  18.5950 2.344e-05 ***
## age[-39]            1 1026.4  1026.4  50.5552 1.264e-11 ***
## neck[-39]           1   83.0    83.0   4.0873   0.04429 *
## Residuals         245 4974.3    20.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(fat.lm4)
```

```
##
## Call:
## lm(formula = body.fat[-39] ~ log(weight)[-39] + log(height)[-39] +
##     wrist[-39] + age[-39] + neck[-39])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.7378  -3.0796  -0.4016   3.2972  11.3939
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      118.71664   36.13380   3.285  0.00117 **
## log(weight)[-39]  60.56758    3.74084  16.191  < 2e-16 ***
## log(height)[-39] -82.87330    9.63657  -8.600 9.64e-16 ***
## wrist[-39]        -2.80081    0.51383  -5.451 1.22e-07 ***
## age[-39]           0.18113    0.02515   7.201 7.33e-12 ***
## neck[-39]         -0.46871    0.23184  -2.022  0.04429 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.506 on 245 degrees of freedom
## Multiple R-squared:  0.6652, Adjusted R-squared:  0.6584
## F-statistic: 97.35 on 5 and 245 DF,  p-value: < 2.2e-16
```

Observe the estimate coefficients of variables, we conclude that `log(weight)` and `age` are postively related with `body.fat`, i.e. a person with larger weight and older age tends to have higher body fat percentage. On the other hand, `log(height), wrist, neck` are negatively related with `body.fat`. Among all, `log(height)` contributes the most, which can be interpreted as "taller people tend to have less body fat percentage when other variable hold constant".

The overall F-test has a p-value much less than 0.05, the model is significant.

Also note that the adjusted R^2 is 0.6584 which is adequate, but definitely leaves some space for improvement.

For multiple one-sample t-tests, the cut-off value is:

```
qt(.975, 250)
```

```
## [1] 1.969498
```

- $H_0 : \beta_0 = 0$ vs. $H_A : \beta_0 \neq 0$.
$$t = \frac{118.71664 - 0}{36.1338} = 3.285$$
- $H_0 : \beta_{\log(\text{weight})} = 0$ vs. $H_A : \beta_{\log(\text{weight})} \neq 0$,
$$t = \frac{60.56758 - 0}{3.74084} = 16.191$$
- $H_0 : \beta_{\log(\text{height})} = 0$ vs. $H_A : \beta_{\log(\text{height})} \neq 0$
$$t = \frac{-82.87330 - 0}{9.63657} = -8.600$$
- $H_0 : \beta_{\text{wrist}} = 0$ vs. $H_A : \beta_{\text{wrist}} \neq 0$
$$t = \frac{-2.80081 - 0}{0.51383} = -5.451$$
- $H_0 : \beta_{\text{age}} = 0$ vs. $H_A : \beta_{\text{age}} \neq 0$
$$t = \frac{0.18113 - 0}{0.02515} = 7.201$$

- $H_0 : \beta_{\text{neck}} = 0$ vs. $H_A : \beta_{\text{neck}} \neq 0$

$$t = \frac{-0.46871 - 0}{0.23184} = -2.022$$

All of those observed t-statistics have absolute values greater than the theoretical t-statistic 1.969498. Therefore, we reject all of those null hypotheses. The results of t-tests agree with that of F-test as we desired.

(e)

```
## The following objects are masked from fat:
##
##     abdomen, age, ankle, bicep, BMI, body.fat, body.fat.siri,
##     case, chest, density, ffweight, forearm, height, hip, knee,
##     neck, thigh, weight, wrist
```

```
##   case body.fat body.fat.siri density age weight height  BMI ffweight neck
## 1    1     12.6          12.3  1.0708  23 154.25  67.75 23.7    134.9 36.2
## 2    2      6.9           6.1  1.0853  22 173.25  72.25 23.4    161.3 38.5
## 3    3     24.6          25.3  1.0414  22 154.00  66.25 24.7    116.0 34.0
## 4    4     10.9          10.4  1.0751  26 184.75  72.25 24.9    164.7 37.4
## 5    5     27.8          28.7  1.0340  24 184.25  71.25 25.6    133.1 34.4
## 6    6     20.6          20.9  1.0502  24 210.25  74.75 26.5    167.0 39.0
##   chest abdomen   hip thigh knee ankle bicep forearm wrist
## 1  93.1    85.2  94.5  59.0 37.3  21.9  32.0    27.4  17.1
## 2  93.6    83.0  98.7  58.7 37.3  23.4  30.5    28.9  18.2
## 3  95.8    87.9  99.2  59.6 38.9  24.0  28.8    25.2  16.6
## 4 101.8    86.4 101.2  60.1 37.3  22.8  32.4    29.4  18.2
## 5  97.3   100.0 101.9  63.2 42.2  24.0  32.2    27.7  17.7
## 6 104.5    94.4 107.8  66.0 42.0  25.6  35.7    30.6  18.8
```

```
##         fit        lwr      upr
## 1  2.152769 -0.1418622  4.44740
## 2 14.626332 13.9233960 15.32927
## 3 22.076989 21.4265424 22.72744
## 4 30.670156 29.4412872 31.89902
```

- The interval for each category is narrow. But for underweight category, we have a negative value for the lower bound of 95% confidnece interval which is unrealistic. This is probably because the underweight category only contains one sample, whose `body.fat` value is 0. (We consider this as an input error, since no human has a zero body fat percent, this is just unrealistic.) So the prediction here is not accurate.
- Generally we should have an impression that, the more data we have in a category, the predicted confidence interval for this category is narrower, i.e. the prediction is more accurate.

To conclude, under such assumption, our model is good for predicting the later 3 categories but definitely not for underweight category.

Appendix

```
# ----------------Q1----------------
# ---------------(a)---------------
moorhen <- read.csv('moorhen.csv', header = TRUE)
attach(moorhen)
```

```r
pairs(moorhen)
cor(moorhen)
# ---------------(b)&(c)---------------
moorhen.lm <- lm(Shield ~ Weight + Stern + Hb + TandT + Adult)
moorhen.loglm <- lm(log(Shield) ~ Weight + Stern + Hb + TandT + Adult)
layout(matrix(c(1,2), 1, 2, byrow = TRUE))
plot(moorhen.lm, which=1, main='residual plot BEFORE log')
plot(moorhen.loglm, which=1, main='residual plot AFTER log')
# ---------------(d)---------------
anova(moorhen.loglm)
summary(moorhen.loglm)
anova(lm(log(Shield) ~ Stern + Adult + Weight + Hb + TandT))
(0.0402 + 0.0001 + 0.0129)/3/0.0573
(0.0001 + 0.0129)/2/0.0573
(0.0129)/0.0573
qf(0.95, 3, 37)
# ---------------(e)---------------
moorhen.loglm2 <- lm(log(Shield) ~ Adult + Stern)
par(mfrow=c(2,2))
plot(moorhen.loglm2, which=1)
plot(fitted(moorhen.loglm2), rstandard(moorhen.loglm2))
abline(0,0, lty=2)
title("IntStud Residuals vs. Fitted Values")
plot(moorhen.loglm2, which=c(2, 4))
# ---------------(f)---------------
par(mfrow=c(1,1))
plot(Shield ~ Stern,
     xlab = "Stern (unit unknown)",
     ylab = "Shield Area (mm^2)",
     #ylim = c(0, 600),
     pch = c(16, 17)[as.numeric(Adult)+1],
     main = "Prediction Intervals of Adult and Juvenile Moorhens",
     col = c("red", "blue")[as.numeric(Adult)+1],
     data = moorhen,
     cex = 1.5)
adults.sterns <- moorhen[moorhen$Adult==1,]$Stern
adults.sterns2 <- (min(adults.sterns) * 10):(round(max(adults.sterns)*10,0)+1)/10
adults.input <- data.frame(Shield=mean(Shield), Weight=mean(Weight),
                           Stern=adults.sterns2,
                           Hb=mean(Hb), TandT=mean(TandT), Adult=1)

juves.sterns <- moorhen[moorhen$Adult!=1,]$Stern
juves.sterns2 <- (round(min(juves.sterns) * 10, 0)-1):(round(max(juves.sterns) * 10, 0)+1)/10
juves.input <- data.frame(Shield=mean(Shield), Weight=mean(Weight),
                          Stern=juves.sterns2,
                          Hb=mean(Hb), TandT=mean(TandT), Adult=0)

adults.pred <- exp(predict(moorhen.loglm2, newdata = adults.input,
                           interval = "prediction"))
juves.pred <- exp(predict(moorhen.loglm2, newdata = juves.input,
                          interval = "prediction"))

lines(adults.sterns2, adults.pred[,"fit"], lty=6, lwd=2, col="blue")
```

```r
lines(adults.sterns2, adults.pred[,"lwr"], lty=5, col="blue")
lines(adults.sterns2, adults.pred[,"upr"], lty=5, col="blue")
lines(juves.sterns2, juves.pred[,"fit"], lty=4, lwd=2, col="red")
lines(juves.sterns2, juves.pred[,"lwr"], lty=3, lwd=2, col="red")
lines(juves.sterns2, juves.pred[,"upr"], lty=3, lwd=2, col="red")
legend( x="topleft",
        legend=c("Adult", "Adult model", "Adult 95% PI",
                 "Juvenile", "Juvenile model", "Juvenile 95% PI"),
        col=c("blue", "blue", "blue", "red", "red", "red"),
        lwd=c(1, 2, 1, 1, 2, 2), lty=c(NA, 6, 5, NA, 4, 3),
        pch=c(17, NA, NA, 16, NA, NA), cex = 0.75)
# ---------------(g)---------------
coef(moorhen.loglm2)
shield.calc <- function(x1, x2){
    exp(as.numeric(moorhen.loglm2$coefficients[1]
                   + moorhen.loglm2$coefficients[2] * x1
                   + moorhen.loglm2$coefficients[3] * x2))
}
ratio.calc <- function(st.input){
    shield.calc(1, st.input)/shield.calc(0, st.input)
}

confint(moorhen.loglm2, "Adult", level = 0.95)
exp(confint(moorhen.loglm2, "Adult", level = 0.95))
# ---------------Q2---------------
# ---------------(a)---------------
library(faraway)

fat <- read.csv('fat.csv', header = TRUE)
fat[42,]$height <- 69.5
attach(fat)
# ---------------(b)---------------
# round(cor(fat[c(2,5:7,10:19)]),4)
# round(cor(fat[c(2,5:7,10:19)]),4)>0.7

vif(lm(body.fat~log(weight)+log(height)+abdomen))
vif(lm(body.fat~log(weight)+log(height)+neck))
vif(lm(body.fat~log(weight)+neck+wrist+log(height)))
vif(lm(body.fat~log(weight)+neck+wrist+hip+log(height)))
vif(lm(body.fat~log(weight)+neck+wrist+chest+log(height)))
vif(lm(body.fat~log(weight)+neck+wrist+log(height)+age))
vif(lm(body.fat~log(weight)+neck+wrist+log(height)+age+forearm))
vif(lm(body.fat~log(weight)+neck+wrist+log(height)+age+forearm+thigh))
vif(lm(body.fat~log(weight)+neck+wrist+log(height)+age+forearm+bicep))
vif(lm(body.fat~log(weight)+neck+wrist+log(height)+age+forearm+bicep+ankle))
vif(lm(body.fat~log(weight)+neck+wrist+log(height)+age+forearm+bicep+ankle+knee))

vif(lm(body.fat~log(weight)+wrist+log(height)+age+forearm+bicep+ankle+knee))

vif(lm(body.fat~log(weight)+wrist+log(height)+age+forearm+bicep+ankle))

vif(lm(body.fat~log(weight)+log(height)+age+forearm+bicep+ankle))
```

```r
vif(lm(body.fat~log(weight)+log(height)+age+forearm+ankle))

# acutally the following model works well too!
vif(lm(body.fat~log(weight)+log(height)+age))
anova(lm(body.fat~log(weight)+log(height)+age))
summary(lm(body.fat~log(weight)+log(height)+age))

# draft model based on correlation matrix intuition and VIF analysis
fat.lm1 <- lm(body.fat ~ log(weight) + log(height) + wrist + neck + age
              + forearm + bicep + ankle + knee)
anova(fat.lm1)
summary(fat.lm1)
# draft model deleted insignificant variables
fat.lm2 <- lm(body.fat ~ log(weight) + log(height) + wrist + neck + age)
anova(fat.lm2)
summary(fat.lm2)
par(mfrow=c(2,2))
plot(fat.lm2, which=1)
plot(fitted(fat.lm2), rstandard(fat.lm2))
abline(0,0, lty=2)
title("IntStud Residuals vs. Fitted Values")
# identify(fitted(fat.lm), rstandard(fat.lm)) # 39, 216
plot(fat.lm2, which=c(2, 4))
par(mfrow=c(1,1))
# ---------------(c)---------------
# case 39 is the suspisious outlier
# also 182 has is 0 as body.fat (but we still need it as 'underweight' category later)

# draft model with case 39 removed (revisiting)
fat.lm3 <- lm(body.fat[-39] ~ log(weight)[-39] + log(height)[-39]
              + wrist[-39] + neck[-39] + age[-39])
anova(fat.lm3)
summary(fat.lm3)

# final model with some order tweaked (neck moved to the last)
fat.lm4 <- lm(body.fat[-39] ~ log(weight)[-39] + log(height)[-39]
              + wrist[-39] + age[-39] + neck[-39])

par(mfrow=c(2,2))
plot(fat.lm4, which=1)
plot(fitted(fat.lm4), rstandard(fat.lm4))
abline(0,0, lty=2)
title("IntStud Residuals vs. Fitted Values")
plot(fat.lm4, which=c(2, 4))
par(mfrow=c(1,1))
# ---------------(d)---------------
anova(fat.lm4)
summary(fat.lm4)
qt(.975, 250)
# ---------------(e)---------------
fat2 <- fat[-39,]
attach(fat2)
head(fat2)
```

```r
underweight <- fat2[which(BMI<18.5),]
normal <- fat2[which(BMI>=18.5 & BMI<25), ]
overweight <- fat2[which(BMI>=25 & BMI<30),]
obese <- fat2[which(BMI>=30),]

UwM <- sapply(underweight[,c(2,5,6,7,10,19)], mean)
NM <- sapply(normal[,c(2,5,6,7,10,19)], mean)
OwM <- sapply(overweight[,c(2,5,6,7,10,19)], mean)
ObM <- sapply(obese[,c(2,5,6,7,10,19)], mean)

df <- data.frame(weight=c(UwM[3],NM[3],OwM[3],ObM[3]),
                 height=c(UwM[4],NM[4],OwM[4],ObM[4]),
                 wrist=c(UwM[6],NM[6],OwM[6],ObM[6]),
                 age=c(UwM[2],NM[2],OwM[2],ObM[2]),
                 neck=c(NM[5],NM[5],OwM[5],ObM[5]))

(ci <- predict(fat.lm4, newdata = df, interval = "confidence"))
```