For HW2 Q1 and Q2, please refer to the other version of solution by Prof. Carol Bigelow. Here we will mainly use R to solve HW2-Q3.

Q3: Let’s download the dataset.

```r
library(foreign)
url <- "http://people.umass.edu/biep640w/datasets/week02.dta"
dat <- read.dta(file = url) # read.dta: Read Stata Binary Files
# dat
head(dat)
##    temp boiling
## 1 210.8 29.211
## 2 210.1 28.559
## 3 208.4 27.972
## 4 202.5 24.697
## 5 200.6 23.726
## 6 200.1 23.369
```

```r
x <- dat$temp
y1 <- dat$boiling
m1 <- lm(y1 ~ x)
# OR: m1 <- lm(boiling~temp, data=dat)
# OR: m1 <- lm(dat$boiling ~ dat$temp)
summary(m1)
```

```r
## Call:
## lm(formula = y1 ~ x)
## ## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8026  0.0057  0.1418  0.5573  1.0022
## ## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -65.34300   4.64381  -14.07  1.73e-14 ***
## x             0.44379   0.02419   18.35  < 2e-16 ***
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```
m1$coefficients  # output the intercept and slope

## (Intercept)    x
## -65.3429977    0.4437942

# confint(m1)  # output the CI for the parameters in the fitted model

anova(m1)

## Analysis of Variance Table
##
## Response: y1
## Df Sum Sq Mean Sq F value Pr(>F)
## x 1 450.56 450.56 336.6 <2.2e-16 ***
## Residuals 29 38.82 1.34
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

plot(x, y1, main="Simple Linear Regression of Y=boiling on X=temp", xlab="temp", ylab="boiling", pch=18)  # pch means "plot character", can be 1 to 18;
abline(m1, col="red", lty=1, lwd=2)  # lty: "line type"; lwd: "line width".

From the above summary(m1) output, we know that:
(a): Parameter Estimates: -65.34 and 0.44, so regression line: Y = -65.34 + 0.44X;
(b): ANOVA (Analysis of Variance) table is obtained above using the anova(m1) code;
(c): R-square=0.9207;
(d): The scatterplot with the fitted line is plotted above.

(2). Now, instead of Y = boiling, we want to use newy = 100*log10(boiling)

```r
newy = 100*log10(y1)
m2 <- lm(newy ~ x)
summary(m2)
```

```
## Call:
## lm(formula = newy ~ x)
## ## Residuals:
## Min 1Q Median 3Q Max
## -14.2548 0.3286 0.6079 0.9923 1.4159
## ## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) -48.85829 11.88807 -4.11 0.000297 ***
## x 0.92615 0.06192 14.96 3.62e-15 ***
## ---
## Signif. codes: 0 '\*'***' 0.001 '***' 0.01 '**' 0.05 '*' 0.1 '.' 1
## ## Residual standard error: 2.962 on 29 degrees of freedom
## Multiple R-squared: 0.8852, Adjusted R-squared: 0.8813
## F-statistic: 223.7 on 1 and 29 DF, p-value: 3.623e-15

m2$coefficients  # output the intercept and slope

```
```
## (Intercept)   x
## -48.8582854  0.9261547
```

```
# confint(m2)  # output the CI for the parameters in the fitted model
anova(m2)
```

```
## Analysis of Variance Table
## ## Response: newy
## Df Sum Sq Mean Sq F value Pr(>|F|)
## x 1 1962.2 1962.25  223.69 3.62e-15 ***
## Residuals 29 254.4  8.77
## ---
## Signif. codes: 0 '\*'***' 0.001 '***' 0.01 '**' 0.05 '*' 0.1 '.' 1
```
From the above summary(m2) output, we know that:

(a): Parameter Estimates: -48.86 and 0.93, so regression line: $100\log_{10}(Y) = -48.86 + 0.93X$;
(b): ANOVA (Analysis of Variance) table is obtained above using the anova(m2) code;
(c): R-square=0.8852;
(d): The scatterplot with the fitted line is plotted above.

Solution (one paragraph of text that is interpretation of analysis):

Did you notice that the scatter plot of these data reveal two outlying values? Their inclusion may or may not be appropriate.

If all $n=31$ data points are included in the analysis, then the model that explains more of the variability in boiling point is $Y=$boiling point modeled linearly in $X=$temp. It has a greater $R^2$ (92.07% vs. 88.52%).

Be careful - It would not make sense to compare the residual mean squares of the two models because the scales of measurement involved are different.