550.413 Assignment 2

Fall 2016

Instruction: This assignment consists of 4 problems. The assignment is due on **Friday, October 14, 2016**, in section. If you cannot make it to section, please leave the assignment under the door at Whitehead 306E and email the course instructor. If possible, please type up your assignments, preferably using LaTeX. For problems with R programming, please also attach a print-out of the **R** code. When asked to perform hypothesis testing, you are free to use any (reasonable) choice of significance level α.

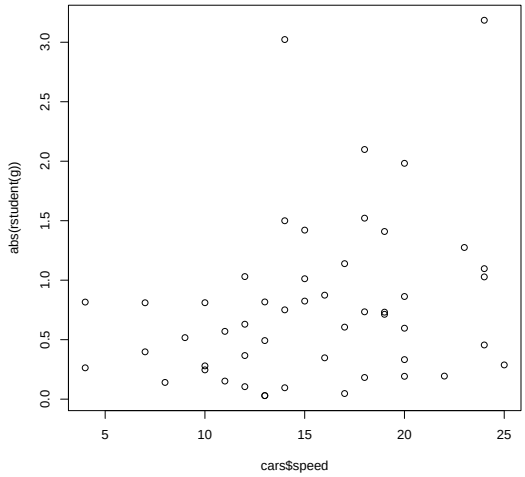
Problem 1: (20pts)

This problem uses the dataset `cars` from the `faraway` library. The `cars` data records the speed of cars and the distances taken to stop. Note that the data were recorded in the 1920s. Using this dataset, first perform a simple linear regression with `dist` as the response variable and `speed` as the predictor variable. Then using this model, answer the following questions

1. Check the constant variance assumption for the errors.
2. Check the normality assumption for the errors.
3. Check for outliers and nominate potential outliers, if they exist.
4. Without using the `anova` command, check the goodness-of-fit of this simple linear regression model.

Solution:

```
library("faraway")
data(cars)
g <- lm(dist ~ speed, cars)
## Plot the externally studentized residuals against the fitted values
plot(cars$speed, abs(rstudent(g)))
```



For part (a), we first fit a model $\texttt{dist} = \beta_0 + \beta_1 \texttt{dist} + \epsilon$. We observe that the plot of the absolute value of the externally studentized residuals `rstudent(g)` against the predictor variable exhibit signs of non-constant variance. It is important to note that we only claim non-constant variance under the assumption that the functional relationship between `dist` and `speed` is indeed of the form $\mathbb{E}[\texttt{dist} \mid \texttt{speed}] = \beta_0 + \beta_1 \texttt{speed}$.

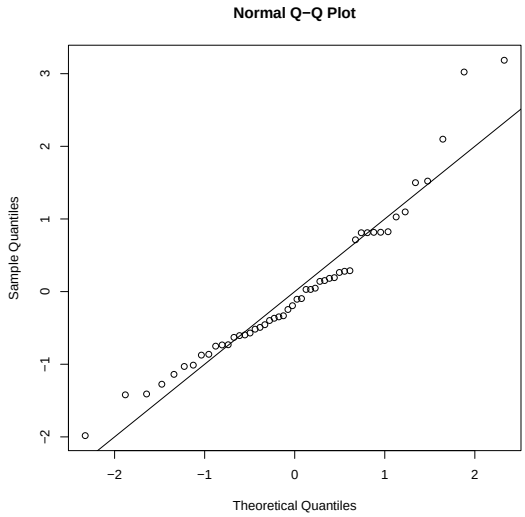
```
## Perform a formal test of non-constant variance using the Brown Forsythe test.
library("lawstat")
levene.test(residuals(g), group = (cars$speed >= 14), location = "median")

##
##  modified robust Brown-Forsythe Levene-type test based on the absolute
##  deviations from the median
##
## data:  residuals(g)
## Test Statistic = 5.9216, p-value = 0.01873

## ``Equivalently''
## note however that the following is a t-test while the
## levene.test is a F-test
I1 <- which(cars$speed < 14)
I2 <- which(cars$speed >= 14)
n1 <- length(I1)
n2 <- length(I2)
n <- n1 + n2
tilde.e1 <- median(residuals(g)[I1])
tilde.e2 <- median(residuals(g)[I2])
bar.d1 <- mean(abs(residuals(g)[I1] - tilde.e1))
```

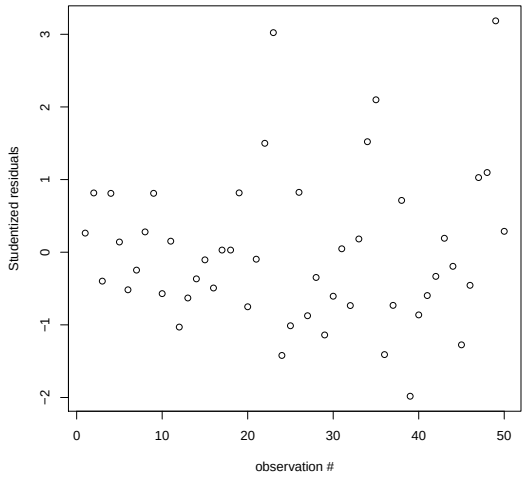
```
bar.d2 <- mean(abs(residuals(g)[I2] - tilde.e2))
s1 <- sd(abs(residuals(g)[I1] - tilde.e1))
s2 <- sd(abs(residuals(g)[I2] - tilde.e2))
s <- sqrt((s1^2*n1 + s2^2*n2)/(n - 2))
t.BF <- (bar.d1 - bar.d2)/(s*sqrt(1/n1 + 1/n2))
cat(paste("|t.BF| = ", abs(t.BF)))

## |t.BF| =  2.39190039030306

cat(paste("p-value is", 2*pt(abs(t.BF), n - 2, lower.tail = FALSE)))

## p-value is 0.0207266510243169
```

A formal Brown-Forsythe test for non-constant variance, where we split the observations into two groups at `speed` 14 miles per hour or faster vs `speed` 13 miles per hour or slower, corroborate this observation. While the Brown-Forsyth test is a "formal" test, we caution that it has limitations. In particular, changing the grouping slightly could lead to conclusions that appear to be quite different. For example, compare the following output

```
levene.test(residuals(g), group = (cars$speed >= 14), location = "median")

##
##  modified robust Brown-Forsythe Levene-type test based on the absolute
##  deviations from the median
##
## data:  residuals(g)
## Test Statistic = 5.9216, p-value = 0.01873

levene.test(residuals(g), group = (cars$speed >= 15), location = "median")

##
##  modified robust Brown-Forsythe Levene-type test based on the absolute
##  deviations from the median
##
## data:  residuals(g)
## Test Statistic = 1.9137, p-value = 0.173
```

```
qqnorm(rstudent(g))
abline(0, 1)
plot(1:nrow(cars), rstudent(g), xlab = "observation #", ylab = "Studentized residuals")
```





For part (b) and (c), a QQ-plot of the studentized residuals against the normal distribution suggests that the residuals are not normally distributed. In particular, the residuals appear to have a slightly heavier right tail compared to the normal distribution. Looking at the values of the residuals, the two largest values, in magnitude, are 3.185 and 3.023, corresponding to observations # 49 and # 23, respectively. "Testing" for outliers using Bonferroni correction at the conventional significance level $\alpha = 0.05$, the (adjusted) critical value is $\texttt{qt}(1 - \frac{0.05}{200}, 48) = 3.734$. We thus conclude that there are no "obvious" outliers.

```
uniq.speed <- unique(cars$speed)
dist.mean <- numeric(length(uniq.speed))
```

```r
for (i in 1:length(uniq.speed)) {
    dist.mean[i] <- mean(cars$dist[cars$speed == uniq.speed[i]])
}
SSPE.partial <- numeric(length(uniq.speed))
for (i in 1:length(uniq.speed)) {
    SSPE.partial[i] <- sum((cars$dist[cars$speed == uniq.speed[i]] - dist.mean[i])^2)
}
df <- data.frame(table(cars$speed), dist.mean, SSPE.partial)
colnames(df) <- c("speed", "Freq", "dist.mean", "SSPE.partial")
df
```

```
##    speed Freq dist.mean SSPE.partial
## 1      4    2   6.00000      32.0000
## 2      7    2  13.00000     162.0000
## 3      8    1  16.00000       0.0000
## 4      9    1  10.00000       0.0000
## 5     10    3  26.00000     128.0000
## 6     11    2  22.50000      60.5000
## 7     12    4  21.50000     107.0000
## 8     13    4  35.00000     204.0000
## 9     14    4  50.50000    1771.0000
## 10    15    3  33.33333     658.6667
## 11    16    2  36.00000      32.0000
## 12    17    3  40.66667     162.6667
## 13    18    4  64.50000    1091.0000
## 14    19    3  50.00000     536.0000
## 15    20    5  50.40000     563.2000
## 16    22    1  66.00000       0.0000
## 17    23    1  54.00000       0.0000
## 18    24    4  93.75000    1256.7500
## 19    25    1  85.00000       0.0000
```

```r
(SSPE <- sum(SSPE.partial))
```

```
## [1] 6764.783
```

```r
(SSE <- sum(residuals(g)^2))
```

```
## [1] 11353.52
```

```r
n <- length(cars$speed)
K <- length(uniq.speed)
(df.SSPE <- n - K)
```

```
## [1] 31
```

```r
(df.SSE <- n - 2)
```

```
## [1] 48
```

```r
(F.stat <- (SSE - SSPE)/SSPE * df.SSPE/(df.SSE - df.SSPE))
```

```
## [1] 1.23695
```

Country	Y	X
Ireland	58	220
Sweden	115	310
Denmark	165	380
United States	190	1280
Switzerland	250	530
Great Britain	465	1145
Norway	90	250
Canada	150	510
Australia	170	455
Holland	245	460
Finland	350	1115

Table 1: Male death per millions (Y) vs per capita cigarette consumption (X) in 1930 for several selected countries.

Finally, for part (d), the goodness-of-fit test can be carried out as follows. The test statistic is $\frac{\texttt{SSE}-\texttt{SSPE}}{\texttt{SSPE}} \times \frac{48}{48-31} = 1.237$ where the degree of freedom for the regression is 48 and the degree of freedom for the pure error is 31. We thus conclude that there is little evidence of a lack of fit of the simple linear regression model, as compared to the saturated model (the model with a predictor variable for each distinct value of the `speed` variable).

Problem 2: (10pts)

Table 1 displays the male death (per millions) due to lung cancer vs per capita cigarette consumption for several countries.

(a) Using the above data, fit a simple linear regression. Do any of the points look particularly influential ? Delete the United States, refit a simple linear regression and note how the regression line changes. Now, put the U.S.A back and delete Great Britain and note how the regression line changes.[1].

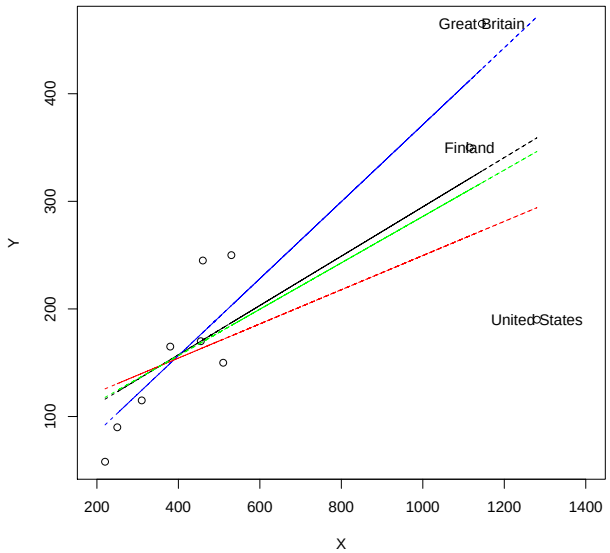
(b) What is a plausible explanation for the outliers or influential points from (a) ?

Solution:

```
library("SenSrivastava")
data(E8.12)
Y <- E8.12$y
X <- E8.12$x
```

[1] If you have better things to do than to manually enter the above data into your **R** session, you can load the dataset as dataset `E8.12` in the package `SenSrivastava`

```r
countries <- E8.12$Country

g <- lm(Y ~ X)
plot(X, Y, xlim = c(200, 1400))
## We can use identify(X, Y, countries) to visually identify outliers or influential points

Yhat <- g$coef[1] + g$coef[2] * X
lines(X, Yhat, lty = 2)
potential.outliers <- countries %in% c("United States", "Great Britain", "Finland")
text(X[potential.outliers], Y[potential.outliers], countries[potential.outliers])

g.noUSA <- lm(y ~ x, data = subset(E8.12, Country != "United States"))
Yhat.noUSA <- g.noUSA$coef[1] + g.noUSA$coef[2] * X
lines(X, Yhat.noUSA, col = "blue", lty = 2)

g.noGB <- lm(y ~ x, data = subset(E8.12, Country != "Great Britain"))
Yhat.noGB <- g.noGB$coef[1] + g.noGB$coef[2] * X
lines(X, Yhat.noGB, col = "red", lty = 2)

g.noFinland <- lm(y ~ x, data = subset(E8.12, Country != "Finland"))
Yhat.noFinland <- g.noFinland$coef[1] + g.noFinland$coef[2] * X
lines(X, Yhat.noFinland, col = "green", lty = 2)
```



We perform a simple linear regression of male death against per capital cigarette consumption. Looking at the plot of male death against per capita

cigarette consumption, we note that three of the data points appear to have high leverage, i.e., they correspond to countries whose per capita cigarette consumption are higher than "average". We can identify these countries are being Finland, Great Britain, and the United States. We then try to fit a regression with those countries removed. In particular, removing the data point corresponding to the United States and rerunning the regression gave us the blue least square regression line. Similarly, removing the data point corresponding to Great Britain gave us the red regression line and removing the data point corresponding to Finland gave us the green regression line. From these regression lines, we conclude that Great Britain and the United States are influential (or high-leverage) points as their removal changes the regression line significantly.

For part (b), there are numerous plausible explanations. One plausible explanation that is interesting (at least in the scope of this class due to its confounding nature) is that there is potentially a sizable number of female smokers in the United States as compared to Great Britain. The response variable Y is the number of *male* deaths per million while the predictor variable is the *per capita* (which includes both male and female) consumption. That is to say, it might be the case that Great Britain has a lot of smokers (with most of them being male) while the USA has a lot of smokers (with a sizable number of them being female). The data for the regression only includes the male smokers in the response variable while the predictor variable includes both male and female, and thus there is a confounding effect. Another plausible explanation is that smokers are in general older in Great Britain. For more on this dataset, see pages 78 through 88 of the book "Data Analysis for Politics and Policy" by Edward R. Tufte.

Problem 3: (10pts)

Install the package `aprean3` to get access to data from the book "Applied Regression Analysis" by Draper and Smith. The following problem uses the dataset `dse13e` which records the growth rate and densities of ice crystals. Ice crystals are introduced into a chamber where the interior is maintained at a fixed temperature (-5 degree centigrade) and humidity. The growth of the crystals with time is observed. The variables here are `t` and `m` denoting the times in seconds from the introduction of the crystals and the mass of the crystals in nanograms, respectively.

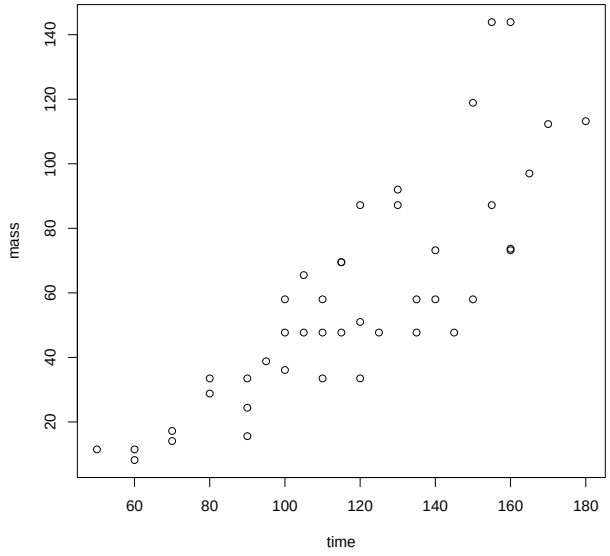
Using the above dataset, answer the following question

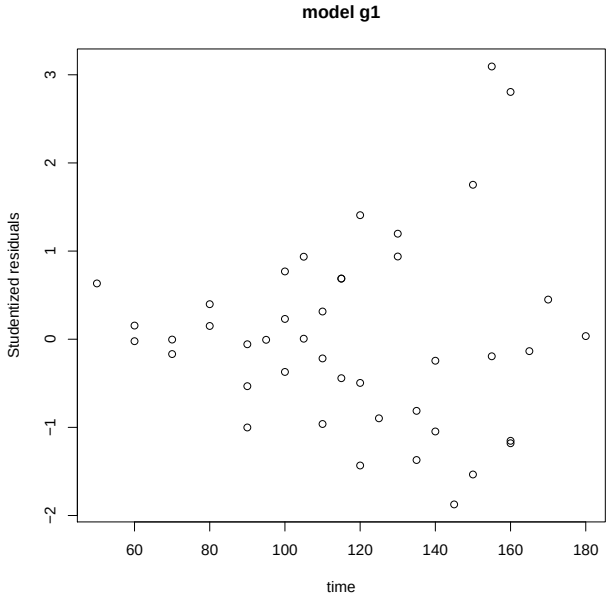
1. Perform a simple linear regression with `m` as the response variable and `t` as a predictor variable. Discuss whether or not this simple linear regression model is appropriate.

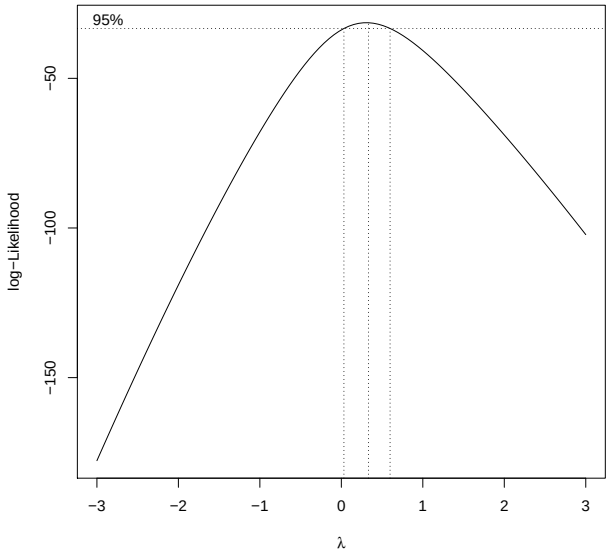
2. Try to see if you can "improve" the simple linear regression model in part (1) by transforming either (or both) the predictor or the response variable.

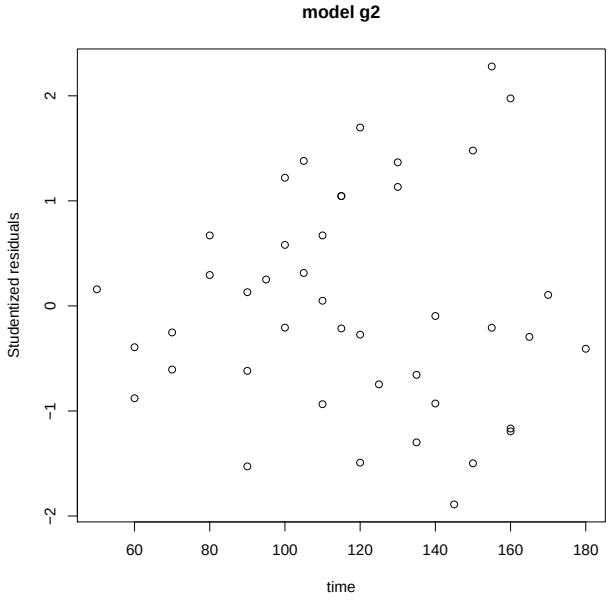
Solution:

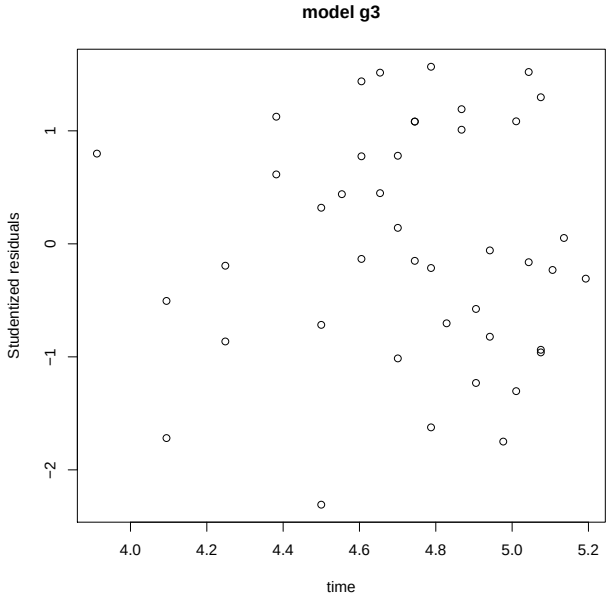
```r
library("aprean3")
library("MASS")
data(dse13e)
plot(dse13e$t, dse13e$m, xlab = "time", ylab = "mass")
g1 <- lm(m ~ t, dse13e)
plot(dse13e$t, rstudent(g1), xlab = "time", ylab = "Studentized residuals", main = "model g1")
boxcox(g1, plotit = TRUE, lambda = seq(from = -3, to = 3, by = 0.1))
g2 <- lm(sqrt(m) ~ t, dse13e)
plot(dse13e$t, rstudent(g2), xlab = "time", ylab = "Studentized residuals", main = "model g2")
g3 <- lm(log(m) ~ log(t), dse13e)
plot(log(dse13e$t), rstudent(g3), xlab = "time", ylab = "Studentized residuals", main = "model g3")
```




model g1
Studentized residuals
time
−2
−1
0
1
2
3
60
80
100
120
140
160
180


95%
log−Likelihood
λ
−150
−100
−50
−3
−2
−1
0
1
2
3





For part (a), we first run a simple linear regression $\mathtt{m} = \beta_0 + \beta_1 \mathtt{t} + \epsilon$. The plot of the predictor variable `t` against the studentized residuals suggests that the posited linear relationship between mass and crystallization time is not appropriate. The Box-Cox likelihood plot and the scatter plot of the variables `m` and `t` suggests a logarithmic or square root transformation for the response variable

$\mathtt{m}$. For part (b), we try fitting the regression model $\sqrt{\mathtt{m}} = \beta_0 + \beta_1 \mathtt{t} + \epsilon$ to the data. The plot of the studentized residuals against the predictor variable, while still exhibiting some sign of non-constant variance, do suggests that our model in part (b) is preferable to our model in part (a). Note that we could also first transform the predictor variable, e.g., by taking $\log \mathtt{t}$ as the predictor variable and then do a Box-Cox transformation to identify suitable transformation of the response variable. This will then yield $\log \mathtt{m} = \beta_0 + \beta_1 + \log \mathtt{t} + \epsilon$ as a suitable candidate model. The plot of the studentized residuals against the predictor variable for this model is similar to that for our model in part (b).

Problem 4: (20pts)

Install the package `aprean3` to get access to data from the book "Applied Regression Analysis" by Draper and Smith. The following problem uses the dataset `dse07c` which records the number of motor vehicle deaths and the number of drivers for 50 states in the US in 1966. We augment the data for the 50 state with the statistics for Washington DC and clean up the variable names as follows.

```
library("aprean3")
data(dse07c)
data(state)
y <- c(dse07c$y, 115)
x1 <- c(dse07c$x1, 35)
x2 <- c(dse07c$x2, 12524)
x3 <- c(dse07c$x3, NA)
x4 <- c(dse07c$x4, "No")
x5 <- c(dse07c$x5, 44)
x6 <- c(dse07c$x6, 23)
df <- data.frame(deaths = y, drivers = x1, density = x2, rural.mileage = x3, more.male = x4,
    january.temp = x5, fuel.consumption = x6, state.name = c(state.name, "DC"))
```

Here the variable `drivers` is in units of 10^4, the variable `density` is number of persons per square mile, `rural.mileage` is the total length of rural road in the state (the unit is in thousand of miles), `january.temp` is the average high temperature in January in the state, and `fuel.consumption` is the total amount of gallons consumed per year (the unit is in 10^7 gallons).

Given this augmented data, answer the following question.

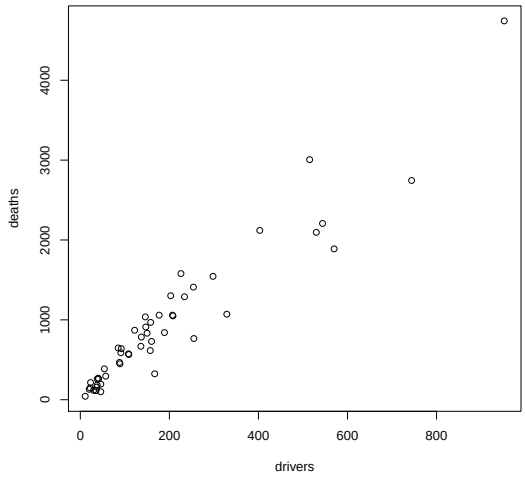
1. Plot the number of deaths against the number of drivers. Argue why a simple linear regression with `log(deaths)` as the response variable against `log(drivers)` as the predictor variable is to be preferred over a simple linear regression with `deaths` as the response variable and `drivers` as the predictor variable.

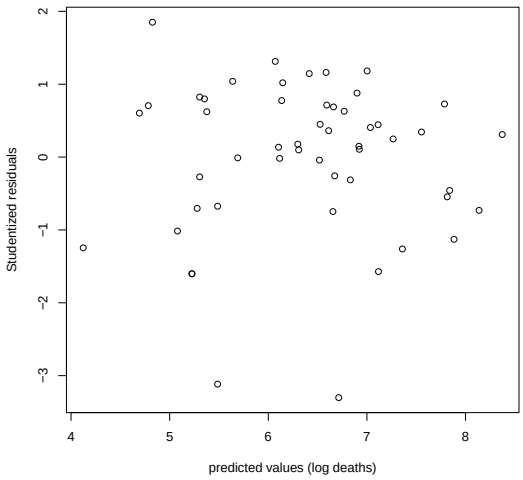
2. Perform a simple linear regression with `log(deaths)` as the response variable against `log(drivers)` as the predictor variable. Plot the residuals

against the fitted value for this regression model. Which of the residuals are potential outliers ?

3. It is claimed that by adding another predictor variable, the new model will no longer exhibit any potential outliers, thereby yielding a model with better fit. Looking at the list of remaining variables, which would be the most logical candidate to be added to the current model ? You are not required to fit a new regression model.

```
library("aprean3")
plot(deaths ~ drivers, df)
g <- lm(log(deaths) ~ log(drivers), df)
plot(fitted(g), rstudent(g), xlab = "predicted values (log deaths)", ylab = "Studentized residuals")
g2 <- lm(log(deaths) ~ log(drivers) + density, df)
anova(g, g2)

## Analysis of Variance Table
##
## Model 1: log(deaths) ~ log(drivers)
## Model 2: log(deaths) ~ log(drivers) + density
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     49 4.7831
## 2     48 4.0939  1   0.68923 8.0812 0.00655 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```





For part (a), we note that the plot of `deaths` against `drivers` indicates large ranges of values for both variables. In particular, the majority of the points are located "near" the origin; the remaining points will therefore be potentially of high leverage. A logarithmic transformation to spread the data points more evenly is thus desirable. For part (b), after fitting the regression model $\texttt{log(deaths)} = \beta_0 + \beta_1 \texttt{log(drivers)} + \epsilon$, we look at the plot of residuals against the fitted values. Two data points appear to be potential outliers, namely observation # 7 and observation # 39. These correspond to the state of Connecticut and Rhode Island. For both of these states, the studentized residual values are negative, suggesting that the number of deaths for these states are less than expected. While an outlier "test" using Bonferroni correction does not indicate any outlier at a significance level of $\alpha = 0.05$, nevertheless, we still want to understand why these two states appear to have less number of deaths per drivers compared to "average". The variable `density` indicates that Connecticut and Rhode Island are two of the most densely populated states. A claim can be made that accidents are less likely to be fatal at lower speed, and more densely populated states are correlated with slower driving conditions. Therefore, the variable `density` is a logical candidate to be added to the model in part (a). Indeed, adding the variable `density` to the model in part (a) and comparing the least square fit between the two models (using the `anova`) command, suggests that the model incorporating the `density` variable is preferable over our original model in part (a). We emphasize that the above discussion are meant only to illustrate the kind of reasoning that maybe useful in understanding our data and the regression output. In general, we can try many things; only a few, if any, of the things we try will yield conclusive evidence.