STAT 757 Assignment 4

DUE X/XX/2018 11:59PM

AG Schissler

2/14/2018

Instructions [20 points]

Modify this file to provide responses to the Ch.4 Exercises in Sheather (2009). You can find some helpful code here: http://www.stat. tamu.edu/~sheather/book/docs/rcode/Chapter4.R. Also address the project milestones indicated below. Please email **both** your .Rmd (or roxygen .R) and one of the following either .HTML, .PDF, or .DOCX using the format SURNAME-FIRSTNAME-Assignment4.Rmd and SURNAME-FIRSTNAME-Assignment4.pdf.

```
data_dir <- "/Users/alfred/OneDrive - University of Nevada, Reno/Teaching/STAT_757/Sheather_data/Data"
```

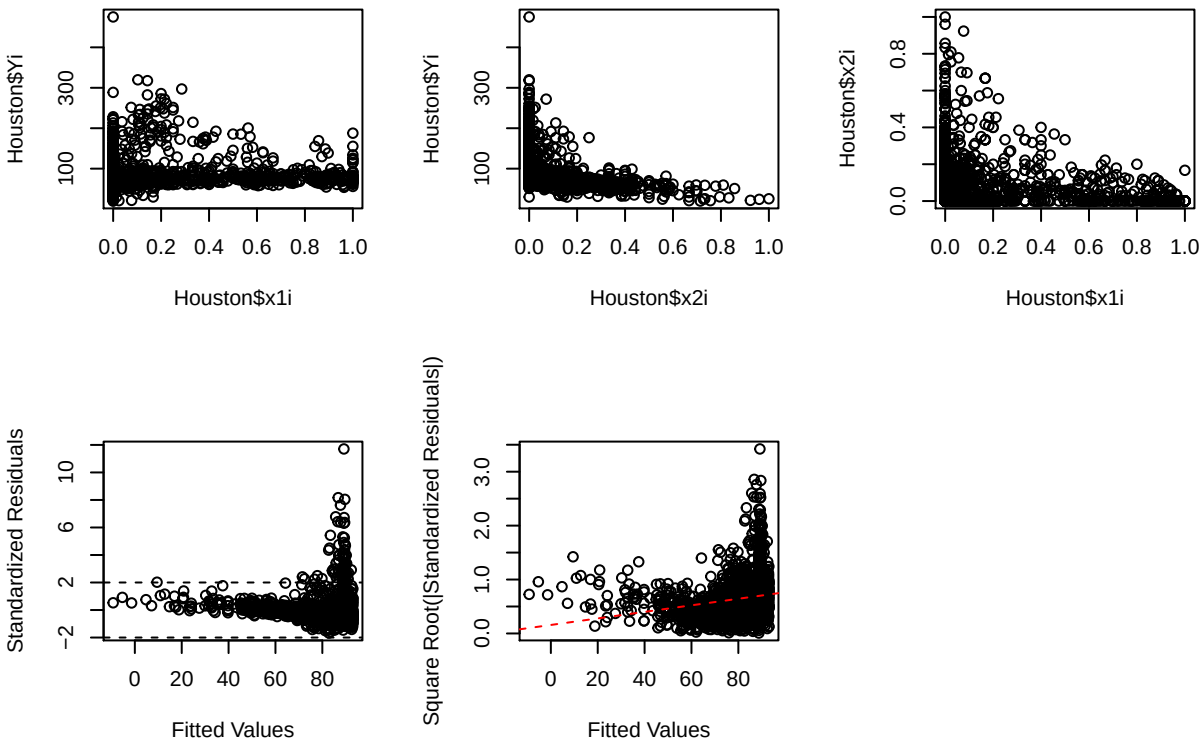
Exercise 4.2.3 [60 points total]

```
Houston <- read.table(file.path(data_dir, "HoustonRealEstate.txt"),header=TRUE)
str(Houston)
```

```
## 'data.frame':    1922 obs. of  4 variables:
##  $ Yi : num  169.2 56.8 25.5 90.7 92.7 ...
##  $ ni : int  7 6 6 5 8 9 5 8 5 7 ...
##  $ x1i: num  0.857 0.167 0 0.8 0.5 0.667 0 0 0 0.571 ...
##  $ x2i: num  0 0.667 1 0.2 0 0 0.2 0.25 0 0.143 ...
```

```
head(Houston)
```

```
##       Yi ni   x1i   x2i
## 1 169.20  7 0.857 0.000
## 2  56.82  6 0.167 0.667
## 3  25.52  6 0.000 1.000
## 4  90.67  5 0.800 0.200
## 5  92.65  8 0.500 0.000
## 6  87.76  9 0.667 0.000
```



Part a [20 points]

Since each observation represents a summary (or aggregation) for a subdivision of homes, each data point should conttribute to the model according to how much information is contained per row. In cases where there are many homes summarizes the weights should be larger and vice versa for subdivisions with few homes. As remarked in Sheather (2009) Section 4.1.5, when the observation is a mean or median of the response values then $w_i = n_i$ is a reasonable choice. There is a well-known result that, for large n, the variance of the estimate of the median from a sample of size n from a distribution with density function $f(x)$ is $\frac{1}{4n[f(m)]^2}$ where m is median, that is the variance is proportional to 1 over the sample size. Since the data for this question are estimated medians calculated from samples of different sizes, it is appropriate to use weighted least squares with the sample sizes as the weights.

Part b [20 points]

Clearly the resdiuals display inconstant variance and non-random appearance, invalidating the model.

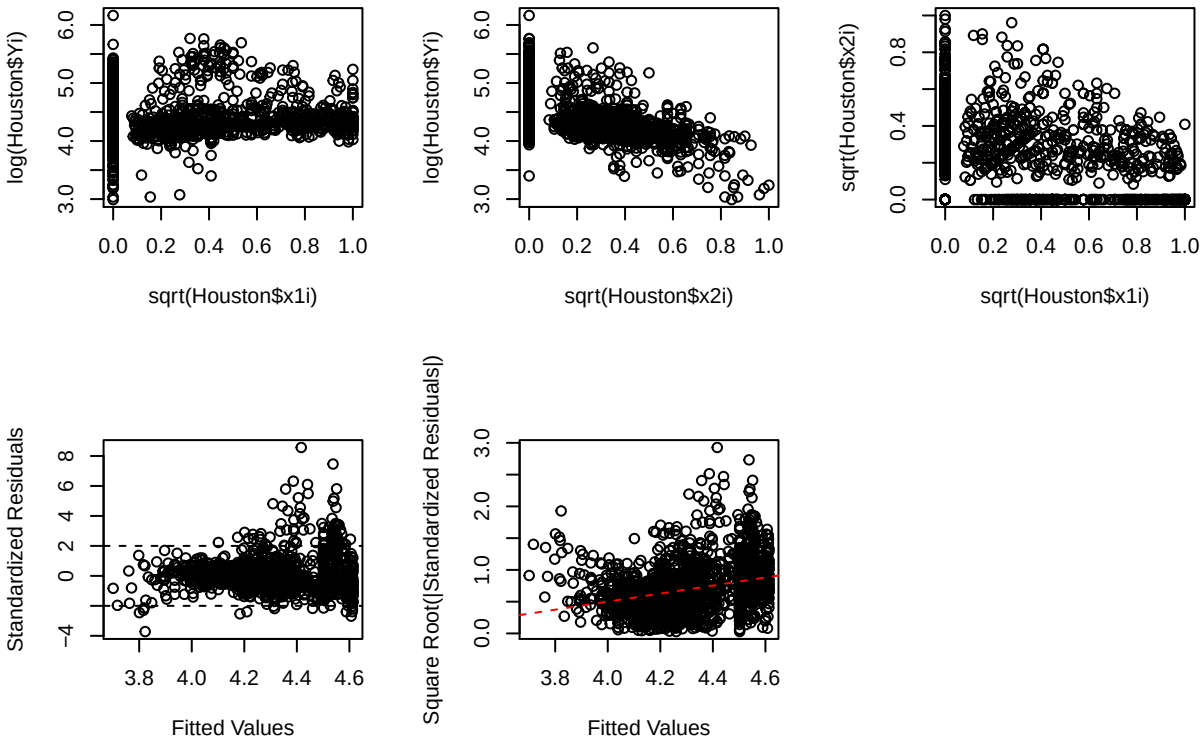
Part c [20 points]

```
#Figure 4.1 on page 123
m2 <- lm(log(Yi) ~ sqrt(x1i) + sqrt(x2i), data = Houston, weights = ni)
summary(m2)
```

```
##
## Call:
## lm(formula = log(Yi) ~ sqrt(x1i) + sqrt(x2i), data = Houston,
##     weights = ni)
##
```

```
## Weighted Residuals:
##    Min     1Q Median     3Q    Max
## -5.035 -0.723 -0.205  0.374 11.598
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.5020     0.0127  354.96  < 2e-16
## sqrt(x1i)     0.1111     0.0197    5.65  1.8e-08
## sqrt(x2i)    -0.8009     0.0344  -23.31  < 2e-16
##
## Residual standard error: 1.36 on 1919 degrees of freedom
## Multiple R-squared:  0.268,  Adjusted R-squared:  0.267
## F-statistic:  352 on 2 and 1919 DF,  p-value: <2e-16
```

```
leverage1 <- hatvalues(m2)
StanRes1 <- rstandard(m2)
absrtsr1 <- sqrt(abs(StanRes1))
residual1 <- m2$residuals
par(mfrow=c(2,3))
plot(sqrt(Houston$x1i), log(Houston$Yi))
plot(sqrt(Houston$x2i), log(Houston$Yi))
plot(sqrt(Houston$x1i), sqrt(Houston$x2i))
plot(m2$fitted.values,StanRes1, ylab="Standardized Residuals",xlab="Fitted Values")
abline(h=2,lty=2)
abline(h=-2,lty=2)
plot(m2$fitted.values,absrtsr1,ylab="Square Root(|Standardized Residuals|)",xlab="Fitted Values")
abline(lsfit(m2$fitted.values,absrtsr1),lty=2,col=2)
```



Transformations only marginally improve the fit. I would also try to categorize the data to account for the zero-inflation in the predictors. The pattern looks very different for when either $x1_i = 0$ or $x2_i = 0$. Moreover these predictors appear only weakly correlated with the outcome variable.

Project milestones [20 points]

1. Review the relevant literature.
2. Identify a gap in knowledge that may be able to be address with your dataset.
3. Update your research question and hypothesis.
4. Draft a preliminary introduction for your written report.

References

Sheather, Simon. 2009. *A Modern Approach to Regression with R.* Springer Science & Business Media.