Stats 195/CME 195 HW#2
Due April 25, 2017

Instructions:
1. Upload R script to Canvas using the filename <your SUNetid>stats195hw2.R). E.g., I would name my submission "hgm7stats195hw2.R".
2. Follow any function/variable/file naming instructions indicated below.
3. Check that your source code loads into memory without throwing errors by using `source`. E.g., I would ensure there are no errors when I run `source(hgm7stats195hw2.R)`.

Problem 1

In these two problems we will compute the training error and test error of a linear model. When you fit a model to data, the model usually better fits the data it was fit to than new data (the training error is usually lower than the test error). You can read up on training versus test error on Wikipedia, etc., but do not need this background to do the problem.

(a) Read the data set

`http://vincentarelbundock.github.io/Rdatasets/csv/Ecdat/Hedonic.csv`

into a data frame in R. The file is in comma-separated value ("csv") format with column labels and a column of row numbers. Take a glance at the resulting data frame to make sure it looks OK. The data set contains a number of measurements (the columns) on the homes in certain regions (the rows).

(b) Take your data frame from (a) and create a new data frame consisting of the columns `mv` and `age`. Call the new data frame `dat`. `mv` refers to median home value in the region and will be our response. `age` is a measure of the typical age of homes in the region and will be our predictor.

(c) Perform a linear regression of `mv` against `age`, but only use the first half of all the observations (the rows) of `dat`, not all of `dat`. We will use the remaining half of the observations to test our linear regression later on. You can separate out the first half of the observations using tools we have already encountered. Another option is to use the function `cut` and/or the `subset=` parameter of `lm`, which we haven't used in class but you can learn about in the help pages.

Create a scatterplot of the points used in the regression, along with the regression line, as we have done in class. Try to annotate the plot with a useful title, axis labels, perhaps a sub-title, etc. Name this plot <your SUNetid>stats195hw2c.png. (The extension may differ if you use a graphics format other than ".png".)

(d) What is the mean sum of squares of the residuals? The residual for a given value of `age` is the difference between, on the one hand, the actual value of `mv` corresponding to that given value of `age` (i.e., the point on your scatter plot where x-axis is equal to the given value of `age`) and, on the other hand, the predicted value of `mv` corresponding to that given value of `age` (i.e., the point on your regression line where the x-axis is equal to the given value of `age`). You can access the residuals in a linear model `my.lm` using `my.lm$residuals`. Given numbers $x_1, \ldots, x_n$, the mean sum of squares is $\frac{1}{n}\sum_1^n x_i^2$. This mean sum of squares, corresponding to the first half of the data set, is the *training error* or *in-sample error. Ans.* 0.07938241.

(e) Apply your linear model from (c) to the second half of the `age` observations in `dat`, i.e., use the model to predict home values of the second half of the observations. You can go about this using `predict.lm(my.lm, new.observations)`, where `my.lm` is your linear model from (c) and `new.observations` is a *data frame* containing the second half of the `age` observations. Another way to go about this is to look at the slope and intercept of the regression line (e.g., `summary(my.lm)`) and compute the value of this line when the independent variable (i.e., the x-values) assumes the values of the `age` observations in the second half of the data set. Either way, you should get a vector of values of the same length as the number of observations in the second half of the `dat`, each entry representing a predicted value of `mv`.

Make a scatterplot of the second half of `dat`, i.e., the second half of the `mv` observations versus the second half of the `age` observations. Use `lines` to plot the predicted values against the second half of the `age` observations. Name this plot <your SUNetid>stats195hw2e.png. (The extension may differ if you use a graphics format other than ".png".)

(f) How well did the linear model predict the second half of the `mv` observations? Take the difference of the predictions in (e) and the second half of the `mv` observations, and obtain the mean sum of squares of these differences. This mean sum of squares, corresponding to the second half of the data set, is the *test error* or *out-of-sample error. Ans.* 0.2098782.

Problem 2

The file `https://haben-michael.github.io/stats195/auth_sms.txt` contains Stanford two-step authentication codes among other control data. For example, line 26 is

```
Stanford Authentication Code: 1911427
```

whereas line 16 is

```
Date: Mon, 06 Feb 2017 12:26:52 -0800.
```

Extract all the authentication codes from the text file. You can assume that the authentication codes all occur on lines with the same format as line 26 above. Print out a table showing the number of occurrences of each digit among all the authentication codes:

```
  0   1   2   3   4   5   6   7   8   9
107 309 112 121 129  94 113 111 112 108
```

Extra Credit. This extra credit may require some prior statistics exposure. Run a chi-squared goodness-of-fit test (`?chisq.test`) to test whether the digits occur uniformly in the authentication codes:

```
Chi-squared test for given probabilities

data:  ***
X-squared = 271.31, df = 9, p-value < 2.2e-16
```

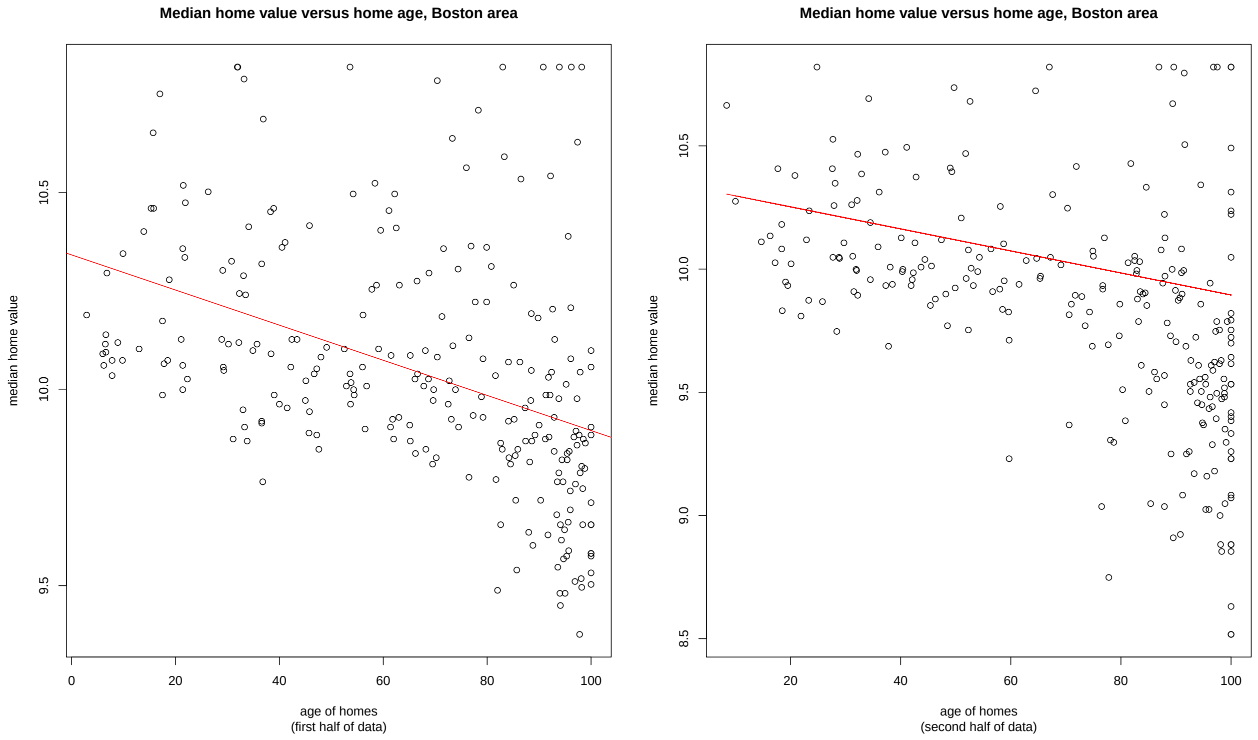

Figure 1: 1(c) and 1(e)