Homework Assignment 3: Old Heteroskedastic

36-402, Data Analysis, Spring 2011

Due at the start of class, 1 February 2011

The data set geyser in the library MASS contains a series of consecutive observations on the "Old Faithful" geyser at Yellowstone National Park, famed for the approximate regularity of its eruptions. There are two columns: duration, the length of the latest eruption of the geyser (in minutes), and waiting, the interval from the end of one eruption to the start of the next (also in minutes).

Begin by obtaining the library (if you don't have it already) and loading the data set. You should be able to reproduce this:

```
> summary(geyser)
    waiting          duration
 Min.   : 43.00   Min.   :0.8333
 1st Qu.: 59.00   1st Qu.:2.0000
 Median : 76.00   Median :4.0000
 Mean   : 72.31   Mean   :3.4608
 3rd Qu.: 83.00   3rd Qu.:4.3833
 Max.   :108.00   Max.   :5.4500
```

1. (5 points) Linearly regress waiting on duration. Plot the data points, together with the regression line. Comment (briefly) on any noteworthy features of the plot.

2. (5 points) Plot the squared residuals from the linear regression versus duration. Comment.

3. (20 points) Using the method presented in the notes to the lecture for 25 January, estimate the variance as a function of duration. Add it to the plot from the previous problem. Does the estimated variance function seem compatible with homoskedastic noise?

4. (10 points) Re-do the linear regression with weighted least squares, making the weights inversely proportional to the estimated variance function. What happens to the linear regression coefficients? Are the changes statistically significant? Does it seem like they matter?

5. (5 points) Do a nonparametric kernel regression of waiting on duration. Plot the results along with the raw data. Comment on how the results differ from the linear regression.

6. (20 points) Repeat the variance estimation using the residuals from the kernel regression. Compare this estimated variance function to the previous one.

7. (25 points) Use `npcdens` to estimate the conditional density of `waiting` given `duration`. Plot the results. (Three-dimensional, contour and level plots are all acceptable. Ask the instructor if you have another idea. You may find the examples in `help(npplot)` useful.)

8. (10 points) Describe how the plot of the conditional density relates to the plots you made in problems 5 and 6.

9. (10 points, extra credit) Suppose the Park Service wanted to provide tourists with estimates of the time until the next eruption of the geyser, including a margin of error. What model would you recommend they use? (Explain.)