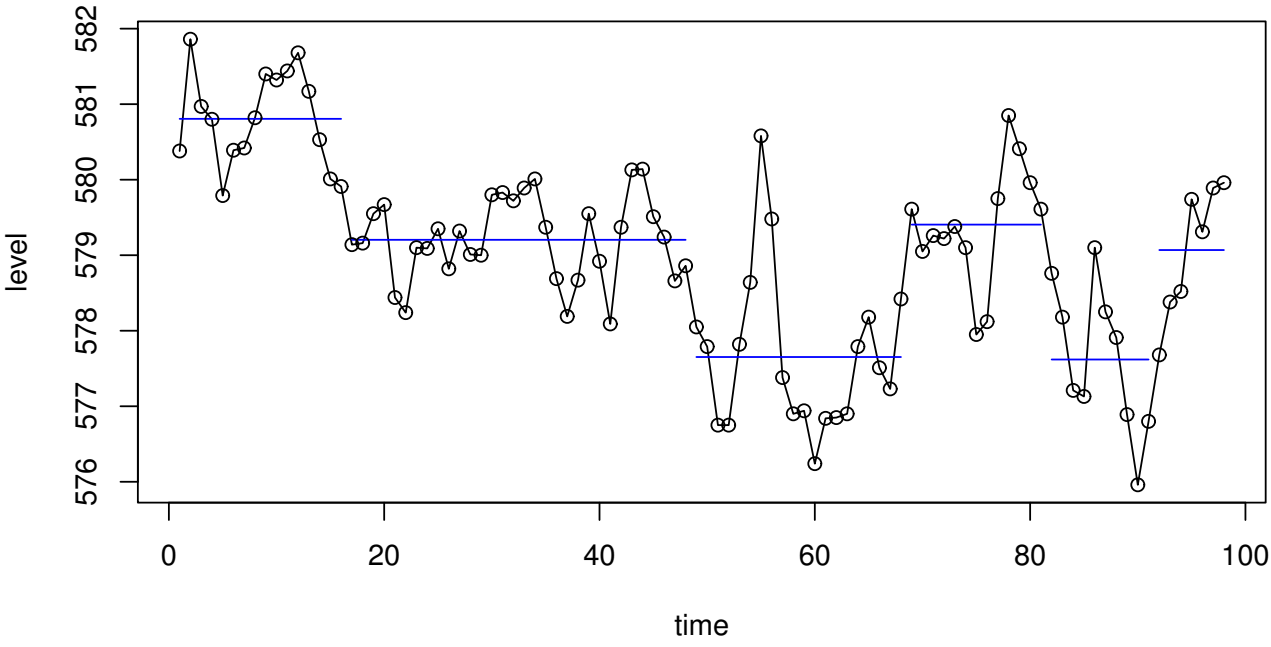
36-350: Data Mining

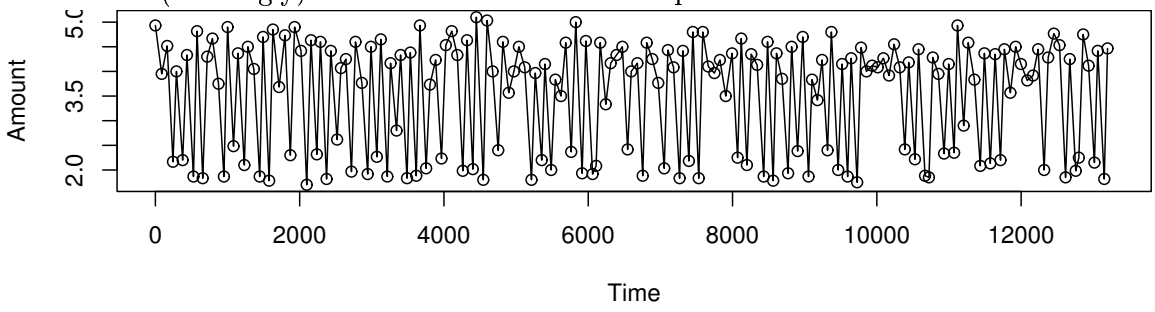
Homework 14
Date: December 1, 2002 **Due: end of lab December 5, 2002**

---

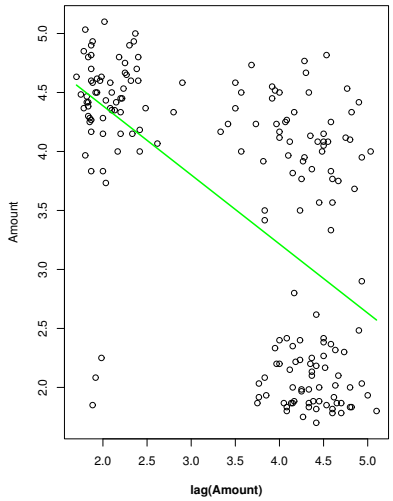
1. In class it was shown how to use Ward's method to find change-points in a time series. A simpler idea for finding change-points is to look for a large difference in the level between consecutive years. Does this idea work on the Lake Huron data? The data is plotted below, along with the result of Ward's method. The individual measurements are shown as circles.



2. The "Old Faithful" geyser in Yellowstone National Park shoots hot water in (seemingly) random amounts at (seemingly) random intervals. Here is a plot of the amount versus time:



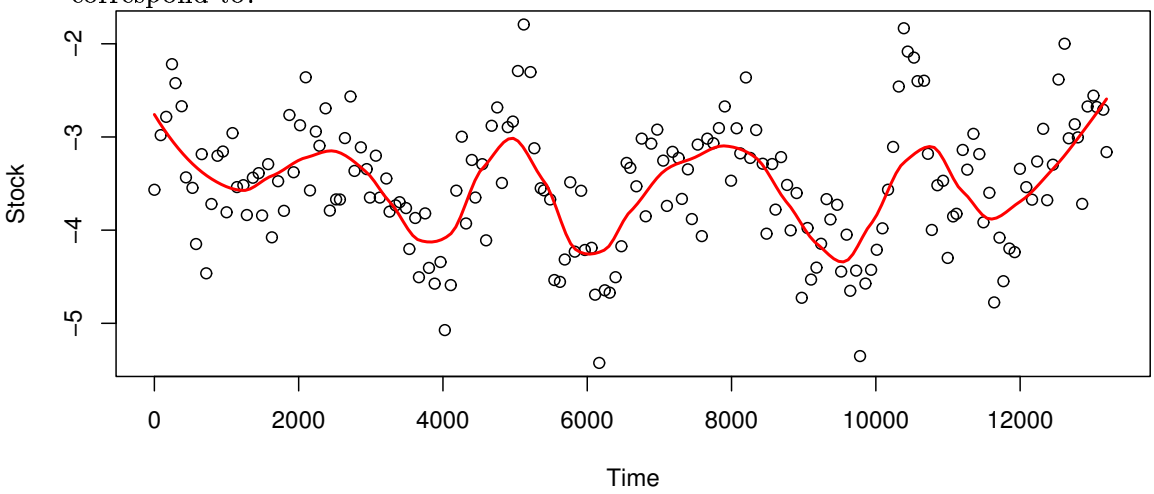
Here is a plot of the amount versus previous amount, with linear fit overlaid:



(a) The eruptions can be divided into "small" and "large". What usually follows a small eruption? What about a large eruption?

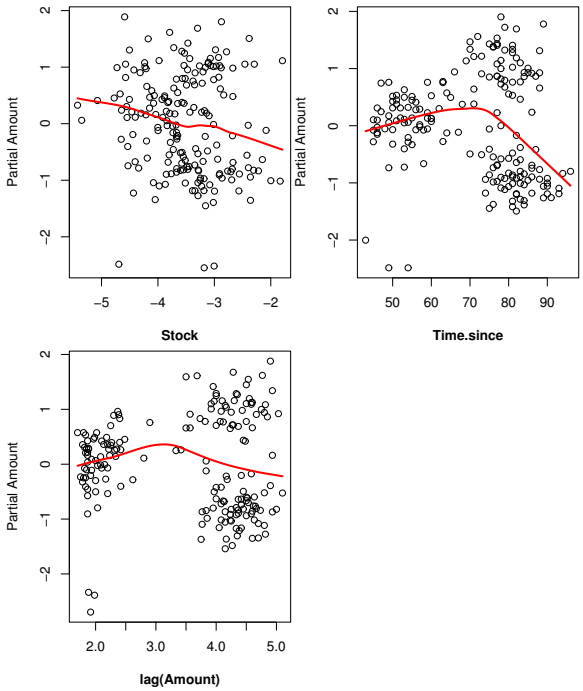
(b) To fit a Markov model, which regression method should be used here: linear or tree? Explain.

(c) To see if a restocking model would work, the cumulative amount of water was regressed on time, and the residuals are plotted below. The slope of the regression, i.e. the "consumption rate", was 0.05 units of water per minute. For the geyser, what does "consumption rate" correspond to?



(d) The amount was regressed on the estimated pre-eruption stock, the time since the last eruption, and the amount of the previous eruption. The coefficients and partial residuals are shown below.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.78610    0.56729  10.200  < 2e-16 ***
Stock       -0.22440    0.10442  -2.149 0.032954 *
Time.since  -0.03908    0.01128  -3.465 0.000661 ***
lag(Amount) -0.08469    0.14712  -0.576 0.565572
```



Which is a better model for this data: a restocking model or a linear Markov model? Explain.

3. In preparation for lab 14, study the R functions for linear regression.