STA 302 H1F / 1001 HF – Fall 2011 – A. Gibbs

Assignment 1

Due: Tuesday, October 11 at 12:00 noon.
You should plan to hand in your assignment in lecture on that day. Assignments will be accepted before lecture starts and at the end of lecture. No late assignments will be accepted without a valid reason. If you cannot attend lecture on that day, you must drop off your assignment before noon in my office, SS 5016A. Do not drop off the assignment after noon. If my office door is closed do not knock. There is no tutorial on Tuesday, October 11.

Presentation of solutions is important. In particular, it is inappropriate to hand in pages of SAS output without explanation or interpretation. The only SAS output you should submit with assignments is relevant plots. Quote relevant numbers from your SAS output as part of your solutions. You do not need to hand in your SAS code.

The Data:
The data for this question are available on the course web site under Assignments or at
`www.utstat.utoronto.ca/alisong/Teaching/1112/Sta302/assignments.html`.

A geyser is a naturally occurring hot spring that shoots hot water into the air from a reservoir. Once the water in the reservoir has depleted, there is a time gap until the next eruption while the reservoir refills.

The geysers in Yellowstone National Park in the U.S. are a popular tourist attraction. Many people come to see the geysers erupt. Old Faithful is the most famous geyser in the park. Since tourists want to see eruptions, park officials want to be able to predict when the next eruption will take place so that they can post a notice for the tourists. Since long eruptions deplete more water from the reservoir, the time interval to the next eruption is greater after an eruption of longer duration.

Research geologists have measured the duration of eruptions and interval to the next eruption for strings of consecutive eruptions at several different times. Because the distribution of intervals between eruptions is bimodal, these data are commonly used as examples in statistics textbooks. We'll look at the most recent data that I could find which was measured in July 1995.

If you've never seen a geyser, there are many videos on YouTube.

The variables in the dataset are:
- the date in July 1995 on which the eruption occurred
- the time of day (in minutes since midnight) at which the eruption began
- the interval (in minutes) from the previous eruption to the current eruption
- the duration (in minutes) of the previous eruption
- a column of 1's that you can ignore

For this assignment, we are interested in building a model (or models) that the park officials can use to inform tourists when to expect the next eruption.

1. The first thing you should always do when given a data set is to take a look at the data. Some SAS code that you might find useful to do this:

```
proc univariate;
  histogram time interval duration;
```

What values does each variable take on? Are there missing values? Can you see the bimodal nature of the data? Are there any surprises? Also look at a scatterplot of interval versus duration. Does a linear model seem appropriate?

We are going to look at all eruptions together and we'll also consider separately eruptions that follow eruptions of long duration and those that follow eruptions of short duration. We'll use 3 minutes as the cut-off, so durations less than 3 minutes are short and greater than 3 minutes are long.

Do not hand in anything for this question.

2. Carry out three simple linear regressions, one for all of the data, one for short durations, and one for long durations. In a table, give the values of the following for each of these regressions:
   - R^2
   - the estimated intercept
   - the estimated slope
   - the estimate of the variance of the error term
   - the p-value for the test with null hypothesis that the slope is 0
   - a 95% confidence interval for the slope

3. R^2 is much larger for the regression on all data than it is for both the regression on the short duration data and the regression on the long duration data. Explain why.

4. Consider the 2 regressions carried out on the short and long duration data separately. We can test to see whether there is a statistically significant difference between the slopes of the 2 regressions using a t-test, similar to the two-sample t-test for the difference between two means. We can estimate the difference in the slopes by $b_{1,short} - b_{1,long}$ where $b_{1,short}$ and $b_{1,long}$ are the estimated slopes for the short and long duration data, respectively. We will also need to find an estimate of the standard deviation of $b_{1,short} - b_{1,long}$; do not assume that the standard deviations of the estimators of the slopes are equal. Under the regression model assumptions and assuming that there is no difference in the slopes, the estimate of the difference in slopes divided by the estimate of the standard deviation of the difference will have approximately a t-distribution with 220 degrees of freedom using Satterthwaite's approximation for the degrees of freedom. (You can take my word for it that the approximate d.f. is 220, but feel free to verify it yourself.) What do you conclude from this t-test? (To estimate the p-value, you can use a t-table. There are lots of t-tables on-line if you don't have one handy.)

5. There are missing values in the dataset. Might this create any problems in making predictions about the time to the next eruption? Why or why not?

6. Conclusions from inferences on regression analyses are valid only if certain conditions hold. Considering what is being measured and how the measurements were taken, might there be any violations of the Gauss-Markov conditions for the regressions? (You do not need to look at plots of the residuals.)

7. Should the Yellowstone officials use one or two equations to calculate their predictions for time until the next eruption? Why?

Marking scheme: Each of questions 2 through 7 is worth 3 marks (for a total of 18). 3 marks will be awarded for complete, correct answers, or answers with only very minor problems. Good answers that are unclear or have some mistakes or are missing some aspects of the solution will be awarded 2 marks. Poor answers that have some value will be awarded one mark. Note that sometimes an answer awarded 3 marks will not be perfect. You should always look at the solutions.