Take-Home Midterm

Statistical Computing, 36-350

Due Wednesday Oct 21, 2015

Your midterm must be submitted in R Markdown format. We will not (indeed, cannot) grade this midterm in another format. Your responses must be supported by both textual explanations and the code you generate to produce your result. (Just examining your various objects in the "Environment" section of R Studio is insufficient—you must use scripted commands.)

Background. There are now over 1,000 confirmed planets outside of our solar system. They have been discovered through a variety of methods, with each method providing access to different information about the planet. Many were discovered by NASA's Kepler space telescope, which observes the "transit" of a planet in front of its host star. In these problems you will use data from the NASA Exoplanet Archive to investigate some of the properties of these exoplanets.

Note: even though this midterm appears to be about a physics data set, you do not need to know any physics whatsoever to succeed. This midterm, as it should be, is really about programming and statistics!

Part 0: Instructions

Make sure you read these.

- You will have been assigned a random partner for this midterm. You must work with this partner, you must submit your own copy of the lab report. In particular, submit a file called "AndrewID1-AndrewID2-midterm.Rmd" , where "AndrewID1" and "AndrewID2" are your and your partner's Andrew IDs. One of the two reports between you and your partner's (chosen at random) will be graded, and both partners will be assigned that same grade. Thus you must work closely together.
- You should download the "midterm.Rmd" file from course webpage and start filling in your answers from there. Display all code that you write, and every output that you're asked to provide, but suppress unnecessary outputs (e.g., printing out of long data tables). Avoid the `echo=FALSE` option or the inline code style.
- As usual, any code that does not knit as HTML (*including* stack space overflow) will get a zero.
- **You may not work with anyone else than your assigned partner; you may only ask clarification questions on Piazza; you may not ask questions for help. This is an exam, not a homework assignment.**

Part I: Getting Data on Planets

The data table provided by NASA is in HTML format. Read in the lines with the following command:

```
dt = readLines("http://www.stat.cmu.edu/~ryantibs/statcomp/exams/planets.htm")
```

1. The column names of the variables in our data table begin are given in line 20 of `planets.htm`. Write code to extract these names from `dt[20]`, and save the names in a vector called `exo.col.names`. Display the entries of this vector; it should have length 11.

2. Lines 21 and onward of `planets.htm` contain data. Write a function that takes as an argument `line`, which will be a string containing a typical line of data in `planets.htm`. The function should extract the data entries in `line`, and return a character vector of length 11, made up of these entries. Display the result of your function, when applied to `dt[21]`.

3. Now create an empty character matrix of dimension 1892 x 11, called `exo.mat`. Fill the rows of `exo.mat` by applying your function from question 2 to the relevant lines in `dt`. (That is, the first row is populated by the data in `dt[21]`, the second row is populated by the data in `dt[22]`, etc.) Assign the column names of `exo.mat` to be `exo.col.names`, and display the last 3 rows of `exo.mat`.

4. You now have a character matrix containing all of the relevant data for your analyses in Parts II and III; but clearly character format is not the most convenient one to work with. Inspecting the contents of `exo.mat` by eye, or the `planets.htm` file by eye, identify the column numbers in `exo.mat` that correspond to numeric data.

5. Define `exo = data.frame(exo.mat)`. This is a data frame that you have created from your character matrix. For all the columns that you determined actually contain numeric data (as per the last question), cast these columns in `exo` to be of numeric type. For all the other columns, cast these columns in `exo` to be of factor type. Display the last 3 rows of `exo`.

6. The column `pl_discmethod` of `exo` documents the method by which the planet was discovered. How many planets were discovered by the *Transit* method? How many were discovered by the *Radial Velocity* method?

Part II: Kepler's Third Law

Kepler's third law states that when the mass M of the host star is much greater than the mass m of the planet, the orbital period T satisfies

$$T^2 \approx \frac{4\pi^2}{GM} a^3.$$

Above, a is the semi-major axis of the planet's elliptical orbit, G is Newton's constant.

7. The orbital period T is found in `pl_orbper`, and the mass of the host star M in `st_mass`, and the semi-major axis a in `pl_orbsmax`. (The variable names here, and henceforth, refer to columns of the data frame `exo`.) We will want to work with these in the next few questions; to prep for that, define `ind.na` to be a Boolean vector, whose entries are `TRUE` whenever at least one of `pl_orbper`, `st_mass`, or `pl_orbsmax` is unobserved (i.e., takes an `NA` value), and `FALSE` otherwise. Then define `ind.obs = !ind.na` as the vector of Booleans such that `pl_orbper`, `st_mass`, and `pl_orbsmax` are all fully observed. How many planets are left, i.e., how many `TRUE` values are there in `ind.obs`?

8. Make a plot of $\log(T)$ versus $\log(a)$ but only for systems where the host star mass is between 0.9 and 1.1 solar masses. Be sure to include only planets where the orbital period, host star mass, and semi-major axis are all observed (i.e., use `ind.obs`, as constructed in last question). Also be sure to properly label the axes. What kind of relationship do you see?

9. Perform a linear regression of a response $\log(T)$ onto predictor variables $\log(a)$ and $\log(M)$, using the same data set as in the last question (planets where the host star mass is between 0.9 and 1.1, and where orbital period, host star mass, and semi-major axis are all observed). What values would you get for the two slopes? Does this match would expect, given Kepler's third law (displayed above)? What value do you get for the intercept? (*Bonus:* does this match what you'd expect, given Kepler's third law?)

10. What are the standard errors of the estimated regression coefficients you found in the last question, as printed out by the `summary()` function?

11. Write a function `reg.coeffs` that takes three arguments `T`, `a`, `M`, standing for the orbital period, semi-major axis, and host star mass, and performs a regression of $\log(T)$ onto $\log(a), \log(M)$, as you implemented in question 9 for the exoplanet data. The output of `reg.coeffs` should be the regression coefficients. Run this function on the exoplanet data from question 9 and verify it gives the same results.

12. Compute a jackknife estimate of the standard error for each regression coefficient found in question 9. Your code here should use either a `for()` loop, or the function `apply()`, and should involve repeatedly calling the function `reg.coeffs()` defined in question 11. You may wish to revisit Homework 5 for a refresher on the jackknife estimates of standard errors. How do your estimates compare to those from question 10?

Part III: Masses from Radii

13. Not all methods of exoplanet detection provide a measurement of the planet's mass. It will be useful to consider the relationship between planetary mass m (in `pl_massj`) and radius R (in `pl_radj`), since the radius is sometimes available even when the mass is not. (For more information on this kind of relationship, see e.g. Wolfgang et al. (2015).) We'll focus our attention on planets with $R \leq 4/11$. How many such planets have a measurement of both their radius and mass available?

14. Plot $\log(m)$ versus $\log(R)$, when these are both observed. Then perform a linear regression of the response $\log(m)$ onto the predictor variable $\log(R)$, again using the data over which these are both observed. Add the estimated regression line to your plot. What is the intercept? What is the coefficient for $\log(R)$? Are they significant, according to the `summary()` function?

15. Plot the residuals (the actual values of $\log(m)$ minus the fitted values) versus $\log(R)$. Do you see any trends and if so what does this say about your fitted linear model?

16. Use your regression model from the last question to predict a mass for all planets in the data set. Add this prediction to your data frame `exo`, under the column name `pl_model_massj`. Note that our model has been developed for a limited range of planet radius—set your prediction to `NA` for planets with a radius outside that range. What is the predicted mass for planets 72, 83, and 92 in the data set?

Bonus: More Exoplanets

17. Download an alternative exoplanet data set, perhaps from Exoplanets.org or the Open Exoplanet Catalogue. Determine how many planets appear both in the new data set and in the data used for this midterm, as well as how many planets appear in only one data set. Estimate the mass-radius relationship using the new data set and the radius range considered in Problem 13. Does this relationship agree with the relationship you found in Problem 14?