Homework 4

Due Saturday, March 19, 11:59pm

Note on Submission Packages

Make sure everything is in one directly and tar'd together using our UCD email names.

Problem A

Here you will analyze the Movie Lens data, 100K version.

Note that successive ratings by the same user are not independent, so our confidence intervals would not be correct if we were to use the data in this raw form. To cope with this problem, let Ai denote the average rating that user i gives to the various movies he/she rates. If for instance user 29 rated 3 movies, then A29 will be the average of those 3 ratings. Our data will now have the samplee size n equal to the number of users. Include in your report your R code for computing the Ai, and an outline of how it works. (I suggest you use **tapply()**, or just **split()**)

Note that we are treating the users here as a random sample from the (rather conceptual) population of all possible users, and the same for the movies.

a) Find an approximate 95% confidence interval for the population mean rating by men.

b) Find the same for women.

c) Find an approximate 95% confidence interval for the difference between the two means.

d) Do a significance test of the hypothesis that the male and female population means are equal.

e) Plot histograms of the male and female ratings on the same graph.

f) Find an approximate 95% confidence interval for the difference between the population mean number of ratings by men and women.

g) Find an approximate 95% confidence interval for the population proportion of users who are men.

h) Using a linear model, estimate the population regression function in which we predict rating from age and gender.

- Form a confidence interval for β_{age}, and test the hypothesis that the coefficient is 0.
- Form a confidence interval for the mean population rating among women of age 28.

Problem B

This problem concerns some fascinating data sets at Stanford, Wordbank, "An open database of children's vocabulary development."

Note that the maintainers of the database have made things convenient for us, remarking,

Wordbank is open access! You can use the wordbankr package to access Wordbank data from R.

Unfortunately, this in turn requires installing other packages, such as the very popular **dplyr**. You'd learn from this, but in this case it is probably not worth the trouble, especially since the functions the package provides don't seem to be too relevant to what you'll be doing in this problem.

Instead, to get the data directly, go to English vocabulary norms, and click on Download Raw Data, to acquire the file **vocabulary_norms_data.csv**.

By the way, there are some nice graphs, and the R code that generated them, throughout this Stanford Wordbank Web site, such as this page. They rely on various R libraries, and thus the code is probably not worth trying to comprehend, but the graphs will give you an idea of what can be done.

This problem is quite open-ended. You will predict vocabulary size from age (in months), birth order, ethnicity, sex and mom's education, using a linear regression model (more from a Description point of view than Prediction). Clearly you will need to form dummy variables from ethnicity and sex, but it will be up to you what to do with birth order and mom's education. You are required to have some graphs, but it is up to you what to graph – univariate measures, bivariate relations and so on.

Some General Issues

- As you know, this Project counts both as a Homework assignment and as a substitute for an in-class Final Exam. All in all, it will require time similar to approximately 1.5 Homework assignments. **START EARLY!** A lot of things you think will be easy, postponable to the last minute, may prove to be much harder than you think.

  A typical good-quality report will run, say, 6-7 pages, excluding code listings, graphs and figures. But there is no magic formula; a really excellent report might be shorter than this, and a mediocre one might be much longer.

- **Please note the open-ended nature of the Project.** As you can see, these Project specs don't take the form of "Step 1, do this, Step 2, do that..."

Writeup:

- You are required to use LATEX to write up your report, with the output being a PDF file. I have a quick tutorial. Also, all the **.tex** files for our own ECS 132 textbook are in **http://heather.cs.ucdavis.edu/ matloff/132/PLN**; by comparing output to input, you can see how to do lots of things in LATEX .

  Make sure that your graphs are in either **.png**, **.jpg** or **.pdf** format. You may find the following code useful:

```
# prints the currently displayed graph to the
# file filename; suffix can be "pdf", "png" or "jpg"
pr2file <- function (filename)
{
    origdev <- dev.cur()
    parts <- strsplit(filename,".",fixed=TRUE)
    nparts <- length(parts[[1]])
    suff <- parts[[1]][nparts]
    if (suff == "pdf") {
        pdf(filename)
    }
    else if (suff == "png") {
        png(filename)
    }
```

```
        else jpeg(filename)
        devnum <- dev.cur()
        dev.set(origdev)
        dev.copy(which = devnum)
        dev.set(devnum)
        dev.off()
        dev.set(origdev)
    }
```

I recommend that you use R's **ggplot2** package to generate your graphs; see my tutorial at the end of my ECS 132 book. Or you may wish to use **lattice**, another popular package (for which there are many tutorials on the Web). In any case, all graphs and figures must be input by your **.tex** file, and appear within your report itself. The code generating the graphs must be R.

- It is of the utmost importance that your report be of **PROFESSIONAL QUALITY**. This means:
  - Clarity! This is very difficult, but quite achievable if you work at it. Ask friends not in the class to read it and tell you if it makes sense. Keep in mind that if you do the writeup at the last minute, the biggest casualty will be clarity.
  - Correct spelling, grammar and usage!
  - Professional-quality presentation! It doesn't necessarily have to be fancy, but should not look sloppy.
  - To be avoided like the plague:
    - Avoid vague use of the word "it." Example: "It is embarrassingly parallel"–exactly WHAT is embarrassingly parallel?
    - "If you wouldn't eat it, don't serve it." This used to be a sign to employees in fast food restaurants, admonishing them not to serve sloppily cooked food. In our case here, "If you wouldn't find the premise of an analysis credible, don't present that analysis."
  - You must include an appendix (use the LATEX **appendix** command), with the following contents.
    - Your code, clearly commented, using the LATEX **listings** package. Also include in a a text portion (i.e. NOT in your code) an explanation of the highlights of your code.
    - Include a brief account of "who did what" in this Project among the various group members. Not everybody need have done exactly equal work, but something of the nature "Transported a USB key to campus" doesn't count as a contribution. :-) (A group acftually said this in their report one time about the contribution of a certain member.)
- Interpret all of your results; don't just say, "The confidence interval is (1.68,1.88)." Assume that the reader has technical background, i.e. understands quantitative issues, but has little or no background in statistics.
- Adhere to the UCD Code of Ethics. Having persons outside your team do any of the work, including the writing, or engaging in paid help of any kind, is unacceptable, and is a serious violation that will be referred to Student Judicial Affairs. There are probably other analyses

of this data on the Web; if you draw any ideas from any of them, that is fine but be sure to cite them.

Important Submission Details

- Your group submits just ONE copy of the report.
- Your writep files must be named **Project.tex** and **Project.pdf**.
- Submit your report, including all files (**.tex, .pdf, .c,.R** etc.) to my **handin** site on CSIF, directory **132project** (see Syllabus). The name of your file must be of the form **email1.email2....tar**, where each **emaili** is the UCD e-mail address of group member i, e.g. **bclinton.gbush.bobama.tar**. Note the periods separating fields. Your **.tar** file must contain only regular files, NO SUBDIRECTORIES.
- Make sure that all partners' names are on the report, and that the e-mail addresses in the file name are EXACTLY the official UCD e-mail addresses for the students. These are the addresses at which you have been receiving your Quiz results. DON'T RISK HAVING A TEAM MEMBER FAILING TO GET CREDIT FOR THE PROJECT.

Grading Criteria

- Technical content of the work.
- Adherence to instructions.
- Professional quality of the work.