Gmail **Andrew Ho <kironide@gmail.com>**

# Week 1 Day 1 Assignment

4 messages

**Jonah Sinick** <jsinick@gmail.com> Mon, Feb 15, 2016 at 9:52 AM
To: david@bolin.at, Ali Bagherpour <ali.bagherp@gmail.com>, Andrew Ho <Kironide@gmail.com>, Chad Groft <clgroft@gmail.com>, Jacob Pekarek <jpekarek@trinity.edu>, Jaiwithani <jaiwithani@gmail.com>, James Cook <cookjw@gmail.com>, Linchuan Zhang <email.linch@gmail.com>, Matthew Gentzel <magw6270@terpmail.umd.edu>, Olivia Schaefer <taygetea@gmail.com>, Tom Guo <tomguo4@gmail.com>, Trevor Murphy <trevor.m.murphy@gmail.com>, Sam Eisenstat <sam.eisenst@gmail.com>

**Warmup**

- Install the packages, "car," "Ecdat", "HistData", "ggplot2" and "GGally."
- Go through the attached file day1Example.R, making sure that you understand the example.
- Write a function cor2(df1, df2 = df1) that takes two dataframes and

  i) Selects their numeric columns

  ii) Computes the correlation matrix obtained by for each pair of columns removing entries for which one or or both of the entires is missing.

  iii) Multiplies the correlation matrix by 100 and then round it, for readability.
- Write a function remove.na.rows(df) that removes rows with missing values from a data frame.

**Exploring the Galton Height Data**

Note: In the discussion below, "run a regression of A against X, Y and Z" should be understood as "first run a regression of A against X, Y and Z individually, perhaps in pairs, then run the regression against all three variables"

1. Aggregate the data by family, first converting gender to a numeric variable.

   a) Plot the average height of children in a family against mother's height, father's height and midparentHeight. You may find multiplot useful for comparing http://www.cookbook-r.com/Graphs/Multiple_graphs_on_one_page_(ggplot2)/.

   b) Compute the correlations between the variables obtained.

   c) Compare multiple R^2 and adjusted R^2 for

   lm(childHeight ~ father + mother)
   vs.
   lm(childHeight ~ midparentHeight)

   and interpret the results.

   d) Run a regression of number of children in a family against all three of father's height, mother's height and average child height. Does the regression capture a statistically significant relationship?

e) Run regressions of average % male children against each of father's height, mother's height and midparent height. What do you find? What if you include childHeight as a predictor?

f) Make a histogram of counts of number of children for each family.

Now consider the original data frame.

1. Run a regression of child's height against mother's height, father's height and/or midparent height, and gender.

   Now run regressions against mother and fathers' heights after restricting to those children of each of the two genders.

2. For those families with two or more children, create a new data frame with rows corresponding to families, columns with heights corresponding to the first and second born children, and columns corresponding to their genders.

   How much additional predictive power comes from including sibling height along with parent heights for the second born child? For the first child? What happens if we restrict to subsets corresponding to the four combinations of gender for first and second born children?

3. Do the same as (2) but restricting to families with 3 or more children.

**day1Example.R**
2K

---

**Jonah Sinick** <jsinick@gmail.com> Mon, Feb 15, 2016 at 2:50 PM
To: david@bolin.at, Ali Bagherpour <ali.bagherp@gmail.com>, Andrew Ho <Kironide@gmail.com>, Chad Groft <clgroft@gmail.com>, Jacob Pekarek <jpekarek@trinity.edu>, Jaiwithani <jaiwithani@gmail.com>, James Cook <cookjw@gmail.com>, Linchuan Zhang <email.linch@gmail.com>, Matthew Gentzel <magw6270@terpmail.umd.edu>, Olivia Schaefer <taygetea@gmail.com>, Tom Guo <tomguo4@gmail.com>, Trevor Murphy <trevor.m.murphy@gmail.com>, Sam Eisenstat <sam.eisenst@gmail.com>

Here is the remainder of today's assignment. No worries if you don't get to this part, or don't complete it – I'm front-loading the assignments with the most crucial material; the goal here is for nobody to be left without something to do.

**Afternoon: The effect of educational expenditures on test scores**

**States dataset**

1. Load the States dataset from the car package, and read about it using help(States), and set df = States.

2. Try computing the correlations between the columns with cor(). What went wrong? Now use your function cor2(). Try also cor(States[-1]) and interpret the results.

   You can display the correlations visually using the library "corrplot." Try

   statesCor = cor(df[-1])
   corrplot(statesCor, method = "pie").

   Interpret the results.

3. Add a "SAT" column defined as the sum of SATV and SATM.

Run each of the following regressions in sequence, each time using summary() to inspect the coefficients, Multiple R-squared, Adjusted R-squared, and interpreting the results.

(i) SAT against pop, percent, dollars and pay
(ii) SAT against pop, dollars and pay
(iii) SAT against dollars and pay
(iv) SAT against dollars
(v) SAT against pay
(vi) percent against pop, dollars and pay.

4. Aggregate the regions using

   aggregate(df[-1], df["region"], FUN = median)

   and compute the correlations between the resulting columns. How do these compare with the correlations in part(1)? What do you think explains the difference?

**Massachusetts Test Score Data Set**

1. Load the MCAS dataset from the car package, and read about it using help(MCAS), and set df = MCAS.

   Find the total number of rows. Remove the rows with missing values, and compute the number of rows of the resulting data frame. Is the number of rows appreciably smaller?

2. Compute the correlations of totsc4 and totsc8 with the other features in the dataset. Why do you think the correlations with totsc8 tend to be larger than the correlations with totsc4?

3. Run a regression of totsc8 against total expenditure per student, totday. What does this say about the effect of spending per student on standardized test scores?

4. Form a new data frame df1 by removing the columns that are not numeric, as well as totsc4. and run a regression of totsc8 against the other features using

   lm(totsc8 ~ . , df1)

   Should we be including "code" as a predictor?

5. Run a regression of totsc8 against the 3 predictors p value < 0.01 in the above regression. Is the predictive power appreciably lower?

6. Use *stepwise regression* by writing

   m0 = lm(totsc8 ~ 1, df1);  m = lm(totsc8 ~ , df1)

   s = step(m0, formula(m), direction = "both)

   How does the resulting model differ from the model above?

   What does this say about the effect of educational expenditure on student test scores for schools in the dataset? Reconcile this with (3).

7. Repeat the above with totsc8 replaced by totsc4 and compare the results.

**California Test Score Data Set**

Explore questions analogous to the ones above for the Caschool dataset in the Ecdat package, and interpret the results.

**Effects on Learning of Small Class Sizes**

Open the Star dataset from the Ecdat package, restricting consideration to those students who were either in regular classes or in small classes. Explore questions analogous to those above, and interpret the results.

---

**Andrew** <kironide@gmail.com> Mon, Feb 15, 2016 at 4:46 PM
Reply-To: Kironide@gmail.com
To: Jeremy Li <h.jeremy.li@gmail.com>

[Quoted text hidden]

---

**Andrew Ho** <kironide@gmail.com> Fri, Apr 29, 2016 at 6:15 PM
Draft To: Jonah Sinick <jsinick@gmail.com>

Notes on this assignment for updating/improvement:

- More explicit instructions, e.g. "Following the example of function definition above, write your own functions to ..."
- The multiplication and rounding of the correlation matrix is a good place to introduce the fact that matrices are just numeric vectors with attributes
- Writing remove.na.rows(df) is complex for beginner R students. This might be a good place to have them explicitly break down the problem into a hierarchical model of steps (e.g. "okay, so an approach might be: we loop through each row; now, for each row, we do X; now, task X requires us to figure out how to do Y and Z; ...")

[Quoted text hidden]