Statistical Analysis Empirical Exercise

Due: Oct. 07 2018 (Sunday), 11:59 PM

This Empirical Exercise is to familiarize you with the R software for statistical computing, and let you try out several basic and common statistical tools we learned in class. Including the following:

- Basic summary statistics
- Confidence intervals
- *t*-tests
- Simple and multiple linear regressions

Each of you need to work on the project and *individually* submit a report on Blackboard by the deadline. **You should also submit your code, preferably in the format of "XXX.R" (if you are using R).** You are welcome to discuss with others, but please do not copy others' results. This instruction will guide you through the entire exercise.
Our instructions are based on R, but feel free to use other software (your code is still needed).

About the data

We will work with a simplified/cleaned version of "Data on Social Networks and Microfinance in Indian Villages", published in Prof. Matthew Jackson and coauthors' websites.
Here is the original link to the data: http://web.stanford.edu/~jacksonm/Data.html. It has data on seven different relations (borrowing and lending rice, borrowing and lending money, going to temple together, etc.) in 75 of villages in India, together with some demographic information. For the current project, we will be mostly working with partial data on village # 1.

Our data file "v1hh.csv" is cleaned/simplified based on "households" (that's why every variable has "hh" in its name)[1] including following variables:

- Each row is specified by a pair of households, "mpid_hh" and "pid_hh". In particular,
    - "mpid_hh" is the id for the first household. For instance, for a number of "01002": the first two numbers represent the village (01) and the next three numbers represent the household (002).
    - "pid_hh" is the id for another household. The two households may have (potential) relations.
    - (some pairs of households are missing due to missing/incorrect data)

[1] Unlike the original data where each observation is based on individual villagers (each of whom belongs to a household), the dataset you get is already an aggregate data based on households. If you are interested in original data on individual level, you are more than welcome to play with "v1.csv".

- "templecompany_hh" is a 0-1 variable. 1 if the pair of households (in that row) has a relation of going to temple together; 0 if not.
- "keroricego_hh" and "lendmoney_hh" are both 0-1 variables, similar to "templecompany_hh". For each of them, a value 1 means the first household (mpid_hh) lends rice/money to the second household (pid_hh) in that row.
- "sameoccupation_hh" = 1 if the two households (mpid_hh and pid_hh) have the same occupation. 0 if not.[2]
- "samecaste_hh" = 1 if the two households (mpid_hh and pid_hh) are in the same caste. 0 if not.[3]

Install R

R is a **free** software available for both *Mac* and *Windows* (as well as *Linux*). Its official website has a lot of basic information: https://www.r-project.org/

R have many versions and can be downloaded from many places. We use "R-Studio" which has a more user-friendly and intuitive interface. It can be downloaded from the following link: https://www.rstudio.com/products/rstudio/download/#download

Open a dataset

First we open a new "R Script". It is a .R file. You can type your codes, commands and notes in this file.
To type a note (which will not be run): add "#" in front of the line. e.g.,

```
## This is a R script for "Statistical Analysis" empirical exercise
## [today's date]
## [your name]
…
```

(note: like the above, our command will be in font `courier`)

You can easily import .txt or .csv data files to R. To do so, you can either click "Import Dataset" in "Environment tab" and choose your data file, or type in the following command:

```
v1hh <- read.csv("/Users/Yiqing/v1hh.csv")
```

[2] There are 45 different occupations in the data, with the majority being "Agriculture labour". If there is some villager in household pid_hh has the same occupation with one villager in household mpid, then sameoccupation = 1.

[3] To know more about caste system in India, see https://en.wikipedia.org/wiki/Caste_system_in_India

in the above example, v1hh on the left is the name for your destination dataframe, and /Users/Yiqing/v1hh.csv is the location of the file, on my computer. (Notice the location is probably different in your computer).

When you wish to run some command, select those command(s) and click "**Run**" (a button at the top-right corner of the current tab).

1. Describing the data

First, to find your sample size, use the following function:

```
length(v1hh$mpid_hh)
```

Function "summary" gives you basic descriptive statistics, including mean, median, 25th and 75th quartiles, min, and max.

You can either run

```
summary(v1hh)
```

which summarizes all variables in your data v1hh; or

```
summary(v1hh$mpid_hh)
```

which summarizes the specific variable mpid_hh.

The command **sd()** displays the standard deviation for a variable X:

```
sd(v1hh$X)
```

and **cor()** shows the correlation coefficient between two variables, X and Y:

```
cor(v1hh$X, v1hh$Y)
```

When conducting a 5%-level test, we often want to compute the 0.025-th and the 0.975-th quantiles (i.e., the 2.5-th and 97.5-th percentiles) of some distribution. If we want the quantiles for the *t* distributions, we may write the following lines for example:

```
N = 30
df = N – 2
qt(0.025, df)
qt(0.025, df)
```

These return the numbers which we would see in the (two-sided) *T* tables.

If our sample is stored as a vector *y*, the following code will calculate the confidence interval:

```
ybar <- mean(y)
```

```
n    <- length(y)
s    <-  sd(y)
se   <- s/sqrt(n)
c    <- qt(0.975,n-2)
CI   <- c(ybar – c*se, ybar + c*se)
```

1.1 Find the sample size.

1.2 Find the mean and standard deviation for the following variables:

templecompany_hh

lendmoney_hh

keroricego_hh

1.3 Find the correlation coefficient between `keroricego_hh and samecaste_hh`

1.4 How many pairs (in total) have the relation "lendmoney_hh" in this village?

1.5 Is the average number relation "lendmoney_hh" greater than 0. 035 at 5% significance level?

Run the t-test below and report/interpret what you find.

```
t.test(v1hh$lendmoney_hh, mu = 0.035, alternative = c("greater"))
```

(lendmoney_hh is the variable you are looking at, mu = 0.035 is the hypothesized mean, and alternative = c("greater")) indicates that this is a one-sided test with ">" in H_a)

1.6. Construct a confidence level for the population mean for lendmoney_hh, at confidence level 95%.

2. Simple linear regressions

"Interacting with similar others?"

Let us examine whether people tend to interact more with "similar others".

First consider the relation "keroricego _hh" as dependent variable, and "samecaste_hh" as the explanatory variable.

We run the following regression:

```
lm1 <- lm(keroricego_hh ~ samecaste_hh, data = v1hh)
```

(in which keroricego_hh is the y-variable, samecaste_hh is the explanatory variable, and v1hh is the data you are using. lm1 <- is to define a variable "lm1" that records the result of your regression)

To display the regression result, run the following:

```
summary(lm1)
```

To plot the OLS regression line:

```
abline(lm1)
```

2.1 Find the intercept and slope. Find the p-values for both. How do you interpret the slope (together with its p-value)?

Compare your answer here to your answer to question 1.3, what can you find?

Plot the regression line, what can you find?

2.2 Repeat what you did in 2.1, with "templecompany_hh" being the new dependent variable (instead of "keroricego_hh").

2.3 Is there any difference(s) between what you found in 2.1 and 2.2? (e.g., in terms of slope, p-value, etc.)

If you found any differences, can you tell some story/intuition to justify those?

3. Multiple linear regressions

"Multiplexity": there can be different kinds of relations. For instance, in this data there are three kinds of relations: (1) lending money, (2) lending rice, and (3) going to temple together. Here we ask the following questions:

Are the relations independent? In particular, (for a pair of households):

Does having one relation (e.g., going to temple together) significantly increase or decrease the likelihood of having another relation (e.g., lending rice)?

First run the following (simple linear) regression:

```
lm3 <- lm(keroricego_hh ~ templecompany_hh + lendmoney_hh, data = v1hh)
```

In this regression, keroricego_hh is our dependent variable, and there are two explanatory variables: templecompany_hh and lendmoney_hh.

Again use the summary function to see the result:

```
summary(lm3)
```

3.1 Report the following information:

Coefficients of the two explanatory variables,

Corresponding p-values,

(Adjusted) R-squared,

3.2 Interpret your result based on coefficients and p-values you found in 3.1.

3.3 For a pair of households such that "templecompany_hh = 1" and "lendmoney_hh = 1", what is the predicted/fitted value (probability) that they also have the relation "keroricego_hh_fit"?

(a) First calculate the fitted value based on what you answered in 3.1.

(b) Verify your answer using the following code in R:

```
keroricego_hh_fit <- predict(lm3)

which(v1hh$templecompany_hh == 1 & v1hh$lendmoney_hh == 1)

keroricego_hh_fit[XXX]
```

(function predict predicts values based on the linear model you specified. It defines a vector, with one predicted value for each row.

The second line of code, with function "which", finds the rows in which the condition(s) are satisfied. Use any of those numbers as your "XXX" in the third line of code, which returns the predicted value of interests.)

References:

Abhijit Banerjee, Arun Chandrasekhar, Esther Duflo, Matthew O. Jackson (2015) ``The Diffusion of Microfinance,'' *Science* Vol. 341 no. 6144, DOI: 10.1126/science.1236498

Matthew O. Jackson, Tomas Rodriguez-Barraquer, Xu Tan (2012) ``Social Capital and Social Quilts: Network Patterns of Favor Exchange,'' *American Economic Review* Vol. 102, Iss. 5, 1857--1897

Chen Cheng, Wei Huang and Yiqing Xing (2018) ``Sustaining Cooperation with Multiple Relationships,'' mimeo.