

Module 4 Lab: data manipulation

BIOSTAT 213: Introduction to Computing in the R Software Environment

Wednesday, February 6, 2019

Save your code in a script file. I suggest adding labels and notes to this script file as comments. You may work in groups of 2–3.

Use the `presidencies.csv` data for the following:

1. Create a new data frame with all of the original variables, subsetted to the presidents from the state of Virginia.
2. Find all of the presidents whose names (`name`) begin with `John`.
3. One president served two non-consecutive terms, and is thus considered to have served two presidencies. Use R to identify this individual. Hint: you might consider revisiting the section on numeric objects in Module 1.

Use the `nhanes.csv` data for the following:

4. List the food security variables (`fs1`, `fs2`, and `fs3`) for individuals 24 years of age (`ridageyr`) or younger with a history of military service (`military`).
5. Fit the following model:

```
> mm<-lm(bpxdi1~bmxbmi+ridageyr+gender,data=ii)
```

Create a data frame, using `data.frame()`, that contains the coefficient estimates, the lower bounds of the 95% confidence intervals, and the upper bounds of the 95% confidence intervals. Your final data frame might look like this:

	estimate	lower	upper
(Intercept)	62.677125492	60.85606519	64.49818579
bmxbmi	0.217201438	0.16648600	0.26791687
ridageyr	-0.005565416	-0.02560865	0.01447782
genderMale	2.539452656	1.83786596	3.24103935

6. How many individuals in the data set are 70 years of age (`ridageyr`) or older *and* have a body mass index (`bmxbmi`) of 30 kg/m² or greater?
7. Find the mean age of individuals who reported an average of 2 hours of sleep (`sleep`) per night.
8. The variable names in the data have varying numbers of characters. For example, the first variable name, `hhid`, has 4 characters. Identify the variable name with the most characters.

Use the `unps.csv` data for the following:

9. Add a variable to the data frame for age group, as a factor with 2 levels: one for individuals younger than 36 months of age, and one for individuals 36 months of age or older. Use names for the levels that can be easily understood without the assistance of an additional data dictionary.