Problem Set 2

Stats 506, Fall 2018

Due: Wednesday October 17, 5pm

Instructions

- Submit the assignment by the due date via canvas. If you intend to utilize late days, please upload partial progress to Canvas and comment that you will utlize late days *before* the assignment is due. In your email please indicate how many days you intend to use. *You do not need to cc me.* **You may use a maximum of *two* late days for this assignment.**
- Use Rmarkdown to create and submit a single pdf with your answers to each question along with supporting evidence in the form of tables and graphs.
- All tables and graphs should be neatly labeled and appear polished.
- Questions 1 and 2 ask you to use Stata. Do all data manipulation and analyses in separate `.do` files named `ps2_q1.do` and `ps2_q2.do`.
- `ps2_q1.do` should write a comma delimited file `recs2015_usage.csv` with the requested point estimates and standard errors.
- Run `ps2_q2.do` in batch mode and produce a `ps2_q2.log` file with the output. Output in the log file should be clearly labeled and referred to in your typed answer to the questions.
- Question 3 asks you to analyze data in *R*. You should submit your code for this problem as `ps2_q3.R`.
- You should submit a single compressed archive (`.zip`) which contains the following files:
  - `ps2.pdf` or `ps2.html`
  - `ps2.Rmd`
  - `ps2_q1.do`
  - `recs2015_usage.csv`
  - `ps2_q2.do`
  - `ps2_q2.log`
  - `ps2_q3.R`

  All files should be executable without errors.
- All files read, sourced, or referred to within scripts should be assumed to be in the same working directory (`./`).
- Your code should be clearly written and it should be possible to assess it by reading it. Use appropriate variable names and comments. Your style will be graded using the style rubric (./StyleRubric.html) [15 points].
- Some of these exercises may require you to use commands or techniques that were not covered in class or in the course notes. You can use the web as needed to identify appropriate approaches. Part of the purpose of these exercises is for you to learn to be resourceful and self sufficient. Questions are welcome at all times, but please make an attempt to locate relevant information yourself first.
- You may wish to review:
  - the tutorial on converting between wide and long data available here (https://stats.idre.ucla.edu/stata/modules/reshaping-data-wide-to-long/),
  - Richard Williams's presentation on "Stata's Margins Command" here (https://www3.nd.edu/~rwilliam/stats/Margins01.pdf).

Question 1 [25 points]

Use Stata to estimate the following national *totals* for residential energy consumption:

- Electricity usage in kilowatt hours
- Natural gas usage, in hundreds of cubic feet
- Propane usage, in gallons
- Fuel oil or kerosene usage, in gallons

In your analysis, be sure to properly weight the individual observations. Use the replicate weights to compute standard errors. At the end of your `.do` file, write the estimates and standard errors to a delimited file `recs2015_usage.csv`.

In your `.Rmd` read `recs2015_usage.csv` and produce a nicely formatted table with estimates and 95% confidence intervals.

Question 2 [35 points]

For this question you should use the 2005-2006 NHANES ORAL Health data available here (https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Examination&CycleBeginYear=2005) and the demographic data available here (https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Demographics&CycleBeginYear=2005). Your analyses for this question should be done in Stata, though you may create plots and format tables using R within Rmarkdown.

For part (b-d), you can ignore the survey aspect of the data and analyze it as if the data were a simple random sample.

a. [5 points] Determine how to read both data sets into Stata and merge them together by the participant id SEQN.

b. [5 points] Use logistic regression to estimate the relationship between age (in months) and the probability that an individual has a primary rather than a missing or permanent upper right 2nd bicuspid. You can recode permanent root fragments as permanent and drop individuals for whom this tooth was not assessed. Use the fitted model to estimate the ages at which 25, 50, and 75% of individuals lose their primary upper right 2nd bicuspid. Round these to the nearest month. Choose a range of representative age values with one year increments by taking the floor (in years) of the 25%-ile and the ceiling (in years) of the 75%-ile.

c. [10 points] In the regression above, control for demographics in the following way:
   - Add gender to the model and retain it if it improves the BIC.
   - Create indicators for each race/ethnicity category using the largest as the reference and collapsing ‘Other Hispanic’ and ‘Other’. In order of group size in the sample, add each category retaining those that improve BIC.
   - Add poverty income ratio to the model and retain it if it improves BIC.

In your pdf document, include a nicely formatted regression table for the final model and an explanation of the model fitting process.

d. [10 points] Use the `margins` command to compute:

   1. Adjusted predctions at the mean (for other values) at each of the representative ages determined in part b.
   2. The marginal effects at the mean of any retained categorical variables at the same representative ages.
   3. The average marginal effect of any retained categorical varialbes at the representative ages.

e. [5 points] Refit your final model from part c using `svy` and comment on the differences. Include a nicely formatted regression table and cite evidence to justify your comments.

You should use the following command to set up the survey weights (ftp://ftp.cdc.gov/pub/health_statistics/nchs/tutorial/nhanes/Continuous/descriptive_mean.do):

```
svyset sdmvpsu [pweight=wtmec2yr], strata(sdmvstra) vce(linearized)
```

Question 3 [30 points]

Repeat part a-d of question 2 using **R**. For part d, you may either use the “margins” package or code the computations yourself.