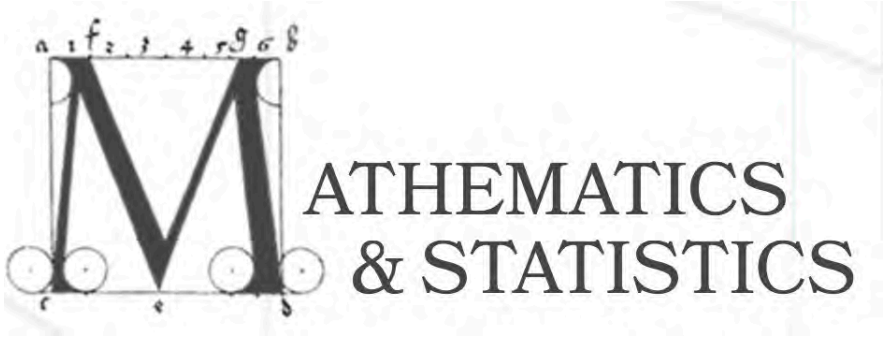
Using CE Microdata in Undergraduate Statistics Courses

Jingchen (Monika) Hu
Mathematics and Statistics Department
Vassar College

2019 Consumer Expenditure Surveys (CE) Microdata Users' Workshop



The CE sample I've used

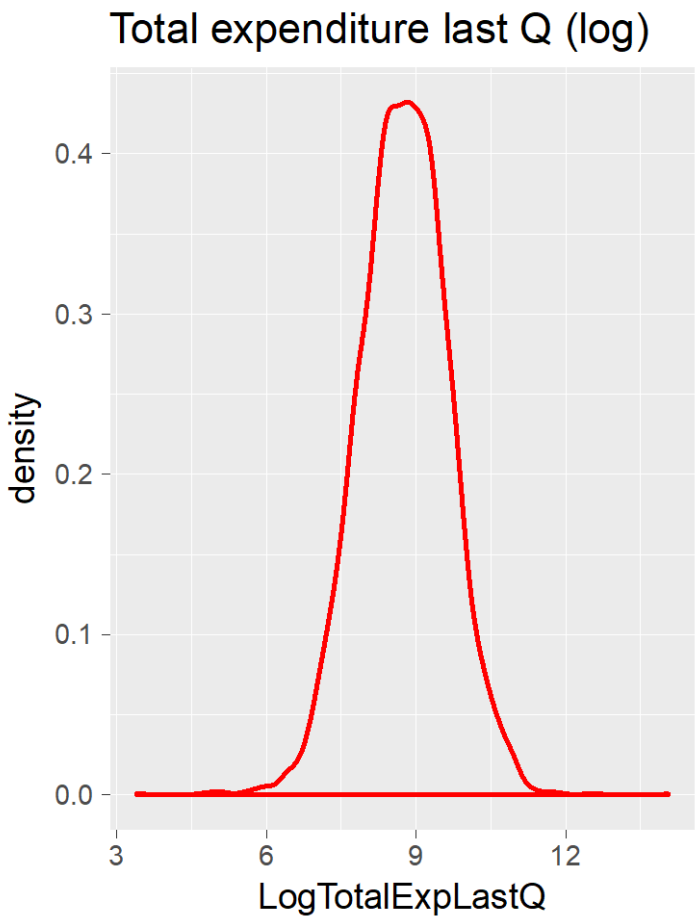
- 2017Q1, Interview Survey, FMLY data files
- Undergraduate upper-level Bayesian Statistics (6 topics)

Variable	Description	Used in topic(s)
log(Expenditure)	Continuous; CU's total expenditures in last quarter (logged)	#3 Bayesian inference for a mean; #4 Gibbs sampler and MCMC; (#5 Bayesian hierarchical modeling); #6 Bayesian linear regression
log(Income)	Continuous; the amount of CU income before taxes in past 12 months (logged)	#6 Bayesian linear regression
rural	Binary; the rural/urban status of CU	#6 Bayesian linear regression
race	Categorical; the race category of the reference person	(#5 Bayesian hierarchical modeling); #6 Bayesian linear regression

Why & how I've used the CE

- The continuous variable that I've been looking for:
  - Clear context; multiple uses to gain familiarity
  - Ideally Normally distributed (transformation)
  - Interesting by itself and with other variables

- Choosing log(Expenditure)
  - Topic #3: Bayesian inference for a mean (itself)
  - Topic #4: Gibbs sampler and MCMC (itself)
  - Topic #6: Bayesian linear regression (log(Income), rural, race)



Data access

- I downloaded fmli171x.csv from the CE website
- I created a subset with: REF_RACE, TOTEXPPQ, BLS_URBN, FINCBTAX
- I provided students with a cleaned CEdata.csv for download

Students' interaction with the CE data

- Topic #3 Bayesian inference for a mean
  - Lecture examples with R script
  - Lab 2 on Bayesian inference for unknown mean of log(Expenditure)
- Topic #4 Gibbs sampler and MCMC
  - Lecture examples with R script
- Topic #6 Bayesian linear regression
  - Lecture examples with R script (SLR with log(Income) as the predictor)
  - Lab 5 on conditional means prior in Bayesian linear regression

Example exercises

Question 1: Assess the statement "the average log total expenditure of a CU is 9 or more". Report on the comparison of the exact solution and approximation by Monte Carlo simulation.

Question 4:

- Step 1: Simulate $S = 1000$ sets of predicted values, each set contains $n = 6208$ predictions.
- Step 2: For each set, calculate the sample mean, $\bar{y}_s$.
- Step 3: Make a plot of $S = 1000$ predicted sample means $\{\bar{y}_s, s = 1, \cdots, S\}$, and compare the sample mean $\bar{y}$ in the CE data sample to the predicted $S = 1000$ sample means. Return $Prob(\bar{y} > \bar{y}_s \mid y)$ and $1 - Prob(\bar{y} > \bar{y}_s \mid y)$ and check the model fitting. Note that if either probability is small, it suggests the model does not describe the data well.

Hint: use the `rnorm` function; you can use the sample precision for `phi` (see sample R script in Section 2); similarly, you can use the sample standard deviation for σ in the prediction step in Equation 5 (sample R script: `sigma = sd(CEdata$LogTotalExpLastQ)`).

Example exercises

Question 1: Provide the hyperparameter values in Equation (7) and Equation (8) in terms the `the_data` for the conditional means prior. You can simply write down R script when specifying the list of `the_data` as your answer.

Question 2: Run the complete JAGS script and perform MCMC diagnostics.

Question 3: Interpret β_0 and β_1 in the context of the CE example.

Question 4: Use your posterior samples of $(\beta_0, \beta_1, \sigma)$ and produce predicted values of future responses at $x = 1, 5, 7, 9$ and make a plot. What can you say about predicted log(Expenditure) for a CU of \$5 log(Income)? (Hint: check out lecture slides page 36 - 39.)

Thoughts and reflection

- Why I like the CE data
  - I like how log(Expenditure) is interesting by itself and it has interesting relationships with other variables (continuous + categorical)
  - I like how rich the dataset is; I feel I can always get more variables

- Using the CE data for multiple topics helps to:
  - Give students the familiarity of the context
  - Cultivate and solidify multivariate thinking
  - Solve interesting research questions (one student's project using ACS and NHIS)

- Challenges and future ideas:
  - After log transformation, things are less intuitive and interpretable
  - Still a bit generic; find more exciting features about the CE data
  - Further explore the use of the CE data for Bayesian hierarchical modeling

Thank you very much!

Jingchen (Monika) Hu

Mathematics and Statistics Department

Vassar College

jihu@vassar.edu