Problem Set 1

Li Ruixue

January 8, 2018

1. **Classify a model from a journal**

   (a) **Find a theoretical or statistical model**

   (b) **Give a detailed citation**

   Chen, Keith. 2013. "The Effect of Language on Economic Behavior: Evident from Savings Rates, Health Behaviors, and Retirement Assets." *American Economic Review* 103(2): 690-731.

   (c) **Write the equation**

$$Pr(save_{it}) = \frac{exp(z_{it})}{1 + exp(z_{it})}$$

$$z_{it} = \beta_0 + \beta_1 StrongFTR + \beta_2 X_{it} + \beta_3 X_t + \beta_4 F_{it}^{ex} + \beta_5 F_t^c$$

   Fixed-effect logic model of an individual's propensity to save (versus not save) in the current year regressed on the FTR (future-time reference) strength of that individual's language (the theory is that if a language has stronger FTR, which means that speakers must use specific forms of the language when talking about the future such as future tense in English, they will feel that the future is further away and less important, and therefore less likely to save) and a set of fixed-effects for country and individual characteristics.

   (d) **List exogenous and endogenous variables**

   endogenous :

   $Pr(save_{it})$: probability of an individual i reporting having net saved in year t

   exogenous :

   X_{it} : characteristics of individual i at time t (country of residence, income decile within that country, marital status, sex, education, age, number of children, survey wave, religion, all interacted)

   X_t : characteristics of a country at time t

   F_{it}^{ex} : a set of fixed effects that can be taken as exogenous, such as age and sex of individual i at time t

   F_t^c : a set of continent fixed effects at time t

   (e) **Classify the model**

   (Generalized) linear model, static, deterministic.

   (f) **a variable or feature that the model is missing that might be valuable**

   This model only looks at current characteristics and saving behaviors, but past saving behaviors may also be important.

2. **Make your own model**

(a) **Write down a model of whether someone decides to get married**

I'd like to use a logistic regression model:

$$Pr(getmarried_{it}) = \frac{exp(z_{it})}{1 + exp(z_{it})}$$

$$\begin{aligned} z_{it} = &\beta_0 + age_{it} \times \beta_1 + sex_{it} \times \beta_2 + education_{it} \times \beta_3 + religion_{it} \times \beta_4 + \\ &income_{it} \times \beta_5 + family_{it} \times \beta_6 + dating_{it} \times \beta_7 + orientation_{it} \times \beta_8 \\ &+ political_{it} \times \beta_9 + language_i \times \beta_1 0 \end{aligned}$$

endogenous:
$Pr(getmarried_{it})$: probability of individual i getting married in year t
exogenous:
$age, education, income, family$: age, year of schooling, income, family environment (lower number for unhappy family and higher score for happy family) of individual i in year t, *numerical.*
$sex, religion, dating, orientation, political, language$: sex, religion, dating status, sexual orientation, political leaning, and language spoken of individual i in year t, *categorical, coded into dummy variables.*

(b) **1 = *getmarried* or 0 = *notgetmarried***

$$y_{it} = \begin{cases} 1, & \text{if } Pr(getmarried_{it}) < 0.5. \\ 0, & \text{otherwise.} \end{cases}$$

(c) **Complete data generating process?**

Yes.

(d) **Key factors that influence this outcome**

I expect all the exogenous variables to have influence on the individual's decision on getting married, especially age and education level.

(e) **Why choose those factors and not others**

I chose these factors based on my intuition and life experience, as well as on some literature I've been exposed in previous studies. For example, some studies have shown that more highly educated individuals tend to get married later, and I think factors such as language and religion can be used as proxies for culture, which will likely affect when people get married.

(f) **How to do a preliminary test**

Since the test is preliminary, I think I'll acquire some data by conducting a small to medium-size survey or extracting relevant information from past surveys, such as national censuses, or general social survey. In a survey, I'll gather information on the variables I specified in the model, as well as whether the individual decides to get married in that year, whereas if I'm getting the information from past surveys, I'll look at individuals' demographic data as needed in the model as well as whether they're married or the year they got married. I can then run the logit regression and see whether some of the variables are significant. I can also divide my dataset into training and test subsets, estimate the coefficients with the training dataset, and predict the outcome for the test dataset to check the accuracy of the prediction.