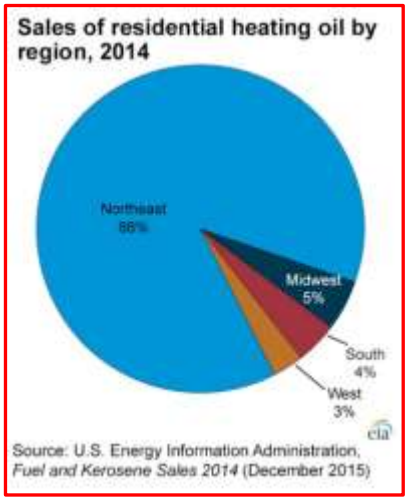
Agricultural and Applied Economics 637
Applied Econometrics II

Assignment V
Estimation of Discrete Choice Models (#1)
(Due: April 20, 2018)

Total Points: 90

For this assignment I would like you to use the lessons learned in Assignment IV to the parameter estimation of a couple of binary discrete choice (DC) models of energy use decisions. Also this assignment will give you some experience on interpreting results obtained from such estimation.

1. (**65 pts**) This file contains a subset of the 2009 RECS data. A mini-version of the 2009 codebook for this data set is here: Assign #5 Codebook.
   a. **(25 pts)** The use of fuel oil for space heating is one of the more inefficient methods (on a BTU basis) to heat a home. A majority of sales of heating oil occurs in the Northeast due to both climatic conditions and the lack of alternative heating fuel sources.[1] Assume the U.S. Department of Energy would like to identify determinants of the use of alternative energy sources for providing household space heating. As a first step in your analysis, you would like to focus on the use of fuel oil. As noted by Lia (2013), in the short run, the type of fuel to use in space heating is already determined. As such, instead of motivating the binary DC model via the difference in latent random utility functions, we will estimate the binary DC as a simple statistical model for predicting the probability that a particular household will use fuel oil in space heating.



   You decide that for this analysis you would like to use the subset of respondents that:

[1] In this figure, the Northeast region is composed of Connecticut, Maine, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island and Vermont.

i. Live in a *Mobile* or *Single-Family* home (i.e., TYPEHUQ = 1, 2, or 3); and
ii. Actually need to heat their home (i.e., HEATHOME=1).

With this subset of data, you would like to estimate the following DC model:

(FUELHEAT=3?) = f(Intercept, NEWENG, MA, ENC, WNC, PACIFIC, MOUNT,METRO, MICRO, HDD30_2, CDD30_2, HDDCDD, HOUSEAGE, OWN, MOBILE)

Where the following are user created variables:

$$\text{NEWENG} \equiv \begin{cases} 1 \text{ if DIVISION} = 1 \\ 0 \text{ Otherwise} \end{cases}; \quad \text{MA} \equiv \begin{cases} 1 \text{ if DIVISION} = 2 \\ 0 \text{ Otherwise} \end{cases};$$

$$\text{ENC} \equiv \begin{cases} 1 \text{ if DIVISION} = 3 \\ 0 \text{ Otherwise} \end{cases}; \text{WNC} \equiv \begin{cases} 1 \text{ if DIVISION} = 4 \\ 0 \text{ Otherwise} \end{cases};$$

$$\text{MOUNT} \equiv \begin{cases} 1 \text{ if DIVISION} = 8 \text{ or } 9 \\ 0 \text{ Otherwise} \end{cases}; \text{PACIFIC} \equiv \begin{cases} 1 \text{ if DIVISION} = 10 \\ 0 \text{ Otherwise} \end{cases}$$

METRO ≡ Home located in Census Metropolitan area[2]

$$= \begin{cases} 1 \text{ if METMIC=1} \\ 0, \text{ otherwise} \end{cases}$$

MICRO ≡ Home located in Census Micropolitan area

$$= \begin{cases} 1 \text{ if METMIC=2} \\ 0, \text{ otherwise} \end{cases}$$

HDD302 ≡ HDD30YR/1000

CDD302 ≡ CDD30YR/1000

HDDCDD ≡ HDD302×CDD302

HOUSEAGE ≡ Age of home with 2009=0

OWN ≡ identifies whether someone in the household owns the residence:

$$= \begin{cases} 1 \text{ if KOWNRENT} = 1 \\ 0 \text{ Otherwise} \end{cases}$$

MOBILE ≡ identifies whether the respondent lives in a mobile home:

$$= \begin{cases} 1 \text{ if TYPEHUQ} = 1 \\ 0 \text{ Otherwise} \end{cases}$$

You can either create the final dataset using EXCEL, MATLAB or whatever method you find the easiest. Personally I used Excel for data management and selection given that this is a relatively small dataset that is easy to manage.

[2] According to the U.S. Census, Metropolitan Statistical Areas must have at least one urbanized area of 50,000 or more inhabitants. Each Micropolitan Statistical Area must have at least one urban cluster of at least 10,000 but less than 50,000 population.

Estimate the parameters of the Probit model shown above using your own MATLAB code. Feel free to use the probit estimation code I distributed in class or the generic maximum likelihood estimation procedures we reviewed earlier. Present estimated coefficient values and associated standard errors. Make sure you show the results of the statistical test of the hypothesis that *the above exogenous variables, as a group, explain a significant portion of the observed pattern of fuel oil use for heating purposes vs. nonuse*.

b. (**20 pts**) Using the model results obtained from the above Probit:
   i. Statistically test the null hypothesis that survey respondent geographic location has no impact on the probability of using fuel oil for heating purposes.
   ii. Using all sample data, what is the ***average*** *discrete change impact* of living in a mobile home on the probability of using fuel oil for space heating purposes?[3] What is the Z-value under the null hypothesis that this ***average*** *discrete change* impact is 0?
   iii. Using all the sample data, test the null hypothesis that the ***average*** *marginal impact* of the 30-yr Heating Degree Day (i.e., HDD302) variable value shows no impact on fuel use probability.
   iv. Test the null hypothesis that there is a statistically significant HDD302 ***average*** *interaction effect*.

c. **(10 pts)** Using the *entire sample*, what is the ***average*** *HDD elasticity* on the probability of heating with fuel oil? Is this elasticity statistically different from 0? [***Hint***: Make sure in your elasticity calculation you use the predicted probability of fuel oil use, conditional on a particular set of explanatory variable values.]

d. **(10 pts)** Evaluate the Household age elasticity using the two methods we have talked about in class: (i) at the *mean of the data* and (ii) estimating the *average elasticity* over all observations. Are these elasticities ***statistically different*** from one another? Explain how you go about answering this question. [***Hint***: Make sure in your elasticity calculation you use the predicted probability of fuel oil use, conditional on a particular set of explanatory variable values.]

2. **(25 pts)** Upon examining the results obtained from the above Probit model you believe that the underlying probability model possesses a heteroscedastic error term. As such, you are concerned about obtaining consistent parameter estimates and correct parameter standard errors. You therefore would like to re-estimate the above Probit model but this time incorporating an heteroscedastic error structure. You assume that the latent error, ε_t, has the following variance specification:

[3] Note that when I use the term **average**, I am referring to you evaluating the statistic across all observations that is then averaged.

$$\sigma_t^2 = \left[\exp\left(Z_t\gamma\right)\right]^2$$

where σ^2_t is the error variance for the t^{th} observation, Z_t is a (T×S) matrix of exogenous variables and γ is an (S×1) vector of error variance coefficients which need to be estimated. [***Hint:*** *Note the squared term in the variance specification*. *Greene, page 714-715 reviews this form of heteroscedastic discrete choice model specification.*]

For this analysis you assume that the Z matrix is composed of the following exogenous variables: INCOME, HDD302, HOUSEAGE, NEWENG where INCOME is annual household income defined in $100,000 units (i.e., to minimize possible scaling problems). You may want to divide HOUSEAGE by 10 to reduce any scaling problems.

a. **(10 pts)** Estimate the heteroscedastic Probit model using the X-matrix accessing in the homoscedastic specification, the assumed error variance functional form shown above and the corresponding Z matrix. Present the typical regression-based statistics with respect to the probit and error variance coefficients. Undertake a likelihood ratio test of whether you have homoscedastic versus heteroscedastic errors. [***Hint:*** *As I noted in class,* ***do not include an intercept term*** *in the error variance expression. You cannot use the standard Probit estimation code to obtain parameter estimates. Use a generic maximum likelihood estimation system to obtain parameter estimates.*]

b. **(15 pts)** Based on the heteroscedastic model results and all of your data, what is the *average HDD302 elasticity* on the probability of a household using fuel oil for space heating purposes? [***Hint:*** *Make sure in your elasticity calculation you use the predicted probability of fuel oil use, conditional on a particular set of explanatory variable values.*]