ECNS 562 –Econometrics II
Eric Belasco
Homework 2
Due Tuesday, February 21

1. Download the dataset labeled "bfrss2010" in order to address the following questions. The following command can be used to download the data.

bfrss2010 = read.table("http://www.montana.edu/ebelasco/ecns562/homework/bfrss2010.dat", header = T)

The data consists of survey responses as part of the CDC's Behavioral Risk Factor Surveillance System (BFRSS) program in 2010. One focus area of this survey is regarding health outcomes. With regard to preventing colorectal cancer, the use of colonoscopy and/or sigmoidoscopy has been recommended for individuals over the age of 50 as a means of preventing colorectal cancer. However, a significant proportion of the population does not comply with this treatment. In order to help examine this issue the following model is proposed:

$$\Pr(Colon2 = 1|X) = f(X = HLTHPLAN, RACE, BMI4, EDUC, INCOME, AGE, RURAL)$$

where *Colon2* is equal to 1 if the individual has had a sigmoidoscopy or colonoscopy; *HLTHPLAN* is a binary variable indicating whether the individual has health care insurance; *RACE* is broken into 4 binary variables (white, black, AI (American Indian), and Asian), *BMI4* is the individual's body mass index based on reported height and weight; *EDUC* is broken into 3 binary variables indicating the highest degree achieved (CollegeGrad, HSGrad, LessThanHS); INCOME is broken into three brackets that include LowIncome (<25,000), MidIncome (25,000 – 75,000), and HighIncome (>75,000) in annual household salaries; *AGE*; and *RURAL* is a binary variable indicating residence in a rural setting.

(a) Use your own function (*this means don't use the lm() function*) to produce OLS estimates for the above regression. Report parameter estimates, standard errors, t-values, and an r-squared measure associated with the linear probability model (LPM) *(Hint: For this question you will need to augment the existing function from the lab in order to compute t-values and r-squared)*.
(b) Discuss the shortcomings of using a LPM model for estimation and the potential benefits from using a nonlinear model such as the logit or probit model.
(c) Produce another function to estimate a logistic (logit) regression model.
(d) Using your newly produced function, estimate the given data using a logistic regression model. Report the same values as reported in part (a). In addition, report the log-likelihood value. (*Hints: Since you cannot compute r-squared with the logit model, compute McFadden's LRI. Also, you will need to use the negative log-likelihood function for optimization, meaning the actual log-likelihood value will be the negative of* ***LL****.*)

(e) Interpret parameter estimates from part (d), keeping in mind that the logit model is non-linear. (*Hint: evaluate the average marginal effect on the population.*)
(f) Given the results in part (e), what efforts might you propose in order to increase the use of colonoscopy and sigmoidoscopy to prevent colorectal cancer.

2. To begin this question please use the R dataset titled '**Travel.Rdat**'. The downloaded file consists of data concerning travel choices made by 210 individuals. Each data frame consists of the one of the variables listed below and includes information regarding individual specific covariates. For each nxJ matrix, the column vectors are ordered Air, Train, Bus, and Car from left to right. (*e.g., this means that the second column corresponds to Train*). In this question, we are interested in finding the impact that different alternative-specific attributes have on travel mode decisions. Independent variables include a generalized cost measure (GC), the terminal waiting time (Ttme), and constants for Air, Train, and Bus. The data contains 4 choice alternatives (travel modes) and 210 observations. The variables are defined in the R file supplied.

(a) Construct your own function that performs maximum likelihood estimation using a conditional logit model. It is recommended that you code this function using three steps. First, compute expected utility associated with each individual choosing each alternative. Second, calculate the probability that individual *i* chooses alternative *j* for each possible combination. Third, compute the negative log-likelihood value.

(b) Estimate a choice model including GC, Ttme, and constants as explanatory variables. Report estimates, standard errors, and t-statistics for this model.

(c) Interpret the impact from an increase on Ttme for Air travel on all modes of transportation. *(Hint: for this question you will need to compute the matrix P given optimal parameter estimates)*

(d) Examine the impact in the predicted outcomes that would result if the generalized cost of driving increased by 25%. What percentage of individuals are predicted to drive after the 25% increase? What about before the increase?

(e) Compute the loss to welfare (consumer surplus) from the increase in increased driving costs. Does this impact (in per person units) seem reasonable? *(Hint: The parameter estimate on GC can be used to normalize changes to expected utility)*

(f) Due to faulty FAA communication technology, the local airport is closed. Discuss the IIA assumption and how it impacts the predicted choices that would be made after one of the transportation modes are no longer available.