Problem Set 6

The purpose of the first part of this problem set is to estimate, interpret the results, and compare the results across different binary dependent variable models. In the second part you will estimate and compare different specifications of an endogenous selection model.

First part will be discussed in week 9 and the second part in week 10 of this term.

The data file for this exercise is on Moodle: `mus16data.dta`. It is a subset of the data used by P. Deb, M. Munkin and P.K. Trivedi (2006): "Bayesian Analysis of Two-Part Model with Endogeneity", Journal of Applied Econometrics, 21, 1081-1100. Data is for 2001 and comes from the Medical Expenditure Survey. Sample has 3,328 observations.

The main outcome variable of interest is ambulatory expenditure (`ambexp`) and the regressors are given below.

Since the expenditure data is skewed, we will be using the logged expenditure variable as our dependent variable. You should read Cameron A.C. and Trivedi, P.K. Micro-econometrics using Stata to see the pros and cons regarding whether to log the dependent variable or not.

Note, there is one individual who has an `expenditure`=1 and this will get coded as 0 when variable is logged. Since it is only one individual, we will ignore the problem by not doing anything. If there are many individuals like this, you will need to see whether you can say why this might be the case.

Dependent variable

- `ambexp`: Ambulatory medical expenditures (excluding dental and outpatient mental). There are 526 individuals with zero expenditure. There is one individual who has expenditure=$1. I am going to assume that this individual did not spend any money.
- `lambexp`: ln(`ambexp`) given `ambexp` > 0 ; missing otherwise
- `dambexp`: 1 if `ambexp` > 0 and 0 otherwise (binary indicator)

Regressors

- `ins`: health insurance measures, either PPO or HMO type insurance
- `totchr`: health status measures: number of chronic diseases
- `age`: age age in years/10
- `female`: 1 for females, zero otherwise
- `educ`: years of schooling of decision maker
- `blhisp`: either black or Hispanic
- `income`: income in USD/1000

Preamble

```
<IPython.core.display.HTML object>
```

Create a do-file for this problem set and include a preamble that sets the directory and opens the data. For example,

```
clear
//or, to remove all stored values (including macros, matrices, scalars, etc.)
*clear all

* Replace $rootdir with the relevant path to on your local harddrive.
cd "$rootdir/problem-sets/ps-6"

cap log close
log using problem-set-6-log.txt, replace

use mus16data.dta, clear
```

Questions

Part 1

1.1. Obtain and comment on the descriptive statistics for ambexp, lambexp, age, female, educ, blhisp, totchr, ins, income.

1.2. Estimate a LP, Probit and a Logit model to explain dambexp. Store the β coefficients and report them in a table.

1.3. Estimate the Marginal Effect at the Mean for each model, and report them in a table. You will want to use the `estpost margins` post-estimation command, with the relevant option for MEM. Pay special attention to the treatment of discrete regressors. Hint: check to see any differences in the estimated MEs based on whether you use factor notation; for example, `i.female` vs `female`.

1.4. Estimate the Average Marginal Effect at the Mean for each model, and report them in a table.You will want to use the `estpost margins` post-estimation command, with the relevant option for AME.

1.5. Check to see how well the prodbit model predicts the outcome using the `estat classification` post-estimation command.

1.6. Construct and interpret the LR test for the omission of income in the probit model. Do this in two ways: (1) using the post estimation `lrtest`; (2) manually recreate (1)'s results (both test-statistic and p-value).

Part 2

Estimate the following models for `lambexp` treating the selection into non-zero `lambexp` value as endogenous using, both Heckman 2-step method and also MLE.

In the main data `lambexp` is missing for values of `ambexp`=0. Before proceeding,

```
replace lambexp = 0 if ambexp==0
```

This will correction will also treat observations with `ambexp=1` as equivalent to =0; however, this is only a single observation.

2.1. Estimate the Heckman 2-step estimator and store the results. In addition, store the Mills ratio as a separate variable. Use `income` as the excluded variable. This means that `income` appears in the selection equation, but NOT the main equation.

2.2. Replicate these results by applying the following steps: (1) estimate the selection equation using a probit model; (2) create the mills ratio; (3) `compare` your mills ratio with the one stored above; (4) estimate the main equation, including the mills ratio.

2.3 Estimate the marginal effects of the selection equation. You can do this using the `estpost margins` command. This should correspond to a probit model estimation above.

2.4. Estimate the Maximum Likelihood version of the Heckmann correction (with an excluded variable) and store the results.

2.5. Now re-estimate the two-step and MLE approach without an excluded variable, storing the results each time. This means that the same set of regressors enter both equations. i.e. include income in the outcome equation.

2.6. Create a table that reports the four models alongside one another and compare the results.

Postamble

```
log close
```