Censoring, Truncation and Selection Models

1. The following data set

   ```
   .  use http://fmwww.bc.edu/ec-p/data/mus/mus18data, clear
   ```

   has information on health care use from

   Deb, P. and P.K. Trivedi (2002). The Structure of Demand for Medical Care: Latent Class versus Two-Part Models", Journal of Health Economics, 21, 601--625.

   We will consider study year 2 and everybody who is 18 years or older.

   (a) Estimate a regression of medical expenditures (med) on age (age), sex (female), family size (num), log income (linc), education of decision maker (educdec), site dummies (site), and the co-insurance rate (coins, the percentage of health cost borne by the insured paid by the patient) how do you interpret the coefficient on coins?

   (b) About 20 percent of the individuals in the data have zero medical expenditures. Estimate a censored regression model using maximum likelihood (i.e. -tobit- in Stata). Use the resulting parameter estimates to estimate the average probability of having positive medical expenditures when coins is 0 and when coins is 100.

   (c) Calculate the average partial effect of increasing coins from 0 to 100 on observed expenditures and compare your answer to (a). How does it compare to predicting levels when coins equals 0 and 100? (Note that $E(y_i) = E(y_i^*|y_i^* > 0)P(y_i^* > 0)$. Also you can obtain the estimate of σ through _b[/:var(e.med)] in Stata.)

   (d) Estimate a truncated regression (ie only using positive expenditures) using maximum likelihood.

   (e) Can you estimate a truncated regression (ie only using positive expenditures) using the control function method?

   (f) Now estimate the same regression using the Heckman selection model and interpret the coefficients.

   (g) Use your estimation results from (f) to calculate the same effects as in (c). What do you conclude?

(h) Is normality a suitable assumption for the residual in your outcome equation? Explain.

(i) Estimate a linear regression of the logarithm of medical expenditures (log(med)) on age (age), how do you interpret the coefficient on age?

(j) Now estimate the same regression as in (i) using the Heckman selection model and interpret the coefficients.

(k) Use your estimation results from (j) to calculate the same effects as in (c) and compare your results to (g). What do you conclude?