Problem Set 7 - Week 50, 2013

Problem 3

We use the dataset RandHIE from package *sampleSelection* coming from the RAND Health Insurance Experiment (RHIE). For more details read the R and Wikipedia help on the experiement. The data extract comes from Deb and Trivedi (2002), who modeled the number of outpatient visits to a medical doctor and to all providers using count data models.

Here instead we model annual health expenditures. The regressors are can be broken down into health insurance variables (logc, idp, lpi, and fmde), socioeconomic characteristics (linc, lfam, xage, female, child, fchild, black and educdec) and health status variables (physlm, disea, hlthg, hlthf and hlthp). The analysis is using only the second year of data.

The dependent variable y is annual individual health expenditures (*meddol*). We are especially interested in the effect of coinsurance rate *logc* on the individual expenditure (http://en.wikipedia.org/wiki/Co-insurance). An econometric model needs to take account of two complications: (1) Health expenditures are zero for 23.2% of the sample and (2) the positive health expenditures are very right-skewed with a mean of $221 that is much larger than the median of 53$. The logarithmic transformation eliminates this skewness, with a mean of 4.07 close to the median of 3.96 and the skewness statistic falls from 24.0 to 0.3. The kurtosis is 3.29, close to the normal value of 3.

We focus on modeling $\ln y$ for those with positive medical expenditures. We model the data with a Tobit II model where the selection is given by y_2 is the indicator of positive expenditure (*binexp*), and y_1 is *lnmeddol*. Note that it is not meaningful to consider the value of y_1 when $y_2 = 0$. In that case the annual individual helth expenditure is 0 with no defined logarithm.

1. Explain why we believe there might be behavioural selection in this data set.

2. Read the data in your memory and understand what each variable mean

```
> library(sampleSelection)
> data( RandHIE )
> ?RandHIE
```

3. Create a subsample using only the variables from year two and make sure that your subsample does not have NA in variable *educdec*

```
> subsample<- RandHIE$year == 2 & !is.na( RandHIE$educdec )
> RandHIE.sub<-data.frame(RandHIE[subsample, ])
```

4. Assumme $X = X_1$ and run Procedure 19.1

```
> selectEq <- binexp ~ logc + idp + lpi + fmde + physlm + disea +
+    hlthg + hlthf + hlthp + linc + lfam + educdec + xage + female +
+    child + fchild + black
> outcomeEq <- lnmeddol ~ logc + idp + lpi + fmde + physlm + disea +
+    hlthg + hlthf + hlthp + linc + lfam + educdec + xage + female +
+    child + fchild + black
> m.1<-heckit(selectEq, outcomeEq, method="2step", data=RandHIE.sub)
> summary(m.1)

--------------------------------------------
Tobit 2 model (sample selection model)
2-step Heckman / heckit estimation
5574 observations (1293 censored and 4281 observed)
39 free parameters (df = 5536)
Probit selection equation:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.2716050  0.1877345  -1.447 0.148023
logc        -0.1187080  0.0269005  -4.413 1.04e-05 ***
idp         -0.1279483  0.0522351  -2.449 0.014338 *
lpi          0.0283091  0.0088793   3.188 0.001439 **
fmde         0.0075319  0.0161584   0.466 0.641142
physlm       0.2732013  0.0743761   3.673 0.000242 ***
disea        0.0224861  0.0035958   6.253 4.32e-10 ***
hlthg        0.0387516  0.0438545   0.884 0.376929
hlthf        0.1920062  0.0836688   2.295 0.021780 *
hlthp        0.6397294  0.2126322   3.009 0.002636 **
linc         0.0518413  0.0168128   3.083 0.002056 **
lfam        -0.0335599  0.0417280  -0.804 0.421284
educdec      0.0363070  0.0076536   4.744 2.15e-06 ***
xage         0.0002631  0.0021606   0.122 0.903070
female       0.4451035  0.0542920   8.198 3.00e-16 ***
child        0.1114890  0.0808338   1.379 0.167877
fchild      -0.4512845  0.0799219  -5.647 1.72e-08 ***
black       -0.6057367  0.0523148 -11.579  < 2e-16 ***
Outcome equation:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.882514   0.469897   6.134 9.14e-10 ***
logc        -0.027921   0.039754  -0.702 0.482495
idp         -0.092290   0.068019  -1.357 0.174893
lpi          0.005223   0.011106   0.470 0.638190
fmde        -0.029521   0.018243  -1.618 0.105667
physlm       0.281495   0.080454   3.499 0.000471 ***
disea        0.021617   0.005040   4.290 1.82e-05 ***
hlthg        0.147403   0.049050   3.005 0.002666 **
hlthf        0.382168   0.096128   3.976 7.11e-05 ***
hlthp        0.833294   0.197449   4.220 2.48e-05 ***
```

```
linc          0.099097   0.025155   3.939 8.27e-05 ***
lfam         -0.144136   0.046807  -3.079 0.002085 **
educdec       0.003364   0.010950   0.307 0.758700
xage          0.005556   0.002255   2.464 0.013778 *
female        0.384632   0.103280   3.724 0.000198 ***
child        -0.256514   0.093677  -2.738 0.006196 **
fchild       -0.392146   0.125089  -3.135 0.001728 **
black        -0.263365   0.157754  -1.669 0.095082 .
Multiple R-Squared:0.1367,        Adjusted R-Squared:0.133
Error terms:
              Estimate Std. Error t value Pr(>|t|)
invMillsRatio   0.2358     0.5018    0.47    0.638
sigma           1.4008         NA      NA       NA
rho             0.1683         NA      NA       NA
--------------------------------------------
```

5. Estimate the model with Maximum likelihood

```
> m.2 <- selection( selectEq, outcomeEq, data =RandHIE.sub )
> summary(m.2)

--------------------------------------------
Tobit 2 model (sample selection model)
Maximum Likelihood estimation
Newton-Raphson maximisation, 6 iterations
Return code 1: gradient close to zero
Log-Likelihood: -10170.11
5574 observations (1293 censored and 4281 observed)
38 free parameters (df = 5536)
Probit selection equation:
              Estimate Std. error t value  Pr(> t)
(Intercept) -0.2141574  0.1842169  -1.163 0.245021
logc        -0.1068027  0.0264766  -4.034 5.49e-05 ***
idp         -0.1087690  0.0509938  -2.133 0.032926 *
lpi          0.0294804  0.0086214   3.419 0.000628 ***
fmde         0.0007403  0.0158738   0.047 0.962803
physlm       0.2848256  0.0722656   3.941 8.10e-05 ***
disea        0.0210805  0.0034967   6.029 1.65e-09 ***
hlthg        0.0576901  0.0427990   1.348 0.177681
hlthf        0.2237238  0.0814547   2.747 0.006022 **
hlthp        0.7984291  0.2048087   3.898 9.68e-05 ***
linc         0.0553122  0.0166179   3.328 0.000873 ***
lfam        -0.0312010  0.0402985  -0.774 0.438785
educdec      0.0314990  0.0074987   4.201 2.66e-05 ***
xage        -0.0006072  0.0021064  -0.288 0.773128
female       0.4093059  0.0532548   7.686 1.52e-14 ***
child        0.0530643  0.0786326   0.675 0.499778
```

```
fchild      -0.3953421  0.0783811  -5.044 4.56e-07 ***
black       -0.5831049  0.0520534 -11.202  < 2e-16 ***
Outcome equation:
             Estimate Std. error t value  Pr(> t)
(Intercept)  2.107745   0.244228   8.630  < 2e-16 ***
logc        -0.076024   0.033746  -2.253  0.02427 *
idp         -0.149720   0.066138  -2.264  0.02359 *
lpi          0.014930   0.010502   1.422  0.15511
fmde        -0.023522   0.019475  -1.208  0.22711
physlm       0.354863   0.075542   4.698 2.63e-06 ***
disea        0.028647   0.003797   7.544 4.54e-14 ***
hlthg        0.155917   0.052177   2.988  0.00281 **
hlthf        0.445122   0.095526   4.660 3.17e-06 ***
hlthp        0.998606   0.187879   5.315 1.07e-07 ***
linc         0.121401   0.023085   5.259 1.45e-07 ***
lfam        -0.158302   0.049746  -3.182  0.00146 **
educdec      0.017595   0.009018   1.951  0.05105 .
xage         0.005738   0.002443   2.349  0.01883 *
female       0.550344   0.063331   8.690  < 2e-16 ***
child       -0.197688   0.097398  -2.030  0.04239 *
fchild      -0.565323   0.097529  -5.796 6.77e-09 ***
black       -0.535868   0.074919  -7.153 8.51e-13 ***
Error terms:
      Estimate Std. error t value Pr(> t)
sigma  1.57005    0.02783   56.42  <2e-16 ***
rho    0.73560    0.03379   21.77  <2e-16 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1
-----------------------------------------
```

6. Interpret your results

   The inverse Mills ratio term is statistically insignificant and low in magnitude with implied $\rho = 0.168$ that is close to zero. This estimator can perform poorly if the inverse Mills ratio term is highly correlated with the other regressors. Here this does not appear to be the case as there is considerable range in the probit model predicted probabilities from 0.15 to 0.99.

   The ML estimates differ considerably from the previous estimates. The errors in the latent variable model are highly correlated with estimate $\rho = 0.736$ that is highly statistically significant. The big difference between the two-step estimates and the ML estimates of σ_{12} or of ρ is best viewed as a problem with the sample selection model. This could be a problem with assumption of $\nu_2 \sim N(0, 1)$. Such fragiality of this kind of sample selection model is not rare, espeically if the same regressors are being used in both parts of the model. In additon, health expenditure data could have

large outliers so that the errors are not normal.

The regressor of most interest is *logc*, the natural logarithm of the coinsurance rate where the coinsurance rate equals the percentage of health cost borne by the insured paid by the patient. The most statistically significant effect is in determining whether or not expenditures are positive, rather than on the size of positive expenditures. If all observations were positive then the coefficient of *logc* in regression on *lonmeddol* equals the price elasticity of demand for health care.