ECMA 31130 HW2

Yijing Zhang & Jeffrey Wang

Question 1

The parameter a of the model captures whether or not there is correlation between R_i and β_i. Intuitively, from the specification of $\log R_i$, ξ_i represents all the unobserved variables that determine the individual i's non-labour income, and if $a \neq 0$, then these unobserved characteristics also enter into determining the individual's disutility from labour.

```
p  = list(gamma = 0.8,beta=1,a=1,rho=1,eta=0.2,delta=-0.2,delta0=-0.1,nu=0.5) # parameters
N=10000  # size of the simulation
set.seed(123456)

simdata = data.table(i=1:N,X=rnorm(N))

# simulating variables
simdata[,X := rnorm(N)]
simdata[,Z := rnorm(N)]
simdata[,u := rnorm(N)]
simdata[,lw := p$eta*X  + Z + 0.2*u ]  # log wage

simdata[,xi := rnorm(N)*0.2]
simdata[,lr := lw + p$delta0+ p$delta*Z + xi]; # log home productivity

simdata[,eps:=rnorm(N)*0.2]
simdata[,beta := exp(p$nu*X  + p$a*xi + eps)]; # heterogenous beta coefficient

# compute decision variables
simdata[, lfp := log(p$rho) + lw >= lr] # labor force participation
simdata[, h   := (p$rho * exp(lw)/beta)^(1/p$gamma)] # hours
simdata[lfp==FALSE,h:=NA][lfp==FALSE,lw:=NA]
simdata[,mean(lfp)]

simdata[, lh := log(h)]
```

Question 2

For a = 0:

```
pander(summary(lm(lh ~ lw + X, simdata)))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.001408	0.003429	-0.4106	0.6814
lw	1.246	0.003594	346.7	0
X	-0.6268	0.003202	-195.8	0

Observations	Residual Std. Error	R^2	Adjusted R^2
6408	0.2487	0.9545	0.9545

For a = 1:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.1415	0.004356	32.49	1.851e-214
lw	1.15	0.004565	251.9	0
X	-0.6111	0.004068	-150.2	0

Observations	Residual Std. Error	R^2	Adjusted R^2
6408	0.316	0.9187	0.9187

We observe that the coefficients for log wage and X are closer to the true values when we set $a = 0$ as opposed to $a = 1$. Because if $a \neq 0$, then the term $a\xi_i$ becomes an omitted variable that enters the regression of log hours on log wage and X, which will absorb its effect on the individual's hours decision from the coefficients of log wage and X. In other words, if an individual's disutility of labour becomes correlated with her non-labour income, then the coefficients in front of log wage and X are biased, since we no longer have the exogeneity condition on the error term of this regression.

Question 3

We begin by deriving the FOC and its log linear form when $e = 1$.

Call $\tilde{h}^*$ and $\tilde{w}^*$ the log observed observed hours and consumption when $lfp = 1$, we write out our log linear form and replace β with our specification:

$$\tilde{h}^* = \frac{1}{\gamma}\log\rho + \frac{1}{\gamma}\tilde{w}^* - \frac{1}{\gamma}(\nu x_i + \epsilon_i + a\xi_i)$$

$$\tilde{h}^* = \frac{1}{\gamma}\tilde{w}^* - \frac{\nu}{\gamma}x_i - \frac{\epsilon_i}{\gamma} - \frac{a\xi_i}{\gamma} + \frac{1}{\gamma}\log\rho$$

Now we take the conditional expectation of hours based on everything we observe

$$E[\tilde{h}^*|\tilde{w}, x_i, z_i, lfp = 1] = \frac{1}{\gamma}\tilde{w} - \frac{\nu}{\gamma}x_i - \frac{1}{\gamma}E[\epsilon_i + a\xi_i|\tilde{w}, x_i, z_i, lfp = 1] + \frac{1}{\gamma}\log\rho$$

We now take a better look at our heckman correction term. Since ϵ_i is independent and mean zero, ϵ_i is independent and mean zero,

$$E[\epsilon_i|\tilde{w}, x_i, z_i, lfp = 1] = 0$$

$$E[a\xi_i|lfp = 1] = aE[\xi_i|\log w_i + \log\rho > \log R] = aE[\xi_i|\log\rho > \delta_0 + \delta z_i + \xi_i] = aE[\xi_i|\xi_i < \log\rho - \delta_0 - \delta z_i]$$

Now let's take $\rho = 1$ to make $\log\rho = 0$, we express the inverse mills ratio:

$$\lambda_i = E[\xi_i|\xi_i < -\delta_0 - \delta z_i] = -\frac{\sigma_{\xi_i}\Phi'(\frac{-\delta_0-\delta z_i}{\sigma_{\xi_i}})}{\Phi(\frac{-\delta_0-\delta z_i}{\sigma_{\xi_i}})}$$

Since $\rho = 0$, we do a probit of lfp on z_i to get δ_0 and δ.

```
fit2 = glm(lfp ~ Z,simdata,family = binomial(link = "probit"))
pander(summary(fit2))
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.5161	0.0155	33.29	5.645e-243
Z	0.9976	0.01973	50.56	0

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	13059 on 9999 degrees of freedom
Residual deviance:	9390 on 9998 degrees of freedom

Question 4

#construct the inverse Mills ratio.

```
simdata[, ai := predict(fit2)]
simdata[, m := dnorm(ai)/pnorm(ai)]

fit3 = summary(lm(lh ~ lw + X + m,simdata))
pander(fit3)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.02577	0.01658	-1.554	0.1202
lw	1.267	0.01117	113.4	0
X	-0.6266	0.004666	-134.3	0
m	0.293	0.02836	10.33	8.009e-25

Table 8: Fitting linear model: lh ~ lw + X + m

Observations	Residual Std. Error	R^2	Adjusted R^2
6405	0.3167	0.921	0.921

Question 5

From Q3 we have

$$E[\xi_i|\xi_i < -\delta_0 - \delta z_i]$$

And since we have $\xi_i \sim -exp$, we can write

$$E[\xi_i|\xi_i < -\delta_0 - \delta z_i] = 1 - \delta_0 - \delta z_i$$

Question 6

```
simulate <- function(p,N,t) {
    set.seed(123456)
    simdata = data.table(i=1:N,X=rnorm(N))

    # simulating variables
    simdata[,X := rnorm(N)]
    simdata[,Z := rnorm(N)]
    simdata[,u := rnorm(N)]
    simdata[,lw := p$eta*X + Z + 0.2*u ]  # log wage

    simdata[,xi := -rexp(N)]
    simdata[,lr := lw + p$delta0 + p$delta*Z + xi]; # log home productivity

    simdata[,eps:=rnorm(N)*0.2]
    simdata[,beta := exp(p$nu*X  + p$a*xi +  eps)]; # heterogenous beta coefficient

    # compute decision variables
    simdata[, lfp := log(p$rho) + lw >= lr] # labor force participation
    simdata[, h   := (p$rho * exp(lw)/beta)^(1/p$gamma)] # hours

    # make hours and wages unobserved in case individual doesn't work
    simdata[lfp==FALSE,h:=NA][lfp==FALSE,lw:=NA]
    simdata[,mean(lfp)]

    # store time
    simdata[, t := t]
    return(simdata)
  }

p  = list(gamma = 0.8,beta=1,a=1,rho=1,eta=0.2,delta=-1.0,delta0=-0.1,nu=0.5) # parameters
N=50000  # size of the simulation

# simulate period 1
sim1  = simulate(p,N,1);
p$eta = 0.4; p$delta0 = -1.1
# simulate period 2 with different eta and delta0 (incuding intercept shifts and variation in wages)
sim2 = simulate(p,N,2);
simdata = rbind(sim1,sim2) # combine the two period

simdata[, lh := log(h)]
fit5 = summary(lm(lh ~ lw + X,simdata))
pander(fit5)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.399	0.004515	309.9	0
lw	1.075	0.00479	224.3	0
X	-0.5727	0.004647	-123.2	0

Table 10: Fitting linear model: lh ~ lw + X

Observations	Residual Std. Error	R^2	Adjusted R^2
86961	1.303	0.3806	0.3806

As we can see from above, if we run the regression with all of our data for both periods, both of our recovered parameters are negatively biased. Because δ_0 contribute to R_i, which influences an individual's lfp, we once again do not have the exogeneity condition if we ran our original regression on $\log w_i$ and x_i.

But if we now construct our heckman correction term using data from both simulations:

```
# construct the IMR, include in the regression and run
fit6 = glm(lfp ~ Z,simdata,family = binomial(link = "probit"))
pander(summary(fit6))
simdata[, ai := predict(fit6)]
simdata[, m := dnorm(ai)/pnorm(ai)]
fit7 = summary(lm(lh ~ lw + X + m,simdata))
pander(fit7)
# Including Plots
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.213	0.006785	178.8	0
lw	1.275	0.007253	175.8	0
X	-0.6352	0.004919	-129.1	0
m	1.003	0.02745	36.53	5.096e-290

Table 12: Fitting linear model: lh ~ lw + X + m

Observations	Residual Std. Error	R^2	Adjusted R^2
86961	1.293	0.39	0.3899

We see that our recovered parameters are now much closer to the true values. This is because we accounted for the heterogeneity in the individual's δ_0 from the two time periods by constructing the inverse mills ratio λ_i. So at each observation or individual i, we are able to correctly calculate the heckman correction term in our cross-section regressions.

Question 7

```
#slice the Z variables into deciles
simdata$Zbin <- cut(simdata$Z, breaks = 10, labels = c("1", "2", "3","4","5","6","7","8","9","10"))
#regress log-hours on log-wage, X, dummies for each of the bins, and time t
fit8 = summary(lm(lh ~ lw + X + factor(Zbin) + factor(t),simdata))
pander(fit8)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.718	0.2319	16.03	9.662e-58
lw	1.176	0.01355	86.79	0
X	-0.6041	0.006038	-100	0

	Estimate	Std. Error	t value	Pr(>|t|)
factor(Zbin)3	-0.8963	0.236	-3.798	0.0001459
factor(Zbin)4	-1.729	0.2299	-7.522	5.442e-14
factor(Zbin)5	-2.139	0.2307	-9.274	1.837e-20
factor(Zbin)6	-2.327	0.2323	-10.02	1.317e-23
factor(Zbin)7	-2.283	0.2346	-9.732	2.258e-22
factor(Zbin)8	-2.257	0.2376	-9.498	2.184e-21
factor(Zbin)9	-2.21	0.2433	-9.082	1.084e-19
factor(Zbin)10	-1.963	0.2712	-7.238	4.575e-13
factor(t)2	-0.2208	0.008893	-24.83	1.22e-135

Table 14: Fitting linear model: lh ~ lw + X + factor(Zbin) + factor(t)

Observations	Residual Std. Error	R^2	Adjusted R^2
86961	1.291	0.3924	0.3923

As we can see from the above results, the recovered coefficients are closer to the estimand of interest. It's because adding Zbin and t into the model as dummy variables captures the heterogeneity of individual's self-selection into labour from different groups and time periods. Equivalently, we are controlling for the group and time effects. But we see that this does not fully capture their effects as the coefficients still have discrepancies from the true values. This is because of our assumption that their composition effects can be fully accounted for by adding them (BDM (98)).