Econometrics 3: Assignment II

Laura B. Gomez Cely 2050216

June 2020

1. Create the logarithms of all variables. Additionally, create the interaction term of the log-cows and log-feed. Produce the summary statistics of all variables
In the panel data there is a missing value in the variable labor in 1997 for the individual 22. Therefore, I deleted this individual and created a balanced panel. After that, I estimate the descriptive variables for Farm, Year, Cows, Land, Milk, Labor, Feed, logCows, logland, loglabor, logfeed and int

Variable	Farm	Year	Cows	Land	Milk	Labor	Feed	lnmilk
Min	1	93	4.5	2	14410	1	3924	9.576
1st Qu	63	94	14.10	8.5	68076	1	25529	11.128
Median	124	95.5	20	12	110033	2	45751	11.609
Mean	124	95.5	22.12	12.97	130970	1.67	57876	11.576
3rd Qu	186	97	27	16	162969	2	73242	12.001
Max.	247	98	82.30	45.10	727281	4	376732	13.497

Variable	logCows	logland	loglabor	logfeed	int
Min	1.504	0.6931	-0.0000003	8.275	13.55
1st Qu	2.646	2.1401	0.0000000	10.148	26.96
Median	2.996	2.4849	0.6931469	10.731	32.18
Mean	2.978	2.4601	0.4549013	10.683	32.14
3rd Qu	3.296	2.7726	0.6931469	10.731	36.69
Max.	4.410	3.8089	1.3862944	12.863	56.63

2. First, estimate a pooled linear regression explaining the logarithm of milk production by the logarithm of the amount of cows, feed, land, and labor available. Are all variables significant at 5 and 10 significance level, respectively?

```
results
========================================
                 Dependent variable:
             ---------------------------
                       lnmilk
----------------------------------------
logCows               0.596***
                       (0.020)

logland                0.023**
                       (0.011)

loglabor                0.022
                       (0.013)

logfeed               0.451***
                       (0.011)

Constant              4.912***
                       (0.077)

----------------------------------------
Observations            1,476
R2                      0.952
Adjusted R2             0.952
F Statistic  7,368.948*** (df = 4; 1471)
========================================
Note:        *p<0.1; **p<0.05; ***p<0.01
```

According to the table, we can see that the log of the amount of spent on animal feed and the log of the average of cows on farm are significant at 1 per cent. The log of hectares of land available for cattle grazing is significant at 5 percent. However, the log of the number of workers on the farm is not significant even at 10 percent.

Therefore, we can conclude that if the average of cows on farm, the amount spent on animal feed, and the number of hectares of land available change by one percent, the number of liters milk increases by 0.5, 0.451 and 0.02 percent. Also, in average, if all variables change by one percent, the percent of milk could increase in 4.9 percent.

3.As the production technology can improve over time, include also time-specific effects and test their joint significance at 10 percent significance level.

```
results
=======================================
                  Dependent variable:
              ---------------------------
                        lnmilk
---------------------------------------
logCows                0.602***
                       (0.020)

logland                0.023**
                       (0.011)

loglabor               0.026*
                       (0.013)

logfeed                0.444***
                       (0.011)

year_94                0.017
                       (0.013)

year_95                0.031**
                       (0.013)

year_96                0.029**
                       (0.013)

year_97                0.024*
                       (0.013)

year_98                0.034***
                       (0.013)

Constant               4.947***
                       (0.079)

---------------------------------------
Observations             1,476
R2                       0.953
Adjusted R2              0.952
F Statistic  3,285.544*** (df = 9; 1466)
=======================================
Note:        *p<0.1; **p<0.05; ***p<0.01
```

After include time dummies, we can see that log of the average of cows on farm and the log of the amount spent on animal feed keep being significant at 1 percent. Also, the number of hectares of land keep being significant at 5 percent. However, after controlling by time, the log of the number of the workers of the farms become significant at 10 percent. It seems that when productivity increases, it reduces the number of workers needed, such as technological improvements, so that is why the effect of work at first is not significant and then it changes to be significant.

```
Lagrange Multiplier Test - time effects (Honda) for balanced panels

data:  model2_1
normal = -1.7356, p-value = 0.9587
alternative hypothesis: significant effects
```

However, the joint test significance indicates that the dummies on time are not significant. It could happen because we are not using the correct specification model. In other words, we can have some omit variables that can give us bias betas.

4. Estimate now the model including the time-specific effects by the random effect estimator. Are coefficients similar? Does the precision of some estimates increase? Repeat the test of significance of time-specific effects – does its p-value also change?

```
results
========================================
                   Dependent variable:
               ---------------------------
                         lnmilk
----------------------------------------
logCows                 0.651***
                        (0.021)

logland                 0.028**
                        (0.014)

loglabor                0.038**
                        (0.017)

logfeed                 0.399***
                        (0.011)

Constant                5.288***
                        (0.081)

----------------------------------------
Observations             1,476
R2                       0.895
Adjusted R2              0.894
F Statistic          12,476.810***
========================================
Note:          *p<0.1; **p<0.05; ***p<0.01
```

After estimating the model with random effects, we can see that the coefficients change just a bit in comparison with the pool model with time controlling.

```
Lagrange Multiplier Test - time effects (Honda) for balanced panels

data:  model2_2
normal = 0.76461, p-value = 0.2223
alternative hypothesis: significant effects
```

Also, we can see that we can not reject the null hypothesis, therefore, the time effects are not significant.

5. Perform the Breusch-Pagan test of the presence of random effects. What can you conclude?

```
Lagrange Multiplier Test - (Breusch-Pagan) for balanced panels

data:  model2_2
chisq = 1576.4, df = 1, p-value < 2.2e-16
alternative hypothesis: significant effects
```

Here we reject the null hypothesis and conclude that random effects is appropriate. This is evidence of significant differences across farms, therefore we can run a random regression.

6. Compute the parameter such that the ordinary least squares applied to data transformed to $y_{it} - \lambda \overline{y}_i$ and $x_{it} - \lambda \overline{x}_i$ is efficient (under the assumptions of the random effects model).

Computed lambda is equal to 0.7210945

7. Attempting to estimate a more flexible production function, add the squares of the (log of) cows and feed variables and their interaction. Evaluate also the correlation between the linear and squared terms. Does the significance of some variables change? What could be the reasons? For comparison, also run regression without the squared variables, but with the interaction term.

```
results
========================================
                   Dependent variable:
               ---------------------------
                         lnmilk
----------------------------------------
logCows                 1.236***
                         (0.330)

logland                  0.024*
                         (0.014)

loglabor                 0.040**
                         (0.017)

logfeed                  -0.402*
                         (0.237)

int                     -0.126**
                         (0.053)

Cows2                   0.127***
                         (0.046)

feed2                   0.055***
                         (0.017)

Constant                8.680***
                         (0.872)

----------------------------------------
Observations              1,476
R2                        0.896
Adjusted R2               0.895
F Statistic           12,632.630***
========================================
Note:          *p<0.1; **p<0.05; ***p<0.01
```

results

	Dependent variable:
	lnmilk
logCows	0.227**
	(0.099)
logland	0.024*
	(0.014)
loglabor	0.036**
	(0.017)
logfeed	0.273***
	(0.031)
int	0.040***
	(0.009)
Constant	6.608***
	(0.311)
Observations	1,476
R2	0.895
Adjusted R2	0.895
F Statistic	12,553.860***
Note:	*p<0.1; **p<0.05; ***p<0.01

After we include the squares of the logcows, logfeed and the interaction, all variables are significance at 10 per cent of significance. However, the sign of logfeed change and now is negative. It means that increasing the cow's feed does not generate more milk production, but when it reaches a point where the quadratic effect exceeds the linear effect (1493.8 pesetas), the effect of the square variable is bigger, increasing the production of milk.

Also, the square of logcows is positive, meaning that increasing the average of cows in one per cent, increase the per cent of milk produces.

Comparing with the model without squares but with interaction, the coefficients change a lot mainly for logcows, logfeed and the interaction. Also, the sign of logfeed now is positive. The possible reason is that the correlation between the variables and their squares is so high that generates bias inside the model

8. Estimate the correlated effects model. Are any time-averaged variables significant? Drop any nonsignificant time-averaged variables. Does the significance of some other variables change compared to the random effect estimation?

```
results
========================================
                 Dependent variable:
                 ---------------------------
                        lnmilk
----------------------------------------
logcowsi                -0.099*
                        (0.055)

loglandi                -0.010
                        (0.032)

loglabori               -0.025
                        (0.038)

logfeedi               0.096***
                        (0.030)

logCows                0.662***
                        (0.025)

logland                 0.035**
                        (0.016)

loglabor                 0.038
                        (0.023)

logfeed                0.382***
                        (0.012)

Constant               4.724***
                        (0.196)

----------------------------------------
Observations             1,476
R2                       0.892
Adjusted R2              0.892
F Statistic           12,161.040***
========================================
Note:        *p<0.1; **p<0.05; ***p<0.01
```

yes, the time average of logfeed is significant at 1 percent and the time average of logcow is significant at 10 per cent. Also, comparing with the random model, the variable loglabor is not significant even at 10 per cent.

9) By testing the hypothesis decide whether the random effects model has to be rejected.

```
Linear hypothesis test

Hypothesis:
logcowsi = 0
loglandi = 0
loglabori = 0
logfeedi = 0

Model 1: restricted model
Model 2: lnmilk ~ logcowsi + loglandi + loglabori + logfeedi + logCows +
    logland + loglabor + logfeed

  Res.Df Df  Chisq Pr(>Chisq)
1   1471
2   1467  4 11.559    0.02095 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From the test we can reject the null hypothesis that $\theta = 0$, therefore correlated random effects is more suitable than random effects.

10)Estimate now consistently the fixed effects model. Are the estimates closer to the random effects or correlated effects results?

the coefficient of the variables are closer to the correlated effects

```
results
========================================
                Dependent variable:
            ---------------------------
                      lnmilk
----------------------------------------
logCows              0.662***
                      (0.025)

logland               0.035**
                      (0.017)

loglabor              0.038
                      (0.023)

logfeed              0.382***
                      (0.012)

----------------------------------------
Observations           1,476
R2                     0.835
Adjusted R2            0.802
F Statistic  1,555.403*** (df = 4; 1226)
========================================
Note:        *p<0.1; **p<0.05; ***p<0.01
```

11) Using the Hausman test, does the random effects have to be rejected in the favor of the fixed effects model

Given the small p-value, we can reject the random effects model and accept the fixed effects model.

```
        Hausman Test

data:  model4_5
chisq = 12.638, df = 4, p-value = 0.01319
alternative hypothesis: one model is inconsistent
```

12)Repeat the Hausman test to choose between the correlated effects and the fixed effects model.

Given the p-value is greater than 0.05, we can not reject the null hypothesis. Therefore, we prefer in this case correlated effects than fixed effects model.

```
Hausman Test

data:  model4_5
chisq = 7.5094e-24, df = 4, p-value = 1
alternative hypothesis: one model is inconsistent
```

m) Given its variation in all models above, is the amount of workers (the log of labor) significant in the finally selected model at 10 percent level?

Given the Hausman test, we choose finally the correlated random effects. In this model the log of the variable labor is not significant even at 10. It is just significant for the case of random effects. Thus, we can conclude that the amount of workers don't affect the liters of milk production in any significance.

Appendix

```
———#Assignment Econometrics 3#————

rm(list = ls())
library(haven)

dairy <- read_dta("C:/Users/lgome/Downloads/dairy.dta")
dairy = as.data.frame(dairy)
View(dairy)
install.packages("plm")
library("plm")
dairy1<-na.omit(dairy)
dairy2 <- make.pbalanced(dairy1, balance.type = "shared.individuals")
any(is.na(dairy2))
is.pbalanced(dairy2)

————#. First Question. Create the logarithms of all variables. Additionally,
  #log-feed. Produce the summary statistics of all variables.————

dairy2$logCows=log(dairy2$cows)
dairy2$logland=log(dairy2$land)
dairy2$loglabor=log(dairy2$labor)
dairy2$logfeed=log(dairy2$feed)
dairy2$int=dairy2$logCows*dairy2$logfeed
summary(dairy2)

————#Question 2——— First, estimate a pooled linear regression explaining the l
  #————of the amount of cows, feed, land, and labor available. Are all variable
  ———#significance level, respectively?#—————
pdata2_1 = pdata.frame(dairy2,index = c("farm","year"))
varnames = names(pdata2_1)
```

```
xnames = c("logCows", "logland", "loglabor", "logfeed")
model2_1 = as.formula(paste("lnmilk ~", paste(xnames, collapse = "+")))
pooling = plm(model2_1, data = pdata2_1, model = "pooling")
summary(pooling)
install.packages("stargazer")
library(stargazer)

stargazer(pooling, type="text", title="results", align=TRUE)

-------#Question 3 ---As the production technology can improve over time, include
  -------#joint significance at 10% significance level.

pdata2_1$year_93 <- 1L * (pdata2_1$year==93)
pdata2_1$year_94 <- 1L * (pdata2_1$year==94)
pdata2_1$year_95 <- 1L * (pdata2_1$year==95)
pdata2_1$year_96 <- 1L * (pdata2_1$year==96)
pdata2_1$year_97 <- 1L * (pdata2_1$year==97)
pdata2_1$year_98 <- 1L * (pdata2_1$year==98)

timelist = c("year_94", "year_95", "year_96", "year_97", "year_98")
xnames = c("logCows", "logland", "loglabor", "logfeed")
xnames = c(xnames, timelist)
model2_1 = as.formula(paste("lnmilk ~", paste(xnames, collapse = "+")))
pooling2_1 = plm(model2_1, data = pdata2_1, model = "pooling")
summary(pooling2_1)
stargazer(pooling2_1, type="text", title="results", align=TRUE)
plmtest(pooling2_1,"time")

#-----Question 4. Estimate now the model including the time-specific effects by t
#----time-specific effects - does its p-value also change?


xnames2_2 = c("logCows", "logland", "loglabor", "logfeed")
model2_2 = as.formula(paste("lnmilk ~", paste(xnames2_2, collapse = "+")))
random2_2 = plm(model2_2, data = pdata2_1, effects="twoways", model = "random")
summary(random2_2)
stargazer(random2_2, type="text", title="results", align=TRUE)
plmtest(random2_2,"time")


#-------Question 5---Perform the Breusch-Pagan test of the presence of random effe

plmtest(random2_2, type=c("bp"))

#----- Question 6 -----Compute the parameter ?? such that the ordinary least squar
#---and xit ??? ??x i is efficient (under the assumptions of the random effects
```

```
summary(random2_2)$ercomp$theta

#----Question 7--Attempting to estimate a more flexible production function, add
#---feed variables and their interaction. Evaluate also the correlation between t
#--------Does the significance of some variables change? What could be the reasons
#-----regression without the squared variables, but with the interaction term.

pdata2_1$Cows2=pdata2_1$logCows*pdata2_1$logCows
pdata2_1$feed2=pdata2_1$logfeed*pdata2_1$logfeed
pdata2_1$int2=pdata2_1$int*pdata2_1$int
cor(pdata2_1$logCows,pdata2_1$Cows2)
cor(pdata2_1$logfeed,pdata2_1$feed)
cor(pdata2_1$int,pdata2_1$int2)

xnames4_1 = c("logCows", "logland", "loglabor", "logfeed","int", "Cows2", "feed2
model4_1 = as.formula(paste("lnmilk ~", paste(xnames4_1, collapse = "+")))
random4_1 = plm(model4_1, data = pdata2_1, effects="twoways", model = "random" )
summary(random4_1)
stargazer(random4_1, type="text", title="results", align=TRUE)

xnames4_2 = c("logCows", "logland", "loglabor", "logfeed","int")
model4_2 = as.formula(paste("lnmilk ~", paste(xnames4_2, collapse = "+")))
random4_2 = plm(model4_2, data = pdata2_1, effets="twoways", model = "random" )
summary(random4_2)
stargazer(random4_2, type="text", title="results", align=TRUE)

#--------Quetion8------Estimate the correlated effects model. Are any time-averaged v

Cowavg=tapply(pdata2_1$logCows, INDEX=pdata2_1$farm, FUN=mean)
pdata2_1$logcowsi = 1
for (i in pdata2_1$farm){pdata2_1$logcowsi[pdata2_1$farm==i]= Cowavg[i]}

Landavg=tapply(pdata2_1$logland, INDEX=pdata2_1$farm, FUN=mean)
pdata2_1$loglandi = 1
for (i in pdata2_1$farm){pdata2_1$loglandi[pdata2_1$farm==i]= Landavg[i]}

Laboravg=tapply(pdata2_1$loglabor, INDEX=pdata2_1$farm, FUN=mean)
pdata2_1$loglabori = 1
for (i in pdata2_1$farm){pdata2_1$loglabori[pdata2_1$farm==i]= Laboravg[i]}

Feedavg=tapply(pdata2_1$logfeed, INDEX=pdata2_1$farm, FUN=mean)
pdata2_1$logfeedi = 1
for (i in pdata2_1$farm){pdata2_1$logfeedi[pdata2_1$farm==i]= Feedavg[i]}

xnames4_3 = c("logcowsi", "loglandi", "loglabori", "logfeedi","logCows", "loglan
```

```
model4_3 = as.formula(paste("lnmilk ~", paste(xnames4_3, collapse = "+")))
random4_3 = plm(model4_3, data = pdata2_1, effect = "twoways", model = "random"
summary(random4_3)
stargazer(random4_3, type="text", title="results", align=TRUE)

xnames4_4 = c( "logfeedi", "logCows", "logfeed", "loglabor" )
model4_4 = as.formula(paste("lnmilk ~", paste(xnames4_4, collapse = "+")))
random4_4 = plm(model4_4, data = pdata2_1, effect="twoways", model = "random" )
summary(random4_4)
stargazer(random4_4, type="text", title="results", align=TRUE)

#---------Question 9 ------By testing the hypothesis H0 : ?? = 0, decide whether the
install.packages("car")
library(car)
linearHypothesis(random4_3, c("logcowsi=0","loglandi=0", "loglabori=0", "logfeed

#-----Question 10--- Estimate now consistently the fixed effects model. Are the es
#correlated effects results?
xnames4_5 = c("logCows", "logland", "loglabor", "logfeed")
model4_5 = as.formula(paste("lnmilk ~", paste(xnames4_5, collapse = "+")))
fixed4_5 = plm(model4_5, data = pdata2_1, model = "within")
summary(fixed4_5)
stargazer(fixed4_5, type="text", title="results", align=TRUE)

#--------Question 11------ Using the Hausman test, does the random effects have to
#model?----#

phtest(fixed4_5,random2_2)

#-----Question 12---- Repeat the Hausman test to choose between the correlated eff

phtest(fixed4_5,random4_3)
```