

# 613\_hw4

## Q1

```
library(Matrix)
DATA<-read.csv("~/Desktop/613/hw4/Koop-Tobias.csv")
set.seed(100)
#Randomly select 5 individuals
vect<-as.vector(sample(1:2178, size=5))
#Set a vector to store the panel dimension of these 5 individuals
freque<-c()
w<-table(DATA$PERSONID)
w<-as.data.frame(w)
w<-as.matrix(w)
for (i in 1:5){
  ans<-w[vect[i],2]
  freque<-c(freque,ans)
}
#Print the outcome
print(vect)
```

```
## [1] 671 561 1202 123 1019
```

```
print(freque)
```

```
## Freq Freq Freq Freq Freq
## "14" " 5" " 8" " 9" " 3"
```

## Q2

```
library(nlme)
model_q2<-gls(LOGWAGE~EDUC+POTEXPER, data=DATA)
summary(model_q2)
```

```
## Generalized least squares fit by REML
##   Model: LOGWAGE ~ EDUC + POTEXPER
##   Data: DATA
##           AIC           BIC      logLik
##   24927.91 24959.08 -12459.95
##
## Coefficients:
##           Value      Std.Error   t-value p-value
## (Intercept) 0.7941911 0.027359284 29.02821      0
## EDUC         0.0938637 0.001929927 48.63590      0
## POTEXPER     0.0374053 0.000898886 41.61293      0
##
## Correlation:
##           (Intr) EDUC
## EDUC      -0.954
## POTEXPER -0.470   0.219
##
## Standardized residuals:
##           Min           Q1           Med           Q3           Max
## -5.16347553 -0.57658469  0.04692351  0.65688942  4.38118968
##
## Residual standard error: 0.4846115
## Degrees of freedom: 17919 total; 17916 residual
```

# Q3

## Between Estimator

```

fact<-as.factor(DATA$PERSONID)
#Calculate average logwage_i overtime
groupDATAwage<-as.matrix(tapply(DATA$LOGWAGE,fact,FUN=mean))
freque<-as.data.frame(table(DATA$PERSONID))
Freque<-as.matrix(freque$Freq)
DATAlogwagenew<-c()
for (i in 1:nrow(Freque)){
  DATAlogwagenew<-c(DATAlogwagenew,as.vector(rep(groupDATAwage[i,1],Freque[i,1])))
}
#Calculate average educ_i overtime
groupDATAeduc<-as.matrix(tapply(DATA$EDUC,fact,FUN=mean))
DATAeducnew<-c()
for (i in 1:nrow(Freque)){
  DATAeducnew<-c(DATAeducnew,as.vector(rep(groupDATAeduc[i,1],Freque[i,1])))
}
#Calculate average potexper_i overtime
groupDATApote<-as.matrix(tapply(DATA$POTEXPER,fact,FUN=mean))
DATApotenew<-c()
for (i in 1:nrow(Freque)){
  DATApotenew<-c(DATApotenew,as.vector(rep(groupDATApote[i,1],Freque[i,1])))
}
DATApersonidnew<-as.matrix(DATA$PERSONID)
constant<-matrix(1,nrow=17919,ncol=1)
#Do Estimation (Between model)
DATA_q3between<-data.frame(DATApersonidnew,DATAlogwagenew,DATAeducnew,DATApotenew)
DATA_q3between<- DATA_q3between[!duplicated(DATA_q3between$DATApersonidnew),]
model_between<-lm(DATAlogwagenew~DATAeducnew+DATApotenew,data=DATA_q3between)
summary(model_between)$coefficients

```

```

##              Estimate  Std. Error  t value    Pr(>|t|)
## (Intercept)  0.84556883  0.077017914  10.978859  2.483879e-27
## DATAeducnew  0.09309987  0.004668494  19.942162  2.128692e-81
## DATApotenew  0.02599874  0.003604887   7.212084  7.571723e-13

```

## Within Estimator

```

#Modify the data
DATAlogwagnew_deta<-as.matrix(DATA$LOGWAGE)-DATAlogwagnew
DATAeducnew_deta<-as.matrix(DATA$EDUC)-DATAeducnew
DATApotenew_deta<-as.matrix(DATA$POTEXPER)-DATApotenew
DATApersonidnew<-as.matrix(DATA$PERSONID)
DATA_q3within<-as.matrix(data.frame(DATApersonidnew,DATAlogwagnew_deta,DATAeducnew_deta,DATApotenew_deta))
#Use ols method to calculate within model estimators
X_q3within<-DATA_q3within[,3:4]
Y_q3within<-as.matrix(DATAlogwagnew_deta)
beta_q3within<-solve((t(X_q3within)%*%X_q3within))%*%(t(X_q3within)%*%Y_q3within)
print(beta_q3within)

```

```

##                [,1]
## DATAeducnew_deta 0.12366202
## DATApotenew_deta 0.03856107

```

```

#Use lm package to check the answer
DATA_q3within<-data.frame(DATApersonidnew,DATAlogwagnew_deta,DATAeducnew_deta,DATApotenew_deta)
model_within<-lm(DATAlogwagnew_deta~DATAeducnew_deta+DATApotenew_deta,data=DATA_q3within)
summary(model_within)$coefficients

```

```

##                Estimate Std. Error      t value      Pr(>|t|)
## (Intercept)    9.231826e-17 0.002348813 3.930422e-14 1.000000e+00
## DATAeducnew_deta 1.236620e-01 0.005400471 2.289837e+01 2.107861e-114
## DATApotenew_deta 3.856107e-02 0.000710902 5.424245e+01 0.000000e+00

```

# First time difference Estimator

```

diff_wage<-c()
diff_educ<-c()
diff_potex<-c()
logwage<-as.matrix(DATA$LOGWAGE)
educ<-as.matrix(DATA$EDUC)
potex<-as.matrix(DATA$POTEXPER)
index<-1
#Take the difference of the data
for (i in 1:2178){
  if(Freque[i]>=2){
    #Take variables at time t-1: logwage_t-1, educ_t-1, potexper_t-1
    i_lag_wage<-logwage[index:(index-1+Freque[i]-1),1]
    i_lag_educ<-educ[index:(index-1+Freque[i]-1),1]
    i_lag_potex<-potex[index:(index-1+Freque[i]-1),1]
    #Take variables at time t: logwage_t, educ_t, potexper_t
    #print(index+1)
    #print(freque[i,1])
    i_t_wage<-logwage[(index+1):(index-1+Freque[i]),1]
    i_t_educ<-educ[(index+1):(index-1+Freque[i]),1]
    i_t_potex<-potex[(index+1):(index-1+Freque[i]),1]
    i_diff_wage<-i_t_wage-i_lag_wage
    i_diff_educ<-i_t_educ-i_lag_educ
    i_diff_potex<-i_t_potex-i_lag_potex
    diff_wage<-c(diff_wage,i_diff_wage)
    diff_educ<-c(diff_educ,i_diff_educ)
    diff_potex<-c(diff_potex,i_diff_potex)
    index<-index+Freque[i,1]
  }
}
#Do estimation (First Difference Model)
DATA_q3diff<-as.matrix(data.frame(diff_wage,diff_educ,diff_potex))
X_q3diff<-DATA_q3diff[,2:3]
Y_q3diff<-DATA_q3diff[,1]
beta_q3diff<-solve((t(X_q3diff)%*%X_q3diff))%*%(t(X_q3diff)%*%Y_q3diff)
#Print out the outcome
print(beta_q3diff)

```

```

##              [,1]
## diff_educ    0.08558180
## diff_potex   0.03796838

```

```

#Use lm package to check the answer
DATA_q3diff<-data.frame(diff_wage,diff_educ,diff_potex)
model_diff<-lm(diff_wage~diff_educ+diff_potex,data=DATA_q3diff)
summary(model_diff)$coefficients

```

##		Estimate	Std. Error	t value	Pr(> t )
##	(Intercept)	-0.0008820122	0.003423167	-0.2576597	7.966729e-01
##	diff_educ	0.0855828031	0.003472454	24.6462057	1.249147e-131
##	diff_potex	0.0379680401	0.000945046	40.1758632	0.000000e+00

**Q3.4 Compare beta\_1 and beta\_2 under different models.**

The beta calculated by using between model and using first difference model are very close.

The coefficient of potexper calculated by using all three types of fixed effect model and that by using the random effect model are very similar as well.

However, the coefficient of education of the random effect model is a little bit different from that calculated by using fixed effect models.

This implies that the individual effect might be somekind correlated with the explaining variables

**Q4**

**Q4.1**

```

###Select 100 samples
set.seed(613)
Sample_q4<-sample(1:2178, 100)
DATA_q4<-matrix(0,nrow=1,ncol=9)
for (i in 1:100){
  WAGE_i<-as.matrix(DATA$LOGWAGE[DATA$PERSONID==Sample_q4[i]])
  EDUC_i<-as.matrix(DATA$EDUC[DATA$PERSONID==Sample_q4[i]])
  POTEXPER_i<-as.matrix(DATA$POTEXPER[DATA$PERSONID==Sample_q4[i]])
  PERSONID_i<-as.matrix(DATA$PERSONID[DATA$PERSONID==Sample_q4[i]])
  ABILITY_i<-as.matrix(DATA$ABILITY[DATA$PERSONID==Sample_q4[i]])
  MOTHERED_i<-as.matrix(DATA$MOTHERED[DATA$PERSONID==Sample_q4[i]])
  FATHERED_i<-as.matrix(DATA$FATHERED[DATA$PERSONID==Sample_q4[i]])
  BRKNHOME_i<-as.matrix(DATA$BRKNHOME[DATA$PERSONID==Sample_q4[i]])
  SIBLINGS_i<-as.matrix(DATA$SIBLINGS[DATA$PERSONID==Sample_q4[i]])
  PERSONID_i<-as.matrix(DATA$PERSONID[DATA$PERSONID==Sample_q4[i]])
  individual<-as.matrix(data.frame(PERSONID_i,WAGE_i,EDUC_i,POTEXPER_i,ABILITY_i,MOTHERED_i,FATHERED_i,BRKNHOME_i,SIBLINGS_i))
  DATA_q4<-rbind(DATA_q4,individual)
}

M<-nrow(DATA_q4)
DATA_q4<-DATA_q4[2:M,]
DATA_q4df<-as.data.frame(DATA_q4)
#Get frequency matrix which stores the frequency of each individual's observations
W_q4<-as.data.frame(table(DATA_q4df$PERSONID_i))
Freque_q4<-as.matrix(W_q4$Freq)
#Construct a likelihood function
likelihood<-function(par,DATA.=DATA_q4){
  likewage<-DATA.[,2]
  likeeduc<-DATA.[,3]
  likepotexper<-DATA.[,4]
  alfa<-par[2:101]
  betal<-par[102]
  beta2<-par[103]
  alfanew<-rep(alfa,Freque_q4)
  alfanew<-as.matrix(alfanew)
  Estimation<-alfanew+likeeduc*betal+likepotexper*beta2
  proEstimation<-dnorm((likewage- Estimation)/par[1])
  proEstimation[proEstimation<0.00001]<-0.00001
  proEstimation[proEstimation>0.99999]<-0.99999
  loglikelihood<--sum(log(proEstimation))
  return(loglikelihood)
}
#Set initial value for the parameter
parameter<-rnorm(102)
parameter<-c(1,parameter)
#Optimize the likelihood function
result_q4.1<-optim(par = parameter,likelihood)
print(result_q4.1)

```

```
## $par
## [1] 1.00000000 -1.25077351 -1.23742224 0.69523430 -1.17679012
## [6] -0.08068324 0.46081104 0.91278708 1.59763959 -1.98707776
## [11] 0.08445674 1.29185357 1.53491371 -0.55306496 0.54951946
## [16] 0.58542189 1.20576095 -1.73014998 -0.25530253 -1.82794356
## [21] 0.56809363 0.10682645 0.48153848 -1.22774092 -1.52015605
## [26] -0.12177822 0.66043497 0.43014612 0.23186502 0.14623862
## [31] -2.55806843 -0.46124848 0.46038107 -0.73495133 -1.10109127
## [36] 0.15664930 -0.83188237 -1.24190168 1.50975916 -0.02710235
## [41] 0.44862196 -1.84793619 -3.97142785 0.28449669 -0.32978955
## [46] -0.15063015 0.31059439 0.29797915 -0.90850221 0.84660544
## [51] -0.08764736 -0.65966377 -1.59920585 0.07589521 -1.89898829
## [56] -1.06989123 -0.55748048 -0.85508621 -0.74909498 1.03530167
## [61] -0.63154088 0.39075821 -0.12786100 1.25552988 0.25942019
## [66] -1.51348038 0.46534419 -1.05643416 0.28022557 0.81562739
## [71] -0.18974537 0.87958593 0.23597572 -1.07435090 2.13786887
## [76] 0.15546938 0.95314884 -0.45760452 2.19346820 -0.07392728
## [81] -1.32512697 1.67854378 -1.91708432 0.43020511 -1.11165067
## [86] 0.73019837 0.45803767 1.14799478 -2.05271300 -1.07523924
## [91] 1.10489036 -0.25848502 1.55271742 -0.67354231 -0.48442545
## [96] -0.59664023 2.17925548 0.23984158 -0.98578039 -1.59232450
## [101] 0.58990485 -1.10913704 -0.10000147
##
## $value
## [1] 9947.168
##
## $counts
## function gradient
## 104 NA
##
## $convergence
## [1] 0
##
## $message
## NULL
```

## Q4.2



```

#Get alfa by using ols method (Sample: 100 individuals)
result_q4.2<-lm(WAGE_i~EDUC_i+POTEXPER_i+factor(PERSONID_i),data=DATA_q4df)
alfa_q4.2<-as.matrix(coef(result_q4.2))
alfa_q4.2<-alfa_q4.2[3:102]
DATA_q4dfNEW<- DATA_q4df[!duplicated(DATA_q4df$PERSONID_i),]
DATA_4.2dfnew<-data.frame(alfa_q4.2,DATA_q4dfNEW)
#Run a regression of estimated individual fixed effets on the invariant variables.
result_q4.2new<-lm(alfa_q4.2~ABILITY_i+MOTHERED_i+FATHERED_i+BRKNHOME_i+SIBLINGS_i,da
ta=DATA_4.2dfnew)
#Print the outcome
summary(result_q4.2new)

```

```

##
## Call:
## lm(formula = alfa_q4.2 ~ ABILITY_i + MOTHERED_i + FATHERED_i +
##      BRKNHOME_i + SIBLINGS_i, data = DATA_4.2dfnew)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.16532 -0.30995  0.02221  0.23708  1.05565
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.101043   0.215052  -5.120 1.62e-06 ***
## ABILITY_i    -0.061407   0.053026  -1.158   0.250
## MOTHERED_i    0.011262   0.019115   0.589   0.557
## FATHERED_i   -0.001082   0.016940  -0.064   0.949
## BRKNHOME_i   -0.153286   0.129872  -1.180   0.241
## SIBLINGS_i    0.028928   0.018830   1.536   0.128
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4222 on 94 degrees of freedom
## Multiple R-squared:  0.04843,    Adjusted R-squared:  -0.002185
## F-statistic: 0.9568 on 5 and 94 DF,  p-value: 0.4484

```

## Q4.3

The standard errors in the previous model are wrong because that there might be some estimation errors which is caused when we generate individual fixed effects.

We can solve the problem by using bootstrap to eliminate the estimation errors generated from individual effects estimation.

```

coeff<-matrix(0,nro=6,ncol=1)
for (i in 1:100){
  #Sampling
  DATA_q4.3<-matrix(0,nrow=1,ncol=10)
  colnames(DATA_q4.3)<-c("PERSONID","EDUC","LOGWAGE","POTEXPER","TIMETRND","ABILITY",
,"MOTHERED","FATHERED","BRKNHOME","SIBLINGS")
  k<-sample(1:2178,100)
  for (i in 1:100){
    trans<-DATA[DATA$PERSONID==k[i],]
    DATA_q4.3<-rbind(DATA_q4.3,trans)
  }
  DATA_q4.3<-as.matrix(DATA_q4.3)
  n<-nrow(DATA_q4.3)
  DATA_q4.3<-as.data.frame(DATA_q4.3[2:n,])
  #Estimate individual effect
  result_q4.3<-lm(LOGWAGE~EDUC+POTEXPER+factor(PERSONID),data=DATA_q4.3)
  alfa_q4.3<-as.matrix(coef(result_q4.3))
  alfa_q4.3<-alfa_q4.3[3:102]
  DATA_q4.3<-DATA_q4.3[!duplicated(DATA_q4.3$PERSONID),]
  DATA_4.3dfnew<-data.frame(alfa_q4.2,DATA_q4.3)
  #Run a regression of estimated individual fixed effects on the invariant variables.
  result_q4.3new<-lm(alfa_q4.3~ABILITY+MOTHERED+FATHERED+BRKNHOME+SIBLINGS,data=DATA_4.3dfnew)
  result<-summary(result_q4.3new)
  result<-as.matrix(coef(result))
  coeff<-cbind(coeff,result[,1])
}
coeff<-as.matrix(coeff)
coeff<-coeff[,2:101]
result_q4.3<-apply(coeff,1,FUN=var)
#Print the outcome
print(sqrt(result_q4.3))

```

## (Intercept)	ABILITY	MOTHERED	FATHERED	BRKNHOME	SIBLINGS
## 0.37335654	0.05385491	0.01952550	0.01468313	0.10900221	0.02330654