Sesión 4 - Comprobación de supuestos y elección del modelo de datos panel

Camilo Forero - Jhan Andrade - Germán Camilo Rodríguez

30/09/2020

Introducción

A la hora de manejar datos panel existen 4 estimadores diferentes, todos basados en derivaciones del estimador de **mínimos cuadrados ordinarios** ó variaciones del estimador de **mínimos cuadrados generalizados** en el caso de efectos aleatorios. Dichos estimadores son:

1. Estimador de mínimos cuadrados combinados.
2. Estimador de primeras diferencias.
3. Estimador de efectos fijos.
4. Estimador de efectos aleatorios.

Todos los anteriores estimadores parten de la siguiente representación:

$$y_{it} = \beta_0 + \beta_1 x_{it1} + \beta_2 x_{it2} + \cdots + \beta_k x_{itk} + a_i + u_{it} \tag{1}$$

Según sea el caso, se puede estimar directamente cuando no se viola el supuesto de exogeneidad, obteniendo el estimador de Mínimos Cuadrados Combinados. Mientras que cuando se viola dicho supuesto de exogeneidad por cuenta del factor de heterogeneidad inobservable, el cual es invariable en el tiempo, se requiere **diferenciar** o aplicar una **transformación intragrupal**, para obtener el estimador de PD y EF, respectivamente. Así pues, con la **diferenciación** o **transformación intragrupal**, el efecto inobservable que genera sesgo de variable omitida desaparece.

Por otra parte, en el caso del estimador de efectos aleatorios,no existe endogeneidad por cuenta del factor de heterogeneidad inobservable, por lo tanto, $cov(a_i, x_{itj}) = 0$. Por el contrario, el estimador de EF y PD parten del supuesto de que existe correlación entre el efecto fijo inobservable a_i y una o más variables regresoras x_{itk}.

Importación de paquetes en R

En primer lugar, para el manejo de datos panel en R se usan los siguientes paquetes:

```
#Paquetes
# install.packages("plm")          #Panel linear models
# install.packages("gplots")       #Tools for plotting data
# install.packages("stargazer")    #Tablas más estéticas
# install.packages("foreing")      #Importar datos
# install.packages("sandwich")     #Estimación con de errores robustos
# install.packages("lmtest")
# install.packages("tseries")
# install.packages("wooldridge")
```

```
library(plm);library(gplots);library(stargazer)
library(foreign);library(sandwich);library(lmtest)
library(tseries);library(wooldridge)
```

La sesión consistirá en explorar cómo estimar y elegir el modelo de datos panel más adecuado, además de cómo comprabar algunos supuestos de dichos modelos por medio de R

Ejemplo 1: Base de datos - Education.

El presente ejemplo está basado en el ejemplo que se encuentra en la sección 14.3 del libro de Wooldridge

En este ejemplo analiza cómo diferentes factores pueden afectar el salario de los trabajadores. Para ello, se estimará la siguiente regresión:

$$lwage = \text{dummies temporales } + black + hisp + exper + exper2 + married + union + yr$$

Para los 4 modelos descritos anteriormente, no obstante, es necesario hacer hay algunas salvedades que se explicarán más adelante.

```
#Base de datos: Wooldridge 14.4 ¿Ha cambiado la educación a lo largo del tiempo?
#Education=read.dta("http://fmwww.bc.edu/ec-p/data/wooldridge/wagepan.dta")
data("wagepan")
Education = wagepan
attach(Education)
#View(Education)
```

Lo primero que hay que hacer al trabajae con datos panel es transformar la base de datos original en un pdata.frame. Lo anterior facilita la estimación de los cuatro modelos mencionados previamente. Generalmente, hay que proveer un índice para cada observación de corte transversal y un índice para el tiempo, en ese órden respectivamente.

```
#Tratar la base de datos como un panel de datos
panel.Education = pdata.frame(Education, index = c("nr","year"))
class(panel.Education)
```

```
## [1] "pdata.frame" "data.frame"
```

```
#help("pdata.frame")
```

Ver las dimensiones del Panel y las variables que no cambian en el tiempo y/o entre individuos:

```
#Para conocer las dimensiones del panel
pdim(panel.Education)
```

```
## Balanced Panel: n = 545, T = 8, N = 4360
```

```
#Para determinar si las variables cambian a lo largo del tiempo o entre individuos
pvar(panel.Education)
```

```
## no time variation:       nr black hisp educ
## no individual variation: year d81 d82 d83 d84 d85 d86 d87
```

Del ejemplo se observa que el panel está balanceado[1] donde en total hay 545 observaciones por panel, es

[1] Un panel balanceado es aquel que para todas las observaciones tiene el mismo número de periodos, i.e. todas las observaciones tienen la misma cantidad de observaciones temporales

decir, en total hay 8 paneles y 4360 observaciones.

De igual forma, se observa que variables como el color de piel *black*, *hisp* y *educ* no cambian en el tiempo, por lo que no pueden incluirse como regresores de los estimadores de primera diferencia y efectos fijos, debido a que están incorporados en el término de heterogeneidad inobservable que no cambia en el tiempo. Sin embargo, sí puede incluirse en la regresión de mínimos cuadrados combinados y en efectos aleatorios.

Las dummies temporales y la variable *year* no varía entre observaciones (es la misma para todos los individuos en un mismo panel) de un mismo panel, pero sí entre observaciones de diferentes paneles.

A continuación, se definirán algunas variables que se utilizarán en las regresiones propuestas:

```
#Definición de las variables
exper2 = expersq
panel.Education$yr = factor(panel.Education$year)#Una Dummy para cada año de la muestra
```

La función factor me permite transformar vectores, ya sean numéricos o de caracter, en variables tipo factor. En este caso, dicha variable factor va a representar dummies temporales. Finalmente, es importante considerar que una variable tipo factor podría describir cualquier variable categórica, sin importar el número de categorias que esta tenga, por ejemplo, raza o sexo.

Para ver las diferentes categorias de la variable *panel.Education$yr* se utiliza el siguiente código:

```
summary(panel.Education$yr)
```

```
## 1980 1981 1982 1983 1984 1985 1986 1987
##  545  545  545  545  545  545  545  545
```

De lo anterior, se muestra que la categoría 1980 tiene 545 observaciones, la categoría 1981 tiene 545 obeservaciones y así sucesivamente para todas las categorías. Dado que todos los años tienen el mismo número de observaciones, se concluye que el panel está balanceado.

Estimaciones de datos panel

La función que se usa para emplear cualquiera de los cuatro modelos previamentes descritos arriba es la función **plm**. Los argumentos de dicha función son:

- Fórmula: Expresión de la formula del modelo estadístico que se va a estimar. En este caso, *lwage~educ+black+hisp+exper+exper2+married+union+yr*
- Data: Base de datos donde se encuentran las variables. En este caso, *panel.Education*
- model: Especifica el tipo de estimador que se quiere utilizar. Hay cuatro opciones:
  - pooling: Para estimaciones de mínimos cuadrados combinados
  - random: Para estimaciones de efectos aleatorios
  - within: Para estimaciones de efectos fijos (estimador intragrupal o within)
  - Between: Es una transformación similar a la de EF, la cual en vez de utilizar la media de cada individuo en todos los periodos, utiliza la media de toda la muestra para cada uno de los años (por ende, se pierde un mayor número de observaciones).
  - fd: Para estimaciones de primeras diferencias

De igual forma, la función **plm** es útil porque sus resultados (el outcome) se pueden utilizar con otros paquetes como **stargazer** y **broom**.

Estimador de mínimos cuadrados combinados

```
Pooled = plm(lwage~educ+black+hisp+exper+exper2+married+union+yr,
             data=panel.Education, model="pooling")
summary(Pooled)
```

```
## Pooling Model
```

```
##
## Call:
## plm(formula = lwage ~ educ + black + hisp + exper + exper2 +
##     married + union + yr, data = panel.Education, model = "pooling")
##
## Balanced Panel: n = 545, T = 8, N = 4360
##
## Residuals:
##     Min.  1st Qu.   Median  3rd Qu.     Max.
## -5.26573 -0.24838  0.03192  0.29475  2.52912
##
## Coefficients:
##                Estimate  Std. Error t-value  Pr(>|t|)
## (Intercept)  0.09205578  0.07827010  1.1761 0.2396076
## educ         0.09134979  0.00523738 17.4419 < 2.2e-16 ***
## black       -0.13923421  0.02357956 -5.9049 3.799e-09 ***
## hisp         0.01601951  0.02079714  0.7703 0.4411788
## exper        0.06723450  0.01369484  4.9095 9.467e-07 ***
## exper2      -0.00241170  0.00081995 -2.9413 0.0032860 **
## married      0.10825295  0.01568942  6.8997 5.962e-12 ***
## union        0.18246128  0.01715677 10.6349 < 2.2e-16 ***
## yr1981       0.05831999  0.03035363  1.9214 0.0547528 .
## yr1982       0.06277442  0.03321407  1.8900 0.0588251 .
## yr1983       0.06201174  0.03666013  1.6915 0.0908072 .
## yr1984       0.09046719  0.04009071  2.2566 0.0240849 *
## yr1985       0.10924630  0.04335248  2.5200 0.0117725 *
## yr1986       0.14195959  0.04642297  3.0580 0.0022421 **
## yr1987       0.17383343  0.04943305  3.5165 0.0004417 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    1236.5
## Residual Sum of Squares: 1002.5
## R-Squared:      0.18928
## Adj. R-Squared: 0.18667
## F-statistic: 72.4588 on 14 and 4345 DF, p-value: < 2.22e-16
```

Estimador de primeras diferencias

```
FD = plm(lwage~educ+black+hisp+exper+exper2+married+union+yr,
             data=panel.Education, model="fd")
summary(FD)
```

```
## Oneway (individual) effect First-Difference Model
##
## Call:
## plm(formula = lwage ~ educ + black + hisp + exper + exper2 +
##     married + union + yr, data = panel.Education, model = "fd")
##
## Balanced Panel: n = 545, T = 8, N = 4360
## Observations used in estimation: 3815
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
```

```
## -4.591254 -0.144894 -0.013479  0.134292  4.841449
##
## Coefficients: (1 dropped because of singularities)
##               Estimate Std. Error t-value  Pr(>|t|)
## (Intercept)  0.1401465  0.0292530  4.7908 1.724e-06 ***
## exper2      -0.0057546  0.0021701 -2.6518  0.008039 **
## married      0.0381433  0.0229385  1.6628  0.096425 .
## union        0.0411497  0.0196922  2.0896  0.036716 *
## yr1981       0.0158512  0.0218647  0.7250  0.468517
## yr1982      -0.0164964  0.0313778 -0.5257  0.599104
## yr1983      -0.0485531  0.0359882 -1.3491  0.177373
## yr1984      -0.0449451  0.0359831 -1.2491  0.211720
## yr1985      -0.0498952  0.0313926 -1.5894  0.112054
## yr1986      -0.0325370  0.0218890 -1.4865  0.137243
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    751.19
## Residual Sum of Squares: 746.37
## R-Squared:      0.0064227
## Adj. R-Squared: 0.0040726
## F-statistic: 2.73293 on 9 and 3805 DF, p-value: 0.0035248
```

Estimador de efectos fijos

```
Fixed = plm(lwage~educ+black+hisp+exper+exper2+married+union+yr,
            data=panel.Education, model="within")
summary(Fixed)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = lwage ~ educ + black + hisp + exper + exper2 +
##     married + union + yr, data = panel.Education, model = "within")
##
## Balanced Panel: n = 545, T = 8, N = 4360
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -4.159280 -0.125273  0.011267  0.154869  1.492088
##
## Coefficients: (1 dropped because of singularities)
##             Estimate  Std. Error t-value  Pr(>|t|)
## exper    0.13214642  0.00982473 13.4504 < 2.2e-16 ***
## exper2  -0.00518550  0.00070444 -7.3612 2.222e-13 ***
## married  0.04668036  0.01831044  2.5494   0.01083 *
## union    0.08000186  0.01931031  4.1430 3.503e-05 ***
## yr1981   0.01904479  0.02036260  0.9353   0.34970
## yr1982  -0.01132198  0.02022754 -0.5597   0.57570
## yr1983  -0.04199552  0.02032053 -2.0667   0.03883 *
## yr1984  -0.03847088  0.02031441 -1.8938   0.05833 .
## yr1985  -0.04324982  0.02024576 -2.1362   0.03272 *
## yr1986  -0.02738194  0.02038633 -1.3432   0.17930
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    572.05
## Residual Sum of Squares: 468.75
## R-Squared:      0.18058
## Adj. R-Squared: 0.061271
## F-statistic: 83.8515 on 10 and 3805 DF, p-value: < 2.22e-16
```

Estimador de efectos aleatorios

```
Random = plm(lwage~educ+black+hisp+exper+exper2+married+union+yr,
             data=panel.Education, model="random")
summary(Random)
```

```
## Oneway (individual) effect Random Effect Model
##    (Swamy-Arora's transformation)
##
## Call:
## plm(formula = lwage ~ educ + black + hisp + exper + exper2 +
##     married + union + yr, data = panel.Education, model = "random")
##
## Balanced Panel: n = 545, T = 8, N = 4360
##
## Effects:
##                   var std.dev share
## idiosyncratic 0.1232  0.3510 0.539
## individual    0.1054  0.3246 0.461
## theta: 0.6429
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -4.567162 -0.144197  0.022999  0.189966  1.551817
##
## Coefficients:
##               Estimate Std. Error z-value  Pr(>|z|)
## (Intercept)  0.0235864  0.1506683  0.1565 0.8756034
## educ         0.0918763  0.0106597  8.6190 < 2.2e-16 ***
## black       -0.1393767  0.0477228 -2.9205 0.0034942 **
## hisp         0.0217317  0.0426063  0.5101 0.6100100
## exper        0.1057545  0.0153668  6.8820 5.902e-12 ***
## exper2      -0.0047239  0.0006895 -6.8513 7.319e-12 ***
## married      0.0639860  0.0167742  3.8145 0.0001364 ***
## union        0.1061344  0.0178539  5.9446 2.771e-09 ***
## yr1981       0.0404620  0.0246946  1.6385 0.1013184
## yr1982       0.0309212  0.0323416  0.9561 0.3390320
## yr1983       0.0202806  0.0415820  0.4877 0.6257435
## yr1984       0.0431187  0.0513163  0.8403 0.4007666
## yr1985       0.0578155  0.0612323  0.9442 0.3450682
## yr1986       0.0919476  0.0712293  1.2909 0.1967494
## yr1987       0.1349289  0.0813135  1.6594 0.0970420 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    656.78
```

```
## Residual Sum of Squares: 538.16
## R-Squared:      0.18062
## Adj. R-Squared: 0.17798
## Chisq: 957.774 on 14 DF, p-value: < 2.22e-16
```

Otras estimaciones:

Vamos a hacer 2 regresiones más: estimar un modelo Pooled con individual & time effects y estimar un modelo de variable binaria. En el primer caso, incluires una dummy para cada periodo de tiempo y una dummy por cada individuo (de esta forma, estaremos controlando por aquellos factores específicos de cada individuo que no cambian en el tiempo). En el segundo caso, únicamente incluiremos una dummy por cada individuo, llegando a resultados muy similares a los obtenidos en el modelo de Efectos Fijos.

#Individual & Time Effects

```
Ind.Plus.Time = lm(lwage~educ+black+hisp+exper+exper2+married+union+yr+factor(nr),data=panel.Education)
```

#Variable Binaria

```
Binary = lm(lwage~educ+black+hisp+exper+exper2+married+union+factor(nr),data=panel.Education)
```

Presentación de resultados:

Para la presentación de resultados es usual emplear la función stargazer, debido a que permite automáticamente generar tablas de alta calidad para ser publicadas. Una de las ventajas del comando stargazer es que permite exportar la tabla, ya sea en formato texto como en formato LaTex para ser insertada directamente en un documento pdf.

```
#Presentación de resultados.
stargazer(Pooled, Random, Fixed, FD,Ind.Plus.Time,Binary, type="text",
          title = "Estimaciones de los modelos de datos panel",
          column.labels=c("OLS","RE","FE", "PD","Ind+Time", "Binary"),keep.stat=c("n","rsq"), style = '
          omit = c("nr","yr1981", "yr1982", "yr1983", "yr1984", "yr1985", "yr1986", "yr1987"))
```

```
##
## Estimaciones de los modelos de datos panel
## ===============================================================================
##                                         lwage
##                                panel                           OLS
##                               linear
##                 OLS       RE        FE        PD     Ind+Time   Binary
##                 (1)       (2)       (3)       (4)       (5)       (6)
## -------------------------------------------------------------------------------
## educ         0.091***  0.092***                      -0.489***   0.033
##               (0.005)   (0.011)                       (0.091)   (0.035)
##
## black        -0.139*** -0.139***                      1.238***  0.849***
##               (0.024)   (0.048)                       (0.186)   (0.153)
##
## hisp           0.016     0.022                         0.220   0.614***
##               (0.021)   (0.043)                       (0.147)   (0.154)
##
## exper        0.067***  0.106***  0.132***             -0.524*** 0.117***
##               (0.014)   (0.015)   (0.010)             (0.110)   (0.008)
##
## exper2       -0.002*** -0.005*** -0.005*** -0.006*** -0.005*** -0.004***
##               (0.001)   (0.001)   (0.001)   (0.002)   (0.001)   (0.001)
```

```
##
## married      0.108***  0.064***   0.047**   0.038*    0.047**   0.045**
##               (0.016)   (0.017)   (0.018)   (0.023)   (0.018)   (0.018)
##
## union        0.182***  0.106***  0.080***   0.041**  0.080***  0.082***
##               (0.017)   (0.018)   (0.019)   (0.020)   (0.019)   (0.019)
##
## Constant       0.092     0.024             0.140***  8.302***    0.364
##               (0.078)   (0.151)            (0.029)   (1.304)   (0.416)
##
## Observations   4,360     4,360     4,360     3,815     4,360     4,360
## R2             0.189     0.181     0.181     0.006     0.621     0.620
## ------------------------------------------------------------------------
## Notes:       ***Significant at the 1 percent level.
##              **Significant at the 5 percent level.
##              *Significant at the 10 percent level.
```

Como los coeficentes de las variables dummies temporales e individuales generalmente no son de interés, se omiten de la tabla mediante *omit = c("nr","yr1981","yr1982","yr1983","yr1984","yr1985","yr1986","yr1987")*

La tabla generada por stargazer para el modelo exportada en LaTex se obtiene cambiando el argumento **type** a LaTex así: type="latex"

Elección del modelo:

Aquí es importante justificar que probar si el supuesto de exogeneidad se satisface es díficil, por lo cual lo que nos puede decir si el término de heterogeneidad inobservable se correlaciona con las regresoras (o no) es la teoría económica y el conocimiento del tema de estudio que estamos analizando, pues no es correcto apoyarse únicamente en los tests, y viceversa.

Prueba Breush-Pagan

La prueba Breush-Pagan es una de las muchas pruebas estándar para verificar si los errores de un modelo presentan o no heterocedasticidad. Tiene sentido aplicar la prueba directamente sobre los resultados del estimador de mínimos cuadrados combinados dado que este estimador es igual que el de mínimos cuadrados ordinarios, salvo que tiene en cuenta la componente temporal de los datos panel además de la componente de corte transversal.

```
#Prueba Breush-Pagan ( Ho: Homocedasticidaad)
bptest(Pooled) # si p-value>5% posiblemente es Pooled
```

```
##
##  studentized Breusch-Pagan test
##
## data:  Pooled
## BP = 28.22, df = 14, p-value = 0.0133
```

De los resultados de la prueba se rechaza la hipótesis nula que el error compuesto $v_{it} = a_i + u_{it}$ sea homocedástico. Por lo anterior, se debe explorar otros modelos de datos panel.

Prueba de multiplicadores de Lagrange

Luego, se aplica una prueba de multiplicadores de Lagrange, el cual parte del supuesto de que no hay correlación entre el término de heterogeneidad inobservable que es fijo en el tiempo y las regresoras (si se cree que dicho supuesto no se satisface, es mejor mirar otro test, pues tanto EA como Pooled no son adecuados, todo depende, de la naturaleza del fenómeno que se esté estudiando). En este caso, si suponemos que no hay

correlación entre a_i y alguna regresora, rechazamos la hipótesis nula (es mejor el Pooled) en favor de la alternativa (Es mejor EA por eficiencia).

```
#Pooled VS Efectos fijos
#Prueba de Multiplicadores de Lagrange de Breusch-Pagan para E.A
plmtest(Random,type = "bp")  #Ho:Mejor Pooled porque var(ai)=0
```

```
##
##  Lagrange Multiplier Test - (Breusch-Pagan) for balanced panels
##
## data:  lwage ~ educ + black + hisp + exper + exper2 + married + union +  ...
## chisq = 3203.6, df = 1, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

```
                              #H1:Se prefiere EA
```

#Test de significancia conjunta de los efectos individuales, temporales o ambos. El propósito de este test es analizar si los factores específicos de cada uno de los individuos (por los cuales estamos controlando) son significativos, es decir, explican la variable dependiente. De igual forma, cuando probamos si los efectos temporales son significativos, se quiere analizar si al controlar por aquellas variables que son comúnes a todos los individuos pero cambian en el tiempo, tienen algún efecto sobre la variable dependiente. En el último caso, se tienen los 2 argumentos ya mencionados.

```
# ¿El efecto es individual, temporal, o ambos?
plmtest(Pooled,"time","bp")     #Ho:Efectos temporales no significativos
```

```
##
##  Lagrange Multiplier Test - time effects (Breusch-Pagan) for balanced
##  panels
##
## data:  lwage ~ educ + black + hisp + exper + exper2 + married + union +  ...
## chisq = 4.0074, df = 1, p-value = 0.0453
## alternative hypothesis: significant effects
```

```
plmtest(Pooled,"individual","bp")     #Ho:Efectos indivuduales no significativos
```

```
##
##  Lagrange Multiplier Test - (Breusch-Pagan) for balanced panels
##
## data:  lwage ~ educ + black + hisp + exper + exper2 + married + union +  ...
## chisq = 3203.6, df = 1, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

```
plmtest(Pooled,"twoways","bp")     #Ho:Efectos temporales e individuales no significativos
```

```
##
##  Lagrange Multiplier Test - two-ways effects (Breusch-Pagan) for
##  balanced panels
##
## data:  lwage ~ educ + black + hisp + exper + exper2 + married + union +  ...
## chisq = 3207.6, df = 2, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

Test de Hausman

Posteriormente, se aplica el test de Hausman. Este test debe ser un complemento al análisis teórico que se plantea al hacer una regresión, pues es la teoría y el razonamiento el que nos da indicios de si pueden haber factores específicos de los individuos (que no cambien en el tiempo) que se puedan correlacionar con

las variables regresoras. Aún así, el test de Hausman nos permite comparar entre EF y EA. La hipótesis nula del test es que ambos modelos son equivalentes y se prefiere EA por eficiencia, mientras que la hipótesis alternativa indica que se prefieren EF.

```
#¿es mejor Efectos fijos o Efectos Aleatorios?
#Test de Hausman para comparar EF vs EA
phtest(Fixed, Random) #Ho: Los modelos son equivalentes estadisticamente, por eficiencia elijo EA.
```

```
##
##  Hausman Test
##
## data:  lwage ~ educ + black + hisp + exper + exper2 + married + union +  ...
## chisq = 31.707, df = 10, p-value = 0.000448
## alternative hypothesis: one model is inconsistent
```

La hipótesis nula del test de Hausman dice que los estimadores de efectos fijos y efectos aleatorios son estadísticamente equivalentes, lo que quiere decir que los coeficientes de los dos estiamdores que se obtiene son muy parecidos y no presentan diferencias estadísticamente significativas. No obstante, bajo los supuestos el modelo de efectos aleatorios el estimador de efectos aleatorios es más eficiente que el de efectos fijos. por lo que si no se rechaza la hipótesis nula se escogería el modelo de efectos aleatorios.

En tanto se rechazí la hipótesis nula en favor de la alternativa, entonces se preferiría el estimador de efectos fijos sobre el de efectos aleatorios. Este resultado, no obstante, debe analizarse en relación al análisis teórico del fenómeno que estamos estudiando.

Un link que explica el test de Hausman se encuentra en el siguiente enlace: https://www.youtube.com/watch?v=54o4-bN9By4

Test de primeras diferencias de Wooldridge

Aplicamos este test cuando (bien sea por cuestiones teóricas y/o por el test de Hausman) se cree que hay correlación entre el factor de heterogeneidad inobservable que está fijo en el tiempo y las variables regresoras (una o más). Con base en esto, se contrastará si el estimador de PD o EF es más eficiente, debido al controlar por el término a_i, ambos son consistentes. Este análisis se llevará a cabo analizando el comportamiento de los residuales de cada modelo.

Cuando se prefiere el estimador de primeras diferencias sobre el de efectos fijos Teóricamente el estimador de primeras diferencias se prefiere cuando el error idiosincrática (en nivel) u_{it} se comporta como una caminata aleatoria. Lo anterior indica que dicho término de error se describe por un proceso:

$$u_{it} = u_{it-1} + r_{it} \text{donde } r_{it} \text{ es un error que no está correlacionado con ningún regresor } x_{itj} \tag{2}$$

Lo anterior es como se supone opera el proceso de generación de datos asociado al término de error. El error r_{it} no tiene importancia en la estimación siempre que no esté correlacionado con ningún regresor x_{itj}, de manera que se pone ahí para ilustrar qué quiere decir que el término de error idiosincrático u_{it} se comporte como una caminata aleatoria[2].

Al diferenciar la ecuación se obtendría que Δu_{it} deja de tener correlación serial, es decir, $cov(\Delta u_{it}, \Delta u_{is} = 0, \quad \forall t \neq s$, por lo que el estimador de primera diferencia sería más eficiente que el de efectos fijos.

Justificación de lo anterior:

[2]En la tercera parte del curso relacionada con el tema de series de tiempo se profundizará sobre caminatas aleatorias

$$\Delta u_{it} = u_{it} - u_{it-1}$$
$$\Delta u_{it} = (u_{it-1} + r_{it}) - (u_{it-1})$$
$$\Delta u_{it} = r_{it}$$

Por lo que $\Delta u_{it} = r_{it}$ y si r_{it} no presenta correlación serial, que es lo usual, entonces Δ_{it} tampoco presenta correlación serial y el estimador de primeras diferencias es más eficiente[3].

Cuando se prefiere el estimador de efectos fijos sobre el de primeras diferencias Teóricamente, el estimador de efectos fijos se prefiere cuando el error idiosincrática u_{it} no presenta correlación serial, es decir cuando $cov(u_{it}, u_{is} = 0, \quad \forall t \neq s$. Lo anterior quiere decir que siempre que cumpla el supuesto EF-6 para el estimador de efectos fijos, se debería usar este modelo, dado que es más eficiente que el modelo de primeras diferencias. Cuando el error idiosincrático u_{it} no presenta correlación serial, los errores Δu_{it} del estimador de primeras diferencias sí presentan correlación serial, y por este motivo, el estimador es menos eficiente que el de efectos fijos.

```
#Test de Primeras diferencias de Wooldridge para comparar EF vs PD
pwfdtest(lwage~exper2+married+union+yr, data=panel.Education,h0= "fe") #H0 = corr(Uij,Uij-1) = 0
```

Test de Wooldridge en R

```
##
##  Wooldridge's first-difference test for serial correlation in panels
##
## data:  plm.model
## F = 23.412, df1 = 1, df2 = 3268, p-value = 1.368e-06
## alternative hypothesis: serial correlation in original errors
```

```
pwfdtest(lwage~exper2+married+union+yr, data=panel.Education,h0= "fd") #H0 = errores diferenciados no c
```

```
##
##  Wooldridge's first-difference test for serial correlation in panels
##
## data:  plm.model
## F = 347.71, df1 = 1, df2 = 3268, p-value < 2.2e-16
## alternative hypothesis: serial correlation in differenced errors
```

```
#La prueba no es concluyente, tanto los erróres diferenciados como sin diferenciar tienen correlación s
```

De los resultados anteriores, ambas hipótesis nulas se rechazan (la primera hipítesis nula asociada a efectos fijos decía que $corr(U_{it}, U_{it-1}) = 0$ mientras que la segunda asociada a primeras diferencias decía que $corr(\Delta U_{it}, \Delta U_{it-1}) = 0$). Como se rechazan ambas hipótesis nulas para ambos modelos el test de Wooldridge no es concluyente y no es posible saber cuál de los dos modelos usar, dado que tanto el error idiosincráticatico como sus diferencias presentan correlación serial.

No obstante, ante este tipo de situaciones y en el desarrollo de política pública el estimador comúnmente más utilizado es el estimador de efectos fijos.

[3]La anterior demostración no se las van a preguntar en clase pero es para que sepan qué quiere decir que u_{it} sea una caminata aleatoria y por qué bajo esas circunstancias el estimador de primeras diferncias es más eficiente

Validación de supuestos:

Pruebas de heterocedasticidad y de autocorrelación serial

Prueba Breusch Pagan para heterocedasticidad Nuevamente se emplea la prueba de Breusch Pagan para la heterocedasticidad de los estimadores de efectos fijos, mínimos cuadrados combinados y efectos aleatorios. Lo anterior tiene sentido, dado que todos son varaciones de estimadores de mínimos cuadrados y para el caso de efectos aleatorios es una variación de mínimos cuadrados generalizados factibles.

```
#Prueba de heterocedasticidad
bptest(Pooled);bptest(Random);bptest(Fixed); bptest(FD)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  Pooled
## BP = 28.22, df = 14, p-value = 0.0133
```

```
##
##  studentized Breusch-Pagan test
##
## data:  Random
## BP = 28.22, df = 14, p-value = 0.0133
```

```
##
##  studentized Breusch-Pagan test
##
## data:  Fixed
## BP = 28.22, df = 14, p-value = 0.0133
```

```
##
##  studentized Breusch-Pagan test
##
## data:  FD
## BP = 28.22, df = 14, p-value = 0.0133
```

Al rechazarse la hipótesis nula para todos los modelos se concluye que todos los residuales de los modelos presentan heterocedasticidad, y por ende, se debe corregir sus errores estándar con errores robustos a la heterocedastidad

Prueba Breusch-Godfrey para autocorrelación de orden p Se emplea la prueba Breusch-Godfrey para la autocorrelación serial de los errores de los modelos de efectos fijos, mínimos cuadrados combinados y efectos aleatorios. Lo anterior tiene sentido, dado que todos son varaciones de estimadores de mínimos cuadrados y para el caso de efectos aleatorios es una variación de mínimos cuadrados generalizados factibles.

```
#Test Breusch-Godfrey para autocorrelación de orden p
bgtest(Pooled);bgtest(Random);bgtest(Fixed);bgtest(FD)
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  Pooled
## LM test = 1097, df = 1, p-value < 2.2e-16
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  Random
## LM test = 1097, df = 1, p-value < 2.2e-16
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  Fixed
## LM test = 1097, df = 1, p-value < 2.2e-16
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  FD
## LM test = 1097, df = 1, p-value < 2.2e-16
```

Al rechazarse la hipótesis nula para todos los modelos se concluye que los errores de los modelos presentan heterocedasticidad, y por ende, se debe corregir sus errores estándar con errores robustos a la correlación serial

Corrección de los errores estándar mediante errores robustos a la heterocedasticidad y a la autocorrelación serial La función **vcovHC** permite calcular la matriz de varianzas y covarianzas para los errores robustos. Se emplea el método arellano, en tanto es el más utilizado, para calcular la matriz de varianzas y covarianzas de los errores robustos tanto a heterocedasticidad como correlación serial.

```
#Correción de correlación serial para EF.
MCOV=vcovHC.plm(Fixed, method=c("arellano"))
MCOV1=vcovHC(Fixed, method="arellano")
coeftest(Fixed,MCOV)
```

```
##
## t test of coefficients:
##
##            Estimate  Std. Error t value  Pr(>|t|)
## exper     0.13214642  0.01198325 11.0276 < 2.2e-16 ***
## exper2   -0.00518550  0.00080857 -6.4132   1.6e-10 ***
## married   0.04668036  0.02096046  2.2271 0.0260011 *
## union     0.08000186  0.02269615  3.5249 0.0004286 ***
## yr1981    0.01904479  0.02267976  0.8397 0.4011146
## yr1982   -0.01132198  0.02117290 -0.5347 0.5928613
## yr1983   -0.04199552  0.02046635 -2.0519 0.0402448 *
## yr1984   -0.03847088  0.02112844 -1.8208 0.0687142 .
## yr1985   -0.04324982  0.01755864 -2.4632 0.0138156 *
## yr1986   -0.02738194  0.01618458 -1.6919 0.0907557 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
coeftest(Fixed,MCOV1)
```

```
##
## t test of coefficients:
##
##            Estimate  Std. Error t value  Pr(>|t|)
## exper     0.13214642  0.01198325 11.0276 < 2.2e-16 ***
## exper2   -0.00518550  0.00080857 -6.4132   1.6e-10 ***
## married   0.04668036  0.02096046  2.2271 0.0260011 *
## union     0.08000186  0.02269615  3.5249 0.0004286 ***
## yr1981    0.01904479  0.02267976  0.8397 0.4011146
## yr1982   -0.01132198  0.02117290 -0.5347 0.5928613
## yr1983   -0.04199552  0.02046635 -2.0519 0.0402448 *
```

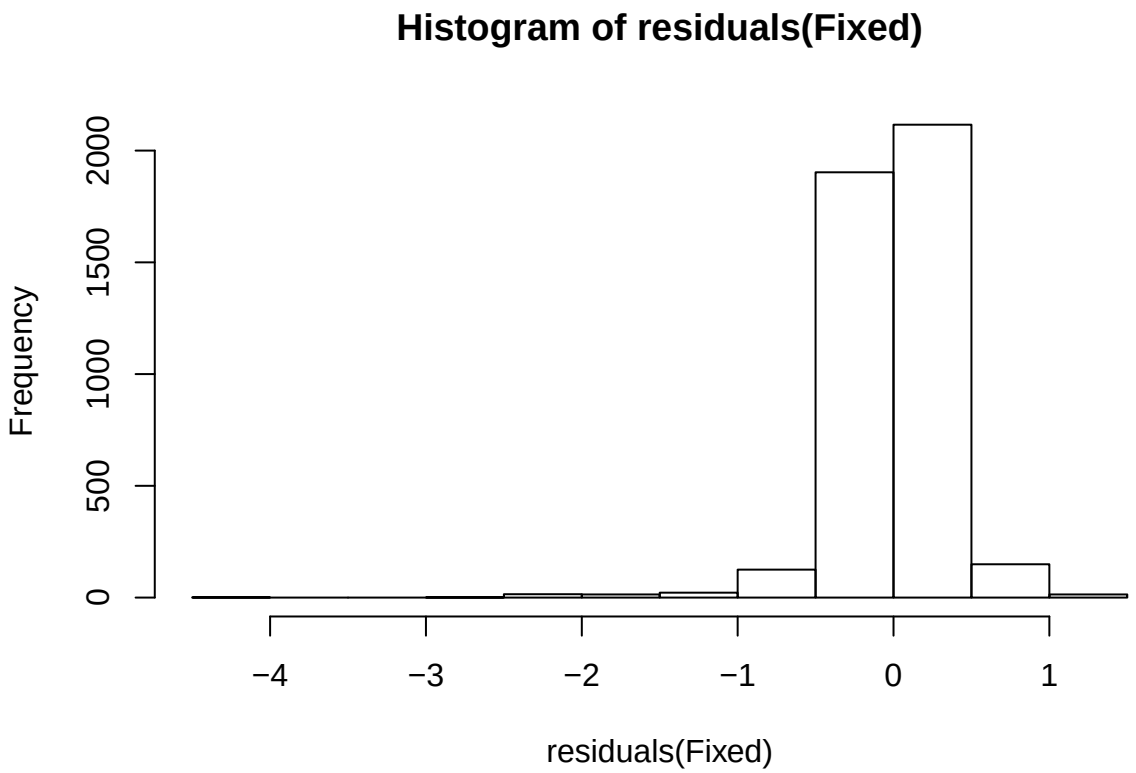
```
## yr1984  -0.03847088  0.02112844 -1.8208 0.0687142 .
## yr1985  -0.04324982  0.01755864 -2.4632 0.0138156 *
## yr1986  -0.02738194  0.01618458 -1.6919 0.0907557 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#help("vcovHC.plm")
```

Se usa **coeftest** para calculasr los errores estándar y estadísticios de prueba asociados a los errores robustos calculados por medio de **vcovHC**. Con estos estos errores recalculados, la inferencia estadística ahora sí es válida.

Verificación de normalidad en los errores En primer lugar, se puede realizar un histograma de los residuales para ver si estos tienen un comportamiento parecido al que se esperaría de una distribución normal.

```
#Análisis de Normalidad.
hist(residuals(Fixed))
```

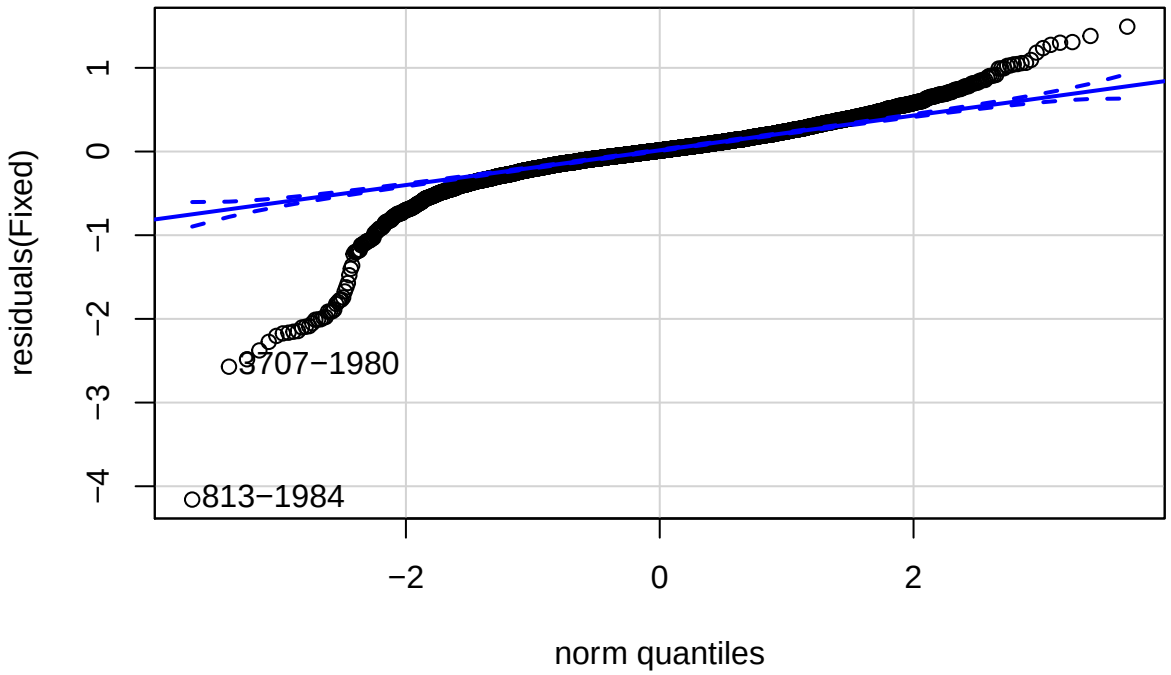


No obstante, ese análisis no siempre es muy precisio por lo que podría usar una qq plot para ver la normalidad de los residuales.

```
library(car)
```

```
## Loading required package: carData
```

```
qqPlot(residuals(Fixed))
```




```
##  813-1984 3707-1980
##       325      1857
```

```
#Es deseable que se ajusten a la linea de tendencia para cumplir normalidad
```

Una gráfica tipo QQ-Plot permite comparar el comportamiento/distribución de los residuales, respecto a una distribución normal teórica. Es decir, se comparan los cuantiles teóricos con los muestrales. En tanto los tests de normalidad clásicos no siempre son confiables porque estos asumen el supuesto de independencia, lo cual no es muy cierto en economía.

Gráficamente puede verse si los residuales se distribuyen normal en la medida en que los datos se agrupen hacia el centro sobre la linea punteada. Asimismo, la distribución de los datos debe ser más o menos simétrica. Finalmente, los puntos deben estar menos concentrados en las colas de la distribución, pues se busca que los datos no se alejen demasiado de la linea punteada.

Por último, está el test de jarque bera que es un test que permite comprobar si los residuales tienen o no un comportamiento normal.

```
jarque.bera.test(residuals(Fixed))
```

```
##
##  Jarque Bera Test
##
## data:  residuals(Fixed)
## X-squared = 52507, df = 2, p-value < 2.2e-16
```

Cómo el test rechaza la hipótesis nula de normalidad en los errores se concluye por el test de jarque bera que aquellos no son normales.

Se debe resaltar que los errores estándar y la inferencia estadística es **asintoticamente** válida así los errores no sean normales, sabiendo que el término **asintoticamente** quiere decir cuando la muestra crece, es decir cuando N, el número de observaciones, tiende a infinito. En estos casos se podría asumir otro supuesto de distribución para los errores, ajustar por valores atípicos o utilizar inferencia randomizada (no depende de la distribución de los datos)

Conclusión

- En la práctica se suele estimar los cuatro modelos para datos panel: mínimos cuadrados combinados, estimador de efectos fijos, estimador de primeras diferncias y estimador de efectos aleatorios.
- No obstante, se suelen utilizar los mútiples criterios expuestos previamente arriba para seleccionar el estimador más conveniente dadas las características de los datos y de la ecuación que se desea estimar. Aún así, recuerden que el supuesto de exogeneidad es un supuesto, por lo que es difícil de probar, de manera que la respuesta está en gran medida en el análisis teórico que se lleve a cabo.

- Además, se deben aplicar las pruebas propuestas para hacer validación de supuestos sobre el modelo seleccionado. Dicha validación de supuestos implica analizar si los errores idiosincráticos u_{it} (o sus diferencias Δu_{it}) son homocedásticos, no presentan correlación serial y son normales.

Ejercicios para la casa:

Queda como tarea completar el siguiente ejercicio:

```
#OTRO EJEMPLO:-------------------------------------------------------------------------------

#En este ejercicio se utiliza la base de datos JTRAIN.RAW para determinar el efecto del
#subsidio a la capacitación laboral en las horas de capacitación por empleado. El modelo
#básico para los tres años es:
#hrsemp~ Bo + S1d88 + S2d89+ B1grant + B2grant_1 + B3lemploy + ai + uit

#Utilizaremos la base de datos Jtrain
data("jtrain")
attach(jtrain)
```

```
## The following objects are masked from Education:
##
##     union, year
```

```
##Definimos el objeto como un panel de datos
panel.jtrain = pdata.frame(jtrain, index = c("fcode","year"))

#-----------------------------------------------------------------------------------------------
#QUEDA COMO EJERCICIO REALIZAR LAS PRUEBAS RESPECTIVAS PARA COMPARAR LOS MODELOS,Así
#COMO LA RESPECTIVA VALIDACIÓN DE SUPUESTOS Y SU RESPECTIVA CORRECCIÓN,CUANDO SEA NECESARIO.
#-----------------------------------------------------------------------------------------------
```