Problem_Set_4

Kristopher C. Toll

April 18, 2018

Theory

Problem 1

Part a

Having aggregate time effects will allow the model to keep track of time changes and shocks correlated with the other regressors. These could include GDP and other business cycle trends.

Part b

v_i could include individual characteristics about the county that are left out of the model. These may be policy differences between counties or what kind of workforce is available, which would make the $\text{cov}(v_i, x_{it}) \neq 0$.

Part c

I would argue that the sign for δ is negative. As the tax rate increases, investments should decrease. Manufactures would seek other counties that are cheaper to produce in.

Part d

It is possible that the $\text{cov}(v_i, x_{it}) \neq 0$. As such, it would be best to run a fixed effects model to account for individual effects not captured in the model. A two-way random effects may also be useful and a Huasman test would let us know which model produced better results.

Problem 2

Let, $\tilde{x}_i = x_{it} - \overline{x}_{it}$ and $\tilde{y}_i = y_{it} - \overline{y}_{it}$

$$\hat{\beta}_{FE} = (\widetilde{X}'\,\widetilde{X})^{-1}\,\widetilde{X}'\,\widetilde{X} = [\sum_{i=1}^{N}(\tilde{x}'_{i1}\,\tilde{x}_{i1} + \tilde{x}'_{i2}\,\tilde{x}_{i2})]^{-1}[\sum_{i=1}^{N}(\tilde{x}'_{i1}\,\tilde{y}_{i1} + \tilde{x}'_{i2}\,\tilde{y}_{i2})]$$

$$\tilde{x}_{i1} = x_{i1} - \overline{x}_i = x_{i1} - \frac{1}{2}(x_{i1} + x_{i2}) = -\frac{1}{2}(x_{i2} - x_{i1})$$

$$\tilde{x}_{i1} = -\frac{1}{2}(\Delta x_i)$$

$$\tilde{x}_{i2} = x_{i2} - \overline{x}_i = x_{i2} - \frac{1}{2}(x_{i1} + x_{i2}) = \frac{1}{2}(x_{i2} - x_{i1})$$

$$\tilde{x}_{i2} = \frac{1}{2}(\Delta x_i)$$

$$\tilde{y}_{i1}=-\frac{1}{2}(\Delta y_i) \\\tilde{y}_{i2}=\frac{1}{2}(\Delta y_i)$$

Thus,

$$\hat{\beta}_{FE} = [\sum_{i=1}^{N}(\tilde{x}'_{i1}\,\tilde{x}_{i1} + \tilde{x}'_{i2}\,\tilde{x}_{i2})]^{-1}[\sum_{i=1}^{N}(\tilde{x}'_{i1}\,\tilde{y}_{i1} + \tilde{x}'_{i2}\,\tilde{y}_{i2})]$$

$$\hat{\beta}_{FE} = [\sum_{i=1}^{N}((-\frac{1}{2}\Delta x_i)'(-\frac{1}{2}\Delta x_i) + (\frac{1}{2}\Delta x_i)'(\frac{1}{2}\Delta x_i))]^{-1}[\sum_{i=1}^{N}((-\frac{1}{2}\Delta x_i)'(-\frac{1}{2}\Delta y_i) + (\frac{1}{2}\Delta x_i)'(\frac{1}{2}\Delta y_i))]$$

$$\hat{\beta}_{FE} = [\sum_{i=1}^{N}(\frac{1}{2}\Delta x_i'\Delta x_i)]^{-1}[\sum_{i=1}^{N}(\frac{1}{2}\Delta x_i'\Delta y_i)] = \frac{1}{2}^{-1}\frac{1}{2}[\sum_{i=1}^{N}(\Delta x_i'\Delta x_i)]^{-1}[\sum_{i=1}^{N}(\Delta x_i'\Delta y_i)]$$

$$\hat{\beta}_{FE} = [\sum_{i=1}^{N}(\Delta x_i'\Delta x_i)]^{-1}[\sum_{i=1}^{N}(\Delta x_i'\Delta y_i)] = \hat{\beta}_{FD}$$

Application

Problem 1

Part a

```
nor_panel <- pdata.frame(norway, index = c("district", "year"))

pool_nor <- plm(log(crime) ~ d78 + clrprc1 + clrprc2, data = nor_panel, index
= c("district", "year"),model = "pooling" )
summary(pool_nor)

## Pooling Model
##
## Call:
## plm(formula = log(crime) ~ d78 + clrprc1 + clrprc2, data = nor_panel,
##     model = "pooling", index = c("district", "year"))
##
## Balanced Panel: n = 53, T = 2, N = 106
##
## Residuals:
```

```
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -1.3631874 -0.2392912  0.0053942  0.3083174  0.9618723
##
## Coefficients:
##               Estimate Std. Error t-value  Pr(>|t|)
## (Intercept)  4.1812196  0.1878879 22.2538 < 2.2e-16 ***
## d78         -0.0547246  0.0944947 -0.5791  0.563779
## clrprc1     -0.0184955  0.0053035 -3.4874  0.000721 ***
## clrprc2     -0.0173881  0.0054376 -3.1978  0.001845 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    39.906
## Residual Sum of Squares: 21.111
## R-Squared:      0.47097
## Adj. R-Squared: 0.45541
## F-statistic: 30.269 on 3 and 102 DF, p-value: 4.445e-14
```

It is evident that as crimes are solved there will be a decrease in the number of crimes committed. For every crime that was solved last year, there is approximately a 1.8% decrease in crimes committed and for two years ago, it is a 1.7% decrease.

```
pbgtest(pool_nor)
```

```
##
##  Breusch-Godfrey/Wooldridge test for serial correlation in panel
##  models
##
## data:  log(crime) ~ d78 + clrprc1 + clrprc2
## chisq = 16.635, df = 2, p-value = 0.0002442
## alternative hypothesis: serial correlation in idiosyncratic errors
```

There is serial correlation in this model. This can overstate the significance of some of the variables and lead to miscalculated stand errors.

Part b

```
fixed_nor <- plm(log(crime) ~ d78 + clrprc1 + clrprc2, data = nor_panel,
index = c("district", "year"), model = "within" )

stargazer(pool_nor, fixed_nor, title = "Panel Regressions", dep.var.labels =
"Log(crime)", column.labels = c("Pooled OLS", "Fixed Effects"), df = TRUE,
digits = 4, type = "text")
```

```
##
## Panel Regressions
## ===============================================================
##                             Dependent variable:
##              -----------------------------------------------
##                                   Log(crime)
##                     Pooled OLS             Fixed Effects
```

```
##                         (1)                      (2)
## ------------------------------------------------------------
## d78                   -0.0547                   0.0857
##                       (0.0945)                 (0.0638)
##
## clrprc1              -0.0185***                -0.0040
##                       (0.0053)                 (0.0047)
##
## clrprc2              -0.0174***                -0.0132**
##                       (0.0054)                 (0.0052)
##
## Constant             4.1812***
##                       (0.1879)
##
## ------------------------------------------------------------
## Observations            106                      106
## R2                     0.4710                   0.4209
## Adjusted R2            0.4554                   -0.2160
## F Statistic  30.2690*** (df = 3; 102) 12.1157*** (df = 3; 50)
## ============================================================
## Note:                                *p<0.1; **p<0.05; ***p<0.01
```

```
#LSDV_nor <- lm(log(crime) ~ d78 + clrprc1 + clrprc2+ factor(district), data
= nor_panel )
#summary(LSDV_nor)
```

The significance level for the clear-up % variable dropped for the prior year and the two-years prior. The two-years prior is still significant at the 95%. As there are only two time periods in this data set and a time-demeaned fixed-effect model was run, it does not make sense to test for serial correlation. The dummy variable should be enough to control for any time trends. However, the significance level could still be misrepresented as there may be heteroskedasticity between the two periods.

```
fixed_robust_nor <- coeftest(fixed_nor, vcov=vcovHC(fixed_nor, method =
"Arellano", type = "HC3"))
print(fixed_robust_nor)
```

```
##
## t-test of coefficients:
##
##           Estimate Std. Error t value Pr(>|t|)
## d78      0.0856556  0.0559941  1.5297  0.13239
## clrprc1 -0.0040475  0.0044538 -0.9088  0.36782
## clrprc2 -0.0131966  0.0048079 -2.7448  0.00839 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Even with the robust errors, we still have a 95% significance for the two-years prior clear-up %.

Part c

```
RE_nor <- plm(log(crime) ~ d78 + clrprc1 + clrprc2, data = nor_panel, index = c("district", "year"), effect = "individual", model = "random" )

stargazer(pool_nor, fixed_nor, fixed_robust_nor, RE_nor, title = "Panel Regressions", dep.var.labels = "Log(crime)", column.labels = c("Pooled OLS", "Fixed Effects","Fixed with Robust Errors","Random Effects"), df = TRUE, digits = 4, type = "text")
```

```
##
## Panel Regressions
## ================================================================================================================
##                                                  Dependent variable:
##              ------------------------------------------------------------------------------------------------
##                                  Log(crime)                                                      log(crime)
##                                    panel                                 coefficient               panel
##                                    linear                                    test                  linear
##                     Pooled OLS             Fixed Effects      Fixed with Robust Errors      Random Effects
##                        (1)                     (2)                     (3)                      (4)
## ----------------------------------------------------------------------------------------------------------------
## d78                  -0.0547                 0.0857                  0.0857                   0.0052
##                      (0.0945)               (0.0638)                (0.0560)                 (0.0592)
##
## clrprc1             -0.0185***               -0.0040                 -0.0040                 -0.0105**
##                      (0.0053)               (0.0047)                (0.0045)                 (0.0043)
##
## clrprc2             -0.0174***              -0.0132**               -0.0132***               -0.0170***
##                      (0.0054)               (0.0052)                (0.0048)                 (0.0045)
##
## Constant            4.1812***                                                               3.8118***
##                      (0.1879)                                                                (0.1869)
##
## ----------------------------------------------------------------------------------------------------------------
## Observations           106                     106                                             106
## R2                    0.4710                  0.4209                                          0.4162
## Adjusted R2           0.4554                  -0.2160                                         0.3990
## F Statistic  30.2690*** (df = 3; 102) 12.1157*** (df = 3; 50)                        24.2357*** (df = 3; 102)
## ================================================================================================================
## Note:                                                                      *p<0.1; **p<0.05; ***p<0.01
```

With the random effects model, both clear-up variables are significant at the 99% level like the pooled regression. There are a lot of 53 different districts in this dataset which could influence the crime rate. Each district could have its own demographics that behave differently from each other. It might make sense to run a Random Effects model as each district could be correlated with the regressors. However, after using a Hausman test, it is evident that the $E(v_i X_{it}) \neq 0$ and the fixed model is the preferred one.

```
phtest(RE_nor, fixed_nor)
```

```
##
##  Hausman Test
##
## data:  log(crime) ~ d78 + clrprc1 + clrprc2
## chisq = 12.183, df = 3, p-value = 0.006783
## alternative hypothesis: one model is inconsistent
```

Problem 2

Part a

All of the time-invariant regressors are education, black and Hispanic. Time-variant variables include experience, married, and whether or not one joins a union.

Part b

```
pool_wage <- plm(log(wage) ~ educ + black + hisp + exper + exper_sqr +
married + union, data = wage_panel, model = "pooling" )
summary(pool_wage)
```

```
## Pooling Model
##
## Call:
## plm(formula = log(wage) ~ educ + black + hisp + exper + exper_sqr +
##     married + union, data = wage_panel, model = "pooling")
##
## Balanced Panel: n = 545, T = 8, N = 4360
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -5.268937 -0.248691  0.033205  0.296163  2.560777
##
## Coefficients:
##                Estimate  Std. Error t-value  Pr(>|t|)
## (Intercept) -0.03470561  0.06456900 -0.5375    0.5910
## educ         0.09938779  0.00467760 21.2476 < 2.2e-16 ***
## black       -0.14384172  0.02355950 -6.1055 1.114e-09 ***
## hisp         0.01569798  0.02081119  0.7543    0.4507
## exper        0.08917906  0.01011105  8.8200 < 2.2e-16 ***
## exper_sqr   -0.00284866  0.00070736 -4.0272 5.742e-05 ***
## married      0.10766559  0.01569647  6.8592 7.897e-12 ***
## union        0.18007255  0.01712053 10.5179 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    1236.5
## Residual Sum of Squares: 1005.8
## R-Squared:      0.18659
## Adj. R-Squared: 0.18528
## F-statistic: 142.613 on 7 and 4352 DF, p-value: < 2.22e-16
```

ν_i may be uncorrelated but ζ_{it} still could be correlated with the other regressors. There are other individual characteristics that could change over time such as promotions, household size, and other obligations that take individuals ways from work which depend on time. Serial correlation may also be present.

Part c

```
RE_wage <- plm(log(wage) ~ educ + black + hisp + exper + exper_sqr + married
+ union, data = wage_panel, effect = "twoway", model = "random" )
#RE_wage2 <- plm(log(wage) ~ educ + black + hisp + exper + exper_sqr +
married + union, data = wage_panel, effect = "time", model = "random" )
#RE_wage3 <- plm(log(wage) ~ educ + black + hisp + exper + exper_sqr +
married + union, data = wage_panel, effect = "individual", model = "random" )
#summary(RE_wage)
stargazer(pool_wage, RE_wage,  title = "Panel Regressions", dep.var.labels =
"Log(crime)", column.labels = c("Pooled OLS","Random Effects"), df = TRUE,
digits = 4, type = "text")
```

```
##
## Panel Regressions
## ===============================================================
##                                  Dependent variable:
##                              ---------------------------
##                                        Log(crime)
##                               Pooled OLS   Random Effects
##                                   (1)           (2)
## ---------------------------------------------------------------
## educ                            0.0994***     0.1005***
##                                 (0.0047)      (0.0090)
##
## black                           -0.1438***    -0.1439***
##                                 (0.0236)      (0.0476)
##
## hisp                             0.0157        0.0205
##                                 (0.0208)      (0.0426)
##
## exper                           0.0892***     0.1151***
##                                 (0.0101)      (0.0089)
##
## exper_sqr                       -0.0028***    -0.0043***
##                                 (0.0007)      (0.0006)
##
## married                         0.1077***     0.0631***
##                                 (0.0157)      (0.0168)
##
## union                           0.1801***     0.1067***
##                                 (0.0171)      (0.0178)
##
## Constant                         -0.0347       -0.1051
##                                 (0.0646)      (0.1133)
##
## ---------------------------------------------------------------
## Observations                      4,360         4,360
## R2                               0.1866        0.1291
## Adjusted R2                      0.1853        0.1277
## F Statistic (df = 7; 4352)  142.6132***    92.1969***
```

```
## ================================================
## Note:                        *p<0.1; **p<0.05; ***p<0.01
```

All the variable in the pooled and random effects models produce very similar results. The magnitudes for each variable between models are different but all of them have the same level of significance across the models. The signs are also the same for each variable.

Part d

```
FE_wage <- plm(log(wage) ~ educ + black + hisp + exper + exper_sqr + married
+ union, data = wage_panel, effect = "twoway", model = "within" )
stargazer(pool_wage, RE_wage, FE_wage, title = "Panel Regressions",
dep.var.labels = "Log(crime)", column.labels = c("Pooled OLS","Random
Effects", "Fixed Effects"), df = TRUE, digits = 4, type = "text")
```

```
##
## Panel Regressions
## ===============================================================================================
##                                          Dependent variable:
##              ---------------------------------------------------------------------------
##                                             Log(crime)
##                       Pooled OLS            Random Effects            Fixed Effects
##                          (1)                     (2)                      (3)
## -----------------------------------------------------------------------------------------------
## educ                  0.0994***               0.1005***
##                        (0.0047)                (0.0090)
##
## black                 -0.1438***              -0.1439***
##                        (0.0236)                (0.0476)
##
## hisp                   0.0157                  0.0205
##                        (0.0208)                (0.0426)
##
## exper                 0.0892***               0.1151***
##                        (0.0101)                (0.0089)
##
## exper_sqr             -0.0028***              -0.0043***               -0.0052***
##                        (0.0007)                (0.0006)                 (0.0007)
##
## married               0.1077***               0.0631***                0.0467**
##                        (0.0157)                (0.0168)                 (0.0183)
##
## union                 0.1801***               0.1067***                0.0800***
##                        (0.0171)                (0.0178)                 (0.0193)
##
## Constant               -0.0347                 -0.1051
##                        (0.0646)                (0.1133)
##
## -----------------------------------------------------------------------------------------------
## Observations            4,360                   4,360                    4,360
## R2                      0.1866                  0.1291                   0.0216
## Adjusted R2             0.1853                  0.1277                   -0.1209
## F Statistic  142.6132*** (df = 7; 4352) 92.1969*** (df = 7; 4352) 27.9590*** (df = 3; 3805)
## ===============================================================================================
## Note:                                                          *p<0.1; **p<0.05; ***p<0.01
```

Under the fixed effects model, Marriage and being part of a union are still significant contributors to wage but the actual impact that they have is much less. Experience is a

redundant variable because as it is demeaned we are letting that distribution of experience be symmetric at zero for each individual. As such, it is not possible to determine the average effect that experience and education will have in the in the regression. Those variable are uniformly distributed with zero as the mean. The average effect will be zero.

part e

```
FE_wage_interactions <- plm(log(wage) ~ educ*d81 + educ*d82 + educ*d83 +
educ*d84 + educ*d85 + educ*d86 + educ*d87 + black + hisp + exper + exper_sqr
+ married + union, data = wage_panel, effect = "twoway", model = "within" )
summary(FE_wage_interactions)
```

```
## Twoways effects Within Model
##
## Call:
## plm(formula = log(wage) ~ educ * d81 + educ * d82 + educ * d83 +
##     educ * d84 + educ * d85 + educ * d86 + educ * d87 + black +
##     hisp + exper + exper_sqr + married + union, data = wage_panel,
##     effect = "twoway", model = "within")
##
## Balanced Panel: n = 545, T = 8, N = 4360
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -4.154622 -0.123038  0.010014  0.154883  1.494760
##
## Coefficients:
##              Estimate  Std. Error t-value  Pr(>|t|)
## exper_sqr -0.00604370  0.00086326 -7.0010 2.992e-12 ***
## married    0.04743368  0.01832770  2.5881  0.009688 **
## union      0.07897592  0.01933281  4.0851 4.497e-05 ***
## educ:d81   0.00499062  0.01222205  0.4083  0.683055
## educ:d82   0.00165097  0.01233043  0.1339  0.893494
## educ:d83  -0.00266214  0.01250977 -0.2128  0.831491
## educ:d84  -0.00982569  0.01275928 -0.7701  0.441299
## educ:d85  -0.00921446  0.01307215 -0.7049  0.480920
## educ:d86  -0.01213816  0.01344189 -0.9030  0.366578
## educ:d87  -0.01578915  0.01386795 -1.1385  0.254969
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    479.09
## Residual Sum of Squares: 468.3
## R-Squared:      0.022505
## Adj. R-Squared: -0.12188
## F-statistic: 8.74424 on 10 and 3798 DF, p-value: 2.59e-14
```

Education will be constant for each individual once they are out of school. For every year that they are out of school, the value of their education will decrease and the effect it will

have on wage will decrease. You could say that the value that education will have on your wage in the future will need to be discounted from the year the individual graduates.

part f

```
wage_panel$union_1 <- lead(wage_panel$union)
FE_wage_ULead <- plm(log(wage) ~ educ + black + hisp + exper + exper_sqr +
married + union + union_1, data = wage_panel, effect = "twoway", model =
"within")
summary(FE_wage_ULead)

## Twoways effects Within Model
##
## Call:
## plm(formula = log(wage) ~ educ + black + hisp + exper + exper_sqr +
##     married + union + union_1, data = wage_panel, effect = "twoway",
##     model = "within")
##
## Balanced Panel: n = 545, T = 7, N = 3815
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -4.052132 -0.116656  0.011444  0.150119  1.479028
##
## Coefficients:
##              Estimate  Std. Error t-value  Pr(>|t|)
## exper_sqr -0.00544481  0.00087715 -6.2074 6.066e-10 ***
## married    0.04487780  0.02088166  2.1491 0.0316960 *
## union      0.07635541  0.02176720  3.5078 0.0004579 ***
## union_1    0.04973565  0.02236180  2.2241 0.0262076 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    424.86
## Residual Sum of Squares: 416.43
## R-Squared:      0.019826
## Adj. R-Squared: -0.14674
## F-statistic: 16.4853 on 4 and 3260 DF, p-value: 2.2212e-13
```

The union lead variable is significant at the 90% level. It seems that if an individual is considering entering a union and perhaps displaying this interest to his or her managers, they may experience an increase in wage by 4%.

Problem 3

Part a

The main purpose of this article was to determine if education is truly a precursor to democracy. Past literature and political thought suggest that having a higher level of educational attainment will lead to more democratic governments. Past models have also

supplied empirical evidence for that realm of thought. However, it is possible that some of those models suffer from omitted bias. The article, "From Education to Democracy" explores what happens when variables to control for omitted bias are introduced into the model. This is done to see if education still increases democracy.

Part b

The authors of "From Education to Democracy" find that when fixed effects on countries are included in their models, the effect of education does not have a significant effect on how democratic a country is. Past literature only examined a cross-sectional correlation between education and democracy as opposed to within variation. Their question about education and democracy was thus phrased as "…[is] a given country (with its other characteristics held constant) more likely to become more democratic as its population becomes more educated?" When the within-country variation is included in the models, there is no clear causal relationship between education and democracy. However, it is possible that there are long-term effects of education on democracy as the authors only looked at short-term effects.

Part c

Country data was obtained at five-year intervals for 108 countries from 1960 to 2000. This was done to do avoid serial correlation that would have come from averaging across the five-year intervals. Average years of schooling where obtain from the total population of 25 and older. The average years of schooling range from 0.04 to 12.18. To measure democracy, The Freedom House Political Rights Index was used to. This index ranges from 1 being least democratic to 7 being the most. The authors transformed this variable so that it lies on a 0 to 1 scale.

Part d

Equation one is $d_{it} = \alpha d_{i,t-1} + \gamma s_{i,t-1} + \mu_t + \nu_{it}$ where d_{it} represents the democracy score, s_{it} the average years of education, μ_t is the full set of time effects for common shocks and trends and ν_{it} is the error term for all other omitted factors. The education variable and the democracy variable are both lagged. This pooled model does not include any individual country effects and it would assume that the $\text{cov}(\nu_{it}, X_{it}) \neq 0$. However, the authors believe that this a false assumption to make and leads to a misleading correlation in γ

Part e

Equation two is $d_{it} = \alpha d_{i,t-1} + \gamma s_{i,t-1} + \mu_t + \delta_i + \nu_{it}$ where the only difference δ_i which is a set of dummy variables. This was included because is it believed that without it, $\text{cov}(\nu_{it}, X_{it}) \neq 0$ and thus within country variation needs to be controlled for. This is why a fixed effect model is used.

Part f

It is believed that democratic persistence in the past will influence the current country's democratic score. This is the reason why $d_{i,t-1}$ is included. It makes sense that if a country

was democratic in the past period it should persistence in the next period. Also, this will help determine if a country will return to some level of democracy.

Part g

Model	Limitations	Strengths
Pooled OlS (i)	This does not control for country effects, there are variables in v_i correlated with X_{it}	This is an easy model to compute
Fixed-effects OLS (ii)	Loss of Degrees of Freedom, we have a parameter for each country and time period. The regressor $d_{i,t-1}$ is also mechanically correlated with the error term	Controls for country variation
Arellano-Bond GMM (iii)	It is possible to over-identify regressors. The model already has a lagged variable and adding an IV may cause endogeneity	The Hansen test show that there is not an overidentifying issue and the mechanical correlation is removed
Fixed-effects OLS (iv)	This model will again have a mechanical correlation between$d_{i,t-1}$ and v_i	including a log of the population should addresses omitted time-varying variable bias
Arellano-Bond GMM (v)	Could cause endogeneity in the model	endogeneity was not found in the model and the mechanical correlation is corrected for.

Fixed-effects OLS (vi)	possible mechanical correlation	including Log GDP will also address omitted time-varying variable bias
Arellano-Bond GMM (vii)	Possible endogeneity	There is no endogeneity and mechanically correlation is removed
Fixed-effects OLS (vii)	Possible mechanical correlation	including both log GDP and Log population will remove omitted variable bias
Arellano-Bond GMM (ix)	Endogeneity may be present	there is no endogeneity and any correlation is removed between $d_{i,t-}$ and v_i

Overall, the authors find that the fixed effects models are preferred because they control for in-country variations that are left out of the pooled model. The pooled model suffers from omitted variable bias, as such, it is not reliable. The Arellano-Bond models do even more to control for heteroskedasticity standard errors.

Part h

i

Including $\log(GDP_{i,t-1})$ will show how wealth can effect democracy. More wealth, which varies with time, could imply more access to education and thus a higher democracy score.

ii

Common Shocks and trends to a country's democracy score would be left out. These could be anything from changes in policies, governments, and political regimes.

iii

I believe that the time indicators should be included. Seeing that they are jointly significant it certainly would not be wise to omit them, as a misleading conclusion could be made between education and democracy.

Problem 4

Part a

The purpose of this article is to determine the advantages on wages for those living in bigger cities. Bigger cities will have a higher cost of living and thus firms tend to pay higher wages to workers for it. However, if this solely is the case, then firms would relocate to areas where they do not have to pay high wages. There must be a benefit to producing in a city despite having to pay higher wages to workers. This paper looks at wages and how workers gain valuable experience to working in a city that benefits firm productivity. This article looks at wages because it will be informative about the productivity advantages of locating in a bigger city.

part b

One of the possible explanations why workers in big cities obtain higher wages is because they have a higher innate ability and they also need compensation for higher living expenses. The Authors instead find that workers in bigger cities have an immediate static premium and then gain more valuable experience that they can take with them when they leave the city. Also, workers with a higher productivity level may choose to live in a bigger city which will also contribute to a larger dispersion in wages. Workers in large and small cities do not necessarily have a different level of abilities but with static gains and learning advantages of living in a bigger city, they will gain higher wages.

part c

The data used in this paper comes from Spain's Continuous Sample of Employment Histories. This dataset contains longitudinal information with social security, income tax, and census records for a 4% non-stratified random sample of workers associated with Spain's Social Security. The authors create a panel with monthly observations of the working life of individuals in the sample. They can control for the individual's labor market status, the occupation, the type of contract, working hours, establishments sector of activity, and the location. Locations are separated into different urban areas in Spain.With the data, they can measure tenure and experience calculated as the number of days the worker has been employed. The dataset also contains basic individual characteristics such as gender, age, and education. It is also known which individuals have pension and unemployment benefits. The samples are restricted to men aged 18 and over with Spanish citizenship born in Spain since 1962. One of the models will include females but the participation rate for that group has been very volatile. This dataset makes it possible to follow workers through time and as they move from city to city.

part d

$$w_{ict} = \sigma_c + \mu_i + \sum_{j=1}^{C} \delta_{jc}\, e_{ijt} + x'_{it}\beta + \varepsilon_{ict}$$

The loge wage of worker i in city c at time t is the dependent variable. σ_c is the city fixed effect, μ_i is the worker fixed effect. e_{ijt} is the experience acquired by worker i in city j up until time t. x_{it} are time varying individual and job characteristics. δ_{jc} are scalars, vector β are parameters and ε_{ict} is an error term. This equation will allow for static earnings premium that come form working in a bigger city. It will control for accumulated experiean nce that comes from working in city j. Having city control effects will control for difference in cities such as size. This model is based on the assumptions that city where the worker acquired experience, will be taken with them and impact their wage where they work now. It also allows for cities with different sizes to provide different kinds of experience and for each worker to have a different level of innate ability to start with. At is point there should be a little left in the error term that would cause omitted time, individual and location varying variable bias.

part e

$$w_{ict} = \sigma_c + x_{it}'\beta + \eta_{ict}$$

This model will ignore worker heterogeneity and the benefits of working in different sized cities. This will equation will only need a pooled panel of worker and assumes that there are no individual and dynamic city effects that would be correlated with the regressors. It assumes that city fixed effects are enough. This model will work if there are not dynamic city gains and all workers are the same.

part f

$$w_{ict} = \sigma_c + \mu_i + X_{it}' + \zeta_{ict}$$

Again, we are still missing the dynamic city size premium and we do not have individual work effect in the model yet. If we apply demeaning we can account for the worker fixed effects. Different works will have different abilities.

$$(w_{ict} - \overline{w}_i) = \sum_{j=1}^{C} \sigma_c \, (l_{ict} - \bar{l}_{ic}) + (x_{it}' - \overline{x}_i)\beta + (\zeta_{ict} - \overline{\zeta}_i)$$

This model is based on the assumption that dynamic city size does not influence wage premiums $\text{cov}(x_{it}, \sum_{j=1}^{C} \sigma_{jc} \, e_{ijt}) = 0$. However, σ_c will still be unbiased if dynamic fixed effects are correlated with the other regressors. Worker fixed effects will account for unobserved work heterogeneity but without the dynamic city variable, there will still be bias in this model.

part g

Section four makes the argument that there are dynamic benefits of bigger cities and the model in section three should be used. Different cities and varying size will impact wages differently and the data set makes it possible to keep track of workers who live and move between different urban areas. Here the authors argue that there are no other unobserved

individual, city and time-varying effects that could be pulled out of the error term which would bias the results. Dynamic city gains are needed to predict.