Department of Economics UN3412
Columbia University Fall 2021

SOLUTIONS to Problem Set 5
Introduction to Econometrics
Seyhan Erden and Tamrat Gashaw
for all sections.

1. The data file rental.dta include rental prices and other variables for college towns in 1980 and in 1990. The idea is to see whether a stronger presence of students affects rental rates. The unobserved effects model is

$$\log(rent_{it}) = \beta_0 + \delta_0\, y90_t + \beta_0 \log(pop_{it}) + \beta_2 \log(avginc_{it}) + \beta_3\, pctstu_{it} + a_i + u_{it}$$

Variables needed are explained in below

Variables in RENTAL.dta

Variable	Definition
pop	City population
avginc	Average income
pctstu	Student population as a percentage of city population (during the school year)
y90	=1 for 1990, zero otherwise.

(a) Estimate the equation by pooled OLS and report the results in standard form. What do you make of the estimate on the 1990 dummy variable?

Using pooled OLS we obtain

$$\log(rent_{it}) = \underset{}{-0.569} + \underset{(0.035)}{0.262}\, y90_t + \underset{(0.023)}{0.041} \log(pop_{it}) + \underset{(0.053)}{0.571} \log(avginc_{it}) + \underset{(0.0010)}{0.0050}\, pctstu_{it}$$
$$\qquad\qquad\qquad\quad (0.535)$$

$n = 128$, $R^2 = 0.861$

The positive and very significant coefficient on *d90* simply means that, other things in the equation fixed, nominal rents grew by over 26% over the 10 year period.

(b) Interpret the sample coefficient of pctstu

The coefficient on *pctstu* means that a one percentage point increase in *pctstu* increases *rent* by half a percent (.5%). The *t* statistic of five shows that, at least based on the usual analysis, *pctstu* is very statistically significant.

(c) Are the standard errors you report in part (a) valid? Explain.

The standard errors from part (i) are not valid, unless we think *ai* does not really appear in the equation. If *ai* is in the error term, the errors across the two time periods for each city are positively correlated, and this invalidates the usual OLS standard errors and *t* statistics.

(d) Now, difference the equation and estimate by OLS. Compare your estimate of β_3 with that of part (a). Does the relative size of the student population appear to affect rental prices?
The equation estimated in differences is

$$\Delta\log(rent) = .386 + .072\ \Delta\log(pop) + .310\ \Delta\log(avginc) + .0112\ \Delta pctstu$$
$$\qquad\qquad (.037)\ \ (.088)\qquad\qquad (.066)\qquad\qquad (.0041)$$

n = 64, R^2 = .322.

Interestingly, the effect of *pctstu* is over twice as large as we estimated in the pooled OLS equation. Now, a one percentage point increase in *pctstu* is estimated to increase rental rates by about 1.1%. Not surprisingly, we obtain a much less precise estimate when we difference (although the OLS standard errors from part (i) are likely to be much too small because of the positive serial correlation in the errors within each city). While we have differenced away *ai*, there may be other unobservables that change over time and are correlated with Δ*pctstu*.

(e) Obtain the heteroskedasticity-robust standard errors for the first-differenced equation in part(d)

The heteroskedasticity-robust standard error on Δ*pctstu* is about .0029, which is actually much smaller than the usual OLS standard error (0.0041). This only makes *pctstu* even more significant (robust *t* statistic ≈ 4). Note that serial correlation is no longer an issue because we have no time component in the first-differenced equation.

(f) Estimate the model by fixed effects to verify that you get identical estimates and standard errors to those in part (d) (use "areg" and "xtreg" commands and report both results)

```
. areg  lrent y90 lpop lavginc pctstu, absorb(city) r

Linear regression, absorbing indicators                 Number of obs =     128
                                                        F(  4,    60) =  691.38
                                                        Prob > F      =  0.0000
                                                        R-squared     =  0.9827
                                                        Adj R-squared =  0.9633
                                                        Root MSE      =  .06373

------------------------------------------------------------------------------
             |               Robust
       lrent |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         y90 |   .3855214   .0487186     7.91   0.000     .2880697    .4829731
        lpop |   .0722456   .0696796     1.04   0.304    -.0671344    .2116255
     lavginc |   .3099605   .0893099     3.47   0.001     .1313141     .488607
      pctstu |   .0112033    .002936     3.82   0.000     .0053305    .0170762
       _cons |   1.409384   1.162326     1.21   0.230    -.9156128    3.734382
-------------+----------------------------------------------------------------
        city |   absorbed                                     (64 categories)
```

```
. xtset city year
       panel variable:  city (strongly balanced)
        time variable:  year, 80 to 90, but with gaps
                delta:  1 unit


. xtreg  lrent y90 lpop lavginc pctstu,fe vce(cluster city)

Fixed-effects (within) regression               Number of obs      =       128
Group variable: city                            Number of groups   =        64

R-sq:  within  = 0.9765                         Obs per group: min =         2
       between = 0.2173                                        avg =       2.0
       overall = 0.7597                                        max =         2

                                                F(4,63)            =    703.09
corr(u_i, Xb)  = -0.1297                        Prob > F           =    0.0000

                                    (Std. Err. adjusted for 64 clusters in city)
------------------------------------------------------------------------------
             |               Robust
       lrent |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         y90 |   .3855214   .0483114     7.98   0.000     .2889788     .482064
        lpop |   .0722456   .0690972     1.05   0.300    -.0658341    .2103252
     lavginc |   .3099605   .0885634     3.50   0.001     .1329806    .4869404
      pctstu |   .0112033   .0029114     3.85   0.000     .0053853    .0170214
       _cons |   1.409384   1.152597     1.22   0.226     -.893896    3.712665
-------------+----------------------------------------------------------------
     sigma_u |  .15905877
     sigma_e |  .06372873
         rho |   .8616755   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

.do file for question 2:

```
regress lrent y90 lpop lavginc pctstu,r
gen dlrent = lrent - lrent[_n-1]
gen dlpop = lpop - lpop[_n-1]
gen dlavginc = lavginc - lavginc[_n-1]
gen dpctstu = pctstu - pctstu[_n-1]
regress dlrent dlpop dlavginc dpctstu,r
areg lrent y90 lpop lavginc pctstu, absorb(city) r
xtset city year
xtreg lrent y90 lpop lavginc pctstu,fe vce(cluster city)
```

2. Use the state-level data on murder rates and executions in murder.dta for the following exercise.

(a) Consider the unobserved effects model

$$mrdrte_{it} = \eta_t + \beta_1 exec_{it} + \beta_2 unem_{it} + a_i + u_{it}$$

where η_t simply denotes different year intercepts and a_i is the unobserved state effect. If past executions of convicted murderers have a deterrent effect, what should be the sign of β_1? What sign do you think β_2 should have? Explain.

(b) Using just the years 1990 and 1993, estimate the equation from part (i) by pooled OLS. Ignore the serial correlation problem in the composite errors. Do you find any evidence for a deterrent effect?

(c) Now, using 1990 and 1993, estimate the equation by fixed effects. You may use first differencing since you are only using two years of data. Is there evidence of a deterrent effect? How strong?

(d) Compute the heteroskedasticity-robust standard error for the estimation in part (ii).

(e) Find the state that has the largest number for the execution variable in 1993. (The variable *exec* is total executions in 1991, 1992, and 1993.) How much bigger is this value than the next highest value?

(f) Estimate the equation using first differencing, dropping Texas from the analysis. Compute the usual and heteroskedasticity-robust standard errors. Now, what do you find? What is going on?

(g) Use all three years of data and estimate the model by fixed effects. Include Texas in the analysis. Discuss the size and statistical significance of the deterrent effect compared with only using 1990 and 1993.

Q#2.

(a) If there is a deterrent effect, then $\beta_1 < 0$. The sign of β_2 is not entirely obvious, although one possibility is that a better economy means less crime in general, including violent crime (such as drug dealing) that would lead to fewer murders. This would imply $\beta_2 > 0$.

(b) The pooled OLS estimates using 1990 and 1993 are

$$\underset{}{mrdrte_{it}} = \underset{(4.43)}{-5.28} - \underset{(2.14)}{2.07\, d93_t} + \underset{(.263)}{.128\, exec_{it}} + \underset{(0.78)}{2.53\, unem_{it}}$$

$N = 51,\ T = 2,\ R^2 = .102.$

There is no evidence of a deterrent effect, as the coefficient on *exec* is actually positive (though not statistically significant).

(c) The first-differenced equation is

$$\underset{}{\Delta mrdrte_i} = \underset{(.209)}{.413} - \underset{(.043)}{.104\, \Delta exec_i} - \underset{(.159)}{.067\, \Delta unem_i}$$

$n = 51,\ R^2 = .110.$

Now, there is a statistically significant deterrent effect: 10 more executions is estimated to reduce the murder rate by 1.04, or one murder per 100,000 people. Is this a large effect? Executions are relatively rare in most states, but murder rates

are relatively low on average, too. In 1993, the average murder rate was about 8.7; a reduction of one would be nontrivial. For the (unknown) people whose lives might be saved via a deterrent effect, it would seem important.

(d) The heteroskedasticity-robust standard error for $\Delta exec_i$ is .017. Somewhat surprisingly, this is well below the nonrobust standard error. If we use the robust standard error, the statistical evidence for the deterrent effect is quite strong ($t \approx -6.1$). See also Computer Exercise 13.12.

(e) Texas had by far the largest value of *exec*, 34. The next highest state was Virginia, with 11. These are three-year totals.

(f) Without Texas in the estimation, we get the following, with heteroskedasticity-robust standard errors in [·]:

$$\begin{array}{rcccc} \Delta mrdrte_i & = .413 & - .067\ \Delta exec_i & - & .070\ \Delta unem_i \\ & (.211) & (.105) & & (.160) \\ & [.200] & [.079] & & [.146] \end{array}$$

$$n = 50,\ R^2 = .013.$$

Now the estimated deterrent effect is smaller. Perhaps more importantly, the standard error on $\Delta exec_i$ has increased by a substantial amount. This happens because when we drop Texas, we lose much of the variation in the key explanatory variable, $\Delta exec_i$.

(g) When we apply fixed effects using all three years of data and all states, we get

$$\begin{array}{rcccccc} mrdrte_{it} & = 1.56\ d90_t & + & 1.73\ d93_t & - .138\ exec_{it} & + & .221\ unem_{it} \\ & (.75) & & (.70) & (.177) & & (.296) \end{array}$$

$$N = 51,\ T = 3,\ R^2 = .073.$$

The size of the deterrent effect is actually slightly larger than when 1987 is not used. However, the *t* statistic is only about –.78. Thus, while the magnitude of the effect is similar, the statistical significance is not. It is somewhat odd that adding another year of data causes the standard error on the *exec* coefficient to increase nontrivially.

3. The file pension.dta contains information on participant-directed pension plans for U.S. workers. Some of the observations are for couples within the same family, so this data set constitutes a small cluster sample (with cluster sizes of two).

(a) Ignoring the clustering by family, use OLS to estimate the model

$$pctstck = \beta_0 + \beta_1 choice + \beta_2 prftshr + \beta_3 female + \beta_4 age + \beta_5 educ + \beta_6 finc25 + \beta_7 finc35 + \beta_8 finc50 + \beta_9 finc75 + \beta_{10} finc100 + \beta_{11} finc101 + \beta_{12} wealth89 + \beta_{13} stckin89 + \beta_{14} irain89 + u$$

where the variables are defined in the data set. The variable of most interest is *choice*, which is a dummy variable equal to one if the worker has a choice in how to allocate pension funds among different investments. What is the estimated effect of *choice*? Is it statistically significant?

(b) Are the income, wealth, stock holding, and IRA holding control variables important? Explain.

(c) Determine how many different families there are in the data set.

(d) Now, obtain the standard errors for OLS that are robust to cluster correlation within a family. Do they differ much from the usual OLS standard errors? Are you surprised?

(e) Estimate the equation by differencing across only the spouses within a family. Why do the explanatory variables asked about in part (ii) drop out in the first-differenced estimation?

(f) Are any of the remaining explanatory variables in part (v) significant? Are you surprised?

Q#3

(a) The OLS estimates are

$$\widehat{pctstck} = 128.54 + 11.74\ choice + 14.34\ prftshr + 1.45\ female - 1.50\ age$$
$$(55.17)\quad (6.23)\quad (7.23)\quad (6.77)\quad (.78)$$

$$+.70\ educ - 15.29\ finc25 + .19\ finc35 - 3.86\ finc50$$
$$(1.20)\quad (14.23)\quad (14.69)\quad (14.55)$$

$$- 13.75\ finc75 - 2.69\ finc100 - 25.05\ finc101 - .0026\ wealth89$$
$$(16.02)\quad (15.72)\quad (17.80)\quad (.0128)$$

$$+ 6.67\ stckin89 - 7.50\ irain89$$
$$(6.68)\quad (6.38)$$

$n = 194,\ R^2 = .108.$

Investment choice is associated with about 11.7 percentage points more in stocks. The *t* statistic is 1.88, and so it is marginally significant.

(b) These variables are not very important. The *F* test for joint significant is 1.03. With 9 and 179 *df*, this gives *p*-value = .42. Plus, when these variables are dropped from the regression, the coefficient on *choice* only falls to 11.15.

(c) There are 171 different families in the sample.

(d) The instructor reported only the cluster-robust standard error for *choice*: 6.20. Therefore, it is essentially the same as the usual OLS standard error. This is not very

surprising, because at least 171 of the 194 observations can be assumed independent of one another. The explanatory variables may adequately capture the within-family correlation.

(e) There are only 23 families with spouses in the data set. Differencing within these families gives

$$\Delta pctstck = \underset{(10.94)}{15.93} + \underset{(15.00)}{2.28}\,\Delta choice - \underset{(16.92)}{9.27}\Delta prftshr + \underset{(21.49)}{21.55}\,\Delta female - \underset{(9.00)}{3.57}\,\Delta age$$

$$\underset{(3.43)}{-1.22}\,\Delta educ$$

$n = 23,\ R^2 = .206,\ \bar{R}^2 = -.028.$

All of the income and wealth variables, and the stock and IRA indicators, drop out, as these are defined at the family level (and therefore are the same for the husband and wife).

(f) None of the explanatory variables is significant in part (v), and this is not too surprising. We have only 23 observations, and we are removing much of the variation in the explanatory variables (except the gender variable) by using within-family differences.

Following questions will not be graded, they are for you to practice and will be discussed at the recitation:

Question 1:

U.S. airlines were deregulated in 1975, allowing them to charge whatever prices they wished and to choose routes for their flights more freely than previously. One anticipated gain from deregulations was cost reduction, to be derived in part by allowing airlines to reduce excess capacity. Baltagi, Griffin and Vadali estimate that airlines did, indeed, reduce excess capacity following deregulations[1]. Their analysis combined data on variable costs and factor shares to efficiently estimate excess capacity for 23 airlines in the years 1971-1986. Data file **deregulate.dta** contain the following variables:

Variable	Description
airline	A number indicating the airline in the observation.
pf	The price of fuel
pl	The price of labor
pm	The price of materials
reg	=1 if the observation is from the regulated period =0 otherwise
stage	Average length of the airline's flights that year
vc	Variable cost (fuel+labor+materials)
y	An index of annual passenger miles flown by the airline
year	The year of the observation

(a) Regress the log of costs on the regulation dummy, year and the natural logs of three price variables and of *stage* (i) using OLS (ii) using firm-specific fixed effects without cluster (iii) with cluster
(b) What is the interpretation of regulation dummy's coefficient?
(c) What is the interpretation of year's coefficient?
(d) Briefly explain why we can conclude that the estimated standard errors reported for OLS are probably incorrect as well as the ones in fixed effects regression without cluster errors?
(e) What does the fixed effects regression imply about the effect of deregulation on airlines' variable cost?
(f) How do you counter the objection that technical change would have reduced airline costs even without the deregulation?
(g) Add the squares of the logged regressors to the fixed effects regression in (a). What does this regression suggest about the conclusions in (e)?
(h) Are the added terms in regression (g), taken together, jointly statistically significant? Show the needed test results.
(i) Some have argued that deregulation enables airlines to better plan their flight. This could mean that more efficient flight lengths were chosen after deregulation. How does this

[1] Badi H. Baltagi, James M. Griffin, and Sharada R. Vadali, "Excess Capacity: A Permanent Characteristic of U.S. Airlines," *Journal of Applied Econometrics* 13, no.5 (1998): 645-657

affect the interpretations in (e) and (g), and how would you take this consideration into account?

a.

OLS results:

```
. reg  lvc reg lpl lpf lpm lstage year, r

Linear regression                                      Number of obs =     268
                                                       F(  6,   261) =  176.21
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.6939
                                                       Root MSE      =  .63537

------------------------------------------------------------------------------
             |               Robust
         lvc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         reg |  -.1044246   .1419933    -0.74   0.463    -.3840228    .1751736
         lpl |   .9137027    .439856     2.08   0.039     .0475846    1.779821
         lpf |  -.4192051   .2361399    -1.78   0.077    -.8841869    .0457766
         lpm |   1.673205   .8965462     1.87   0.063    -.0921797    3.438589
      lstage |    1.31977   .0511289    25.81   0.000     1.219092    1.420448
        year |  -.0688188   .0515875    -1.33   0.183    -.1703994    .0327618
       _cons |  -5.619306    3.28694    -1.71   0.089     -12.0916    .8529905
------------------------------------------------------------------------------
```

Fixed Effects WITHOUT CLUSTER:

```
. xtreg  lvc reg lpl lpf lpm lstage year, fe

Fixed-effects (within) regression               Number of obs      =       268
Group variable: airline                         Number of groups   =        23

R-sq:  within  = 0.9324                         Obs per group: min =         5
       between = 0.4710                                        avg =      11.7
       overall = 0.6358                                        max =        16

                                                F(6,239)           =    549.18
corr(u_i, Xb)  = 0.1966                         Prob > F           =    0.0000

------------------------------------------------------------------------------
         lvc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         reg |  -.0103893   .0404275    -0.26   0.797     -.090029    .0692504
         lpl |     .12939   .1013104     1.28   0.203    -.0701854    .3289654
         lpf |   .0880113   .0603803     1.46   0.146    -.0309343    .2069569
         lpm |   .3837664   .2618333     1.47   0.144    -.1320294    .8995621
      lstage |   .8636402   .0610974    14.14   0.000     .7432821    .9839984
        year |   .0458234   .0130707     3.51   0.001     .0200749     .071572
       _cons |  -4.573632   .9142783    -5.00   0.000    -6.374705   -2.772559
-------------+----------------------------------------------------------------
     sigma_u |  .74149963
     sigma_e |  .15044171
         rho |  .96046374   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(22, 239) =   200.74             Prob > F = 0.0000
```

Fixed Effects WITH CLUSTER ERRORS

```
. xtset  airline year
       panel variable:  airline (unbalanced)
        time variable:  year, 71 to 86, but with gaps
                delta:  1 unit

. xtreg  lvc reg lpl lpf lpm lstage year, fe vce(cluster airline)

Fixed-effects (within) regression               Number of obs      =       268
Group variable: airline                         Number of groups   =        23

R-sq:  within  = 0.9324                         Obs per group: min =         5
       between = 0.4710                                        avg =      11.7
       overall = 0.6358                                        max =        16

                                                F(6,22)            =    221.76
corr(u_i, Xb)  = 0.1966                         Prob > F           =    0.0000

                                (Std. Err. adjusted for 23 clusters in airline)
------------------------------------------------------------------------------
             |               Robust
         lvc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         reg |  -.0103893   .0352599    -0.29   0.771    -.0835138    .0627352
         lpl |      .12939   .1389062     0.93   0.362    -.1586838    .4174638
         lpf |   .0880113   .0784058     1.12   0.274    -.0745924    .2506149
         lpm |   .3837664   .1888784     2.03   0.054    -.0079436    .7754763
      lstage |   .8636402   .1855688     4.65   0.000     .4787941    1.248486
        year |   .0458234   .0147504     3.11   0.005      .015233    .0764138
       _cons |  -4.573632   1.777618    -2.57   0.017    -8.260187   -.8870769
-------------+----------------------------------------------------------------
     sigma_u |  .74149963
     sigma_e |  .15044171
         rho |  .96046374   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

b. Suppose that β_0 is the intercept of our multiple regression model for cost. Then, β_0 is the intercept for deregulated period and $\beta_0 + \beta_{reg}$ is the intercept for regulated period. The regulation dummy's coefficient, β_{reg}, is the difference in average cost between regulated period and deregulated period.

c. When *year* increases by 1, the coefficient tells us by how much variable cost is estimated to fall or rise. The coefficient of *year* tells us the time effect all else constant. For example, according to the OLS results, holding the other factors fixed, one more year is predicted to reduce $\ln(vc)$ by 0.068, which means 6.8% decrease in variable cost.

d. Take a close look at the results. The standard errors of the OLS estimates is bigger relative to the other estimation methods for all variables. The estimated standard

error would be incorrect if the regressors are considerably collinear. If this is the case, the variance of the OLS estimates of the coefficient of the collinear variables are quite large. Fixed effects without cluster errors are also larger than the ones with cluster because there is autocorrelation over time within the same airline since errors overtime would be correlated for each airline (but not necessarily across airlines)

e. The estimated coefficients for *reg* is -.0103, indicating that on average, the variable costs during regulated period are 1.03 percent lower than the costs during deregulated period. So the airlines' variable costs became higher as U.S. airlines were deregulated.

f. If tech change is a smooth change then it will be captured by "year" variable. But if tech change happens as suddenly as regulation then it cannot be accounted by "year" variable hence it will cause omitted variable bias.

g.

```
. xtreg  lvc reg lpl lpf lpm lstage lpl2 lpf2 lpm2 lstage2 year, fe vce(cluster
>  airline)

Fixed-effects (within) regression               Number of obs      =       268
Group variable: airline                         Number of groups   =        23

R-sq:  within  = 0.9385                         Obs per group: min =         5
       between = 0.4484                                        avg =      11.7
       overall = 0.6246                                        max =        16

                                                F(10,22)           =    158.01
corr(u_i, Xb)  = 0.2217                         Prob > F           =    0.0000

                                (Std. Err. adjusted for 23 clusters in airline)
------------------------------------------------------------------------------
             |               Robust
         lvc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         reg |   .0523131   .0262211     2.00   0.059    -.0020662    .1066923
         lpl |   .1602116   .1623721     0.99   0.335    -.1765276    .4969508
         lpf |   .0102717   .2519246     0.04   0.968    -.5121878    .5327313
         lpm |  -5.049218   7.213431    -0.70   0.491    -20.00896    9.910523
      lstage |   1.432682   1.354852     1.06   0.302    -1.377109    4.242472
        lpl2 |  -.4498118   .1330767    -3.38   0.003    -.7257959   -.1738277
        lpf2 |  -.0179439   .0672491    -0.27   0.792    -.1574101    .1215222
        lpm2 |   .5261826   .7548597     0.70   0.493    -1.039301    2.091666
     lstage2 |  -.0526781   .1142455    -0.46   0.649    -.2896087    .1842525
        year |   .0465786   .0304395     1.53   0.140    -.0165491    .1097063
       _cons |   7.859137   16.95165     0.46   0.647    -27.29643    43.01471
-------------+----------------------------------------------------------------
     sigma_u |  .76036594
     sigma_e |   .1447302
         rho |  .96503636   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

If we add the squares of the logged regressors the fixed effects regression in (a), the estimated coefficients for *reg* is 0.052. This indicates that the variable costs during regulated period are approximately 5.2 percent higher than the costs during deregulated

period. In other words, it suggests that deregulation did contribute to the reduction on airlines' variable cost.

h. Yes they are, see below

. test lpl2 lpf2 lpm2 lstage2

(1) lpl2 = 0
(2) lpf2 = 0
(3) lpm2 = 0
(4) lstage2 = 0

F(4, 22) = 6.00
Prob > F = 0.0020

i. Regression (e) and (g) give opposite results about regulation, in (e) we are not controlling for efficiency of the flight length variable but in (g) by adding the stage squared term we may be better addressing the efficiency of the flight length so we are seeing the true impact of the regulation.

Do file for question 5:

```
use deregulate.dta, clear
sum vc
gen lvc=log(vc)
gen lpl=log(pl)
gen lpf=log(pf)
gen lpm=log(pm)
gen lstage=log(stage)
xtset airline year
reg  lvc reg  year  lpl lpf lpm lstage, r
xtreg  lvc reg  year  lpl lpf lpm lstage, fe
xtreg  lvc reg  year  lpl lpf lpm lstage, fe vce(cluster airline)
gen lpl2=lpl^2
gen lpf2=lpf^2
gen lpm2=lpm^2
gen lstage2=lstage^2
xtreg  lvc reg  year  lpl lpf lpm lstage  lpl2 lpf2 lpm2 lstage2, fe
```

Question 2: SW Exercise 10.1

(a) With a $1 increase in the beer tax, the expected number of lives that would be saved is 0.45 per 10,000 people. Since New Jersey has a population of 8.1 million, the expected number of lives saved is $0.45 \times 810 = 364.5$. The 95% confidence interval is $(0.45 \pm 1.96 \times 0.22) \times 810 = [15.228, 713.77]$.

(b) When New Jersey lowers its drinking age from 21 to 18, the expected fatality rate increases by 0.028 deaths per 10,000. The 95% confidence interval for the change in death rate is $0.028 \pm 1.96 \times 0.066 = [-0.1014, 0.1574]$. With a population of 8.1 million, the number of fatalities will increase by $0.028 \times 810 = 22.68$ with a 95% confidence interval $[-0.1014, 0.1574] \times 810 = [-82.134, 127.49]$. When real income per capita in new Jersey increases by 1%, the expected fatality rate increases by 0.0181 deaths per 10,000

(c) When real income per capita in new Jersey increases by 1%, the expected fatality rate increases by 1.81 deaths per 10,000. The 90% confidence interval for the change in death rate is $1.81 \pm 1.64 \times 0.47 = [1.04, 2.58]$. With a population of 8.1 million, the number of fatalities will increase by $1.81 \times 810 = 1466.1$ with a 90% confidence interval $[1.04, 2.58] \times 810 = [840, 2092]$.

(d) The low p-value (or high F-statistic) associated with the F-test on the assumption that time effects are zero suggests that the time effects should be included in the regression.

(e) The difference in the significance levels arises primarily because the estimated coefficient is higher in (5) than in (4). However, (5) leaves out two variables (unemployment rate and real income per capita) that are statistically significant. Thus, the estimated coefficient on Beer Tax in (5) may suffer from omitted variable bias. The results from (4) seem more reliable. In general, statistical significance should be used to measure reliability only if the regression is well-specified (no important omitted variable bias, correct functional form, no simultaneous causality or selection bias, and so forth.)

(f) Define a binary variable west which equals 1 for the western states and 0 for the other states. Include the interaction term between the binary variable west and the unemployment rate, west × (unemployment rate), in the regression equation corresponding to column (4). Suppose the coefficient associated with unemployment rate is β, and the coefficient associated with
west × (unemployment rate) is γ. Then β captures the effect of the unemployment rate in the eastern states, and $\beta + \gamma$ captures the effect of the unemployment rate in the western states.
The difference in the effect of the unemployment rate in the western and eastern states is γ. Using the coefficient estimate $(\hat{\gamma})$ and the standard error $\mathrm{SE}(\hat{\gamma})$, you can calculate the t-statistic to test whether γ is statistically significant at a given significance level.

Question 3: SW Exercise 10.5

Let $D2_i = 1$ if $i = 2$ and 0 otherwise; $D3_i = 1$ if $i = 3$ and 0 otherwise … $Dn_i = 1$ if $i = n$ and 0 otherwise. Let $B2_t = 1$ if $t = 2$ and 0 otherwise; $B3_t = 1$ if $t = 3$ and 0 otherwise … $BT_t = 1$ if $t = T$ and 0 otherwise. Let $\beta_0 = \alpha_1 + \mu_1$; $\gamma_i = \alpha_i - \alpha_1$ and $\delta_t = \mu_t - \mu_1$.

Question 4: SW Empirical Exercise 10.2

The solutions will reference the following regression results.

	(1)	**(2)**	**(3)**	**(4)**	**(5)**
log_GDPPC	0.236 (0.012) [0.212, 0.259]	0.235 (0.012) [0.211, 0.259]	0.083 (0.031) [0.021, 0.146]	0.054 (0.042) [-0.030, 0.137]	0.025 (0.054) [-0.057, 0.120]
Controls	No	No	No	No	Yes
State Effects	No	No	Yes	Yes	Yes
Time Effects	No	Yes	No	Yes	Yes
***F*-Statistics and *p*-values testing exclusion of groups of variables**					
Time Effects		9.31 (0.000)		5.73 (0.00)	4.61 (0.000)
Age Variables					2.12 (0.08)
Age, educ., & pop. variables					1.44 (0.21)

Controls: Age, education and population variables

a. The dataset is unbalanced because data are available over different years for different countries. As an example, data on *Dem_ind* for Andorra are available for 1970, 1995, and 2000, but for Afghanistan data are available for 1960, 1965, … , 2000.

b. (i) Values of *Dem_ind* range from 0.0 to 1.0. The mean in 0.50 and the standard deviation is 0.37. The 10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th} percentiles are 0.0, 0.17, 0.5, 0.83, and 1.0.

(ii) The value for the U.S. in 2000 is *Dem_ind* = 1.0. The average for the nine years in the sample is 0.99.

(iii) The value for Libya in 2000 is *Dem_ind* = 0.0. The average for the nine years in the sample is 0.11.

(iv) This involves looking through the data set. Do it. It's fun and interesting!

c. (i) The coefficient is 0.24 with a standard error of 0.01. The 95% confidence interval is 0.21 to 0.26. The coefficient is large, as described below.

(ii) A 20% increase in GDP per capita implies that *log_gdp* increases by approximately 0.20, so that *Dem_ind* is predicted to increase by approximately 0.20×0.24 = 0.048, or about 1/10 of the standard deviation in the dataset. The 95% confidence for the effect is (approximately) 0.20×0.21 to 0.20×0.26 or 0.042 to 0.052.

(iii) Clustered standard errors are needed because of country-specific omitted factors in the regressions. The clustered standard error for *Dem_ind* is 0.012; the unclustered standard error is smaller (0.007) because it ignores the positive within-country correlation of the errors.

d. (i) Countries have different histories, institutions, social structures, and religions. All affect their preference for demography and may also be correlated with economic development, and therefore per-capital income.

(ii) The estimated coefficient falls by a factor of 3, to 0.083 with a standard error of 0.032. The estimated effect, while significantly smaller is still statistically significant at the 1% significance level. The estimated effect in (c.ii) also falls by a factor of 3.

(iii) Data for Azerbaijan is available only for 2000. These data are completely absorbed the country-specific fixed effect and therefore have no effect on the estimated coefficient on *log_gdppc*.

(iv) The demand for democracy is contagious and sweeps across countries (remember the "Arab Spring" of 2012). To the extent that these events coincide with global changes in economic conditions they are correlated with *log_gdppc* and will therefore lead to omitted variable bias.

(v) The estimated coefficient falls further to 0.05, approximately 1/5 of the value that omits time and country fixed effects. The estimate is not statistically significant at the 10% level.

(vi) Regression (5) in the table includes population, age, and education. Jointly, these variables are not statistically significant in the regression, although the age variables are significant at the 10% level. When these variables are included, the estimated coefficient on *log_gdppc* falls further to 0.03 with a standard error of 0.05.

e. After controlling for omitted variables – particularly country-fixed effects – there is little evidence of an income effect on the demand for democracy.

Do file:

```
gen y1960 =(year==1960);
gen y1965 =(year==1965);
gen y1970 =(year==1970);
gen y1975 =(year==1975);
gen y1980 =(year==1980);
gen y1985 =(year==1985);
gen y1990 =(year==1990);
gen y1995 =(year==1995);
gen y2000 =(year==2000);
```

```
xtset code year;
summarize;
summarize dem_ind, detail;
tabulate code, summarize(dem_ind);
reg dem_ind log_gdppc, vce(cluster code);
reg dem_ind log_gdppc, robust;
reg dem_ind log_gdppc y1965 y1970 y1975 y1980 y1985 y1990 y1995 y2000,
vce(clustercode);
test y1965 y1970 y1975 y1980 y1985 y1990 y1995 y2000;
xtreg dem_ind log_gdppc, fe vce(cluster code);
xtreg dem_ind log_gdppc if code ~= 11, fe vce(cluster code);
xtreg dem_ind log_gdppc y1965 y1970 y1975 y1980 y1985 y1990 y1995 y2000, fe
vce(cluster code);
test y1965 y1970 y1975 y1980 y1985 y1990 y1995 y2000;
xtreg dem_ind log_gdppc log_pop age_median y1965 y1970 y1975 y1980 y1985 y1990 y1995
y2000, fe vce(cluster code);
xtreg dem_ind log_gdppc log_pop educ age_2 age_3 age_4 age_5 y1965 y1970 y1975 y1980
y1985 y1990 y1995 y2000, fe vce(cluster code);
test y1965 y1970 y1975 y1980 y1985 y1990 y1995 y2000;
test age_2 age_3 age_4 age_5;
test age_2 age_3 age_4 age_5 educ
```