[redacted]

Name: Ben Chu
[redacted]

ben-chu@berkeley.edu

Problem 1. Efficient markets hypothesis

a)

Based on the results for the two OLS regressions, the correlation between dkr and lnetincome is close to zero. In regression 1, the coefficient on dkr is . 0597546, while in regression 2, the coefficient on dkr is .062906. This is not a large change, which means that it is unlikely that the two are correlated with each other (likely that they are uncorrelated). The correlation should be close to zero since the coefficient did not have a huge change from one regression to the next, which shows that the variation within the regression 1 dkr was not captured within the relationship between return and lnetincome. Additionally, the correlation is not negative since the coefficient on dkr did slightly increase when moving from regression 1 to regression 2.

b)

The coefficient on lnetincome in Regression #2 can be interpreted as a 1% increase in the net income of a firm in 2013 is associated (association is not causation) with an increase in return of holding their stock for one year (Jan 14' to Dec 14') of 4.551175. This means .01*4.551175 = .04551175%.

c)

Use regression #3 to examine whether EMH holds. It is given that EMH = stock prices reflect all of the information available to investors. If EMH holds, then using information from 2013 should not permit us to predict 2014 stock returns.

Null hypothesis H_0: $\beta_1 = \beta_2 = \beta_3 = 0$

Alternative hypothesis H_1: at least one $\beta_{1,} \beta_{2,} \beta_3 \neq 0$
(at least one does not equal zero)

$\beta_{1,} \beta_{2,} \beta_3$ = the population regression coefficients for dkr, lnetincome, and lsalary

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

d)

Using Stata, test at the 5% level:

test dkr lnetincome lsalary

Since multiple regressors, use f-statistic from regression #3 = 11.01. Stata tests for joint hypothesis that regressors all = 0.

The null hypothesis from this output matches the null hypothesis for EMH.

An f-statistic of 11.01 is much greater than the 2.60 which corresponds with the 5% critical value for 3df (given in regression #3: F(3, 138)).

Prob > F = 0.0000 means an incredibly small p-value much lower than the specified 5% level of < 0.05. Given the small p-value below the 5% significance level, we must reject the efficient markets null hypothesis.

e)

The result obtained in part (d) after examining the validity of EMH led us to reject the null hypothesis based on the very small p-value. At the 5% significance level, at least one $\beta_{1,}\beta_{2,}\beta_3 \neq 0$, which means that at least one does not equal zero, which implies that at least one of the variables in question: dkr, lnetincome, or lsalary can be used to predict the returns of firms' stocks. Using this framework, there is reason to doubt the absolute validity of EMH because the stock returns in 14' do not necessarily reflect the information that is available to the public from 13'; the previous year.

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

f)

At least two reasons why there might be imperfect multicollinearity present in regression 3.

1) A significant f-statistic within regression #3 given by the small p-value of 0.0000, but when given an individual t-statistic, lnetincome is the only one that is statistically significant with a p-value of .044.

2) The coefficient size associated with lnetincome changes a lot when including the regressor lsalary. Additionally, the R^2 increases even though lsalary is most likely correlated with the returns to the stock.

3) The standard error of lnetincome is over 5 times greater in regression #3 than regression #2 after adding the regressor lsalary. Regressor #2 did not include lsalary, while regressor #3 did.

g)

The true statement based on a comparison of regression #2 and regression #3 is:

(iii) lnetincome and lsalary are highly correlated

When including the regressor lsalary in Regression #3, the lnetincome coefficient increases. Conversely, the coefficient regression on lsalary is negative. Finally, the coefficient on dkr did not have a huge change from regression #2 to regression #3.

h)

Not including a company that went out of business before the end of the year and not being able to enter it into the sample could have a significant impact on the estimated coefficient relative to the population regression, such as selectivity bias. If the companies who had very bad returns and go out of business are not represented in the sample, selectivity bias could. If the sample

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

does exhibit selectivity, then the coefficients may be closer to zero than they would be if the true value was represented.

Problem 2. Airlines and antitrust

a)

regress lfare dist passen concen, robust

```
. regress lfare dist passen concen, robust

Linear regression                               Number of obs     =      1,149
                                                F(3, 1145)        =     233.18
                                                Prob > F          =     0.0000
                                                R-squared         =     0.3697
                                                Root MSE          =     .33151

------------------------------------------------------------------------------
             |               Robust
       lfare |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dist |   .4272306   .0182638    23.39   0.000     .3913964    .4630648
      passen |  -.0580609   .0143404    -4.05   0.000    -.0861973   -.0299245
      concen |   .1874532   .0611786     3.06   0.002     .0674185    .3074879
       _cons |   4.656021   .0509217    91.43   0.000     4.556111    4.755931
------------------------------------------------------------------------------
```

b)

The interpretation of the coefficient on passen is that an average of one thousand more passengers per day on the origin destinations pairs can be associated (association is not

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

causation) with a change of -.0580609 in lfare. This means that a one unit (positive) change in passen is associated with a -.0580609 change in lfare. Change in passengers to origin destination is associated with a dependent variable change. The passen variable is in units of thousands of passengers.

The change in log of lfare, the logarithm of the average fare on the given route, equals a 5.80609% or about 5.8% decline in the price of the average fare. The average fare among all of the markets for airlines within the sample is about $188 given the concen coefficient of .1874532. Since -.0580609*188 = -10.904, which means that the decrease in price should be about $11.

c)

The partial effect of the market share (concen) of the largest carriers on air fares is the regressor concen, gives a positive coefficient of .1874532. Since the coefficient is positive, as long as the OLS assumptions hold, then a 10% increase in the market share equals 100*.1874532*.1 = 1.874532%. This means a 1.874532% increase in fares since concen is measured as a fraction.

The answer is consistent with the hypothesis that the firms use their market power to charge higher prices because of the small p-value of .002. The small p-value is below even the 1% significance level (as well as 5% and 10% levels) gives strong support for the null hypothesis.

d)

To test whether market power is used the same way on more popular and less popular routes: model, hypothesis, estimation, test. Dummy variable in Stata will be a popular variable, which will be the origin destination above the mean average daily passengers sampled, which is .67. Anything >.67 in thousands of passengers. Once popular is generated, interact popular with concen by generating a new variable and multiplying them together, concen*popular. Finally, we will regress the variables lfare dist passen concen concpop against each other and call for robust.

log(fare) = $\beta_0 + \beta_1$dist $+ \beta_2$passen $+ \beta_3$concen $+ \beta_4$popular*concen + u

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

gen popular = passen >.670

gen concpop = popular*concen

regress lfare dist passen concen concpop, robust

```
. regress lfare dist passen concen concpop, robust

Linear regression                               Number of obs     =      1,149
                                                F(4, 1144)        =     204.78
                                                Prob > F          =     0.0000
                                                R-squared         =     0.3934
                                                Root MSE          =     .32538
```

lfare	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
dist	.4216869	.0180954	23.30	0.000	.3861831	.4571908
passen	-.0025564	.0142658	-0.18	0.858	-.0305465	.0254336
concen	.2469424	.0606471	4.07	0.000	.1279504	.3659344
concpop	-.2851556	.0468132	-6.09	0.000	-.377005	-.1933061
_cons	4.636875	.0495973	93.49	0.000	4.539563	4.734187

The coefficient dealing with the interaction term concpop, which is popular*concen. The estimate of concpop is -.2851556. Concpop is extremely statistically significant at any significance level due to the small p-value of 0.000 well below the .05 cutoff for the 95% level. These regression results mean that the market power is not essentially utilized in the same way on the more popular routes or even less frequent usage of market power. For example, a 10% increase of the market share for a particular airline is associated with (association is not causation) a 2.851556% less increase (since negative) on the fares for popular routes vs. the fares

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

on non-popular routes. This may mean that larger airlines may be able to capitalize on economies of scale.

e)

The results would not be "externally valid" if applied to the airline industry in one of these other two regions for a few reasons. "External validity" is threatened since the U.S, European, and Asain markets differ so much in geography, taste, culture, practice, etc. The European and Asian markets may have differing distances between prime destinations, which may change the time and fuel input and may thereby harm "external validity." Shorter distances may be cheaper (especially since Europe is comparatively smaller than Asia or the U.S) than longer distance flights. Long distance flights have no alternative modes competing with them; however, short flights compete with cars, boats, and trains, which may also impact "external validity" if these forms are prevalent in one region and not the other. The results may not be "externally valid" if other regions or countries have different business or economic structures and practices that would impact the fares-concentration relationship. Another reason why the results would not be "externally valid" if applied to a different region is differences in legal practices or technicalities between the two regions, which impact the fares-concentration relationship. Another reason why the results would not be "externally valid" if applied to a different region is differences in the social climate or norms which would further impact the fares-concentration relationship. This means that results from the regression may not hold in a region and time period with a more regulated airline industry. Such an industry would be characterized by more concentration because of a strong national airline system. Conversely, this is not the case in the United States. The business-structure between two such regions differ significantly. This may result in different demands for airline travel between the market in the United States and other markets: Europe and Asia.

f)

The threats to "internal validity" are sample selectivity, omitted variables, simultaneous causality, specification error, and errors in variables.

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

1) Sample selectivity - Sample selectivity is a potential threat to “internal validity.” Sample selectivity is if the observations taken for a sample are drawn in a manner which is in some way related to the dependent variable. This will mean that coefficient estimates derived from the sample may be systematically different from the coefficient within the actual population regression. Within the context of this problem, sample selectivity may happen if some or all of the cities included in the sample are drawn only because they are already above a minimum level of population: a population cutoff. A population cutoff to be included in the sample is sample selectivity because drawing to take only cities above the population would be related in some way to the dependent variables. If this sample selectivity threat to “internal validity” happens, then the coefficient on concen would not be accurate within the context of the problem because smaller cities would not be included, which may also cause smaller airliners to not be included in the sample.

2) Omitted variables - Omitted variables are a potential threat to “internal validity.” Omitted variables are variables within the regression model which impact the price of air fares which are not included in the regression model, yet they are correlated with the variables that are included within the regression model. Omitted variables within the context of this problem would be regional taxes on airlines, price of alternative transportation (competition from other modes), and the price variation of the destination city. Essentially any variable which would impact the fare price NOT included in the variables presented. The variables omitted from the regression would be a threat to “internal validity” and cause bias if they are correlated with passengers or in any other way impact the price of fares.

3) Simultaneous causality - Simultaneous causality is a potential threat to “internal validity.” Simultaneous causality is when an independent variable within the regression model may also serve as a dependent variable within another regression model, which would make the variable essentially endogenous. An example of simultaneous causality within the context of this problem is that the concentration on a given route is calculated by the cost of serving that particular route (supply-side conditions). At the same time, it would be expected that the costs of serving a route are linked to the price of air fares; however, given the fact that costs are not included as a variable within the regression being estimated, it is assumed that they are part of the leftover or

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

error term and invariably correlated with the concen variable. This will cause the concen coefficient to suffer from omitted variable bias.

4) Specification error - Specification error is a potential threat to "internal validity." A specification error means that there may be a relationship that is not in log-linear specification within the context of the regression model stated. An example of specification error within the context of this problem is if the concentration of the market share of biggest airline carriers on the route is directly related to the price of airfare prices; however, the relation is at a diminishing rate. If the relationship is diminishing (or even conversely an increasing in rate), then it is quadratic, rather than linear (as specified by logarithm of the average fare on the route). The specification of log-linear is wrong if it is diminishing, not linear, which represents a specification error in the model.

5) Errors in variables - Errors in variables is a potential threat to "internal validity." Measuring explanatory variables wrong or mis-measuring explanatory variables will make it so that the coefficient on the variable that was measured wrong will be biased towards zero. Mismeasuring an explanatory so that the coefficient on the variable that was measured wrong will be biased towards zero is called the "attenuation effect." Within the context of this problem, error in variables could happen if the data is mismeasured. For example, if data recorded for the largest airlines share of the market is not efficiently calculated because the calculation only looked at one or two variables (ex. market cap or prices of the stock), then the measurement would not be considered to be measured correctly. This will threaten "internal validity" because the coefficient on the variable that was measured wrong will be biased towards zero. Another mis-measuring error may happen if a better way to measure the share of the airline carriers market considers all of the carriers, including the smaller airliners. Either of these situations would threaten internal validity by causing errors in variables via mis-measurement of explanatory, which causes mismeasured coefficients to be biased towards zero. Bad data measurement!

Problem 3. World Health Organization (WHO)

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

a)

keep if year == 1997
gen hc3sq = hc3^2
regress dale hexp hc3 hc3sq , robust

```
. keep if year == 1997
(560 observations deleted)

. gen hc3sq = hc3^2

. regress dale hexp hc3 hc3sq , robust

Linear regression                               Number of obs     =        140
                                                F(3, 136)         =     119.27
                                                Prob > F          =     0.0000
                                                R-squared         =     0.6679
                                                Root MSE          =     7.2699
```

dale	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
hexp	.0059464	.000802	7.41	0.000	.0043604	.0075324
hc3	7.697244	.9630129	7.99	0.000	5.792828	9.601661
hc3sq	-.4167271	.0755841	-5.51	0.000	-.5661992	-.2672549
_cons	25.91128	2.720703	9.52	0.000	20.53093	31.29164

It appears that the relationship is both quadratic (given the hc3^2) and strong (given the high R-squared). This means that the life expectancy increases as the hc3 = Educational attainment (tertiary schooling) increases; however, the rate at which it increases is decreasing, which means that it slows in growth for higher values, which signifies a quadratic relationship. Since the p-values for each coefficient are 0.000, they are all statistically significant within the context of this regression, which means that both quadratic and linear coefficients are extremely statistically significant.

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

b)

regress dale hexp hc3 hc3sq gini tropics popden pubthe gdpc voice geff , robust

```
. regress dale hexp hc3 hc3sq gini tropics popden pubthe gdpc voice geff , robust

Linear regression                               Number of obs     =        140
                                                F(10, 129)        =      64.11
                                                Prob > F          =     0.0000
                                                R-squared         =     0.7235
                                                Root MSE          =     6.8106
```

dale	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
hexp	-.0019312	.0015318	-1.26	0.210	-.0049619	.0010994
hc3	6.781209	.9243331	7.34	0.000	4.952393	8.610024
hc3sq	-.3826294	.0732444	-5.22	0.000	-.5275452	-.2377136
gini	-16.16471	10.33849	-1.56	0.120	-36.61967	4.290248
tropics	-3.12001	1.789764	-1.74	0.084	-6.661101	.4210809
popden	-.0000846	.0001293	-0.65	0.514	-.0003403	.0001712
pubthe	-.0508495	.028834	-1.76	0.080	-.1078982	.0061992
gdpc	.0004559	.0001854	2.46	0.015	.0000891	.0008226
voice	.8410275	1.120596	0.75	0.454	-1.376098	3.058153
geff	1.911934	1.307101	1.46	0.146	-.6741981	4.498066
_cons	41.07172	5.065922	8.11	0.000	31.04867	51.09477

The effect of including the additional controls on the coefficients of the other included regressors has a small impact on the OLS coefficient of the variables for education. The f-statistic leads us to reject the joint test that all of the additional controls added after part (a) are equal to zero. Testing the coefficients yields an f-statistic of 5.47.

test gini tropics popden pubthe gdpc voice geff

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

```
. test gini tropics popden pubthe gdpc voice geff

 ( 1)  gini = 0
 ( 2)  tropics = 0
 ( 3)  popden = 0
 ( 4)  pubthe = 0
 ( 5)  gdpc = 0
 ( 6)  voice = 0
 ( 7)  geff = 0

       F(  7,   129) =    5.47
            Prob > F =    0.0000
```

c)

Using factor variables:

regress dale hexp c.hc3 c.hc3#c.hc3 gini tropics popden pubthe gdpc voice geff , robust

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

```
. regress dale hexp c.hc3 c.hc3#c.hc3 gini tropics popden pubthe gdpc voice geff , robust

Linear regression                               Number of obs     =        140
                                                F(10, 129)        =      64.11
                                                Prob > F          =     0.0000
                                                R-squared         =     0.7235
                                                Root MSE          =     6.8106
```

dale	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
hexp	-.0019312	.0015318	-1.26	0.210	-.0049619	.0010994
hc3	6.781208	.9243331	7.34	0.000	4.952393	8.610024
c.hc3#c.hc3	-.3826294	.0732444	-5.22	0.000	-.5275452	-.2377135
gini	-16.16471	10.33849	-1.56	0.120	-36.61967	4.290248
tropics	-3.12001	1.789764	-1.74	0.084	-6.661101	.4210811
popden	-.0000846	.0001293	-0.65	0.514	-.0003403	.0001712
pubthe	-.0508494	.028834	-1.76	0.080	-.1078982	.0061993
gdpc	.0004559	.0001854	2.46	0.015	.0000891	.0008226
voice	.8410274	1.120596	0.75	0.454	-1.376098	3.058153
geff	1.911934	1.307101	1.46	0.146	-.6741982	4.498065
_cons	41.07172	5.065922	8.11	0.000	31.04867	51.09477

margins , dydx(hc3)

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

```
. margins , dydx(hc3)

Average marginal effects                        Number of obs     =        700
Model VCE    : Robust

Expression   : Linear prediction, predict()
dy/dx w.r.t. : hc3
```

	dy/dx	Delta-method Std. Err.	t	P>|t|	[95% Conf.	Interval]
hc3	2.038887	.1053184	19.36	0.000	1.832103	2.24567

The difference in marginal effects of education between the quadratic specification in question b) and the linear specification in part a) is large. For the linear specification in part a), the marginal effect of hc3, the educational attainment (tertiary schooling) is equal to its coefficient of 7.697244. There is a marginal effect of about 2 (2.038887) when the quadratic term is included in the regression.

d)

Interacting the 3 of the coefficients in the population regression with an indicator of the OECD membership.

Test if the coefficients on the interaction variables are actually jointly zero. In order to jointly test if the coefficients are jointly zero, allow the constant term _cons to either force it to be the same or different from the two subsamples. No need for factor variables.

gen oecdhexp = oecd*hexp

gen oecdhc3 = oecd*hc3

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

gen oecdhc3sq = oecd*hc3sq

regress dale hexp hc3 hc3sq oecd oecdhexp oecdhc3 oecdhc3sq , robust

```
. gen oecdhexp = oecd*hexp

. gen oecdhc3 = oecd*hc3

.
. gen oecdhc3sq = oecd*hc3sq

. regress dale hexp hc3 hc3sq oecd oecdhexp oecdhc3 oecdhc3sq , robust

Linear regression                               Number of obs     =        140
                                                F(7, 132)         =     102.13
                                                Prob > F          =     0.0000
                                                R-squared         =     0.6924
                                                Root MSE          =     7.1022
```

dale	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
hexp	.0105243	.0027291	3.86	0.000	.0051258	.0159228
hc3	6.328638	1.301461	4.86	0.000	3.754221	8.903056
hc3sq	-.3155583	.109694	-2.88	0.005	-.5325438	-.0985727
oecd	14.33995	5.490885	2.61	0.010	3.478441	25.20147
oecdhexp	-.0078453	.0027855	-2.82	0.006	-.0133552	-.0023354
oecdhc3	-.1518451	1.881249	-0.08	0.936	-3.873142	3.569452
oecdhc3sq	-.0692694	.139721	-0.50	0.621	-.3456514	.2071126
_cons	28.38834	3.104459	9.14	0.000	22.24742	34.52927

test oecd oecdhexp oecdhc3 oecdhc3sq

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

```
. test oecd oecdhexp oecdhc3 oecdhc3sq

 ( 1)  oecd = 0
 ( 2)  oecdhexp = 0
 ( 3)  oecdhc3 = 0
 ( 4)  oecdhc3sq = 0

       F(  4,   132) =   30.99
            Prob > F =    0.0000
```

regress dale hexp hc3 hc3sq oecdhexp oecdhc3 oecdhc3sq , robust

```
. regress dale hexp hc3 hc3sq oecdhexp oecdhc3 oecdhc3sq , robust

Linear regression                               Number of obs     =        140
                                                F(6, 133)         =     114.50
                                                Prob > F          =     0.0000
                                                R-squared         =     0.6917
                                                Root MSE          =     7.0832
```

dale	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
hexp	.010544	.0027265	3.87	0.000	.0051512	.0159369
hc3	6.244195	1.283488	4.87	0.000	3.705505	8.782885
hc3sq	-.3088645	.108325	-2.85	0.005	-.5231272	-.0946017
oecdhexp	-.0079034	.002775	-2.85	0.005	-.0133922	-.0024145
oecdhc3	3.287132	.7946714	4.14	0.000	1.715303	4.858962
oecdhc3sq	-.2687377	.0852747	-3.15	0.002	-.4374077	-.1000677
_cons	28.61435	3.052606	9.37	0.000	22.57641	34.65228

test oecdhexp oecdhc3 oecdhc3sq

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

```
. test oecdhexp oecdhc3 oecdhc3sq

 ( 1)  oecdhexp = 0
 ( 2)  oecdhc3 = 0
 ( 3)  oecdhc3sq = 0

       F(  3,   133) =   14.12
            Prob > F =    0.0000
```

e)

The panel data set that includes all five years is a "balanced" panel. The panel is balanced and has 140 countries who report their data for each one of the 5 years (i.e every country is observed every year = balanced). This gives a total sample size of 700 since 140 (countries) x 5 (years) = 140*5 = 700.

f)

An example of a time-invariant variable (variable that will be the same regardless of when it is observed) that would result in the different life expectancy across countries is that the degree to risk of death within an environment or situation differs across countries. Various regions have different existential and fundamental threats to life. Within the context of this problem, a variable reported as tropics has more infectious diseases because the climate and environment lends itself to be infections and diseases. A region that is considered to fall under "tropics" will not change with time (aside from extreme unusual circumstances). These infections (and region) may have a higher threat to life than a region without the high risk of diseases and infections. The coefficient for tropics when regressed is -3.12001, which is a negative coefficient, which means that the tropics is correlated with a much lower life expectancy. This is evident in the expanded

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

regression. Adding a dummy variable may also show a correlation between poorer countries and poorer people. Another example would be a variable named "desert" and the inherent dangers which would come along with the desert which do not necessarily vary across time.

g)

When including the fixed effects for each country in the sample and incorporating a dummy indicator. The fixed effect of only the first country is included in the total of the 140 countries in the sample. This means country = 103, the first country indicated by the Stata output.

regress dale hexp hc3 hc3sq i.ncountry, robust

```
. regress dale hexp hc3 hc3sq i.ncountry, robust

Linear regression                               Number of obs     =        700
                                                F(142, 557)       =   15334.33
                                                Prob > F          =     0.0000
                                                R-squared         =     0.9988
                                                Root MSE          =     .47231
```

dale	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
hexp	.0013422	.0002462	5.45	0.000	.0008587	.0018258
hc3	2.211138	.6367876	3.47	0.001	.9603389	3.461936
hc3sq	-.0344191	.0488191	-0.71	0.481	-.1303111	.061473
ncountry						
103	-.5422769	1.015914	-0.53	0.594	-2.537768	1.453215
_cons	50.16896	2.147919	23.36	0.000	45.94995	54.38798

Coefficients on education and the square of education are both closer to zero. Additionally, another change is that the quadratic coefficient is not statistically significant anymore.

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

h)

regress dale hexp hc3 hc3sq i.ncountry , cluster(ncountry)

```
. regress dale hexp hc3 hc3sq i.ncountry , cluster(ncountry)

Linear regression                               Number of obs     =        700
                                                F(2, 139)         =          .
                                                Prob > F          =          .
                                                R-squared         =     0.9988
                                                Root MSE          =     .47231

                               (Std. Err. adjusted for 140 clusters in ncountry)
```

dale	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
hexp	.0013422	.000356	3.77	0.000	.0006384	.0020461
hc3	2.211138	1.072283	2.06	0.041	.0910444	4.331231
hc3sq	-.0344191	.0810333	-0.42	0.672	-.1946363	.1257982
ncountry						
103	-.5422769	1.625003	-0.33	0.739	-3.755197	2.670643

The standard errors change when computed clustered by country. The OLS coefficients remain unchanged when computed clustered by country. The size of the standard errors has increased in size with each of the three regressors: increased SE for hexp, hc3, and hc3sq.

i)

An example of an entity-invariant variable, which is excluded from the estimated regression model in part (a), that would result in variation in life expectancy over time are advances in

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

tactics to prevent diseases, technology, and medicine, which all spread to many other countries over time. These are time-invariant variables which have been excluded from the regression and result in a variation in life expectancy in all countries across the world. Advancements eventually spread to all countries in the world, even underdeveloped nations, so the technology will reach everyone regardless of their own innovations or advancements. This would be relevant in the example where the diseases in the tropic environment are necessarily time-invariant.

j)

In order to include the year fixed effects, we must use the dummy factor variable. Upon analyzing the Stata output, it is evident that after about half of a year of increasing life expectancy between 93' and 94', the difference from the year 93' continues to fall with each year after, which is not consistent with the reasoning about entity-invariant variables since you would assume that medical practices and disease-prevention-techniques would advance and thereby continue the upward trend in life expectancy. The egressor of education (explanatory) is extremely statistically significant with a 0.000 p-value, well below the 10%, 5%, and 1% significance level. To test for an alternate excluded category, we can choose 1995 as the "baseline" year (or any other year that we want to test) by plugging it into the i in i.year within the regression formula below.

regress comp hexp hc3 hc3sq i.year, robust

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

```
. regress comp hexp hc3 hc3sq i.year, robust

Linear regression                               Number of obs     =        700
                                                F(7, 692)         =     270.96
                                                Prob > F          =     0.0000
                                                R-squared         =     0.7034
                                                Root MSE          =     6.7749
```

comp	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
hexp	.0082305	.0004745	17.35	0.000	.0072989	.009162
hc3	6.394964	.4129277	15.49	0.000	5.584222	7.205705
hc3sq	-.3388748	.0318409	-10.64	0.000	-.4013912	-.2763584
year						
1994	.531221	.9306397	0.57	0.568	-1.295995	2.358437
1995	.4392526	.9344017	0.47	0.638	-1.39535	2.273855
1996	.2760003	.9380137	0.29	0.769	-1.565694	2.117695
1997	.1318851	.9441007	0.14	0.889	-1.72176	1.985531
_cons	46.2106	1.171935	39.43	0.000	43.90963	48.51158

k)

"Perform a joint hypothesis test using the F stat under the assumption of homoscedasticity."

Test for post-estimation. Run a test each year from 1993, 1994, 1995, 1996, 1997.
Run a test on:

- 1993.year
- 1994.year
- 1995.year

Name: Julian Bertalli
SID: 3033589368
Section ID: 103
jjbertalli@berkeley.edu

Name: Ben Chu
SID: 3032040857
Section ID: 110
ben-chu@berkeley.edu

- 1996.year
- 1997.year

test 1993.year 1994.year 1995.year 1996.year 1997.year

```
. test 1993.year 1994.year 1995.year 1996.year 1997.year

 ( 1)  1993.year = 0
 ( 2)  1994.year = 0
 ( 3)  1995.year = 0
 ( 4)  1996.year = 0
 ( 5)  1997b.year = 0
       Constraint 5 dropped

       F(  4,   692) =    0.14
            Prob > F =    0.9689
```

Test to see if all time fixed effects jointly = 0
F(4, 692) = .14
Prob > F = .9689