Notebook

April 16, 2021

Question 1.a. Based on the results for the two OLS regressions, what is the sign of the correlation between `dkr` and `lnetincome`? Alternatively, is there not enough information to determine the sign of the correlation?

The sign of the correlation seems to be positive as per the r2 statistic… however thinking about how net income is affected by debt, it does not make any sense (as debt goes up, net income goes down?).

Question 1.b. Interpret the coefficient on `lnetincome` in Regression 2.

The lnetincome tells us that as we multiply by whatever the net income is, it has a +4.551 weight on the return, when in combination with the dkr.

Question 1.c. Suppose that you use Regression 3 to examine whether EMH holds. What are the null and alternative hypotheses?

Null: B1 = B2 = B3 = 0 Alternate: B1 != B2 != B3 != 0

Question 1.d. Carry out the test in part (c) at the 5% level. Do you reject or fail to reject the null hypothesis?

Use the F statistic, which gives us a p value of 0.0000, and thus we can reject the null and accept the alternate.

Question 1.e. Interpret the result you obtained in part (d), in light of your task of examining the validity of EMH.

Because we reject the null, we can assume that EMH doesnt hold because B1, B2, and B3 should be equal to zero and thus do not hold.

Question 1.f. Provide (at least) two reasons why there might be imperfect multicollinearity present in Regression 3.

Not all the data is from the same year than the rest of the data, and the data might be inconsistent with how it should be analysed. The values have also been log(x) on some of them, so the scales might be off and inconsistent with the EMH.

Question 1.g. Which of the following statements is true based on a comparison of Regression 2 and Regression 3? - (i) `dkr` and `lnetincome` are highly-correlated. - (ii) `dkr` and `lsalary` are highly-correlated. - (iii) `lnetincome` and `lsalary` are highly-correlated. - (iv) All of the above. - (v) None of the above.

i

Question 1.h. The sample of 142 stocks only include companies that were traded on the NYSE as of the end of 2013. A company that went out of business, for instance, before the end of that year could not enter the sample. How would this sampling affect the estimated coefficient relative to the population regression?

It would cause errors because for this to hold true, we must take into account ALL variables contributing to the total return, so therefore the 142 stocks will help us fine tune our datasets.

Question 2.a. Regress `lfare` on `dist`, `passen` and `concen`, with robust standard errors. Make sure the cell below (and all regression questions in this assignment) shows your regression results like you've done in previous assignments, otherwise we cannot give credit. This assignment will be a little less guided. Make sure do use different variable names for each separate coding part to avoid unexpected errors from reusing variables. Refer to previous assignments if you need a refresher on how we performed different regressions. *Don't forget to add a constant to your regressions.*

[4]:
```
dist =  af['dist']
passen = af['passen']
concen = af['concen']
X_const = sm.add_constant(np.stack([passen, dist, concen], axis = 1))
model_X = sm.OLS(af['lfare'], X_const)
results_X = model_X.fit()
results_X.summary()
```

[4]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                  lfare   R-squared:                       0.370
Model:                            OLS   Adj. R-squared:                  0.368
Method:                 Least Squares   F-statistic:                     223.9
Date:                Fri, 16 Apr 2021   Prob (F-statistic):          2.75e-114
Time:                        11:08:52   Log-Likelihood:                -359.76
No. Observations:                1149   AIC:                             727.5
Df Residuals:                    1145   BIC:                             747.7
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          4.6560      0.051     90.629      0.000       4.555       4.757
x1            -0.0581      0.012     -4.930      0.000      -0.081      -0.035
x2             0.4272      0.019     22.548      0.000       0.390       0.464
x3             0.1875      0.059      3.168      0.002       0.071       0.304
==============================================================================
Omnibus:                       67.392   Durbin-Watson:                   1.367
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               28.039
Skew:                           0.131   Prob(JB):                     8.16e-07
Kurtosis:                       2.281   Cond. No.                         13.8
```

```
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
"""
```

Question 2.b. What is the interpretation of the coefficient on `passen`?

The coeficcient on passen is -0.0581 so therefore as the amount of passengers go up, their fare tends to go down due to that. This makes sense as it would be cheaper per capita to fly the plane as more people are fit in th plane.

Question 2.c. Based on your OLSEs, and assuming the OLS assumptions hold, what is the partial effect of the market share of the largest carrier on air fares? Is your answer consistent with the hypothesis that firms use their market power to charge higher prices?

I believe that the partial effect of market share is quite strong, as it is .20 * whatever that percent value will be. Distance is the largest partaker however, the only other contributing factor to raising it will be this percentage.

Question 2.d. How would you test whether market power is used the same way on more popular and less popular routes? Write down the model and the hypothesis, carry out the estimation and the test.

This question is for your code, the next is for your explanation.

[5]:
```
dist =  af['dist']
passen = af['passen']
concen = af['concen']
X_const = sm.add_constant(np.stack([passen, dist, concen, concen*passen], axis
 ↪= 1))
model_X = sm.OLS(af['lfare'], X_const)
results_X = model_X.fit()
results_X.summary()
```

[5]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                  lfare   R-squared:                       0.395
Model:                            OLS   Adj. R-squared:                  0.393
Method:                 Least Squares   F-statistic:                     186.8
Date:                Fri, 16 Apr 2021   Prob (F-statistic):          2.86e-123
Time:                        11:08:54   Log-Likelihood:                -336.12
No. Observations:                1149   AIC:                             682.2
Df Residuals:                    1144   BIC:                             707.5
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
```

```
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          4.5397      0.053     85.533      0.000       4.436       4.644
x1             0.1581      0.033      4.755      0.000       0.093       0.223
x2             0.4245      0.019     22.854      0.000       0.388       0.461
x3             0.4281      0.068      6.334      0.000       0.296       0.561
x4            -0.4455      0.064     -6.932      0.000      -0.572      -0.319
==============================================================================
Omnibus:                       53.030   Durbin-Watson:                   1.389
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               24.778
Skew:                           0.137   Prob(JB):                     4.16e-06
Kurtosis:                       2.335   Cond. No.                         17.9
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
"""
```

Question 2.e. Explain.

lfare = b0 + B1(passen) + b2(dist + B3(concen) + B4(passenXconcen)

Null: B4 = zero. Alternate: B4 != zero.

The t statistic is extremely far from the value we need it to be (around -6.932) so therefore the P value is extreemely small so at a signficance level of .01 we reject the null and accept the alternate that there is in fact a higher concen from more popular flights.

Question 2.f. We need to question whether the results of the regression in part (d) is revealing a causal relationship between concentration and airfares. In particular, we are concerned whether our estimation results on U.S. data are valid for other markets, such as Europe and Asia. Give one reason why the results would not be "externally valid" if applied to the airline industry in one of these other two regions.

This is not universally applical. First of all, there is a lot of domestic travel within the United States by plane that is easily replacable by train in Europe (through the european union). We cannot apply it to other countries as well because the way that American companies make money vary greatly from those in other countries.

Question 2.g. We are also aware of several potential threats to "internal validity" of the results. For each one of the five main internal validity threats, describe one possibility that could plausibly lead to that particular threat.

1. Omitted Variable bias We can have for instance capacity of plane contribute, but we wouldnt know if we do not include this variable in our regression.
2. Functional form misspecification This can happen when we for instance take data that is incompariable.
3. Measurement error and errors-in-variables bias There can be a misreading of fares for instance, because they vary for different parts of the airplane such as first class.

4. Sample selection bias Sometimes we are given a small group that is not actually representative of the full sample.
5. Simultaneous causality As holidays get closer, prices tend to go up according to demand and it doesnt reflect this as well.

Question 3.a. Create a new variable for the dataset that is the square of educational attainment (`hc3`). Then regress life expectancy (`dale`) on health expenditures (`hexp`), the educational attainment in the country (`hc3`), and its square (the variable you created). For now, select rows from 1997 and use only these rows in the regression. Use robust standard errors and *don't forget to add a constant term.* Comment on whether you think the relationship between life expectancy and education is linear or quadratic and why you came to that conclusion.

This question is for your code, the next is for your explanation.

[7]:
```
who['sqhc3'] = np.sqrt(who['hc3'])
who_97 = who[who['year'] == 1997]
X_const = sm.add_constant(np.stack([who_97['hexp'], who_97['hc3'],␣
 ↪who_97['sqhc3']], axis = 1))
model_X = sm.OLS(who_97['dale'], X_const)
results_X = model_X.fit()
results_X.summary()
```

[7]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                   dale   R-squared:                       0.659
Model:                            OLS   Adj. R-squared:                  0.651
Method:                 Least Squares   F-statistic:                     87.51
Date:                Fri, 16 Apr 2021   Prob (F-statistic):           1.35e-31
Time:                        11:08:55   Log-Likelihood:                -476.25
No. Observations:                 140   AIC:                             960.5
Df Residuals:                     136   BIC:                             972.3
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.2673      9.103      0.139      0.889     -16.734      19.269
x1             0.0057      0.001      4.811      0.000       0.003       0.008
x2            -4.6738      1.848     -2.529      0.013      -8.329      -1.019
x3            34.0501      8.324      4.091      0.000      17.589      50.512
==============================================================================
Omnibus:                       22.470   Durbin-Watson:                   1.686
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               29.396
Skew:                          -0.913   Prob(JB):                     4.14e-07
Kurtosis:                       4.306   Cond. No.                     1.69e+04
==============================================================================
```

```
Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 1.69e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

Question 3.b. Explain.

It looks like overall that the sqrth3c has a higher affect on the data than the actual h3c variable. Also, the data seems to barely be affected by health expenditure as well.

Question 3.c. To the specification in part (a), add the additional control variables: `gini`, `tropics`, `popden`, `pubthe`, `gdpc`, `voice`, and `geff`. Test whether these additional regressors are jointly significant (we do the F-test for you in this part, you just have to interpret it). What effect does inclusion of these additional controls have on the coefficients of the other included regressors?

This question is for your code, the next is for your explanation.

[16]:
```
# This is the code for your regression.
# We give you starter code for this one so that we know what the variable name␣
 ↪is
# for the regression results, which we use in the code cell below.
const = sm.add_constant(who_97[['hexp', 'hc3', 'sqhc3', 'gini', "tropics",␣
 ↪'popden','pubthe','gdpc','voice','geff']])

model_3b = sm.OLS(who_97['dale'], const)
results_3b = model_3b.fit()
results_3b.summary()
```

[16]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                   dale   R-squared:                       0.719
Model:                            OLS   Adj. R-squared:                  0.697
Method:                 Least Squares   F-statistic:                     33.01
Date:                Fri, 16 Apr 2021   Prob (F-statistic):           6.36e-31
Time:                        11:30:48   Log-Likelihood:                -462.65
No. Observations:                 140   AIC:                             947.3
Df Residuals:                     129   BIC:                             979.7
Df Model:                          10
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         17.2656      9.268      1.863      0.065      -1.071      35.602
hexp          -0.0028      0.003     -1.084      0.280      -0.008       0.002
```

```
hc3          -4.7855      1.883     -2.542      0.012      -8.510      -1.061
sqhc3        32.0681      8.345      3.843      0.000      15.557      48.579
gini        -15.4376      8.293     -1.861      0.065     -31.846       0.971
tropics      -3.2664      1.563     -2.090      0.039      -6.359      -0.174
popden    -7.181e-05      0.000     -0.352      0.726      -0.000       0.000
pubthe       -0.0468      0.036     -1.305      0.194      -0.118       0.024
gdpc          0.0005      0.000      2.088      0.039    2.65e-05       0.001
voice         0.9255      1.109      0.835      0.405      -1.268       3.119
geff          1.9539      1.397      1.399      0.164      -0.810       4.717
==============================================================================
Omnibus:                       18.527   Durbin-Watson:                   1.826
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               23.823
Skew:                          -0.770   Prob(JB):                     6.71e-06
Kurtosis:                       4.308   Cond. No.                     2.26e+05
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 2.26e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

Question 3.d. Explain.

Because of our values for our F, we can assume that this situation is really rare when we control for all these variables.

Question 3.e. Return to the simpler regression specification in part (a). We want see if the determinants of life expectancy are different for rich and poor countries. Use membership in the "Organization of Economic Cooperation & Development" (`oecd`) as the indicator of a rich country. The OECD had 30 member countries during this time period. Perform a test of the hypothesis that all three of the coefficients in the population regression are equal for OECD and non-OECD countries.

Hint: You will need to create three new variables.

This question is for your code, the next is for your explanation.

[25]:
```
who['ohexp'] = who['oecd']*who['hexp']
who['oh'] = who['oecd']*who['hc3']
who['osq'] = who['oecd']*who['sqhc3']
X_const = sm.add_constant(who[['hexp','hc3','sqhc3','ohexp','oh','osq']])
model_X = sm.OLS(who['dale'], X_const)
results_X = model_X.fit()
results_X.summary()
```

[25]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   dale   R-squared:                       0.707
Model:                            OLS   Adj. R-squared:                  0.704
Method:                 Least Squares   F-statistic:                     278.3
Date:                Fri, 16 Apr 2021   Prob (F-statistic):          8.17e-181
Time:                        11:41:09   Log-Likelihood:                -2318.5
No. Observations:                 700   AIC:                             4651.
Df Residuals:                     693   BIC:                             4683.
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         10.7647      3.872      2.780      0.006       3.163      18.366
hexp           0.0124      0.001     10.047      0.000       0.010       0.015
hc3           -2.9246      0.843     -3.469      0.001      -4.580      -1.269
sqhc3         25.0043      3.694      6.769      0.000      17.752      32.257
ohexp         -0.0095      0.001     -6.563      0.000      -0.012      -0.007
oh            -4.6066      0.862     -5.343      0.000      -6.299      -2.914
osq           16.1157      2.524      6.386      0.000      11.161      21.070
==============================================================================
Omnibus:                       80.788   Durbin-Watson:                   0.360
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              122.337
Skew:                          -0.796   Prob(JB):                     2.72e-27
Kurtosis:                       4.290   Cond. No.                     2.34e+04
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 2.34e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

Question 3.f. Explain.

Null: B4,B5,B6 = 0 Alternate: B4,B5,B6 != 0

By being able to check out the t stat p value, all of them are very low <0.001 so therefore we can assume that OECD does in fact affect the life expectancy of an individual living in said county.

Question 3.g. Give an example of a time-invariant variable that would result in different life expectancy across countries.

Quality of water in different countries.

Question 3.h. Estimate the regression having a fixed effect for each country in the sample. We have defined the endogenous and exogenous variables for you, you just have to fill in the rest. Notice how we converted the `country` variable to a set of dummy variables for each country. You

can ignore the coefficients for every country variable. What change took place in the coefficients on the education variables? Explain why you think there was a change in these coefficients.

This question is for your code, the next is for your explanation.

[27]:
```
# .get_dummies transforms a categorical variable into a dataframe of dummy␣
 ↪variables,
# one for each category. The prefix and prefix_sep part just makes sure the␣
 ↪variable
# names are strings and not integers.
countries = pd.get_dummies(who['country'], prefix='', prefix_sep='')
# This just joins the dummy dataframe with the original
who_country = who[['dale', 'hexp', 'hc3', 'sqhc3']].join(countries)
y_3h = who_country['dale']
# Here we drop country 191, since otherwise there would be perfect colinearity␣
 ↪in
# the columns. We also have to drop dale since that's the endogenous variable we
# regress on.
X_3h = sm.add_constant(who_country.drop(columns=['dale', '191']))
model_3h = sm.OLS(y_3h, X_3h)
results_3h = model_3h.fit()
results_3h.summary()
```

[27]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                   dale   R-squared:                       0.999
Model:                            OLS   Adj. R-squared:                  0.999
Method:                 Least Squares   F-statistic:                     3326.
Date:                Fri, 16 Apr 2021   Prob (F-statistic):               0.00
Time:                        11:49:32   Log-Likelihood:                -387.46
No. Observations:                 700   AIC:                             1061.
Df Residuals:                     557   BIC:                             1712.
Df Model:                         142
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         21.7037      3.436      6.316      0.000      14.954      28.453
hexp           0.0014      0.000      3.601      0.000       0.001       0.002
hc3            0.8494      0.742      1.145      0.253      -0.608       2.307
sqhc3          4.3208      3.130      1.380      0.168      -1.828      10.470
6             29.0807      0.401     72.482      0.000      28.293      29.869
7             23.5946      0.844     27.941      0.000      21.936      25.253
8             25.0202      0.781     32.016      0.000      23.485      26.555
10            24.9246      1.233     20.216      0.000      22.503      27.346
11            28.5143      0.776     36.757      0.000      26.991      30.038
```

14	5.3044	0.560	9.479	0.000	4.205	6.404
15	26.7060	0.961	27.787	0.000	24.818	28.594
16	13.1275	0.741	17.720	0.000	11.672	14.583
17	7.0452	0.697	10.104	0.000	5.676	8.415
18	18.3774	0.473	38.861	0.000	17.449	19.306
19	21.5627	1.074	20.079	0.000	19.453	23.672
20	27.9699	0.362	77.235	0.000	27.259	28.681
21	15.8392	0.792	19.989	0.000	14.283	17.396
23	19.7473	0.967	20.429	0.000	17.849	21.646
25	15.8556	0.377	42.085	0.000	15.116	16.596
26	24.4484	0.320	76.314	0.000	23.819	25.078
27	22.7962	0.778	29.317	0.000	21.269	24.324
30	0.1251	0.302	0.414	0.679	-0.468	0.718
31	6.8522	0.523	13.094	0.000	5.824	7.880
32	24.8030	1.232	20.138	0.000	22.384	27.222
34	25.7819	1.105	23.336	0.000	23.612	27.952
35	28.5096	0.579	49.240	0.000	27.372	29.647
36	24.2112	0.501	48.341	0.000	23.227	25.195
37	13.5283	0.488	27.699	0.000	12.569	14.488
38	8.8146	0.310	28.431	0.000	8.206	9.424
39	9.1854	0.371	24.787	0.000	8.458	9.913
40	26.2723	0.352	74.698	0.000	25.581	26.963
41	14.1092	0.373	37.822	0.000	13.377	14.842
42	24.6694	0.329	74.889	0.000	24.022	25.316
43	28.7827	0.419	68.755	0.000	27.960	29.605
45	28.2981	0.692	40.886	0.000	26.939	29.658
46	22.4337	1.223	18.348	0.000	20.032	24.835
47	24.4153	1.030	23.709	0.000	22.393	26.438
50	21.9370	1.303	16.836	0.000	19.378	24.496
51	26.5290	0.328	80.937	0.000	25.885	27.173
53	22.0677	0.492	44.833	0.000	21.101	23.035
54	24.7928	0.316	78.434	0.000	24.172	25.414
56	31.3879	0.622	50.456	0.000	30.166	32.610
57	19.9428	0.902	22.115	0.000	18.172	21.714
58	4.6674	0.611	7.633	0.000	3.466	5.869
59	24.6702	1.074	22.964	0.000	22.560	26.780
60	17.4702	0.844	20.688	0.000	15.811	19.129
61	30.8203	0.777	39.665	0.000	29.294	32.347
64	27.4836	0.916	29.989	0.000	25.683	29.284
65	21.8474	1.338	16.328	0.000	19.219	24.476
66	10.5970	0.301	35.251	0.000	10.007	11.187
68	17.1303	0.535	32.006	0.000	16.079	18.182
69	7.8658	0.698	11.265	0.000	6.494	9.237
70	8.5068	0.311	27.341	0.000	7.896	9.118
71	30.0833	0.785	38.322	0.000	28.541	31.625
73	20.8723	0.318	65.681	0.000	20.248	21.496
74	22.2811	0.444	50.133	0.000	21.408	23.154

75	25.4362	0.312	81.512	0.000	24.823	26.049
76	25.0184	0.836	29.925	0.000	23.376	26.661
77	12.5820	0.320	39.357	0.000	11.954	13.210
78	20.4654	1.003	20.411	0.000	18.496	22.435
79	22.5852	0.368	61.343	0.000	21.862	23.308
80	17.0120	0.329	51.662	0.000	16.365	17.659
81	26.9001	0.787	34.184	0.000	25.354	28.446
82	24.8377	0.310	80.226	0.000	24.230	25.446
83	21.7247	0.317	68.442	0.000	21.101	22.348
84	27.0421	0.847	31.914	0.000	25.378	28.706
85	25.0506	1.035	24.201	0.000	23.017	27.084
86	31.3070	0.670	46.714	0.000	29.991	32.623
87	31.2824	0.335	93.514	0.000	30.625	31.939
88	21.3756	0.493	43.335	0.000	20.407	22.345
89	29.5432	0.971	30.410	0.000	27.635	31.451
90	18.0340	0.594	30.356	0.000	16.867	19.201
91	6.4632	0.303	21.355	0.000	5.869	7.058
96	19.0047	1.314	14.461	0.000	16.423	21.586
97	23.6647	0.523	45.283	0.000	22.638	24.691
99	21.2186	0.526	40.364	0.000	20.186	22.251
103	24.2869	0.487	49.864	0.000	23.330	25.244
104	3.7786	0.299	12.629	0.000	3.191	4.366
105	22.1136	0.803	27.544	0.000	20.537	23.691
106	26.1631	0.953	27.440	0.000	24.290	28.036
107	19.0597	0.950	20.052	0.000	17.193	20.927
108	28.0423	0.505	55.580	0.000	27.051	29.033
110	20.9418	0.764	27.407	0.000	19.441	22.443
112	16.7197	0.328	50.928	0.000	16.075	17.365
113	24.3336	0.646	37.667	0.000	23.065	25.603
116	5.4992	1.125	4.888	0.000	3.289	7.709
117	32.6247	0.418	78.119	0.000	31.804	33.445
118	19.1861	0.377	50.931	0.000	18.446	19.926
120	6.7280	1.093	6.153	0.000	4.580	8.876
121	9.3797	0.400	23.433	0.000	8.593	10.166
122	19.9838	0.439	45.560	0.000	19.122	20.845
123	-1.8276	0.359	-5.085	0.000	-2.534	-1.122
124	22.8339	0.491	46.506	0.000	21.869	23.798
125	1.5962	0.310	5.151	0.000	0.988	2.205
126	1.8414	1.113	1.654	0.099	-0.345	4.028
127	3.8206	0.302	12.670	0.000	3.228	4.413
128	23.2999	0.299	78.015	0.000	22.713	23.887
129	27.3084	0.946	28.879	0.000	25.451	29.166
130	27.6813	0.845	32.771	0.000	26.022	29.341
131	18.4376	0.529	34.843	0.000	17.398	19.477
133	21.3365	1.441	14.810	0.000	18.507	24.166
134	32.2501	0.445	72.437	0.000	31.376	33.125
135	20.2468	0.313	64.623	0.000	19.631	20.862

136	23.7399	0.806	29.444	0.000	22.156	25.324
137	19.5174	0.546	35.723	0.000	18.444	20.591
138	18.2630	0.658	27.746	0.000	16.970	19.556
141	21.9594	1.119	19.620	0.000	19.761	24.158
143	30.1299	0.488	61.726	0.000	29.171	31.089
144	26.2062	0.363	72.261	0.000	25.494	26.919
145	24.7972	0.501	49.454	0.000	23.812	25.782
146	19.1933	1.145	16.761	0.000	16.944	21.443
147	17.3342	1.070	16.203	0.000	15.233	19.436
148	0.4522	0.322	1.406	0.160	-0.179	1.084
149	30.4385	0.319	95.285	0.000	29.811	31.066
150	12.4420	0.661	18.833	0.000	11.144	13.740
151	12.9902	0.446	29.100	0.000	12.113	13.867
152	30.2460	0.479	63.105	0.000	29.305	31.188
155	26.9204	0.305	88.353	0.000	26.322	27.519
159	21.5782	1.248	17.287	0.000	19.126	24.030
160	24.2438	0.912	26.595	0.000	22.453	26.034
161	26.8375	1.072	25.039	0.000	24.732	28.943
162	1.7515	0.335	5.233	0.000	1.094	2.409
164	21.0462	0.413	51.018	0.000	20.236	21.857
166	8.4735	0.312	27.122	0.000	7.860	9.087
167	22.3748	0.434	51.590	0.000	21.523	23.227
168	15.3669	0.794	19.355	0.000	13.807	16.926
169	13.9671	0.751	18.610	0.000	12.493	15.441
171	19.3878	1.077	18.002	0.000	17.272	21.503
172	24.0778	0.640	37.641	0.000	22.821	25.334
173	25.2607	0.319	79.149	0.000	24.634	25.888
174	27.5736	0.302	91.204	0.000	26.980	28.167
175	4.8546	0.354	13.728	0.000	4.160	5.549
176	4.4649	0.585	7.638	0.000	3.317	5.613
177	20.8299	1.030	20.214	0.000	18.806	22.854
178	25.8343	0.629	41.090	0.000	24.599	27.069
179	18.9778	1.624	11.689	0.000	15.789	22.167
180	20.9234	0.612	34.174	0.000	19.721	22.126
182	28.4090	0.356	79.903	0.000	27.711	29.107
183	16.3838	0.864	18.959	0.000	14.686	18.081
185	15.3062	1.359	11.265	0.000	12.637	17.975
186	20.7497	0.789	26.297	0.000	19.200	22.300
188	2.4244	0.404	5.994	0.000	1.630	3.219
190	-4.2545	0.366	-11.626	0.000	-4.973	-3.536

Omnibus:	99.732	Durbin-Watson:	1.399
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1106.312
Skew:	0.095	Prob(JB):	5.85e-241
Kurtosis:	9.156	Cond. No.	3.35e+05

```
Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 3.35e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

Question 3.i. Explain.

This happens when we look at the data of countries seem to have time as not as much of a feature and rather the inherent features as they change over time is what is causing this change.

Question 3.j. Give 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.

The way that human capital is changing overtime. As we gain more money per individuals, there is more money and a stronger economy at work.

Question 3.k. Perform regression with time fixed effects. Are the results consistent with your reasoning about the entity-invariant variables? The procedure for this question will be similar to 3.h. Drop the dummy variable for 1993 for this question.

This question is for your code, the next is for your explanation.

[29]:
```
# .get_dummies transforms a categorical variable into a dataframe of dummy␣
 ↪variables,
# one for each category. The prefix and prefix_sep part just makes sure the␣
 ↪variable
# names are strings and not integers.
countries = pd.get_dummies(who['year'], prefix='', prefix_sep='')
# This just joins the dummy dataframe with the original
who_country = who[['dale', 'hexp', 'hc3', 'sqhc3']].join(countries)
y_3h = who_country['dale']
# Here we drop country 191, since otherwise there would be perfect colinearity␣
 ↪in
# the columns. We also have to drop dale since that's the endogenous variable we
# regress on.
X_3h = sm.add_constant(who_country.drop(columns=['dale', '1993']))
model_3h = sm.OLS(y_3h, X_3h)
results_3h = model_3h.fit()
results_3h.summary()
```

[29]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                   dale   R-squared:                       0.672
Model:                            OLS   Adj. R-squared:                  0.669
Method:                 Least Squares   F-statistic:                     202.5
Date:                Fri, 16 Apr 2021   Prob (F-statistic):          8.44e-163
```

```
Time:                        11:57:46   Log-Likelihood:                -2357.7
No. Observations:                 700   AIC:                             4731.
Df Residuals:                     692   BIC:                             4768.
Df Model:                           7
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.8240      3.742      0.487      0.626      -5.524       9.172
hexp           0.0059      0.001     11.153      0.000       0.005       0.007
hc3           -4.8406      0.768     -6.304      0.000      -6.348      -3.333
sqhc3         34.4088      3.432     10.026      0.000      27.671      41.147
1994          -0.1197      0.844     -0.142      0.887      -1.777       1.538
1995          -0.2216      0.844     -0.262      0.793      -1.879       1.436
1996          -0.3647      0.845     -0.432      0.666      -2.023       1.294
1997          -0.4827      0.845     -0.571      0.568      -2.142       1.176
==============================================================================
Omnibus:                       86.787   Durbin-Watson:                   0.352
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              128.842
Skew:                          -0.857   Prob(JB):                     1.05e-28
Kurtosis:                       4.218   Cond. No.                     1.53e+04
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 1.53e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

Question 3.l. Explain.

Because of the variables given to us and how small they are, there is basically no way that time had an effect on the life expectancy from 1994 to 1997

Question 3.m. Perform a test that all time fixed effects are jointly equal to zero. Remember that we excluded 1993. What is the result of your test?

This question is for your code, the next is for your explanation.

[30]:
```
results_3h.f_test("1994, 1995, 1996, 1997").summary()
```

[30]:
```
'<F test: F=array([[0.1028111]]), p=0.9815064997603732, df_denom=692, df_num=4>'
```

Question 3.n. Explain.

At a signicance level of .05, we can assume that our null hypothesis (which is that they are all jointly equal to zero) still holds true and we cannot reject the null and accept the alternate because the p value is too high (0.98).