# DSE ECONOMETRICS PROJECT

NAME: DEBASMITA SAHA

CLASS ROLL NO.: ECON045

REGISTRATION NO.: 21201220045

PAPER NAME: APPLIED ECONOMETRICS (DSE)

# LIFE EXPECTANCY AROUND THE WORLD

Answer 1:

According to The World Bank, ***"Life expectancy at birth indicates the number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life".***

According to World Health Organization (WHO), ***"The average number of years that a newborn could expect to live, if he or she were to pass through life exposed to the sex- and age-specific death rates prevailing at the time of his or her birth, for a specific year, in a given country, territory, or geographic area".***

**For 2019:-**

```
.
. ***FOR 2019:
.
. *QUESTION 1:
. *describing all the variables used in the data
. desc

Contains data
  obs:           208
 vars:             8
 size:        17,888
--------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
--------------------------------------------------------------------------------
CountryName     str30   %30s                  Country Name
lexp            double  %14.2f                lexp
co2             double  %14.2f                co2
pcrate          double  %14.2f                pcrate
pun             double  %14.2f                pun
hexp            double  %14.2f                hexp
gdppc           double  %14.2f                gdppc
stat            double  %14.2f                stat
--------------------------------------------------------------------------------
Sorted by:
     Note: Dataset has changed since last saved.

.
```

```
. *regressing lexp with all other explanatory variables to get their relation
. ///with each other
> reg lexp co2 pcrate pun hexp gdppc stat

      Source |       SS           df       MS      Number of obs   =        74
-------------+----------------------------------   F(6, 67)        =     18.78
       Model |  1604.27303         6  267.378838   Prob > F        =    0.0000
    Residual |  954.088994        67  14.2401342   R-squared       =    0.6271
-------------+----------------------------------   Adj R-squared   =    0.5937
       Total |  2558.36202        73  35.0460551   Root MSE        =    3.7736

------------------------------------------------------------------------------
        lexp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         co2 |  -.4834918   .2793186    -1.73   0.088    -1.041014    .0740304
      pcrate |   .1161309   .0380442     3.05   0.003     .0401944    .1920674
         pun |  -.0895493   .0694143    -1.29   0.201    -.2281008    .0490021
        hexp |  -.0398276   .2026607    -0.20   0.845    -.4443401    .3646848
       gdppc |   .0004022   .0000884     4.55   0.000     .0002257    .0005786
        stat |   .0521342     .04474     1.17   0.248    -.0371672    .1414356
       _cons |    53.7987   4.457649    12.07   0.000     44.90119     62.6962
------------------------------------------------------------------------------

.
end of do-file
```

Here, *lexp* is the dependent (or explained) variable. *co2, pcrate, pun, hexp, gdppc* and *stat* are the independent (or explanatory) variables. All the variables mentioned above are continuous in nature. If we regress *lexp* with all other explanatory variables, then,

a. *co2* is statistically insignificant at 5% level of significance; which means as *co2* emission increases in each country, *lexp* decreases by 0.4833 years.
b. *pcrate* is also statistically significant at 5% level of significance; as *pcrate* increases by 1 unit, *lexp* increases by 0.116 years for each country.
c. *pun* is statistically insignificant at 5% level of significance, thus, as *pun* increases by 1%, *lexp* decreases by 0.0895 years for each country.
d. *hexp* is statistically insignificant at 5% level of significance; as *hexp* increases by 1% of GDP, *lexp* decreases by 0.0398 years.
e. *gdppc* is statistically significant at 5%  level of significance which indicates that, as *gdppc* increases by 1 unit, *lexp* increases by 0.0004 years.
f. *stat* is statistically insignificant at 5% level of significance; thus; as *stat* increases by 1 unit, *lexp* increases by 0.052 years for each country.
g. the constant term (= 53.79) is statistically significant at 5% level of significance.

Answer 2:

```
. *QUESTION 2:
. *generating a new variable deve which stores the value 0 if developing and
. /// 1 if developed
> *the development threshold for GDP(PPP) per capita of a developed country is
. ///atleast US $22,000
> *comparing gdppc of our data with the given value, we get
. gen deve = 0 if gdppc <= 22000
(90 missing values generated)

. replace deve = 1 if gdppc > 22000
(90 real changes made)

. replace deve = . if(missing(gdppc))
(20 real changes made, 20 to missing)
```

Here, we generated a new variable named *deve* which took the value ‘0’ if the country is developing and ‘1’if the country is developed. We also replaced the value of *deve* with ‘.’ when it satisfies neither of the conditions mentioned above.

We had compared the value of *gdppc* with the given value (= 22000). This is the development threshold for GDP(PPP) per capita for a developed country (= US $22,000).

[1]

[1] Source for the Development Threshold for GDP (PPP) per capita [https://en.wikipedia.org/wiki/Developed_country#:~:text=Another%20commonly%20used%20measure%20of,fit%20three%20out%20of%20four.&text=World%20map%20showing%20country%20classifications,(last%20updated%20April%202023).]

Answer 3:

```
. *QUESTION 3:
. *Finding the mean of all the variables except stat score and testing its
. ///significance for both developing and developed countries
> mean lexp co2 pcrate pun hexp gdppc if deve == 0

Mean estimation                   Number of obs   =         66

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
        lexp |   69.97071   .7210385      68.53069    71.41072
         co2 |   2.057725   .2319248      1.594539    2.520911
      pcrate |   90.38655   1.785383       86.8209    93.95221
         pun |   11.26515   1.142248      8.983926    13.54638
        hexp |   5.741879   .2935322      5.155655    6.328103
       gdppc |   9999.214    819.077      8363.404    11635.02
--------------------------------------------------------------

. mean lexp co2 pcrate pun hexp gdppc if deve == 1

Mean estimation                   Number of obs   =         40

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
        lexp |   79.76691   .4994316      78.75672    80.77711
         co2 |   7.869315   .7295501      6.393661     9.34497
      pcrate |   98.70799   .9928123      96.69984    100.7161
         pun |      2.875   .1943348      2.481921    3.268079
        hexp |   7.668487    .432519      6.793635    8.543339
       gdppc |   47506.83   3048.122      41341.42    53672.24
--------------------------------------------------------------
```

Here, we are finding the mean of all the variables except stat score for both developing and developed countries.

```
. ttest lexp, by(deve)
```

Two-sample t test with equal variances

| Group | Obs | Mean | Std. Err. | Std. Dev. | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| 0 | 125 | 69.21329 | .5656498 | 6.324158 | 68.09371 | 70.33287 |
| 1 | 63 | 79.63789 | .449573 | 3.568375 | 78.7392 | 80.53657 |
| combined | 188 | 72.70664 | .5413007 | 7.421942 | 71.6388 | 73.77448 |
| diff | | -10.4246 | .8589884 | | -12.11921 | -8.729986 |

```
    diff = mean(0) - mean(1)                                      t = -12.1359
Ho: diff = 0                                     degrees of freedom =      186

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

Here, Life expectancy at birth (*lexp*) varies between -12.11 and -8.72

Standard error for the difference in life expectancy at birth is very less (= 0.858)

The null hypothesis is given by,

Ho: diff = 0
From the t-test, we find that for,

a) Ha: diff < 0: Ho is rejected at 1% level of significance
b) Ha: diff != 0: Ho is rejected at 1% level of significance
c) Ha: diff > 0: Ho is accepted at 1% level of significance

```
. ttest co2, by(deve)
```

Two-sample t test with equal variances

| Group | Obs | Mean | Std. Err. | Std. Dev. | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| 0 | 124 | 2.057643 | .1831721 | 2.039718 | 1.695066 | 2.420221 |
| 1 | 56 | 8.757779 | .7606969 | 5.692534 | 7.233308 | 10.28225 |
| combined | 180 | 4.14213 | .3534802 | 4.742435 | 3.444606 | 4.839655 |
| diff | | -6.700135 | .577988 | | -7.840726 | -5.559545 |

```
    diff = mean(0) - mean(1)                                      t = -11.5922
Ho: diff = 0                                     degrees of freedom =      178

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

Here, CO2 emission (*co2*) varies between -7.841 and -5.55

Standard error for the difference in CO2 emission is very less (= 0.577)

The null hypothesis is given by,

Ho: diff = 0

From the t-test, we find that for,

a) Ha: diff < 0: Ho is rejected at 1% level of significance
b) Ha: diff != 0: Ho is rejected at 1% level of significance
c) Ha: diff > 0: Ho is accepted at 1% level of significance

```
. ttest pcrate, by(deve)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      80    89.53725    1.703021    15.23229    86.14747    92.92703
       1 |      47    98.87777    .8734635    5.988164    97.11958     100.636
---------+--------------------------------------------------------------------
combined |     127    92.99398     1.18742    13.38154    90.64411    95.34384
---------+--------------------------------------------------------------------
    diff |           -9.340525    2.323499               -13.93902   -4.742031
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -4.0200
Ho: diff = 0                                     degrees of freedom =      125

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0001          Pr(T > t) = 1.0000
```

Here, primary completion rate (*pcrate*) varies between -13.93 and -4.74

Standard error for the difference in primary completion rate is 2.323

The null hypothesis is given by,

Ho: diff = 0

From the t-test, we find that for,

a) Ha: diff < 0: Ho is rejected at 1% level of significance
b) Ha: diff != 0: Ho is rejected at 1% level of significance
c) Ha: diff > 0: Ho is accepted at 1% level of significance

```
. ttest pun, by(deve)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |     102    12.31765    1.093659    11.04542    10.14812    14.48717
       1 |      52    3.009615    .1831363    1.320615    2.641954    3.377277
---------+--------------------------------------------------------------------
combined |     154    9.174675    .8083227    10.03102    7.577761    10.77159
---------+--------------------------------------------------------------------
    diff |            9.308032    1.539719                6.266018    12.35005
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   6.0453
Ho: diff = 0                                     degrees of freedom =      152

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

Here, percentage of undernourished population *(pun)* varies between 6.266 and 12.35

Standard error for the difference in percentage of undernourished population is 1.53

The null hypothesis is given by,

Ho: diff = 0
From the t-test, we find that for,

a) Ha: diff < 0: Ho is accepted at 1% level of significance
b) Ha: diff != 0: Ho is rejected at 1% level of significance
c) Ha: diff > 0: Ho is rejected at 1% level of significance

```
. ttest hexp, by(deve)

Two-sample t test with equal variances
```

| Group | Obs | Mean | Std. Err. | Std. Dev. | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| 0 | 121 | 5.927667 | .2598077 | 2.857884 | 5.413266 | 6.442068 |
| 1 | 55 | 7.68739 | .371519 | 2.755259 | 6.942539 | 8.432241 |
| combined | 176 | 6.47758 | .2212076 | 2.93465 | 6.041002 | 6.914158 |
| diff | | -1.759723 | .4596435 | | -2.666918 | -.8525287 |

```
    diff = mean(0) - mean(1)                                      t =  -3.8285
Ho: diff = 0                                     degrees of freedom =      174

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0001         Pr(|T| > |t|) = 0.0002          Pr(T > t) = 0.9999
```

Here, health expenditure (% of GDP) (*hexp*) varies between -2.66 and -8.52

Standard error for the difference in health expenditure (% of GDP) is very less (=0.459)

The null hypothesis is given by,

Ho: diff = 0
From the t-test, we find that for,

a) Ha: diff < 0: Ho is rejected at 1% level of significance
b) Ha: diff != 0: Ho is rejected at 1% level of significance
c) Ha: diff > 0: Ho is accepted at 10% level of significance

```
. ttest gdppc, by(deve)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |     125    9509.119    590.3338    6600.133    8340.683    10677.55
       1 |      63    49745.26    2762.249    21924.68     44223.6    55266.92
---------+--------------------------------------------------------------------
combined |     188    22992.51    1711.827    23471.39    19615.53    26369.48
---------+--------------------------------------------------------------------
    diff |           -40236.14    2125.673               -44429.67   -36042.61
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t = -18.9287
Ho: diff = 0                                     degrees of freedom =      186

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000

.
```

Here, GDP per capita (*gdppc*) varies between -44429.67 and -36042.61

Standard error for the difference in GDP per capita is very high (=2125.673)

The null hypothesis is given by,

Ho: diff = 0

From the t-test, we find that for,

a) Ha: diff < 0: Ho is rejected at 1% level of significance
b) Ha: diff != 0: Ho is rejected at 1% level of significance
c) Ha: diff > 0: Ho is accepted at 1% level of significance

## Answer 4:

```
. *QUESTION 4:
. *Taking log of GDP per capita
. gen lgdppc = ln(gdppc)
(20 missing values generated)

. *creating a dummy variable from stat score
. gen ss = 1 if stat <= 25
(207 missing values generated)

. replace ss = 2 if stat > 25 & stat <= 50
(29 real changes made)

. replace ss = 3 if stat > 50 & stat <= 75
(70 real changes made)

. replace ss = 4 if stat > 75 & stat <= 100
(42 real changes made)

. replace ss = . if (missing(stat))
(0 real changes made)

.
end of do-file
```

Here, we took log of *gdppc* and stored it in *lgdppc*. Next, we are generating s new dummy variable from stat score named *'ss'* which stores the values '1', '2', '3', '4' and '.'under the following conditions given.

# Answer 5:

```
. *QUESTION 5:
. *running a regression on lexp and all other explanatory variables including
. ///the dummy variable
> reg lexp co2 pcrate pun hexp lgdppc stat i.ss

      Source |       SS           df       MS      Number of obs   =        74
-------------+----------------------------------   F(8, 65)        =     16.68
       Model |  1720.50292         8  215.062865   Prob > F        =    0.0000
    Residual |  837.859097        65    12.89014   R-squared       =    0.6725
-------------+----------------------------------   Adj R-squared   =    0.6322
       Total |  2558.36202        73  35.0460551   Root MSE        =    3.5903

------------------------------------------------------------------------------
        lexp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         co2 |  -.4530016   .2487928    -1.82   0.073    -.9498751    .0438719
      pcrate |   .0826651   .0374112     2.21   0.031     .0079497    .1573805
         pun |   .0115544    .074892     0.15   0.878    -.1380152     .161124
        hexp |  -.0621712   .1938881    -0.32   0.750    -.4493924    .3250501
      lgdppc |   4.367693   .9802781     4.46   0.000     2.409942    6.325443
        stat |   .0072442   .0777578     0.09   0.926    -.1480489    .1625374
             |
          ss |
          3  |  -1.564096   2.296125    -0.68   0.498    -6.149775    3.021583
          4  |   1.400875   3.568512     0.39   0.696    -5.725936    8.527685
             |
       _cons |   24.51965   10.02219     2.45   0.017     4.503957    44.53533
------------------------------------------------------------------------------

.
```

Here, we are regressing life expectancy at birth (*lexp*) with *co2, pcrate, pun, hexp, lgdppc, stat*; which are the explanatory variables in our data along with the newly created dummy variable named *'ss'*. Here, while regressing, we considered *'i.ss'* for the dummy variable. This is because ***i.*** stands for the intercept dummy of the categorical variable *ss* in our data. All other variables are continuous in nature.

Now, from the regression, we find that,

- *co2, pun, hexp, stat* are statistically insignificant at 5% level of significance
- *pcrate and lgdppc* are statistically significant at 5% level of significance
- if *co2* increases by 1 kiloton (kt), *lexp* decreases by 0.453 years
- if *pcrate* increases by 1 unit, *lexp* increases by 0.082 years

- if *pun* increases by 1 member, *lexp* increases by 0.115 years
- if *hexp* increases by 1 percent, *lexp* decreases by 0.062 years
- if *lgdppc* increases by 1 US $, *lexp* increases by 4.367 years
- if *stat* increases by 1 score, *lexp* increases by 0.0072 years

For the dummy variable *ss*,

- compared to s = 1 and s = 2, if stat score lies between 50 and 75, *lexp* decreases by 1.564 years
- compared to s = 1 and s = 2, if stat score lies between 75 and 100, *lexp* increases by 1.4 years

However, the dummy variable s is statistically insignificant at 5% level of significance.

The constant term of the regression is 24.51 which is statistically significant at 5% level of significance.

## Answer 6:

Regression Diagnostics:-

```
. predict lexp1
(option xb assumed; fitted values)
(134 missing values generated)

. predict res, residuals
(134 missing values generated)

. *checking for normality
. *(jarque-bera test)
. jb res
Jarque-Bera normality test:  1.353 Chi(2)  .5084
Jarque-Bera test for Ho: normality:
```

Jarque-Bera test is a no (no parameter) test where the null hypothesis is presented as,

Ho: normality

The value of Jarque-Bera normality test is 1.353

Here, the value of Chi(2) = 0.5084 with 2 degrees of freedom

This is statistically insignificant at 5% level of significance and we reject the null hypothesis of normality.

```
. histogram res, normal
(bin=8, start=-10.254249, width=2.1857467)

.
end of do-file

. graph save Graph "F:\Stata MP 14.2\MD PROJECT\histogram 2019.gph"
(file F:\Stata MP 14.2\MD PROJECT\histogram 2019.gph saved)

. do "C:\Users\student\AppData\Local\Temp\STD00000000.tmp"

. rvfplot, yline(0)

.
end of do-file

. graph save Graph "F:\Stata MP 14.2\MD PROJECT\rvfplot 2019.gph"
(file F:\Stata MP 14.2\MD PROJECT\rvfplot 2019.gph saved)
```

We had plotted a histogram where we found, the histogram is very mean-centric (i.e. it is leptokurtic in nature).

We had also plotted a scatterplot of residuals versus fitted values where we found there is no such relation between the two axes. It is not depicting any particular shape (or a pattern).

Thus, there is no presence of heteroscedasticity in our data.

```
. *checking for homoscedasticity
. *# graphical method
. rvfplot, yline(0)

.
. *# formal test
. hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of lexp

         chi2(1)      =     6.50
         Prob > chi2  =   0.0108
```

After performing the graphical method for detecting heteroscedasticity, we then performed the formal method for detecting heteroscedasticity.

Here, we found that for the Breusch-Pagan/ Cook-Weisberg test for heteroscedasticity, the null hypothesis is presented as,

Ho: Constant variance

The chi2(1) value is 6.50 which is greater than the critical value (= 3.84), thus we reject the null hypothesis of constant variance. The chi2(1) value is statistically insignificant at 5% level of significance.

Thus, there is a presence of heteroscedasticity in our data.

```
. reg lexp co2 pcrate pun hexp lgdppc stat i.ss, robust
```

Linear regression

| | | |
|---|---|---|
| Number of obs | = | 74 |
| F(8, 65) | = | 24.83 |
| Prob > F | = | 0.0000 |
| R-squared | = | 0.6725 |
| Root MSE | = | 3.5903 |

| lexp | Coef. | Robust Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| co2 | -.4530016 | .1957965 | -2.31 | 0.024 | -.8440343 | -.0619689 |
| pcrate | .0826651 | .0315637 | 2.62 | 0.011 | .0196281 | .1457021 |
| pun | .0115544 | .0899919 | 0.13 | 0.898 | -.1681718 | .1912806 |
| hexp | -.0621712 | .2131474 | -0.29 | 0.771 | -.4878559 | .3635135 |
| lgdppc | 4.367693 | .9961031 | 4.38 | 0.000 | 2.378337 | 6.357048 |
| stat | .0072442 | .0852536 | 0.08 | 0.933 | -.163019 | .1775075 |
| ss | | | | | | |
| 3 | -1.564096 | 2.731703 | -0.57 | 0.569 | -7.019684 | 3.891492 |
| 4 | 1.400875 | 3.853447 | 0.36 | 0.717 | -6.294991 | 9.09674 |
| _cons | 24.51965 | 9.840201 | 2.49 | 0.015 | 4.867408 | 44.17188 |

Next, we performed regression with robust standard errors. "Robust" standard errors is a technique to obtain unbiased standard errors of OLS coefficients under heteroscedasticity.

Now, some of the coefficients are statistically significant at 5% level of significance.

```
*checking for multicollinearity
vif
```

| Variable | VIF | 1/VIF |
|---|---|---|
| co2 | 2.29 | 0.437030 |
| pcrate | 1.56 | 0.640425 |
| pun | 2.68 | 0.373562 |
| hexp | 1.17 | 0.857127 |
| lgdppc | 4.39 | 0.227771 |
| stat | 5.48 | 0.182362 |
| ss | | |
| 3 | 7.57 | 0.132158 |
| 4 | 18.06 | 0.055363 |
| Mean VIF | 5.40 | |

We can use vif command after the regression to check for multicollinearity. As a rule of thumb, a variable whose vif value is > 10 may merit further investigation.

Here, the vif value = 5.40 which is less than 10, thus there is no multicollinearity issue in our data and it is a full column rank matrix.

```
. *descriptive statistics
. desc lexp - stat

              storage   display    value
variable name   type    format     label      variable label
------------------------------------------------------------------------------------------
lexp            double  %14.2f                lexp
co2             double  %14.2f                co2
pcrate          double  %14.2f                pcrate
pun             double  %14.2f                pun
hexp            double  %14.2f                hexp
gdppc           double  %14.2f                gdppc
stat            double  %14.2f                stat

. sum lexp - stat

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        lexp |        206    72.96525    7.449954      52.91   85.18049
         co2 |        188    4.040135    4.674155   .0337149    31.8772
      pcrate |        132    92.64288     13.5823   54.72869   120.4473
         pun |        160     9.60625    10.62906        2.5       54.8
        hexp |        180    6.458008     2.94654   2.017115   20.88979
-------------+---------------------------------------------------------
       gdppc |        188    22992.51    23471.39   760.6041   128031.2
        stat |        142     64.5227    16.66675   16.66667   96.66667


 corr lexp - stat
(obs=74)

             |     lexp      co2   pcrate      pun     hexp    gdppc     stat
-------------+---------------------------------------------------------------
        lexp |   1.0000
         co2 |   0.5126   1.0000
      pcrate |   0.5587   0.4250   1.0000
         pun |  -0.5989  -0.5379  -0.4863   1.0000
        hexp |   0.0901   0.0810   0.1475   0.1192   1.0000
       gdppc |   0.7161   0.7761   0.4275  -0.6208   0.1387   1.0000
        stat |   0.5399   0.5004   0.4057  -0.5066   0.1236   0.5763   1.0000
```

There is high correlation among the variables as the values are greater than or equal to 0.5

```
. *principle component analysis (PCA)
. factor lexp - stat, pcf
(obs=74)

Factor analysis/correlation                        Number of obs    =         74
    Method: principal-component factors            Retained factors =          2
    Rotation: (unrotated)                          Number of params =         13
```

| Factor | Eigenvalue | Difference | Proportion | Cumulative |
|---|---|---|---|---|
| Factor1 | 3.76368 | 2.69533 | 0.5377 | 0.5377 |
| Factor2 | 1.06835 | 0.37921 | 0.1526 | 0.6903 |
| Factor3 | 0.68915 | 0.15493 | 0.0984 | 0.7887 |
| Factor4 | 0.53422 | 0.10624 | 0.0763 | 0.8651 |
| Factor5 | 0.42798 | 0.06027 | 0.0611 | 0.9262 |
| Factor6 | 0.36770 | 0.21877 | 0.0525 | 0.9787 |
| Factor7 | 0.14893 | . | 0.0213 | 1.0000 |

```
    LR test: independent vs. saturated:  chi2(21) =  238.08 Prob>chi2 = 0.0000

Factor loadings (pattern matrix) and unique variances
```

| Variable | Factor1 | Factor2 | Uniqueness |
|---|---|---|---|
| lexp | 0.8343 | -0.0054 | 0.3040 |
| co2 | 0.7987 | -0.0157 | 0.3618 |
| pcrate | 0.6822 | 0.1208 | 0.5201 |
| pun | -0.7851 | 0.3296 | 0.2749 |
| hexp | 0.1306 | 0.9697 | 0.0426 |
| gdppc | 0.8836 | 0.0229 | 0.2188 |
| stat | 0.7418 | 0.0627 | 0.4459 |

a. There are two factors (retained factors = 2) which satisfies the kaiser criterion. The first 2 eigenvalues are 3.76 and 1.06.
b. We performed the factor analysis, where we considered rotation (unrotated) to understand the factor loadings.
c. We perform LR test: independent vs saturated (correlated to each other)
d. The null hypothesis is Ho: independence (i.e. no correlation) (or, zero correlation)
e. The value of chi2(21) = 238.08 and Prob>chi2 = 0.0000
f. The null hypothesis, Ho is rejected at 1% level of significance; thus, running factor analysis is a good idea.

Now,

1. *lexp* has as much as 83.43% commonness with factor 1
2. *co2* has as much as 79.87% commonness with factor 1
3. *pcrate* has as much as 68.22% commonness with factor 1
4. *hexp* has as much as 13.06% commonness with factor 1
5. *gdppc* has as much as 88.36% commonness with factor 1
6. *stat* has as much as 74.18% commonness with factor 1

```
. estat kmo

Kaiser-Meyer-Olkin measure of sampling adequacy
```

| Variable | kmo |
|---|---|
| lexp | 0.7967 |
| co2 | 0.7646 |
| pcrate | 0.7933 |
| pun | 0.8593 |
| hexp | 0.3286 |
| gdppc | 0.7294 |
| stat | 0.9377 |
| Overall | 0.7907 |

```
. factortest lexp - stat

Determinant of the correlation matrix
Det                 =     0.035


Bartlett test of sphericity

Chi-square          =           234.723
Degrees of freedom =                21
p-value             =             0.000
H0: variables are not intercorrelated


Kaiser-Meyer-Olkin Measure of Sampling Adequacy
KMO                =     0.791
```

Here, kmo value is 0.7907 which is close to 1, thus we have adequate data to run factor analysis.

For the Bartlett's test of Sphericity,

1. the null hypothesis is Ho: variables are not intercorrelated
2. the value of chi-square is 234.723 with 21 degrees of freedom and the p-value is 0.0000
3. we reject the null hypothesis that variables are not intercorrelated at 5% level of significance.
4. Therefore, factor analysis is a good idea.

```
. *scree plot
. screeplot

. screeplot, yline(1)

.
end of do-file

. graph save Graph "D:\Stata MP 14.2\MD PROJECT\screeplot 2019.gph"
(file D:\Stata MP 14.2\MD PROJECT\screeplot 2019.gph saved)
```

We had then plotted the screeplot which is a graphical method of detecting and dropping the eigenvalues which are less than 1 since these provide less information than is provided by a single variable.

```
. *rotation
. *# orthogonal rotation
. rotate, varimax

Factor analysis/correlation                     Number of obs    =         74
    Method: principal-component factors         Retained factors =          2
    Rotation: orthogonal varimax (Kaiser off)   Number of params =         13
```

| Factor | Variance | Difference | Proportion | Cumulative |
|---|---|---|---|---|
| Factor1 | 3.73652 | 2.64102 | 0.5338 | 0.5338 |
| Factor2 | 1.09550 | . | 0.1565 | 0.6903 |

```
LR test: independent vs. saturated:  chi2(21) =  238.08 Prob>chi2 = 0.0000
```

Rotated factor loadings (pattern matrix) and unique variances

| Variable | Factor1 | Factor2 | Uniqueness |
|---|---|---|---|
| lexp | 0.8306 | 0.0784 | 0.3040 |
| co2 | 0.7963 | 0.0646 | 0.3618 |
| pcrate | 0.6666 | 0.1887 | 0.5201 |
| pun | -0.8142 | 0.2492 | 0.2749 |
| hexp | 0.0326 | 0.9779 | 0.0426 |
| gdppc | 0.8768 | 0.1115 | 0.2188 |
| stat | 0.7317 | 0.1368 | 0.4459 |

Factor rotation matrix

| | Factor1 | Factor2 |
|---|---|---|
| Factor1 | 0.9950 | 0.1004 |
| Factor2 | -0.1004 | 0.9950 |

This is an orthogonal rotation.

Here, variance = 3.736 has been extracted by factor 1 and variance = 1.095 has been extracted by factor 2

The cumulative variance is 0.6903

However, all the variables are loaded on both the factors

```
. rotate, varimax blanks(.49)

Factor analysis/correlation                        Number of obs    =         74
    Method: principal-component factors            Retained factors =          2
    Rotation: orthogonal varimax (Kaiser off)      Number of params =         13
```

| Factor | Variance | Difference | Proportion | Cumulative |
|---|---|---|---|---|
| Factor1 | 3.73652 | 2.64102 | 0.5338 | 0.5338 |
| Factor2 | 1.09550 | . | 0.1565 | 0.6903 |

```
    LR test: independent vs. saturated:  chi2(21) =  238.08 Prob>chi2 = 0.0000

Rotated factor loadings (pattern matrix) and unique variances
```

| Variable | Factor1 | Factor2 | Uniqueness |
|---|---|---|---|
| lexp | 0.8306 | | 0.3040 |
| co2 | 0.7963 | | 0.3618 |
| pcrate | 0.6666 | | 0.5201 |
| pun | -0.8142 | | 0.2749 |
| hexp | | 0.9779 | 0.0426 |
| gdppc | 0.8768 | | 0.2188 |
| stat | 0.7317 | | 0.4459 |

```
    (blanks represent abs(loading)<.49)

Factor rotation matrix
```

| | Factor1 | Factor2 |
|---|---|---|
| Factor1 | 0.9950 | 0.1004 |
| Factor2 | -0.1004 | 0.9950 |

We had considered 0.49 as the rule of thumb. Thus, values which are less than 0.49 are represented by blanks in the pattern matrix.

Therefore, *lexp* to *pun* and *gdppc* and *stat* are now loaded on factor 1

But, *hexp* is now loaded on factor 2

```
. *saving factor score
. predict pc1 pc2
(regression scoring assumed)

Scoring coefficients (method = regression; based on varimax rotated factors)
```

| Variable | Factor1 | Factor2 |
|---|---|---|
| lexp | 0.22105 | 0.01723 |
| co2 | 0.21262 | 0.00669 |
| pcrate | 0.16898 | 0.13073 |
| pun | -0.23852 | 0.28605 |
| hexp | -0.05657 | 0.90658 |
| gdppc | 0.23143 | 0.04489 |
| stat | 0.19020 | 0.07815 |

```
. estat common

Correlation matrix of the varimax rotated common factors
```

| Factors | Factor1 | Factor2 |
|---|---|---|
| Factor1 | 1 | |
| Factor2 | 0 | 1 |

```
. corr pc1 pc2
(obs=74)
```

| | pc1 | pc2 |
|---|---|---|
| pc1 | 1.0000 | |
| pc2 | 0.0000 | 1.0000 |

We find that, there is no correlation in orthogonal rotation.

```
. *# oblique rotation
. rotate, promax

Factor analysis/correlation                        Number of obs    =         74
    Method: principal-component factors           Retained factors =          2
    Rotation: oblique promax (Kaiser off)          Number of params =         13
```

| Factor | Variance | Proportion | Rotated factors are correlated |
|---|---|---|---|
| Factor1 | 3.76003 | 0.5371 | |
| Factor2 | 1.14107 | 0.1630 | |

LR test: independent vs. saturated: chi2(21) = 238.08 Prob>chi2 = 0.0000

Rotated factor loadings (pattern matrix) and unique variances

| Variable | Factor1 | Factor2 | Uniqueness |
|---|---|---|---|
| lexp | 0.8306 | 0.0255 | 0.3040 |
| co2 | 0.7970 | 0.0138 | 0.3618 |
| pcrate | 0.6585 | 0.1470 | 0.5201 |
| pun | -0.8355 | 0.3030 | 0.2749 |
| hexp | -0.0307 | 0.9819 | 0.0426 |
| gdppc | 0.8750 | 0.0558 | 0.2188 |
| stat | 0.7274 | 0.0906 | 0.4459 |

Factor rotation matrix

| | Factor1 | Factor2 |
|---|---|---|
| Factor1 | 0.9993 | 0.1643 |
| Factor2 | -0.0368 | 0.9864 |

This is oblique rotation.

Here, variance = 3.76 has been extracted by factor 1 and variance = 1.141 has been extracted by factor 2

There is correlation in oblique rotation as the rotated factors are correlated.

However, all the variables are loaded on both the factors

```
. rotate, promax blanks(.44)

Factor analysis/correlation                        Number of obs    =        74
    Method: principal-component factors            Retained factors =         2
    Rotation: oblique promax (Kaiser off)          Number of params =        13
```

| Factor | Variance | Proportion | Rotated factors are correlated |
|---|---|---|---|
| Factor1 | 3.76003 | 0.5371 | |
| Factor2 | 1.14107 | 0.1630 | |

```
    LR test: independent vs. saturated:  chi2(21) =  238.08 Prob>chi2 = 0.0000
```

Rotated factor loadings (pattern matrix) and unique variances

| Variable | Factor1 | Factor2 | Uniqueness |
|---|---|---|---|
| lexp | 0.8306 | | 0.3040 |
| co2 | 0.7970 | | 0.3618 |
| pcrate | 0.6585 | | 0.5201 |
| pun | -0.8355 | | 0.2749 |
| hexp | | 0.9819 | 0.0426 |
| gdppc | 0.8750 | | 0.2188 |
| stat | 0.7274 | | 0.4459 |

(blanks represent abs(loading)<.44)

Factor rotation matrix

| | Factor1 | Factor2 |
|---|---|---|
| Factor1 | 0.9993 | 0.1643 |
| Factor2 | -0.0368 | 0.9864 |

We had considered 0.44 as the rule of thumb. Thus, values which are less than 0.44 are represented by blanks in the pattern matrix.

Therefore, *lexp* to *pun* and *gdppc* and *stat* are now loaded on factor 1

But, *hexp* is now loaded on factor 2

```
. *saving factor score
. predict pc3 pc4
(regression scoring assumed)
```

Scoring coefficients (method = regression; based on promax(3) rotated factors)

| Variable | Factor1 | Factor2 |
|---|---|---|
| lexp | 0.22170 | 0.03144 |
| co2 | 0.21261 | 0.02037 |
| pcrate | 0.17697 | 0.14134 |
| pun | -0.21981 | 0.27009 |
| hexp | 0.00131 | 0.90105 |
| gdppc | 0.23382 | 0.05970 |
| stat | 0.19479 | 0.09024 |

```
. estat common
```

Correlation matrix of the promax(3) rotated common factors

| Factors | Factor1 | Factor2 |
|---|---|---|
| Factor1 | 1 | |
| Factor2 | .1279 | 1 |

```
. corr pc3 pc4
(obs=74)
```

| | pc3 | pc4 |
|---|---|---|
| pc3 | 1.0000 | |
| pc4 | 0.1279 | 1.0000 |

There is some amount of correlation in oblique rotation, however no presence of multicollinearity in the data.

```
. linktest

      Source |       SS           df       MS      Number of obs   =        74
-------------+----------------------------------   F(2, 71)        =     73.10
       Model |  1722.0889          2  861.044448   Prob > F        =    0.0000
    Residual |  836.273124        71  11.7784947   R-squared       =    0.6731
-------------+----------------------------------   Adj R-squared   =    0.6639
       Total |  2558.36202        73  35.0460551   Root MSE        =     3.432

------------------------------------------------------------------------------
        lexp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        _hat |   .0641803   2.551769     0.03   0.980    -5.023902    5.152263
      _hatsq |    .006723   .0183225     0.37   0.715     -.029811     .043257
       _cons |   32.40201   88.50066     0.37   0.715    -144.0633    208.8673
------------------------------------------------------------------------------
```

Here, we have performed model misspecification tests. For the linktest, we get,

a. *lexp* has been regressed on *hat* and *hatsq*
b. here, *hat* is significant as it should be at 1% level of significance and *hatsq* in insignificant at 1% level of significance.

Thus, model is highly misspecified.

```
. ovtest

Ramsey RESET test using powers of the fitted values of lexp
       Ho:  model has no omitted variables
                  F(3, 62) =      0.69
                  Prob > F =      0.5627
```

For the ovtest, we get,

a. the null hypothesis is given by, Ho: model has no omitted variables
b. F(3,62) = 0.69 < 2.68 (the critical value of F(3,62) at 5% level of significance); thus we reject the null hypothesis that the model has no omitted variables at 5% level of significance.
c. Prob > F = 0.5627 > 0.05; thus it is insignificant at 5% level of significance.
d. Thus, the model has a lot of explanatory variables which are omitted variables.

Answer 1:

# For 2020:

```
. ***FOR 2020:
.
. *QUESTION 1:
. *describing all the variables used in the data
. desc

Contains data
  obs:           208
 vars:             8
 size:        17,888
----------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
----------------------------------------------------------------------------------
CountryName     str30   %30s                  Country Name
lexp            double  %14.2f                lexp
co3             double  %14.2f                co3
pcrate          double  %14.2f                pcrate
pun             double  %14.2f                pun
hexp            double  %14.2f                hexp
gdppc           double  %14.2f                gdppc
stat            double  %14.2f                stat
----------------------------------------------------------------------------------
Sorted by:
     Note: Dataset has changed since last saved.
```

```
. *regressing lexp with all other explanatory variables to get their relation
. ///with each other
> reg lexp co3 pcrate pun hexp gdppc stat

      Source |       SS           df       MS      Number of obs   =        60
-------------+----------------------------------   F(6, 53)        =     18.95
       Model |  1247.01478         6  207.835797   Prob > F        =    0.0000
    Residual |  581.312867        53  10.9681673   R-squared       =    0.6821
-------------+----------------------------------   Adj R-squared   =    0.6461
       Total |  1828.32765        59  30.9886042   Root MSE        =    3.3118

------------------------------------------------------------------------------
        lexp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         co3 |  -.3202865   .2918929    -1.10   0.277    -.9057496    .2651766
      pcrate |   .1247831   .0398583     3.13   0.003     .0448375    .2047288
         pun |  -.1328368   .0717487    -1.85   0.070    -.2767466    .0110729
        hexp |   .1481946    .194948     0.76   0.451    -.2428216    .5392108
       gdppc |   .0003295   .0000836     3.94   0.000     .0001618    .0004972
        stat |   .0342244   .0388422     0.88   0.382    -.0436831    .1121319
       _cons |   53.60646   4.744244    11.30   0.000     44.09071    63.12221
------------------------------------------------------------------------------
```

Here, *lexp1* is the dependent (or explained) variable. *co3*, *pcrate1*, *pun1*, *hexp1*, *gdppc1* and *stat1* are the independent (or explanatory) variables. All the variables mentioned above are continuous in nature. If we regress *lexp1* with all other explanatory variables, then,

a. *co3* is statistically insignificant at 5% level of significance; which means as *co3* emission increases in each country, *lexp1* decreases by 0.3208 years.
b. *pcrate1* is also statistically significant at 5% level of significance; as *pcrate1* increases by 1 unit, *lexp1* increases by 0.124 years for each country.
c. *pun1* is statistically insignificant at 5% level of significance, thus, as *pun1* increases by 1%, *lexp1* decreases by 0.132 years for each country.
d. *hexp1* is statistically insignificant at 5% level of significance; as *hexp1* increases by 1% of GDP, *lexp1* decreases by 0.148 years.
e. *gdppc1* is statistically significant at 5% level of significance which indicates that, as *gdppc1* increases by 1 unit, *lexp1* increases by 0.0003 years.
f. *stat1* is statistically insignificant at 5% level of significance; thus; as *stat1* increases by 1 unit, *lexp1* increases by 0.034 years for each country.
g. the constant term (=53.60) is statistically significant at 5% level of significance.

Answer 2:

```
. *QUESTION 2:
. *generating a new variable deve which stores the value 0 if developing and
. /// 1 if developed
> *the development threshold for GDP(PPP) per capita of a developed country is
. ///atleast US $22,000
> *comparing gdppc of our data with the given value, we get
. gen deve = 0 if gdppc <= 22000
(84 missing values generated)

. replace deve = 1 if gdppc > 22000
(84 real changes made)

. replace deve = . if(missing(gdppc))
(21 real changes made, 21 to missing)
```

Here, we generated a new variable named *deve* which took the value ‘0’ if the country is developing and ‘1’if the country is developed. We also replaced the value of *deve* with ‘.’ when it satisfies neither of the conditions mentioned above.

We had compared the value of *gdppc* with the given value (= 22000). This is the development threshold for GDP(PPP) per capita for a developed country (= US $22,000).

Answer 3:

```
. *QUESTION 3:
. *Finding the mean of all the variables except stat score and testing its
. ///significance for both developing and developed countries
> mean lexp co3 pcrate pun hexp gdppc if deve == 0

Mean estimation                   Number of obs    =         52
```

| | Mean | Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| lexp | 69.99173 | .7648911 | 68.45615 | 71.52732 |
| co3 | 2.061425 | .2467412 | 1.566071 | 2.556778 |
| pcrate | 92.79357 | 1.94169 | 88.89547 | 96.69168 |
| pun | 10.20577 | 1.207201 | 7.782213 | 12.62933 |
| hexp | 6.289619 | .3329459 | 5.621202 | 6.958035 |
| gdppc | 10138.76 | 877.9201 | 8376.259 | 11901.26 |

```
. mean lexp co3 pcrate pun hexp gdppc if deve == 1

Mean estimation                   Number of obs    =         40
```

| | Mean | Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| lexp | 79.11657 | .5190952 | 78.0666 | 80.16654 |
| co3 | 7.09856 | .7912014 | 5.498204 | 8.698916 |
| pcrate | 98.7742 | .9477576 | 96.85718 | 100.6912 |
| pun | 2.98 | .2237845 | 2.527353 | 3.432647 |
| hexp | 8.688491 | .4654086 | 7.747113 | 9.629869 |
| gdppc | 46002.95 | 3122.254 | 39687.6 | 52318.31 |

Here, we are finding the mean of all the variables except stat score for both developing and developed countries.

```
. *executing t-test among two group of countries
. ttest lexp, by(deve)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |     124    68.50131    .5489023    6.112317    67.41479    69.58783
       1 |      63    79.10329    .4643617    3.685756    78.17505    80.03154
---------+--------------------------------------------------------------------
combined |     187     72.0731     .539755    7.381038    71.00827    73.13793
---------+--------------------------------------------------------------------
    diff |           -10.60198    .8387973               -12.25682   -8.947144
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t = -12.6395
Ho: diff = 0                                     degrees of freedom =      185

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

Here, Life expectancy at birth (*lexp*) varies between -12.25 and -8.94

Standard error for the difference in life expectancy at birth is very less (= 0.838)

The null hypothesis is given by,

Ho: diff = 0

From the t-test, we find that for,

a) Ha: diff < 0: Ho is rejected at 1% level of significance
b) Ha: diff != 0: Ho is rejected at 1% level of significance
c) Ha: diff > 0: Ho is accepted at 1% level of significance

```
. ttest co3, by(deve)
```

Two-sample t test with equal variances

| Group | Obs | Mean | Std. Err. | Std. Dev. | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| 0 | 122 | 1.807898 | .1608685 | 1.77685 | 1.489416 | 2.126379 |
| 1 | 57 | 8.187992 | .7786513 | 5.878689 | 6.628165 | 9.747818 |
| combined | 179 | 3.839548 | .349803 | 4.680045 | 3.149253 | 4.529842 |
| diff | | -6.380094 | .5805183 | | -7.525722 | -5.234466 |

```
    diff = mean(0) - mean(1)                                      t = -10.9903
Ho: diff = 0                                     degrees of freedom =      177

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

Here, CO3 emission (*co3*) varies between -7.525 and -5.234

Standard error for the difference in CO3 emission is very less (= 0.58)

The null hypothesis is given by,

Ho: diff = 0
From the t-test, we find that for,

a) Ha: diff < 0: Ho is rejected at 1% level of significance
b) Ha: diff != 0: Ho is rejected at 1% level of significance
c) Ha: diff > 0: Ho is accepted at 1% level of significance

```
. ttest pcrate, by(deve)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      64    90.85485    1.956955    15.65564    86.94419    94.76551
       1 |      46    99.10757     .848447    5.754448    97.39871    100.8164
---------+--------------------------------------------------------------------
combined |     110    94.30599     1.25066    13.11703    91.82722    96.78476
---------+--------------------------------------------------------------------
    diff |           -8.252725    2.420257               -13.05009   -3.455356
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -3.4099
Ho: diff = 0                                     degrees of freedom =      108

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0005         Pr(|T| > |t|) = 0.0009          Pr(T > t) = 0.9995
```

Here, primary completion rate (*pcrate*) varies between -13.05 and -3.455

Standard error for the difference in primary completion rate is 2.42

The null hypothesis is given by,

Ho: diff = 0
From the t-test, we find that for,

a) Ha: diff < 0: Ho is rejected at 1% level of significance
b) Ha: diff != 0: Ho is rejected at 1% level of significance
c) Ha: diff > 0: Ho is accepted at 1% level of significance

```
. ttest hexp, by(deve)

Two-sample t test with equal variances

------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |     119    6.370955    .2724702    2.972299     5.83139    6.910521
       1 |      56    8.610337    .3858615    2.887523    7.837053    9.383621
---------+--------------------------------------------------------------------
combined |     175    7.087557     .235727    3.118375    6.622305     7.55281
---------+--------------------------------------------------------------------
    diff |           -2.239382    .4773391               -3.18154   -1.297223
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -4.6914
Ho: diff = 0                                     degrees of freedom =      173

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

Here, health expenditure (% of GDP) (*hexp*) varies between -3.181 and -1.297

Standard error for the difference in health expenditure (% of GDP) is very less (=0.477)

The null hypothesis is given by,

Ho: diff = 0
From the t-test, we find that for,

a) Ha: diff < 0: Ho is rejected at 1% level of significance
b) Ha: diff != 0: Ho is rejected at 1% level of significance
c) Ha: diff > 0: Ho is accepted at 10% level of significance

```
. ttest pun, by(deve)

Two-sample t test with equal variances
```

| Group | Obs | Mean | Std. Err. | Std. Dev. | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| 0 | 100 | 13.061 | 1.132858 | 11.32858 | 10.81316 | 15.30884 |
| 1 | 53 | 3.035849 | .1945762 | 1.416536 | 2.645403 | 3.426295 |
| combined | 153 | 9.588235 | .8369678 | 10.35272 | 7.934643 | 11.24183 |
| diff | | 10.02515 | 1.564907 | | 6.93321 | 13.11709 |

```
    diff = mean(0) - mean(1)                                      t =   6.4062
Ho: diff = 0                                     degrees of freedom =      151

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

Here, percentage of undernourished population *(pun)* varies between 6.933 and 13.11

Standard error for the difference in percentage of undernourished population is 1.564

The null hypothesis is given by,

Ho: diff = 0
From the t-test, we find that for,

a) Ha: diff < 0: Ho is accepted at 1% level of significance
b) Ha: diff != 0: Ho is rejected at 1% level of significance
c) Ha: diff > 0: Ho is rejected at 1% level of significance

```
. ttest gdppc, by(deve)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |     124    8843.649    530.7116    5909.755    7793.138     9894.16
       1 |      63    46908.57    2505.692    19888.31    41899.76    51917.38
---------+--------------------------------------------------------------------
combined |     187    21667.66    1602.762    21917.44    18505.73    24829.59
---------+--------------------------------------------------------------------
    diff |           -38064.92    1931.069               -41874.67   -34255.17
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t = -19.7118
Ho: diff = 0                                     degrees of freedom =      185

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

Here, GDP per capita (*gdppc*) varies between -41874.67 and -34255.17

Standard error for the difference in GDP per capita is very high (=1931.069)

The null hypothesis is given by,

Ho: diff = 0
From the t-test, we find that for,

a) Ha: diff < 0: Ho is rejected at 1% level of significance
b) Ha: diff != 0: Ho is rejected at 1% level of significance
c) Ha: diff > 0: Ho is accepted at 1% level of significance

<u>Answer 4:</u>

```
. *QUESTION 4:
. *Taking log of GDP per capita
. gen lgdppc = ln(gdppc)
(21 missing values generated)

. *creating a dummy variable from stat score
. gen ss = 1 if stat <= 25
(207 missing values generated)

. replace ss = 2 if stat > 25 & stat <= 50
(33 real changes made)

. replace ss = 3 if stat > 50 & stat <= 75
(70 real changes made)

. replace ss = 4 if stat > 75 & stat <= 100
(38 real changes made)

. replace ss = . if (missing(stat))
(0 real changes made)
```

Here, we took log of *gdppc* and stored it in *lgdppc*. Next, we are generating s new dummy variable from stat score named *'ss'* which stores the values '1', '2', '3', '4' and '.'under the following conditions given.

Answer 5:

```
. *QUESTION 5:
. *running a regression on lexp and all other explanatory variables including
. ///the dummy variable
> reg lexp co3 pcrate pun hexp lgdppc stat i.ss

      Source |       SS           df       MS      Number of obs   =        60
-------------+----------------------------------   F(8, 51)        =     16.90
       Model |  1327.55074         8  165.943843   Prob > F        =    0.0000
    Residual |  500.776904        51  9.81915498   R-squared       =    0.7261
-------------+----------------------------------   Adj R-squared   =    0.6831
       Total |  1828.32765        59  30.9886042   Root MSE        =    3.1336

------------------------------------------------------------------------------
        lexp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         co3 |  -.3639859   .2703337    -1.35   0.184    -.9067036    .1787317
      pcrate |   .0627868   .0395569     1.59   0.119    -.0166271    .1422007
         pun |  -.0655449   .0744067    -0.88   0.383    -.2149226    .0838328
        hexp |   .2051085   .1828109     1.12   0.267    -.1618997    .5721168
      lgdppc |   4.121248   .9174177     4.49   0.000     2.279455    5.963041
        stat |   .0064151   .0785306     0.08   0.935    -.1512417    .1640719
             |
          ss |
          3  |  -2.546789        2.1    -1.21   0.231    -6.762716    1.669137
          4  |  -.6976435   3.460376    -0.20   0.841    -7.644638    6.249351
             |
       _cons |   28.46513   8.894389     3.20   0.002      10.6089    46.32137
------------------------------------------------------------------------------
```

Here, we are regressing life expectancy at birth (*lexp*) with *co3, pcrate, pun, hexp, lgdppc, stat*; which are the explanatory variables in our data along with the newly created dummy variable named *'ss'*. Here, while regressing, we considered *'i.ss'* for the dummy variable. This is because ***i.*** stands for the intercept dummy of the categorical variable *ss* in our data. All other variables are continuous in nature.

Now, from the regression, we find that,

- *co3, pcrate, pun, hexp, stat* are statistically insignificant at 5% level of significance
- *lgdppc* is statistically significant at 5% level of significance
- if *co3* increases by 1 kiloton (kt), *lexp* decreases by 0.363 years
- if *pcrate* increases by 1 unit, *lexp* increases by 0.062 years

- if *pun* increases by 1 member, *lexp* decreases by 0.065 years
- if *hexp* increases by 1 percent, *lexp* increases by 0.205 years
- if *lgdppc* increases by 1 US $, *lexp* increases by 4.12 years
- if *stat* increases by 1 score, *lexp* increases by 0.0064 years

For the dummy variable *ss*,

- compared to s = 1 and s = 2, if stat score lies between 50 and 75, *lexp* decreases by 2.54 years
- compared to s = 1 and s = 2, if stat score lies between 75 and 100, *lexp* decreases by 0.69 years

However, the dummy variable s is statistically insignificant at 5% level of significance.

The constant term of the regression is 28.46 which is statistically significant at 5% level of significance.

## Answer 6:

## Regression Diagnostics:

```
. predict lexp1
(option xb assumed; fitted values)
(148 missing values generated)

. predict res, residuals
(148 missing values generated)

.
end of do-file

. do "C:\Users\DILIP\AppData\Local\Temp\STD00000000.tmp"

. *checking for normality
. *(jarque-bera test)
. jb res
Jarque-Bera normality test:  5.176 Chi(2)  .0752
Jarque-Bera test for Ho: normality:

. histogram res, normal
(bin=7, start=-8.1841908, width=1.9835954)

.
end of do-file

. graph save Graph "D:\Stata MP 14.2\MD PROJECT\histogram 2020.gph"
(file D:\Stata MP 14.2\MD PROJECT\histogram 2020.gph saved)
```

Jarque-Bera test is a no (no parameter) test where the null hypothesis is presented as,

Ho: normality

The value of Jarque-Bera normality test is 5.176

Here, the value of Chi(2) = 0.0752 with 2 degrees of freedom

This is statistically insignificant at 5% level of significance and we reject the null hypothesis of normality.

```
. *checking for homoscedasticity
. *# graphical method
. rvfplot, yline(0)

.
end of do-file

. graph save Graph "D:\Stata MP 14.2\MD PROJECT\rvfplot 2020.gph"
(file D:\Stata MP 14.2\MD PROJECT\rvfplot 2020.gph saved)
```

We had plotted a histogram where we found, the histogram is very mean-centric (i.e. it is leptokurtic in nature).

We had also plotted a scatterplot of residuals versus fitted values where we found there is no such relation between the two axes. It is not depicting any particular shape (or a pattern).

Thus, there is no presence of heteroscedasticity in our data.

```
. *# formal test
. hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of lexp

         chi2(1)      =     0.62
         Prob > chi2  =   0.4293

. reg lexp co3 pcrate pun hexp lgdppc stat i.ss, robust

Linear regression                               Number of obs     =         60
                                                F(8, 51)          =      30.07
                                                Prob > F          =     0.0000
                                                R-squared         =     0.7261
                                                Root MSE          =     3.1336
```

| lexp | Coef. | Robust Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| co3 | -.3639859 | .1892717 | -1.92 | 0.060 | -.7439648 | .0159929 |
| pcrate | .0627868 | .0363616 | 1.73 | 0.090 | -.0102121 | .1357857 |
| pun | -.0655449 | .0769133 | -0.85 | 0.398 | -.2199549 | .0888651 |
| hexp | .2051085 | .1695959 | 1.21 | 0.232 | -.1353694 | .5455865 |
| lgdppc | 4.121248 | .7351396 | 5.61 | 0.000 | 2.645393 | 5.597102 |
| stat | .0064151 | .0943056 | 0.07 | 0.946 | -.1829113 | .1957415 |
| ss | | | | | | |
| 3 | -2.546789 | 2.266373 | -1.12 | 0.266 | -7.096723 | 2.003144 |
| 4 | -.6976435 | 3.832307 | -0.18 | 0.856 | -8.39132 | 6.996033 |
| _cons | 28.46513 | 7.376782 | 3.86 | 0.000 | 13.65563 | 43.27464 |

After performing the graphical method for detecting heteroscedasticity, we then performed the formal method for detecting heteroscedasticity.

Here, we found that for the Breusch-Pagan/ Cook-Weisberg test for heteroscedasticity, the null hypothesis is presented as,

Ho: Constant variance

The chi2(1) value is 0.62 which is less than the critical value (= 3.84), thus we accept the null hypothesis of constant variance. The chi2(1) value is statistically insignificant at 5% level of significance.

Thus, there is no presence of heteroscedasticity in our data.

```
. *checking for multicollinearity
. vif
```

| Variable | VIF | 1/VIF |
|---|---|---|
| co3 | 2.15 | 0.464895 |
| pcrate | 1.67 | 0.597545 |
| pun | 2.39 | 0.417640 |
| hexp | 1.17 | 0.851671 |
| lgdppc | 4.02 | 0.248666 |
| stat | 7.21 | 0.138754 |
| ss | | |
| 3 | 6.67 | 0.149937 |
| 4 | 18.11 | 0.055221 |
| Mean VIF | 5.43 | |

We can use vif command after the regression to check for multicollinearity. As a rule of thumb, a variable whose vif value is > 10 may merit further investigation.

Here, the vif value = 5.43 which is less than 10, thus there is no multicollinearity issue in our data and it is a full column rank matrix

```
. *descriptive statistics
. desc lexp - stat

              storage   display    value
variable name   type    format     label      variable label
------------------------------------------------------------------------
lexp            double  %14.2f                lexp
co3             double  %14.2f                co3
pcrate          double  %14.2f                pcrate
pun             double  %14.2f                pun
hexp            double  %14.2f                hexp
gdppc           double  %14.2f                gdppc
stat            double  %14.2f                stat

. sum lexp - stat

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        lexp |        206    72.34202    7.407253     52.777   85.49756
         co3 |        188    3.775203    4.620423   .0325848   31.72684
      pcrate |        114    94.53861    12.96707   51.19454   115.6307
         pun |        160    9.935625    10.81931        2.5       53.1
        hexp |        180    7.064762    3.123236   2.007863   21.53917
-------------+---------------------------------------------------------
       gdppc |        187    21667.66    21917.44   751.2009   120010.2
        stat |        142    63.10251     16.4038    22.2222   94.44447

. corr lexp - stat
(obs=60)

             |     lexp      co3   pcrate      pun     hexp    gdppc     stat
-------------+---------------------------------------------------------------
        lexp |   1.0000
         co3 |   0.5105   1.0000
      pcrate |   0.5966   0.4166   1.0000
         pun |  -0.6510  -0.5468  -0.5315   1.0000
        hexp |   0.3070   0.0487   0.2383  -0.1197   1.0000
       gdppc |   0.7326   0.7035   0.4039  -0.6133   0.2625   1.0000
        stat |   0.5020   0.4686   0.2505  -0.4473   0.2247   0.5798   1.0000
```

There is high correlation among the variables as the values are greater than or equal to 0.5

```
. *principle component analysis (PCA)
. factor lexp - stat, pcf
(obs=60)

Factor analysis/correlation                        Number of obs    =         60
    Method: principal-component factors            Retained factors =          2
    Rotation: (unrotated)                          Number of params =         13
```

| Factor | Eigenvalue | Difference | Proportion | Cumulative |
|---|---|---|---|---|
| Factor1 | 3.77218 | 2.76884 | 0.5389 | 0.5389 |
| Factor2 | 1.00334 | 0.19348 | 0.1433 | 0.6822 |
| Factor3 | 0.80986 | 0.32320 | 0.1157 | 0.7979 |
| Factor4 | 0.48665 | 0.07225 | 0.0695 | 0.8674 |
| Factor5 | 0.41441 | 0.06945 | 0.0592 | 0.9266 |
| Factor6 | 0.34496 | 0.17636 | 0.0493 | 0.9759 |
| Factor7 | 0.16860 | . | 0.0241 | 1.0000 |

```
    LR test: independent vs. saturated:  chi2(21) =  189.00 Prob>chi2 = 0.0000

Factor loadings (pattern matrix) and unique variances
```

| Variable | Factor1 | Factor2 | Uniqueness |
|---|---|---|---|
| lexp | 0.8657 | 0.1027 | 0.2400 |
| co3 | 0.7620 | -0.3576 | 0.2916 |
| pcrate | 0.6728 | 0.1375 | 0.5285 |
| pun | -0.8010 | 0.1736 | 0.3282 |
| hexp | 0.3399 | 0.9001 | 0.0743 |
| gdppc | 0.8730 | -0.0743 | 0.2324 |
| stat | 0.6857 | -0.0159 | 0.5295 |

g. There are two factors (retained factors = 2) which satisfies the kaiser criterion. The first 2 eigenvalues are 3.77 and 1.00.

h. We performed the factor analysis, where we considered rotation (unrotated) to understand the factor loadings.

i. We perform LR test: independent vs saturated (correlated to each other)

j. The null hypothesis is Ho: independence (i.e. no correlation) (or, zero correlation)

k. The value of chi2(21) = 189.00 and Prob>chi2 = 0.0000

l. The null hypothesis, Ho is rejected at 1% level of significance; thus, running factor analysis is a good idea.

Now,

7. *lexp* has as much as 86.57% commonness with factor 1
8. *co2* has as much as 76.20% commonness with factor 1
9. *pcrate* has as much as 67.28% commonness with factor 1
10. *hexp* has as much as 33.99% commonness with factor 1
11. *gdppc* has as much as 87.30% commonness with factor 1
12. *stat* has as much as 68.57% commonness with factor 1

```
. estat kmo

Kaiser-Meyer-Olkin measure of sampling adequacy
```

| Variable | kmo |
|---|---|
| lexp | 0.7967 |
| co3 | 0.7738 |
| pcrate | 0.7572 |
| pun | 0.8958 |
| hexp | 0.6630 |
| gdppc | 0.7593 |
| stat | 0.9112 |
| Overall | 0.8020 |

```
. factortest lexp - stat

Determinant of the correlation matrix
Det                  =     0.036


Bartlett test of sphericity

Chi-square           =           185.677
Degrees of freedom =                  21
p-value              =             0.000
H0: variables are not intercorrelated


Kaiser-Meyer-Olkin Measure of Sampling Adequacy
KMO                 =     0.802
```

Here, kmo value is 0.8020 which is close to 1, thus we have adequate data to run factor analysis.

For the Bartlett's test of Sphericity,

1. the null hypothesis is Ho: variables are not intercorrelated
2. the value of chi-square is 185.677 with 21 degrees of freedom and the p-value is 0.0000
3. we reject the null hypothesis that variables are not intercorrelated at 5% level of significance.
4. Therefore, factor analysis is a good idea.

```
. *scree plot
. screeplot

. screeplot, yline(1)

.
end of do-file

. graph save Graph "D:\Stata MP 14.2\MD PROJECT\screeplot 2020.gph"
(file D:\Stata MP 14.2\MD PROJECT\screeplot 2020.gph saved)
```

We had then plotted the screeplot which is a graphical method of detecting and dropping the eigenvalues which are less than 1 since these provide less information than is provided by a single variable.

```
. *rotation
. *# orthogonal rotation
. rotate, varimax

Factor analysis/correlation                        Number of obs    =         60
    Method: principal-component factors            Retained factors =          2
    Rotation: orthogonal varimax (Kaiser off)      Number of params =         13
```

| Factor | Variance | Difference | Proportion | Cumulative |
|---|---|---|---|---|
| Factor1 | 3.54968 | 2.32384 | 0.5071 | 0.5071 |
| Factor2 | 1.22584 | . | 0.1751 | 0.6822 |

LR test: independent vs. saturated:  chi2(21) =  189.00 Prob>chi2 = 0.0000

Rotated factor loadings (pattern matrix) and unique variances

| Variable | Factor1 | Factor2 | Uniqueness |
|---|---|---|---|
| lexp | 0.8011 | 0.3439 | 0.2400 |
| co3 | 0.8321 | -0.1269 | 0.2916 |
| pcrate | 0.6062 | 0.3226 | 0.5285 |
| pun | -0.8174 | -0.0606 | 0.3282 |
| hexp | 0.0708 | 0.9595 | 0.0743 |
| gdppc | 0.8582 | 0.1762 | 0.2324 |
| stat | 0.6621 | 0.1791 | 0.5295 |

Factor rotation matrix

| | Factor1 | Factor2 |
|---|---|---|
| Factor1 | 0.9590 | 0.2835 |
| Factor2 | -0.2835 | 0.9590 |

This is an orthogonal rotation.

Here, variance = 3.549 has been extracted by factor 1 and variance = 1.225 has been extracted by factor 2

The cumulative variance is 0.6822

However, all the variables are loaded on both the factors

```
. rotate, varimax blanks(.49)

Factor analysis/correlation                        Number of obs    =         60
    Method: principal-component factors            Retained factors =          2
    Rotation: orthogonal varimax (Kaiser off)      Number of params =         13
```

| Factor | Variance | Difference | Proportion | Cumulative |
|---|---|---|---|---|
| Factor1 | 3.54968 | 2.32384 | 0.5071 | 0.5071 |
| Factor2 | 1.22584 | . | 0.1751 | 0.6822 |

```
LR test: independent vs. saturated:  chi2(21) =  189.00 Prob>chi2 = 0.0000
```

Rotated factor loadings (pattern matrix) and unique variances

| Variable | Factor1 | Factor2 | Uniqueness |
|---|---|---|---|
| lexp | 0.8011 | | 0.2400 |
| co3 | 0.8321 | | 0.2916 |
| pcrate | 0.6062 | | 0.5285 |
| pun | -0.8174 | | 0.3282 |
| hexp | | 0.9595 | 0.0743 |
| gdppc | 0.8582 | | 0.2324 |
| stat | 0.6621 | | 0.5295 |

(blanks represent abs(loading)<.49)

Factor rotation matrix

| | Factor1 | Factor2 |
|---|---|---|
| Factor1 | 0.9590 | 0.2835 |
| Factor2 | -0.2835 | 0.9590 |

We had considered 0.49 as the rule of thumb. Thus, values which are less than 0.49 are represented by blanks in the pattern matrix.

Therefore, *lexp* to *pun* and *gdppc* and *stat* are now loaded on factor 1

But, *hexp* is now loaded on factor 2

```
. *saving factor score
. predict pc1 pc2
(regression scoring assumed)

Scoring coefficients (method = regression; based on varimax rotated factors)
```

| Variable | Factor1 | Factor2 |
|---|---|---|
| lexp | 0.19108 | 0.16319 |
| co3 | 0.29474 | -0.28450 |
| pcrate | 0.13219 | 0.18198 |
| pun | -0.25269 | 0.10572 |
| hexp | -0.16789 | 0.88582 |
| gdppc | 0.24293 | -0.00543 |
| stat | 0.17883 | 0.03629 |

```
. estat common

Correlation matrix of the varimax rotated common factors
```

| Factors | Factor1 | Factor2 |
|---|---|---|
| Factor1 | 1 | |
| Factor2 | 0 | 1 |

```
. corr pc1 pc2
(obs=60)
```

| | pc1 | pc2 |
|---|---|---|
| pc1 | 1.0000 | |
| pc2 | -0.0000 | 1.0000 |

We find that, there is no correlation in orthogonal rotation.

```
. *# oblique rotation
. rotate, promax

Factor analysis/correlation                        Number of obs    =         60
    Method: principal-component factors            Retained factors =          2
    Rotation: oblique promax (Kaiser off)           Number of params =         13
```

| Factor | Variance | Proportion | Rotated factors are correlated |
|---|---|---|---|
| Factor1 | 3.73232 | 0.5332 | |
| Factor2 | 1.47741 | 0.2111 | |

```
    LR test: independent vs. saturated:  chi2(21) =  189.00 Prob>chi2 = 0.0000
```

Rotated factor loadings (pattern matrix) and unique variances

| Variable | Factor1 | Factor2 | Uniqueness |
|---|---|---|---|
| lexp | 0.7820 | 0.2159 | 0.2400 |
| co3 | 0.8827 | -0.2764 | 0.2916 |
| pcrate | 0.5827 | 0.2278 | 0.5285 |
| pun | -0.8402 | 0.0799 | 0.3282 |
| hexp | -0.0661 | 0.9800 | 0.0743 |
| gdppc | 0.8658 | 0.0325 | 0.2324 |
| stat | 0.6617 | 0.0697 | 0.5295 |

Factor rotation matrix

| | Factor1 | Factor2 |
|---|---|---|
| Factor1 | 0.9928 | 0.4138 |
| Factor2 | -0.1200 | 0.9104 |

This is oblique rotation.

Here, variance = 3.732 has been extracted by factor 1 and variance = 1.477 has been extracted by factor 2

There is correlation in oblique rotation as the rotated factors are correlated.

However, all the variables are loaded on both the factors

```
. rotate, promax blanks(.44)

Factor analysis/correlation                        Number of obs    =         60
    Method: principal-component factors            Retained factors =          2
    Rotation: oblique promax (Kaiser off)          Number of params =         13
```

| Factor | Variance | Proportion | Rotated factors are correlated |
|---|---|---|---|
| Factor1 | 3.73232 | 0.5332 | |
| Factor2 | 1.47741 | 0.2111 | |

```
    LR test: independent vs. saturated:  chi2(21) =  189.00 Prob>chi2 = 0.0000

Rotated factor loadings (pattern matrix) and unique variances
```

| Variable | Factor1 | Factor2 | Uniqueness |
|---|---|---|---|
| lexp | 0.7820 | | 0.2400 |
| co3 | 0.8827 | | 0.2916 |
| pcrate | 0.5827 | | 0.5285 |
| pun | -0.8402 | | 0.3282 |
| hexp | | 0.9800 | 0.0743 |
| gdppc | 0.8658 | | 0.2324 |
| stat | 0.6617 | | 0.5295 |

```
    (blanks represent abs(loading)<.44)

Factor rotation matrix
```

| | Factor1 | Factor2 |
|---|---|---|
| Factor1 | 0.9928 | 0.4138 |
| Factor2 | -0.1200 | 0.9104 |

We had considered 0.44 as the rule of thumb. Thus, values which are less than 0.44 are represented by blanks in the pattern matrix.

Therefore, *lexp* to *pun* and *gdppc* and *stat* are now loaded on factor 1

But, *hexp* is now loaded on factor 2

```
. *saving factor score
. predict pc3 pc4
(regression scoring assumed)

Scoring coefficients (method = regression; based on promax(3) rotated factors)

    ----------------------------------
        Variable |  Factor1   Factor2
    -------------+--------------------
            lexp |  0.21556   0.18812
             co3 |  0.24330  -0.24086
          pcrate |  0.16062   0.19856
             pun | -0.23158   0.06964
            hexp | -0.01818   0.85396
           gdppc |  0.23864   0.02832
            stat |  0.18238   0.06075
    ----------------------------------


. estat common

Correlation matrix of the promax(3) rotated common factors

    ----------------------------------
         Factors |  Factor1   Factor2
    -------------+--------------------
         Factor1 |        1
         Factor2 |    .3016         1
    ----------------------------------

. corr pc3 pc4
(obs=60)

             |      pc3      pc4
-------------+------------------
         pc3 |   1.0000
         pc4 |   0.3016   1.0000
```

There is some amount of correlation in oblique rotation, however no presence of multicollinearity in the data.

```
. *model specification tests
. linktest
```

| Source | SS | df | MS |
|---|---|---|---|
| Model | 1247.12113 | 2 | 623.560566 |
| Residual | 581.206515 | 57 | 10.1966055 |
| Total | 1828.32765 | 59 | 30.9886042 |

| | | |
|---|---|---|
| Number of obs | = | 60 |
| F(2, 57) | = | 61.15 |
| Prob > F | = | 0.0000 |
| R-squared | = | 0.6821 |
| Adj R-squared | = | 0.6710 |
| Root MSE | = | 3.1932 |

| lexp | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| _hat | 1.262007 | 2.567952 | 0.49 | 0.625 | -3.880227 | 6.404242 |
| _hatsq | -.0018839 | .0184524 | -0.10 | 0.919 | -.0388342 | .0350665 |
| _cons | -9.068041 | 89.05248 | -0.10 | 0.919 | -187.3926 | 169.2565 |

Here, we have performed model misspecification tests. For the linktest, we get,

a. *lexp* has been regressed on *hat* and *hatsq*
b. here, *hat* is significant as it should be at 1% level of significance and *hatsq* in insignificant at 1% level of significance.

Thus, model is highly misspecified.

```
. ovtest

Ramsey RESET test using powers of the fitted values of lexp
       Ho:  model has no omitted variables
                 F(3, 50) =      0.53
                 Prob > F =      0.6610
```

For the ovtest, we get,

a. the null hypothesis is given by, Ho: model has no omitted variables
b. F(3,62) = 0.6610 < 2.68 (the critical value of F(3,62) at 5% level of significance); thus we reject the null hypothesis that the model has no omitted variables at 5% level of significance.
c. Prob > F = 0.53 > 0.05; thus it is insignificant at 5% level of significance.
d. Thus, the model has a lot of explanatory variables which are omitted variables.

Answer 7:

If we compare the data given for the years 2019 and 2020, we see that,

- *pcrate* and *gdppc* are statistically significant at 5% level of significance for both the years.
- *lexp*, *co2*, *pun*, *hexp*, *stat* are statistically insignificant at 5% level of significance for 2019.
- *lexp*, *co3*, *pun*, *hexp*, *stat* are statistically insignificant at 5% level of significance for 2020.
- There is significant presence of heteroscedasticity in the data for 2019, but no presence of heteroscedasticity in the data for 2020.
- There is no presence of multicollinearity in the data for both the years.
- There is significant presence of omitted variable bias in the data for both the years.