Econometric Analysis Lab-1(HS49002)
Group Project

Factors Affecting Corruption in Developing and Emerging Countries

Aayush Joshi, 18HS200**01**

Dipanshu, 18HS200**14**

Hritik Mehta, 18HS200**18**

Saket Mahajan, 18HS200**28**

1. ABSTRACT

A multiple regression model with social and economic parameters/ indices of a country can tell significantly about its corruption. We attempt to do so by taking in consideration 123 countries (developed, developing and underdeveloped) and creating cross-sectional datasets of the same and performing econometrics tests.

2. INTRODUCTION

Our research aims to construct a model able to **explain the level of corruption**, dependent variable, in developing and emerging countries given a set of economic parameters as the independent variables. Economic and social indices of a country are becoming increasingly important in explaining a developing and emerging country's level of corruption.

At first, a simple-regression model is developed in order to show the effect of GDP per capita on the level of corruption. To get a better picture, we further build a multiple-regression analysis model where GDP per capita, HDI, and SPI are used as independent variables in order to predict the level of corruption.

GDP per capita and corruption have always been thought to be correlated to each other, it's masked sometimes to prevent social unrest or decreased foreign direct investment. We all know that corruption prevents an efficient allocation of resources, slowing down the growth of countries. So, we further choose to add essential growth variables such as human development index (HDI), and the social progress index (SPI). Typically, the more of each of the components within the three measures, the greater quality of life resulting in a greater awareness of the government's practices by the public.

In the past it's seen how corruption affects a nation's economy, our study decides to analyze the opposite relation; could corruption be explained by a nation's economy? We try to answer this question empirically.

3. DATA

The purpose of this paper is trying to explain corruption in countries by a set of economic parameters and measures of development; therefore the **measure of corruption** will be the dependent variable and following economic/ social indices are independent variables.

Data Sources (Hyperlinks)

- World CPI : Control of Corruption from Worldwide Governance Indicators (WGI)
- GDP per Capita 2019 (by -World Bank) : The World Bank was used as a source to determine each country's GDP per capita
- SPI : Gathered from The Social Progress Imperative Organization webpage.
- Unemployment : Gathered from Trading Economics
- Gini index : Gathered from the CIA World Factbook
- Government Type : Wikipedia
- Personal Income Tax Rate : Gathered from Trading Economics
- HDI : We used the Developing Programme of the United Nations' annual HDI (Human Development Index) data

3.1 Variable description

Variable	Description	Expected Relation
Control of Corruption (cpi)	The index ranges from -2.5(weak, very corrupt) to 2.5 (strong, transparent government).	Target Variable
GDP per Capita (ppp)	Gross Domestic Product per capita, measured in US dollars, recorded by the World Bank.	Positive
Human development Index (hdi, hdigrwoth)	Measured by life expectancy at birth, education index, and GNI per capita.	Positive
SPI (spi)	Composed of 3 dimensions: Basic Human Needs, Foundations of Wellbeing, and Opportunity	Positive
Unemployment (unemployment)	Levels of unemployment	Negative
Gini index (gini)	The Gini index (ranging from 0, being income distributed with perfect equality, to 100, being income distributed with perfect inequality) measures the degree of inequality in the distribution of family income in a country.	Negative
Government Type (head of state, constitutional form)	Three dummy variables were created for the four most common government types for these countries. The base model is the one where the country has an Absolute Monarchy (It’s 1, rest are 0).	NA
Personal Income Tax Rate (tax)	This variable represents the percent of income taxed.	Negative

3.2 Data Description

```
. summarize cpi spi ppp unempolyment  gini hdi hdigrowth tax, separator(10)
```

Variable	Obs	Mean	Std. Dev.	Min	Max
cpi	123	.0236585	1.016172	-1.42	2.17
spi	123	70.88577	14.89185	31.29	92.73
ppp	123	24675.9	23661.71	988	121293
unempolyment	123	9.153496	7.20796	.1	33.89
gini	123	37.66098	7.737482	24.2	63
hdi	123	.7480976	.1499248	.397	.957
hdigrowth	123	.0066195	.0038023	.0003	.019
tax	123	30.03756	13.16779	0	57.2

.

```
. tabulate headofstate , generate( headofstate )
```

Head of state	Freq.	Percent	Cum.
Ceremonial	50	40.65	40.65
Executive	73	59.35	100.00
Total	123	100.00	

```
. tabulate constitutionalform , generate( constitutionalform )
```

Constitutional form	Freq.	Percent	Cum.
Absolute monarchy	2	1.63	1.63
Constitutional monarchy	21	17.07	18.70
Provisional	1	0.81	19.51
Republic	99	80.49	100.00
Total	123	100.00	

4. METHODOLOGY

We have constructed **a model to explain the level of corruption**, dependent variable, in developing and emerging countries **given a set of economic parameters as the independent variables**: country's gross domestic product (GDP) per capita, human development index (HDI), the social progress index (SPI), the type of government in the country, included as a dummy variable, the level of unemployment in the country, the income tax rate collected in the country, and the Gini index which measures the degree of inequality in the distribution of family income in a country

After checking the descriptive statistics of all the variables and standardizing them; we have **built a multiple-regression analysis model using OLS** to analyse the level of corruption with mentioned independent variables. Then we checked the significance of the model and all the variables. From there **we picked up significant variables** and we checked for **severity of multicollinearity** and **heteroscedasticity**. Later robustness tests (**restricted –F tests**) are conducted on the obtained models to check whether the independent variables are jointly significant.

4.1 Initial Model

In our initial model, without manipulating the data, both variables and data were used as it is.

Model 1.1: $cpi = \beta_0 + \beta_1(spi) + \beta_2(ppp) + \beta_3(hdi) + \beta_4(unempolyment) + \beta_5(gini) + \beta_6(hdigrowth) + \beta_7(tax) + \beta_8(constitutionalform2) + \beta_9(constitutionalform3) + \beta_{10}(constitutionalform4) + \beta_{11}(headofstate2)$

Our first dummy variable set, head of state has two values where ceremonial is headofstate1 and executive is headofstate 2. And the second dummy variable set, which has four values Absolute monarchy, Constitutional monarchy, provisional, republic which are constitutionalform1, 2, 3 and 4 respectively.

note: headofstate1 omitted because of collinearity

Source	SS	df	MS
Model	105.53932	11	9.5944836
Residual	20.4385327	111	.184130925
Total	125.977852	122	1.03260535

Number of obs =	123
F(11, 111) =	52.11
Prob > F =	0.0000
R-squared =	0.8378
Adj R-squared =	0.8217
Root MSE =	.4291

cpi	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
spi	.0582794	.0110344	5.28	0.000	.036414	.0801448
ppp	.0000241	3.10e-06	7.75	0.000	.0000179	.0000302
unempolyment	.0070352	.0061262	1.15	0.253	-.0051043	.0191747
gini	.0011342	.0062125	0.18	0.855	-.0111762	.0134447
hdi	-3.414467	1.123858	-3.04	0.003	-5.641467	-1.187466
hdigrowth	20.98907	15.64057	1.34	0.182	-10.00376	51.98191
tax	.0092395	.0036173	2.55	0.012	.0020716	.0164074
headofstate2	-.0283816	.0999802	-0.28	0.777	-.2264991	.169736
headofstate1	0	(omitted)				
constitutionalform2	-.0393889	.3868375	-0.10	0.919	-.8059333	.7271555
constitutionalform3	-.682962	.5796401	-1.18	0.241	-1.831558	.4656336
constitutionalform4	-.3766897	.3729671	-1.01	0.315	-1.115749	.3623695
_cons	-2.338131	.601402	-3.89	0.000	-3.529849	-1.146413

At first, we get a significant model with decent R value, the coefficients of variables spi, ppp, hdi, tax turn out to be significant at 5% level. But, all the dummy variables and unemployment, gini, hdigrowth are statistically insignificant.

Model 1.2: $cpi = \beta_0 + \beta_{01}(spi) + \beta_{02}(ppp) + \beta_{03}(hdi) + \beta_{04}(unempolyment) + \beta_{05}(gini) + \beta_{06}(hdigrowth) + \beta_{07}(tax)$

```
. regress cpi spi ppp unempolyment gini hdi hdigrowth tax

      Source |       SS           df       MS      Number of obs   =       123
-------------+----------------------------------   F(  7,    115)  =     75.83
       Model |   103.54472         7  14.7921029   Prob > F        =    0.0000
    Residual |  22.4331323       115  .195070715   R-squared       =    0.8219
-------------+----------------------------------   Adj R-squared   =    0.8111
       Total |  125.977852       122  1.03260535   Root MSE        =    .44167

------------------------------------------------------------------------------
         cpi |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         spi |   .0547275   .0108923     5.02   0.000     .0331519    .0763031
         ppp |   .0000264   2.94e-06     8.98   0.000     .0000206    .0000323
unempolyment |   .0069426   .0062485     1.11   0.269    -.0054345    .0193198
        gini |   .0009568   .0059331     0.16   0.872    -.0107956    .0127092
         hdi |  -3.156112   1.128131    -2.80   0.006    -5.390723   -.9215011
   hdigrowth |   18.62057   15.93686     1.17   0.245    -12.94729    50.18842
         tax |   .0109551   .0034893     3.14   0.002     .0040434    .0178668
       _cons |  -2.698799   .5038204    -5.36   0.000    -3.696771   -1.700828
------------------------------------------------------------------------------
```

After removing all the dummy variables, we again regress the model and find it to be **significant** and a bit of reduced but decent value of R-square value. The reduction in R-Square may be because of dummy variables multicollinearity.

We now check for the problem of multicollinearity and heteroscedasticity in the above model.

```
. vif

    Variable |       VIF       1/VIF
-------------+----------------------
         hdi |     17.89    0.055894
         spi |     16.46    0.060771
         ppp |      3.03    0.329497
   hdigrowth |      2.30    0.435450
         tax |      1.32    0.757399
        gini |      1.32    0.758692
unempolyment |      1.27    0.788227
-------------+----------------------
    Mean VIF |      6.23
```

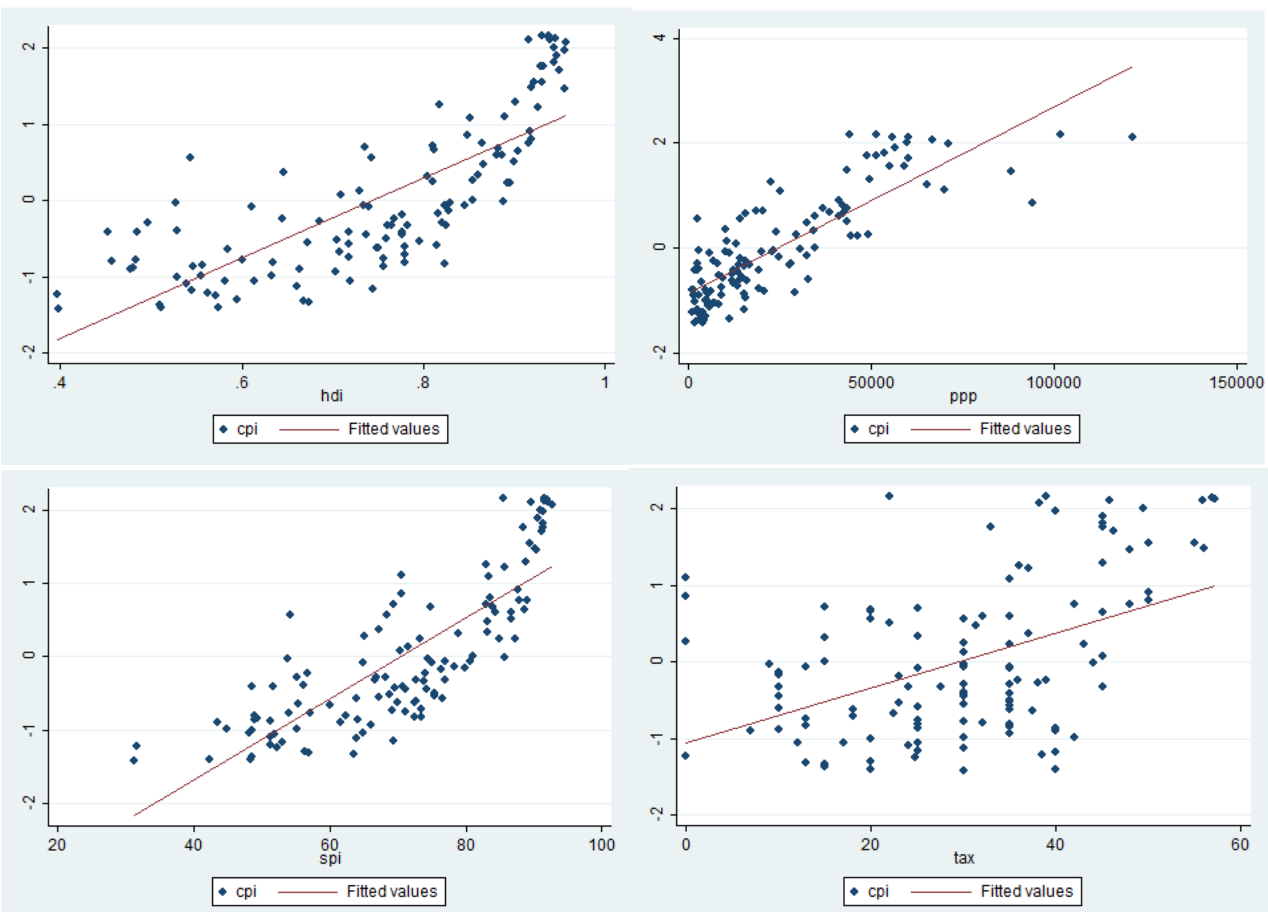
```
. estat hettest

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

         chi2(1)      =     5.40
         Prob > chi2  =   0.0202
```

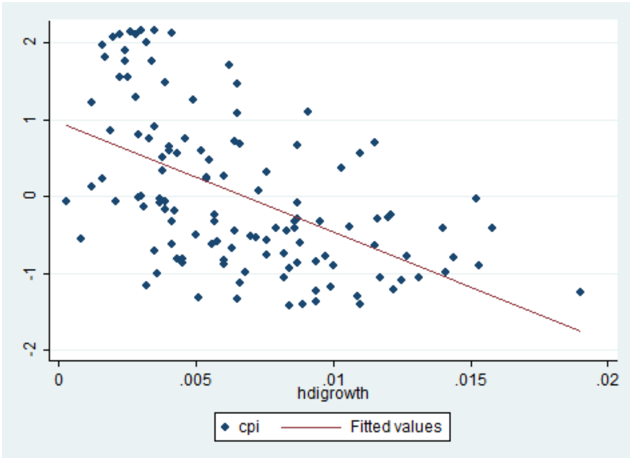
The model is exposed to multicollinearity and heteroscedasticity.

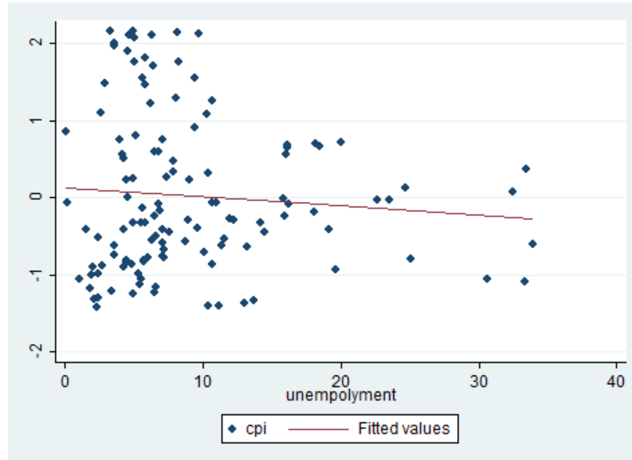
Before removing other insignificant independent variables we wanted to see their relationship with the other independent variables and the dependent variables. Hence we have obtained the following correlation matrix and scatter plots.

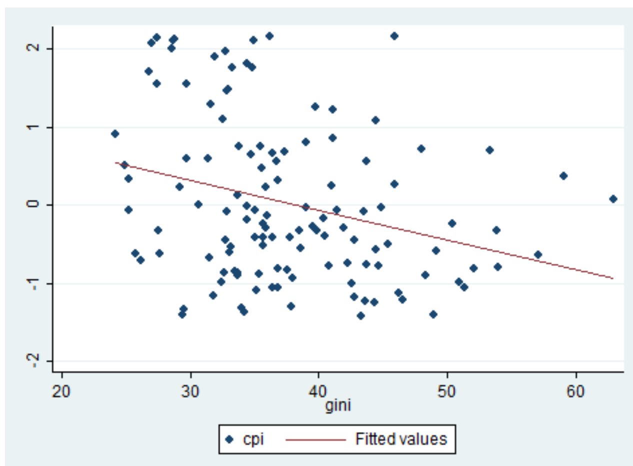
	cpi	spi	ppp	unempo~t	hdi	hdigro~h	gini
cpi	1.0000						
spi	0.8122*	1.0000					
ppp	0.8282*	0.7498*	1.0000				
unempolyment	-0.0833	-0.0919	-0.2389*	1.0000			
hdi	0.7778*	0.9591*	0.8039*	-0.1536	1.0000		
hdigrowth	-0.5348*	-0.7205*	-0.5716*	0.2430*	-0.7179*	1.0000	
gini	-0.2901*	-0.3873*	-0.3245*	0.3245*	-0.3998*	0.3762*	1.0000
tax	0.4680*	0.3724*	0.2794*	-0.0186	0.2818*	-0.1683	-0.1082

*significance at 5%




2
1
0
-1
-2
0
.005
.01
.015
.02
hdigrowth
cpi
Fitted values


2
1
0
-1
-2
0
10
20
30
40
unempolyment
cpi
Fitted values


2
1
0
-1
-2
20
30
40
50
60
gini
cpi
Fitted values

Model 1.3: $cpi = \beta_0 + \beta_{01}(spi) + \beta_{02}(ppp) + \beta_{03}(hdi) + \beta_{04}(tax)$

```
. regress cpi spi ppp hdi tax
```

Source	SS	df	MS
Model	102.8075	4	25.7018751
Residual	23.1703521	118	.196358916
Total	125.977852	122	1.03260535

Number of obs =	123
F(4, 118) =	130.89
Prob > F =	0.0000
R-squared =	0.8161
Adj R-squared =	0.8098
Root MSE =	.44312

cpi	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
spi	.0543684	.010428	5.21	0.000	.0337182	.0750186
ppp	.0000257	2.91e-06	8.85	0.000	.00002	.0000315
hdi	-3.454254	1.119393	-3.09	0.003	-5.670956	-1.237551
tax	.0113752	.0034672	3.28	0.001	.0045093	.0182411
_cons	-2.222953	.2964467	-7.50	0.000	-2.809998	-1.635908

Model is significant with a good R-square value. "Spi", "ppp", "hdi", "tax" and intercept have significant coefficients with "hdi" having negative relationship with "cpi" and intercept is also negative.

Variable	VIF	1/VIF
hdi	17.50	0.057145
spi	14.98	0.066741
ppp	2.94	0.340064
tax	1.30	0.772182
Mean VIF	9.18	

```
. estat hettest

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

         chi2(1)      =     4.05
         Prob > chi2  =   0.0441
```

Multicollinearity is present and Heteroscedasticity is also present at 5% **significance** level.

4.2 Corrective Measures

The last model i.e. **model 1.3 has multicollinearity and heteroscedasticity issues**. Possibility is there that "spi" and "hdi" variables are collinear because both measure development opportunities in a country, this can be also observed from the Mean VIF table. We have assumed our model to have statistically significant multicollinearity if our VIF exceeds above '5' or our Tolerance comes out to be less than '0.2'.

As corrective measures we start with **standardizing** all the independent variables i.e. dividing them by their standard deviation. **cpi_2, spi_2, hdi_2, tax_4** are standardized.

```
. regress cpi_2 spi_4 ppp_4 hdi_4 tax_4

      Source |       SS           df       MS      Number of obs   =       123
-------------+----------------------------------   F(4, 118)       =    130.89
       Model |  102.807501         4  25.7018754   Prob > F        =    0.0000
    Residual |  23.1703508       118  .196358905   R-squared       =    0.8161
-------------+----------------------------------   Adj R-squared   =    0.8098
       Total |  125.977852       122  1.03260535   Root MSE        =    .44312

------------------------------------------------------------------------------
       cpi_2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       spi_4 |   .8096461   .1552915     5.21   0.000     .5021266    1.117166
       ppp_4 |   .6089966   .0687963     8.85   0.000     .4727612    .7452319
       hdi_4 |  -.5178784   .1678247    -3.09   0.003     -.850217   -.1855397
       tax_4 |   .1497866   .0456547     3.28   0.001     .0593778    .2401953
       _cons |  -2.222953   .2964467    -7.50   0.000    -2.809998   -1.635908
------------------------------------------------------------------------------
```

```
. vif

    Variable |       VIF       1/VIF
-------------+----------------------
       hdi_4 |     17.50    0.057145
       spi_4 |     14.98    0.066741
       ppp_4 |      2.94    0.340064
       tax_4 |      1.30    0.772182
-------------+----------------------
    Mean VIF |      9.18
```

Again “Spi”, “ppp”, “hdi”, “tax” and intercept have significant coefficients with “hdi” having negative relationship with “cpi” and intercept in this model is also negative. Goodness is the same after standardization. Here we could resolve heteroscedasticity but multicollinearity couldn’t be resolved.

After trying different functional forms we came to the conclusion that the multicollinearity is due to “spi” and “hdi”. But removing “hdi” or “spi” makes the R-square value to drop so we move to do **restricted F-Test** for “hdi” and “spi”.

```
. test spi_4=hdi_4

 ( 1)  spi_4 - hdi_4 = 0

       F(  1,   118) =   17.63
            Prob > F =    0.0001
```

The F statistic is highly significant, which means we reject the hypothesis that the two effects are equal. For now we move forward without dropping any of the variables.

Now we move forward to **log transformation**

lnppp = Ln(ppp),

lnspi = Ln(spi),

lnhdi = Ln(hdi),

lntax = Ln(tax)

Model 1.4: $cpi = \beta_0 + \beta_1(\ln of\ spi) + \beta_2(\ln of\ ppp) + \beta_3(\ln of\ hdi) + \beta_4(\ln of\ tax)$

```
. . regress cpi_3 lnspi_3 lnppp_3 lnhdi_3 lntax_3

      Source |       SS           df       MS      Number of obs   =       119
-------------+----------------------------------   F(4, 114)       =     84.86
       Model |  91.6850145         4  22.9212536   Prob > F        =    0.0000
    Residual |  30.7906277       114  .270093225   R-squared       =    0.7486
-------------+----------------------------------   Adj R-squared   =    0.7398
       Total |  122.475642       118  1.03792917   Root MSE        =     .5197

------------------------------------------------------------------------------
       cpi_3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lnspi_3 |   4.538417   .7800237     5.82   0.000     2.993196    6.083638
     lnppp_3 |   .8405465   .1701302     4.94   0.000       .50352    1.177573
     lnhdi_3 |  -5.506323   1.254633    -4.39   0.000    -7.991741   -3.020906
     lntax_3 |   .5182982   .1080979     4.79   0.000     .3041571    .7324394
       _cons |  -30.72871   4.204241    -7.31   0.000    -39.05728   -22.40014
------------------------------------------------------------------------------

. vif

    Variable |       VIF       1/VIF
-------------+----------------------
     lnhdi_3 |     31.28    0.031966
     lnppp_3 |     15.57    0.064217
     lnspi_3 |     13.22    0.075657
     lntax_3 |      1.12    0.892354
-------------+----------------------
    Mean VIF |     15.30

. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of cpi_3

         chi2(1)      =     1.38
         Prob > chi2  =   0.2408

.
```

In log transformation we drop those countries which had **tax of 0%**

By log transformation we could tackle heteroscedasticity but the severity of multicollinearity has increased.

We finally decided to drop the "**hdi**" variable and perform our modelling and testing.

```
==============================================================================
* Farrar-Glauber Multicollinearity Tests
==============================================================================
  Ho: No Multicollinearity - Ha: Multicollinearity

* (1) Farrar-Glauber Multicollinearity Chi2-Test:
    Chi2 Test =   220.1657    P-Value > Chi2(3) 0.0000

* (2) Farrar-Glauber Multicollinearity F-Test:
```

Variable	F_Test	DF1	DF2	P_Value
lnspi_3	306.256	116.000	2.000	0.003
lnppp_3	309.075	116.000	2.000	0.003
lntax_3	3.460	116.000	2.000	0.250

```
* (3) Farrar-Glauber Multicollinearity t-Test:
```

Variable	lnsp~3	lnpp~3	lnta~3
lnspi_3	.		
lnppp_3	24.748	.	
lntax_3	2.442	2.628	.

Just for more strength of the analysis, we also conducted **Farror Glauber test of multicollinearity** in which we conducted chi square test, F- test and t-tests and found that there is severe multicollinearity by the chi square test, then conducted F-test for the location of multicollinearity, then t-test for the pattern of multicollinearity.

4.3 Final Model

$$\textbf{Model 2: } cpi = \beta_0 + \beta_1(lnspi) + \beta_2(ppp) + \beta_3(tax)$$

```
. regress cpi_2 ppp_4 lnspi tax_4
```

Source	SS	df	MS
Model	99.7000694	3	33.2333565
Residual	26.2777829	119	.220821705
Total	125.977852	122	1.03260535

Number of obs =	123
F(3, 119) =	150.50
Prob > F =	0.0000
R-squared =	0.7914
Adj R-squared =	0.7862
Root MSE =	.46992

cpi_2	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ppp_4	.5628419	.0601945	9.35	0.000	.4436508	.682033
lnspi	1.341857	.2654914	5.05	0.000	.8161578	1.867557
tax_4	.2159999	.0451832	4.78	0.000	.1265327	.3054671
_cons	-6.740734	1.067496	-6.31	0.000	-8.854483	-4.626986

As per the **Mean VIF** and **Breusch-Pagan test**, our current model is free of multicollinearity and heteroscedasticity

Variable	VIF	1/VIF
lnspi	2.07	0.482723
ppp_4	2.00	0.499538
tax_4	1.13	0.886601
Mean VIF	1.73	

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of cpi_2

         chi2(1)      =     0.72
         Prob > chi2  =   0.3974
```

For confirming that heteroscedasticity is not present in the model, we also verify it with **White's test**. Our model also passed it stating homoscedasticity presence.

```
. . estat imtest, white

White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(9)      =      6.71
         Prob > chi2  =    0.6671

Cameron & Trivedi's decomposition of IM-test
```

Source	chi2	df	p
Heteroskedasticity	6.71	9	0.6671
Skewness	7.98	3	0.0463
Kurtosis	0.00	1	0.9973
Total	14.70	13	0.3267

For confirming absence of multicollinearity, we perform **Farrar-Glauber Multicollinearity Tests**

```
==============================================================================
* Farrar-Glauber Multicollinearity Tests
==============================================================================
  Ho: No Multicollinearity - Ha: Multicollinearity

* (1) Farrar-Glauber Multicollinearity Chi2-Test:
    Chi2 Test =    97.2896    P-Value > Chi2(3) 0.0000

* (2) Farrar-Glauber Multicollinearity F-Test:
```

Variable	F_Test	DF1	DF2	P_Value
ppp_4	60.111	120.000	2.000	0.016
lnspi	64.295	120.000	2.000	0.015
tax_4	7.674	120.000	2.000	0.122

```
* (3) Farrar-Glauber Multicollinearity t-Test:
```

Variable	ppp_4	lnspi	tax_4
ppp_4	.		
lnspi	10.912	.	
tax_4	3.188	3.834	.

We find that multicollinearity is present but the reason behind seems to be the **sample-set**. On trying with **different samples**, multicollinearity was changing a lot.

4.4 Robustness Test

Restricted Model #1: cpi= ß0 + ß1lnSPI + u

The significant F-test, 58.52, means that the collective contribution of these two variables is significant. One way to think of this, is that there is a significant difference between a model with **ppp_4** and **tax_4** as compared to a model without them, i.e., there is a significant difference between the “full” model and the “reduced” models.

```
. test ppp_4 tax_4

 ( 1)  ppp_4 = 0
 ( 2)  tax_4 = 0

       F(  2,   119) =   58.52
            Prob > F =    0.0000
```

Restricted Model #2: cpi= ß0 + ß1ppp+ u

Since the F-Statistic is again greater than the critical value obtained previously (about 3.10 at a 5% confidence), the null hypothesis is rejected and it can be concluded that **ln(spi)** and **tax_4** are also jointly significant at a 5% confidence level.

```
. test lnspi tax_4

 ( 1)  lnspi = 0
 ( 2)  tax_4 = 0

       F(  2,   119) =   30.08
            Prob > F =    0.0000
```

5. CONCLUSION

We can conclude from our research that **ln(spi)** is positively correlated with the Control of Corruption index in the simple regression model. This is because as SPI increases, wellbeing and social progress of the nation increases implying that corruption is more controlled (approaching a score of 2.5 which is typically seen in a strong, transparent government). Similar relationship can be seen with **per Capita GDP** and **Personal Income Tax.**

We also saw that our final model didn't show symptoms of heteroscedasticity but the severity of multicollinearity of the model kept on changing when we changed the sample size. So we chose the significant model with the best goodness of fit and least multicollinearity. To tackle multicollinearity, we had to drop **"hdi"** variable after speculating its high collinearity with **"spi"**

An important caveat to note here is that even though we didn't choose the model with **"hdi", "unemployment" as** our final model, the effects of unemployment, human development and government type should not be ruled out when explaining corruption in developing and emerging countries. Different methodologies to conduct the measurement of each of these excluded variables could represent a greater impact on corruption than what is discovered.