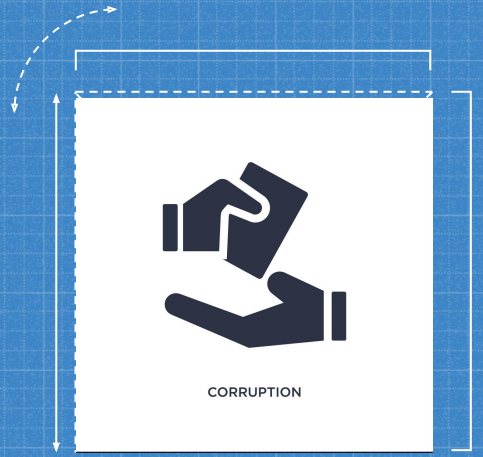
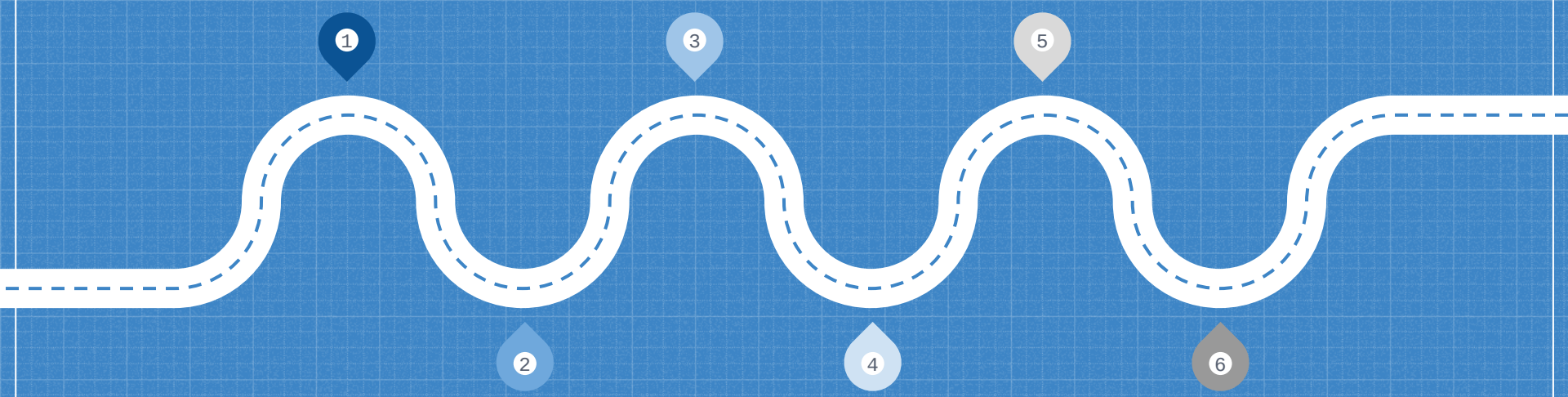
Econometric Analysis-1 (HS49002)

Factors Affecting Corruption in Developing and Emerging Countries

By-
Aayush Joshi, 18HS20001
Dipanshu, 18HS20014
Hritik Mehta, 18HS20018
Saket Mahajan, 18HS20028



ROADMAP


1
Introduction/ Motivation towards the topic
2
Data Sets
3
Initial Regression Modelling
4
Handling the Regression Problems
5
Final Model
6
Learning outcomes

1

INTRODUCTION

Underlying theory and motivation

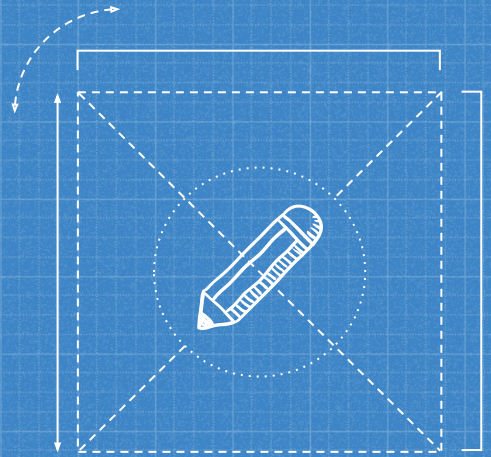
- Economic and social indices of a country are becoming increasingly important in explaining a developing and emerging country’s level of corruption.
- We aim to construct a model able to explain the level of corruption, dependent variable, in countries given a set of economic parameters as the independent variables.
- GDP per capita and corruption have always been thought to be correlated to each other, we further choose to add essential growth variables such as human development index (HDI), and the social progress index (SPI).
- 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.

- 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)
  - 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
  - The Gini index which measures the degree of inequality in the distribution of family income in a country

2

DATASET DESCRIPTION

Variable Description
And Descriptive Statistics



Variable Description

Control of Corruption (cpi)

According to the WGI, this index “reflects perception of the extent to which public power is exercised for private gain. The index ranges from -2.5(weak, very corrupt) to 2.5 (strong, transparent government).

Expected Relation: Target Variable

GDP per Capita (ppp)

Gross Domestic Product per capita from the year 2019, measured in US dollars, recorded by World Bank

Expected Relation: Positive

Human development Index (hdi, hdigrwoth)

Measured by life expectancy at birth, education index, and GNI per capita. Growth rate is calculated using two consecutive years’ HDI.

Expected Relation: Positive

Social Progress Index (SPI)

Composed of 3 dimensions: Basic Human Needs, Foundations of Wellbeing, and Opportunity.

Expected Relation: Target Variable

Unemployment (unemployment)

Levels of unemployment from the year 2019 gathered from Trading Economics given as a percentage for each country included in the study.

Expected Relation: 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.

Expected Relation: Negativ

Government Type (constitutional form)

The four main government types for the list of countries are Constitutional Monarchy, Republic, Provisional, Absolute Monarchy.

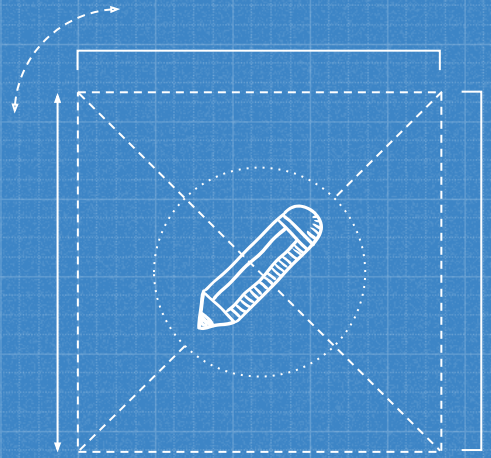
Government Type (head of state)

The two main types for the head of the state are ceremonial and executive. For eg. US have executive president while India have Ceremonial.

Personal Income Tax Rate (tax)

This variable represents the percent of income taxed in each of the developing countries.

Expected Relation: Positive



Dataset 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	

- Dummy variables were created for the most common government types and head of states for these countries.
- Our first dummy variable set, head of state has two values where ceremonial is headofstate1 and executive is headofstate 2.
- The second dummy variable set, which has four values Absolute monarchy, Constitutional monarchy, provisional, republic which are constitutionalform1, 2, 3 and 4 respectively.

3

REGRESSION MODELLING

Initial multiple Regression model
and statistical tests

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

```
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

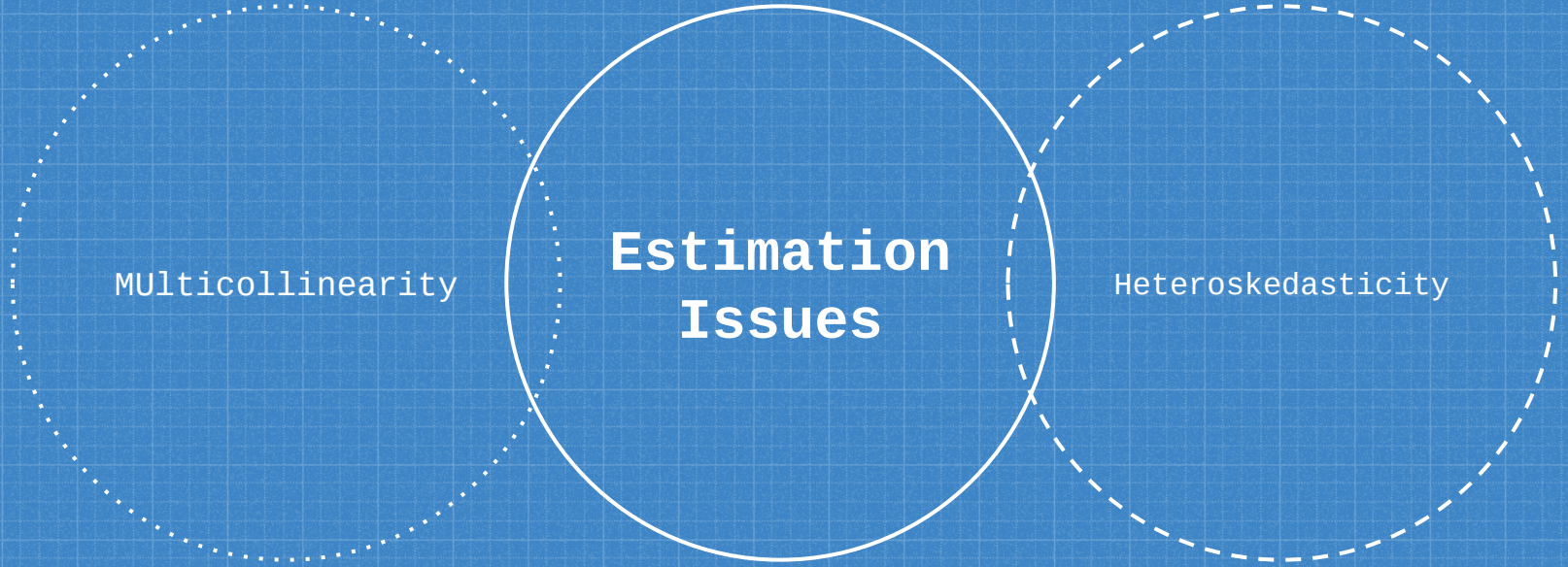
Model 1.2: $cpi = \beta_0 + \beta_{01}(spi) + \beta_{02}(ppp) + \beta_{03}(hdi) + \beta_{04}(unempolyment) + \beta_{05}(gini) + \beta_{06}(hdigrowth) + \beta_{07}(tax)$

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

Source	SS	df	MS
Model	103.54472	7	14.7921029
Residual	22.4331323	115	.195070715
Total	125.977852	122	1.03260535

Number of obs	= 123
F(7, 115)	= 75.83
Prob > F	= 0.0000
R-squared	= 0.8219
Adj R-squared	= 0.8111
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


MUlticollinearity
Estimation Issues
Heteroskedasticity

```
. 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 current model is exposed to multicollinearity and heteroscedasticity. We also conform these using other tests.

	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

HDI, SPI and Per Capita are collinear but we expect HDI-SPI collinearity is increasing the severity of multicollinearity.

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

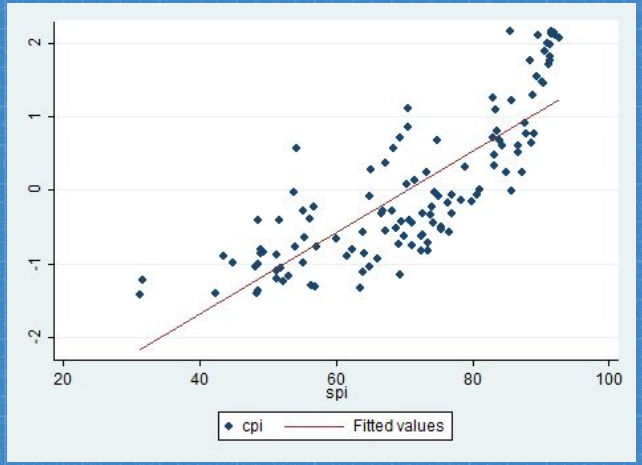
cpi	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
spi	.0543684	.010428	5.21	0.000	.0337182	.0750186
ppp	.0000257	2.91e-06	8.85	0.000	.00002	.0000315
hdi	-3.454254	1.119393	-3.09	0.003	-5.670956	-1.237551
tax	.0113752	.0034672	3.28	0.001	.0045093	.0182411
_cons	-2.222953	.2964467	-7.50	0.000	-2.809998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

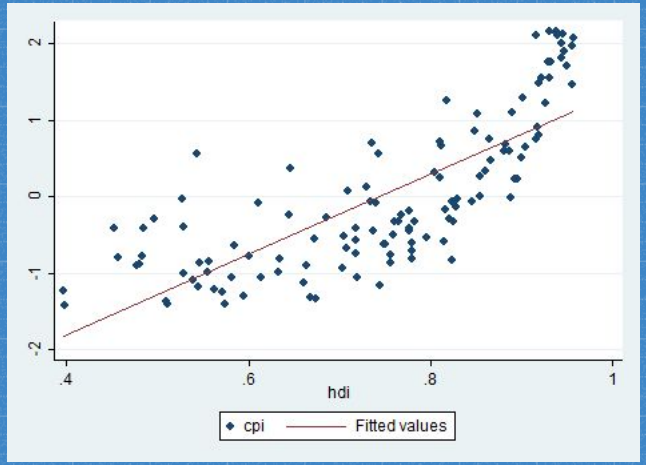
Variable	VIF	1/VIF
hdi	17.50	0.057145
spi	14.98	0.066741
ppp	2.94	0.340064
tax	1.30	0.772182
Mean VIF	9.18	

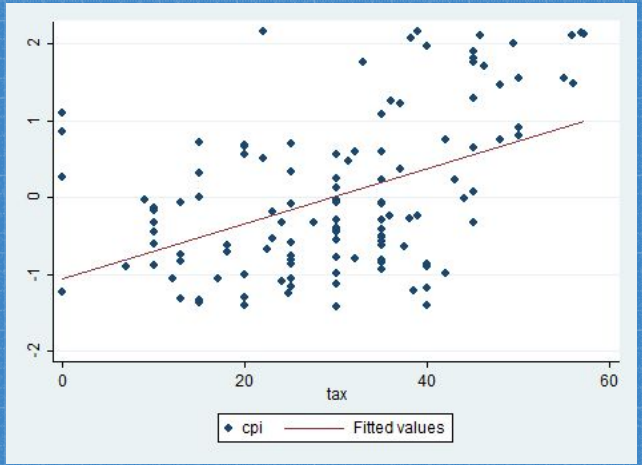
```
. estat hettest

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

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

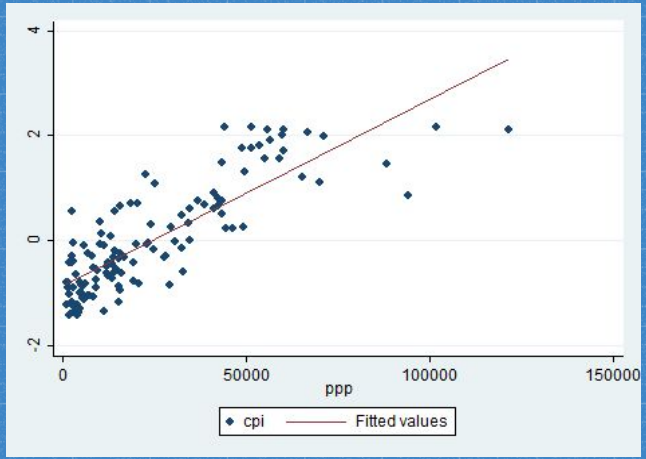
The model is still exposed to multicollinearity and heteroscedasticity.

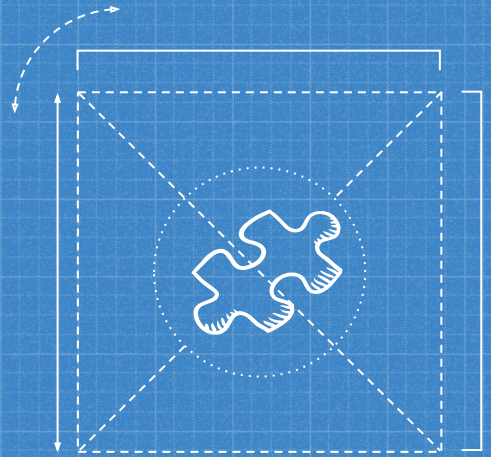
For now, we consider to drop either HDI or SPI


2
1
0
-1
-2
20
40
60
80
100
spi
cpi
Fitted values


2
1
0
-1
-2
.4
.6
.8
1
hdi
cpi
Fitted values


2
1
0
-1
-2
0
20
40
60
tax
cpi
Fitted values


4
2
0
-2
0
50000
100000
150000
ppp
cpi
Fitted values



Corrective Measures

```
. regress cpi_2 spi_4 ppp_4 hdi_4 tax_4
```

Source	SS	df	MS		
Model	102.807501	4	25.7018754	Number of obs =	123
Residual	23.1703508	118	.196358905	F(4, 118) =	130.89
				Prob > F =	0.0000
				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

Model is free from heteroscedasticity but multicollinearity problem persists

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		
Model	91.6850145	4	22.9212536	Number of obs =	119
Residual	30.7906277	114	.270093225	F(4, 114) =	84.86
				Prob > F =	0.0000
Total	122.475642	118	1.03792917	R-squared =	0.7486
				Adj R-squared =	0.7398
				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

Model is free from heteroscedasticity but multicollinearity problem persists and Mean VIF has increased.

Models not shown

- “HDI” was dropped because dropping “SPI” was decreasing Adj R-square.
- Different combinations of variables after ratio transformation, log-transformation and squaring was tried.
- Use of ln(ppp) or ln(tax) was decreasing the Adj R-square.
- Use of ln(ppp) and ln(spi) simultaneously was increasing Mean VIF.

Final Model

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

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
```

```
. . 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

As per Mean VIF Model is free from multicollinearity

Also Breusch-Pagan and White’s test suggest absence of significant heteroskedastic.

Restricted f-test

Restricted Model #1:

cpi= ß0 + ß1lnSPI + u

```
. 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

```
. test lnspi tax_4

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

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

Since the F-Statistic is greater than the critical value obtained, the collective contribution of these two variables is significant. there is a significant difference between the “full” model and the “reduced” models.

3

CONCLUSION

Findings and Caveats

Findings

- Our final model explains cpi using ln(spi), gdp per capita and personal income tax at 5% level of significance.
- ln(spi) is positively correlated with the Control of Corruption index in the simple regression model.
- Similar relationship can be seen with per Capita GDP and Personal Income Tax.
- In the past it’s seen how corruption affects a nation’s economy, our study found the opposite relationship is also true.

Learning Outcomes

1. Practical Application of using data from different sources and combining them for estimating a multiple regression model.
2. Application of statistical tests to check for different issues associated with estimating a regression model.
3. Severe Multicollinearity can be tackled using dropping some variables if log-transformation doesn’t work.
4. Heteroscedasticity can be tackled by standardising all the independent variables.

Limitations

- We have also used developed countries in our dataset which usually aren’t ideal for study on corruption.
- Only 123 countries were included in dataset out of cpi data available for 200+ countries.
- We didn’t choose the model with “hdi”, “unemployment” as our final model.

Thanks!

ANY QUESTIONS?