EMPIRICAL PROJECT REPORT

Aditi Kaushik Jayachandran (aj3235)

Swostik Pati (sp6441)

Kevin Chae (yc5076)

I. INTRODUCTION

Research Question

How does GDP per capita PPP ($ constant International 2017) change with the change in the Labor Force Participation Rate for the year 2016?

Hypothesis

We hypothesize that change in Labor Force Participation Rate will impact GDP per capita PPP of countries with positive linear correlation.

Rationale - Relevance of the Topic

Economics teaches us that when more people are employed in the labor force and are involved in production, more goods are produced, and the output of a nation increases.

Since GDP is a measure of output, it would only make sense for GDP to increase as more people are employed.

The variables GDP per capita PPP and total labor force participation rate would allow us to compare these two metrics across countries to see if economic theory applies and fits correctly in the real world.

All three of us are inclined towards business and economics as students, we found it particularly interesting.

II. DATA AND SAMPLE DESCRIPTION

Data Source – World Bank

A credible and reliable source as data is ‘‘gathered by the Bank's country management units and data obtained from official sources’’.

We use the data from the “World Development Indicator” series.

World Bank applies and follows ethical means of data collection.

Sample Used

While selecting countries, we picked 212 nations that we could use for our study.

Many countries had missing observations for either the IV or DV, so we used 175 nations as data points for our sample.

Our sample is as reliable and close to fair conditions for testing as possible, we randomly selected the countries and tried to be representative. Used countries from all continents in equal proportion.

Used countries with high and low GDP values, also mixed population sizes. The sample made full use of available information.

By using the maximum number of countries in the sample we reduced variability in the sample as much as possible.

We checked to see which year had the lowest number of missing values/observations - this was the year 2016. Hence, 2016 became our year of focus for the investigation.

III. Empirical Strategy

Research Methods:

We use the following research methods to conduct our study.

1. Preliminary Analysis - Summarizing the data, plotting histograms, understanding variables and outliers, etc.
2. Intermediate Analysis - Computing correlation and variance between variables, weighting, running regressions, plotting scatterplots and regression lines, etc.
3. Supplementary Analysis – Transforming variable into the logarithmic scale, performing Intermediate Analysis with the Logarithmic variable, plotting two-way graphs for comparative study, quartile division, etc.

Variables Used

We picked two indicators to compare in this investigation:

1. Independent variable (regressor): Labor force participation rate (LFPR).
   Expressed as a proportion of the population aged 15-64 that is 'economically active'.
2. Dependent variable- GDP per capita (purchasing power parity- PPP, USD International 2017).
   All adjusted to the same scale for the strength of each currency. Brings each value to a comparable level.
   Unit: international dollars converted by purchasing power parity (PPP) conversion factor

We also used a third parameter – population – while regressing the values to ensure that each country's IV and DV values were weighted by population. A high population country would therefore affect the data more than a less populated one.

IV. Results

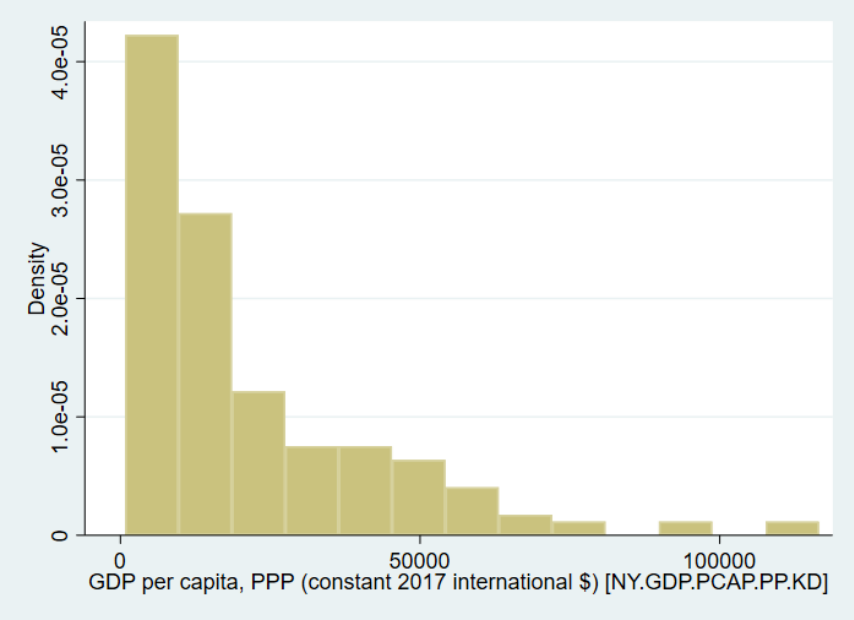
Summarizing Variables

Given below is the table summarizing the sample parameters – like mean, variance, etc. - of both the independent and dependent variables.

	IV: Labor force participation rate	DV: GDP per capita PPP
Mean	67.88	21018.08
Standard Deviation	10.78	21713.51
Variance	116.28	14040.31
Skewness	-0.56	1.72

Histogram and Outliers

We plotted the data points of GDP per capita on a histogram to observe the data. Most of the data countries lied in the region until 50000 USD. But we also observed some countries lying at the extreme end of the histogram. It is evident that any GDP value above 70,000 tended to be in the outlier zone. Even though these data points had the ability to drive our observation away from the desired one, but we didn't exclude them so as to study their effects as well.



Correlation

Correlation between LFPR and GDP per capita PPP = 0.3368.

There is a weak positive linear relationship between the variables. As labor force participation rate increases, GDP per capita PPP also increases.

Correlation may be weak since there tends to be a lot of outliers and this causes the countries' data points to be further away from the regression line. As correlation is not very strong, it is unlikely that we can establish causation between IV and DV.

When we compute the covariance we find it to be positive. This shows that both variables move in the same direction

```
. corr GDP LFPR
(obs=175)

             |      GDP      LFPR
-------------+------------------
         GDP |   1.0000
        LFPR |   0.3368   1.0000
```

Weighting by Population

While finding the regression for this investigation, we considered the fact that our sample contains nations with different sized populations. For example, nations like India/ China as well as nations like Belize/ Malta were part of the study.

Hence, we weighted IV and DV analytically using population. Weighting by population allowed the regression and other statistic measures to be more greatly impacted by nations that have large populations.

Without weighting, all nations would have equal shares and weightage reflected in the statistic measures, which is not desirable.

We chose population as the parameter used to weight the data since it affects both variables, labor force participation rate as well as GDP. A good example is how India has a large GDP but equally large population so the GDP per capita would be lower.

Regression

$y = \beta x + \alpha + \varepsilon$

	GDP vs LFPR (non-weighted)	GDP vs LFPR (weighted)
α	-26205.41	-17941.37
β	687.15	509.66
R^2	0.1134	0.1222
p value	0.01	0.011
Predicted Reg Eq	y = 687.16x -26205.41	y = 509.66x - 17941.37

As we can see there is a significant difference in the IV/DV relationship when we weight the sample data by population.

Non-weighted

Interpreting beta:

One percent increase in the Labor Force Participation Rate in the proportion of population aged 15-64 causes an increase in GDP by 687.15 USD.

Interpreting alpha:

If no one is 'economically active', the GDP will be -26205.41 USD (based on estimates).

P&R^2 Values:

P-value is 0.010, meaning the regression model is very significant. There is a significant association between the IV and DV variables.

R squared is 0.1134, meaning that 11.34% of the variability is explained by the regression model.

Weighted

Interpreting beta

One percent increase in the Labor Force Participation Rate in the proportion of population aged 15-64 causes an increase in GDP by 509.66 USD.

Interpreting alpha

If no one is 'economically active', the GDP will be -17941.37 USD dollars (based on estimates).

P&R^2 Values

P-value is 0.011, meaning that the regression model is significant.

R squared is 0.1222, meaning that 12.22% of the variability is explained by the regression model.

Logarithmic Variable

On STATA, we generated a new variable LGDP (GDP in the logarithmic scale) from GDP.

LGDP gives us GDP of the countries in terms of the proportion or percentage.

This variable will enable us to know the percentage change in GDP when LFPR changes.

LGDP will be more useful when we compare the labor force participation rate to GDP as both are proportions, and so they are more comparable.

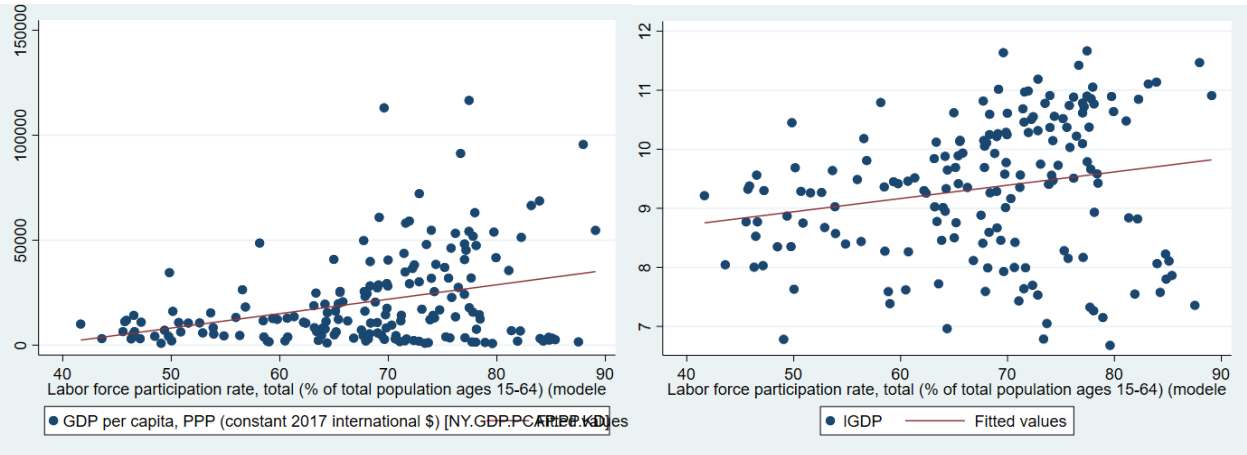
Regressions of LGDP vs LFPR and GDP vs LFPR

$y = \beta x + \alpha + \varepsilon$

Unweighted

	GDP vs LFPR (unweighted)	LGDP vs LFPR (unweighted)
α	-26205.41	7.81
β	687.1517	0.022
R^2	0.1134	0.0416
p value	0.00	0.007
Estimated reg	y = 687.15x -26205.41	y = 0.022x + 7.81

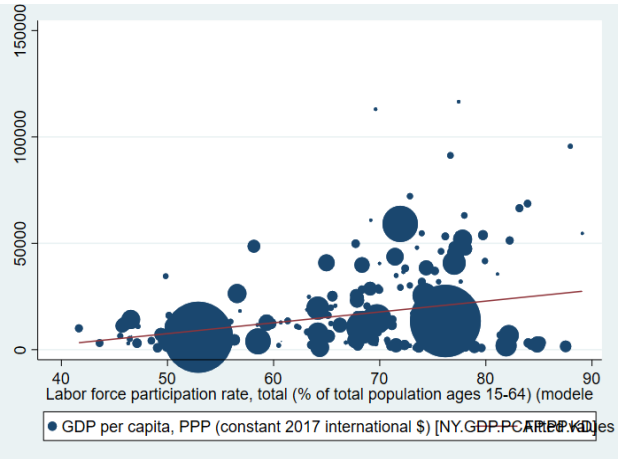
GDP vs Labor (unweighted)	LGDP vs Labor (unweighted)

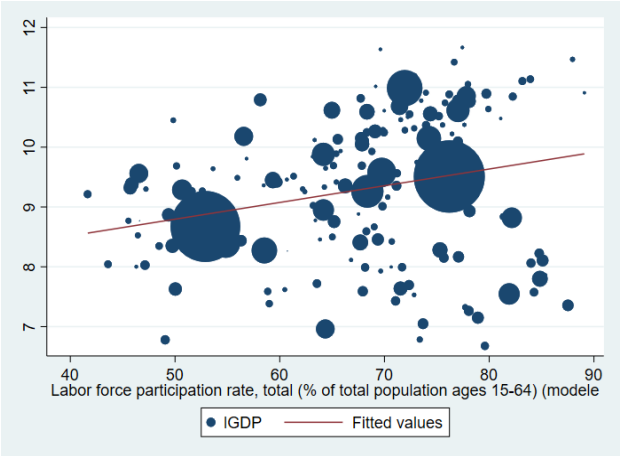


Weighted

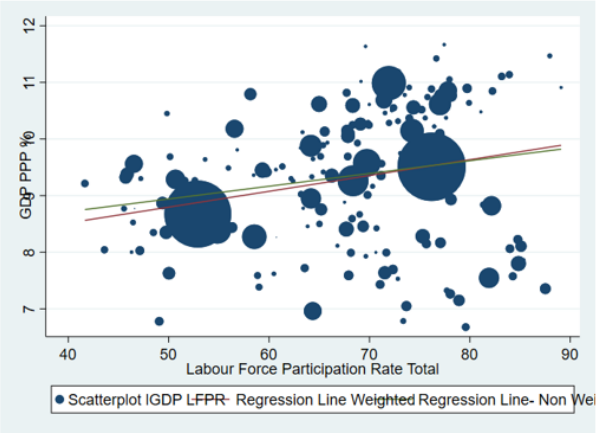
	GDP vs LFPR (weighted)	LGDP vs LFPR (weighted)
α	-17941.37	7.40
β	509.66	0.028
R^2	0.1222	0.105
p value	0.00	0.00
Estimated reg	y = 509.66x -17941.37	y = 0.028x +7.40

GDP vs Labor(weighted)	LGDP vs Labor(weighted)

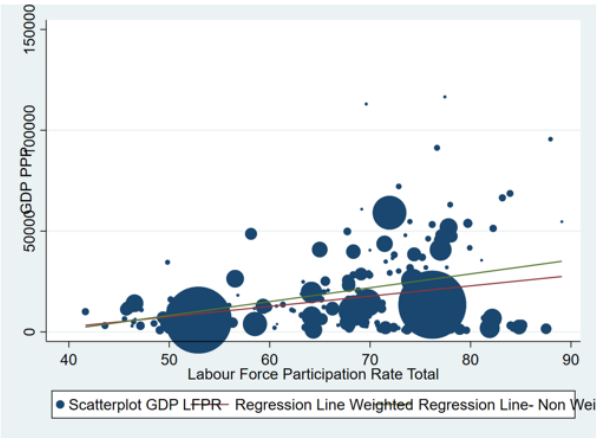




Summarization Using Two Way Scatters



LGDP vs LFPR



GDP vs LFPR

Given above are the weighted regression of both GDP vs LFPR and LGDP vs LFPR. The regression of lines of each are also plotted. The red line in both graphs signifies the weighted regression and the green line signifies the unweighted regression. The comparison between both graphs shows us that in case of LGDP vs LFPR, it is much easier to observe both the variables. Also, even though there isn't much difference between the weighted and unweighted regressions of both, we observe differences further down the study as we divide out samples into individual quartiles.

Splitting the Sample into Quartiles Based on Population

Since we have weighted the data based on population, we have a more accurate regression that shows us how IV affects DV, relative to the sizes of countries' populations in the sample.

But we do not know which quartile of the population affects regression/correlation the most.

For this reason, we split up the population variable into 4 quartiles and split the data into 4 groups based on the population. We now analyze the regression and graphs for each group separately.

GDP vs LFPR (weighted) divided into 4 quartiles

	Q1 25%	Q2 50%	Q3 75%	Q4 100%	Overall
α	-33531.54	-68495.19	-21008.35	-16626.89	**-17941.37**
β	954.28	1381.52	535.55	487.79	**509.66**
R^2	0.056	0.285	0.103	0.122	**0.1222**
p value	0.615	0.002	0.186	0.190	**0.011**
Corr	0.2370	0.5336	0.3212	0.3494	**0.2039**

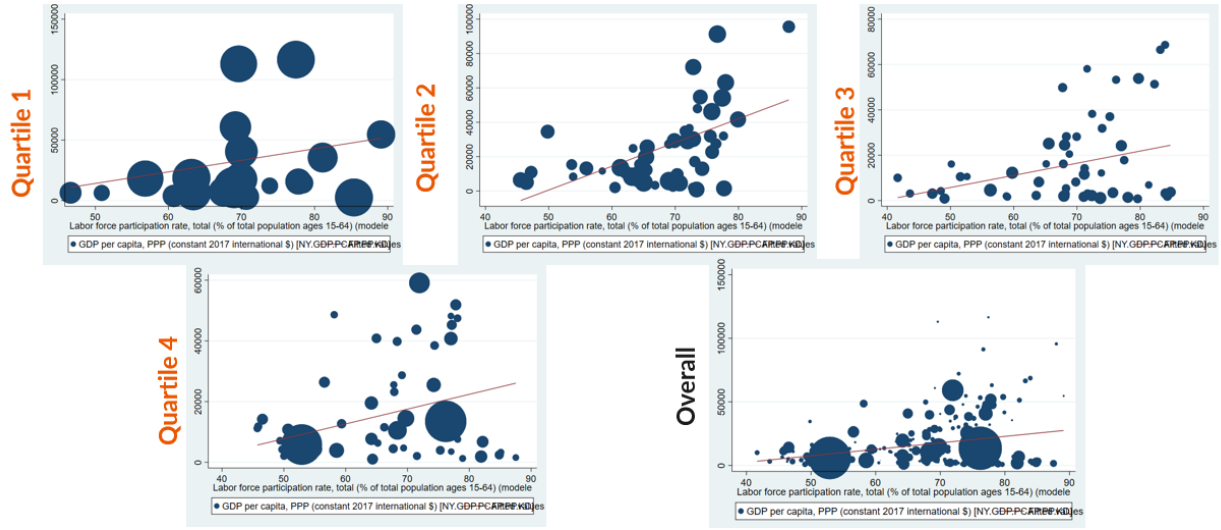
LGDP vs LFPR (weighted) divided into 4 quartiles

	Q1 25%	Q2 50%	Q3 75%	Q4 100%	Overall
α	8.57	6.57	7.53	7.36	**7.40**
β	0.018	0.045	0.021	0.029	**0.028**
R^2	0.020	0.135	0.030	0.127	**0.105**
p value	0.001	0.00	0.00	0.00	**0.00**
Corr	0.1413	0.3676	0.1740	0.3559	**0.3240**

Given above are the individual parameters of the regression and correlation of GDP vs LFPR (weighted) and LGDP vs LFPR (weighted) for each quartile. We find that the countries in quartile 2(lower mid population countries) make the most significant contribution to overall correlation while the countries in quartile 1(very low population countries) and quartile 4(very high population countries) making the least contribution and sort of act as outliers in case of our study. The P - values show that the regression is more significant in case of LGDP vs LFPR than it is for GDP vs LFPR.

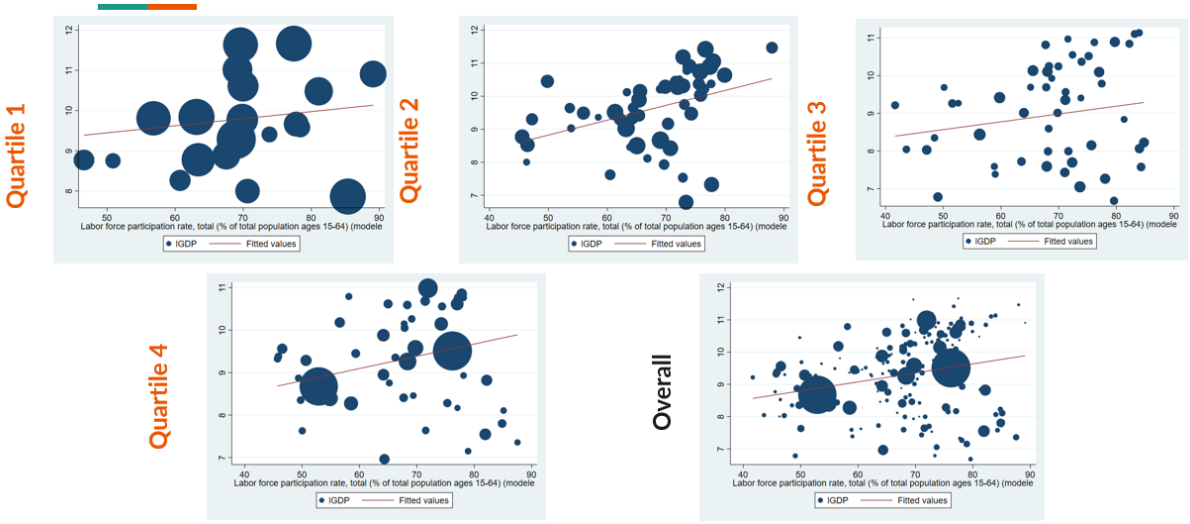
Scatter Plots- GDP vs LFPR (weighted)

Given below are the individual scatterplots of GDP vs LFPR (weighted) for the four quartiles. We observe that the regression line of quartile 3 is most close to the regression line of the overall data.



Scatter Graphs – LGDP vs LFPR (weighted)

Given below are the individual scatterplots of LGDP vs LFPR (weighted) for the four quartiles. We observe that the regression line of quartile 3 is most close to the regression line of the overall data.



V. Conclusion

After running several regressions, summarizing data, generating scatterplots, calculating correlations, using methods of random sampling and weighting we come to the following conclusions:

- Change in Labor Force Participation Rate weighted by population positively impacts change in GDP per capita PPP of countries with a correlation of 0.3.
- The correlation between them is linear but weak stemming from the presence of several outlier data points that lie far from our line of best fit.
- Segregating countries based on population, we observe that countries lying on the second quartile impact most to the correlation, meaning that our hypothesis holds true most for these countries.
- On the other hand the countries lying on the first and last quartiles impact least to the correlation, meaning that these countries serve as outliers in case of our regression.

Potential Sources of Bias

Every regression has an error term that has to be justified with potential variables or effects that bias the sample results.

Since we are not fully aware of the methods used by World Bank we cannot assert any sources of bias within the method of data collection. However, we can predict a potential lurking (omitted) variable.

The availability or abundance of natural resources is a potential lurking variable:

- A discovery of natural resources would lead to increased production/ extracting/ mining it, resulting in an increase in GDP and hence GDP per capita.
- A discovery of resources would lead to the formation of more jobs needed in the extracting/ mining/ associated industries. Assuming labor will always fill open jobs, and discovery of resources would increase LFPR.

Limitations of the Study

1. Only 175 variables (countries) were used and data was collected for 1 year -> relatively small data pool. Improvement -> increase the sample size by collecting data over multiple years (not just 2016), increasing the countries involved may be unfeasible as data was not available.

2. Labor force participation rate shows the proportion of people 'economically active', as stated by the World Bank. One major assumption of the study is that LFPR reflects employment levels. It is unclear exactly what being 'economically active' entails; it may not actually be employment.

3. There might be lurking variables present in the error term which will bias the regression and it is beyond our scope to calculate them and maybe much different from our predicted regression.

Strengths of the Study

1. Compared labor force participation rate and GDP per capita. These are both processed data variables, hence they can effectively be compared to reach conclusions about how IV affects DV. Using LGDP puts GDP in a proportion, and labor is already in a proportion so data is comparable.

2. Weighting the sample by population made the data more reflective of the real world. Nations with large populations would occupy a higher weight in the sample, making the results more representative.

3. The sample was split into 4 sets to identify which quartile affected the regression the most. This was useful in seeing how the weight of each country (by population) directly affected the regression model.

4. The data source used was reliable and credible. It contained data that was collected through a systematic methodology, increasing the validity of the findings.