Predicting GDP per capita for U.S. States using State-level Factors

Jatin Suri

Abstract - GDP per capita is an important indicator of economic performance and is helpful to compare economic wellbeing between areas. A majority of literature in economics focuses on predicting a country's GDP per capita using macroeconomic variables such as interest rates, inflation, unemployment, etc. There, however, is a lack of papers that study GDP per capita for U.S. states, which could be a beneficial resource for state policymakers. This paper uses multiple regression analysis to look at state-level factors that may be valuable for predicting GDP per capita and comments on these factors' significance.

I. Introduction

State policymakers and economists often track key economic indicators such as inflation, employment figures, home sales, etc., to ensure that their state's economy is healthy. One such indicator that is primarily used to measure economic performance is GDP which measures the total monetary value of goods and services produced in a state. Since states with a higher population are likely to have a higher GDP, economists are more interested in GDP per capita, which is GDP divided by the population of the state. GDP per capita is a better indicator for economic performance because, unlike GDP, it removes bias towards populous states.

Since GDP per capita is an excellent indicator of a state's economic performance relative to other U.S states, factors that lead to a higher GDP per capita should be of high interest to state policymakers because they can plan to improve these factors. Therefore, determining what these factors are and which are most significant will be a beneficial resource to improve state policies.

II. Literature Review

The majority of literature estimates GDP per capita for countries using macroeconomic variables. For instance, Desai et al. (2016) build a multivariable regression model to predict GDP in India using CPI, interest rate, manufacturing index, and oil prices. The paper finds inflation and the manufacturing index to be statistically significant. Additionally, Samiyu (2021) finds disposable income, unemployment rate, and house price index to have a significant relationship with GDP per capita. However, there is a lack of papers that speak directly to state-level factors that would impact GDP per capita for the U.S. States.

There are likely a variety of economic and social variables that can help explain the variance in GDP per capita. Valliere and Peterson (2010) find a relationship between entrepreneurship and economic growth, which suggests that factors such as the number of patents or startups in a state could be useful in predicting GDP. Additionally, Ivanova and Masarova (2013) find a link between road infrastructure and economic development, implying that factors such as road quality or housing conditions may predict GDP per capita.

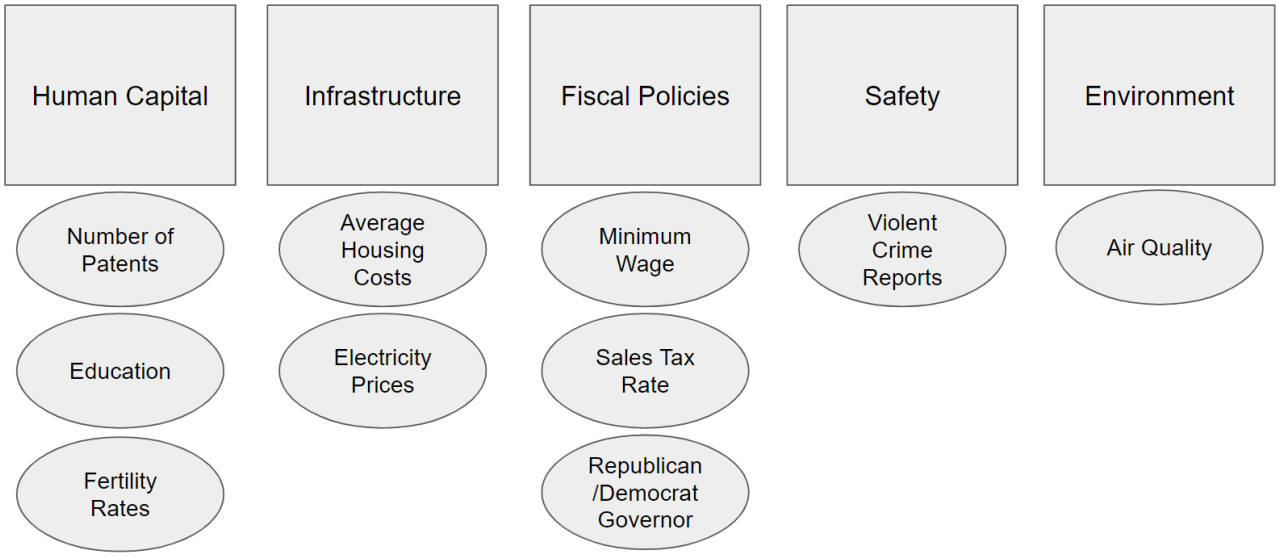
III. Theory

Economic literature shows that macroeconomic variables can be used to predict GDP per capita for countries. This idea can be applied to U.S. states as they are similar in structure to countries with minute differences. Instead of using macroeconomic variables, this paper analyzes state-levels factors (social and economic) that may correlate with GDP per capita.

Since GDP per capita is defined as an output, it may be modeled using the production function. The production function is expressed as $Q = f(K, L, P, H)$, where K, L, P, H are the factors of production: Physical capital (K), Labor (L), Land (P), and Entrepreneurship (H). Using similar

logic, state-level factors can be expressed by breaking down a state into different factors, such as those that relate to human capital, infrastructure, fiscal policies, safety, and the environment.

Figure 1: Variable Selection



Each factor is decomposed into variables related to the factor, and data for each variable is collected to determine which are significant in predicting GDP per capita.

Table 1: Hypotheses

Variable	Positive/Negative Correlation with GDP per capita	Rationale
Number of Patents	Positive	A greater number of patents most probably indicates a higher level of technology in the state which leads to greater output.
Education	Positive	More educated individuals often have a higher income and thus spend more.
Fertility Rates	Negative	A higher fertility rate may mean that families have to take more time off work which leads to less production. Also, a greater number of kids suggests a lower average education among them as the costs of education for more kids is higher.

Average Housing Costs	Positive	Cost of houses and land is more expensive in urban/developed areas.
Electricity Prices	Negative	Developed states can invest in renewable and cheap energy. Lower electricity prices incentivize innovation.
Minimum Wage	Positive	A higher minimum wage relative to other states may suggest that the state has a higher GDP as people have more income to spend.
Sales Tax Rate	Positive	Higher taxes suggest higher government income which could lead to more government spending.
Republican/Democrat Governor	Unknown	Unclear if Republican or Democratic policies lead to greater GDP per capita.
Violent Crime Reports	Negative	A state with more GDP can invest in crime prevention.
Air Quality	Unknown	States that produce a lot of goods may emit a lot of pollution; however, richer states may also have the renewable technology/regulations to focus on better air quality.

IV: Model Specification

A multiple linear regression model is used to predict GDP per capita, and can be represented as follows:

$$GDP_c = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

where GDP_c is the response variable; β_0 is the intercept; n is the number of explanatory variables; β_n is the slope for x_n, x is the independent variable; ϵ is the error term. In multiple regression, the main null hypothesis is that there is no relationship between the x and y variables, while the alternative hypothesis is that there is some relationship between the two. These hypotheses can be illustrated as follows:

$$H_0: \beta_N = 0$$

$$H_a: \beta_N \neq 0$$

$$\text{where } 1 \leq N \leq n$$

When using regression, it is important to verify that the Guess-Markov assumptions of linearity are met to guarantee the validity of ordinary least squares. Since regression is being used to predict GDP per capita, GDP per capita for the state is the response variable, while the other 12 variables are the explanatory variables.

V. Data

The data includes 51 observations accounting for all 50 states and the District of Columbia. Washington D.C. is included in the data set because even though it is not a state, all of the data sources include D.C., and it houses a population greater than some states such as Wyoming or Vermont. All of the data is from 2018 because there is more data available, and a pre-covid data set would be a better indicator of "normal" data. Furthermore, all of the data is in nominal dollar values.

There are no significant data limitations as all observations for each variable are accounted for. Additionally, most of the data is derived from government organizations such as the FBI, Census Bureau, NSF, etc., so it is most likely credible.

Table 2: Variable Descriptions

Variables	Description	Data Source
$\boldsymbol{y}$ = gdpc	GDP per Capita (nominal $)	Bureau of Economic Analysis
$\boldsymbol{x_1}$ = educhs	Percent of population with high school diploma or higher	U.S. Census Bureau
$\boldsymbol{x_2}$ = educba	Percent of population with a BA or higher	U.S. Census Bureau
$\boldsymbol{x_3}$ = educprof	Percent of population with a professional degree or higher	U.S. Census Bureau
$\boldsymbol{x_4}$ = fertrate	General fertility rate per 1,000 women aged 15-44	Centers for Disease Control and Prevention
$\boldsymbol{x_5}$ = avghouse	Average home price (nominal $)	Zillow
$\boldsymbol{x_6}$ = crime	Reported violent crime rates	Federal Bureau of Investigation

x_7 = patent	# of patents per 1,000 in science and engineering	National Science Foundation
x_8 = minwage	State minimum wage (nominal $)	Labor Law Center
x_9 = rdgov	State governor party affiliation: Republican = 0, Democrat = 1	Ballotpedia
x_{10} = airquality	Air quality index	U.S. Environmental Protection Agency
x_{11} = saletax	Sales tax rate	Tax Foundation
x_{12} = elecprice	Average electricity prices (nominal $)	U.S. Energy Information Administration

Table 3: Descriptive Statistics

Variables	Observations	Mean	Std. dev.	Min	Max
y = gdpc	51	56789.88	20163.01	35015	178442
x_1 = educhs	51	0.900	.0270004	.84	.95
x_2 = educba	51	0.327	.0663278	.21	.6
x_3 = educprof	51	0.126	.0416088	.08	.34
x_4 = fertrate	51	59.68431	5.769831	47.2	73.6
x_5 = avghouse	51	307774.8	141361.4	117768	730511
x_6 = crime	51	396.109	175.8936	108.6	999.8
x_7 = patent	51	3262.431	6720.664	50	46145
x_8 = minwage	51	8.607	1.510797	7.25	13.25
x_9 = rdgov	51	0.471	.5041008	0	1
x_{10} = airquality	51	42.304	5.248236	21.2	51.2
x_{11} = saletax	51	0.0517	.0198702	0	.07
x_{12} = elecprice	51	11.219	4.020397	7.71	29.18

VI. Model Validation and Results

Model 1: (Initial Parameters)

$$gdp_c = \beta_0 + \beta_1(educhs) + \beta_2(educba) + \beta_3(educprof) + \beta_4(fertrate) + \beta_5(avghouse) + \beta_6(crime) + \beta_7(patent) + \beta_8(minwage) + \beta_9(rdgov) + \beta_{10}(airquality) + \beta_{11}(saletax) + \beta_{12}(elecprice)$$

The first model includes all of the initial explanatory variables; however, it violates the Gauss-Markov assumptions of heteroskedasticity and multicollinearity. It also results in a model specification error. The model violations are tested using the Breusch-Pagan/Cook-Weisberg test, variance inflation factor, and Ramsey RESET test, respectively.

Model 2:

$$\ln(gdp_c) = \beta_0 + \beta_1(educhs) + \beta_2(educba) + \beta_3(educprof) + \beta_4(fertrate) + \beta_5(avghouse) + \beta_6(crime) + \beta_7(patent) + \beta_8(minwage) + \beta_9(rdgov) + \beta_{10}(airquality) + \beta_{11}(saletax) + \beta_{12}(elecprice)$$

Model 2 includes a log-transformation of the response variable (GDP_c), which fixes the specification error. It also drops a few variables to reduce multicollinearity and improve the adjusted-R squared value. The log-transformation reduces heteroskedasticity slightly; however, the p-value from the Breusch-Pagan test is still below the 5% significance level. As a result, robust standard errors are applied.

Table 4: Model 2 Regression Results

lgdp	Coefficient	Robust std. err.	t	P >|t|	[95% conf. interval]	
educba	3.190878	.4486985	7.11	0.000***	2.285991	4.095765
avghouse	-6.81e-08	1.90e-07	-0.36	0.722	-4.51e-07	3.15e-07
crime	.0002921	.0001346	2.17	0.036**	.0000206	.0005636
patent	4.28e-06	1.83e-06	2.34	0.024**	5.87e-07	7.97e-06
minwage	.0190487	.0192224	0.99	0.327	-.0197169	.0578143
rdgov	.0362146	.0392559	0.92	0.361	-.0429525	.1153817
fertilityrate	.0123611	.0045815	2.70	0.010***	.0031215	.0216006
_cons	8.837124	.3909808	22.60	0.000***	8.048636	9.625612
*,**,*** indicates significance at the 10%, 5%, 1% level respectively						
Observations = 51	F(7, 43) = 18.91	Prob > F = 0.0000	R-squared = 0.7280	Root MSE = .14184		

Model 3: Final Model

$$\ln(gdpc) = \beta_0 + \beta_1(educba) + \beta_2(crime) + \beta_3(patent) + \beta_4(fertilityrate)$$

Table 5: Final Model Regression Results

lgdp	Coefficient	Robust std. err.	t	P >|t|	[95% conf. interval]	
educba	3.366427	.4028551	8.36	0.000***	2.555522	4.177332
crime	.000314	.0001367	2.30	0.026**	.0000389	.0005891
patent	5.02e-06	1.36e-06	3.70	0.001***	2.29e-06	7.75e-06
fertrate	.0106317	.0042796	2.48	0.017**	.0020172	.0192461
_cons	9.031796	.3282641	27.51	0.000***	8.371035	9.692558
*,**,*** indicates significance at the 10%, 5%, 1% level respectively						
Observations = 51	F(4, 46) = 29.95	Prob > F = 0.0000	R-squared = 0.7178	$\sqrt{MSE}$ = .13968		

The final model includes the statistically significant variables at a 5% significance level. The three variables: *avghouse*, *minwage*, and *rdgov* were dropped from the initial model because their inclusion only contributed to an additional 0.01% explanation of variance in GDP per capita. There also was no linking economic theory that suggested that these variables needed to be included in the model.

The model should be checked with the Gauss-Markov assumptions. A variance inflation factor test results in VIF values less than two for all variables suggesting that the model does not have a multicollinearity problem.

Table 6: Multicollinearity Test - Variance Inflation Factor

Variable	VIF	1/VIF
fertrate	1.47	0.681196
educba	1.43	0.699018
crime	1.03	0.966872
patent	1.03	0.968111
Mean VIF	1.24	

The model is tested for heteroskedasticity using the Breusch-Pagan/Cook-Weisberg test. Since the p-value > 0.05, the null hypothesis that there is homoscedasticity is accepted. Therefore, the model does not have a heteroskedasticity problem. As a result, robust standard errors are not needed; however, they are included in Table 4.

Table 7: Heteroskedasticity Test – Breusch-Pagan/Cook-Weisberg

$\boldsymbol{H_0}$: Constant Variance
chi2(1) = 0.34
Prob > chi2 = 0.5622

The model is also checked for a specification error by running the Ramset RESET. Since the p-value > 0.05, the null hypothesis that there is no specification error is accepted.

Table 8: Specification Error Test – Ramsey RESET

$\boldsymbol{H_0}$: Model has no omitted variables
F(3, 43) = 2.41
Prob > F = 0.0798

The model checks the critical Gauss-Markov assumptions and explains approximately 72% of the variance in GDP per capita.

VII. Discussion

All four of the statistically significant variables (*educba*, *fertrate*, *crime*, *patents*) have a positive correlation with $\ln(gdp_c)$. Understandably, *educba* and *patents* have a positive correlation because people with higher education can earn more money and consume more, resulting in a higher GDP. Additionally, a higher number of patents within a population suggests more innovation and R&D is occurring within a state. Therefore, the state probably has greater technology and hence a better means of production.

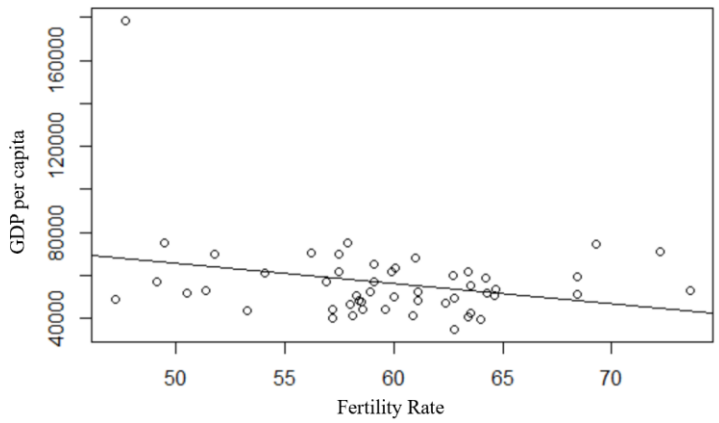
On the other hand, it is odd that *fertrate* and *crime* are negatively correlated with GDP. If a state has more income, then the state can implement more crime prevention strategies and spend more on law enforcement. This would suggest that crime and GDP have a positive relationship; however, it is also possible that a negative relationship might develop. Crime is more likely to happen in impoverished areas and GDP per capita does not explain income inequality within a region. As result, states with a higher GDP per capita may have more areas of income inequality and thus greater crime rates.

If a state has a higher fertility rate, then that suggests that women are having a greater number of children on average. Since families are more likely to take time off of work when having

children, the opportunity cost of more kids includes the loss of working in the economy. Furthermore, the average education among more children in a household will most likely be lower than the average education with fewer children because parents will split their educational spending among more people.

The positive correlation between GDP per capita and fertility rate in the final model is surprising because the data suggests that in a 1-variable linear regression, there is a negative relationship.

Figure 2: Fertility Rate Plot



The inclusion of other variables in the multiple regression model is causing the fertility rate coefficient to appear positive. This is a large limitation of the model.

Table 9: Standardized Beta Coefficients (β^*) - Economic Significance

Variable	β^*
educba	$11.07 \cdot 10^{-6}$
crime	$2.74 \cdot 10^{-6}$
patents	$1.67 \cdot 10^{-6}$
fertrate	$3.04 \cdot 10^{-6}$

The strength of the effect of the significant variables can be measured by the standardized beta coefficients. The *educba* variable has the greatest weight, while *fertrate*, *patents*, and *crime* have

much smaller weights. As a result, this regression indicates that education plays a leading role in determining a state's GDP per capita.

Limitations

Even though GDP per capita adjusts for population size, including the District of Columbia is problematic because the size of the region is miniscule relative to the U.S. States. Therefore, it accounts for more urban industrialization and hence greater GDP. Removing D.C. as an outlier can potentially enhance the predictive power of the model.

Another limitation of the model is the sample size. The data is cross-sectional and only includes the 51 observations for 2018. Potentially, this data set could be expanded to include GDP per capita for other years by calculating simple averages. An increased sample size would result in a better model.

The number of state-level factor categories is also a potential limitation of this study. Only five categories are selected (human capital, infrastructure, fiscal policies, safety, and the environment), but more categories could be created, such as government integrity/corruption, life satisfaction, etc.

Finally, a better distribution of state-level factors can be created to understand which categories are the most important for state policymakers to focus on. For instance, in this, there are multiple factors from human capital, while there is only one factor within environment.

VIII. Conclusion

This paper develops a multiple regression model to better understand significant factors that can predict GDP per capita for U.S. states. The results indicate that education, fertility rates, crime,

and patents are all useful in helping predict GDP per capita. Additionally, education plays the largest role in explaining GDP per capita. Therefore, state policymakers should increase their efforts to improve education for their citizens because better education may result in higher state production.

IX: Bibliography

Dave Valliere & Rein Peterson (2009) Entrepreneurship and economic growth: Evidence from emerging and developed countries, Entrepreneurship & Regional Development, 21:5-6, 459-480, DOI: 10.1080/08985620802332723

Eva Ivanova & Jana Masarova (2013). Importance of road infrastructure in the Economic Development and competitiveness. *ECONOMICS AND MANAGEMENT*, *18*(2). https://doi.org/10.5755/j01.em.18.2.4253

Mutiu Samiyu (2021). Multiple regression model for predicting GDP using macroeconomic variables (part 1). *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3895177

Nirav Desai (2016). A multiple variable regression model for gross domestic product growth rate prediction in India using key macroeconomic indicators. *ISOR Journal of Economics and Finance.* DOI: 10.9790/5933-0702034751

X: Appendix

Dataset

https://docs.google.com/spreadsheets/d/1qiOCvfyTUbambSTB4GBl5XmMgAEyYs-d/edit?usp=sharing&ouid=108625195237286430469&rtpof=true&sd=true

β* coefficients calculations

$\beta^* = |b_{ols}\frac{s_x}{s_y}|$

educba: $\beta^* = |3.366427 \cdot \frac{.0663278}{20163.01}| = 11.07 \cdot 10^{-6}$

crime: $\beta^* = \left|.000314 \cdot \frac{175.8936}{20163.01}\right| = 2.74 \cdot 10^{-6}$

patents: $\beta^* = |5.02e-06 \cdot \frac{6720.664}{20163.01}| = 1.67 \cdot 10^{-6}$

fertilityrate: $\beta^* = |.0106317 \cdot \frac{5.769831}{20163.01}| = 3.04 \cdot 10^{-6}$