Financial Engineering and Risk Management (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/ferm.2023.061109
ISSN 2523-2576 Vol. 6 Num. 11

Relationship between higher education level and GDP per capita of different American States

Heming Wang

Louis Olin Business School, Washington University in St, St. Louis, USA

Keywords: Higher education, GDP per capita, American States, Education level

Abstract: Education has long been considered a vital factor that determines a person's income level. This paper aims to explore if this effect still exists after being magnified to a state-wide scale, and, if it still does, the extent to which this effect can be observed. The percentage of the population aged 25 or older that are bachelor's degree holders, by state, is used as a measure of the education level of a state, and this paper attempts to find the relationship between this value and GDP per capita of a state. Furthermore, this paper takes other variables into account, which are unemployment and urban population, to better model the effect of education on GDP per capita.

1. Introduction

There are abundant ways to increase income according to economic theory, and education is amongst one of those that works by increasing labor productivity. In general, for individuals the higher a person's education level is, the higher that person's income level is likely going be. This effect, when magnified to a state-scale, is supposedly going to be that the higher a state's education level is, the better its economic outlook is going to be. Whether this stands true is still required to be proven. However, "education" itself is hardly any quantity or value that can be used and plugged into the calculation, so it needs further interpretation. It can be represented by many measures, such as the number of people/ percentage of the population that has a bachelor's degree or higher, secondary education enrollment rate, etc. It would be reasonable to assume that high-income positions that are filled by holders of bachelor's degrees or higher contribute the most to the real GDP, therefore this paper utilizes the data of the percentage of the population aged 25 or older that are bachelor's degree holders. This chosen age group is the potential long-term participants in the labor force, so they have the strongest impact on a state's output.

This paper hypothesizes that a state's GDP per capita is positively related to its education, meaning that as the education level of a state increases, the GDP per capita of that state increases consequently. The primary independent variable is the percentage of the population aged 25 or older that are bachelor's degree holders. This variable is considered to potentially have an effect on a state's GDP per capita, which is the dependent variable of the models created in this paper. The above-mentioned relationship comes through two ways of justification. First, since it is a general trend that the higher one's education is, the higher income one can earn, when it comes to a group of individuals, this trend should be expected to exist. If that is the case, a state with more holders of a college degree would be expected to have higher per-person GDP. The second way of justification

is that a highly educated crowd is the largest contributor to a state's GDP, so if a state has a large crowd of highly educated individuals, the states' GDP would be boosted greatly and the per-person GDP would also increase since the population does not change.

However, other factors play important roles in this process. To get to the relationship between education and GDP, there is still one more process needed to connect the two ends, and that is employment. That is also the reason that the multilinear regression model takes into account employment factors such as unemployment rate, the participation of owners of a college degree or higher in the labor force.

This topic has its significance in helping the policymaker in making decisions on future educational expenditure. If the relationship between higher education level and GDP per capita is observable and obvious, then it indicates that investment in education is an important booster for not only the economy but also the quality of life for individuals. Furthermore, if this topic is extended and aligned with other studies that focus on state-specific economic development models, more useful results could potentially be generated. That could include other policies that could complement the investment in higher education in terms of boosting GDP per capita or even the economic development patterns that do not require a fairly large crowd of highly educated, skillful individuals to achieve high GDP per capita such as the ones that could explain the high GDP per capita and relatively low percentage of highly educated individuals appeared in Alaska's economy.

2. Literature Review

One of those that explore the economic impact of universities was written by Valero and van Reenen (2018). In this paper, the authors pointed out the overlapping trend of an increasing number of universities/ increasing number of universities per million person and mean growth in GDP per capita. This trend can be shown by a scatter plot that plots the mean growth of GDP per capita and the number of universities per million people and the trend line following it. They modeled the hypothesized relationship with cross-sectional regressions and found a strong and positive correlation between GDP per capita and universities. Further results suggest that a 10% increase in the number of universities in a region is related to about 0.4% higher GDP per person.[1]

Another paper contributing to this area of research is written by Aghion et. al. (2009). This paper focuses on the effect of all educational sectors, instead of college. The authors built complicated models accounting for the effect of education investment on GDP and introduced effects such as migration of skilled labor into their models. In a conclusion, the paper finds support for the hypothesis that some investments in education do stimulate economic growth.

What's more, another piece of work composed by Odit et. al. (2010) also provides interesting insight into the area. They see education attainment as a contributor to the quality of human capital, which is an engine for economic growth.

The authors utilized the Cobb-Douglas production function with constant returns to scale where human capital is used as an independent factor of production in the human capital augmented growth model.[2] As for their conclusion, the calculated results suggest that education is productivity-enhancing rather than "a device that individuals use to signal their level of ability to the employer", which is interesting.

All the papers reviewed have made a common point in their conclusion, which is that education does have a positive impact on the economy, whether it be from a technology innovation and R&D perspective or a labor productivity perspective. They provide an important base and guideline for this paper.

3. Data

This paper aims to find the relationship between a state's real GDP per capita, and its education level, measured by the percentage of the population with a bachelor's degree or higher. For GDP per capita, the data used here is measured in chained 2012 dollars and it has the advantage of being inflation- adjusted. It is a measure of each state's per-person GDP that is based on inflation-adjusted national prices for the goods and services produced within a state. For education level, the measure is the percentage of the population of a state aged 25 and older that has a bachelor's degree or higher. The year of both data sets is chosen to be 2019, which has the advantage of being relevant and having an intact set at the same time. The data of real GDP per capita is obtained from U.S. Bureau of Economic Analysis (BEA) which is an authoritative and credible official source, especially for economic data.[3] The data of the percentage of the population aged 25 and over who have completed a bachelor's degree or higher in the U.S. in 2019, by state is also obtained from BEA.

Other variables that also play a role in affecting the GDP per capita of states are unemployment, urban population, labor participation rate, and total labor force. The unemployment rate defines how big or how saturated the labor pool is, so if a state has a high education level but also high unemployment, the effect of education on GDP per capita is going to be offset to some extent. Labor participation rate defines the proportion of the working-age population that is either working or actively looking for work so the higher the rate, the higher the likelihood that the highly skilled employees are participating in the generation of GDP. The total labor force can also affect GDP per capita. The sample sizes are the number of the states in the US which is 50 plus Washington D.C. The year for the controlled variables' data is also 2019, for the sake of consistency with other variables except for the urban population as the census is only decennial and the latest data is not yet available. As for the source, the unemployment rate by state is obtained from FRED, the urban population figures are obtained from U.S. Census Bureau, the labor participation rate is obtained from U.S. Bureau of Labor Statistics, which is namely the most authoritative source for employment-related data, and the total workforce size by state is obtained from Bureau of Labor Statistics. Figure 1 is an overview of all the variables used in the model that is going to be generated next.

	Description	Year	Units	Source
gdpp	GDP per capita (by state)	2019	2012 dollar value	U.S. Bureau of Economic Analysis
educ	The percentage of the population aged 25 or older that are bachelor's degree holders (by state)	2019	Percentage	U.S. Bureau of Economic Analysis
un	Unemployment rate (by state)	2019	Percentage	FRED
urbanp	Urban population (by state)	2010	Percentage	U.S. Census Bureau
labpar	Labor participation rate (by state)	2019	Percentage	FRED
labf	Labor force (by state)	2019	person	Bureau of Labor Statistics

Figure 1: Brief Description of All the Data Used in This Paper

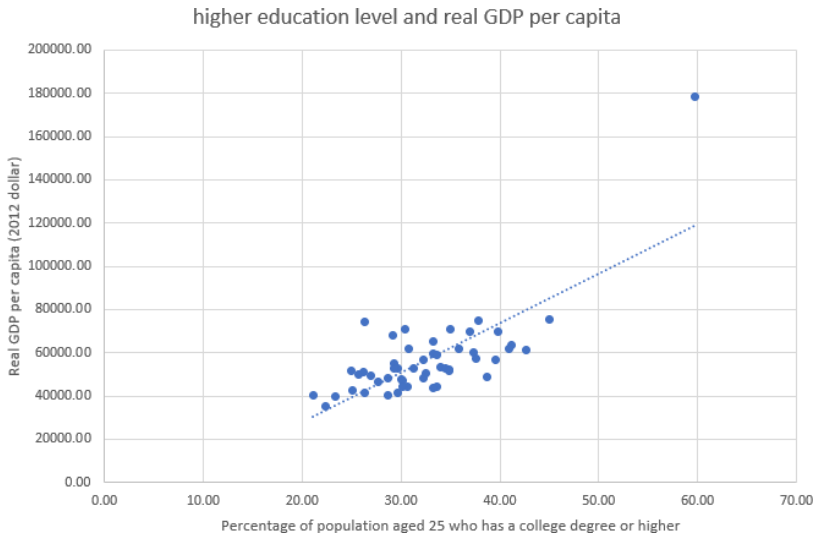

Figure 2: Scattered Plot of Higher Education Level and Real GDP Per Capita

Figure 2 is a scattered plot generated with a linearly best-fitted trendline. Observing just from here, the relationship appears to be weak, but the trend still contains potential. There is one noticeable outlier at the top right corner of the graph and that is D.C. which has a significantly higher education level and GDP per capita than any other states.

One important step before proceeding to regression is making sure that the model meets Gauss Markov Assumptions.

1) Linear in Parameters:

$y = \beta 0 + \beta 1x1 + \beta 2x2 + \ldots + \beta kxk + u$ is a general form of linear regression model. Here loggdpp, which stands for the log of GDP per capita, takes up the place of y on the left-hand side of the equation. And educ, which stands for the percentage of population aged 25 or older that are bachelor's degree holders, is representing x.

$\beta 0$ is the constant here, and $\beta 1$ to βk are the coefficients and u is the unobserved error. This equation does satisfy the first condition for being linear in parameters.

2) Random Sampling:

Since the topic of this paper is about states in US, the sampling done here is among the states. The sample size is 51, including all 50 states and the District of Columbia. Although technically all the data available is sampled, the sampling is still considered random plus the fact that the data include the state with high and low GDP per capita and high and low college degree rates, so the sampling includes a fair range of data.

3) No Perfect Collinearity:

The assumption of no perfect collinearity means that there is no perfect linear relationship (one-to-one) among the independent variables. STATA is utilized to test this assumption. As seen in Fig. 3, no correlation between independent variables is perfect, meaning this model is safe in terms of assumption 3.

	loggdpp	educ	un	urbanp	labpar	loglabf
loggdpp	1.0000					
educ	0.7612	1.0000				
un	0.1408	-0.1436	1.0000			
urbanp	0.5324	0.4997	0.1557	1.0000		
labpar	0.6015	0.6321	-0.3835	0.2613	1.0000	
loglabf	-0.0544	0.0469	0.0588	0.4299	-0.1631	1.0000

Figure 3: Test for Perfect Collinearity

4) Zero Condition Mean:

The zero condition means tests whether, given the values of the independent variables, the expected value for the error term is always 0. One way to test the zero conditional mean is through the residual plot. Judging by the looking of the residual plot, which is shown as Figure 4, from the multiple regression, one can tell that the points are scattered above and below the zero level roughly randomly without an obvious pattern. Therefore, it is safe to conclude that the model passes the test of zero condition mean.

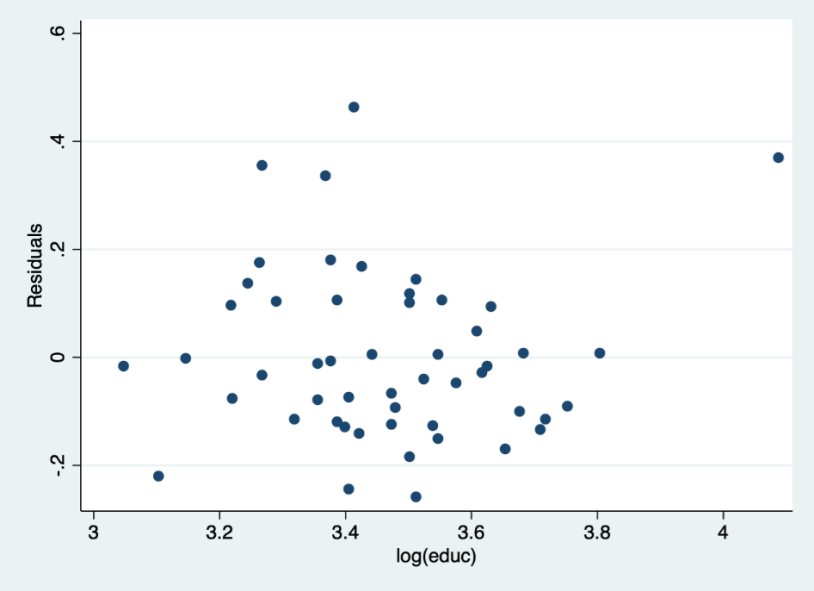

Figure 4: The Residual Plot of Multi Regression Model

5) Homoskedasticity:

The homoskedasticity assumption assumes that different samples have the same variance, so the variance of u should be the same for all. This assumption can also be tested by looking at the residual versus predicted value plot. As the points are roughly evenly distributed, the variance would also approximate a common value. Therefore, this assumption holds.

3.1. Results

3.1.1. Simple Linear Regression model (Model 1)

Here a simple regression model is generated as a first approach to the hypothesized relationship.

$$\log(gdpp) = \beta 0 + \beta 1 \log(educ) + u$$

gdpp: the GDP per capita; educ: the percentage of the population aged 25 and older that has a college degree or higher; u: unobserved error

After regression with STATA, the coefficient $\beta 1$ and constant $\beta 0$ are generated.

Estimated Equation:

$$\log(gdpp) = 4.326 + 0.0126 educ$$

This simple regression model has an adjusted R-squared value of 0.5794 meaning the relationship between logeduc and loggdpp is relatively moderate, which is not ideal for drawing conclusions on the relationship. Another thing to look at here in the table is the coefficient for educ, which is positive, meaning that an increase in educ is followed by an increase in log GDP per capita, proving that the relationship is positive, so part of the initial hypothesis is proven here.

This is the first attempt at modeling the hypothesized relationship, which isn't particularly satisfying for it only provides a limited amount of useful information. That leads the way to a more complicated multi regression model, accounting for the omitted variables in the simple model. The next model will supposedly be better at providing insights into the relationship between educ and loggdpp.

3.1.2. Multi Linear Regression model (Model 2)

Now taking into account of all the controlled variables, the model becomes a multi regression one which has the form of:

$$\log(gdpp) = \beta_0 + \beta_1 educ + \beta_2 un + \beta_3 urbanp + \beta_4 labpar - \beta_5 \log(labf) + u$$

gdpp: the GDP per capita; educ: the percentage of the population aged 25 and older that has a college degree or higher; un: the unemployment rate; un: the unemployment rate; urbanp: urban population; labpar: labor force participation rate; labf: total labor force u: unobserved error

Estimated Equation:

$$\log(gdpp) = 3.8068 + 0.0083 educ + 0.0427 un + 0.0014 urbanp + 0.0094 labpar - 0.0316 \log(labf)$$

Since more variable that contributes to the relationship are added, the R-squared value increased consequently to a much higher 0.7406, which is a significant increase. This indicates that the new model performs better at constructing the overall relationship. The coefficients got from this regression attempt are all positive except for the coefficient of log(labf). Further improvement is thus needed to be made to this model to correctly indicate education's effect on GDP per capita, and that could be removing the log(labf) rate variable and other ones that might not be as significant, like urban populations.

3.1.3. Multi Linear Regression model (Model 3)

$$\log(gdpp) = \beta_0 + \beta_1 educ + \beta_2 un + \beta_4 labpar + u$$

Source	SS	df	MS	Number of obs	=	51
				F(3, 47)	=	39.81
Model	.430481883	3	.143493961	Prob > F	=	0.0000
Residual	.169409513	47	.003604458	R-squared	=	0.7176
				Adj R-squared	=	0.6996
Total	.599891396	50	.011997828	Root MSE	=	.06004

loggdpp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.0095253	.0016768	5.68	0.000	.006152	.0128986
un	.0497983	.0114319	4.36	0.000	.0268004	.0727963
labpar	.0107416	.0030555	3.52	0.001	.0045946	.0168885
_cons	3.564069	.186281	19.13	0.000	3.18932	3.938818

Figure 5: Multi Regression Model Table (Model 3)

This is the new model after removing variables that seem to be relatively insignificant and it can be observed from the table that the R-squared value is lower for this model, which is 0.7176. It is still unsure whether this model with fewer variables is superior to model 2 or the other way, but a clear picture will be shown after running statistical tests to examine those models.

In the next stage, each variable in the three models will be tested upon its statistical significance.

In figure 5, a summary of the variables is given. The first figure in each box represents the coefficient of the corresponding variable on the left and the stars right next to it indicate its significance. One star means that variable is only significant at 10% level of significance, two means it's significant at 10% and 5% level, and three means the variable is significant all the way to 1% level. The figure in the parentheses below it is the standard error of the variable.

Dependent Variable log(gdpp)			
Independent Variables	Model 1	Model 2	Model 3
educ	0.013*** (0.0015)	0.008*** (0.0018)	0.010*** (0.0017)
unem		0.043***	0.050***
		(0.1176)	(0.114)
urbanp		0.001** (0.0008)	
labpar		0.009*** (0.0031)	0.011*** (0.0031)
log(labf)		-0.316* (0.0221)	
intercept	4.326 (0.0512)	3.807 (0.2552)	3.564 (0.1863)
No. of obs.	51	51	51
R-squared	0.5794	0.7406	0.7176

*Significant at 10%, **5%, ***1%

Figure 6: Regression Models Summary

It can be interpreted from the figure 6 that the urbanp and log (labf) are the two least significant variables, so it makes sense to construct a new model without them and examine the new model's performance. One more step is needed to determine whether model 3 is better, having removed those two variables.

4. Extensions Robustness Test

Since urbanp and log(labf) are removed, it is necessary to find out if they are truly insignificant to the model. In this step, an F-test has to be performed. If the rejection of the null hypothesis, which would be that $\beta 3 = \beta 5 = 0$, can't be concluded from the results, then it is safe to say that urban population and labor force are jointly insignificant in the construction of the model here.

H0: $\boldsymbol{\beta_3} = \mathbf{0}, \boldsymbol{\beta_5} =$

H1: H0 not true

Unrestricted Model:

$$\log(gdpp) = \beta 0 + \beta 1 educ + \beta 2 un + \beta 3 urbanp + \beta 4 labpar - \beta 5 \log(labf) + u =$$

Restricted Model:

$$\log(gdpp) = \beta 0 + \beta 1 educ + \beta 2 un + \beta 4 labpar + u \quad F = (SSRr - SSRur)/SSRur * q/(n-k-1)$$

$$= [(0.7406 - 0.7176)/(1 - 0.7406)] * (45/2)$$

$$= 1.99 < 3.204 \ (c)$$

Fail to reject the null hypothesis

Therefore, log (labf) and urbanp are jointly insignificant.

In conclusion, this F-test testified that removing those two variables is the right choice to make to get a more refined model.

5. Conclusions

This paper conducted a study on the relationship between higher education level and GDP per capita of different American States. This paper hypothesizes a positive relationship between the two variables.

(The percentage of the population aged 25 and over who have completed a bachelor's degree or higher and GDP per capita), meaning that the more holders of bachelor's degree a state has, the higher it's GDP per capita would be. This hypothesis is supported by the result of the study as the coefficient of educ in the models is positive. Granted, the factors behind rises and falls of a state's GDP per capita are countless and many of which are extremely hard to model. Some states, such as Alaska and South Dakota, are falling behind most of the other states in terms of college education rates, but they have easily the two of the top GDP per capita figures. This phenomenon is not a disproval of the relationship between education and GDP per capita, but it rather tells the fact that different state runs their own economy differently. Alaska and South Dakota both have abundant oil and gas and the firms tapping those contribute largely to the overall GDP.

The attempts to add and remove variables to refine the model definitely improved the model as it is obvious that the R squared value increased significantly and the variables removed are tested to be insignificant, meaning that the attempts are rather successful.

This study shows that higher education level is vital to the growth of a state's GDP per capita, which has many implications. It also means that pursuing higher education does seem to improve people's quality of life. The main takeaway at the individual's level is that the general rule is that investment in pursuing college degrees does pay off. When it comes to policymakers' perspectives, the point is that government expenditure in constructing a wider and better college education sector would generate higher GDP and would also benefit the society since it also makes people better off on average. Policymakers have long been instilled with the importance of ensuring primary and secondary education sector's vitality and pushing for higher enrollment rate and sometimes the contribution of higher education to the economic growth is easily ignored. The results from this study could potentially make them consider the idea of putting more G spending into building more and better colleges, which serves as a pillar to both the economic growth and a highly skilled labor pool.

References

[1] Aghion, P., Vandenbussche, J., Hoxby, C., & Boustan, L. (2009, March). The Causal Impact of Education on Economic Growth: Evidence from U.S. Retrieved December 11, 2021, from https://scholar. harvard.edu/ files/ aghion/files/causal_impact_of_education.pdf

[2] Odit, M. P., Dookhan, K., & Fauzel, S. (2010). The impact of education on economic growth: The case of mauritius. International Business & Economics Research Journal (IBER), 9(8). doi:10.19030/iber.v9i8.620

[3] Valero, A., & Van Reenen, J. (2019). The economic impact of universities: Evidence from across the Globe. Economics of Education Review, 68, 53-67. doi:10.1016/j.econedurev.2018.09.001

Appendix

List of Observations

District of Columbia	Texas	Wisconsin	New Mexico
Massachusetts	Colorado	Rhode Island	Florida
New York	Minnesota	Louisiana	Arizona
Alaska	Nebraska	Utah	Montana
North Dakota	Hawaii	Oklahoma	Maine
California	New Hampshire	Georgia	Kentucky
Connecticut	Virginia	Nevada	South Carolina

Washington	Pennsylvania	Indiana	Alabama
Wyoming	Iowa	Vermont	Idaho
Delaware	Kansas	North Carolina	West Virginia
New Jersey	South Dakota	Tennessee	Arkansas
Maryland	Oregon	Michigan	Mississippi
Illinois	Ohio	Missouri	

STATA Output

```
. regress loggdpp educ
```

Source	SS	df	MS	Number of obs	=	51
				F(1, 49)	=	67.51
Model	.347606012	1	.347606012	Prob > F	=	0.0000
Residual	.252285384	49	.005148681	R-squared	=	0.5794
				Adj R-squared	=	0.5709
Total	.599891396	50	.011997828	Root MSE	=	.07175

loggdpp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.0126377	.0015381	8.22	0.000	.0095469	.0157286
_cons	4.325594	.051176	84.52	0.000	4.222752	4.428436

```
. regress loggdpp educ un urbanp labpar loglabf
```

Source	SS	df	MS	Number of obs	=	51
				F(5, 45)	=	25.70
Model	.444307301	5	.08886146	Prob > F	=	0.0000
Residual	.155584095	45	.003457424	R-squared	=	0.7406
				Adj R-squared	=	0.7118
Total	.599891396	50	.011997828	Root MSE	=	.0588

loggdpp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.0083492	.0018195	4.59	0.000	.0046845	.012014
un	.042718	.0117609	3.63	0.001	.0190303	.0664058
urbanp	.0014537	.000757	1.92	0.061	-.000071	.0029784
labpar	.0093673	.0031247	3.00	0.004	.0030738	.0156607
loglabf	-.0315792	.022121	-1.43	0.160	-.0761332	.0129749
_cons	3.806824	.2552162	14.92	0.000	3.292792	4.320856

```
. regress loggdpp educ un labpar
```

Source	SS	df	MS	Number of obs	=	51
				F(3, 47)	=	39.81
Model	.430481883	3	.143493961	Prob > F	=	0.0000
Residual	.169409513	47	.003604458	R-squared	=	0.7176
				Adj R-squared	=	0.6996
Total	.599891396	50	.011997828	Root MSE	=	.06004

loggdpp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.0095253	.0016768	5.68	0.000	.006152	.0128986
un	.0497983	.0114319	4.36	0.000	.0268004	.0727963
labpar	.0107416	.0030555	3.52	0.001	.0045946	.0168885
_cons	3.564069	.186281	19.13	0.000	3.18932	3.938818

	loggdpp	educ	un	urbanp	labpar	loglabf
loggdpp	1.0000					
educ	0.7612	1.0000				
un	0.1408	-0.1436	1.0000			
urbanp	0.5324	0.4997	0.1557	1.0000		
labpar	0.6015	0.6321	-0.3835	0.2613	1.0000	
loglabf	-0.0544	0.0469	0.0588	0.4299	-0.1631	1.0000

. sum

Variable	Obs	Mean	Std. Dev.	Min	Max
loggdpp	51	4.737906	.1095346	4.544254	5.251497
educ	51	32.62549	6.597661	21.1	59.7
un	51	3.578431	.8119886	2.3	5.5
urbanp	51	73.82353	15.09472	38.7	100
labpar	51	63.75882	3.879803	55.1	71.5
loglabf	51	6.293534	.446053	5.471292	7.286771