An Empirical Studies in Income Per Capita across States

Zhenzhen Ge and Linfeng He

December 22, 2016

Abstract

As politicians arguing over what makes America great again, we take an empirical study on the state level and explore the correlation between several factors. We conduct diagnostics on several linear regression assumptions and run stepwise model selection to decide the final model structure. Eventually we arrived to the conclusion that college graduation rate played a leading role in determining the income per capita in each state. Therefore, we strongly recommend to increase spending on educations.

1 Introduction

The income-per-capita is one of the most important factors to evaluate the overall strength of the economy of a region. To boost the income-per-capita, the politicians shall make efficient policies that have the direct impacts on the most related factors with the income-per-capita. However, since the income-per-capita is related to many socio-economical factors, such as population and education, it is not clear which factors impact the most on the income-per-capita. In our project, we exam the relationship between the personal income and various other socio-economical factors using public accessible socio-economical dataset and statistical tools. The sources of information included, but not limited to, various government agencies such as FBI, United Census Bureau, National Center for Education Statistics, etc. Our study suggests that the bachelor graduation rates have quite substantial amount of explanatory powers over the income per capital among states. We believed our findings thus shall benefit the upper level policy-makers in decision making which would eventually result in a greater level of our citizens' welfare.

2 Data Description

As it has been iterated, our study would be centered on the socio-economical factors of each state. Thus, numbers like the population of every state, would play a significant role in our study. We have collected and selected eight variables which we believe which would be important to achieving our desirable outcomes as following:

1. Pop: The population of the state[1]
2. Percap.Consum: Personal Consumption Expenditures by state, 2015 [2]
3. LifeExp: Average life expectancy from birth (all values expressed in years) [3]
4. HS.Rate: High school graduation rate, latest [4]
5. BA.Rate: Bachelor school graduation rate, latest [5]
6. Area: Land area of the state [6]

[1] Wikipedia: https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_population
[2] Bureau of Economic Analysis: http://www.bea.gov/newsreleases/regional/pce/pce_newsrelease.htm
[3] Institute for Health Metrics and Evaluation: http://www.healthdata.org/us-health/data-download
[4] U.S. Department of Education: https://nces.ed.gov/programs/digest/d15/tables/dt15_104.85.asp?current=yes
[5] U.S. Department of Education: https://nces.ed.gov/programs/digest/d15/tables/dt15_104.85.asp?current=yes
[6] Bureau of U.S. Census: https://www.census.gov/geo/reference/geoguide.html

7. Violent.Crime: The violent crime figures include the offenses of murder, rape, robbery and aggravated assault [7]

8. Percap.Income: State Personal Income [8]

3 Preliminary Analysis

When we completed the data gathering, we first take a glance at the overall data, such as the distribution of personal income, bachelor graduation rate, and other variables by states, to get a general understanding of the data.

As we have speculated that, there are inequalities in population (obviously), area, consumption per capital, bachelor degree rate, violent crime rate and income per capital among states. Whereas the life expectancy and high school graduation rate seem to be more constant (since they are quite substantial already).

Moreover, before we start model fitting, we split all states into west and east groups and conduct permutation test. No difference of distribution is detected, then we proceed with model building.

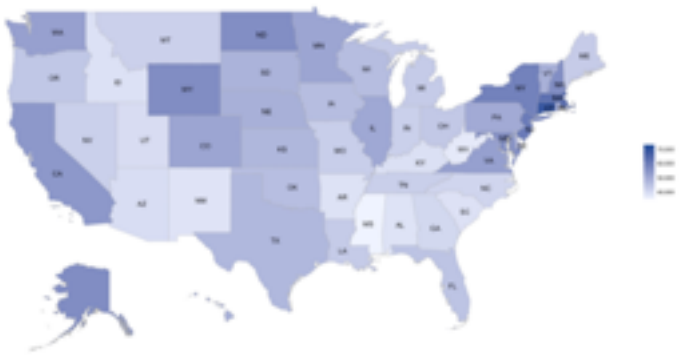


Figure 1: **Per Capita Income in the USA.**

4 Methodologies

4.1 Model Fitting and Hypothesis Test

Using Lease Squares Method, we fit models with per capita income as the response variable, and all others as predictors(1). By performing hypothesis test on all predictors and evaluating p-values, we can see two variables(bachelor graduation rate and personal consumption) are significant at level 0.5.

$$\begin{aligned} Percap.Income = \beta_0 + \beta_1 * Pop + \beta_2 * Percap.Consum + \beta_3 * LifeExp \\ + \beta_4 * HS.Rate + \beta_5 * BA.Rate + \beta_6 * Area + \beta_7 * Violent.Crime \end{aligned} \tag{1}$$

[7] FBI Uniform Crime Report: https://ucr.fbi.gov/crime-in-the-u.s/2015/crime-in-the-u.s.-2015/tables/table-4

[8] Bureau of Economic Analysis: http://www.bea.gov/newsreleases/regional/spi/sqpi_newsrelease.htm

4.2 Diagnostics

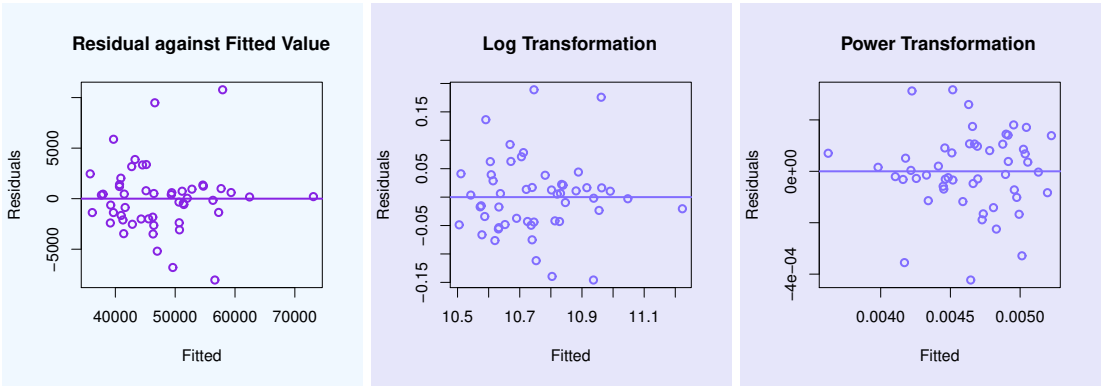
We want to check whether our assumptions of this linear model are correct. So we run series of tests to verify.

4.2.1 Constant Variance

We plot residuals of per capital income against fitted value of model (1) as well as other variables to check whether constant variance holds under current model. The plot shows certain trend in Fig. 2 (a). So we use both power transformation(2) and log transformation(3) to see whether we can eliminate the trend. After the transformation, we get improved plots which have more constant variances, however the improvement is minor.

$$\begin{aligned} Percap.Income^{-1/2} = \beta_0 + \beta_1 * Pop + \beta_2 * Percap.Consum + \beta_3 * LifeExp \\ + \beta_4 * HS.Rate + \beta_5 * BA.Rate + \beta_6 * Area + \beta_7 * Violent.Crime \end{aligned} \tag{2}$$

$$\begin{aligned} \log(Percap.Income) = \beta_0 + \beta_1 * Pop + \beta_2 * Percap.Consum + \beta_3 * LifeExp \\ + \beta_4 * HS.Rate + \beta_5 * BA.Rate + \beta_6 * Area + \beta_7 * Violent.Crime \end{aligned} \tag{3}$$



(a) (b) (c)

Figure 2: **Residual against Fitted Value** (a) Model(1). (b) Model(2). (c) Model(3).

4.2.2 Normality

We continue to check the three models(before and after transformation) whether normality holds. Using the tool of QQ plot, which is residuals of our fitted models against theoretical quantiles of normal distribution, we can see the residuals may not follow normal distribution very well, but have slight heavy tail issues even after model transformations.

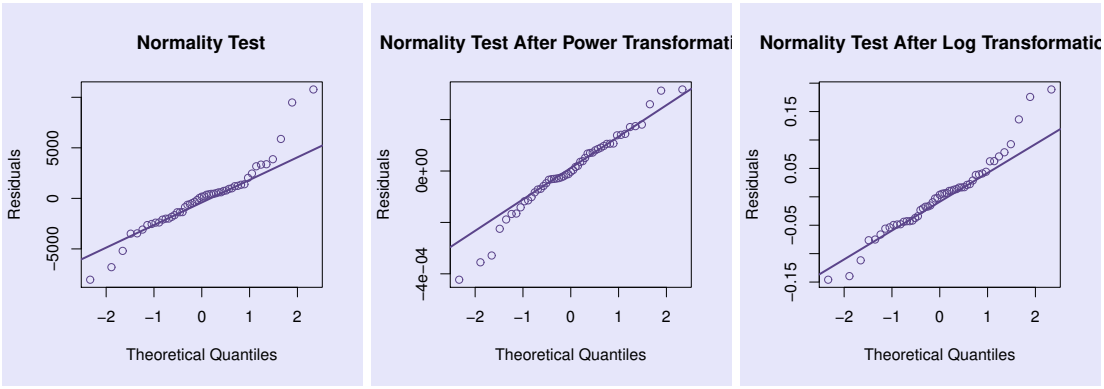
4.2.3 Outliers

As it happens often that heavy tails might result in outliers issues. We look at possible outliers by plotting their cook's distance. Despite of some dots which might seem far up, the values are all smaller than 1, therefore we may assume no outliers among the states.

4.3 Robust Regression

We may still have heavy tail issues in our data, we attempt to use robust method to improve our model. First we try the Huber's estimator on model (1). After the model fitting and hypothesis test of predictors, per capita consumption and bachelor graduation rate(Percap.Consum and BA.Rate) are significant according to this method. It's similar to what we have in original linear model (1).

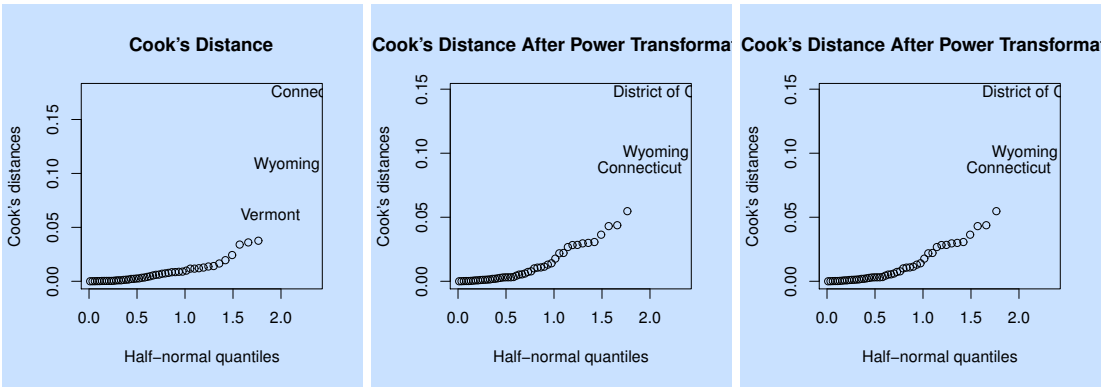
Then we try Huber's estimator on model (2) and (3), and we get almost same results: Percap.Consum and BA.Rate are significant under the method.



(a) (b) (c)

Figure 3: **Normality Test.** (a) Model(1). (b) Model(2). (c) Model(3).

In fact we also try the Least Trimmed Squares. However, this method will throw away a few data to fit the regression model. We get a somehow different result from this regression method, where besides per capita consumption and bachelor graduation rate, life expectancy is also significant. Since our data is not large, we would not proceed with this model.



(a) (b) (c)

Figure 4: **Cook's Distances.** (a) Model(1). (b) Model(2). (c) Model(3).

4.4 Model Selection

4.4.1 Step One

Robust regression doesn't give us more meaningful results compared to least square method, so we will implement model selection on the previous 3 models using Information Criterion method. We perform stepwise selection on model (1), and get below result:

$$Percap.Income = \beta_0 + \beta_1 * Pop + \beta_2 * Percap.Consum + \beta_3 * BA.Rate \tag{4}$$

Also on the model (2) with power transformation :

$$Percap.Income^{-1/2} = \beta_0 + \beta_1 * Pop + \beta_2 * Percap.Consum + \beta_3 * BA.Rate \tag{5}$$

And on the model (3) with log transformation :

$$\log(Percap.Income) = \beta_0 + \beta_1 * Pop + \beta_2 * Percap.Consum + \beta_3 * BA.Rate \tag{6}$$

4.4.2 Step Two

AIC gives quite similar results of model selection on the three models. Seems population, per capita consumption, and bachelor graduation rate are three critical factors of individual income. However, we know consumption is more of the outcome of income, not the other way round. Therefore, we decide to remove consumption and run model selection again.

Without any transformation, we end up with below model after removing consumption manually:

$$Percap.Income = \beta_0 + \beta_1 * BA.Rate + \beta_2 * Area \tag{7}$$

If we take power or log transformation and remove consumption, we have below model:

$$Percap.Income^{-1/2} = \beta_0 + \beta_1 * HS.Rate + \beta_2 * BA.Rate + \beta_3 * Violent.Crime \tag{8}$$

$$\log(Percap.Income) = \beta_0 + \beta_1 * HS.Rate + \beta_2 * BA.Rate + \beta_3 * Violent.Crime \tag{9}$$

Now we can see Area and Violent.Crime coming into the picture as significant predictors.

4.4.3 Step Three

Since Area and Violent.Crime may have positive relations with Pop, and in our model (4), (5), (6) we all have Pop as important predictor, we want to eliminate the effect of population by scaling the original Area and Violent.Crime using Pop as denominator.

After we apply the scaled data, without transformation, we have below models after stepwise selection:

$$Percap.Income = \beta_0 + \beta_1 * BA.Rate + \beta_2 * Area \tag{10}$$

If we apply the scaled data and also take power or log transformation, we will have:

$$Percap.Income^{-1/2} = \beta_0 + \beta_1 * Pop + \beta_1 * HS.Rate + \beta_2 * BA.Rate + \beta_3 * Area \tag{11}$$

$$\log(Percap.Income) = \beta_0 + \beta_1 * Pop + \beta_1 * HS.Rate + \beta_2 * BA.Rate + \beta_3 * Area \tag{12}$$

After scaling Area is still selected. And if we look into the transformed model, we can say population and high school graduation rate also have significant impact on per capita income of the states.

5 Conclusion

We have investigated the relationship between the income-per-capita and various socio-economical factors with various public accessible data set and statistical tools. Our investigation have shown that the income per capita is highly related to the bachelor degree rate of the state residents, as well as high school graduation rate. Furthermore, per capita income level is partially determined by the state's preliminary conditions of population and area. Therefore, in practice, investing in education will boost economy of the state and benefit the personal income. In return, the active economic environment will attract more population to the state and form a beneficial cycle.