The Wealth of States
(ST 352 Final Project)
By Noah Bean

What determines the gross domestic product of a US state? As a student of economics, I am curious about what makes an economy successful and in the time of the COVID-19, preserving the American economy is a high priority. One good measure of the health of an economy is Gross domestic product, defined as "the total monetary or market value of all the finished goods and services produced within a country's borders in a specific time period. As a broad measure of overall domestic production, it functions as a comprehensive scorecard of a given country's economic health" (Investopedia). For this analysis, I selected 8 variables which I believed would affect GDP (from the year 2018 measured in current dollars) for a US state, including the population of the state (a greater number of workers), the percentage of the population living under the poverty line (fewer "productive" workers and greater difficulty in job performance associated with poverty), the unemployment rate (part of population not working), the crime rate per 100,000 people (destruction of the means of production and human capital damage), the percentage of the population holding a Bachelor's degree (more ""productive" workers), the median age of the population (older, more experience workers), and the primary political party affiliation encoded as 1 for democrat or 0 for republican (an interesting mystery to me) (sources under "Sources" section). This data was collected through various online databases and compiled into a single file (sources). The way that the 20 states were sampled was through a simple random sampling procedure (source).

Once, the data was collected, the full model was analyzed for the properties of a linear model:

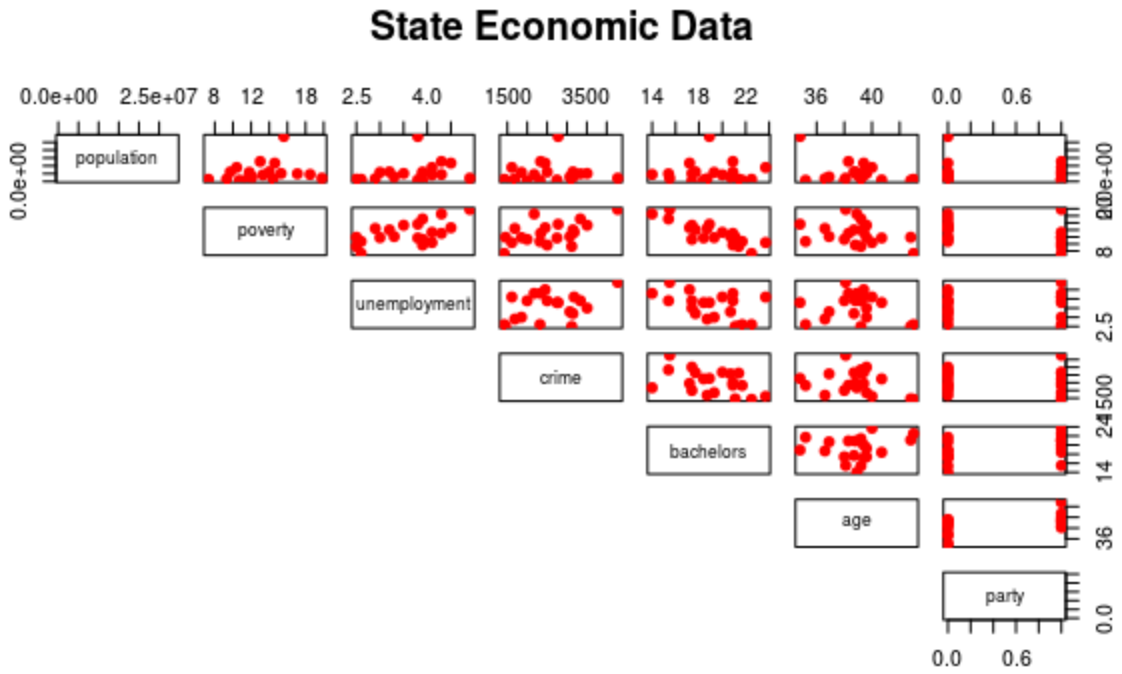
```
lm(formula = gdp ~ population + poverty + unemployment + bachelors +
    crime + age + party, data = complete_state_economic_data)

Residuals:
   Min     1Q Median     3Q    Max
-32692  -5462    -60   6886  23857

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.810e+05  2.008e+05   1.399  0.18713
population       6.403e-02  8.605e-04  74.408  < 2e-16
poverty         -4.592e+03  3.220e+03  -1.426  0.17933
unemployment     2.238e+04  8.917e+03   2.509  0.02743
bachelors_percent 6.230e+03  4.192e+03   1.486  0.16302
```

crime	-1.522e+01	7.447e+00	-2.044	0.06352
age	-1.082e+04	3.198e+03	-3.383	0.00544
party	4.378e+04	1.422e+04	3.080	0.00954

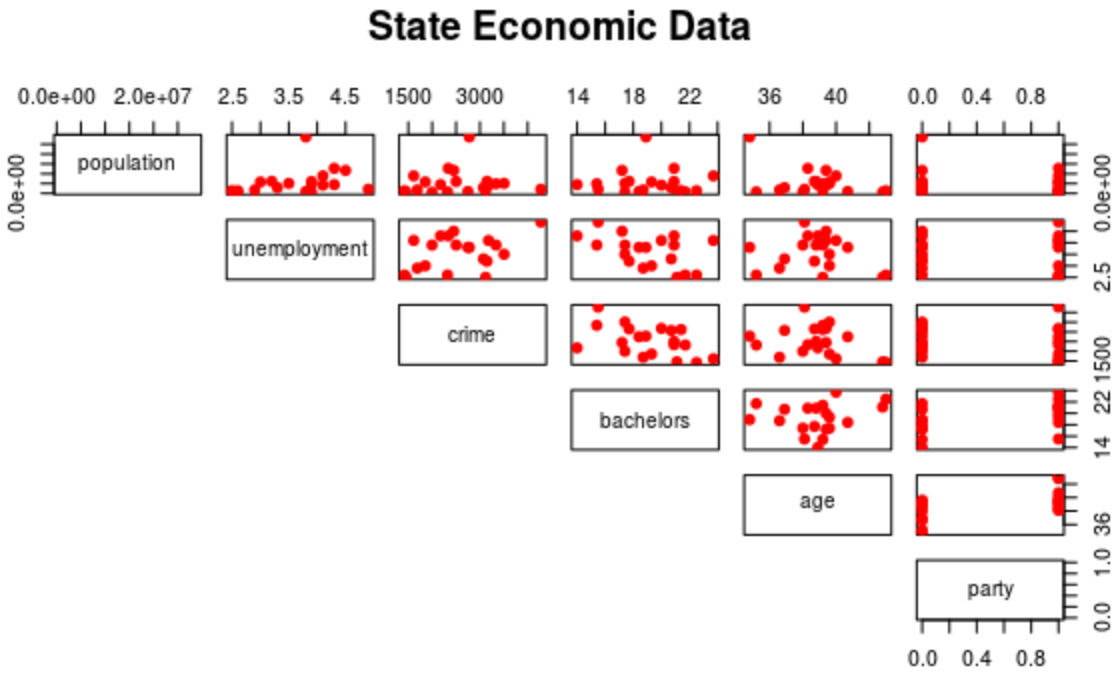
Residual standard error: 19400 on 12 degrees of freedom
Multiple R-squared: 0.9987, Adjusted R-squared: 0.9979
F-statistic: 1296 on 7 and 12 DF, p-value: 2.648e-16

From the printout, the F-statistic indicates that at least one of the variables is significantly correlated with a state's GDP. However, from the individual t-tests, multiple variables appear to not be significant. Therefore, a backwards selection process occurred to reduce the model to the minimum.

From the correlation matrix



It would appear that poverty and crime are highly correlated, and since poverty has a high p-value (0.17933)from its t-statistic, it was removed from the model. The resulting model showed that none of the variables were correlated.



lm(formula = gdp ~ population + unemployment + crime + bachelors_percent + age + party, data = complete_state_economic_data)

Residuals:

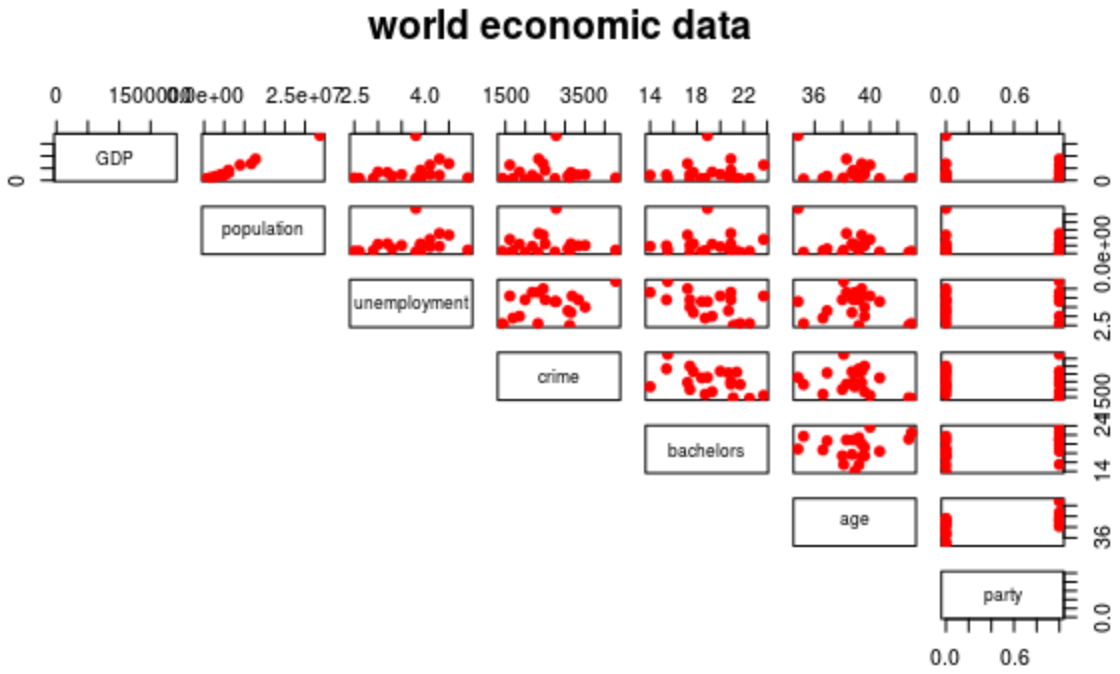
Min	1Q	Median	3Q	Max
-31280	-5118	-501	9932	32503

Coefficients:

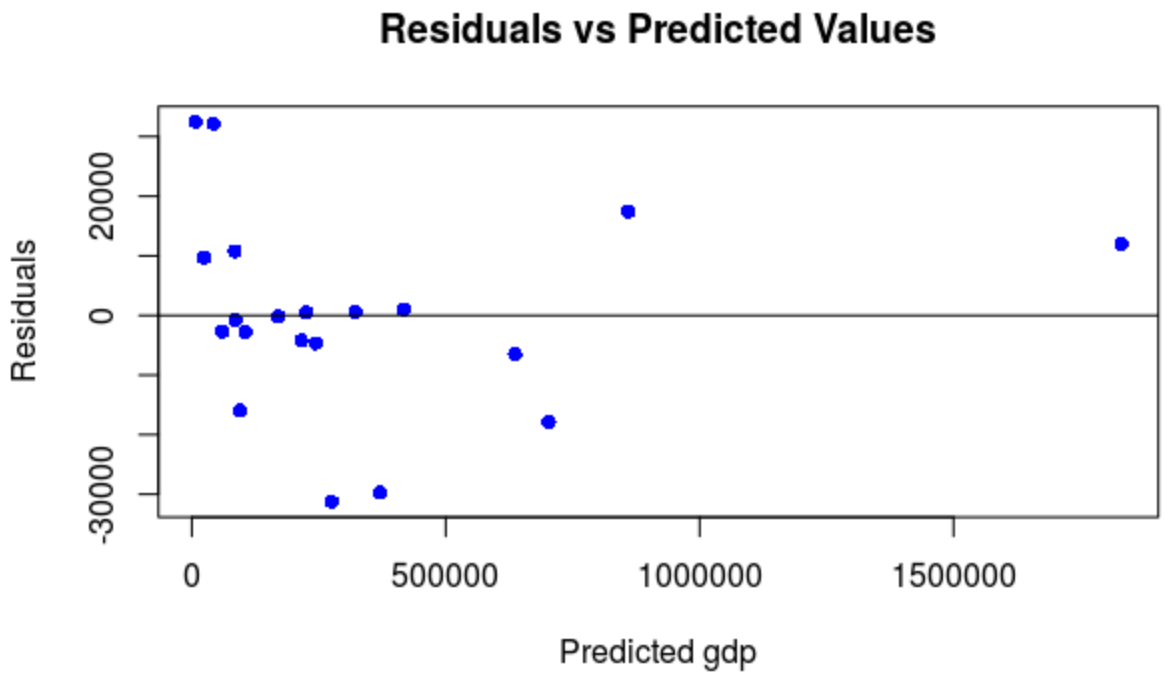
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.206e+05	1.729e+05	0.698	0.49764
population	6.369e-02	8.598e-04	74.077	< 2e-16
unemployment	2.035e+04	9.146e+03	2.225	0.04442
crime	-1.742e+01	7.571e+00	-2.301	0.03862
bachelors_percent	1.051e+04	3.039e+03	3.459	0.00423
age	-9.943e+03	3.261e+03	-3.049	0.00931
party	4.169e+04	1.469e+04	2.838	0.01397

Residual standard error: 20160 on 13 degrees of freedom
Multiple R-squared: 0.9985, Adjusted R-squared: 0.9977
F-statistic: 1400 on 6 and 13 DF, p-value: < 2.2e-16

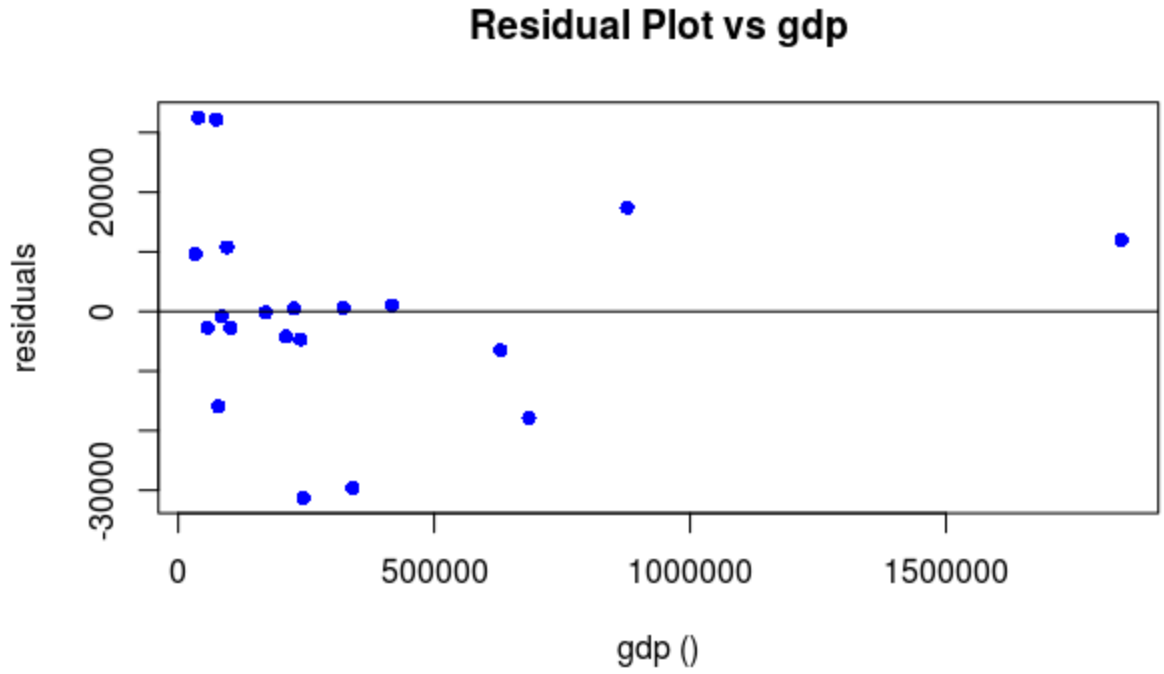
The model was then tested for the conditions of linearity:

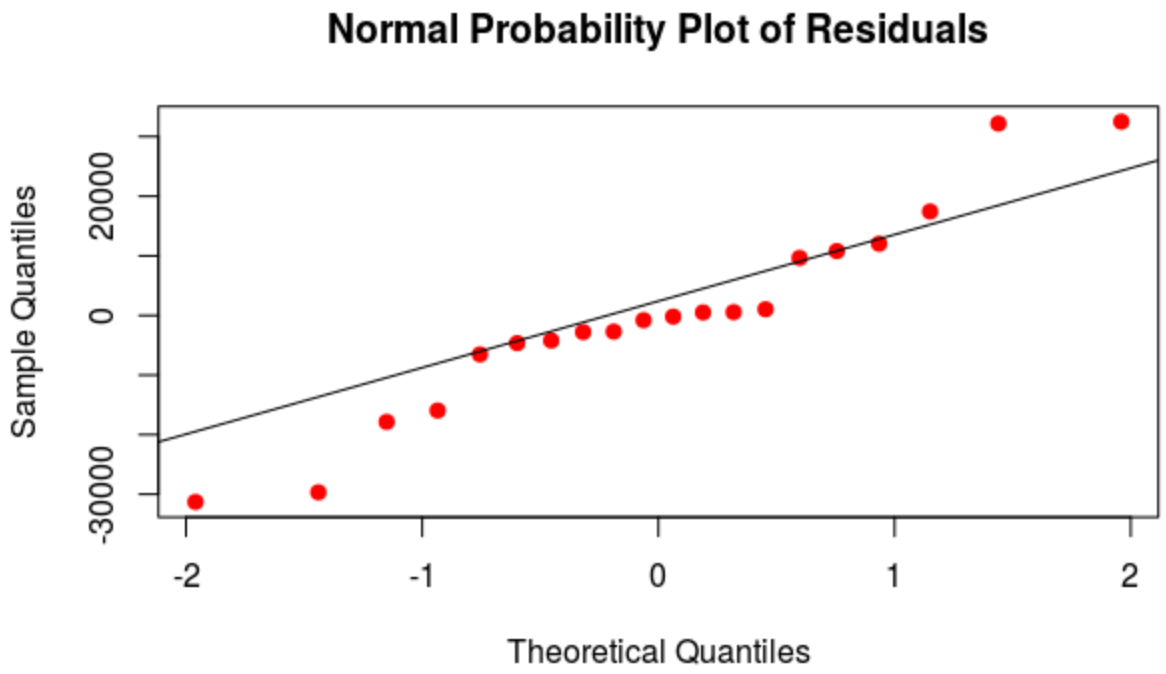


With the exception of population, none of the variables appears to have a strong correlation with GDP, but there are no obvious non-linear relationships.



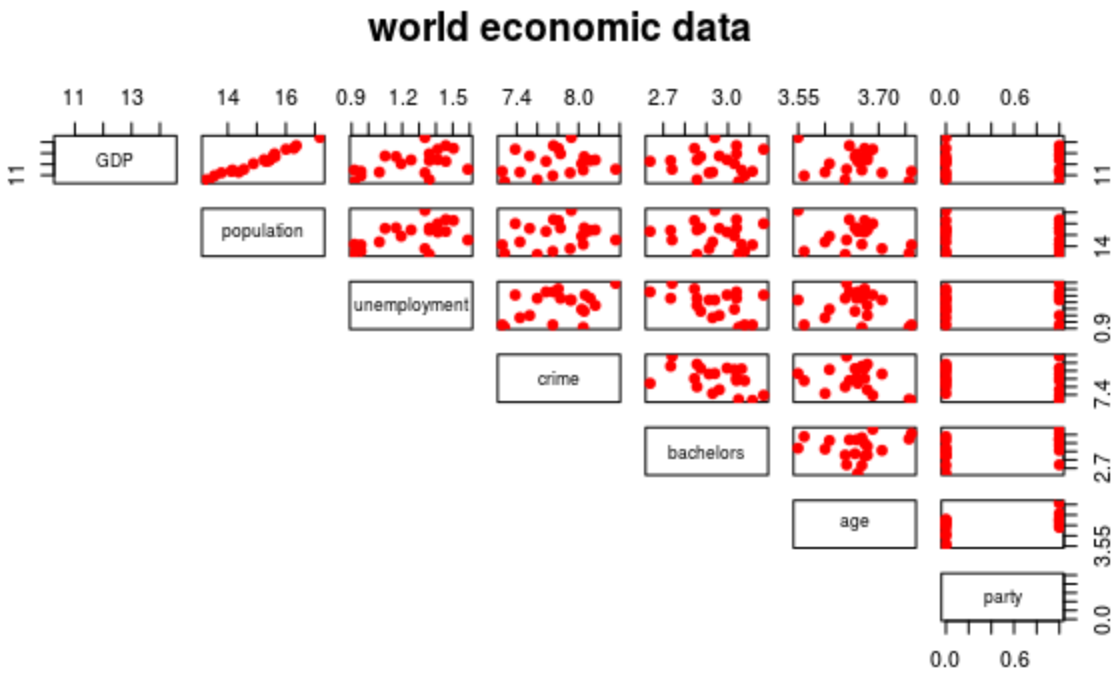
The residuals are approximately varied above and below the line with no noticeable patterns.



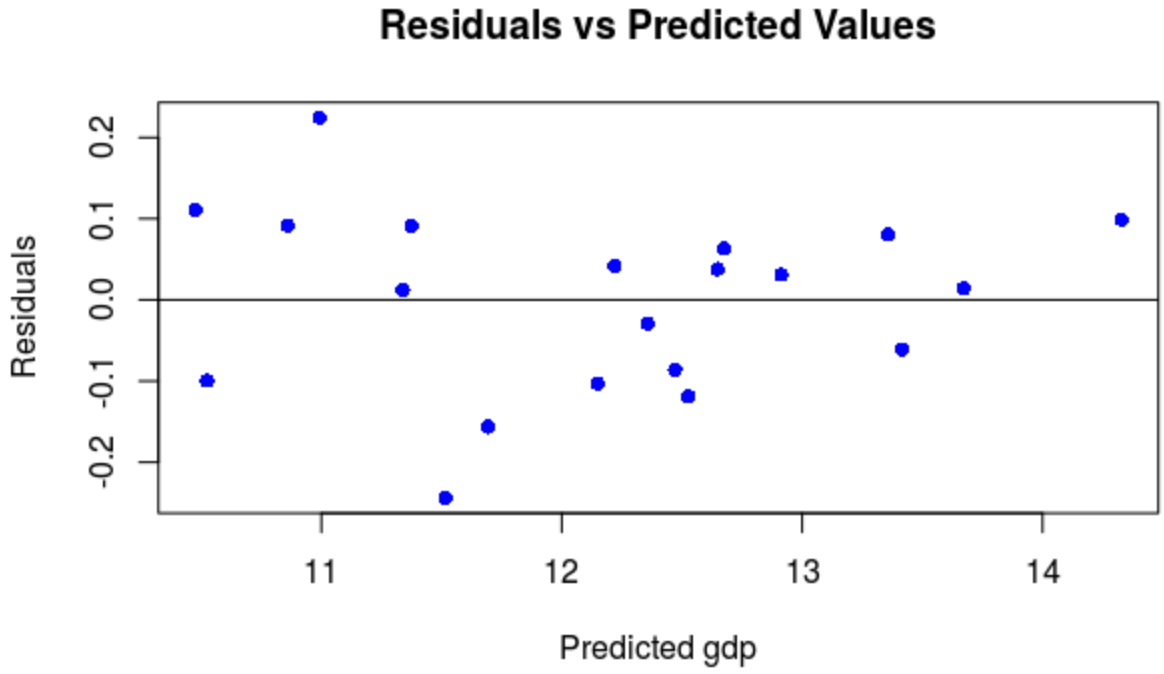


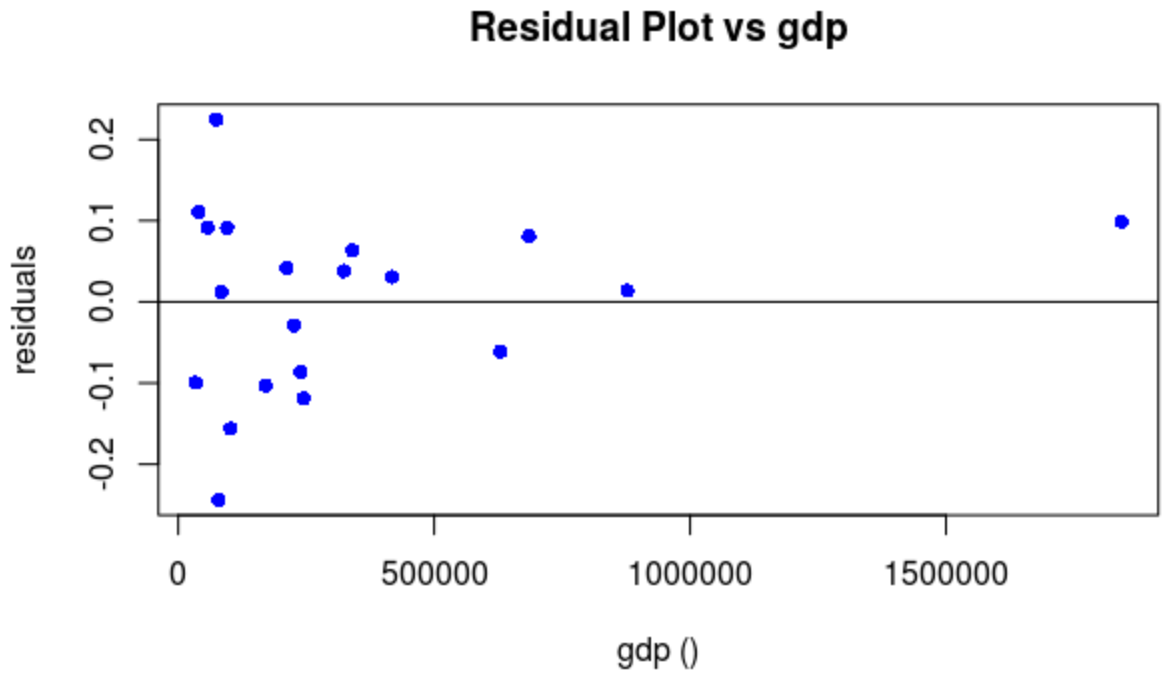
The normal probability distribution of the residuals was unsatisfactory since they appear to not be linear.

To see if a transformation would improve the results, first, the explanatory variable was transformed with a natural logarithm, then, the explanatory variables were transformed, but these results were unsatisfactory due to high non-linearity in the scatter plot. With a transformation on both the response and the explanatory variable, the results were:

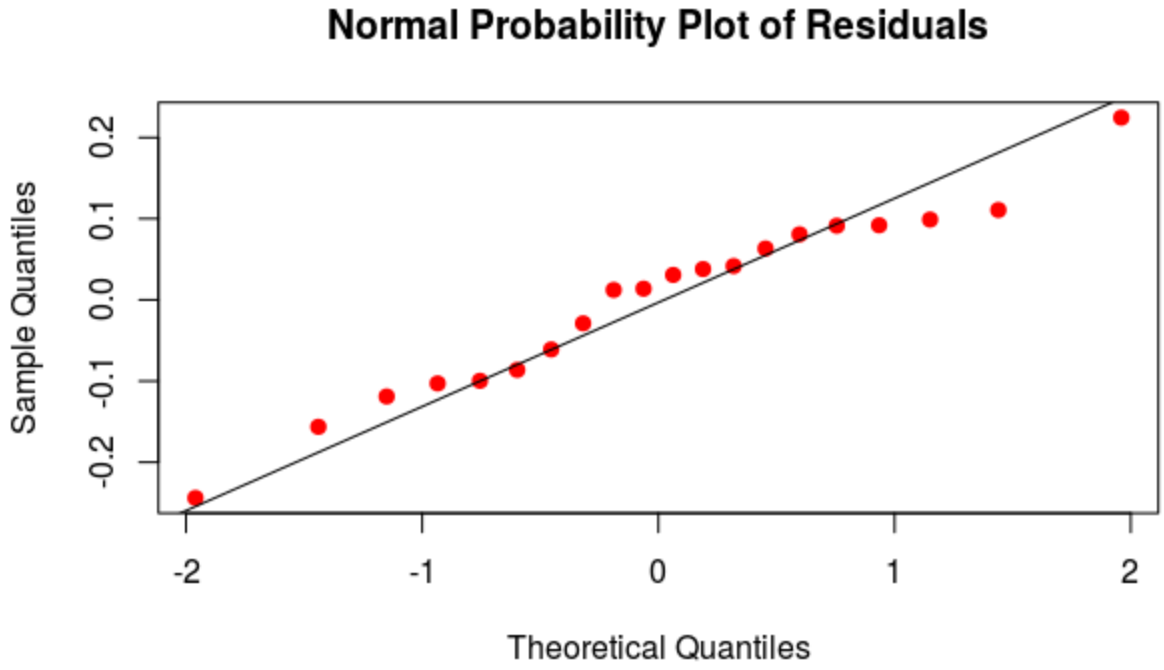


The relationship between the explanatory variables and the response variable is now much stronger and more linear.





The residuals are in a “funnel” shape, which is concerning, but the result is arguably better in variance than the model without transformation.



The normal probability plot looks better, but is still not completely linear.

Thus, it would appear that the transformation on all the variables increases the reliability of the model, albeit complicating the model as well.

For the analysis of the data, both the original model and the fully transformed model will be presented because of a significant effect on the t-tests for the variables. The F-test was performed to see if at least one of the explanatory variables had an effect on the response variable, thus the null hypothesis is H0: $\beta i = 0$ for i = 1- 8 and the alternative hypothesis is $\beta i \neq 0$ for at least some i , where β is the slope of the explanatory variables. The F-test results from the untransformed data were 1400 on 6 and 13 degrees of freedom with a p-value of 2.2e-16 or < 0.0001. The F-test from the transformed data was 201.3 on 6 and 13 degrees of freedom with a p-value of 4.708e-12 or < 0.0001. Thus, there is strong evidence (no matter the model) to indicate that at least one of the variables has a significant effect on the GDP of a state (p-value < 0.0001). However, if a transformation is truly necessary, the new model:

```
lm(formula = log(gdp) ~ log(population) + log(unemployment) +
    log(crime) + log(bachelors_percent) + log(age) + party, data =
    complete_state_economic_data)

Residuals:
      Min       1Q   Median       3Q      Max
-0.24393 -0.08958  0.02228  0.08338  0.22450

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        -3.123538   4.148780  -0.753   0.4649
log(population)     0.951683   0.035664  26.685 9.73e-13
log(unemployment)   0.397396   0.226370   1.756   0.1027
log(crime)         -0.005824   0.124510  -0.047   0.9634
log(bachelors_percent)  1.043514   0.373318   2.795   0.0152
log(age)           -0.688232   0.824415  -0.835   0.4189
party               0.026582   0.096463   0.276   0.7872

Residual standard error: 0.1344 on 13 degrees of freedom
Multiple R-squared:  0.9894,  Adjusted R-squared:  0.9844
F-statistic: 201.3 on 6 and 13 DF,  p-value: 4.708e-12
```

has t-tests that invalidate the statistical significance of the variables unemployment, crime, age, and party, meaning that only the population and percentage of the population with Bachelor's degrees have an effect on the GDP.

Thus, with the model constructed, the least-squares regression equation for the original model is $\hat{y}$ = 1.206e+05 + 6.403e-02*x1 + -4.592e+03*x2 + -1.742e+01*x3 + 1.051e+04*x4 + -9.943e+03x5 + 4.378e+04x6 and the transformed model is ln($\hat{y}$) = -3.123538 + 0.951683*x1 + 0.397396*x2 + -0.005824*x3 + 1.043514*x4 + -0.688232*x5 + 0.026582*x6 with all of the explanatory variables transformed with a natural log. In this model, x1 is the population, x2 is the unemployment rate, x3 is the crime rate per 100,000 people, x4 is the percentage of the population holding Bachelor's degrees, x5 is the median age of the population, and x6 is the primary political party affiliation coded as 1 for Democrat or 0 for republican so a blue state has 4.378e+04 more GDP, on average, than a red state. No control variable was used for the categorical variable of party affiliation. From the original model, the most significant variables were the unemployment rate and the percentage of the population who held Bachelor's degrees. In the transformed model, none of the variables were significant except for the population number and the percentage of the population who held Bachelor's degrees (both positively correlated).

In conclusion, there appears to be multiple variables that have an effect on the GDP of a state, so state governments, as well as the Federal government, will need to cater to these factors in order to restart the United States economy as quickly as possible, and also strengthen it in the long term. A surprising discovery from this study was that political party affiliation appears to have an effect on a state's GDP. This effect could be fleshed out in further research by quantifying the degree to which a state is polarized. Another discovery, not quite as surprising, was that poverty and crime are highly correlated and this could be explored in further research to reduce crime rates.

Sources

GDP by state in USD for the 4th quarter of 2018
https://apps.bea.gov/itable/iTable.cfm?ReqID=70&step=1#reqid=70&step=1&isuri=1

Population by state in 2018
https://worldpopulationreview.com/states/

Percentage of state population with a Bachelor's degree in 2018
Percentage of state population living below the poverty line in 2018
https://www.chronicle.com/interactives/almanac-2018

unemployment rate of states in 2018
http://www.dlt.ri.gov/lmi/laus/us/annavg.htm

Number of crimes committed per 100,000 people in states during 2018
https://www.statista.com/statistics/301549/us-crimes-committed-state/

Party affiliation of states in 2017 (highest percentage chose as the state affiliation)
https://news.gallup.com/poll/226643/2017-party-affiliation-state.aspx

Median age of states in 2018
https://www.considerable.com/entertainment/trivia/median-age-us-states

Definition of Gross Domestic Product
https://www.investopedia.com/terms/g/gdp.asp

Tool for simple random sample of states
https://www.randomlists.com/random-us-states?dup=false&qty=20

Data set

state	population	gdp	poverty	unemployment	crime	bachelors	age	party
Alabama	4887871	225859.4	17.1	3.9	3336.81	15.4	39.2	0
Delaware	967171	74347.1	11.7	3.8	2748.01	18.4	40.7	1
Idaho	1754208	78707.3	14.4	2.9	1688.45	18.7	36.6	0
Illinois	12741080	877010.4	13	4.3	2336.96	20.9	38.3	1
Kansas	2911505	170242.9	12.1	3.3	3072.91	20.7	36.9	0
Kentucky	4468402	211073.3	18.5	4.3	2174.42	14	38.9	0

Maryland	6042718	417975.6	9.7	3.9	2501.92	20.9	38.8	1
Missouri	6126452	322971.9	14	3.2	3149.15	17.7	38.7	0
New_Hampshire	1356458	85075.4	7.3	2.6	1421.64	22.5	43	1
New_Jersey	8908520	629995.9	10.4	4.1	1612.98	23.7	40	1
New_Mexico	2095428	102370.5	19.8	4.9	4276.26	15.5	38.1	1
North_Dakota	760077	56967.6	10.7	2.6	2320.82	21.7	35.2	0
Ohio	11689442	685299.4	14.6	4.5	2450.78	17.2	39.4	0
Oregon	4190713	244543.6	13.3	4.1	3179.51	20	39.4	1
South_Carolina	5084127	239176.3	15.3	3.5	3505.93	17.4	39.6	0
Texas	28701845	1842447.1	15.6	3.8	2778.07	18.9	34.8	0
Vermont	626299	33669.6	11.9	2.5	1455.06	21.1	42.8	1
Wisconsin	5813568	340853.5	11.8	3	1855.35	19.3	39.6	1
Wyoming	577737	39612.6	11.3	3.9	1997.28	17.4	38	0

Hawaii 1420491 95368.5 9.3 2.5 3118.92 21.4 39.2

*a higher quality .CSV file is available upon request