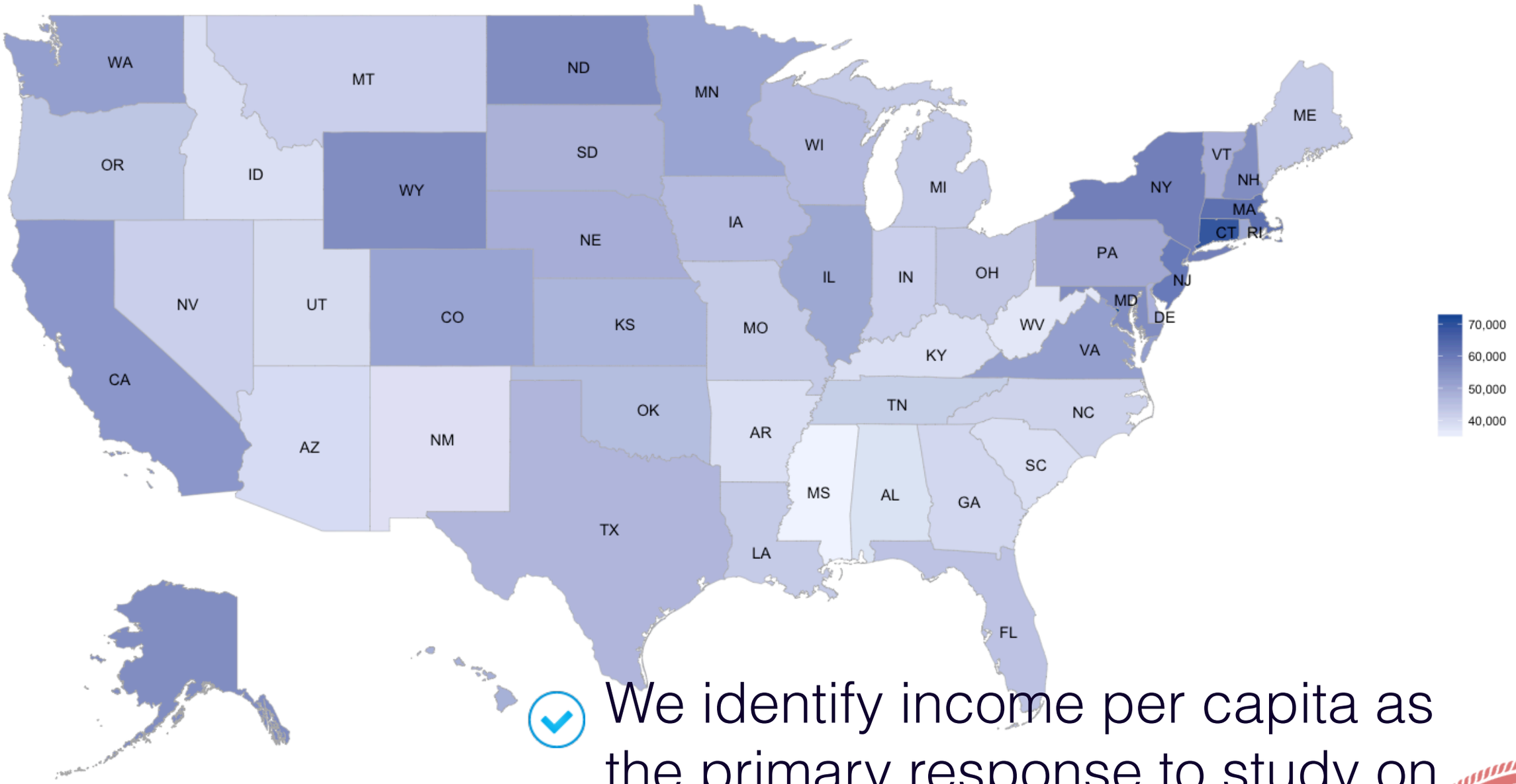
Statistical Analysis for Strategic Planning

What determines the outcome of each states?

Mission statement

Income per capita



We identify income per capita as the primary response to study on

1. Problem Identification

DATA

- Pop: The population of the state
- Percap.Consum: Personal Consumption Expenditures by state
- LifeExp: Average life expectancy from birth
- HS.Rate: High school graduation rate
- BA.Rate: Bachelor school graduation rate
- Area: Land area of the state
- Violent.Crime: The violent crime figures
  include the offenses of murder, rape, robbery and aggravated assault

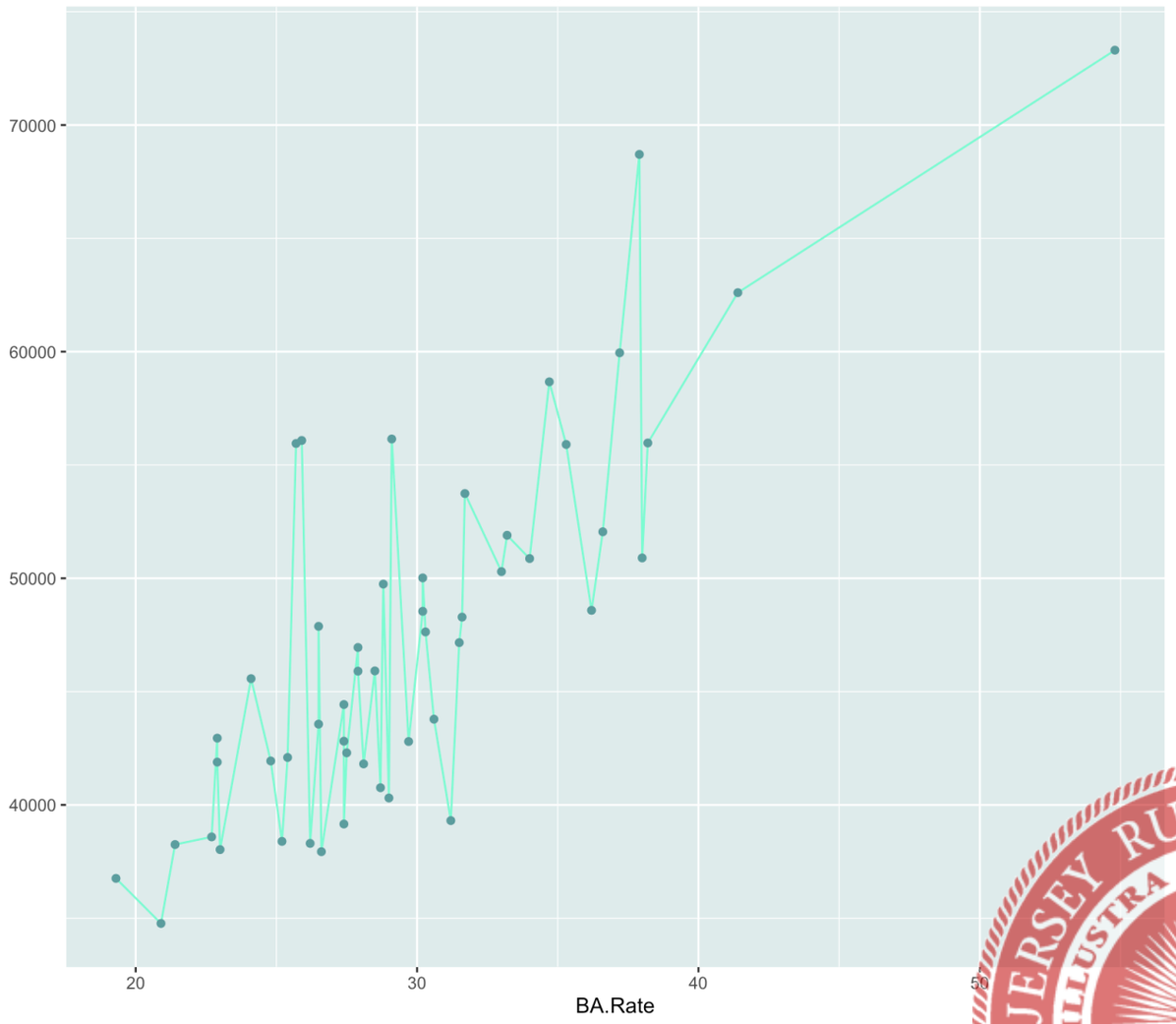
	State	Pop	Percap.Consum	LifeExp	HS.Rate	BA.Rate	Area	Violent.Crime	Percap.Income
1	California	39144818	39715	80.4	82.1	31.7	423967	166883	53740
2	Texas	27469114	35527	78.1	82.3	27.9	695662	113227	46947
3	Florida	20271272	37020	79.0	87.2	27.4	170312	93626	44429
4	New York	19795791	45272	79.9	85.8	34.7	141297	75165	58669
5	Illinois	12859995	39859	78.6	88.4	33.0	149995	49354	50294
6	Pennsylvania	12802503	39498	78.2	89.5	28.8	119280	40339	49744
7	Ohio	11613423	36460	77.4	89.3	26.5	116098	33898	43565
8	Georgia	10214860	33353	76.8	85.6	29.0	153910	38643	40306
9	North Carolina	10042802	32501	77.3	86.4	28.7	139391	34852	40759
10	Michigan	9922576	37775	77.8	89.9	27.4	250487	41231	42812
11	New Jersey	8958013	47256	79.7	89.0	37.2	22591	22879	59949
12	Virginia	8382993	40195	78.5	88.6	36.6	110787	16399	52051

Data Snippet

Linear Regression Result

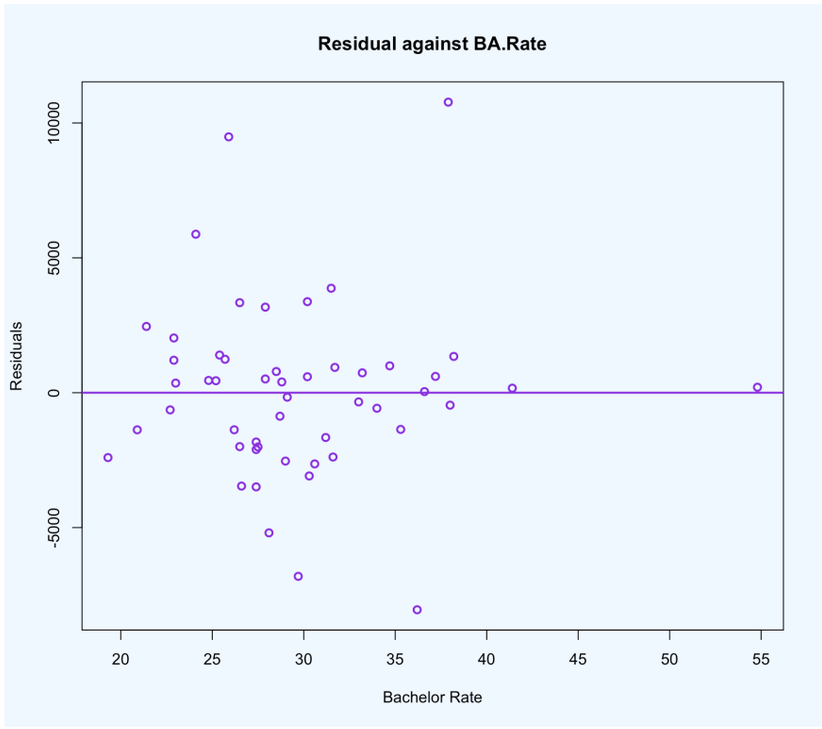
```
Call:
lm(formula = Percap.Income ~ Pop + Percap.Consum + LifeExp +
    HS.Rate + BA.Rate + Area + Violent.Crime, data = data_raw)

Residuals:
    Min      1Q  Median      3Q     Max
-8046.2 -1912.6   168.8  1100.9 10773.0

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.496e+04  3.051e+04   1.146   0.2582
Pop            2.715e-04  3.712e-04   0.731   0.4685
Percap.Consum  1.092e+00  1.650e-01   6.620 4.57e-08 ***
LifeExp       -4.612e+02  4.889e+02  -0.943   0.3507
HS.Rate       -4.320e+01  2.877e+02  -0.150   0.8813
BA.Rate        3.138e+02  1.450e+02   2.164   0.0361 *
Area          -5.923e-04  2.184e-03  -0.271   0.7875
Violent.Crime -3.135e-02  8.929e-02  -0.351   0.7272
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3519 on 43 degrees of freedom
Multiple R-squared:  0.8433,    Adjusted R-squared:  0.8178
F-statistic: 33.05 on 7 and 43 DF,  p-value: 2.592e-15
```

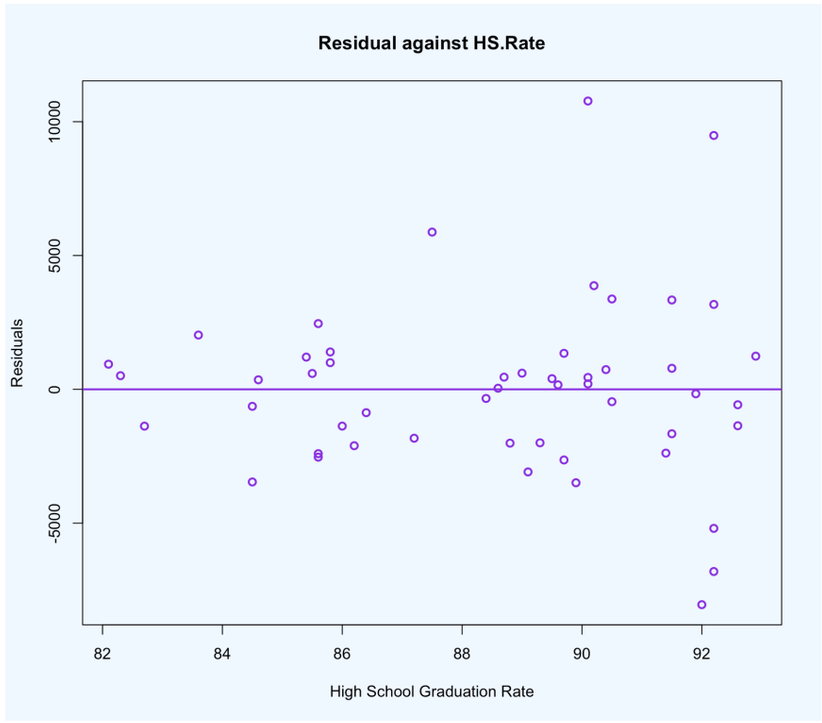
Income against Bachelor Rate

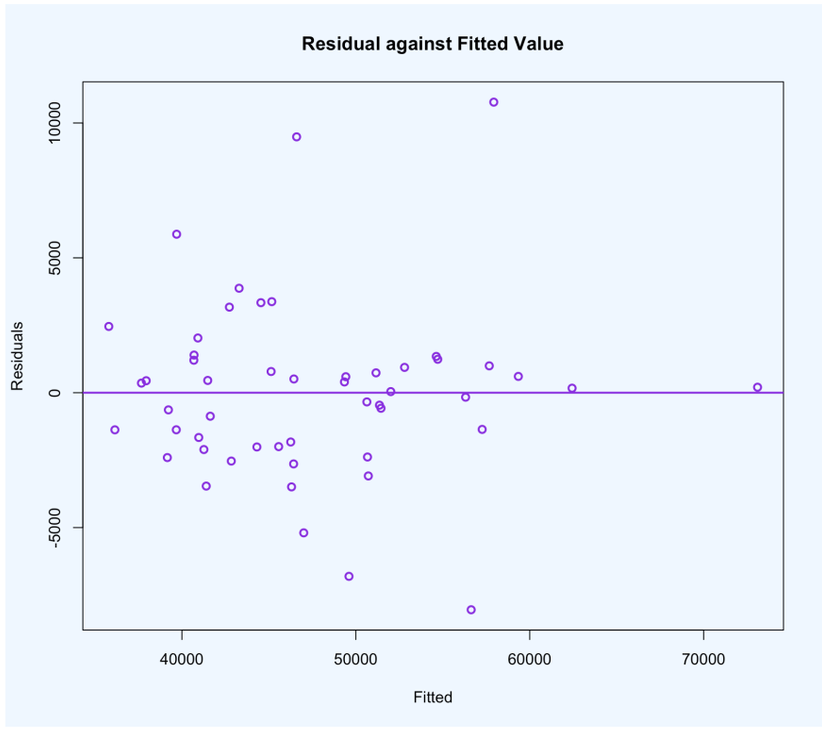


3. Methodology

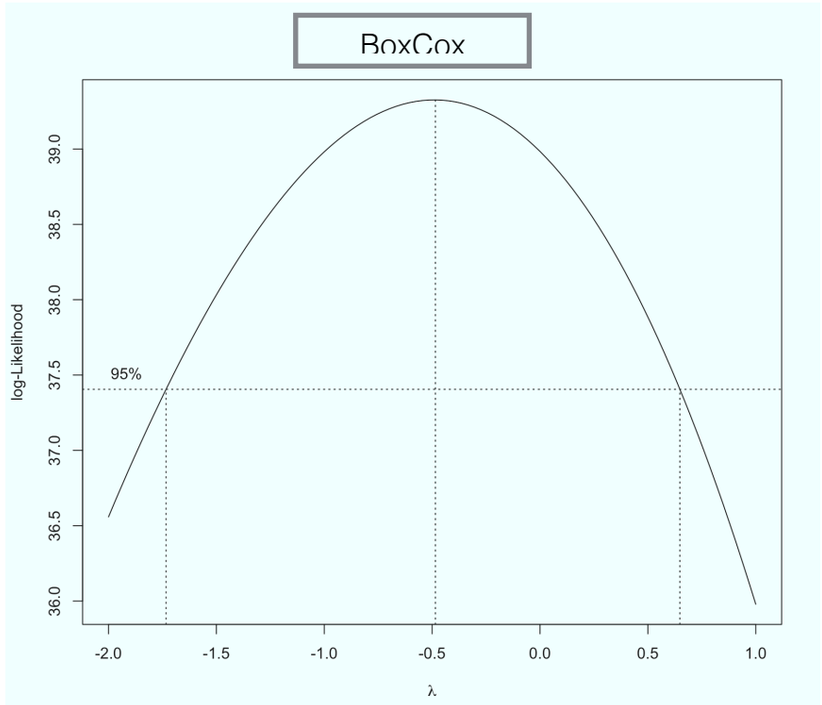
Diagnostics I - Constant Variance

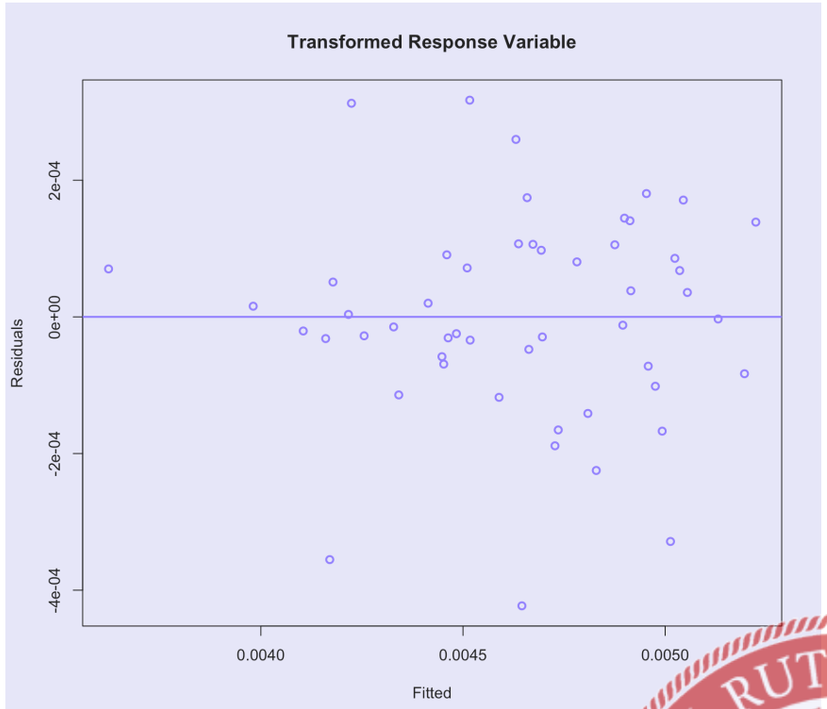






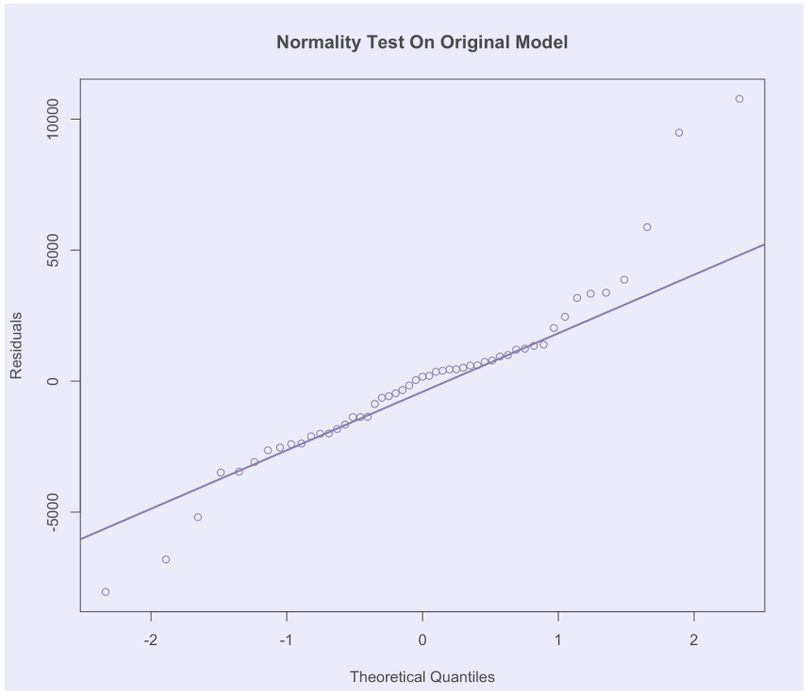
Residual Plots show **non-constant** variance, so we use **BoxCox** to identify the transformation method: we choose **-0.5** as the power of response variable

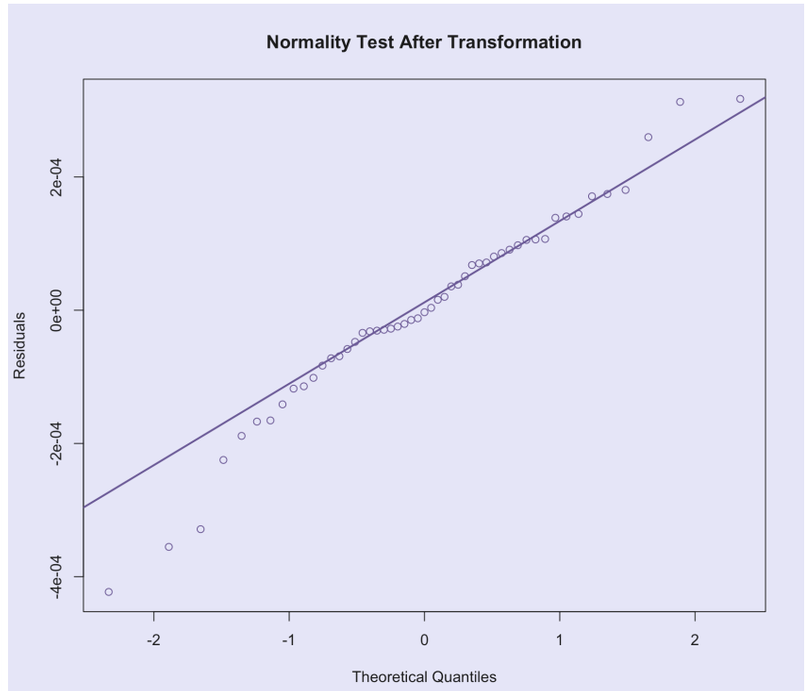




3. Methodology

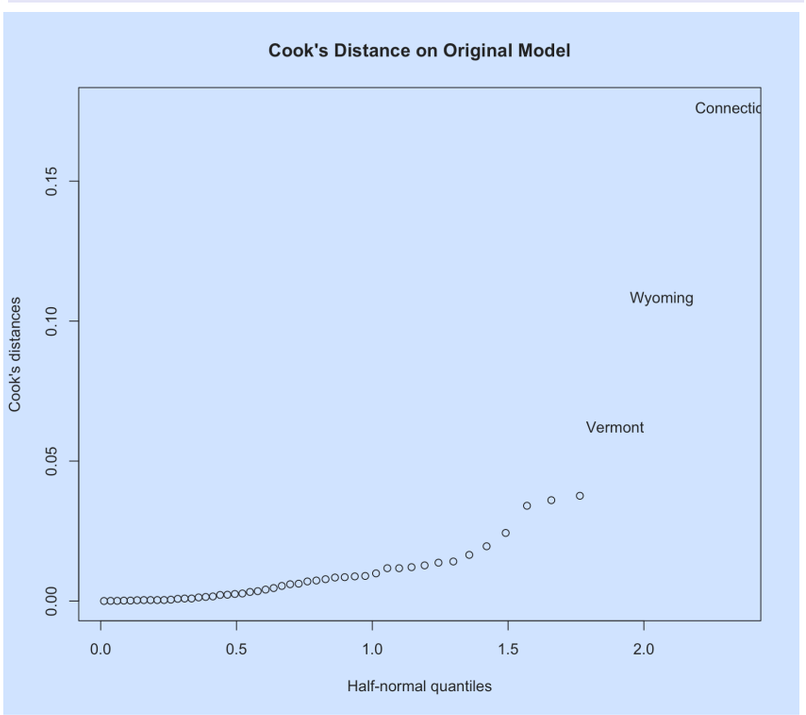
Diagnostics II - Normality & Outliers

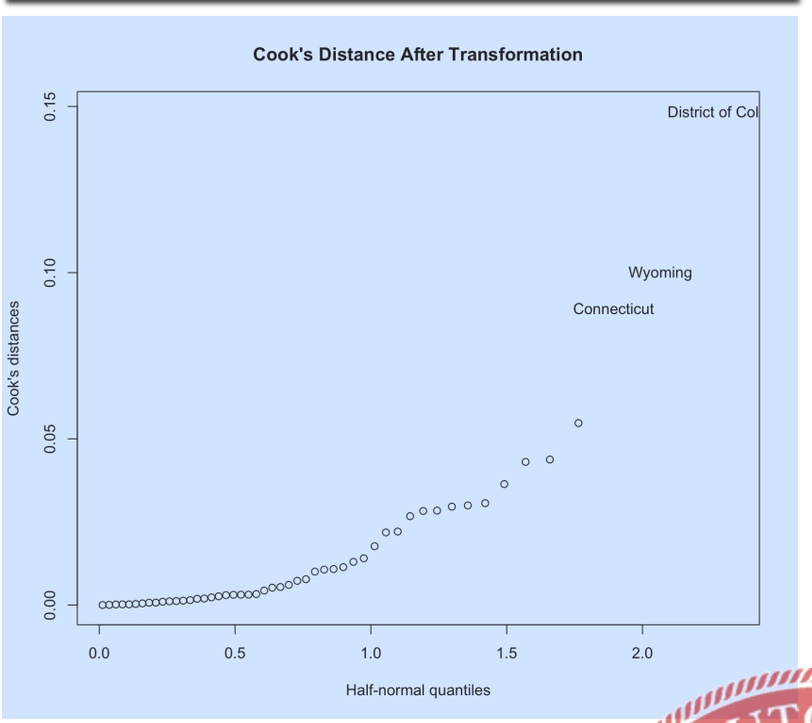




We detect
heavy tail issue
via Normality Test

Heavy tail might be associated with outliers. Fortunately,
NOT in this case

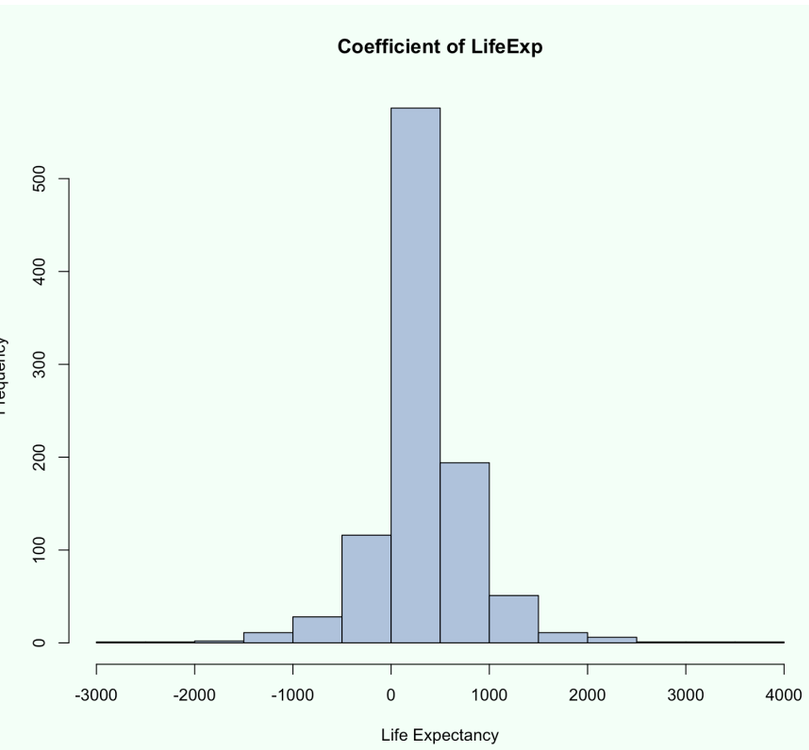


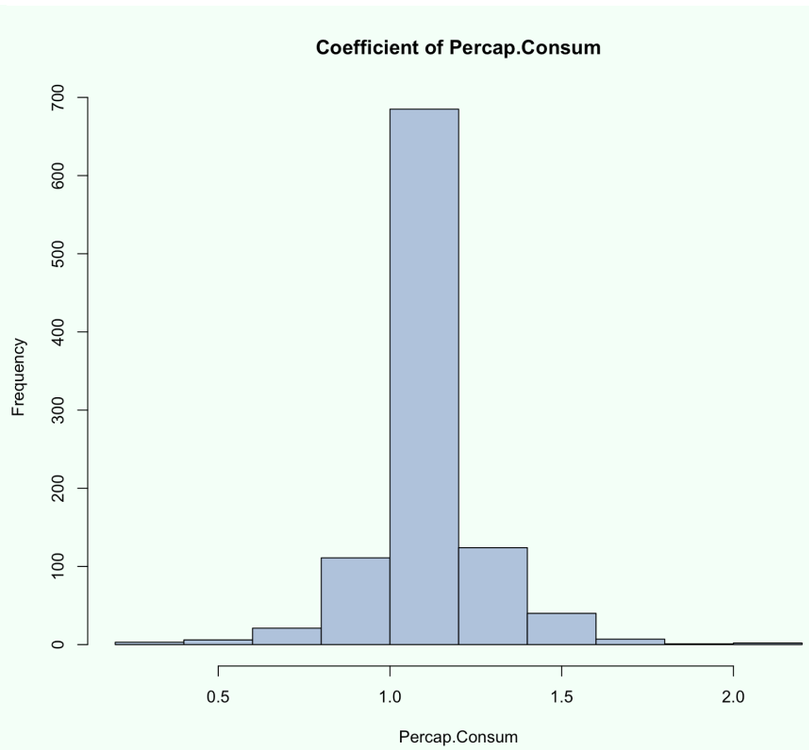


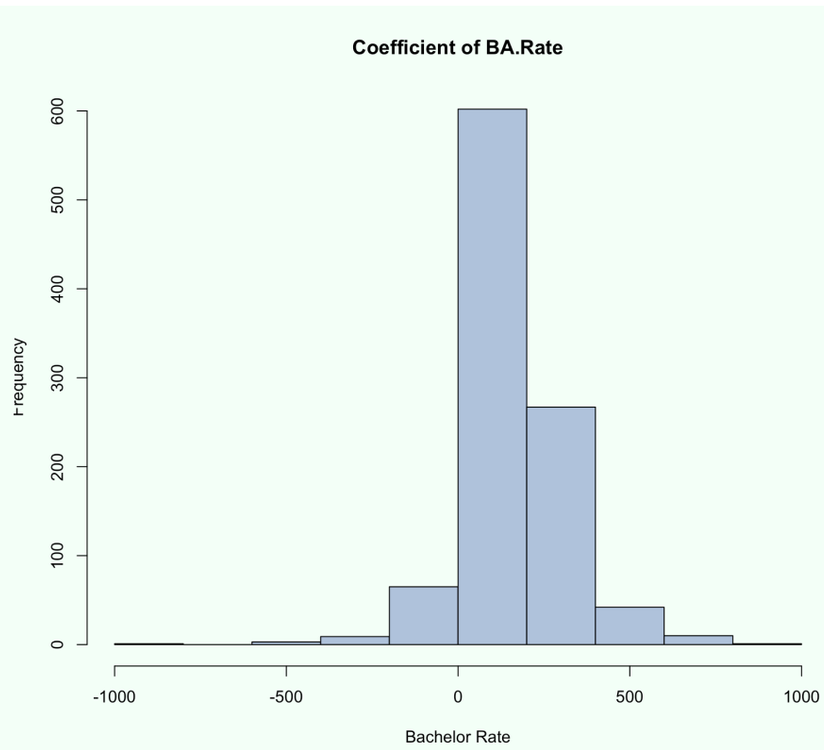
3. Methodology

Re-test Least Trimmed Squares

	(Intercept)	Pop	Percap.Consum	LifeExp	HS.Rate	BA.Rate	Area	Violent.Crime
2.5%	-77255.01	-0.0002331703	0.7851889	-750.0682	-780.6295	-136.8832	-0.009158213	-0.2762241
97.5%	53753.79	0.0012583106	1.4755198	1453.7132	402.4070	497.2017	0.011081916	0.1178079







Using this robust method, we discovered
a few more predictors which are significantly **non-zero**

3. Methodology

Using backward selection or information criterion method, we get same outcome after selection process. Also, the original model and transformed model get same results.

```
Call:
lm(formula = Percap.Income ~ Pop + Percap.Consum + BA.Rate, data = data_raw)

Coefficients:
  (Intercept)            Pop   Percap.Consum        BA.Rate
   -1.687e+03      1.257e-04       1.001e+00      3.251e+02


Call:
lm(formula = Percap.Income ~ BA.Rate + Area, data = data_raw)

Coefficients:
(Intercept)       BA.Rate          Area
  1.347e+04     1.114e+03     4.112e-03
```

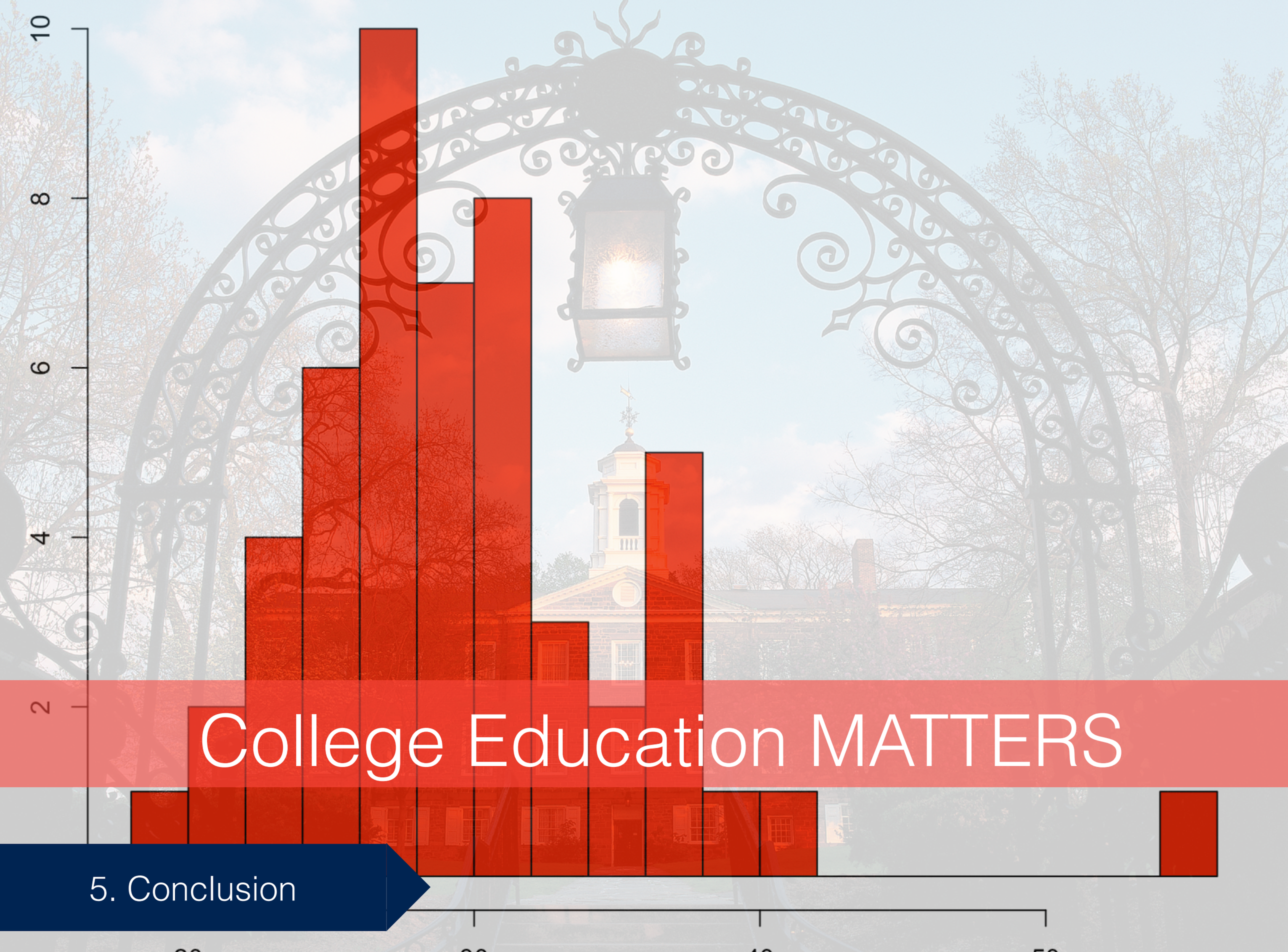
Wait a minute, we are not done yet!
Since we know consumption is highly related to income, and it's probably the outcome, not the cause, we should remove this predictor then select our model.

4. Model Selection

After we delete redundant predictors, we fit the final linear model

```
Call:
lm(formula = Percap.Income ~ BA.Rate + Area, data = data_raw)

Residuals:
    Min      1Q  Median      3Q     Max
-9839.3 -2614.4  -205.6  2875.9 13082.6

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.347e+04  3.528e+03   3.819 0.000385 ***
BA.Rate     1.114e+03  1.126e+02   9.898 3.54e-13 ***
Area        4.112e-03  2.747e-03   1.497 0.140925
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4824 on 48 degrees of freedom
Multiple R-squared:  0.6712,    Adjusted R-squared:  0.6575
F-statistic: 48.99 on 2 and 48 DF,  p-value: 2.553e-12
```

4. Model Selection



College Education MATTERS

5. Conclusion

One last thing...

- Crime Rate and
  High School Graduation Rate
- Life Expectancy and
  Personal Consumption

5. Conclusion

Thanks for Watching

ANY QUESTIONS?