INST 627

Project Report

Analyzing factors associated with the uninsured population in the US

Team:

Marina Cascaes Cardoso

Sanjna Srivatsa

Snigdha Petluru

Point Allocation Agreement: All three students have agreed to divide the points evenly among all members.

1. Introduction

Having a health insurance is strongly advised by the government of the United States as it has multiple benefits. It ensures stability and alleviates financial stress during times of need. People who opt for a health insurance also have better access to healthcare as they can undertake essential medical procedures and not worry about the huge bills that accompany most treatments. People without health insurance tend to reminisce about the costs they will incur and end up avoiding most medical procedures and more often than not, they refrain from gaining access to some ailments that might require immediate attention.

Problem

Our project aims to determine the extent to which various socioeconomic and demographic parameters influence people's decision to not opt for a health insurance in various states within the United States. By analyzing the mentioned problem statement, we can recognize the part of population most deprived of healthcare and infer the main reason for people opting not enrolling for health-insurance. It might be helpful to restructure policies and schemes considering some biases in the design of health-insurance schemes.

Research Question

Can income, education and employment percentages be used to predict the percentage of uninsured people in various states in the U.S.?

In simple terms we wanted to know which of these factors influence the rate of enrollment into health insurance among US population. From preliminary readings and observation of data sets, we would like to explore the relationship between not having a health insured and factors like income, education and employment. A key area of interest lies in identifying trends associated with these factors in various states to ascertain if they play a part in determining if people opt for a health insurance. In this context, the sub-questions that we would like to explore specifically would be:

1. Do states with higher percentage of low income population have a higher percentage of uninsured people?
2. Do states with higher percentage of people with less or no education have a higher percentage of uninsured people?
3. Do states with higher unemployment rates have a higher percentage of uninsured people?

Importance

It is important to identify and address the reasons compelling people against taking an insurance, especially when most State governments aim to provide better health services, be it in terms of accessibility, or enhanced infrastructure to provide for the growing needs of its constituent population. There could be certain segments of the population that can simply not afford the costs associated with health insurance. Often times, lack of awareness about the benefits or the necessity of a health insurance can lead to people not opting for an insurance in the first place. Analyzing these factors would provide valuable insights to both governments and insurance providers about any necessary changes to make to existing norms and prices.

2. Method

Source data

a) Dataset: We retrieved multiple data sets from the open source data website **factfinder.census.gov.** We created an aggregate dataset containing a compilation of state wise statistics derived from these data sets, assimilated and collected by the U.S. Census Bureau.

b) Population: All States within the U.S. Our analysis will be held at a state level.

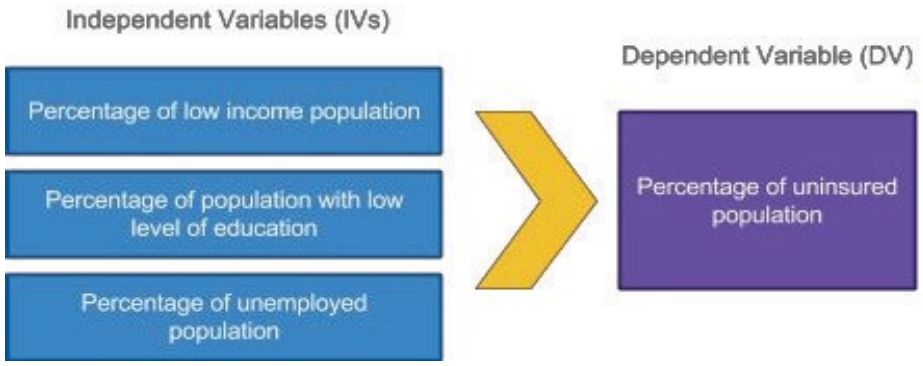
c) Study Type: Our analysis of datasets would be an Observational study.

Design

a) Sample

We have considered the statistics of various states for the year 2013, as the sample for this project. The datasets are obtained from the surveys conducted by the US Government, the description provided in the information file accompanying the data says that the data has been retrieved randomly. The sample in itself, is representative of population estimates and extensive due to the fact that it considers various aspects of the people that it has surveyed, and comprises of data from all states within the US. More importantly, since this data source belongs to the Census of the government, we have assumed that it is thorough in its collection and will have negligible error rates. The analysis is at the State level, where various counts and percentages of population following a certain characteristic are tallied for each state. For this analysis, a non-random sampling procedure was selected and we chose statistics specific to the year 2013 from a pool of statistics for other years. The reason we chose 2013 was because we wanted to know implications on health insurance post ObamaCare.

b) Variables



c) Type of Measure

Independent (Predictor) Variable [IV]	**Type of Measure**
Percentage of low income population	Ratio
Percentage of people with low level of education	Ratio
Percentage of unemployed population	Ratio

Dependent (Outcome) Variable [DV]	**Type of Measure**
Percentage of uninsured population	Ratio

Data Cleaning

In order to answer the research questions, we obtained various datasets belonging to the year 2013. The assembled dataset contains the following variables:

- State – This is used as a primary key to identify each state in the United States
- PERCENT_POVERTY – *Predictor* variable *indicating* percentage of people with low income
- PERCENT_HIGHSCHOOLPLUS - variable indicating percentage of people with greater or equal to high school education
- PERCENT_LESS_HIGH – *Predictor* variable *calculated* by 100 –PERCENT_HIGHSCHOOLPLUS, *indicating* percentage of people with low level of education (i.e. less than high school)
- EST_UNEMPLOYMENT – *Predictor* variable describing unemployment rate, indicating percentage of unemployed population
- TOT_POP – Total population of the state
- UNINSURED_POP – Total uninsured population of the state
- PERCENT_UNINSURED_POP – *Outcome* variable *calculated* by (UNINSURED_POP/TOT_POP)*100, *indicating* percentage of uninsured population

Cleaning of the data included considering only those states present within all three datasets, and states not present in all datasets relating to predictor and outcome variables were removed. A subset of each dataset has been considered, specifically, only those variables that are of interest to our research questions.

Insights into Statistical Approach

a) Test Method

We use multiple regression in order to answer the main research question: *Can income, education and employment percentages be used to predict the percentage of uninsured people in various states in the U.S.?*

Multiple Linear Regression is an ideal choice when we want to perform predictive or causal analysis. This is especially true for the project, where we are trying to determine which of the predictor variables (PERCENT_POVERTY, PERCENT_LESS_HIGH, EST_UNEMPLOYMENT) cause or affect the outcome variable (PERCENT_UNINSURED_POP). Given than we have more than 2 predictor variables- all of which are ratio scaled, and an outcome variable which is also ratio scaled, this method is the best way to obtain desired results that could lead to an equation describing the outcome variable in terms of the predictors.

In order to answer each sub question, simple linear regression test is used for each predictor – outcome pair. This test is required as for each question, we try to define a relationship between the predictor – outcome pair. We are trying to describe the outcome variable (PERCENT_UNINSURED_POP) in terms of each predictor (PERCENT_POVERTY, PERCENT_LESS_HIGH, EST_UNEMPLOYMENT) respectively.

b) Expected results

With this research, we are looking to find the relationship between our independent variables and how they affect state populations that do not enroll in health insurance plans. We have tested how income of these populations are reflecting on insurance enrollment in different states, as well as education levels and unemployment rates.

We started off with the following hypotheses:

- Income can be a major factor in decision making about enrolling in health insurance plans and both low income and very high income groups would, for different reasons, would not enroll into health insurance plans.
- Less access to information or lack of formal education can increase the lack of awareness on how important and beneficial enrolling in a health insurance plan can be. States with large percentage of uneducated population might have the largest percentages of uninsured population.
- Lack of a fixed income may lead to people not considering health insurance plans for a financial reason since they do not have access to work related health-coverage benefits usually obtained through a job. We expect to see that states with higher unemployment rates are the ones with higher uninsured percentage of its population as well.

3. Results

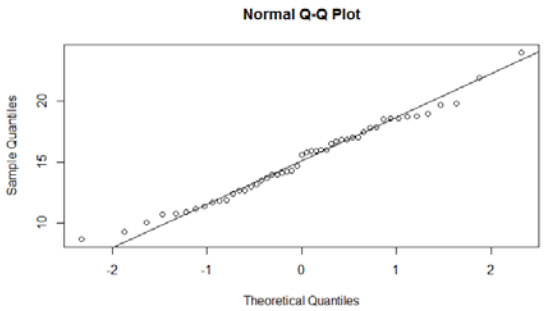
Exploratory Data Analysis

Measure of Central Tendency

Variable	Mean	Range	Standard Deviation
PERCENT_POVERTY	15.138	8.7 – 24.0	3.359
PERCENT_LESS_HIGH	11.877	6.5 – 18.3	3.142
EST_UNEMPLOYMENT	7.777	2.6 – 11	1.856
PERCENT_UNINSURED_POP	13.499	3.7 – 22.1	3.790

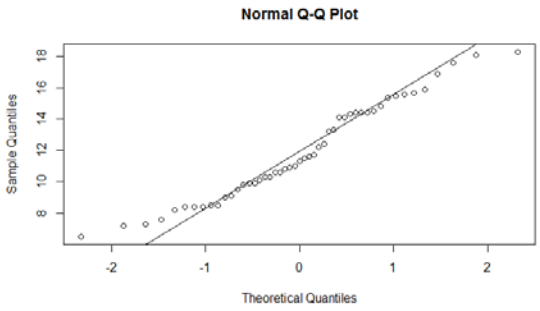
Distribution of variables

a) PERCENT_POVERTY



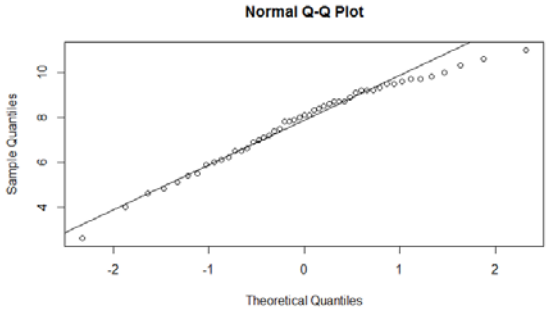
This variable is not normally distributed and is skewed.

b) PERCENT_LESS_HIGH



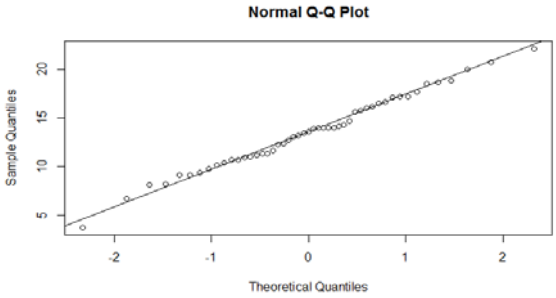
This variable is not normally distributed and is skewed. It is the least normally distributed variable

c) EST_UNEMPLOYMENT



This variable is not normally distributed and is skewed.

d) PERCENT_UNINSURED_POP



This variable is nearly normally distributed but is slightly skewed. It is considerably more normal compared to other variables.

Simple Linear Regression

This method is incorporated for each predictor – outcome pair in this analysis to answer the 3 segments/ sub questions that were described earlier in this document.

We want to obtain an equation of the form for each predictor outcome pair:

OUTCOME = B_0 + B_1 * PREDICTOR

For all these tests, we assumed

Null Hypothesis: B = 0, i.e. outcome variable cannot be predicted by predictor variable.

Alternative Hypothesis: B≠0 i.e. outcome variable can be predicted by predictor variable.

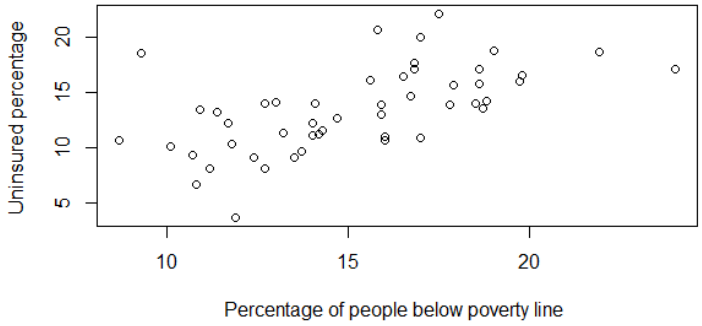
1. Do states with higher percentage of low income population have a higher percentage of uninsured people?

Null hypothesis H_0: B = 0, i.e. percentage of uninsured people cannot be predicted by income.

Alternative Hypothesis H_a: B≠0 i.e. percentage of uninsured people can be predicted by income.

We execute a scatterplot for the Predictor variable (income level) vs Outcome variable (uninsured

percentage)



We find the covariance of PERCENT_POVERTY and PERCENT_UNINSURED_POP is 7.533556.

The covariance only tells us the skew but not the strength of the relationship. Based on the scatterplot and the covariance value we can conclude that poverty level has a positive relationship (covariance) with percentage of uninsured population. This means that PERCENT_UNINSURED_POP can be predicted by PERCENT_POVERTY.

Degrees of freedom	1 and 47
F statistic	25.31
p-value	7.569e-06
Adjusted R squared value	0.3362
Significant (Yes/No)	Yes

Therefore we reject the Null hypothesis, and consider the Alternative Hypothesis.

Summary of regression:

	B_0	**B_1**
Coefficient value	3.3929	0.6676
t-value	1.650	5.031
p-value	0.106	7.57e-06
Significant (Yes/No)	No	Yes

From this test, obtain the linear regression equation

PERCENT_UNINSURED_POP = 0.6676*PERCENT_POVERTY

With every unit increase in PERCENT_POVERTY, there is a 0.6676 unit increase in

PERCENT_UNINSURED_POP. *(For detailed results, refer to Appendix Section 2)*

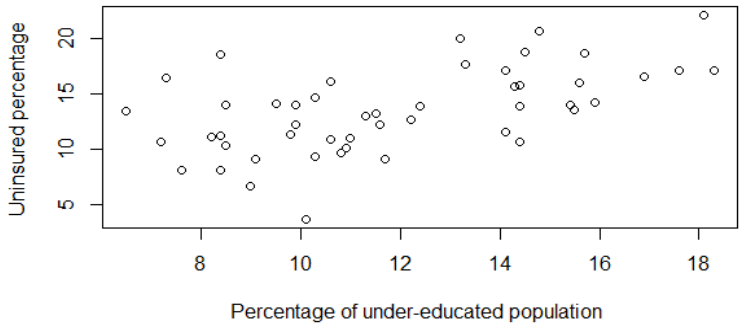
Based on this test, we can conclude that greater the percentage of low income population, higher is the percentage of uninsured people.

2. Do states with higher percentage of people with less or no education have a higher percentage of uninsured people?

Null hypothesis: B = 0, i.e. percentage of uninsured people cannot be predicted by education level.

Alternative Hypothesis: B≠0 i.e. percentage of uninsured people can be predicted by education level.

We execute a scatterplot for the Predictor variable (education level) vs Outcome variable (uninsured percentage)



The covariance of PERCENT_LESS_HIGH and PERCENT_UNINSURED_POP is 6.740258.

The covariance only tells us the skew but not the strength of the relationship. Based on the scatterplot and the covariance value we can conclude that education has a positive relationship (covariance) with percentage of uninsured population. This means that PERCENT_UNINSURED_POP can be predicted by PERCENT_LESS_HIGH.

Degrees of freedom	1 and 47
F statistic	22.15
Adjusted R squared value	0.3058
p-value	2.254e-05
Significant (Yes/No)	Yes

Therefore we reject the Null hypothesis and consider the Alternative Hypothesis.

Summary of regression:

	B_0	B_1
Coefficient value	*5.3897*	*0.6828*
t-value	3.026	4.703
p-value	0.00417	2.254e-05
Significant (Yes/No)	Yes	Yes

From this test, obtain the linear regression equation

PERCENT_UNINSURED_POP = 5.3897 + 0.6828*PERCENT_LESS_HIGH

This means that with every unit increase in PERCENT_LESS_HIGH, there is a 0.6828 unit increase in PERCENT_UNINSURED_POP. *(For detailed results, refer to Appendix Section 3)*

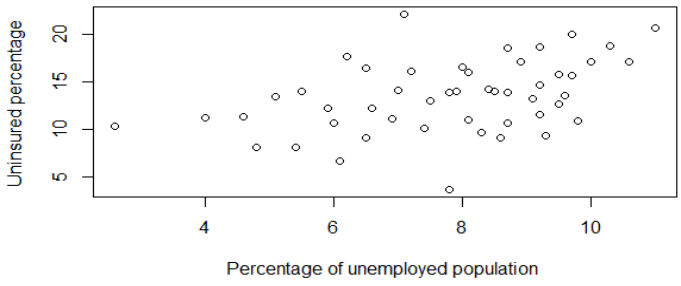
Based on this test, we can conclude that greater the percentage of people with less or no education, higher is the percentage of uninsured people.

3. Do states with higher rates of unemployment have a higher percentage of uninsured people?

Null hypothesis: B = 0, i.e. percentage of uninsured people cannot be predicted by unemployment level

Alternative Hypothesis: B≠0 i.e. percentage of uninsured people can be predicted by unemployment level

We execute a scatterplot for the Predictor variable (unemployment) vs Outcome variable (percent uninsured).



The covariance of EST_UNEMPLOYMENT and PERCENT_UNINSURED_POP is 2.939325

Since the covariance only tells us the skew but not the strength of the relationship, we can conclude that unemployment level has a positive relationship (covariance) with percentage of uninsured population. This means that PERCENT_UNINSURED_POP can be predicted by EST_UNEMPLOYMENT.

Degrees of freedom	1 and 47
F statistic	10.07
Adjusted R squared value	0.159
p-value	0.002654
Significant (Yes/No)	Yes

Therefore we reject the Null hypothesis and consider the Alternative Hypothesis.

Summary of regression:

	B_0	B_1
Coefficient value	6.7899	0.8627
t-value	3.127	3.174
p-value	0.00303	0.00265
Significant (Yes/No)	Yes	Yes

From this test, obtain the linear regression equation

PERCENT_UNINSURED_POP = 6.7899 + 0.8627*EST_UNEMPLOYMENT

With every unit increase in EST_UNEMPLOYMENT, there is a 0.8627 unit increase in PERCENT_UNINSURED_POP. *(For detailed results, refer to Appendix Section 4)*

Based on this test, we can conclude that greater the percentage of unemployment rates, higher is the percentage of uninsured people.

Summary of results:

From this set of independent single linear regressions, we have obtained the following equations:

PERCENT_UNINSURED_POP = 0.6676*PERCENT_POVERTY

PERCENT_UNINSURED_POP = 5.3897 + 0.6828*PERCENT_LESS_HIGH

PERCENT_UNINSURED_POP = 6.7899 + 0.8627*EST_UNEMPLOYMENT

Multiple Linear Regression

Multiple Linear Regression is used to obtain an equation of the form:

PERCENT_UNINSURED_POP = B_0 + B_1 * PERCENT_POVERTY + B_2 * PERCENT_LESS_HIGH + B_3 * EST_UNEMPLOYMENT

Main Research Question: *"Can income, education and employment be used to predict the percentage of uninsured people in various states in the U.S.?"*

For all these tests, we assumed the following:

Null Hypothesis H_0: Percentage of uninsured people cannot be predicted by any of the predictor variables - percentage of people below the poverty level, percentage of people with less education, and unemployment rate.

$B_1 = B_2 = \ldots B_p = 0$, where p = {1, number of predictors in the model} i.e. outcome variable cannot be predicted by any predictor variable.

Alternative Hypothesis H_a: Percentage of uninsured people can be predicted by atleast one of the predictors – percentage of people below the poverty level, percentage of people with less education, and unemployment rate.

$B_j \neq 0$ for at least one of j, j belongs to {1, number of predictors in the model} i.e. outcome variable can be predicted by at least 1 predictor variable.

We considered the following 7 possible models to perform a multiple linear regression test:

Model 1: Percentage of uninsured population (PERCENT_UNINSURED_POP) is related to only percentage of people below the poverty level (PERCENT_POVERTY).

Model 2: Percentage of uninsured population (PERCENT_UNINSURED_POP) is related to only the percentage of people with less education (PERCENT_LESS_HIGH).

Model 3: Percentage of uninsured population (PERCENT_UNINSURED_POP) is related to only the unemployment rate (EST_UNEMPLOYMENT).

Model 4: Percentage of uninsured population (PERCENT_UNINSURED_POP) is related to only percentage of people below the poverty level (PERCENT_POVERTY) and percentage of people with less education (PERCENT_LESS_HIGH).

Model 5: Percentage of uninsured population (PERCENT_UNINSURED_POP) is related to only percentage of people with less education (PERCENT_LESS_HIGH) and unemployment rate (EST_UNEMPLOYMENT).

Model 6: Percentage of uninsured population (PERCENT_UNINSURED_POP) is related to only

(PERCENT_POVERTY) and unemployment rate (EST_UNEMPLOYMENT).

Model 7: Percentage of uninsured population (PERCENT_UNINSURED_POP) is related to percentage of people below the poverty level (PERCENT_POVERTY), percentage of people with less education (PERCENT_LESS_HIGH) and unemployment rate (EST_UNEMPLOYMENT).

We conducted ANOVA tests to compare these models against each other. The summary of these results are:

Model A	**Model B**	**Df1**	**Df2**	**F(DF1,Df2)**	**P**	**Significant: TRUE/FALSE (P < 0. 05)**	**Better Model**	**Effective Model Order**
1	4	1	46	2.0897	0.1551	FALSE	1	1>4>2
2	4	1	46	4.2908	0.0440	TRUE	4	
2	5	1	46	2.9395	0.5926	FALSE	2	2>5>3
3	5	1	46	10.085	0.0027	TRUE	5	
1	6	1	46	1.2285	0.2735	FALSE	1	1>6>3
3	6	1	46	13.84	0.0005	TRUE	6	
4	7	1	45	1.1535	0.6185	FALSE	4	4>6>7>5
5	7	1	45	4.1615	0.0472	TRUE	7	
6	7	1	45	1.0765	0.3050	FALSE	6	
1	7	2	45	1.1535	0.3247	FALSE	1	1>2>7>3
2	7	2	45	2.2359	0.1185	FALSE	2	
3	7	2	45	7.47	0.0016	TRUE	7	

We obtained that Model 1 > Model 4 > Model 2 > Model 6 > Model 7 > Model 5 > Model 3.
Model 1 was the best model to explain the percentage of uninsured population and can be given by the equation: PERCENT_UNINSURED = B_0 + B_1*PERCENT_POVERTY

From the summary of Model 1, the following values were obtained:

	B_0	**B_1**
Coefficient value	3.3929	0.6676
t-value	1.650	5.031
p-value	0.106	7.57e-06
Significant (Yes/No)	No	Yes

Hence, PERCENT_UNINSURED = 0.6676*PERCENT_POVERTY describes the best model for this data *(for detailed results, refer to Appendix Section 5).* With every unit increase in the percentage of population below the poverty line, there is 0.6676 unit increase in the percentage of uninsured population.

When we tested for the assumptions for multiple linear regression for Model 7 *(for detailed results, refer to Appendix Section 6),* we found that it satisfied all assumptions except for constancy of errors, multicollinearity and linearity. Since all the predictors were strongly correlated, it was indicative in our final outcome where we chose Model 1 as a representative model which eliminated the other strongly correlated predictors. This implies that poverty level is a better and more significant indicator than unemployment rates and percentage of population with less education.

4. Conclusion

Limitations

One of the major limitation for this analysis is that it considers data only of a single year (2013) rather than a time series. This limits extension of the analysis to another year or over a time frame. Another aspect to consider is the fact that time ranges are not necessarily comparable because of the incorporation of ObamaCare that made insurance mandatory in the United States.

Another limitation arises from the fact that all three predictor variables are strongly correlated. This resulted in a multiple linear regression that included only one predictor. Also, un-normalized values were used for these tests, but it did not affect the results of the test as much, due to the fact that the final model obtained contained only one predictor. This was verified by using normalized values and it was found that the final outcome remained the same.

Conclusion

Based on the various tests performed, we can derive the following conclusions –

a) States with higher percentage of low income population have a higher percentage of uninsured people. With every unit increase in the percentage of population below the poverty line (PERCENT_POVERTY), there is a 0.6676 unit increase in the percentage of uninsured population (PERCENT_UNINSURED_POP).

b) States with higher percentage of people with less or no education have a higher percentage of uninsured people. With every unit increase in the percentage of population with less education (PERCENT_LESS_HIGH), there is a 0.6828 unit increase in the percentage of uninsured population (PERCENT_UNINSURED_POP).

c) States with higher unemployment rates have a higher percentage of uninsured people. With every unit increase in the unemployment rate (EST_UNEMPLOYMENT), there is a 0.8627 unit increase in the percentage of uninsured population (PERCENT_UNINSURED_POP).

However, when we tried to compute a multiple regression model, we found that only one predictor was part of the final model chosen. This predictor was percentage of people below the poverty line (PERCENT_POVERTY), and the outcome variable- percentage of uninsured population (PERCENT_UNINSURED_POP) was related to it by the equation:

PERCENT_UNINSURED = 0.6676 * PERCENT_POVERTY.

This implies that education and unemployment often reflect on the poverty levels and hence, a more inclusive predictor would be the poverty level, which would consider both education and unemployment. Based on this analysis, it would be best to keep poverty levels in check and focus on schemes that provide free health care to poorer sections of the population, or substantially reduce the costs of getting insurance for not just an individual but for the entire family.

Implications

Our analysis indicates that people may not be able to afford even the costs associated with insurances, and enforcing a mandatory rule that issues fines in the absence of an insurance may not be a compelling enough reason or strategy to increase subscription for a health insurance. A better approach could be to target specific groups of the population to spread awareness about the benefits and necessity of an insurance. This could help people with no formal education to gain information that they would have otherwise gained in a college or professional educational setting. Alternatively, providing better employment opportunities could help people sustain the costs for insurance, and in turn, health care. In effect, these findings can help the government contemplate necessary policy changes and introduce programs focused on specific segments of the society.

Future Scope

Being able to obtain and analyze data over a larger time frame than just a single year could provide a more reliable relationship between the predictor-outcome pairs. This could also be extended to the pre-ObamaCare years when insurance wasn't mandatory and post ObamaCare timeframe that has mandated insurance. Improving the model could be done by using a different form of regression (non-linear), or using a log based transformation of the variables to eliminate some of the limitations that the current tests face. Another way to enhance the findings would be conduct a Partial Least Squares Regression (PLS) or Principal Components Analysis, as these methods can segment the predictors into smaller, unrelated subsets.

5. Appendix

1. R Code

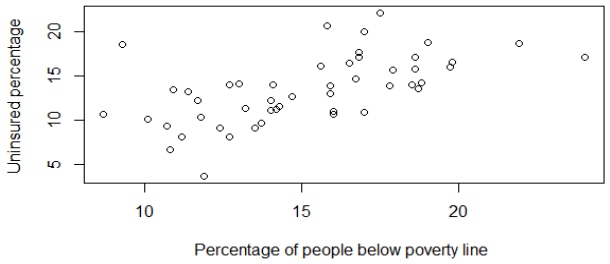
```
# Read data
d = read.csv("data assimilated.csv")

# 1.	Do states with higher percentage of low income population have a higher percentage of uninsured people?
# Simple linear regression

plot(d$PERCENT_POVERTY,d$PERCENT_UNINSURED_POP,xlab="Percentage of people below poverty line ",ylab="Uninsured percentage ")
cov(d$PERCENT_POVERTY,d$PERCENT_UNINSURED_POP)
m1 = lm(PERCENT_UNINSURED_POP~PERCENT_POVERTY,data=d)
summary(m1)



# 2.	Do states with higher percentage of people with less or no education have a higher percentage of uninsured people?
# Simple linear regression

plot(d$PERCENT_LESS_HIGH,d$PERCENT_UNSINSURED_POP,xlab="Percentage of under-educated population ",ylab="Uninsured percentage ")
cov(d$PERCENT_LESS_HIGH,d$PERCENT_UNSINSURED_POP)
m2 = lm(PERCENT_UNINSURED_POP~PERCENT_LESS_HIGH,data=d)
summary(m2)

# 3.	Do states with higher unemployment rates have a higher percentage of uninsured people?
# Simple linear regression

plot(d$EST_UNEMPLOYMENT,d$PERCENT_UNINSURED_POP,xlab="Percentage of unemployed population ",ylab="Uninsured percentage ")
```

```
cov(d$EST_UNEMPLOYMENT,d$PERCENT_UNSINSURED_POP)
m3 = lm(PERCENT_UNINSURED_POP~EST_UNEMPLOYMENT,data=d)
summary(m3)

# Main research question: Can income, education and employment be used to predict the percentage of uninsured people in various states in the U.S.?
# Multiple linear regression

m1 = lm(PERCENT_UNINSURED_POP~PERCENT_POVERTY,data=d)
m2 = lm(PERCENT_UNINSURED_POP~PERCENT_LESS_HIGH,data=d)
m3 = lm(PERCENT_UNINSURED_POP~EST_UNEMPLOYMENT,data=d)
m4 = lm(PERCENT_UNINSURED_POP~PERCENT_POVERTY+PERCENT_LESS_HIGH,data=d)
m5 = lm(PERCENT_UNINSURED_POP~PERCENT_LESS_HIGH+EST_UNEMPLOYMENT,data=d)
m6 = lm(PERCENT_UNINSURED_POP~PERCENT_POVERTY+EST_UNEMPLOYMENT,data=d)
m7 =
lm(PERCENT_UNINSURED_POP~PERCENT_POVERTY+PERCENT_LESS_HIGH+EST_UNEMPLOYMENT,data=d)

summary(m1)
summary(m2)
summary(m3)
summary(m4)
summary(m5)
summary(m6)
summary(m7)

anova(m1,m4)
anova(m2,m4)
anova(m2,m5)
anova(m3,m5)
anova(m1,m6)
anova(m3,m6)
anova(m4,m7)
anova(m5,m7)
```

```
anova(m6,m7)
anova(m1,m7)
anova(m2,m7)
anova(m3,m7)

# Tests for assumptions of linear regression
# generate predicted and residual values

pred=m1$fitted.values
res=m1$residuals

# Test for Independence of errors

durbinWatsonTest(m1)

# Test for Constancy of errors

plot(pred,res)

# Test for Normality of errors

qqnorm(res)
qqline(res)
shapiro.test(res)

#Test for non-linearity

plot(d$PERCENT_POVERTY,d$PERCENT_UNINSURED_POP,xlab="Percentage of people below poverty line ",ylab="Uninsured percentage ")
plot(d$PERCENT_LESS_HIGH,d$PERCENT_UNSINSURED_POP,xlab="Percentage of under-educated population ",ylab="Uninsured percentage ")
plot(d$EST_UNEMPLOYMENT,d$PERCENT_UNINSURED_POP,xlab="Percentage of unemployed population ",ylab="Uninsured percentage ")
```

2. Simple Linear Regression Question 1 Output

plot(d$PERCENT_POVERTY,d$PERCENT_UNINSURED_POP,xlab="Percentage of people below poverty line ",ylab="Uninsured percentage ")



cov(d$PERCENT_POVERTY,d$PERCENT_UNINSURED_POP)

[1] 7.533556

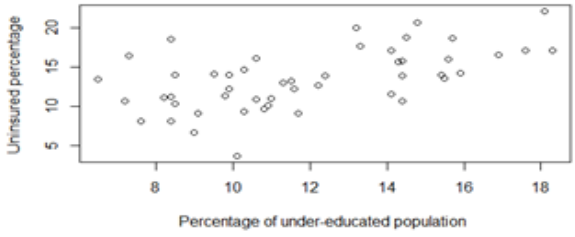
m1 = lm(PERCENT_UNINSURED_POP~PERCENT_POVERTY,data=d)

summary(m1)

```
> d = read.csv("data assimilated.csv")
> m1 = lm(PERCENT_UNINSURED_POP~PERCENT_POVERTY,data=d)
> summary(m1)

Call:
lm(formula = PERCENT_UNINSURED_POP ~ PERCENT_POVERTY, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-7.6076 -1.7657 -0.4682  2.0237  8.9294

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.3928     2.0567   1.650    0.106
PERCENT_POVERTY  0.6676     0.1327   5.031 7.57e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.088 on 47 degrees of freedom
Multiple R-squared:   0.35,     Adjusted R-squared:  0.3362
F-statistic: 25.31 on 1 and 47 DF,  p-value: 7.569e-06
```

3. Simple Linear Regression Question 2 Output

plot(d$PERCENT_LESS_HIGH,d$PERCENT_UNSINSURED_POP,xlab="Percentage of under-educated population ",ylab="Uninsured percentage ")



cov(d$PERCENT_LESS_HIGH,d$PERCENT_UNSINSURED_POP)

[1] 6.740258

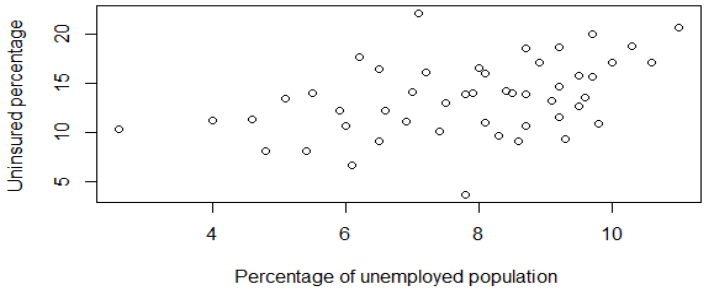
m2 = lm(PERCENT_UNINSURED_POP~PERCENT_LESS_HIGH,data=d)

summary(m2)

```
> m2 = lm(PERCENT_UNINSURED_POP~PERCENT_LESS_HIGH,data=d)
> summary(m2)

Call:
lm(formula = PERCENT_UNINSURED_POP ~ PERCENT_LESS_HIGH, data = d)

Residuals:
   Min     1Q Median     3Q    Max
-8.556 -1.950 -0.074  2.219  7.406

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       5.3897     1.7813   3.026  0.00401 **
PERCENT_LESS_HIGH   0.6828     0.1451   4.706 2.25e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.158 on 47 degrees of freedom
Multiple R-squared:  0.3203,    Adjusted R-squared:  0.3058
F-statistic: 22.15 on 1 and 47 DF,  p-value: 2.254e-05
```

4. Simple Linear Regression Question 3 Output

plot(d$EST_UNEMPLOYMENT,d$PERCENT_UNINSURED_POP,xlab="Percentage of unemployed population ",ylab="Uninsured percentage ")



cov(d$EST_UNEMPLOYMENT,d$PERCENT_UNSINSURED_POP)

[1] 2.939325

m3 = lm(PERCENT_UNINSURED_POP~EST_UNEMPLOYMENT,data=d)

summary(m3)

```
> m3 = lm(PERCENT_UNINSURED_POP~EST_UNEMPLOYMENT,data=d)
> summary(m3)

Call:
lm(formula = PERCENT_UNINSURED_POP ~ EST_UNEMPLOYMENT, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-9.7892 -2.7448  0.3805  2.2484  9.2133

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       6.7899     2.1716   3.127  0.00303 **
EST_UNEMPLOYMENT   0.8627     0.2718   3.174  0.00265 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.476 on 47 degrees of freedom
Multiple R-squared:  0.1765,    Adjusted R-squared:  0.159
F-statistic: 10.07 on 1 and 47 DF,  p-value: 0.002654
```

5. Multiple Linear Regression Main Research Question Output

```
> m1 = lm(PERCENT_UNINSURED_POP~PERCENT_POVERTY,data=d)
> m2 = lm(PERCENT_UNINSURED_POP~PERCENT_LESS_HIGH,data=d)
> m3 = lm(PERCENT_UNINSURED_POP~EST_UNEMPLOYMENT,data=d)
> m4 = lm(PERCENT_UNINSURED_POP~PERCENT_POVERTY+PERCENT_LESS_HIGH,data=d)
> m5 = lm(PERCENT_UNINSURED_POP~PERCENT_LESS_HIGH+EST_UNEMPLOYMENT,data=d)
> m6 = lm(PERCENT_UNINSURED_POP~PERCENT_POVERTY+EST_UNEMPLOYMENT,data=d)
> m7 = lm(PERCENT_UNINSURED_POP~PERCENT_POVERTY+PERCENT_LESS_HIGH+EST_UNEMPLOYMENT,data=d)
```

Model 1: PERCENT_UNINSURED_POP is related to only PERCENT_POVERTY.

i.e.; PERCENT_UNINSURED_POP = B_0+ B_1*PERCENT_POVERTY

```
> summary(m1)

Call:
lm(formula = PERCENT_UNINSURED_POP ~ PERCENT_POVERTY, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-7.6076 -1.7657 -0.4682  2.0237  8.9294

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.3928     2.0567   1.650    0.106
PERCENT_POVERTY   0.6676     0.1327   5.031 7.57e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.088 on 47 degrees of freedom
Multiple R-squared:   0.35,     Adjusted R-squared:  0.3362
F-statistic: 25.31 on 1 and 47 DF,  p-value: 7.569e-06
```

Model 2: PERCENT_UNINSURED_POP is related to only PERCENT_LESS_HIGH.

i.e.; PERCENT_UNINSURED_POP = B_0+ B_1* PERCENT_LESS_HIGH

```
> summary(m2)

Call:
lm(formula = PERCENT_UNINSURED_POP ~ PERCENT_LESS_HIGH, data = d)

Residuals:
   Min     1Q Median     3Q    Max
-8.556 -1.950 -0.074  2.219  7.406

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       5.3897     1.7813   3.026  0.00401 **
PERCENT_LESS_HIGH 0.6828     0.1451   4.706 2.25e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.158 on 47 degrees of freedom
Multiple R-squared:  0.3203,    Adjusted R-squared:  0.3058
F-statistic: 22.15 on 1 and 47 DF,  p-value: 2.254e-05
```

Model 3: PERCENT_UNINSURED_POP is related to only EST_UNEMPLOYMENT.

i.e.; PERCENT_UNINSURED_POP = B_0+ B_1* EST_UNEMPLOYMENT

```
> summary(m3)

Call:
lm(formula = PERCENT_UNINSURED_POP ~ EST_UNEMPLOYMENT, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-9.7892 -2.7448  0.3805  2.2484  9.2133

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       6.7899     2.1716   3.127  0.00303 **
EST_UNEMPLOYMENT  0.8627     0.2718   3.174  0.00265 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.476 on 47 degrees of freedom
Multiple R-squared:  0.1765,    Adjusted R-squared:  0.159
F-statistic: 10.07 on 1 and 47 DF,  p-value: 0.002654
```

Model 4: PERCENT_UNINSURED_POP is related to only PERCENT_POVERTY and

PERCENT_LESS_HIGH.

i.e.; PERCENT_UNINSURED_POP = B_0+ B_1*PERCENT_POVERTY +B_2*PERCENT_LESS_HIGH

```
> summary(m4)

Call:
lm(formula = PERCENT_UNINSURED_POP ~ PERCENT_POVERTY + PERCENT_LESS_HIGH,
    data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-7.7948 -2.1130 -0.6112  1.5134  8.6801

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.1155     2.0423   1.526    0.134
PERCENT_POVERTY   0.4326     0.2089   2.071    0.044 *
PERCENT_LESS_HIGH 0.3228     0.2233   1.446    0.155
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.053 on 46 degrees of freedom
Multiple R-squared:  0.3783,    Adjusted R-squared:  0.3513
F-statistic: 13.99 on 2 and 46 DF,  p-value: 1.788e-05
```

Model 5: PERCENT_UNINSURED_POP is related to only PERCENT_LESS_HIGH and EST_UNEMPLOYMENT.

i.e.; PERCENT_UNINSURED_POP = B_0+ B_1* PERCENT_LESS_HIGH+B_2* EST_UNEMPLOYMENT

```
> summary(m5)

Call:
lm(formula = PERCENT_UNINSURED_POP ~ PERCENT_LESS_HIGH + EST_UNEMPLOYMENT,
    data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-8.6815 -2.1221 -0.0642  2.0826  7.0046

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       4.8208     2.0824   2.315  0.02513 *
PERCENT_LESS_HIGH 0.6145     0.1935   3.176  0.00267 **
EST_UNEMPLOYMENT  0.1774     0.3294   0.539  0.59264
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.182 on 46 degrees of freedom
Multiple R-squared:  0.3246,    Adjusted R-squared:  0.2952
F-statistic: 11.05 on 2 and 46 DF,  p-value: 0.0001203
```

Model 6: PERCENT_UNINSURED_POP is related to only PERCENT_POVERTY and EST_UNEMPLOYMENT.

i.e.; PERCENT_UNINSURED_POP = B_0+ B_1* PERCENT_POVERTY +B_2* EST_UNEMPLOYMENT

```
> summary(m6)

Call:
lm(formula = PERCENT_UNINSURED_POP ~ PERCENT_POVERTY + EST_UNEMPLOYMENT,
    data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-7.9060 -2.0232 -0.1828  1.4938  8.1153

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.3187     2.2690   1.022  0.31217
PERCENT_POVERTY   0.5776     0.1553   3.720  0.00054 ***
EST_UNEMPLOYMENT  0.3132     0.2826   1.108  0.27346
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.081 on 46 degrees of freedom
Multiple R-squared:  0.367,     Adjusted R-squared:  0.3394
F-statistic: 13.33 on 2 and 46 DF,  p-value: 2.71e-05
```

Model 7: PERCENT_UNINSURED_POP is related to PERCENT_POVERTY, PERCENT_LESS_HIGH and EST_UNEMPLOYMENT.

i.e.; PERCENT_UNINSURED_POP = B_0 + B_1*PERCENT_POVERTY + B_2*PERCENT_LESS_HIGH + B_3* EST_UNEMPLOYMENT

```
> summary(m7)

Call:
lm(formula = PERCENT_UNINSURED_POP ~ PERCENT_POVERTY + PERCENT_LESS_HIGH +
    EST_UNEMPLOYMENT, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-7.9127 -2.0980 -0.2512  1.4345  8.3104

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.6183     2.2855   1.146   0.2580
PERCENT_POVERTY   0.4298     0.2107   2.040   0.0472 *
PERCENT_LESS_HIGH 0.2637     0.2542   1.038   0.3050
EST_UNEMPLOYMENT  0.1598     0.3187   0.501   0.6185
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.078 on 45 degrees of freedom
Multiple R-squared:  0.3817,    Adjusted R-squared:  0.3405
F-statistic: 9.262 on 3 and 45 DF,  p-value: 6.952e-05
```

ANOVA Tests for Models

```
> anova(m1,m4)
Analysis of Variance Table

Model 1: PERCENT_UNINSURED_POP ~ PERCENT_POVERTY
Model 2: PERCENT_UNINSURED_POP ~ PERCENT_POVERTY + PERCENT_LESS_HIGH
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     47 448.24
2     46 428.76  1    19.478 2.0897 0.1551
> anova(m2,m4)
Analysis of Variance Table

Model 1: PERCENT_UNINSURED_POP ~ PERCENT_LESS_HIGH
Model 2: PERCENT_UNINSURED_POP ~ PERCENT_POVERTY + PERCENT_LESS_HIGH
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     47 468.75
2     46 428.76  1    39.988 4.2902 0.04397 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> anova(m2,m5)
Analysis of Variance Table

Model 1: PERCENT_UNINSURED_POP ~ PERCENT_LESS_HIGH
Model 2: PERCENT_UNINSURED_POP ~ PERCENT_LESS_HIGH + EST_UNEMPLOYMENT
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     47 468.75
2     46 465.81  1    2.9395 0.2903 0.5926


> anova(m4,m7)
Analysis of Variance Table

Model 1: PERCENT_UNINSURED_POP ~ PERCENT_POVERTY + PERCENT_LESS_HIGH
Model 2: PERCENT_UNINSURED_POP ~ PERCENT_POVERTY + PERCENT_LESS_HIGH +
    EST_UNEMPLOYMENT
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     46 428.76
2     45 426.38  1    2.3819 0.2514 0.6185


> anova(m1,m7)
Analysis of Variance Table

Model 1: PERCENT_UNINSURED_POP ~ PERCENT_POVERTY
Model 2: PERCENT_UNINSURED_POP ~ PERCENT_POVERTY + PERCENT_LESS_HIGH +
    EST_UNEMPLOYMENT
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     47 448.24
2     45 426.38  2    21.86 1.1535 0.3247
> anova(m2,m7)
Analysis of Variance Table

Model 1: PERCENT_UNINSURED_POP ~ PERCENT_LESS_HIGH
Model 2: PERCENT_UNINSURED_POP ~ PERCENT_POVERTY + PERCENT_LESS_HIGH +
    EST_UNEMPLOYMENT
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     47 468.75
2     45 426.38  2    42.37 2.2359 0.1186
```

```
> anova(m1,m6)
Analysis of Variance Table

Model 1: PERCENT_UNINSURED_POP ~ PERCENT_POVERTY
Model 2: PERCENT_UNINSURED_POP ~ PERCENT_POVERTY + EST_UNEMPLOYMENT
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     47 448.24
2     46 436.58  1    11.659 1.2285 0.2735
> anova(m3,m6)
Analysis of Variance Table

Model 1: PERCENT_UNINSURED_POP ~ EST_UNEMPLOYMENT
Model 2: PERCENT_UNINSURED_POP ~ PERCENT_POVERTY + EST_UNEMPLOYMENT
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     47 567.94
2     46 436.58  1    131.36 13.84 0.0005401 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> anova(m3,m5)
Analysis of Variance Table

Model 1: PERCENT_UNINSURED_POP ~ EST_UNEMPLOYMENT
Model 2: PERCENT_UNINSURED_POP ~ PERCENT_LESS_HIGH + EST_UNEMPLOYMENT
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     47 567.94
2     46 465.81  1    102.13 10.085 0.002668 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> anova(m5,m7)
Analysis of Variance Table

Model 1: PERCENT_UNINSURED_POP ~ PERCENT_LESS_HIGH + EST_UNEMPLOYMENT
Model 2: PERCENT_UNINSURED_POP ~ PERCENT_POVERTY + PERCENT_LESS_HIGH +
    EST_UNEMPLOYMENT
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     46 465.81
2     45 426.38  1    39.431 4.1615 0.04725 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> anova(m6,m7)
Analysis of Variance Table

Model 1: PERCENT_UNINSURED_POP ~ PERCENT_POVERTY + EST_UNEMPLOYMENT
Model 2: PERCENT_UNINSURED_POP ~ PERCENT_POVERTY + PERCENT_LESS_HIGH +
    EST_UNEMPLOYMENT
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     46 436.58
2     45 426.38  1      10.2 1.0765  0.305
> anova(m3,m7)
Analysis of Variance Table

Model 1: PERCENT_UNINSURED_POP ~ EST_UNEMPLOYMENT
Model 2: PERCENT_UNINSURED_POP ~ PERCENT_POVERTY + PERCENT_LESS_HIGH +
    EST_UNEMPLOYMENT
  Res.Df    RSS Df Sum of Sq    F  Pr(>F)
1     47 567.94
2     45 426.38  2    141.56 7.47 0.00158 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We obtain that:

- For models 1, 4: F (1, 46) = 2.0897, p=0.1551 >0.05. Hence addition of predictors has no effect. m1 is a better model than m4.

- For models 2, 4: F (1, 46) = 4.2908, p=0.04397 <0.05. Hence addition of predictors has an effect. m4 is a better model than m2.

Hence Model 1 > Model 4 > Model 2

- For models 2, 5: F (1, 46) = 2.9395, p=0.5926 >0.05. Hence addition of predictors has no effect. m2 is a better model than m5.
- For models 3, 5: F (1, 46) = 10.085, p=0.002668<0.05. Hence addition of predictors has an effect. m5 is a better model than m3.

Hence Model 2 > Model 5 > Model 3

- For models 3, 6: F (1, 46) = 13.84, p=0.0005401 <0.05. Hence addition of predictors has an effect. m6 is a better model than m3.
- For models 1, 6: F (1, 46) = 1.2285, p=0.2735 >0.05. Hence addition of predictors has no effect. m1 is a better model than m6.

Hence Model 1 > Model 6 > Model 3

- For models 4, 7: F (1, 45) = 0.2514, p=0.6185 >0.05. Hence addition of predictors has no effect. m4 is a better model than m7.
- For models 5, 7: F (1, 45) = 4.1615, p=0.04725 <0.05. Hence addition of predictors has an effect. m7 is a better model than m5.
- For models 6, 7: F (1, 45) = 1.0765, p=0.305 >0.05. Hence addition of predictors has no effect. m6 is a better model than m7.

Hence Model 4 and Model 6 > Model 7 > Model 5

- For models 1, 7: F (2, 45) = 1.1535, p=0.3247 >0.05. Hence addition of predictors has no effect. m1 is a better model than m7.
- For models 2, 7: F (2, 45) = 2.2359, p=0.1186 >0.05. Hence addition of predictors has no effect. m2 is a better model than m7.
- For models 3, 7: F (2, 45) = 7.47, p=0.00158 <0.05. Hence addition of predictors has an effect. m7 is a better model than m3.

Hence we can conclude that Model 1 > Model 4 > Model 2 > Model 6 > Model 7 > Model 5 > Model 3

In effect, Model 1 i.e.; m1 is the best model that has is more significant in predicting the outcome variable.

PERCENT_UNINSURED = B_0 + B_1*PERCENT_POVERTY

From the summary (m1), we obtain

The significance test for Intercept t-value = 1.650, p = 0.106. Hence B_0 is not significant.

The significance test for PERCENT_POVERTY t-value = 5.031, p = 7.57e-06. Hence B_1 is significant.

$B_1 = 0.6676$

Therefore,

PERCENT_UNINSURED = 0.6676*PERCENT_POVERTY describes the best model for this data.

```
> summary(m1)

Call:
lm(formula = PERCENT_UNINSURED_POP ~ PERCENT_POVERTY, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-7.6076 -1.7657 -0.4682  2.0237  8.9294

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.3928     2.0567   1.650    0.106
PERCENT_POVERTY   0.6676     0.1327   5.031 7.57e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.088 on 47 degrees of freedom
Multiple R-squared:   0.35,     Adjusted R-squared:  0.3362
F-statistic: 25.31 on 1 and 47 DF,  p-value: 7.569e-06
```

6. Tests for assumption of multiple linear regression

- We generate the predicted values and residuals for Model 1

```
> pred = m1$fitted.values
> res = m1$residuals
```

- To test for independence of errors, we run the Durbin Watson Test.

  Here, our null hypothesis H_0: there are no auto correlations.

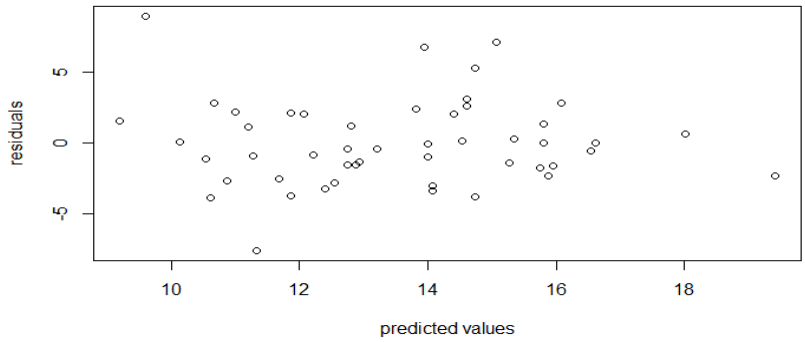
  Alternative hypothesis H_a: there is atleast one auto correlation.

  Based on this test, DW statistic = 1.838082, p-value = 0.51>0.05. Hence we fail to reject H_0. There are no auto correlations. The assumption of independence of errors is not violated.

```
> durbinWatsonTest(m1)
 lag Autocorrelation D-W Statistic p-value
   1      0.06648767      1.838082    0.51
 Alternative hypothesis: rho != 0
```

- To check for constant errors,

```
> plot(pred,res)
```



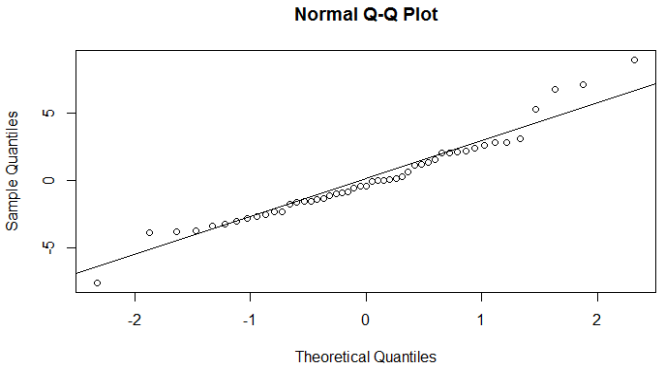
Residuals do not have constant variance across predictor values. Heteroscedasticity is occurring as this assumption is violated. Hence we could apply a log transformation or choose an alternative model that better explains the relationship between the predictor and outcome variable.

- To check for normality of errors,

  We plot a qq plot of the residuals. The plot looks nearly normal, hence we perform the Shapiro –Wilk test for normality. Here H_0: The data follows a normal distribution. H_a: The data does not follow a normal distribution.

  This test results in W = 0.95673, p = 0.06951 >0.05. We fail to reject H_0. This indicates that the errors are normally distributed.

```
> qqnorm(res)
> qqline(res)
> shapiro.test(res)

        Shapiro-Wilk normality test

data:  res
W = 0.95673, p-value = 0.06951
```



- To check for multicollinearity ,

		PERCENT_ POVERTY	PERCENT_ LESS_HIGH	EST_ UNEMPLOYMENT	PERCENT_ UNINSURED_POP
PERCENT_POVERTY	Pearson Correlation	1	.778**	.523**	.592**
	Sig. (2-tailed)		.000	.000	.000
	Sum of Squares and Cross-products	541.656	394.243	155.573	361.611
	Covariance	11.285	8.213	3.241	7.534
	N	49	49	49	49
PERCENT_LESS_HIGH	Pearson Correlation	.778**	1	.655**	.566**
	Sig. (2-tailed)	.000		.000	.000
	Sum of Squares and Cross-products	394.243	473.845	182.375	323.532
	Covariance	8.213	9.872	3.799	6.740
	N	49	49	49	49
EST_UNEMPLOYMENT	Pearson Correlation	.523**	.655**	1	.420**
	Sig. (2-tailed)	.000	.000		.003
	Sum of Squares and Cross-products	155.573	182.375	163.545	141.088
	Covariance	3.241	3.799	3.407	2.939
	N	49	49	49	49
PERCENT_UNINSURED_POP	Pearson Correlation	.592**	.566**	.420**	1
	Sig. (2-tailed)	.000	.000	.003	
	Sum of Squares and Cross-products	361.611	323.532	141.088	689.651
	Covariance	7.534	6.740	2.939	14.368
	N	49	49	49	49

Looking at the matrix we can observe the following correlations:

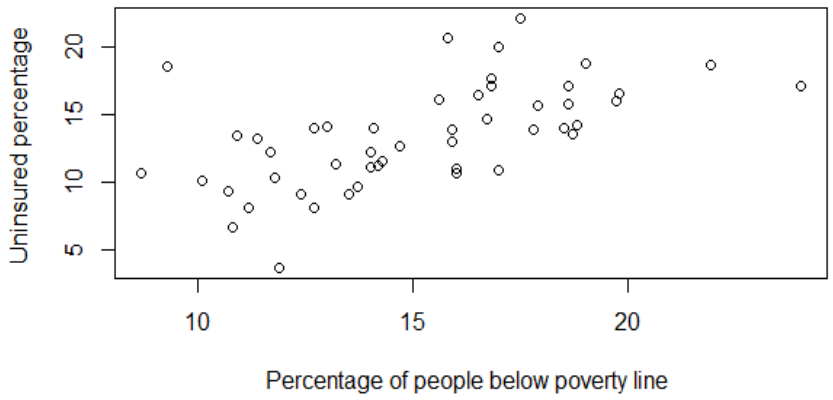
- PERCENT_POVERTY and PERCENT_LESS_HIGH are strongly positive correlated (r = 0.778).
- PERCENT_POVERTY and PERCENT_UNEMPLOYMENT are strongly positive correlated (r = 0.523).
- PERCENT_ UNEMPLOYMENT and PERCENT_LESS_HIGH are strongly positive correlated (r =0.655).

We can conclude that all the predictors are strongly correlated and this can results in violation in the multicollinearity assumption, if model 7 is used. This has been overcome by Model 1 which eliminates the other strongly correlated predictors. An alternative would be to use different regression techniques like **Partial Least Squares Regression (PLS) or Principal Components Analysis.**

- To check for non-linearity

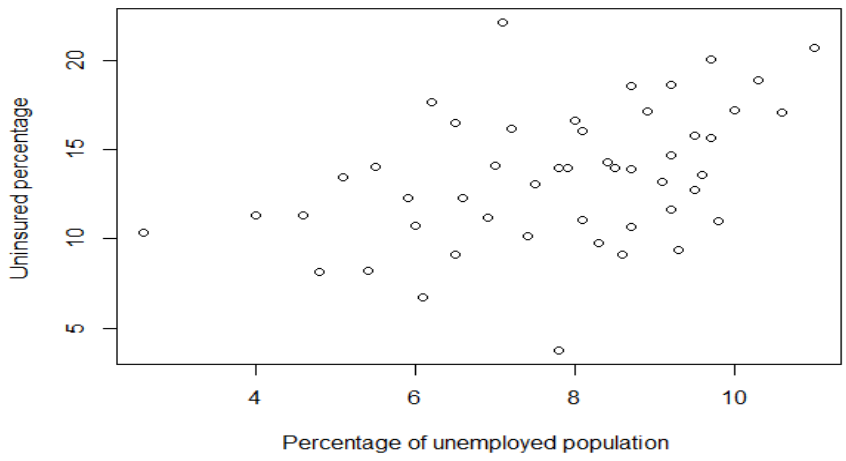
We will inspect scatterplots individually

a) PERCENT_POVERTY



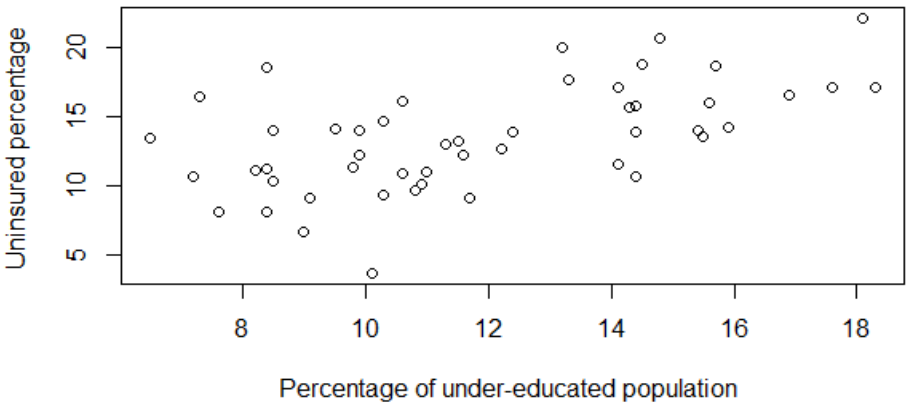
This scatterplot shows a positive association but an imperfect linear association. But this graph displays linear tendencies

b) EST_UNEMPLOYMENT



This scatterplot shows a positive association but no linear association.

c) PERCENT_LESS_HIGH



This scatterplot shows a positive association but an imperfect linear association. But this graph displays linear tendencies.