US CENSUS ANALYSIS

STAT 515 (Final Project)

Ashok Vardhan Kari

December 17, 2015

1 TABLE OF CONTENTS

2 Introduction 2
3 Hypothesis 2
4 Data Collection 2
4.1 Data Sources 3
5 Data Cleansing 3
6 Analysis 4
6.1 Hypothesis 1 4
6.1.1 TC Maps 5
6.1.2 Micromaps 8
6.1.3 Linear Regression 10
6.1.4 Plotting regression output using VISREG2D 12
6.2 Hypothesis 2 13
6.2.1 Net Migration influencing factors using Micromaps 13
6.2.2 Linear Regression 14
7 Conclusion 16

2 INTRODUCTION

Objective of the project is to analyze US Census data and draw insights from it. Approach of the analysis will look to be more concept driven than objective driven as most of the project is primarily wrapped around the visualization concepts taught in the class.

3 HYPOTHESIS

1. Factors that led to population rise/ fall from 2000-2010 are the same for 2010-2014.
2. Rise in weekly wages and CO2 emissions are significant factors for rise/ fall in Net Migration.

4 DATA COLLECTION

I have created the dataset by extracting variables from 15 datasets.

My data set contain 42 variables. Most of them taken from different datasets. It took me about two full weeks to finalize and extract all the necessary factors.

This is how my dataset looks like.

State	P_2000	P_2010	P_2014	Pop%_Change_1	Pop%_Change_2	GDP_2000	GDP_2010	GDP_2014	GDP%_Cha	GDP%_Cha
Alabama	4447207	4779736	4849377	7.48	1.46	120428	176287	200414	46	14
Alaska	626933	710231	736732	13.29	3.73	26932	52490	56647	95	8
Arizona	5130247	6392017	6731484	24.59	5.31	166108	248110	286554	49	15
Arkansas	2673293	2915918	2966369	9.08	1.73	69111	105195	120035	52	14
California	33871653	37253956	38802500	9.99	4.16	1377014	1964588	2305921	43	17
Colorado	4302086	5029196	5355866	16.9	6.5	178331	257810	305871	45	19
Connecticut	3405650	3574097	3596677	4.95	0.63	166995	231060	250569	38	8
Delaware	783559	897934	935614	14.6	4.2	41677	57369	63404	38	11
District of Colu	572086	601723	658893	5.18	9.5	60458	104175	116378	72	12
Florida	15982571	18801310	19893297	17.64	5.81	490538	731278	838939	49	15
Georgia	8186653	9687653	10097343	18.33	4.23	304942	409747	474696	34	16
Hawaii	1211497	1360301	1419561	12.28	4.36	41247	67451	76171	64	13
Idaho	1293957	1567582	1634464	21.15	4.27	38416	55576	63235	45	14
Illinois	12419927	12830632	12880580	3.31	0.39	492922	653597	736285	33	13
Indiana	6080827	6483802	6596855	6.63	1.74	205807	282262	318085	37	13
Iowa	2926538	3046355	3107126	4.09	1.99	95021	141552	169707	49	20
Kansas	2688925	2853118	2904021	6.11	1.78	87446	127967	144407	46	13
Kentucky	4042193	4339367	4413457	7.35	1.71	115126	165550	187788	44	13
Louisiana	4469035	4533372	4649676	1.44	2.57	134251	232694	251672	73	8
Maine	1274779	1328361	1330089	4.2	0.13	36684	51336	54324	40	6
Maryland	5296647	5773552	5976407	9	3.51	192934	314107	346857	63	10
Massachusetts	6349364	6547629	6745408	3.12	3.02	289554	398347	455732	38	14
Michigan	9938823	9883640	9909877	-0.56	0.27	351996	385800	448243	10	16
Minnesota	4919631	5303925	5457173	7.81	2.89	192948	271973	317237	41	17
Mississippi	2844754	2967297	2994079	4.31	0.9	66171	95258	104753	44	10
Missouri	5596564	5988927	6063589	7.01	1.25	187707	256576	279835	37	9
Montana	902200	989415	1023579	9.67	3.45	21884	37315	44135	71	18

4.1 Data Sources

As per the general analysis of census, the top influencing factors for population rise/ fall are

1. Birth Rate
2. Death Rate
3. Migration
4. Employment
5. Salary
6. Lifestyle

In quest for the data for above specified factors, I found the following data sources available for public.

- US Census Bureau (Census Data)
- Migration Policy Institute (Migration Data)
- American Fact Finder (Employment)
- Wikipedia (CO2 Emissions)
- The Disaster Center (Crime Data)
- Bureau of Labor Statistics (Income/ Wages Data)

5 Data Cleansing

- There wasn't much need for cleaning except for formatting the data.
- Few columns were created in order to calculate the percentage change over a period of time.
- Excel has been used for most of the cleaning.
- Some of the data preprocessing tasks such as breaking down the dataset for a particular analysis has been done using R.

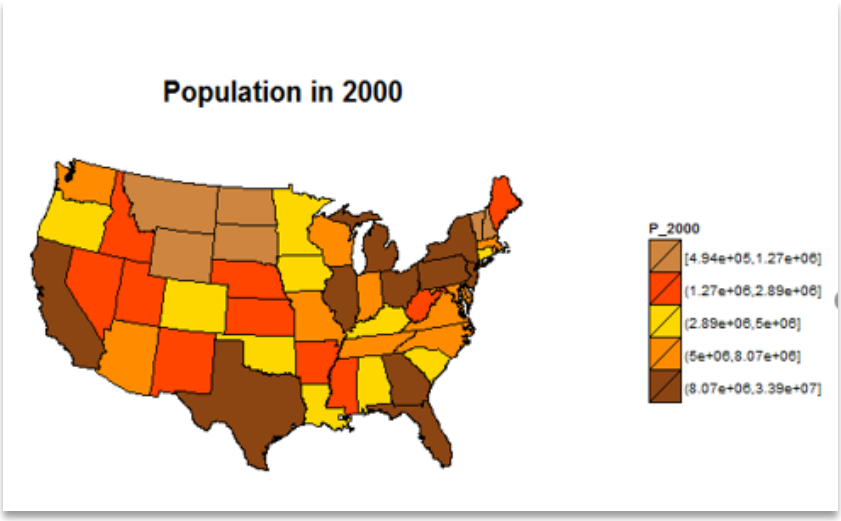
6 ANALYSIS

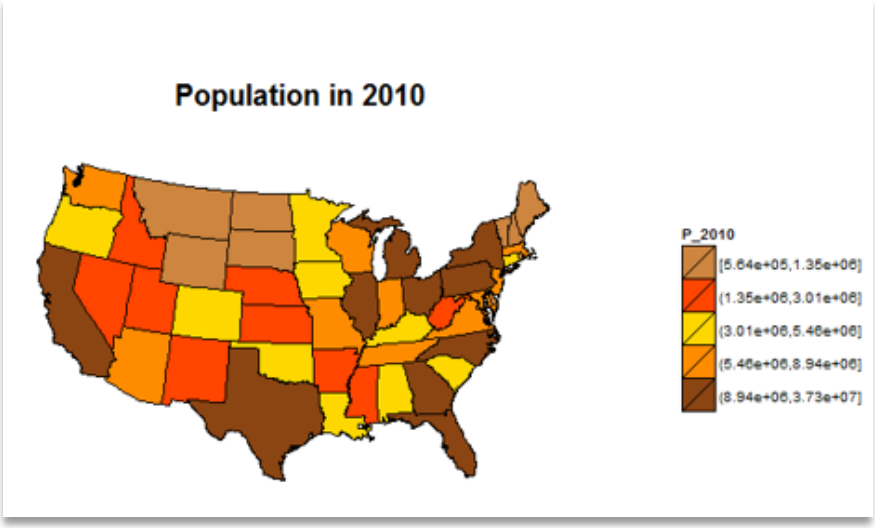
6.1 HYPOTHESIS 1

Factors that led to population rise/ fall in the period 2000-2010 are same for 2010-2014.

A simple choropleth map (using ggplot).

A slight difference in the North Eastern region.





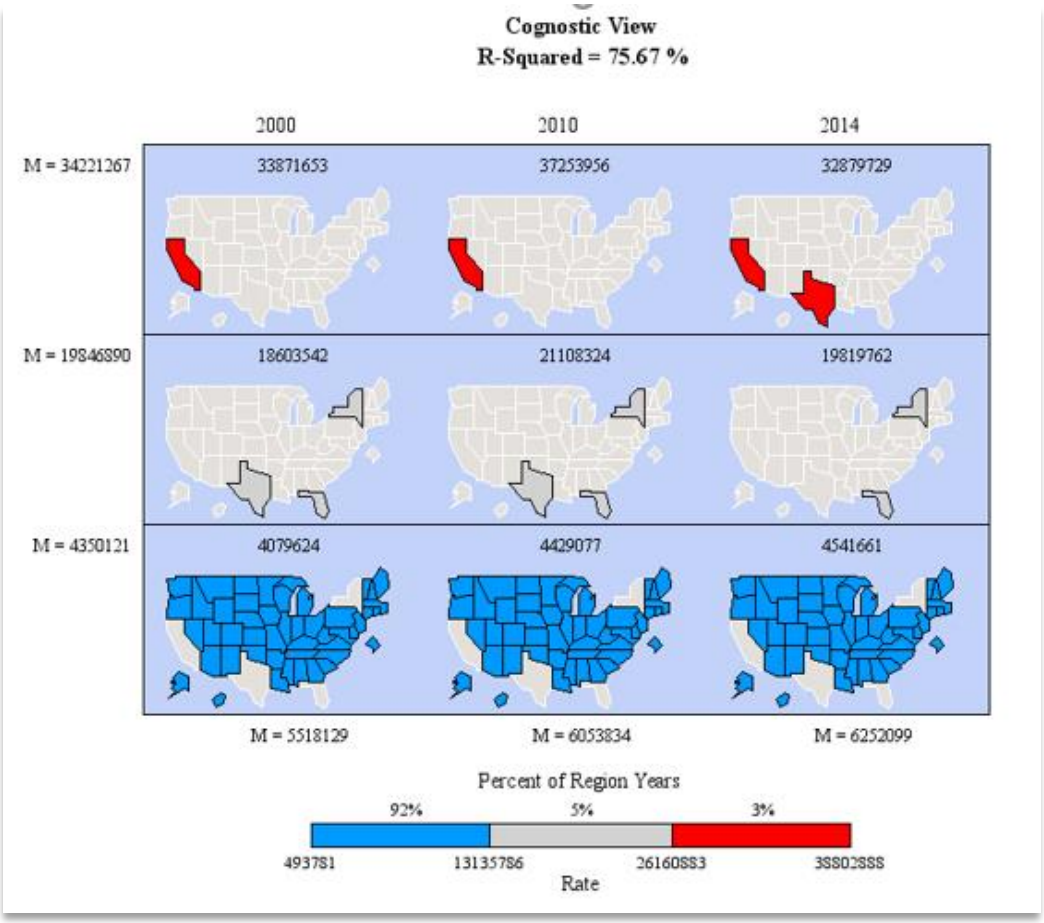
Not a very good resource for comparison.

6.1.1 TC Maps

6.1.1.1 Congnostic View

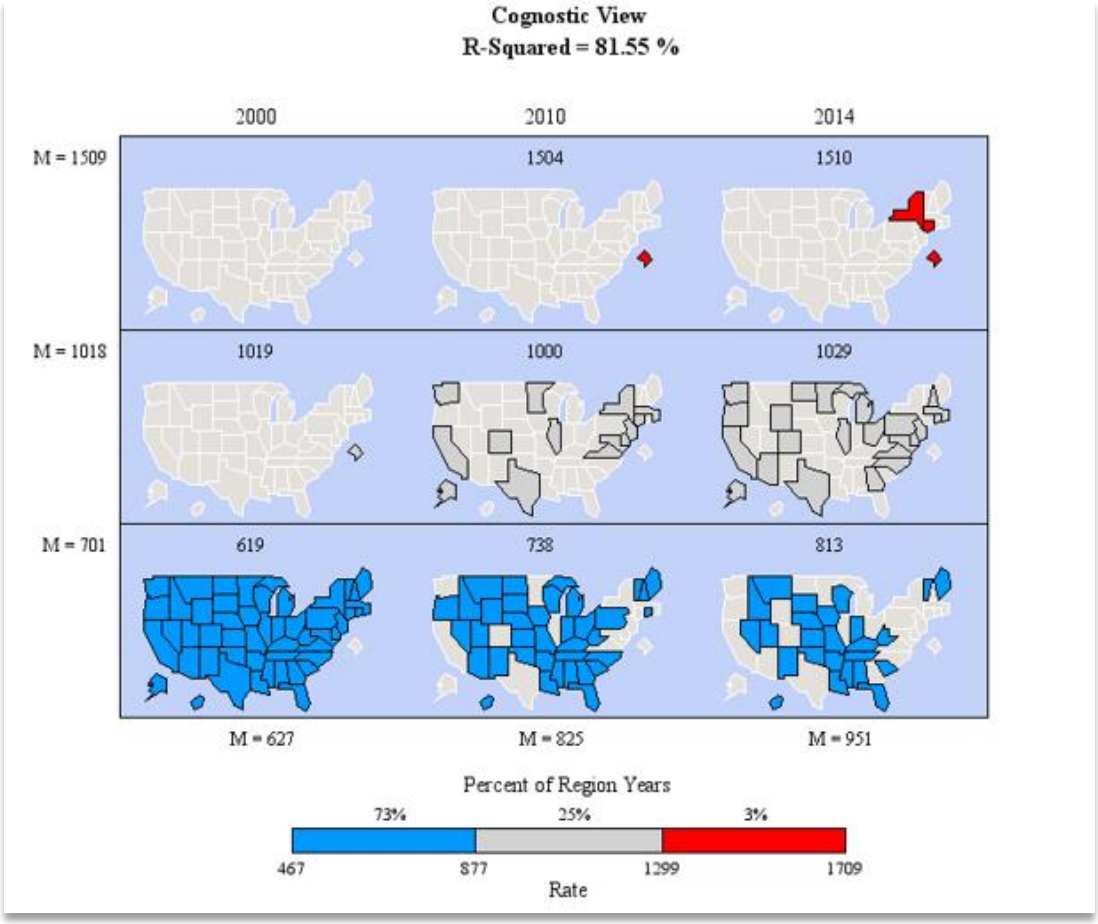
Cognostic view in TCMaps has the functionality of grid search algorithm. It sets the slider thresholds and partitions the data into low, middle and high for each period.

- ◎ High R-Squared indicates the stability of values over time relative to the partition.
- ◎ High Threshold has only 3% of data.
- ◎ Medium Threshold holds about 5% of data
- ◎ Low Threshold holds the maximum data of 92%



Cognostic View of Average weekly wages for years 2000, 2010, 2014.

1. New York has made its way from low (in 2000) to high threshold (in 2014)

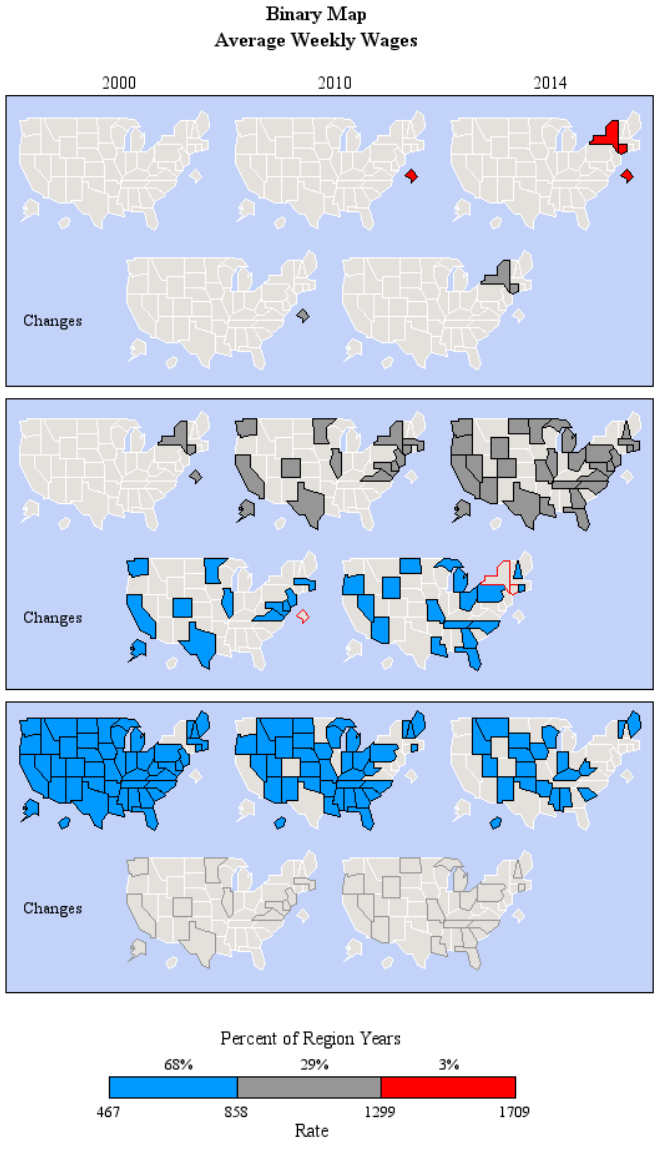


6.1.1.2 Binary Map

Average Weekly Wages for years 2000, 2010, 2014.

This map explains us two things.

1. Detailed view of changes happened between years.
2. Divides the data based on the slider. (it helps in categorizing the states based on the slider limits)



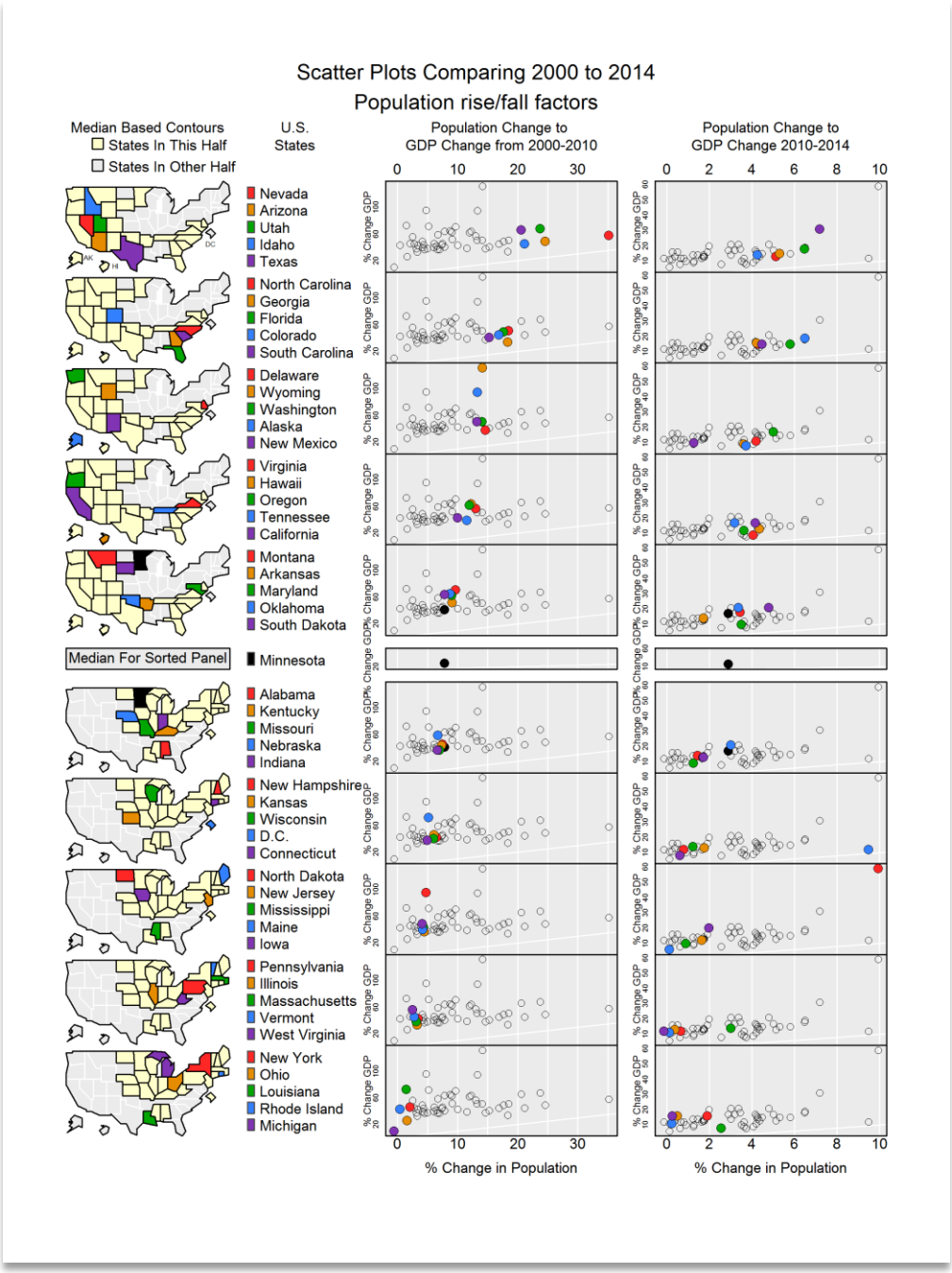
Average Weekly Wage for years 2000, 2010, 2014

6.1.2 Micromaps

6.1.2.1 Scatter Plot

Scatter plots are extremely helpful is analyzing a trend. In this case a clear trend has been observed.

- ❑ Percent Change in Population to Percent Change in GDP.
- ❑ Linear Trend (Except for North Dakota and Connecticut).

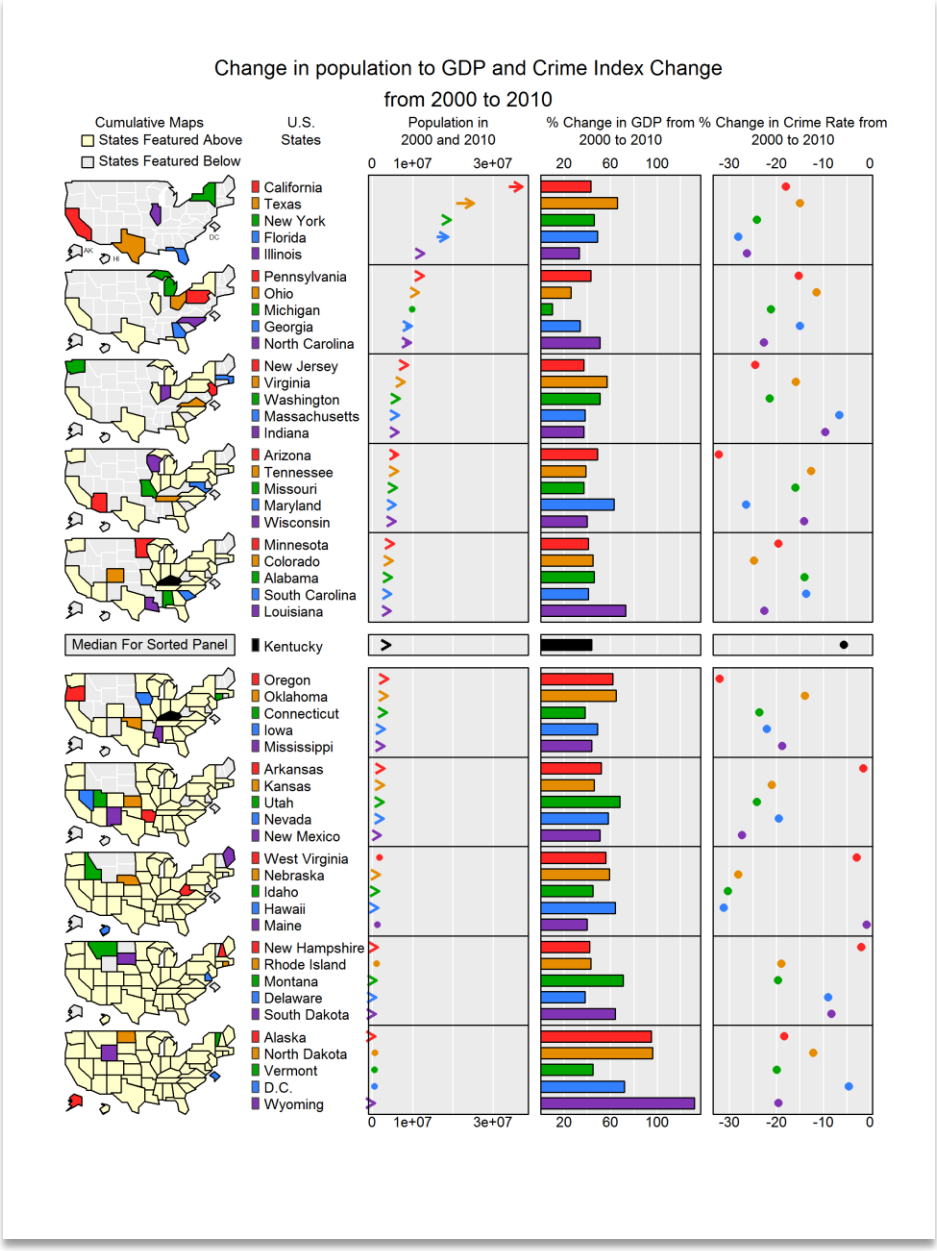


6.1.2.2 Cumulative Map

- ❑ Change in population to change in GDP and Crime Index.
- ❑ Though Wyoming has the least increment in population, it contributed highest to the economy.
- ❑ Maine has highest crime rate.

Anyways this doesn't show us any significant linear trend.

So, I've built a linear regression model to further illustrate this linear relationship.



6.1.3 Linear Regression

6.1.3.1 Linear Regression of Population to several factors – Year 2000

R-Squared – 0.9927 (Very high correlation)

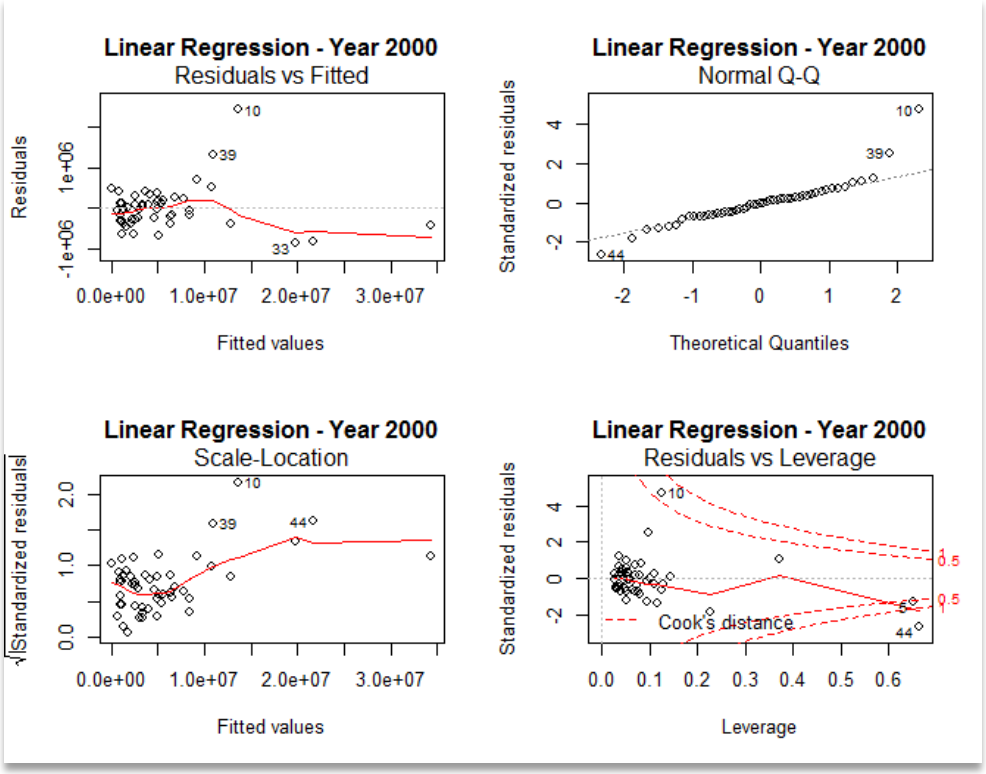
P-Value – 0.00000000000000022

Regressand

Population

Regressors

GDP, Crime, Birth Rate, Death Rate, Net Migration, Income Median, Average Weekly Wage



6.1.3.2 Linear Regression of Population to several factors – Year 2010

R-Squared – 0.9996 **(Very high correlation)**

P-Value – 0.00000000000000022

Left Plot 1

This plot shows the residuals (the vertical distance from a point to the regression line) versus the fitted values. Smooth Curve (red line) very close to gray dashed line. This is expected when the presence of high correlation there between variables.

Right Plot 1

This plot evaluates that the errors are normally distributed. If, the points lie very close to the dashed line, then errors are normally distributed, which is true in our case.

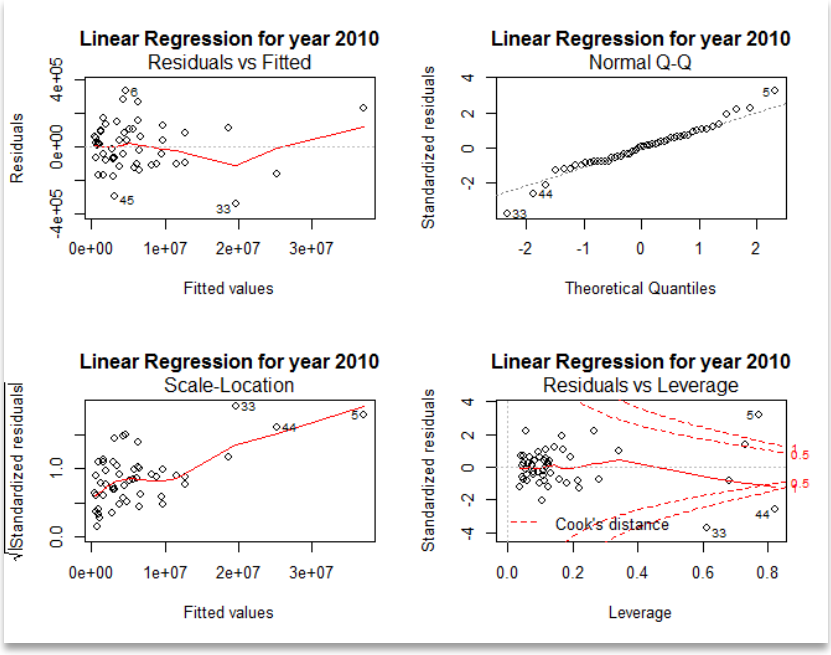
Left Plot 2

No homoscedasticity (the variance in the residuals doesn't change as a function of x) as the red line is not exactly flat.

Right Plot 2

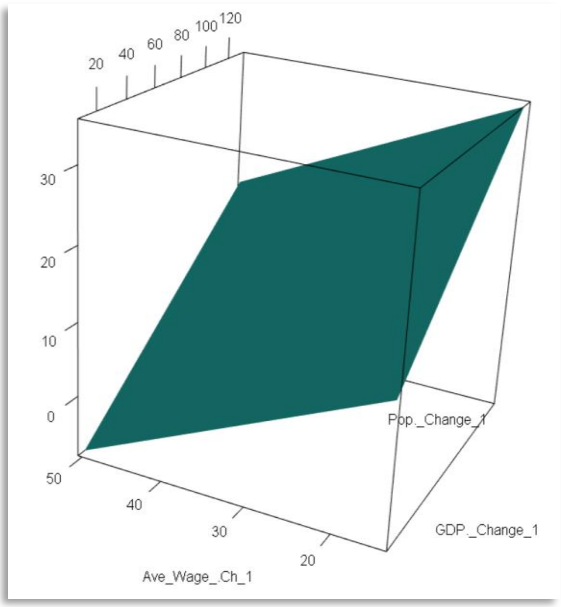
Standardized residuals centered around zero.

Cooks distance not > 0.5.

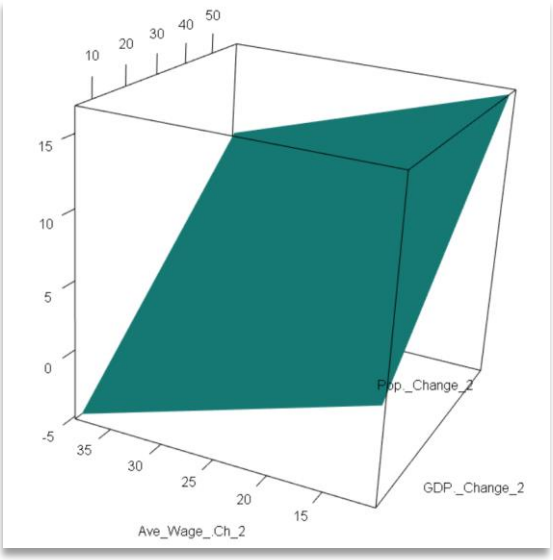


6.1.4 Plotting regression output using VISREG2D

Change in Population to Change in GDP and Weekly Wages 2000-2010



Change in Population to Change in GDP and Weekly Wages 2010-2014



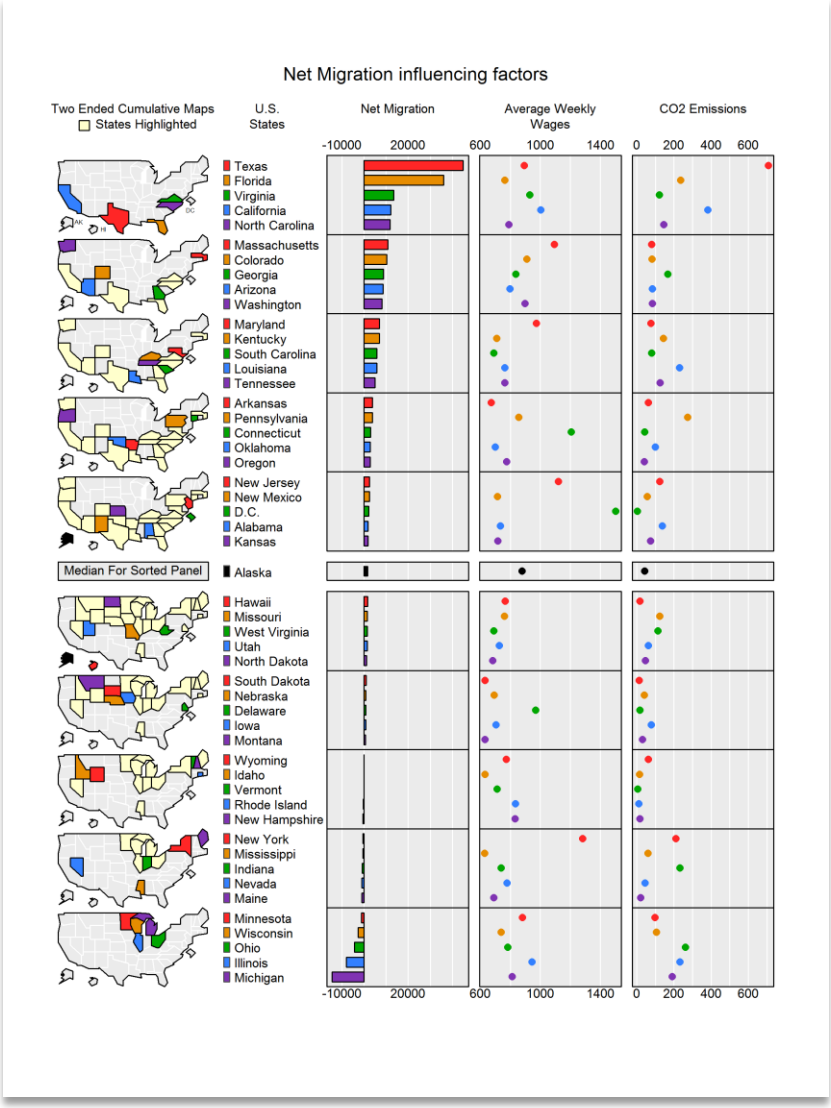
These plots tell us that the plane built using the variables "% Change in Weekly Wages" and "% Change in GDP" are both similar, which means a linear relationship is existing.

6.2 HYPOTHESIS 2

Rise in weekly wages and CO2 emissions are significant factors for rise/ fall in Net Migration.

6.2.1 Net Migration influencing factors using Micromaps

- ❑ CO2 Emissions has more linearity to Net Migration, compared to Average Weekly Wages.
- ❑ Texas has highest migration as CO2 emissions are high.
- ❑ This doesn't show us any significant linear trend.
- ❑ So, I've built a linear regression model to further illustrate this linear relationship.



6.2.2 Linear Regression

6.2.2.1 Linear Regression of Migration to several factors – Year 2010

R-Squared – 0.4266

P-Value – 0.00009685

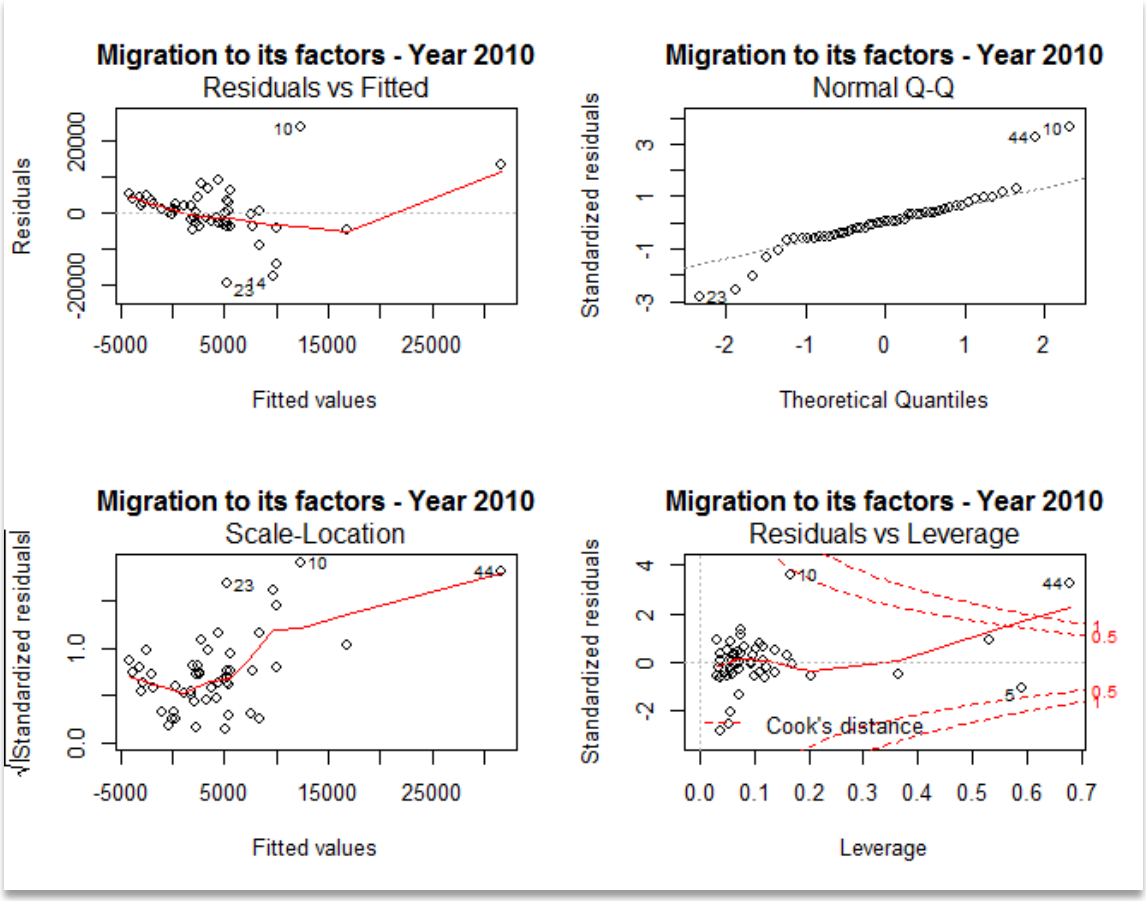
Regressand

Net Migration

Regressors

Crime, Income Mean, Average Weekly Wage, Employment, CO2 Emissions

The red line in the Residuals Vs Fitted plot lay close to the grey dashed line, which is expected in the case of high correlation.



6.2.2.2 Linear Regression of Migration to several factors – Year 2014

R-Squared – 0.4688

P-Value – 0.00001977

Left Plot 1

Smooth Curve (red line) not very close to gray dashed line.

Right Plot 1

Points lie very close to the dashed line except for the edges.

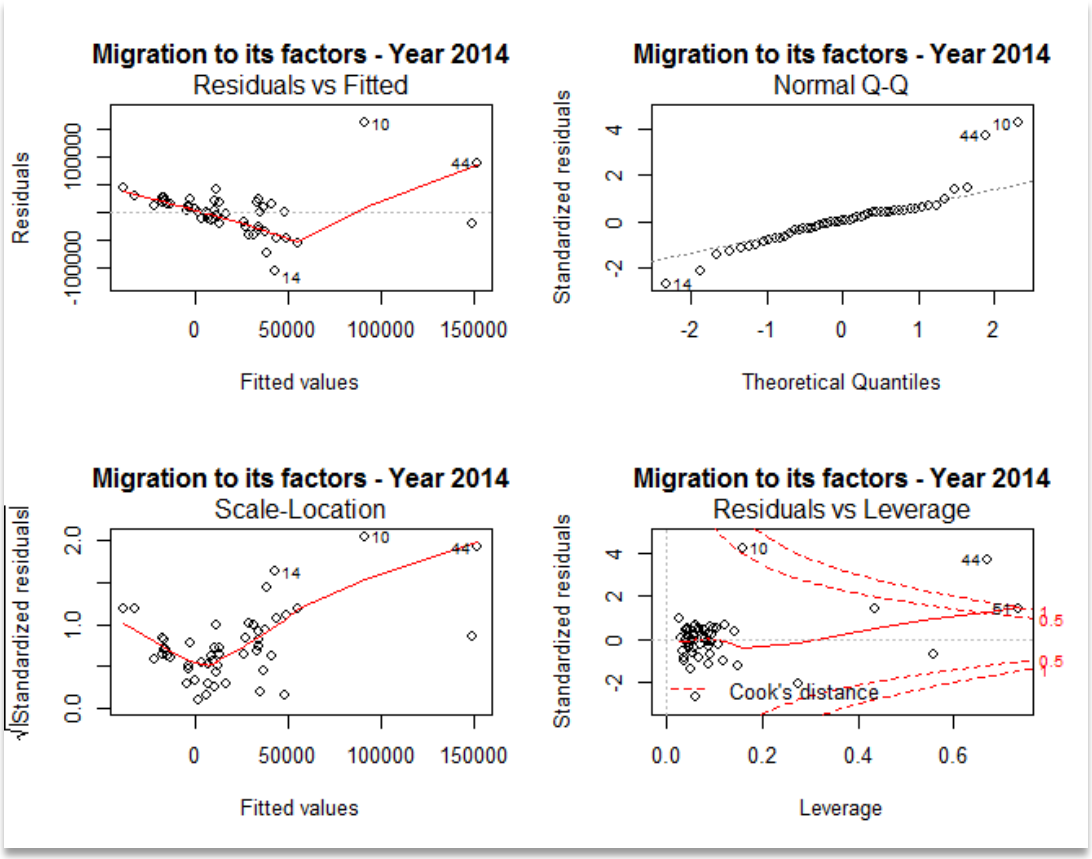
Left Plot 2

No homoscedasticity, as the red line is not flat.

Right Plot 2

Standardized residuals centered around zero.

Cooks distance not > 0.5.



7 CONCLUSION

- Factors for population growth in the period 2000-2010 are also influencing for the period 2010-2014.
- I can partly conclude that, Average Weekly Wages, CO2 Emissions and Income mean are playing a significant role in the rise/ fall of Net Migration.