Will *you* pay off your loans during a recession?

A Statistical Exploration of Relevant Factors that Eroded Credit Payments in 2008

Introduction

The purpose of this analysis is to determine whether certain variables can provide relevant information about credit delinquency rates in the event of a housing crisis/general economic downturn. In 2008, due to overleveraging by large institutions in asset-backed-securities among other factors, the real estate market in the United States crumbled. When such a collapse takes place, people are forced to rework their spending habits. What's more, such events can lead to borrowers not being able to pay off their loans. This exploration will provide statistical evidence suggesting that certain factors can increase fragility in lending markets. Though there continue to be credible claims that severe economic decline has yet to unfold, some government officials and economists have suggested that the United States is headed towards another recession (if not in one already). This project will assist in displaying which areas are most at risk to see credit delinquency rates rise during the upcoming economic downturn. Three questions that summarize the intent of this analysis are as follows:

1. In times of increased economic distress, which areas (whether regions or states) are most adversely affected?
    a. Understanding which regions get hit the hardest during a recession when it comes to its inhabitants ability to pay off their loans is crucial if further exploration is to take place. That said, just identifying regions that performed poorly last time around is NOT sufficient in portraying what will happen in the future. That said, examining these areas most impacted will give a good starting point before the individual variables are tested for statistical significance.
2. Which factors led to these areas being so drastically affected?
    a. This is the meat of the exploration. The statistical analysis done to determine the importance of the explanatory variables will unlock key insights as to why one area may have seen higher levels of default than others. Answering this question will provide information that can be applied to where the country stands currently.
3. Knowing the factors that are statistically relevant, which states (or districts) are most susceptible to a rise in credit delinquency rates currently?
    a. This is where the relevance of the analysis is most evident. Using the answers to the previous two questions, one can surmise which areas might be vulnerable during an economic downturn.

A recession will undoubtedly create numerous obstacles for citizens all over the country, and this analysis peels back the curtain on those who may be at high risk of defaulting on their loan payments. Though the project might be useful for a lending firm who is looking to take their business out of risky areas, it is even more useful for an individual's own extrapolations and decision making. Lending standards are much stricter now than they were back in 2008, so although firms may glean some insight from this analysis, it is more relevant to individuals. For example, if it was determined that household income has nothing to do with credit delinquency rates, but population does, that might be relevant for a student looking at potential places to live after graduation. Or what if the data suggested that Midwest areas see high increases in default rate during a recession? Would students want to live there going into an economic crisis? Rather than focusing on the corporate benefits, the project provides insights to those looking to get ahead of the curve and plan for their future, taking into account the most relevant factors that might reduce the likelihood of defaulting on loans. Lending standards are much tighter than they

were in 2008, but understanding the factors that led to so many people being unable to pay off their loans is still crucial.

Data Summary

We chose to collect data that represented the entire population of interest (US citizens) split up into observations categorized by state (or DC). In order to collect data that reflects the entire nation, we had to mainly use data sets provided by government or government backed organizations. Details and drawbacks of these sources are listed below.

The Census Bureau - We used data provided by the Census Bureau on poverty/income and statistical groupings of states/counties in order to acquire the data for two of our explanatory variables. The data pertaining to poverty and income were collected in 2008 through household surveys and programs where they measure poverty based on cash resources and the statistical groupings of states and counties was published in 1994 in the Census Bureau's geographic areas reference manual. For the most part, Census data is extremely reliable, but no census is 100% accurate and one potential drawback is that the census has used the same definition for poverty since the mid 1960s. Perhaps an updated definition would be more applicable to our population. Even though the geographic data was published a while ago, since we are only using this data to group states by region this should not present an issue. The data for both of these data sets– while not having an explicitly stated reason for collection– were collected in order to fulfill the Census Bureau's main goal: to serve the nation and help the nation and other communities make decisions.

The National Mortgage Database - We used data provided by the National Mortgage Database on mortgage/credit delinquency in order to acquire the data for our response variable. The data was collected from 2008 to 2021 through nationwide credit repositories who collected it because the statistic reflects the health of the overall economy. One drawback specific to this data source is that the data started being collected in 2008, so their accuracy and methodology of data collection may have evolved in a positive way over the last 15 years.

National Center for Education Statistics - We used data provided by National Center for Education Statistics on household income and highschool graduation rates in order to acquire the data for two of our explanatory variables. The data pertaining to household income was collected from 1990-2010 and the data pertaining to high school graduation rates was collected from 1990-2007. No specific collection method was outlined for either of these datasets but it can be presumed that it was collected through survey as this is the main data procurement method listed in the Nation Center for Education Statistic's about us tab. The data was collected for the purpose of fulfilling a congressional mandate. One drawback of this data is that it was also collected through a census so it may be relatively inaccurate.

Bureau of Labor Statistics - We used data provided by the Bureau of Labor Statistics on unemployment in order to acquire the data for one of our explanatory variables. The data was collected in the years 2007 and 2008 through state postings, sampling, and modeling and was collected to better inform political and community decisions. Little to no drawbacks other than normal error and some error caused by rounding.

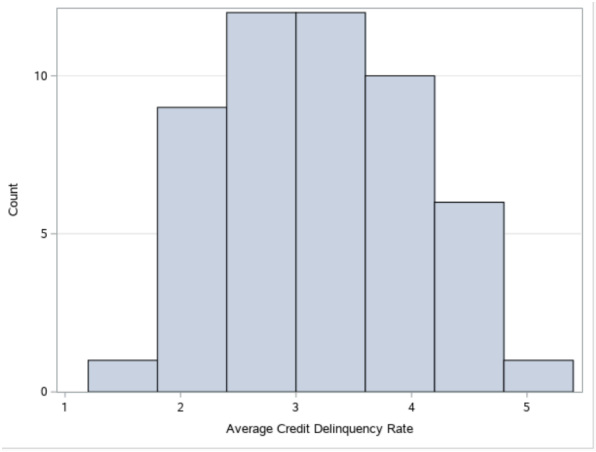
Federal Election Commision - We used data provided by the Federal Election Commision in order to acquire the presidential election outcomes by state for 2008 which we used as one of our explanatory variables. The data is collected and stored every year by the Federal Election Commision who gets the data through polling to be later used as a historical reference. One drawback is that the polls are subject to nonresponse bias.

Federal Bureau of Investigation - We used data provided by the Federal Bureau of Investigation in order to acquire the drug abuse violation statistics for 2008 which we used as one of our explanatory variables. The data was the raw amount of drug abuse violations in every state and DC and to account for population we divided the amount of drug abuse violations in every state by the states population. The data was collected by the FBI for historical reference purposes and was collected/compiled through data reported to the FBI through state governments. One potential drawback of this data is its data on the District of Columbia. They weren't able to get a report directly from a state government so the data listed here– as explained at the bottom of the data table– is data procured from the Metro Transit Police. This amount of drug abuse violations is weirdly low compared to the population of the DC area so it may be of concern as we proceed.

Exploratory Data Analysis

Response Variable - Average Credit Delinquency Rate

We want to visually and numerically describe the response variable so that we have an understanding of the distribution.



Analysis Variable : Average Credit Delinquency Rate				
Mean	Std Dev	Minimum	Maximum	N
3.1994118	0.8633410	1.5400000	5.2500000	51

Graphical Summaries

[Figure 1] We want to visually see how average credit delinquency rates correlate with median house income as we think that median house income could be a useful predictor. In addition, we want to see if a state being Democratic or Republican affects these two variables.

[Figure 2] We want to visualize how credit delinquency rates differ based on what region of the country the state is located in.

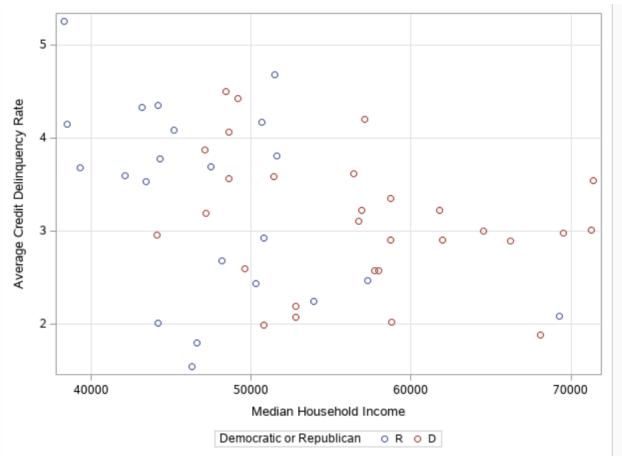


Figure 1

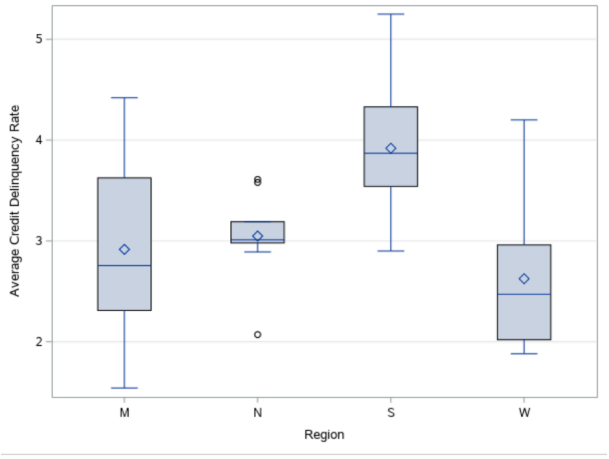


Figure 2

Numerical Summaries

Based on what we know about credit delinquency rates, we hypothesized that a state's unemployment rate and highschool graduation rate would be correlated with the state's credit delinquency rate. We test this hypothesis by looking at the mathematical correlation between these variables individually and credit delinquency rate.

Pearson Correlation Coefficients, N = 51	
	Average Credit Delinquency Rate
Unemployment Rate	0.52730

Figure 3

Pearson Correlation Coefficients, N = 51	
	Average Credit Delinquency Rate
Highschool Graduation Rate	-0.50033

Figure 4

Conclusion

In our exploratory data analysis, we used graphical and numerical summaries to attempt to determine if there was any relationship between our explanatory variables and our response variable of average credit delinquency rate in order to eventually determine if these explanatory variables will be useful predictors for the response. Prior to analyzing the explanatory variables, we displayed the response variable both graphically and numerically. The histogram of the response variable shows a somewhat symmetric normal distribution without any obvious skew. In the 51 observations that were analyzed, it was found that the mean of response variable is delinquency rate of approximately 3.2, while the minimum is a rate of approximately 1.54.

For the qualitative variable of political party (Democratic or Republican) and the quantitative variable of median household income, we created a scatter plot to identify whether or not median household income would be useful in predicting average credit delinquency rates by observing the correlation between the two variables. The political affiliation of each observation was also represented in the scatterplot to see if there is any correlation. Visually, we found that there seems to be a negative linear relationship between household income and average credit delinquency rates, however the relationship looks very weak. For the case of political parties, we

noticed that republicans tended to have a higher household income than the democrats in our data, and seemed to have a lower credit delinquency rate on average than democrats in the data. Although the relationships between the variables are too weak to draw any definite conclusions, it seems that a relationship does exist between them based on the graphical summary we created. This suggests that both Political Party and Median Household Income are relevant enough to continue with further statistical analysis.

For the qualitative variable of region, we used a graphical representation of 4 boxplots, one for each region, in order to compare the values contained in the 5 number summaries of each region with regard to average credit delinquency rate. It was found that the region South had the highest credit delinquency rate of the four regions, as determined by the mean, median, and maximum which were all clearly larger than the other regions (approximately 3.80 and 3.90 for the mean and median), and the West had the lowest delinquency rate as determined by the mean and median. As seen by the means and medians of the four regions, there is a distinction in the average credit delinquency rate depending on the region where the state is located, this suggests that the explanatory variable, Region, is relevant to include in further exploration.

For the quantitative variables of a state's high school graduation rate and unemployment rate, we determined the correlation coefficient between each of the two variables and our response variable of average credit delinquency rate to see if there is any relationship between them. The correlation coefficient between graduation rate and credit delinquency rate was 0.52730, which allows us to determine that there is a positive correlation between the two variables, however it is not a strong relationship. The correlation coefficient of -0.50033 that was obtained by unemployment rate and credit delinquency rate shows that there is a negative correlation between the two variables, however it is not very strong, similar to the previous correlation coefficient described. As unemployment rises, so does credit delinquency. The inverse is true with high school graduation rates (which makes intuitive sense). As we hypothesized, relationships do seem to exist between graduation/unemployment rate and average credit delinquency rate, justifying them for inclusion in further analysis.

For the quantitative explanatory variable of poverty percentage, we believe that there could be a relationship between more impoverished families and credit delinquency rate, as families living in greater poverty are more likely to default on their credit card payments. For the quantitative variable of population, we hypothesized that states with larger populations would have more opportunities for education about financial literacy, so it is possible that they would be less likely to have a high delinquency rate. For the explanatory variable of crime rate, we hypothesized that areas with high rates of crime would also be areas of greater poverty, meaning that there could be a positive relationship between crime rate and credit card delinquency rate.

In order to continue with the model building process we would fit the model and determine its adequacy. Then progress by dropping any variables deemed statistically insignificant followed by running further tests to determine the strength of each variable. Testing for interactions would be necessary to recognize the prevalence of limiting factors such as multicollinearity.

Appendix & Data Dictionary

Obs	Name	Average Credit Delinquency Rate	Median Household Income	Unemployment Rate	Highschool Graduation Rate	Poverty Percentage	Population	Democratic or Republican	Region	Drug Abuse Level	DummyR	DummyS	DummyW	DummyN	DummyMedium	DummyHigh
1	Alabama	4.33	43200	5	69	15.9	4,718,206	R	S	Low	1	1	0	0	0	0
2	Alaska	2.08	69300	6.7	69.1	9.2	687,455	R	W	Low	1	0	1	0	0	0
3	Arizona	3.8	51600	5.5	70.7	14.7	6,280,362	R	W	Medium	1	0	1	0	1	0
4	Arkansas	3.68	39300	5.1	76.4	17.3	2,874,554	R	S	Medium	1	1	0	0	1	0
5	California	3.22	61800	7.2	71.2	13.3	36,604,337	D	W	High	0	0	1	0	0	1
6	Colorado	2.57	57700	4.9	75.4	11.2	4,889,730	D	W	Low	0	0	1	0	0	0
7	Connecticut	2.98	69500	5.7	82.2	9.1	3,545,579	D	N	Medium	0	0	0	1	1	0
8	Delaware	3.35	58700	4.8	72.1	10.3	883,874	D	S	Medium	0	1	0	0	1	0
9	District of Columbia	2.9	58700	7	56	16.9	580,236	D	S	Low	0	1	0	0	0	0
10	Florida	4.5	48400	6.2	66.9	13.3	18,527,305	D	S	High	0	1	0	0	0	1
11	Georgia	4.68	51500	6.2	65.4	14.7	9,504,843	R	S	Low	1	1	0	0	0	0
12	Hawaii	1.88	68100	3.9	76	9.3	1,332,213	D	W	Low	0	0	1	0	0	0
13	Idaho	2.68	48200	4.9	80.1	12.5	1,534,320	R	W	Low	1	0	1	0	0	0
14	Illinois	3.22	56900	6.5	80.4	12.2	12,747,038	D	M	Low	0	0	0	0	0	0
15	Indiana	4.06	48600	5.9	74.1	12.9	6,424,806	D	M	Low	0	0	0	0	0	0
16	Iowa	2.59	49600	4.1	86.4	11.4	3,016,734	D	M	Low	0	0	0	0	0	0
17	Kansas	2.92	50800	4.4	79	11.3	2,808,076	R	M	Low	1	0	0	0	0	0
18	Kentucky	3.59	42100	6.4	74.4	17.3	4,289,878	R	S	Low	1	1	0	0	0	0
19	Louisiana	3.77	44300	4.6	63.5	17.6	4,435,586	R	S	Medium	1	1	0	0	1	0
20	Maine	3.19	47200	5.4	79.1	12.6	1,330,509	D	N	Medium	0	0	1	0	1	0

Variable Name	Variable Description	Units
Average Credit Delinquency Rate	This is the response variable and it provides information on the percentage of loans within a financial institution's loan portfolio whose payments are delinquent (past due). The data was sourced from The National Mortgage Database and gives a value for each state's average credit delinquency rate in 2008. This variable provides a useful gauge of the financial wellbeing of an observation's inhabitants.	The units for our response variable are integers in percentage terms. For example, a value of 5.1 would be the value correlated with an observation that has an average credit delinquency rate of 5.1% This is a quantitative variable.
Median Household Income	This is one of the explanatory variables and it provides information on the average income of an observation's inhabitants (blocking families together into households). The data is provided for each of the 51 observations and is from 2008. This data was sourced by NCES and is relevant as it provides a metric for measuring absolute units of income.	The units here are dollars in nominal terms. For example if the median household income variable is equal to 50,000, that means that the observation correlated with that statistic has a median household income of $50,000. This is a quantitative variable.
Unemployment Rate	This data was sourced by the National Bureau of Labor Statistics and gives details about the percentage of persons unemployed in a specific observation. The data is sourced for each of the 51 observations and is from 2008. This variable provides relevant information about the health of the labor force within each of the observations.	The units here are in percentage terms. For example, a value of 3.9 would be correlated with an unemployment rate of 3.9%. This is a quantitative variable.
High School Graduation Rate	This data was sourced by the NCES and is from the year of 2007-2008. It provides relevant information about the status of	The units here are in percentage terms once again. For example, a value of 70.1 would be correlated with a high school graduation rate of 70.1%. This is a quantitative variable.

Variable Name	Variable Description	Units
	academic literacy in each of the observations right before the crisis unfolded.	
Poverty Percentage	This variable was sourced from the United States Census Bureau and is for each observation in 2008. The poverty threshold is adjusted for family size and composition to determine who is in poverty. If the family's income is below that threshold then they are deemed to be impoverished. This data is relevant because there may be a connection between likelihood to miss loan payments and whether or not poverty is prevalent in our observation.	The units used to describe poverty are in percentage terms. For example, a value of 34.5 would represent an observation with 34.5% of its inhabitants deemed to be below the poverty threshold. This is a quantitative variable.
Population	This data is sourced from the US Census Bureau in 2008 and gives a number of persons in each observation. The data was included in an effort to determine whether larger states (or states with higher populations) have better programs in place for financial literacy, among other reasons a larger state might have a noticeable difference in credit delinquency rates than a state of a much smaller size.	The units here are in number of people. For example, a population of 10,999 has 10,999 people within that observation.
Political Party	This variable was sourced by the Federal Elections Commission from 2008. The data shows whether an observation voted Red or Blue during the 2008 presidential election. It is particularly important for this analysis because it helps determine whether or not the financial literacy of one party is greater than that of the other (all else equal).	The levels for this variable are as follows: DummyR = 1 if Republican, 0 if otherwise (Democrat). Since none of the observations voted independent, it is easy to assign an observation with one of these two levels. This is a qualitative variable.

Variable Name	Variable Description	Units
Region	This variable was sourced by the US Census Bureau and categorizes each observation into one of 4 regions: South, West, Midwest and North. The groupings were done in the 1990s, but even though state governments disagree about the categorizations, these observations are the same as they were 30 years ago. Observations don't enter and exit regions, they are either in the region or they are not. This variable is relevant in the analysis because it can be used in tandem with the response variable to determine whether or not region affects financial literacy/default rates.	The base level here is Midwest. There are 4 levels total. DummyS = 1 if south, 0 otherwise DummyW = 1 if west, 0 otherwise DummyN = 1 if north, 0 otherwise This is a qualitative variable.
Drug-Related Crime level	The drug-related crime level was calculated by the number of drug violations divided by population multiplied by 100. This provides an estimate of the percentage of drug-related crimes in each observation. Areas with more crime may be less financially literate as a whole. The crime variable was sourced from the Federal Bureau of Investigation (FBI) and is from the relevant time period of 2008. We hypothesize that drug violations are associated with poor decision-making, which might lead to lower financial literacy and generally higher rates of credit delinquency.	The base level is 'low' drug-related crimes. There are 3 levels total. DummyMedium = 1, If 0.4 <= Crime <= 0.7, 0 otherwise DummyHigh = 1, If Crime > 0.7, 0 otherwise

Works Cited

[1]"Averaged Freshman Graduation Rates for Public Secondary Schools, by State or Jurisdiction: Selected Years, 1990-91 through 2007-08." *National Center for Education Statistics (NCES) Home Page, a Part of the U.S. Department of Education*, https://nces.ed.gov/programs/digest/d10/tables/dt10_112.asp?referrer=report.

[2]Bureau, US Census. "State and County Estimates for 2008." *Census.gov*, 8 Oct. 2021, https://www.census.gov/data/datasets/2008/demo/saipe/2008-state-and-county.html.

[3]"Federal Elections 2008." *FEC.gov*, https://www.fec.gov/introduction-campaign-finance/election-and-voting-information/federal-elections-2008/.

[4]*Geographic Regions*. https://www2.census.gov/geo/pdfs/reference/GARM/Ch6GARM.pdf.

[5]"Median Household Income, by State." *National Center for Education Statistics (NCES) Home Page, a Part of the U.S. Department of Education*, https://nces.ed.gov/programs/digest/d11/tables/dt11_025.asp.

[6]"Mortgages 30-89 Days Delinquent." *Consumer Financial Protection Bureau*, https://www.consumerfinance.gov/data-research/mortgage-performance-trends/mortgages-30-89-days-delinquent/.

[7]“REGIONAL AND STATE UNEMPLOYMENT, 2008 ANNUAL AVERAGES .” *BLS*, 27 Feb. 2009, https://www.bls.gov/news.release/archives/srgune_02272009.pdf.

[8]Abadi, Mark. “Even the US Government Can't Agree on How to Divide up the States into Regions.” *Business Insider*, Business Insider, https://www.businessinsider.com/regions-of-united-states-2018-5.

[9]Owyang, Michael T, and Ashley Stewart. “US Recession: What Key Economic Indicators Say.” *Saint Louis Fed Eagle*, Federal Reserve Bank of St. Louis, 26 Sept. 2022, https://www.stlouisfed.org/on-the-economy/2022/sep/us-recession-what-key-economic-indicators-say.

[10]Rugaber, Christopher. “Federal Reserve Chair Jerome Powell Says Inflation Fight May Cause a Recession.” *PBS*, Public Broadcasting Service, 22 Sept. 2022, https://www.pbs.org/newshour/economy/federal-reserve-chair-jerome-powell-says-inflation-fight-may-cause-a-recession.

[11] “Crime in the US 2008.” *Federal Bureau of Investigation*, https://ucr.fbi.gov/crime-in-the-u.s/2008

[12] Bureau, US Census. “About Poverty in the U.S. Population.” *Census.gov*, 13 Sept. 2022, https://www.census.gov/topics/income-poverty/poverty/about.html.

Average Credit Delinquency Rate	This is the response variable and it provides information on the percentage of loans within a financial institution's loan portfolio whose payments are delinquent (past due). The data was sourced from The National Mortgage Database.	**Quantitative** (Continuous)
Median Household Income	This is one of the explanatory variables and it provides information on the average income of an observation's inhabitants (blocking families together into households). The data is provided for each of the 51 observations and is from 2008. This data was sourced by NCES.	**Quantitative** (Continuous)
Unemployment Rate	This data was sourced by the National Bureau of Labor Statistics and gives details about the percentage of persons unemployed in a specific observation. The data is sourced for each of the 51 observations and is from 2008.	**Quantitative** (Continuous)
High School Graduation Rate	This data was sourced by the NCES and is from the year of 2007-2008. Rate at which high school students completed their programs.	**Quantitative** (Continuous)
Poverty Percentage	This variable was sourced from the United States Census Bureau and is for each observation in 2008. The poverty threshold is adjusted for family size and composition to determine who is in poverty. If the family's income is below that threshold then they are deemed to be impoverished.	**Quantitative** (Continuous)
Population	This data is sourced from the US Census Bureau in 2008 and gives a number of persons in each observation.	**Quantitative** (Discrete)
Political Party	This variable was sourced by the Federal Elections Commission from 2008. The data shows whether an observation voted Red or Blue during the 2008 presidential election.	**Qualitative** The base level is Democrat. There are 2 levels total. **DummyR** = 1 if Republican, 0 if otherwise (Democrat).
Region	This variable was sourced by the US Census Bureau and categorizes each observation into one of 4 regions: South, West, Midwest and North.	**Qualitative** The base level here is Midwest. There are 4 levels total. **DummyS** = 1 if south, 0 otherwise **DummyW** = 1 if west, 0 otherwise **DummyN** = 1 if north, 0 otherwise
Drug-Related Crime level	The drug-related crime level was calculated by the number of drug violations divided by population multiplied by 100. This provides an estimate of the percentage of drug-related crimes in each observation.	**Qualitative** The base level is 'low' drug-related crimes. There are 3 levels total. **DummyMedium** = 1 if $0.4 \leq \text{Crime} \leq 0.7$, 0 otherwise **DummyHigh** = 1 if Crime > 0.7, 0 otherwise

$$E(Average\ Credit\ Delinquency\ Rate) = \beta_0 + \beta_1 Median\ Household\ Income + \beta_2 Unemployment\ Rate + \beta_3 High\ School\ Graduation\ Rate + \beta_4 DummyR + \beta_5 DummyS + \beta_6 DummyW + \beta_7 DummyN + \beta_8 DummyMedium + \beta_9 DummyHigh$$