NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

CRIME TREND PREDICTION USING REGRESSION MODELS FOR SALINAS, CALIFORNIA

by

Jarrod S. Shingleton

June 2012

Thesis Co-Advisors: Bard Mansager
Hong Zhou

Approved for public release; distribution is unlimited

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE		Form Approved OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	June 2012	Master's Thesis

4. TITLE AND SUBTITLE	5. FUNDING NUMBERS
Crime Trend Prediction Using Regression Models for Salinas, California	
6. AUTHOR(S) Jarrod Shingleton	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A	**10. SPONSORING/MONITORING AGENCY REPORT NUMBER**

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. IRB Protocol number _____.

12a. DISTRIBUTION / AVAILABILITY STATEMENT	12b. DISTRIBUTION CODE
Approved for public release; distribution is unlimited.	A

13. ABSTRACT (maximum 200 words)

Salinas, California has been battling an above average crime rate for over 30 years. This is due primarily to two rival gangs in Salinas: the Norteños and the Sureños. The city and the surrounding community have implemented many methods to mitigate the crime level, from community involvement to the inception of a gang task force. As of yet, none of the efforts have had long-lasting effects.

In a 2009 thesis, Jason A. Clarke and Tracy L. Onufer postulated that various socio-economic variables are influential on the crime level in Salinas. They characterized "crime" as a summation of homicides, assaults and robberies reported. Their thesis determined that "to lower overall violence levels, officials in Salinas should focus on: reducing the unemployment rate, the number of vacant housing units, and the high school dropout rate; and increasing the high school graduation rate and average daily attendance."

A deeper examination of the data could lead not only to assumptions about how to lower crime rates, but also to a means of predicting future crime rates by using various methods of multiple value regression.

14. SUBJECT TERMS Salinas, Crime, Regression	15. NUMBER OF PAGES
	89
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UU

NSN 7540-01-280-5500

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release; distribution is unlimited

CRIME TREND PREDICTION USING REGRESSION MODELS FOR SALINAS, CALIFORNIA

Jarrod S. Shingleton
Captain, United States Army
B.S., Sam Houston State University, 2003

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN APPLIED MATHEMATICS

from the

NAVAL POSTGRADUATE SCHOOL
June 2012

Author: Jarrod S. Shingleton

Approved by: Bard Mansager
Thesis Co-Advisor

Hong Zhou
Thesis Co-Advisor

Dr. Carlos Borges
Chair, Department of Applied Mathematics

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Salinas, California has been battling an above average crime rate for over 30 years. This is due primarily to two rival gangs in Salinas: the Norteños and the Sureños. The city and the surrounding community have implemented many methods to mitigate the crime level, from community involvement to the inception of a gang task force. As of yet, none of the efforts have had long-lasting effects.

In a 2009 thesis, Jason A. Clarke and Tracy L. Onufer postulated that various socio-economic variables are influential on the crime level in Salinas. They characterized “crime” as a summation of homicides, assaults and robberies reported. Their thesis determined that “to lower overall violence levels, officials in Salinas should focus on: reducing the unemployment rate, the number of vacant housing units, and the high school dropout rate; and increasing the high school graduation rate and average daily attendance.”

A deeper examination of the data could lead not only to assumptions about how to lower crime rates, but also to a means of predicting future crime rates by using various methods of multiple value regression.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

My lovely wife, Courtney

The NPS Mathematics Department

Dr. Lyn Whitaker

Rebecca Lorentz, NPS Defense Analysis Deparmtment

Karen Ivey and all of the Officers and Staff of the SPD

Georgina Mendoza, Senior Dep. City Attorney at City of Salinas

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

The Salinas Police Department (SPD), in conjunction with the community leadership in Salinas, has been working tirelessly to mitigate gang related crime. Numerous efforts are currently in practice to reduce the city crime rate, from community involvement to the making of a gang task force in association with the surrounding county police offices. All of these efforts are derived from experience and, as seen in other cities, with no statistical model to predict future levels of violent crime in Salinas. This study's purpose is to give Salinas a tool to predict crime using socio-economic statistics easily attainable from public sources.

A. PREVIOUS RESEARCH

In December 2009, Jason Clarke and Tracy Onufer completed a NPS thesis entitled "Understanding Environmental Factors that Affect Violence in Salinas, California" (Onufer & Clark, 2009). Their research compared nine environmental factors: economy; population; housing; education; police force; prison influence; gang rivalry; social service programs; and community involvement against the yearly violence rate in Salinas to determine which environmental factors, if any, are correlated with the violence levels in Salinas. Clarke and Onufer considered violence a combination of reported homicides, robberies, and assaults.

The resulting recommendation of Clarke and Onufer's research was summarized to follow Mayor Dennis Donahue's "four-fold [strategy]: prevention, intervention, a newly envisioned and expanded police department and enhanced community engagement and mobilization" (Stahl, 2009, para. 51). Clarke and Onufer showed that violence was highly correlated with education and dropout rate. This led Clarke and Onufer to conjecture that with an increased emphasis on education and prevention, violence rates would decrease. Intervention was postulated to be established through vocational, education, counseling, and rehabilitation programs. Clarke and Onufer suggested opening a police

substation in the center of the gang territory as a police expansion strategy. Their final recommendation was to start a Mayor's Gang Prevention Task Force (MGPTF) like San Jose to enhance community engagement and mobilization.

Clarke and Onufer's correlation analysis lead to the conclusion that there are four highly correlated environmental factors to the Salinas violence rate: unemployment rate, number of vacant housing units, high school dropout rate, and daily school attendance rate.

B. RESEARCH OBJECTIVE

Using previously established environmental variables, multi-variable regression models were created to predict future violence levels using statistical analysis techniques. A program was also written to allow for automatic regression for further exploration and analysis of the Salinas environmental data.

C. BACKGROUND

1. History of Violence in Salinas

Small tribes of Native Americans inhabited the City of Salinas until around 1822. In 1822, Mexico gained independence from Spain and outside settlers began to arrive in Salinas. From the 1820s to the 1890s, the Salinas Valley was used primarily for ranching and wheat and barley growth. After the 1890s, advances in irrigation and agricultural practices introduced the sugar beet industry to Salinas. In the 1920s, sugar beets and beans gave way to the farming of lettuce because of the ice bunkered railroad, allowing fresh produce shipment nationwide. The area continues to grow lettuce and other green vegetables to this day (Seavey, 2010).

The success of the farming industry helped give rise to the nickname "The Salad Bowl of the World" to Salinas, fueling a "$2 billion agriculture industry which supplies 80% of the country's lettuce and artichokes, along with many

other crops" (History of Salinas, 2012,para 3). Every year, thousands of migrant workers travel to Salinas from Mexico to work on the farms during the harvest season.

The population and the racial demographic in Salinas has greatly changed from 1980 to 2011. The population in Salinas in 1980 was 80,479 (Chapman, 1982, p. 18) and increased to 150,441 by the 2010 census (*State & County QuickFacts: Salinas, California*, 2012). In 1980, Salinas was 38.1% Hispanic and increased to 75.0% in the 2010 census (McFarlane, 2012).

The Salinas gang problem can be traced back to the 1950s. In his 2009 90-day report, Police Chief Fetherolf quoted the 1950 Police Chief McIntyre's statement "Gang fights will not be tolerated in the City of Salinas" (Fetherolf, 2009,p. 6). The 90-day report goes on to mention various instances of Salinas's violence prior to his 2009 report. As displayed in Figures 1–3, the crime in Salinas has steadily increased and maintained a higher level than the national average over the past two decades. In 2009, Salinas had a record breaking 29 homicides, about four times the national average, followed by 19 homicides in 2010, again about four times the national average. In 2009, Salinas was ranked 4th in California for homicides per capita (Fetherolf, 2010, p. 2). Figures 1, 2 and 3 show a comparison of Salinas homicides, robberies and assaults, respectively, compared to national averages. Homicides, assaults, and robberies were added together and used as one statistic. They were labeled as *violence* in this study.

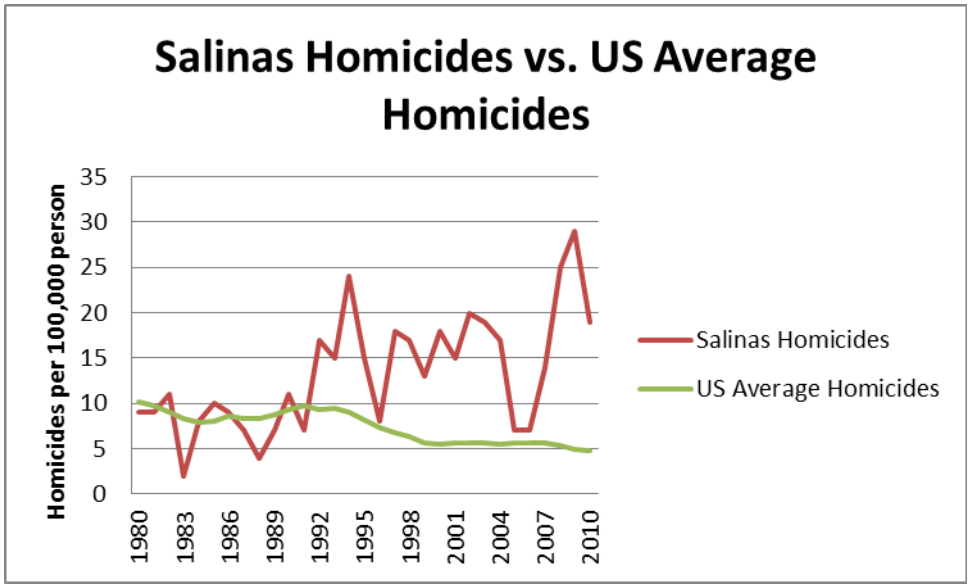

Figure 1. Salinas Homicides versus U.S. Average Homicides 1980–2010

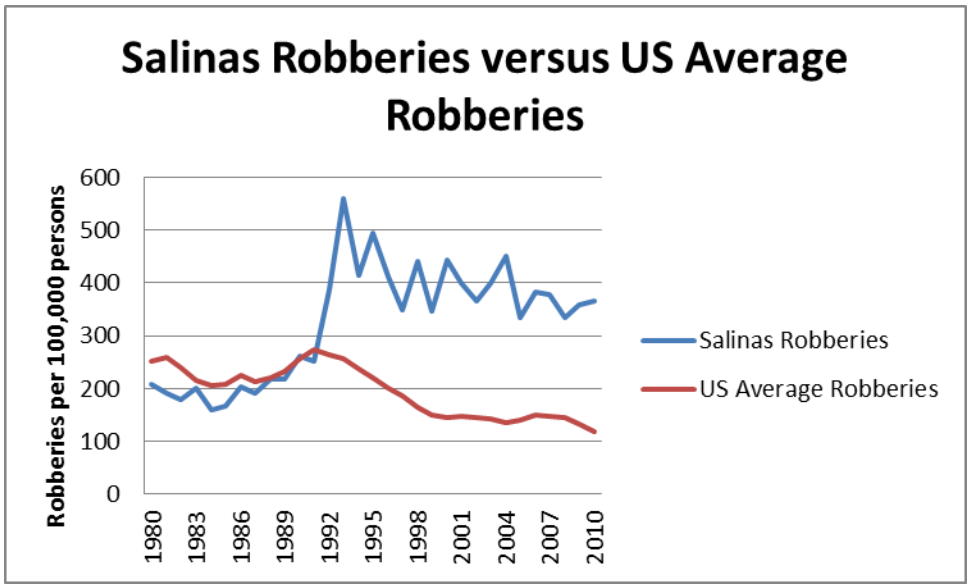

Figure 2. Salinas Robberies versus U.S. Average Robberies 1980–2010

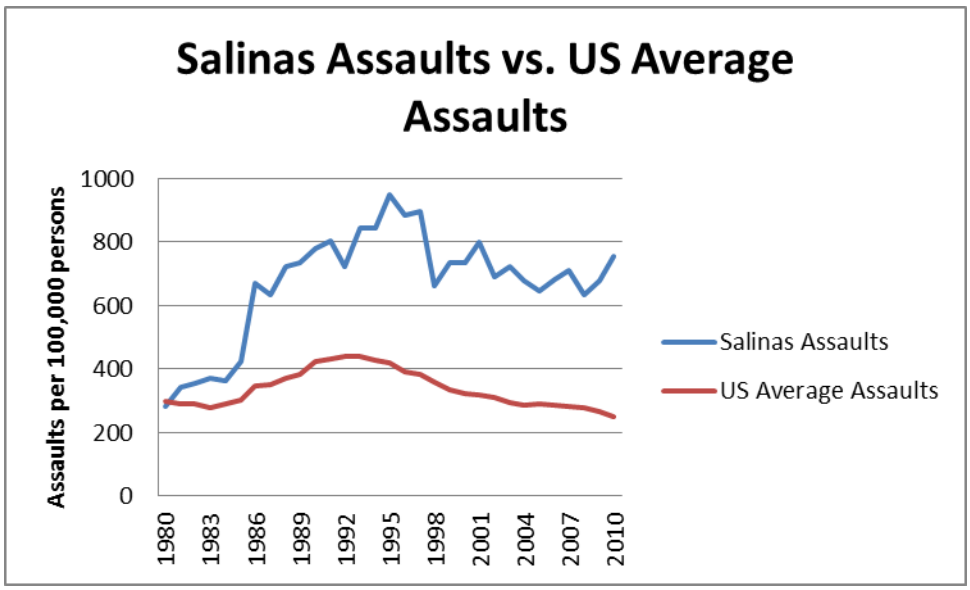

Figure 3. Salinas Assaults versus U.S. Average Assaults 1980–2010

Much of the violence in Salinas is gang-related, stemming from two feuding gangs, the Norteños and the Sureños. These two gangs often live in Salinas physically separated by only one or two city blocks. Both of the gangs vie for control of the drug and prostitution trade within the city limits.

The Sureños, Spanish for Southerner, are a Hispanic gang originating from "a prison dispute between the Mexican Mafia (La Eme) and Nuestra Familia (NF)" (Sureños, 2005). The original members of the gang were associated with the urban Hispanic population, distinguishing themselves from the rural farm working Hispanics. However, since its inception, the Sureños have turned into one of the largest gangs in the United States. Sureños have migrated from California over the past decade and are now living and active in almost every state in the country (Morales et al., 2008, p. 8).

The Norteños are associated with the Nuestra Familia gang. The Norteños, Spanish for "Our Family," are rumored to have started their rivalry with the Sureños because a member of the Sureños stole a pair of shoes from a Norteños member in prison. This incident started the conflict between the

Sureños and the Norteños that continues today (Hennessey, 2003). Norteños are most widely linked with the rural Hispanics in Northern California.

Bakersfield, California is widely accepted as the division point between Northern and Southern California gangs. However, the gangs often ignore traditional boundary lines because of familial ties or gang related opportunities. Salinas is ideally situated to accommodate both gangs because of its proximity to large agricultural and mid-sized urbanized areas. Drug trafficking through Salinas is also very common due to its location between San Francisco and Los Angeles.

There are 11 factors identified in Chief Fetheroff's 2009 90-day report, which contribute to the Salinas gang-crime problem:

- The close proximity of two state prisons to the city (incarcerated gangsters directing gangster activities outside of the prisons)
- The effects of poverty, exacerbated by a sagging economy
- Dysfunctional or struggling families, providing too little juvenile supervision
- Lack of positive adult male role models
- Multi-generational gang families
- Lack of effective teacher/student attachments with at-risk youth
- Inadequate education and an elevated high school drop-out rate
- Drug sales money enticing youth into gang affiliation
- Increased violence in media and video games, desensitizing youth to the impacts of violence
- Limited opportunities for after-school recreation
- Migrating gangsters infiltrating and victimizing law abiding, hardworking seasonal farm workers, many of whom are fearful victims of unreported crime (Fetherolf, 2009, pp. 7–8).

The close approximation to two different prisons plays a key role in the perpetuity of gang-related violence in Salinas. Salinas Valley State Prison and the Correctional Training Facility are located about 25 miles southwest of Salinas outside of Soledad. The California Prison System is running at 200% capacity as of 2010. Being overcrowded, the California Department of Corrections and Rehabilitation (CDCR) is taking steps to reduce the overcrowding. Some of these steps include inmates being placed out of state, non-revocable parole, and inmate population reduction (Actions CDCR Has Taken to Reduce Overcrowding, 2012*)*.

A report by Rand Corporation in their Record on Research about Criminal Behavior (2009) estimates that a prisoner will commit an average of 13 crimes after released early from prison. The current court ordered capacity for the California Prison Systems is 147% by November, 2012. This is a reduction from 168,830 inmates to 117,000 inmates in a one-year period (Burke & Cavanaugh, 2011). This equates to around 50,000 inmates released. With each early release potentially committing crimes, this could result in as many as 500,000 crimes in California. The early release program sends the inmates back into the community in which they were arrested and Salinas could see a percentage of this increase in crime rate from the early release program.

The recidivism rate in California as of 2010 is 67.5% within three years of release (Cate, 2010, p. 32). Many of the gangs in America, to include the Norteños and the Sureños , have strong ties to the prison system and are still primarily run from the leadership that is incarcerated.

2. Current and Past Efforts to Reduce Violence in Salinas

In 1995, the Clinton Administration “awarded the Salinas Police Department nearly $1 million as part of the COPS [Community Oriented Policing Service] Youth Firearms Violence Initiative” (Success Stories, 2011). Salinas Police Department used the money to create a permanent anti-gang task force. During this same period, the Salinas Police Department also created a “Violence

Suppression Unit (VSU) to take firearms away from youth and gang members" (Success Stories, 2011). Finally, Salinas also instituted "Peace Builders," which was to encourage non-violent behavior for elementary school-aged children. Finally, Salinas also instituted a 20-city block clean-up program to remove clutter and garbage from the streets. These efforts did result in a 50% decrease of homicides from 1995 to 1996 and a slight reduction in robberies and assaults. However, as of 2002, the homicide rate was back up to a record high of 20.

In 2010, Salinas instituted Operation Ceasefire and Operation Knockout. Operation Ceasefire was a program successfully instituted in 1996 in Boston, Massachusetts. Operation Ceasefire was a "direct law enforcement attack effort on illicit firearms traffickers supplying youths with guns and an attempt to generate a strong deterrent to gang violence" (Record on Research about Criminal Behavior Corrected, 2009). The local law enforcement made it clear that there would be zero tolerance on gang related activity. When the officers received reports of gang-related activity, the Boston Police Department held gang crackdowns, arresting gang offenders. The result of the policy is that the "Gang violence in Boston declined abruptly" and "it was unnecessary to repeat the crackdowns or move out gradually along the gang network as originally planned" (Record on Research about Criminal Behavior Corrected, 2009).

As of May 2010, the Salinas Police Department had two Operation Ceasefire call-ins, inviting community gang members to meet with personnel who assisted the gang member to leave behind their gang life. "Those agreeing to take part in the program are offered employment opportunities, training and personal services – from résumé-building to tattoo removal" (Solan a, 2010). The call-ins are also useful in informing the gang members of the zero tolerance for gang- related activities in the community.

Operation Knockout was "an eight-month operation … aimed at apprehending members of the Norteños and Sureños gangs that turned Salinas into a hub of murder, robbery and drug dealing" (San Francisco Citizen, 2010). The operation was a multi-organization operation led by the California

Department of Justice's Bureau of Narcotics Enforcement in collaboration with the Salinas Police Department and other local agencies. The culmination of the operation resulted in 44 arrest warrants, leading to 37 arrests and the seizure of over 50 pounds of illegal drugs and paraphernalia. Operation Knockout was also an attempt to "cripple the gang's grip on younger gang members in the area and make existing gang-violence intervention efforts, such as Ceasefire, more effective" (Reynolds, 2010).

D. REGRESSION ANALYSIS TO PREDICT CRIME

Police departments nationwide use some form of crime prediction. The first instance of formal crime analysis was instituted by August Vollmer in the early 1900's (Boba, 2005, p. 20). Vollmer's method involved the "use of pin mapping, the regular review of police reports, and the formation of patrol districts based on crime volume" (Grassie et al., 1977). This method of crime analysis lasted into the 1970s.

In 1968, the Omnibus Crime Control and Safe Streets Acts greatly increased awareness to the analysis of crime statistics and crime prediction. The act authorized grants to the States to fund efforts to reduce crime rates (P.L. 90-351). This shift of attention from crime prosecution to crime prediction led to many police departments adopting crime analysis techniques, finally concluding in the creation of the International Association of Crime Analysts in 1991 and the implementation of Compstat, a data-and-mapping driven strategy at police management for increasing the awareness of crime analysis (Boba, 2005, p. 23).

Currently, most police agencies use some type of crime analysis in every-day operations. In a survey of over 17,000 agencies, Mamalian and La Vigne found that 73% of agencies use crime analysis to fulfill the Unified Crime Report and around 52% calculate statistical reports on criminal activity. However, out of all of the agencies, only 13% use some type of computerized crime analysis, the majority of agencies preferring the more conventional pushpin maps to the more advanced computerized techniques (Mamalian & LaVigne, 1999).

For those agencies that use some type of crime analysis, the emphasis is on the short-term, tactical goals of the police department as opposed to the long-term strategic uses of the data. The 2003 report by O'Shea and Nicholls states that the view of the police officers is, first and foremost, the apprehension of criminals and secondly the "sophisticated police tactical and strategic decision outcomes and solutions to chronic crime problems" (O'Shea & Nicholls, 2003, p. 25). The report goes on to state that the data and methods could better be served as a deterrent tool as opposed to an apprehension tool and this effort would take a cooperation between police officials and academics.

Current forecasting models are primarily built around crime mapping using a geographic information system (GIS). This system is a computerized version of the pushpin map model defined as "a set of computer-based tools that allows the user to modify, visualize, query, and analyze geographic and tabular data" (Boba, 2005, p. 37). GIS is a computerized tool to assist in departmental crime mapping. Crime mapping uses geographical information to conduct special analysis of crime problems to assist in resource allocation for police agencies (Boba, 2005, p. 37).

A 1998 study by Diana Ehlers and Gideon Pimstone used various factors to predict crime rate per 100,000 in the United States. These factors included:

- higher unemployment and increased economic deprivation
- political instability
- urbanisation patterns
- successful implementation of a crime prevention campaign which calls on people to report crime
- increased public awareness of crime
- improved police detection resulting in greater recording of crime (Ehlers, 1998).

The conclusion of their research was that statistical methods could be used to predict crime trends. Regression analysis and correlation are useful tools to predict crime patterns and through these methods, policy makers can be given a “statistical glimpse of the future” (Ehlers, 1998).

In “Forecasting Crime, a City-Level Analysis,” John V. Pepper (2007) explored the ability of different regression models to predict crime rates. In his study, Pepper used two primary variables to predict homicide rates: “the percent of the population that are 18 year old males and the fraction of the population (per 100,000) that are incarcerated” (Pepper, 2007, p. 4). His research concentrated on linear regression models and used lag regression techniques, which use previous homicide levels to predict future homicide levels, as well as other variables in the model. Pepper’s research concluded that naïve walk prediction, a method of prediction that uses previously witnessed statistics and used by many police departments, does well for very short-term prediction, but regression analysis out performs naïve walk prediction for long-range forecasting.

Dr. Wayne Osgood took a different regression based-approach to predicting aggregate crime rates. Osgood explored using Poisson and negative-binomial regression for crime rate predictions in his 2000 study (Osgood, 2000). His research argues, “Poisson regression analysis explicitly addresses the heterogeneous residual variance that presented a problem for [ordinary least squares] regression analysis of crime rates” (Osgood, 2000, p. 27). Osgood then went on to explain that negative binomial regression may be the best method because negative binomial does not have the problem of increased variance that occurs in Poisson regression. This method allows for a more varied approach at crime analysis.

Linear, Poisson, and negative binomial regression were all used in this study in an attempt to find the best regression tool for crime rate prediction in Salinas. All three of the methods are discussed, in detail, in the next section.

THIS PAGE INTENTIONALLY LEFT BLANK

II. SUMMARY OF REGRESSION

A. ORDINARY LEAST SQUARES METHOD

There are two different variable types associated with regression. The first is the class of independent variables or regressors. Independent variables are observed through research and study. The other type of variable is the dependent variable or response variable. The purpose of regression is to model and investigate the relationship between the dependent variable and the independent variable. Equivalently the errors $\varepsilon_i, ..., \varepsilon_n$ are independent and normally distributed with mean 0 and variance σ^2. The βs are estimated by minimizing the error or residual sums of squares:

$$S(\beta_0, \beta_1, \cdots, \beta_k) = \sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} \right) \right)^2 \tag{1}$$

To find the minimum of (2) with respect to β, the derivative of the function in (2), with respect to each of the βs, is set to zero and solved. This gives the following equations:

$$\left. \frac{\delta S}{\delta \beta_0} \right|_{\hat{\beta}_0, \hat{\beta}_1, \cdots \hat{\beta}_k} = -2 \sum_{i=1}^{n} \left(y_i - \left(\hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) \right) = 0, j = 0,1,2...k, \tag{2}$$

and

$$\left. \frac{\delta S}{\delta \beta_j} \right|_{\hat{\beta}_0, \hat{\beta}_1, \cdots \hat{\beta}_k} = -2 \sum_{i=1}^{n} \left(y_i - \left(\hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) \right) x_{ij} = 0, j = 1,2,\ldots,k. \tag{3}$$

The $\hat{\beta} s$, the solutions to (3) and (4), are the least squares estimates of the βs.

It is useful to express both the *n* equations in (1) and the *k*+1 equations in (3) and (4) (which are based on linear function of the βs) in matrix form. The model (1) can be expressed as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{4}$$

where $\mathbf{y}$ is the *n*x1 vector of observations, $\mathbf{X}$ is an *n*x(*k*+1) matrix of independent variables (and an extra column of 1s for the intercept β_0), $\boldsymbol{\beta}$ is a (*k*+1)x1 vector of coefficients and $\boldsymbol{\varepsilon}$ is an *n*x1 vector of independent and identically distributed errors associated with (1).

In order to find the $\hat{\boldsymbol{\beta}}$, the (*k*+1)x1 vector of $\hat{\beta}s$ and the estimate of $\boldsymbol{\beta}$ that minimizes the error, (2) in matrix form is:

$$\begin{aligned} S(\boldsymbol{\beta}) &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{y}^T\mathbf{y} - \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{y}^T\mathbf{y} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \end{aligned} \tag{5}$$

with a superscript "*T*" denoting the transpose of a matrix or vector. The expression $\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y}$ is a scalar. Therefore, the least-squares estimator must satisfy the (*k*+1) equations (3) and (4) written in matrix form as:

$$\left.\frac{\delta \mathbf{S}}{\delta \boldsymbol{\beta}}\right|_{\hat{\boldsymbol{\beta}}} = -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{0} \tag{6}$$

where **0** is the (*k*+1)x1 vector of 0's. This equation can be simplified to:

$$\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y} \tag{7}$$

Under appropriate conditions (i.e. $\mathbf{X}^T\mathbf{X}$ is not singular), this formula will finally net the least squares coefficients:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{8}$$

These coefficients can then be used for predicting or estimating the expected dependent variable for values of the independent variables that do not need to be in the sample used to estimate β.

B. RELATIONSHIPS AMONG VARIABLES

1. Correlation

In practice, there are often many candidate independent variables that can be used in the regression equation. One of the most difficult tasks of an analyst is to determine which of these to use. In order to determine the variables to use in a regression, the relationship between the dependent and independent variables must be established. The relationship between the independent variables must also be examined. An important relationship for this study is correlation and is defined as the linear relationship between two variables. This relationship is measured between pairs of observed variables. For example, simple linear regression with one independent variable, observations are the pairs $(x_1, y_1),(x_2, y_2),...,(x_n, y_n)$. The correlation between these two variables is measured using the sample correlation coefficient with the formula:

$$r = \frac{\sum_{i=1}^{n} y_i (x_i - \bar{x})}{\sqrt{\left[\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2\right]}} \tag{9}$$

with $\bar{x}$ and $\bar{y}$ representing respectively the mean of the observed independent and dependent variables. The coefficient r takes values between -1 and 1, inclusive. A result of -1 implies a perfect negative relationship between the two variables wherein an increase of one variable indicates a decrease in the other. A positive relationship indicates an increase or decrease in both variables simultaneously. A result near zero indicates no or a very small linear relationship between the variables.

Ideally, a good regression fit will include a dependent variable highly correlated with the independent variables, with a correlation value between 0.5 and 1 or between -1 and -0.5. A good regression fit will also have independent variables with very low correlation with a sample correlation for any pair of independent variables, between -0.5 and 0.5. Including highly correlated

independent variables does not add to the regression and can lead to a non-generalized, overfit regression model. Having highly correlated independent variables in the regression is called multicollinearity which can lead to difficulty in interpretation and, when extreme, will cause $\mathbf{X}^T\mathbf{X}$ of equation (9) to be ill-conditioned. Further, perfect linear dependence among the independent variables will cause $\mathbf{X}^T\mathbf{X}$ to be singular and give infinite least squares estimates of $\hat{\boldsymbol{\beta}}$ in equation (8).

2. Transforming Dependent and Independent Variables

Oftentimes, a straight line will not be the best fit of the dependent variables as a function of the independent variables. Therefore, the variables, either dependent or independent must be transformed or adapted. Some common transformations of variables are:

- Take the variable to a power
- Use the natural log function on the variable
- Invert the variable
- Multiply several variables together (interactions)

After transformation, the variables will then be put back into the regression. As a standard of practice, the original variable will be left in the regression with any transformations.

C. VARIABLE SELECTION FOR REGRESSION

In this section, the focus is directed to the most pressing issue of the study, that of selecting the independent variables. Various methods are used to test the adequacy of the regression model. Should too many variables be added into the model, the model could be overfit and only applicable to the given dataset. There are various methods to determine the goodness-of-fit for the regression. The methods used in this analysis were hypothesis tests for the regression, hypothesis tests for each of the coefficients, R-squared for the

regression, and hypothesis tests based on the analysis of variance for the regression. These tools are also used as a basis for the goodness-of-fit for the regression model.

1. Hypothesis Test for Regression

The hypothesis test for regression can be performed for each of the coefficients separately and for the entirety of the regression. The hypothesis test for on a single coefficient tests:

$$\begin{aligned} H_0 &: \beta_j = 0, \\ H_1 &: \beta_j \neq 0 \end{aligned} \tag{10}$$

The equation used to test this hypothesis is:

$$F^* = \left(\frac{\hat{\beta}_i}{se(\hat{\beta}_i)} \right)^2 \tag{11}$$

where $\hat{\beta}_i$ is the *i*th coefficient to be tested and $se(\hat{\beta}_i)$ is the standard error of that coefficient, calculated from:

$$\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{ii}([\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}]^T[\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}])/(n-k)} \tag{12}$$

with *k* being the number of β parameters in the model including the intercept β_0, and where $(X^TX)^{-1}_{ii}$ denotes the (*i*)th diagonal element of square matrix $(X^TX)^{-1}$.

The hypothesis test for the regression tests:

$$H_0 : \beta_1 = ...\beta_k = 0 \tag{13}$$

by using the equation:

$$F^* = \frac{\left[(\mathbf{y}-\bar{\mathbf{y}})^T(\mathbf{y}-\bar{\mathbf{y}})\right] - \left[(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})\right]/(k)}{\left[(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})\right]/(n-k)} \tag{14}$$

where $\bar{\mathbf{y}}$ represents the *n*x1 constant vector where each element is the average of $y_i,...,y_n$. Under the null hypothesis, F^* has an F-distribution with *k* and *n-k* degrees of freedom. Using the F-distribution with the calculated F-statistic one can find the probability of seeing a value in the F-distribution of the size of the F-

statistic or larger. A small p-value indicates that at least one of the coefficients is not zero. A large p-value indicates that there is not enough evidence from the data to show any relationship between the dependent and independent variables. Graphically, the p-value is shown in Figure 4.

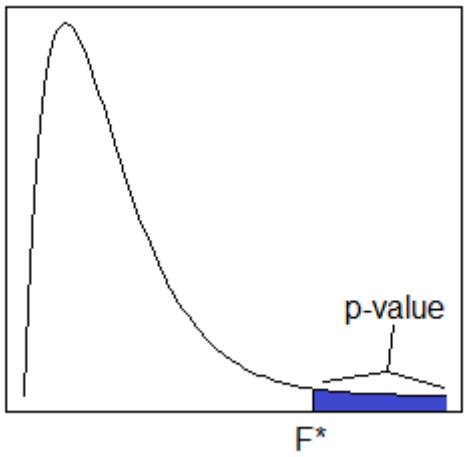

Figure 4. Graphical Representation of the p-value

2. R-Squared (R^2)

R-Squared, often abbreviated R^2, which is also called the coefficient of determination or the percentage of variance explained is equated by the following formula:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{SSR}{SST} \tag{15}$$

where SSR is the sum of squares regression and SST is the sum of squares total and $\hat{y}_i$ is the fitted or predicted value for the *i*th observation.

R^2 is the ratio of the sum of predicted values minus the mean of the observed dependent values squared over total sum of squares. Ideally, this number should be as close to 1 as possible, signifying that the predicted values for the dependent variable are very close to the actual values for the dependent

variable. An R^2 value near 1 indicates that most of the variability in the observed *y* values is accounted for in the model.

3. Residual Standard Error

One important test to ensure the validity of a regression model is to study the Residual Standard Error (RSE) for the model. The equation for the RSE for the model is:

$$RSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k}} \tag{16}$$

This equation is a method of estimating how far the fitted values are from the actual observed values. This value is also used in cross-validation of the model, a tool used to negate model overfitting and to be explained in another section.

4. Analysis of Variance

The Analysis of Variance (ANOVA) organizes the computation of the test statistics for a sequence of hypothesis tests. The most common ANOVA tests the sequence of hypothesis which adds coefficients into the model, one at a time in order to test the increased significance of the model with the independent variable added. The sequence of hypothesis tested:

$$\begin{aligned} &H_0 : \beta_0 \\ &H_1 : \beta_0 + \beta_1 x_1 \\ &\vdots \\ &H_k : \beta_0 + \ldots + \beta_k x_k \end{aligned} \tag{17}$$

An F-statistic is calculated for each step of the ANOVA with $H_n(\hat{y}_i)$ being the predicted value of the *i*th observation based on the model in the *j*th hypothesis.

$$F = \frac{\sum_{i=1}^{n}(y_i - H_{j-1}(\hat{y}_i))^2 - \sum_{i=1}^{n}(y_i - H_j(\hat{y}_i))^2}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 / n - k}, j = 1,2...,k. \quad (18)$$

As with the p-value test, if the F-statistic is very large, implying a very small p-value, the *n*th independent variable should be left in the regression. As before, p-value is the probability of observing an F-statistic as large or larger than are computed from the data. A small p-value corresponding to the test statistic for the test of the null hypothesis H_{j-1} against the alternative H_j indicates that the *j*th regressor is needed in the regression equation when the previous *j*-1 regressors are already accounted for in the model.

One thing to note about the ANOVA is the p-value of an independent variable may be large, but may still be left in the regression. This is because, as a general rule of regression, hierarchical terms are left in the regression if higher power terms have a lower p-value. For example: if a squared term has a very low p-value, but the linear term has a high p-value, the linear term will be left in the regression.

D. GENERALIZED LINEAR MODELS

Generalized linear models (GLM) include linear regression explained in the previous section. GLMs are a "unifying approach to regression and experimental design models, uniting the usual normal-theory linear regression models and nonlinear model" (Montgomery, Peck & Vining, 2006, pp. 454–455) where the dependent variable can have a distribution from a family of distribution other than normal, such as Poisson, exponential, or binomial.

It is still necessary to estimate the coefficients in order to predict the dependent variable for a GLM. However, a GLM will have an additional, called the link function, which gives the relationship between the expected dependent variable and the linear function of the independent variables. A GLM takes the form:

$$g\left[E(y_i)\right] = \mathbf{x}_i^T\boldsymbol{\beta}$$
$$E(y_i) = g^{-1}\left(\mathbf{x}_i^T\boldsymbol{\beta}\right) \tag{19}$$

where the function *g* is called the link function and x_i is the (*k*+1)x1 vector of 1s and the values of the k independent variables for the *i*th observation.

Two non-linear approaches were used in this study, Poisson regression and negative-binomial regression. Both of these methods will be explained in the following sections.

1. Poisson Regression

Poisson regression is based on the fact that the dependent variable may have a count based distribution. The formula for a Poisson distribution is

$$f(y) = \frac{e^{-\mu}\mu^y}{y!}, \quad y = 0,1,... \tag{20}$$

where μ>0 and represents the mean of *y*. The variance for the Poisson distribution is also identically μ.

For Poisson regression, the log function is used as the link function, therefore for $i = 1,...,n$ where each y_i has a Poisson distribution with mean μ_i, the Poisson regression model is expressed as:

$$E(y_i) = \mu_i$$
$$g(\mu_i) = \mathbf{x}_i^T\boldsymbol{\beta} \tag{21}$$
$$\mu_i = g^{-1}(\mathbf{x}_i^T\boldsymbol{\beta})$$

Substituting the log link function gives:

$$\ln(\mu_i) = \mathbf{x}_i^T\boldsymbol{\beta}$$
$$\mu_i = e^{\mathbf{x}_i^T\boldsymbol{\beta}} \tag{22}$$

The maximum-likelihood estimation (MLE) approach must be employed in order to estimate the βs for the regression. Finding the MLE for the Poisson regression starts with the expression for the likelihood of observing ***y*** as a function of $\boldsymbol{\beta}$ (where $y_i,...,y_n$ are assumed independent):

$$L(\mathbf{y},\boldsymbol{\beta}) = \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n} \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}$$
$$= \frac{\prod_{i=1}^{n} e^{-\mu_i} \prod_{i=1}^{n} \mu_i^{y_i}}{\prod_{i=1}^{n} y_i!} \tag{23}$$

Taking the natural log of both sides gives:

$$\ln(L(\mathbf{y},\boldsymbol{\beta})) = \ln\left(\frac{\prod_{i=1}^{n} e^{-\mu_i} \prod_{i=1}^{n} \mu_i^{y_i}}{\prod_{i=1}^{n} y_i!} \right)$$
$$= \sum_{i=1}^{n} \ln(e^{-\mu_i}) + \sum_{i=1}^{n} \ln(\mu_i^{y_i}) - \sum_{i=1}^{n} \ln(y_i!) \tag{24}$$
$$= -\sum_{i=1}^{n} \mu_i + \sum_{i=1}^{n} y_i \ln(\mu_i) - \sum_{i=1}^{n} \ln(y_i!)$$
$$= \sum_{i=1}^{n} \left[y_i \left(\mathbf{x}_i^T \boldsymbol{\beta} \right) - e^{\mathbf{x}_i^T \boldsymbol{\beta}} - \ln(y_i!) \right]$$

Denoting $\ln(L(\mathbf{y},\boldsymbol{\beta}))$ by $l(\mathbf{y},\boldsymbol{\beta})$, as in ordinary least squares, the goal is to solve the following equation:

$$\frac{\partial l(\mathbf{y},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left[y_i \mathbf{x}_i^T - \mathbf{x}_i^T e^{\mathbf{x}_i^T \boldsymbol{\beta}} \right] = \sum_{i=1}^{n} \mathbf{x}_i^T \left(y_i - e^{\mathbf{x}_i^T \boldsymbol{\beta}} \right) = \mathbf{0} \tag{25}$$

There is no analytical approach to solving for the coefficients in the above equation. Therefore, at this point, some type of numerical method, such as the Newton-Raphson technique using iteratively reweighted least squares, is used to estimate the coefficients (Montgomery, Peck & Vining, 2006, p. 575).

a) Goodness-of-Fit with the Poisson Model

Goodness-of-fit for a Poisson model is measured using the residual deviance instead of R^2 or the residual standard error used in linear regression. The formula for residual deviance for Poisson regression is:

$$D = 2\sum_{i=1}^{n} \left(y_i \ln\left(\frac{y_i}{\hat{y}_i} \right) - (y_i - \hat{y}_i) \right) \tag{26}$$

The residual deviance should be as small as possible. For Poisson regression, the residual deviance, ideally, will be close to or less than the number of observations minus the number of parameters, or the residual degrees of freedom of the model. If the residual deviance is too much greater than the residual degrees of freedom, the model may not be a good fit and must be modified.

2. Negative Binomial Regression

A Poisson model assumes that the variance of the dependent variable will equal the mean of the dependent variable. Oftentimes, this assumption does not hold and the dependent variable is more variable than can be accounted for by the independent variables in the Poisson regression. As a means to remedy this, the negative-binomial regression technique can be employed. The negative binomial distribution is closely related to the Poisson distribution in that the negative binomial is a measure of instances of an event until reaching a concluding event. The formula for the negative binomial distribution is:

$$f(y) = \frac{\Gamma(r+y)}{y!\Gamma(r)}(1-p)^r p^y \qquad y = 0,1,2... \tag{27}$$

with

$$\begin{aligned} E(y) &= \frac{pr}{1-p} \\ Var(y) &= \frac{pr}{(1-p)^2} \end{aligned} \tag{28}$$

where $r>0$ and, when it is an integer, r can be interpreted as the number of failures and y is the number of successes required to get exactly r failures and p is the probability of a success. The function $\Gamma(x)$ is the gamma function and is a continuous version of the choose function:

$$\Gamma(x)=\int_0^\infty t^{z-1}e^{-t}dt \tag{29}$$

To use the negative binomial for regression and make the distribution comparable to the Poisson regression, some adjustments to the distribution must be made. Let $\mu = E(y)$. Then:

$$\mu = \frac{pr}{1-p} \Rightarrow p = \frac{\mu}{\mu + r} \tag{30}$$

Substituting p and 1-p for functions of μ and r in (28) gives:

$$\begin{aligned} f(y) &= \frac{\Gamma(r+y)}{y!\Gamma(r)}\left(1-\frac{\mu}{\mu+r}\right)^r\left(\frac{\mu}{\mu+r}\right)^y = \frac{\Gamma(r+\mu)}{y!\Gamma(r)}\left(\frac{\mu+r-\mu}{\mu+r}\right)^r\left(\frac{\mu}{\mu+r}\right)^y \\ &= \frac{\Gamma(r+y)}{y!\Gamma(r)}\left(\frac{r}{\mu+r}\right)^r\left(\frac{\mu}{\mu+r}\right)^y = \frac{\mu^y}{y!}\frac{\Gamma(r+y)}{\Gamma(r)(\mu+r)^y}\frac{1}{\left(1+\frac{\mu}{r}\right)^r} \end{aligned} \tag{31}$$

Further, to see the relationship between the negative binomial and Poisson distributions, let $r \to \infty$ and $y \to 0$. This implies no chance of a success and continual counting of failures and leads to:

$$\lim_{r\to\infty} \frac{\mu^y}{y!}\frac{\Gamma(r+y)}{\Gamma(r)(\mu+r)^y}\frac{1}{\left(1+\frac{\mu}{r}\right)^r} = \frac{\mu^y}{y!}\cdot 1 \cdot e^{-\mu} \tag{32}$$

which leads back to the Poisson distribution.

Estimating the parameters for the negative binomial regression, which also uses the log link function, is similar to estimating the parameters for Poisson regression:

$$g(u_i) = \mathbf{x}_i^T\boldsymbol{\beta} \Rightarrow u_i = g^{-1}(\mathbf{x}_i^T\boldsymbol{\beta}) = e^{\mathbf{x}_i^T\boldsymbol{\beta}} \tag{33}$$

Now, the MLE for the negative binomial can be found by first expressing the likelihood of observing y and equivalently the log-likelihoood as:

$$\begin{aligned} L(\mathbf{y},\boldsymbol{\beta}) &= \prod_{i=1}^n \frac{\mu_i^{y_i}}{y_i!}\frac{\Gamma(r+y_i)}{\Gamma(r)(\mu_i+r)^{y_i}}\frac{1}{\left(1+\frac{\mu_i}{r}\right)^r} \\ l(\mathbf{y},\boldsymbol{\beta}) &= \ln(L(\mathbf{y},\boldsymbol{\beta})) = \ln\left(\prod_{i=1}^n \frac{\mu_i^{y_i}}{y_i!}\frac{\Gamma(r+y_i)}{\Gamma(r)(\mu+r)^{y_i}}\frac{1}{\left(1+\frac{\mu_i}{r}\right)^r}\right) \\ &= \sum_{i=1}^n y_i \ln\mu_i + \ln(\Gamma(r+y_i)) + \ln(1) - \ln(y_i!) - \ln(\Gamma(r)) - \ln((\mu_i+r)^{y_i}) - r\ln\left(1+\tfrac{\mu_i}{r}\right) \\ &= \sum_{i=1}^n y_i x_i + \ln(\Gamma(r+y_i)) - \ln(y_i!) - \ln(\Gamma(r)) - y_i\ln(x_i^T\boldsymbol{\beta}+r) - r\ln\left(1+\tfrac{x_i^T\boldsymbol{\beta}}{r}\right) \end{aligned} \tag{34}$$

Now, to find the MLEs $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$, take the derivative with respect to $\boldsymbol{\beta}$:

$$\frac{\partial l(\mathbf{y},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} y_i x_i + \frac{y_i x_i}{x_i^T \boldsymbol{\beta} + r} - \frac{x_i}{1 + \frac{x_i^T \boldsymbol{\beta}}{r}} = \mathbf{0} \tag{35}$$

As with the Poisson regression, there is no closed-form solution to this equation, leading to the use of a numerical method to solve for the parameters for the regression equation.

In this study, the negative binomial regression is a means to expand upon the Poisson regression. When a negative binomial regression was used, a Poisson model was first fit to the data. The Poisson coefficients were then used as the starting values for the numeric computation of the negative binomial regression coefficients.

E. CROSS VALIDATION

Cross validation is a method used to ensure a regression model of any type is not overfit. An overfit model will only work for the observed regressors and will not be as useful for predicting future outcomes of the data set. This is a danger when many independent variables are available to predict the dependent variables and *n* is small and moderate. To cross validate a regression, the observed response values are randomly broken into *m* subgroups. The regression is then refit to m-*1* of the original subgroups and the values for the *m*th group is estimated. In the extreme version of cross validation, called jack-knife, *m*=*n.* The regression model is fit to *n*-1 observations and used to predict the one observation that is left out. This is repeated *n* times. Let $\hat{y}_{(i)} i = 1,...,n$ represent the predicted value of the *i*th observation obtained in this manner. The residuals of the estimated group are calculated, giving the cross-validation score of

$$\sqrt{\frac{\sum_{i=1}^{n} (y_{(i)} - \hat{y}_{(i)})^2}{n}} \tag{36}$$

This score is compared to the residual standard error (RSE) for the complete model described in equation (17).

If the cross-validation score is much higher than for the equated model, the model is said to be overfit. An overfit model may have too many variables or too many interactions, giving the regression the illusion of a very good fit when, in fact, the model is very good, but only for the given observation and not good for other data.

For this study, the linear regression and Poisson models were both checked for overfitting. Because the negative binomial regressions are fit after each Poisson regression and use the same regressors as do the corresponding Poisson regression, they are not cross validated to check for overfitting.

III. DATA ANALYSIS

The data collected and researched in this study is specific to Salinas, California and thus any conclusions made about the data will be specific to that region of California.

A. DATA COLLECTION

In an effort to continue the work of Clark and Onufer, the data assessed in this research closely resembles the variables used in their 2009 thesis. The data has been updated to use any more current statistics pertaining to Salinas. All available 2010 data was added to the past data.

All of the other data was collected from online federal resources. The purpose of this research was to give Salinas Police Department (SPD) an easily accessible tool to estimate crime levels with readily accessible data. Therefore, all data used in this research is publicly accessible.

It is also important to note that inflation was taken into account with the study, but did not significantly affect the trend of the financial data and was therefore not input into the calculations.

In order to estimate Salinas violence trends, two different types of variables are needed. Independent variables are the environmental factors effecting violence, such as Salinas Police Department budget, unemployment level, and prison statistics. The dependent variable is what is to be predicted, in this case, violence levels.

A 2006 study by the Department of Justice showed that aggravated assault, auto theft, burglary, drug sales, theft, and robbery are the most likely criminal offences perpetrated by youth gangs (Egley & O'Donnell, 2008). Therefore, this study uses a summation of reported homicides, aggravated assaults, and robberies as reported yearly from Salinas to the Federal Bureau of

Investigation to predict future crime trends. This data is easily obtained from either the Salinas Police Department web page or the Federal Bureau of Investigation web page.

B. CORRELATION OF VARIABLE ANALYSIS

In order to accurately formulate a regression for prediction, the correlation between all of the variables was explored. The variables with the highest correlation to the dependent variable were used for regression formulation. Because the study is a time-based regression, there is a possibility that some of the variables will affect other variables one or even two years later. Therefore, not only was direct correlation explored, but also correlation with violence shifted one and two years in the future. Table 1 shows the resulting correlations.

	No Shift	One Year Shift	Two Year Shift
Population	0.687	0.635	0.572
Drop Outs	0.310	0.09993	0.311
Drop Out Rate	0.478	0.283	0.482
SPD Budget	0.459	0.423	0.403
SPD Employees	0.272	0.194	0.133
Sworn Police	-0.552	-0.705	-0.373
CDCR Capacity	0.791	0.729	0.655
CDCR Population	0.776	0.712	0.644
CDCR Overpopulation Percentage	0.834	0.804	0.799
Parole Population	0.815	0.767	0.713
Parks and Recreation Budget	0.242	0.298	0.352
Library Budget	0.572	0.550	0.549
Unemployment Percentage	0.516	0.701	0.776
Number of Vacant Units	-0.526	-0.631	-0.724
Personnel Per Household	-0.242	-0.399	-0.651

Table 1. Correlation of Independent Variables and Violence

Examining the correlation led to the use of population, SPD Budget, sworn police with a one year shift, CDCR Overpopulation Percentage, parole population, unemployment percentage with a two year shift, number of vacant units with a two year shift and personnel per household with a two year shift. With the choice of variables, it was necessary to examine the correlation

between the variables. If two variables are highly correlated, only one of the variables will assist in the regression. The correlation between variables is described in Table 2.

	P	S	O	Pa	SP	U	N	H
Population (P)	1.000	0.891	0.866	0.948	0.66	-0.80	0.92	0.963
SPD Budget (S)	0.891	1.000	0.684	0.769	0.93	-0.83	0.96	0.714
CDCR Overpopulation Percentage (O)	0.866	0.684	1.000	0.909	-0.07	-0.41	0.64	0.832
Parole Population (Pa)	0.948	0.769	0.909	1.000	-0.10	-0.61	0.69	0.850
Sworn Police (SP)	0.659	0.934	-0.065	-0.10	1.000	-0.665	0.86	0.350
Unemployment Percentage (U)	-0.80	-0.827	-0.408	-0.61	-0.67	1.000	-0.89	-0.70
Number of Vacant Units (N)	0.922	0.963	0.643	0.686	0.864	-0.89	1.000	0.844
Personnel Per Household (H)	0.963	0.714	0.832	0.850	0.350	-0.70	0.844	1.000

Table 2. Correlation between Independent Variables for OLS Regression to Predict Violence

Examining the correlation between pairs of dependent variables led to the removal of population and parole population from the regression formulation. Those two variables were highly correlated with other variables and were accounted for by the other variables.

A graphical depiction of the correlation for the chosen variables was also examined in Figure 5:

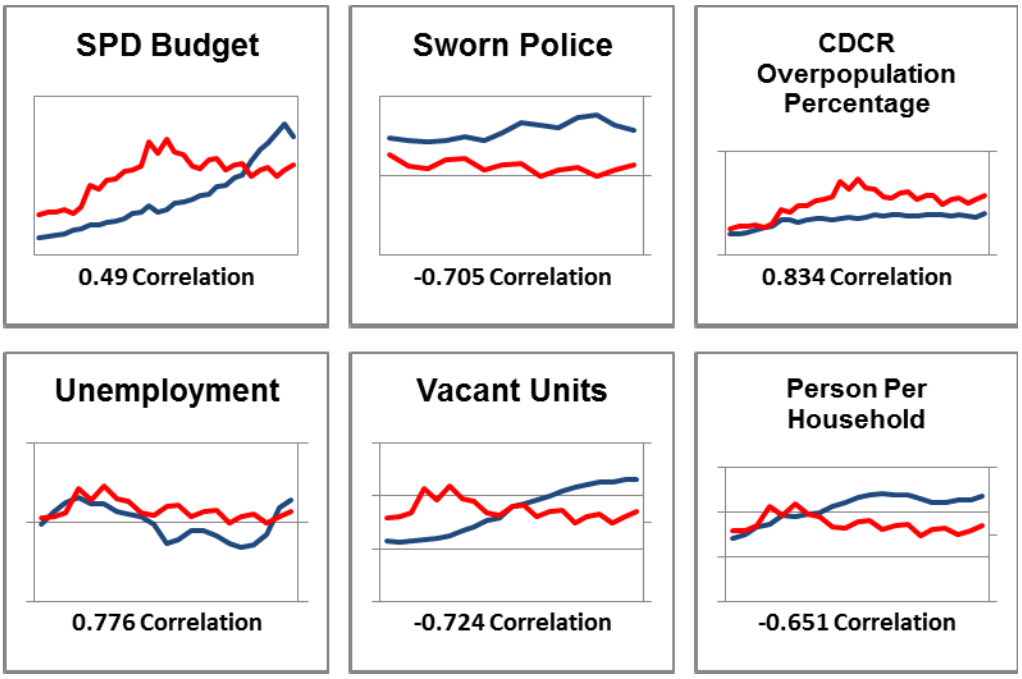

Figure 5. Graphical Representation of Correlation between Independent Variables and Violence. Violence in Red

In each of the panes in Figure 5, violence (red) is plotted against time. Each of the blue lines represents an independent variable (as labeled in each panel) plotted against time.

Most of the variables for the study, be it budgets or manpower, increased over time. This trend similarity indicates that many of the variables will not assist in a regression because of the similarity of correlation between independent variables.

It is also of interest to the City of Salinas to predict future homicide rates using the economic variables. Therefore, the correlation between the variables and Salinas homicide rates with shifts in years was also calculated and displayed in Table 3.

	No Shift	One Year Shift	Two Years Shift
Population	0.610	0.593	0.582
Drop Outs	0.065	0.209	0.154
Drop Out Rate	-0.052	0.171	0.213
SPD Budget	0.586	0.618	0.618
SPD Employees	0.502	0.561	0.556
Sworn Police	0.103	0.250	0.215
CDCR Capacity	0.614	0.625	0.641
CDCR Population	0.596	0.613	0.625
CDCR Overpopulation Percentage	0.470	0.504	0.523
Parole Population	0.588	0.651	0.671
Parks and Recreation Budget	0.557	0.541	0.333
Library Budget	0.773	0.568	0.329
Unemployment Percentage	0.248	0.002	-0.259
Number of Vacant Units	0.262	0.219	0.169
Persons Per Household	0.340	0.161	-0.076

Table 3. Correlation between Dependent Variables and Homicide Events in Salinas

An inspection of the correlation lead to the use of population, SPD Budget, CDCR Overpopulation Percentage, parks and recreation budget, and library budget. None of the variables were shifted because there was not enough of a correlation disparity to warrant a shift. The correlation between these variables were also explored and displayed in Table 4.

	P	S	O	R	L
Population (P)	1.000	0.898	0.879	0.638	0.841
SPD Budget (S)	0.898	1.000	0.700	0.745	0.746
CDCR Overpopulation Percentage (O)	0.879	0.700	1.000	0.537	0.734
Parks and Recreation Budget (R)	0.638	0.745	0.537	1.000	0.751
Library Budget (L)	0.841	0.746	0.734	0.751	1.000

Table 4. Correlation between Independent Variables for Homicide Regression

All of the independent variables of the homicide regression are highly correlated, but all variables were kept for the initial exploration of regression for

homicide prediction. A graphical representation of the correlation between the variables and homicide is displayed in Figure 6.

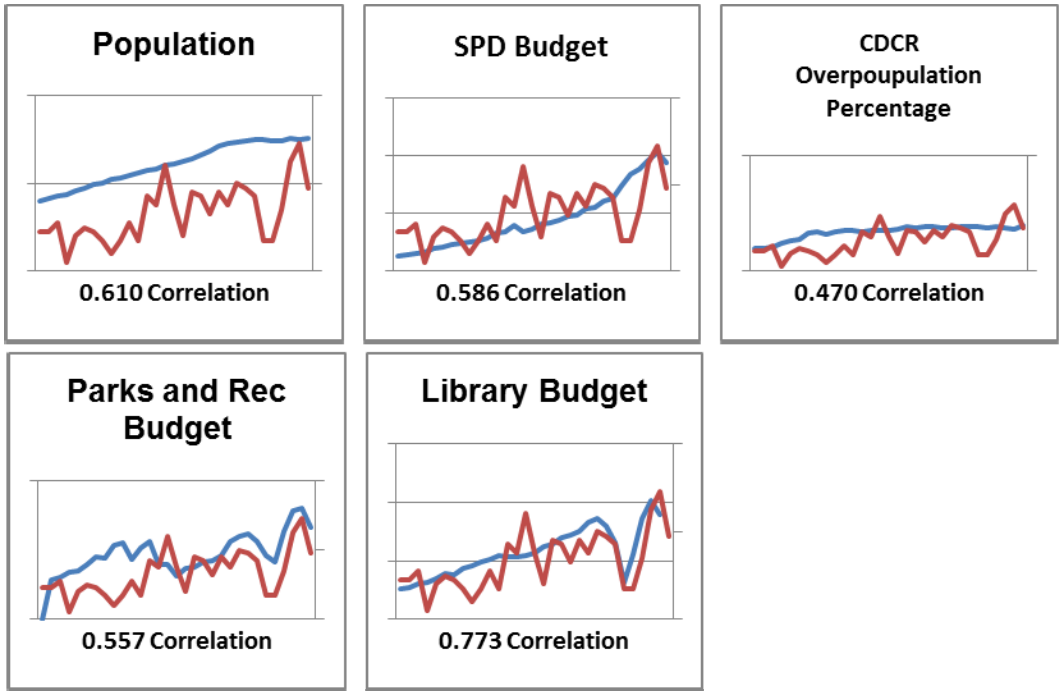

Figure 6. Graphical Correlation between Homicide and Independent Variables. Homicides in Red

Assaults, defined in this study and reported to the FBI, are defined as aggravated assaults and consist of assaults with a weapon involved. These crimes could have easily escalated into homicides. Because of this, regression analysis was used to predict the amount of assaults and homicides in Salinas.

The correlations of the independent variables against homicide and assault numbers with one and two-year shifts are in Table 5.

	No Shift	One Year Shift	Two Year Shift
Population	0.589	0.520	0.436
Drop Outs	0.362	0.276	0.152
Drop Out Rate	0.498	0.441	0.343
SPD Budget	0.370	0.325	0.288
SPD Employees	0.156	0.063	-0.028
Sworn Police	-0.479	-0.521	-0.194
CDCR Capacity	0.693	0.610	0.507
CDCR Population	0.684	0.596	0.501
CDCR Overpopulation Percentage	0.827	0.766	0.729
Parole Population	0.728	0.649	0.564
Parks and Recreation Budget	0.245	0.257	0.309
Library Budget	0.493	0.435	0.399
Unemployment Percentage	0.575	0.690	0.773
Number of Vacant Units	-0.672	-0.676	-0.685
Personnel Per Household	-0.486	-0.522	-0.602

Table 5. Correlation of Variables against Assaults and Homicides

Examination of the correlation of assaults and homicides against the possible regressors led to the use of CDCR Overpopulation percentage, parole population, unemployment percentage with a two year shift, number of vacant units, and person per household with a two year shift. The correlation between these possible variables is displayed in Table 6.

	O	Pa	U	N	H
CDCR Overpopulation Percentage (O)	1.000	0.909	-0.408	0.770	0.832
Parole Population (Pa)	0.909	1.000	-0.605	0.805	0.850
Unemployment Percentage (U)	-0.408	-0.605	1.000	-0.859	-0.701
Number of Vacant Units (N)	0.770	0.805	-0.859	1.000	0.909
Personnel Per Household (H)	0.832	0.850	-0.701	0.909	1.000

Table 6. Correlation between Independent Variables for Homicide and Assault Regression

With an examination of the inter-correlation between possible regressor variables, it was decided to not use personnel per household or number of vacant units in the formulation of the regression to predict assaults and homicides. The graphical depiction of correlation between the chosen regressors and assaults and homicides is graphed in Figure 7.

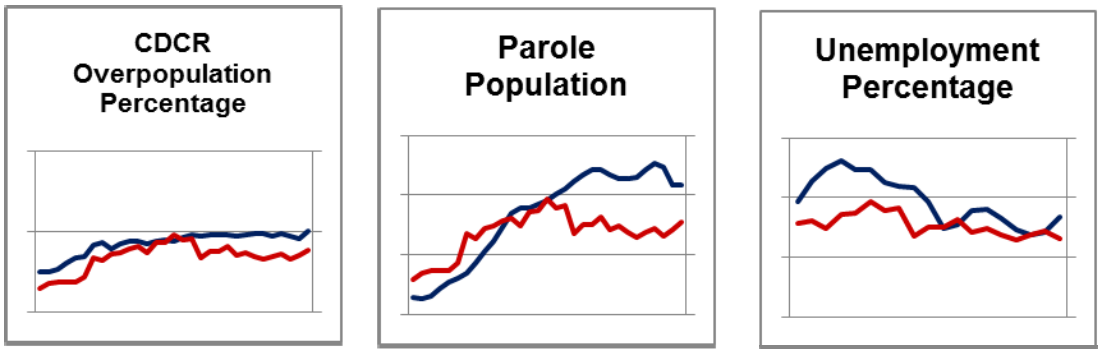

Figure 7. Graphical Correlation between Homicide and Assaults and Independent Variables. Homicides and Assaults in Red

After all variables were chosen, all three of the dependent variables, violence, homicides, and assaults and homicides, were input into linear, Poisson, and negative binomial models in an attempt to predict possible future criminal activity levels in Salinas.

C. REGRESSION ANALYSIS

In order to estimate violence levels for Salinas, California, the regression techniques outlined in Chapter II were applied to the data used in the Onufer and Clark (2009) thesis.

1. Violence Prediction using Ordinary Least Squares

The initial regression for all three variables was ordinary least squares (OLS), or general linear regression. The method used to find the optimal regression was backward elimination, wherein all of the variables of interest identified in the previous section were included in the regression and removed if the variable did not add to the quality of the regression.

An initial regression included person per household included and it was found that person per household added nothing to the model and was therefore taken out of the initial model. The second regression equation for violence prediction with OLS regression was:

$$\hat{y}=-1.827\text{x}10^{3}+9.384\text{x}10^{-6}(\text{SPD Budget})$$
$$-8.354\text{x}10^{2}(\text{CDCR Overpopulation Percentage})$$
$$+50.62(\text{Unemployment with Two Year Shift}) \quad (37)$$
$$+1.712(\text{Number of Vacant Units with Two Year Shift})$$
$$-13.93(\text{SPD Sworn Police with One Year Shift})$$

This initial regression seems to be a decent fit for the violence levels. The p-values for the coefficients are given in Table 7.

	Estimates	Standard Error	P-value
Intercept	$1.83\text{x}10^{3}$	$9.46\ \text{x}10^{2}$	0.09465
SPD Budget	$9.38\ \text{x}10^{-6}$	$5.87\ \text{x}10^{-6}$	0.154
CDCR Over Population Percentage	$-8.35\ \text{x}10^{2}$	$4.01\ \text{x}10^{2}$	0.076
Unemployment	-1.39 x10	2.75	0.0381
Number Vacant Units	5.06 x10	1.99 x10	0.125
SPD Badged Police	1.71	$9.82\ \text{x}10^{-1}$	0.00147

Table 7. P-Values for the initial OLS regression

This regression equated an R^2 value of 0.8449. This also coincides with a good OLS fit for violence prediction.

With five variables being included into the model, there is worry that the regression may overfit the regression. Therefore, cross-validation was used on this model and will be used on all subsequent models to check for overfitting. The RSE for this model was 34.8 and the cross-validation score was 42.59. The model may be slightly overfit, but is still within acceptable means and will be used to predict violence future violence levels in Salinas.

Although a good fit for the data was quickly derived from the initial values, there was interest in attempting to fit another OLS regression with other variables of interest and a more complete data set. The regression in equation (42) was limited by the observations of SPD Badged Police, being only recorded from 1997–2010.

This new regression fit consisted of SPD budget and CDCR Overpopulation percentage. The formula for this regression is:

$$-4.259\text{x}10^{2}-6.239\text{x}10^{-6}(\text{SPD Budget}) \\ +9.036\text{x}10^{2}(\text{CDCR Overpopulation Percentage}) \tag{38}$$

with the following p-values displayed in Table 8.

	Estimates	Standard Error	P-value
Intercept	$-3.34\text{x}10^{2}$	$1.73\text{x}10^{2}$	0.0635
SPD Budget	$-3.23\text{x}10^{-6}$	$3.14\text{x}10^{-6}$	0.0858
CDCR Over Population Percentage	$8.22\text{x}10^{2}$	$1.18\text{x}10^{2}$	$6.21\text{x}10^{-8}$

Table 8. Second Fit OLS Regression for Violence P-Values

With an R^2 of 0.7285. Although the R^2 for this regression is not as high, this model emphasizes different variables, which may be useful to the City of Salinas. The RSE for this model was 143.5 with a cross-validation score of 149.4174. These two scores are very close together so this model is not overfit.

A graphical depiction of the fit of the two different models is shown in Figures 8 and 9.

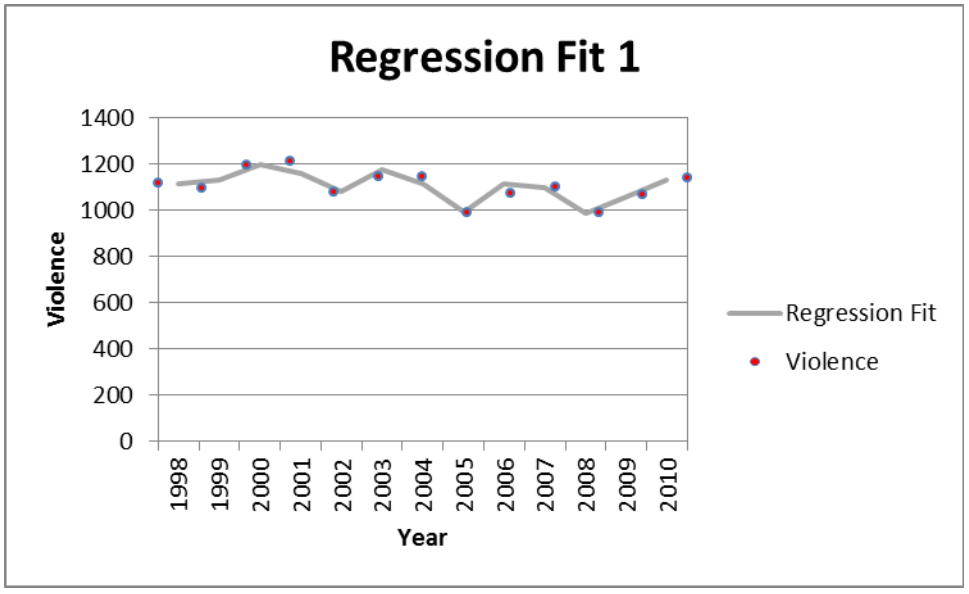

Figure 8. Regression Fit for Formula (39)

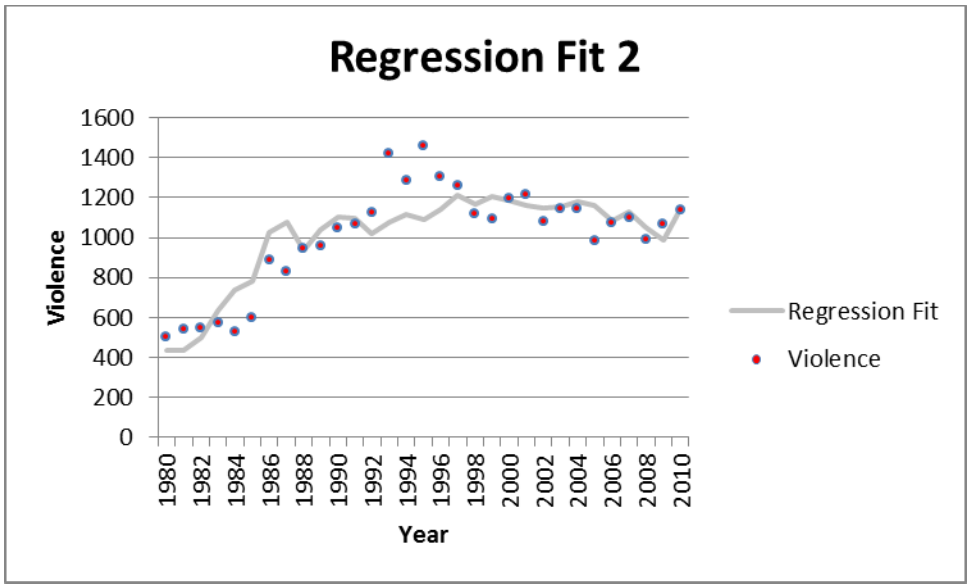

Figure 9. Regression Fit for Formula (40)

With an inspection of the regression fit through the actual data and examination of the RSE for each of the models, it was determined that formula (39) was the best fit for the OLS prediction of violence levels in Salinas. A Poisson model was then formulated to see if Poisson regression would be a better means of predicting violence.

2. Violence Prediction using Poisson Regression

As stated by Osgood (2000), Poisson regression is often useful for exploring crime rates. Therefore, in this study, all response variables were analyzed with Poisson regression, in addition to OLS regression.

The starting method to Poisson regression was similar to OLS regression. The initial Poisson model used the same regressors as the OLS model:

$$\begin{aligned}\hat{y}=\exp(\ & 7.679-8.695\times10^{-9}(\text{SPD Budget})\\ & -1.287\times10^{-2}(\text{SPD Sworn Police})\\ & -7.743\times10^{-1}(\text{CDCR Overpopulation Percentage})\\ & +4.666\times10^{-2}(\text{Unemployment})\\ & +1.582\times10^{-3}(\text{Number of Vacant Units}))\end{aligned} \tag{39}$$

with the p-values for the regressors displayed in Table 9.

	Estimate	Standard Error	P-value
Intercept	7.68	8.30×10^{-1}	2.00×10^{-16}
SPD Budget	8.70×10^{-9}	5.14×10^{-9}	0.091
CDCR Over Population Percentage	-7.74×10^{-1}	3.51×10^{-1}	0.0276
Number Vacant Units	-1.29×10^{-2}	2.44×10^{-3}	0.0637
Unemployment	4.67×10^{-2}	1.73×10^{-2}	0.00701
SPD Badged Police	1.58×10^{-3}	8.5×10^{-4}	1.28×10^{-7}

Table 9. Initial Poisson Regression for Violence

Ideally, for a Poisson regression, the residual deviance should be close to the degrees of freedom. For this regression, the deviance was 7.48 for 7 degrees of freedom. This shows not only is this a good fit using the Poisson regression, but that the data is not overdispersed.

This regression was also tested for over fitting. The cross-validation score for the data was 1646.29 and the RSE for the model was 1217.46. These numbers suggest overfitting, but the numbers are within acceptable means. A graphical representation of this fit is shown in Figure 10.

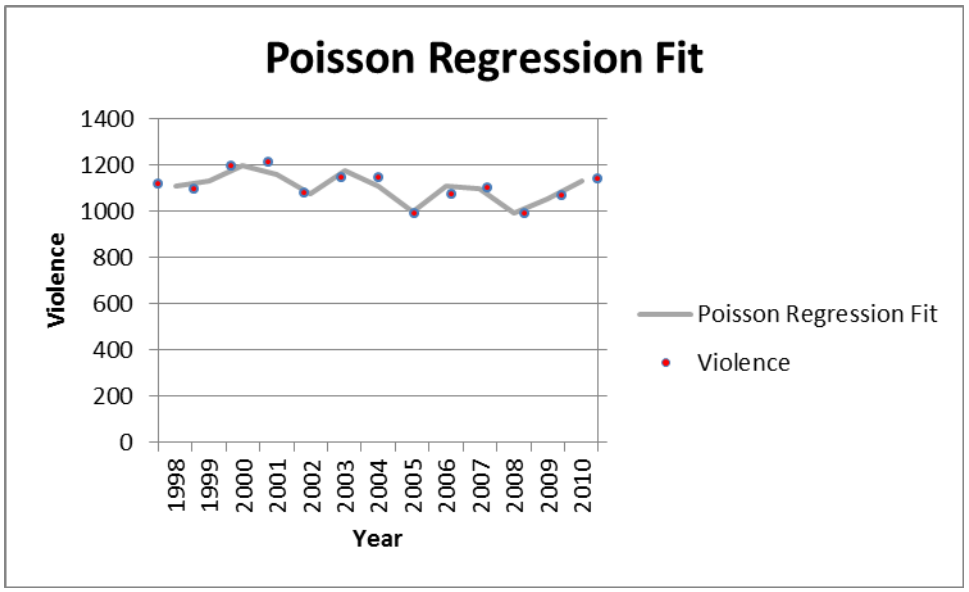

Figure 10. Poisson Regression fit for Violence Prediction

The Poisson regression seems to predict violence well and there is no evidence of overdispersion. However, for completeness the negative binomial was also explored.

3. Violence Prediction using Negative Binomial Regression

Osgood states that Poisson regression is valid for crime rate exploration, but that negative binomial regression can also be used and may be more efficient with the ability of negative binomial to reduce the error caused by overdispersion with which a Poisson model cannot compensate. The initial model for the negative binomial regression was identical to the final model for the Poisson regression. The coefficients from the Poisson regression were also used as the initial guess for the numerical method to determine the coefficients for the negative binomial model. The resulting model was exactly same model as the Poisson model. This indicates that the negative binomial regression is not an improvement over the Poisson model and was not used to predict violence in this study.

4. Homicide Prediction using Ordinary Least Squares

With homicide levels almost five times the national average, the leadership in Salinas is constantly finding ways to decrease the violence levels in their city. To do this, it would be useful to see the factors that influence homicide levels in Salinas. Clarke and Onufer did this in their 2009 thesis, showing the economic factors correlated with violence. Taking this a step further, the city can predict future homicide levels and estimate the change on homicide levels by focusing on different economic variables.

The initial regression method to predict homicides is the same as violence levels. However, different regressors were more highly correlated with homicides than violence. The initial OLS model was:

$$\hat{y}=1.152-7.363\text{x}10^{-5}(\text{Population}) +1.805\text{x}10^{-1}(\text{SPD Budget}) -3.959(\text{CDCR Overpopulation Percentage}) -1.745\text{x}10^{-6}(\text{Parks and Rec Funding}) +8.184\text{x}10^{-6}(\text{Library Funding}) \quad (40)$$

This equation was not a very good fit to predict homicide rates in Salinas, as shown by the p-values for the coefficients in the regression. Only the p-value for library funding is small enough for a good regression fit. The p-values for the regression are given in Table 10.

	Estimate	Standard Error	P-value
Intercept	1.15x10	9.78	0.2505
Population	$-7.36\text{x}10^{-5}$	$1.57\text{x}10^{-4}$	0.64332
SPD Budget	$1.81\text{x}10^{-7}$	$2.46\text{x}10^{-7}$	0.47005
CDCR Over Population Percentage	-3.96	6.85	0.56862
Parks and Rec Fund	$-1.75\text{x}10^{-6}$	$2.42\text{x}10^{-6}$	0.47756
Library Fund	$8.18\ \text{x}10^{-6}$	$2.33\text{x}10^{-6}$	0.00177

Table 10. P-values for Initial OLS Model to Predict Homicide Rates

Although the regression equated an R^2 of 0.6288 some of the variables are not needed with the presence of other variables in the regression. Exploring the model in greater detail and eliminating unnecessary variables find that the only independent variable necessary to predict future homicide rates based on the trends of past homicide rates is, surprisingly, the Salinas library funding. The model derived from this is:

$$\hat{y}=-0.5852+6.130\text{x}10^{-6}(\text{Library Funding}) \quad (41)$$

Although the R^2 is slightly reduced from the previous model at 0.5969, it is still high enough to show an adequate fit. R^2 will increase with more variables, whether or not the variables are necessary. The p-value for library funding in the model was $5.67\text{x}10^{7}$. This model is not overfit. The RSE for the model was 4.21 and the cross-validation score was 4.33. A graphical depiction of the fit of this model to the homicide levels is displayed in Figure 11.

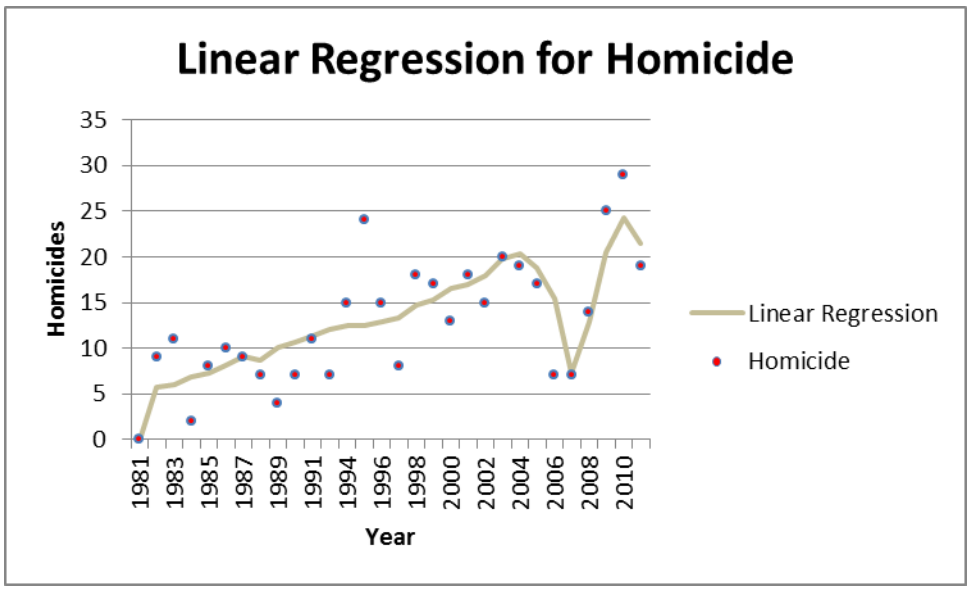

Figure 11. Graphical Depiction of OLS Regression to Predict Homicide Rates in Salinas

The OLS fit for homicide is good, but homicides could easily have a count based distribution, so an exploration of a Poission fit for homicide prediction is advisable and is covered in the next section.

5. Homicide Prediction using Poisson Regression

The Poisson model derived for homicide prediction started with the same variables the OLS model started with. The initial Poisson model was:

$$\begin{aligned}\hat{y} = \exp(&2.040 - 2.433\times10^{-6}(\text{Population}) \\ &+ 4.777\times10^{-1}(\text{SPD Budget}) \\ &- 0.1496(\text{CDCR Overpopulation Percentage}) \\ &- 1.437\times10^{-7}(\text{Parks and Rec Funding}) \\ &+ 5.718\times10^{-7}(\text{Library Funding}))\end{aligned} \tag{42}$$

Much like the OLS model, this first fit for the Poisson regression to predict homicide levels was not a very good fit. The p-values for the regression are given in Table 11.

	Estimate	Standard Error	P-value
Intercept	2.04	6.22×10^{-1}	0.001048
Population	-2.43×10^{-6}	9.15×10^{-6}	0.790286
SPD Budget	4.78×10^{-9}	1.33×10^{-8}	0.720189
CDCR Over Population Percentage	-1.50×10^{-1}	4.74×10^{-1}	0.752115
Parks and Rec Fund	-1.44×10^{-7}	1.54×10^{-7}	0.349927
Library Fund	5.72×10^{-7}	1.56×10^{-7}	0.000235

Table 11. P-values for Poisson Regression for Homicide Levels

The Poisson model, just like the OLS model, reduced to a regression with Salinas library funding as the single regressor. The model derived was:

$$\hat{y}=\exp(1.522+4.417 \times 10^{-7}(\text{Library Funding})) \quad (43)$$

The regression was checked for overfitting, but was found to not be overfit with a cross-validation score of 17.557 and a model RSE of 16.96. The graphical interpretation of the model is shown in Figure 12.

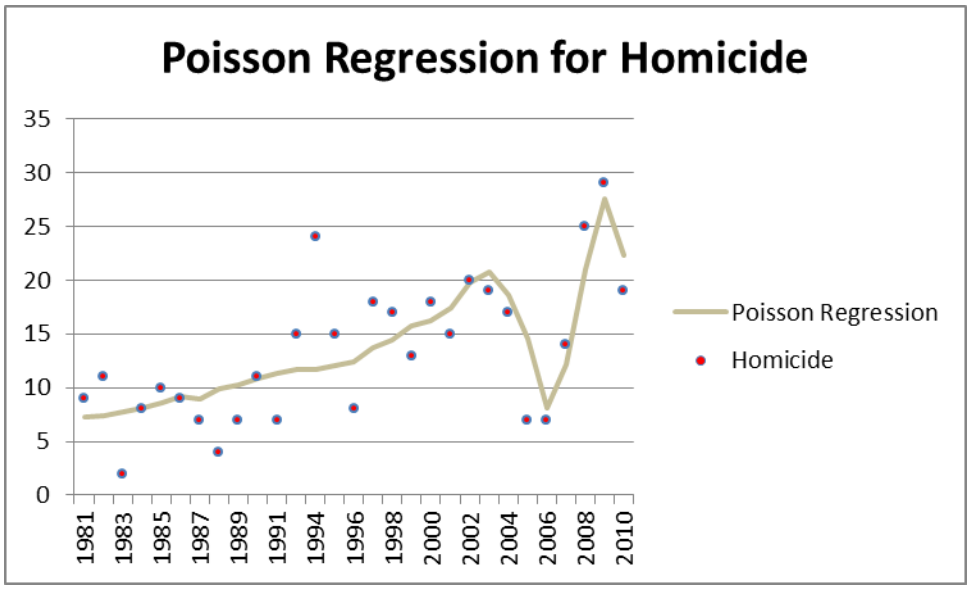

Figure 12. Poisson Regression fit for Homicide

The Poisson regression fit was an acceptable fit for the data, but there is the possibility of overdispersion. The residual deviance for the model is 41.196 with 28 degrees of freedom. Therefore, the negative binomial regression was explored as an alternative to the Poisson regression.

6. Homicide Prediction using Negative Binomial Regression

The starting point for the negative binomial regression for homicide prediction was the fit for the Poisson regression. This gave an equation of:

$$\hat{y}=\exp(1.520+4.425\text{x}10^{-7}(\text{Library Funding})) \quad (44)$$

This model is similar to the Poisson model with a lower residual deviance. The residual deviance of the negative binomial model was 34.36. This is a significant decrease from the Poisson model and suggests a better fit than the Poisson model for predicting homicide levels. The graphical fit is displayed in Figure 13.

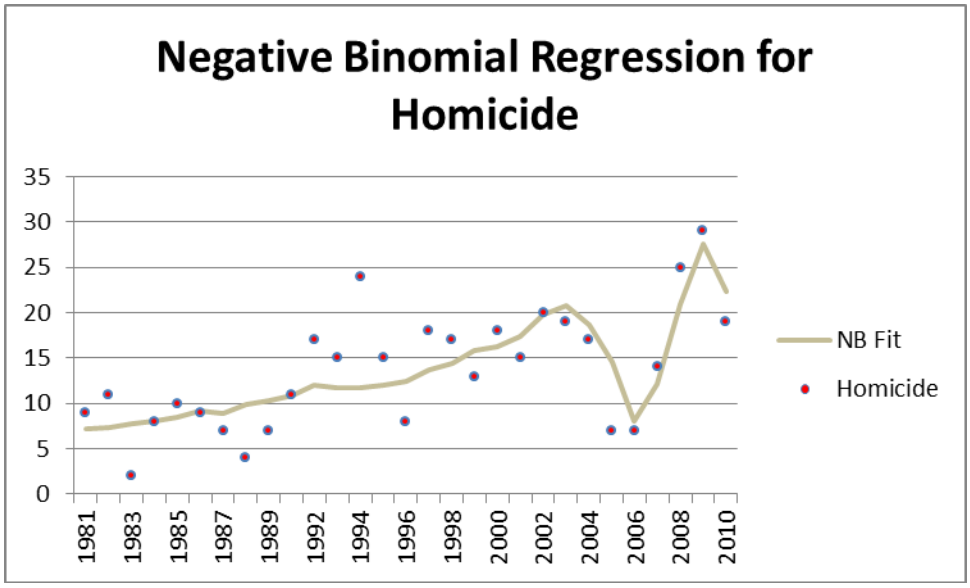

Figure 13. Negative Binomial Fit for Homicide

7. Assault and Homicide Prediction using Ordinary Least Squares

The initial fit to predict assaults and homicides was:

$$\hat{y}= 548.549-115.333(\text{CDCR Overpopulation Percentage}) -0.002447(\text{Parole Population}) +27.882(\text{Unemployment with two year shift}) \quad (45)$$

with p-values of:

	Estimate	Standard Error	P-value
Intercept	5.49×10^{2}	4.31×10^{2}	0.22266
CDCR Over Population Percentage	1.15×10^{2}	2.98×10^{2}	0.70399
Parolee Population	-2.45×10^{-3}	2.27×10^{-3}	0.29801
Unemployment with 2 year shift	2.79x10	8.85	0.00662

Table 12. P-values for OLS Regression for Assault and Homicide prediction

The R^2 for the initial regression fit was 0.635. Although this seems to be a good fit, the regression to predict assault and homicide levels was dominated by unemployment with a 2-year shift, much like homicides and library funding. However, unlike the homicide regression, this equation benefited from an addition of a squared unemployment term. The ideal OLS regression equation was:

$$\hat{y}=1330.385 -155.163(\text{Unemployment with two year shift}) +9.593(\text{Unemployment with two year shift})^2 \quad (46)$$

This regression was not overfit, having a cross-validation score of 56.82 and an RSE for the model of 52.31. This equation resulted in an R^2 of 0.7117 and a graphical interpretation shown in Figure 14.

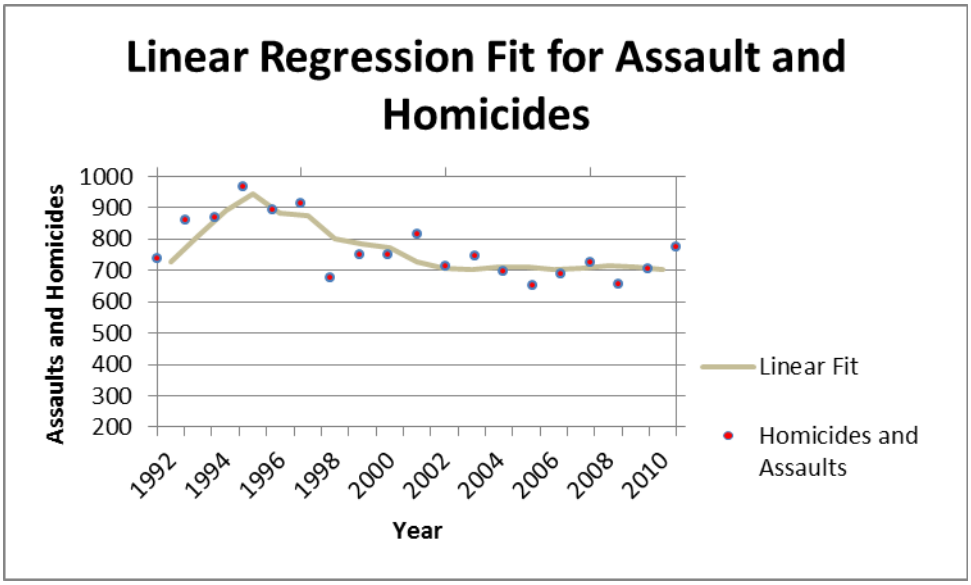

Figure 14. OLS Regression for Assaults and Homicides

This was a good fit for the prediction. Poisson and negative binomial models were also explored.

8. Assault and Homicide Prediction using Poisson Regression

The initial fit for the Poisson regression to predict assaults and homicides was:

$$\hat{y}= \exp(6.324-1.690x10^{-1}(\text{CDCR Overpopulation Percentage}) \\ -3.226x10^{-6}(\text{Parole Population}) \\ +3.622x10^{-2}(\text{Unemployment with two year shift})) \quad (47)$$

with p-values of:

	Estimate	Standard Error	P-value
Intercept	6.32	$2.55x10^{-1}$	$2.00x10^{-16}$
CDCR Over Population Percentage	$1.69x10^{-1}$	$1.78x10^{-1}$	0.3413
Parolee Population	$-3.23x10^{-6}$	$1.36x10^{-6}$	0.0174
Unemployment with 2 year shift	$3.62x10^{-2}$	$5.25x10^{-3}$	$5.42x10^{-12}$

Table 13. P-values for OLS Regression for Assault and Homicide prediction

This model suggests that CDCR Overpopulation does not add to the regression. This variable was taken out. A squared term of unemployment was added to the model to give the following equation for the optimal Poisson regression fit:

$$\hat{y}= \exp(7.470 - 1.791x10^{-6}(\text{Parole Population}) \\ -1.730x10^{-1}(\text{Unemployment with two year shift}) \\ +1.063x10^{-2}(\text{Unemployment with two year shift})^2) \quad (48)$$

The model is not overfit, with a cross-validation score of 3056.97 and a model RSE of 2639.76. A graphical representation of the model is given by Figure 15.

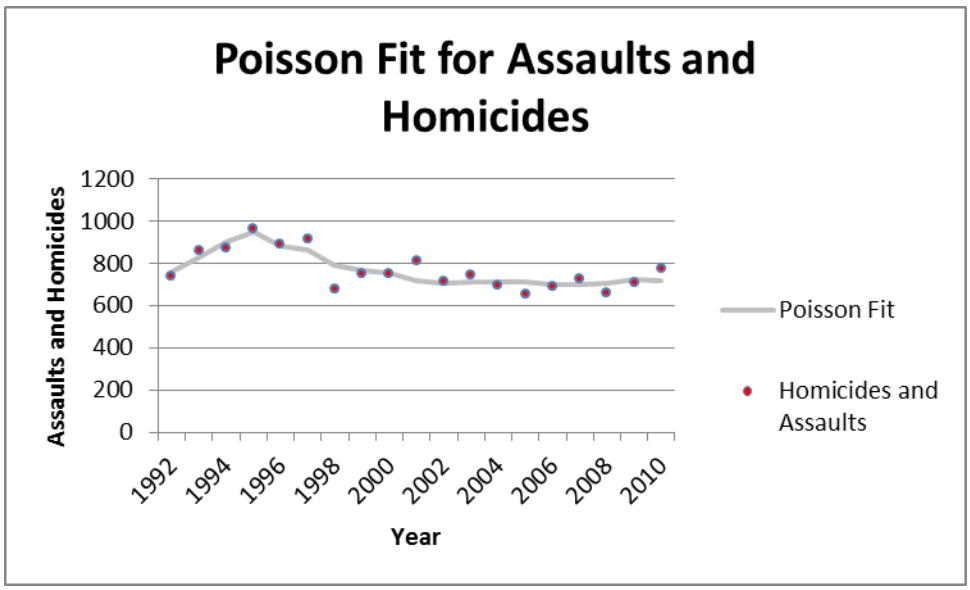

Figure 15. Poisson Fit for Assaults and Homicides

This equation has a residual deviance of 52.856 with 15 degrees of freedom. There is strong evidence of overdispersion, which will be remedied with the negative binomial regression.

9. Assault and Homicide Prediction using Negative Binomial Regression

The initial fit, using the Poisson regression as a basis for the negative binomial regression was:

$$\begin{aligned}\hat{y} = \exp(&7.448 - 1.771\text{x}10^{-6}(\text{Parole Population}) \\ &-1.687\text{x}10^{-1}(\text{Unemployment with two year shift}) \\ &+1.04\text{x}10^{-2}(\text{Unemployment with two year shift})^2)\end{aligned} \tag{49}$$

This model also seems very close to the Poisson model but reduces the residual deviance from 52.856 to 19.16, implying a much better fit for the data using the negative binomial over the Poisson regression. The visual for the fit is depicted in Figure 16.

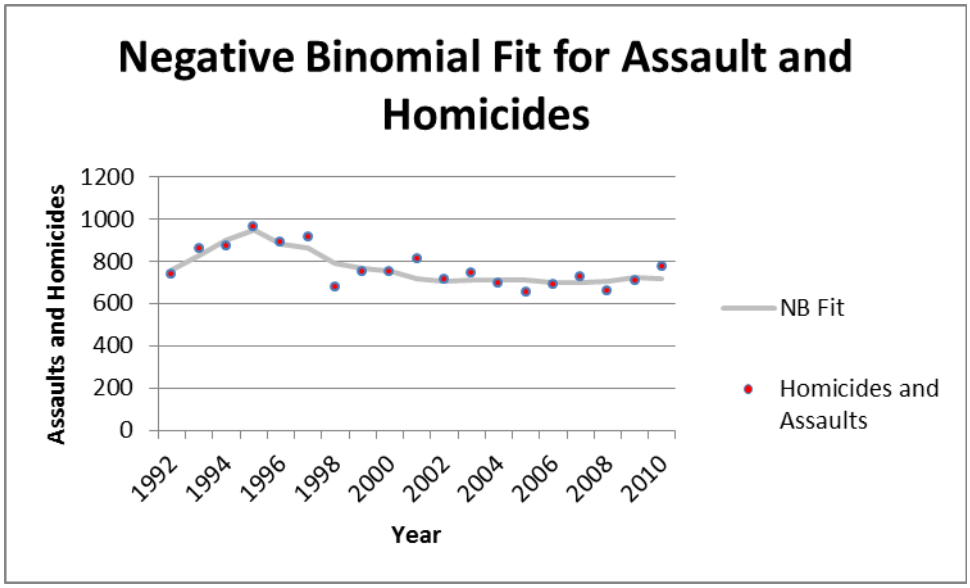

Figure 16. Negative Binomial Fit for Assault and Homicides

D. PREDICTION RESULTS USING REGRESSION MODELS

To test the validity of the models, 2011 data was gathered or, if the data was not yet published, estimated with the best available knowledge and past trends. 2011 and 2012 were both predicted, with the 2012 prediction results using model equated using 2011 data and is detailed in Appendix B. Although the models will predict as many years in the future as data is input, it is unwise to attempt to predict too far into the future with a regression model as changes to policy or environmental factors could change at any time. None of these models should be used to estimate more than one or two years of violence levels in Salinas.

Violence prediction for 2011 is detailed in Table 14.

	Salinas Violence
OLS Regression Model	1810.954
Poisson Regression Model	2117.753
Negative Binomial Regression Model	2117.753

Table 14. 2011 Violence Prediction based on the Derived Models

Homicide prediction is displayed in Table 15.

	Salinas Homicides
OLS Regression Model	22.64106
Poisson Regression Model	24.43126
Negative Binomial Regression Model	24.45902

Table 15. 2011 Homicide Prediction based on the Derived Models

Finally, assaults and homicide predictions for 2011 are in Table 16.

	Salinas Assaults and Homicides
OLS Regression Model	835.2094
Poisson Regression Model	836.4053
Negative Binomial Regression Model	836.1911

Table 16. 2011 Homicide and Assault Prediction based on the Derived Models

The actual counts for 2011 are detailed in Table 17.

Violence	1083
Homicides	15
Assaults and Homicides	709

Table 17. Observed 2011 Crime Statistics

The predicted numbers are higher than the actual numbers for 2011 crime statistics in Salinas.

IV. CONCLUSION AND FUTURE WORK

All three of the model types had similar predictions, with the exception being the model for violence. The different types of regression for predicting violence yielded varying results. The linear model for violence gave a closer prediction to the 2011 crime rates than the other two models. The Poisson and negative binomial models assume that the data has a Poisson distribution and, should the data not have a Poisson distribution, OLS regression will be as good, if not better, at prediction. This being the case, this study found that ordinary least squares models are adequate to predict crime trends in Salinas for all three of the explored dependent variables.

Several economic variables are highly correlated with crime statistics in Salinas. These variables could be used to predict future crime rates in Salinas based on past trends and observations in the city. However, these numbers do not take into account policy changes enacted in the city, such as Operation Ceasefire. These operations are difficult to numerically quantify in a study, and can very well be responsible for the reduction in crime levels in Salinas.

According to all of the models derived in this study, crime in Salinas should be on the rise in all categories. Salinas saw an increase in crime statistics from 2008-2010, but a reduction in crime in 2011. The most obvious conclusion to draw between the disparity between the statistical models and the actual crime levels is that Salinas is moving in the correct direction for crime prevention and gang reduction.

These results lend heavy credence to a continuation to the current crime prevention methods in Salinas to include the gang task force, Operation Ceasfire, coordination with CASP, and any other methods of crime reduction.

Although the 2011 predictions were incorrect, the 2012 calculations seem quite feasible and are listed in Appendix B. These numbers were derived from models that took into account 2011 economic and violence levels. The 2011

violence levels and environmental data may have aided in future predications and have included the efforts of the crime prevention tactics employed by Salinas.

A. FUTURE WORK

Future work in this area would apply the equated models of this thesis in different communities with smaller and larger populations to see if the models predict violence in communities other than Salinas. All of the variables in the study are present in any community, with the difference being the large gang presence in Salinas. It could be a significant study to see if the level of environmental variables effect crime levels, no matter the type of population.

Another topic for future work is further exploration of the prison overpopulation problem compared to crime rates in California. As of the date of this thesis, prison overcrowding is a very important governmental topic and California prison population and crime could be compared to neighboring states prison population and crime rates to see if there are correlations between crime rates and prison populations.

B. RECOMMENDATIONS

Salinas California should maintain its current level of diligence in crime deterrence. There is some variable not taken into account in this study that must account for the drop in crime rates from 2010 to 2011 in Salinas. Currently, Salinas's budget predictions for 2013 show a considerable reduction to the police force, down to 88 patrol police. This will decrease the police presence in Salinas to from 157 about 143 police officers. The population of Salinas does not show any sign of decreasing. This reduction in officers will equate to a ratio of around 1 officer per 1050 people in Salinas. For a city Salinas's size, the Bureau of Justice Statistics estimates the average to be 1.9 officers per 1000 residents (Reaves, 2007, p. 9).

A reduction in police force can have a devastating effect on crime rates in an already crime ridden community. In September of 2011, Governor Chris Christie of New Jersey balanced the New Jersey budget. The consequences of the balance were a reduction by 103 officers from the Trenton, NJ police force. As a result, Trenton PD has seen a drastic increase in crime rates, having almost a shooting a day, up from one a week, as reported from Sergeant Mark Kieffer, a 16-year veteran of Trenton PD (Glass, 2012). Although the proposed cuts in Salinas are not as drastic, the repercussions could be as dire as in Trenton.

No single variable in the study should be concentrated on as the fix to the crime problem. It may seem preposterous to think that increasing the Salinas library budget will increase the number of homicides and this could well be a case of correlation having little or no link to causation. However, the statistically observed correlation of homicides and library funding could also be an artifact of perceptions by the citizens of Salinas concerning city funding policies. The leadership of Salinas would be wise to consider possible second or third order effects during budget negotiations.

Many times in the study, the overpopulation of the California prison system was a variable of great concern. Overcrowding in prisons means early parole for prisoners. The parolees, when released, go back to their previous residence, which is also the place they committed the crime to lead to their prison sentence in the first place. The unemployment level in Salinas is currently around 15% and the recidivism rate for California is around 70%. These factors point towards a continual high level of crime in Salinas.

Finally, all findings from this study will be given to the Salinas leadership in an effort to assist Salinas's crime problem in any way possible. Although the predictions are just that, predictions, these tools can be used to help guide the administration and the budgeting department for Salinas into the future.

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

Actions CDCR Has Taken to Reduce Overcrowding. (2012). Retrieved 4/27, 2012 from http://www.cdcr.ca.gov/News/docs/FS-Actions-ReduceInmatePop.pdf.

Boba, R. (2005). Crime Analysis and Crime Mapping. Thousand Oaks: Sage Publications, Inc.

Braga, A. A., & Pierce, G. L. (2005). Disrupting Illegal Firearms Markets in Boston: The Effects of Operation Ceasefire on the Supply of New Handguns to Criminals. Criminology & Public Policy, 4(4), 717.

Burke, M., & Cavanaugh, M. (2011). New Calif. Prison Plan. Retrieved 4/27, 2012, from http://www.kpbs.org/news/2011/jun/07/new-calif-prison-plan/.

Cate, M. (2010). Corrections Year at a Glance. Sacramento, CA: California Department of Corrections and Rehabilitation.

Chapman, B. (1982). 1980 Census of Population. U.S. Department of Commerce.

Dobson, A. J., & Barnett, A. G. (2008). An Introduction to Generalized Linear Models Third Edition. Boca Raton: Taylor & Francis Group, LLC.

Egley, A. J., & O'Donnell, C. E. (2008). Highlights of the 2006 National Youth Gang Survey. OJJDP, 5.

Ehlers, D., & Pimstone, G. (1998). Predicting crime: A statistical glimpse of the future? Nedbank ISS Crime Index, 2(2).

Fetherolf, L. (2009). 90-day Report to the Community. Salinas, California: Salinas Police Department.

Fetherolf, L. (2010). Report to the Community. Salinas, CA: City of Salinas.

Glass, I. (2012). What Kind of Country. This American Life, 459, 4/27/2012.

Gorr, W., & Olligschlaeger, A. (2002). Crime Hot Spot Forecasting: Modeling and Comparative Evaluation, Final Project Report. (No. 195167).U.S. Department of Justice.

Grassie, R. P., Waymire, R. V., Burrows, J. W., Anderson, C. L., & Wallace, W. D. (1977). Integrated Criminal Apprehension Program - Crime Analysis - Executive Manual [Abstract].

Hennessey, V. (2003). An End to the Cycle. Monterey Herald.

History of Salinas. (2009). Retrieved 4/27, 2012, from http://www.ci.salinas.ca.us/visitors/history.cfm.

Kennedy, D. M., Braga, A. A., & Piehl, A. M. (2001). Reducing Gun Violence The Boston Gun Project's Operation Ceasefire. U.S. Department of Justice.

Mamalian, C. A., & LaVigne, N. G. (1999). The Use of Computerized Crime Mapping by Law Enforcement: Survey Results. National Institute of Justice Research Preview,

McFarlane, A. M. (2012). SOCDS Census Data: Output for Monterey city, CA. Retrieved 4/27, 2012, from http://socds.huduser.org/Census/Census_java.html.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2006). Introduction to Linear Regression Analysis (4th ed.). Hoboken: John Wiley & Sons, Inc.

Morales, G., Eways, A., Novotny, B., & Schoville, C. (2008). *Sureños 2008: EME.* Phoenix, AZ: Rocky Mountain Information Network.

Omnibus Crime Control and Safe Streets act of 1968, P.L. 90-351, (1968).

Onufer, T. L., & Clark, J. A. (2009). Understanding Environmental Factors that Affect Violence in Salinas, California. (Unpublished Master of Science in Defense Analysis). Naval Postgraduate School, Monterey, CA.

Osgood, W. D. (2000). Poisson-Based Regression Analysis of Aggregate Crime Rates. Journal of Quantitative Criminology, 16(1).

O'Shea, T. C., & Nicholls, K. (2003). Crime Analysis in America Finding and Recommendations. U.S. Department of Justice.

Pepper, J. V. (2007). Forecasting Crime: A City Level Analysis. University of Virginia.

Record on Research About Criminal Behavior Corrected. (2009). Retrieved 4/27, 2012, from http://www.rand.org/news/press/2009/07/24.html.

Reaves, B. A. (2007). Local Police Departments, 2007. Bureau of Justice Statistics.

Reynolds, J. (2011, 25 May). Operation Knockout: Gang raid targets Nuestra Familia in Salinas. Monterey Herald.

San Francisco Citizen. (2010, 22 April). Jerry Brown Takes Down: “Operation Knockout” Arrests 94 Norteños and Sureños in Salinas. Message posted to http://sfcitizen.com/blog/2010/04/22/jerry-brown-takes-down-operation-knockout-arrests-94-in-the-salinas-area.

Seavey, K. (2010). A Short History of Salinas, California. Retrieved 5/19, 2010, from http://www.mchsmuseum.com/salinasbrief.html.

Simple Linear Regression Analysis. (2012). Retrieved 4/27, 2012, from http://www.weibull.com/DOEWeb/simple_linear_regression_analysis.htm.

Solana, K. (2010). Salinas Ceasefire call-in draws 30 gang members, a few juveniles. Retrieved 4/27, 2012, from http://www.thecalifornian.com/article/20100506/NEWS09/5060301/Salinas-Ceasefire-call-draws-30-gang-members-few-juveniles.

Stahl, Z. (2009). Saving Salinas. Retrieved 4/27, 2012, from http://donohueformayor.com/node/97.

State & County QuickFacts: Salinas, California. (2012). Retrieved 4/27, 2012, from http://quickfacts.census.gov/qfd/states/06/0664224.html.

Success Stories. (2012). Retrieved 4/27, 2012, from http://www.guncite.com/success.htm.

Sureños. (2005). Retrieved 05/03, 2012, from http://www.sampsonsheriff.com/otherforms/20051011_surenos.pdf.

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX A. DATA

All data in the study was taken from government sources, when available, and from the past research from Clark and Onufer, when the government sources did not have the data. All of the links to data sources are listed below the table.

Year	Population	Homicides	Robbery	Assault	Violence
1980	80479	9	208	283	500
1981	82700	9	191	344	544
1982	85300	11	179	356	546
1983	87600	2	200	372	574
1984	91100	8	159	362	529
1985	94600	10	167	424	601
1986	98300	9	204	672	885
1987	100800	7	192	633	832
1988	103900	4	217	722	943
1989	105400	7	217	734	958
1990	108777	11	262	778	1051
1991	111184	7	253	805	1065
1992	114736	17	388	722	1127
1993	116686	15	560	844	1419
1994	120885	24	414	846	1284
1995	121960	15	494	950	1459
1996	124972	8	412	884	1304
1997	127369	18	348	895	1261
1998	132449	17	440	661	1118
1999	136797	13	346	737	1096
2000	142685	18	443	734	1195
2001	144728	15	399	799	1213
2002	146659	20	367	692	1079
2003	148117	19	399	725	1143
2004	149838	17	452	678	1147
2005	149626	7	335	645	987
2006	148707	7	383	683	1073
2007	148782	14	378	711	1103
2008	150898	25	334	633	992
2009	150215	29	359	678	1066

2010	150441	19	365	755	1139
2011	150441	15	374	694	1083

Year	Drop Outs	Drop Out Rate	SPD Budget	SPD Employees	Sworn Police
1980	NA	NA	5342123	NA	NA
1981	NA	NA	5726778	NA	NA
1982	NA	NA	6170072	173	NA
1983	NA	NA	6645438	177	NA
1984	NA	NA	7715577	180.5	NA
1985	NA	NA	8331832	180.5	NA
1986	NA	NA	9315054	184	NA
1987	NA	NA	9607781	180	NA
1988	NA	NA	10177311	184	NA
1989	NA	NA	10649699	184	NA
1990	NA	NA	11456256	186	NA
1991	NA	NA	13144292	186	NA
1992	157	2.6	13634881	187	NA
1993	237	3.7	15683718	181	NA
1994	157	2.4	13612478	179	NA
1995	198	2.9	14471238	188	NA
1996	330	4.7	16393545	193	NA
1997	281	3.8	16929407	198	147
1998	342	4.5	17575700	198	144
1999	256	3.2	18852899	199	143
2000	203	2.5	19288170	213	145
2001	225	2.6	21713995	221	149
2002	204	2.3	22040439	222	145
2003	75	0.08	24224300	224	154
2004	124	1.3	25241659	222	167
2005	85	0.9	29704910	232	164
2006	124	1.3	33356709	238	161
2007	180	2.6	35416564	255	174
2008	147	1.5	38380314	251	177
2009	264	2.8	41187794	251	164
2010	276	2.9	37360500	230	157
2011	276	2.9	39852481	210	157

Year	CDCR Capacity	CDCR Population	CDCR Overpopulation Percentage	Parole Population
1980	23534	23371	0.99307385	14650
1981	23800	26372	0.99307385	13952
1982	24611	31319	1.07155337	16072
1983	25703	35965	1.2184959	22202
1984	26792	40524	1.34237832	27000
1985	29042	45528	1.39535845	30726
1986	32097	53620	1.67056111	34771
1987	36465	62949	1.72628548	43355
1988	44124	69695	1.57952588	52788
1989	47120	79849	1.69458829	61665
1990	51013	90405	1.77219532	73096
1991	54042	95930	1.77510085	85470
1992	57986	98386	1.6967199	89453
1993	61983	109654	1.76909798	88858
1994	66183	118968	1.79756131	92958
1995	70717	125585	1.77588133	96110
1996	73121	135294	1.85027557	100934
1997	75952	146656	1.93090373	105449
1998	79877	150731	1.88703882	111875
1999	79873	154440	1.93356954	117612
2000	80272	154014	1.91865158	121414
2001	80467	153649	1.90946599	121820
2002	79957	151579	1.89575647	117138
2003	80187	153783	1.91780463	114136
2004	80980	157895	1.94980242	113768
2005	81008	158837	1.96075696	115001
2006	87370	166547	1.90622639	121808
2007	84653	166277	1.96421863	126906
2008	84066	160169	1.90527681	123597
2009	84241	154749	1.83697962	109026
2010	84596	168830	2.00413101	108656
2011	84130	136619	1.623903483	100490

Year	Parks and Rec Fund	Library Fund	Unemployment	Number of Vacant Units	Person Per Household
1980	NA	NA	NA	NA	NA
1981	1443405	1032802	NA	NA	NA
1982	1533507	1077642	NA	NA	NA
1983	1704186	1203748	NA	NA	NA
1984	1734921	1278394	NA	NA	NA
1985	1956253	1410963	NA	NA	NA
1986	2257021	1584859	NA	NA	NA
1987	2223182	1518322	NA	NA	NA
1988	2674380	1748197	NA	NA	NA
1989	2768221	1825048	NA	NA	NA
1990	2178310	1955405	9.7	1228	3.21
1991	2587574	2059120	11.4	1227	3.24729
1992	2785467	2180700	12.5	1230	3.33079
1993	1992787	2121146	13.1	1234	3.36624
1994	1966659	2119116	12.4	1240	3.46189
1995	1542992	2191539	12.3	1251	3.45003
1996	1850056	2263639	11.3	1269	3.47405
1997	1871278	2489376	11	1281	3.49622
1998	2048408	2589735	10.8	1305	3.56309
1999	2131408	2799503	9.7	1314	3.60341
2000	2296515	2868795	7.4	1360	3.662
2001	2817339	3020075	7.8	1370	3.69
2002	2974138	3316832	8.9	1385	3.702
2003	3081427	3423623	9	1400	3.7
2004	2795909	3170427	8.3	1418	3.699
2005	2285817	2614595	7.3	1433	3.654
2006	2082617	1278414	6.9	1441	3.614
2007	3153973	2207708	7.1	1450	3.601
2008	3893586	3440113	8.4	1452	3.637
2009	3976221	4061128	11.8	1463	3.643
2010	3302147	3587431	12.8	1462	3.685
2011	1571796	3788695	12.4	1462	4

All links verified as of 20 April 2012

Salinas Population 1980.
http://www.dof.ca.gov/research/demographic/reports/estimates/e-4/1971–80/counties-cities/

Salinas Population 1981–1989. Retreived 20 April 2012:
http://www.dof.ca.gov/research/demographic/reports/estimates/e-4/1981–90/

Salinas Population 1981–1990
http://www.dof.ca.gov/research/demographic/reports/estimates/e-4/1981–90/

Salinas population 1990–2000
http://www.dof.ca.gov/research/demographic/reports/estimates/e-8/

Salinas population 2000–2010
http://www.dof.ca.gov/research/demographic/reports/estimates/e-5/2001-10/view.php

Vacant Houses: 1990–2000. Used total houses minus Occupied houses
http://www.dof.ca.gov/research/demographic/reports/estimates/e-8/

Vacant Houses: 2001–2010. Used total houses minus occupied houses
http://www.dof.ca.gov/research/demographic/reports/estimates/e-5/2001–10/view.php

Person Per Household 1990–2000
http://www.dof.ca.gov/research/demographic/reports/estimates/e-8/

Person per household 2001–2010
http://www.dof.ca.gov/research/demographic/reports/estimates/e-5/2001–10/view.php

Prison and parolee populations: 1980–2009:
http://www.cdcr.ca.gov/Reports_Research/Offender_Information_Services_Branch/Annual/CalPrisArchive.html

Prison and parolee Population: 2009–2011
http://www.cdcr.ca.gov/Reports/CDCR-Annual-Reports.html

Police, Library, and Parks and Recreation 1981–2004:
Clark and Onufer thesis

Police, Library, and Parks and Recreation 2005-2011: Taken from column labeled “Actual.” 2011 taken from column labeled “Budget.” 2012 taken from column labeled “Adopted.”
http://www.ci.salinas.ca.us/services/finance/budget.cfm
SPD Employee and SPD Sworn Police numbers taken from Clark and Onufer Thesis.

Crime Stats 1985–2010:
http://www.ucrdatatool.gov/Search/Crime/Local/RunCrimeJurisbyJurisLarge.cfm

Crime statistics 2011:
http://www.salinaspd.com/statistics.html

School drop-out information 1991–2010. Taken from column Grade 9-12 total drop-outs. 2011–2012 not currently published and estimated to be approximately 2010 level:
http://dq.cde.ca.gov/dataquest/DropoutReporting/DropoutsByGrade.aspx?cDistrictName=SALINAS%20UNION%20HIGH%20%20%20%20%20%20%20%20%20%20%20%20&cCountyCode=27&cDistrictCode=2766159&cSchoolCode=0000000&Level=District&TheReport=GradeOnly&ProgramName=All&cYear=2009-10&cAggSum=DTotGrade&cGender=B

Unemployment 1990-2011. 2012 estimated from current 2012 data:
http://www.calmis.ca.gov/file/lfhist/LabForce-CAMSACo.txt

APPENDIX B. DERIVED MODELS AND PREDICTIONS

There were nine different models derived in the study. The models were all equated using the data found in Appendix I. All of the models were equated in R.

A. VIOLENCE MODELS

OLS Model	$\hat{y}=1.827x10^{3}+9.384x10^{-6}$(SPD Budget) $-8.354x10^{2}$(CDCR Overpopulation Percentage) +50.62(Unemployment with Two Year Shift) +1.712(Number of Vacant Units with Two Year Shift) -13.93(SPD Sworn Police with One Year Shift)
Poisson Model	$\hat{y}=\exp(7.679-8.695x10^{-9}$(SPD Budget) $-1.287x10^{-2}$(SPD Sworn Police with one year shift) $-7.743x10^{-1}$(CDCR Overpopulation Percentage) $+4.666x10^{-2}$(Unemployment with two year shift) $+1.582x10^{-3}$(Number of Vacant Units with two year shift))
Negative Binomial Model	Same as Poisson Model

B. HOMICIDE MODELS

OLS Model	$\hat{y}=-0.5852+6.130x10^{-6}$(Library Funding)
Poisson Model	$\hat{y}=\exp(1.522+4.417x10^{-7}$(Library Funding))
Negative Binomial Model	$\hat{y}=\exp(1.520+4.425x10^{-7}$(Library Funding))

C. HOMICIDE AND ASSAULT MODELS

OLS Model	..
Poisson Model	$\hat{y}$= exp(7.470 − 1.791x10^{-6}(Parole Population) −1.730x10^{-1}(Unemployment with two year shift) +1.063x10^{-2}(Unemployment with two year shift)^2)
Negative Binomial Model	$\hat{y}$= exp(7.448 − 1.771x10^{-6}(Parole Population) −1.687x10^{-1}(Unemployment with two year shift) +1.04x10^{-2}(Unemployment with two year shift)^2)

D. PREDICTIONS USING MODELS:

The predictions for 2011 were made with the derived models. The prediction for 2012 were made with models reformulated using the available 2011 data.

2011	OLS	Poisson	Negative Binomial	Recorded Levels
Violence	1789.337	2077.991	2077.991	1083
Homicides	22.64	24.43	24.46	15
Homicides and Assaults	881.40	879.80	879.02	664

2012	OLS	Poisson	Negative Binomial	Recorded Levels
Violence	1093.26	1094.10	1094.10	NA
Homicides	21.64318	22.92571	23.03067	NA
Homicides and Assaults	901.64993	916.76166	914.25029	NA

APPENDIX C. R-CODE

```
s=read.table("clipboard",header=T)
############Correlation for Violence Level
corallvio=setcor(s$Violence,s,2)
names(corallvio)=c("NoShift","OneYear","TwoYears")
###Correlation between possible regressors for Violence####
##This consists of: Population, SPD Budget,
###sworn police with a one year shift, CDCR Overpopulation Percentage,
###parole population, unemployment percentage with a two year shift,
###number of vacant units with a two year shift
###personnel per household with a two year shift.
cormat=data.frame(s$Population[3:31],s$SPDBudget[3:31],s$CDCRPercentage[
3:31],
s$ParolePop[3:31],
s$Police[2:30],s$Unemployment[1:29],
s$Vacant[1:29],s$PersonPerHouse[1:29])
corbetween=cor(cormat,use="pairwise.complete.obs")
######It is cleaner to work with the a new dataframe used only for violence. The
big dataset can be used
viodata=data.frame(s$Violence[3:31],s$SPDBudget[3:31],s$CDCRPercentage[3:
31],
s$Police[2:30],s$Unemployment[1:29],
s$Vacant[1:29])
names(viodata)=c("Violence","SPDBudget","CDCRPercentage","Police","Unempl
oyment","Vacant")
######Correlation for Homidices
corallhom=setcor(s$Homicide,s,iter=2)
colnames(corallhom)=c("NoShift","OneYear","TwoYears")
###Correlation between possible regressors for Homicide####
##This consists of: Population, SPD Budget,
###CDCR Overpopulation Percentage,
###Parks and Recreation budget
```

```
###library budget
##homdata will be used for the regression
homdata=data.frame(s$Homicide,s$Population,s$SPDBudget,s$CDCRPercenta
ge,
s$ParksandRec,s$Library)
names(homdata)=c("Homicide","Population","SPDBudget","CDCRPercentage","
ParksandRec","Library")
homcorbetween=cor(homdata,use="pairwise.complete.obs")
######Correlation for Assaults and Homicide
HA=s$Homicide+s$Assault
corallHA=setcor(HA,s,iter=2)
colnames(corallhom)=c("NoShift","OneYear","TwoYears")
###Correlation between possible regressors for Homicide and Assaults####
##This consists of: CDCR Overpopulation Percentage,
###parole population
###unemployment percentage with a two year shift
###number of vacant units
###person per household with a two year shift
##HAdata will be used for he regression
HA=s$Homicide+s$Assault
HAdata=data.frame(HA[3:31],s$CDCRPercentage[3:31],s$ParolePop[3:31],s$Un
employment[1:29],
s$Vacant[3:31],s$PersonPerHouse[1:29])
HAcorbetween=cor(HAdata,use="pairwise.complete.obs")
###Since we are not going to use Vacant Units or Person Per Household, lets
put the first 2 obs back into the data set
HAdata=data.frame(HA[3:31],s$CDCRPercentage[3:31],s$ParolePop[3:31],s$Un
employment[1:29])
names(HAdata)=c("HA","CDCRPercentage","ParolePop","Unemployment")
############Regression fitting for Violence
###Linear to start
### SPD Budget, Sworn Police (1 year Shift), CDCR Overpopulation Percentage
## Unemployment (2 year shift), Vacant Units (2 year shift), Person per
household (2 year shift)
```

```
###first fit:
violm=lm(Violence~., data=viodata)
###To cross-validate, all of the rows with NAs must be taken out. this is the first 16 rows for this data set
crossviodata=viodata[17:29,]
violm=lm(Violence~.-PersonPerHouse,data=crossviodata)
###I like to run cross validation 10 times and take the mean
xstat=1:10
for(i in 1:10){
xstat[i]=xval(violm)}
mean(xstat)
###Second fit:
vio2lm=lm(Violence~SPDBudget+CDCRPercentage,data=s)
xstat=1:10
for(i in 1:10){
xstat[i]=xval(vio2lm)}
mean(xstat)
#####Predictions for the graphing of the two fits:
pviolm=as.matrix(predict(violm))
pvio2lm=as.matrix(predict(vio2lm))
####Violence Poisson Regression
viopoiss=glm(Violence~.-PersonPerHouse,data=viodata,family=poisson)
crossglm(viopoiss)
pviopoiss=as.matrix(predict(viopoiss,type='response'))
####Violence NB Regression
vionb=glm.nb(Violence~.-
PersonPerHouse,data=viodata,link=log,start=viopoiss$coefficients)
#############Regression fitting for Homicide############################
###Linear to start
###first fit:
homlm=lm(Homicide~., data=homdata)
####Final Fit:
```

```
homlm=lm(Homicide~Library,data=homdata)
###Cross-validate
###I like to run cross validation 10 times and take the mean
xstat=1:10
for(i in 1:10){
xstat[i]=xval(homlm)}
mean(xstat)
#####Predictions for the graphing of the linear model:
phomlm=as.matrix(predict(homlm))
####Violence Poisson Regression##
hompoiss=glm(Homicide~.,data=homdata,family=poisson)
###First one is no good
##Here is the final model
hompoiss=glm(Homicide~Library,data=homdata,family=poisson)
crossglm(hompoiss)
phompoiss=as.matrix(predict(hompoiss,type='response'))
####Violence NB Regression
homnb=glm.nb(Homicide~Library,data=homdata,link=log,start=hompoiss$coefficients)
phomnb=as.matrix(predict(homnb,type="response"))
#############Regression fitting for Homicide and Assaults##########################
###Linear to start
###first fit:
HAlm=lm(HA~., data=HAdata)
####Final Fit:
HAdata=HAdata[complete.cases(HAdata[,]),]
HAlm=lm(HA~Unemployment+I(Unemployment^2),data=HAdata)
###To cross-validate, all of the rows with NAs must be taken out.
###I like to run cross validation 10 times and take the mean
xstat=1:10
for(i in 1:10){
```

```
xstat[i]=xval(HAlm)}
mean(xstat)
#####Predictions for the graphing of the linear model:
pHAlm=as.matrix(predict(HAlm))
####Homicide and Assault Poisson Regression##
HApoiss=glm(HA~.,data=HAdata,family=poisson)
###First one is no good
##Here is the final model

HApoiss=glm(HA~ParolePop+Unemployment+I(Unemployment^2),data=HAdata,
family=poisson)
crossglm(HApoiss)
pHApoiss=as.matrix(predict(HApoiss,type='response'))
####Violence NB Regression
HAnb=glm.nb(HA~ParolePop+Unemployment+I(Unemployment^2),data=HAdata
,link=log,start=HApoiss$coefficients)
pHAnb=as.matrix(predict(HAnb,type="response"))
##########This is the future predictions
s2011=read.table("clipboard",header=T)
plmvio2011=predict(violm,newdata=data.frame(SPDBudget=s2011$SPDBudget,
CDCRPercentage=s2011$CDCRPercentage,
Police=s2011$Police,Unemployment=s2011$Unemployment,Vacant=s2011$Vac
ant, PersonPerHouse=s2011$PersonPerHouse))
pviolm2011=predict(violm,newdata=s2011)
pviopoiss2011=predict(viopoiss, newdata=s2011,type='response')
phomlm2011=predict(homlm,newdata=s2011)
phompoiss2011=predict(hompoiss,newdata=s2011,type='response')
phomnb2011=predict(homnb,newdata=s2011,type='response')
pHAlm2011=predict(HAlm,newdata=s2011)
pHApoiss2011=predict(HApoiss,newdata=s2011,type='response')
pHAnb2011=predict(HAnb,newdata=s2011,type='response')
###########Here are the models with 2011 data added in and regenerated: It
doesn't change the models very much
```

```
s=rbind(s,s2011)
viodata=data.frame(s$Violence[3:32],s$SPDBudget[3:32],s$CDCRPercentage[3:
32],
s$Police[2:31],s$Unemployment[1:30],
s$Vacant[1:30])
names(viodata)=c("Violence","SPDBudget","CDCRPercentage","Police",
"Unemployment","Vacant")
viodata=viodata[complete.cases(viodata[,]),]
homdata=data.frame(s$Homicide,s$Library)
names(homdata)=c("Homicide","Library")
homdata=homdata[complete.cases(homdata[,]),]
HA=s$Homicide+s$Assault
HAdata=data.frame(HA[3:32],s$ParolePop[3:32],s$Unemployment[1:30])
names(HAdata)=c("HA","ParolePop","Unemployment")
HAdata=HAdata[complete.cases(HAdata[,]),]
violm=lm(Violence~.,data=viodata)
viopoiss=glm(Violence~.,data=viodata,family=poisson)
vionb=glm.nb(Violence~.,data=viodata,link=log,start=viopoiss$coefficients)
homlm=lm(Homicide~Library,data=homdata)
hompoiss=glm(Homicide~Library,data=homdata,family=poisson)
crossglm(hompoiss)
homnb=glm.nb(Homicide~Library,data=homdata,link=log,start=hompoiss$coeffici
ents)
HAlm=lm(HA~Unemployment+I(Unemployment^2),data=HAdata)
HApoiss=glm(HA~ParolePop+Unemployment+I(Unemployment^2),data=HAdata,
family=poisson)
HAnb=glm.nb(HA~ParolePop+Unemployment+I(Unemployment^2),data=HAdata
,link=log,start=HApoiss$coefficients)
#Code for 2012 predictions. There are no 2012 actual numbers yet. most of the
2012 data is estimated
s2012=read.table('clipboard',header=T)
pviolm2012=predict(violm,newdata=s2012)
pviopoiss2012=predict(viopoiss, newdata=s2012,type='response')
```

```
phomlm2012=predict(homlm,newdata=s2012)
phompoiss2012=predict(hompoiss,newdata=s2012,type='response')
phomnb2012=predict(homnb,newdata=s2012,type='response')
pHAlm2012=predict(HAlm,newdata=s2012)
pHApoiss2012=predict(HApoiss,newdata=s2012,type='response')
pHAnb2012=predict(HAnb,newdata=s2012,type='response')
predictions=as.matrix(c(pviolm2012,pviopoiss2012,phomlm2012,phompoiss2012
,phomnb2012,
pHAlm2012,pHApoiss2012,pHAnb2012))
```

THIS PAGE INTENTIONALLY LEFT BLANK

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California