Crime Trend Prediction Using Regression Models for Salinas, California

Shingleton, Jarrod S.

Monterey, California. Naval Postgraduate School

http://hdl.handle.net/10945/7416
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
Crime Trend Prediction Using Regression Models for Salinas, California

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.

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
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
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.
# TABLE OF CONTENTS

I. INTRODUCTION ............................................................................................. 1  
   A. PREVIOUS RESEARCH ...................................................................... 1  
   B. RESEARCH OBJECTIVE ................................................................. 2  
   C. BACKGROUND ................................................................................... 2  
      1. History of Violence in Salinas ................................................. 2  
      2. Current and Past Efforts to Reduce Violence in Salinas...... 7  
   D. REGRESSION ANALYSIS TO PREDICT CRIME ............................... 9  

II. SUMMARY OF REGRESSION ..................................................................... 13  
   A. ORDINARY LEAST SQUARES METHOD ........................................ 13  
   B. RELATIONSHIPS AMONG VARIABLES .......................................... 15  
      1. Correlation .............................................................................. 15  
      2. Transforming Dependent and Independent Variables……… 16  
   C. VARIABLE SELECTION FOR REGRESSION .............................. 16  
      1. Hypothesis Test for Regression ........................................... 17  
      2. R-Squared ($R^2$) .................................................................. 18  
      3. Residual Standard Error .......................................................... 19  
      4. Analysis of Variance .............................................................. 19  
   D. GENERALIZED LINEAR MODELS ................................................... 20  
      1. Poisson Regression .............................................................. 21  
      2. Negative Binomial Regression ............................................. 23  
   E. CROSS VALIDATION ........................................................................ 25  

III. DATA ANALYSIS ......................................................................................... 27  
   A. DATA COLLECTION ......................................................................... 27  
   B. CORRELATION OF VARIABLE ANALYSIS ..................................... 28  
   C. REGRESSION ANALYSIS ................................................................ 34  
      1. Violence Prediction using Ordinary Least Squares .......... 34  
      2. Violence Prediction using Poisson Regression............... 37  
      3. Violence Prediction using Negative Binomial Regression. 39  
      4. Homicide Prediction using Ordinary Least Squares ....... 39  
      5. Homicide Prediction using Poisson Regression .............. 41  
      6. Homicide Prediction using Negative Binomial Regression .......... 43  
      7. Assault and Homicide Prediction using Ordinary Least Squares ................................................................. 43  
      8. Assault and Homicide Prediction using Poisson Regression ................................................................. 45  
      9. Assault and Homicide Prediction using Negative Binomial Regression ................................................................. 46  
   D. PREDICTION RESULTS USING REGRESSION MODELS .............. 47  

IV. CONCLUSION AND FUTURE WORK.......................................................... 49  
   A. FUTURE WORK................................................................................. 50
B. RECOMMENDATIONS ........................................................................................................ 50

LIST OF REFERENCES .......................................................................................................... 53

APPENDIX A. DATA ............................................................................................................... 57

APPENDIX B. DERIVED MODELS AND PREDICTIONS ...................................................... 63
   A. VIOLENCE MODELS ...................................................................................................... 63
   B. HOMICIDE MODELS ..................................................................................................... 63
   C. HOMICIDE AND ASSAULT MODELS ......................................................................... 64
   D. PREDICTIONS USING MODELS: .................................................................................. 64

APPENDIX C. R-CODE ......................................................................................................... 65

INITIAL DISTRIBUTION LIST ............................................................................................. 73
LIST OF FIGURES

Figure 1. Salinas Homicides versus U.S. Average Homicides 1980–2010 .......... 4
Figure 2. Salinas Robberies versus U.S. Average Robberies 1980–2010 .......... 4
Figure 3. Salinas Assaults versus U.S. Average Assaults 1980–2010 .......... 5
Figure 4. Graphical Representation of the p-value ............................................ 18
Figure 5. Graphical Representation of Correlation between Independent Variables and Violence. Violence in Red ............................................ 30
Figure 6. Graphical Correlation between Homicide and Independent Variables. Homicides in Red .............................................................. 32
Figure 7. Graphical Correlation between Homicide and Assaults and Independent Variables. Homicides and Assaults in Red .................... 34
Figure 8. Regression Fit for Formula (39) .......................................................... 36
Figure 9. Regression Fit for Formula (40) .......................................................... 37
Figure 10. Poisson Regression fit for Violence Prediction ................................. 38
Figure 11. Graphical Depiction of OLS Regression to Predict Homicide Rates in Salinas .......................................................... 41
Figure 12. Poisson Regression fit for Homicide ................................................ 42
Figure 13. Negative Binomial Fit for Homicide .................................................. 43
Figure 14. OLS Regression for Assaults and Homicides .................................. 44
Figure 15. Poisson Fit for Assaults and Homicides .......................................... 46
Figure 16. Negative Binomial Fit for Assault and Homicides ........................... 47
# LIST OF TABLES

Table 1. Correlation of Independent Variables and Violence ........................... 28
Table 2. Correlation between Independent Variables for OLS Regression to Predict Violence ................................................................. 29
Table 3. Correlation between Dependent Variables and Homicide Events in Salinas ............................................................................... 31
Table 4. Correlation between Independent Variables for Homicide Regression ......................................................................................... 31
Table 5. Correlation of Variables against Assaults and Homicides .......... 33
Table 6. Correlation between Independent Variables for Homicide and Assault Regression ........................................................................ 33
Table 7. P-Values for the initial OLS regression ........................................... 35
Table 8. Second Fit OLS Regression for Violence P-Values ......................... 36
Table 9. Initial Poisson Regression for Violence ......................................... 38
Table 10. P-values for Initial OLS Model to Predict Homicide Rates ............ 40
Table 11. P-values for Poisson Regression for Homicide Levels .................. 42
Table 12. P-values for OLS Regression for Assault and Homicide prediction 44
Table 13. P-values for OLS Regression for Assault and Homicide prediction 45
Table 14. 2011 Violence Prediction based on the Derived Models ............... 47
Table 15. 2011 Homicide Prediction based on the Derived Models .......... 48
Table 16. 2011 Homicide and Assault Prediction based on the Derived Models 48
Table 17. Observed 2011 Crime Statistics ................................................... 48
ACKNOWLEDGMENTS

My lovely wife, Courtney

The NPS Mathematics Department

Dr. Lyn Whitaker

Rebecca Lorentz, NPS Defense Analysis Department

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

Georgina Mendoza, Senior Dep. City Attorney at City of Salinas
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
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” (Solana, 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 everyday 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.
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 $\varepsilon_i, \ldots, \varepsilon_n$ are independent and normally distributed with mean 0 and variance $\sigma^2$. The $\beta$s are estimated by minimizing the error or residual sums of squares:

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

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

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

and

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

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

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 $\beta$s) in matrix form. The model (1) can be expressed as
\[ \mathbf{y} = \mathbf{X}\beta + \mathbf{\varepsilon} \]  \hspace{1cm} (4)

where \( \mathbf{y} \) is the \( n \times 1 \) vector of observations, \( \mathbf{X} \) is an \( n \times (k+1) \) matrix of independent variables (and an extra column of 1s for the intercept \( \beta_0 \)), \( \beta \) is a \( (k+1) \times 1 \) vector of coefficients and \( \mathbf{\varepsilon} \) is an \( n \times 1 \) vector of independent and identically distributed errors associated with (1).

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

\[
S(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) \\
= y^T \mathbf{y} - \beta^T \mathbf{X}^T \mathbf{y} - \mathbf{y}^T \mathbf{X}\beta + \beta^T \mathbf{X}^T \mathbf{X}\beta \\
= y^T \mathbf{y} - 2\beta^T \mathbf{X}^T \mathbf{y} + \beta^T \mathbf{X}^T \mathbf{X}\beta
\]  \hspace{1cm} (5)

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

\[
\frac{\partial S}{\partial \beta} \bigg|_{\hat{\beta}} = -2\mathbf{X}^T \mathbf{y} + 2\hat{\beta}^T \mathbf{X}^T \mathbf{X}\hat{\beta} = 0
\]  \hspace{1cm} (6)

where \( \mathbf{0} \) is the \( (k+1) \times 1 \) vector of 0’s. This equation can be simplified to:

\[
\mathbf{X}^T \mathbf{X}\hat{\beta} = \mathbf{X}^T \mathbf{y}
\]  \hspace{1cm} (7)

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

\[
\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}
\]  \hspace{1cm} (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 \( \beta \).
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 \((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\). The correlation between these two variables is measured using the sample correlation coefficient with the formula:

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

(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 $X^T X$ of equation (9) to be ill-
conditioned. Further, perfect linear dependence among the independent
variables will cause $X^T X$ to be singular and give infinite least squares estimates
of $\hat{\beta}$ 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:

\[ H_0 : \beta_j = 0, \quad H_1 : \beta_j \neq 0 \]  

(10)

The equation used to test this hypothesis is:

\[ F^* = \left( \frac{\hat{\beta}_i}{se(\hat{\beta}_i)} \right)^2 \]  

(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:

\[ \sqrt{(X^T X)^{-1}} \left( [y - X\hat{\beta}]^T [y - X\hat{\beta}] \right) / (n-k) \]  

(12)

with \( k \) being the number of \( \beta \) parameters in the model including the intercept \( \beta_0 \), and where \( (X^T X)^{-1} \) denotes the \( (i) \)th diagonal element of square matrix \( (X^T X)^{-1} \).

The hypothesis test for the regression tests:

\[ H_0 : \beta_1 = \ldots = \beta_k = 0 \]  

(13)

by using the equation:

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

(14)

where \( \bar{y} \) represents the \( nx1 \) constant vector where each element is the average of \( y_i, \ldots, y_n \). 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](image)

2. **R-Squared (R²)**

R-Squared, often abbreviated R², which is also called the coefficient of determination or the percentage of variance explained is equated by the following formula:

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

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:

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

$$H_0 : \beta_0$$
$$H_1 : \beta_0 + \beta_1 x_i$$
$$\vdots$$
$$H_k : \beta_0 + \ldots + \beta_k x_k$$  \hspace{1cm} (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.
\[
F = \frac{\sum_{i=1}^{n}(y_i - H_{j-1}(\hat{y}_i))^2 - \sum_{i=1}^{n}(y_i - H_j(\hat{y}_i))^2}{\sum_{i=1}^{n}(y_i - \hat{y}_j)^2 / n - k}, \quad j = 1, 2, \ldots, k. \quad (18)
\]

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:
\[ g[E(y_i)] = x_i^T \beta \]
\[ E(y_i) = g^{-1}(x_i^T \beta) \]  \hspace{1cm} (19)

where the function \( g \) is called the link function and \( x_i \) is the \((k+1) \times 1\) 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
\[ f(y) = \frac{e^{-\mu} \mu^y}{y!}, \quad y = 0, 1, \ldots \]  \hspace{1cm} (20)
where \( \mu > 0 \) and represents the mean of \( y \). The variance for the Poisson distribution is also identically \( \mu \).

For Poisson regression, the log function is used as the link function, therefore for \( i = 1, \ldots, n \) where each \( y_i \) has a Poisson distribution with mean \( \mu_i \), the Poisson regression model is expressed as:

\[ E(y_i) = \mu_i \]
\[ g(\mu_i) = x_i^T \beta \]
\[ \mu_i = g^{-1}(x_i^T \beta) \]  \hspace{1cm} (21)

Substituting the log link function gives:
\[ \ln(\mu_i) = x_i^T \beta \]
\[ \mu_i = e^{x_i^T \beta} \]  \hspace{1cm} (22)

The maximum-likelihood estimation (MLE) approach must be employed in order to estimate the \( \beta \)s for the regression. Finding the MLE for the Poisson regression starts with the expression for the likelihood of observing \( y \) as a function of \( \beta \) (where \( y_1, \ldots, y_n \) are assumed independent):
\[ L(y, \beta) = \prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!} \]
\[ = \frac{\prod_{i=1}^{n} e^{-\mu_i} \prod_{i=1}^{n} \mu_i^{y_i}}{\prod_{i=1}^{n} y_i!} \quad \text{(23)} \]

Taking the natural log of both sides gives:

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

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

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

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:

\[ D = 2 \sum_{i=1}^{n} \left[ y_i \ln \left( \frac{y_i}{\hat{y}_i} \right) - (y_i - \hat{y}_i) \right] \quad \text{(26)} \]
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:

\[ f(y) = \frac{\Gamma(r + y)}{y!\Gamma(r)} (1 - p)^r p^y \quad y = 0, 1, 2, \ldots \]  

with

\[ E(y) = \frac{pr}{1-p} \]  
\[ \text{Var}(y) = \frac{pr}{(1-p)^2} \]  

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:

\[ \Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt \]  

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:
\[ \mu = \frac{pr}{1-p} \Rightarrow p = \frac{\mu}{\mu + r} \quad (30) \]

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

\[
f(y) = \frac{\Gamma(r+y)}{y!\Gamma(r)} \left(1 - \frac{\mu}{\mu + r}\right)^r \left(\frac{\mu}{\mu + r}\right)^y = \frac{\Gamma(r+\mu)}{y!\Gamma(r)} \left(\frac{\mu + r - \mu}{\mu + r}\right)^r \left(\frac{\mu}{\mu + r}\right)^y \]
\[= \frac{\Gamma(r+y)}{y!\Gamma(r)} \left(\frac{r}{\mu + r}\right)^r \left(\frac{\mu}{\mu + r}\right)^y = \frac{\mu^y}{y!\Gamma(r)(\mu + r)^y} \frac{1}{\left(1 + \frac{\mu}{r}\right)^r} \quad (31)\]

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:

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

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:

\[ g(u_i) = x_i^T \beta \Rightarrow u_i = g^{-1}(x_i^T \beta) = e^{x_i^T \beta} \quad (33) \]

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

\[
L(y, \beta) = \prod_{i=1}^{n} \frac{\mu_i^{y_i}}{y_i!} \frac{\Gamma(r+y_i)}{\Gamma(r)(\mu_i + r)^{y_i}} \frac{1}{\left(1 + \frac{\mu_i}{r}\right)^r} \]

\[
l(y, \beta) = \ln(L(y, \beta)) = \ln \left( \prod_{i=1}^{n} \frac{\mu_i^{y_i}}{y_i!} \frac{\Gamma(r+y_i)}{\Gamma(r)(\mu_i + r)^{y_i}} \frac{1}{\left(1 + \frac{\mu_i}{r}\right)^r} \right) \]
\[
= \sum_{i=1}^{n} y_i \ln \mu_i + \ln(\Gamma(r+y_i)) + \ln(1) - \ln(y_i!) - \ln(\Gamma(r)) - \ln((\mu_i + r)^{y_i}) - r \ln(1 + \frac{\mu_i}{r}) \]
\[
= \sum_{i=1}^{n} y_i x_i + \ln(\Gamma(r+y_i)) - \ln(y_i!) - \ln(\Gamma(r)) - y_i \ln(x_i^T \beta + r) - r \ln(1 + \frac{x_i^T \beta}{r})
\]
Now, to find the MLEs $\hat{\beta}$ for $\beta$, take the derivative with respect to $\beta$:

$$\frac{\partial l(y, \beta)}{\partial \beta} = \sum_{i=1}^{n} y_i x_i + \frac{y_i x_i}{x_i' \beta + r} - \frac{x_i}{1 + \frac{x_i' \beta}{r}} = 0$$  \hspace{1cm} (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 jackknife, $m=n$. 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 $\hat{y}_{(i)}^j = 1, ..., n$ 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

$$\sqrt{\frac{\sum_{i=1}^{n} \left( y_{(i)} - \hat{y}_{(i)} \right)^2}{n}}$$  \hspace{1cm} (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.

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

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.

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

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:
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.
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.

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.

![Graphical Correlation between Homicide and Independent Variables.](image)

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.
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.

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

Table 5. Correlation of Variables against Assaults and Homicides

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

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

<table>
<thead>
<tr>
<th></th>
<th>Estimates</th>
<th>Standard Error</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Intercept</strong></td>
<td>1.83 x 10³</td>
<td>9.46 x 10²</td>
<td>0.09465</td>
</tr>
<tr>
<td><strong>SPD Budget</strong></td>
<td>9.38 x 10⁻⁶</td>
<td>5.87 x 10⁻⁶</td>
<td>0.154</td>
</tr>
<tr>
<td><strong>CDCR Over Population Percentage</strong></td>
<td>-8.35 x 10²</td>
<td>4.01 x 10²</td>
<td>0.076</td>
</tr>
<tr>
<td><strong>Unemployment</strong></td>
<td>-1.39 x 10</td>
<td>2.75</td>
<td>0.0381</td>
</tr>
<tr>
<td><strong>Number Vacant Units</strong></td>
<td>5.06 x 10</td>
<td>1.99 x 10</td>
<td>0.125</td>
</tr>
<tr>
<td><strong>SPD Badged Police</strong></td>
<td>1.71</td>
<td>9.82 x 10⁻¹</td>
<td>0.00147</td>
</tr>
</tbody>
</table>

**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:
with the following p-values displayed in Table 8.

<table>
<thead>
<tr>
<th>Estimates</th>
<th>Standard Error</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-3.34x10^{-2}</td>
<td>1.73x10^{-2}</td>
</tr>
<tr>
<td>SPD Budget</td>
<td>-3.23x10^{-6}</td>
<td>3.14x10^{-6}</td>
</tr>
<tr>
<td>CDCR Over Population Percentage</td>
<td>8.22x10^{-2}</td>
<td>1.18x10^{-2}</td>
</tr>
</tbody>
</table>

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.
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:

\[
\hat{y} = \exp(7.679 - 8.695 \times 10^{-9} \text{ (SPD Budget)} - 1.287 \times 10^{-2} \text{ (SPD Sworn Police)} - 7.743 \times 10^{-1} \text{ (CDCR Overpopulation Percentage)} + 4.666 \times 10^{-2} \text{ (Unemployment)} + 1.582 \times 10^{-3} \text{ (Number of Vacant Units)}) \tag{39}
\]

with the p-values for the regressors displayed in Table 9.
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 overfitting. 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 the 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:
\[ \hat{y} = 1.152 - 7.363 \times 10^{-5} \text{ (Population)} + 1.805 \times 10^{-1} \text{ (SPD Budget)} - 3.959 \text{ (CDCR Overpopulation Percentage)} - 1.745 \times 10^{-6} \text{ (Parks and Rec Funding)} + 8.184 \times 10^{-6} \text{ (Library Funding)} \]  

(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.

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Standard Error</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.15x10</td>
<td>9.78</td>
<td>0.2505</td>
</tr>
<tr>
<td>Population</td>
<td>-7.36x10^{-5}</td>
<td>1.57x10^{-4}</td>
<td>0.64332</td>
</tr>
<tr>
<td>SPD Budget</td>
<td>1.81x10^{-7}</td>
<td>2.46x10^{-7}</td>
<td>0.47005</td>
</tr>
<tr>
<td>CDCR Over Population Percentage</td>
<td>-3.96</td>
<td>6.85</td>
<td>0.56862</td>
</tr>
<tr>
<td>Parks and Rec Fund</td>
<td>-1.75x10^{-8}</td>
<td>2.42x10^{-8}</td>
<td>0.47756</td>
</tr>
<tr>
<td>Library Fund</td>
<td>8.18x10^{-8}</td>
<td>2.33x10^{-8}</td>
<td>0.00177</td>
</tr>
</tbody>
</table>

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:

\[ \hat{y} = -0.5852 + 6.130 \times 10^{-6} \text{ (Library Funding)} \]

(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 5.67x10^7. 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.
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:

$$\hat{y} = \exp(2.040 - 2.433 \times 10^{-6} \text{ (Population)} + 4.777 \times 10^{-1} \text{ (SPD Budget)} - 0.1496 \text{ (CDCR Overpopulation Percentage)} - 1.437 \times 10^{-7} \text{ (Parks and Rec Funding)} + 5.718 \times 10^{-7} \text{ (Library Funding)})$$

(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.
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:

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

(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.

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:

\[ \hat{y} = \exp(1.520 + 4.425 \times 10^{-7} \text{(Library Funding)}) \]  

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.

![Negative Binomial Regression for Homicide](image)

**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:

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

with p-values of:
## Table 12. P-values for OLS Regression for Assault and Homicide prediction

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Standard Error</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Intercept</strong></td>
<td>5.49x10²</td>
<td>4.31x10²</td>
<td>0.22266</td>
</tr>
<tr>
<td><strong>CDCR Over Population Percentage</strong></td>
<td>1.15x10²</td>
<td>2.98x10²</td>
<td>0.70399</td>
</tr>
<tr>
<td><strong>Parolee Population</strong></td>
<td>-2.45x10⁻³</td>
<td>2.27x10⁻³</td>
<td>0.29801</td>
</tr>
<tr>
<td><strong>Unemployment with 2 year shift</strong></td>
<td>2.79x10</td>
<td>8.85</td>
<td>0.00662</td>
</tr>
</tbody>
</table>

The R² 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:

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

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 R² of 0.7117 and a graphical interpretation shown in Figure 14.

![Linear Regression Fit for Assault and Homicides](image.png)

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:

\[
\hat{y} = \exp(6.324 - 1.690 \times 10^{-1} (\text{CDCR Overpopulation Percentage}) \\
- 3.226 \times 10^{-6} (\text{Parole Population}) \\
+ 3.622 \times 10^{-2} (\text{Unemployment with two year shift}))
\]

with p-values of:

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Standard Error</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>6.32</td>
<td>2.55x10^{-1}</td>
<td>2.00x10^{-16}</td>
</tr>
<tr>
<td>CDCR Over Population Percentage</td>
<td>1.69x10^{-1}</td>
<td>1.78x10^{-1}</td>
<td>0.3413</td>
</tr>
<tr>
<td>Parolee Population</td>
<td>-3.23x10^{-6}</td>
<td>1.36x10^{-6}</td>
<td>0.0174</td>
</tr>
<tr>
<td>Unemployment with 2 year shift</td>
<td>3.62x10^{-2}</td>
<td>5.25x10^{-3}</td>
<td>5.42x10^{-12}</td>
</tr>
</tbody>
</table>

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:

\[
\hat{y} = \exp(7.470 - 1.791 \times 10^{-6} (\text{Parole Population}) \\
- 1.730 \times 10^{-1} (\text{Unemployment with two year shift}) \\
+ 1.063 \times 10^{-2} (\text{Unemployment with two year shift}^2)
\]

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.
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:

\[
\hat{y} = \exp(7.448 - 1.771 \times 10^{-6} \text{Parole Population})
- 1.687 \times 10^{-1} \text{(Unemployment with two year shift)}
+ 1.04 \times 10^{-2} \text{(Unemployment with two year shift)^2)}
\]  

(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.

<table>
<thead>
<tr>
<th>Model</th>
<th>Salinas Violence</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS Regression Model</td>
<td>1810.954</td>
</tr>
<tr>
<td>Poisson Regression Model</td>
<td>2117.753</td>
</tr>
<tr>
<td>Negative Binomial Regression Model</td>
<td>2117.753</td>
</tr>
</tbody>
</table>

Table 14. 2011 Violence Prediction based on the Derived Models
Homicide prediction is displayed in Table 15.

<table>
<thead>
<tr>
<th>Model</th>
<th>Salinas Homicides</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS Regression Model</td>
<td>22.64106</td>
</tr>
<tr>
<td>Poisson Regression Model</td>
<td>24.43126</td>
</tr>
<tr>
<td>Negative Binomial Regression Model</td>
<td>24.45902</td>
</tr>
</tbody>
</table>

Table 15. 2011 Homicide Prediction based on the Derived Models

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

<table>
<thead>
<tr>
<th>Model</th>
<th>Salinas Assaults and Homicides</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS Regression Model</td>
<td>835.2094</td>
</tr>
<tr>
<td>Poisson Regression Model</td>
<td>836.4053</td>
</tr>
<tr>
<td>Negative Binomial Regression Model</td>
<td>836.1911</td>
</tr>
</tbody>
</table>

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

The actual counts for 2011 are detailed in Table 17.

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Violence</td>
<td>1083</td>
</tr>
<tr>
<td>Homicides</td>
<td>15</td>
</tr>
<tr>
<td>Assaults and Homicides</td>
<td>709</td>
</tr>
</tbody>
</table>

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 Ceasefire, 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.
LIST OF REFERENCES


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.

<table>
<thead>
<tr>
<th>Year</th>
<th>Population</th>
<th>Homicides</th>
<th>Robbery</th>
<th>Assault</th>
<th>Violence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1980</td>
<td>80479</td>
<td>9</td>
<td>208</td>
<td>283</td>
<td>500</td>
</tr>
<tr>
<td>1981</td>
<td>82700</td>
<td>9</td>
<td>191</td>
<td>344</td>
<td>544</td>
</tr>
<tr>
<td>1982</td>
<td>85300</td>
<td>11</td>
<td>179</td>
<td>356</td>
<td>546</td>
</tr>
<tr>
<td>1983</td>
<td>87600</td>
<td>2</td>
<td>200</td>
<td>372</td>
<td>574</td>
</tr>
<tr>
<td>1984</td>
<td>91100</td>
<td>8</td>
<td>159</td>
<td>362</td>
<td>529</td>
</tr>
<tr>
<td>1985</td>
<td>94600</td>
<td>10</td>
<td>167</td>
<td>424</td>
<td>601</td>
</tr>
<tr>
<td>1986</td>
<td>98300</td>
<td>9</td>
<td>204</td>
<td>672</td>
<td>885</td>
</tr>
<tr>
<td>1987</td>
<td>100800</td>
<td>7</td>
<td>192</td>
<td>633</td>
<td>832</td>
</tr>
<tr>
<td>1988</td>
<td>103900</td>
<td>4</td>
<td>217</td>
<td>722</td>
<td>943</td>
</tr>
<tr>
<td>1989</td>
<td>105400</td>
<td>7</td>
<td>217</td>
<td>734</td>
<td>958</td>
</tr>
<tr>
<td>1990</td>
<td>108777</td>
<td>11</td>
<td>262</td>
<td>778</td>
<td>1051</td>
</tr>
<tr>
<td>1991</td>
<td>111184</td>
<td>7</td>
<td>253</td>
<td>805</td>
<td>1065</td>
</tr>
<tr>
<td>1992</td>
<td>114736</td>
<td>17</td>
<td>388</td>
<td>722</td>
<td>1127</td>
</tr>
<tr>
<td>1993</td>
<td>116686</td>
<td>15</td>
<td>560</td>
<td>844</td>
<td>1419</td>
</tr>
<tr>
<td>1994</td>
<td>120885</td>
<td>24</td>
<td>414</td>
<td>846</td>
<td>1284</td>
</tr>
<tr>
<td>1995</td>
<td>121960</td>
<td>15</td>
<td>494</td>
<td>950</td>
<td>1459</td>
</tr>
<tr>
<td>1996</td>
<td>124972</td>
<td>8</td>
<td>412</td>
<td>884</td>
<td>1304</td>
</tr>
<tr>
<td>1997</td>
<td>127369</td>
<td>18</td>
<td>348</td>
<td>895</td>
<td>1261</td>
</tr>
<tr>
<td>1998</td>
<td>132449</td>
<td>17</td>
<td>440</td>
<td>661</td>
<td>1118</td>
</tr>
<tr>
<td>1999</td>
<td>136797</td>
<td>13</td>
<td>346</td>
<td>737</td>
<td>1096</td>
</tr>
<tr>
<td>2000</td>
<td>142685</td>
<td>18</td>
<td>443</td>
<td>734</td>
<td>1195</td>
</tr>
<tr>
<td>2001</td>
<td>144728</td>
<td>15</td>
<td>399</td>
<td>799</td>
<td>1213</td>
</tr>
<tr>
<td>2002</td>
<td>146659</td>
<td>20</td>
<td>367</td>
<td>692</td>
<td>1079</td>
</tr>
<tr>
<td>2003</td>
<td>148117</td>
<td>19</td>
<td>399</td>
<td>725</td>
<td>1143</td>
</tr>
<tr>
<td>2004</td>
<td>149838</td>
<td>17</td>
<td>452</td>
<td>678</td>
<td>1147</td>
</tr>
<tr>
<td>2005</td>
<td>149626</td>
<td>7</td>
<td>335</td>
<td>645</td>
<td>987</td>
</tr>
<tr>
<td>2006</td>
<td>148707</td>
<td>7</td>
<td>383</td>
<td>683</td>
<td>1073</td>
</tr>
<tr>
<td>2007</td>
<td>148782</td>
<td>14</td>
<td>378</td>
<td>711</td>
<td>1103</td>
</tr>
<tr>
<td>2008</td>
<td>150898</td>
<td>25</td>
<td>334</td>
<td>633</td>
<td>992</td>
</tr>
<tr>
<td>2009</td>
<td>150215</td>
<td>29</td>
<td>359</td>
<td>678</td>
<td>1066</td>
</tr>
<tr>
<td>Year</td>
<td>Drop Outs</td>
<td>Drop Out Rate</td>
<td>SPD Budget</td>
<td>SPD Employees</td>
<td>Sworn Police</td>
</tr>
<tr>
<td>------</td>
<td>-----------</td>
<td>---------------</td>
<td>------------</td>
<td>---------------</td>
<td>--------------</td>
</tr>
<tr>
<td>1980</td>
<td>NA</td>
<td>NA</td>
<td>5342123</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>1981</td>
<td>NA</td>
<td>NA</td>
<td>5726778</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>1982</td>
<td>NA</td>
<td>NA</td>
<td>6170072</td>
<td>173</td>
<td>NA</td>
</tr>
<tr>
<td>1983</td>
<td>NA</td>
<td>NA</td>
<td>6645438</td>
<td>177</td>
<td>NA</td>
</tr>
<tr>
<td>1984</td>
<td>NA</td>
<td>NA</td>
<td>7715577</td>
<td>180.5</td>
<td>NA</td>
</tr>
<tr>
<td>1985</td>
<td>NA</td>
<td>NA</td>
<td>8331832</td>
<td>180.5</td>
<td>NA</td>
</tr>
<tr>
<td>1986</td>
<td>NA</td>
<td>NA</td>
<td>9315054</td>
<td>184</td>
<td>NA</td>
</tr>
<tr>
<td>1987</td>
<td>NA</td>
<td>NA</td>
<td>9607781</td>
<td>180</td>
<td>NA</td>
</tr>
<tr>
<td>1988</td>
<td>NA</td>
<td>NA</td>
<td>10177311</td>
<td>184</td>
<td>NA</td>
</tr>
<tr>
<td>1989</td>
<td>NA</td>
<td>NA</td>
<td>10649699</td>
<td>184</td>
<td>NA</td>
</tr>
<tr>
<td>1990</td>
<td>NA</td>
<td>NA</td>
<td>11456256</td>
<td>186</td>
<td>NA</td>
</tr>
<tr>
<td>1991</td>
<td>NA</td>
<td>NA</td>
<td>13144292</td>
<td>186</td>
<td>NA</td>
</tr>
<tr>
<td>1992</td>
<td>157</td>
<td>2.6</td>
<td>13634881</td>
<td>187</td>
<td>NA</td>
</tr>
<tr>
<td>1993</td>
<td>237</td>
<td>3.7</td>
<td>15683718</td>
<td>181</td>
<td>NA</td>
</tr>
<tr>
<td>1994</td>
<td>157</td>
<td>2.4</td>
<td>13612478</td>
<td>179</td>
<td>NA</td>
</tr>
<tr>
<td>1995</td>
<td>198</td>
<td>2.9</td>
<td>14471238</td>
<td>188</td>
<td>NA</td>
</tr>
<tr>
<td>1996</td>
<td>330</td>
<td>4.7</td>
<td>16393545</td>
<td>193</td>
<td>NA</td>
</tr>
<tr>
<td>1997</td>
<td>281</td>
<td>3.8</td>
<td>16929407</td>
<td>198</td>
<td>147</td>
</tr>
<tr>
<td>1998</td>
<td>342</td>
<td>4.5</td>
<td>17575700</td>
<td>198</td>
<td>144</td>
</tr>
<tr>
<td>1999</td>
<td>256</td>
<td>3.2</td>
<td>18852899</td>
<td>199</td>
<td>143</td>
</tr>
<tr>
<td>2000</td>
<td>203</td>
<td>2.5</td>
<td>19288170</td>
<td>213</td>
<td>145</td>
</tr>
<tr>
<td>2001</td>
<td>225</td>
<td>2.6</td>
<td>21713995</td>
<td>221</td>
<td>149</td>
</tr>
<tr>
<td>2002</td>
<td>204</td>
<td>2.3</td>
<td>22040439</td>
<td>222</td>
<td>145</td>
</tr>
<tr>
<td>2003</td>
<td>75</td>
<td>0.08</td>
<td>24224300</td>
<td>224</td>
<td>154</td>
</tr>
<tr>
<td>2004</td>
<td>124</td>
<td>1.3</td>
<td>25241659</td>
<td>222</td>
<td>167</td>
</tr>
<tr>
<td>2005</td>
<td>85</td>
<td>0.9</td>
<td>29704910</td>
<td>232</td>
<td>164</td>
</tr>
<tr>
<td>2006</td>
<td>124</td>
<td>1.3</td>
<td>33356709</td>
<td>238</td>
<td>161</td>
</tr>
<tr>
<td>2007</td>
<td>180</td>
<td>2.6</td>
<td>35416564</td>
<td>255</td>
<td>174</td>
</tr>
<tr>
<td>2008</td>
<td>147</td>
<td>1.5</td>
<td>38380314</td>
<td>251</td>
<td>177</td>
</tr>
<tr>
<td>2009</td>
<td>264</td>
<td>2.8</td>
<td>41187794</td>
<td>251</td>
<td>164</td>
</tr>
<tr>
<td>2010</td>
<td>276</td>
<td>2.9</td>
<td>37360500</td>
<td>230</td>
<td>157</td>
</tr>
<tr>
<td>2011</td>
<td>276</td>
<td>2.9</td>
<td>39852481</td>
<td>210</td>
<td>157</td>
</tr>
<tr>
<td>Year</td>
<td>CDCR Capacity</td>
<td>CDCR Population</td>
<td>CDCR Overpopulation Percentage</td>
<td>Parole Population</td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>---------------</td>
<td>-----------------</td>
<td>-------------------------------</td>
<td>------------------</td>
<td></td>
</tr>
<tr>
<td>1980</td>
<td>23534</td>
<td>23371</td>
<td>0.99307385</td>
<td>14650</td>
<td></td>
</tr>
<tr>
<td>1981</td>
<td>23800</td>
<td>26372</td>
<td>0.99307385</td>
<td>13952</td>
<td></td>
</tr>
<tr>
<td>1982</td>
<td>24611</td>
<td>31319</td>
<td>1.07155337</td>
<td>16072</td>
<td></td>
</tr>
<tr>
<td>1983</td>
<td>25703</td>
<td>35965</td>
<td>1.2184959</td>
<td>22202</td>
<td></td>
</tr>
<tr>
<td>1984</td>
<td>26792</td>
<td>40524</td>
<td>1.34237832</td>
<td>27000</td>
<td></td>
</tr>
<tr>
<td>1985</td>
<td>29042</td>
<td>45528</td>
<td>1.39535845</td>
<td>30726</td>
<td></td>
</tr>
<tr>
<td>1986</td>
<td>32097</td>
<td>53620</td>
<td>1.67056111</td>
<td>34771</td>
<td></td>
</tr>
<tr>
<td>1987</td>
<td>36465</td>
<td>62949</td>
<td>1.72628548</td>
<td>43355</td>
<td></td>
</tr>
<tr>
<td>1988</td>
<td>44124</td>
<td>69695</td>
<td>1.57952588</td>
<td>52788</td>
<td></td>
</tr>
<tr>
<td>1989</td>
<td>47120</td>
<td>79849</td>
<td>1.69458829</td>
<td>61665</td>
<td></td>
</tr>
<tr>
<td>1990</td>
<td>51013</td>
<td>90405</td>
<td>1.77219532</td>
<td>73096</td>
<td></td>
</tr>
<tr>
<td>1991</td>
<td>54042</td>
<td>95930</td>
<td>1.77510085</td>
<td>85470</td>
<td></td>
</tr>
<tr>
<td>1992</td>
<td>57986</td>
<td>98386</td>
<td>1.6967199</td>
<td>89453</td>
<td></td>
</tr>
<tr>
<td>1993</td>
<td>61983</td>
<td>109654</td>
<td>1.76909798</td>
<td>88858</td>
<td></td>
</tr>
<tr>
<td>1994</td>
<td>66183</td>
<td>118968</td>
<td>1.79756131</td>
<td>92958</td>
<td></td>
</tr>
<tr>
<td>1995</td>
<td>70717</td>
<td>125585</td>
<td>1.77588133</td>
<td>96110</td>
<td></td>
</tr>
<tr>
<td>1996</td>
<td>73121</td>
<td>135294</td>
<td>1.85027557</td>
<td>100934</td>
<td></td>
</tr>
<tr>
<td>1997</td>
<td>75952</td>
<td>146656</td>
<td>1.93090373</td>
<td>105449</td>
<td></td>
</tr>
<tr>
<td>1998</td>
<td>79877</td>
<td>150731</td>
<td>1.88703882</td>
<td>111875</td>
<td></td>
</tr>
<tr>
<td>1999</td>
<td>79873</td>
<td>154440</td>
<td>1.93356954</td>
<td>117612</td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>80272</td>
<td>154014</td>
<td>1.91865158</td>
<td>121414</td>
<td></td>
</tr>
<tr>
<td>2001</td>
<td>80467</td>
<td>153649</td>
<td>1.90946599</td>
<td>121820</td>
<td></td>
</tr>
<tr>
<td>2002</td>
<td>79957</td>
<td>151579</td>
<td>1.89575647</td>
<td>117138</td>
<td></td>
</tr>
<tr>
<td>2003</td>
<td>80187</td>
<td>153783</td>
<td>1.91780463</td>
<td>114136</td>
<td></td>
</tr>
<tr>
<td>2004</td>
<td>80980</td>
<td>157895</td>
<td>1.94980242</td>
<td>113768</td>
<td></td>
</tr>
<tr>
<td>2005</td>
<td>81008</td>
<td>158837</td>
<td>1.96075696</td>
<td>115001</td>
<td></td>
</tr>
<tr>
<td>2006</td>
<td>87370</td>
<td>166547</td>
<td>1.90622639</td>
<td>121808</td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>84653</td>
<td>166277</td>
<td>1.96421863</td>
<td>126906</td>
<td></td>
</tr>
<tr>
<td>2008</td>
<td>84066</td>
<td>160169</td>
<td>1.90527681</td>
<td>123597</td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>84241</td>
<td>154749</td>
<td>1.83697962</td>
<td>109026</td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td>84596</td>
<td>168830</td>
<td>2.00413101</td>
<td>108656</td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>84130</td>
<td>136619</td>
<td>1.623903483</td>
<td>100490</td>
<td></td>
</tr>
<tr>
<td>Year</td>
<td>Parks and Rec Fund</td>
<td>Library Fund</td>
<td>Unemployment</td>
<td>Number of Vacant Units</td>
<td>Person Per Household</td>
</tr>
<tr>
<td>------</td>
<td>-------------------</td>
<td>--------------</td>
<td>--------------</td>
<td>------------------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>1980</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>1981</td>
<td>1443405</td>
<td>1032802</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>1982</td>
<td>1533507</td>
<td>1077642</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>1983</td>
<td>1704186</td>
<td>1203748</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>1984</td>
<td>1734921</td>
<td>1278394</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>1985</td>
<td>1956253</td>
<td>1410963</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>1986</td>
<td>2257021</td>
<td>1584859</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>1987</td>
<td>2223182</td>
<td>1518322</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>1988</td>
<td>2674380</td>
<td>1748197</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>1989</td>
<td>2768221</td>
<td>1825048</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>1990</td>
<td>2178310</td>
<td>1955405</td>
<td>9.7</td>
<td>1228</td>
<td>3.21</td>
</tr>
<tr>
<td>1991</td>
<td>2587574</td>
<td>2059120</td>
<td>11.4</td>
<td>1227</td>
<td>3.24729</td>
</tr>
<tr>
<td>1992</td>
<td>2785467</td>
<td>2180700</td>
<td>12.5</td>
<td>1230</td>
<td>3.33079</td>
</tr>
<tr>
<td>1993</td>
<td>1992787</td>
<td>2121146</td>
<td>13.1</td>
<td>1234</td>
<td>3.36624</td>
</tr>
<tr>
<td>1994</td>
<td>1966659</td>
<td>2119116</td>
<td>12.4</td>
<td>1240</td>
<td>3.46189</td>
</tr>
<tr>
<td>1995</td>
<td>1542992</td>
<td>2191539</td>
<td>12.3</td>
<td>1251</td>
<td>3.45003</td>
</tr>
<tr>
<td>1996</td>
<td>1850056</td>
<td>2263639</td>
<td>11.3</td>
<td>1269</td>
<td>3.47405</td>
</tr>
<tr>
<td>1997</td>
<td>1871278</td>
<td>2489376</td>
<td>11</td>
<td>1281</td>
<td>3.49622</td>
</tr>
<tr>
<td>1998</td>
<td>2048408</td>
<td>2589735</td>
<td>10.8</td>
<td>1305</td>
<td>3.56309</td>
</tr>
<tr>
<td>1999</td>
<td>2131408</td>
<td>2799503</td>
<td>9.7</td>
<td>1314</td>
<td>3.60341</td>
</tr>
<tr>
<td>2000</td>
<td>2296515</td>
<td>2868795</td>
<td>7.4</td>
<td>1360</td>
<td>3.662</td>
</tr>
<tr>
<td>2001</td>
<td>2817339</td>
<td>3020075</td>
<td>7.8</td>
<td>1370</td>
<td>3.69</td>
</tr>
<tr>
<td>2002</td>
<td>2974138</td>
<td>3316832</td>
<td>8.9</td>
<td>1385</td>
<td>3.702</td>
</tr>
<tr>
<td>2003</td>
<td>3081427</td>
<td>3423623</td>
<td>9</td>
<td>1400</td>
<td>3.7</td>
</tr>
<tr>
<td>2004</td>
<td>2795909</td>
<td>3170427</td>
<td>8.3</td>
<td>1418</td>
<td>3.699</td>
</tr>
<tr>
<td>2005</td>
<td>2285817</td>
<td>2614595</td>
<td>7.3</td>
<td>1433</td>
<td>3.654</td>
</tr>
<tr>
<td>2006</td>
<td>2082617</td>
<td>1278414</td>
<td>6.9</td>
<td>1441</td>
<td>3.614</td>
</tr>
<tr>
<td>2007</td>
<td>3153973</td>
<td>2207708</td>
<td>7.1</td>
<td>1450</td>
<td>3.601</td>
</tr>
<tr>
<td>2008</td>
<td>3893586</td>
<td>3440113</td>
<td>8.4</td>
<td>1452</td>
<td>3.637</td>
</tr>
<tr>
<td>2009</td>
<td>3976221</td>
<td>4061128</td>
<td>11.8</td>
<td>1463</td>
<td>3.643</td>
</tr>
<tr>
<td>2010</td>
<td>3302147</td>
<td>3587431</td>
<td>12.8</td>
<td>1462</td>
<td>3.685</td>
</tr>
<tr>
<td>2011</td>
<td>1571796</td>
<td>3788695</td>
<td>12.4</td>
<td>1462</td>
<td>4</td>
</tr>
</tbody>
</table>


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

Salinas population 2000–2010

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

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

Person per household 2001–2010

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

http://www.cdcr.ca.gov/Reports/CDCR-Annual-Reports.html

Police, Library, and Parks and Recreation 1981–2004:
Clark and Onufer thesis
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%20%20%20%20%20%20%20%20%20%20&cCountyCode=27&cDistrictCode=2766159&cSchoolCode=000000&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

<table>
<thead>
<tr>
<th>Model</th>
<th>Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS Model</td>
<td>$\hat{y}=1.827 \times 10^3 + 9.384 \times 10^{-6} \text{ (SPD Budget)}$</td>
</tr>
<tr>
<td></td>
<td>-8.354 \times 10^2 \text{ (CDCR Overpopulation Percentage)}</td>
</tr>
<tr>
<td></td>
<td>+50.62 \text{ (Unemployment with Two Year Shift)}</td>
</tr>
<tr>
<td></td>
<td>+1.712 \text{ (Number of Vacant Units with Two Year Shift)}</td>
</tr>
<tr>
<td></td>
<td>-13.93 \text{ (SPD Sworn Police with One Year Shift)}</td>
</tr>
<tr>
<td>Poisson Model</td>
<td>$\hat{y}=\exp(7.679 - 8.695 \times 10^{-9} \text{ (SPD Budget)}$</td>
</tr>
<tr>
<td></td>
<td>-1.287 \times 10^{-2} \text{ (SPD Sworn Police with one year shift)}</td>
</tr>
<tr>
<td></td>
<td>-7.743 \times 10^{-1} \text{ (CDCR Overpopulation Percentage)}</td>
</tr>
<tr>
<td></td>
<td>+4.666 \times 10^{-2} \text{ (Unemployment with two year shift)}</td>
</tr>
<tr>
<td></td>
<td>+1.582 \times 10^{-3} \text{ (Number of Vacant Units with two year shift)}</td>
</tr>
<tr>
<td>Negative Binomial Model</td>
<td>Same as Poisson Model</td>
</tr>
</tbody>
</table>

B. HOMICIDE MODELS

<table>
<thead>
<tr>
<th>Model</th>
<th>Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS Model</td>
<td>$\hat{y}=-0.5852+6.130 \times 10^{-6} \text{ (Library Funding)}$</td>
</tr>
<tr>
<td>Poisson Model</td>
<td>$\hat{y}=\exp(1.522+4.417 \times 10^{-7} \text{ (Library Funding)})$</td>
</tr>
<tr>
<td>Negative Binomial Model</td>
<td>$\hat{y}=\exp(1.520+4.425 \times 10^{-7} \text{ (Library Funding)})$</td>
</tr>
</tbody>
</table>
C. HOMICIDE AND ASSAULT MODELS

<table>
<thead>
<tr>
<th>Model</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS Model</td>
<td>..</td>
</tr>
<tr>
<td>Poisson Model</td>
<td>( \hat{y} = \exp(7.470 - 1.791 \times 10^{-6}\text{ (Parole Population)} - 1.730 \times 10^{-1}\text{ (Unemployment with two year shift)} + 1.063 \times 10^{-2}\text{ (Unemployment with two year shift)}^2) )</td>
</tr>
<tr>
<td>Negative Binomial Model</td>
<td>( \hat{y} = \exp(7.448 - 1.771 \times 10^{-6}\text{ (Parole Population)} - 1.687 \times 10^{-1}\text{ (Unemployment with two year shift)} + 1.04 \times 10^{-2}\text{ (Unemployment with two year shift)}^2) )</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>2011</th>
<th>OLS</th>
<th>Poisson</th>
<th>Negative Binomial</th>
<th>Recorded Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Violence</td>
<td>1789.337</td>
<td>2077.991</td>
<td>2077.991</td>
<td>1083</td>
<td></td>
</tr>
<tr>
<td>Homicides</td>
<td>22.64</td>
<td>24.43</td>
<td>24.46</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>Homicides and Assaults</td>
<td>881.40</td>
<td>879.80</td>
<td>879.02</td>
<td>664</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>2012</th>
<th>OLS</th>
<th>Poisson</th>
<th>Negative Binomial</th>
<th>Recorded Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Violence</td>
<td>1093.26</td>
<td>1094.10</td>
<td>1094.10</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>Homicides</td>
<td>21.64318</td>
<td>22.92571</td>
<td>23.03067</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>Homicides and Assaults</td>
<td>901.64993</td>
<td>916.76166</td>
<td>914.25029</td>
<td>NA</td>
<td></td>
</tr>
</tbody>
</table>
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","Unemployment","Vacant")

########Correlation for Homicides

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

```r
homdata = data.frame(s$Homicide, s$Population, s$SPDBudget, s$CDCRPercentage,
                     s$ParksandRec, s$Library)
```

```r
names(homdata) = c("Homicide", "Population", "SPDBudget", "CDCRPercentage", 
"ParksandRec", "Library")
```

```r
homcorbetween = cor(homdata, use = "pairwise.complete.obs")
```

#### Correlation for Assaults and Homicide

```r
HA = s$Homicide + s$Assault
corallHA = setcor(HA, s, iter = 2)
```

```r
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 the regression

```r
HA = s$Homicide + s$Assault
```

```r
HAdata = data.frame(HA[3:31], s$CDCRPercentage[3:31], s$ParolePop[3:31], s$Unemployment[1:29],
                     s$Vacant[3:31], s$PersonPerHouse[1:29])
```

```r
HAcorbetween = cor(HAdata, use = "pairwise.complete.obs")
```

##### Since we are not going to use Vacant Units or Person Per Household, let's put the first 2 obs back into the data set

```r
HAdata = data.frame(HA[3:31], s$CDCRPercentage[3:31], s$ParolePop[3:31], s$Unemployment[1:29])
```

```r
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#
HApoisss=glm(HA~.,data=HAdata,family=poisson)

First one is no good

Here is the final model

HApoisss=glm(HA~ParolePop+Unemployment+I(Unemployment^2),data=HAdata,
family=poisson)
crossglm(HApoisss)
pHApoisss=as.matrix(predict(HApoisss,type='response'))

####Violence NB Regression
HAnb=glm.nb(HA~ParolePop+Unemployment+I(Unemployment^2),data=HAdata,
link=log,start=HApoisss$coefficients)
pHAnb=as.matrix(predict(HAnb,type="response"))

This is the future predictions
s2011=read.table("clipboard",header=T)
plmvio2011=predict(viom,newdata=data.frame(SPDBudget=s2011$SPDBudget,
CDCRPercentage=s2011$CDCRPercentage,
Police=s2011$Police,Unemployment=s2011$Unemployment,Vacant=s2011$Vacant,
PersonPerHouse=s2011$PersonPerHouse))
pviolm2011=predict(viom,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$coefficients)
homlm=lm(Homicide~Library,data=homdata)
hompoiss=glm(Homicide~Library,data=homdata,family=poisson)
crossglm(hompoiss)
homnb=glm.nb(Homicide~Library,data=homdata,link=log,start=hompoiss$coefficients)
HAlm=lm(HA~Unemployment+I(Unemployment^2),data=HADATA)
HApoiiss=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=HApoiiss$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)
pvioipoiss2012=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')
PHAIm2012 = predict(HAIm,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,
PHAIm2012,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