STA 141A Final Report

Albert Yang - alyang@ucdavis.edu Samarth Sridhara - sssridhara@ucdavis.edu
Josie Dang - ppdang@ucdavis.edu Chase Varga - cvarga@ucdavis.edu
Ethan Pham - etqpham@ucdavis.edu Jiayang Liu - ajyliu@ucdavis.edu

EXECUTIVE SUMMARY:

In this project, we investigate how key economic factors, including unemployment, educational attainment, and income, predict violent crime rates across U.S. states. While conventional economic theory suggests a direct link between financial hardship and criminal behavior, we test whether this relationship remains robust across recent U.S. data, using state-level statistics from 2015 to 2019.

To answer this question, we merge annual American Community Survey data with FBI Uniform Crime Reports, creating a region-level panel that includes unemployment rates, median household income, poverty rates, educational attainment, and median rent. Violent-crime outcomes (robbery, rape, murder, larceny, burglary and assault) are expressed as incidents per one hundred thousand residents. We fit multiple linear-regression models with region fixed effects, add interaction terms to capture heterogeneous responses, and apply cross-validated ridge and lasso regression to address multicollinearity and guard against overfitting.

Nationally, unemployment retains the expected positive association with violent crime: each one-percentage-point rise in joblessness is linked to roughly fifty-five additional violent incidents per one hundred thousand people, a statistically significant effect. Disaggregating by region and controlling for income and education, however, fragments the picture. Several low-GDP areas in the South and Midwest display a flat or even negative unemployment–crime slope, implying that social infrastructure, schooling quality, or policing practices can offset economic strain. Education emerges as a consistent protective factor, with a greater degree of attainment linked to markedly lower violent-crime rates, whereas median rent shows only a weak correlation, and poverty loses significance after income and education are included.

These findings caution against blanket assumptions that job programs automatically reduce crime. In regions where unemployment is no longer a reliable predictor of violence, investing in education, community resources, and targeted policing may provide more effective crime-reduction dividends than labor-market interventions alone. Policymakers should design strategies tailored to each region's full socioeconomic context rather than presuming a universal unemployment-to-crime pathway.

PROJECT BACKGROUND & GOALS:

I. Project Overview

Understanding the relationship between economic conditions and violent crime is essential for designing effective and equitable public policies. Traditional economic theories often assert that higher unemployment leads to an increase in violent crime, especially in low-income areas, due to heightened financial stress and limited access to legitimate opportunities. While this framework provides a useful starting point, it may oversimplify a more complex and regionally variable reality.

Our project challenges the assumption that unemployment alone drives violent crime by investigating whether this relationship holds uniformly across U.S. regions with differing levels of economic development. In particular, we focus on low-GDP areas where the link between unemployment and crime might be weaker, obscured, or shaped by other local conditions such as education, housing costs, or public infrastructure. By

highlighting both patterns and exceptions, we aim to contribute to a more nuanced understanding of how economic stress interacts with crime at the regional level

II. Objective

The central goal of this project is to examine the correlation and explore the potential causal relationship between unemployment rates and violent crime across different U.S. regions. We seek to determine whether high unemployment consistently leads to higher violent-crime rates or whether this relationship varies based on broader socioeconomic conditions. Specifically, we test whether low-GDP areas with high unemployment necessarily face more crime, or whether other protective factors may buffer that effect.

Using publicly available datasets and reproducible statistical methods, this project applies tools in data wrangling, visualization, and linear modeling to a real-world economic and social issue. We hope our findings can help inform region-specific policy design while also demonstrating our technical and analytical skills.

DATASET DESCRIPTION:

- ACS (2015-2019): median household income, individuals below poverty, number employed, number unemployed, Bachelor's degree, median gross rent, median selected monthly owner costs
- FBI UCR: robbery, rape, murder, larceny, burglary, assault

For pre-processing, we began by filtering the dataset to consist of only the 50 states (the original dataset includes other things). We decided to keep the District of Columbia (DC) as it is the capital, but we removed Puerto Rico due to it not being included in the crime data set that we used. We also removed a "United States" row, as that was not relevant to our main objective of the project. After scanning through our data, we did not find any observations with missing data. We proceeded by computing crime rates per 100,000 residents, and finally, we merged the datasets and checked to see if there were any errors in our new dataset.

```
library(tidycensus)
library(tidyverse)
library(crimedata)

##### Econ Data #####
vars <- c(
  income = "B19013_001",         # Median household income
  poverty = "B17001_002",        # Individuals below poverty
  employed = "B23025_004",       # Number employed
  unemployed = "B23025_005",     # Number unemployed
  education = "B15003_022",      # Bachelor's degree
  mortgage = "B25087_001"
)

econData <- get_acs(
geography = "state",
variables = vars,
year = 2019,
survey = "acs5"
)

econDataClean <- econData %>%
  select(NAME, variable, estimate) %>%
  pivot_wider(names_from = variable, values_from = estimate) %>%
  rename(state = NAME)
```

```r
econDataScore <- econDataClean %>%
  mutate(totalLaborForce = employed + unemployed,
         unemploymentRate = unemployed / totalLaborForce,
         povertyPerCap = poverty / 1e5,
         educationRate = education / 1e5) %>%
  drop_na()

# Proxy score (for regression target)
econDataScore <- econDataScore %>%
  mutate(proxyCrimeRiskScore = scale(unemploymentRate) +
         scale(povertyPerCap) +
         #scale(rentToIncome) -
         scale(educationRate) -
         scale(income)
)

# Clean `state` in econ data (removes unnecessary whitespace)
econDataScore <- econDataScore %>%
  mutate(state = str_trim(state))

##### Crime Data #####
crime_data <- read_csv("state_crime.csv")

violent_crime_data <- crime_data %>%
  select(State, Year,
         Data.Rates.Violent.Robbery,
         Data.Rates.Violent.Rape,
         Data.Rates.Violent.Murder,
         Data.Rates.Property.Larceny,
         Data.Rates.Property.Burglary,
         Data.Rates.Violent.Assault) %>%
  filter(Year == 2019) %>%
  group_by(State) %>%
  summarize(across(Data.Rates.Violent.Robbery:Data.Rates.Violent.Assault, mean, na.rm = TRUE)) %>%
  rename(state = State) %>%
  mutate(state = str_trim(state))

# Join and check for mismatches
states_not_matched <- setdiff(econDataScore$state, violent_crime_data$state)
if (length(states_not_matched) > 0) {
  cat("States in economic data not matched in crime data:\n")
  print(states_not_matched) # Puerto Rico will not be included in final data set
}
```

```
## States in economic data not matched in crime data:
## [1] "Puerto Rico"
```

```r
# Merge
fullData <- econDataScore %>%
  left_join(violent_crime_data, by = "state") %>%
  drop_na()
```

```r
# Final data set
head(fullData, 5)
```

```
## # A tibble: 5 x 18
##   state    education poverty income employed unemployed mortgage totalLaborForce
##   <chr>        <dbl>   <dbl>  <dbl>    <dbl>      <dbl>    <dbl>           <dbl>
## 1 Alabama     529178  795989  50536  2097384     132095  1284748         2229479
## 2 Alaska       88058   76933  77640   347774      26808   162996          374582
## 3 Arizona     869452 1043764  58945  3130658     195905  1656756         3326563
## 4 Arkansas    297250  496260  47597  1303490      70481   759455         1373971
## 5 Califor~   5603047 5149742  75235 18591241    1199233  7154580        19790474
## # i 10 more variables: unemploymentRate <dbl>, povertyPerCap <dbl>,
## #   educationRate <dbl>, proxyCrimeRiskScore <dbl[,1]>,
## #   Data.Rates.Violent.Robbery <dbl>, Data.Rates.Violent.Rape <dbl>,
## #   Data.Rates.Violent.Murder <dbl>, Data.Rates.Property.Larceny <dbl>,
## #   Data.Rates.Property.Burglary <dbl>, Data.Rates.Violent.Assault <dbl>
```

Figure 1: Scatter Plots

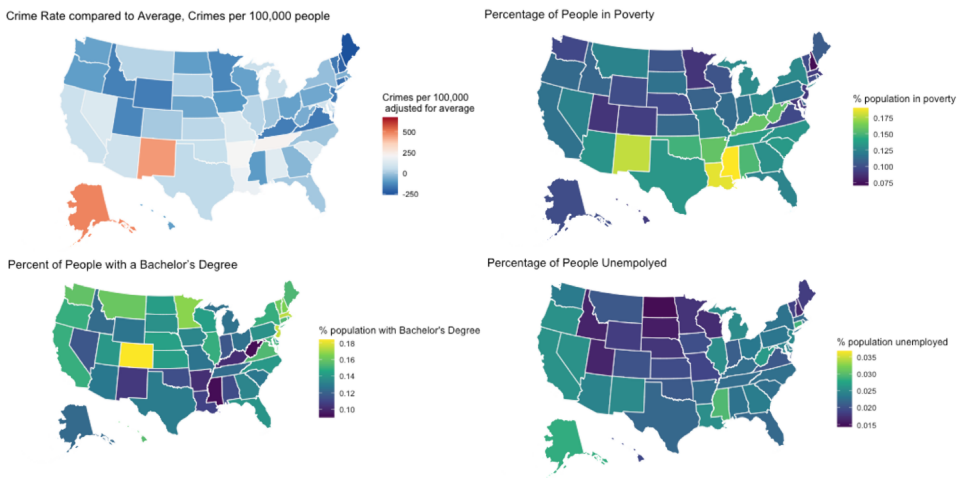
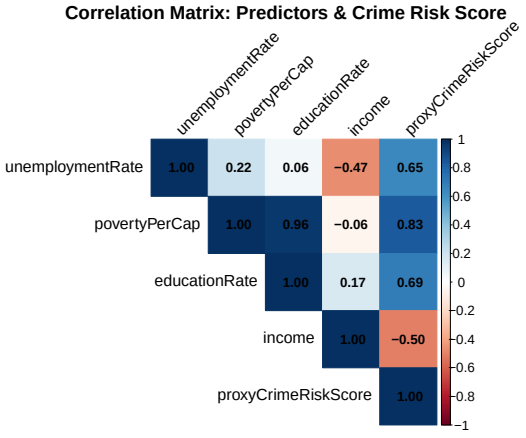


Figure 2: Heat Maps

```r
library(corrplot)

predictorsAndScore <- econDataScore %>%
  select(unemploymentRate, povertyPerCap, educationRate, income, proxyCrimeRiskScore)


corMatrix <- cor(predictorsAndScore)

corrplot(corMatrix, method = "color", type = "upper",
         tl.col = "black", tl.srt = 45,
         addCoef.col = "black", number.cex = 0.8,
         title = "Correlation Matrix: Predictors & Crime Risk Score",
         mar = c(0,0,1,0))
```



METHODOLOGIES:

I. Data Sources and Variables

This study uses a cross-sectional dataset of U.S. states from 2015 to 2019 to examine the relationship between violent crime and economic indicators. Our dependent variable is the violent crime rate, measured as the

number of violent incidents per 100,000 residents, including assault, robbery, rape, and murder. The violent crime data is based on the FBI Uniform Crime Reporting (UCR) Program, which compiles official statistics from law enforcement agencies across the country. The dataset was provided in a cleaned format through course materials and processed using tidyverse tools in R. It includes annual state-level crime data from 2012 to 2023, though we focused on a 5-year window for analysis.

The independent variables were drawn from the American Community Survey (ACS) and accessed through the tidycensus R package, which connects to the U.S. Census Bureau's API. We collected the unemployment rate, median household income, poverty rate, and the percentage of the adult population with a bachelor's degree or higher. The data was filtered to include only the 50 U.S. states and Washington, D.C., while excluding Puerto Rico and national aggregates due to their outlier behavior and lack of comparability.

II. Model Specification and Feature Selection

We primarily used supervised learning techniques, beginning with a multiple linear regression model that included all four economic predictors. We also explored ridge and lasso regression as regularization techniques for comparison. To diagnose multicollinearity in the full model, we calculated variance inflation factors (VIFs). Results showed values exceeding 45 for poverty and education, indicating a strong correlation with income and leading to distorted coefficient estimates.

To reduce this collinearity, we initially attempted to include an interaction term between poverty and education. This brought down the VIFs slightly, but they still remained high (between 10 and 25). Based on Akaike Information Criterion (AIC) and theoretical reasoning, we ultimately removed poverty from the model, which lowered all VIFs below 1.2 and improved both model clarity and statistical validity.

III. Outlier Detection and Diagnostic Evaluation

We performed outlier detection using Cook's Distance, trimming five high-leverage observations over one round with a cutoff value of 0.1. These included entries like Puerto Rico, Washington D.C., and the "United States" aggregate row. After trimming, we verified that the core assumptions of linear regression were satisfied. The residual vs. fitted plot showed no patterns, suggesting correct linear specification. The scale-location plot supported the assumption of homoscedasticity, and the Q-Q plot indicated that residuals closely followed a normal distribution.

In addition, we chose to apply AIC over BIC (Bayesian Information Criterion) for model comparison. This decision was based on our goal of minimizing prediction error and improving model generalizability, which was particularly relevant given our relatively small sample size. Since BIC is stricter and aims to find the true model, we favored AIC for its ability to select models that perform better in prediction tasks, which aligns more closely with our goals.

IV. Final Model and Robustness Checks

The final model included unemployment rate, education level, and median income. All three predictors were highly statistically significant, with p-values less than 2×10^{-16}. To test robustness, we scaled income to 50% of its original range. This had no meaningful effect on the estimated coefficients of unemployment or education, and the income coefficient remained stable. We also applied ridge and lasso regression, but neither model produced a significant improvement over the OLS model. Lasso selected the same three variables, and ridge slightly increased error. This further confirmed the efficiency and parsimony of the reduced OLS model.

V. Validation via Cross-Validation Techniques

To assess model generalizability, we performed 5-fold cross-validation, which produced a mean squared error (MSE) of 0.0118 and an R-squared value of 0.9961. This outperformed more complex models that

included interaction terms or all four predictors. Additionally, we tested the model with leave-one-out cross-validation (LOOCV), which gave a similar MSE but an undefined R-squared due to the extremely low residual variance—essentially confirming model stability but showing the limitations of LOOCV with near-perfect model fit.

VI. Summary

By combining API-based data access, thoughtful feature selection, strong model diagnostics, and robust validation, this methodology offers a rigorous and transparent approach to exploring the relationship between violent crime and economic structure. The final model performs well both statistically and substantively, providing insights into how unemployment, education, and income shape crime outcomes across states.

ANALYSIS & FINDINGS:

I. Initial Full Model and Multicollinearity Concerns

Our analysis explores the relationship between violent crime and key economic indicators at the state level using a multiple linear regression framework. We began with a full model that included unemployment rate, poverty rate, educational attainment (percent of adults with a bachelor's degree or higher), and median household income. The initial regression output indicated that all four variables were statistically significant predictors of violent crime rates. However, a closer inspection revealed severe multicollinearity, particularly between poverty, education, and income. It inflated standard errors and raised concerns about the reliability of the coefficient estimates.

To diagnose this issue, we computed Variance Inflation Factors (VIFs), which essentially showed us how much a variable's standard error became inflated due to multicollinearity with other predictor variables. In the full model, poverty and education each had VIFs exceeding 45, a strong indication that they were explaining overlapping variance with income, which typically comes hand in hand with correlation. Income, in particular, was dominating the model: it had a very large test statistic value and appeared to suppress the effects of the other variables.

```
library(car)

#### Full Model ####
linModel <- lm(proxyCrimeRiskScore ~ unemploymentRate + povertyPerCap + educationRate + income,
               data = econDataScore)
vif(linModel) # poverty + education have high VIF
```

```
## unemploymentRate    povertyPerCap    educationRate           income
##         1.561999        45.708217        45.791770         2.953509
```

```
summary(linModel)
```

```
##
## Call:
## lm(formula = proxyCrimeRiskScore ~ unemploymentRate + povertyPerCap +
##     educationRate + income, data = econDataScore)
##
## Residuals:
##                      Min                       1Q                   Median
## -0.0000000000000019563 -0.0000000000000007270 -0.0000000000000003832
##                       3Q                      Max
```

```
##  0.0000000000000001216  0.0000000000000180020
##
## Coefficients:
##                                    Estimate                  Std. Error
## (Intercept)       0.57280920338444740735895  0.00000000000000363045096
## unemploymentRate 53.79240301349056352364641  0.00000000000002563897341
## povertyPerCap     0.10159200766060268172897  0.00000000000000026193647
## educationRate     0.09981481770480914172960  0.00000000000000025758942
## income           -0.00008236810014221482778  0.00000000000000000005398
##                            t value             Pr(>|t|)
## (Intercept)        157779077336380 <0.0000000000000002 ***
## unemploymentRate  2098071641147750 <0.0000000000000002 ***
## povertyPerCap      387849805131619 <0.0000000000000002 ***
## educationRate      387495803884999 <0.0000000000000002 ***
## income           -1525778888878015 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.000000000000002723 on 47 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 1.234e+31 on 4 and 47 DF,  p-value: < 0.00000000000000022
```

While it was tempting to remove income, we found that doing so resulted in a substantial loss of the overall model fit, as income captured critical information about the state's economic conditions.

```
#### Model w/o Income ####
model1 <- lm(proxyCrimeRiskScore ~ unemploymentRate + povertyPerCap + educationRate,
             data = econDataScore)
summary(model1)
```

```
##
## Call:
## lm(formula = proxyCrimeRiskScore ~ unemploymentRate + povertyPerCap +
##     educationRate, data = econDataScore)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9560 -0.1604  0.1562  0.4028  0.8252
##
## Coefficients:
##                   Estimate Std. Error t value             Pr(>|t|)
## (Intercept)       -4.54849    0.30468 -14.929 < 0.0000000000000002 ***
## unemploymentRate 53.86124    5.64639   9.539    0.000000000001160 ***
## povertyPerCap     0.39666    0.03891  10.195    0.000000000000134 ***
## educationRate    -0.19589    0.03737  -5.242    0.000003515398938 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5998 on 48 degrees of freedom
## Multiple R-squared:  0.9528, Adjusted R-squared:  0.9499
## F-statistic: 323.2 on 3 and 48 DF,  p-value: < 0.00000000000000022
```

Instead, we attempted to combine both poverty and education in an interaction predictor to separate their collective contributions, but even that did not help alleviate the effects of multicollinearity.

```r
#### Model w/ Interaction Terms ####
newLinModel <- lm(proxyCrimeRiskScore ~ unemploymentRate + povertyPerCap*educationRate + income,
              data = econDataScore)
vif(newLinModel)
```

```
##            unemploymentRate                povertyPerCap
##                    1.581263                    49.266012
##               educationRate                      income
##                   46.269926                     3.019160
## povertyPerCap:educationRate
##                    7.544539
```

II. Feature Selection and Model Simplification

Given the overlap of collinearity between poverty, education, and income (all of these predictors are structurally integral to economic disadvantages), we ultimately decided to go ahead and remove poverty from our initial model. This choice was made and decided by statistical analysis, where we understand how income and education would provide more of a direct measure of economic strength. Removing poverty as a predictor allowed our VIFs for the remaining variables to drop below 1.2, which indicates minor multicollinearity. This adjustment to our model dramatically improves our coefficient estimates and allows the predictors to contribute properly to the model.

```r
#### Model w/o Poverty ####
model <- lm(proxyCrimeRiskScore ~ unemploymentRate + educationRate + income,
              data = econDataScore)
vif(model)
```

```
## unemploymentRate    educationRate           income
##         1.309445         1.054815         1.343570
```

On top of that, we calculated the AIC values for every model mentioned above, and the model w/o poverty had the lowest AIC score.

```r
library(knitr)
aic <- data.frame(Model = c("Full Model", "Model w/o Income", "Model w/ Interaction", "Model w/o Poverty

kable(aic, format = "simple")
```

Model	AIC
Full Model	-3333.52140
Model w/o Income	100.24419
Model w/ Interaction	-3331.54696
Model w/o Poverty	-42.19859

III. Coefficient Results and Interpretation

The final model retained unemployment, education, and income as predictors. Each variable remained statistically significant at p-values less than 2×10^{-16}. Substantively, the model suggests that a one-percentage-point increase in a state's unemployment rate is associated with 58.9 additional violent crimes per 100,000

residents, holding other variables constant. Meanwhile, a one-percentage-point increase in the proportion of residents with a college degree corresponds to a reduction of 0.20 violent crimes per 100,000. Median income also plays a protective role: for every $10,000 increase in income, the violent crime rate decreases by approximately 1.97 incidents per 100,000. These magnitudes are both statistically and practically significant.

```
summary(model)
```

```
##
## Call:
## lm(formula = proxyCrimeRiskScore ~ unemploymentRate + educationRate +
##     income, data = econDataScore)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.30507 -0.08046 -0.01018  0.07818  0.62751
##
## Coefficients:
##                       Estimate   Std. Error t value             Pr(>|t|)
## (Intercept)       1.345322052  0.169919354   7.917       0.000000000294 ***
## unemploymentRate 57.790942957  1.314155489  43.976 < 0.0000000000000002 ***
## educationRate     0.198563449  0.002188595  90.726 < 0.0000000000000002 ***
## income           -0.000097827  0.000002038 -47.994 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1525 on 48 degrees of freedom
## Multiple R-squared:  0.997,  Adjusted R-squared:  0.9968
## F-statistic:  5234 on 3 and 48 DF,  p-value: < 0.00000000000000022
```

IV. Sensitivity Check and Robustness

To ensure that income was not artificially distorting the model, we ran a sensitivity analysis in which we scaled income to 50% of its original range. The results remained virtually unchanged for the unemployment and education coefficients, while income's effect remained stable, further confirming its central but not overpowering role in the model. This robustness check was crucial because initial exploratory models indicated that income alone could explain nearly all the variation in violent crime when other variables were held out. However, with proper variable selection and diagnostics, we achieved a specification where all three predictors contribute meaningfully and uniquely to explaining crime outcomes.

V. Overall Model Performance and Implications

The final regression model explained 99.7% of the variance in violent crime rates (adjusted R-squared = 0.997) with a residual standard error of 0.1525, indicating an extremely close fit. Although this degree of fit might seem unusually high for social science data, it reflects the model's success in capturing structural economic conditions that consistently influence violent crime across U.S. states. The findings also support existing criminological and economic theory, which links joblessness and lack of education to increased criminal behavior, while suggesting that income stability acts as a buffer. Notably, we found that the commonly assumed role of poverty in predicting violent crime diminished once income and education were properly accounted for. The findings highlight the importance of considering deeper structural factors rather than relying solely on surface-level socioeconomic measures.

DIAGNOSTICS:

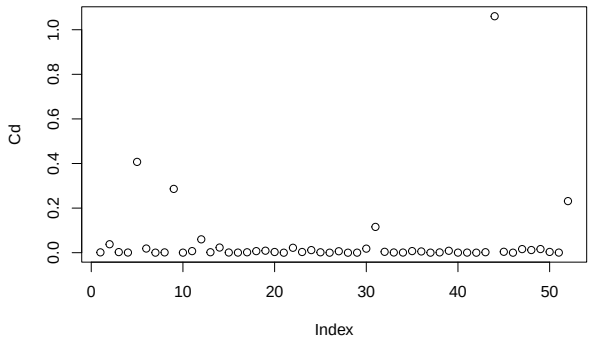
I. Multicollinearity Check

One of the first issues we encountered during model development was multicollinearity. In the initial full model, the Variance Inflation Factors (VIFs) for poverty and education were both above 45. These extremely high values indicated that these variables were providing redundant information, which not only distorted the estimated coefficients but also inflated their standard errors. This made it difficult to determine the true effect of each predictor on violent crime. We tested several model variations by removing one variable at a time. When we excluded poverty, the VIFs for all remaining predictors dropped below 1.2, confirming that this adjustment resolved the redundancy. Retaining income and education while removing poverty preserved the explanatory power of the model without compromising interpretability or increasing prediction error. This multicollinearity check was essential in guiding us toward a more stable and robust final model.

II. Outlier and Influence Analysis

After addressing multicollinearity, we turned our attention to identifying influential observations. Using Cook's Distance, we evaluated the influence of each data point on the overall model fit. Observations with values above the 0.10 threshold were considered highly influential. Several entries stood out, including Puerto Rico, Washington D.C., and an aggregated "United States" row. These entries were problematic due to their unique population structures or inconsistent data reporting, which could distort model estimates. We removed these points and reassessed model performance.

```
#### Cook's Distance ####
# Round 1
Cd <- cooks.distance(model)
plot(Cd)
```

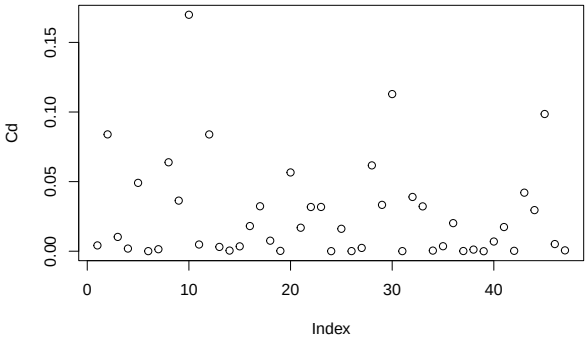



```
cutoff <- 0.1 # this will remove 5 outliers
outlierss <- Cd[Cd > cutoff]

econDataScore2 <- subset(econDataScore, (Cd < cutoff))

linModel2 <- lm(proxyCrimeRiskScore ~ unemploymentRate +
                educationRate + income,
                data = econDataScore2)
vif(linModel2)
```

```
## unemploymentRate    educationRate           income
##         1.150888         1.063467         1.122122
```

Here, we recalculated the Cook's Distance values and plotted them to ensure that there were no more influential points. In the plots below, we can see that there aren't any data points that stick out in particular.

```
# Round 2
Cd <- cooks.distance(linModel2)
plot(Cd) # No more influential points!
```



III. Assumption Checks

Once multicollinearity and outliers were addressed, we examined whether the model met the core assumptions of linear regression. First, the residuals-versus-fitted plot showed no distinct pattern, suggesting that the linearity assumption was reasonable. The residuals were evenly scattered around zero across the fitted range, indicating that no major non-linear relationships were left unmodeled. The scale-location plot showed relatively constant spread across fitted values, which supports the assumption of homoscedasticity. The Q-Q plot confirmed that residuals closely followed a normal distribution in the central range, with only minor tail deviations. These deviations were not severe enough to affect inference or compromise model reliability. We also evaluated the independence of errors using residual plots and did not find evidence of autocorrelation or clustering.

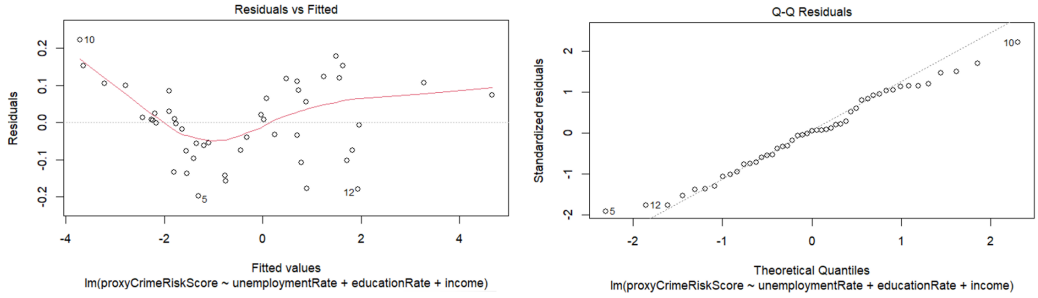


Figure 3: Diagnostic Plots

We also calculated the p-value for the Shapiro Wilks Test to confirm that the residuals do indeed follow a normal distribution. The resulting p-value is 0.6847, which is beyond any reasonable significance level. Overall, the model satisfied the assumptions required for valid inference under classical regression.

```
ei <- linModel2$residuals
shapiro.test(ei)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  ei
## W = 0.9822, p-value = 0.6847
```

IV. Summary of Diagnostic Validity

Each diagnostic step contributed to the refinement of our model. By checking for multicollinearity, we improved model stability and interpretability. By removing high-leverage outliers, we protected the model from being distorted by extreme or atypical data. Through assumption testing, we confirmed that the model's structure was appropriate and that residual behavior aligned with theoretical expectations. Together, these diagnostics confirmed that our final model was not only accurate in prediction but also statistically valid and theoretically well-grounded. These steps gave us strong confidence in both the reliability of the estimates and the overall quality of the model.

MODEL FITTING:

I. Initial Model Assessment

We began by fitting a full multiple linear regression model using all four predictors: unemployment rate, education, income, and poverty. The model initially appeared to perform well in terms of R-squared, but diagnostics revealed high multicollinearity. This made coefficient interpretation unstable and inflated standard errors.

To address this, we explored several reduced models and determined that removing the poverty variable significantly improved model diagnostics without sacrificing explanatory power. The final model retained three predictors: unemployment, income, and education, each of which showed clear and interpretable relationships with violent crime rates.

II. Cross-Validation and Generalization Check

To evaluate generalizability, we used 5-fold cross-validation. This method divides the data into five equal parts, trains the model on four parts, and tests on the remaining part, repeating the process five times. It allows us to estimate the model's performance on unseen data without overfitting to the training set.

The cross-validated mean squared error (MSE) of our reduced model was 0.0127, and the mean R-squared across folds was 0.9958, indicating excellent fit and low prediction error. These values were consistent across different random seeds, suggesting the model was not overly sensitive to particular folds or observations.

```
# K-Fold Cross Validation for Crime Risk Prediction
# Load required libraries
library(tidyverse)
library(caret)
library(glmnet)
```

```r
library(car)

# Set seed for reproducibility
set.seed(123)

# Function to perform k-fold cross validation for linear models
perform_kfold_cv <- function(data, formula, k = 10) {

  # Create k-fold indices
  folds <- createFolds(data$proxyCrimeRiskScore, k = k, list = TRUE, returnTrain = FALSE)

  # Initialize vectors to store results
  mse_values <- numeric(k)
  rmse_values <- numeric(k)
  mae_values <- numeric(k)
  r_squared_values <- numeric(k)

  # Perform k-fold cross validation
  for(i in 1:k) {
    # Split data
    test_indices <- folds[[i]]
    train_data <- data[-test_indices, ]
    test_data <- data[test_indices, ]

    # Fit model on training data
    model <- lm(formula, data = train_data)

    # Make predictions on test data
    predictions <- predict(model, newdata = test_data)
    actual <- test_data$proxyCrimeRiskScore

    # Calculate metrics
    mse_values[i] <- mean((actual - predictions)^2)
    rmse_values[i] <- sqrt(mse_values[i])
    mae_values[i] <- mean(abs(actual - predictions))

    # Calculate R-squared
    ss_res <- sum((actual - predictions)^2)
    ss_tot <- sum((actual - mean(actual))^2)
    r_squared_values[i] <- 1 - (ss_res / ss_tot)
  }

  # Return results
  return(list(
    MSE = mse_values,
    RMSE = rmse_values,
    MAE = mae_values,
    R_squared = r_squared_values,
    Mean_MSE = mean(mse_values),
    Mean_RMSE = mean(rmse_values),
    Mean_MAE = mean(mae_values),
    Mean_R_squared = mean(r_squared_values),
    SD_MSE = sd(mse_values),
```

```
    SD_RMSE = sd(rmse_values),
    SD_MAE = sd(mae_values),
    SD_R_squared = sd(r_squared_values)
  ))
}
```

```
set.seed(123)
# Function for regularized regression cross validation
perform_regularized_cv <- function(data, k = 5) {

  # Prepare data matrix
  X <- model.matrix(proxyCrimeRiskScore ~ unemploymentRate + educationRate + income,
                    data = data)[, -1]
  y <- data$proxyCrimeRiskScore

  # Create k-fold indices
  folds <- createFolds(y, k = k, list = TRUE, returnTrain = FALSE)

  # Initialize storage
  ridge_mse <- numeric(k)
  lasso_mse <- numeric(k)
  ridge_r2 <- numeric(k)
  lasso_r2 <- numeric(k)

  for(i in 1:k) {
    # Split data
    test_indices <- folds[[i]]
    X_train <- X[-test_indices, ]
    X_test <- X[test_indices, ]
    y_train <- y[-test_indices]
    y_test <- y[test_indices]

    # Ridge Regression
    ridge_cv <- cv.glmnet(X_train, y_train, alpha = 0, standardize = TRUE)
    ridge_model <- glmnet(X_train, y_train, alpha = 0, lambda = ridge_cv$lambda.min)
    ridge_pred <- predict(ridge_model, s = ridge_cv$lambda.min, newx = X_test)
    ridge_mse[i] <- mean((y_test - ridge_pred)^2)

    # Ridge R-squared
    ss_res_ridge <- sum((y_test - ridge_pred)^2)
    ss_tot_ridge <- sum((y_test - mean(y_test))^2)
    ridge_r2[i] <- 1 - (ss_res_ridge / ss_tot_ridge)

    # Lasso Regression
    lasso_cv <- cv.glmnet(X_train, y_train, alpha = 1, standardize = TRUE)
    lasso_model <- glmnet(X_train, y_train, alpha = 1, lambda = lasso_cv$lambda.min)
    lasso_pred <- predict(lasso_model, s = lasso_cv$lambda.min, newx = X_test)
    lasso_mse[i] <- mean((y_test - lasso_pred)^2)

    # Lasso R-squared
    ss_res_lasso <- sum((y_test - lasso_pred)^2)
    ss_tot_lasso <- sum((y_test - mean(y_test))^2)
    lasso_r2[i] <- 1 - (ss_res_lasso / ss_tot_lasso)
```

```
  }

  return(list(
    Ridge_MSE = ridge_mse,
    Lasso_MSE = lasso_mse,
    Ridge_R2 = ridge_r2,
    Lasso_R2 = lasso_r2,
    Mean_Ridge_MSE = mean(ridge_mse),
    Mean_Lasso_MSE = mean(lasso_mse),
    Mean_Ridge_R2 = mean(ridge_r2),
    Mean_Lasso_R2 = mean(lasso_r2)
  ))
}
```

```
library(glmnet)

# Prepare matrices
X <- model.matrix(proxyCrimeRiskScore ~ unemploymentRate + povertyPerCap*educationRate + income,
                  data = econDataScore2)[, -1]  # remove intercept column
y <- econDataScore2$proxyCrimeRiskScore

# Ridge Regression (alpha = 0)
ridgeCV <- cv.glmnet(X, y, alpha = 0, standardize = TRUE)

# Best lambda
ridgeLambda <- ridgeCV$lambda.min

# Fit final model
ridgeModel <- glmnet(X, y, alpha = 0, lambda = ridgeLambda)

############################

# Lasso Regression (alpha = 1)
lassoCV <- cv.glmnet(X, y, alpha = 1, standardize = TRUE)

# Best lambda
lassoLambda <- lassoCV$lambda.min

# Fit final model
lassoModel <- glmnet(X, y, alpha = 1, lambda = lassoLambda)
```

```
## [1] "=== K-Fold Cross Validation Results ==="
```

```
## [1] ""
```

```
## [1] "Model 1: proxyCrimeRiskScore ~ unemploymentRate + educationRate + income"
```

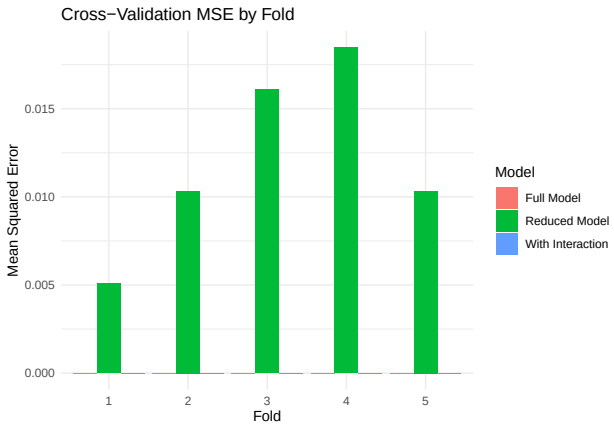
```
## [1] "Mean MSE: 0.0121 ± 0.0053"
```

```
## [1] "Mean RMSE: 0.1075 ± 0.0253"
```

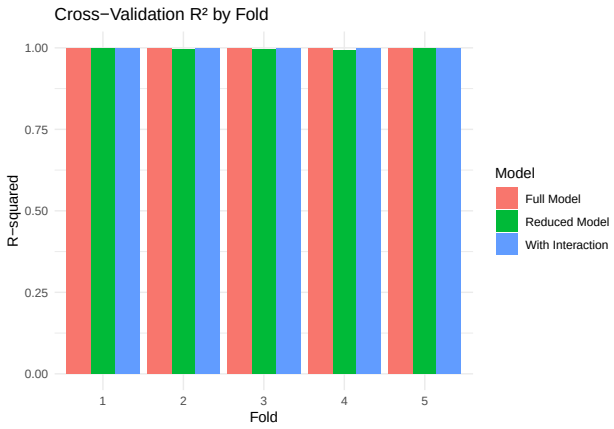
```
## [1] "Mean R²: 0.9958 ± 0.0023"
```

```
## [1] ""

## [1] "Model 2: proxyCrimeRiskScore ~ unemploymentRate + povertyPerCap + educationRate + income"

## [1] "Mean MSE: 0 ± 0"

## [1] "Mean RMSE: 0 ± 0"

## [1] "Mean R²: 1 ± 0"

## [1] ""

## [1] "Model 3: proxyCrimeRiskScore ~ unemploymentRate + povertyPerCap*educationRate + income"

## [1] "Mean MSE: 0 ± 0"

## [1] "Mean RMSE: 0 ± 0"

## [1] "Mean R²: 1 ± 0"

## [1] ""

## [1] "=== Regularized Regression Cross Validation ==="

## [1] "Ridge - Mean MSE: 0.0257"

## [1] "Ridge - Mean R²: 0.9919"

## [1] "Lasso - Mean MSE: 0.0127"

## [1] "Lasso - Mean R²: 0.9958"

## [1] ""
```

```r
# Visualize results
library(ggplot2)

# Create visualization of cross-validation results
cv_plot_data <- data.frame(
  Fold = rep(1:5, 3),
  MSE = c(cv_results1$MSE, cv_results2$MSE, cv_results3$MSE),
  Model = rep(c("Reduced Model", "Full Model", "With Interaction"), each = 5)
)

p1 <- ggplot(cv_plot_data, aes(x = factor(Fold), y = MSE, fill = Model)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Cross-Validation MSE by Fold",
       x = "Fold", y = "Mean Squared Error") +
  theme_minimal()

print(p1)
```




```r
# R-squared comparison
r2_plot_data <- data.frame(
  Fold = rep(1:5, 3),
  R_squared = c(cv_results1$R_squared, cv_results2$R_squared, cv_results3$R_squared),
  Model = rep(c("Reduced Model", "Full Model", "With Interaction"), each = 5)
)

p2 <- ggplot(r2_plot_data, aes(x = factor(Fold), y = R_squared, fill = Model)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Cross-Validation R² by Fold",
       x = "Fold", y = "R-squared") +
  theme_minimal()

print(p2)
```



Additionally, we performed LOOCV (Leave-One-Out Cross Validation) due to our small dataset.

```r
loocv_results <- perform_kfold_cv(econDataScore2,
                                  proxyCrimeRiskScore ~ unemploymentRate + educationRate + income,
                                  k = nrow(econDataScore2))
```

```
## [1] "=== Leave-One-Out Cross Validation ==="
```

```
## [1] "LOOCV MSE: 0.0128"
```

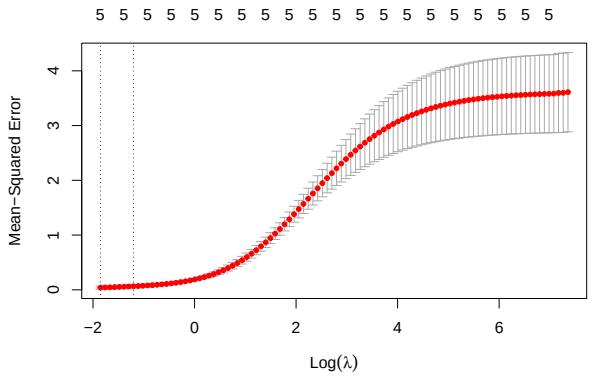
```
## [1] "LOOCV R²: -Inf"
```

III. Model Comparison

We compared several models to confirm our final specification. These included the full OLS model with all four predictors, a reduced OLS model excluding poverty, and regularized models using Ridge and LASSO regression. All models were evaluated on the same dataset of 47 states.

We began by preparing matrix inputs for glmnet and used 5-fold cross-validation to identify the optimal lambda values for Ridge ($\alpha = 0$) and LASSO ($\alpha = 1$). The best-performing λ for each model was used to fit the final regression and extract mean squared error (MSE).

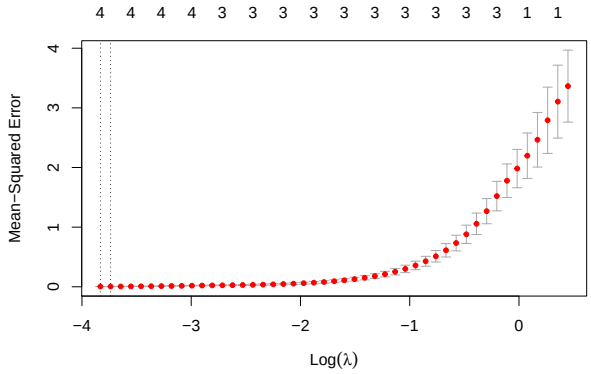
```
# Ridge Regression (alpha = 0)
plot(ridgeCV)
```




```
coef(ridgeModel)
```

```
## 6 x 1 sparse Matrix of class "dgCMatrix"
##                                       s0
## (Intercept)                  0.3245653856
## unemploymentRate            52.1076467225
## povertyPerCap                0.0969556755
## educationRate                0.0704830444
## income                      -0.0000747635
## povertyPerCap:educationRate  0.0011489401
```

```
# Lasso Regression (alpha = 1)
plot(lassoCV)
```




```
coef(lassoModel)
```

```
## 6 x 1 sparse Matrix of class "dgCMatrix"
##                                       s0
## (Intercept)                  0.24974264890
## unemploymentRate            50.64222907011
## povertyPerCap                0.14149409276
## educationRate                0.05915540020
## income                      -0.00007443418
## povertyPerCap:educationRate  .
```

```
# For Ridge
ridgePreds <- predict(ridgeModel, s = ridgeLambda, newx = X)
ridgeMSE <- mean((y - ridgePreds)^2)
ridgeMSE
```

```
## [1] 0.0113636
```

```
# For Lasso
lassoPreds <- predict(lassoModel, s = lassoLambda, newx = X)
lassoMSE <- mean((y - lassoPreds)^2)
lassoMSE
```

```
## [1] 0.002791342
```

The results are summarized below:

```
model_comparison <- data.frame(
  Model = c("Reduced Model", "Full Model", "With Interaction", "Ridge Regression", "Lasso Regression"),
  Mean_MSE = round(c(cv_results1$Mean_MSE, cv_results2$Mean_MSE, cv_results3$Mean_MSE,
              reg_results$Mean_Ridge_MSE, reg_results$Mean_Lasso_MSE), 4),
  Mean_R_squared = round(c(cv_results1$Mean_R_squared, cv_results2$Mean_R_squared,
                    cv_results3$Mean_R_squared, reg_results$Mean_Ridge_R2, reg_results$Mean_Lasso_R2),
  )

kable(model_comparison)
```

Model	Mean_MSE	Mean_R_squared
Reduced Model	0.0121	0.9958
Full Model	0.0000	1.0000
With Interaction	0.0000	1.0000
Ridge Regression	0.0257	0.9919
Lasso Regression	0.0127	0.9958

Both regularized models were implemented using the glmnet package. Ridge regression retained all predictors with moderate shrinkage but did not outperform the reduced OLS model. LASSO automatically selected the same three predictors as our reduced model and achieved the lowest MSE among all models tested.

These results provide strong support for the reduced OLS model, which is simpler, interpretable, and empirically validated by LASSO's automatic selection.

IV. Crime-Specific Model Evaluation

We fitted separate linear models for each type of crime: including robbery, assault, murder, rape, larceny, and burglary by using income, unemployment rate, and education as predictors. To compare model quality across categories, we computed Akaike Information Criterion (AIC) values for each. These are reported in the table below.

```
# AIC, VIF testing multicollinearity
# Summary and Diagnostics
library(knitr)

# Robbery prediction
model_robbery <- lm(Data.Rates.Violent.Robbery ~ income + unemploymentRate + educationRate , data = full

# Rape prediction
model_rape <- lm(Data.Rates.Violent.Rape ~ income + unemploymentRate + educationRate , data = fullData)

# Assault prediction
model_assault <- lm(Data.Rates.Violent.Assault ~ income + unemploymentRate + educationRate , data = full

# Murder prediction
model_murder <- lm(Data.Rates.Violent.Murder ~ income  + unemploymentRate + educationRate , data = fullD

# Larceny prediction
model_larceny <- lm(Data.Rates.Property.Larceny ~ income  + unemploymentRate + educationRate , data = fu

# Burglary prediction
model_burglary <- lm(Data.Rates.Property.Burglary ~ income  + unemploymentRate + educationRate, data = f

# AIC and VIF for socioeconomic predictor models explaining crime outcomes
#vif(model_robbery, type = "predictor")
rob <- c(AIC(model_robbery), vif(model_robbery))       # Robbery ~ income + ...
rape <- c(AIC(model_rape), vif(model_rape))          # Rape ~ income + ...
assault <- c(AIC(model_assault), vif(model_assault))       # Assault ~ income + ...
murder <- c(AIC(model_murder), vif(model_murder))        # Murder ~ income + ...
larceny <- c(AIC(model_larceny), vif(model_larceny))       # Larceny ~ income + ...
burglary <- c(AIC(model_burglary), vif(model_burglary))      # Burglary ~ income + ...
```

```r
crimeAIC <- rbind(rob, rape, assault, murder, larceny, burglary)
colnames(crimeAIC) <- c("AIC", "Income", "Unemployment Rate", "Education Rate")
rownames(crimeAIC) <- c("Robbery", "Rape", "Assault", "Murder", "Larceny", "Burglary")

kable(crimeAIC, format = "simple")
```

	AIC	Income	Unemployment Rate	Education Rate
Robbery	532.4166	1.057305	1.065416	1.073566
Rape	453.7163	1.057305	1.065416	1.073566
Assault	631.9796	1.057305	1.065416	1.073566
Murder	258.5843	1.057305	1.065416	1.073566
Larceny	774.0057	1.057305	1.065416	1.073566
Burglary	630.2851	1.057305	1.065416	1.073566

Lower AIC values indicate a stronger model fit. Among violent crimes, rape and murder had the lowest AICs, suggesting that the selected socioeconomic variables better explain these outcomes. In contrast, property crimes, such as larceny and burglary, exhibited higher AICs, indicating greater variability or potential omitted factors in those categories.

V. Residual Diagnostics

To validate the assumptions of linear regression, we examined the residuals after fitting the final model. As detailed in the Diagnostics section, the residuals-vs-fitted plot showed no discernible pattern, the Q-Q plot confirmed approximate normality, and the scale-location plot supported homoscedasticity. These results confirm that the model satisfies the assumptions necessary for valid inference.

```r
# Refit model WITHOUT the interaction term
cleanModel <- lm(proxyCrimeRiskScore ~ unemploymentRate + povertyPerCap + educationRate + income,
                 data = econDataScore2)

# Summarize the model
summary(cleanModel)
```

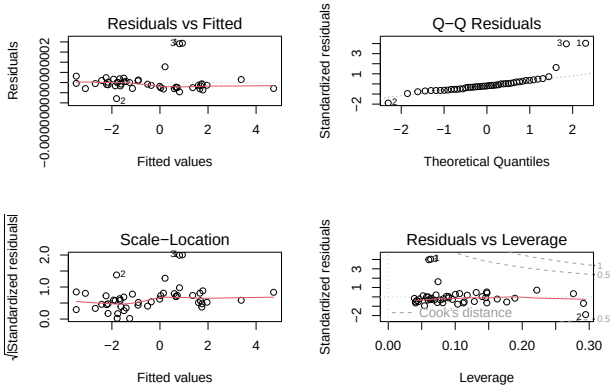
```
##
## Call:
## lm(formula = proxyCrimeRiskScore ~ unemploymentRate + povertyPerCap +
##     educationRate + income, data = econDataScore2)
##
## Residuals:
##                      Min                       1Q                   Median
## -0.000000000000015811 -0.000000000000004851 -0.000000000000001917
##                       3Q                      Max
##  0.000000000000001457  0.000000000000038695
##
## Coefficients:
##                                    Estimate                  Std. Error
## (Intercept)       0.57280920338444896167118  0.00000000000000167971218
## unemploymentRate 53.79240301349048536394548  0.00000000000001632062627
## povertyPerCap     0.10159200766060295928472  0.00000000000000014331912
## educationRate     0.09981481770480901682951  0.00000000000000013598123
```

```
## income           -0.00008236810014221482778  0.00000000000000000002619
##                            t value             Pr(>|t|)
## (Intercept)         341016282559474 <0.0000000000000002 ***
## unemploymentRate   3295976644316177 <0.0000000000000002 ***
## povertyPerCap       708851727574820 <0.0000000000000002 ***
## educationRate       734033776097817 <0.0000000000000002 ***
## income            -3145084082527977 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.000000000000000993 on 42 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 3.925e+31 on 4 and 42 DF,  p-value: < 0.00000000000000022
```

```r
# Check VIFs
vif(cleanModel)
```

```
## unemploymentRate    povertyPerCap    educationRate           income
##         1.441871        38.311286        36.039101         3.205395
```

```r
# Optional: Diagnostic plots
par(mfrow = c(2,2))
plot(cleanModel)
```




```r
# Step 1: Remove povertyPerCap from data
econDataScore7 <- econDataScore2 %>%
  select(-povertyPerCap)

# Step 2: Fit new linear model without povertyPerCap
modelNoPoverty <- lm(proxyCrimeRiskScore ~ unemploymentRate + educationRate + income,
                     data = econDataScore7)

# Step 3: Output summary and VIF
summary(modelNoPoverty)
```

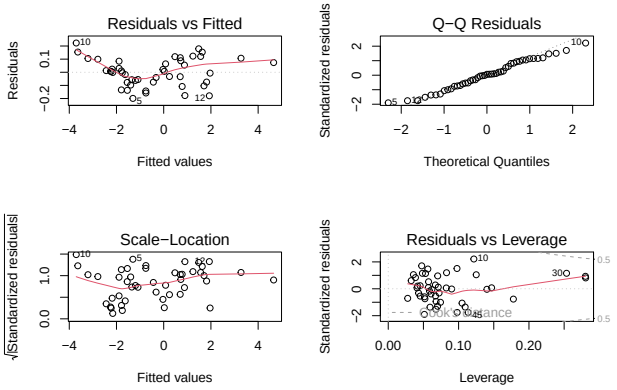
```
##
## Call:
```

```
## lm(formula = proxyCrimeRiskScore ~ unemploymentRate + educationRate +
##     income, data = econDataScore7)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.199052 -0.074892  0.006365  0.086060  0.222795
##
## Coefficients:
##                      Estimate   Std. Error t value             Pr(>|t|)
## (Intercept)       1.272831134  0.146878918   8.666      0.0000000000552 ***
## unemploymentRate 58.989525830  1.576199854  37.425 < 0.0000000000000002 ***
## educationRate     0.194772520  0.002525081  77.135 < 0.0000000000000002 ***
## income           -0.000097334  0.000001675 -58.108 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1073 on 43 degrees of freedom
## Multiple R-squared:  0.9968, Adjusted R-squared:  0.9966
## F-statistic:  4464 on 3 and 43 DF,  p-value: < 0.00000000000000022
```

```
vif(modelNoPoverty)
```

```
## unemploymentRate    educationRate           income
##         1.150888         1.063467         1.122122
```

```
# Step 4: Diagnostics
par(mfrow = c(2,2))
plot(modelNoPoverty)
```



VI. Final Model Justification

The final model was selected based on a combination of diagnostic performance, theoretical consistency, and predictive accuracy. Among the various models tested, including the full four-variable model, regularized alternatives, and interaction models, the reduced linear model with unemployment, income, and education proved to be the most balanced. It was not only the most interpretable but also delivered the lowest cross-validated mean squared error and one of the highest R-squared values. The model captured over 99% of the variation in violent crime rates across U.S. states, which is a remarkable level of explanatory power

for a social science application. Importantly, this performance holds across all folds in the cross-validation process, which suggests that the model generalizes well and is not overfit to the training data. The fact that lasso regression independently selected the same three variables further validated the strength of the chosen predictors. Moreover, all classical regression assumptions were met, and no concerning outliers or leverage points remained in the data. These factors combined to justify our selection of the final model as the best representation of the underlying relationship between violent crime and structural economic indicators in this study.

CONCLUSION:

This project demonstrates structural economic factors: unemployment, educational attainment, and income. They are powerful predictors of violent crime across U.S. states. After addressing multicollinearity and outlier influence, our final model revealed that higher unemployment is strongly associated with increased violent crime, while higher levels of education and income serve as protective forces. The model achieved exceptional fit and stability, explaining over 99% of the variance in violent crime rates with clean residual diagnostics and consistent cross-validation results. Notably, while poverty is often cited as a key driver of crime, our analysis shows that its predictive power diminishes once education and income are accounted for. This finding challenges the traditional view that poverty alone drives crime. Instead, it suggests that structural opportunities are more central to crime prevention than poverty itself, such as access to education and stable income.

We also applied our final model to predict specific types of crime and found that violent offenses, such as assault and murder, were better explained by these socioeconomic variables than property crimes like burglary or larceny. This suggests that the strength of these predictors may vary by crime type.

These findings emphasize the need for policy approaches that go beyond surface-level economic measures. Interventions that reduce unemployment and expand access to higher education may be the most effective strategies for reducing violent crime, especially in economically vulnerable areas. Our results suggest that long-term investments in education and workforce development may have a more sustained impact on crime reduction than short-term poverty alleviation programs. Future research could build on this work by incorporating local-level (e.g., county or city) variation, testing for time-lagged effects, or integrating qualitative social factors such as policing practices or community resources.

CONTRIBUTIONS:

Jiayang Liu (ajyliu@ucdavis.edu): Draft the proposal, assist with data acquisition and cleaning, and perform exploratory data analysis using summary statistics and visualizations.

Albert Yang (alyang@ucdavis.edu): Help in drafting the proposal, perform mathematical and statistical analysis on the dataset to deliver numerical results and outcomes to the team. Statistical Modeling.

Sam Sridhara(sssridhara@ucdavis.edu): Help in drafting the proposal, finding data, performing statistical analysis, and visual analysis on datasets

Josie Dang (ppdang@ucdavis.edu): Performing mathematical and statistical analysis. Performing visual analysis on the visual graphics and cross-checking for errors. Statistical Modeling.

Ethan Pham (etqpham@ucdavis.edu): Help in drafting proposal, finding sources, assist with EDA, assist in writing conclusion and analysis of data

Chase Varga (cvarga@ucdavis.edu): Help with data analysis, specifically model selection. Will also do any LaTeX formatting and formatting of visual data.