STAT-UB 103 Regression Project

Team Members: Geoffrey Budiman, Bowen Zhang

Introduction:

In this study, we aim to investigate the relationship between the violent crime rate and various socio-economic and legislative factors. We will be using the "More Guns, Less Crime?" dataset found online in the following link https://vincentarelbundock.github.io/Rdatasets/datasets.html. Our dataset comprises data from multiple states over several years, totaling 1173 observations, which provides a robust sample for analysis. The response variable that we have chosen for this study is the column titled ‘violent’, which has data for the violent crime rate per 100,000 individuals. We have selected three predictor variables: 'law', which indicates whether a state has a shall-carry law in effect; 'income', the real per capita personal income within the state in US dollars; and 'density', the population per square mile, divided by 1,000.

Before conducting any computations, we initially believe that all three factors have a significant effect on the violent crime rate. Our preliminary guess would be that the violent crime rate will be higher in states with a shall-carry law in effect, that have lower real per capita personal income, and that have a denser population. However, to fully understand this relationship, we will perform a rigorous statistical analysis, considering many possible levels of confidence, to gain insight on how these three predictor variables truly impact the rate of violent crimes. This analysis will involve us constructing a Regression Model in Rstudio using the dataset.

Preliminary Analysis:

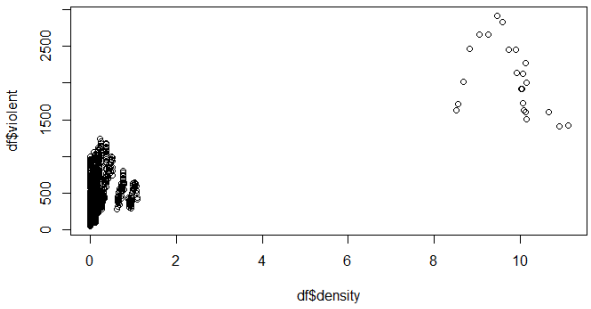
We first compute important descriptive statistics for the quantitative variables in the dataset

- violent
    - Mean: 503.0747
    - Variance: 111741.2
- income
    - Mean: 13724.8
    - Variance: 6525687
- density
    - Mean: 0.3520382
    - Variance: 1.837304

The correlation matrix for these three variables is given below

```
          violent    income   density
violent 1.0000000 0.4079864 0.6647260
income  0.4079864 1.0000000 0.3432839
density 0.6647260 0.3432839 1.0000000
```

Additionally, we have constructed two scatterplots below. The first one shows the violent crime rate against real per capita personal income and the second one shows the violent crime rate against population density.



Inference:

Before our analysis, we first conduct some preprocessing of our dataset. The code that we are using to do so is included in the Appendix at the end of this document.

1. Our first step is constructing a Regression Model in Rstudio using the following code

```
model1 <- lm(violent~., data = df)
summary(model1)
```

```
Call:
lm(formula = violent ~ ., data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-693.35 -190.08  -29.67  151.62  875.86

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.027e+02  3.969e+01   2.587   0.0098 **
income       2.748e-02  2.889e-03   9.515  < 2e-16 ***
lawyes      -1.106e+02  1.626e+01  -6.803 1.63e-11 ***
density      1.422e+02  5.479e+00  25.957  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 237.1 on 1169 degrees of freedom
Multiple R-squared:  0.4984,    Adjusted R-squared:  0.4971
F-statistic: 387.1 on 3 and 1169 DF,  p-value: < 2.2e-16
```

   Hence for the regression model

   $y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$

   We estimated the following coefficients

   $\alpha = 1.027\times10^{2}$, $\beta_1 = 2.748\times10^{-2}$, $\beta_2 = -1.106\times10^{2}$, $\beta_3 = 1.422\times10^{2}$

   Thus, our estimated regression model is as follows

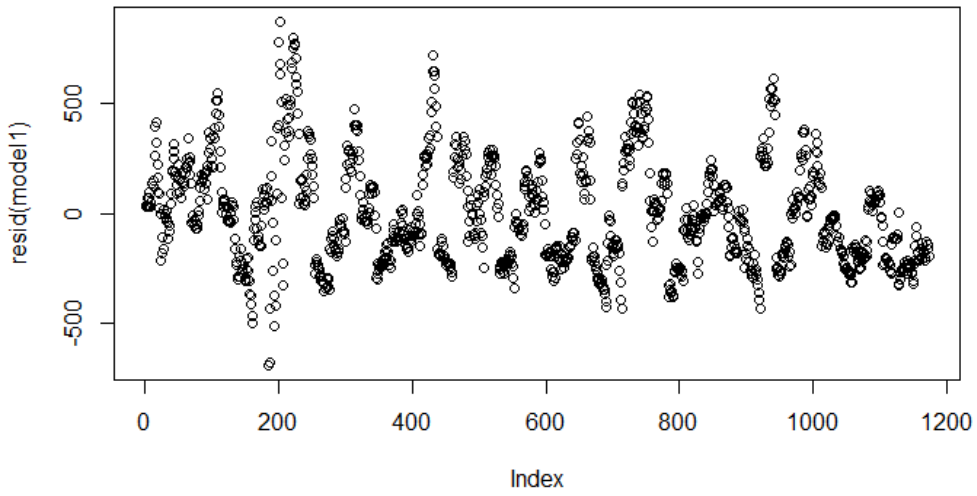
   $y_i = 1.027\times10^{2} + 2.748\times10^{-2} income_i - 1.106\times10^{2} law_i + 1.422\times10^{2} density_i + \varepsilon_i$

2. To compute an estimate for the variance of the residuals, we use the following code

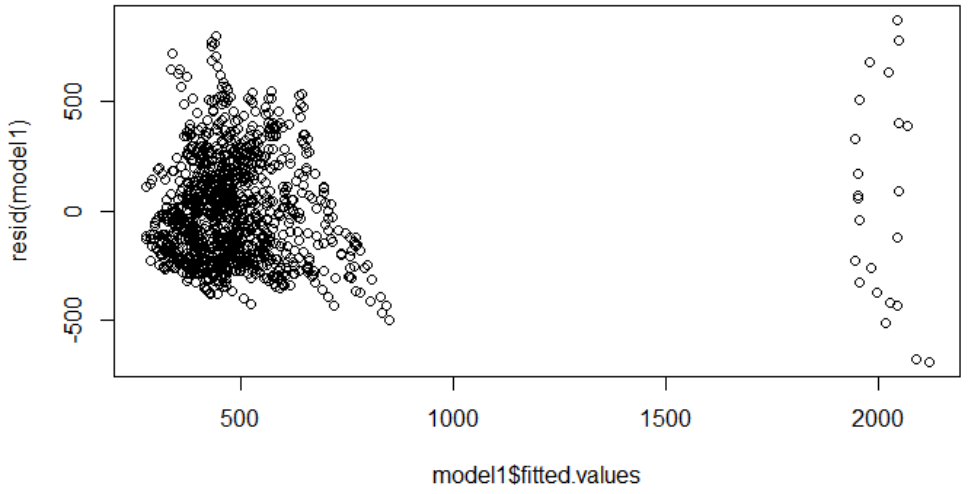
```
var(model1$residuals)
```

   This gives us 56053.3

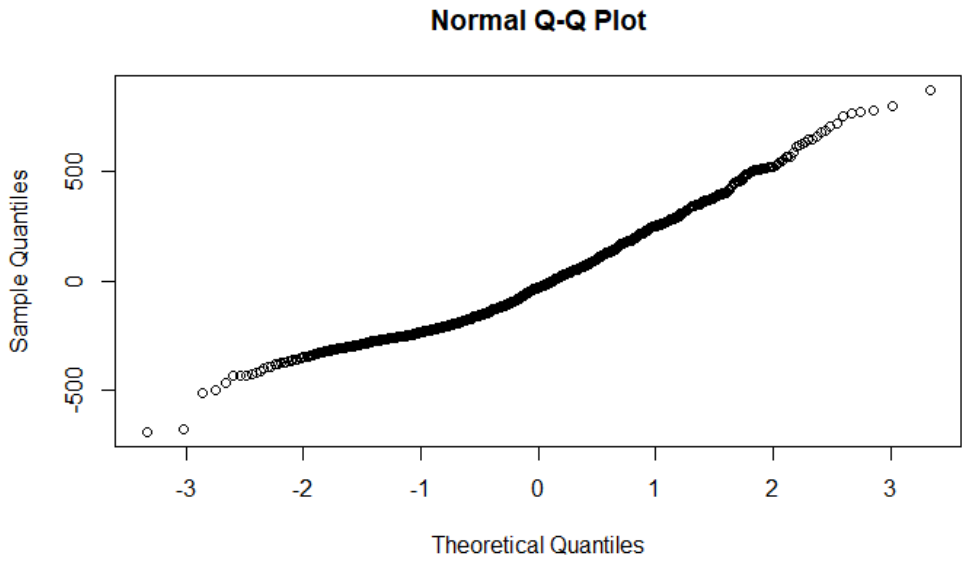
3. To perform residual analysis, we will need to construct the following graphs



This first graph is a scatterplot of the residuals plotted against the observation number. We can see from it that the residuals of the models do not seem to exhibit any sort of pattern, hence indicating that the residuals are independent of each other.



This second graph is a scatterplot of the residuals plotted against the value of the predictor variable.We can see from it that the residuals of the models do not seem to exhibit any sort of pattern, hence indicating that the residuals are independent of each other.



This third graph is a Q-Q Plot of the residuals. In it, we see that the dots mostly follow a straight line but there are some points at both ends that seem to deviate from the straight line. This suggests a potential violation of the normality assumption of our model.

From our residual analysis, we have found no issues with the assumptions that the residuals should be independent of each other. However, we have identified a potential violation of the normality assumption of our model. Regardless, we will continue with our analysis.

4. To see if there is any evidence of collinearity between the predictor variables, we refer to the following correlation matrix between the 2 quantitative predictor variables.

```
            income   density
income  1.0000000 0.3432839
density 0.3432839 1.0000000
```

There is no evidence of collinearity among the predictor variables that we have chosen for this study. Hence, we can proceed with these variables.

5. To compute a 95% confidence interval for each coefficient in the regression model, we use the following code

```
confint(model1, level=0.95)
```

```
                    2.5 %       97.5 %
(Intercept)   24.81246296 180.53953426
income         0.02181621   0.03315085
lawyes      -142.48461250 -78.69521130
density      131.45842707 152.95659132
```

The 95% confidence interval for the intercept tells us that we are 95% confident that the true value of the intercept is between 24.81246296 and 180.53953426.

The 95% confidence interval for the slope coefficient in front of income tells us that we are 95% confident that the true value of the slope coefficient is between 0.02181621 and 0.03315085. Since the entire range is positive, this indicates that there is a positive relationship between the violent crime rate and real per capita income.

The 95% confidence interval for the slope coefficient in front of lawyes tells us that we are 95% confident that the true value of the slope coefficient is between -142.48461250 and -78.69521130. Since the entire range is negative, this indicates that states with a shall carry law in effect in that year results in a lower violent crime rate.

The 95% confidence interval for the slope coefficient in front of density tells us that we are 95% confident that the true value of the slope coefficient is between 131.45842707 and 152.95659132. Since the entire range is positive, this indicates that there is a positive relationship between the violent crime rate and population density.

6. To compute a 99% confidence interval for each coefficient in the regression model, we use the following code

```
confint(model1, level=0.99)
```

```
                   0.5 %       99.5 %
(Intercept)    0.28474921 205.06724801
income         0.02003096   0.03493611
lawyes      -152.53172994 -68.64809386
density      128.07236956 156.34264882
```

The 99% confidence interval for the intercept tells us that we are 95% confident that the true value of the intercept is between 0.28474921 and 205.06724801.

The 99% confidence interval for the slope coefficient in front of income tells us that we are 99% confident that the true value of the slope coefficient is between 0.02003096 and 0.03493611. Since the entire range is positive, this indicates that there is a positive relationship between the violent crime rate and real per capita income.

The 99% confidence interval for the slope coefficient in front of lawyes tells us that we are 99% confident that the true value of the slope coefficient is between -152.53172994 and -68.64809386. Since the entire range is negative, this indicates that states with a shall carry law in effect in that year results in a lower violent crime rate.

The 99% confidence interval for the slope coefficient in front of density tells us that we are 99% confident that the true value of the slope coefficient is between 128.07236956 and 156.34264882. Since the entire range is positive, this indicates that there is a positive relationship between the violent crime rate and population density.

7. As a recap, our preliminary guess was that the violent crime rate will be higher in states with a lower real per capita personal income, a shall-carry law in effect, and a denser population. We will now conduct hypothesis testing for each of our predictions using information from the model summary in part 1 of our Inference.

   First, we conduct our hypothesis testing for the income predictor variable

   $H_0: \beta_1 = 0 \quad H_a: \beta_1 < 0$

   $RR: t_{value} <- t_{.995,1171}$

   We can calculate $t_{.995,1171}$ using the code 'qt(0.995, 1171)' which gives us 2.580034

   $t_{value} = 9.515 >- 2.580034$

   Since the rejection region does not hold, we fail to reject the null hypothesis that the slope coefficient is equal to 0 in favor of the alternative hypothesis that it is less than 0 with a 99% confidence level. Therefore, we cannot prove that the true coefficient is equal to our initial guess. Next, we conduct our hypothesis testing for the law predictor variable

   $H_0: \beta_2 = 0 \quad H_a: \beta_2 > 0$

   $RR: t_{value} > t_{.995,1171}$

   $t_{value} =- 6.803 < 2.580034$

   Since the rejection region does not hold, we fail to reject the null hypothesis that the slope coefficient is equal to 0 in favor of the alternative hypothesis that it is greater than 0 with a 99% confidence level. Therefore, we cannot prove that the true coefficient is equal to our initial guess. Next, we conduct hypothesis testing for the density predictor variable

   $H_0: \beta_3 = 0 \quad H_a: \beta_3 > 0$

   $RR: t_{value} > t_{.995,1171}$

   $t_{value} = 25.957 > 2.580034$

   Since the rejection region holds, we reject the null hypothesis that the slope coefficient is equal to 0 in favor of the alternative hypothesis that it is greater than 0 with a 99% confidence level. Therefore, we are 99% confident that the true coefficient is equal to our initial guess.

8. Next, we will conduct hypothesis tests for each predictor variable to test whether the predictor variable has an impact on the response variable.

   First, we conduct our hypothesis testing for the income predictor variable

   $H_0: \beta_1 = 0 \quad H_a: \beta_1 \neq 0$

   $RR: p_{value} < 1 - CL = 0.01$

   $p_{value} < 2 \times 10^{-16} < 0.01$

   Since the rejection region holds, we reject the null hypothesis that the slope coefficient is equal to 0 in favor of the alternative hypothesis that it is not equal to 0 with a 99% confidence level. Therefore, we are 99% confident that income has an impact on the violent crime rate. Next, we conduct our hypothesis testing for the law predictor variable

   $H_0: \beta_2 = 0 \quad H_a: \beta_2 \neq 0$

   $RR: p_{value} < 1 - CL = 0.01$

   $p_{value} < 1.63 \times 10^{-11} < 0.01$

   Since the rejection region holds, we reject the null hypothesis that the slope coefficient is equal to 0 in favor of the alternative hypothesis that it is not equal to 0 with a 99% confidence level. Therefore, we are 99% confident that law has an impact on the violent crime rate. Next, we conduct our hypothesis testing for the density predictor variable

   $H_0: \beta_3 = 0 \quad H_a: \beta_3 \neq 0$

   $RR: p_{value} < 1 - CL = 0.01$

   $p_{value} < 2 \times 10^{-16} < 0.01$

   Since the rejection region holds, we reject the null hypothesis that the slope coefficient is equal to 0 in favor of the alternative hypothesis that it is not equal to 0 with a 99% confidence level. Therefore, we are 99% confident that law has an impact on the violent crime rate.

9. Now, we will conduct a hypothesis test to test whether any of the predictor variables have an impact on the response variable.

   H_0: $\beta_1 = \beta_2 = \beta_3 = 0$ H_a: at least one of the predictor variables have an impact

   $RR: p_{value} < 1 - CL = 0.05$

   $p_{value} < 2.2 \times 10^{-16} < 0.05$

   Since the rejection region holds, we reject the null hypothesis that the slope coefficients for all three of the predictor variables are equal to 0 in favor of the alternative hypothesis that at least one of the coefficients is not equal to 0 with a 95% confidence level. Therefore, we are 95% confident that at least one of the predictor variables has an impact on the response variable.

10. Our next step is to split the dataset into a Train and Test set where the Train set contains 90% of the observations. We do this with the following code

```
n <- nrow(df)
trainIndex <- sample(1:n, size = round(0.9*n),
replace=FALSE)
train <- df[trainIndex,]
test <- df[-trainIndex,]
```

    Next, we estimate the model using the Train set and obtain the MSE of the model using the Test set

```
model2 <- lm(violent~., data = train)
est <- predict(model2, newdata = test)
res = test$violent-est
mse <- sum(res^2)/(length(est)-2)
```

    We find that the MSE is equal to 68269.16

11. Next, we compute the MSE of the regression model using n-fold cross validation with the three predictor variables using the following code

```
model_glm <- glm(violent~., data = df)
cv.glm(df,model_glm,K=nrow(df))$delta[1]
```

We find that the MSE is equal to 56692.43 which is lower than the MSE for the model we used in part 10.

12. Now, we compute the MSE of the regression model using n-fold cross validation with only income and density using the following code

```
model_glm_2 <- glm(violent~.-law, data = df)
cv.glm(df,model_glm_2,K=nrow(df))$delta[1]
```

We find that the MSE is equal to 58814.82 which is higher than the MSE for the model we used in part 11. Hence, we should use the law predictor variable in our model to predict the response variable.

Conclusion:

From the results of our hypothesis tests, we can conclude that all three predictor variables we have chosen—real per capita income within the state in US dollars; whether the state has a shall-issue carry law in effect; and the population density of the state—have a statistically significant impact on the violent crime rate. Additionally, from the MSEs that we have computed, we find that the model that uses all three predictor variables results in less errors than the model that only uses two of them. This gives further evidence that we should use all three of the predictor variables, income, law, and density, in a model to predict the response variable, the violent crime rate.

Appendix:

```
# Load in libraries
library(tidyverse)
library(caret)
library(boot)

df <- Guns %>% select(violent, income, law, density)

# Preliminary Analysis
mean(df$violent)
var(df$violent)
mean(df$income)
var(df$income)
mean(df$density)
var(df$density)

# Draw scatterplots
plot(df$violent,df$income)
plot(df$violent,df$density)

# regression model
model1 <- lm(violent~., data = df)

# 1 Estimate the coefficients in the Regression Model
summary(model1)

# 2 Compute an estimate for the variance of the residuals
var(model1$residuals)
```

```
# 3 Perform residual analysis
plot(resid(model1))
plot(model1$fitted.values, resid(model1))
qqnorm(resid(model1))

# 4 See if there is evidence of collinearity (check correlation)
cor(df[,c(1,2,4)])

# 5 Compute a 95% Confidence Interval for each coefficient
confint(model1, level=0.95)

# 6 Compute a 99% Confidence Interval for each coefficient
confint(model1, level=0.99)

# 7 Conduct a 99% CL Hypothesis test for each coefficient
summary(model1)
# income for example
# H0: beta_1 = 10
# Ha: beta_1 not= 10
alpha <- 0.01
t_value = (27.484-10)/2.889
p_value <- 2*pt(t_value, length(df)-2, lower.tail = F)
p_value < alpha

# 8, 9 Conduct a 99% CL Hypothesis test for each predictor
variable
summary(model1)
```

```
# 10 Split the dataset into a “Train” and “Test”
n <- nrow(df)
trainIndex <- sample(1:n, size = round(0.9*n), replace=FALSE)
train <- df[trainIndex,]
test <- df[-trainIndex,]

model2 <- lm(violent~., data = train)
est <- predict(model2, newdata = test)

res = test$violent-est
mse <- sum(res^2)/(length(est))
mse

# 11 declare the dataset
model_glm <- glm(violent~., data = df)
cv.glm(df,model_glm,K=nrow(df))$delta[1]



# 12 declare the dataset
aov(model_glm) # remove the variable with least variance
model_glm_2 <- glm(violent~.-law, data = df)

summary(model_glm_2)
cv.glm(df,model_glm_2,K=nrow(df))$delta[1]
```