State-level crime occurrance in the US in relation to household firmarm ownership: a depth study with respect to 2019 data

Tural Sadigov   Felix Pu   Andrew Gou

November 16, 2022

According to CDC's annual statistics, on average, 321 people are shot dead as victims of criminal activies in the US everyday. Yet, federal gun-control program has always faced resistence with disagreements over association between firearms and crime occurrence. In this paper, firearm data from FBI National Instant Criminal BackgroundCheck System and the state-level crime data from FBI Unified Crime Reporting Program in 2019 are used to test existence of potential association between gun possession and occurrence of crimes across the US. Three single linear regression models using the number of permits as explanatory variable and one multivariate regression model with a nuanced total gun explanatory variable are generated. The result suggests that there is a considerable positive correlation between the number of permits and the number of crimes. Moreover, the number of guns is also correlated with the total crime with a strong association, indicating that firearm ownership plays a crucial role in victimization.

Table of contents

1 Background and Significance

Gun control has always been a controversial topic in the United States. On one hand, "American gun owners - an estimated 80 million of them - see their firearms as protection to defend their own freedoms and property" (Smith 2022) under the Second Amendment, which notes "the right of the people to keep and bear Arms, shall not be infringed" ("Second Amendment" 2002). On the other hand, tragedies such as the 2018 Stoneman Douglas High School Shooting, where 17 young adults were murdered under "a pattern of disciplinary issues and unnerving behavior," (Rozsa, Berman, and Merle 2018) led other Americans to strongly advocate restrictive gun control policies.Tragedies related to firearms reignited the debate over gun control in the US. But realistically, many believe that "the argument over guns has simply become too politically divisive and culturally entrenched to allow for meaningful change" (Smith 2022). Even the the new gun laws New Jersey recently enacted in response to the Supreme Court's ruling are facing ongoing litigation and legal scrutiny (Han 2022). The tremendous resistance against gun rights sets the motivation behind composing this paper, specifically whether their resistance towards guns based on the reason that guns are associated with crimes can be backed by statistical analysis. Would more firearms lead to more crimes as anti-gun individuals contend for or is it unrelated that occurrences of crimes is independent from presence of firearms. With this question, this paper uses data collected by federal agencies at 2019 to investigate the potential association between gun possession and occurrences of crime across United States. This paper is written to answer to what degree is prevalence of gun possession associated with occurrence of crime: a focus study on United States based on 2019 data.

There are prior researches conducted pertaining to this topic and other relevant topics of this field. For example, Gary Kleck and E. Britt Patterson (Kleck and Patterson 1993) conclude in a prior study that the results are generally negative for the violence control effectiveness of gun control. Yet, the significance of the few supportive results should not be overlooked. There do appear to be some gun controls which work. In addition, Philip J. Cook (Cook 1981) studies the effect of Gun availability on violent crime patterns. And Cook's research concludes that it is reasonable to suppose that moderate gun control measures have the potential to discourage some violent criminals - potential or active - from obtaining guns. Miller (Miller, Hemenway, and Azrael 2007)examines the role played by household firearms in homicide victimization, whom concludes existence of potential positive association between percentage of household firearm ownership and crime rates across the United States based on data collected in the early 2000s.

However, this research is significant for three reasons. Firstly, most of the prior researches are done based on historical data when the partisan divide on gun rights and regulations was not as significant, such as a study conducted by Branas(Branas et al. 2009). Therefore, one of the goals of this paper is to update this field with analysis based on recent data. Secondly, prior researches did not emphasize the specifics of guns owned in terms of their type and actual number but substituted by a vague percentage of household firearm ownership. Thirdly, different from prior studies, this research uses number of permits as the initial explanatory

variable instead of number of firearms presented at each state. Permit is intentionally chosen because this research intends to extend the magnitude of the explanatory variable by including all the individuals with the potential of owning firearms, thereby holistically examining whether there is an association between a broadly defined magnitude of 'gun possession' (numbers of permit) and occurrence of crimes across the United States in 2019, in which the null hypothesis and the alternative hypothesis are given as the following. There are four tests conducted in this paper, thereby the significance level is set at 0.0125 as calculated by 0.05/4 to avoid p-hacking practice.

Hypotheses:

Null Hypothesis: There is no association between gun possession and occurrence of crime across the southern states in 2019

Alternative Hypothesis: There is association between gun possession and occurrence of crime across the southern states in 2019

Significance Level/Alpha: 0.0125

2 Methods

Raw Data Description:

The "Firearm checks" data (Singer-Vine 2019) is recorded by the FBI's National Instant Criminal Background Check System from November 1998 – September 2022. Data is collected by all the registered sales and license administrations. Before every permit grant and sale, personnel would call the FBI or other designated agencies to ensure that customer does not have a criminal record. Otherwise, customers are ineligible to obtain the permit nor make a purchase. And when purchasing firearms, the type of gun purchased will be recorded as well.

The "state crime" data from 1960 to 2019 (C.Bart et al. 2022) is collected by the FBI's Unified Crime Reporting Program (UCR) using its Summary Reporting System along with the assistance from the U.S. Department of Justice and the Federal Bureau of Investigation. The data contains two main categories: property and violent crime. Property crime refers to burglary, larceny, and motor related crime while violent crime refers to assault, murder, rape, and robbery.

Methods:

The analytic method used in research is regression analysis, both simple linear regression and multivariate linear regression, under R Studio (R Core Team 2021). Before, conducting regression analysis, data within the scope of interest is selected to create new data frame with aid of R Studio package (Wickham et al. 2019). Both raw firearm checks data set and state crime data set are first limited to the year of 2019, the year of this research's interest. Because the firearm check data is categorized by months, the first assumption behind this paper is that

the data of December 2019 from the firearm date frame will be used as the proxy for the entire 2019. December is chosen for two reasons: it is the last month of the year when permit holders of earlier months would maintain their qualifications through the year while new permits of the year are approved; December is also the month when people take their Christmas break from school and work that opens up more chances of gun related crimes.

After filtering the data frame by year, six columns of variable are chosen in total and bind to create the final data frame to be analyzed in this research. Five variables are chosen from the firearm checks data which are month, state, permit, handgun, and long-gun. The key explanatory variable 'permit' is defined as 'the number of people with a gun permit,' while the two 'gun' named variable are defined with respect to their number of presences at each state. Variables 'handgun' and 'long-gun' are chosen with purpose to nuance the regression analysis after generating the machine-learning models, which will be explained towards the end of this section. Two additional variables are chosen from the state_crime data set and added together to create a new variable: total number of recorded crimes happened at 2019, which includes the total number of recorded violent crimes: assaults, murders, rapes, and robberies and the total number of the property-related crime: burglaries, larcenies, and motor crimes.

After the data frame is set, the analysis starts by plotting the two variables of interest: number of permits in each state and the total number of reported crime in each state into a scatter-plot to check if the two variables present features of linearity that support the usage of liner regression analysis. The scatter plot would be used as the first preliminary result before applying machine learning models with R Studio packages [Schloerke et al. (2021)](Tang, Horikoshi, and Li 2016)[Kuhn and Wickham (2020)](Couch et al. 2021)(Xie 2022). There are three models applied in this research, in which model 1 is the classic linear regression model without any modification:

$model1 : \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$

$H_0 : \hat{\beta}_1 = 0$

$H_a : \hat{\beta}_1 \neq 0$

model 2 is a logarithmic y-variable-transformed linear regression model:

$model2 : \hat{ln(y)} = \hat{\beta}_0 + \hat{\beta}_1 x_1$

$H_0 : \hat{\beta}_1 = 0$

$H_a : \hat{\beta}_1 \neq 0$

model 3 is another logarithmic model with both x and y-variables-transformed:

$model3 : \hat{ln(y)} = \hat{\beta}_0 + \hat{\beta}_1 ln(x_1)$

$H_0 : \hat{\beta}_1 = 0$

$H_a : \hat{\beta}_1 \neq 0$

where x_1 is defined as the number of permits registered at a given state.

By applying feature (predictor) engineering to variables under the frame of regression analysis, this research would generate three models with different hypotheses and adjusted-r-squared values. On one hand, for each of the models, null hypothesis is given as non-association between explanatory and response variables, while alternative hypothesis is presence of association between the two variables.

In addition to the three linear regression models, one additional multivariate regression model is created based the best-fitted linear model from the three. Specifically, the model with the highest adjusted-r-squared value would be nuanced by narrowing the magnitude of the explanatory variable from the number of permits to the actual number of guns (handguns + long_guns) presented at each state. While using the permit variable broadens the scope of analysis, the replacement of the explanatory variable would enhance this research's precision to capture the association between the actual gun possession and prevalence of crimes, which is this research's ultimate purpose. Model 4 is the multivariate regression model without any modification:

$model4 : \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_2 + \hat{\beta}_2 x_3$

x_2 is defined as the number of handgun registered at a given state

x_3 is defined as the number of long-gun registered at a given state

$H_0 : \beta_1 = \beta_2 = 0$

$H_0 : \beta_1 \neq \beta_2 \neq 0$

For each of the fitted models, both independence of residuals and normality are analyzed with corresponding plots and summary statistics given. The four prediction sets are compiled together along the raw data set to conclude which of the prediction is the best fit of the true relationship between prevalence of firearms and occurrence of crimes across the United States at 2019.

3 Results

Before discussing the main results, it would be useful to understand the descriptive statistics of the variables. Based on figure 1, among the six variables of interest, the first two variables are categorical ('month' being ordinal and 'state' being nominal) and the rest are discrete numerical.

The two variables of interest are permit and total crime, thereby the two are chosen to generate a scatter-plot as shown in figure 2. Based on the the scatter-plot, there is a sign of moderate positive linearity that approves the feasibility of linear regression analysis.

```
     month              state               permit            handgun          long_gun       state_crime_2019_total
 Length:51           Length:51           Min.   :    0.0   Min.   :    0   Min.   :    0   Min.   :  10150
 Class :character    Class :character    1st Qu.:  851.5   1st Qu.: 3699   1st Qu.: 3878   1st Qu.:  41247
 Mode  :character    Mode  :character    Median : 3446.0   Median : 8815   Median :10957   Median : 103895
                                         Mean   : 6736.7   Mean   :15257   Mean   :12258   Mean   : 160217
                                         3rd Qu.: 9763.0   3rd Qu.:21011   3rd Qu.:16712   3rd Qu.: 197154
                                         Max.   :33683.0   Max.   :69732   Max.   :53386   Max.   :1095445
```

Figure 1: Descriptive Statistics
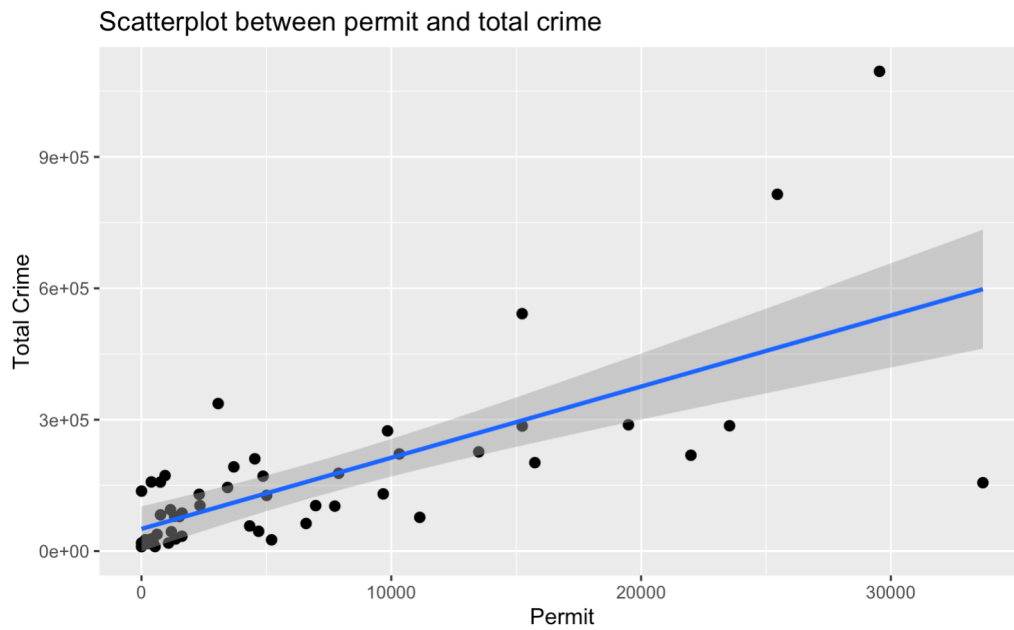
Figure 2: Scatter-plot (model1)

According to figure 3, the fitted p-value from model 1 is almost 0, which is smaller than the given alpha at 0.0125. Therefore, the null hypothesis is rejected and there is association between number of permits and crimes.

term <chr>	estimate <dbl>	std.error <dbl>	statistic <dbl>	p.value <dbl>
(Intercept)	50748.46940	25543.609051	1.986738	5.255800e-02
permit	16.24967	2.394261	6.786925	1.418753e-08

Figure 3: Summary statistics and formula of model 1 simple linear regression

$model1 : \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$

$H_0 : \hat{\beta}_1 = 0$

$H_a : \hat{\beta}_1 \neq 0$

Continuing the analysis, as demonstrated by figure 4, the residuals plot suggests there is a weak and subtle negative correlation between the residuals and fitted values, which barely meets the homoscedasticity requirement. On the other hand, the normal QQ plot suggests that most points align to the QQ line well. There are just a few points at the two ends that do not follow the regression line, which satisfies the normality check of the linear regression analysis.
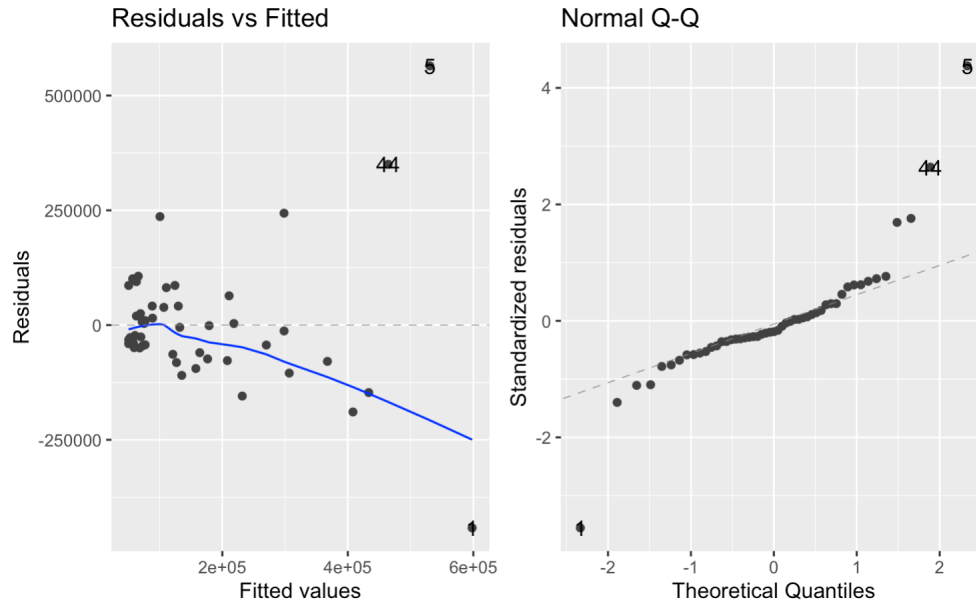

Figure 4: Residual independence and normality plots of simple linear regression model 1

After applying the logarithmic transformation to both explanatory and response variable of interest, the scatter-plot maintains features of moderate positive linearity as displayed by figure 5.
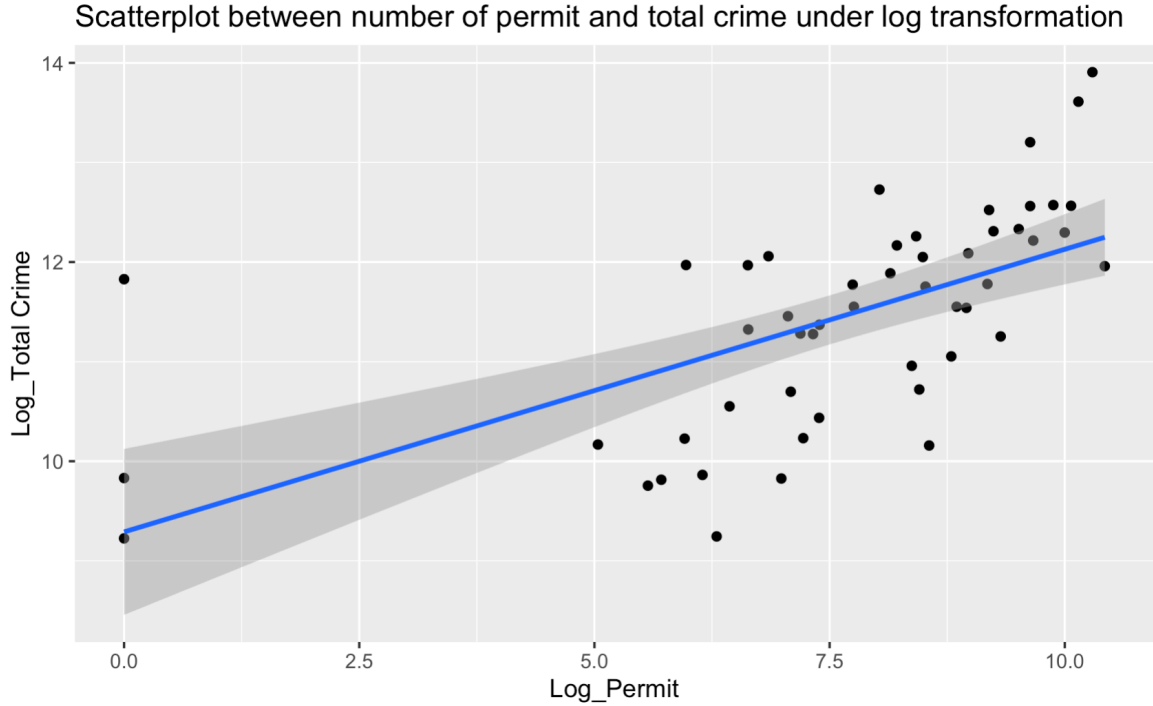

Figure 5: Scatter-plot (model 3)

However, under model 3, both the residuals plot and normal QQ plot have less desirable features in comparison to the plots generated under model 1. According to figure 6, there is a more obvious polynomial correlation between the residuals and fitted values, which does not meet the constant variance requirement. On the other hand, the normal QQ plot also demonstrate more points that fit less accordingly to the QQ line. Therefore, model 3 is generally a less accurate prediction in comparison to model 1 under the frame of linear regression analysis.

Among the three linear regression models, according to figure 7, 8, 9, model 1 has shown the highest adjusted-r_squared value (0.474). So that, model 1 is chosen as the base template to implement multivariate linear regression analysis. As described in the method section, model 4 is a multivariate linear regression that nuances the explanatory variable from number of permits to number of firearms, which is the sum of handgun and long-gun. Because model 4 is featured based on model 1, similar approach of analysis is conducted. The updated variables of interest are total number of guns and total crime, thereby the two are chosen to plot a scatter-plot as shown in figure 10. This scatter-plot indicates stronger positive linearity in comparison to that of model 1, which also approves the feasibility of linear regression analysis.
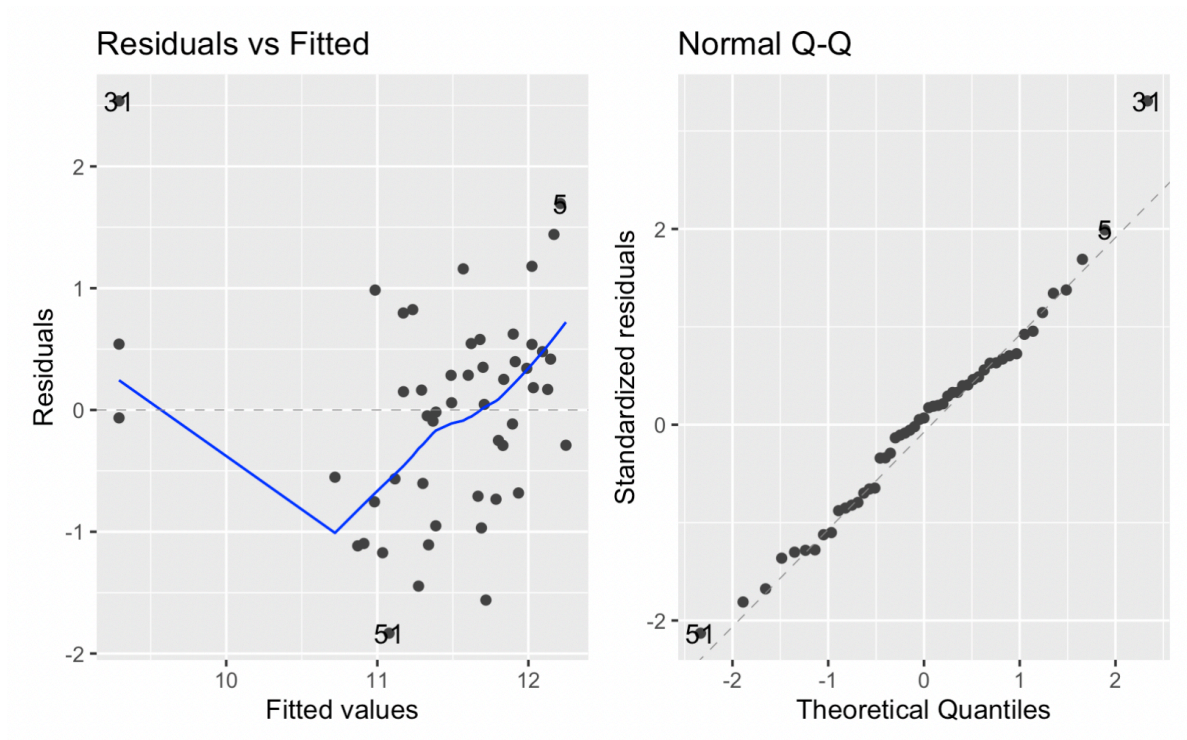
Figure 6: Residual independence and normality plots of simple linear regression model 3

r.squared <dbl>	adj.r.squared <dbl>	sigma <dbl>	statistic <dbl>	p.value <dbl>
0.4845…	0.4740294	14145…	46.06235	1.418753e−…

Figure 7: Summary statistics from simple linear regression model 1

r.squared <dbl>	adj.r.squared <dbl>	sigma <dbl>	statistic <dbl>	p.value <dbl>
0.4317…	0.4201405	0.8325…	37.22779	1.632581e−…

Figure 8: Summary statistics from simple linear regression model 2

r.squared <dbl>	adj.r.squared <dbl>	sigma <dbl>	statistic <dbl>	p.value <dbl>
0.376709	0.3639887	0.8719…	29.61496	1.672696e−…

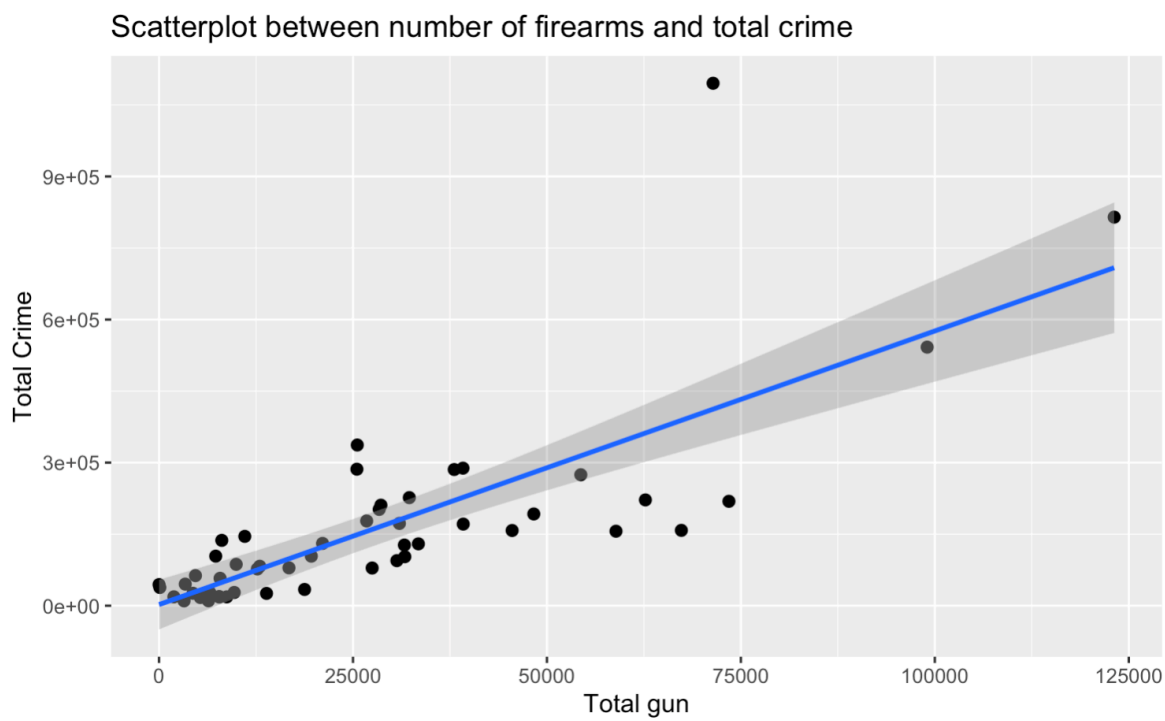Figure 9: Summary statistics from simple linear regression model 3

Figure 10: Scatter-plot (model 4)

Given the p-values of the fitted coefficients of the full model, and specifying the significance level at $\alpha = 0.0125$, when the two $\hat{\beta}_1$ are viewed separately, null hypotheses $H_0 : \beta_1 = 0$ can not be rejected because, as illustrated by figure 11, the given p-value for x_2 is 0.4645, which is much larger than the set alpha.

term <chr>	estimate <dbl>	std.error <dbl>	statistic <dbl>	p.value <dbl>
(Intercept)	−15149.261684	27315.206109	−0.5546091	0.581737239
handgun	1.760594	2.388064	0.7372473	0.464560689
long_gun	12.114525	3.737215	3.2415920	0.002163586

Figure 11: Summary statistics of model 4 multi-linear regression (separated variables)

r.squared <dbl>	adj.r.squared <dbl>	sigma <dbl>	statistic <dbl>	p.value <dbl>
0.6120499	0.5958853	123987	37.86363	1.350891e−10

Figure 12: Summary statistics and formula of model 4 multi-linear regression (combined variables)

$model4 : \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_2 + \hat{\beta}_2 x_3$

x_2 is defined as the number of handgun registered at a given state

x_3 is defined as the number of long-gun registered at a given state

where x_2 and x_3 are assumed to be positively correlated

On the other hand, when we treat the two $\hat{\beta}_1$ as one, the summary statistics from figure 12 suggests that we can reject the null hypothesis because the calculated p-value is almost 0, which is less than the set alpha. Superficially, the results appear contradictory, but they are actually not because the calculated p-value takes both x_2 and x_3 into consideration when fitting. The value change of either x_2 or x_3 would affect the other and have a collective effect on the result. On the other hand, separation of the two $\hat{\beta}$ only take consideration of one x-variable respectfully, which generates a less accurate prediction of the linear relationship as x_2 and x_3 are correlated, which means a change of either x_2 or x_3 would also affect the value of the other that have a collective effect on the result.

With respect to the residuals and normality checks, as demonstrated by figure 13, the residuals plot demonstrates the weakest correlation between the residuals and fitted values among the four models. In addition, the normal QQ plot also depicts that majority of points lie on the QQ line. Deviation from the QQ line is greatly reduced at both ends. Both plots best satisfy the conditions of the linear regression analysis, which also suggests that this model would potentially best fit the reality.
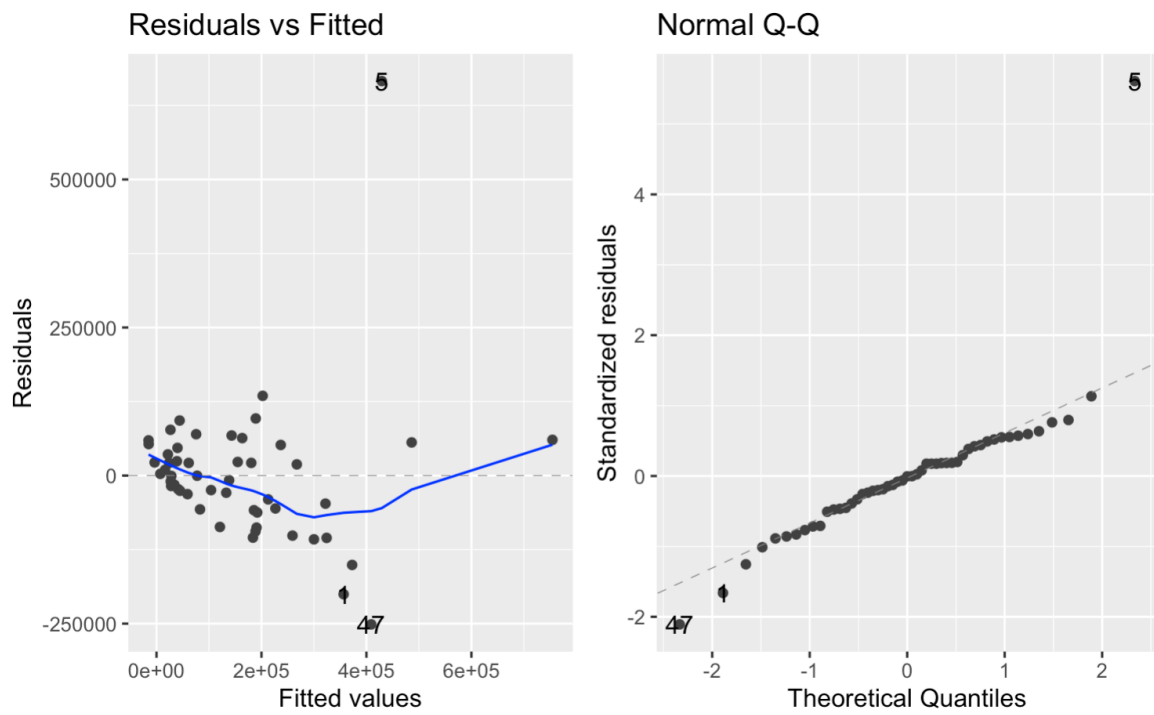
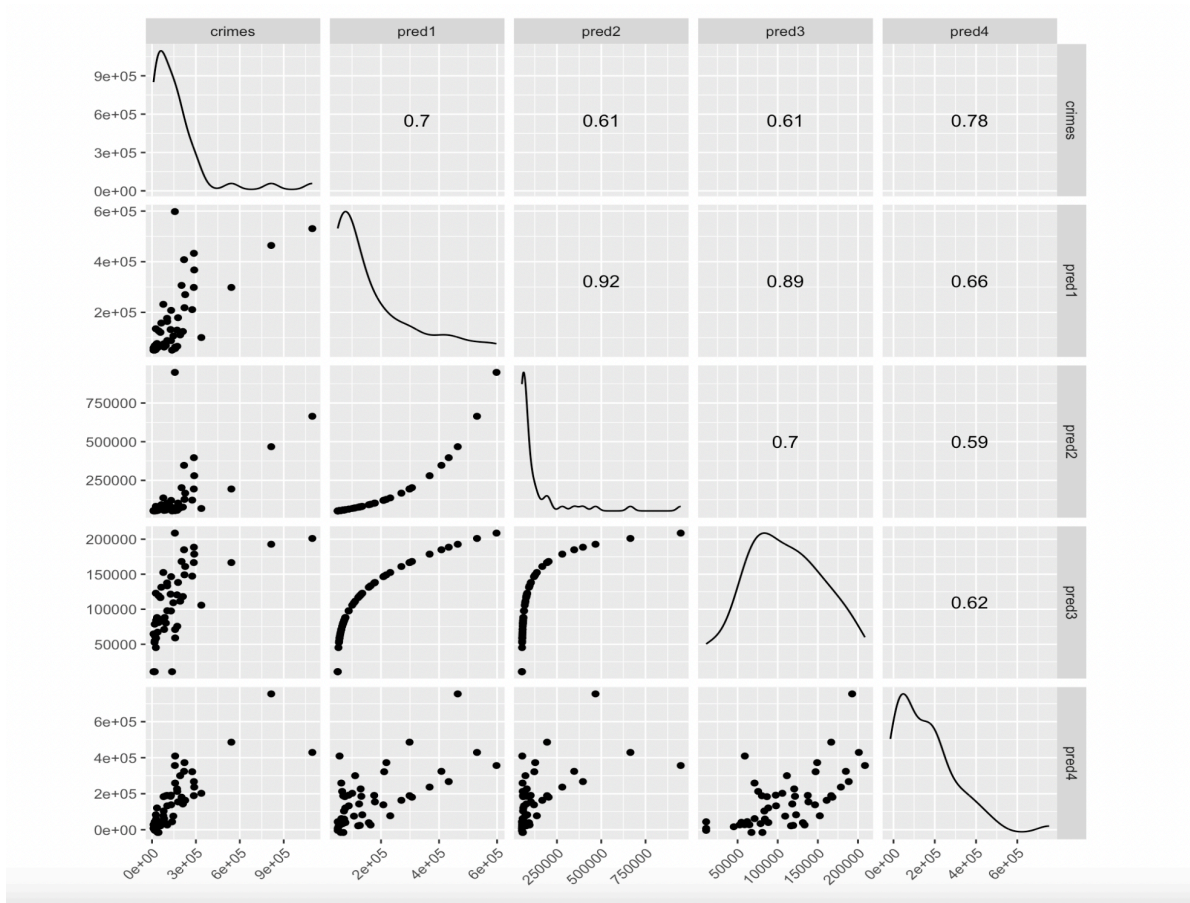Figure 13: Residual independence and normality plots of simple linear regression model 4

Figure 14: Summarized results based on 4 predictions of regression models

13

Summarizing all the results we have generated, figure 14 displays pairwise scatter-plots of the true number of crimes at each state versus all four predictions for the training set. Based on the pairwise plot, prediction 4 has the most accurate prediction because it has the highest correlation coefficient with the actual number of crimes. In addition, the qualitative representation of relationship between prediction 4 and number of crimes is closest to that of the actual graphic representation between the two 'crimes' variable. Lastly, because the scatter plot between volume and prediction 4 data best approximates a positive linear correlation, which also suggest prediction 4 to be the most accurate.

4 Discussion/Conclusions

In this project, we aim to assess the association between gun possession and the crime rates across U.S. in 2019. Based on the four models we created, we can conclude that there is a positive correlation between number of guns possessed and the total crime rate. Among all the models, model 4, which predicts the association between crime rate and the number of handguns and long guns combined, shows the strongest correlation with adjusted $r = 0.78$ and $r^2 = 0.60$. Meanwhile, model 1 which predicts the association between crime rate and the number of people who own a permit also shows a relatively strong results with adjusted $r = 0.70$ and $r^2 = 0.47$. Intuitively, people who have a permit doesn't necessarily imply that they already have possessed a gun. Hence, there might exist an over-prediction and it makes sense that model 4 will be a better predictor than model 1.

The gun ownership has always been a highly controversial topic, especially when it comes to the enactment of laws or constitutions. If the congress decided to make a new gun regulation law, and they look at the results of our project, the republicans, on the one hand, might say that r^2 of 0.60 is not large enough to be persuasive for passing a new law on gun-control. On the other hand, the democrats will claim that this is more than enough to be taken into consideration. Nevertheless, if discussions are only within the field of social science, r^2 of 0.60 is a considerably strong indicator.

Unfortunately, several limitations of our study may result in the inaccuracy of the outputs. Firstly, our measure does not provide information on any firearms which can be obtained in secondary market transfers or any illegal transactions. Moreover, the model used for fitting a linear regression might not be the most appropriate estimation. Instead, we can try to apply functions other than the natural logarithm transformation. It is clear that more studies and predictions are in need in the future to better approximate the association between firearms and criminal activities in reality.

14

5 Appendix

```r
library(tidyverse)
library(infer)
library(ggfortify)
library(tidymodels)
library(GGally)

#raw firearm data
remote = 'https://raw.githubusercontent.com/'
account = 'BuzzFeedNews/nics-firearm-background-checks/'
folder = 'master/data/'
file  = 'nics-firearm-background-checks.csv'
url = paste(remote, account, folder, file, sep = '')
firearm <- read.csv(url)

#raw crime data
remote = 'https://raw.githubusercontent.com/'
account = 'AndrewGou1013/'
file = 'State-crime-final/main/state_crime_final.csv'
url = paste(remote, account, file, sep = '')
state_crime <- read.csv(url)

df_Dec <-
  firearm %>%
  filter(month == '2019-12')

#deleted Guam, Virgin Islands, Marine Islands, Puerto Rico
new_df_Dec <- df_Dec[-c(12, 22, 42, 50), ]

new_df_Dec_month <-
new_df_Dec %>%
select(month)
as.factor('month')


[1] month
Levels: month

new_df_Dec_state <-
new_df_Dec %>%
select(state)
as.factor('state')
```

```
[1] state
Levels: state

new_df_Dec_permit <-
new_df_Dec %>%
select(permit)

new_df_Dec_handgun <-
new_df_Dec %>%
select(handgun)

new_df_Dec_long_gun <-
new_df_Dec %>%
select(long_gun)

state_crime_2019_final <-
  state_crime %>%
  filter(Year == 2019) %>%
  mutate(state_crime_2019_total = Data.Totals.Property.All + Data.Totals.Violent.All) %>%
  select(state_crime_2019_total)

df_firearm_final <-
  new_df_Dec_month %>%
  bind_cols(new_df_Dec_state) %>%
  bind_cols(new_df_Dec_permit) %>%
  bind_cols(new_df_Dec_handgun) %>%
  bind_cols(new_df_Dec_long_gun) %>%
  bind_cols(state_crime_2019_final)

Scatterplot_between_permit_and_total_Violence <-
  df_firearm_final %>%
  ggplot(aes(x = permit+1,
             y = state_crime_2019_total)) +
  geom_point(size = 2) +
  geom_smooth(method = 'lm', lwd = 1, se = TRUE) +
 xlab('Permit') +
 ylab('Total Crime') +
 ggtitle('Scatterplot between permit and total crime')
Scatterplot_between_permit_and_total_Violence
```
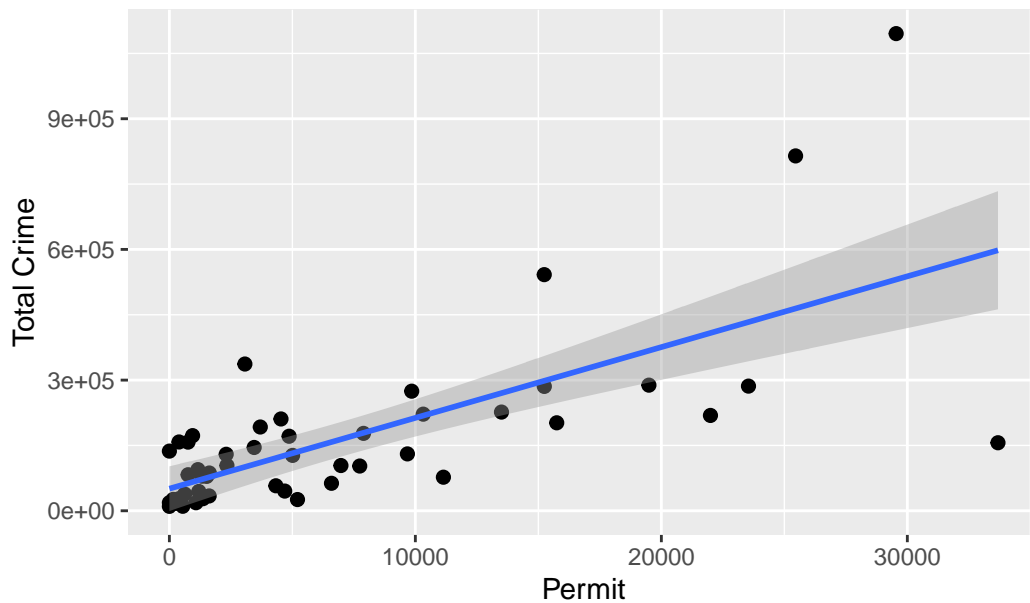
Scatterplot between permit and total crime


```
model_1 <-
  linear_reg() %>%
  set_engine('lm') %>%
  set_mode('regression') %>%
  fit(state_crime_2019_total ~ permit+1,
      data = df_firearm_final)

# tidy output
tidy(model_1)


# A tibble: 2 x 5
  term         estimate std.error statistic      p.value
  <chr>           <dbl>     <dbl>     <dbl>        <dbl>
1 (Intercept)  50748.    25544.        1.99 0.0526
2 permit          16.2       2.39      6.79 0.0000000142

# glance at metrics
glance(model_1)


# A tibble: 1 x 12
  r.squared adj.r.squa~1  sigma stati~2 p.value    df logLik   AIC   BIC devia~3
      <dbl>        <dbl>  <dbl>   <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>
1     0.485        0.474 1.41e5    46.1 1.42e-8     1  -676. 1358. 1364. 9.80e11
```
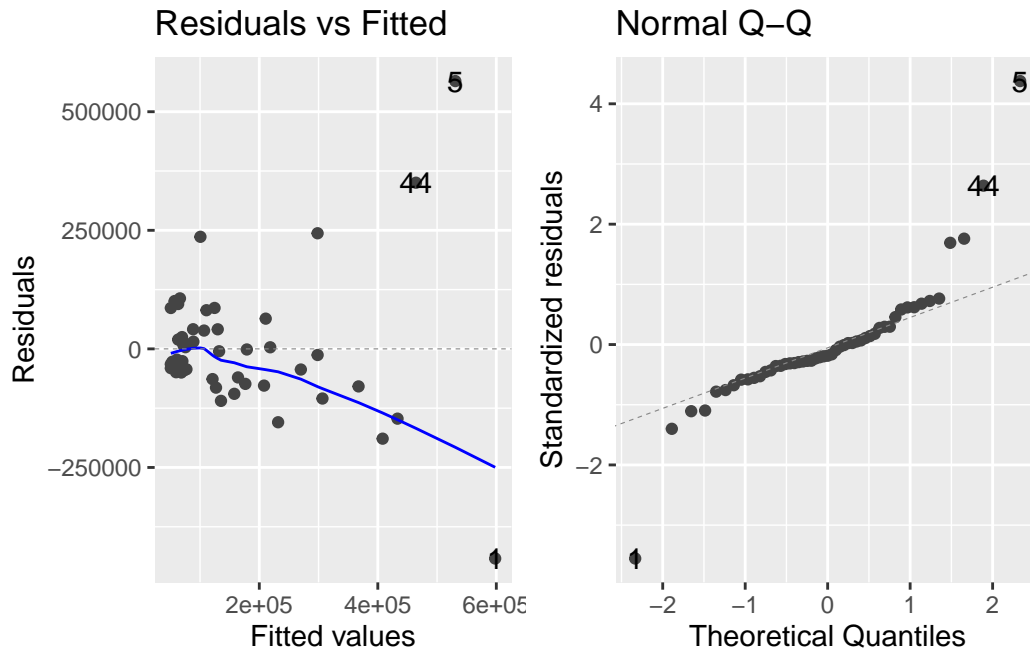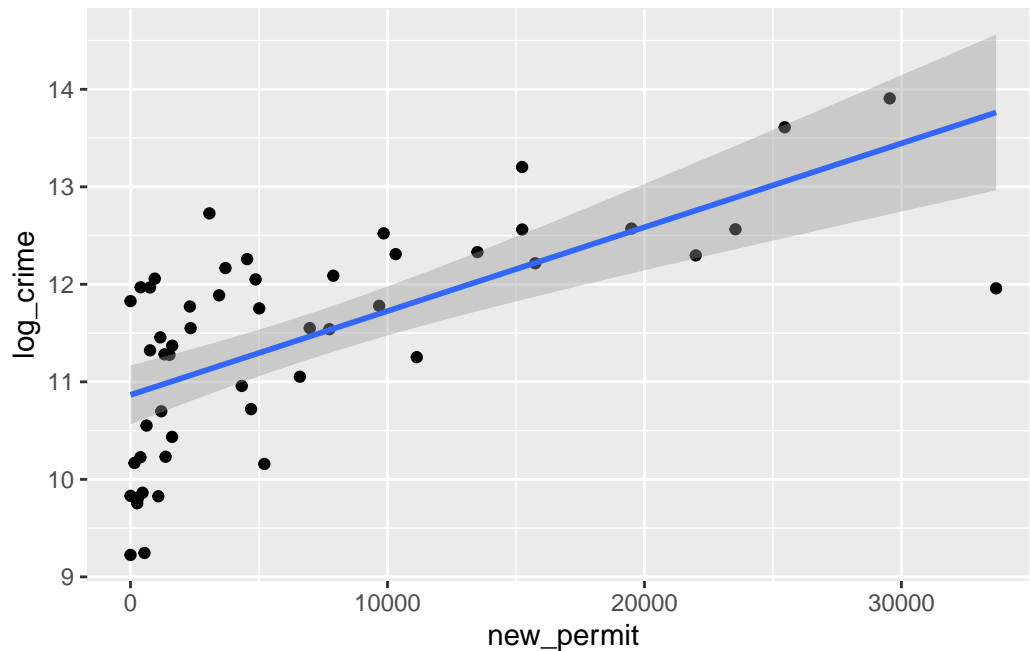
```
# ... with 2 more variables: df.residual <int>, nobs <int>, and abbreviated
#   variable names 1: adj.r.squared, 2: statistic, 3: deviance

# residual analysis and accuracy of the model
autoplot(model_1, c(1,2))
```



```
test_mod <-
  df_firearm_final %>%
  mutate(log_crime = log(state_crime_2019_total),
         new_permit = permit+1 )

test_mod %>%
  ggplot(aes(x = new_permit, y = log_crime)) +
  geom_point() +
  geom_smooth(method = 'lm')
```

```
model_2 <-
  linear_reg() %>%
  set_engine('lm') %>%
  set_mode('regression') %>%
  fit(log_crime ~ new_permit,
      data = test_mod)

# tidy output
tidy(model_2)
```

```
# A tibble: 2 x 5
  term            estimate  std.error statistic  p.value
  <chr>              <dbl>      <dbl>     <dbl>    <dbl>
1 (Intercept)   10.9        0.150         72.3  1.90e-51
2 new_permit     0.0000860  0.0000141      6.10 1.63e- 7
```
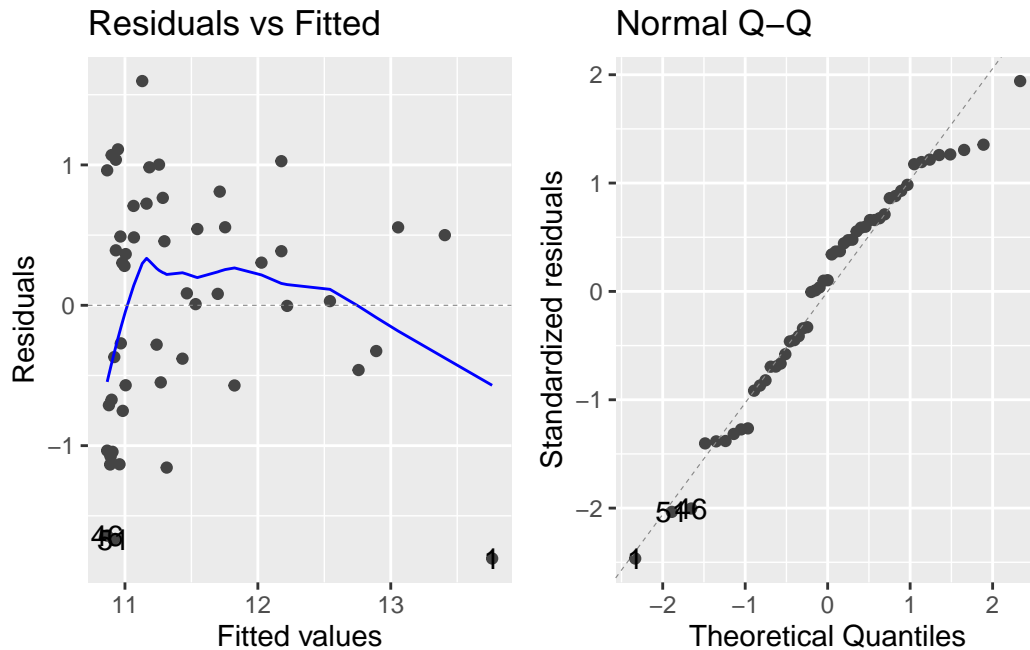
```
# glance at metrics
glance(model_2)
```

```
# A tibble: 1 x 12
  r.squ~1 adj.r~2 sigma stati~3 p.value    df logLik   AIC   BIC devia~4 df.re~5
    <dbl>   <dbl> <dbl>   <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>   <int>
1   0.432   0.420 0.833    37.2 1.63e-7     1  -62.0  130.  136.    34.0      49
```

```
# ... with 1 more variable: nobs <int>, and abbreviated variable names
#   1: r.squared, 2: adj.r.squared, 3: statistic, 4: deviance, 5: df.residual

# residual analysis and accuracy of the model
autoplot(model_2, c(1,2))
```
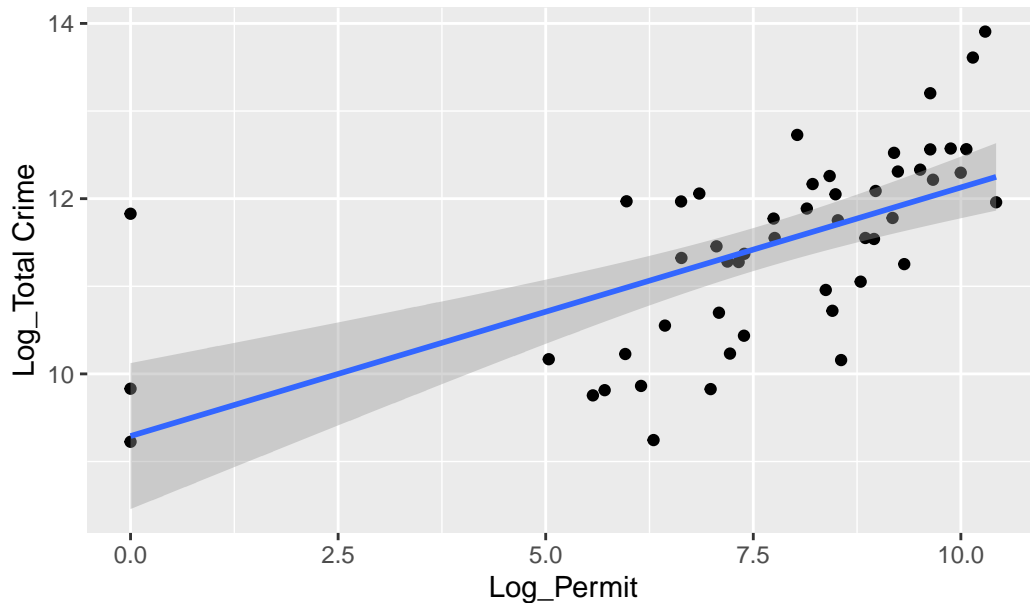


```
test_mod_again <-
  test_mod %>%
  mutate(log_permit = log(new_permit))

test_mod_again %>%
  ggplot(aes(x = log_permit, y = log_crime)) +
  geom_point() +
  geom_smooth(method = 'lm') +
  xlab('Log_Permit') +
  ylab('Log_Total Crime') +
  ggtitle('Scatterplot between number of permit and total crime under log transformation')
```

Scatterplot between number of permit and total crime under log


```r
# fit new model
model_3 <-
  linear_reg() %>%
  set_engine('lm') %>%
  set_mode('regression') %>%
  fit(log_crime ~ log_permit,
      data = test_mod_again)

# tidy output
tidy(model_3)
```

```
# A tibble: 2 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)    9.29      0.414      22.4  2.07e-27
2 log_permit     0.284     0.0521      5.44 1.67e- 6
```
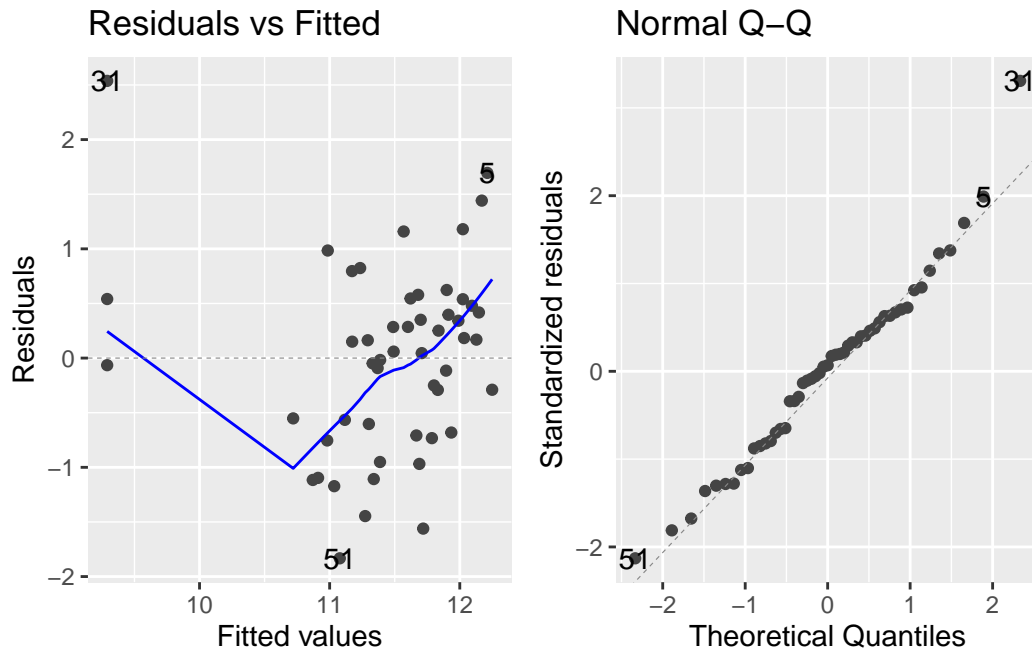
```r
# glance at metrics
glance(model_3)
```

```
# A tibble: 1 x 12
  r.squ~1 adj.r~2 sigma stati~3 p.value    df logLik   AIC   BIC devia~4 df.re~5
    <dbl>   <dbl> <dbl>   <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>   <int>
```

```
1   0.377    0.364 0.872    29.6 1.67e-6      1  -64.4  135.  141.    37.3      49
# ... with 1 more variable: nobs <int>, and abbreviated variable names
#   1: r.squared, 2: adj.r.squared, 3: statistic, 4: deviance, 5: df.residual

# residual analysis
autoplot(model_3, c(1,2))
```

Residuals vs Fitted   ## Normal Q–Q


```
glance(model_1)


# A tibble: 1 x 12
  r.squared adj.r.squa~1  sigma stati~2 p.value    df logLik   AIC   BIC devia~3
      <dbl>        <dbl>  <dbl>   <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>
1     0.485        0.474 1.41e5    46.1 1.42e-8     1  -676. 1358. 1364. 9.80e11
# ... with 2 more variables: df.residual <int>, nobs <int>, and abbreviated
#   variable names 1: adj.r.squared, 2: statistic, 3: deviance

glance(model_2)


# A tibble: 1 x 12
  r.squ~1 adj.r~2 sigma stati~3 p.value    df logLik   AIC   BIC devia~4 df.re~5
    <dbl>   <dbl> <dbl>   <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>   <int>
1   0.432   0.420 0.833    37.2 1.63e-7     1  -62.0  130.  136.    34.0      49
# ... with 1 more variable: nobs <int>, and abbreviated variable names
#   1: r.squared, 2: adj.r.squared, 3: statistic, 4: deviance, 5: df.residual
```

```r
glance(model_3)
```

```
# A tibble: 1 x 12
  r.squ~1 adj.r~2 sigma stati~3 p.value    df logLik   AIC   BIC devia~4 df.re~5
    <dbl>   <dbl> <dbl>   <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>   <int>
1   0.377   0.364 0.872    29.6 1.67e-6     1  -64.4  135.  141.    37.3      49
# ... with 1 more variable: nobs <int>, and abbreviated variable names
#   1: r.squared, 2: adj.r.squared, 3: statistic, 4: deviance, 5: df.residual
```

```r
Scatterplot_between_handgun_and_total_Violence <-
  df_firearm_final %>%
  ggplot(aes(x = handgun + long_gun,
             y = state_crime_2019_total)) +
  geom_point(size = 2) +
  geom_smooth(method = 'lm', lwd = 1, se = TRUE) +
 xlab('Total gun') +
 ylab('Total Crime') +
 ggtitle('Scatterplot between number of firearms and total crime')

model_4 <-
  linear_reg() %>%
  set_engine('lm') %>%
  set_mode('regression') %>%
  fit(state_crime_2019_total ~ handgun + long_gun,
      data = df_firearm_final) # fits the model

# tidy output
tidy(model_4)
```

```
# A tibble: 3 x 5
  term         estimate std.error statistic p.value
  <chr>           <dbl>     <dbl>     <dbl>   <dbl>
1 (Intercept)  -15149.    27315.     -0.555 0.582
2 handgun          1.76      2.39     0.737 0.465
3 long_gun        12.1       3.74     3.24  0.00216
```
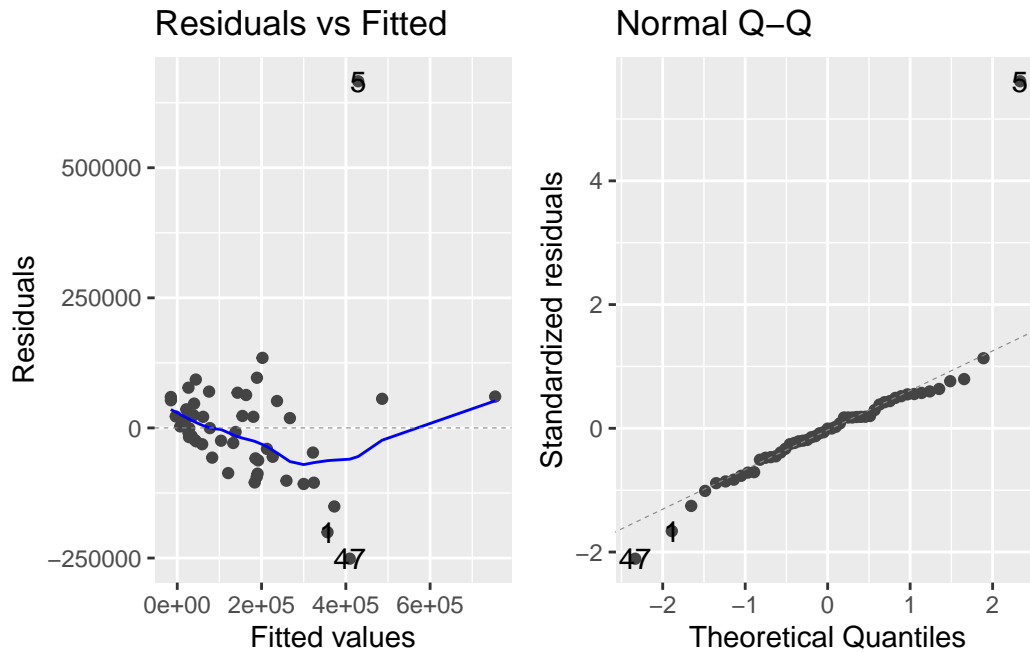
```r
# glance at metrics
glance(model_4)
```

```
# A tibble: 1 x 12
  r.squared adj.r.squ~1  sigma stati~2 p.value    df logLik   AIC   BIC devia~3
      <dbl>       <dbl>  <dbl>   <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>
```

```
1      0.612         0.596 1.24e5       37.9 1.35e-10      2  -669. 1346. 1354. 7.38e11
# ... with 2 more variables: df.residual <int>, nobs <int>, and abbreviated
#   variable names 1: adj.r.squared, 2: statistic, 3: deviance
```

```
# residual analysis
autoplot(model_4, c(1,2))
```



```
test_preds <-
  test_mod_again %>%
  select(-month, -handgun, -long_gun, -log_crime, -new_permit, -log_permit, -permit) %>%
  bind_cols(pred1 = predict(model_1, new_data = test_mod_again),
            pred2 = exp(predict(model_2, new_data = test_mod_again)),
            pred3 = exp(predict(model_3, new_data = test_mod_again)),
            pred4 = predict(model_4, new_data = test_mod_again))

# change column names
colnames(test_preds) = c('state', 'crimes', 'pred1', 'pred2', 'pred3', 'pred4')


# display the data
glimpse(test_preds)
```

```
Rows: 51
```

```
Columns: 6
$ state  <chr> "Alabama", "Alaska", "Arizona", "Arkansas", "California", "Colo~
$ crimes <int> 156179, 27637, 210779, 103893, 1095445, 171127, 57408, 26046, 3~
$ pred1  <dbl> 598086.07, 57020.84, 124456.97, 88756.45, 530828.69, 129835.61,~
$ pred2  <dbl> 948479.16, 54149.69, 77369.93, 64051.33, 664450.13, 79603.64, 7~
$ pred3  <dbl> 208684.01, 58757.65, 118148.12, 97910.44, 201062.75, 120532.70,~
$ pred4  <dbl> 356533.68, 28246.33, 143124.50, 132855.70, 429281.06, 226601.07~
```

```
test_preds %>%
  GGally::ggscatmat() +
  theme(axis.text.x = element_text(angle = 45, hjust =1))
```
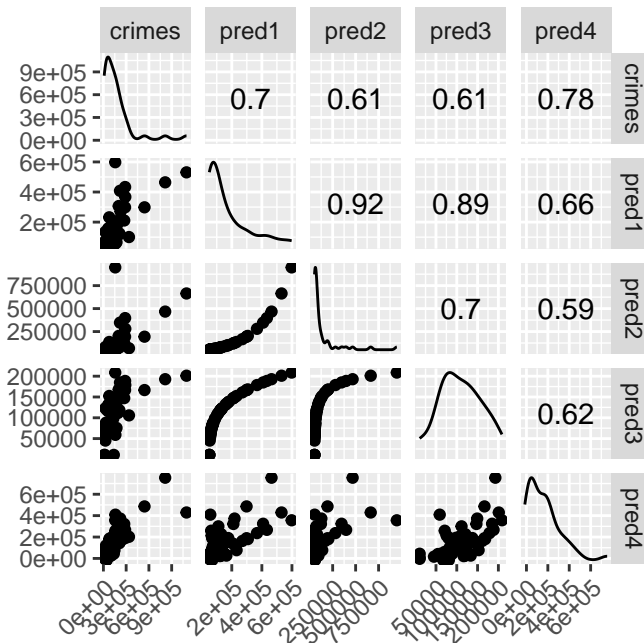

Bibliography

Branas, Charles C., Therese S. Richmond, Dennis P. Culhane, Thomas R. Ten Have, and Douglas J. Wiebe. 2009. "Investigating the Link Between Gun Possession and Gun Assault." *American Journal of Public Health* 99 (11): 2034–40. https://doi.org/10.2105/ajph.2008.143099.

C.Bart, Austin, Dennis Kafura, Clifford A. Shaffer, Javier Tibau, Luke Gusukuma, and Eli. Tilevich. 2022. "State Crime CSV File from the CORGIS Dataset Project." https://corgis-edu.github.io/corgis/csv/state_crime/.

Cook, Philip J. 1981. "The Effect of Gun Availability on Violent Crime Patterns." https://www.jstor.org/stable/1044071.

Couch, Simon P., Andrew P. Bray, Chester Ismay, Evgeni Chasnovski, Benjamin S. Baumer, and Mine Çetinkaya-Rundel. 2021. "{Infer}: An {r} Package for Tidyverse-Friendly Statistical Inference" 6: 3661. https://doi.org/10.21105/joss.03661.

Han, Daniel. 2022. "New Jersey Poised to Enact 'Nation's Strongest' Gun Law After Supreme Court Ruling," October. https://www.politico.com/news/2022/10/13/lawmakers-pitch-nations-strongest-gun-bill-in-response-to-supreme-court-00061693.

Kleck, Gary, and E. Britt Patterson. 1993. "The Impact of Gun Control and Gun Ownership Levels on Violence Rates." https://www.jstor.org/stable/23365752.

Kuhn, Max, and Hadley Wickham. 2020. "Tidymodels: A Collection of Packages for Modeling and Machine Learning Using Tidyverse Principles." https://www.tidymodels.org.

Miller, Matthew, David Hemenway, and Deborah Azrael. 2007. "State-Level Homicide Victimization Rates in the US in Relation to Survey Measures of Household Firearm Ownership, 2001–2003." *Social Science & Medicine* 64 (3): 656–64. https://doi.org/10.1016/j.socscimed.2006.09.024.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rozsa, Lori, Mark Berman, and Renae Merle. 2018. "Accused South Florida School Shooter Confessed to Rampage That Killed 17 People, Police Say," February. https://www.washingtonpost.com/news/post-nation/wp/2018/02/15/florida-school-shooting-suspect-booked-on-17-counts-of-murder-premeditated/.

Schloerke, Barret, Di Cook, Joseph Larmarange, Francois Briatte, Moritz Marbach, Edwin Thoen, Amos Elberg, and Jason Crowley. 2021. "GGally: Extension to 'Ggplot2'." https://CRAN.R-project.org/package=GGally.

"Second Amendment." 2002. https://doi.org/10.4135/9781483302799.n327.

Singer-Vine, Jeremy. 2019. "FBI NICS Firearm Background Check Data in 2019." https://github.com/BuzzFeedNews/nics-firearm-background-checks.

Smith, Sarah. 2022. "Texas Shooting: America's Gun Control Debate That Never Goes Away." https://www.bbc.com/news/world-us-canada-61585716.

Tang, Yuan, Masaaki Horikoshi, and Wenxuan Li. 2016. "Ggfortify: Unified Interface to Visualize Statistical Result of Popular r Packages" 8. https://doi.org/10.32614/RJ-2016-060.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the {Tidyverse}" 4: 1686. https://doi.org/10.21105/joss.01686.

Xie, Yihui. 2022. "Knitr: A General-Purpose Package for Dynamic Report Generation in r." https://yihui.org/knitr/.