Modeling Crime in Los Angeles

Julia McNerney, Yue Wan and Xukan Zhang
University of Notre Dame
Notre Dame, IN 46556, USA

1. Background and Motivation

Los Angeles is consistently ranked as having some of the highest crime rates among major U.S. cities. According to crime data collected by the FBI, Los Angeles's crime rate ranks high among cities with populations greater than one million (FBI, 2019). As the second largest city in the U.S., having such a high crime rate puts more people at risk compared to smaller cities. Thus, analyzing crime in Los Angeles has been the subject of many criminological and statistical studies.

Most studies that model crime in Los Angeles focus on modeling gang-related crime. Using self-exciting point processes and temporal graph analysis, researchers have been able to accurately represent gang violence in Los Angeles (Mike Egesdal, 2010). Other studies have researched risk factors associated with gang violence, but have been unsuccessful in forecasting gang violence (Valasik, 2018). Few studies have explored Los Angeles crime more generally, and none have focused on forecasting.

This paper aims to examine how crime trends differ over time for different groups of victims. As the most diverse city in the U.S. according to CNBC, Los Angeles offers a unique setting to investigate how different groups of people are affected by crime over time (Wells, 2015). The results can help inform policy makers in creating more appropriate measures to fight crime as well as give law enforcement a deeper understanding of who is most affected by crime in the city. In addition, with most research about crime in Los Angeles being focused on gang violence, looking instead at the victims of crime will help the general public be more aware of what threats they face and when.

2. Dataset Selection and Analysis

We download the LA crime dataset from (Kaggle, 2023). The data used in this analysis consists of all reported crimes in Los Angeles from January 1, 2020 through October 8, 2023, downloaded from kaggle.com. There are over 700,000 observations and numerous features that describe each reported crime. The selected features used in this analysis are Crime Type, Victim Sex and Descent.

3. Monthly Analysis

We first look at monthly data, hope to get a general overview of the crime count trend. We implement data from Jan 2020 to Sep 2023, which is 45 data points in total.

3.1 Modeling Overall Monthly Crime Count

From the monthly crime count time series plot, we can see fluctuation through the four years, with a noticeable decrease starting at March 2020,which is exactly the Covid pandemic breakout time in California (Mohler et al., 2020), and the number of crimes began to increase and back to normally high rates at around mid 2022, which is close to the date where Joe Biden announced the end of pandemic (Archie, 2022).

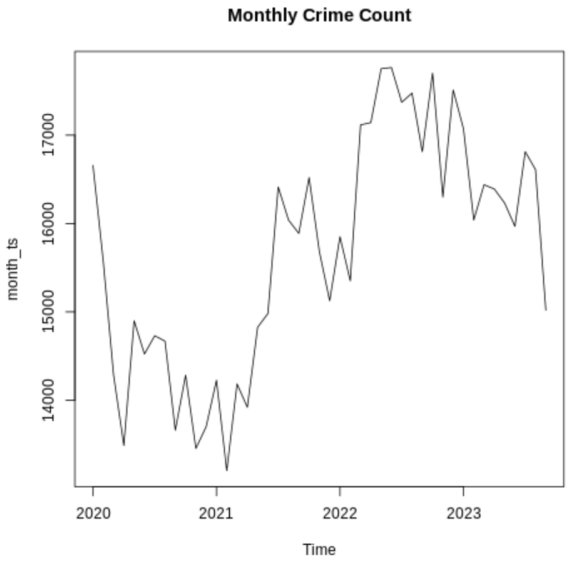


Figure 1: Time series plots of overall monthly crime.

We fit ARIMA(1,1,0) model for the overall monthly crime counts.The detailed ACF-PACF plots and diagnostic plots are in Appendix A.1. We do first order differencing to attain stationary for the data, and the AR(1) component suggests a potential short-term autocorrelation in the data. Thus, we dive into victim's profile and segment the data by victim's gender.

3.2 Modeling Monthly Crime by Victim Gender

The first analysis about the characteristics of crime victims we performed was segmented by the victim's gender based on monthly crime counts. The time series plots of each gender was first inspected for any noticeable differences between the two groups. The time series plots of the data have similar overall trends, but enough differences to warrant further analysis. Both time series were differenced to remove these trends. The ACF and PACF plots of both groups suggested an AR(1) model to be appropriate for the data because of the tapering behavior of the ACF plots and the PACF plots cutting off at lag 1 (Appendix B). Fitting this model to both groups produced adequate diagnostic plots with all significant coefficients.

While ARIMA(1,1,0) was selected for both male and female victims, it seems that this model fits male victims better, as seen by the lower p values of the Ljung-Box statistic and one significant lag in the ACF of residuals in the female diagnostic plots. These results

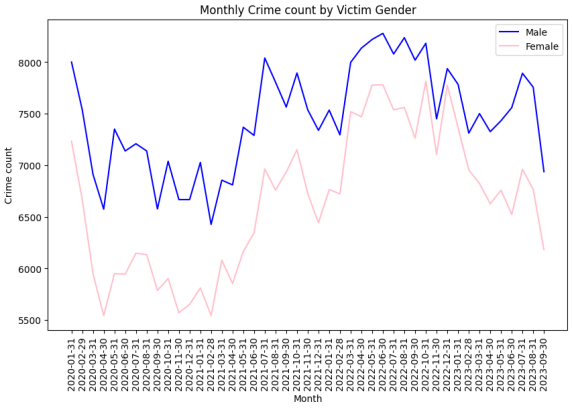


Figure 2: Time series plots of male and female victims.

indicate that a more complex model may be needed to more accurately model crime against female victims while a simple ARIMA(1,1,0) model is adequate for modeling crimes with male victims. However, to obtain a general, simple model of crimes against female victims, it appears that an ARIMA(1,1,0) model can work.

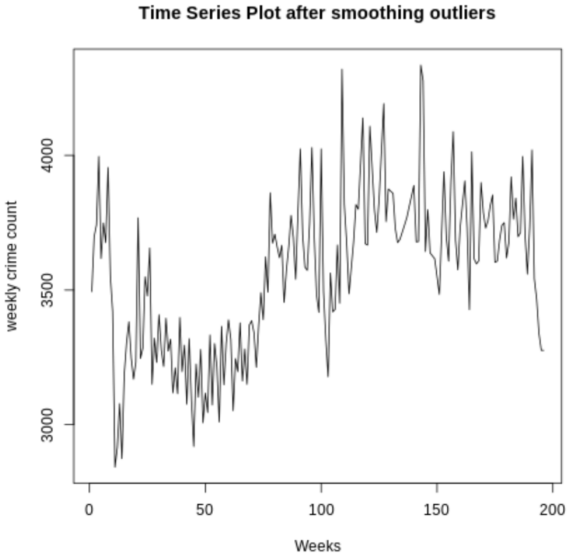
4. Weekly Analysis

Although we have fitted several models in Section 3 that can predict future crime counts, the models fitted all belong to the same model type ARIMA(1,1,0). This means that whether we focus on overall crime count or gender-specific crime count, the predictions will show similar trends and we cannot provide more informative interpretations of how victim gender may affect crime. The limited data for monthly analysis (45 months) may account for this similarity in the three models fitted in Section 3.

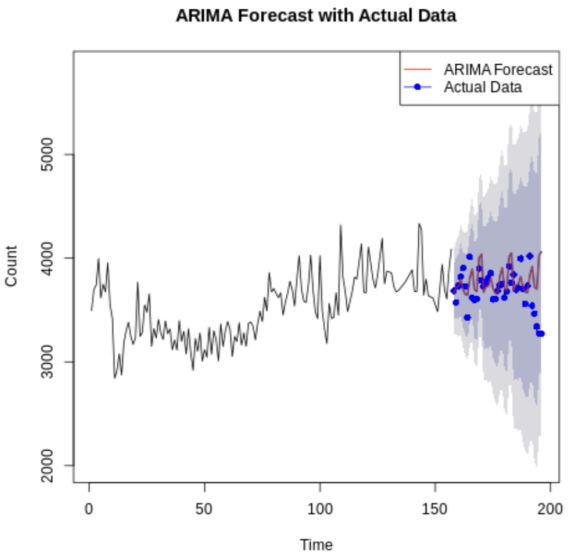
Therefore, we grouped the daily data into weekly data for more data points (196 weeks). We then split the weekly data into a training dataset and testing dataset, with 80 % of data (157 weeks) as the training dataset and 20 % of data (39 weeks) as the testing dataset. After fitting our models, the predictions for the next 39 weeks will be generated and compared with the testing dataset to decide whether the fitted models generate satisfying predictions.

4.1 Modeling Overall Weekly Crime Count

We first smoothed the data to deal with outliers. We defined outliers as data that satisfy the formula $|week_{ts} - mean(week_{ts})| > 2 * std(week_{ts})$, and we used linear interpolation to smooth the outliers. Figure 3a shows the smoothed time series plot of overall weekly crime count.



(a) Time series plot of overall weekly crime



(b) 39 week prediction of Overall Weekly crimes

Appendix 9a is the non-seasonal differenced ACF and PACF plots. We spotted seasonality, but the interval between spikes was not even. We observed spikes at lag 4, 9, 13,17,22,26..., with the interval following a 4,5,4 seasonality. We take the sum of 4,5,4 as the seasonality and continue to fit and diagnose our seasonal ARIMA model (Appendix 9b and 9c). The final model fitted is $ARIMA(0,1,1) \times (0,1,1)_{13}$. Then we used the last 39 weeks of data to test our model's performance. Figure 3b suggests that our model can correctly predict the trend of the future data.

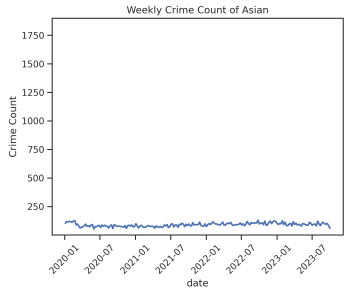
4.2 Modeling Weekly Crime by Victim Descent

The model fitted in Section 4.1 shows a seasonal pattern with a period of 13 weeks for the overall crime count. To analyze if and how this seasonality is affected by victim characteristics, we separated our weekly crime count data set according to victim descent and built separate models for each crime descent.

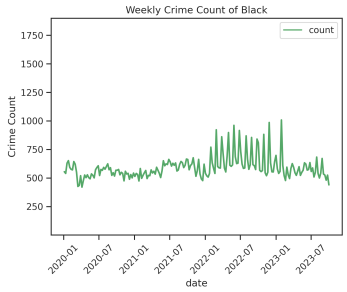
We focused on four different victims descents: A (Asian), B (Black), H (Hispanic), and W (White), which accounted for almost 80% of the entire data set.

Figure 4 are the time series plots of weekly crime count for each of the four victim descents. The four plots displayed obvious differences in average crime counts as well as seasonality. Hispanic victims have the highest average crime count and Asian victims have the lowest average crime count. Asian victim crime count does not show clear patterns of seasoanlity. The other three descents show possible seasonality, which needs to be further examined through ACF & PACF plots and the diagonostic plots of fitted models.

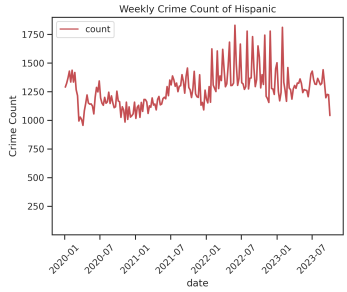
For Asian victims, the ACF & PACF plots of the differenced data show no pattern of seasonality and suggested an MA(1) model since its ACF plot cuts off after lag 1 and PACF plot tails off (Appendix C.1). The final model fitted is ARIMA(1,1,0), which suggests that the future predictions will display a constant trend, which overlaps with the observed data well (Figure 5a).



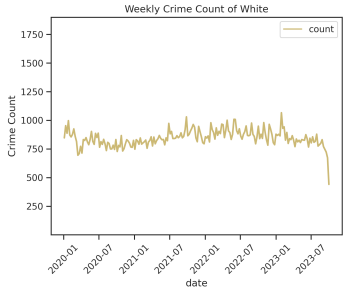
(a) Time series plot of Asian victims



(b) Time series plot of Black victims



(c) Time series plot of Hispanic victims



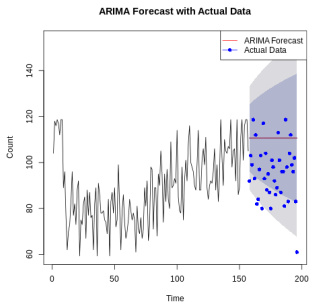
(d) Time series plot of White victims

Figure 4: Time series plot of weekly crime count by victim descent.

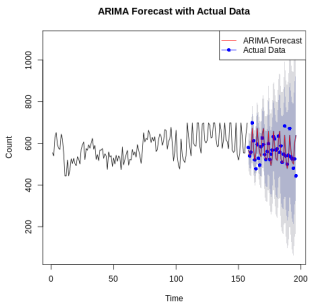
For Black victims, the ACF plot of the differenced data shows slow decay at seasonal lags with seasonality S=13, suggesting a seasonality with S=13. After taking first difference and seasonal difference, the ACF & PACF plots suggested a nonseasonal MA(1) model since its ACF plot cuts off after lag 1 and PACF plot tails off (Appendix C.2). The seasonal components cannot be determined by the ACF & PACF plots, thus we experimented with different values of the seasonal parameters and choose the model that satisfies the time series assumption. The final model fitted is ARIMA$(1,1,0) \times (1,1,0)_{13}$. The future predictions overlap with the observed data well (Figure 5b), displaying a decreasing trend with seasonality.

For Hispanic victims, the ACF plot of the differenced data shows slow decay at seasonal lags with seasonality S=4, suggesting a seasonality with S=4. After taking first difference and seasonal difference, both the nonseasonal and seasonal components cannot be determined by the ACF & PACF plots (Appendix C.3), thus we experimented with different values of the seasonal parameters and choosed the model that satisfies time series assumption. The final model fitted is ARIMA$(2,1,1) \times (1,1,0)_4$. However, the future predictions of this model does not overlap well with the observed data (Figure 5c). The predictions show a constant trend with seasonality, while the seasonality in the observed data is not obvious and the predictions have higher average value than the observed. This discrepancy between the predicted and observed corresponds to the problems of the diagnostic plots of the fitted model (Appendix C.3), suggesting the need for a different type of model that can better fit this data.

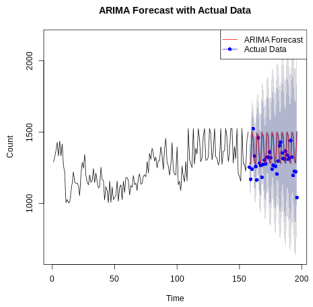
For White victims, the ACF plot of the differenced data shows slow decay at seasonal lags with seasonality S=13, suggesting a seasonality with S=13. After taking first difference and seasonal difference, both the nonseasonal and seasonal components cannot be determined by the ACF & PACF plots (Appendix C.4), thus we experimented with different values of the seasonal parameters and choosed the model that satisfies time series assumption. The final model fitted is ARIMA$(0,1,2) \times (2,1,0)_{13}$. The future predictions overlap with the observed data well (Figure 5d), displaying an decreasing trend with seasonality.



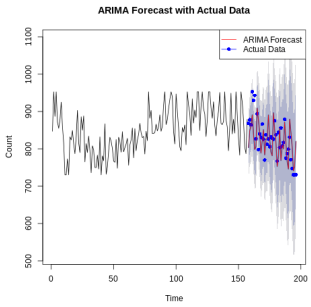
(a) 39 week prediction of Asian victims



(b) 39 week prediction of Black victims



(c) 39 week prediction of Hispanic victims



(d) 39 week prediction of White victims

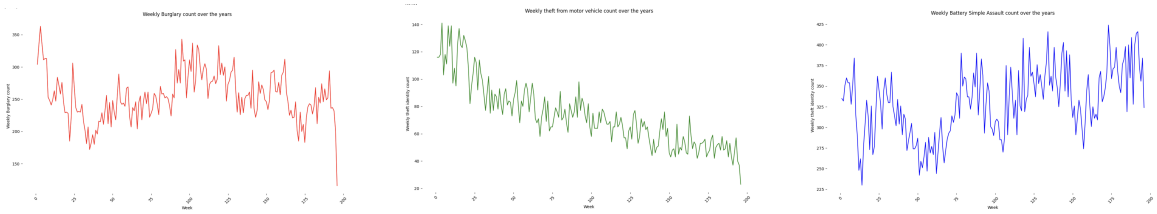
Figure 5: 39 week prediction weekly crime count by victim descent.

Comparing the results between the four different victim descents, we concluded that victim descents is a significant predictor when predicting future crime. Despite their difference in the magnitude of crime counts (with Hispanic having the largest and Asian having the smallest), different victim descents also have different future crime patterns. Black and White victims may be exposed to less future crimes, while Asian and Hispanic may have equal possibility to be affected by crime in the future when compared to present. The clear seasonality for Black, Hispanic, and White victims also shows that despite the general trends, crime counts fluctuates on a seasonal basis, with period of 13 weeks for Black and White victims and 4 weeks for Hispanic victims.

4.3 Modeling Weekly Crime by Crime Type

There are 210 crime types in total. We selected battery simple assault, burglary, and theft from motor vehicle as our subjects of analysis, each composes of 8%, 6.2% and 1.8% among overall crime. They are chosen because of their high crime rate and commonality. According to Statista (Statista, 2022), these crimes are the ones that people worry the most in their everyday city life, and they are closely related to personal safety and property security.

Figure 6 shows the trend of weekly crime count versus different crime type. Note that even though there is an observable seasonality trend in the time series plot of theft from motor vehicle and battery simple assault, we failed to find a proper seasonality lag that could make the diagnostic plots look good. Therefore, we fit crime types to a non-seasonal ARIMA model and find that all three of them follow the MA model with first order differencing.



(a) Weekly Burglary Crime Count

(b) Weekly Theft from Motor Vehicle Crime Count

(c) Weekly Battery Simple Assault

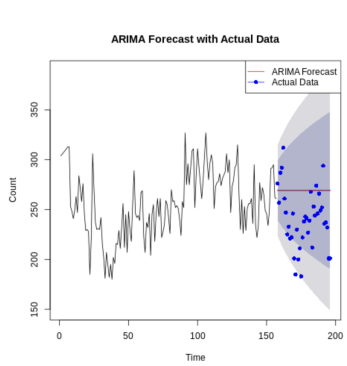
Figure 6: Weekly Crime Count by Crime Type

We fit Burglary and Battery Simple Assault crime to an ARIMA(0,1,1) model, and fit Theft from Motor Vehicles to an ARIMA(0,1,2) model. Detailed ACF and PACF plots, diagnostic plots and analysis are in Appendix D. Figure 7 shows the 39 week prediction of crime count versus different crime type. Since each crime type is a differenced MA model, the predicted values are a constant trend. For theft from motor vehicle crime, the predicted value is in the middle of the actual data, indicating the accuracy of fitted model. For each crime type, the first predicted data is close to the actual data, which suggests a short term dependency between the predicted data and it's closest one MA(1) or two MA(2) white noise.

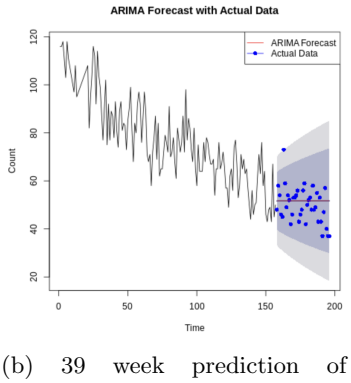
Note that no AR model is observed, which reveals that the current data has a lack of long term dependency on the previous data. For theft from motor vehicle with ARIMA(0,1,2) model, the difference in the MA order indicates a larger autocorrelation within the white noise. A more comprehensive model should be dived into to explore the trend and inter-dependency of this crime type, since it also shows a consistent decreasing and seasonal-like trend.

4.4 Hidden Seasonality behind crime types

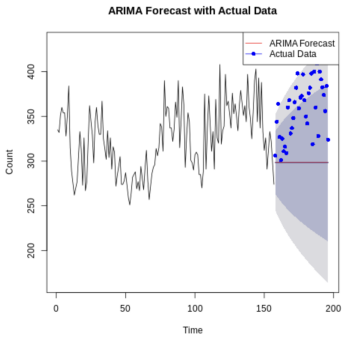
Although we fail to fit proper seasonality data from weekly crime count by type, we can still observe trend and seasonality within each crime type. Figure 6c shows that during the middle to late middle of a year, the crime count each week increases and reach the highest point. Interestingly, a report (Johnson, 2023) found violent crimes (like murder, rape and aggravated assault) are more likely to happen during the summer than any other season and



(a) 39 week prediction of Burglary ARIMA(0,1,1)



(b) 39 week prediction of Theft from Motor Vehicle ARIMA(0,1,2)



(c) 39 week prediction of Battery Simple Assault ARIMA(0,1,1)

Figure 7: 39 Week prediction by Crime Type

it is long believed that heat increases aggression. This can explain the high crime rate each summer regarding battery simple assault. We can also observe a decreasing trend for theft from motor vehicle. According to Los Angeles Almanac (Almanac, 2022), the motorcycle registration data in Los Angeles County shows a decreasing trend from 2020 to 2022, which reflects the decrease in crime rates.

5. Summary

This report analyzes crime trends in Los Angeles, focusing on how they differ for different groups of victims. The data used covers all reported crimes from 2020 to 2023, categorized by victim sex, descent, and crime type. Both monthly and weekly analyses were conducted using ARIMA models. The monthly analysis revealed that the overall crime count decreases during the pandemic and returns to higher rates in mid-2022. In addition, separate models were fit on crimes on male and female victims, but showed little difference in both groups' relationship with time because an ARIMA(1,1,0) model was adequate for both. Analyzing trends at the weekly level revealed seasonal patterns for both overall crime count and for some specific types of crime. This analysis found that victim descent significantly affects future crime predictions, with different patterns for each group being found. Lastly, crime types like battery, burglary, and theft from motor vehicle showed consistent trends and potential seasonality at the weekly level.

These findings can inform policy makers in developing targeted crime prevention strategies for different victim groups. For example, strategies aimed at preventing crime against men or women may be effective in preventing crime overall because of the little difference found between the two groups. However, the differences found between the different victims' descents could be used to better predict spikes in crimes against a specific race, allowing law enforcement to allocate resources more efficiently. Further research is needed to explore the hidden seasonality observed in some crime types and to develop more comprehensive models. Overall, this report provides valuable insights into the complex dynamics of crime in Los Angeles and highlights the importance of considering victim characteristics when analyzing and predicting crime trends.

6. Contributions

- Julia: Wrote Background/Motivation, Dataset Selection & Analysis, and Modeling Crime by Victim Gender sections. Fitted Models. Edited final draft.
- Elena: Weekly Crime Count by Victim Descent
- Yue: Overall Crime Count monthly and weekly, Crime Type

References

Los Angeles Almanac. Motor vehicle registrations los angeles county. *DMV Statistics*, 2022.

Ayana Archie. Joe biden says the covid-19 pandemic is over. this is what the data tells us. *NPR*, 2022.

FBI. Offenses known to law enforcement, by state by city, 2019. *FBI:UCR*, 2019.

Arianna Johnson. Here's why warm weather causes more violent crimes—from mass shootings to aggravated assault. *Forbes*, 2023.

Kaggle. Crime data from 2020 to present. *Kaggle*, 2023.

Kym Louie Jeremy Neuman Mike Egesdal, Chris Fathauer. Statistical modeling of gang violence in los angeles. *SIAM Undergrad. Res*, 2010.

George Mohler, Andrea L. Bertozzi, Jeremy Carter, Martin B. Short, Daniel Sledge, George E. Tita, Craig D. Uchida, and P. Jeffrey Brantingham. Impact of social distancing during covid-19 pandemic on crime in los angeles and indianapolis. *Journal of Criminal Justice*, 68:101692, 2020. ISSN 0047-2352. doi: https://doi.org/10.1016/j.jcrimjus.2020.101692. URL https://www.sciencedirect.com/science/article/pii/S0047235220301860.

Statista. Offenses known to law enforcement, by state by city, 2019. *Statista*, 2022.

Matthew Valasik. Gang violence predictability: Using risk terrain modeling to study gang homicides and gang assaults in east los angeles. *Journal of Criminal Justice*, 58:10–21, 2018.

Jane Wells. American diversity: Cities where it works. *CNBC*, 2015.

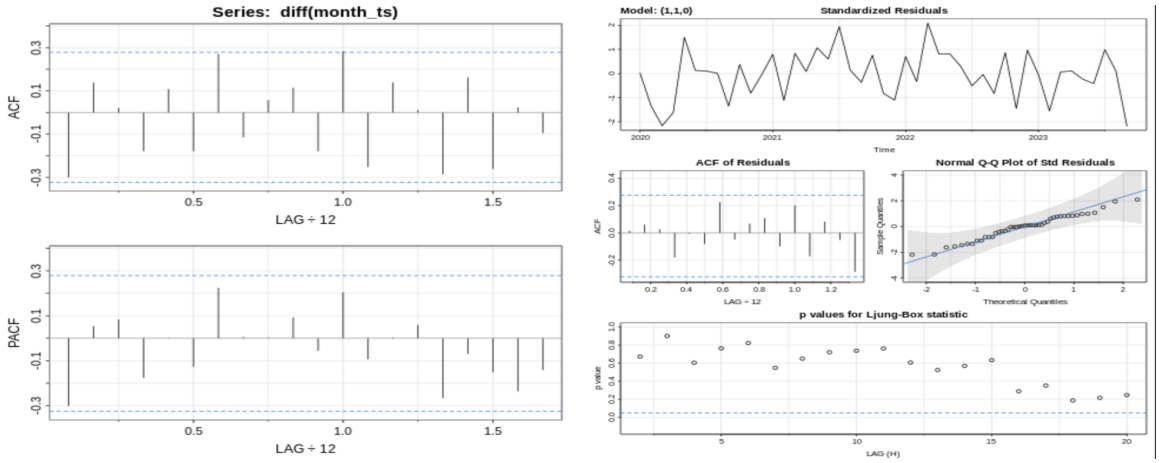
Appendices

A. Overall Crime Count Monthly and Weekly

A.1 Monthly Crimes

Figure 8a is the ACF and PACF plots of the overall monthly crime count after taking first difference. Suggesting a ARIMA(1,1,0) model for the orginal data.

All diagnostic plots in Figure 8b shows no violation of the assumption of time series models, suggesting that ARIMA(1,1,0) is a suitable model.



(a) ACF & PACF plot of the differenced data

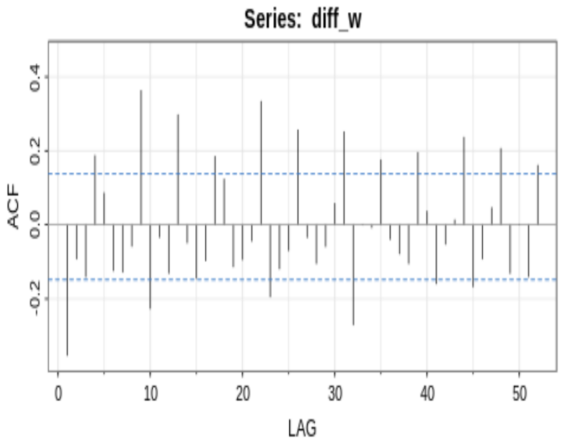
(b) Diagostic plots of the ARIMA(1,1,0) model.

Figure 8

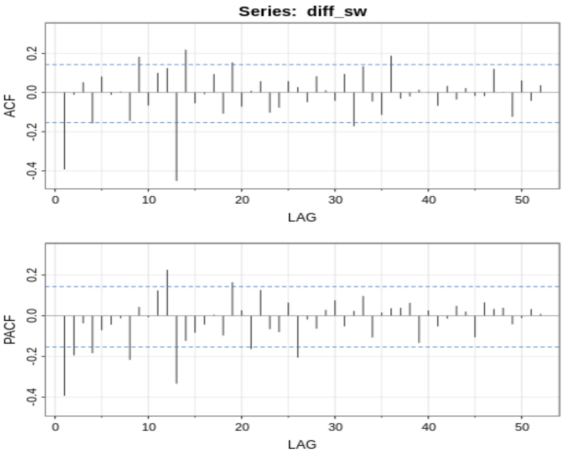
A.2 Weekly Crimes

Figure 9a is the ACF and PACF plots of the overall weekly crime count after taking non-seasonal first difference. Figure 9b is the ACF and PACF plots of the overall weekly crime count after taking first difference with lag 13. For the seasonal pattern, we observed a tails off pattern for lags at 13, 26..., and a cut off after lag 13 in ACF plot, indicating a seasonal ARIMA$(0,1,1)_{13}$. The non-seaonal part shows a similar pattern, with ACF plot cuts off after first lag and PACF plot tails off at the first few lags, which shows a non-seasonal ARIMA(0,1,1). Combining them, we get ARIMA$(0,1,1) \times (0,1,1)_{13}$ model.

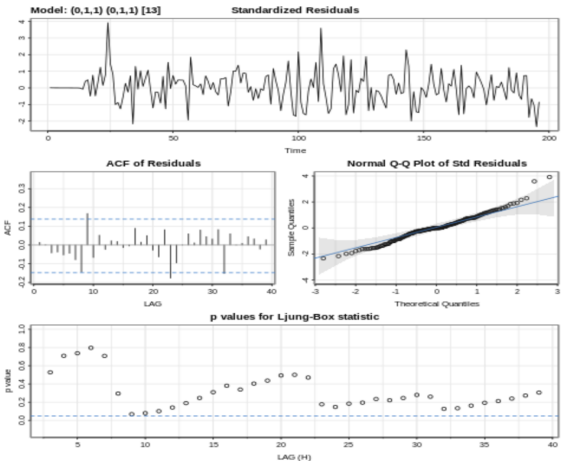
All diagnostic plots in Figure 9c shows no violation of the assumption of time series models, although two lags or the ACF residuals are significant for our fitted model, all the p-values of the Ljung-Box statistic are significant.



(a) ACF & PACF plot of the non-seasonal differenced data



(b) ACF & PACF plot of the seasonal differenced data

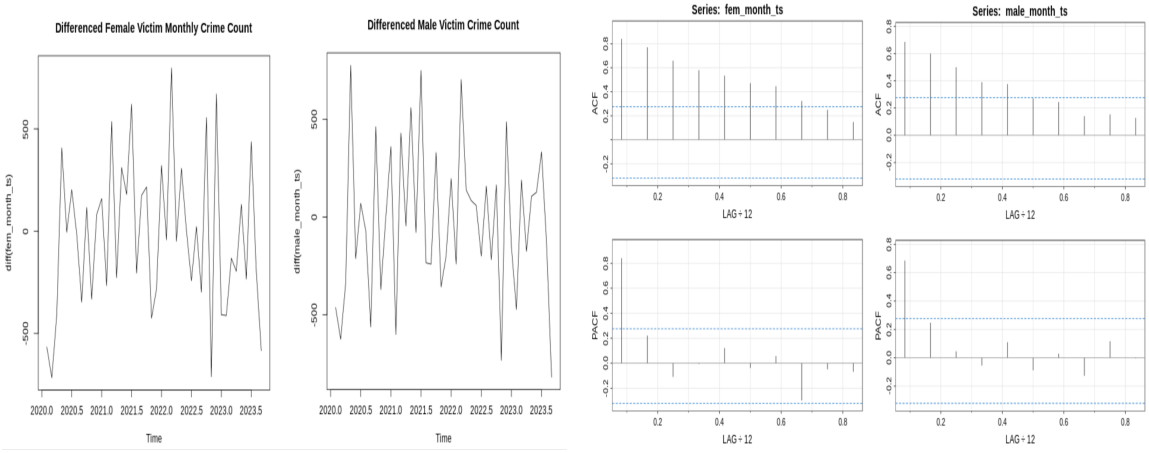


(c) ARIMA$(0, 1, 1) \times (0, 1, 1)_{13}$ Diagnostic Plot

Figure 9

B. Monthly Crime Count by Victim Gender

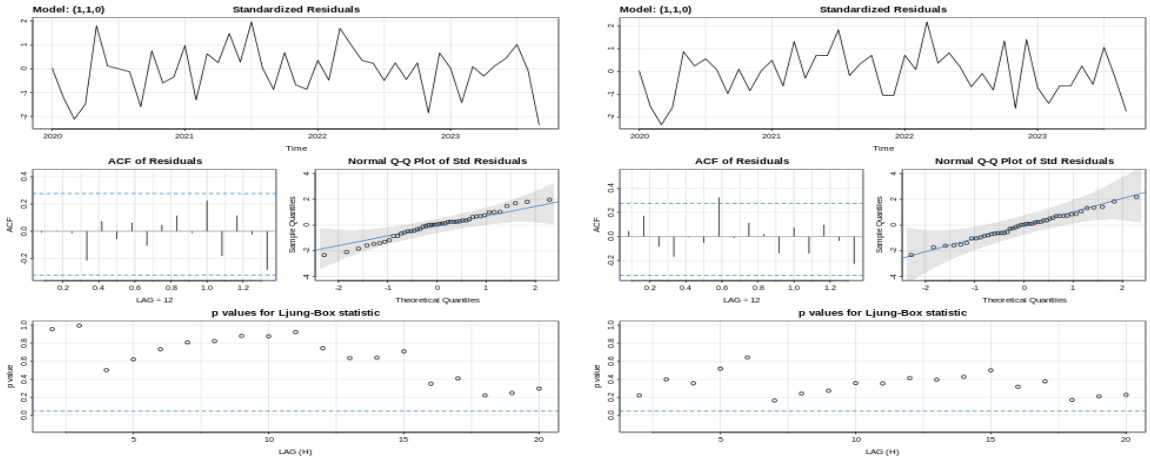
The differenced time series plots Figure 10a of the male and female crime counts indicate some differences in behavior. The differenced data does not seem to have a trend in either of the groups. The ACF and PACF plots 10b of the male and female victims time series both suggest an AR(1) model would be appropriate for this data.



(a) Differenced data of Female and Male Victims (b) ACF & PACF Plots of the differenced data

Figure 10

The diagnostic plots Figure 11 of ARIMA(1,1,0) on both the male and female victims indicate that this is an appropriate model for both segments.



(a) Diagnostic Plots of Male Victims (b) Diagnostic Plots of Female Victims

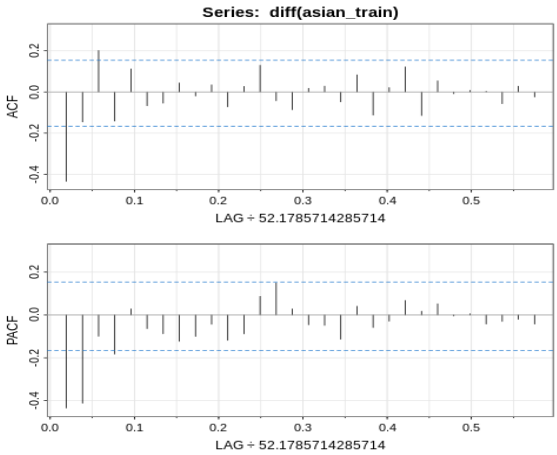
Figure 11

C. Weekly Crime Count by Victim Descent

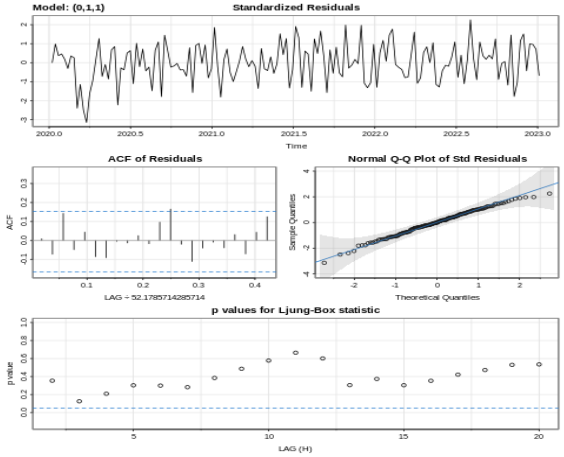
C.1 Asian Victims

Figure 12a is the ACF and PACF plots of the weekly crime count of Asian victims after taking first difference. Suggesting a ARIMA(0,1,1) model for the orginal data.

All diagnostic plots in Figure 12b shows no violation of the assumption of time series models, suggesting that ARIMA(0,1,1) is a suitable model.



(a) ACF & PACF plot of the differenced data



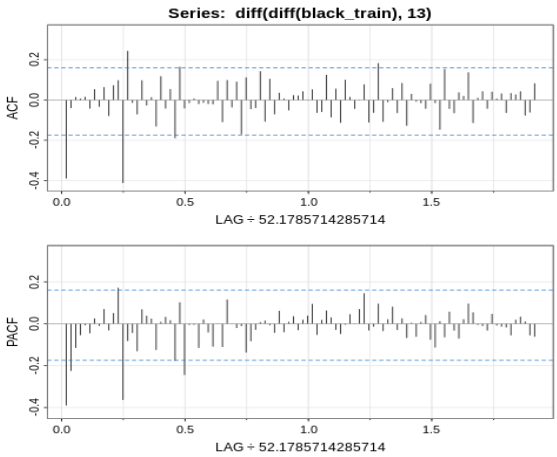
(b) Diagostic plots of the ARIMA(0,1,1) model.

Figure 12

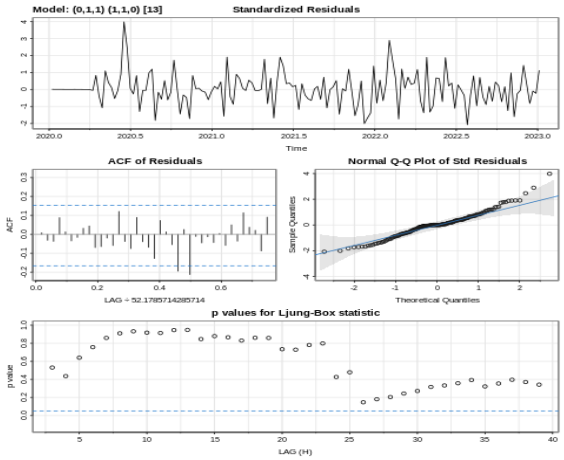
C.2 Black Victims

Figure 13a is the ACF and PACF plots of the weekly crime count of Black victims after taking first difference and seasonal difference with S=13. Suggesting a nonseasonal ARIMA(0,1,1) model for the orginal data. The seasonal components were determined by trying different seasonal parameters, the final model fitted is ARIMA$(0,1,1)\times(1,1,0)_{13}$.

Although two lags for the ACF of residuals of the fitted model are significant and the QQ-plot shows slightly deviation from the straight line, the diagnostic plots in Figure 13b are acceptable in general, suggesting that ARIMA(0,1,1) is a suitable model.



(a) ACF & PACF plot of the differenced data



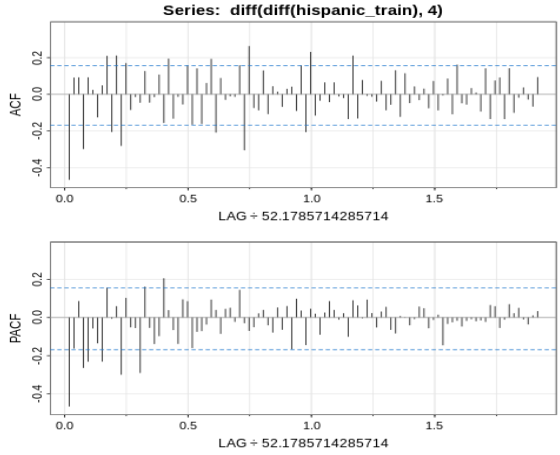
(b) Diagnostic plots of the ARIMA$(0,1,1)\times(1,1,0)_{13}$ model.

Figure 13

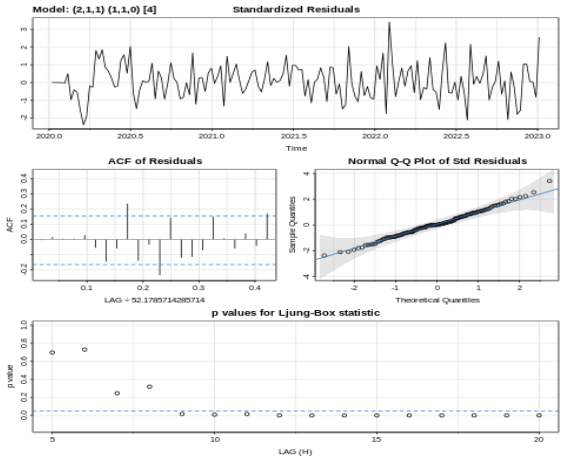
C.3 Hispanic Victims

Figure 14a is the ACF and PACF plots of the weekly crime count of Hispanic victims after taking first difference and seasonal difference with S=4. The final model fitted is $\text{ARIMA}(2,1,1) \times (1,1,0)_4$.

Although two lags for the ACF of residuals of the fitted model are significant and over half of the p-values of the Ljung-Box statistic is not significant in Figure 14b, this model is the best one we can found for SARIMA models. This suggests that other types of models may be more suitable to fit this data.



(a) ACF & PACF plot of the differenced data



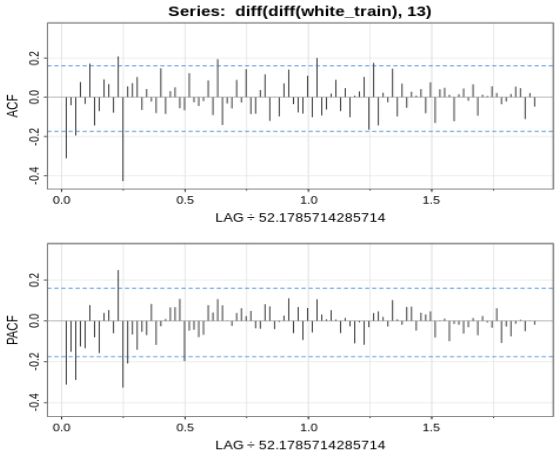
(b) Diagnostic plots of the $\text{ARIMA}(2,1,1) \times (1,1,0)_4$ model.

Figure 14

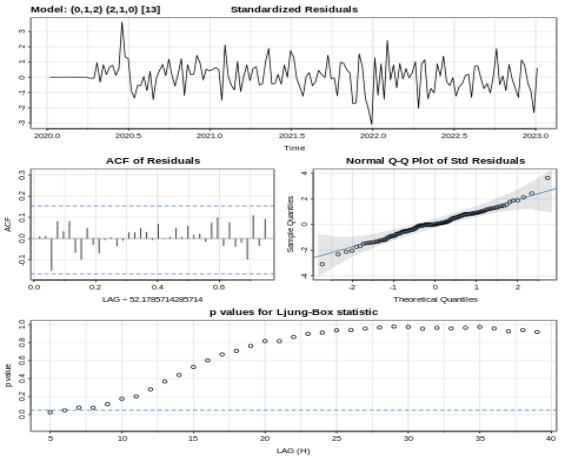
C.4 White Victims

Figure 15a is the ACF and PACF plots of the weekly crime count of Black victims after taking first difference and seasonal difference with S=13. The final model fitted is $\text{ARIMA}(0,1,2) \times (2,1,0)_{13}$.

Although one p-value of the Ljung-Box statistic is not significant, the diagnostic plots in Figure 15b are acceptable in general, suggesting that $\text{ARIMA}(0,1,2) \times (2,1,0)_{13}$ is a suitable model.



(a) ACF & PACF plot of the differenced data



(b) Diagnostic plots of the ARIMA$(0,1,2)\times(2,1,0)_{13}$ model.

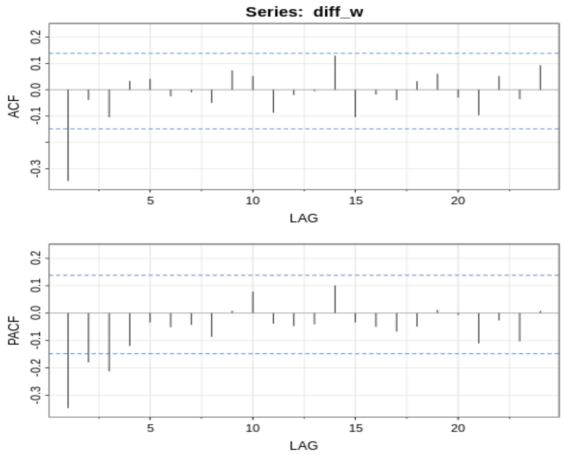
Figure 15

D. Weekly Crime Count by Crime Type

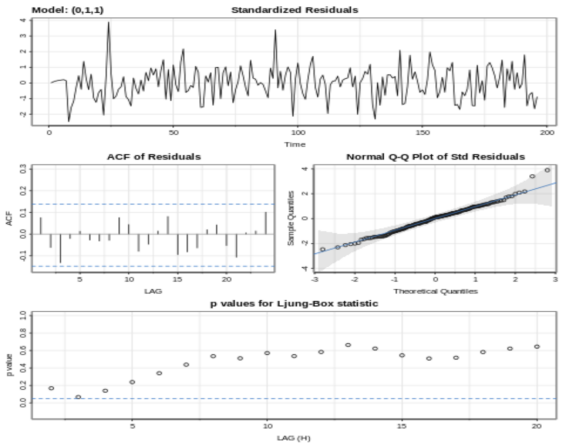
D.1 Burglary

Figure 16a is the ACF and PACF plots of the weekly crime count of Burglary after taking first difference. Suggesting a ARIMA(0,1,1) model for the orginal data.

Although there is one p-value that is not significant in Ljung-Box statistic, all other diagnostic plots in Figure 16 shows no violation of the assumption of time series models, suggesting that ARIMA(0,1,1) is still a suitable model.



(a) ACF & PACF plot of the differenced data



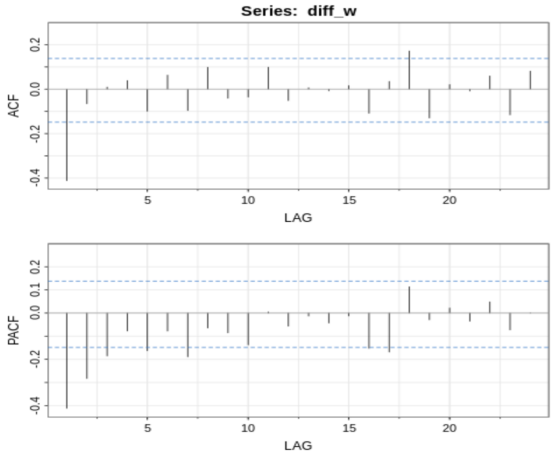
(b) Diagnostic plots of the ARIMA(0,1,1) model.

Figure 16

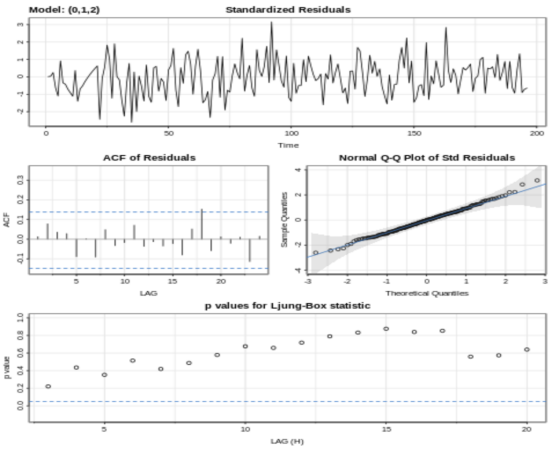
D.2 Theft from Motor Vehicle

Figure 17a is the ACF and PACF plots of the weekly crime count of Theft from Motor Vehicle after taking first difference. Suggesting an ARIMA(0,1,1) model for the orginal data.

All diagnostic plots in Figure 17 shows no violation of the assumption of time series models, suggesting that ARIMA(0,1,2) is a suitable model.



(a) ACF & PACF plot of the differenced data



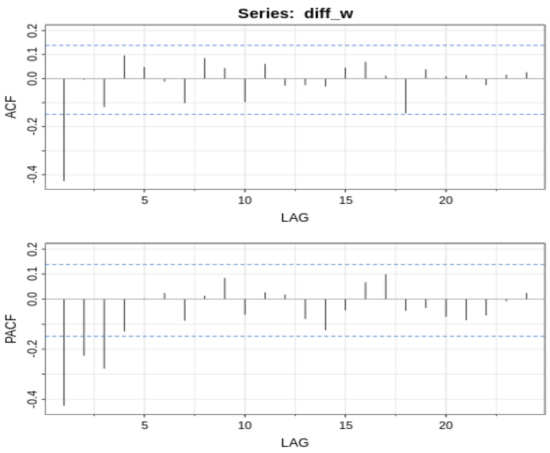
(b) Diagnostic plots of the ARIMA(0,1,2) model.

Figure 17

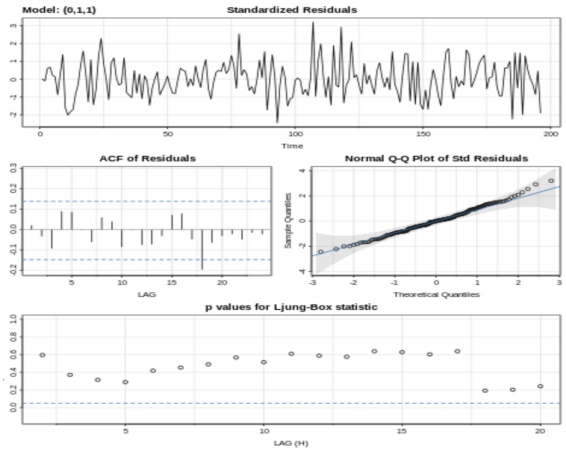
D.3 Battery Simple Assault

Figure 18a is the ACF and PACF plots of the weekly crime count of Battery Simple Assault after taking first difference. Suggesting a ARIMA(0,1,1) model for the orginal data.

Although there is a spike of one lag for the ACF of Recisuals plot18, all the p-values of Ljung-Box statistic are significant, suggesting that ARIMA(0,1,1) is still a suitable model.



(a) ACF & PACF plot of the differenced data



(b) Diagnostic plots of the ARIMA(0,1,1) model.

Figure 18