STA 137: GD Data Report

Huachao Lin

Introduction

The monthly handgun sales and firearms related deaths data (GD data) in California is the monthly observations at 227 times point in 1980-1998. By plotting these data, we can observe that acf decays slowly over time lag, which means adjacent observations are not independent. This leads us to use the time series analysis instead of some traditional statistical methods whose assumption is that adjacent observations are independent and identically distributed.

The analysis of the monthly handgun sales and firearms related deaths time series data could be useful for evaluating the relation between monthly handgun sales and firearms related deaths and forecasting the firearms related deaths. This will help evaluate the safety and gun policy in California.

Material and Methodology

Let variable Y_t represents firearms related deaths and variable Z_t represents monthly handgun sales. As shown in Figure 1, both time series do not seem stationary since the mean of each series seems to change with time. As shown in Figure 2, ACF decays slowly and there is only few peaks in PACF for both series. This also implies a trend in the data. Transformation is required for the time series. By comparing the two datasets, we can observe there seems to be a correlation between two time series. Drawing a scatterplot matrix as Figure 3 shown, an obvious pattern can be observed. The time series analysis consists of four major parts: regression, transformation, model building and model diagnose. The detailed descriptions are as follows.

- Regression

When we want to explore regression with autocorrelated errors to construct a linear model between Y_t and Z_t, the first step is to run an ordinary regression of Y_t on Z_t. It can be implemented in a form of $Y_t = \beta*Z_t$ + error or $Y_t = \beta*Z_{t-3}$ + error, depending on the cross correlation function between two series. Once the model is fitted, we need to get the residuals of the model. Then the residuals are considered to be a new time series which should be dealt with as a normal time series analysis.

- Transformation

Constructing a time plot of the data and identifying the dependence order of the model helps us to understand whether log transformation and/or differentiating transformation are needed. If it is not stationary, for example, the variability in the data change with time, transforming the data to stabilize the variance is needed. If the sample ACF, $\rho\hat{}(h)$, decays to zero slowly, differencing may be needed. If differencing is called for, inspect the time plot of ∇X_t. Differencing again if necessary until the time plot is stationary.

- Model building

After time series being transformed properly, the next step is to identify preliminary values of the autoregressive order, p and the moving average order, q. The d is differencing times in last step. Possible p and q values can be estimated by looking that ACF and PACF of transformed data tails off and/or cut off with lag. If there are seasonal components, P,Q values should also be estimated based on ACF and PACF. With these possible values, we can start to estimate the parameters.Fit possible SARIMA models with these possible values.

- Model Diagnose and model selection

With these possible fitting models, it is important to select which one is the best model and to diagnose whether this model fit the data well. Calculating the criteria (AIC,AICc,BIC), the model with the smallest criteria values will be determined as the best model. But it is also important to look at the p-value, residuals, QQ plot of the model to judge whether it is a good model. It is important to observe whether there is an obvious pattern in the plot of the standardized residuals and whether ACF shows a dependence structure.

Results

- Monthly handgun sales (Z_t) Result

Because Z_t time series is not stationary, I apply log and differentiating transformation to the data. The time plot and QQ plot for transformed data are as shown in Figure 4. We can observe that applying log transformation does not seem to stabilize the variance and the deviation from normal distribution in QQ plot is large. Therefore the log transformation doesn't seem appropriate for this series. $\nabla(Z_t)$ looks like stationary and looks like a normal distribution except one outlier.

Figure 5 shows the ACF and PACF of $\nabla(Z_t)$, from which we can observe that the acf for $\nabla(Z_t)$ does not decay quickly at lags multiple of 12 and PACF has a large peak in 12. So I do a differentiating transformation of lag 12 again to remove the seasonality. The time plot, ACF and PACF of $\nabla\nabla_{12}(Z_t)$ are shown in Figure 6. Lag 12 trend is removed and there are no apparent deviation from stationarity. The acf seems cut off and pacf seems tail off(especially after lag2), and the acf at lags multiple of 12 seems to tails off and pacf at lags multiple of 12 seems to tail off or cut off at lag 3*12.

Now I have three candidate models:

SARIMA(0,1,2)x(1,1,1)$_{[12]}$, SARIMA(0,1,2)x(1,1,0)$_{[12]}$ and SARIMA(0,1,2)x(3,1,0)$_{[12]}$. The AIC, AICc or BIC of three model are as follows :

	AIC	AICc	BIC
SARIMA(0,1,2)x(1,1,1)$_{[12]}$	7.566597	7.567405	7.641397
SARIMA(0,1,2)x(1,1,0)$_{[12]}$	7.867013	7.867496	7.926853
SARIMA(0,1,2)x(3,1,0)$_{[12]}$	7.621455	7.622673	7.711214

we can look at these criterias and pick SARIMA(0,1,2)x(1,1,1)$_{[12]}$ as our best model with the smallest criteria values. The diagnostics plot is as shown in Figure 7. We can see there is no apparent trend or pattern in the plot of the standardized residuals but there are some outliers with residuals exceeding three standard deviations. ACF shows no apparent significant dependence structure. Q statistics is not significant at any lag. P-values are small. Normality assumption seems to be appropriate with the exception of outliers. Overall, it seems like a good model and the parameters for this SARIMA(0,1,2)x(1,1,1)$_{[12]}$ model are:

```
          ma1      ma2     sar1     sma1
      -0.2166  -0.2959  0.1232   -1.000
s.e.   0.0671   0.0682  0.0715    0.069
```

- Firearms related deaths (Yt) Result

Because Yt time series is not stationary, I apply log and differentiating transformation to the data. The time plot and QQ plot for transformed data are as shown in Figure 8. We can observe that applying log transformation does not seem to stabilize the variance and the deviation from normal distribution in QQ plot is large. Therefore the log transformation doesn't seem appropriate for this series. ∇(Yt) looks like stationary and looks like a normal distribution.

Figure 9 shows the ACF and PACF of ∇(Yt), from which we can observe that the acf for ∇(Yt) does not decay quickly at lags multiple of 12 and PACF has a large peak in 12. So I do a differentiating transformation of lag 12 again to remove the seasonality. The time plot, ACF and PACF of $\nabla \nabla_{12}$(Yt) are shown in Figure 10. Lag 12 trend is removed and there are no apparent deviation from stationarity. The acf seems cut off or tail off and pacf seems tail off, and the acf at lags multiple of 12 seems to cut off and pacf at lags multiple of 12 seems to tail off .

Now I have two candidate models: SARIMA(0,1,1)x(0,1,1)$_{[12]}$ and SARIMA(1,1,1)x(0,1,1)$_{[12]}$. The AIC, AICc or BIC of two model are as follows :

	AIC	AICc	BIC
SARIMA(0,1,1)x(0,1,1)$_{[12]}$	-1.856231	-1.855991	-1.811351
SARIMA(1,1,1)x(0,1,1)$_{[12]}$	-1.855884	-1.855401	-1.796044

I pick SARIMA(0,1,1)x(0,1,1)$_{[12]}$ as our best model with the smallest value. The diagnostics plot is as shown in Figure 11. We can see there is no apparent trend or pattern in the plot of the standardized residuals. ACF shows no apparent significant dependence structure. Q statistics is not significant at any lag. P-values are small except a few points. Normality assumption seems to be appropriate. Overall, it seems like a good model and the parameters for the SARIMA(0,1,1)x(0,1,1)$_{[12]}$ model are:

```
          ma1      sma1
      -0.6286  -1.0000
s.e.   0.0643   0.1064
```

- Regression with autocorrelated error result

As shown in Figure 12, sample cross correlation function(ccf) between Y_t and Z_t are computed. The ccf between two time series are really large in some lags, which implies that two time series are highly correlated. The ccf seems to be largest at lag 5. Hence I run an ordinary regression of Y_{t+5} on Z_t. And let variable X_t represent the residual of this regression.

As shown in Figure 13, the X_t time series is not stationary, I apply log transformation to the data but it does not seem to stabilize the variance and the deviation from normal distribution in QQ plot is large. Therefore the log transformation doesn't seem appropriate for this series. From Figure 14 of ACF and PACF plot of X_t, ACF seems decay slow at lags multiple of 12. So I do a differentiating transformation at lag 12.

Figure 15 shows the time plot, ACF and PACF of $\nabla_{12}(X_t)$, from which we can observe that $\nabla_{12}(X_t)$ shows a clear pattern with time. So I do a differentiating transformation again to remove the trend. The time plot, ACF and PACF of $\nabla \nabla_{12}(X_t)$ are shown in Figure 16, Figure 17, respectively. Trend is removed and there are no apparent deviation from stationarity. The acf seems cut off and pacf seems tail off, and the acf at lags multiple of 12 seems to cut off or tails off and pacf at lags multiple of 12 seems to tail off.

Now I have two candidate models: SARIMA(0,1,1)x$(0,1,1)_{[12]}$ and SARIMA(0,1,1)x$(1,1,1)_{[12]}$.The diagnostics plot for two models are as shown in Figure 18,19 respectively. We can see there is no apparent trend or pattern in the plot of the standardized residuals. ACF shows no apparent significant dependence structure. Q statistics is not significant at any lag. P-values are small. Normality assumption seems to be appropriate for both models. The AIC, AICc or BIC of two models are as follows :

	AIC	AICc	BIC
SARIMA(0,1,1)x$(0,1,1)_{[12]}$	-1.839984	-1.839478	-1.779214
SARIMA(0,1,1)x$(1,1,1)_{[12]}$	-1.831602	-1.830756	-1.755640

we can look at criteria above and pick SARIMA(0,1,1)x$(0,1,1)_{[12]}$ as our best model with the smallest value. The parameters for this SARIMA(0,1,1)x$(0,1,1)_{[12]}$ model are:

```
          ma1     sma1   xreg
      -0.6347  -1.0000  2e-04
s.e.   0.0670   0.1039  5e-04
```

Conclusion and Discussion

To summarize, the best model I found for monthly handgun sales is a SARIMA$(0,1,2)\text{x}(1,1,1)_{[12]}$ model, whose expression is:

$$(1+0.1232B^{12})(1-B)(1-B^{12})Z_t = (1-0.2166B-0.2959B^2)(1-1.000B^{12})\omega_t \tag{1}$$

The best model for firearms related deaths is a SARIMA$(0,1,1)\text{x}(0,1,1)_{[12]}$ model:

$$(1-B)(1-B^{12})Y_t = (1-0.6286B)(1-1.000B^{12})\omega_t \tag{2}$$

The best model for regression with autocorrelated error is a SARIMA$(0,1,1)\text{x}(0,1,1)_{[12]}$ model:

$$(1-B)(1-B^{12})Y_{t+5}=2e{-}04*(1-B)(1-B^{12})Z_t+(1-0.6347B)(1-1.000B^{12})\omega_t \tag{3}$$

Where $\omega_t \sim WN(0,\sigma^2)$.

We observed from diagnosis plot (Fig 7,11,18,19) that normality assumption seems to be appropriate for each of these models. There is no apparent trend or pattern in the plot of the standardized residuals. ACF shows no apparent significant dependence structure. P-values are small. However, we can also observe from the model that the estimated σ^2 (MSE) for model (1), (2), (3) are 136.7, 0.006843, 0.006855 respectively. The mean values of gun sales and death, however, are 120.1643, 1.238282, respectively. Therefore, The model for monthly handgun sales will produce a huge error when forecasting while the prediction error for firearms related death are relatively low for both model (2) and (3),which is also shown in prediction plot for model (1) and (2) in Figure 20. We can observe the error expand more quickly in model (1) than model (2). This might be because the variance of monthly gun sales is large or because only using time series to predict gun sales is not enough, other variables should also be considered.

What I have to mention is that, in my analysis, I run regression Y_{t+5} on Z_t instead of Y_t on Z_t. The reason is that I found ccf is locally highest in lag -4 and lag 5. Hence, it might be better to regress with lag 5. Regression Y_{t-4} on Z_t might give us an interesting result of how future gun sale related to previous death, which might lead to a smaller prediction error for handgun sales. Another interesting thing we can get from scatter plot in Figure 3 is that there seems to be some second order term in the relation between two variables. Due to time constraints, I am not going to evaluate how the second term impact the model.

From this project, I learned how to do a time series analysis, the main steps are plotting and identifying whether a transformation is needed, building several model candidates based on the behavior of acf and pacf, using model selection criteria to select the best model and diagnosing the models.

Appendix

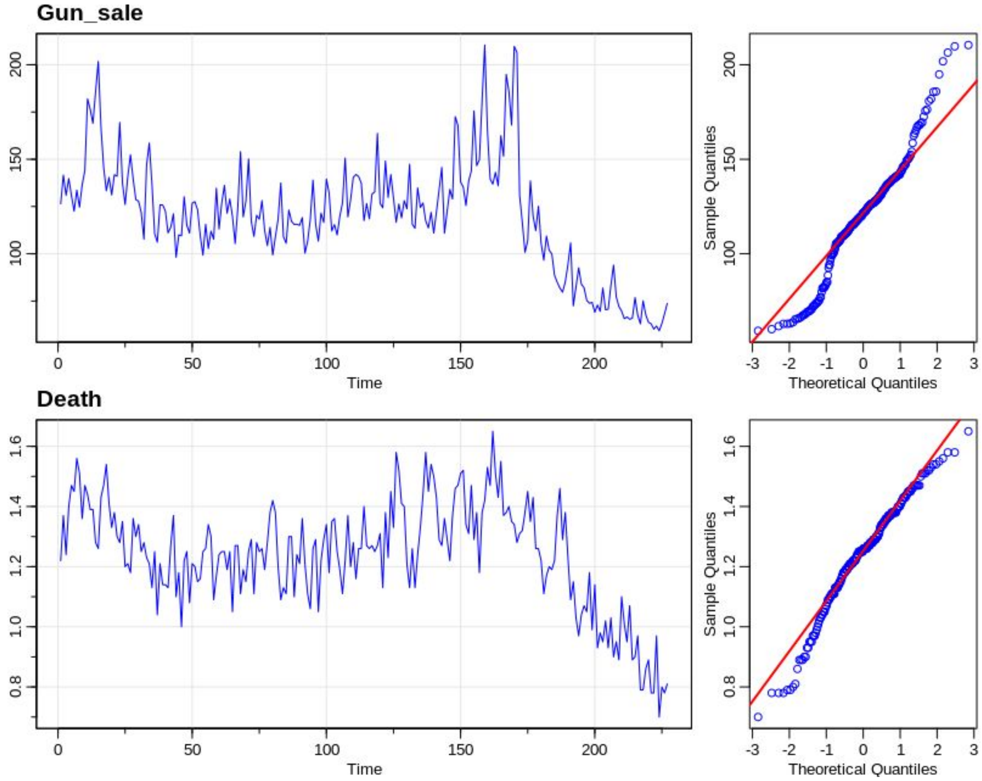
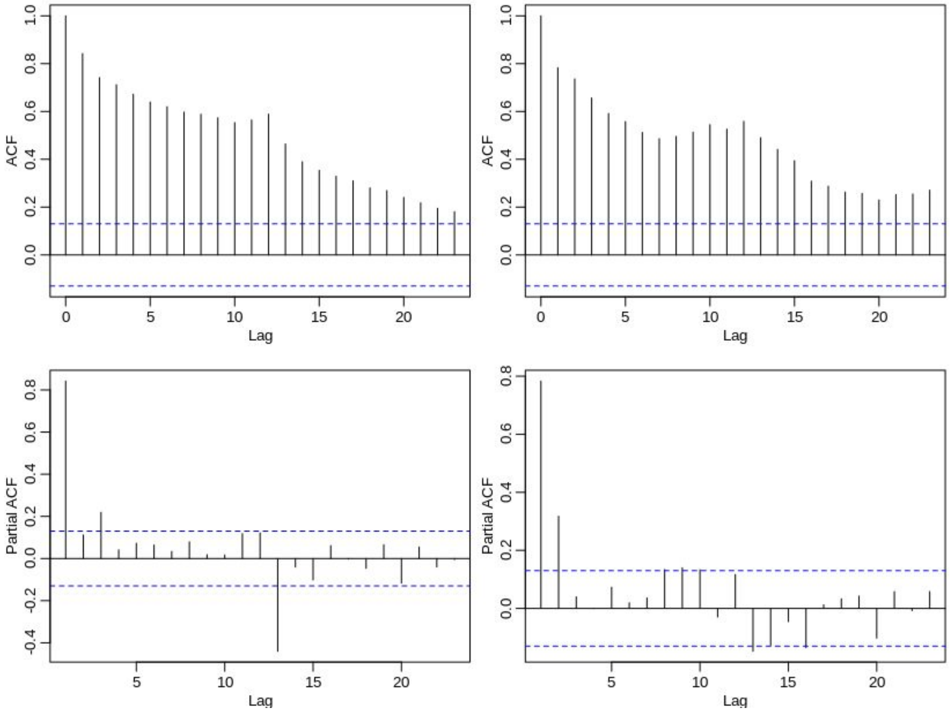


Figure 1: *Variations over time*



(a) Gun_sale (b) Death

Figure 2*: ACF and PACF of times series (a) Gun_sale,(b) Death*

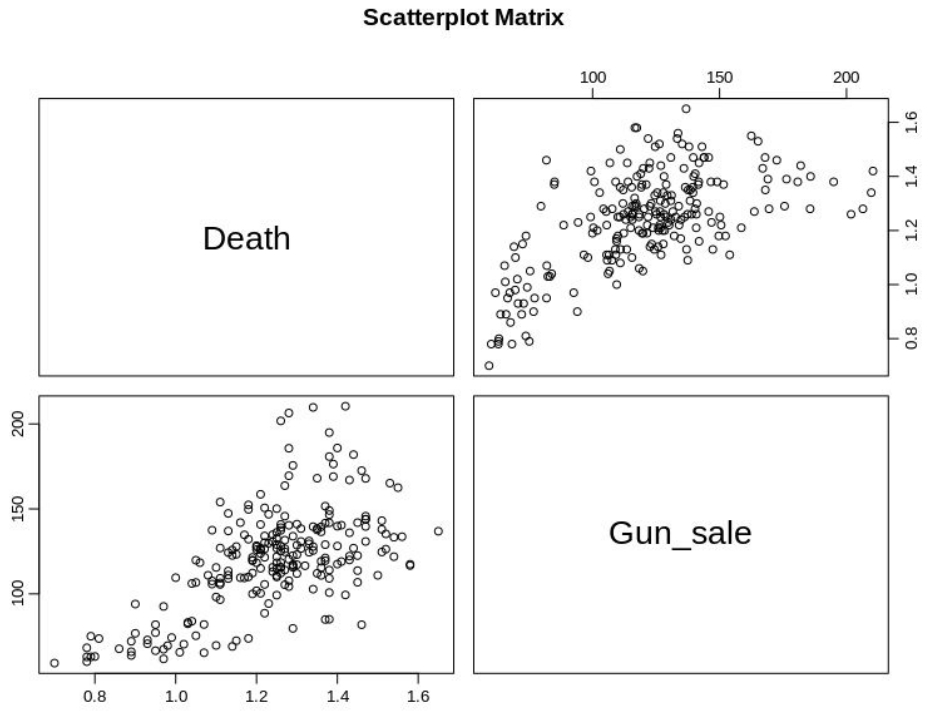


Figure 3: *Scatter matrix between Death and gun_sale*

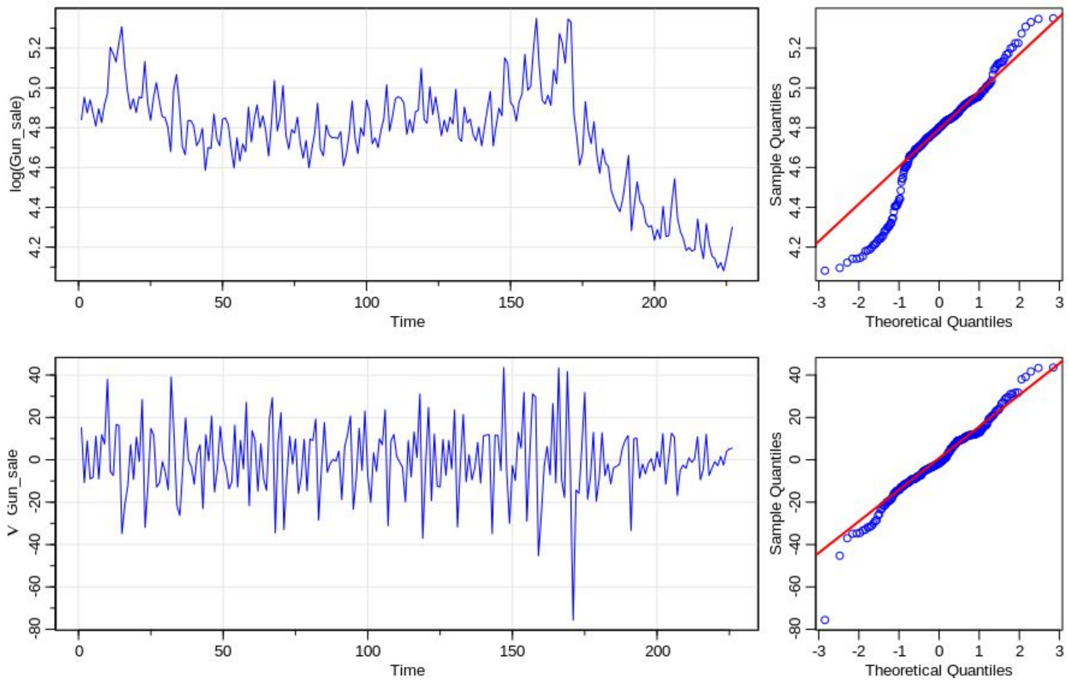


Figure 4: *Time plot and QQ plot of log(Gun_sale) and* ∇*(Gun_sale)*

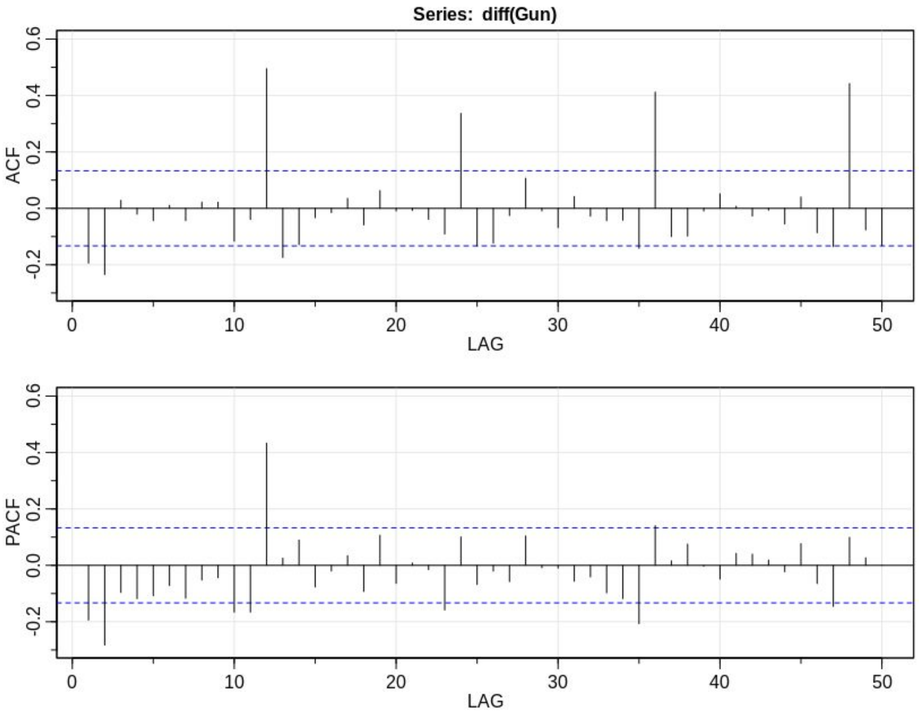


***Figure 5**: ACF and PACF of ∇(Gun_sale)*

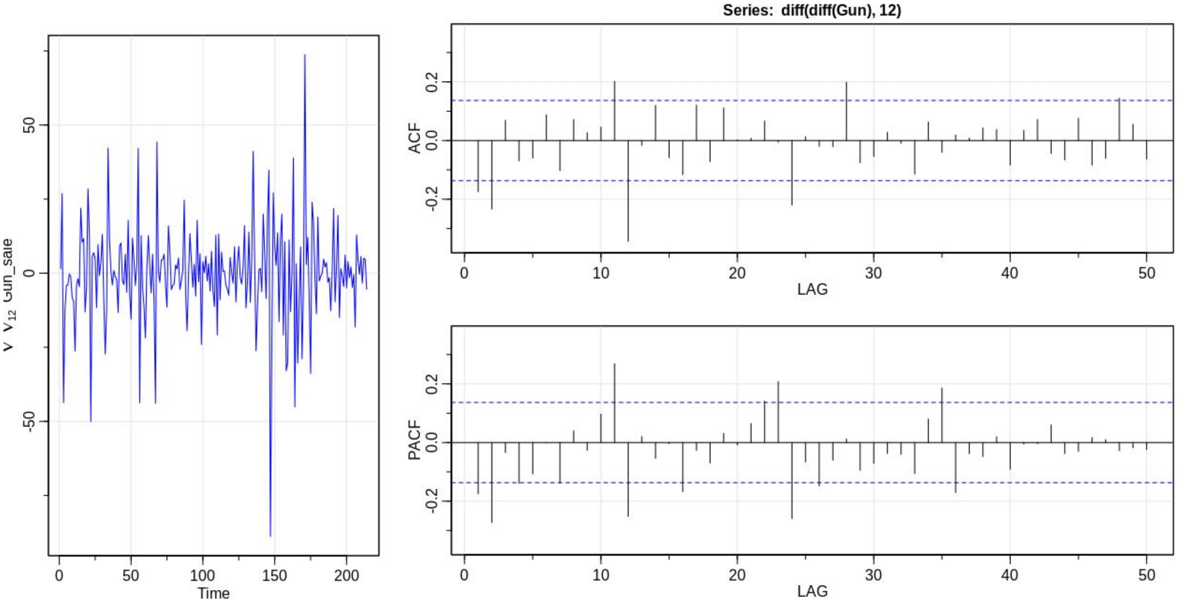


***Figure 6:** ACF and PACF of* $\nabla\nabla_{12}$*(Gun_sale)*

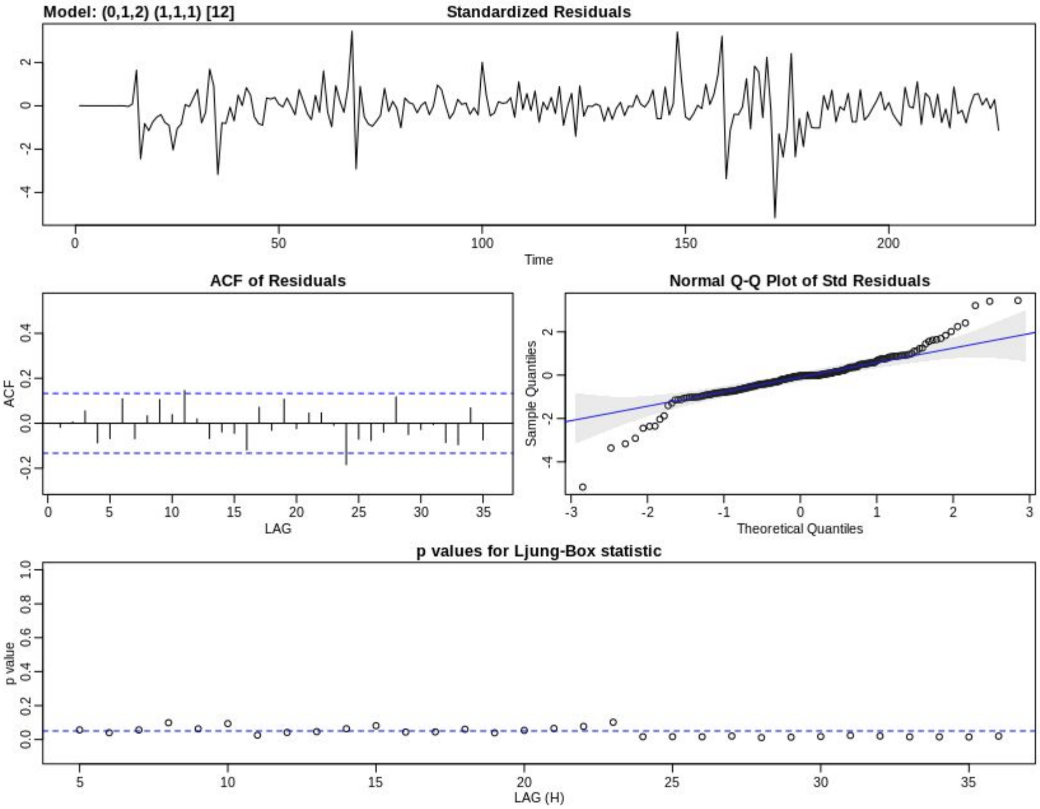


***Figure** 7: Diagnose plot for SARIMA model of Gun_sale data*

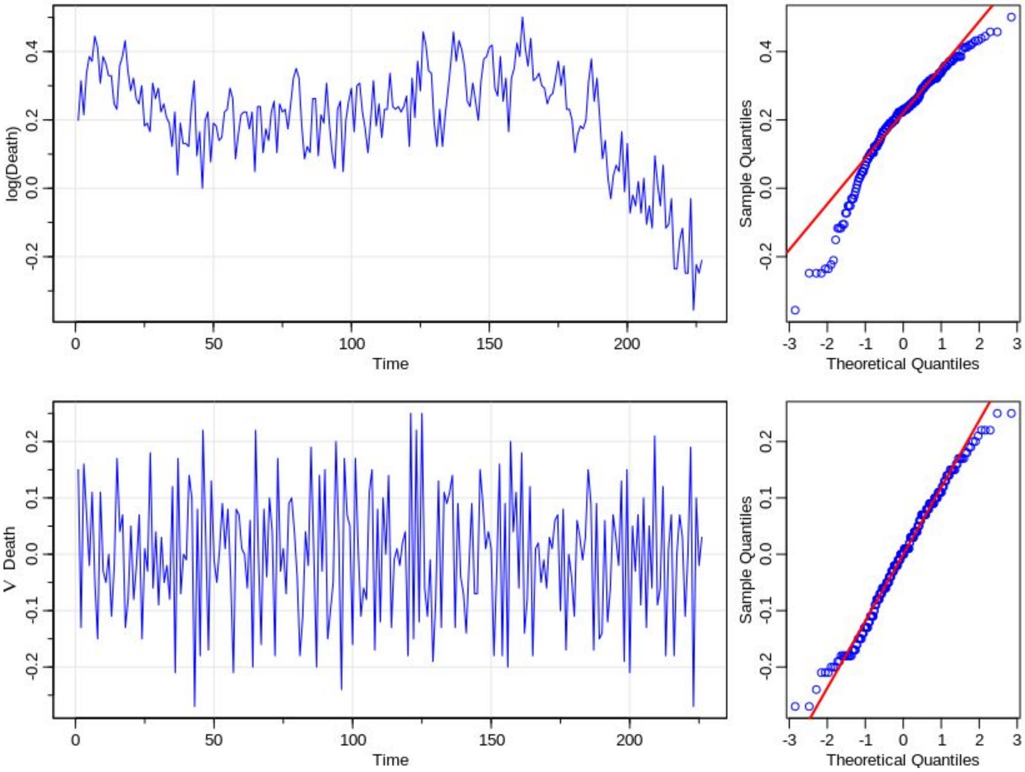


***Figure 8**: Time plot and QQ plot of log(Death) and ∇(Death)*

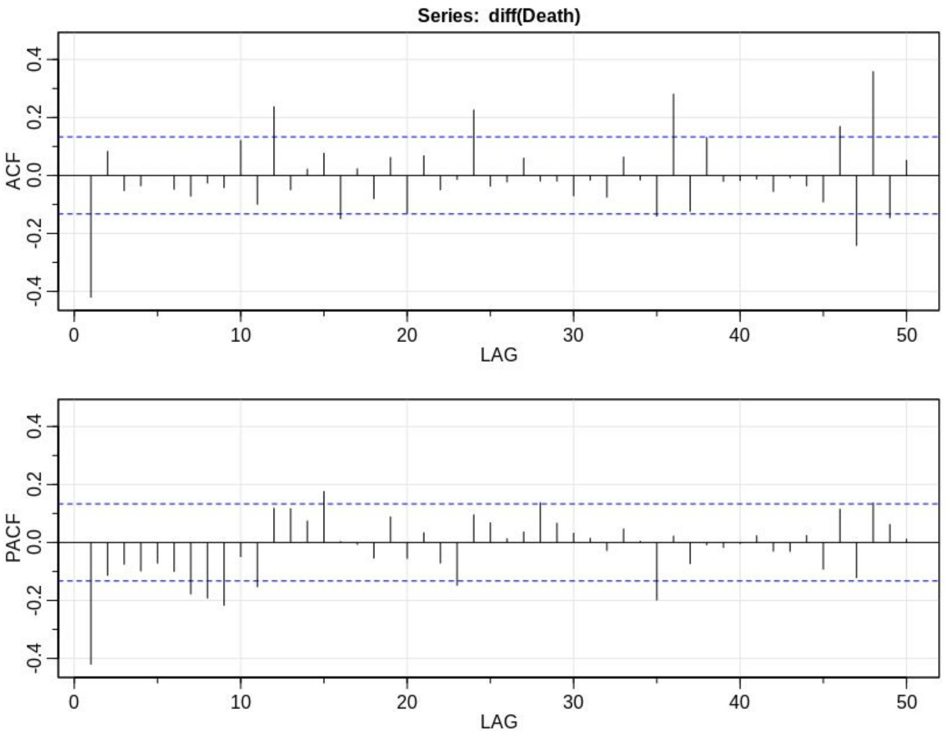


Figure 9: *ACF and PACF of ∇(Death)*

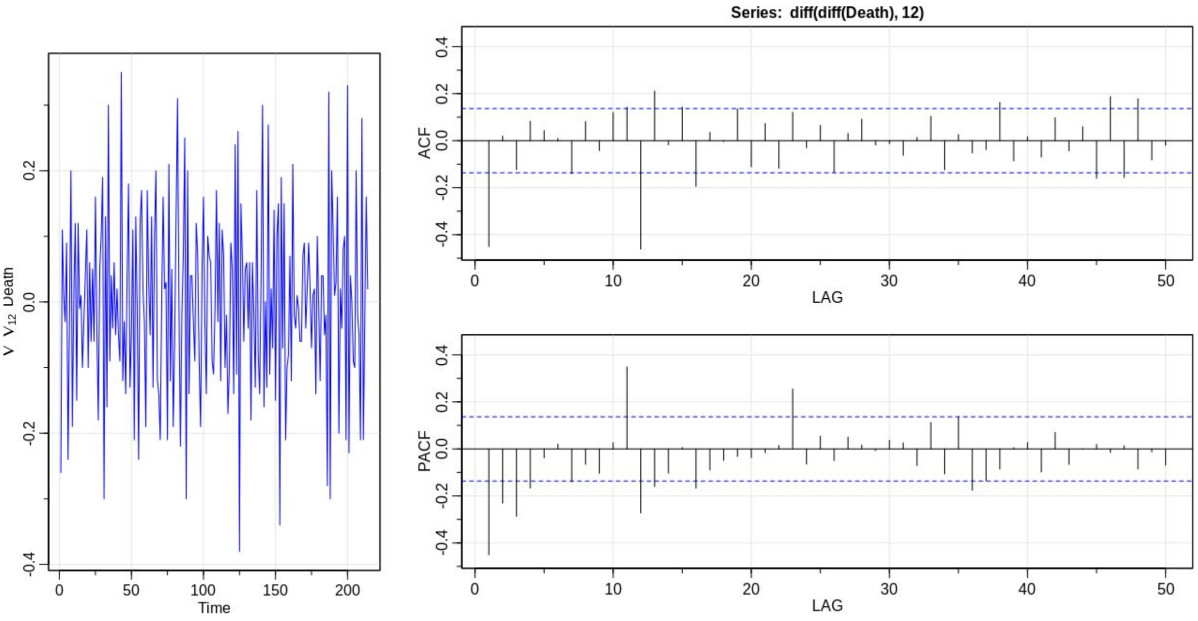


Figure 10: *ACF and PACF of ∇∇$_{12}$(Death)*

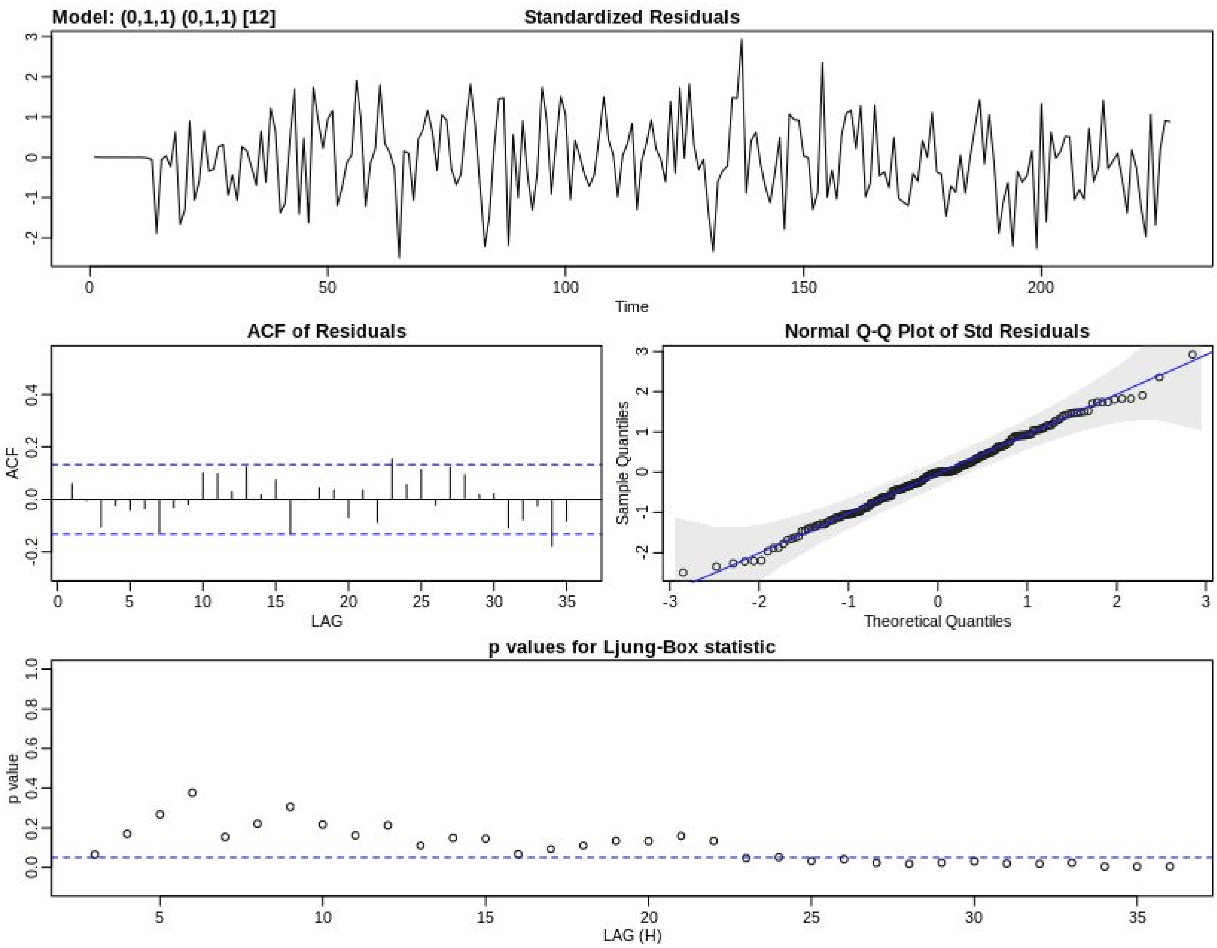


***Figure 11**: Diagnose plot for SARIMA model of Death data*

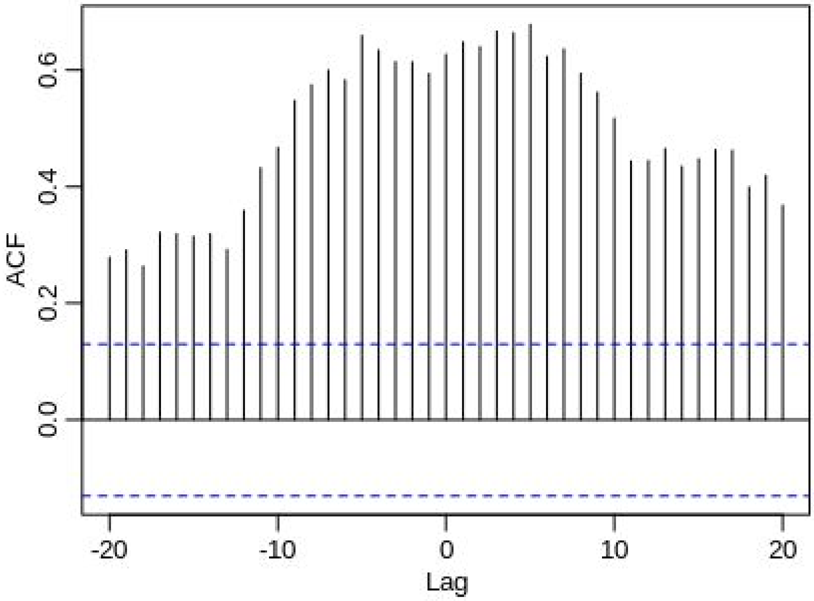


***Figure 12:** Cross correlation function between gun_sale and death*

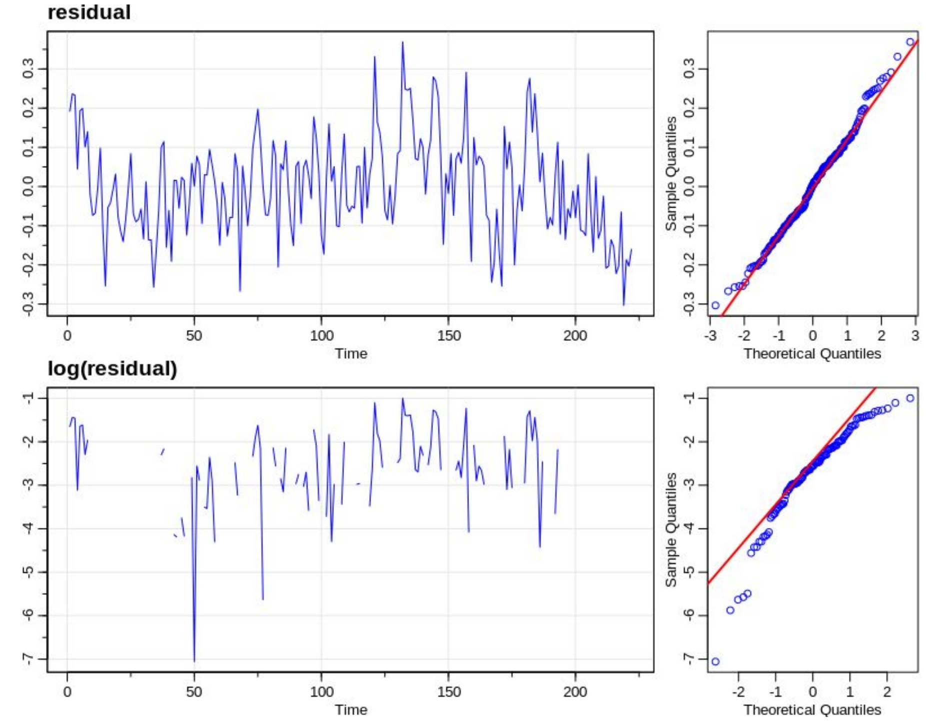


Figure 13: *Time plot and QQ plot of residual and log(Residual)*

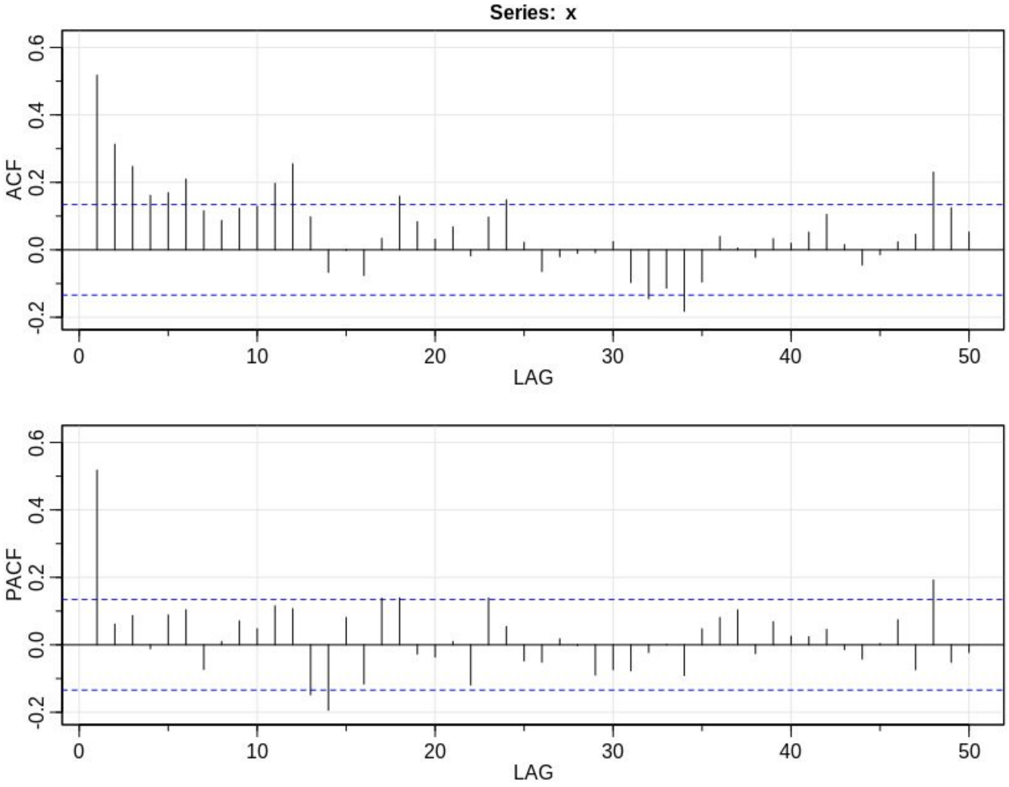


Figure 14: *ACF and PACF of residuals*

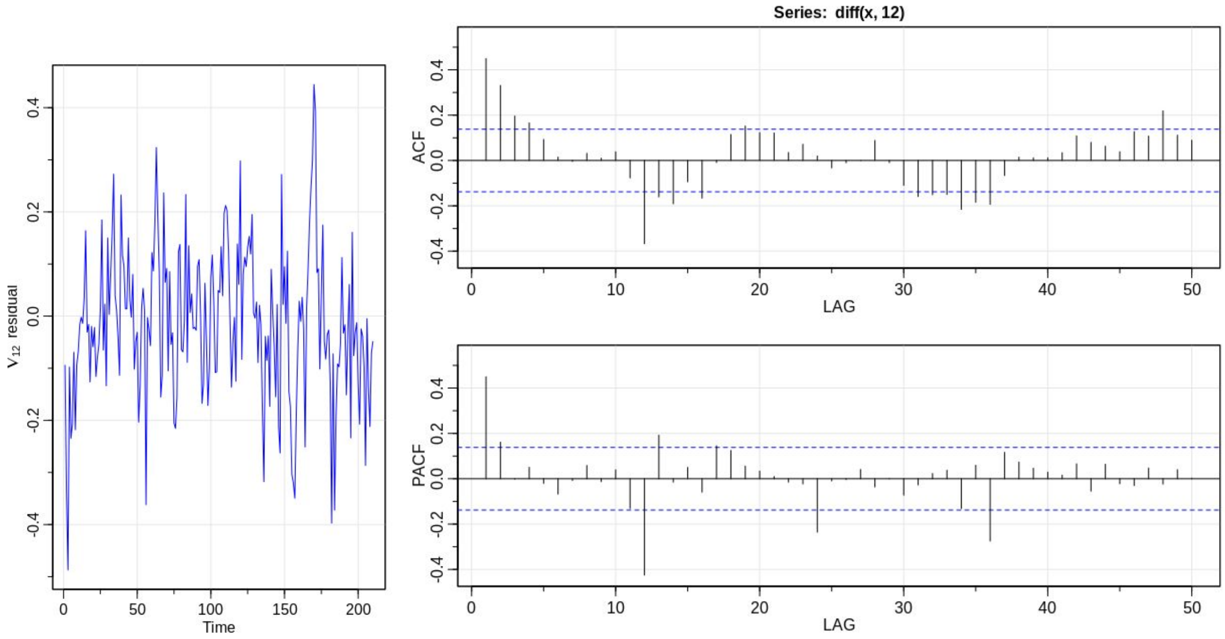


Figure 15: *Time plot, ACF and PACF of* ∇_{12}*(residuals)*

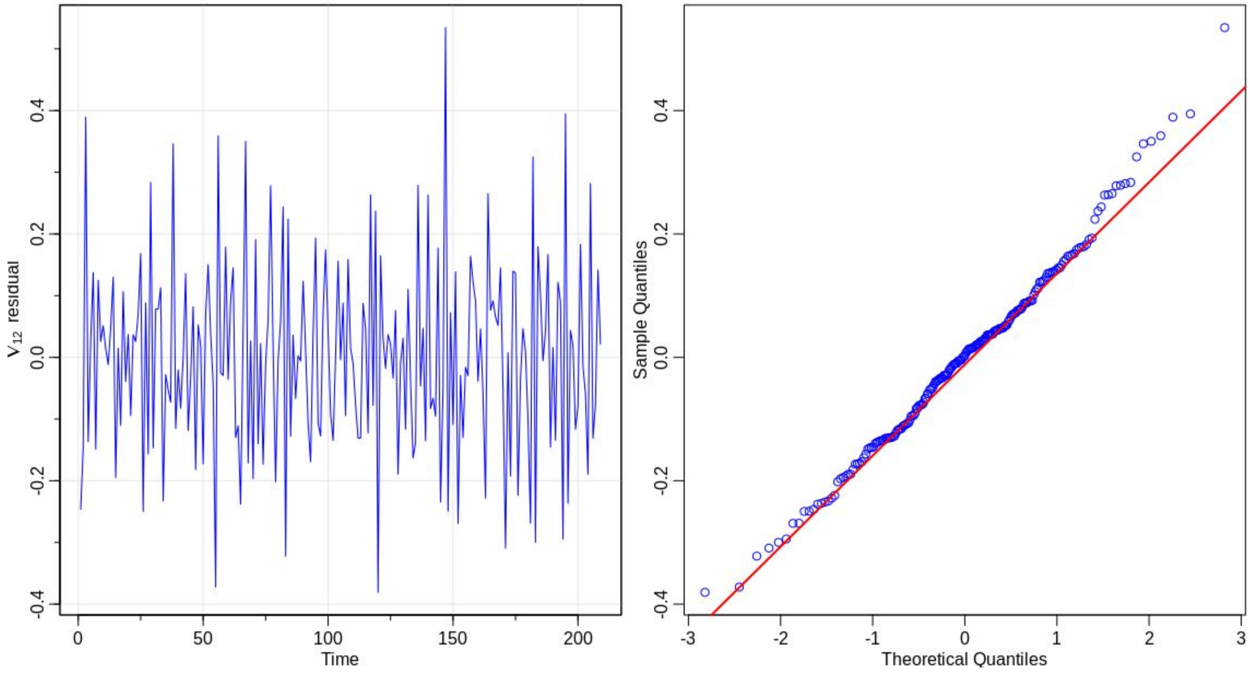


Figure 16: *Time plot and QQ plot of* $\nabla\nabla_{12}$*(residuals)*

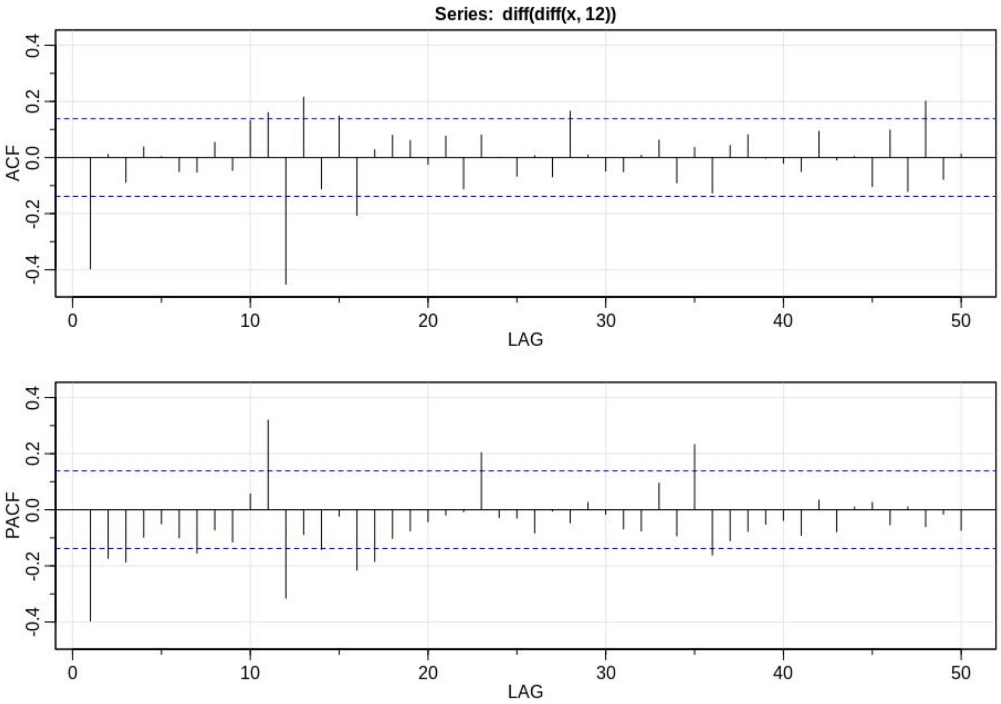


Figure 17*: ACF and PACF of* $\nabla\nabla_{12}$*(residuals)*

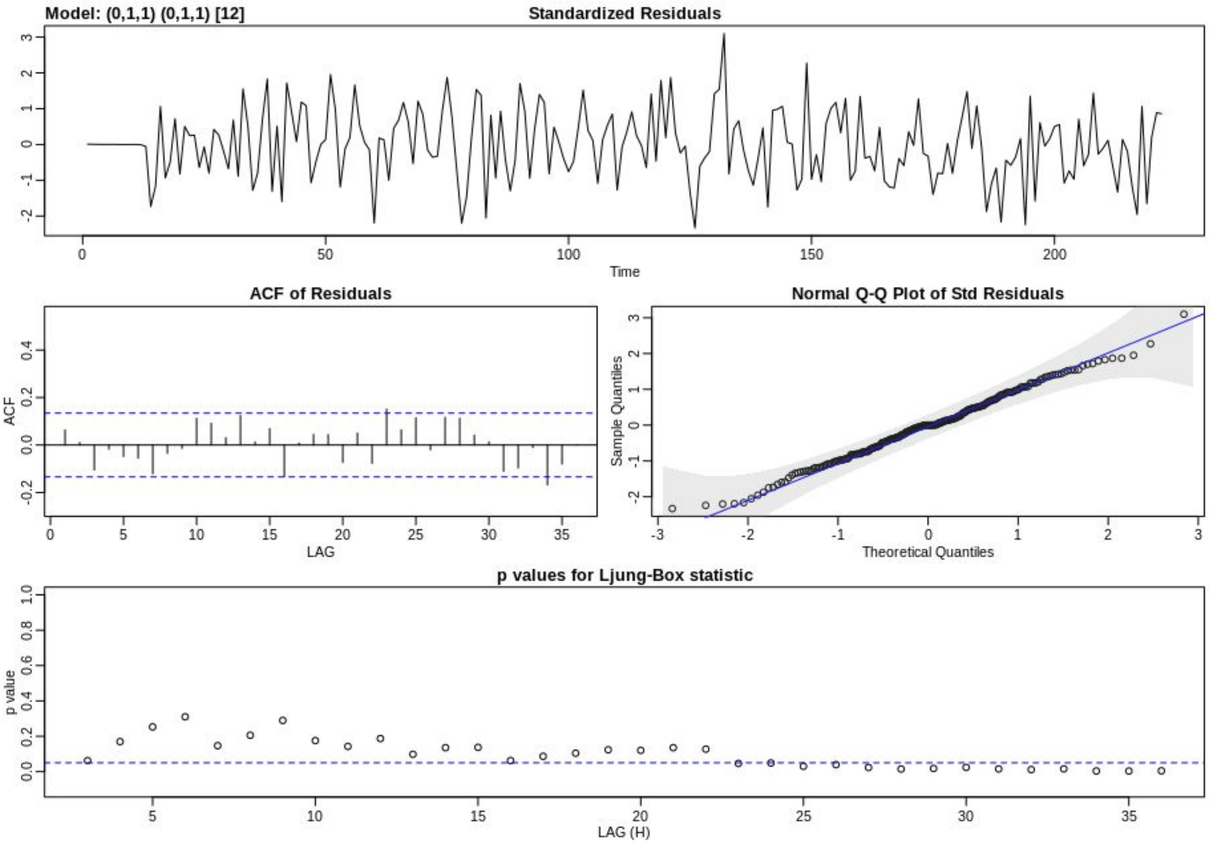


Figure 18*: Diagnose plot of* SARIMA(0,1,1)x(0,1,1)$_{[12]}$

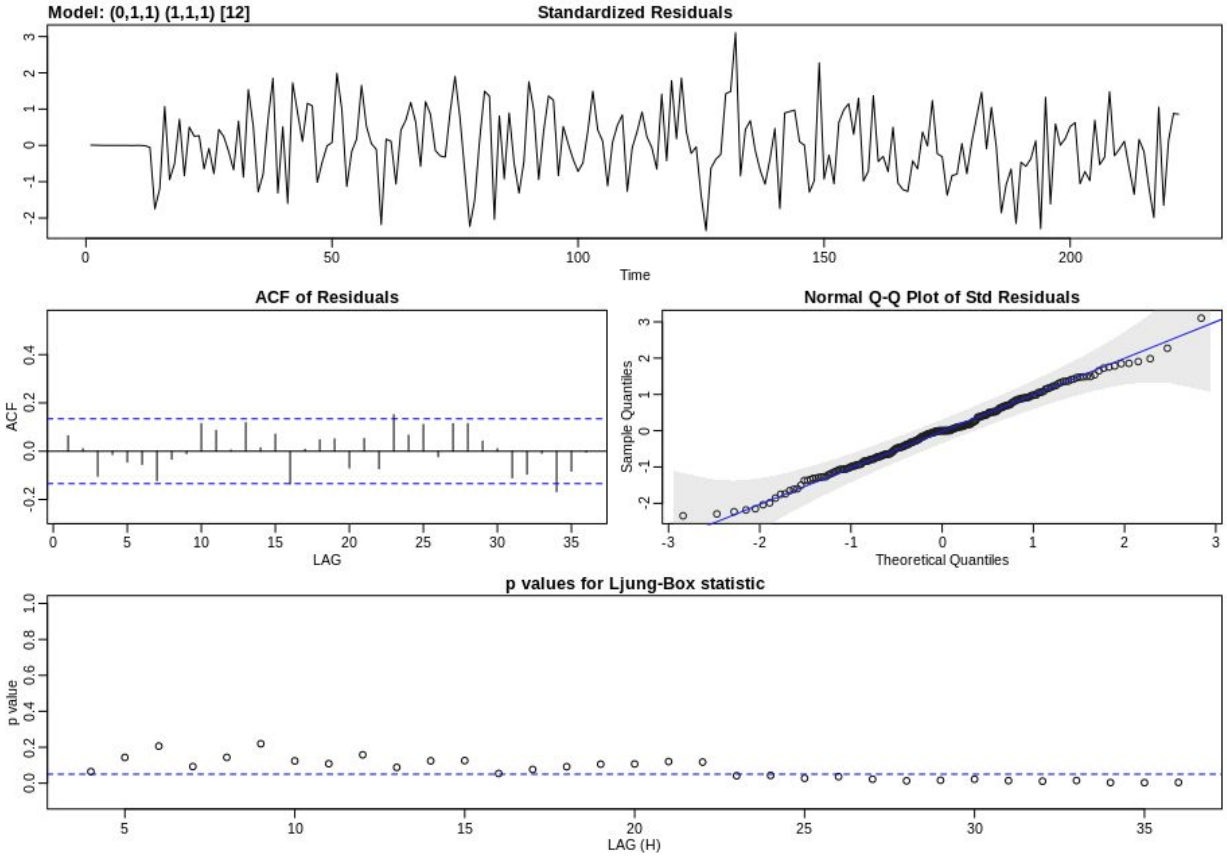


Figure 19: *Diagnose plot of* SARIMA(0,1,1)x(1,1,1)$_{[12]}$

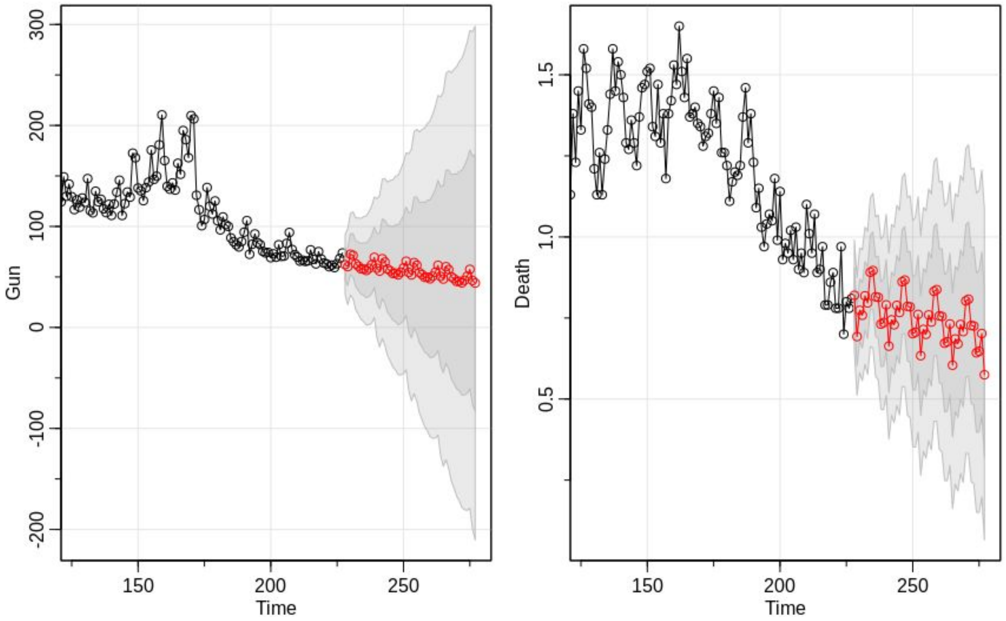


Figure 20: *Prediction plot for model(1) and model(2)*

R code from here:

```
## ------------------------------------------------------------------------  ##
# read data
library(astsa)
GD_data = read.table("/home/huachao/Documents/course/STA137/Final
Project/GD.dat.txt",header = FALSE)
Gun = GD_data[,1]
Death = GD_data[,2]
## plot the data
layout(matrix(1:4,2), widths=c(2.5,1))
par(mgp=c(1.6,.6,0), mar=c(2,2,.5,0)+.5)
tsplot(Gun, main="", ylab="", col=4, margin=0)
mtext("Gun_sale", side=3, line=.5, cex=1.2, font=2, adj=0)
tsplot(Death, main="", ylab="", col=4, margin=0)
mtext("Death", side=3, line=.5, cex=1.2, font=2, adj=0)
qqnorm(Gun, main="", col=4); qqline(Gun, col=2, lwd=2)
qqnorm(Death, main="", col=4); qqline(Death, col=2, lwd=2)

par(mfrow = c(2,2))
acf(Gun)
acf(Death)
pacf(Gun)
pacf(Death)
# correlation part plot
pairs(~Death+Gun_sale,data=GD_data,main="Scatterplot Matrix")
ccf(Death,Gun)
## ------------------------------------------------------------------------  ##
## transformation         # do Gun first
## plot the data
layout(matrix(1:4,2), widths=c(2.5,1))
par(mgp=c(1.6,.6,0), mar=c(2,2,.5,0)+.5)
tsplot(log(Gun), main="", ylab="log(Gun_sale)", col=4, margin=0)
tsplot(diff(Gun), ylab=expression(nabla~Gun_sale), main="", col=4, margin=0)
qqnorm(log(Gun), main="", col=4); qqline(log(Gun), col=2, lwd=2)
qqnorm(diff(Gun), main="", col=4); qqline(diff(Gun), col=2, lwd=2)

# do Death second
tsplot(log(Death), main="", ylab="log(Death)", col=4, margin=0)
tsplot(diff(Death), main="", ylab=expression(nabla~Death), col=4, margin=0)
```

```
qqnorm(log(Death), main="", col=4); qqline(log(Death), col=2, lwd=2)
qqnorm(diff(Death), main="", col=4); qqline(diff(Death), col=2, lwd=2)
## ----------------------------------------------------------------------------  ##
# build model
par(mfrow=c(2,1))
acf2(diff(Gun),50)
tsplot(diff(diff(Gun),12), ylab=expression(nabla~nabla[12]~Gun_sale), main="", col=4,
margin=0)
acf2(diff(diff(Gun),12),50)
## ----------------------------------------------------------------------------  ##
# model selection and diagnose
m1=sarima(Gun, p=0, d=1, q=2,D=1,P=1,Q=1,S=12)
m2=sarima(Gun, p=0, d=1, q=2,D=1,P=1,S=12)
m3=sarima(Gun, p=0, d=1, q=2,D=1,P=3,S=12)
c(m1$AIC,m1$AICc,m1$BIC)
c(m2$AIC,m2$AICc,m2$BIC)
c(m3$AIC,m3$AICc,m3$BIC)
## ----------------------------------------------------------------------------  ##
# do death part
acf2(diff(Death),50)
tsplot(diff(diff(Death),12), ylab=expression(nabla~nabla[12]~Death), main="", col=4,
margin=0)
acf2(diff(diff(Death),12),50)
m4=sarima(Death, p=0, d=1, q=1,D=1,Q=1,S=12)
m5=sarima(Death, p=1, d=1, q=1,D=1,Q=1,S=12)
c(m4$AIC,m4$AICc,m4$BIC)
c(m4$AIC,m4$AICc,m4$BIC)
## ----------------------------------------------------------------------------  ##
## do autocorrelated part
fit1 = lm(Death[6:227]~Gun[1:222])
x = resid(fit1)
layout(matrix(1:4,2), widths=c(2.5,1))
par(mgp=c(1.6,.6,0), mar=c(2,2,.5,0)+.5)
tsplot(x, main="", ylab="", col=4, margin=0)
mtext("residual", side=3, line=.5, cex=1.2, font=2, adj=0)
tsplot(log(x), main="", ylab="", col=4, margin=0)
mtext("log(residual)", side=3, line=.5, cex=1.2, font=2, adj=0)
qqnorm(x, main="", col=4); qqline(x, col=2, lwd=2)
qqnorm(log(x), main="", col=4); qqline(log(x1), col=2, lwd=2)
```

```
acf2(x,50)
par(mfrow=c(1,2))
tsplot(diff(x,12), ylab=expression(nabla[12]~residual), main="", col=4, margin=0)
acf2(diff(x,12),50)
tsplot(diff(diff(x,12)), ylab=expression(nabla[12]~residual), main="", col=4, margin=0)
qqnorm(diff(diff(x,12)), main="", col=4); qqline(diff(diff(x,12)), col=2, lwd=2)
acf2(diff(diff(x,12)),50)
# fit
m6 = sarima(Death[6:227],0,1,1,0,1,1,S=12,xreg=Gun[1:222])
m7 = sarima(Death[6:227],0,1,1,1,1,1,S=12,xreg=Gun[1:222])
## model selection
c(m6$AIC,m6$AICc,m6$BIC)
c(m7$AIC,m7$AICc,m7$BIC)

## prediction plot
par(mfrow=c(1,2))
sarima.for(Gun,0,1,2,1,1,0,S=12,n.ahead = 50)
sarima.for(Death,0,1,1,0,1,1,S=12,n.ahead = 50)
```