Objective 2 Analysis

Chance Robinson

9/27/2019

Contents

Exploratory Data Analysis

Library Imports

Load the csv data

```
train <- read_csv('../../../data/train.csv')
test <- read_csv('../../../data/test.csv')
```

Data Dictionary

Column Name	Type	Description
1. datetime	Date	YYYY-MM-DD HH24 (example: 2011-01-01 04:00:00)
2. season	Integer	(1-4)
3. holiday	Integer	(0 or 1)
4. workingday	Integer	(0 or 1)
5. weather	Integer	(1-4)
6. temp	Float	temparture in Celcius
7. atemp	Float	"feels like" temperature in Celsius
8. humidity	Integer	relative humidity
9. windspeed	Float	wind speed
10. casual	Integer	count of casual users
11. registered	Integer	count of registered users
12. count	Integer	count of total users `response variable`

Factors

- season
  - 1 = Dec 21 ~ March 20 (Spring)
  - 2 = March 21 ~ Jun 20 (Summer)
  - 3 = June 21 ~ Sept 20 (Fall)
  - 4 = Sept 21 ~ Dec 20 (Winter)
- holiday
  - 0 = No
  - 1 = Yes
- workingday
  - 0 = No
  - 1 = Yes

```
train$season <- factor(train$season, labels = c("Spring", "Summer", "Fall", "Winter"))
test$season <- factor(test$season, labels = c("Spring", "Summer", "Fall", "Winter"))

table(train$season)
```

```
##
## Spring Summer   Fall Winter
##   2686   2733   2733   2734
```

```
train$holiday <- factor(train$holiday, labels = c("No", "Yes"))
test$holiday <- factor(test$holiday, labels = c("No", "Yes"))

table(train$holiday)
```

```
##
##    No   Yes
## 10575   311
```

```
train$workingday <- factor(train$workingday, labels = c("No", "Yes"))
test$workingday <- factor(test$workingday, labels = c("No", "Yes"))

table(train$workingday)
```

```
##
##   No  Yes
## 3474 7412
```

```
train$weather <- factor(train$weather, labels = c("Great", "Good", "Average", "Poor"))
test$weather <- factor(test$weather, labels = c("Great", "Good", "Average", "Poor"))


# table(train$weather)
```

Split Date-Time (Both)

- Year, Month, Day and Hour

```
# library(lubridate)

train <- train %>%
  mutate(year = as.factor(format(datetime, format = "%Y")),
         month = as.numeric(format(datetime, format = "%m")),
         day = as.numeric(format(datetime, format = "%d")),
         hour = as.factor(format(datetime, format = "%H")))

test <- test %>%
  mutate(year = as.factor(format(datetime, format = "%Y")),
         month = as.numeric(format(datetime, format = "%m")),
         day = as.numeric(format(datetime, format = "%d")),
         hour = as.factor(format(datetime, format = "%H")))
```

Convert Months to Ordered Factor (Both)

```
train$month <-month(train$datetime, label = TRUE, abbr = FALSE)
test$month <-month(test$datetime, label = TRUE, abbr = FALSE)
```

```r
# need to convert the datetime column to a string for rbind function
train$datetime <-as.character(train$datetime)
test$datetime <-as.character(test$datetime)
```

Modeling

- psuedo code

- Loop through years (train and test)

- Loop through months (train and test)

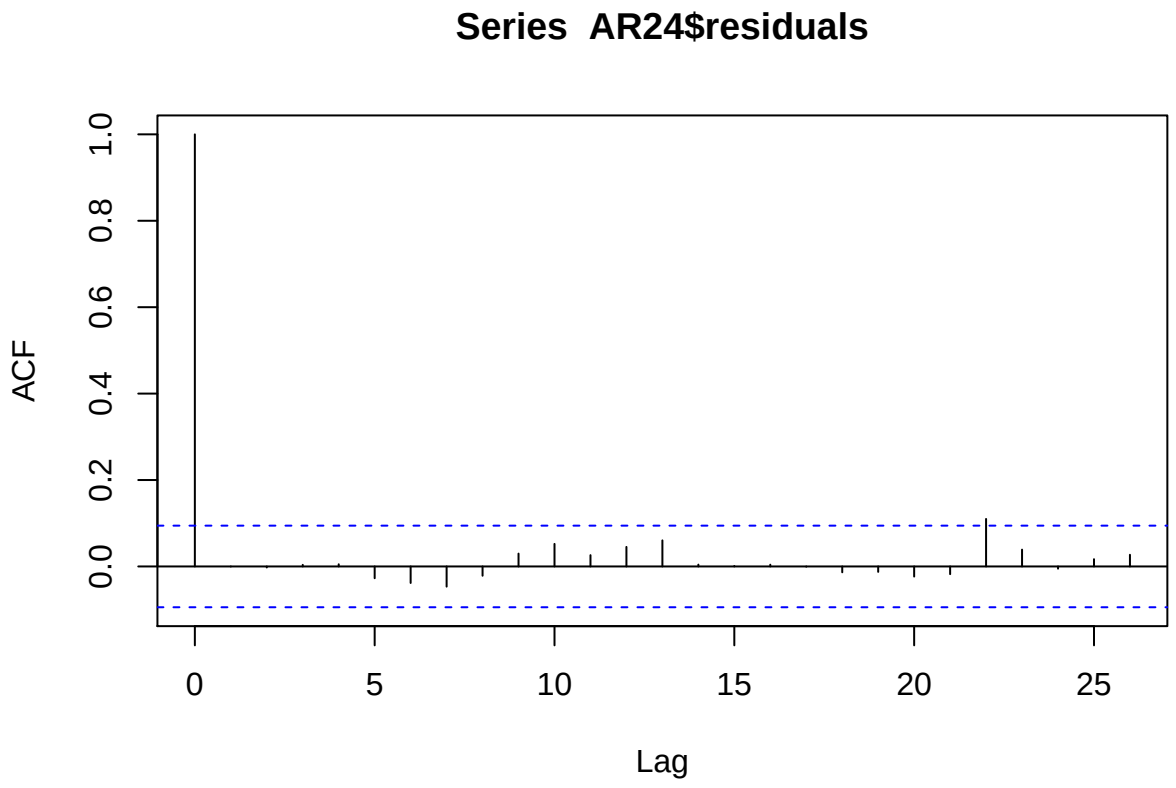
- fit AR model

- Forcast x number of observations based on nrow from test dataframe and impute the count from the time

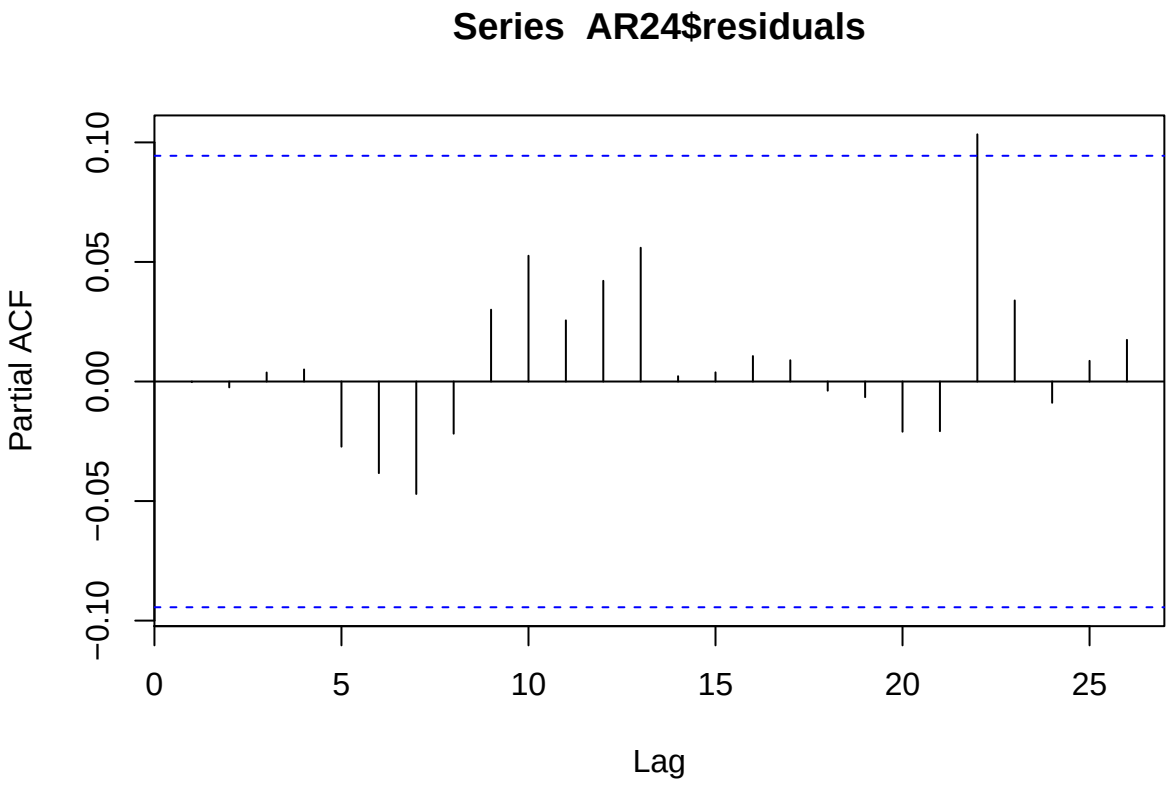
2011

January

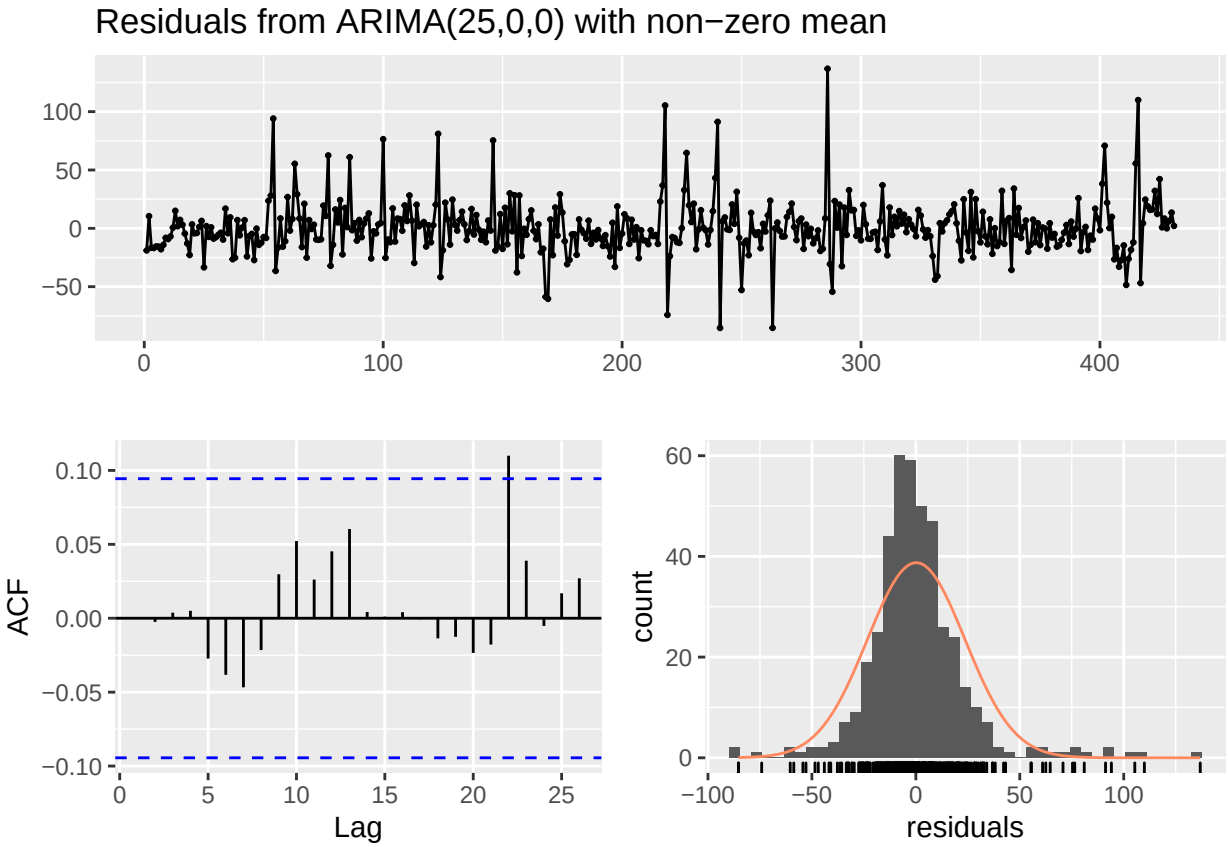
```r
train1 <- train %>%
  filter(year == '2011' & month == 'January') %>%
  select(datetime, count)

test1 <- test %>%
  filter(year == '2011' & month == 'January') %>%
  mutate(count = NA) %>%
  select(datetime, count)

# head(train1)
# head(test1)

AR24 <- arima(train1$count,order=c(25,0,0))
# tsdisplay(residuals(AR24),lag.max=25,main="AR(24) Resid. Diagnostics")

number = nrow(test1)

acf(AR24$residuals)
```




```
pacf(AR24$residuals)
```

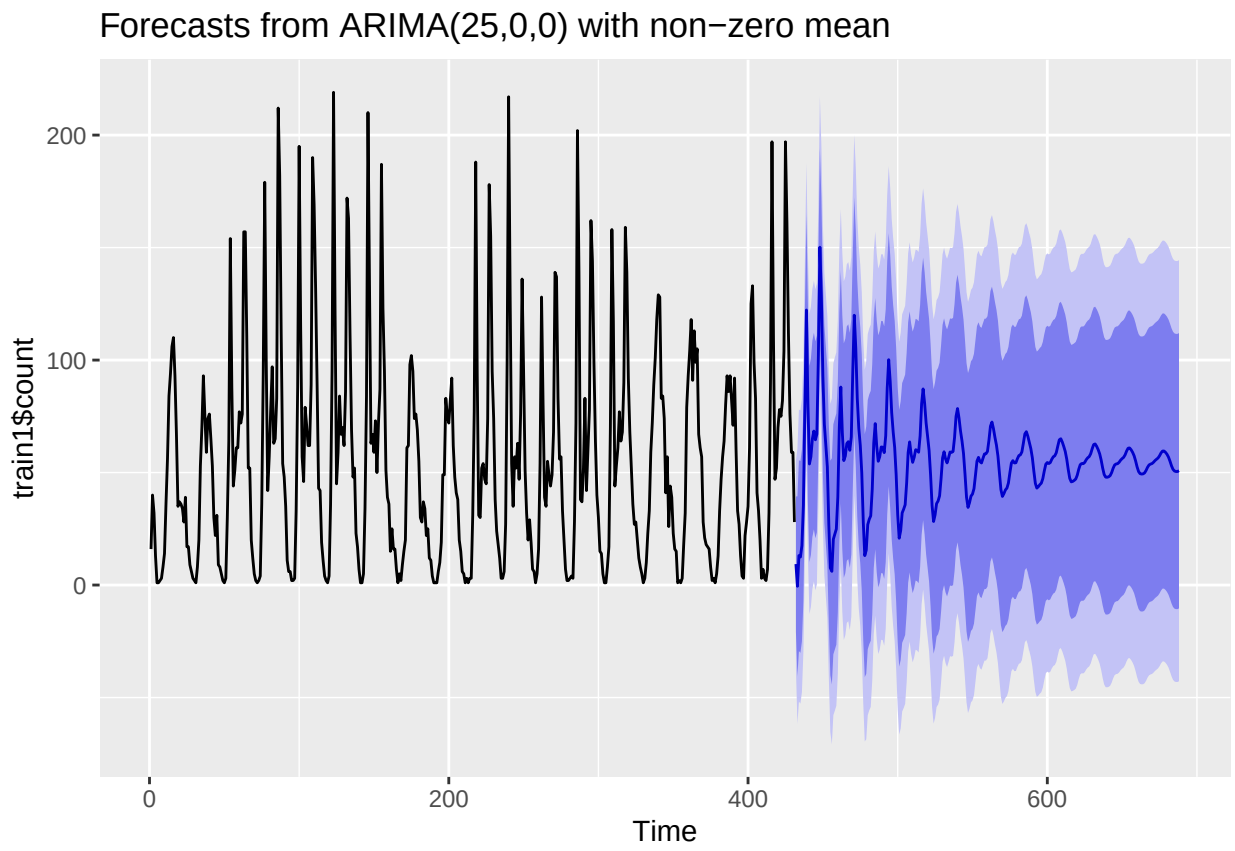



```
checkresiduals(AR24)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0) with non-zero mean
## Q* = 14.338, df = 3, p-value = 0.00248
##
## Model df: 26.   Total lags used: 29
```

```
fcst <- forecast(AR24, h=number)

autoplot(fcst)
```



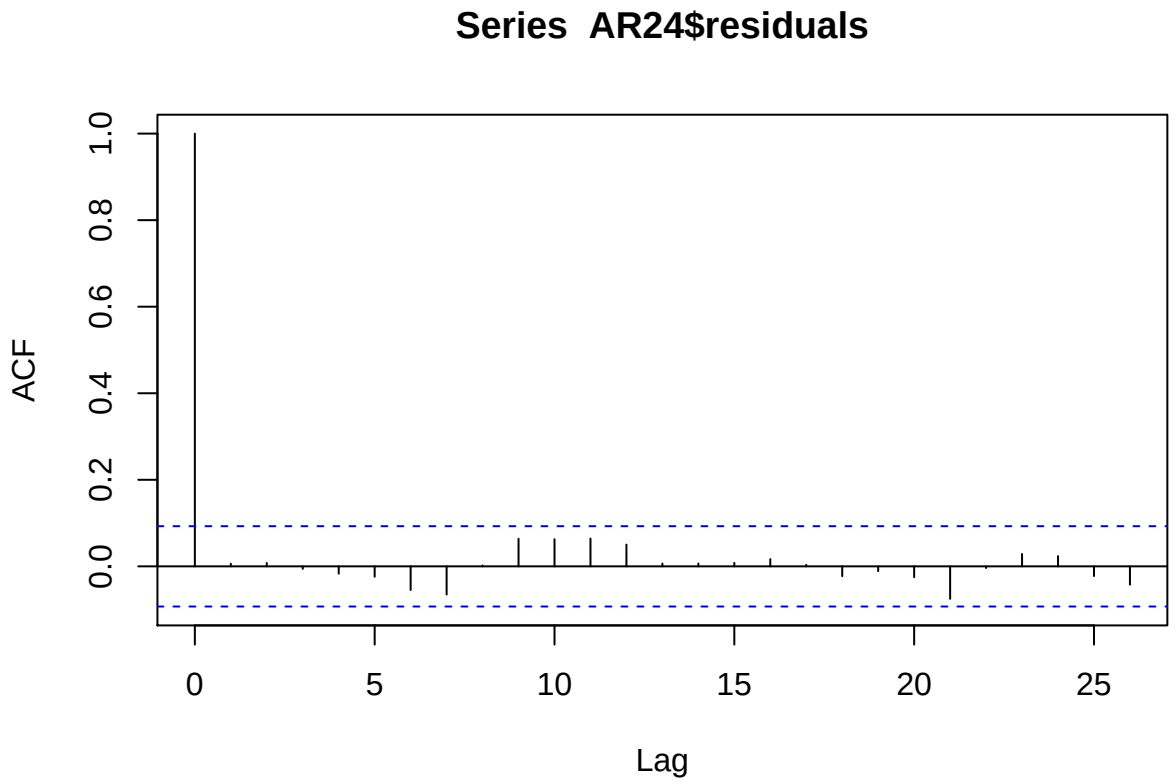

```
# point estimate (mean)
test1$count <- round(fcst$mean)
# test1

RMSLE(y_pred = floor(ifelse(fcst$fitted < 0, 0, round(fcst$fitted))), y_true = train1$count)
```

```
## [1] 0.7602042
```

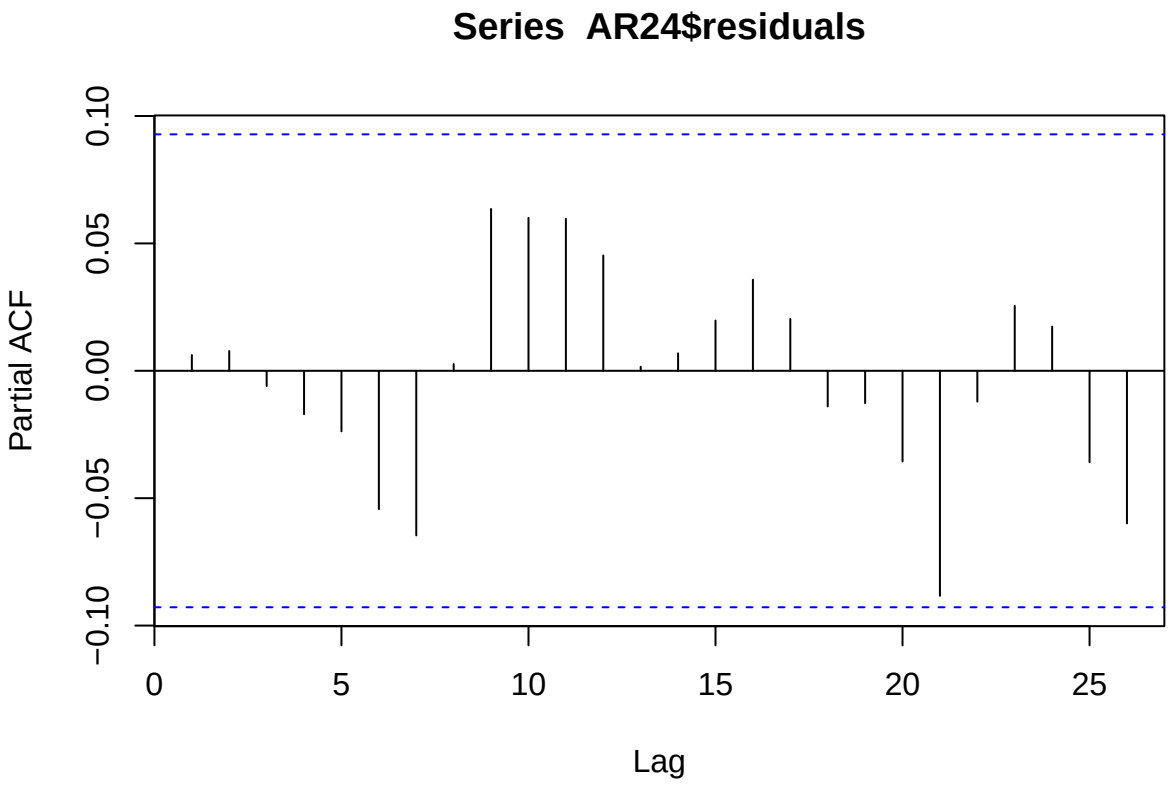
February

```
train2 <- train %>%
  filter(year == '2011' & month == 'February') %>%
  select(datetime, count)

test2 <- test %>%
  filter(year == '2011' & month == 'February') %>%
  mutate(count = NA) %>%
  select(datetime, count)

# head(train2)
# head(test2)

AR24 <- arima(train2$count,order=c(25,0,0))
# tsdisplay(residuals(AR24),lag.max=25,main="AR(24) Resid. Diagnostics")

number = nrow(test2)
```

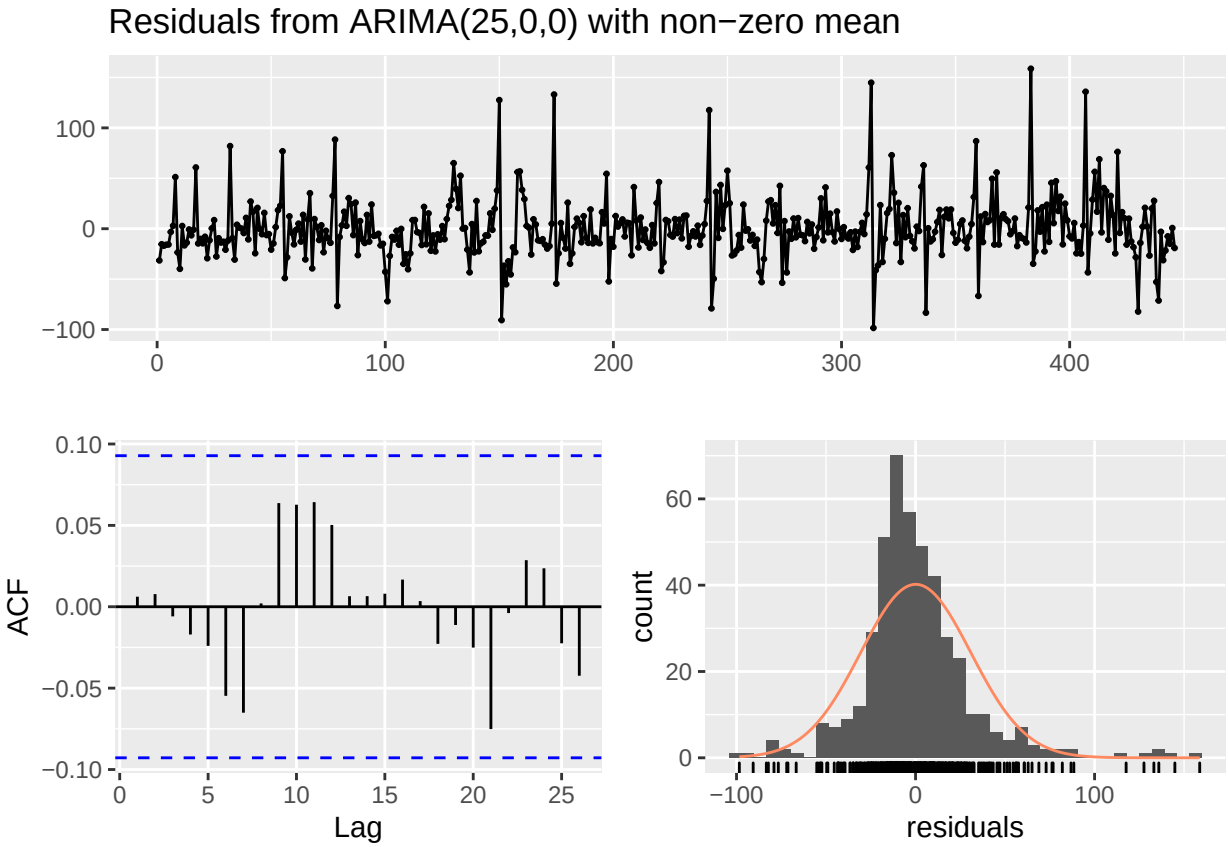
```r
acf(AR24$residuals)
```




```r
pacf(AR24$residuals)
```

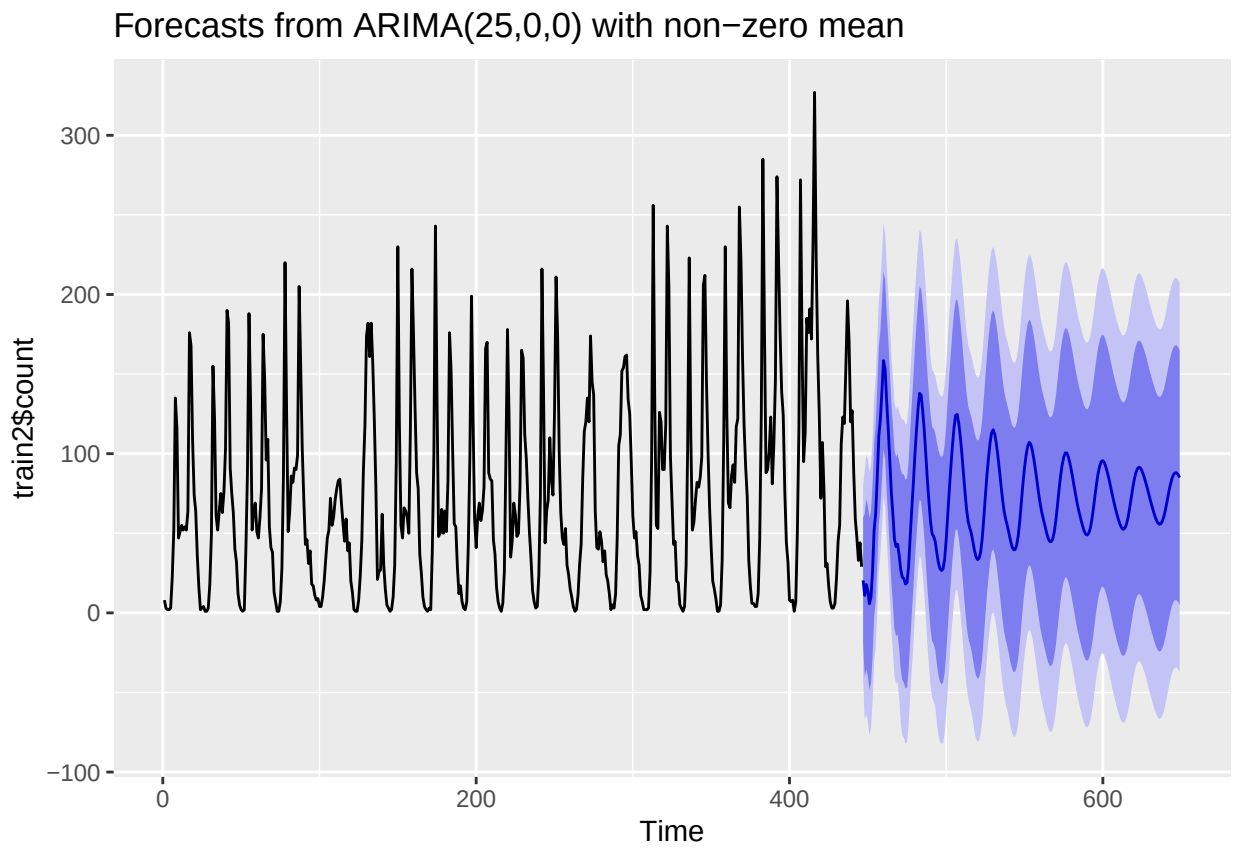



```
checkresiduals(AR24)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0) with non-zero mean
## Q* = 15.914, df = 3, p-value = 0.001181
##
## Model df: 26.   Total lags used: 29
```

```
fcst <- forecast(AR24, h=number)

autoplot(fcst)
```



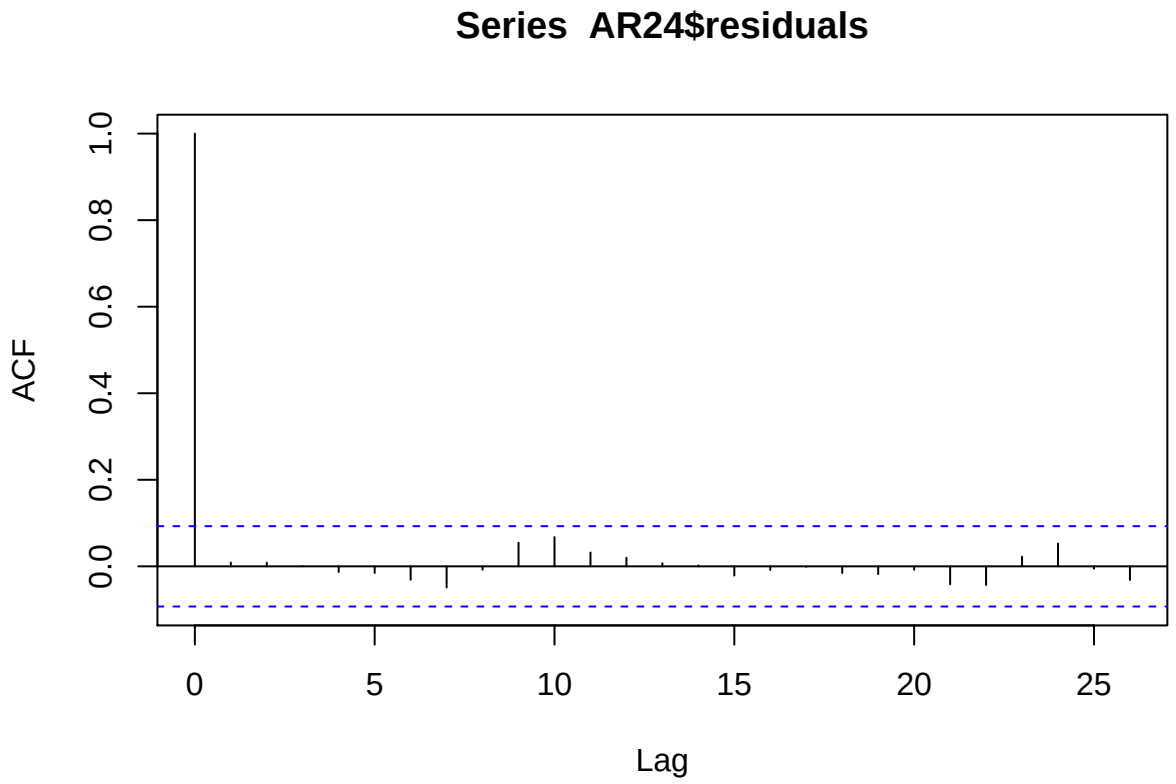

```
# point estimate (mean)
test2$count <- round(fcst$mean)


RMSLE(y_pred = floor(ifelse(fcst$fitted < 0, 0, round(fcst$fitted))), y_true = train2$count)
```

```
## [1] 0.7576135
```

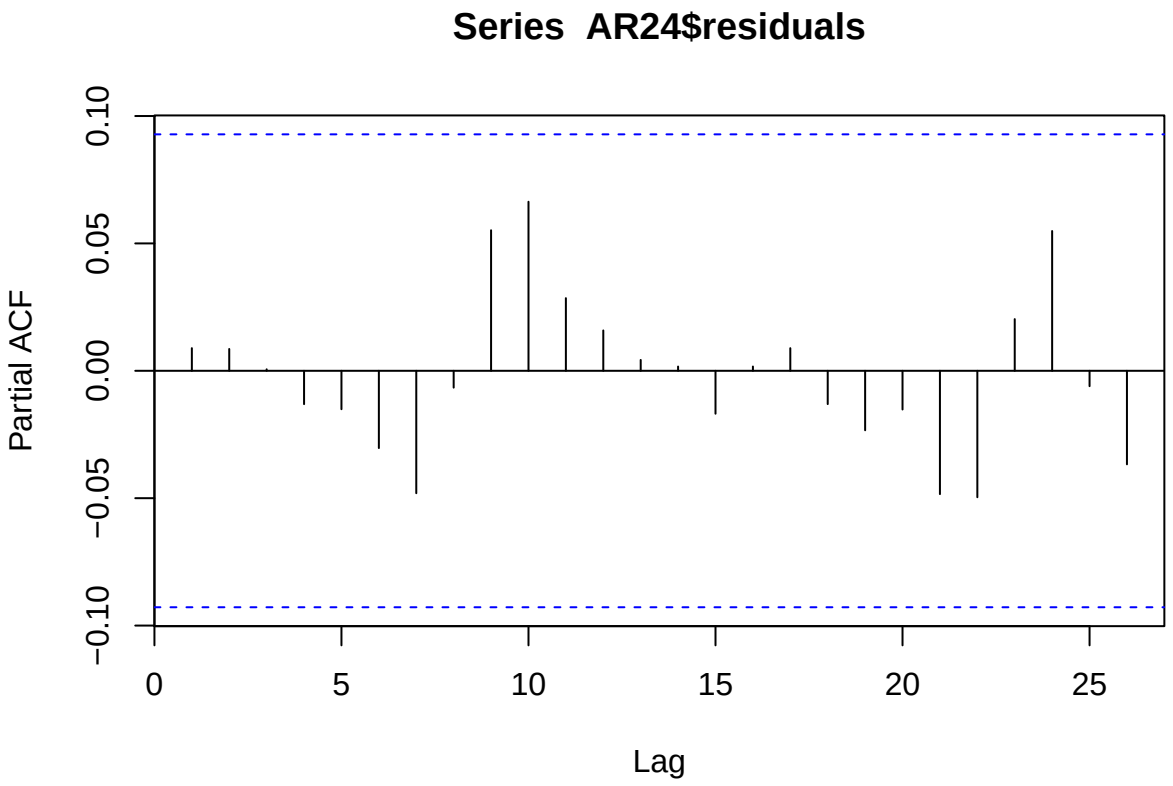
March

```
train3 <- train %>%
  filter(year == '2011' & month == 'March') %>%
  select(datetime, count)

test3 <- test %>%
  filter(year == '2011' & month == 'March') %>%
  mutate(count = NA) %>%
  select(datetime, count)

# head(train3)
# head(test3)

AR24 <- arima(train3$count,order=c(25,0,0))
# tsdisplay(residuals(AR24),lag.max=25,main="AR(24) Resid. Diagnostics")

number = nrow(test3)
```

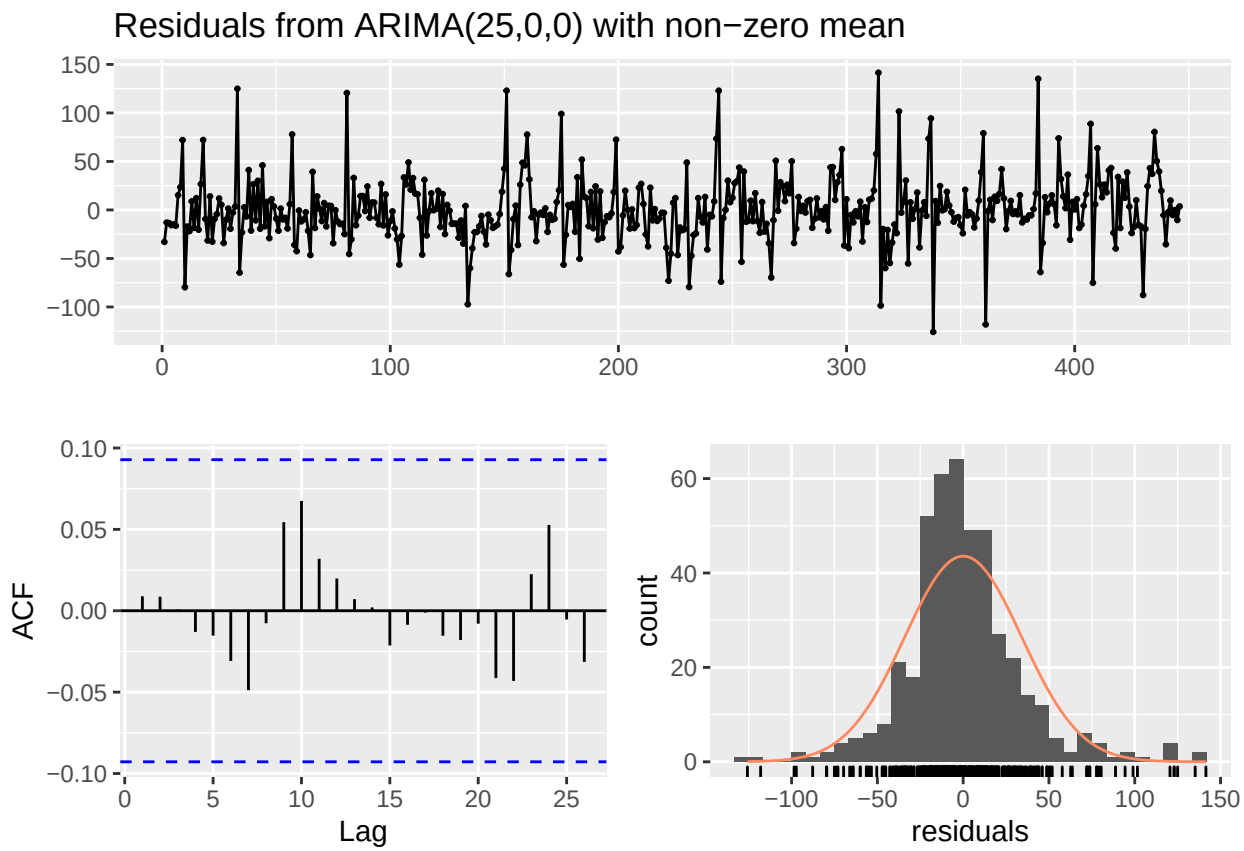
```r
acf(AR24$residuals)
```




```r
pacf(AR24$residuals)
```

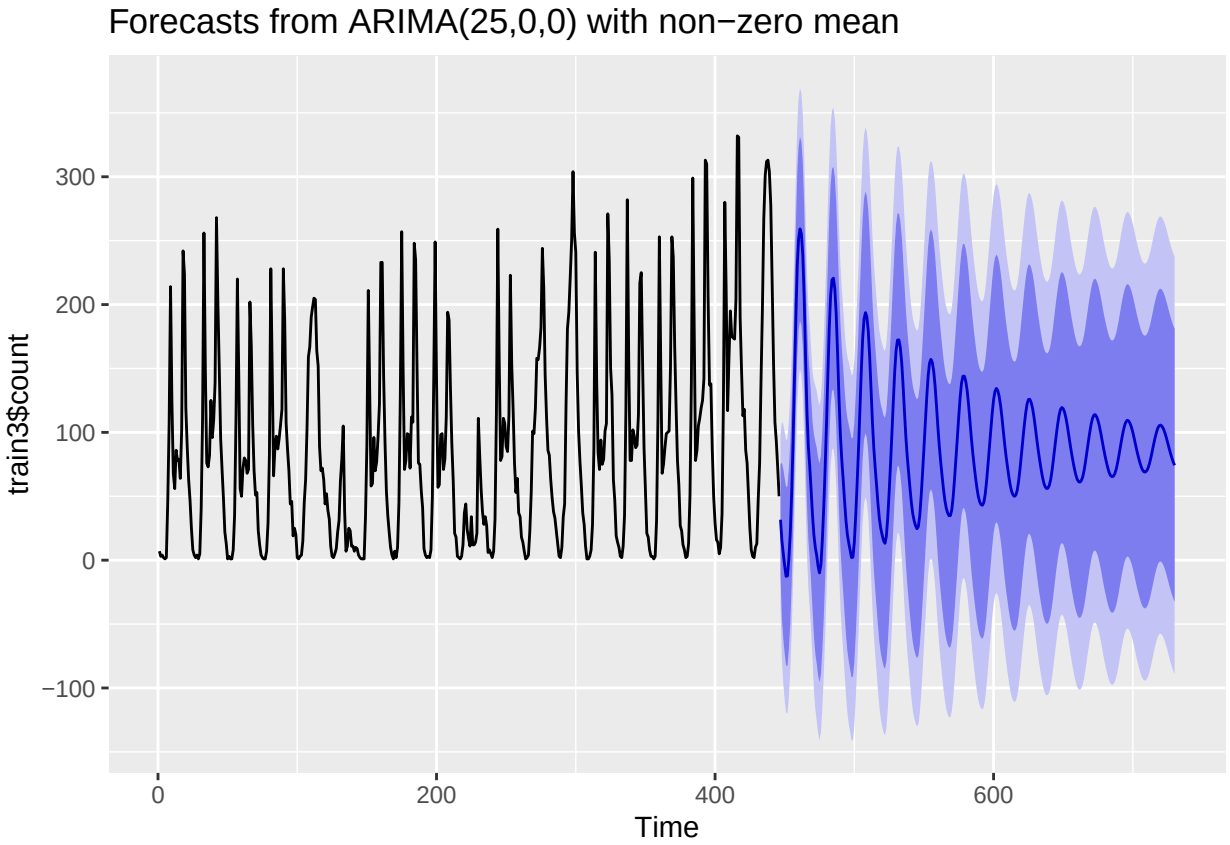



```
checkresiduals(AR24)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0) with non-zero mean
## Q* = 10.342, df = 3, p-value = 0.01587
##
## Model df: 26.   Total lags used: 29
```

```
fcst <- forecast(AR24, h=number)

autoplot(fcst)
```



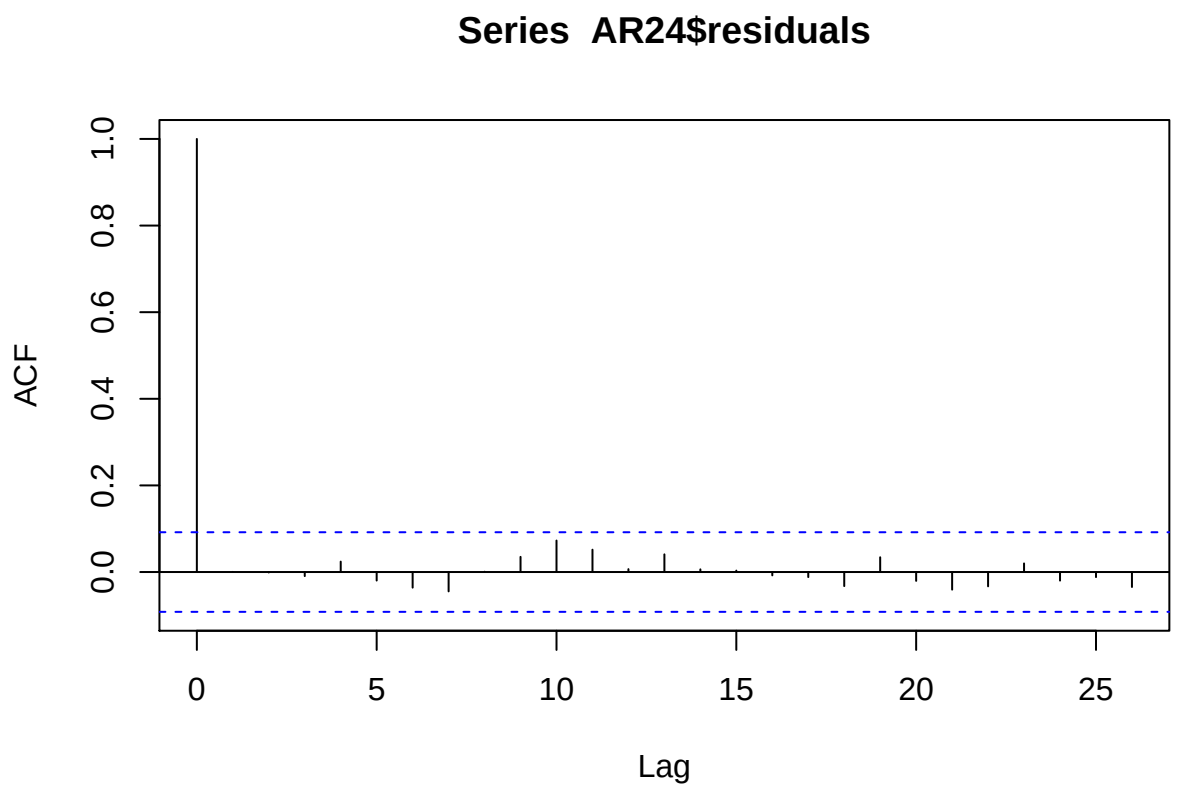

```
# point estimate (mean)
test3$count <- round(fcst$mean)

# test3

RMSLE(y_pred = floor(ifelse(fcst$fitted < 0, 0, round(fcst$fitted))), y_true = train3$count)
```

```
## [1] 0.8099865
```

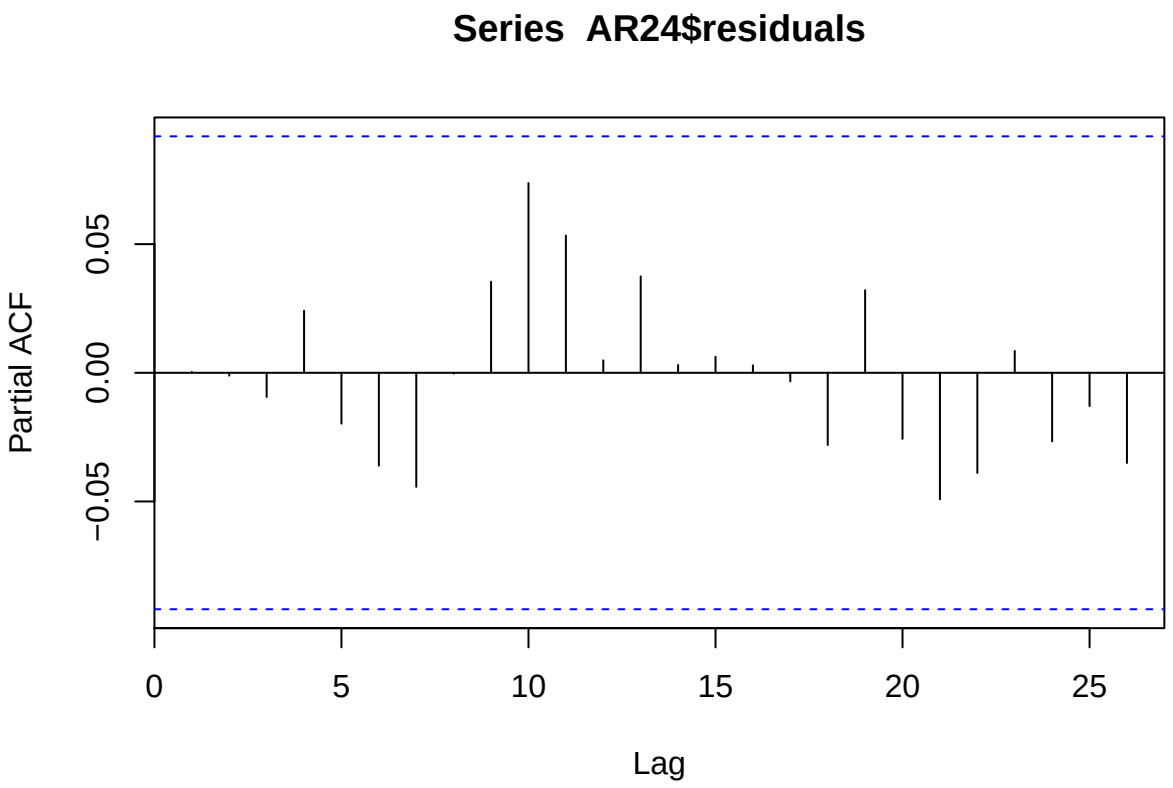
April

```
train4 <- train %>%
  filter(year == '2011' & month == 'April') %>%
  select(datetime, count)

test4 <- test %>%
  filter(year == '2011' & month == 'April') %>%
  mutate(count = NA) %>%
  select(datetime, count)

# head(train4)
# head(test4)

AR24 <- arima(train4$count,order=c(25,0,0))
# tsdisplay(residuals(AR24),lag.max=25,main="AR(24) Resid. Diagnostics")
```

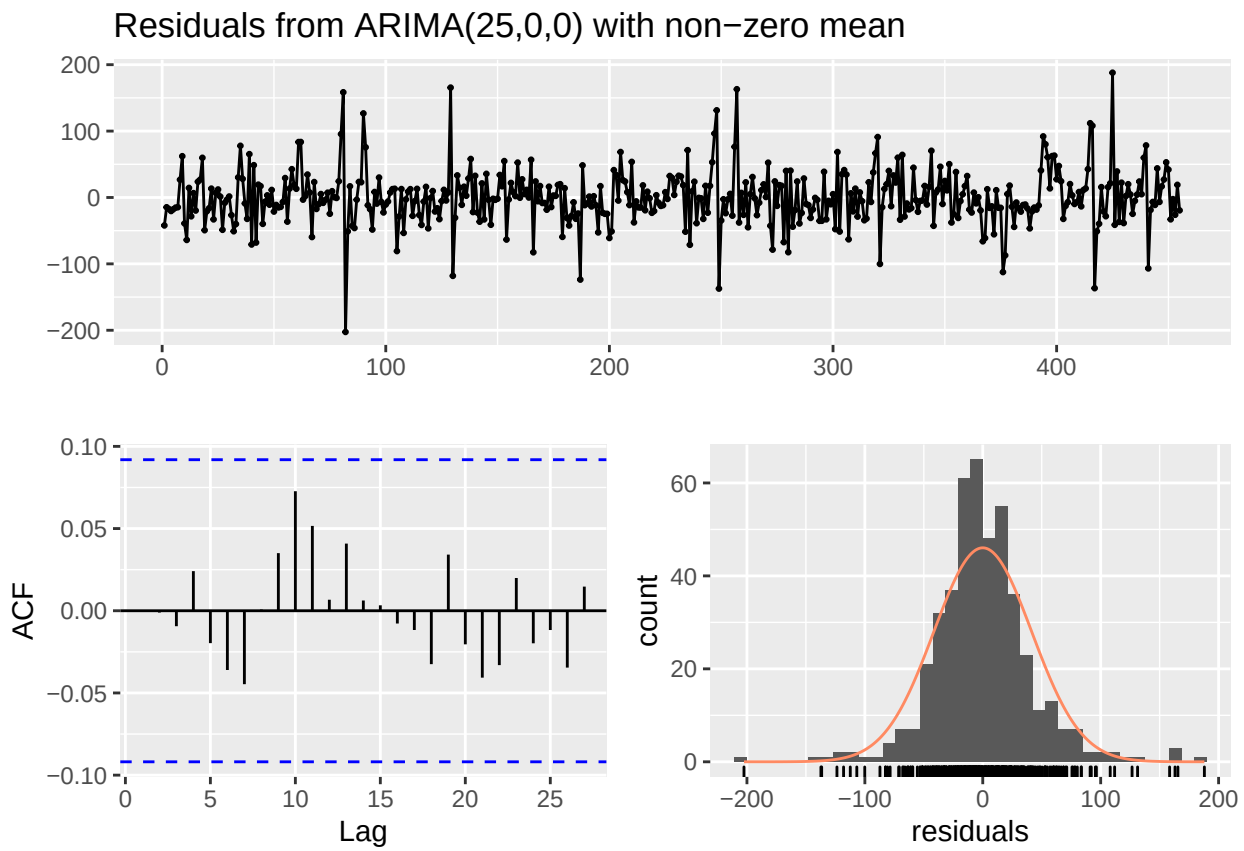
```r
number = nrow(test4)

acf(AR24$residuals)
```




```r
pacf(AR24$residuals)
```

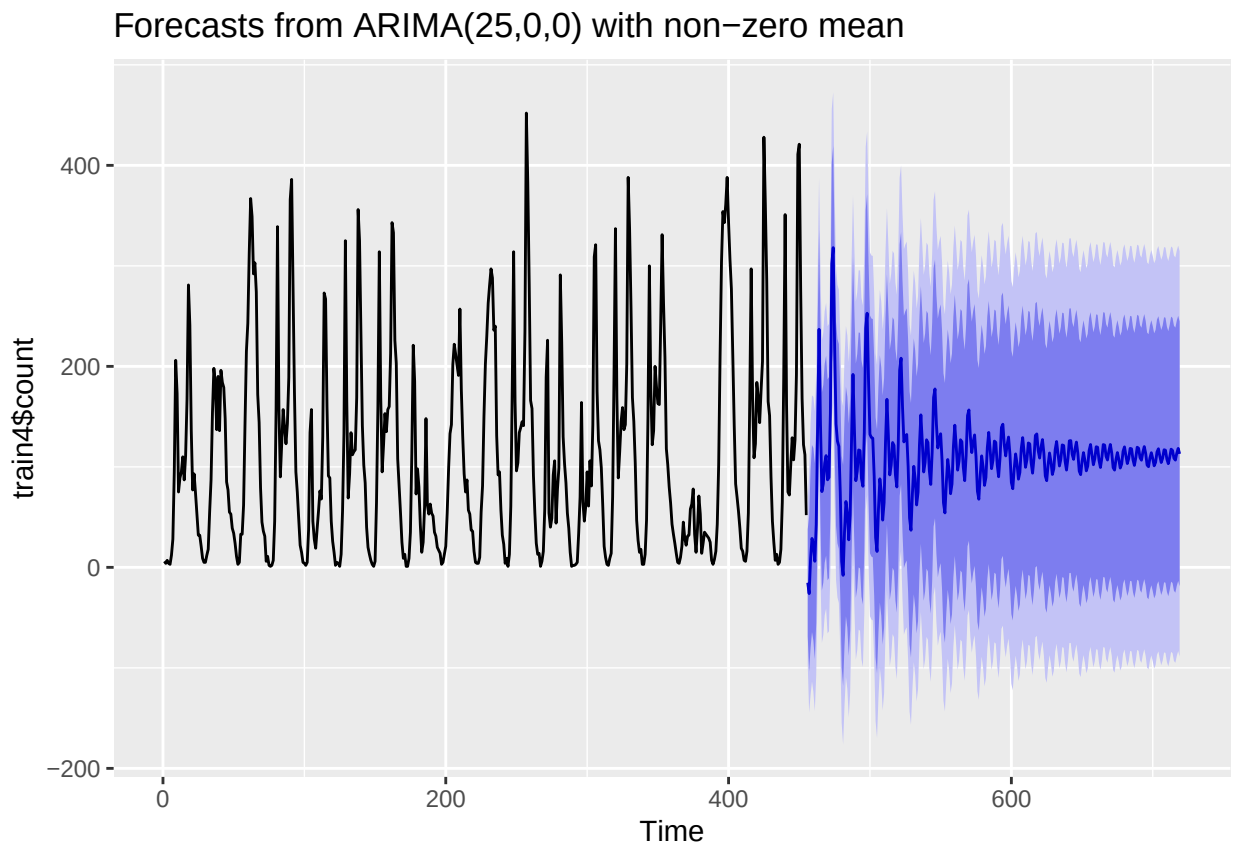



```
checkresiduals(AR24)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0) with non-zero mean
## Q* = 16.514, df = 3, p-value = 0.0008894
##
## Model df: 26.   Total lags used: 29
```

```
fcst <- forecast(AR24, h=number)

autoplot(fcst)
```



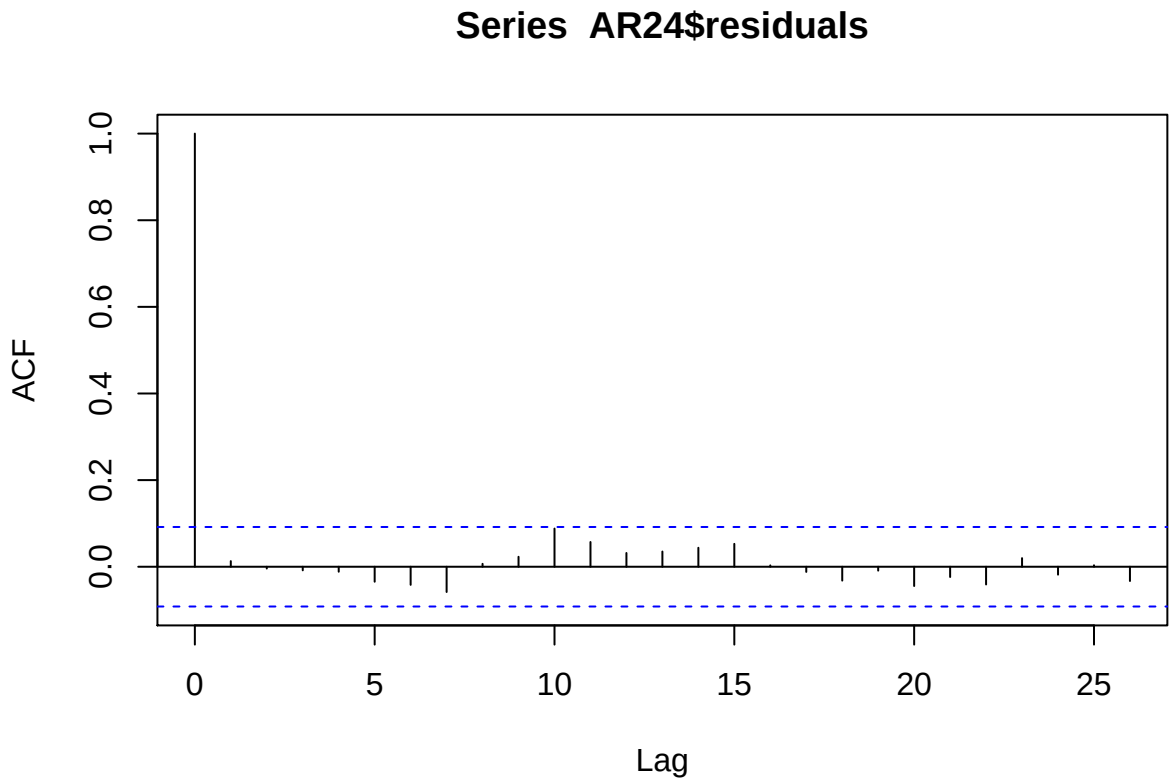

```
# point estimate (mean)
test4$count <- round(fcst$mean)


RMSLE(y_pred = floor(ifelse(fcst$fitted < 0, 0, round(fcst$fitted))), y_true = train4$count)
```

```
## [1] 0.8225528
```

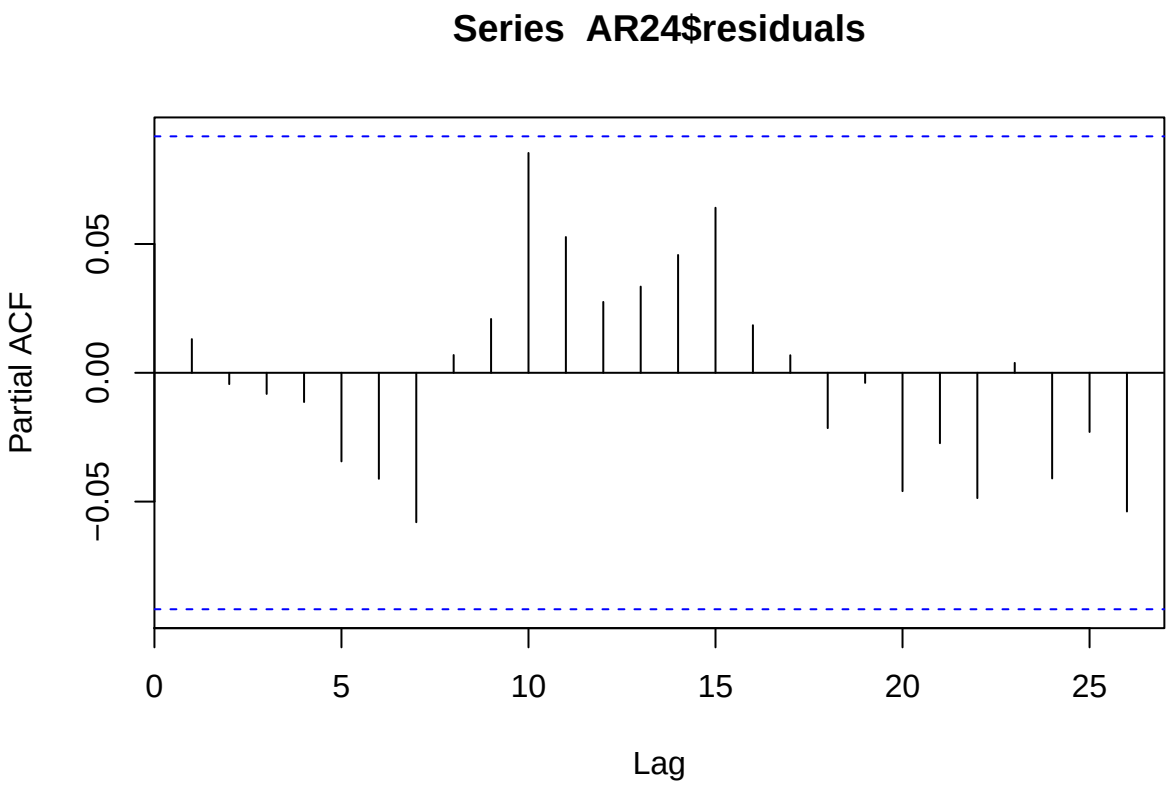
May

```
train5 <- train %>%
  filter(year == '2011' & month == 'May') %>%
  select(datetime, count)

test5 <- test %>%
  filter(year == '2011' & month == 'May') %>%
  mutate(count = NA) %>%
  select(datetime, count)

# head(train5)
# head(test5)

AR24 <- arima(train5$count,order=c(25,0,0))
# tsdisplay(residuals(AR24),lag.max=25,main="AR(24) Resid. Diagnostics")

number = nrow(test5)
```

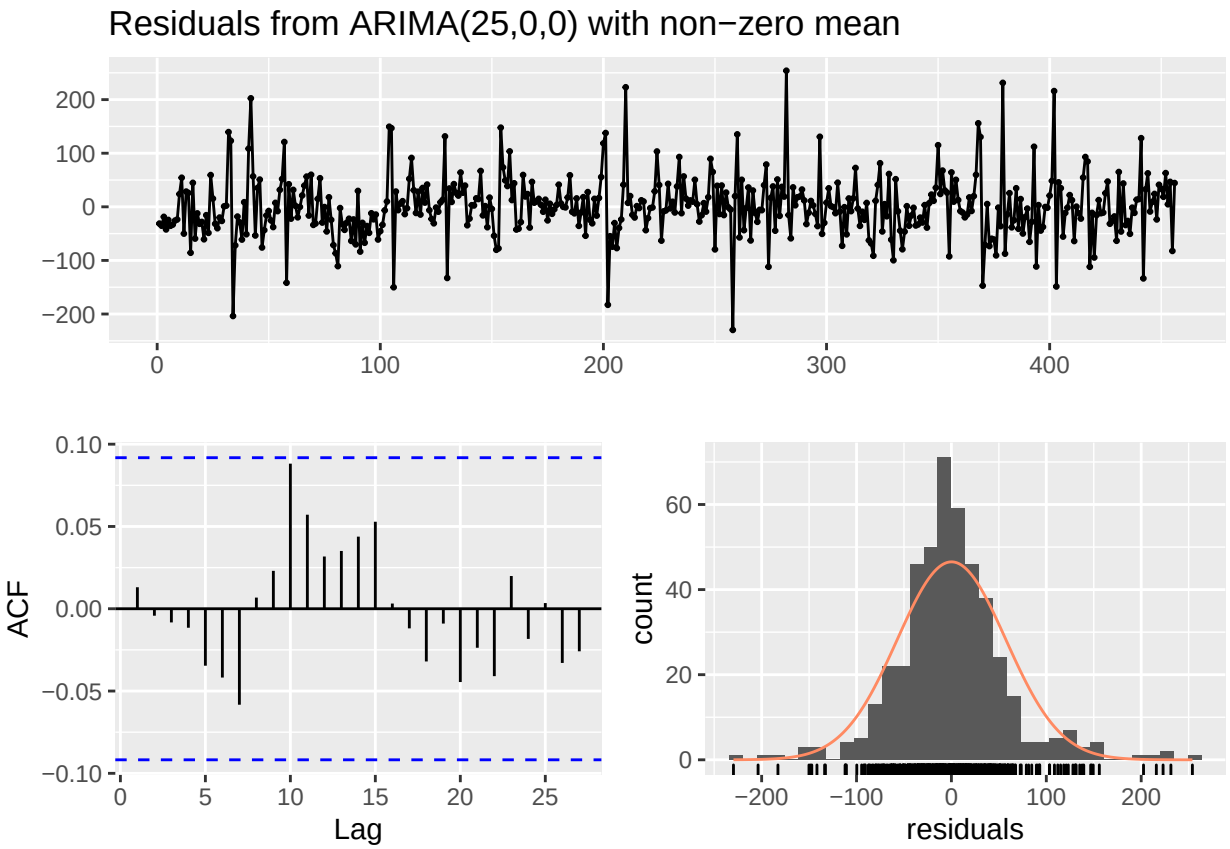
```r
acf(AR24$residuals)
```




```r
pacf(AR24$residuals)
```

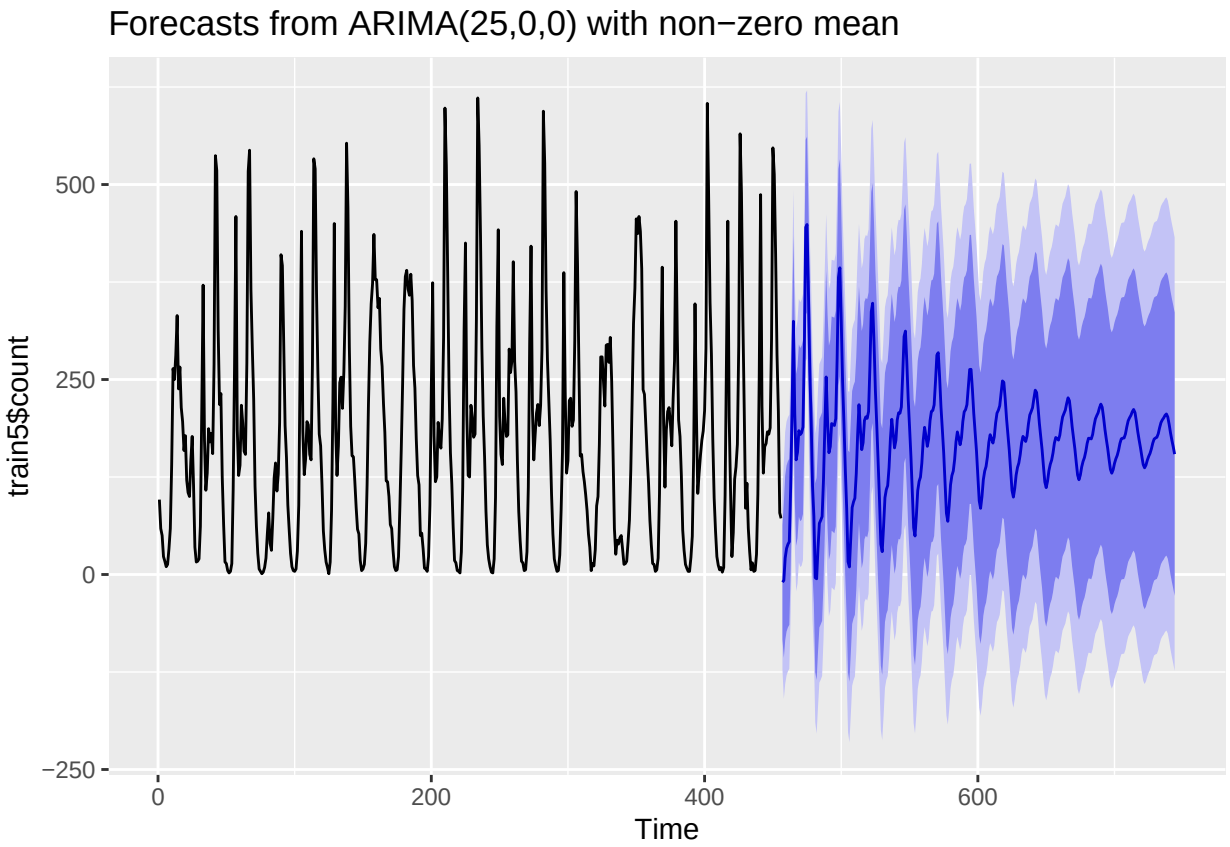



```
checkresiduals(AR24)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0) with non-zero mean
## Q* = 17.175, df = 3, p-value = 0.0006504
##
## Model df: 26.   Total lags used: 29
```

```
fcst <- forecast(AR24, h=number)

autoplot(fcst)
```



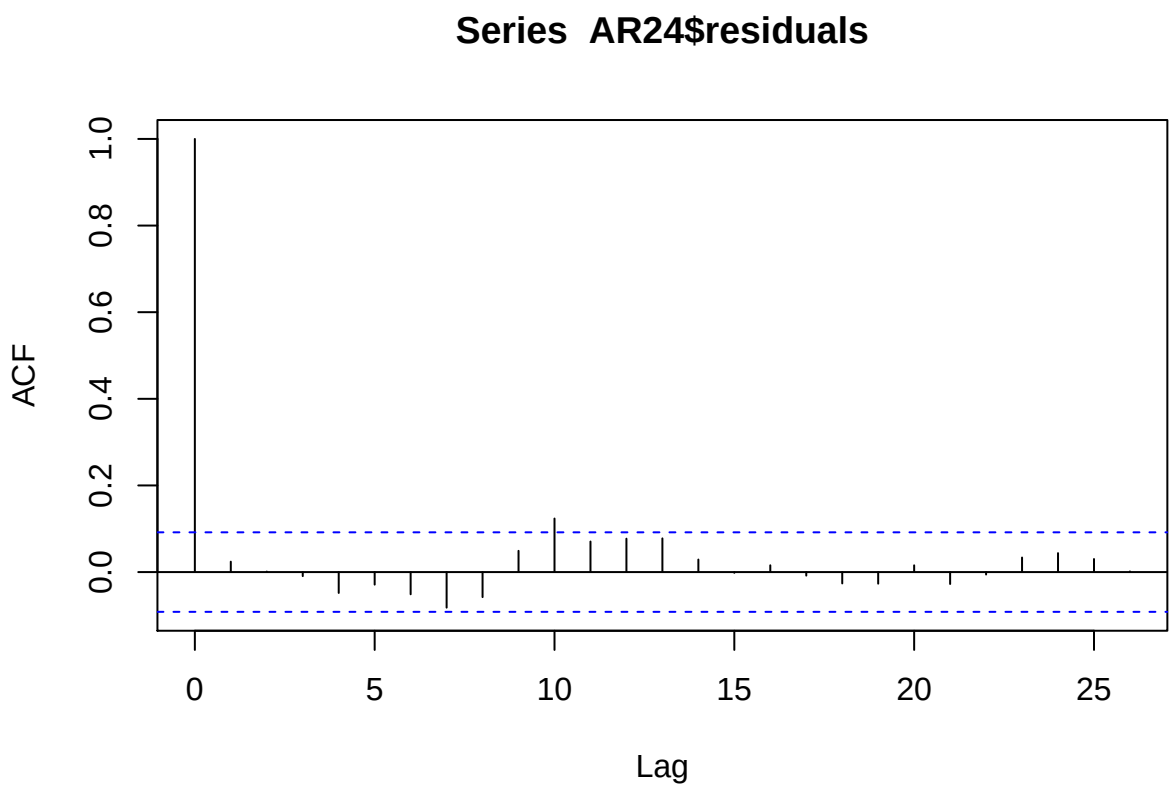

```
# point estimate (mean)
test5$count <- round(fcst$mean)

# test5

RMSLE(y_pred = floor(ifelse(fcst$fitted < 0, 0, round(fcst$fitted))), y_true = train5$count)
```

```
## [1] 0.7820472
```

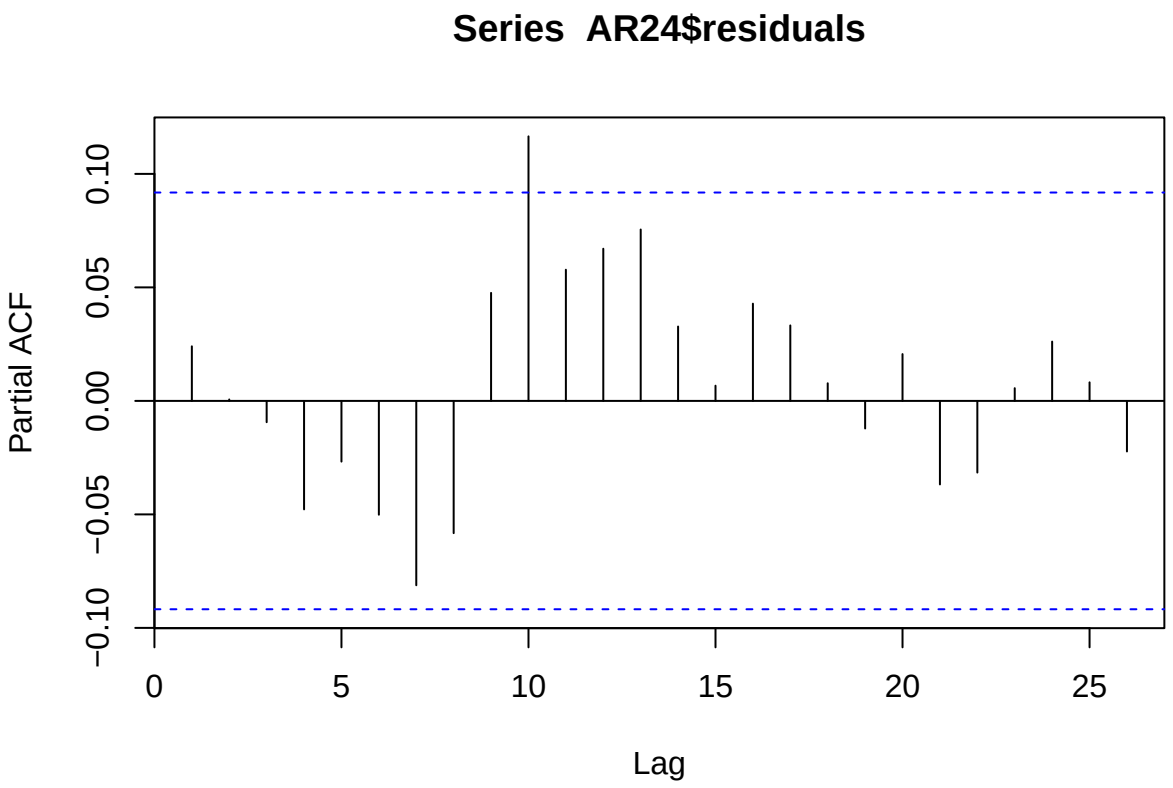
June

```
train6 <- train %>%
  filter(year == '2011' & month == 'June') %>%
  select(datetime, count)

test6 <- test %>%
  filter(year == '2011' & month == 'June') %>%
  mutate(count = NA) %>%
  select(datetime, count)

# head(train6)
# head(test6)

AR24 <- arima(train6$count,order=c(25,0,0))
# tsdisplay(residuals(AR24),lag.max=25,main="AR(24) Resid. Diagnostics")
```

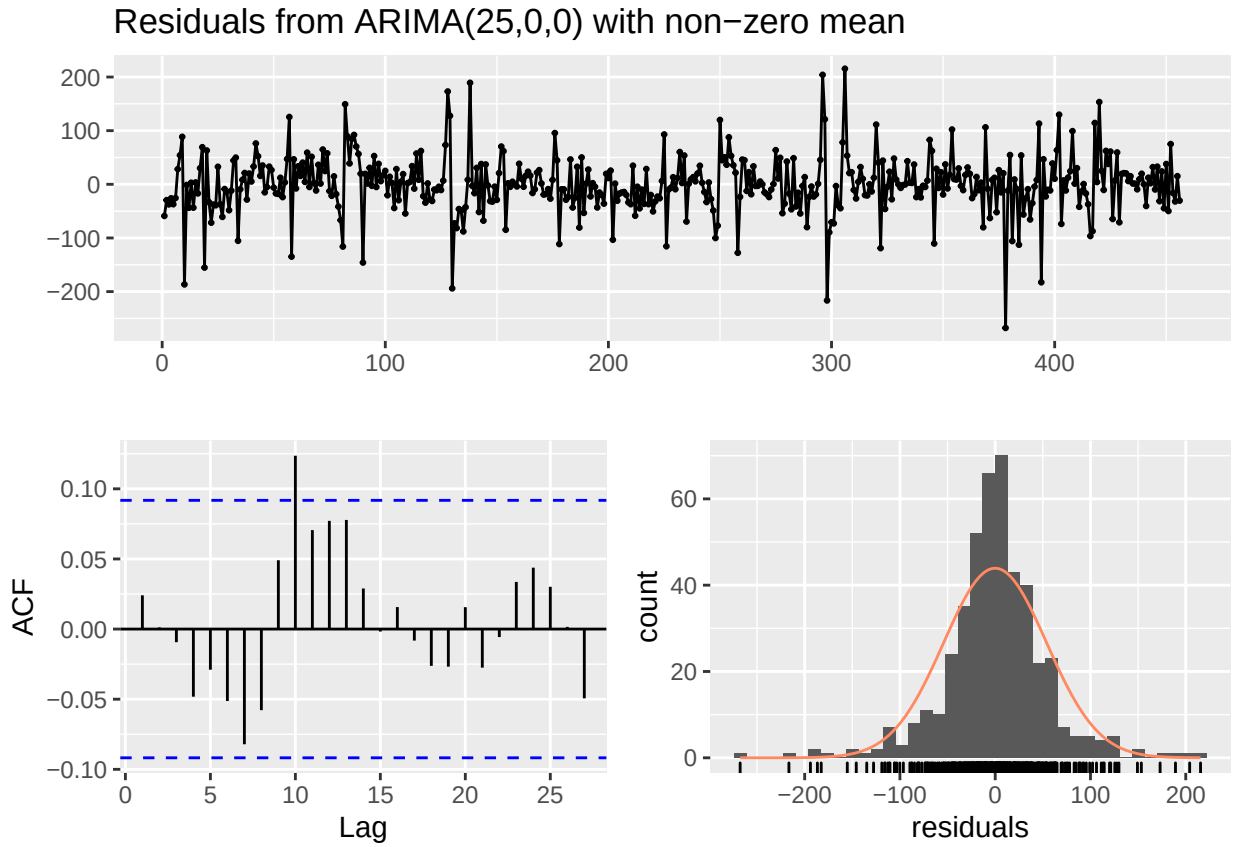
```r
number = nrow(test6)

acf(AR24$residuals)
```




```r
pacf(AR24$residuals)
```

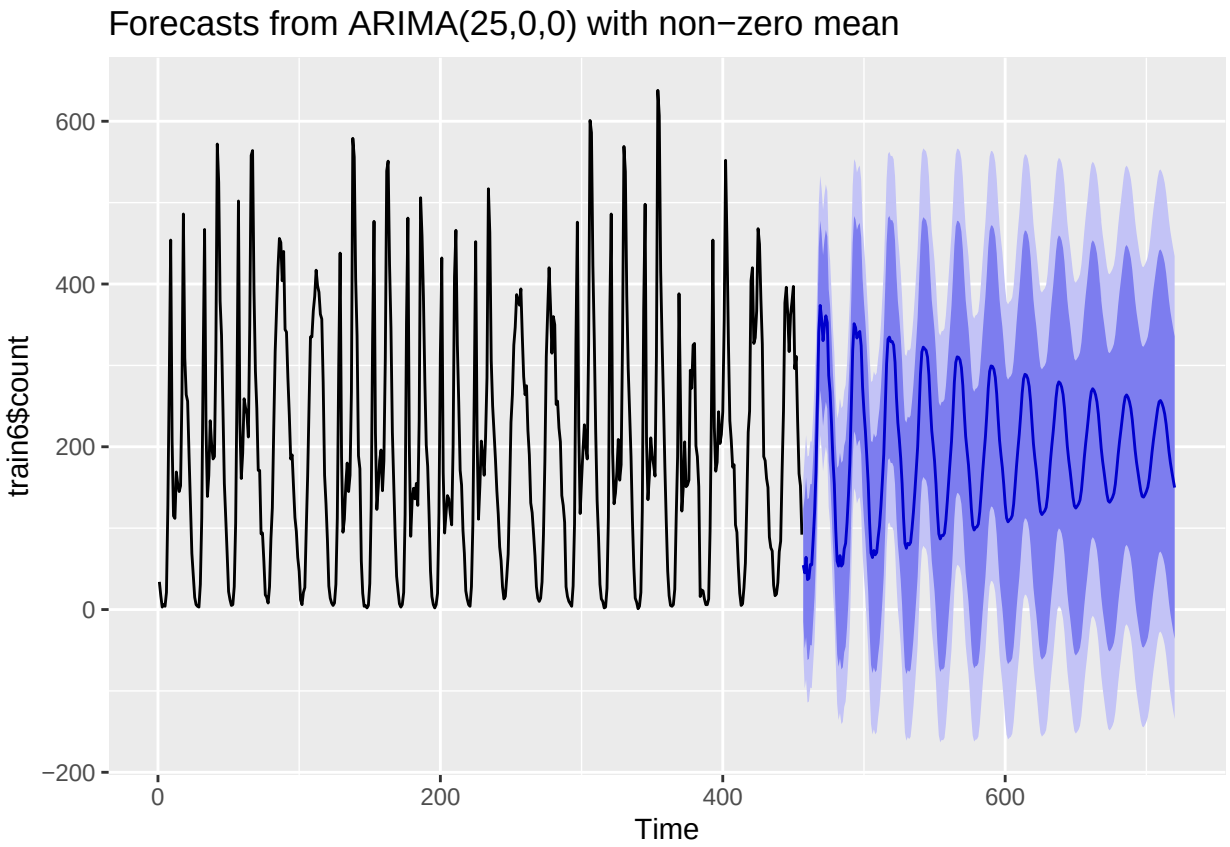



```
checkresiduals(AR24)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0) with non-zero mean
## Q* = 29.762, df = 3, p-value = 1.549e-06
##
## Model df: 26.   Total lags used: 29
```

```
fcst <- forecast(AR24, h=number)

autoplot(fcst)
```



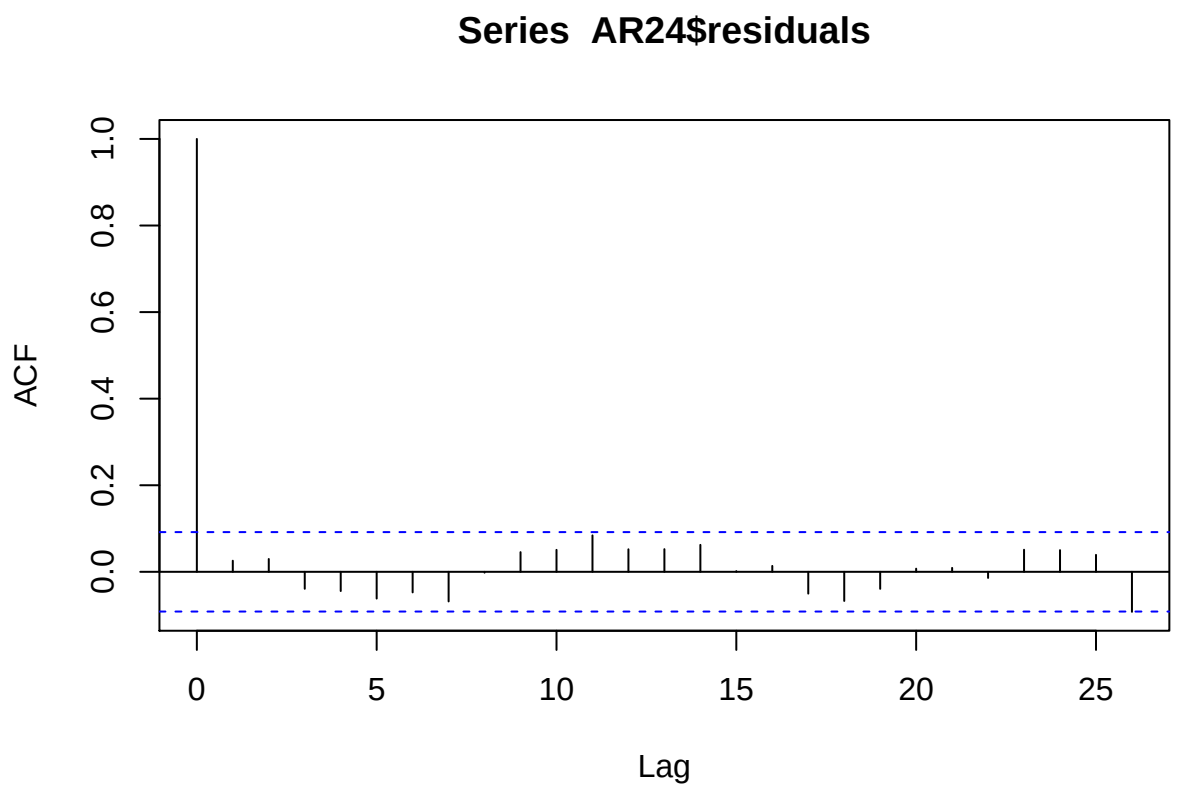

```
# point estimate (mean)
test6$count <- round(fcst$mean)

# test6

RMSLE(y_pred = floor(ifelse(fcst$fitted < 0, 0, round(fcst$fitted))), y_true = train6$count)
```

```
## [1] 0.6827719
```

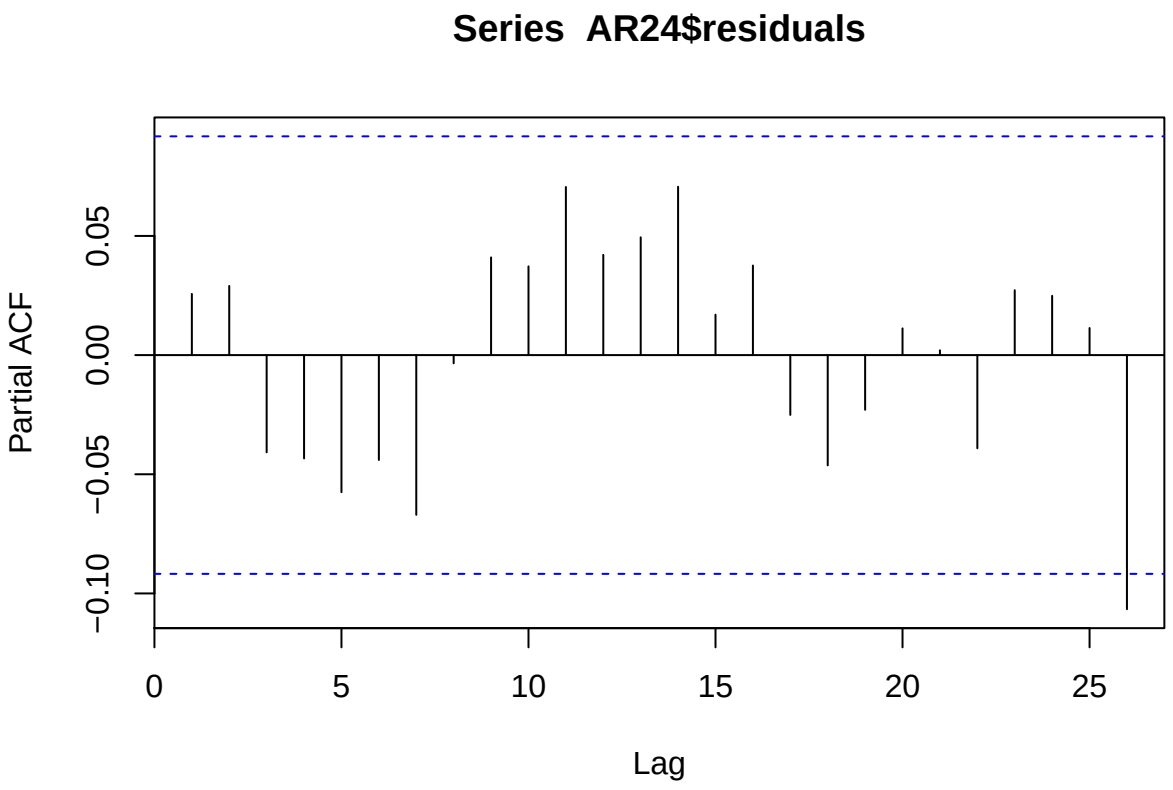
July

```
train7 <- train %>%
  filter(year == '2011' & month == 'July') %>%
  select(datetime, count)

test7 <- test %>%
  filter(year == '2011' & month == 'July') %>%
  mutate(count = NA) %>%
  select(datetime, count)

# head(train7)
# head(test7)

AR24 <- arima(train7$count,order=c(25,0,0))
# tsdisplay(residuals(AR24),lag.max=25,main="AR(24) Resid. Diagnostics")
```

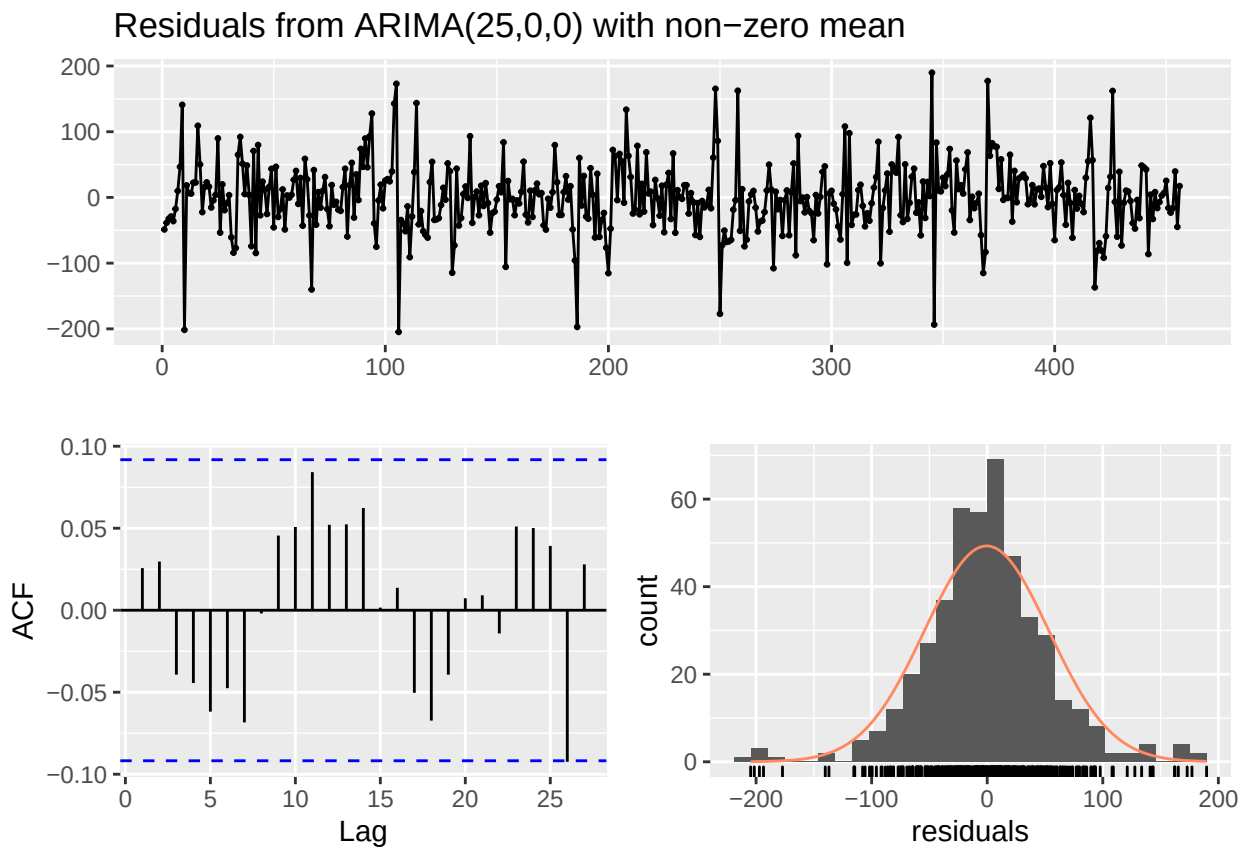
```r
number = nrow(test7)

acf(AR24$residuals)
```




```r
pacf(AR24$residuals)
```

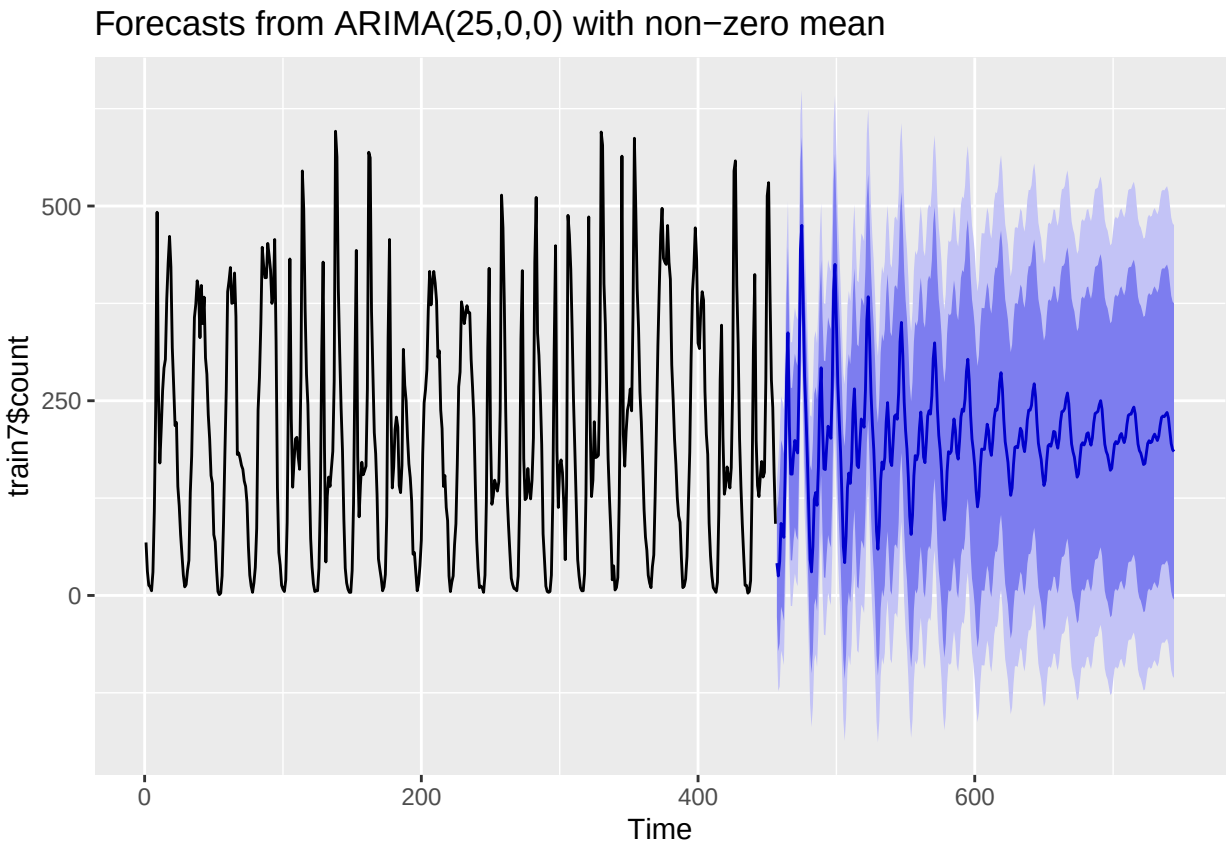



```
checkresiduals(AR24)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0) with non-zero mean
## Q* = 29.837, df = 3, p-value = 1.494e-06
##
## Model df: 26.   Total lags used: 29
```

```
fcst <- forecast(AR24, h=number)

autoplot(fcst)
```



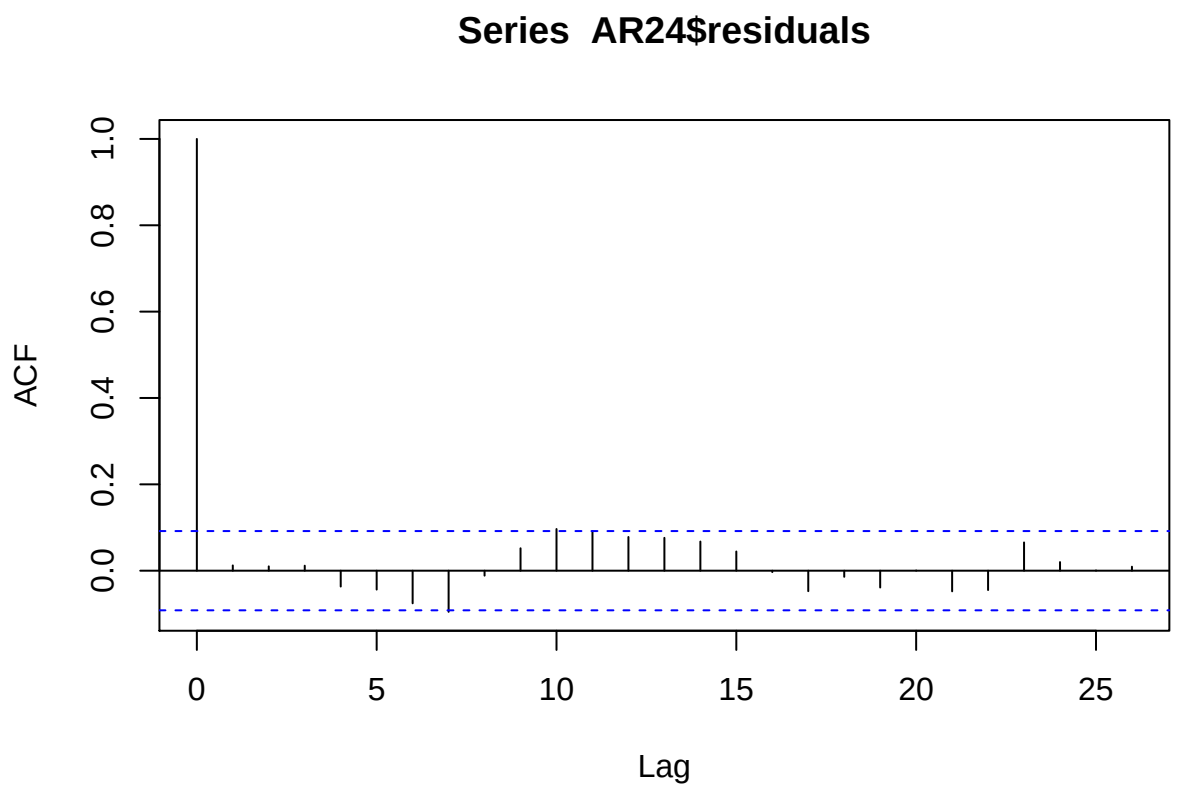

```
# point estimate (mean)
test7$count <- round(fcst$mean)

# test7

RMSLE(y_pred = floor(ifelse(fcst$fitted < 0, 0, round(fcst$fitted))), y_true = train7$count)
```

```
## [1] 0.6671488
```

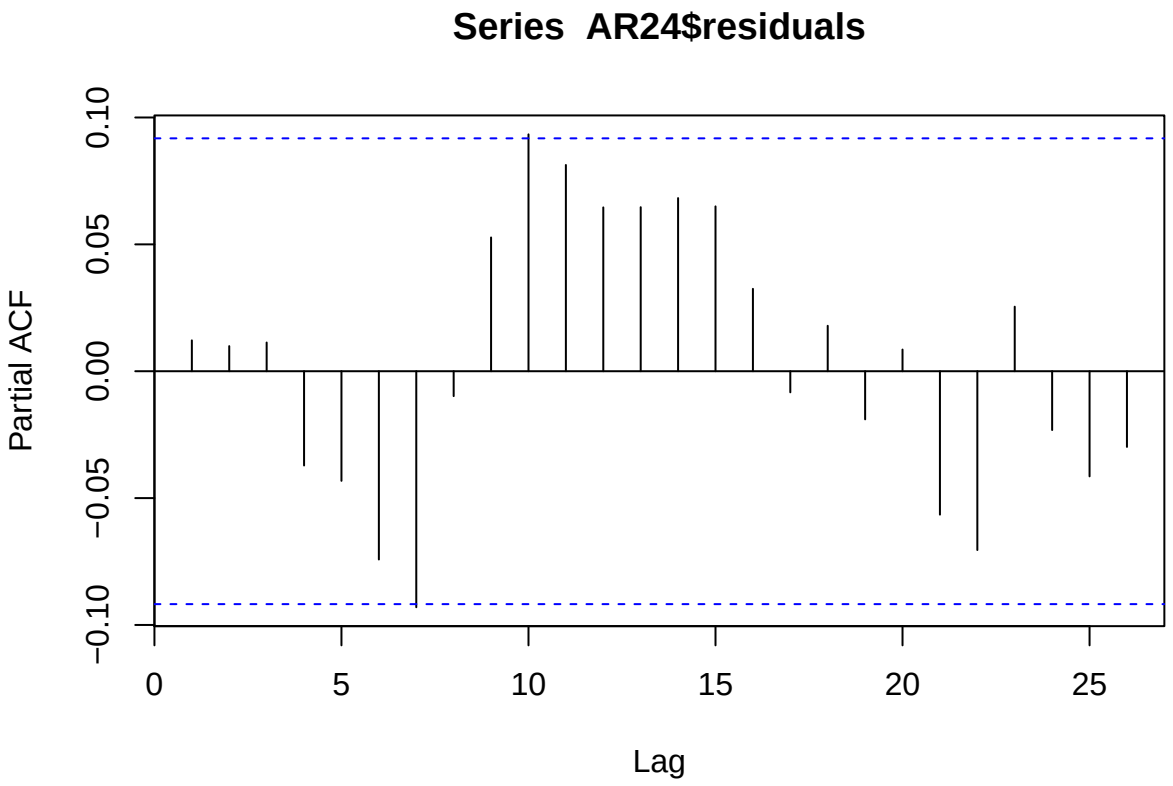
August

```
train8 <- train %>%
  filter(year == '2011' & month == 'August') %>%
  select(datetime, count)

test8 <- test %>%
  filter(year == '2011' & month == 'August') %>%
  mutate(count = NA) %>%
  select(datetime, count)

# head(train8)
# head(test8)

AR24 <- arima(train8$count,order=c(25,0,0))
# tsdisplay(residuals(AR24),lag.max=25,main="AR(24) Resid. Diagnostics")
```

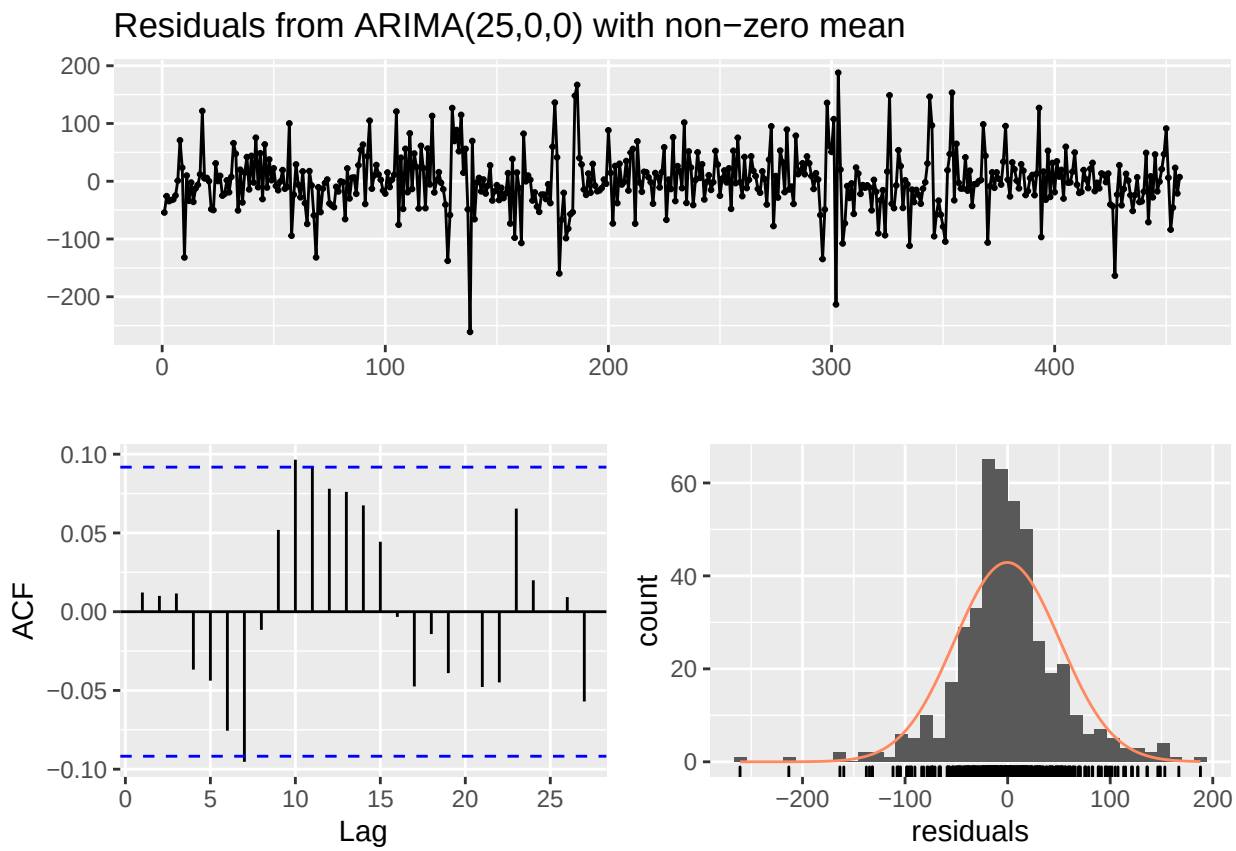
```r
number = nrow(test8)

acf(AR24$residuals)
```




```r
pacf(AR24$residuals)
```

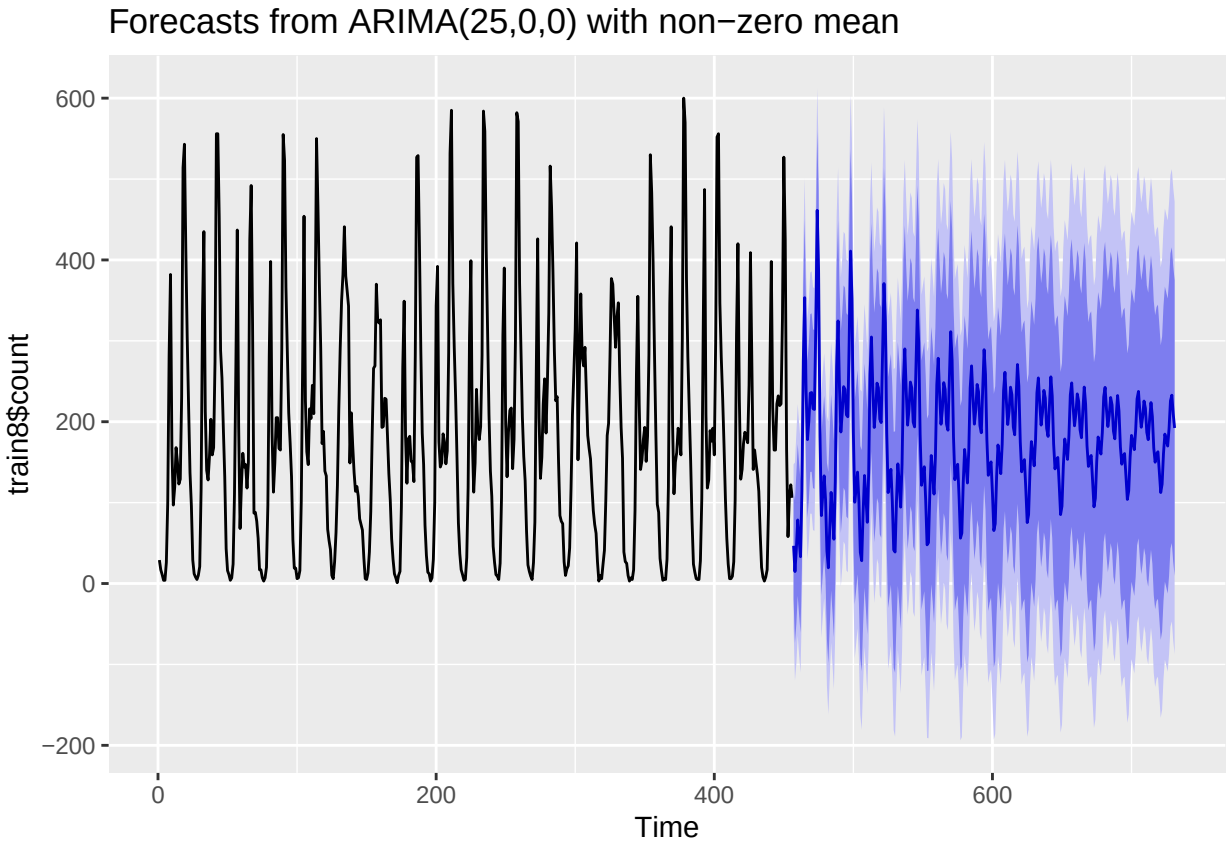



```
checkresiduals(AR24)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0) with non-zero mean
## Q* = 37.588, df = 3, p-value = 3.455e-08
##
## Model df: 26.   Total lags used: 29
```

```
fcst <- forecast(AR24, h=number)

autoplot(fcst)
```



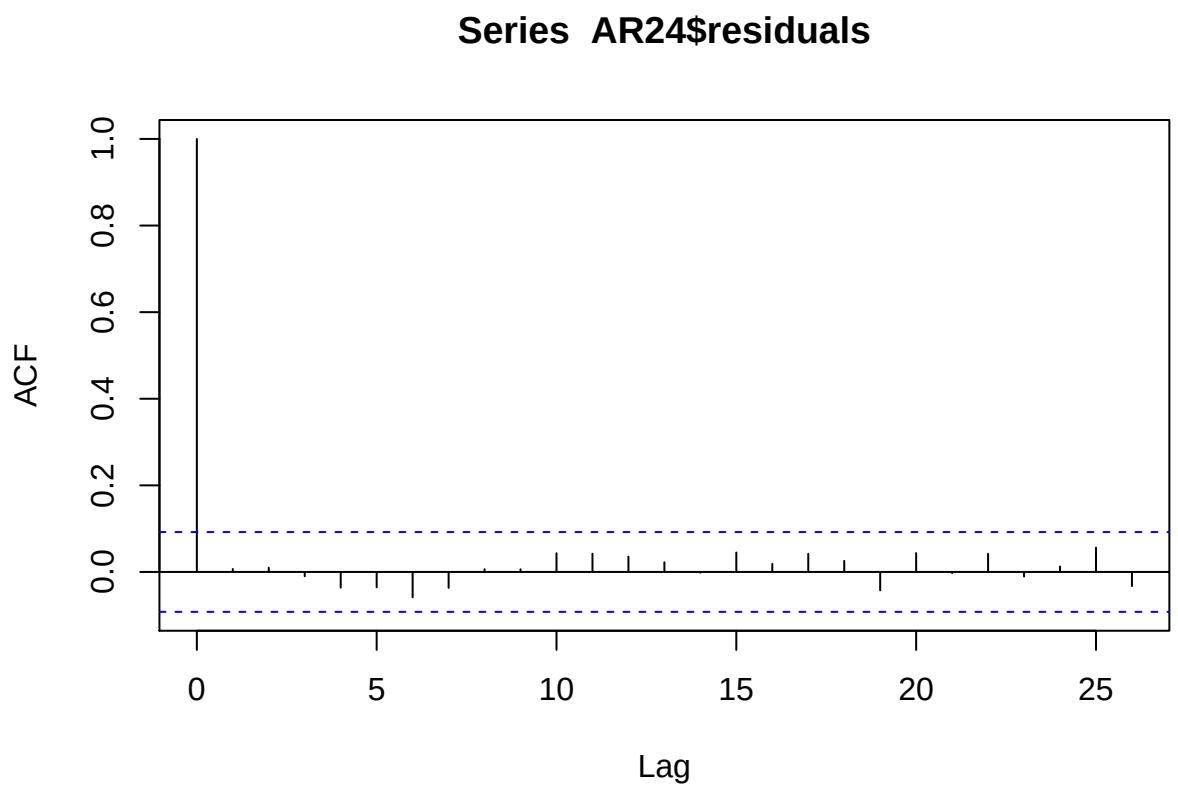

```
# point estimate (mean)
test8$count <- round(fcst$mean)

# test8

RMSLE(y_pred = floor(ifelse(fcst$fitted < 0, 0, round(fcst$fitted))), y_true = train8$count)
```

```
## [1] 0.7082002
```

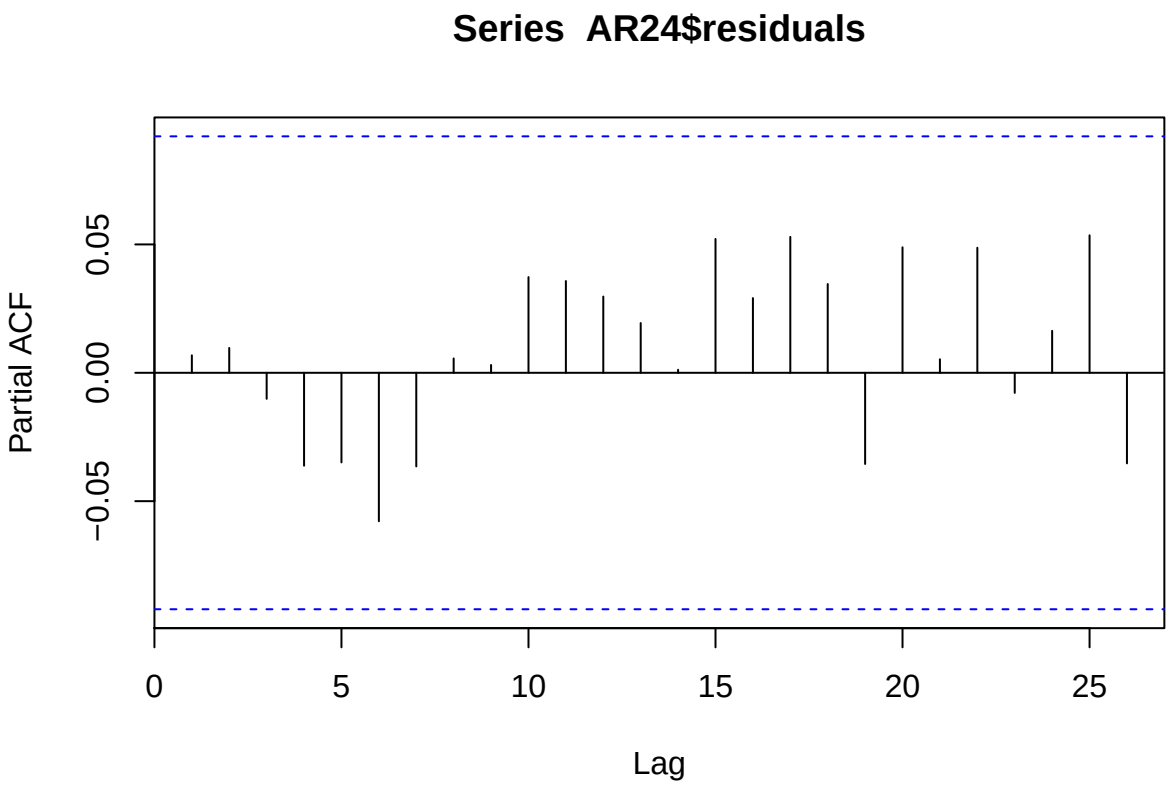
September

```
train9 <- train %>%
  filter(year == '2011' & month == 'September') %>%
  select(datetime, count)

test9 <- test %>%
  filter(year == '2011' & month == 'September') %>%
  mutate(count = NA) %>%
  select(datetime, count)

# head(train9)
# head(test9)

AR24 <- arima(train9$count,order=c(25,0,0))
# tsdisplay(residuals(AR24),lag.max=25,main="AR(24) Resid. Diagnostics")
```

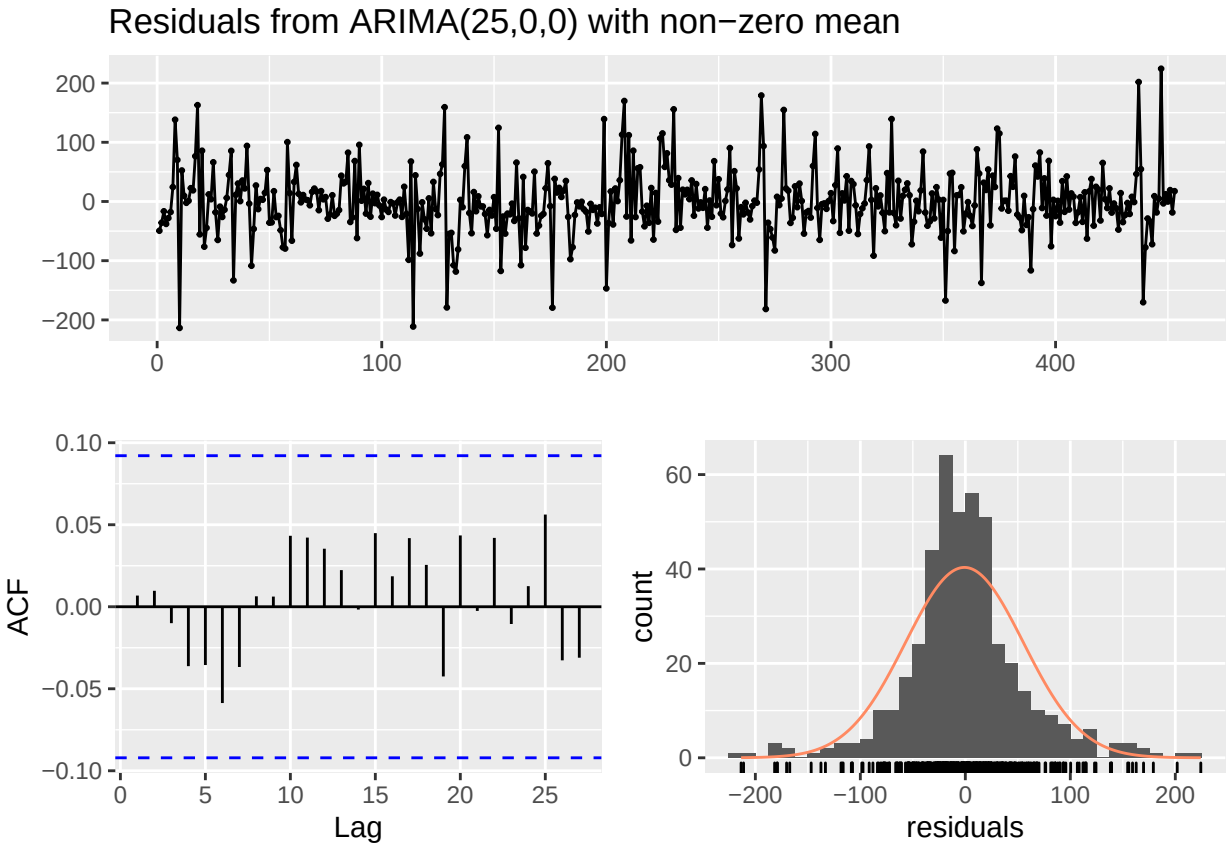
```r
number = nrow(test9)

acf(AR24$residuals)
```




```r
pacf(AR24$residuals)
```

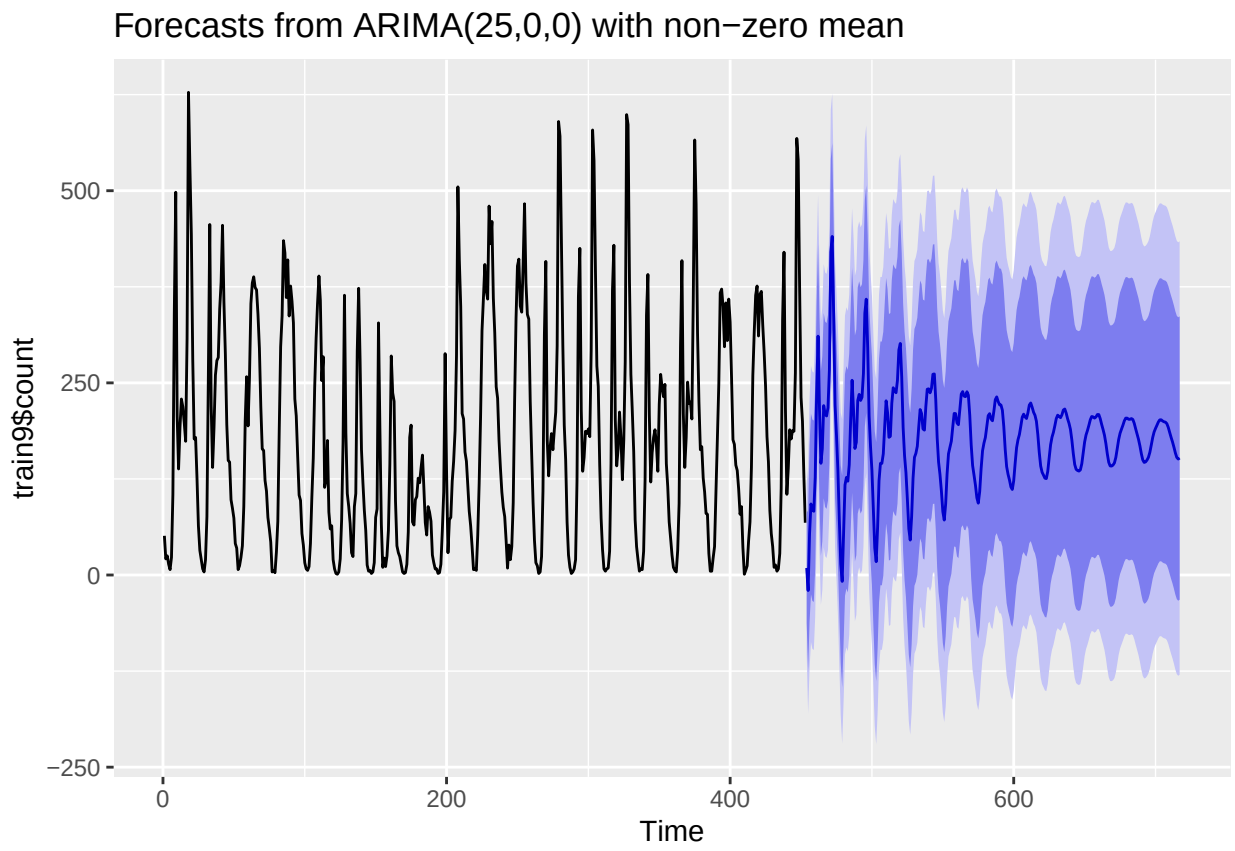



```
checkresiduals(AR24)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0) with non-zero mean
## Q* = 13.656, df = 3, p-value = 0.003413
##
## Model df: 26.   Total lags used: 29
```

```
fcst <- forecast(AR24, h=number)

autoplot(fcst)
```



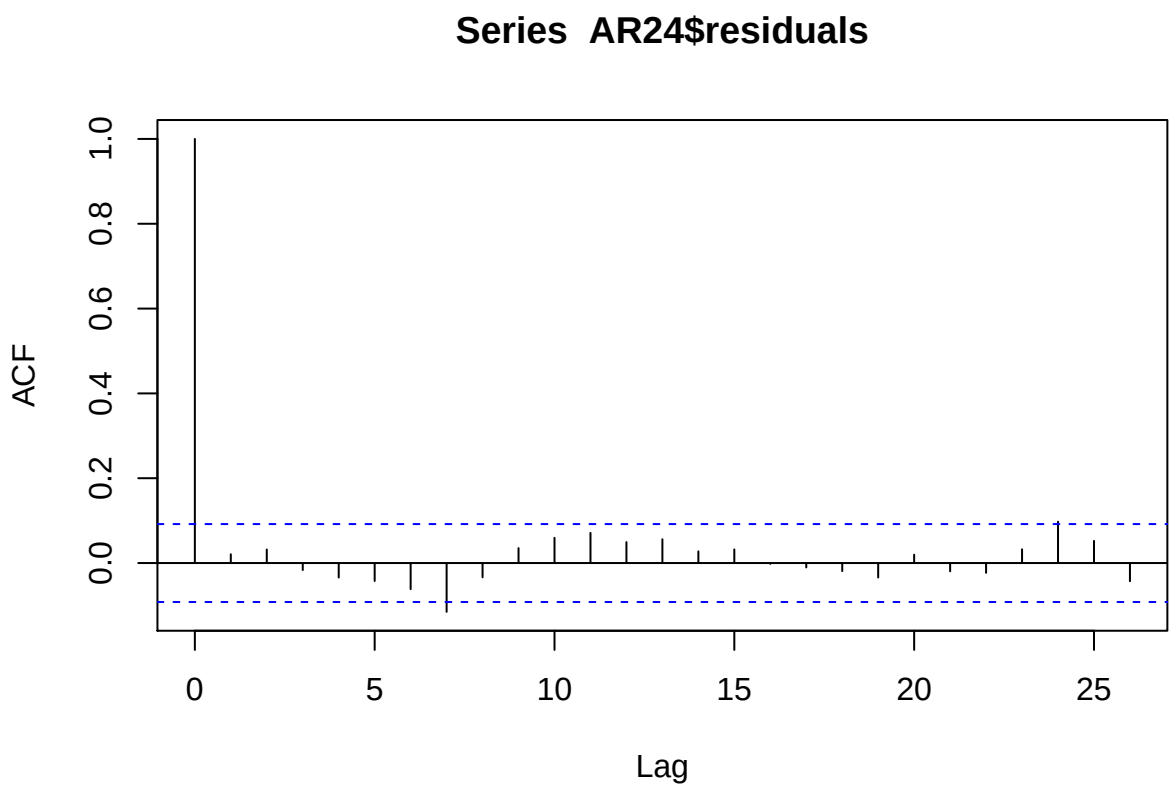

```
# point estimate (mean)
test9$count <- round(fcst$mean)

# test9

RMSLE(y_pred = floor(ifelse(fcst$fitted < 0, 0, round(fcst$fitted))), y_true = train9$count)
```

```
## [1] 0.7600216
```

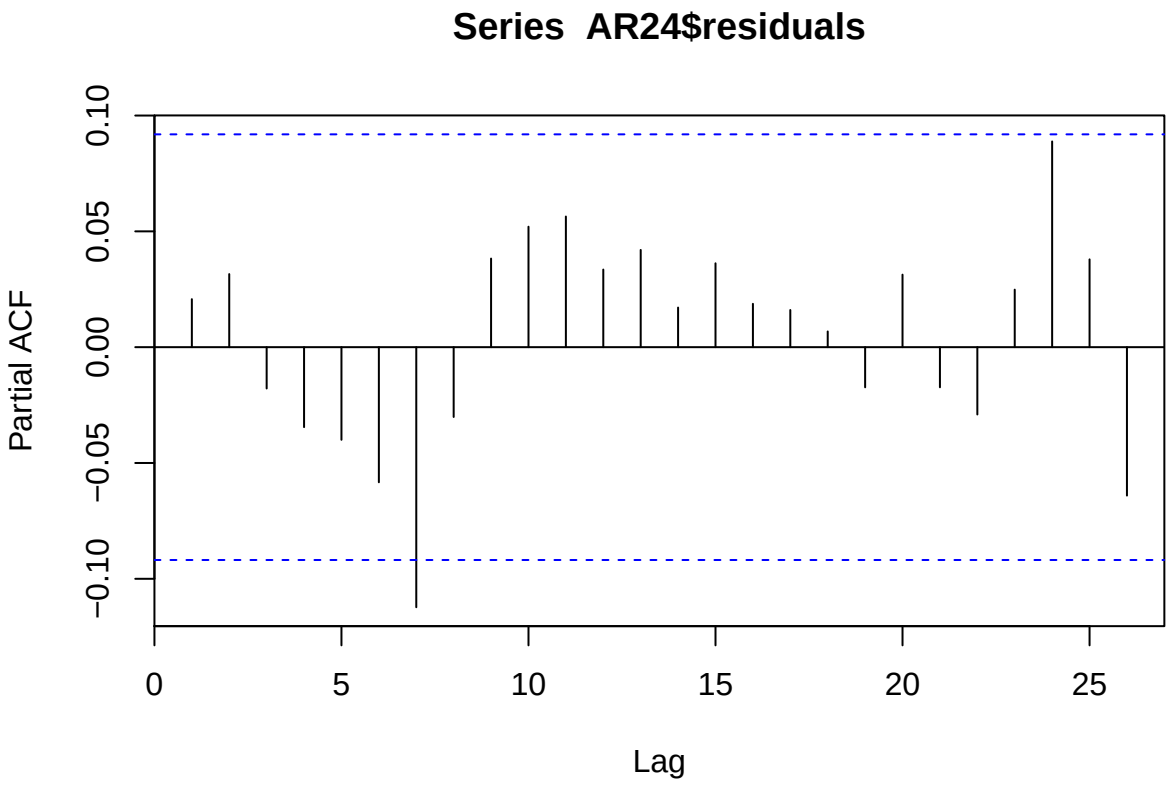
October

```
train10 <- train %>%
  filter(year == '2011' & month == 'October') %>%
  select(datetime, count)

test10 <- test %>%
  filter(year == '2011' & month == 'October') %>%
  mutate(count = NA) %>%
  select(datetime, count)

# head(train10)
# head(test10)

AR24 <- arima(train10$count,order=c(25,0,0))
# tsdisplay(residuals(AR24),lag.max=25,main="AR(24) Resid. Diagnostics")
```

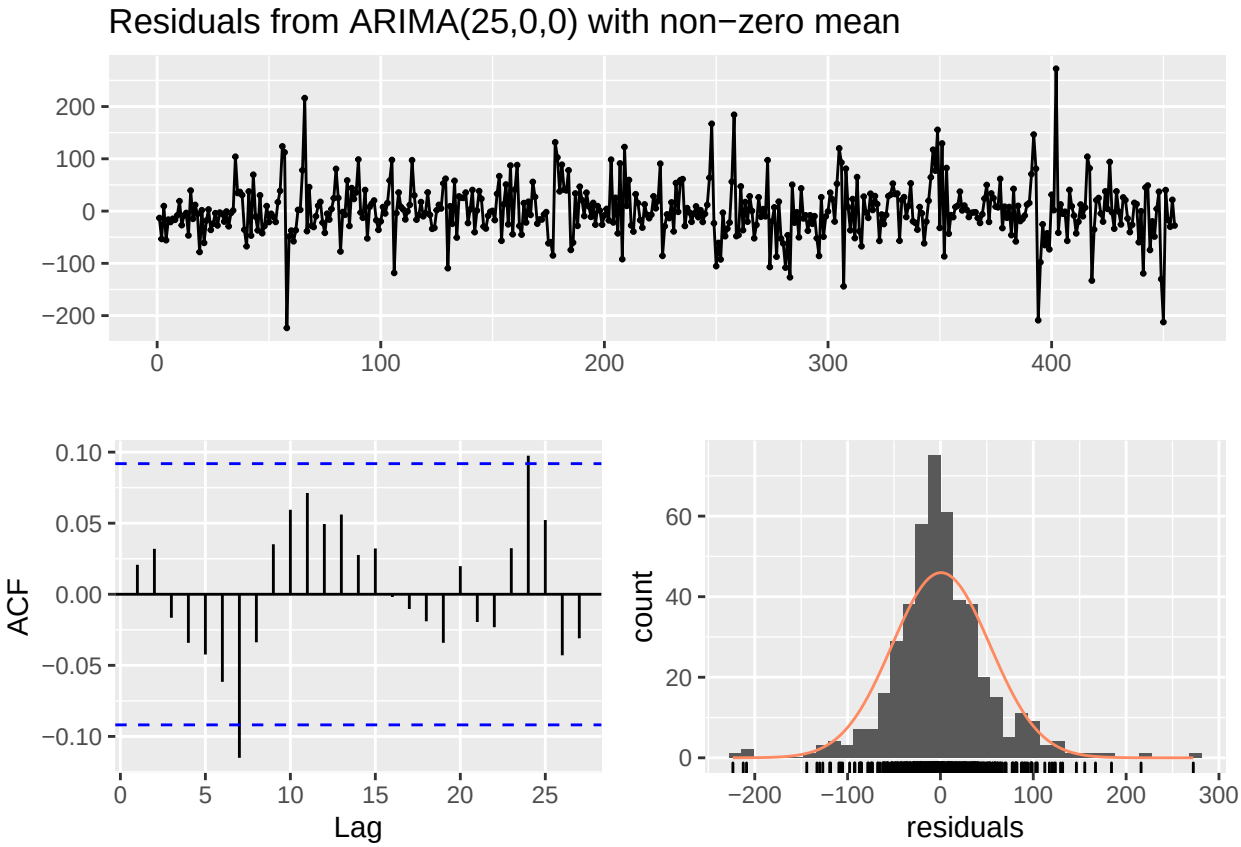
```r
number = nrow(test10)

acf(AR24$residuals)
```




```r
pacf(AR24$residuals)
```

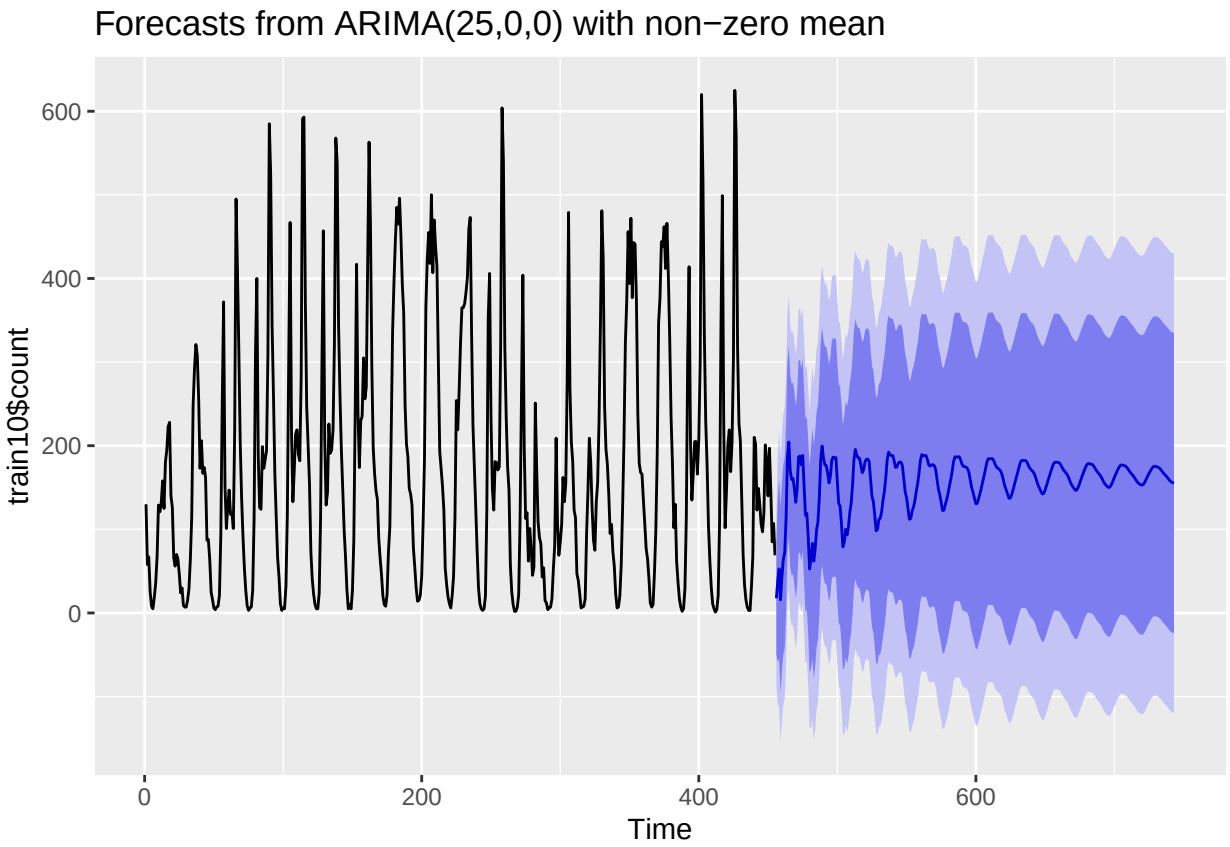



```
checkresiduals(AR24)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0) with non-zero mean
## Q* = 29.574, df = 3, p-value = 1.696e-06
##
## Model df: 26.   Total lags used: 29
```

```
fcst <- forecast(AR24, h=number)

autoplot(fcst)
```



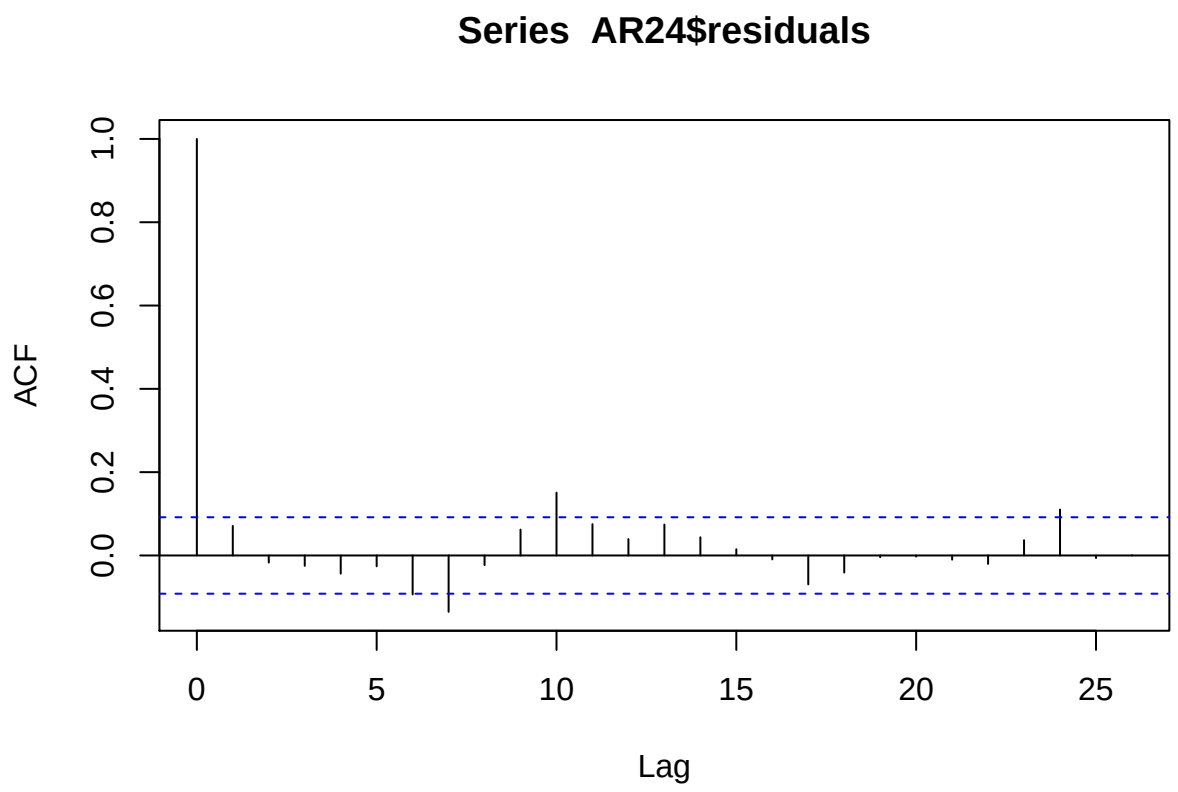

```
# point estimate (mean)
test10$count <- round(fcst$mean)

# test10

RMSLE(y_pred = floor(ifelse(fcst$fitted < 0, 0, round(fcst$fitted))), y_true = train10$count)
```

```
## [1] 0.6762384
```

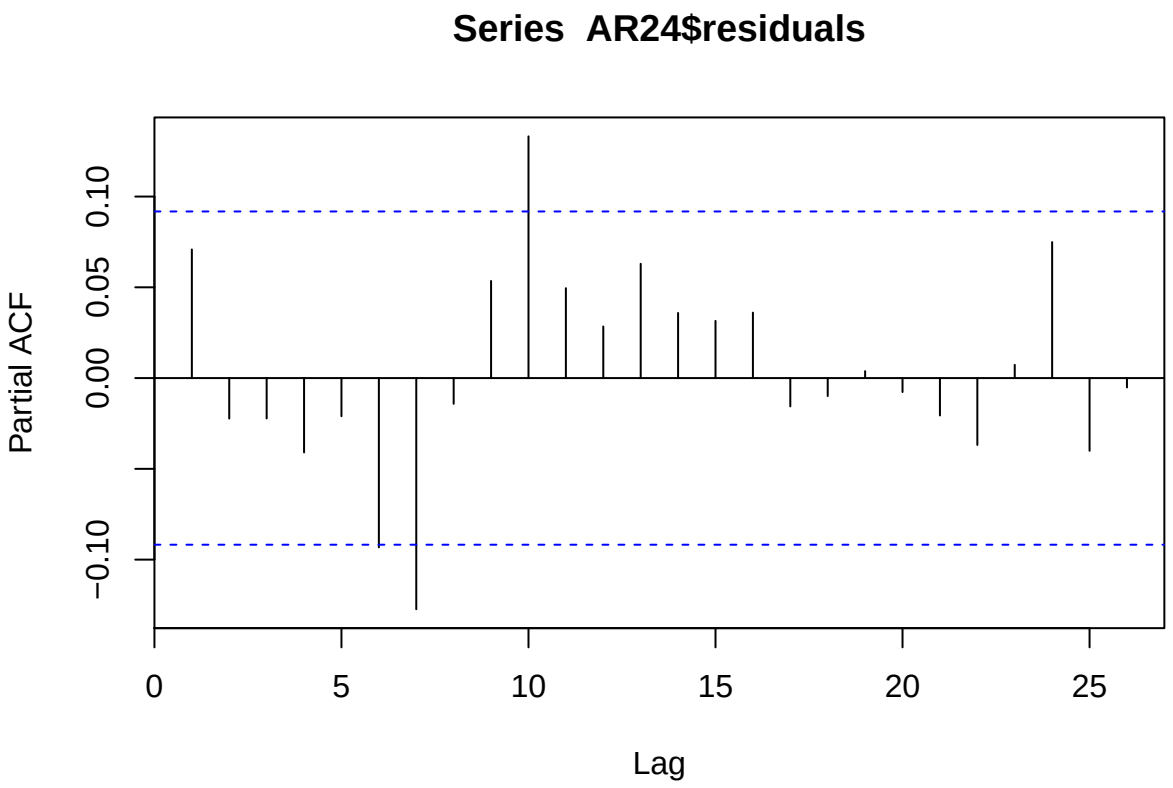
November

```
train11 <- train %>%
  filter(year == '2011' & month == 'November') %>%
  select(datetime, count)

test11 <- test %>%
  filter(year == '2011' & month == 'November') %>%
  mutate(count = NA) %>%
  select(datetime, count)

# head(train11)
# head(test11)

AR24 <- arima(train11$count,order=c(25,0,0))
# tsdisplay(residuals(AR24),lag.max=25,main="AR(24) Resid. Diagnostics")
```

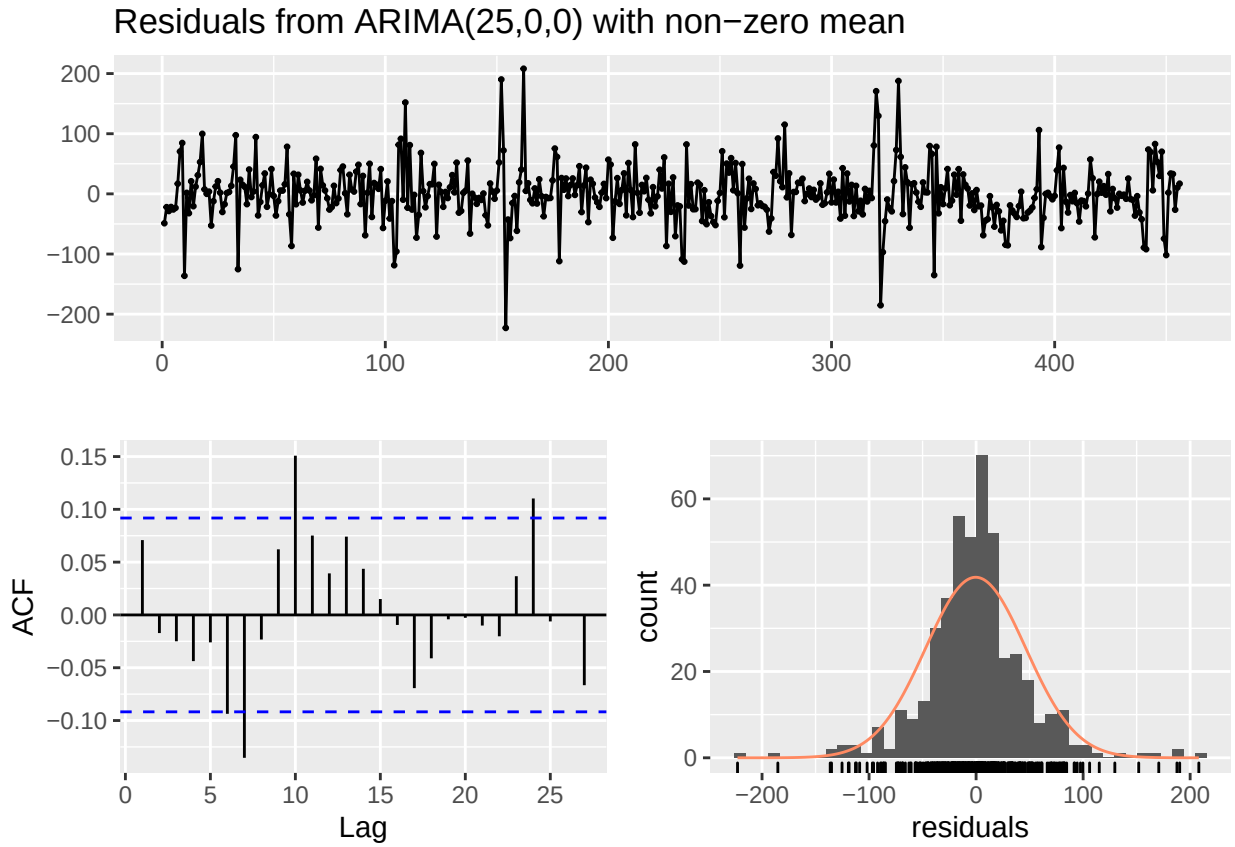
```
number = nrow(test11)

acf(AR24$residuals)
```




```
pacf(AR24$residuals)
```

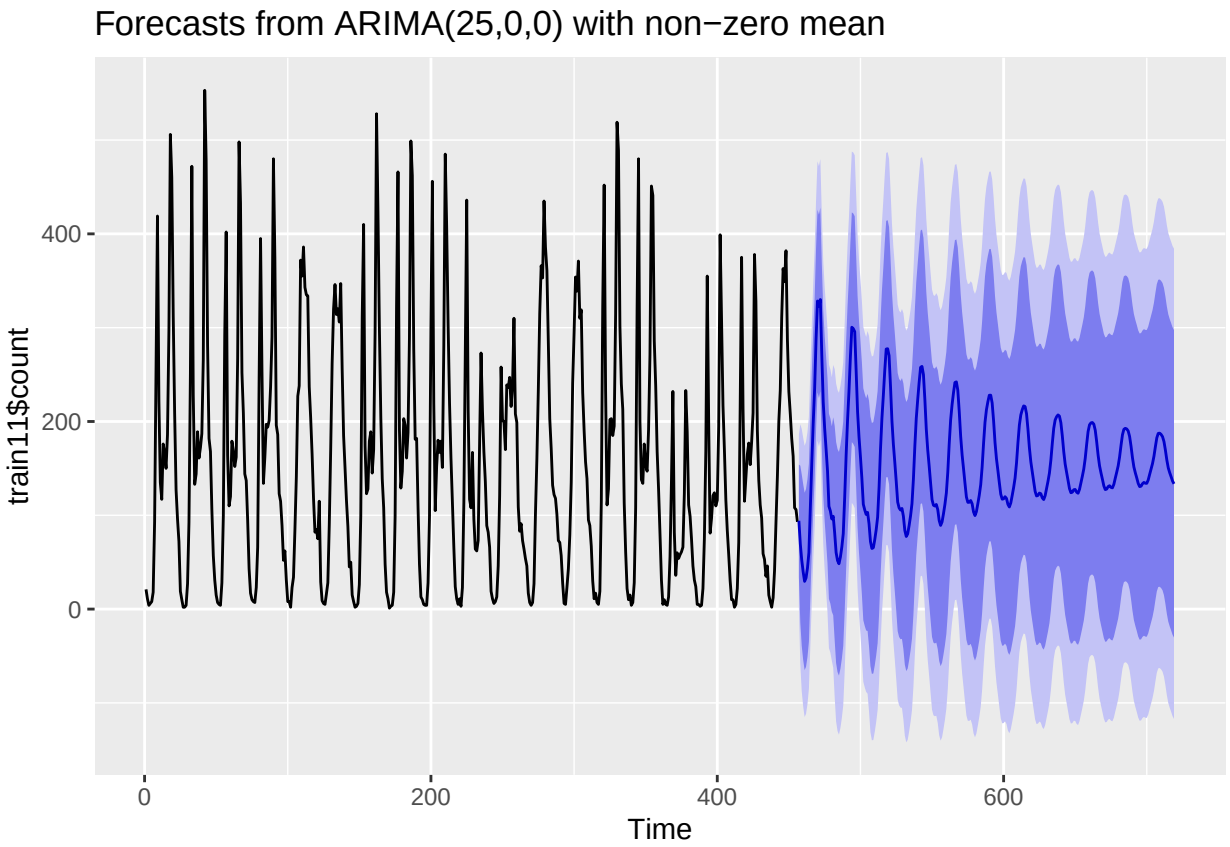



```
checkresiduals(AR24)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0) with non-zero mean
## Q* = 49.722, df = 3, p-value = 9.156e-11
##
## Model df: 26.   Total lags used: 29
```

```
fcst <- forecast(AR24, h=number)

autoplot(fcst)
```



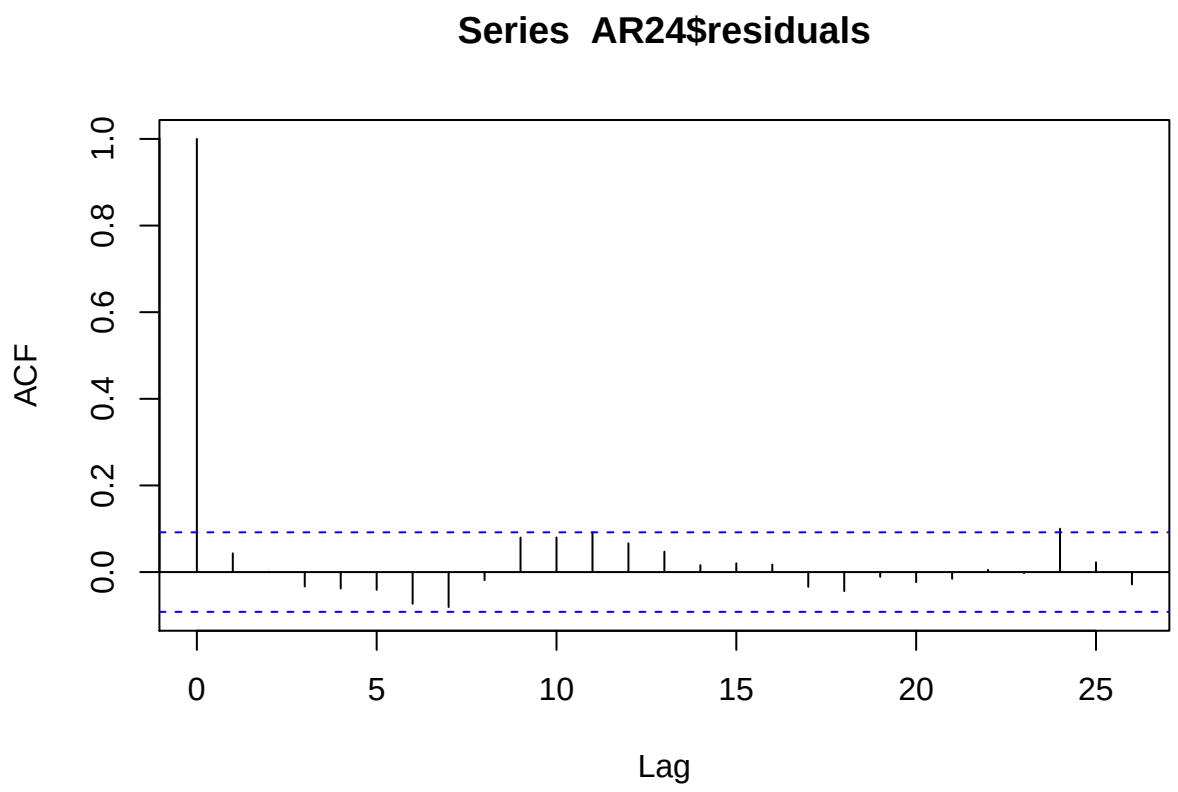

```
# point estimate (mean)
test11$count <- round(fcst$mean)

# test11

RMSLE(y_pred = floor(ifelse(fcst$fitted < 0, 0, round(fcst$fitted))), y_true = train11$count)
```

```
## [1] 0.7449703
```

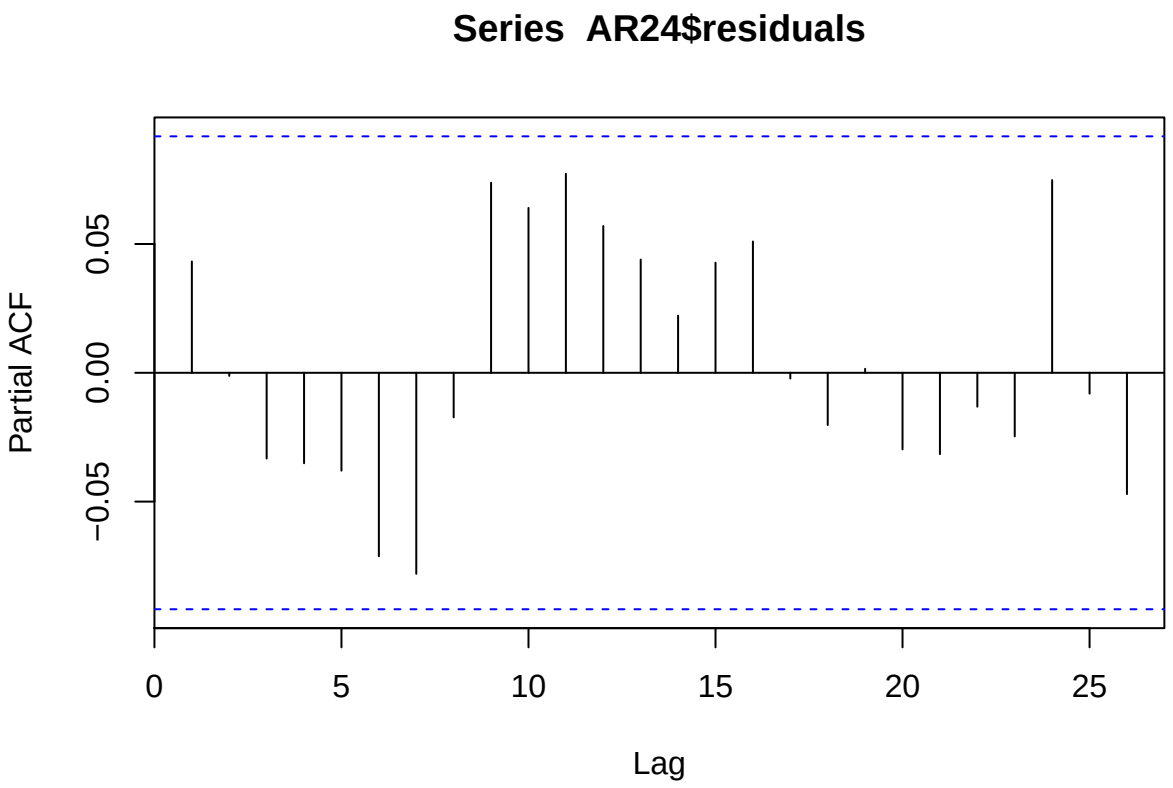
December

```
train12 <- train %>%
  filter(year == '2011' & month == 'December') %>%
  select(datetime, count)

test12 <- test %>%
  filter(year == '2011' & month == 'December') %>%
  mutate(count = NA) %>%
  select(datetime, count)

# head(train12)
# head(test12)

AR24 <- arima(train12$count,order=c(25,0,0))
# tsdisplay(residuals(AR24),lag.max=25,main="AR(24) Resid. Diagnostics")
```

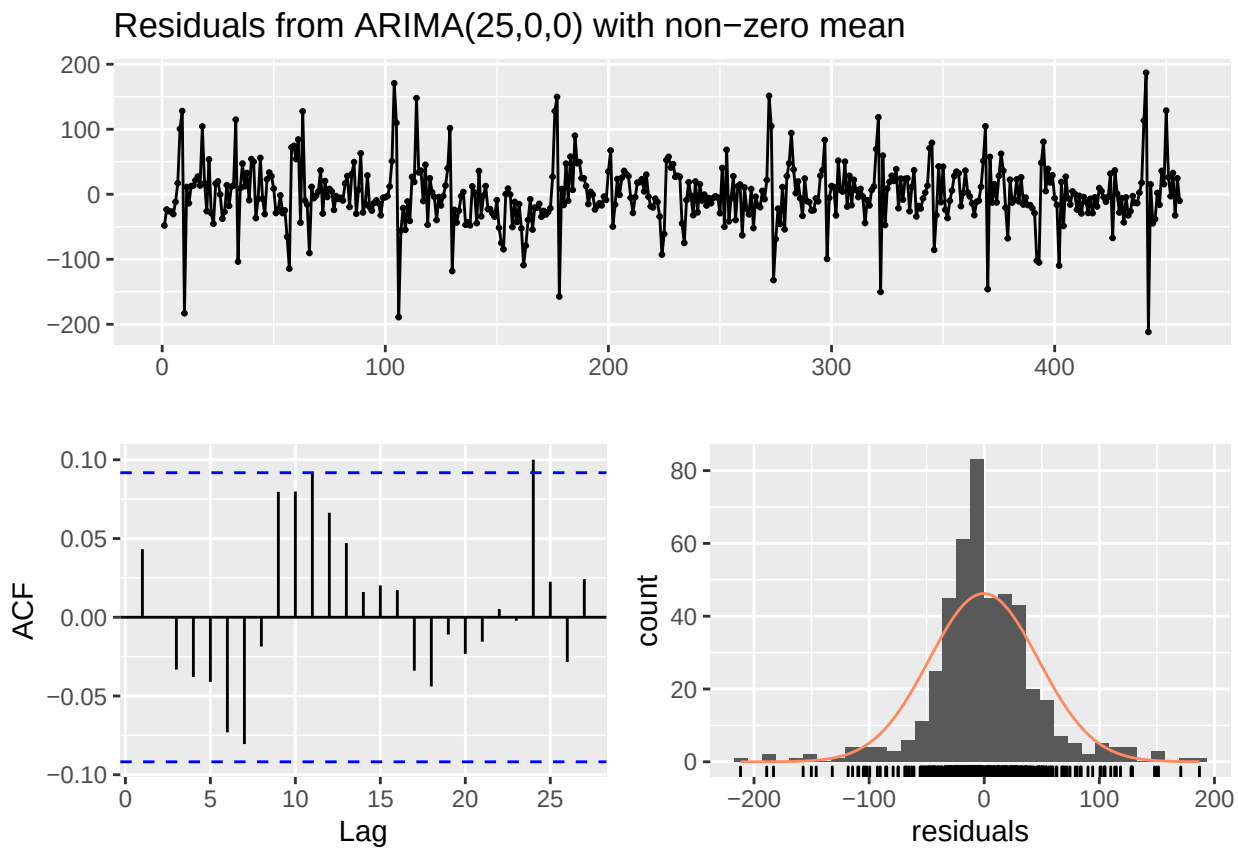
```r
number = nrow(test12)

acf(AR24$residuals)
```




```r
pacf(AR24$residuals)
```

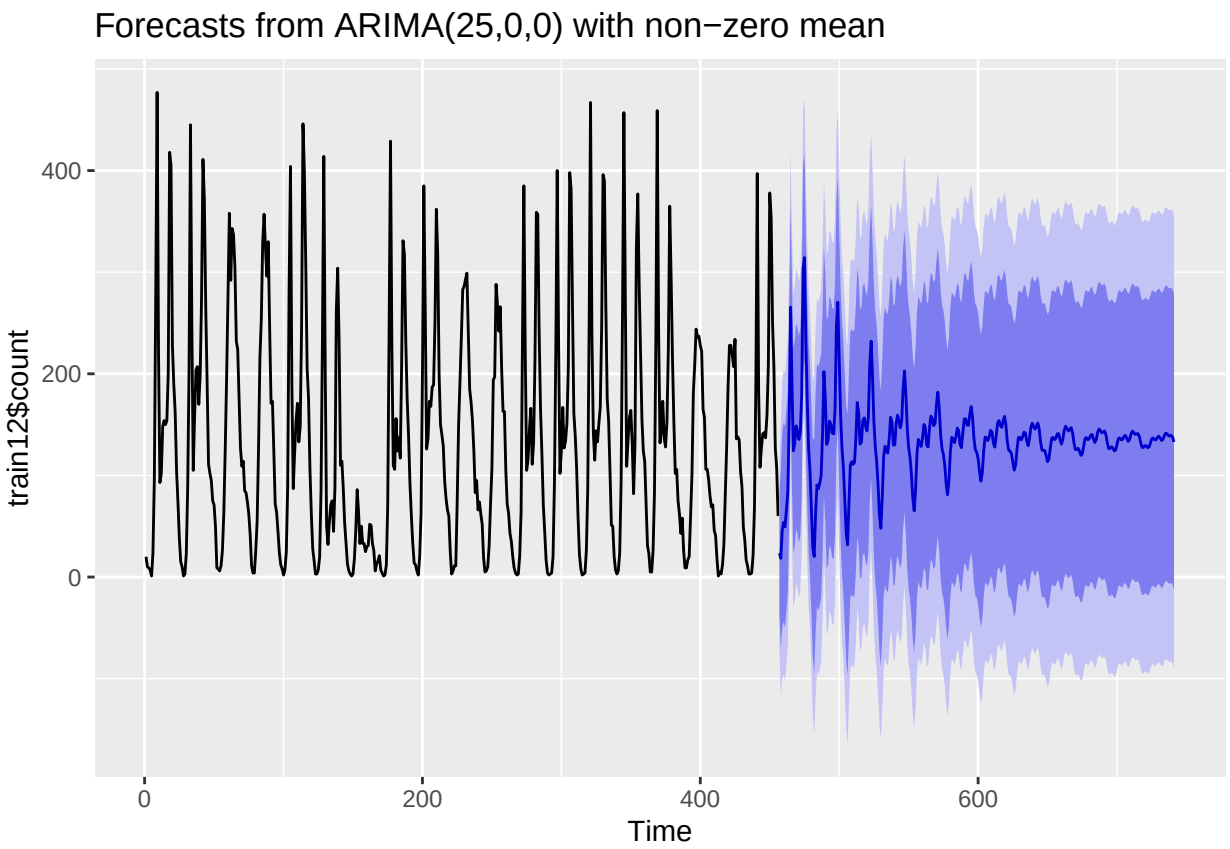



```
checkresiduals(AR24)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0) with non-zero mean
## Q* = 29.7, df = 3, p-value = 1.596e-06
##
## Model df: 26.   Total lags used: 29
```

```
fcst <- forecast(AR24, h=number)

autoplot(fcst)
```




```
# point estimate (mean)
test12$count <- round(fcst$mean)

# test12

RMSLE(y_pred = floor(ifelse(fcst$fitted < 0, 0, round(fcst$fitted))), y_true = train12$count)
```

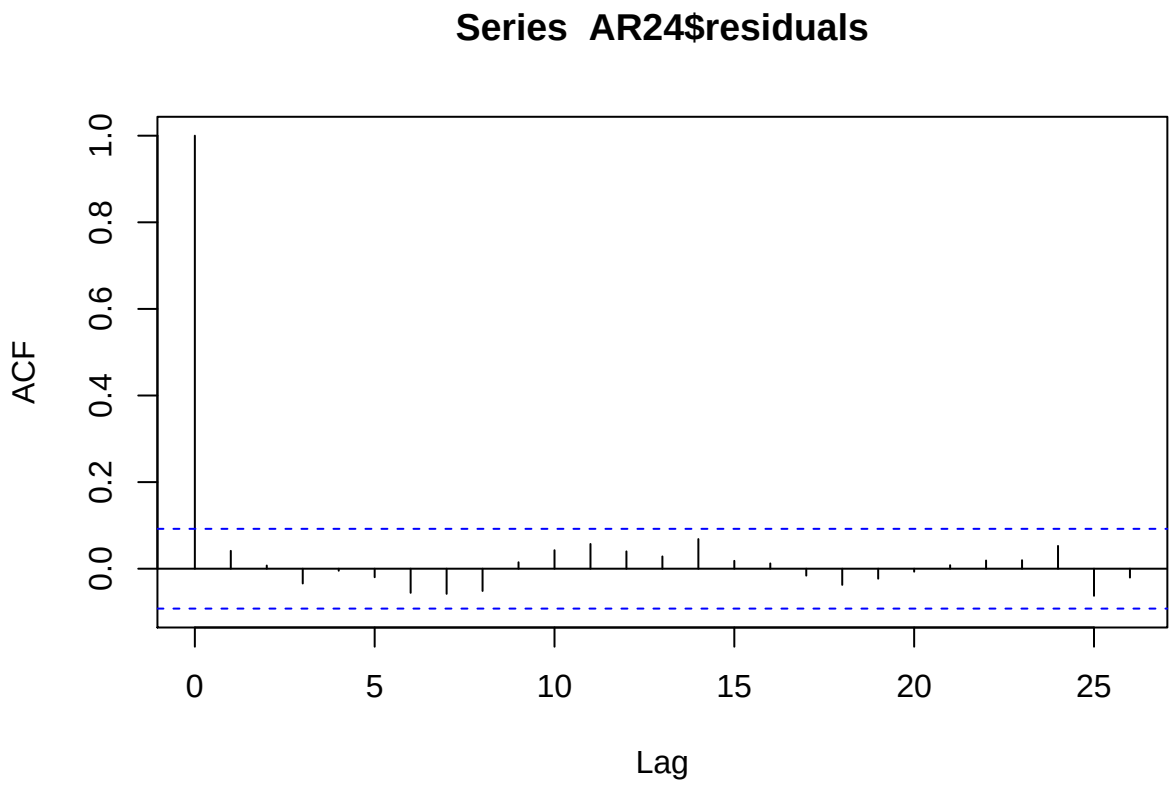
```
## [1] 0.6857358
```

2012

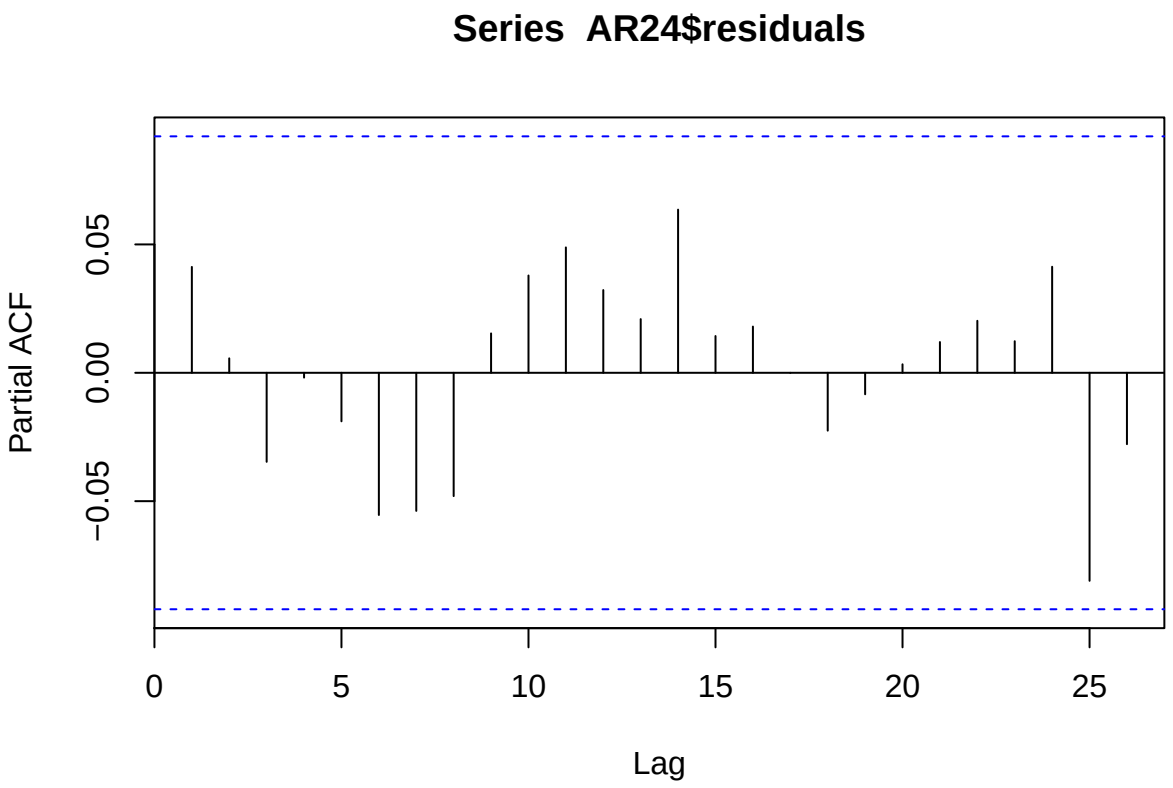
January

```
train13 <- train %>%
  filter(year == '2012' & month == 'January') %>%
  select(datetime, count)

test13 <- test %>%
  filter(year == '2012' & month == 'January') %>%
  mutate(count = NA) %>%
  select(datetime, count)

# head(train13)
# head(test13)

AR24 <- arima(train13$count,order=c(25,0,0))
```

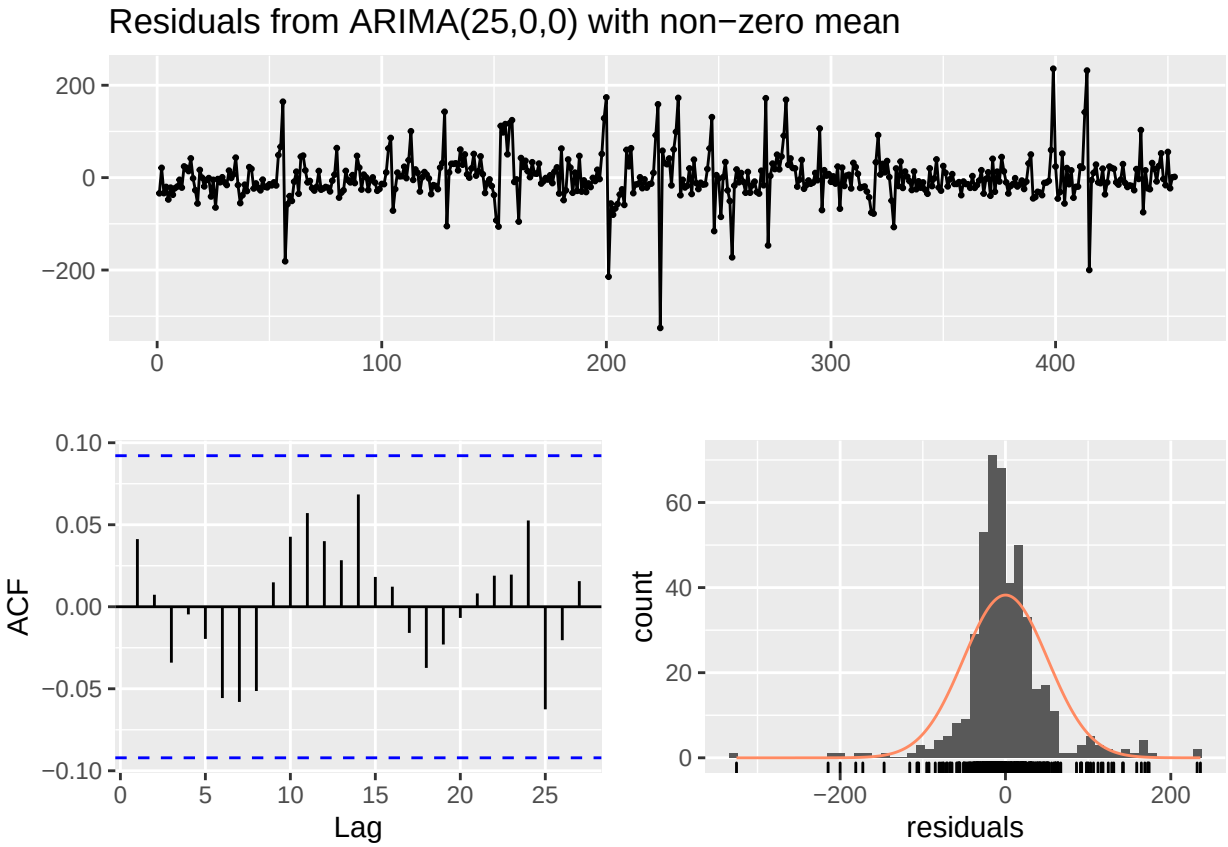
```r
# tsdisplay(residuals(AR24),lag.max=25,main="AR(24) Resid. Diagnostics")

number = nrow(test13)

acf(AR24$residuals)
```




```r
pacf(AR24$residuals)
```

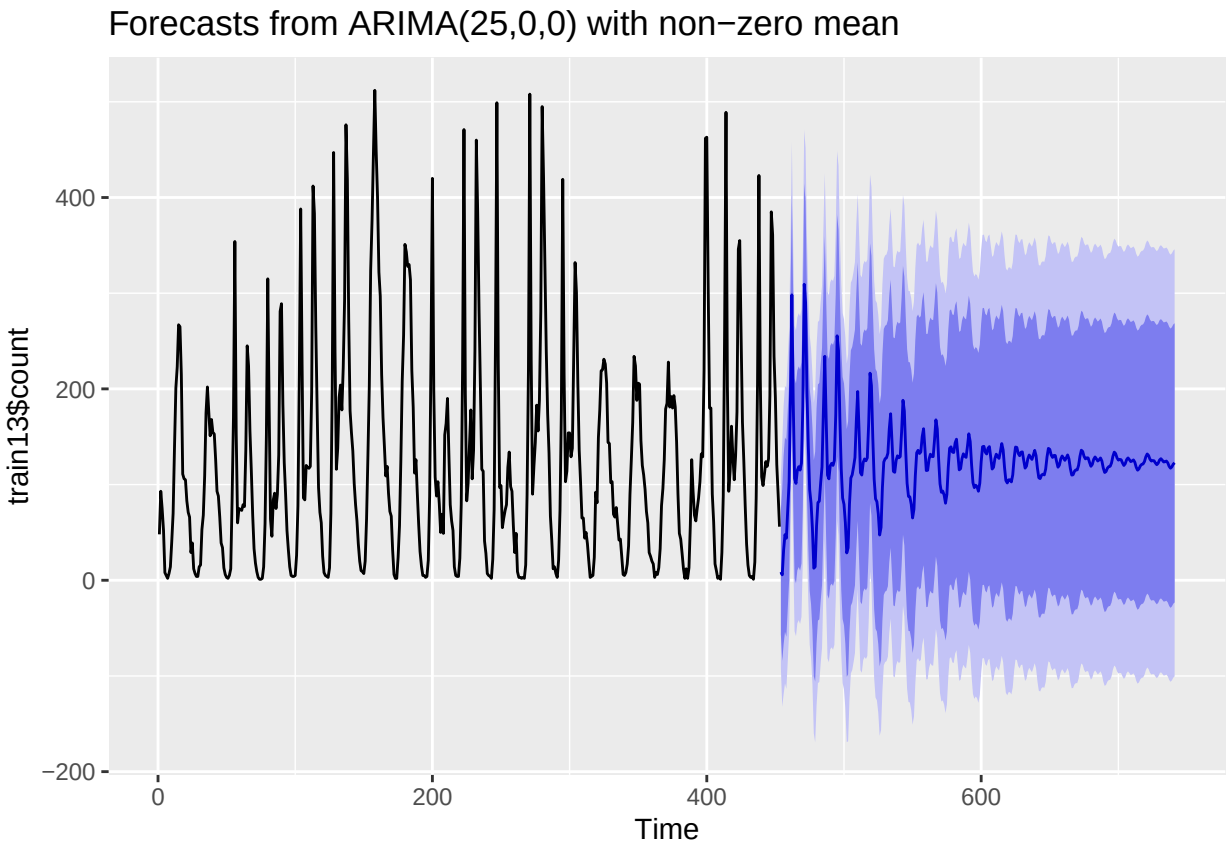



```
checkresiduals(AR24)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0) with non-zero mean
## Q* = 19.095, df = 3, p-value = 0.0002613
##
## Model df: 26.   Total lags used: 29
```

```
fcst <- forecast(AR24, h=number)

autoplot(fcst)
```



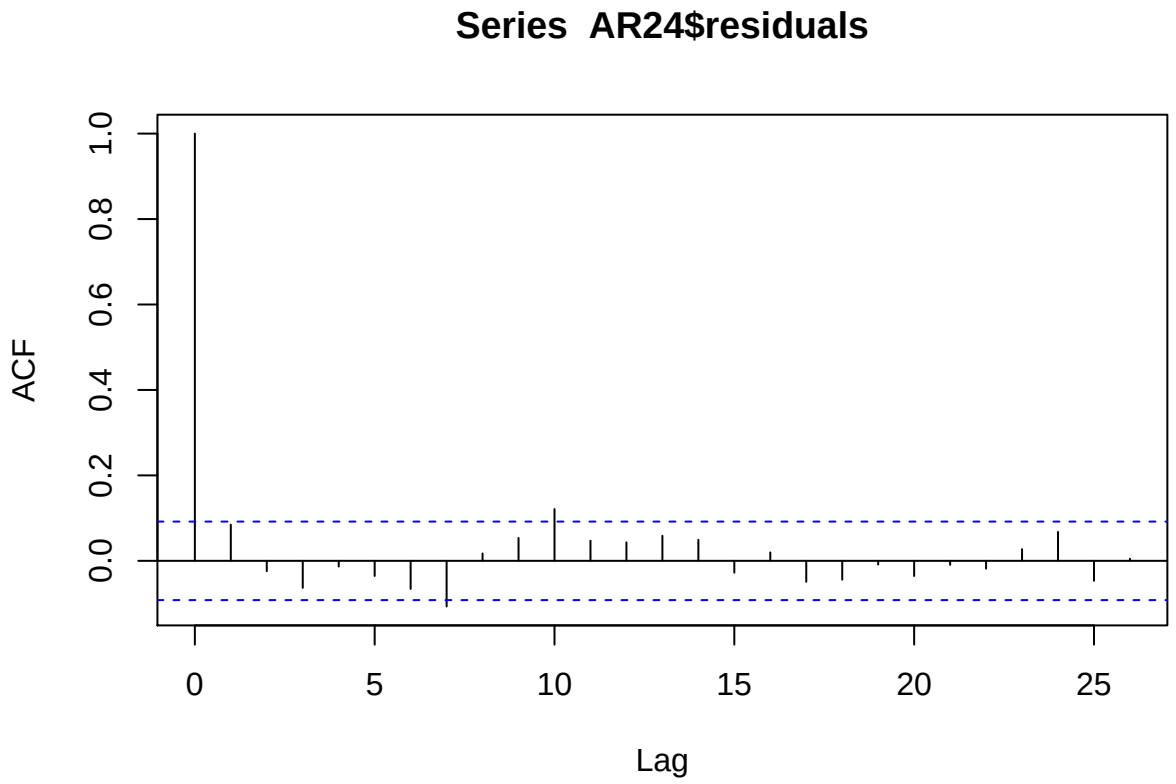

```
# point estimate (mean)
test13$count <- round(fcst$mean)
# test13

RMSLE(y_pred = floor(ifelse(fcst$fitted < 0, 0, round(fcst$fitted))), y_true = train13$count)
```

```
## [1] 0.7707937
```

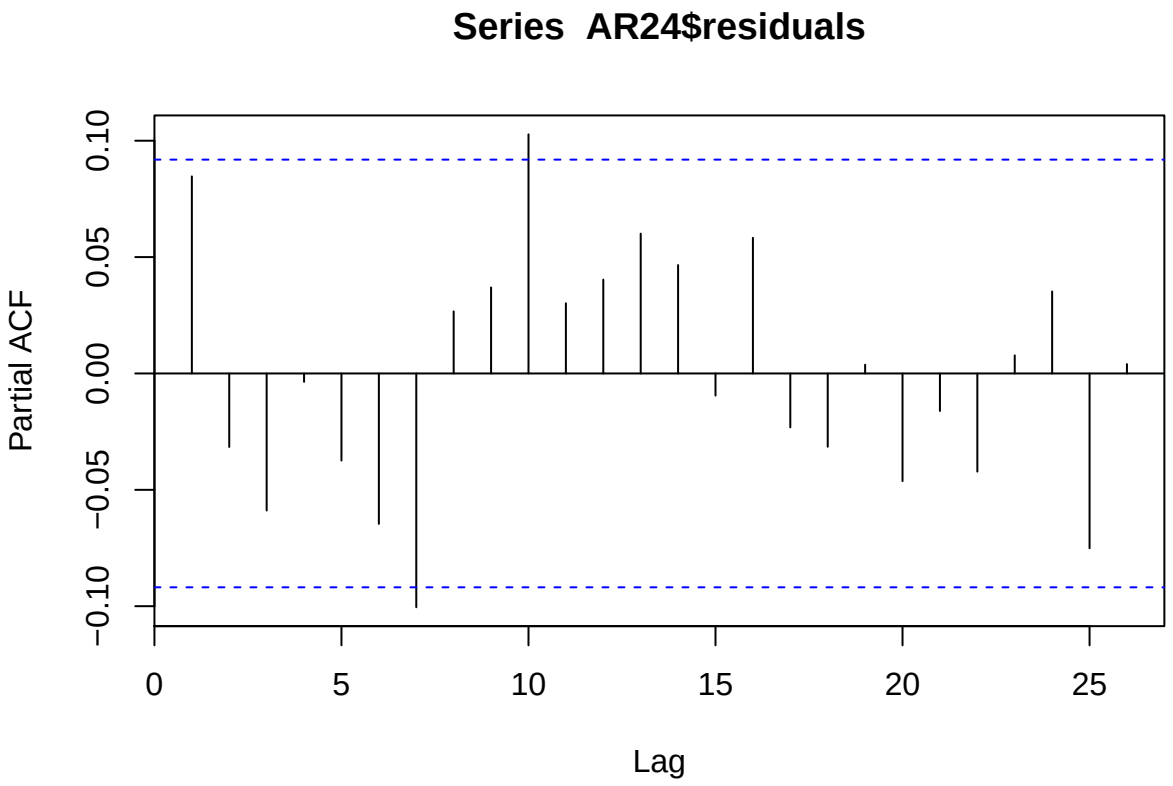
February

```
train14 <- train %>%
  filter(year == '2012' & month == 'February') %>%
  select(datetime, count)

test14 <- test %>%
  filter(year == '2012' & month == 'February') %>%
  mutate(count = NA) %>%
  select(datetime, count)

# head(train14)
# head(test14)

AR24 <- arima(train14$count,order=c(25,0,0))
# tsdisplay(residuals(AR24),lag.max=25,main="AR(24) Resid. Diagnostics")

number = nrow(test14)
```

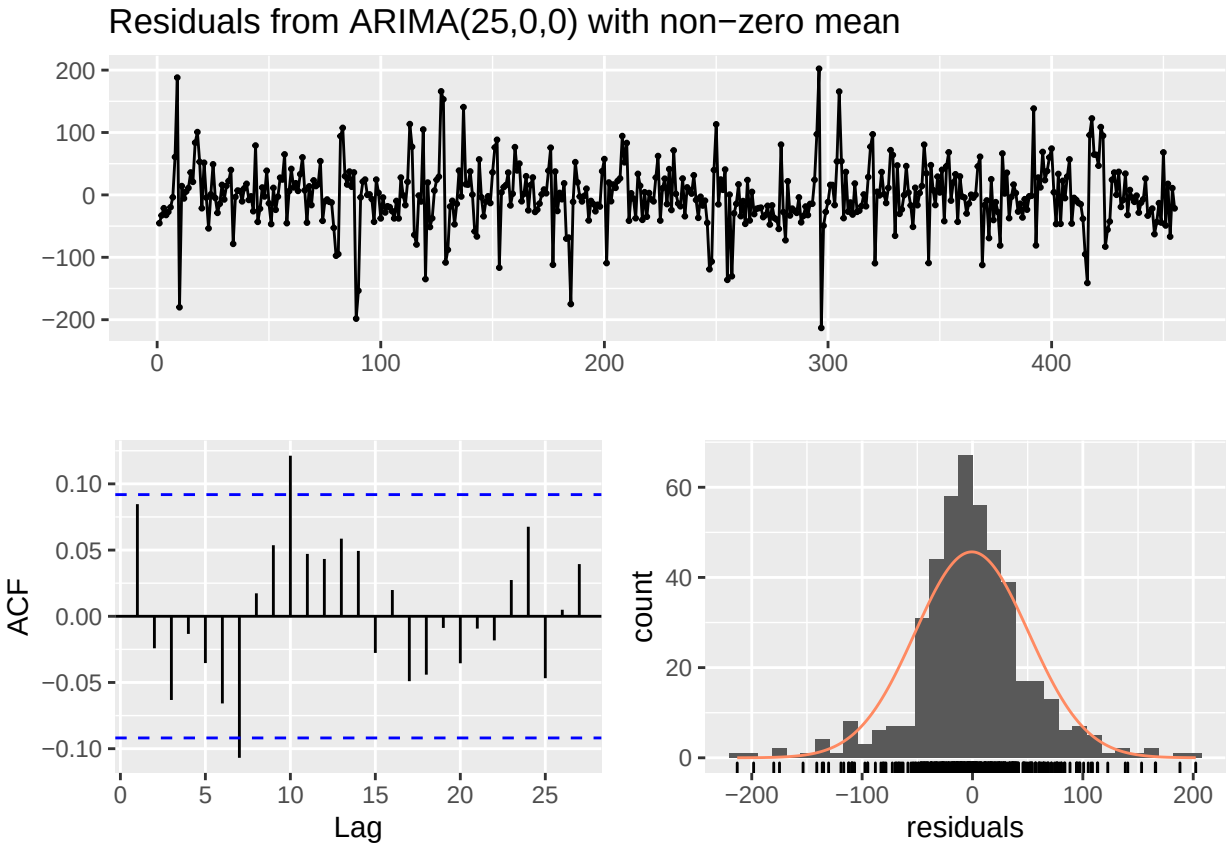
```r
acf(AR24$residuals)
```




```r
pacf(AR24$residuals)
```

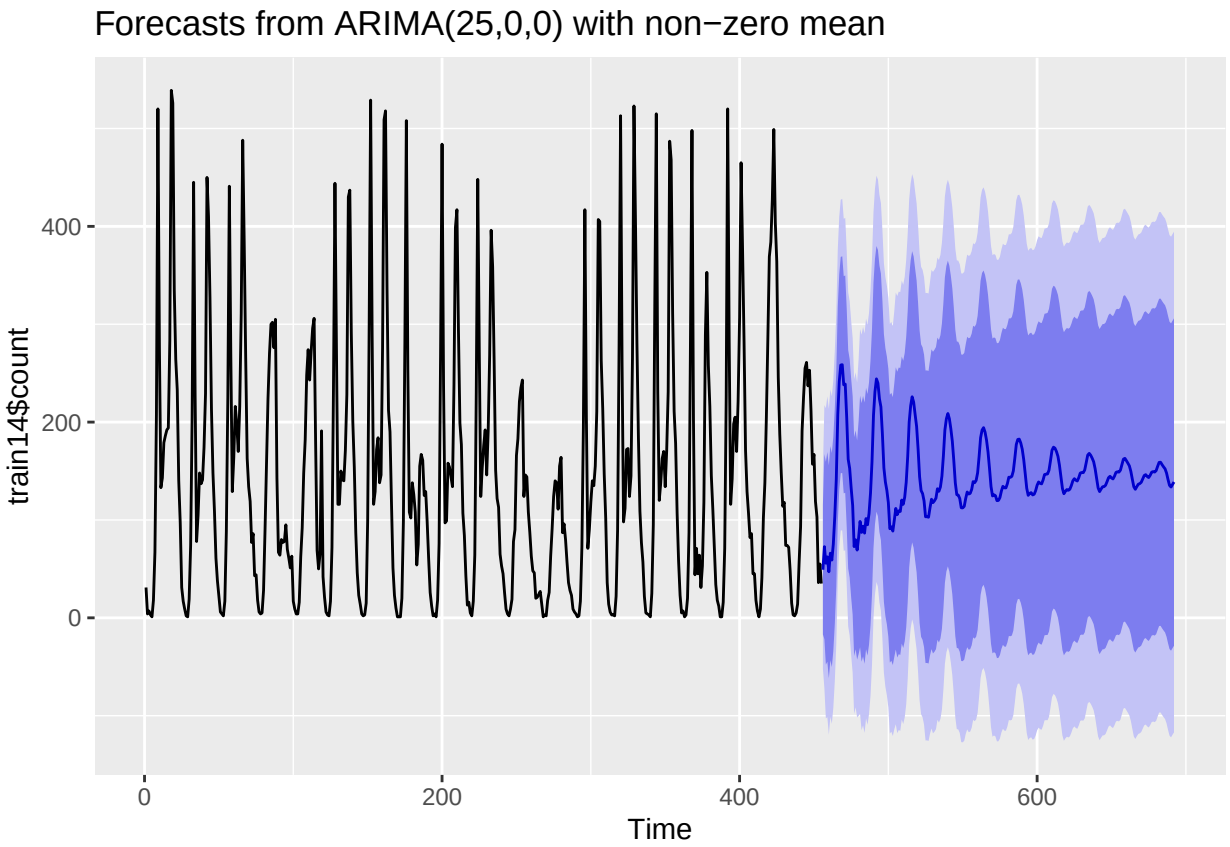



```
checkresiduals(AR24)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0) with non-zero mean
## Q* = 34.488, df = 3, p-value = 1.563e-07
##
## Model df: 26.   Total lags used: 29
```

```
fcst <- forecast(AR24, h=number)

autoplot(fcst)
```



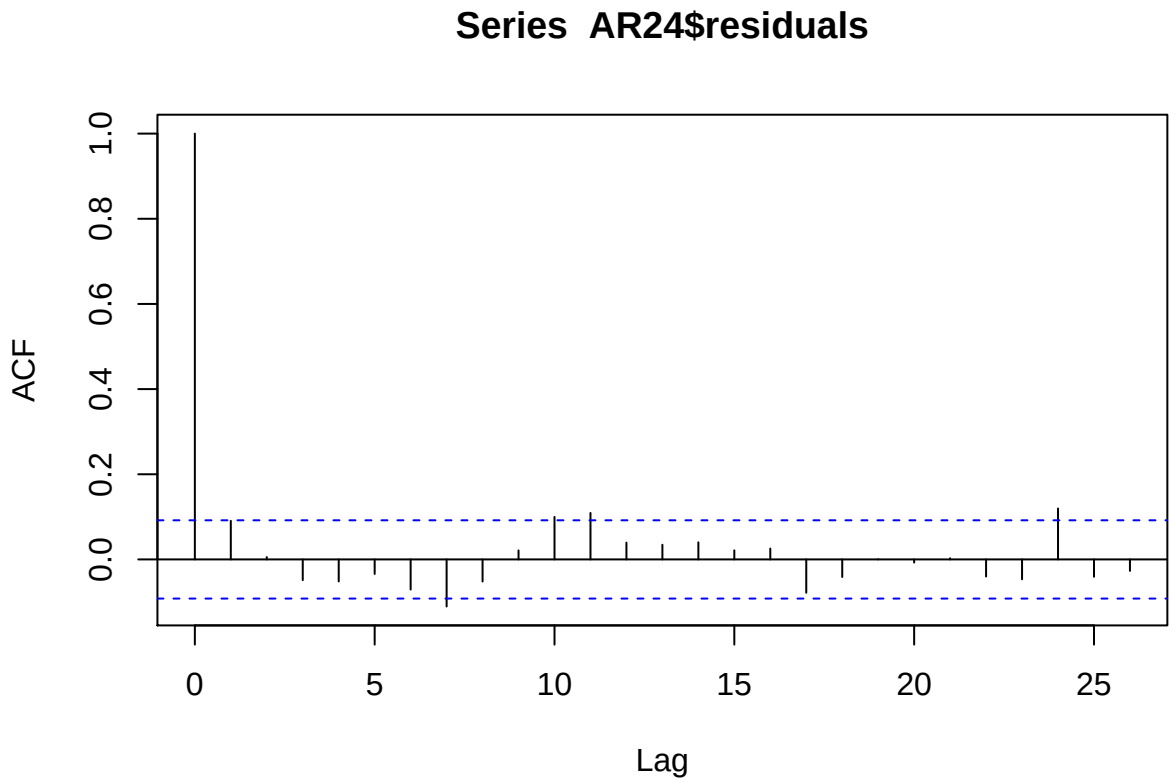

```
# point estimate (mean)
test14$count <- round(fcst$mean)


RMSLE(y_pred = floor(ifelse(fcst$fitted < 0, 0, round(fcst$fitted))), y_true = train14$count)
```

```
## [1] 0.8620118
```

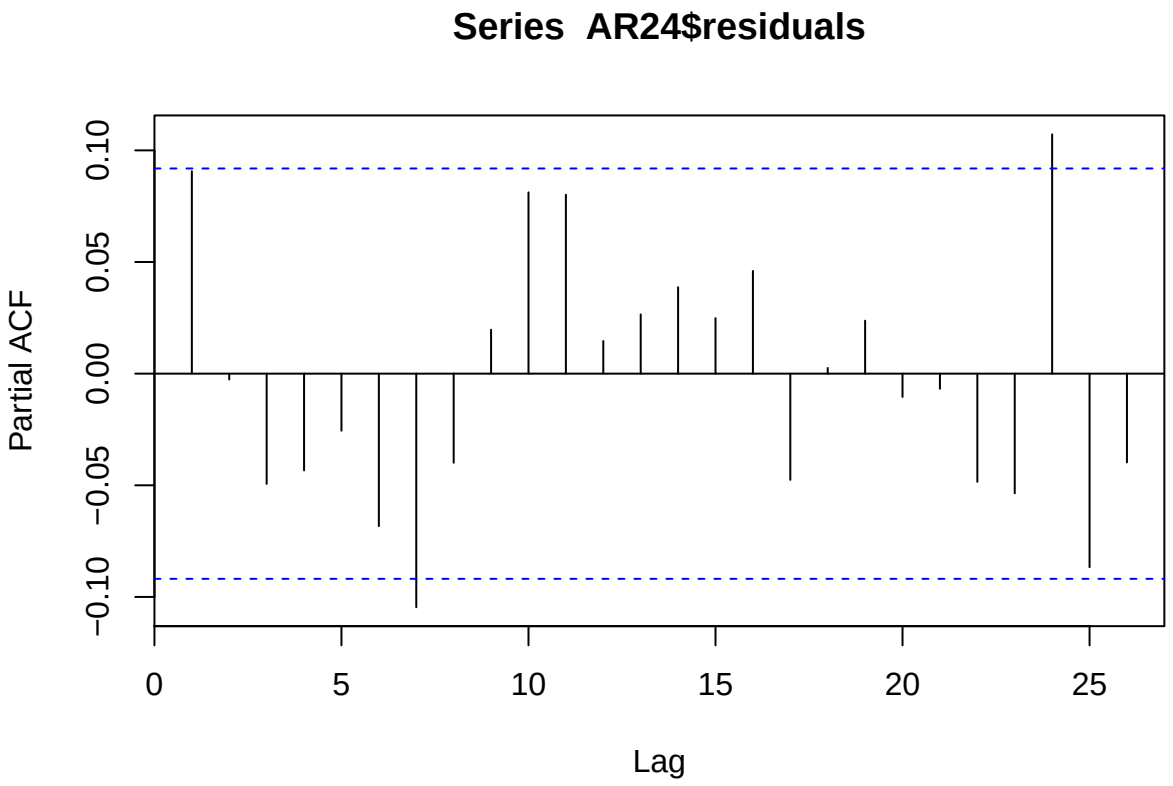
March

```
train15 <- train %>%
  filter(year == '2012' & month == 'March') %>%
  select(datetime, count)

test15 <- test %>%
  filter(year == '2012' & month == 'March') %>%
  mutate(count = NA) %>%
  select(datetime, count)

# head(train15)
# head(test15)

AR24 <- arima(train15$count,order=c(25,0,0))
# tsdisplay(residuals(AR24),lag.max=25,main="AR(24) Resid. Diagnostics")

number = nrow(test15)
```

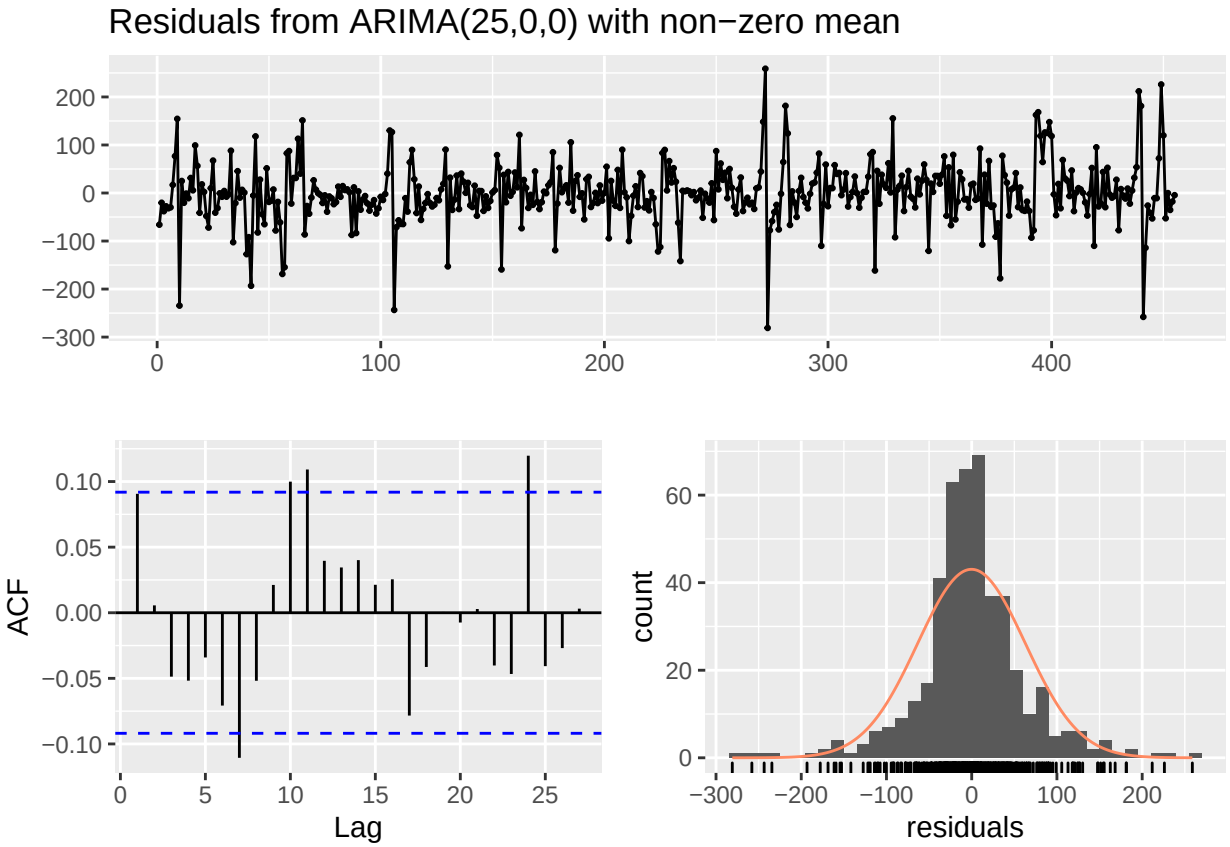
```r
acf(AR24$residuals)
```




```r
pacf(AR24$residuals)
```

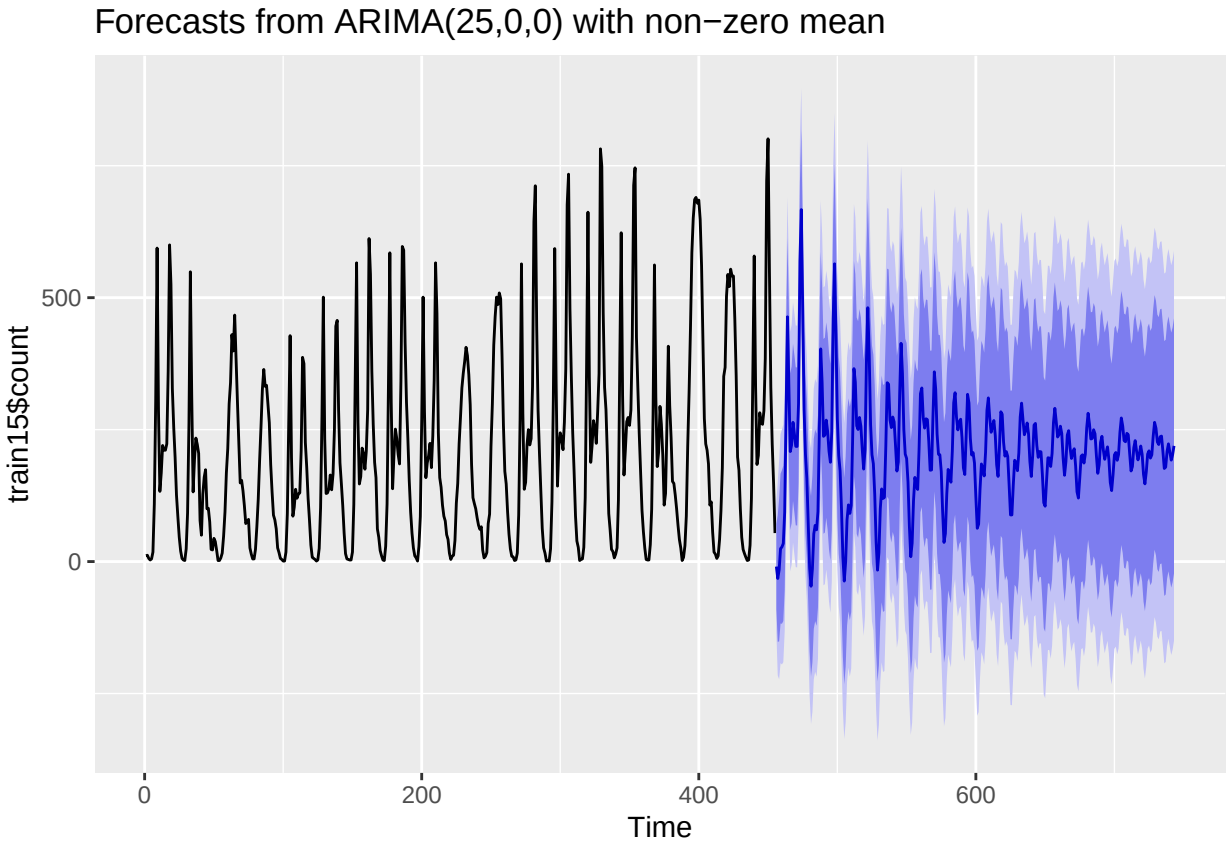



```
checkresiduals(AR24)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0) with non-zero mean
## Q* = 42.759, df = 3, p-value = 2.769e-09
##
## Model df: 26.   Total lags used: 29
```

```
fcst <- forecast(AR24, h=number)

autoplot(fcst)
```



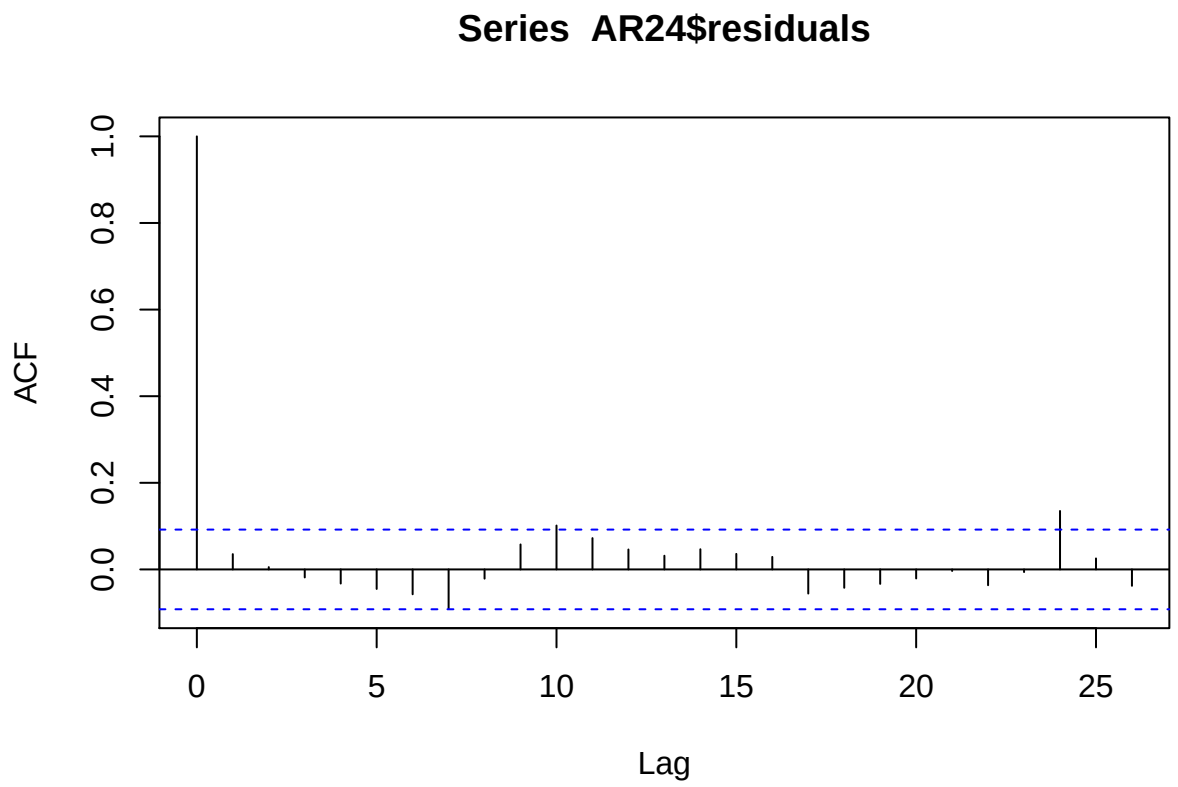

```
# point estimate (mean)
test15$count <- round(fcst$mean)

# test15


RMSLE(y_pred = floor(ifelse(fcst$fitted < 0, 0, round(fcst$fitted))), y_true = train15$count)
```

```
## [1] 0.8112976
```

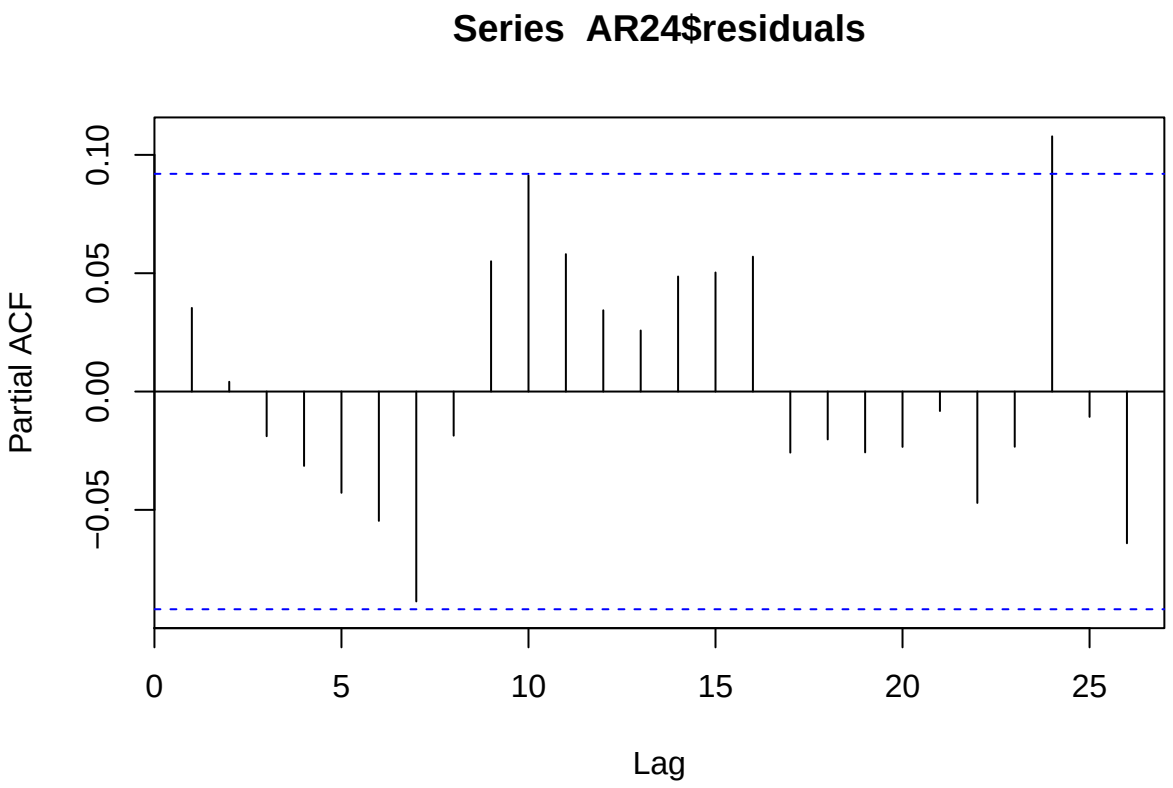
April

```
train16 <- train %>%
  filter(year == '2012' & month == 'April') %>%
  select(datetime, count)

test16 <- test %>%
  filter(year == '2012' & month == 'April') %>%
  mutate(count = NA) %>%
  select(datetime, count)

# head(train16)
# head(test16)

AR24 <- arima(train16$count,order=c(25,0,0))
# tsdisplay(residuals(AR24),lag.max=25,main="AR(24) Resid. Diagnostics")
```

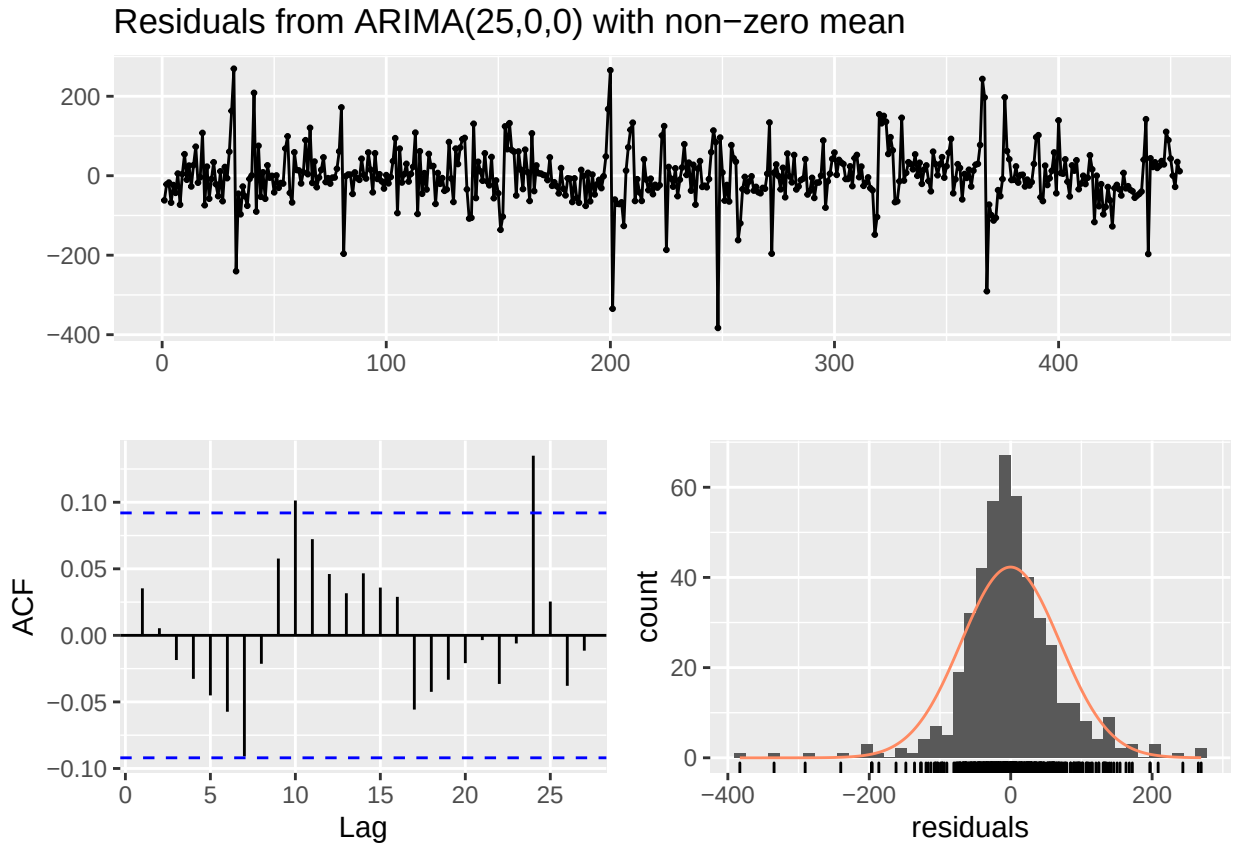
```r
number = nrow(test16)

acf(AR24$residuals)
```




```r
pacf(AR24$residuals)
```

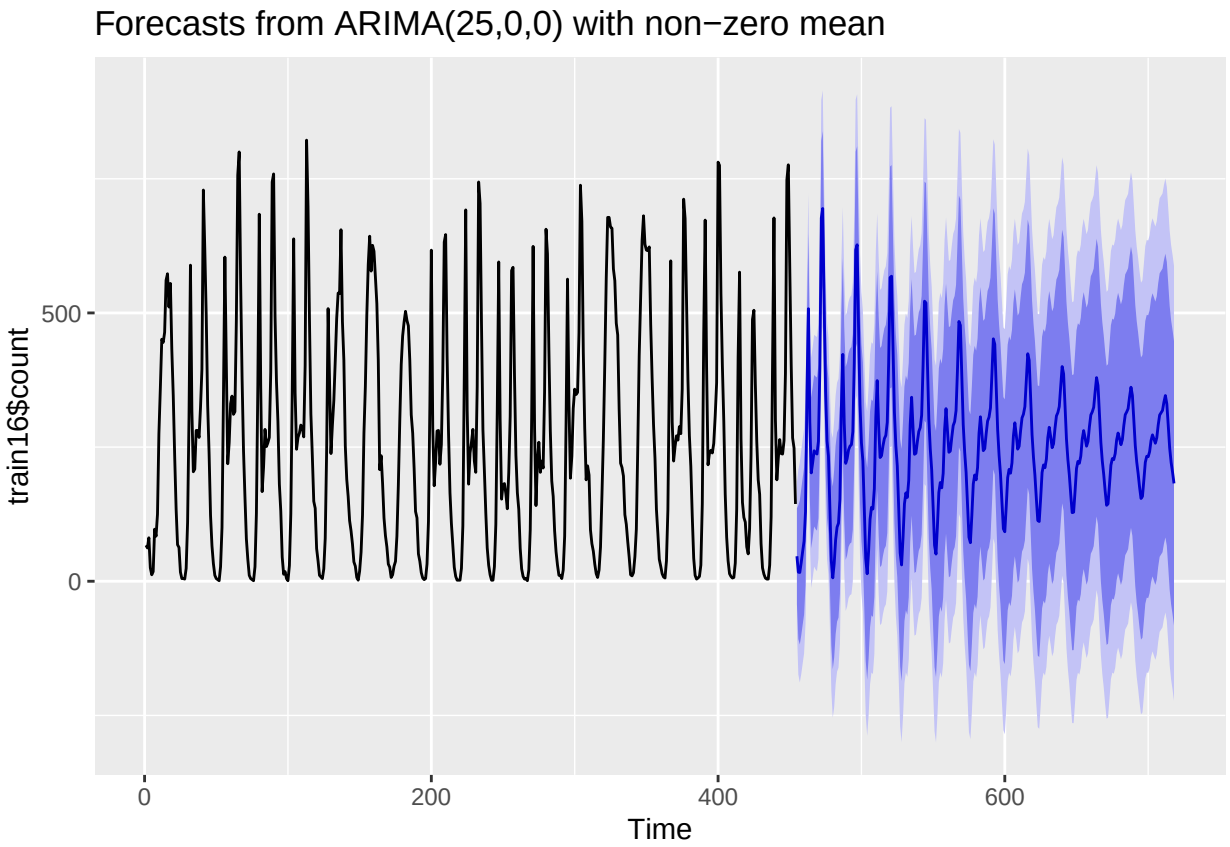



```
checkresiduals(AR24)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0) with non-zero mean
## Q* = 34.022, df = 3, p-value = 1.96e-07
##
## Model df: 26.   Total lags used: 29
```

```
fcst <- forecast(AR24, h=number)

autoplot(fcst)
```



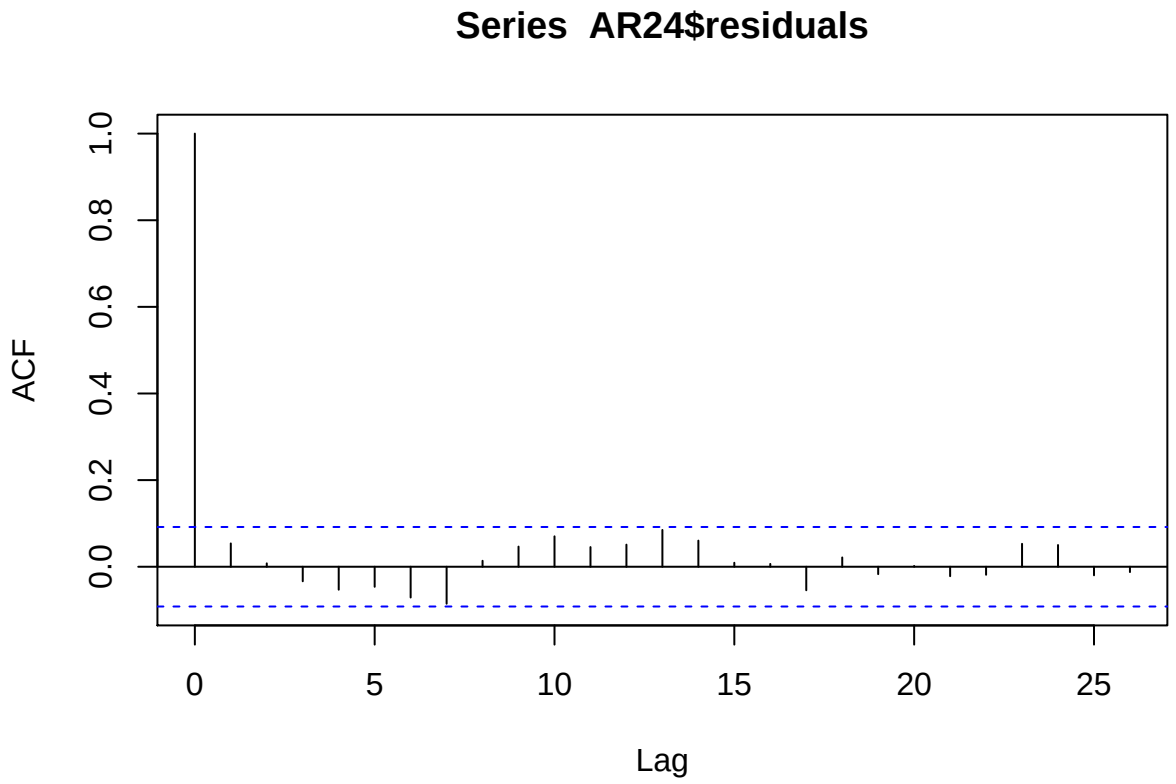

```
# point estimate (mean)
test16$count <- round(fcst$mean)


RMSLE(y_pred = floor(ifelse(fcst$fitted < 0, 0, round(fcst$fitted))), y_true = train16$count)
```

```
## [1] 0.775059
```

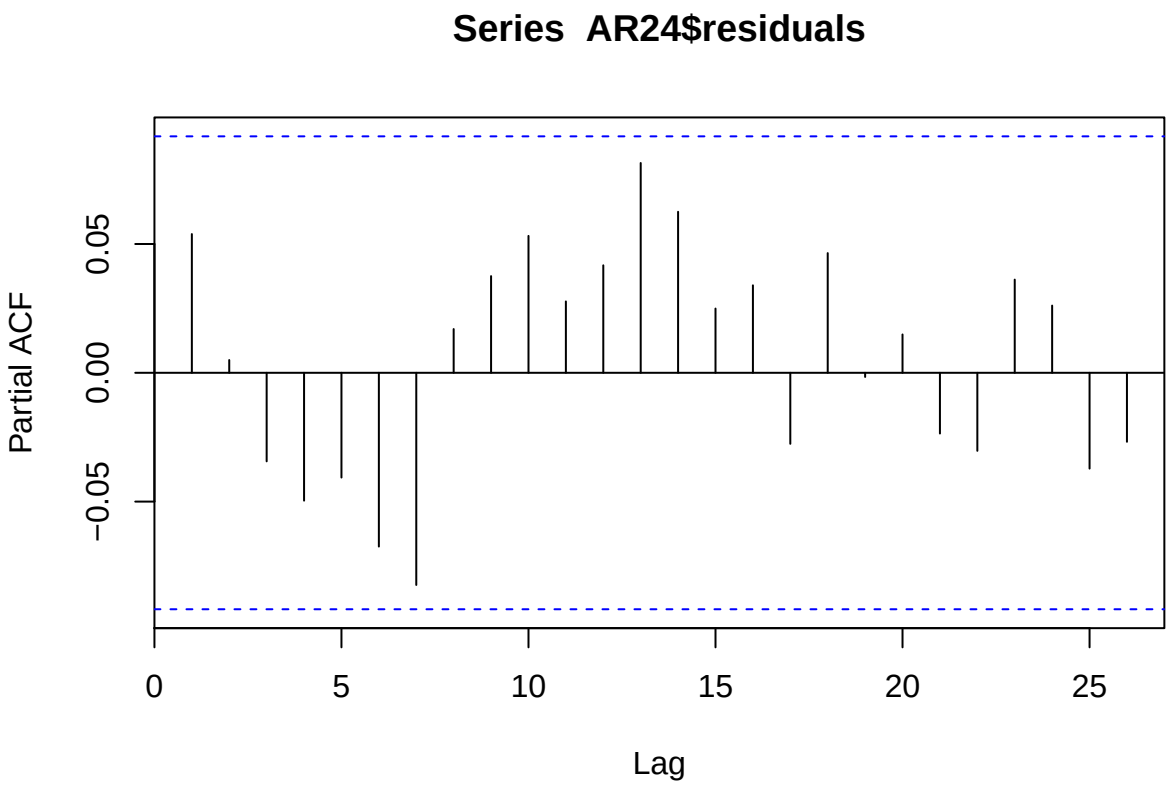
May

```
train17 <- train %>%
  filter(year == '2012' & month == 'May') %>%
  select(datetime, count)

test17 <- test %>%
  filter(year == '2012' & month == 'May') %>%
  mutate(count = NA) %>%
  select(datetime, count)

# head(train17)
# head(test17)

AR24 <- arima(train17$count,order=c(25,0,0))
# tsdisplay(residuals(AR24),lag.max=25,main="AR(24) Resid. Diagnostics")

number = nrow(test17)
```

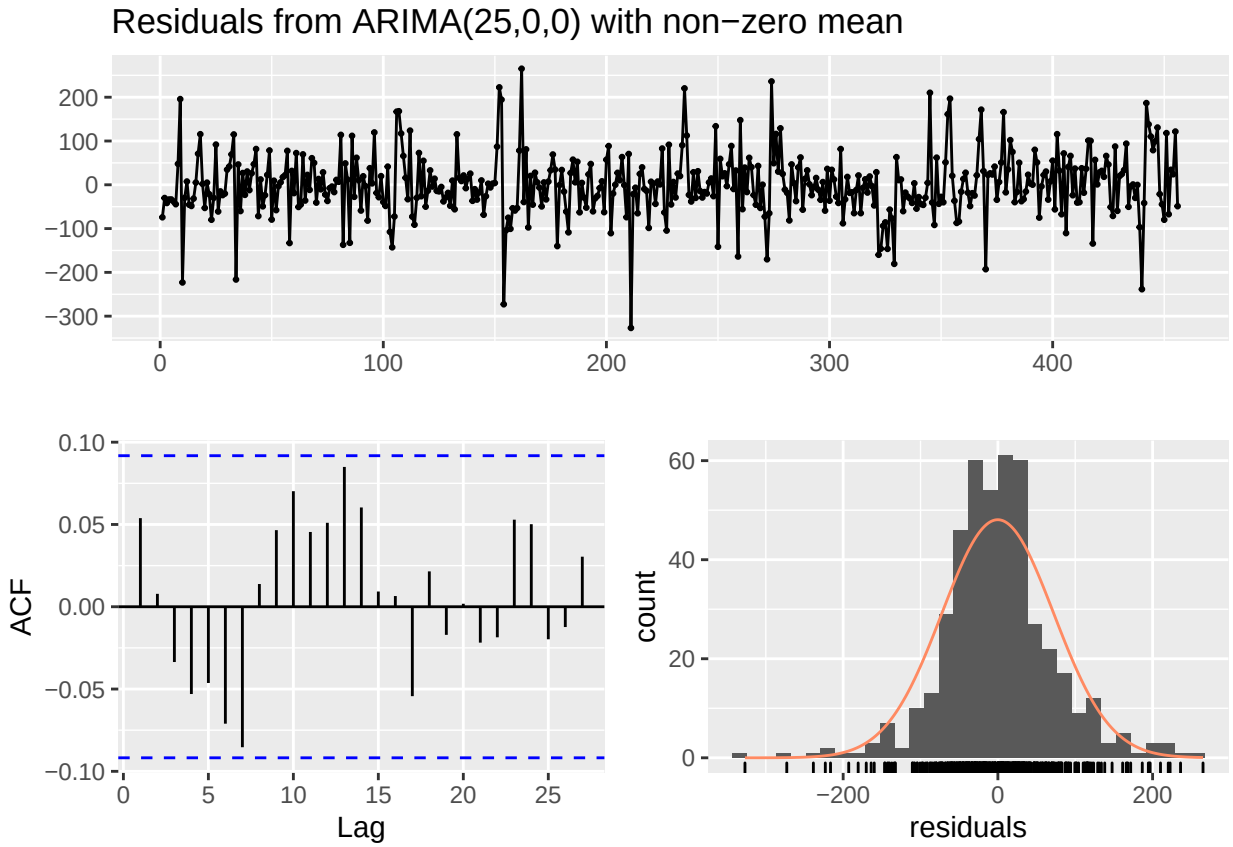
```r
acf(AR24$residuals)
```




```r
pacf(AR24$residuals)
```

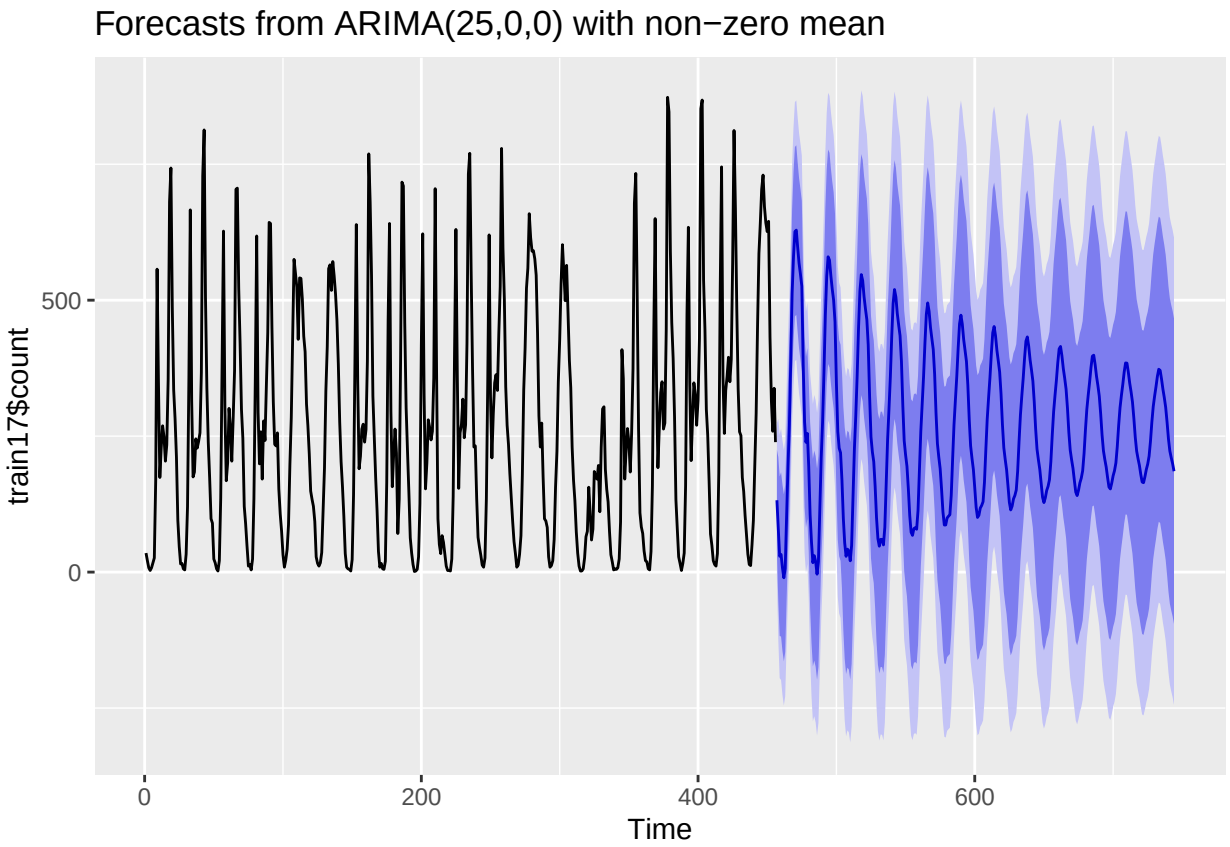



```
checkresiduals(AR24)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0) with non-zero mean
## Q* = 28.85, df = 3, p-value = 2.407e-06
##
## Model df: 26.   Total lags used: 29
```

```
fcst <- forecast(AR24, h=number)

autoplot(fcst)
```



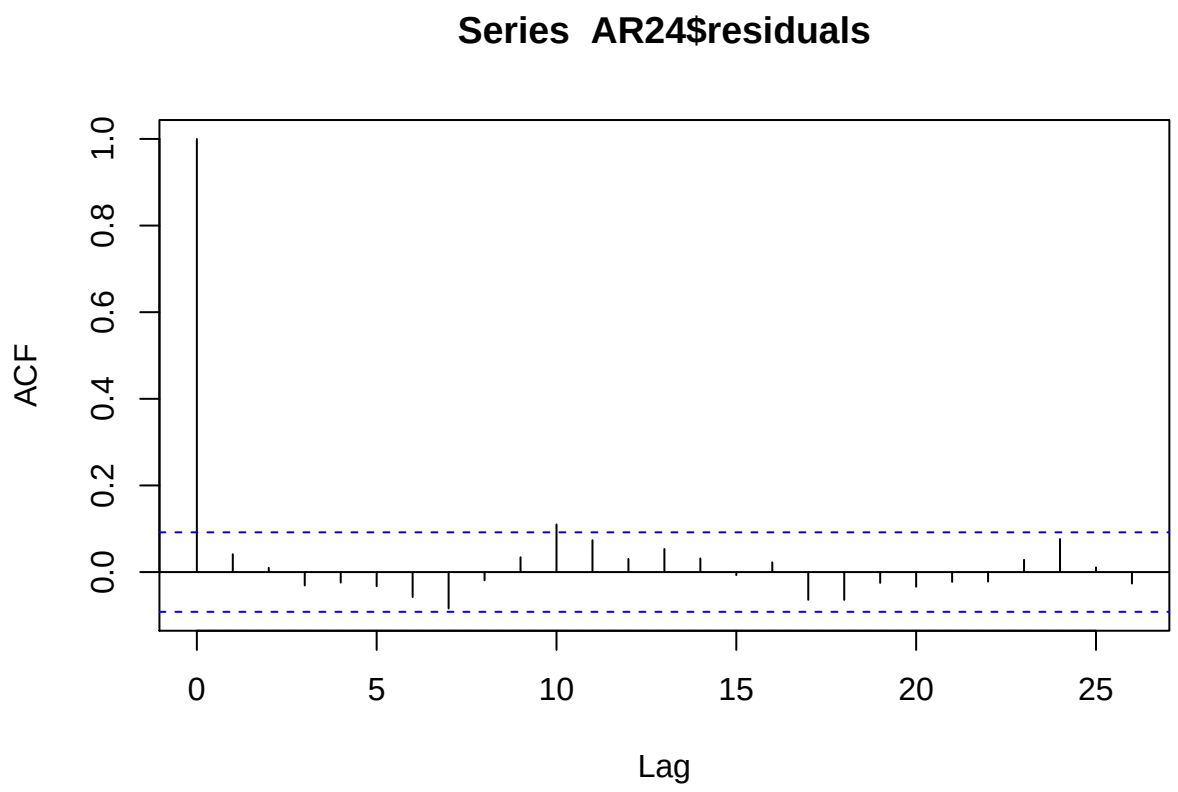

```
# point estimate (mean)
test17$count <- round(fcst$mean)

# test5

RMSLE(y_pred = floor(ifelse(fcst$fitted < 0, 0, round(fcst$fitted))), y_true = train17$count)
```

```
## [1] 0.7627352
```

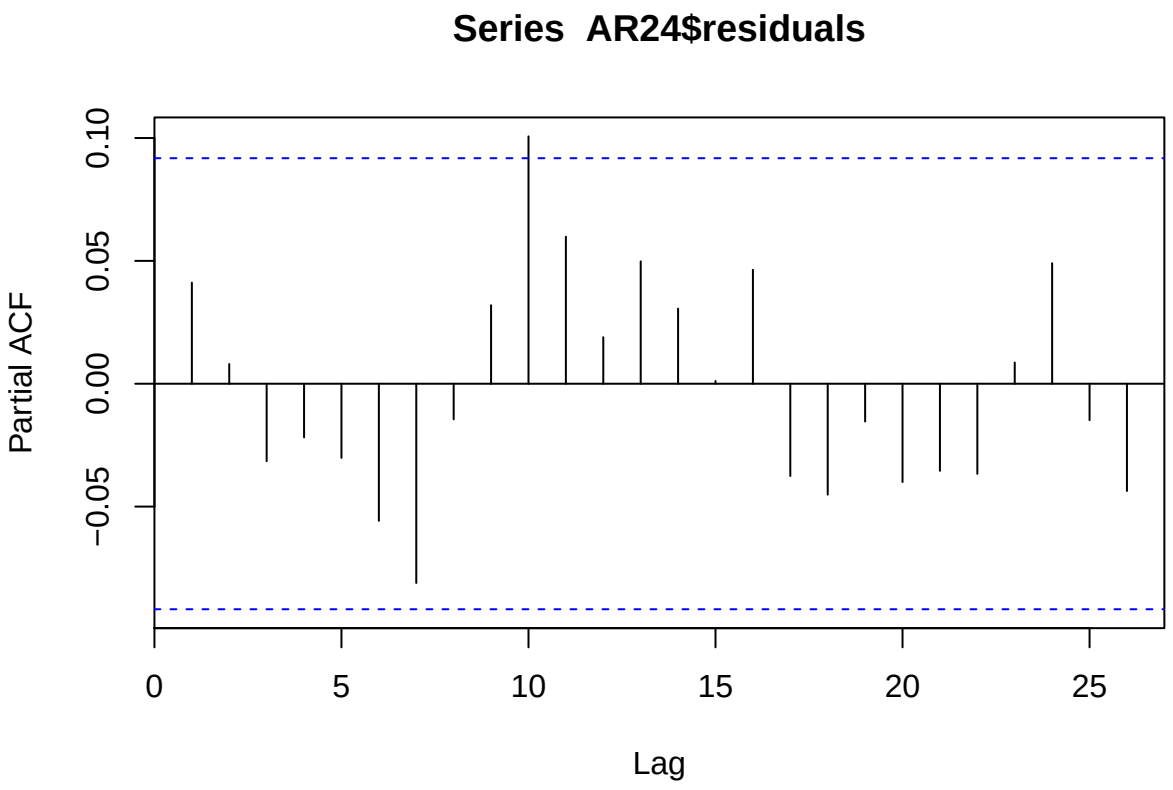
June

```
train18 <- train %>%
  filter(year == '2012' & month == 'June') %>%
  select(datetime, count)

test18 <- test %>%
  filter(year == '2012' & month == 'June') %>%
  mutate(count = NA) %>%
  select(datetime, count)

# head(train18)
# head(test18)

AR24 <- arima(train18$count,order=c(25,0,0))
# tsdisplay(residuals(AR24),lag.max=25,main="AR(24) Resid. Diagnostics")
```

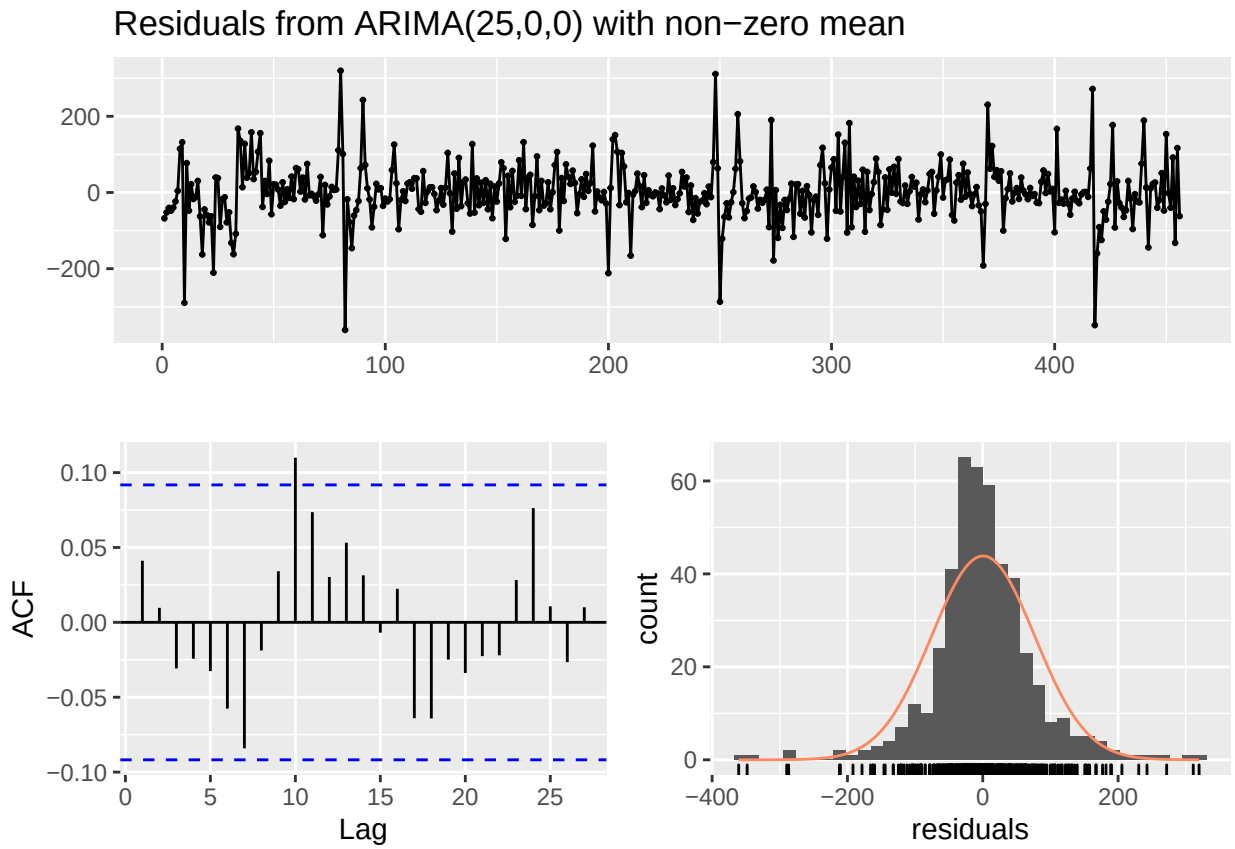
```r
number = nrow(test18)

acf(AR24$residuals)
```




```r
pacf(AR24$residuals)
```

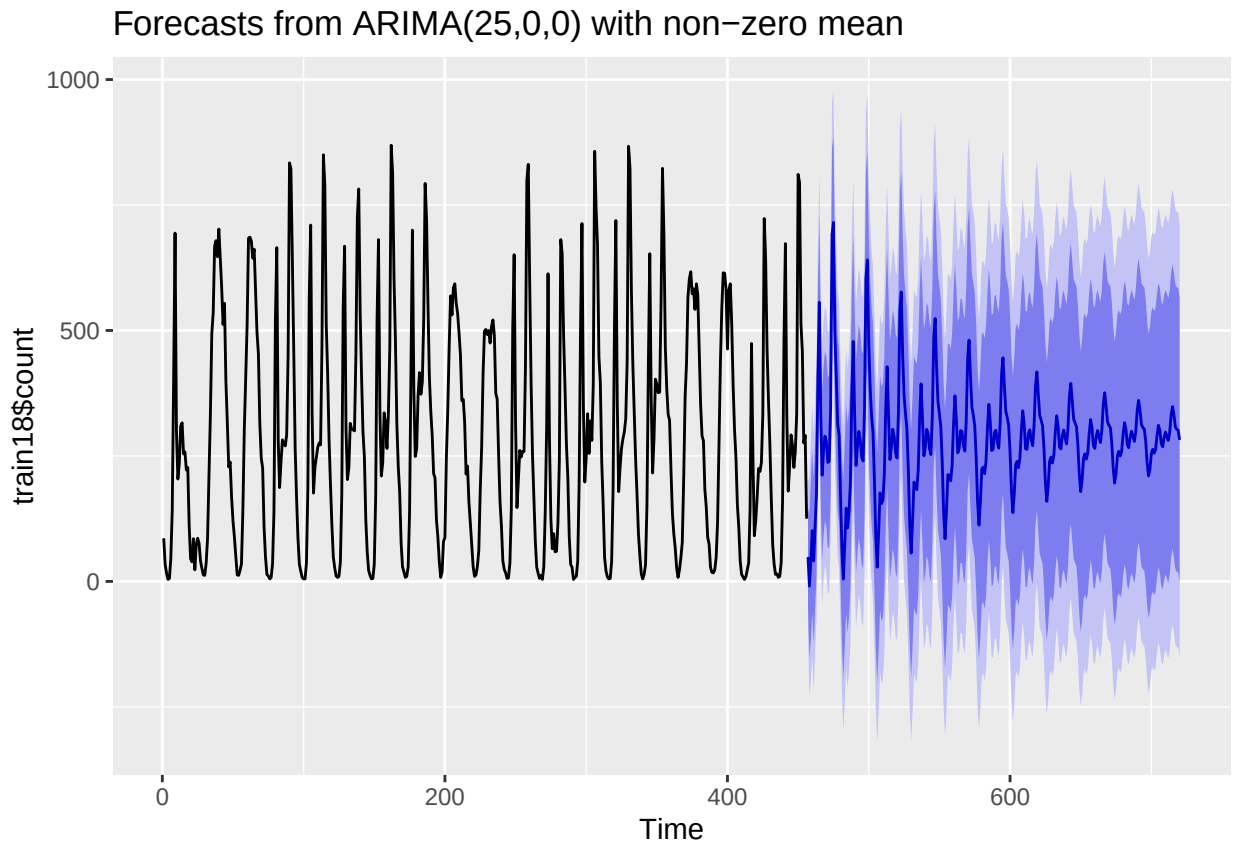



```
checkresiduals(AR24)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0) with non-zero mean
## Q* = 27.266, df = 3, p-value = 5.178e-06
##
## Model df: 26.   Total lags used: 29
```

```
fcst <- forecast(AR24, h=number)

autoplot(fcst)
```



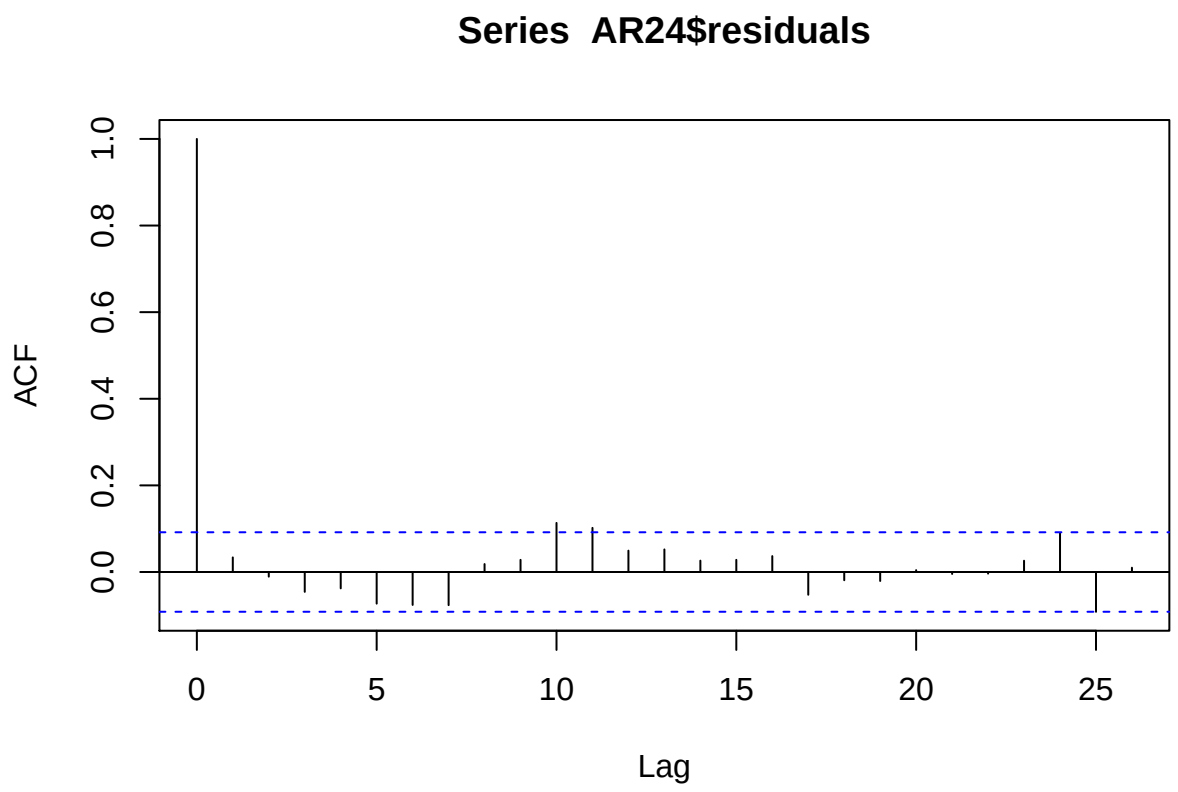

```
# point estimate (mean)
test18$count <- round(fcst$mean)

# test18

RMSLE(y_pred = floor(ifelse(fcst$fitted < 0, 0, round(fcst$fitted))), y_true = train18$count)
```

```
## [1] 0.6855907
```

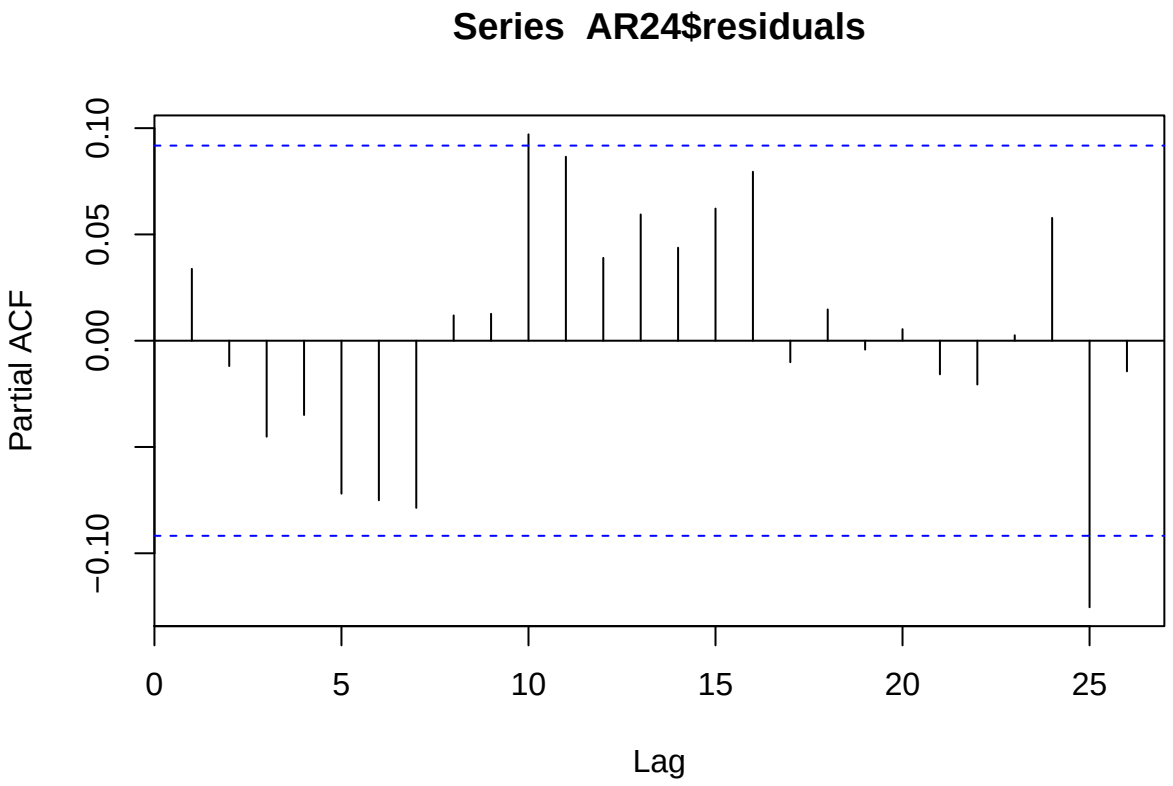
July

```
train19 <- train %>%
  filter(year == '2012' & month == 'July') %>%
  select(datetime, count)

test19 <- test %>%
  filter(year == '2012' & month == 'July') %>%
  mutate(count = NA) %>%
  select(datetime, count)

# head(train19)
# head(test19)

AR24 <- arima(train19$count,order=c(25,0,0))
# tsdisplay(residuals(AR24),lag.max=25,main="AR(24) Resid. Diagnostics")
```

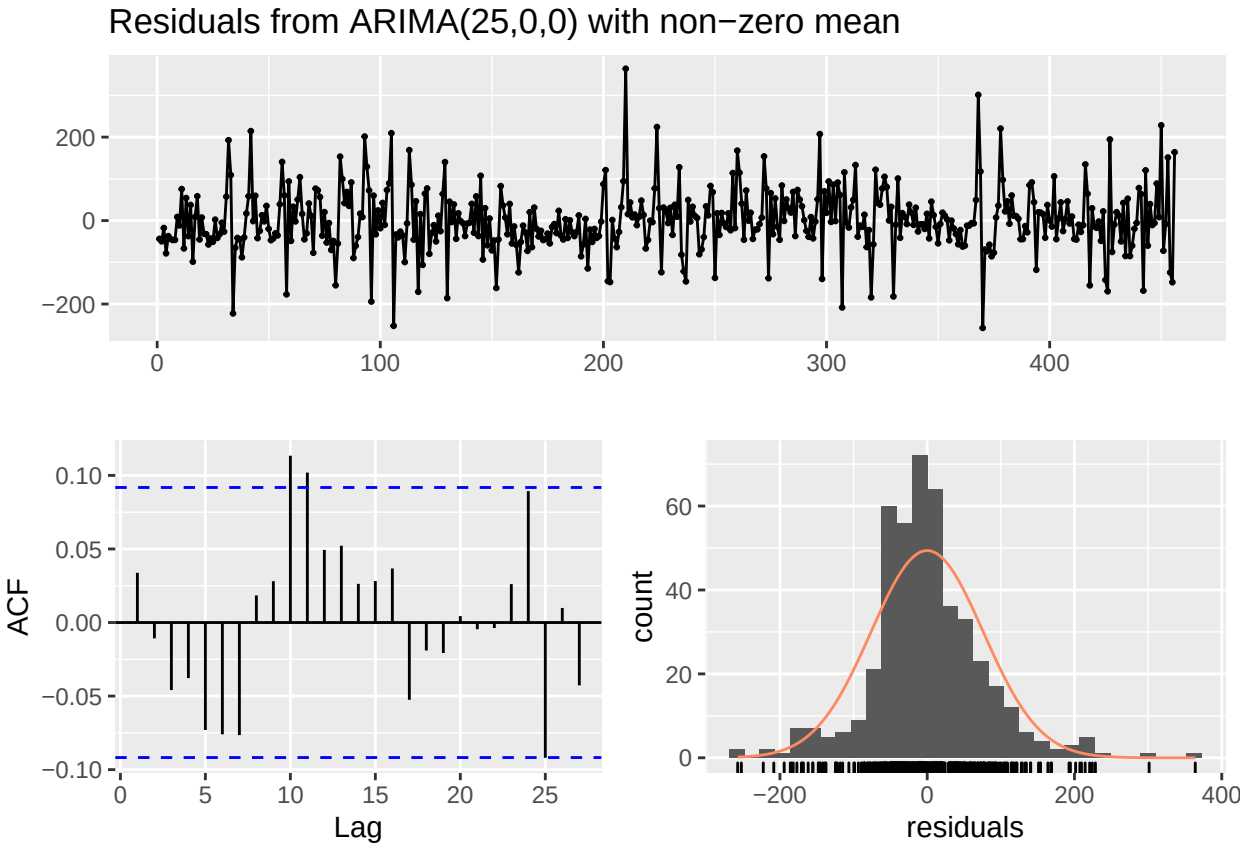
```r
number = nrow(test19)

acf(AR24$residuals)
```




```r
pacf(AR24$residuals)
```

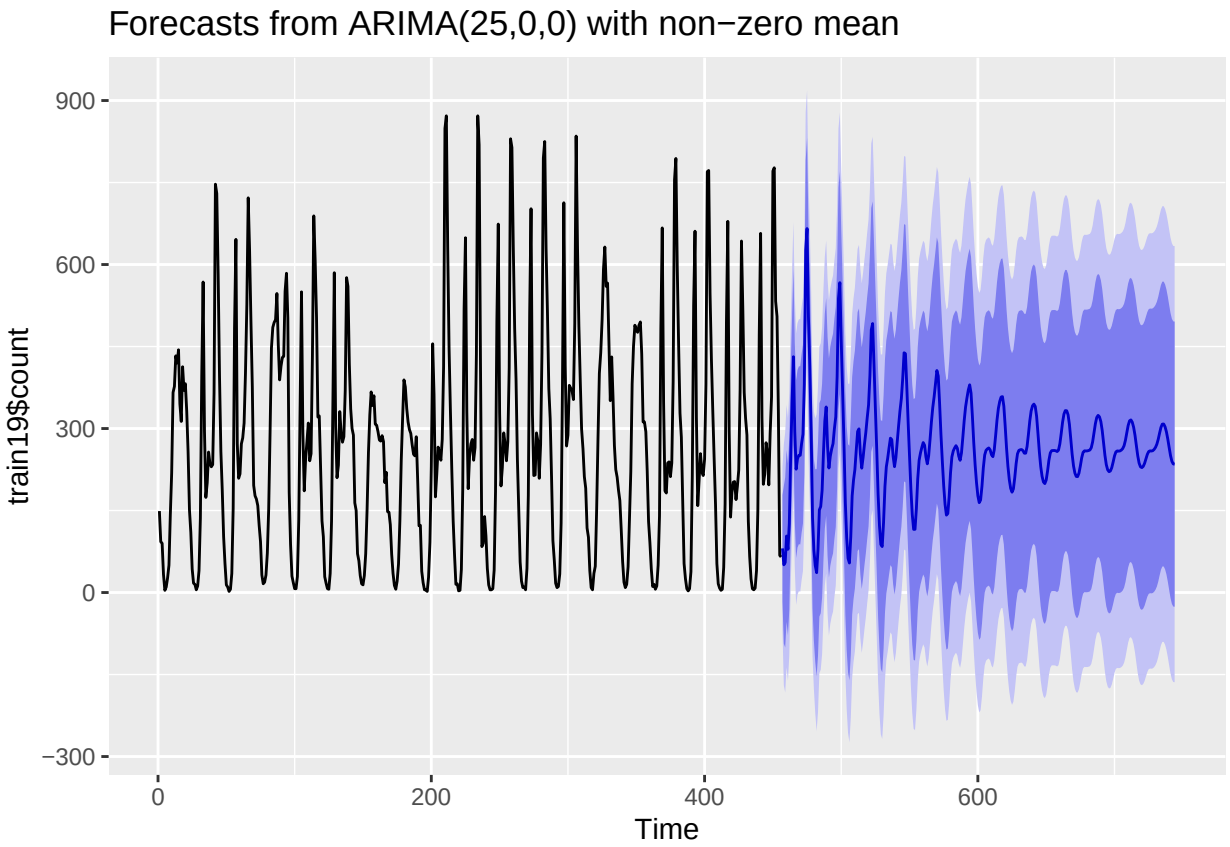



```
checkresiduals(AR24)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0) with non-zero mean
## Q* = 37.927, df = 3, p-value = 2.929e-08
##
## Model df: 26.   Total lags used: 29
```

```
fcst <- forecast(AR24, h=number)

autoplot(fcst)
```



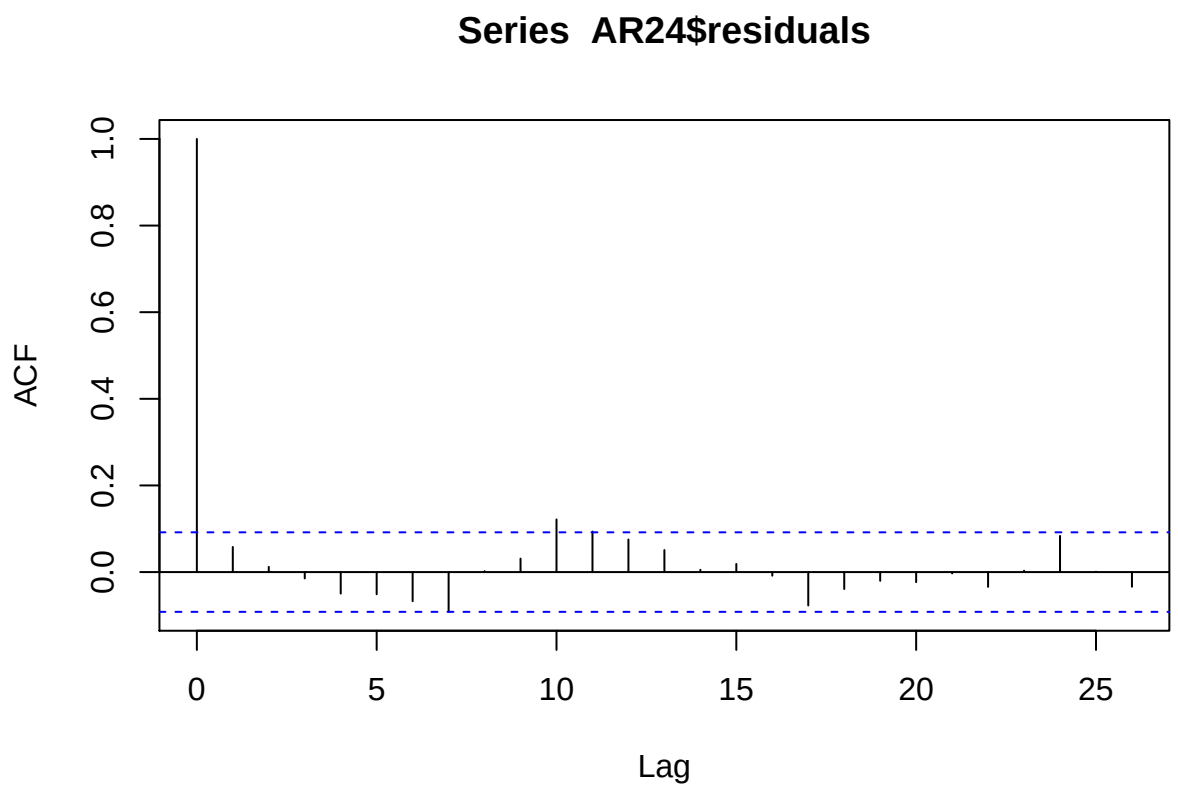

```
# point estimate (mean)
test19$count <- round(fcst$mean)

# test19

RMSLE(y_pred = floor(ifelse(fcst$fitted < 0, 0, round(fcst$fitted))), y_true = train19$count)
```

```
## [1] 0.7432611
```

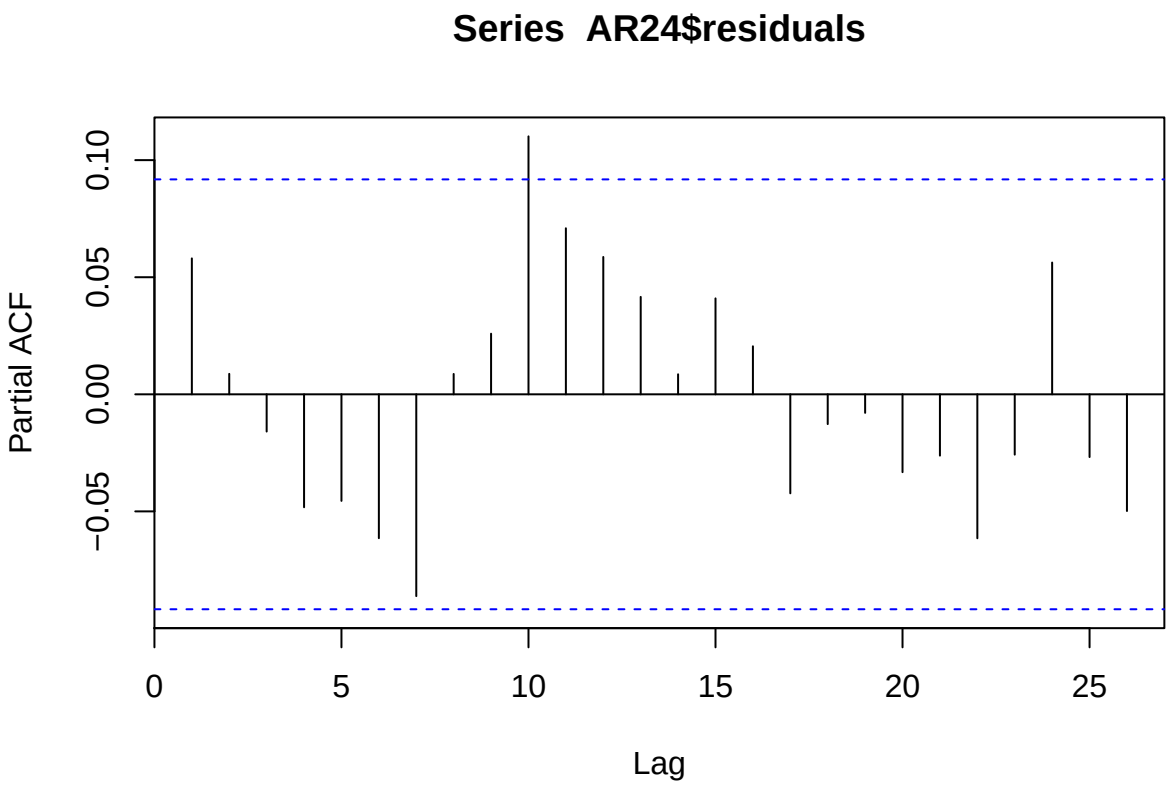
August

```
train20 <- train %>%
  filter(year == '2012' & month == 'August') %>%
  select(datetime, count)

test20 <- test %>%
  filter(year == '2012' & month == 'August') %>%
  mutate(count = NA) %>%
  select(datetime, count)

# head(train20)
# head(test20)

AR24 <- arima(train20$count,order=c(25,0,0))
# tsdisplay(residuals(AR24),lag.max=25,main="AR(24) Resid. Diagnostics")
```

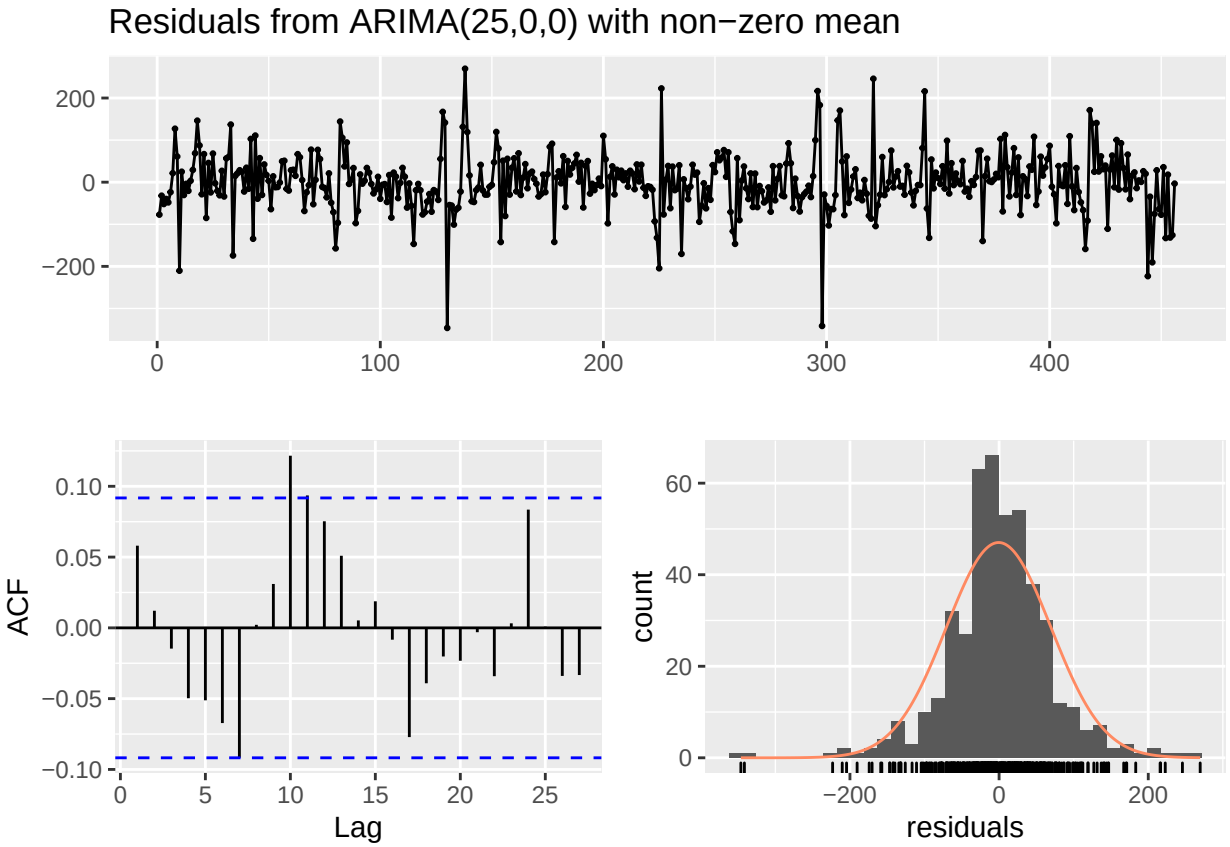
```r
number = nrow(test20)

acf(AR24$residuals)
```




```r
pacf(AR24$residuals)
```

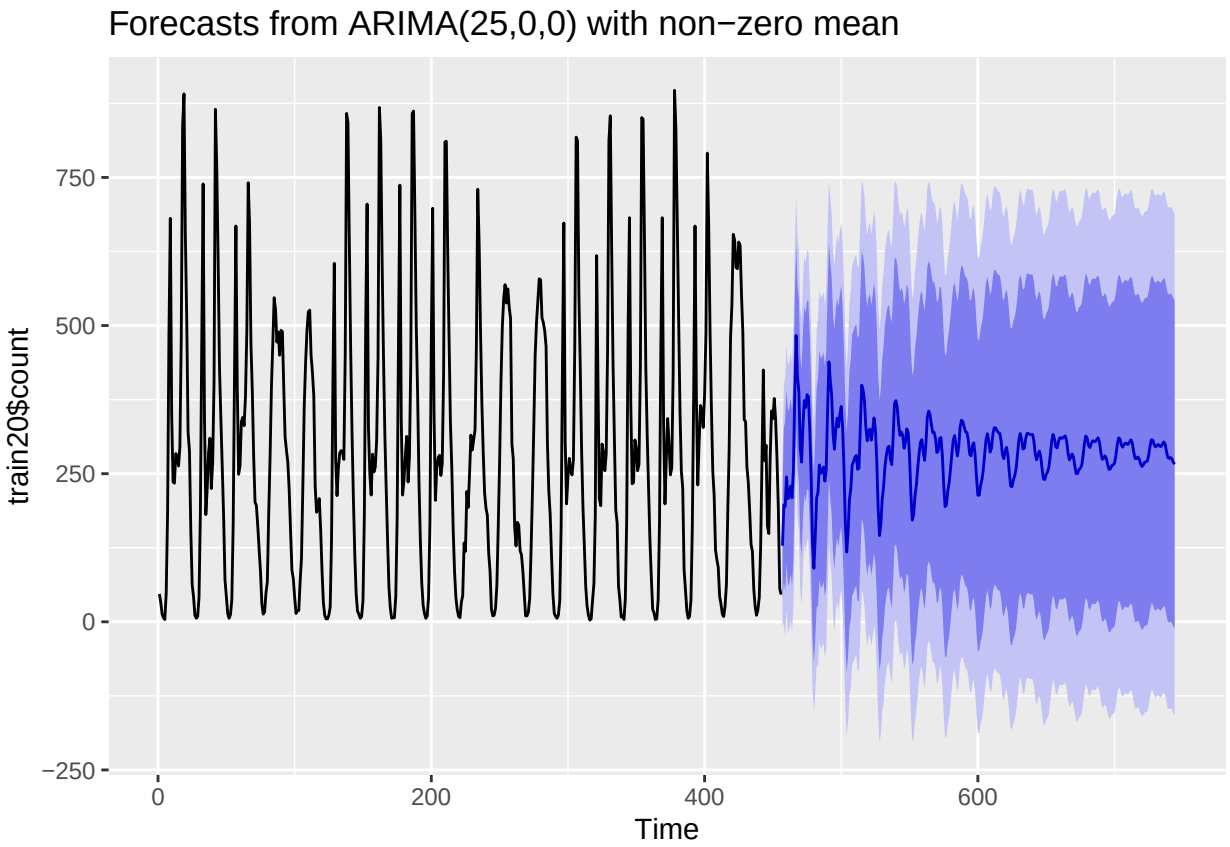



```
checkresiduals(AR24)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0) with non-zero mean
## Q* = 35.45, df = 3, p-value = 9.786e-08
##
## Model df: 26.   Total lags used: 29
```

```
fcst <- forecast(AR24, h=number)

autoplot(fcst)
```



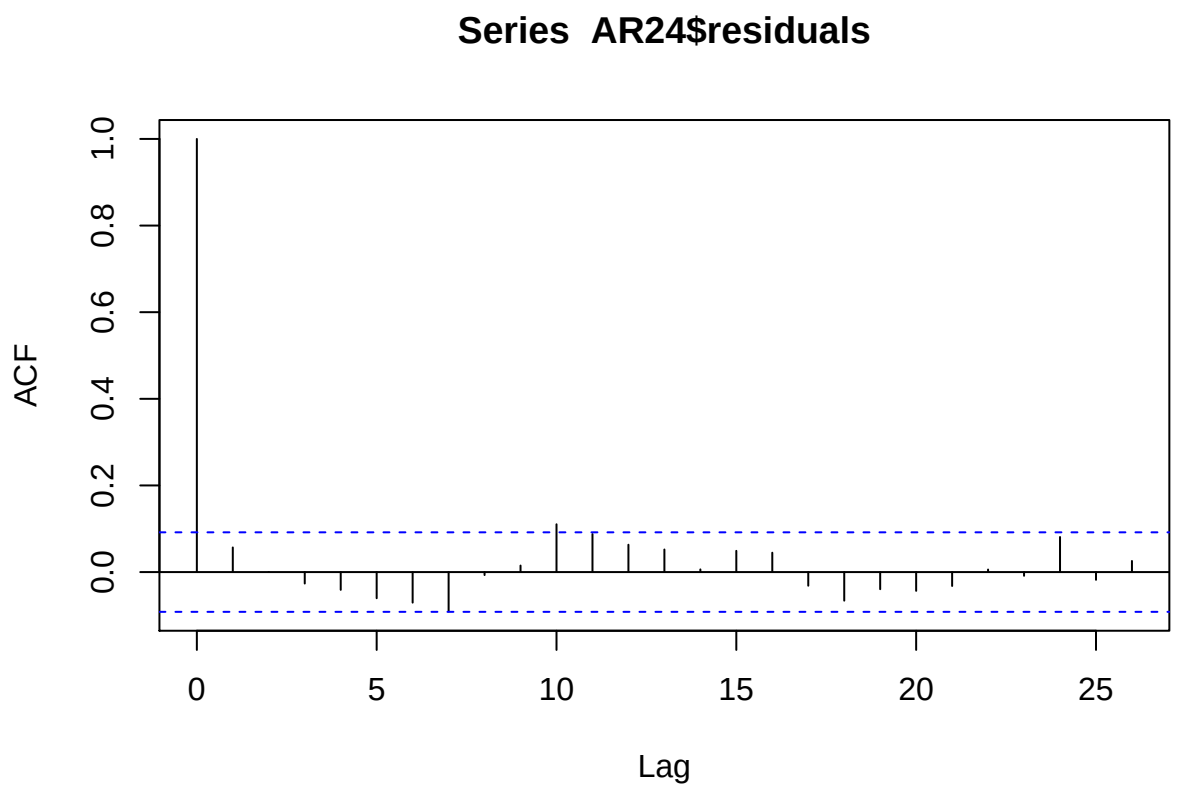

```
# point estimate (mean)
test20$count <- round(fcst$mean)

# test20

RMSLE(y_pred = floor(ifelse(fcst$fitted < 0, 0, round(fcst$fitted))), y_true = train20$count)
```

```
## [1] 0.7654981
```

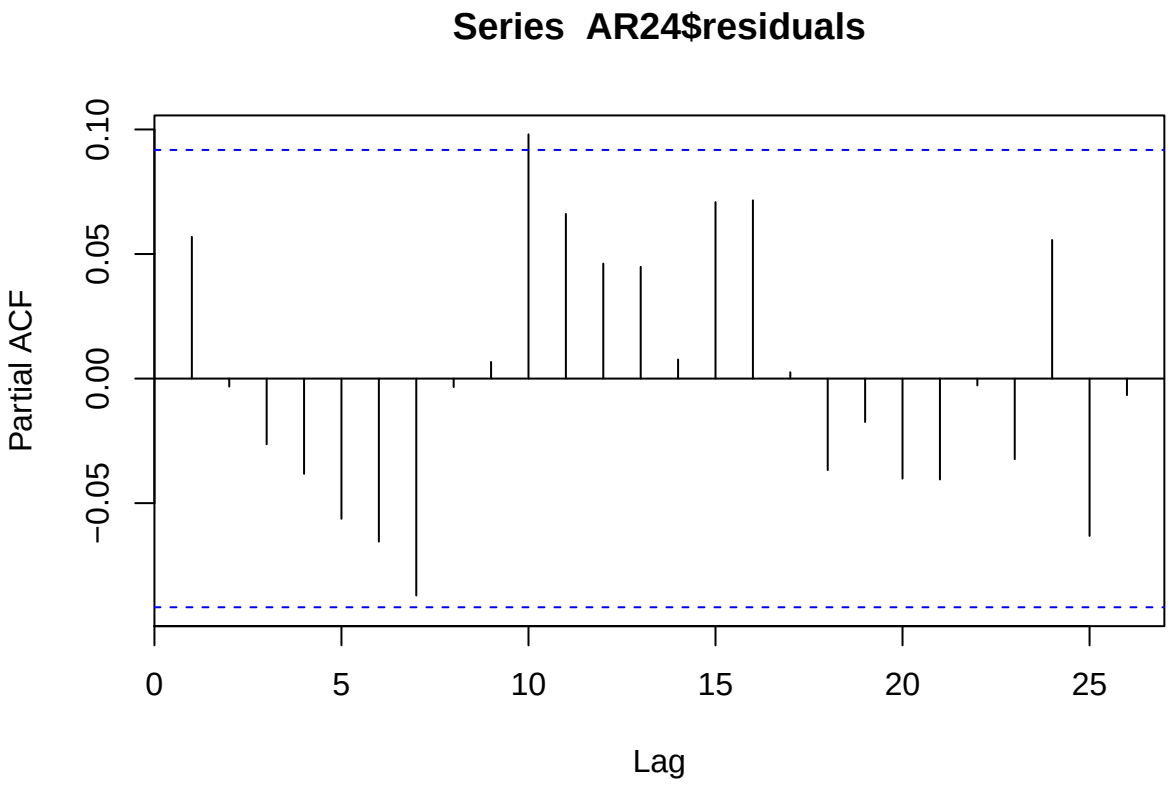
September

```
train21 <- train %>%
  filter(year == '2012' & month == 'September') %>%
  select(datetime, count)

test21 <- test %>%
  filter(year == '2012' & month == 'September') %>%
  mutate(count = NA) %>%
  select(datetime, count)

# head(train21)
# head(test21)

AR24 <- arima(train21$count,order=c(25,0,0))
# tsdisplay(residuals(AR24),lag.max=25,main="AR(24) Resid. Diagnostics")
```

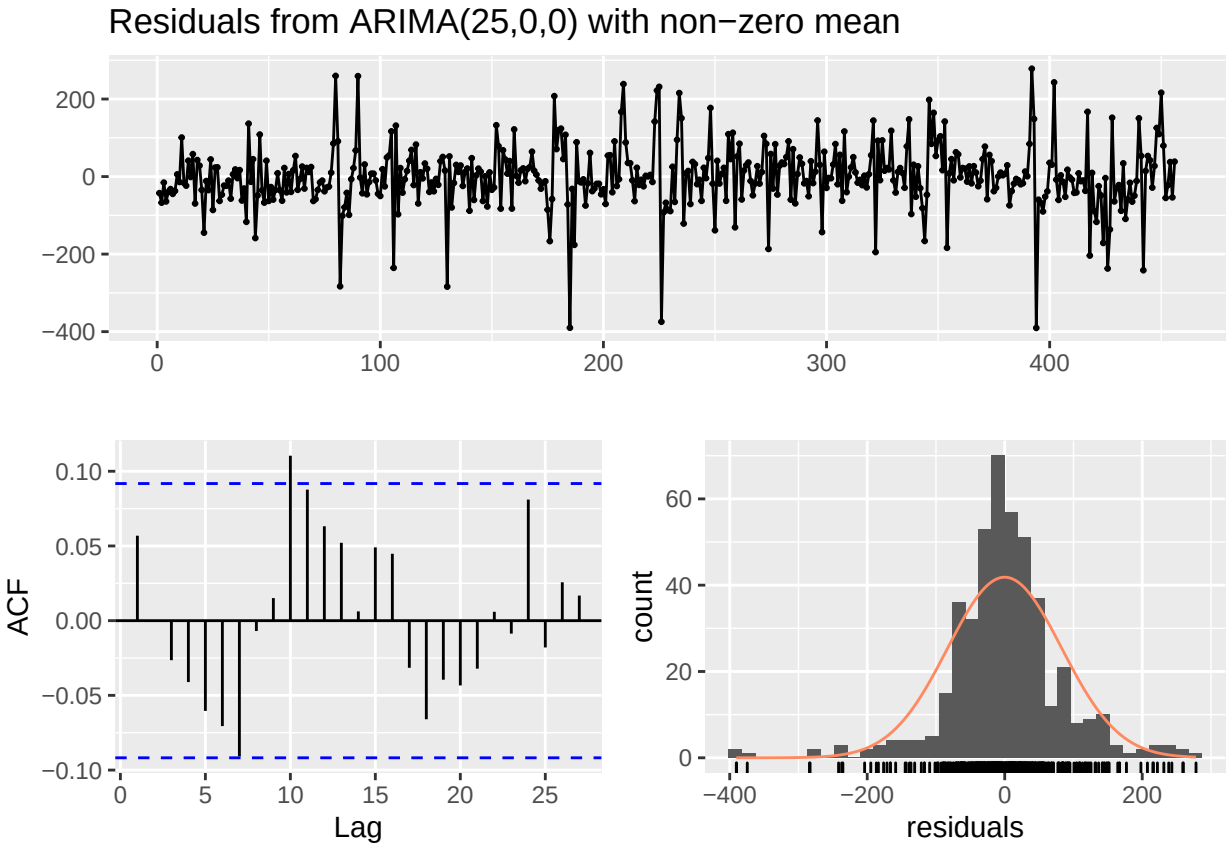
```r
number = nrow(test21)

acf(AR24$residuals)
```




```r
pacf(AR24$residuals)
```

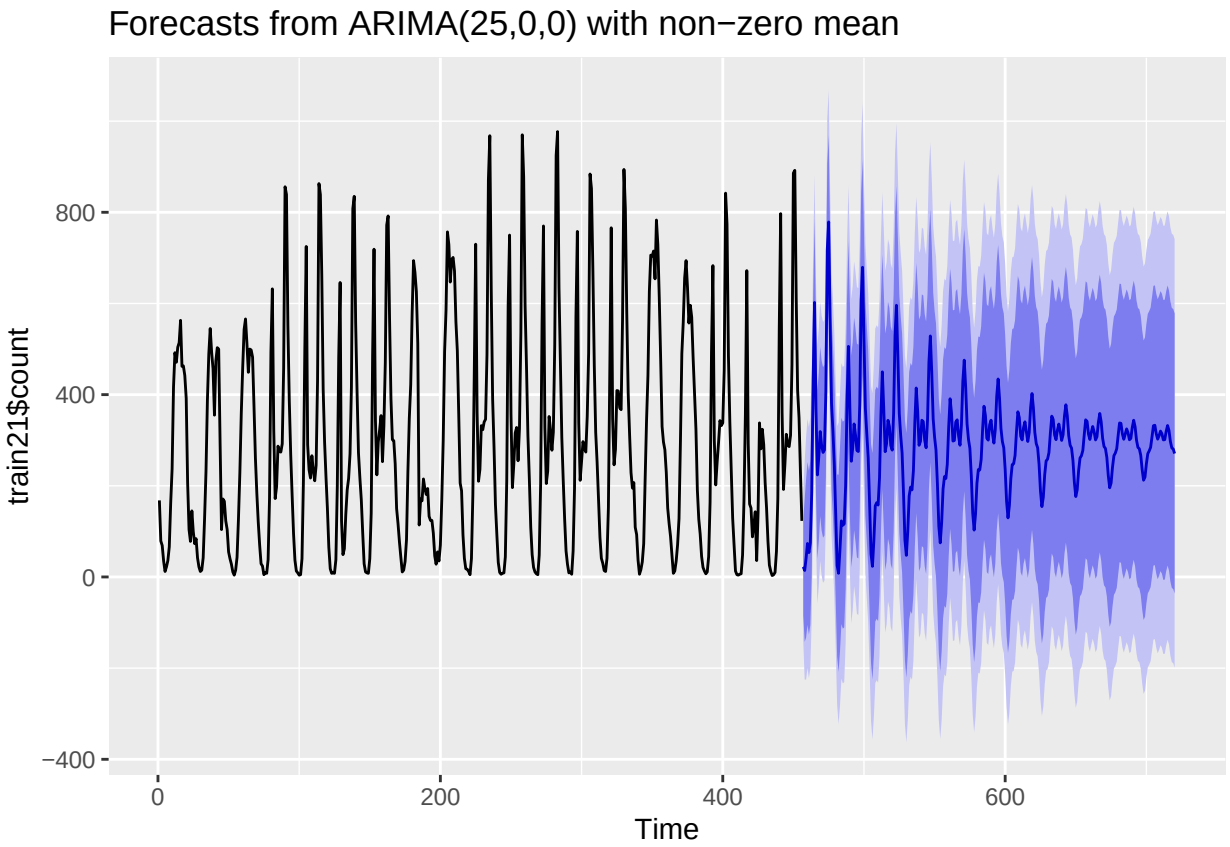



```
checkresiduals(AR24)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0) with non-zero mean
## Q* = 33.797, df = 3, p-value = 2.186e-07
##
## Model df: 26.   Total lags used: 29
```

```
fcst <- forecast(AR24, h=number)

autoplot(fcst)
```



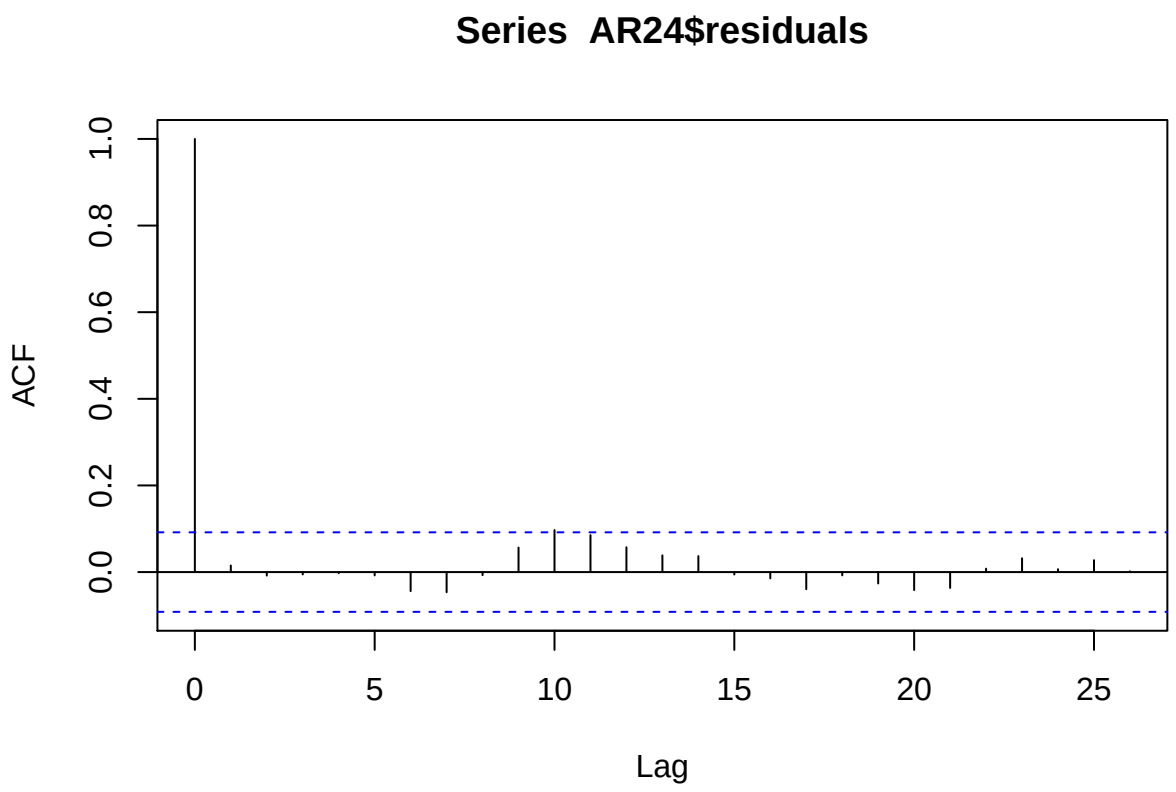

```
# point estimate (mean)
test21$count <- round(fcst$mean)

# test21

RMSLE(y_pred = floor(ifelse(fcst$fitted < 0, 0, round(fcst$fitted))), y_true = train21$count)
```

```
## [1] 0.7018709
```

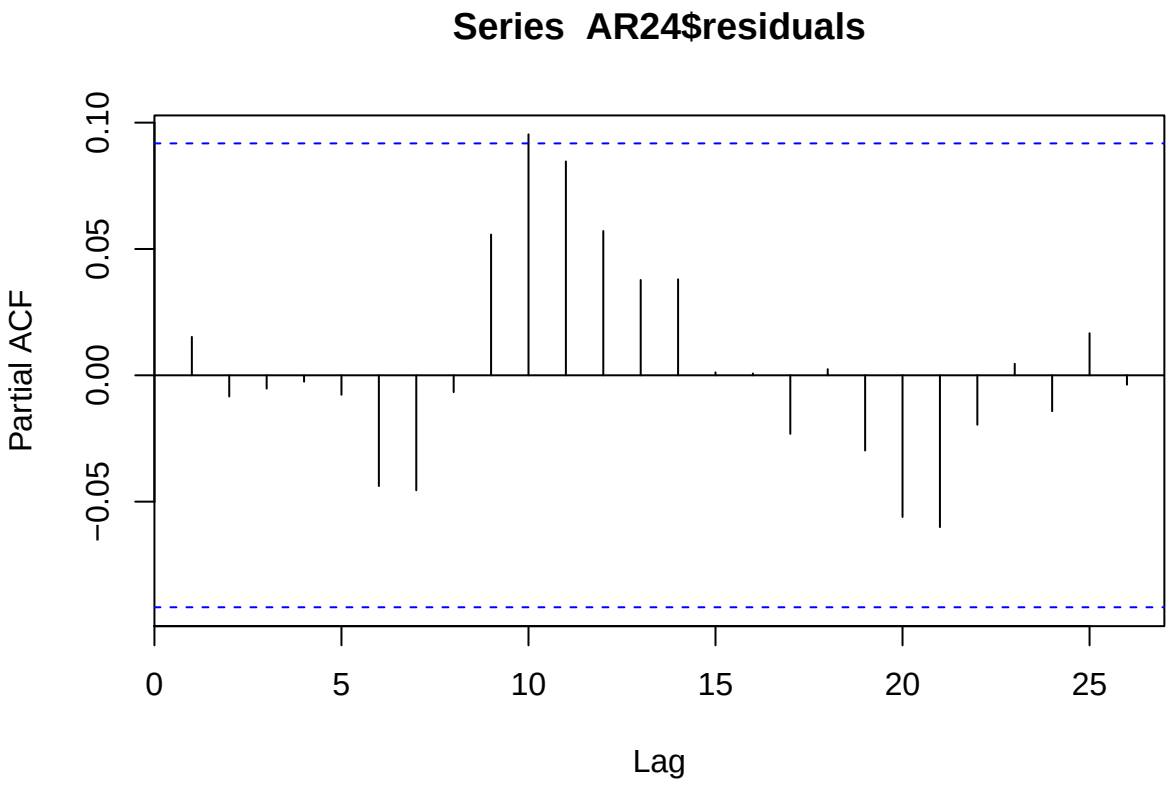
October

```
train22 <- train %>%
  filter(year == '2012' & month == 'October') %>%
  select(datetime, count)

test22 <- test %>%
  filter(year == '2012' & month == 'October') %>%
  mutate(count = NA) %>%
  select(datetime, count)

# head(train22)
# head(test22)

AR24 <- arima(train22$count,order=c(25,0,0))
# tsdisplay(residuals(AR24),lag.max=25,main="AR(24) Resid. Diagnostics")
```

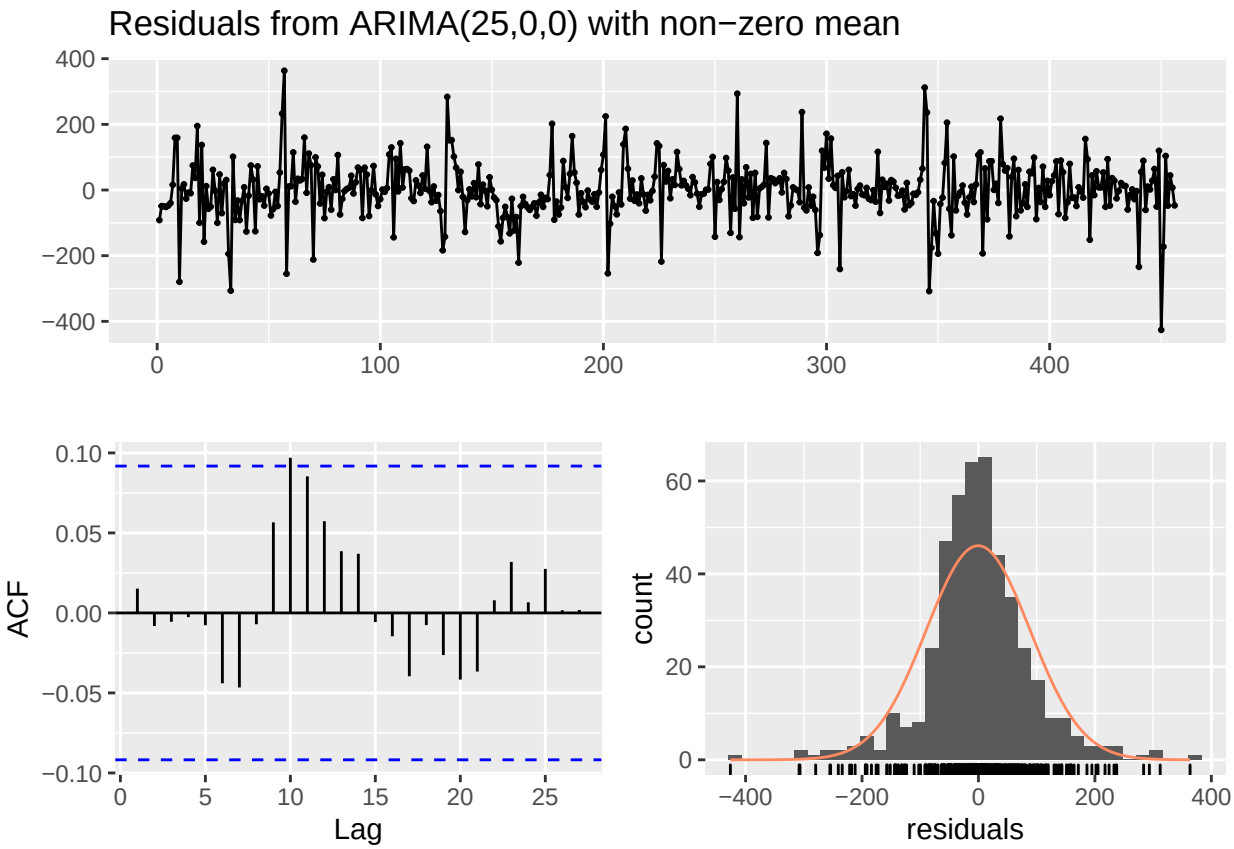
```r
number = nrow(test22)

acf(AR24$residuals)
```




```r
pacf(AR24$residuals)
```

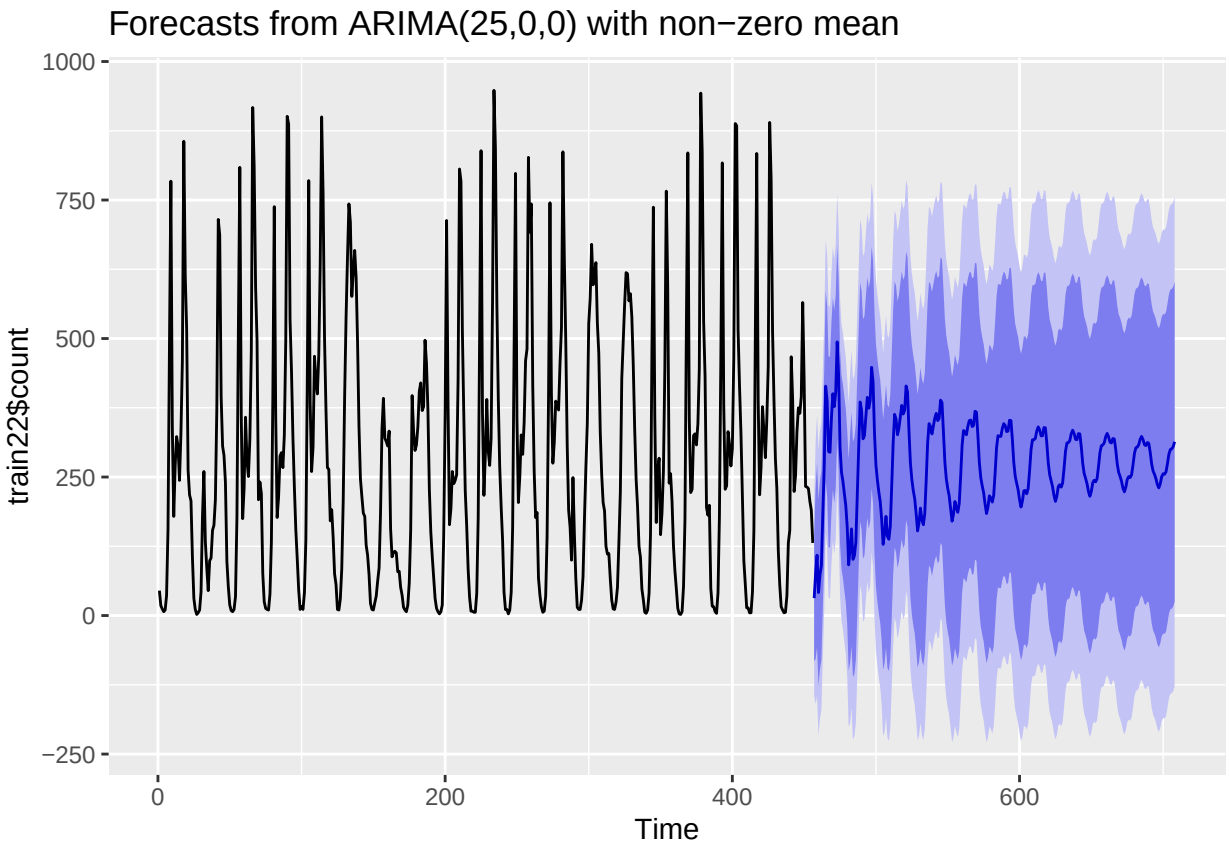



```
checkresiduals(AR24)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0) with non-zero mean
## Q* = 19.986, df = 3, p-value = 0.0001709
##
## Model df: 26.   Total lags used: 29
```

```
fcst <- forecast(AR24, h=number)

autoplot(fcst)
```



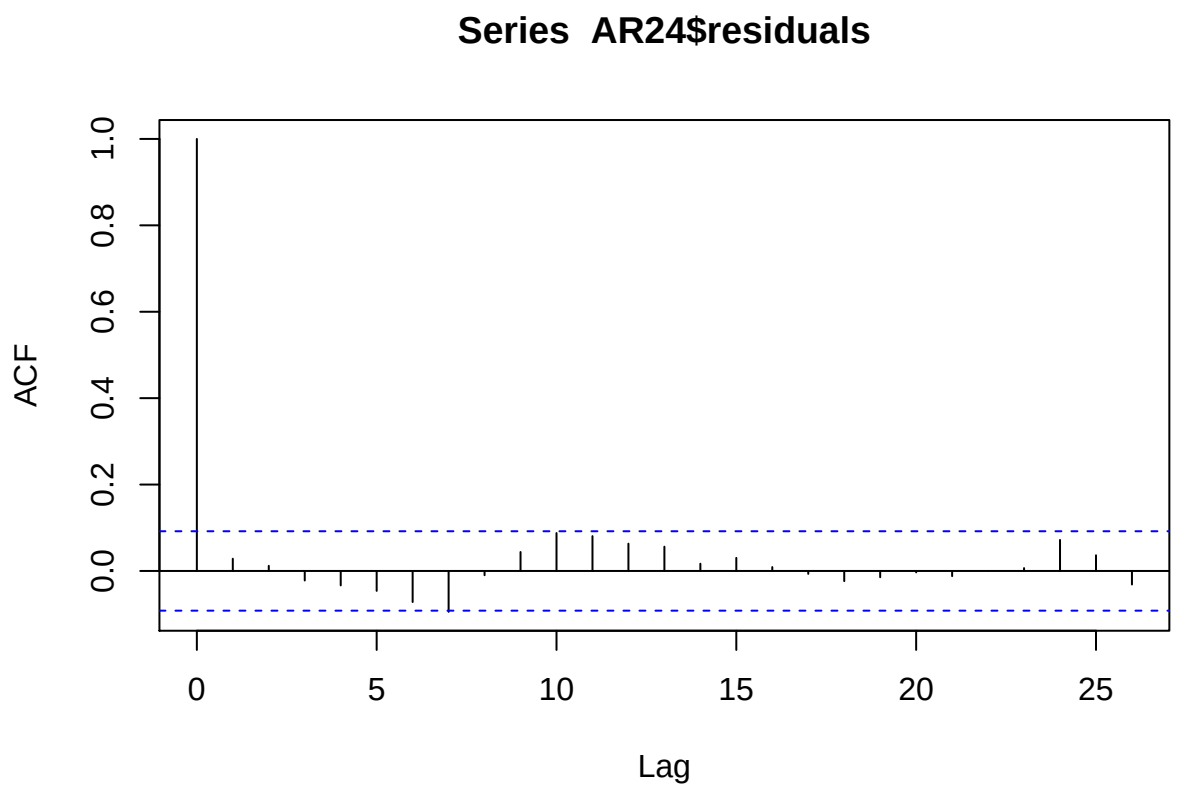

```
# point estimate (mean)
test22$count <- round(fcst$mean)

# test22

RMSLE(y_pred = floor(ifelse(fcst$fitted < 0, 0, round(fcst$fitted))), y_true = train22$count)
```

```
## [1] 0.8302162
```

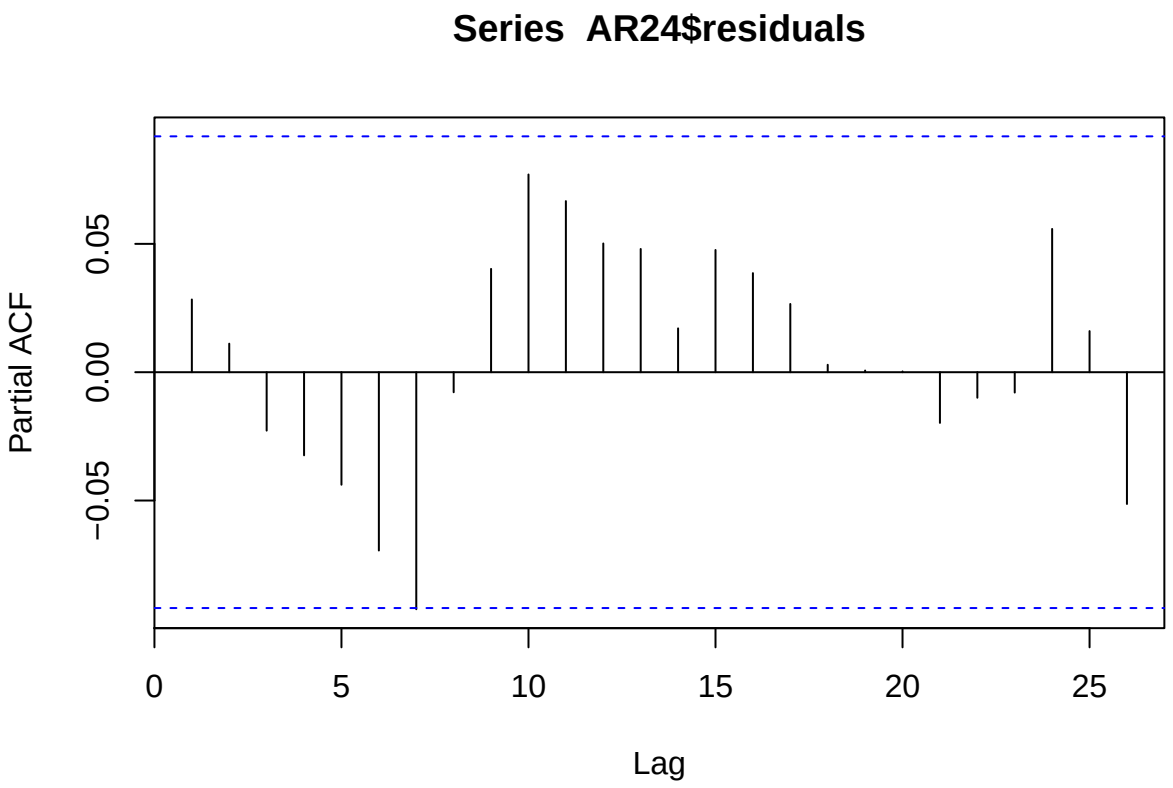
November

```
train23 <- train %>%
  filter(year == '2012' & month == 'November') %>%
  select(datetime, count)

test23 <- test %>%
  filter(year == '2012' & month == 'November') %>%
  mutate(count = NA) %>%
  select(datetime, count)

# head(train23)
# head(test23)

AR24 <- arima(train23$count,order=c(25,0,0))
# tsdisplay(residuals(AR24),lag.max=25,main="AR(24) Resid. Diagnostics")
```

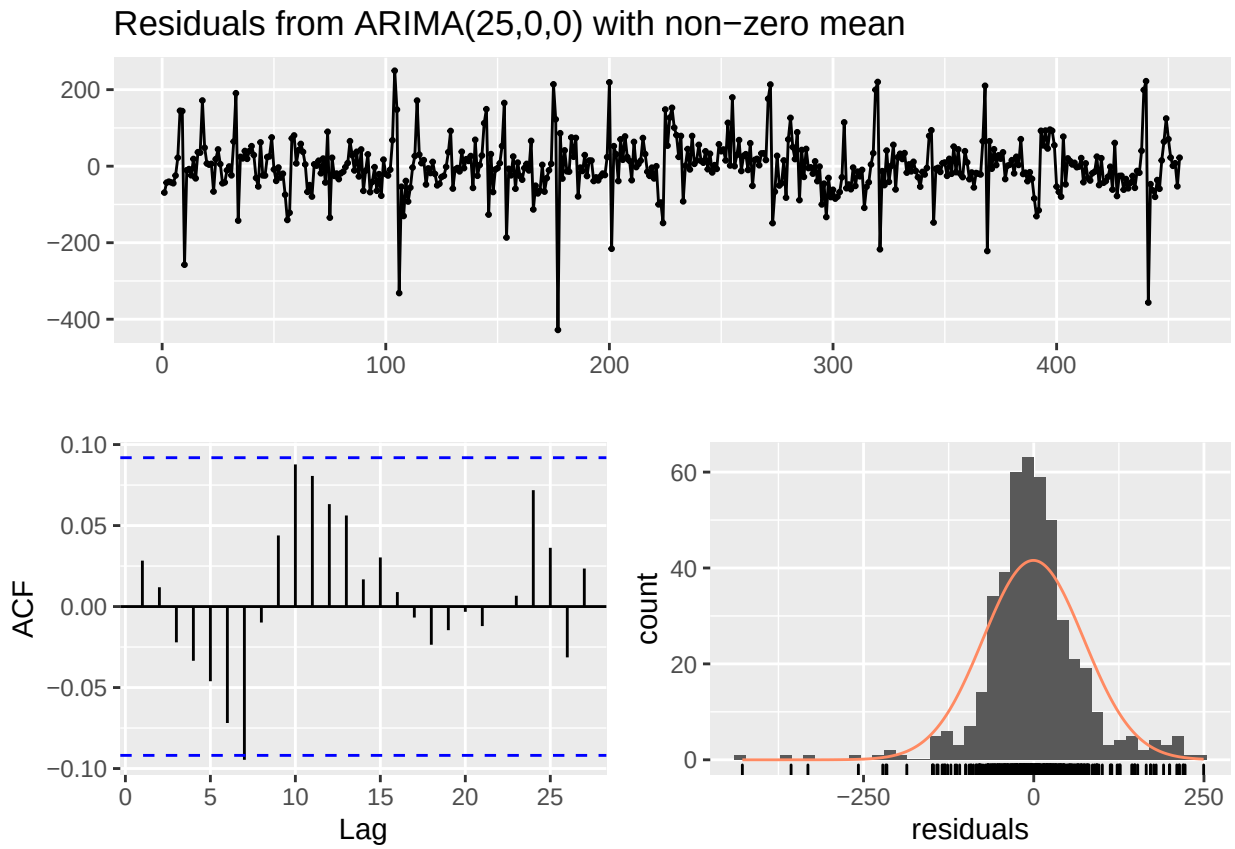
```r
number = nrow(test23)

acf(AR24$residuals)
```




```r
pacf(AR24$residuals)
```

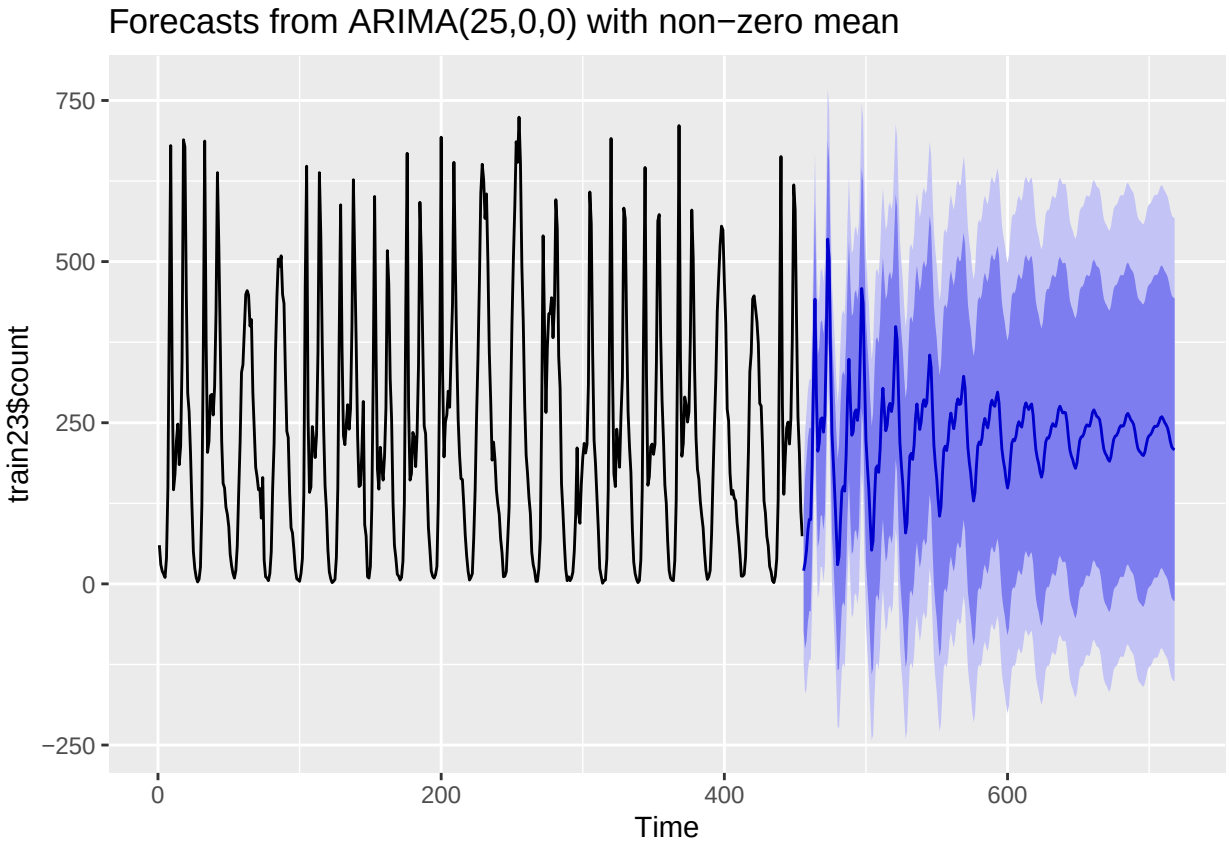



```
checkresiduals(AR24)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0) with non-zero mean
## Q* = 24.672, df = 3, p-value = 1.808e-05
##
## Model df: 26.   Total lags used: 29
```

```
fcst <- forecast(AR24, h=number)

autoplot(fcst)
```



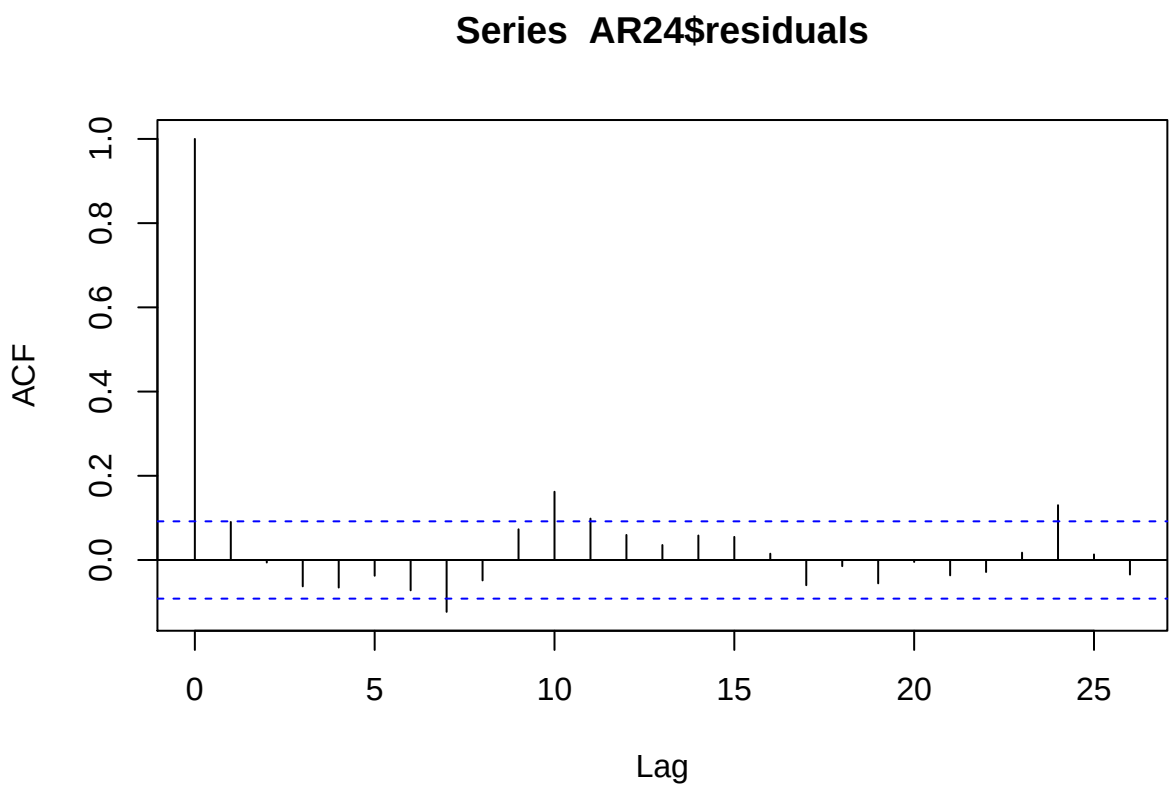

```
# point estimate (mean)
test23$count <- round(fcst$mean)

# test23

RMSLE(y_pred = floor(ifelse(fcst$fitted < 0, 0, round(fcst$fitted))), y_true = train23$count)
```

```
## [1] 0.747518
```

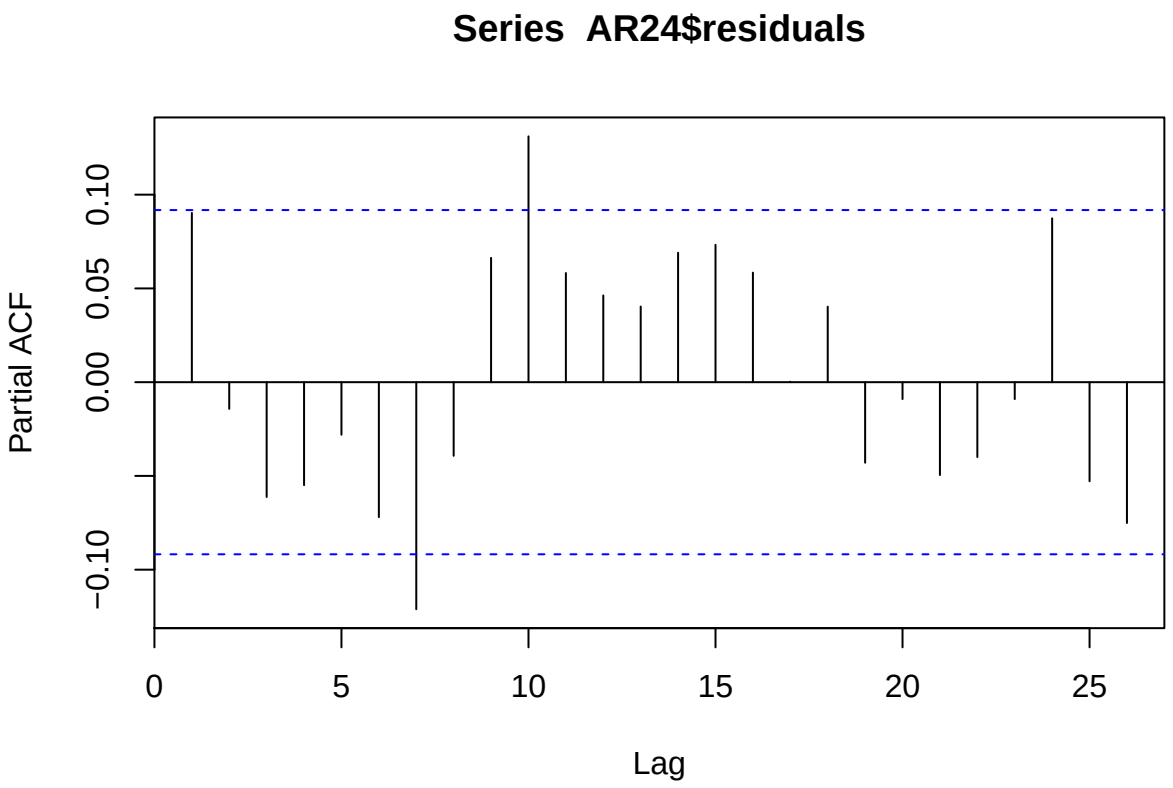
December

```
train24 <- train %>%
  filter(year == '2012' & month == 'December') %>%
  select(datetime, count)

test24 <- test %>%
  filter(year == '2012' & month == 'December') %>%
  mutate(count = NA) %>%
  select(datetime, count)

# head(train24)
# head(test24)

AR24 <- arima(train24$count,order=c(25,0,0))
# tsdisplay(residuals(AR24),lag.max=25,main="AR(24) Resid. Diagnostics")
```

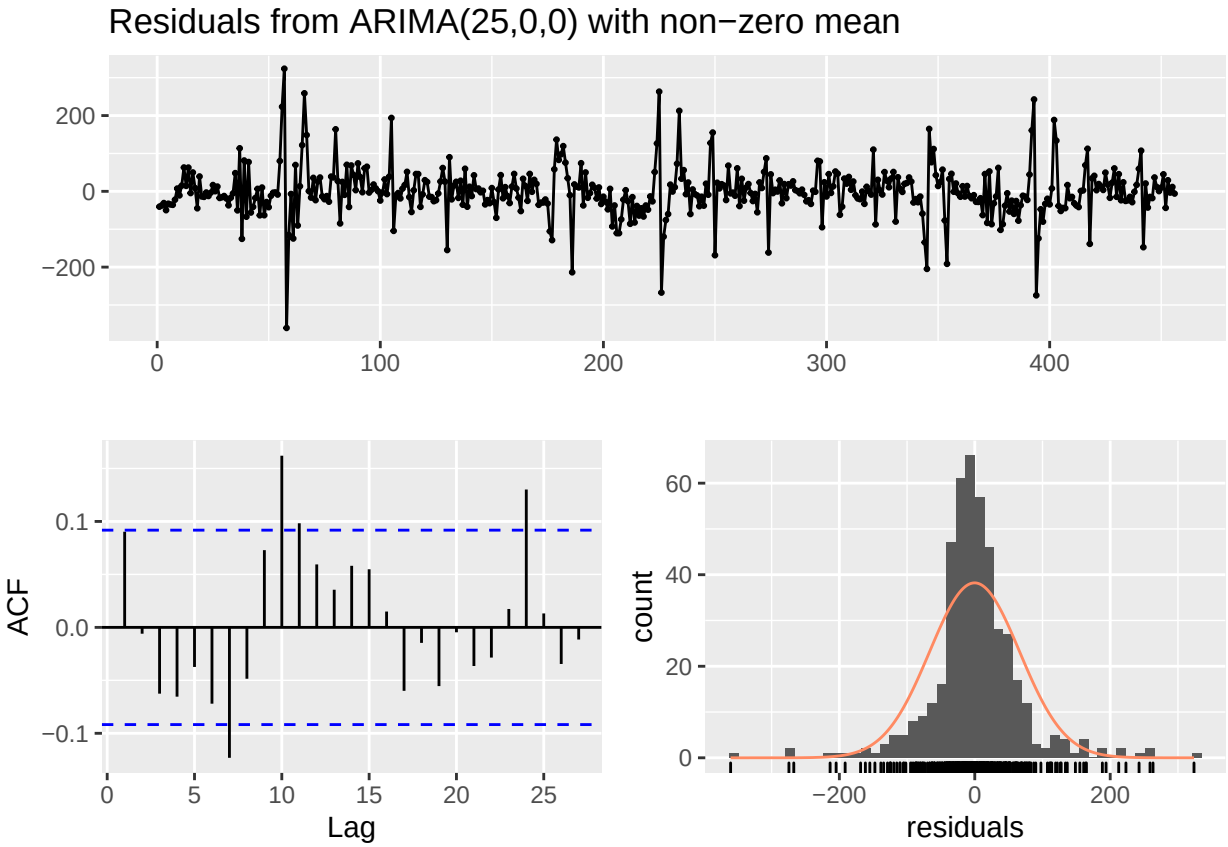
```r
number = nrow(test24)

acf(AR24$residuals)
```




```r
pacf(AR24$residuals)
```

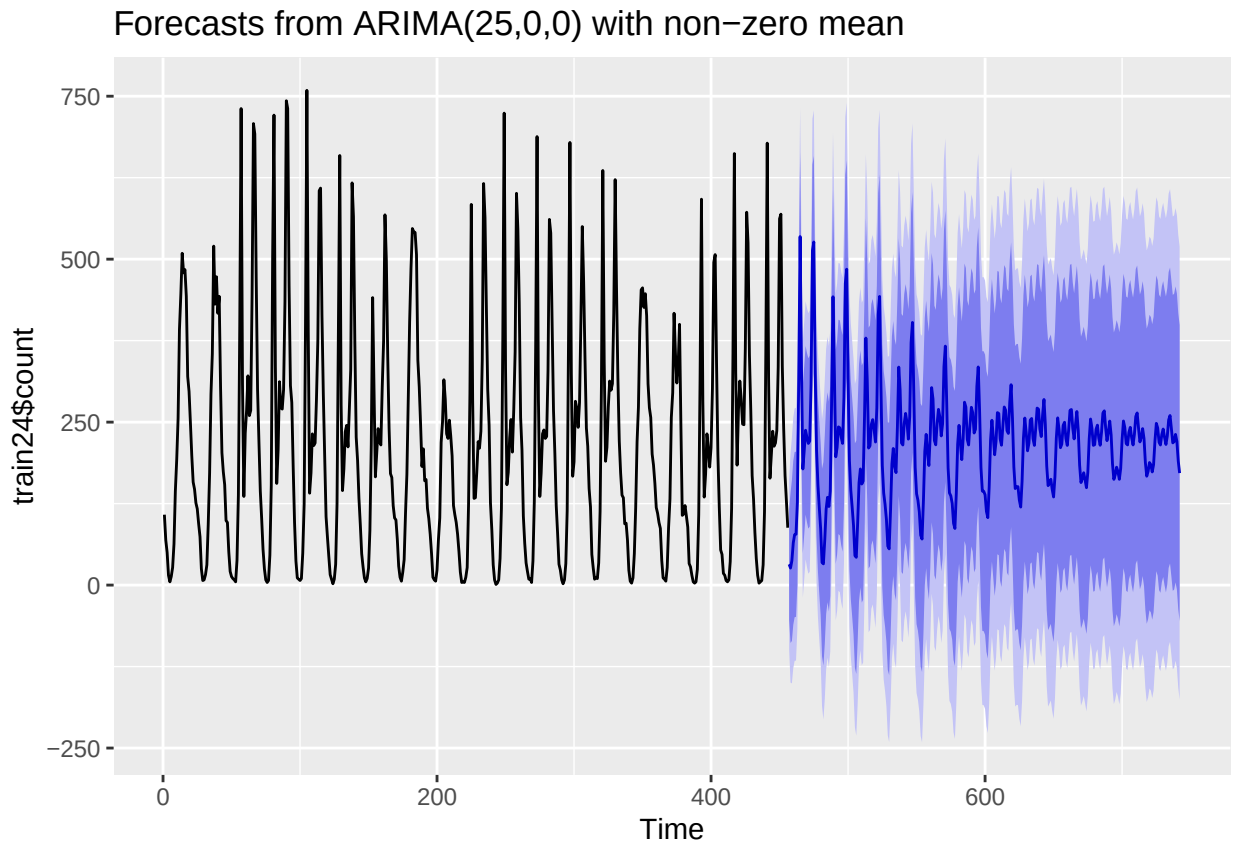



```
checkresiduals(AR24)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0) with non-zero mean
## Q* = 56.929, df = 3, p-value = 2.661e-12
##
## Model df: 26.   Total lags used: 29
```

```
fcst <- forecast(AR24, h=number)

autoplot(fcst)
```




```
# point estimate (mean)
test24$count <- round(fcst$mean)

# test24

RMSLE(y_pred = floor(ifelse(fcst$fitted < 0, 0, round(fcst$fitted))), y_true = train24$count)
```

```
## [1] 0.6427754
```

Combine all of the individual data frames

```
combined <- data.frame(datetime=character(),
                 count=double(),
                 stringsAsFactors=FALSE)



combined <- bind_rows(test1, test2, test3, test4, test5, test6, test7, test8, test9, test10, test11, te
                      test13, test14, test15, test16, test17, test18, test19, test20, test21, test22, t

combined <- combined %>%
  mutate(count = floor(ifelse(count < 0, 0, count)))

# write.csv(combined, file = "~/Desktop/ts_kaggle_submission.csv", row.names = F)
```

RMSLE: Root Mean Squared Logarithmic Error Loss

```
# RMSLE(y_pred = floor(ifelse(fcst$fitted < 0, 0, round(fcst$fitted))), y_true = train2$count)
```

Submit

```
# Kaggle Score:  RMSLE = 1.33332
score = (1 - (3008 / 3251)) * 100

# We only beat ~7% of all submissions
score
```

```
## [1] 7.474623
```