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

```r
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
```

```r
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
```

```r
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
```

```r
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)
```

```
##
##   Great    Good Average    Poor
##    7192    2834     859       1
```

Split Date-Time (Both)

- Year, Month, Day and Hour

```r
# library(lubridate)

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

Modeling

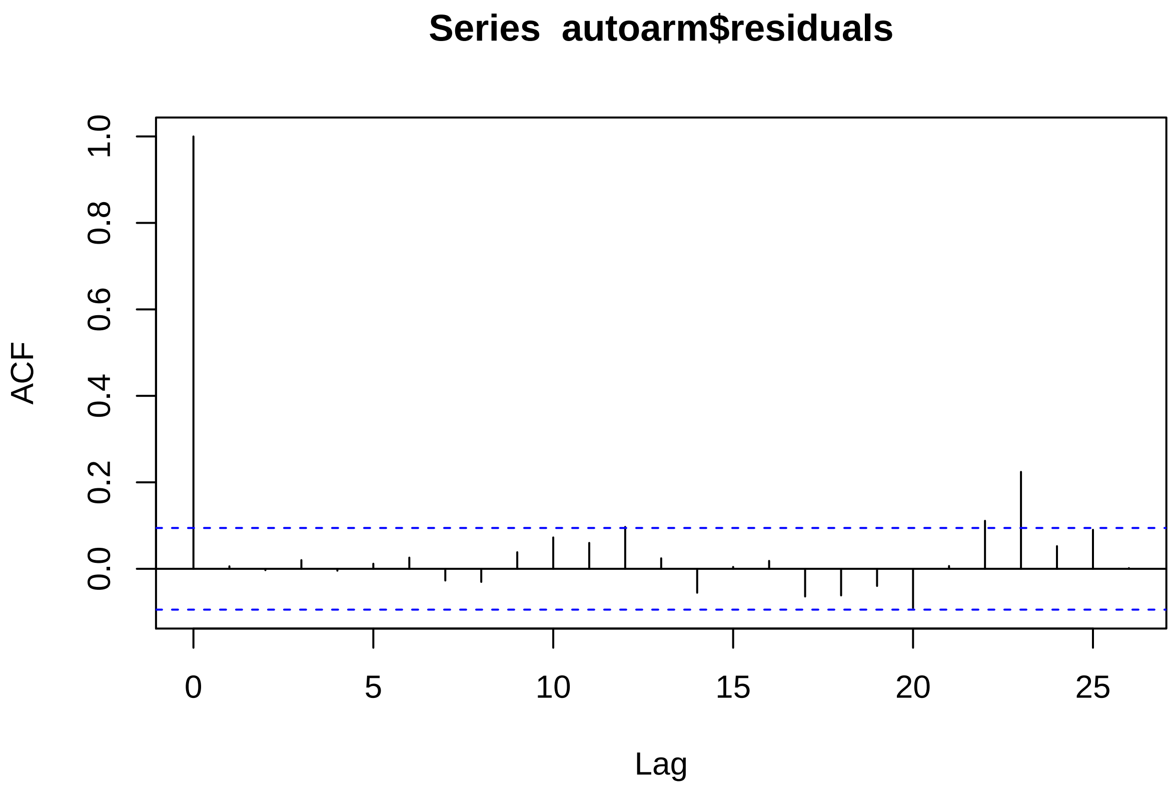
```
- psuedo code

- 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

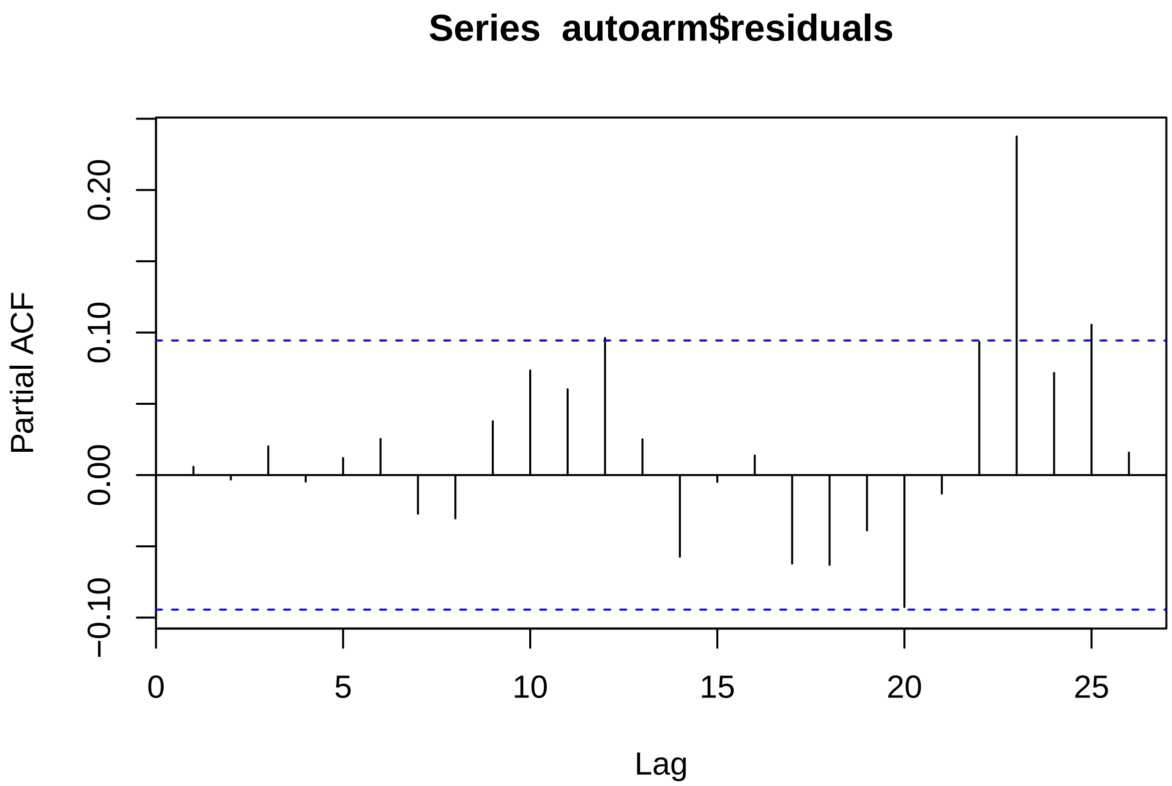
January

Auto Arima

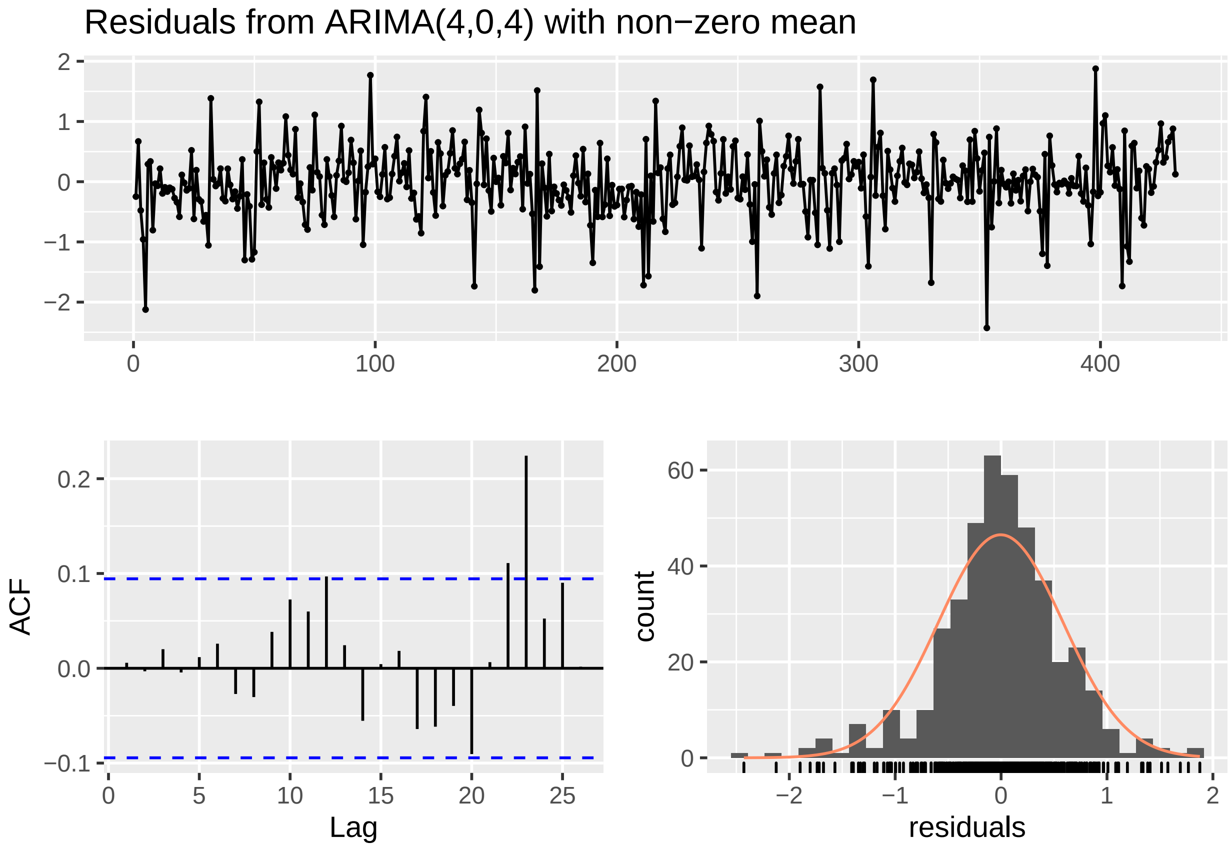
```
train1arm <- train %>%
  filter(year == '2011' & month == 'January') %>%
  select(datetime, count)

test1arm <- test %>%
  filter(year == '2011' & month == 'January') %>%
  mutate(count = NA) %>%
  select(datetime, count)


### Log the response variable
train1arm$count = log(train1arm$count)

# head(train25)
# head(test25)

autoarm <- auto.arima(train1arm$count, D=1)

# ?auto.arima

number = nrow(test1arm)

acf(autoarm$residuals)
```



```
pacf(autoarm$residuals)
```

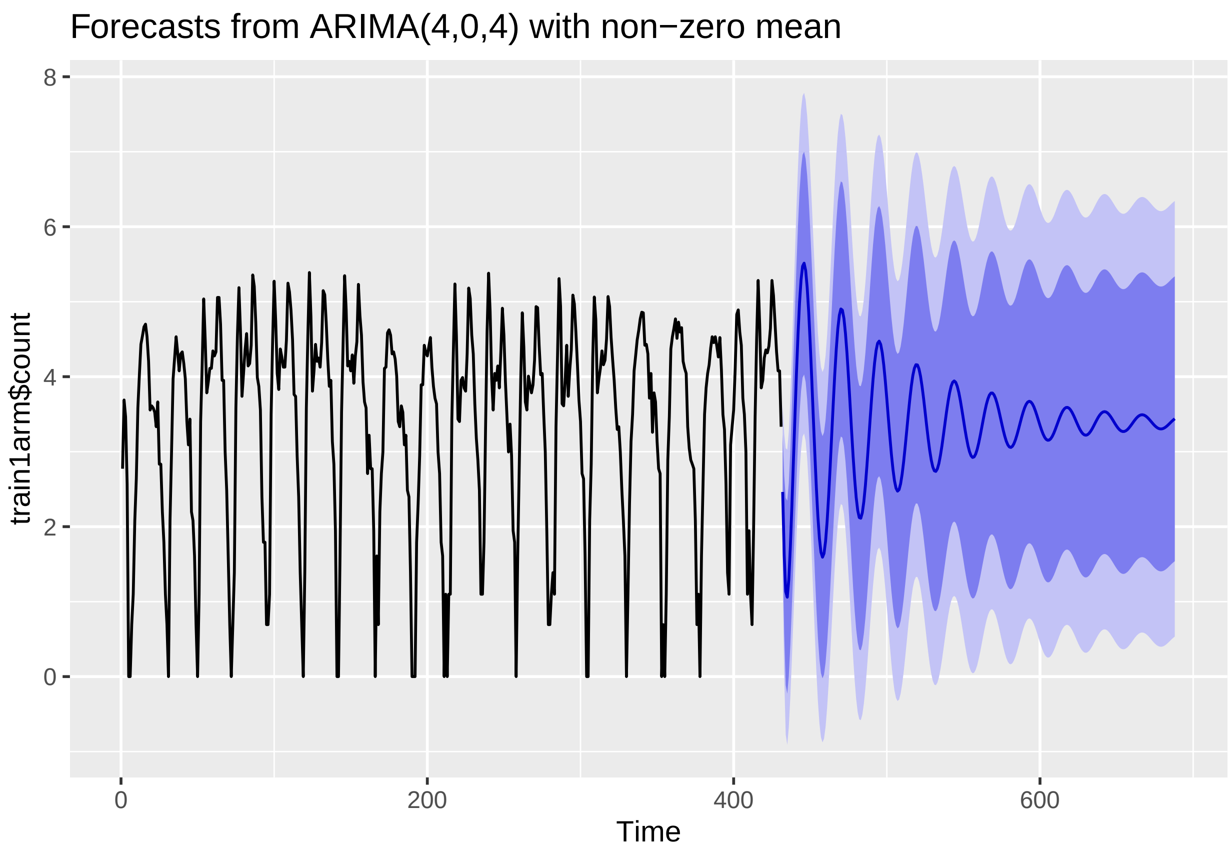


```
checkresiduals(autoarm)
```



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(4,0,4) with non-zero mean
## Q* = 10.049, df = 3, p-value = 0.01816
##
## Model df: 9.   Total lags used: 12
```

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

autoplot(fcst)
```



```r
# point estimate (mean)
test1arm$count <- fcst$mean

RMSLE(y_pred = fcst$fitted, y_true = train1arm$count)
```

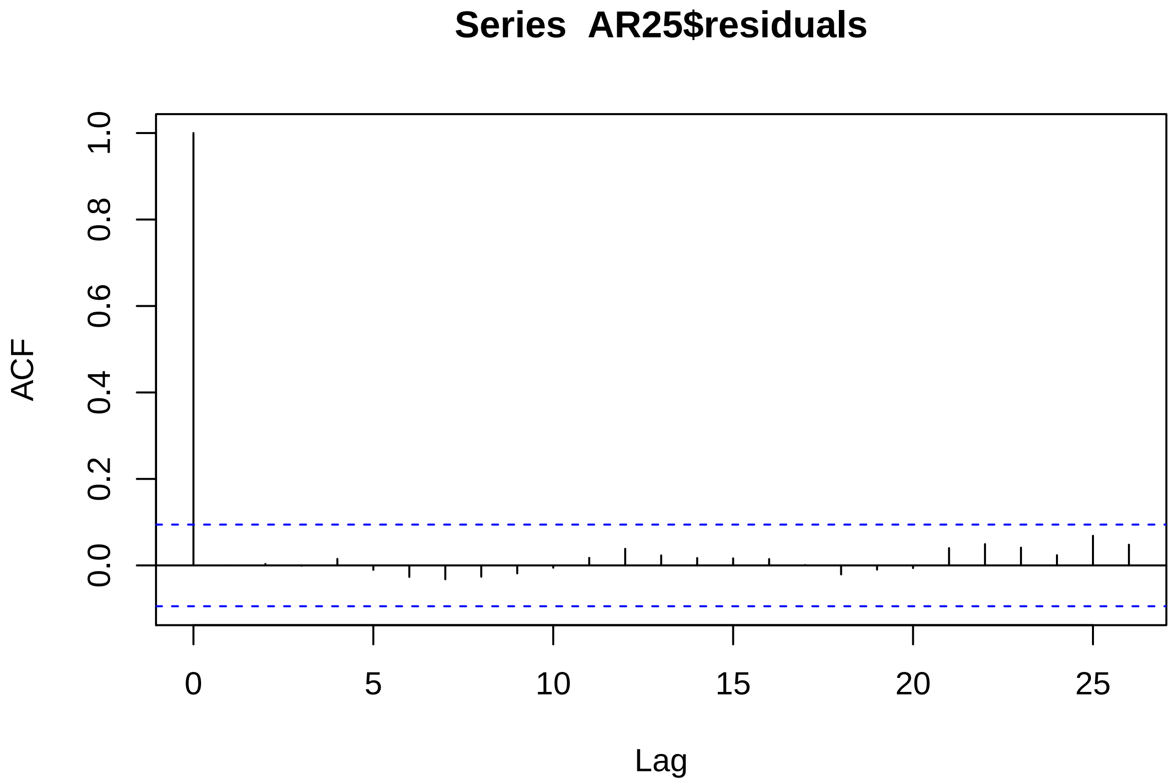
```
## [1] 0.2418941
```

```r
summary(autoarm)
```

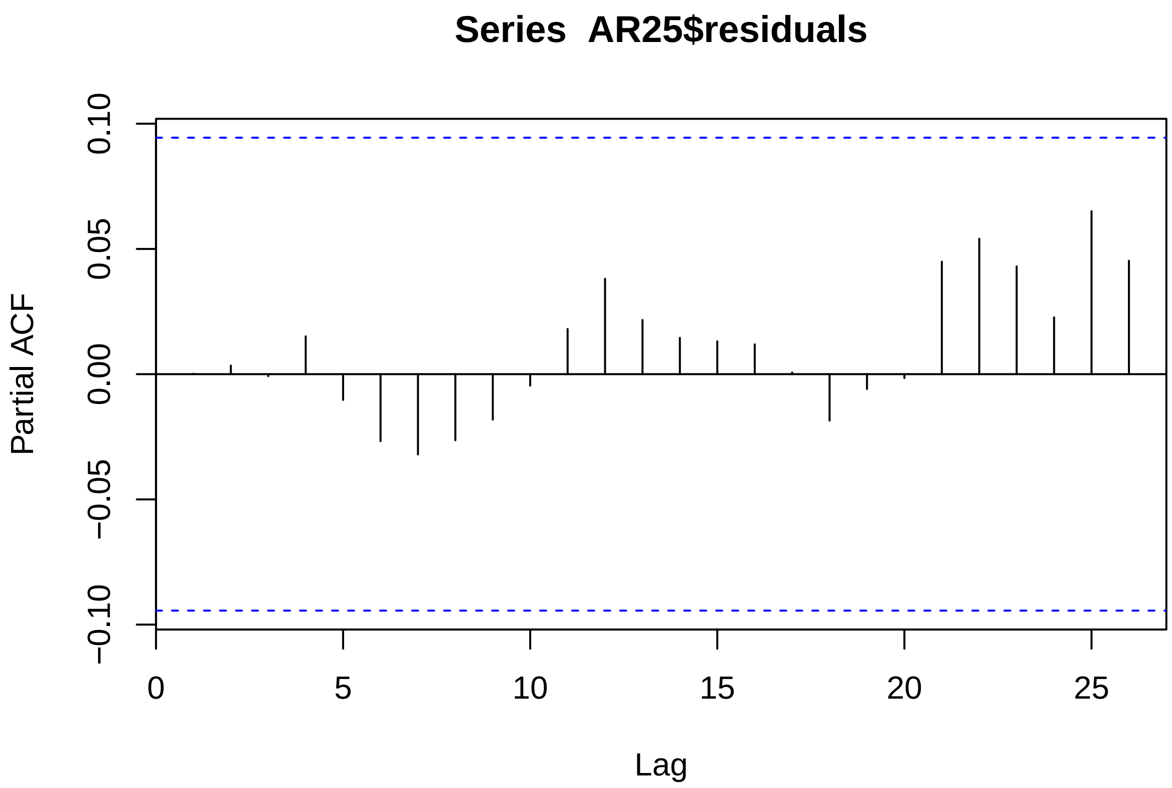
```
## Series: train1arm$count
## ARIMA(4,0,4) with non-zero mean
##
## Coefficients:
##          ar1      ar2     ar3      ar4      ma1     ma2      ma3     ma4
##       2.7868  -2.9306  1.3916  -0.2738  -1.8234  1.1141  -0.5814  0.3398
## s.e.  0.1292   0.3497  0.3333   0.1123   0.1228  0.2204   0.1350  0.0625
##         mean
##       3.3896
## s.e.  0.0535
##
## sigma^2 estimated as 0.3541:  log likelihood=-385.78
## AIC=791.57   AICc=792.09   BIC=832.23
##
## Training set error measures:
##                        ME      RMSE       MAE MPE MAPE      MASE
## Training set -0.00352749 0.5888569 0.4286433 NaN  Inf 0.7797332
##                       ACF1
## Training set 0.005790901
```

AR 25

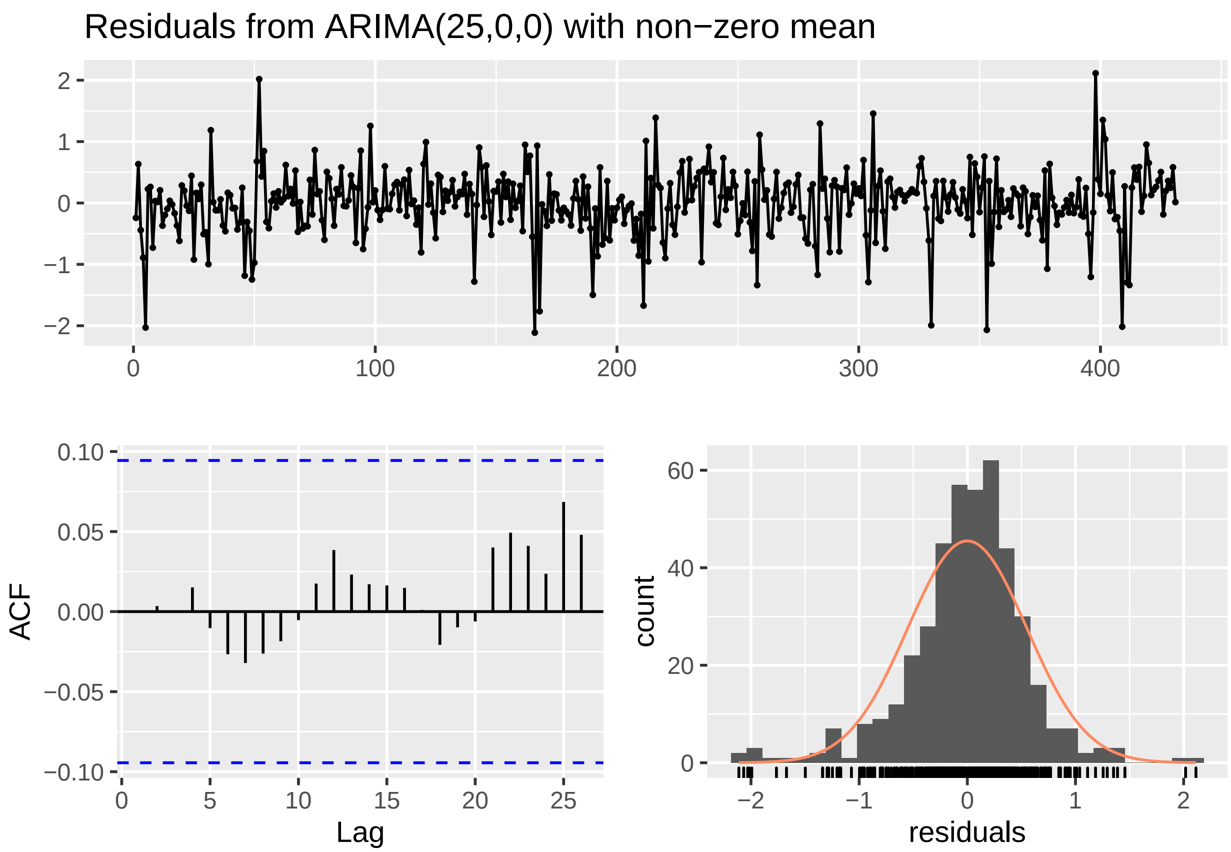
```
train1 <- train %>%
  filter(year == '2011' & month == 'January') %>%
  select(datetime, count)

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


### Log the response variable
train1$count = log(train1$count)

# head(train1)
# head(test1)

AR25 <- arima(train1$count,order=c(25,0,0))

number = nrow(test1)

acf(AR25$residuals)
```



```
pacf(AR25$residuals)
```

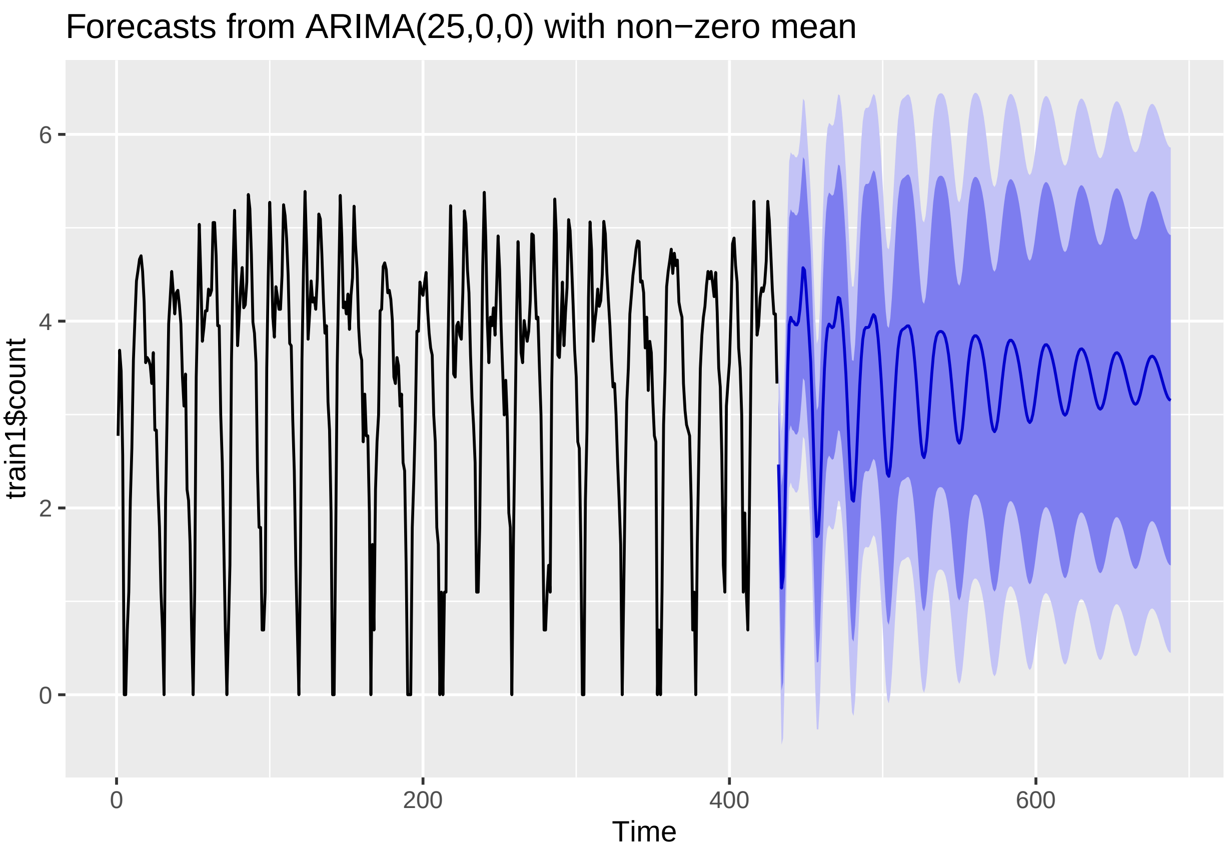


```
checkresiduals(AR25)
```



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0) with non-zero mean
## Q* = 11.131, df = 3, p-value = 0.01104
##
## Model df: 26.   Total lags used: 29
```

```
fcst <- forecast(AR25, h=number)

autoplot(fcst)
```



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


RMSLE(y_pred = fcst$fitted, y_true = train1$count)
```

```
## [1] 0.2265693
```

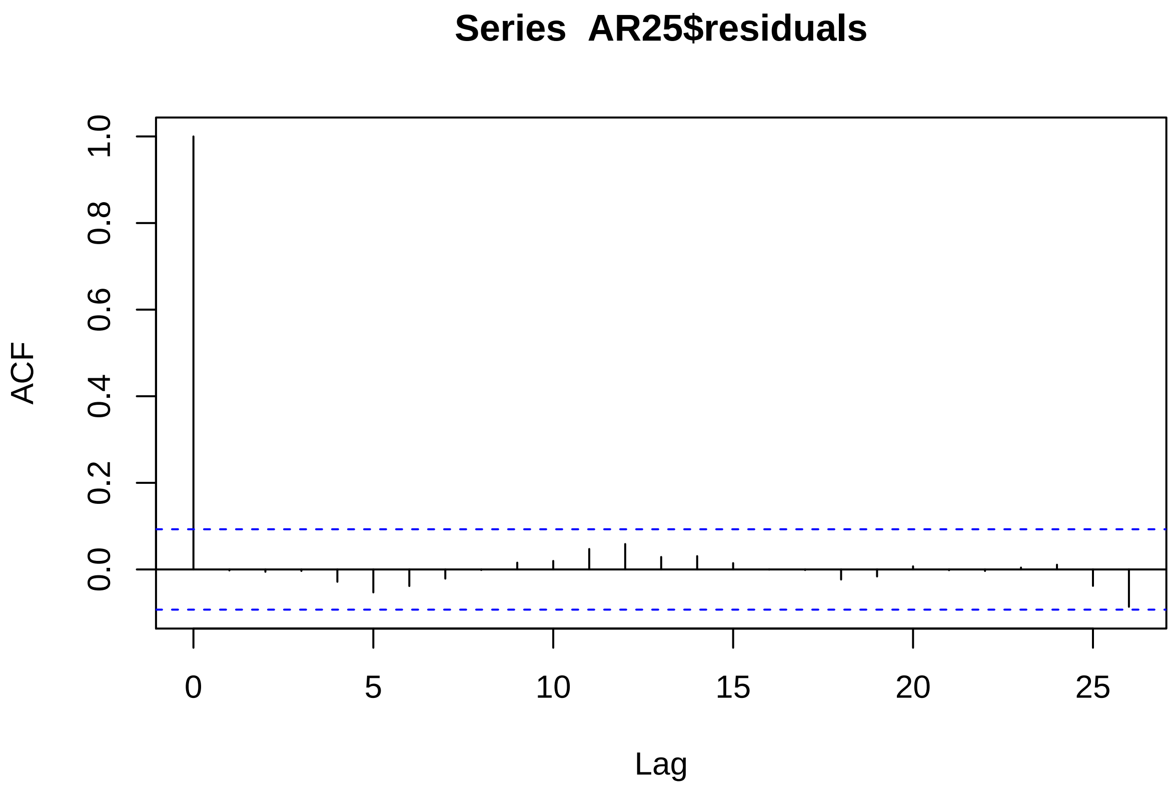
```r
summary(AR25)
```

```
##
## Call:
## arima(x = train1$count, order = c(25, 0, 0))
##
## Coefficients:
##          ar1      ar2      ar3      ar4     ar5      ar6      ar7     ar8
##       0.8974  -0.0200  -0.3072  -0.0068  0.1252  -0.0417  -0.1243  0.0083
## s.e.  0.0483   0.0644   0.0639   0.0656  0.0663   0.0663   0.0665  0.0667
##          ar9     ar10     ar11    ar12     ar13     ar14    ar15   ar16
##       0.0568  -0.0293  -0.0557  0.0395  -0.0611  -0.0674  0.0715  0.018
## s.e.  0.0663   0.0664   0.0661  0.0663   0.0659   0.0659  0.0659  0.066
##          ar17    ar18    ar19     ar20    ar21    ar22    ar23     ar24
##       -0.1017  0.0035  0.0501  -0.0692  0.0555  0.1176  0.1545  -0.1514
## s.e.   0.0659  0.0661  0.0660   0.0661  0.0659  0.0664  0.0656   0.0664
##         ar25  intercept
##       0.0543     3.3852
## s.e.  0.0499     0.0683
```

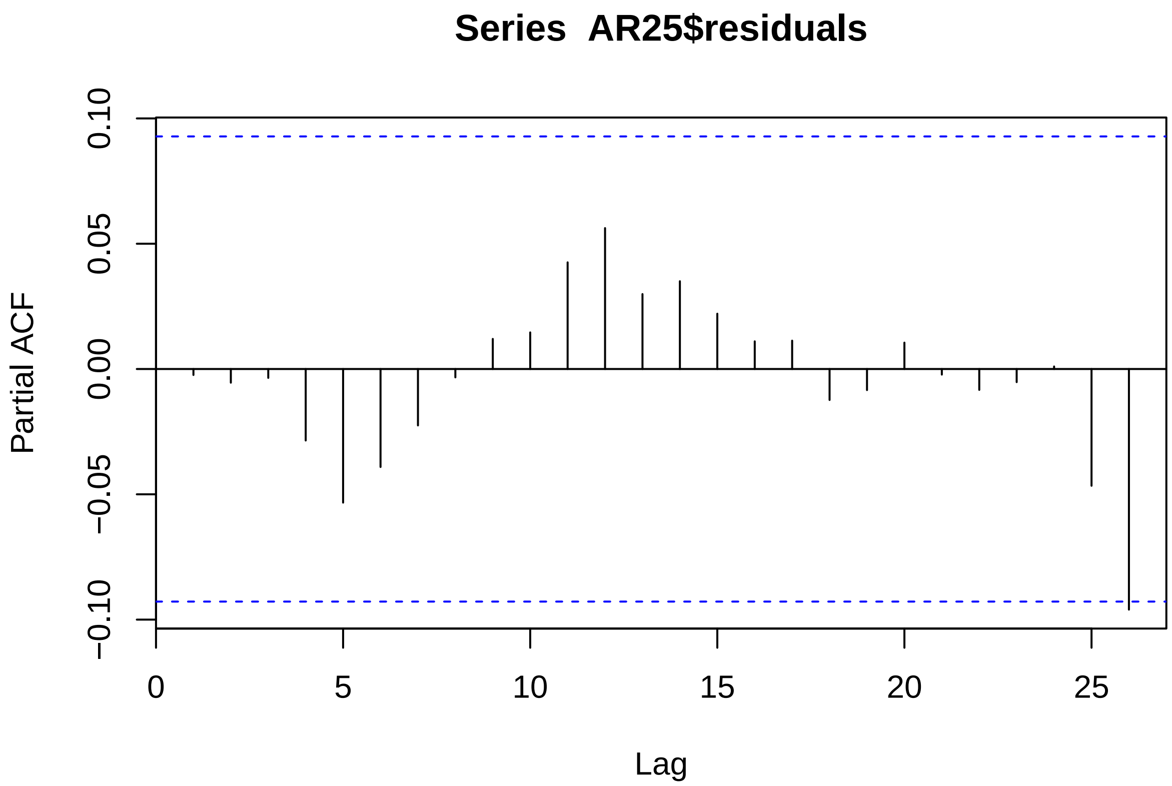
```
##
## sigma^2 estimated as 0.3026:  log likelihood = -357.19,  aic = 768.38
##
## Training set error measures:
##                         ME      RMSE       MAE MPE MAPE      MASE
## Training set -0.0002183009 0.5500744 0.3973957 NaN  Inf 0.7228916
##                      ACF1
## Training set 0.0001902056
```

February

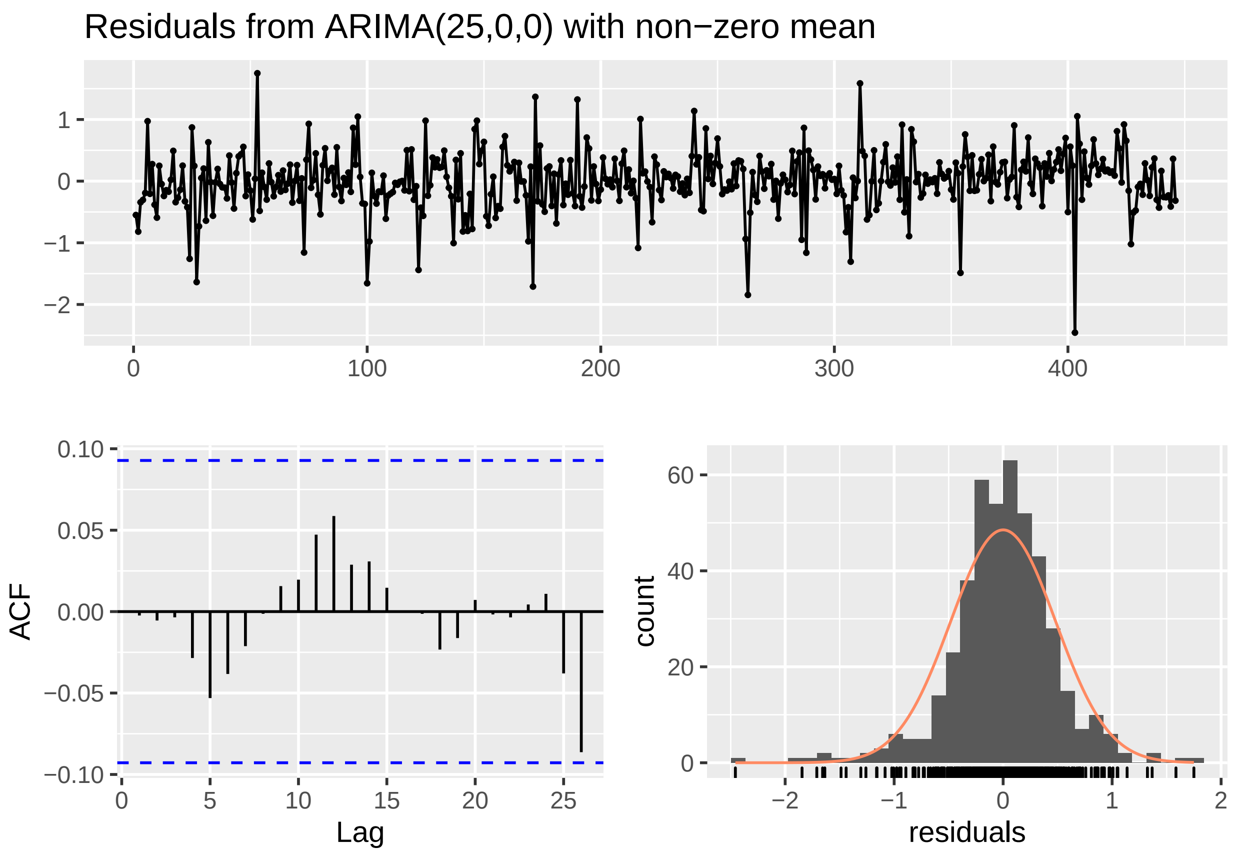
```
train2 <- train %>%
  filter(year == '2011' & month == 'February') %>%
  select(datetime, count)

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

### Log the response variable
train2$count = log(train2$count)

# head(train2)
# head(test2)

AR25 <- arima(train2$count,order=c(25,0,0))

number = nrow(test2)

acf(AR25$residuals)
```



```r
pacf(AR25$residuals)
```

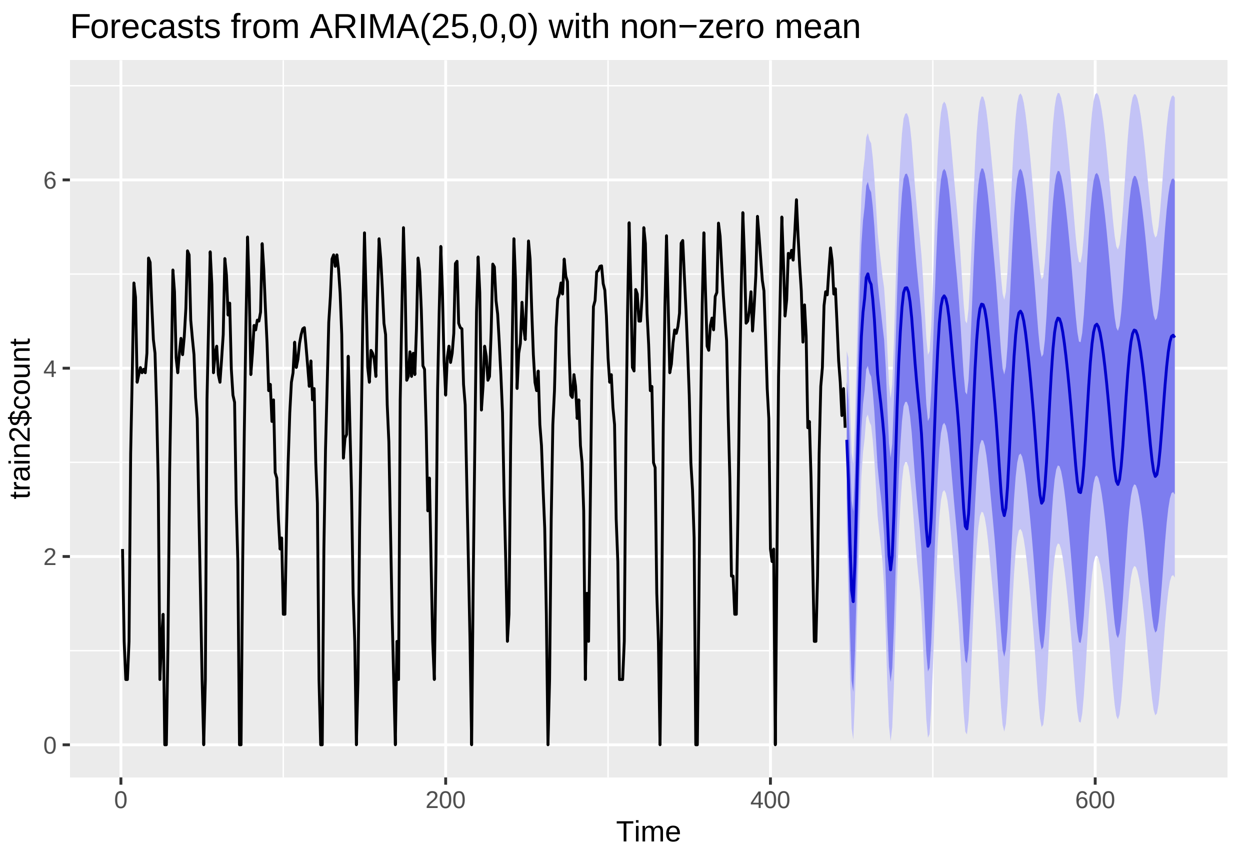


```
checkresiduals(AR25)
```



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0) with non-zero mean
## Q* = 12.022, df = 3, p-value = 0.007309
##
## Model df: 26.   Total lags used: 29
```

```
fcst <- forecast(AR25, h=number)

autoplot(fcst)
```



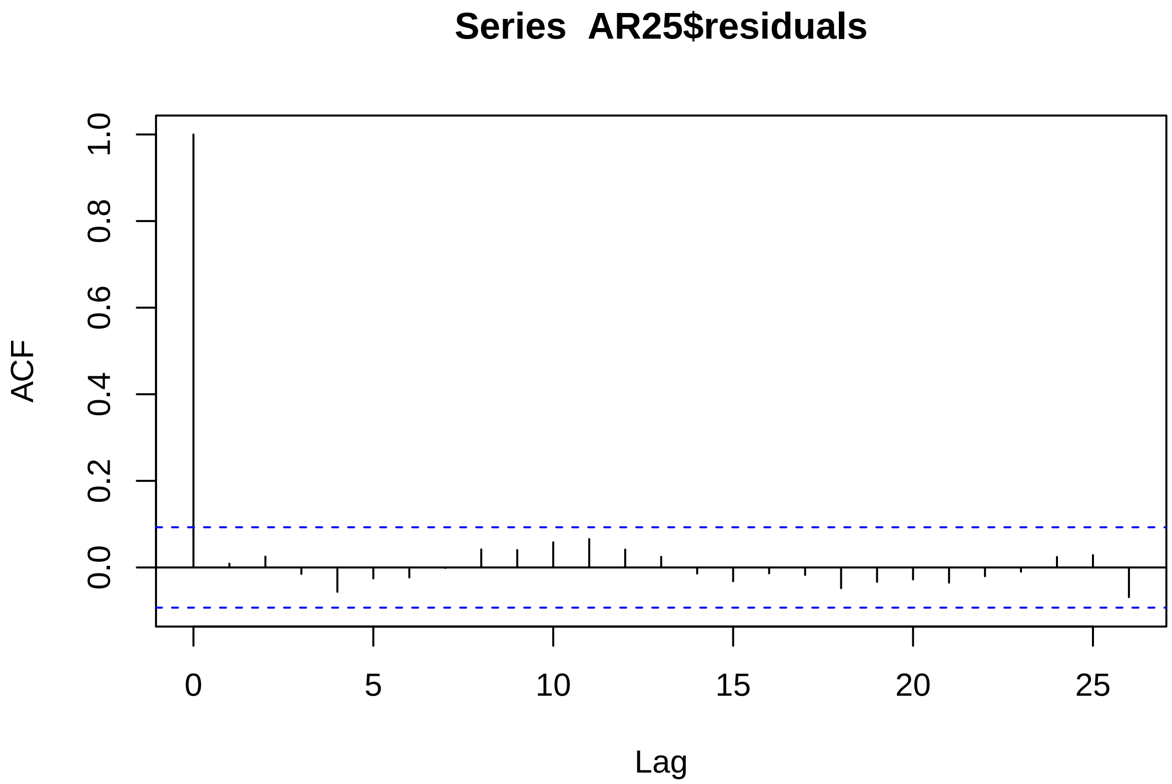
```
# point estimate (mean)
test2$count <- fcst$mean


RMSLE(y_pred = fcst$fitted, y_true = train2$count)
```

```
## [1] 0.190118
```

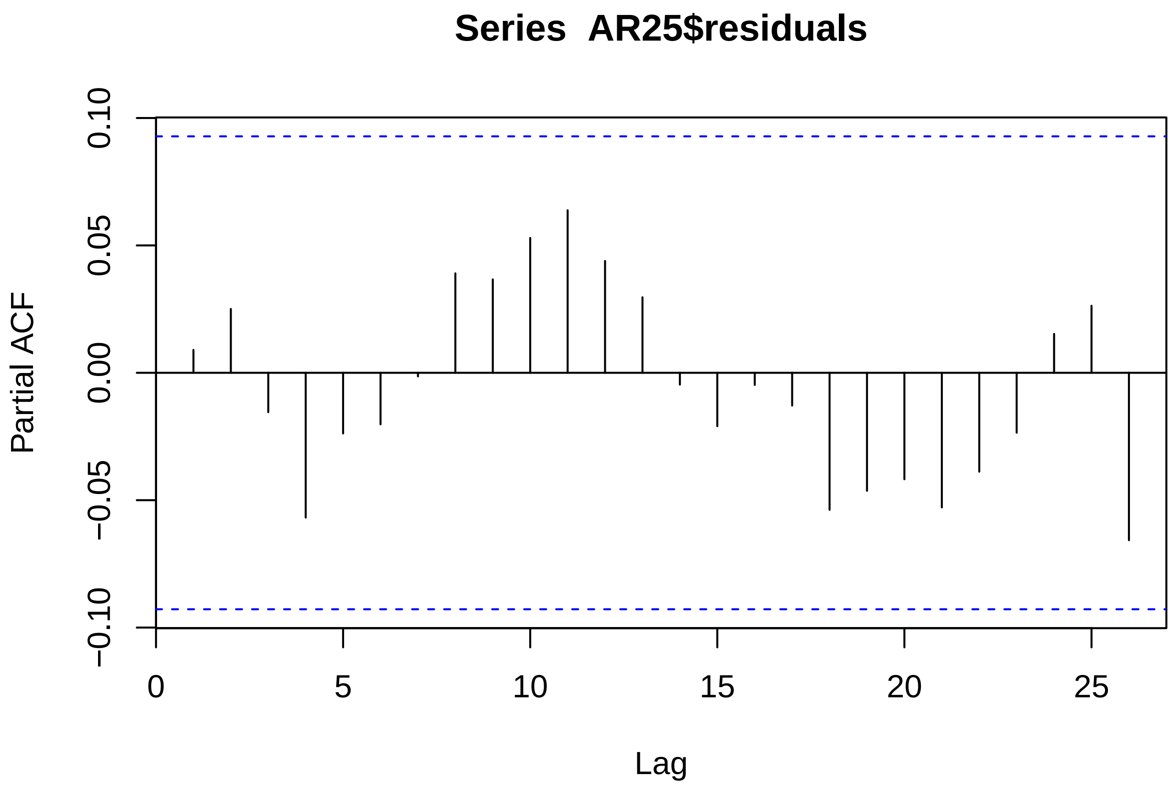
March

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

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

### Log the response variable
train3$count = log(train3$count)

# head(train3)
# head(test3)

AR25 <- arima(train3$count,order=c(25,0,0))
```

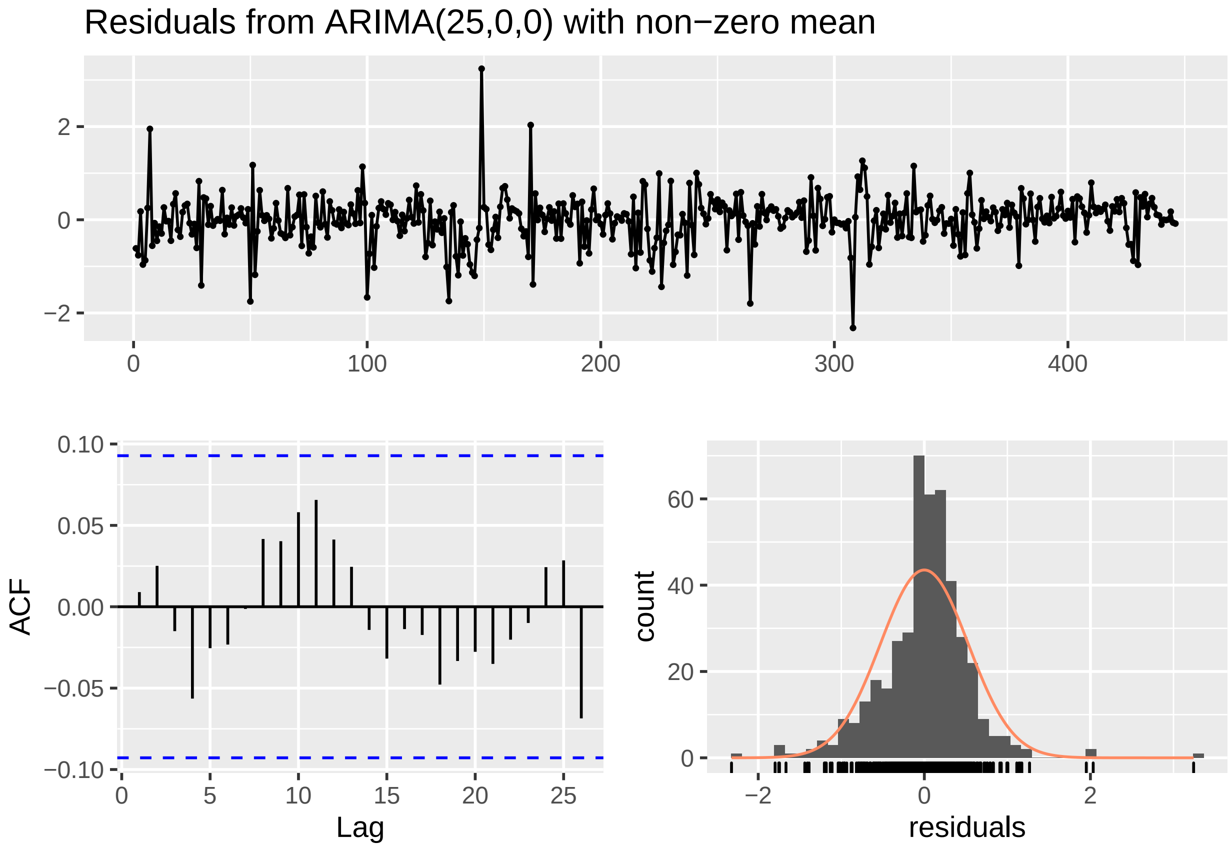
```
number = nrow(test3)

acf(AR25$residuals)
```



```
pacf(AR25$residuals)
```

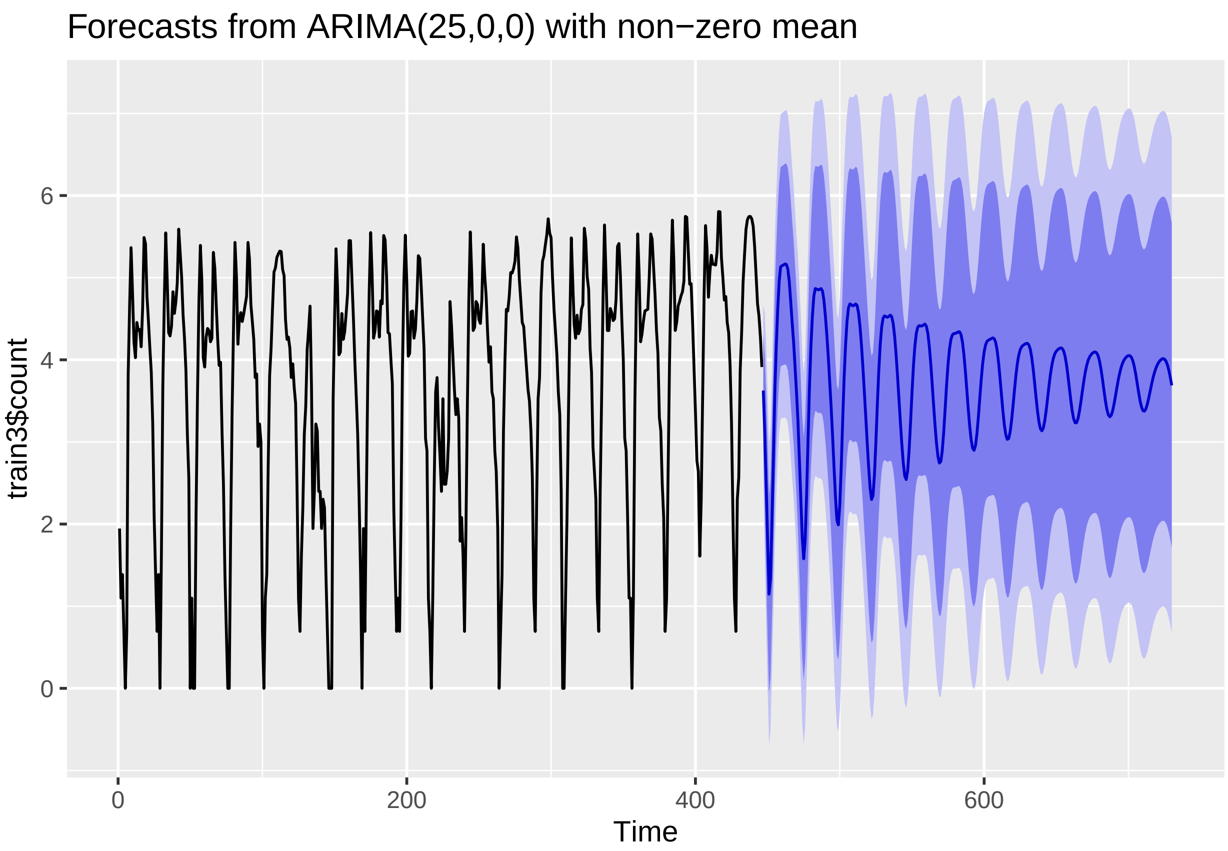


```r
checkresiduals(AR25)
```



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0) with non-zero mean
## Q* = 18.587, df = 3, p-value = 0.0003327
##
## Model df: 26.   Total lags used: 29
```

```
fcst <- forecast(AR25, h=number)

autoplot(fcst)
```



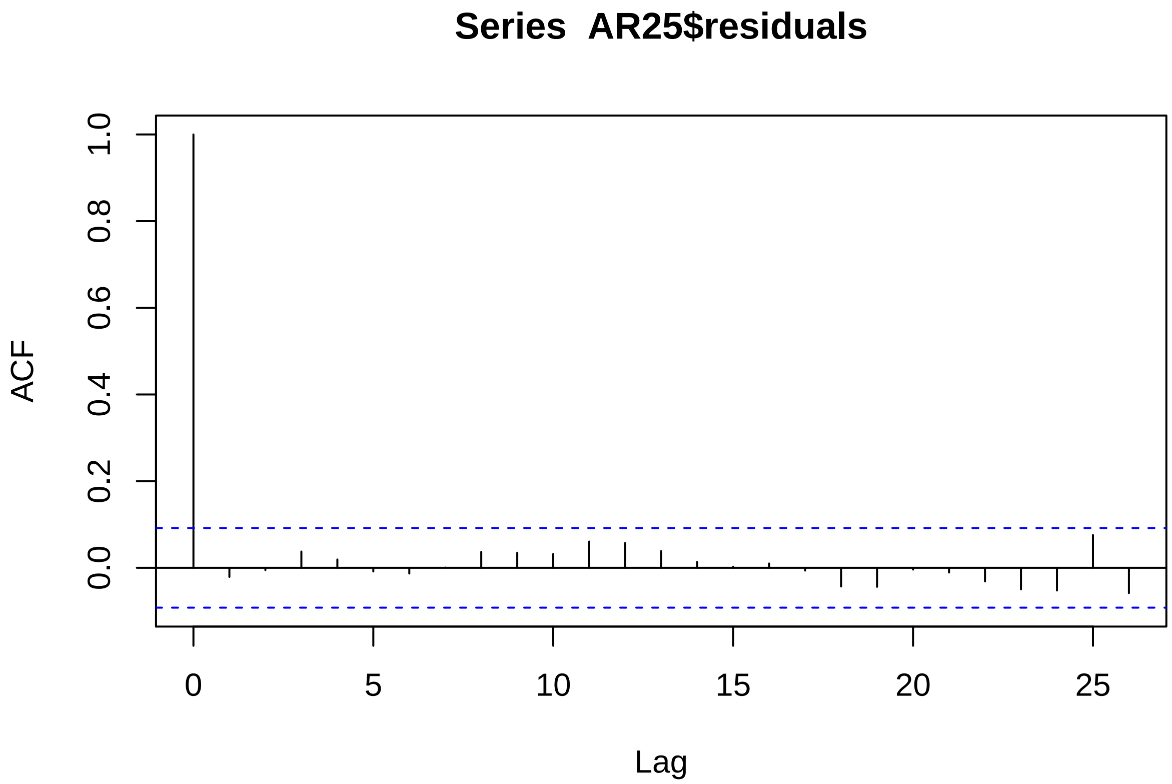
```
# point estimate (mean)
test3$count <- fcst$mean


RMSLE(y_pred = fcst$fitted, y_true = train3$count)
```

```
## [1] 0.2105017
```

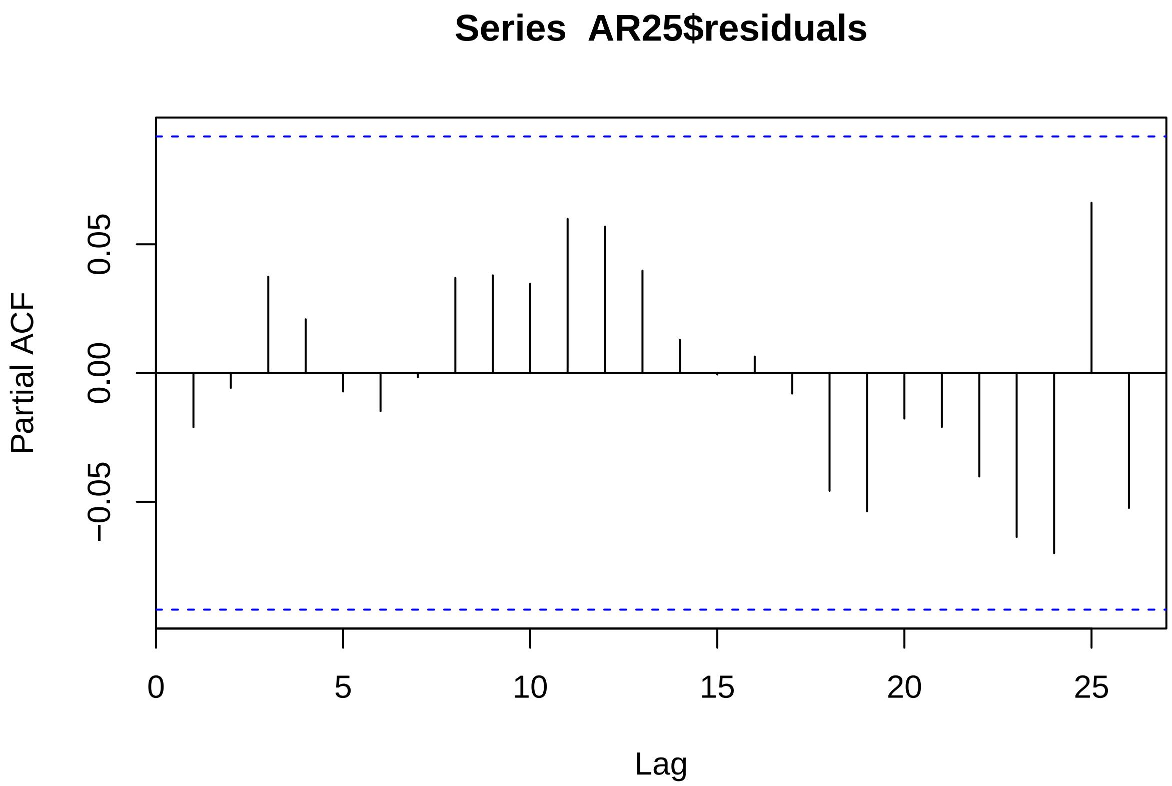
April

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

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

### Log the response variable
train4$count = log(train4$count)

# head(train4)
# head(test4)

AR25 <- arima(train4$count,order=c(25,0,0))
```

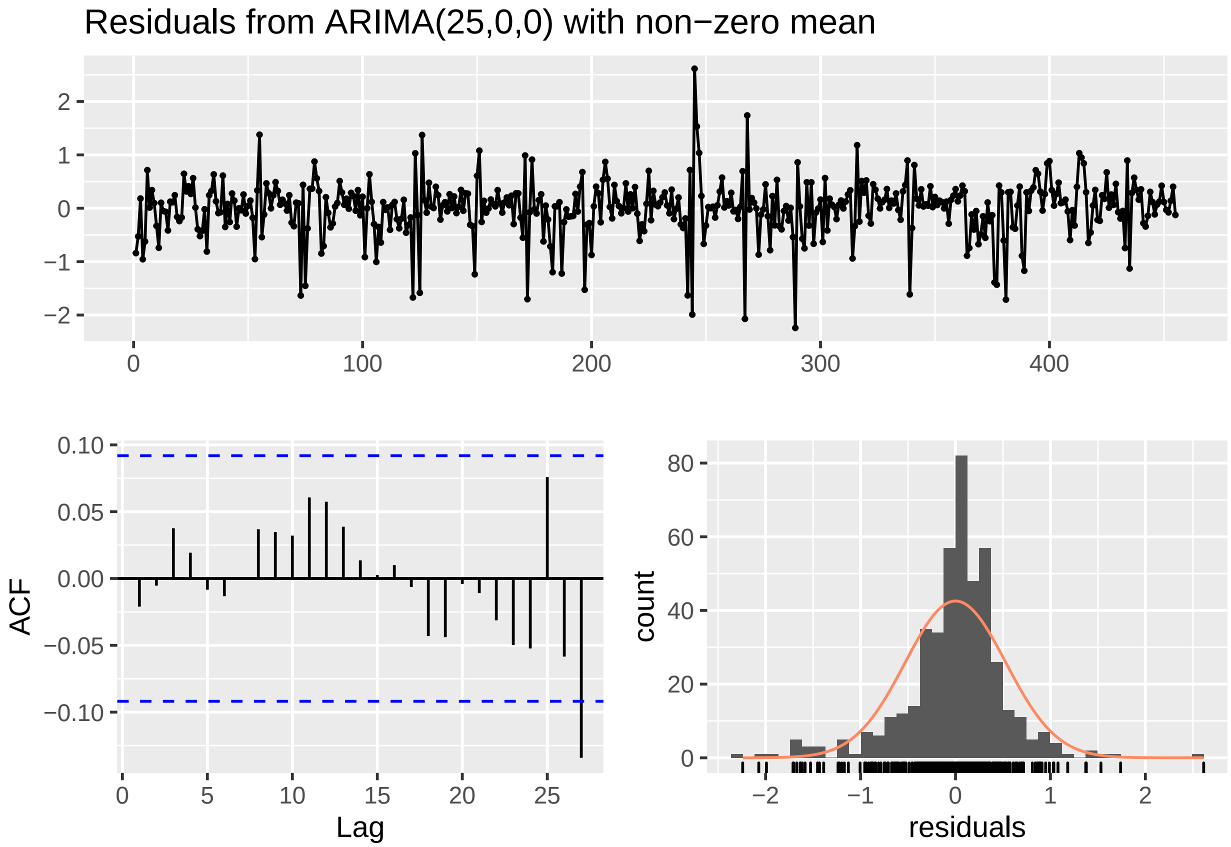
```
number = nrow(test4)

acf(AR25$residuals)
```



```
pacf(AR25$residuals)
```

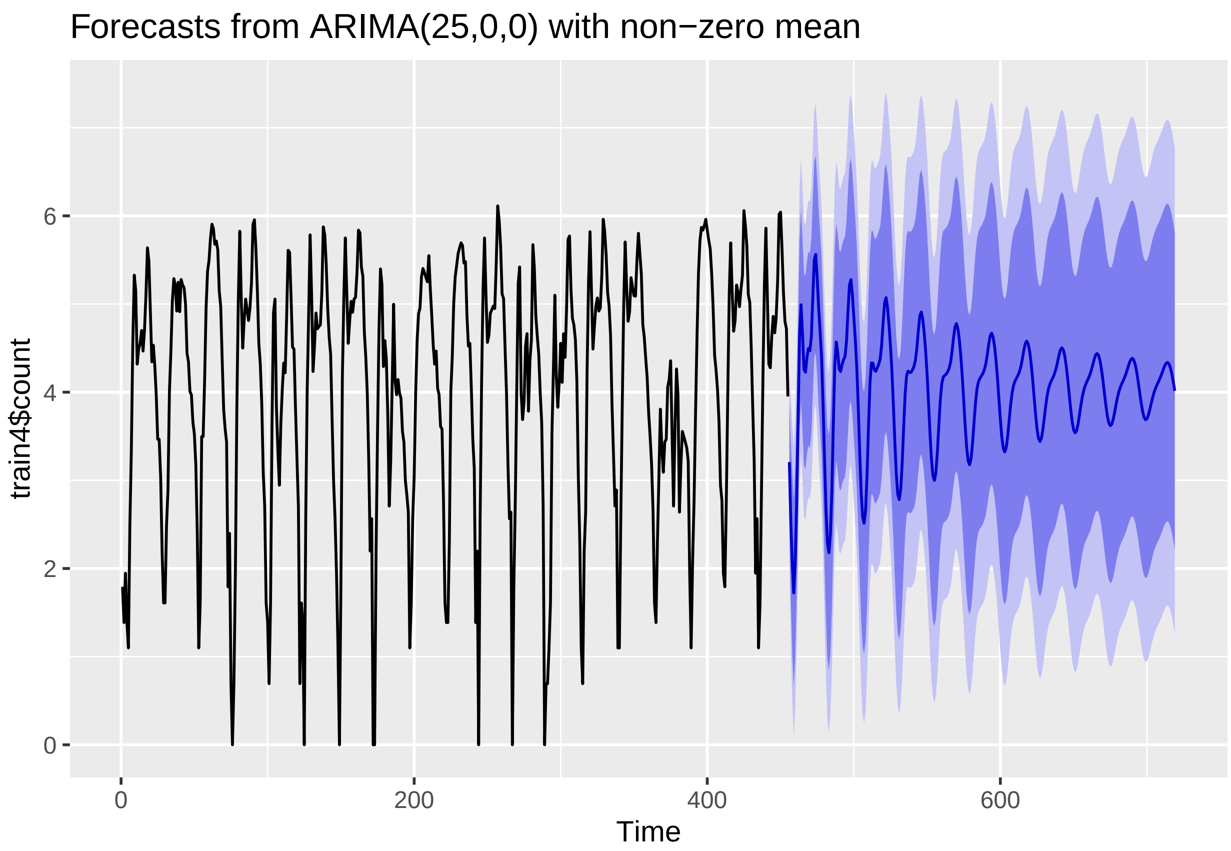


```r
checkresiduals(AR25)
```



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0) with non-zero mean
## Q* = 26.129, df = 3, p-value = 8.961e-06
##
## Model df: 26.   Total lags used: 29
```

```
fcst <- forecast(AR25, h=number)

autoplot(fcst)
```



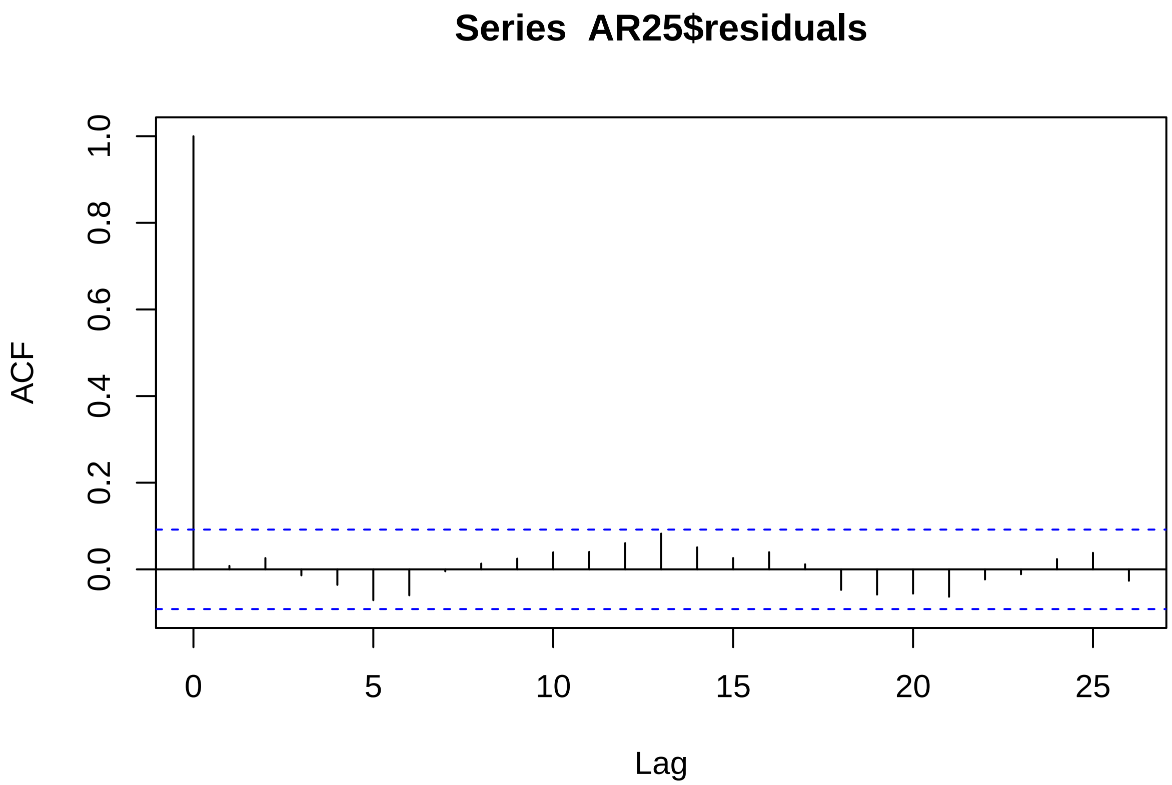
```
# point estimate (mean)
test4$count <- fcst$mean


RMSLE(y_pred = fcst$fitted, y_true = train4$count)
```

```
## [1] 0.1956949
```

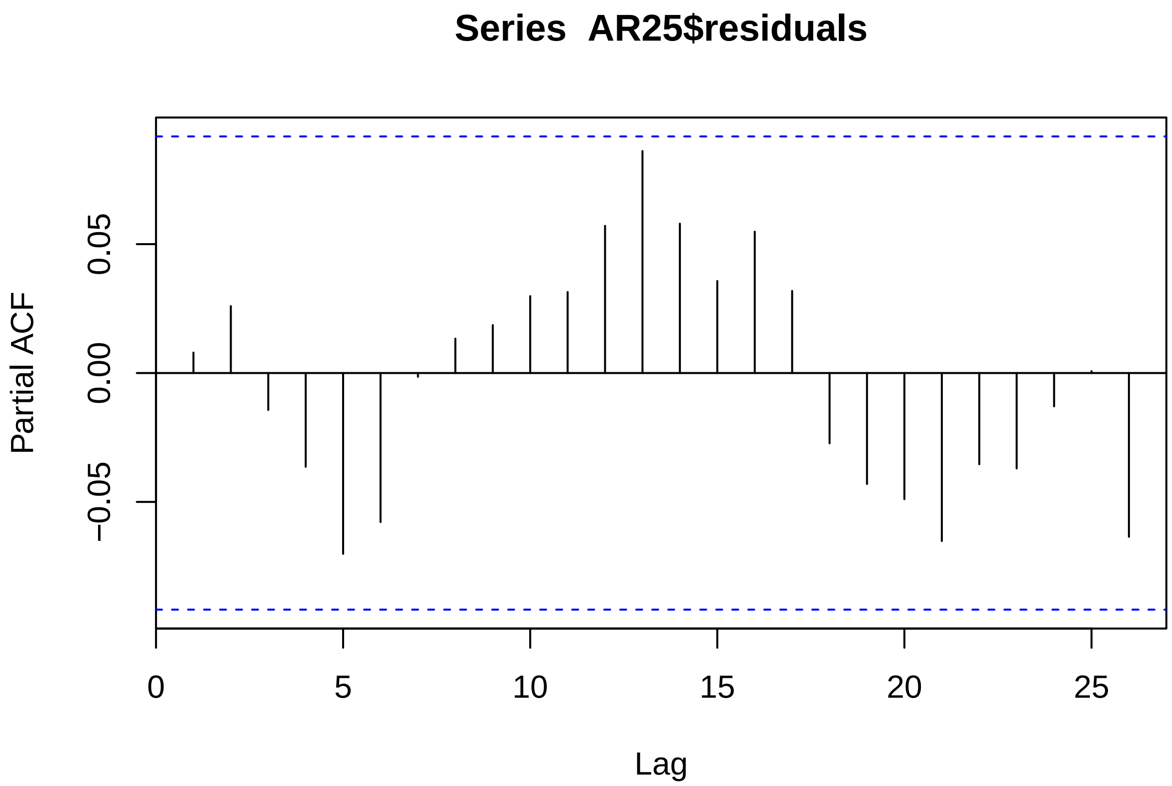
May

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

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

### Log the response variable
train5$count = log(train5$count)

# head(train5)
# head(test5)

AR25 <- arima(train5$count,order=c(25,0,0))
```

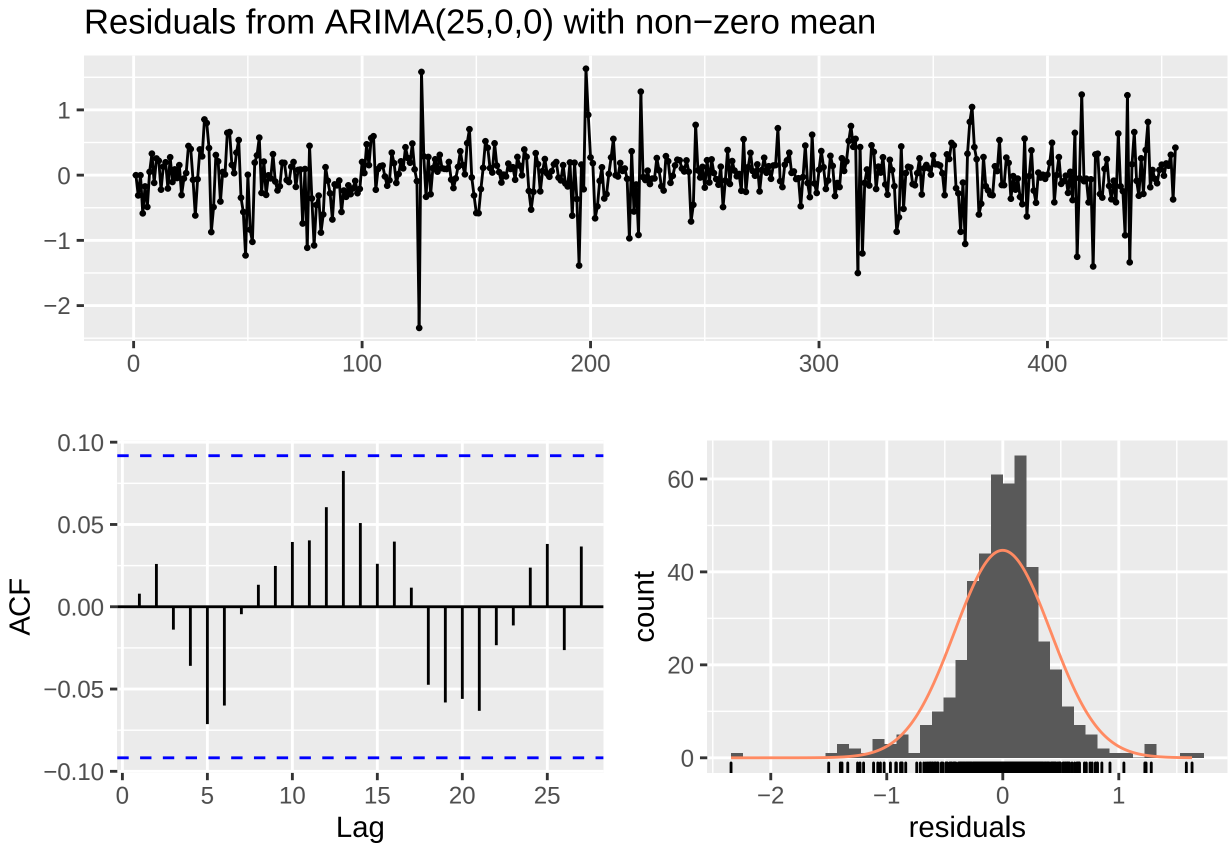
```
# tsdisplay(residuals(AR25),lag.max=25,main="AR(24) Resid. Diagnostics")

number = nrow(test5)

acf(AR25$residuals)
```



```
pacf(AR25$residuals)
```

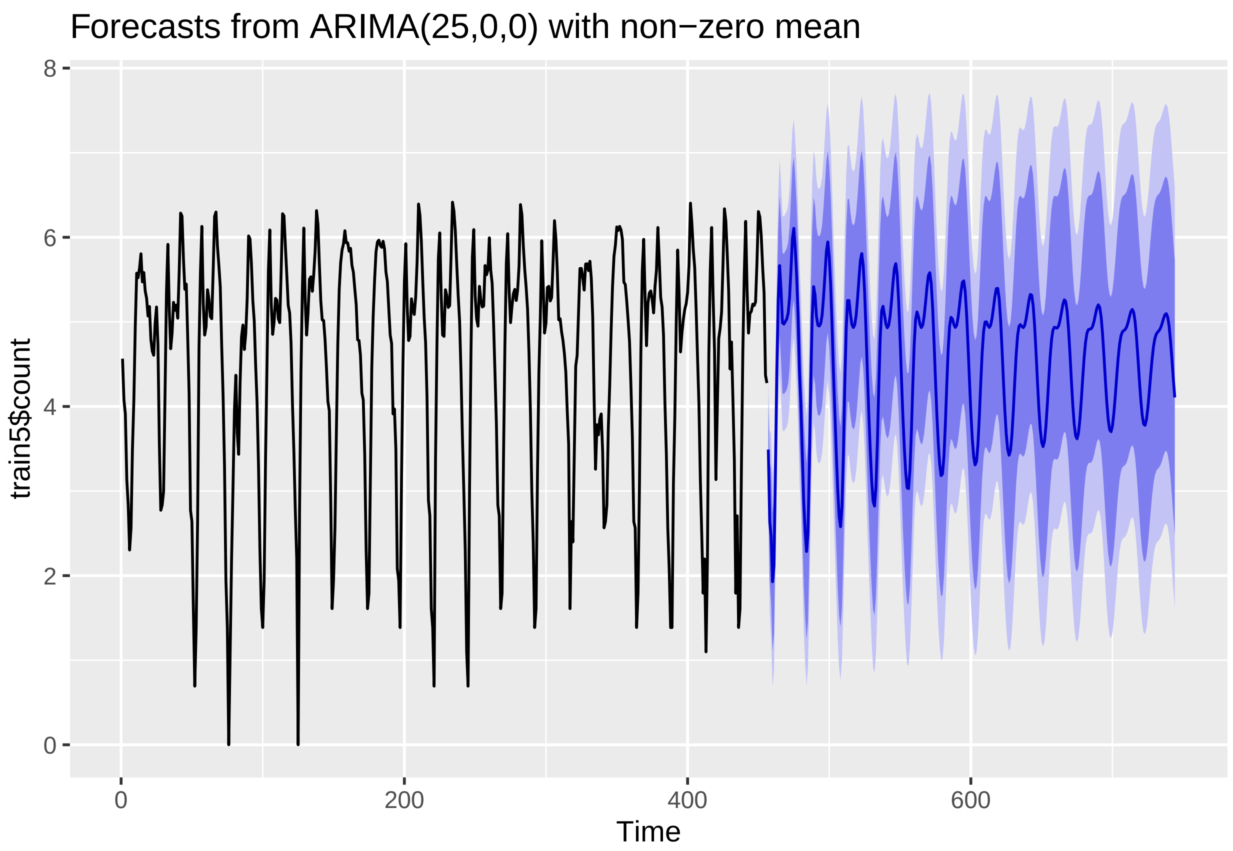


```
checkresiduals(AR25)
```



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0) with non-zero mean
## Q* = 24.97, df = 3, p-value = 1.566e-05
##
## Model df: 26.   Total lags used: 29
```

```
fcst <- forecast(AR25, h=number)

autoplot(fcst)
```



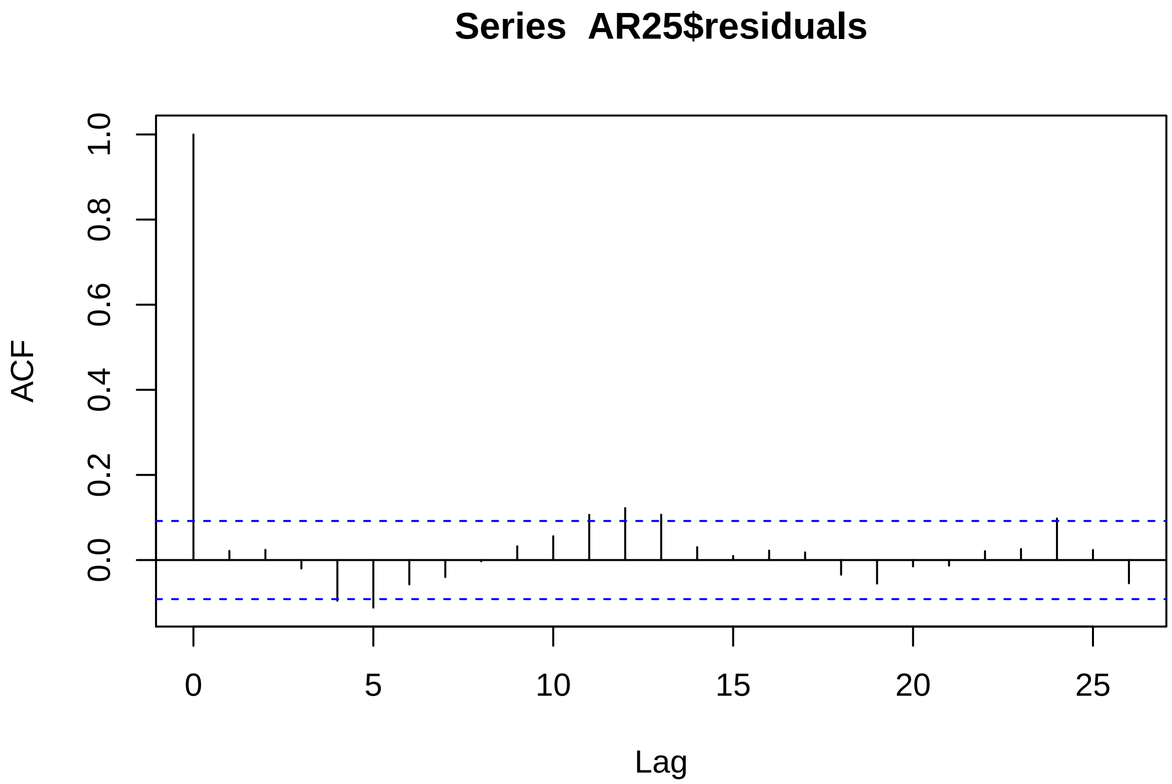
```
# point estimate (mean)
test5$count <- fcst$mean

RMSLE(y_pred = fcst$fitted, y_true = train5$count)
```

```
## [1] 0.1216449
```

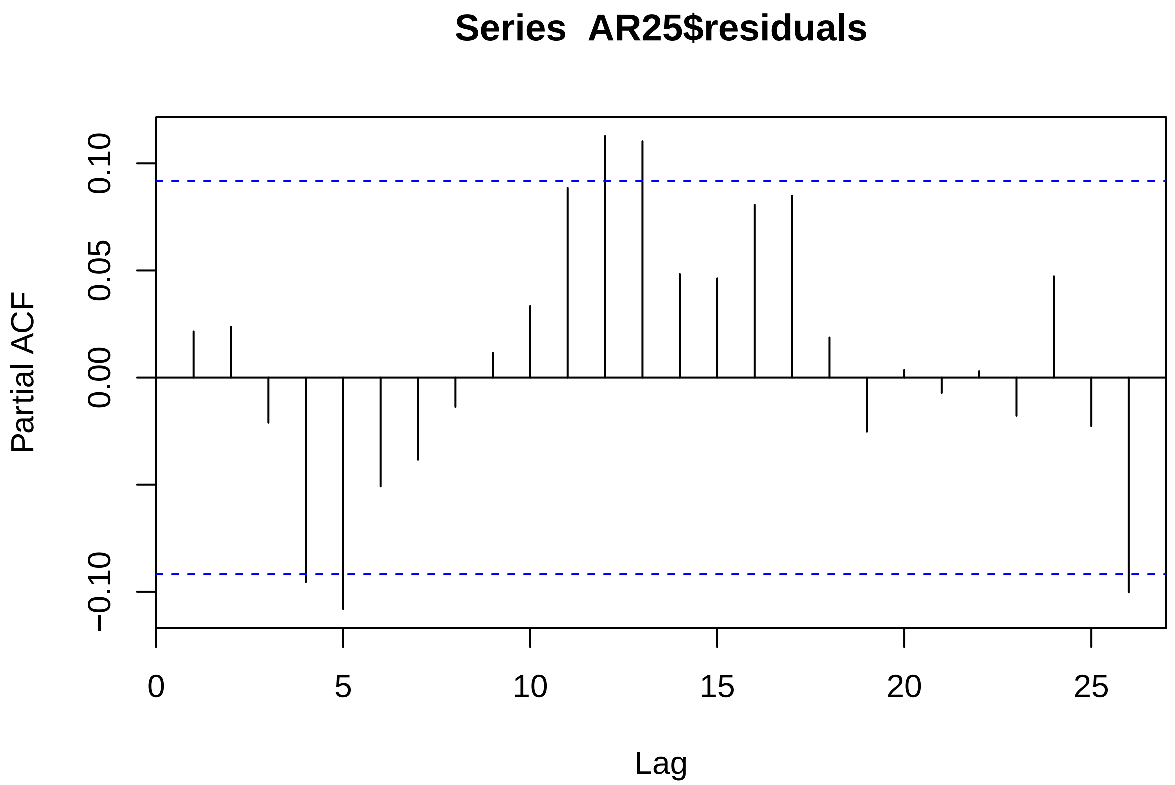
June

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

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

### Log the response variable
train6$count = log(train6$count)

# head(train6)
# head(test6)

AR25 <- arima(train6$count,order=c(25,0,0))
# tsdisplay(residuals(AR25),lag.max=25,main="AR(24) Resid. Diagnostics")
```

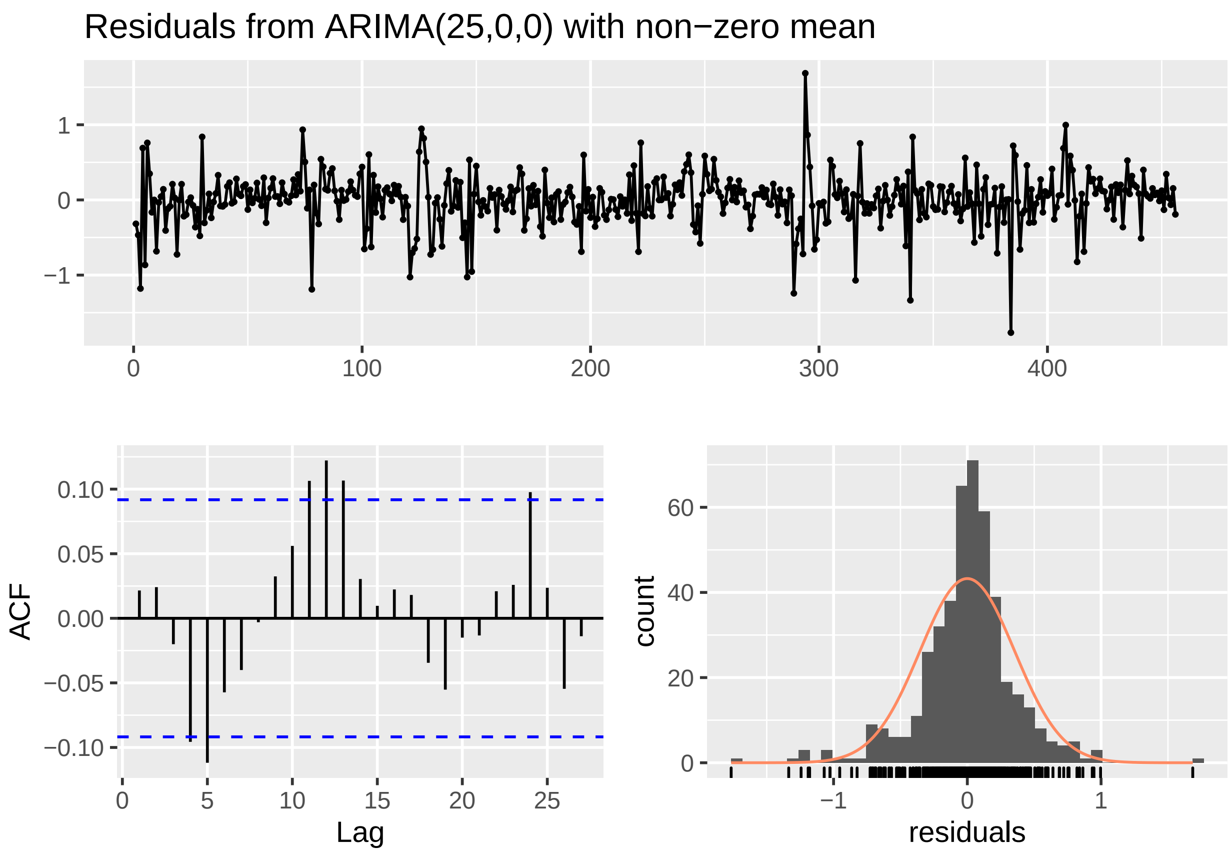
```
number = nrow(test6)

acf(AR25$residuals)
```



```
pacf(AR25$residuals)
```

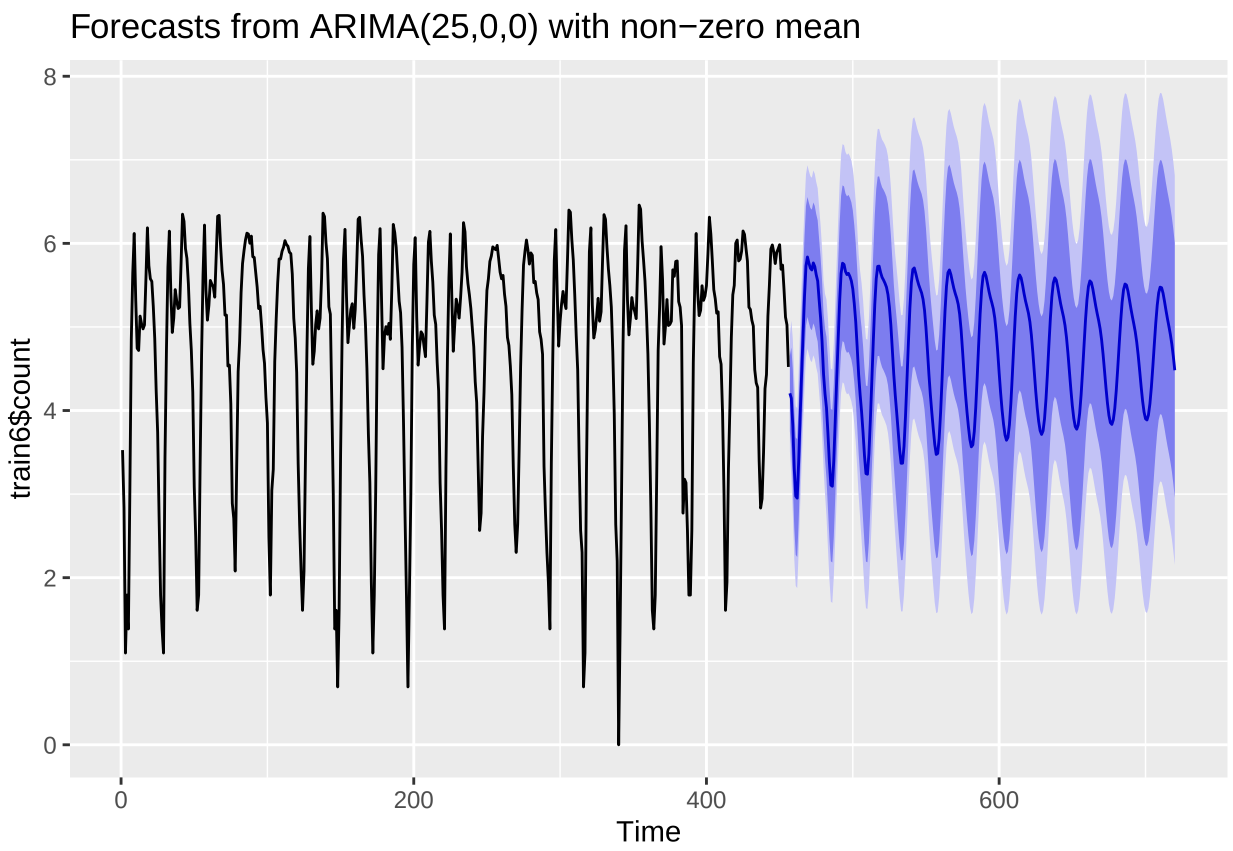


```
checkresiduals(AR25)
```



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0) with non-zero mean
## Q* = 43.898, df = 3, p-value = 1.587e-09
##
## Model df: 26.   Total lags used: 29
```

```
fcst <- forecast(AR25, h=number)

autoplot(fcst)
```



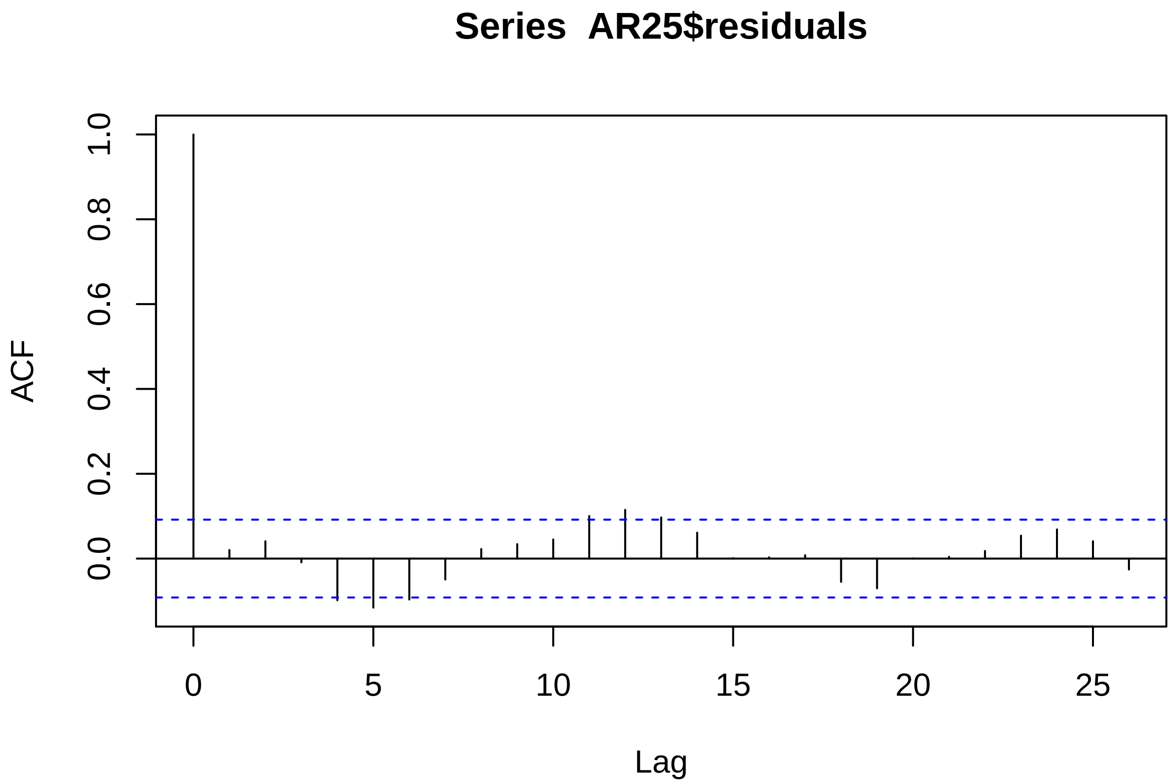
```
# point estimate (mean)
test6$count <- fcst$mean

RMSLE(y_pred = fcst$fitted, y_true = train6$count)
```

```
## [1] 0.09993809
```

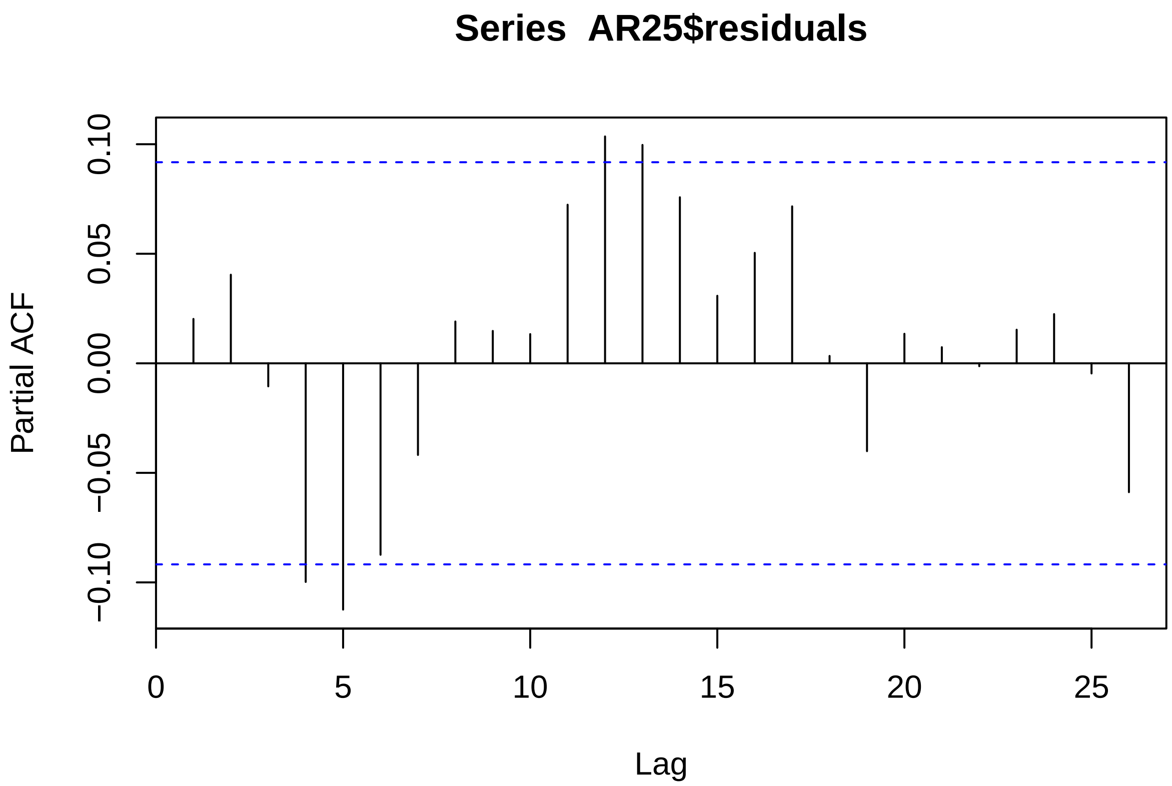
July

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

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


### Log the response variable
train7$count = log(train7$count)

# head(train7)
# head(test7)

AR25 <- arima(train7$count,order=c(25,0,0))
```

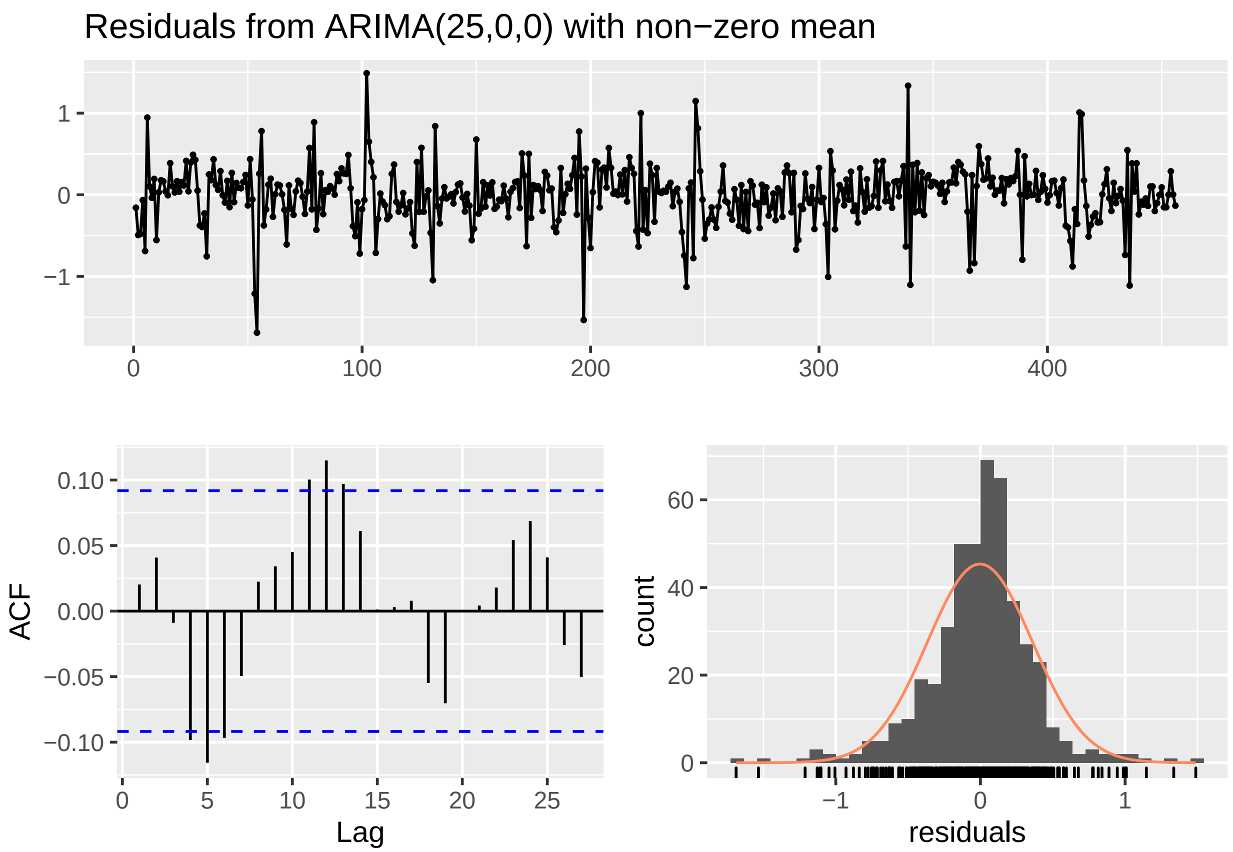
```r
number = nrow(test7)

acf(AR25$residuals)
```



```r
pacf(AR25$residuals)
```

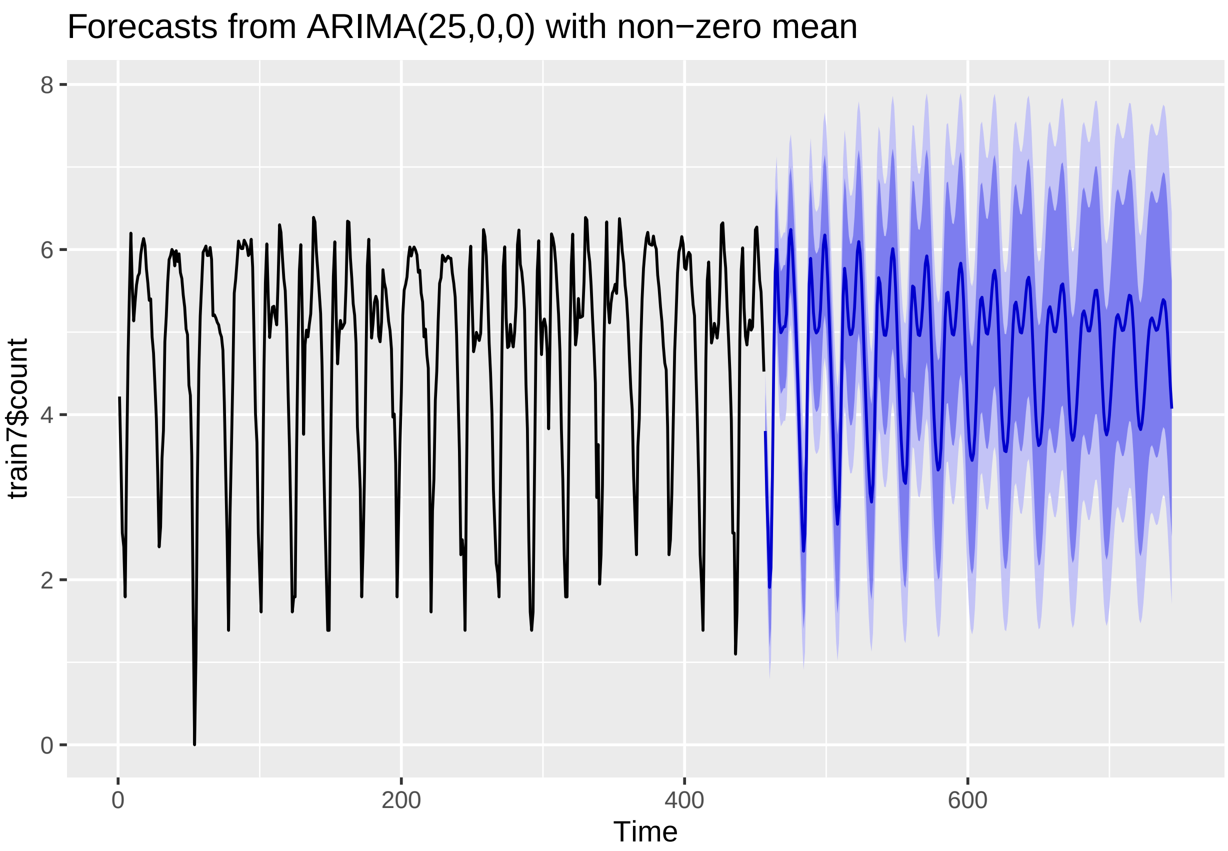


```
checkresiduals(AR25)
```



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0) with non-zero mean
## Q* = 46.28, df = 3, p-value = 4.944e-10
##
## Model df: 26.   Total lags used: 29
```

```
fcst <- forecast(AR25, h=number)

autoplot(fcst)
```



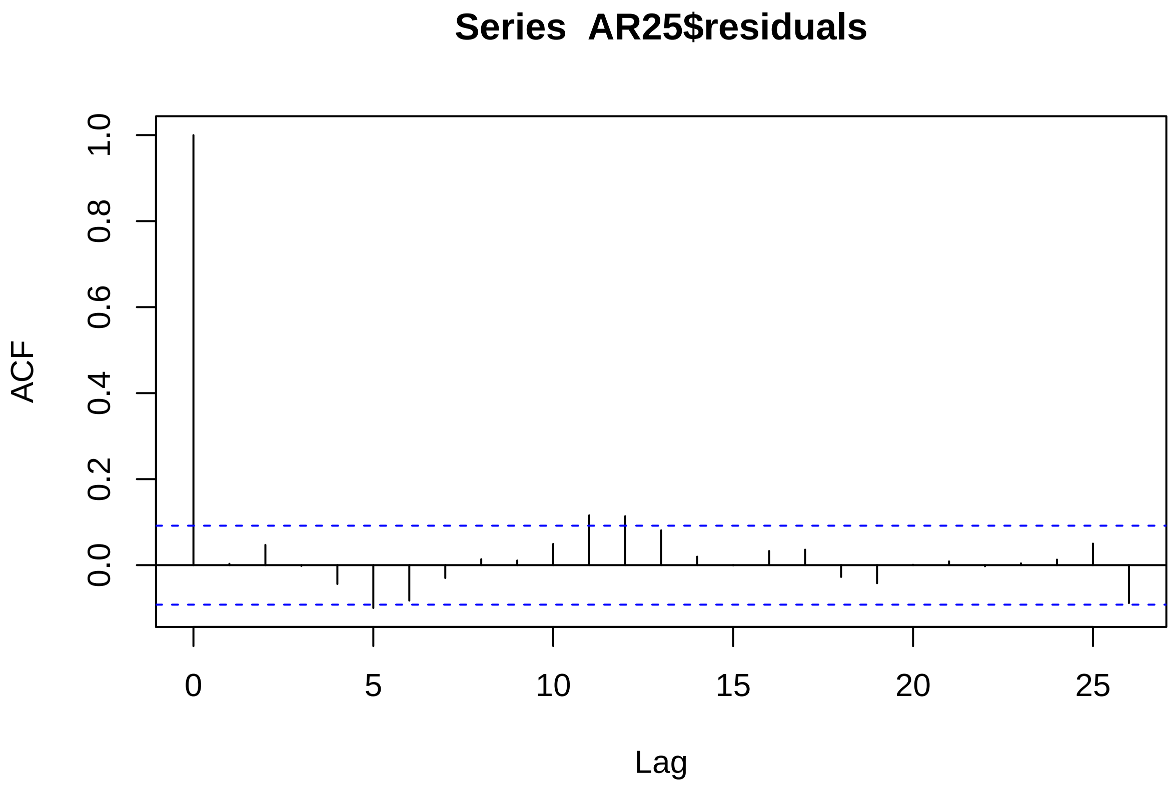
```r
# point estimate (mean)
test7$count <- fcst$mean


RMSLE(y_pred = fcst$fitted, y_true = train7$count)
```

```
## [1] 0.09859652
```

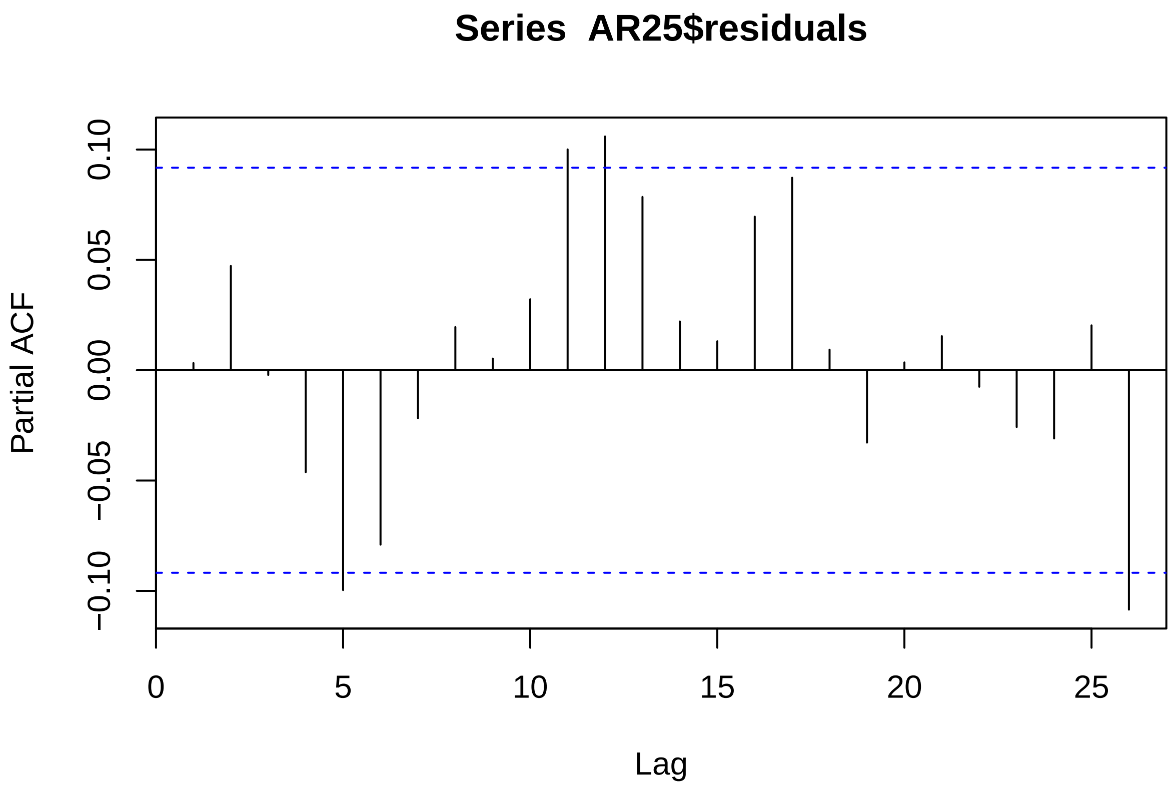
August

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

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



### Log the response variable
train8$count = log(train8$count)

# head(train8)
# head(test8)
```

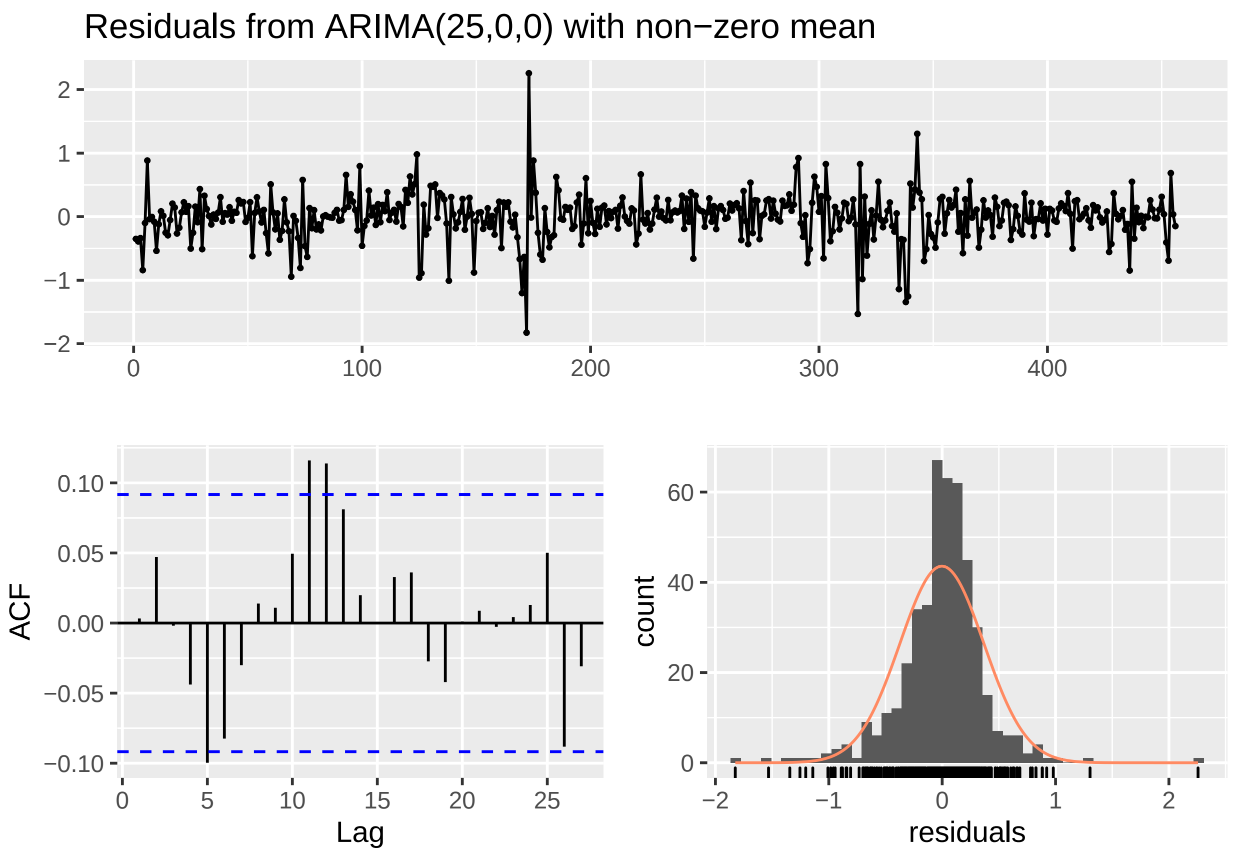
```
AR25 <- arima(train8$count,order=c(25,0,0))

number = nrow(test8)

acf(AR25$residuals)
```



```
pacf(AR25$residuals)
```

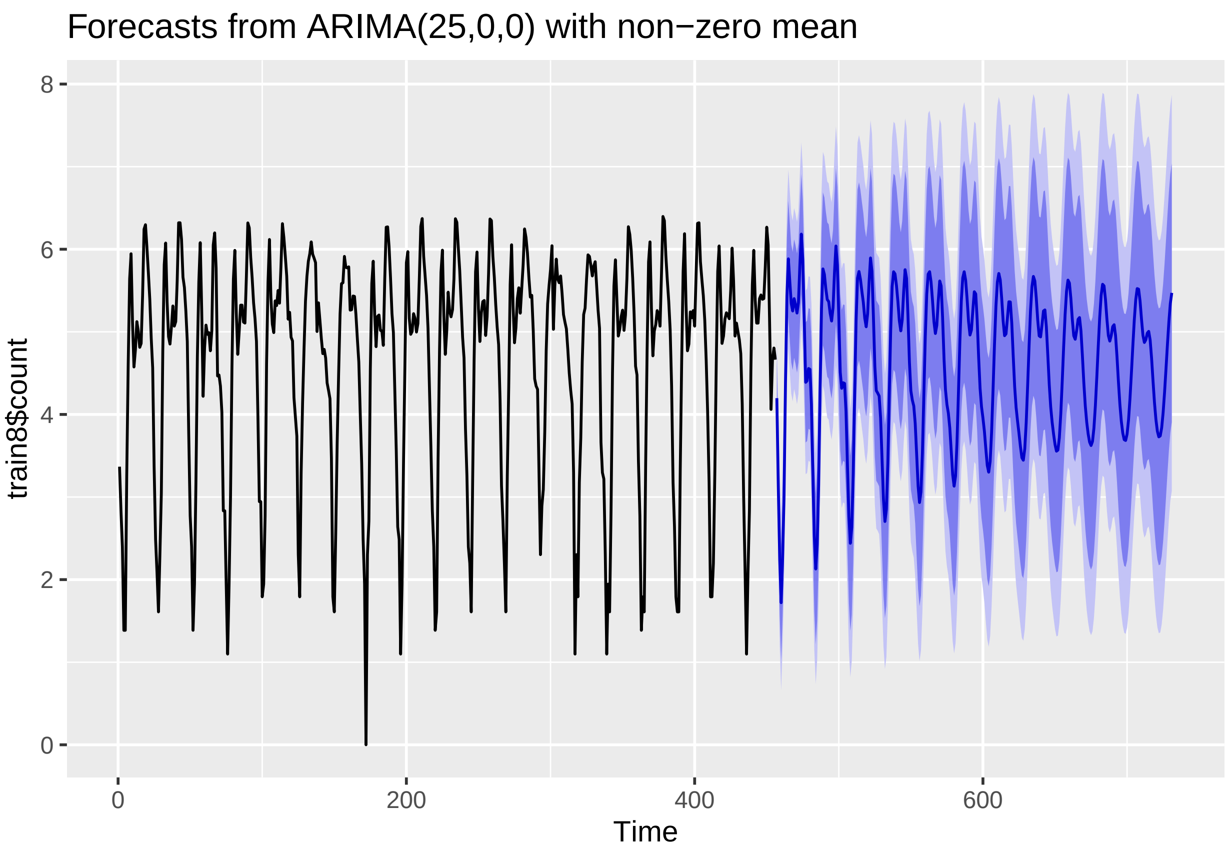


```r
checkresiduals(AR25)
```



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0) with non-zero mean
## Q* = 35.14, df = 3, p-value = 1.138e-07
##
## Model df: 26.   Total lags used: 29
```

```
fcst <- forecast(AR25, h=number)

autoplot(fcst)
```



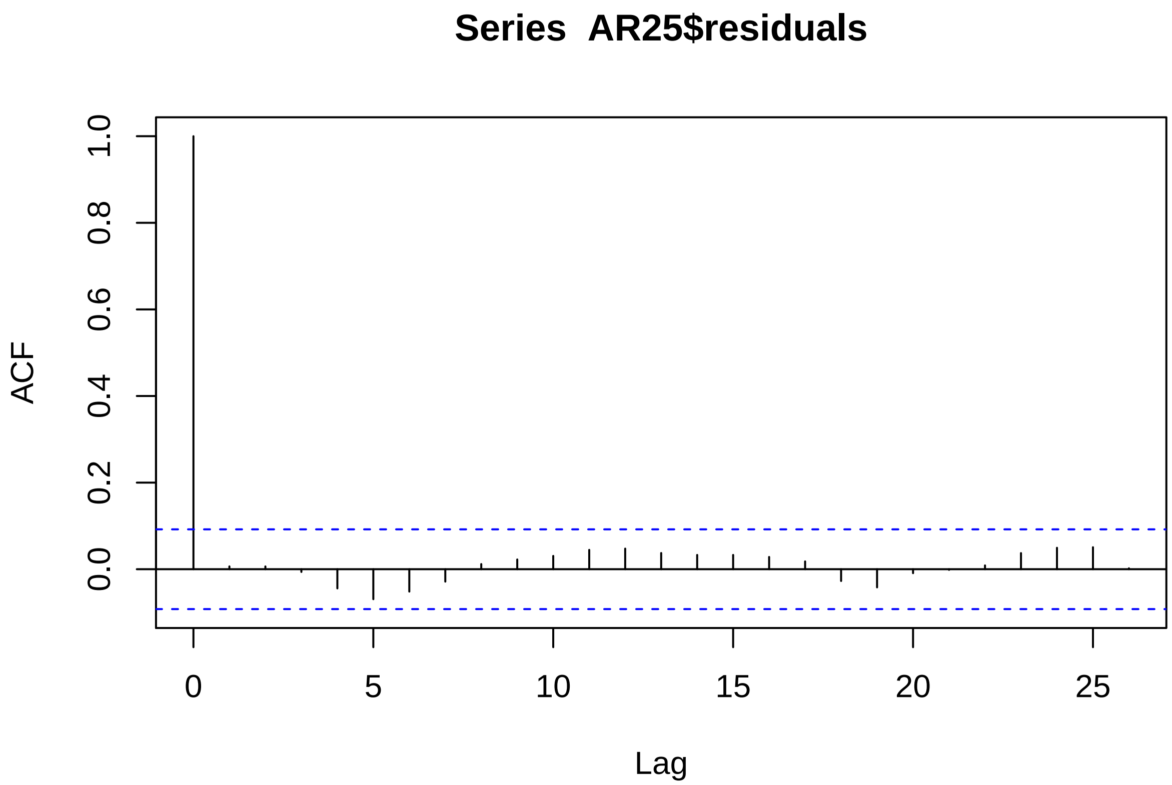
```
# point estimate (mean)
test8$count <- fcst$mean

RMSLE(y_pred = fcst$fitted, y_true = train8$count)
```

```
## [1] 0.1118663
```

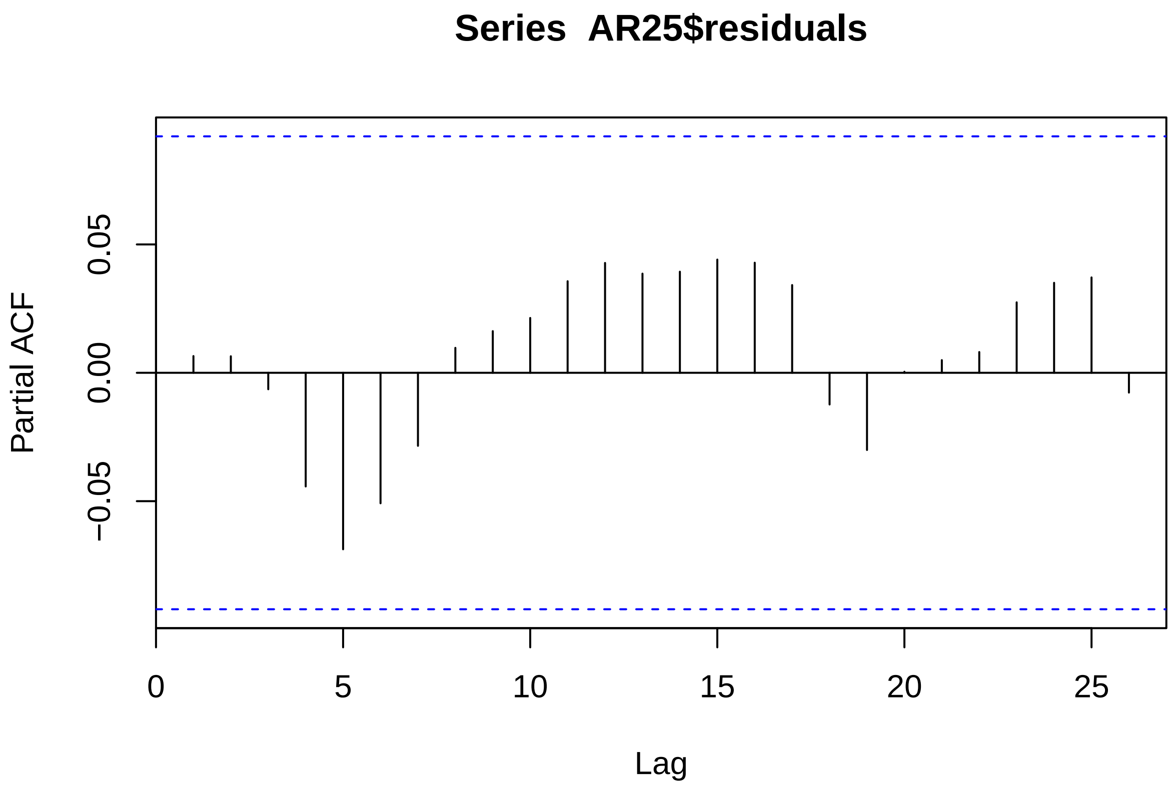
September

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

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



### Log the response variable
train9$count = log(train9$count)

# head(train9)
# head(test9)
```

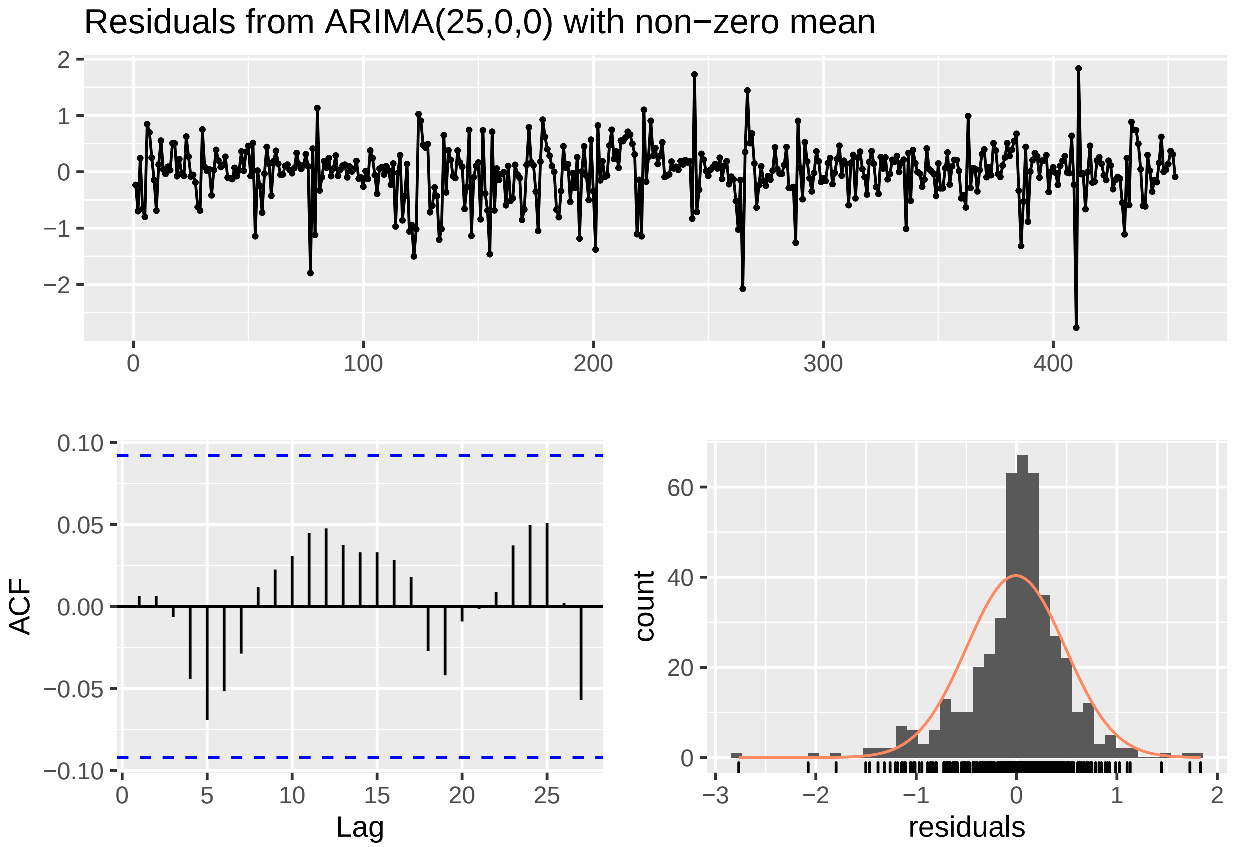
```
AR25 <- arima(train9$count,order=c(25,0,0))

number = nrow(test9)

acf(AR25$residuals)
```



```
pacf(AR25$residuals)
```

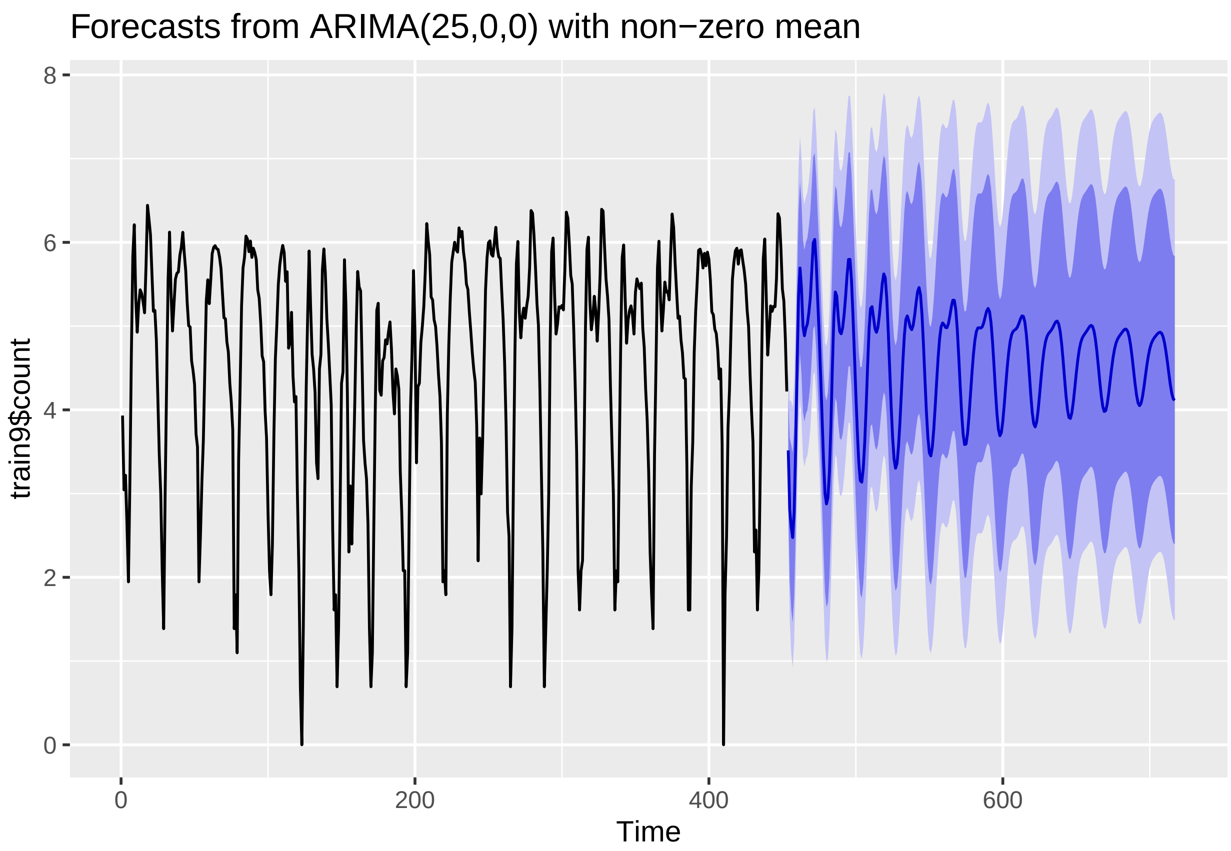


```
checkresiduals(AR25)
```



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0) with non-zero mean
## Q* = 16.133, df = 3, p-value = 0.001065
##
## Model df: 26.   Total lags used: 29
```

```
fcst <- forecast(AR25, h=number)

autoplot(fcst)
```



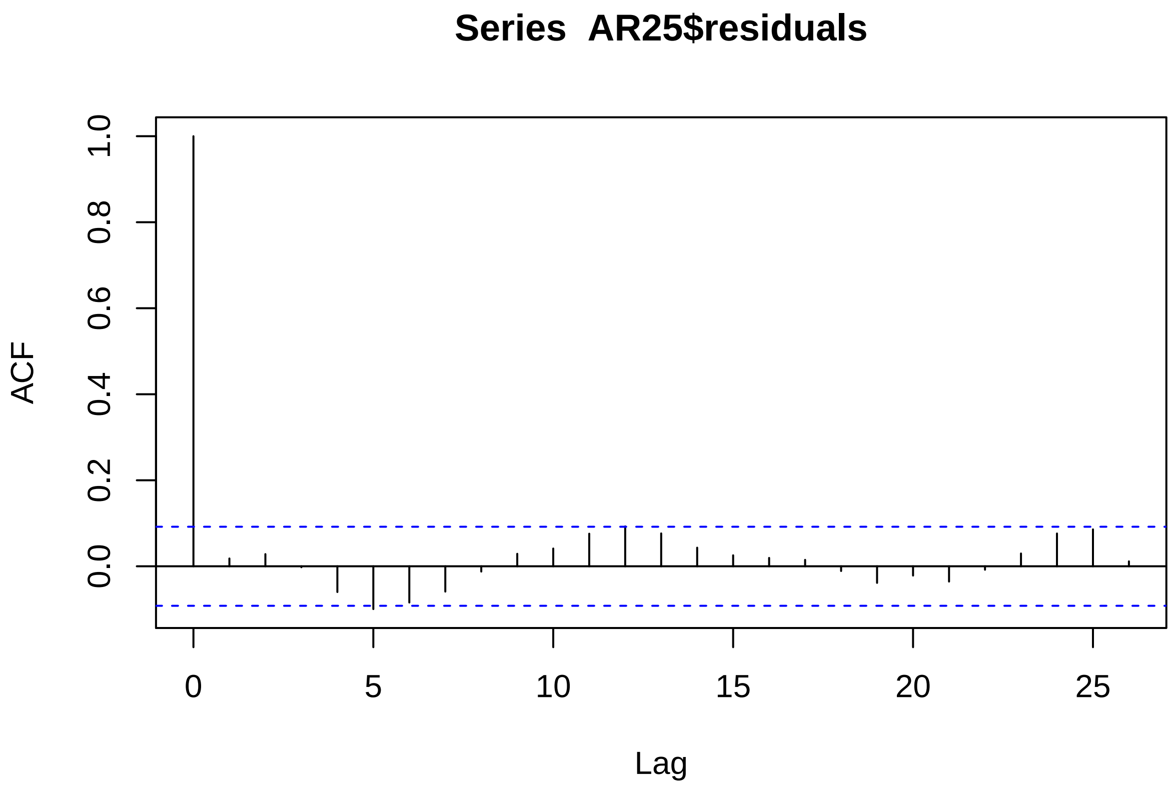
```
# point estimate (mean)
test9$count <- fcst$mean


RMSLE(y_pred = fcst$fitted, y_true = train9$count)
```

```
## [1] 0.1544213
```

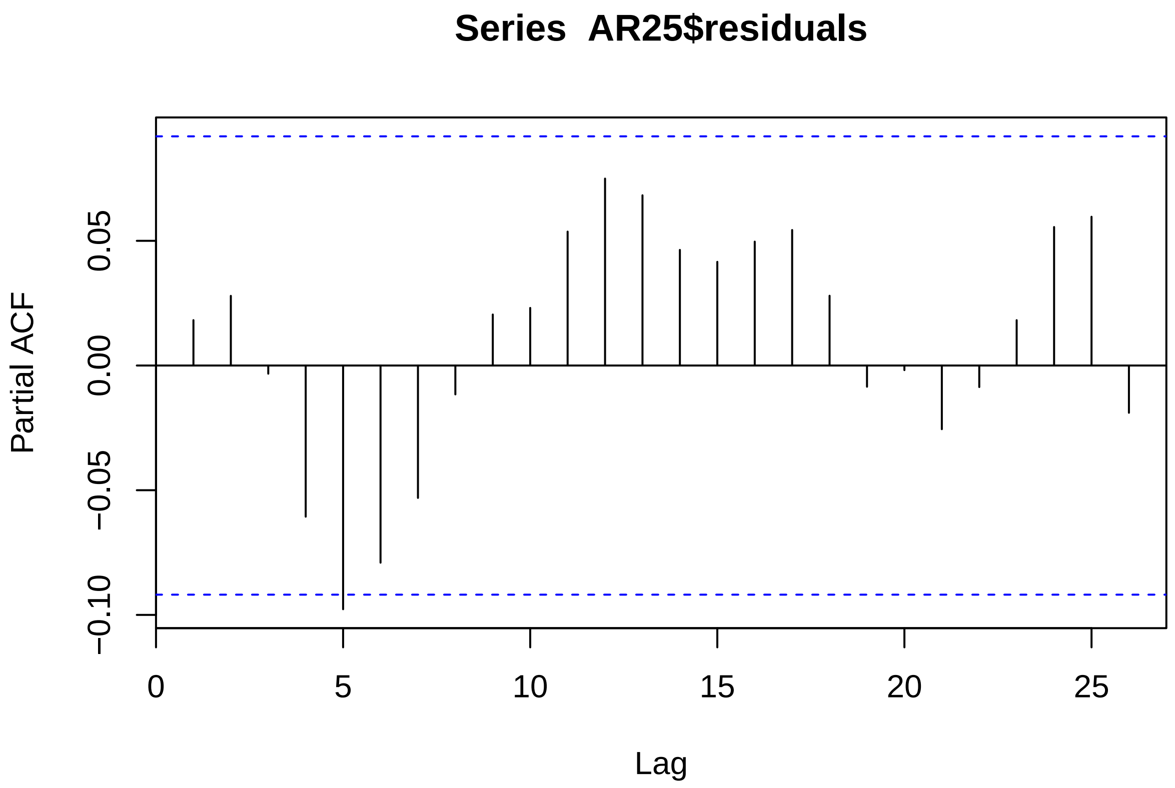
October

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

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


### Log the response variable
train10$count = log(train10$count)

# head(train10)
# head(test10)
```

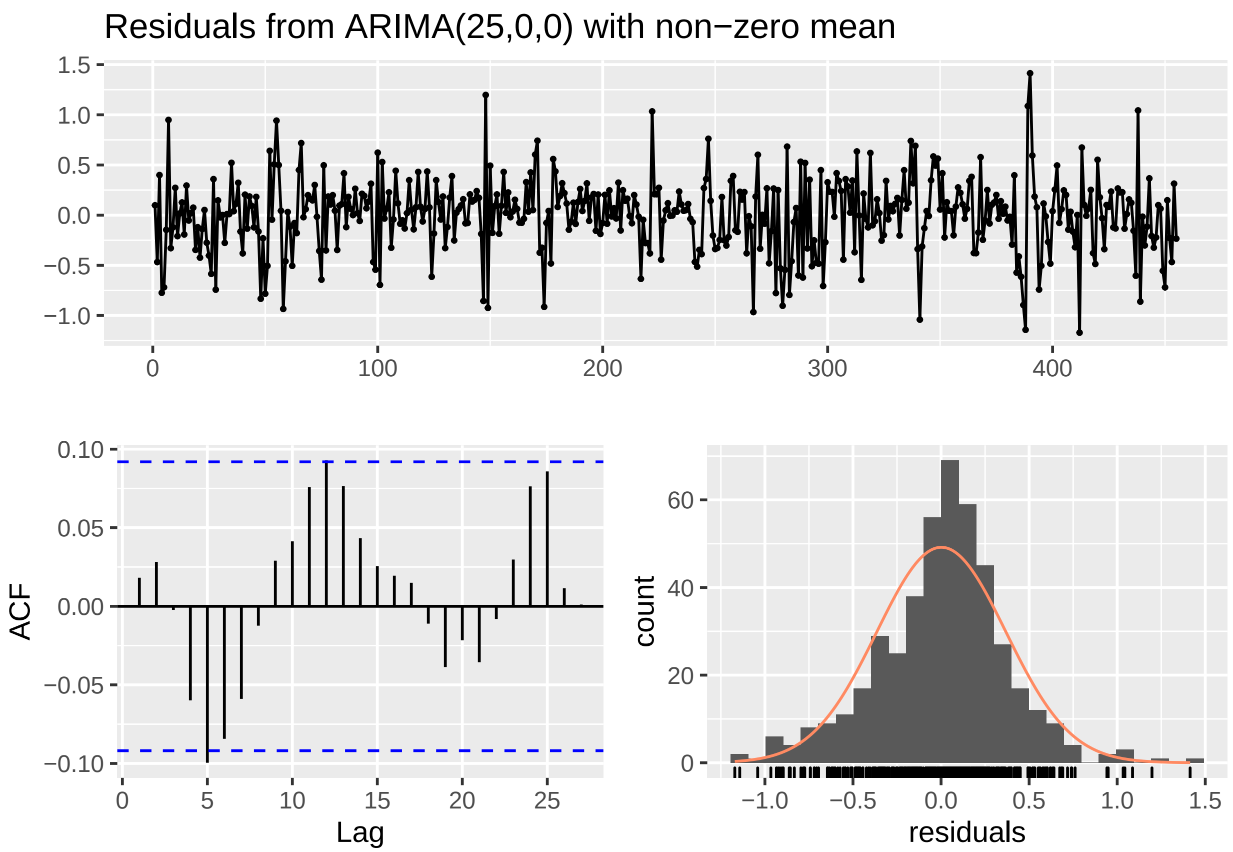
```
AR25 <- arima(train10$count,order=c(25,0,0))

number = nrow(test10)

acf(AR25$residuals)
```



```
pacf(AR25$residuals)
```

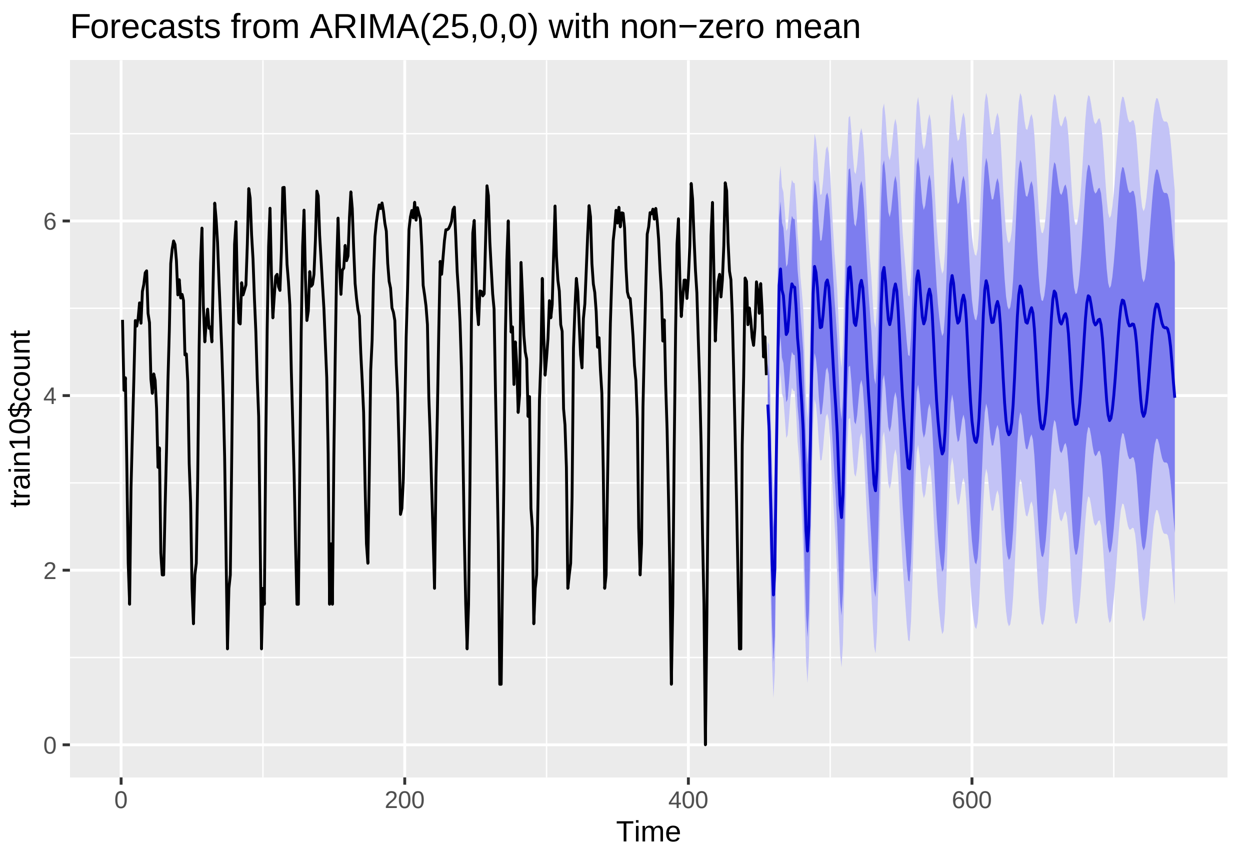


```r
checkresiduals(AR25)
```



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0) with non-zero mean
## Q* = 32.663, df = 3, p-value = 3.793e-07
##
## Model df: 26.   Total lags used: 29
```

```
fcst <- forecast(AR25, h=number)

autoplot(fcst)
```



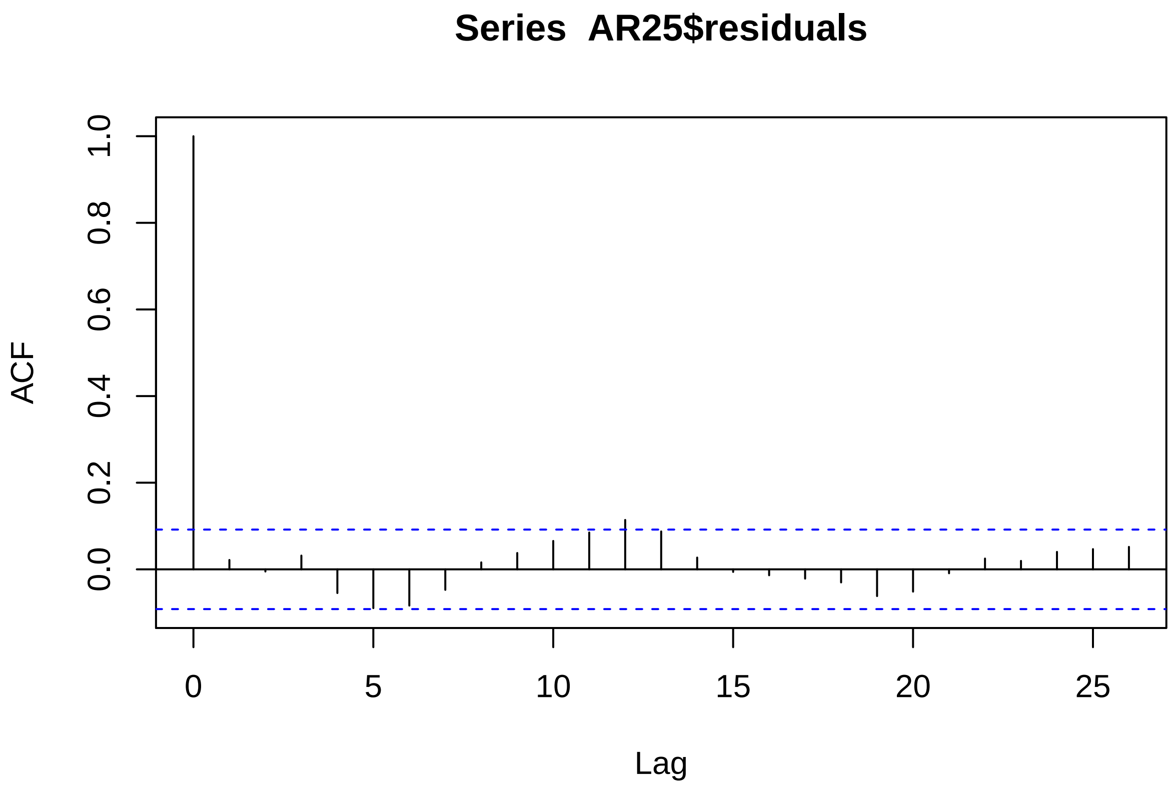
```
# point estimate (mean)
test10$count <- fcst$mean


RMSLE(y_pred = fcst$fitted, y_true = train10$count)
```

```
## [1] 0.1022587
```

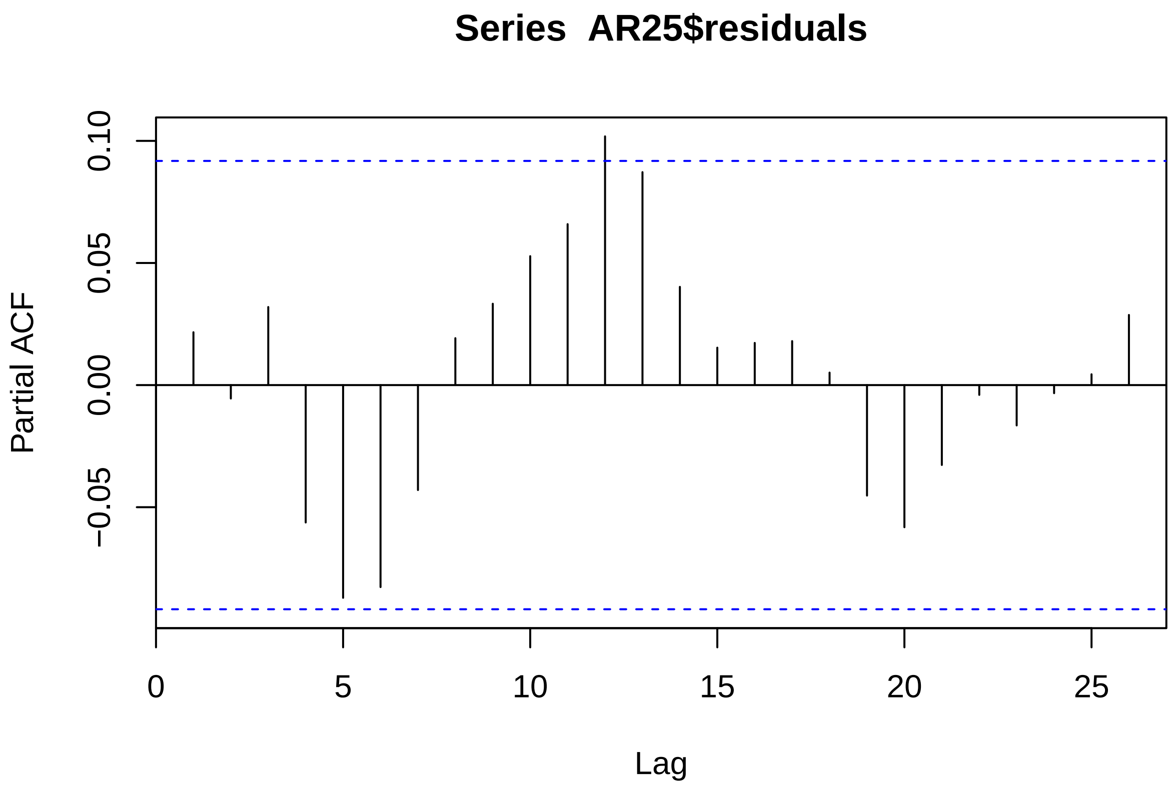
November

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

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


### Log the response variable
train11$count = log(train11$count)

# head(train11)
# head(test11)
```

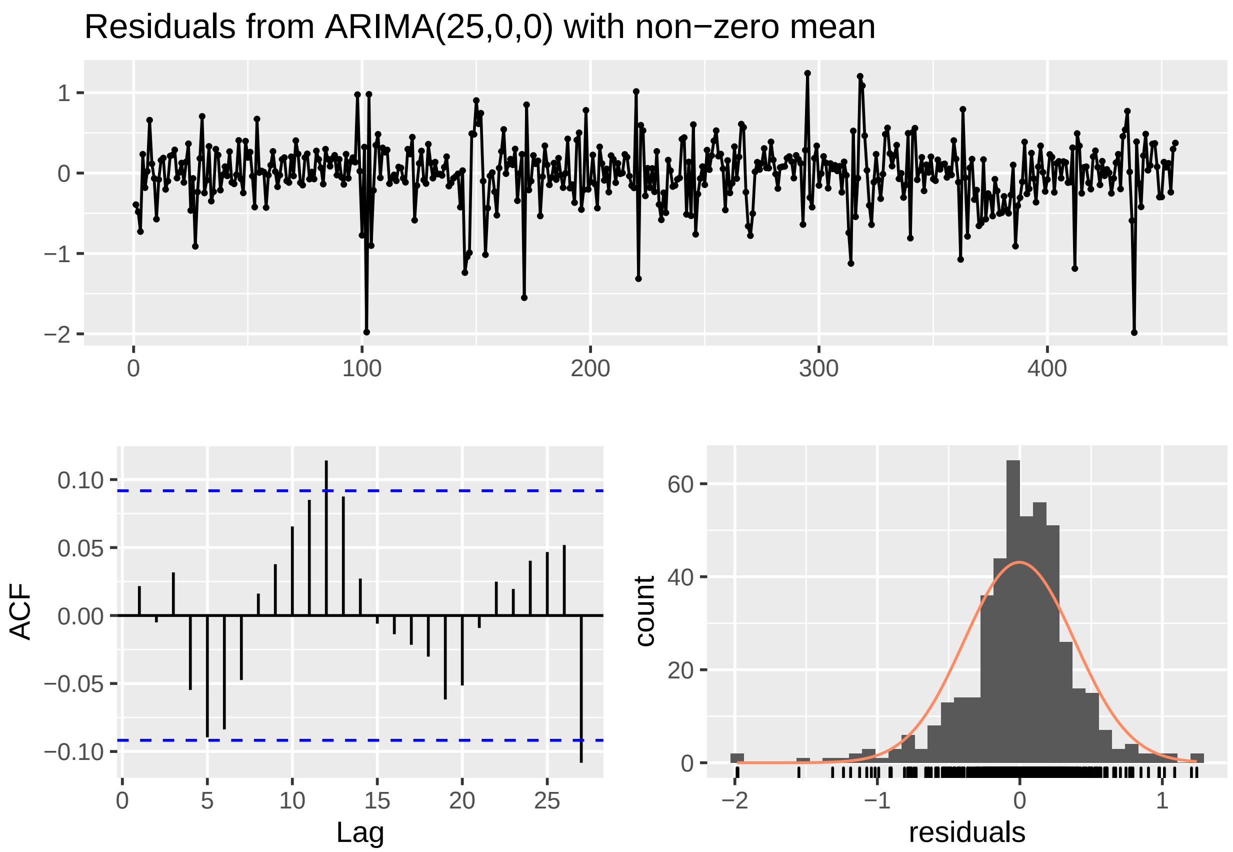
```
AR25 <- arima(train11$count,order=c(25,0,0))

number = nrow(test11)

acf(AR25$residuals)
```



```
pacf(AR25$residuals)
```

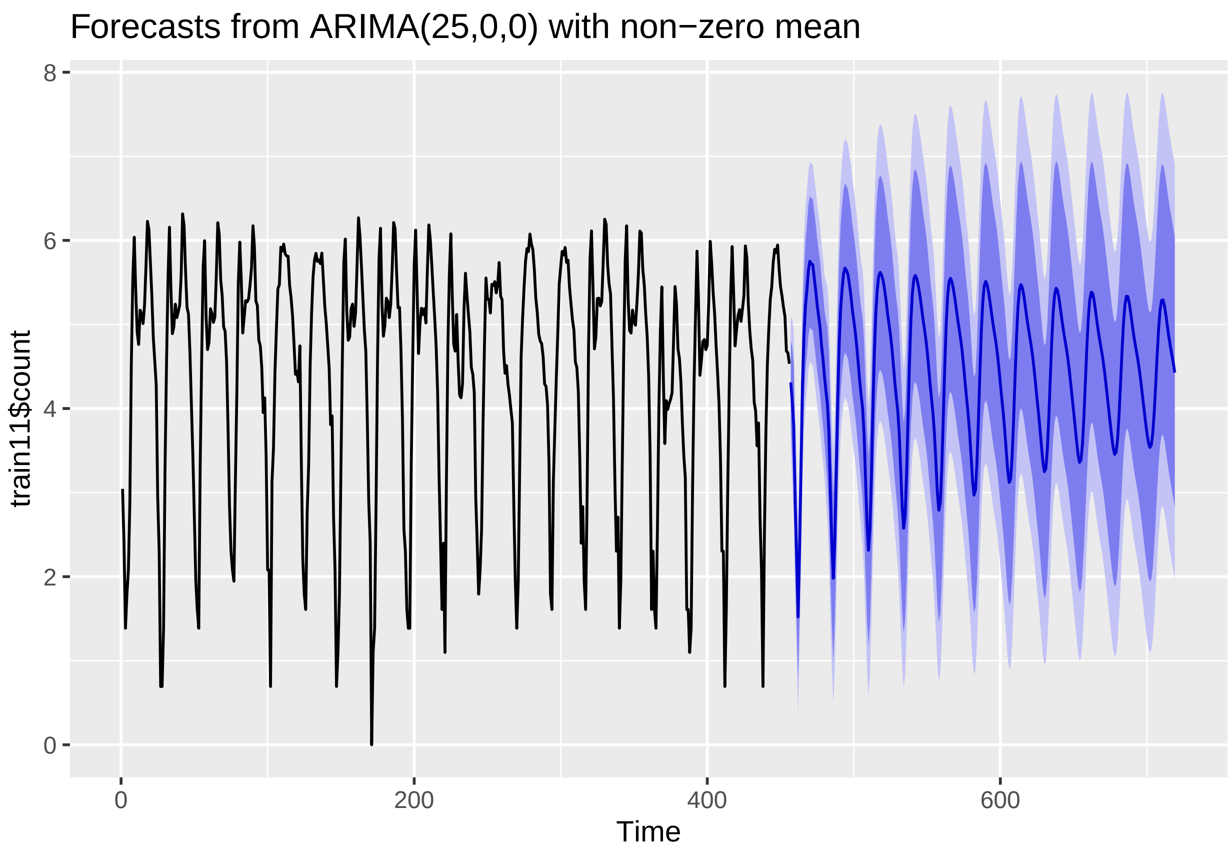


```r
checkresiduals(AR25)
```



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0) with non-zero mean
## Q* = 39.617, df = 3, p-value = 1.285e-08
##
## Model df: 26.   Total lags used: 29
```

```
fcst <- forecast(AR25, h=number)

autoplot(fcst)
```



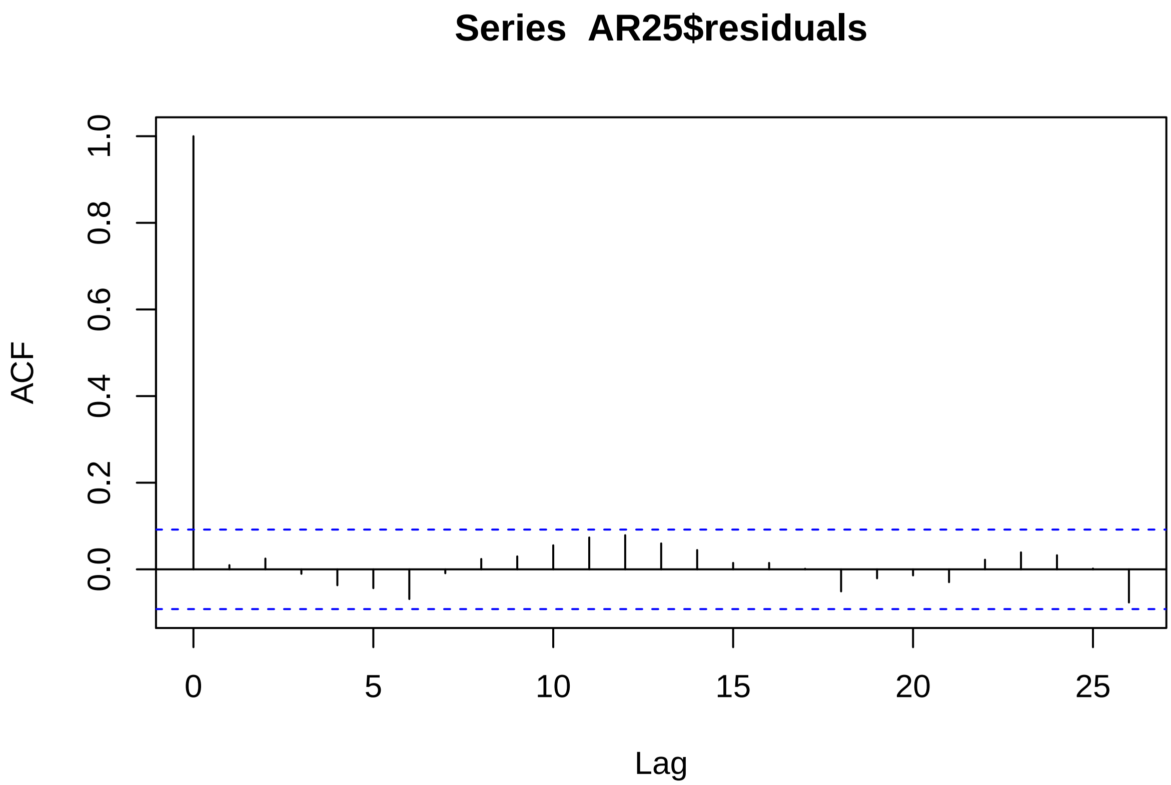
```r
# point estimate (mean)
test11$count <- fcst$mean


RMSLE(y_pred = fcst$fitted, y_true = train11$count)
```

```
## [1] 0.1204845
```

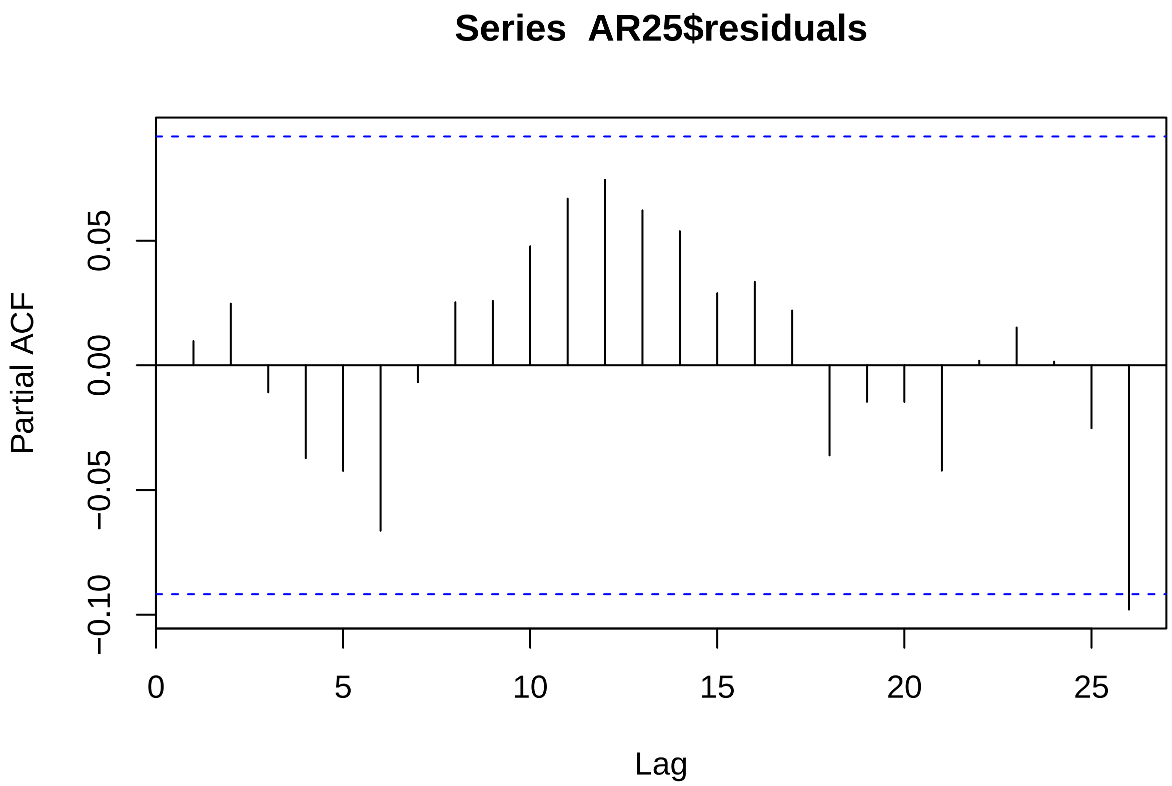
December

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

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


### Log the response variable
train12$count = log(train12$count)

# head(train12)
# head(test12)
```

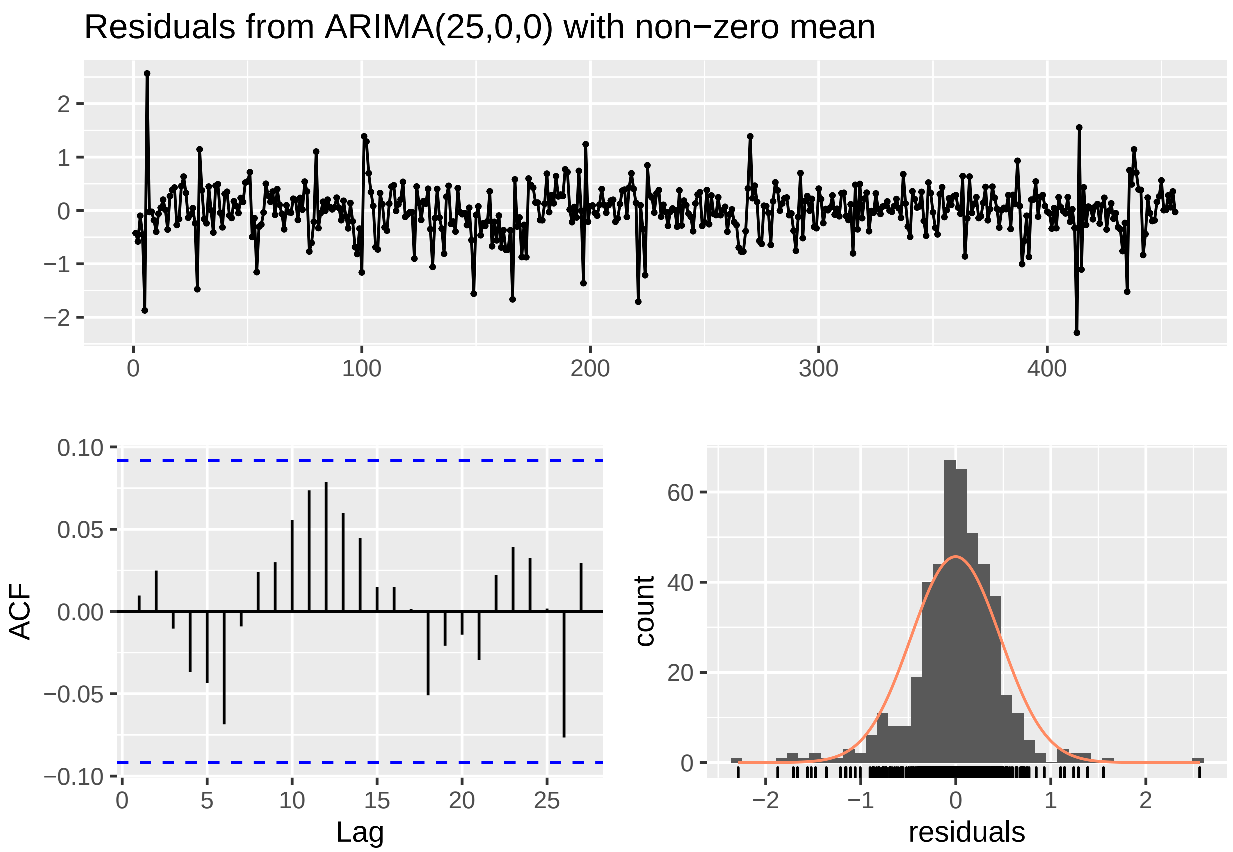
```
AR25 <- arima(train12$count,order=c(25,0,0))

number = nrow(test12)

acf(AR25$residuals)
```



```
pacf(AR25$residuals)
```

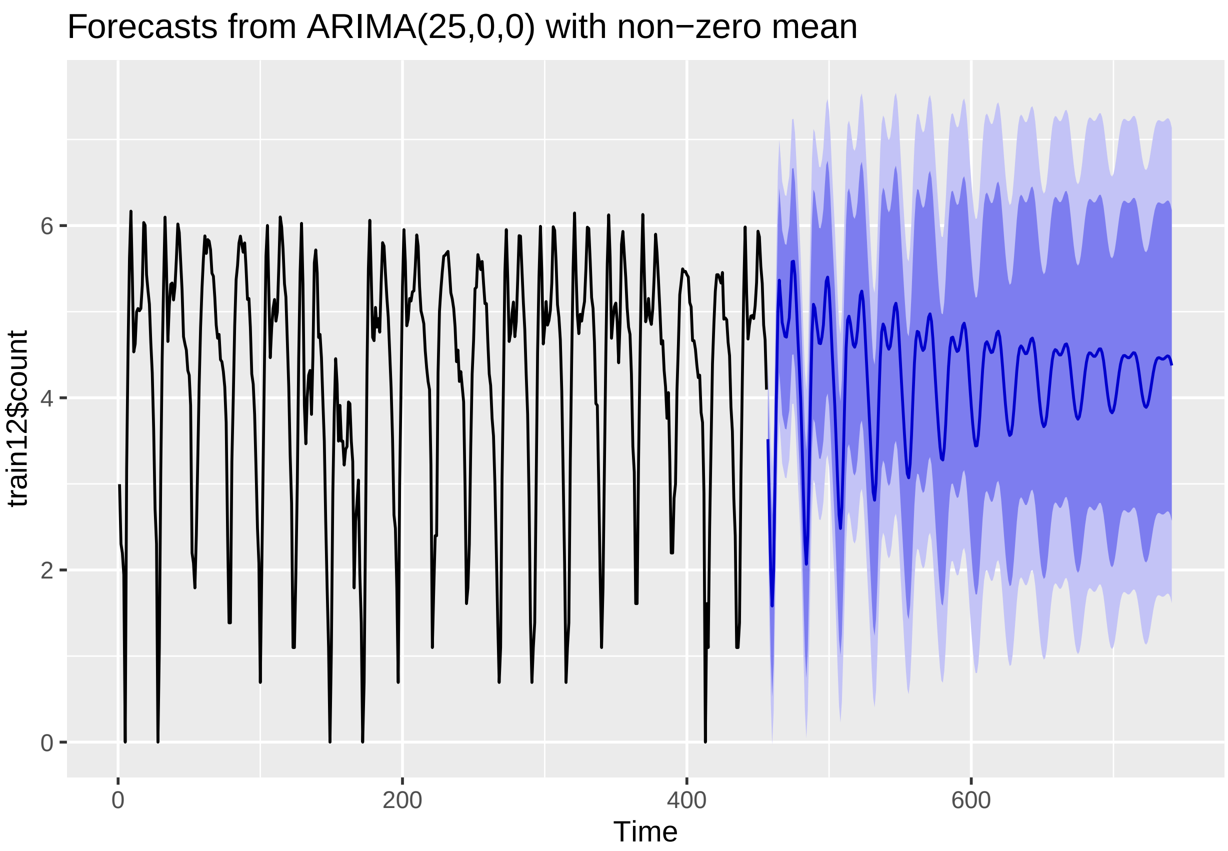


```r
checkresiduals(AR25)
```



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0) with non-zero mean
## Q* = 22.855, df = 3, p-value = 4.33e-05
##
## Model df: 26.   Total lags used: 29
```

```
fcst <- forecast(AR25, h=number)

autoplot(fcst)
```



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

RMSLE(y_pred = fcst$fitted, y_true = train12$count)
```

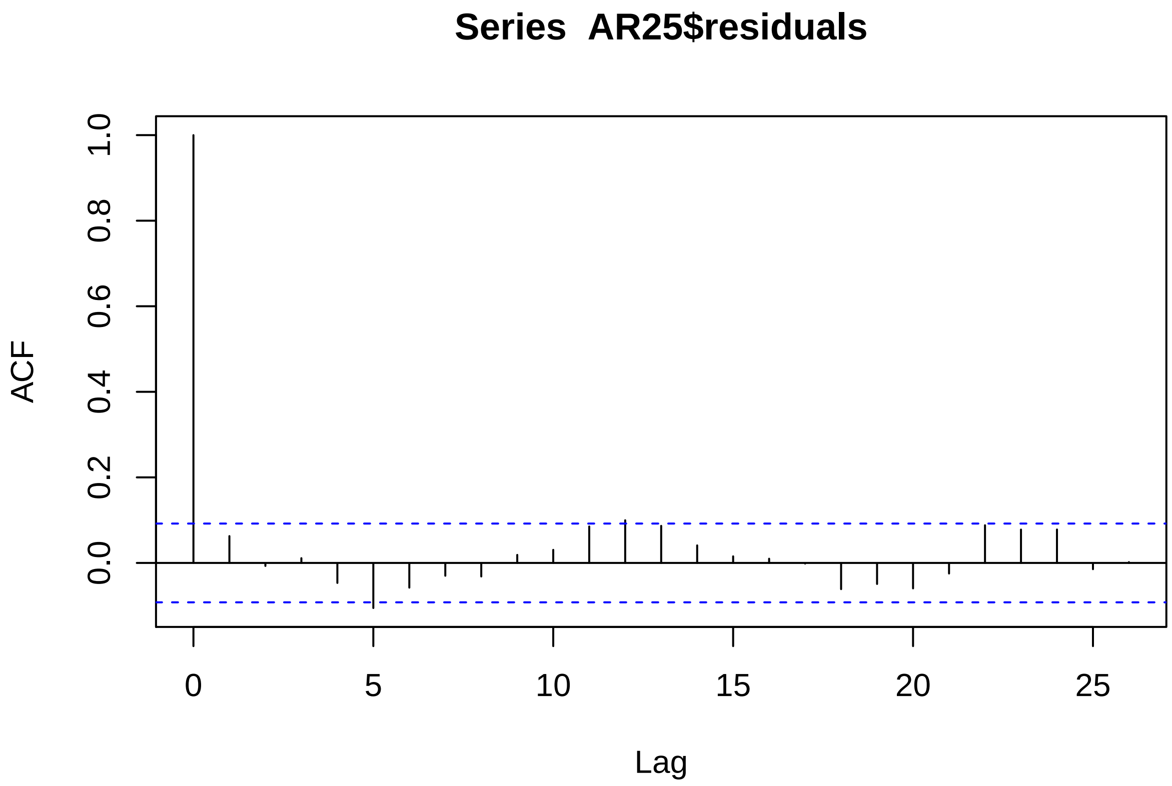
```
## [1] 0.1716392
```

2012

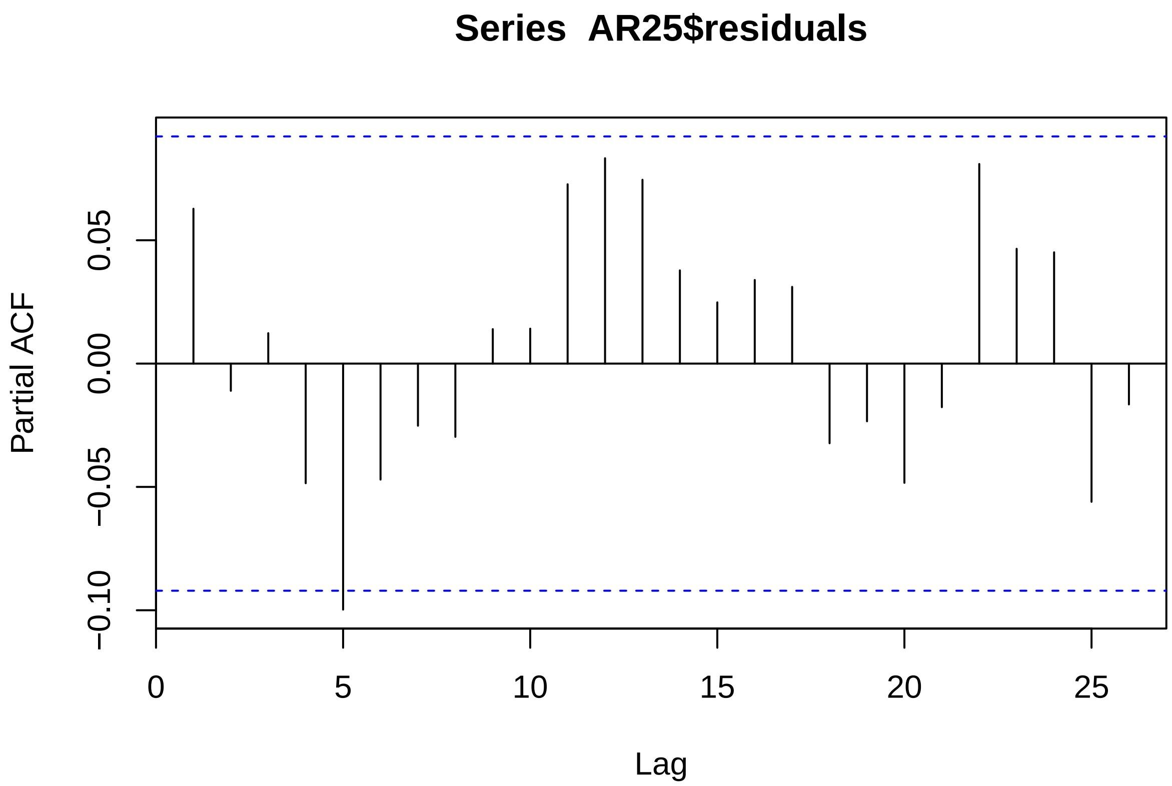
January

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

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


### Log the response variable
train13$count = log(train13$count)

# head(train13)
# head(test13)
```

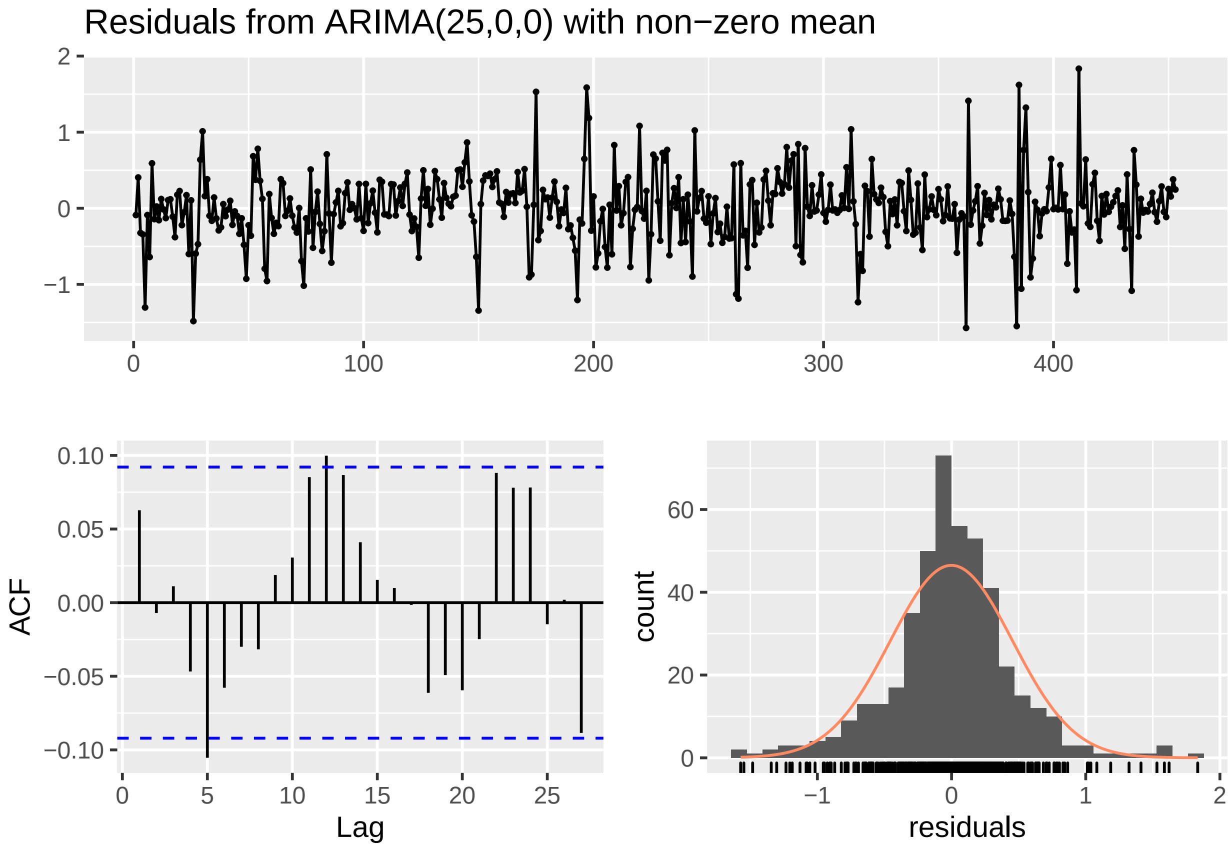
```
AR25 <- arima(train13$count,order=c(25,0,0))

number = nrow(test13)

acf(AR25$residuals)
```



```
pacf(AR25$residuals)
```

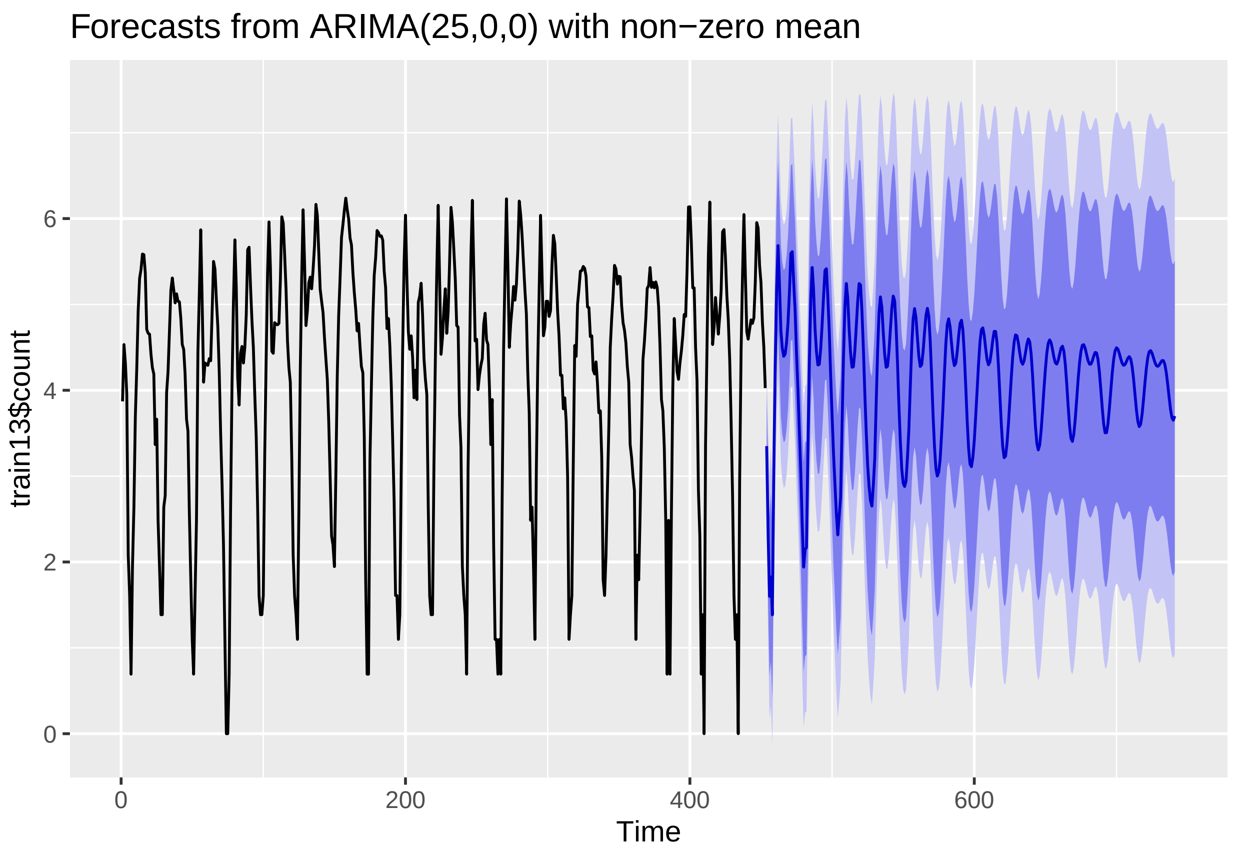


```
checkresiduals(AR25)
```



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0) with non-zero mean
## Q* = 44.308, df = 3, p-value = 1.298e-09
##
## Model df: 26.   Total lags used: 29
```

```
fcst <- forecast(AR25, h=number)

autoplot(fcst)
```



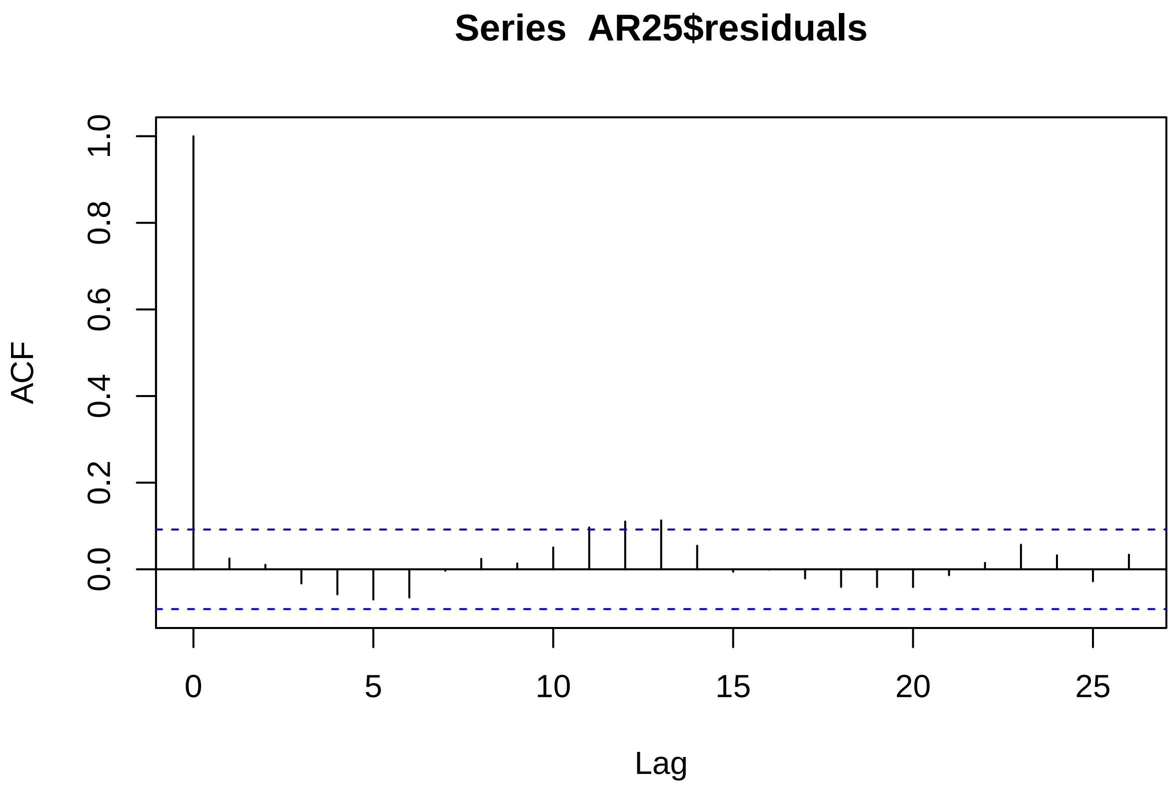
```
# point estimate (mean)
test13$count <- fcst$mean


RMSLE(y_pred = fcst$fitted, y_true = train13$count)
```

```
## [1] 0.1433694
```

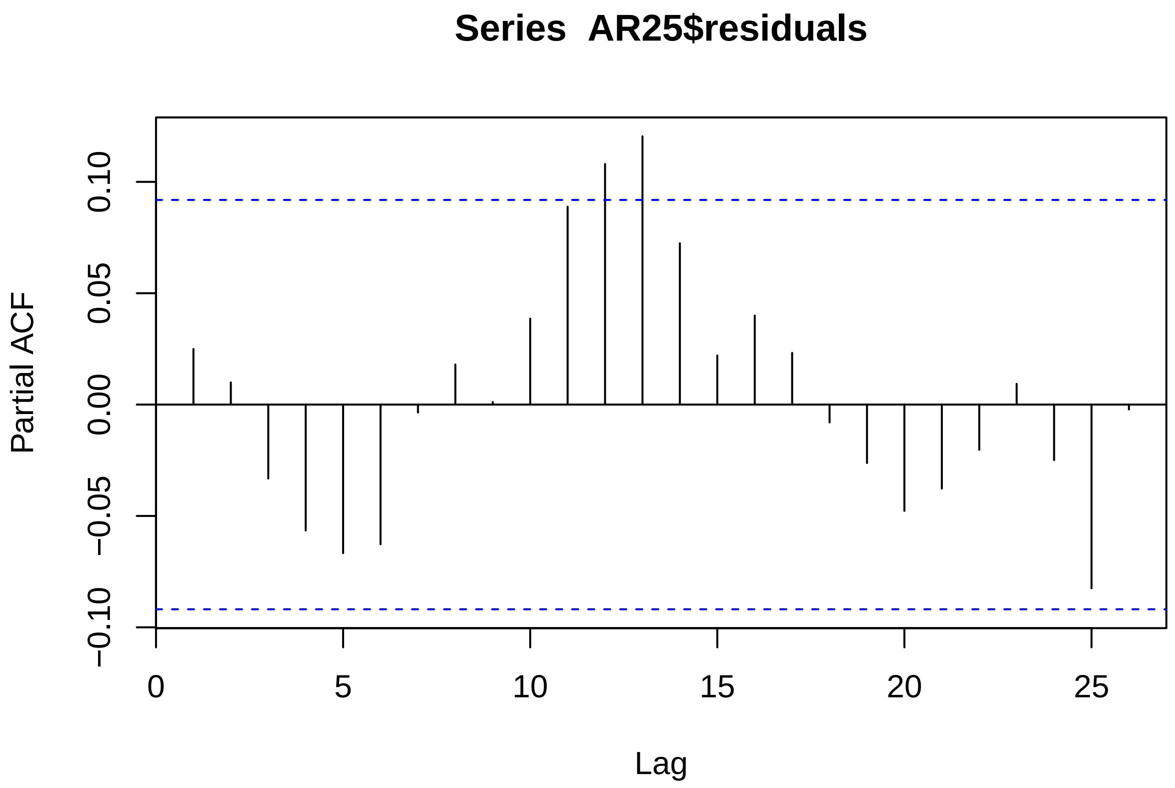
February

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

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


### Log the response variable
train14$count = log(train14$count)

# head(train14)
# head(test14)
```

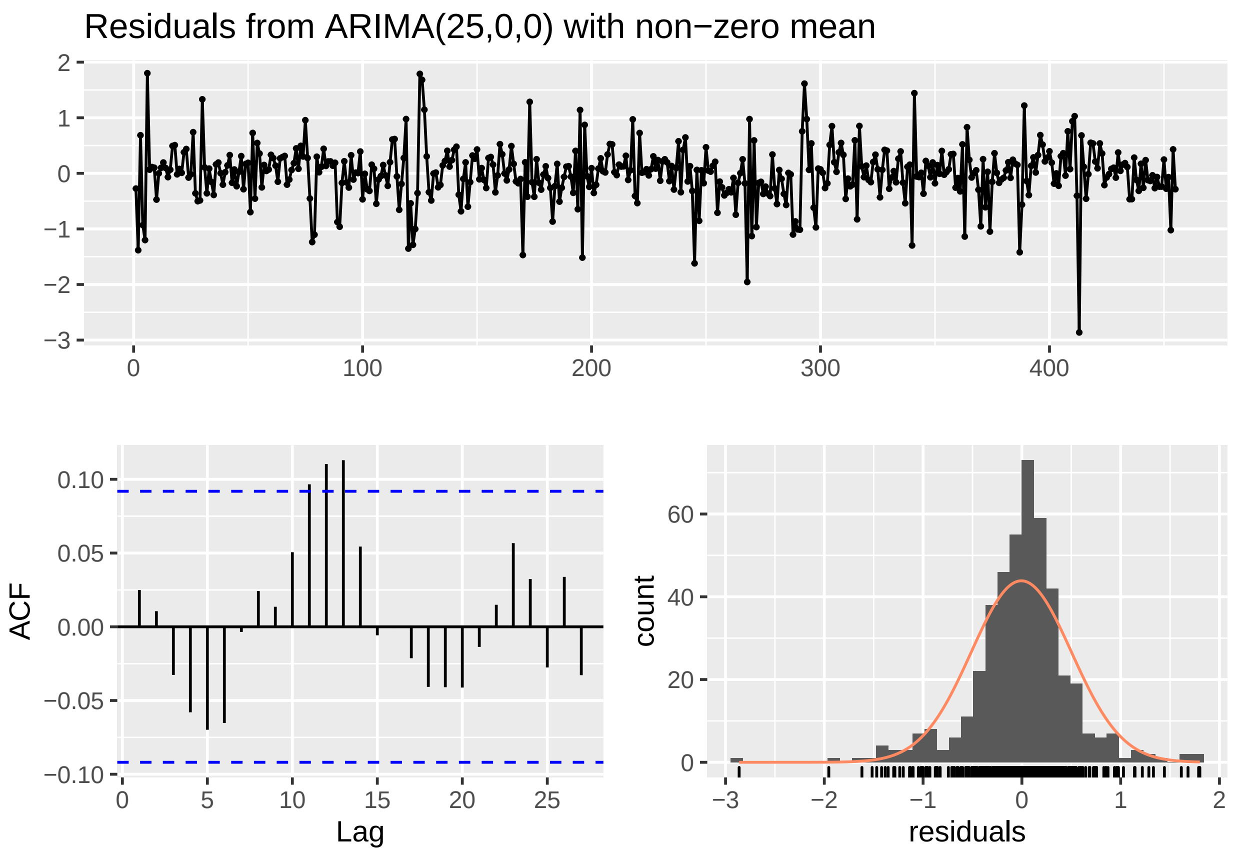
```
AR25 <- arima(train14$count,order=c(25,0,0))

number = nrow(test14)

acf(AR25$residuals)
```



```
pacf(AR25$residuals)
```

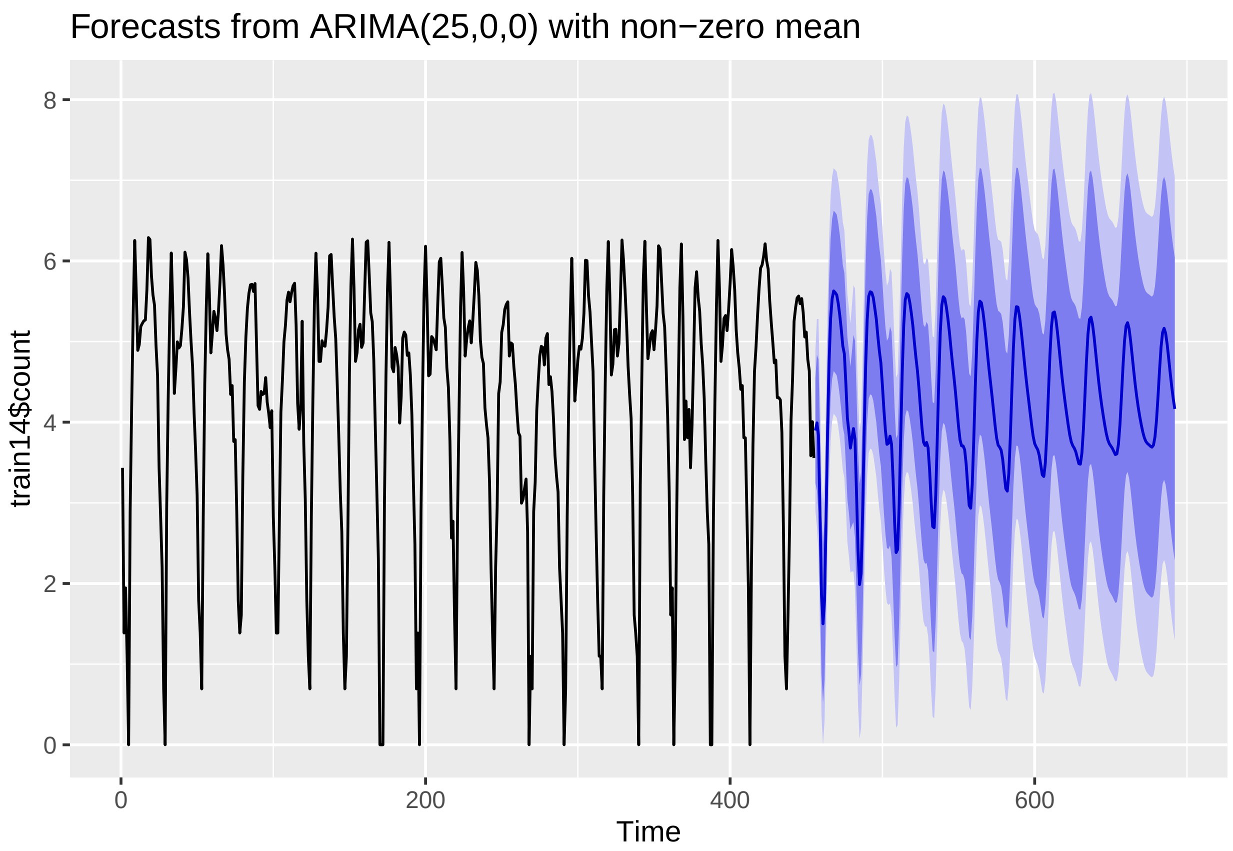


```
checkresiduals(AR25)
```



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0) with non-zero mean
## Q* = 32.137, df = 3, p-value = 4.897e-07
##
## Model df: 26.   Total lags used: 29
```

```
fcst <- forecast(AR25, h=number)

autoplot(fcst)
```



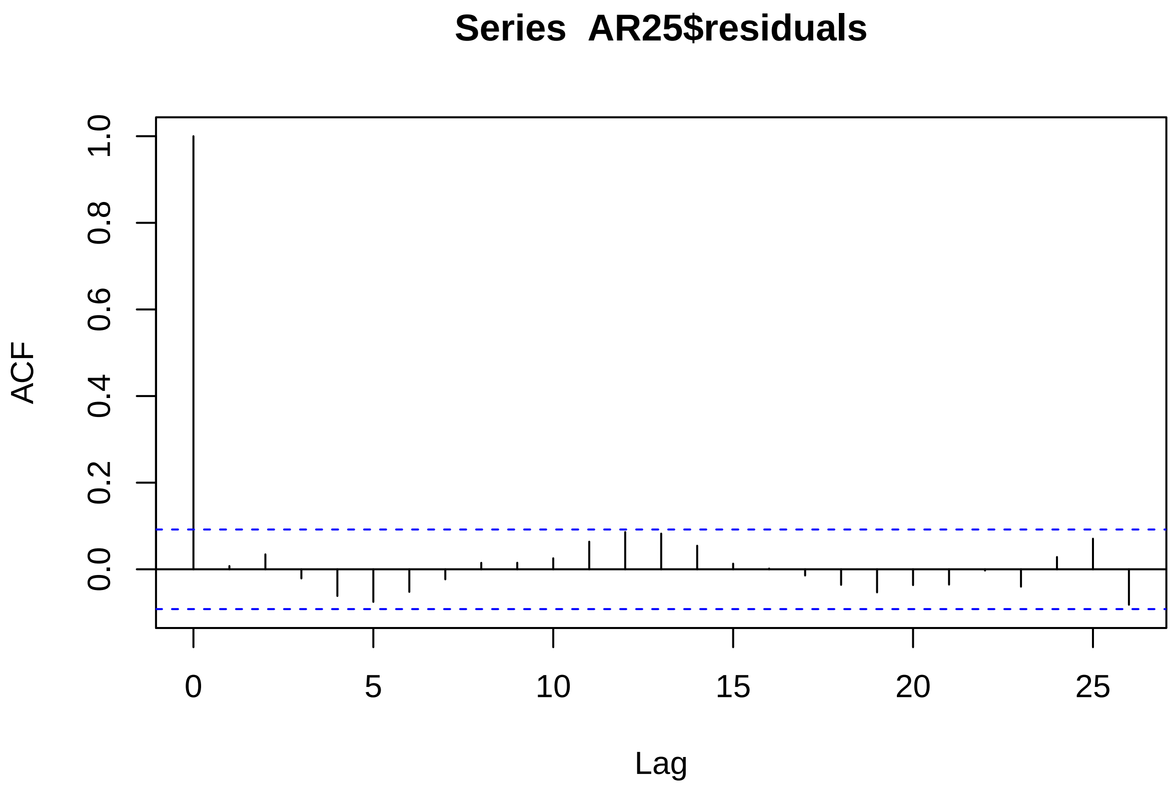
```r
# point estimate (mean)
test14$count <- fcst$mean


RMSLE(y_pred = fcst$fitted, y_true = train14$count)
```

```
## [1] 0.1921697
```

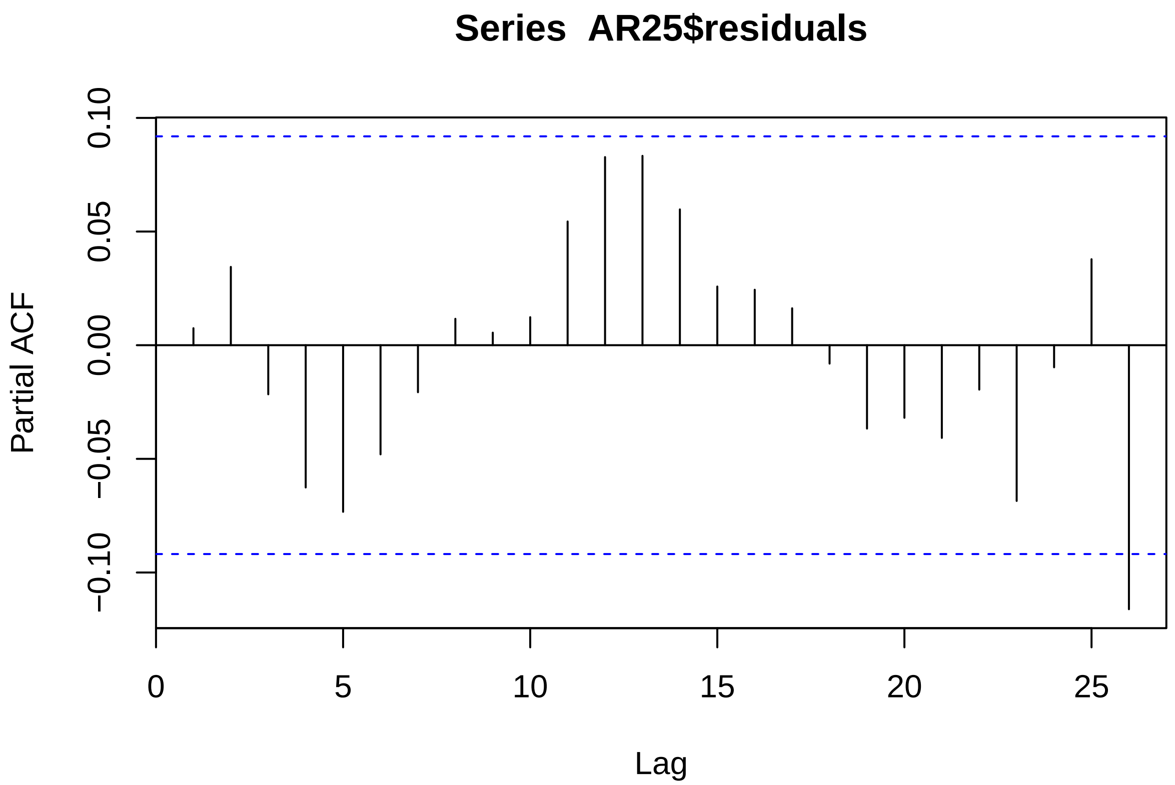
March

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

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


### Log the response variable
train15$count = log(train15$count)

# head(train15)
# head(test15)
```

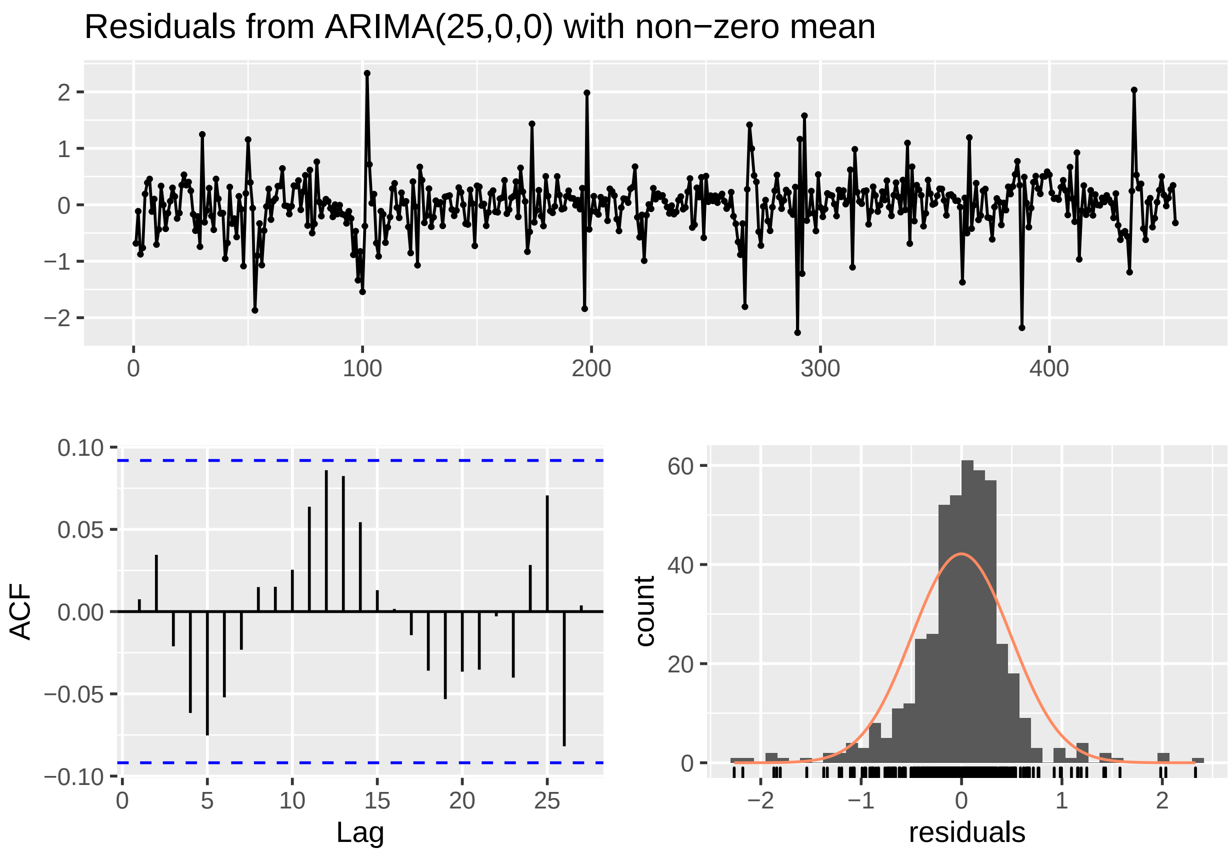
```
AR25 <- arima(train15$count,order=c(25,0,0))

number = nrow(test15)

acf(AR25$residuals)
```



```
pacf(AR25$residuals)
```

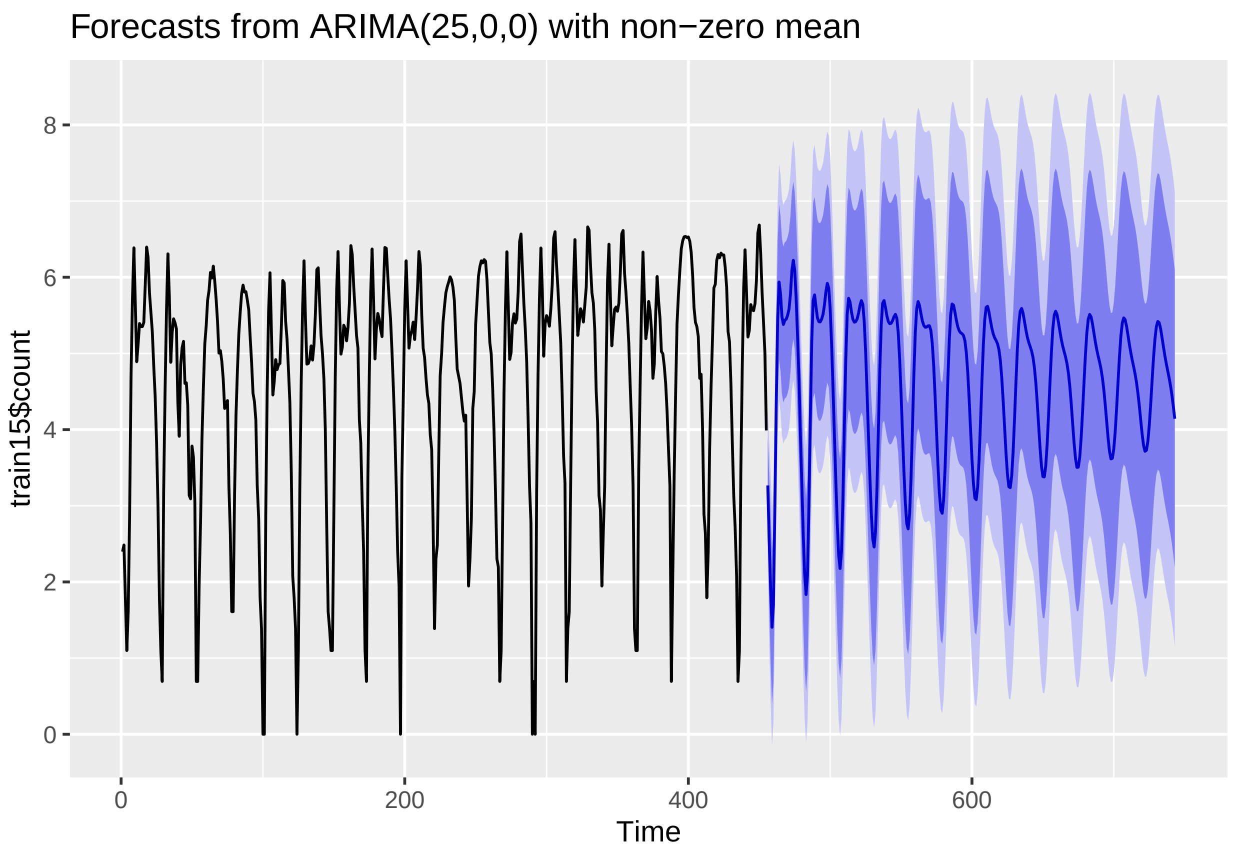


```r
checkresiduals(AR25)
```



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0) with non-zero mean
## Q* = 27.515, df = 3, p-value = 4.592e-06
##
## Model df: 26.   Total lags used: 29
```

```
fcst <- forecast(AR25, h=number)

autoplot(fcst)
```



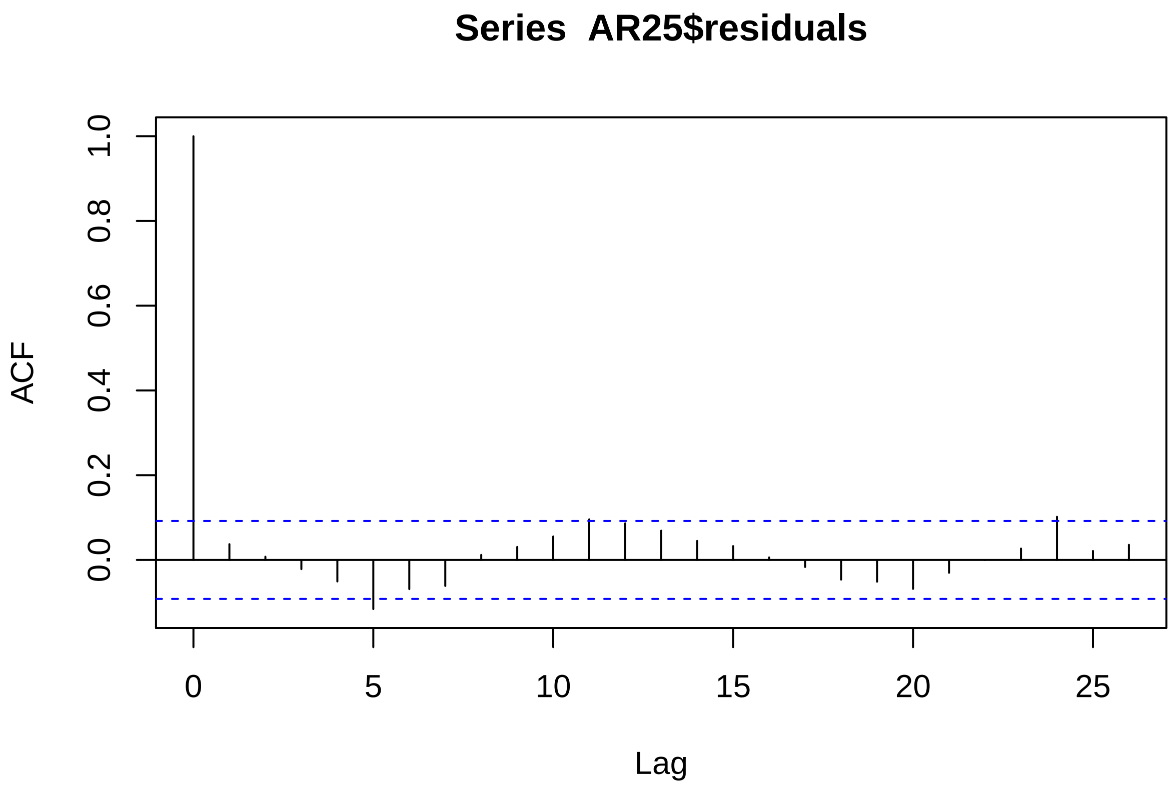
```
# point estimate (mean)
test15$count <- fcst$mean


RMSLE(y_pred = fcst$fitted, y_true = train15$count)
```

```
## [1] 0.1794584
```

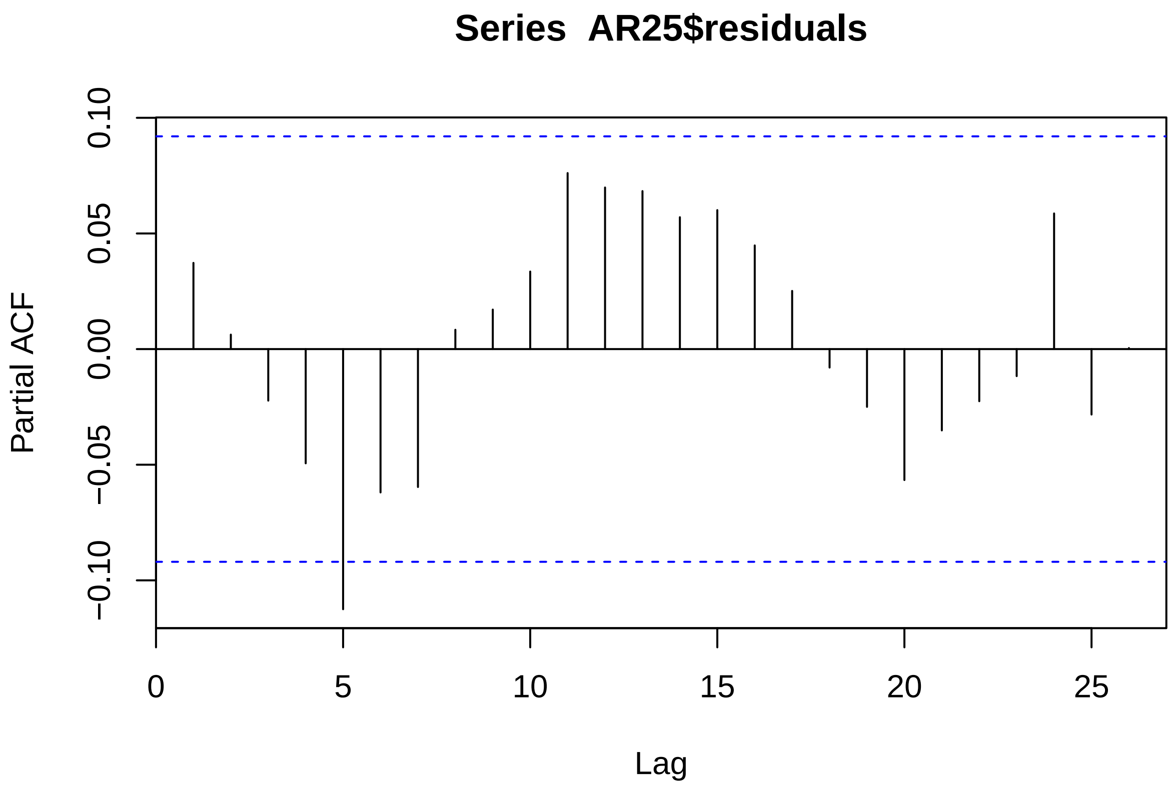
April

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

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


### Log the response variable
train16$count = log(train16$count)

# head(train16)
# head(test16)
```

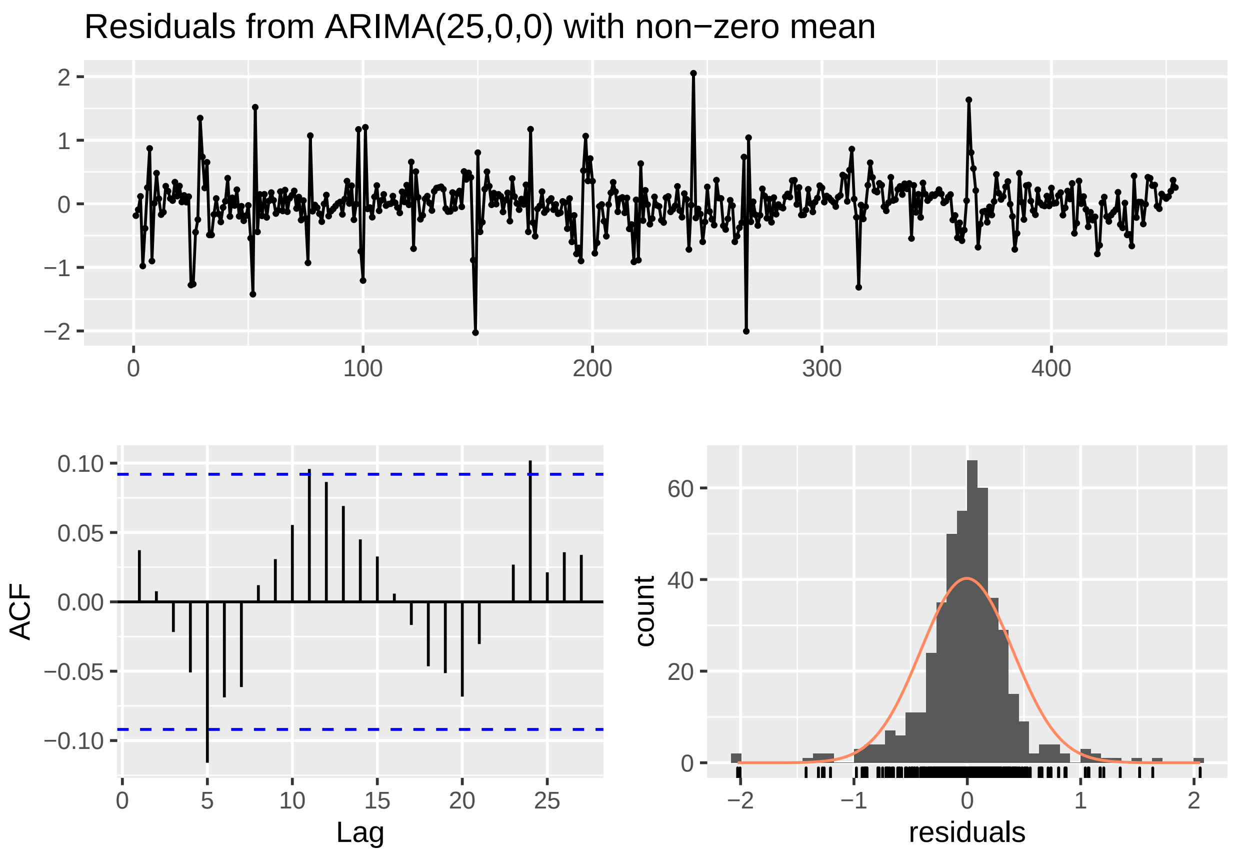
```
AR25 <- arima(train16$count,order=c(25,0,0))

number = nrow(test16)

acf(AR25$residuals)
```



```
pacf(AR25$residuals)
```

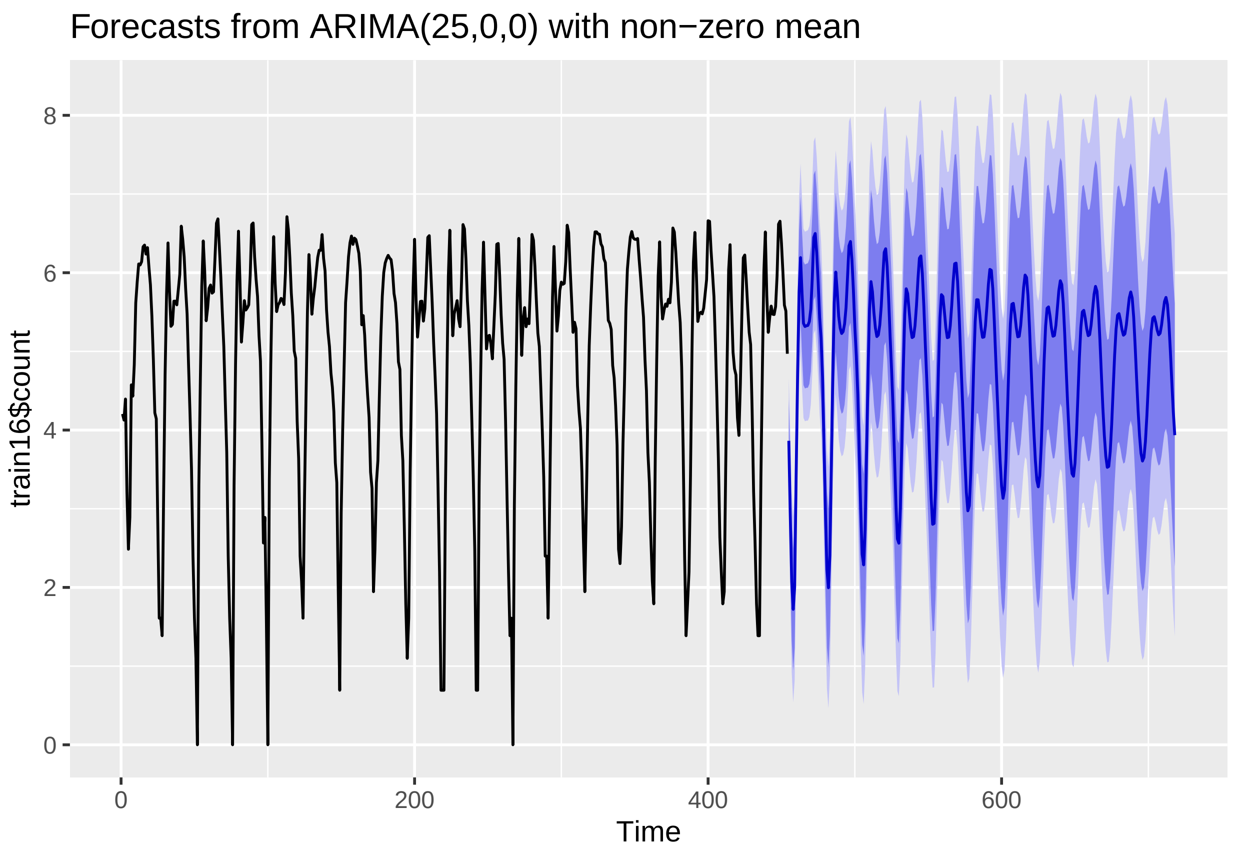


```
checkresiduals(AR25)
```



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0) with non-zero mean
## Q* = 42.662, df = 3, p-value = 2.903e-09
##
## Model df: 26.   Total lags used: 29
```

```
fcst <- forecast(AR25, h=number)

autoplot(fcst)
```



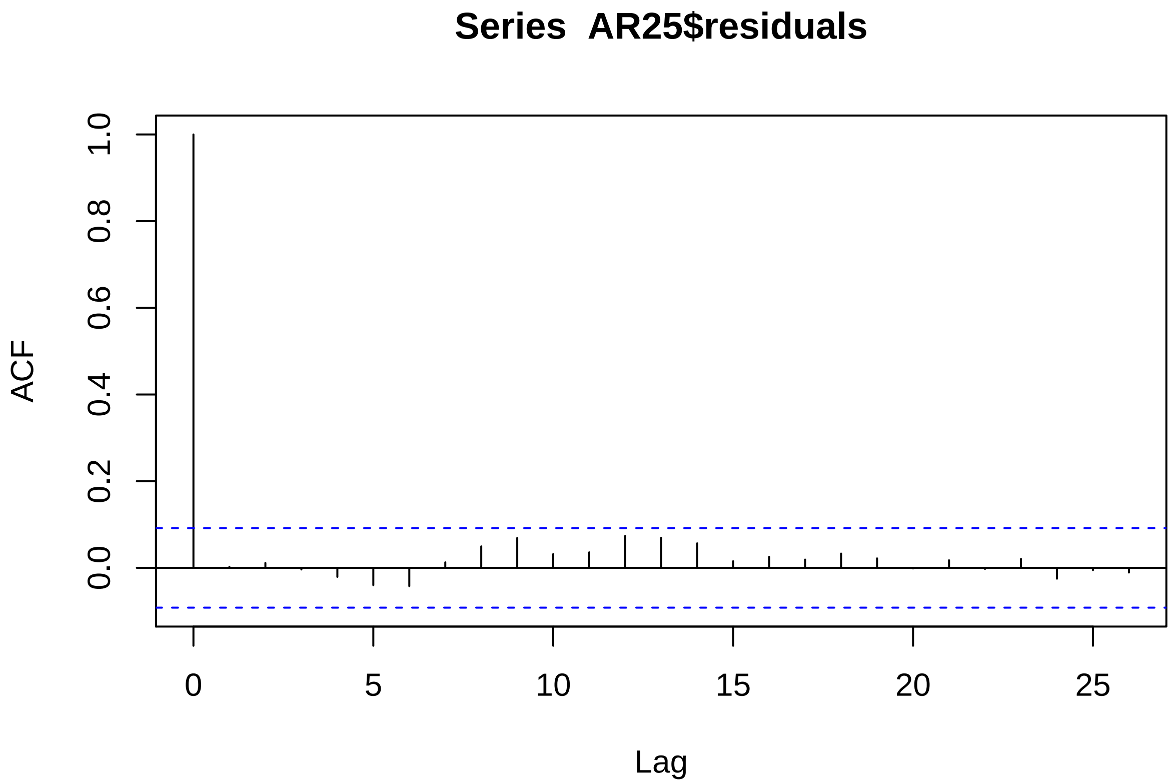
```
# point estimate (mean)
test16$count <- fcst$mean

RMSLE(y_pred = fcst$fitted, y_true = train16$count)
```

```
## [1] 0.1338244
```

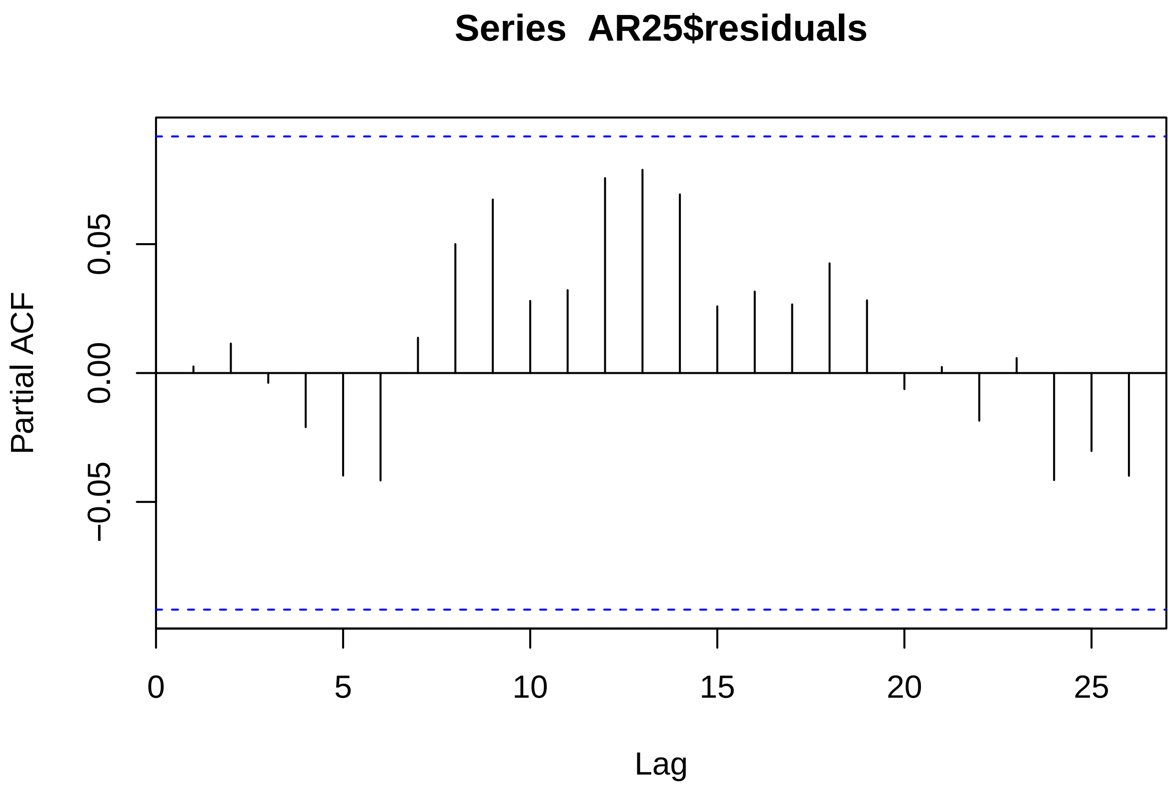
May

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

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


### Log the response variable
train17$count = log(train17$count)

# head(train17)
# head(test17)

AR25 <- arima(train17$count,order=c(25,0,0))
```

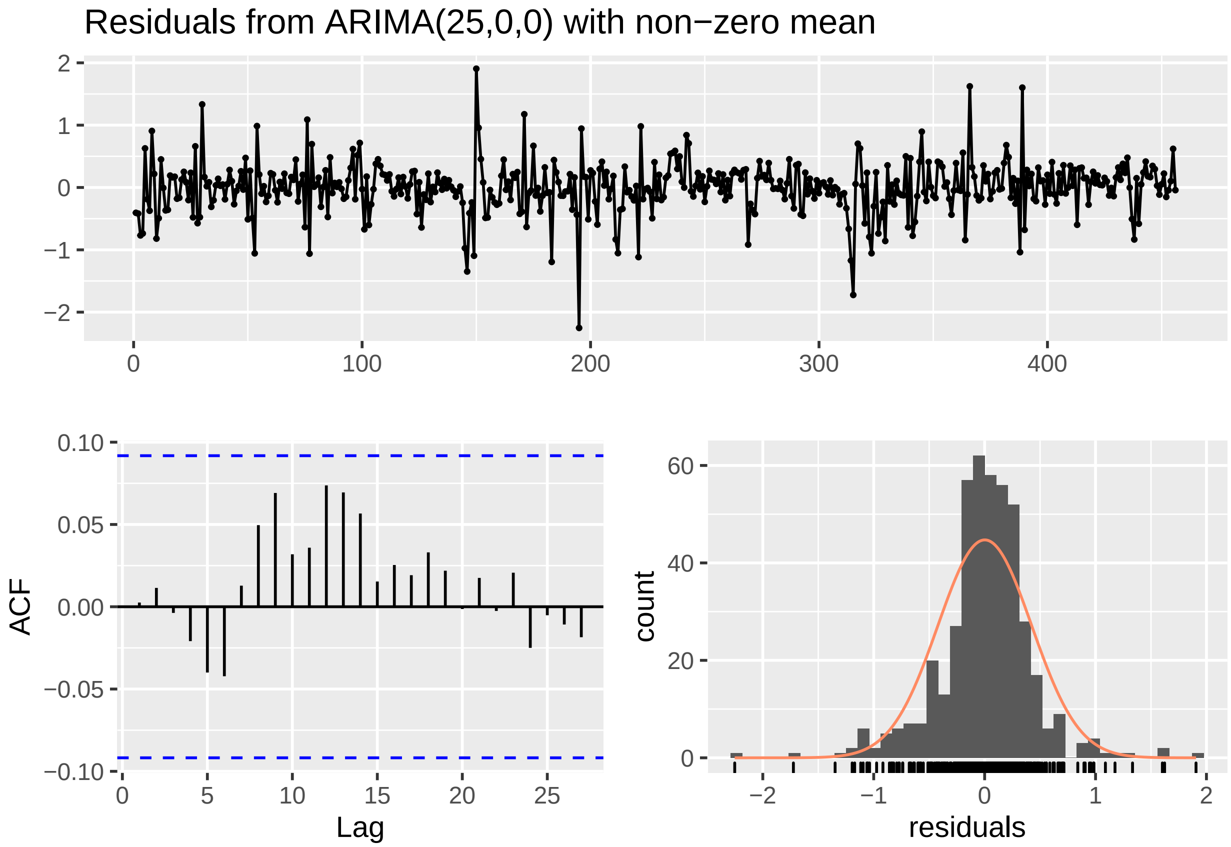
```
number = nrow(test17)

acf(AR25$residuals)
```



```
pacf(AR25$residuals)
```

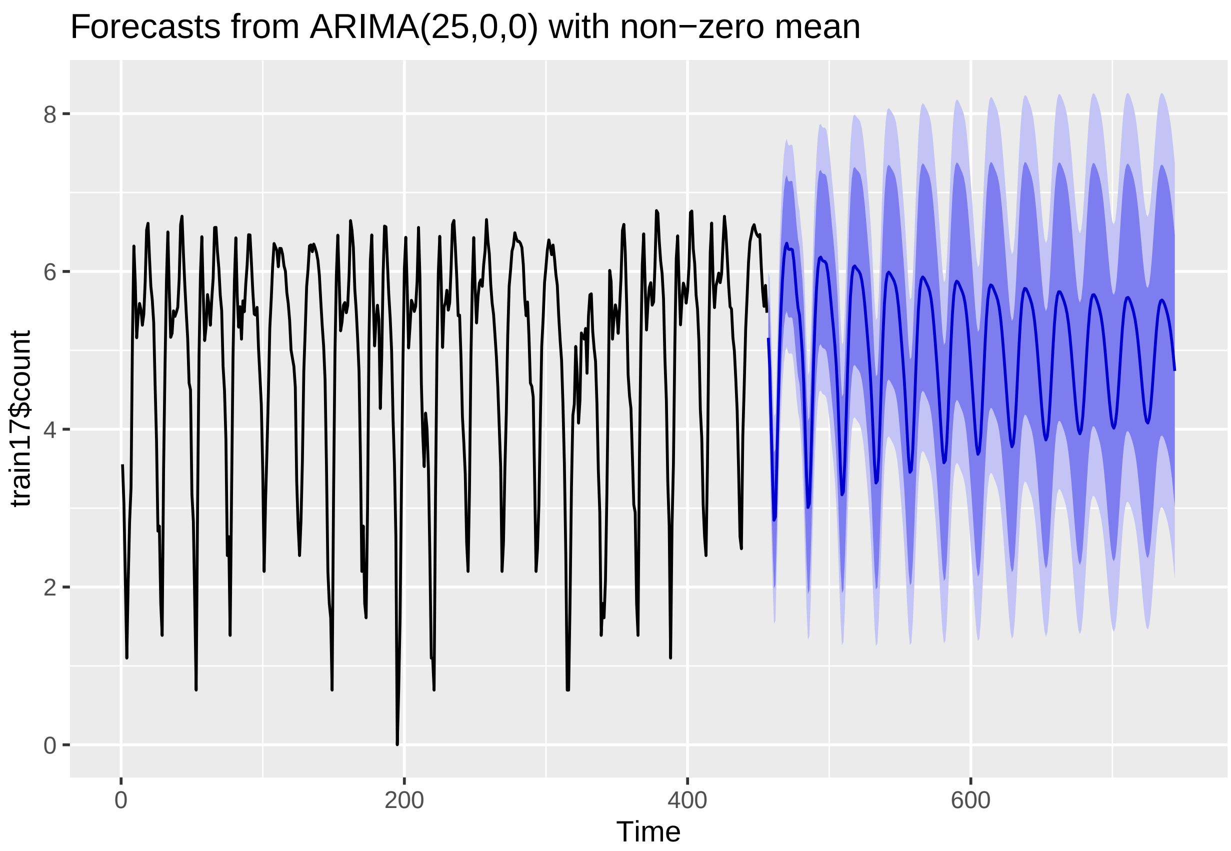


```
checkresiduals(AR25)
```



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0) with non-zero mean
## Q* = 16.911, df = 3, p-value = 0.0007372
##
## Model df: 26.   Total lags used: 29
```

```
fcst <- forecast(AR25, h=number)

autoplot(fcst)
```



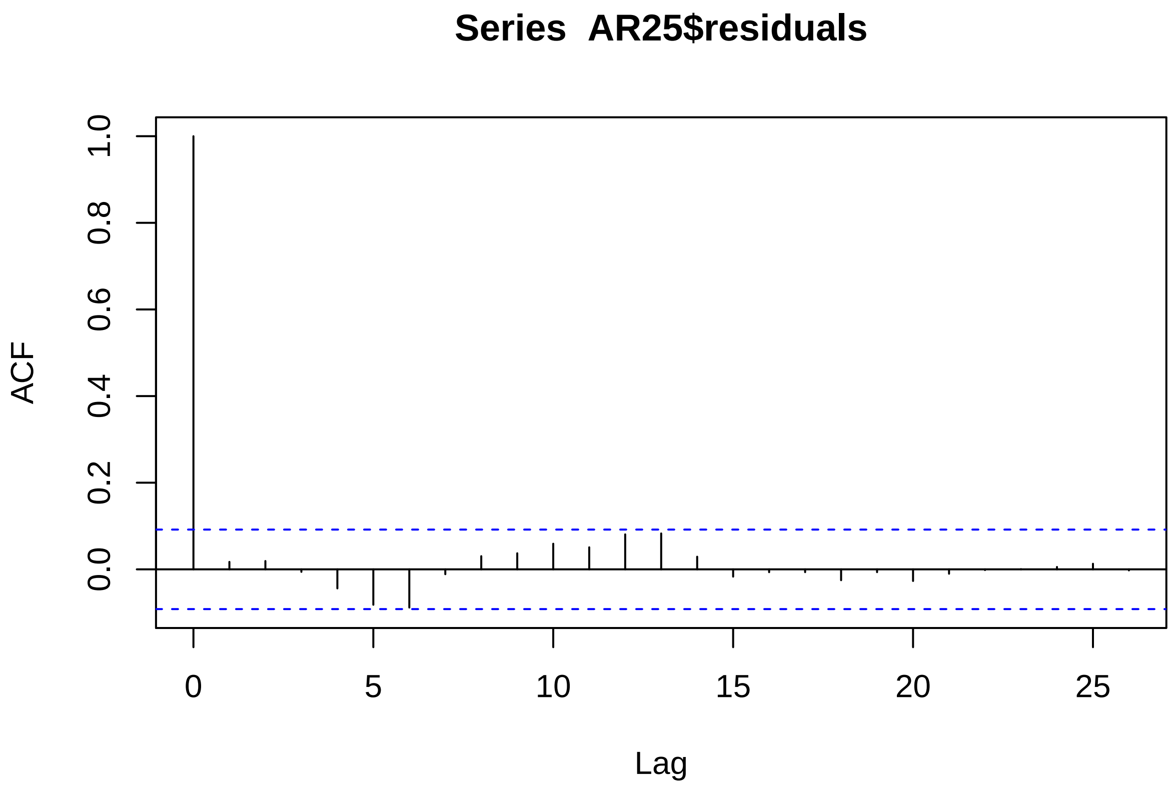
```r
# point estimate (mean)
test17$count <- fcst$mean


RMSLE(y_pred = fcst$fitted, y_true = train17$count)
```

```
## [1] 0.1279052
```

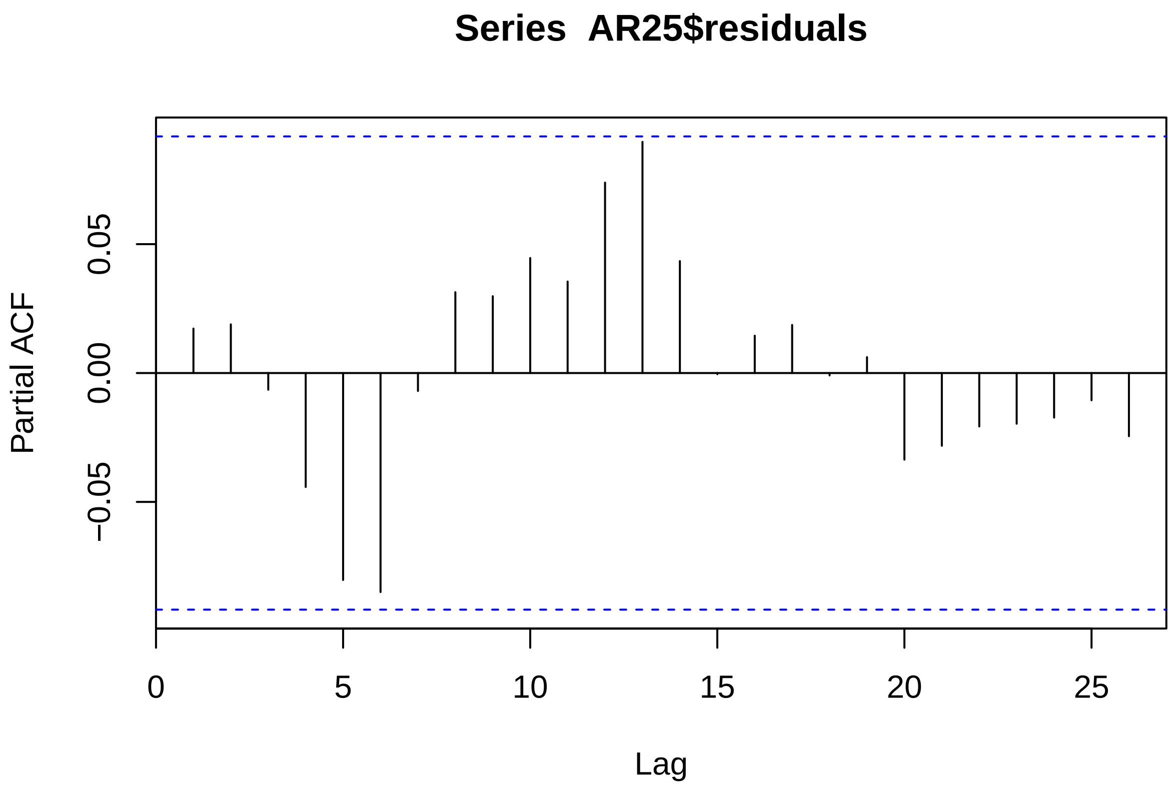
June

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

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


### Log the response variable
train18$count = log(train18$count)

# head(train18)
# head(test18)
```

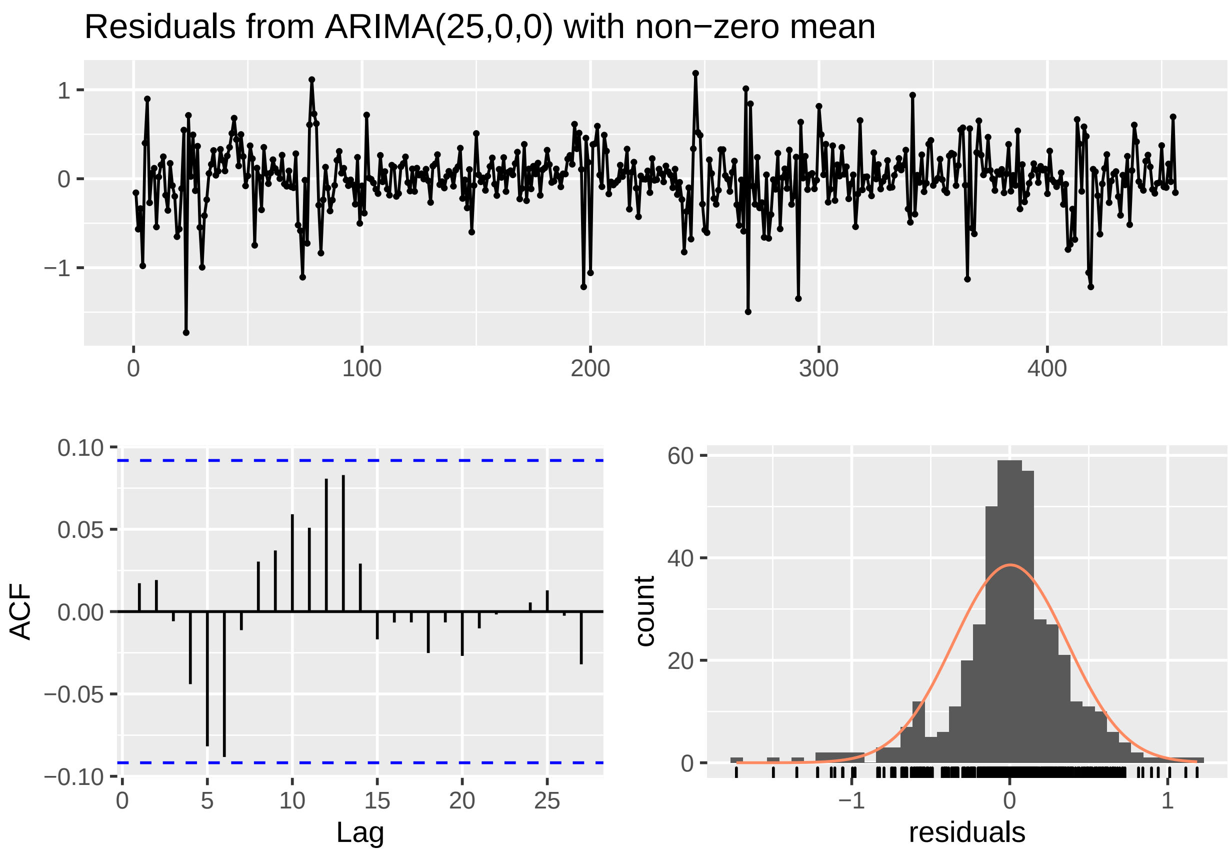
```
AR25 <- arima(train18$count,order=c(25,0,0))

number = nrow(test18)

acf(AR25$residuals)
```



```
pacf(AR25$residuals)
```

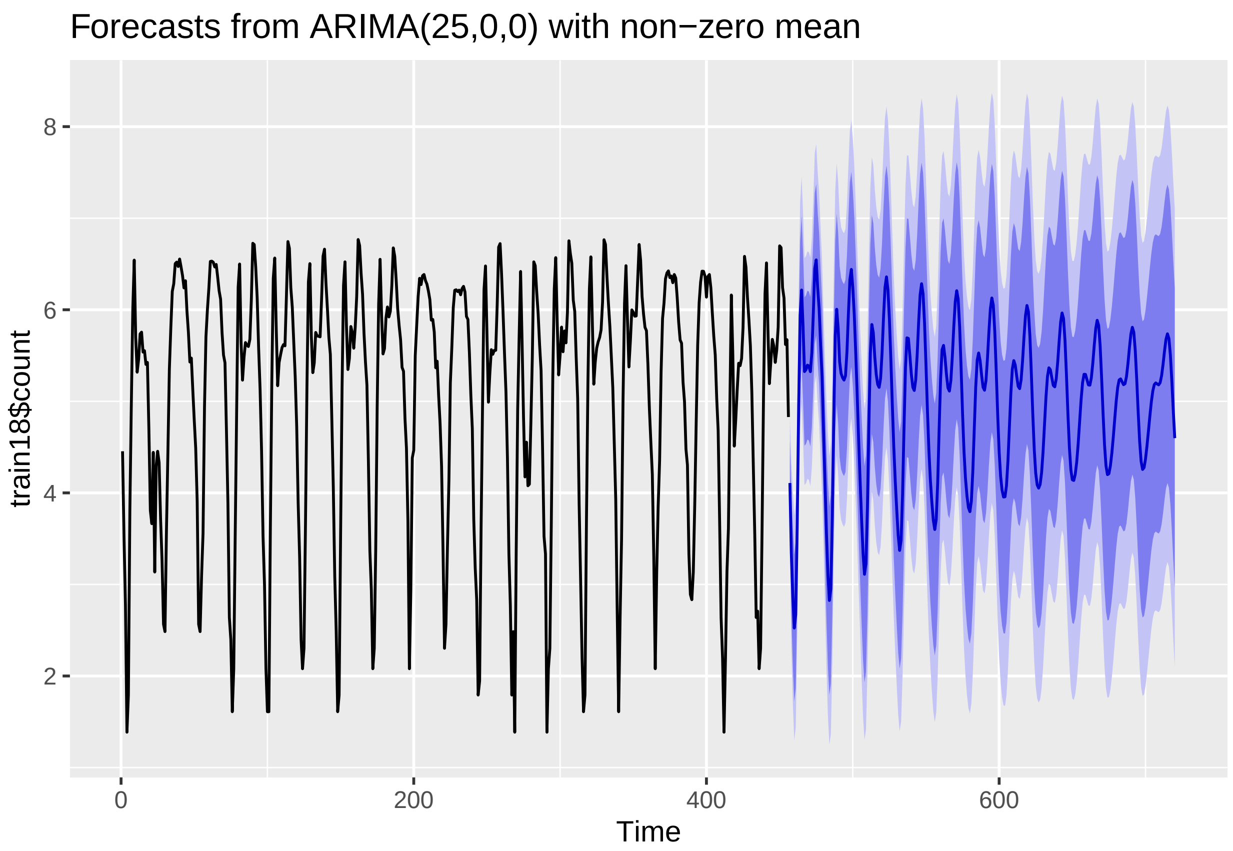


```
checkresiduals(AR25)
```



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0) with non-zero mean
## Q* = 20.769, df = 3, p-value = 0.0001176
##
## Model df: 26.   Total lags used: 29
```

```
fcst <- forecast(AR25, h=number)

autoplot(fcst)
```



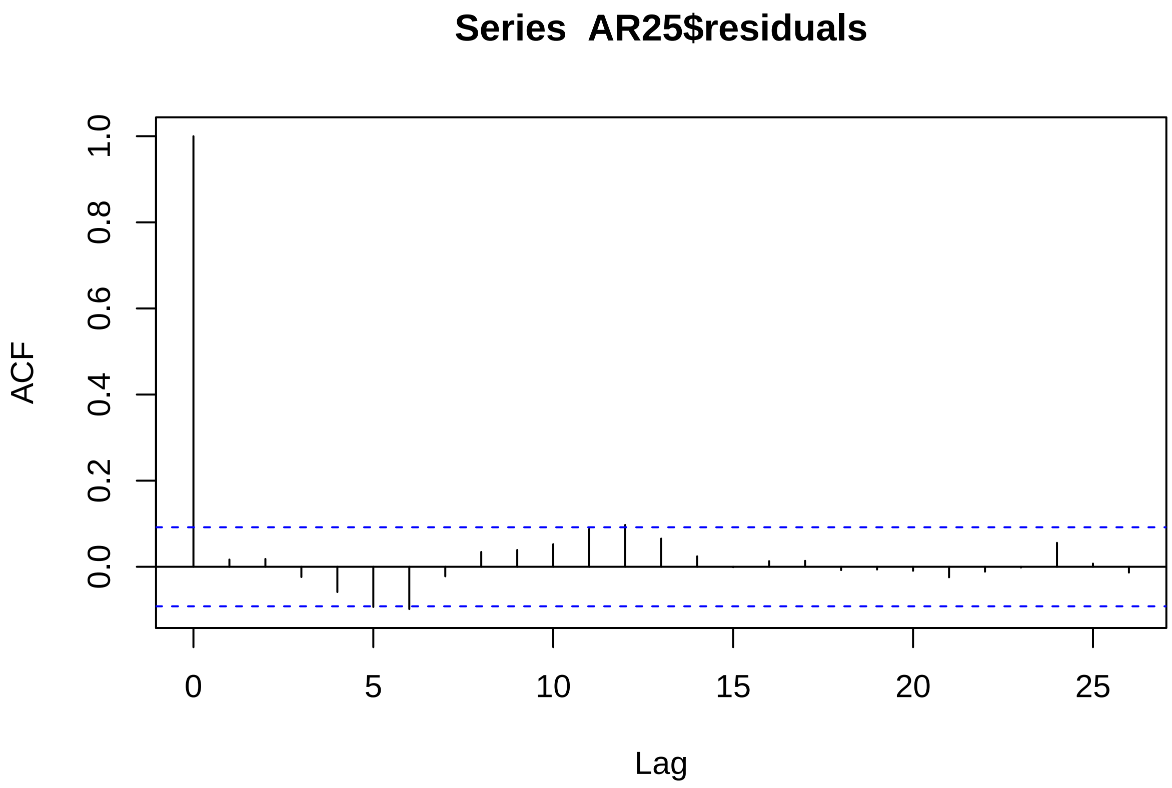
```
# point estimate (mean)
test18$count <- fcst$mean


RMSLE(y_pred = fcst$fitted, y_true = train18$count)
```

```
## [1] 0.08421316
```

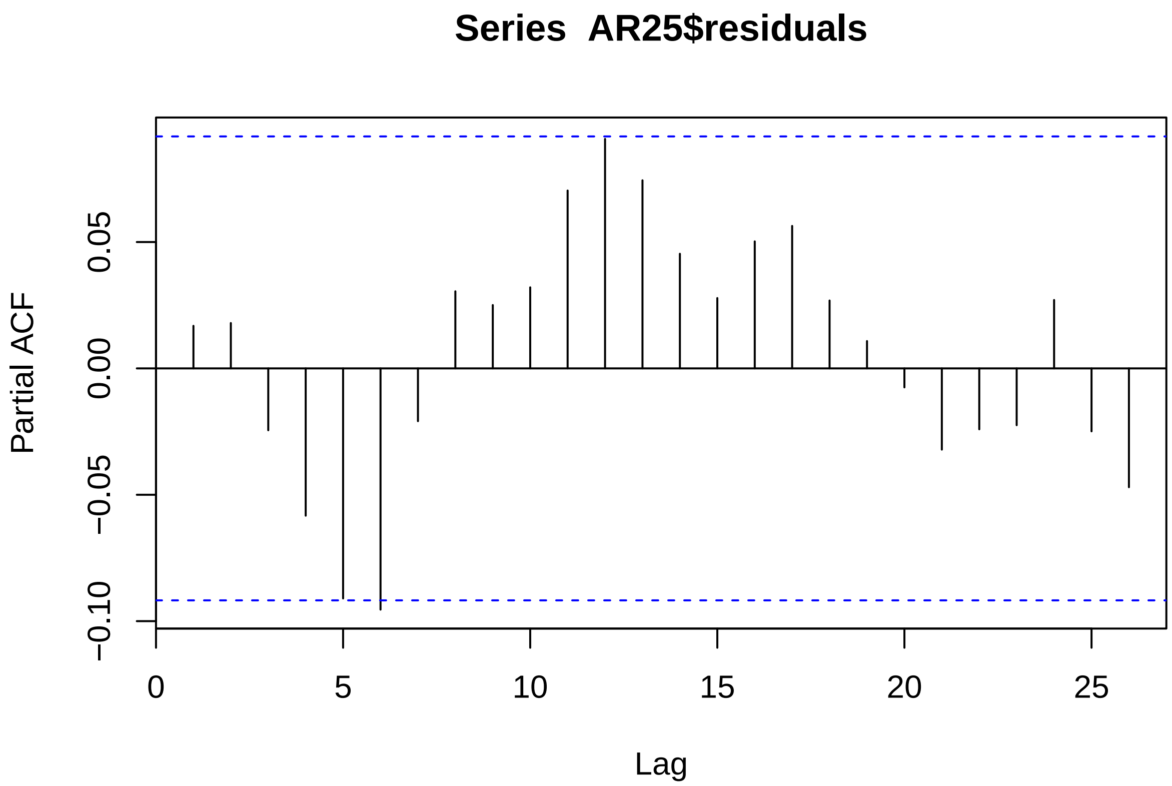
July

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

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


### Log the response variable
train19$count = log(train19$count)

# head(train19)
# head(test19)
```

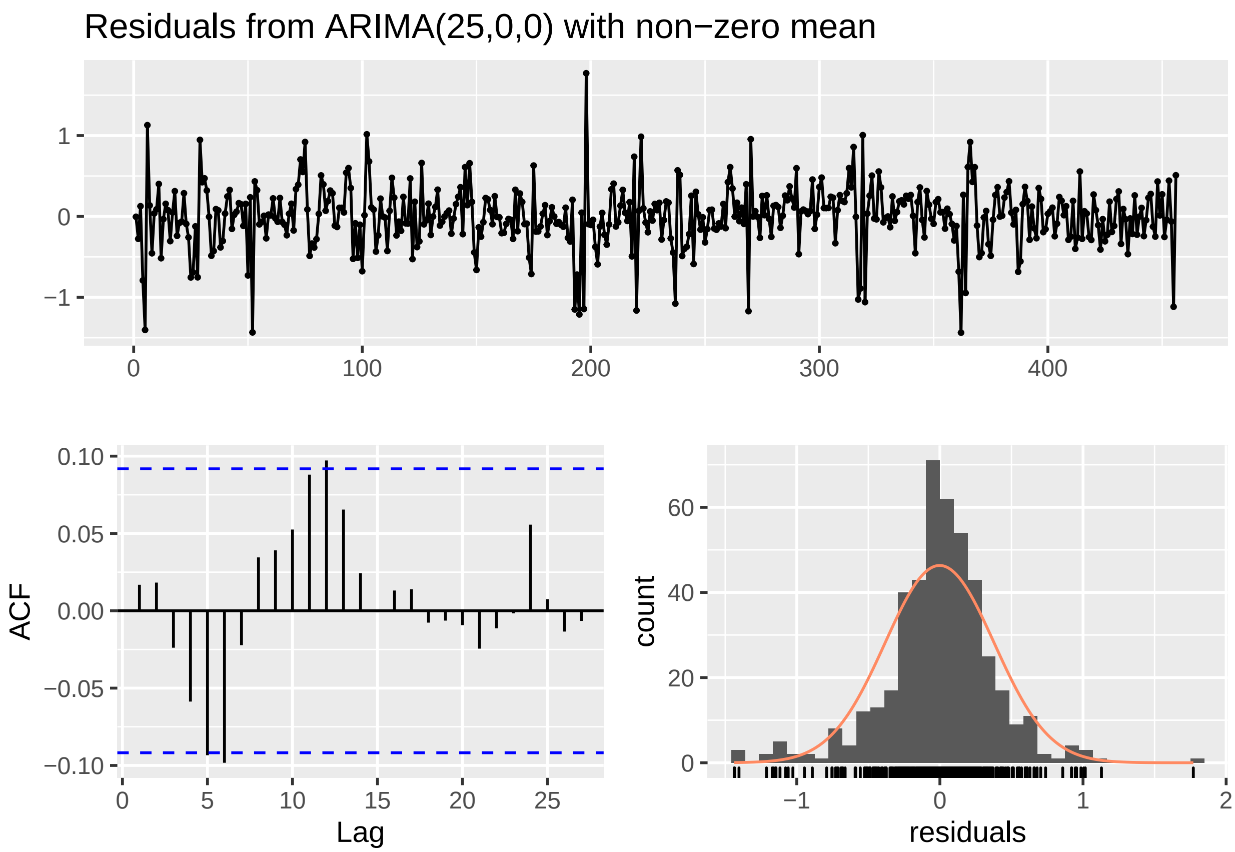
```
AR25 <- arima(train19$count,order=c(25,0,0))

number = nrow(test19)

acf(AR25$residuals)
```



```
pacf(AR25$residuals)
```

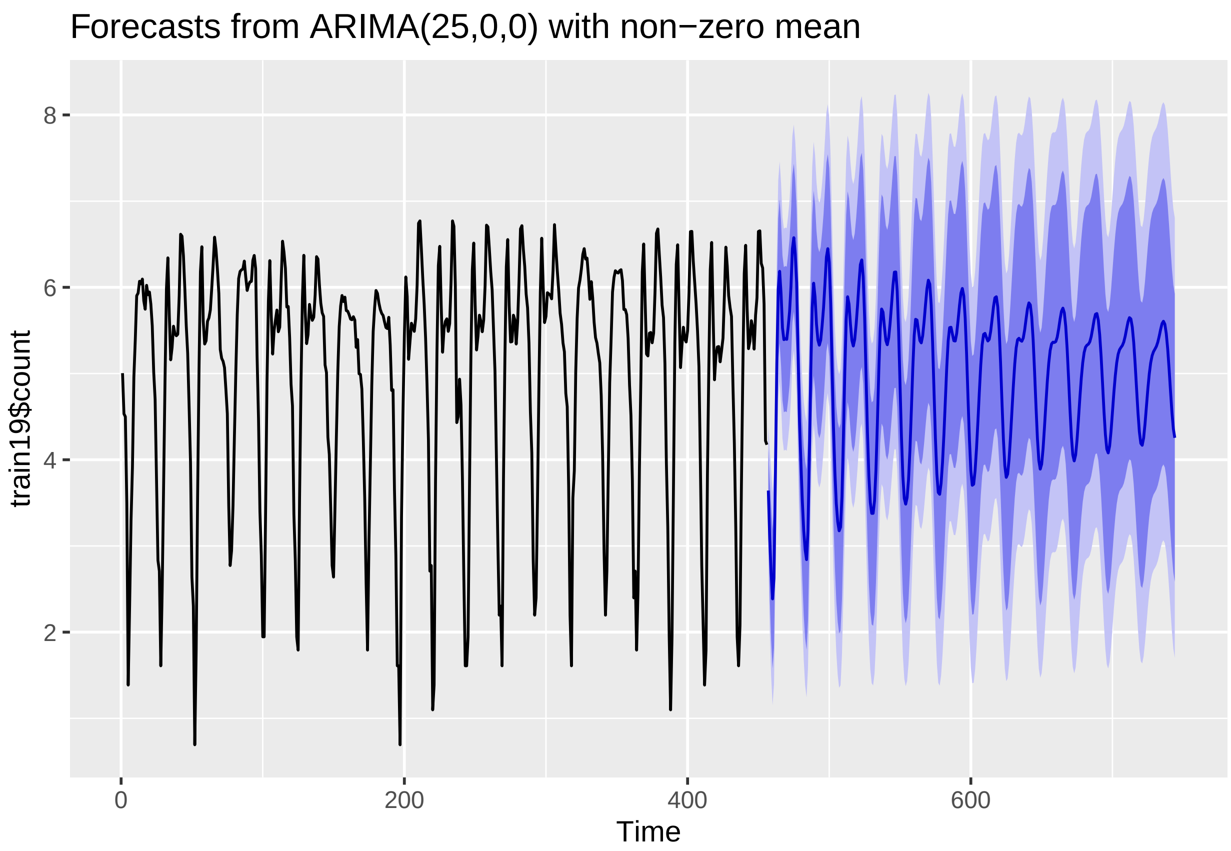


```
checkresiduals(AR25)
```



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0) with non-zero mean
## Q* = 29.505, df = 3, p-value = 1.754e-06
##
## Model df: 26.   Total lags used: 29
```

```
fcst <- forecast(AR25, h=number)

autoplot(fcst)
```



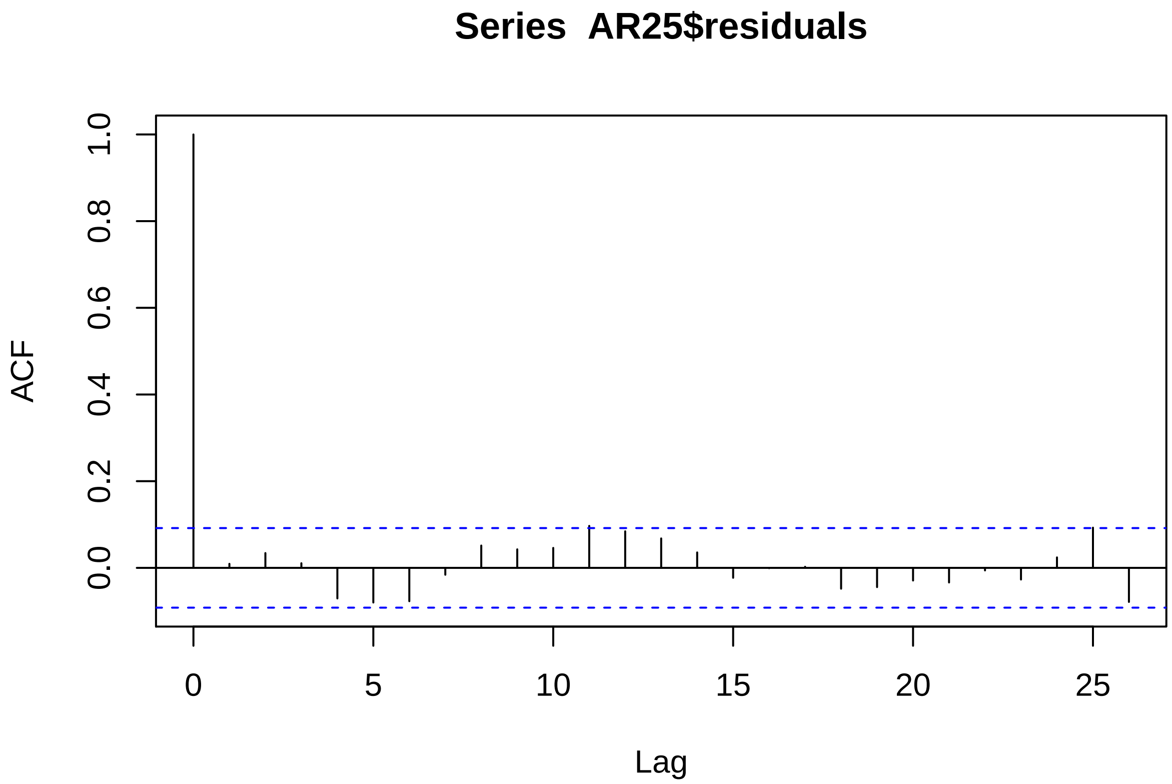
```
# point estimate (mean)
test19$count <- fcst$mean

RMSLE(y_pred = fcst$fitted, y_true = train19$count)
```

```
## [1] 0.09721031
```

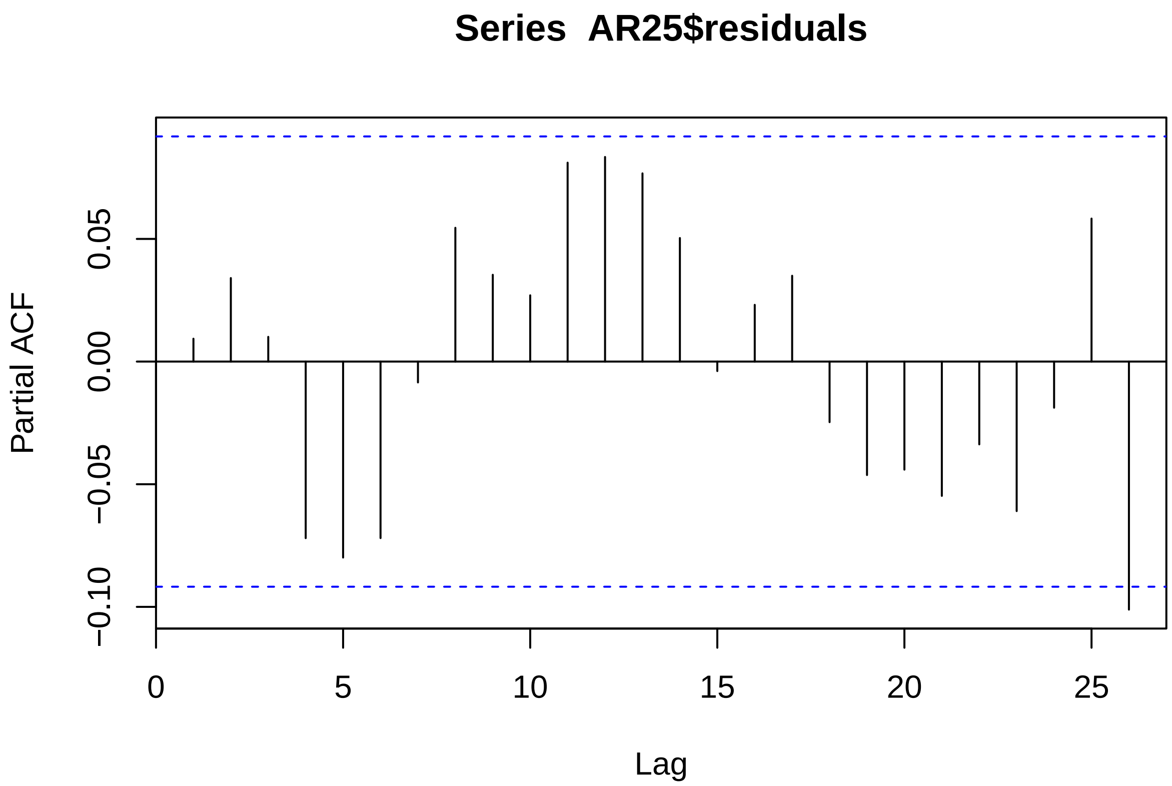
August

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

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


### Log the response variable
train20$count = log(train20$count)

# head(train20)
# head(test20)

AR25 <- arima(train20$count,order=c(25,0,0))
```

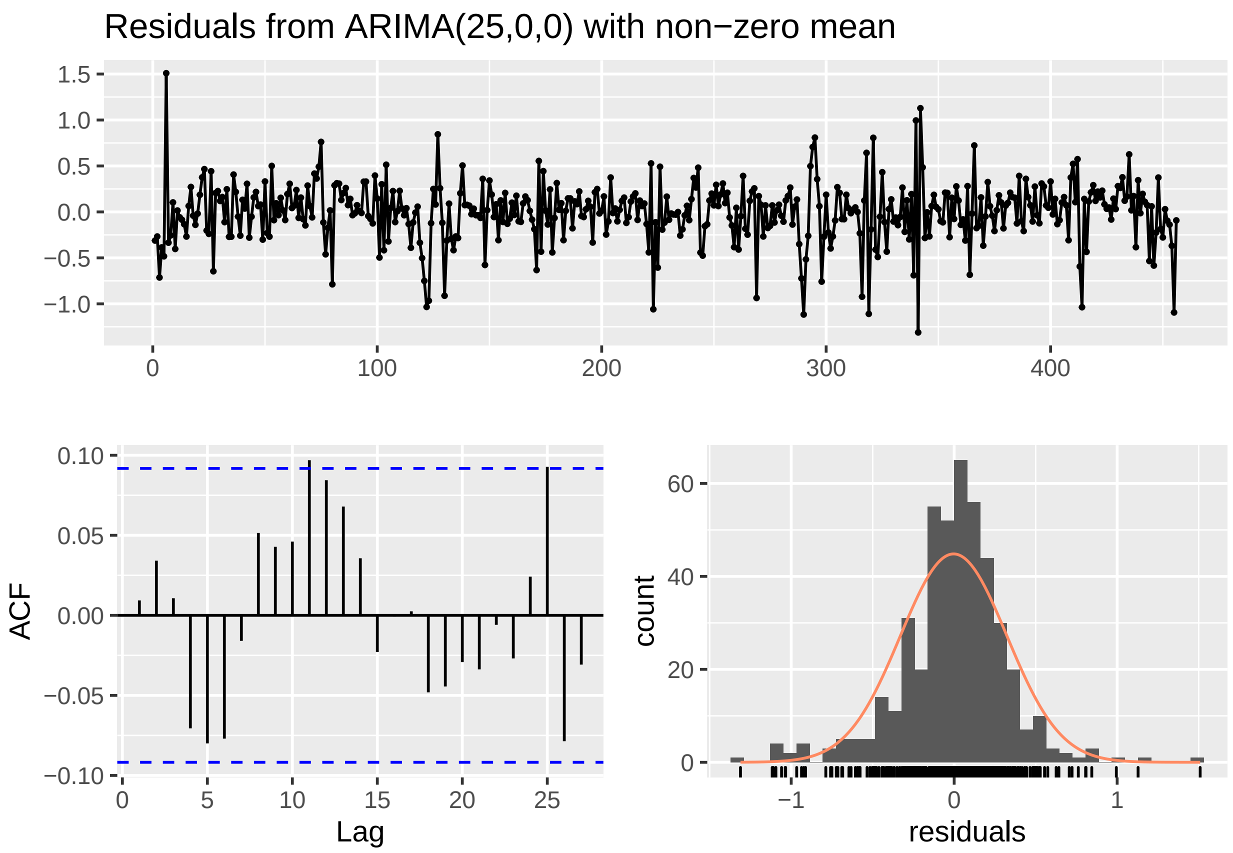
```
number = nrow(test20)

acf(AR25$residuals)
```



```
pacf(AR25$residuals)
```

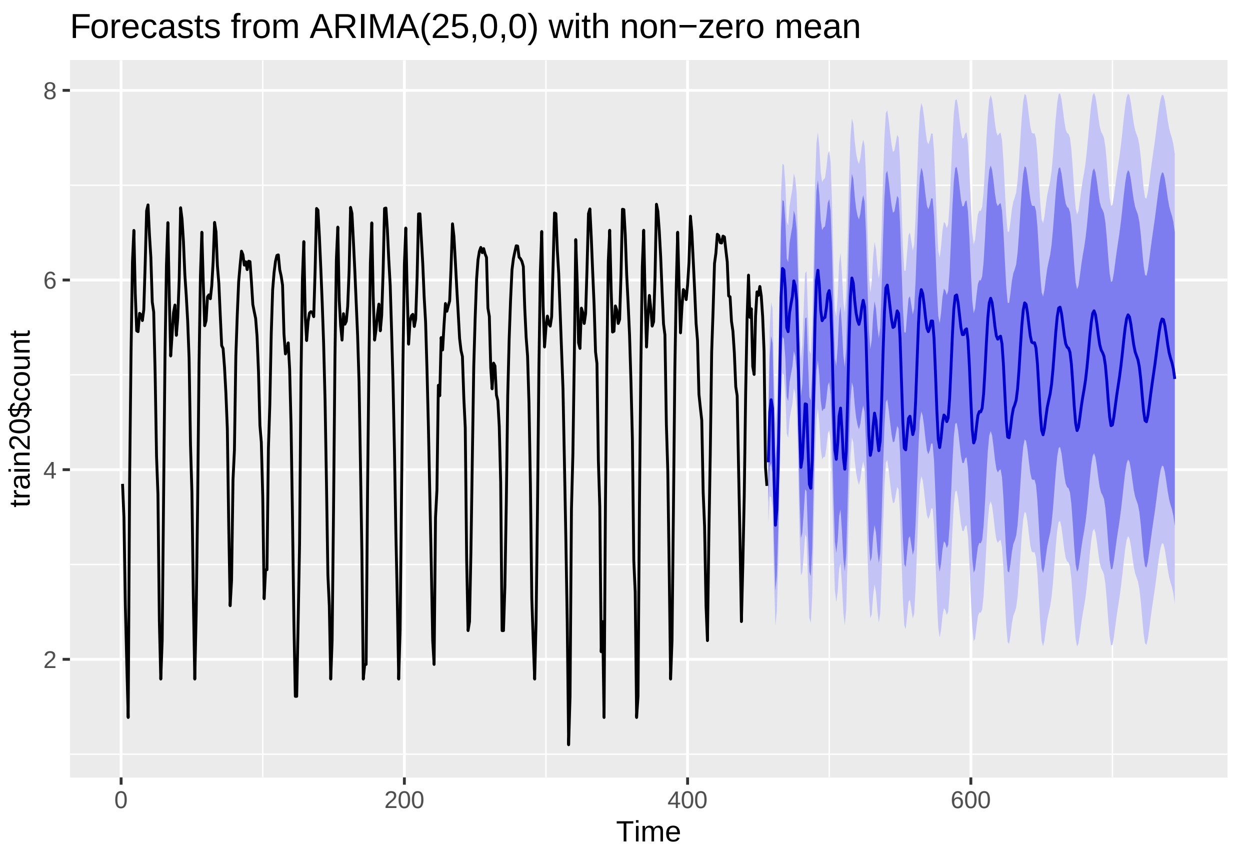


```
checkresiduals(AR25)
```



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0) with non-zero mean
## Q* = 35.287, df = 3, p-value = 1.06e-07
##
## Model df: 26.   Total lags used: 29
```

```
fcst <- forecast(AR25, h=number)

autoplot(fcst)
```



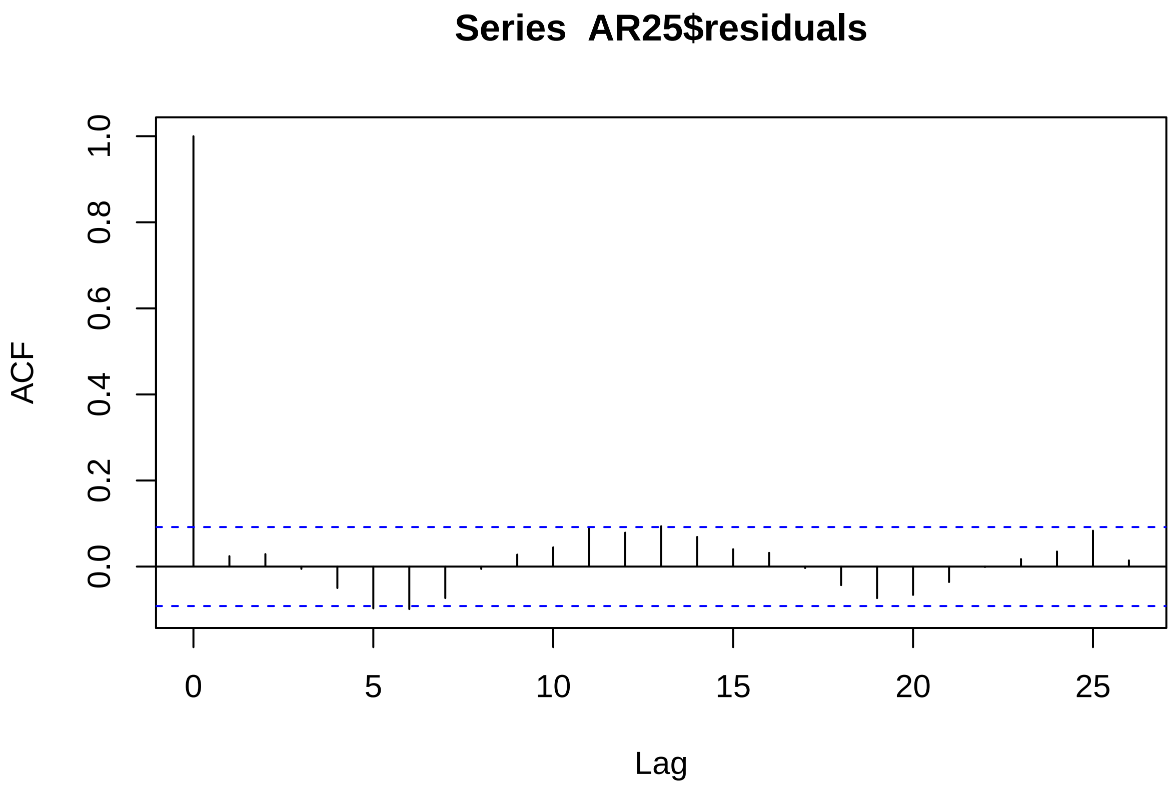
```
# point estimate (mean)
test20$count <- fcst$mean


RMSLE(y_pred = fcst$fitted, y_true = train20$count)
```

```
## [1] 0.07543401
```

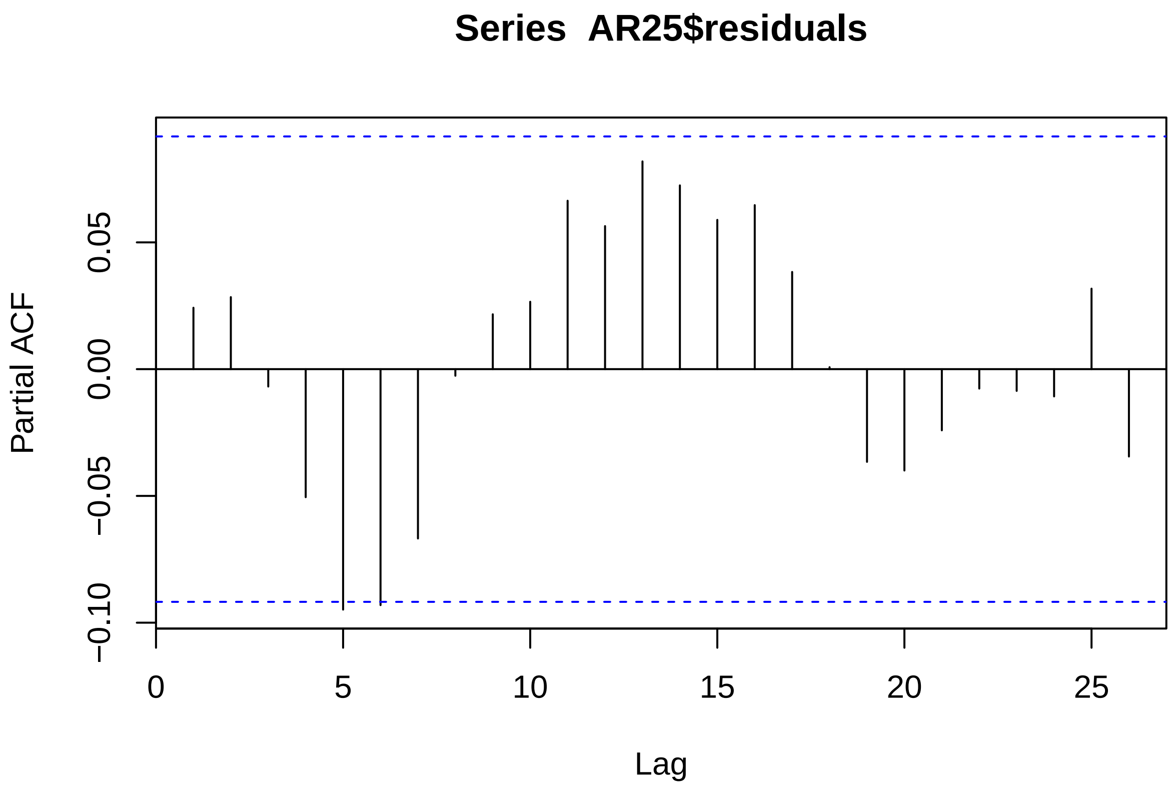
September

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

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


### Log the response variable
train21$count = log(train21$count)

# head(train21)
# head(test21)
```

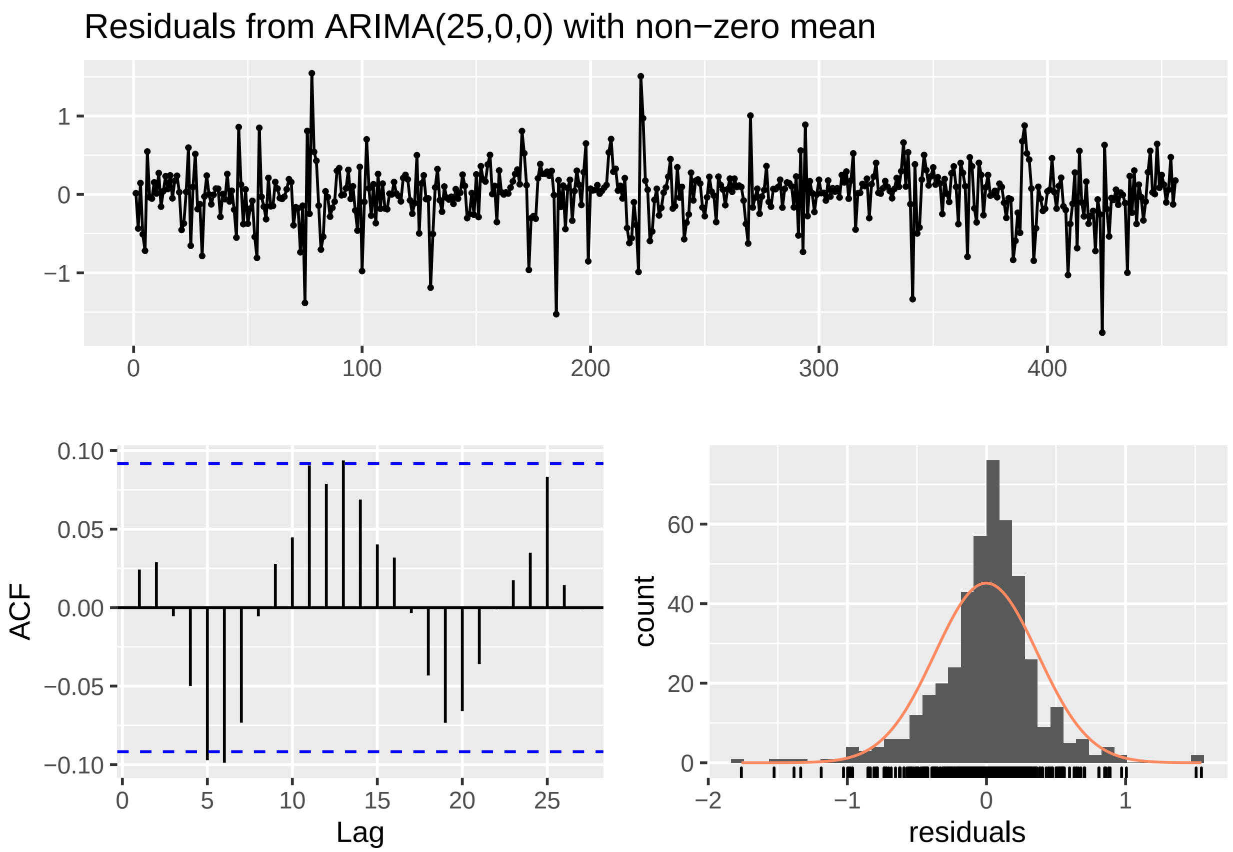
```
AR25 <- arima(train21$count,order=c(25,0,0))

number = nrow(test21)

acf(AR25$residuals)
```



```
pacf(AR25$residuals)
```

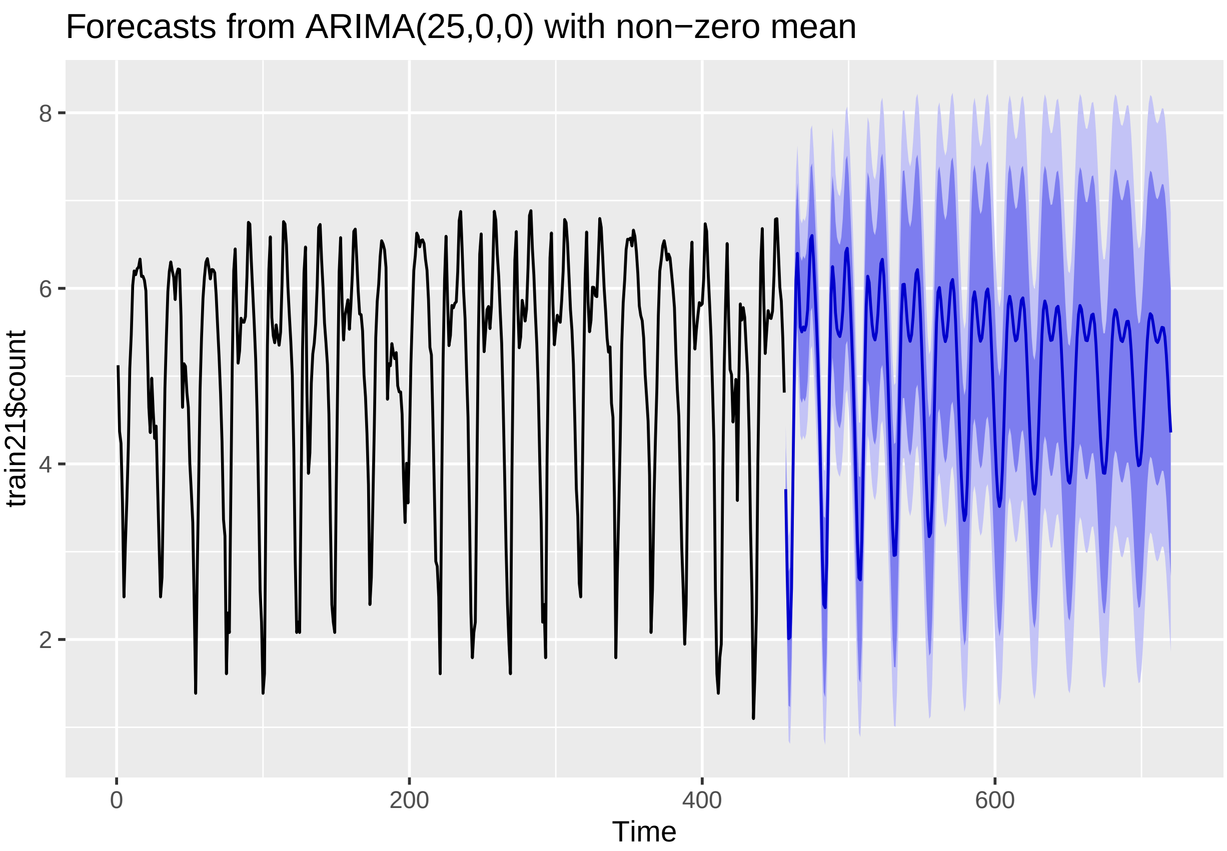


```r
checkresiduals(AR25)
```



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0) with non-zero mean
## Q* = 39.388, df = 3, p-value = 1.437e-08
##
## Model df: 26.   Total lags used: 29
```

```
fcst <- forecast(AR25, h=number)

autoplot(fcst)
```



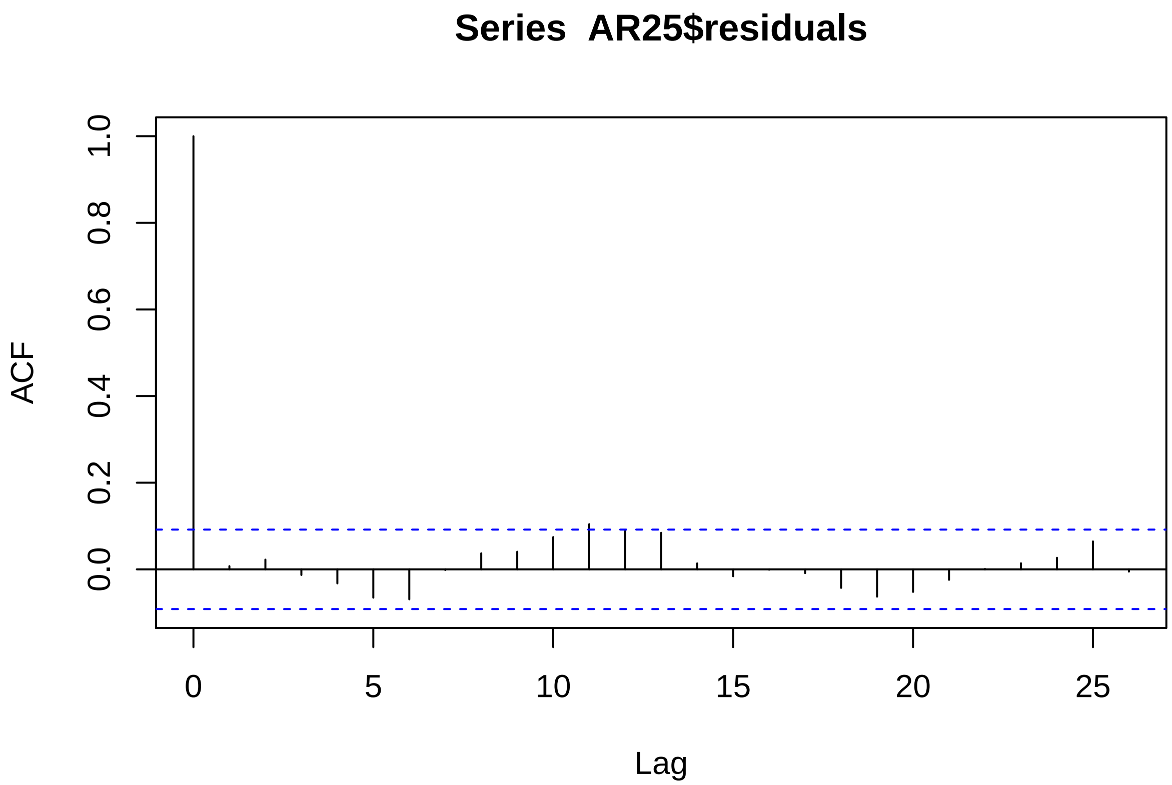
```
# point estimate (mean)
test21$count <- fcst$mean


RMSLE(y_pred = fcst$fitted, y_true = train21$count)
```

```
## [1] 0.08502199
```

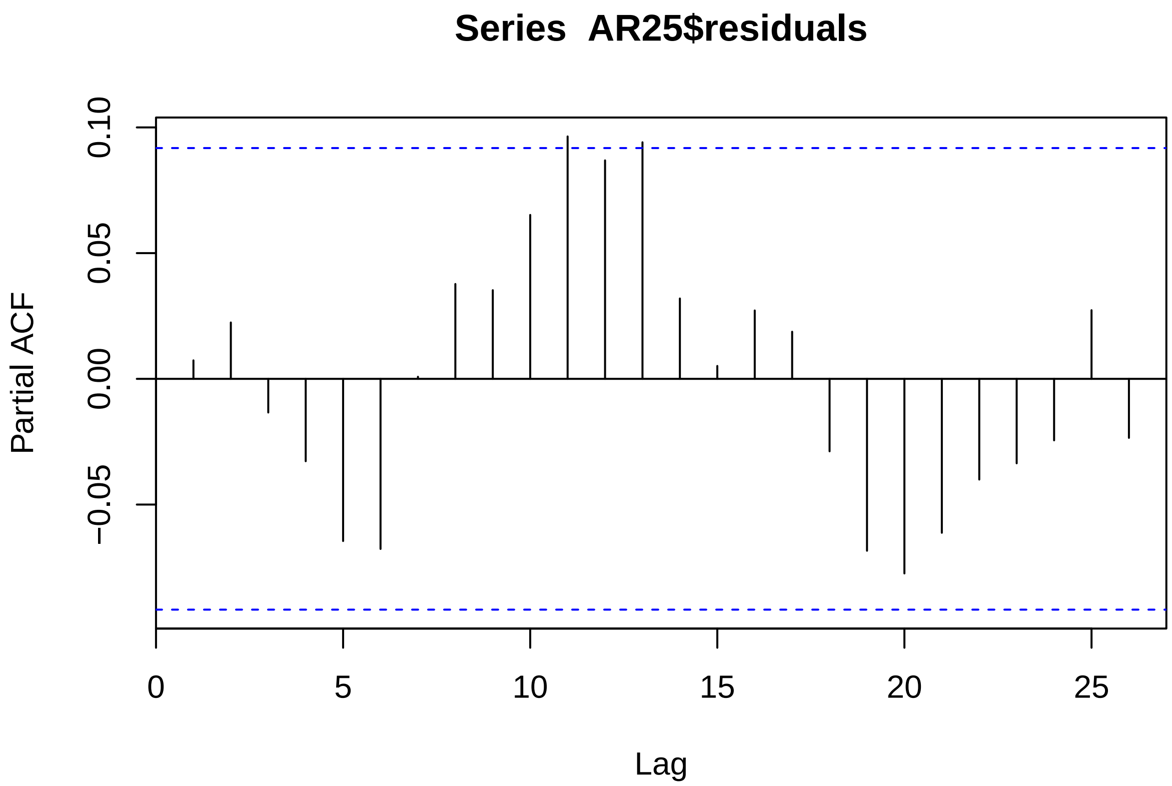
October

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

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


### Log the response variable
train22$count = log(train22$count)

# head(train22)
# head(test22)
```

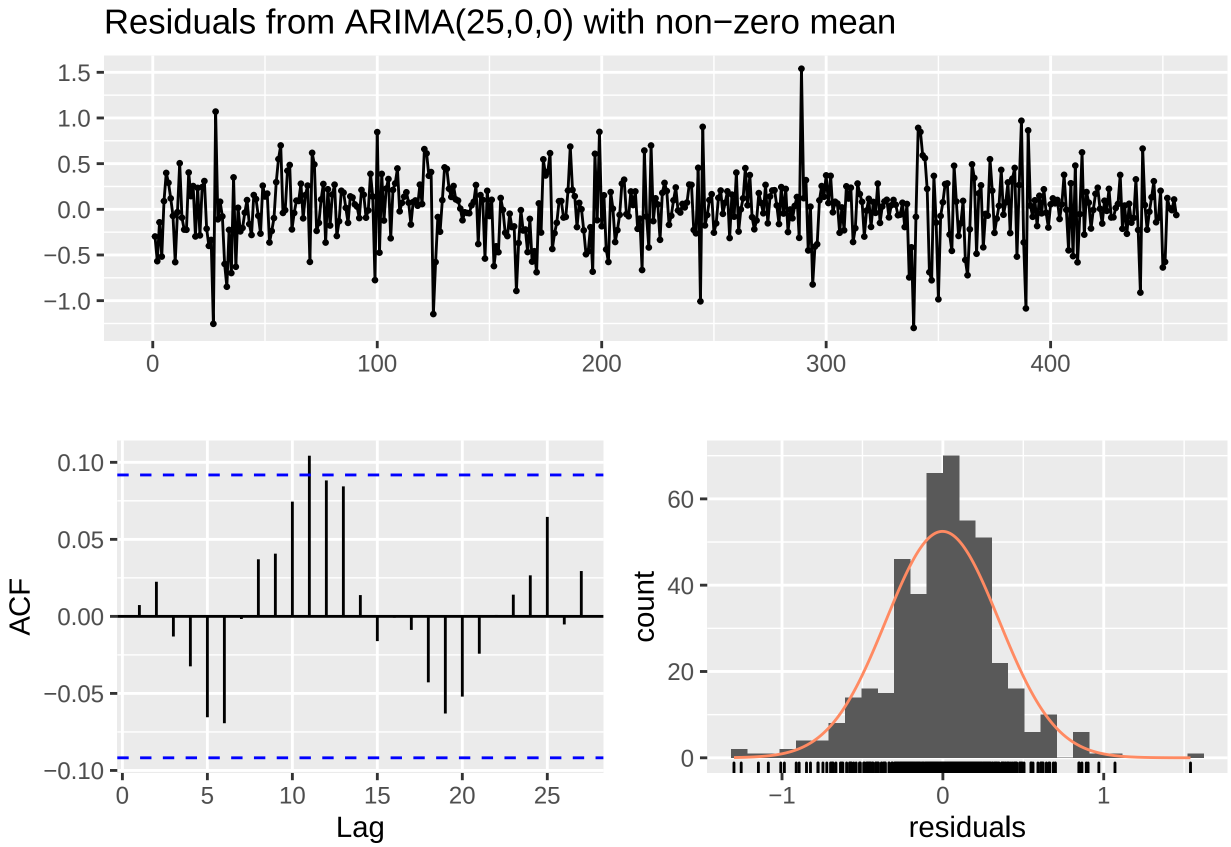
```r
AR25 <- arima(train22$count,order=c(25,0,0))

number = nrow(test22)

acf(AR25$residuals)
```



```r
pacf(AR25$residuals)
```

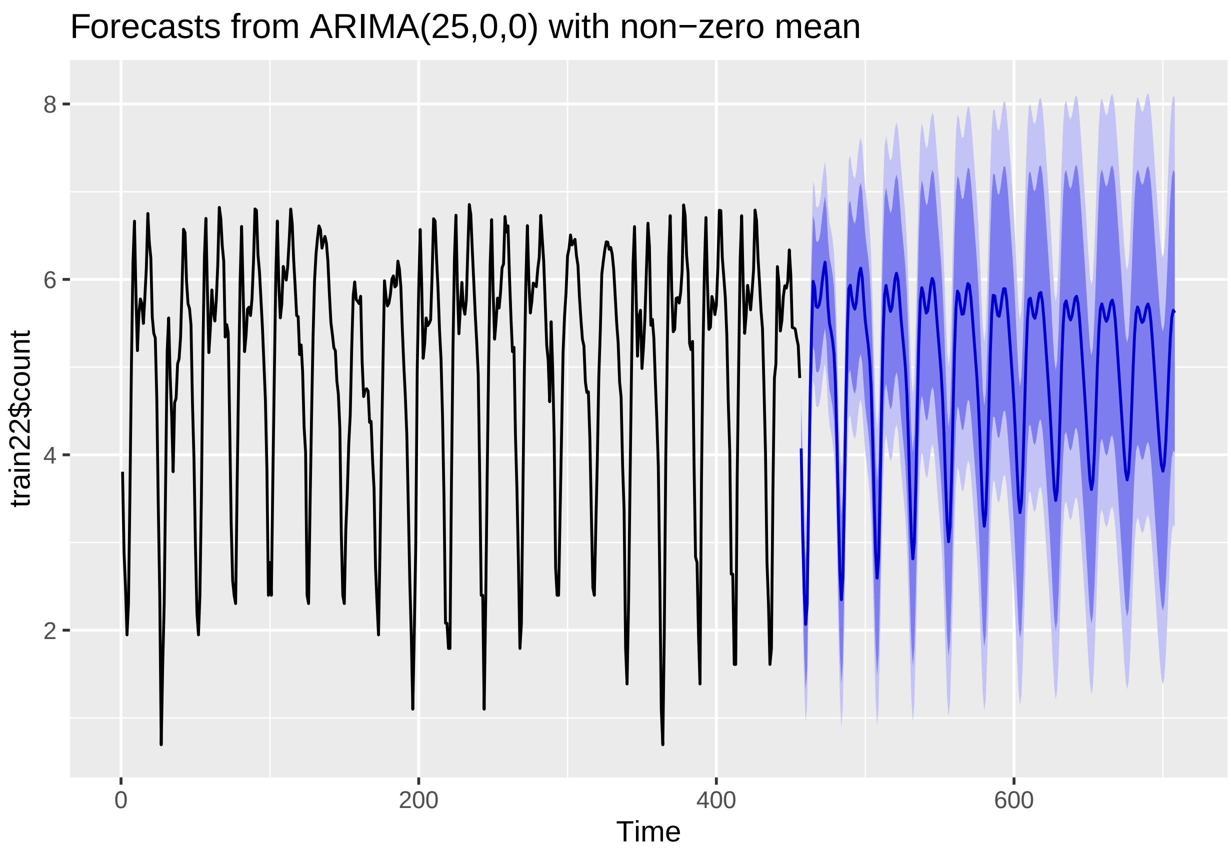


```r
checkresiduals(AR25)
```



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0) with non-zero mean
## Q* = 31.328, df = 3, p-value = 7.251e-07
##
## Model df: 26.   Total lags used: 29
```

```
fcst <- forecast(AR25, h=number)

autoplot(fcst)
```



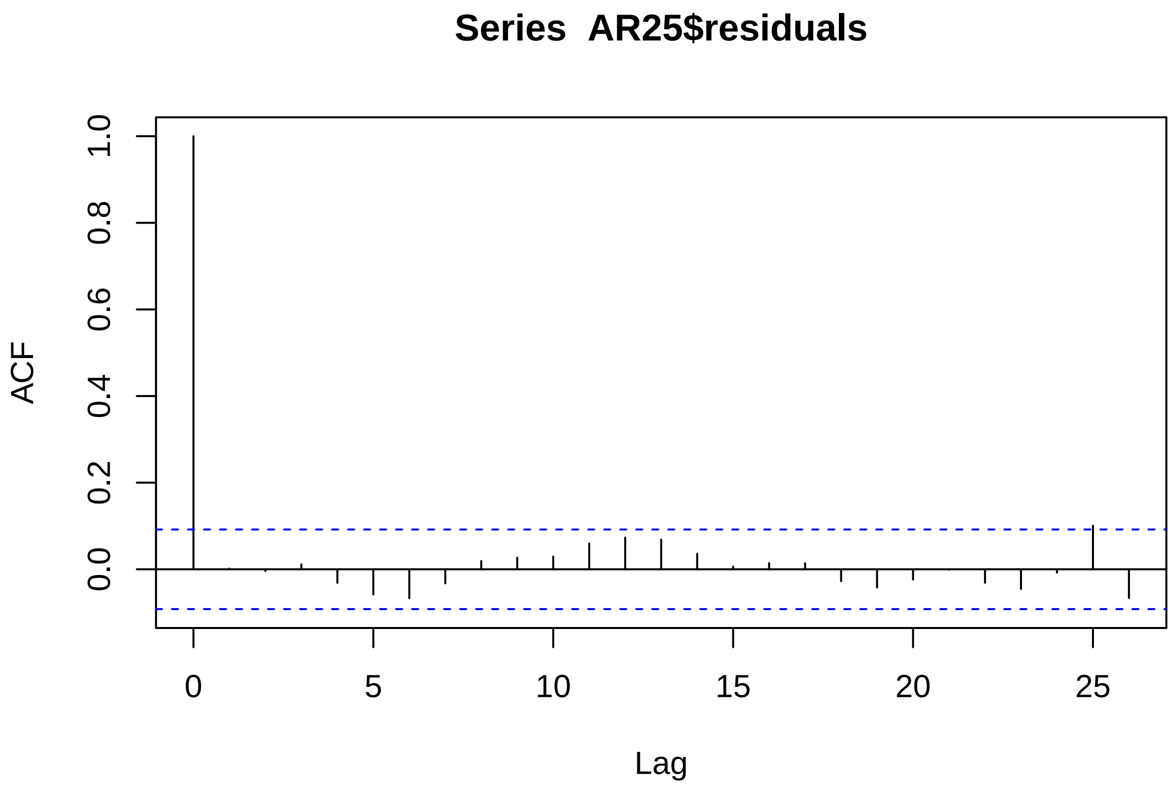
```
# point estimate (mean)
test22$count <- fcst$mean

RMSLE(y_pred = fcst$fitted, y_true = train22$count)
```

```
## [1] 0.08617625
```

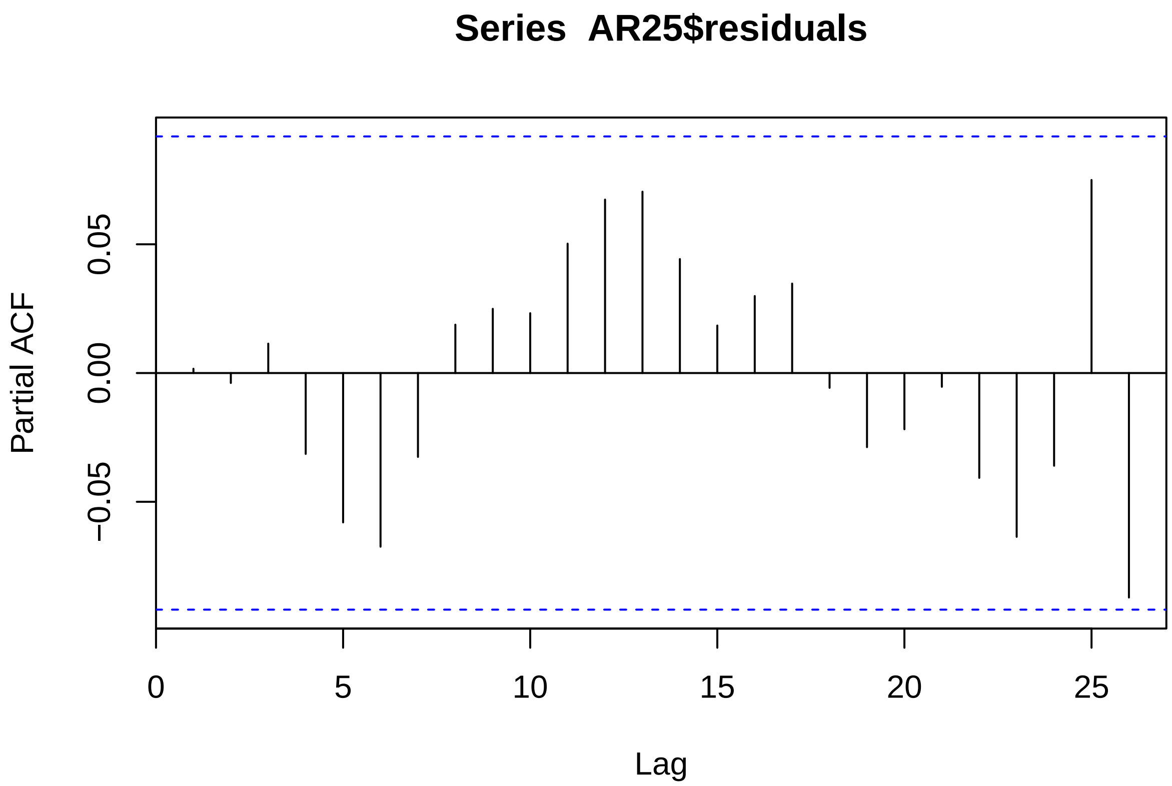
November

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

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


### Log the response variable
train23$count = log(train23$count)

# head(train23)
# head(test23)

AR25 <- arima(train23$count,order=c(25,0,0))
```

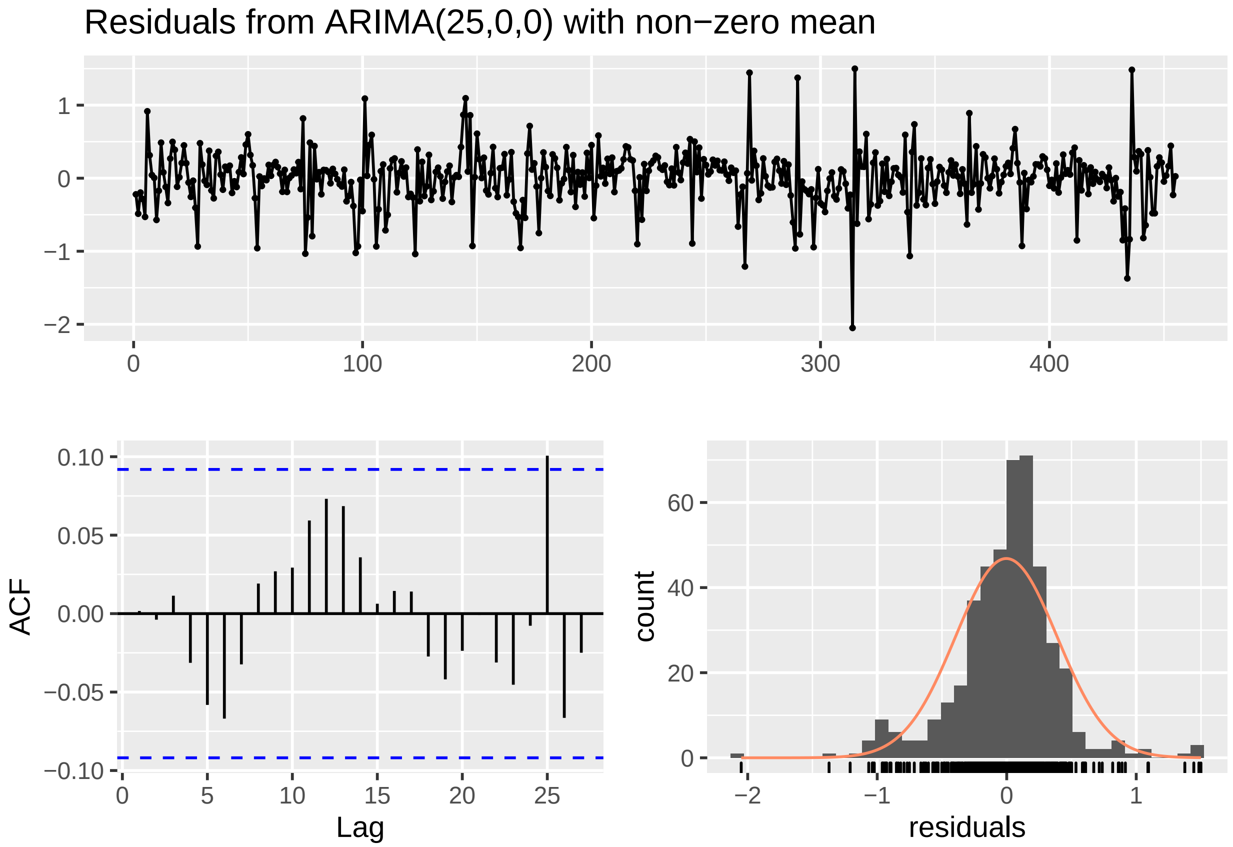
```
# tsdisplay(residuals(AR25),lag.max=25,main="AR(24) Resid. Diagnostics")

number = nrow(test23)

acf(AR25$residuals)
```



```
pacf(AR25$residuals)
```

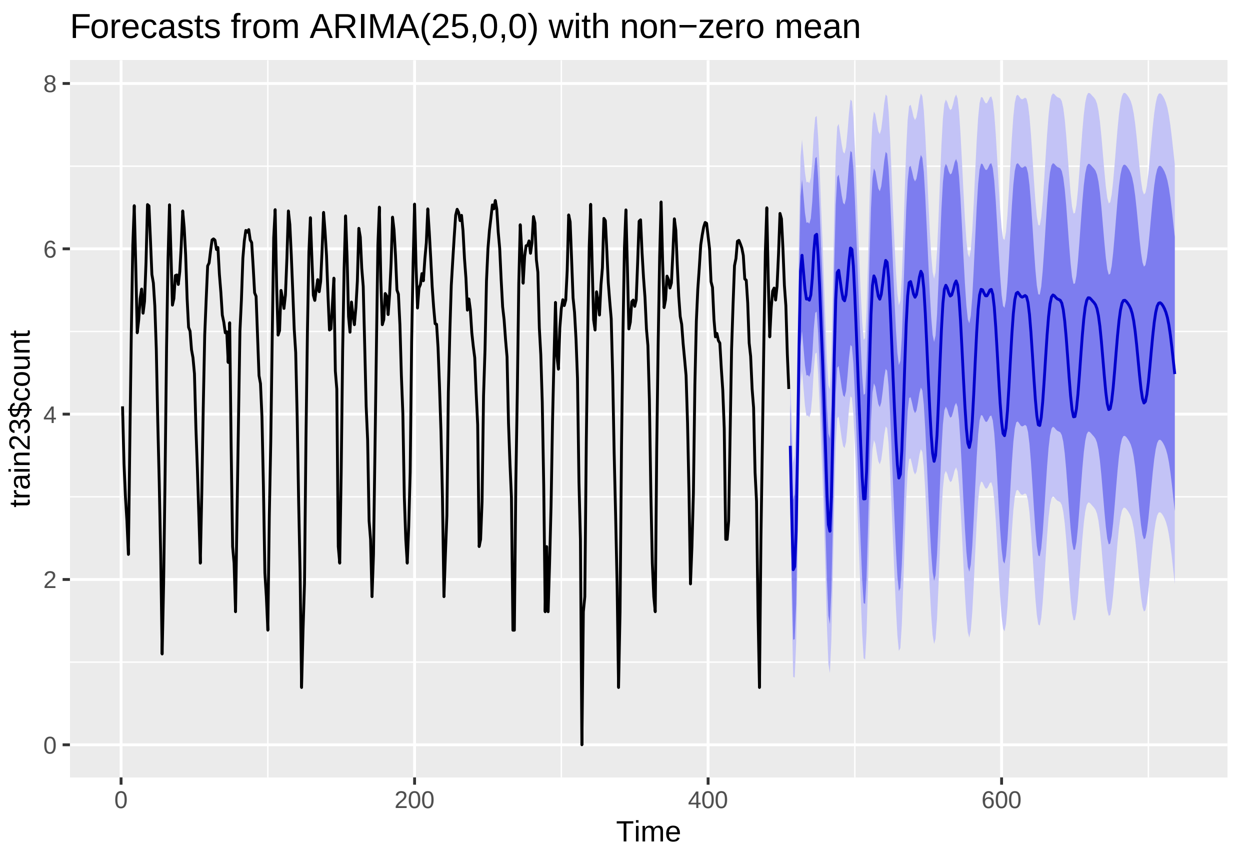


```r
checkresiduals(AR25)
```



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0) with non-zero mean
## Q* = 25.308, df = 3, p-value = 1.332e-05
##
## Model df: 26.   Total lags used: 29
```

```
fcst <- forecast(AR25, h=number)

autoplot(fcst)
```



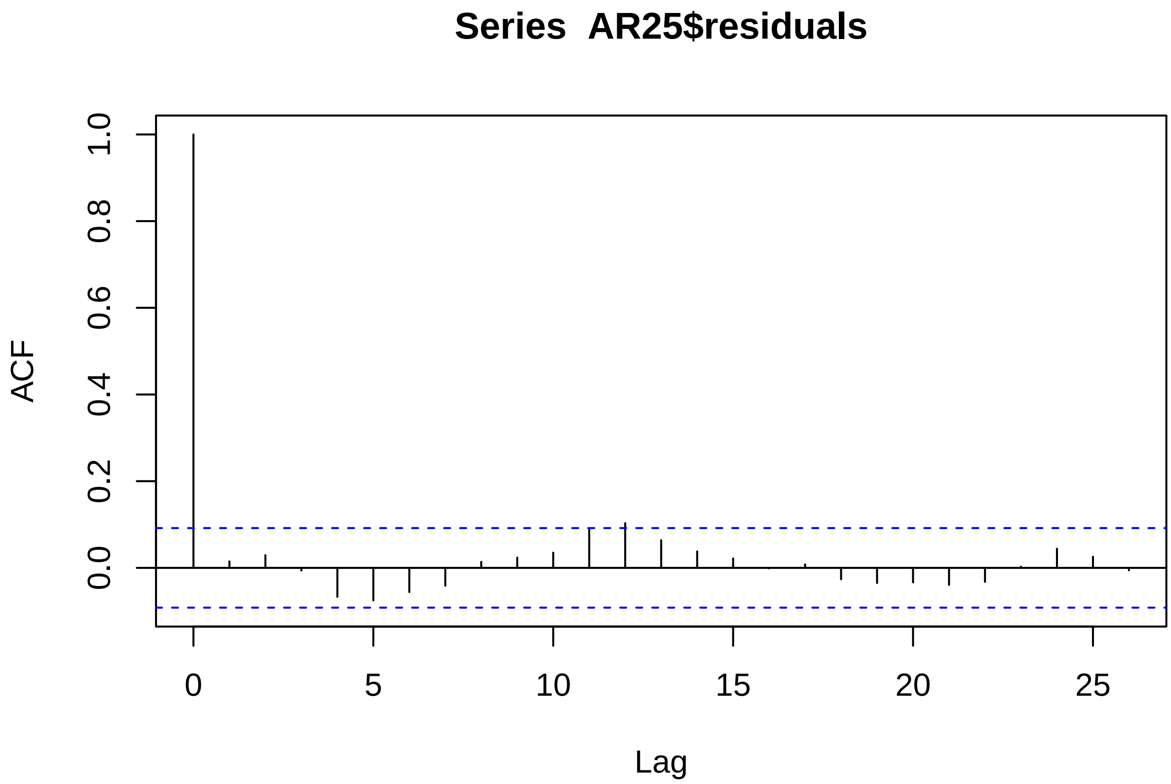
```
# point estimate (mean)
test23$count <- fcst$mean

RMSLE(y_pred = fcst$fitted, y_true = train23$count)
```

```
## [1] 0.1179183
```

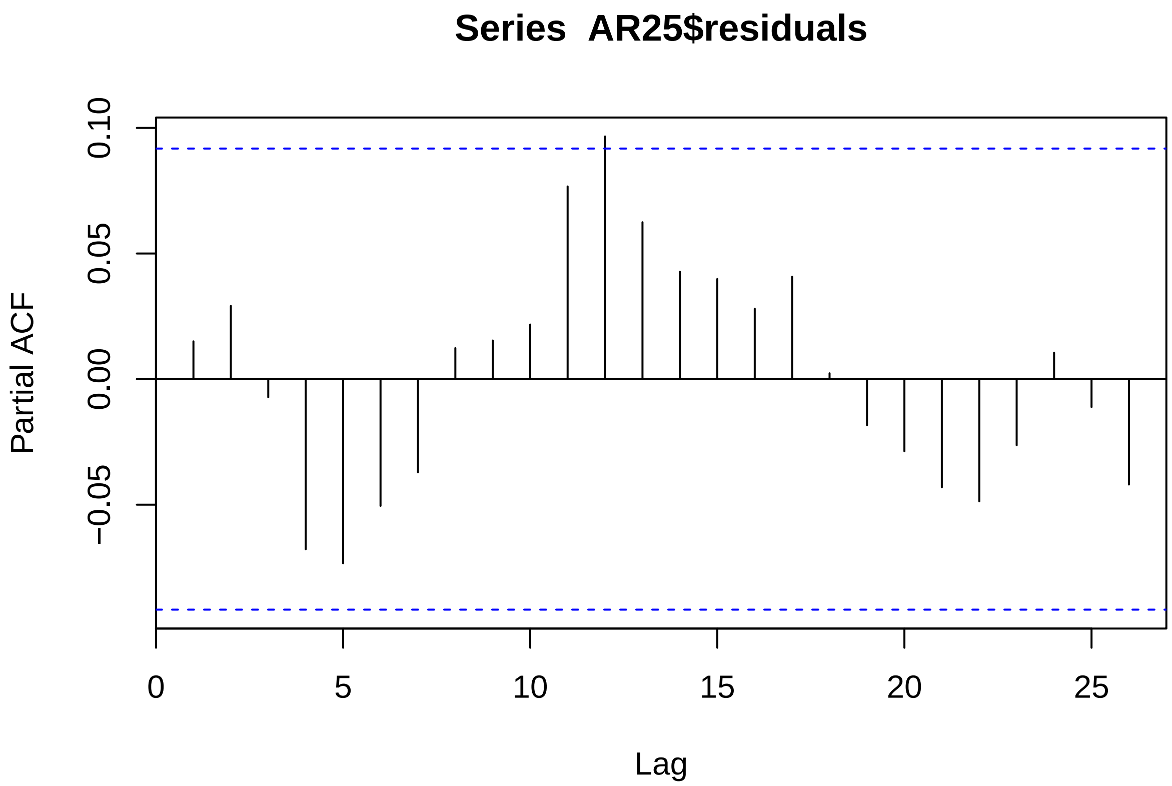
December

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

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


### Log the response variable
train24$count = log(train24$count)

# head(train24)
# head(test24)

AR25 <- arima(train24$count,order=c(25,0,0))
```

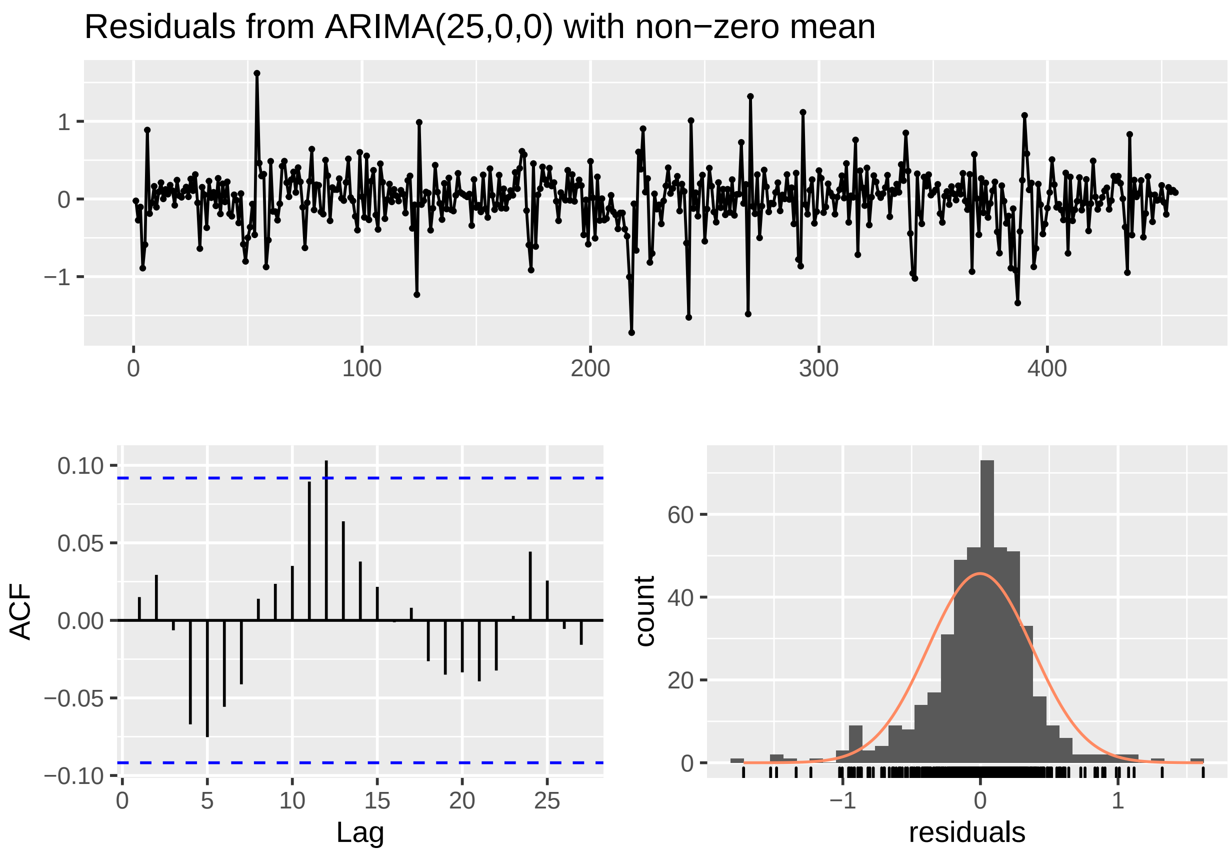
```
number = nrow(test24)

acf(AR25$residuals)
```



```
pacf(AR25$residuals)
```

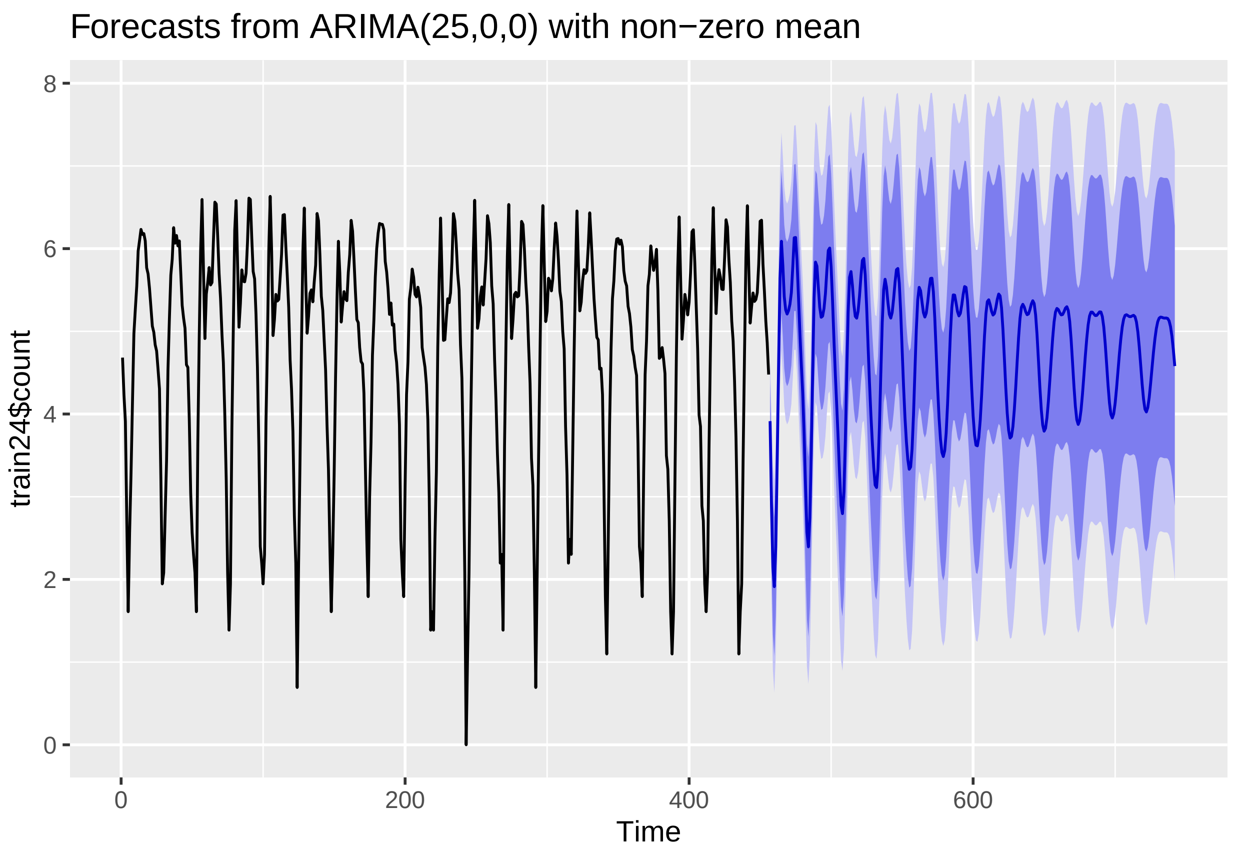


```r
checkresiduals(AR25)
```



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(25,0,0) with non-zero mean
## Q* = 25.753, df = 3, p-value = 1.074e-05
##
## Model df: 26.   Total lags used: 29
```

```
fcst <- forecast(AR25, h=number)

autoplot(fcst)
```



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


RMSLE(y_pred = fcst$fitted, y_true = train24$count)
```

```
## [1] 0.1124551
```

```
summary(AR25)
```

```
##
## Call:
## arima(x = train24$count, order = c(25, 0, 0))
##
## Coefficients:
##          ar1      ar2      ar3     ar4     ar5     ar6      ar7     ar8
##       1.0452  -0.2584  -0.1778  0.0355  0.0058  0.0160  -0.0903  0.0397
## s.e.  0.0464   0.0674   0.0668  0.0670  0.0675  0.0677   0.0676  0.0677
##           ar9     ar10     ar11     ar12     ar13     ar14     ar15
##       -0.0476  0.0199  -0.0489  -0.0039  -0.0262  -0.0168  -0.0182
## s.e.   0.0677  0.0675   0.0676   0.0677   0.0672   0.0675   0.0674
##          ar16     ar17     ar18     ar19     ar20     ar21     ar22     ar23
##       -0.0501  0.0483  -0.0461  -0.0234  0.0072  0.0113  -0.1163  0.3276
## s.e.   0.0675  0.0675   0.0675   0.0674  0.0674  0.0676   0.0682  0.0680
##          ar24     ar25  intercept
##       0.0745  -0.1376     4.7708
## s.e.  0.0688   0.0472     0.0411
```

```
##
## sigma^2 estimated as 0.1441:  log likelihood = -211,  aic = 475.99
##
## Training set error measures:
##                        ME      RMSE       MAE  MPE MAPE      MASE
## Training set -0.002368542 0.3795712 0.2678918 -Inf  Inf 0.5288532
##                    ACF1
## Training set 0.01503052
```

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, tes
                      test13, test14, test15, test16, test17, test18, test19, test20, test21, test22, t


combined <- combined %>%
  mutate(count = round(exp(count)))


# combined
# write.csv(combined, file = "C:\\Users\\Chance\\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 = 1.01847
score = (1 - (2776 / 3246)) * 100

# We only beat ~14% of all submissions
score
```

```
## [1] 14.47936
```