Forcasting issues

Forcast Padawan 2

November 17, 2016

The goal of this experiment is to design the best model to forcaste the number of issue in the per day in the comming two weeks. We think that sthis could help Open Source organisation to manage there human ressources.

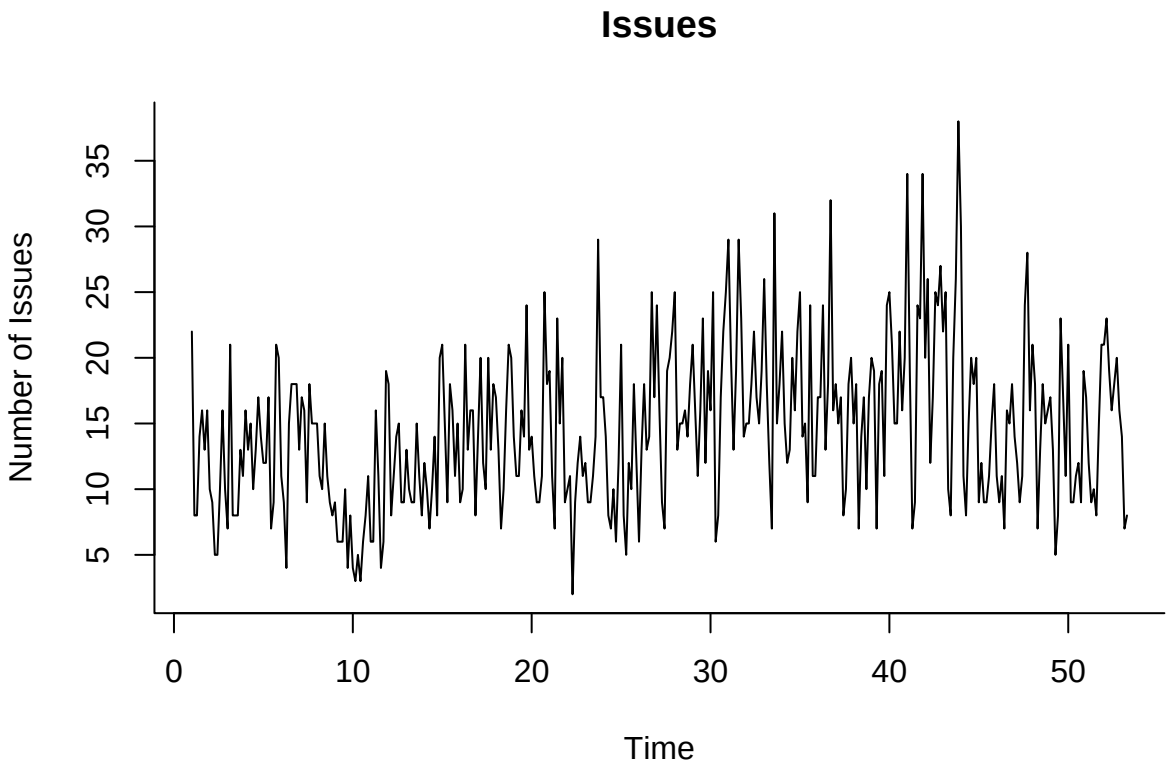
Load the data

```
#install.packages('forecast')
library('forecast')
library(knitr)
#load the data frame
issues.csv <- read.csv("issues/julialang_julia.csv")
commits.csv <- read.csv("commits/julialang_julia.csv")

issues.csv$date = as.POSIXlt(as.Date(issues.csv$date,format='%m/%d/%Y'))
commits.csv$date = as.POSIXlt(as.Date(commits.csv$date,format='%m/%d/%Y'))
```

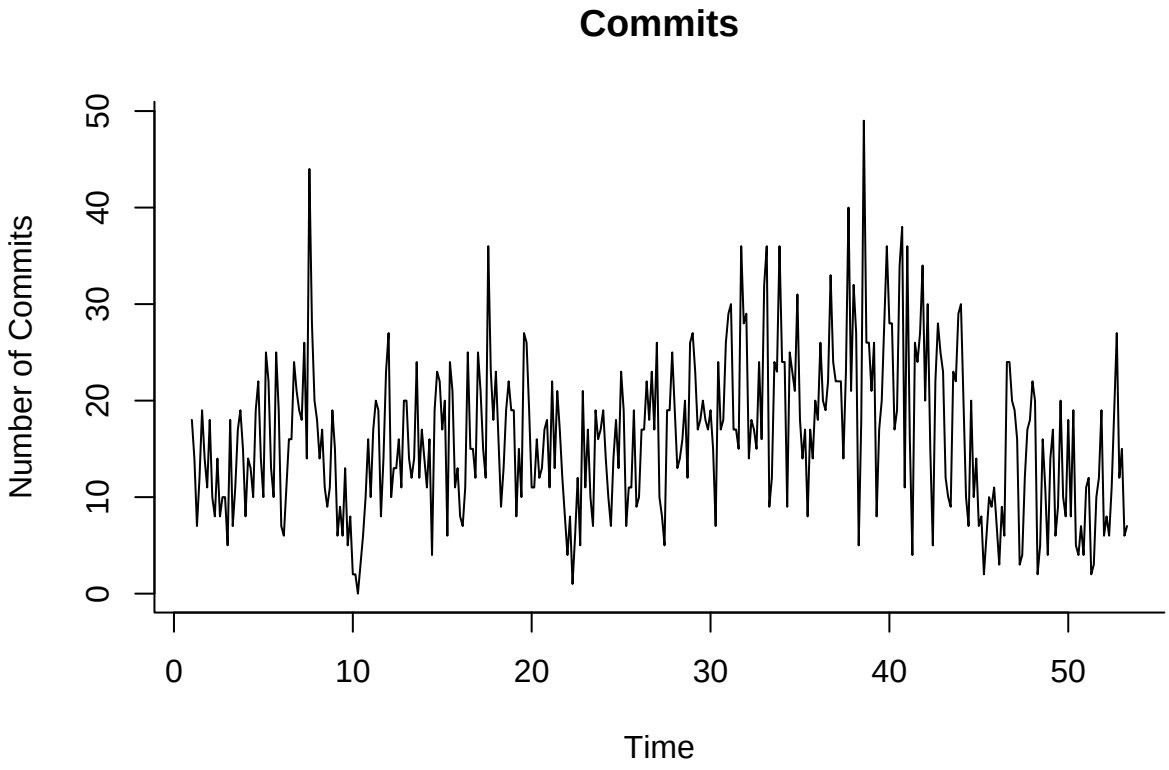
keep the last 12 months

```
to_date <- issues.csv$date[length(issues.csv$date)]
from_date <- to_date
from_date$year <- from_date$year - 1

issues.csv <- subset(issues.csv, date <= to_date & date >= from_date)
commits.csv <- subset(commits.csv, date <= to_date & date >= from_date)
```

```
#loading issues and commits into a ts object
issues.ts <- ts(issues.csv$number_of_issues, frequency = 7)

commits.ts <- ts(commits.csv$number_of_commits, frequency = 7)
plot(issues.ts, main = 'Issues', bty = 'l', ylab = 'Number of Issues')
```




```
plot(commits.ts, main = 'Commits', bty = 'l', ylab = 'Number of Commits')
```




```
time <- time(issues.ts)
```

```
n.valid <- 21
n.train <- length(issues.ts) - n.valid

train.issues.ts <- window(issues.ts, start=time[1], end=time[n.train])
valid.issues.ts <- window(issues.ts,
                    start=time[n.train+1],
                    end=time[n.train+n.valid])

train.commits.ts <- window(commits.ts, start=time[1], end=time[n.train])
valid.commits.ts <- window(commits.ts,
                      start=time[n.train+1],
                      end=time[n.train+n.valid])
```

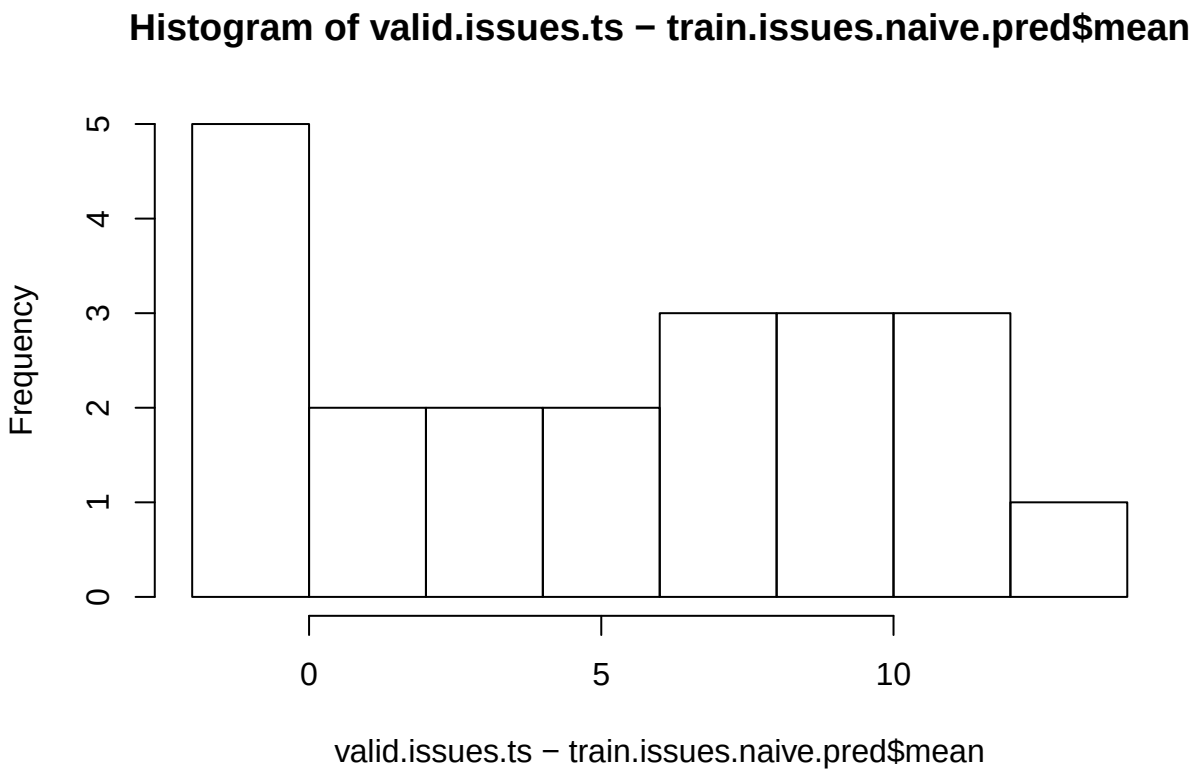
Naive Forecast

Naive

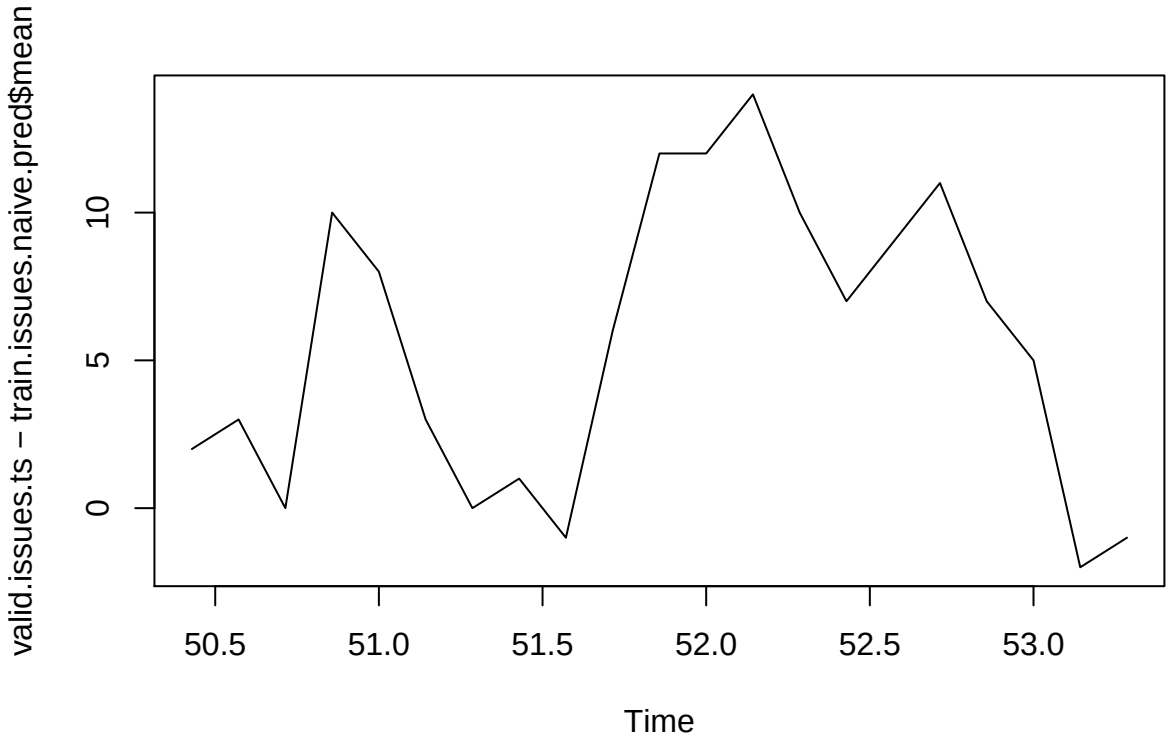
```
train.issues.naive.pred <- naive(train.issues.ts, h=n.valid)
kable(accuracy(train.issues.naive.pred, valid.issues.ts))
```

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Training set	-0.0376812	6.718998	5.260870	-13.08450	42.03105	1.011591	-0.2812264	NA
Test set	5.5238095	7.361418	5.904762	29.53819	34.64023	1.135402	0.5978010	1.293618

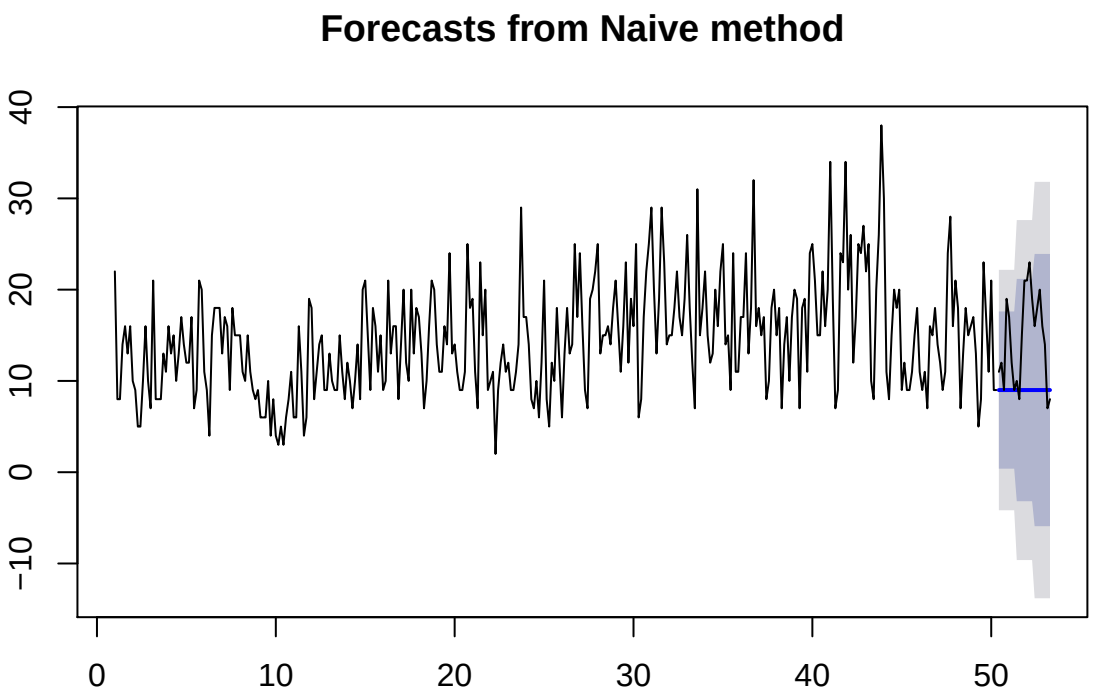
```
hist(valid.issues.ts - train.issues.naive.pred$mean)
```




```
plot(valid.issues.ts - train.issues.naive.pred$mean)
```




```
plot(train.issues.naive.pred)
lines(valid.issues.ts)
```

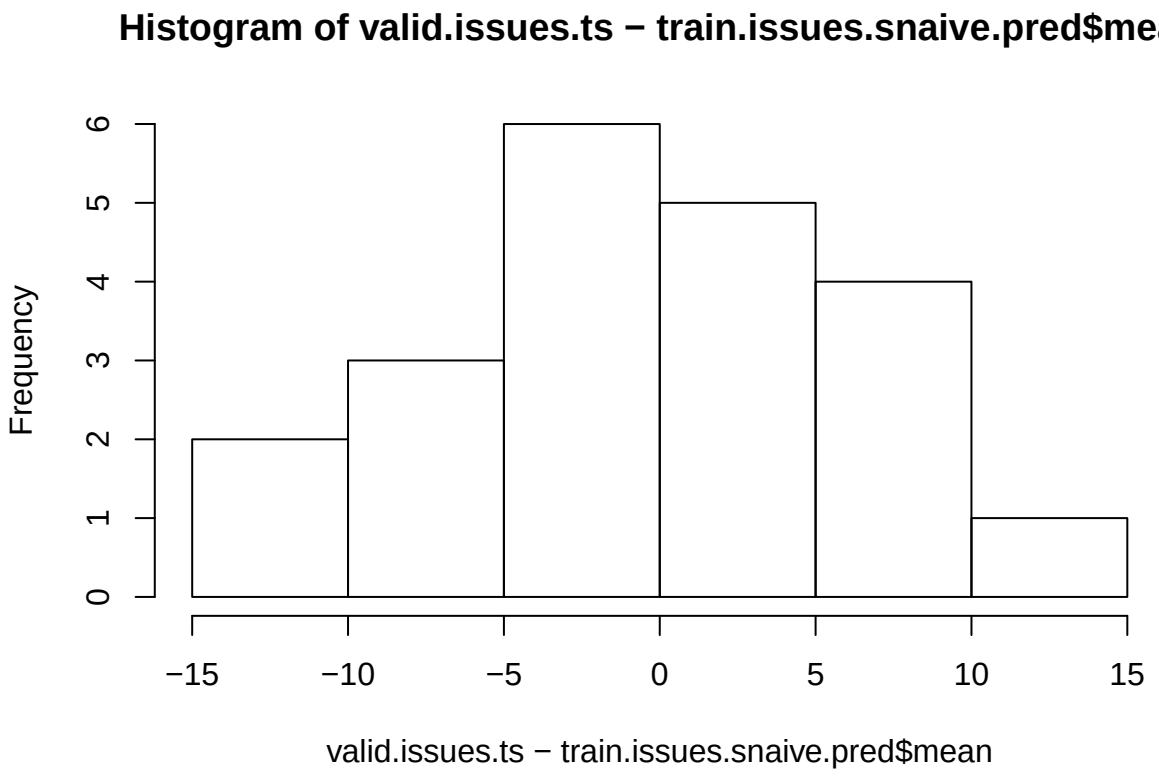


Seasonal Naive

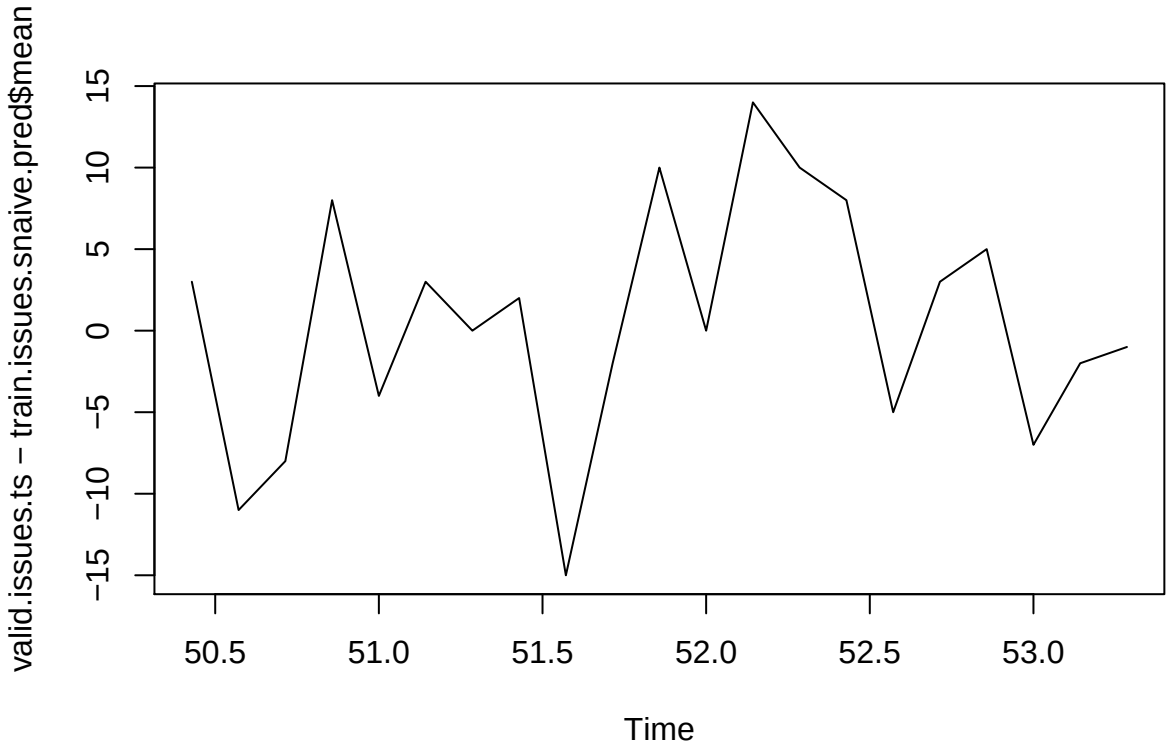
```
train.issues.snaive.pred <- snaive(train.issues.ts, h=n.valid)
kable(accuracy(train.issues.snaive.pred, valid.issues.ts))
```

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Training set	0.0029499	6.552038	5.200590	-12.839076	42.21488	1.000000	0.1720590	NA
Test set	0.5238095	7.201190	5.761905	-7.239015	42.64360	1.107933	0.0766326	1.489315

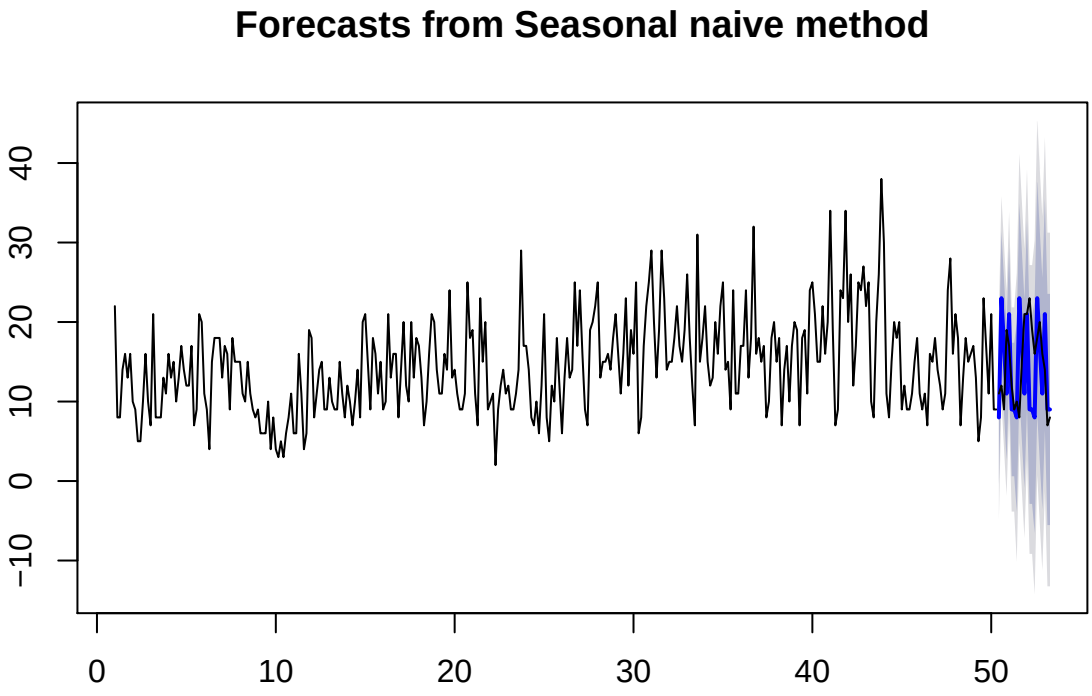
```
hist(valid.issues.ts - train.issues.snaive.pred$mean)
```




```
plot(valid.issues.ts - train.issues.snaive.pred$mean)
```




```
plot(train.issues.snaive.pred)
lines(valid.issues.ts)
```



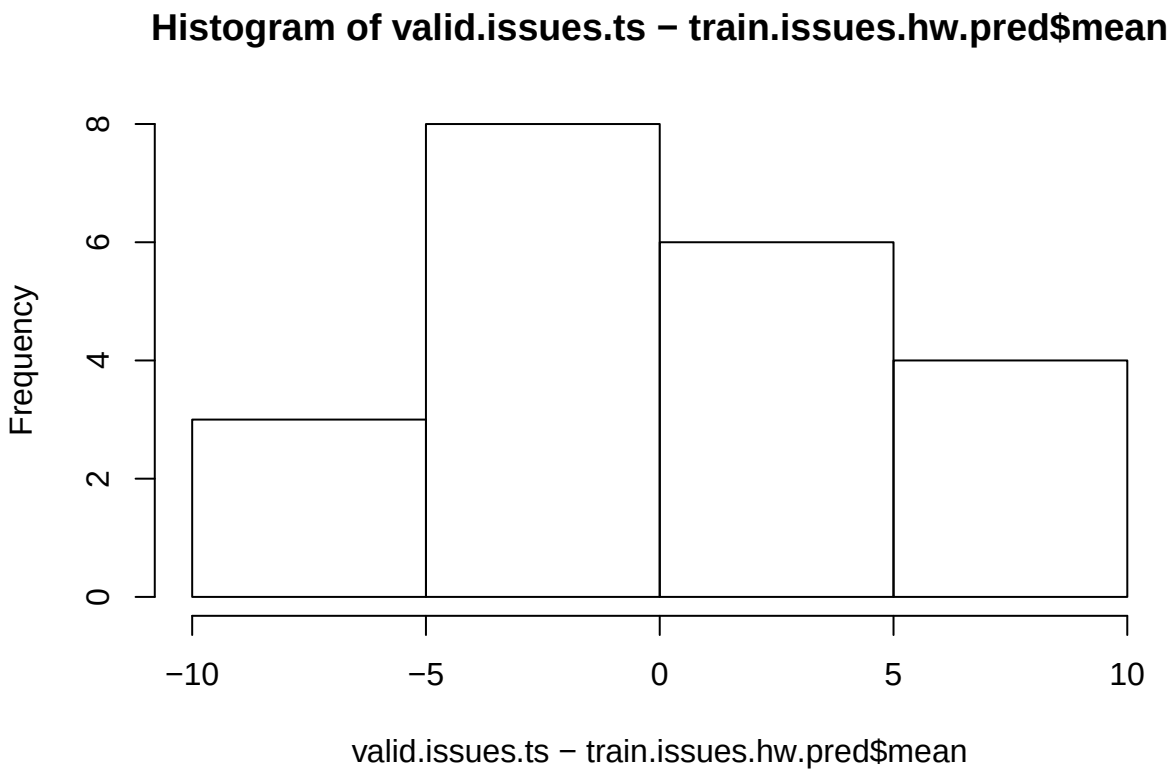
Smoothing

Holt Winter

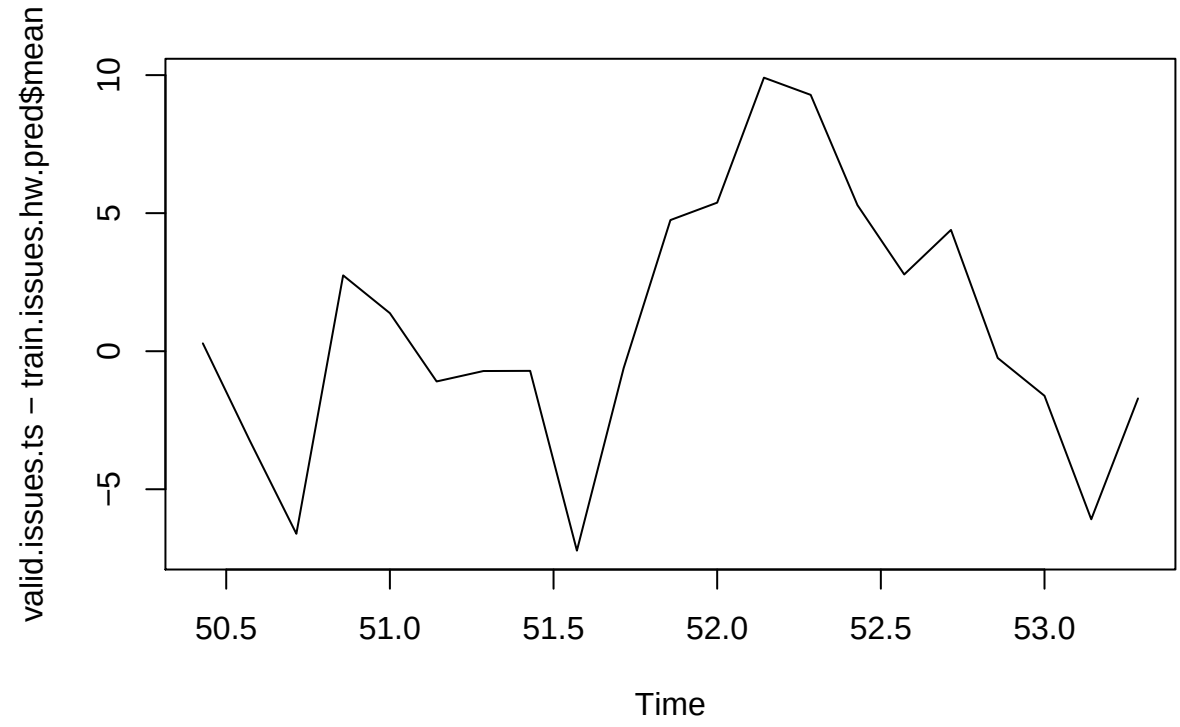
```
train.issues.hw.pred <- hw(train.issues.ts, hw = "ZAA", h = n.valid)
kable(accuracy(train.issues.hw.pred, valid.issues.ts))
```

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Training set	-0.0265046	4.977548	3.872611	-13.061762	32.46036	0.7446484	0.077384	NA
Test set	0.7779834	4.639578	3.621891	-4.980661	27.43986	0.6964384	0.602180	0.837638

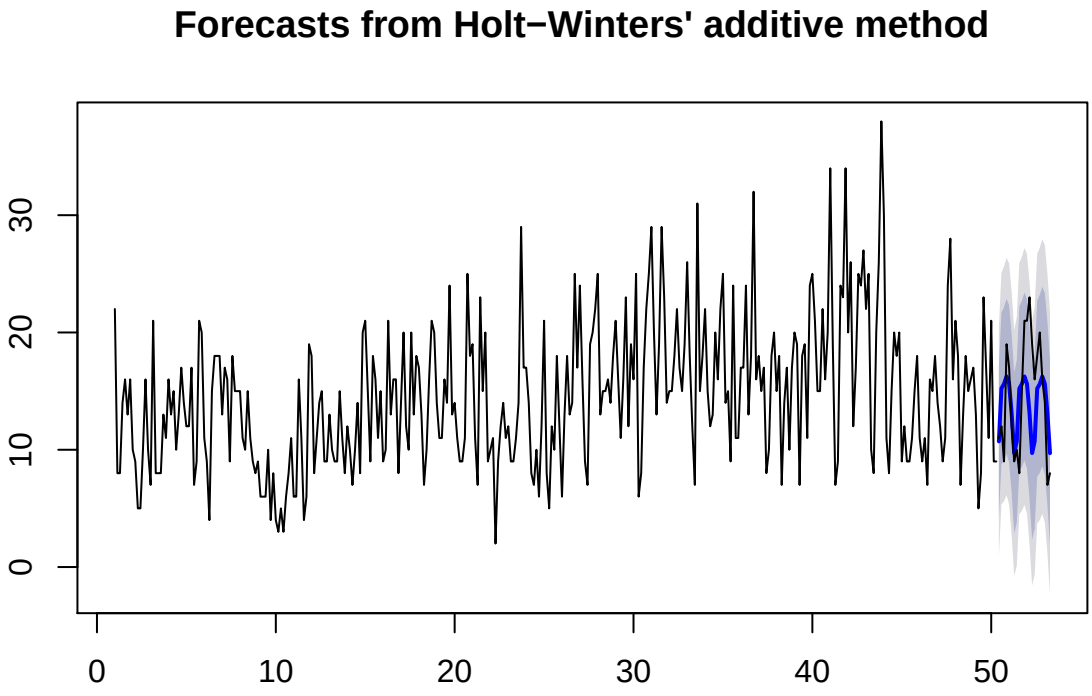
```
hist(valid.issues.ts - train.issues.hw.pred$mean)
```




```
plot(valid.issues.ts - train.issues.hw.pred$mean)
```




```
plot(train.issues.hw.pred)
lines(valid.issues.ts)
```

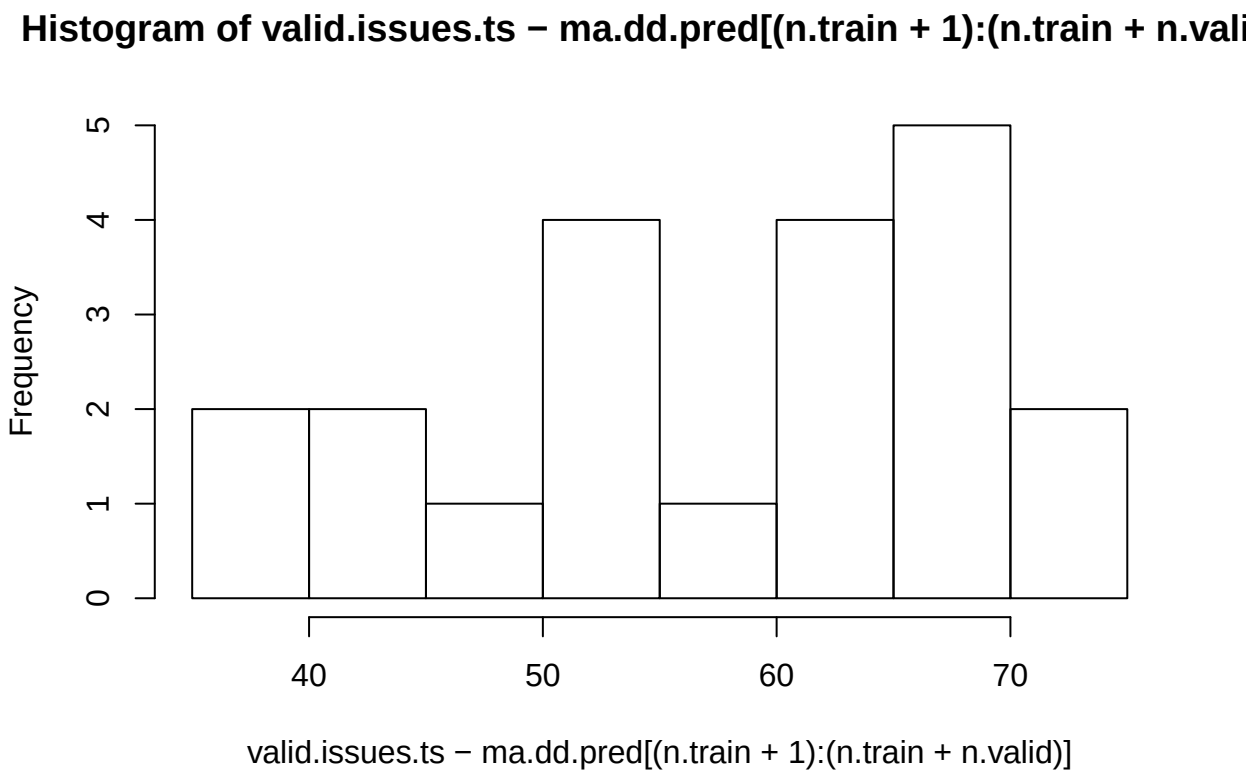


Double differencing

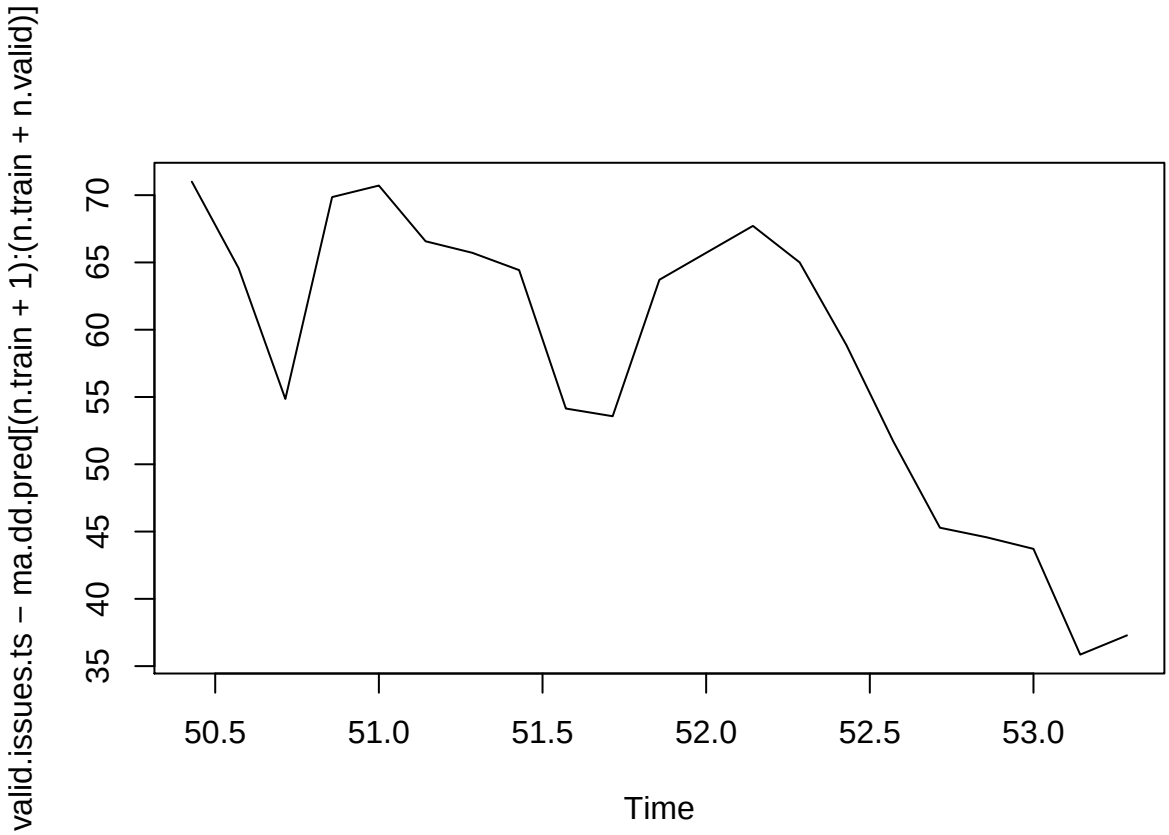
```
train.issues.d1 <- diff(train.issues.ts, lag = 1)
train.issues.d1.d7 <- diff(train.issues.d1, lag = 7)

ma.trailing <- rollmean(train.issues.d1.d7, k = 7, align = "right")
last.ma <- tail(ma.trailing, 1)
ma.trailing.pred <- ts(c(train.issues.d1.d7[1:6], ma.trailing, rep(last.ma, n.valid)), start=c(2,2), fr

ma.dd.pred.d1 <- diffinv(ma.trailing.pred, lag = 7, xi=train.issues.d1[1:7])
ma.dd.pred <- diffinv(ma.dd.pred.d1, lag = 1, xi=train.issues.ts[1])

kable(accuracy(ma.dd.pred[(n.train+1):(n.train+n.valid)], valid.issues.ts))
```

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Test set	57.85034	58.8597	57.85034	440.1622	440.1622	0.7177956	12.1168

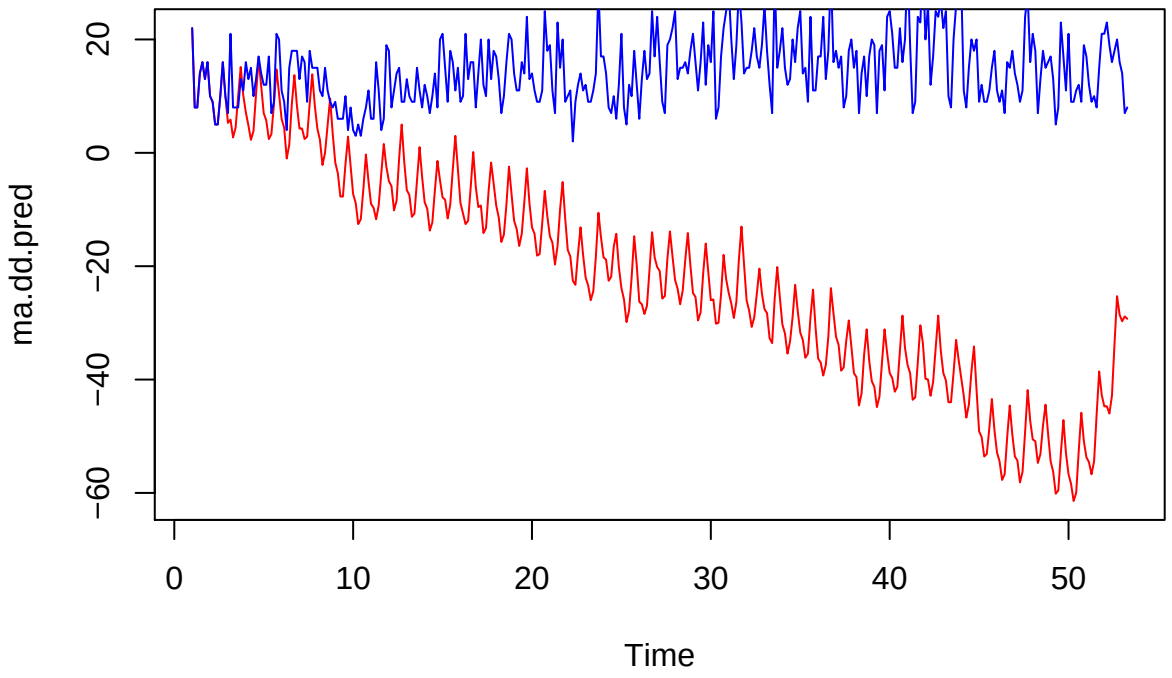
```
hist(valid.issues.ts - ma.dd.pred[(n.train+1):(n.train+n.valid)])
```




```
plot(valid.issues.ts - ma.dd.pred[(n.train+1):(n.train+n.valid)])
```



```
plot(ma.dd.pred,col='red')
lines(issues.ts,col='blue')
```



Regression

Linear regression

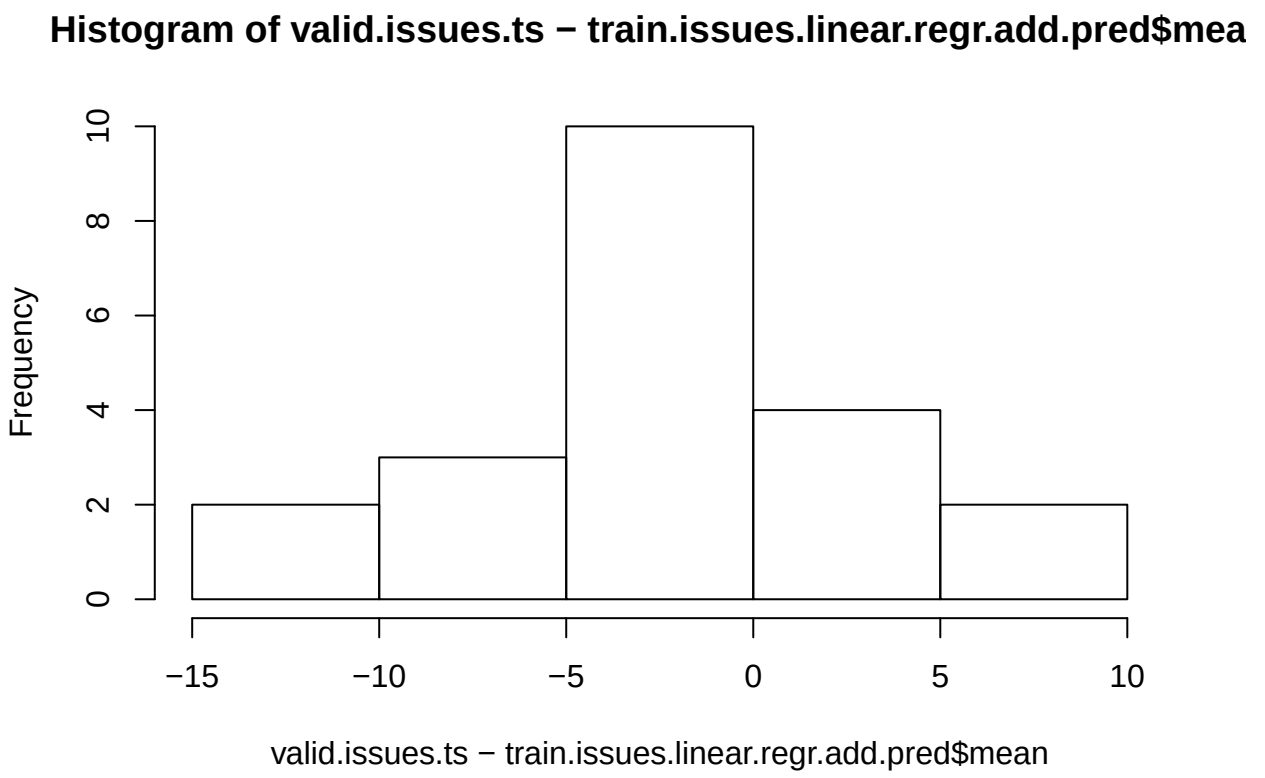
```
train.issues.linear.regr.add.m <- tslm(train.issues.ts ~ trend + season, lambda = 0)
train.issues.linear.regr.add.m
```

```
##
## Call:
## tslm(formula = train.issues.ts ~ trend + season, lambda = 0)
##
## Coefficients:
## (Intercept)        trend      season2      season3      season4
##    2.483884     0.001444    -0.191684    -0.488595    -0.354584
##     season5      season6      season7
##   -0.011783    -0.018838     0.016470
```

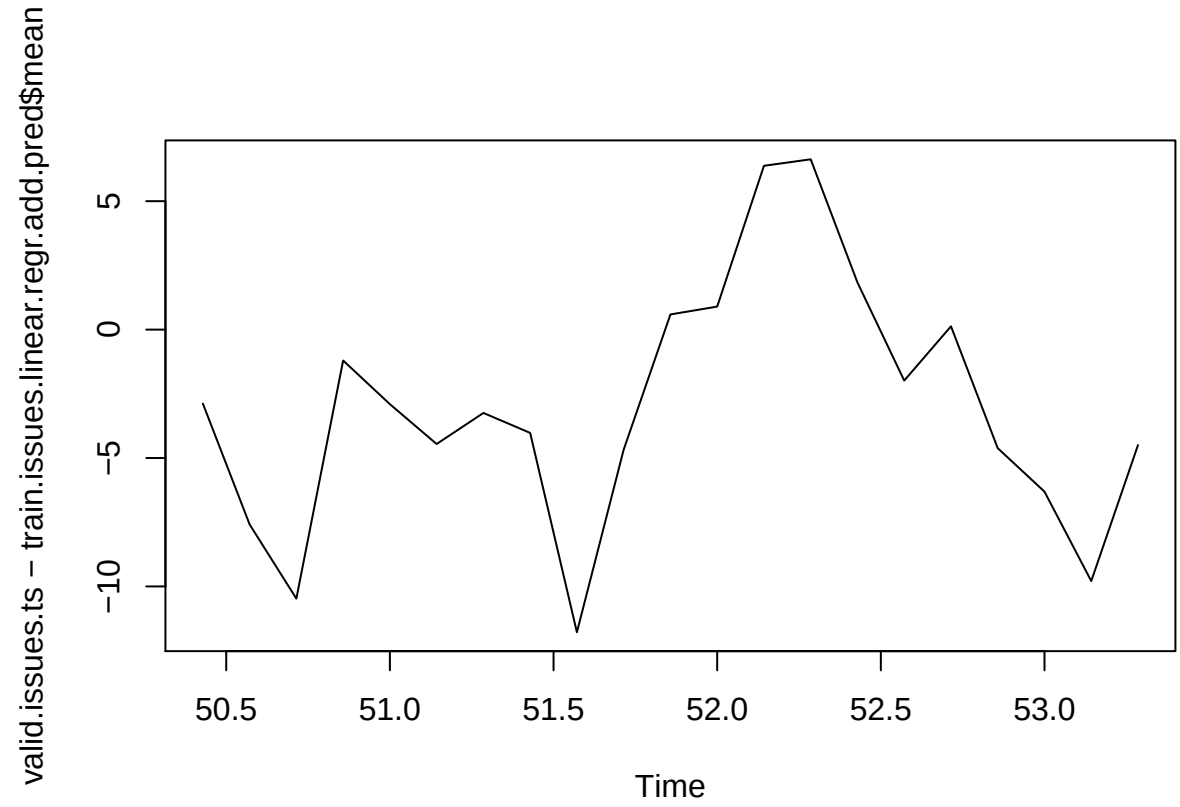
```
train.issues.linear.regr.add.pred <- forecast(train.issues.linear.regr.add.m , h=n.valid)

kable(accuracy(train.issues.linear.regr.add.pred, valid.issues.ts))
```

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Training set	0.9509424	5.238848	4.086120	-8.514036	33.40185	0.7857032	0.2766360	NA
Test set	-3.0456515	5.618531	4.613154	-34.294688	42.08708	0.8870443	0.5703958	1.276547

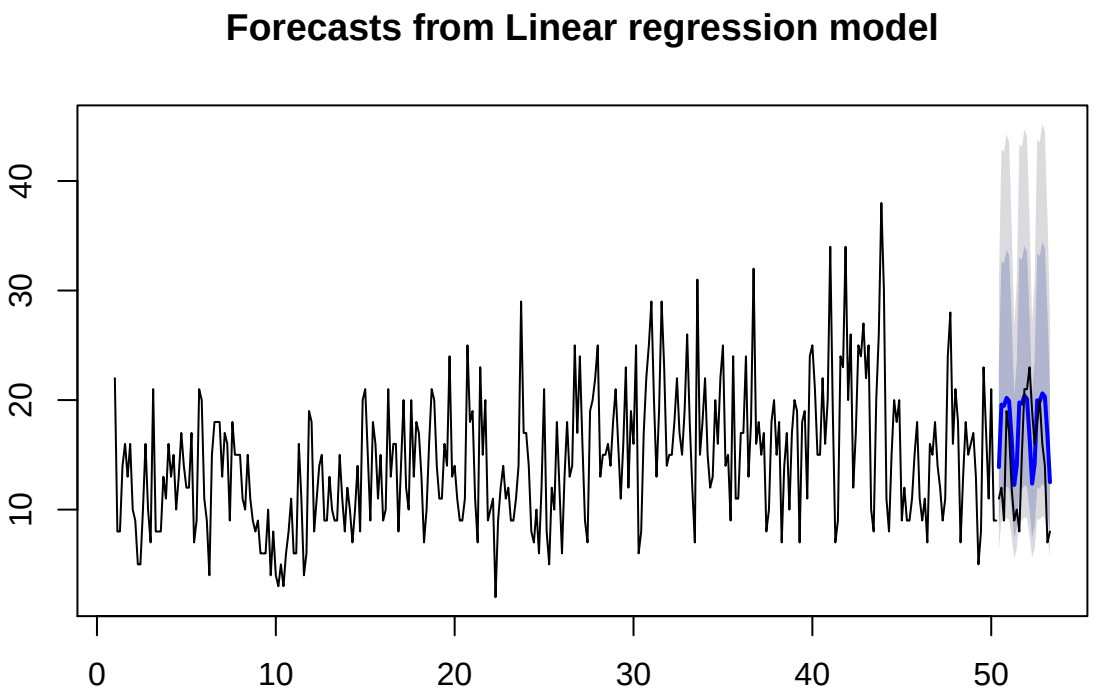
```
hist(valid.issues.ts - train.issues.linear.regr.add.pred$mean)
```




```
plot(valid.issues.ts - train.issues.linear.regr.add.pred$mean)
```



```
plot(train.issues.linear.regr.add.pred)
lines(valid.issues.ts)
```



exponential regression

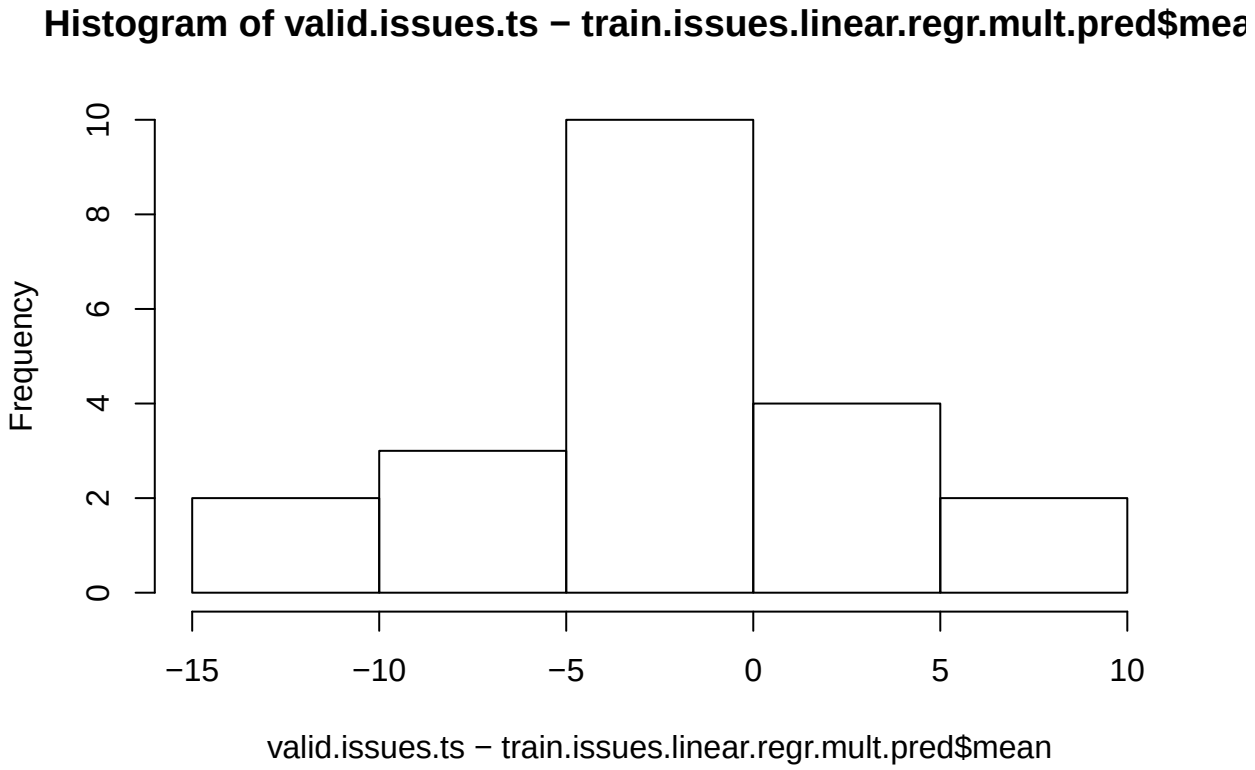
```
train.issues.linear.regr.mult.m <- tslm(train.issues.ts ~ trend + season, lambda = 0)
train.issues.linear.regr.mult.m
```

```
##
## Call:
## tslm(formula = train.issues.ts ~ trend + season, lambda = 0)
##
## Coefficients:
## (Intercept)        trend      season2      season3      season4
##    2.483884     0.001444    -0.191684    -0.488595    -0.354584
##     season5      season6      season7
##   -0.011783    -0.018838     0.016470
```

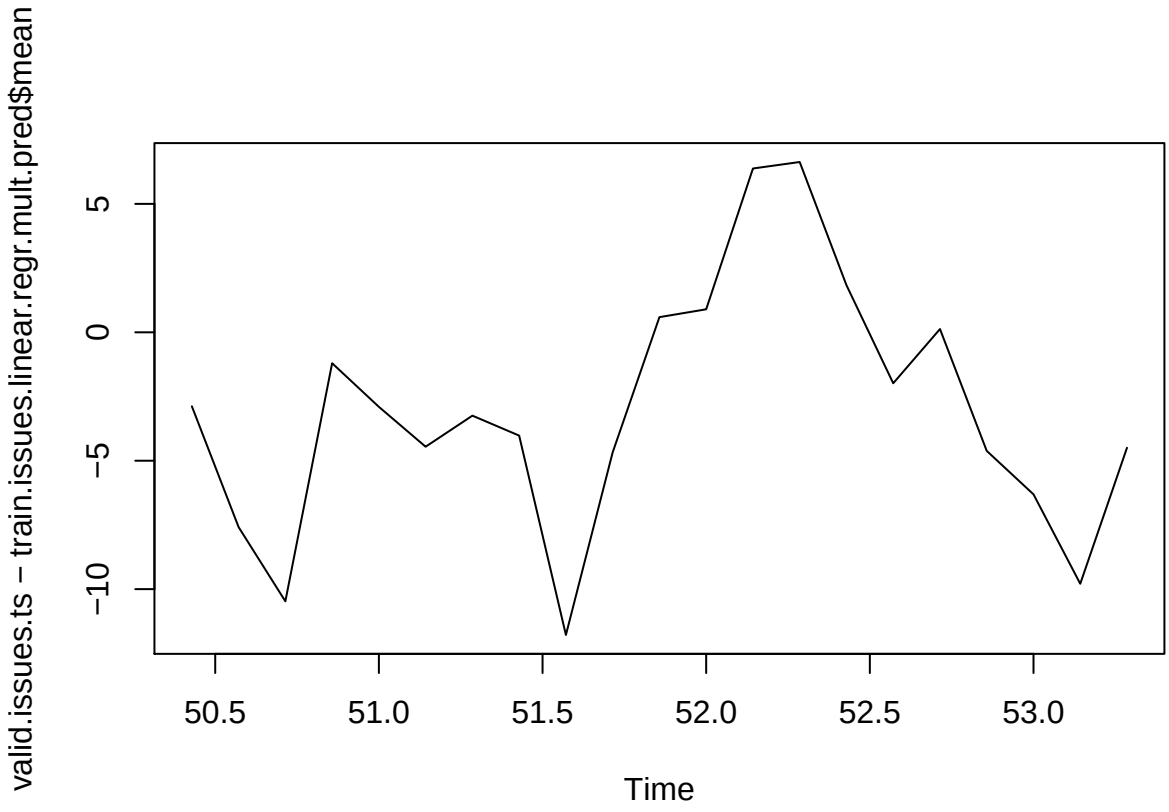
```
train.issues.linear.regr.mult.pred <- forecast(train.issues.linear.regr.mult.m , h=n.valid)

kable(accuracy(train.issues.linear.regr.mult.pred, valid.issues.ts))
```

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Training set	0.9509424	5.238848	4.086120	-8.514036	33.40185	0.7857032	0.2766360	NA
Test set	-3.0456515	5.618531	4.613154	-34.294688	42.08708	0.8870443	0.5703958	1.276547

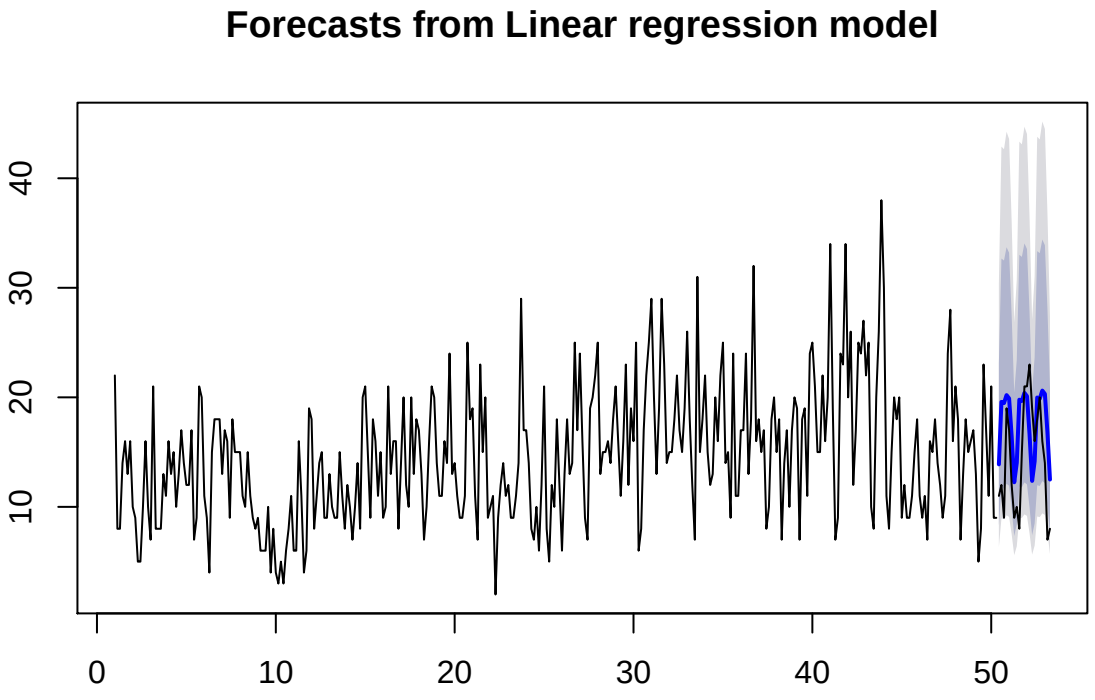
```
hist(valid.issues.ts - train.issues.linear.regr.mult.pred$mean)
```




```
plot(valid.issues.ts - train.issues.linear.regr.mult.pred$mean)
```

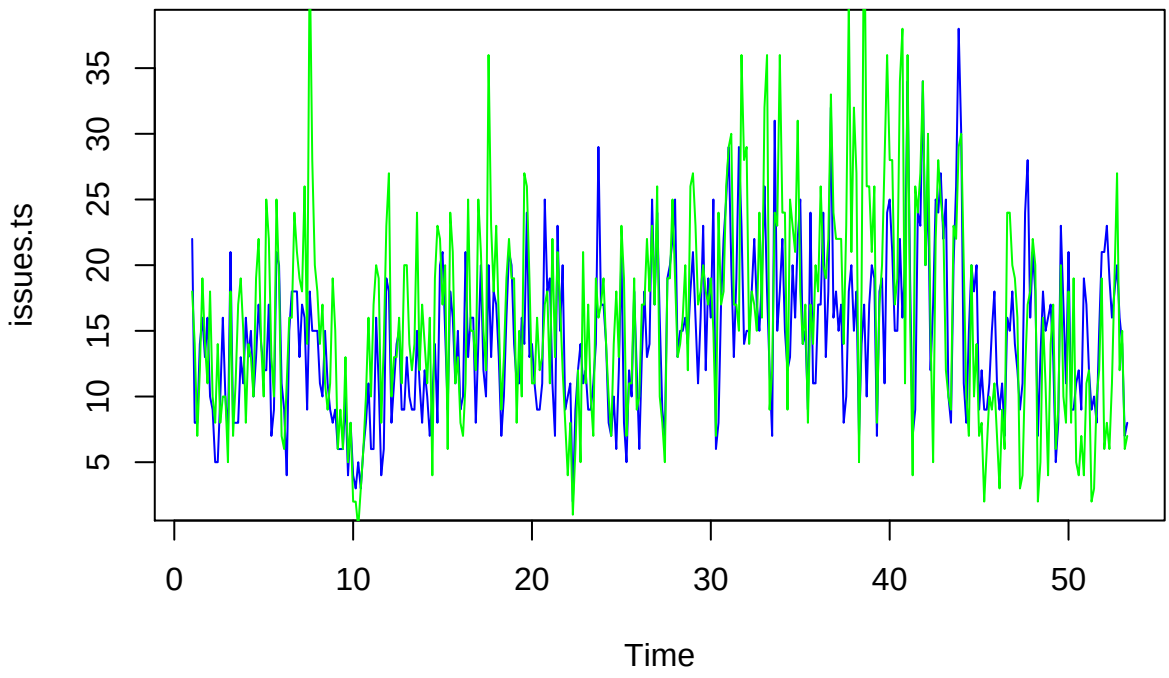



```
plot(train.issues.linear.regr.mult.pred)
lines(valid.issues.ts)
```

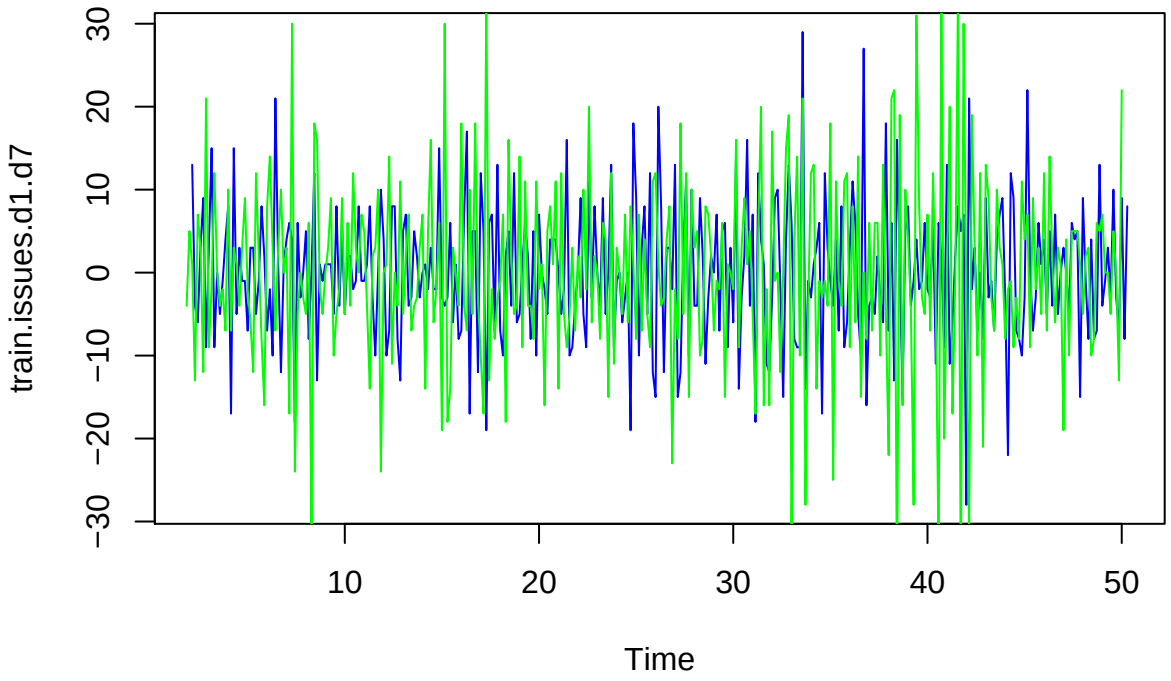


external regression

```
plot(issues.ts, col='blue')
lines(commits.ts, col='green')
```




```
train.commits.d1 <- diff(train.commits.ts, lag = 1)
train.commits.d1.d7 <- diff(train.commits.d1, lag = 7)

plot(train.issues.d1.d7, col='blue')
lines(lag(train.commits.d1.d7,2), col='green')
```




```
train.issues.arima.ext.m <- Arima(train.issues.ts, order=c(1,0,0), seasonal=c(1,0,0), xreg=train.commits
train.issues.arima.ext.m
```

```
## Series: train.issues.ts
## ARIMA(1,0,0)(1,0,0)[7] with non-zero mean
##
## Coefficients:
##          ar1    sar1  intercept  train.commits.ts
##       0.2050  0.2187     7.9198            0.4004
## s.e.  0.0548  0.0565     0.7248            0.0366
##
## sigma^2 estimated as 21.16:  log likelihood=-1017.15
## AIC=2044.3   AICc=2044.48   BIC=2063.53
```