Project1

Alice Ding

2024-06-18

Data Importing and Indexing

```r
data_start_ind <- 1
data_end_ind <- 1622
forecast_stary_ind <- 1623
forecast_end_ind <- 1722

path <- paste(getwd(), '/Data Set for Class.xls', sep="")
sheet_name <- 'S01'

# Read the specified sheet from the Excel file
s01 <- read_excel(path, sheet = sheet_name)
```

Data Visualization

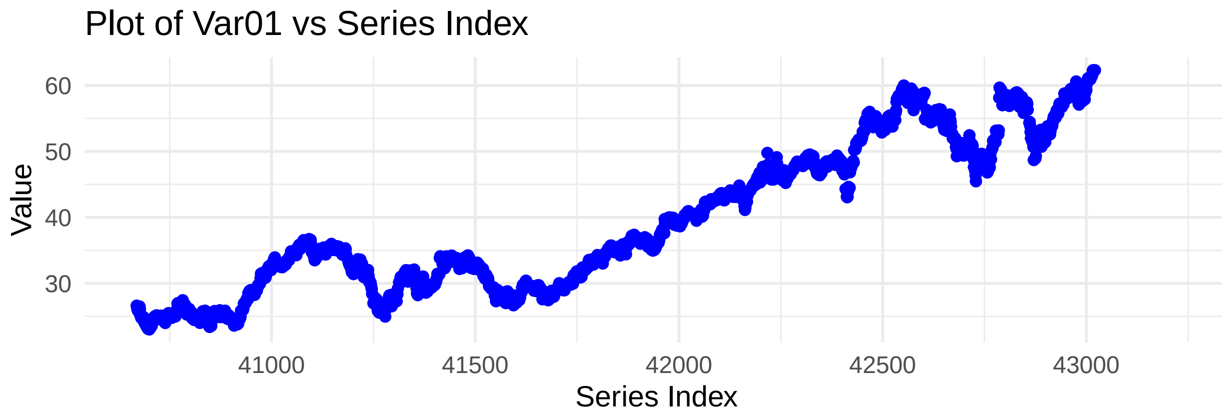
```r
var1_plot <- ggplot(s01, aes(x = SeriesInd, y = Var01)) +
  geom_point(color = "blue") +
  labs(title = "Plot of Var01 vs Series Index", x = "Series Index", y = "Value") +
  theme_minimal()
```

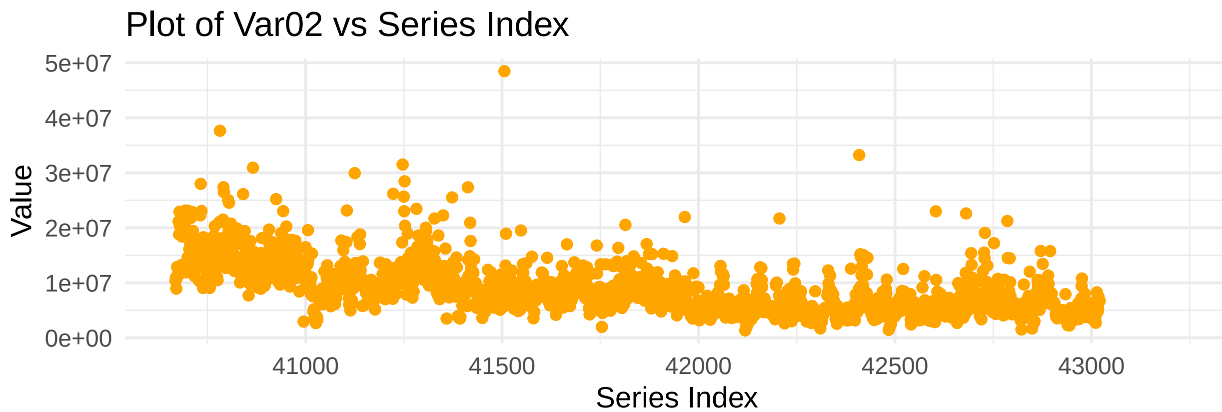
```r
var2_plot <- ggplot(s01, aes(x = SeriesInd, y = Var02)) +
  geom_point(color = "orange") +
  labs(title = "Plot of Var02 vs Series Index", x = "Series Index", y = "Value") +
  theme_minimal()
```

```r
grid.arrange(var1_plot, var2_plot, nrow = 2)
```

```
## Warning: Removed 142 rows containing missing values or values outside the scale range
## (‘geom_point()‘).
```

```
## Warning: Removed 140 rows containing missing values or values outside the scale range
## (‘geom_point()‘).
```





Data Imputation

I'm using linear imputation, so creating a line of best fit between the last two known points and filling in missing values along that line. This works for Var01, however for Var02, I will impute the median given how it contains more static.

```
data_range <- which(s01$SeriesInd < 43022)
na_var1 <- which(is.na(s01$Var01[data_range]))
```

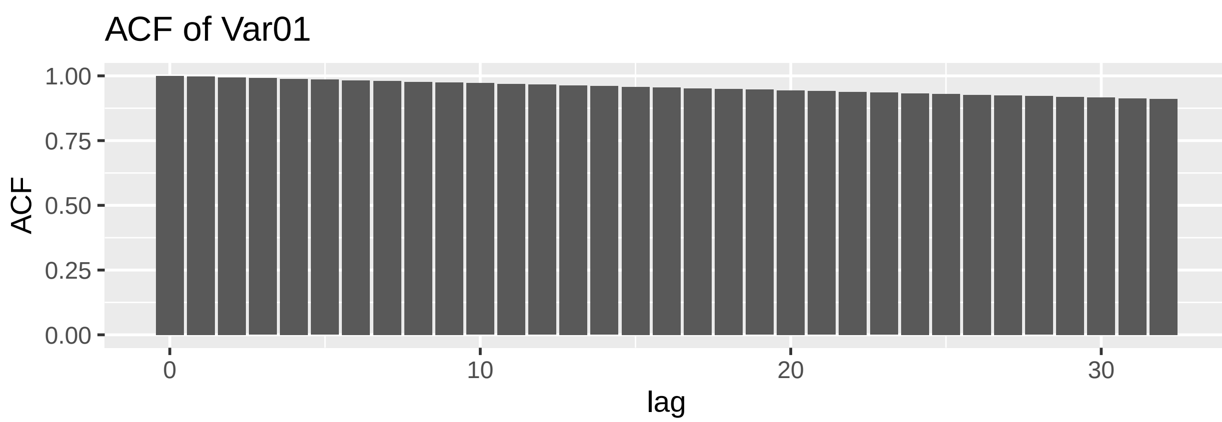
```
imputed_var1 <- approx(x = s01$SeriesInd[data_range], y = s01$Var01[data_range],
                              xout = s01$SeriesInd[data_range])$y
```

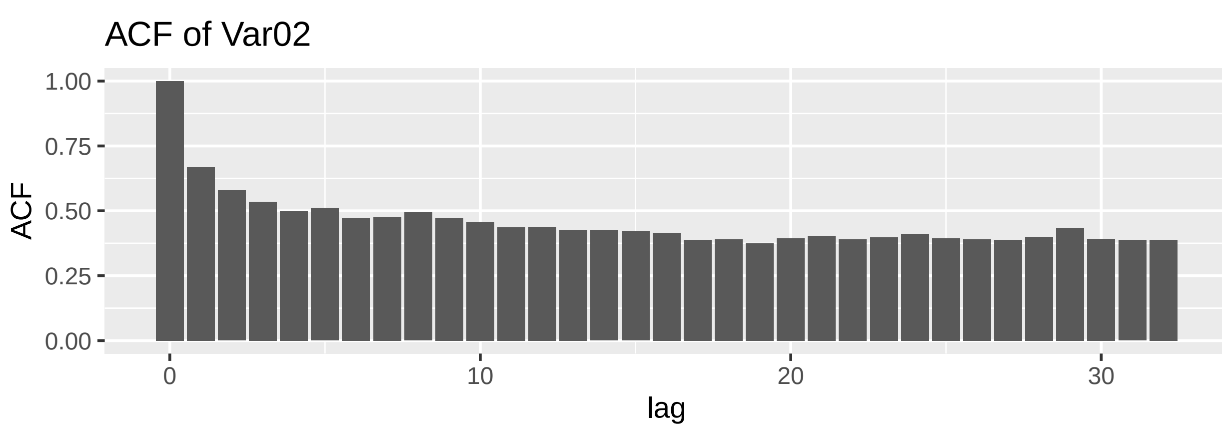
```
s01$Var01[data_range][na_var1] <- imputed_var1[na_var1]
s01 <- s01 |>
  mutate(Var02 = replace_na(Var02, median(Var02, na.rm=TRUE)))
```

Values to forecast: 43022 - 43221 index numbers: 1623 - 1762

Checking for Stationarity

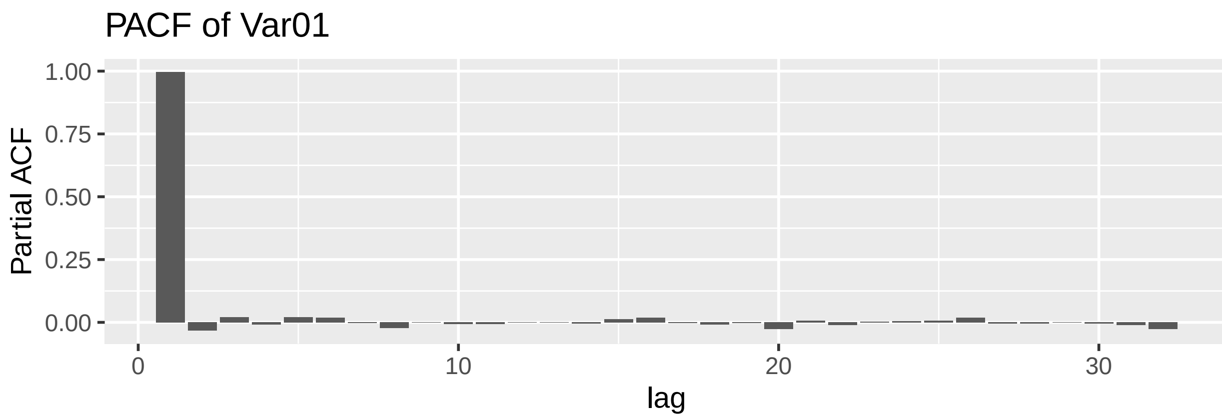
```r
acf_var1 <- acf(s01$Var01[data_range], plot = FALSE)
acf_var2 <- acf(s01$Var02[data_range], plot = FALSE)

acf_var1_df <- data.frame(lag = acf_var1$lag, acf = acf_var1$acf)
acf_var2_df <- data.frame(lag = acf_var2$lag, acf = acf_var2$acf)

acf1 <- ggplot(acf_var1_df, aes(x = lag, y = acf)) +
  geom_bar(stat = "identity") +
  labs(title = "ACF of Var01", y = 'ACF')

acf2 <- ggplot(acf_var2_df, aes(x = lag, y = acf)) +
  geom_bar(stat = "identity") +
  labs(title = "ACF of Var02", y = 'ACF')

grid.arrange(acf1, acf2, nrow=2)
```

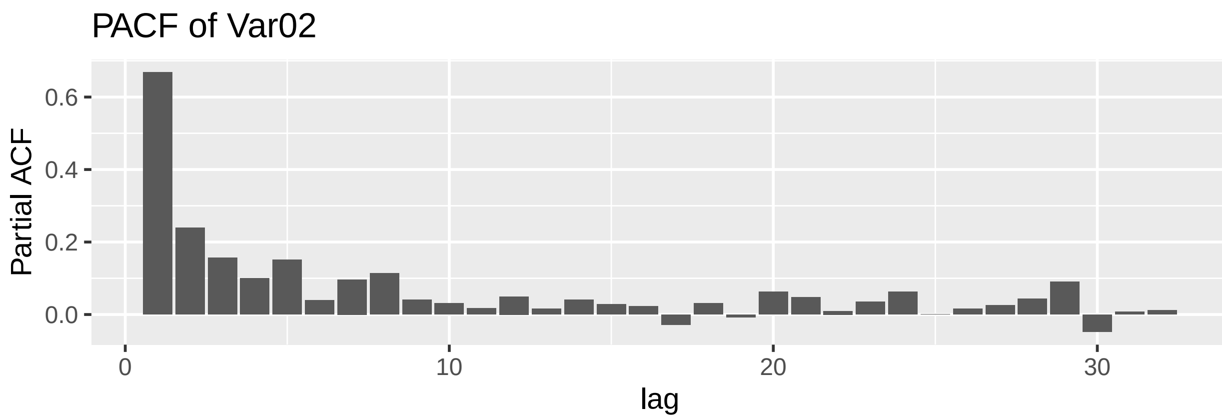





```r
pacf_var1 <- pacf(s01$Var01[data_range], plot = FALSE)
pacf_var2 <- pacf(s01$Var02[data_range], plot = FALSE)

pacf_var1_df <- data.frame(lag = pacf_var1$lag, pacf = pacf_var1$acf)
pacf_var2_df <- data.frame(lag = pacf_var2$lag, pacf = pacf_var2$acf)

pacf1 <- ggplot(pacf_var1_df, aes(x = lag, y = pacf)) +
  geom_bar(stat = "identity") +
  labs(title = "PACF of Var01", y = 'Partial ACF')
```

```r
pacf2 <- ggplot(pacf_var2_df, aes(x = lag, y = pacf)) +
  geom_bar(stat = "identity") +
  labs(title = "PACF of Var02", y = 'Partial ACF')

grid.arrange(pacf1, pacf2, nrow=2)
```





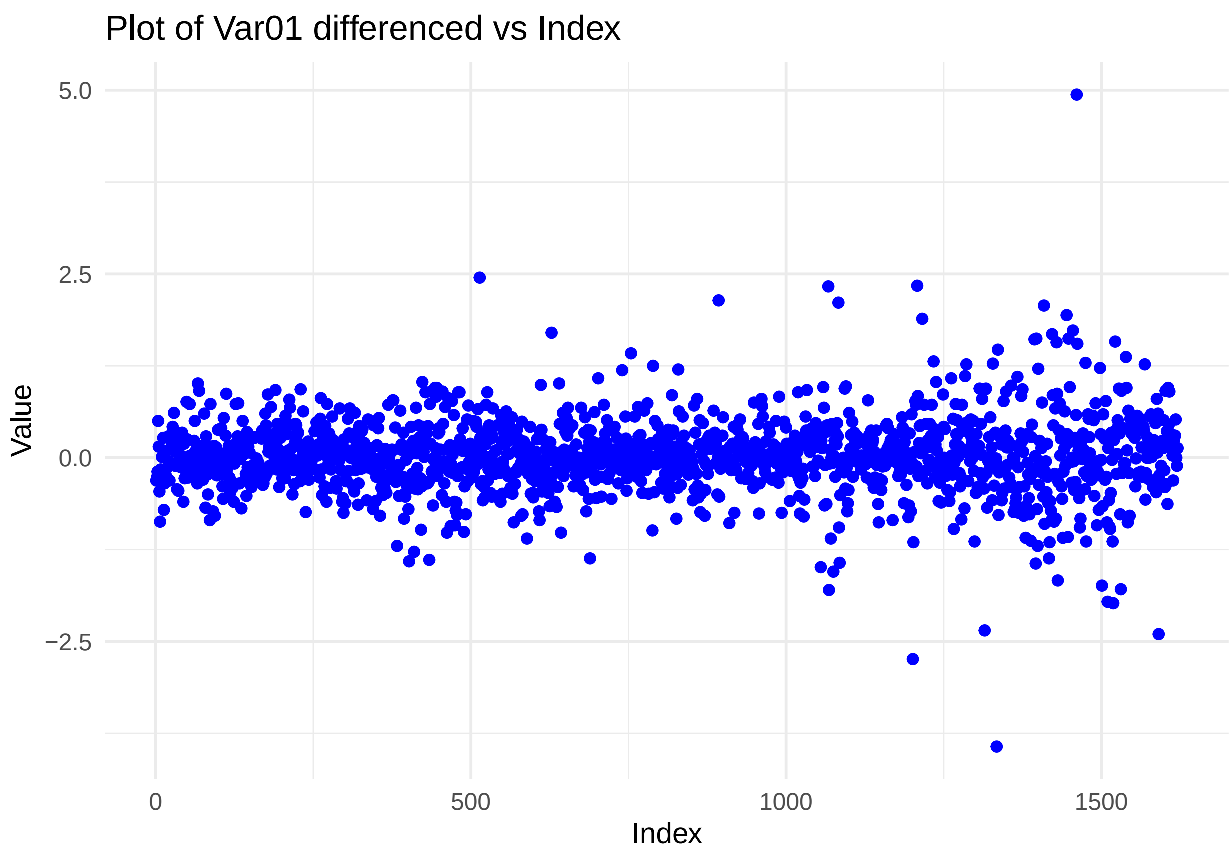
Var01 is non-stationary while Var02 is stationary.

We will preforming differencing to make Var01 stationary.

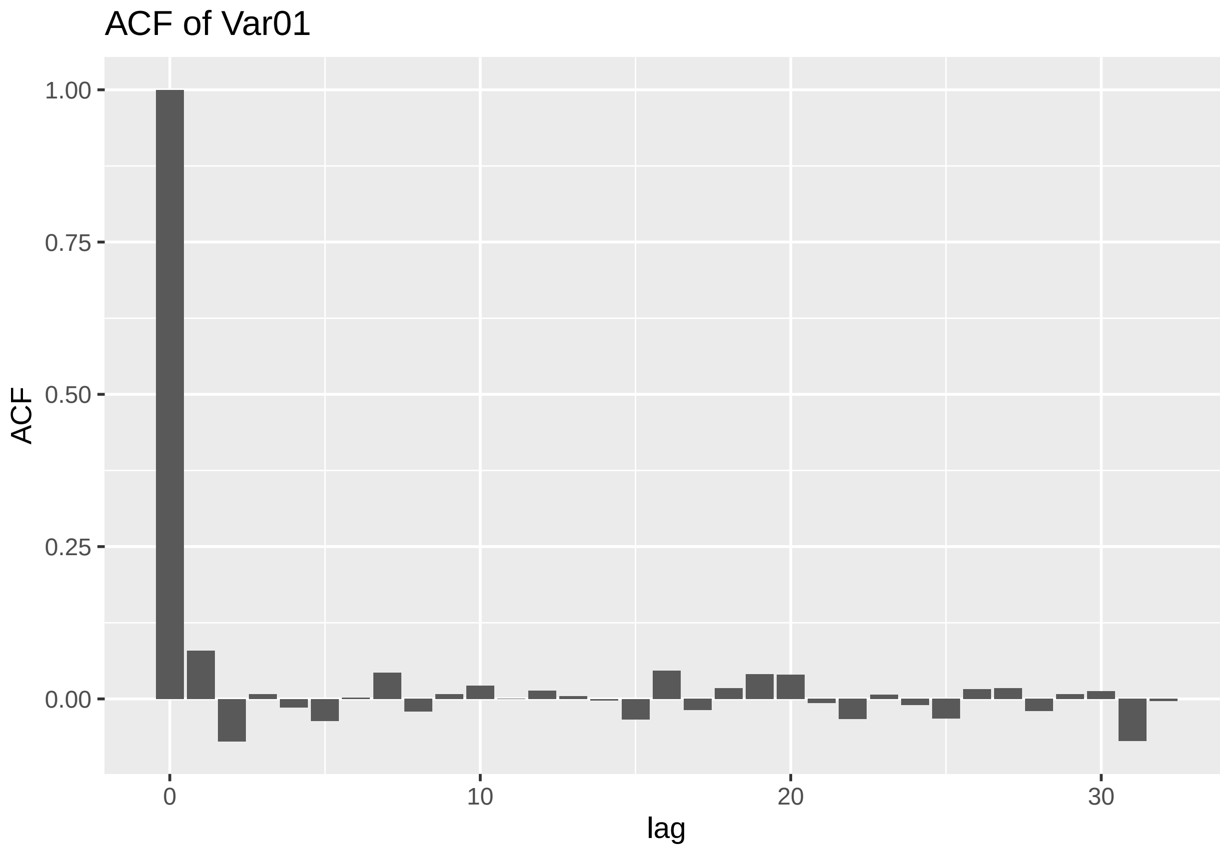
```r
var1_diff <- diff(s01$Var01[data_range], differences = 1)

var1_diff_df <- data.frame(Index = seq_along(var1_diff), Value = var1_diff)
```

```r
var1_plot <- ggplot(var1_diff_df, aes(x = Index, y = Value)) +
  geom_point(color = "blue") +
  labs(title = "Plot of Var01 differenced vs Index", x = "Index", y = "Value") +
  theme_minimal()
```

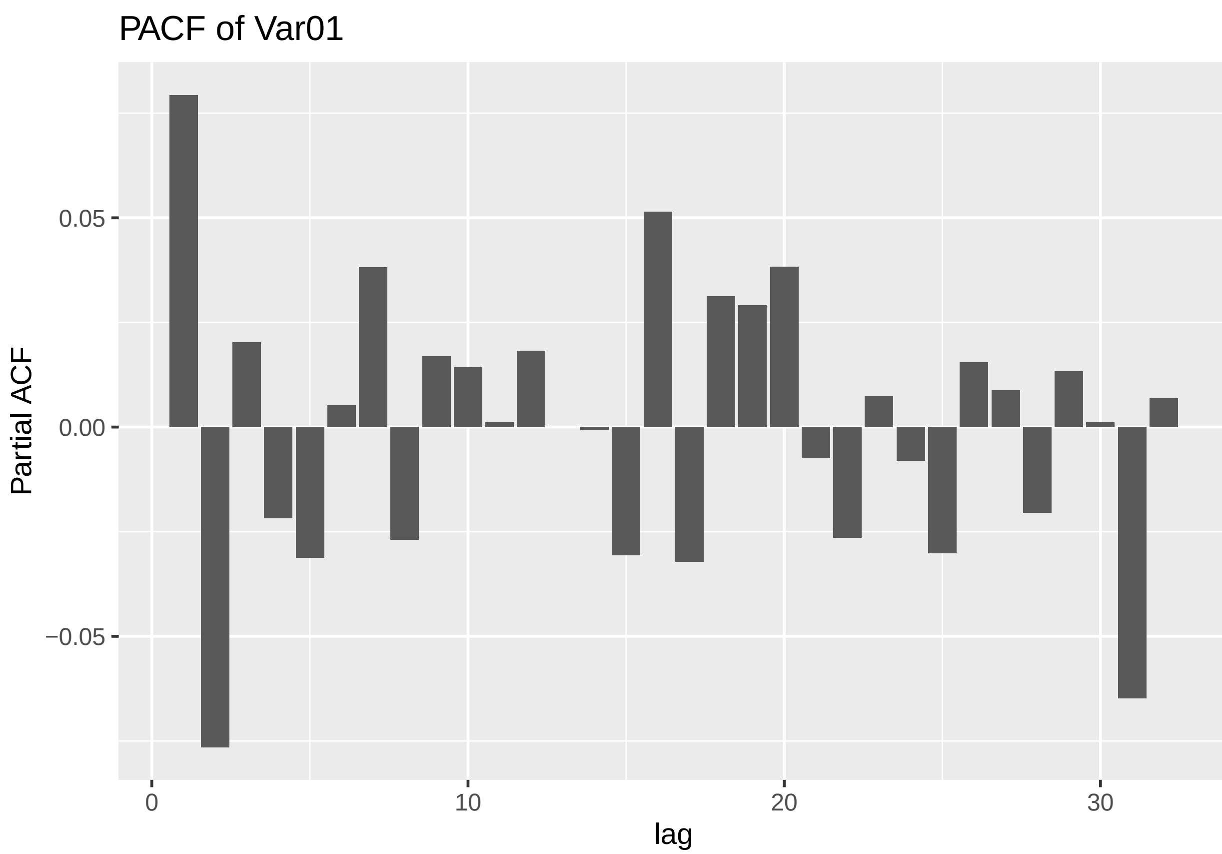
```r
grid.arrange(var1_plot, nrow = 1)
```




```
acf_var1 <- acf(var1_diff, plot = FALSE)

acf_var1_df <- data.frame(lag = acf_var1$lag, acf = acf_var1$acf)

acf1 <- ggplot(acf_var1_df, aes(x = lag, y = acf)) +
  geom_bar(stat = "identity") +
  labs(title = "ACF of Var01", y = 'ACF')

grid.arrange(acf1, nrow = 1)
```




```r
pacf_var1 <- pacf(var1_diff, plot = FALSE)

pacf_var1_df <- data.frame(lag = pacf_var1$lag, pacf = pacf_var1$acf)

pacf1 <- ggplot(pacf_var1_df, aes(x = lag, y = pacf)) +
  geom_bar(stat = "identity") +
  labs(title = "PACF of Var01", y = 'Partial ACF')

grid.arrange(pacf1, nrow=1)
```

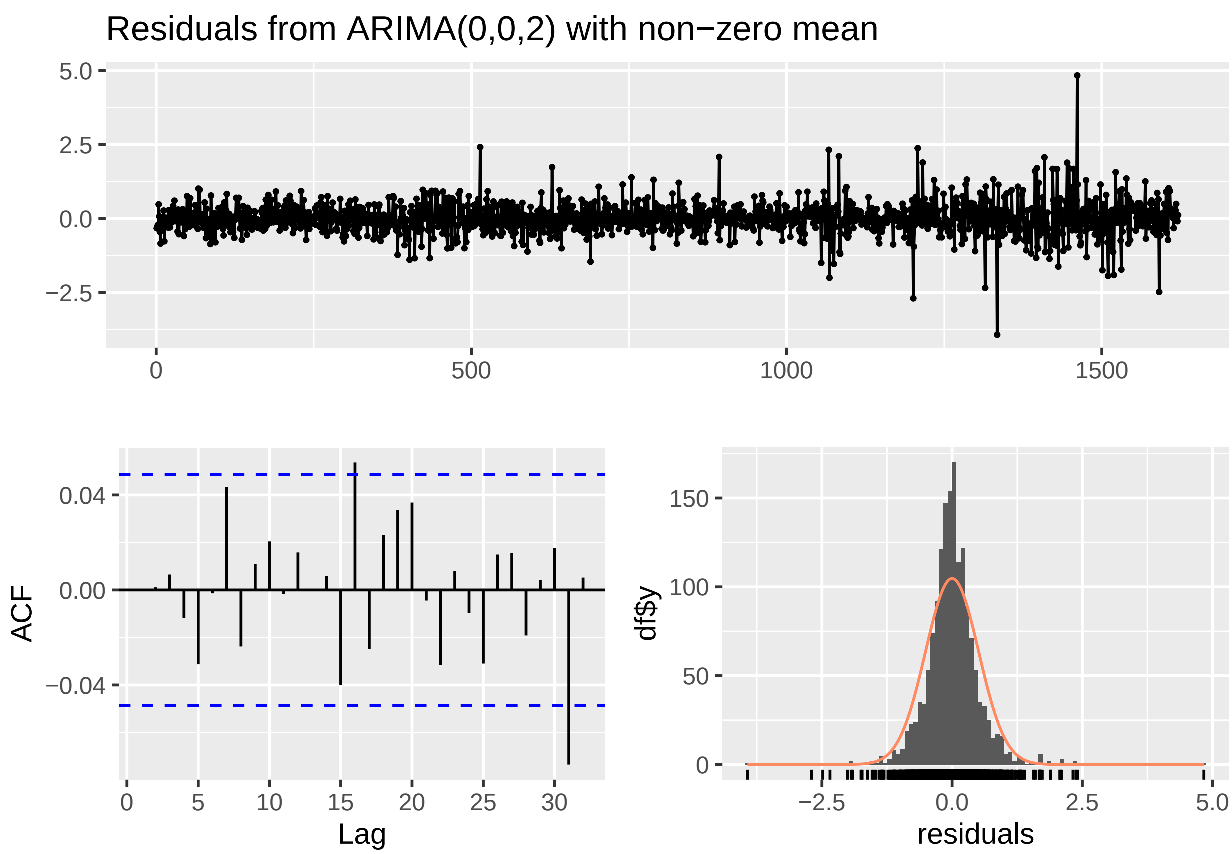


Forecasting

```
fit_var1 <- auto.arima(var1_diff, stationary = TRUE)
summary(fit_var1)
```

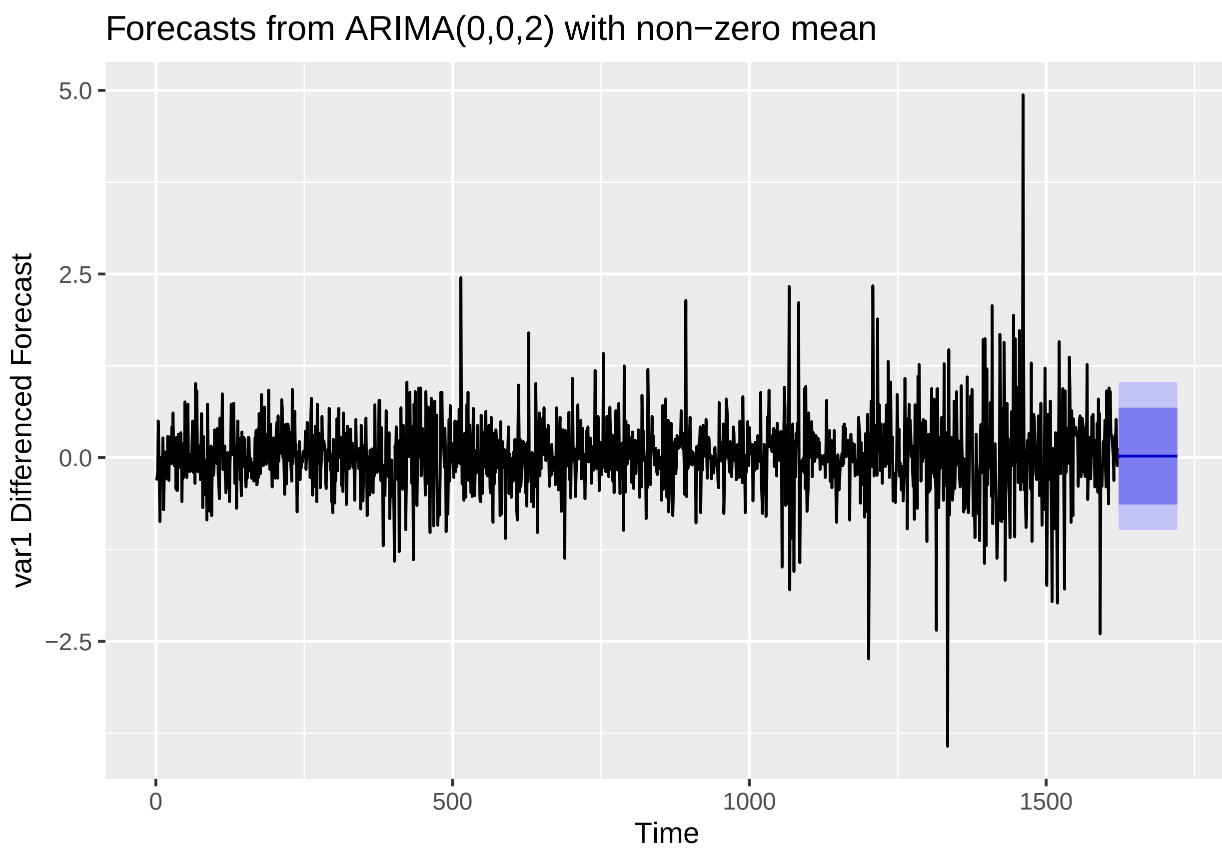
```
## Series: var1_diff
## ARIMA(0,0,2) with non-zero mean
##
## Coefficients:
##          ma1      ma2    mean
##       0.0875  -0.0731  0.0220
## s.e.  0.0248   0.0250  0.0129
##
## sigma^2 = 0.2612:  log likelihood = -1210.39
## AIC=2428.77   AICc=2428.79   BIC=2450.33
##
## Training set error measures:
##                        ME      RMSE       MAE MPE MAPE      MASE          ACF1
## Training set 2.867594e-06 0.5105564 0.3473101 NaN  Inf 0.7076254 -0.0003393583
```

```
checkresiduals(fit_var1)
```

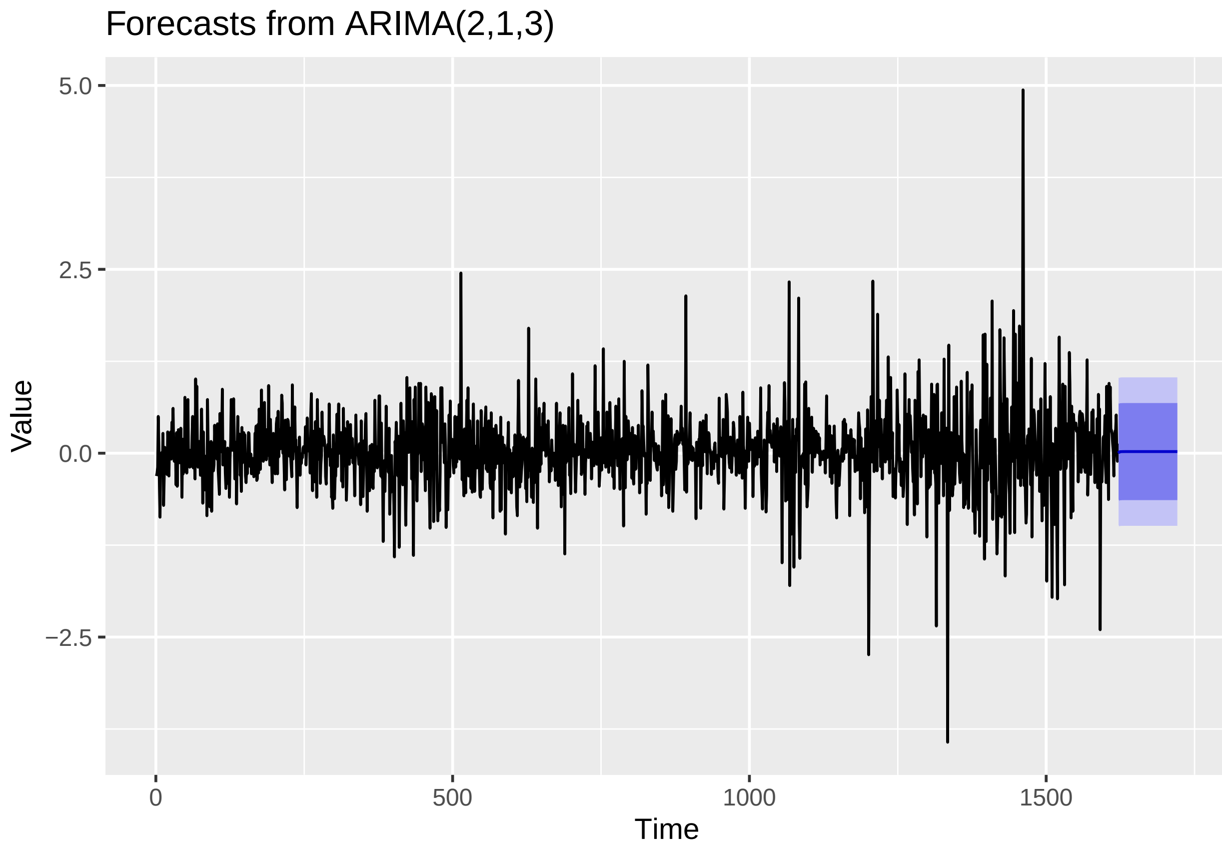



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(0,0,2) with non-zero mean
## Q* = 6.7689, df = 8, p-value = 0.5618
##
## Model df: 2.   Total lags used: 10
```

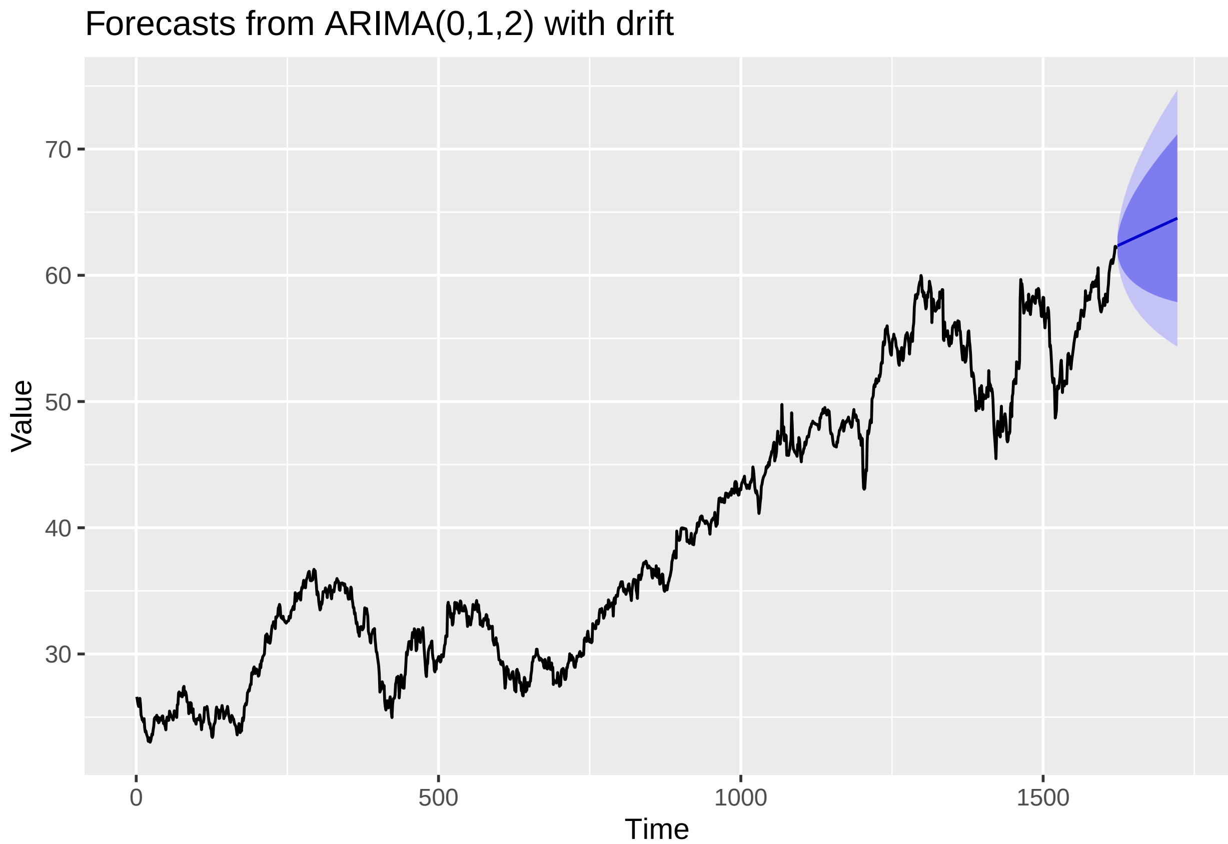
```
fc_var1 <- forecast(fit_var1, h=100)
autoplot(fc_var1) + ylab('var1 Differenced Forecast')
```



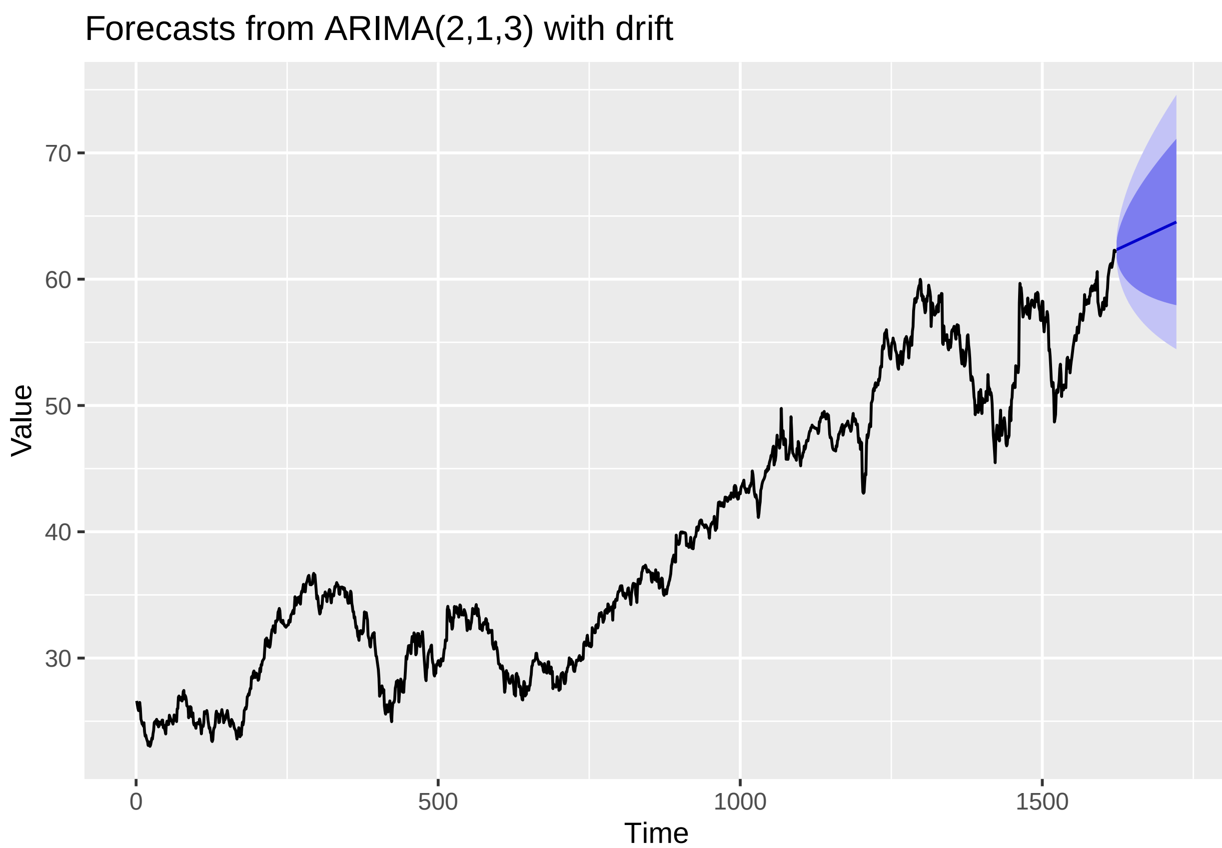

```
fit <- Arima(var1_diff, order=c(2,1,3), include.constant=FALSE)
fc <- forecast(fit, h=100)
autoplot(fc) + ylab('Value')
```




```
fit <- auto.arima(s01$Var01[data_range])
fc <- forecast(fit, h=100)
autoplot(fc) + ylab('Value')
```



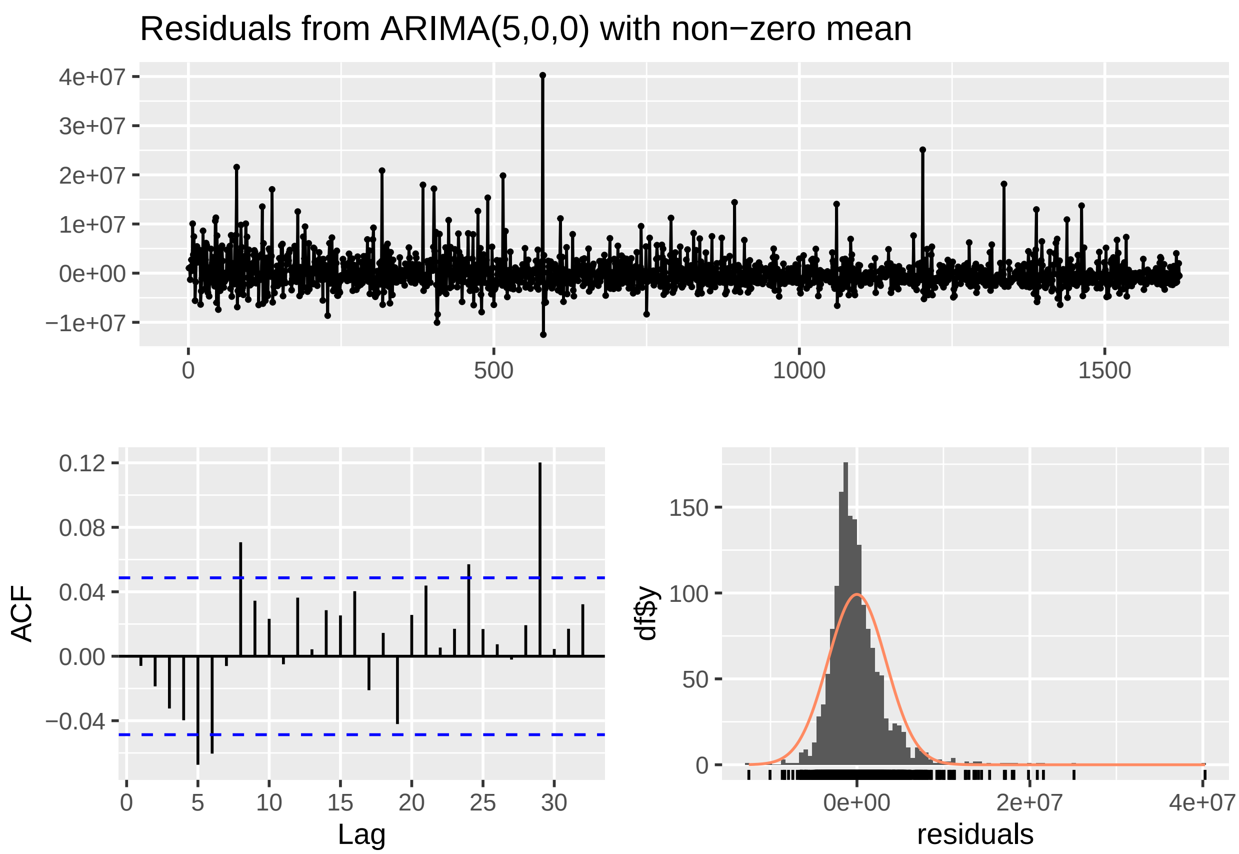

```r
fit <- Arima(s01$Var01[data_range], order=c(2,1,3), include.drift=TRUE)
fc <- forecast(fit, h=100)
autoplot(fc) + ylab('Value')
```




```
fit_var2 <- auto.arima(s01$Var02[data_range], stationary = TRUE)
summary(fit_var2)
```

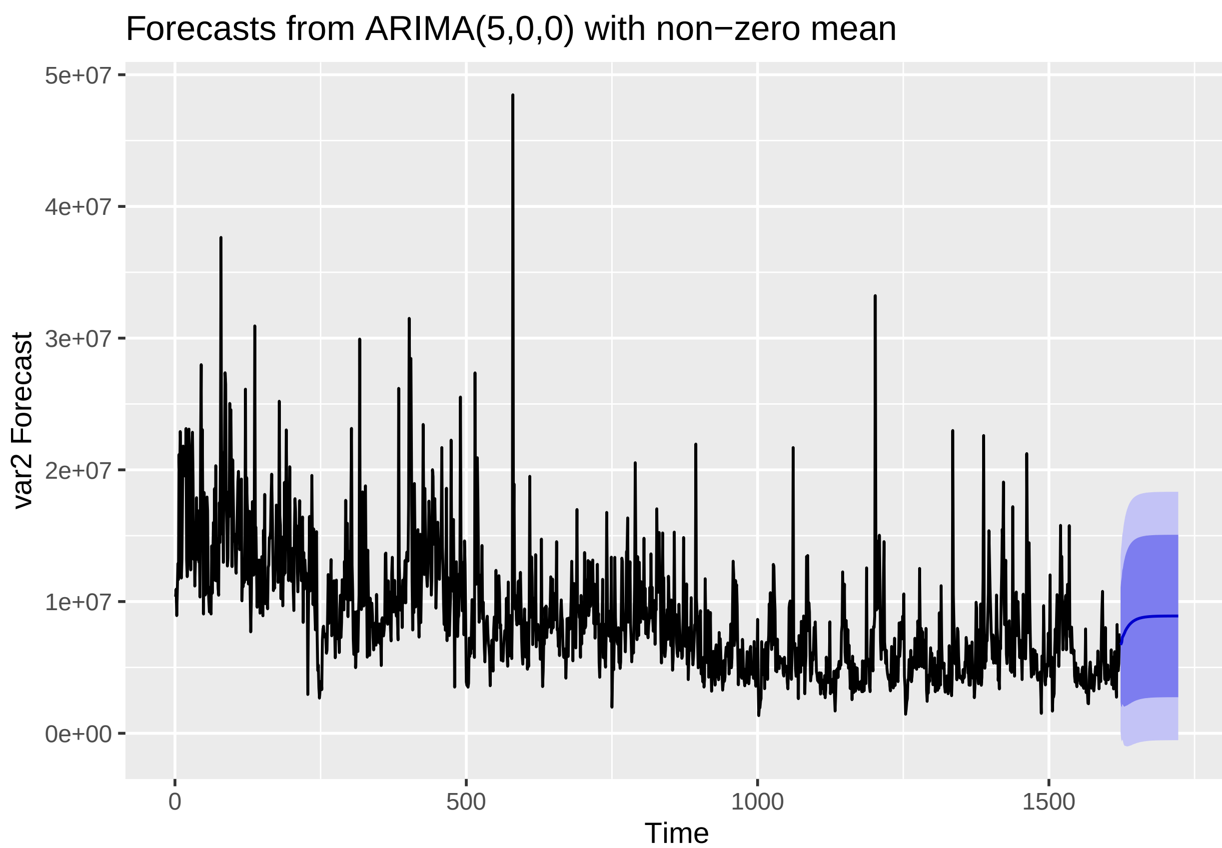
```
## Series: s01$Var02[data_range]
## ARIMA(5,0,0) with non-zero mean
##
## Coefficients:
##          ar1     ar2     ar3     ar4     ar5       mean
##       0.4398  0.1268  0.0875  0.0313  0.1514  8905853.1
## s.e.  0.0245  0.0268  0.0269  0.0268  0.0245   575822.3
##
## sigma^2 = 1.142e+13:  log likelihood = -26683.01
## AIC=53380.01   AICc=53380.08   BIC=53417.75
##
## Training set error measures:
##                     ME    RMSE     MAE       MPE     MAPE     MASE         ACF1
## Training set -3637.317 3373476 2243121 -12.14313 28.10484 0.888227 -0.005976824
```

```
checkresiduals(fit_var2)
```

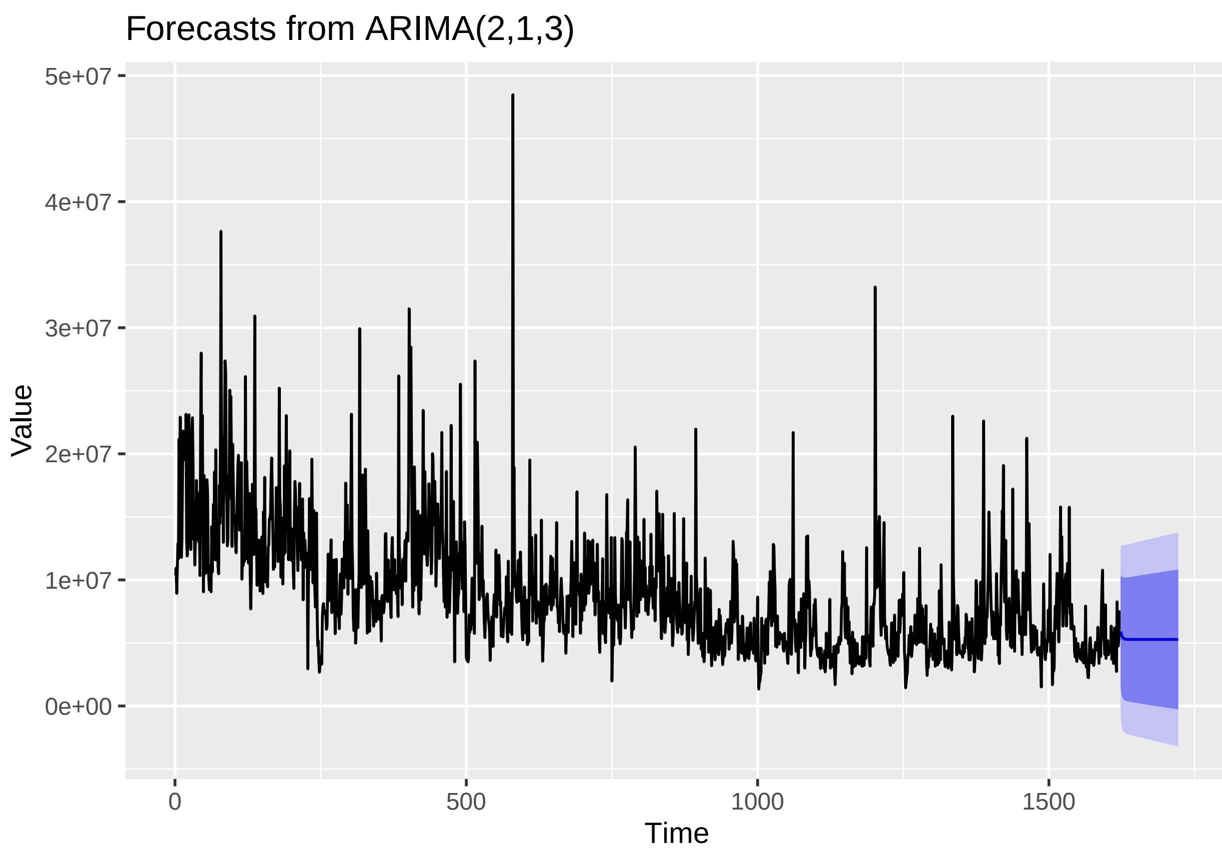



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(5,0,0) with non-zero mean
## Q* = 29.281, df = 5, p-value = 2.043e-05
##
## Model df: 5.   Total lags used: 10
```

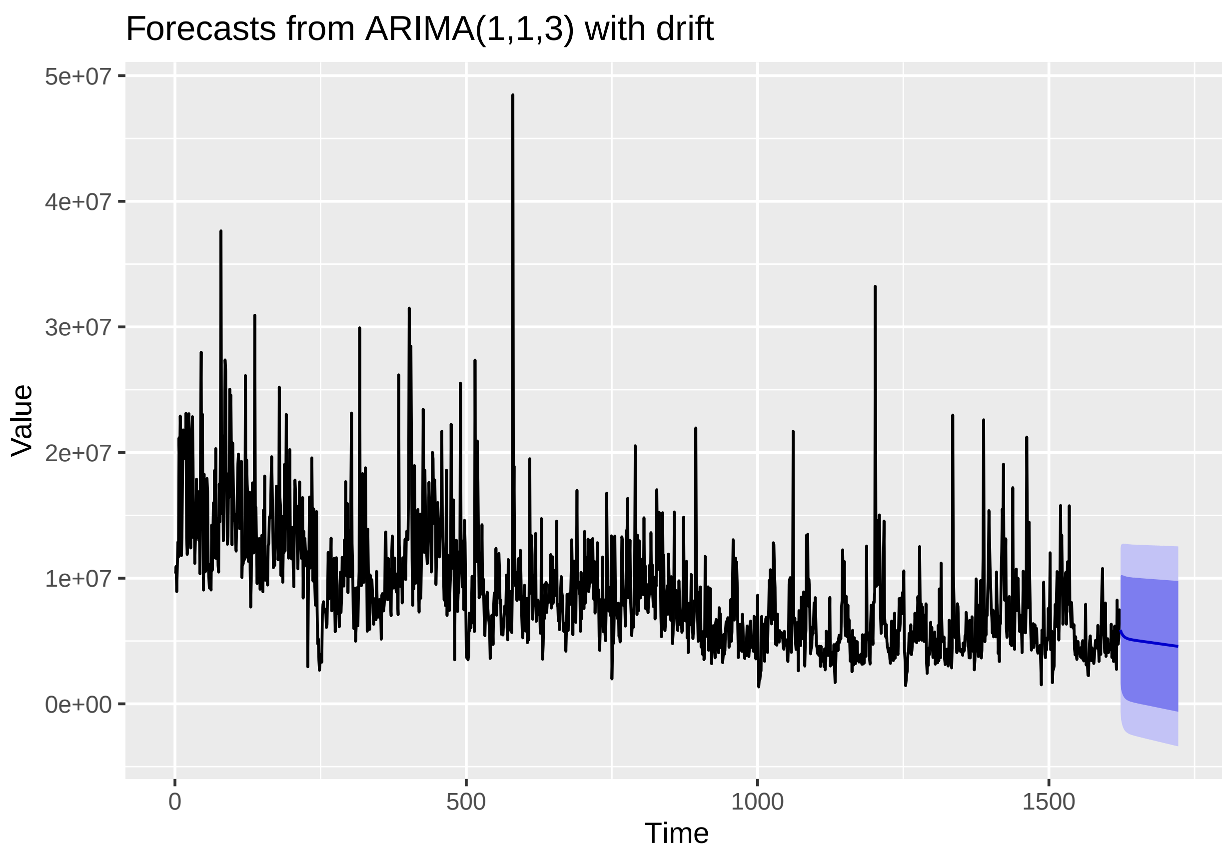
```
fc_var2 <- forecast(fit_var2, h=100)
autoplot(fc_var2) + ylab('var2 Forecast')
```



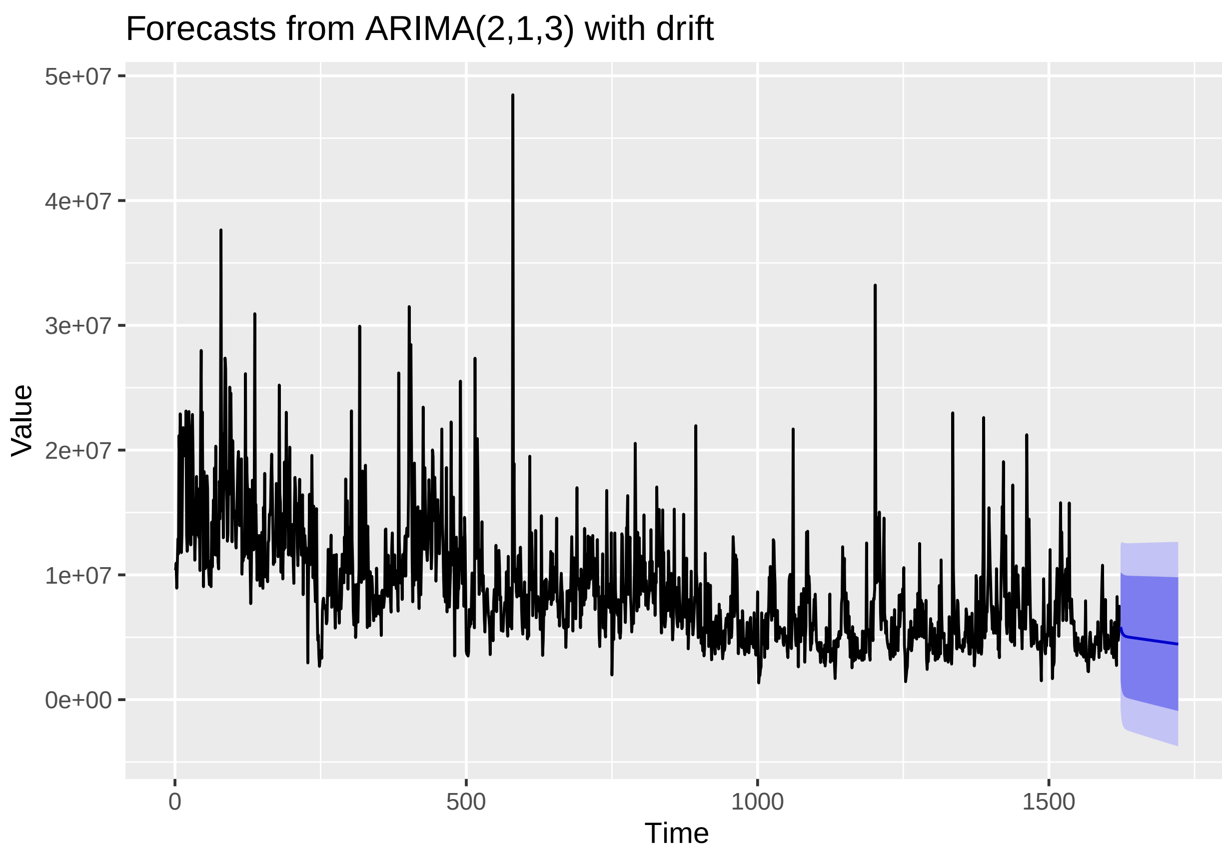

```
fit <- Arima(s01$Var02[data_range], order=c(2,1,3), include.constant=FALSE)
fc <- forecast(fit, h=100)
autoplot(fc) + ylab('Value')
```




```r
fit <- auto.arima(s01$Var02[data_range])
fc <- forecast(fit, h=100)
autoplot(fc) + ylab('Value')
```




```r
fit <- Arima(s01$Var02[data_range], order=c(2,1,3), include.drift=TRUE)
fc <- forecast(fit, h=100)
autoplot(fc) + ylab('Value')
```



Forecasts from ARIMA(2,1,3) with drift
5e+07
4e+07
3e+07
2e+07
1e+07
0e+00
Value
0
500
1000
1500
Time