Modélisation du taux de chômage

Yannick Le Pen

2023-2024

Importation de la série du taux de chômage à partir de DB.nomics

On importe les observations trimestrielles du taux de chômage aux Etats-Unis du 1994:T1 au 23:T1 (117 observations) à partir du site DB.nomics. Les données proviennent du Bureau of Labor Statistics.

```
rm(list = ls())
library(data.table)
library(rdbnomics)
df<-rdb(ids = "BLS/ln/LNS13327708Q") # ids = identifiant de la série sur DB.nomics
```

On crée un dataframe contenant uniquement les dates (colonne "period") et la série du taux de chômage (colonne "value").

```
df_u<-df[,c("period","value")]
colnames(df_u)<-c('date','taux')
head(df_u)
```

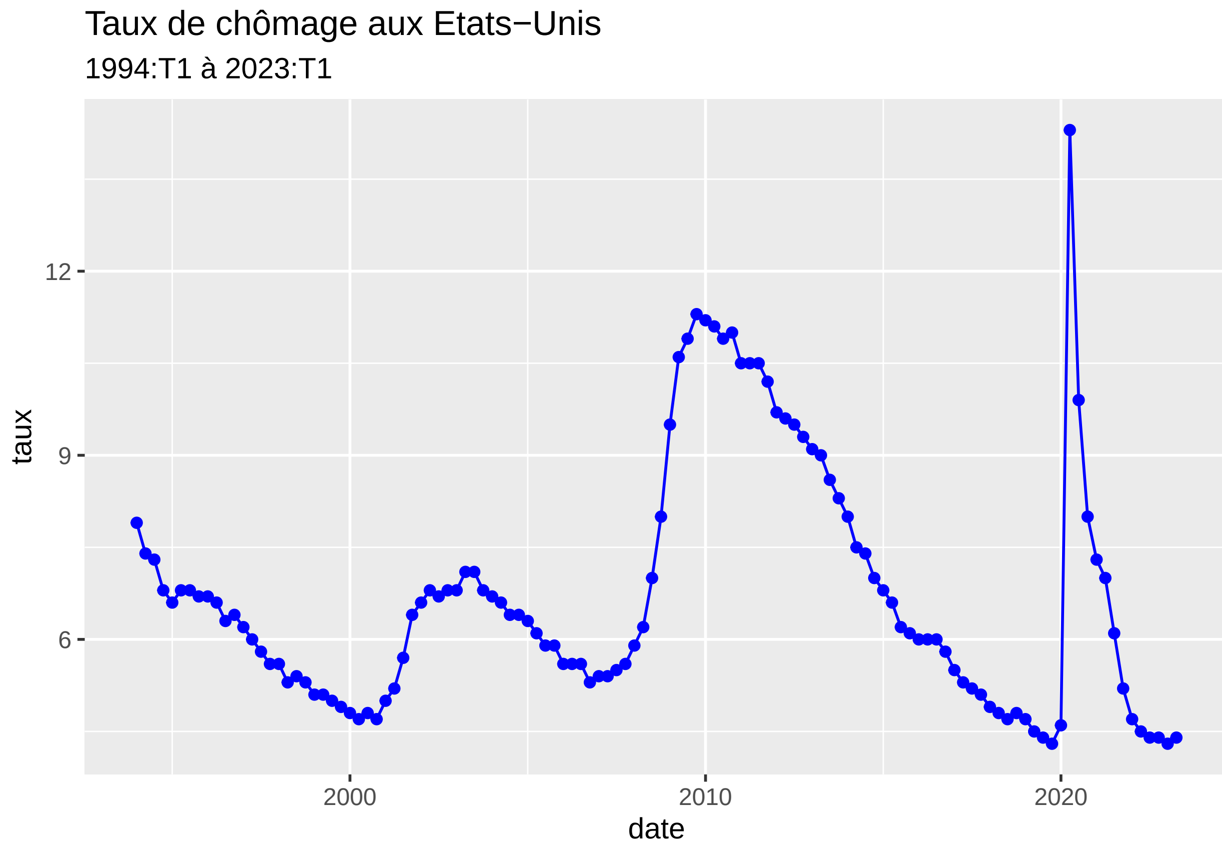
```
##          date taux
## 1: 1994-01-01  7.9
## 2: 1994-04-01  7.4
## 3: 1994-07-01  7.3
## 4: 1994-10-01  6.8
## 5: 1995-01-01  6.6
## 6: 1995-04-01  6.8
```

```
tail(df_u,n=5)
```

```
##          date taux
## 1: 2022-04-01  4.5
## 2: 2022-07-01  4.4
## 3: 2022-10-01  4.4
## 4: 2023-01-01  4.3
## 5: 2023-04-01  4.4
```

Représentation graphique Le taux de chômage aux Etats-Unis fluctue autour de 6% du début de l'échantillon jusqu'à l'année 2008 à partir de laquelle il augmente très rapidement jusqu'au niveau de 11%, sous l'effet de la crise des subprimes. On observe aussi l'épisode de la crise du Covid avec une hausse instantanée du taux de chômage autour de 14% suivie d'un retour rapide à des niveaux très bas. La période du covid peut avoir un impact sur les résultats des tests statistiques. Au vu du graphique, il est assez difficile de trancher sur la stationnarité du taux de chômage.

```
library(ggplot2)
p_us<-ggplot(data=df_u,aes(x=date,y=taux))+geom_point(color='blue')+geom_line(color='blue')
p_us+labs(x="date",y="taux",title = "Taux de chômage aux Etats-Unis",subtitle = "1994:T1 à 2023:T1")
```

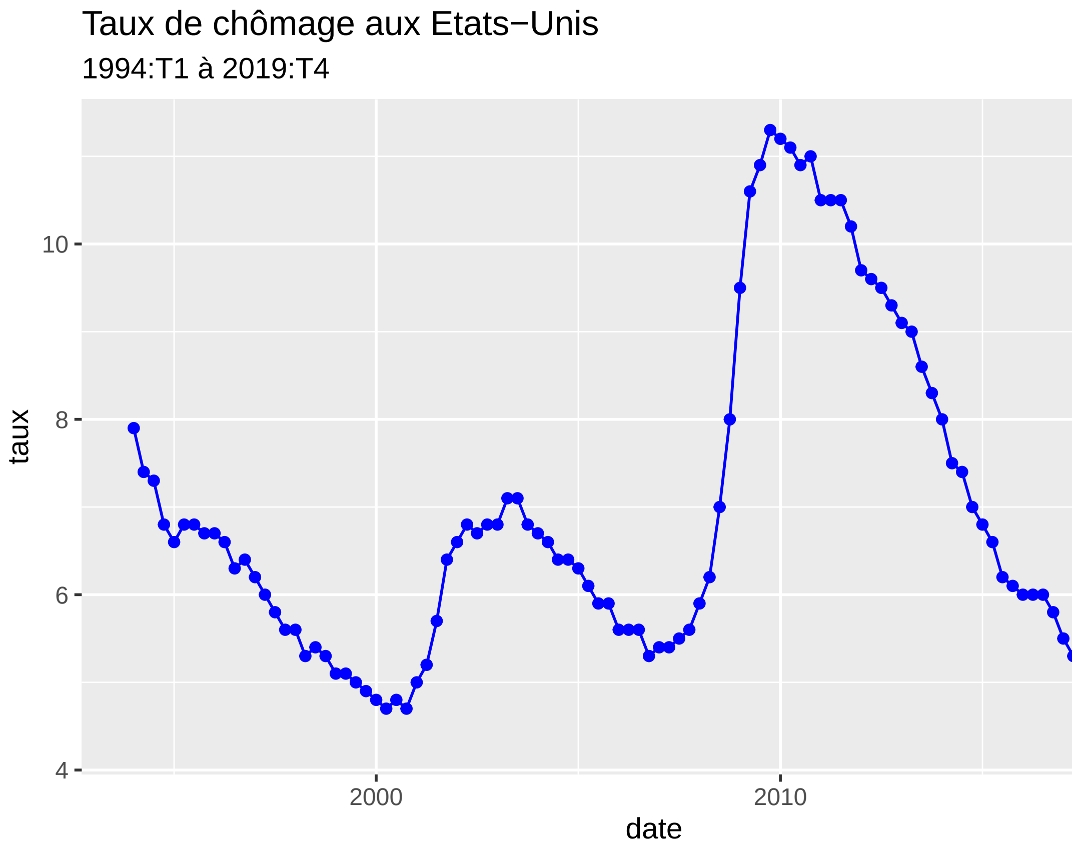


Estimation sans la période Covid : 1994:T1 à 2019:T4

Dans un premier temps, on exclut les observations à partir du dernier trimestre de l'année 2019.

```
df_u_19T1=df_u[df_u$date<="2019-10-01",]
```

```
library(ggplot2)
p_us<-ggplot(data=df_u_19T1,aes(x=date,y=taux))+geom_point(color='blue')+geom_line(color='blue')
p_us+labs(x="date",y="taux",title = "Taux de chômage aux Etats-Unis",subtitle = "1994:T1 à 2019:T4")
```



Représentation graphique

Le graphique montre que pendant la grande récession de 2008-2009, le taux de chômage a atteint des niveaux nettement plus élevés que depuis le début de l'échantillon.

```
statdes0<-summary(df_u_19T1$taux)
sprintf("Ecart type %f",sd(df_u_19T1$taux))
```

Statistiques descriptives et autocorrélogrammes

```
## [1] "Ecart type 1.848130"
```

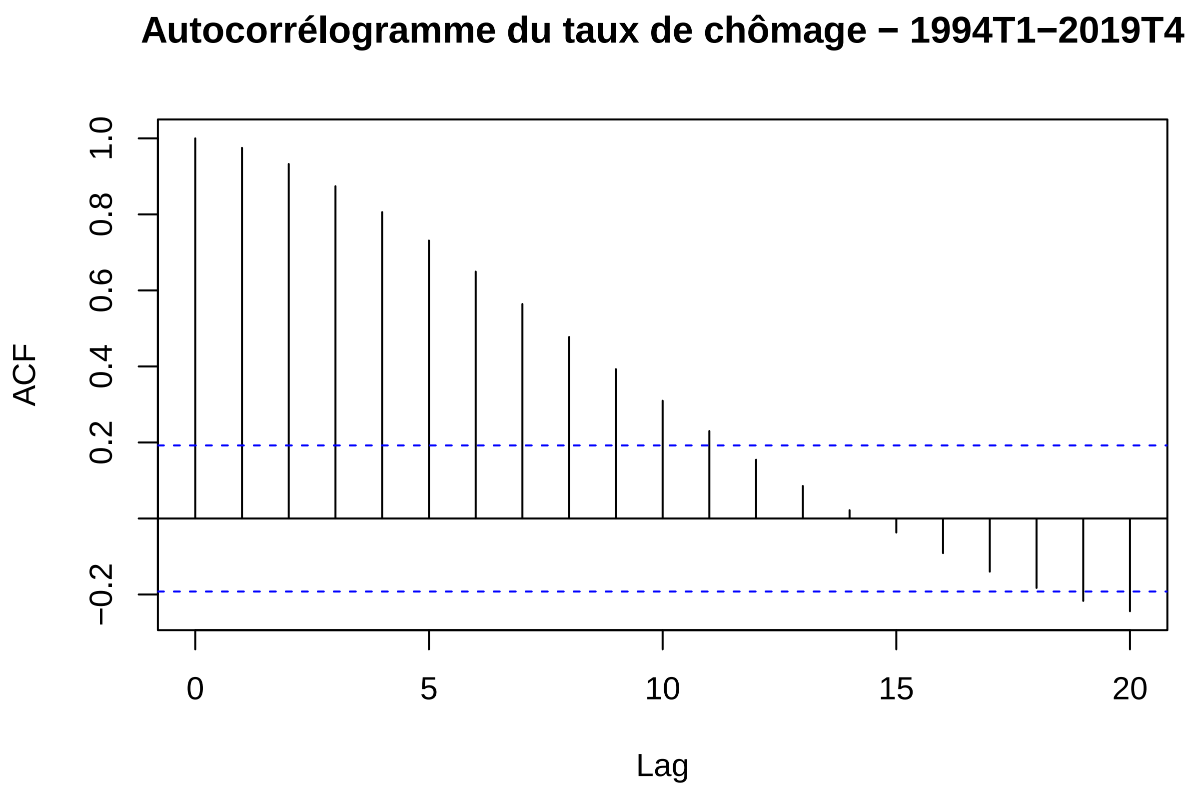
La library FinTS permet aussi de calculer des statistiques descriptives d'une série temporelle.

```
library(FinTS)
stat_des1<-FinTS.stats(df_u_19T1$taux)
stat_des1
```

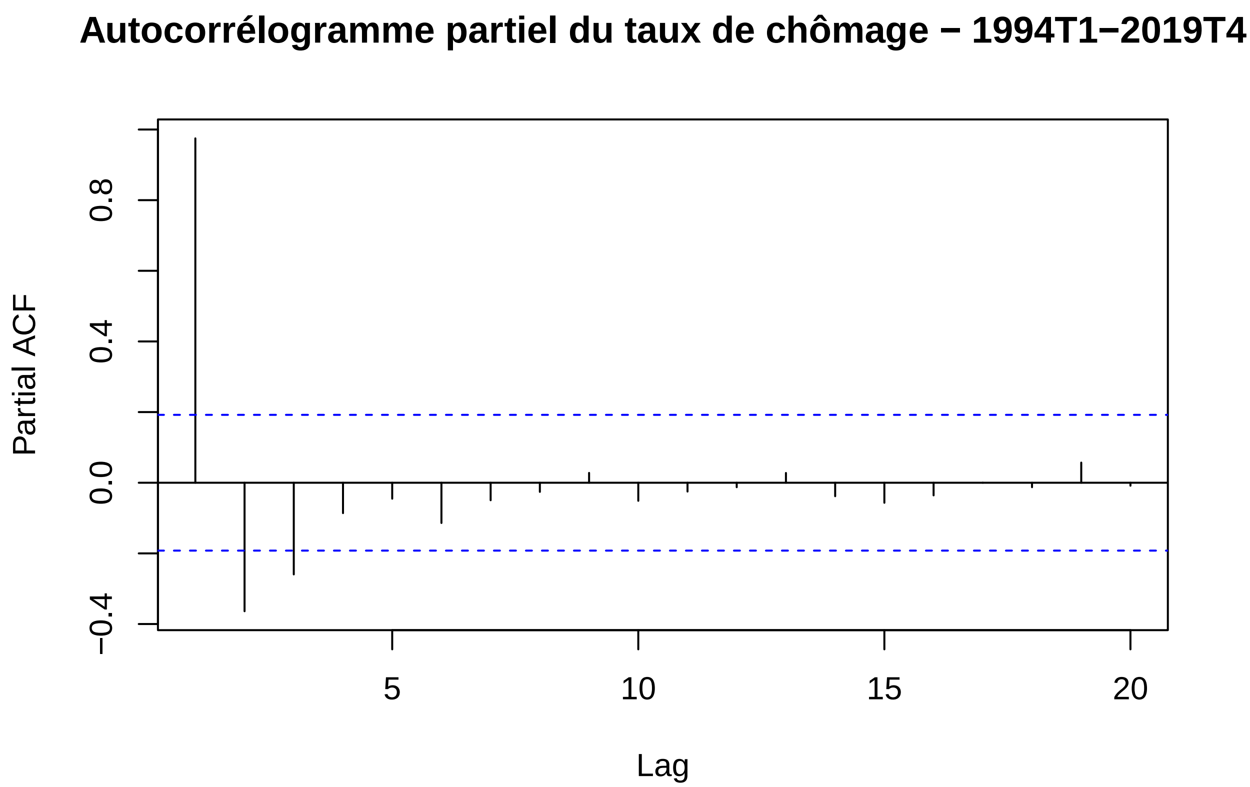
```
##   Start Size     Mean Standard.Deviation Skewness Excess.Kurtosis Minimum
## 1     1  104 6.768269            1.84813 1.038192      0.07256283     4.3
##   Maximum
## 1    11.3
```

‘

```
acf_u_19T1<-acf(x=df_u_19T1$taux,main='Autocorrélogramme du taux de chômage - 1994T1-2019T4')
```




```
pacf(x=df_u_19T1$taux,main='Autocorrélogramme partiel du taux de chômage - 1994T1-2019T4')
```



Commentaires

- Les autocorrélations sont positives et significativement jusqu'au retard 11.
- Elles sont proches de 1 pour les retards 1, 2, 3, 4 5 puis décroissent vers 0. On en déduit que le taux de chômage est assez autocorrélé. Néanmoins, il est assez difficile de tirer une conclusions sur la stationnarité ou non-stationnarité du taux de chômage. le profil des autocorrélations pourrait correspondre à une série stationnaire.
- Les autocorrélations partielles sont significatives pour les retards 1, 2 et 3.

Le profil des autocorrélations suggère d'esimer un modèle AR(3)

Estimation et tests de validation d'un modèle ARMA

On va écarter les 4 dernières observations de l'échantillon, c'est-à-dire l'année 2019. Dans notre cas, cela revient à estimer le modèle avec les 100 premières observations. L'objectif est de comparer les valeurs prédites par le modèle final aux valeurs réalisées. Ainsi, on pourra évaluer la qualité des prévisions en-dehors de l'échantillon ("out-of-sample) utilisé pour l'estimer.

On utilise la librairie forecast. En se basant sur l'analyse des autocorrélations, on spécifie un modèle AR(3) :

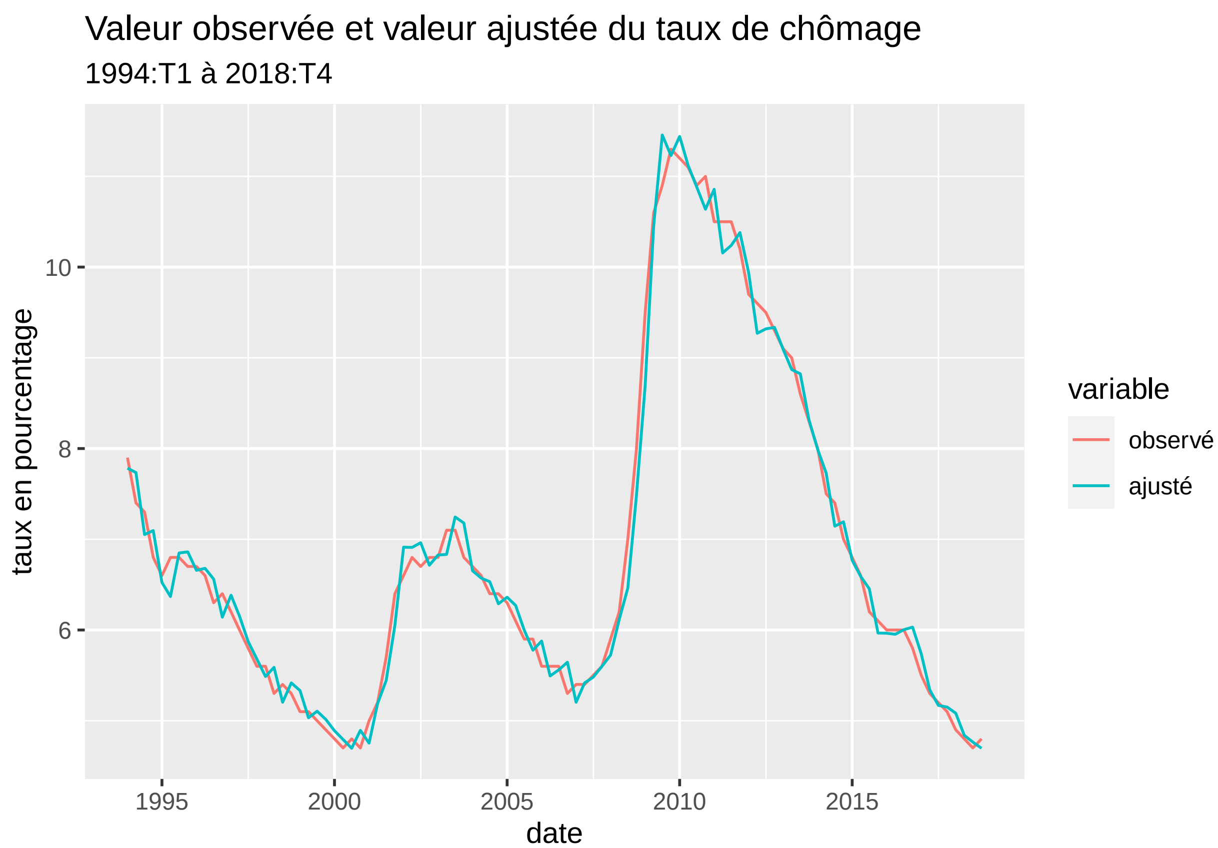
- order=c(p,d,q) où :
  - p : nombre de retards de la partie AR,
  - d : nombre de différenciations pour obtenir un série stationnaire,
  - q : nombre de retards pour la partie MA
- Arima inclut une constante par défaut

```
library(forecast)
# Estimation d'un  modèle
```

```r
AR3<-Arima(y = df_u_19T1$taux[1:100],order=c(3,0,0))
# Affichage de l'estimation
summary(AR3)
```

```
## Series: df_u_19T1$taux[1:100]
## ARIMA(3,0,0) with non-zero mean
##
## Coefficients:
##          ar1      ar2      ar3    mean
##       1.4712  -0.2221  -0.2827  6.9381
## s.e.  0.0966   0.1777   0.0971  0.6218
##
## sigma^2 = 0.05001:  log likelihood = 7.4
## AIC=-4.8   AICc=-4.16   BIC=8.22
##
## Training set error measures:
##                        ME      RMSE       MAE       MPE     MAPE     MASE
## Training set -0.004696549 0.2191125 0.1681198 -0.185275 2.454551 0.774133
##                     ACF1
## Training set -0.02262365
```

```r
library(ggplot2)
library(reshape2)

df_plot<-data.frame(df_u_19T1$date[1:100],df_u_19T1$taux[1:100],F=as.matrix(AR3$fitted))
colnames(df_plot)<-c("date","observé","ajusté")

data_melt=melt(as.data.frame(df_plot),id.vars = 1)

p_F_us<-ggplot(data=data_melt,aes(x=date))+geom_line(aes(x=date,y=value,color= variable))
p_F_us+labs(y="taux en pourcentage",title = "Valeur observée et valeur ajustée du taux de chômage",
            subtitle = "1994:T1 à 2018:T4")
```



Test de significativité des coefficients

1. Avec la librairie lmtest

```
library(lmtest)
coeftest(AR3)
```

```
##
## z test of coefficients:
##
##             Estimate Std. Error z value  Pr(>|z|)
## ar1         1.471196   0.096591 15.2312 < 2.2e-16 ***
## ar2        -0.222085   0.177671 -1.2500  0.211308
## ar3        -0.282668   0.097106 -2.9109  0.003604 **
## intercept   6.938081   0.621805 11.1580 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Commentaires

1. On vérifie la significativité de la constante :

$$\begin{cases} H_0 \ : \ c = 0 \\ H_a \ : \ c \neq 0 \end{cases}$$

- La statistique de test $t_{\hat{c}} = \frac{\hat{c}}{\hat{\sigma}_{\hat{c}}} \to N(0,1)$ sous H_0 car On dispose de 104 observations.
- Les seuils critiques sont :
- `seuil = 2.57 pour un risque de première espèce égal à 1\%`

- `seuil = 1.96 pour un risque de première espèce égal à 5\%`
- $t_{\hat{c}} = 10.107 > 2.57$: la constante est significative pour un risque de première espèce de 1 %
- La probabilité critique est inférieure à 1% pour la constante.

2.On vérifie la significativité celle des coefficients de l'AR(3) :

$$\begin{cases} H_0 : \phi_i = 0 \\ H_a : \phi_i \neq 0 \end{cases}$$

- $|t_{\hat{\phi}_1}| > 2.57$ et $|t_{\hat{\phi}_3}| > 2.57$: les coefficients de retards 1 et 3 de l'AR sont significativement différents de 0 pour un risque de première espèce de 1%.
- Les probabilités critiques pour $\hat{\phi}_1$ et $\hat{\phi}_3$ sont inférieures à 1% : on rejette l'hypothèse nulle de non-significativité pour un risque de première espèce de 1%.
- $|t_{\hat{\phi}_2}| = 1.33 < 1.96$: le coefficient du second retard de l'AR(3) n'est pas significativement différent de 0
- La probabilité critique de $\hat{\phi}_2$ est égale à 21.1 % : ce coefficient n'est pas significatif pour un risque de 5%.
- Le coefficient $\hat{\phi}_3$ du dernier retard est significativement différent de 0 : il n'y a pas de retard superflu.

2. Calcul des tstat directement

```
AR3$tstat<-AR3$coef/sqrt(diag(AR3$var.coef))# tstat of the ARMA coefficient
AR3$tstat
```

```
##       ar1       ar2       ar3 intercept
## 15.231223 -1.249978 -2.910920 11.157962
```

Estimation de spécifications alternatives On estime un AR(4) pour vérifier s'il n'est pas nécessaire d'ajouter des retards supplémentaires

```
AR4<-Arima(y = df_u_19T1$taux[1:100],order=c(4,0,0))
summary(AR4)
```

```
## Series: df_u_19T1$taux[1:100]
## ARIMA(4,0,0) with non-zero mean
##
## Coefficients:
##          ar1      ar2      ar3      ar4    mean
##       1.4609  -0.2332  -0.2223  -0.0401  6.9451
## s.e.  0.1000   0.1799   0.1813   0.1018  0.6020
##
## sigma^2 = 0.05046:  log likelihood = 7.48
## AIC=-2.96   AICc=-2.05   BIC=12.67
##
## Training set error measures:
##                        ME      RMSE       MAE        MPE    MAPE     MASE
## Training set -0.004596516 0.2189354 0.1670426 -0.1843019 2.44139 0.769173
##                     ACF1
## Training set -0.00766017
```

```
coeftest(AR4)
```

```
##
## z test of coefficients:
##
##            Estimate Std. Error z value Pr(>|z|)
## ar1        1.460889   0.100021 14.6058   <2e-16 ***
## ar2       -0.233222   0.179885 -1.2965   0.1948
```

```
## ar3       -0.222349   0.181283 -1.2265   0.2200
## ar4       -0.040142   0.101844 -0.3941   0.6935
## intercept  6.945071   0.602005 11.5366   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Commentaires

- On voit que le coefficient du quatrième retard ar4 n'est pas significativement différent de 0 $|t_{\hat{\phi}_4}| = 0.693 < 1.96$

On estime un ARMA(3,1) pour vérifier s'il n'est pas nécessaire d'ajouter des retards MA supplémentaires

```r
ARMA31<-Arima(y = df_u_19T1$taux[1:100],order=c(3,0,1))
summary(ARMA31)
```

```
## Series: df_u_19T1$taux[1:100]
## ARIMA(3,0,1) with non-zero mean
##
## Coefficients:
##          ar1      ar2      ar3      ma1    mean
##       1.5409  -0.3383  -0.2344  -0.0758  6.9422
## s.e.  0.2680   0.4573   0.2022   0.2653  0.6101
##
## sigma^2 = 0.05049:  log likelihood = 7.44
## AIC=-2.88   AICc=-1.98   BIC=12.75
##
## Training set error measures:
##                        ME      RMSE       MAE        MPE     MAPE      MASE
## Training set -0.004633258 0.2190182 0.1675776 -0.1846136 2.447797 0.7716363
##                     ACF1
## Training set -0.01420855
```

```r
coeftest(ARMA31)
```

```
##
## z test of coefficients:
##
##            Estimate Std. Error z value  Pr(>|z|)
## ar1        1.540930   0.267982  5.7501 8.918e-09 ***
## ar2       -0.338255   0.457332 -0.7396    0.4595
## ar3       -0.234379   0.202171 -1.1593    0.2463
## ma1       -0.075757   0.265319 -0.2855    0.7752
## intercept  6.942189   0.610092 11.3789 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Commentaires

- On voit que le coefficient du premier retard ma1 n'est pas significativement différent de 0 $|t_{\hat{\phi}_4}| = 0.7752 < 1.96$. Il n'est pas nécessaire d'ajouter un retard MA.

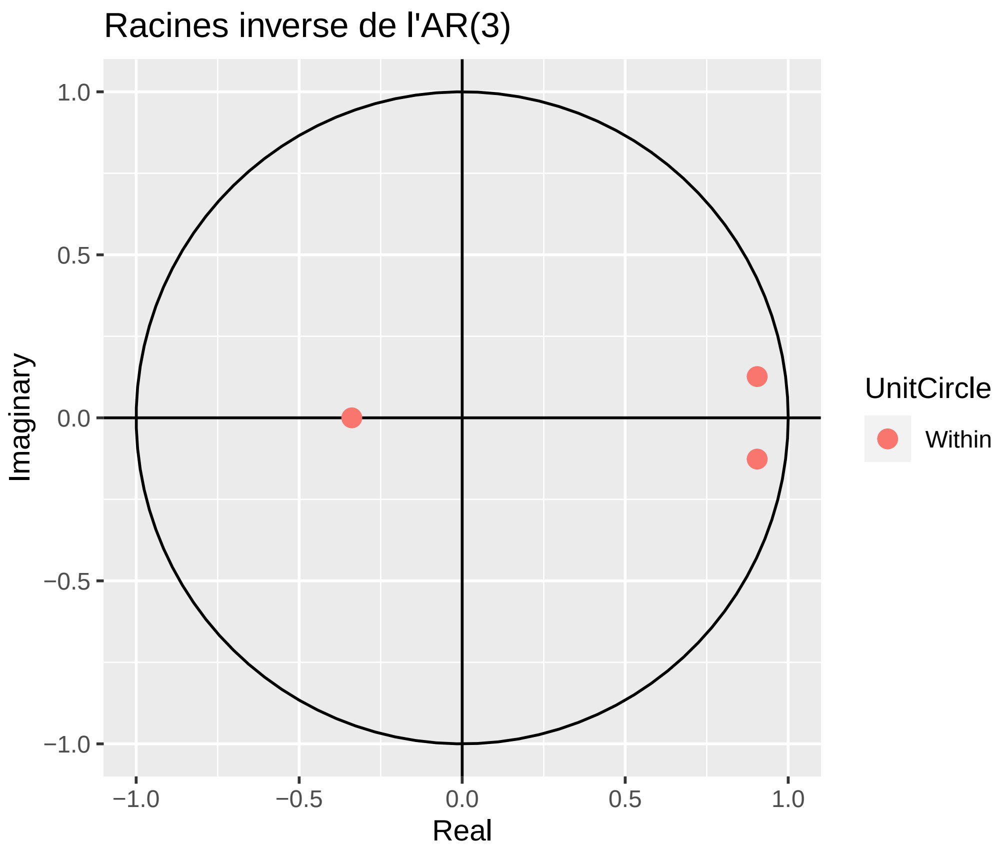
Interprétation des coefficients Dans la présentation de l'estimation, la constante reportée représente l'espérance de la série. Dans notre exemple les résultats peuvent s'écrire comme :

$$\begin{cases} US\_taux_t = 6.9381 + \hat{u}_t \\ \hat{u}_t = 1.4712 \times \hat{u}_{t-1} - 0.2221 \times \hat{u}_{t-2} - 0.2827 \times \hat{u}_{t-3} + \hat{\epsilon}_t \end{cases}$$

$6.9381 = E(US\_taux_t)$ et $\hat{u}_t = US\_taux_t - 6.9381$ est le taux de chômage corrigé de sa moyenne.

Analyse des racines On représente les **inverses** des racines dans le cercle unitaire. La condition de stationnarité impose que ces inverses doivent être de module strictement inférieur à 1, c'est-à-dire dans le cercle unitaire. La condition de stationnarité est donc satisfaite.

```
autoplot(AR3,main="Racines inverse de l\'AR(3)")
```



Tests sur les résidus

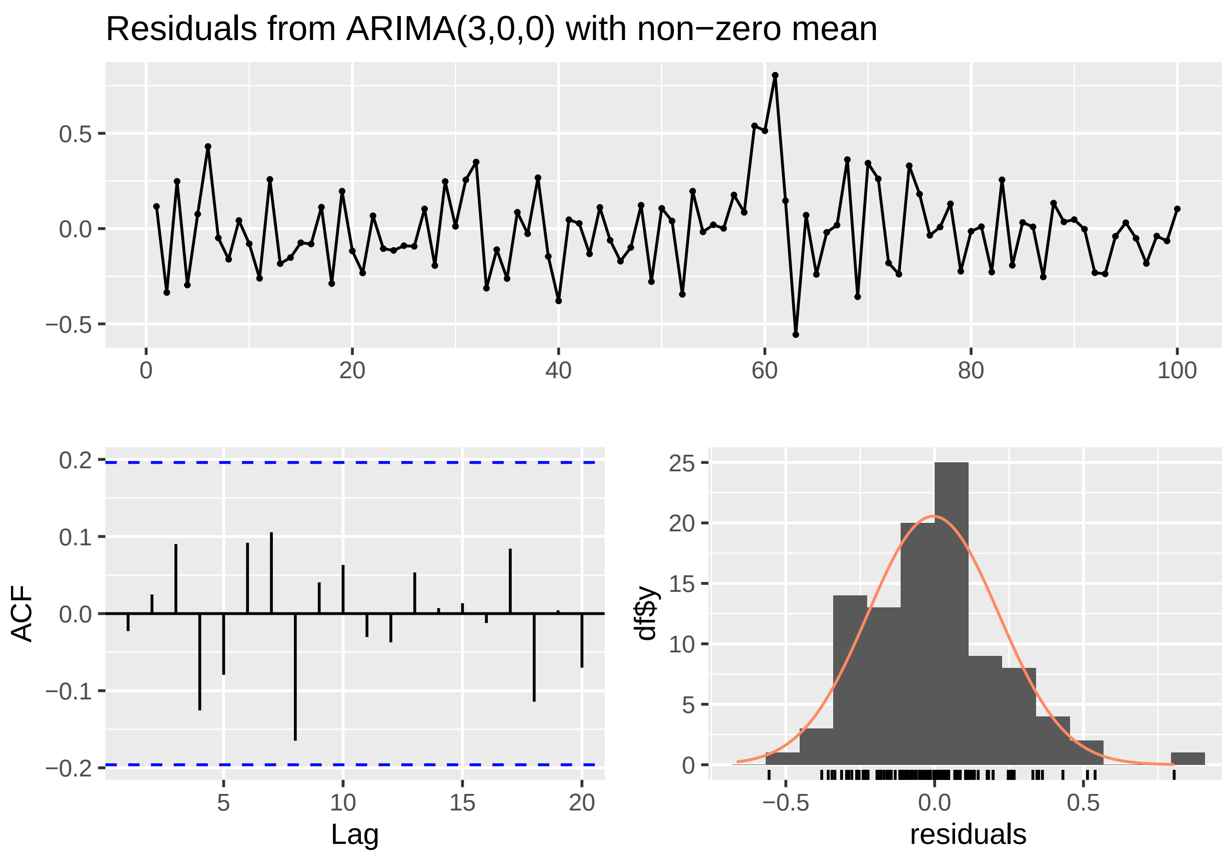
Les tests de vérification des résidus sont :

1. Le test d'absence d'autocorrélation de Ljung-Box,
2. Le test de normalité de Jarque et Bera,
3. Le test d'absence d'effet GARCH de Engle et Granger.

Tests d'absence d'autocorrélation des résidus On teste l'absence d'autocorrélation des résidus jusqu'à l'ordre 10. Les hypothèses du test sont :

$$\left\{ \begin{array}{l} H_0 \ : \ \rho(1) = \rho(2) = \ldots = \rho(10) = 0 \\ H_a \ : \ \exists i \in \{1, \cdots, 10\} \text{ tel que } \rho(i) \neq 0 \end{array} \right.$$

```
test_resAR3<-checkresiduals(AR3,lag = 10)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(3,0,0) with non-zero mean
## Q* = 9.1111, df = 7, p-value = 0.2448
##
## Model df: 3.   Total lags used: 10
```

Commentaires

1. L'autocorrélogramme montre que toutes les autocorrélations sont situées dans les bornes de l'intervalle de confiance à 95% : aucune des autocorrélations n'est significativement différente de 0.

2. La statistique de test de Ljung-Box est égale à Q*(10) = 9.1111.

   - Sous H_0, LB_stat suit une loi $\chi^2(10-3) = \chi^2(7)$.
   - Les seuils critiques sont donc tirés de la loi $\chi^2(7)$:
     - $Q^*(10)_{0.90} = 12.02$ pour un risque de première espèce de 10%
     - $Q^*(10)_{0.95} = 14.07$ pour un risque de première espèce de 5%
     - $Q^*(10)_{0.99} = 18.48$ pour un risque de première espèce de 1%
   - Q*(10) = 9.1111 est inférieure à ces seuils critiques. On ne rejette l'hypothèse nulle d'absence d'autocorrélation des résidus pour un risque de première espèce de 1%.
   - La probabilité critique est égale à 0.2665. Elle est supérieure aux risques de première espèce habituel (1%, 5% et 10%). On ne rejette pas l'hypothèse nulle d'absence d'autocorrélation des résidus.

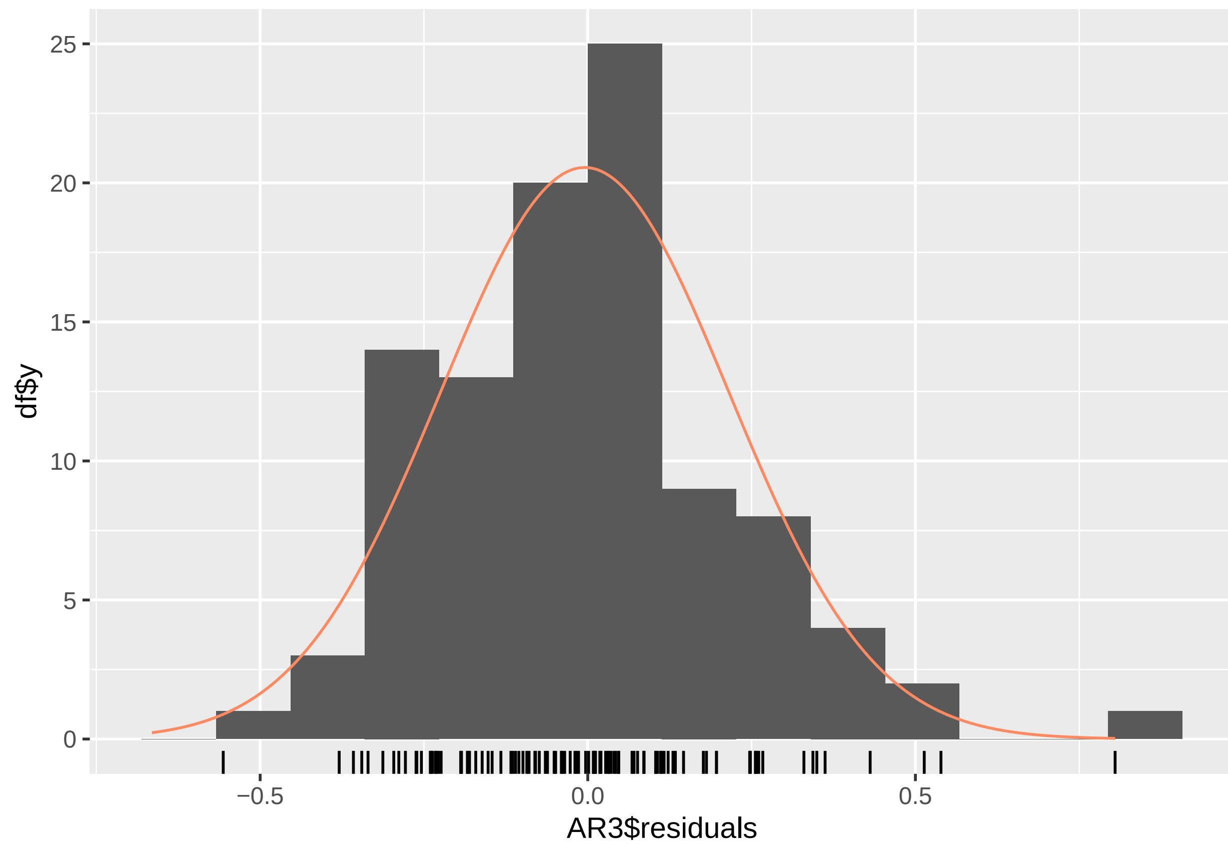
Test de l'hypothèse de normalité des résidus On représente :

1. l'histogramme des résidus (rectangles gris)

2. l'histogramme d'une loi normale avec la même moyenne et le même écart type que celui des résidu (trait rouge)

On peut remarquer que l'h

```r
gghistogram(AR3$residuals, add.normal = TRUE)
```



On peut voir qu'il existe des différences entre les deux histogrammes ce qui permet de supposer que les résidus ne suivent pas une loi normale.

```r
library(moments)
sprintf("Skewness des residus : %f",skewness(AR3$residuals))
```

Test de l'hypothèse en normalité de Jarque and Bera

```
## [1] "Skewness des residus : 0.605464"
```

```r
sprintf("kurtosis des residus : %f",kurtosis(AR3$residuals))
```

```
## [1] "kurtosis des residus : 4.120144"
```

- Le skewness estimé est positif : on a une asymétrie du côté des valeurs positives
- Le kurtosis estimé est supérieur à 3 : c'est l'indice d'une loi non gaussienne

```r
library(tseries)
jarque.bera.test(AR3$residuals)
```

```
##
##  Jarque Bera Test
```

```
##
## data:  AR3$residuals
## X-squared = 11.338, df = 2, p-value = 0.003452
```

Les hypothèses du test sont :

$$\begin{cases} H_0 : S(X) = 0 \; et \; K(X) = 3 \\ H_1 : S(X) \neq 0 \; ou \; K(X) \neq 3 \end{cases}$$

* La statistique du test de Jarque et Bera est égale à est JB_stat = 11.338. * Sous H_0, LB_stat suit une loi $\chi^2(2)$. * Les seuils critiques sont : + $\chi^2_{0.90}(2) = 4.61$ pour un risque de première espèce de 10% + $\chi^2_{0.95}(2) = 5.99$ pour un risque de première espèce de 5% + $\chi^2_{0.99}(2) = 9.21$ pour un risque de première espèce de 1%

- JB_stat = 11.338 est supérieure aux seuils critiques pour des risques de première espèce de 1%, 5% et 10% : on rejette l'hypothèse d'une loi normale pour les résidus.
- La probabilité critique est égale à 0.001 : on rejette l'hypothèse nulle de loi normale à 10%, 5% et 1%.
- Le rejet de l'hypothèse de normalité peut s'expliquer par les pics du taux de chômage observés en 2010(grande récession). Le valeur du taux de chômage est alors beaucoup plus élevée que lors de période précédente. Ceci peut expliquer le kurtosis supérieur à 3

Test de Engle Granger d'absence d'effet ARCH On utilise la fonction ArchTest de la library FinTS (il existe d'autres fonctions dans les packages ATSA et NortsTest notamment)

On teste l'hypothèse d'absence d'effet ARCH avec quatre retards. La régression estimée est :

$$\hat{\epsilon}_t^2 = \gamma_0 + \gamma_1 \hat{\epsilon}_{t-1}^2 + \gamma_2 \hat{\epsilon}_{t-2}^2 + \gamma_3 \hat{\epsilon}_{t-3}^2 + \gamma_4 \hat{\epsilon}_{t-4}^2 + v_t$$

Les hypothèses du test sont :

$$\begin{cases} H_0 : \gamma_1 = ... = \gamma_4 = 0 \Rightarrow pas \, d'effet \, ARCH \\ H_a : \gamma_1 \neq 0 \, ou \, ... \gamma_4 \neq 0 \Rightarrow effet \, ARCH \end{cases}$$

Les résultats du test d'absence d'effet ARCH figurent ci-dessous :

```r
library(FinTS)
ArchTest(AR3$residuals,lags=4,demean = FALSE)
```

```
##
##  ARCH LM-test; Null hypothesis: no ARCH effects
##
## data:  AR3$residuals
## Chi-squared = 25.704, df = 4, p-value = 3.631e-05
```

- La statistique de test est ARCH_LM = 25.704.
- Sous H_0, LB_stat suit une loi $\chi^2(2)$. Les seuils critiques sont :
  - $\chi^2_{0.90}(2) = 4.61$ pour un risque de première espèce de 10%
  - $\chi^2_{0.95}(2) = 5.99$ pour un risque de première espèce de 5%
  - $\chi^2_{0.99}(2) = 9.21$ pour un risque de première espèce de 1%
- ARCH_LM = 23.13 est supérieure aux seuils critiques : on ne rejette l'hypothèse nulle d'absence d'effet ARCH pour les résidus pour les risques de première espèce habituel (1%, 5% et 10%)
- La probabilité critique est égale à 0.000 : on rejette l'hypothèse nulle d'absence d'effet ARCH
- La hausse exceptionnelle du taux de chômage en 2010 explique ce rejet de l'hypothèse d'homoscédasticité.

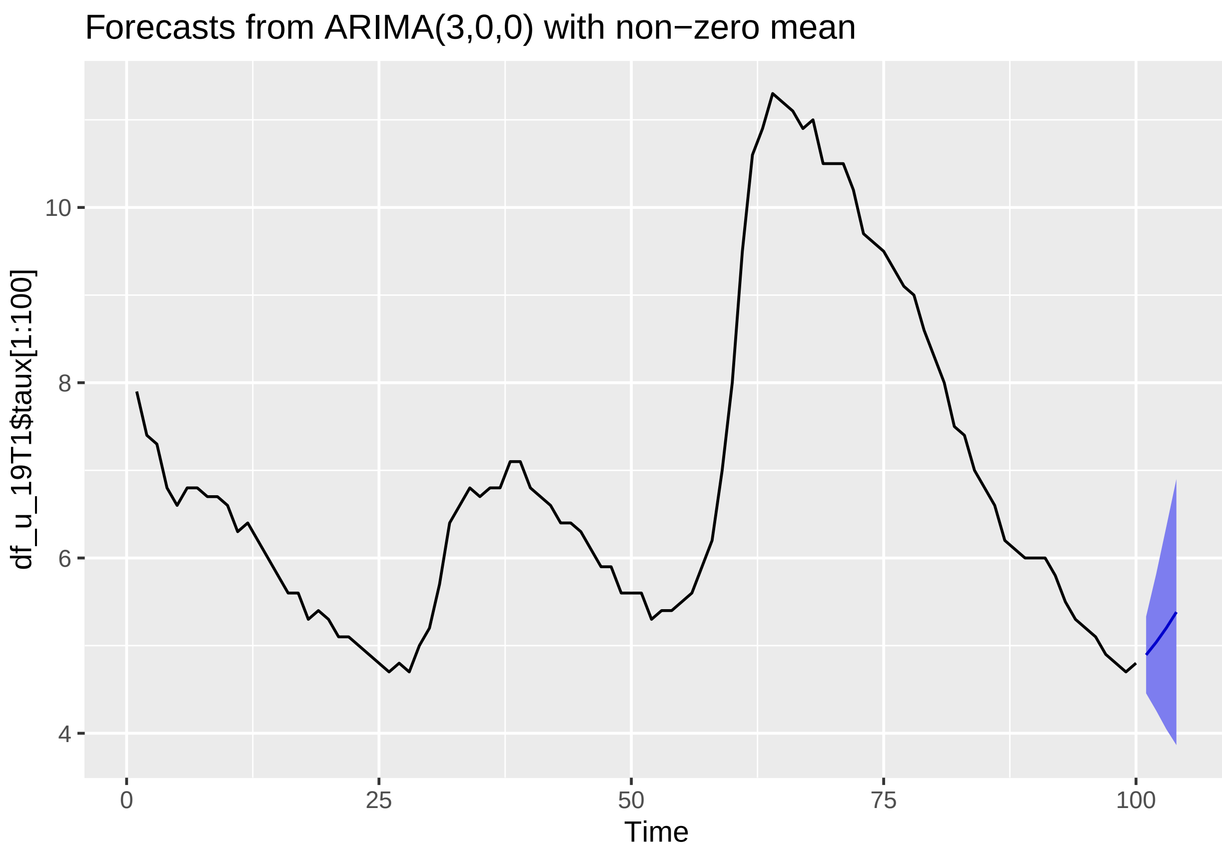
Prévision

Calcul des prévision et représentation graphique On calcule les prévisions pour l'année 2019 (soit à l'horizon h=4) ainsi que leur intervalle de confinance à 95 %.

```r
library(forecast)
forecast_AR3<-forecast(AR3,h=4,level=95)
forecast_AR3
```

```
##     Point Forecast    Lo 95    Hi 95
## 101       4.893957 4.455648 5.332265
## 102       5.038243 4.258545 5.817942
## 103       5.201385 4.046954 6.355815
## 104       5.382795 3.865004 6.900586
```

```r
autoplot(forecast_AR3)
```



$$\begin{cases} US\_taux_t = 6.9381 + \hat{u}_t \\ \hat{u}_t = 1.4712 \times \hat{u}_{t-1} - 0.2221 \times \hat{u}_{t-2} - 0.2827 \times \hat{u}_{t-3} + \hat{\epsilon}_t \end{cases}$$

$$\Rightarrow US\_taux_t - 6.9381 = 1.4712 \times (US\_taux_{t-1} - 6.9381) - 0.2221 \times (US\_taux_{t-2} - 6.9381) - 0.2827 \times (US\_taux_{t-3} - 6$$

$$\Rightarrow US\_taux_t = (1 - 1.4712 + 0.2221 + 0.2827) \times 6.9381 + 1.4712 \times US\_taux_{t-1} - 0.2221 \times US\_taux_{t-2} - 0.2827$$

$$\Rightarrow US\_taux_t = 0.2331 + 1.4712 \times US\_taux_{t-1} - 0.2221 \times US\_taux_{t-2} - 0.2827 \times US\_taux_{t-3} + \hat{\epsilon}_t$$

Calcul de la prévision pour 2019:T1 (h=1)

Les prévisions du de taux de chômage vont dépendre uniquement des trois dernières observations passées

```r
tail(forecast_AR3$mean,n=4)
```

```
## Time Series:
## Start = 101
```

```
## End = 104
## Frequency = 1
## [1] 4.893957 5.038243 5.201385 5.382795
```

```
tail(df_u_19T1[98:100,])
```

```
##          date taux
## 1: 2018-04-01  4.8
## 2: 2018-07-01  4.7
## 3: 2018-10-01  4.8
```

$$US\_taux^a_{2019:T1} = 0.2331 + 1.4712 \times US\_taux_{2018:T4} - 0.2221 \times US\_taux_{2018:T3} - 0.2827 \times US\_taux_{2018:T2}$$
$$\Rightarrow US\_taux^a_{2019:T1} = 0.2331 + 1.4712 \times 4.8 - 0.2221 \times 4.7 - 0.2827 \times 4.8 = 4.8940$$

Calcul de la prévision pour 2019:T2 (h=2)

$$US\_taux^a_{2019:T2} = 0.2331 + 1.4712 \times US\_taux^a_{2019:T1} - 0.2221 \times US\_taux_{2018:T4} - 0.2827 \times US\_taux_{2018:T3}$$
$$\Rightarrow US\_taux^a_{2019:T1} = 0.2331 + 1.4712 \times 4.8940 - 0.2221 \times 4.8 - 0.2827 \times 4.7 = 5.0384$$

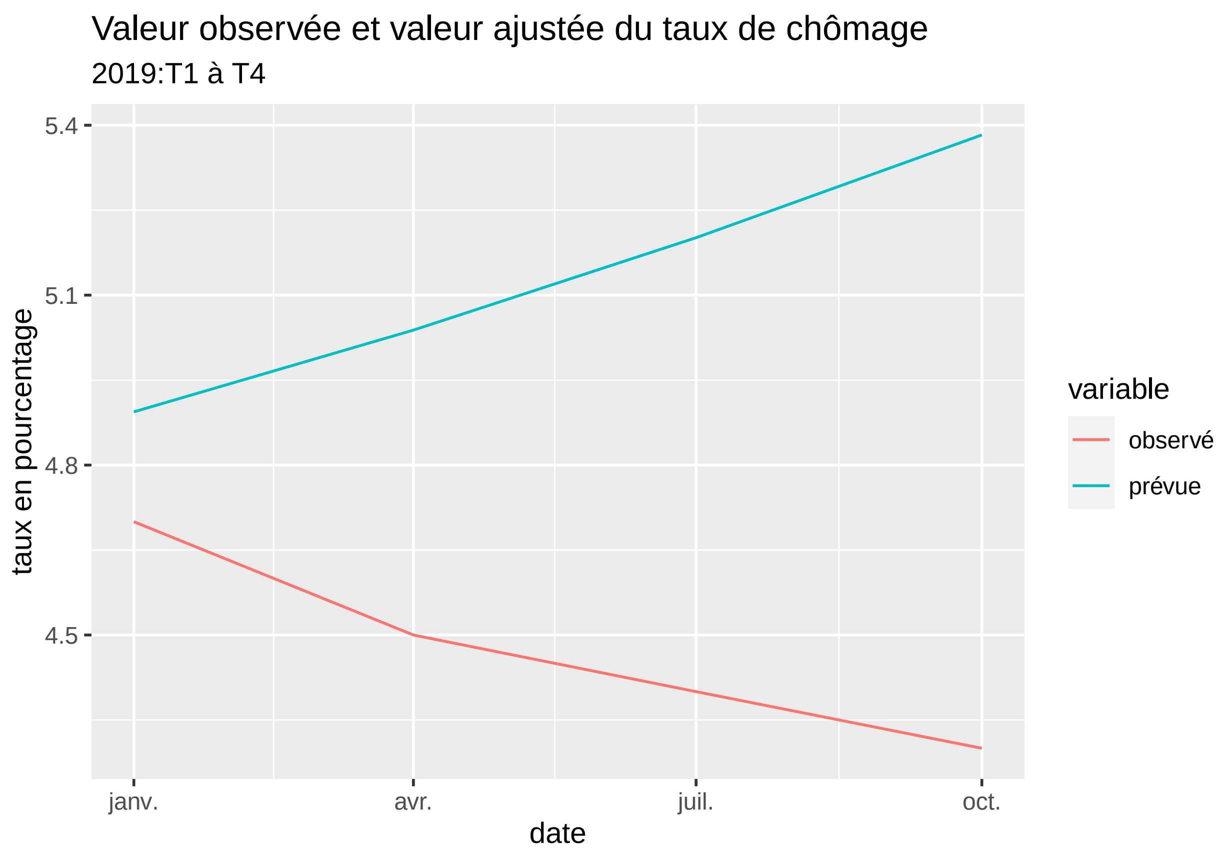
Calcul de la prévision pour 2019:T3 (h=3)

$$US\_taux^a_{2019:T3} = 0.2331 + 1.4712 \times US\_taux^a_{2019:T2} - 0.2221 \times US\_taux^a_{2019:T1} - 0.2827 \times US\_taux_{2018:T4}$$
$$\Rightarrow US\_taux^a_{2019:T1} = 0.2331 + 1.4712 \times 5.0384 - 0.2221 \times 4.8940 - 0.2827 \times 4.8 = 5.2016$$

Calcul de la prévision pour 2019:T4 (h=4)

$$US\_taux^a_{2019:T4} = 0.2331 + 1.4712 \times US\_taux^a_{2019:T3} - 0.2221 \times US\_taux^a_{2019:T2} - 0.2827 \times US\_taux^a_{2019:T1}$$
$$\Rightarrow US\_taux^a_{2019:T1} = 0.2331 + 1.4712 \times 5.2016 - 0.2221 \times 5.0384 - 0.2827 \times 4.8940 = 5.3831$$

```
df_forc<-data.frame(df_u_19T1$date[101:104],df_u_19T1$taux[101:104],as.matrix(forecast_AR3$mean))

colnames(df_forc)<-c("date","observé","prévue")


data_F_melt=melt(as.data.frame(df_forc),id.vars = 1)

p_g_us<-ggplot(data=data_F_melt,aes(x=date))+geom_line(aes(x=date,y=value,color= variable))
p_g_us+labs(y="taux en pourcentage",title = "Valeur observée et valeur ajustée du taux de chômage",
            subtitle = "2019:T1 à T4")
```



Le modèle a tendance à surestimer l'évolution future du taux de chômage par rapport à ses valeurs observées.

```
library(Metrics)
sprintf("MAE out of sample %f",mae(df_forc$observé,df_forc$prévue))
```

Evaluation des la qualité des prévision hors échantillon

```
## [1] "MAE out of sample 0.654095"
```

```
sprintf("RMSE out of sample %f",rmse(df_forc$observé,df_forc$prévue))
```

```
## [1] "RMSE out of sample 0.731776"
```

La MAE et la RMSE hors échantillon sont plus élevées que leur valeur dans l'ensemble d'estimation (training set). Les valeurs hors échantillon permette d'évaluer plus précisément la qualité prédictive du modèle estimé.

Ajout de variables indicatrices pour l'année 2008

Estimation du modèle AR(3) et tests de spécification Le NBER date la "Grande Récession'' de 2007:T4 à 2009:T2. On définit des variables indicatrices individuelles pour les trimestres de la période et on les ajoute au modèle AR(3).

```
df_u_19T1$du_07T4<-as.numeric(df_u_19T1$date=="2007-10-01")
df_u_19T1$du_08T1<-as.numeric(df_u_19T1$date=="2008-01-01")
df_u_19T1$du_08T2<-as.numeric(df_u_19T1$date=="2008-04-01")
df_u_19T1$du_08T3<-as.numeric(df_u_19T1$date=="2008-07-01")
df_u_19T1$du_08T4<-as.numeric(df_u_19T1$date=="2008-10-01")
df_u_19T1$du_09T1<-as.numeric(df_u_19T1$date=="2009-01-01")
```

```r
df_u_19T1$du_09T2<-as.numeric(df_u_19T1$date=="2009-04-01")
AR3_DU<-Arima(y = df_u_19T1$taux[1:100],order=c(3,0,0),xreg=as.matrix(df_u_19T1[1:100,c("du_07T4","du_08
summary(AR3_DU)
```

```
## Series: df_u_19T1$taux[1:100]
## Regression with ARIMA(3,0,0) errors
##
## Coefficients:
##          ar1     ar2      ar3  intercept  du_07T4  du_08T1  du_08T2  du_08T3
##       1.3695  0.0147  -0.4161     6.9857  -0.2732  -0.4862  -0.8770  -0.8307
## s.e.  0.0965  0.1795   0.0977     0.6010   0.1656   0.2192   0.2518   0.2197
##       du_08T4
##       -0.7372
## s.e.   0.1724
##
## sigma^2 = 0.04399:  log likelihood = 16.32
## AIC=-12.63   AICc=-10.16   BIC=13.42
##
## Training set error measures:
##                        ME      RMSE       MAE        MPE     MAPE      MASE
## Training set -0.004025853 0.2000762 0.1623374 -0.1342593 2.400261 0.7475073
##                     ACF1
## Training set -0.05642515
```

```r
library(lmtest)
coeftest(AR3_DU)
```

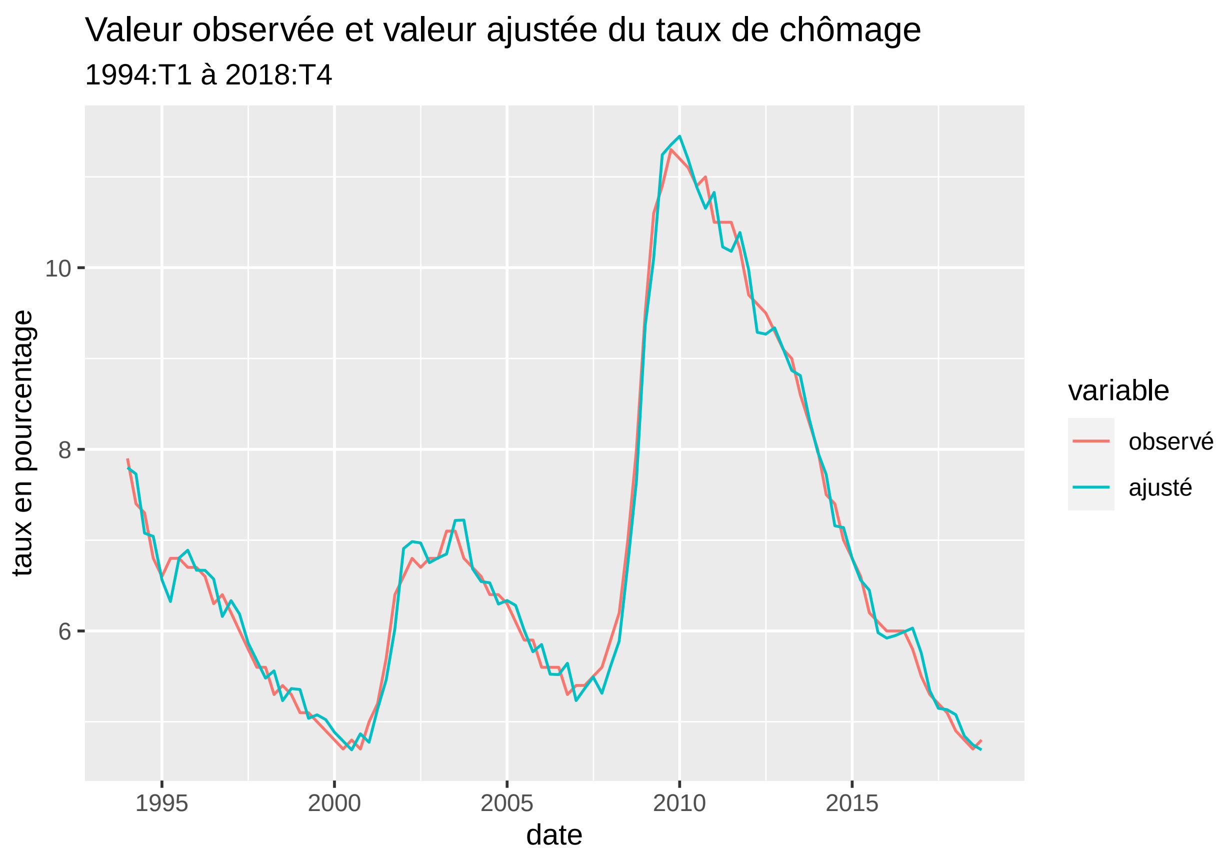
```
##
## z test of coefficients:
##
##            Estimate Std. Error z value  Pr(>|z|)
## ar1        1.369548   0.096488 14.1940 < 2.2e-16 ***
## ar2        0.014652   0.179471  0.0816 0.9349336
## ar3       -0.416099   0.097695 -4.2592 2.052e-05 ***
## intercept  6.985706   0.601042 11.6227 < 2.2e-16 ***
## du_07T4   -0.273223   0.165582 -1.6501 0.0989278 .
## du_08T1   -0.486212   0.219196 -2.2182 0.0265441 *
## du_08T2   -0.876987   0.251756 -3.4835 0.0004949 ***
## du_08T3   -0.830660   0.219716 -3.7806 0.0001564 ***
## du_08T4   -0.737217   0.172392 -4.2764 1.899e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Commentaires

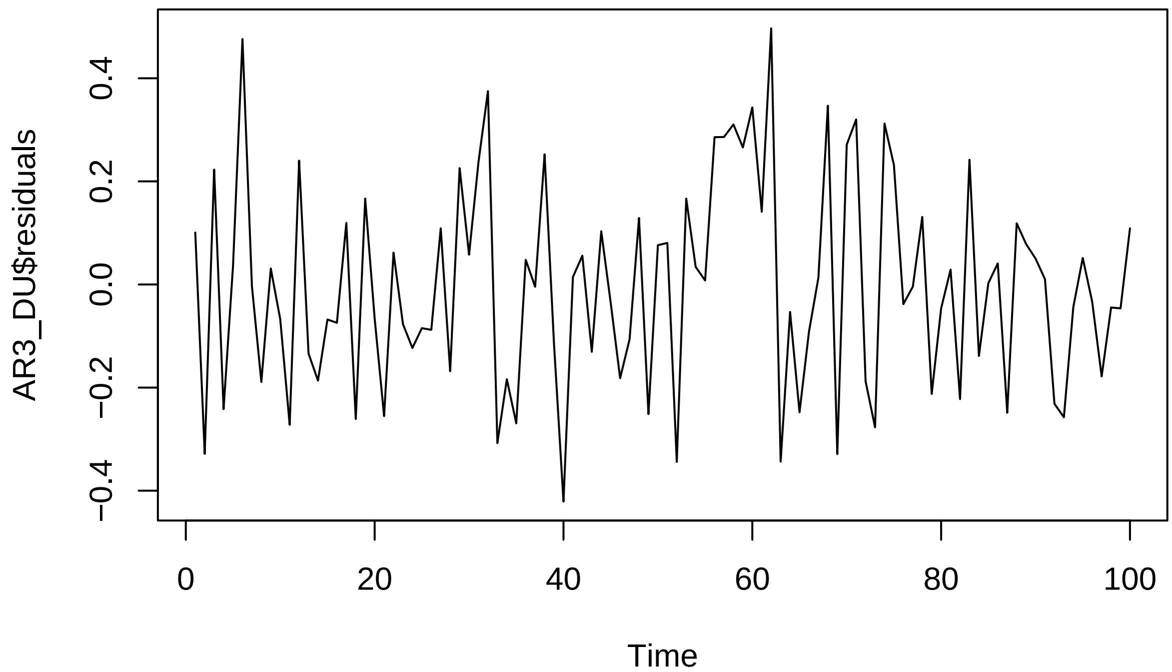
- Le coefficient de l'AR(2) n'est plus significatif. La constante et les coefficients de l'AR(1) et l'AR(3) sont significatifs à 1% et leur ordre de grandeur est proche de celui du modèle sans indicatrices.
- Toutes les indicatrices sont significatives au moins à 10% (du_07T4), à 5% (du_08T1) ou à 1%
- Le critère AIC est plus petit que pour la régression sans indicatrice, le critère BIC est plus élevé.
- La RMSE et la MAE reportée sont les valeurs pour l'échantillon d'estimation ("Training set"). On va comparer les comparer à leur valeur hors échantillon ("out-of-sample") en calculant des prévisions pour des observations n'ayant pas été utilisées pour estimer le modèle.

On représente ci-dessous les valeurs observées et les valeurs prédites par le modèle. On voit que les deux courbes sont assez proches, ce qui permet de penser que la qualité de l'ajustement n'est pas trop mauvaise.

```r
library(ggplot2)
library(reshape2)

df_plot<-data.frame(df_u_19T1$date[1:100],df_u_19T1$taux[1:100],F=as.matrix(AR3_DU$fitted))
colnames(df_plot)<-c("date","observé","ajusté")

data_melt=melt(as.data.frame(df_plot),id.vars = 1)

p_F_us<-ggplot(data=data_melt,aes(x=date))+geom_line(aes(x=date,y=value,color= variable))
p_F_us+labs(y="taux en pourcentage",title = "Valeur observée et valeur ajustée du taux de chômage",
            subtitle = "1994:T1 à 2018:T4")
```




```r
library(tseries)
library(FinTS)
plot(AR3_DU$residuals)
```




```
jarque.bera.test(AR3_DU$residuals)
```

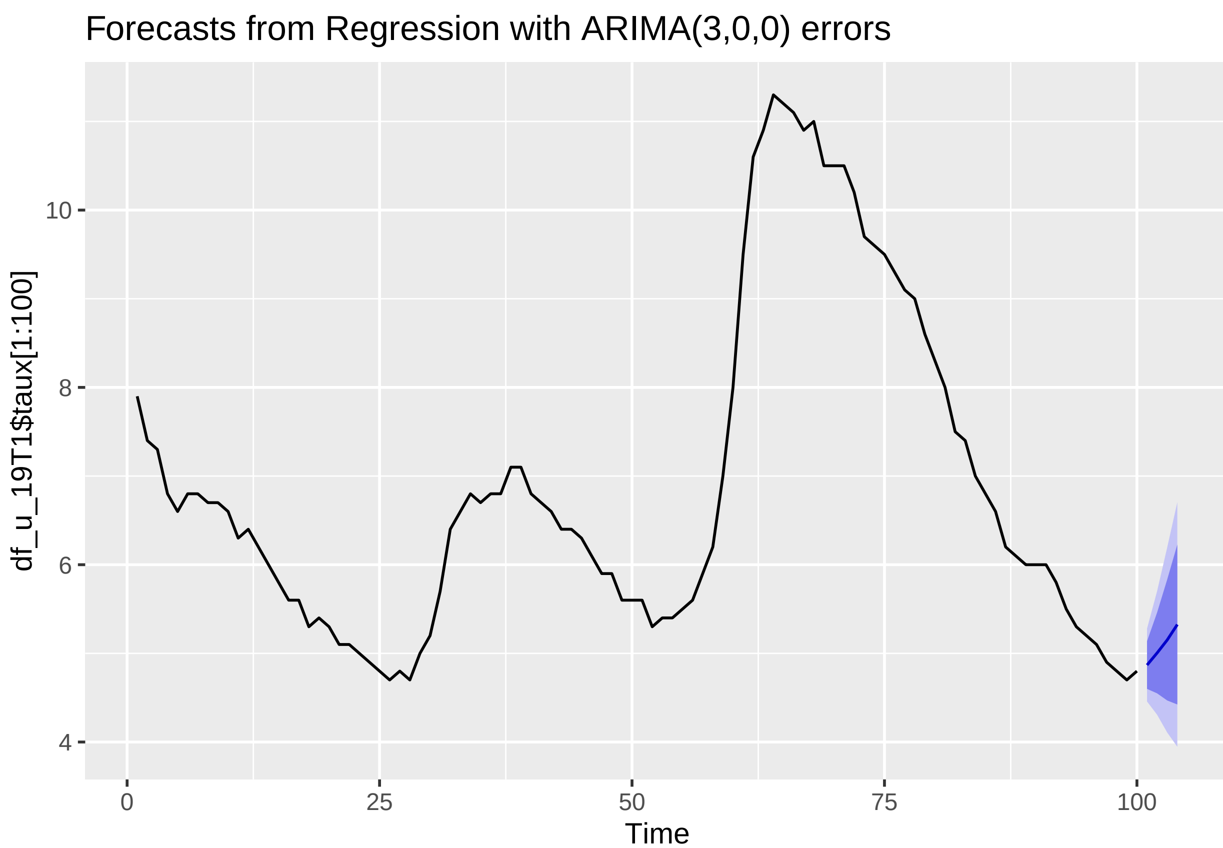
```
##
##  Jarque Bera Test
##
## data:  AR3_DU$residuals
## X-squared = 2.1188, df = 2, p-value = 0.3467
```

```
ArchTest(AR3_DU$residuals,lags=4,demean = FALSE)
```

```
##
##  ARCH LM-test; Null hypothesis: no ARCH effects
##
## data:  AR3_DU$residuals
## Chi-squared = 2.538, df = 4, p-value = 0.6378
```

Commentaires * On ne rejette pas l'hypothèse nulle de loi normale (test de Jarque et Bera) pour un risque de première espèce de 1%. * On ne rejette pas l'hypothèse nulle d'absence d'effet ARCH (test de Engle-Granger) pour un risque de première espèce de 1%. * Les variables indicatrices pour la période de la Grande Récession permettent d'obtenir des résidus compatibles avec la loi normale et sans effet ARCH.

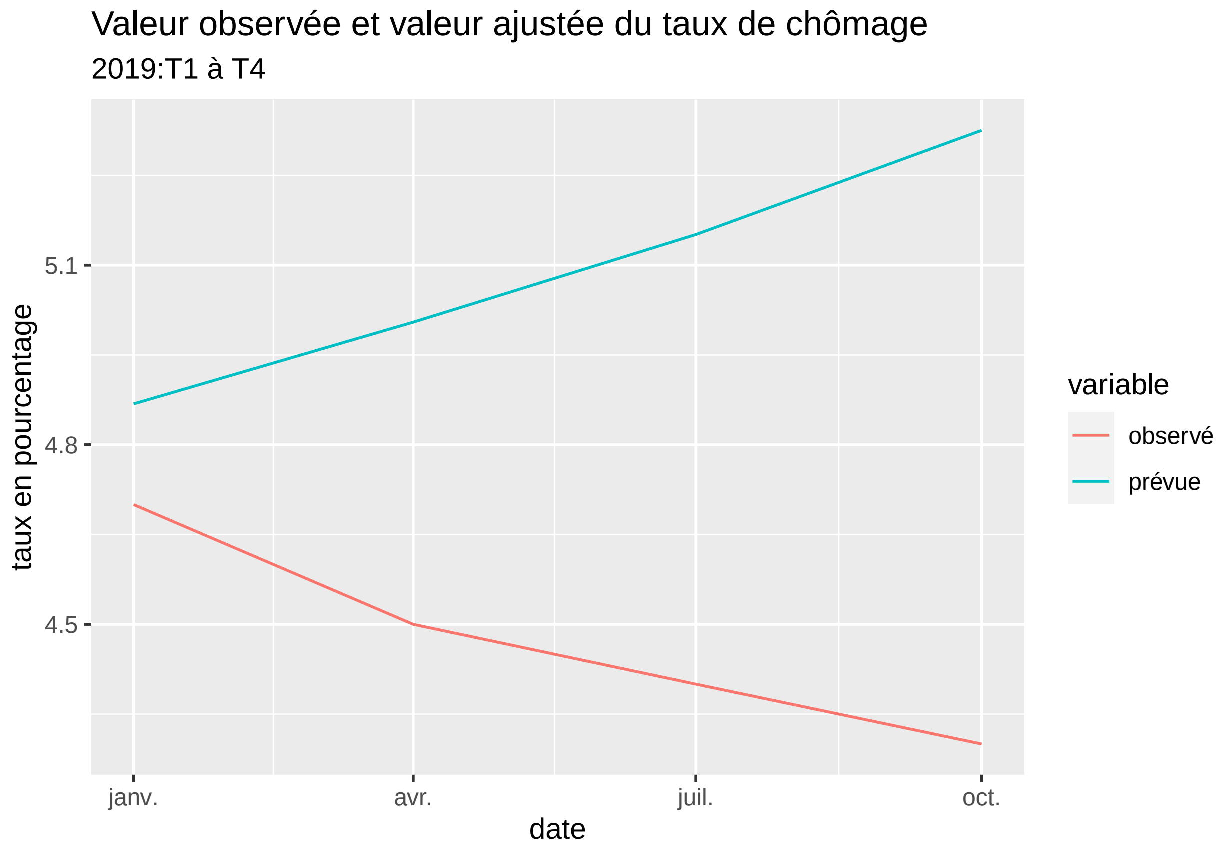
```
library(forecast)
forecast_DU<-forecast(AR3_DU,h=4, xreg = as.matrix(df_u_19T1[101:104,c("du_07T4","du_08T1","du_08T2","du
autoplot(forecast_DU)
```



Prévision

Le calcul des prévisions se fait exactement comme avec le modèle sans indicatrices. Dans notre cas, les indicatrices sont définies par rapport à des dates données. Leurs valeurs futures sont égales à 0.

```
df_forc_DU<-data.frame(df_u_19T1$date[101:104],df_u_19T1$taux[101:104],as.matrix(forecast_DU$mean))

colnames(df_forc_DU)<-c("date","observé","prévue")


data_F_DU_melt=melt(as.data.frame(df_forc_DU),id.vars = 1)

p_g_us<-ggplot(data=data_F_DU_melt,aes(x=date))+geom_line(aes(x=date,y=value,color= variable))
p_g_us+labs(y="taux en pourcentage",title = "Valeur observée et valeur ajustée du taux de chômage",
           subtitle = "2019:T1 à T4")
```



Comme avec le modèle précédent, les valeurs prévues sont plus élevées que les valeurs observées.

```
library(Metrics)
sprintf("MAE out of sample %f",mae(df_forc_DU$observé,df_forc_DU$prévue))
```

```
## [1] "MAE out of sample 0.612411"
```

```
sprintf("RMSE out of sample %f",rmse(df_forc_DU$observé,df_forc_DU$prévue))
```

```
## [1] "RMSE out of sample 0.688988"
```

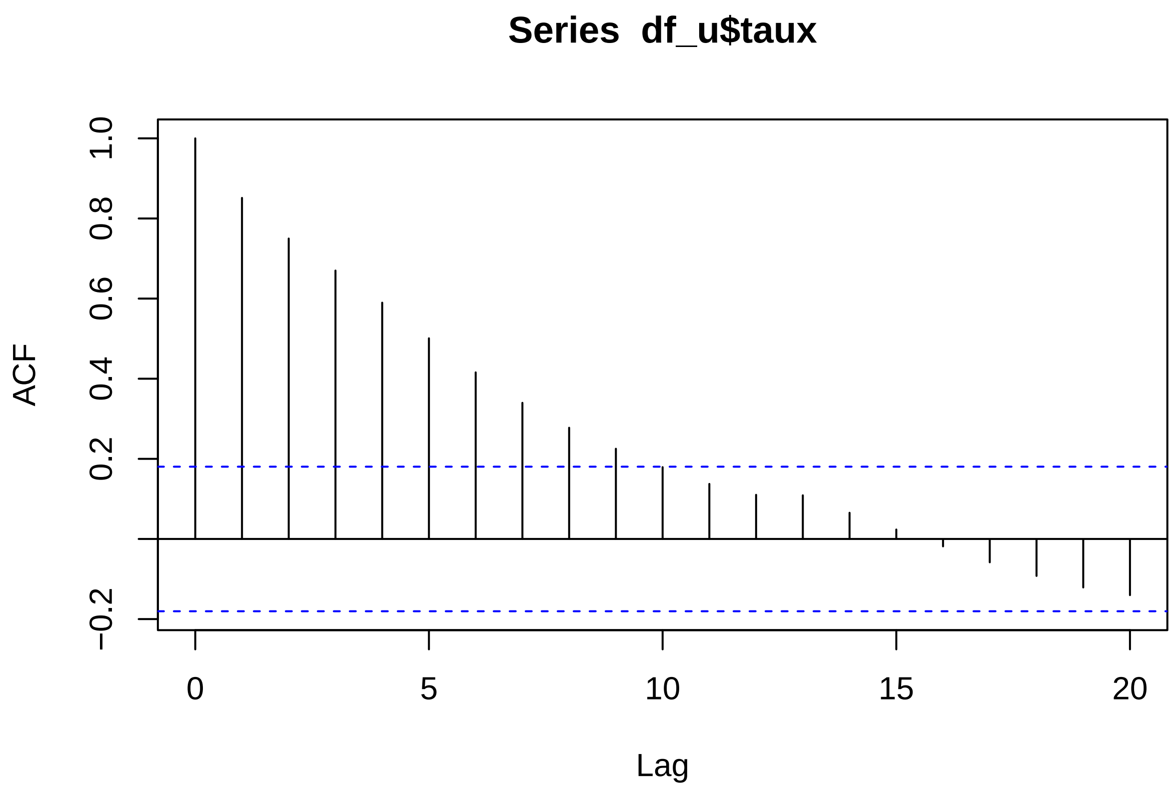
La RMSE et la MAE sont plus faibles pour les prévisions obtenues du modèle sans indicatrices. Les indicatrices ont donc améliorer les capacités prédictives du modèles même si l'évolution future du taux de chômage est surestimée.

Estimation sur la totalité de l'échantillon 1994:T1 à 2023:T1

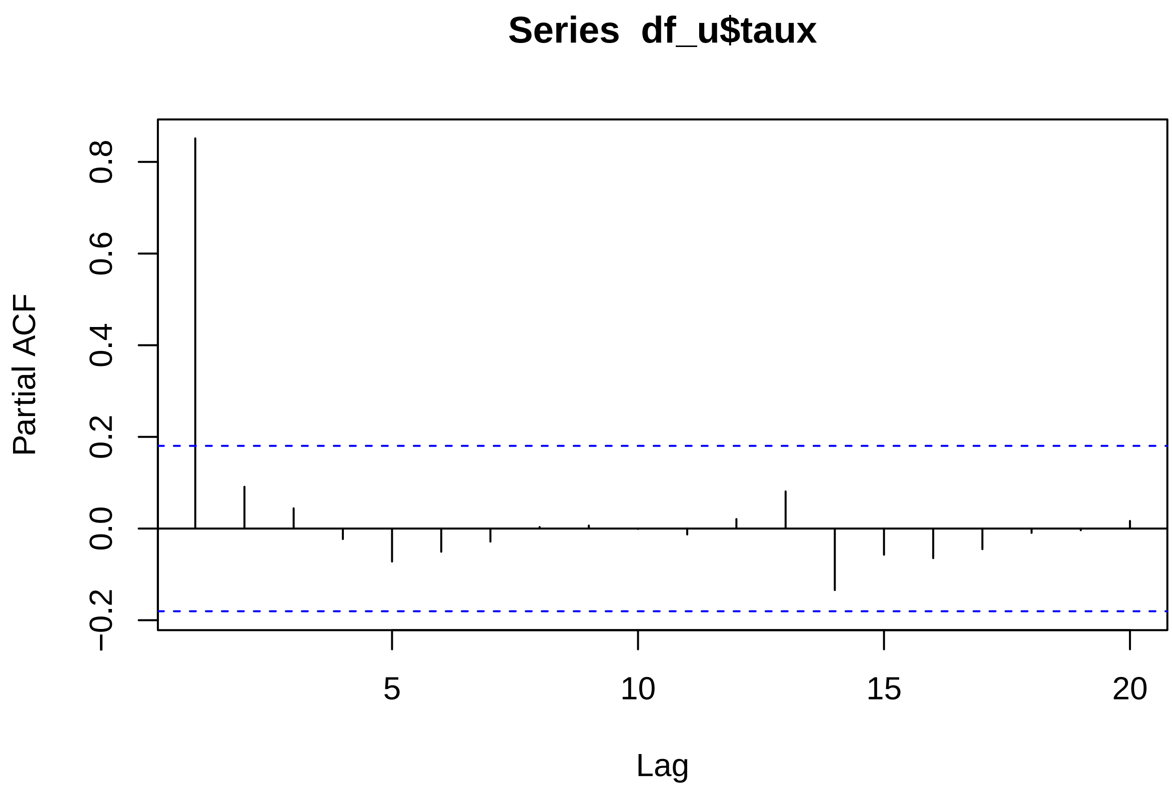
L'échantillon total est caractérisé notamment par la hausse brutale et temporaite du taux de chômage pendant l'épidémie de covid. Après cette période, le taux de chômage revient à des niveaux comparables à ses valeurs passées. On devra sans doute introduire des variables indicatrices pour prendre en compte les effets de l'épidémie de covid.

Première estimation On commence par estimer l'autocorrélogramme.

```
acf(x = df_u$taux)
```




```
pacf(x = df_u$taux)
```



Commentaires

- Pas de modification des autocorrélations par rapport à la période 1994:T1-2019:T4
- Seule l'autocorrélation partielle à l'ordre 1 est maintenant significitive.

On estime dans un premier temps un AR(1)

```
AR1_C<-Arima(y = df_u$taux[1:114],order=c(1,0,0))
summary(AR1_C)
```

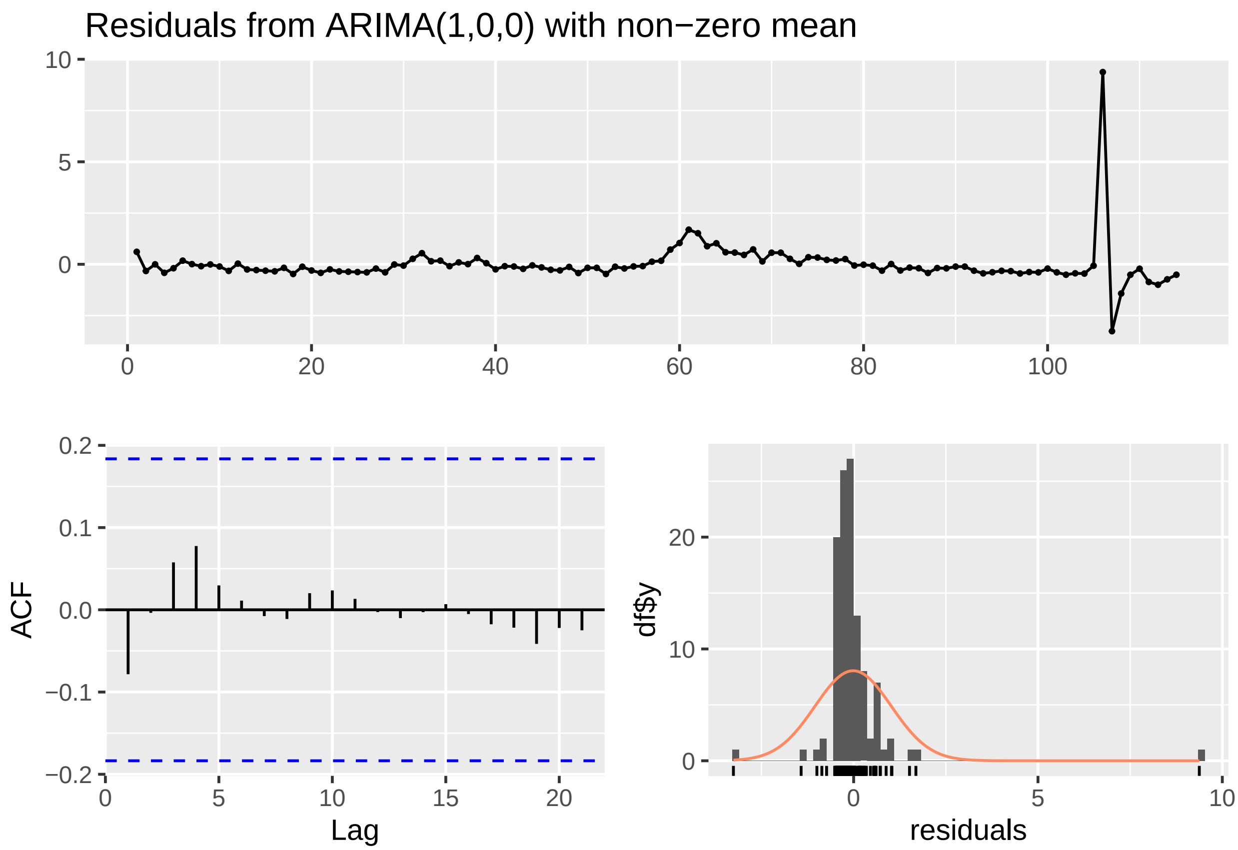
```
## Series: df_u$taux[1:114]
## ARIMA(1,0,0) with non-zero mean
##
## Coefficients:
##          ar1    mean
##       0.8492  6.7483
## s.e.  0.0483  0.6110
##
## sigma^2 = 1.081:  log likelihood = -165.82
## AIC=337.64   AICc=337.86   BIC=345.85
##
## Training set error measures:
##                       ME     RMSE       MAE       MPE    MAPE     MASE
## Training set -0.01333289 1.030493 0.4379158 -1.822542 5.83886 1.183839
##                     ACF1
## Training set -0.07833147
```

```
coeftest(AR1_C)
```

```
##
## z test of coefficients:
##
##           Estimate Std. Error z value  Pr(>|z|)
## ar1       0.849237   0.048324  17.574 < 2.2e-16 ***
## intercept 6.748346   0.610995  11.045 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Commentaires * La constante et le coefficient ar1 sont significativement différents de 0

```
checkresiduals(AR1_C)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(1,0,0) with non-zero mean
## Q* = 2.1064, df = 9, p-value = 0.9897
##
## Model df: 1.   Total lags used: 10
```

Commentaires * Le test de Ljung-Box avec 10 retards ne rejette pas l'hypothèse de nullité des 10 premières autocorrélations.

```
jarque.bera.test(AR1_C$residuals)
```

```
##
##  Jarque Bera Test
##
```

```
## data:  AR1_C$residuals
## X-squared = 17061, df = 2, p-value < 2.2e-16
```

```
ArchTest(AR1_C$residuals,lags=4,demean = FALSE)
```

```
##
##  ARCH LM-test; Null hypothesis: no ARCH effects
##
## data:  AR1_C$residuals
## Chi-squared = 1.3248, df = 4, p-value = 0.8572
```

Commentaires * Le test de Jarque et Bera rejette l'hypothèse nulle de normalité des résidus * Le test ARCH rejette l'hypothèse nulle d'absence d'effet ARCH.

Ajout de variables indicatrices

On va essayer d'améliorer le modèle en ajoutant des variables indicatrices :

- pour la crise de 2007-2008 comme précédemment
- pour la crise du covid

L'ajout de ces indicatrices va nous obliger à augmenter le nombre de retards et de passer un AR(2). En effet, l'AR(1) avec les indicatrices présente de l'autocorrélation des résidus.

```
df_u$du_07T4<-as.numeric(df_u$date=="2007-10-01")
df_u$du_08T1<-as.numeric(df_u$date=="2008-01-01")
df_u$du_08T2<-as.numeric(df_u$date=="2008-04-01")
df_u$du_08T3<-as.numeric(df_u$date=="2008-07-01")
df_u$du_08T4<-as.numeric(df_u$date=="2008-10-01")
df_u$du_09T1<-as.numeric(df_u$date=="2009-01-01")
df_u$du_09T2<-as.numeric(df_u$date=="2009-04-01")


df_u$du_20T1<-as.numeric(df_u$date=="2020-01-01")
df_u$du_20T2<-as.numeric(df_u$date=="2020-04-01")
df_u$du_20T3<-as.numeric(df_u$date=="2020-07-01")
df_u$du_20T4<-as.numeric(df_u$date=="2020-10-01")


AR2_DU_C<-Arima(y = df_u$taux[1:114],order=c(2,0,0),xreg=as.matrix(df_u[1:114,c("du_09T1","du_09T2","du_
summary(AR2_DU_C)
```

```
## Series: df_u$taux[1:114]
## Regression with ARIMA(2,0,0) errors
##
## Coefficients:
##          ar1      ar2  intercept  du_09T1  du_09T2  du_20T2  du_20T3  du_20T4
##       1.7189  -0.7500     6.6785   0.4125   0.4928   8.9798   3.7109   1.0614
## s.e.  0.0610   0.0613     0.6486   0.1440   0.1440   0.1661   0.2198   0.1664
##
## sigma^2 = 0.05604:  log likelihood = 4.14
## AIC=9.71   AICc=11.44   BIC=34.34
##
## Training set error measures:
##                        ME      RMSE       MAE        MPE     MAPE      MASE
## Training set -0.005706512 0.2282715 0.1853179 -0.2592227 2.796242 0.5009789
##                     ACF1
```

```
## Training set -0.05851768
```

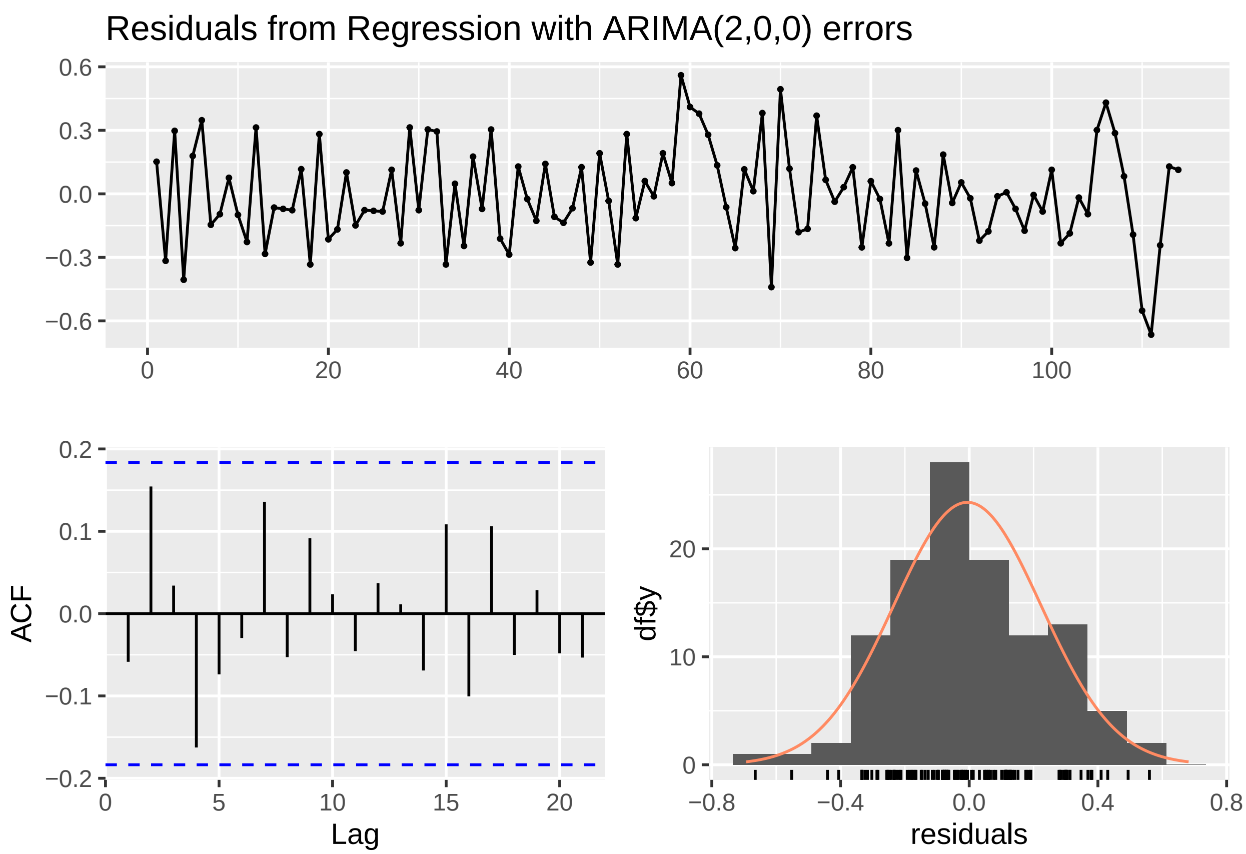
```r
coeftest(AR2_DU_C)
```

```
##
## z test of coefficients:
##
##            Estimate Std. Error  z value  Pr(>|z|)
## ar1        1.718875   0.061019  28.1696 < 2.2e-16 ***
## ar2       -0.749988   0.061325 -12.2297 < 2.2e-16 ***
## intercept  6.678455   0.648620  10.2964 < 2.2e-16 ***
## du_09T1    0.412524   0.144014   2.8645  0.004177 **
## du_09T2    0.492827   0.144042   3.4214  0.000623 ***
## du_20T2    8.979821   0.166141  54.0493 < 2.2e-16 ***
## du_20T3    3.710918   0.219786  16.8843 < 2.2e-16 ***
## du_20T4    1.061392   0.166367   6.3798 1.773e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Commentaires

- Les trois indicatrices relatives à l'année 2020T2, T3 et T4 sont significatives et positives pour un risque de première espèce de 1%
- Seules les indicatrices pour 2009:T1 et 2009:T2 s'avèrent significatives quand on prend en compte la totalité des observations pour un risque de première espèce de 1%.
- La constante et les coefficients ar1 et ar2 sont significatifs pour un risque de première espèce de 1%
- Les critères AIC et BIC sont nettement plus petits que pour le AR(1) sant indicatrices.

```r
checkresiduals(AR2_DU_C)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from Regression with ARIMA(2,0,0) errors
## Q* = 11.056, df = 8, p-value = 0.1985
##
## Model df: 2.   Total lags used: 10
```

Commentaires

- Le test de Ljung-Box ne rejette pas l'hypothèse d'absence d'autocorrélation des résidus.

```
jarque.bera.test(AR2_DU_C$residuals)
```

```
##
##  Jarque Bera Test
##
## data:  AR2_DU_C$residuals
## X-squared = 0.13527, df = 2, p-value = 0.9346
```

```
ArchTest(AR2_DU_C$residuals,lags=4,demean = FALSE)
```

```
##
##  ARCH LM-test; Null hypothesis: no ARCH effects
##
## data:  AR2_DU_C$residuals
## Chi-squared = 22.9, df = 4, p-value = 0.0001326
```

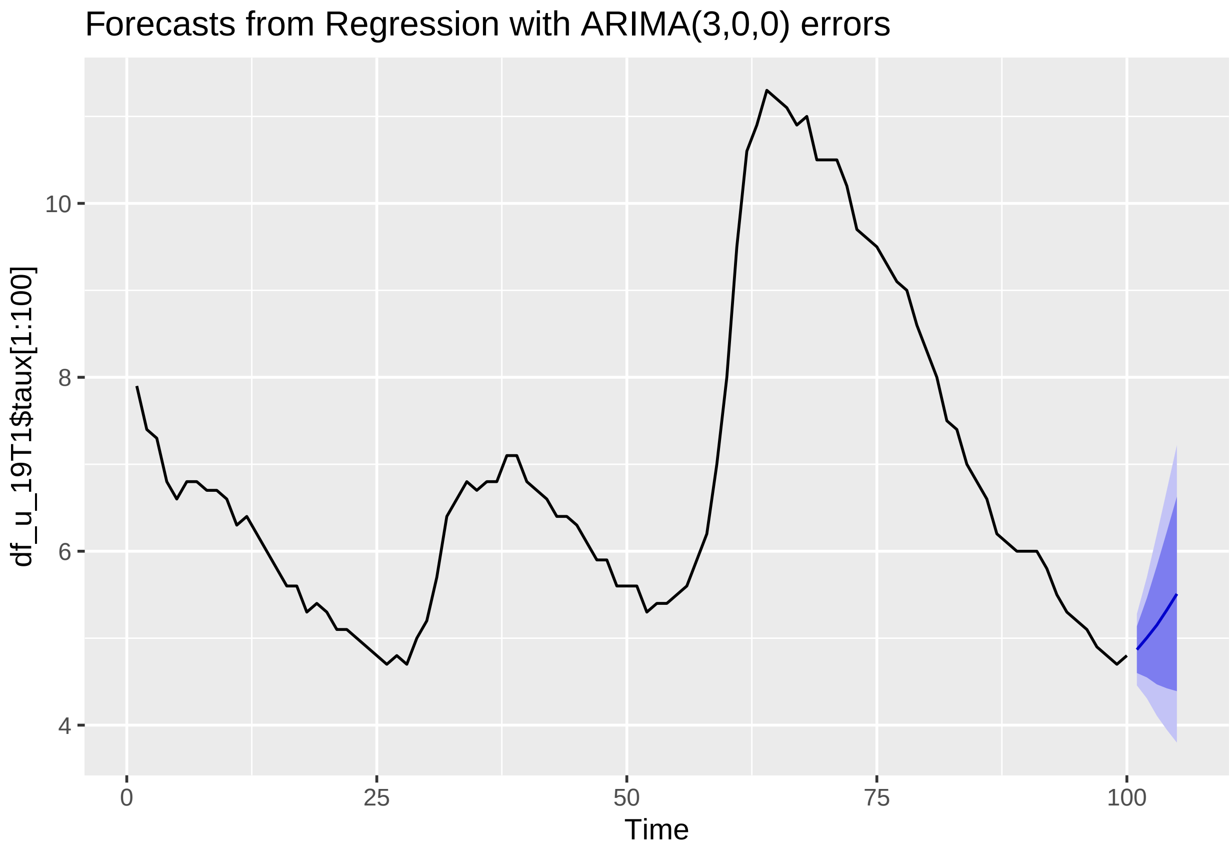
Commentaires

* Le test de Jarque et Bera ne rejette plus l'hypothèse de normalités des résidus
* Le test ARCH rejette encore l'hypothèse nulle d'absence d'effet ARCH. Malgrè l'introduction de vari

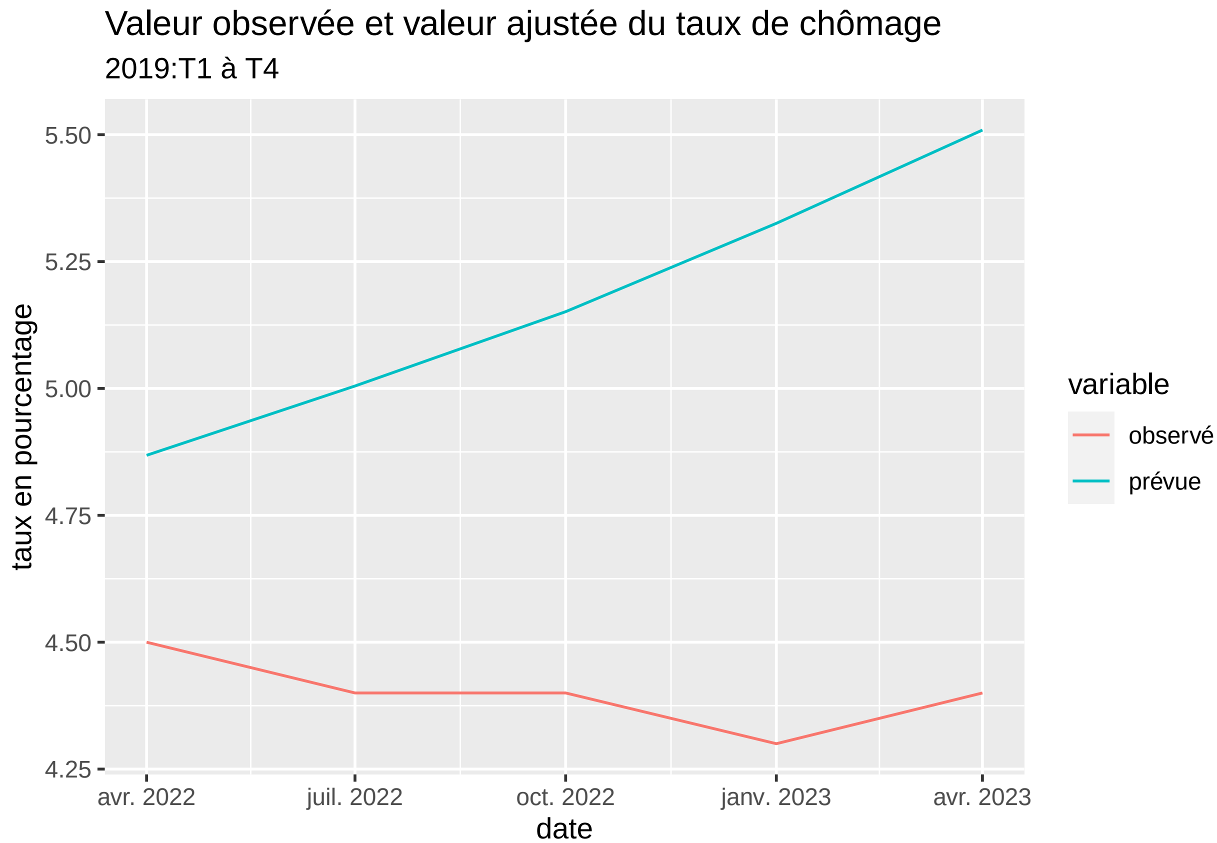
Prévision

On calcule les prévisions à partir du modèle AR(2) avec les indicatrices

```
library(forecast)
forecast_DU_C<-forecast(AR3_DU,h=4, xreg = as.matrix(df_u[114:118,c("du_09T1","du_09T2","du_20T2","du_20
autoplot(forecast_DU_C)
```




```
df_forc_DU<-data.frame(df_u$date[114:118],df_u$taux[114:118],as.matrix(forecast_DU_C$mean))

colnames(df_forc_DU)<-c("date","observé","prévue")


data_F_DU_melt=melt(as.data.frame(df_forc_DU),id.vars = 1)

p_g_us<-ggplot(data=data_F_DU_melt,aes(x=date))+geom_line(aes(x=date,y=value,color= variable))
p_g_us+labs(y="taux en pourcentage",title = "Valeur observée et valeur ajustée du taux de chômage",
            subtitle = "2019:T1 à T4")
```



Le modèle surestime l'évolution future du taux de chômage.

```
library(Metrics)
sprintf("MAE out of sample %f",mae(df_forc_DU$observé,df_forc_DU$prévue))
```

```
## [1] "MAE out of sample 0.771754"
```

```
sprintf("RMSE out of sample %f",rmse(df_forc_DU$observé,df_forc_DU$prévue))
```

```
## [1] "RMSE out of sample 0.818200"
```

La MAE et la RMSE du modèle out-of-sample sont plus élevés que leur valeur "in sample" (RMSE et MAE in training)

Méthodes d'estimations alternatives On présente quelques fonctions et packages alternatifs.

auto.arima Dans la librairie forecast, la fonction auto.arima permet de sélectionner le meilleur modèle en spcifiant le nombre de retards maximum pour la composante AR (max.p), la composante MA (max.q), le nombre de différenciation de la série (max.d). Ici on a contraint max.d=0 mais sa valeur par défaut est max.d=1

```
best_ARMA <-auto.arima(df_u_19T1$taux,max.p = 4,max.q=4,max.d=0)
class(best_ARMA)
```

```
## [1] "forecast_ARIMA" "ARIMA"          "Arima"
```

```
summary(best_ARMA)
```

```
## Series: df_u_19T1$taux
```

```
## ARIMA(3,0,0) with non-zero mean
##
## Coefficients:
##          ar1      ar2      ar3    mean
##       1.4785  -0.2325  -0.2760  6.7872
## s.e.  0.0949   0.1745   0.0955  0.6715
##
## sigma^2 = 0.049:  log likelihood = 8.73
## AIC=-7.46   AICc=-6.85   BIC=5.76
##
## Training set error measures:
##                        ME      RMSE      MAE        MPE     MAPE      MASE
## Training set -0.005086007 0.2170553 0.166647 -0.2014136 2.468176 0.7802108
##                     ACF1
## Training set -0.02324285
```

```r
arimaorder(best_ARMA)
```

```
## p d q
## 3 0 0
```

```r
library(tseries)
order=c(3,0) # order of the ARMA model c(ARlag,MAlag)
AR2_b = arma(x=df_u_19T1$taux,order=c(2,0),include.intercept=TRUE)
summary(AR2_b) # to display the results
```

Librairie tseries

```
##
## Call:
## arma(x = df_u_19T1$taux, order = c(2, 0), include.intercept = TRUE)
##
## Model:
## ARMA(2,0)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.47452 -0.15826 -0.03582  0.12876  0.84298
##
## Coefficient(s):
##            Estimate  Std. Error  t value Pr(>|t|)
## ar1         1.66830     0.07012   23.794   <2e-16 ***
## ar2        -0.68907     0.07058   -9.763   <2e-16 ***
## intercept   0.13417     0.08524    1.574    0.116
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Fit:
## sigma^2 estimated as 0.05106,  Conditional Sum-of-Squares = 5.16,  AIC = -8.23
```

```r
library(FinTS)
AR2_fin = ARIMA(x = df_u_19T1$taux,order=c(3,0,0),type=c("Ljung-Box"))
summary(AR2_fin)
```

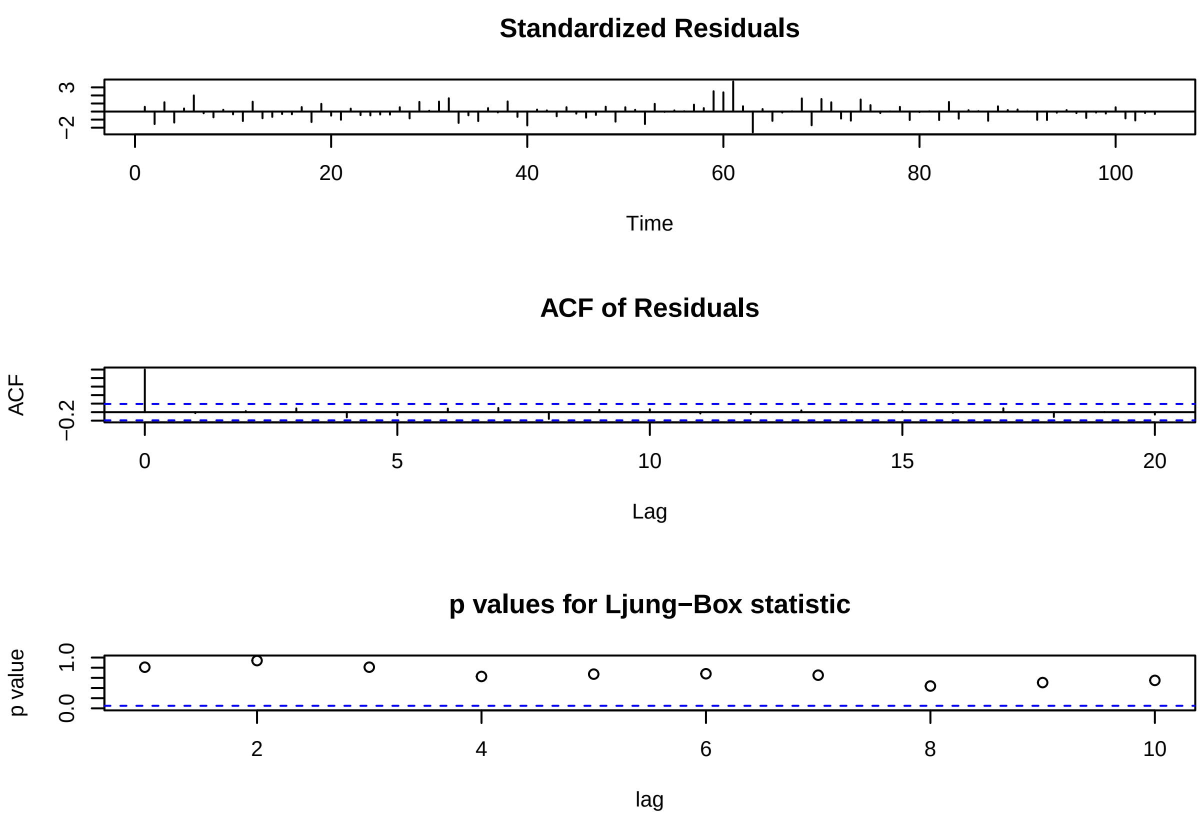
Librairie FinTS

```
##
## Call:
## stats::arima(x = x, order = order, seasonal = seasonal, xreg = xreg, include.mean = include.mean,
##     transform.pars = transform.pars, fixed = fixed, init = init, method = method,
##     n.cond = n.cond, optim.control = optim.control, kappa = kappa)
##
## Coefficients:
##          ar1      ar2      ar3  intercept
##       1.4785  -0.2325  -0.2760     6.7872
## s.e.  0.0949   0.1745   0.0955     0.6715
##
## sigma^2 estimated as 0.04711:  log likelihood = 8.73,  aic = -7.46
##
## Training set error measures:
##                 ME RMSE MAE MPE MAPE
## Training set NaN  NaN NaN NaN  NaN
```

```r
AR2_fin$Box.test
```

```
##
##  Box-Ljung test (lag = 5)
##
## data:  fit$resid
## X-squared = 3.1726, df = 2, p-value = 0.2047
```

```r
tsdiag(AR2_fin)
```



```r
#plotArmaTrueacf(AR2_fin)
#ArchTest(AR2_fin$residuals,lags = 2)
```