TD2 Modélisation ARMA

Arthur et Florian

Sommaire

Chargement des packages nécessaires

```
library(tidyverse)
library(forecast)
```

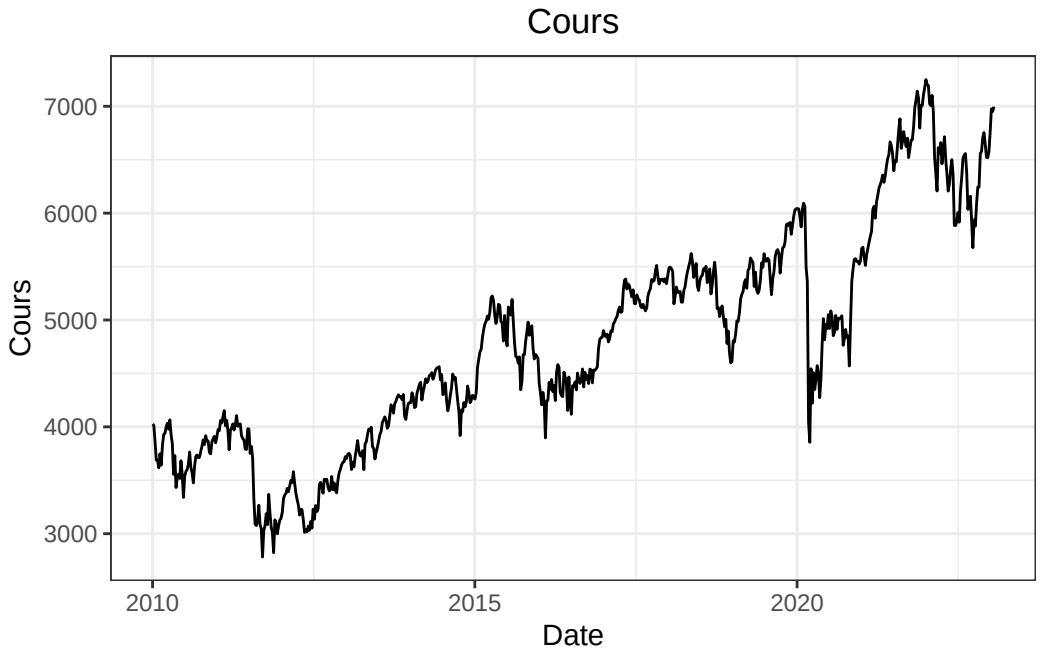
I. Estimation de processus autorégressifs via les MCO

```
CAC40 <- read_csv2("data/CAC40_2010_2023.csv")
```

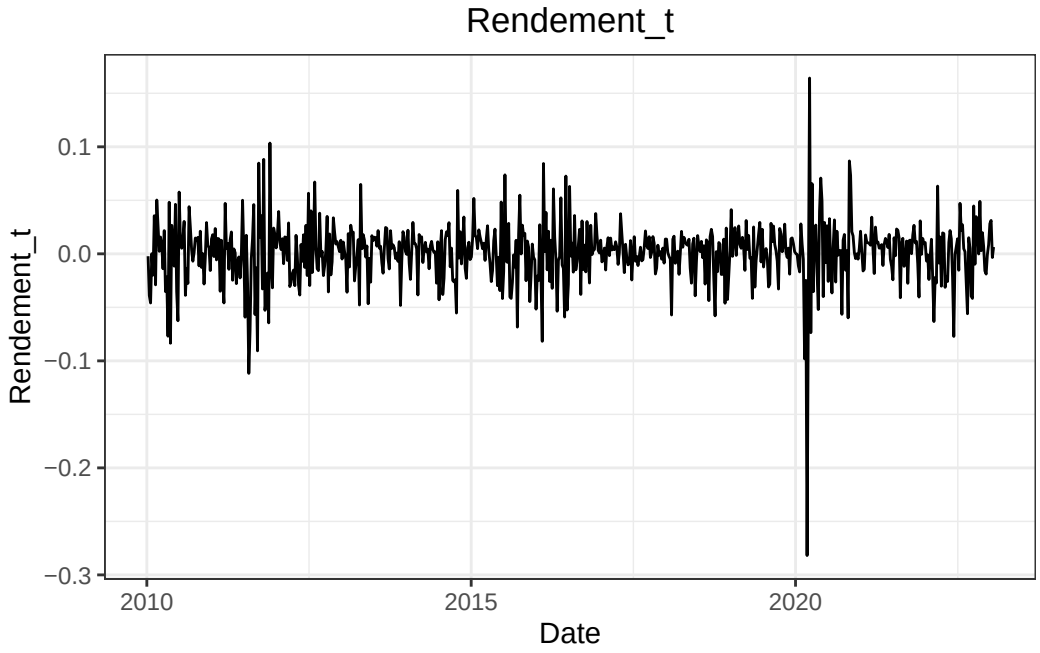
```
## Préparation des données

CAC40 <- CAC40 |>
  mutate(
    Date = as_date(Date, format = "%d/%m/%Y"),
    CAC40 = as.numeric(CAC40)
  ) |>
  rename(Cours = CAC40)
```

```
# Calcul du rendement en t, t-1 et t-2

CAC40 <- CAC40 |>
  mutate(
    Rendement_t = log(Cours / lag(Cours)),
    Rendement_t_1 = lag(Rendement_t),
    Rendement_t_2 = lag(Rendement_t, 2)
  )
```

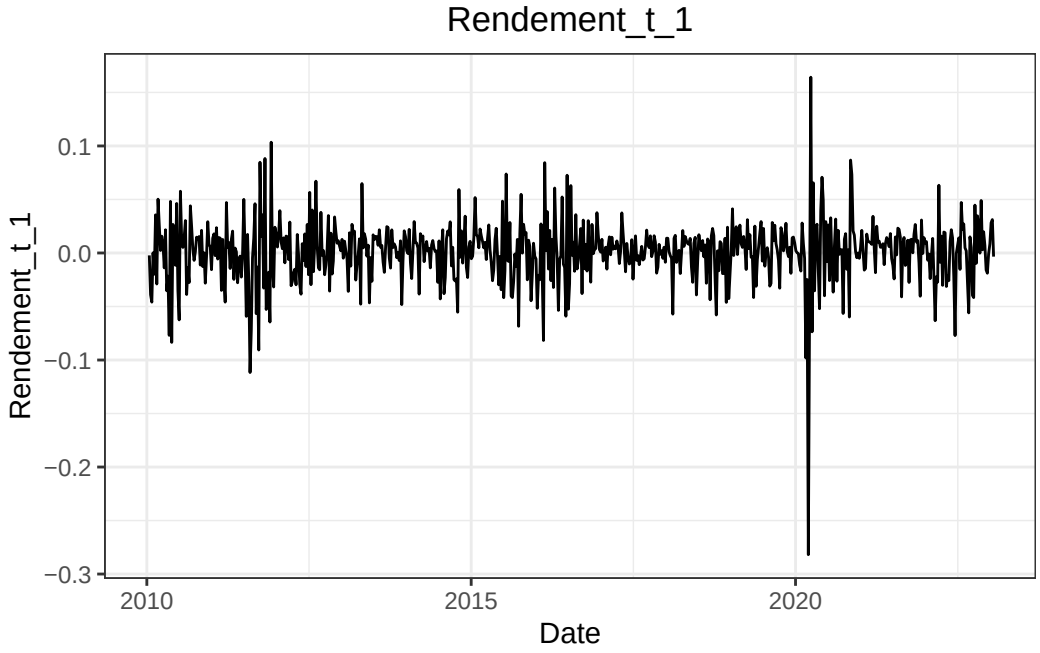
```
# Graphiques des rendements

colonnes <- c("Cours", "Rendement_t", "Rendement_t_1", "Rendement_t_2")

graphique_rendement <- map(colonnes, function(i) {
  CAC40 |>
    ggplot(aes(x = Date, y = .data[[i]])) +
    geom_line() +
    theme_bw() +
    ggtitle(i) +
    theme(plot.title = element_text(hjust = 0.5))
})

walk(graphique_rendement, print)
```



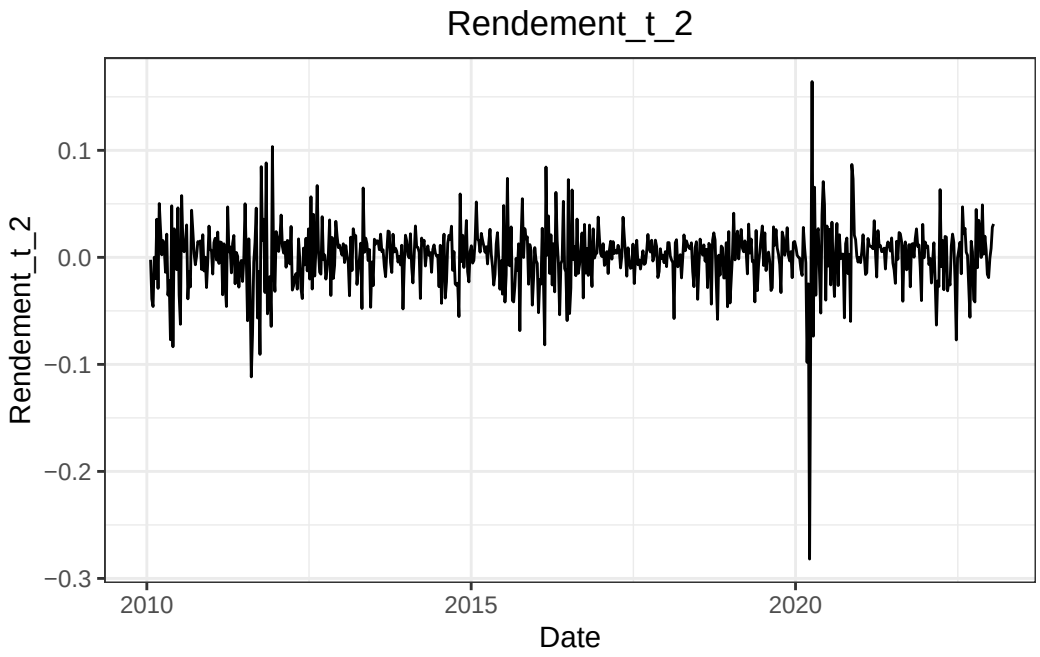

```
Warning: Removed 1 row containing missing values or values outside the scale range
(`geom_line()`).
```



```
Warning: Removed 2 rows containing missing values or values outside the scale range
(`geom_line()`).
```




```
Warning: Removed 3 rows containing missing values or values outside the scale range
(`geom_line()`).
```




```
# Transformation des colonnes en séries temporelles

CAC40 <- CAC40 |>
  mutate(across(Cours:Rendement_t_2, ~ ts(.x, start = c(2010, 1), frequency = 52)))
```

I.1. Estimation des modèles AR(1) et AR(2) en utilisant les séries retardées comme régresseurs et en utilisant les MCO.

```r
# Modèle des MCO pour AR(1)

lm_modele_ar1 <- lm(Rendement_t ~ Rendement_t_1, data = CAC40)

summary(lm_modele_ar1)
```

```
Call:
lm(formula = Rendement_t ~ Rendement_t_1, data = CAC40)

Residuals:
      Min        1Q    Median        3Q       Max
-0.284879 -0.013137  0.003678  0.014780  0.159234

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.0008852  0.0011241   0.787   0.4313
Rendement_t_1 -0.0856821  0.0382642  -2.239   0.0255 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0293 on 678 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.007341,  Adjusted R-squared:  0.005877
F-statistic: 5.014 on 1 and 678 DF,  p-value: 0.02547
```

La p-value de 0,02547 du test de Fisher indique que l'hypothèse nulle est rejetée au seuil de 5 %, ce qui signifie qu'au moins une variable est statistiquement significative.

Ensuite, la p-value de 0,0255 du test de Student associé au coefficient de la partie autorégressive montre que l'hypothèse nulle est rejetée au seuil de 5 %. Le rendement décalé d'une période (Rendement_t_1) est donc statistiquement significatif au seuil de 5 %.

```r
# Modèle des MCO pour AR(2)

lm_modele_ar2 <- lm(Rendement_t ~ Rendement_t_1 + Rendement_t_2, data = CAC40)

summary(lm_modele_ar2)
```

```
Call:
lm(formula = Rendement_t ~ Rendement_t_1 + Rendement_t_2, data = CAC40)

Residuals:
      Min        1Q    Median        3Q       Max
-0.286071 -0.013222  0.003772  0.014743  0.155965

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.0009542  0.0011255   0.848   0.3968
Rendement_t_1 -0.0868462  0.0384068  -2.261   0.0241 *
Rendement_t_2 -0.0111547  0.0384071  -0.290   0.7716
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0293 on 676 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.007521,  Adjusted R-squared:  0.004584
F-statistic: 2.561 on 2 and 676 DF,  p-value: 0.07796
```

La p-value de 0,07796 du test de Fisher indique que l'hypothèse nulle n'est pas rejetée au seuil de 5 %, ce qui signifie qu'aucune des variables testées n'est statistiquement significative à ce niveau de confiance.

Le modèle AR(1) est donc le meilleur, compte tenu du test de Fisher et de la significativité du coefficient autorégressif.

I.2. Estimation des modèles AR(1) et AR(2) en recourant directement à une fonction R disponible dans un des packages proposés.

```r
# Fonction arima pour estimer le modèle AR(1) (p = 1, d = 0, q = 0)

fn_modele_ar1 <- Arima(CAC40$Rendement_t, order = c(1, 0, 0))

fn_modele_ar1
```

```
Series: CAC40$Rendement_t
ARIMA(1,0,0) with non-zero mean

Coefficients:
          ar1   mean
      -0.0856  8e-04
s.e.   0.0382  1e-03

sigma^2 = 0.0008574:  log likelihood = 1439.18
AIC=-2872.37   AICc=-2872.33   BIC=-2858.8
```

Le modèle AR(1) à un AIC de -2872,37 et un BIC de -2858,8.

```r
# Fonction arima pour estimer le modèle AR(2) (p = 2, d = 0, q = 0)

fn_modele_ar2 <- Arima(CAC40$Rendement_t, order = c(2, 0, 0))

fn_modele_ar2
```

```
Series: CAC40$Rendement_t
ARIMA(2,0,0) with non-zero mean

Coefficients:
          ar1      ar2   mean
      -0.0865  -0.0112  8e-04
s.e.   0.0383   0.0383  1e-03

sigma^2 = 0.0008585:  log likelihood = 1439.23
AIC=-2870.45   AICc=-2870.39   BIC=-2852.36
```

Le modèle AR(2) à un AIC de -2870,45 et un BIC de -2852,36.

Le modèle AR(1), ayant un AIC et un BIC plus faibles, est donc le meilleur.

I.3. Conclusion

En utilisant la méthode des MCO et le modèle ARIMA sur la série, nos résultats convergent vers la même conclusion : le modèle AR(1) est le meilleur. La méthode des MCO montre que ce modèle AR(1) donne un coefficient significatif, contrairement au modèle AR(2) (où l'hypothèse nulle du test de Fisher est acceptée). De plus, l'estimation avec le modèle ARIMA indique que ce modèle présente une meilleure qualité de prévision, avec des critères AIC et BIC plus faibles.

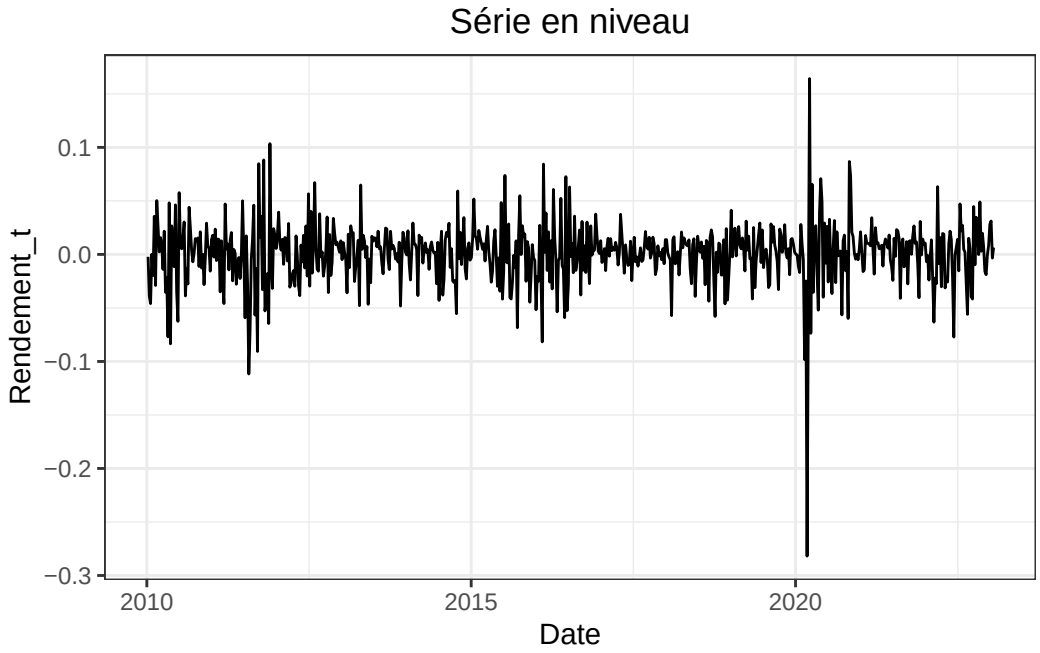
II. Modélisation ARMA du rendement d'un indice boursier

II.1. Analyse complète de la série des rendements

```
# 1. Graphique de la série en niveau
graphique_niveau <- CAC40 |>
  ggplot() +
  aes(x = Date, y = Rendement_t) +
  geom_line(na.rm = TRUE) +
  theme_bw() +
  ggtitle("Série en niveau") +
  theme(plot.title = element_text(hjust = 0.5))

# 2. Graphique de la série en différence première
graphique_difference <- CAC40 |>
  mutate(diff_Rendement_t = c(NA, diff(Rendement_t))) |>
  ggplot() +
  aes(x = Date, y = diff_Rendement_t) +
  geom_line(na.rm = TRUE) +
  theme_bw() +
  ggtitle("Série en différence première") +
  theme(plot.title = element_text(hjust = 0.5))

# 3. Autocorrélogramme (ACF)
graphique_acf <- CAC40$Rendement_t |>
  ggAcf() +
  ggtitle("Autocorrélogramme (ACF)") +
  theme(plot.title = element_text(hjust = 0.5))

# 4. Autocorrélogramme partiel (PACF)
graphique_pacf <- CAC40$Rendement_t |>
  ggPacf() +
  ggtitle("Autocorrélogramme partiel (PACF)") +
  theme(plot.title = element_text(hjust = 0.5))
```

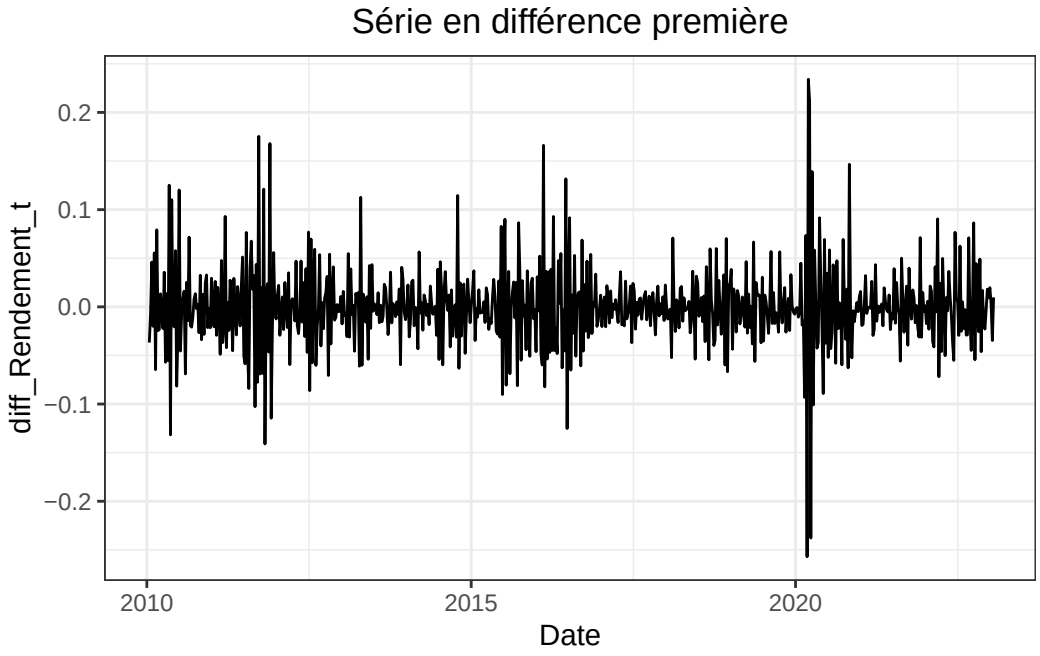
```
print(graphique_niveau)
```

```
Don't know how to automatically pick scale for object of type <ts>. Defaulting
to continuous.
```



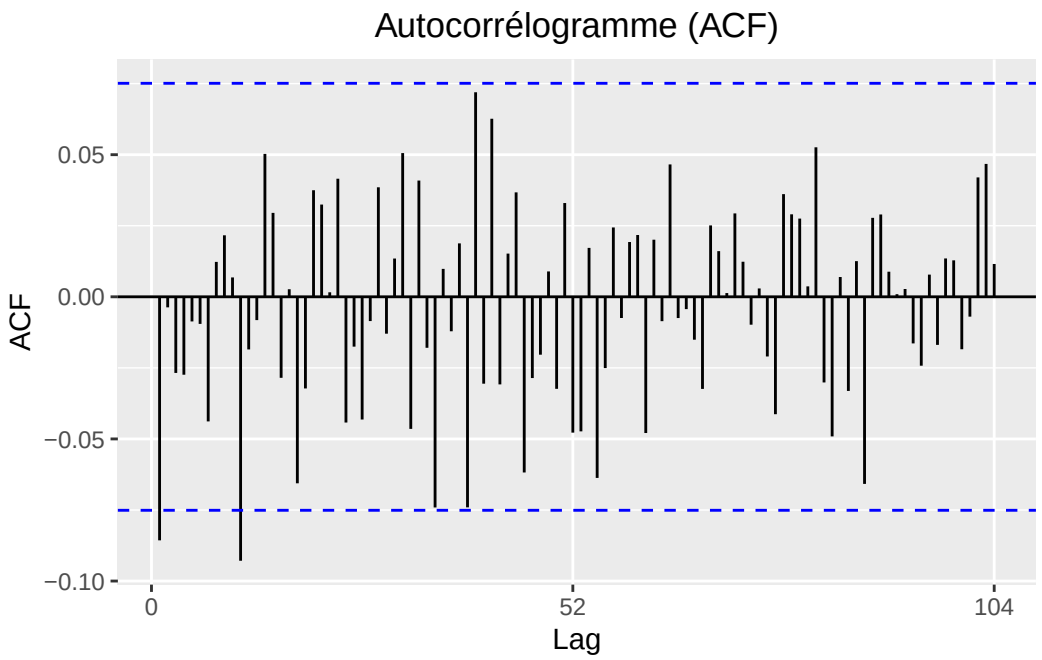
Le premier graphique représente la série des rendements en niveau. On observe une volatilité globalement stable sur la période, avec cependant des pics marqués autour de 2020, probablement liés à un événement de marché, comme la crise du COVID-19. La série semble présenter des périodes de volatilité accrue et d'autres plus calmes. La stationnarité de la série n'est pas évidente à première vue.

```r
print(graphique_difference)
```



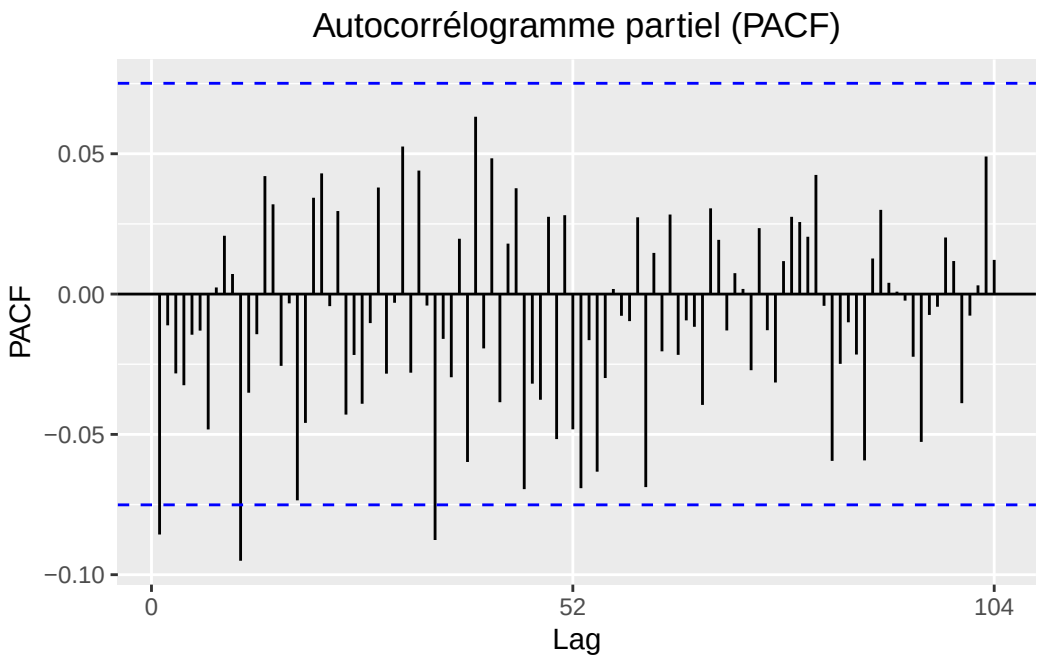
Le deuxième graphique montre la série en différence première, c'est-à-dire la variation des rendements d'une période à l'autre. On constate que la structure de la volatilité change légèrement : certaines périodes affichent une amplitude plus marquée, mais la tendance générale reste similaire, avec toujours un choc visible autour de 2020.

```r
print(graphique_acf)
```



L'autocorrélogramme montre des coefficients d'autocorrélation faibles.

```
print(graphique_pacf)
```



L'autocorrélogramme partiel indique des coefficients proches de zéro, avec quelques valeurs significatives aux premiers retards.

L'ACF et le PACF montrent peu d'autocorrélation persistante, indiquant un processus proche du bruit blanc. De plus, seuls quelques coefficients semblent significatifs, ce qui indique une faible dépendance aux valeurs passées pour la prédiction. Ce qui suggère un modèle ARMA de faible ordre (de faibles paramètres p et q).

II.2. Estimation des meilleurs modèles

```
# Toutes les combinaisons de p et q
TIBBLE <- expand_grid(p = factor(0:10), q = factor(0:2))

# Critères AIC et BIC
CRITERES <- TIBBLE |>
  mutate(
    modele = pmap(
      list(p, q),
      ~ Arima(na.omit(CAC40$Rendement_t), order = c(..1, 0, ..2))
    ),
    AIC = map_dbl(modele, AIC),
    BIC = map_dbl(modele, BIC)
  ) |>
  select(-modele)

CRITERES
```

```
# A tibble: 33 x 4
   p     q        AIC    BIC
   <fct> <fct>  <dbl>  <dbl>
 1 0     0     -2877. -2859.
 2 0     1     -2878. -2856.
 3 0     2     -2876. -2849.
 4 1     0     -2878. -2856.
 5 1     1     -2876. -2849.
 6 1     2     -2874. -2843.
 7 2     0     -2876. -2849.
 8 2     1     -2874. -2843.
 9 2     2     -2872. -2836.
10 3     0     -2874. -2843.
# i 23 more rows
```

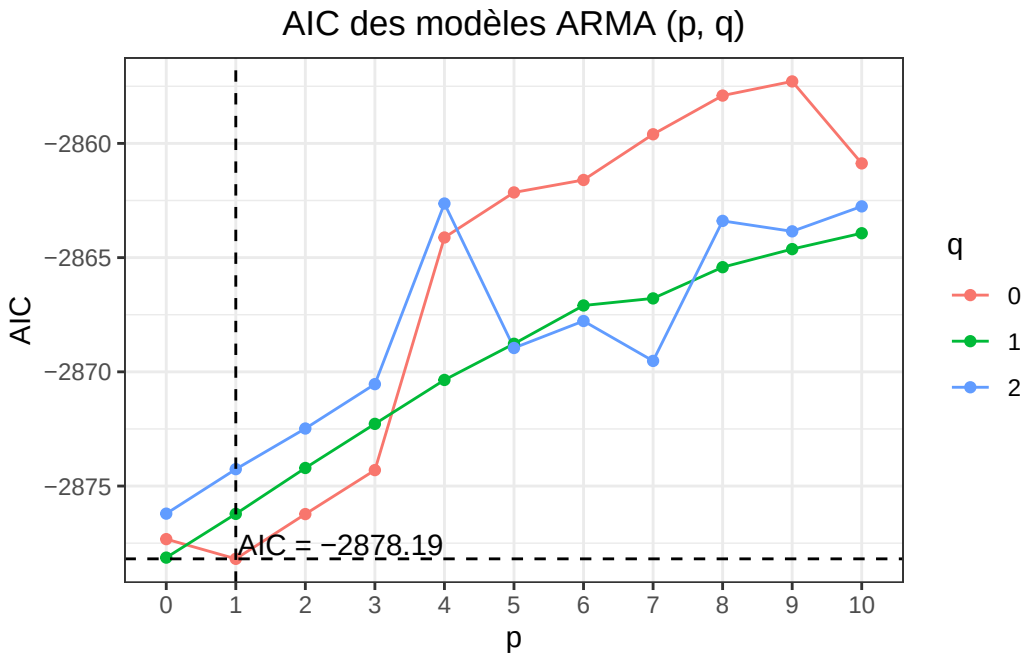
II.2.1. AIC

```
# AIC minimum

CRITERES_min_aic <- CRITERES |>
  filter(AIC == min(AIC))

CRITERES_min_aic
```

```
# A tibble: 1 x 4
  p     q         AIC    BIC
  <fct> <fct>   <dbl>  <dbl>
1 1     0     -2878. -2856.
```

```
# Graphique représentant le AIC de chaque modèle ARMA(p, q)

CRITERES |>
  ggplot(aes(x = p, y = AIC, color = q, group = q)) +
  geom_line() +
  geom_point() +
  geom_hline(
    aes(yintercept = CRITERES_min_aic$AIC),
    color = "black", linetype = "dashed"
  ) +
  geom_vline(
    aes(xintercept = CRITERES_min_aic$p),
    color = "black", linetype = "dashed"
  ) +
  geom_text(
    data = CRITERES_min_aic,
    aes(x = p, y = AIC, label = paste("AIC =", round(AIC, 2))),
    vjust = -0.2, hjust = -0.01,
    color = "black"
  ) +
  labs(
    title = "AIC des modèles ARMA (p, q)",
    x = "p",
    y = "AIC",
    color = "q"
  ) +
```

```
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))
```



Le modèle qui minimise l'AIC est le modèle ARMA(1,0).

```
# Modèle ARMA(1,0)

arma_1_0 <- Arima(CAC40$Rendement_t, order = c(1, 0, 0))

arma_1_0
```

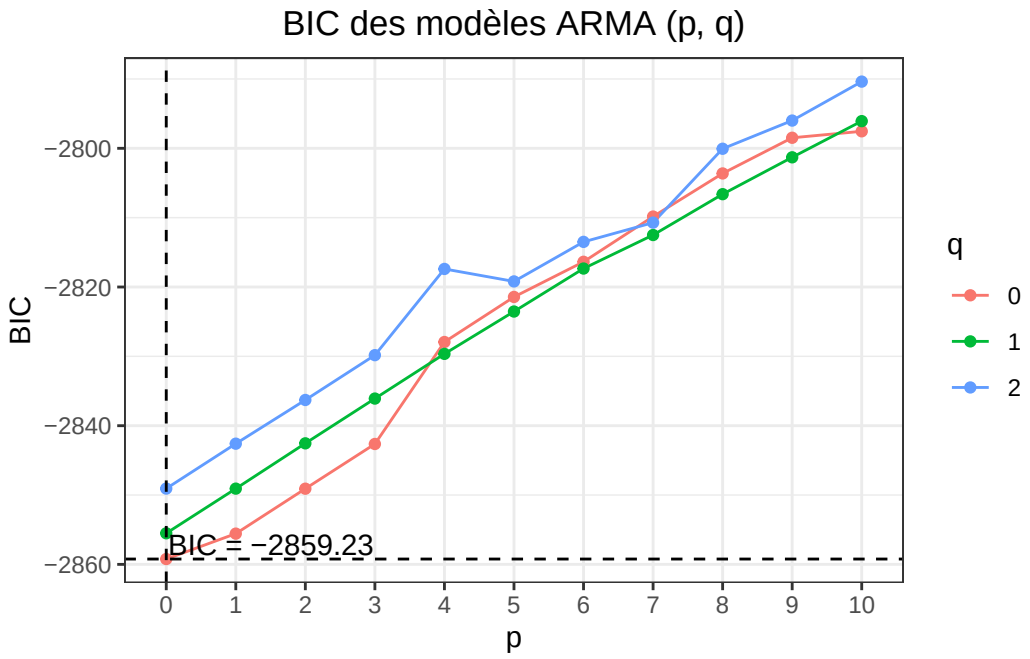
```
Series: CAC40$Rendement_t
ARIMA(1,0,0) with non-zero mean

Coefficients:
          ar1   mean
      -0.0856  8e-04
s.e.   0.0382  1e-03

sigma^2 = 0.0008574:  log likelihood = 1439.18
AIC=-2872.37   AICc=-2872.33   BIC=-2858.8
```

II.2.2. BIC

```
# BIC minimum

CRITERES_min_bic <- CRITERES |>
  filter(BIC == min(BIC))

CRITERES_min_bic
```

```
# A tibble: 1 x 4
  p     q        AIC    BIC
  <fct> <fct>  <dbl>  <dbl>
1 0     0     -2877. -2859.
```

```
# Graphique représentant le BIC de chaque modèle ARMA(p, q)
CRITERES |>
  ggplot(aes(x = p, y = BIC, color = q, group = q)) +
  geom_line() +
  geom_point() +
  geom_hline(
    aes(yintercept = CRITERES_min_bic$BIC),
    color = "black", linetype = "dashed"
  ) +
  geom_vline(
    aes(xintercept = CRITERES_min_bic$p),
    color = "black", linetype = "dashed"
  ) +
  geom_text(
    data = CRITERES_min_bic,
    aes(x = p, y = BIC, label = paste("BIC =", round(BIC, 2))),
    vjust = -0.2, hjust = -0.01,
    color = "black"
  ) +
  labs(
    title = "BIC des modèles ARMA (p, q)",
    x = "p",
    y = "BIC",
    color = "q"
  ) +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))
```



Le modèle qui minimise le BIC est le modèle ARMA(0,0).

```
# Modèle ARMA(0,0)

arma_0_0 <- Arima(CAC40$Rendement_t, order = c(0, 0, 0))

arma_0_0
```

```
Series: CAC40$Rendement_t
ARIMA(0,0,0) with non-zero mean

Coefficients:
        mean
      0.0008
s.e.  0.0011

sigma^2 = 0.0008625:  log likelihood = 1436.68
AIC=-2869.36   AICc=-2869.34   BIC=-2860.31
```

Le modèle choisi selon le critère AIC offre une meilleure qualité de prévision, car il pénalise moins la complexité que le critère BIC, notamment sur le lag de la moyenne mobile. En revanche, le critère BIC, en appliquant une pénalisation plus forte, limite le risque de surajustement aux erreurs passées.

II.3. Tests diagnostics

II.3.1. Meilleur modèle selon l'AIC : ARMA(1,0)

II.3.1.1. Test de Ljung-Box

```
# Test de Ljung-Box sur les résidus pour les 10 premiers lag

resultat_Ljung_Box <- tibble(
  lag = 1:10
) %>%
  mutate(
    test_result = map(lag, ~ Box.test(residuals(arma_1_0), lag = .x, type = "Ljung-Box")),
    X_squared = map_dbl(test_result, ~ .x$statistic),
    df = map_dbl(test_result, ~ .x$parameter),
    p_value = map_dbl(test_result, ~ .x$p.value),
    significativite = ifelse(
      p_value < 0.1,
      "Rejet de H0",
      "Non-rejet de H0"
    )
  ) %>%
  select(lag, X_squared, df, p_value, significativite)

resultat_Ljung_Box
```

```
# A tibble: 10 x 5
     lag X_squared    df p_value significativite
   <int>     <dbl> <dbl>   <dbl> <chr>
 1     1  0.000794     1   0.978 Non-rejet de H0
 2     2  0.125        2   0.939 Non-rejet de H0
 3     3  0.738        3   0.864 Non-rejet de H0
 4     4  1.39         4   0.845 Non-rejet de H0
 5     5  1.49         5   0.914 Non-rejet de H0
 6     6  1.63         6   0.950 Non-rejet de H0
 7     7  2.98         7   0.887 Non-rejet de H0
 8     8  3.06         8   0.931 Non-rejet de H0
 9     9  3.44         9   0.944 Non-rejet de H0
10    10  3.44        10   0.969 Non-rejet de H0
```

Pour ces dix premiers lags, l'hypothèse nulle n'est pas rejetée, ce qui indique une absence d'autocorrélation significative des résidus au seuil de 10 %. Le modèle est donc probablement bien spécifié.

II.3.1.2. Test de Bartlett

```
# Fonction

test_bartlett <- function(modele) {
  # Taille de l'échantillon
  T <- length(na.omit(CAC40$Rendement_t))

  # Coefficients du modèle
  coefficients <- modele$coef[-length(modele$coef)]

  # Calcul de l'écart-type en utilisant la formule de Bartlett
  ecart_type <- sqrt((1 / T) * (1 + 2 * sum(coefficients^2)))

  # Calcul de la statistique t pour chaque coefficient
  t_stats <- coefficients / ecart_type

  # Valeur critique
  valeur_critique <- 1.96

  # Tibble avec les résultats
  resultats <- tibble(
    coefficient = coefficients,
    ecart_type = ecart_type,
    t_stats = t_stats,
    significativite = ifelse(
      abs(t_stats) > valeur_critique,
      "Significatif",
      "Non significatif"
    )
  )

  return(resultats)
}
```

```
# Test de Bartlett

test_bartlett(arma_1_0)
```

```
# A tibble: 1 x 4
  coefficient ecart_type t_stats significativite
        <dbl>      <dbl>   <dbl> <chr>
1     -0.0856     0.0386   -2.22 Significatif
```

La valeur absolue de la statistique de test étant supérieure à 1,96, l'hypothèse nulle est rejetée, ce qui indique que le coefficient est significatif au seuil de 5 %.

II.3.2. Meilleur modèle selon le BIC : ARMA(0,0)

II.3.2.1. Test de Ljung-Box

```
# Test de Ljung-Box sur les résidus pour les 10 premiers lag

resultat_Ljung_Box <- tibble(
  lag = 1:10
) %>%
  mutate(
    test_result = map(lag, ~ Box.test(residuals(arma_0_0), lag = .x, type = "Ljung-Box")),
    X_squared = map_dbl(test_result, ~ .x$statistic),
    df = map_dbl(test_result, ~ .x$parameter),
    p_value = map_dbl(test_result, ~ .x$p.value),
    significativite = ifelse(
      p_value < 0.05,
      "Rejet de H0",
      "Non-rejet de H0"
    )
  ) %>%
  select(lag, X_squared, df, p_value, significativite)

resultat_Ljung_Box
```

```
# A tibble: 10 x 5
     lag X_squared    df p_value significativite
   <int>     <dbl> <dbl>   <dbl> <chr>
 1     1      5.02     1  0.0250 Rejet de H0
 2     2      5.03     2  0.0808 Non-rejet de H0
 3     3      5.52     3  0.137  Non-rejet de H0
 4     4      6.04     4  0.196  Non-rejet de H0
 5     5      6.09     5  0.298  Non-rejet de H0
 6     6      6.15     6  0.406  Non-rejet de H0
 7     7      7.48     7  0.381  Non-rejet de H0
 8     8      7.58     8  0.475  Non-rejet de H0
 9     9      7.91     9  0.543  Non-rejet de H0
10    10      7.94    10  0.635  Non-rejet de H0
```

Le rejet de l'hypothèse nulle au premier lag dans un modèle ARMA(0,0) indique qu'il existe une autocorrélation significative des résidus, ce qui suggère que le modèle est mal spécifié. Un modèle ARMA(0,0) devrait capturer un bruit blanc, où les résidus sont indépendants et sans autocorrélation. Ajouter un terme autoregressif ou de moyennes mobiles d'ordre supérieur pourrait permettre de mieux capturer la structure temporelle des données.

II.4. Conclusion

Le meilleur modèle est donc le modèle ARMA(1,0), soit AR(1). Ce modèle présente l'AIC minimal, il est bien spécifié selon le test de Ljung-Box, et le coefficient de la partie autorégressive est significatif selon le test de Bartlett.