TP Séries chronologiques

Laurent LIN & Guillaume SHI

24/02/2019

Partie 1 - Estimation du modèle

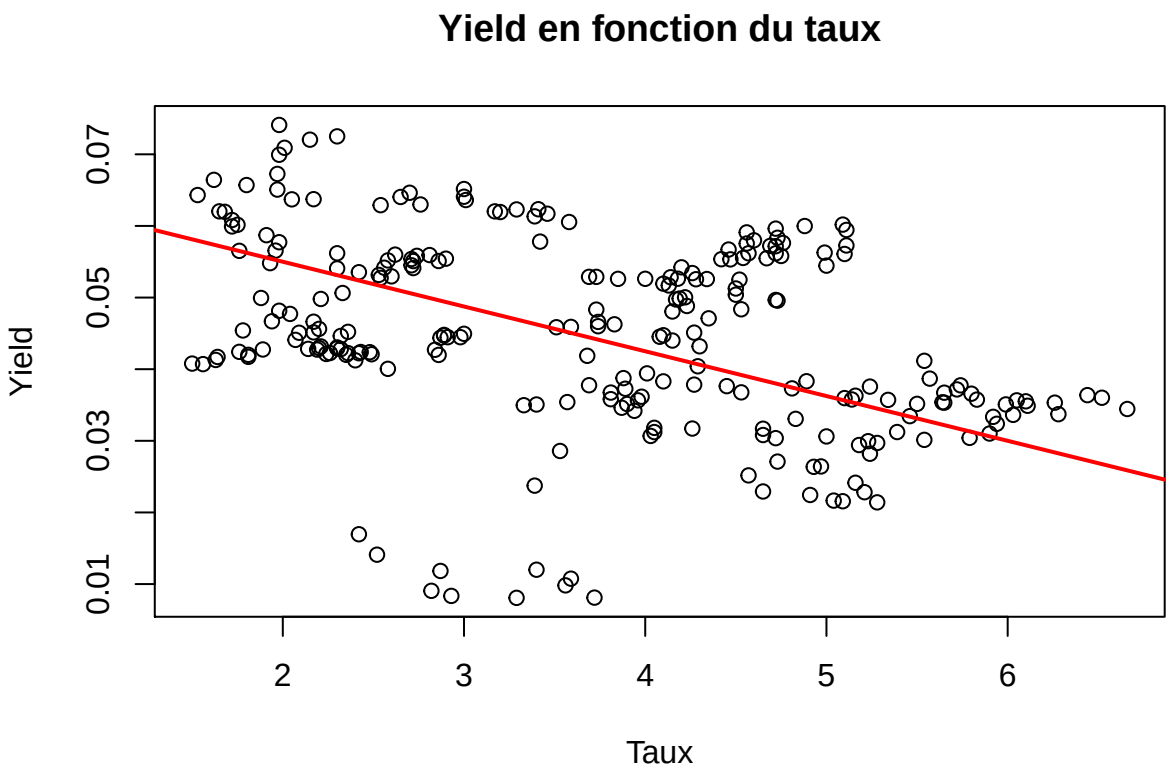
Question 1

Calculer le earning yield de l'indice du S&P500. Tracer un nuage de points liant le earning yield au taux sans risque. L'ajustement linéaire est-il justifié ? Utiliser la commande "abline" pour tracer cet ajustement.

L'earning yield est défini comme étant le rapport entre l'earning et le prix d'une action.

A première vue, on pourrait dégager une tendance linéaire avec une droite passant par 0.06 pour un taux à 2% et par 0.035 pour un taux à 6%, soit l'équation linéaire pour la régression : $y = -0.00625x + 0.0675$.

```
data <- fread("0219_td_centrale.csv", header = TRUE)
data$yield <- data$earnings/data$price

plot(data$rates, data$yield, xlab = "Taux", ylab = "Yield", main = "Yield en fonction du taux")
abline(0.0675, -0.00625, col = "red", lwd = 2)
```



Question 2

Rappeler le principe de la méthode des moindres carrées ordinaires. Rappeler les hypothèses sous-jacentes de l'estimateur des MCO.

On considère la régression linéaire du vecteur des variables d'intérêt $\mathbf{y}$ en fonction du vecteur des variables explicatives $\mathbf{X}$: il s'agit de trouver le vecteur $\boldsymbol{\gamma}$ tel que $\mathbf{y} = \mathbf{X}^T\boldsymbol{\gamma} + \boldsymbol{\epsilon}$ où $\boldsymbol{\epsilon}$ représente un bruit blanc. Les hypothèses sous-jacentes sont qu'il existe une relation linéaire entre $\mathbf{y}$ et $\mathbf{X}$, $\mathbf{X}$ doit être de rang maximum (non-colinéarité des variables explicatives), $\mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{0}$ (indépendance des erreurs) et $\mathbb{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T) = \sigma I_n$ avec σ constante (homoscédasticité).

La méthode des moindres carrés ordinaires (MCO) consiste à trouver le vecteur $\boldsymbol{\gamma}$ par minimisation de la quantité suivante : $\Delta = (\mathbf{y} - \mathbf{X}^T\boldsymbol{\gamma})^2$. Ce problème de minimisation se résout en dérivant la quantité Δ par rapport à chacune des composantes de $\boldsymbol{\gamma}$ et en écrivant que la quantité obtenue est nulle :

$$\forall 1 \leq i \leq n, \frac{\partial \Delta}{\partial \boldsymbol{\gamma}_i} = 0$$

Question 3

Calculer les expressions des coefficients des estimateurs des moindres carrés et montrer que la variance de beta tend vers 0 quand T vers vers l'infini.

On garde les notations de la question précédente. Ici, on souhaite calibrer $\boldsymbol{\gamma} = (\alpha \quad \beta)^T$ afin d'avoir une relation du type $\mathbf{y} = \frac{E_t}{P_t} = \mathbf{X}^T\boldsymbol{\gamma}$ si $\mathbf{X} = (1 \quad r_t)^T$ désigne le vecteur des variables explicatives. Etant en présence de variables aléatoires, la quantité à minimiser est

$$\mathbb{E}((\frac{E_t}{P_t} - (\alpha + \beta r_t))^2) = \mathbb{E}(\frac{E_t}{P_t}) - 2\mathbb{E}(\frac{E_t}{P_t}(\alpha + \beta r_t)) + \mathbb{E}((\alpha + \beta r_t)^2) = \Delta$$

- En dérivant par rapport à la première composante de $\boldsymbol{\gamma}$, c'est-à-dire α, on trouve

  $$-2\mathbb{E}(\frac{E_t}{P_t}) + 2\mathbb{E}(\alpha + \beta r_t) = 0$$

  soit $\alpha = \mathbb{E}(\frac{E_t}{P_t}) - \beta\mathbb{E}(r_t)$, ou encore

  $$\widehat{\alpha} = \overline{(\frac{E}{P})} - \widehat{\beta}\overline{r}$$

  en prenant la moyenne empirique des earning yields $\overline{(\frac{E}{P})}$ comme estimateur de $\mathbb{E}(\frac{E_t}{P_t})$ et $\overline{r}$ pour $\mathbb{E}(r_t)$.

- La dérivation par rapport à la seconde composante, β, donne

  $$-2\mathbb{E}(\frac{E_t}{P_t}r_t) + 2\mathbb{E}(r_t(\alpha + \beta r_t)) = 0 = -\mathbb{E}(\frac{E_t}{P_t}r_t) + \alpha\mathbb{E}(r_t) + \beta\mathbb{E}(r_t^2)$$

  soit encore en substituant à α sa valeur en fonction de β trouvée précédemment :

  $$-\mathbb{E}(\frac{E_t}{P_t}r_t) + \mathbb{E}(r_t)(\mathbb{E}(\frac{E_t}{P_t}) - \beta\mathbb{E}(r_t)) + \beta\mathbb{E}(r_t^2) = 0$$

d'où

$$-\mathbb{E}(\frac{E_t}{P_t}r_t) + \mathbb{E}(\frac{E_t}{P_t})\mathbb{E}(r_t) - \beta\mathbb{E}(r_t)^2 + \beta\mathbb{E}(r_t^2) = 0 = -Cov(\frac{E_t}{P_t}, r_t) + \beta Var(r_t)$$

d'où l'on tire ainsi

$$\widehat{\beta} = \frac{Cov(\frac{E}{P}, r)}{\sigma_r^2}$$

Montrons que $\lim_{T\to\infty} \text{Var}(\hat{\beta}) = 0$

- Pour calculer $\widehat{\beta}$, il faut utiliser les estimateurs empiriques des fonctions de la moyenne ($\widehat{x} = \frac{1}{T+1}\sum_{t=0}^{T} x_i$), la variance et de la covariance, ce qui donne : $\widehat{\beta} = \frac{\sum_{t=0}^{T}(r_t-\bar{r})(\frac{E_t}{P_t}-\overline{(\frac{E_t}{P_t})})}{\sum_{t=0}^{T}(r_t-\bar{r})^2}$.
- En injectant l'équation (2) décrivant le modèle vérifié par les données, on obtient : $\text{Var}(\widehat{\beta}) = \text{Var}(\frac{\sum_{t=0}^{T}(r_t-\bar{r})(\alpha+\beta r_t+\epsilon_t-\overline{(\frac{E_t}{P_t})})}{\sum_{t=0}^{T}(r_t-\bar{r})^2})$
- On remarque qu'au numérateur, les ϵ_t sont les seules variables aléatoires, le reste peut être supprimé de la variance, on obtient donc : $\text{Var}(\widehat{\beta}) = \text{Var}(\frac{\sum_{t=0}^{T}(r_t-\bar{r})\epsilon_t}{\sum_{t=0}^{T}(r_t-\bar{r})^2})$
- Par hypothèse, les ϵ_t sont i.i.d. donc on obtient finalement :

$$\text{Var}(\widehat{\beta}) = \sum_{t=0}^{T}\text{Var}(\frac{(r_t-\hat{r})\epsilon_t}{\sum_{t=0}^{T}(r_t-\bar{r})^2}) = \frac{\sum_{t=0}^{T}(r_t-\bar{r})^2\text{Var}(\epsilon_t)}{(\sum_{t=0}^{T}(r_t-\bar{r})^2)^2} = \sigma_\epsilon^2\frac{\sum_{t=0}^{T}(r_t-\bar{r})^2}{(\sum_{t=0}^{T}(r_t-\bar{r})^2)^2} = \frac{\sigma_\epsilon^2}{\sum_{t=0}^{T}(r_t-\bar{r})^2}$$

- Comme le dénominateur est une somme de termes positifs, lorsque $T \to \infty$, $\sum_{t=0}^{T}(r_t-\bar{r})^2 \to \infty$ et donc $\lim_{T\to\infty} \text{Var}(\hat{\beta}) = 0$

Question 4 - Interprétation des résultats de l'estimateur des MCO

A l'aide de la fonction "lm" du logiciel R, estimer par les MCO les coefficients de l'équation 2. Commenter vos résultats en particulier le signe du coefficient bêta et sa significativité. Qu'indiquent les statistiques de Student et de Fisher ainsi que le coefficient de détermination ?

```
fit = lm(data$yield ~ data$rates)
summary(fit)
```

```
##
## Call:
## lm(formula = data$yield ~ data$rates)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.039266 -0.007419 -0.000533  0.008763  0.022890
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0586936  0.0023539  24.935  < 2e-16 ***
## data$rates  -0.0037816  0.0006005  -6.297 1.38e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01244 on 247 degrees of freedom
## Multiple R-squared:  0.1383, Adjusted R-squared:  0.1349
## F-statistic: 39.66 on 1 and 247 DF,  p-value: 1.375e-09
```

- $\widehat{\beta} = -0.0036653$ est de signe négatif, ce qui signifie que le earning yield diminue avec l'augmentation du taux sans risque.
- $\widehat{\alpha} = 0.0586936$ (l'`intercept`) est de signe positif, cela signifie que le earning yield vaut 0.0586936 lorsque le taux sans risque est nul.
- La statistique de Student (`t-value`) est un critère qui permet de statuer sur l'hypothèse nulle suivante : $H_0 = \{$le coefficient considéré vaut 0$\}$. On peut alors observer la p-value associée (`Pr(>|t|)`) : ici, sa valeur vaut `2.94e-08` pour $\widehat{\beta}$ et `< 2e-16` pour $\widehat{\alpha}$, elle est donc inférieure aux seuils de significativité classiques (0.1%, 1%, 5% etc.) et nous pouvons rejeter donc H_0 et dire que ces coefficients sont significativement différents de 0.

- Le F-test considère l'hypothèse nulle suivante : $H_0 = \{$les coefficients (autre que l'`intercept`) valent tous 0$\}$, il s'agit d'une comparaison avec le modèle constitué uniquement d'une estimation de l'`intercept`, i.e. l'ordonnée à l'origine. La p-value associée à la statistique de Fisher valant `2.944e-08`, elle est inférieure aux seuils de significativité classiques (0.1%, 1%, 5% etc.) c'est-à-dire qu'au moins une variable n'est pas significativement différente de 0 donc on peut rejeter cette hypothèse nulle, ie que le modèle dans l'ensemble est significatif.

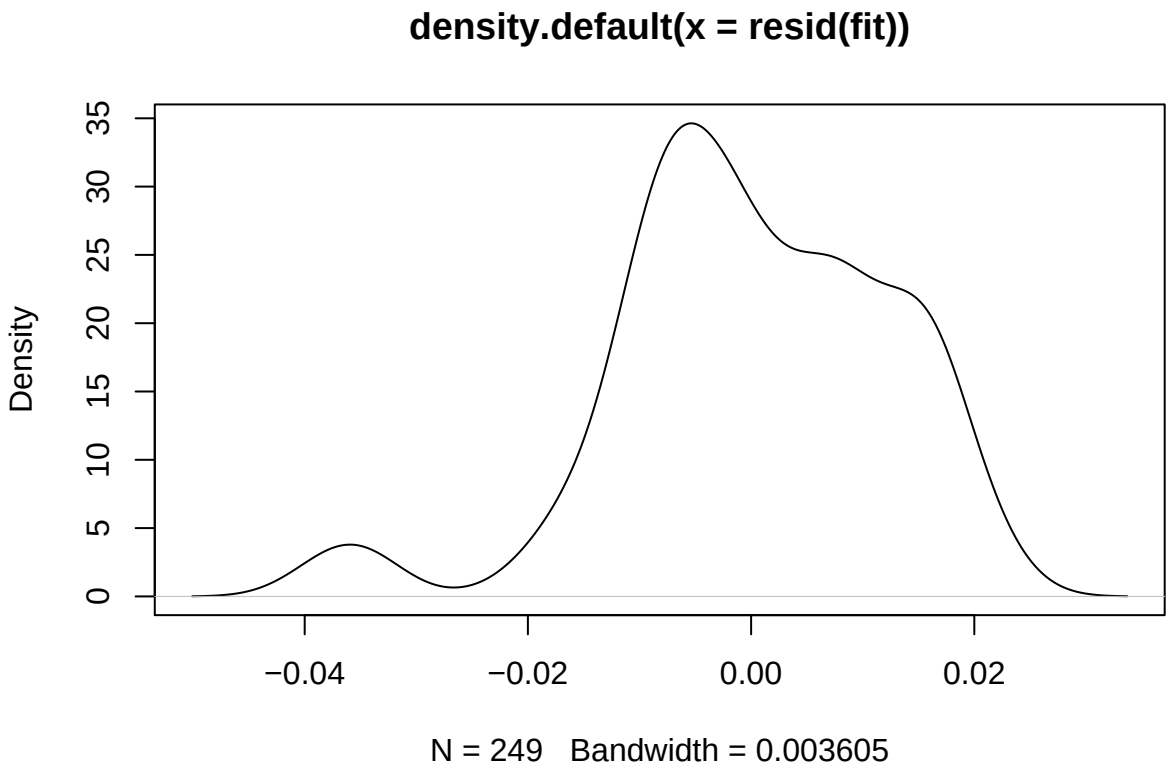
Le coefficient de détermination R^2 (`R-squared = 0.1229`) et le coefficient de détermination ajusté (`Adjusted R-squared = 0.1191`) sont tous les deux très faibles, indiquant que seulement une faible part d'environ 12% de la variable en question $\frac{E_t}{P_t}$ est expliquée par le modèle, et donc que la qualité de la régression est médiocre.

Question 5 - Résidus estimés

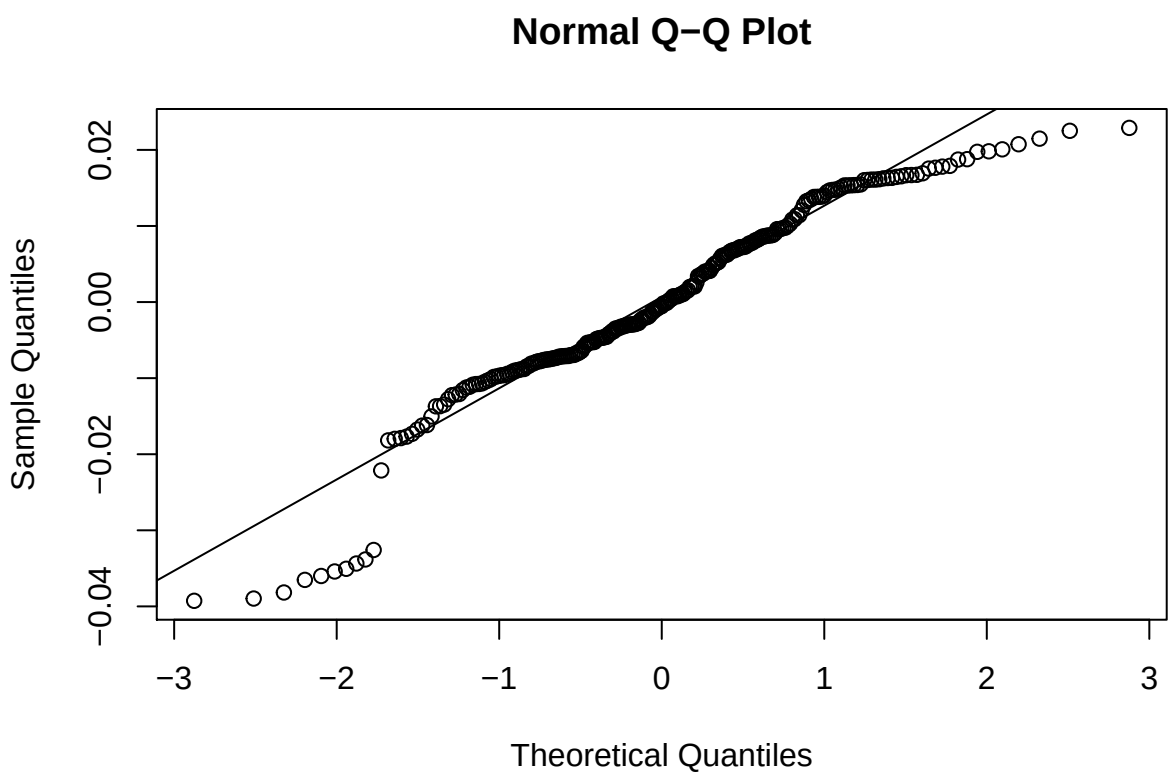
Réaliser une étude complète des résidus estimés : tracer leur densité, étudier leur normalité et vérifier l'existence/l'absence d'autocorrélation et d'hétéroscédasticité.

- Normalité des résidus du modèle : En visualisant la densité des résidus, nous voyons que celle-ci n'a pas la forme de cloche symétrique que possède une distribution normale classique. Cette idée est confirmée par le graphe quantile-quantile (qq-plot) : nous voyons que les points s'éloignent en queue de distribution. Le test de Shapiro-Wilk est effectué avec comme hypothèse nulle $H_0 = \{$l'échantillon est issu d'une population normalement distribuée$\}$, la p-valeur associée étant de $1.233e-07 < 0.05$, nous pouvons rejeter H_0 : il est probable que les données ne soient pas issues d'une population normalement distribuée.

```
plot(density(resid(fit)))
```




```
qqnorm(resid(fit))
qqline(resid(fit))
```

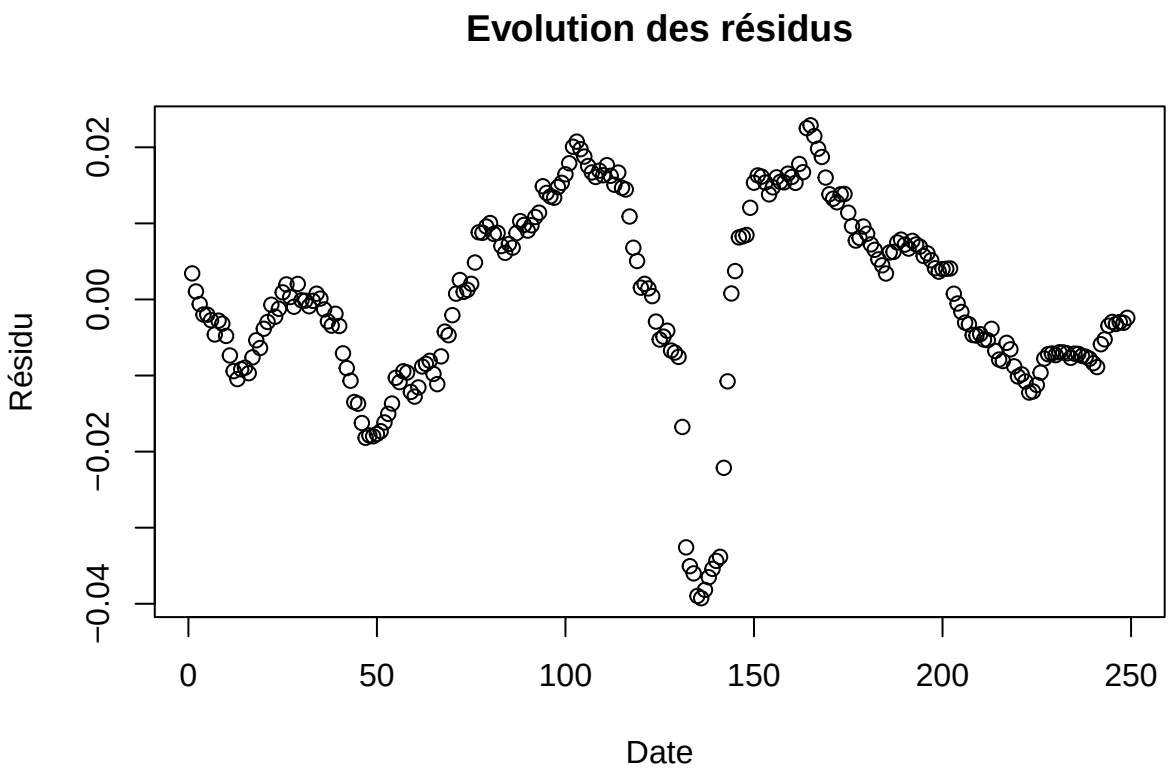


```r
shapiro.test(resid(fit))
```

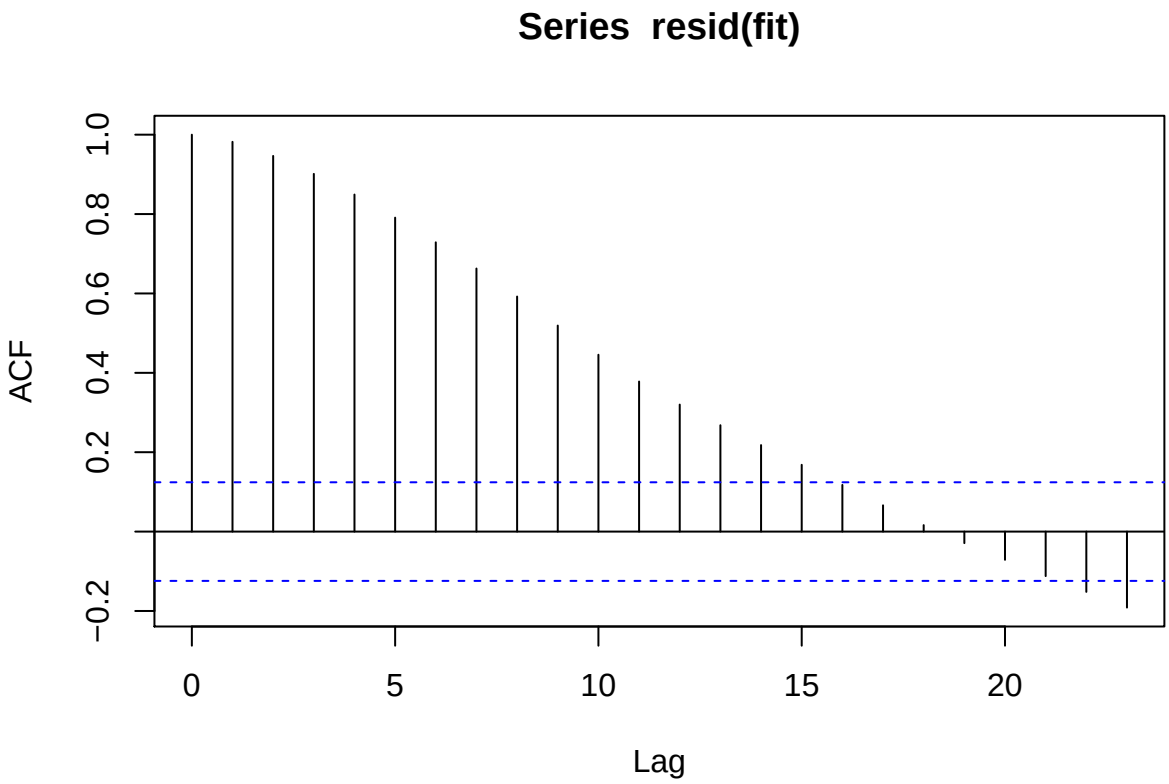
```
##
##  Shapiro-Wilk normality test
##
## data:  resid(fit)
## W = 0.94914, p-value = 1.233e-07
```

- Autocorrélation des résidus : on constate qu'il y a existe une importante autocorrélation des résidus : la fonction d'autocorrélation prend des valeurs importantes (en dehors des bandes bleues donc non nulles de manière statistiquement significative) pour des lags allant de 1 à 14 jours. Cela signifie que le modèle n'est pas optimal car les résidus peuvent être utilisés pour prédire les prochains résidus : il y a de l'information qui échappe au modèle actuel.

```r
plot(resid(fit), main = "Evolution des résidus", xlab = "Date", ylab = "Résidu")
```




```r
acf(resid(fit))
```



- Hétéroscédasticité des résidus : L'observation sur les résidus nous fait penser que ceux-ci sont hétéroscédastiques : la variance des résidus a augmenté puis diminué au cours du temps. Le test de Breusch-Pagan donne une p-valeur de 0.1577, on rejeterait l'hypothèse nulle H0 = {les données sont homoscédastiques} avec un seuil de confiance de 20%, ce qui est élevé comparé à tous les autres seuils de confiance obtenus précédemment, donc il est préférable de dire qu'il est impossible de conclure sur l'homoscédasticité des résidus.

```
bptest(fit)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  fit
## BP = 1.9966, df = 1, p-value = 0.1577
```

Conclusion : Outre le fait que le modèle ait un coefficient de détermination faible, l'étude des résidus a montré que ces derniers ne suivent pas une distribution normale, il y a présence d'autocorrélation et il est impossible de conclure sur l'homoscédasticité. Modéliser le earning yield en utilisant uniquement le taux sans risque ne permet pas d'avoir une modélisation où les hypothèses sous-jacentes de l'estimateur des MCO sont bien respectées, il faudrait améliorer le modèle.

Partie 2 - Estimation d'une nouvelle spécification et comparaison

Question 6

Calculer le taux d'intérêt réel. Estimer par la méthode des moindres carrés ordinaires cette nouvelle spécification. Améliore-t-elle le pouvoir explicatif du modèle ? Justifier votre réponse.

Le taux d'intérêt réel est calculé en utilisant l'indice des prix à la consommation (CPI) sur 12 mois de la façon suivante :

$$\pi_t = 100 \frac{cpi_t - cpi_{t-12}}{cpi_{t-12}}$$

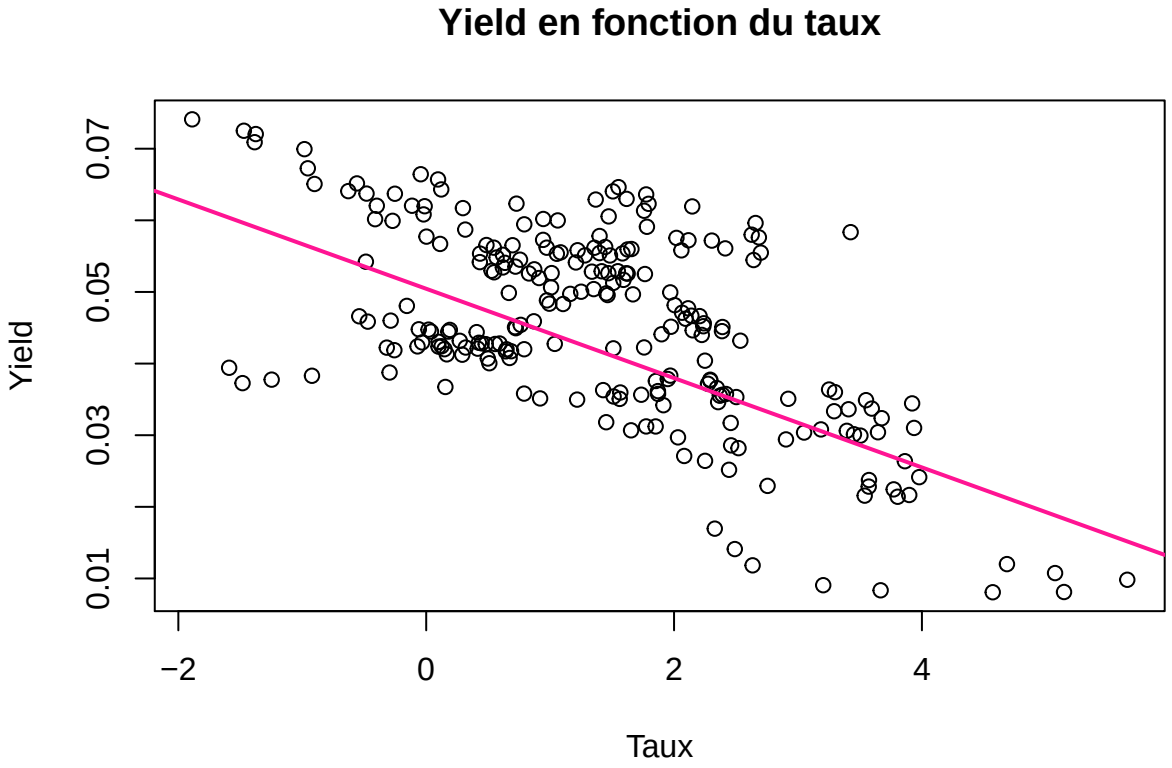
On omet donc les 12 premiers mois pour pour la construction du modèle Fed corrigé :

$$\frac{E_t}{P_t} = \alpha + \beta(r_t - \pi_t) + \epsilon_t$$

```
inflation = 100 * (data$cpi[-(1:12)]/data$cpi[1:(length(data$cpi) -
    12)] - 1)  # retrouver l'inflation à partir du CPI

data = data[-(1:12), ]

data$inflation = inflation

data$real_rates = data$rates - data$inflation

fit2 = lm(data$yield ~ data$real_rates)

summary(fit2)
```

```
##
## Call:
## lm(formula = data$yield ~ data$real_rates)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.026014 -0.008460  0.001518  0.007572  0.025657
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.0540445  0.0009952   54.30   <2e-16 ***
## data$real_rates -0.0062351  0.0005060  -12.32   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01058 on 235 degrees of freedom
## Multiple R-squared:  0.3925, Adjusted R-squared:   0.39
## F-statistic: 151.9 on 1 and 235 DF,  p-value: < 2.2e-16
```

```
plot(data$real_rates, data$yield, xlab = "Taux", ylab = "Yield", main = "Yield en fonction du taux")
abline(0.050445, -0.0062351, col = "deeppink", lwd = 2)
```



Les résultats de la régression sont les suivants :

- $\widehat{\beta} = -0.0062351$
- $\widehat{\alpha} = 0.0540445$

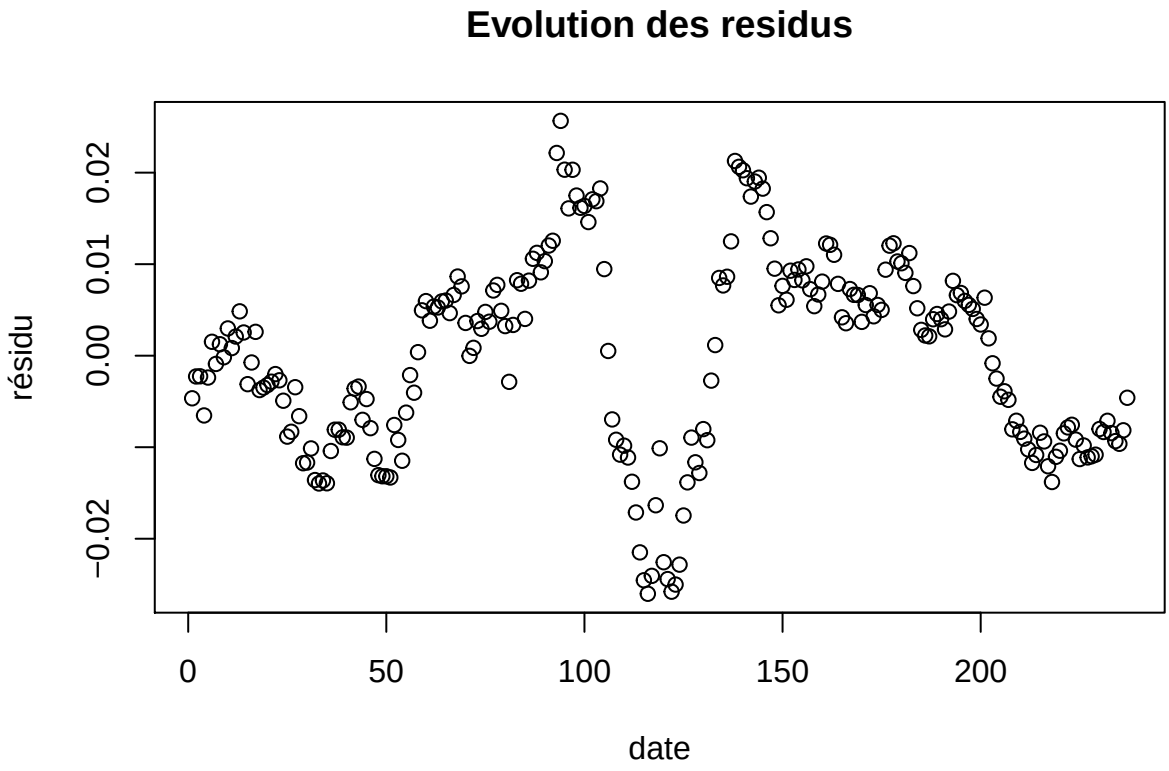
Ces coefficients sont de même signe et assez proches de ceux du modèle précédent. En ce qui concerne la significativité des coefficients, la statistique de Student ainsi que le test Fisher suggèrent des p-valeurs de l'ordre de $2e-16$ et donc que les coefficients sont significativement différents de 0.

Concernant le pouvoir explicatif du modèle :

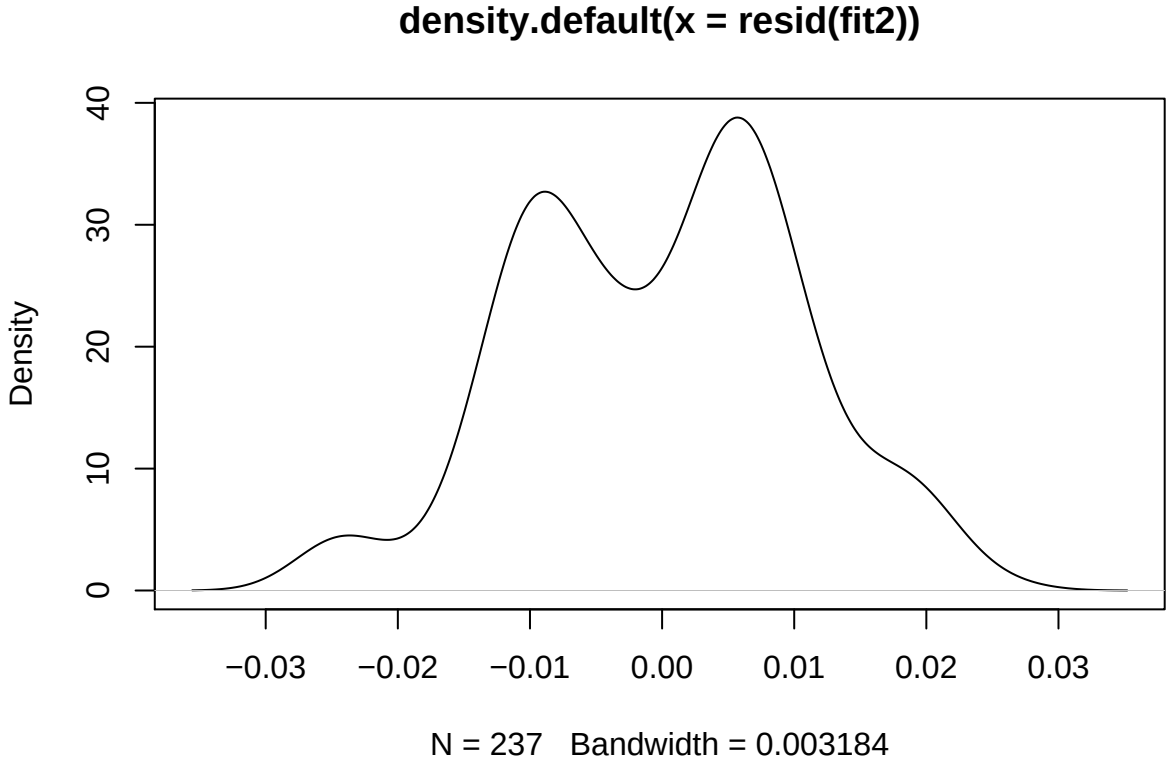
- Le coefficient de détermination R^2 est à présent de 0.39, ce qui est une amélioration nette par rapport à précédemment où il était de 0.14.
- Toutefois, une analyse plus poussée des résidus montre que les résidus ne suivent toujours pas une distribution normale, et ne sont pas décorrélés et il est impossible de conclure sur leur homoscédasticité. La fonction de densité présente deux pics, et les points en queue de distribution qui s'éloignent de la valeur théorique du diagramme quantile-quantile et le test de Shapiro-Wilk donnant une p-valeur très faible de 0.004569 permettent de rejeter l'hypothèse nulle selon laquelle les données suivraient une distribution normale. La fonction d'autocorrélation montre que les résidus ne sont pas décorrélés. Le test de Breusch-Pagan donne une p-valeur de 0.7441 donc nous ne pouvons pas conclure sur l'homoscédasticité des résidus.

L'amélioration du pouvoir explicatif du modèle se traduit alors uniquement par un meilleur coefficient de détermination, mais le modèle linéaire choisi ne permet toujours pas d'obtenir des résidus conformes aux hypothèses de modélisation du modèle MCO.

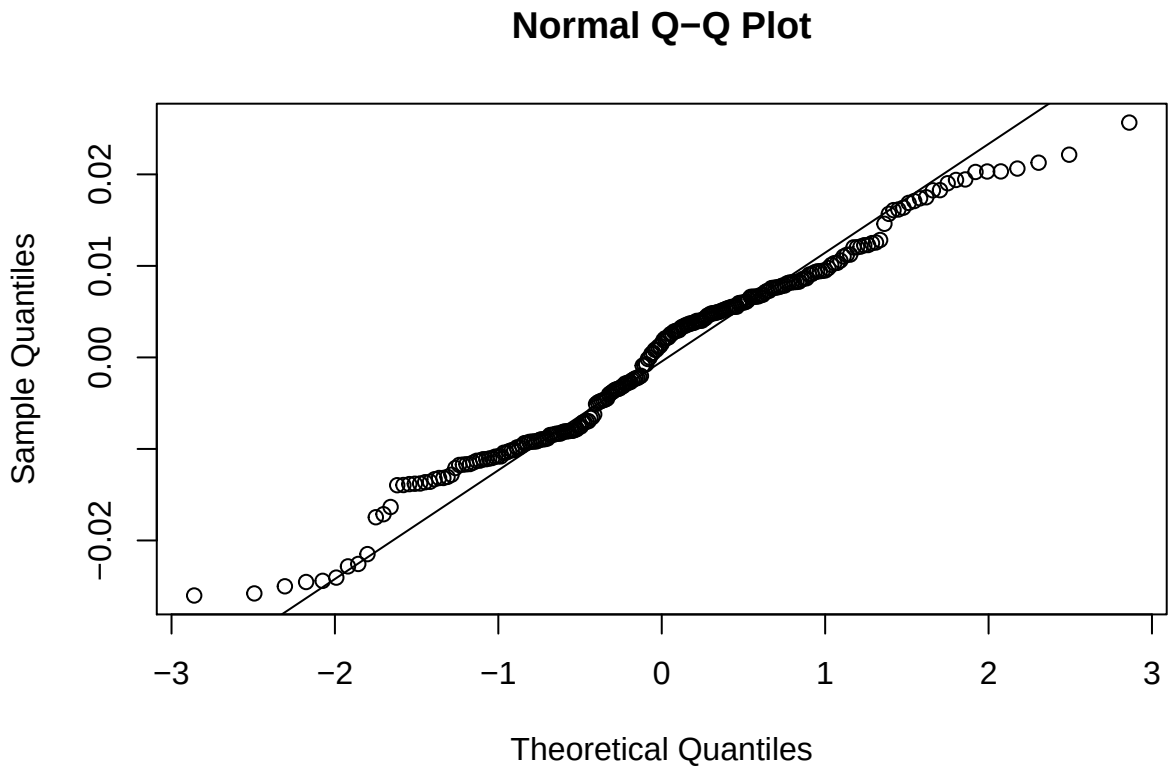
```r
plot(resid(fit2), main = "Evolution des residus", xlab = "date", ylab = "résidu")
```




```r
plot(density(resid(fit2)))
```



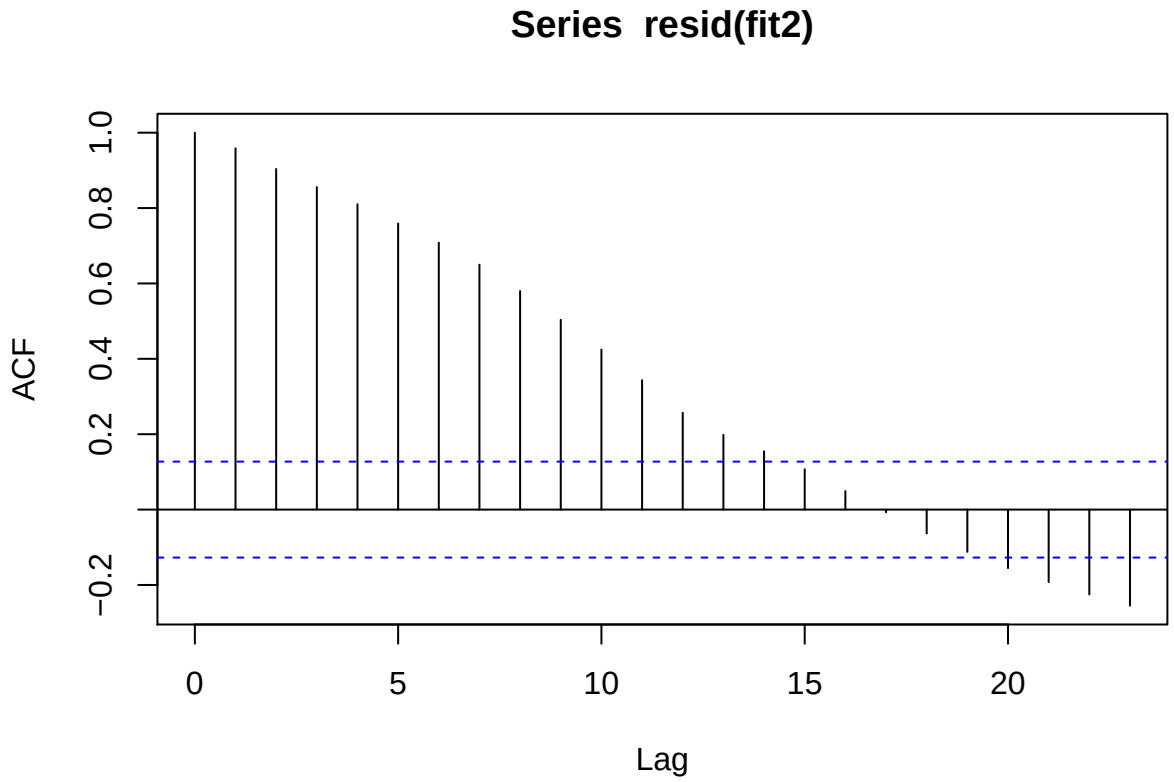
```r
qqnorm(resid(fit2))
qqline(resid(fit2))
```




```r
shapiro.test(resid(fit2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(fit2)
## W = 0.98223, p-value = 0.004569
```

```r
acf(resid(fit2))
```




```
bptest(fit2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  fit2
## BP = 0.10656, df = 1, p-value = 0.7441
```

Partie 3 - Estimation d'une nouvelle spécification : modèle ARMA(p,q)

Question 7

Présenter le modèle ARMA et ses principales propriétés de manière précise et succinte.

Le modèle ARMA est un modèle mathématique visant à modéliser le comportement d'une série temporelle en fonction de ses valeurs historiques (c'est la partie autorégressive –AR– du modèle) mais aussi en fonction d'un bruit blanc qui viendrait perturber les données (d'où une moyenne mobile dans le modèle –MA–), dans le but de prédire les valeurs que prendrait la série temporelle dans le futur.

Pour une variable représentée par une série temporelle, on introduit généralement les ordres p et q et on parle d'un modèle ARMA(p,q) pour modéliser une valeur au temps t à partir des p valeurs précédentes et de q bruits blancs indépendants et identiquement distribués :

$$X_t = \phi_1 X_{t-1} + ... + \phi_p X_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + ... + \theta_q \epsilon_{t-q}$$

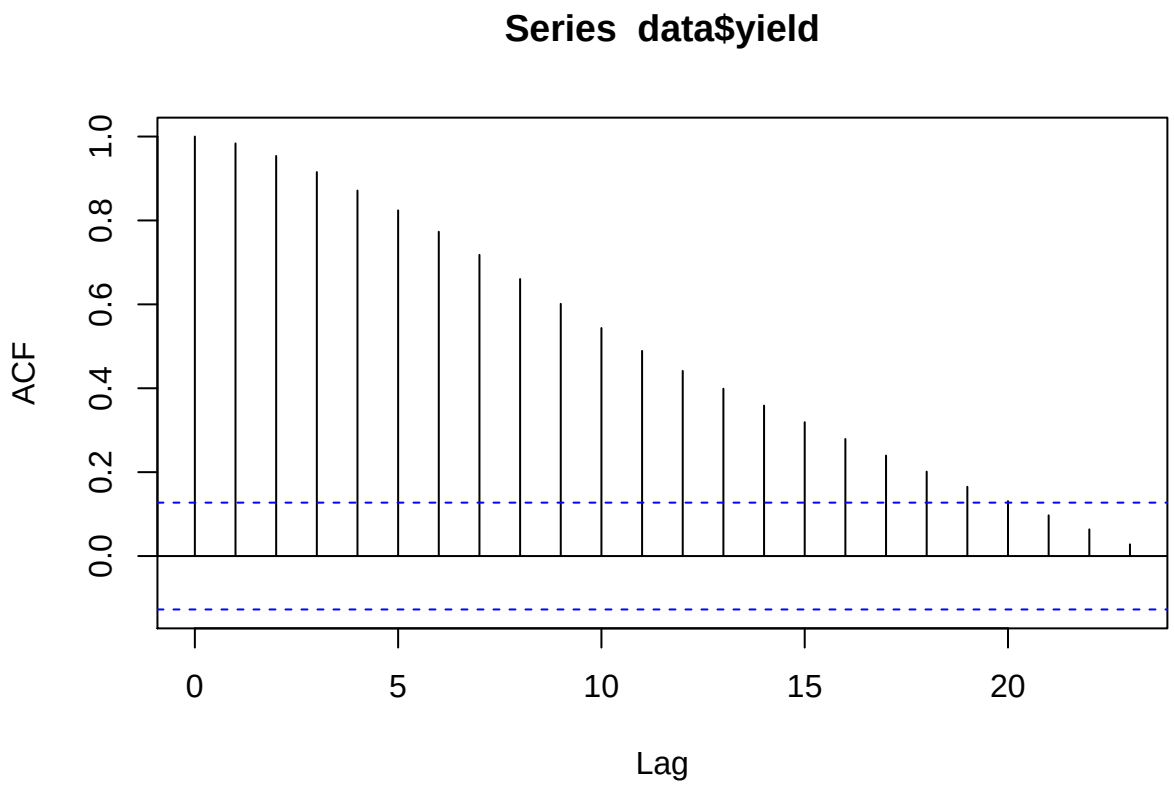
Question 8

Identifier l'ordre du modèle ARMA à l'aide de deux méthodes différentes (on privilégiera une approche parcimonieuse).

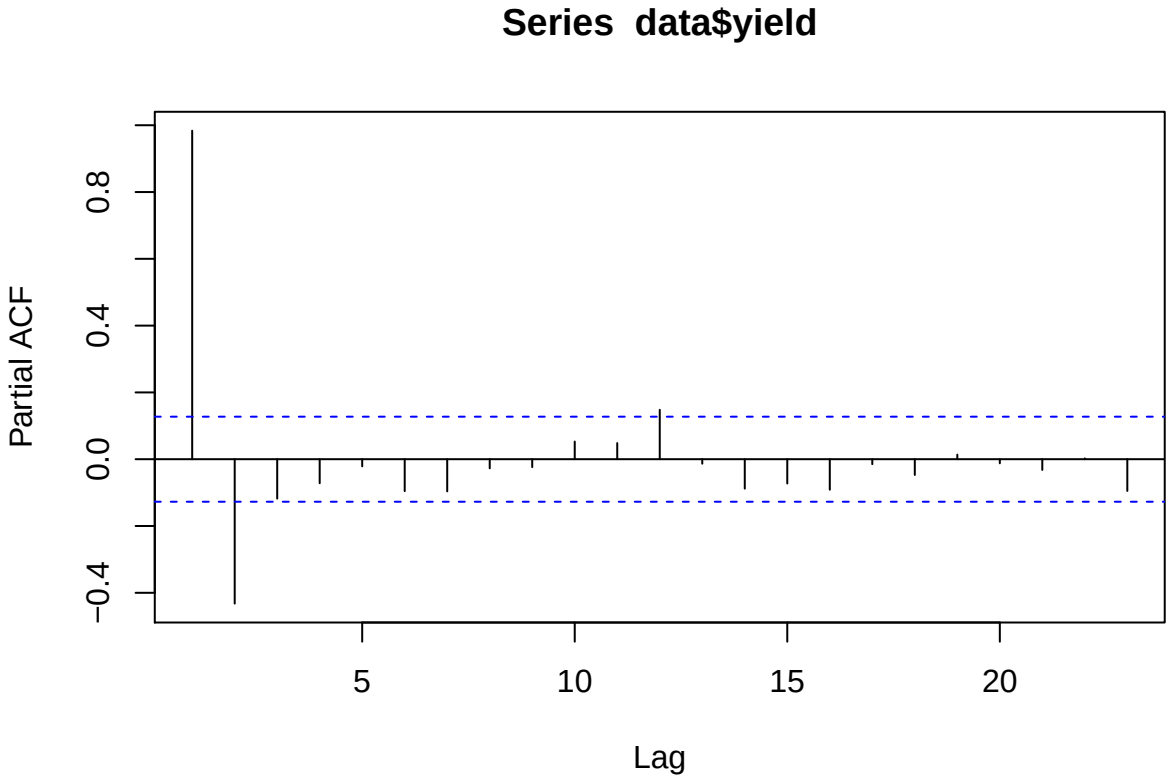
Première approche

La première approche serait de regarder la fonction d'autocorrélation ainsi que la fonction d'autocorrélation partielle. En effet, celles-ci donnent des informations sur les ordres q et p respectivement du modèle que l'on pourrait utiliser pour modéliser la série temporelle des yields.

```
acf(data$yield)  # fonction d'autocorrélation
```




```
pacf(data$yield)  # fonction d'autocorrélation partielle
```



La fonction d'autocorrélation sur le premier graphe ne nous apprend pas grand-chose sur la dépendance entre X_t et d'éventuels bruits blancs car pratiquement tous les pics sont hors de la zone délimitée par les bandes bleues en pointillés. Les pics dans cette zone peuvent être considérés comme nuls. Or, on vo de manière statistiquement significativeit que les pics sont très au-delà de cette zone, et ceci pour des lags jusqu'à environ 19 : on ne peut donc pas présumer de la valeur de q.

En revanche, à l'aide de la fonction d'autocorrélation partielle présentée sur le second graphe, on observe que deux pics ne se situent pas entre les pontillés bleus : ceux qui correspondent à un lag de 1 et à un lag de 2. On a donc deux pics qui se distringuent, et **on peut ainsi supposer que l'ordre pour la partie autorégressive du modèle, p, vaut 2**.

On a considéré que le pic du lag à 12 était nul bien qu'il reflète une saisonnalité des données : la valeur du yield présente une certaine corrélation par rapport à sa valeur il y a 12 mois. Néanmoins, pour des raisons de simplicité, on ne traite pas cette corrélation.

Deuxième approche

La seconde approche serait une approche "force brute" : on teste toutes les combinaisons possibles de p et q (pour p et q entiers inférieurs à une valeur limite, typiquement 4), et on regarde quels modèles permettent de mieux modéliser les données empiriques, via la fonction `arima` par exemple. Pour cela, on s'attache à regarder des critères d'information tels que l'AIC (Akaike Information Criteria) ou le BIC (Bayesian Information Criteria) qui donnent une estimation de la qualité du modèle fitté. Ces critères font une sorte de compromis entre la complexité des modèles utilisés et la qualité d'estimation de ceux-ci par rapport aux données empiriques. On choisit ensuite le modèle qui donne le plus petit critère.

La fonction `arima` ne renvoyant que l'AIC pour un modèle, on calcule le BIC à partir de la fonction `BIC` de R.

```r
order_max = 4

order_p = c()
order_q = c()
aic = c()
bic = c()

for (p in 0:order_max) {
    for (q in 0:order_max) {
        order_p = c(order_p, p)
        order_q = c(order_q, q)
        model = arima(data$yield, order = c(p, 0, q))

        aic = c(aic, model$aic)
        bic = c(bic, BIC(model))
    }
}

arma_models = data.frame(order_p, order_q, aic, bic)

arma_models[which(arma_models$aic == min(arma_models$aic)), ]  # modèle avec AIC minimal
```

```
##    order_p order_q       aic       bic
## 13       2       2 -2279.704 -2258.895
```

```r
arma_models[which(arma_models$bic == min(arma_models$bic)), ]  # modèle avec BIC minimal
```

```
##    order_p order_q       aic       bic
## 11       2       0 -2279.087 -2265.215
```

Le modèle minimisant le critère AIC est un modèle ARMA(2,2) tandis que le modèle qui minimise le critère BIC est un modèle ARMA(2,0). On voit ainsi que les deux critères d'information ne sont pas équivalents : en effet, le critère BIC prend en compte la taille de l'échantillon. Quel que soit le critère à minimiser, le modèle correspondant donne un ordre p égal à 2, ce qui confirme la conclusion de la première approche.

En fonction du critère à minimiser, on choisira plutôt un modèle ARMA(2,0) ou un modèle ARMA(2,2).

Question 9

Estimer le modèle identifié à l'aide de la fonction auto.arima et vérifier la qualité de votre estimation.

```r
arima_data = auto.arima(data$yield, d = 0, D = 0, ic = "aic")
mean(data$yield)
```

```
## [1] 0.04517566
```

```r
summary(arima_data)
```

```
## Series: data$yield
## ARIMA(2,0,2) with non-zero mean
##
## Coefficients:
##          ar1      ar2      ma1      ma2    mean
##       1.8257  -0.8388  -0.4021  -0.1616  0.0446
```

```
## s.e.  0.0847   0.0821   0.1113   0.0849  0.0040
##
## sigma^2 estimated as 3.711e-06:  log likelihood=1145.85
## AIC=-2279.7   AICc=-2279.34   BIC=-2258.9
##
## Training set error measures:
##                       ME        RMSE        MAE        MPE     MAPE
## Training set 3.12455e-05 0.001906076 0.00127265 -0.1796656 3.339317
##                    MASE        ACF1
## Training set 0.9218602 0.005506432
```

```r
arima_data_bic = auto.arima(data$yield, d = 0, D = 0, ic = "bic")
summary(arima_data_bic)
```

```
## Series: data$yield
## ARIMA(2,0,0) with non-zero mean
##
## Coefficients:
##          ar1      ar2    mean
##       1.4864  -0.5081  0.0441
## s.e.  0.0555   0.0556  0.0053
##
## sigma^2 estimated as 3.753e-06:  log likelihood=1143.54
## AIC=-2279.09   AICc=-2278.91   BIC=-2265.22
##
## Training set error measures:
##                        ME        RMSE         MAE        MPE     MAPE
## Training set 4.593964e-05 0.001925084 0.001307974 -0.2324159 3.502333
##                    MASE        ACF1
## Training set 0.9474475 -0.03979978
```

La fonction `auto.arima` identifie le meilleur modèle pour un critère (`ic`) donné en cherchant tous les modèles possibles. Pour les 2 critères on retrouve les mêmes résultats qu'en utilisant notre propre approche "force brute".

Pour le meilleur modèle selon l'AIC, la qualité d'estimation est plutôt bonne car elle donne une Root Mean Square Error (RMSE) de 0.002, soit environ $0.002/0.045 \approx 4\%$ de la moyenne des yields, de même pour la Mean Absolute Error (MAE) qui est de 0.001, soit environ $0.001/0.045 \approx 2\%$. Avec BIC, les RMSE et le MAE sont du même ordre de grandeur donc on en conclut que la qualité d'estimation est de même plutôt bonne.

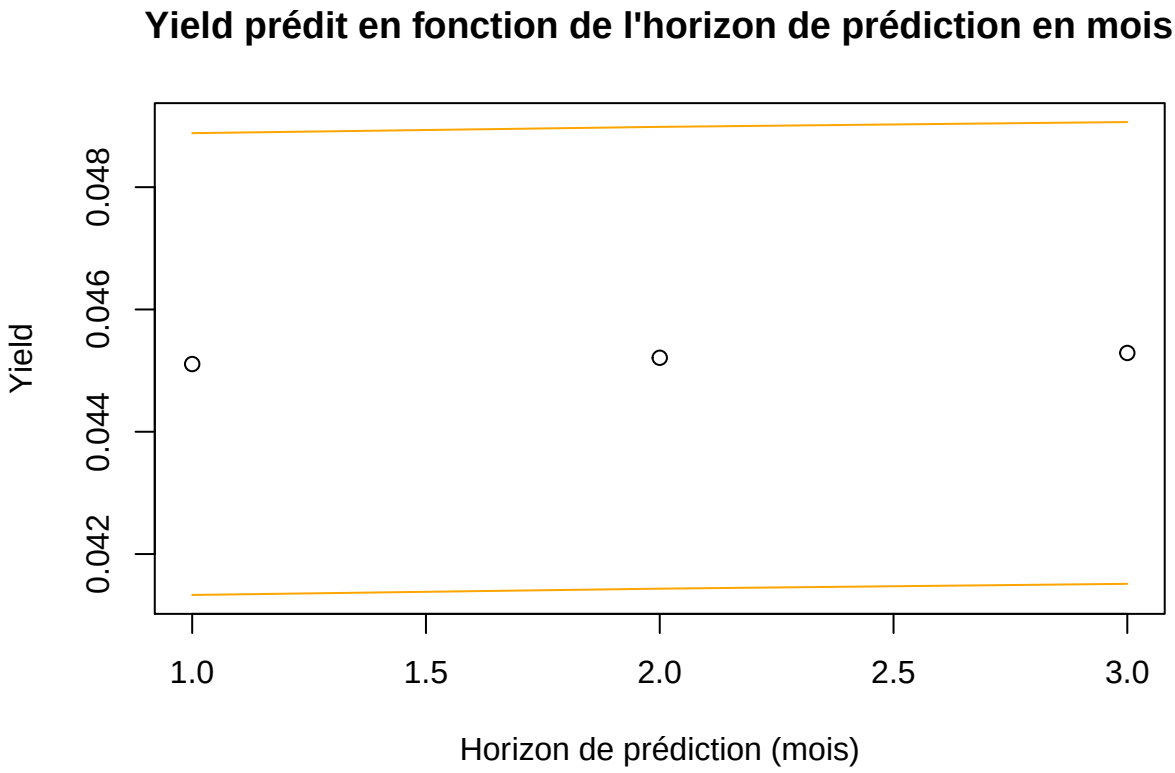
Question 10

Effectuer une prévision à l'aide de la commande predict sur horizon de trois périodes. Donner l'intervalle de confiance à 95%.

```r
sigma2 = arima_data$sigma2

prediction = predict(arima_data, n.ahead = 3)

prediction_upper = prediction$pred + 1.96 * sqrt(sigma2)  # intervalle de confiance à 95%
prediction_lower = prediction$pred - 1.96 * sqrt(sigma2)
```

On présente ci-dessous les prédictions ainsi que l'intervalle de confiance à 95% représenté par les droites en orange.

```r
plot(1:3, prediction$pred[1:3], xlab = "Horizon de prédiction (mois)",
    ylab = "Yield", main = "Yield prédit en fonction de l'horizon de prédiction en mois",
    ylim = c(min(prediction_lower), max(prediction_upper)))
lines(1:3, prediction_lower, col = "orange")
lines(1:3, prediction_upper, col = "orange")
```



On donne ci-dessous un tableau récapitulatif de la dernière observation du yield ainsi que les 3 prochaines prédictions.

Table 1: Dernier yield observé et 3 prochaines prédictions utilisant un modèle ARMA(2,2).

	Statut	Valeur	Intervalle de confiance
01/09/2018	Observation	0.044939	NA
01/10/2018	Prédiction	0.045107	[0.04133106, 0.04888296]
01/11/2018	Prédiction	0.045210	[0.04143399, 0.04898590]
01/12/2018	Prédiction	0.045289	[0.04151282, 0.04906472]

Question 11

Comparer les deux derniers modèles estimés en utilisant ces deux critères (MAE et RMSE). Quel est celui qui affiche la meilleure performance ?

MAE

```
MAE_fit2 = mean(abs(fit2$residuals))
MAE_fit2
```

```
## [1] 0.008861187
```

On trouve une MAE pour la régression linéaire (par rapport aux taux d'intérêt réels) de 0.008861187, elle est donc supérieure au 0.00127265 pour la MAE du modèle ARMA(2,2). Au sens de la MAE, c'est donc le modèle ARMA qui est meilleur.

RMSE

```
RMSE_fit2 = sqrt(mean(fit2$residuals^2))
RMSE_fit2
```

```
## [1] 0.01053751
```

La RMSE pour la régression linéaire est de 0.01053751 alors que le modèle ARMA(2,2) donne une RMSE de 0.001906076. Le modèle ARMA est donc meilleur que la régression linéaire au sens de la RMSE.

Ainsi, que ce soit au sens de la RMSE ou de la MAE, le modèle ARMA(2,2) semble être un meilleur modèle que la régression linéaire.

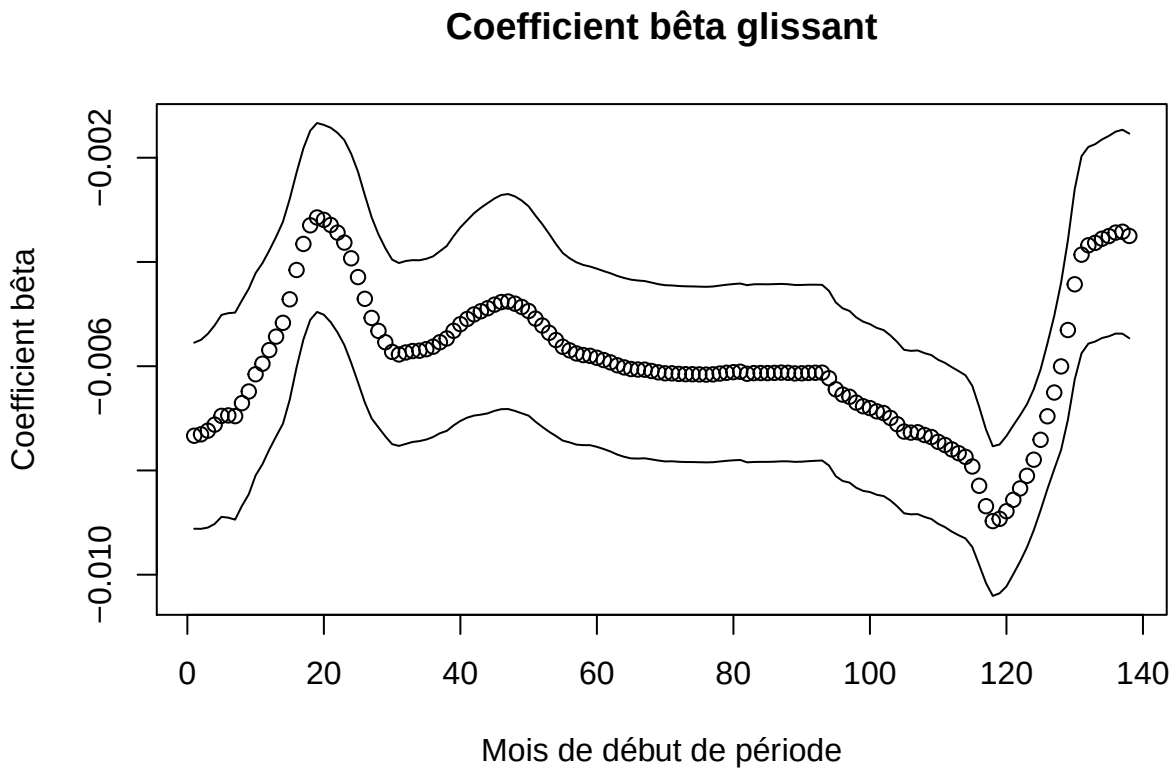
Question 12

A l'aide d'une routine que vous développez sous R, estimer les coefficients conformément à la méthode glissante. Tracer les courbes des coefficients β_i estimés ainsi que leur intervalle de confiance au seuil de 95%. Commenter.

On prend une période de 100 mois sur laquelle on estime les β_i : β_i est donc le coefficient de r_t dans l'équation de régression $\frac{E_t}{P_t} = \alpha + \beta r_t$ résolue à l'aide uniquement des données disponibles entre le mois i et le mois $i + 99$. β_1 porte donc sur les données entre le mois 1 et le mois 100; β_2 sur celles entre le mois 2 et le mois 101 etc.

```
period_move = 1
period_length = 100

nperiods = (length(data$price) - period_length)%/%period_move

beta = c()
lower = c()
upper = c()

for (i in 0:nperiods) {

    yields = data$yield[(i * period_move + 1):(i * period_move + period_length)]
    real_rates = data$real_rates[(i * period_move + 1):(i * period_move +
        period_length)]
    fit_i = lm(yields ~ real_rates)
    confidence = confint(fit_i)

    beta = c(beta, fit_i$coefficients[2])
```

```
    lower = c(lower, confidence[2, 1])
    upper = c(upper, confidence[2, 2])
}

plot(0:nperiods + 1, beta, ylim = c(min(lower), max(upper)), xlab = "Mois de début de période",
    ylab = "Coefficient bêta", main = "Coefficient bêta glissant")
lines(upper)
lines(lower)
```



Globalement, le coefficient β reste négatif et oscille autour d'une valeur moyenne à -0.006, ce qui est très proche de la valeur trouvée (-0.0062351) à la question 6 en utilisant l'ensemble des données. On remarque toutefois que $widehat\beta$ baisse notablement en valeur autour du 120e mois, c'est-à-dire si l'on essaie de régresser le yield par rapport aux taux sur des données postérieures à fin 2007. Cela peut s'expliquer par le fait que la crise de 2008 a notablement dégradé les taux qui sont devenus négatifs et moins volatils, d'où un dénominateur plus faible dans l'expression

$$\widehat{\beta} = \frac{Cov(\frac{E}{P}, r)}{\sigma_r^2}$$

ce qui pourrait expliquer pourquoi β a augmenté en valeur absolue.

Question 13

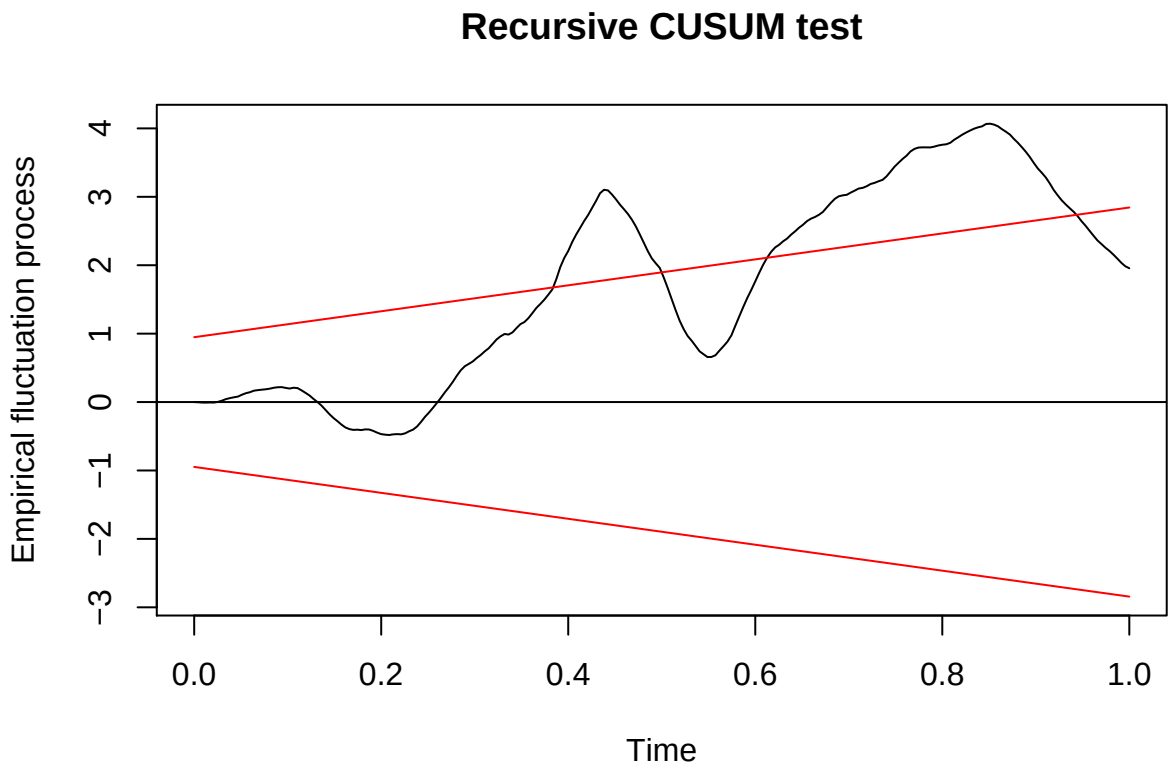
Présenter le test CUSUM. Implémenter le test de CUSUM et commenter vos résultats.

Le test CUSUM est un test fondé sur la somme cumulée (**cu**mulative **sum** en anglais) des résidus récursifs : au fur et à mesure que le modèle "grossit" avec un nombre croissant de données pour estimer le coefficient β, on en déduit les résidus récursifs qui représentent l'erreur successive entre les données observées et les données modélisées.

La somme partielle est la statistique de test; elle permet de détecter tout changement structurel dans l'estimation : lorsque la somme sort d'un certain intervalle de stabilité, on décrète qu'il y a eu un changement significatif dans l'estimation, et qu'il y a donc instabilité du modèle.

Lorsque l'on effectue un test CUSUM, l'hypothèse nulle est la constance des coefficients estimés par le modèle. Sous cette hypothèse nulle, il y a instabilité du modèle dès lors que la statistique de test sort de l'intervalle de stabilité.

```
cusum_data = efp(data$yield ~ data$real_rates)
plot(cusum_data)
```




```
sctest = sctest(data$yield ~ data$real_rates)
sctest
```

```
##
##  Recursive CUSUM test
##
## data:  data$yield ~ data$real_rates
## S = 1.6536, p-value = 3.449e-05
```

On voit que la statistique de test n'est pas entièrement contenue dans l'intervalle de stabilité délimité par les droites en rouge. **Il y a donc instabilité du modèle**, ce qui est confirmé par la fonction `sctest` qui donne une p-valeur très basse ($3.449e-05$), d'où un rejet de l'hypothèse nulle : le coefficient β estimé n'est en fait pas constant sur toute la durée d'observation.