TP Séries chronologiques

Laurent LIN & Guillaume SHI

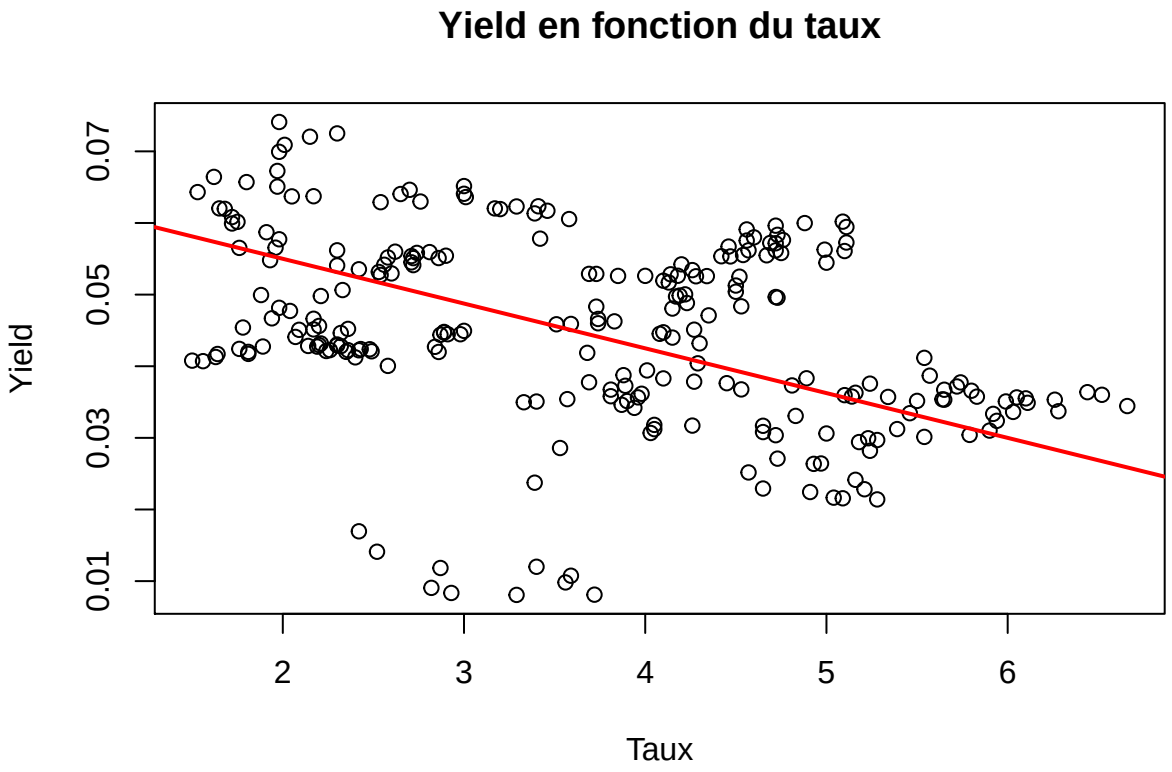
24/02/2019

Question 1

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("/Users/guillaumeshi/Desktop/OMA/SCH/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

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. L'hypothèse sous-jacente est donc que $\mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{0}$ et $\mathbb{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T) = \sigma I_n$ avec σ constante.

La méthode des moindres carrés ordinaires (MCO) consiste à trouver le vecteur $\boldsymbol{\gamma}$ par minimisation de la quantité suivante : $\Delta = (\mathbf{y} - \mathbf{X}\boldsymbol{\gamma})^2$. 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

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

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

  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

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

```
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:
```

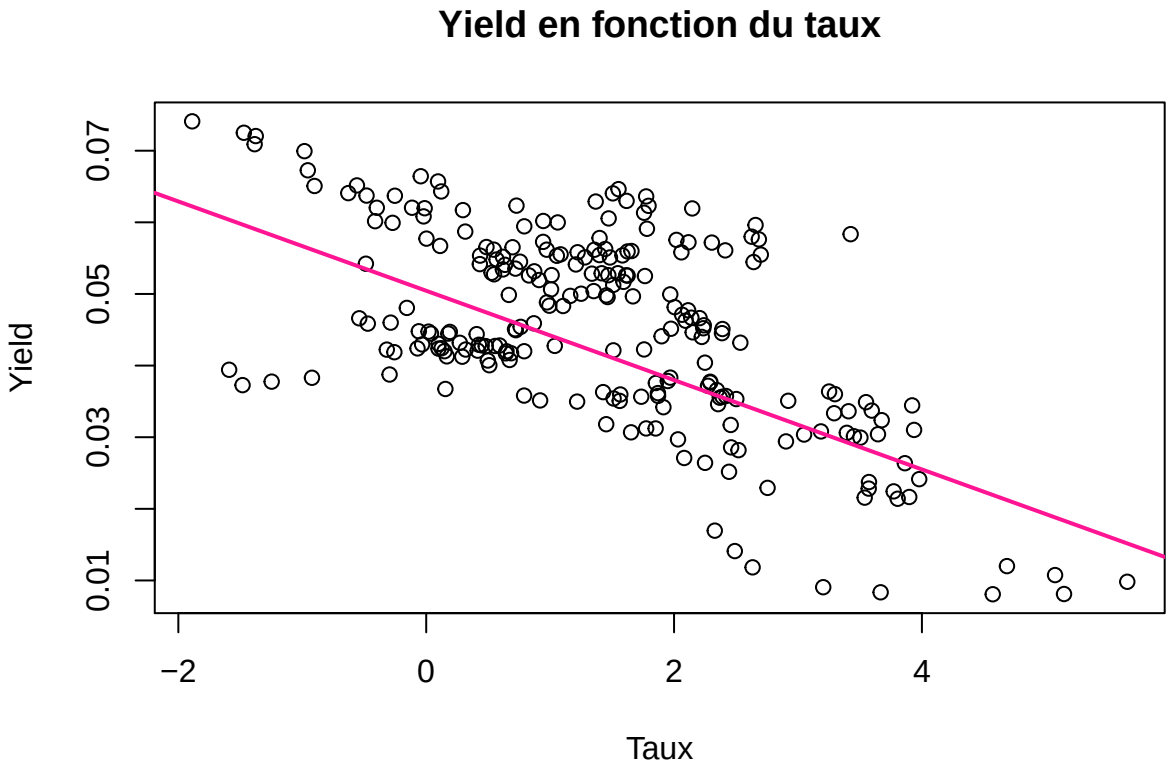
```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0586936  0.0023539  24.935  < 2e-16 ***
## data$rates  -0.0037816  0.0006005  -6.297 1.38e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01244 on 247 degrees 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
```

Le $R^2 = 0.14$ n'est pas très bon... La statistique de Student pour l'ordonnée à l'origine et le coefficient directeur donnent des p-values très petites, on a donc significativité à tous les niveaux de confiance usuels. La statistique de Fisher est de 39.66, d'où une p-valeur très basse elle aussi, on a aussi significativité.

```
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)
```



Le R^2 est de 0.39, ce qui est une amélioration nette par rapport à précédemment où il était de 0.14.

Question 7

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 :

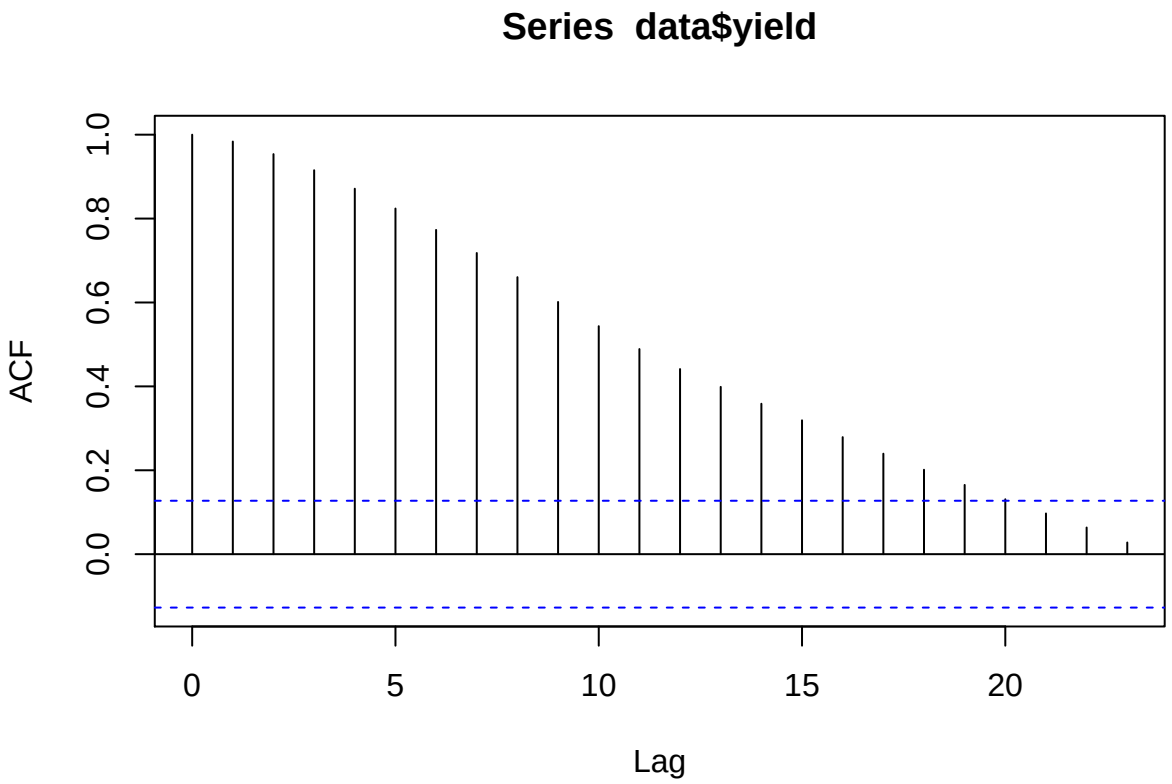
$$X_t = \phi_1 X_{t-1} + ... + \phi_p X_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + ... + \theta_q \epsilon_{t-q}$$

Question 8

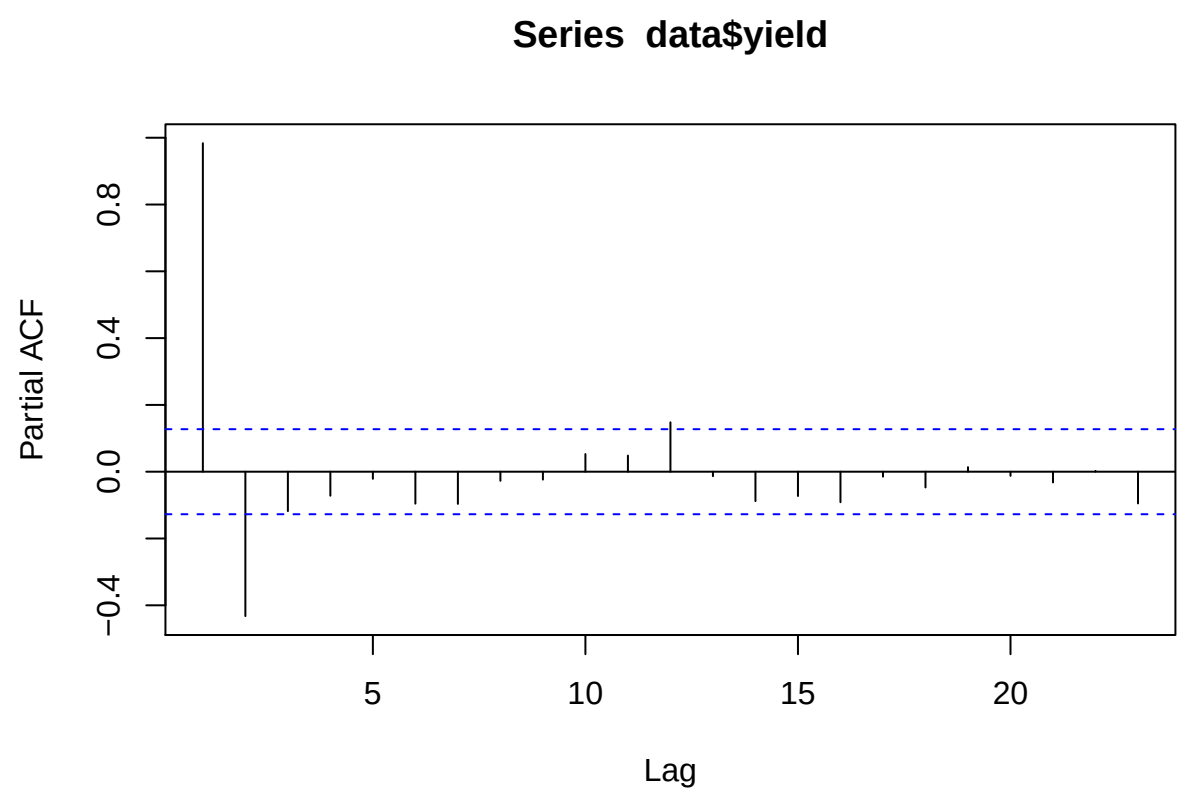
Première approche

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
```




```r
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 voit 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 distinguent, 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..

```
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
```

```
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

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

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

```
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
```

Le modèle suggéré est un ARMA(2, 2), ce qui confirme la conclusion des deux approches de la question 8 concernant l'ordre p. Quant à l'ordre q, il semble qu'en accord avec la deuxième approche, ce soit l'AIC qui ait été privilégié par la fonction `auto.arima`.

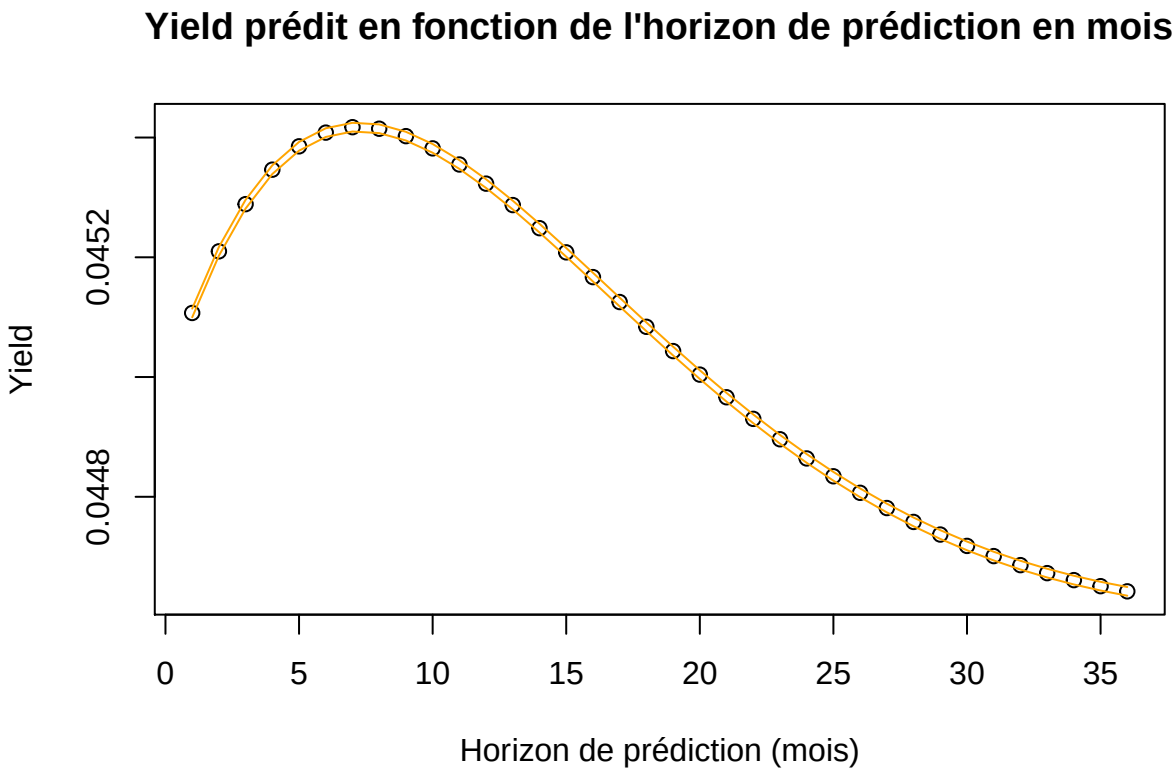
La qualité d'estimation est plutôt bonne car elle donne une Root Mean Square Error (RMSE) de 0.002, soit environ $0.002/0.045 \approx 4\%$ de la moyenne des yields, de même pour la Mean Absolute Error (MAE) qui est de 0.001, soit environ $0.001/0.045 \approx 2\%$.

Question 10

```
sigma2 = arima_data$sigma2

prediction = predict(arima_data, n.ahead = 36)

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

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

```
plot(1:36, prediction$pred[1:36], 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(1:36, prediction_lower, col = "orange")
lines(1:36, prediction_upper, col = "orange")
```



Question 11

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
```

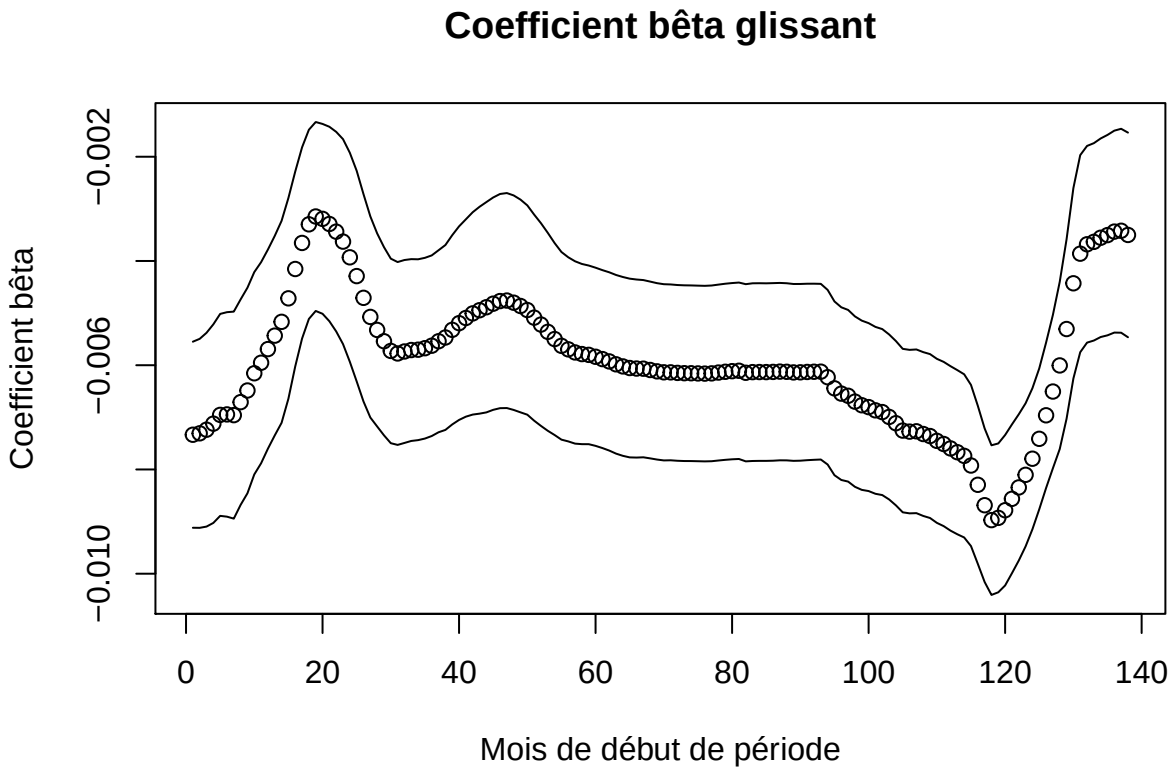
La RMSE pour la régression linéaire est de 0.01053751 alors que le modèle ARMA(2,2) donne une RMSE de 0.001906076. Le modèle ARMA est donc meilleur que la régression linéaire au sens de la RMSE.

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

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. On remarque toutefois que β 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

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

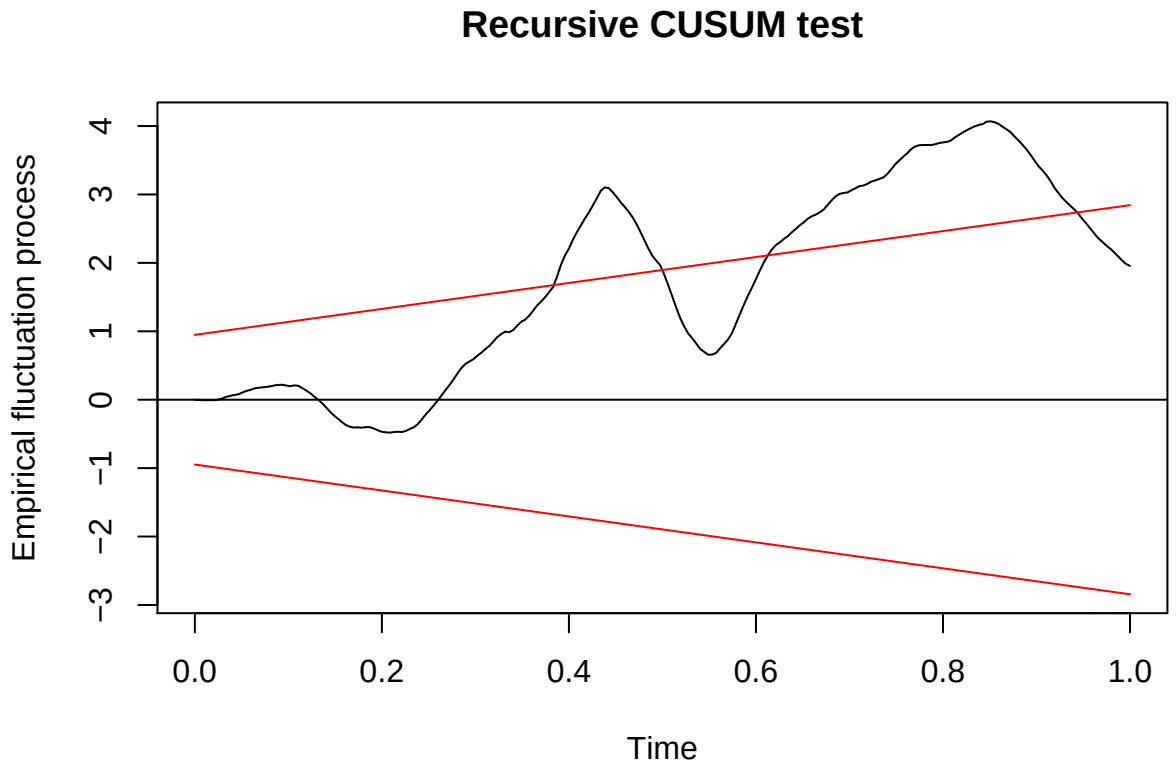
ce qui pourrait expliquer pourquoi β a augmenté en valeur absolue.

Question 13

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, d'où un rejet de l'hypothèse nulle : le coefficient β estimé n'est en fait pas constant sur toute la durée d'observation.