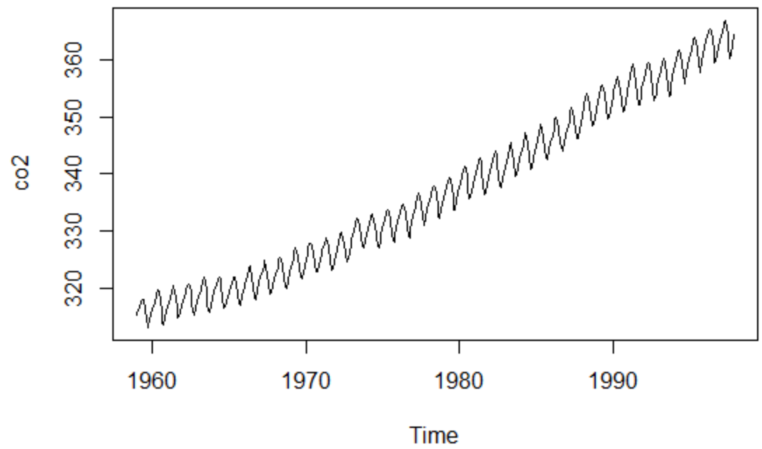
Nous allons étudier le jeu de données nommé *co*2 qui recense les émissions de CO_2 par mois de l'année 1959 à 1997.

1 Etude de la série

On commence par afficher la série afin de la décrire un peu plus et d'en déduire des modèles qui nous permettront de faire la meilleure prévision possible.



En voyant l'allure de la série on peut déjà dire que :

1. elle a une tendance linéaire croissante,
2. elle a une saisonnalité annuelle, ie de 12 mois,
3. elle n'est pas hétéroscédastique puisque la variance est constante dans le temps.

Puisque la série ne présente pas d'hétéroscadasticité nous n'avons pas besoin de la passer au log.

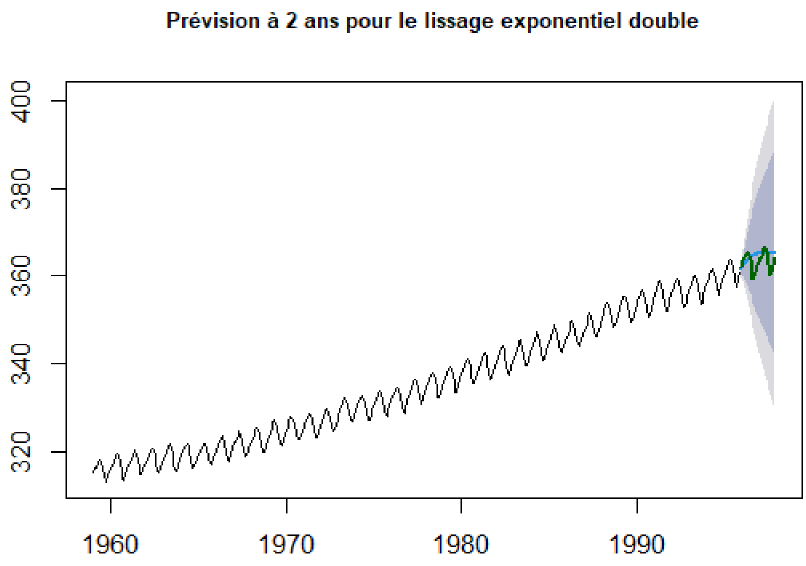
2 Modélisation par lissage exponentiel

Dans cette partie nous allons modéliser la série par un lissage exponentiel double, puis par un lissage de Holt-Winter pour un modèle additif et un modèle multiplicatif. Pour cela, nous allons séparer la série en deux parties :

1. Une première partie pour la modélisation (de 1959 à 1995)
2. Une deuxième partie pour tester notre modèle (de 1996 à 1997)

2.1 Lissage exponentiel double

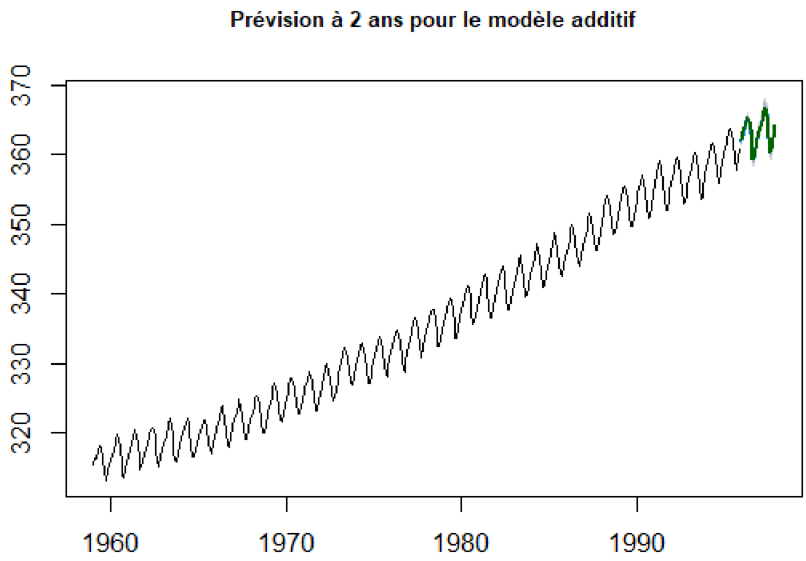
Pour ce modèle là on obtient une prévision à 2 ans comme ci-dessous :



On voit que la courbe bleue, qui correspond à la prévision, ne donne pas du tout une bonne prévision puisque qu'elle ne suit pas la courbe verte, qui correspond à ce qu'il s'est passé en 1996 et 1997. On peut déjà exclure la modélisation par un lissage exponentiel double.

2.2 Lissage de Holt-Winter pour un modèle additif

Dans ce cas là on obtient le graphique ci-dessous :



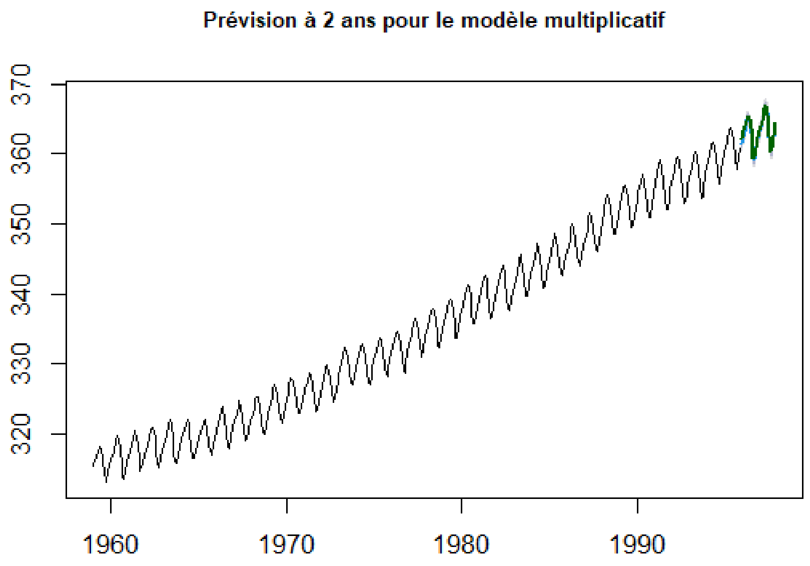
Cette fois-ci on voit que la prévision (tracée en bleue) suit parfaitement la courbe verte.
Pour évaluer la performance de la prévision on va utiliser 3 critères qui sont le RMSE, le MAPE, et l'AIC.
On obtient :

```
> print(RMSE_AAA)
[1] 0.352872
> print(MAPE_AAA)
[1] 0.0007323709
> AIC(fit_LHW_AAA)
[1] 1628.01
```

On utilisera ces valeurs pour les comparer à celles que l'on obtiendra dans la sous-partie suivante.

2.3 Lissage de Holt-Winter pour un modèle multiplicatif

Dans ce cas là on obtient le graphique ci-dessous :



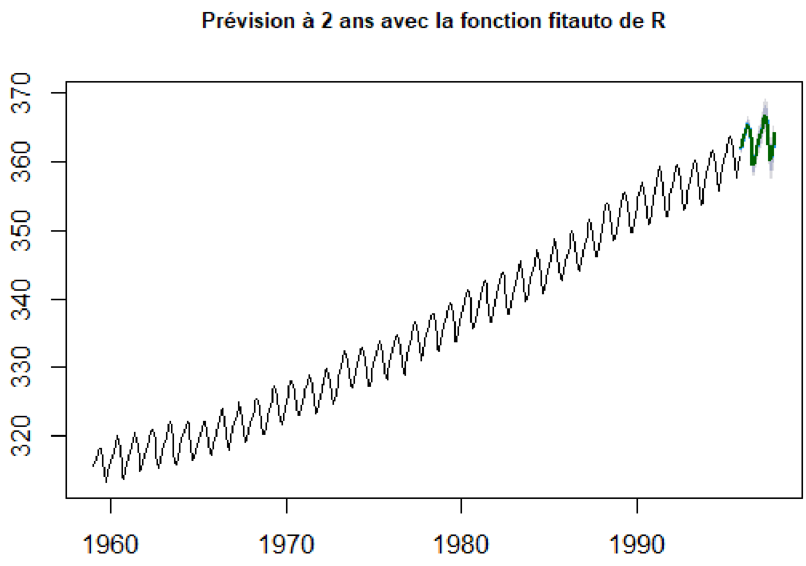
On voit que la prévision suit, là aussi, bien la courbe verte. pour déterminer quel modèle (additif ou multiplicatif) est le meilleur, on va calculer le RMSE, le MAPE et l'AIC.

```
> print(RMSE_MMM)
[1] 0.4490259
> print(MAPE_MMM)
[1] 0.00098458
> AIC(fit_LHW_MMM)
[1] 2181.843
```

On remarque que le RMSE, le MAPE et l'AIC sont supérieurs à ceux obtenus pour le modèle additif. On en conclut donc que le modèle additif donne une meilleure prévision que le modèle multiplicatif.

2.4 Prévision à l'aide de la fonction fitauto de R

Avec cette fonction on obtient la prévision suivante, en bleue, qui suit bien la courbe verte :



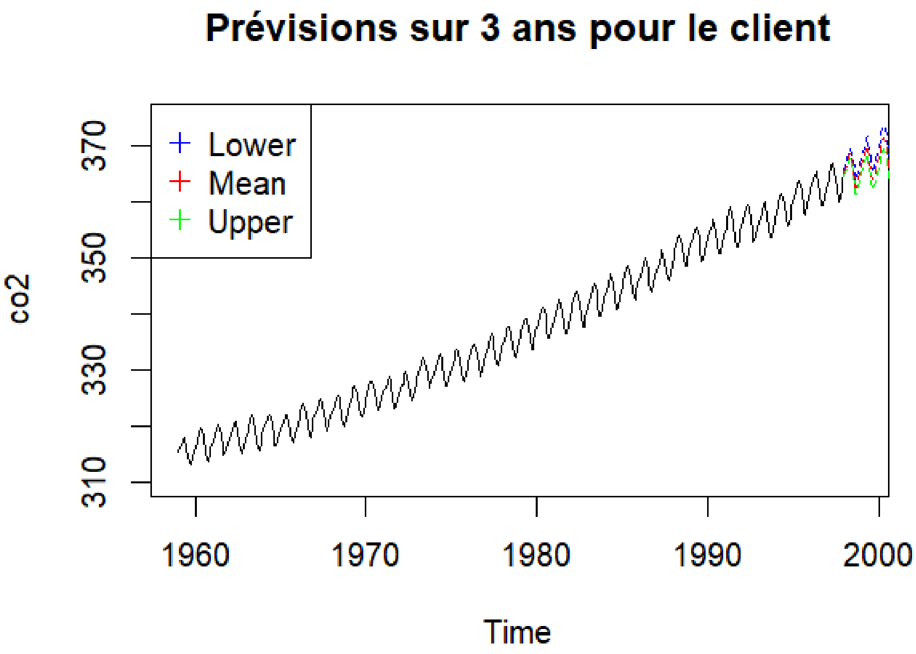
Pour voir si ce modèle nous donne une bonne prévision nous allons regarder le RMSE, le MAPE et l'AIC :

```
> RMSE.predauto
[1] 0.3697787
> MAPE.predauto
[1] 0.0007307653
> AIC(fitauto)
[1] 1600.527
```

Ici, les RMSE et le MAPE obtenus sont supérieurs à ceux que l'on avait obtenus pour le modèle additif, par contre l'AIC est meilleur pour cette prévision. Nous faisons le choix de prendre le modèle avec le meilleur RMSE et le meilleur MAPE puisque nous sommes uniquement intéressés par la prévision.

2.5 Prévision finale sur 3 ans, pour le client

On choisit le meilleur modèle que nous avons obtenu c'est à dire le lissage de Holt-Winter pour un modèle additif. On refait une modélisation sur l'ensemble de notre jeux de données (partie modélisation et partie test) et on va prédire pour les 3 prochaines années (de 1998 à 2001). On obtient les résultats suivants :



3 Modélisation par régression linéaire

Dans cette partie, nous allons modéliser la série par une régression linéaire. Pour cela, nous allons séparer la série en deux parties :

1. Une première partie pour la modélisation (de 1959 à 1995)
2. Une deuxième partie pour tester notre modèle (de 1996 à 1997)

Le modèle de régression linéaire s'écrit de la façon suivante : $\forall$t $\in (1, .., T)$, Xt $= \sum_{i=1}^{n} \alpha_i T_i^t + \sum_{j=1}^{p} \beta_j S_t^j + \epsilon_t$ avec

- $T_t = at + b$ (tendance linéaire)
- $S_t = \sum_{j=1}^{12} \left[\alpha_j cos\left(\frac{2\pi jt}{12}\right) + \beta_j sin\left(\frac{2\pi jt}{12}\right)\right] = \sum_{j=1}^{6} \alpha_j cos\left(\frac{2\pi jt}{12}\right) + \sum_{j=1}^{5} \beta_j sin\left(\frac{2\pi jt}{12}\right)$ car $sin(\pi t) = 0$ (saisonnalité)

Nous avons ainsi le système matriciel suivant :

$$\begin{pmatrix} X_1 \\ . \\ . \\ . \\ . \\ . \\ . \\ X_T \end{pmatrix} = \begin{pmatrix} 1 & 1 & \cos\left(\frac{2\pi}{12}\right) & \cos\left(\frac{4\pi}{12}\right) & \dots & \cos(\pi) & \sin\left(\frac{2\pi}{12}\right) & \dots & \sin\left(\frac{10\pi}{12}\right) \\ . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . \\ 1 & T & \cos\left(\frac{2\pi T}{12}\right) & \cos\left(\frac{4\pi T}{12}\right) & \dots & \cos(\pi T) & \sin\left(\frac{2\pi T}{12}\right) & \dots & \sin\left(\frac{10\pi T}{12}\right) \end{pmatrix} \begin{pmatrix} a \\ b \\ \alpha_1 \\ \vdots \\ \alpha_6 \\ \beta_1 \\ \vdots \\ \beta_5 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ . \\ . \\ . \\ . \\ . \\ . \\ \epsilon_T \end{pmatrix}$$

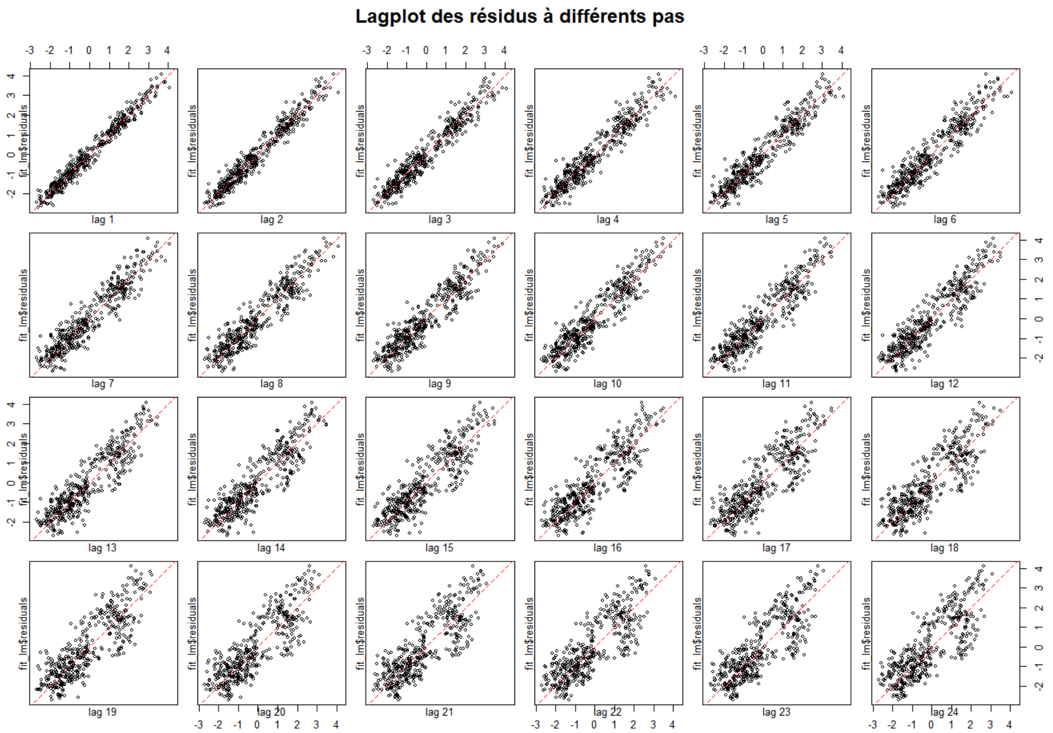
L'objectif est d'estimer les paramètres α_1, α_2, ... , α_6 et β_1, β_2, ..., β_5 en minimisant $\sum_{t=1}^{T} \epsilon_t^2$

3.1 Régression classique

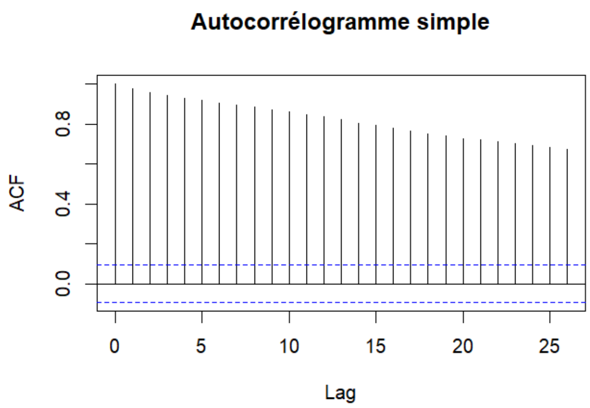
On fait une régression classique, en faisant notre matrice régresseur définit précédemment et on obtient :

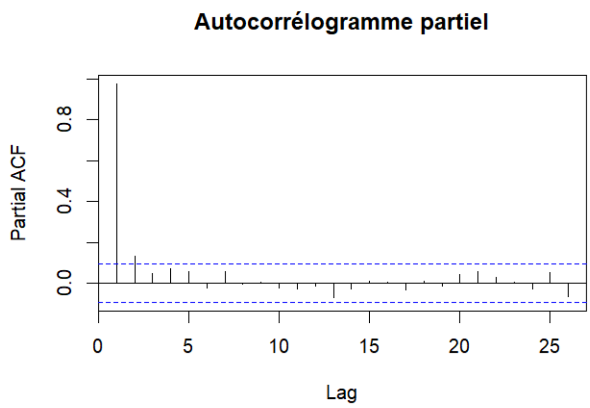
```
> AIC(fit_lm)
[1] 1691.748
```

On effectue un lagplot pour vérifier que les ϵ_t sont des bruits blancs.



Les résidus forment une droite ce qui indique que les résidus ne sont pas des bruits blancs. Le modèle n'est pas valide. On trace les autocorrélogrammes simple et partiel des résidus.





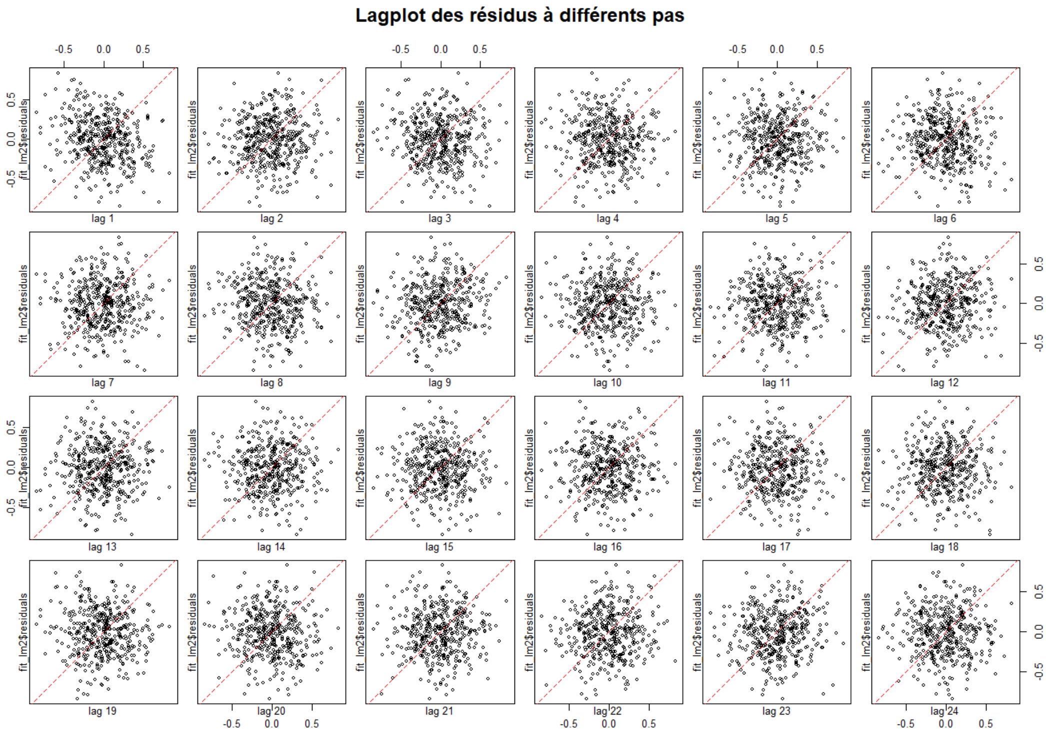
On observe encore une fois que les résidus ne sont pas des bruits blancs puisque les pics ne sont pas tous situés entre les deux droites bleues. De plus, on remarque une décroissance sur l'ACF ainsi que le PACF vaut 0 pour $h > p = 1$. Les résidus sont donc des AR(1).

3.2 Régression linéaire avec les epsilon qui suivent un AR(1)

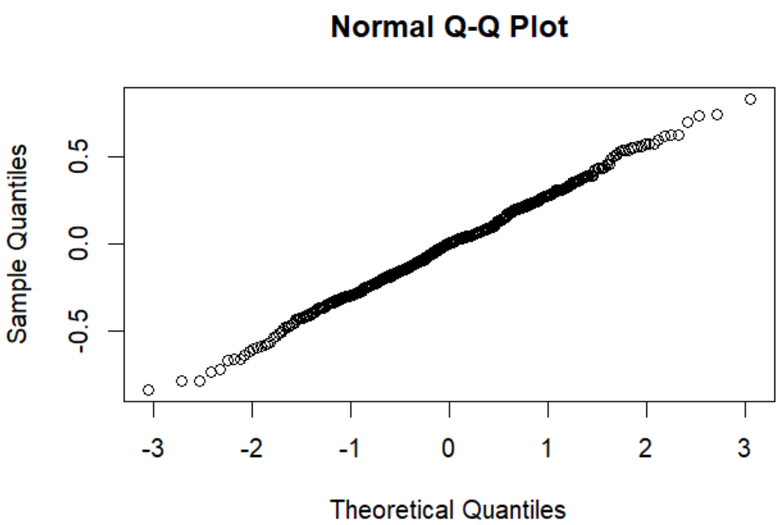
Nous avons maintenant $X_t = T_t + S_t + \epsilon_t$ avec $\epsilon_t = \phi\epsilon_{t-1} + \eta_t$
On obtient :

```
> AIC(fit_lm2)
[1] 191.751
```

Ce modèle est bien meilleur que le précédent puisque l'AIC est bien inférieur à celui qu'on a obtenu précédemment. On effectue un lagplot pour vérifier que les ϵ_t sont des bruits blancs.



Cette fois-ci les résidus ne forment plus une droite ce qui indique que les résidus sont maintenant des bruits blancs.
De plus, on fait un qqnorm :



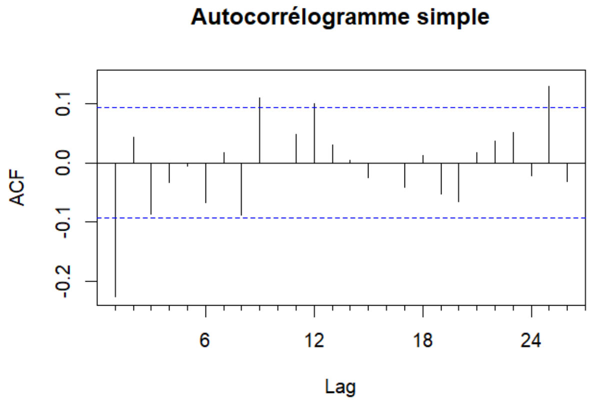
On remarque que les résidus suivent une loi normale ce qui confirme que les résidus sont des bruits blancs.
Par ailleurs, on fait le test de portmanteau :

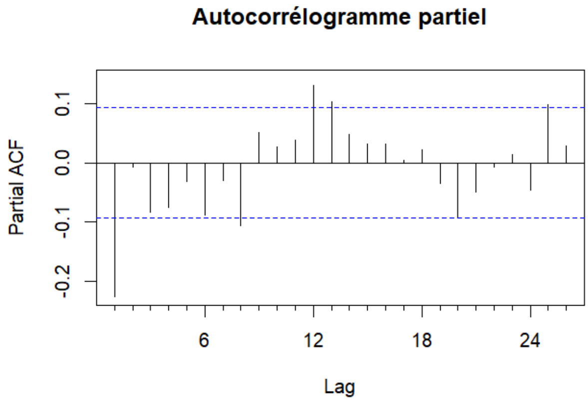
```
> Box.test(fit_lm2$residuals,lag=1,type="Ljung")

        Box-Ljung test

data:  fit_lm2$residuals
X-squared = 22.963, df = 1, p-value = 1.652e-06
```

Avec ce test, on ne peut pas dire que les résidus ne sont pas des bruits blancs puisque la p_value est inférieur à 0.05.
Ce test confirme encore une fois que les résidus sont bien des bruits blancs. Le modèle est donc validé.
Traçons maintenant les autocorrélogrammes simple et partiel des résidus.





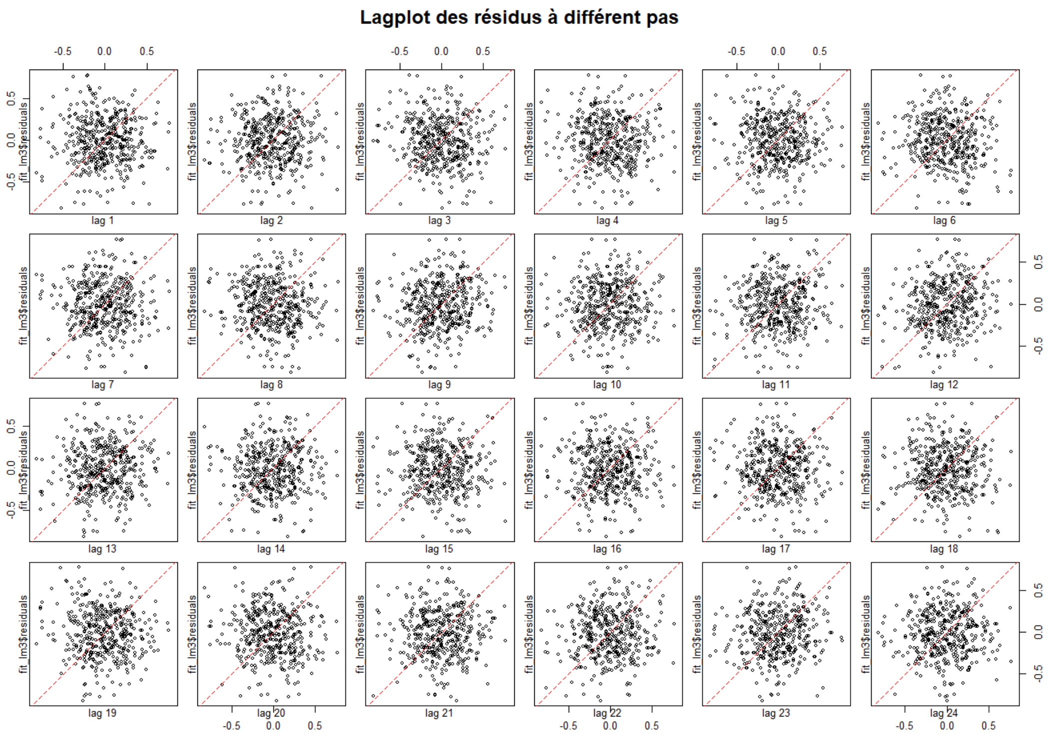
On observe encore un pic en 1, une petite décroissance exponentielle du PACF et on peut aussi dire que l'ACF vaut pour $h > p = 1$ et on décide de faire un ARMA(1,1)

3.3 Régression linéaire avec les epsilons qui suivent un ARMA(1,1)

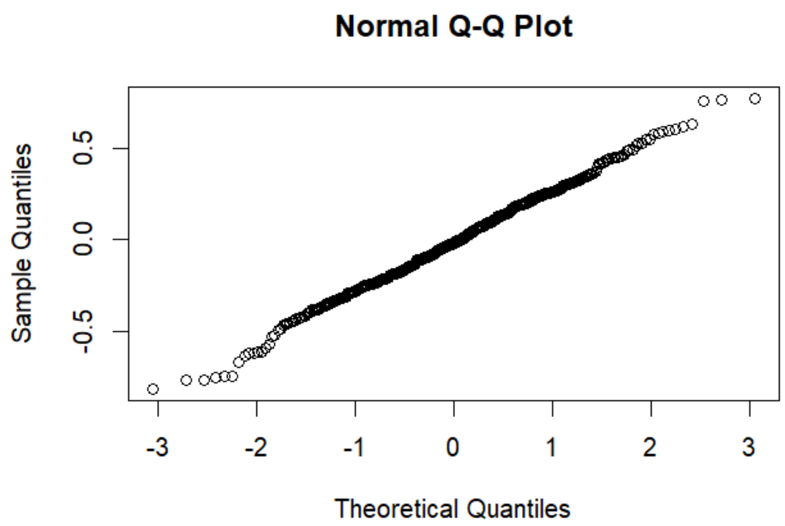
En faisant une régression linéaire avec les résidus modélisés par un ARMA(1,1), on obtient comme performance :

```
> AIC(fit_lm3)
[1] 169.9511
```

Ce modèle est encore meilleur que le précédent.
On effectue un lagplot pour vérifier que les ϵ sont des bruits blancs.

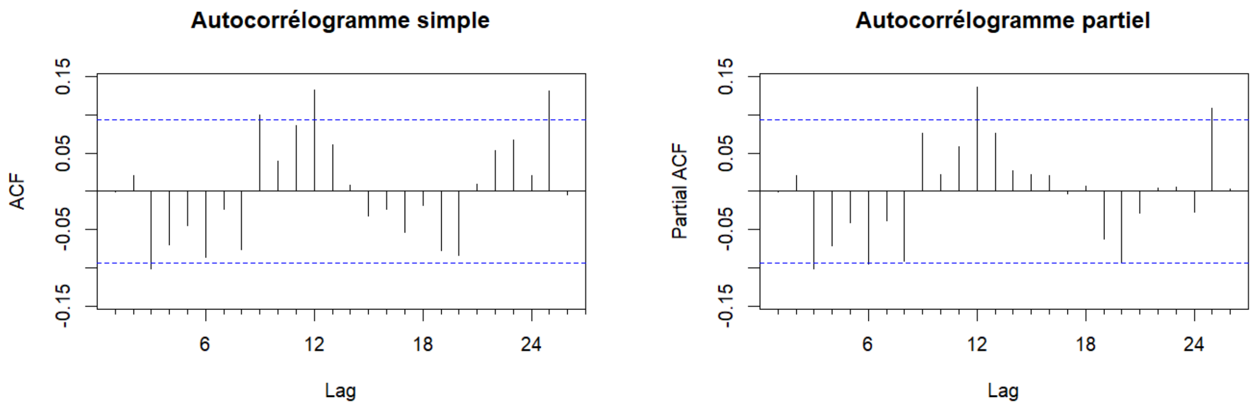


Les résidus ne forment pas une droite ce qui indique que les résidus sont des bruits blancs.
De plus, on fait un qqnorm :



On remarque que les résidus suivent une loi normale. Ainsi, le modèle est validé.

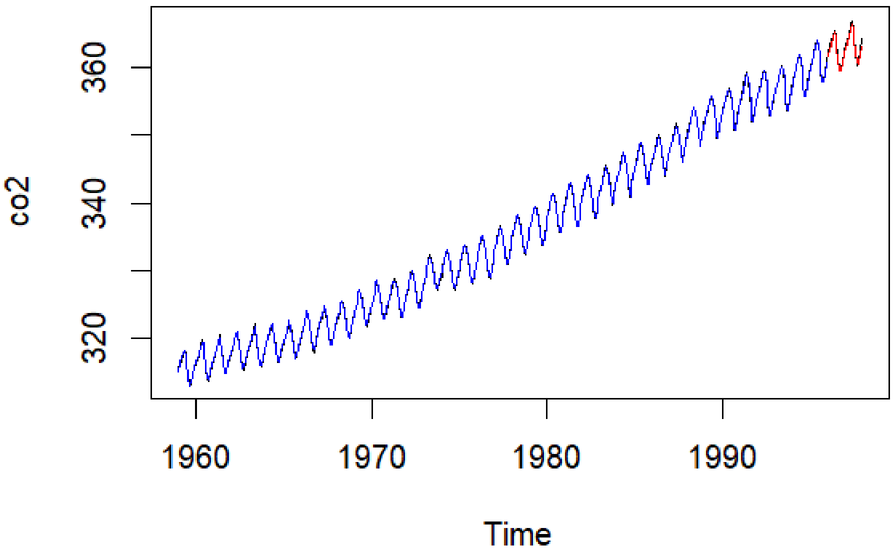
On trace les autocorrélogrammes simple et partiel des résidus.



Tous les pics sont entre les lignes en pointillés bleus, on garde donc ce modèle qui est le plus performant.

3.4 Prévision sur le jeu de données tests

Nous allons utiliser le dernier modèle de régression linéaire avec les résidus qui sont modélisés par un ARMA(1,1). Maintenant que nous avons entraîné notre modèle sur la partie modélisation, on utilise la fonction forecast pour prédire sur le jeu de données test pendant une durée de 2 ans. On obtient les résultats suivants :



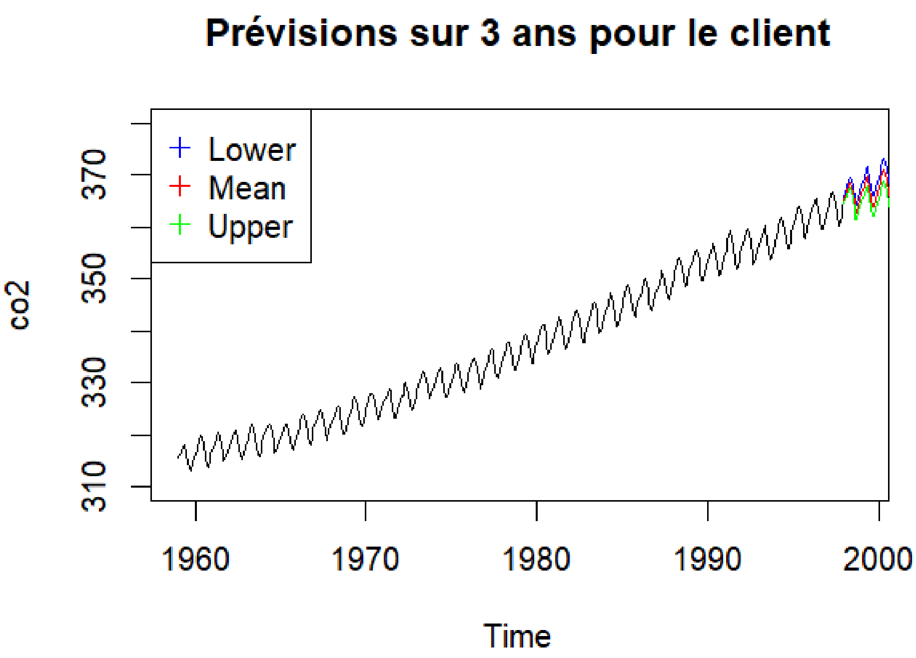
La courbe en noir est la courbe de CO2 initiale. Ensuite, la courbe en bleu correspond à la modélisation de la première partie de la série et la courbe en rouge correspond à la prévision sur 2 ans de notre jeu de données test. Nous pouvons également afficher les erreurs :

```
> print(RMSE.lm)
[1] 0.4925445
> print(MAPE.lm)
[1] 0.001116106
```

On observe qu'au niveau des erreurs, la régression linéaire est légèrement moins performante que pour Holt-Winter pour un modèle additif.

3.5 Prévision finale sur 3 ans, pour le client

On refait une modélisation sur l'ensemble de notre jeu de données (partie modélisation et partie test) et on va prédire pour les 3 prochaines années (de 1998 à 2001). On obtient les résultats suivants :

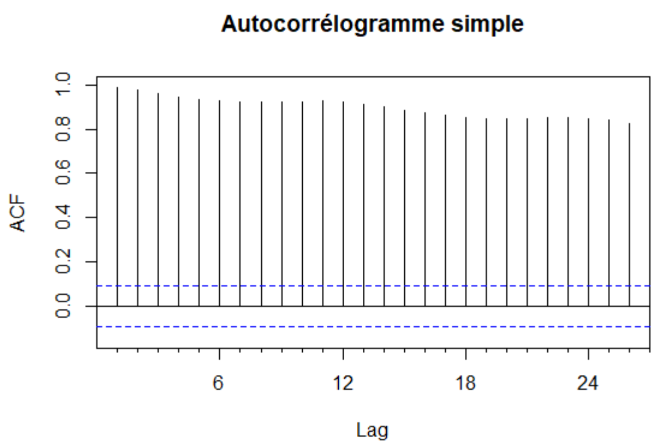


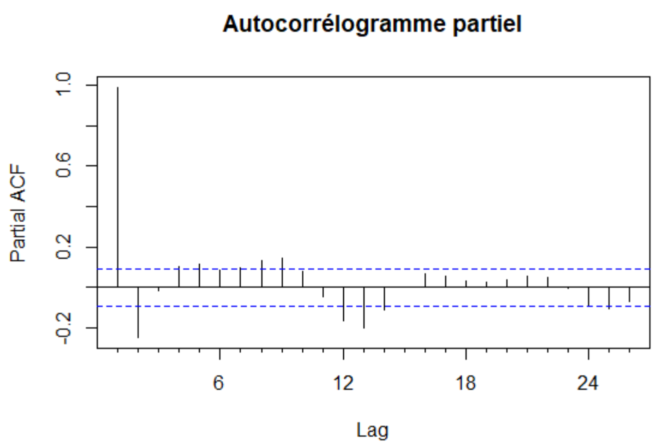
4 Modélisation par SARIMA

Dans cette partie nous allons tester plusieurs modèles afin de trouver le meilleur modèle SARIMA pour notre série.

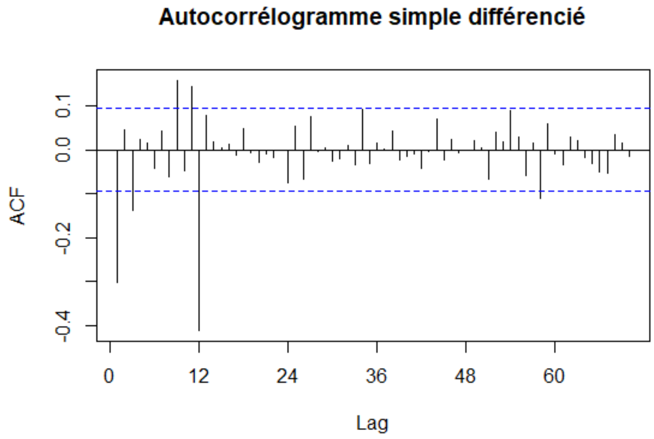
4.1 Etude de la série

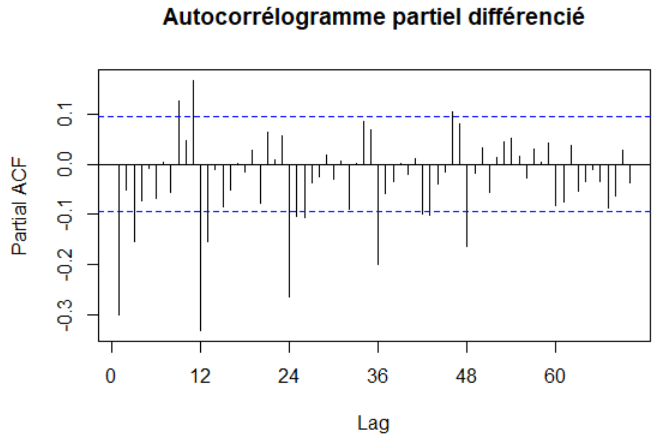
Commençons par tracer les autocorrélogrammes simple et partiel afin de trouver la tendance et la saisonnalité.





Le graphique de l'autocorrélogramme simple montre des pics proches de 1 et proches les uns des autres ce qui veut dire que l'on a une tendance linéaire. Le graphique de l'autocorrélogramme partiel montre une saisonnalité de 12.
On va maintenant supprimer la tendance linéaire et la saisonnalité de 12 en utilisant des opérateurs différenciation.
On obtient les deux graphiques suivants :





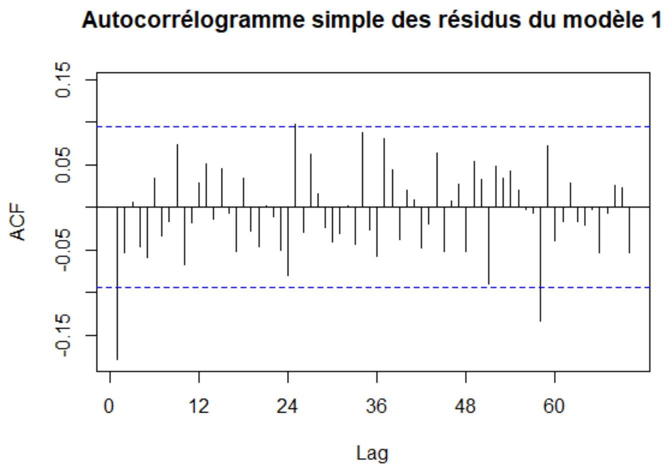
On remarque que l'autocorrélogramme simple s'annule à partir de 12 et que l'autocorrélogramme partiel décroit exponentiellement. Ces caractéristiques sont propres à un MA(12).
Pour la suite, on notera $Y_t = (I - B)\left(I - B^{12}\right) X_t$ la série à laquelle on a enlevé la tendance et la saisonnalité.

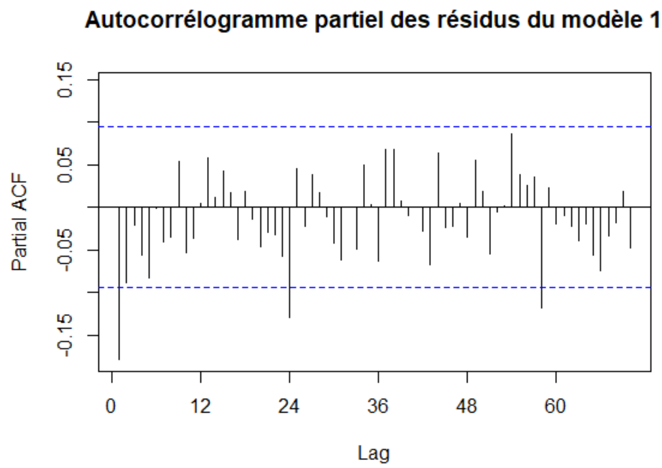
4.2 Modèle 1 : Moyenne Mobile d'ordre 12 non parcimonieux

Le SARIMA s'écrit : $Y_t = \left(I + \theta_1 B^1 + ... + \theta_{12} B^{12}\right) \epsilon_t$
Ce premier modèle nous donne un AIC de :

```
> print(mod1$aic)
[1] 194.8758
```

On va maintenant regarder si les résidus sont des bruits blancs à l'aide de l'autocorrélogramme simple et partiel :





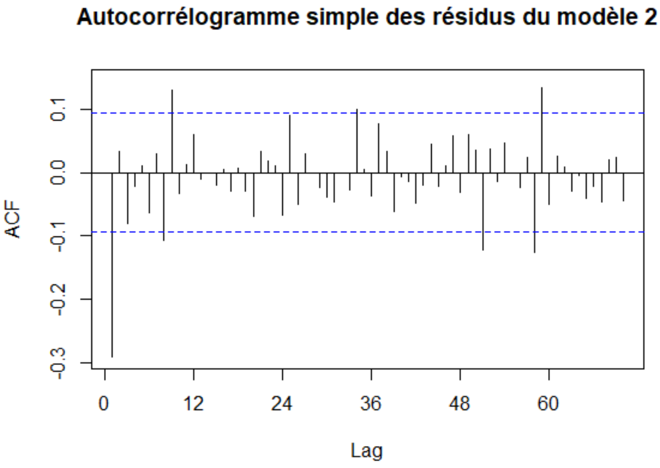
On voit que pour ces deux graphiques, on a des pics qui ne sont pas compris entre les deux droites bleues ce qui montre bien que les résidus ne sont pas des bruits blancs. On en conclut que ce modèle 1 ne convient pas. Donc modéliser la série par un MA(12) n'est pas un bon moyen pour faire des prévisions.

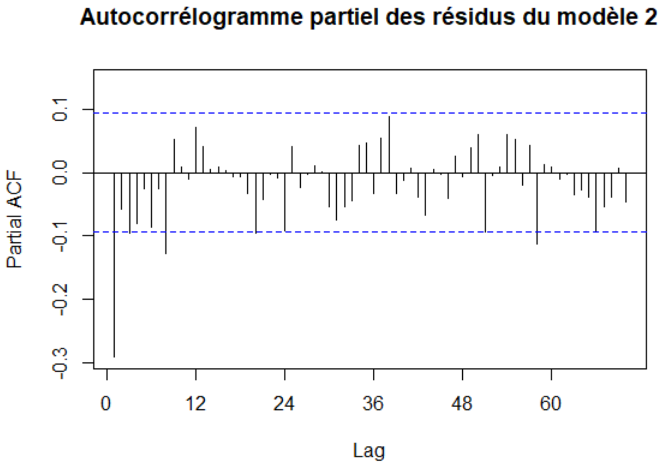
4.3 Modèle 2 : Moyenne Mobile d'ordre 12 parcimonieux

Cette fois-ci le SARIMA s'écrit : $Y_t = \left(I + \theta_{12} B^{12}\right) \epsilon_t$
Ce deuxième modèle nous donne un AIC de :

```
> print(mod2$aic)
[1] 202.2602
```

On va regarder si les résidus sont des bruits blancs à l'aide de l'autocorrélogramme simple et partiel :





On remarque qu'il y a un pic qui dépasse largement les droites bleues en 1 pour l'autocorrélogramme simple et partiel. Ce modèle MA(12) parcimonieux ne convient pas pour faire des prévisions sur notre série co2.

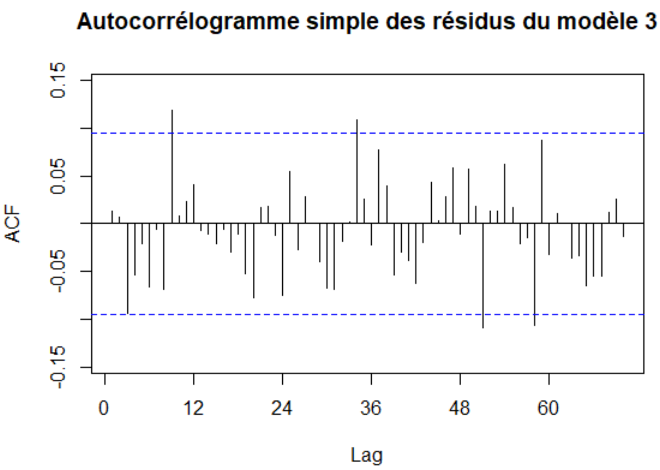
Il faut que l'on prenne en compte le pic présent en 1. C'est que nous allons faire dans le modèle 3. On remarque également une décroissance exponentielle de l'autocorrélogramme partiel et l'autocorrélogramme simple s'annule à partir de $p = 1$ ce qui est caractristique d'un MA(1). On va donc rajouter $p = 1$ dans le prochain modèle.

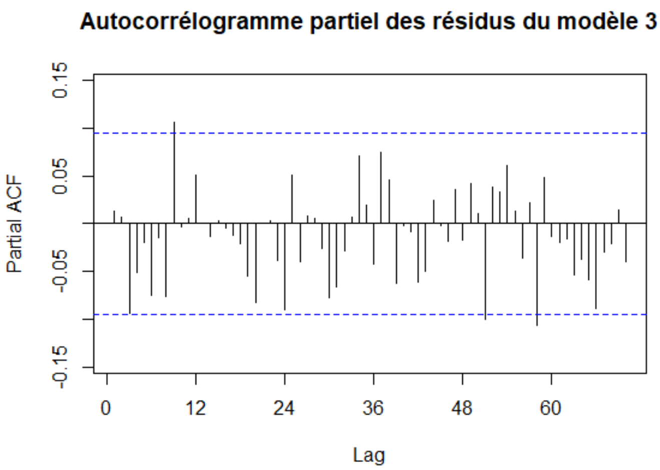
4.4 Modèle 3

Dans ce modèle le SARIMA s'écrit : $Y_t = (I + \theta_1 B)\left(I + \theta_{12} B^{12}\right)\epsilon_t$
Ce deuxième modèle nous donne un AIC de :

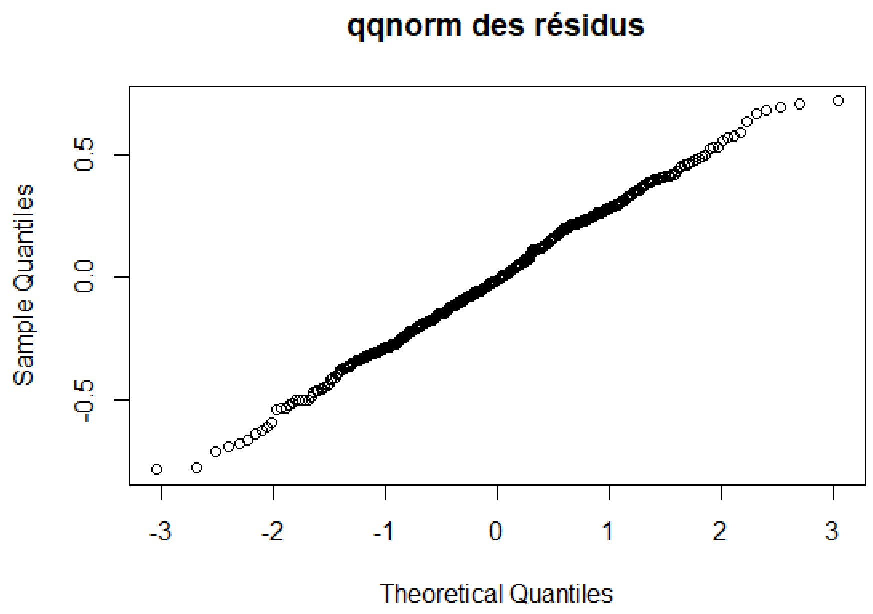
```
> print(mod3$aic)
[1] 159.7686
```

On voit déjà que ce dernier AIC est bien meilleur que ceux obtenus au modèle 1 et au modèle 2. Traçons maintenant l'autocorrélogramme simple et l'autocorrélogramme partiel de ce modèle.

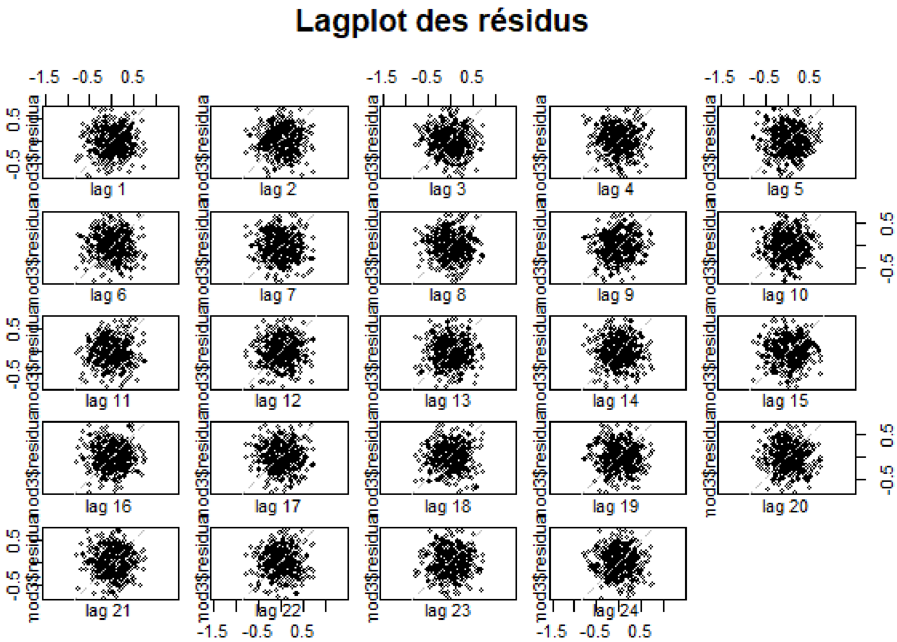




Nous n'avons plus de pic en 1 pour l'autocorrélogramme simple et partiel. Et on peut voir que les pics sont compris entre les droites bleues. On va alors vérifier que les résidus sont bien des bruits blancs en faisant un qqnorm et un lagplot.
Le qqnorm nous donne :



On remarque que les résidus sont bien normaux puisqu'on obtient une droite.
Faisons un lagplot pour confirmer que les résidus sont bien des bruits blancs.



On remarque que l'on n'obtient pas de droite ce qui signifie bien que les résidus sont des bruits blancs. Le modèle 3 est validé.

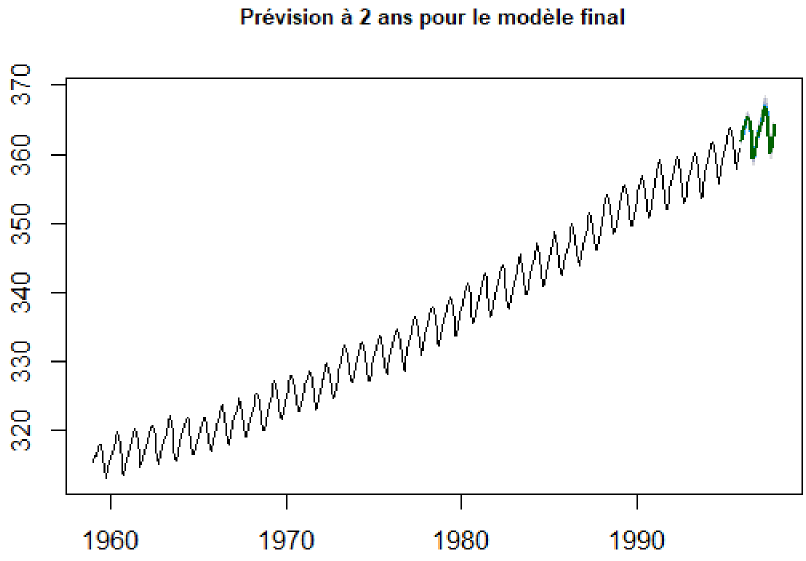
Il reste donc a considérer la tendance linéaire et la saisonnalité de 12 dans le modèle. C'est ce que l'on va faire dans le modèle final.

4.5 Modèle final

Ce modèle s'écrit : $(I - B)\left(I - B^{12}\right) X_t = (I + \theta_1 B)\left(I + \theta_{12} B^{12}\right) \epsilon_t$
Avec ce modèle on obtient un AIC de :

```
> print(mod_final.M$aic)
[1] 159.5804
```

L'AIC obtenu est meilleur que celui du modèle 3.
On va maintenant faire des prévisions sur la série coupée en deux parties pour vérifier l'exactitude de la prévision avant d'appliquer ce modèle final sur la série complète.
On obtient le graphique suivant :



Graphiquement, on voit bien que la prévision suit bien la courbe verte qui représente ce qu'il s'est réellement passé en 1996 et 1997.
On obtient le RMSE et le MAPE suivant :

```
> print(RMSE_final)
[1] 0.3540707
> print(MAPE_final)
[1] 0.00076138
```

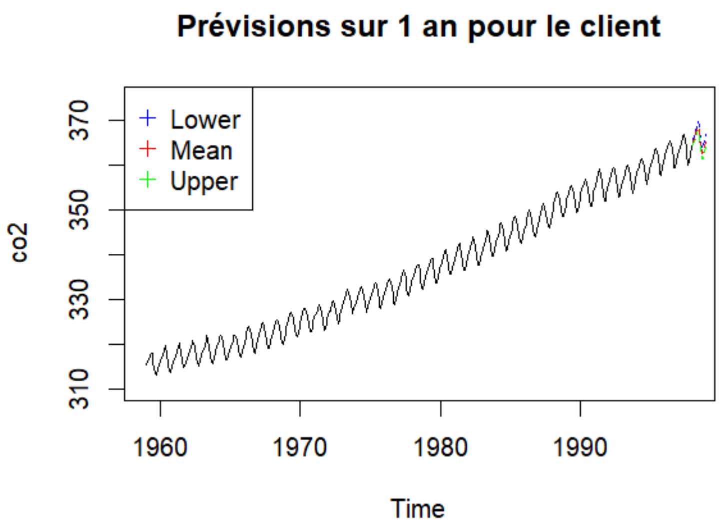
Nous avons de bons résultats pour le RMSE et le MAPE.
On va maintenant s'intéresser au modèle complet dans lequel nous considèrerons la série complète. Nous pourrons alors faire des prévisions sur plusieurs années.

4.6 Modèle complet : pour le client

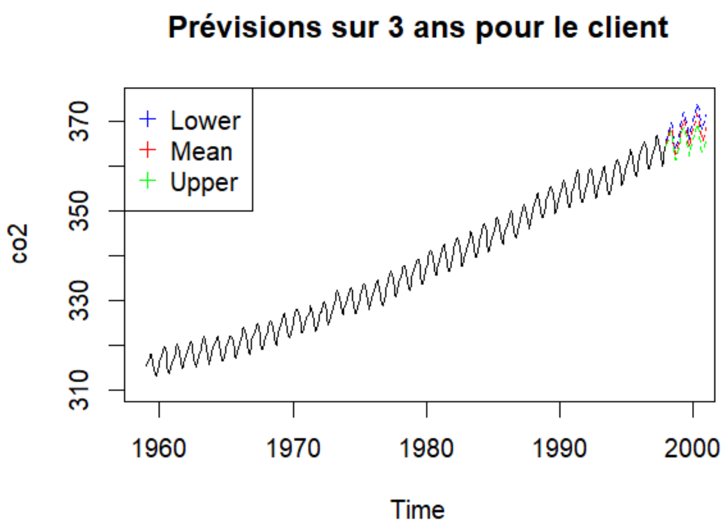
Pour ce modèle, on a un AIC de :

```
> print(mod_final$aic)
[1] 178.1557
```

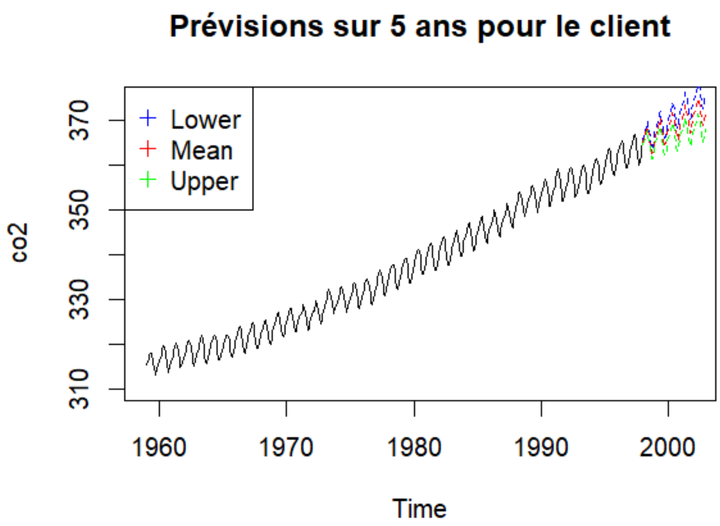
Faisons maintenant des prévisions sur 1 an, 3 ans et 5 ans.
Pour les prévisions à 1 an on obtient le graphique ci-dessous :



Pour les prévisions à 3 ans on obtient le graphique ci-dessous :



Pour les prévisions à 5 ans on obtient le graphique ci-dessous :

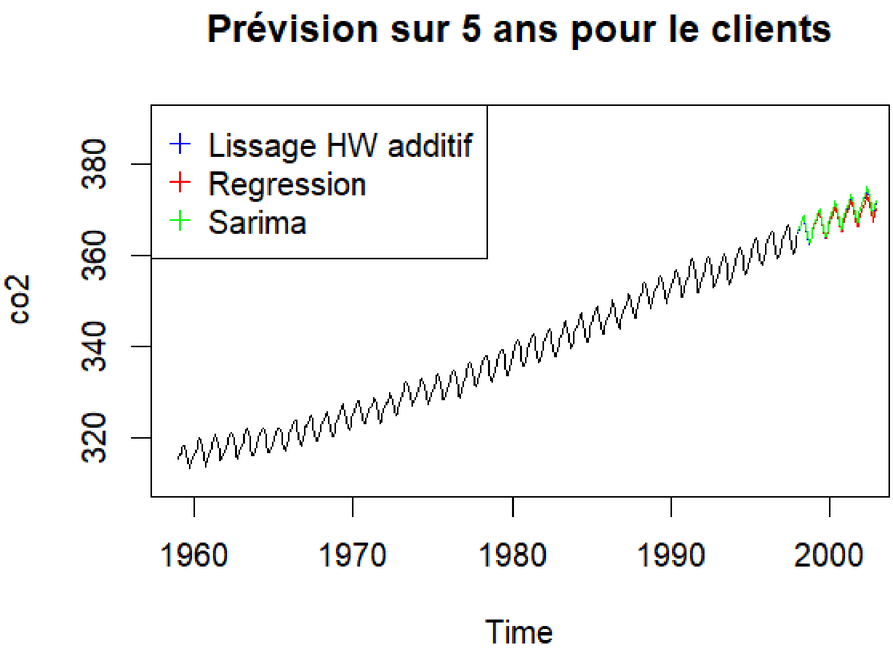


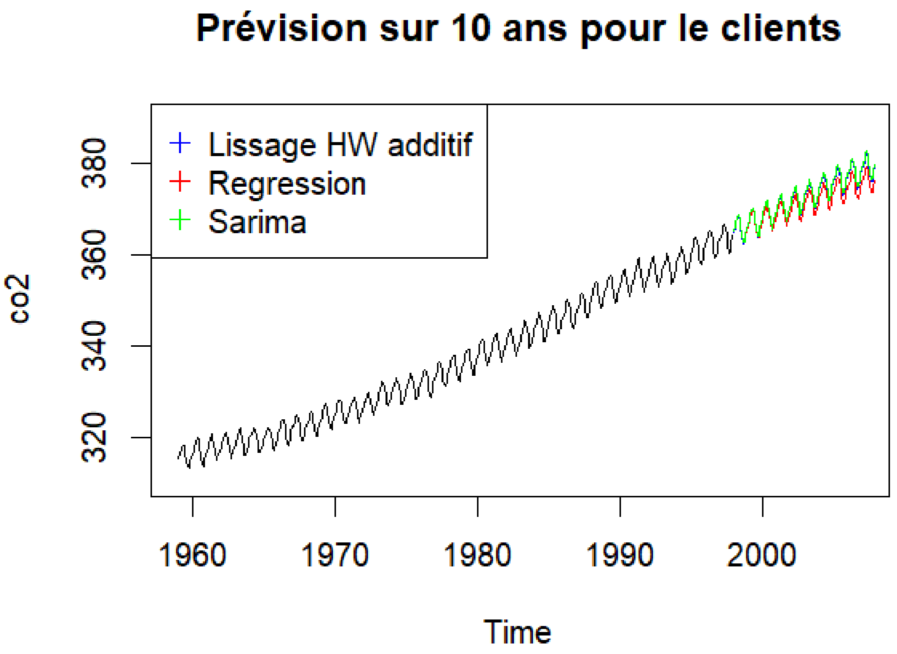
5 Comparaisons des 3 modèles et prévisions sur plusieurs années

Nous avons sélectionné trois modèles :

- Lissage de Holt-Winter pour un modèle additif
- Régression linéaire avec les epsilon modélisés par un ARMA(1,1)
- SARIMA(0,1,1)(0,1,1)12

Nous allons afficher les prévisions pour les différents modèles sur un même graphique pour différentes années :





Il est difficile de distinguer les trois courbes sur les graphiques. En effet, la série temporelle du jeu de données CO2 est très facile à prédire et les trois modèles font des bonnes prédictions. Au vu des calculs des critères AIC ainsi que des erreurs faites dans les parties précédent, on privilégiera le modèle SARIMA(0,1,1)(0,1,1)12, c'est-à-dire le cas où le modèle s'écrit : $(I-B)\left(I-B^{12}\right)X_t=(I+\theta_1 B)\left(I+\theta_{12}B^{12}\right)\epsilon_t$