Etude dans le temps des décès dût aux accidents de la route au Royaume-Uni

Master 1 MAS Rennes - Série Temporelle

BERNARD Baptiste, MONFRET Dylan, RAKOTOSON Loïc

Pour le 15 mai 2020

Contextualisation & Méthodologie

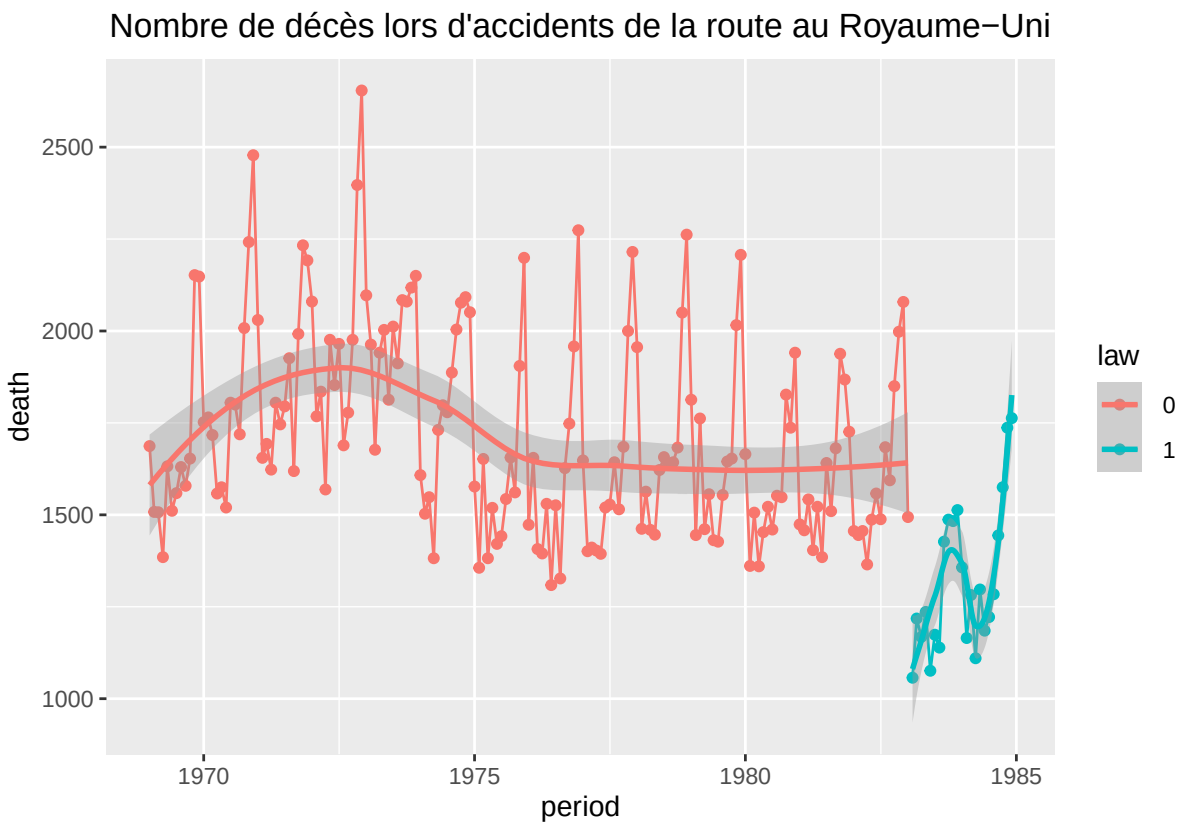
Les données sont les enregistrements mensuels du nombre de morts, `death`, sur les accidents de la route au Royaume-Uni entre Janvier 1969 et Décembre 1984. La loi sur le port obligatoire de la ceinture de sécurité, `law`, a été introduite en Février 1983. La variable `law` prend alors la valeur 0 pour les mois où la loi n'est pas en vigueur, 1 lorsqu'elle est en vigueur.

L'objectif de cette étude sera de proposer diverses modélisations de la série temporelle, qui permettraient par exemple de "simuler" le nombre de décès sur les routes britanniques arpès 1984.

```
period <-
  seq(as.Date('1969-01-01'), as.Date('1984-12-31'), by = "month")

ukdeath <-
  read_delim("data.txt", delim = "  ", col_types = "if") %>%
  mutate(death_log = log(death),
         period = period)
```

```
##      death       law        death_log          period
##  Min.   :1057    0:169   Min.   :6.963   Min.   :1969-01-01
##  1st Qu.:1462    1: 23   1st Qu.:7.287   1st Qu.:1972-12-24
##  Median :1631            Median :7.397   Median :1976-12-16
##  Mean   :1670            Mean   :7.406   Mean   :1976-12-15
##  3rd Qu.:1851            3rd Qu.:7.523   3rd Qu.:1980-12-08
##  Max.   :2654            Max.   :7.884   Max.   :1984-12-01
```

```
## # A tibble: 10 x 4
##    death law   death_log period
##    <int> <fct>     <dbl> <date>
##  1  1687 " 0"       7.43 1969-01-01
##  2  1508 " 0"       7.32 1969-02-01
##  3  1507 " 0"       7.32 1969-03-01
##  4  1385 " 0"       7.23 1969-04-01
##  5  1632 " 0"       7.40 1969-05-01
##  6  1511 " 0"       7.32 1969-06-01
##  7  1559 " 0"       7.35 1969-07-01
##  8  1630 " 0"       7.40 1969-08-01
##  9  1579 " 0"       7.36 1969-09-01
## 10  1653 " 0"       7.41 1969-10-01
```

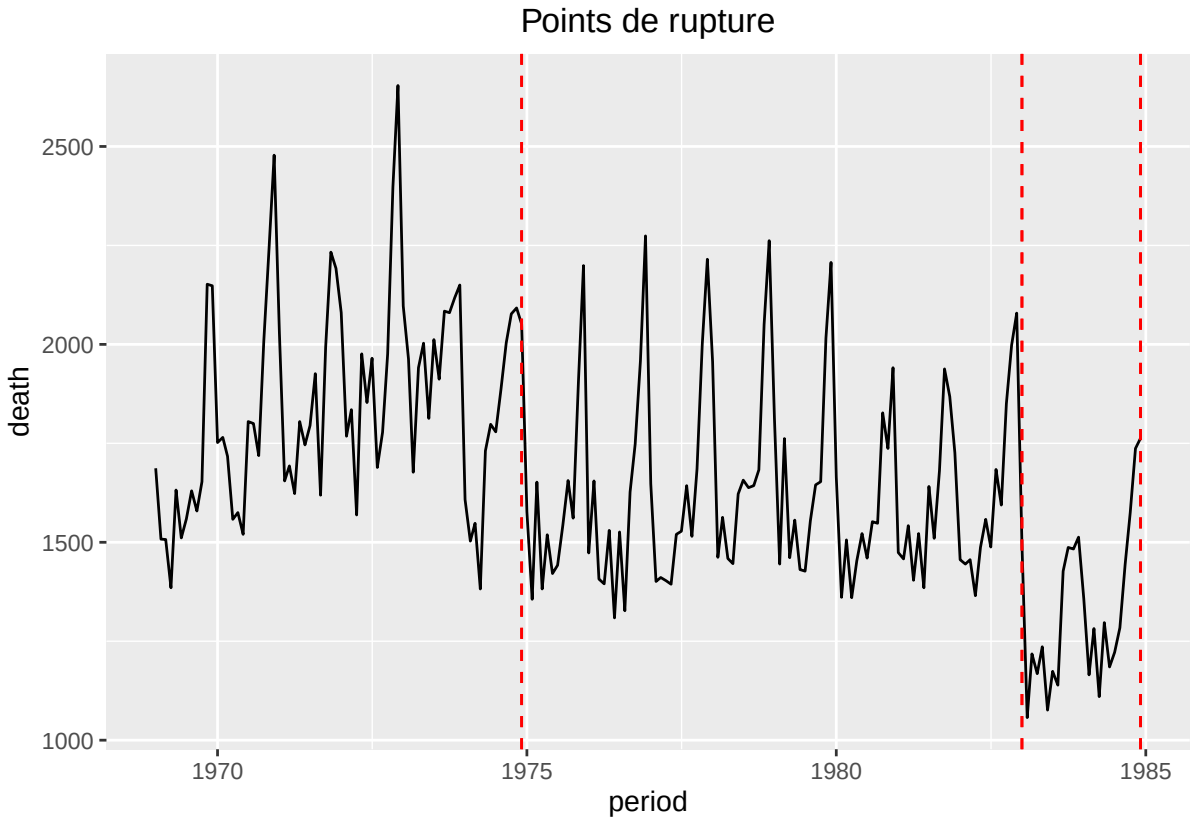


Une première observation de la série, avec mise en évidence des périodes d'application de la loi sur le port de la ceinture, permet de visualiser deux éléments constitutifs de la série :

- Le premier étant la périodicité du nombre de mort sur les routes, avec des accroissements significatifs à chaque fin d'année.
- Le second étant la tendance générale de la série, bien mis en relief par la régression locale (méthode non paramétrique, *LOESS*). Le nombre de décès chaque année à tendance à **s'accroitre entre 1969 et 1973**, à **décroitre entre 1973 et 1976**, à **stagner de 1976 à 1983**, avant de **chuter brutalement avec la mise en application de la loi sur le port de la ceinture en Février 1983**.

Intuitivement, nous pourrions construire nos modèles autour d'une saisonalité des décès sur les routes par rapport fin d'année (période des fêtes), soit **un pique de décès tous les 12 mois** ; et en prenant en compte 2 à 3 phases de la série temporelle (2 phases si l'on se ramène uniquements à la période sans la loi et la période avec application de celle-ci).

C'est point de rupture sont aussi visualisable avec le package **`changepoint`** et sa fonction `cpt.meanvar`, et cela sans prendre en compte la variable `law`.



Pour la suite de l'étude, nous travaillerons avec le nombre de décès passé au log. Une variabilté réduite peut nous garantir un meilleur ajustement en concervant la forme de la série.

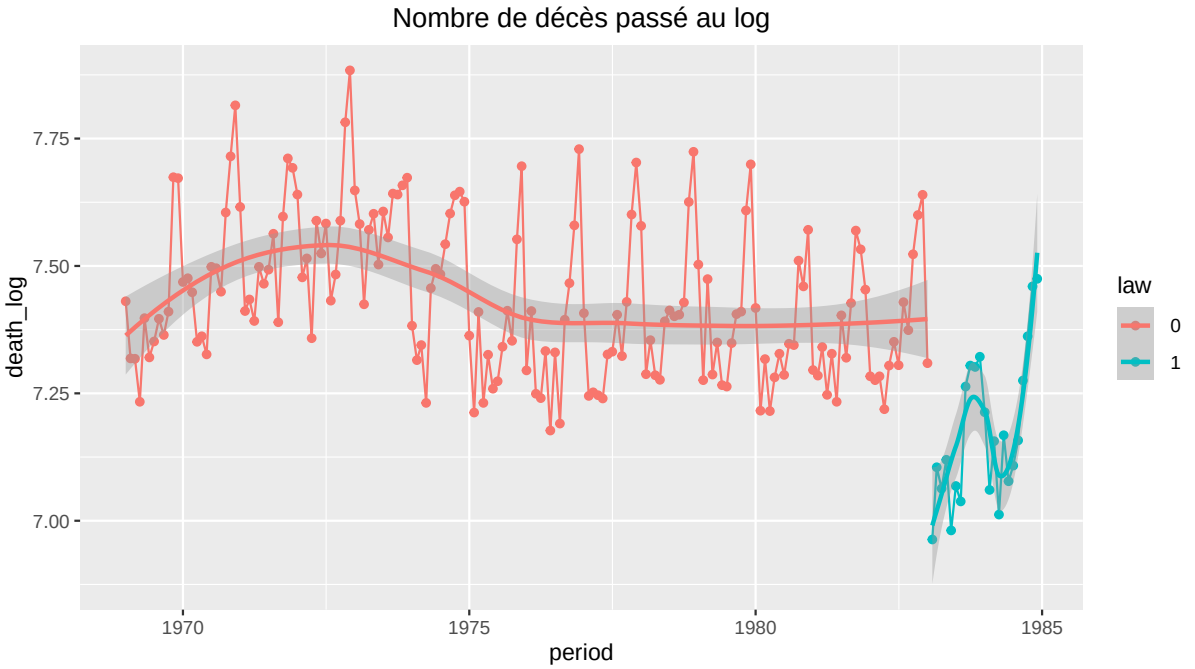
```
var(ukdeath$death)
```

```
## [1] 83874.51
```

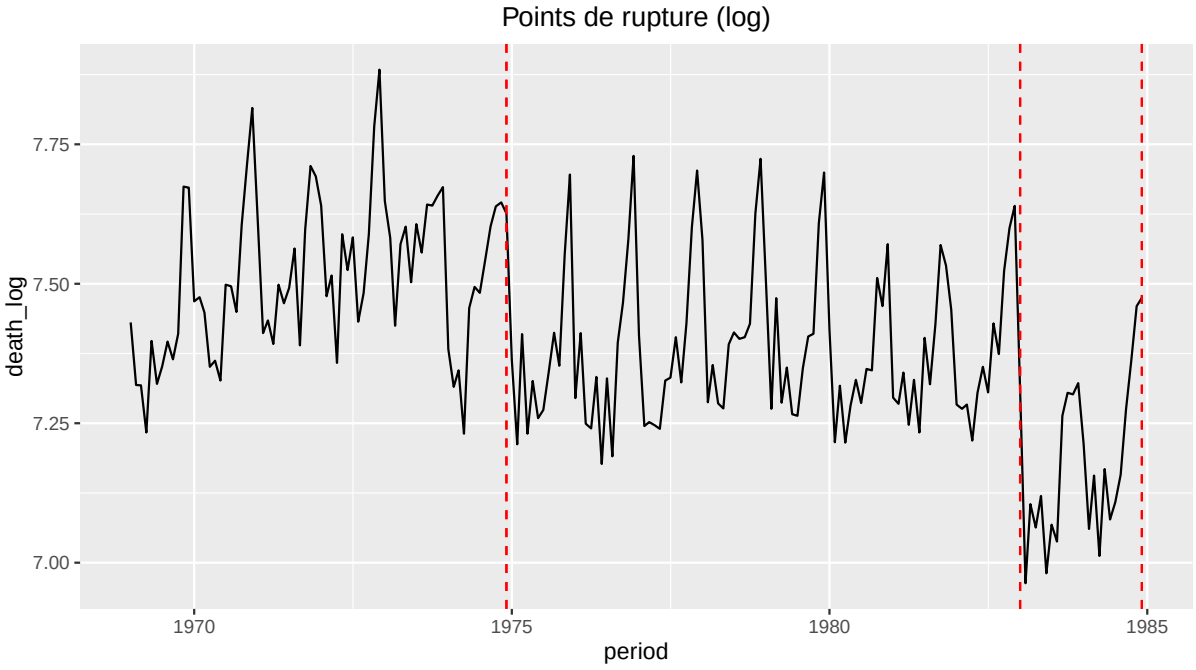
```
var(ukdeath$death_log)
```

```
## [1] 0.02935256
```

```
ggplot(ukdeath) +
  aes(x = period, y = death_log, color = law) +
  geom_point() + geom_line() + stat_smooth(method = "loess") +
  labs(title = "Nombre de décès passé au log") +
  theme(plot.title = element_text(hjust = 0.5))
```




```
ts_ukdeath_log <-
  ts(
    data = ukdeath$death_log,
    start = c(1969, 1),
    frequency = 12
  )

ts_ukdeath_log %>%
  changepoint::cpt.meanvar(method = "PELT", minseglen = 11)  %>%
  autoplot() +
  labs(title = "Points de rupture (log)", x = "period", y = "death_log") +
  theme(plot.title = element_text(hjust = 0.5))
```



Visualiser et confirmer la saisonalité

Pour construire les modèles adéquats à notre problème, nous devons avoir une idée précise des caractéristiques de la série. Il s'agit ici de confirmer si la saisonalité des décès est bien de 12 mois, et de l'autocorrélation des données (lien ou similitude entre l'été 1963 et l'été 1973, par exemple).

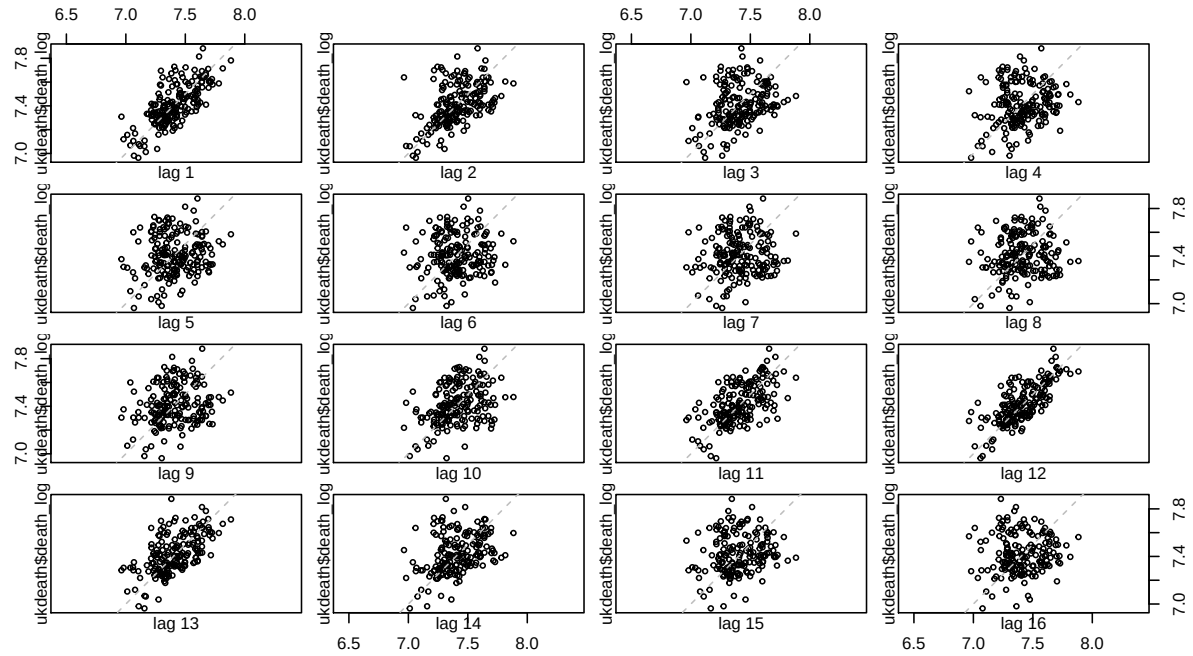
Lag Plot, *Month Plot* et autocorrélations de la série

Les représentations graphiques de la variable d'intérêt par rapport au temps sont de bons indicateurs pour émettre des hypothèses quant à la saisonalité d'une série temporelle.

- Le **Lag Plot**, mettant en perspective la valeur d'une variable à un instant t et à un instant $t-h$ permet de visualiser le lien linéaire entre deux instants et l'on peut répondre au question du type : *"Est-ce qu'un même schéma se répète à 6 mois / 2 ans / 10 secondes d'intervalles" ? Cela se traduit par le plus ou moins bon alignement des points sur la première bissectrice.
- Le **Month Plot** décompose la série par mois. C'est asssez utile lorsque les données sont renseignés mensuellement, on peut alors l'évolution de la variable par mois et au cours des années, avec par défaut l'indication de la valeur moyenne pour chaque mois. On peut alors constater les mois clés de la saisonalité d'une série.
- Les graphiques de représentation de l'autocorrélation et de l'autocorrélation partielle, respectivement **ACF** et **PACF**, mettent eux en évidence le lien entre une valeur prise à un instant t et celui à un instant $t-h$. L'autocorrélation partielle prend en compte en plus de cela le temps écoulé, en incluant les décalages précédents, donc les valeurs à $t-h-k, k < h$.

Et donc, en ce qui nous concerne, le `lag.plot` mais bien en évidence le lien linéaire entre le nombre de décès à un mois t et le nombre de décès à $t-12$ mois (et légèrement à $t-1$, mais probablement puisque les comportements individuelles ne change pas radicalement d'un mois sur l'autre).

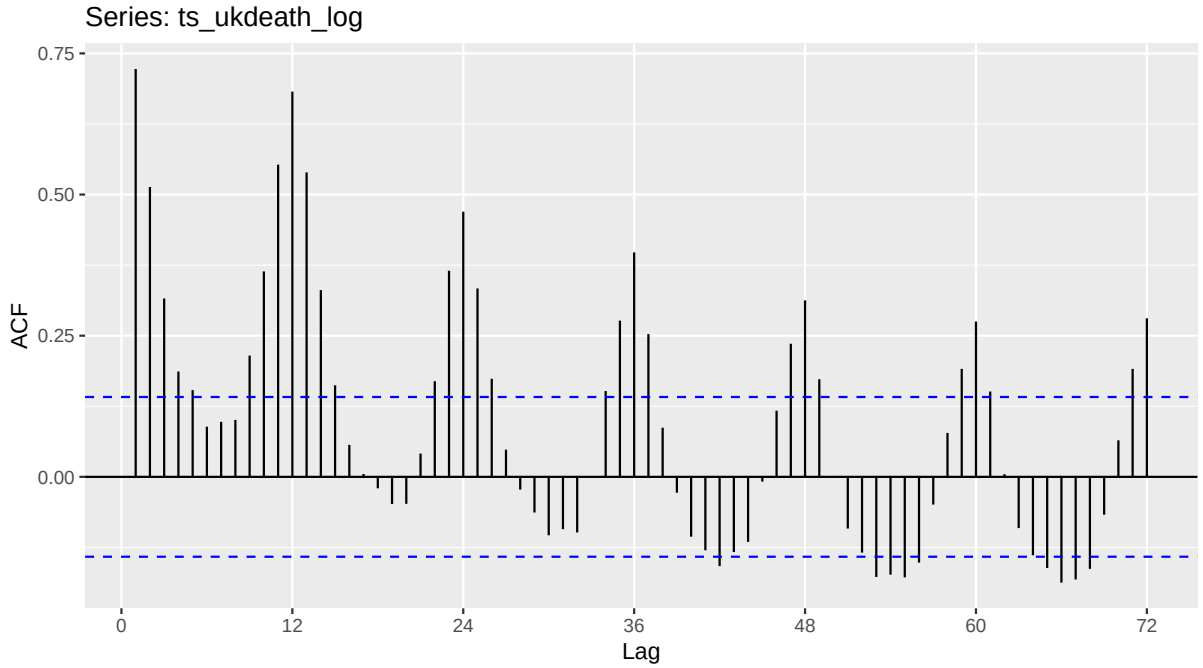
```
lag.plot(ukdeath$death_log, lags = 16)
```



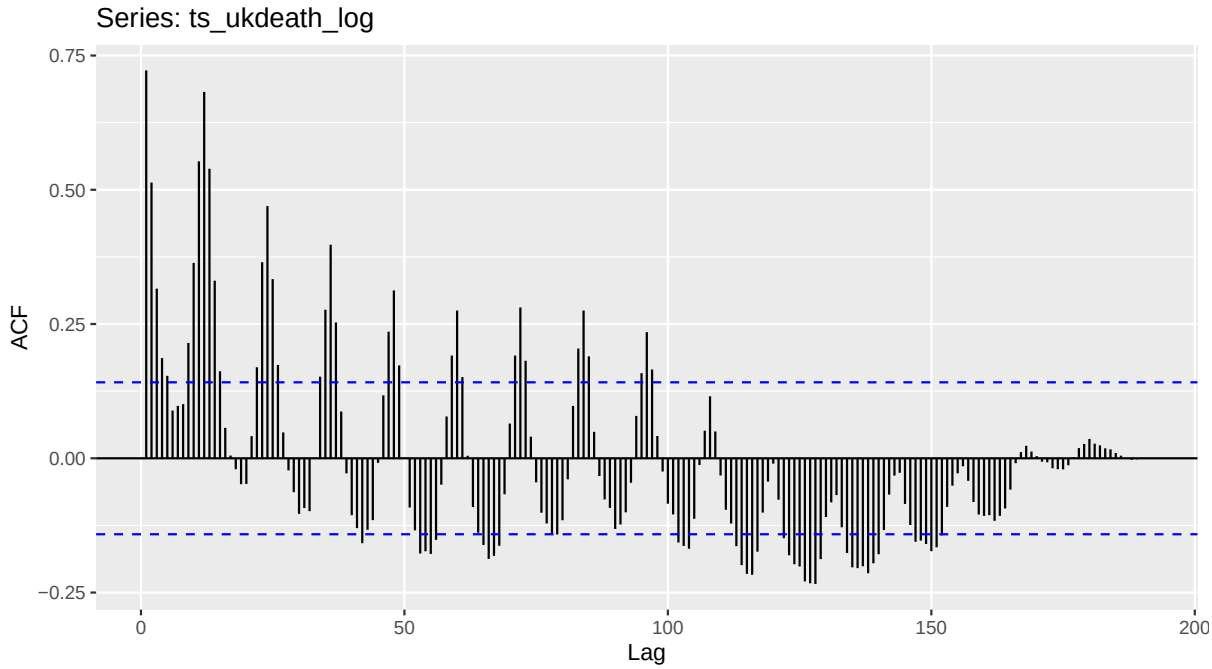
C'est une information confirmé également par l'ACF. L'autocorrélation fini par s'annuler plus significativement qu'une passer le décalage seuil de environ 150 mois, correspondant à l'entrée en vigueur de la loi sur le port de la ceinture, et l'effondrement du nombre de décès.

Le PACF montre également des signes de la saisonalité, mais avec pour nuance supplémentaire que l'ensemble des valeurs prises précédement par la série n'influe pas les valeurs futures. Les valeurs approchant 0 après un décalage seuil de 25 mois.

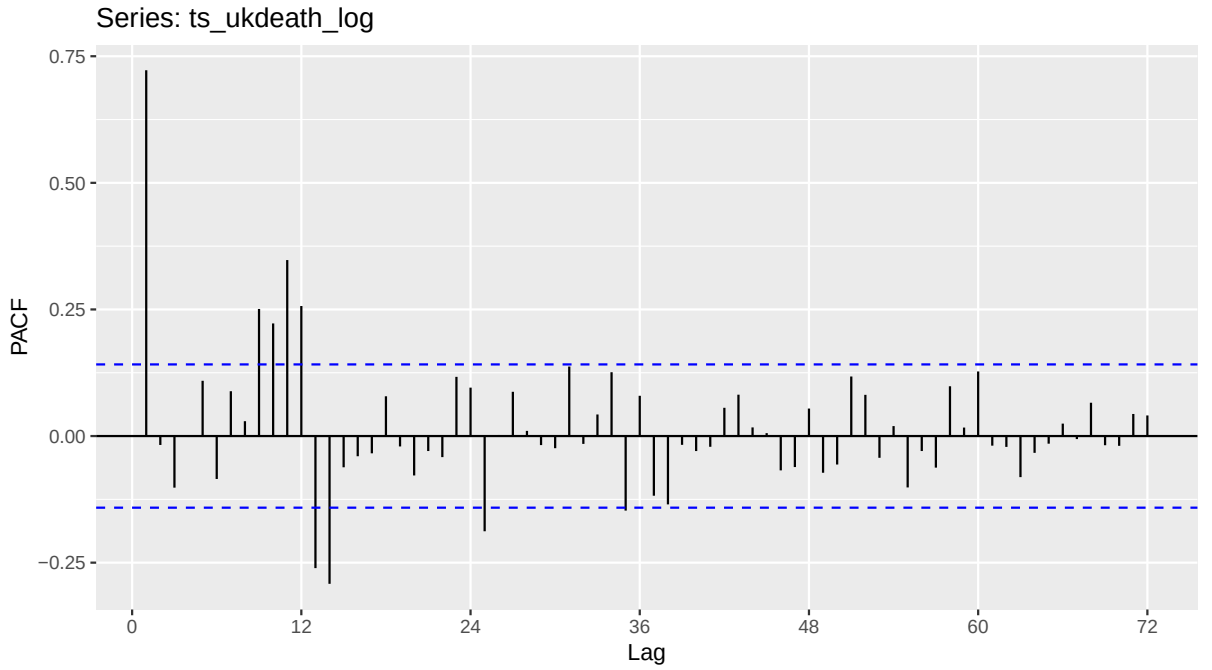
```
ggAcf(ts_ukdeath_log, 72)
```




```
ggAcf(ts_ukdeath_log, 191)
```



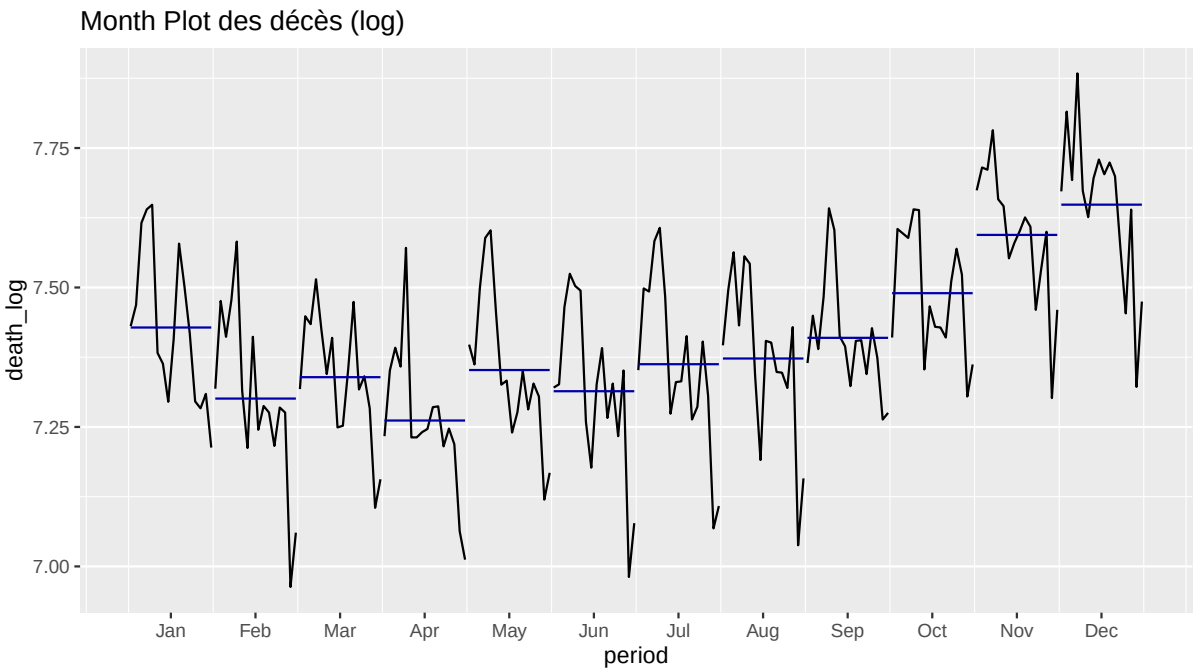

```
ggPacf(ts_ukdeath_log, lag.max = 72)
```



Enfin le Month Plot fini de démontrer la tendance des décès par mois :

- En année : une diminution brutale du nombre de décès vers les dernières années de mesure du nombre de décès.
- En mois : un accroissement du nombre de décès pendant l'hiver et les fêtes de fin d'année (de octobre à janvier) et une période plus calme avec les beaux jours (de mars à août).

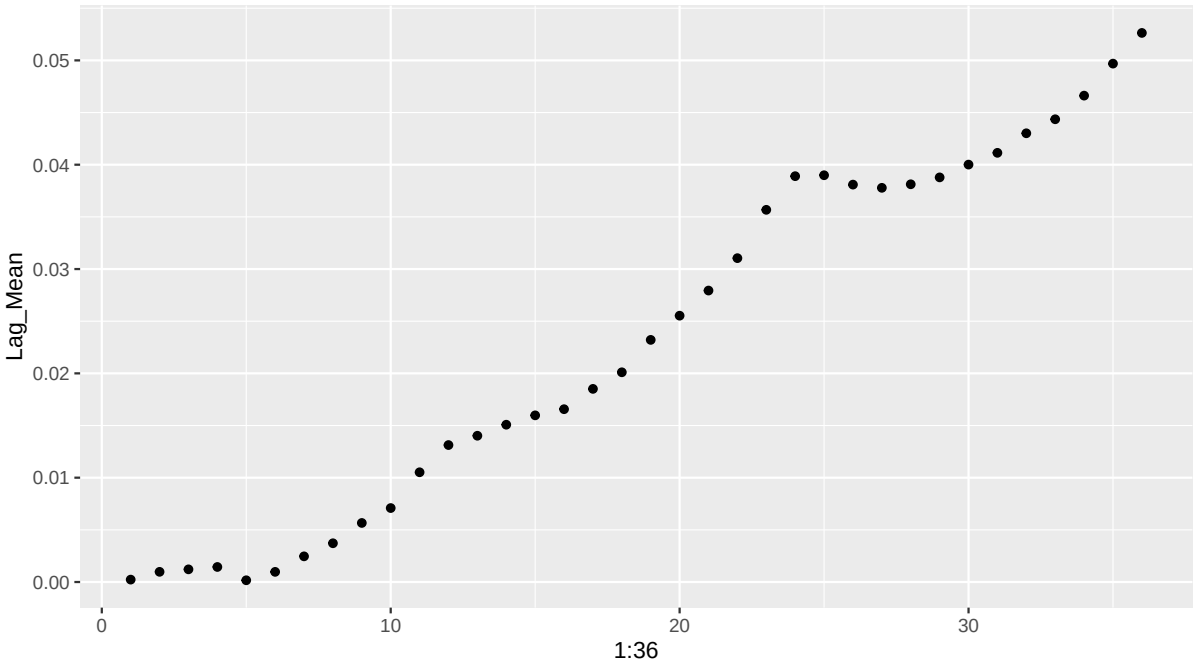
```
ggmonthplot(ts_ukdeath_log) +
  labs(title = "Month Plot des décès (log)", x = "period", y = "death_log")
```



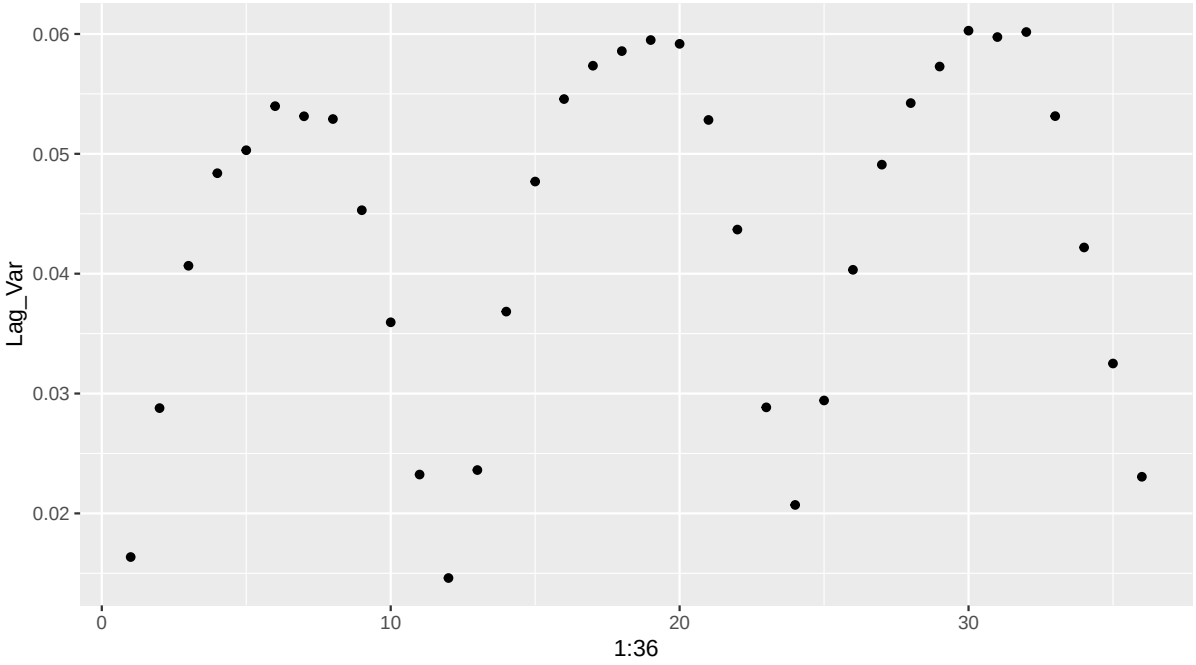
Opérateurs de différence

La période (lag)

```
Lag_Mean <- NULL
Lag_Var <- NULL

for (ind in 1:36) {
  diff <- diff(ukdeath$death_log, ind, 1)
  Lag_Mean[ind] <- abs(mean(diff))
  Lag_Var[ind] <- var(diff)
}

ggplot() + aes(y = Lag_Mean, x = 1:36) + geom_point()
```



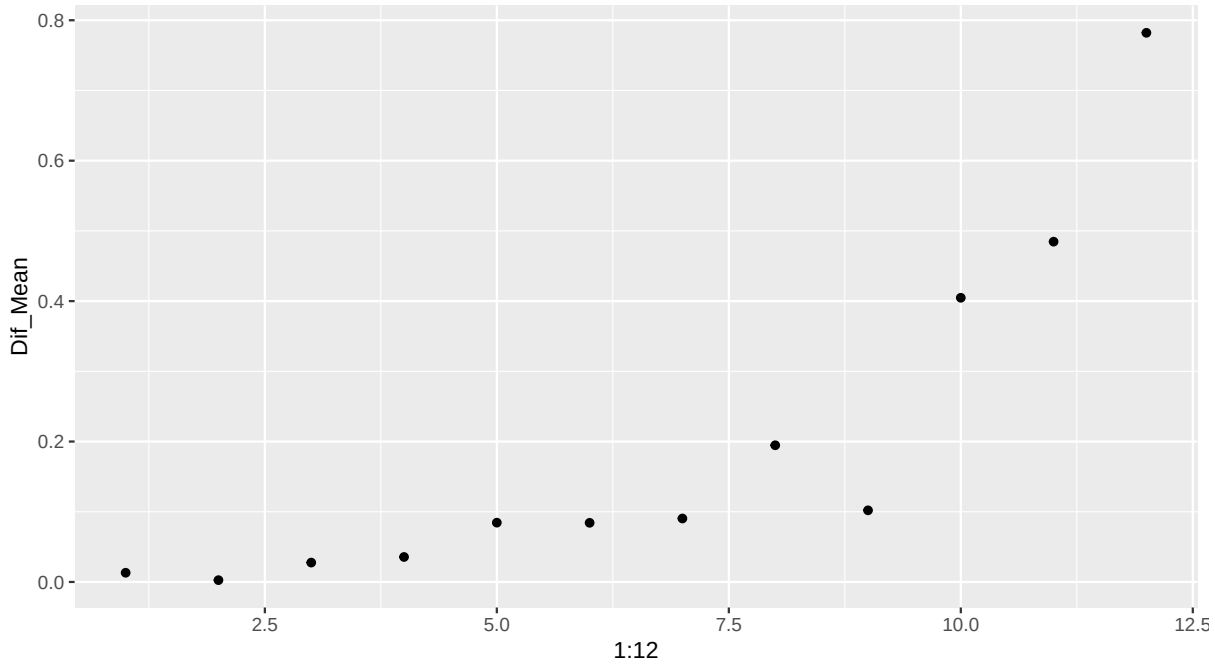

```
ggplot() + aes(y = Lag_Var, x = 1:36) + geom_point()
```



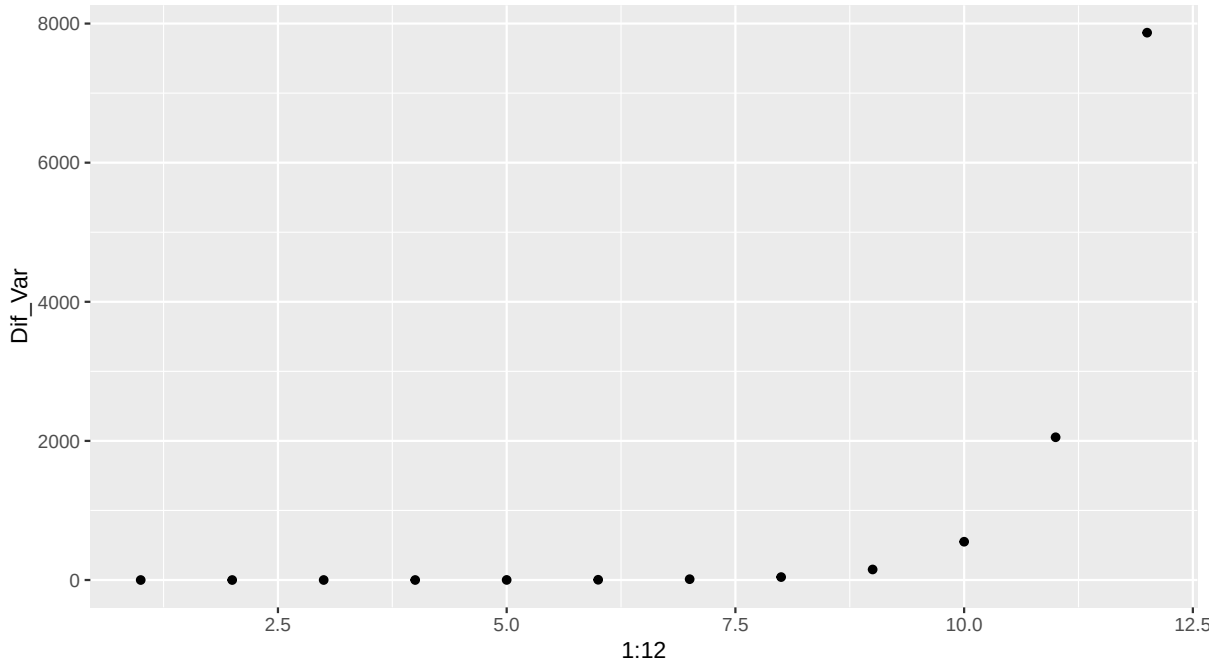
Le degré (differencies)

```
Dif_Mean <- NULL
Dif_Var <- NULL

for (ind in 1:12) {
  diff <- diff(ukdeath$death_log, 12, ind)
  Dif_Mean[ind] <- abs(mean(diff))
```

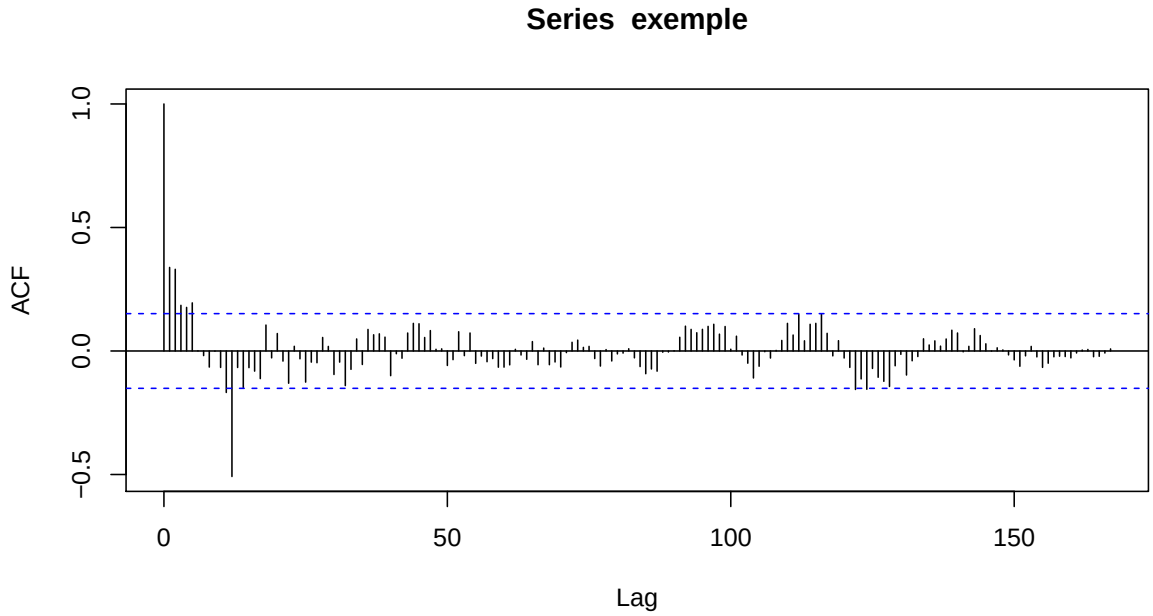
```
  Dif_Var[ind] <- var(diff)
}

ggplot() + aes(y = Dif_Mean, x = 1:12) + geom_point()
```



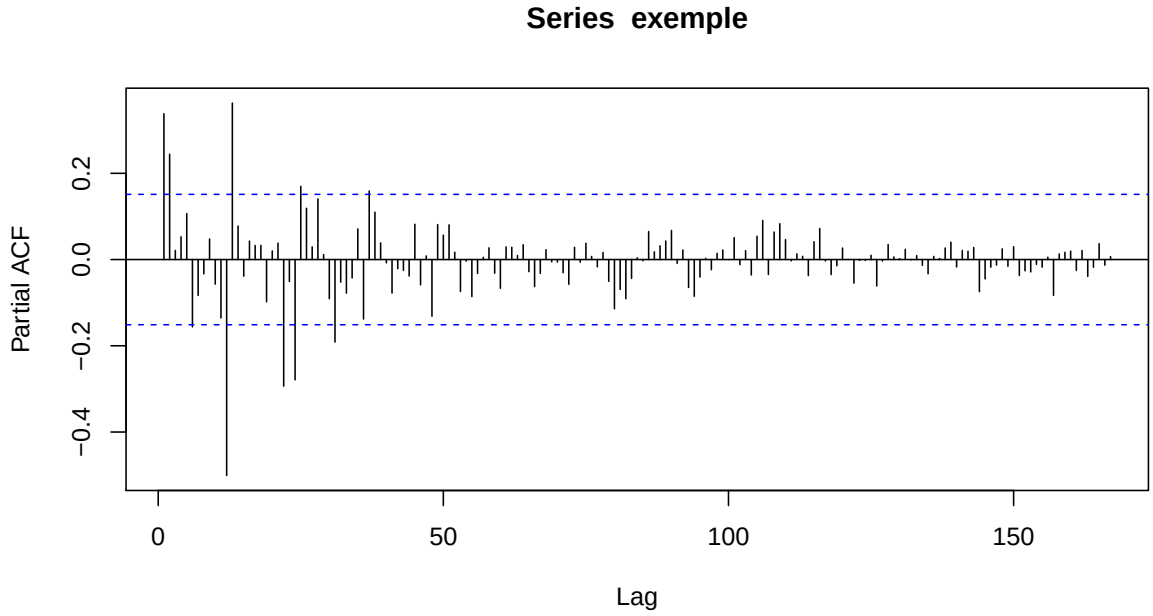

```
ggplot() + aes(y = Dif_Var, x = 1:12) + geom_point()
```




```
exemple <- diff(ukdeath$death_log, 12 , 2)
acf(exemple, 191)
```




```
pacf(exemple, 191)
```



Modélisations

Modèle additif

lag : 12 diff : 1 ou 2

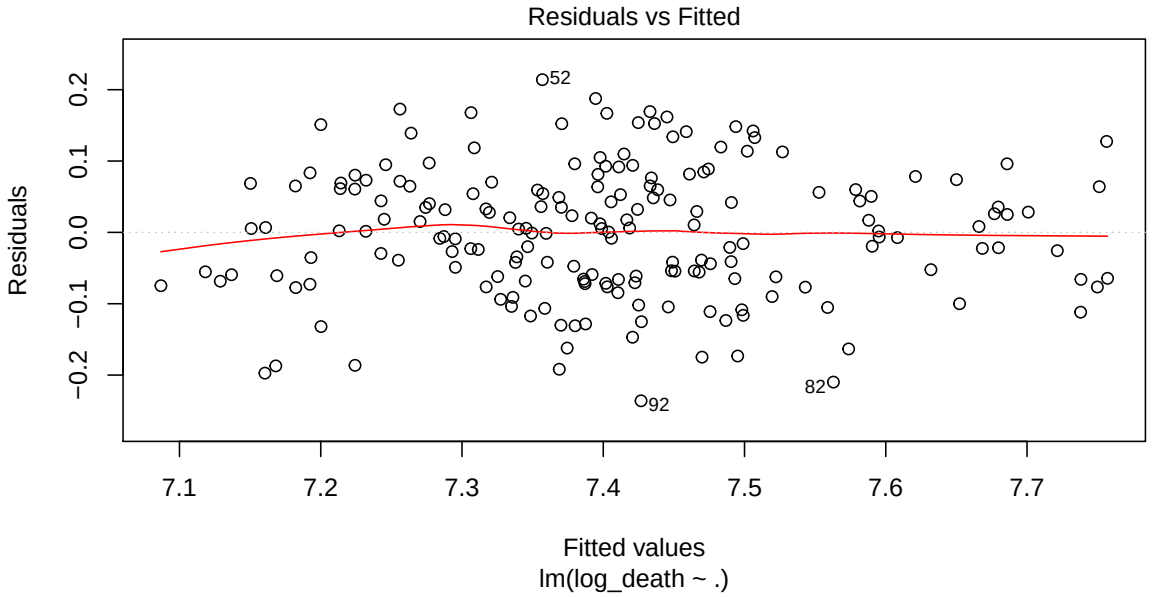
```
t <- 1:192
sinusoides <- t %o% c(rep(1:5, 2)) * pi / 6
```

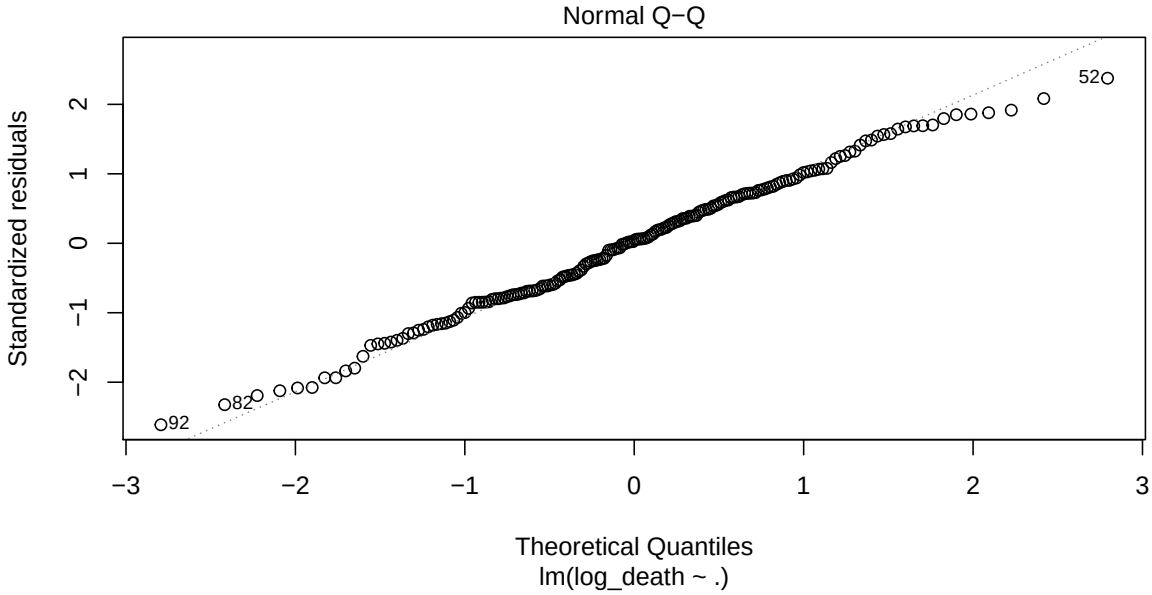
```r
sinusoides[, 1:5] <- sin(sinusoides[, 1:5])
sinusoides[, 6:10] <- cos(sinusoides[, 6:10])
sinusoides <- as.data.frame(sinusoides)
names(sinusoides) <-
  c(paste("sin_", 1:5, sep = ""), paste("cos_", 1:5, sep = ""))
```

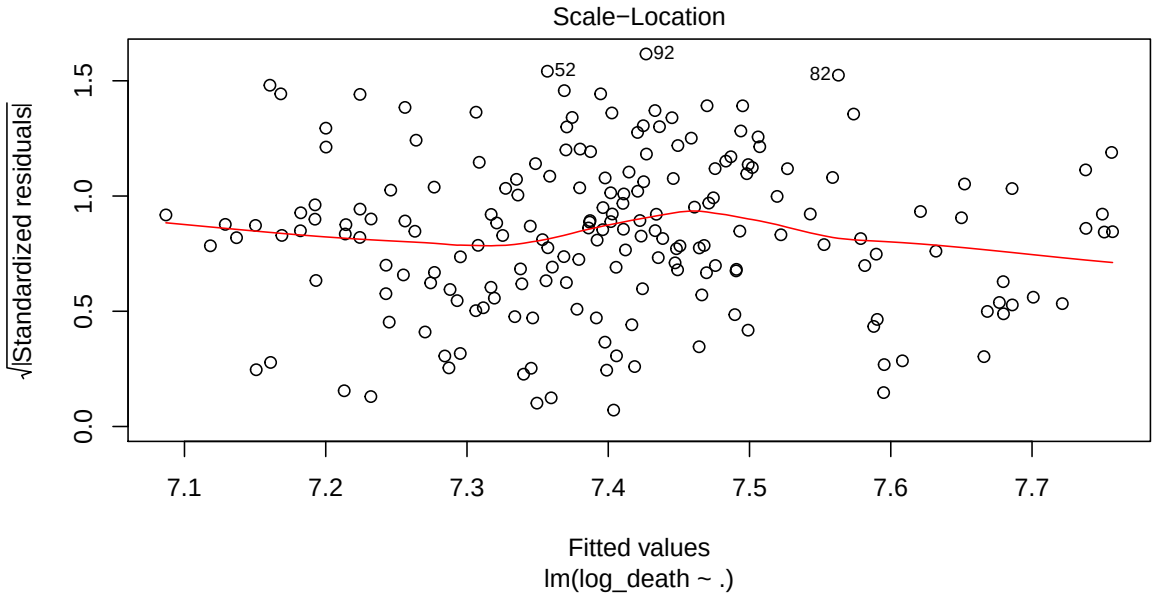
```r
log_death <- ukdeath$death_log
df <- data.frame(log_death, t, t ^ 2, t ^ 3)
df <- cbind(df, sinusoides)
ModAddifitf <- lm(data = df, log_death ~ .)
summary(ModAddifitf)
```

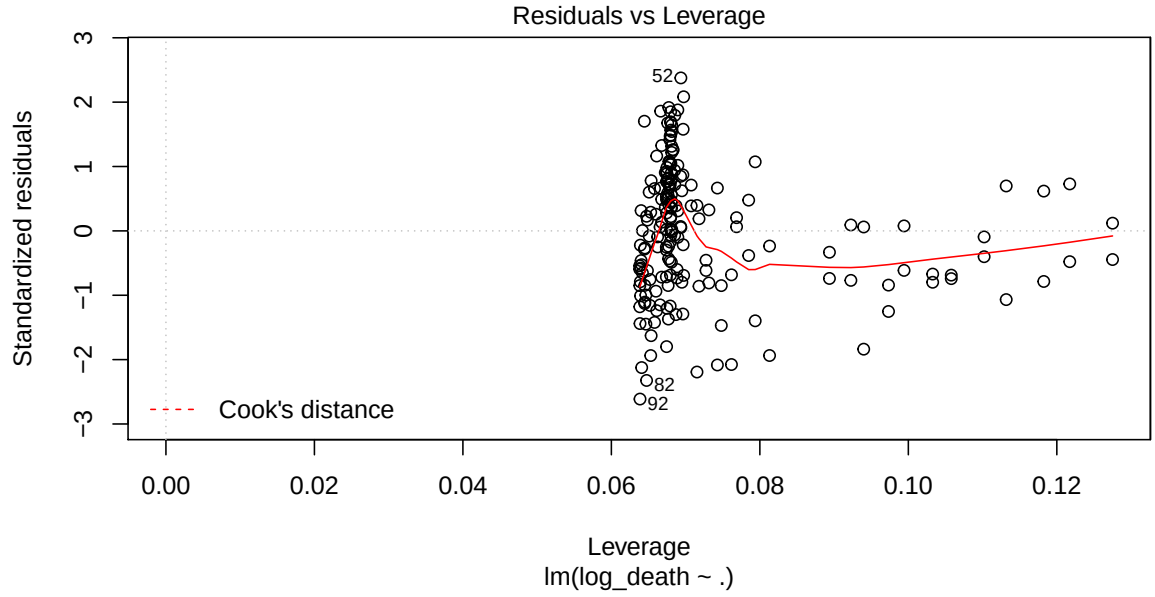
```
##
## Call:
## lm(formula = log_death ~ ., data = df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.236069 -0.064984  0.003402  0.064214  0.214065
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.460e+00  2.760e-02 270.258  < 2e-16 ***
## t            2.068e-03  1.236e-03   1.673 0.096058 .
## t.2         -2.903e-05  1.486e-05  -1.954 0.052268 .
## t.3          5.937e-08  5.062e-08   1.173 0.242437
## sin_1       -7.471e-02  9.567e-03  -7.809 4.77e-13 ***
## sin_2       -3.595e-02  9.539e-03  -3.768 0.000223 ***
## sin_3       -1.897e-02  9.534e-03  -1.990 0.048128 *
## sin_4       -1.199e-02  9.532e-03  -1.258 0.210199
## sin_5        1.651e-02  9.531e-03   1.732 0.085045 .
## cos_1        1.147e-01  9.534e-03  12.034  < 2e-16 ***
## cos_2        6.297e-02  9.534e-03   6.606 4.42e-10 ***
## cos_3        3.101e-02  9.534e-03   3.253 0.001367 **
## cos_4        2.309e-02  9.534e-03   2.422 0.016458 *
## cos_5        2.564e-02  9.534e-03   2.689 0.007841 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09338 on 178 degrees of freedom
## Multiple R-squared:  0.7231, Adjusted R-squared:  0.7029
## F-statistic: 35.76 on 13 and 178 DF,  p-value: < 2.2e-16
```

```r
plot(ModAddifitf)
```

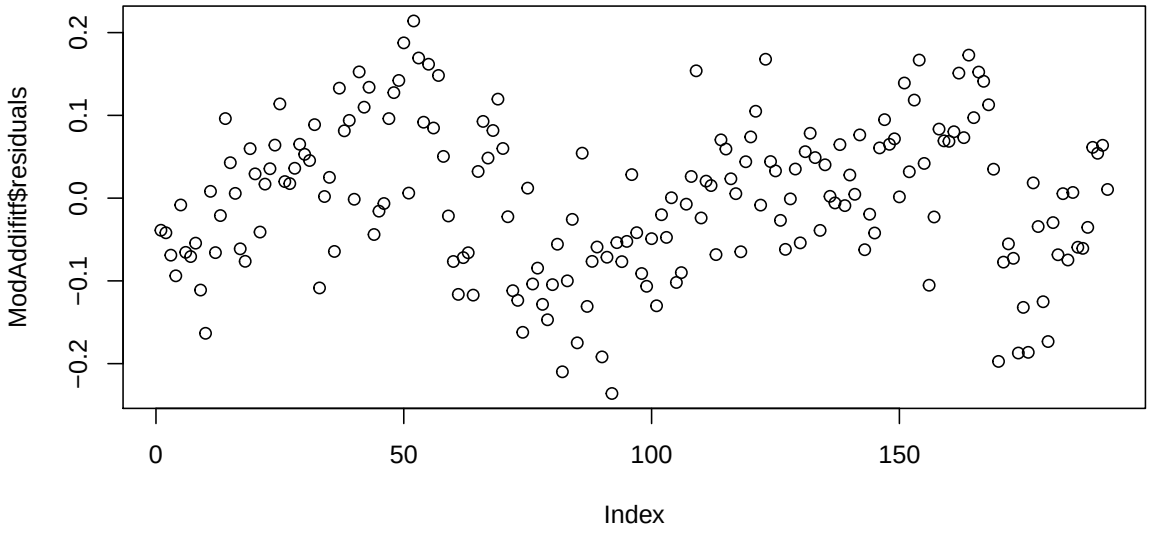

Residuals vs Fitted
Residuals
Fitted values
lm(log_death ~ .)
52
92
82
7.1
7.2
7.3
7.4
7.5
7.6
7.7
−0.2
−0.1
0.0
0.1
0.2


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(log_death ~ .)
52
92
82
−3
−2
−1
0
1
2
3





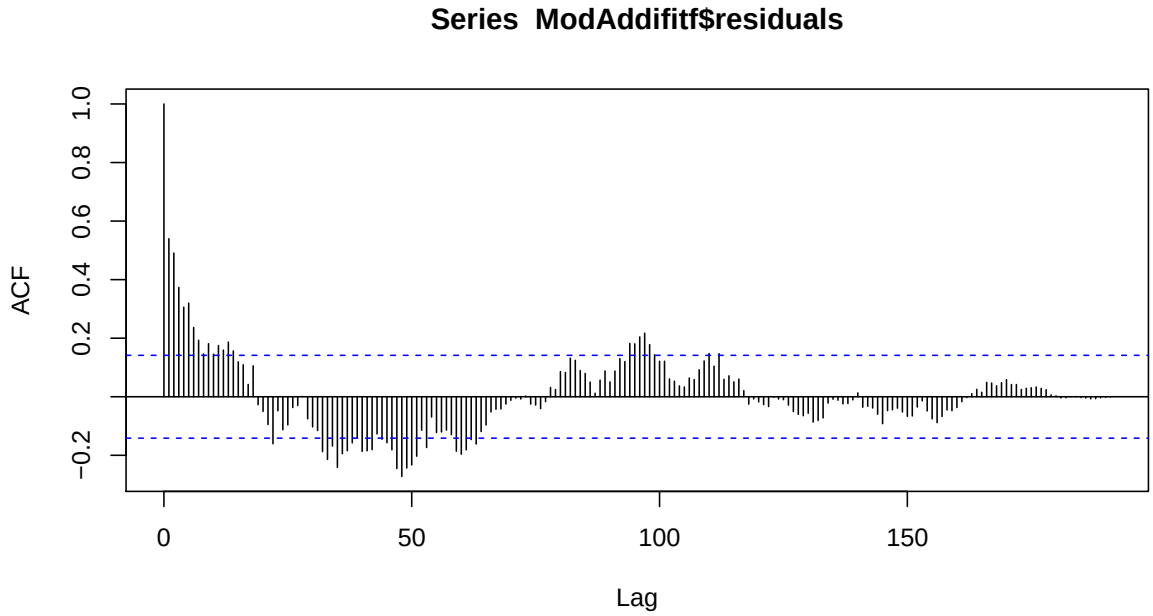

```
plot(ModAddifitf$residuals)
```




```
t.test(ModAddifitf$residuals)
```

```
##
##  One Sample t-test
##
## data:  ModAddifitf$residuals
## t = 2.5758e-16, df = 191, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.01283293  0.01283293
## sample estimates:
##    mean of x
## 1.675797e-18
```

```
acf(ModAddifitf$residuals, 192)
```




```
MA <- arima(log_death, c(1, 2, 1))
plot(MA$residuals)
```