Linear Regression for Predictive Modeling

Katie Schilling - 501130072

07/02/2022

```
library(readxl)
library(caTools)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
require(lmtest)
```

```
## Loading required package: lmtest

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

Import bothe Clean datasets

```
Covid_monthly <- read.csv("C:/Users/Katie Schilling/Downloads/covid_monthly_clean.csv")
Vital_Events <- read.csv("C:/Users/Katie Schilling/Downloads/vital_events_clean.csv")
```

Combine the vital events data with the Covid Monthly data

```
Final_dataset <- merge(x=Vital_Events, y=Covid_monthly, all = TRUE)
```

Check the data and ensure data merged properly

```
summary(Final_dataset)
```

```
##      Date               Births        Marriages         Deaths
##  Length:336         Min.   :10020   Min.   :  597   Min.   : 5926
##  Class :character   1st Qu.:11260   1st Qu.: 2596   1st Qu.: 6706
##  Mode  :character   Median :11818   Median : 3559   Median : 7326
##                     Mean   :11763   Mean   : 5085   Mean   : 7500
##                     3rd Qu.:12288   3rd Qu.: 7627   3rd Qu.: 8094
##                     Max.   :13398   Max.   :11532   Max.   :11390
##                     NA's   :6       NA's   :6       NA's   :6
##   Stillbirths         Covid
##  Min.   :  0.00   Min.   :    11
##  1st Qu.: 73.00   1st Qu.: 22889
##  Median : 90.50   Median : 75935
##  Mean   : 90.74   Mean   : 95144
##  3rd Qu.:114.00   3rd Qu.:134128
##  Max.   :156.00   Max.   :395815
##  NA's   :6        NA's   :313
```

Change the N/A in the Covid Positive Cases to 0 so that the data is not omitted from the predictions

```
Final_dataset$Covid[is.na(Final_dataset$Covid)] = 0
```

See if there are anymore NA's in the dataset

```
Final_dataset %>% filter_all(any_vars(is.na(.)))
```

```
##        Date Births Marriages Deaths Stillbirths  Covid
## 1 2021-07-1     NA        NA     NA          NA  15968
## 2 2021-08-1     NA        NA     NA          NA  67913
## 3 2021-09-1     NA        NA     NA          NA 125560
## 4 2021-10-1     NA        NA     NA          NA  91834
## 5 2021-11-1     NA        NA     NA          NA  75935
## 6 2021-12-1     NA        NA     NA          NA 395815
```

Remove rows with NA as they will skew the results

```
Final_dataset <- na.omit(Final_dataset)
```

Check for NA's to confirm all have been removed

```
Final_dataset %>% filter_all(any_vars(is.na(.)))
```

```
## [1] Date        Births      Marriages   Deaths      Stillbirths Covid
## <0 rows> (or 0-length row.names)
```

```
summary(Final_dataset)
```

```
##      Date               Births        Marriages         Deaths
##  Length:330         Min.   :10020   Min.   :  597   Min.   : 5926
##  Class :character   1st Qu.:11260   1st Qu.: 2596   1st Qu.: 6706
##  Mode  :character   Median :11818   Median : 3559   Median : 7326
##                     Mean   :11763   Mean   : 5085   Mean   : 7500
##                     3rd Qu.:12288   3rd Qu.: 7627   3rd Qu.: 8094
##                     Max.   :13398   Max.   :11532   Max.   :11390
##   Stillbirths         Covid
##  Min.   :  0.00   Min.   :     0
##  1st Qu.: 73.00   1st Qu.:     0
##  Median : 90.50   Median :     0
##  Mean   : 90.74   Mean   :  4289
##  3rd Qu.:114.00   3rd Qu.:     0
##  Max.   :156.00   Max.   :237308
```

```
Final_dataset$Date <- as.Date(Final_dataset$Date,"%Y-%m-%d")
```

```
glimpse(Final_dataset)
```

```
## Rows: 330
## Columns: 6
## $ Date        <date> 1994-01-01, 1994-02-01, 1994-03-01, 1994-04-01, 1994-05-0~
## $ Births      <int> 11631, 11254, 13003, 12576, 13240, 13072, 13045, 12982, 12~
## $ Marriages   <int> 2078, 2650, 2557, 3967, 6493, 7754, 9264, 9194, 8540, 7400~
## $ Deaths      <int> 8094, 6428, 6503, 6224, 6483, 6187, 6196, 5926, 6062, 6515~
## $ Stillbirths <int> 75, 62, 73, 74, 67, 66, 70, 79, 60, 59, 56, 43, 78, 84, 75~
## $ Covid       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
```

Normalize the Data

```
Final_dataset$Births <- scale(Final_dataset$Births, center = T, scale = T)
Final_dataset$Deaths <- scale(Final_dataset$Deaths, center = T, scale = T)
Final_dataset$Marriages <- scale(Final_dataset$Marriages, center = T, scale = T)
Final_dataset$Stillbirths <- scale(Final_dataset$Stillbirths, center = T, scale = T)
Final_dataset$Covid <- scale(Final_dataset$Covid, center = T, scale = T)
```

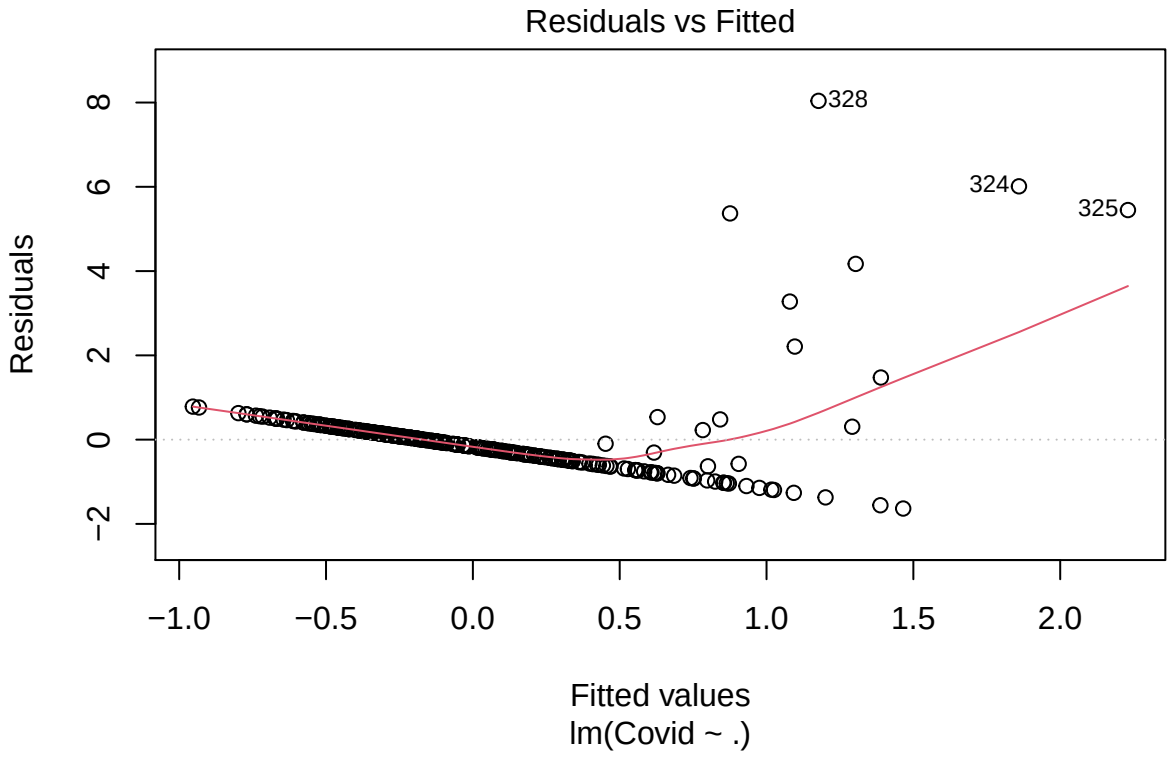
Multiple Linear Regression

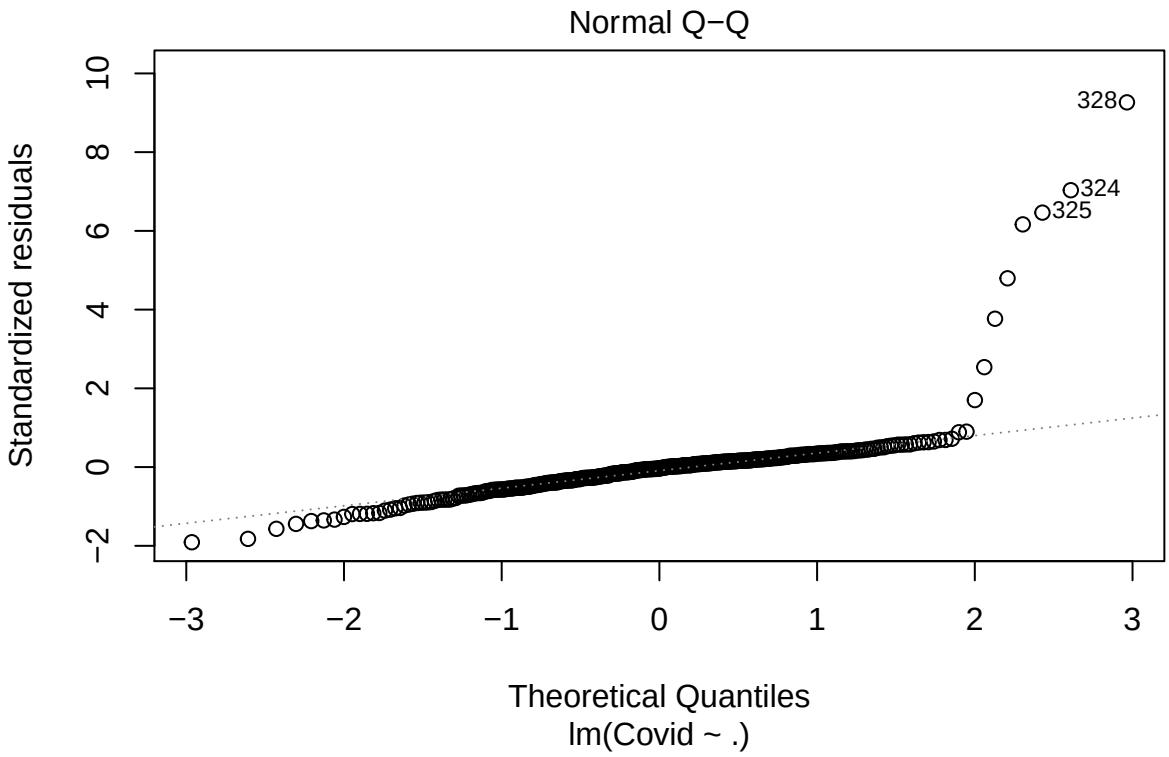
```
MLR_1 <- lm(Covid ~ ., data = Final_dataset)

summary(MLR_1)
```

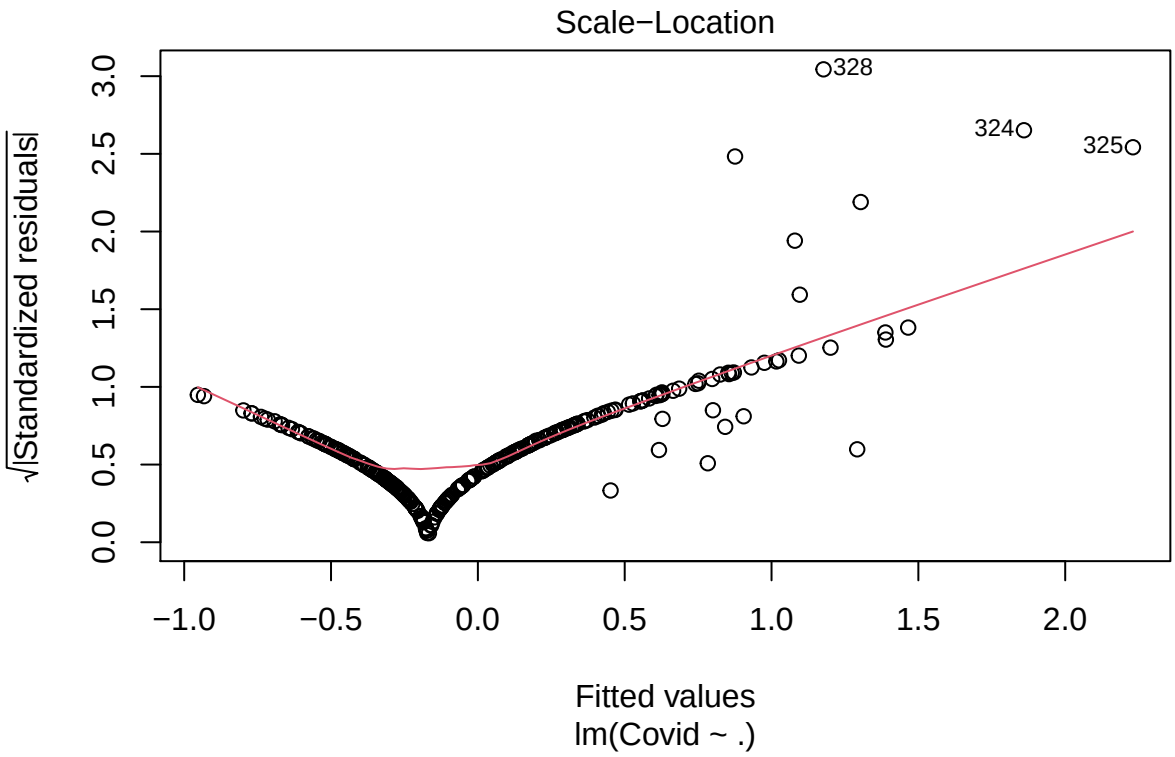
```
##
## Call:
## lm(formula = Covid ~ ., data = Final_dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6350 -0.3439 -0.0302  0.1834  8.0398
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

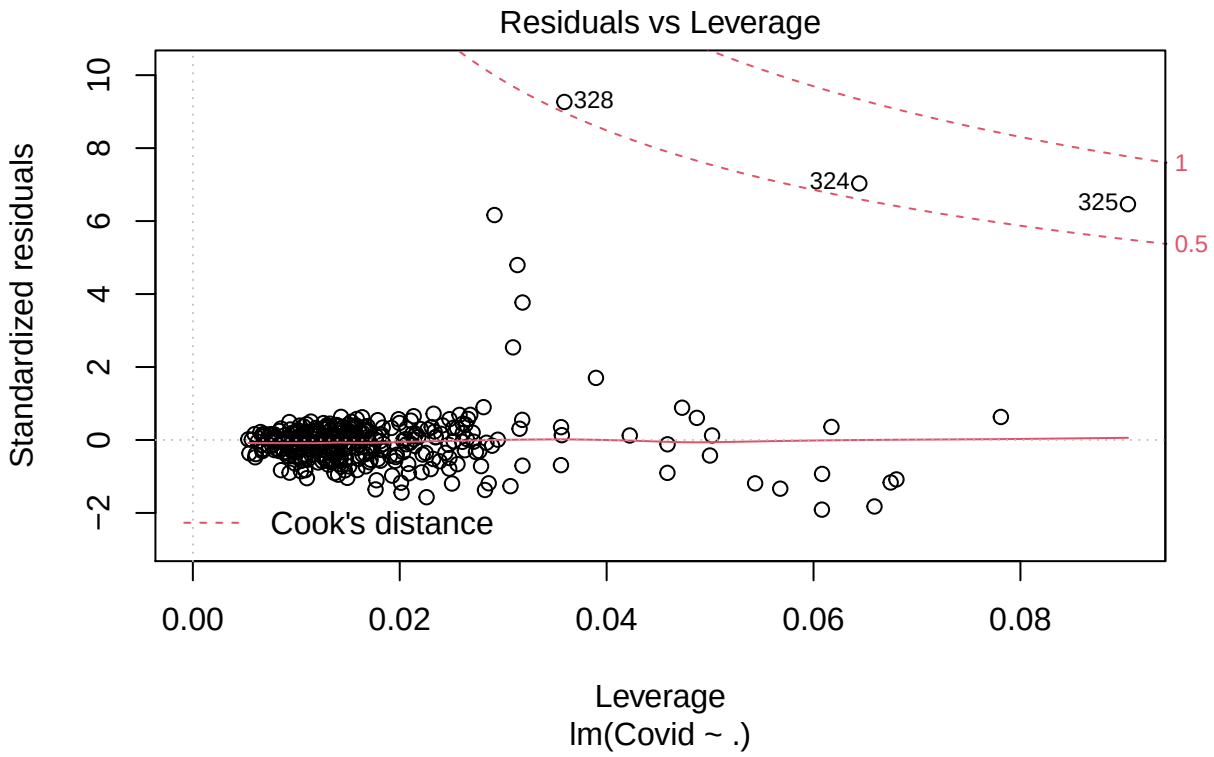
```
## (Intercept) -5.726e-01  4.806e-01  -1.191    0.234
## Date         4.158e-05  3.472e-05   1.197    0.232
## Births      -4.302e-02  6.717e-02  -0.641    0.522
## Marriages    1.168e-01  7.444e-02   1.570    0.117
## Deaths       5.453e-01  1.019e-01   5.351 1.66e-07 ***
## Stillbirths -3.633e-01  7.236e-02  -5.021 8.49e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8838 on 324 degrees of freedom
## Multiple R-squared:  0.2308, Adjusted R-squared:  0.219
## F-statistic: 19.45 on 5 and 324 DF,  p-value: < 2.2e-16
```

```
plot(MLR_1)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Covid ~ .)
328
324
325
−2
0
2
4
6
8
10
−3
−2
−1
0
1
2
3

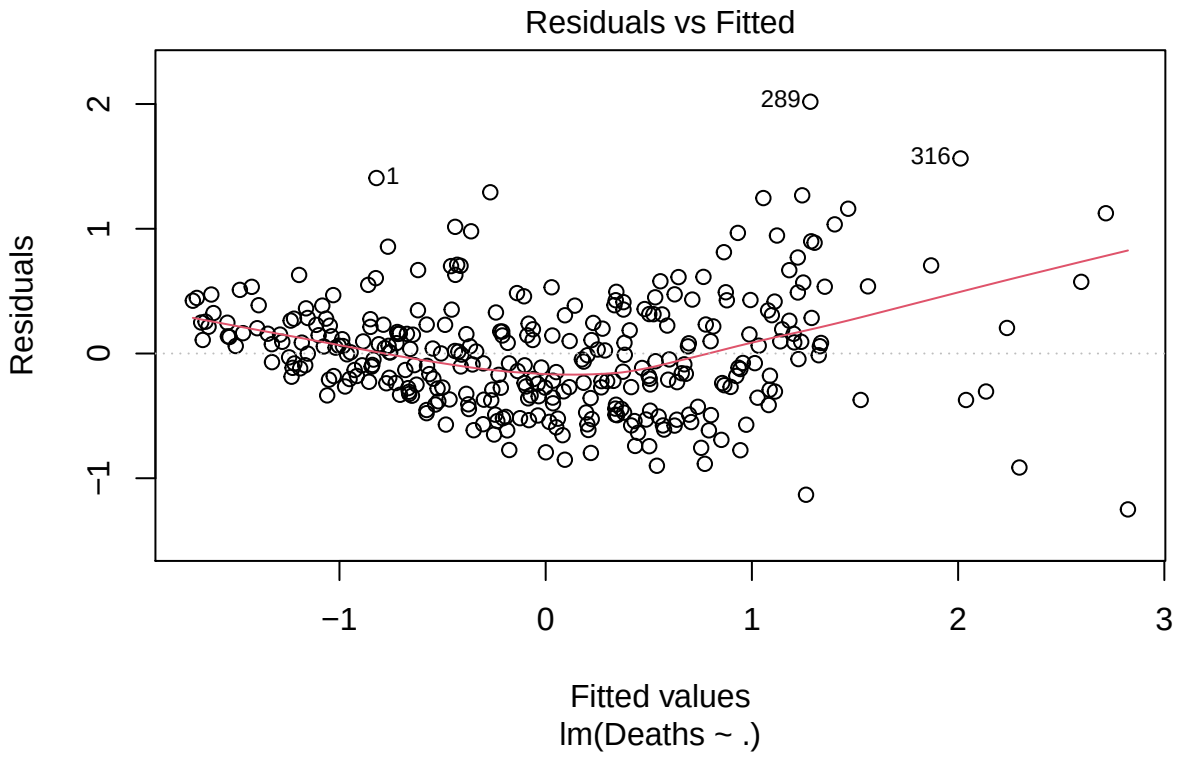

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
328
324
325
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
Fitted values
lm(Covid ~ .)

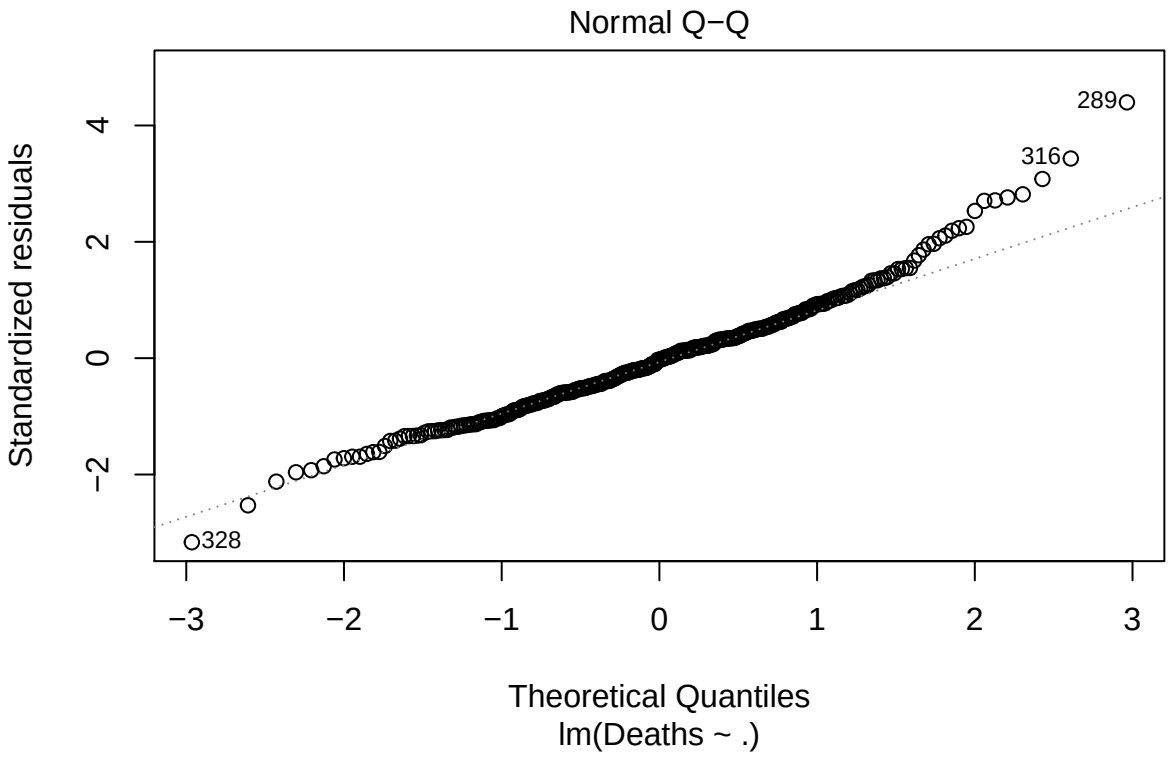


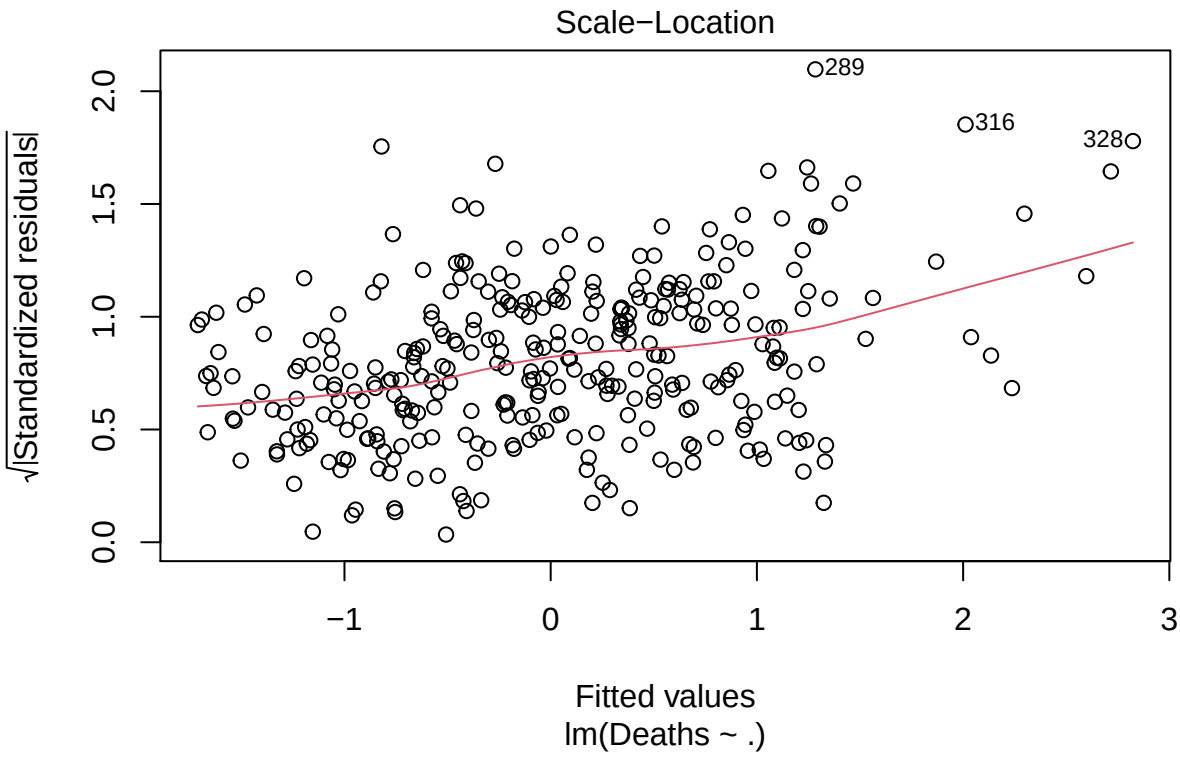

```
MLR_2 <- lm(Deaths ~ ., data = Final_dataset)

summary(MLR_2)
```

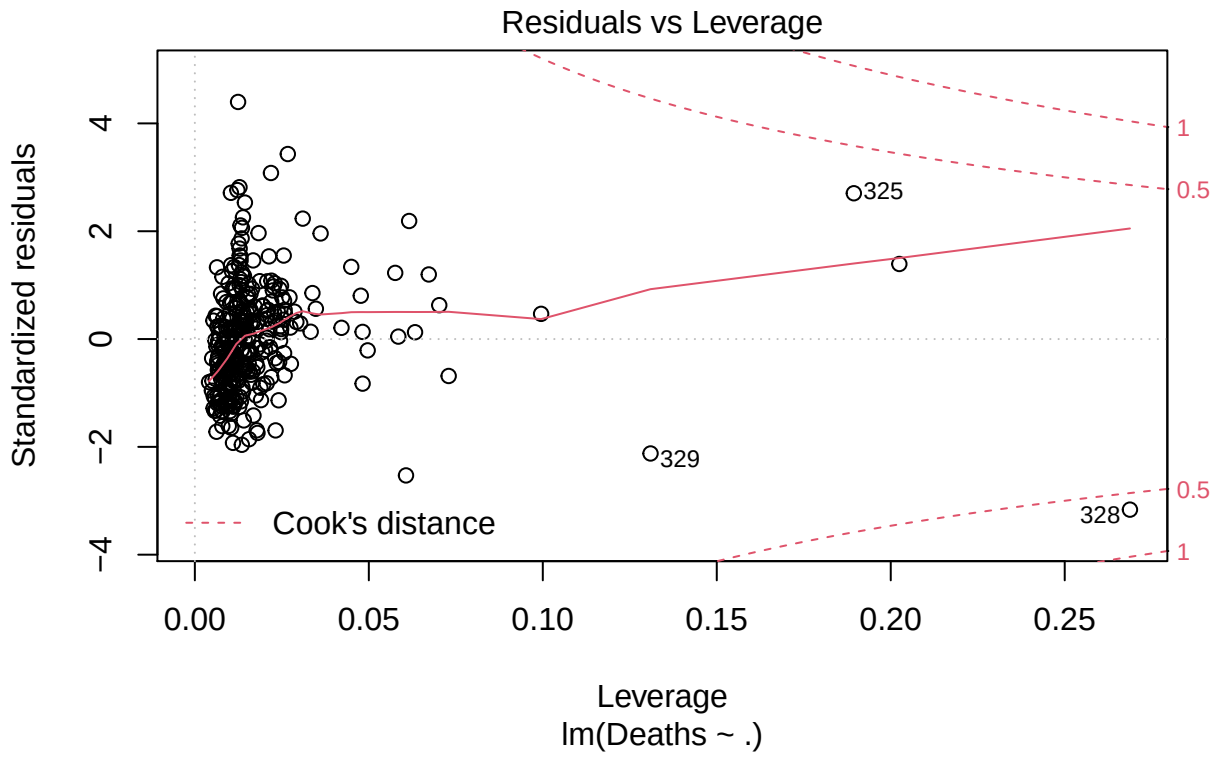
```
##
## Call:
## lm(formula = Deaths ~ ., data = Final_dataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.24974 -0.30443 -0.01222  0.24370  2.01840
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.072e+00  1.850e-01 -16.604  < 2e-16 ***
## Date         2.231e-04  1.331e-05  16.763  < 2e-16 ***
## Births      -8.377e-03  3.512e-02  -0.239   0.8116
## Marriages   -3.537e-01  3.374e-02 -10.483  < 2e-16 ***
## Stillbirths  9.965e-02  3.886e-02   2.564   0.0108 *
## Covid        1.489e-01  2.783e-02   5.351 1.66e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4618 on 324 degrees of freedom
## Multiple R-squared:   0.79,  Adjusted R-squared:  0.7867
## F-statistic: 243.7 on 5 and 324 DF,  p-value: < 2.2e-16
```

```
plot(MLR_2)
```


Residuals vs Fitted
289
316
1
Residuals
Fitted values
lm(Deaths ~ .)
−1
0
1
2
3
−1
0
1
2


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Deaths ~ .)
289
316
328
−3
−2
−1
0
1
2
3
−2
0
2
4

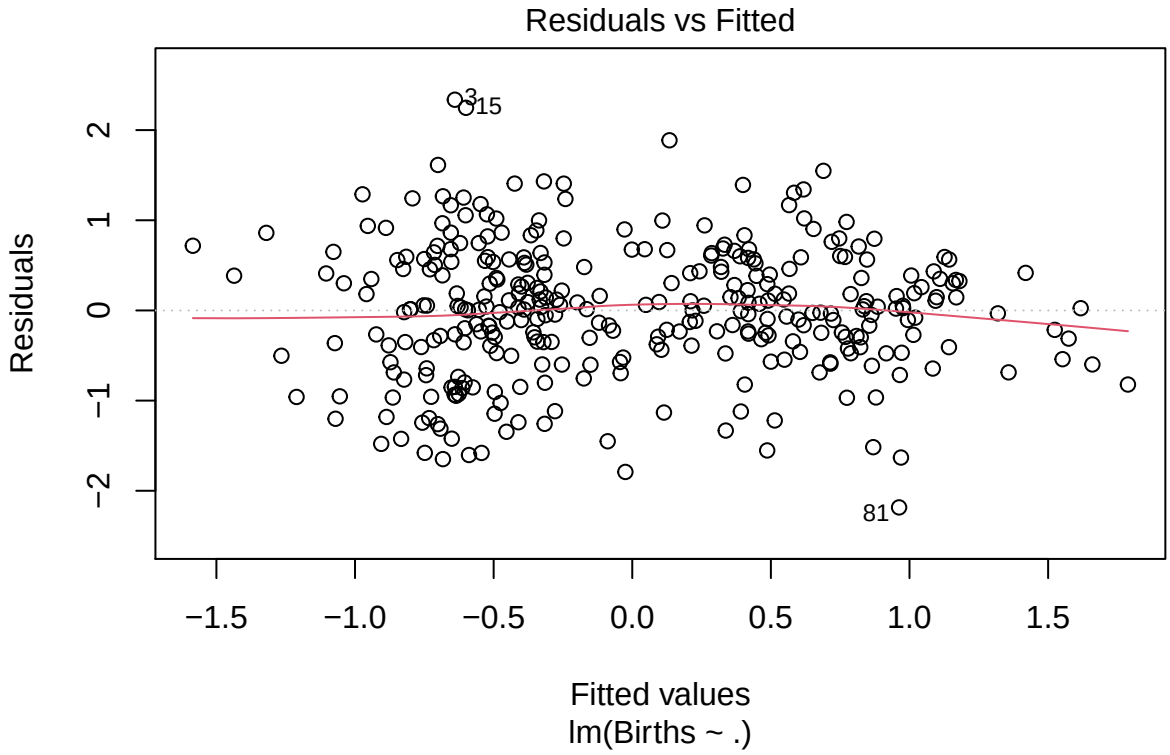

Scale−Location
289
316
328
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
−1
0
1
2
3
Fitted values
lm(Deaths ~ .)

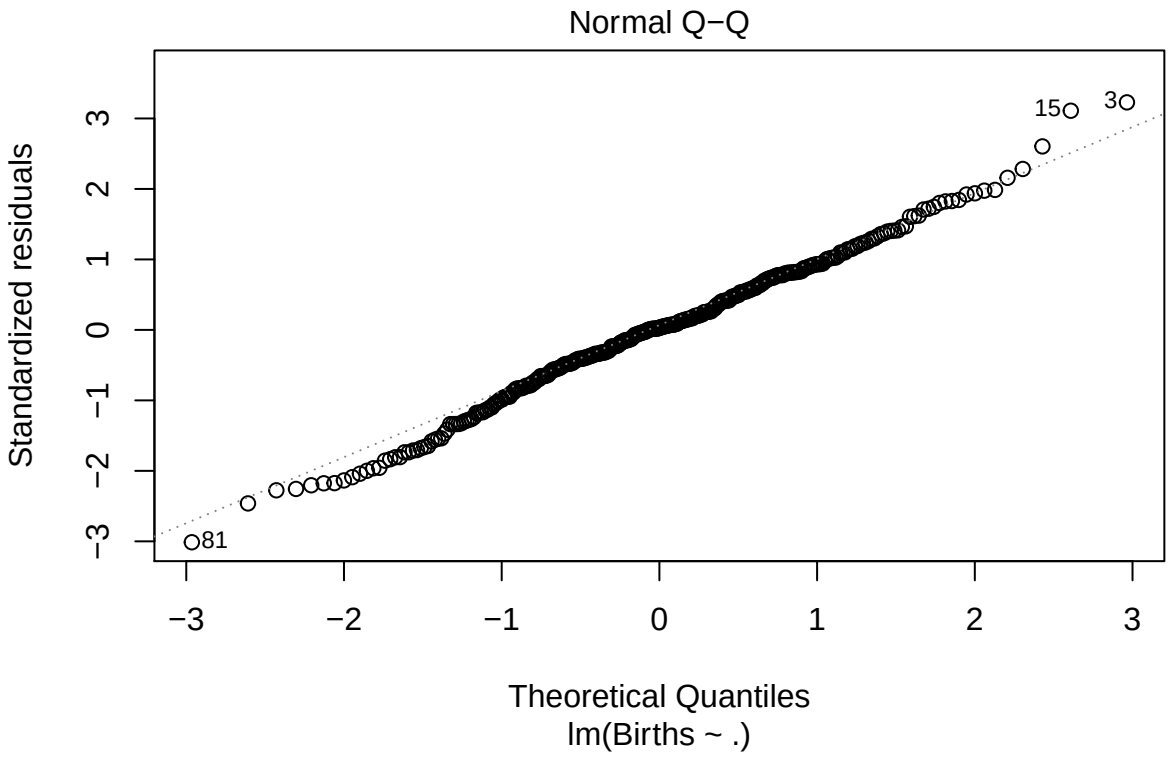


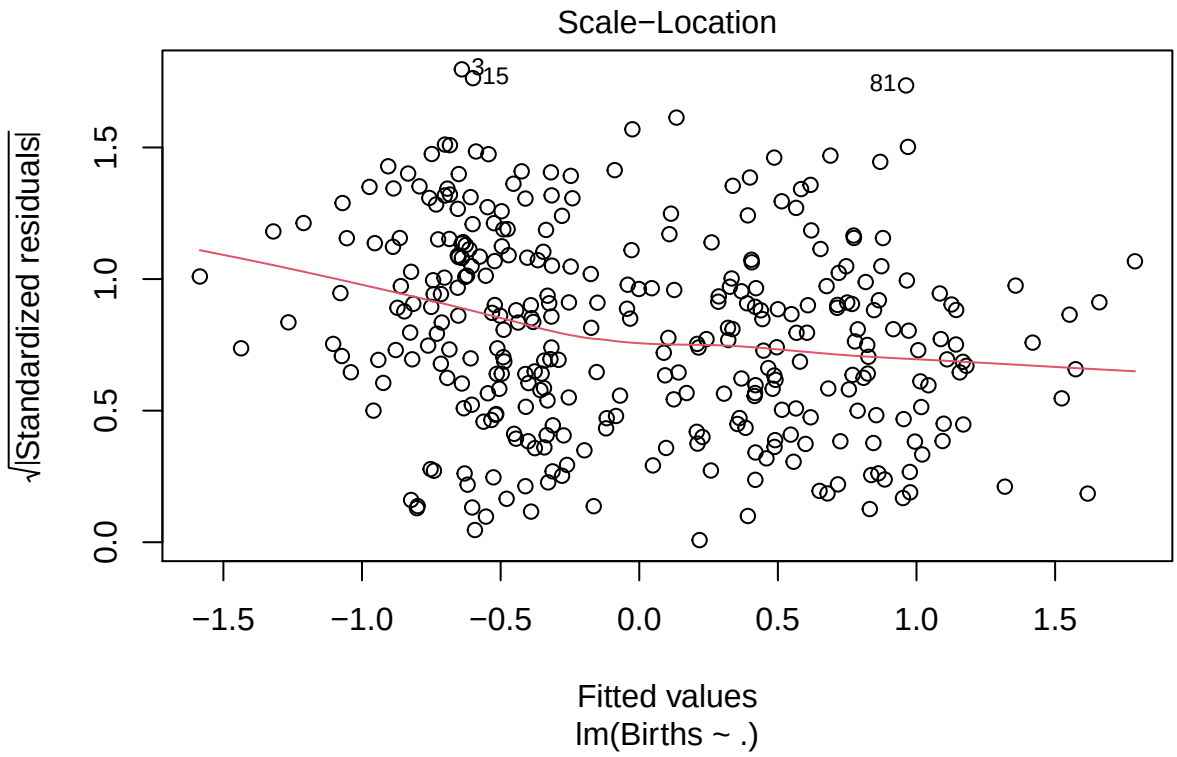

```
MLR_3 <- lm(Births ~ ., data = Final_dataset)

summary(MLR_3)
```

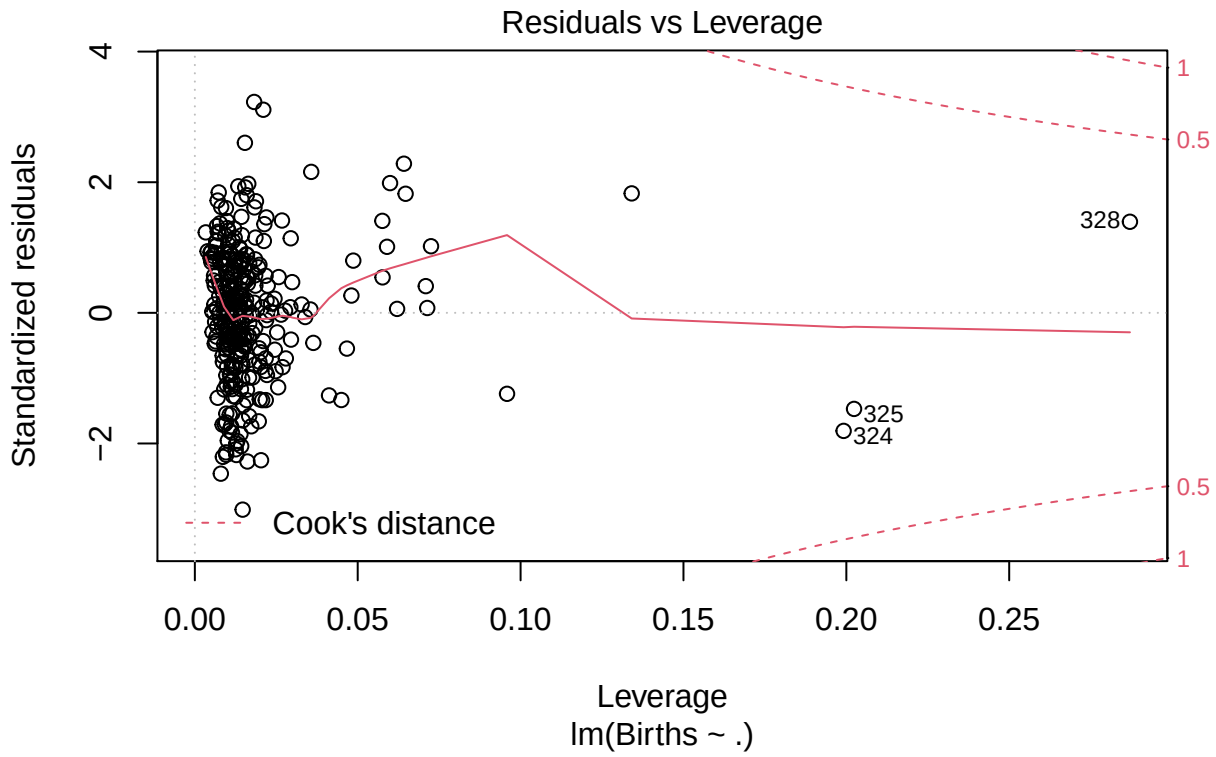
```
##
## Call:
## lm(formula = Births ~ ., data = Final_dataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.18449 -0.40753  0.02254  0.50720  2.33685
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.696e-01  3.979e-01   0.678    0.499
## Date        -1.957e-05  2.874e-05  -0.681    0.496
## Marriages    6.109e-01  5.160e-02  11.838  < 2e-16 ***
## Deaths      -2.096e-02  8.786e-02  -0.239    0.812
## Stillbirths  3.468e-01  5.903e-02   5.876 1.04e-08 ***
## Covid       -2.939e-02  4.589e-02  -0.641    0.522
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7305 on 324 degrees of freedom
## Multiple R-squared:  0.4745, Adjusted R-squared:  0.4664
## F-statistic: 58.52 on 5 and 324 DF,  p-value: < 2.2e-16
```

```
plot(MLR_3)
```


Residuals vs Fitted
Residuals
Fitted values
lm(Births ~ .)
3
15
81
−2
−1
0
1
2
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Births ~ .)
15
3
81
−3
−2
−1
0
1
2
3

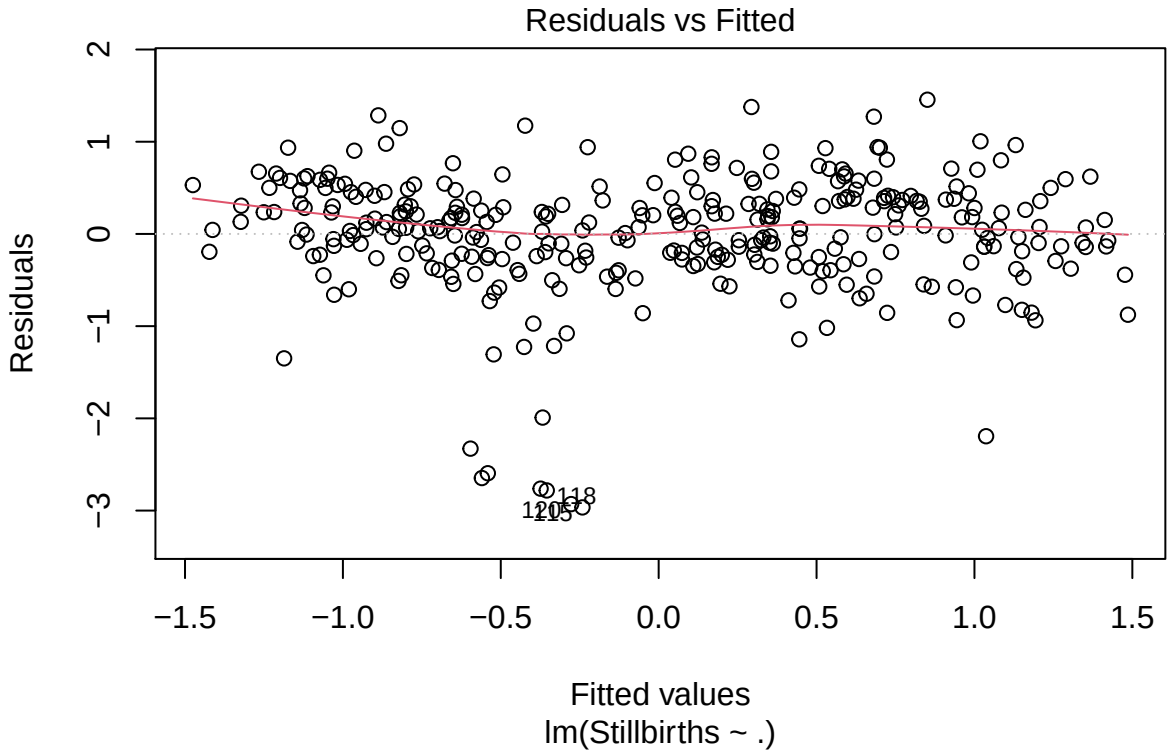

Scale−Location
3
15
81
√|Standardized residuals|
0.0
0.5
1.0
1.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Fitted values
lm(Births ~ .)

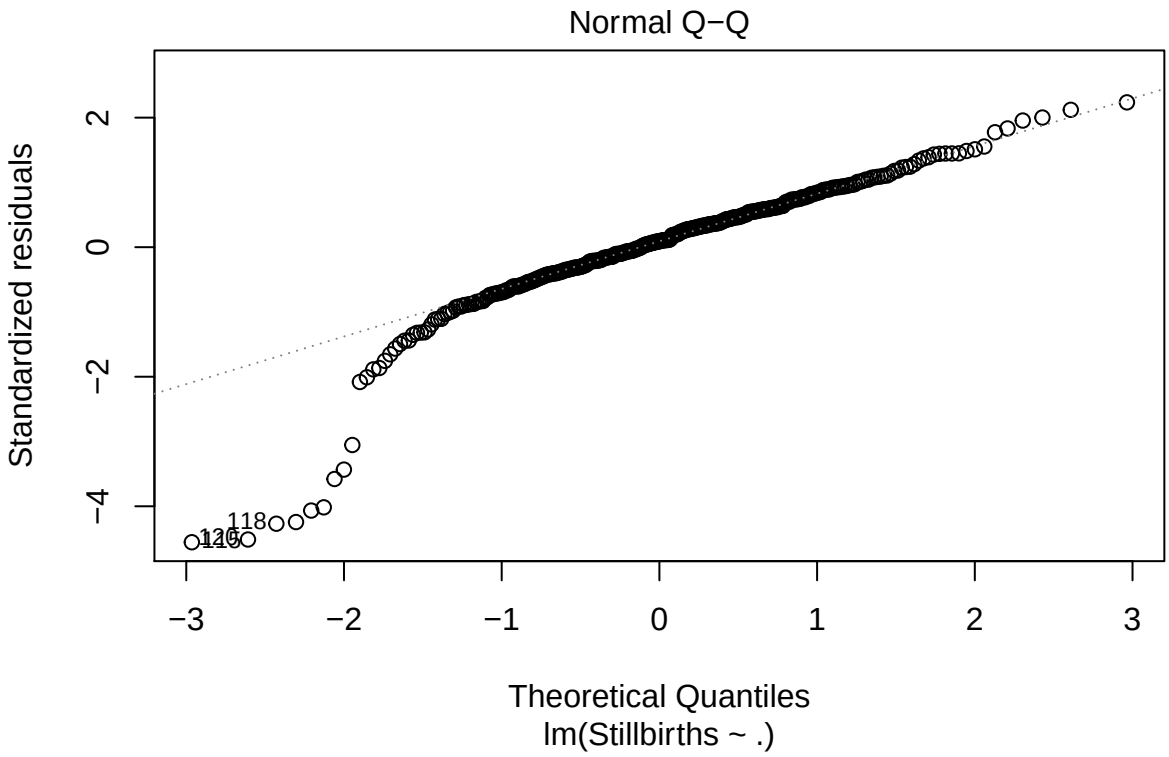


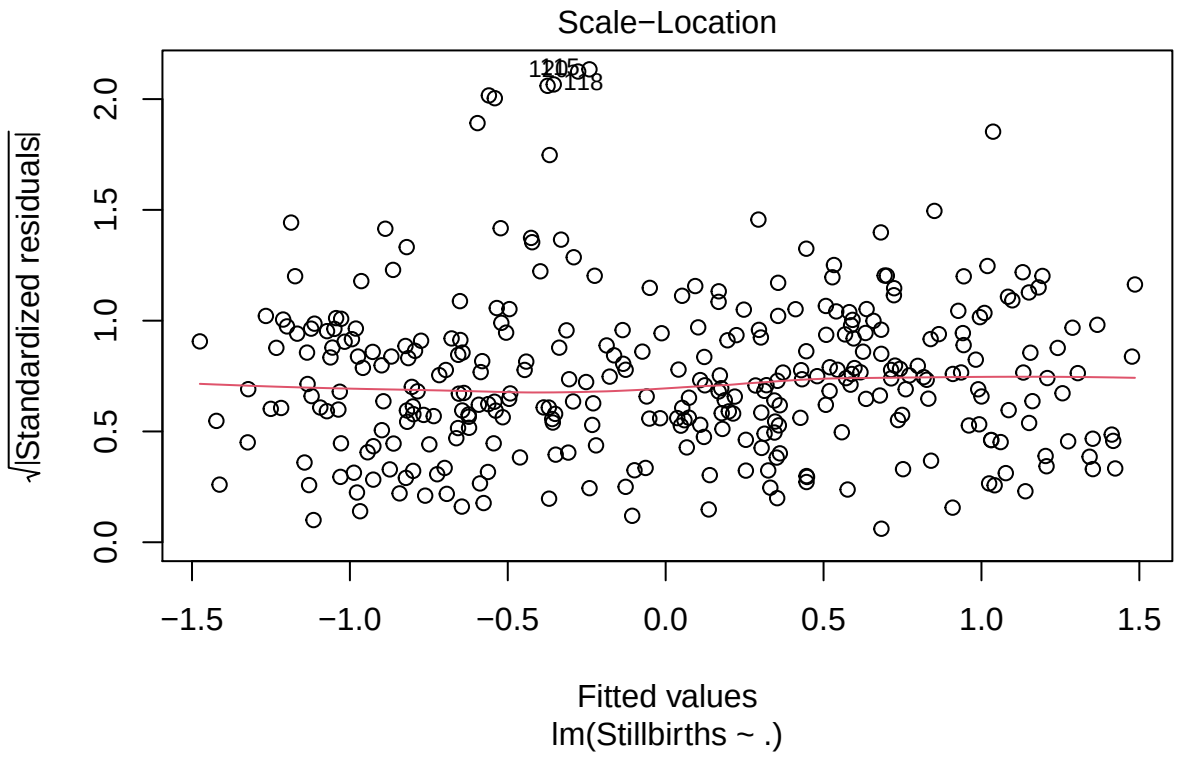

```
MLR_4 <- lm(Stillbirths ~ ., data = Final_dataset)

summary(MLR_4)
```

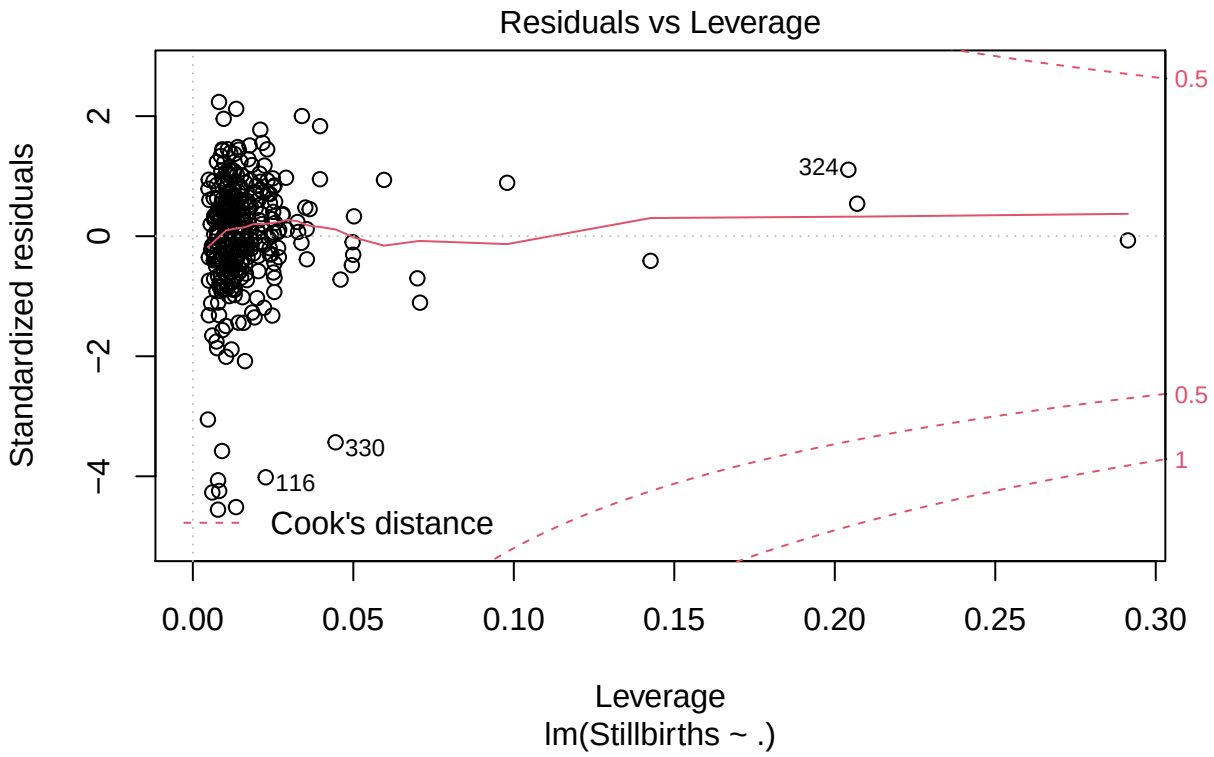
```
##
## Call:
## lm(formula = Stillbirths ~ ., data = Final_dataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.96543 -0.26250  0.05965  0.38107  1.45550
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.652e+00  3.244e-01  -8.175 6.74e-15 ***
## Date         1.925e-04  2.341e-05   8.226 4.75e-15 ***
## Births       2.777e-01  4.725e-02   5.876 1.04e-08 ***
## Marriages   -9.123e-02  5.503e-02  -1.658   0.0983 .
## Deaths       1.996e-01  7.784e-02   2.564   0.0108 *
## Covid       -1.987e-01  3.958e-02  -5.021 8.49e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6536 on 324 degrees of freedom
## Multiple R-squared:  0.5793, Adjusted R-squared:  0.5728
## F-statistic: 89.24 on 5 and 324 DF,  p-value: < 2.2e-16
```

```
plot(MLR_4)
```


Residuals vs Fitted
Residuals
Fitted values
lm(Stillbirths ~ .)
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2
1
0
−1
−2
−3
118


Normal Q−Q
Standardized residuals
2
0
−2
−4
118
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(Stillbirths ~ .)

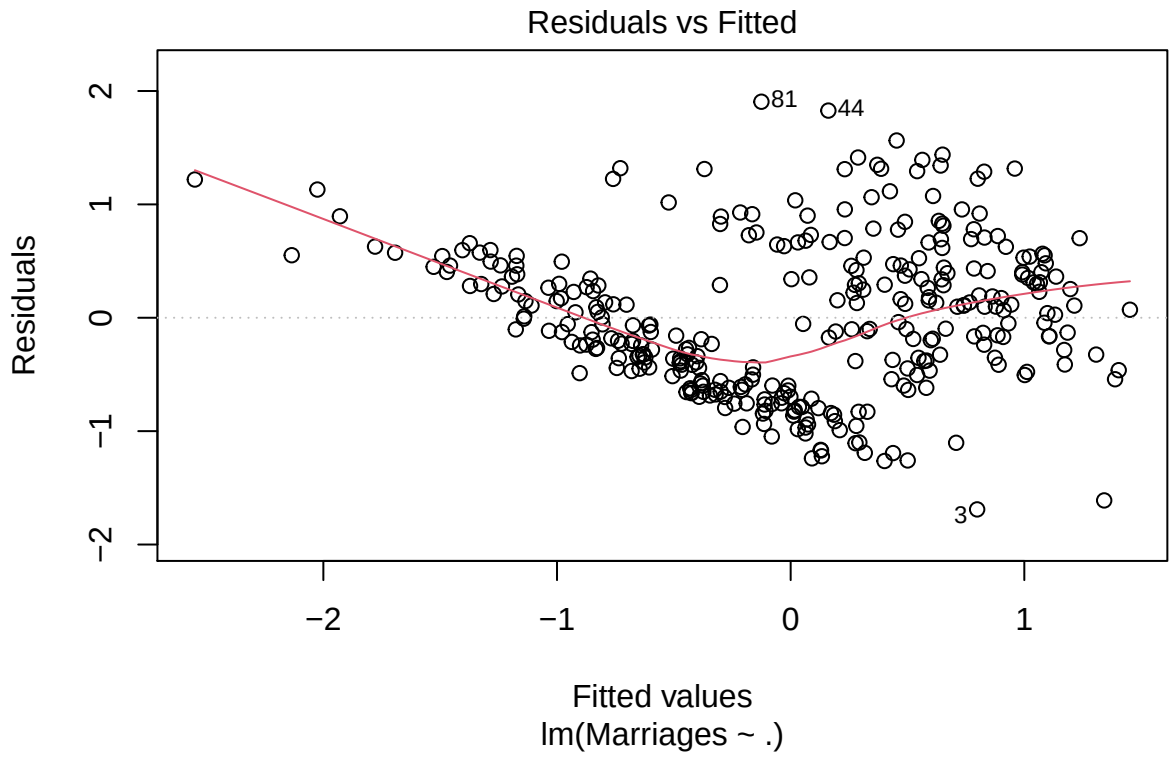

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Fitted values
lm(Stillbirths ~ .)

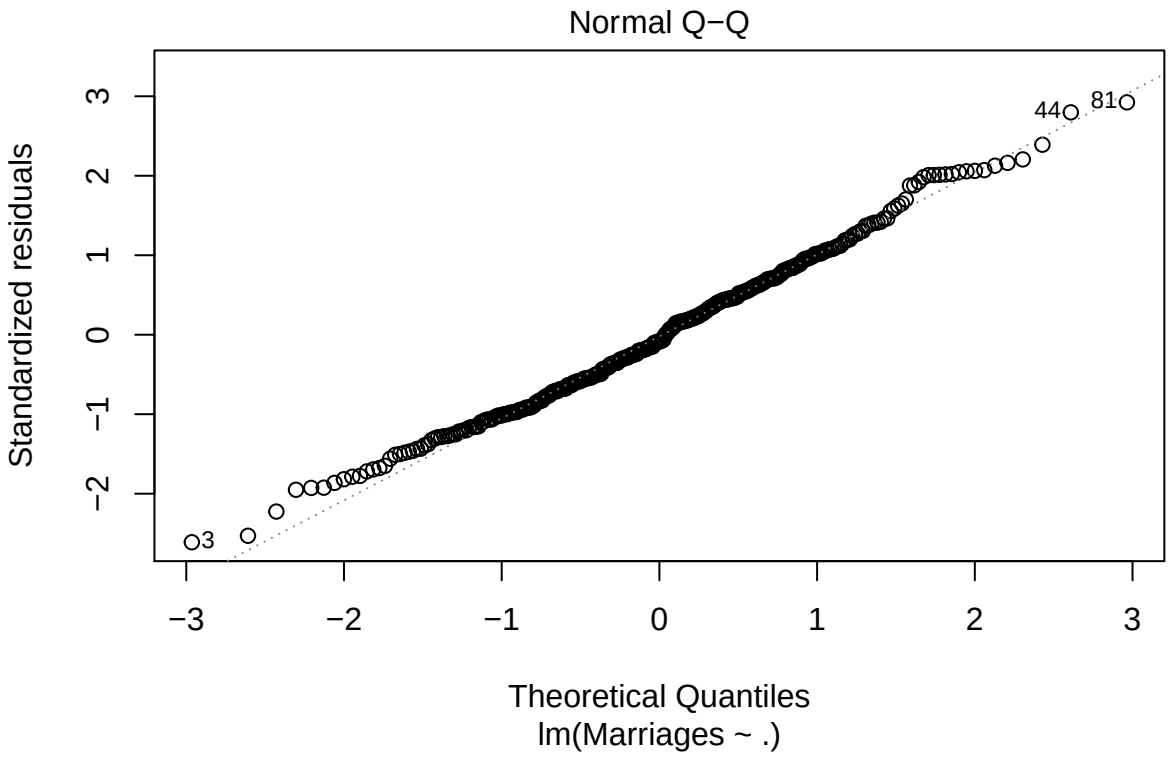


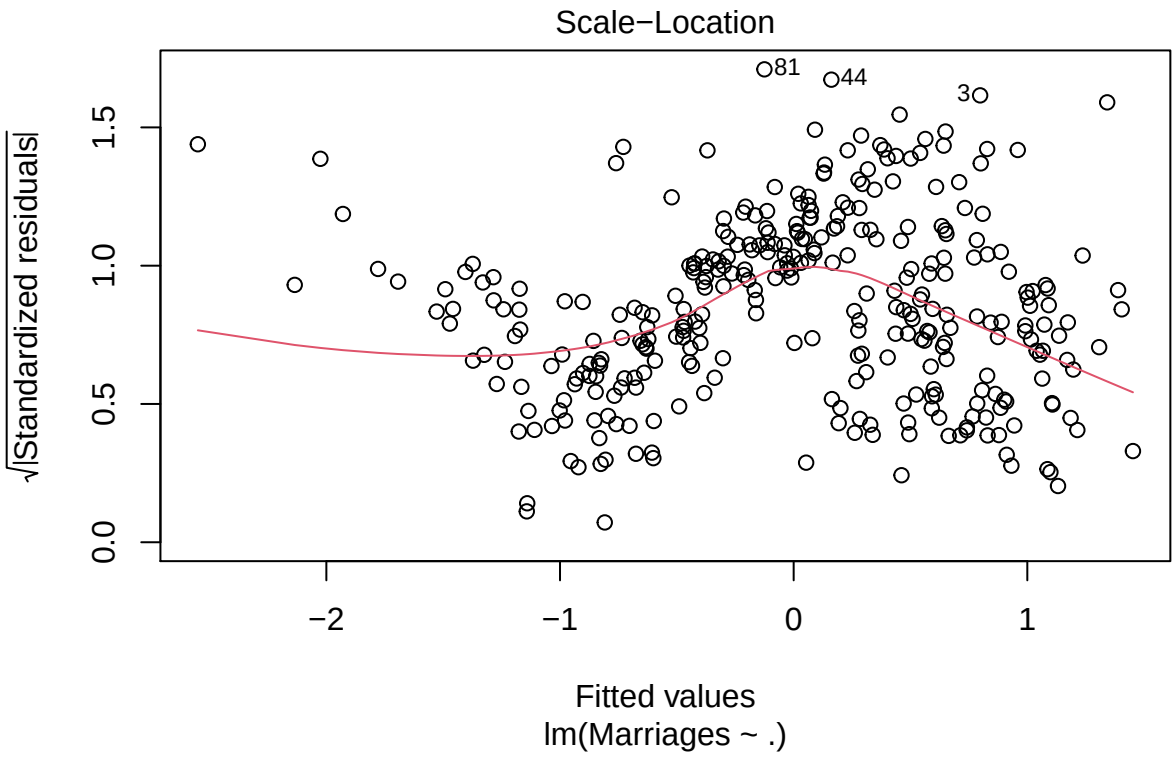

```
MLR_5 <- lm(Marriages ~ ., data = Final_dataset)

summary(MLR_5)
```

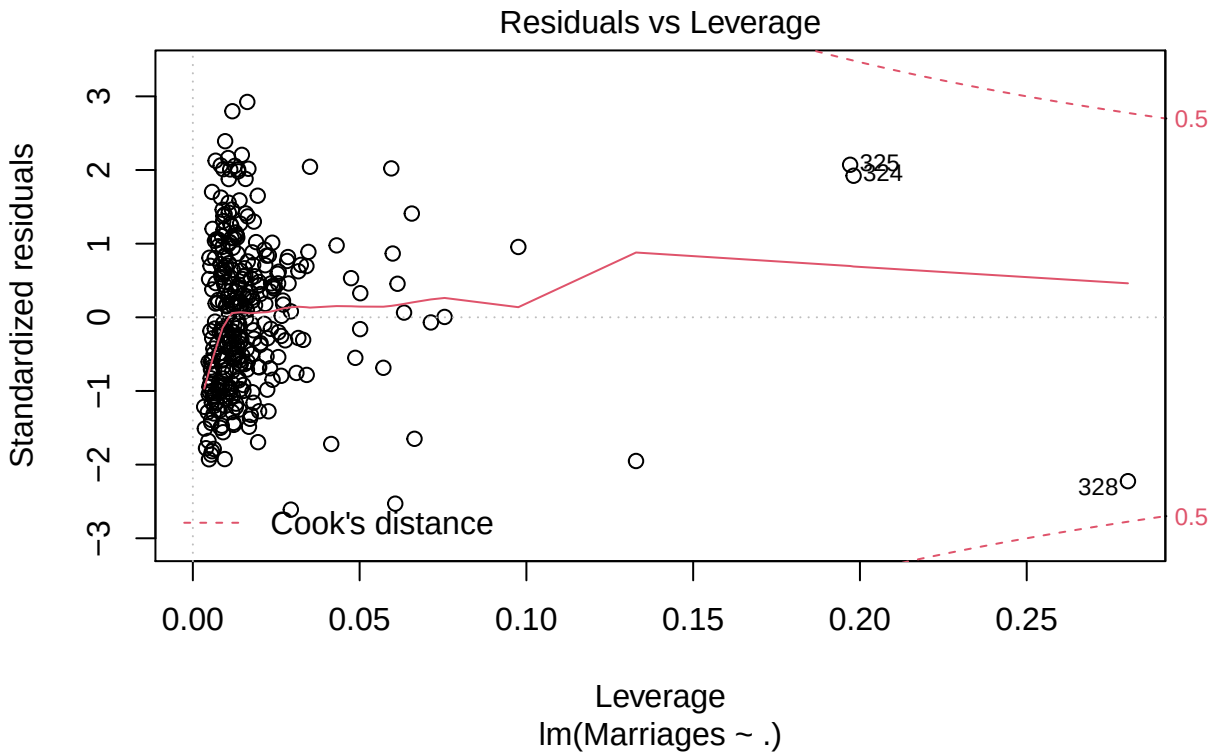
```
##
## Call:
## lm(formula = Marriages ~ ., data = Final_dataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.69025 -0.46904 -0.05514  0.44035  1.90542
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.206e+00  3.365e-01  -6.556 2.18e-10 ***
## Date         1.602e-04  2.429e-05   6.594 1.74e-10 ***
## Births       4.943e-01  4.175e-02  11.838  < 2e-16 ***
## Deaths      -7.160e-01  6.830e-02 -10.483  < 2e-16 ***
## Stillbirths -9.221e-02  5.562e-02  -1.658   0.0983 .
## Covid        6.459e-02  4.115e-02   1.570   0.1175
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6571 on 324 degrees of freedom
## Multiple R-squared:  0.5748, Adjusted R-squared:  0.5682
## F-statistic:  87.6 on 5 and 324 DF,  p-value: < 2.2e-16
```

```
plot(MLR_5)
```


Residuals vs Fitted
Residuals
Fitted values
lm(Marriages ~ .)
81
44
3
−2
−1
0
1
2


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Marriages ~ .)
3
44
81
−3
−2
−1
0
1
2
3

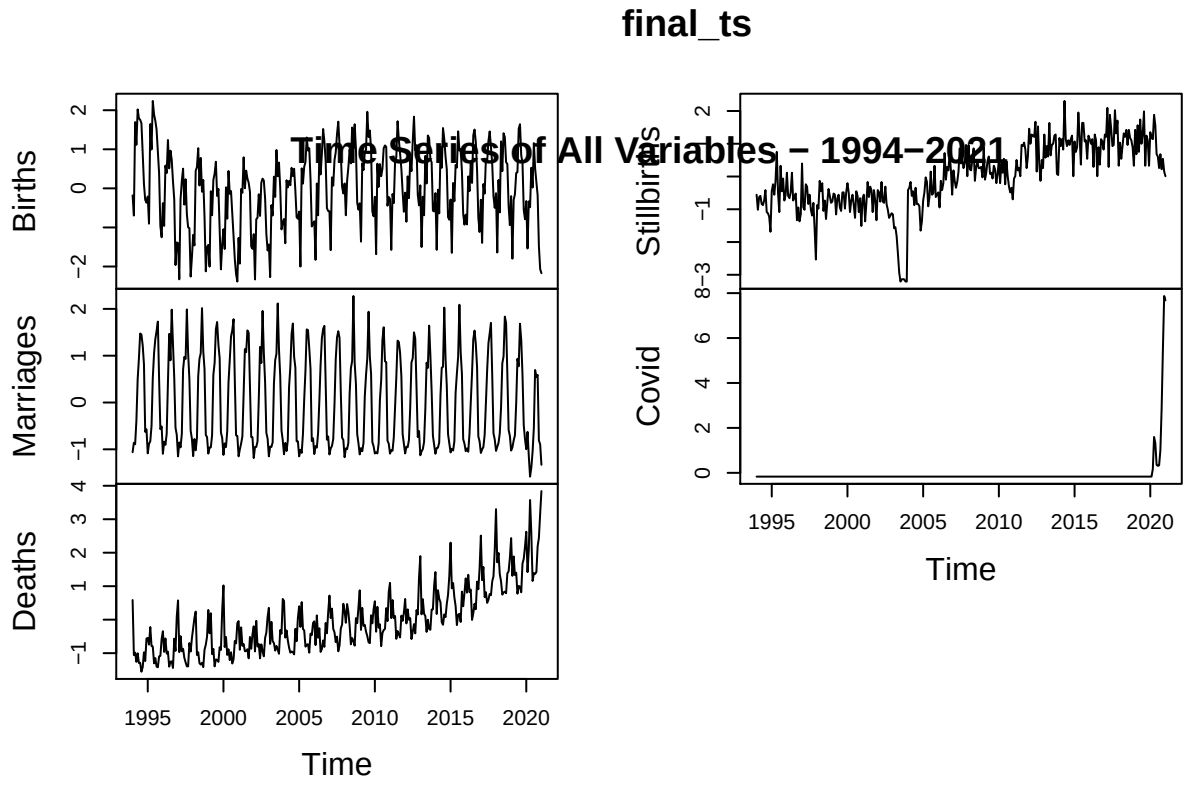

Scale−Location
81
44
3
√|Standardized residuals|
0.0
0.5
1.0
1.5
−2
−1
0
1
Fitted values
lm(Marriages ~ .)



convert Final data set to a time series . Date selection is from 1994-2021 When selecting the starting point of 1964, which is when the data starts, there appears to be a moment in which the data begins to repeat itself. The original time frame selected was too long, and needed to be reduced.

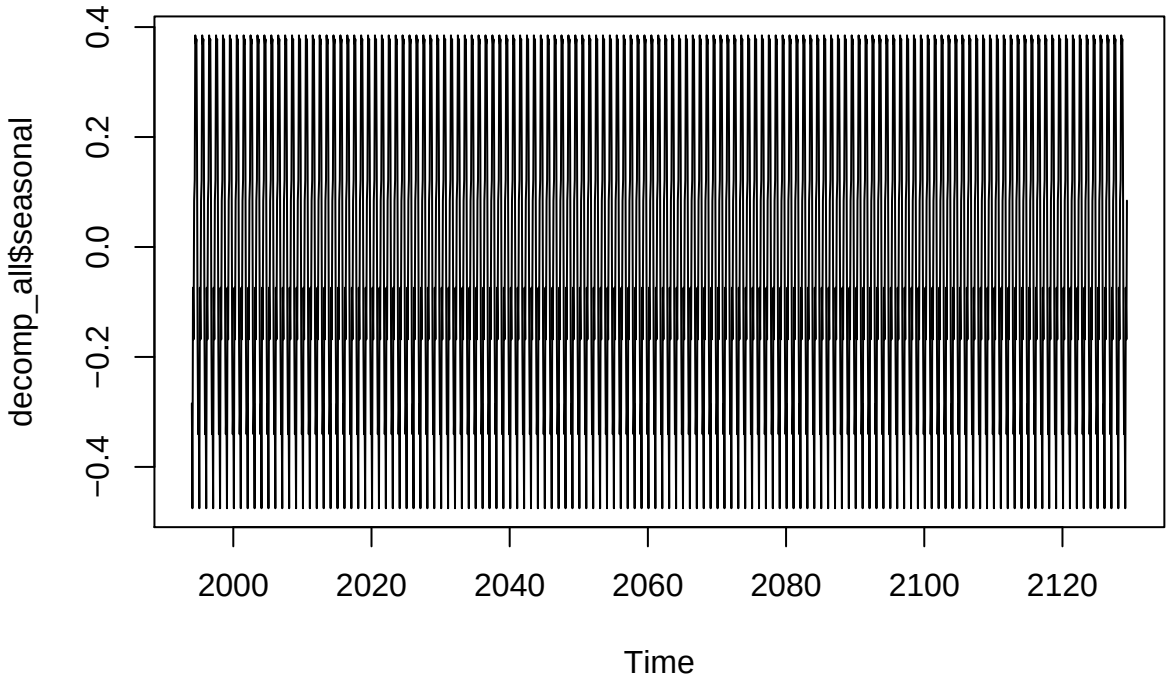
```
final_ts <- ts(Final_dataset[2:6], start = 1994, end = 2021, frequency = 12)
```

```
plot(final_ts)
title(main = "Time Series of All Variables - 1994-2021", line = -1)
```

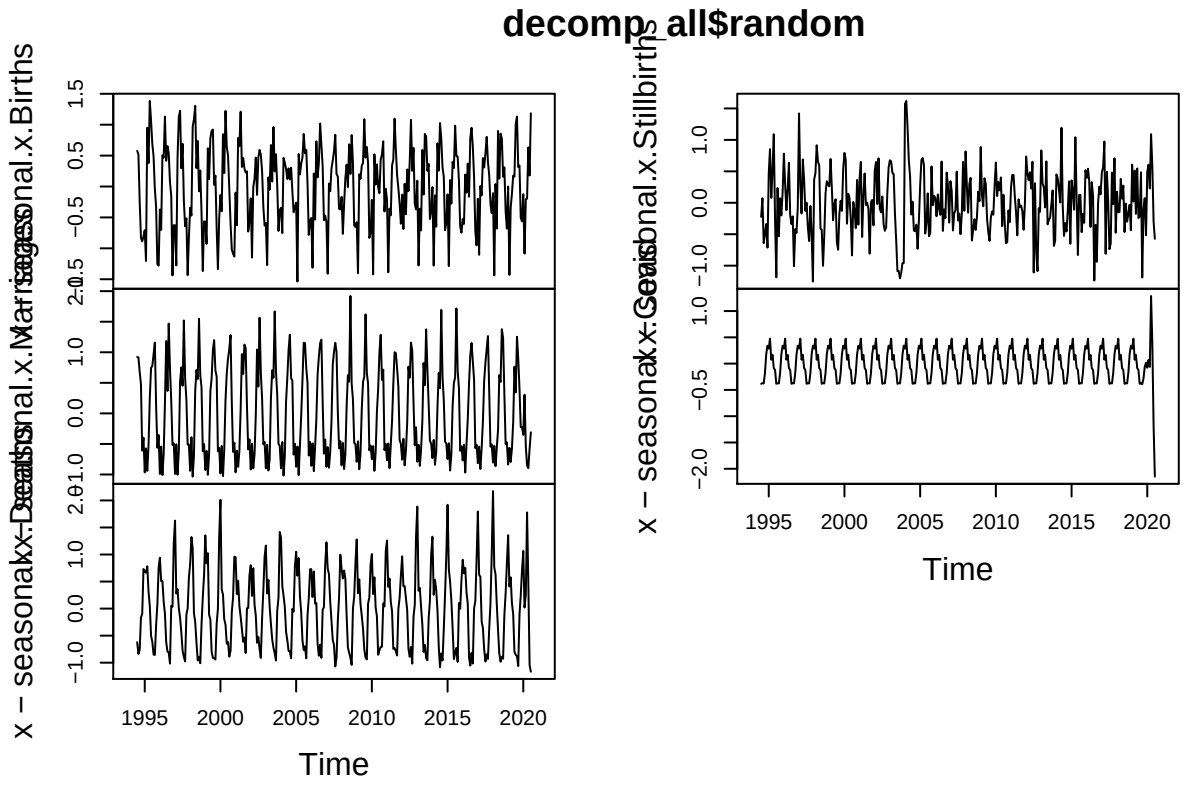


View decomposition of thetime series

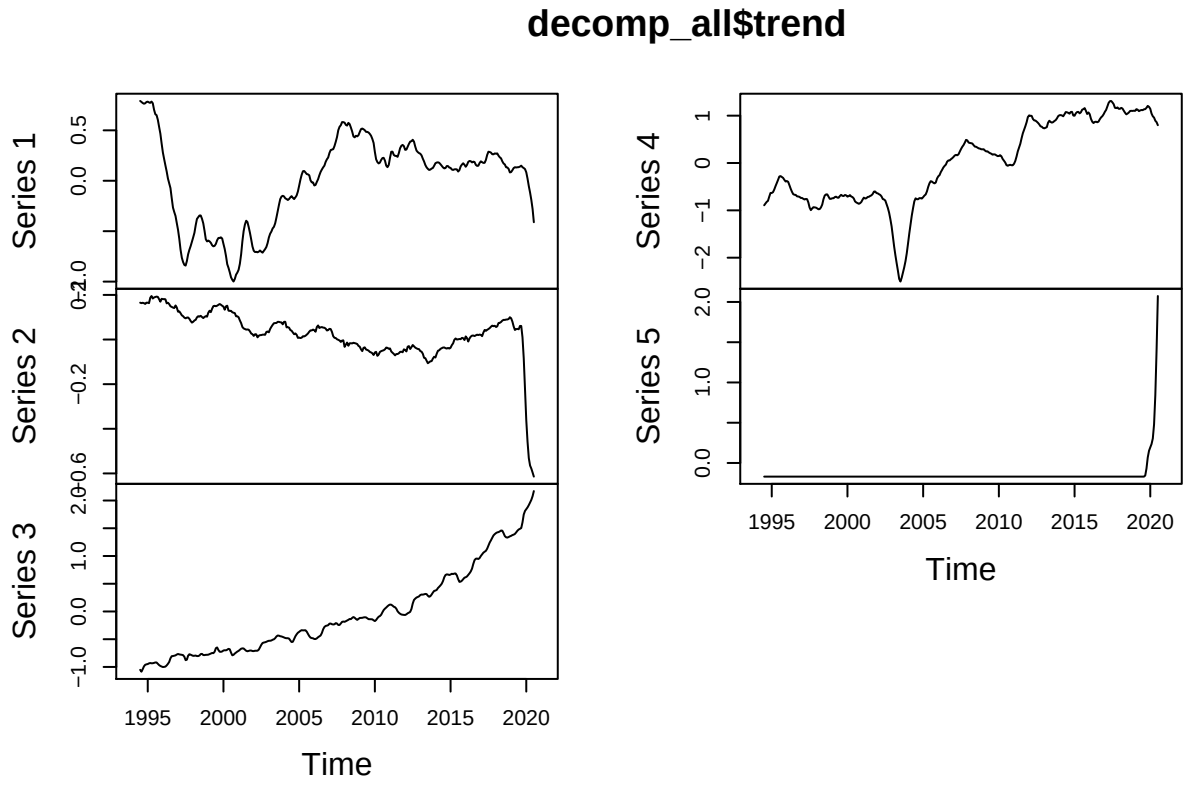
```
decomp_all <- decompose(final_ts)
plot(decomp_all$seasonal)
```




```
plot(decomp_all$random)
```




```
plot(decomp_all$trend)
```

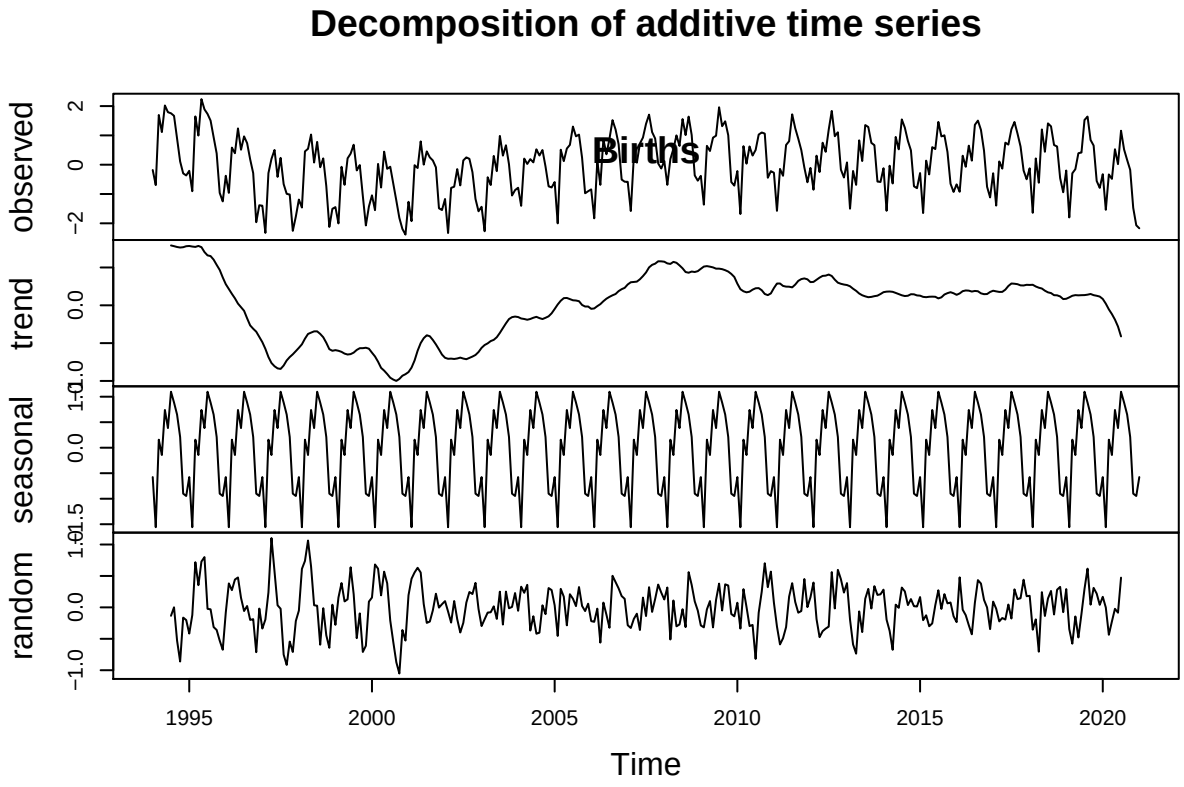


decompose time series data for each variable

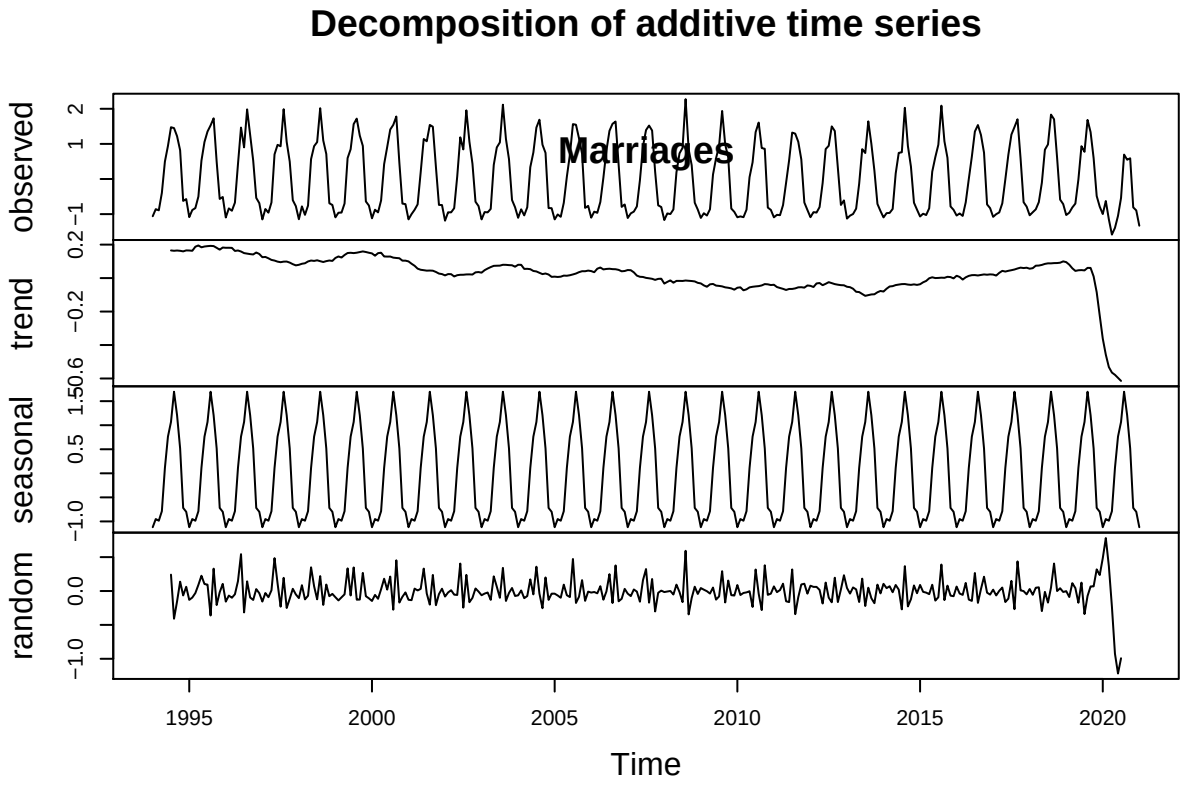
```
Births_ts_decomp <- decompose(final_ts[,1])
Marriages_ts_decomp <- decompose(final_ts[,2])
Deaths_ts_decomp <- decompose(final_ts[,3])
Stillbirths_ts_decomp <- decompose(final_ts[,4])
Covid_ts_decomp <- decompose(final_ts[,5])
```

plot decomposition of the time series for each vairable

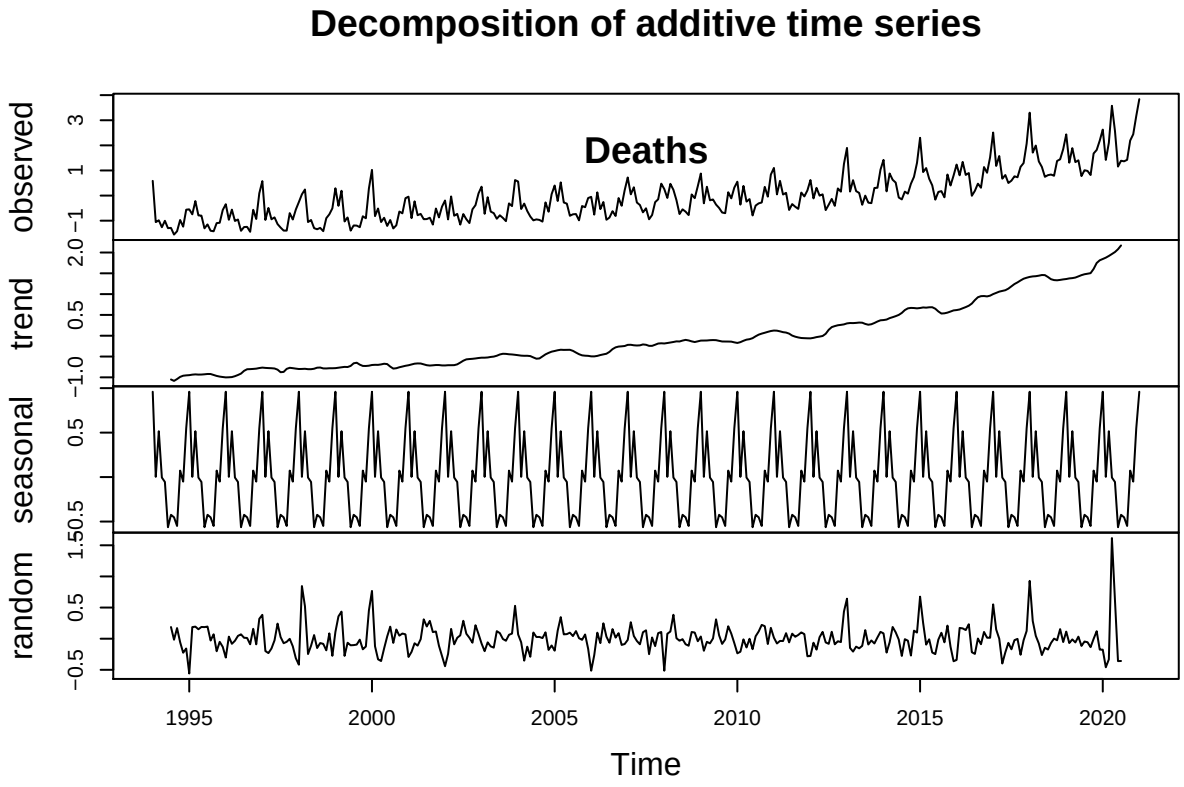
```
plot(Births_ts_decomp)
title("Births", line = -1)
```



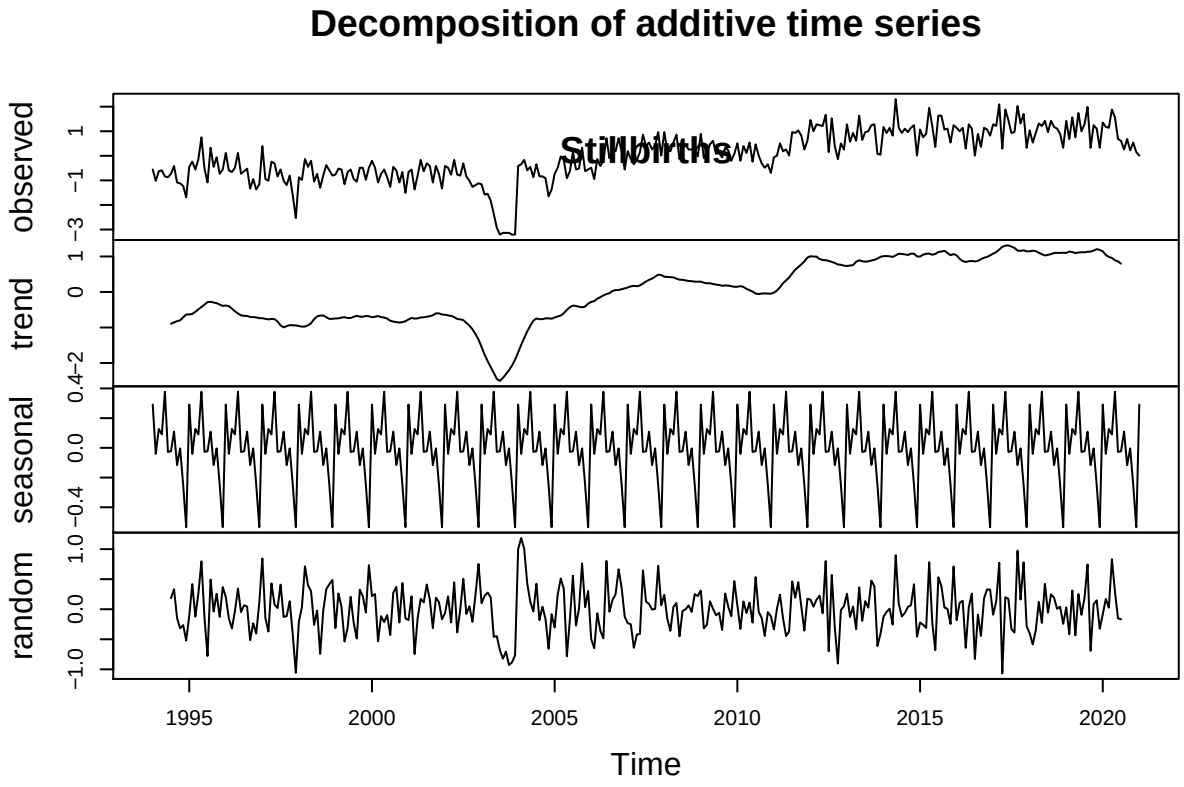

```
plot(Marriages_ts_decomp)
title("Marriages", line = -1)
```



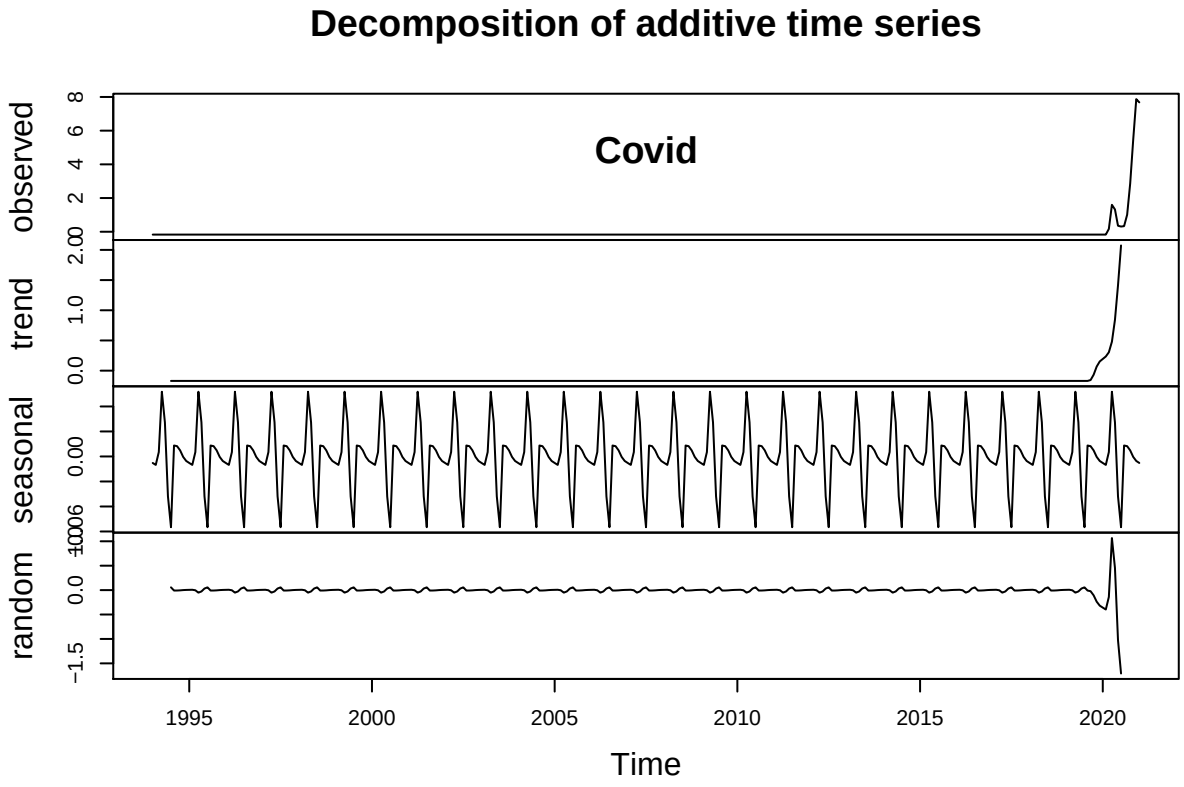

```
plot(Deaths_ts_decomp)
title("Deaths", line = -1)
```




```
plot(Stillbirths_ts_decomp)
title("Stillbirths", line = -1)
```




```
plot(Covid_ts_decomp)
title("Covid", line = -1)
```


Decomposition of additive time series
Covid
observed
trend
seasonal
random
Time
1995
2000
2005
2010
2015
2020