Nonparametric Analysis of US Dairy Production and Consumption

Robustness

Teo Bucci* Filippo Cipriani† Gabriele Corbo‡ Andrea Puricelli§

2023-02-17

Contents

1 Load libraries and data

```
library(robustbase)
library(splines)
library(mgcv)
```

```
data_path = file.path('data_updated_2021')
output_path = file.path('output')

data_infl <-
    read.table(
        file.path(data_path, 'production_facts_inflated.csv'),
        header = T,
        sep = ';'
    )
```

*teo.bucci@mail.polimi.it
†filippo.cipriani@mail.polimi.it
‡gabriele.corbo@mail.polimi.it
§andrea3.puricelli@mail.polimi.it

2 Robust regression

Define the formula for the regression.

```
formula = avg_price_milk ~ avg_milk_cow_number + milk_per_cow +
  milk_cow_cost_per_animal + milk_volume_to_buy_cow_in_lbs
```

Perform the regression.

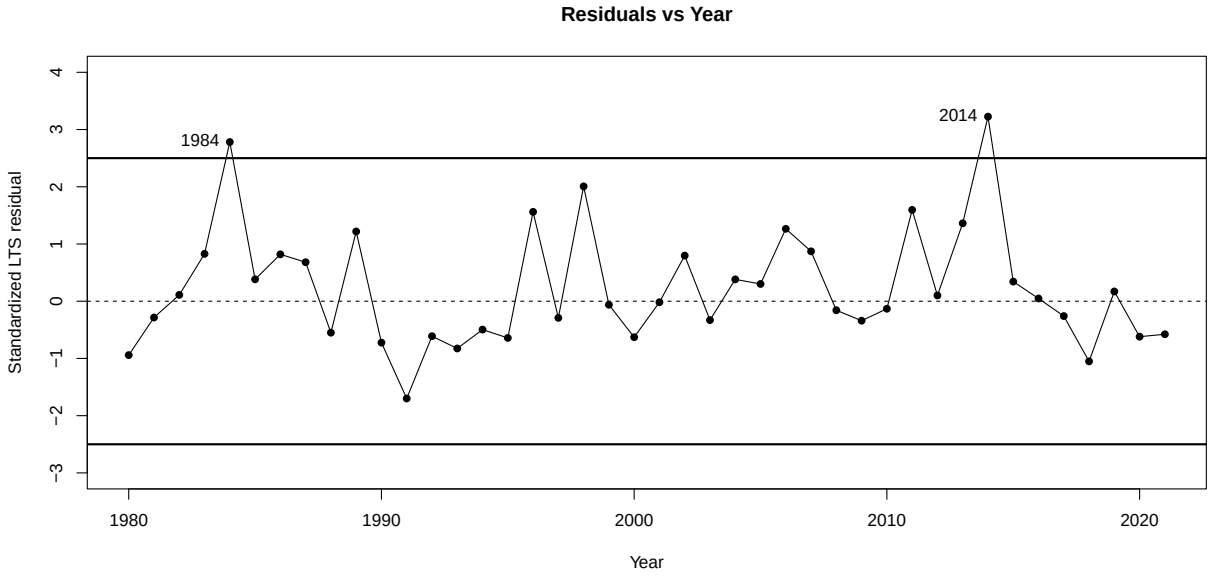
```
fit_lts <- ltsReg(formula ,
                  alpha = .75,
                  mcd = TRUE,
                  data = data_infl)
```

3 Plot diagnostic

```
thresh = sqrt(qchisq(0.975, ncol(data_infl)))
```

3.1 Residual versus year (index)

```
# plot(fit_lts, which="rindex")
plot(
    data_infl$year,
    fit_lts$resid,
    ylim = c(-3, 4),
    main = "Residuals vs Year",
    xlab = "Year",
    ylab = "Standardized LTS residual",
    type = "l"
)
points(
    data_infl$year,
    fit_lts$resid,
    pch = 16
)
abline(h = c(-2.5, 2.5), lwd = 2)
abline(h = 0, lty = 2)
text(
    data_infl$year,
    fit_lts$resid,
    labels = ifelse(abs(fit_lts$resid) > 2.5, data_infl$year, ""),
    pos = 2
)
```



The overall outliers are

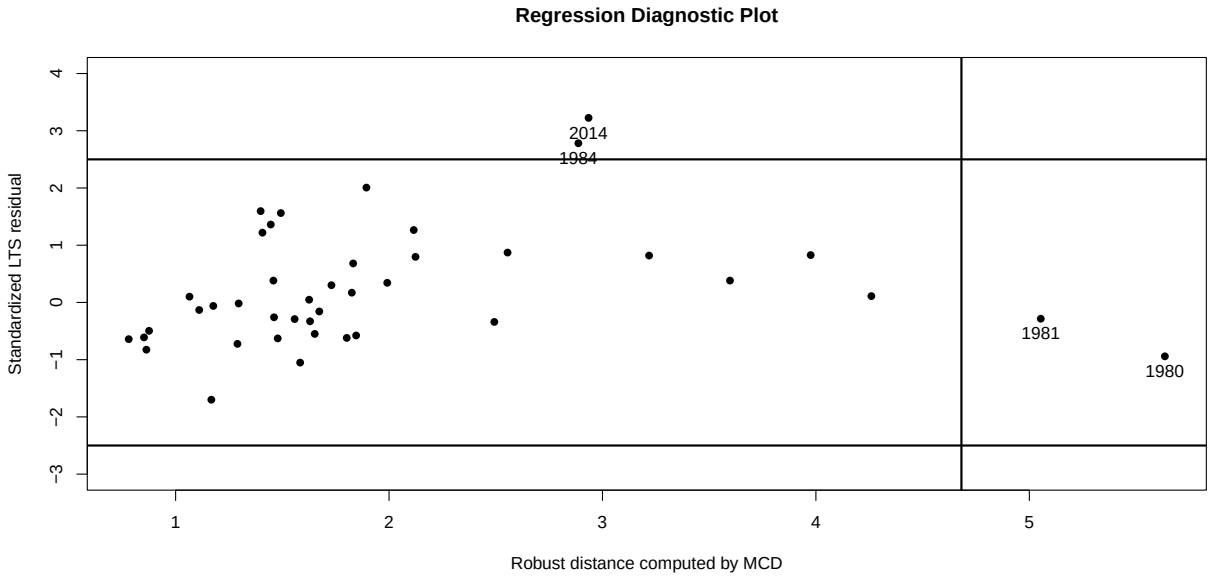
```
data_infl$year[which(abs(fit_lts$resid) > 2.5)]
```

```
## [1] 1984 2014
```

We can now proceed to classify them as *vertical outliers* or *bad leverages*.

3.2 Outlier map

```
# plot(fit_lts, which="rdiag")
plot(
    fit_lts$RD,
    fit_lts$resid,
    ylim = c(-3, 4),
    pch = 16,
    main = "Regression Diagnostic Plot",
    xlab = "Robust distance computed by MCD",
    ylab = "Standardized LTS residual"
)
abline(h = c(-2.5, 2.5), v = thresh, lwd = 2)
text(
    fit_lts$RD,
    fit_lts$resid,
    labels = ifelse(abs(fit_lts$resid) > 2.5 |
                        fit_lts$RD > thresh, data_infl$year, ""),
    pos = 1
)
```



The bad leverages are

```
data_infl$year[which(abs(fit_lts$resid) > 2.5 & fit_lts$RD > thresh)]
```

```
## integer(0)
```

The vertical outliers are

```
data_infl$year[which(abs(fit_lts$resid) > 2.5 & fit_lts$RD < thresh)]
```

```
## [1] 1984 2014
```