LOWESS for outlier detection

Outliers

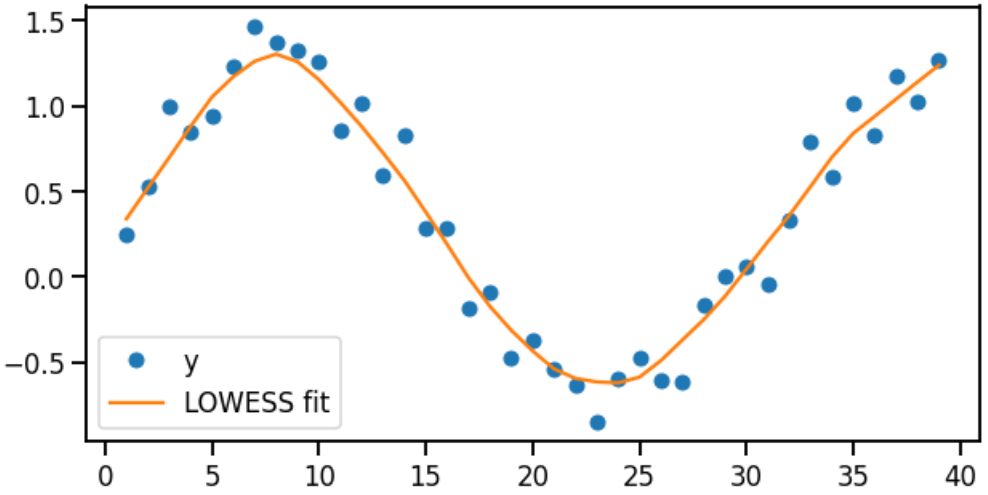
Contents

FIT TIME SERIES USING LOWESS

USE RESIDUALS TO IDENTIFY OUTLIERS

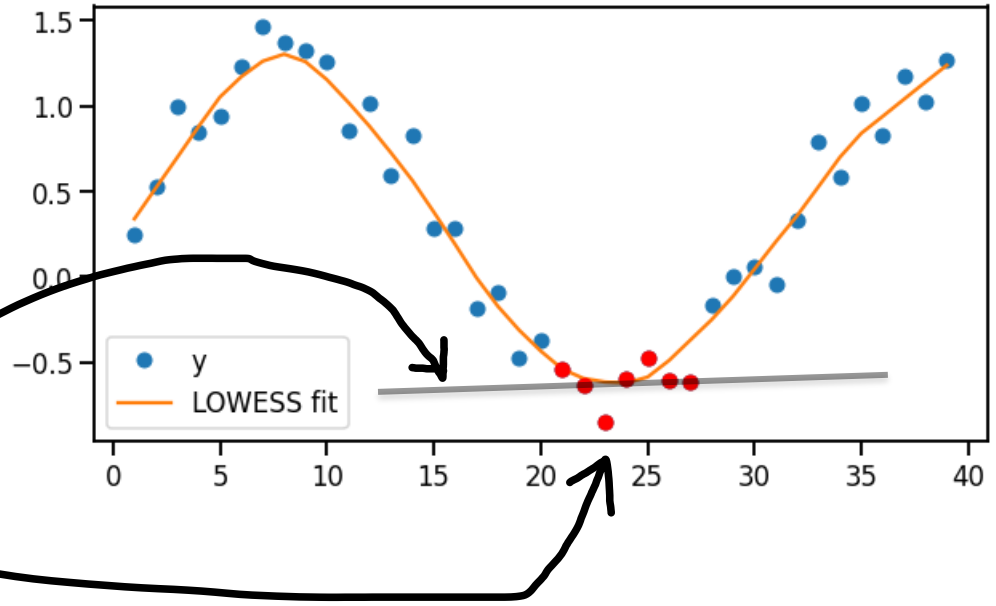
LOWESS recap

- **Lo**cally **We**ighted **S**catterplot **S**moothing
- Non-parametric smooth curve fitting
- The LOWESS curve at point (x,y) is obtained from a weighted linear regression built from a subset of data close to (x,y)
- Gives less weight to data further away from (x,y)



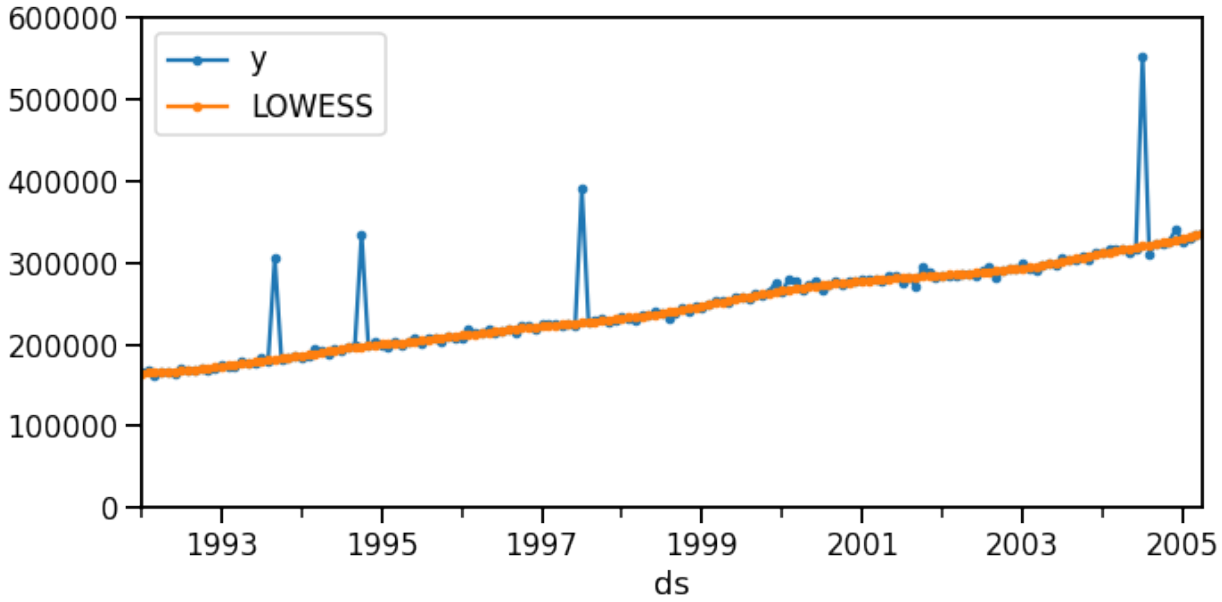
LOWESS recap

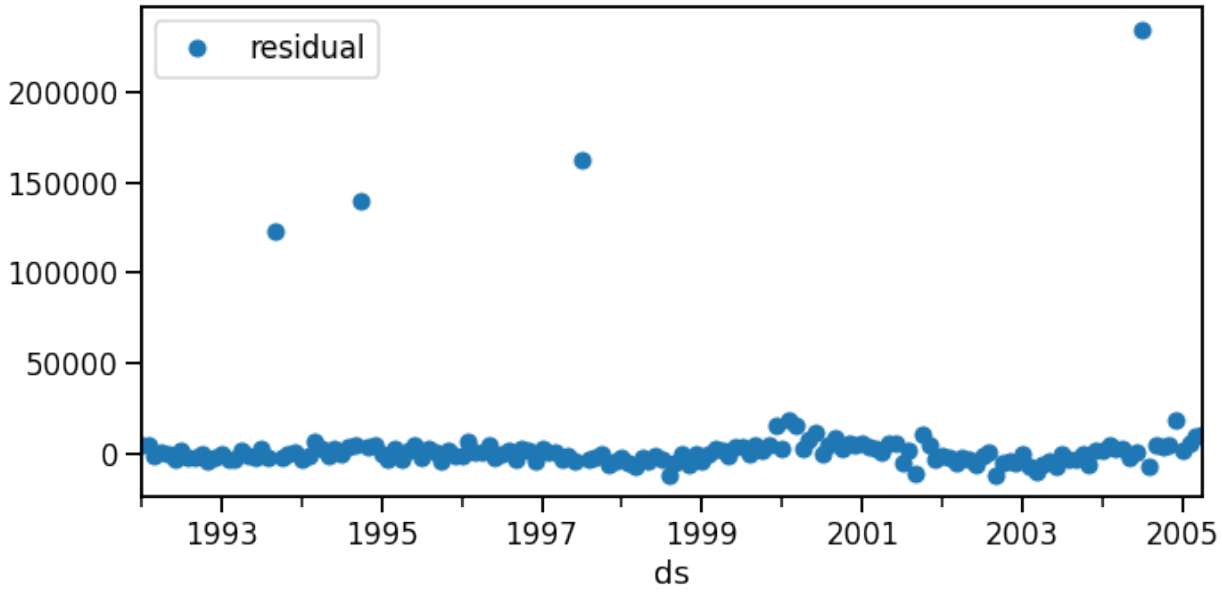
- **Lo**cally **We**ighted **S**catterplot **S**moothing
- Non-parametric smooth curve fitting
- The LOWESS curve at point (x,y) is obtained from a **weighted linear regression** built from **a subset of data** close to (x,y)
- Gives less weight to data further away from (x,y)



Consider residuals from fitting a LOWESS curve

$$e_t = y_t - \hat{y}_t$$
$$\hat{y}_t = LOWESS(\boldsymbol{y}, t)$$



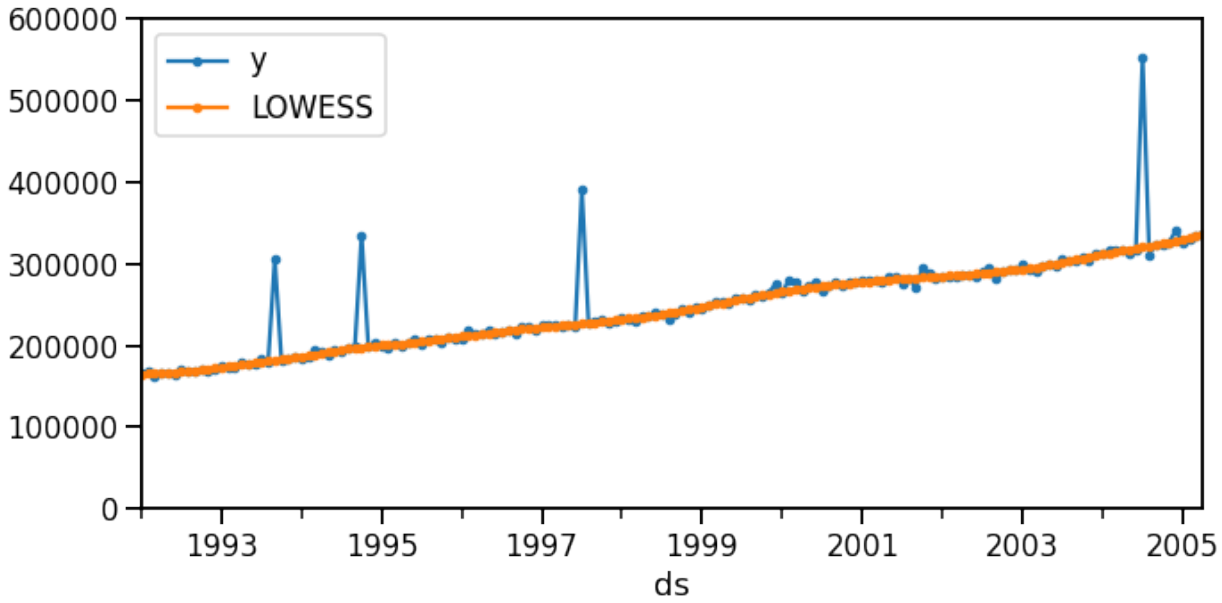


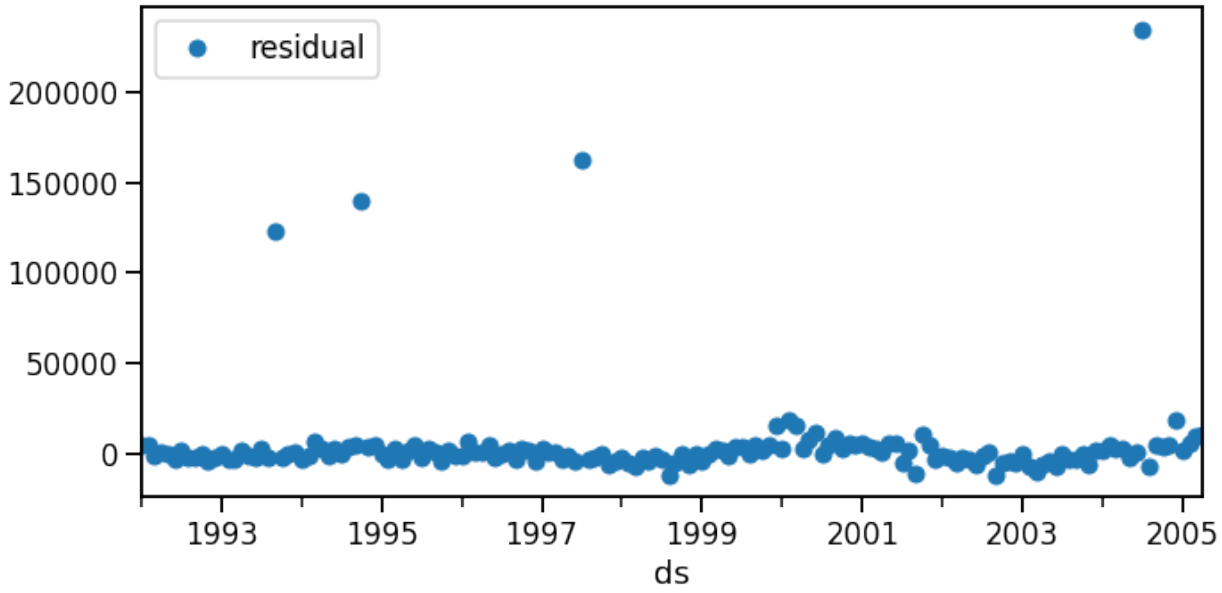
- The residuals look stationary

- Determine outliers using IQR:

$$e_t > \delta_{upper} = Q3 + \alpha \times IQR$$
$$e_t < \delta_{lower} = Q1 - \alpha \times IQR$$

- We set $\alpha = 3$ so that only more extreme outliers are detected



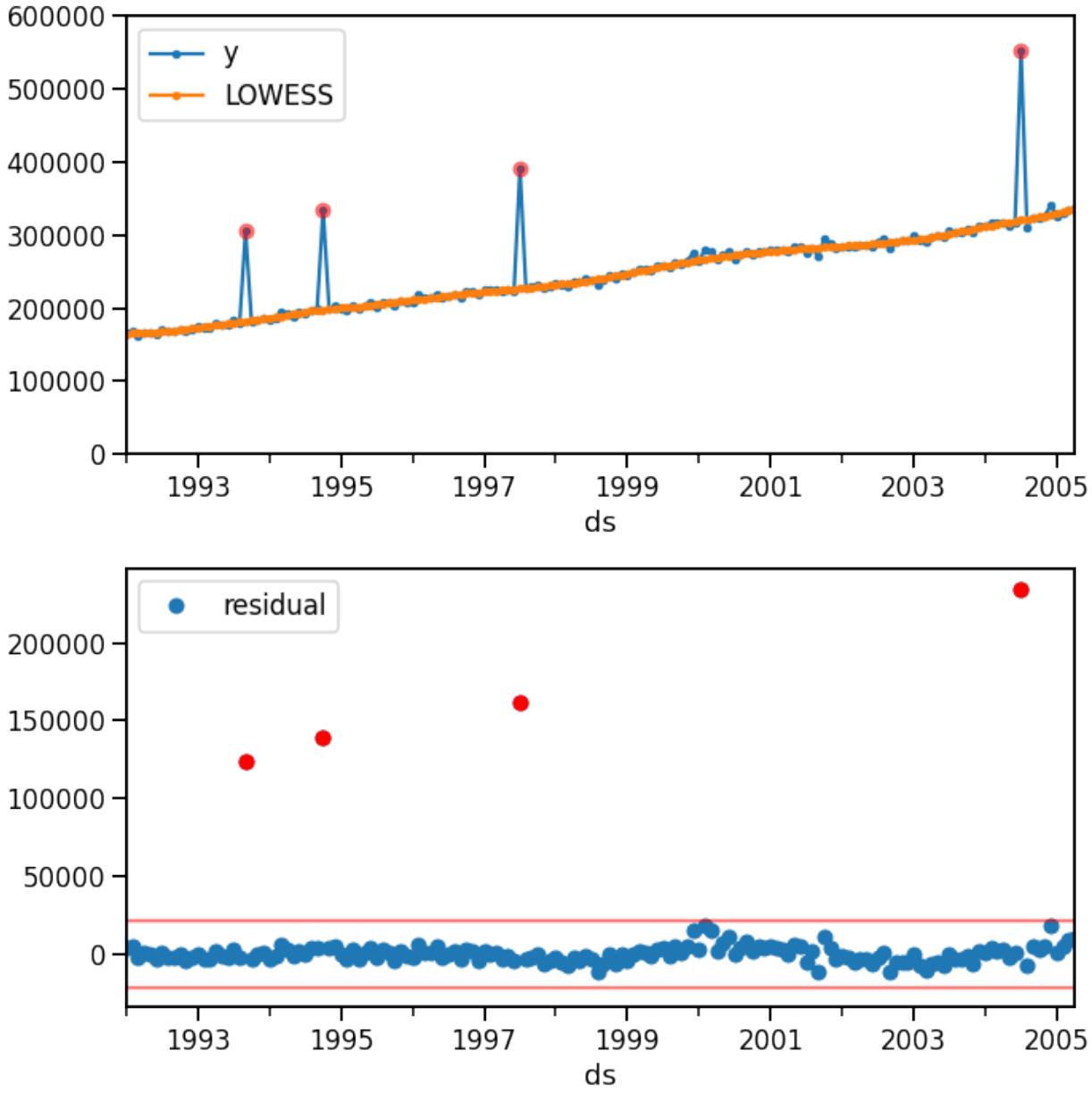


- The residuals look stationary

- Determine outliers using IQR:

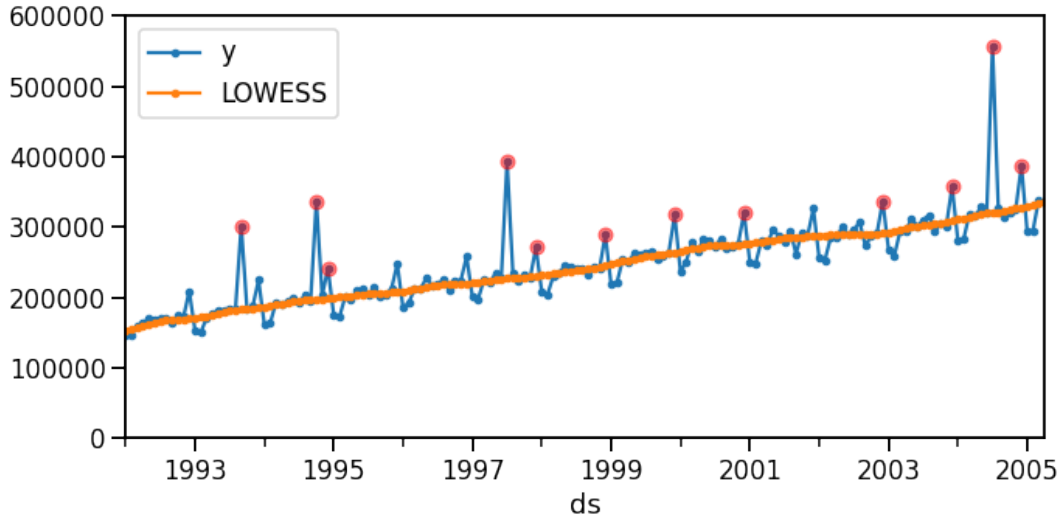
$$e_t > \delta_{upper} = Q3 + \alpha \times IQR$$
$$e_t < \delta_{lower} = Q1 - \alpha \times IQR$$

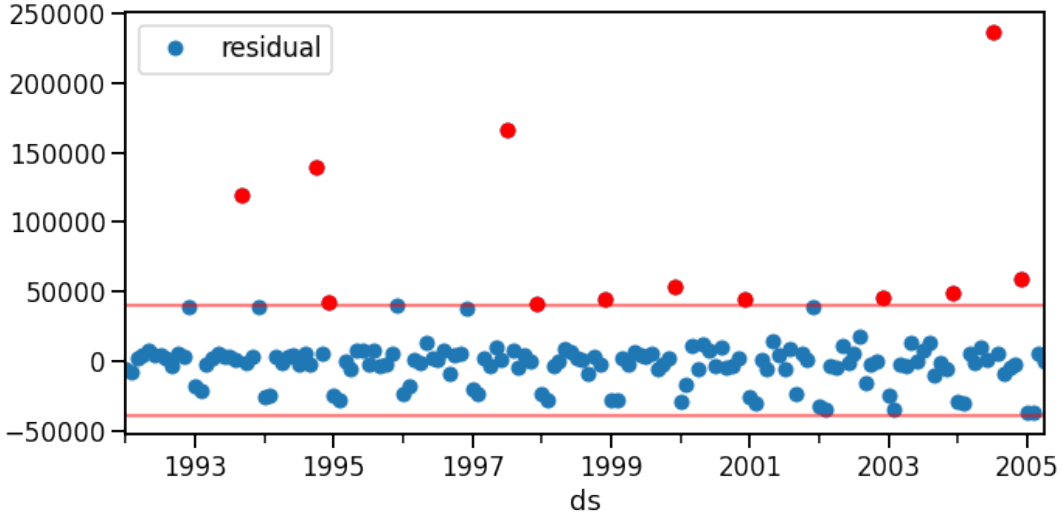
- We set $\alpha = 3$ so that only more extreme outliers are detected



Seasonality can still be an issue

- LOWESS captures the trend but not seasonality here
- So seasonal spikes are picked up as outliers
- Solution: De-seasonalize first or use STL decomposition





LOWESS- summary

- Parameters:
  - LOWESS parameters (fraction of data for window size)
  - Threshold parameter α

- Pros:
  - Robust to outliers
  - Can handle missing data or non-uniform sampling
  - No missing data at edges
  - Captures rapid changes in the trend

- Cons:
  - Computationally more intensive
  - Need to remove seasonality

Summary

LOWESS can extract trends and be used to compute an expected value for a time series

The residuals can be used to identify outliers

Seasonal spikes need to be handled beforehand