Overview

Outliers

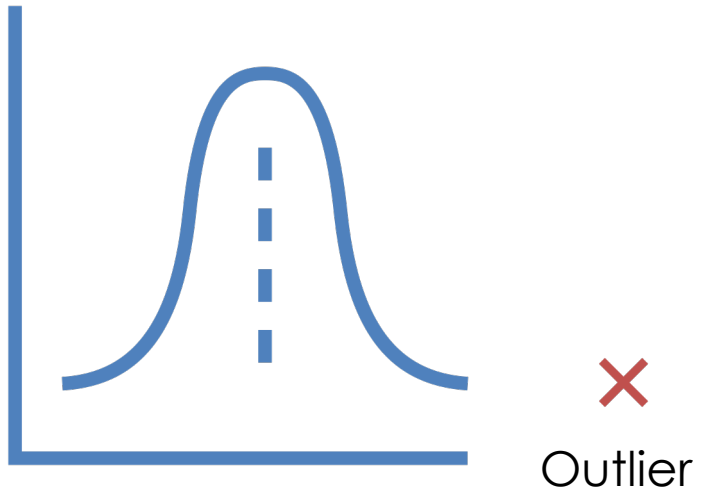
Contents

WHAT IS AN OUTLIER?

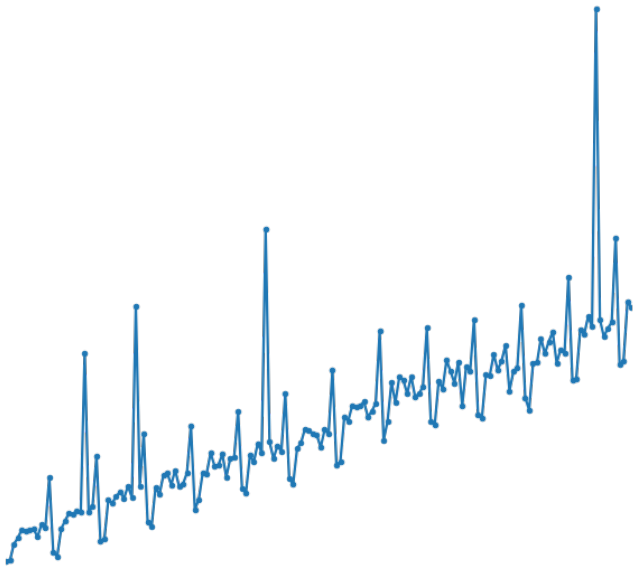
METHODS TO IDENTIFY OUTLIERS

What is an outlier?

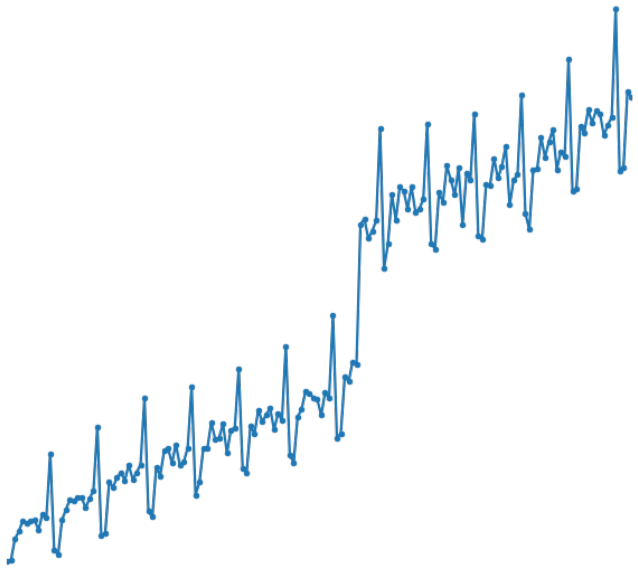
- Data that is very different from other observations
- Suspicion that data generating process is different for these points
- Cause of the outlier determines how we handle it



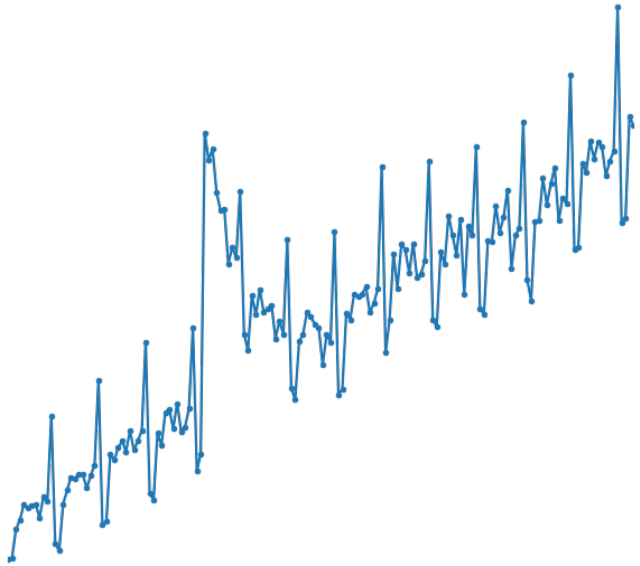
Outliers in time series data



Point outliers



Level shift outlier



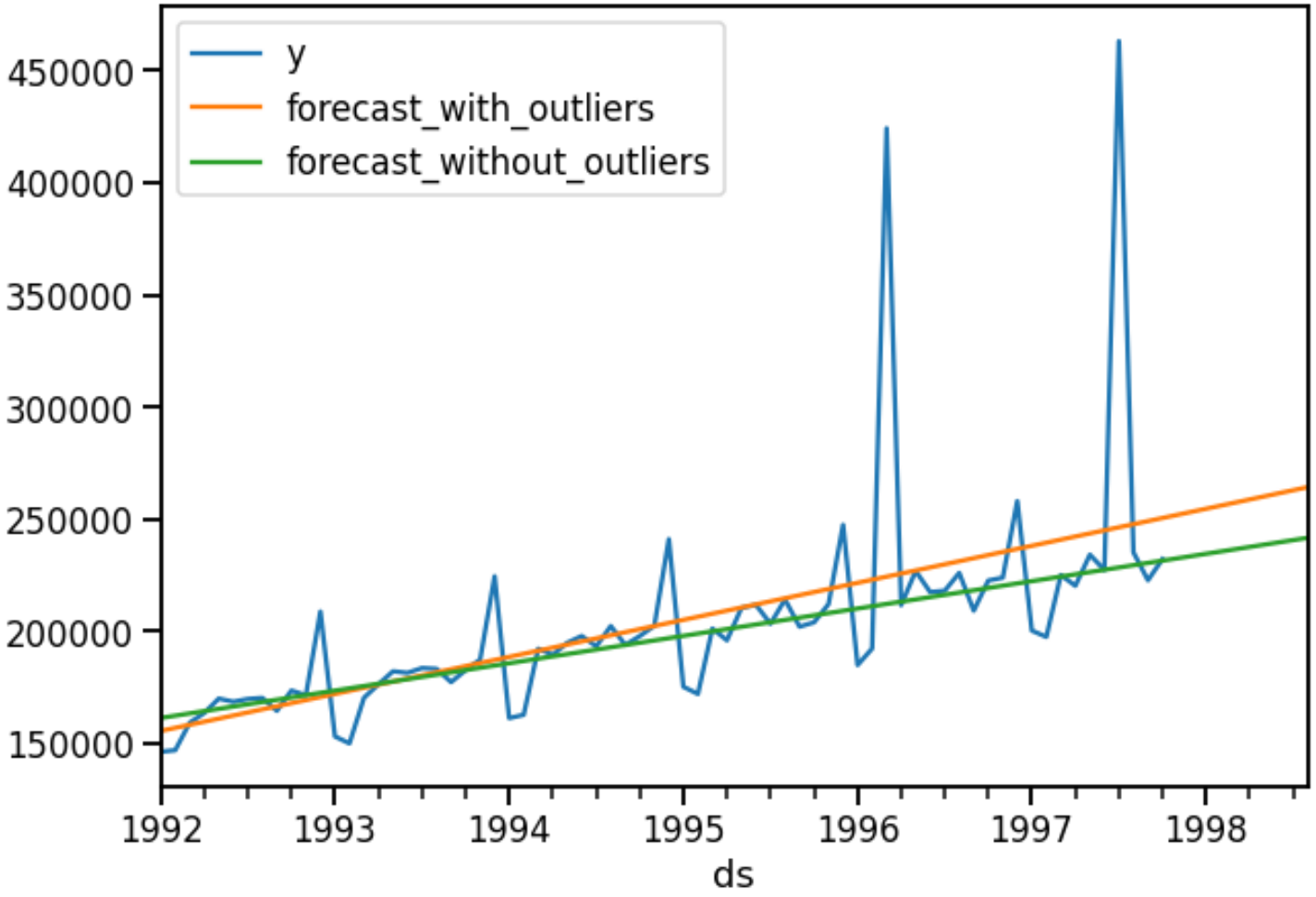
Transient shift outlier

Subsequence outlier *

* Blázquez-García, Ane, et al. "A review on outlier/anomaly detection in time series data." arXiv preprint arXiv:2002.04236 (2020)

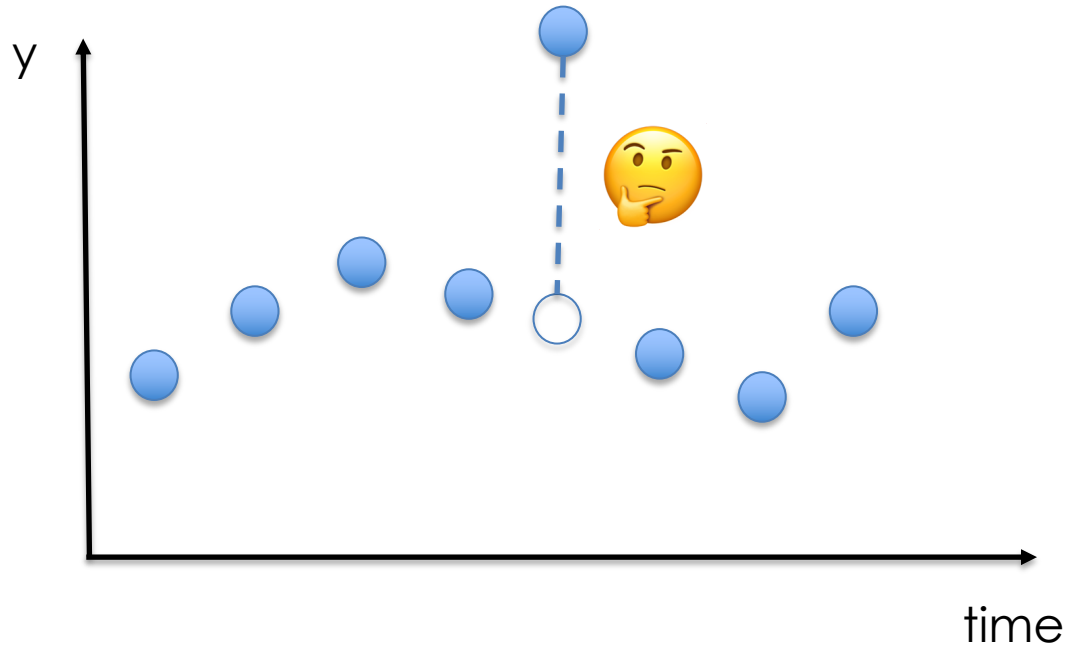
Why are outliers a problem?

Outliers can bias models which result in worse forecasts



Estimation methods

- Examine each actual: y_t
- Compute expected value: $\hat{y}_t$
- Is actual very different than expected?
- If yes, flag as an outlier
- Formally: $|y_t - \hat{y}_t| > \delta$, where δ is a threshold to select outliers

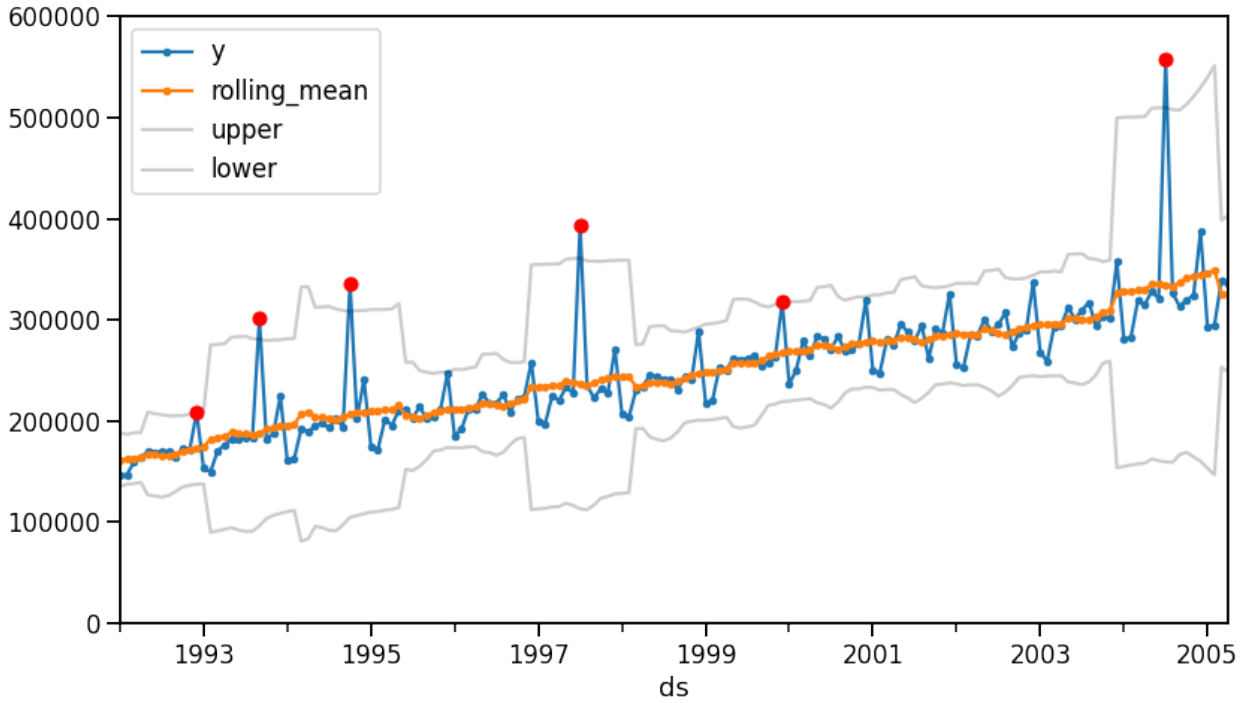


Estimation methods to identify outliers

1. Rolling mean
2. Rolling median
3. LOWESS residuals
4. STL residuals

Estimation methods to identify outliers

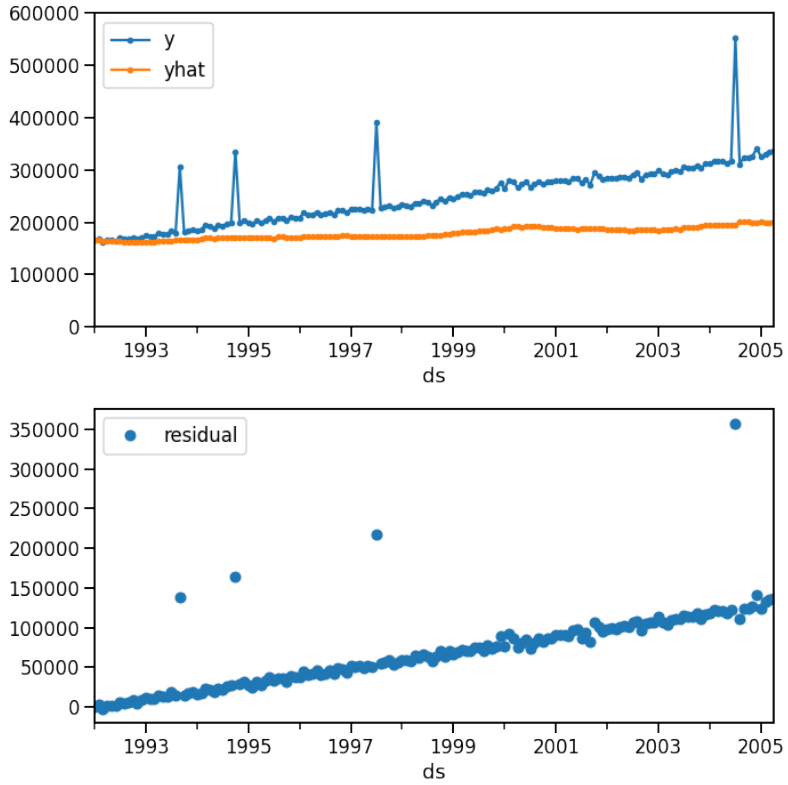
1. Rolling mean
2. Rolling median
3. LOWESS residuals
4. STL residuals



De-seasonalize the data first if seasonal!

Estimation methods to identify outliers

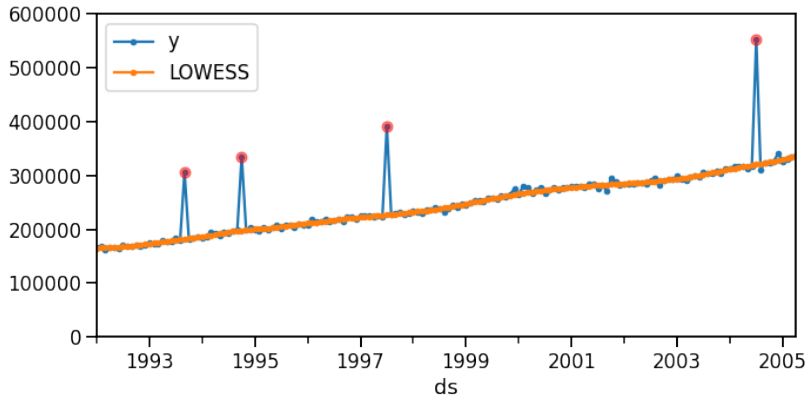
1. Rolling mean
2. Rolling median
3. LOWESS residuals
4. STL residuals

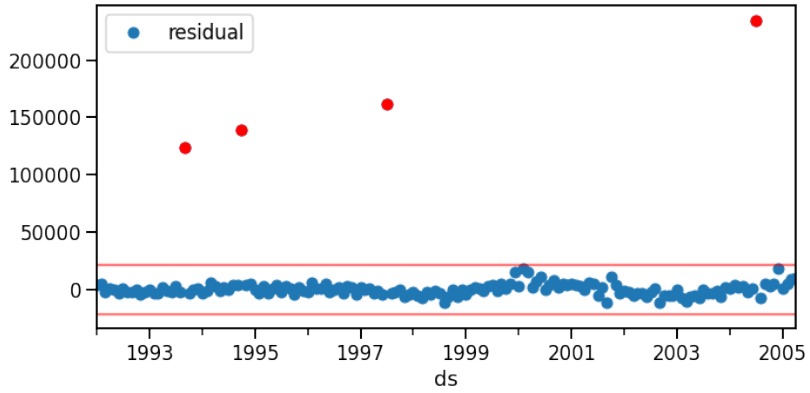


Check for stationary residuals!

Estimation methods to identify outliers

1. Rolling mean
2. Rolling median
3. LOWESS residuals
4. STL residuals





Check for stationary residuals!

Practical tips

- Strong seasonality: STL residuals
- Long term trends: LOWESS residuals
- Faster results: rolling mean or rolling median
- Adjust sensitivity depending on the use case:
  - Do you only care about extreme outliers?
  - What is the cost of a false positive? (e.g., manual review time but reduced business risk)
- Understand the nature of the outlier
  - Known cause which can occur in the future: Model the outlier as a feature
  - Random or uncontrollable: Treat as missing value and impute
- Sense check time series plots before and after imputation (even on a subsample)

Methods shown here are for time series: can be feature or target

Date	y	temperature	Foot fall
2015-01-03	18	23	0
2015-01-04	10000	26	20
2015-01-05	15	-3	2
2015-01-06	7	24	3

- Features which take discrete values and low volumes (i.e., count data) lack structure for STL and LOWESS.
- If the data is count like consider using rolling statistics methods or if the data looks stationary use extreme value analysis methods (e.g., IQR rule)