Robust linear forecasting of trends in water quality

Amanda Clare
Dept of Computer Science
Aberystwyth University

afc@aber.ac.uk

Katherine Martin
Dŵr Cymru Welsh Water

katherine.martin@dwrcymru.com

Trend forecasting

Purposes:

- Determine whether treatment plant capacity will be sufficient in future
- Determine whether levels of contaminants are increasing over time

Approach:

- Consider the values of a statistic (e.g. 95th percentile) over time

Models:

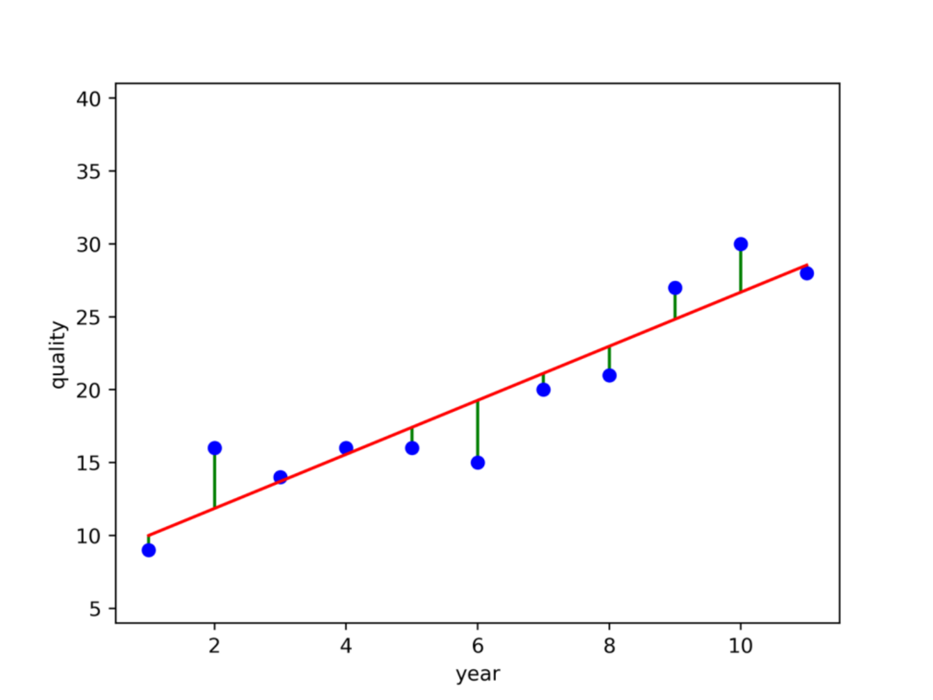
- Linear models
- Autoregressive models
- Non-linear models

Ordinary least squares linear regression

Finds a line of best fit.

Line chosen to minimise **the sum of the squares of residuals** (distances between y values and line).

Often used to make a predictive model (and extrapolate for future trend)



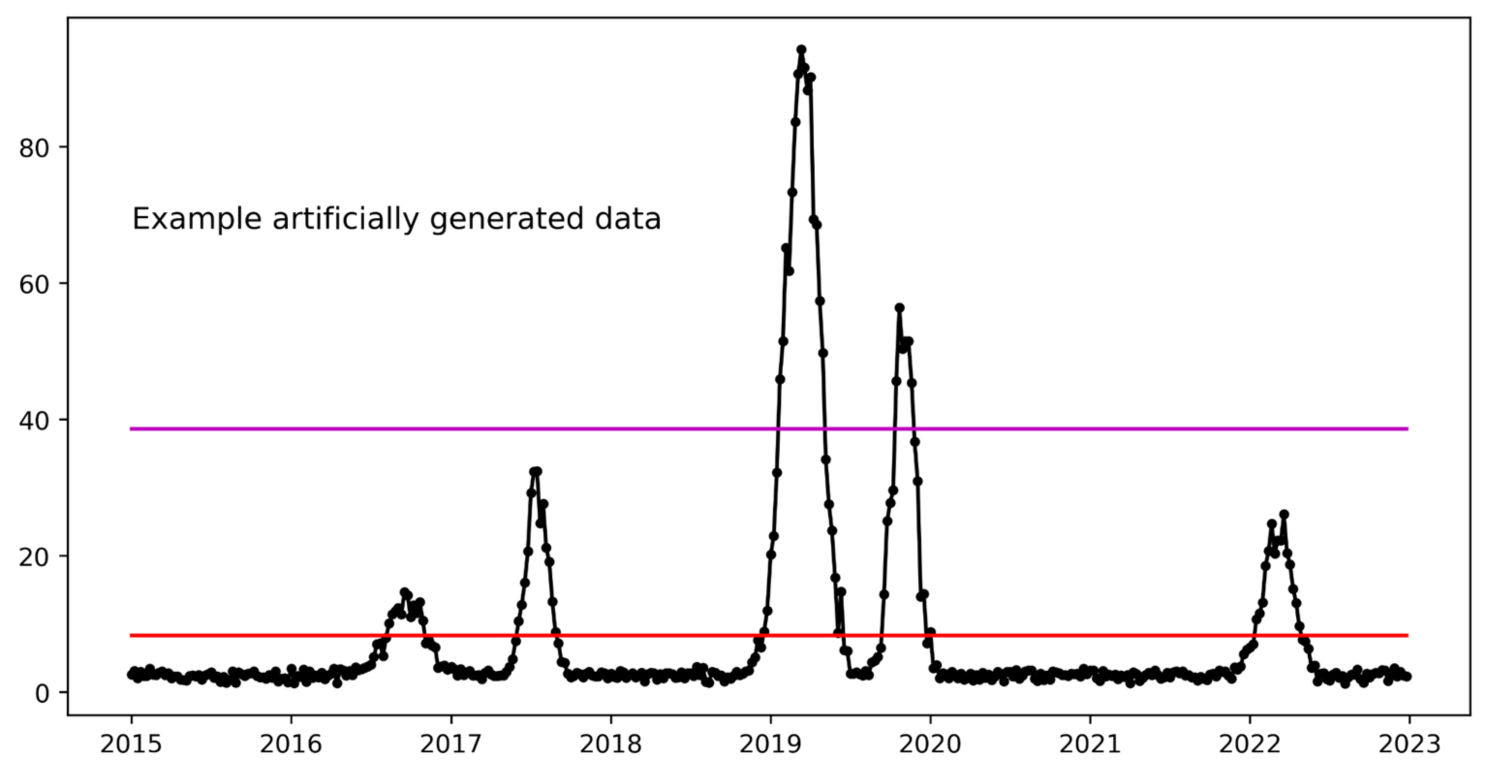
Outliers/Anomalies

Outlier values could be caused by unusual weather, onset of global pandemic, temporarily faulty equipment, etc.

Some measurements are then abnormally high or low

Ideally these should not affect prediction of future trends, as are caused by one-off events.

Mean and 2x standard deviation when outliers present



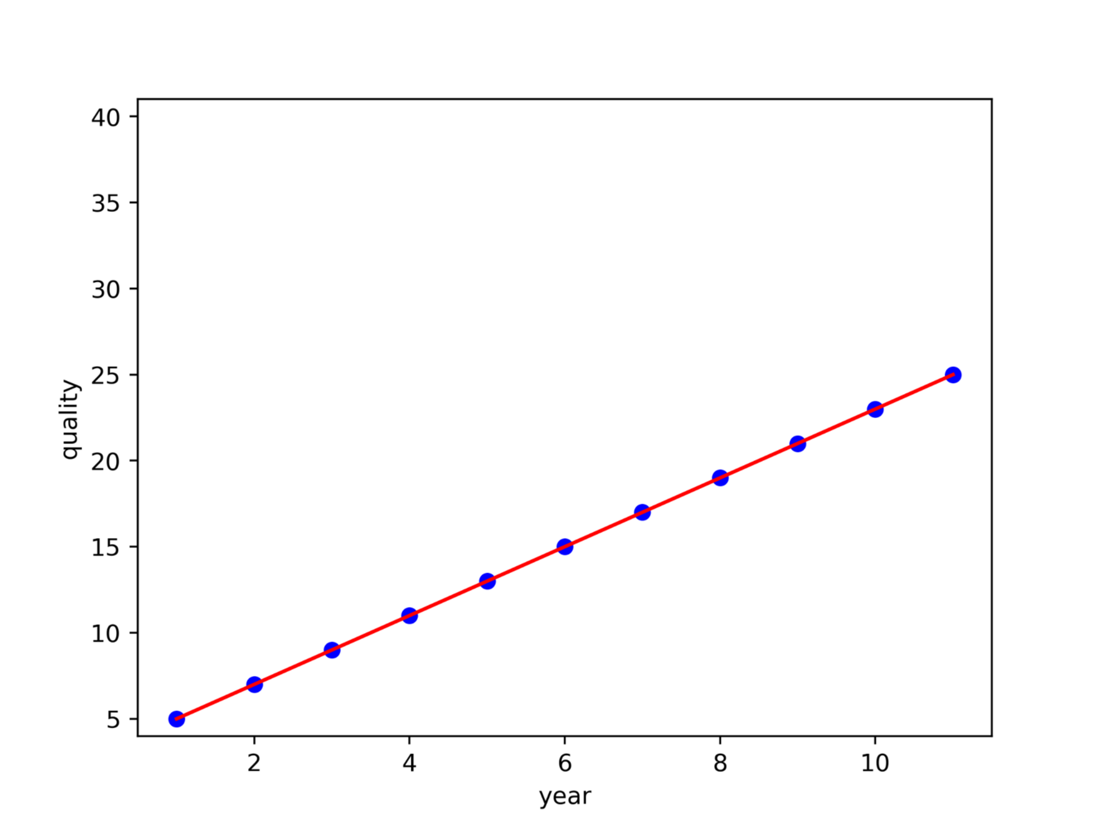
How do outliers affect your data analysis?

- Mean is distorted in the presence of outliers
- So is the variance, and the standard deviation
- So is least squares linear regression, PCA, and many other standard data analysis techniques.

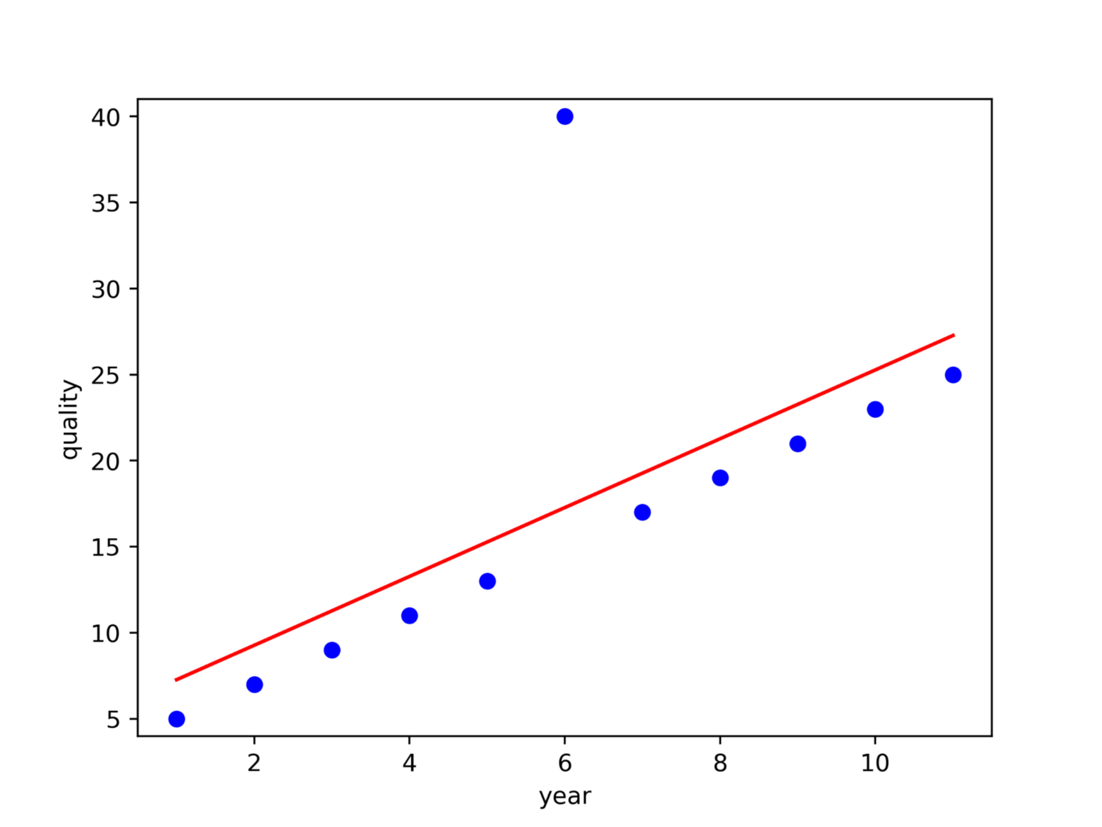
All of these rely on averaging a sum over all the data, which includes any outliers.

What happens to our linear regression models?

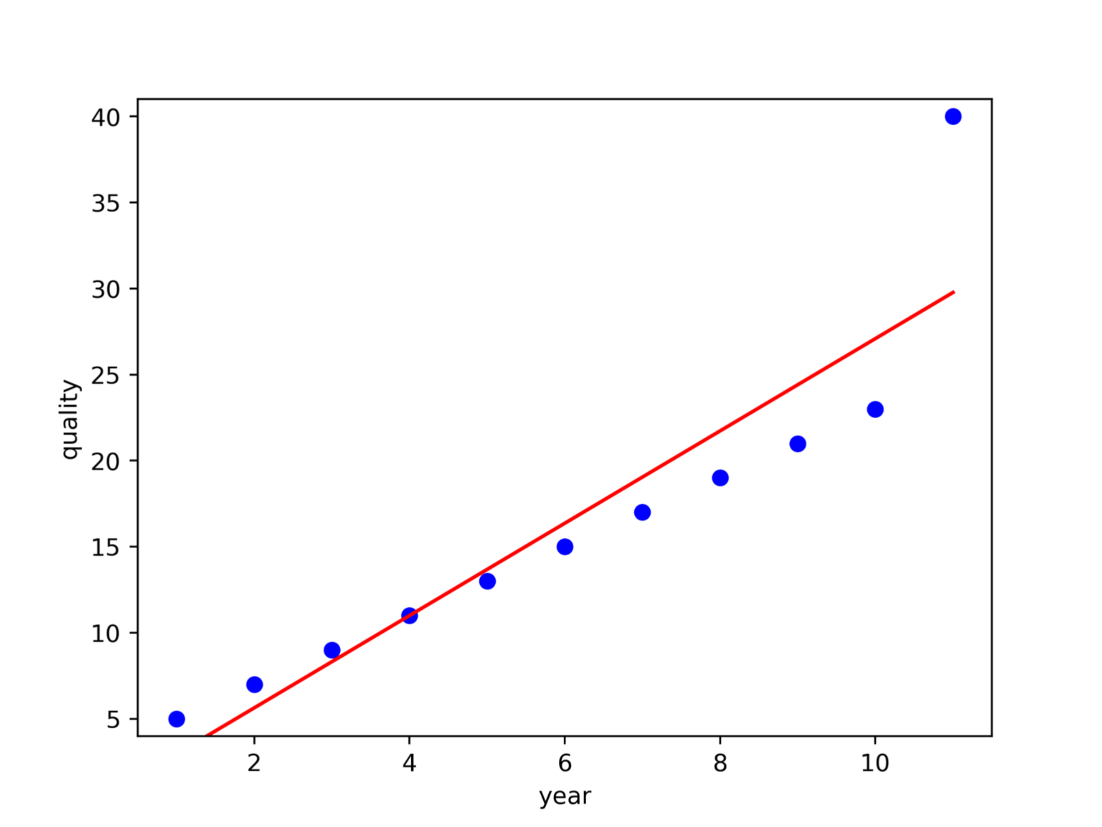
A simple data series



Outlier in middle of series moves the line of best fit



Outlier at end of series changes gradient of line



Using Median instead of Mean

Median is the middle value of your sorted set of values in a data set.

For example [1,1,2,3,3,3,4,6,15]

Median = 3

Mean = 4.22

(note: the median does not rely on summing all the data)

Robust decision making

The median is an example of a **robust** statistic.

If some of your data is:

- corrupted,
- recorded badly,
- measured incorrectly,
- or just has unusual outliers

How will your analysis be affected?

If we set some values to arbitrarily large numbers, how reliable is your result?

Robust model:

- It should be a reasonably good quality model
- Small deviations from the model assumptions should impair the model performance only slightly
- Somewhat larger deviations from the model should not cause a catastrophe

Paraphrased from: Huber, Peter J. (1981), Robust statistics, New York: John Wiley & Sons, Inc., ISBN 0-471-41805-6

The median is robust, but not the mean

For example [1,1,2,3,3,3,4,6,15]

Median = 3

Mean = 4.22

[1,1,2,3,3,3,4,15,1000]

Median = 3

Mean = 114.7

If one value is set to 1000, the mean is completely ruined, but the median is not.

Median has a high **breakdown point**

How many values can you corrupt without your model becoming useless?

Median: just under half

- The median will change, but won't be arbitrarily far from the right amount

Mean: no values

- Just one outlier going wild can arbitrarily corrupt your model, especially for small data sets.

Are there regression methods that are more robust, and use the median?

Theil-Sen estimator (robust line fitting)

Finds a robust line of best fit to your data using medians.

How: Want to find a line $y = mx + c$

Find m using median.

Find c using median.

This method first published in 1950 (Theil) and updated in 1968 (Sen).

Sen, Pranab Kumar (1968), "Estimates of the regression coefficient based on Kendall's tau", Journal of the American Statistical Association, 63 (324): 1379–1389, doi:10.2307/2285891. - cited by over 10,000 papers

Find median slope (m)

Find the median slope of all slopes made from all pairs of data with different x-coordinates.

For each `i, j` where `i < j` we find

$$\mathtt{slope}_{ij} = (y_i - y_j) / (x_i - x_j)$$

Then take

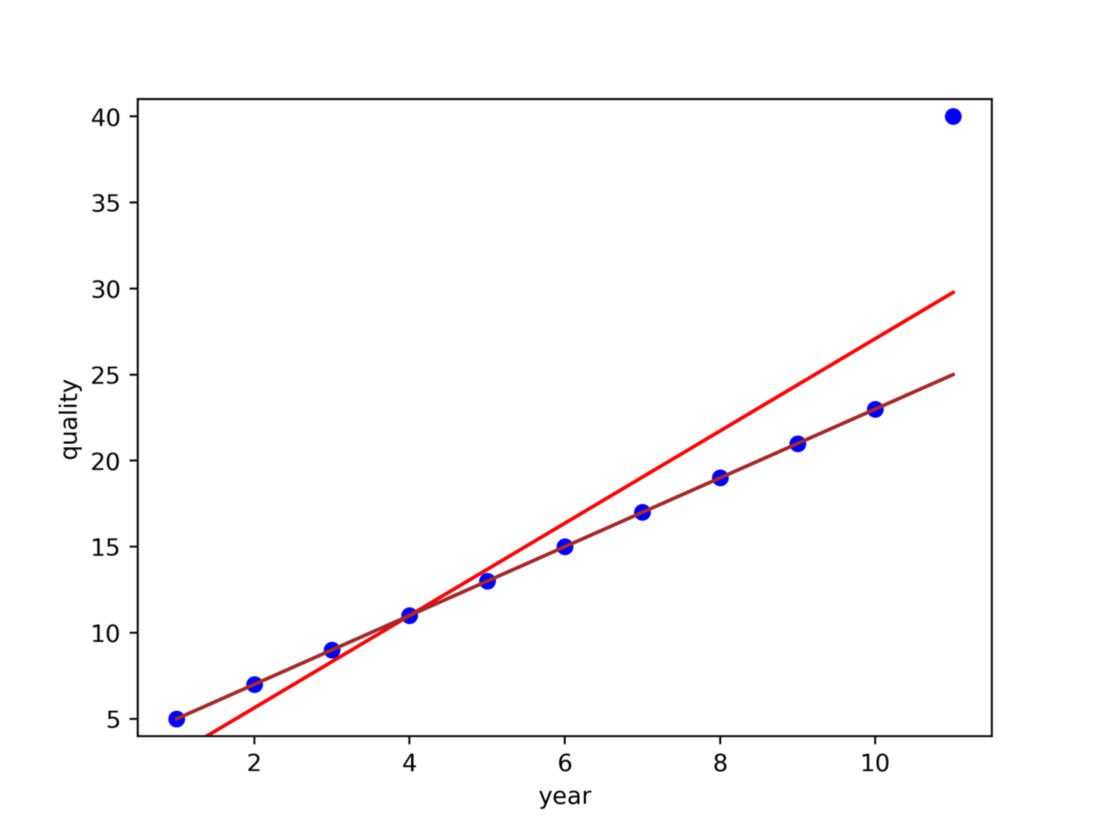
$$m = \mathrm{median}(\mathtt{slope}_{ij})$$

Now find median intercept c

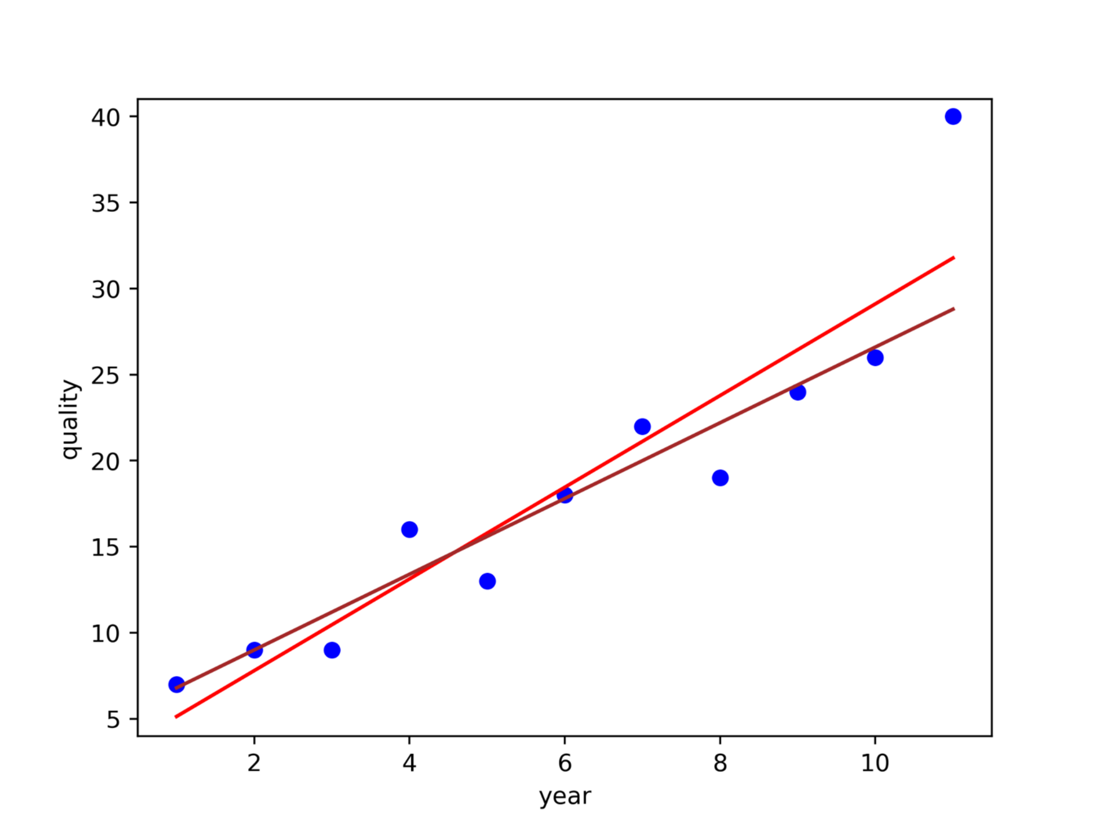
For each data point in turn, find the intercept of a line of slope m passing through this point ($c_i = y_i - mx_i$)

Take the median of these intercepts c_i to give $c = \mathrm{median}(c_i)$.

Theil-Sen (brown) fits better than least squares (red)



Theil-Sen (brown) fits well even when noisy data



Choose robust methods to model your trends

Least squares regression does not handle outliers well.

Robust statistics should be more widely considered and adopted.

Theil-Sen is a robust method to find a line of best fit to model your trend.

Theil-Sen has a breakdown point of 29.3% (can handle several outliers).

Other robust regression methods exist (Seigel 1982, Rousseau 1984, Yohai 1987)

Email: afc@aber.ac.uk