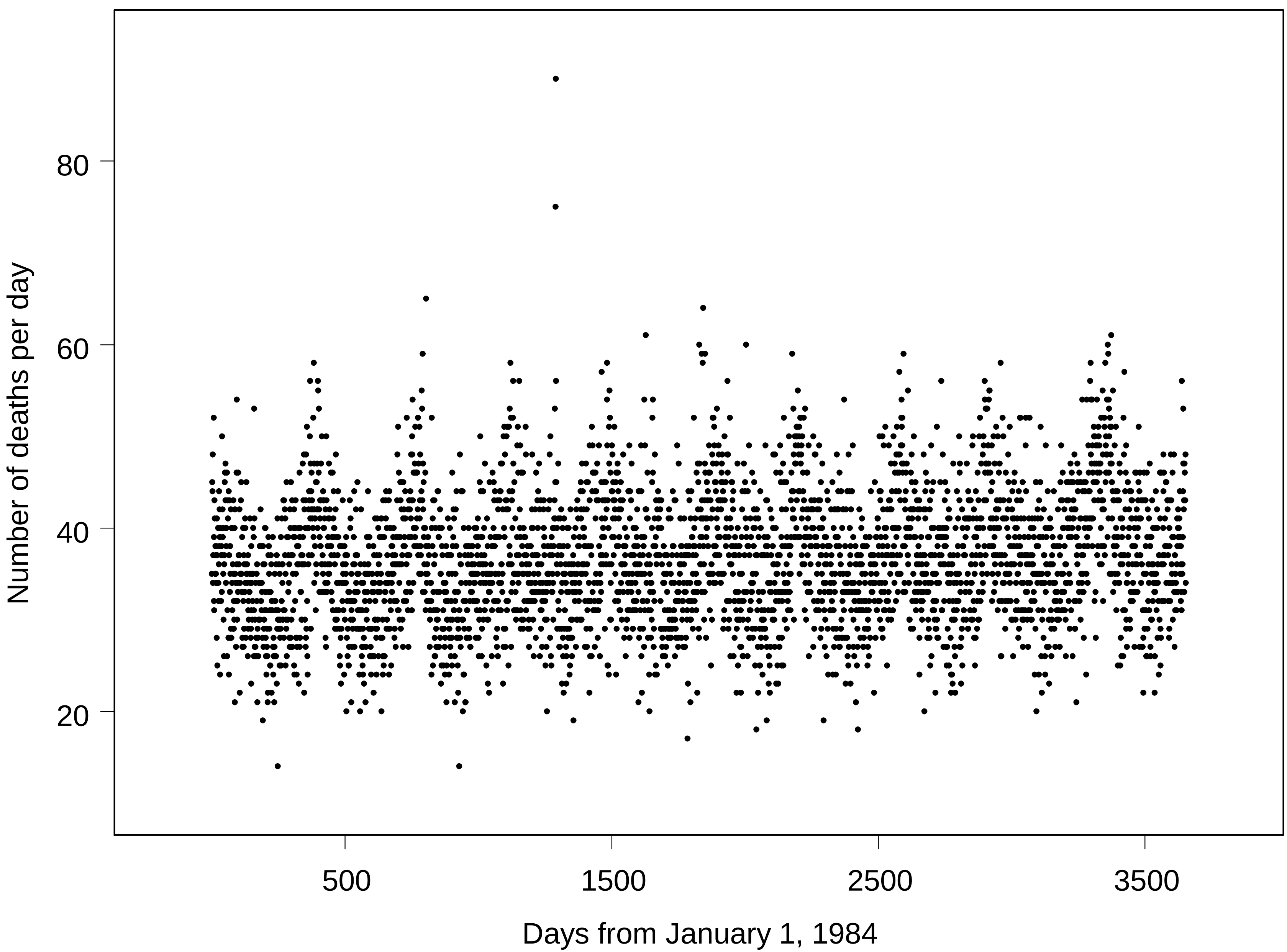
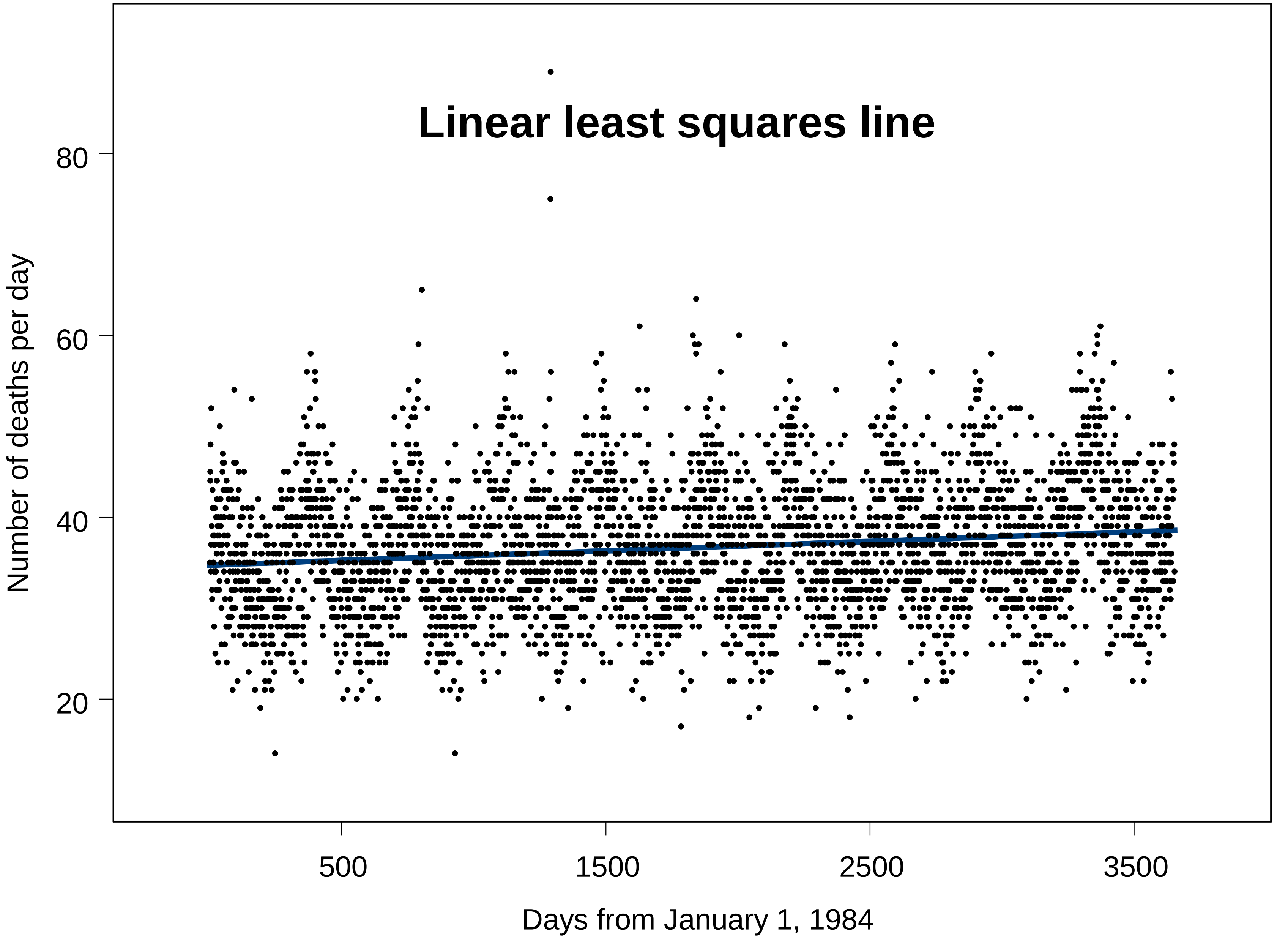
Data Visualization and Model Building in Regression Analyses: Use of Generalized Additive Models in Epidemiology

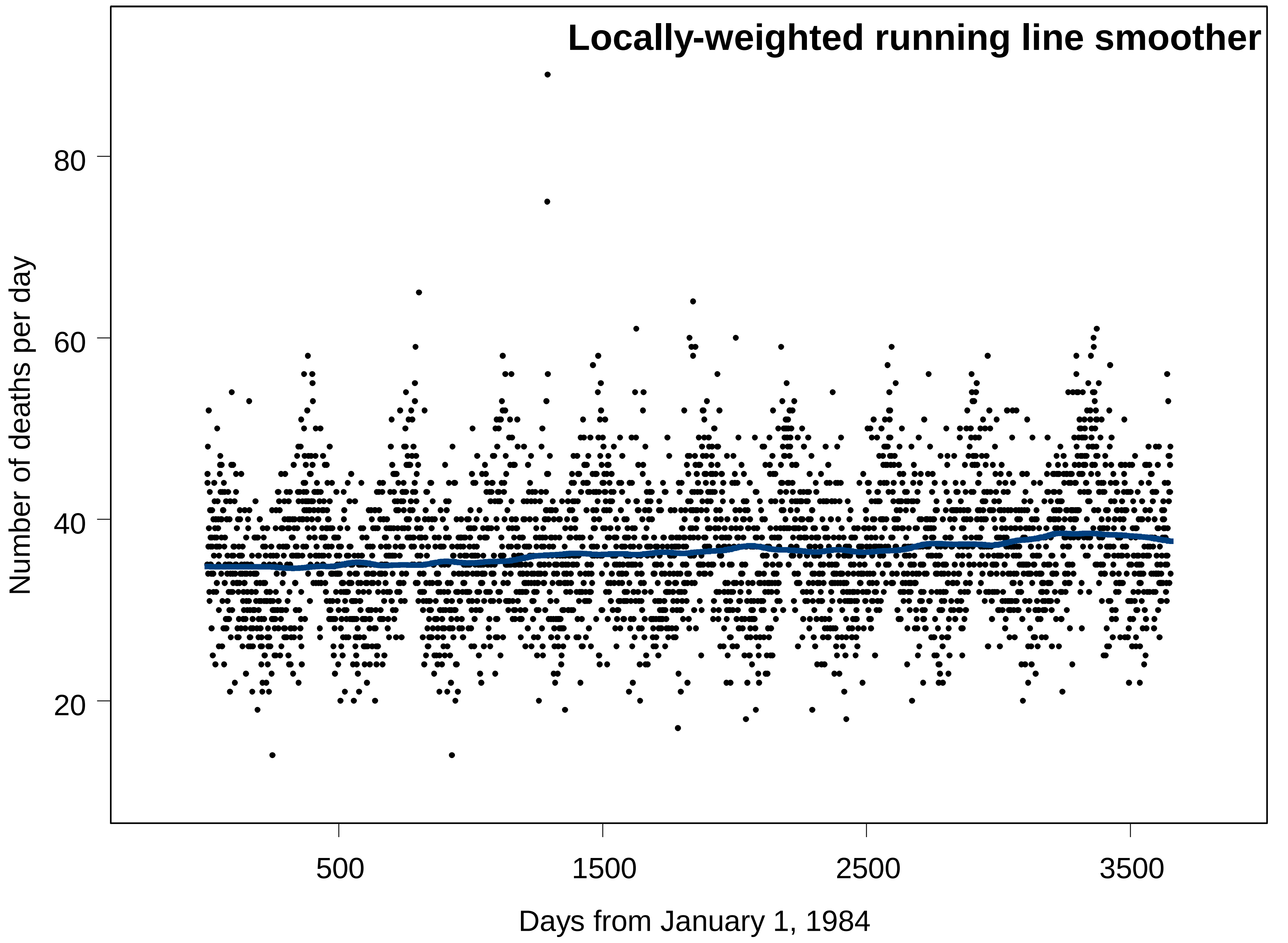
Mark Goldberg
McGill University



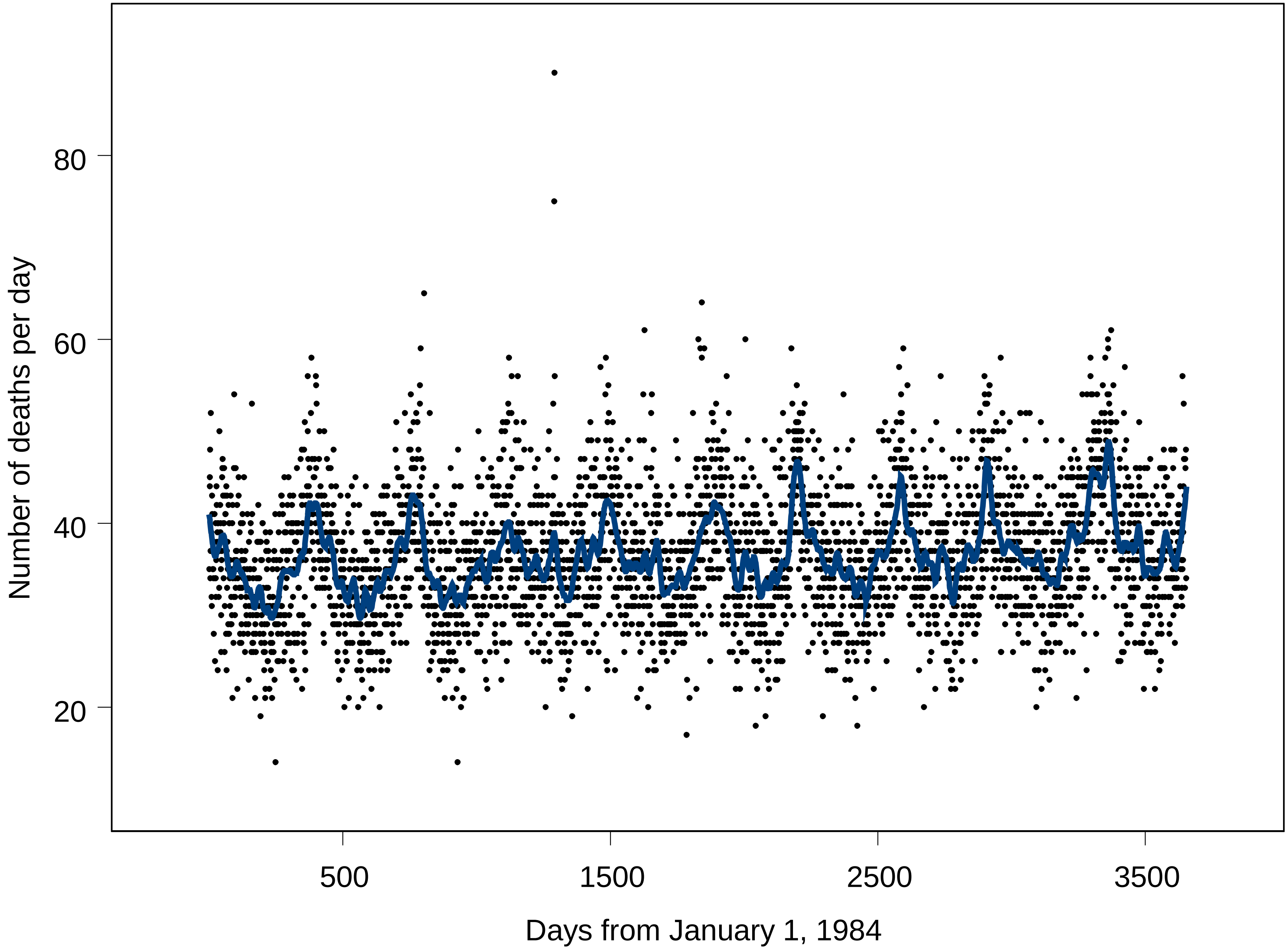
80
60
40
20
Number of deaths per day
500
1500
2500
3500
Days from January 1, 1984


Linear least squares line
Number of deaths per day
20
40
60
80
500
1500
2500
3500
Days from January 1, 1984

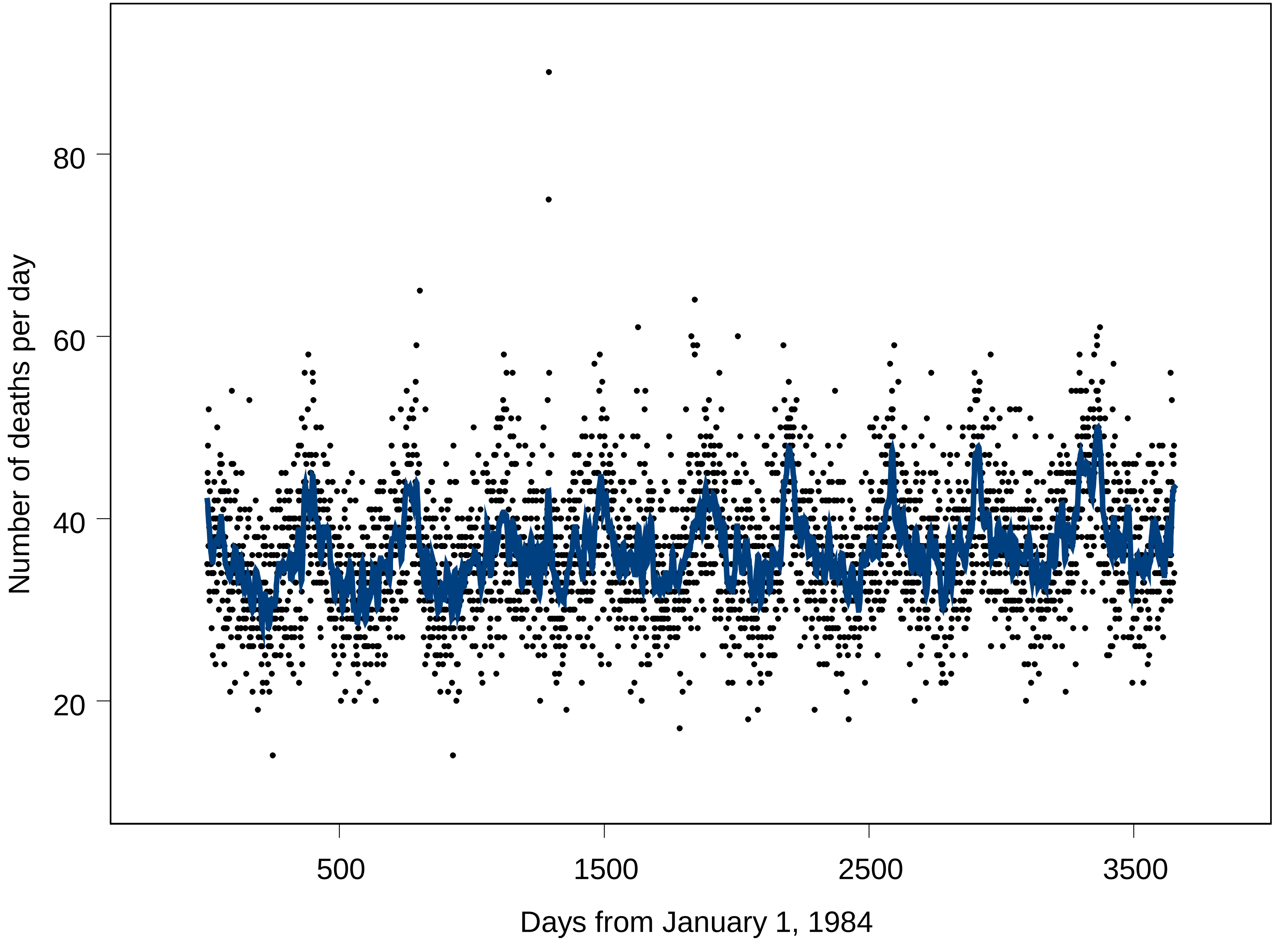
Loess smooth using 20% of the data



Loess smooth using 1% of the data



Loess smooth using 0.5% of the data



Smoothers

Smoothers

- **Definition:** A mathematical function that transforms the relationship between a continuous variable (x) and a response variable (y).
- The result of the operation is a function that is less variable than the original variable.
- It is nonparametric, as it is not based on a rigid mathematical function.

Locally-weighted Running-line Smoothers (LOESS)

- For each data point (x_0), loess uses the k nearest neighbouring points.
- D = distance from x_0 to the furthest point in the neighbourhood
- Each adjacent point in the neighbourhood is given a weight

The weight function is:

$W(u) = (1-u^3)^3$ for $0 \le u < 1$

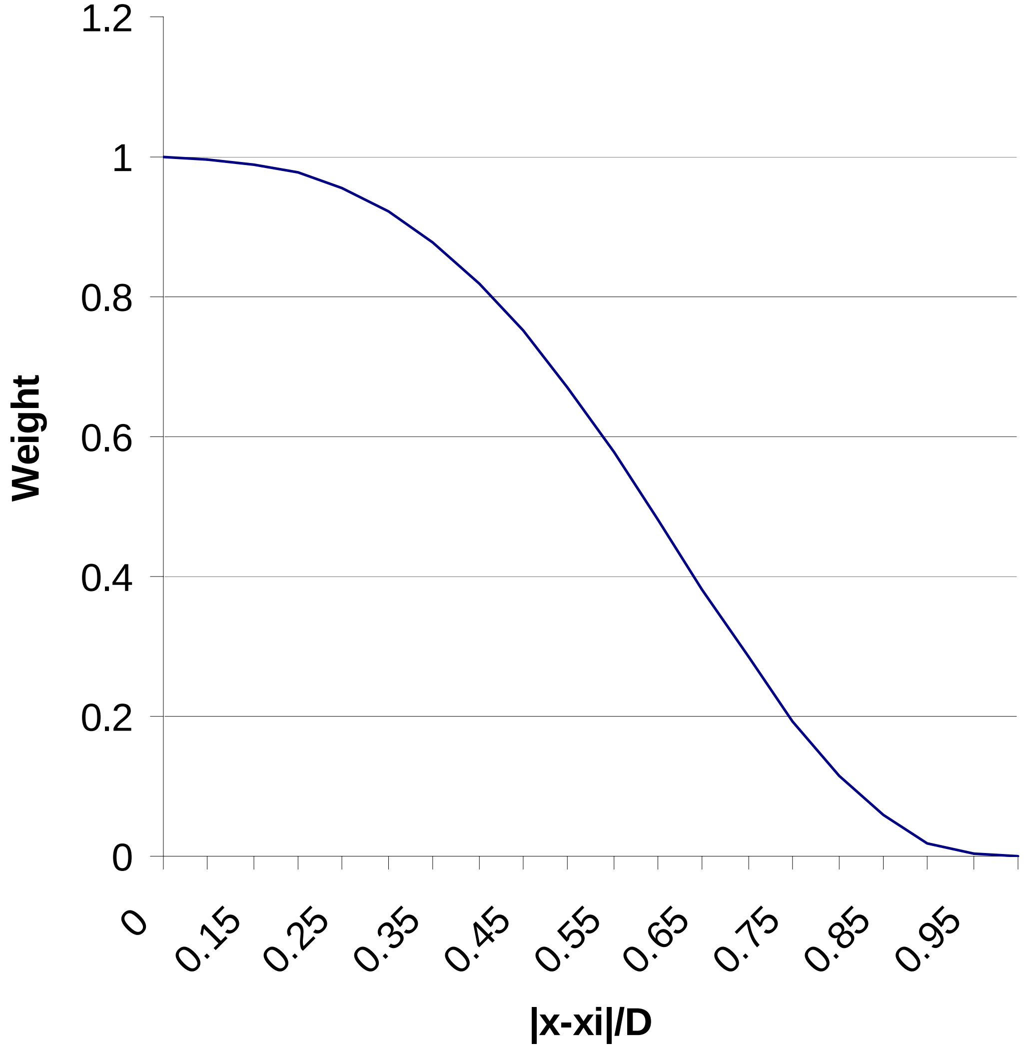
$= 0$ otherwise

where $u = |x_0 - x_i| / D$

W is a maximum at $u=0$ $(x=x_0)$

W is a minimum at $u=1$

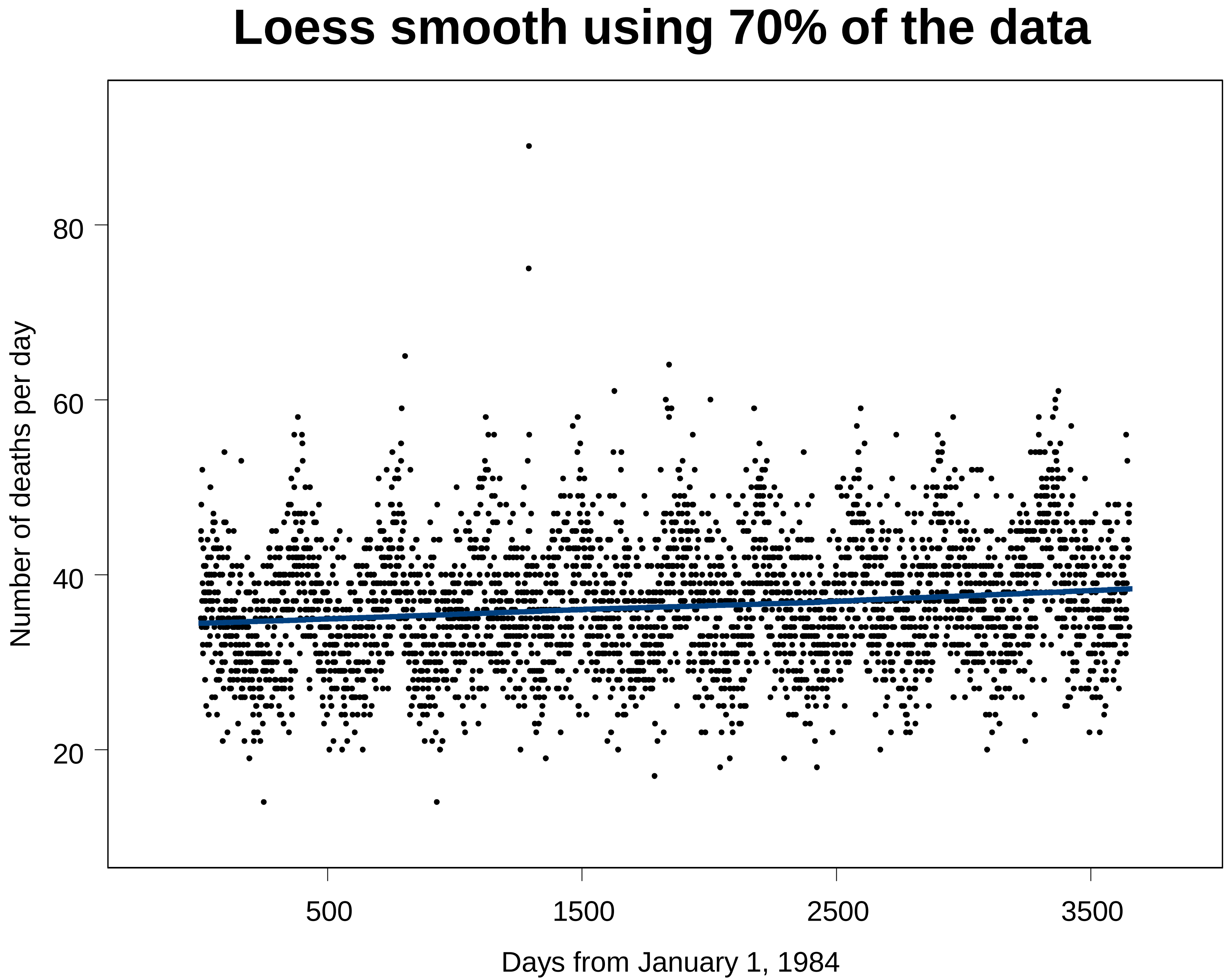
Weight Function for LOESS Smoothers



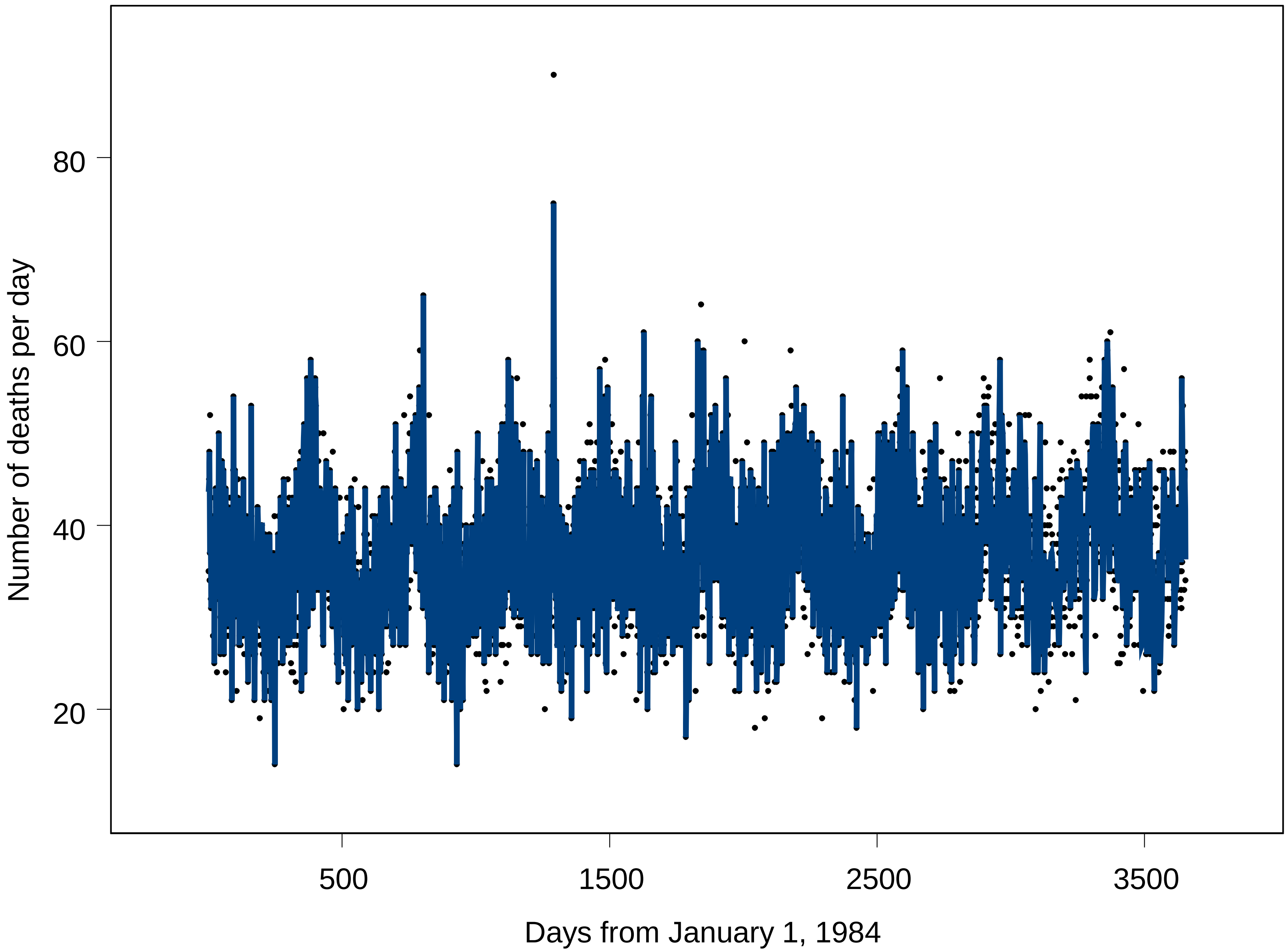
- Weighted linear least squares is then carried out in the neighbourhood of points and the smooth value at x_0 is just the fitted value from the regression equation.
- Polynomial regression can be used instead.

- **Span:** The percentage of data points used as nearest neighbours (in percent of total n).

- **Advantages:**
  - Handles end-points nicely
  - Can easily tune the smoothness using the percent of data points to be included in the neighbourhood
  - Excellent for interactions (e.g., plotting three dimensional surfaces)


Loess smooth using 70% of the data
80
60
40
20
Number of deaths per day
500
1500
2500
3500
Days from January 1, 1984

Loess smooth using 0.1% of the data



Regression Splines

- Divide the data by a sequence of cutpoints (knots).
- Carry out polynomial regression (usually cubic) within each of these regions.
- Each polynomial must join up at each knot in a smooth fashion.
- This is achieved by ensuring that the first and second derivatives are continuous at the knots.

- The type of polynomial (e.g., cubic) and the number of knots define effectively the number of degrees of freedom used.

- NB: Cubic has four degrees of freedom.

- **Disadvantages:**
  - Must specify the number and/or position of the knots in advance.
  - Small numbers of knots can lead to spurious nonlocal behaviour.
  - Smoothness cannot be varied continuously, as with LOESS.
  - Not great for multi-dimensional plots

Degrees of Freedom

- Degrees of freedom is an indication of the amount of smoothing
  - More smoothing: fewer df or higher span
- Df is not necessarily an integer

Approximate Correspondence between Degrees of Freedom and Span

Df	Span
2.5	1
4	0.5
5	0.3
6	0.22
10	0.18

Lots of smoothing

From Hastie and Tibshirani, page 53, Figure 3.5