# Chapter 14

Fundamentals of Data Visualization

May 1, 2023

# Visualizing Trends

- When making scatter plots or time series, we are often more interested in the overarching trend of the data than in the specific detail.
- By drawing the trend we can create a visualization that helps the reader immediately see key features of the data.
- There are two fundamental approaches to determining a trend:
  - We can smooth the data by some method.
  - We can fit a curve with a defined functional form.
- Once we have identified a trend we can look specifically at deviations from the trend.
- Or we can separate the data into multiple components, including the underlying trend, any existing cyclical components, and episodic components or random noise.

# Dow Jones for 2009

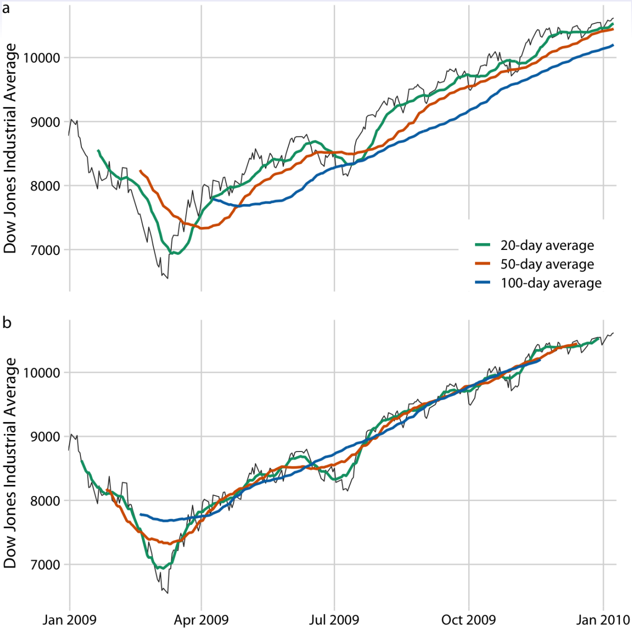

- How can we show the dominant trends without the noise?

# Smoothing by moving average

- To generate a moving average, we take a time window, say the first 20 days in the time series.
- Calculate the average price over these 20 days.
- Then move the time window by one day, so it now spans the 2nd to 21st day.
- Calculate the average over these 20 days.
- Move the time window again, and so on.
- The result is a new time series consisting of a sequence of averaged prices.

- Financial analysts usually plot the smooth curve at the end point.

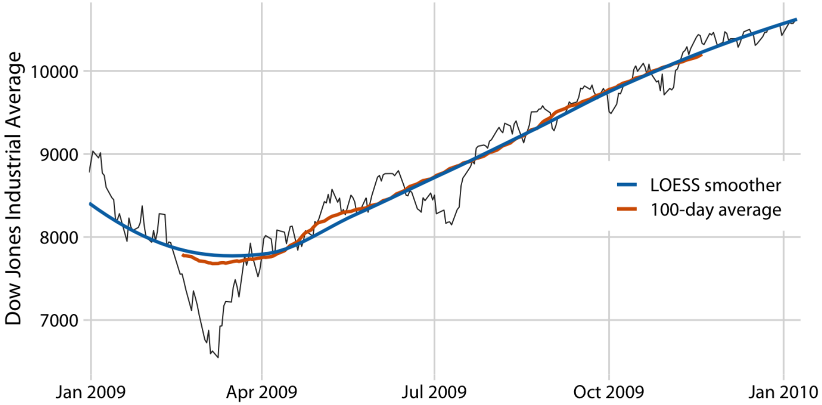
- Statisticians usually plot the smooth curve at the center of the window.



# Moving average

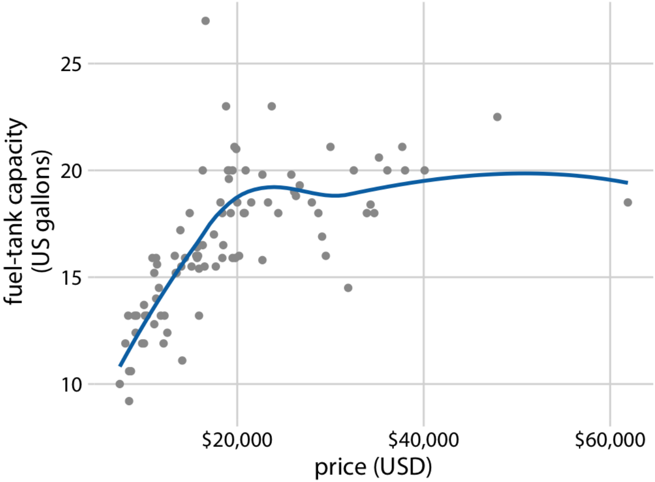
- Rresults in a curve that is shorter than the original curve.
- Parts are missing at the beginning or the end or both.
- The more the series is smoothed the shorter the smoothed curve.
- It is not necessarily really very smooth.
- Wiggles are caused by individual data points that enter or exit the averaging window.
- Since all data points in the window are weighted equally, individual data points at the window boundaries can have visible impact on the average.

# LOESS smoothing



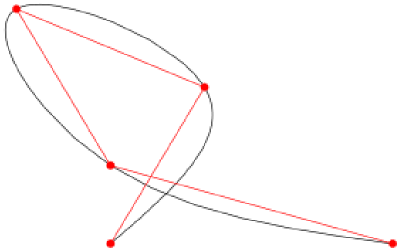

- Fit low-order polynomials to subsets of the data.
- Weight points by proximity to center of subset.
- Amount of smoothing controlled by parameters.

# LOESS can be used on non-time series




- Fit looks good to human eye.
- Fit comes from many separate regressions, can be slow.

# Splines



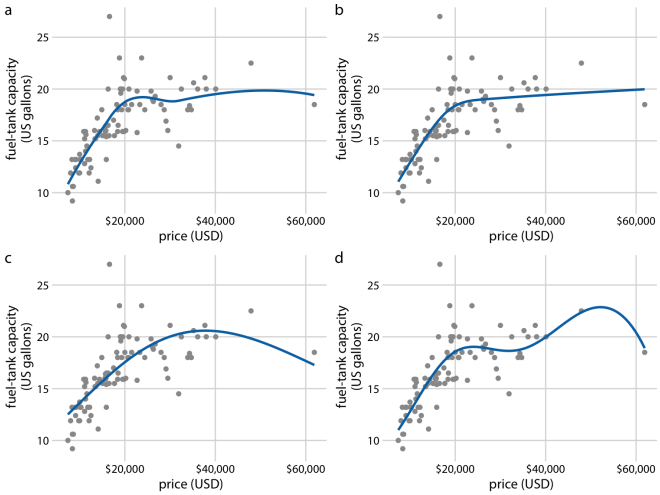
https://mathworld.wolfram.com/CubicSpline.html

- A cubic spline is a spline constructed of piecewise third-order polynomials which pass through a set of $m$ control points, called the **knots**.
- The second derivative of each polynomial is commonly set to zero at the endpoints, since this provides a boundary condition that completes the system of $m - 2$ equations.

# Splines

- A spline is a piecewise polynomial function that is highly flexible yet always looks smooth.
- The **knots** in a spline are the endpoints of the individual spline segments.
- If we fit a spline with $k$ segments, we need to specify $k + 1$ knots.
- Spline fitting is computationally efficient.
- There is a bewildering array of different types of splines, including cubic splines, B-splines, thin-plate splines, Gaussian process splines, and many others.
- The specific choice of the type of spline and number of knots used can result in widely different smoothing functions for the same data.

# Splines




(a) LOESS smoother. (b) Cubic regression splines with 5 knots.
(c) Thin-plate regression spline with 3 knots. (d) Gaussian process spline with 6 knots.

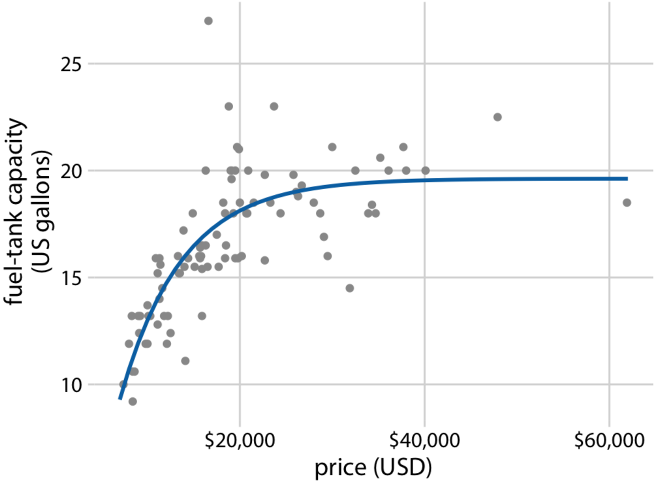
# Smoothers

- Most data visualization software provides smoothing.
- Either a LOESS or a spline, or both.
- The smoothing method may be referred to as a GAM, a generalized additive model.
- The output of the smoothing feature is highly dependent on the model that is fit.
- Unless you try out a number of different choices you may never realize to what extent the results you see depend on the default choices made by your software.
- **Be careful when interpreting the results from a smoothing function.**
- **The same dataset can be smoothed in many different ways.**

# Showing trends with a functional form

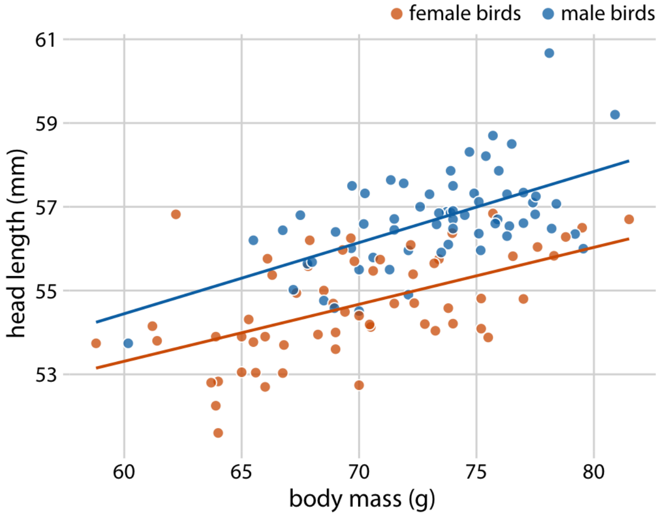
- General-purpose smoothers are somewhat unpredictable for any given dataset.
- These smoothers also do not provide parameter estimates that have a meaningful interpretation.
- Whenever possible, it is preferable to fit a curve with a specific functional form that is appropriate for the data and that uses parameters with clear meaning.

$$y = A - B\exp(-mx)$$



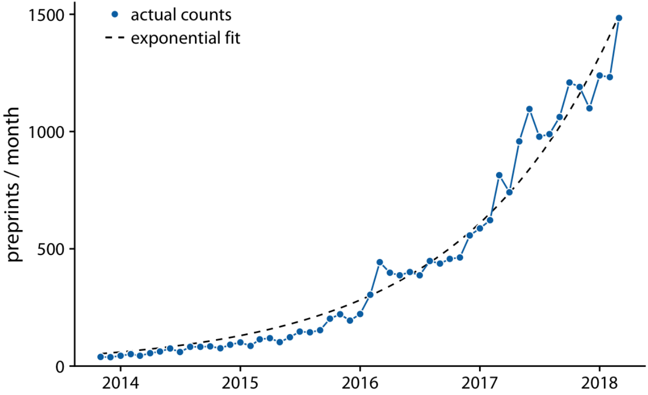

- Fitted parameters: $A = 19.6$, $B = 29.2$, $m = 0.00015$

$$y = A + mx$$




- Approximately linear relationships between two variables are surprisingly common in real-world datasets.

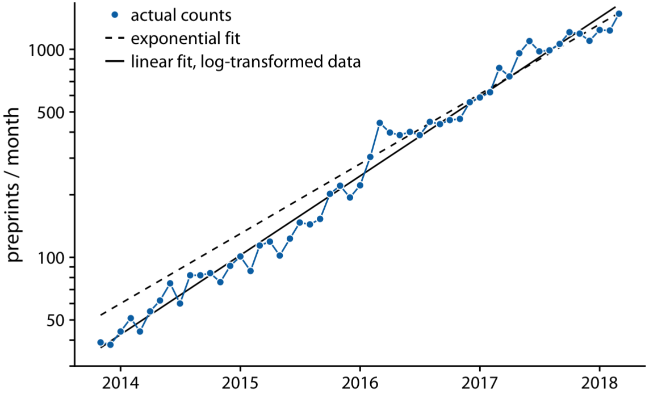
# Finding non-linear relationships




$$y = 60 \exp(0.77(x - 2014))$$

- Percentage growth each year = exponential growth

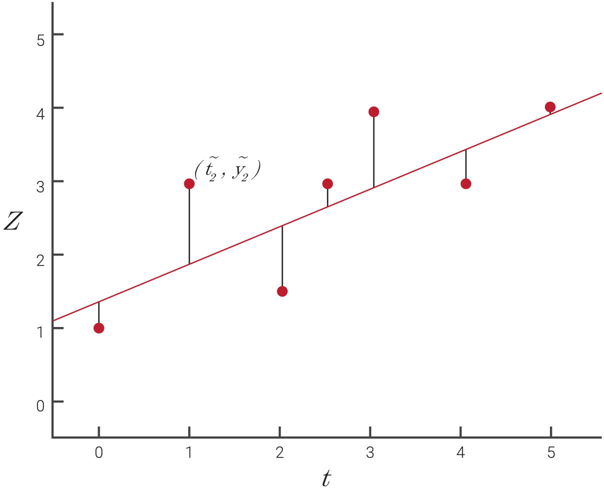
# Transform and look for linear relations




$$y = 43\exp(0.88(x - 2014))$$

- **It is usually better to fit a straight line to transformed data than to fit a nonlinear curve to untransformed data.**

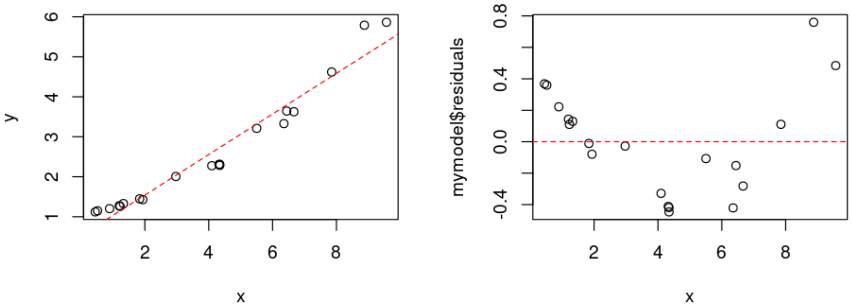
# Least squares regression




https:

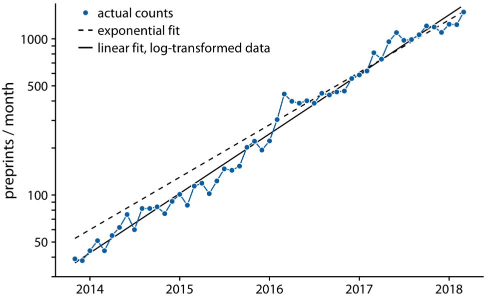
//www.vectornav.com/resources/inertial-navigation-primer/math-fundamentals/math-leastsquares

# Residuals



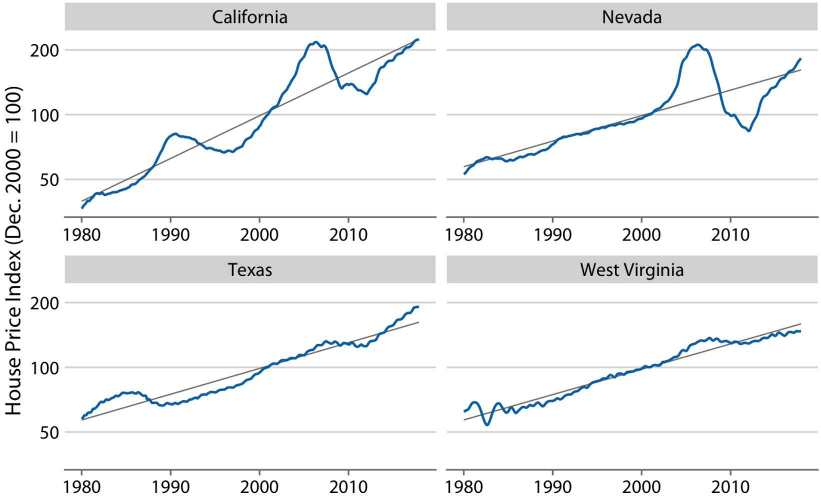

```
x <- runif(20)*10
y <- 1.2^x + runif(20)*x*0.1
plot(y ~ x)
mymodel <- lm(y~x)
abline(mymodel$coefficients, lty="dashed", col="red")
plot(mymodel$residuals ~ x)
abline(h=0,lty="dashed", col="red")
```

# Types of plots



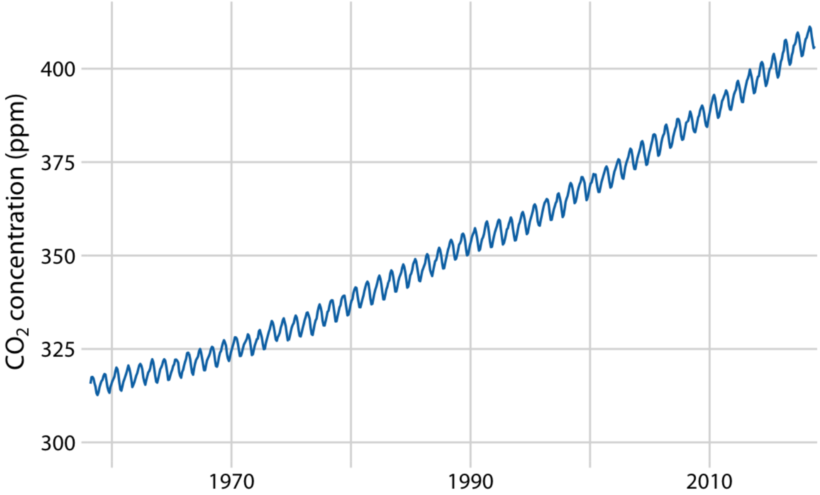

| type | straightens |
|---|---|
| log-linear | $y \sim \exp(x)$ |
| log-log | $y \sim x^{\alpha}$ |
| linear-log | $y \sim \log(x)$ |

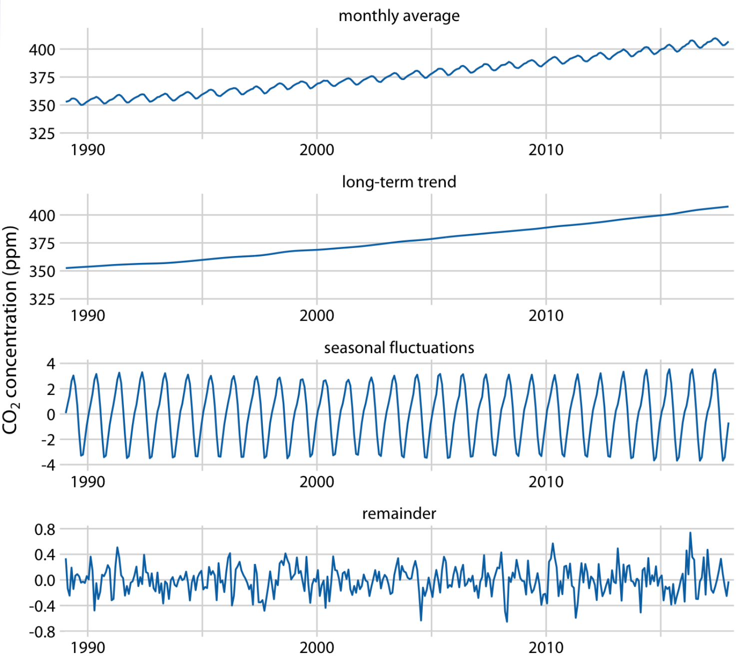
# Detrending




- Divide by the fit value

# Seasonal decomposition of Time series by LOESS




monthly average
long-term trend
seasonal fluctuations
remainder
CO2 concentration (ppm)
400
375
350
325
4
2
0
-2
-4
0.8
0.4
0.0
-0.4
-0.8
1990
2000
2010