Rate of Change
Part 2: Fitting and Using Lines

INFO-1301, Quantitative Reasoning 1

University of Colorado Boulder

October 31, 2016

Prof. Michael Paul

Prof. William Aspray

Interpreting Linear Functions

Fishermen in the Finger Lakes Region have been recording the dead fish they encounter while fishing in the region. The Department of Environmental Conservation monitors the pollution index for the Finger Lakes Region. The model for the number of fish deaths *y* for a given pollution index *x* is $y = 9.607x + 111.958$.

What is the meaning of the slope?

What is the meaning of the y-intercept?

Interpreting Linear Functions

Fishermen in the Finger Lakes Region have been recording the dead fish they encounter while fishing in the region. The Department of Environmental Conservation monitors the pollution index for the Finger Lakes Region. The model for the number of fish deaths *y* for a given pollution index *x* is $y = 9.607x + 111.958$.

What can we do with this function?

- Estimate fish deaths for pollution values that we've never measured

Interpolation and Extrapolation

$y = 9.607x + 111.958$ *x* is the pollution index

Suppose we came up with this formula as an approximation after measuring fish deaths when the pollution index was: 1.1, 1.8, 2.5, 3.0, 3.9, 5.2

- What if we wanted to know deaths at $x=3.5$?

Interpolation is when we use our linear function to estimate a value at a point *in between* points we have already measured

Interpolation and Extrapolation

$y = 9.607x + 111.958$ *x* is the pollution index

Suppose we came up with this formula as an approximation after measuring fish deaths when the pollution index was: 1.1, 1.8, 2.5, 3.0, 3.9, 5.2

- What if we wanted to know deaths at x=7.0?

Extrapolation is when we use our linear function to estimate a value at a point *outside of* points we have already measured

- Why might this fail?

Interpolation and Extrapolation

In general, interpolation will be more accurate than extrapolation

Sometimes you can intuitively reason about when and why extrapolation will or will not work

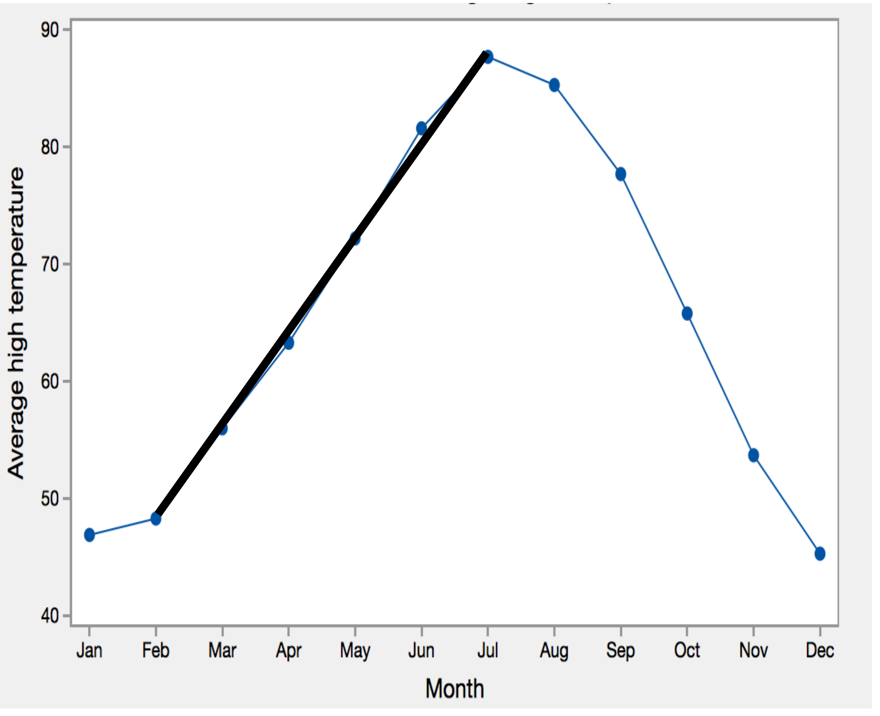
- You have some knowledge of chemistry and biology that there is an upper limit to how much pollution fish can take before they all die

Interpolation and Extrapolation

In general, interpolation will be more accurate than extrapolation

Sometimes you can see what is reasonable by examining your data

- Visually see regions where the line is a good or bad fit



Interpolation and Extrapolation

The average lifespan of Americans has increased over time. The average lifespan of an American in a given year is approximately $y = 0.2x + 73$, where x is the number of years since 1960.

What was the approximate lifespan in 1980?

What will be the approximate lifespan in 2020?

What will be the approximate lifespan in 2400?

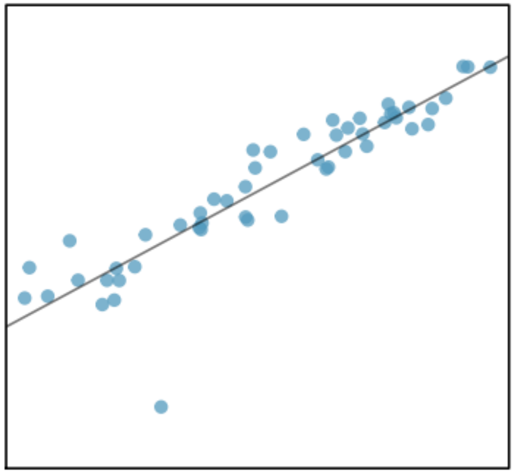
- Unknown in 2400. This model is likely a bad approximation for that large of an interval.

Fitting Linear Functions

Where does a linear function such as "y = 9.607x + 111.958" come from?

Want to pick slope and y-intercept ($y=mx+b$) such that the line is as close as possible to the true data points

- Want to minimize distance from each point to the line
- We'll be more concrete later in the semester



Fitting Linear Functions

The process of picking the parameters of a function (e.g., m and b) to make it is close as possible to a set of data points is **regression**

If the function is linear (i.e., a line) then this is **linear regression**

Statistical software such as MiniTab Express can perform linear regression automatically

Practice

Regression in MiniTab Express.

Differencing

A special type of slope is the **difference** in y-value between consecutive points

- Assuming consistent interval of width 1 along *x*

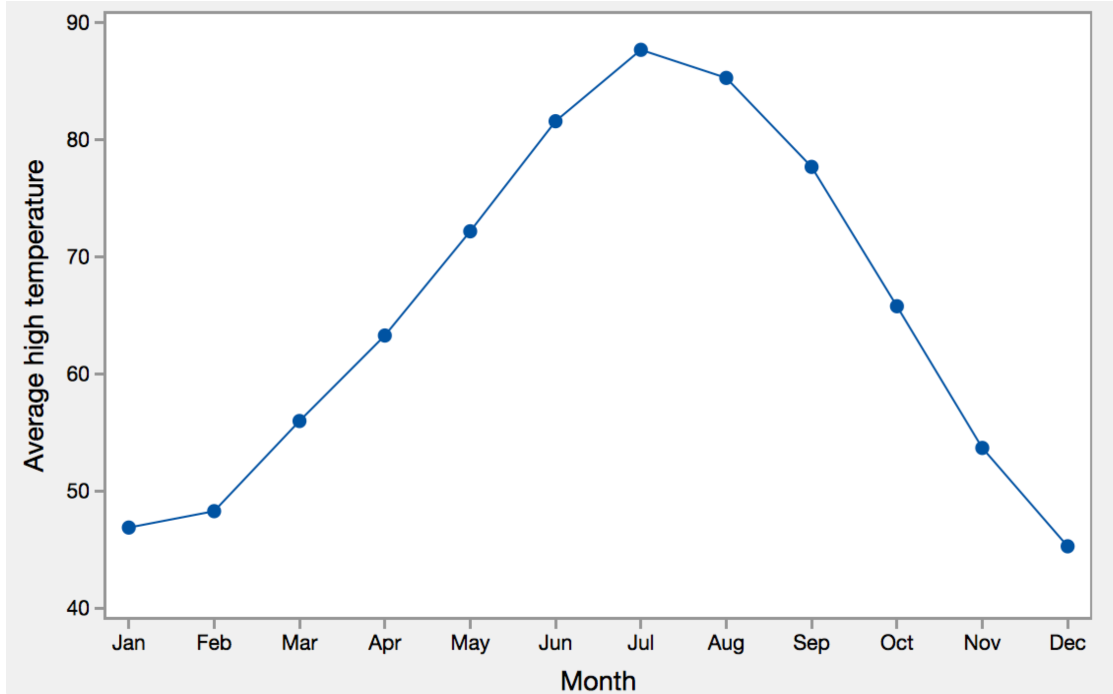
For two adjacent points:

$y_i - y_{i-1}$ e.g., $y_2 - y_1$

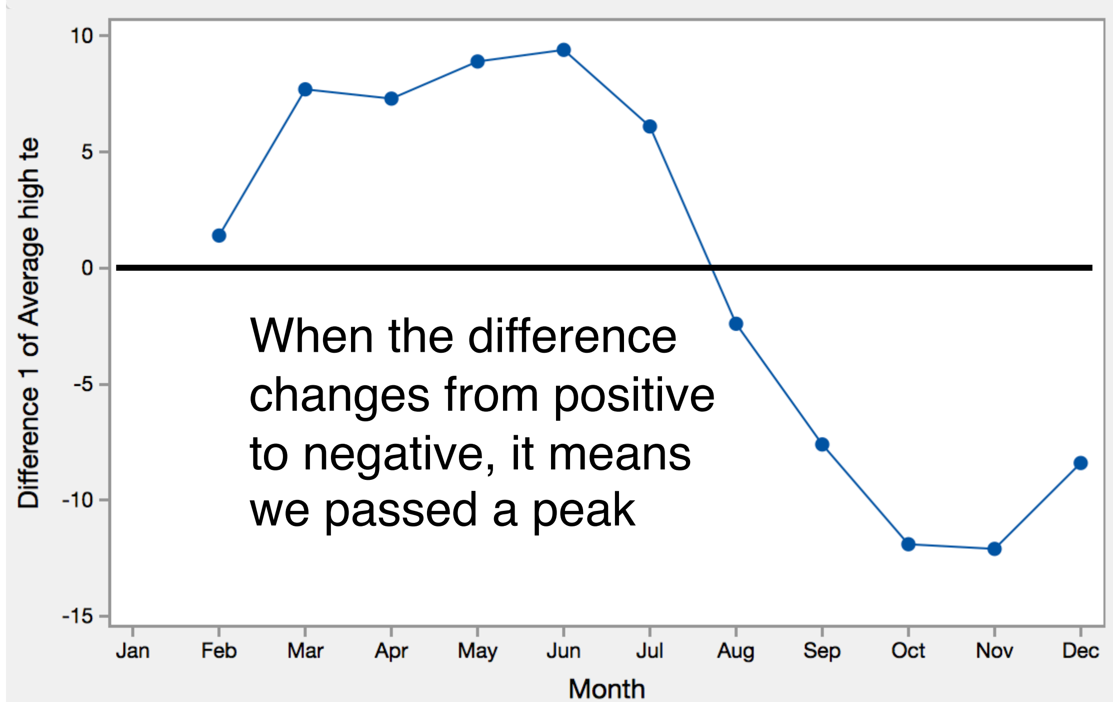
$y_5 - y_4$

The sign of the difference tells you whether it increased or decreased from the previous point

Original:



Difference:



Differencing

Positive difference means increasing

Negative difference means decreasing

Change from positive to negative means there is a peak

Change from negative to positive means there is a trough

Maxima and Minima

Whenever there is a peak in the data, this is a **maximum**

The **global** maximum is the highest peak in the entire data set (the largest y-value)

A **local** maximum is any peak, when the rate of change between to consecutive points (or difference) switches from positive to negative

Maxima and Minima

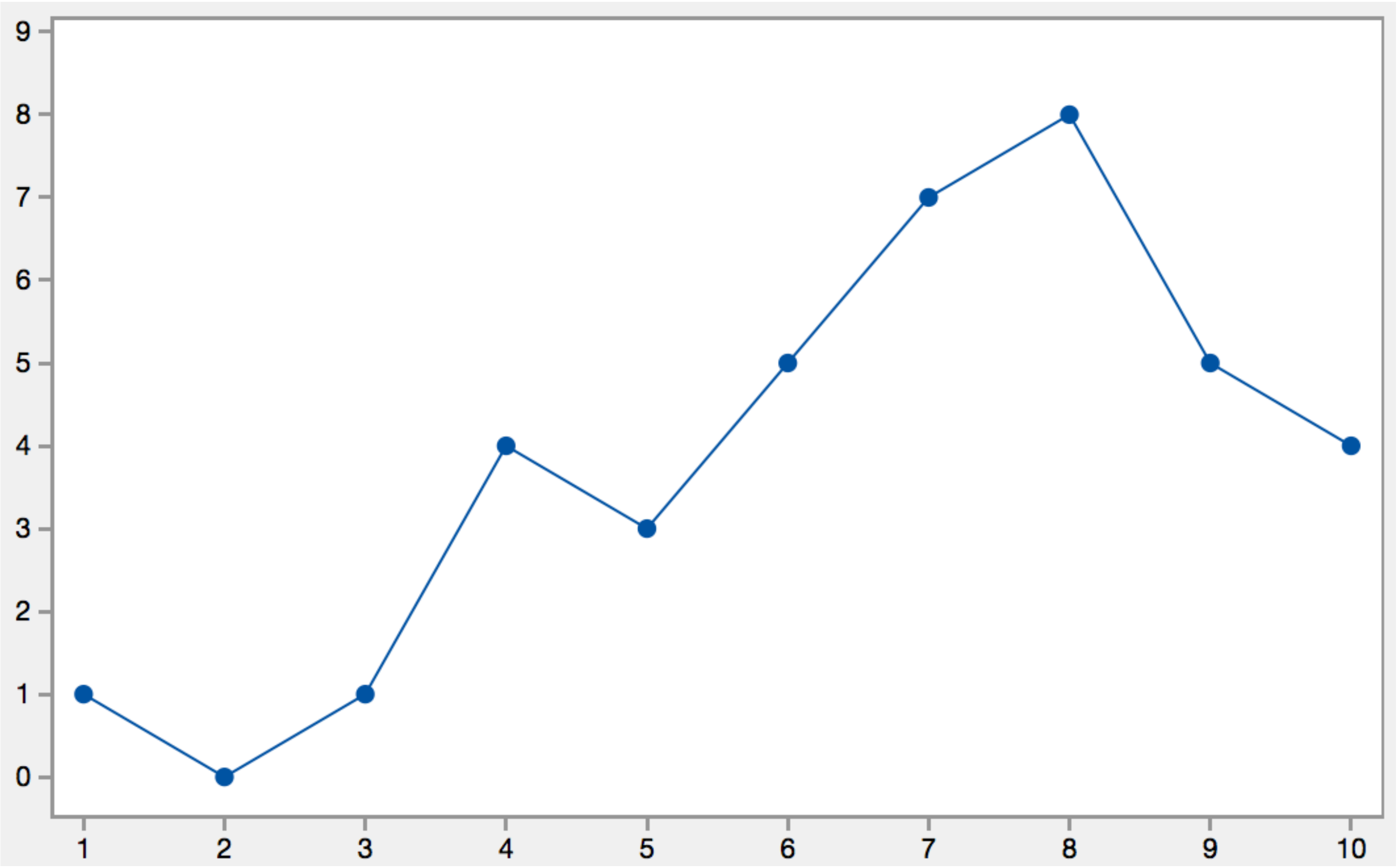
Whenever there is a trough in the data, this is a **minimum**

The **global** minimum is the lowest trough in the entire data set (the smallest y-value)

A **local** minimum is any trough, when the rate of change between to consecutive points (or difference) switches from negative to positive

Practice

Identify all global and local maxima and minima



Residuals

The **residual** of a point (x_i, y_i) is the difference between the true y_i value and the value you estimated based on your best-fit line:

$$e_i = y_i - (mx_i + b)$$

Also referred to as the **error** of your line at that point

The *size* of a residual is its absolute value: $|e_i|$

Residuals

The average residual size can tell you the average error you will make if you estimate new data points (e.g., interpolation or extrapolation)

This is only true if the new data points follow the same pattern as the data you originally observed

- More likely to be true for interpolation than extrapolation

Revisiting Correlation

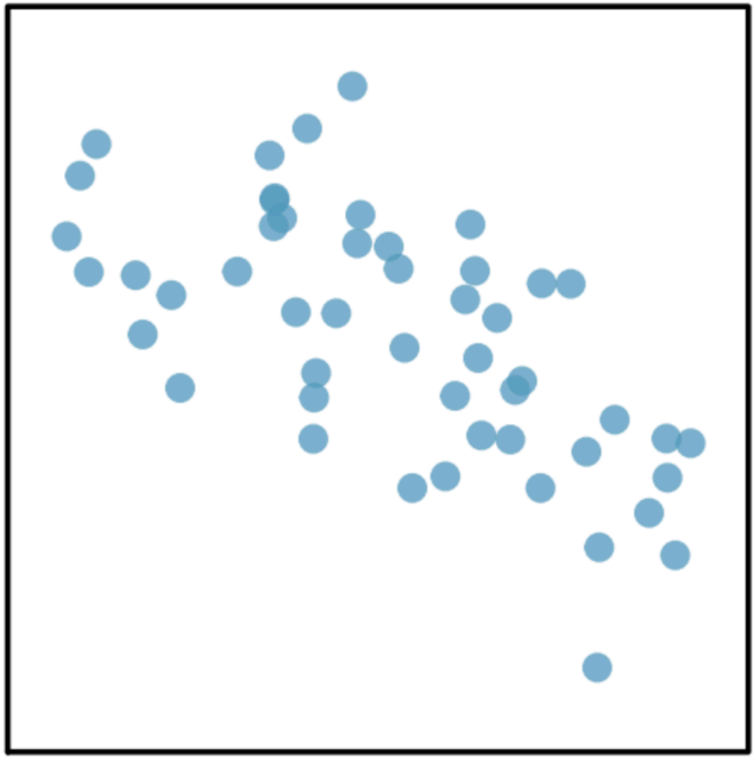
The Pearson correlation measures how strongly data points are related *linearly*

A perfect correlation of 1 or -1 occurs if all residuals from the best-fit line are 0

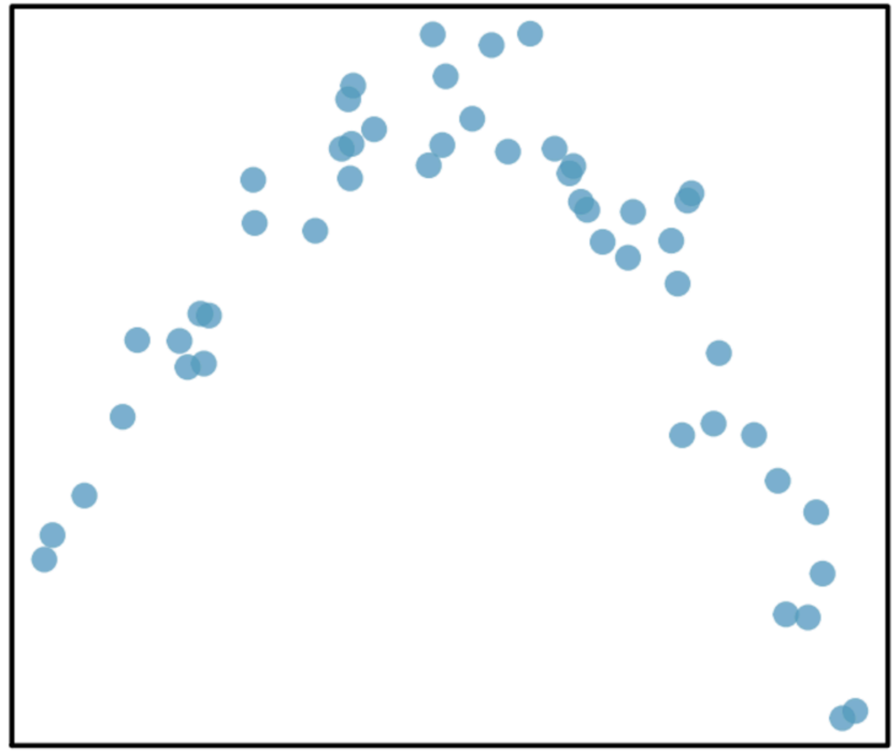
- No error → perfect linear fit
- 1 if slope is positive, -1 if slope is negative

As a rule of thumb: the higher the absolute value of correlation, the smaller residual sizes

Revisiting Correlation



R = −0.64



R = −0.23