Linear Regression
Part 1: Fitting Lines

INFO-1301, Quantitative Reasoning 1

University of Colorado Boulder

April 19, 2017

Prof. Michael Paul

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

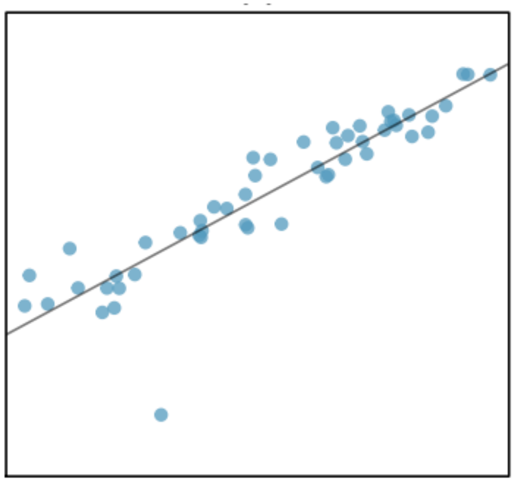
- Why might this fail?

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 next time



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.

Revisiting Correlation

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

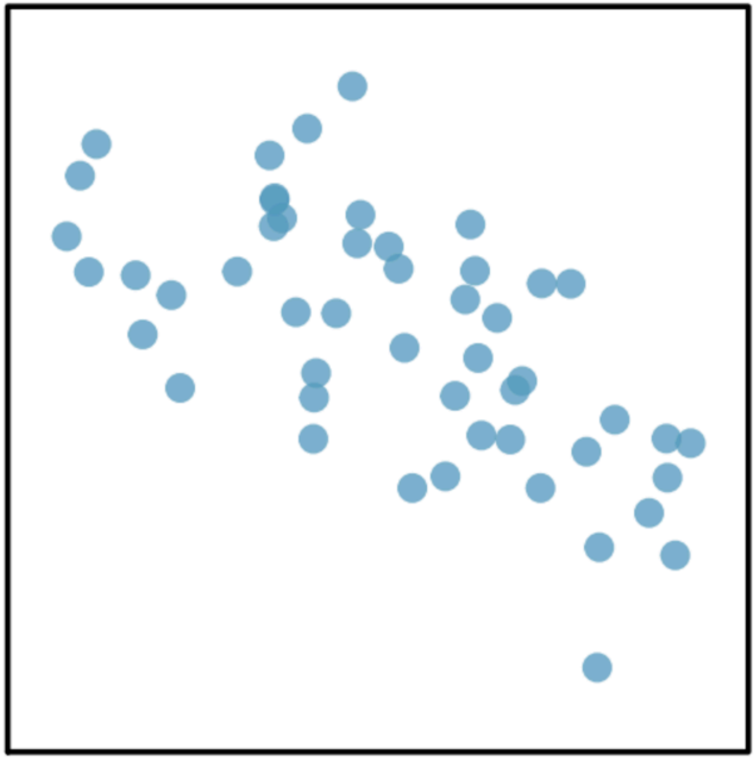
A perfect correlation of 1 or -1 occurs the best-fit line exactly matches every data point

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

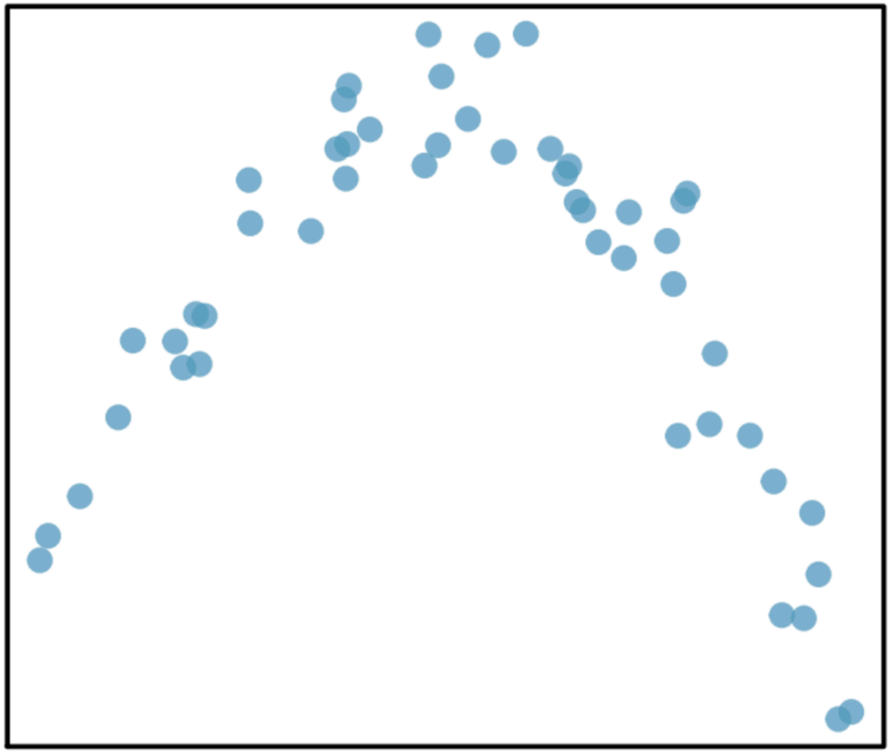
The size of the correlation tells you how well the points can be approximated with a line

The sign of the correlation tells you the slope

Revisiting Correlation



R = –0.64



R = –0.23

R squared

A common metric for measuring the quality of the fit of a line is called R^2

This is the square of the correlation (sometimes called R) between the true Y values and the Y values that you estimate with the y=mx+b line