The next few days

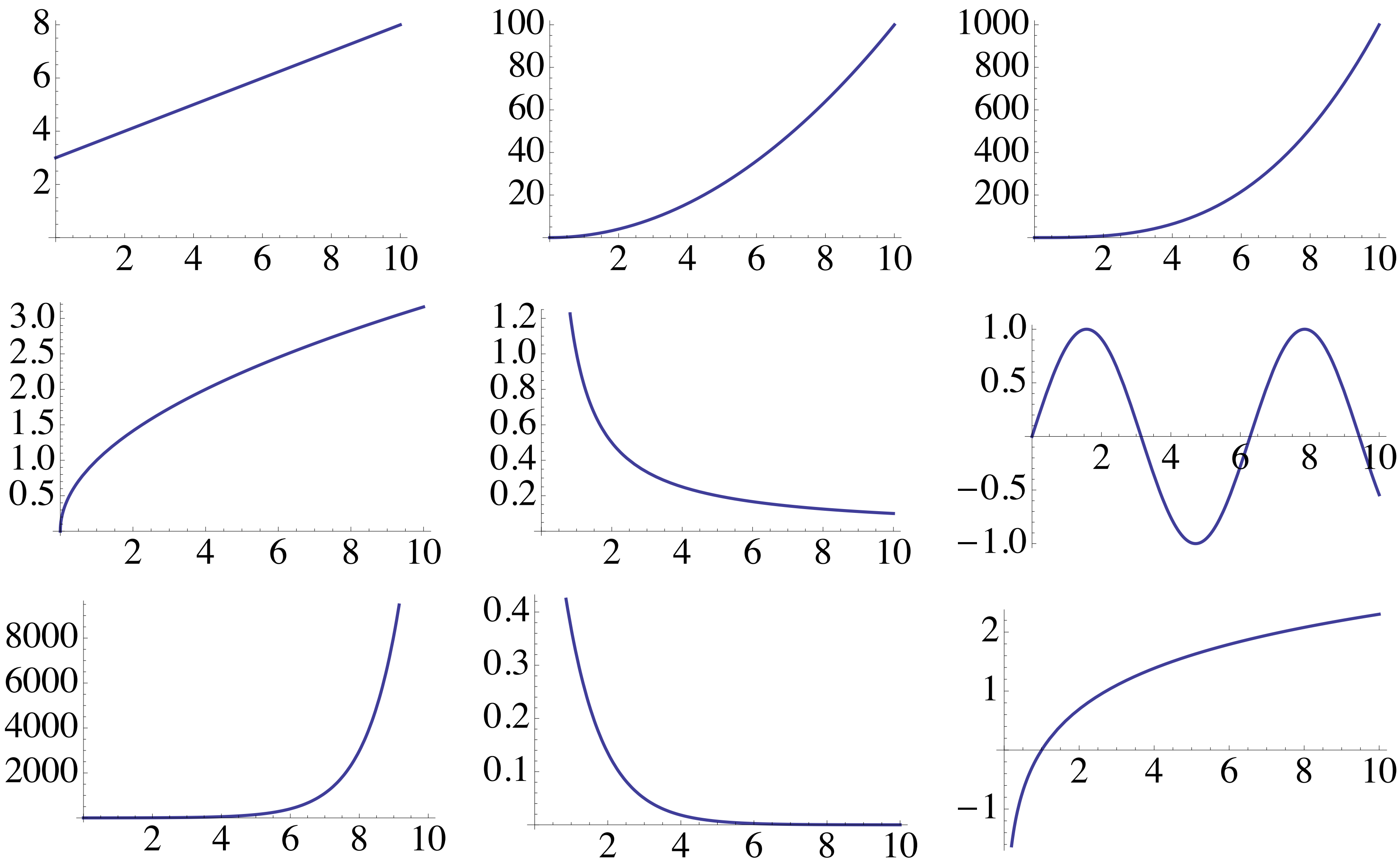
Goals: Understand function fitting, introduce *Mathematica*

Frame of reference:

- Formulation. Suppose the problem has been properly formulated.
  - Problem statement is precise and clear.
  - Dependent variable(s) and independent variable determined.
- Now we need a mathematical model; one type is a function.
  - We collect data*, plot it, and notice a pattern. $y \approx Cx^k$???
  - Simplifying assumption: The independent variable is a (simple) function of the dependent variables.
- Math. Manipulation. Determine the best function of this type.
  - Now: Visually. Later: Using a computer
- Evaluation. Does this function fit the data well?

★ For a real world question, there is more evaluation to do.

Functions you should recognize on sight



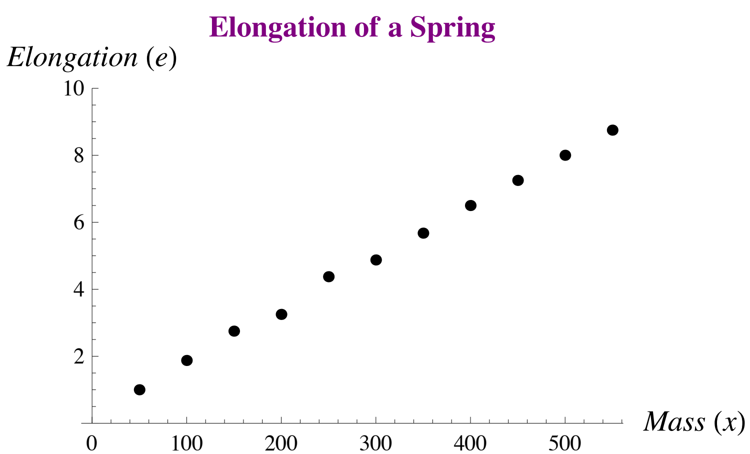
TWPS: What are these functions? What is the most general equation?

Springs and Elongations

Example: Modeling Spring Elongation

Take your favorite spring. Attach different masses.
How much does it stretch from rest? [Its **elongation**.]

When we plot the data, we get the following **scatterplot**.



What do you notice? ______________________________

mass x	elong y
50	1.000
100	1.875
150	2.750
200	3.250
250	4.375
300	4.875
350	5.675
400	6.500
450	7.250
500	8.000
550	8.750

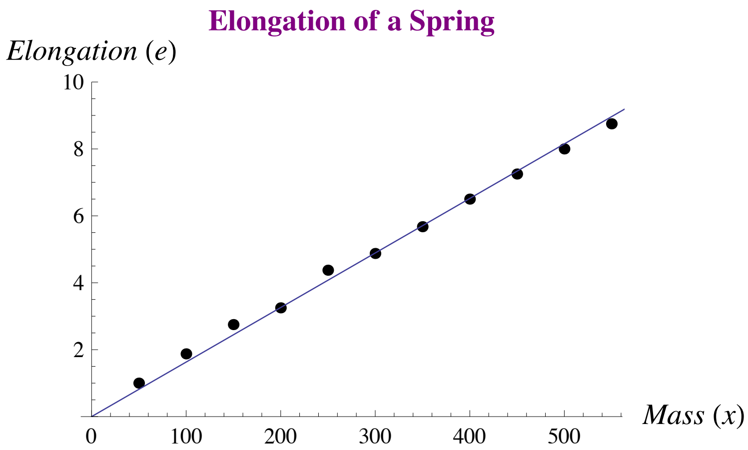
Proportionality

When data seems to lie on a line through the origin, we expect the two variables to be **proportional**; in this case, $y = kx$ for some constant k.

We need to find this **constant of proportionality** k.

So: Estimate the slope of the line. **How?**

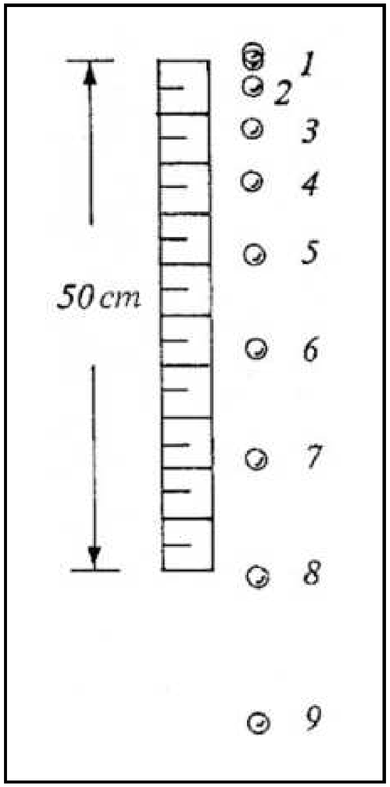
1. Guesstimating



2. Mathematically: **Linear Regression / Least Squares**
   (For another day)

Fitting Gravity Data

Example. Modeling the dropping of a golf ball

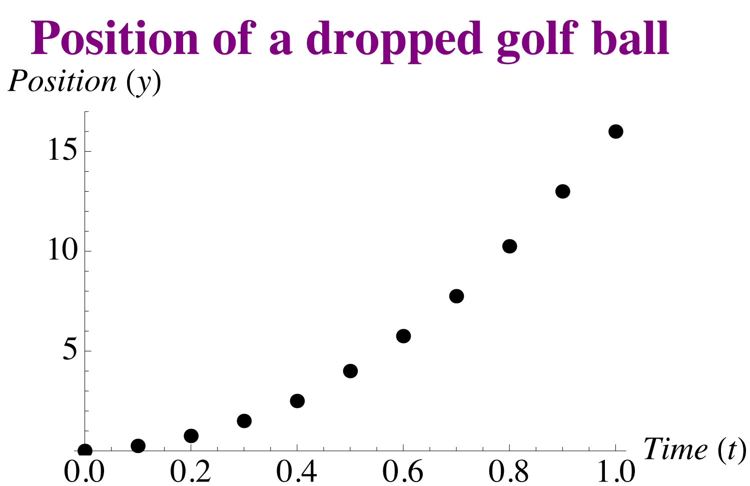


Source: practicalphysics.org

Let's use an experiment to test the gravity model from last time.

Use a camera to record the position every tenth of a second.

Data would be similar to the table→
It's plotted in the scatterplot below.



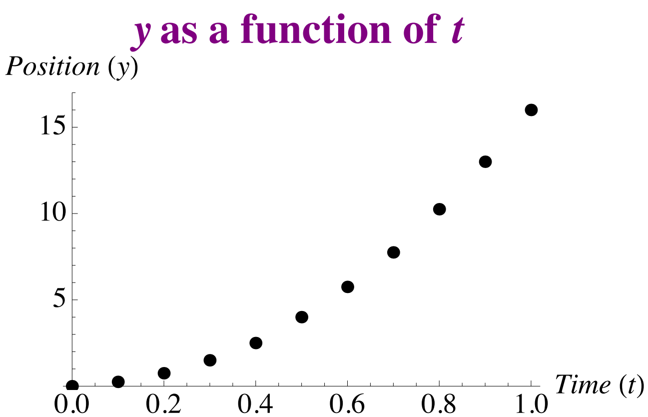
t	y
0.0	0.00
0.1	0.25
0.2	0.75
0.3	1.50
0.4	2.50
0.5	4.00
0.6	5.75
0.7	7.75
0.8	10.25
0.9	13.00
1.0	16.00

[Ignore data on p. 25.]
[It's BAD data.]

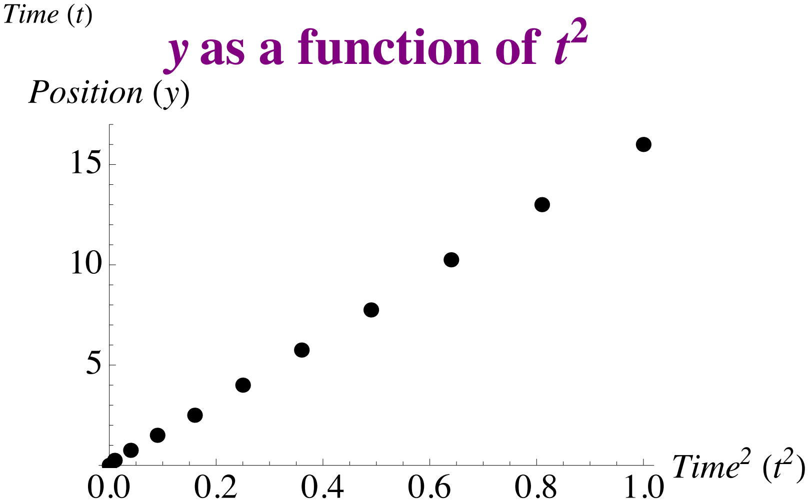
Fitting Gravity Data

These data seem to fit a $\underset{\text{(type of function)}}{\underline{\hspace{4cm}}}$. How can we be sure?

1. Plot distance as a function of t^2.
   Next, estimate the constant of proportionality.



transform



t	t^2	y
0.0		0.00
0.1		0.25
0.2		0.75
0.3		1.50
0.4		2.50
0.5		4.00
0.6		5.75
0.7		7.75
0.8		10.25
0.9		13.00
1.0		16.00

This implies
$y \approx \_\_ t^2$.

Fitting Gravity Data

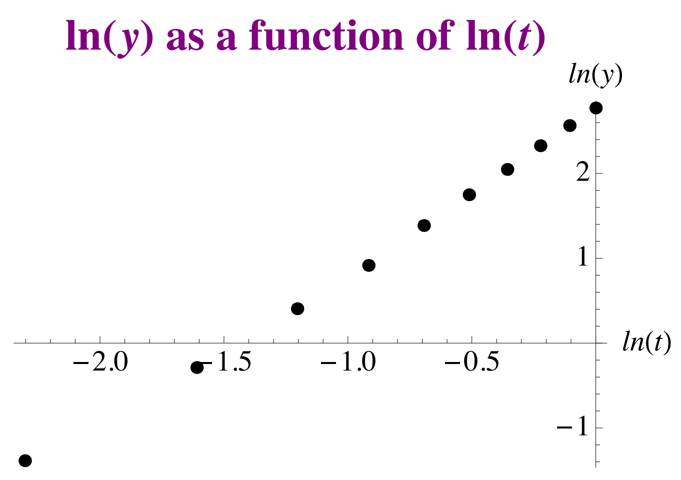
When fitting data to a function $y = Ct^k$, an alternate method is:

2 ⋆ Plot the log of distance as a function of log of time. ⋆

- WHY? Suppose $y = Ct^k$. Taking a logarithm of both sides,
  $\ln y = \ln(Ct^k) =$

Conclusion: To approximate C and k,

- First, calculate $\ln y$ and $\ln t$ for each datapoint.
- Fit the transformed data to a line.
  - The slope is an approximation for k.
  - The y-intercept approximates $\ln C$.



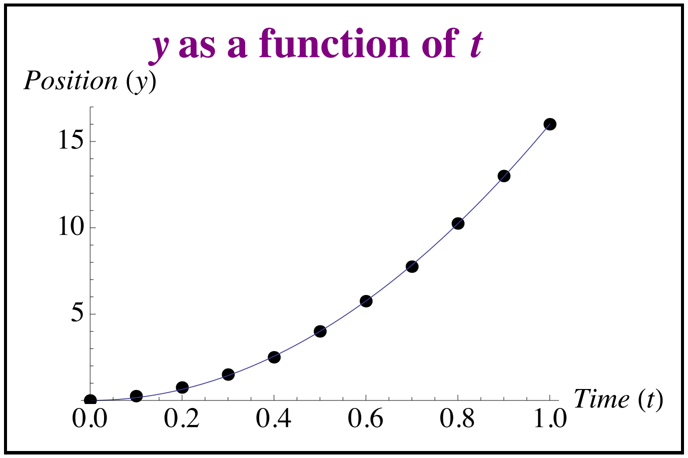
$\ln y \approx 2 \ln t + 2.8$

Fitting Gravity Data

We have determined that our gravity model

$$y(t) = 16t^2$$

appears to model the dropping of a golf ball.



Example. Raindrops—Our model gives their position as $y(t) = 16t^2$.

A raindrop falling from 1024 feet would land after $t = 8$ seconds.

However, an experiment shows that the fastest drop takes 40 seconds, and that drops fall at different rates depending on their size.

Even if we have a good model for one situation doesn't mean it will apply everywhere. **We always need to question our assumptions.**

—Extensive gravity discussion in Section 1.3.—

Modeling Population Growth

Example. Modeling the size of a population.

We would like to build a **simple** model to predict the size of a population in 10 years.

- A very **macro**-level question.

Definitions: Let t be time in years; $t = 0$ now.
$P(t)$ = size of population at time t.
$B(t)$ = number of births between times t and $t+1$.
$D(t)$ = number of deaths between times t and $t+1$.

Therefore, $P(t+1) =$ ____________________.

Definitions imply

$P(4) =$
$B(\frac{1}{2}) =$
$B(5) - D(5)$
$=$

Assumption: The birth rate and death rate stay constant.

That is, the birth rate $b = \frac{B(t)}{P(t)}$ and death rate $d = \frac{D(t)}{P(t)}$ are constants.

Assumption: No migration.

Population Growth

Therefore, $$P(t+1) = P(t)\left[\frac{P(t)}{P(t)} + \frac{B(t)}{P(t)} - \frac{D(t)}{P(t)}\right].$$

Under our assumptions, $$P(t+1) = P(t)[1+b-d].$$

$$\begin{aligned} \text{This implies: } P(1) &= \underline{\hspace{4cm}}, \\ P(2) &= \underline{\hspace{4cm}}, \ldots \\ \text{In general, } P(n) &= \underline{\hspace{4cm}}. \end{aligned}$$

Definition. The **growth rate** of a population is $r = (1+b-d)$. This constant is also called the **Malthusian parameter**.

A model for the size of a population is
$$P(t) = P(0)r^t,$$
where $P(0)$ and r are constants.

Applying the Malthusian Model

Approximate US Population at: http://www.census.gov/main/www/popclock.html

Example 1. Suppose that the current US population is 315,400,000. Assume that the birth rate is 0.02 and the death rate is 0.01. What will the population be in 10 years?

Answer. Use $P(t) = P(0)r^t$:

Refinement. Approx. US Growth Rate at http://www63.wolframalpha.com/input/?i=US+birth+rate

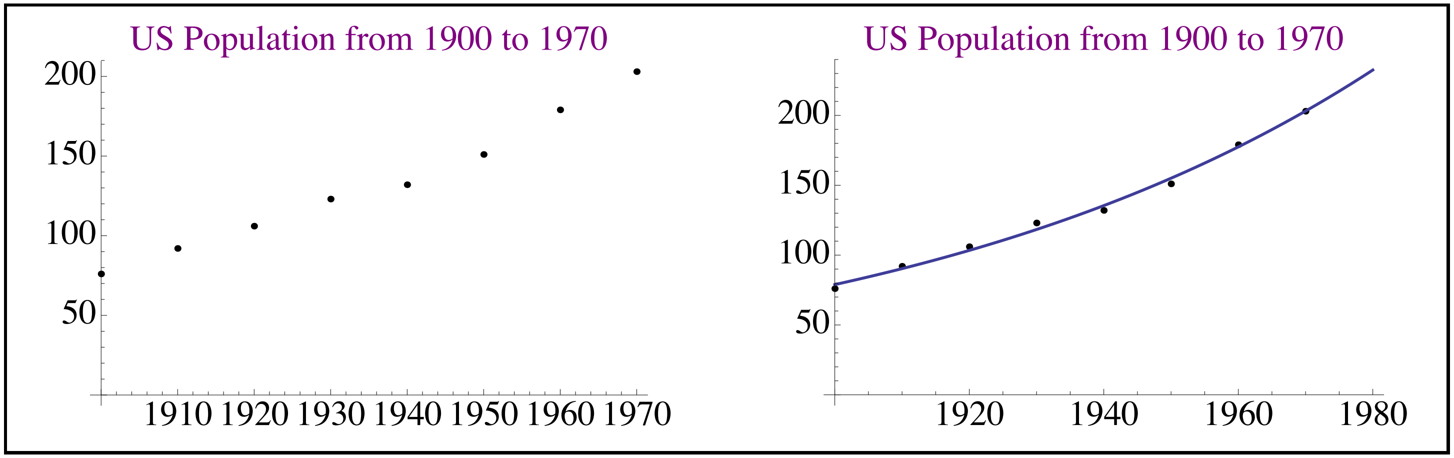
Resource: Wolfram Alpha, integrable directly into *Mathematica*.

Example 2. How long will it take the population to double?

Answer. Use $P(t) = P(0)r^t$:

Determining constants of exponential growth

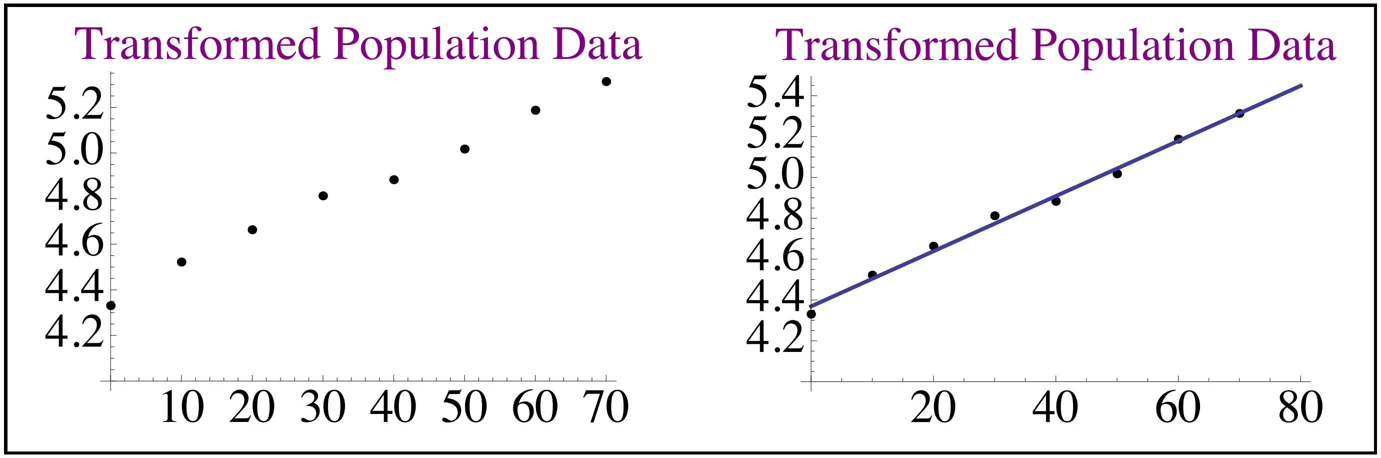
Goal: Given population data, determine model constants.



- Take the logarithm of both sides of $P(t) = P(0)r^t$.
- We have $\ln[P(t)] =$ ____________________.
- A linear fit for $P(t)$ vs. t gives values for _________ and ______.
- Exponentiate each value to find the values for $P(0)$ and r.

Determining constants of exponential growth

Here we plot $\ln[P(t)]$ as a function of t:



The line of best fit is approximately $\ln[P(t)] = 4.4 + 0.0135t$.

Therefore our model says $P(t) \approx e^{4.4}(e^{0.0135})t = 81.5 \cdot (1.014)^t$.

Analysis:
- History indicates we should split the interval $[1900, 1970]$.
- We have to be careful when trying to **extrapolate**!

$\star$ Important: Transformations distort distances between points, so verification of a fit should always take place on y versus x axes. $\star$

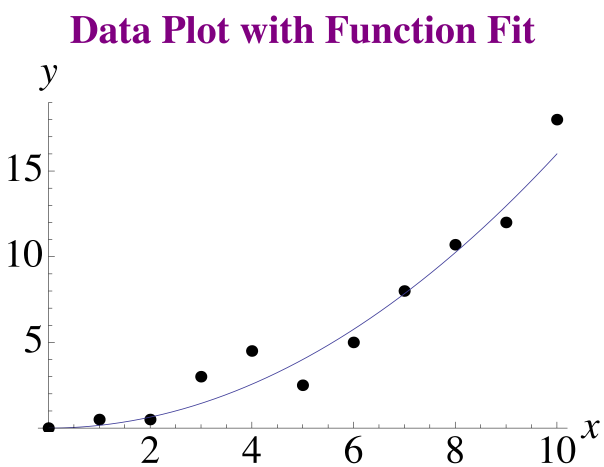
Residuals

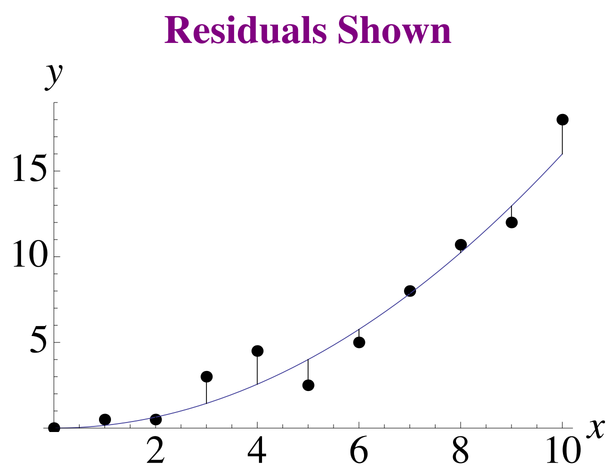
Once you determine a function of best fit, then you should verify that it fits well. One way to do this is to look at the residual plot.

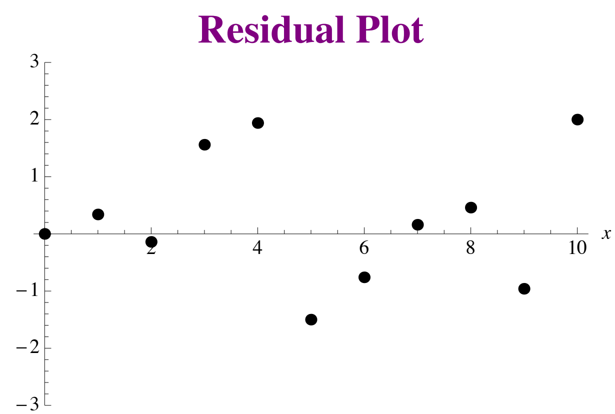
Definition: Given a point (x_i, y_i) and a function fit $f(x)$, the **residual** r_i is the error between the actual and predicted values.

Mathematically, $r_i = y_i - f(x_i)$.

Definition: A **residual plot** is a plot of the points (x_i, r_i).



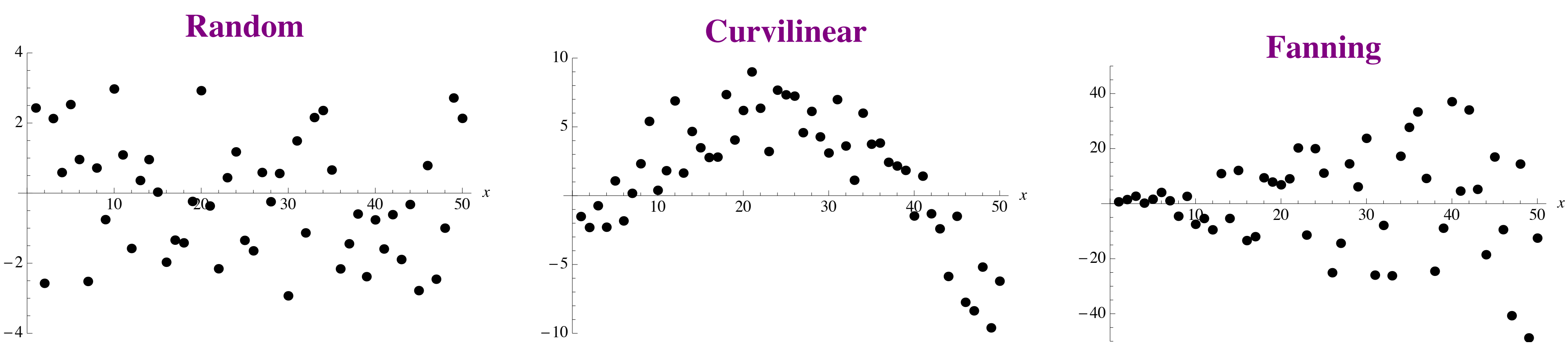




Residuals

The structure of the points in the residual plot give clues about whether the function fits the data well. Three common appearances:

1. **Random** : Residuals are randomly scattered at a consistent distance from axis. Indicates a good fit, as on previous page.
2. **Curvilinear**: Residuals appear to follow a pattern. Indicates that some aspect of model behavior is not taken into account.
3. **Fanning**: Residuals small at first and get larger (or vice versa). Indicates non-constant variability (model better for small x?).



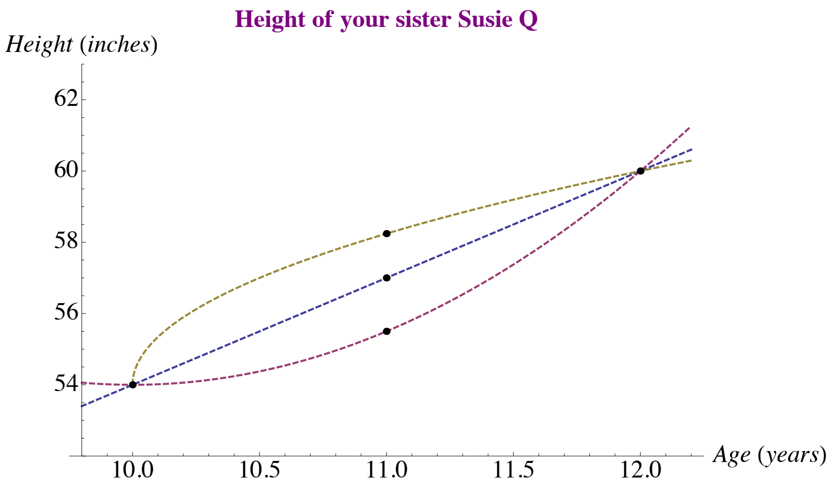
Interpolation vs. Extrapolation

Suppose you have collected a set of *known* data points (x_i, y_i), and you would like to estimate the y-value for an *unknown* x-value.

The name for such an estimation depends on the placement of the x-value relative to the *known* x-values.

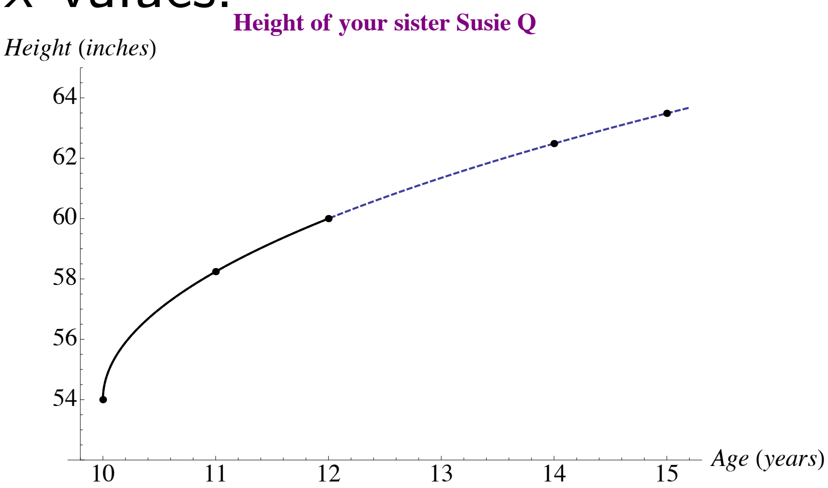
Interpolation

Inserting one or more x-values between known x-values.



Extrapolation

Inserting one or more x-values outside of the range of known x-values.

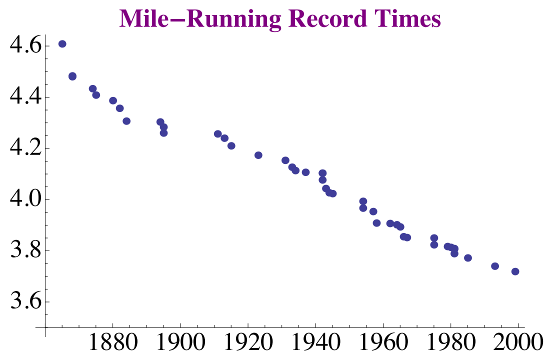


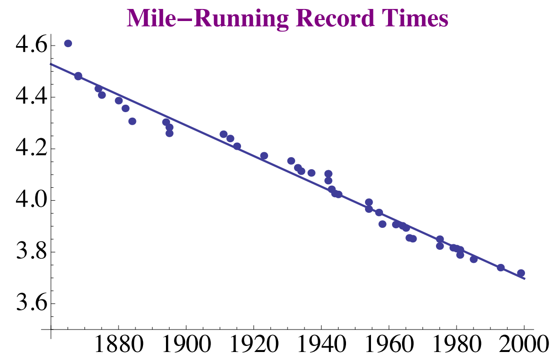
Interpolation vs. Extrapolation

- The most common method for interpolation is taking a weighted average of the two nearest data points; suppose $x_1 < x < x_2$, then,
$$f(x) \approx y_1 + \frac{y_2 - y_1}{x_2 - x_2}(x - x_1).$$
- In both interpolation and extrapolation, when you have a function f that is a good fit to the data, simply plug in $y = f(x)$.
- Confidence in approximated values depends on confidence in your data and your model.
- Confidence in extrapolated data is higher when closer to the range of known x-values.

Extrapolation: Running the Mile (p. 162)

Below is a plot of the years in which a record was broken for running a mile and the record-breaking time.





The data appears to fit the line $T(t) = 15.5639 - 0.00593323t$.

Solve for $T(t) = 0$: You get $t \approx 2623$.

Conclusion: In the year 2623, the record will be zero minutes!

- Note the lack of realistic assumptions behind the data.
- Always be careful when you extrapolate!