Avoiding Using Least Squares

Justification for fitting data visually:

- Large simplifications in model development mean that eyeballing a fit is reasonable.
- Mathematical methods do not necessarily imply a *better* fit!
- You can make objective judgements that computers cannot; you know which data points should be taken more seriously.
- Mathematics give precise answers; every answer is fallible.

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 $t + 1$.
and therefore, $P(t + 1) =$ ______________________.

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 assumption, $$P(t+1) = P(t)[1+b-d].$$

This implies: $P(1) = \underline{\hspace{4cm}}$,

$P(2) = \underline{\hspace{4cm}}, \ldots$

In general, $P(n) = \underline{\hspace{4cm}}$.

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

Example 1. Suppose that the current US population is 308,690,000. Assume that the birth rate is 0.02 and the death rate is 0.01. What will the population be in 10 years?

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

Refinement. Approximate US Growth Rate

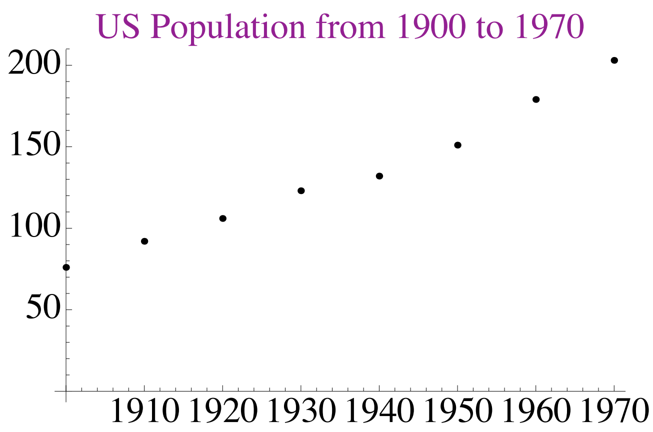
Resource: Wolfram Alpha, integrable directly into *Mathematica*.

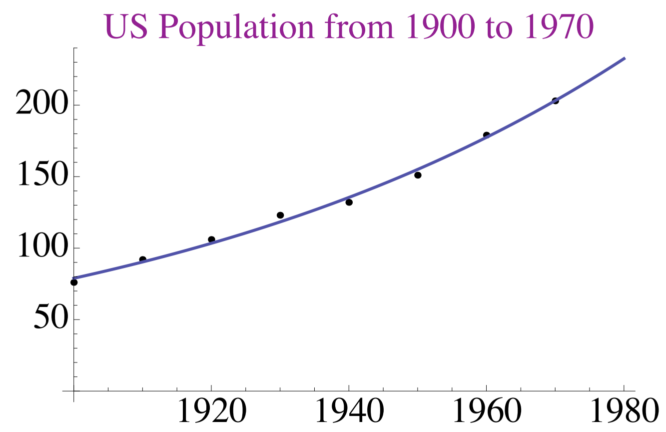
Example 2. How long will it take the population to double?

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

Determining constants of exponential growth

Goal: Given population data, determine model constants.

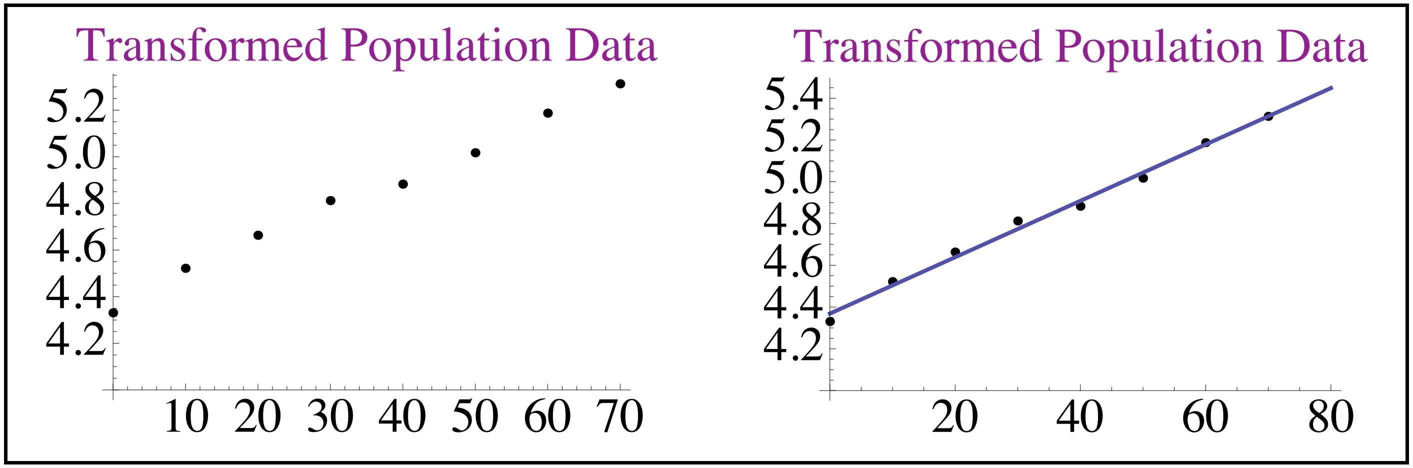




- Take the logarithm of both sides of $P(t) = P(0)r^t$.
- We have $\log[P(t)] =$ ________________________.
- A linear fit for $P(t)$ vs. t gives values for __________ and ______.
- Exponentiate each value to find the values for $P(0)$ and r.

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 $P(t) \approx e^{4.4+0.0135t} = 81.5 \cdot (1.014)^t$.

Analysis: ▶ History indicates we should split the interval $[1900, 1970]$.

▶ What might go wrong when trying to **extrapolate**?

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

Regression

We've done a bunch of curve fitting using our eyes.

If we have confidence in our data, we may wish to do a **regression**,
a method for fitting a curve through a set of points
by following a goodness-of-fit criterion.

Goal: Formulate mathematically what we do internally:
Make the discrepancies between the data and the curve small.

- Make the sum of the set of **absolute deviations** small. (Pic!)
  minimize over all f the sum: $\displaystyle\sum_{(x_i,y_i)} \left|y_i - f(x_i)\right|$

- Make the largest of the set of absolute deviations small.
  minimize over all f the value: $\displaystyle\max_{(x_i,y_i)} \left|y_i - f(x_i)\right|$

One or the other might make more sense depending on the situation.

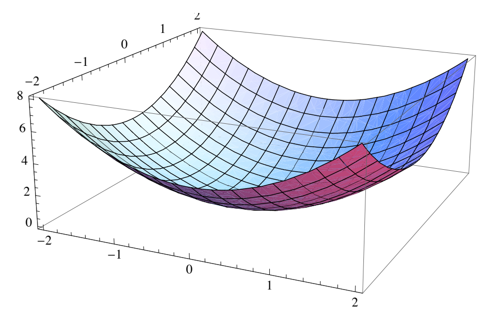
Least Squares

A regression method often used is called *least squares*.

$$\text{minimize over all } f \text{ the sum: } \sum_{(x_i, y_i)} \big(y_i - f(x_i)\big)^2$$

- A middle ground, giving weight to all discrepancies and more weight to those that are further from the curve.
- Easy to analyze mathematically because this is a smooth function.

Calculating minima of smooth functions:



- Differentiate with respect to each variable, and set equal to zero.
- Solve the resulting system of equations.
- Check to see if the solutions are local minima.

Least Squares Example

Example. Use the least-squares criterion to fit a line $y = mx + b$ to the data: $\{(1.0, 3.6), (2.1, 2.9), (3.5, 2.2), (4.0, 1.7)\}$.
Intution / Expectations?

Solution. We need to calculate the sum $S = \sum_{(x_i, y_i)} \left[y_i - (mx_i + b)\right]^2$.
$S = (3.6–1.0m–b)^2 + (2.9–2.1m–b)^2 + (2.2–3.5m–b)^2 + (1.7–4.0m–b)^2$

Expanding, $S = 29.1 - 20.8b + 4b^2 - 48.38m + 21.2bm + 33.66m^2$

Calculating the partial derivatives and setting equal to zero:

$$\begin{cases} \frac{\partial S}{\partial b} = -20.8 + 8b + 21.2m = 0 \\ \frac{\partial S}{\partial m} = -48.38 + 21.2b + 67.32m = 0 \end{cases}$$

Solving the system of equations gives: $\{b = 4.20332, m = -0.605027\}$

That is, the line that gives the least-squares fit for the data is

$$y = -0.605027x + 4.20332.$$

Notes on the Method of Least Squares

- Least squares becomes messy when there are many data points.
- We chose least squares because it was easy. Is it really the "right" method for the job?
- Least squares isn't always easy, for example: $y = Ce^k$.
- You can use least squares on transformed data, but the result is NOT a least-squares curve for the original data.
- Multivariable least squares can also be done: $w = ax + by + cz + d$ (Would want to minimize: ______________________________.)
- Least squares measures distance vertically. A better measure would probably be perpendicular distance.
- You need to understand the concept of least squares and know how to do least squares by hand for small examples.
- We'll learn how to use *Mathematica* to do this for us!