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.

Regression

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

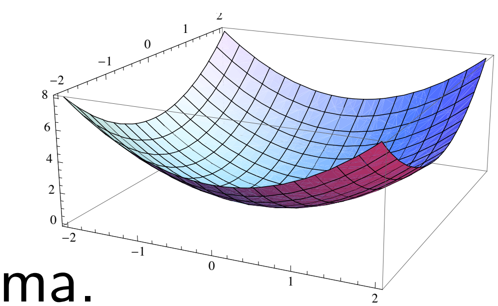
A regression method often used is called *least squares*.

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

- 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: (You know how!)

- 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: $y = Ce^{kx}$.
- 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!

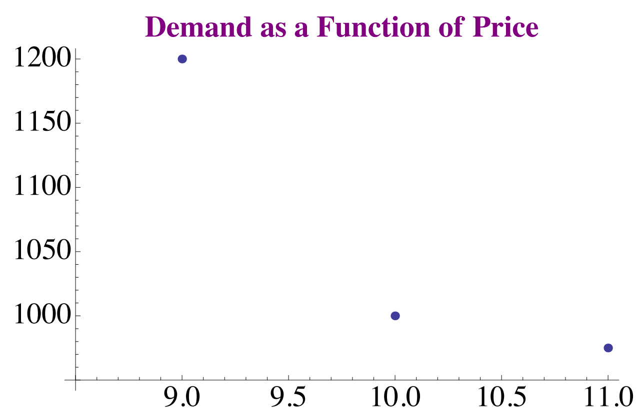
Price – Demand Curve (p. 111–114)

Example. A company is trying to determine how demand for a new product depends on its price and collect the following data:

price p	\$9	\$10	\$11
demand d	1200/mo.	1000/mo.	975/mo.

The company has reason to believe that price and demand are **inversely proportional**, that is, $d = \frac{c}{p}$ for some constant c.

$\rightarrow$ Use the method of least squares to determine this constant c.



Price – Demand Curve (p. 111–114)

Solution. Since $f(p) = \dfrac{c}{p}$, then the sum $S = \sum_{(p_i, d_i)} \left[d_i - \left(\dfrac{c}{p_i} \right) \right]^2$.
Specifying datapoints gives

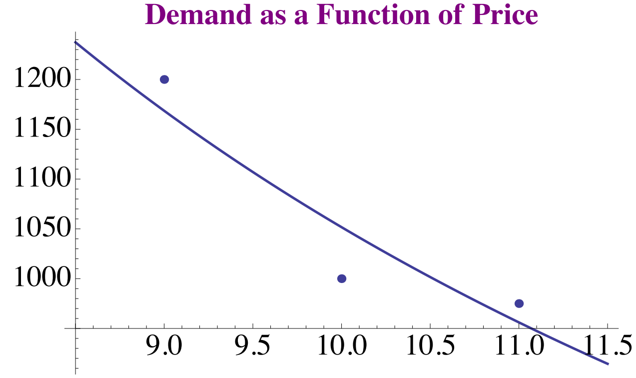
$$S = \left[1200 - \frac{c}{9}\right]^2 + \left[1000 - \frac{c}{10}\right]^2 + \left[975 - \frac{c}{11}\right]^2$$

Setting the derivative equal to zero gives

$$\frac{dS}{dc} = \frac{-2}{9}\left[1200 - \frac{c}{9}\right] + \frac{-2}{10}\left[1000 - \frac{c}{10}\right] + \frac{-2}{11}\left[975 - \frac{c}{11}\right] = 0$$

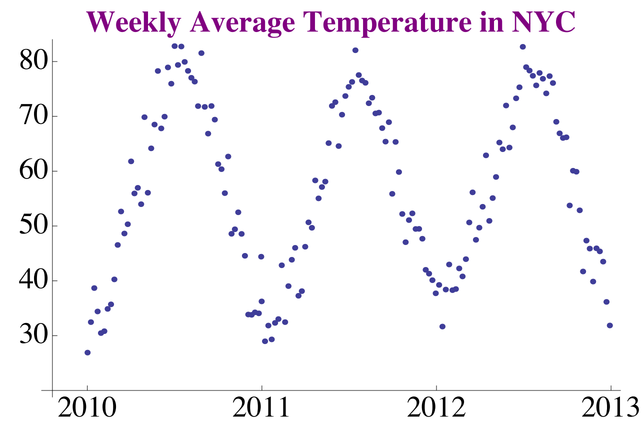
Solving for c gives $c \approx 10517$.

Demand as a Function of Price



New York City Temperature (similar to p. 158)

The graph of average weekly temperature in New York City from Jan. 2010 to Dec. 2012 gives the distinct impression of a ________.

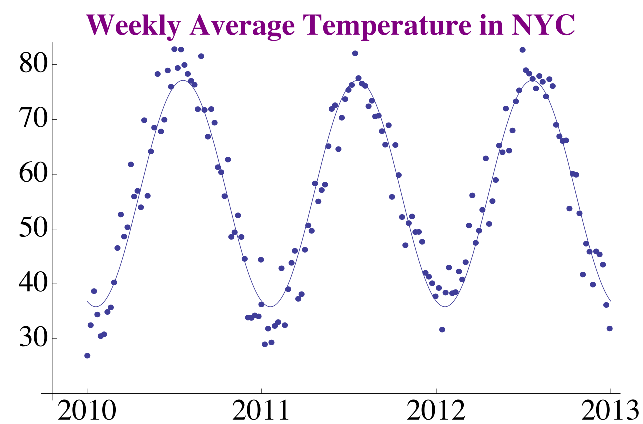


We need to determine the constants in:

$$Temp(t) = A + B\sin(C(t - D)).$$

Mathematica has a hard time finding all four constants at once. Using knowledge of the seasons, we can make assumptions about C and D. We can assume that $C =$ ____.

For D, find when the sine passes through zero. Since January is coldest and July is hottest, the zero should occur in April; guess $D \approx 0.3$.



Fitting to $Temp(t) = A + B\sin[2\pi(t - 0.3)]$ gives: $Temp(t) = 56.5 + 20.6\sin[2\pi(t - 0.3)]$