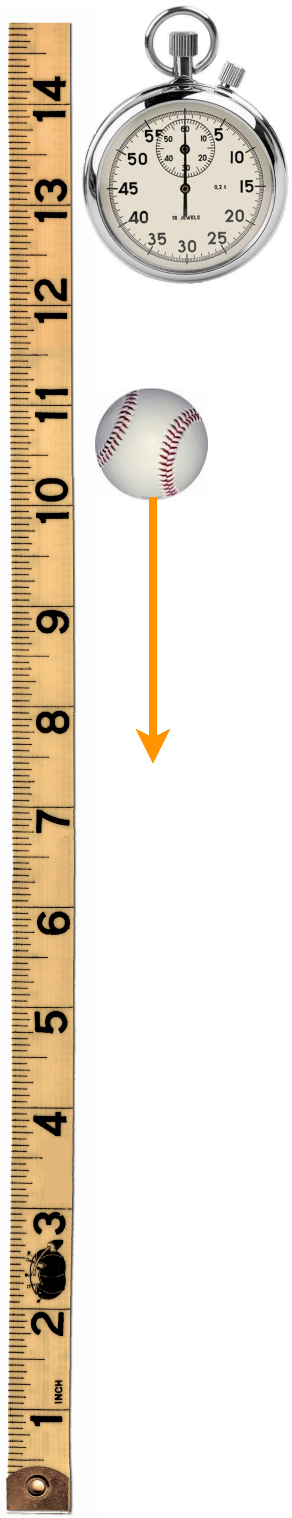
Linear Regression

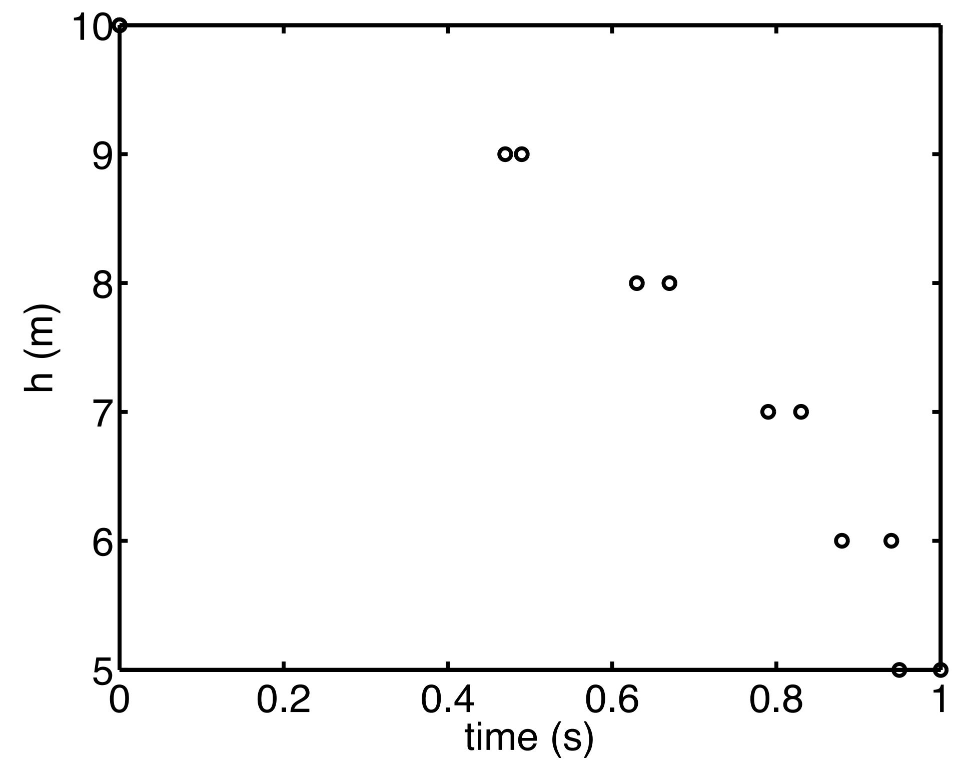
CHEN 1703

THE UNIVERSITY OF UTAH

Motivating Example - Gravity



Experiment 1			Experiment 2	
t (s)	h (m)		t (s)	h (m)
0	10		0	10
0.49	9		0.47	9
0.63	8		0.67	8
0.83	7		0.79	7
0.88	6		0.94	6
0.95	5		1.0	5



$v_o = 0$

From theory: $h = h_o + \cancel{v_o} t + \frac{1}{2} g t^2$

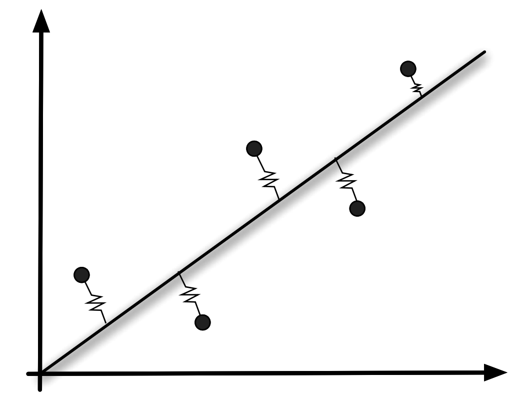
Can we use this data to find the value of g?

(note g=9.80 m/s^2 on earth)

Linear *Least-Squares* Regression

Problem: given $f(x_i)$ at points x_i, we want to find some constant(s) in $f(x)$ to "best" fit the data.

Solution: try to write this as a *linear* problem.



A linear problem, $y=ax+b$, with m observations that we want to use to determine the "best" a and b:

$$
\begin{aligned}
y_1 &= a\,x_1 + b \\
y_2 &= a\,x_2 + b \\
&\vdots \\
y_m &= a\,x_m + b
\end{aligned}
\quad \Longrightarrow \quad
\underbrace{\begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_m & 1 \end{bmatrix}}_{A}
\underbrace{\begin{pmatrix} a \\ b \end{pmatrix}}_{\phi}
=
\underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}}_{b}
$$

m equations,
2 unknowns (a,b)
(cannot solve it)

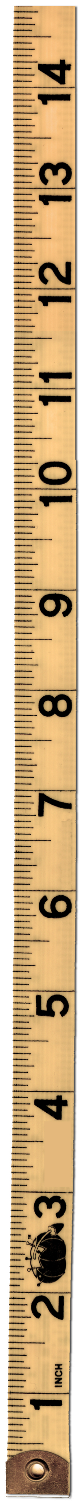
Re-cast this system to minimize the error between $f(x)$ and the observations (x_i,y_i)...

$$\boxed{A^\mathsf{T} A\phi = A^\mathsf{T} b}$$

"Normal" Equations

```
x = [ 0.0 1.0 2.0 3.0 ];
y = [ 0.1 0.9 2.2 2.9 ];

n = length(x);

A = [x' ones(n,1)];
b = y';

AA = A'*A;
bb = A'*b;

phi=AA\bb;
```

Gravity Example Revisited



$$h = h_o + \frac{1}{2} g t^2$$

Find the value of g from this data.

h_o is known precisely.

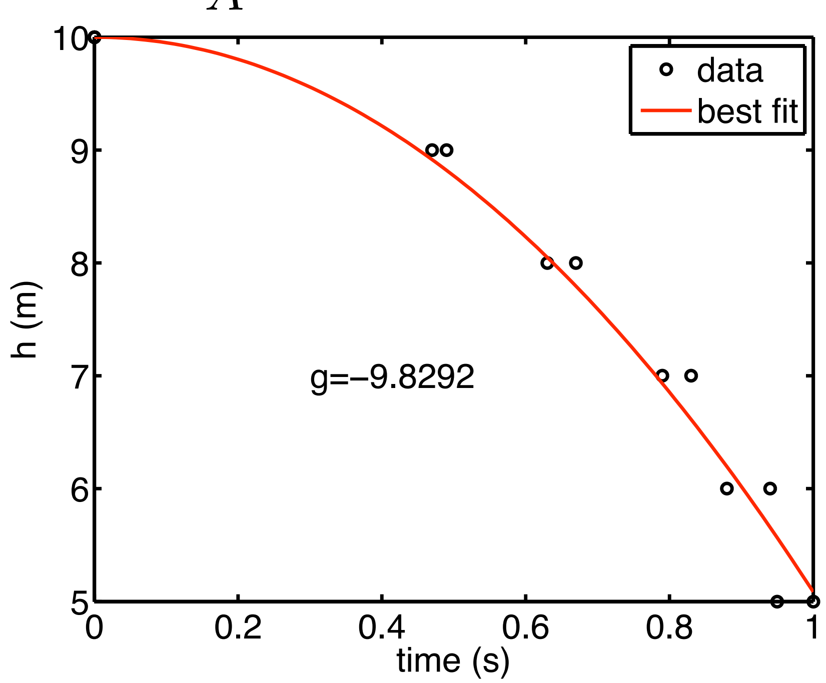
$$y = ax$$

$$y = h - h_o, \quad x = \frac{1}{2} t^2, \quad a = g$$

t (s)	h (m)	t (s)	h (m)
0	10	0	10
0.49	9	0.47	9
0.63	8	0.67	8
0.83	7	0.79	7
0.88	6	0.94	6
0.95	5	1.0	5

$$\underbrace{\begin{bmatrix} t_1^2/2 \\ t_2^2/2 \\ \vdots \\ t_m^2/2 \end{bmatrix}}_{A} \underbrace{\begin{pmatrix} g \end{pmatrix}}_{\phi} = \underbrace{\begin{pmatrix} h_1 - h_o \\ h_2 - h_o \\ \vdots \\ h_m - h_o \end{pmatrix}}_{b}$$

$$A^\mathsf{T} A \phi = A^\mathsf{T} b$$




```
t = [ 0 0.49 0.63 0.83 0.88 0.95 ...
      0 0.47 0.67 0.79 0.94 1.0];
h = [ 10 9 8 7 6 5 ...
      10 9 8 7 6 5 ];
h0 = 10;
n = length(t);

A = [(t'.^2)/2];
b = h'-h0;

AA = A'*A;
bb = A'*b;

phi=AA\bb;
```

THE UNIVERSITY OF UTAH

Example - Reaction Rate Constant

Pre-exponential factor

Activation energy (J/mol)

$$k = A \exp\left(\frac{-E_a}{RT}\right)$$

rate "constant"

Gas constant R=8.314 J/mol-K

Temperature (K)

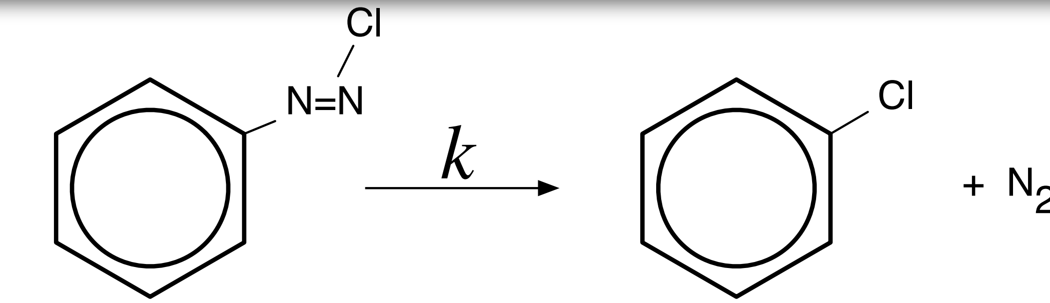
$$\ln(k) = \ln(A) - \frac{E_a}{RT}$$

$\boxed{\text{recall: } \ln(ab) = \ln(a) + \ln(b)}$

$$y = a_0 + a_1 x$$

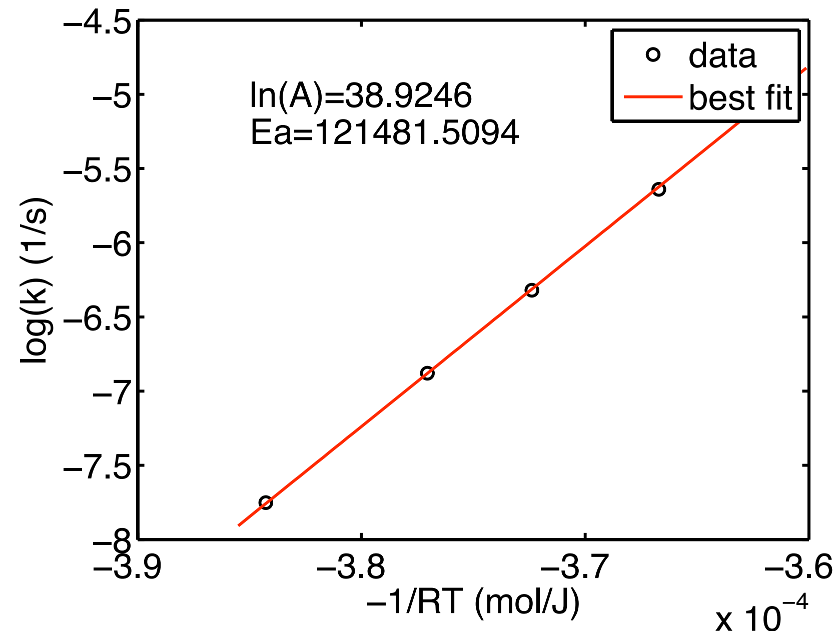
$$y = \ln(k), \qquad x = \frac{-1}{RT},$$

$$a_0 = \ln(A), \qquad a_1 = E_a$$



Benzenediazonium chloride

Chlorobenzene



T (K)	k ($1/s$)
313	0.00043
319	0.00103
323	0.00180
328	0.00355
333	0.00717

$$\underbrace{\begin{bmatrix} 1 & \frac{-1}{RT_1} \\ 1 & \frac{-1}{RT_2} \\ \vdots & \vdots \\ 1 & \frac{-1}{RT_m} \end{bmatrix}}_{Q} \underbrace{\begin{pmatrix} a_0 \\ a_1 \end{pmatrix}}_{\phi} = \underbrace{\begin{pmatrix} \ln(k_1) \\ \ln(k_2) \\ \vdots \\ \ln(k_m) \end{pmatrix}}_{b} \quad \Rightarrow \quad Q^T Q \phi = Q^T b$$

solve for $\phi = \begin{pmatrix} a_0 \\ a_1 \end{pmatrix}$

Note: need to calculate A from a_0.

General Polynomial Regression

$$p = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots + a_n x^n$$

Given m ($m>n$) observations (x_i,p_i), find a_j.

One equation for each observation (m equations)

$$\underbrace{\begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^n \\ 1 & x_2 & x_2^2 & \cdots & x_2^n \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_m & x_m^2 & \cdots & x_m^n \end{bmatrix}}_{A} \underbrace{\begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}}_{\phi} = \underbrace{\begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ \vdots \\ p_m \end{pmatrix}}_{b}$$

$$A^{\mathsf{T}} A \phi = A^{\mathsf{T}} b$$

NOTE: this is a *linear* problem for the coefficients, a_i.

THE UNIVERSITY OF UTAH

Recap - Regression

Given a set of observations, and a function
that you wish to determine parameters for:

- Write the function in a form where the parameters enter *linearly* as polynomial coefficients.
  - May need to rearrange function a bit.
  - Sometimes logarithm functions can help accomplish this.
- Once you have the function in polynomial form, solve the Normal Equations:

$$p(x) = \sum_{i=0}^{n} a_i x^i \rightarrow \underbrace{\begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^n \\ 1 & x_2 & x_2^2 & \cdots & x_2^n \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_m & x_m^2 & \cdots & x_m^n \end{bmatrix}}_{A} \underbrace{\begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}}_{\phi} = \underbrace{\begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ \vdots \\ p_m \end{pmatrix}}_{b} \rightarrow \boxed{A^\mathsf{T} A \phi = A^\mathsf{T} b}$$

- Finally, from ϕ, calculate the parameters in the original equation.

THE UNIVERSITY OF UTAH

Regression - MATLAB

- Do it "manually" - the way that we just showed.
- Polynomial regression: `p=`**`polyfit`**`(x,y,n)`
  - gives the "best fit" for a polynomial of order n through the data.
  - if `n==(length(x)-1)` then you get an *interpolant*.
  - if `n<(length(x)-1)` then you get a *least-squares* fit.
  - You still must get the problem into a polynomial form.

Linear Least Squares Regression Using Excel

- Convert data to polynomial form
- Plot converted data
- Right-click line & choose "Add Trendline"
  - Choose the appropriate trendline type
  - Under "options" choose "Display Equation on Chart"
    - Also set y-intercept value if known - otherwise it will be calculated as a parameter.
- Convert parameters back to obtain them for the original equation.

THE UNIVERSITY OF UTAH

The "R²" Value

How well does the regressed line fit the data?

$\hat{\phi}_i$ Value predicted by the function.

ϕ_i Observed value (data).

$$\bar{\phi} = \frac{1}{n}\sum_{i=1}^{n}\phi_i$$ Average value of φ

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(\phi_i - \hat{\phi}_i\right)^2}{\sum_{i=1}^{n}\left(\phi_i - \bar{\phi}\right)^2}$$

Measure of how well the line fits the data.

$R^2=1 \Rightarrow$ Perfect fit

THE UNIVERSITY OF UTAH