Regression
(multiple Regression to come later)

Dan Frey
Assistant Professor of Mechanical Engineering and Engineering Systems

Concept Question

In hospital (A), 45 babies are born each day (on average) and in the smaller hospital (B) about 15 babies are born each day (on average).

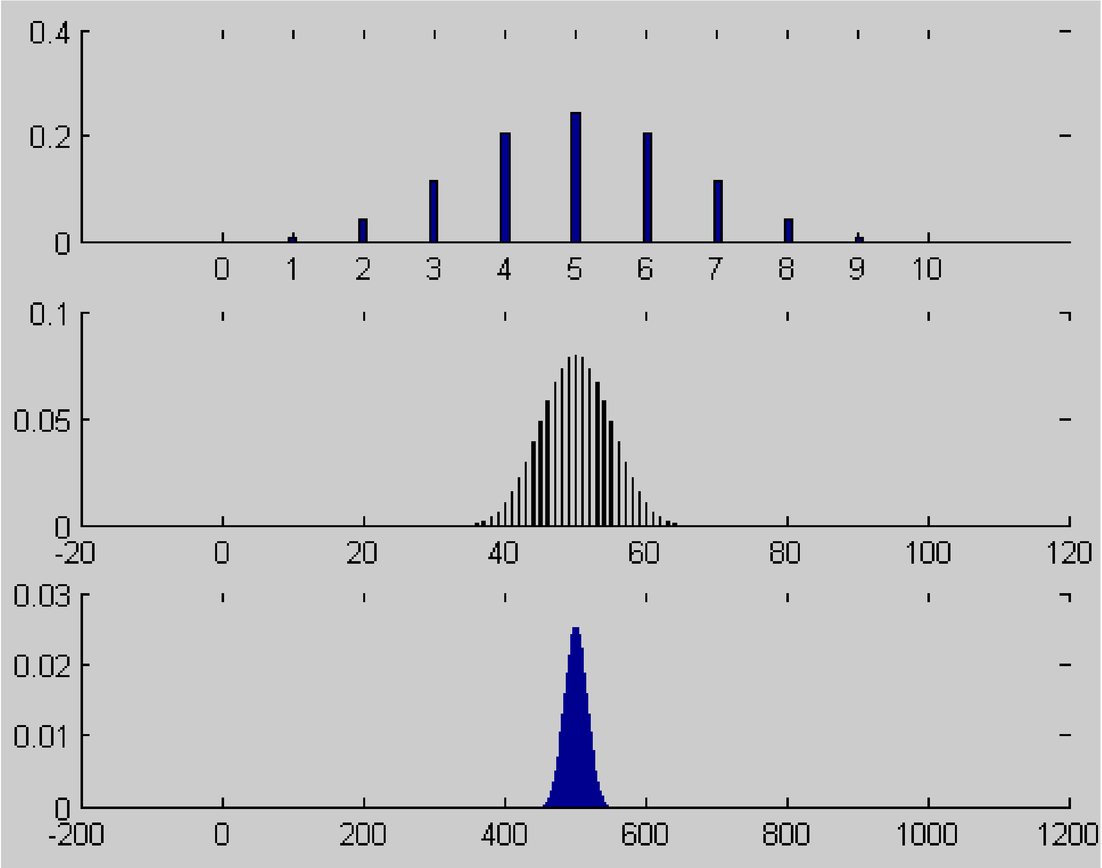
Let's model births as a Bernoulli process and both hospitals have p=0.5 (baby boys and girls equally probable).

For a period of a year, each hospital recorded the days on which more than 60% of the babies were boys.

1) Hospital A probably recorded more days with >60% boys
2) Hospital B probably recorded more days with >60% boys
3) Hospital A and B are probably about the same

The Binomial Distribution

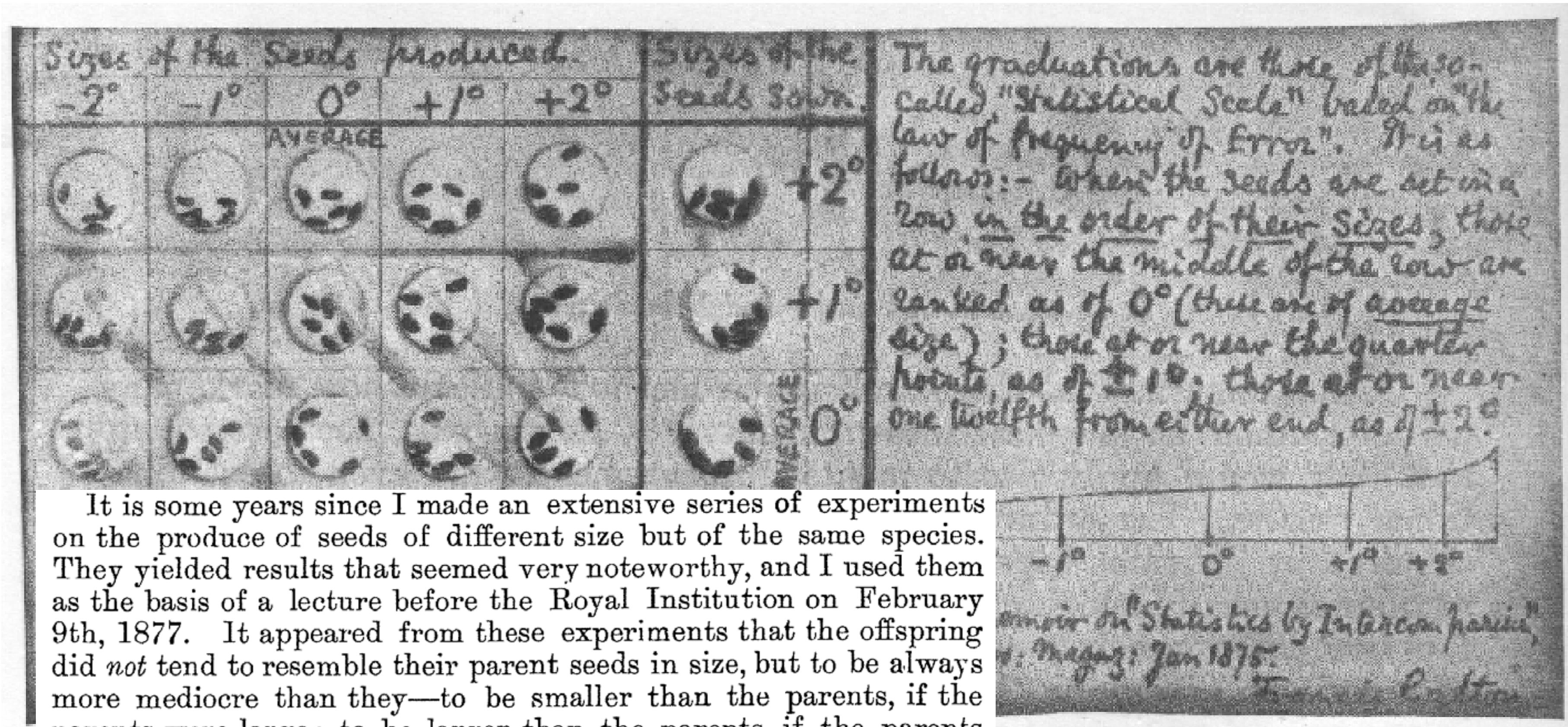
```
n=10;
x = 0:n;
y = binopdf(x,n,0.5);
subplot(3,1,1); bar(x,y,0.1)

n=100;
x = 0:n;
y = binopdf(x,n,0.5);
subplot(3,1,2); bar(x,y,0.1)

n=1000;
x = 0:n;
y = binopdf(x,n,0.5);
subplot(3,1,3); bar(x,y,0.1)
```



Plan for Today

- Regression
  - History / Motivation
  - The method of least squares
  - Inferences based on the least squares estimators
  - Checking the adequacy of the model
  - The Bootstrap
  - Non-linear regression

Regression Toward the Mean

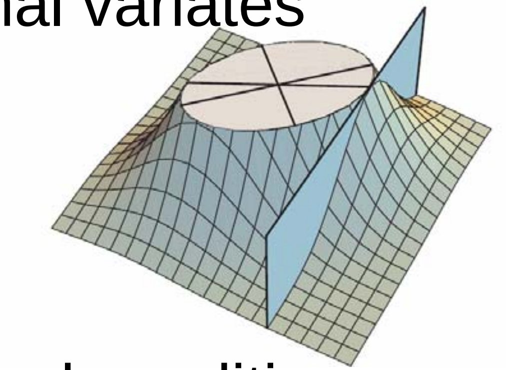


It is some years since I made an extensive series of experiments on the produce of seeds of different size but of the same species. They yielded results that seemed very noteworthy, and I used them as the basis of a lecture before the Royal Institution on February 9th, 1877. It appeared from these experiments that the offspring did *not* tend to resemble their parent seeds in size, but to be always more mediocre than they—to be smaller than the parents, if the parents were large; to be larger than the parents, if the parents were very small. The point of convergence was considerably below the average size of the seeds contained in the large bagful I bought at a nursery garden, out of which I selected those that were sown, and I had some reason to believe that the size of the seed towards which the produce converged was similar to that of an average seed taken out of beds of self-planted specimens.

Galton, Francis, 1886, "'Regression towards mediocrity in hereditary stature," J*ournal of the Anthropological Institute* **15**:246-63.

Regression Toward the Mean

Consider the joint pdf of two standard normal variates



$$f(x,y)=\frac{1}{2\pi\sqrt{1-\rho^2}}e^{-\frac{x^2-2\rho xy-y^2}{2(1-\rho^2)}}$$

Now, let's say you make an observation x and condition your marginal distribution of y

$$f(y|x)=\frac{f(x,y)}{f_x(x)}=\frac{1}{2\pi\sqrt{1-\rho^2}}e^{-\frac{x^2-2\rho xy-y^2}{2(1-\rho^2)}}\bigg/\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$$

$$y\sim N(\rho x,1-\rho^2)$$

the mean of y is less far from the mean than the observed value X

Regression Toward the Mean

“... while attempting to teach flight instructors that praise is more effective than punishment for promoting skill-learning...one of the most seasoned instructors in the audience raised his hand and made his own short speech..."On many occasions I have praised flight cadets for clean execution of some aerobatic maneuver, and in general when they try it again, they do worse. On the other hand, I have often screamed at cadets for bad execution, and in general they do better the next time. So please don't tell us that reinforcement works and punishment does not, because the opposite is the case." ...because we tend to reward others when they do well and punish them when they do badly, and because there is regression to the mean, it is part of the human condition that we are statistically punished for rewarding others and rewarded for punishing them.”

Kahneman, D., 2002, Bank of Sweden "Nobel" Prize Lecture

What is Linear Regression?

1. Form a probabilistic model

independent variable

theoretical parameters

random

random

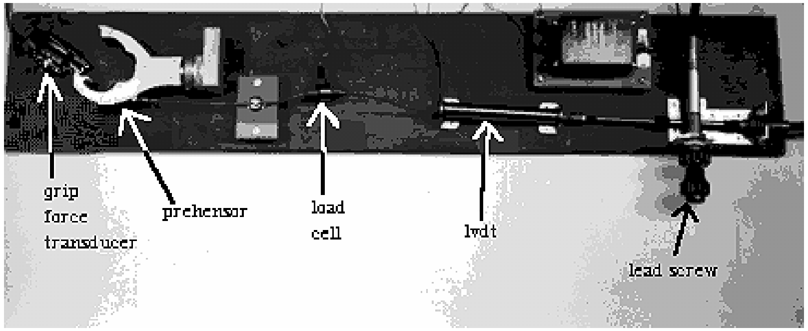
$E(\varepsilon)=0$

$$Y = \alpha + \beta x + \varepsilon$$

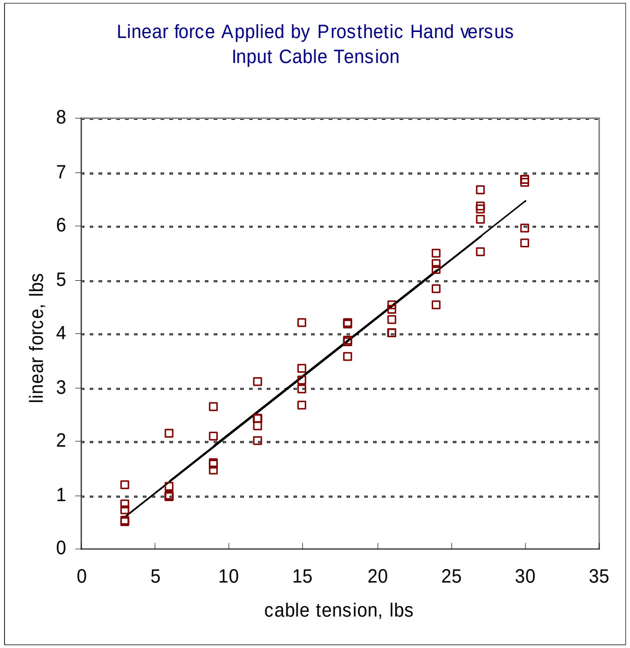
2. Get a sample of data in pairs (X_i, Y_i), $i=1...n$
3. Estimate the parameters of the model from the data

What can we do with Regression?

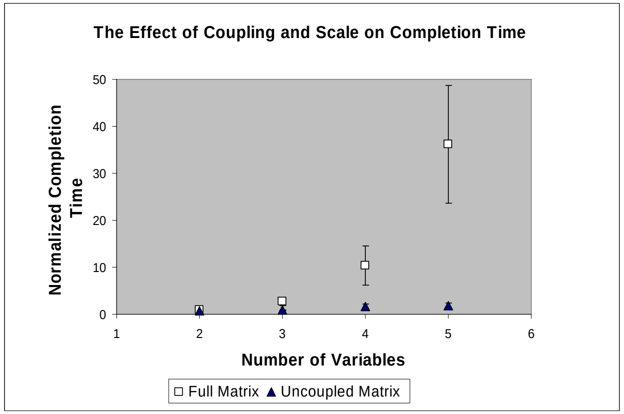
Calibrate a measuring device



Characterize a Product



Evaluate a Conjecture

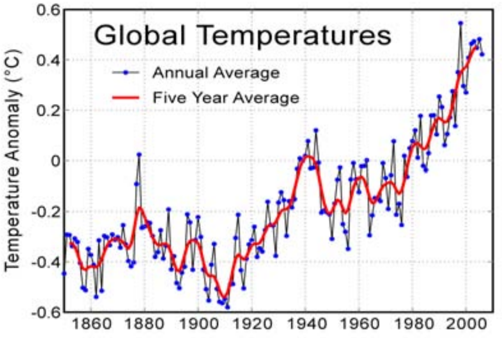


What can we do with Regression?

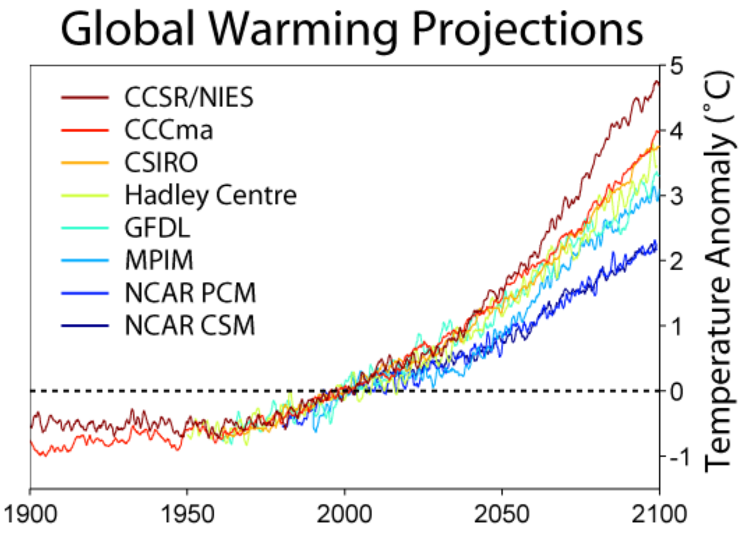
Diagnose a Problem

Photo and screen shot removed due to copyright restrictions.
Calibration apparatus by Renishaw plc.

Suggest a Trend



Extrapolate

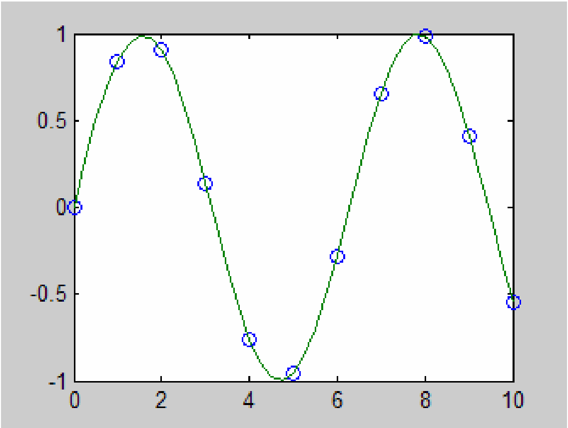


Source: Wikipedia. Courtesy of globalwarmingart.com.

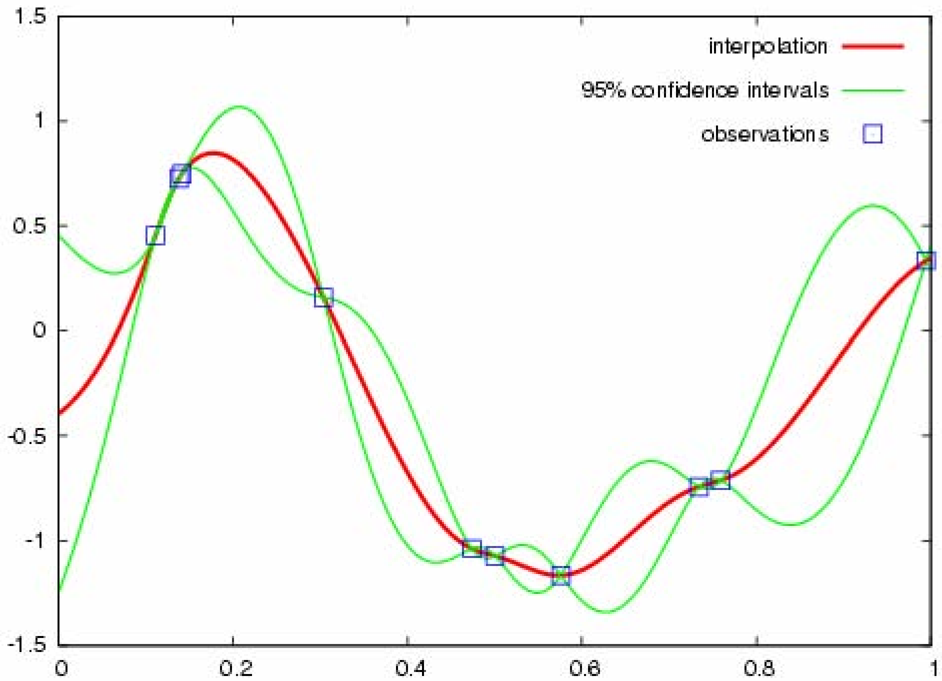
Interpolation is Different

- Fits a curve to a set of points
- But assume <u>no error</u> in the points themselves
- Enable estimates at values other than the observed ones

```
x = 0:10;
y = sin(x);
xx = 0:.25:10;
yy = spline(x,y,xx);
plot(x,y,'o',xx,yy)
```



spline interpolation



Kriging

Plan for Today

- Regression
  - History / Motivation
  - The method of least squares
  - Inferences based on the least squares estimators
  - Checking the adequacy of the model
  - The Bootstrap
  - Non-linear regression

Regression Curve of Y on x

theoretical
parameters

random

random
$E(\varepsilon)=0$

$$Y = \alpha + \beta x + \varepsilon$$

independent
variable

Regression Curve vs Prediction Equation

random

theoretical parameters

Random $E(\varepsilon)=0$

Regression Curve $Y = \alpha + \beta x + \varepsilon$

independent variable

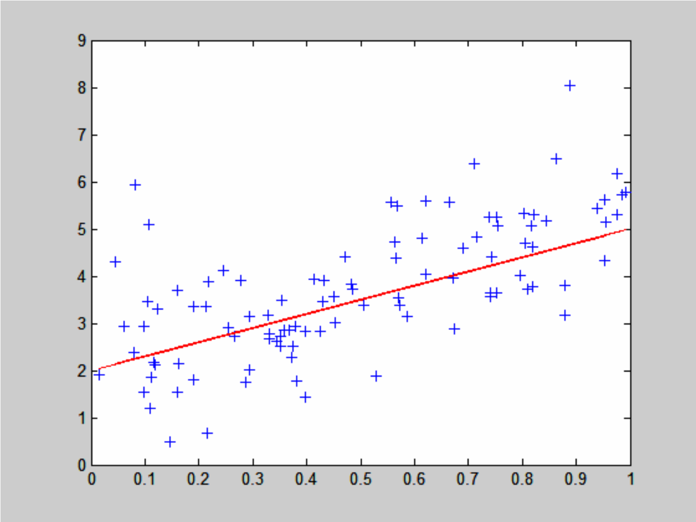
Prediction Equation $\hat{y} = a + bx$

estimated $E(Y|X)$

computed estimates of theoretical parameters α and β

Matlab Code Simulating the Probability model

```
hold on
alpha=2;
beta=3;
eps_std=1;
for trial=1:100
x(trial) = random('Uniform',0,1,1,1);
eps= random('Normal',0, eps_std,1,1);
Y(trial)=alpha+beta*x(trial)+eps;
end
plot(x,Y,'+')
hold on
plot(x,alpha+beta*x,'-','Color','r')
```



The Method of Least Squares

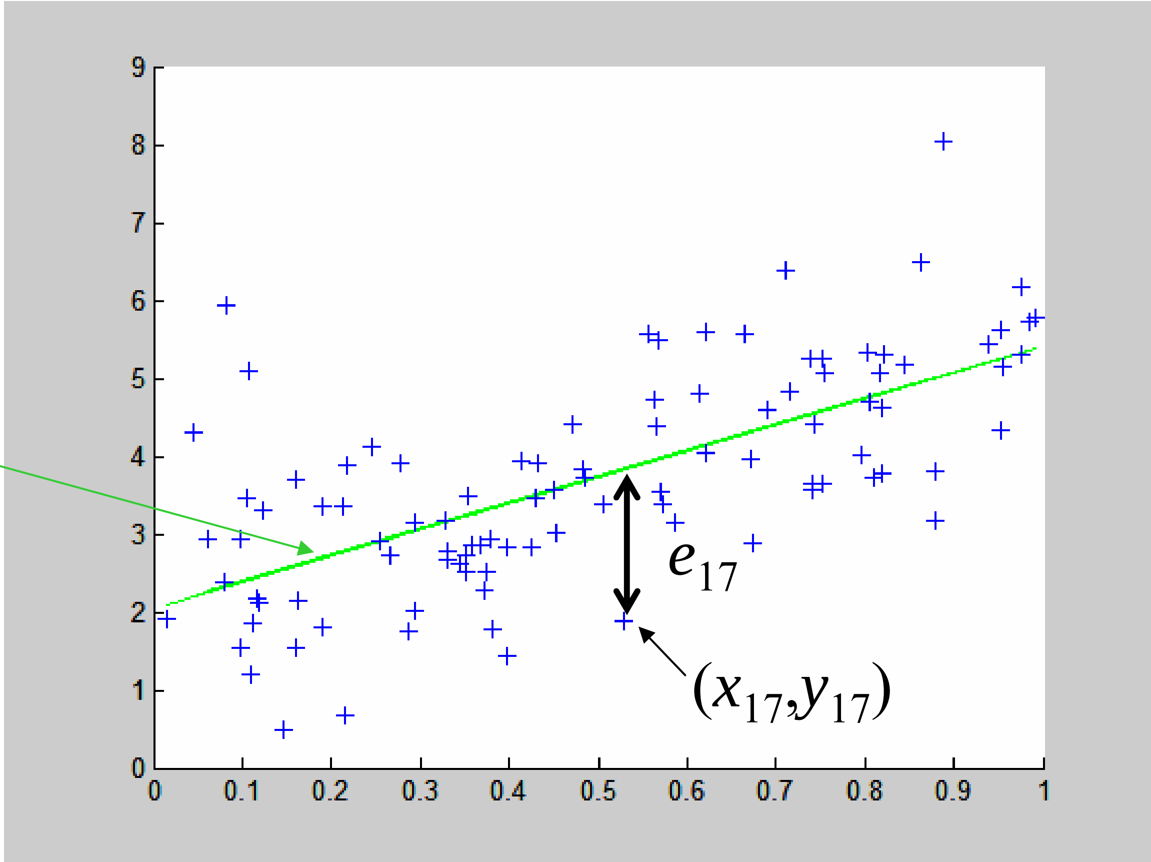
Given a set of *n* data points (*x*,*y*) pairs

There exists a <u>unique</u> line $\hat{y} = a + bx$

that minimizes the *residual sum of squares*

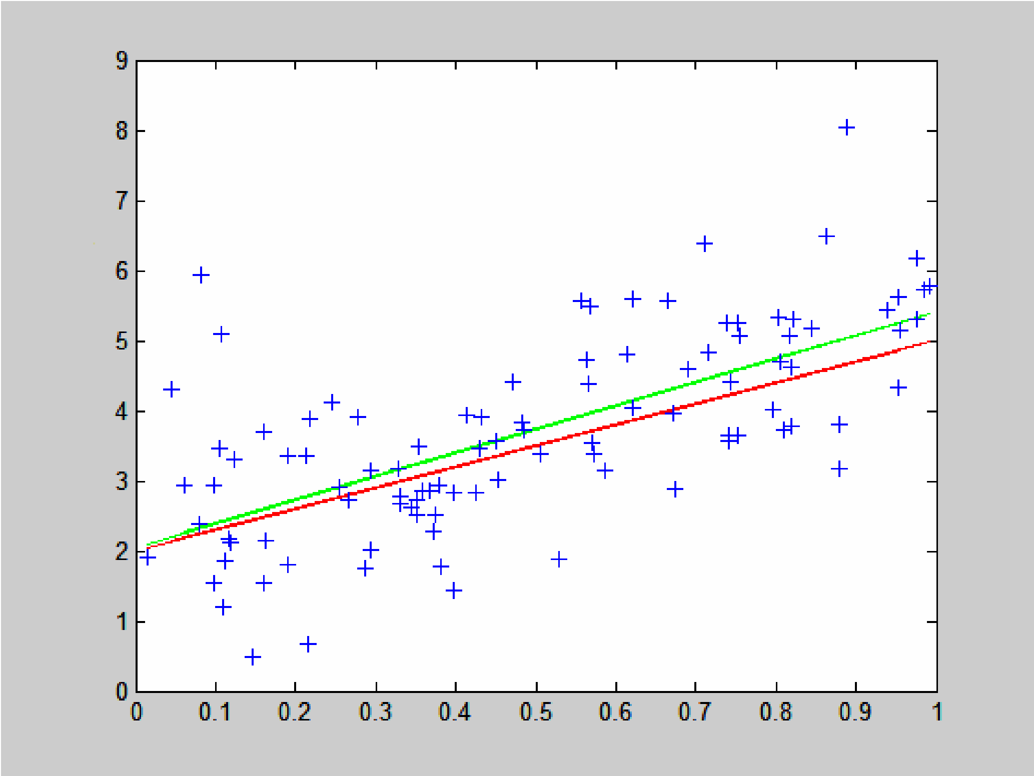
$$\sum_{i=1}^{n} e_i^{\,2} \qquad e_i = y_i - \hat{y}_i$$

$$s_e^{\,2} = \frac{1}{n-2}\sum_{i=1}^{n} e_i^{\,2}$$



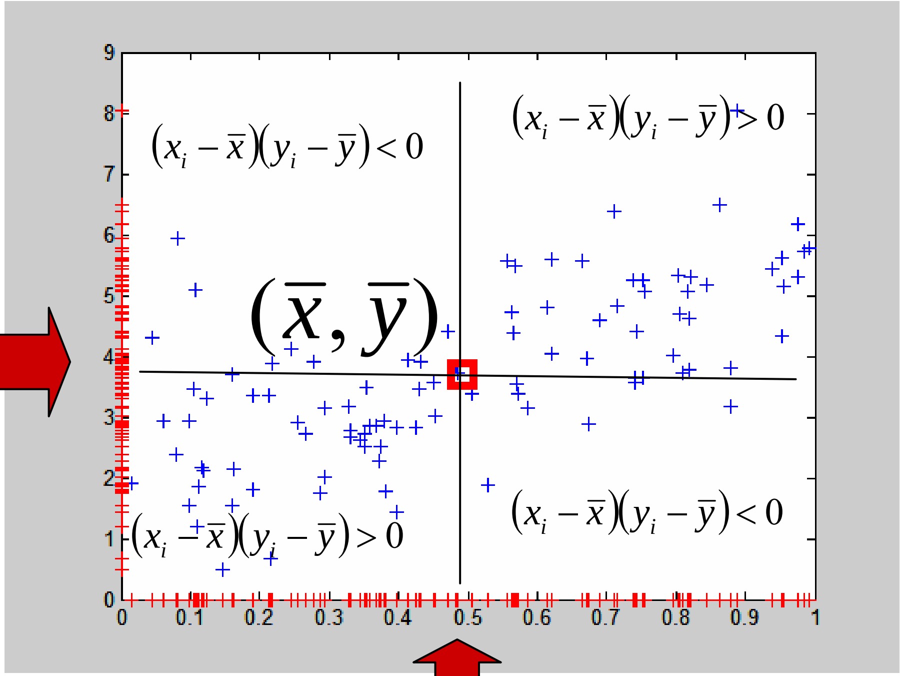
Matlab Code for Regression

```
p = polyfit(x,Y,1)
y_hat=polyval(p,x);
plot(x,y_hat,'-','Color', 'g')
```



Computing Least Squares Estimators

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$



What are the values of *a* and *b* for the regression line?

$$b = \frac{S_{xy}}{S_{xx}}$$

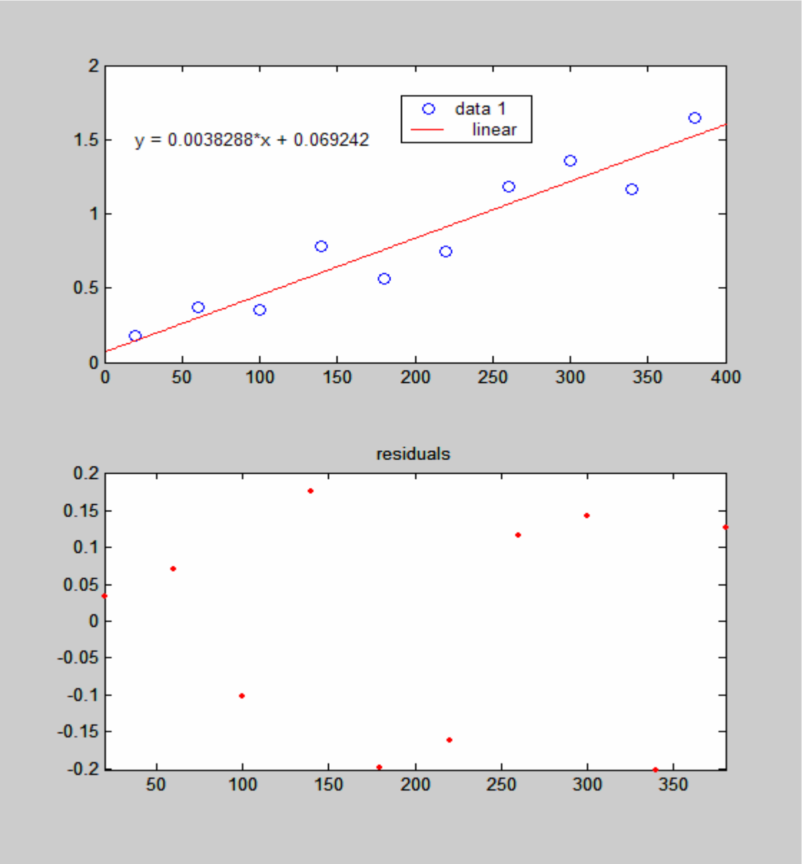
$$a = \bar{y} - b\bar{x}$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Example – Evaporation vs Air Velocity

Air vel (cm/sec)	Evap coeff. (mm^2/sec)
20	0.18
60	0.37
100	0.35
140	0.78
180	0.56
220	0.75
260	1.18
300	1.36
340	1.17
380	1.65



Concept Question

You are seeking to calibrate a load cell. You wish to determine the regression line relating voltage (in Volts) to force (in Newtons). What are the units of a, b, S_{xx}, and S_{xy} respectively?

1) N, N, N, and N
2) V, V, V^2, and V^2
3) V, V/N, N^2, and VN
4) V/N, N, VN, and V^2
5) None of the variables have units

Regression Curve vs Prediction Equation

theoretical parameters

Random $E(\varepsilon)=0$

random

Regression Curve $Y = \alpha + \beta x + \varepsilon$

independent variable

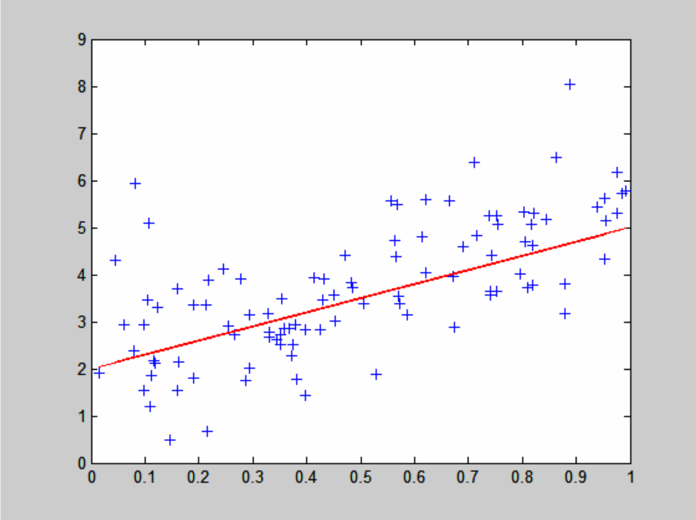
Prediction Equation $\hat{y} = a + bx$

estimated $E(Y|x)$

computed estimates of theoretical parameters α and β

Matlab Code Simulating the Probability model

```
hold on
alpha=2;
beta=3;
eps_std=1;
for trial=1:100
x(trial) = random('Uniform',0,1,1,1);
eps= random('Normal',0, eps_std,1,1);
Y(trial)=alpha+beta*x(trial)+eps;
end
plot(x,Y,'+')
hold on
plot(x,alpha+beta*x,'-','Color','r')
```



Why is the Least Squares Approach Important?

There are other criteria that also provide reasonable fits to data (e.g. minimize the max error)

BUT, if the data arise from the model below, then least squares method provides an unbiased, minimum variance estimate of α and β

random

theoretical parameters

independent variable

$$Y = \alpha + \beta x + \varepsilon$$

random

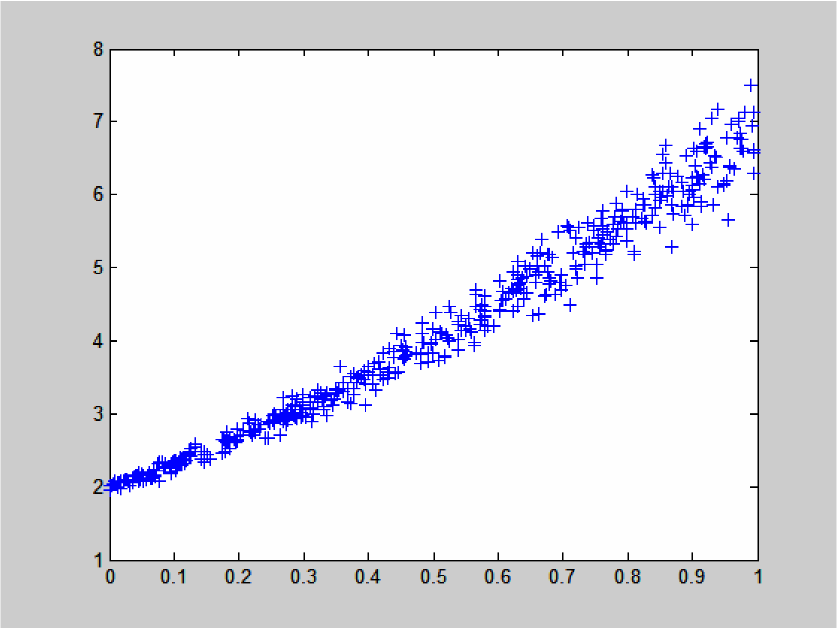
Plan for Today

- Regression
  - History / Motivation
  - The method of least squares
  - Inferences based on the least squares estimators
  - Checking the adequacy of the model
  - The Bootstrap
  - Non-linear regression

Assumptions Required for Inferences to be Discussed

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

- The Y_i are
  - independent
  - normally distributed
  - with means $\alpha + \beta X_i$
  - and common variance (homoscedastic)



Inferences Based on the Least Squares Estimators

$$t = \frac{(b-\beta)}{s_e}\sqrt{S_{xx}}$$

is a random variable having the t distribution with n-2 degrees of freedom

Evaporation vs Air Velocity Hypothesis Tests

Air vel (cm/sec)	Evap coeff. (mm^2/sec)
20	0.18
60	0.37
100	0.35
140	0.78
180	0.56
220	0.75
260	1.18
300	1.36
340	1.17
380	1.65

Air vel (cm	Evap coeff. (mm2/sec)
20	0.18
60	0.37
100	0.35
140	0.78
180	0.56
220	0.75
260	1.18
300	1.36
340	1.17

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.934165
R Square	0.872665
Adjusted R Square	0.854474
Standard Error	0.159551
Observations	9

ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regression	1	1.221227	1.221227	47.97306	0.000226
Residual	7	0.178196	0.025457		
Total	8	1.399422			

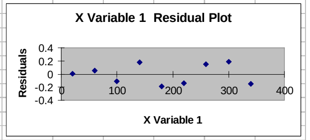
	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	0.102444	0.106865	0.958637	0.369673	-0.15025	0.355139	-0.15025	0.355139
X Variable 1	0.003567	0.000515	6.926259	0.000226	0.002349	0.004784	0.002349	0.004784

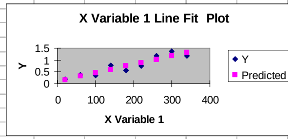
RESIDUAL OUTPUT

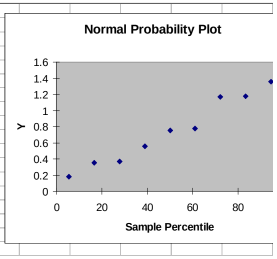
Observation	*Predicted Y*	*Residuals*	*dard Residuals*
1	0.173778	0.006222	0.041691
2	0.316444	0.053556	0.35884
3	0.459111	-0.10911	-0.73108
4	0.601778	0.178222	1.194149
5	0.744444	-0.18444	-1.23584
6	0.887111	-0.13711	-0.91869
7	1.029778	0.150222	1.006539
8	1.172444	0.187556	1.256685
9	1.315111	-0.14511	-0.97229

PROBABILITY OUTPUT

Percentile	*Y*
5.555556	0.18
16.66667	0.35
27.77778	0.37
38.88889	0.56
50	0.75
61.11111	0.78
72.22222	1.17
83.33333	1.18
94.44444	1.36



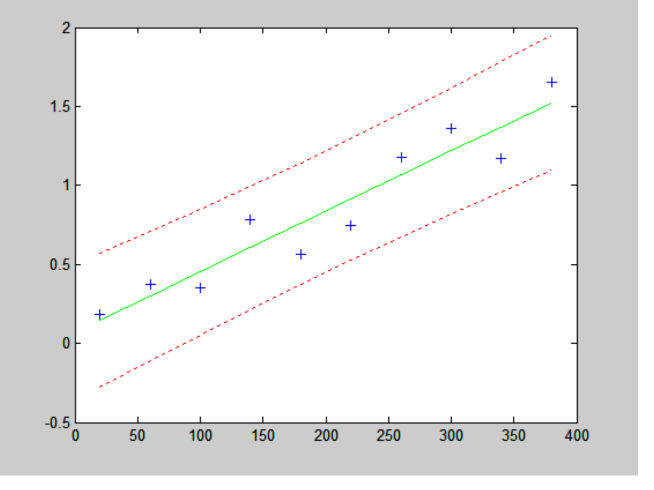




Evaporation vs Air Velocity Confidence Intervals for Prediction

Air vel (cm/sec)	Evap coeff. (mm^2/sec)
20	0.18
60	0.37
100	0.35
140	0.78
180	0.56
220	0.75
260	1.18
300	1.36
340	1.17
380	1.65

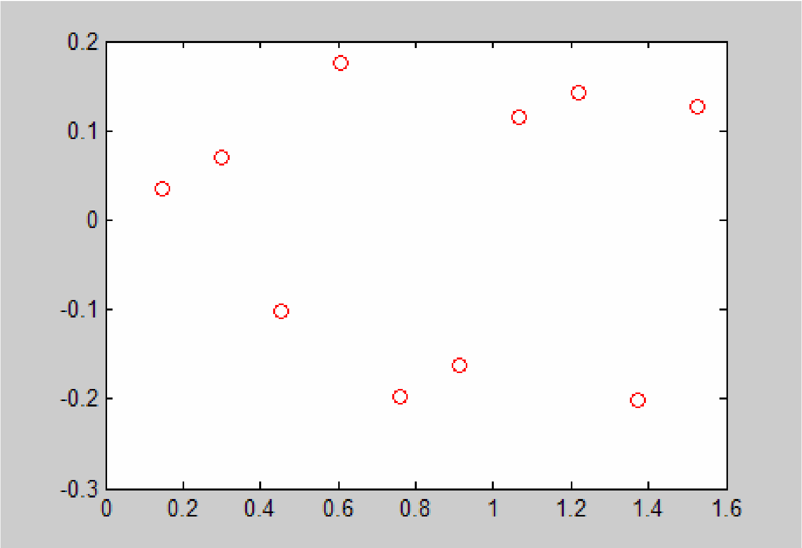
```
[p,S] = polyfit(x,y,1);
alpha=0.05;
[y_hat,del]=polyconf(p,x,S,alpha);
plot(x,y,'+',x,y_hat,'g')
hold on
plot(x,y_hat+del,'r:')
plot(x,y_hat-del,'r:')
```



Checking the Assumptions

- Plot the residuals
  - Check for patterns
  - Check for uniform variance

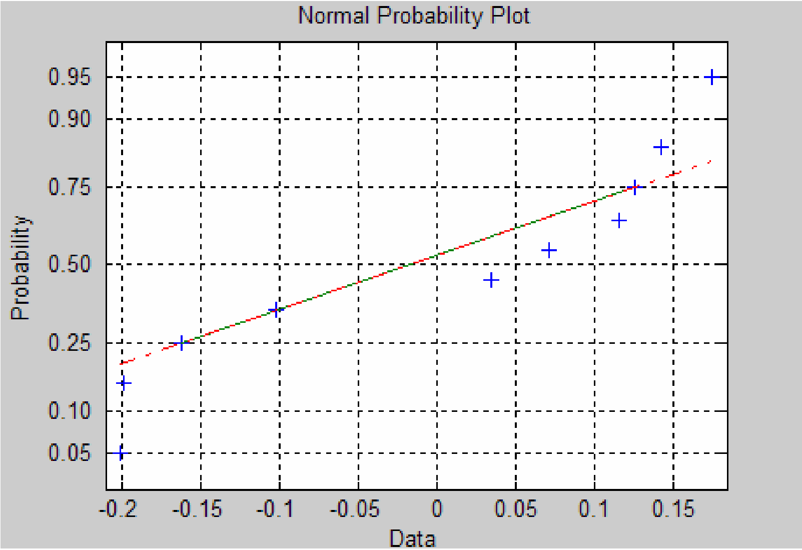
```
hold off
e=y-y_hat;
plot(y_hat, e, 'or')
```



Checking the Assumptions

- Normal scores-plot of the residuals
  - Check for linearity
  - If there are outliers
    - check sensitivity of results
    - try to identify “special causes”

```
hold off
normplot(e)
```



Plan for Today

- Regression
  - History / Motivation
  - The method of least squares
  - Inferences based on the least squares estimators
  - Checking the adequacy of the model
  - The Bootstrap
  - Non-linear regression

The Bootstrap

- A random smaple of size n is observed from a completely unspecified probability distribution F

$$X_i = x_i, \quad X_i \sim_{\text{ind}} F$$

- Given a random variable $R(\mathbf{X},F)$, estimate the sampling distribution on R on the basis of he observed data $\mathbf{x}$

Efron, B., 1979, "Bootstrap Methods: Another Look at the Jackknife," *Annals of Statistics* 7:1-26.

The Bootstrap

- Construct a sample probability distribution ,
  putting mass 1/n at each point $x_1, x_2, \ldots, x_n$
- With $\hat{F}$ fixed, draw a random sample $\mathbf{X}^*$ of size n from $\hat{F}$
- Approximate the sampling distribution as the bootstrap distribution

$$R^* = R(\mathbf{X}^*, \hat{F})$$

- "...shown to work satistfactorily on a variety of estimation problems."

Efron, B., 1979, "Bootstrap Methods: Another Look at the Jackknife," *Annals of Statistics* 7:1-26.

The Bootstrap

- In the Acknowledgements:

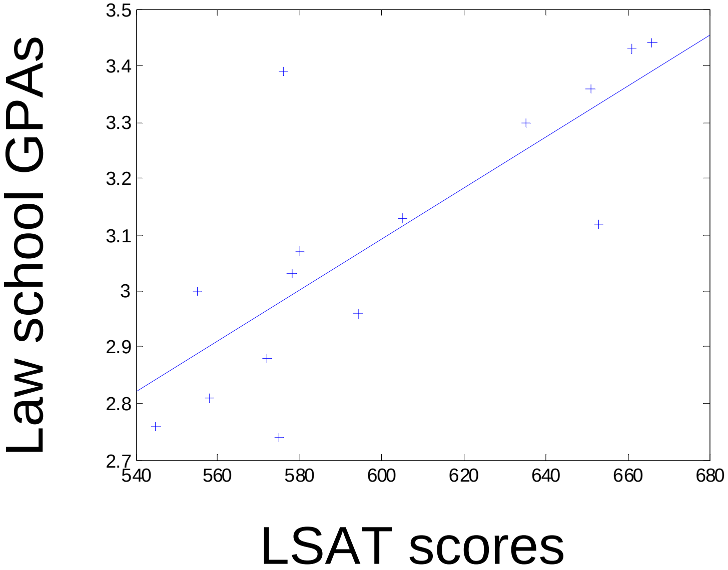
  ...I also wish to thank the many freinds who suggested names more colorful than *Bootstrap*, including *Swiss Army Knife*, *Meat Axe*, *Swan-Dive*, *Jack-Rabbit*, and my personal favorite, the *Shotgun*, which, to paraphrase Tukey, "can blow the head off any problem if the statistician can stand the resulting mess."

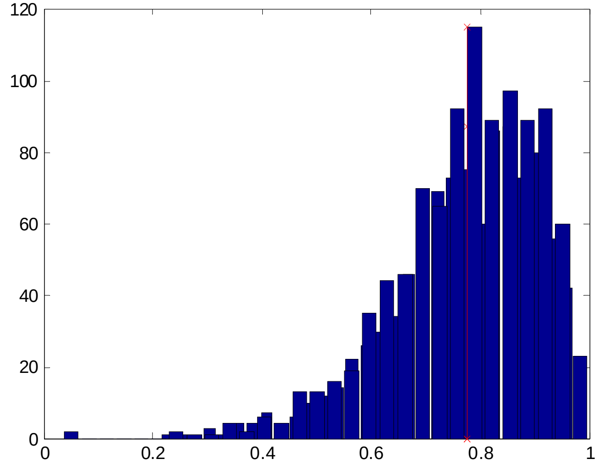
Efron, B., 1979, "Bootstrap Methods: Another Look at the Jackknife," *Annals of Statistics* 7:1-26.

The Bootstrap

- Sampling with replacement

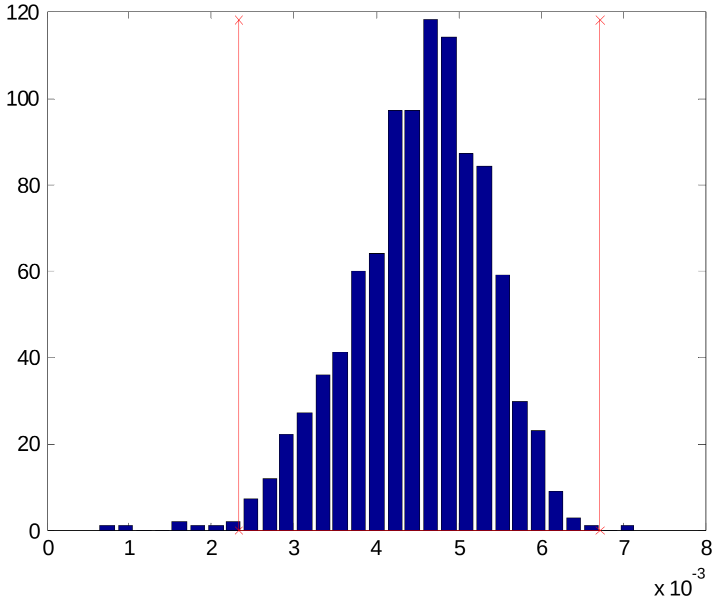
```
load lawdata
figure(1); plot(lsat,gpa,'+')
lsline
rhohat = corr(lsat,gpa);
rhos1000 =
bootstrp(1000,'corr',lsat,gpa);
[n,xout]=hist(rhos1000,30);
figure(2); bar(xout,n); hold on;
plot([rhohat rhohat],[0 max(n)],'-rx');
```

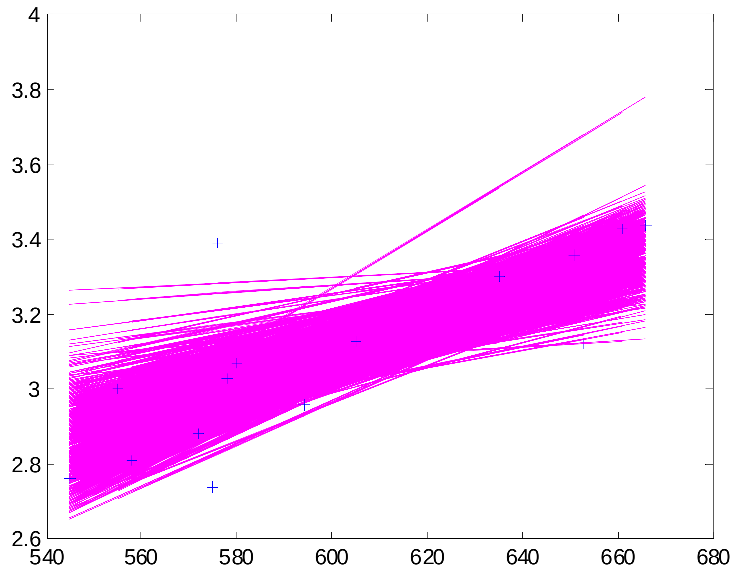




The Bootstrap for Regression

```
load lawdata; [n m]=size(lsat);
[b bint]=regress(gpa, [ones(n,1) lsat]);
for t=1:1000
  samp_repl=floor(n*rand(size(lsat)))+1;
  x=[ones(n,1) lsat(samp_repl)];
y=gpa(samp_repl);
  b_boot = regress(y,x);
  int(t)=b_boot(1); slope(t)=b_boot(2);
end
[bin_n,xout]=hist(slope,30);
figure(3); bar(xout,bin_n); hold on;
plot([bint(2,1) bint(2,1) bint(2,2)
bint(2,2)],[max(bin_n) 0 0 max(bin_n)],'-rx');
figure(4); hold on;
for t=1:1000;
  plot(lsat,slope(t)*lsat+int(t),'m');
end
plot(lsat,gpa,'+');
```





Plan for Today

- Regression
  - History / Motivation
  - The method of least squares
  - Inferences based on the least squares estimators
  - Checking the adequacy of the model
  - The Bootstrap
  - Non-linear regression

Polynomial Regression

Linear regression curve

$$Y = \alpha + \beta x + \varepsilon$$

Polynomial regression curve

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_p x^p + \varepsilon$$

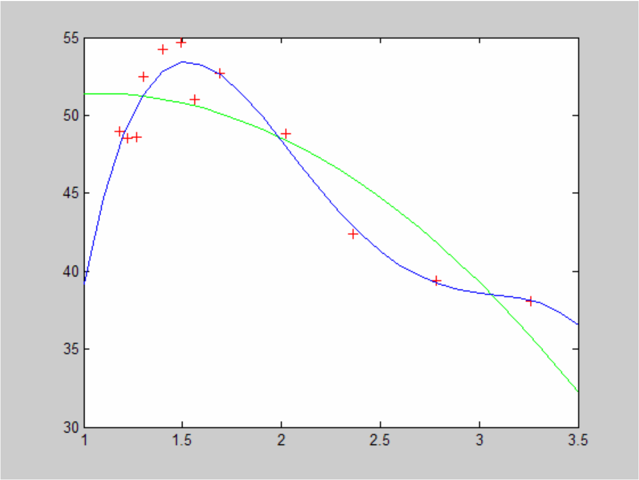
Often used for locally approximating "well behaved" functions because of Taylor's series approximation

$$f(x) \approx f(x_0) + (x - x_0)\left.\frac{\mathrm{d}f}{\mathrm{d}x}\right|_{x=x_0} + (x - x_0)^2 \left.\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}\right|_{x=x_0} + \text{h.o.t}$$

Beware of Over Fitting

Current (Amps)	Efficiency
1.18	49.0%
1.22	48.5%
1.27	48.6%
1.3	52.5%
1.4	54.2%
1.49	54.7%
1.56	51.0%
1.69	52.7%
2.02	48.8%
2.36	42.4%
2.78	39.4%
3.26	38.1%

```
p2= polyfit(I,e,2)
p4 = polyfit(I,e,4)
I2=1:0.1:3.5;
e_hat2=polyval(p2,I2);
e_hat4=polyval(p4,I2);
plot(I,e,'+r',I2,e_hat2,'-g', I2,e_hat4,'-b')
```

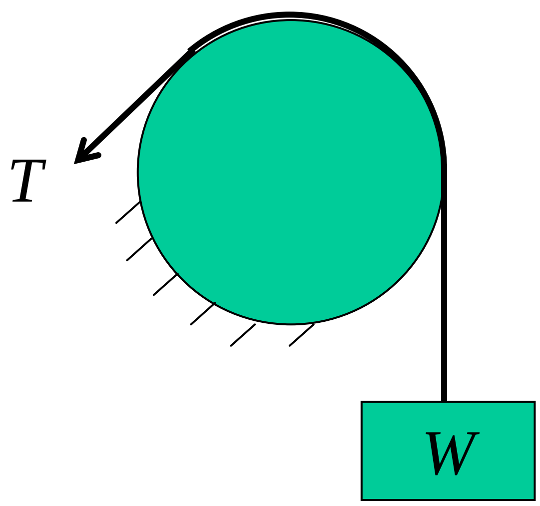


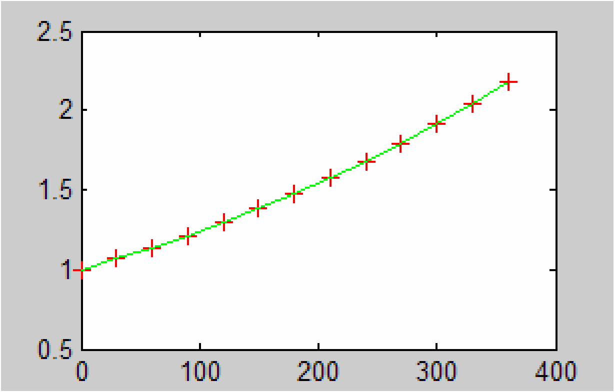
Exponential Curve Fitting

Theta	T/W
0	1
30	1.06708
60	1.13966
90	1.215042
120	1.296548
150	1.38352
180	1.436327
210	1.57536
240	1.701036
270	1.7938
300	1.914129
330	2.002529
360	2.179542

The Capstan Equation

$$T = We^{\mu\Theta}$$



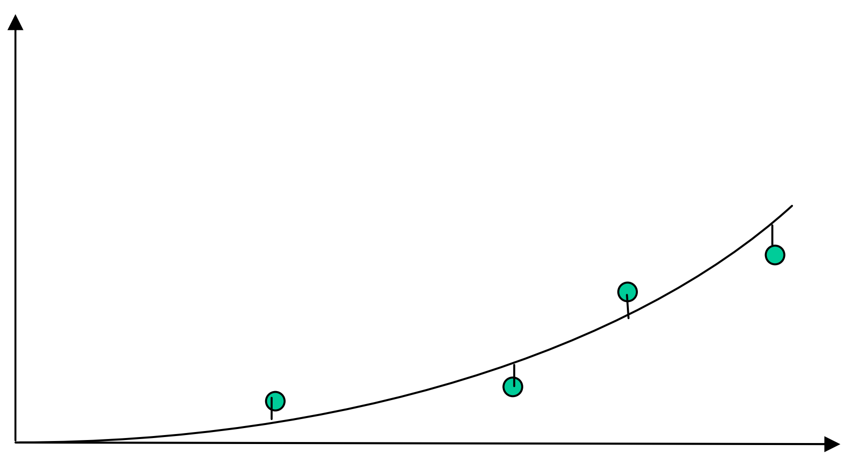



```
lTW= log(TW);
p = polyfit(theta,lTW,1);
lTW_hat=polyval(p,theta);
TW_hat=exp(lTW_hat);
plot(theta,TW,'+r',theta,TW_hat,'-g')
```

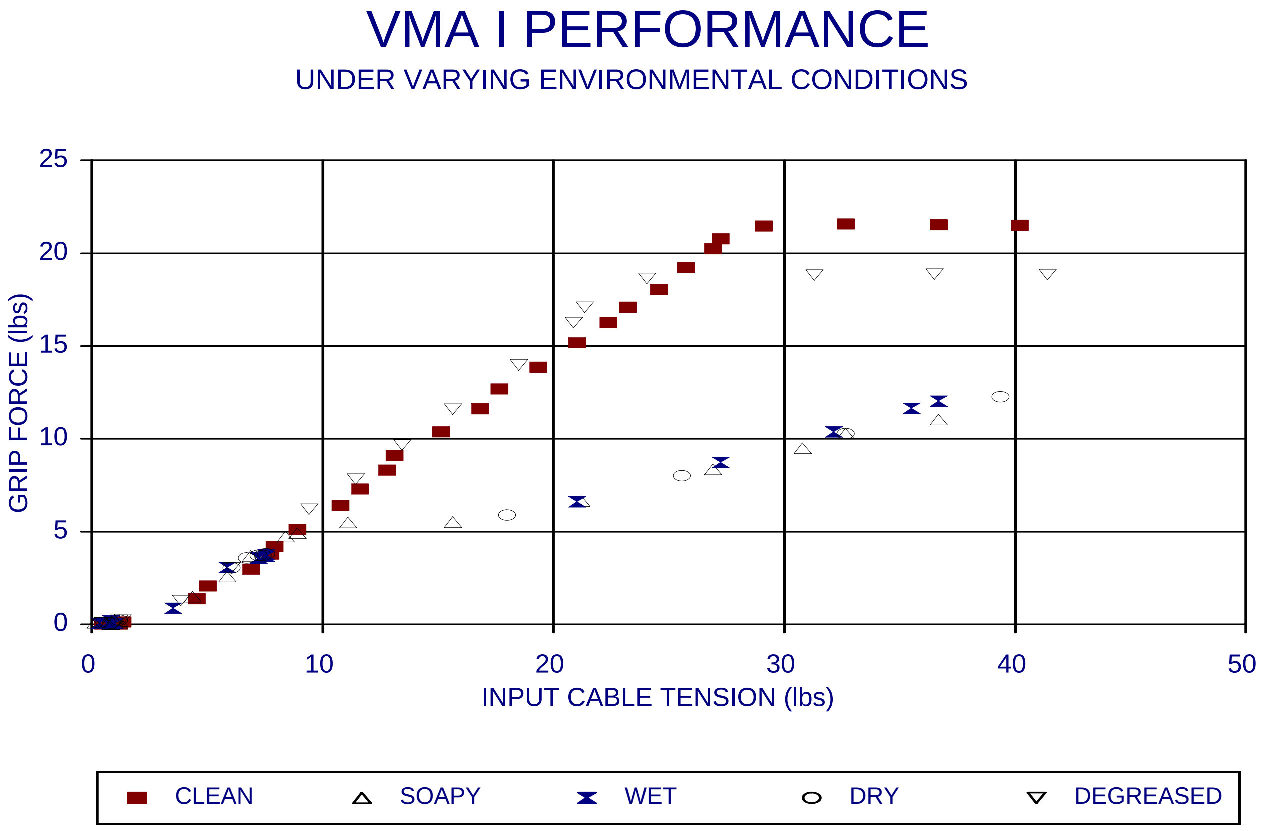
Concept Question

When you carry out an exponential regression by transforming the dependent variable, the resulting regression curve minimizes the sum squared error of the residuals as plotted here.

1) TRUE
2) FALSE



What about a system like this?



Next Steps

- Friday, 13 April
  - Session to support the term project
  - Be prepared to stand up and talk for 5 minutes about your ideas and your progress
- 16-17 April, No classes (Patriot's Day)
- Wednesday 18 April
  - Efron, "Bayesians, Frequentists, and Scientists"
  - Analysis of Variance
- Friday 20 April, recitation (by Frey)
- Monday 23 April, PS#6 due