



Predictive Statistics (Trending)
a Tutorial
CMG Brazil

Ray Wicks
561-236-5846
RayWicks@us.ibm.com
RayWicks@yahoo.com

IBM 2008

Trade Marks, Copyrights & Stuff

This presentation is copyright by Ray Wicks 2008.

Many terms are trademarks of different companies and are owned by them.

This session is sponsored by **IBM**

- On foils that appear in this presentation are not in the handout. This is to prevent you from looking ahead and spoiling my jokes and surprises.

IBM 2008

Abstract

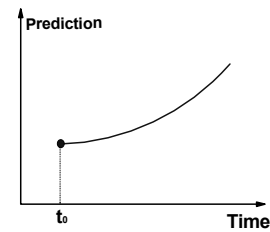
Predictive Statistics (Trending) – A Tutorial

This session reviews some of the trending techniques which can be useful in capacity planning. The introduction of the basic statistical concept of regression analysis will be examined. The simple linear regression analysis will be shown.

This session is sponsored by **IBM**

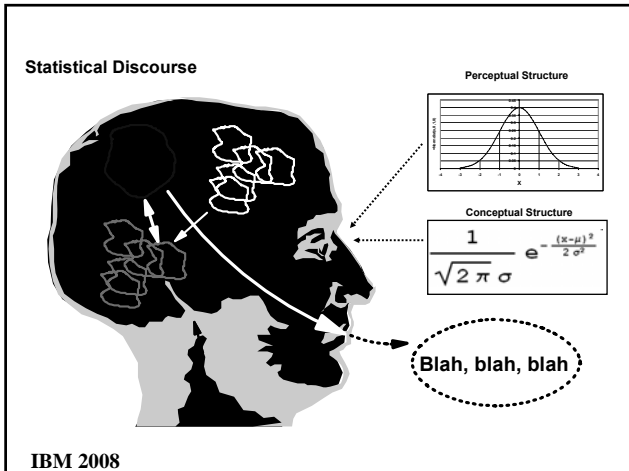
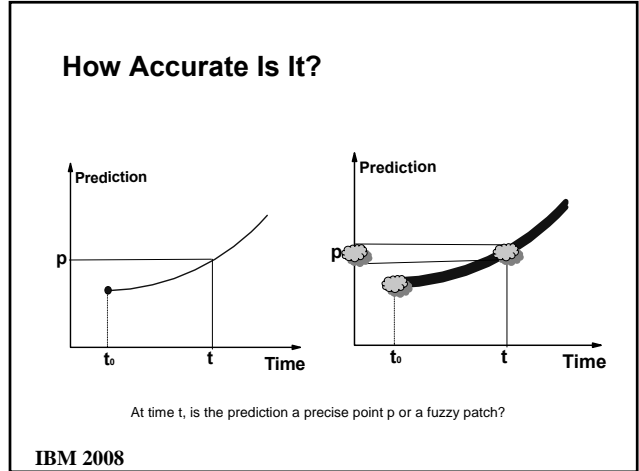
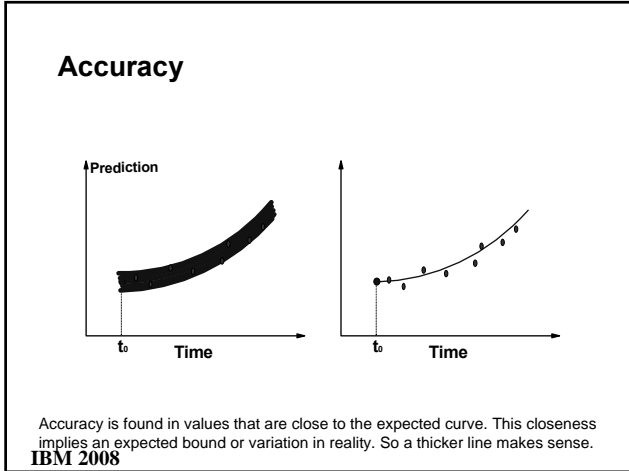
IBM 2008

How Accurate Is It?



Starting from an initial point of maybe dubious accuracy, we apply a growth rate (also dubious) and then recommend actions costing lots of money.

IBM 2008



A Conversation

You: The answer is 42.67.
 Them: I measured it and the answer is 42.663!
 You: Give me a break.
 Them: I just want to be exact.
 You: OK the answer is around 42.67.
 Them: How far around.
 You: ????

IBM 2008

Confidence Interval or How Thick is the Line?

$P[\mu - 2\sigma < X < \mu + 2\sigma] = 0.954$

$P[\mu - 1.96\sigma < X < \mu + 1.96\sigma] = 0.95$ or 95%

[L,U] is called the 100(1-α)% confidence interval.

1-α is called the level of confidence associated with [L,U]

IBM 2008

Confidence Interval

$[\mu - 1.96 \sigma/n , \mu + 1.96 \sigma/n]$

$[\mu - z_{\alpha/2} \sigma/n , \mu + z_{\alpha/2} \sigma/n]$

Using a Standard Normal Probability table, 95% confidence (2 tail) is found by looking for a z score of 0.025.

In Excel: =Confidence(μ, σ, n)

=Confidence(0.5,1,100) = 1.96

IBM 2008

Summary

Given a list of numbers X=(X_i) i=1 to n

Term	Formula	Excel	PS View
Count (number of items)	n	=Count(X)	Number of points plotted
Average	$\bar{X} = \text{Sum}(X)/n$	=Average(X)	Center of gravity
Median§	$X[\text{ROUND DOWN } 1+n*0.5]$	=MEDIAN(X)	Middle number
Variance	$V = \frac{\sum(X_i - \bar{X})^2}{n}$	=Var(X)	Spread of data
Standard Deviation	$s = \text{SQRT}(V)$	=Stnd(X)	Spread of data
Coefficient of Variation (Std/Avg)	$CV = s/\bar{X}$		Spread of data around average
Minimum	First in Sorted list	=MIN(X)	Bottom of plot
Maximum	Last in Sorted list	=MAX(X)	Top of plot
Range	[Minimum,Maximum]		Distance between top and bottom
90th percentile§	$X[\text{ROUND DOWN } 1+n*0.9]$	=Percentile(X,0.9)	10% from the top
Confidence interval	Look in book	=Confidence(0.05,s,n)	Expected Variability of average (a thick line)

§= Percentile formulae assume a sorted list; Low to high.

IBM 2008

Linear Regression (for Trending)

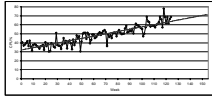
$y = 3.0504x + 385.42$

$R^2 = 0.7881$

Obtain a useful fit of the data ($y = mx+b$) and then extend the values of X to obtain predicted values of Y. But remember as Niels Bohr said: "Prediction is very hard to do. Especially about the future."

IBM 2008

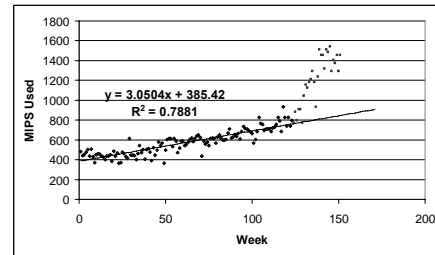
Trending Assumptions & Questions



- The future will be like the past.
- How much history is too much?
- You should look at Era segments.
- Shape and scale of graph can be interesting.
- You may need more than numbers.... The business and technical environment?
- Be smart and lazy.... What questions are you answering?

IBM 2008

Reality



Linear regression's predictions assume that the future looks like the past.

IBM 2008

Coding Implementation

The Butterfly Effect

Algorithm 1:

$X_{n+1} = s * X_n$ if $X_n < 0.5$
 $X_{n+1} = s * (1 - X_n)$ otherwise
 In Excel: cell X_{n+1} is =IF($X_n < 0.5$, $S * X_n$, $S * (1 - X_n)$)

Algorithm 2:

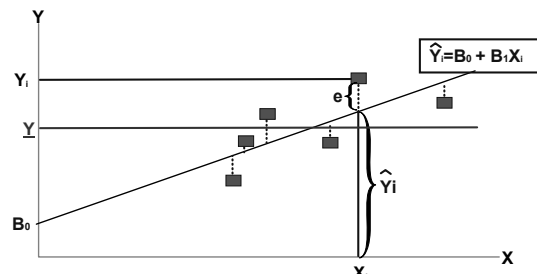
$X_{n+1} = s * (0.5 - |X_n - 0.5|)$
 In Excel: cell X_{n+1} is =S*(0.5-ABS($X_n - 0.5$))

Mathematically Equal.

(Ref. Chaos Under Control, section on Butterfly Effect.)

IBM 2008

Linear Fit for $\{X_i, Y_i\}$



$$\text{Goodness of Fit } R^2 = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2}$$

On the line would be perfect.
 Next to that would be a line with minimum error (e).
 Actually minimum e^2 is better.

IBM 2008

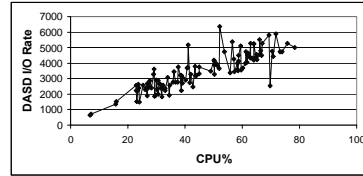
Excel Help

Search Excel Help for *R Squared* return:

RSQ: Returns the square of the Pearson product moment correlation coefficient through data points in known_y's and known_x's. For more information, see PEARSON. The r-squared value can be interpreted as the proportion of the variance in y attributable to the variance in x.

IBM 2008

Correlation



$$\begin{aligned} \text{Correlation} &= \text{COV}(X,Y) / \sigma_x \sigma_y \\ &= \sigma_{xy}^2 / \sigma_x \sigma_y \\ &= E[(x-\mu_x)(y-\mu_y)] / \sigma_x \sigma_y \\ \text{Correlation } &\in [-1,1] \\ \text{=CORREL(CPU\%,DASDIO)} &= 0.86 \end{aligned}$$

IBM 2008

Briefly: Correlation is not Causality

Cause → Effect (sufficient cause)
 ~Effect → ~Cause (necessary cause)

R² or CORR(C,E) may indicate a linear relationship without there being a causal connection.

In cities of various sizes:

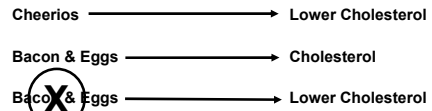
- C = number of TVs is highly correlated with E = number of murders.
- C = religious events is highly correlated with E = number of suicides.

IBM 2008

Causality & Correlation

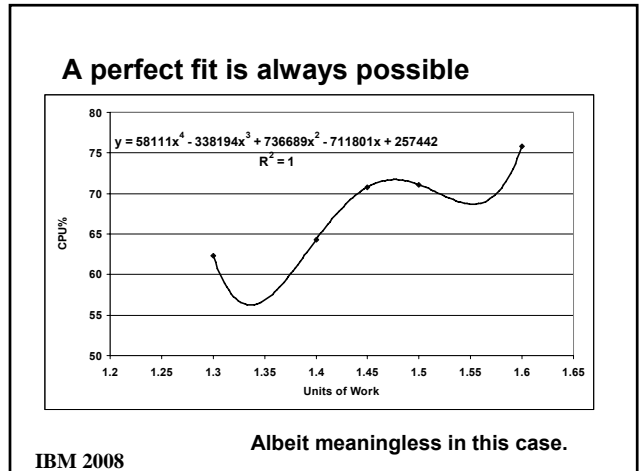
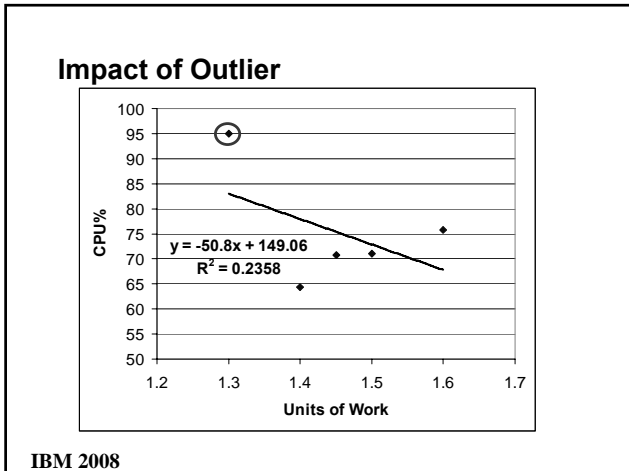
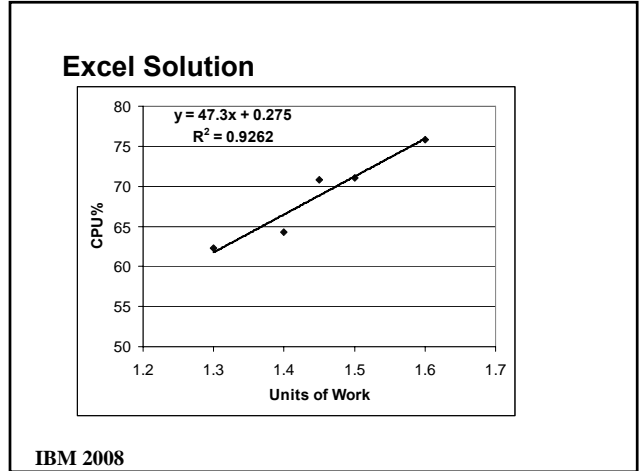
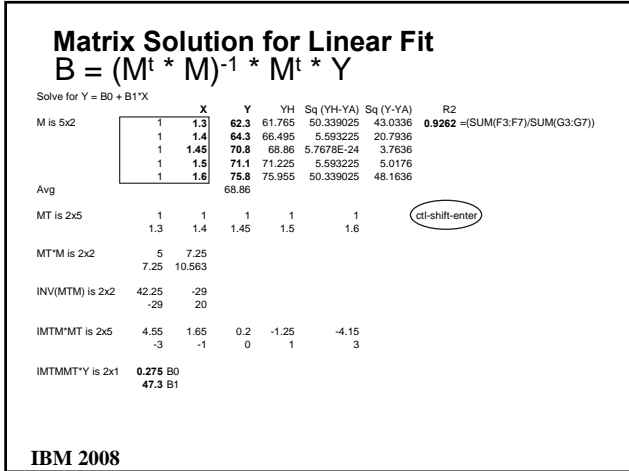
Claim: Eating Cheerios will lower your cholesterol
 Cause → Effect
 Cause: Eating Cheerios
 Effect: Lower Cholesterol

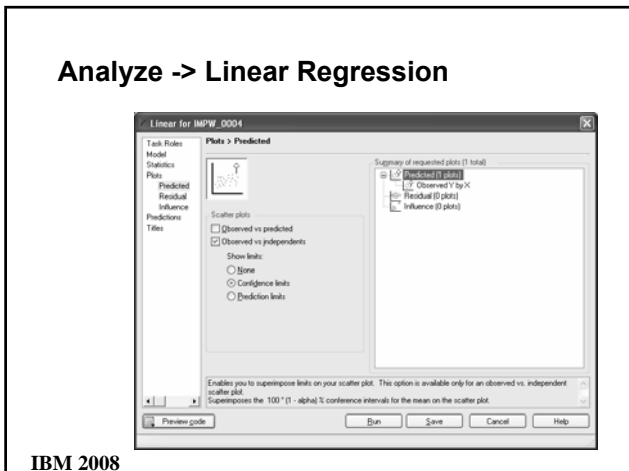
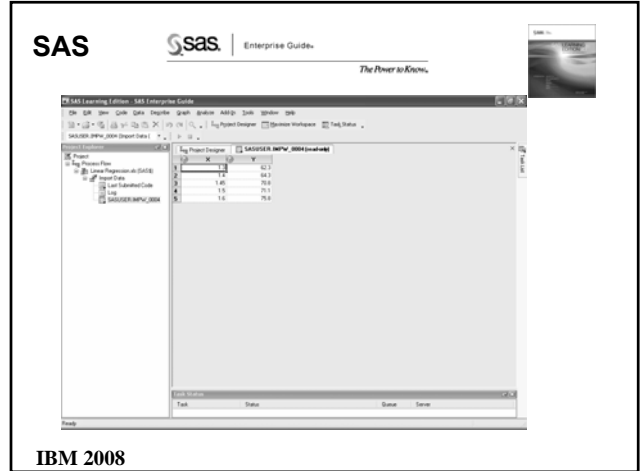
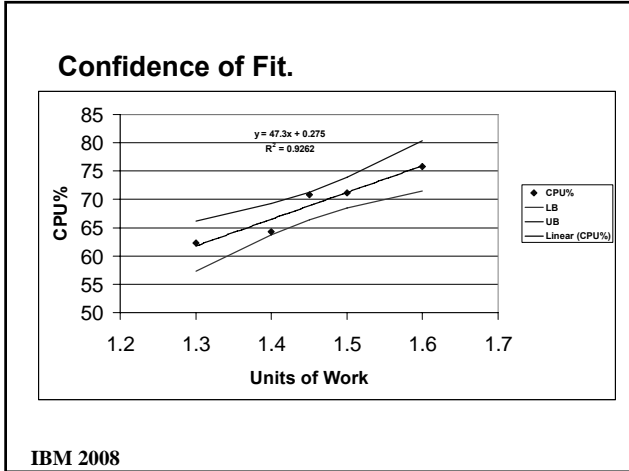
Test: Real cause
 Intervening Variable



There is a correlation between Eating Cheerios and lower Cholesterol but is there a causal relationship?

IBM 2008



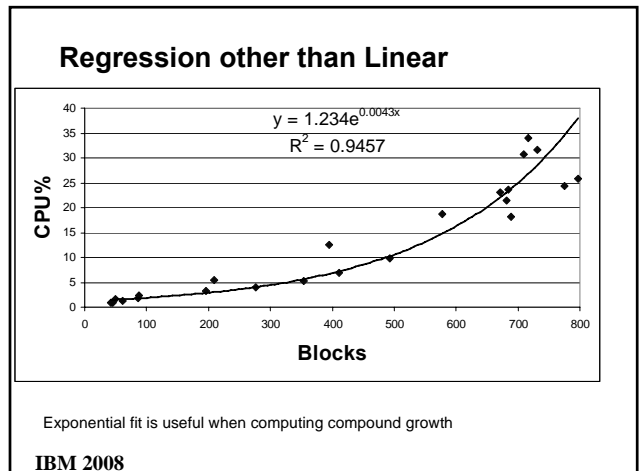
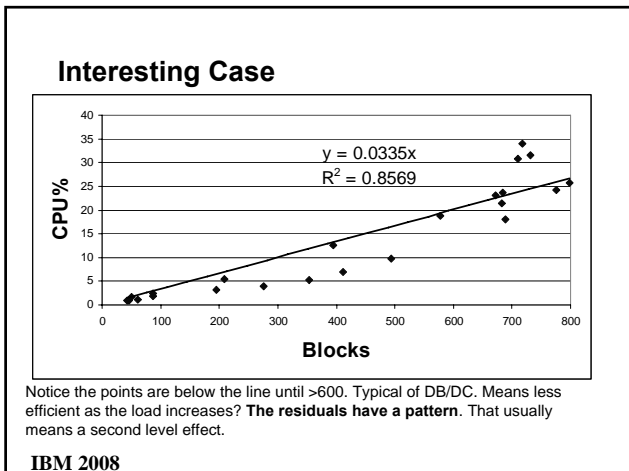
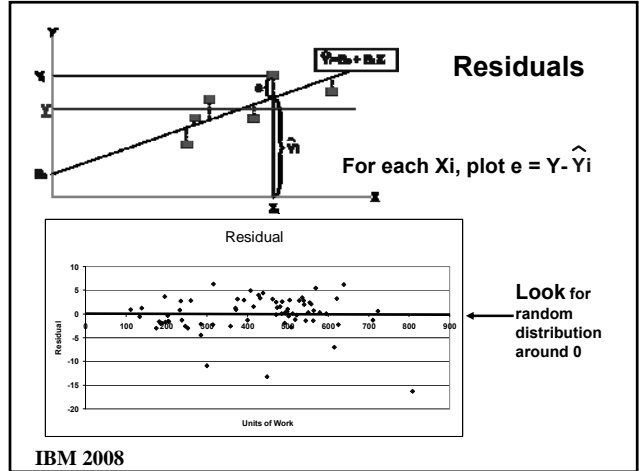
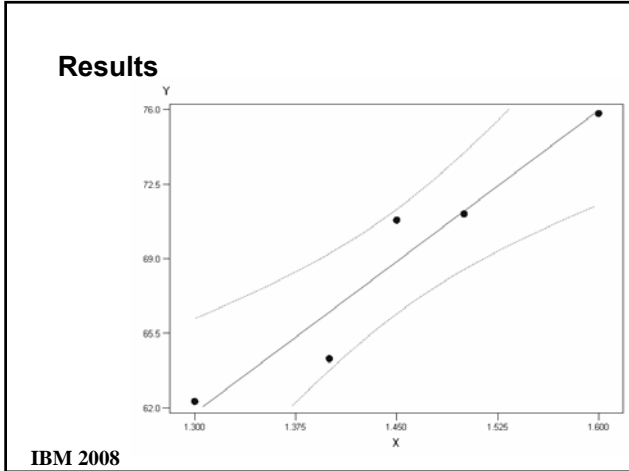


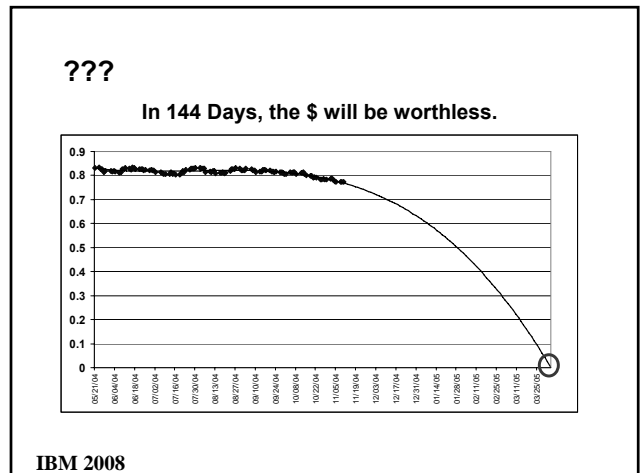
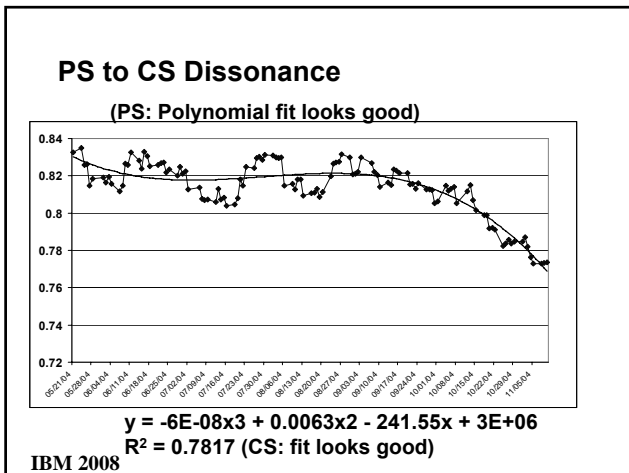
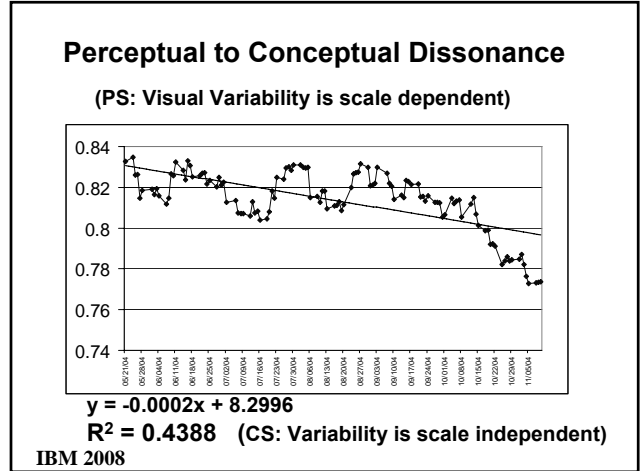
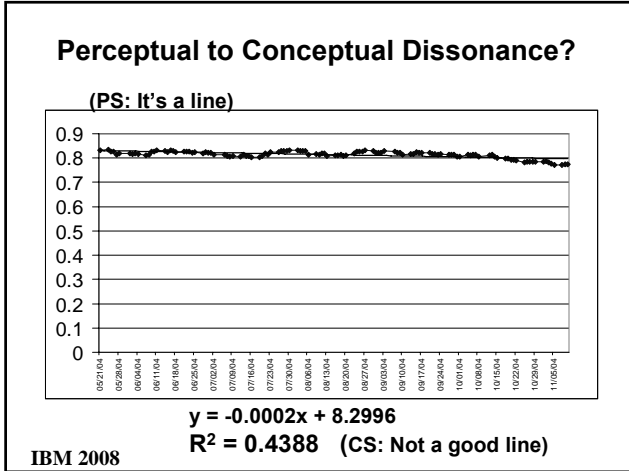
Run

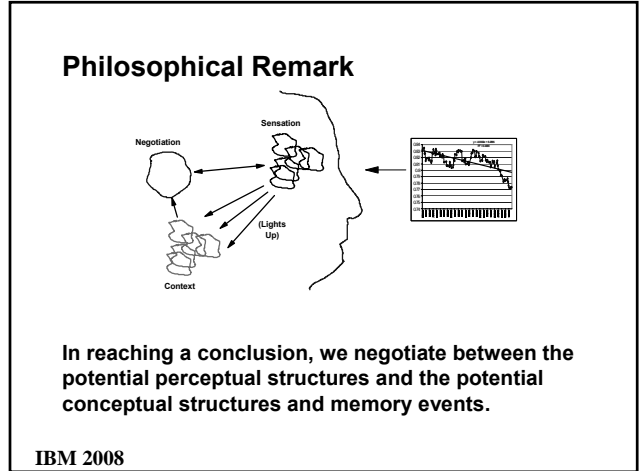
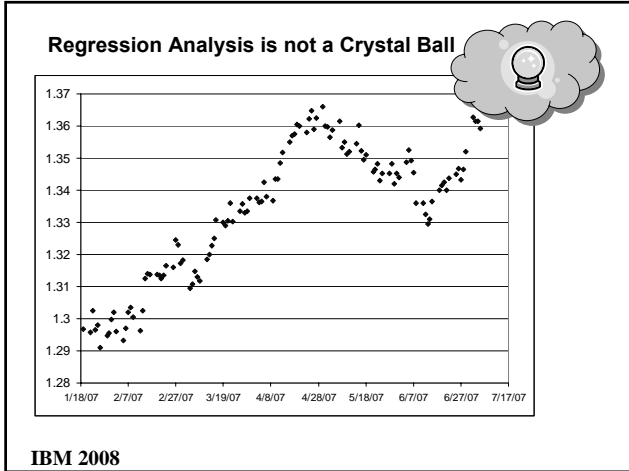
Root MSE	1.72313	R-Square	0.9262
Dependent Mean	68.86000	Adj R-Sq	0.9017
Coeff Var	2.50236		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	Intercept	1	0.27500	11.20033	0.02	0.9820
X	X	1	47.30000	7.70606	6.14	0.0087

IBM 2008







Model Building: Which is Best?

X1	X2	X3	X4	Y
7	26	6	60	78.5
1	29	15	52	74.3
11	56	8	20	104.3
11	31	8	47	87.6
7	52	6	33	96.9
11	55	9	22	109.2
3	71	17	6	102.7
1	31	22	44	72.5
2	54	18	22	93.1
21	47	4	26	115.9
1	40	23	34	83.8
11	66	9	12	113.3
10	68	8	12	109.4

Stepwise procedure to find the best combination of variables.
 $Y = b + a1X1$
 $Y = b + a1X1 + a2X2$
 $Y = b + a2X2 + a3X3$

 $Y = b + a1X1 + a2X2 + a3X3 + a4X4$

Using Hald Data from Draper

IBM 2008

Stepwise Results

Stepwise Analysis
Table of Results for General Stepwise

X4 entered.

	df	SS	MS	F	Significance F
Regression	1	1831.89816	1831.89816	22.7985202	0.000576232
Residual	11	883.8669169	80.3515379		
Total	12	2715.765077			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	117.5679312	5.262206511	22.34194552	1.62424E-10	105.9858927	129.1499696
X4	-0.738161808	0.154595996	-4.774779597	0.000576232	-1.078425302	-0.397898315

X1 entered.

	df	SS	MS	F	Significance F
Regression	2	2841.000965	1320.500482	176.6269631	1.58106E-08
Residual	10	74.76211216	7.476211216		
Total	12	2715.765077			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	103.0973816	2.123983606	48.5395154	3.32834E-13	98.36485128	107.829912
X4	-0.613953628	0.048644552	-12.62122063	1.81489E-07	-0.722340445	-0.505566811
X1	1.439958285	0.13841864	10.40307211	1.10528E-06	1.131546793	1.748369777

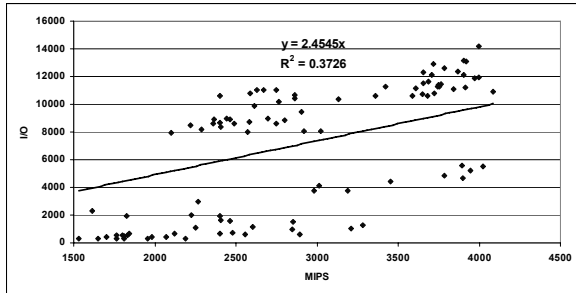
No other variables could be entered into the model. Stepwise ends.

Using Add-In from Levine

IBM 2008

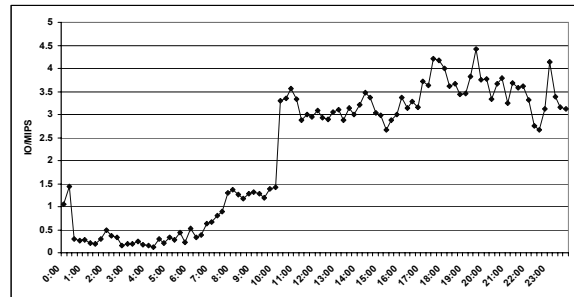
**Looking for I/O = F(MIPS).
Don't give up too quickly**

Y intercept forced to 0.



IBM 2008

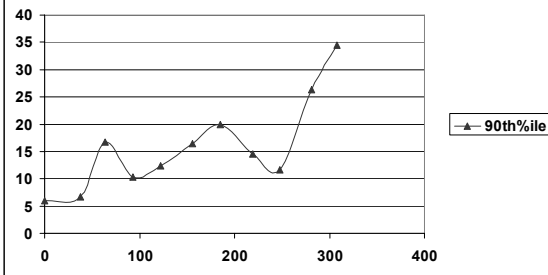
Look at ratio in time



IBM 2008

Trending: What to Do?

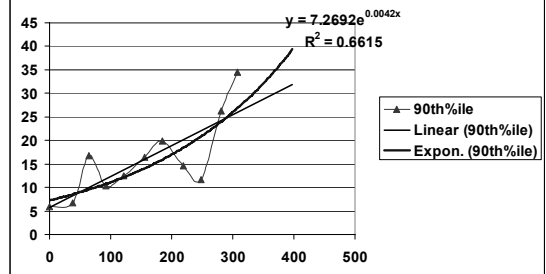
Average In & Ready



IBM 2008

Options?

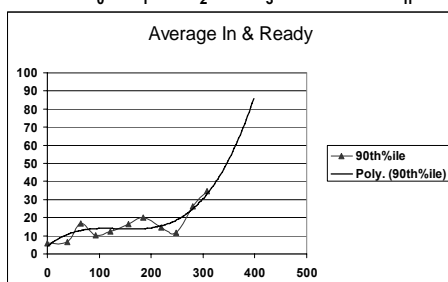
Average In & Ready



IBM 2008

How About A Polynomial?

$$Y = b_0 + b_1X + b_2X^2 + b_3X^3 + \dots + b_nX^n$$



A polynomial can be made to fit about any wandering data within the bounds of the data [min,max]. Beyond the bounds, any prediction is suspect.

IBM 2008

Time Series

A time series is a sequence of observations which are ordered in time (or space). If observations are made on some phenomenon throughout time, it is most sensible to display the data in the order in which they arose, particularly since successive observations will probably be dependent. Time series are best displayed in a scatter plot. The series value X is plotted on the vertical axis and time t on the horizontal axis. Time is called the independent variable (in this case however, something over which you have little control).

There are two kinds of time series data:

1. Continuous, where we have an observation at every instant of time e.g. lie detectors, electrocardiograms. We denote this using observation X at time t , $X(t)$.
2. Discrete, where we have an observation at (usually regularly) spaced intervals. We denote this as X_t .

See http://www.cas.lancs.ac.uk/glossary_v1.1/tsd.html#timeseries

IBM 2008

Bibliography

- *Applied Regression Analysis*, Draper & Smith, Wiley. Definitive source for regression analysis. Highly technical.
- *Statistical Concepts and Methods*, Bhattacharyya & Johnson, Wiley, 1977. This has both a discussion of meaning and the formulae.
- *Applied Statistics for Engineers and Scientists*, Levine, Ramsey & Smidt, Prentice Hall, 2001. This has a good approach to statistics and Excel implementations. CD comes with the book which has some powerful Excel Add-ins.
- *The Art of Computer Systems Performance Analysis*, by Raj Jain, Wiley. I like this one. For performance analysis and capacity planning, it is thorough and complete. A very good reference. It may be hard to find.
- *Chaos Under Control*, by Peak & Frame, Freeman & Co.
- <http://www.itl.nist.gov/div898/handbook/pmc/pmc.htm> is a good web site to explore statistics.

IBM 2008