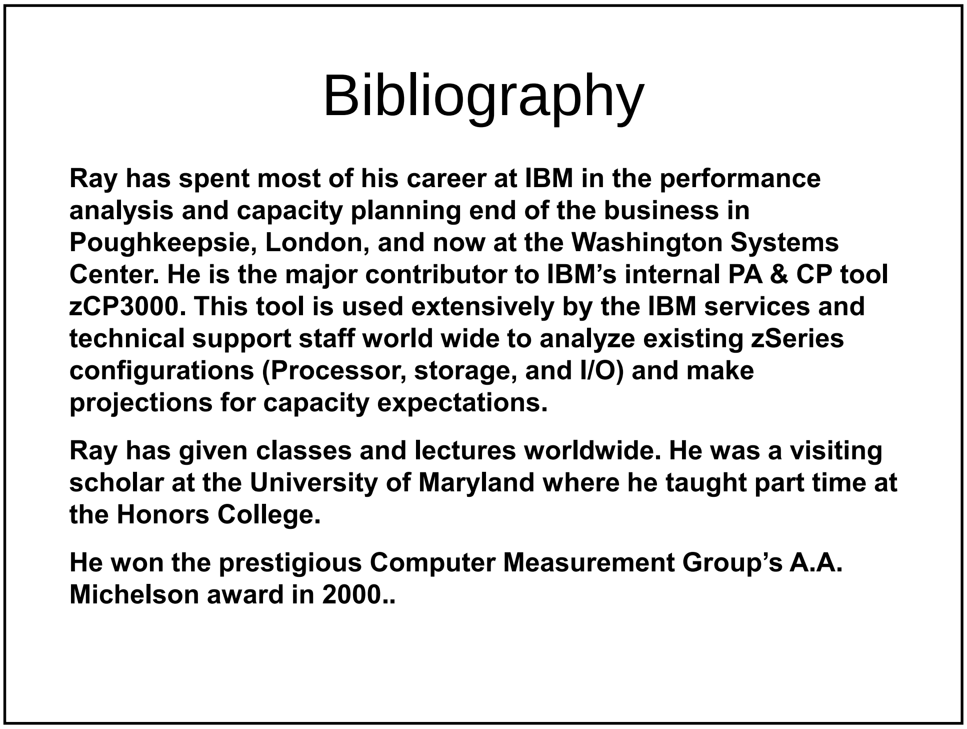
Tracking and Trending for Capacity Planning and Performance Analysis

Ray Wicks
ATS at the Washington Systems Center
561-236-5846
RayWicks@us.ibm.com

Abstract

This session covers the technicalities of simple linear Regression Analysis and the extension of this into multivariate analysis found in Time Series. The approach is generally intuitive so that one can learn what is being said and what it means. You'll see the principles of how-to and the evaluation of different regressions.

The examples used will generally be taken from system data (utilizations, rates). We will look at the reasons for both tracking and trending along with the reasons why such activities can fail. The simpler examples will use EXCEL.

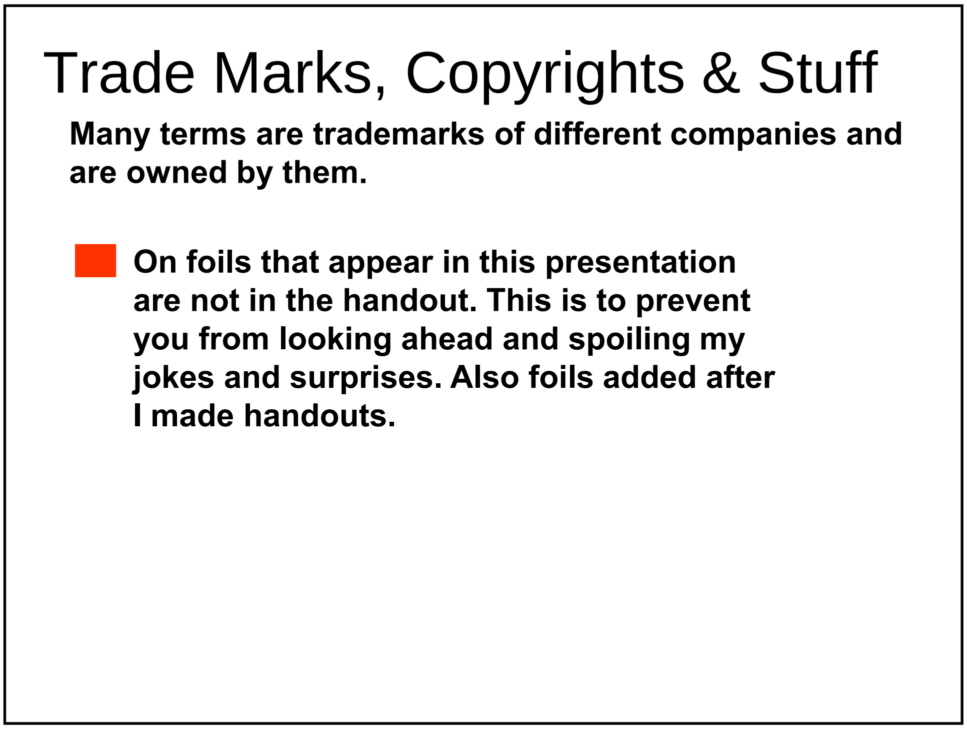


Bibliography

Ray has spent most of his career at IBM in the performance analysis and capacity planning end of the business in Poughkeepsie, London, and now at the Washington Systems Center. He is the major contributor to IBM’s internal PA & CP tool zCP3000. This tool is used extensively by the IBM services and technical support staff world wide to analyze existing zSeries configurations (Processor, storage, and I/O) and make projections for capacity expectations.

Ray has given classes and lectures worldwide. He was a visiting scholar at the University of Maryland where he taught part time at the Honors College.

He won the prestigious Computer Measurement Group’s A.A. Michelson award in 2000..



Trade Marks, Copyrights & Stuff

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

- **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. Also foils added after I made handouts.**

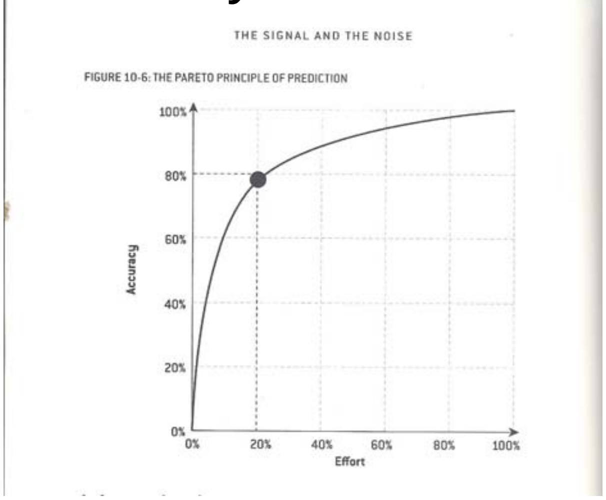
The Knowing the Future a.k.a. Prediction

Niels Bohr said: "Prediction is very hard to do. Especially about the future."

Karl Popper was asked: Will the future be like the past?
"*I do not know that the future will be like the past; on the contrary, I have good reason to expect that it will be different in many ways*"

Two issues: Accuracy and Variation.

Accuracy and Effort

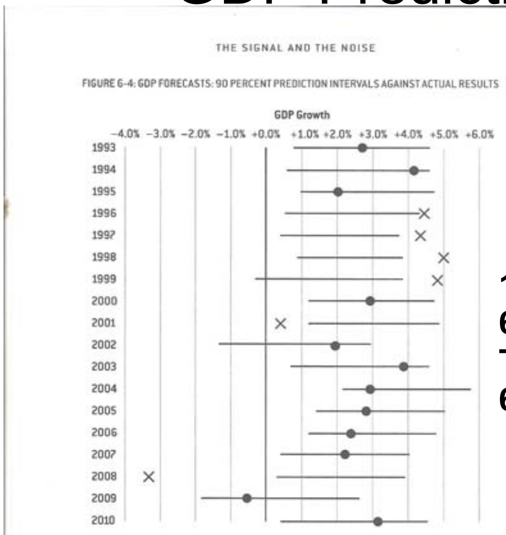


THE SIGNAL AND THE NOISE

FIGURE 10-6: THE PARETO PRINCIPLE OF PREDICTION

Message: Complete accuracy is hard, may not be needed and costs a lot. Do your questions need that accuracy?

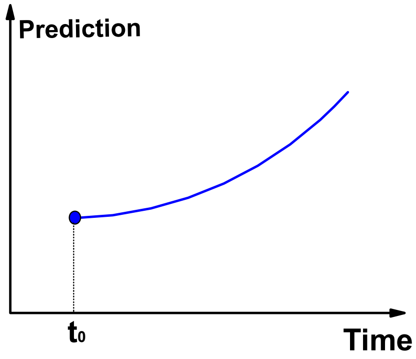
GDP Prediction



18 Predictions,
6 Wrong,
That's a Grade of
67%.

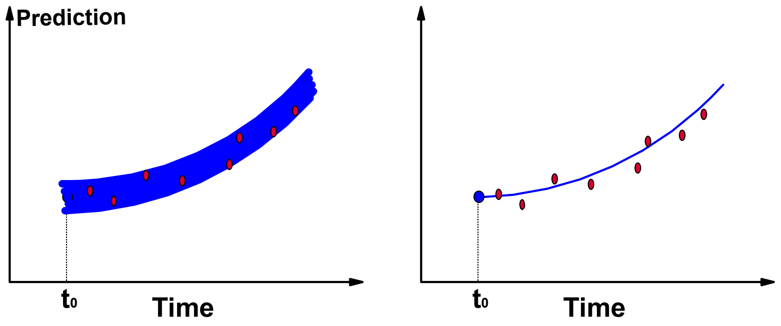
Message: Ask yourself, how is my decision effected by the prediction being wrong?

How Accurate Is It?



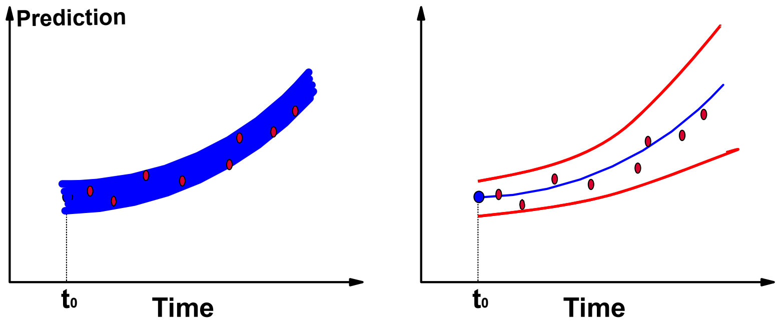
Starting from an initial point of maybe dubious accuracy, we apply a growth rate (also dubious) and then recommend actions costing lots of money.

Accuracy

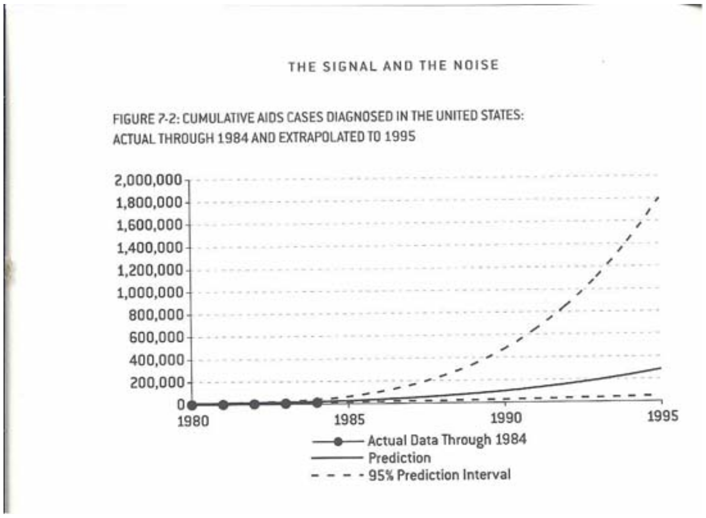


Accuracy is found in values that are close to the expected curve. This closeness implies an expected bound or variation in reality. So a thicker line makes sense.

Rather than a thick line…



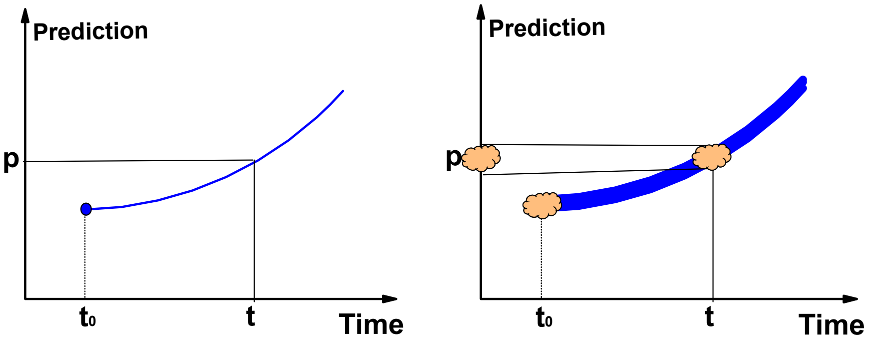
t_0 is Now and errors compound in time.



THE SIGNAL AND THE NOISE

FIGURE 7-2: CUMULATIVE AIDS CASES DIAGNOSED IN THE UNITED STATES: ACTUAL THROUGH 1984 AND EXTRAPOLATED TO 1995

How Accurate Is It?



At time t, is the prediction a precise point p or a fuzzy patch?

Accuracy

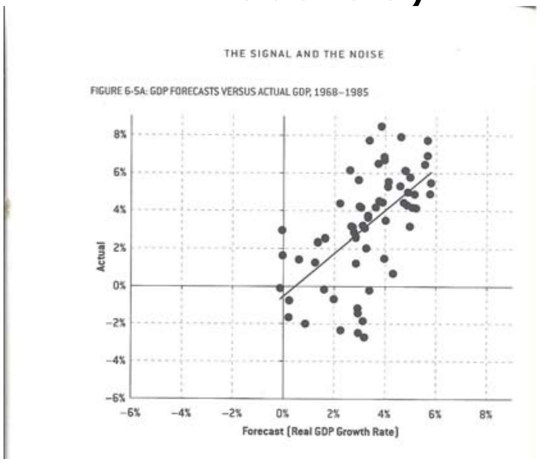

THE SIGNAL AND THE NOISE

FIGURE 6-5A: GDP FORECASTS VERSUS ACTUAL GDP, 1968–1985

Message: How far off the line is too far? And, I better track this stuff.

A Conversation

You: The answer is 42.67.

Them: I measured it and the answer is 42.663!

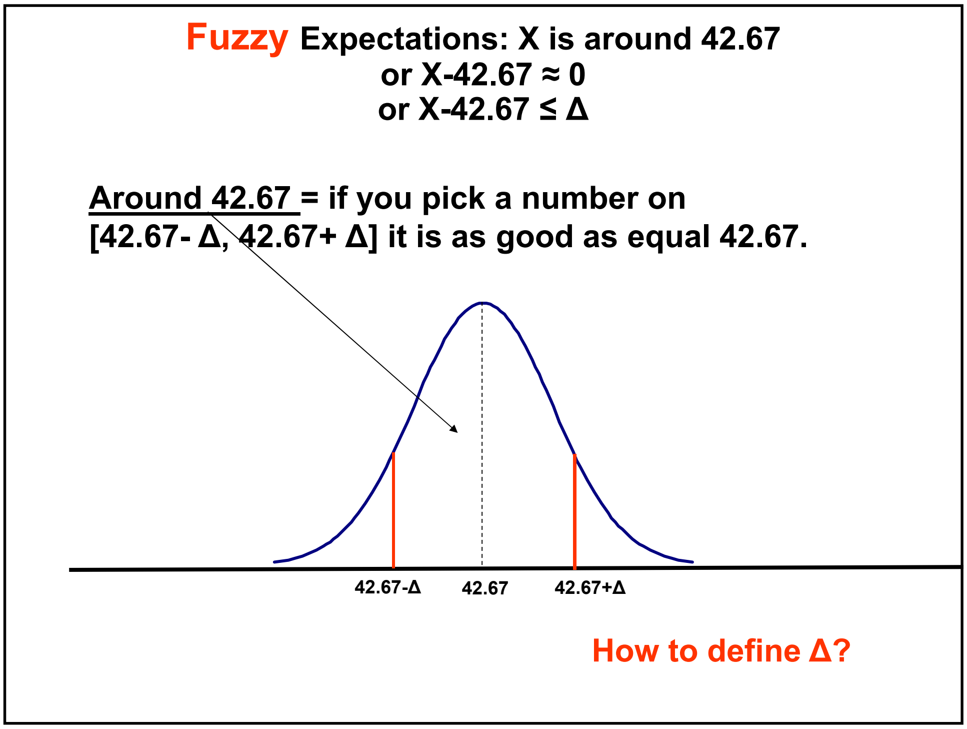
You: Give me a break.

Them: I just want to be exact.

You: OK the answer is *around* 42.67.

Them: How far around.

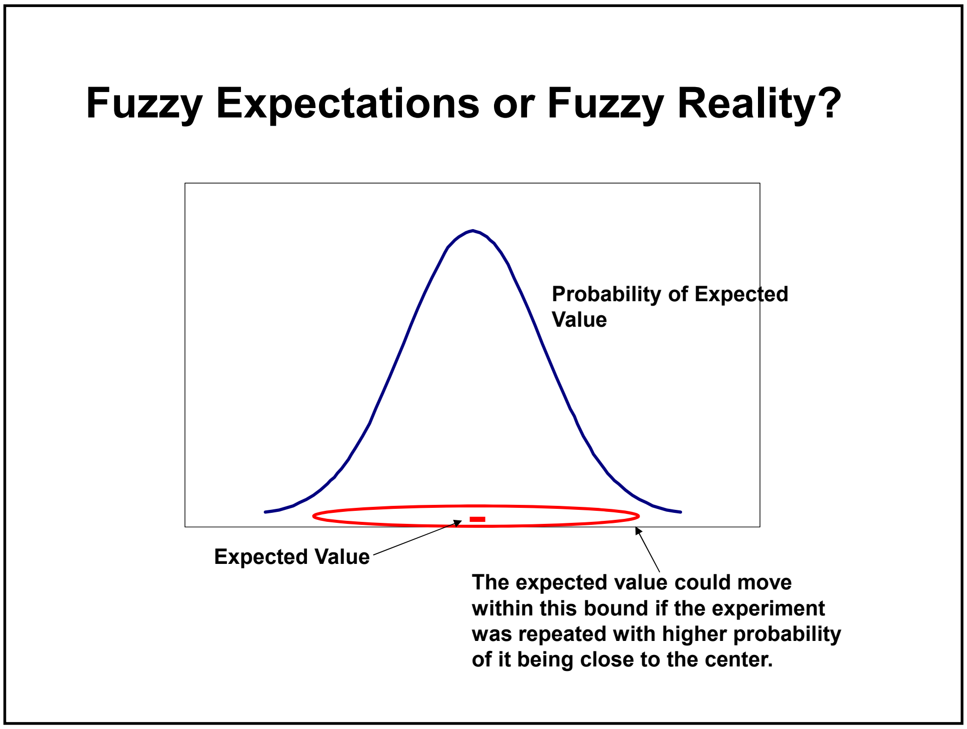
You: ????



Fuzzy Expectations: X is around 42.67
or X-42.67 ≈ 0
or X-42.67 ≤ Δ

Around 42.67 = if you pick a number on [42.67- Δ, 42.67+ Δ] it is as good as equal 42.67.

How to define Δ?



Fuzzy Expectations or Fuzzy Reality?

The expected value could move within this bound if the experiment was repeated with higher probability of it being close to the center.

Confidence Interval

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

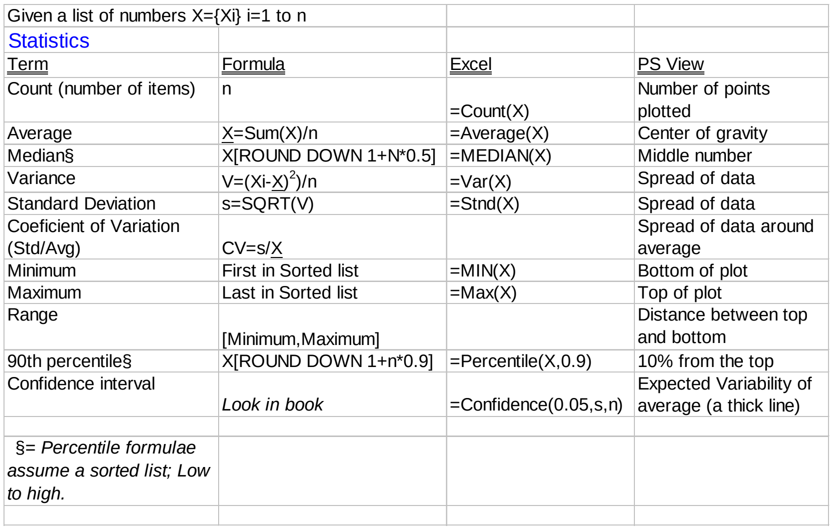
$[\ \mu - z_{\alpha/2}\ \sigma/n\ \ ,\ \mu + z_{\alpha/2}\ \ \sigma/n\ ]$

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)

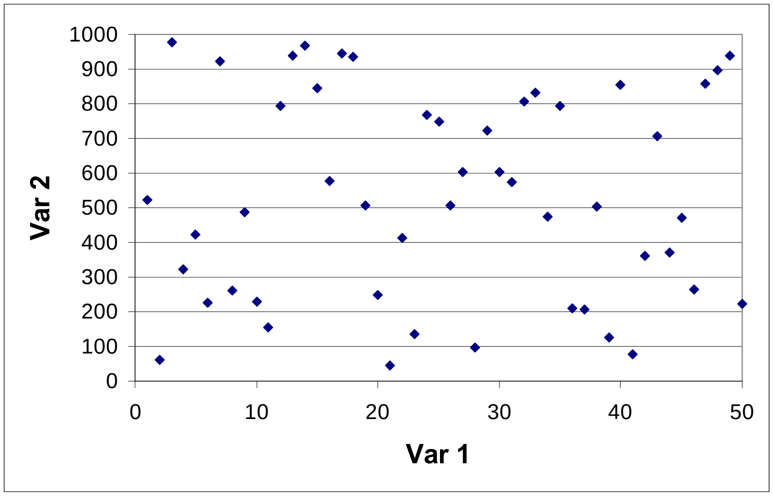
=Confidence(0.5,1,100) = 1.96

Summary



Given a list of numbers X={Xi} i=1 to n			
Statistics			
Term	Formula	Excel	PS View
Count (number of items)	n	=Count(X)	Number of points plotted
Average	X=Sum(X)/n	=Average(X)	Center of gravity
Median§	X[ROUND DOWN 1+N*0.5]	=MEDIAN(X)	Middle number
Variance	$V=(Xi-X)^2)/n$	=Var(X)	Spread of data
Standard Deviation	s=SQRT(V)	=Stnd(X)	Spread of data
Coeficient of Variation (Std/Avg)	CV=s/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[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.*			

Correlation & Prediction



Random with correlation = 0

The Intent of regression analysis

Given a set of paired observations $\{(x_i,y_i)\}$ for i=1 to n
The goal is to develop a function that uses X as a predictor of Y.
$\underline{Y}$ = f(X) such that $y_i-\underline{y}_i$ is minimal.
Or Yi = $\underline{\text{Y}}$i + e where e is the error term.

Question: Does X cause (correlate, act as a predictor) of Y?

A concern when X is Time. Given $\{(t_i,y_i)\}$, can time be a cause? If T is peak daily period and Y is CPU%, does time of day cause CPU% level? No it is a correlate.

Briefly: Correlation is not Causality

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

R^2 or CORR(C,E) may indicate a linear relationship without there being a causal connection.

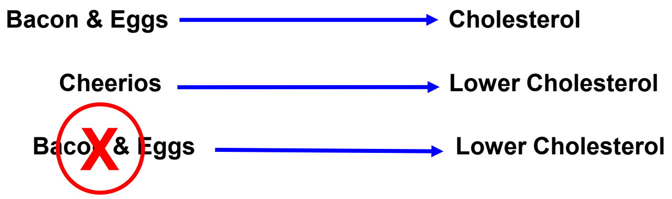
In cities of various sizes:

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

Causality & Correlation

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

Test: Real cause
Intervening Variable

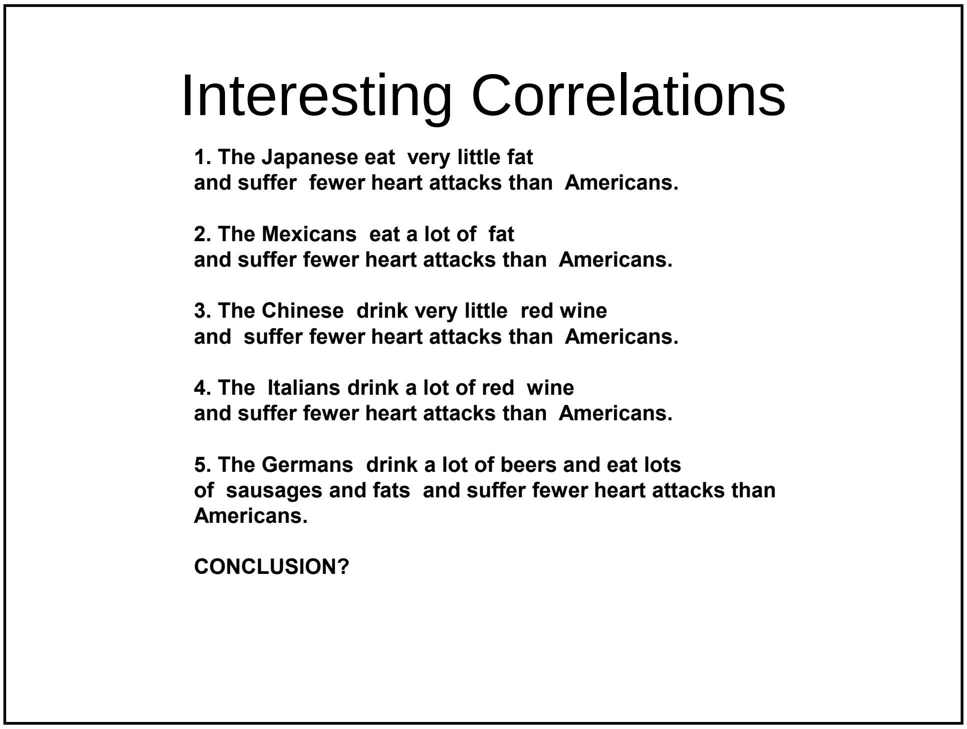


Bacon & Eggs → Cholesterol

Cheerios → Lower Cholesterol

Bacon & Eggs → Lower Cholesterol

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



Interesting Correlations

1. The Japanese eat very little fat
and suffer fewer heart attacks than Americans.

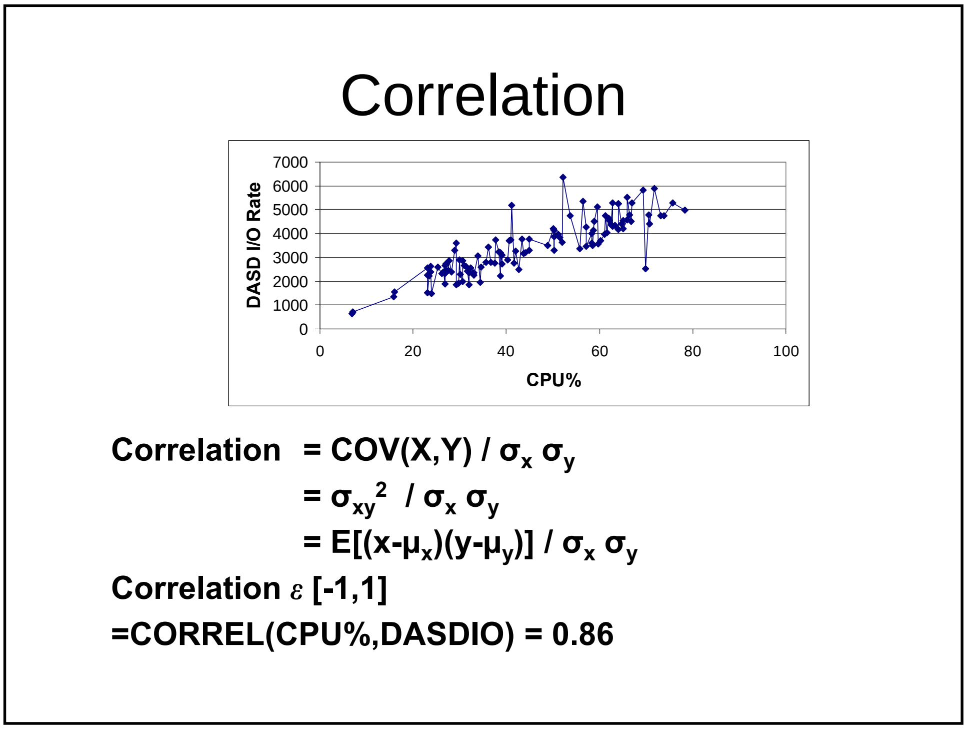
2. The Mexicans eat a lot of fat
and suffer fewer heart attacks than Americans.

3. The Chinese drink very little red wine
and suffer fewer heart attacks than Americans.

4. The Italians drink a lot of red wine
and suffer fewer heart attacks than Americans.

5. The Germans drink a lot of beers and eat lots
of sausages and fats and suffer fewer heart attacks than
Americans.

CONCLUSION?



Correlation

Correlation $= COV(X,Y) / \sigma_x \sigma_y$

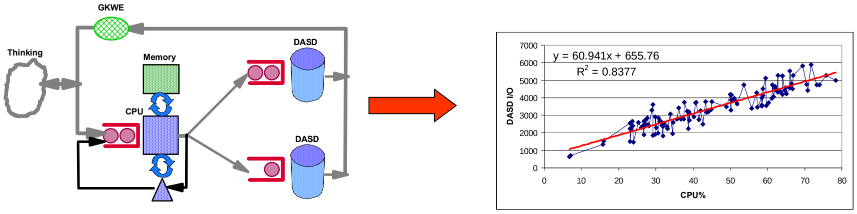
$= \sigma_{xy}^2 / \sigma_x \sigma_y$

$= E[(x-\mu_x)(y-\mu_y)] / \sigma_x \sigma_y$

Correlation ε [-1,1]

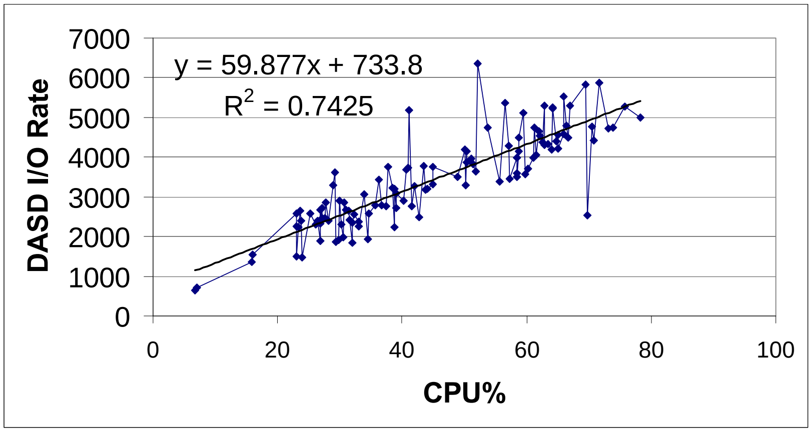
=CORREL(CPU%,DASDIO) = 0.86

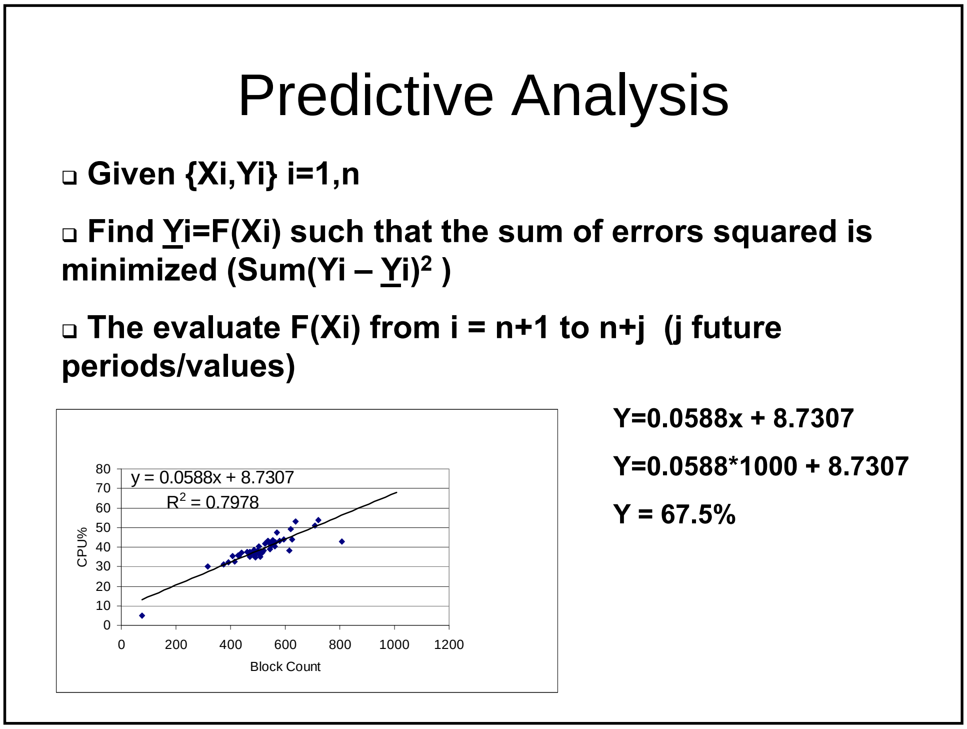
The B.S. Model



Our B.S. model anticipates a correlation between CPU time and DASD rate.

Linear Fit





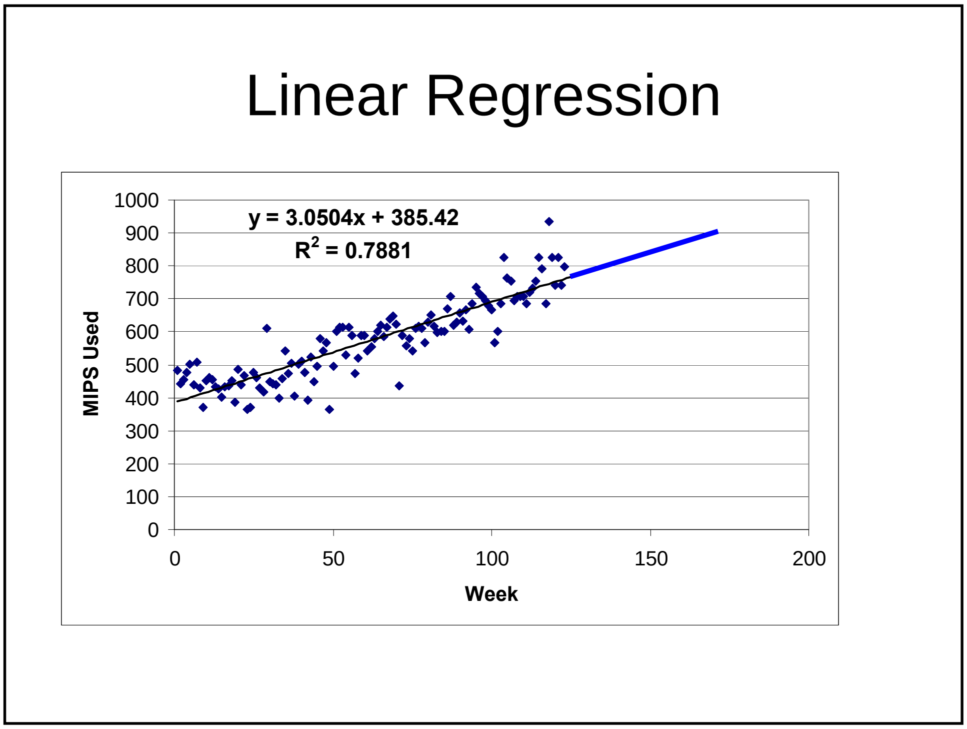
Predictive Analysis

- Given {Xi,Yi} i=1,n
- Find Yi=F(Xi) such that the sum of errors squared is minimized (Sum(Yi – Yi)^2)
- The evaluate F(Xi) from i = n+1 to n+j (j future periods/values)

Y=0.0588x + 8.7307

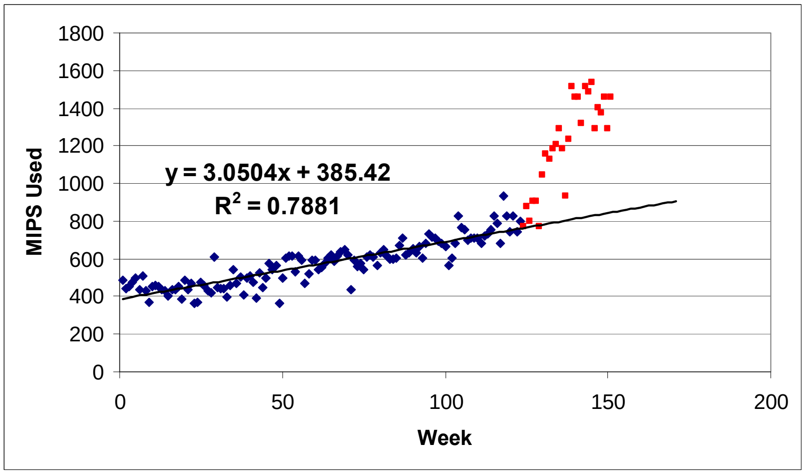
Y=0.0588*1000 + 8.7307

Y = 67.5%



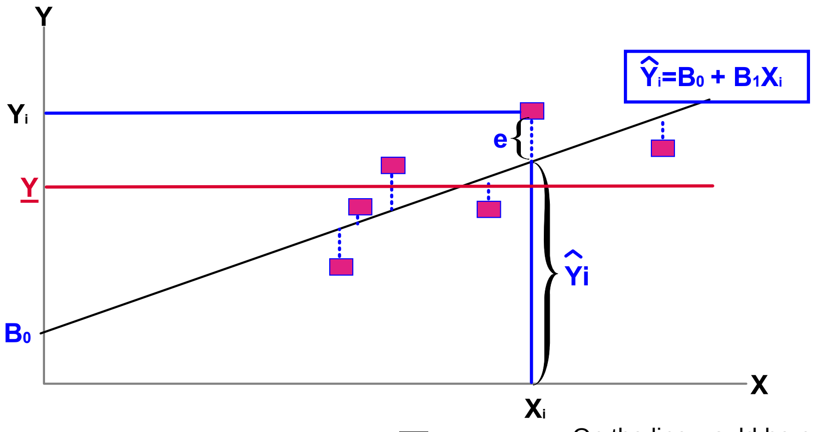
Linear Regression

Reality



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

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



Goodness of Fit $$R^2 = \frac{\sum (\hat{Y}_i - \underline{Y})_2}{\sum (Y_i - \underline{Y})_2}$$

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

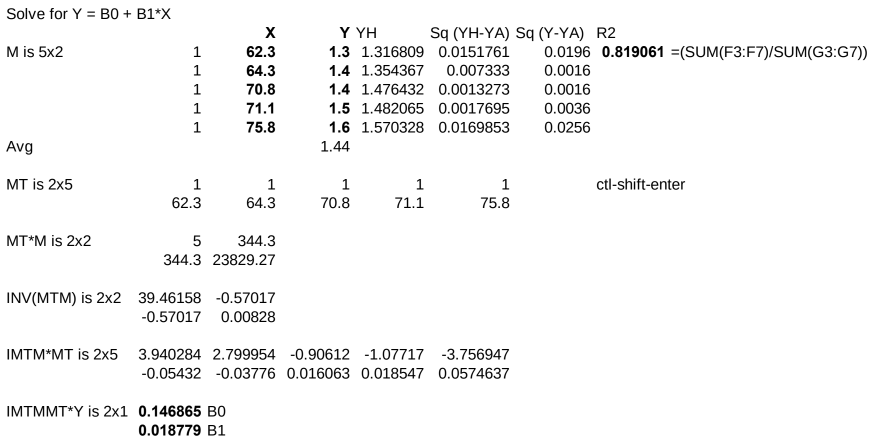
Excel Help

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.

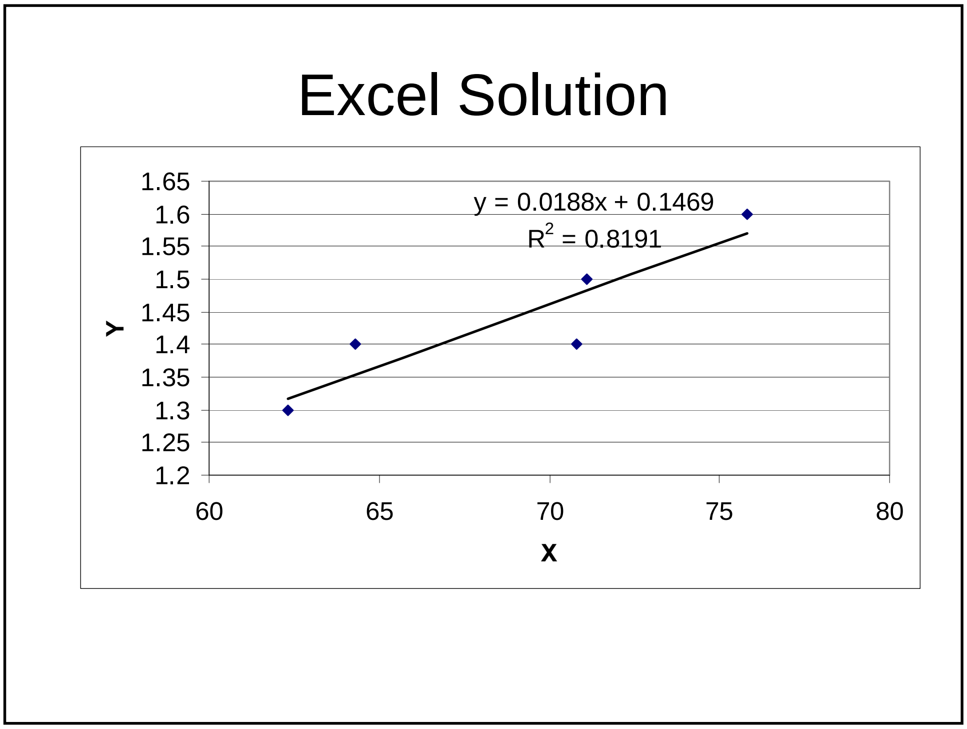
Matrix Solution for Linear Fit

$B = (M^t * M)^{-1} * M^t * Y$

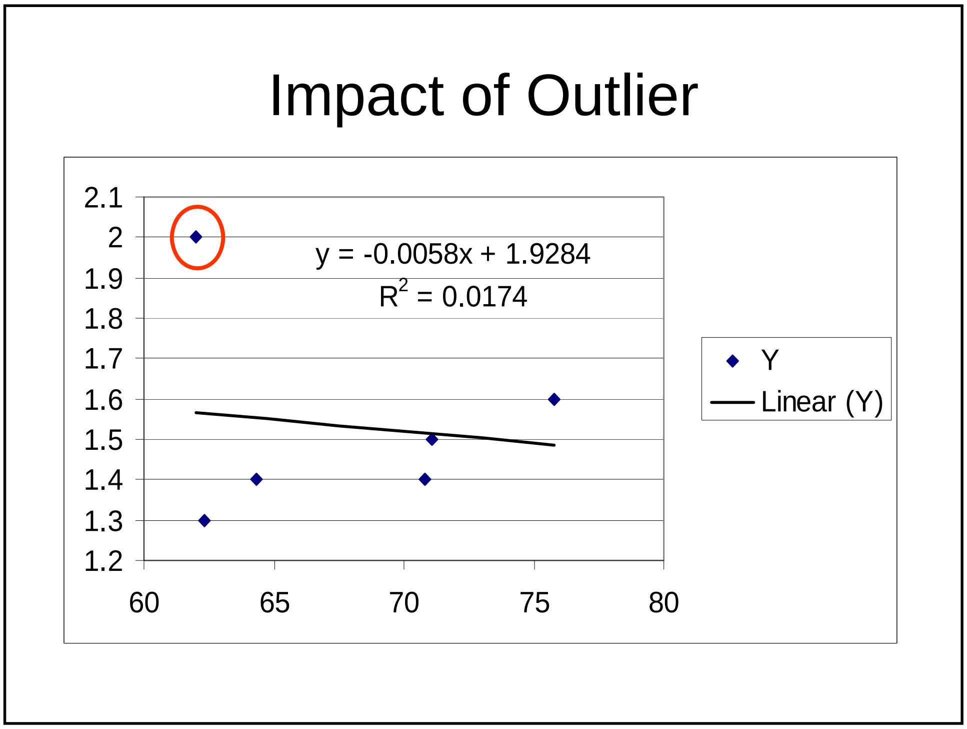


Solve for Y = B0 + B1*X

		X	**Y**	YH	Sq (YH-YA)	Sq (Y-YA)	R2	
M is 5x2	1	**62.3**	**1.3**	1.316809	0.0151761	0.0196	**0.819061**	=(SUM(F3:F7)/SUM(G3:G7))
	1	**64.3**	**1.4**	1.354367	0.007333	0.0016		
	1	**70.8**	**1.4**	1.476432	0.0013273	0.0016		
	1	**71.1**	**1.5**	1.482065	0.0017695	0.0036		
	1	**75.8**	**1.6**	1.570328	0.0169853	0.0256		
Avg			1.44					
MT is 2x5	1	1	1	1	1		ctl-shift-enter	
	62.3	64.3	70.8	71.1	75.8			
MT*M is 2x2	5	344.3						
	344.3	23829.27						
INV(MTM) is 2x2	39.46158	-0.57017						
	-0.57017	0.00828						
IMTM*MT is 2x5	3.940284	2.799954	-0.90612	-1.07717	-3.756947			
	-0.05432	-0.03776	0.016063	0.018547	0.0574637			
IMTMMT*Y is 2x1	**0.146865**	B0						
	0.018779	B1						

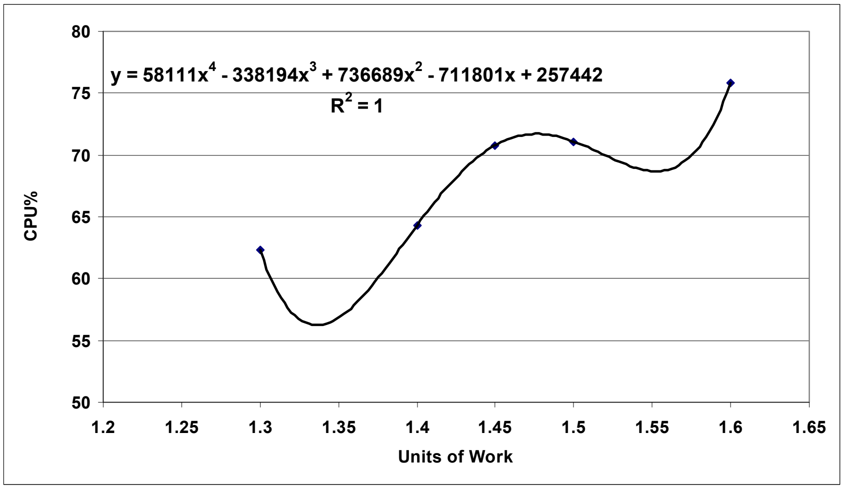


Excel Solution



Impact of Outlier

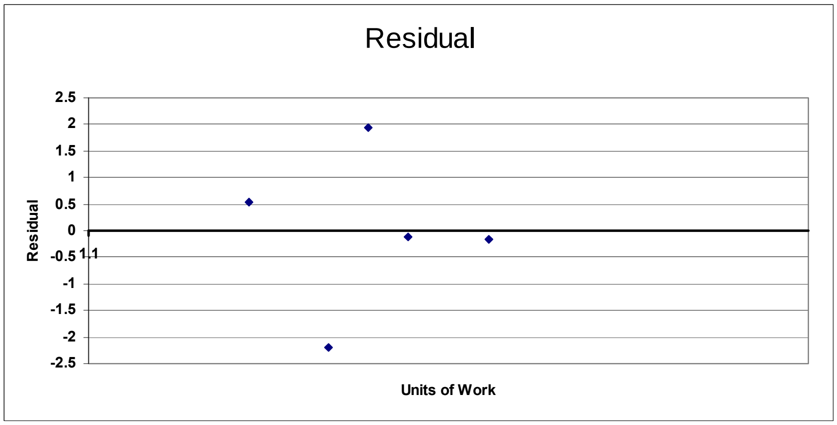
A perfect fit is always possible



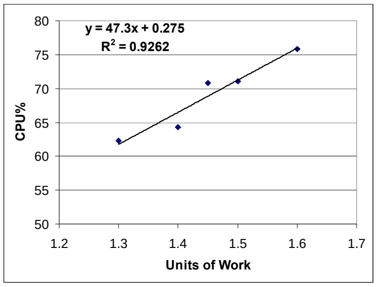
Albeit meaningless in this case.

Goodness of Fit.

Residual = Yi – Ypredict
The plot of residuals should show points randomly distributed around 0.



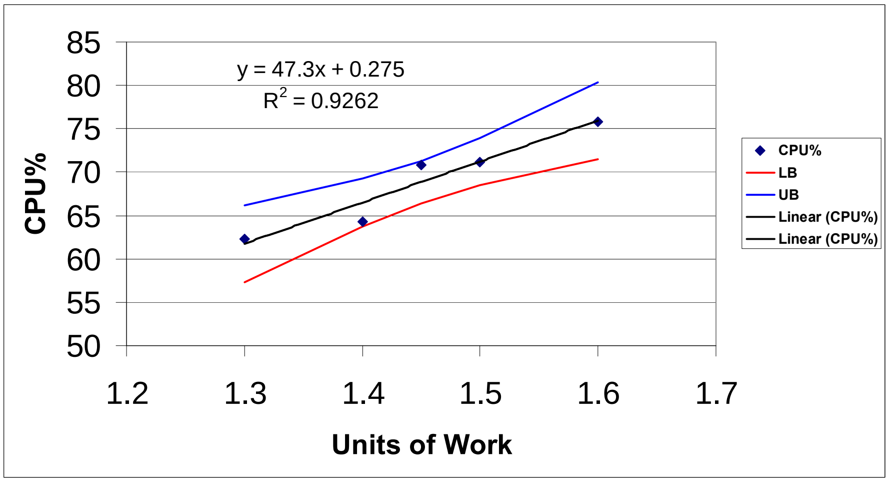
EXCEL Solution



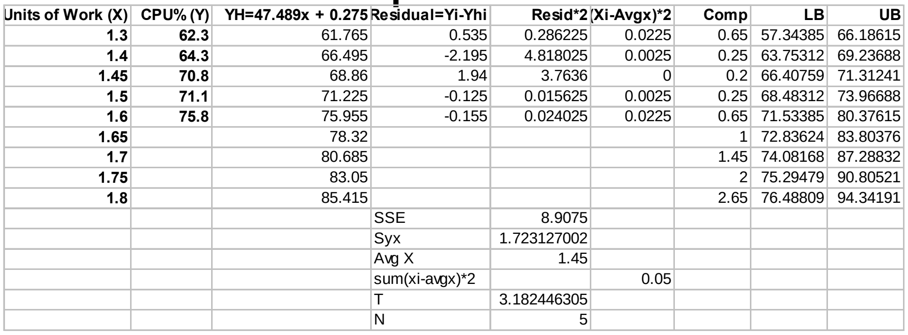
Units of Work (X)	CPU% (Y)	YH=47.3x + 0.275	Residual=Yi-Yhi	Resid*2
1.3	**62.3**	61.765	0.535	0.286225
1.4	**64.3**	66.495	-2.195	4.818025
1.45	**70.8**	68.86	1.94	3.7636
1.5	**71.1**	71.225	-0.125	0.015625
1.6	**75.8**	75.955	-0.155	0.024025

SSE	8.9075
Syx	1.723127002
Avg X	1.45
sum(xi-avgx)*2	
T	3.182446305
N	5

Solution with Bounds



Computations

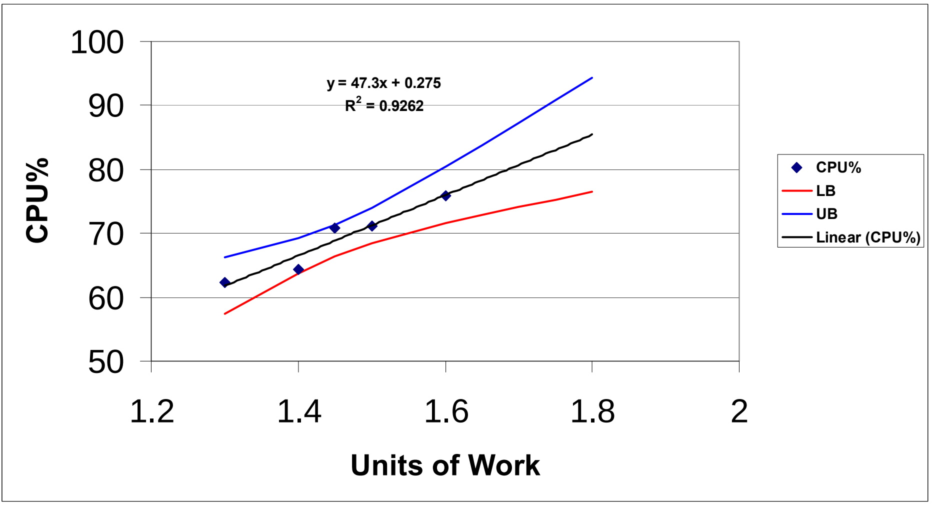


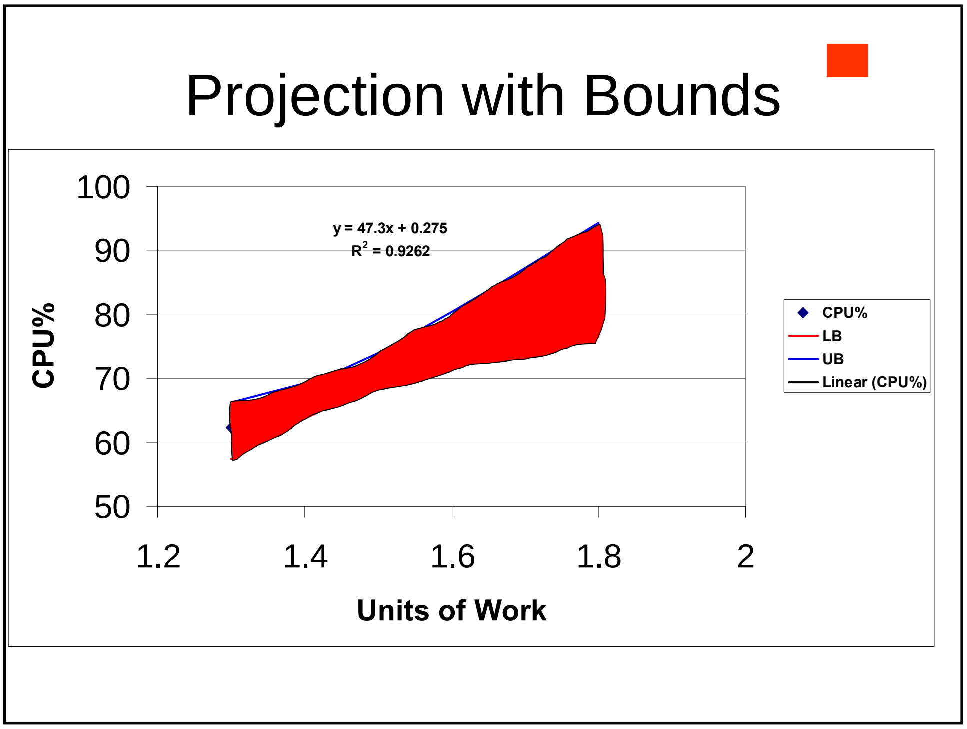
Units of Work (X)	CPU% (Y)	YH=47.489x + 0.275	Residual=Yi-Yhi	Resid*2	(Xi-Avgx)*2	Comp	LB	UB
1.3	**62.3**	61.765	0.535	0.286225	0.0225	0.65	57.34385	66.18615
1.4	**64.3**	66.495	-2.195	4.818025	0.0025	0.25	63.75312	69.23688
1.45	**70.8**	68.86	1.94	3.7636	0	0.2	66.40759	71.31241
1.5	**71.1**	71.225	-0.125	0.015625	0.0025	0.25	68.48312	73.96688
1.6	**75.8**	75.955	-0.155	0.024025	0.0225	0.65	71.53385	80.37615
1.65		78.32				1	72.83624	83.80376
1.7		80.685				1.45	74.08168	87.28832
1.75		83.05				2	75.29479	90.80521
1.8		85.415				2.65	76.48809	94.34191
			SSE	8.9075				
			Syx	1.723127002				
			Avg X	1.45				
			sum(xi-avgx)*2		0.05			
			T	3.182446305				
			N	5				

Comp: =(1/ROWS(X))+(POWER(A2-Avgx,2))/$F$14

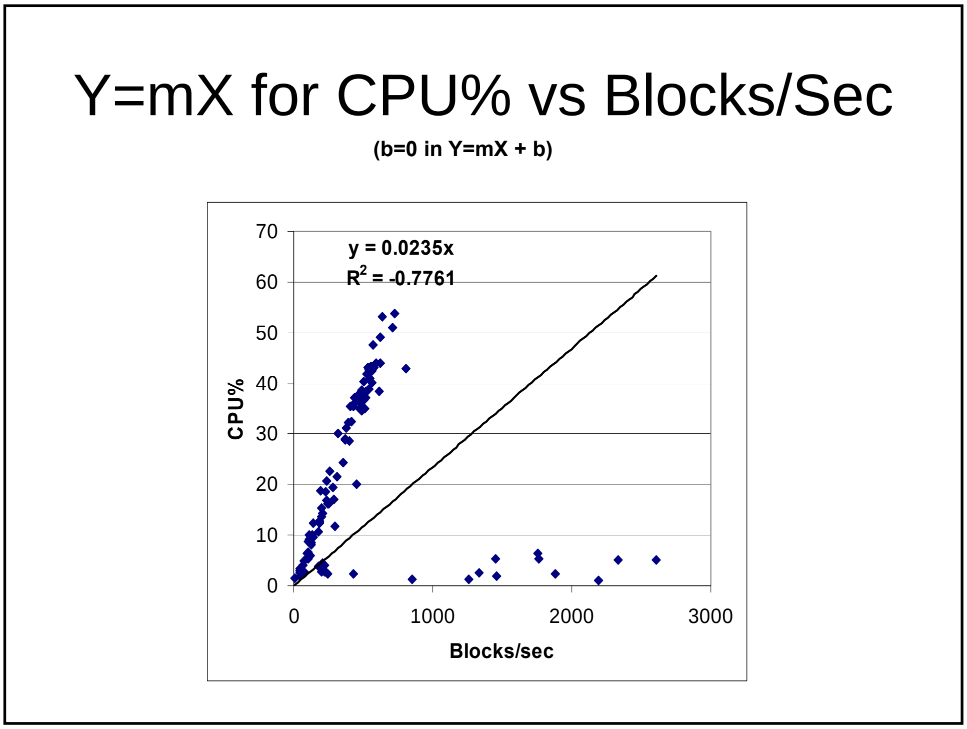
LB: =$C2-T*Sxy*SQRT($G2)

Projection with Bounds





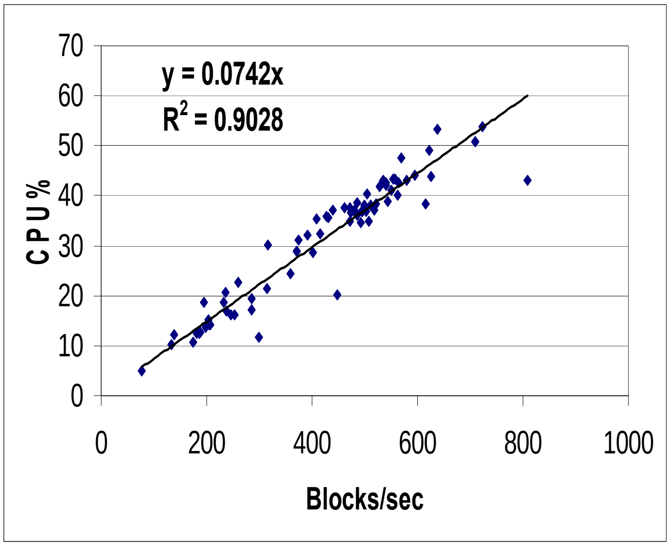
Projection with Bounds



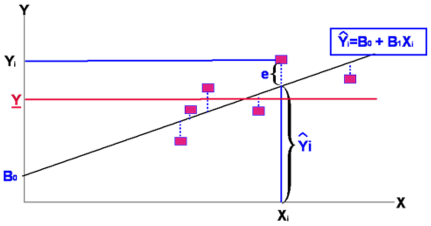
Y=mX for CPU% vs Blocks/Sec

(b=0 in Y=mX + b)

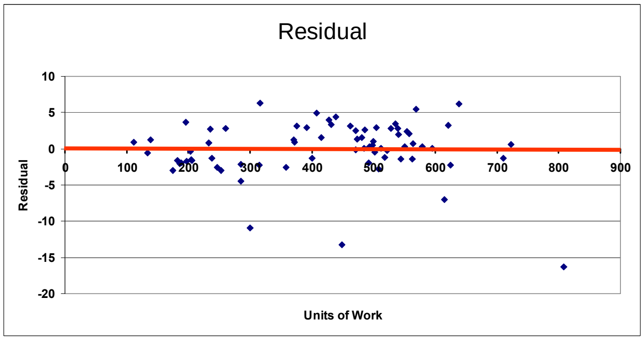
Filtered Data CPU%>10

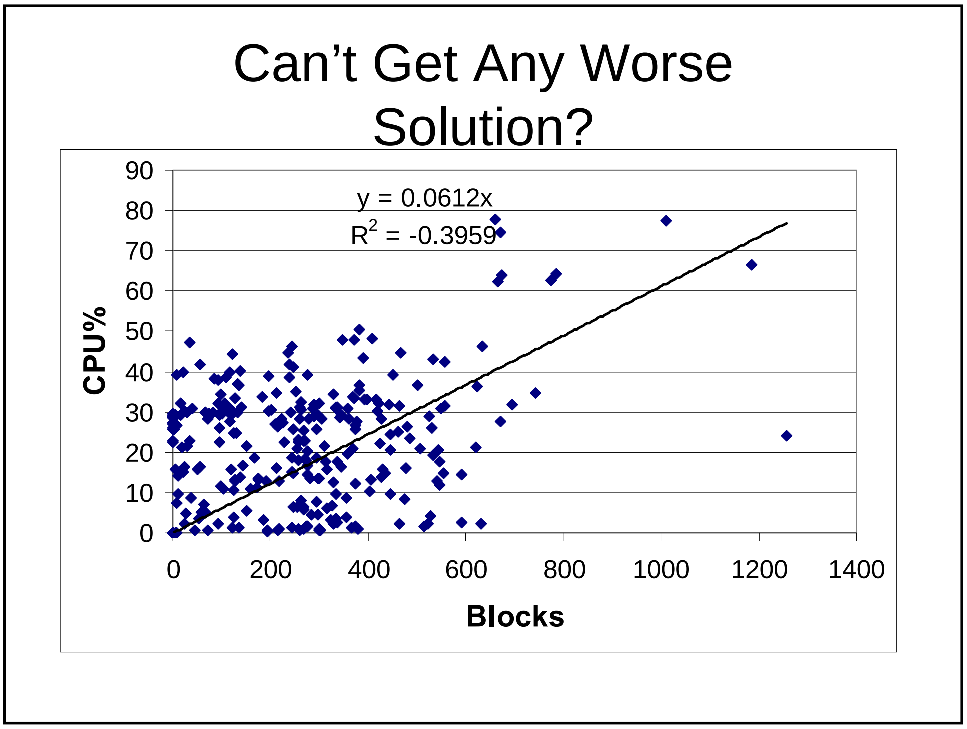


Residuals

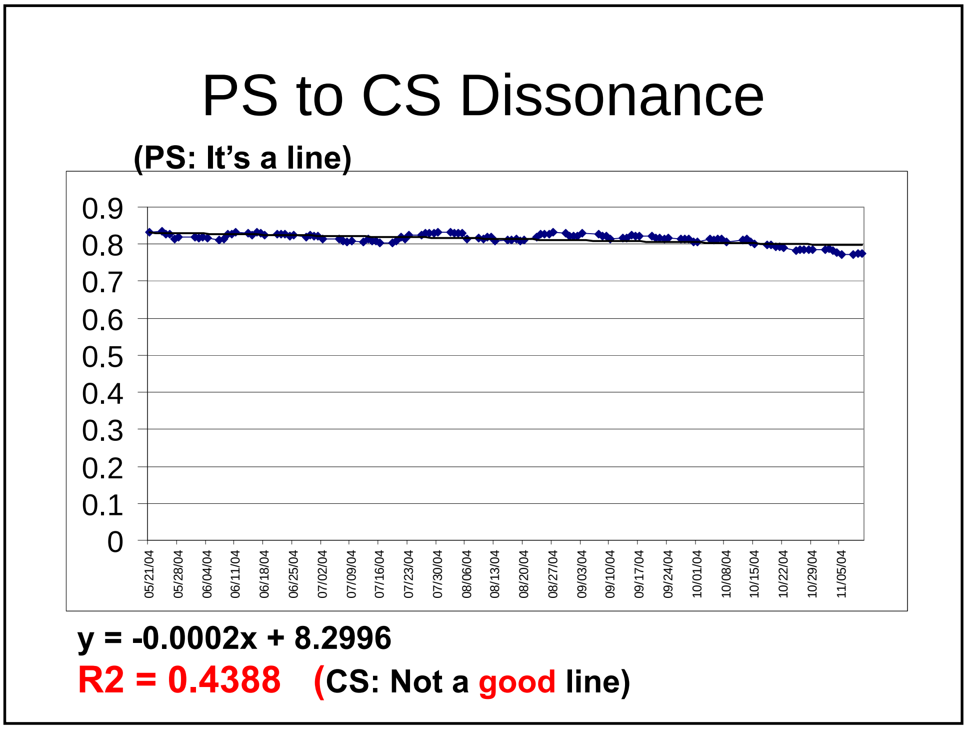


For each Xi, plot e = Y-$\hat{Y}i$





Can’t Get Any Worse Solution?

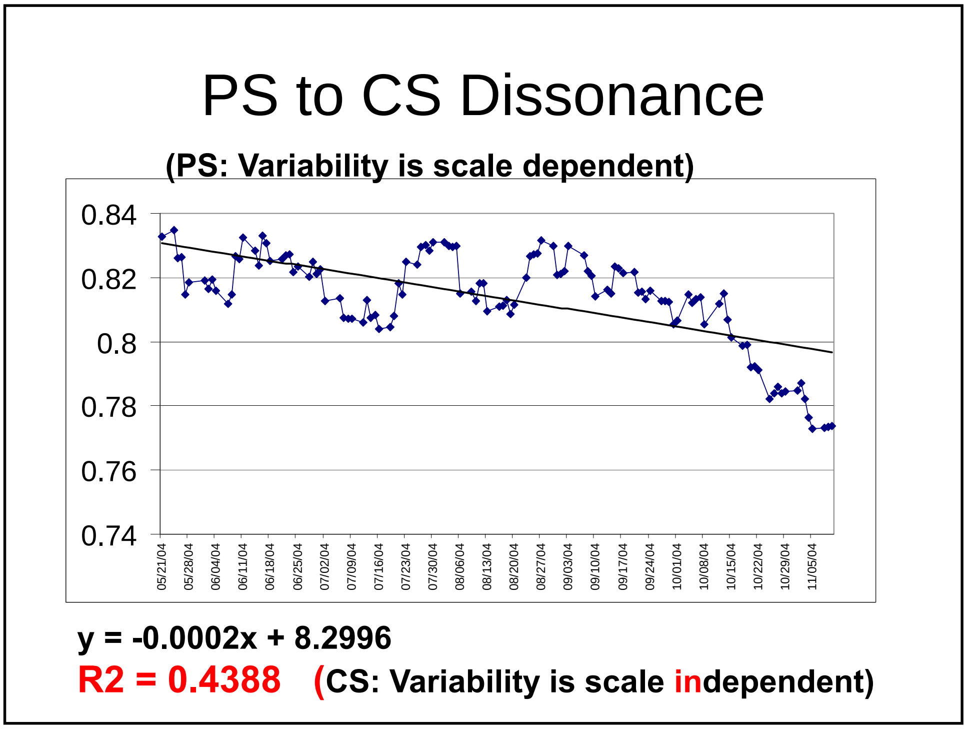


PS to CS Dissonance

(PS: It’s a line)

y = -0.0002x + 8.2996

R2 = 0.4388 (CS: Not a good line)

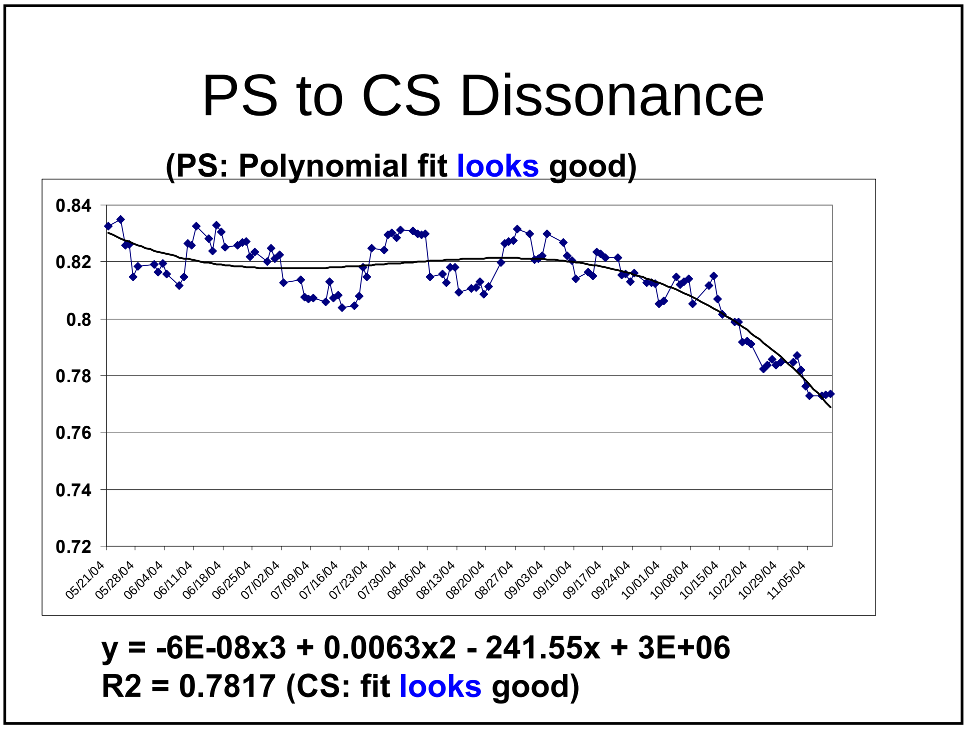


PS to CS Dissonance

(PS: Variability is scale dependent)

y = -0.0002x + 8.2996

R2 = 0.4388 (CS: Variability is scale independent)

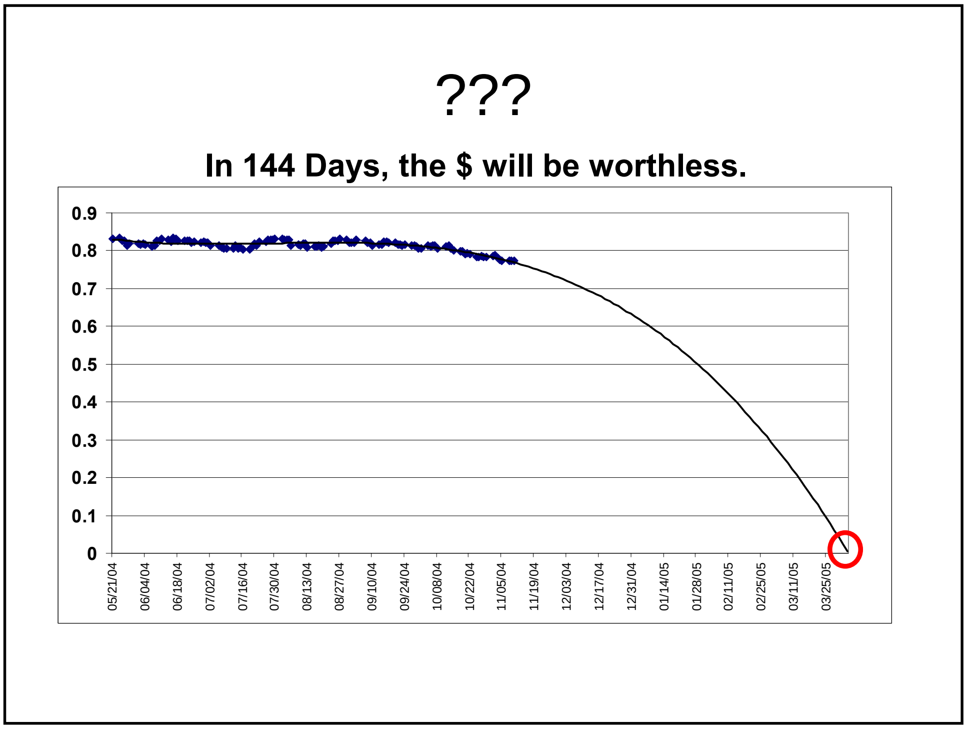


PS to CS Dissonance

(PS: Polynomial fit looks good)

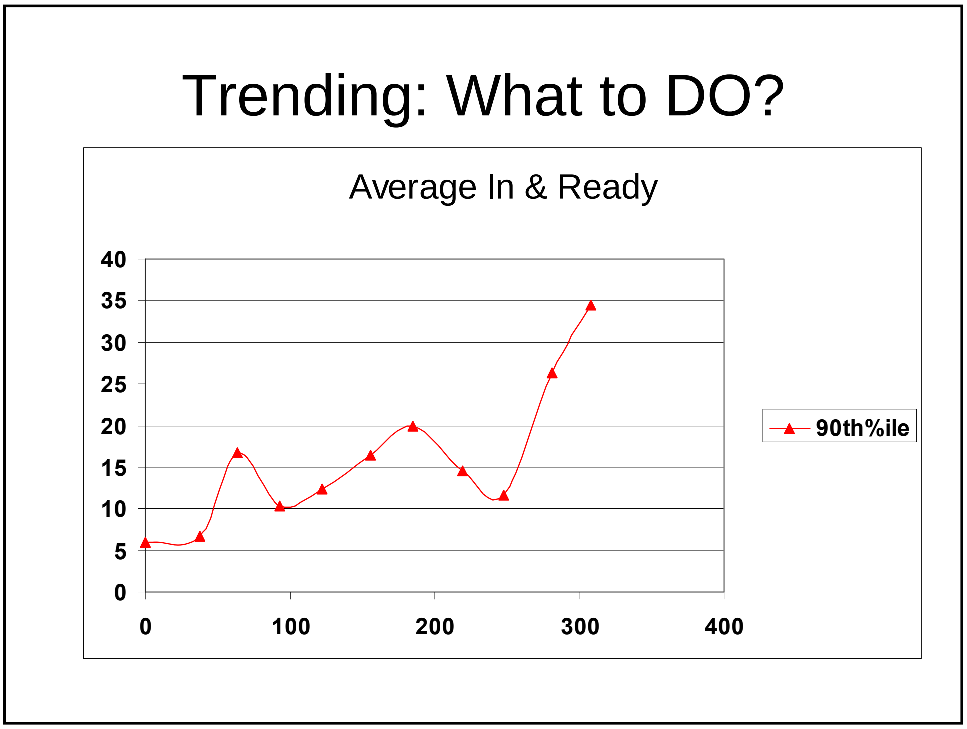
y = -6E-08x3 + 0.0063x2 - 241.55x + 3E+06

R2 = 0.7817 (CS: fit looks good)

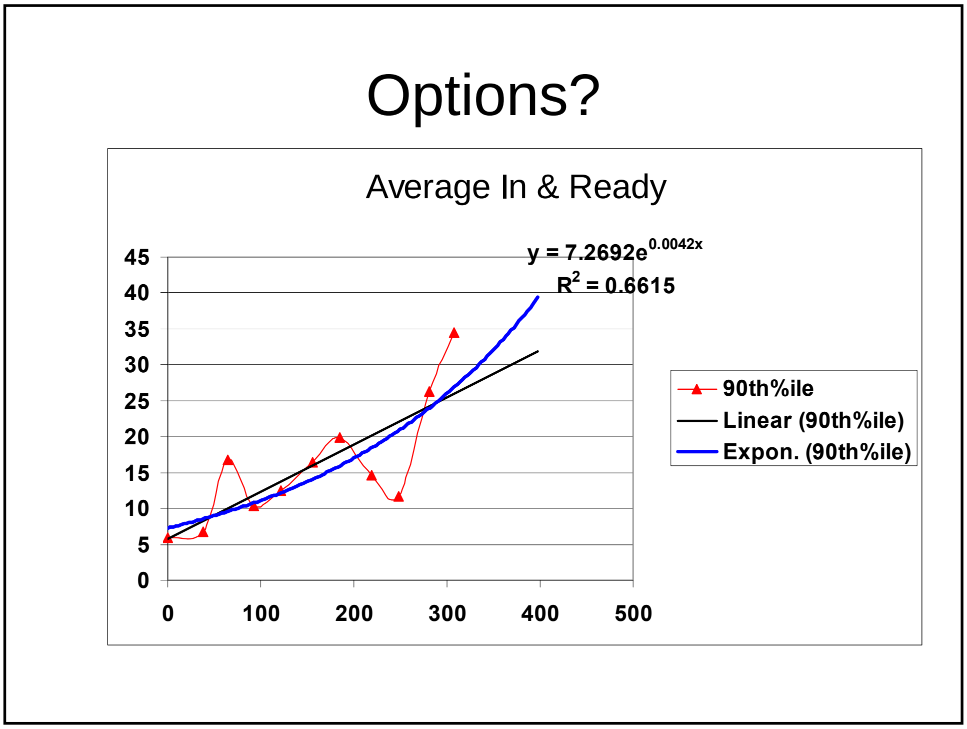


???

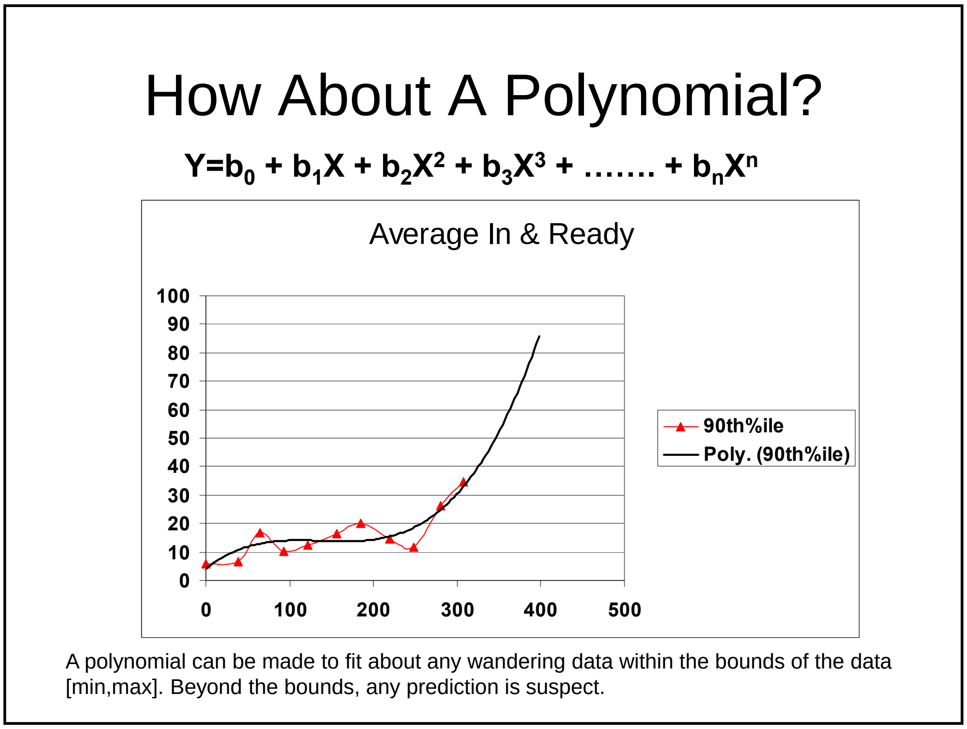
In 144 Days, the $ will be worthless.



Trending: What to DO?



Options?



How About A Polynomial?

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

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.

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:

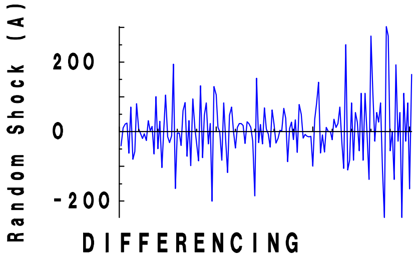
1. Continuous, where we have an observation at every instant of time e.g. lie detectors, electrocardiograms. We denote this using observation X at time t, X(t).

2. Discrete, where we have an observation at (usually regularly) spaced intervals. We denote this as Xt.

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

Time Series Models (Briefly)

(Box-Jenkins Analysis)



Correlation at Lag K

(Z_t , Z_{t-k})

DIFFERENCING

$Z_t = Y_t - Y_{t-1}$

AUTOREGRESSIVE (AR) MODELS

$Z_t = A_t - G_1 Z_{t-1} - G_2 Z_{t-2} - \ ...$

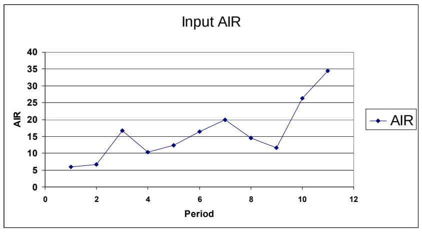
MOVING AVERAGE (MA) MODELS

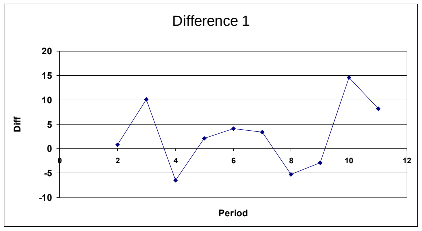
$Z_t = A_t - F_1 A_{t-1} - F_2 A_{t-2} - \ ...$

OR ARMA OR ARIMA

Poor Mans Time Series

INDEX	AIR	Diff 1
1	**5.9**	
2	**6.7**	0.8
3	**16.8**	10.1
4	**10.3**	-6.5
5	**12.4**	2.1
6	**16.5**	4.1
7	**19.9**	3.4
8	**14.6**	-5.3
9	**11.7**	-2.9
10	**26.3**	14.6
11	**34.5**	8.2





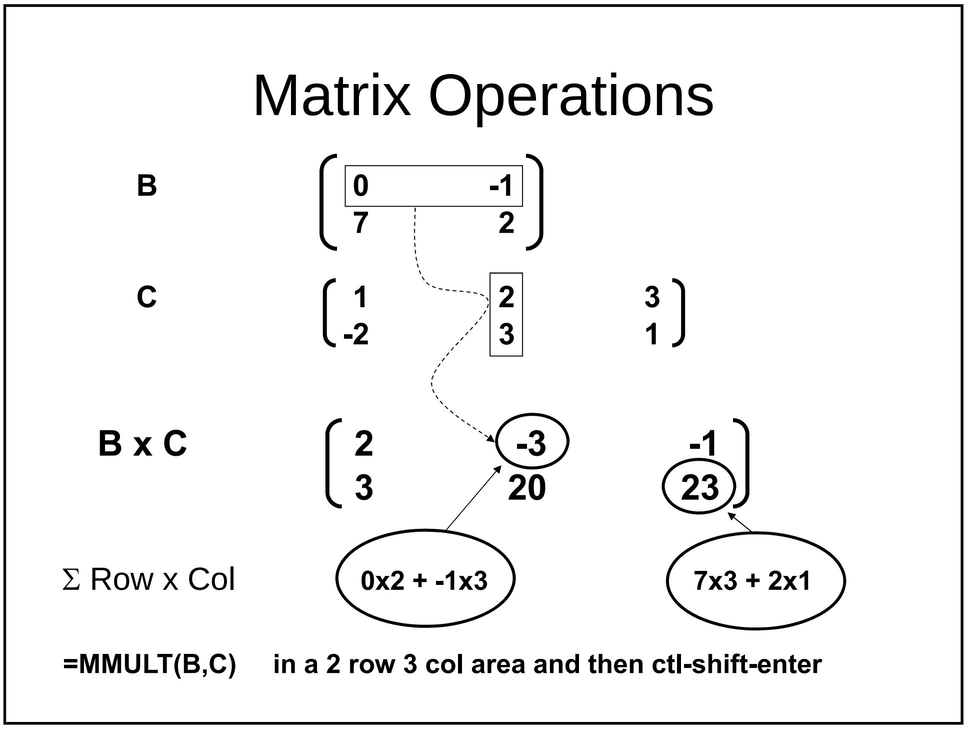
Ref: TSERDAT.xls

Matrix Operations

$$\mathbf{A} \quad \begin{pmatrix} \mathbf{1} & \mathbf{2} \\ \mathbf{-3} & \mathbf{2.5} \end{pmatrix}$$

$$\mathbf{B} \quad \begin{pmatrix} \mathbf{0} & \mathbf{-1} \\ \mathbf{7} & \mathbf{2} \end{pmatrix}$$

$$\mathbf{A+B} \quad \begin{pmatrix} \mathbf{1} & \mathbf{1} \\ \mathbf{4} & \mathbf{4.5} \end{pmatrix}$$



Matrix Operations

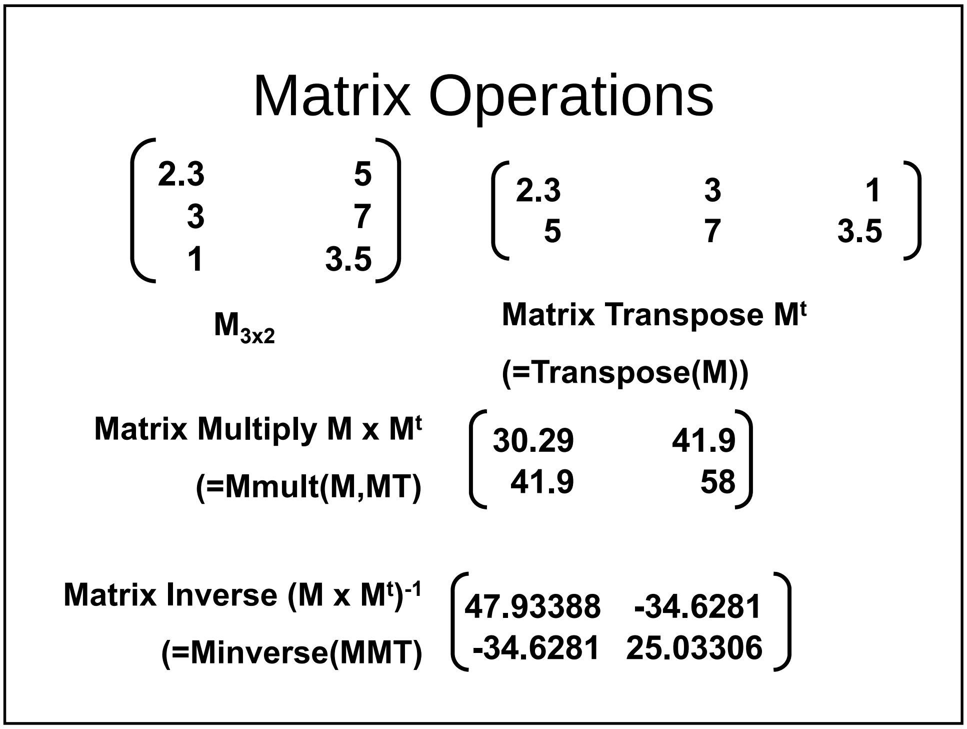
B $\begin{pmatrix} 0 & -1 \\ 7 & 2 \end{pmatrix}$

C $\begin{pmatrix} 1 & 2 & 3 \\ -2 & 3 & 1 \end{pmatrix}$

B x C $\begin{pmatrix} 2 & -3 & -1 \\ 3 & 20 & 23 \end{pmatrix}$

Σ Row x Col **0x2 + -1x3** **7x3 + 2x1**

=MMULT(B,C) in a 2 row 3 col area and then ctl-shift-enter



Matrix Operations

$\begin{pmatrix} 2.3 & 5 \\ 3 & 7 \\ 1 & 3.5 \end{pmatrix}$ $\begin{pmatrix} 2.3 & 3 & 1 \\ 5 & 7 & 3.5 \end{pmatrix}$

M_{3x2} **Matrix Transpose M^t**

(=Transpose(M))

Matrix Multiply M x M^t

(=Mmult(M,MT) $\begin{pmatrix} 30.29 & 41.9 \\ 41.9 & 58 \end{pmatrix}$

Matrix Inverse $(M \times M^t)^{-1}$

(=Minverse(MMT) $\begin{pmatrix} 47.93388 & -34.6281 \\ -34.6281 & 25.03306 \end{pmatrix}$

Matrix Operations

$$\begin{pmatrix} 47.93388 & -34.6281 \\ -34.6281 & 25.03306 \end{pmatrix} \times \begin{pmatrix} 30.29 & 41.9 \\ 41.9 & 58 \end{pmatrix}$$

$$\begin{array}{cc} 1 & 4.54747E-13 \\ 0 & 1 \end{array}$$

The Ugly Part

INDEX	AIR	Diff 1
1	5.9	
2	6.7	0.8
3	16.8	10.1
4	10.3	-6.5
5	12.4	2.1
6	16.5	4.1
7	19.9	3.4
8	14.6	-5.3
9	11.7	-2.9
10	26.3	14.6
11	34.5	8.2

Y = b0 + b1X1 + b2X2 +b3X3
Or
X4 = b0 + b1X1 + b2X2 +b3X3

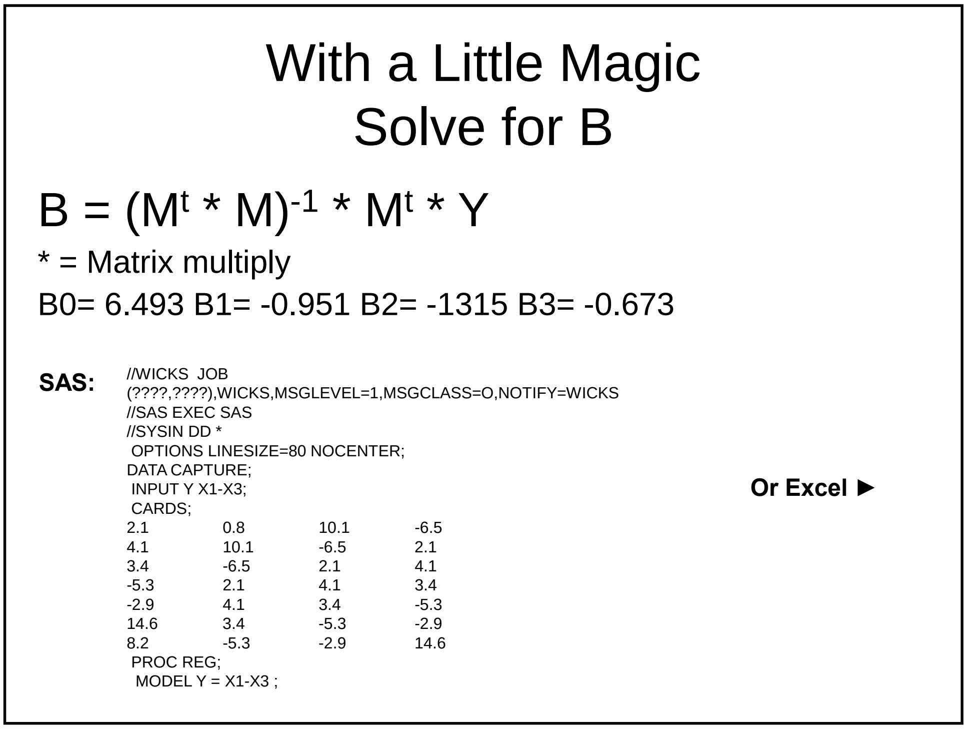
Y

Y
2.1
4.1
3.4
-5.3
-2.9
14.6
8.2

M

X0	X1	X2	X3
1	0.8	10.1	-6.5
1	10.1	-6.5	2.1
1	-6.5	2.1	4.1
1	2.1	4.1	3.4
1	4.1	3.4	-5.3
1	3.4	-5.3	-2.9
1	-5.3	-2.9	14.6

From the input variable AIR, form the pair wise difference sequence Diff 1 = $x_n - x_{n-1.}$ Then build the matrix M for order 3 solution.



With a Little Magic Solve for B

$B = (M^t * M)^{-1} * M^t * Y$

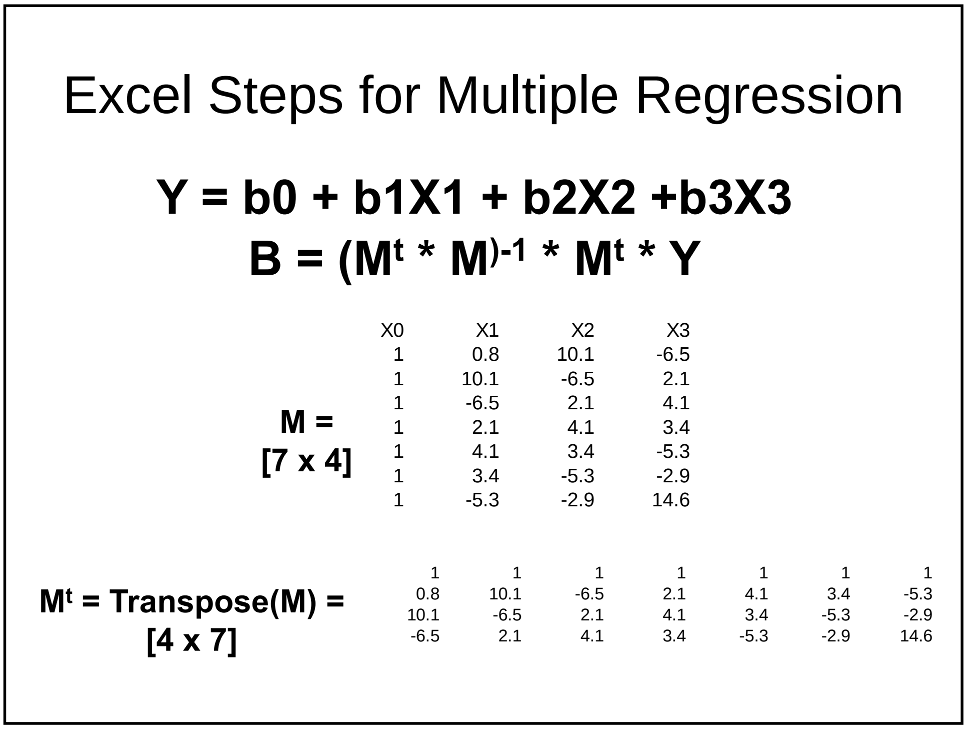
* = Matrix multiply

B0= 6.493 B1= -0.951 B2= -1315 B3= -0.673

SAS:

```
//WICKS  JOB
(????,????),WICKS,MSGLEVEL=1,MSGCLASS=O,NOTIFY=WICKS
//SAS EXEC SAS
//SYSIN DD *
 OPTIONS LINESIZE=80 NOCENTER;
DATA CAPTURE;
 INPUT Y X1-X3;
 CARDS;
2.1        0.8        10.1       -6.5
4.1        10.1       -6.5       2.1
3.4        -6.5       2.1        4.1
-5.3       2.1        4.1        3.4
-2.9       4.1        3.4        -5.3
14.6       3.4        -5.3       -2.9
8.2        -5.3       -2.9       14.6
 PROC REG;
  MODEL Y = X1-X3 ;
```

Or Excel ►



Excel Steps for Multiple Regression

$$Y = b0 + b1X1 + b2X2 + b3X3$$

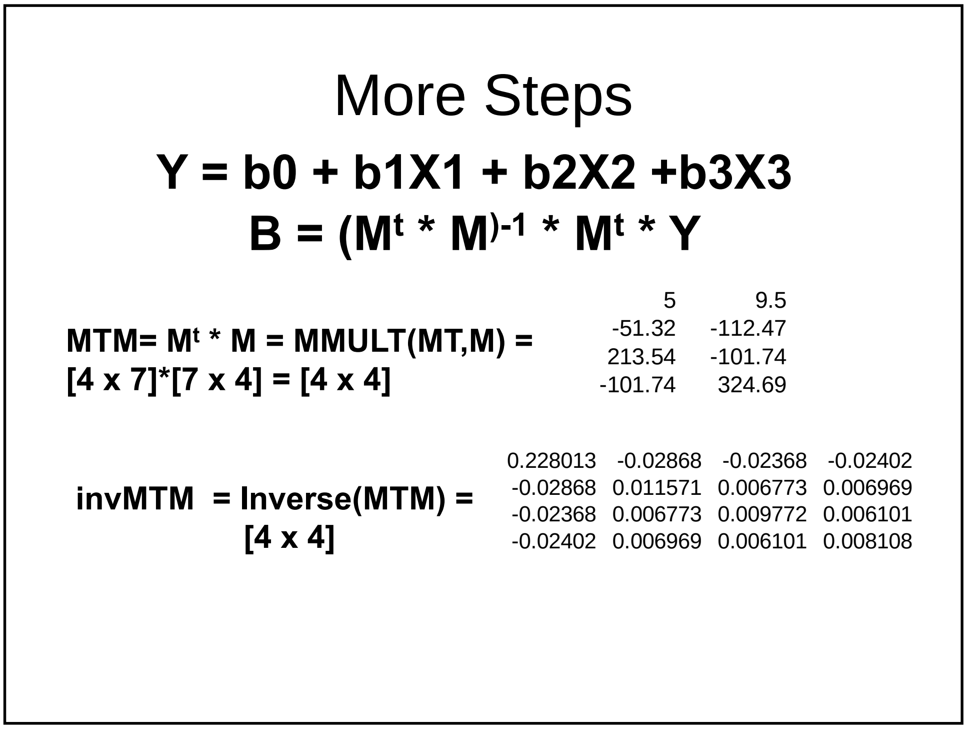
$$B = (M^t * M^{)-1} * M^t * Y$$

M = [7 x 4]

$$\begin{array}{cccc} X0 & X1 & X2 & X3 \\ 1 & 0.8 & 10.1 & -6.5 \\ 1 & 10.1 & -6.5 & 2.1 \\ 1 & -6.5 & 2.1 & 4.1 \\ 1 & 2.1 & 4.1 & 3.4 \\ 1 & 4.1 & 3.4 & -5.3 \\ 1 & 3.4 & -5.3 & -2.9 \\ 1 & -5.3 & -2.9 & 14.6 \end{array}$$

M^t = Transpose(M) = [4 x 7]

$$\begin{array}{ccccccc} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0.8 & 10.1 & -6.5 & 2.1 & 4.1 & 3.4 & -5.3 \\ 10.1 & -6.5 & 2.1 & 4.1 & 3.4 & -5.3 & -2.9 \\ -6.5 & 2.1 & 4.1 & 3.4 & -5.3 & -2.9 & 14.6 \end{array}$$



More Steps

$Y = b0 + b1X1 + b2X2 +b3X3$

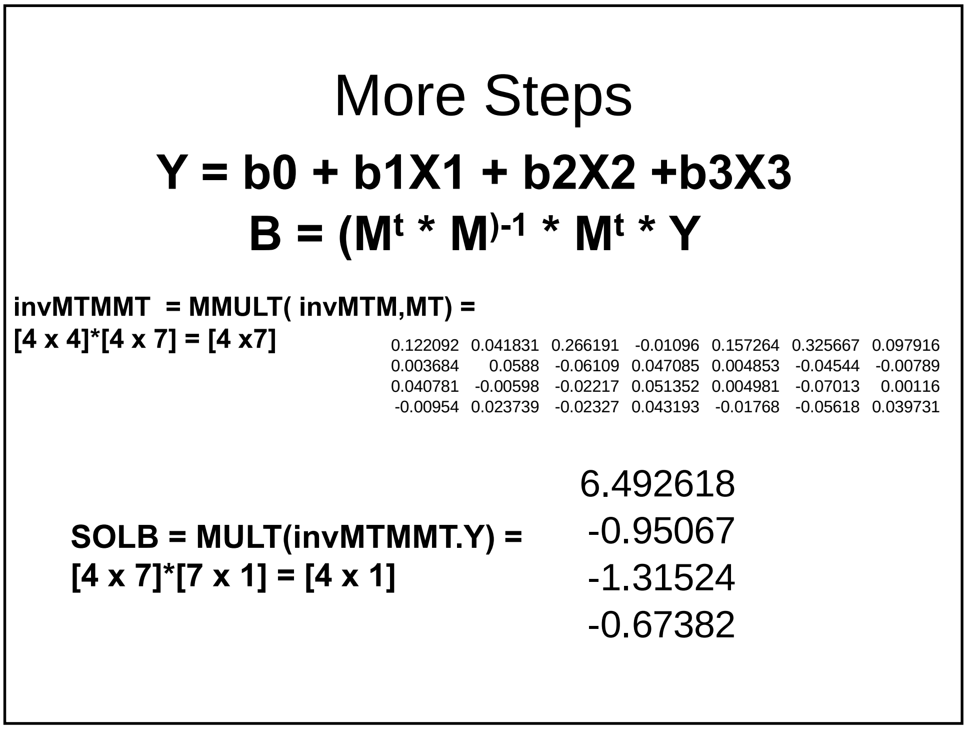
$B = (M^t * M^{)-1} * M^t * Y$

MTM= M^t * M = MMULT(MT,M) =
[4 x 7]*[7 x 4] = [4 x 4]

5	9.5
-51.32	-112.47
213.54	-101.74
-101.74	324.69

invMTM = Inverse(MTM) =
[4 x 4]

0.228013	-0.02868	-0.02368	-0.02402
-0.02868	0.011571	0.006773	0.006969
-0.02368	0.006773	0.009772	0.006101
-0.02402	0.006969	0.006101	0.008108



More Steps

$Y = b0 + b1X1 + b2X2 +b3X3$

$B = (M^t * M^{)-1} * M^t * Y$

invMTMMT = MMULT(invMTM,MT) =
[4 x 4]*[4 x 7] = [4 x7]

0.122092	0.041831	0.266191	-0.01096	0.157264	0.325667	0.097916
0.003684	0.0588	-0.06109	0.047085	0.004853	-0.04544	-0.00789
0.040781	-0.00598	-0.02217	0.051352	0.004981	-0.07013	0.00116
-0.00954	0.023739	-0.02327	0.043193	-0.01768	-0.05618	0.039731

SOLB = MULT(invMTMMT.Y) =
[4 x 7]*[7 x 1] = [4 x 1]

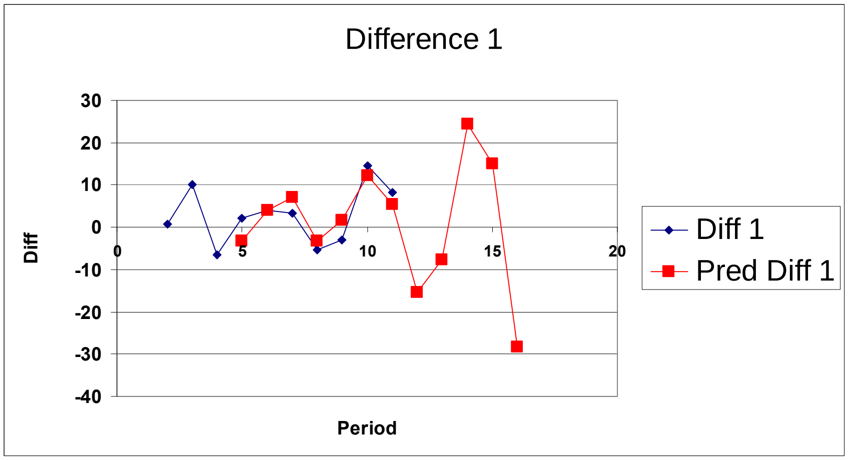
6.492618
-0.95067
-1.31524
-0.67382

The Prediction

$$X_n = 6.493 - 0.951X_{n-1} -1315X_{n-2} -0.673X_{n-3}$$

INDEX	AIR	Diff 1		Pred Diff 1		Pred AIR
1	5.9					
2	6.7	0.8				
3	16.8	10.1				
4	10.3	-6.5				
5	12.4	2.1	+	-3.17		7.13
6	16.5	4.1		4.02	=	16.42
7	19.9	3.4		7.15		23.65
8	14.6	-5.3		-3.19		16.71
9	11.7	-2.9		1.69		16.29
10	26.3	14.6		12.19		23.89
11	34.5	8.2		5.51		31.81
12				-15.48		19.02
13				-7.74		11.28
14				24.27		35.55
15				15.04		50.59
16				-28.20		22.39

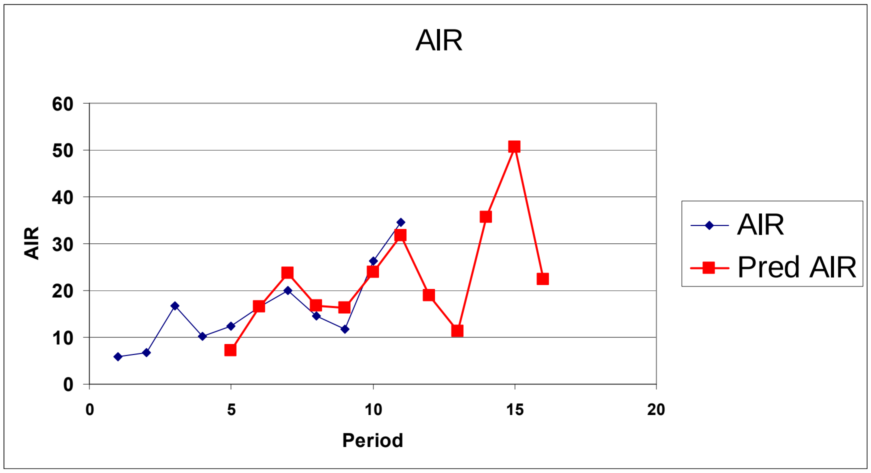
Diff 1 Plot



Prediction for AIR

$X_n = 6.493 - 0.951X_{n-1} -1315X_{n-2} -0.673X_{n-3}$

$R^2 = 0.89$



Bibliography

❑ 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.

❑ *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.

❑ *Applied Regression Analysis*, by Draper & Smith. This is the classic in regression analysis. It can get a little deep. However, if you like a full treatment with derivations of the formulae, this is it.

❑ The Signal and the Noise, by Nate Silver. An interesting book on real life prediction.