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.
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.

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

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 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.
How Accurate Is It?

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

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

\[
\left[ \mu - 1.96 \frac{\sigma}{\sqrt{n}}, \mu + 1.96 \frac{\sigma}{\sqrt{n}} \right]
\]

\[
\left[ \mu - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \mu + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]
\]

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

In Excel: \(=\text{Confidence}(\mu, \sigma, n)\)

\(=\text{Confidence}(0.5, 1, 100) = 1.96\)

Summary

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

§= Percentile formulae assume a sorted list: Low to high.
The Intent of regression analysis

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

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

A concern when X is Time. Given \{\langle t_i, y_i \rangle \}, 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)

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.

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

\[
\text{Correlation} = \frac{\text{COV}(X,Y)}{\sigma_x \sigma_y} = \frac{\sigma_{xy}^2}{\sigma_x \sigma_y} = \frac{E[(x-\mu_x)(y-\mu_y)]}{\sigma_x \sigma_y}
\]

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

The linear equation for the data is:

\[ y = 59.877x + 733.8 \]

with a correlation coefficient of:

\[ R^2 = 0.7425 \]
Predictive Analysis

- Given \( \{X_i, Y_i\} \ i=1,n \)
- Find \( Y_i=F(X_i) \) such that the sum of errors squared is minimized \( \text{sum}(Y_i - \hat{Y_i})^2 \)
- Evaluate \( F(X_i) \) from \( i = n+1 \) to \( n+j \) (j future periods/values)

\[
y = 0.0588x + 8.7307 \\
R^2 = 0.7978
\]

-\( Y = 0.0588 \times 1000 + 8.7307 \)
-\( Y = 67.5\%

Linear Regression

\[
y = 3.0504x + 385.42 \\
R^2 = 0.7881
\]
Linear regression’s predictions assume that the future looks like the past.

Linear Fit for \( \{X_i, Y_i\} \)

\[
y = B_0 + B_1 X_i
\]

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

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

\[ B = (M^t \cdot M)^{-1} \cdot M^t \cdot Y \]

Solve for Y = B0 + B1X

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>YH</th>
<th>Sq(YH-YA)</th>
<th>Sq(Y-YA)</th>
<th>R2</th>
</tr>
</thead>
<tbody>
<tr>
<td>62.3</td>
<td>1.3</td>
<td></td>
<td>1.316809</td>
<td>0.0196</td>
<td>0.819061 = {SUM(F3:F7)/SUM(G3:G7)}</td>
</tr>
<tr>
<td>64.3</td>
<td>1.4</td>
<td></td>
<td>1.354367</td>
<td>0.007333</td>
<td>0.0016</td>
</tr>
<tr>
<td>70.8</td>
<td>1.4</td>
<td></td>
<td>1.476432</td>
<td>0.001323</td>
<td>0.0016</td>
</tr>
<tr>
<td>71.1</td>
<td>1.5</td>
<td></td>
<td>1.482065</td>
<td>0.001769</td>
<td>0.0036</td>
</tr>
<tr>
<td>75.8</td>
<td>1.6</td>
<td></td>
<td>1.570328</td>
<td>0.016885</td>
<td>0.0296</td>
</tr>
</tbody>
</table>

Avg 1.44

MT is 2x5

<table>
<thead>
<tr>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>ctl-shift-enter</th>
</tr>
</thead>
<tbody>
<tr>
<td>62.3</td>
<td>64.3</td>
<td>70.8</td>
<td>71.1</td>
<td>75.8</td>
<td></td>
</tr>
</tbody>
</table>

MT*M is 2x2

| 5 | 344.3 |
| 344.3 | 23829.27 |

INV(MTM) is 2x2

| 38.46158 | -0.57017 |
| -0.57017 | 0.00828 |

IMTM*MT is 2x5

| 3.940284 | 2.799954 | -0.90612 | -1.07717 | -3.75847 |
| -0.05432 | -0.03776 | 0.016063 | 0.018547 | 0.0574637 |

IMTMM*Y is 2x1

| 0.146865 | 0.018779 |
| 0.018779 | B1 |
**Excel Solution**

\[
y = 0.0188x + 0.1469 \\
R^2 = 0.8191
\]

**Impact of Outlier**

\[
y = -0.0058x + 1.9284 \\
R^2 = 0.0174
\]
A perfect fit is always possible

\[ y = 58111x^4 - 338194x^3 + 736689x^2 - 711801x + 257442 \]

\[ R^2 = 1 \]

Albeit meaningless in this case.

Goodness of Fit.

Residual = Yi – Ypredict

The plot of residuals should show points randomly distributed around 0.
## EXCEL Solution

**Equation:**
\[ y = 47.3x + 0.275 \]

**\( R^2 \):** 0.9262

### Data Table

<table>
<thead>
<tr>
<th>Units of Work (X)</th>
<th>CPU% (Y)</th>
<th>( YH = 47.3x + 0.275 )</th>
<th>Residual (( Yi - Yhi ))</th>
<th>Resid*2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.3</td>
<td>62.3</td>
<td>61.765</td>
<td>0.535</td>
<td>0.286225</td>
</tr>
<tr>
<td>1.4</td>
<td>64.3</td>
<td>66.495</td>
<td>-2.195</td>
<td>4.618025</td>
</tr>
<tr>
<td>1.45</td>
<td>70.8</td>
<td>68.86</td>
<td>1.94</td>
<td>3.7636</td>
</tr>
<tr>
<td>1.5</td>
<td>71.1</td>
<td>71.225</td>
<td>-0.125</td>
<td>0.015625</td>
</tr>
<tr>
<td>1.6</td>
<td>75.8</td>
<td>75.955</td>
<td>-0.155</td>
<td>0.024025</td>
</tr>
</tbody>
</table>

**SSE:** 8.9075

**Syx:** 1.723127002

**Avg X:** 1.45

**\( T \):** 3.182446305

**N:** 5

---

## Solution with Bounds

**Equation:**
\[ y = 47.3x + 0.275 \]

**\( R^2 \):** 0.9262

---

### Diagram

- **CPU%**
- **LB**
- **UB**
- **Linear (CPU%)**
- **Linear (CPU%)**
### Computations

<table>
<thead>
<tr>
<th>Units of Work (X)</th>
<th>CPU% (Y)</th>
<th>$Y = 47.489x + 0.275 \text{ Residual}=Y_i-Y_{hi}$</th>
<th>Resid$2 = (X_i-Avgx)^2$</th>
<th>Comp</th>
<th>LB</th>
<th>UB</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.3</td>
<td>62.3</td>
<td>61.765 0.535 0.286225 0.0225 0.65 57.34385 66.16815</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.4</td>
<td>64.3</td>
<td>66.495 -2.195 4.816025 0.0025 0.25 63.75312 69.23688</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.45</td>
<td>70.8</td>
<td>68.86 1.94 3.7636 0 0.2 66.40759 71.31241</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td>71.1</td>
<td>71.225 -0.125 0.015625 0.0025 0.25 68.48312 73.96688</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.6</td>
<td>75.8</td>
<td>75.955 -0.155 0.024025 0.0225 0.65 71.53385 80.37615</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.65</td>
<td>78.32</td>
<td>1 1.45 74.08168 87.2832</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.7</td>
<td>80.68</td>
<td>1.45 75.29479 90.80521</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.75</td>
<td>83.05</td>
<td>2 75.29479 90.80521</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.8</td>
<td>80.415</td>
<td>2.65 76.46899 94.34191</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

$\text{LB: } = \$C2-T*\text{Sxy}^\ast\text{SQRT(}\$G2\$)$

### Projection with Bounds

![Projection with Bounds](image)

$y = 47.3x + 0.275$

$R^2 = 0.99922$
Projection with Bounds

\[ y = 47.3x + 0.275 \]
\[ R^2 = 0.9262 \]

Units of Work

CPU%

Y=mX for CPU% vs Blocks/Sec

(\(b=0\) in Y=mX + b)

\[ y = 0.0235x \]
\[ R^2 = -0.7761 \]
Filtered Data CPU%>10

\[ y = 0.0742x \]
\[ R^2 = 0.9028 \]

Residuals

For each Xi, plot \( e = Y - \hat{Y}_i \)
Can’t Get Any Worse Solution?

\[ y = 0.0612x \]
\[ R^2 = -0.3959 \]

PS to CS Dissonance

y = -0.0002x + 8.2996
\[ R^2 = 0.4388 \] (CS: Not a good line)
**PS to CS Dissonance**

*(PS: Variability is scale dependent)*

\[ y = -0.0002x + 8.2996 \]

\[ R^2 = 0.4388 \]  (CS: Variability is scale independent)

---

**PS to CS Dissonance**

*(PS: Polynomial fit looks good)*

\[ y = -6E-08x^3 + 0.0063x^2 - 241.55x + 3E+06 \]

\[ R^2 = 0.7817 \]  (CS: fit looks good)
In 144 Days, the $ will be worthless.

Trending: What to DO?

Average In & Ready

90th%ile
Options?

Average In & Ready

\[ y = 7.2692e^{0.0042x} \]
\[ R^2 = 0.6615 \]

How About A Polynomial?

\[ Y=b_0 + b_1X + b_2X^2 + b_3X^3 + \ldots + b_nX^n \]

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)

**Differencing**

\[ Z_t = Y_t - Y_{t-1} \]

**Autoregressive (AR) Models**

\[ Z_t = \alpha_t - G_1 Z_{t-1} - G_2 Z_{t-2} - \ldots \]

**Moving Average (MA) Models**

\[ Z_t = \alpha_t - F_1 \alpha_{t-1} - F_2 \alpha_{t-2} - \ldots \]

or ARMA or ARIMA

\[ \text{Correlation at Lag } K \]

\[ (Z_t, Z_{t-k}) \]
**Poor Mans Time Series**

<table>
<thead>
<tr>
<th>INDEX</th>
<th>AIR</th>
<th>Diff 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.9</td>
<td>0.8</td>
</tr>
<tr>
<td>2</td>
<td>6.7</td>
<td>10.1</td>
</tr>
<tr>
<td>3</td>
<td>16.8</td>
<td>-6.5</td>
</tr>
<tr>
<td>4</td>
<td>10.3</td>
<td>2.1</td>
</tr>
<tr>
<td>5</td>
<td>12.4</td>
<td>1.2</td>
</tr>
<tr>
<td>6</td>
<td>16.5</td>
<td>4.1</td>
</tr>
<tr>
<td>7</td>
<td>19.9</td>
<td>3.4</td>
</tr>
<tr>
<td>8</td>
<td>14.6</td>
<td>-5.3</td>
</tr>
<tr>
<td>9</td>
<td>11.7</td>
<td>-2.9</td>
</tr>
<tr>
<td>10</td>
<td>26.3</td>
<td>14.6</td>
</tr>
<tr>
<td>11</td>
<td>34.5</td>
<td>8.2</td>
</tr>
</tbody>
</table>

Ref: TSERDAT.xls

---

**Matrix Operations**

\[ A = \begin{pmatrix} 1 & 2 \\ -3 & 2.5 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & -1 \\ 7 & 2 \end{pmatrix} \]

\[ A + B = \begin{pmatrix} 1 & 1 \\ 4 & 4.5 \end{pmatrix} \]
### Matrix Operations

#### Matrix Multiplication

**B**

\[
\begin{pmatrix}
0 & -1 \\
7 & 2
\end{pmatrix}
\]

**C**

\[
\begin{pmatrix}
1 & 2 & 3 \\
-2 & 3 & 1
\end{pmatrix}
\]

**B x C**

\[
\begin{pmatrix}
2 & -3 & -1 \\
3 & 20 & 23
\end{pmatrix}
\]

\[
\sum \text{Row x Col} = 0\times2 + (-1\times3) = 7\times3 + 2\times1
\]

\[
= \text{MMULT}(B, C) \text{ in a 2 row 3 col area and then ctl-shift-enter}
\]

#### Matrix Transpose

\[
\text{Matrix Transpose } M^t = \text{Transpose}(M)
\]

**Matrix Multiply**

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

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

\[
\text{Matrix Multiply } M \times M^t = \text{Mmult}(M, M^t)
\]

\[
\begin{pmatrix}
30.29 & 41.9 \\
41.9 & 58
\end{pmatrix}
\]

**Matrix Inverse**

\[
\text{Matrix Inverse } (M \times M^t)^{-1} = \text{Minverse}(M \times M^t)
\]

\[
\begin{pmatrix}
47.93388 & -34.6281 \\
-34.6281 & 25.03306
\end{pmatrix}
\]
### Matrix Operations

\[
\begin{pmatrix}
47.93388 & -34.6281 \\
-34.6281 & 25.03306
\end{pmatrix} \times \begin{pmatrix}
30.29 \\
41.9
\end{pmatrix}
\]

\[
\begin{pmatrix}
1 \\
0
\end{pmatrix} + 4.54747 \times 10^{-13}
\]

\[
\begin{pmatrix}
41.9 \\
58
\end{pmatrix}
\]

### The Ugly Part

\[
Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3
\]

Or

\[
X_4 = b_0 + b_1X_1 + b_2X_2 + b_3X_3
\]

From the input variable AIR, form the pair wise difference sequence

\[
\text{Diff } 1 = x_n - x_{n-1}
\]

Then build the matrix M for order 3 solution.
With a Little Magic
Solve for B

\[ B = (M^t \ast M)^{-1} \ast M^t \ast Y \]

\(* = \text{Matrix multiply}\)

\[ B_0 = 6.493 \]
\[ B_1 = -0.951 \]
\[ B_2 = -1315 \]
\[ B_3 = -0.673 \]

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

\[ Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 \]

\[ B = (M^t \ast M)^{-1} \ast M^t \ast Y \]

\[ X_0 \quad X_1 \quad X_2 \quad X_3 \]
\[ 1 \quad 0.8 \quad 10.1 \quad -6.5 \]
\[ 1 \quad 10.1 \quad -6.5 \quad 2.1 \]
\[ 1 \quad -6.5 \quad 2.1 \quad 4.1 \]
\[ 1 \quad 2.1 \quad 4.1 \quad 3.4 \]
\[ 1 \quad 4.1 \quad 3.4 \quad -5.3 \]
\[ 1 \quad 3.4 \quad -5.3 \quad -2.9 \]
\[ 1 \quad -5.3 \quad -2.9 \quad 14.6 \]

\[ M = \begin{bmatrix} 7 \times 4 \end{bmatrix} \]

\[ M^t = \text{Transpose}(M) = \begin{bmatrix} 4 \times 7 \end{bmatrix} \]
More Steps

\[ Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 \]

\[ B = (M^t \ast M)^{-1} \ast M^t \ast Y \]

MTM = \( M^t \ast M = \text{MMULT(MT,M)} = [4 \times 7]^*[7 \times 4] = [4 \times 4] \)

invMTM = \( \text{Inverse(MTM)} = [4 \times 4] \)

\[ B = (Mt \ast M)^{-1} \ast Mt \ast Y \]

invMTMMT = \( \text{MMULT( invMTM,MT)} = [4 \times 4]^*[4 \times 7] = [4 \times 7] \)

SOLB = \( \text{MULT(invMTMMT,Y)} = [4 \times 7]^*[7 \times 1] = [4 \times 1] \)
The Prediction

\[ X_n = 6.493 - 0.951X_{n-1} - 1315X_{n-2} - 0.673X_{n-3} \]

<table>
<thead>
<tr>
<th>INDEX</th>
<th>AIR</th>
<th>Diff 1</th>
<th>Pred Diff 1</th>
<th>Pred AIR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.9</td>
<td>0.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>6.7</td>
<td>0.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>16.8</td>
<td>10.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>10.3</td>
<td>-6.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>12.4</td>
<td>2.1</td>
<td>-3.17</td>
<td>7.13</td>
</tr>
<tr>
<td>6</td>
<td>16.5</td>
<td>4.1</td>
<td>4.02</td>
<td>16.42</td>
</tr>
<tr>
<td>7</td>
<td>19.9</td>
<td>3.4</td>
<td>7.15</td>
<td>23.65</td>
</tr>
<tr>
<td>8</td>
<td>14.6</td>
<td>-5.3</td>
<td>-3.19</td>
<td>16.71</td>
</tr>
<tr>
<td>9</td>
<td>11.7</td>
<td>-2.9</td>
<td>1.69</td>
<td>16.29</td>
</tr>
<tr>
<td>10</td>
<td>26.3</td>
<td>14.6</td>
<td>12.19</td>
<td>23.89</td>
</tr>
<tr>
<td>11</td>
<td>34.5</td>
<td>8.2</td>
<td>5.51</td>
<td>31.81</td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td>-15.48</td>
<td>19.02</td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td>-7.74</td>
<td>11.28</td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td>24.27</td>
<td>35.55</td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td>15.04</td>
<td>50.59</td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td>-26.20</td>
<td>22.39</td>
</tr>
</tbody>
</table>
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.
- 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.