Correlation in Temperatures on the First and Last Days of March

Daniel Kerner
Civilized Software Inc.
12109 Heritage Park Circle
Silver Spring MD 20906
phone:301-962-3711
email:csi@civilized.com
URL:www.civilized.com

August 4, 2010

An old adage regarding weather in the month of March states:

```
March comes in like a lion and out like a lamb, or
in like a lamb and out like a lion.
```

If we associate lion-like weather with lower temperatures and lamb-like weather with higher temperatures, then the adage can be understood to be a statement regarding the relationship between the temperature on the first and last days of March: the average temperature on the first day of March and the average temperature on the last day of March are *anti*-correlated; meaning when the average temperature on the first day of the month is high, the average temperature on the last day of the month is low, or *visa-versa*.

In this paper, we demonstrate the use of the MLAB mathematical modeling computer program available at `http://www.civilized.com` to compute various measures of correlation between the average temperatures on the first and last day of March, and test the veracity of the adage.

The average temperatures in Fahrenheit degrees on the first day and last day of March for New York City, New York, over the last ten years was obtained from:

```
http://www.wunderground.com/history/airport/KNYC
```

After starting the MLAB program, these numbers are stored in a three column matrix `M` with the following MLAB command:

```
M = SHAPE(10,3,LIST(2001, 30, 42,\
                    2002, 37, 54,\
                    2003, 36, 36,\
                    2004, 54, 43,\
                    2005, 36, 48,\
                    2006, 34, 60,\
                    2007, 40, 52,\
                    2008, 40, 47,\
                    2009, 32, 49,\
                    2010, 42, 51))
```

This example of the `SHAPE` operator assigns 30 numbers to the matrix with ten rows and three columns. The first column of the matrix `M` contains year values; the second column contains the average temperature on the first day of March for the given year; and the third column contains the average temperature on the last day of March for the given year.

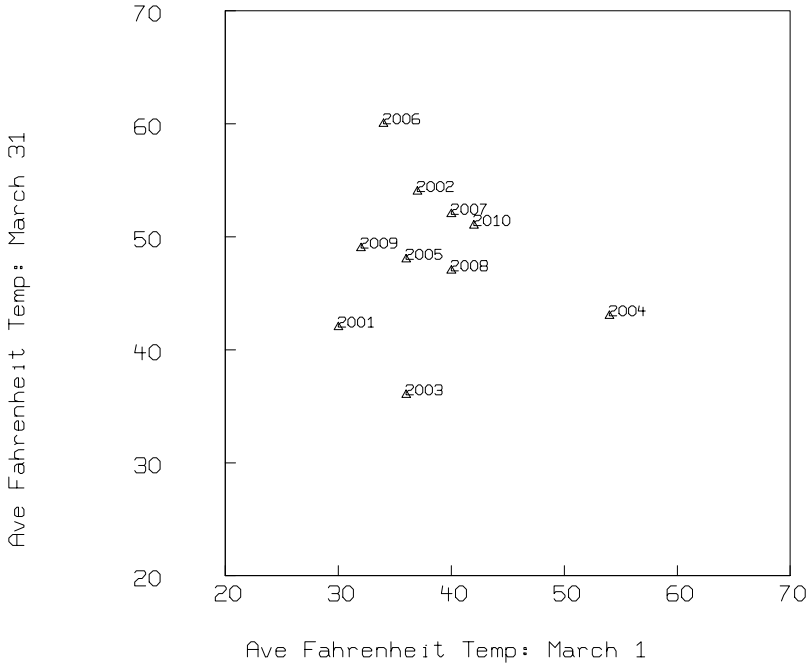
MLAB provides several methods for quantifying correlation between two data sets. These methods begin with the null hypothesis that no association exists between the two data sets. Here we show four methods: linear regression, Pearson product-moment coefficient, Spearman rank correlation coefficient, and Kendall's τ coefficient.

We begin by generating a scatter plot of the March temperature data with the following MLAB commands:

```
DRAW M COL (2,3) LINETYPE NONE POINTTYPE TRIANGLE PTSIZE .01 FFRACT \
   LABEL 2001:2010 LABELSIZE .01 FFRACT
VIEW
```

The `DRAW` command generates a graph in the plane with coordinates of points given by the second and third columns of matix `M`. `LINETYPE NONE` specifies that no lines will be drawn to connect the points. The phrase `POINTTYPE TRIANGLE` causes each point to be drawn as a triangle. By virtue of the clause `PTSIZE .01 FFRACT`, the base of each triangle is drawn with a size that is 0.01 frame fraction units–which means 0.01 times the width of the full graphics window. The expression `LABEL 2001:2010` indicates the numbers 2001, 2002, 2003, ..., and 2010 are to be drawn as labels on each successive point. The size of each label is also 0.01 times the width of the graphics window due to the clause `LABELSIZE .01 FFRACT`.

Following the `VIEW` command, the following graph is obtained:



If the adage is true, we would expect the scatter plot to show data points lying close to a line with negative slope, i.e. a line extending from the upper left to the lower right of the graph.

The first measure of association we consider is linear regression. We find the best-fitting straight line for the data in the previous graph with the following MLAB commands:

```
FCT F(T) = A+B*T  /* define the linear function */
A = 1; B = 1;  /* supply initial values of slope A and intercept B */
FIT (A,B), F TO M COL (2,3) /* find the best fitting parameter values */
```

Note comments delimited by `/*` and `*/` are ignored by MLAB. The `FCT` statement defines the linear function. In this case, the function takes one argument, `T`, and has two parameters, `A` and `B`. The asterisk symbol, `*`, in the function definition denotes multiplication and the plus sign, `+`, denotes addition.

The second statement assigns the value 1 to the parameters, `A` and `B`. Establishing initial values of the parameters is necessary in order to evaluate the function `F` during the `FIT` operation in the next statement.

The third statement is the `FIT` command. The `FIT` command shown causes MLAB to find the values of the parameters `A` and `B` in the function `F` which minimize the sum-of-squares:

$$S = \sum_{i=1}^{n} (F(M[i,2]) - M[i,3])^2 = \sum_{i=1}^{n} (A + B * M[i,2] - M[i,3])^2$$

with n=10. MLAB returns the following information:

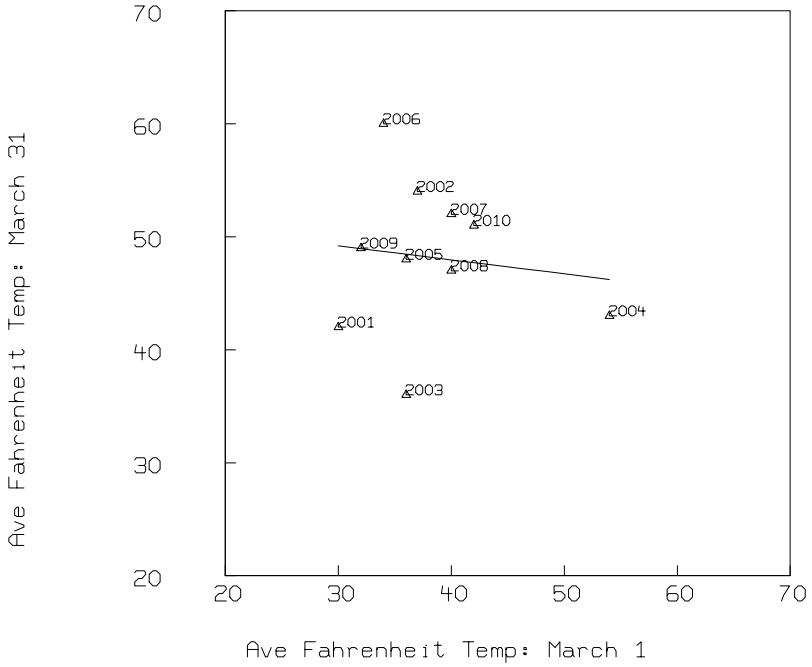
```
final parameter values
      value                error                  dependency    parameter
    52.92368451          13.66498553           0.9728637491   A
  -0.1239812209         0.3537612101           0.9728637491   B
2 iterations
CONVERGED
best weighted sum of squares = 4.053761e+02
weighted root mean square error = 7.118428e+00
weighted deviation fraction = 9.813054e-02
R squared = 1.512113e-02
```

The best-fitting line segment is added to the previous scatter plot with the following command:

```
DRAW POINTS(F,30:60!20)
```

The `POINTS` operator in MLAB evaluates the function specified by the first argument at the values specified by the second argument. The second argument, `30:60!20`, is a vector of twenty equally spaced values starting with `30` and ending with `60`.

Another `VIEW` command results in the following display:



Although there is wide deviation by the temperature data from the best-fit line segment, the slope of the best-fit line is, nonetheless, negative, indicating anti-correlation in the data.

The Pearson product-moment coefficient provides another measure of association in two data sets. The coefficient is computed as:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where all three summations are for i=1:10, and $x_1, x_2, ...x_n$ are samples of the first population–in our case, the average temperatures on the first day of the month; $y_1, y_2, ..., y_n$ are samples from the second population– in our case the average temperatures on the last day of the month; $\bar{x}$ is the average of the x_i values; and $\bar{y}$ is the average of the y_i values. With some algebra, one can show that the Pearson r product-moment coefficient ranges in value from -1 to 1. The value -1 is obtained if there is exact linear *anti*-correlation, the value 1 is obtained for exact linear correlation, and values near 0 indicate no correlation or anti-correlation.

The Pearson product-moment coefficient for the temperature data in matrix `M` can be computed with the MLAB command:

```
CORR(M COL (2,3))
```

MLAB responds:

```
    : a  2 by 2 matrix

  1: 1             -0.12296801
  2: -0.12296801   1
```

So the correlation coefficient is equal to -0.122968.

MLAB can also compute the correlation coeffient in the context of a so-called hypothesis test, using the hypothesis test function, `PEART`.

```
PEART(M COL 2,M COL 3)
```

MLAB responds as follows:

```
[correlation-coefficient test: is the underlying correlation
r, of which the correlation R of the input data x[1:n] and
y[1:n] is a sample, plausibly zero?]
null hypothesis H0: r = 0.
Then R*sqrt((n-2)/(1-R^2)) is approximately distributed as
Student's t with n-2 degrees of freedom.
The sample R-value = -0.122968
The sample t-value = -0.350466

The probability P(t < -0.350466) = 0.367519
This means that a value of t smaller than -0.350466 arises
about 36.751909 percent of the time, given H0.

The probability P[t > -0.350466] = 0.632481
This means that a value of t greater than -0.350466 arises
about 63.248091 percent of the time, given H0.

The probability P[t < -0.350466 or t > 0.350466] = 0.735038
This means that a value of t more extreme than 0.350466  arises
about 73.503818 percent of the time, given H0.

    : a  5 by 1 matrix

  1: -0.12296801
  2: -.350465869
  3: .367519089
  4: .632480911
  5: .735038179
```

The R-value of -0.122968 returned indicates weak anti-correlation.

The Spearman correlation coefficient is computed from relative ranks of the n samples. We replace the values of March 1st temperatures with their relative ranks; 1 is the rank of the highest temperature and n is the rank of the lowest temperature. The values of March 31st temperatures are also replaced by their relative ranks.

The Spearman correlation coefficient is then computed as:

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

where d_i is the difference between the ranks of the i-th pair of ranks. Like the Pearson product-moment correlation coefficient, the Spearman correlation coefficient is -1 for exact, anti-correlation; 1 for exact, correlation; and near 0 for no correlation at all.

This expression is evaluated with MLAB as follows:

```
SPEART(M COL 1,M COL 2)
```

MLAB responds to this command with:

```
[Spearman rank-correlation test: is the correlation R between
the ranks of the paired data d1[] and d2[] plausibly zero?]
null hypothesis H0: R = 0
The sample R-value = 0.036810

The probability P(R < 0.036810) = 0.554188
This means that a value of R smaller than 0.036810 arises
about 55.418816 percent of the time, given H0.

The probability P[R > 0.036810] = 0.445812
This means that a value of R greater than 0.036810 arises
about 44.581184 percent of the time, given H0.

The probability P[R < -0.036810 or R > 0.036810] = 0.891624
This means that a value of R more extreme than 0.036810 arises
about 89.162368 percent of the time, given H0.

    : a  4 by 1 matrix

  1: .036809816
  2: .554188161
  3: .445811839
  4: .891623677
```

The Spearman correlation coefficient for the first and last days' temperatures in March has a value of 0.0368, indicating the data are correlated, not anti-correlated.

Another measure of correlation/anti-correlation is given by the Kendall paired-sample τ coefficient. As with the Spearman correlation coefficient, computing the Kendall paired-sample τ coefficient begins by ranking the observations for each population separately. One then tallies the pairs of ranks that are equal, i.e. concordant, with the pairs of ranks that are unequal, i.e. discordant. The measure of correlation is then computed as:

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{\frac{1}{2}n(n-1)}$$

We can evaluate the Kendall τ coefficient for the first and last day of March temperature data with the MLAB command:

```
KEN1T(M COL 2,M COL 3)
```

MLAB responds:

```
[Kendall's tau correlation-coefficient test: is the tau correlation
of the input data x[] and y[] plausibly zero?]
null hypothesis H0: tau = 0.
The sample kappa = -1
The sample tau-value = -0.022222, variance = 123.000000

The probability P(tau < -0.022222) = 0.464077
This means that a value of tau smaller than -0.022222 arises
about 46.407727 percent of the time, given H0.

The probability P[tau > -0.022222] = 0.535923
This means that a value of tau greater than -0.022222 arises
about 53.592273 percent of the time, given H0.

The probability P[tau < -0.022222 or tau > 0.022222] = 0.928155
This means that a value of tau more extreme than 0.022222 arises
about 92.815454 percent of the time, given H0.

    : a  5 by 1 matrix

  1: -1
  2: -2.22222222E-2
  3: .464077268
  4: .535922732
  5: .928154537
```

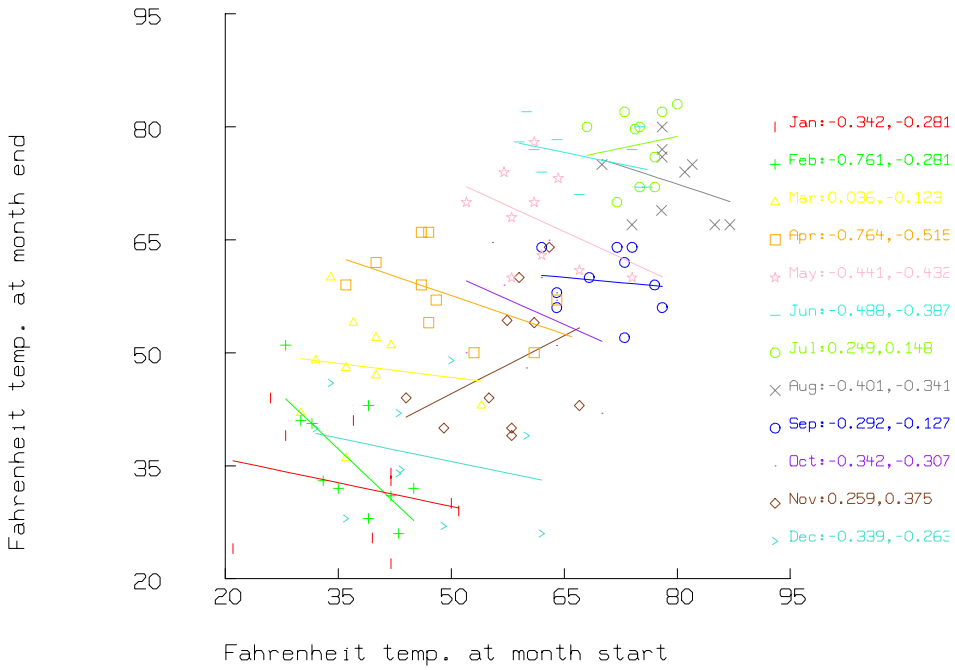
Kendall's τ coefficient, like the Pearson product-moment coefficient and the Spearman correlation coefficient, is 1 for exact linear correlation, and -1 for exact *anti*-linear correlation. The negative value of -0.02222 observed here indicates a tendency toward anti-correlation.

The March average daily temperature data for New York City on the first and last days of the month in the last decade have been seen above to be weakly anti-correlated by the linear regression, Pearson product moment, and Kendall τ measures; but linearly correlated by the Spearman correlation coefficient.

In order to gain some perspective on this result, we consider three further tests. First we consider correlation between temperatures on the first and last day of every month in the calendar. The adage regarding March weather would lead one to believe that the anti-correlation observed for the month of March is more extreme than any other month. With additional data from the web-site given above and commands similar to those given above, the following table of correlation coefficients and graph are obtained:

Correlation measures for 1st and last days of month

Month	Regression	Pearson	Spearman	Kendall
Jan	-0.2092	-0.2804	-0.3416	-0.2667
Feb	-0.9491	-0.7142	-0.7607	-0.6000
Mar	-0.1240	-0.1230	0.0368	-0.0222
Apr	-0.3433	-0.5175	-0.7127	-0.3611
May	-0.5704	-0.5801	-0.6228	-0.4444
Jun	-0.2091	-0.4503	-0.5555	-0.3611
Jul	-0.0346	-0.0283	0.0522	0.0833
Aug	-0.3798	-0.3071	-0.4102	-0.1667
Sep	0.0120	0.0154	-0.1140	0.0278
Oct	-0.4867	-0.3580	-0.4828	-0.3056
Nov	0.4007	0.2784	0.1610	0.1667
Dec	-0.2061	-0.2622	-0.3384	-0.2222

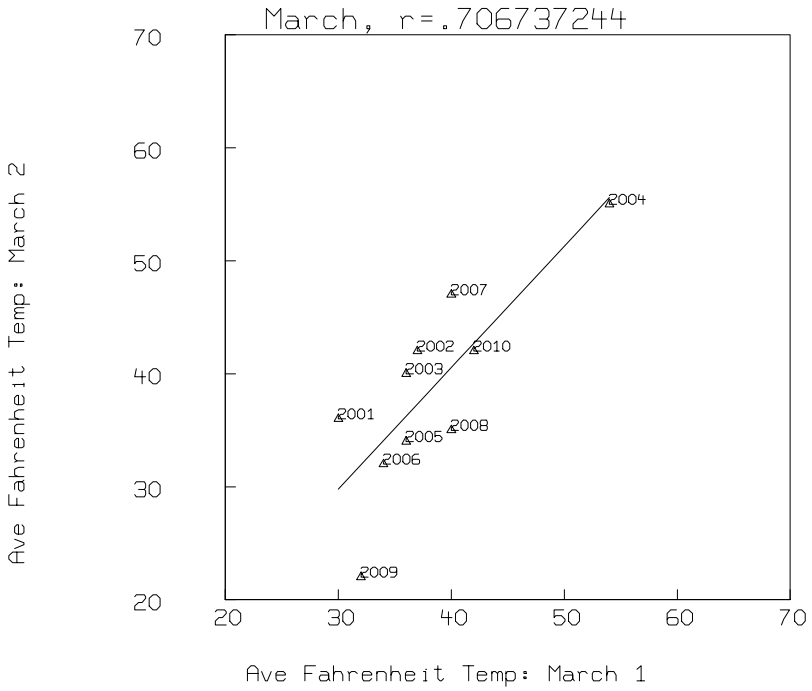


The numbers in the legend to the right side of the plot are the Spearman rank coefficient and the Pearson product-moment coefficient for each month. The line segments drawn are best-fit line-segments with colors corresponding to monthly data in the same color/symbol combination.

Note that for all months except March, July, September, and November, all four measures of association indicate anti-correlation. The month of February exhibits the largest magnitude anti-correlation by all four measures. Therefore the adage would appear to apply more to the month of February, than to March.

Another test easily done with MLAB is to consider correlation in temperatures on the first and second days of March. As weather conditions between the first and second days of a month would likely exhibit less variation than between the first and last days of a month, we would expect correlation–as opposed to anti-correlation, in the correlation measures for first-second day temperature data.

The preceeding expectation is confirmed in the following scatter plot:

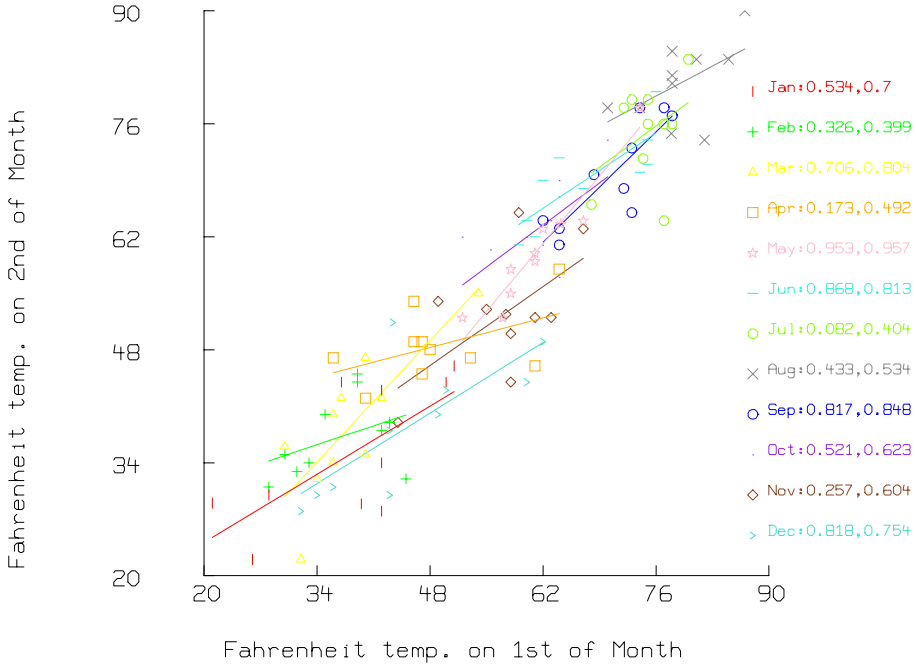


The slope of the best-fit line segment is 1.0756–a positive value confirming the expectation of correlation.

With commands similar to those above we find the linear regression slope, Pearson product-moment coefficient, Spearman rank correlation coefficient, and Kendall τ coefficient for the remaining months of the year, and generate the scatter plot for data.

Correlation measures forr 1st and 2nd days of month

Month	Regression	Pearson	Spearman	Kendall
Jan	0.6028	0.7005	0.5343	0.4667
Feb	0.3340	0.3993	0.3262	0.2667
Mar	1.0756	0.8041	0.7067	0.6222
Apr	0.2611	0.4919	0.1810	0.1667
May	1.2204	0.9603	0.9576	0.9444
Jun	0.6682	0.8140	0.9034	0.8056
Jul	0.9065	0.4874	0.1503	0.2222
Aug	0.6674	0.5267	0.3955	0.3889
Sep	0.9749	0.8172	0.7990	0.6389
Oct	0.7450	0.6229	0.5172	0.4722
Nov	0.9045	0.7059	0.3644	0.2778
Dec	0.6245	0.7548	0.8190	0.7333



It is clear that temperatures of the first and second day of the month exhibit a magnitude of correlation that is greater than the magnitude of anti-correlation exhibited in the temperatures of the first and last day of the month.

Finally, we can use MLAB to develop a simple, predictive model of temperature variation with time. We define a sinusoidal function with parameters for the amplitude `A`, frequency `B`, phase `C`, and DC offset `D`, and use the `FIT` command to find least-squares estimates of the parameters:

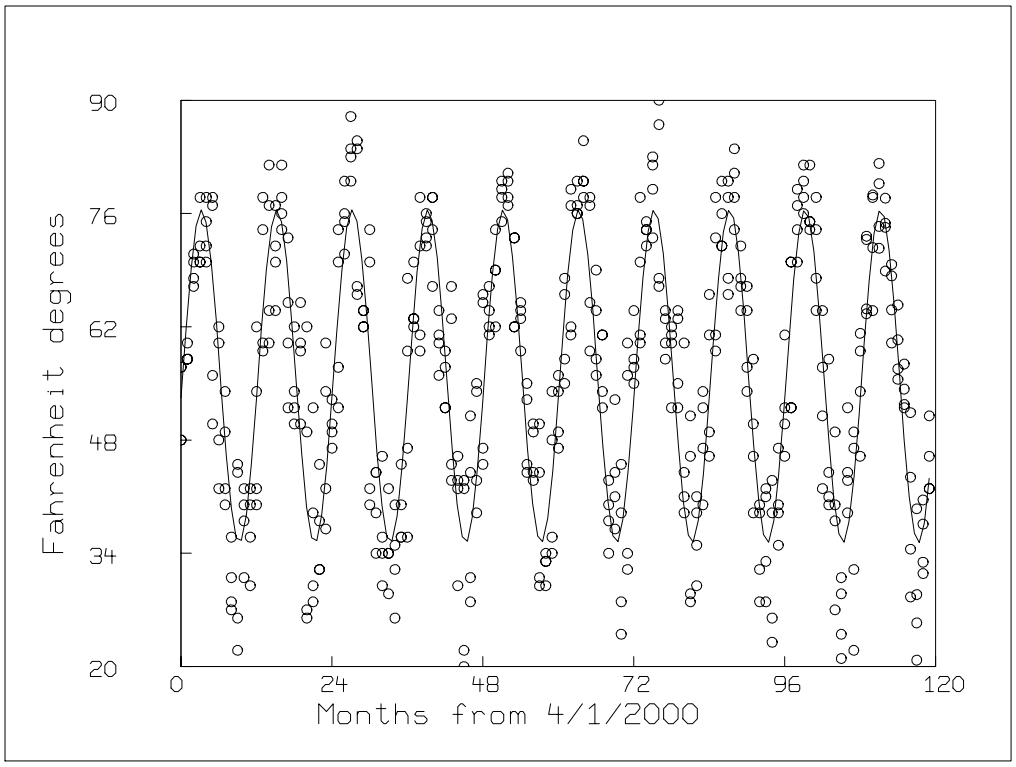
```
FCT Z(T) = A*SIN(B*T+C)+D
A = 35; B = 2*PI/12; C = 4; D = 55;
FIT (A,B,C,D), F TO MD
```

The response from MLAB is:

```
final parameter values
      value              error                dependency    parameter
   -20.59653183        0.4956031298        0.0004216946698   A
   0.5235442569      0.0006981114834          0.7425741944   B
    3.013703067       0.04763598198           0.7424864215   C
    55.87355095        0.3505327307        0.000898282782   D
7 iterations
```

```
CONVERGED
best weighted sum of squares = 2.804885e+04
weighted root mean square error = 7.676337e+00
weighted deviation fraction = 9.971410e-02
R squared = 7.840235e-01
```

A plot of the data and fitted function appears as follows:



For more information about MLAB, please visit `http://www.civilized.com`.