Time Series Project – Boston Marathon

Introduction

The first Boston Marathon was run on April 19, 1897. John J. McDermott won the first race with a time just shy of 3 hours. The 2010 race was completed in just under 2 hours, 6 minutes. I thought it would be interesting to fit a model to the data and see when the winner is expected to come in less than 2 hours.

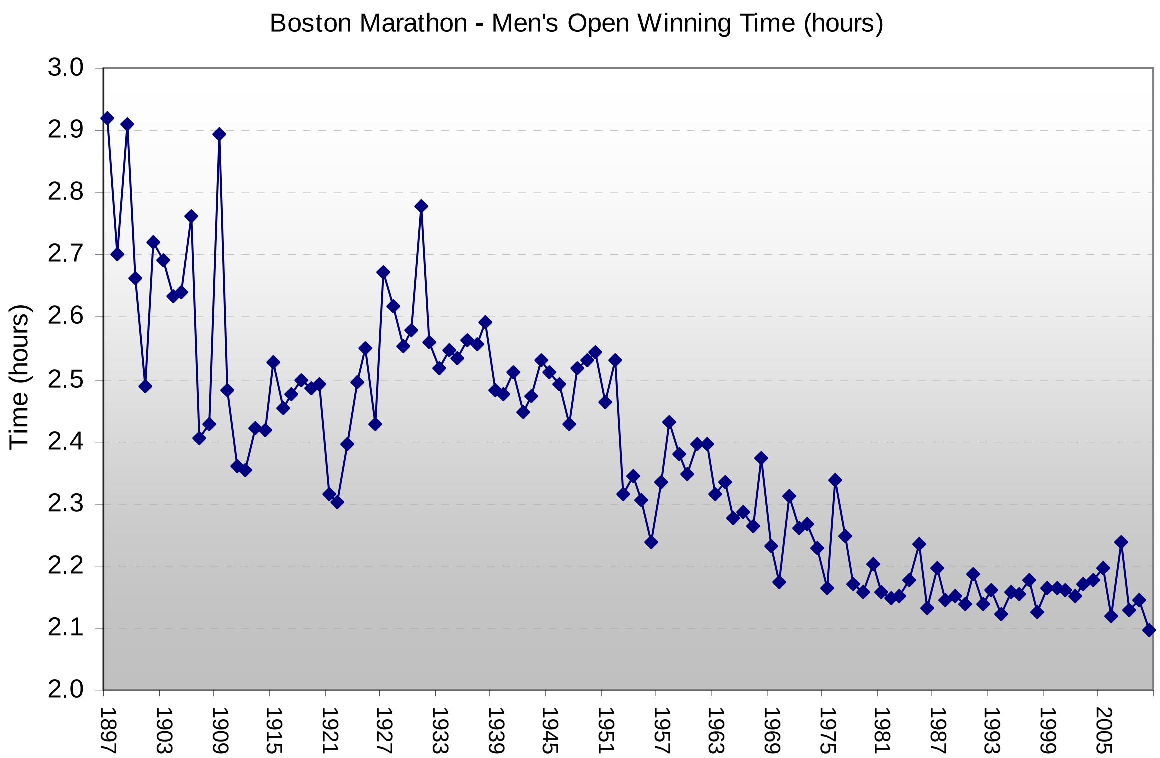
Data

After quite a bit of Google searching, I came across a great website that has lots of interesting stats including historical finishing times for the Boston Marathon:

http://www.hickoksports.com/history/alphindx.shtml

Model Specification

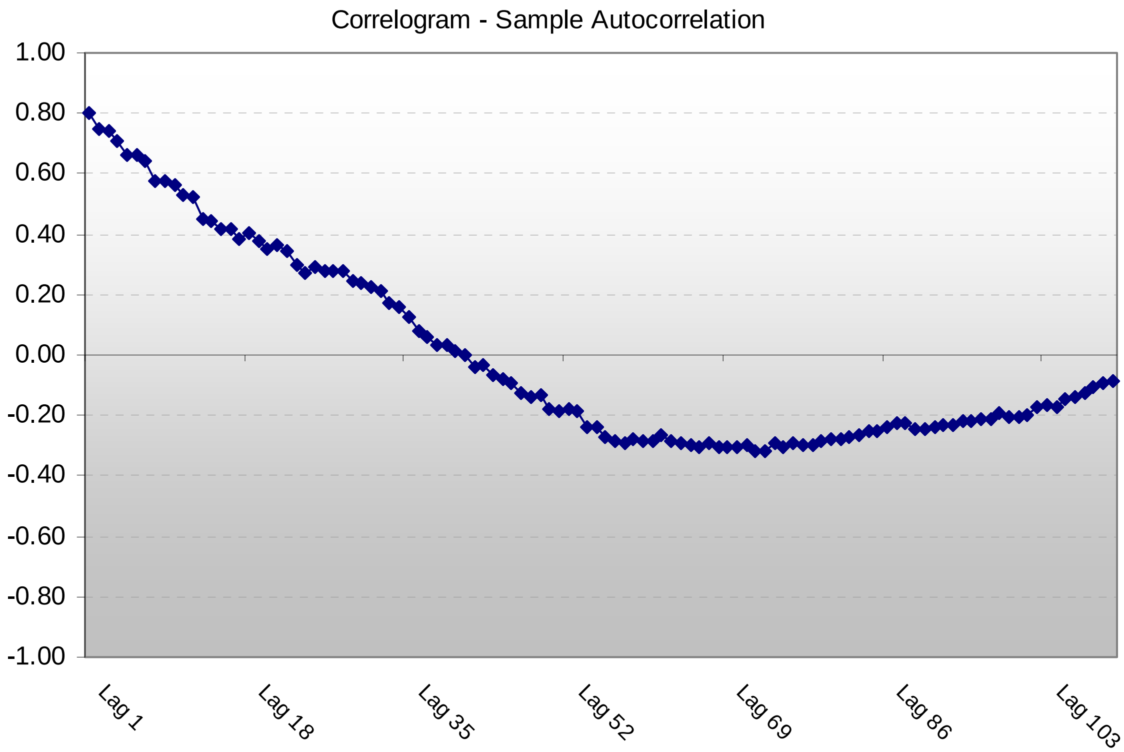
It's not surprising that the finishing time has gone down over the years. A larger number of participants and developments in running gear and technique have likely played a role in decreasing the winning times close to the 2-hour mark.



Seasonality doesn't seem to be a factor, but the data does need to be tested for stationarity. The stationarity test is based on the autocorrelation:

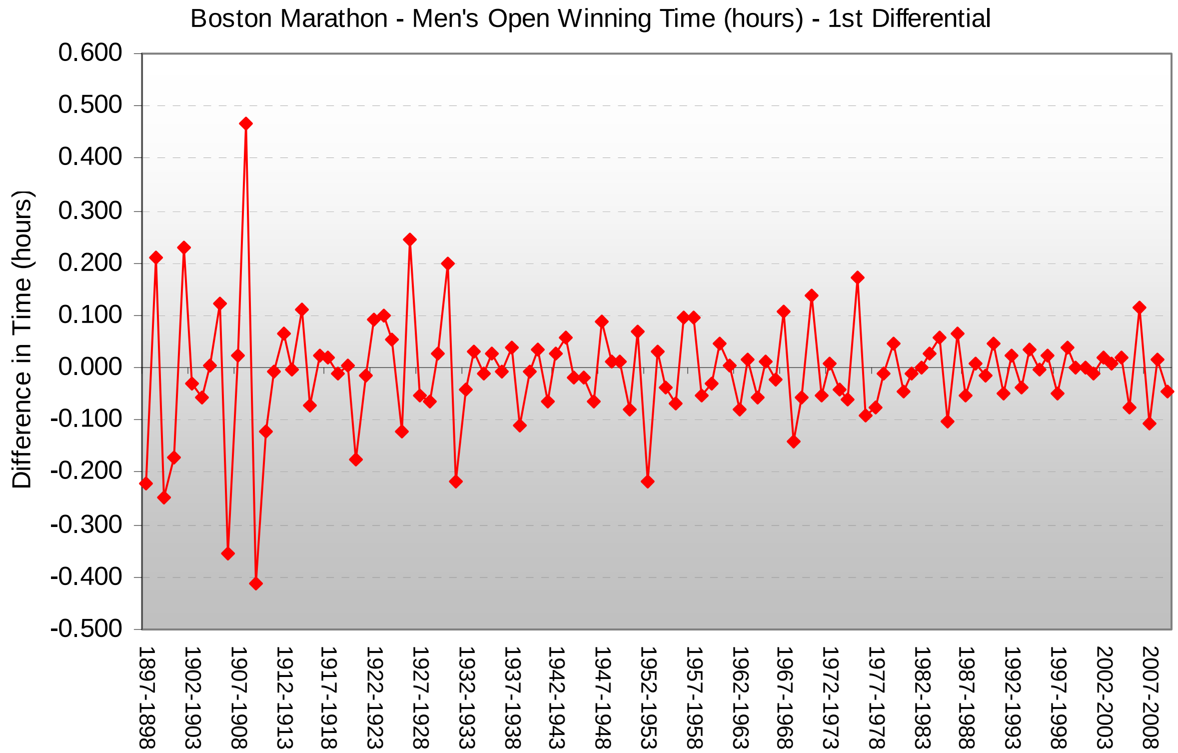
$$\hat{\rho}_k = \frac{\hat{\gamma}_k}{\hat{\gamma}_0}$$

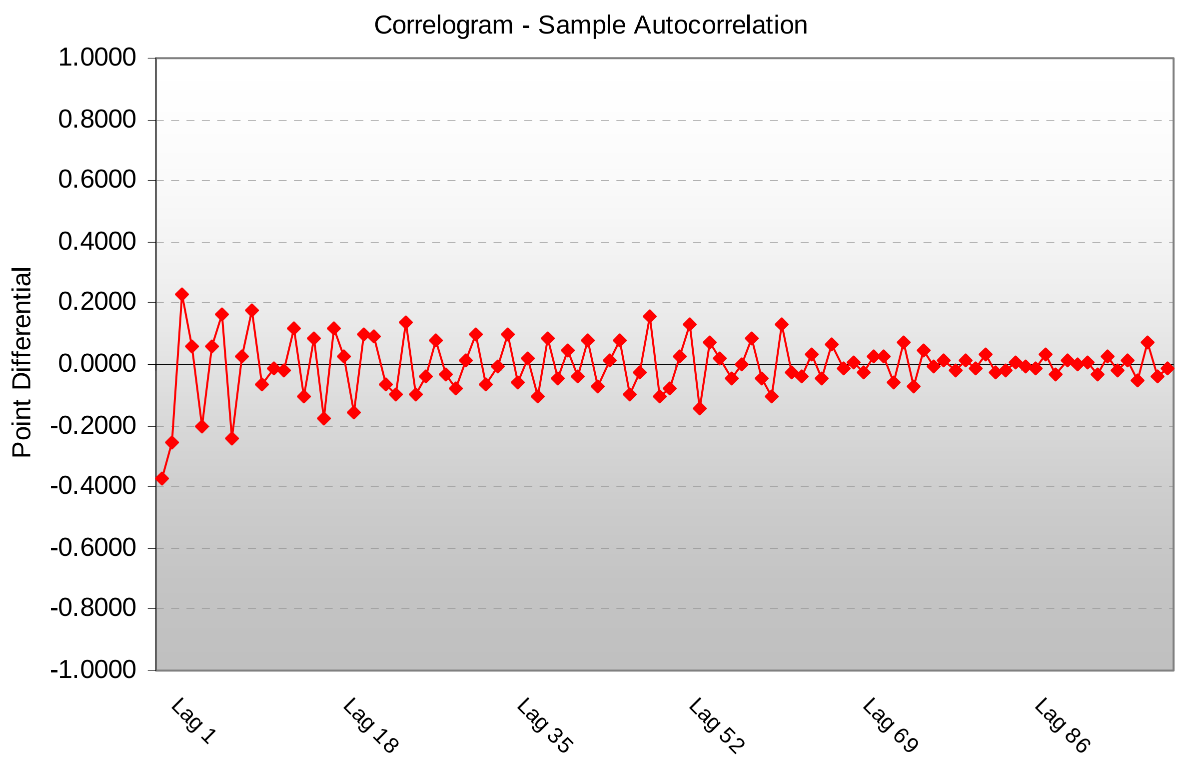
where $\hat{\rho}_k$ is the ratio of sample covariance to variance at lag k. If the data is stationary, then we expect to see the correlogram drop and remain close to zero fairly quickly:



Not surprisingly, the series for the Boston Marathon does not seem to be stationary.

By taking the first difference, we can see if the resulting series is stationary:





After taking the first difference, the series' correlogram rapidly approaches zero and appears to be stationary.

Model Parameterization

Using the stationary first differential series, the Boston Marathon winning time can be described using an autoregressive model:

$$Y_t = \delta + \sum_{i=1}^{p} \phi_i Y_{t-i} + \varepsilon_t$$

where Y_t is the data at time *t*

δ is a constant

ϕ_i is the coefficients for lag *i* data

ε_t is the error term at time *t*

p is the order of auto-regression

Using Excel, for p=(1, 2, 3) the equation is as follows:

AR(1) = -0.007938 - 0.370401 Y_{t-1} + ε_t

AR(2) = -0.013086 - 0.518298 Y_{t-1} - 0.445744 Y_{t-2} + ε_t

AR(3) = -0.012322 - 0.540221 Y_{t-1} - 0.531690 Y_{t-2} - 0.108443 Y_{t-3} + ε_t

The absolute value of the sum of ϕ_i is less than one for the models, as well as the individual components.

	R^2	Adj. R^2
AR(1)	0.141823	0.134022
AR(2)	0.308158	0.295346
AR(3)	0.327718	0.308691

AR(3) has the highest R^2 and adjusted R^2.

Durbin-Watson

$$d = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}.$$

	DWS
AR(1)	2.278
AR(2)	1.984
AR(3)	2.068

For the Durbin-Watson statistic, values close to two reinforce the null hypothesis indicating no serial correlation among the residuals.

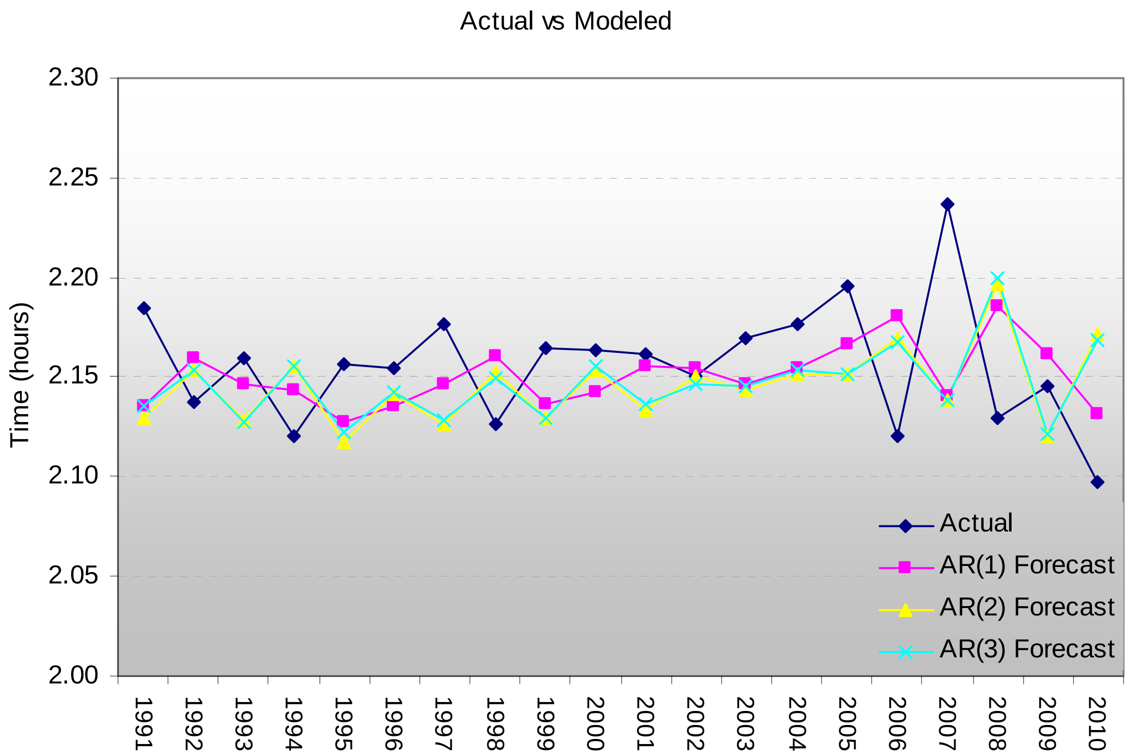
Box-Pierce test

$$Q = T\sum_{k=1}^{s} r_k^2.$$

	BP
AR(1)	90.01
AR(2)	50.02
AR(3)	49.30

For the Box-Pierce test, at the 10% significance level with 110 degrees of freedom the Q statistic is less than the chi-squared statistic reinforcing the null hypothesis that the residuals are a white-noise process.

Modeled Results vs Actual



Betting Time

The AR(3) model has the highest R^2 and adjusted R^2, a Durbin Watson statistic near 2 indicating a failure to reject the null hypothesis of serial correlation, and a Box-Pierce Q statistic reflecting at the 10% significance level that the residuals are white noise. Therefore, it's a sure bet the model is correct and money should be wagered! The AR(3) model forecasts that 2032 will be the first year a runner completes the Boston Marathon in under 2 hours. I'm calling Vegas now…