Christopher J. Pailes
Time Series
Summer 2011
Project

Time Series Analysis of the Average Number of Stolen Bases Per Game by the Texas Rangers

Introduction and Data

This project endeavors to fit an ARIMA model to the time series of the average number of stolen bases per game (i.e. SB/GS) by month for the Texas Rangers Major League Baseball team. The data used for the analysis was taken from BASEBALL-REFERENCE.COM (the following is the specific URL for Texas Rangers team statistics: http://www.baseball-reference.com/teams/TEX/batteam.shtml; by selecting a particular year and then selecting "Splits", one can view totals by month). Using data for regular season games only (i.e. Spring Training as well as postseason games were excluded), the averages were calculated from 1996 through 2011.

The Major League Baseball regular season usually occurs between April 1st and September 30th. However, depending upon scheduling year to year, sometimes a few games are played in March and/or October. Since the observations are averages and not totals, for the purposes of this analysis, March data was combined with April data and October data was been combined with September data. As a result, the time series consists of only six months of data per year (i.e. April through September), with the remaining months ignored. Using six months of data per year from 1996 through 2010 gave 90 total data points that were used to develop an appropriate ARIMA model to then forecast 2011 values.

Please refer to the provided Excel workbook named "Time Series Analysis.xlsx" for the background work behind the ARIMA model development.

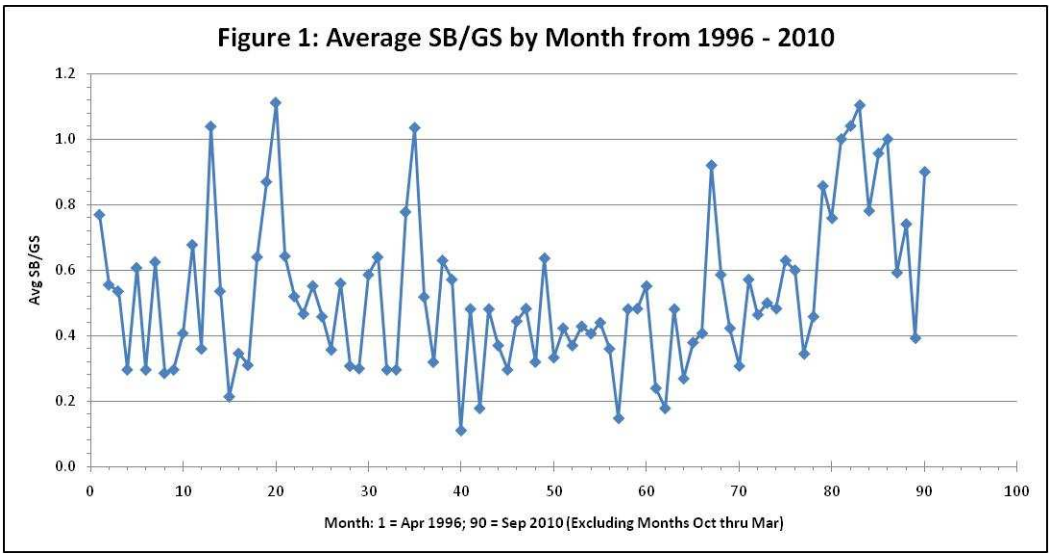
Confirmation That Time Series is Not White Noise

I began the process of developing an ARIMA model by first testing the time series data to confirm that it was not merely a white noise process. This was accomplished by first determining the sample autocorrelations for lags 1 through 89 and then calculating the Box-Pierce Q statistic from these sample autocorrelations (see the tab "Original Time Series" in the provided Excel workbook). For the first 54 lags, the BPQ statistic is greater than the corresponding χ^2 critical value at a 10% significance level for degrees of freedom equal to the corresponding lag. This result suggests that we reject the null hypothesis that the time series is a white noise process.

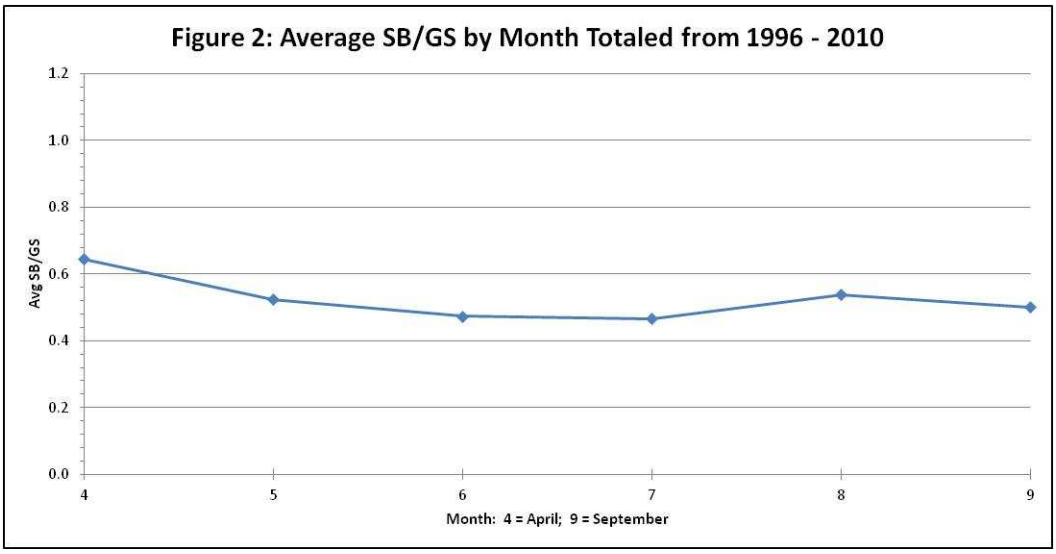
However, since the remaining 35 BPQ statistics did not exceed the χ^2 critical value, an additional test was performed. Specifically, knowing that the Ljung-Box test is actually more reliable than the Box-Pierce test, the corresponding Ljung-Box Q* statistics were calculated for all 89 lags as well. Reviewing the results, it was found that the LBQ* statistics exceed the corresponding χ^2 critical value at a 10% significance level for all 89 lags. Therefore, the null hypothesis that the time series is a white noise process can be confidently rejected.

Development of Stationary Time Series

The 90 data points from the original time series were plotted in Figure 1 in order to identify any patterns such as seasonality or trends that should be accounted for when developing the ARIMA model.



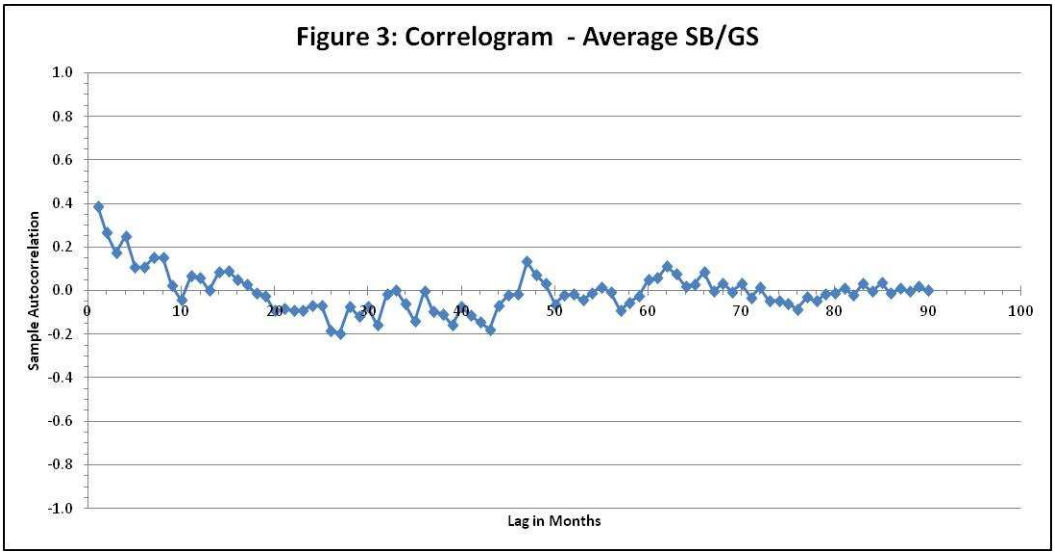
The time series itself shows significant volatility and any patterns that may exist are difficult to decipher without additional analysis. Therefore, in order to check for seasonality, I calculated the monthly averages of stolen bases per game for 1996 through 2010 combined. These averages were then plotted in Figure 2. As can be seen from the plot, there are only minor fluctuations in the averages from month to month indicating that there is no significant seasonality that needs to be accounted for in the model.

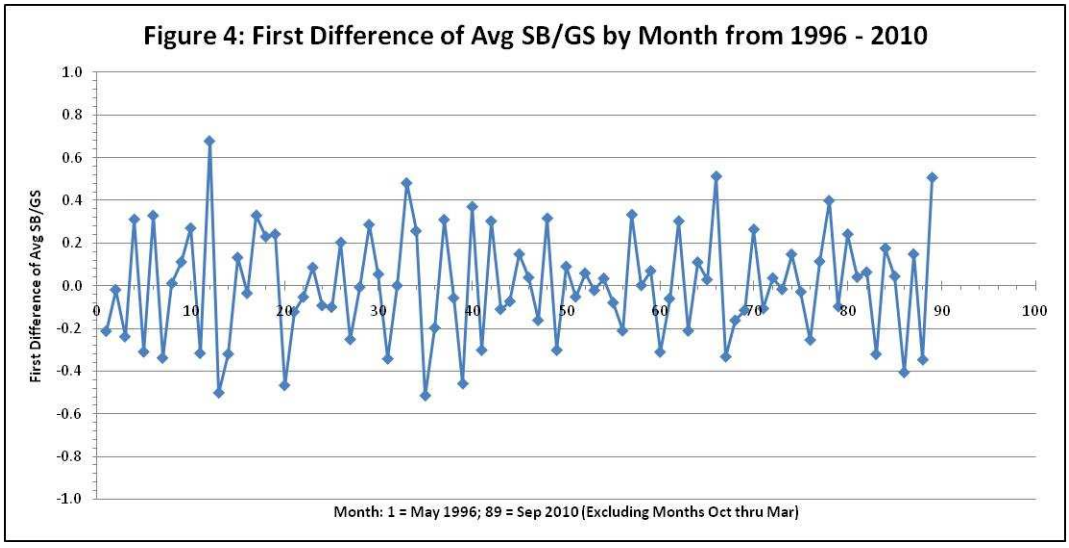


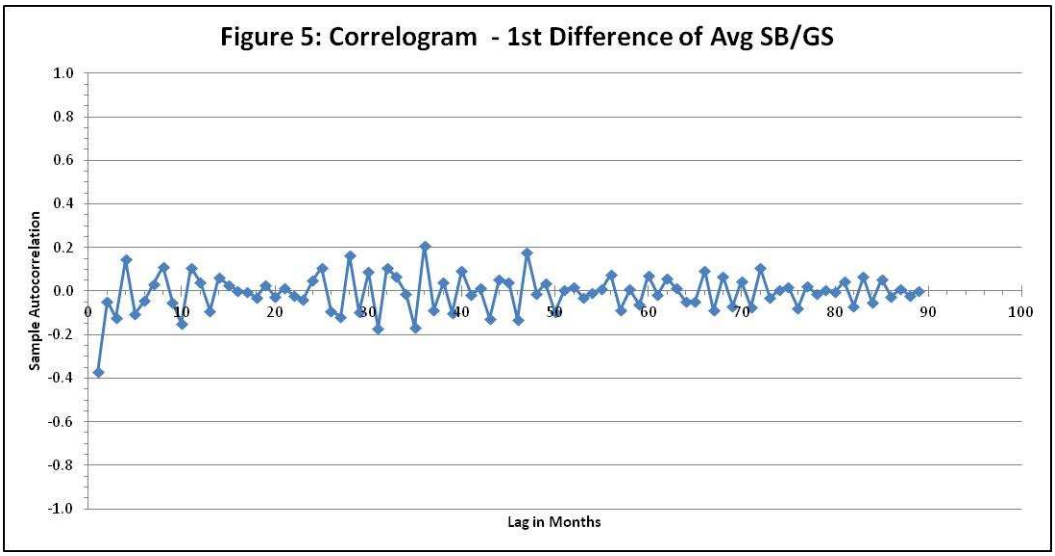
In order to continue evaluating the time series and determine if it's stationary, I created a correlogram which is a plot of sample autocorrelations by lag (see Figure 3). Because the sample autocorrelations do not reach zero until approximately lag 10 and don't remain near zero until around lag 70, the original time series is not stationary and a transformation is needed. There appears to be a trend causing the time series to be non-stationary. In order to eliminate it, first differences were taken (see the tab "1st Diff Time Series" for the related work). Figures 4 and 5 below respectively are plots of the corresponding time series as well as correlogram.

Figure 4 at first glance appears to indicate that the time series of first differences may be stationary because the level of the time series remains consistent over time. Further confirmation of this assumption is provided by reviewing Figure 5 and observing that the sample autocorrelations (in absolute value) quickly decline to zero and stay near zero for the remaining lags.

Additionally, knowing that the standard deviation of a white noise process with 89 observations is $\frac{1}{\sqrt{89}} = 10.6\%$ and then taking the absolute value of the sample autocorrelations for lags 2 through 89, it is found that only 15 exceed one standard deviation and none exceed two standard deviations. This confirms the general impression that the sample autocorrelations remain near zero for lags 2 through 89. Thus, since the sample autocorrelations decline quickly to zero and remain near zero, the time series is stationary and no additional transformations are necessary.







Model Fitting

The next step involved fitting various ARMA models to the time series of first differences.

AR(1) Model

The first model to be fit was an AR(1) model. This was done by regressing ΔY_t onto ΔY_{t-1} by using Excel's LINEST() function. Note, the use of this function was chosen over the Regression tool in the Analysis ToolPak because LINEST() was found to be less cumbersome to use. See the tab "AR(1) Fit & Test" for the corresponding calculations. The following is the model that resulted:

$$\Delta Y_t = -0.3893\Delta Y_{t-1} + e_t + 0.0022$$

AR(2) Model

The next model to be fit was an AR(2) model. This was accomplished by regressing ΔY_t onto ΔY_{t-1} and ΔY_{t-2}. Again, this regression was performed using Excel's LINEST() function and the corresponding work can be found in the "AR(2) Fit & Test" tab. The following is the model that resulted:

$$\Delta Y_t = -0.4745\Delta Y_{t-1} - 0.2252\Delta Y_{t-2} + e_t + 0.0032$$

MA(1) Model

To fit the data to an MA(1) model, the equation for the autocorrelation at lag 1 of an MA(1) process was solved for θ:

$$\rho_1 = \frac{-\theta}{1+\theta^2} \quad \rightarrow \quad \theta = \frac{-1 \pm \sqrt{1-4\rho_1^2}}{2\rho_1}$$

The parameter θ was then estimated by substituting the sample autocorrelation at lag 1 for the real autocorrelation in the equation above. Since this involves solving a quadratic which gives two values, the value of θ whose absolute value was less than 1 was taken as the estimated parameter. Additionally, θ_0 was estimated as the mean of the time series (see the tab "MA(1) Fit & Test" for the corresponding work). The following is the model that resulted:

$$\Delta Y_t = e_t - 0.4462e_{t-1} + 0.0015$$

ARMA(1,1) Model

Lastly, the data was fit to an ARMA(1,1) model. The parameter ϕ was estimated from the equation $\phi = \frac{\rho_k}{\rho_{k-1}}$ where the sample autocorrelations at lags 1 and 2 were substituted in place of the actual autocorrelations at lags 1 and 2. Then, the estimated value of ϕ and the sample autocorrelation at lag 1 were substituted into the equation $\rho_1 = \frac{(1-\theta\phi)(\phi-\theta)}{1-2\phi\theta+\theta^2}$ which was then solved for θ using Excel's Goal Seek tool. Finally, θ_0 was calculated as the mean of the time series multiplied by $1-\phi$ (see the tab "ARMA(1,1) Fit & Test" for the corresponding work). The following is the model that resulted:

$$\Delta Y_t = 0.1302\Delta Y_{t-1} + e_t - 0.6298e_{t-1} + 0.0013$$

Model Diagnostics

After fitting AR(1), AR(2), MA(1), and ARMA(1,1) models to the first differences of the time series data, the models were checked for goodness-of-fit with the ultimate goal being to choose the model that most closely fits the data.

Durbin-Watson Statistic

The first test performed involved the calculation of the Durbin-Watson statistic which tests for first-order serial correlation among the residuals (the details behind the calculation of the Durbin-Watson statistic can be found in the corresponding tabs for each model). The calculated statistics are shown in the following table:

Model	DWS
AR(1)	2.146
AR(2)	2.093
MA(1)	2.171
ARMA(1,1)	2.160

In all four cases, the DWS is close to two indicating no first-order serial correlation among the residuals. More generally speaking, this indicates the residuals may come from a white noise process which is a requirement of an ARIMA model to be a good fit. Therefore, any of the four models may be a good fit; however, additional testing is required to determine the optimal model.

Box-Pierce Q Statistic

The next test for goodness-of-fit involved the calculation of the Box-Pierce Q statistic which is a more direct check for whether or not the residuals from the fitted models are a white noise process. Given that there are roughly 90 data points, the BPQ statistics at the approximate halfway point of lag 45 were pulled and summarized in the table below:

	Box-Pierce Q Test at Lag 45		
Model	Degrees of Freedom	BPQ	χ^2 for α = 10%
AR(1)	44	41.429	56.369
AR(2)	43	30.929	55.230
MA(1)	44	31.324	56.369
ARMA(1,1)	43	28.717	55.230

For all four models, the BPQ statistic is less than the corresponding χ^2 critical value at a 10% significance level. Therefore, we do not reject the null hypothesis that the residuals come from a white noise process. Given this result, any of the four models may be a good fit. Ergo, an additional distinguishing factor must be taken into account in order to determine the model that is the best fit.

Sample Autocorrelation Versus Implied Autocorrelation

The final test performed involved the comparison of the sample autocorrelations of the time series against the autocorrelations implied by the parameters estimated for the models of the time series. It is easy to see from Figure 5 above that the sample autocorrelation drops to around zero immediately when going from lag 1 and to lag 2. It then continues to stay near zero for the remaining lags. The only reason the sample autocorrelation is not exactly zero is because of random fluctuations in the actual time series. Nonetheless, this type of behavior is indicative of an MA(1) model which has an implied autocorrelation of zero starting at lag 2. On the other hand, models with auto-regressive components exhibit geometric decline which does not appear to be occurring here.

Conclusion and 2011 Forecast

Because the tests above for the AR(2) and ARMA(1,1) models did not indicate a materially better fit than the AR(1) and MA(1) models, by the Principle of Parsimony, the AR(2) and ARMA(1,1) models are rejected as possible models. Additionally, because the implied autocorrelation of an AR(1) model declines geometrically but instead the sample autocorrelation drops off suddenly, the AR(1) model is rejected as a possible fit. Thus, the model that fits best is the MA(1) model:

$$\Delta Y_t = e_t - 0.4462e_{t-1} + 0.0015$$

Given that the first differences of the original time series were modeled by a stationary MA(1) process, the original time series itself follows an ARIMA(0,1,1) process with the following equation:

$$Y_t = Y_{t-1} + e_t - 0.4462e_{t-1} + 0.0015$$

As a final form of validation of the model chosen, the average stolen bases per game by month for the 2011 season were forecasted and plotted along with the actual averages (see Figure 6). Although the forecasted values aren't too far off from the actual, they are a little low and the model doesn't appear to do a good job of forecasting future values. However, reviewing the data on stolen bases, it turns out that since 1996, there's been only one season in which the Rangers stole more bases than in 2011. Therefore, the larger than desired difference between the forecasted and actual values may be merely due to random fluctuation, and the MA(1) model still appears to be a good fit.