Alex Sarver

Dr. Kristin McCullough

STAT 578

2 December 2016

Analysis of Chicago Cubs Ticket Sales Time Series

One of the most statistically impossible events occurred in 2016, the Chicago Cubs won the World Series. As a group of dutiful statisticians, our group set out with the purpose to analyze the number of tickets sold at Wrigley Field from 1925 to 2014 and see if we could model and predict ticket sales. Examining the plot of ticket sales (see Figure 1) from 1925 to 2014 yields two interesting observations. The first observation was a sharp decline in ticket sales followed by an astronomical jump to a value higher than what was previously observed. A general recession of the U.S. economy followed by a strike shortened 1981 season is what lead to the drastic drop. The astronomical leap was caused by a combination of the U.S. economy improving, the Baby Boom generation reaching early adulthood, and an expansion of the bleachers over the catwalk in Wrigley Field. Considering this drastic change in conditions, I decided to cut off the data and only use the years from 1985 on.

Examining the plot of the ticket sales from 1985 on (Figure 2), there was another sharp decrease followed by a shape increase of ticket sales; this time the decrease was 1994 and the increase wasn’t until 1996. The depressed ticket sales were the result of a major player strike that shortened both the 1994 and 1995 season. The sharp increase from about 1996 to 1998 was a result of the combination of the introduction of interleague play during the regular season, recovery from the 1994 player’s strike, and the historic race to break Mickey Mantle’s single season home run record by both Mark McGuire and Sammy Sosa (who played for the Chicago Cubs at the time). Since the strike shortened both the 1994 and 1995 season, I removed both years from my time plot, which yielded the plot given in Figure 3.

After removing the outliers and shortening my time series, the result was a time series that was not stationary. Along with the increasing trend shown in figure 3, the Dickey-Fuller test returned a p-value that was greater than 0.10. Since I did not have a stationary series, I decided to transform my data until the time series was stationary. I took the log of my time series and examined the plot (Figure 4). The plot suggests that it is still not stationary, and the Dickey-Fuller Test confirmed it. I took the first difference of the time series and the first difference of the logs and examined the time plots of both (Figure 5 and 6 respectively). After confirming with the Dickey-Fuller test, I concluded that both data sets were stationary and proceeded to model both.

For the time series of the first differences, examining a plot of the ACF (Figure 7) showed a possible significant correlation at lag 4. I noticed that lag 8, while not as much as lag 4, was higher than the correlation of the surrounding lags. This shows that the time series could be modeled with a Seasonal IMA (1,1) with a period of 4. Examining the EACF (Figure 8) showed that I could model the time series using an ARIMA (5,1,1). An analysis of the ARIMA subset models showed that the best subset would have been and ARIMA (6,1,5) containing only the following terms: AR1, AR3, AR4, AR6, MA1, MA5, and the intercept.

For the time series of the first difference of the log, the ACF (Figure 9) showed nearly the same thing as the ACF for the first difference, thus I concluded that a Seasonal IMA (1,1) with a period 4 could be the best model as well. While checking for any subset models that might work, I got an interesting graph (Figure 10). Essentially, it showed that ARI models with autoregressive degree of 2 to 6 and a differencing degree of 1. I decided to do is when I use R to fit an ARI (6,1); if I found any evidence of insignificant coefficients, I would reduce the autocorrelation degree by one and try and fit that model. I planned to keep going until I found a model where all the coefficients were significant.

Figure 11 shows the result of me fitting a seasonal IMA (1,1) with period 4, ARIMA (5,1,1) and a sub-setted ARIMA (6,1,5) with the first difference data. The ARIMA (5,1,1) and sub-setted ARIMA (6,1,5) models have insignificant coefficients, so I eliminated those models from consideration. The seasonal IMA model did not have an insignificant coefficient, so I kept that model for further analysis. Figure 12 shows the results for seasonal IMA (1,1) with period 4 for the first difference of the log data and Figure 13 shows the results for ARI models (6,1) to (3,1) for the first difference of the log data. The seasonal IMA (1,1) model with the first difference of the logs did not have an insignificant coefficient, so I kept that model in consideration. The AIC for the seasonal IMA (1,1) using the first difference of the logs was a lot lower than the model of the first difference (742.22 versus -51.04). When examining the ARI models, starting with ARI (6,1), I kept getting insignificant coefficients. I kept reducing the auto-regressive order until I got a model with all significant coefficients, which turned out to be an ARI (3,1). So, I kept the ARI (3,1) model in consideration.

In my examination of the residuals, I ended up concluding that all three models that I kept in consideration could be used to model the data. Figure 14 shows the analysis of the residuals for seasonal IMA (1,1) for the first difference data, Figure 15 shows the analysis of the residuals for the seasonal IMA (1,1) for the first difference of the logs data, and Figure 16 shows the analysis of the residuals of the ARI (3,1) model of the first difference of the logs data. Looking at all three figures, I did not see anything that I would consider to be inappropriate and all the residuals seemed to follow a normal distribution. I did a Shapiro-Wilks test to see if the distributions of all three sets of residuals were normal. All three tests did show that all three sets of residuals were in fact normal; with the seasonal IMA (1,1) of the first difference of the data having a p-value of 0.5977, the seasonal IMA (1,1) of the first difference of the logs have a p-value of 0.3748, and the ARI (3,1) has a p-value of 0.7331. I picked the two models with the residuals with the most normal distributions, the seasonal IMA (1,1) of the first difference and the ARI (3,1) model.

Figure 17 shows a plot of the observed time series and of the Seasonal IMA (1,1) with period 4. The model doesn’t vary as much as the observed data, however it does appear to model the data well. Figure 18 shoes a plot of the observed time series and the ARI (3,1) model. The model does vary a lot closer to the way that the observed times series varies. The model does seem to tail towards an increasing trend at the end.

Figure 19 shows the forecast of the ARI (3,1) for the next ten years and Figure 20 shows the untransformed and back-transformed figures. In looking at the forecasts of the ARI (3,1) model, the standard error is typically greater than the actual predictions given by the model. Another troubling sign is that the model stops producing negative exponents (which would signify a decrease) and shows an exponential growth in ticket sales, but, we know there must be an upper limit (the capacity of Wrigley Field is the easiest constraint to consider).

The forecast given by the seasonal IMA (1,1) model is very similar. Figure 21 shows that the standard errors are several orders of magnitude larger than the actual predictions that are given. Figure 22 gives the untransformed and the back transformed forecasts. We know that the forecasts will converge toward the process mean, which for the seasonal IMA (1,1) model is after year 5. I would not trust either model past 5 years. Between the two models, I would say that the ARI (3,1) is probably the better model to use because the AIC is low compared to the other models and the residuals are the most normal.

Since the data cuts off at 2014, we can compare both models’ estimates to the actual totals from 2015 and 2016 (Which is given by Figure 23). The ARI (3,1) was off the actual sales amount of tickets in 2015 by only 155,246 tickets. I feel that the shortcoming of both models is that it fails to consider the performance of the team, which I suspect is the best predictor of ticket sales. Since the Cubs did extremely well in 2015 and 2016 compared to years past, it makes sense that both models underestimated the ticket sales. If the Cubs did worse than average, I would suspect my models would overestimate ticket sales. A way to improve my model is to use regression and incorporate the number of wins that a team had in the current year and years past. To summarize I would say that my model gives a decent forecast for two to three years into the future, however, it would have to be revised after every year to ensure that it stays current with existing trends. To conclude, I would say that ticket sales will increase for the Cubs over the next 2-3 years and will decline when the Cubs eventually do.

Figure 1



Figure 2



Figure 3



Figure 4



Figure 5



Figure 6



Figure 7



Figure 8

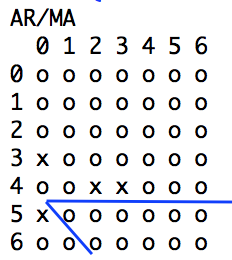


Figure 9



Figure 10



Figure 11

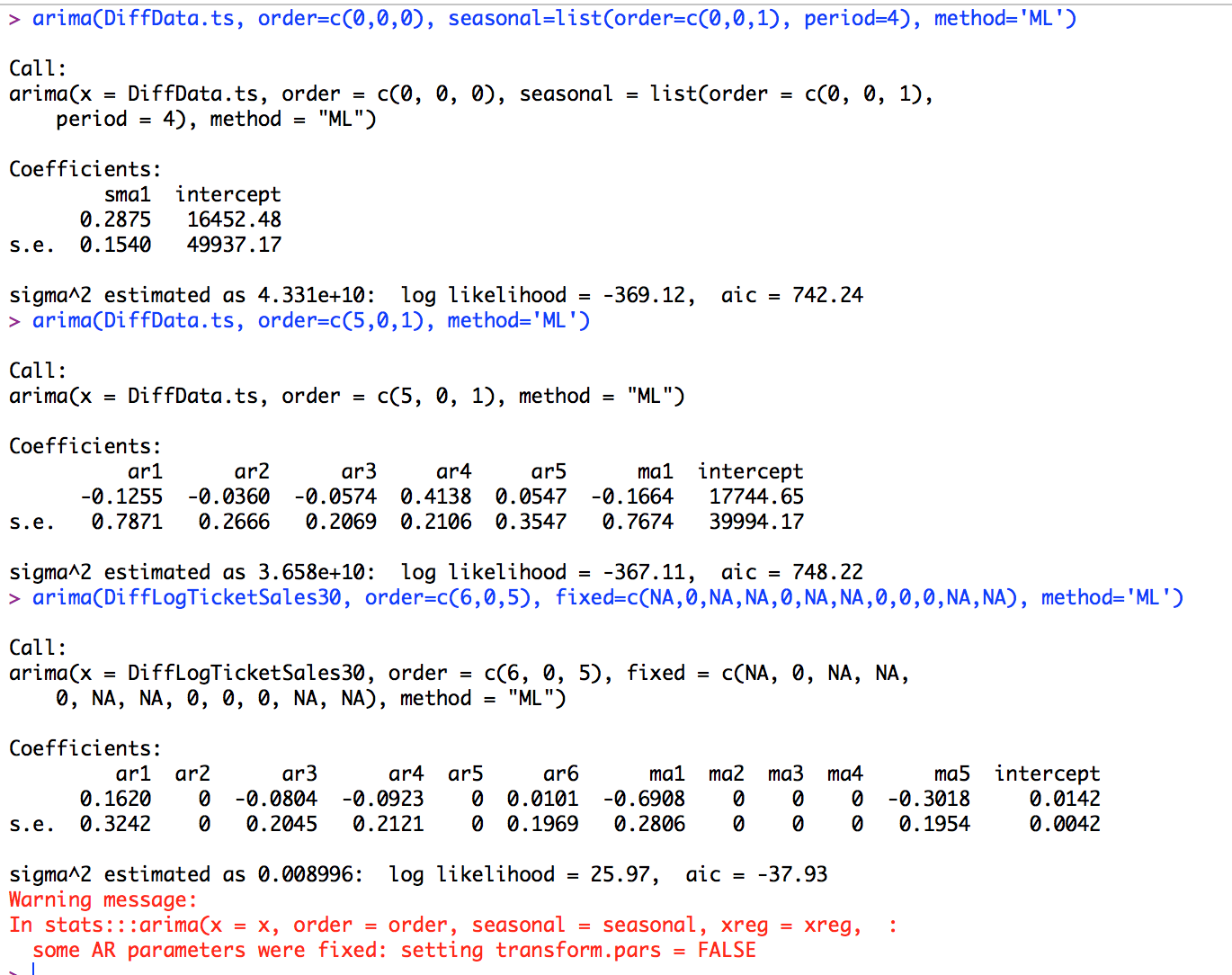


Figure 12

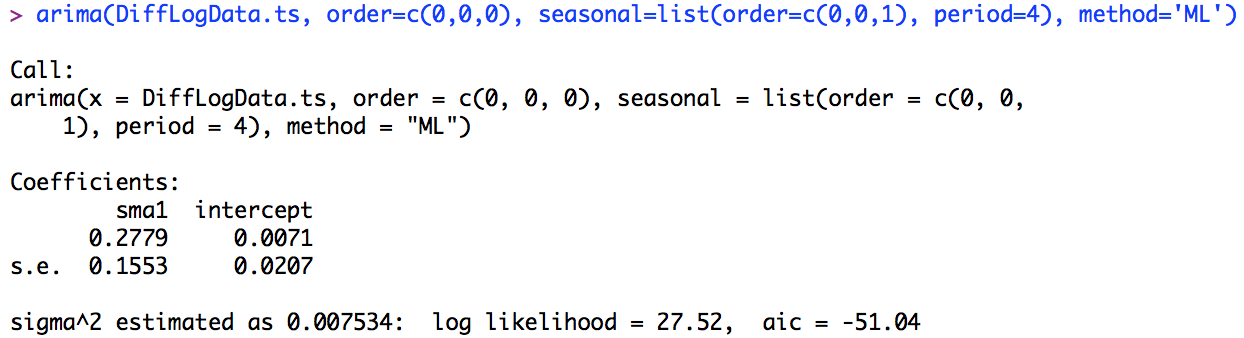


Figure 13

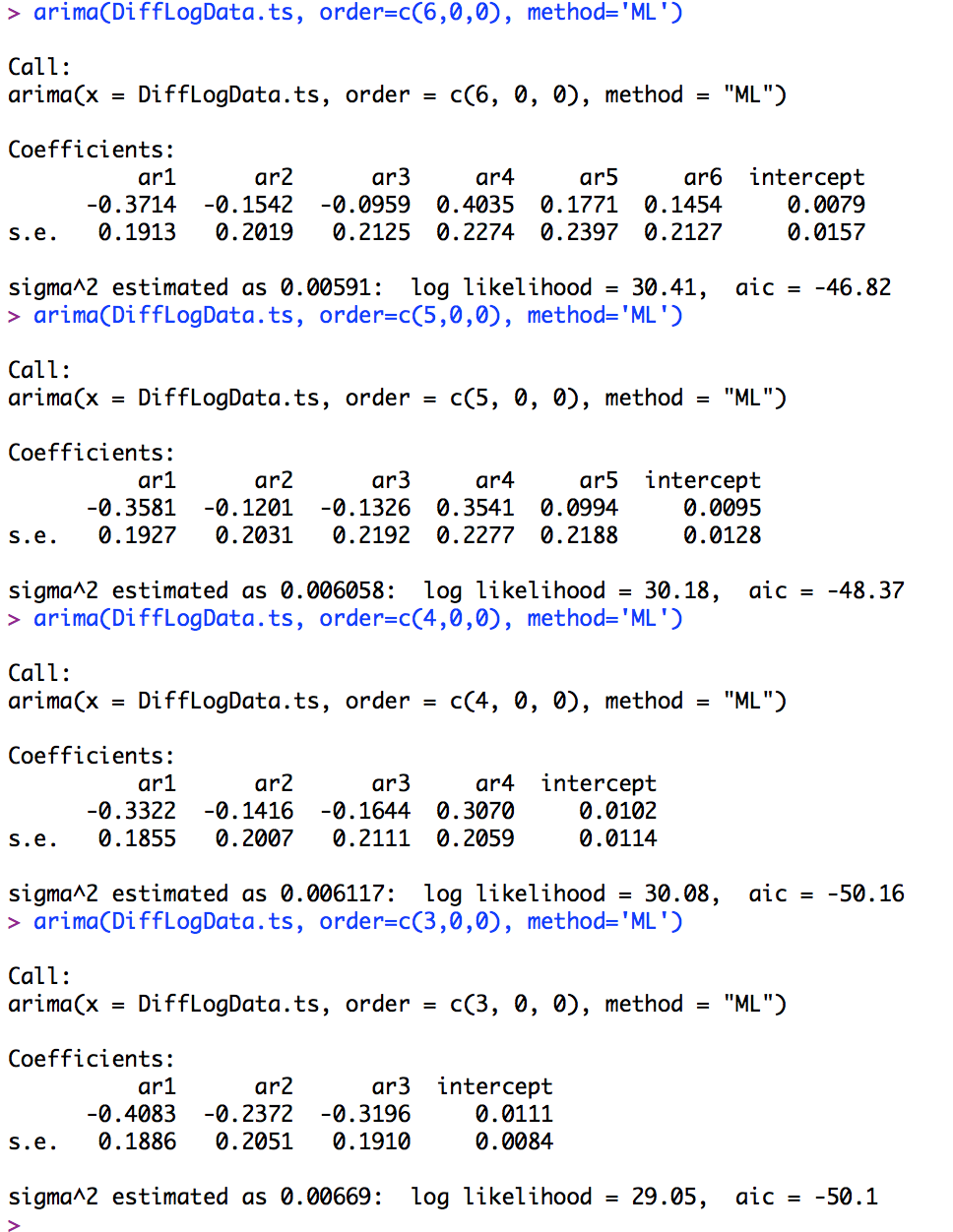


Figure 14



Figure 15



Figure 16



Figure 17



Figure 18



Figure 19

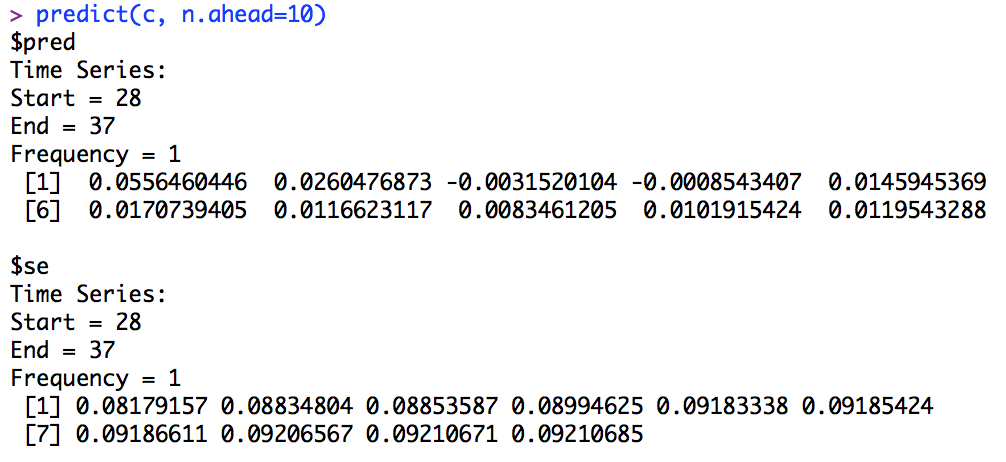


Figure 20

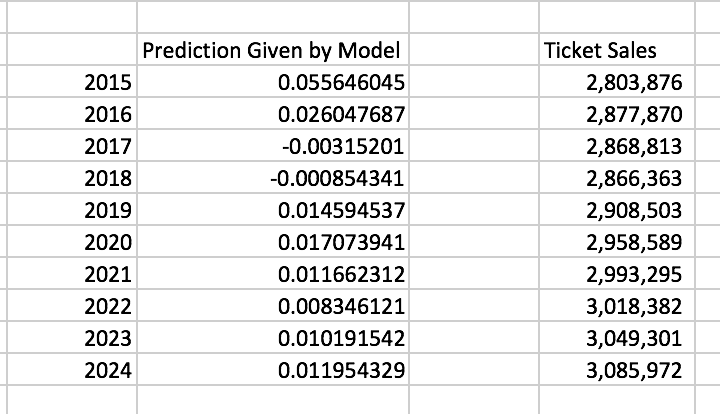


Figure 21

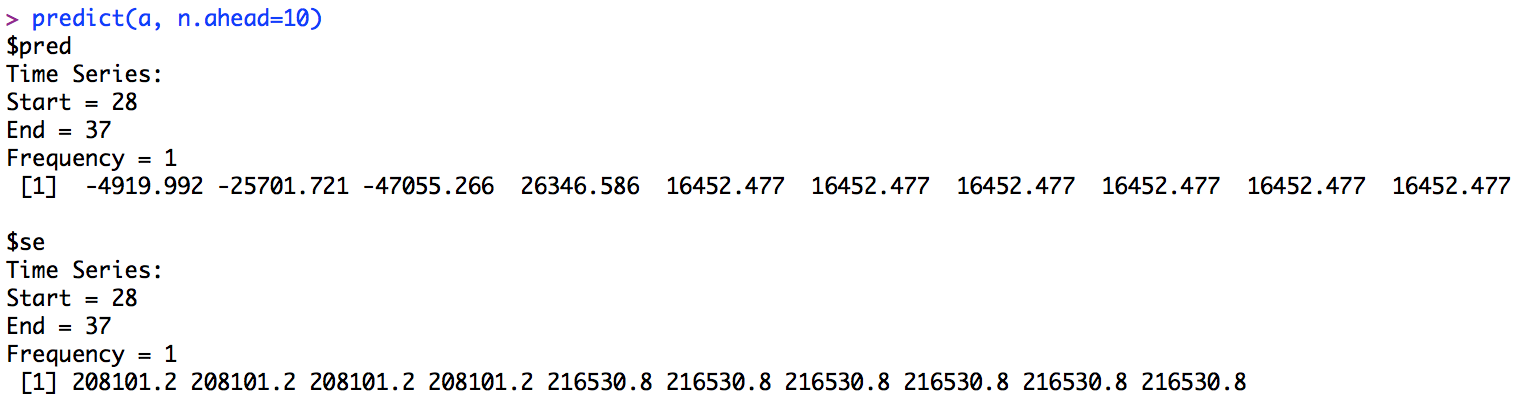


Figure 22

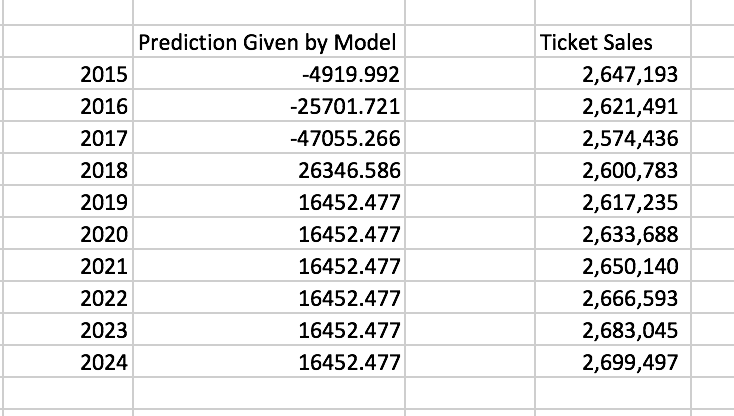


Figure 23

