**SARIMA Models – OSU Example**

Example: OSU enrollment data (osu\_enroll\_MB.R, osu\_enroll.csv)

The Office of Planning, Budget, and Institutional Research (OPBIR) makes fall semester enrollment forecasts in their annual Student Profile report. At the beginning of the course, we examined an O’Colly article which discussed these forecasts. In particular, the forecast for fall 2000 was 743 students too high resulting in a $1.8 million tuition shortfall. According to an official OPBIR document, forecasts were made using the following method:

New freshmen projections are based on an expected market share of projected Oklahoma ACT test takers. On-campus sophomore, junior and senior projections are based on cohort survival rates using an average of the previous three years. OSU-Tulsa undergraduates are projected to increase approximately 5% per year.

According to OPBIR, the enrollment data was collected as close to the final Drop and Add Day as possible.

For this analysis, I am going to use data up to the spring 2002 semester only to simulate what would be available for wanting to forecast fall 2002. Below are the estimated ACF and PACF plots of the data.

> osu.enroll <- read.csv(file = "OSU\_enroll.csv",

 stringsAsFactors = TRUE)

> head(osu.enroll)

 t Semester Year Enrollment date

1 1 Fall 1989 20110 1989-08-31

2 2 Spring 1990 19128 1990-02-01

3 3 Summer 1990 7553 1990-06-01

4 4 Fall 1990 19591 1990-08-31

5 5 Spring 1991 18361 1991-02-01

6 6 Summer 1991 6702 1991-06-01

> tail(osu.enroll)

 t Semester Year Enrollment date

35 35 Spring 2001 20004 2001-02-01

36 36 Summer 2001 7558 2001-06-01

37 37 Fall 2001 21872 2001-08-31

38 38 Spring 2002 20922 2002-02-01

39 39 Summer 2002 7868 2002-06-01

40 40 Fall 2002 22992 2002-08-31

> #Suppose it was early spring 2002 and you wanted to

 forecast fall 2002 so use data only up to spring 2002

> x <- osu.enroll$Enrollment[1:38]

> #ACF and PACF of x\_t

> par(mfcol = c(1,2))

> acf(x = x, type = "correlation", lag.max = 20, ylim =

 c(-1,1), main = expression(paste("Estimated ACF plot

 for ", x[t])), xlim = c(1,20))

> pacf(x = x, lag.max = 20, ylim = c(-1,1), xlab = "h",

 main = expression(paste("Estimated PACF plot for ",

 x[t])))



The autocorrelations are very slowly going to 0 for lag 3, 6, 9, … . Therefore, first differences should be examined for s = 3.

Suppose instead an ARIMA(0,0,0)x(1,0,0)3 model was fit to the data. This may happen if the ACF plot was interpreted as tailing off to 0 and noticing the significant lag 3 partial autocorrelation.

> mod.fit.000.100 <- arima(x = x, order = c(0, 0, 0),

 seasonal = list(order = c(1, 0, 0), period = 3))

Error in arima(x = x, order = c(0, 0, 0), seasonal = list(order = c(1, :

 non-stationary seasonal AR part from CSS

R gives an error message saying the model cannot be fit due to non-stationarity. For s = 3, first differences, (1-B3), could be interpreted as having a 1 coefficient on B3.

Below are the ACF and PACF plots after taking the first differences with s = 3.

> #ACF and PACF of (1-B^3)\*x\_t

> acf(x = diff(x = x, lag = 3, differences = 1), type =

 "correlation", lag.max = 20, ylim = c(-1,1), main =

 expression(paste("Est. ACF for ", (1-B^3)\*x[t], "

 data")), xlim = c(1,20))

> pacf(x = diff(x = x, lag = 3, differences = 1), lag.max

 = 20, ylim = c(-1,1), xlab = "h", main =

 expression(paste("Est. PACF for ", (1-B^3)\*x[t],

 " data")))



For the non-seasonal part, the ACF is tailing off to 0 and the PACF has a significant value at lag = 1. For the seasonal part, it looks like there may be a significant autocorrelation at lag 9. Possibly due to the small season length (s = 3), it may be hiding other significant values. Because I am not for sure, I decided to add an AR(1) term to the model for the non-seasonal part. Below is part of the output and code.

> mod.fit.100.010 <- arima(x = x, order = c(1, 0, 0),

 seasonal = list(order = c(0, 1, 0), period = 3))

> mod.fit.100.010

Call:

arima(x = x, order = c(1, 0, 0), seasonal = list(order = c(0, 1, 0), period = 3))

Coefficients:

 ar1

 0.6989

s.e. 0.1278

sigma^2 estimated as 130905: log likelihood = -256.19, aic = 516.37

> #Need to run function in examine.mod.R file first

> examine.mod(mod.fit.obj = mod.fit.100.010, mod.name =

 "ARIMA(1,0,0)x(0,1,0)\_3", max.lag = 20)

$z

 ar1

5.468627

$p.value

 ar1

4.535362e-08



Notes:

* Does the model make sense for this problem?
* The test of Ho:ϕ1=0, Ha:ϕ1≠0 produces a test statistic of 0.6989/0.1278 = 5.47. Therefore, ϕ1≠0.
* The residual ACF and PACF plots do not show any significantly different from 0 values.
* The Ljung-Box-Pierce test also results in all “don’t reject Ho’s” for its test of non-zero autocorrelation.
* There are no extreme standardized residuals.
* There is some evidence against normality in the histogram and QQ-Plot. However, because the sample size is fairly small, I am not surprised by this and decided not to investigate transformations.
* Because everything is o.k. with this model, one could use the final estimated model of an ARIMA(1,0,0)×(0,1,0)3:

(1-0.6989B)(1-B3)xt = wt with  = 130,905

I still have some concerns about whether or not this is the best model that I could use. Below are some items that I tried first starting with just the (1-B3)xt series.

1. Remember seeing the marginally large autocorrelation at lag 9 for the (1-B3)xt series? Given s = 3, I decided to investigate adding seasonal MA terms to a model. The ARIMA(0,0,0)×(0,1,1)3 model results in a significant Θ1 and the residual ACF and PACF plots indicate either ϕ1 or θ1 should be added to the model. When ϕ1 is added to the model, Θ1 has a p-value of 0.11 and the AIC is 516 which is a little smaller than the 516.37 for the ARIMA(1,0,0)×(0,1,0)3 model. When θ1 is added, Θ1 is no longer significant (p-value = 0.25) and there are marginally significant groups of autocorrelations as found by the Ljung-Box-Pierce test.
2. Due to the significant autocorrelation at lag 1, I decided to start with a θ1. The term is significant, but the Ljung-Box-Pierce test results in smaller p-values (some marginally significant?) than for the ARIMA(1,0,0)×(0,1,0)3 model. The ARIMA(0,0,1)×(0,1,0)3 model’s AIC is 519.2 which is larger than the 516.37 for the ARIMA(1,0,0)×(0,1,0)3 model. When both θ1 and ϕ1 are in the model, θ1 is no longer significant.
3. When Φ1 is in the model alone, it is marginally significant (p-value = 0.086) and the residual ACF and PACF’s indicate changes need to be made to the non-seasonal part of the model. When ϕ1 is added, Φ1 has a p-value of 0.15 and ϕ1 is highly significant. The AIC is 516.35.
4. Below is a plot of (1-B3)xt:



This could indicate nonstationarity still exists and that differencing is needed. If one examined (1-B)(1-B3)xt, the following results





While there is no longer the small upward trend in the plot of the data over time, the estimated ACF and PACF have no significant values. There does appear to be too many negative partial autocorrelations indicating there is some dependence left in the series. Also, the Ljung-Box-Pierce test indicates some marginally significant tests. I tried a variety of ARIMA models with the data, but none produced significant AR or MA parameters. Also, I am a little concerned about doing the (1-B) difference considering the ARIMA(1,0,0)×(0,1,0)3 estimates ϕ1 to be 0.6989. Using the standard error, a 95% C.I. for ϕ1 is 0.6989±1.96×0.1278 = (0.4484, 0.9494). So, there is not strong evidence to indicate ϕ1 = 1. I did try to add additional terms to the model, but ϕ1 and θ1 are non-significant. Also, Θ1 has a p-value of 0.11 and Φ1 has a p-value of 0.15.

Below is the code and output for the forecasts using the ARIMA(1,0,0)×(0,1,0)3 model.

> #Forecasts 6 time periods into the future

> fore.mod <- predict(object = mod.fit.100.010, n.ahead =

 6, se.fit = TRUE)

> fore.mod

$pred

Time Series:

Start = 39

End = 44

Frequency = 1

[1] 8199.614 22320.440 21235.426 8418.675 22473.548

 21342.437

$se

Time Series:

Start = 39

End = 44

Frequency = 1

[1] 361.8086 441.4208 475.4894 679.4434 759.4145 795.5628

> pred.mod <- x - mod.fit.100.010$residuals

> #Calculate 95% C.I.s

> low <- fore.mod$pred - qnorm(p = 0.975, mean = 0, sd =

 1) \* fore.mod$se

> up <- fore.mod$pred + qnorm(p = 0.975, mean = 0, sd =

 1) \* fore.mod$se

> data.frame(low, up)

 low up

1 7490.482 8908.745

2 21455.271 23185.609

3 20303.484 22167.368

4 7086.990 9750.360

5 20985.123 23961.973

6 19783.162 22901.711

Below is a plot of the forecasts (see program for code).



Note that time = 40 is Fall 2002. Therefore, the forecasted enrollment for the Fall 2002 semester is 22,320. The actual enrollment was 22,992. The 95% C.I. is 21,455 ≤ x40 ≤ 23,185. Perhaps this interval is too wide to be helpful. Alternatively, perhaps OSU could use the lower bound in its enrollment projections to be conservative.

Below is a table comparing the forecasts by OPBIR and my SARIMA model. The forecasts from my model are done using an ARIMA(1,0,0)×(0,1,0)3, but parameter estimates are found using data only up to the spring previous to the fall enrollment.

|  |  |  |
| --- | --- | --- |
|  | **OPBIR** |  |
| **Semester** | **Actual** | **Forecast** | **Error** |  |
| **Fall 1998** | 20,466 | 19,490 | 976 |  |
| **Fall 1999** | 21,087 | 21,072 | 15 |  |
| **Fall 2000** | 21,252 | 21,995 | -743 |  |
| **Fall 2001** | 21,872 | 21,918 | -46 |  |
| **Fall 2002** | 22,992 | 22,377 | 615 |  |
|  |
|  | **SARIMA model** |
| **Semester** | **Actual** | **Forecast** | **Error** | **95% C. I.** |
| **Fall 1998** | 20,466 | 19,478.77 | 987 | (18,741, 20,216) |
| **Fall 1999** | 21,087 | 21,083.52 | 3 | (20,205, 21,962) |
| **Fall 2000** | 21,252 | 21,256.57 | -5 | (20,411, 22,102) |
| **Fall 2001** | 21,872 | 21,320.11 | 552 | (20,485, 22,155) |
| **Fall 2002** | 22,992 | 22,320.44 | 672 | (21,455, 23,185) |

Notice how good the model did in fall 2000 ☺, which was the year discussed in the O’Colly article.

How do you measure which forecasting method is more accurate overall? One way is to look at the mean square error.

|  | **MSE** |
| --- | --- |
| OPBIR | 377,038 |
| SARIMA model | 346,046 |