Modelling Politician Approval Ratings

Group: Yash Bhate, Aditya Kotak, Hubert Luo, Aniket Mandalik, Vinit Parikh

November 22, 2019

1 Executive Summary

Using a dataset of approval ratings for a politician, the goal was to forecast their approval ratings for the next 10 days. In our analysis, we used both first order differencing and Local Polynomial Regression (LOESS) as tools to achieve stationarity and eventually settled on LOESS to forecast the next 10 days worth of approval ratings.

2 Exploratory Data Analysis

We start by first visualizing our data:

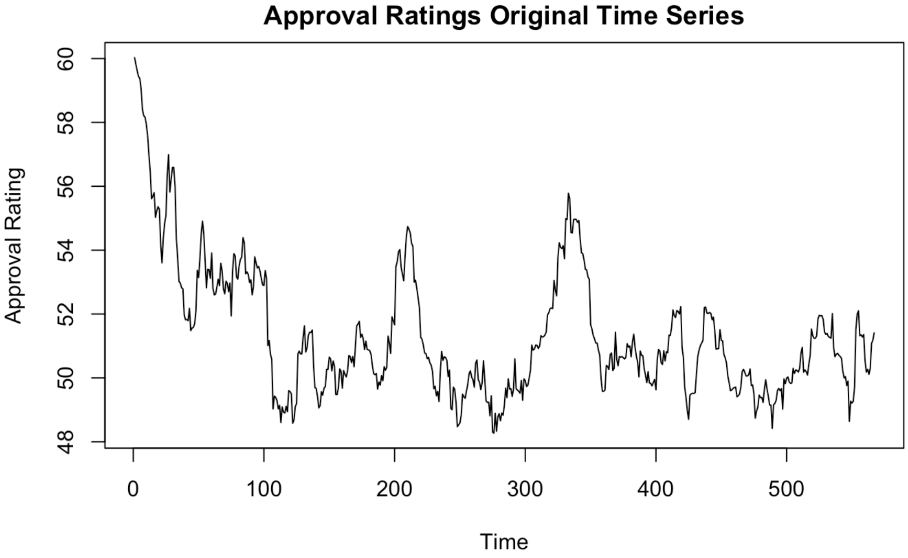


Figure 1: Raw Approval Ratings Data

The data does not look like stationary white noise, and there seem to be peaks indicating some sort of cyclical trend in the data. This could imply that differencing might be helpful to get rid of these patterns.

The variance appears to be constant over time which implies homoscedasticity of the data, so we do not need variance stabilizing transformations.

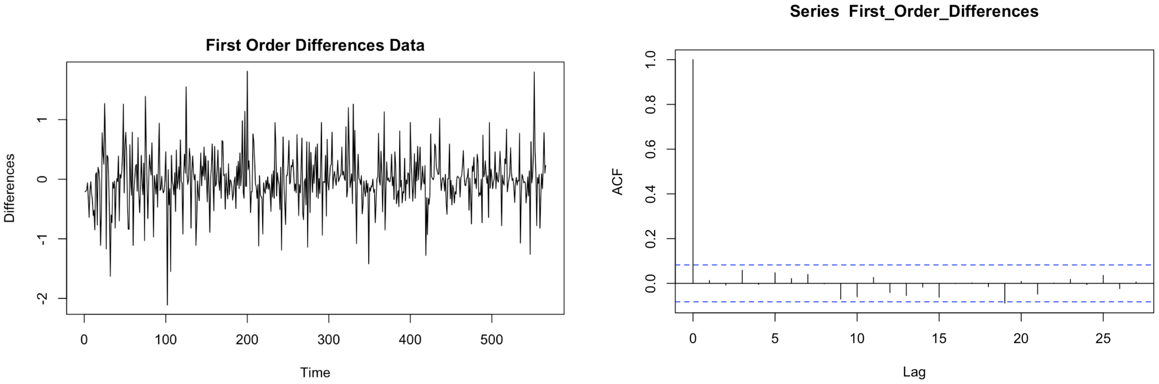
With that in mind, it could be useful to look for general trends in the data. From the data, we can see that there is a slight downward trend. This means that it could be useful to use some sort of regression model to "de-trend" the data in pursuit of stationarity.

3 Modeling a Deterministic Function of Time

Using EDA to inform our decision, we have two main approaches to pursuing stationarity. First is **Approach 1: First Order Differences** and second is **Approach 2: Local Polynomial Regressions (LOESS)**

3.1 First Order Differences

We decided to take a first order difference to see if we can get rid of some weak cyclic pattern. This is what the first order differences and the corresponding ACF plot ended up looking like:



(a) First Order Differences Raw Data

(b) ACF of the First Order Differences

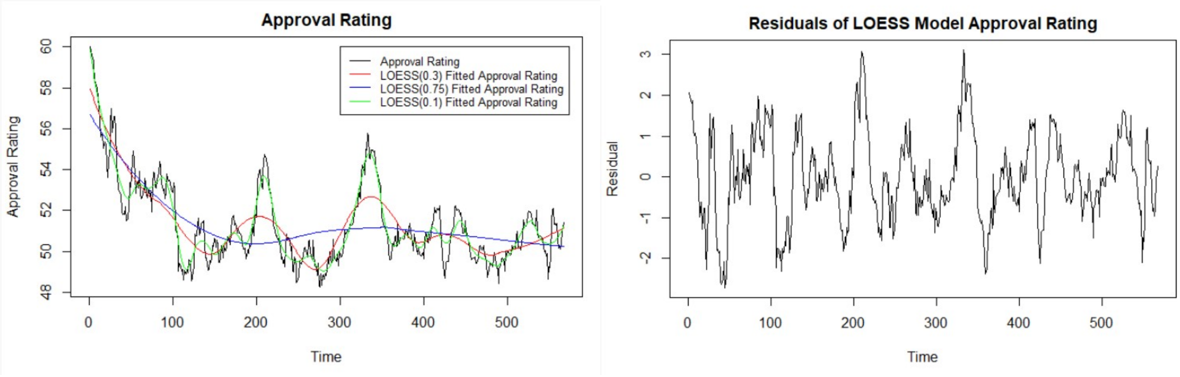
Figure 2: First Order Differences and ACF

From this we can see the data looks like white noise since there are no significant spikes in the ACF.

3.2 Local Polynomial Regression (LOESS)

Our justification for using LOESS instead of traditional Linear/Quadratic model was that since the data being modeled is based on a politician's approval rating, it is likely affected much more heavily by what happened in the last few days rather than what had happened hundreds of days ago. The inspiration for this decision was from fivethirtyeight.com's President Trump approval rating model. As a result, we felt that a local estimate of values based on only a small window of past approval ratings may be a more accurate representation of the trends observed in order to "de-trend" the data.

We had to select the window to look back, which is the span parameter of LOESS. For our model, we selected span=(0.3) as it was large enough to avoid overfitting on an extremely small time period, but small enough to accurately reflect trends in the data in a local time period. We used cross-validation to justify this decision.[1] Figure 3 below demonstrate the fit of the local polynomial regression (with various span parameters) over the raw approval rating data as well as the residuals of the model with parameter 0.3:



(a) Comparing the fit of the LOESS Models

(b) Residuals of the LOESS(0.3) Model

Figure 3: LOESS Model and Reisduals

Although there are spikes in the residuals, it appears to be neither occurring at a consistent-enough interval nor similar enough in magnitude to conclude any seasonality. The residuals appear to be roughly homoskedastic as well, mitigating the need for a variance-transforming function. Therefore, the residuals

[1] We made a visual assessment regarding the fit of the various LOESS model based on the cross-validation.

of the LOESS model appear to be roughly stationary and can be modeled using an ARIMA model in section 3, thus achieving stationary.

4 ARIMA Model Selection

4.1 First Order Difference w/ ARIMA(0,0,0)

As we saw in part 2, the first order differences were a great fit because the differences ended up looking like white noise. When we diagnose this model, this is what we observe:

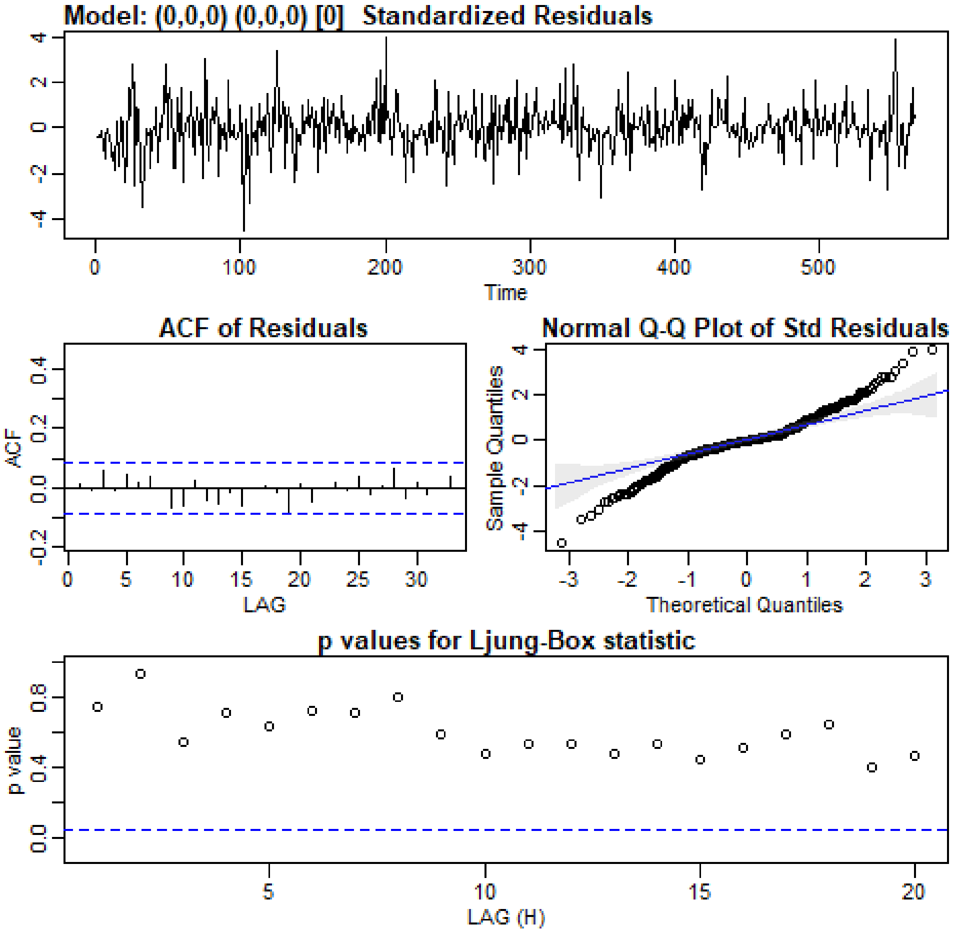
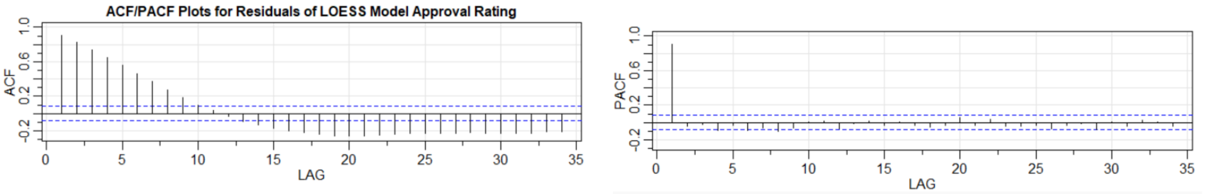


Figure 4: Diagnostics for First Order Differenced Model

Since we already established that the data looks like white noise, it is appropriate to fit ARIMA(0,0,0). We can see from the Ljung-Box plots that the p-values are insignificant and the ACF of residuals does not appear outside the blue bands therefore this model fits well.

4.2 LOESS w/ ARIMA(1,0,0)

Figure 5 below shows the ACF and PACF plots for the residuals of the LOESS model we used in part 2. These plots below demonstrate a roughly consistent decrease in the ACF as the lag increases and the PACF has a spike at 1, with all other values falling within the blue bars (insignificant). As a result, this points to the first of two potential candidate ARIMA models for the noise: ARIMA(1,0,0)



(a) ACF OF LOESS Residuals

(b) PACF of the LOESS Residuals

Figure 5: ACF and PACF of the LOESS Residuals

After fitting an ARIMA(1,0,0) to the residuals, we diagnosed this model with the help of the following plots:

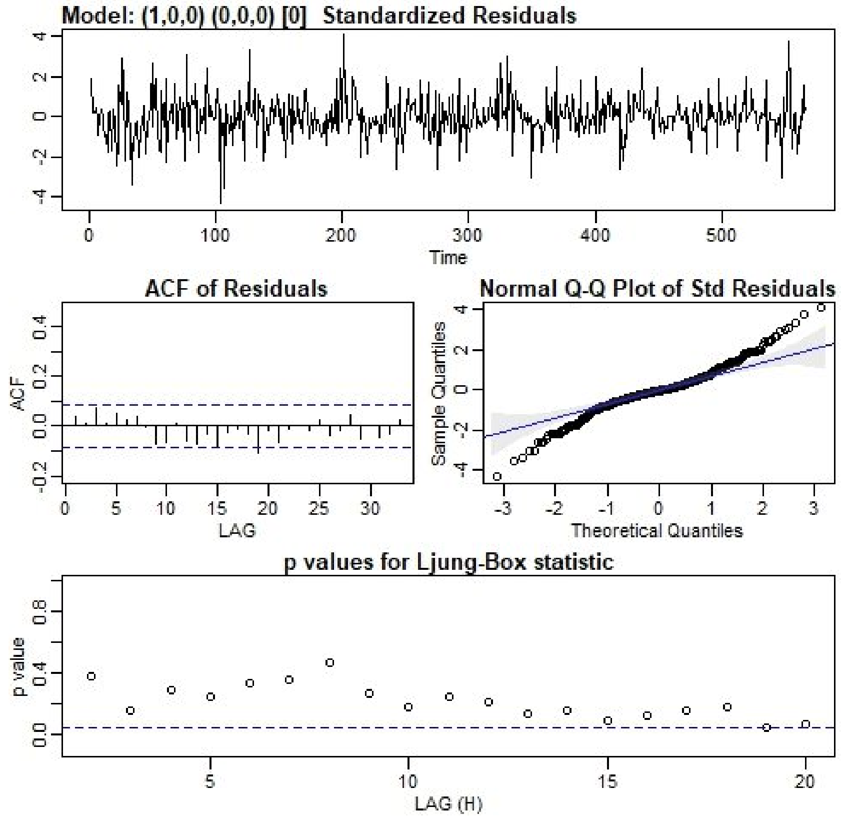


Figure 6: Diagnostics for LOESS w/ ARIMA(1,0,0)

These plots demonstrate that the standardized residuals appear to be roughly stationary, the ACF of the residuals of this ARIMA model has most values within the blue bands or at worse slightly outside. The QQ plot demonstrates the residuals are not exactly normally distributed and are in fact more closely concentrated around the mean of zero. The Ljung-Box test shows p-values which are mostly not significant but still relatively small, indicating the model is a decent, but not great, fit for the residuals observed in the LOESS model.

4.3 LOESS w/ ARIMA(0,1,0)

The second ARIMA model for the LOESS residuals to be examined is the ARIMA(0,1,0) model, the first differences of the LOESS residuals. Figure 7 below is the diagnostics of this model and they demonstrate that the standardized residuals also look roughly stationary and the ACF of the residuals of this ARIMA model again have values of fairly small magnitude, with most again falling within the blue bands. In addition, the QQ plot again demonstrates these residuals are light-tailed, with most of the residuals falling around the mean of zero relative to the normal distribution. The biggest difference comes in the values of the Ljung-Box p-values as they are much higher than those for the ARIMA(1,0,0) model across the board. The p-values are highest for small lags, before decreasing as the lag increases. Therefore, the ARIMA(0,1,0) model is a good fit for the residuals observed in the LOESS data.

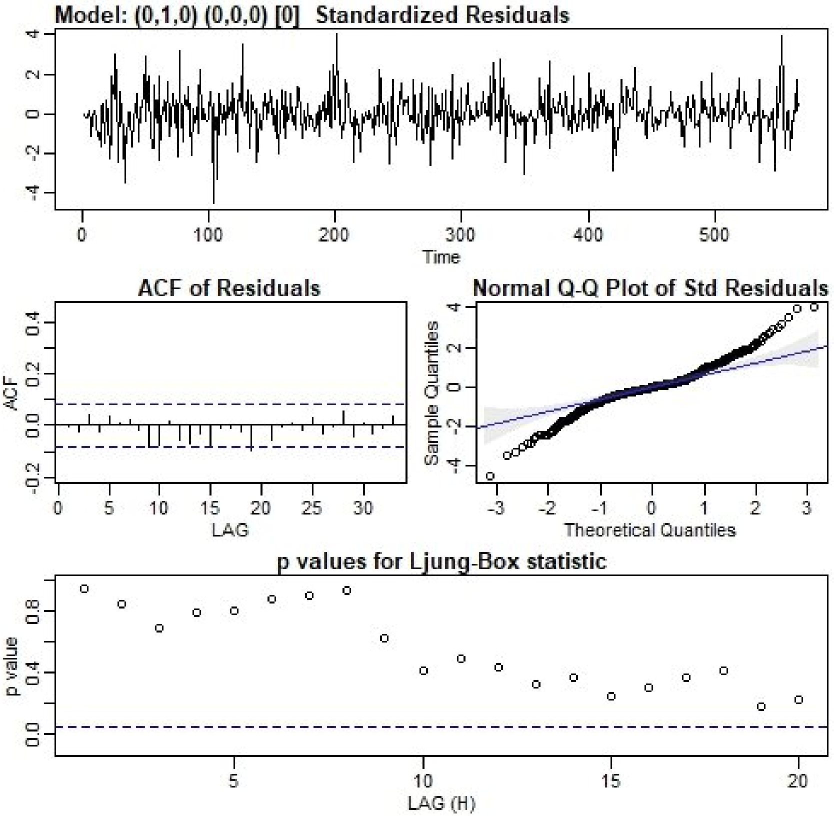


Figure 7: Diagnostics for ARIMA(0,1,0) model on the LOESS residuals

4.4 Model Selection

We picked LOESS w/ ARIMA(1,0,0). This was because the AIC, BIC and AICc for this model were marginally better than the other two. As a comparsion, the BIC for the 3 models are listed below

Model	BIC
First Order Difference with ARIMA(0,0,0)	1.29
LOESS with ARIMA(1,0,0)	1.24
LOESS with ARIMA(0,1,0)	1.27

Table 1: These are the Bayesian Information Criterion for our models of interest

You can see that the model we selected is the best!.

5 Results

The model we selected used the LOESS with a span parameter of 0.3 to account for the trend and a ARIMA(1,0,0) model for the residuals. Our ARIMA model for the residuals of the LOESS regression is defined in equation (1).

$$(1 - \phi_1 B)X_t = Z_t \tag{1}$$

5.1 Estimation of model parameters

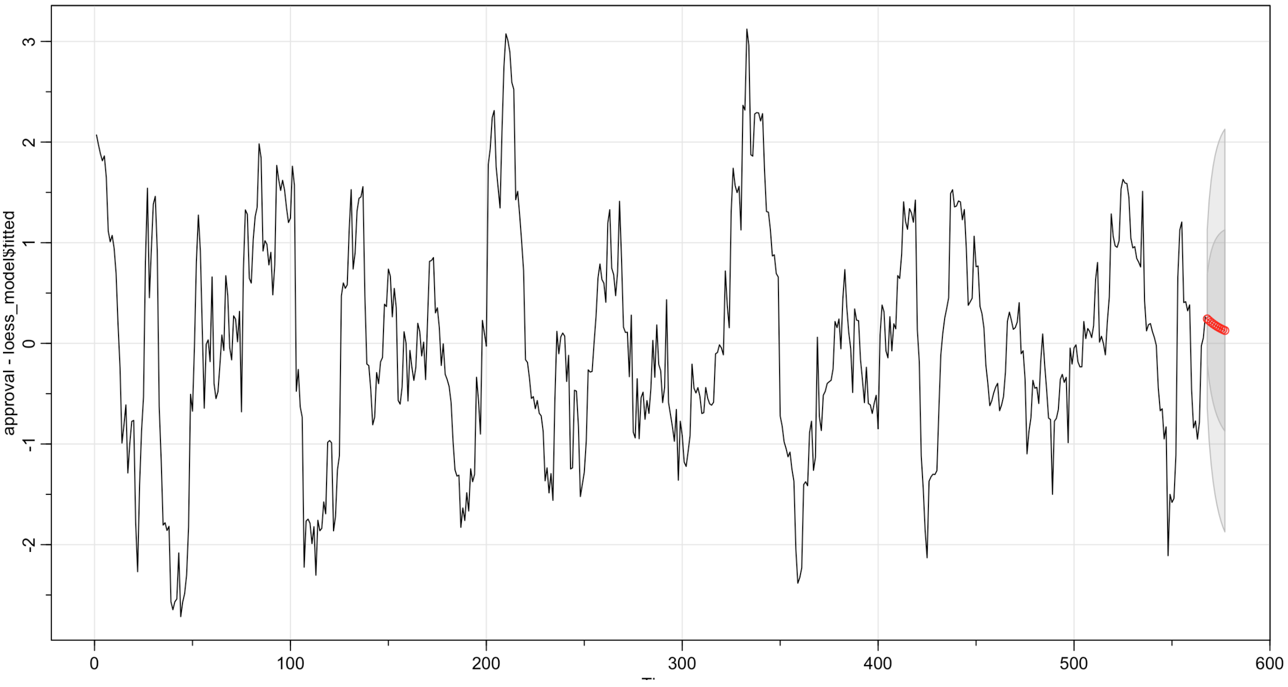
Given that in (Table 2). Parameters in the LOESS model and ARIMA(1,0,0)

Parameter	Estimate (s.e)
degree	2
span	0.3
ϕ_1	0.915 (0.016)

Table 2: These are our parameter estimates and corresponding standard errors the LOESS and the ARIMA model in equation 1.

5.2 Prediction

Find below the time series of the entire data with the next 10 predictions along with a band defined due to the presence of ARIMA. The shades of gray indicate degrees of uncertainty and values closer to the red are darker



The predicted values for the next 10 periods are: 51.16851, 51.18979, 51.21108, 51.23237, 51.25365, 51.27493, 51.29620 51.31746, 51.33871 51.35993

We now must cross-validate our model to ensure we pick the right one for our forecasts. We first decided to visually alter the spans in our loess model and plot our results above. As shown previously, higher spans (0.5) tend to underfit the data and lower spans (0.1) overfit the data. This led us to visually believe that somewhere between 0.2 to 0.4 is probably a good span value to use but we then tested this by minimizing the SSE and then trying to optimize for the span. (See Appendix for the code.) We found that the optimal span is indeed 0.3 and used that in our models. We also see that for our ARIMA(1,0,0), the Ljung Box statistics also show us that the p-values for all values are in a good range, validating our ideas about this being a good model for the data. We then computed all of the Information Criterion in order to see which model minimized the AIC, AICC, and BIC. This led us to using the ARIMA(1,0,0) on the LOESS model with a span of 0.3 as all the Information Criteria were indeed lower than the other models. Our forecasts show that the incumbent mayor should most likely remain above 51 percentage points and thus win the local election in 10 days.

6 Appendix

6.1 LOESS and Forecasting Code

```
library(astsa)
library(forecast)
#library(msir)
elections_data = read.csv('politics.csv')[c('Date','Approval')]
approval = elections_data[['Approval']]
n = length(approval)
time = 1:n

plot(approval, type='l', main='Approval Rating', xlab='Time', ylab='Approval Rating')
#model1 = sarima(approval,p=0,d=1,q=0,S=0,P=0,D=0,Q=0)
X_t = diff(approval)
plot(X_t, type='l', main='First Differences', xlab="Time", ylab="Approval Rating")
#this is the first order difference with ARIMA(0,0,0)
model11 = sarima(X_t,p=0,d=0,q=0,S=0,P=0,D=0,Q=0)
sarima.for(approval, n.ahead=10,p=0,d=1,q=0,S=0,P=0,D=0,Q=0)

loess_model = loess(approval~time, degree=2, span=0.3,
    control = loess.control(surface ="direct"))
resid3 = approval-loess_model$fitted
plot(approval, type='l', main='Approval Rating', xlab='Time', ylab='Approval Rating')
lines(time, loess_model$fitted, col="red")
legend(300, 60, legend=c("Approval Rating","LOESS Fitted Approval Rating"),
    col=c("black", "red","blue"),lty=rep(1,2), cex=0.8)

plot(resid3, type='l', main='Residuals of LOESS Model Approval Rating',
    xlab='Time', ylab='Residual')
acf2(resid3, main='ACF/PACF Plots for Residuals of LOESS Model Approval Rating')

pred_time = (length(time)+1):(length(time)+10)
pred = predict(loess_model, newdata=pred_time)
plot(approval, type='l')
lines(pred_time, pred, col='red')
```

6.2 First Order Difference Code

```
df <- read.csv(file="politics.csv", header=TRUE, sep=",")
First_Order_Differences = diff(df$Approval)
#sarima(First_Order_Differences, p=0,d=0,q=0,S=0,P=0,D=0,Q=0)
plot(df$Approval, main='Approval Ratings Original Time Series',
    xlab='Time', ylab='Approval Rating', type='l')

plot(First_Order_Differences, type='l', main='First Order Differences Data',
    ylab='Differences', xlab='Time')
acf(First_Order_Differences)
pacf(First_Order_Differences)
```