Yuwei Chen (999070073)
Ming Yang (998942775)
Section: STA 137 A01

STA 137 Project: Monthly employment figures for the motion picture industry

March 8, 2016

Abstract

This report uses the time series additive decomposition model $Y_t = m_t + s_t + X_t$ to forecast and analyze an univariate time series. The data description section will provide a summary of the data. In our "Deterministic components" and "Timer series model" data analysis we will display and discuss the procedures taken to assess the univariate time series. The "Forecasting" section, we use the deterministic components and time series model found in the previous parts to forcast the next ten time points. And the "conclusion" section will discuss and summarize our finding from our analysis. Lastly, we will place R code in the Appendix.

Description of the data

We use time series dataset called "motion", which can be found in the "fma" library. This dataset contains Monthly employment figures for the motion picture industry from Jan 1955 to Dec 1970. By looking at the salary data a person can gain insight and knowledge into the behavior of salary changes. This knowledge would be valuable to anyone who interested in working in motion picture industry. This dataset was drawn from the source: "Makridakis, Wheelwright and Hyndman (1998) Forecasting: methods and applications, John Wiley & Sons: New York. Exercise 7.9"

Here is the plot for the original data:

By observing the dataset as a whole, we can fit a quadratic trend on the dataset from Jan 1955 to Dec 1970. The monthly overall salary decreases roughly from 1955 to 1963, then increases afterwards since 1963 to 1970 in a relatively slower rate. Which means the salary for motion picture industries had declined from 1955 to 1963, and then the trend changed and salaries increase again from 1963 to 1970. There are obvious seasonal flucuations on the salary throughout the duration we observed in our dataset in every 12 months. By observing the dataset, we don't find any "off-trend" points/outliers. The fluctuation amplitude is consistent as time increases, so we conclude that transformation is unnessassary for this time series data.

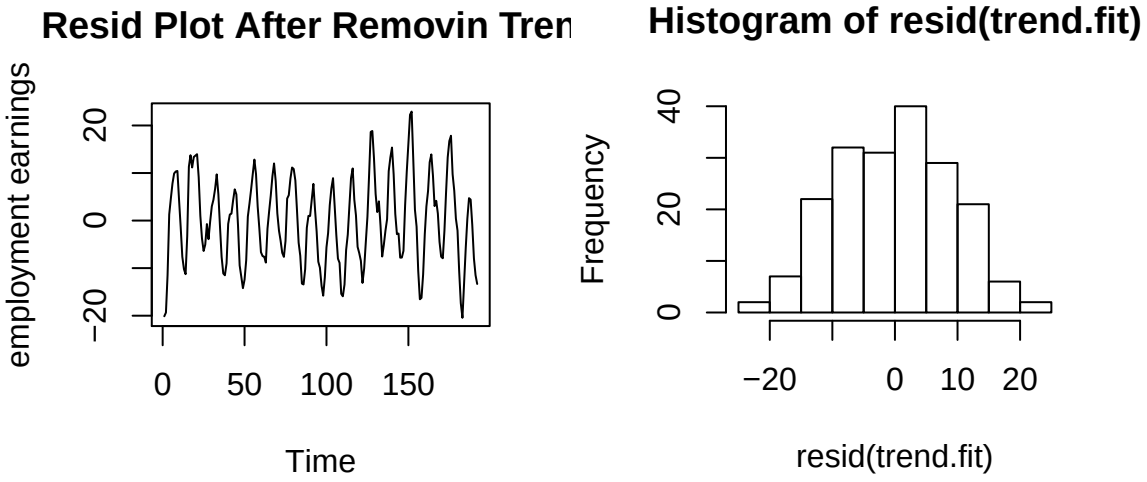
Deterministic components

By observing the data, we can see that there are two obvious componentss in the dataset, one is trend component and the other one is seaonsal component. For overall trend, we can fit a quadratic trend on the dataset from Jan 1955 to Dec 1970. In other words, we can fit a polynomial of order 2 based on our observation on the plot. After the trend component is removed, we can recognize an apparent seasonality in the residuals, with period d approximately 12 months, which is a year. For seasonal component, we can fit a sum of harmonics, which contains 6 sets of harmonics.

By observing the data, the underlying trend is visible, it looks most approprate to fit a quadratic model to the data. **So first of all, let's fit a polynomial of order 2:**

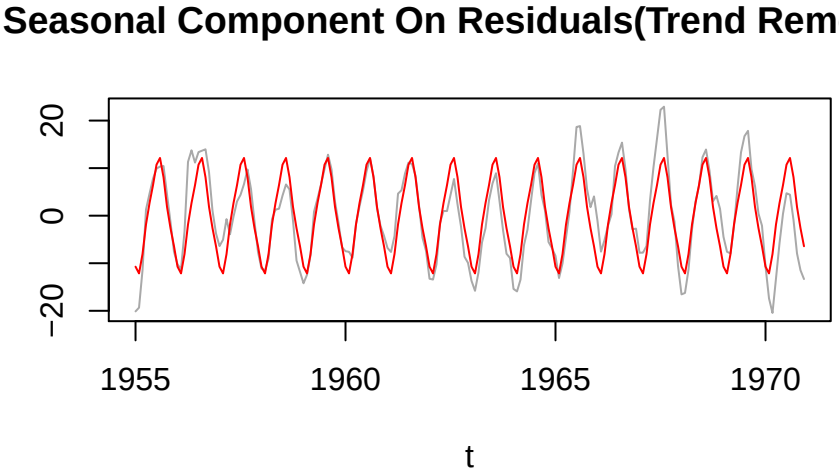
Since the r-squared adjusted value is .7341, about 73.41% of the variation in the data can be explained by the quadratic model. To further assess whether the polynomial model is a good fit, a diagnostics of the residuals can detect possible problems.

Then, let's plot the residuals after trend being removed:



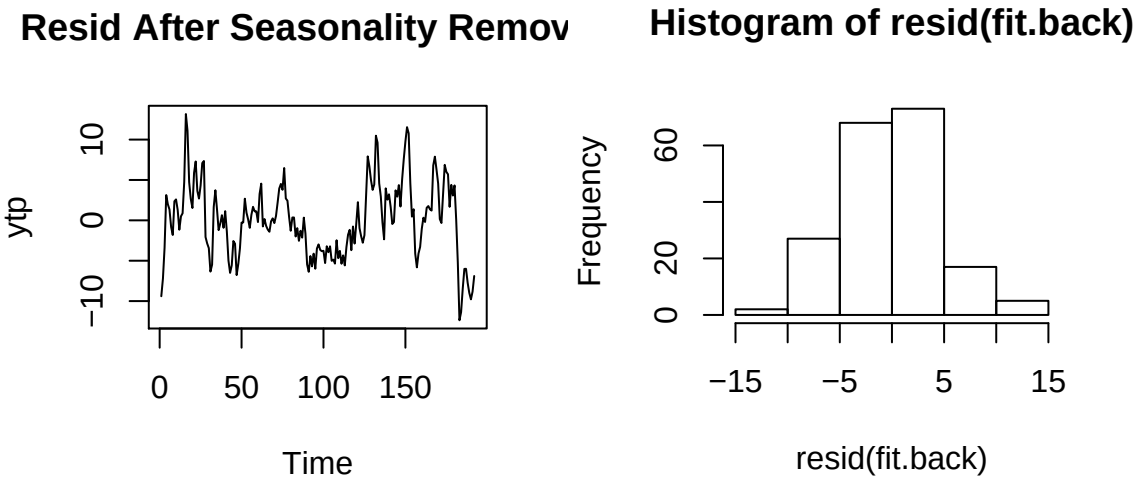
By observing the residuals plot after removing the underlying trend, we can't find any obvious trend in the plot. The plot above on the right is a histogram plot for the residuals after overall trend is removed. In the presentation of the histogram of the residuals, which appear fairly normally distributed. And the mean is roughly at 0.

After the underlying trend is removed, it looks most appropriate to fit a sum of harmonic model. **So secondly, let's remove the seasonality component by fitting a sum of harmonic:**



Since the r-squared adjusted value is .8699, about 86.99% of the variation in the data can be explained by the sum of harmonic fit. To further assess whether the sum of harmonic model is a good fit, a diagnostics of the residuals can detect possible problems.

Then, let's plot the residuals after seasonal trend being removed:



By observing the residuals plot after removing the seasonality trend, all the plots are fluctuate around 0. The plot above on the right is a histogram plot for the residuals after seasonal trend is removed. In the presentation of the histogram of the residuals, which appear roughly normally distributed. And the mean is roughly at 0.

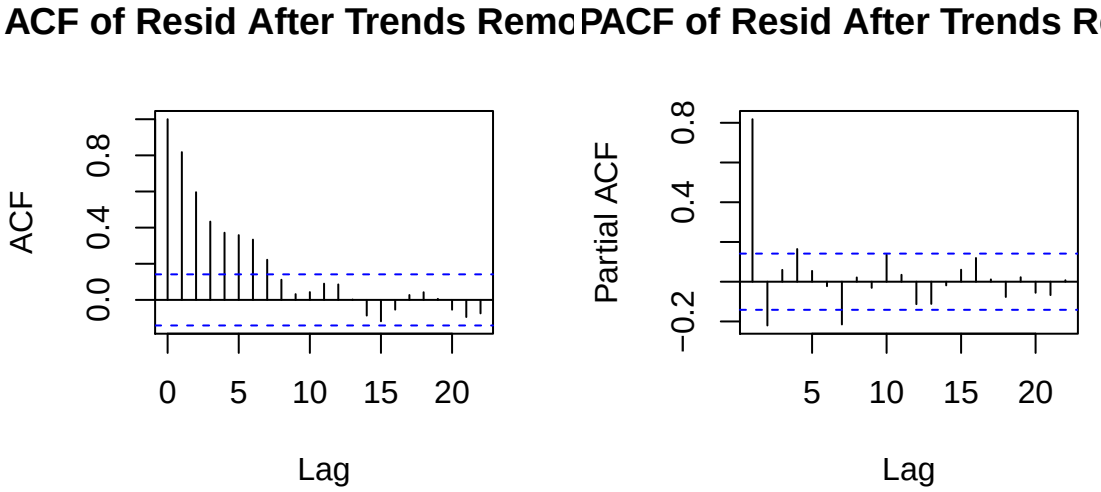
Time series model

After de-seasonalizing and de-trending the original data, we are able to retrieve residuals. Before fitting an ARMA model, the residuals must resemble a stationary process. From the figure above, the residuals look stationary. But in order to make sure that our data is stationary, we decide to generate a KPSS test.

Let's first do a KPSS test and see whether if we need differencing the data:

```
##
##  KPSS Test for Level Stationarity
##
## data:  ytp
## KPSS Level = 0.15853, Truncation lag parameter = 3, p-value = 0.1
```

According to the output, we can see that p-value is 0.1, which is greater than 0.05. And large p-value in the KPSS test indicates stationarity.



The ACF plot drops to zero and the data resembles a stationary process. Some of the values in the ACF and PACF plot are not within the 95% limits, which indicates an ARMA model may best fit the data.

Let's use AIC criteria to find the best model:

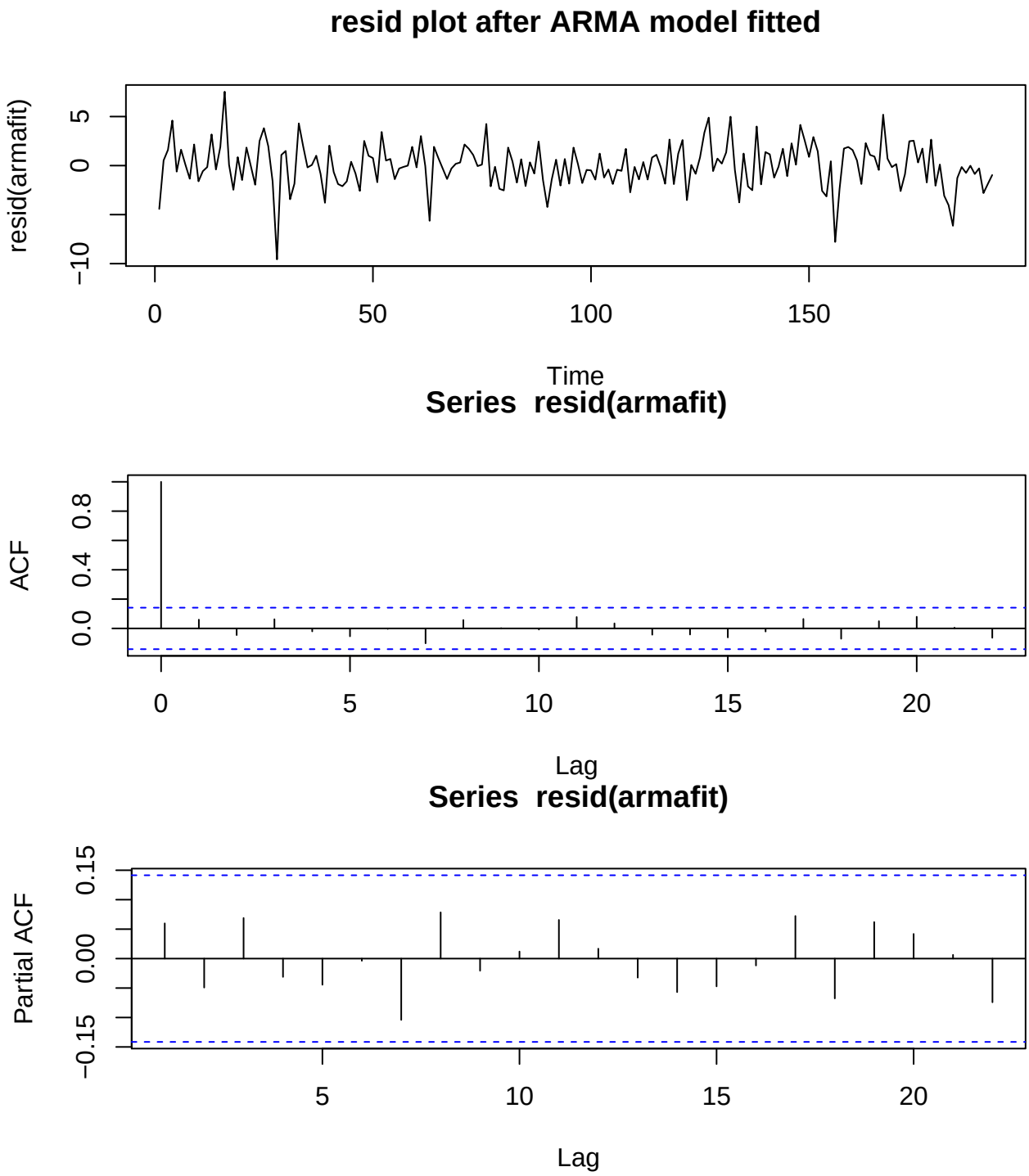
```
## Series: ytp
## ARIMA(4,0,3) with zero mean
##
## Coefficients:
##          ar1      ar2      ar3     ar4      ma1     ma2     ma3
##       0.9475  -0.2527  -0.6430  0.6744  -0.0331  0.1987  0.7175
## s.e.  0.2339   0.4223   0.4212  0.2042   0.2012  0.1812  0.1859
##
## sigma^2 estimated as 5.178:  log likelihood=-430.3
## AIC=877.83   AICc=878.61   BIC=903.89
```

According to the output, we can see that by using 'auto.arima' demand, we found that ARIMA (4,0,3) is the best model among all models with the max.order of 8, which it has the smallest AIC value. Here is the expression for the best fitted model - general:

$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}$

plug in the values:

$X_t + 0.9475 X_{t-1} - 0.2527 X_{t-2} - 0.6430 X_{t-3} + 0.6744 X_{t-4} = Z_t - 0.0331 Z_{t-1} + 0.1987 Z_{t-2} + 0.7175 Z_{t-3}$



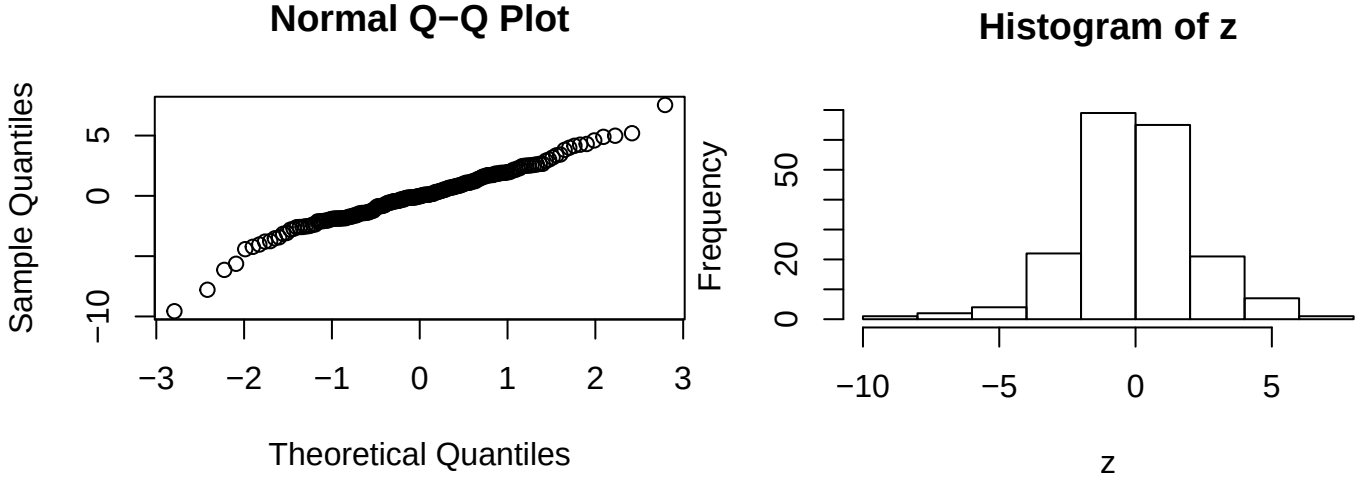

```
##
##  Box-Ljung test
##
## data:  armafit$residuals
## X-squared = 13.47, df = 24, p-value = 0.9577
```

In the plots above, there are three plots including the residual plot after the ARMA model is fitted. By observing the ACF and PACF of the residuals after the time series model is fitted, the remaining noise appears

to be white noise. Because all values are within 95% limits. So we can say that there isn't any dependence structure remaining in these residuals. Also by looking at the values we generated from Ljung-Box test, we can see that the p-value is 0.9582, which is greater than 0.05 (significant), then we reject the null hypothesis, and we can conclude that the noise is independent.

Forecasting

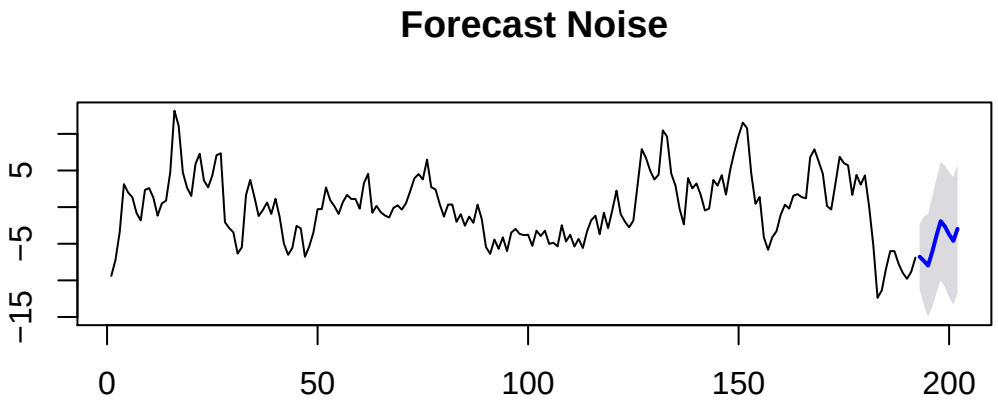
In this section, our goal is to forecast the next ten time points by using the models we found in the previous sections. Here is our original plan: In order to get promising forecasting points, we planned to take out the last year's data, which is the 1970 data. We firstly want to forecast the 1970 data and compare with the actual data from 1970 and see if the actual data fall in our forecasting range (confidence interval), then we will use the models to forecast the next ten time points. We leave out a year/a cycle from the beginning of the analysis. Then make a prediction for the last year data - testing data (1970), based on the model fitted for the first 15 years. So that we can test whether if our model is reasonable for predicting future points by comparing our predicted values to the actual values. But later we found that it's hard to keep the local variable and global variable the same for the whole time, so we didn't take out 1970 data in our final report.



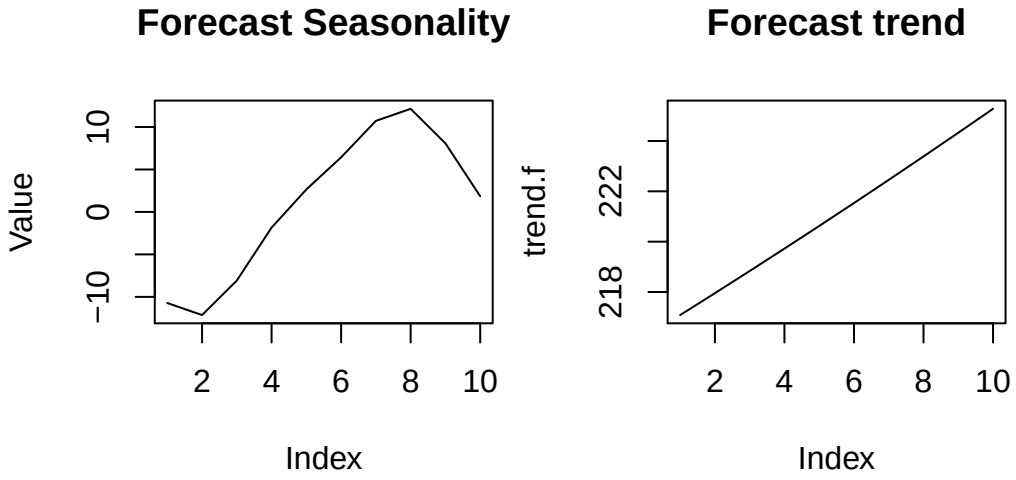

```
##
##  Shapiro-Wilk normality test
##
## data:  z
## W = 0.97489, p-value = 0.001583
```

Let's look at the qqplot and histogram of the residuals to see if normality can be assumed But since we have a small p-value for the Shapiro-Wilk test, so we reject the hypothesis that the residuals are normal. Since we are rejecting normaility, the forecasts intervals may not be correct.

Let's make the forecast for noise, seasonality and the overall trend:

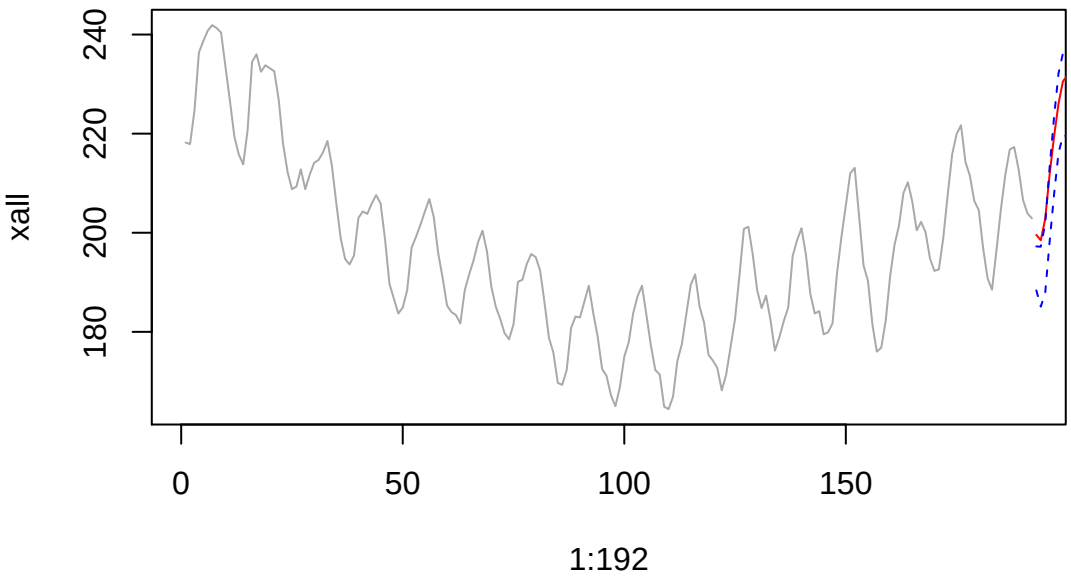


Above plot represents the forecasting for the noise based on our ARIMA(4,0,3) model. It shows that the trend of the noise is converging towards zero, which is a reasonable prediction.



The left plot above shows values of prediction for seasonality, and it shows a going down then going up then going down trend, which matches the sum of harmonic model we have fitted for the seasonality component. The right plot shows values of prediction for trend component, and it shows a increasing trend, which mathces the second half of our polynomial of order 2 model. According to the plots, we can say those predictions are reasonable for future forecasting.

Then we combine all the components and forecast the next ten time points:



In the plot above, lightgrey line represents the original time series data, two dashed blue lines show the lower bond and upper bond of our confidence intervel with level .95; the red solid line shows our prediction for the future ten time points based on our fitted model.

Conclusions

An analysis of this type would allows an individual to predict the salary or a motion picture company to predict the figures of their employees. Also for motion picture investors, they can use this evidence to see how much money they should invest on motion picture industries and what they should expect from the motion picture job demands.

Reference

· Makridakis, Wheelwright and Hyndman (1998) Forecasting: methods and applications, John Wiley & Sons: New York. Exercise 7.9.

· "Employment and earnings, US 1909–1978", Department of Labor, 1979.

Appendix

```
require(fma)
data(motion)
x = as.vector(motion)
n = length(x)
t = 1:n
plot(t, x, type = "l", xlab = "time", ylab = "employment earnings",
     main = "Monthly employment figures - Motion Picture Firm")
plot(t, x, type = "l", xlab = "time", ylab = "employment earnings",
     main = "Monthly employment figures - Motion Picture Firm")
#Remove trend component
t1 = t
t2 = t^2
```

```
trend.fit = lm(x~I(t2)+t1)
lines(t, fitted(trend.fit), col="red")

#remove the trend
yt = resid(trend.fit)
ts.plot(yt, ylab = "employment earnings")
title('Resid Plot After Removin Trend')
hist(resid(trend.fit))
#Remove seanality
n = length(t)
t = 1:length(yt)
t = (t) / n
d = 12
n.harm = 6 #set to [d/2]
harm = matrix(nrow=length(t), ncol=2*n.harm)
for(i in 1:n.harm){
  harm[,i*2-1] = sin(n/d * i *2*pi*t)
  harm[,i*2] = cos(n/d * i *2*pi*t)
}
colnames(harm) = paste0(c("sin", "cos"), rep(1:n.harm, each = 2))

#fit on all of the sines and cosines
dat = data.frame(yt, harm)
fit = lm(yt~., data = dat)

full = lm(yt~.,data = dat)
reduced = lm(yt~1, data=dat)
fit.back = step(full, scope = formula(reduced),
                direction = "both", trace = 0)
t = as.vector(time(motion))
plot(t, yt, type="l", col="darkgrey", ylab="")
lines(t, fitted(fit.back), col="red")
title('Seasonal Component On Residuals(Trend Removed)')

#Residuals after seasonality removed
ytp = resid(fit.back)
ts.plot(ytp)
title('Resid After Seasonality Removed')
hist(resid(fit.back))

kpss.test(ytp)

#ACF and PACF figure of the residuals after trend and seanality are removed
acf(ytp, main = 'ACF of Resid After Trends Removed')
pacf(ytp, main = 'PACF of Resid After Trends Removed')

#Fitting Model
require(forecast)
auto.arima(ytp, allowmean = FALSE, stepwise = FALSE,
           trace = FALSE, allowdrift = FALSE, max.order = 8, ic = 'aic')

#ARMA(4,3)
armafit = arima(ytp, order = c(4,0,3), include.mean = F)
```

```r
ts.plot(resid(armafit))
title('resid plot after ARMA model fitted')
acf(resid(armafit))
pacf(resid(armafit))

#Ljong-Box Test
Box.test(armafit$residuals, type ='Ljung-Box',
         lag = min(2*d, floor(n/5)))

# check normality
z = resid(armafit)
qqnorm(z)
hist(z)
shapiro.test(z)

x = as.vector(motion)
n = length(x)
t = 1:n
#Forecast the noise
noise.f = forecast(armafit, 10, level = .95)
plot(noise.f, main = 'Forecast Noise')

#Predict the seasonality
season.f = fitted(fit.back)[1:10]
plot(season.f, main = 'Forecast Seasonality', ylab = 'Value', type = "l")

#Predict the trend
t.f = 193:202
t.f2 = t.f^2

trend.f = predict(trend.fit, newdata = data.frame(t1=t.f, t2 = t.f2))
plot(trend.f, type = 'l')
title('Forecast trend')

# add all components together to get the overall forecast
x.f = trend.f + season.f + noise.f$mean
xall = as.vector(motion)

# predict the future time points
plot(1:192, xall, type = "l", col = "darkgrey")
lines(193:202, x.f, col = "red")
lines(193:202, x.f+noise.f$lower, col = "blue", lty = 2)
lines(193:202, x.f+noise.f$upper, col = "blue", lty = 2)
```