Predicting the average weekly hours of Production and Nonsupervisory employees

Divya Krishnan
November 15 2018

Abstract

The project attempts to forecast working hours of blue collar jobs like production and nonsupervisory employees. Since the early industrial revolution in the 18th century the working trends of production employees have always been an interesting trend to study. The project aims to further study this interesting trend using ARIMA model and attempt to forecast their future working hours.

1. A brief Introduction to the problem and Data description

Since the early industrial revolution in the 18th century blue collar jobs like production and nonsupervisory have been an integral source on the economic and have impacted the socio-political climate of the country. So much that Blue-collar workers have played a large role in electoral politics. In the 2016 United States Presidential election, many attributed President Donald Trump's victories in the states of Ohio, Pennsylvania, and Michigan to blue-collar workers.

As many blue-collar jobs involve manual labor and relatively unskilled workers, automation poses a threat of unemployment for blue-collar workers. One study from the MIT Technology Review estimates that 83% of jobs that make less than $20 per hour are threatened by automation. Thus in this report I try to study if the working hours have actually been affected due to the new computer revolution.

The data set contains weekly working hours of production and nonsupervisory employees from 2000 to 2018 detailing the average of the weekly hours they worked in each month. It contains only private sector employees and the data had been seasonally adjusted.

2. Data Visualization and Model Selection

One can say that the prominence of automation and AI has happened not even a decade back and we can see that there has been a decline in the working hours of the employees.
Figure 1 shows time series plot of the data.

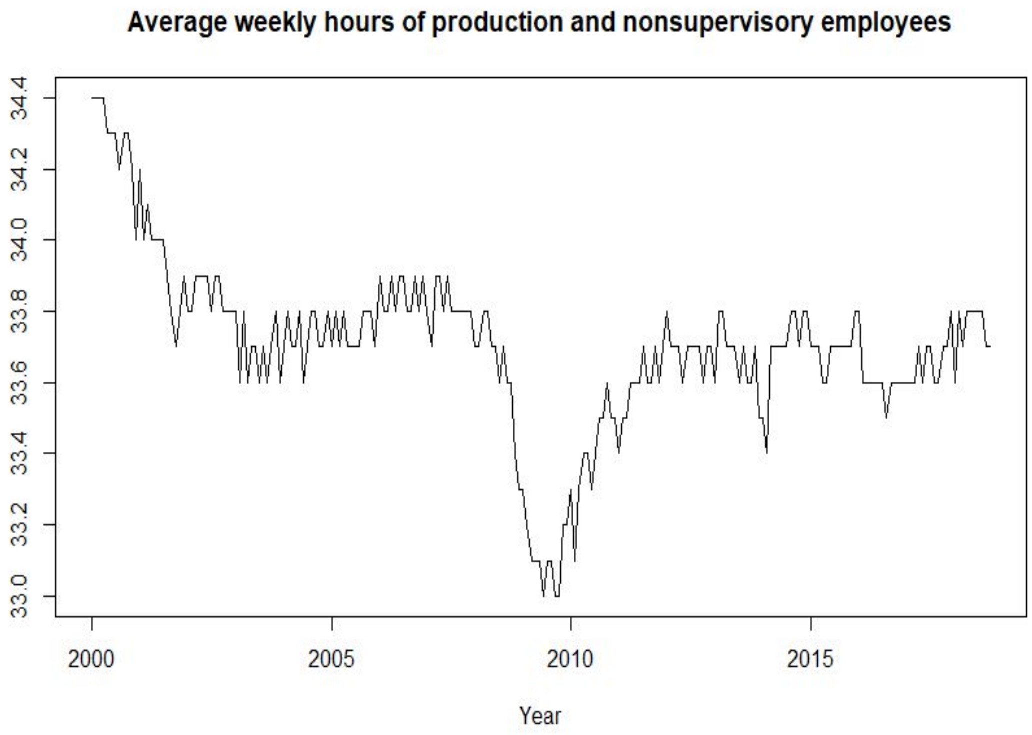


Figure 1

The above figure shows a general decline in the working hours and a sharp decline from 2009-2010. This was due to the recession that caused a depression for Blue Collared workers.

The ACF and PACF plot for the raw data is given in the below figures:

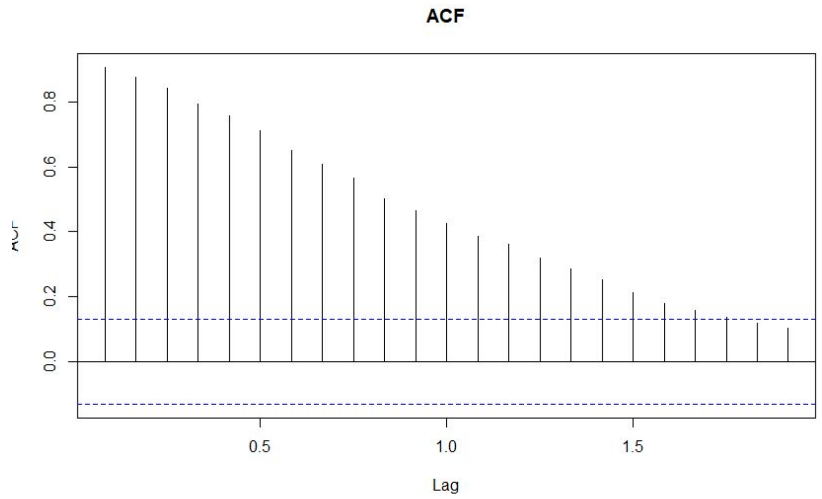


Figure 2a

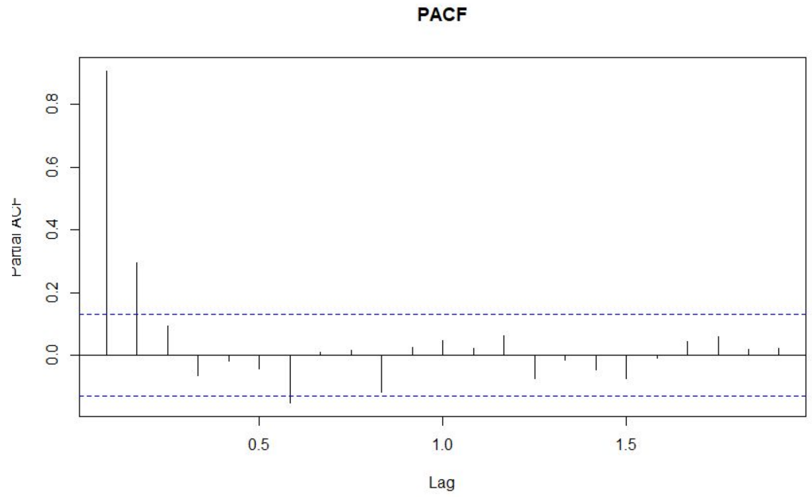
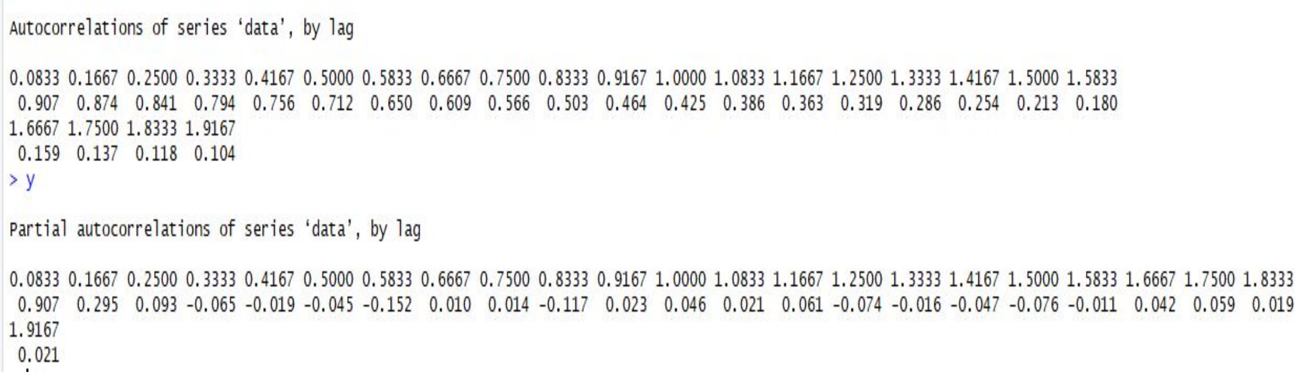


Figure 2b



```
Autocorrelations of series ‘data’, by lag

0.0833 0.1667 0.2500 0.3333 0.4167 0.5000 0.5833 0.6667 0.7500 0.8333 0.9167 1.0000 1.0833 1.1667 1.2500 1.3333 1.4167 1.5000 1.5833
 0.907  0.874  0.841  0.794  0.756  0.712  0.650  0.609  0.566  0.503  0.464  0.425  0.386  0.363  0.319  0.286  0.254  0.213  0.180
1.6667 1.7500 1.8333 1.9167
 0.159  0.137  0.118  0.104
> y

Partial autocorrelations of series ‘data’, by lag

0.0833 0.1667 0.2500 0.3333 0.4167 0.5000 0.5833 0.6667 0.7500 0.8333 0.9167 1.0000 1.0833 1.1667 1.2500 1.3333 1.4167 1.5000 1.5833 1.6667 1.7500 1.8333
 0.907  0.295  0.093 -0.065 -0.019 -0.045 -0.152  0.010  0.014 -0.117  0.023  0.046  0.021  0.061 -0.074 -0.016 -0.047 -0.076 -0.011  0.042  0.059  0.019
1.9167
 0.021
```

Table 1

2b Box- Cox Transformation

The ACF plot shows patterns of a non stationary data as it decays very slowly. Figure 1 shows non uniform variance in the data hence we decide on doing Box- Cox transformation which suggests $\Lambda = 1.9996$, so we do Box Cox transformation on our data rendering the data to look like figure 3

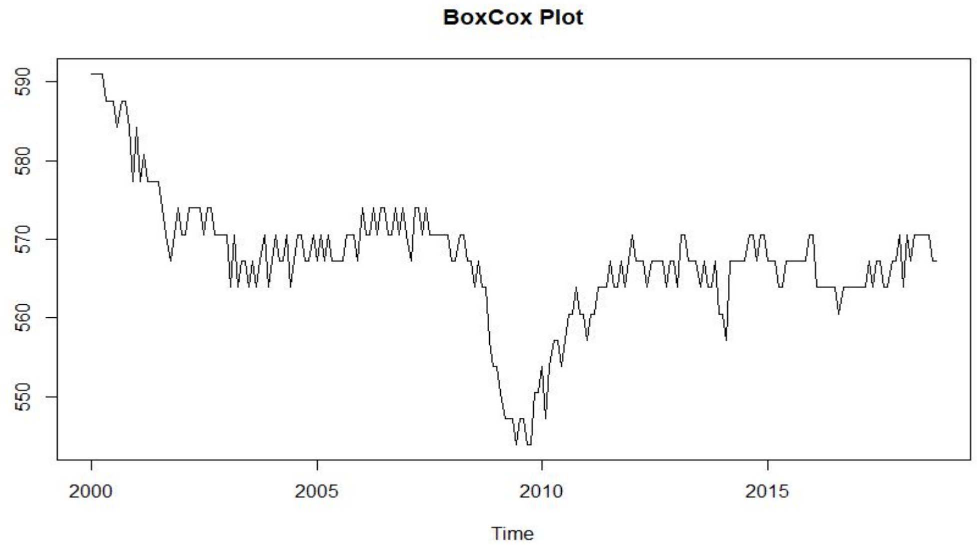


Figure 3

The ACF and PACF plot of the Box Cox transformed data are shown in figure 4

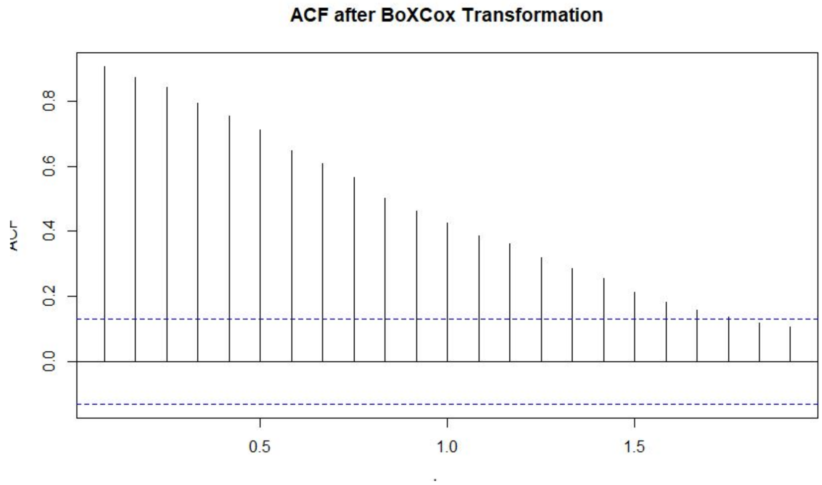


Figure 4a

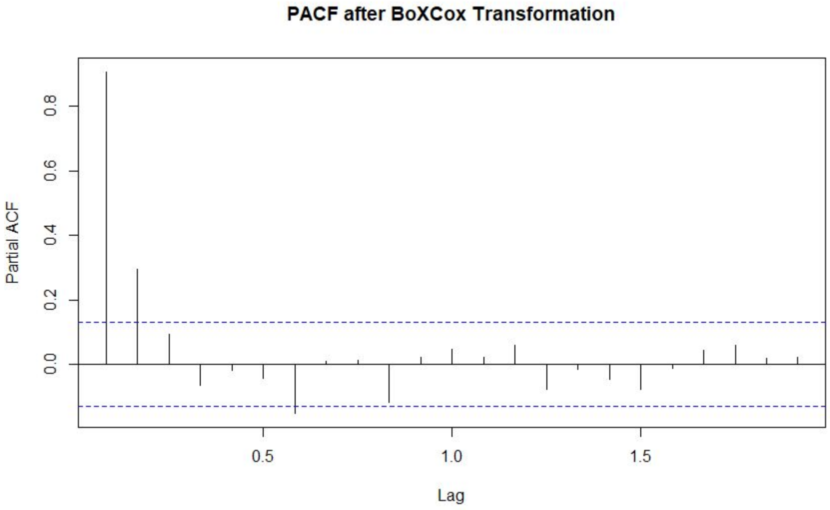
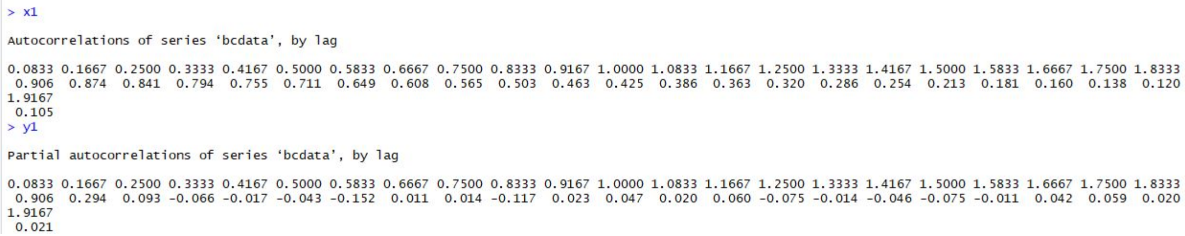


Figure 4b



```
> x1

Autocorrelations of series ‘bcdata’, by lag

0.0833 0.1667 0.2500 0.3333 0.4167 0.5000 0.5833 0.6667 0.7500 0.8333 0.9167 1.0000 1.0833 1.1667 1.2500 1.3333 1.4167 1.5000 1.5833 1.6667 1.7500 1.8333
 0.906  0.874  0.841  0.794  0.755  0.711  0.649  0.608  0.565  0.503  0.463  0.425  0.386  0.363  0.320  0.286  0.254  0.213  0.181  0.160  0.138  0.120
1.9167
 0.105
> y1

Partial autocorrelations of series ‘bcdata’, by lag

0.0833 0.1667 0.2500 0.3333 0.4167 0.5000 0.5833 0.6667 0.7500 0.8333 0.9167 1.0000 1.0833 1.1667 1.2500 1.3333 1.4167 1.5000 1.5833 1.6667 1.7500 1.8333
 0.906  0.294  0.093 -0.066 -0.017 -0.043 -0.152  0.011  0.014 -0.117  0.023  0.047  0.020  0.060 -0.075 -0.014 -0.046 -0.075 -0.011  0.042  0.059  0.020
1.9167
 0.021
```

Table 2

2c) Differencing

The Box Cox clearly doesn't fix the non stationarity of the data and doing Augmented Dickey-Fuller Test gives p-value 0.1646, indicating that it is necessary to take differencing
On taking the first difference we get the following ACF and PACF plots:

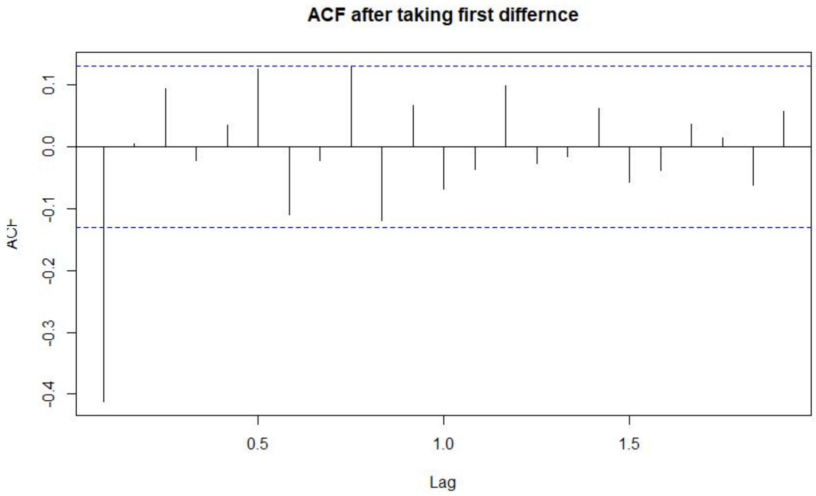


Figure 5a

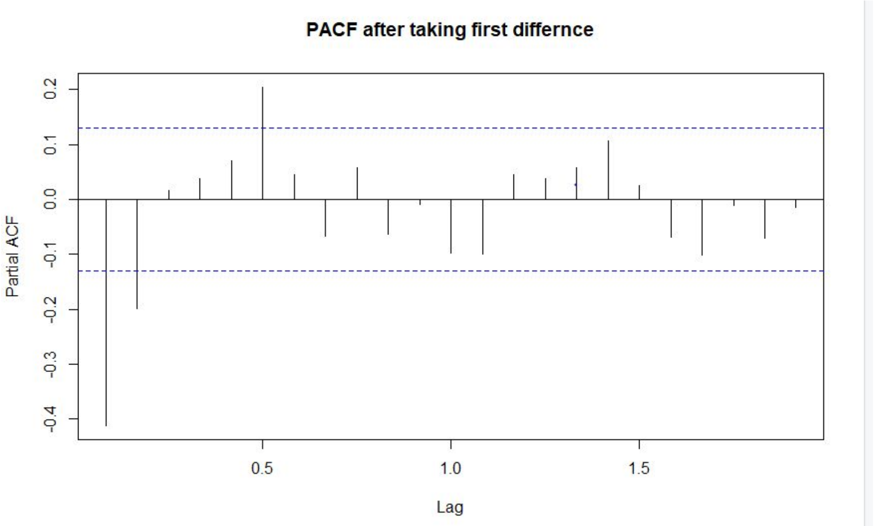
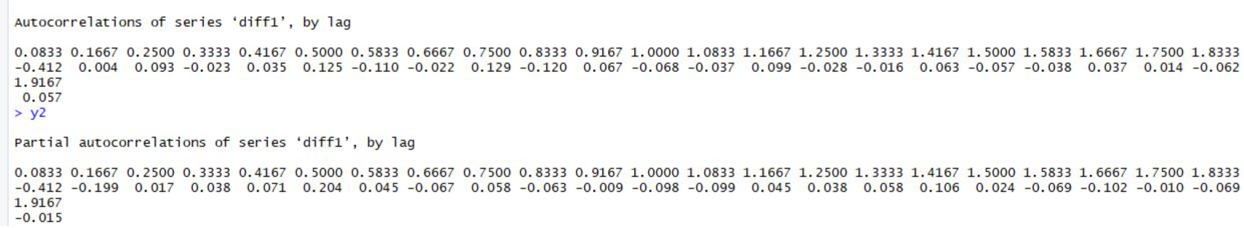


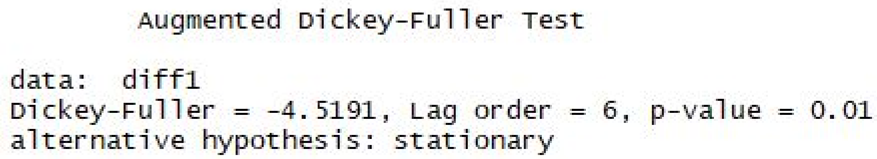
Figure 5b



```
Autocorrelations of series ‘diff1’, by lag

0.0833 0.1667 0.2500 0.3333 0.4167 0.5000 0.5833 0.6667 0.7500 0.8333 0.9167 1.0000 1.0833 1.1667 1.2500 1.3333 1.4167 1.5000 1.5833 1.6667 1.7500 1.8333
-0.412  0.004  0.093 -0.023  0.035  0.125 -0.110 -0.022  0.129 -0.120  0.067 -0.068 -0.037  0.099 -0.028 -0.016  0.063 -0.057 -0.038  0.037  0.014 -0.062
1.9167
 0.057
> y2

Partial autocorrelations of series ‘diff1’, by lag

0.0833 0.1667 0.2500 0.3333 0.4167 0.5000 0.5833 0.6667 0.7500 0.8333 0.9167 1.0000 1.0833 1.1667 1.2500 1.3333 1.4167 1.5000 1.5833 1.6667 1.7500 1.8333
-0.412 -0.199  0.017  0.038  0.071  0.204  0.045 -0.067  0.058 -0.063 -0.009 -0.098 -0.099  0.045  0.038  0.058  0.106  0.024 -0.069 -0.102 -0.010 -0.069
1.9167
-0.015
```

Tabe 3

By taking the Augmented Dickey-Fuller Test again on the difference data we get p = 0.01 Indicating the data is now stationary and no more differencing is needed.



```
        Augmented Dickey-Fuller Test

data:  diff1
Dickey-Fuller = -4.5191, Lag order = 6, p-value = 0.01
alternative hypothesis: stationary
```

2d) Model selection

From the ACF and PACF plot, we can say that it's an ARIMA(p,d,q) model. We try different values of p and q and compare the BIC of each model.

ARMA (p,q)	BIC error
1,1	1123.32
2,3	1123.07
3,2	1124.34
1,2	1119.23
5,5	1136.72

Note that the BIC errors are really close and we can't decide on the best model that fits the data from just using BIC values.

So we decide to compare the models by cross validation. Which means for data from 2000 to 2017 and find the MSE by comparing the predicted weekly working hours in 2018 with the actual values in 2018

We consider the top 3 models for cross validation - ARIMA (1,1), ARIMA(2,3) and ARIMA(1,2)

ARMA (p,q)	Cross Validation Error Rate
1,1	36.7859
2,3	32.3315
1,2	44.897

After comparing the cross validation error we decide that ARIMA(2,1,3) fits the model the best to give us better forecast.

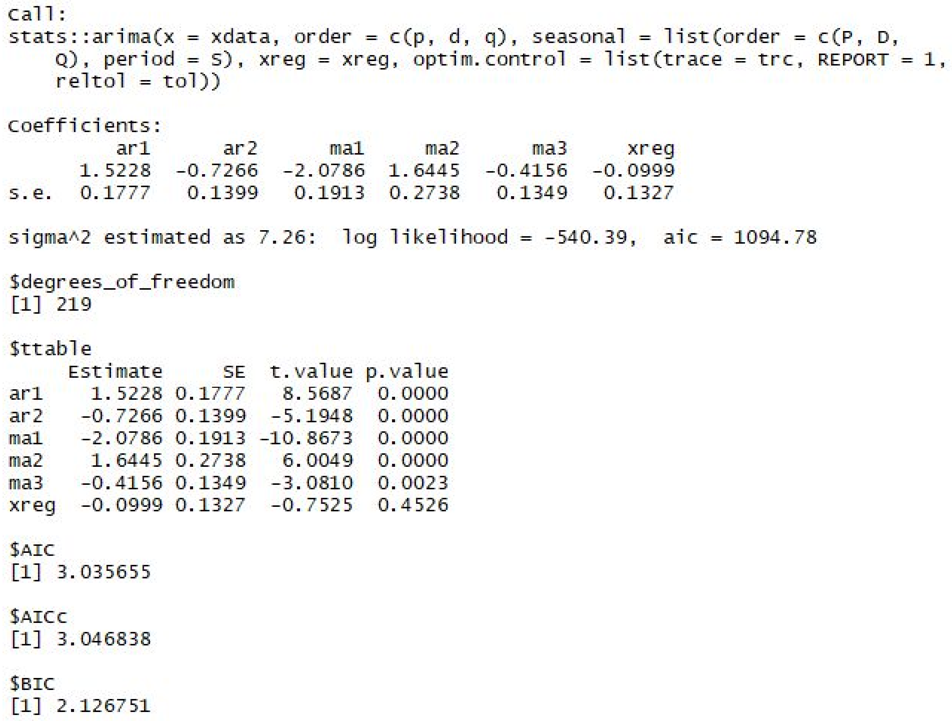
$$\phi(B)(1-B)xt = \delta + \theta(B)\omega t$$

$\phi B = 1 - 1.5228B + 0.7266B^2$

$\theta B = 1 - 2.0786B + 1.6445B^2 - 0.4156B^3$

$\delta = -0.0999(1 + 2.0786 - 1.6445 + 0.04156)$

As the p values for all the coefficients are significant I decide on keeping all of them.



```
Call:
stats::arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D,
    Q), period = S), xreg = xreg, optim.control = list(trace = trc, REPORT = 1,
    reltol = tol))

Coefficients:
         ar1      ar2      ma1     ma2      ma3     xreg
      1.5228  -0.7266  -2.0786  1.6445  -0.4156  -0.0999
s.e.  0.1777   0.1399   0.1913  0.2738   0.1349   0.1327

sigma^2 estimated as 7.26:  log likelihood = -540.39,  aic = 1094.78

$degrees_of_freedom
[1] 219

$ttable
     Estimate     SE  t.value p.value
ar1    1.5228 0.1777   8.5687  0.0000
ar2   -0.7266 0.1399  -5.1948  0.0000
ma1   -2.0786 0.1913 -10.8673  0.0000
ma2    1.6445 0.2738   6.0049  0.0000
ma3   -0.4156 0.1349  -3.0810  0.0023
xreg  -0.0999 0.1327  -0.7525  0.4526

$AIC
[1] 3.035655

$AICc
[1] 3.046838

$BIC
[1] 2.126751
```

2e) Diagnostic Checking

On diagnostic checking for ARIMA(2,1,3) , The histogram in Figure suggests that the normality assumption is valid for the model but the The QQ-plot shows that some points at the tail do not lies in the straight line, indicating there may exist outliers however the standard residuals are fairly constant as they are pretty random.

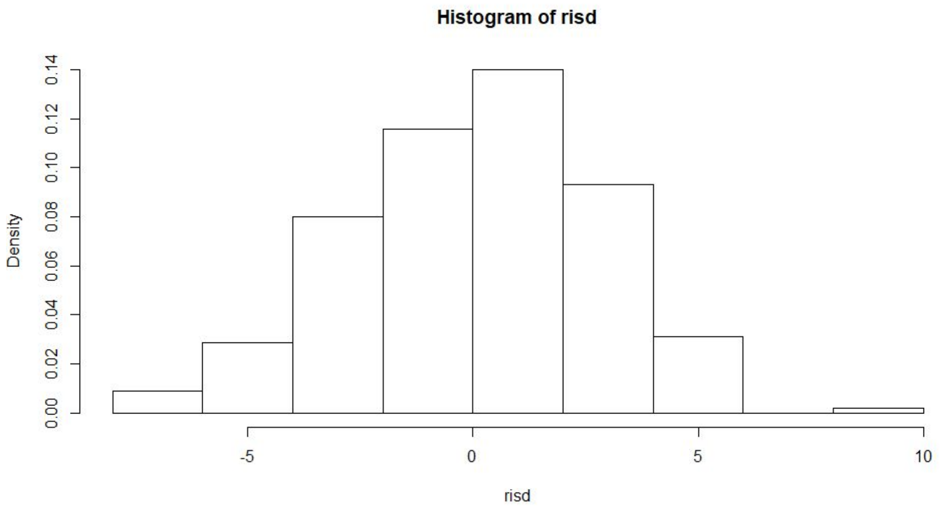


Figure 6

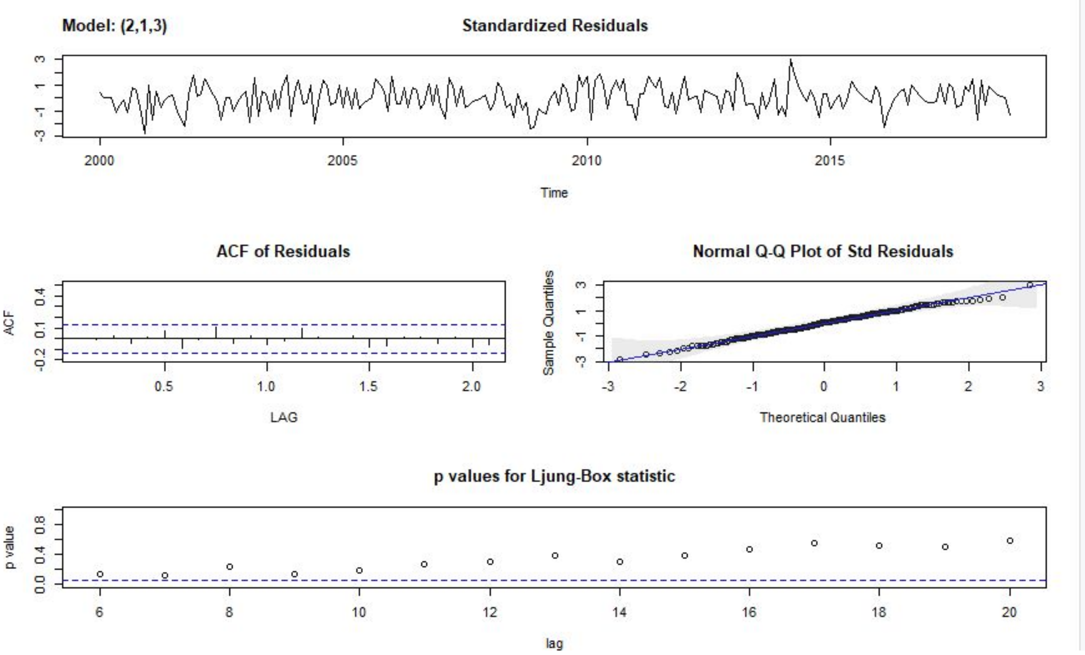


Figure 7

The ACF of residue shows pattern for white noise and Thus, the model passed the diagnostics show.

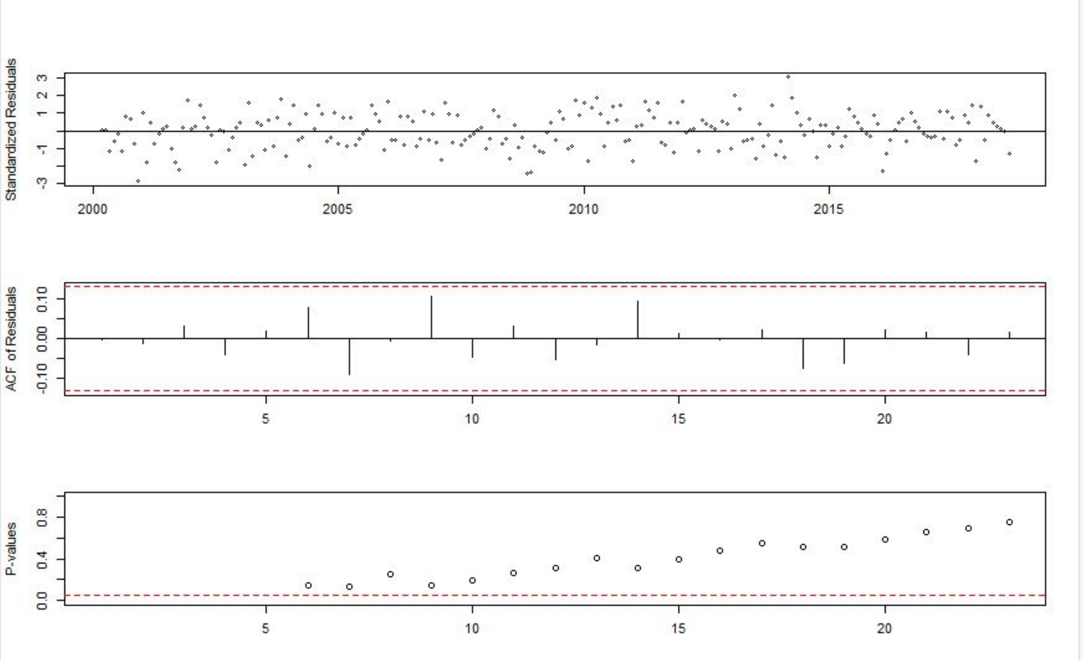
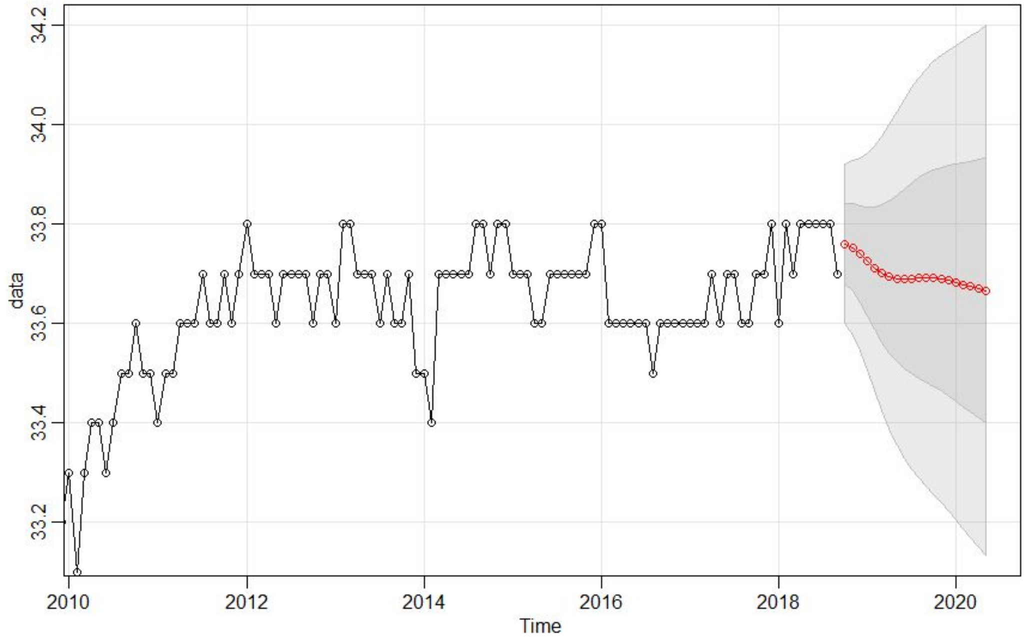
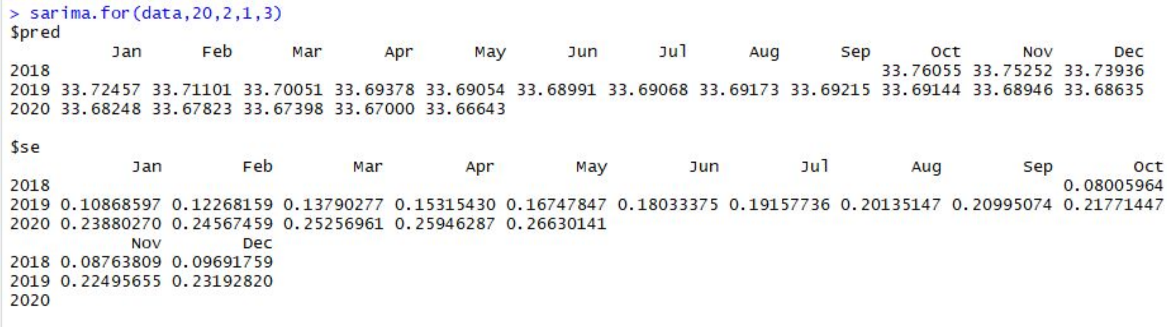


Figure 8

3.Forecasting

After fitting the ARMA(2,3) on the data I predict the working hours of the employees for the next 20 months. He trend can be seen on Figure 9 and the predicted values can be seen on table 4





```
> sarima.for(data,20,2,1,3)
$pred
          Jan      Feb      Mar      Apr      May      Jun      Jul      Aug      Sep      Oct      Nov      Dec
2018                                                                                     33.76055 33.75252 33.73936
2019 33.72457 33.71101 33.70051 33.69378 33.69054 33.68991 33.69068 33.69173 33.69215 33.69144 33.68946 33.68635
2020 33.68248 33.67823 33.67398 33.67000 33.66643

$se
            Jan        Feb        Mar        Apr        May        Jun        Jul        Aug        Sep        Oct
2018                                                                                                   0.08005964
2019 0.10868597 0.12268159 0.13790277 0.15315430 0.16747847 0.18033375 0.19157736 0.20135147 0.20995074 0.21771447
2020 0.23880270 0.24567459 0.25256961 0.25946287 0.26630141
            Nov        Dec
2018 0.08763809 0.09691759
2019 0.22495655 0.23192820
2020
```

Table 4

4.Conclusion

From the plot it can be noticed that the general trend on the working hours is decreasing although it not a steep decline. So we can conclude that although rise in automation can posses a threat for production employees but its effects won't be felt at least in the near future

Appendix

Code

```
library(astsa)
library(forecast)
library(tidyr)
data = read.table("C:\\Users\\Divya Krishnan\\Desktop\\proj1.csv")

data = ts(data, frequency =12, start = c(2000,1))
plot(data,main = 'Average weekly hours of production and nonsupervisory employees',xlab =
"Year",ylab= "Weekly income of employees from 2000 to 2018")
x = acf(data, main = "ACF")
y = pacf(data, main = "PACF")
lambda = BoxCox.lambda(data)
bcdata = BoxCox(data,lambda)
plot(bcdata,main =  "BoxCox Plot")
x1=acf(bcdata, main = "ACF after BoXCox Transformation")
y1=pacf(bcdata,main = "PACF after BoXCox Transformation")
adf.test(bcdata)


diff1 = diff(bcdata)
x2 = acf(diff1, main = "ACF after taking first differnce")
y2 = pacf(diff1,main = "PACF after taking first differnce")
adf.test(diff1)

#model building

eacf(diff1)
nobs = length(bcdata)
sarima(bcdata, 2,1,3,xreg = 1:nobs)
mdl_arma = arima(bcdata,order = c(2,1,3), xreg = 1:nobs)
tsdiag(mdl_arma)
risd = residuals(mdl_arma)
hist(risd, probability =  T)
qqnorm(risd)
qqline(risd)
plot(bcdata, risd)
pred = predict(mdl_arma,20,(nobs+1): (nobs+20))
fore.pred = InvBoxCox(pred$pred, 1.999)
fore.se = InvBoxCox(pred$se, 1.999)
plot(forecast(Arima(data, order = c(2,1,3),lambda = 1.999)))
ts.plot(data,fore.pred,col = 1:2)
sarima.for(data,20,2,1,3)

#model Selection

nobs = length(bcdata)
mdl1 = arima(bcdata, order = c(2,1,3),xreg = 1:nobs)
BIC(mdl1)
mdl2 = arima(bcdata, order = c(1,1,2),xreg = 1:nobs)
BIC(mdl2)
mdl3 = arima(bcdata, order = c(1,1,1),xreg = 1:nobs)
BIC(mdl3)
mdl4 = arima(bcdata, order = c(5,1,5),xreg = 1:nobs)
BIC(mdl4)
mdl4 = arima(bcdata, order = c(3,1,2),xreg = 1:nobs)
BIC(mdl4)
```

BIBLIOGRAPHY

1. https://data.bls.gov/cgi-bin/surveymost

2.https://www.technologyreview.com/s/603465/the-relentless-pace-of-automation/