STAT443 Assignment4

Yiming Shen 20891774

29/03/2023

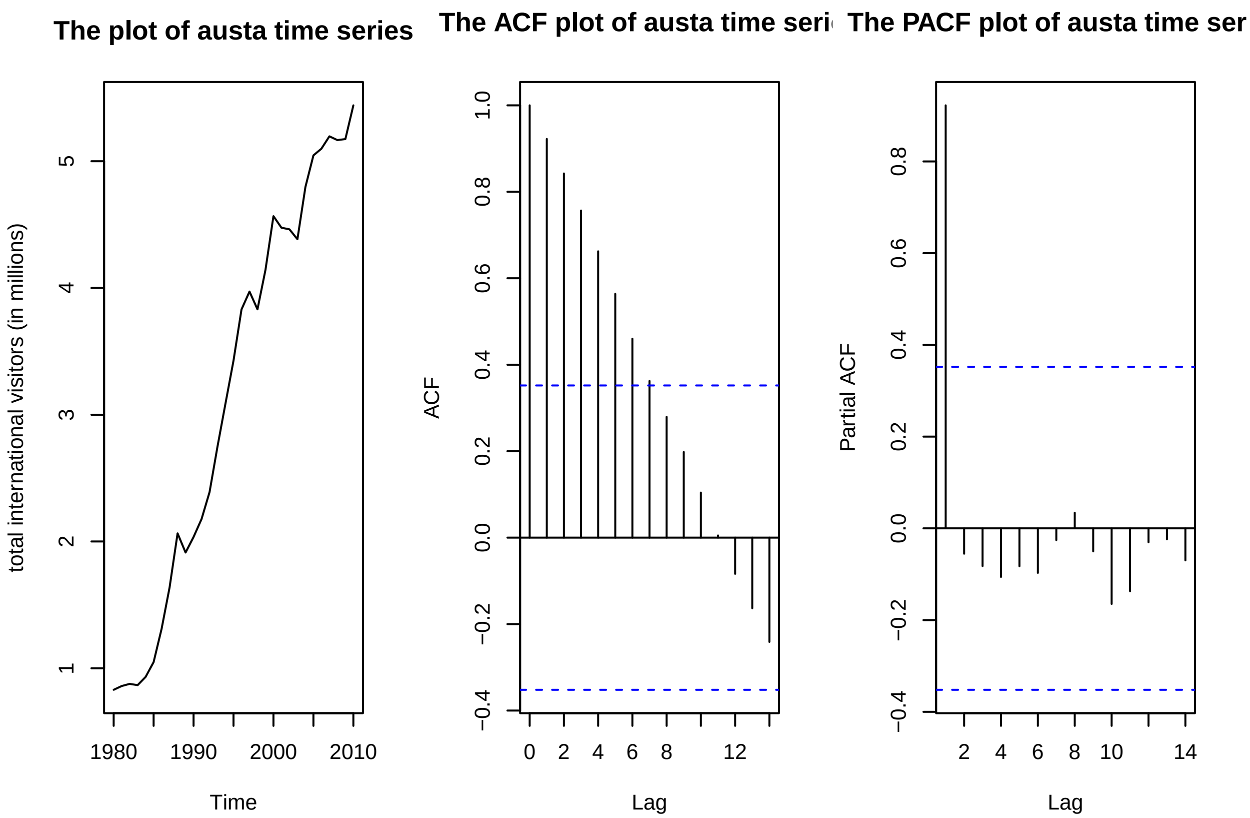
Problem 4

(a)

```
library(fpp)
```

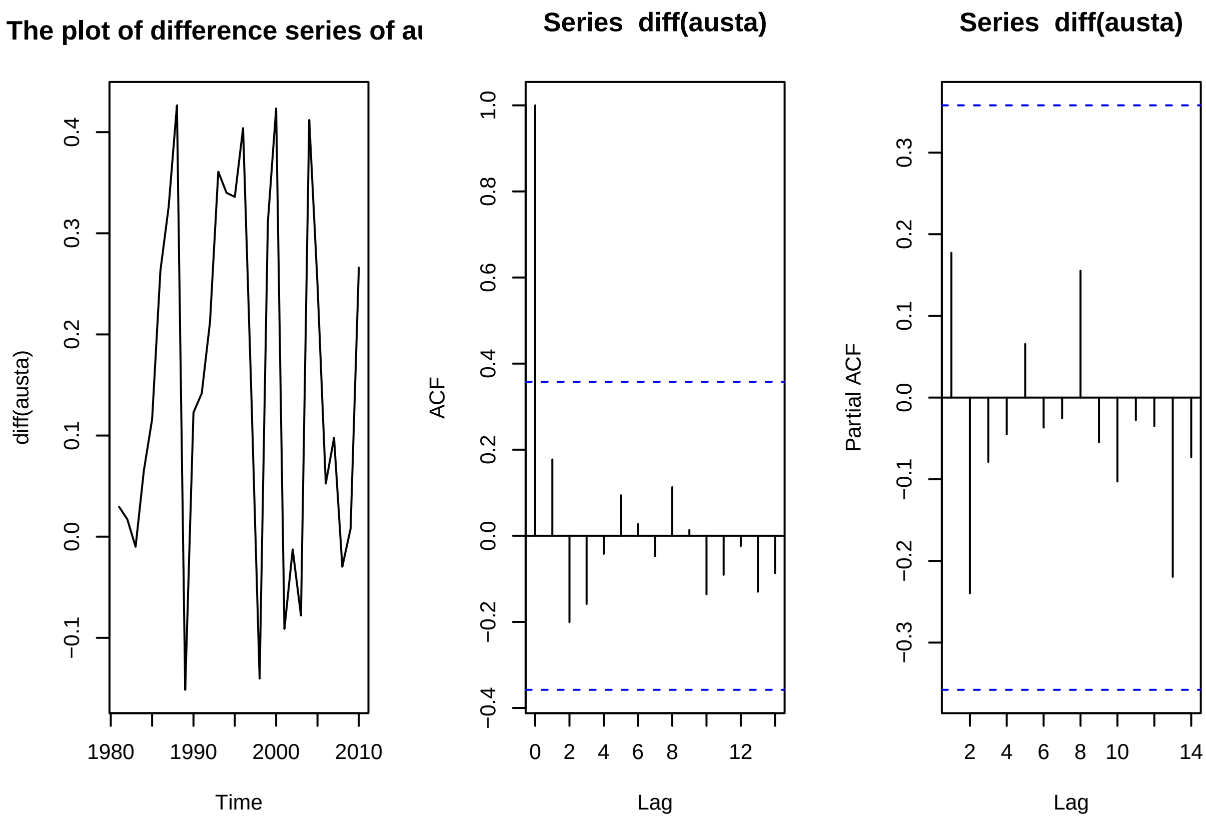
```
## Loading required package: forecast

## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo

## Loading required package: fma

## Loading required package: expsmooth

## Loading required package: lmtest

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: tseries
```

```
par(mfrow=c(1,3))
# plot the series
plot(austa, main="The plot of austa time series",
     ylab="total international visitors (in millions)")
# plot ACF
acf(austa, main="The ACF plot of austa time series")
# plot PACF
pacf(austa, main="The PACF plot of austa time series")
```




```
# comments:
# Based on the time series plot, we found that there is a increasing trend over
# 1980 to 2015, so it is not stationary.
# Based on the ACF plot, we found that ACF has a slow decay and cut off after
# lag7.
# Based on the PACF plot, we found that except significant spike at lag1, other
# spikes stay within the band.
```

```
par(mfrow=c(1,3))
# plot difference series
plot(diff(austa), main="The plot of difference series of austa")
# plot its ACF
acf(diff(austa))
# plot its PACF
pacf(diff(austa))
```



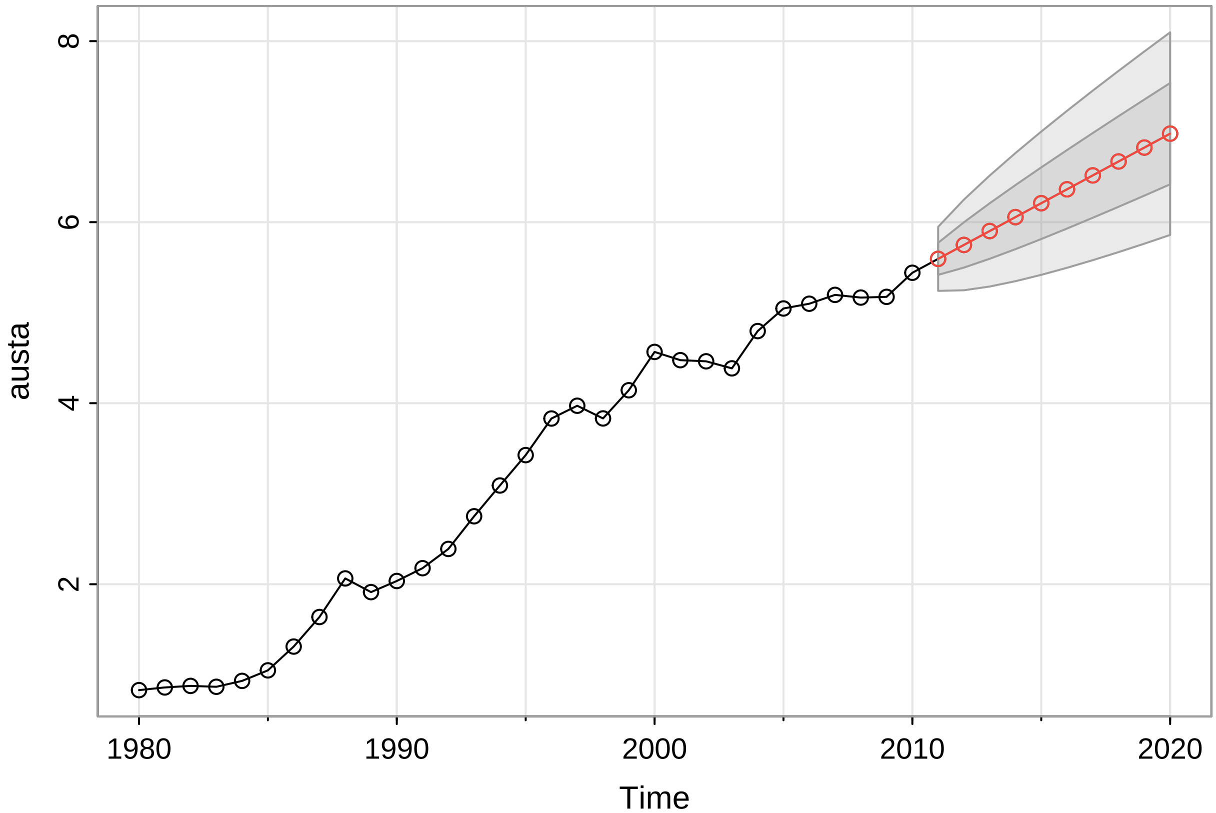

```
# comments:
# Based on the plot of difference series, there is no obvious trend and it
# looks quite stationary.
# Based on its ACF and PACF plot, almost (except lag0 in ACF) all spikes stay
# within the band, so there is no significant spikes for the first difference
# series.
```

(b)

```
# How to settled on the model:
# We notice that the first difference series is relative stationary and looks
# like a white noise process. So we fit a ARIMA(0,1,0) model to the series.
library(astsa)
```

```
##
## Attaching package: 'astsa'

## The following object is masked from 'package:fpp':
##
##     oil

## The following objects are masked from 'package:fma':
##
##     chicken, sales

## The following object is masked from 'package:forecast':
##
##     gas
```

```r
# forecast 10 steps ahead
model <- sarima.for(austa, n.ahead=10, p=0, d=1, q=0)
```




```r
model
```

```
## $pred
## Time Series:
## Start = 2011
## End = 2020
## Frequency = 1
##  [1] 5.594594 5.748294 5.901994 6.055694 6.209394 6.363094 6.516794 6.670494
##  [9] 6.824194 6.977894
##
## $se
## Time Series:
## Start = 2011
## End = 2020
## Frequency = 1
##  [1] 0.1769883 0.2502993 0.3065528 0.3539766 0.3957579 0.4335311 0.4682671
##  [8] 0.5005986 0.5309649 0.5596862
```

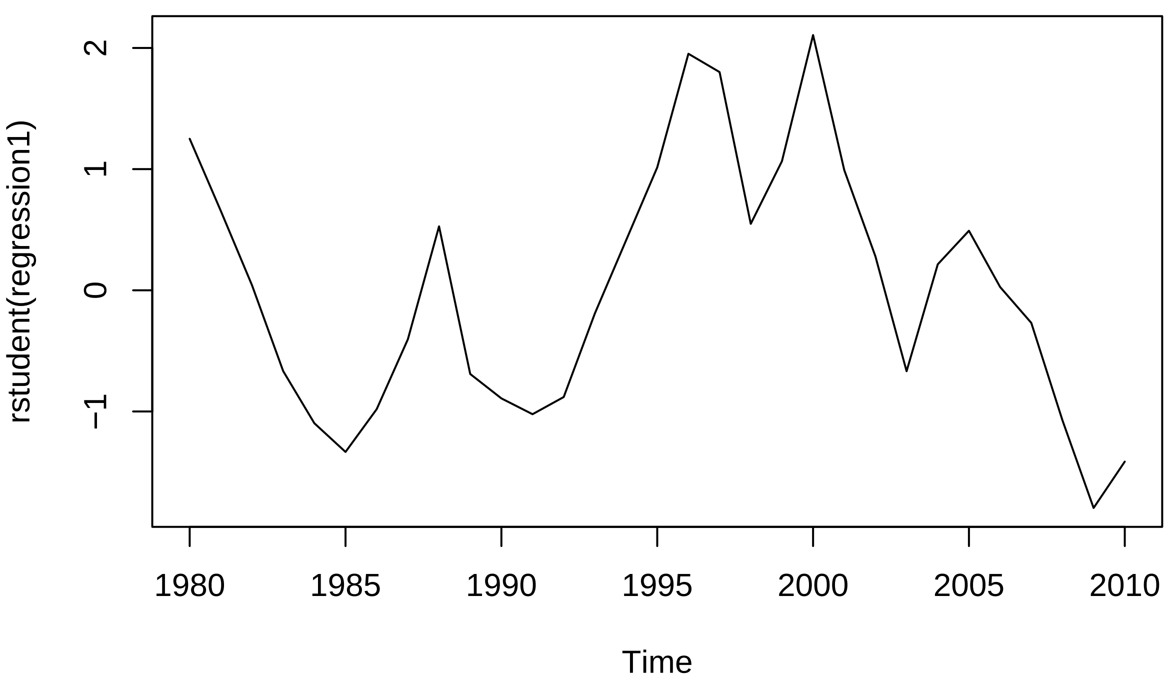
```r
# construct 95% prediction interval
c <- 1.96
lower_bound <- model$pred-c*model$se
upper_bound <- model$pred+c*model$se
cbind(model$pred,lower_bound, upper_bound)
```

```
## Time Series:
## Start = 2011
```

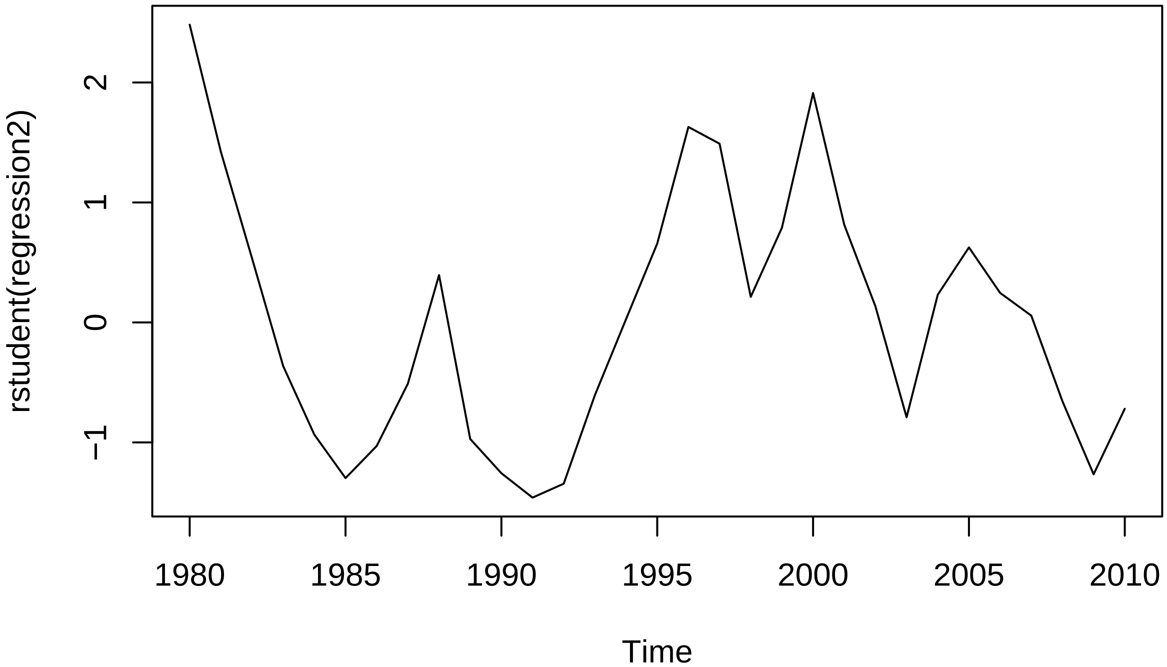
```
## End = 2020
## Frequency = 1
##      model$pred lower_bound upper_bound
## 2011   5.594594    5.247697    5.941491
## 2012   5.748294    5.257707    6.238881
## 2013   5.901994    5.301151    6.502837
## 2014   6.055694    5.361900    6.749488
## 2015   6.209394    5.433708    6.985079
## 2016   6.363094    5.513373    7.212815
## 2017   6.516794    5.598990    7.434597
## 2018   6.670494    5.689321    7.651667
## 2019   6.824194    5.783503    7.864885
## 2020   6.977894    5.880909    8.074879
```

(c)

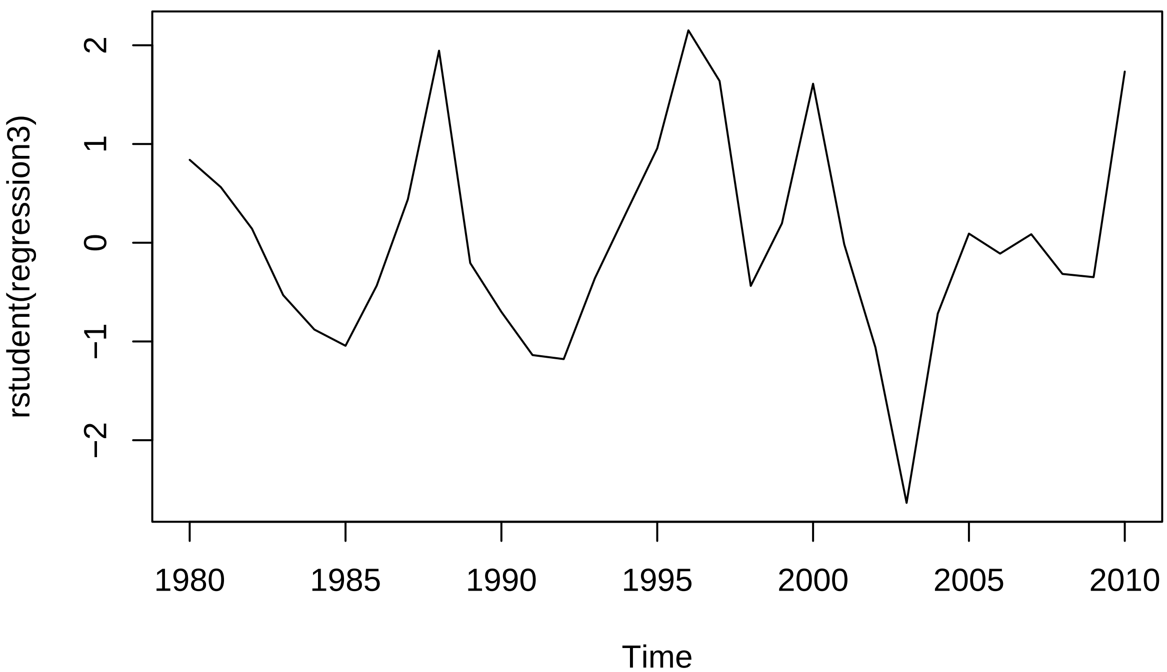
```
time.vec <- 1:length(austa)
time.vec2 <- time.vec^2
time.vec3 <- time.vec^3
time.vec4 <- time.vec^4
time.vec5 <- time.vec^5

# fitting models
regression1 <- lm(austa~time.vec, na.action=NULL)
regression2 <- lm(austa~time.vec+time.vec2, na.action=NULL)
regression3 <- lm(austa~time.vec+time.vec2+time.vec3, na.action=NULL)
regression4 <- lm(austa~time.vec+time.vec2+time.vec3+time.vec4, na.action=NULL)
regression5 <- lm(austa~time.vec+time.vec2+time.vec3+time.vec4+time.vec5,
                  na.action=NULL)

# Based on Residual plot
plot(rstudent(regression1))
```



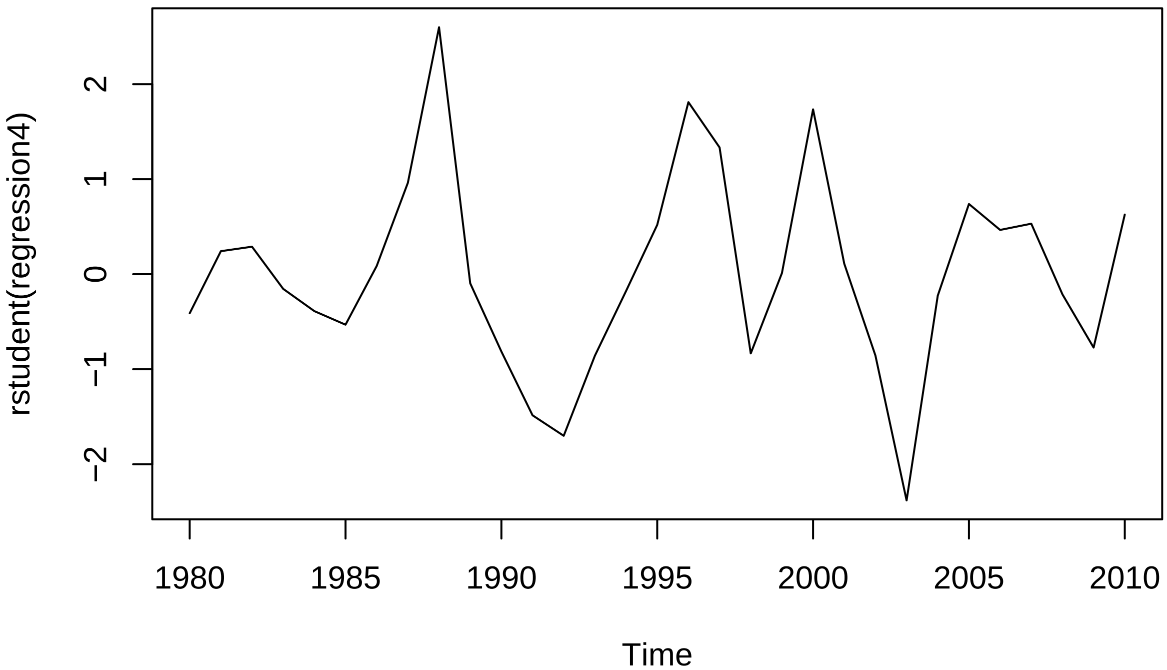
```
plot(rstudent(regression2))
```



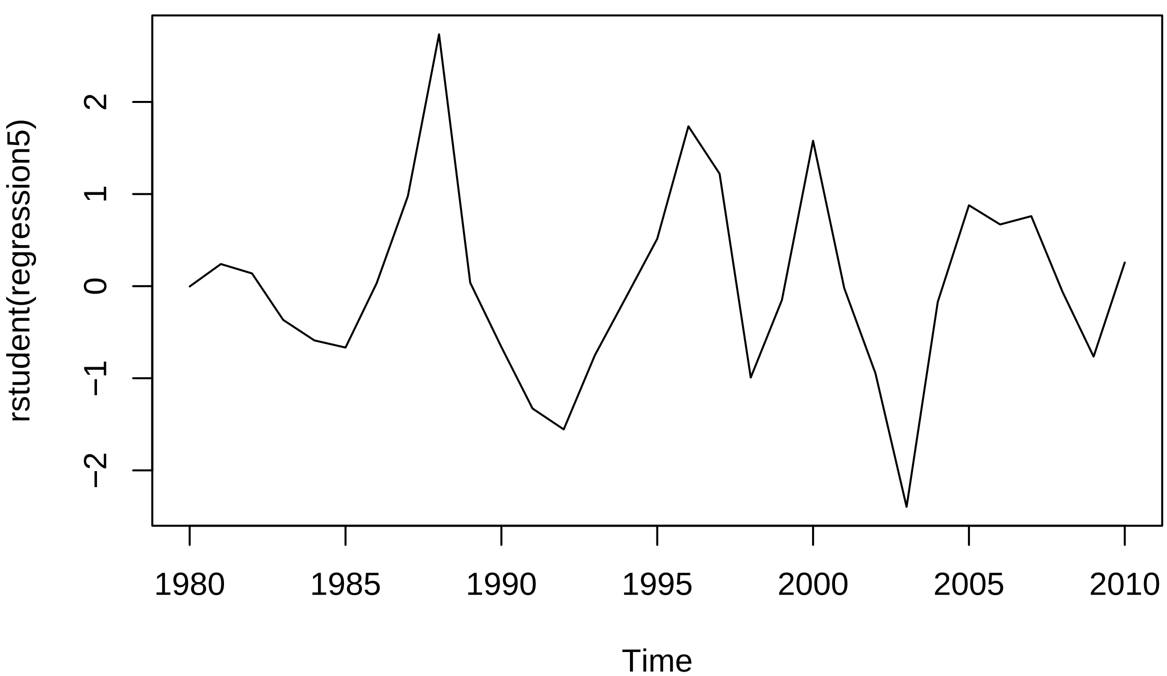

```
plot(rstudent(regression3))
```




```
plot(rstudent(regression4))
```




```
plot(rstudent(regression5))
```




```
# Based on AIC
AIC(regression1)
```

```
## [1] 11.33923
```

```
AIC(regression2)
```

```
## [1] 9.026575
```

```
AIC(regression3)
```

```
## [1] -12.8224
```

```
AIC(regression4)
```

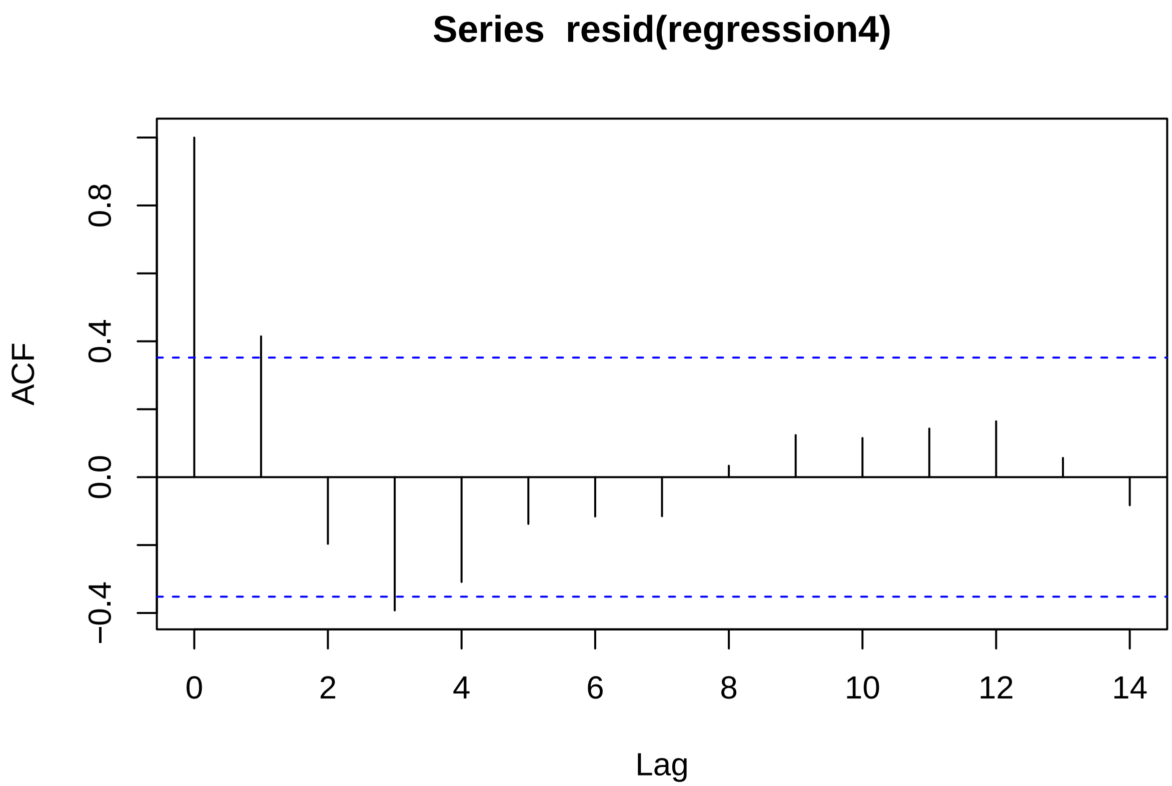
```
## [1] -17.66838
```

```
AIC(regression5)
```

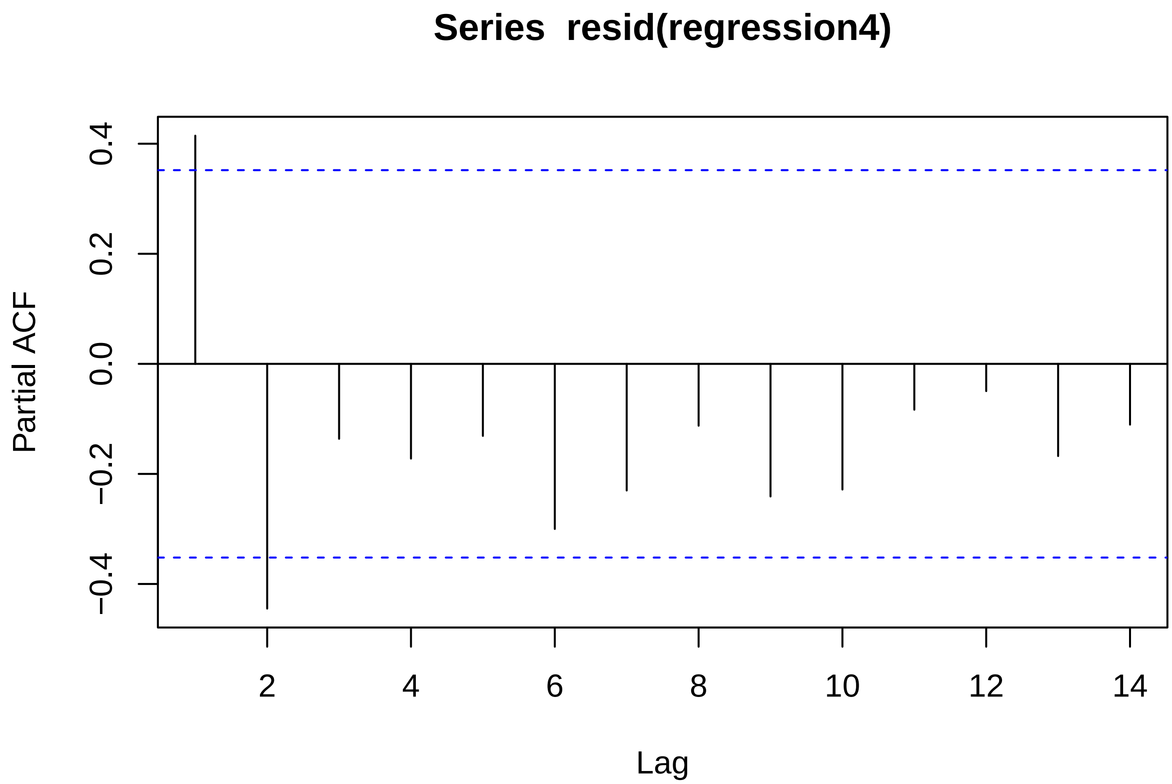
```
## [1] -16.34848
```

```
# fitting trend+noise model:
# Based on the residual plots, we found that there seems to be decreasing trend
# in regression1 and regression2, and for regression 3,4,5, they all do not have
# obvious trend.

# Then based on their AIC, we found that model_reg4 has the smallest AIC.
# Therefore, based on both residual plot and AIC, regression4
# (time.vec+time.vec2+time.vec3+time.vec4) is appropriate.
```

```
acf(resid(regression4))
```



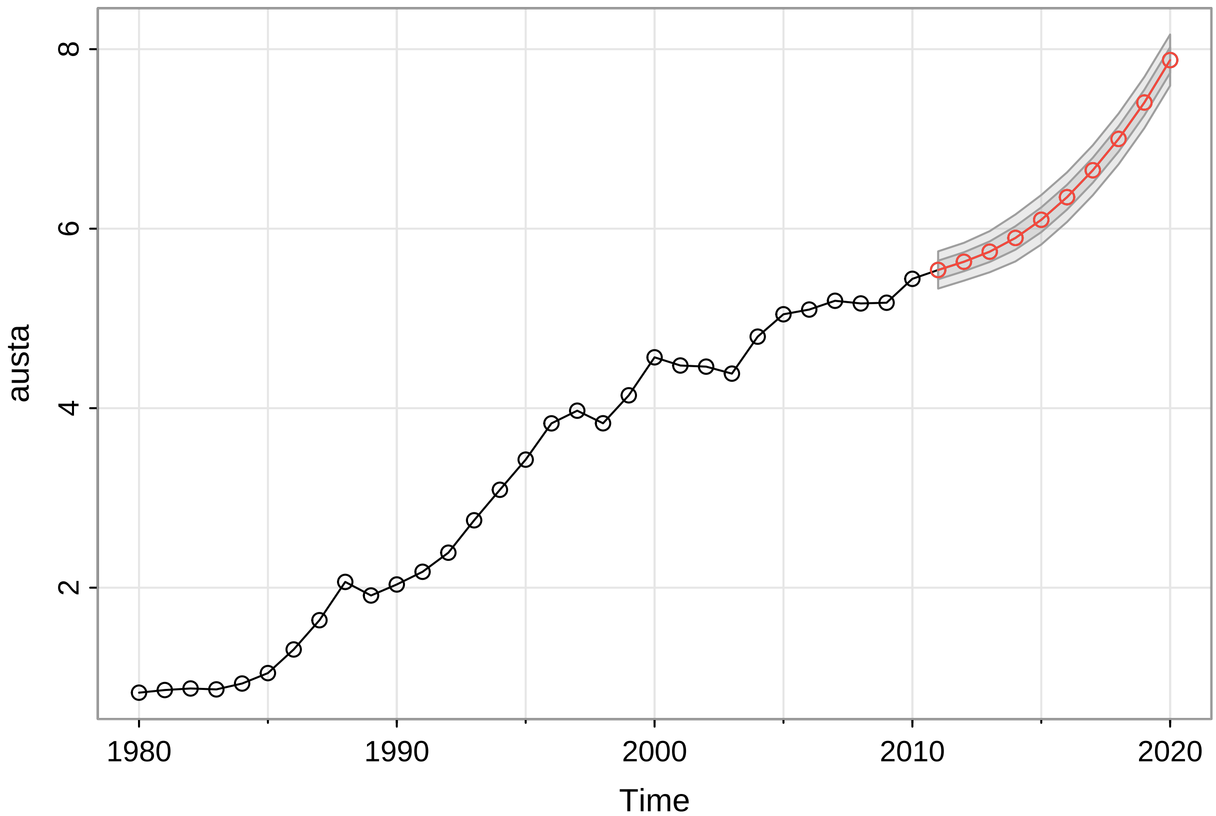

```
pacf(resid(regression4))
```




```
# How to settled on the model:
# Based on the ACF and PACF plot of reg4, we found that its ACF cut off after
# lag1 and PACF cut off after lag2, so we fit a ARIMA(2,0,1) to the series.
```

```
predict_time = 32:41
predict_data = cbind(time.vec=predict_time,
                     time.vec2=predict_time^2,
                     time.vec3=predict_time^3,
                     time.vec4=predict_time^4)
# forecast 10 steps ahead
model2 <- sarima.for(austa, n.ahead = 10,
                     p=2, d=0, q=1,
                     xreg=model.matrix(regression4)[,-1],
                     newxreg = predict_data)
```

```
## Warning in log(s2): NaNs produced

## Warning in log(s2): NaNs produced

## Warning in log(s2): NaNs produced

## Warning in log(s2): NaNs produced
```




```r
model2
```

```
## $pred
## Time Series:
## Start = 2011
## End = 2020
## Frequency = 1
##  [1] 5.539765 5.631367 5.743567 5.896603 6.099000 6.350941 6.650713 6.999946
##  [9] 7.405515 7.878398
##
## $se
## Time Series:
## Start = 2011
## End = 2020
## Frequency = 1
##  [1] 0.1038075 0.1053323 0.1145138 0.1307160 0.1377850 0.1381310 0.1390896
##  [8] 0.1413095 0.1424743 0.1425601
```

```r
# construct 95% prediction interval
c <- 1.96
lower_bound <- model2$pred-c*model2$se
upper_bound <- model2$pred+c*model2$se
cbind(model2$pred,lower_bound,upper_bound)
```

```
## Time Series:
## Start = 2011
## End = 2020
## Frequency = 1
##      model2$pred lower_bound upper_bound
```

```
## 2011    5.539765    5.336302    5.743228
## 2012    5.631367    5.424916    5.837819
## 2013    5.743567    5.519120    5.968014
## 2014    5.896603    5.640400    6.152807
## 2015    6.099000    5.828942    6.369059
## 2016    6.350941    6.080205    6.621678
## 2017    6.650713    6.378098    6.923329
## 2018    6.999946    6.722979    7.276912
## 2019    7.405515    7.126265    7.684764
## 2020    7.878398    7.598980    8.157816
```

(d)

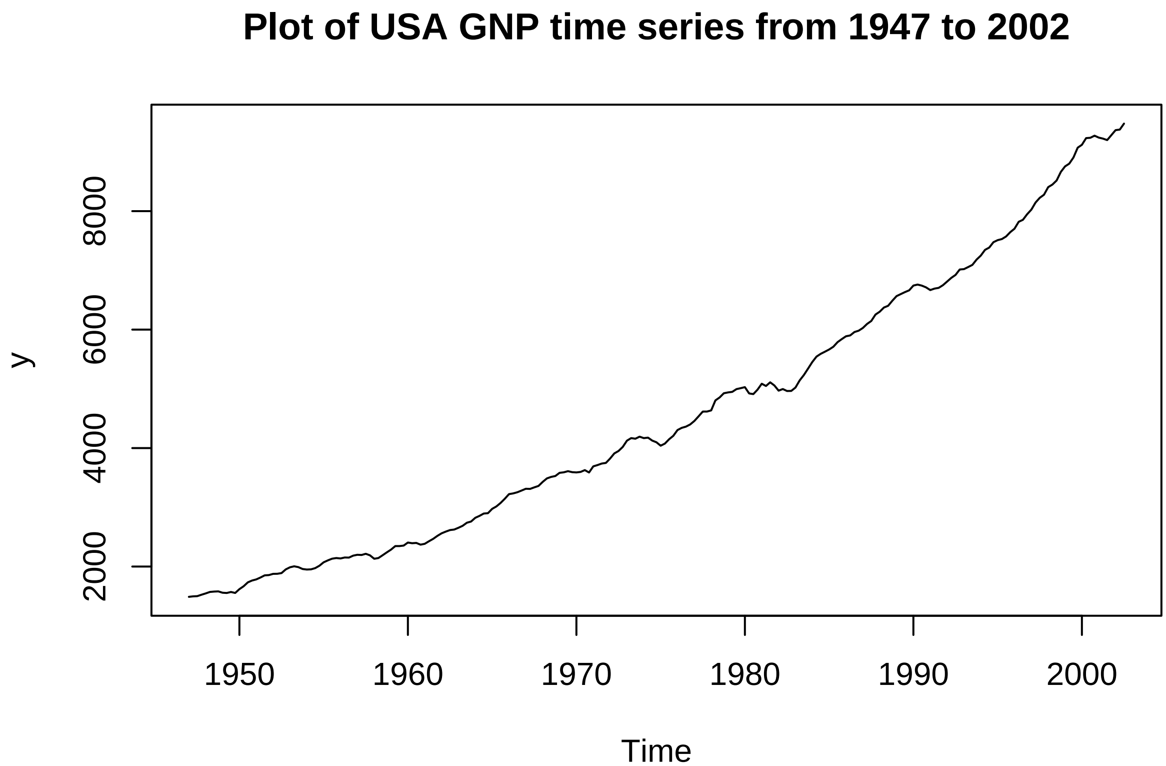
Difference: The ARIMA forecast is linear trend while the trend+noise forecast is non-linear trend. Besides, the prediction interval of trend+noise forecast is much narrower then ARIMA forecast. Therefore, I think that trend+noise forecast is better. Because the narrow forecast interval means that the forecast is more accurate.

Problem 5

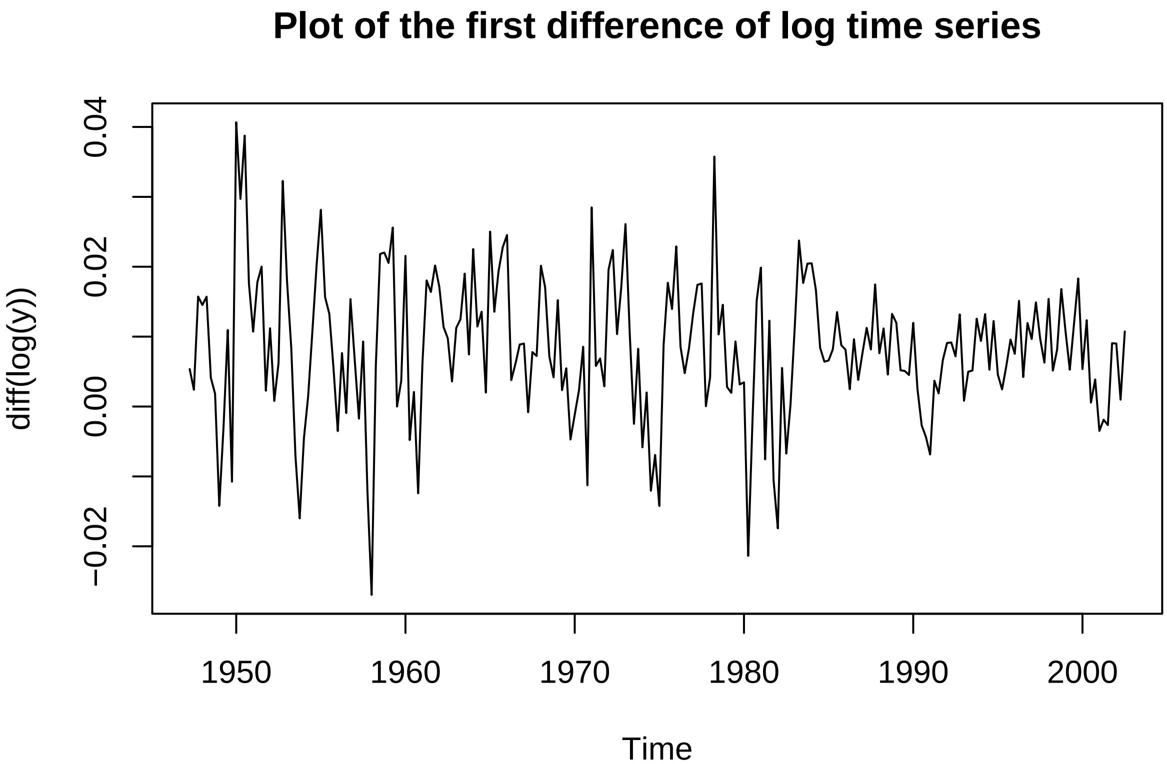
```
load("usGNP.RData")
```

(a)

```
# plot raw data
plot(y, main="Plot of USA GNP time series from 1947 to 2002")
```



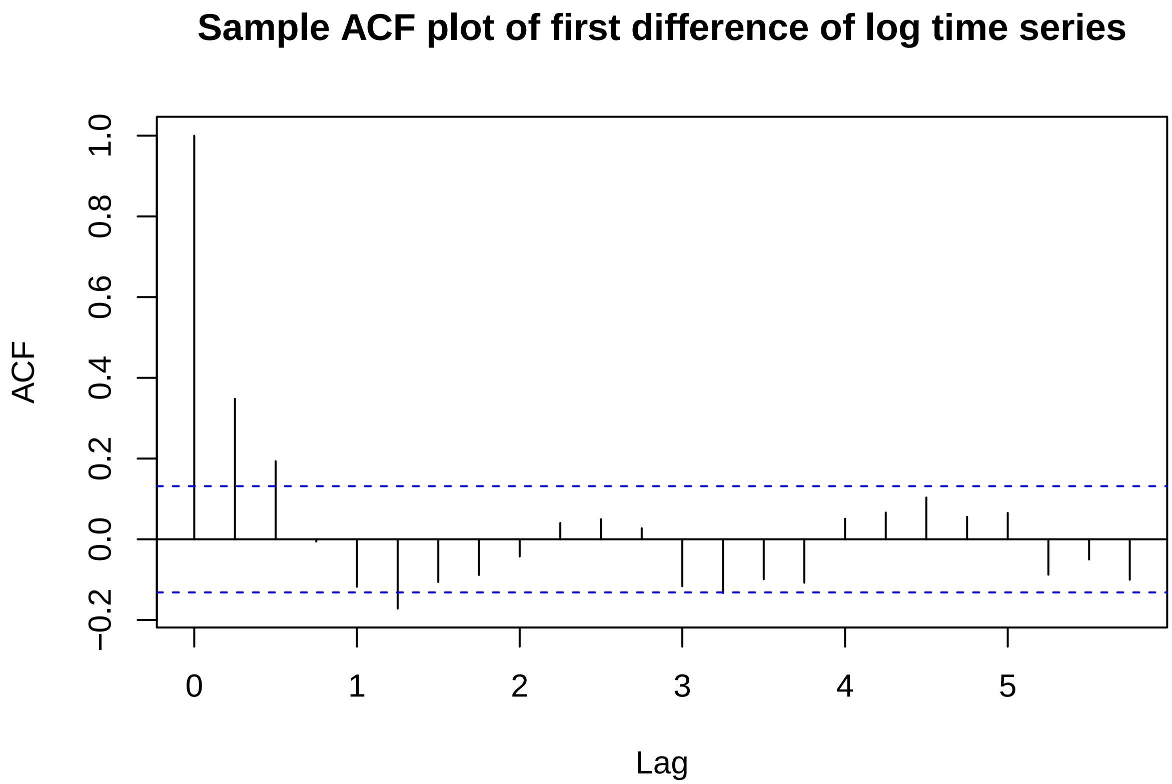

```
# plot first difference of log time series
plot(diff(log(y)), main="Plot of the first difference of log time series")
```



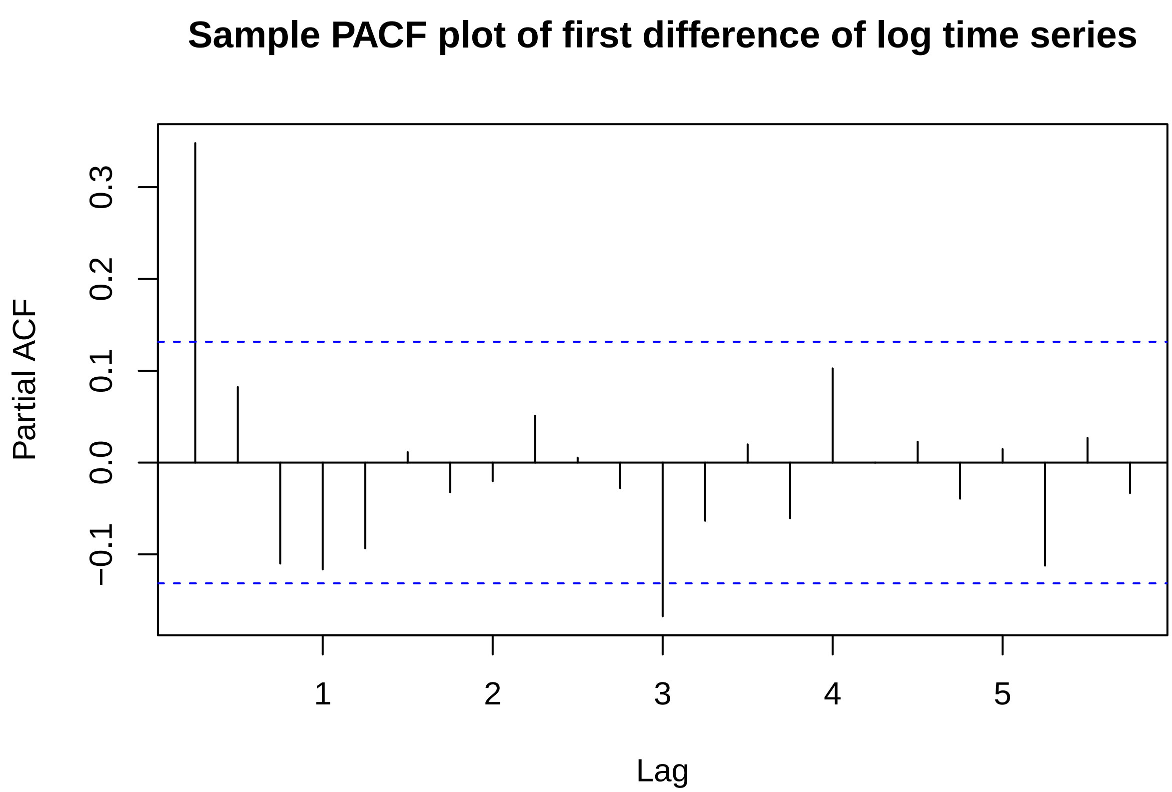

```
# Comments:
# Based on the plot of raw data, we found that there is a very obvious
# increasing trend, so the time series is non-stationary.
# Based on the first difference of log time series, we found that there is no
# obvious trend and seasonality so it looks relative stationary.
# The meaning of first difference of log(y) in real life is GNP's growth rate.
```

(b)

```
# ACF plot
acf(diff(log(y)), main="Sample ACF plot of first difference of log time series")
```




```
# PACF plot
pacf(diff(log(y)), main="Sample PACF plot of first difference of log time series")
```



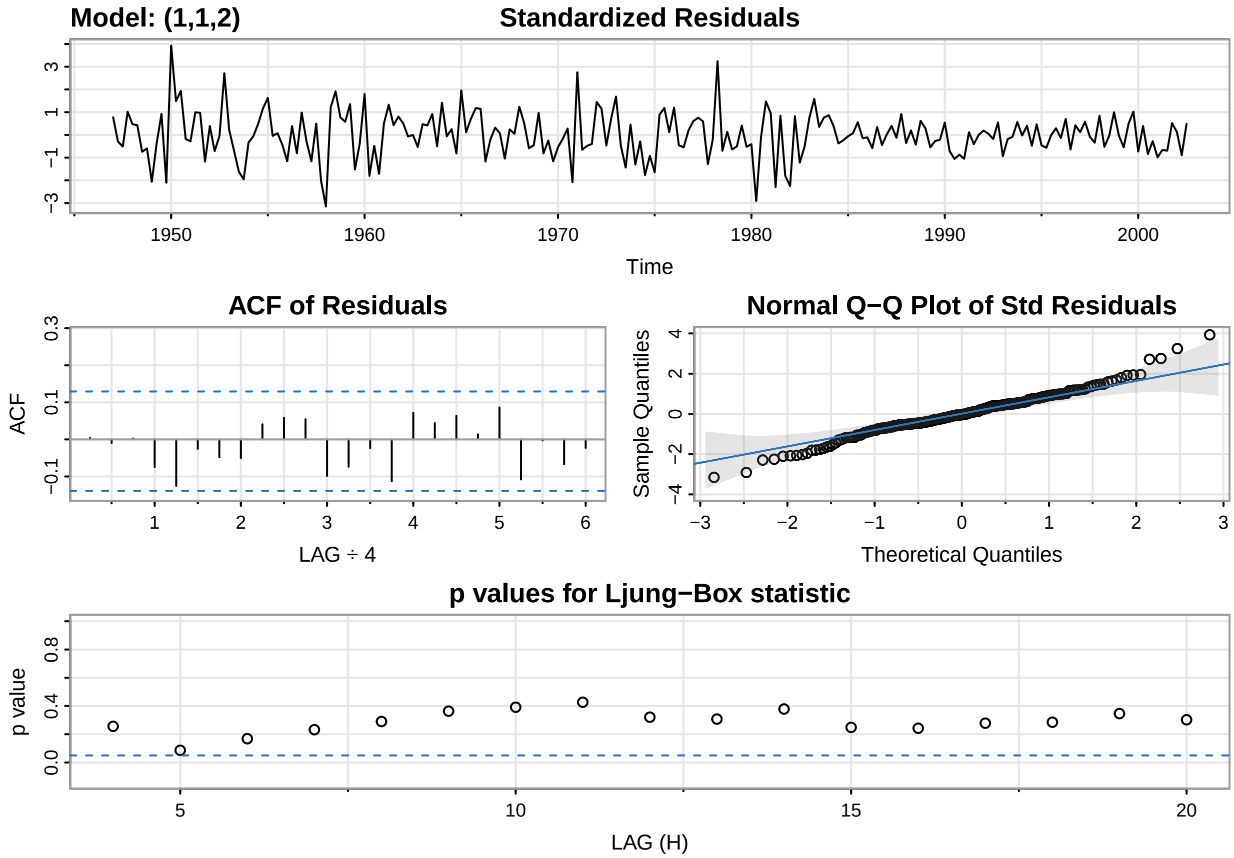

```
# Comments:
# Based on the sample ACF plot, we found that spikes cut off at lag2.
```

```
# Based on the sample PACF plot, we found that spikes cut off at lag1.
# Therefore, we suggest an ARIMA(1,1,2) model on the logrithmic data.
```

(c)

```
library(astsa)
model <- sarima(log(y), p=1, d=1, q=2)
```

```
## initial  value -4.589567
## iter   2 value -4.593469
## iter   3 value -4.661378
## iter   4 value -4.662245
## iter   5 value -4.662354
## iter   6 value -4.662395
## iter   7 value -4.662567
## iter   8 value -4.662643
## iter   9 value -4.662676
## iter  10 value -4.662678
## iter  10 value -4.662678
## final  value -4.662678
## converged
## initial  value -4.664307
## iter   2 value -4.664310
## iter   3 value -4.664311
## iter   4 value -4.664313
## iter   5 value -4.664315
## iter   6 value -4.664315
## iter   7 value -4.664316
## iter   8 value -4.664316
## iter   9 value -4.664316
## iter   9 value -4.664316
## iter   9 value -4.664316
## final  value -4.664316
## converged
```




```
model
```

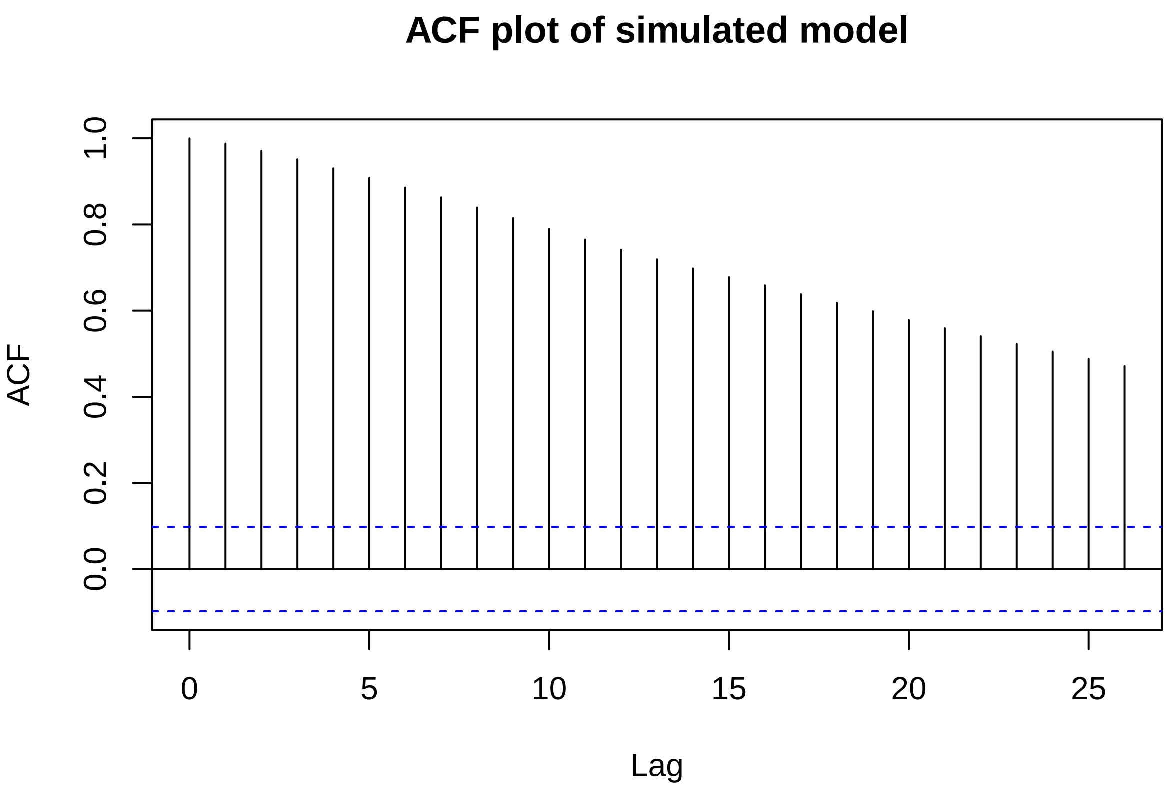
```
## $fit
##
## Call:
## arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D, Q), period = S),
##     xreg = constant, transform.pars = trans, fixed = fixed, optim.control = list(trace = trc,
##         REPORT = 1, reltol = tol))
##
## Coefficients:
##          ar1     ma1     ma2  constant
##       0.2407  0.0761  0.1623    0.0083
## s.e.  0.2066  0.2026  0.0851    0.0010
##
## sigma^2 estimated as 8.877e-05:  log likelihood = 720.47,  aic = -1430.95
##
## $degrees_of_freedom
## [1] 218
##
## $ttable
##          Estimate     SE t.value p.value
## ar1        0.2407 0.2066  1.1651  0.2453
## ma1        0.0761 0.2026  0.3754  0.7078
## ma2        0.1623 0.0851  1.9083  0.0577
## constant   0.0083 0.0010  8.0774  0.0000
##
## $AIC
```

```
## [1] -6.445711
##
## $AICc
## [1] -6.44488
##
## $BIC
## [1] -6.369074
```

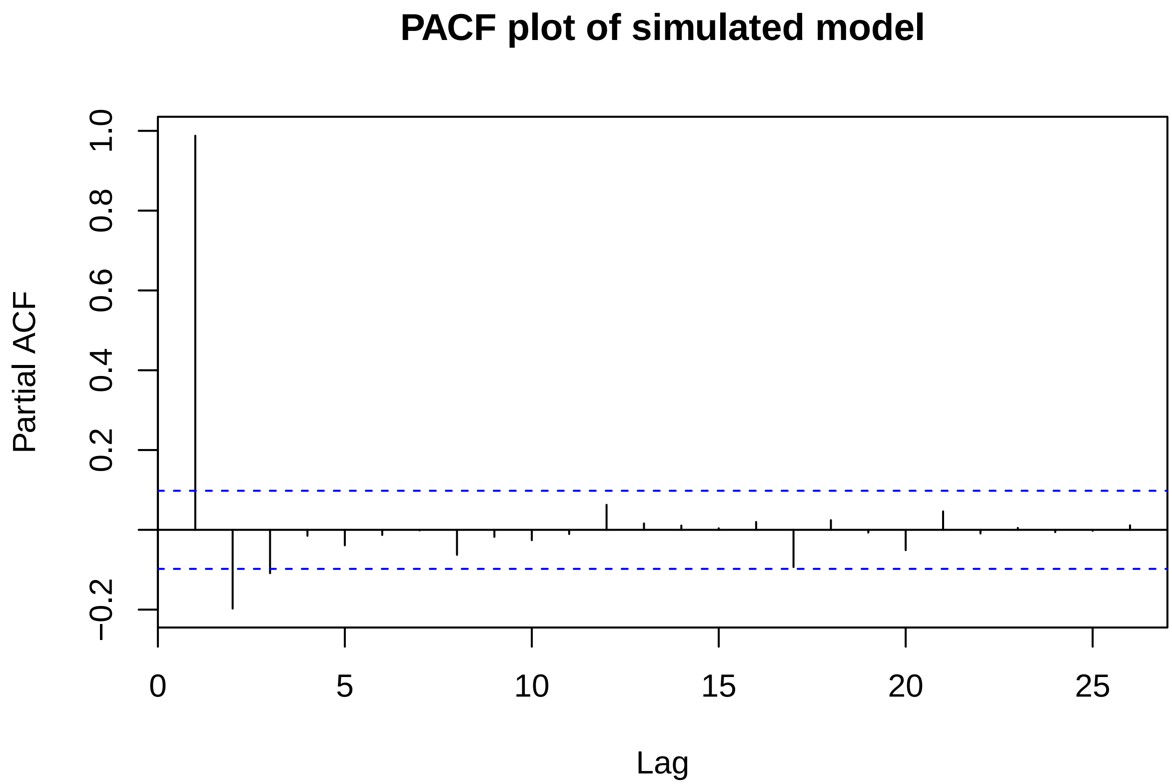
```
# diagnostics:
# Based on plot of standardized residuals, there is no obvious trend,
# seasonality.
# Based on plot of ACF, except lag0, there is no significant spike.
# Based on QQplot, most points distribute along qqline roughly, except a
# few points at two ends.
# Based on LjungBox test, p-value above 0.05 shows that there is no
# evidence to reject hypothesis of white noise.
```

(d)

```
# simulate ARIMA model
model_sim <- arima.sim(list(order=c(1,1,2),ar=0.2407,ma=c(0.0761, 0.1623)),
                       n=400)
acf(model_sim, main="ACF plot of simulated model")
```



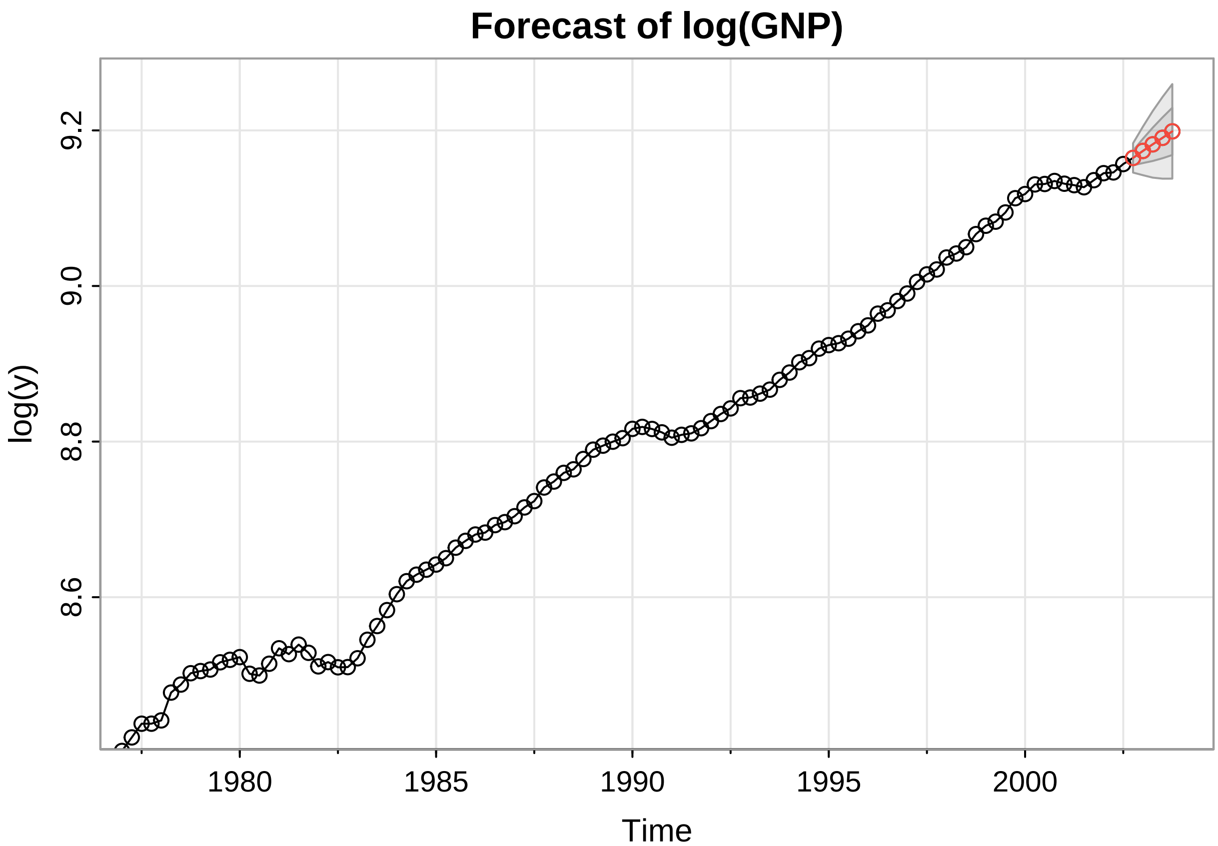

```
pacf(model_sim, main="PACF plot of simulated model")
```




```
# Comments:
# Based on the ACF plot, the values decay slowly.
# Based on the PACF plot, there are significant spikes at lag1 and lag2.
```

(e)

```
# forecast of log(GNP)
model_forecast <- sarima.for(log(y), n.ahead = 5, p=1, d=1, q=2,
                             main="Forecast of log(GNP)")
```




```r
# Switch to forecast of GNP
exp(model_forecast$pred)
```

```
##          Qtr1     Qtr2     Qtr3     Qtr4
## 2002                            9553.013
## 2003 9639.100 9721.190 9802.819 9884.854
```

```r
# construct 95% forecasted interval
c <- 1.96
# lower bound
lower_bound <- exp(model_forecast$pred-c*model_forecast$se)
# upper bound
upper_bound <- exp(model_forecast$pred+c*model_forecast$se)
# summary
cbind(model_forecast$pred, lower_bound, upper_bound)
```

```
##         model_forecast$pred lower_bound upper_bound
## 2002 Q4            9.164612    9378.215    9731.069
## 2003 Q1            9.173583    9349.228    9937.960
## 2003 Q2            9.182063    9322.100   10137.365
## 2003 Q3            9.190425    9311.473   10320.093
## 2003 Q4            9.198759    9313.366   10491.409
```