Forecasting US GNP based on selected time series data

Last name: Zhang
First name: Min
Student ID: 1004709201

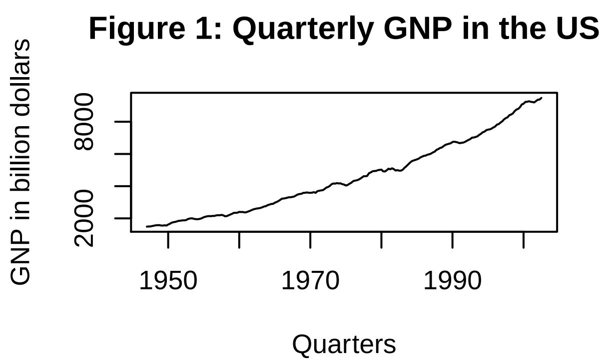
April 22, 2022

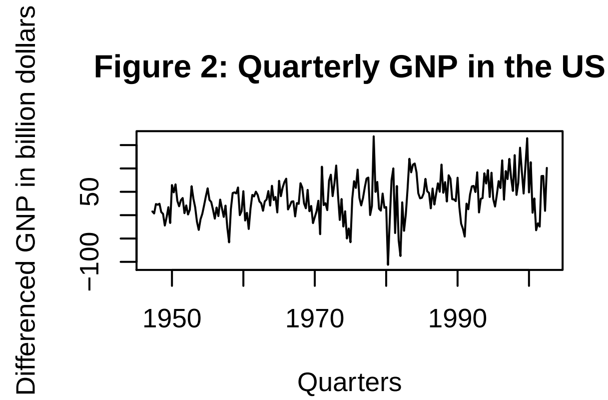
Abstract:

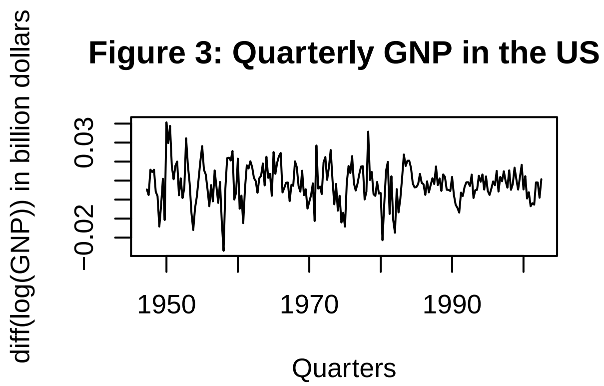
This academic project will be analyzing adjusted quarterly GNP (Gross National Product) in the United States from January 1947 to March 2002. Overall, we find that Gross National Product in billion dollars has an upward trend, with some setbacks in the economic recession cycle. We also projected GNP for the next 10 quarters by using selected ARIMA model, and the result is generally fit to the actual GNP number in the US.

Introduction:

Gross National Product, by definition, measures total values of all final products and services created by a country's residents domestically and overseas. This type of economic measurement led us to think about the validity of the model. Therefore, we used data set 'gnp' from package 'astsa' to illustrate quarterly US GNP after World War II until the beginning of 21st century. Since the data is recorded quarterly in a year, we are interested if there are some quarters that have higher GNP than the rest. When we firstly plotted data set 'gnp', we found it is not stationary which may affect our further analysis, we then did transformation to make it relatively stationary. We also manipulate time series data and forecast next 10 quarters of GNP to compare with the actual recorded numbers.
Statistical Methods:

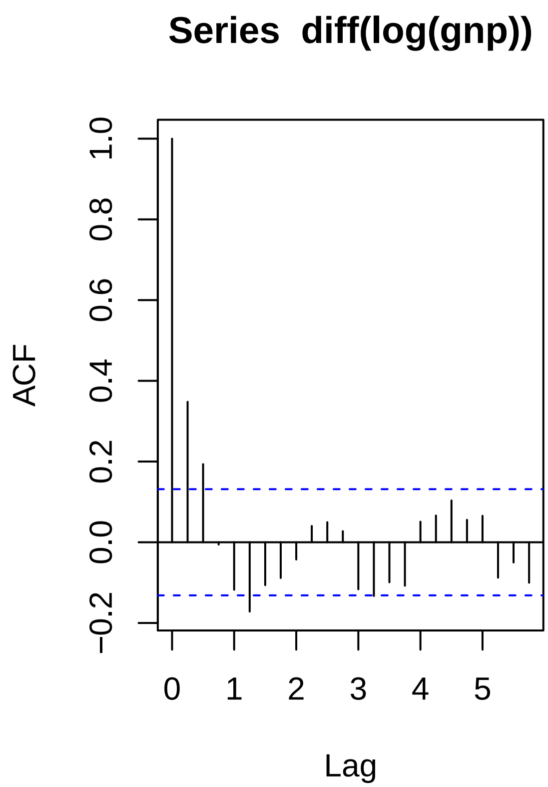


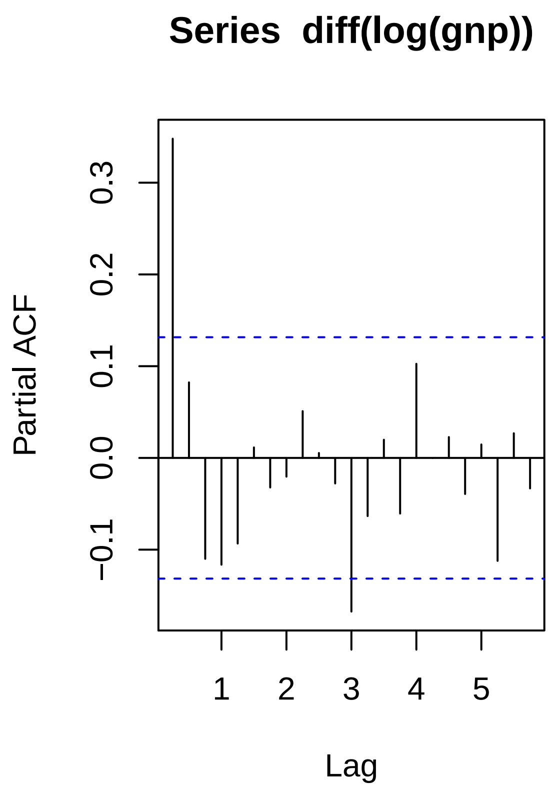




By illustrating the seasonally adjusted quarterly GNP in the US, we found out from Figure 1 that the original dataset 'gnp' is not stationary, as the mean of this dataset increases as time passes by. Thus, we chose to do a differencing shown in Figure 2 that the variance of this manipulated dataset changes over time, we therefore stabilized variance by applying a log transformation, and we used time series data from Figure 3 to investigate quarterly GNP in the US by forming a meaningful ARIMA model and forecasting based on this ARIMA model.

Results:

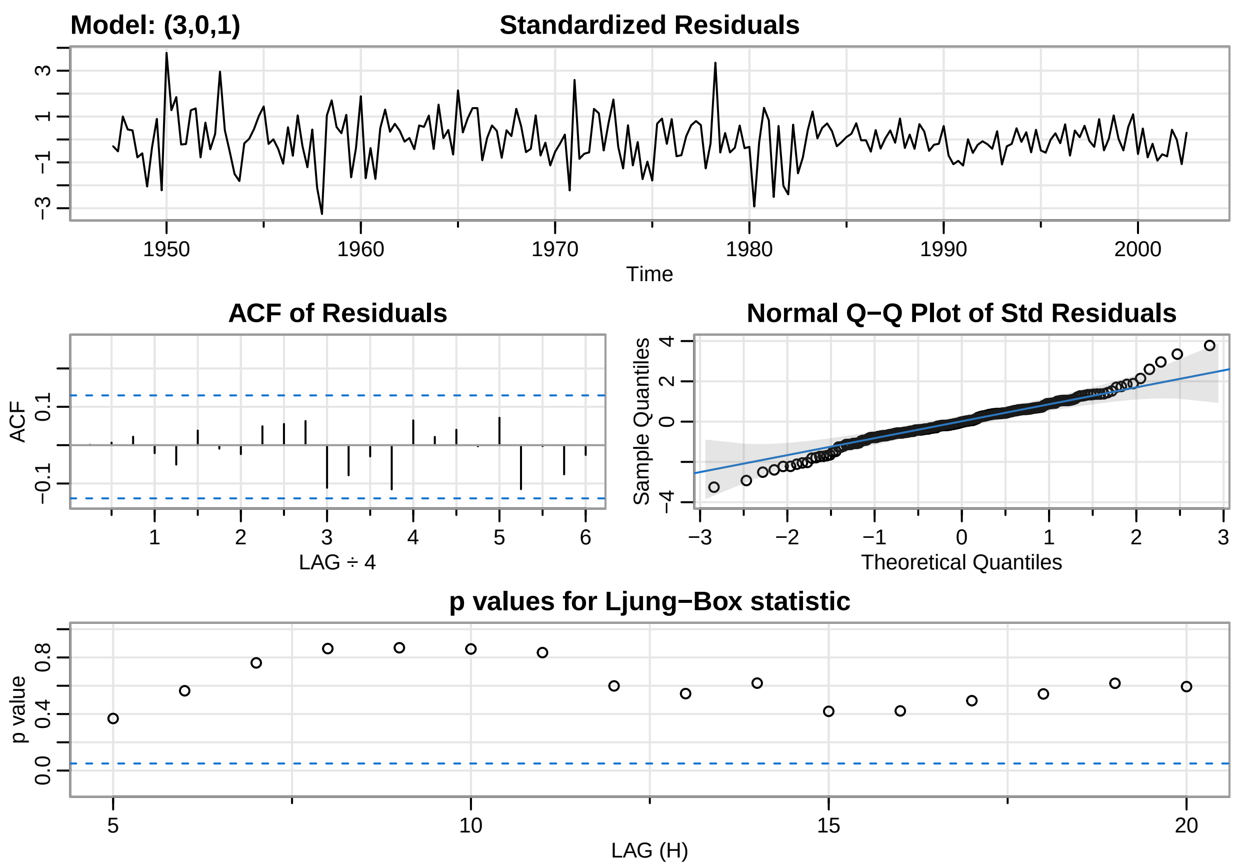




We proposed potential ARIMA models by plotting ACF and PACF of diff(log(gnp)). As shown above, we can observe that PACF cuts off after first and third lag while ACF cuts off after first lag. Thus, we selected two potential ARIMA models - ARIMA(1,1,1) and ARIMA(3,1,1) and will do a further diagnostic analysis to decide the final ARIMA model that will be used for forecast.

```
## initial  value -4.586998
## iter   2 value -4.587095
## iter   3 value -4.657097
## iter   4 value -4.659286
## iter   5 value -4.660052
## iter   6 value -4.660118
## iter   7 value -4.661984
## iter   8 value -4.666296
## iter   9 value -4.667301
## iter  10 value -4.668659
## iter  11 value -4.669466
## iter  12 value -4.669883
## iter  13 value -4.669955
## iter  14 value -4.670078
## iter  15 value -4.670107
## iter  16 value -4.670109
## iter  17 value -4.670111
## iter  18 value -4.670114
## iter  19 value -4.670118
## iter  20 value -4.670120
```

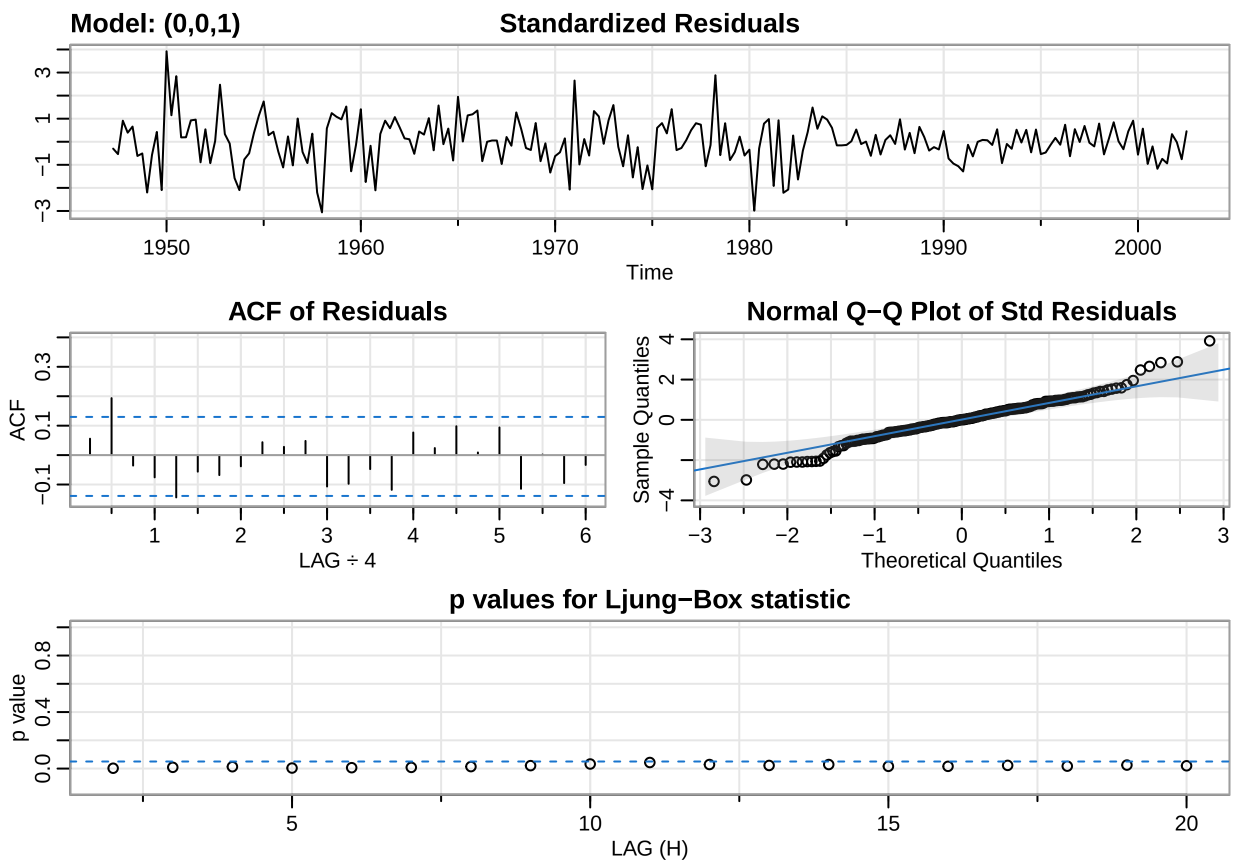
```
## iter  21 value -4.670120
## iter  21 value -4.670120
## iter  21 value -4.670120
## final  value -4.670120
## converged
## initial  value -4.674085
## iter   2 value -4.674119
## iter   3 value -4.674185
## iter   4 value -4.674396
## iter   5 value -4.674626
## iter   6 value -4.674884
## iter   7 value -4.675049
## iter   8 value -4.675160
## iter   9 value -4.675178
## iter  10 value -4.675187
## iter  11 value -4.675195
## iter  12 value -4.675202
## iter  13 value -4.675202
## iter  13 value -4.675202
## iter  13 value -4.675202
## final  value -4.675202
## converged
```




```
## $fit
##
## Call:
```

```
## arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D, Q), period = S),
##     xreg = xmean, include.mean = FALSE, transform.pars = trans, fixed = fixed,
##     optim.control = list(trace = trc, REPORT = 1, reltol = tol))
##
## Coefficients:
##          ar1      ar2      ar3      ma1   xmean
##       1.0905  -0.1251  -0.1739  -0.7881  0.0084
## s.e.  0.3028   0.1529   0.0777   0.3113  0.0006
##
## sigma^2 estimated as 8.682e-05:  log likelihood = 722.89,  aic = -1433.78
##
## $degrees_of_freedom
## [1] 217
##
## $ttable
##       Estimate     SE t.value p.value
## ar1     1.0905 0.3028  3.6016  0.0004
## ar2    -0.1251 0.1529 -0.8182  0.4141
## ar3    -0.1739 0.0777 -2.2389  0.0262
## ma1    -0.7881 0.3113 -2.5320  0.0120
## xmean   0.0084 0.0006 12.9429  0.0000
##
## $AIC
## [1] -6.458473
##
## $AICc
## [1] -6.457222
##
## $BIC
## [1] -6.366509
```

```
## initial  value -4.591629
## iter   2 value -4.636774
## iter   3 value -4.641036
## iter   4 value -4.641183
## iter   5 value -4.641184
## iter   5 value -4.641184
## iter   5 value -4.641184
## final  value -4.641184
## converged
## initial  value -4.641011
## iter   2 value -4.641011
## iter   3 value -4.641012
## iter   3 value -4.641012
## iter   3 value -4.641012
## final  value -4.641012
## converged
```

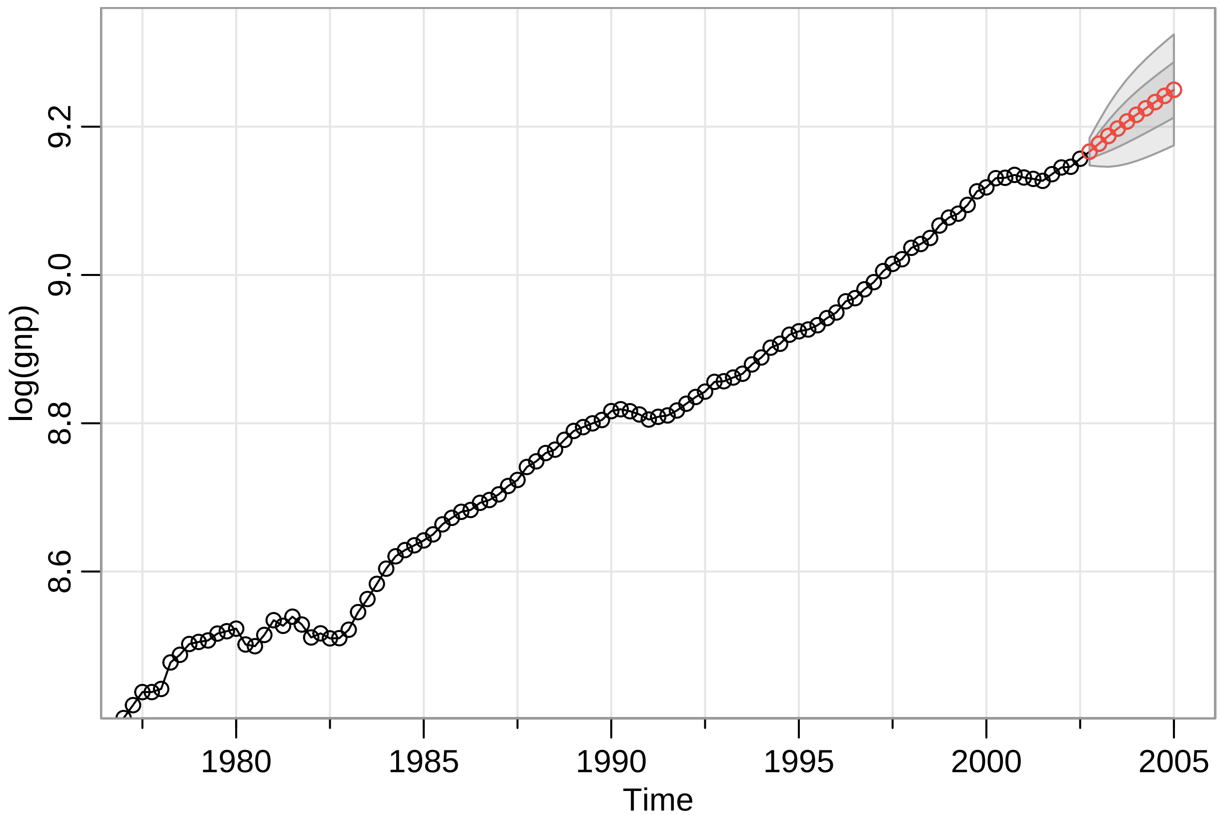



```
## $fit
##
## Call:
## arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D, Q), period = S),
##     xreg = xmean, include.mean = FALSE, transform.pars = trans, fixed = fixed,
##     optim.control = list(trace = trc, REPORT = 1, reltol = tol))
##
## Coefficients:
##          ma1   xmean
##       0.2719  0.0083
## s.e.  0.0549  0.0008
##
## sigma^2 estimated as 9.305e-05:  log likelihood = 715.3,  aic = -1424.6
##
## $degrees_of_freedom
## [1] 220
##
## $ttable
##       Estimate     SE t.value p.value
## ma1     0.2719 0.0549  4.9546       0
## xmean   0.0083 0.0008 10.1062       0
##
## $AIC
## [1] -6.417119
##
## $AICc
```

```
## [1] -6.416872
##
## $BIC
## [1] -6.371137
```

```
##              AIC       BIC
## model1 -6.458473 -6.366509
## model2 -6.417119 -6.371137
```

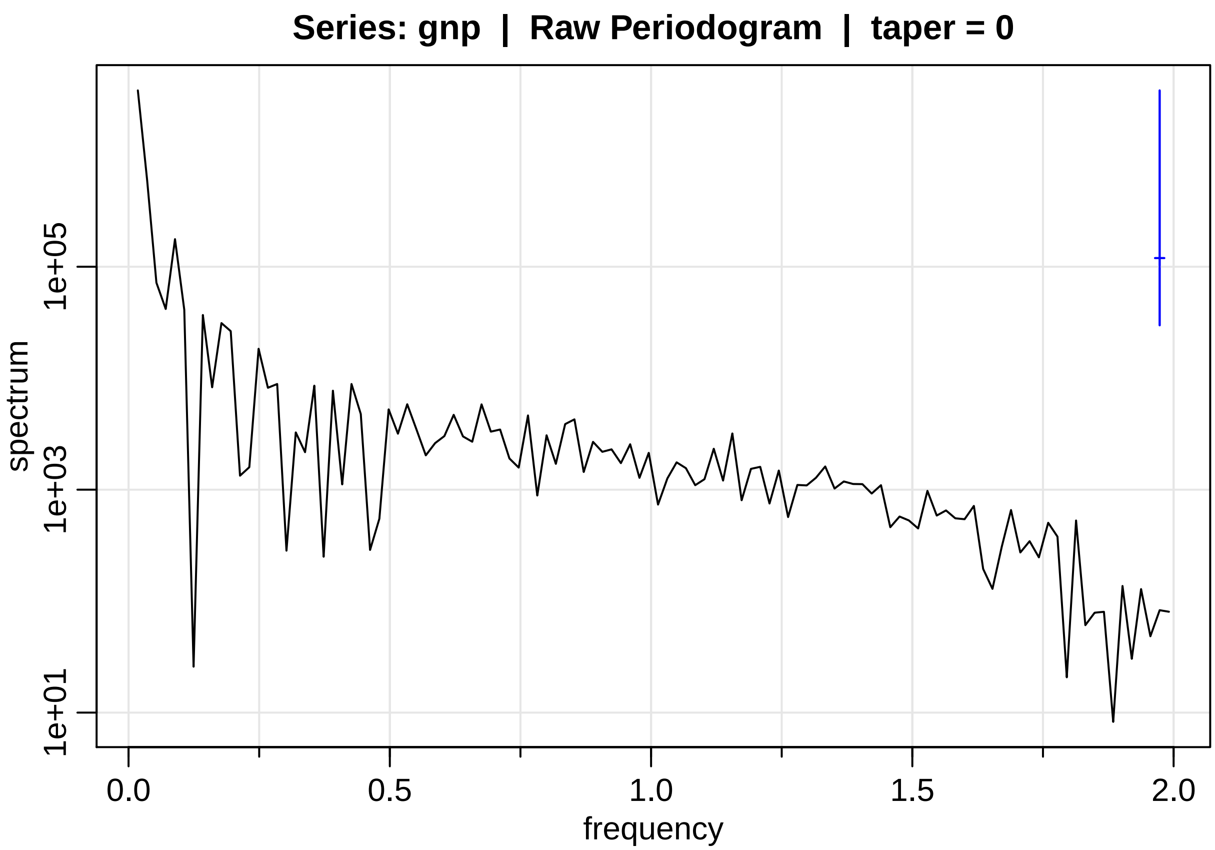
To decide which ARIMA model for forecast, we used sarima knowledge to plot diagnostics plots. As observed from model(3,0,1) that the residuals show no significant auto correlations, as they are all cut off. The QQ plot shows the condition of residuals that normal distribution is satisfied, with some potential outliers at both tails. Moreover, the p-values for the Ljung-Box plot are all above the threshold of 0.05, suggesting no significance for residuals which is a good fit for the model. In contrast, the ACF residuals of model(0,0,1) did not entirely cut off and the p-values for the Ljung-Box plot suggests the correlation of residuals as the values are mostly below 0.05 cutoff. The AIC and BIC of both models show significant similarity. Thus, we decide to choose ARIMA(3,1,1) to further investigate and forecast.




```
##              Lower     Upper
## 2002 Q4 9396.409  9745.953
## 2003 Q1 9387.068  9967.255
## 2003 Q2 9383.875 10179.673
## 2003 Q3 9397.903 10370.664
## 2003 Q4 9424.642 10540.251
## 2004 Q1 9461.766 10691.251
## 2004 Q2 9506.313 10828.052
## 2004 Q3 9556.159 10955.054
```

```
## 2004 Q4 9609.753 11076.011
## 2005 Q1 9666.059 11193.813
```

As a result, we will use model ARIMAR(3,1,1) to forecast next 10 quarters of GNP in the US, with 95% confidence level. It can be seen that GNP in the US would have a gradual increase in next 10 quarters since March 2002, and we are 95% confident that the quarterly GNP will fall into the range listed above.




```
##    frequency       period     spectrum
##       0.0178      56.2500 3819548.6471
```

```
##   frequency      period    spectrum
##      0.0356     28.1250 598276.5530
```

```
##   frequency      period    spectrum
##      0.0889     11.2500 176734.1797
```

```
##  spectrum
## 150864113
```

```
## spectrum
##  1035422
```

we also performed a periodogram analysis listed above and found the first three predominant periods, with 95% confidence interval at frequencies 0.0178, 0.0356 and 0.0889. Overall, the periodogram has a continuous decreasing trend.

Discussions:

There are limitations in this model that can be improved if we have a larger dataset. First of all, in the previous QQ plot, there are some potential outliers that lie at both tail which may influence the dependence order of ARIMA model. We thus could try to identify these outliers and remove those extreme values. Throughout the model selection and forecast, we found that the quarterly GNP in the US from 1947 to 2002 depends on the economic performance both globally and domestically, as GNP measures both final products and services produced in the country and overseas. Thus, during economic recession periods, GNP in the US would also decrease. Moreover, we could potentially forecast more time periods as more forecast data would allow us to verify the model accuracy. For example, we could forecast the GNP during 2008 financial crisis period and to see if our model could accurately predict the quarterly GNP during that time.