**Midterm Exam**

**Oliver Yuan**

**4/10/2017**

This exam is open book, open notes. However, you must complete the exam on your own. You may not work with other students to complete the exam.

All of the information needed to answer the first two questions has been provided. You do not need to do any additional analysis in order to address the questions. However, I have included the International Bank Loan data set (from question 2) in case it helps you to be able to look at the data on your own.

Question 3 requires you to analyze the Steel series. The Steel data set is available in SAS and .csv form on Sakai. You may use whatever software you prefer to analyze the data. You may use any of the techniques that we have discussed to analyze the data.

If you have questions during the exam period, please don’t hesitate to e-mail me, and I will do my best to respond promptly.

NOTE: We will **not** be meeting in the classroom on Tuesday, April 10th to give you time to work on the exam.

The exam is due Wednesday, April 11th at 1:00pm. You should submit the exam on Sakai.

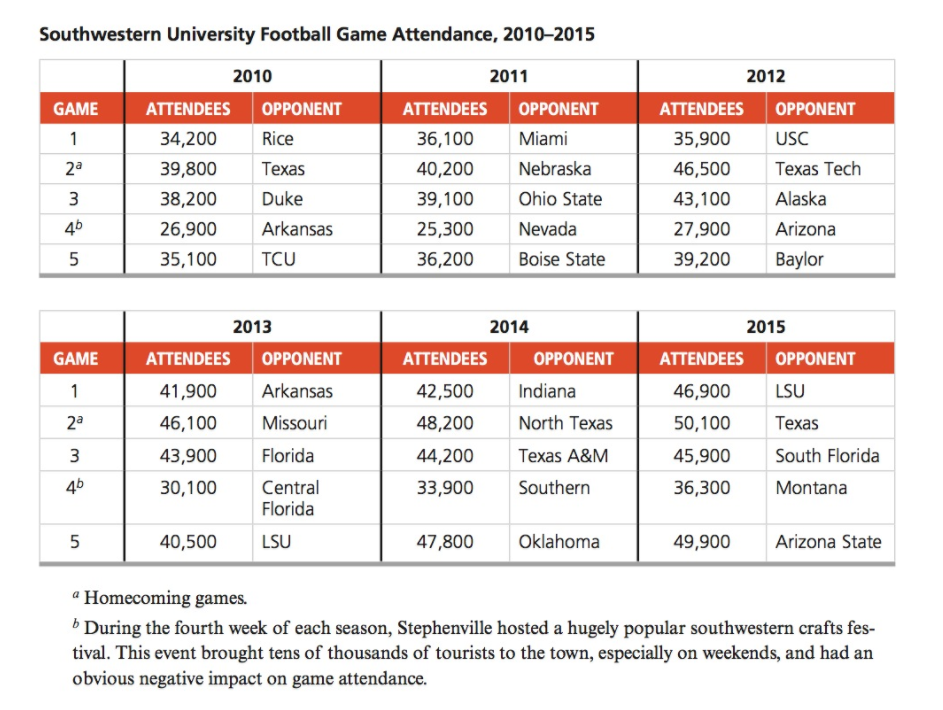
1. Southwestern University (SWU), a large state college in Stephenville, Texas, enrolls close to 20,000 students. The school is a dominant force in the small city, with more students during fall and spring than permanent residents. Always a football powerhouse, SWU is usually in the top 20 in college football rankings. Since the legendary Phil Flamm was hired as its head coach in 2009 (in hopes of reaching the elusive number 1 ranking), attendance at the five Saturday home games each year increased. Prior to Flamm’s arrival, attendance generally averaged 25,000 to 29,000 per game. Season ticket sales bumped up by 10,000 just with the announcement of the new coach’s arrival. The immediate issue facing SWU, however, was not NCAA ranking. It was capacity. The existing SWU stadium, built in 1953, has seating for 54,000 fans. The following table shows attendance at each game for 2010-2015. One of Flamm’s demands upon joining SWU had been a stadium expansion, or possibly even a new stadium. With attendance increasing, SWU administrators began to face the issue head-on. SWU’s president decided it was time for his vice president of development to forecast when the existing stadium would “max out” (e.g., consistently exceed maximum capacity.
2. In order to develop a forecasting model that could be used to project attendance for the next 3 – 5 years, you first need to choose a time interval for the forecast. Based on the information provided, what time interval would you choose as the basis for the forecast? (5 points)

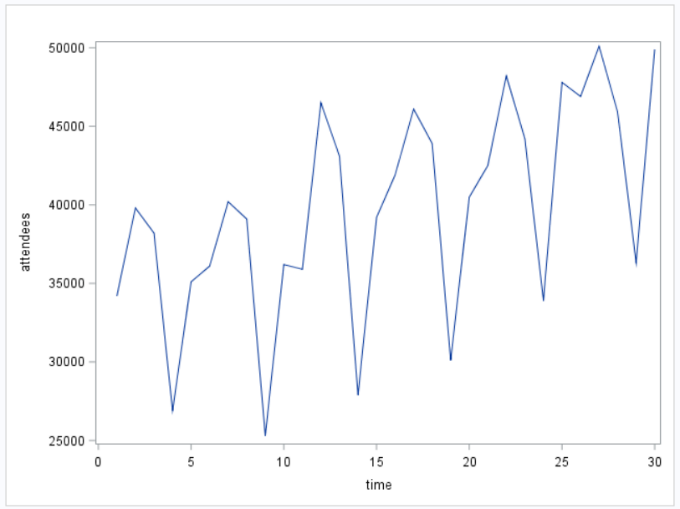
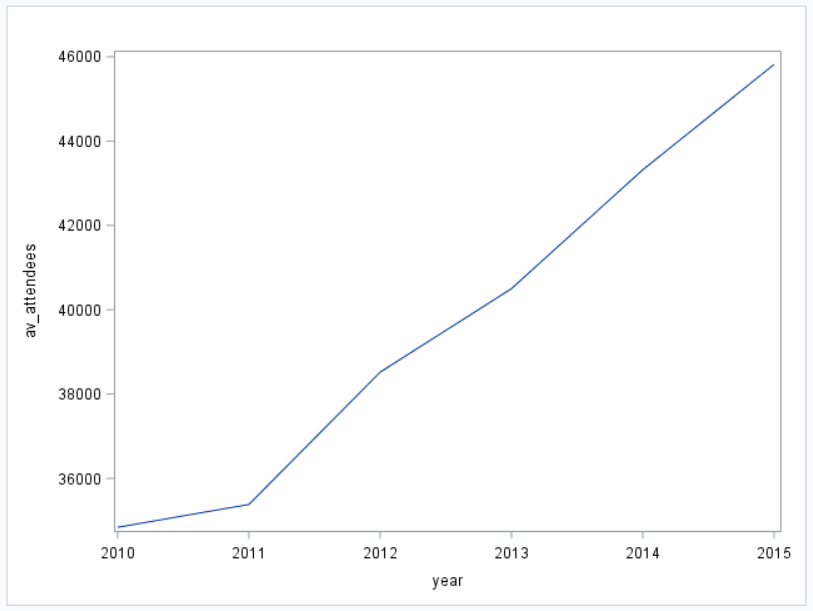
**I will choose to aggregate data by year. As we know, NCAA games are held every year so season tickets are sold every year. Therefore, we can observe change from season to season by aggregating data yearly. What’s more, we are developing model to project attendance for the next 3 – 5 years, which indicates that we want to forecast attendance for the whole season. In addition, we don’t have information about when games are held within a year so aggregating by year seems to be our only choice.**

1. Suppose that you also need to aggregate the data to the desired time interval before building the forecast model. You can aggregate the data using the total, mean, minimum, or maximum attendance. Given the objectives of the university, what aggregation method would you choose and why? (5 points)

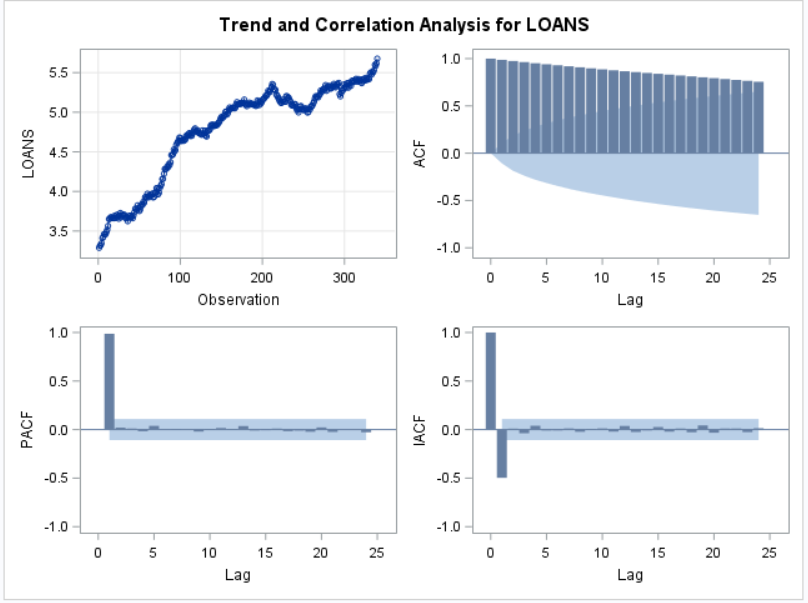
**I will aggregate the data using mean. Our goal is to forecast when attendance will consistently exceed maximum capacity. Using total or maximum attendance is not appropriate because it can’t reflect most of situations. If we use minimum attendance, we will have already exceeded capacity for a while when our prediction exceeds total capacity. By using mean attendance, we can get a sense of overall attendance in a year, which can indicate if attendance exceed maximum capacity most of the time.**

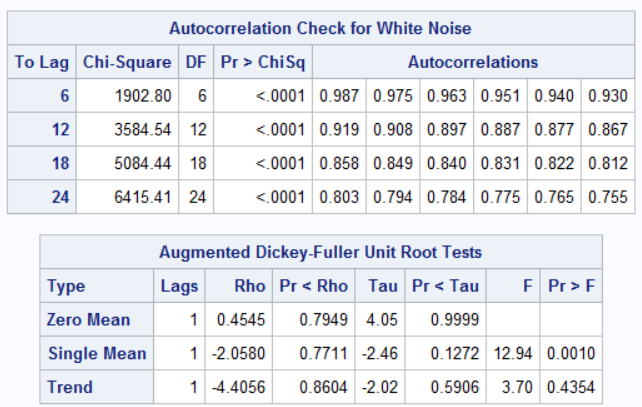
1. Based solely on the information provided below, what other concerns might you have about using these data to forecast attendance at home games for the next 3 – 5 years? (5 points)
2. **As we can see from the data, attendance will drop in the 4th game in each year due to the crafts festival. By aggregating data, we may ignore this pattern, which certainly has effect on our attendance.**
3. **The original data contains 30 points. By aggregating data by year, we only have 5 data points which we can build our model on. Without sufficient data, the model will become less reliable.**
4. **Who opponents are will affect attendance. We can’t take opponent into account when we use forecast model.**



1. The following output shows the series plot, ACF, PACF, and IACF for the *log* of the monthly interbank loans in billions of dollars. The data were extracted from the Federal Reserve website.





1. Based on the output shown above, does this series appear to be white noise? Be sure to support your answer with information from the output. (8 points)

**This series is not white noise. We can see that all p-values in autocorrelation check for white noise are significant. Since the null hypothesis is the series is stationary, we can reject the null hypothesis based on the result of the test.**

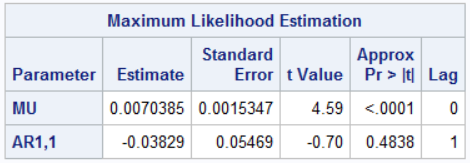
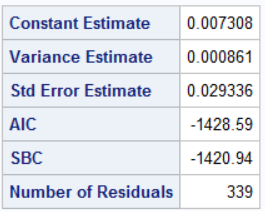
1. Based on the output shown above, does the series appear to be stationary or non-stationary? Give two pieces of evidence from the output to support your answer. (10 points)

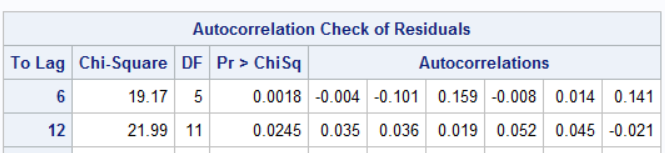
**From the ACF plot, we can see that autocorrelations decay to zero very slowly, which indicates that the series is non-stationary. Also, if we check the result of D-F test, we can see that the p-value for trend is greater than 0.05, which indicates that the series is non-stationary.**

Now, suppose that you take the first difference of the log loan value series. The series plot, ACF, PACF, and IACF for the differenced series is shown below. Note that the Ljung-Box test indicates that this series is not white noise, so we need to determine the values of p and q for the model that best fits the series.

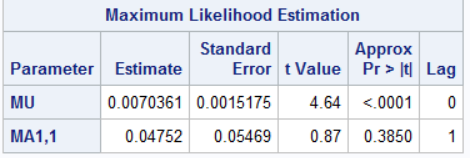
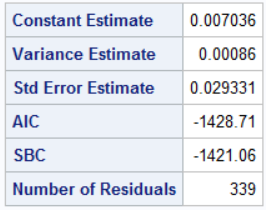
In order to help determine the “best” values of p and q, I have fit the following models to the data: (p=1, q=0), (p=0, q=1), (p=3, q=0), (p=0, q=3), (p=1, q=1), (p=1, q=3). The output from each model is shown below, and a table summarizing the AIC, SBC, RMSE, and MAPE for each model is included at the end.

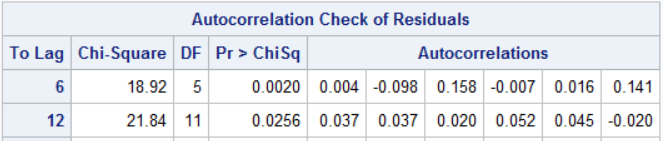
**Model 1: p=1, q=0**

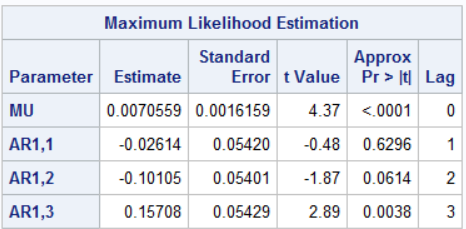
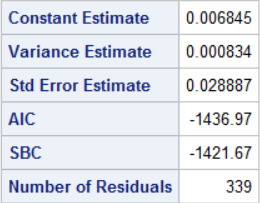


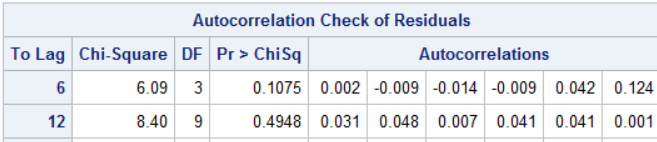
**Model 2: p=0, q=1**

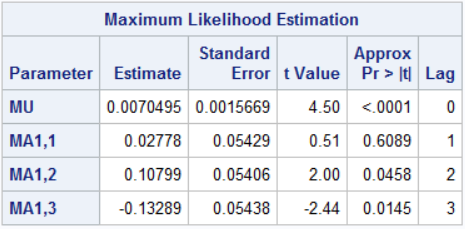
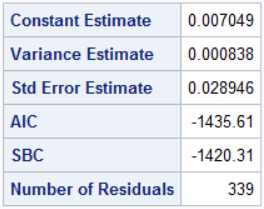


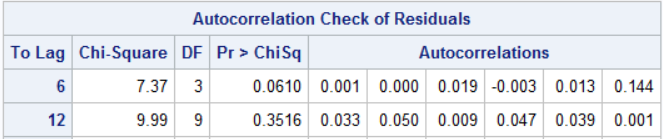
**Model 3: p=3, q=0**

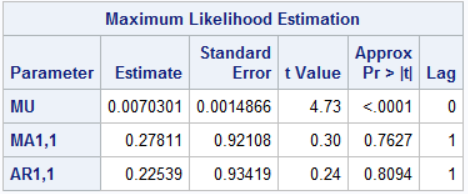
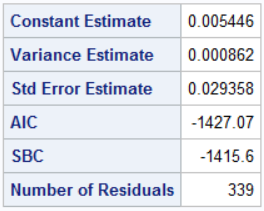


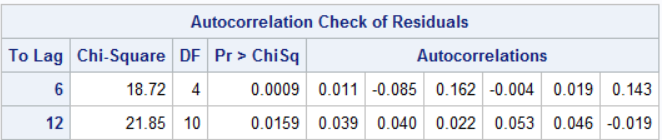
**Model 4: p=0, q=3**

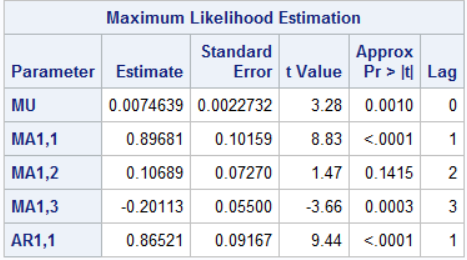
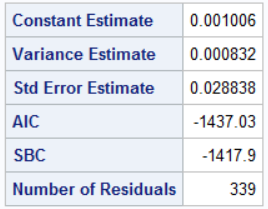


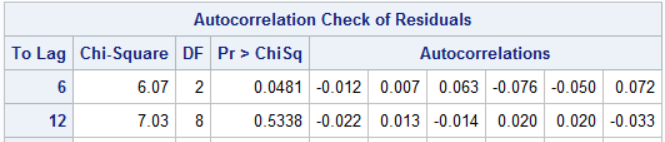
**Model 5: p=1, q=1**



**Model 6: p=1, q=3**

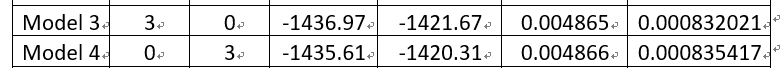


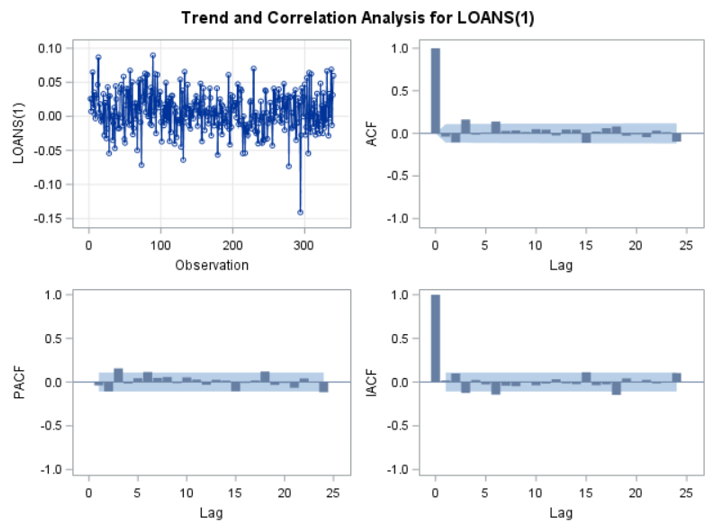
**Summary Table for Model Comparison and Selection:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **p** | **q** | **AIC** | **SBC** | **MAPE** | **RMSE** |
| Model 1 | 1 | 0 | -1428.59 | -1420.94 | 0.004901 | 0.000858047 |
| Model 2 | 0 | 1 | -1428.71 | -1421.06 | 0.004901 | 0.000857751 |
| Model 3 | 3 | 0 | -1436.97 | -1421.67 | 0.004865 | 0.000832021 |
| Model 4 | 0 | 3 | -1435.61 | -1420.31 | 0.004866 | 0.000835417 |
| Model 5 | 1 | 1 | -1427.07 | -1415.6 | 0.004066 | 0.000859347 |
| Model 6 | 1 | 3 | -1437.03 | -1417.9 | 0.004839 | 0.000829493 |

1. List three reasons why Model 5 should not be recommended as a viable model for this time series. (12 points)
2. **Model 5 has the largest AIC and SBC.**
3. **Both the AR and the MA terms in this model is not significant.**
4. **Based on the autocorrelation test on residual of model 5, the residual is not white noise.**
5. Of the remaining models, which model would you recommend as being the “best” model? Why? Support for your answer using the information provided above. (10 points)

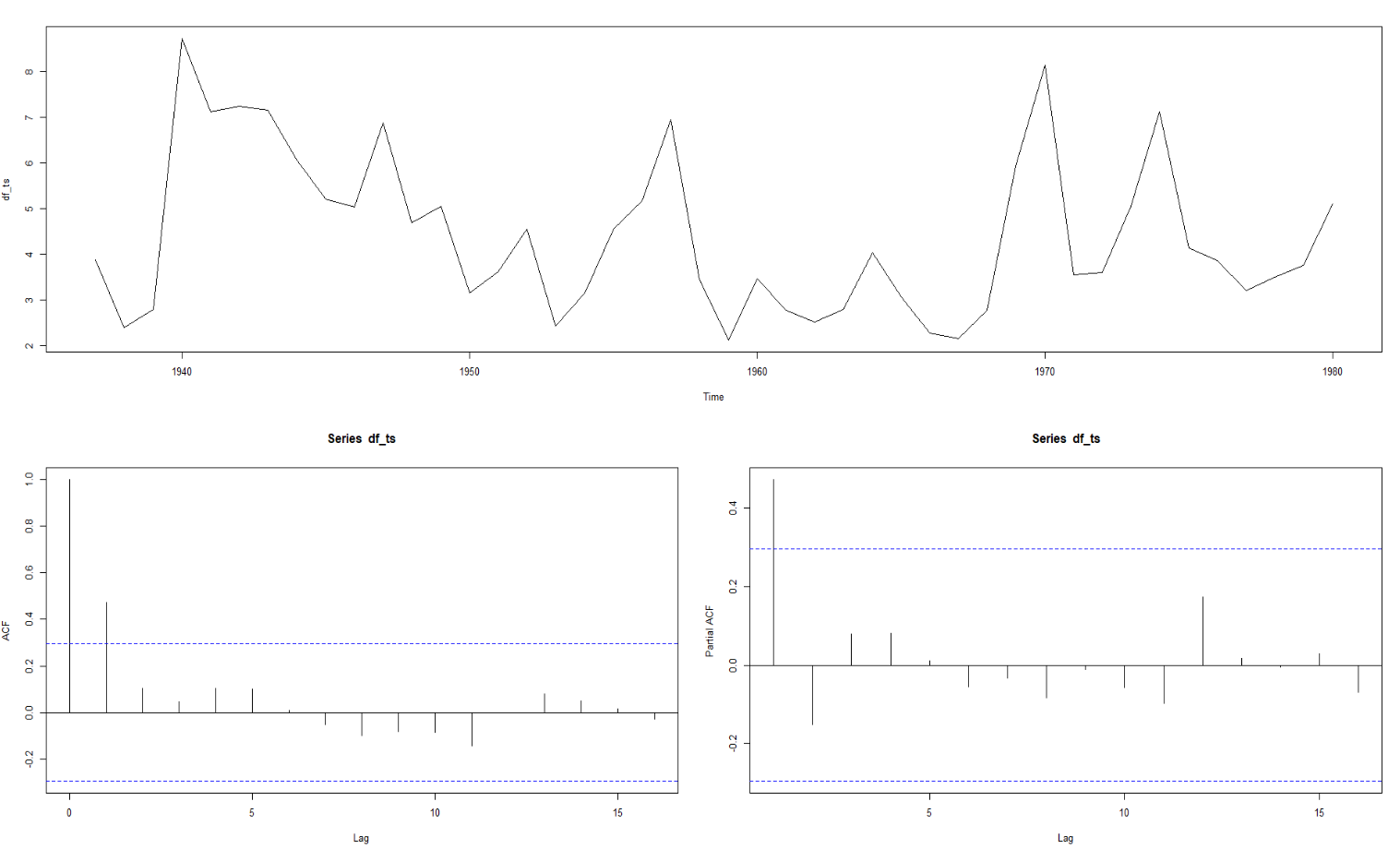
**I will exclude model 1, model 2 and model 6 directly because their residuals are not white noise. Model 3 is better than model 4 since model 3 has smaller AIC, SBC, MAPE and RMSE. To sum up, I will choose model 3 as my final model.**



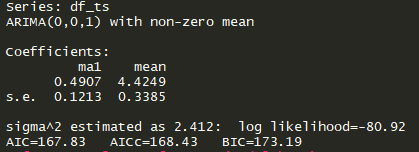


1. Download the U.S. iron and steel export yearly series (Fairchild Publications 1981) from Sakai. The data set is available in both SAS (steel.sas7bdat) and .csv (mt\_steel.csv) format. The series is composed of the weight (in million tons) of iron and steel exports (excluding scraps) from 1937 to 1980. For the purposes of this exercise, assume that the series is stationary (e.g., do not difference the series). Based on these data:

* Choose a model for the data (e.g., determine the value of p and q)
* Estimate the model
* Use the model to forecast iron and steel exports for 1981, 1982, and 1983



**Based on the ACF and PACF plot above, we can see the second term in ACF plot is significant, which suggests a MA(1) term. After I used auto.arima function in R, I got result below:**

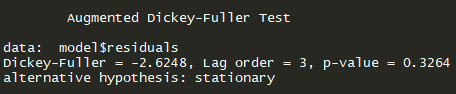
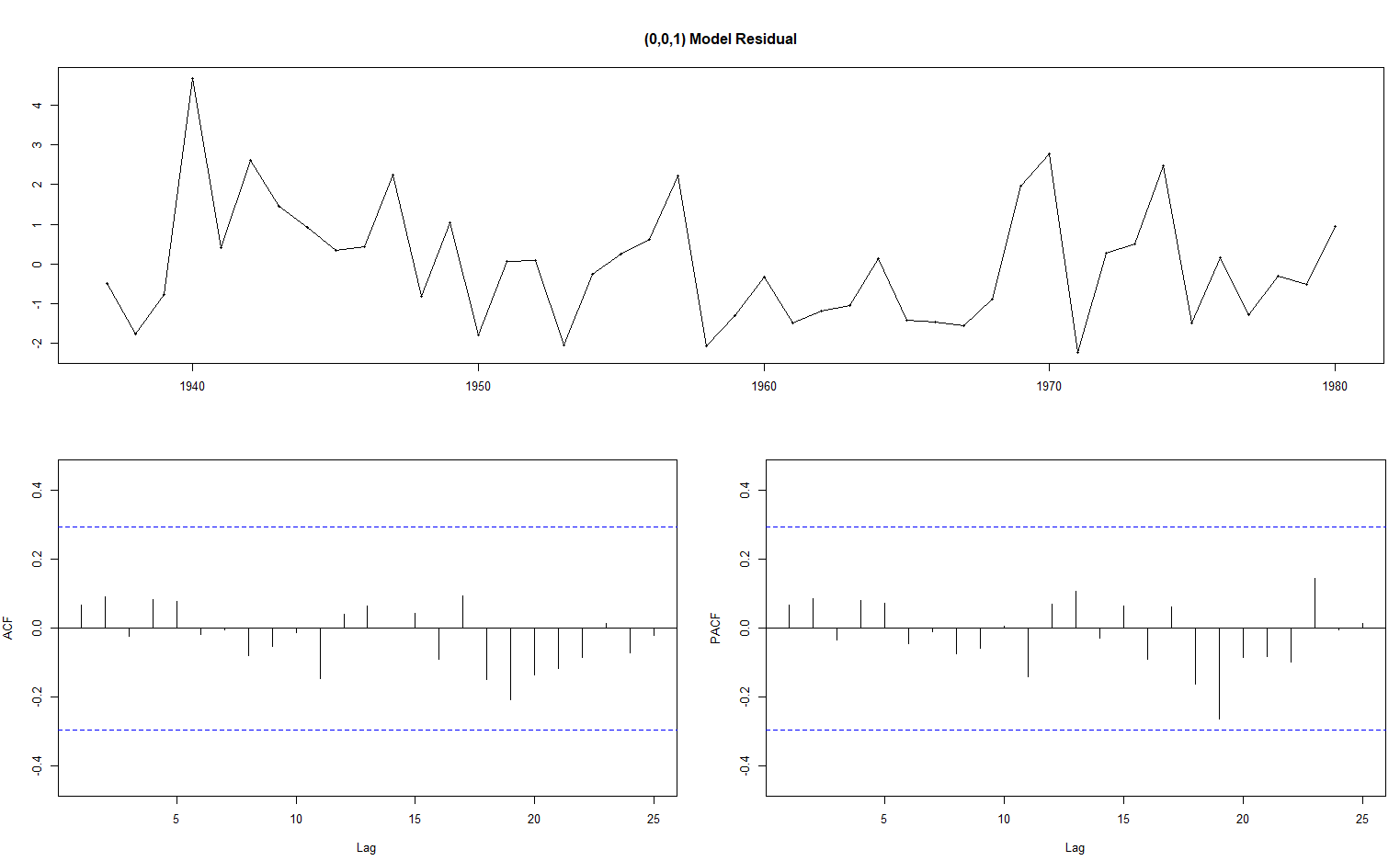


**Therefore, I choose (0,0,1) as (p,d,q) to fit my model.**

**Accuracy Table**

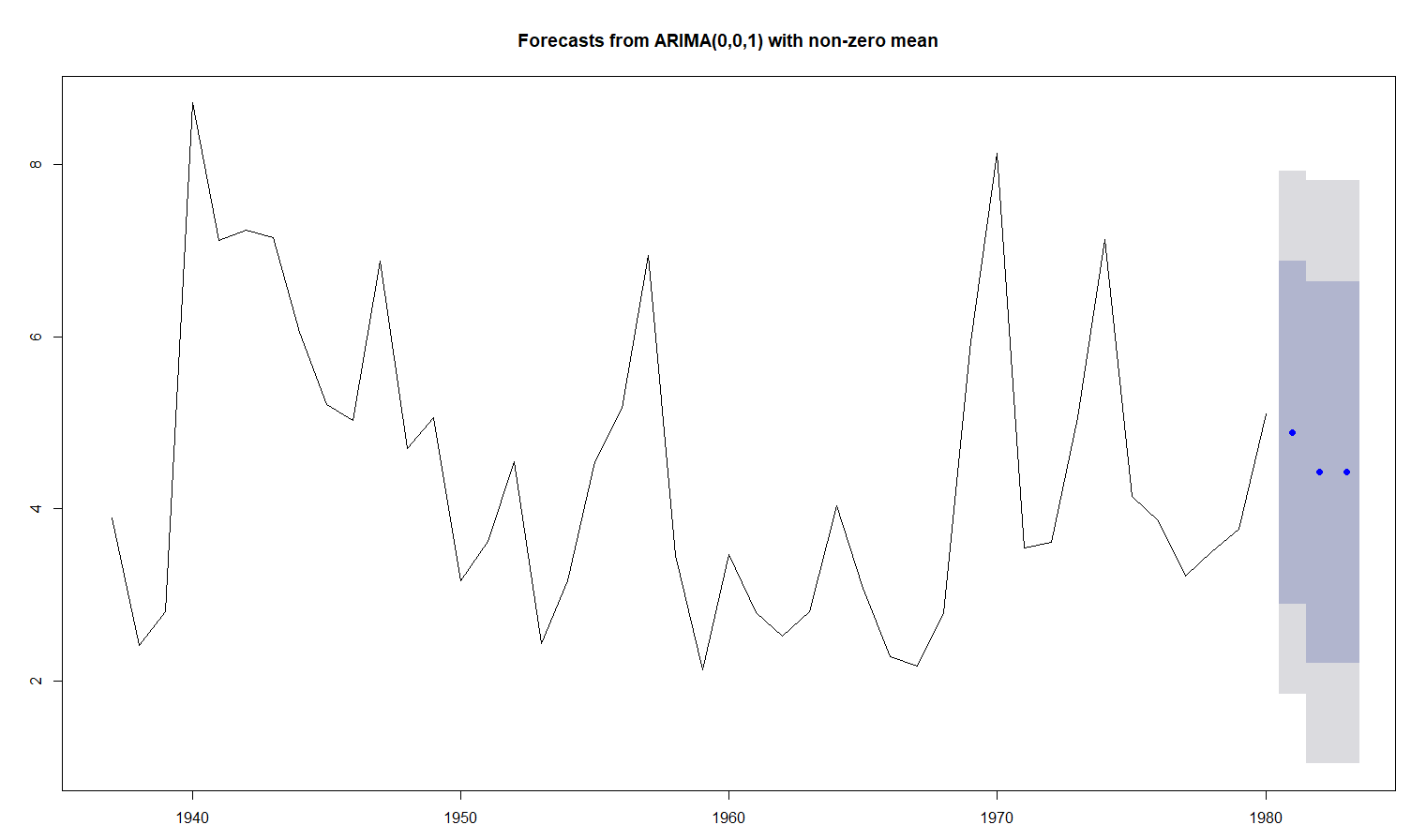
****

**Residual:**

****

**Based on residual plot and the result of D-F test(p-value = 0.3264), our residual is white noise, which is great.**

**Prediction:**

****

|  |  |  |
| --- | --- | --- |
| **1981** | **1982** | **1983** |
| **4.886352** | **4.424851** | **4.424851** |

Your answer to this question should include the values of p and q along with a series plot, plots of the ACF and PACF, and some justification for why you chose these values (note that this could be as simple as referencing an automated selection technique – it does not need to be long). You should also include a table of parameter estimates, at least one measure of model accuracy (AIC, SBC, RMSE, or MAPE), and a test of white noise on the residuals from your model. Finally, you should include a table showing the forecast values for the 3 times periods requested above.

You may choose which software you wish to use to answer this question.

Question 3 is worth a total of 45 points.