Forecasting National bankruptcy rates

Sheri Nguyen, Juan Pablo Oberhauser, Evelyn Peng

December 13, 2016

Contents

Problem Description

According to the Office of the Superintendent of Bankruptcy Canada, "The consumer insolvency rate is defined as the number of consumer insolvencies per thousand residents aged 18 years or above." bankruptcy rates have gained prevalence and importance sine the 2008 economic downturn since most of the issues arising in that time came from unpaid mortgages and the bundling of sub-prime loans that were unpaid. Being able to accurately predict economic indicators is paramount to understand and prepare for many macro- and micro-economic events.

In the following sections, we will utilize bankruptcy rates from the country of Canada from January 1987 to December 2010 to forecast this same economic indicator into the future. Figure 1 below showcases said economic indicator through time. We can see that the indicator is volatile, has a clear upward trend, has some spikes (likely due to external economic factors), and could have some seasonality.

Given that most of the bankruptcy filings are caused by economic downturns, it is no surprise that there is seasonality and trend in the data, and it is also not surprising that the largest bankruptcy rates are seen at the end of 2008 and beginning of 2009. Also, some of the largest overall increases in bankruptcy rates are seen around 1990, 1995, and 2008, all years of very bad economic performance.

Knowing that there is an intricate relationship between the indicator we will be trying to predict and several other economic factors, it could be important to analyze other economic indicators. Some of those are given in the train data. These include Unemployment Rate, Population, and the Housing Price Index.
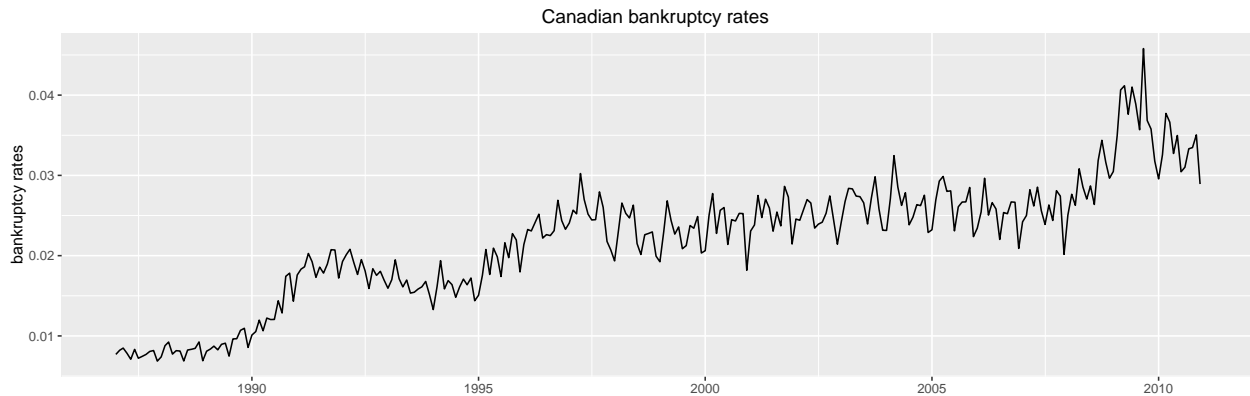
Figure 1: January 2011 – December 2012 bankruptcy rates

Methods to Solve Problem

The main idea behind the approach to solve the problem of forecasting bankruptcy rates into the future is to exploit the fact that the bankruptcy rate tomorrow is very likely to be similar to that of today. Similarly, we can exploit the fact that the rate today is likely to be very similar to that of yesterday. Furthermore, we can exploit other relationships in the data to gain insights into the future like the fact that bankruptcy rates in July tend to go up or that every January sees a decline in bankruptcy rates. We will now use plots and other metrics to see which are these kinds of relationships in the data and once we extract them, use them to predict into the future.

External Regressors Approach

One approach is to use other data that highly correlates to bankruptcy rates as external regressors. For example, if we know from economic theory that bankruptcy rates spike after people have a hard time paying off their debt, we can use other indicators like the House Price Index to tell us whether the bankruptcy rates are likely to go up or down when the House Price Index moves up or down. We will further explore these external regressors in the Final Method section to see if they should be included in a model to predict future bankruptcy rates.

It is important to remember that when adding external regressors, we want to keep the number of variables as small as possible to keep the complexity of the model down and the number of parameters to estimate small as well. Below are some of the potential variables we could use to help with the forecasting of bankruptcy rates.

Given that the regressors in the Possible External Regressors plot look correlated, we might want to include these in the models we test. If we decide to go this "multivariate" route, we can choose one of two approaches. The first approach works if we believe that the other variables we believe are correlated with bankruptcy rates are not affected by bankruptcy rates, but they **do** affect bankruptcy rates (i.e exogenous). That method is ARIMAX, which is an extension of AR, MA, and ARMA models whose intuition is described in the following section.

Another possible approach to model and forecast bankruptcy rates is we treat the external regressors (variables such as Unemployment Rate or Population) as variables that are affected by bankruptcy rates and bankruptcy rates are also affected by these external regressors (i.e endogenous variables). This method is called vector auto-regression. For both of these multivariate approaches we need to make sure that certain assumptions about the data or the modeling results (residuals) are met. The results and discussion of these assumptions will be covered in the Technical Appendix section of this report.
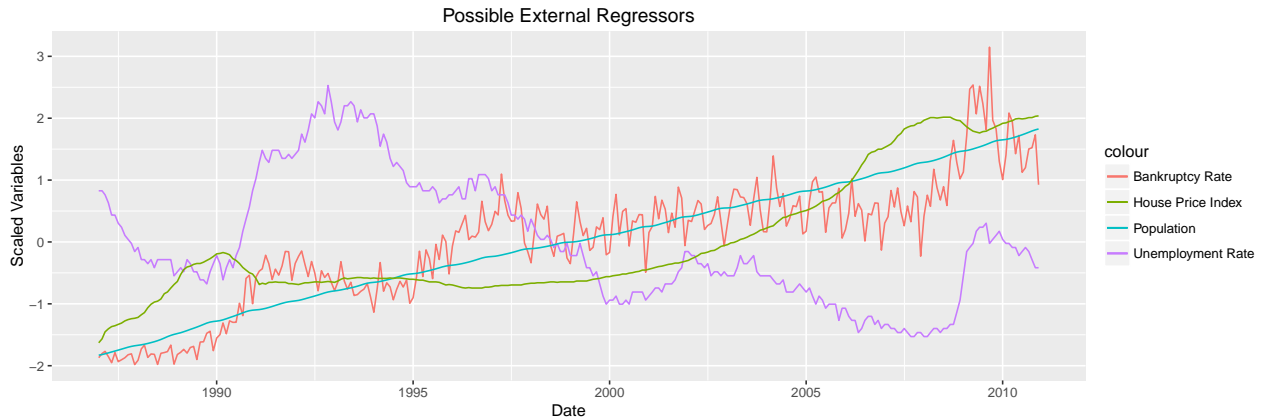
Figure 2: Possible External Regressors

One variable/univariate approaches

The main approaches to model time series data are AR, MA, and its combination ARMA models. The intuition behind MA models is that the time series is related to an error term and the error term one lag before that. For AR models, the intuition is that there is a dependence between an observation of the data and that same data point p steps before it. ARMA models are a combination of those two approaches.

A completely different approach is EX¡exponential Smoothing. In exponential smoothing we predict data points in the future through a set of recursive equations that do not require any distributional assumptions.

To solve the bankruptcy rates problem, we have employed every one of these different approaches and the following sections introduce the best model and why we think it is so.

Stationary Data

Stationary in a time series happens when the average values, or the overall variability of the time series are not changing over time. This basically means that we do not see an evident upward or downward trend, that we do not observe seasonality (spikes every July for example), and that we also do not observe large chunks of high volatility in the data. If we have a stationary time series (which the bankruptcy rates data does not look stationary), we can employ several methods which include AR, MA, ARMA or single exponential smoothing.

Stationary in a time series model is important since we are looking for some aspect of the data to **not** depend on time. In order to build a model with any kind of accuracy, we require an assumption that something does not vary with time. The Box-Jenkins methods for modeling time series belong to a class of models that adequately describe a stationary time series.

To illustrate stationarity and how we deal with it in our final model (described in the following section), below is an example of the raw time series of bankruptcy rates and what it looks like after one round of ordinary differencing.

Final Method Chosen

After looking closely at the available methods to solve this problem, it is clear that the best modeling approach will be to include an external regressor that helps predict movements in bankruptcy rates. Looking back at Figure 2, it looks like unemployment rate can help explain the big dips and large spikes in bankruptcy rates. For example, right after 1990 there is a very significant spike in the unemployment rate. This is
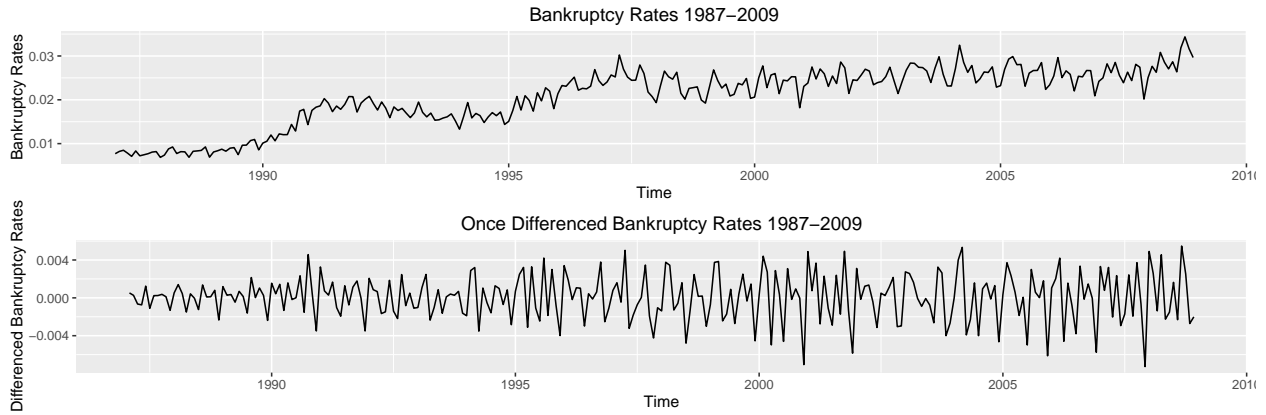
Figure 3: Non-stationary data after one ordinary differenceing

followed very closely by a very large spike in the bankruptcy rate starting at the same exact time period. This makes sense from an economic stand point because many of the bankruptcies re cause by people who are unable to pay credit cards or medical bills, so it makes sense that when there is a significant increase in the unemployment rate, bankruptcies will spike since people who are unemployed are less likely to be able to pay their debt. even though it is not a perfect correlation between bankruptcy rates and unemployment rates this will be the best predictor to help account for large spikes and significant decline in the bankruptcy rates.

The final model is an ARIMAX model which uses unemployment rate as an external regressor, does ordinary differencing on the log of bankruptcy rates, and uses 2 lags for the error and 5 lags for the inter-dependency of bankruptcy rates today and bankruptcy rates in the past. The final model also uses the log of the bankruptcy data to model the changes in rates from one month to the next as opposed to the raw rates.
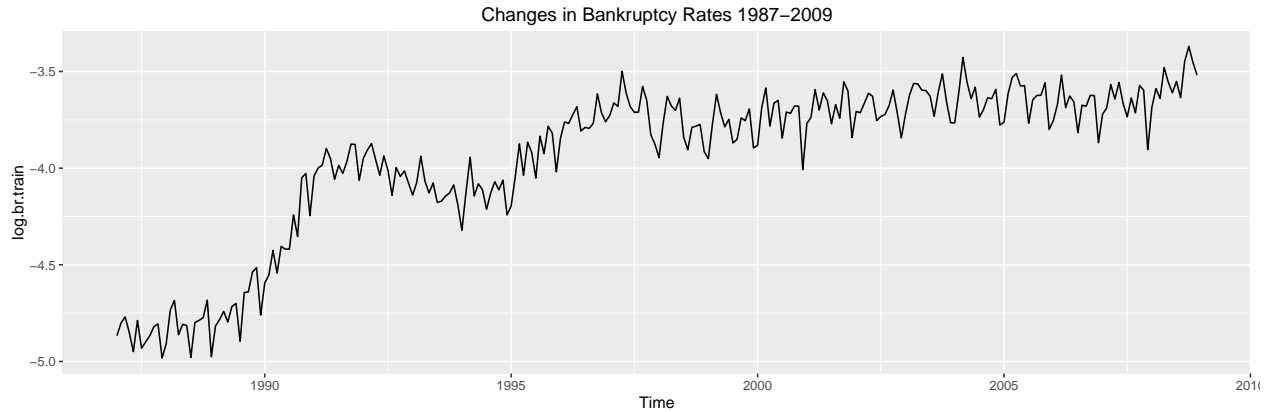

Figure 4: Changes in banruptcy rates from 1987-2009

Using this model to forecast into the future, the results look promising.

Given that we have monthly data from 1987-2010, to find an optimal model we split said data into two chunks, one from 1987-2008 and the other from 2009-2010. This will help us validate how the model is performing in terms of forecasting values into the "future", while still being able to find the error between what the model predicts and observed data. This model is doing a very good job of predicting into the future since when we predict the years 2009 and 2010, the model is on average missing the true rates by only **0.0059**.

Using the fact that we have those two years' data to validate the forecasts of our model, we tried all different kinds of models described in the "Methods to Solve Problem" section above. Fitting different
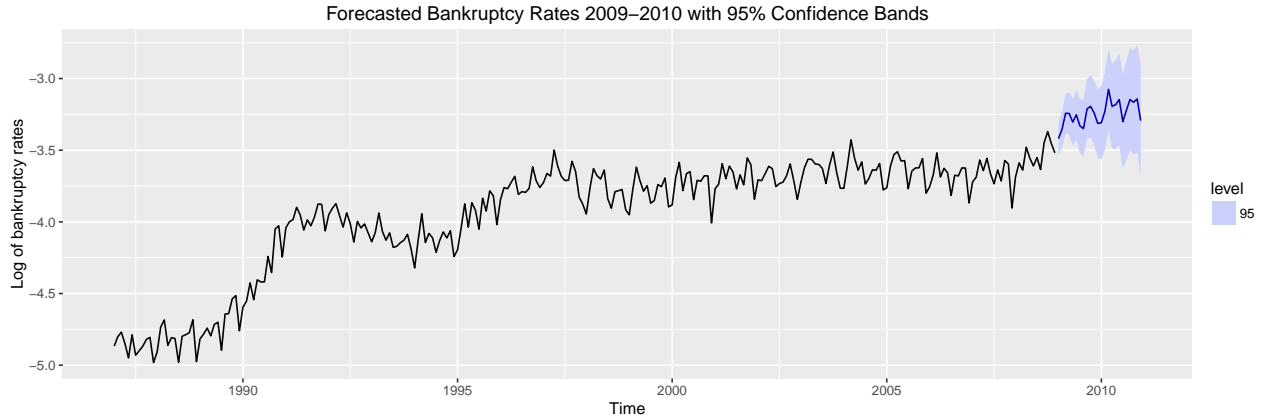
Figure 5: Forecasted bankruptcy rates from 2009-2010

models, we were able to compare 1. whether the models satisfied the error assumptions we imposed of the errors not depending on time or on each other and 2. the predictive power and prediction accuracy of each model by way of root mean squared error.

As seen in Figure 6, the model is doing a very good job of capturing the ups and downs of bankruptcy rates. We expect its predictive power to be similar when predicting to the future. As can be seen in Figure 5, the forecasts have a 95% confidence interval that still captures the overall behavior of changes in bankruptcy rates, so we can be confident that this model will predict well into the future.

```
## Don't know how to automatically pick scale for object of type ts. Defaulting to continuous.
```
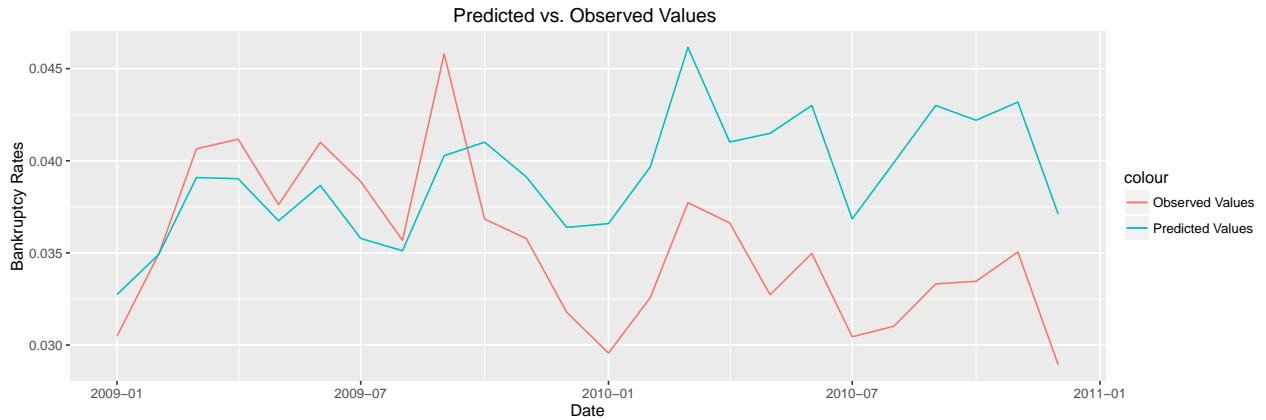

Figure 6: Observed vs. Predicted Bankruptcy Rates 2009-2010

- and an acknowledgement of any limitations associated with your method.

In terms of limitations of this model, this model will become less and less precise the further into the future we try to forecast bankruptcy rates. The upside is that if we keep predictions to about two years in the future for monthly data, we can keep adding observed values of the rates and always have a moving prediction window of two years into the future.

In terms of modeling assumptions, this model satisfies all of them. The errors terms (what we predict - actual observed values) are not correlated with each other. The errors do not depend on time (i.e there is not a series of errors that are positive followed by a series of errors that are negative). In other words, we are certain that there is a part of this model that is not dependent on time. The results for the formal and informal tests for these assumptions can be found in the technical appendix below.

Results

Visual Results

Monthly Exports in China
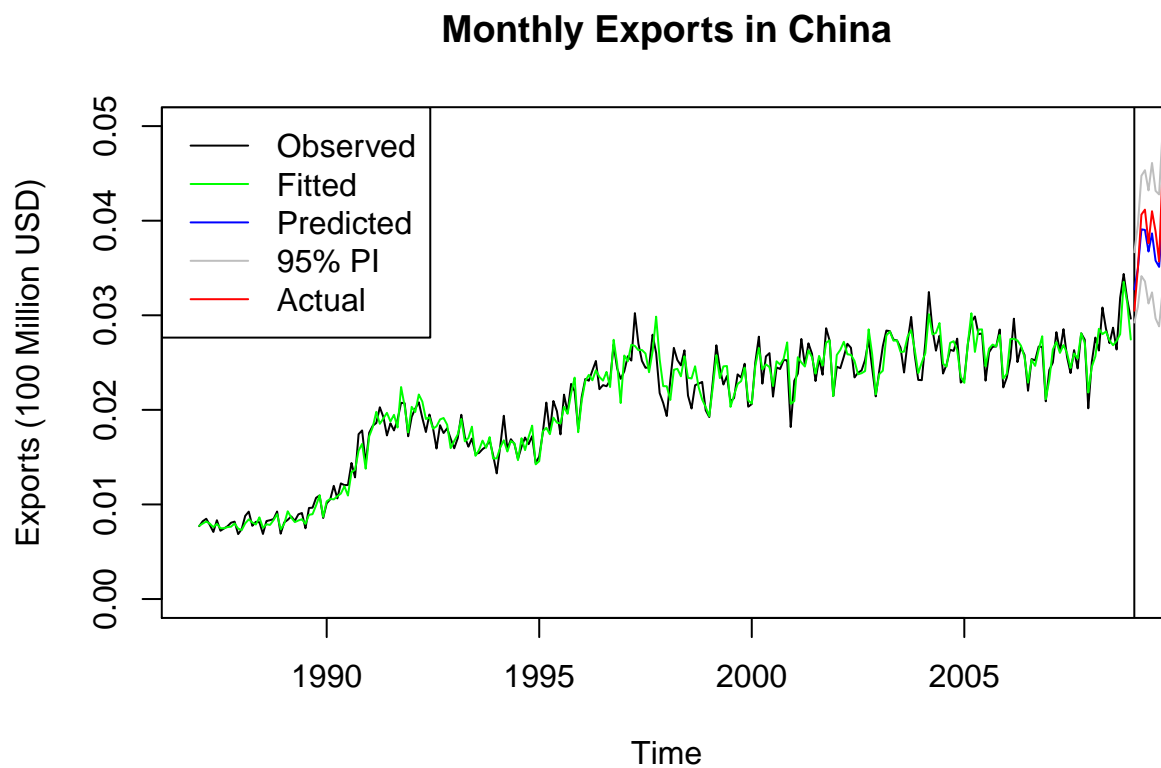
Tabular form Results

| observed	predicted	Date
0.0304815	0.03273570	2009-01-01
0.0349375	0.03491532	2009-02-01
0.0406471	0.03908764	2009-03-01
0.0411709	0.03902389	2009-04-01
0.0376177	0.03674052	2009-05-01
0.0410042	0.03866162	2009-06-01
0.0388748	0.03578512	2009-07-01
0.0356858	0.03511638	2009-08-01
0.0457978	0.04027539	2009-09-01
0.0368357	0.04100985	2009-10-01
0.0357764	0.03912313	2009-11-01
0.0317768	0.03638398	2009-12-01
0.0295630	0.03658995	2010-01-01
0.0325692	0.03966422	2010-02-01
0.0377229	0.04614726	2010-03-01
0.0366266	0.04101925	2010-04-01
0.0327276	0.04148942	2010-05-01
0.0349727	0.04300134	2010-06-01
0.0304477	0.03684668	2010-07-01

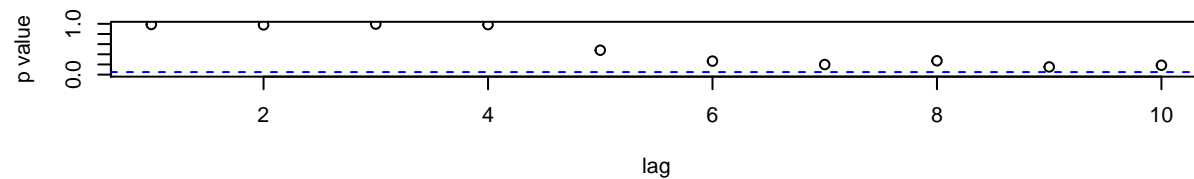| observed	predicted	Date
0.0310200	0.03990518	2010-08-01
0.0333156	0.04299992	2010-09-01
0.0334606	0.04219872	2010-10-01
0.0350488	0.04318679	2010-11-01
0.0289219	0.03709572	2010-12-01

Technical Appendix

Standardized Residuals
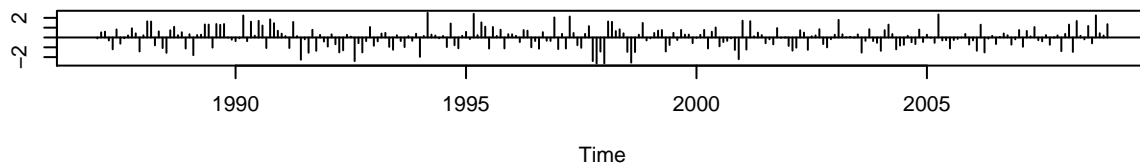

ACF of Residuals
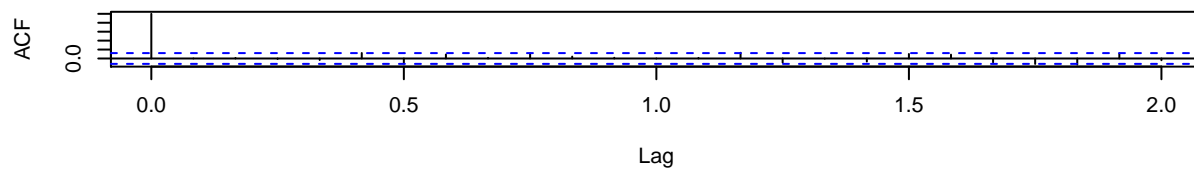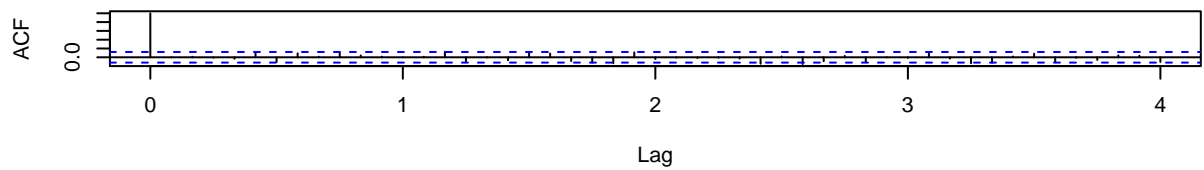

p values for Ljung–Box statistic

```
##
##  Shapiro-Wilk normality test
##
## data:  model$residuals
## W = 0.99221, p-value = 0.1795
##
##
##  One Sample t-test
##
## data:  model$residuals
## t = -0.21732, df = 263, p-value = 0.8281
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.007698603  0.006168152
## sample estimates:
```

```
##     mean of x
## -0.0007652253
##
##
##  modified robust Brown-Forsythe Levene-type test based on the
##  absolute deviations from the median
##
## data:  model$residuals
## Test Statistic = 1.1346, p-value = 0.3356
```
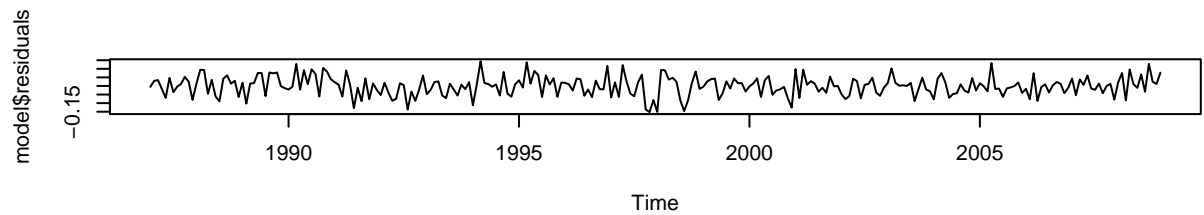

Series model$residuals

Series model$residuals

!