US Korea Exchange Rates
STA206 Final Project

Clark Fitzgerald clarkfitzg@gmail.com
Amy Kim atykim@ucdavis.edu

December 15, 2014

Abstract

We analyze the exchange rate between the United States and South Korea (hereafter Korea) using monthly country level economic data.

1 Introduction

1.1 Background

In today's highly connected global society people and money move between countries. The exchange rates between countries determine the relative value of that money.

We analyze data on exchange rate between the US and Korea from 1999 to 2014. In 1997 the IMF crisis in Korea perturbed the economic indicators, but in looking at the data we can see that it has been stable since 1999.

1.2 Questions

1. How do exchange rates behave?
2. Can we predict the exchange rate between two countries?
3. When is the best time to exchange currency from won (Korean currency) to dollar or vice versa?
4. Can we find a linear relationship that can be applied over other countries' currency?

1.3 Motivation

Sometimes it's necessary to transfer funds between countries. This is true for both individuals, families, and organizations. The exchange rate can vary substantially over time. For example, in Figure 1 we observe the exchange rate between Korea and the US varying between 900 and 1400 Korean Won per US Dollar. A quick calculation on this data implies that during this period a fixed amount of Korean currency could have been worth \$100,000 or \$158,000, depending on when it was exchanged.

Hence if one has a significant amount of capital to move from one country to another and some flexibility around the timing then it makes sense to do it when the exchange rates are favorable for the transfer. For example, you may want to sell a house in Korea and put that money towards a house in the US.

1.4 Data

We use data from Quandl.com, a service that provides clean, documented data from a wide variety of sources including the US Federal Reserve, World Bank, and the National Bank of Korea. `Exchange rate` between Korean Won and US dollar is the primary Y variable.

For each country we will examine the following variables:

- `gdp` The gross domestic product. Units: USD Million
- `unemployment rate` The percentage of the labor force who are unemployed and actively seeking work. Units: Percent (monthly)
- `exports` The total value of the goods and services produced domestically and purchased by foreign entities. Units: USD Million (Monthly)
- `imports` The total value of a country's imports of physical goods and payments to foreigners for services like shipping and tourism. Units: USD Million (Monthly)
- `interest rate` The monthly average of the central bank policy rate. This is the interest rate the central bank charges on loans to commercial banks. Units: Percent (Monthly)
- `inflation rate` The growth rate of the prices. (Monthly)
- `consumer produce index` The Consumer Price Index (CPI) is a measure of inflation related to the cost of living. Units: Index Points 2010=100, NSA (Monthly)
- `debt` The total amount of public and private debt owned by foreign creditors. Units: USD Million Current Prices, NSA (quarterly)
- `gdp deflator` The relative difference between the real and nominal GDPs. Units: Index Points NSA
- `goverment spending` The yearly expenditure of the federal government. Units: local currency

- `political party` Categorical variable for the political party of the president.

2 Methods and Results

2.1 Dynamic Data

The data used in this analysis is unusual- it's generated dynamically from live, high quality sources, and is self updating. The data is loaded directly from Quandl using a REST (Representational State Transfer) web API (Application programming interface), and then cached locally, limiting network dependence. Using this service makes it easy to repeat the analysis in the future, or conduct similar analyses between different countries.

2.2 Reproducibility

Above we mentioned that the data and all analysis in the report is self updating. This is accomplished through the use of GNU `make`, which describes a DAG (Directed Acyclic Graph) of dependencies for the final report. `make` detects file modifications and will lazily run all commands and scripts required for the final output.

What does this mean? Suppose that we need more variables to do the regression. Thanks to the use of industry standard ISO codes for the countries and proper parameters in our code base this becomes a simple task. We've stored string templates for the country level variables in a file called `template.txt`. The data collection and preprocessing steps are automated, so we can add more variables simply by adding a row to `template.txt` and entering the command `make` from the shell prompt, which causes the following sequence of events:

1. `make` detects the changed template file
2. The script `download.R` runs, updating the cache
3. The script `preprocess.R` runs, applying the appropriate transformations and joining the tables to a form suitable for analysis
4. All other scripts which depend on the preprocessed data run
5. The report output is produced

In this case the output is a PDF, but it could just as easily be a web page on the department server, or an upload into another REST API.

If we were doing this analysis for a client and need to run it again a year later with a new year of data then all we have to do is type `make`. We'll still need to interpret plots and models, but the most time consuming work is done.

One other aspect of reproducibility is version control. In this project we've used Git to collect snapshots of the project at every stage. The project is hosted publicly on Github[1]. A look at the log file shows what work happened when and why, which is important for establishing provenance. In particular it shows that we've been active on this project every day since December 1st. Here's an example of a log entry:

```
commit e2beaf01b0a1457e7167989c1fae575c2a676f19
Author: Clark Fitzgerald <clarkfitzg@gmail.com>
Date:   Mon Dec 8 20:51:04 2014 -0800

    implemented local caching, decoupled download step
```

Through a little engineering the data analysis process can be made transparent, automated, extensible, and fully reproducible.

Modern high level programming languages like R provide incredible tools. By using their capabilities we can

2.3 Exploratory Data Analysis

We started out with some general summary statistics and plotting. The first issue was missing data, which we had anticipated. While most of the variables were reported monthly, debt was reported on a quarterly basis and Our simple approach for dealing with this was to make a plot of these variables over time. If they are approximately linear then we could linearly interpolate with respect to time. The scaled plots Figure 2 and Figure 3 do appear to be mostly linear, so we interpolated. In this plots one can observe that debt_USA is decreasing; this is because debt carries a sign. The value of debt_USA on

When beginning this project we knew that the time dependency and multicollinearity would be the biggest obstacles, since in this case the assumption that the errors $\epsilon_i \sim N(0, \sigma^2)$ does not hold. Figure 4 shows what the original data looks like. We observe clear patterns emerging, with some even approaching a continuous mathematical functional relationship- ie the patterns between gdp, gdp_deflator, and debt for Korea. Figure 5 shows a histogram of all the pairwise sample correlations for the original data set- they are highly correlated. In fact, 30 % of the columns have absolute sample correlation greater than 0.9.

Our first approach for overcoming this time dependence was to create a new data frame from the original data by taking the difference between each subsequent row (monthly difference). More formally, if our original design matrix X is $n \times p$ then we can define a new $n - 1 \times p$ matrix $\tilde{X}$ by setting

$$\tilde{X_{i,*}} = X_{(i+1),*} - X_{i,*}$$

for $i = 1, 2, \ldots, n - 1$ where $X_{i,*}$ represents the ith row of X.

As can be observed in Figure 6, this approach did indeed eliminate the most obvious patterns, but at the expense of interpretability. We attempted regression on this transformed data,

[1] https://github.com/clarkfitzg/stats206_project

but were unsuccessful.

2.4 Ratios

Because exchange rate is a ratio of Korean currency / US dollar we decided to make transformed variables consisting of the Korean economic indicator divided by the corresponding US indicator. We also included `inflation` untransformed after some preliminary analysis indicated that it was important. `Date` was intentionally left out of the model in order to be able to interpret the model in terms of the other parametrs. However, the effects of time are still present through the variables that are linear over time.

2.5 Analysis

Our first step when fitting the model was to separate the data into training and validation sets. We checked to make sure that the variables looked approximately the same in both training and validation sets.

We used forward stepwise selection to choose a model. Based on the criteria from the output of the `leaps regsubsets` in the appendix we chose to explore a model with an intercept and 5 predictor variables. The one with 5 predictors had the lowest BIC of all the models, and the Mallow's CP at 5.39 was significantly better than the model with 4 predictors which had a CP value of 13.91. Since 5.39 is close to 5 and 13.91 is pretty far from 4 we conclude that the model with 5 variables is correct, meaning that it has little bias.

The variables under further investigation are: `gdp`, `interest_rate`, `inflation_KOR`, `inflation_USA`, and `govspending`.

The box - cox transformation in figure 7 suggests a transformation of $1/x^2$, but this extreme and is due to the outliers. We used the multiplicative inverse transformation and examined the residual plots figure 8 to see why this is the case.

We see that the points 113 and 117 are the primary outlying cases. Looking up the corresponding dates reveals that these are the points occuring during October 2008 and February 2009, during the subprime mortgage crisis which shook the global economy. There are no points between 113 and 117 because the test train split didn't select them. In the validation set we discovered similar results for this period. The subprime mortgage crisis is the reason that they are outliers.

Attempting to reference [1].

3 Conclusions and Discussion

Our use of dynamic data caused trouble because the data was a little too dynamic.

4 Appendix

Model Selection Output:

	sse	bic	r2	r2a	cp
1	0.00	-46.47	0.37	0.37	411.04
2	0.00	-178.62	0.80	0.79	55.28
3	0.00	-185.95	0.82	0.81	41.04
4	0.00	-206.51	0.85	0.85	13.91
5	0.00	-212.36	0.86	0.86	5.39
6	0.00	-210.43	0.87	0.86	4.72
7	0.00	-207.14	0.87	0.86	5.33
8	0.00	-202.47	0.87	0.86	7.22
9	0.00	-197.85	0.87	0.86	9.06
10	0.00	-193.12	0.87	0.86	11.00

List of Figures

1 Exchange rate between US and South Korea from 1999 to 2014 7
2 Scaled variables with NA values 8
3 Scaled variables with NA values 8
4 Scatterplot of untransformed data 9
5 Histogram of all pairwise sample correlations 10
6 Scatterplot of data after row difference transformation 11
7 Box cox procedure on training set- affected by outliers 12
8 Cooks distance for residuals of training data 13

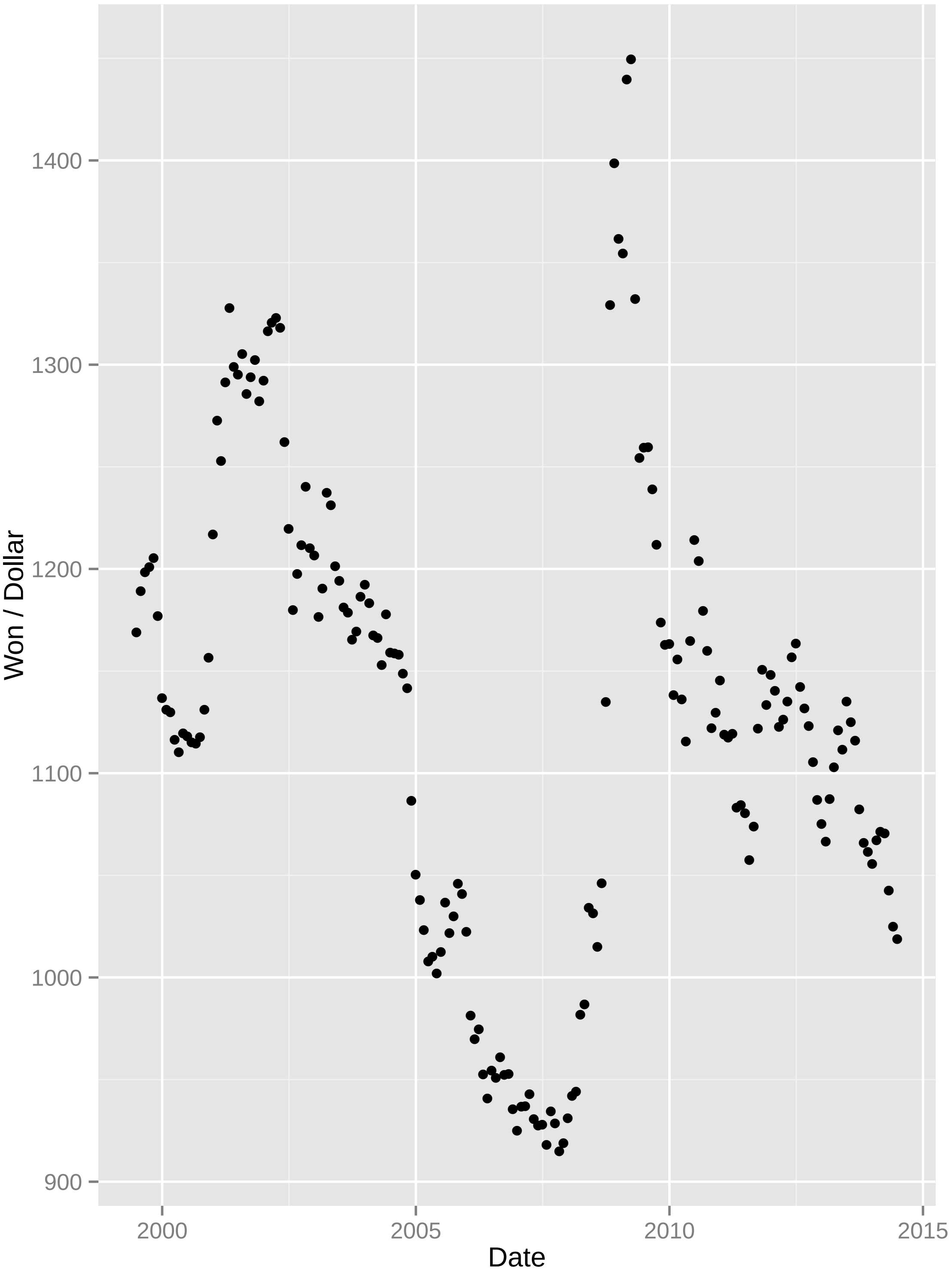

Figure 1: Exchange rate between US and South Korea from 1999 to 2014

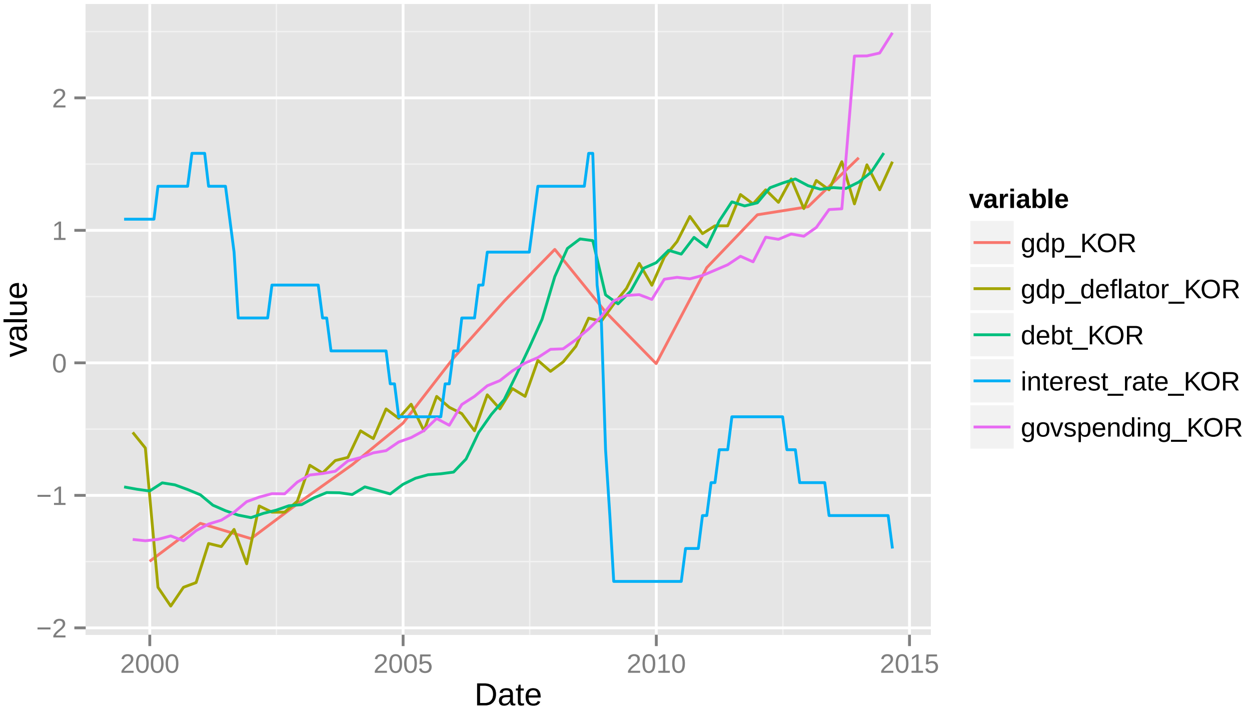

Figure 2: Scaled variables with NA values

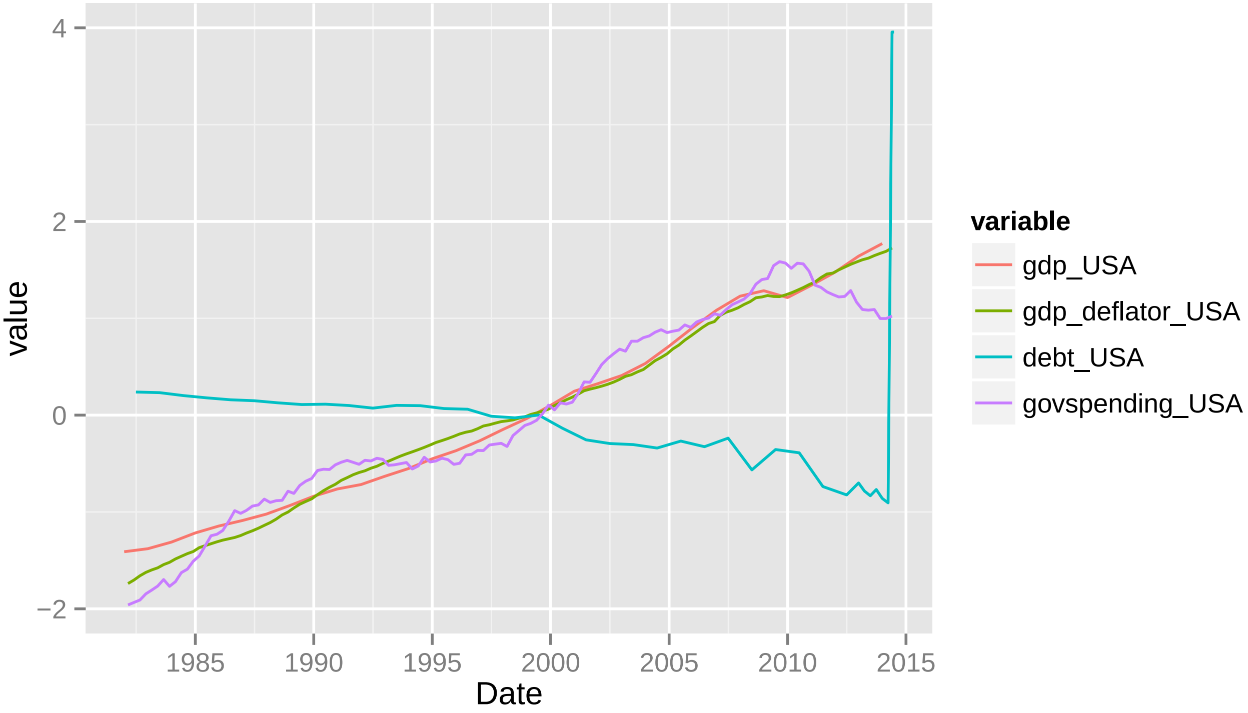

Figure 3: Scaled variables with NA values

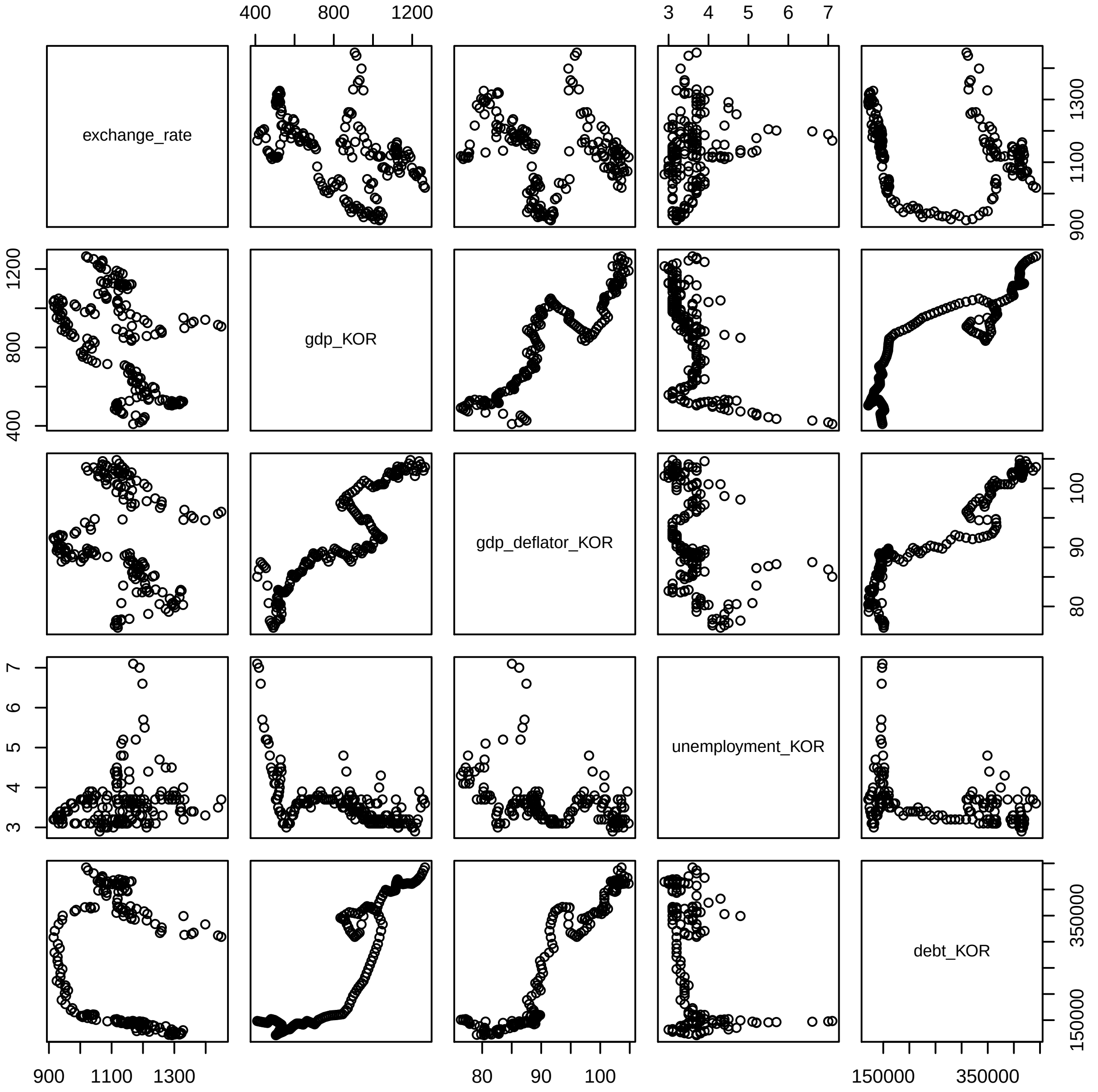

Figure 4: Scatterplot of untransformed data

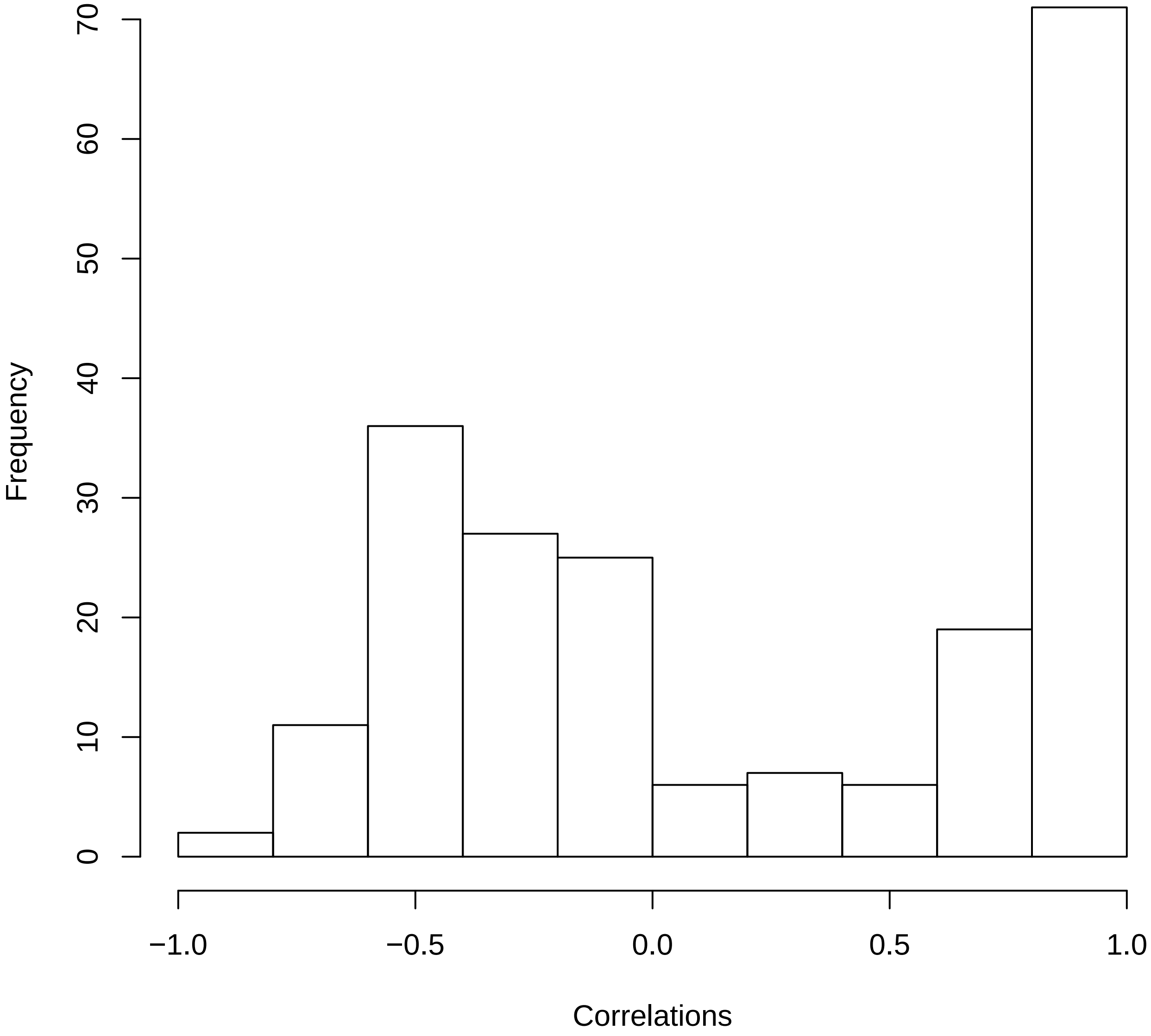

Figure 5: Histogram of all pairwise sample correlations

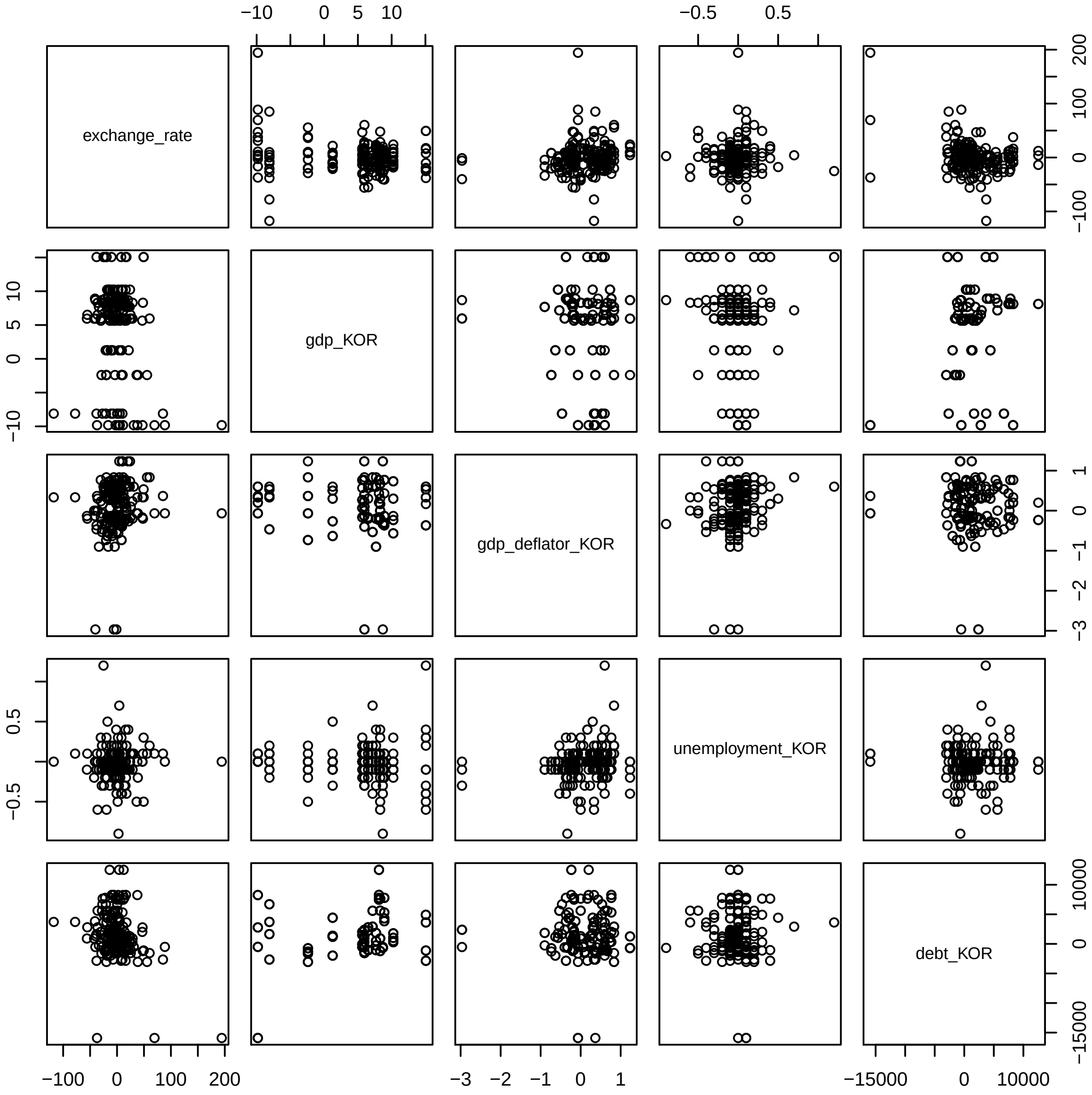

Figure 6: Scatterplot of data after row difference transformation

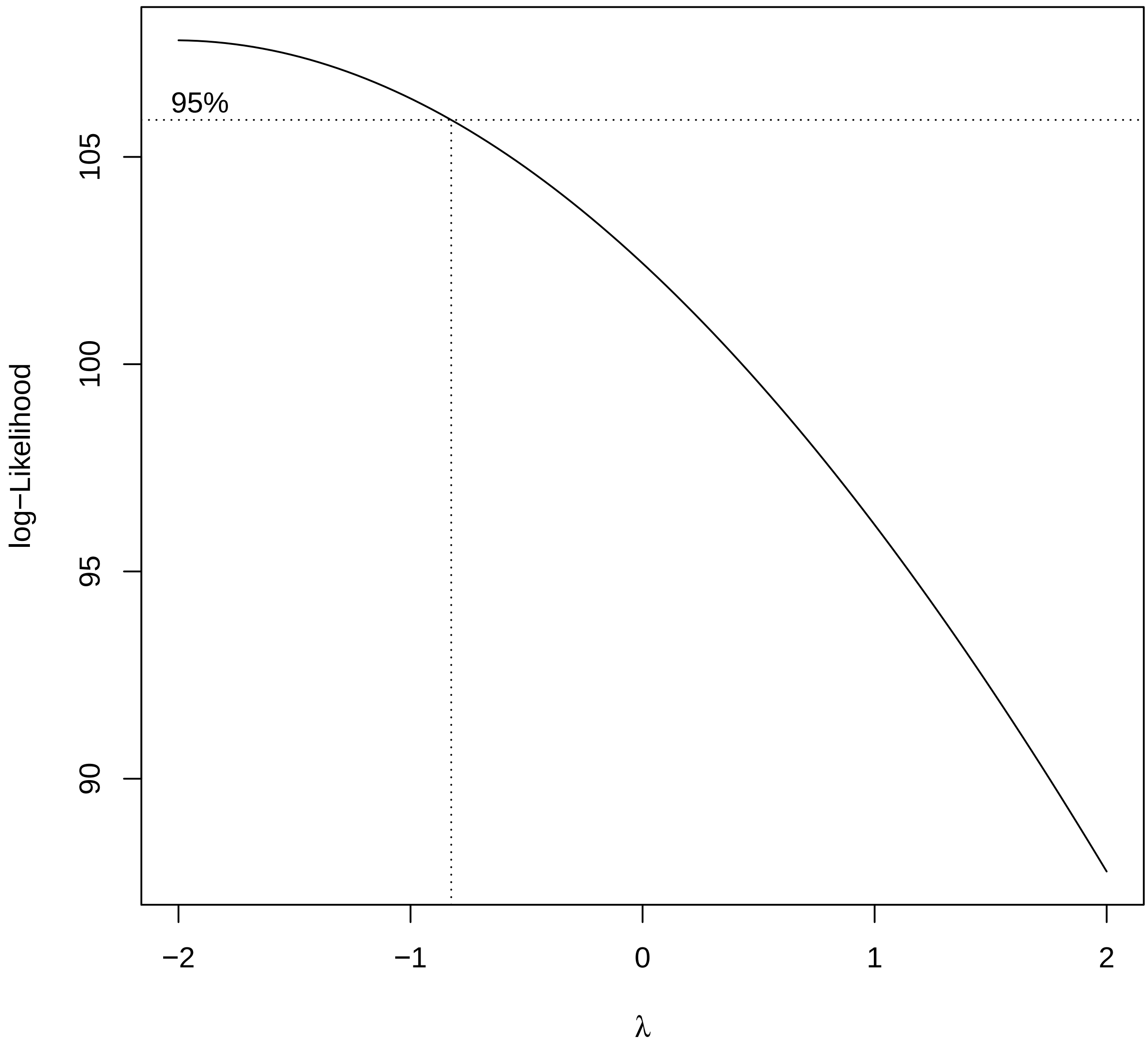

Figure 7: Box cox procedure on training set- affected by outliers

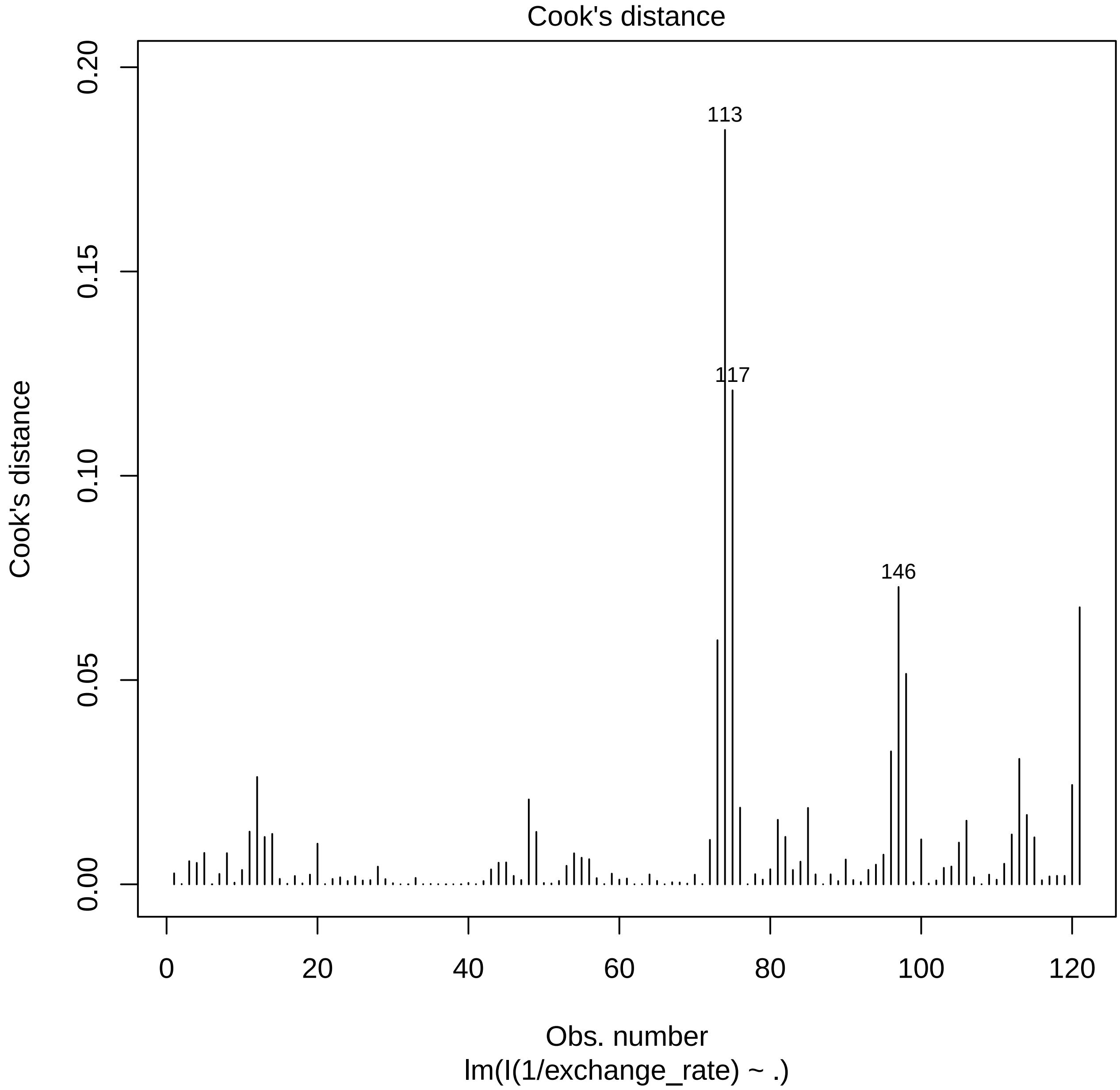

Figure 8: Cooks distance for residuals of training data

References

[1] Leslie Lamport, *LaTeX: a document preparation system*. Addison Wesley, Massachusetts, 2nd edition, 1994.