**China`s GDP and its factors**

**Technical Report**

**Lu Zhenyang**

To test statistical factors and form a regression model of China`s GDP, I collect data of various economical indicators.

**Data sources:**

1. International Monetary Bureau:

<http://www.imf.org>

2. National statistics Bureau of China:

<http://www.stats.gov.cn>

**Dependent Variables:**

Gross Domestic Product of China (GDP)

**Independent Variables:**

Fixed Assets Investment (UFA)

Broad Money Supply (M2)

Total of Import and Export (TIE)

Shanghai Composite Index (SHI)

Gross Domestic Product of USA (GUSA)

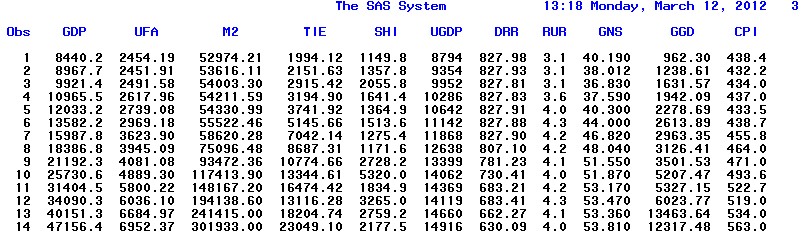
Currency Rate of U.S. Dollar and Chinese Yuan (DDR)

Registered Unemployment Rate (RUR)

Gross National Savings (GNS)

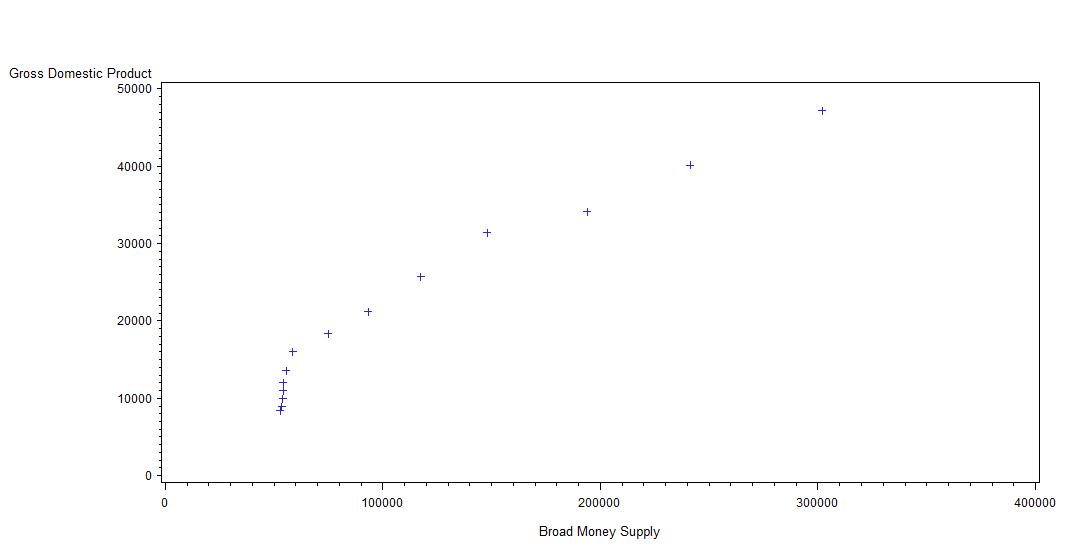
General Government Gross Debt (GGD)

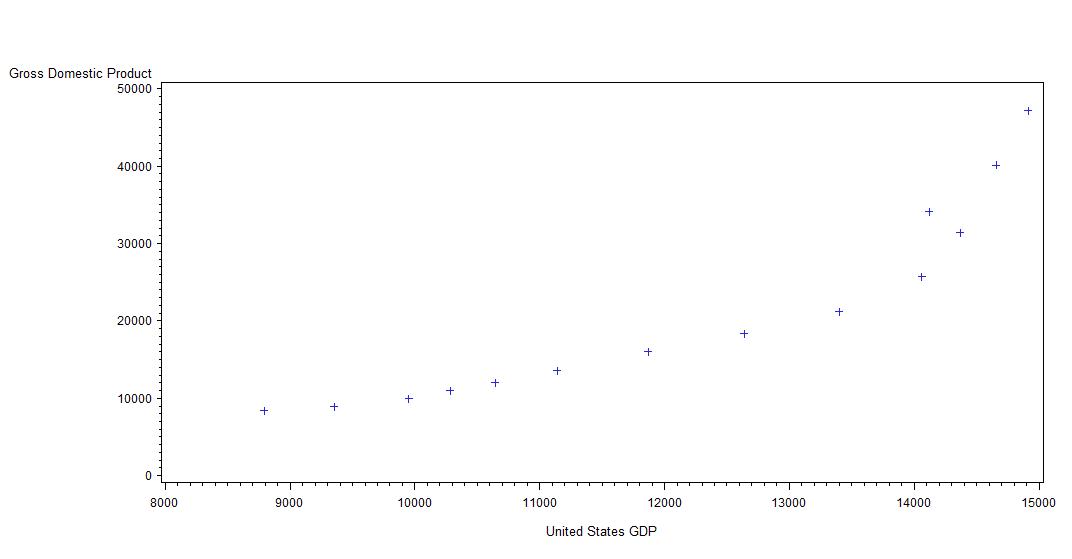
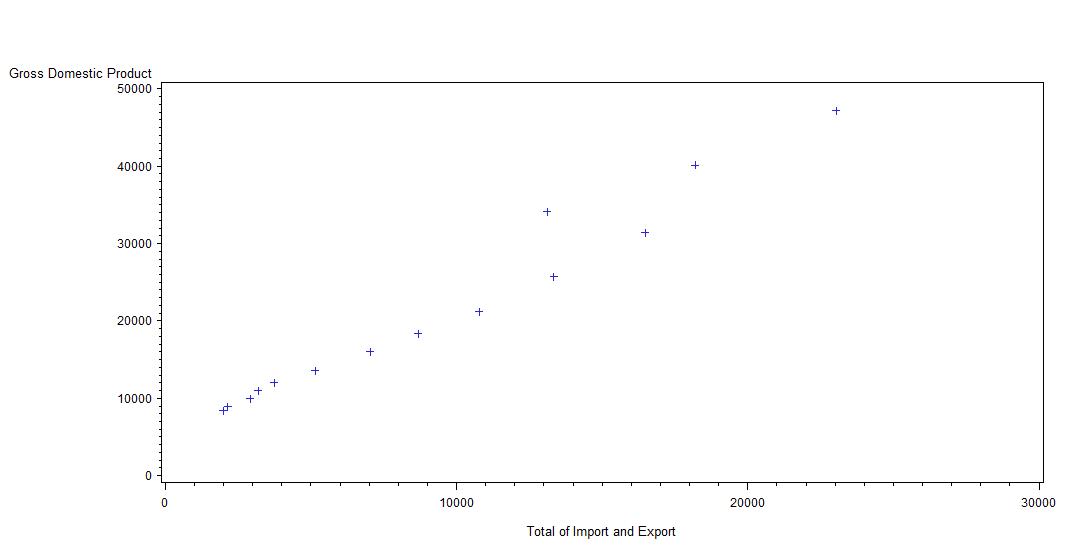
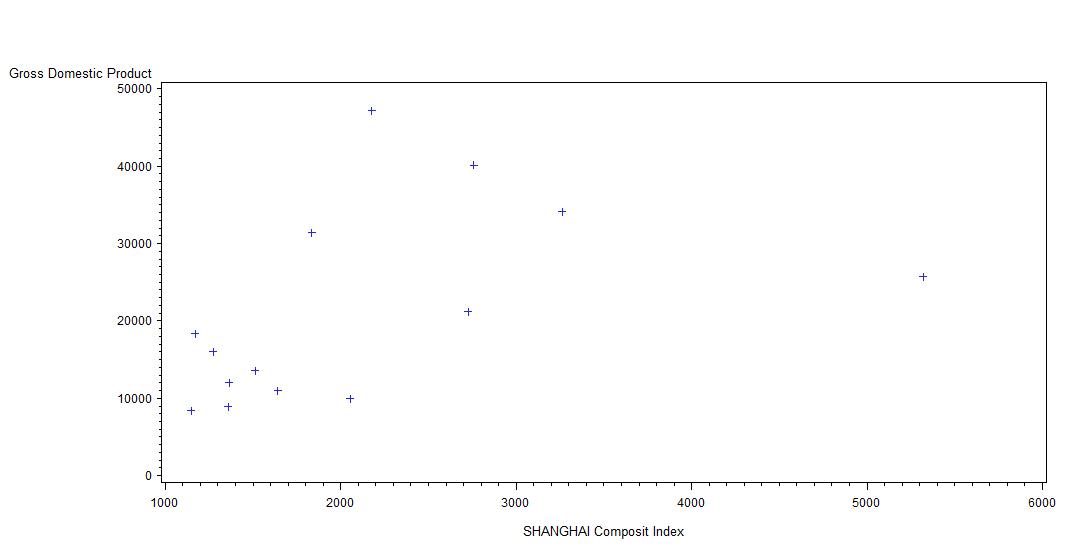
Consumer Price Index (CPI)

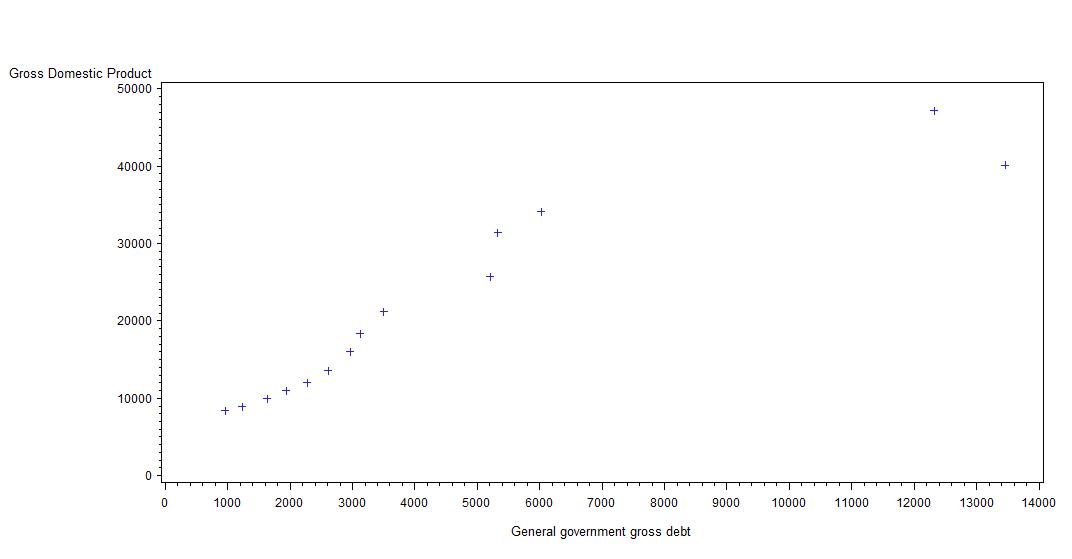
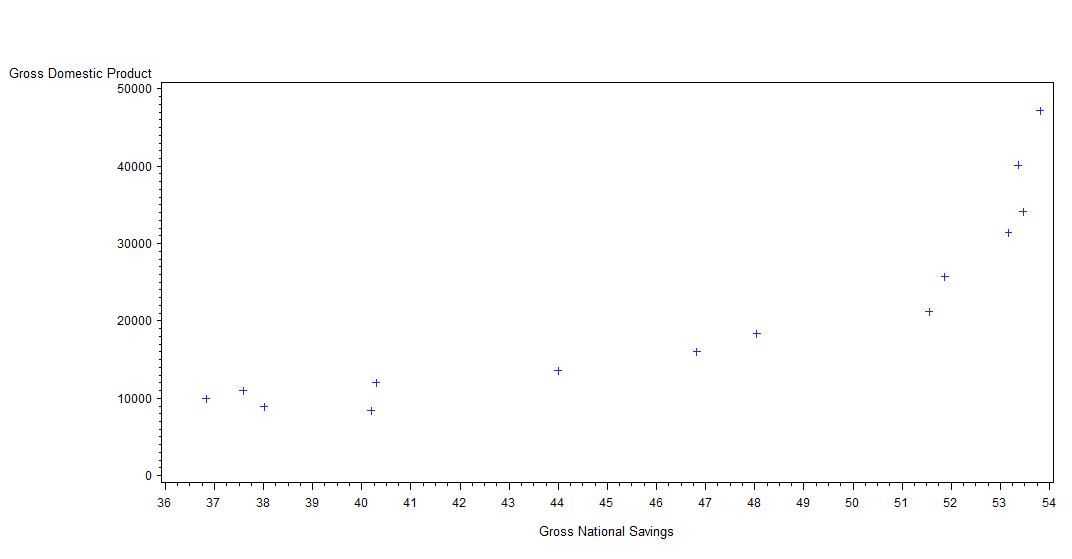
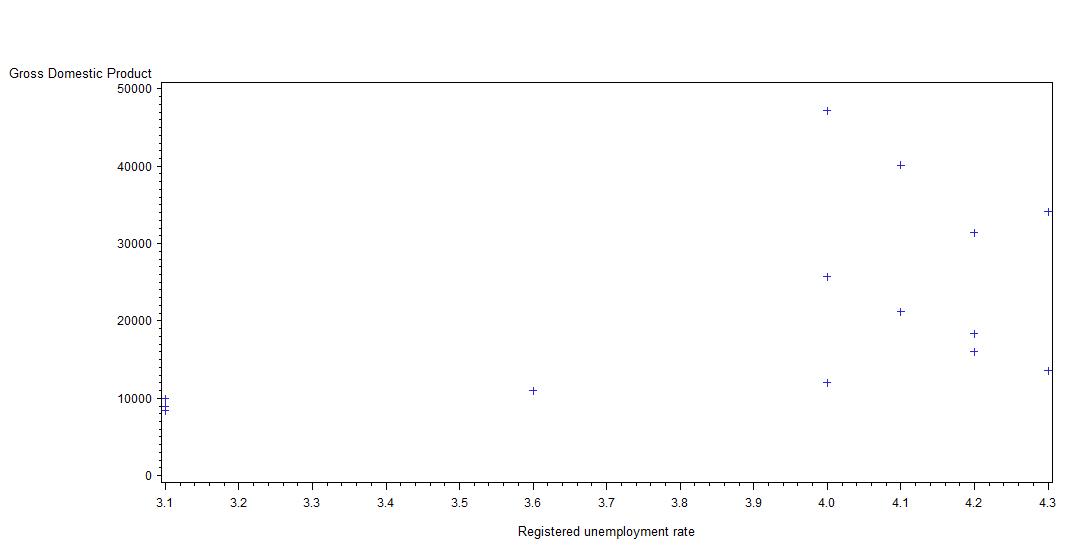
**Dataset:** 

**First step regression analysis:**

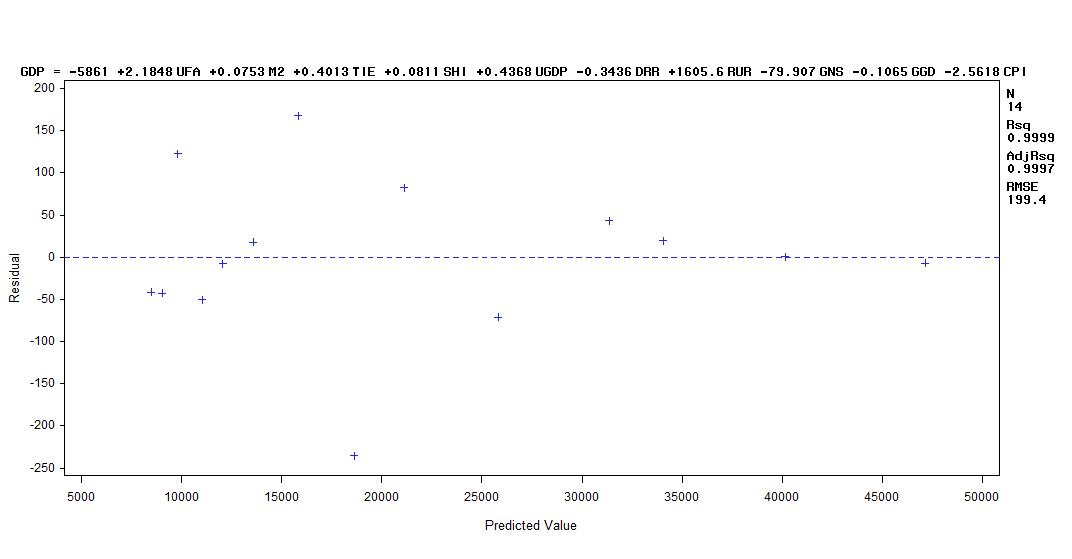
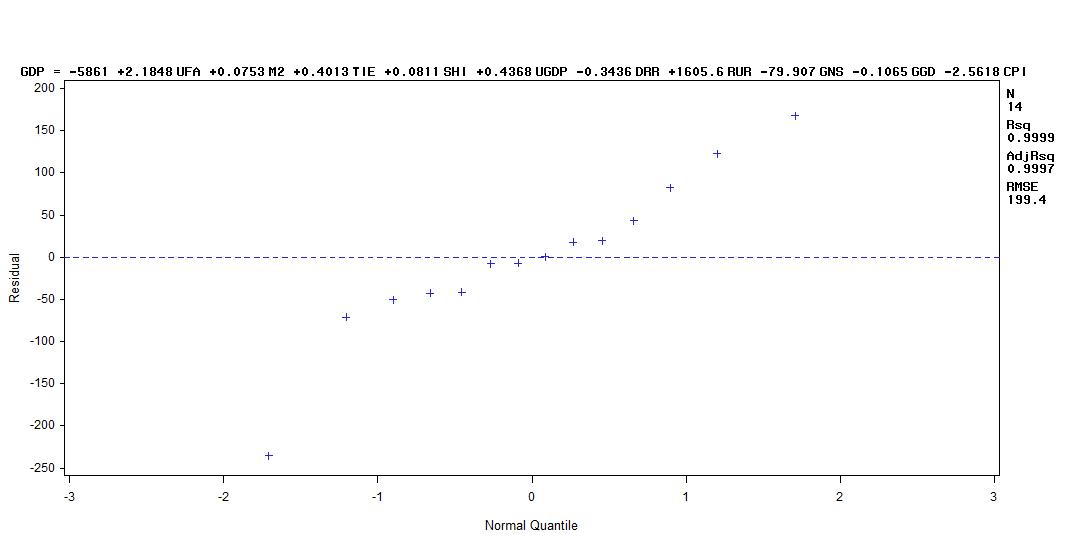
Independent variables that are not linear with dependent variables:







**Residual plot:**

****

**REG OUTPUT:**

Analysis of Variance

Sum of Mean

Source DF Squares Square F Value Pr > F

Model 10 2045016321 204501632 5143.48 <.0001

Error 3 119278 39759

Corrected Total 13 2045135599

Root MSE 199.39751 R-Square 0.9999

Dependent Mean 21286 Adj R-Sq 0.9997

Coeff Var 0.93673

Parameter Estimates

Parameter Standard Variance

Variable DF Estimate Error t Value Pr > |t| Inflation

Intercept 1 -5860.98033 22514 -0.26 0.8115 0

UFA 1 2.18481 1.33267 1.64 0.1997 1592.99395

M2 1 0.07532 0.00543 13.87 0.0008 63.45167

TIE 1 0.40129 0.21456 1.87 0.1582 681.98931

SHI 1 0.08111 0.11734 0.69 0.5391 5.78998

UGDP 1 0.43683 0.42908 1.02 0.3836 273.02145

DRR 1 -0.34365 7.43593 -0.05 0.9660 101.15513

RUR 1 1605.64366 329.10953 4.88 0.0165 7.36511

GNS 1 -79.90740 50.09319 -1.60 0.2089 36.57681

GGD 1 -0.10647 0.14866 -0.72 0.5256 109.28842

CPI 1 -2.56181 50.74356 -0.05 0.9629 1658.69084

**Comments on first step:**

Seven independent variables are not linear with dependent variable, say China GDP. This is mainly because those data are from economy, which is different from physics or natural science, data have huge variations. What`s more, those data cannot be measured very precisely. Outliers are generated often by economic policies by Chinese Government and economic fluctuation, so we do not want to omit those outliers.

The residual vs. predicted value plot shows a heterscedastic relationship. As a result, we are going to use natural log transform to the regression model.

Also, those ten independent variables show a strong VIF, saying that some of them are multicollinear. We have to delete some of the variables, too.

**Second Step on Regression:**

We then use three Regression Model Selection methods to delete redundant variables applying SAS source code:

model logGDP= logpmi logMSR logcpi logiov logaov logufa logm2 logrpi logtie logtrl/ selection = stepwise slstay= 0.05 slenter= 0.1;

model logGDP= logpmi logMSR logcpi logiov logaov logufa logm2 logrpi logtie logtrl/ selection = backward slstay=0.05;

model logGDP= logpmi logMSR logcpi logiov logaov logufa logm2 logrpi logtie logtrl/ selection = forward slenter = 0.1;

**The output we got using STEPWISE SELECTION:**

Summary of Stepwise Selection

Variable Variable Number Partial Model

Step Entered Removed Label Vars In R-Square R-Square C(p) F Value Pr > F

1 logtie 1 0.9818 0.9818 281.215 648.63 <.0001

2 logggd 2 0.0094 0.9912 133.083 11.71 0.0057

3 logrur 3 0.0030 0.9942 87.0603 5.16 0.0464

5 logm2 3 0.0045 0.9974 35.4016 17.42 0.0019

6 logcpi 4 0.0019 0.9993 6.5617 26.28 0.0006

**The output we got using BACKWARD SELECTION:**

Summary of Backward Elimination

Variable Number Partial Model

Step Removed Label Vars In R-Square R-Square C(p) F Value Pr > F

1 logugdp 9 0.0000 0.9998 9.1004 0.10 0.7721

2 logufa 8 0.0000 0.9998 7.3147 0.28 0.6268

3 logtie 7 0.0000 0.9997 6.0592 1.12 0.3378

4 loggns 6 0.0001 0.9997 5.1821 1.66 0.2451

5 logshi 5 0.0002 0.9995 5.6359 3.31 0.1115

6 logdrr 4 0.0002 0.9993 6.5617 3.07 0.118

**The output we got using FORWARD SELECTION:**

Summary of Forward Selection

Variable Number Partial Model

Step Entered Label Vars In R-Square R-Square C(p) F Value Pr > F

1 logtie 1 0.9818 0.9818 281.215 648.63 <.0001

2 logggd 2 0.0094 0.9912 133.083 11.71 0.0057

3 logrur 3 0.0030 0.9942 87.0603 5.16 0.0464

4 logm2 4 0.0035 0.9977 33.4264 13.38 0.0053

5 logcpi 5 0.0017 0.9994 8.0210 21.88 0.0016

6 logdrr 6 0.0001 0.9995 7.6241 2.20 0.1815

7 logshi 7 0.0002 0.9997 6.8917 3.35 0.1169

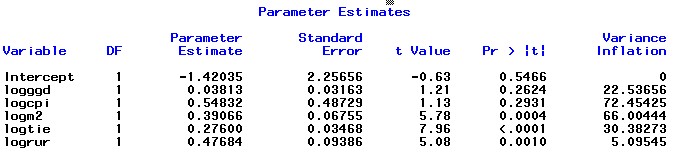
8 loggns 8 0.0001 0.9998 7.3147 2.38 0.1836

**Comments on second step regression:**

Using Forward, Backward and Stepwise Model Selection methods, we obtain three models. Since those three models behave equally in terms of Rsquare, we want to keep the parsimonious model, the one that is selected by STEPWISE SELECTION:

**Log(GDP) = Log(TIE) + Log(GGD)+ Log(RUR)+ Log(M2)+ Log(CPI);**

However, VIF for this model is still large, as shown:



We still have make it short by deleting the one(s) has/have the largest P-Norm! We leave out three variables like the following order:

**1. Log (GGD);**

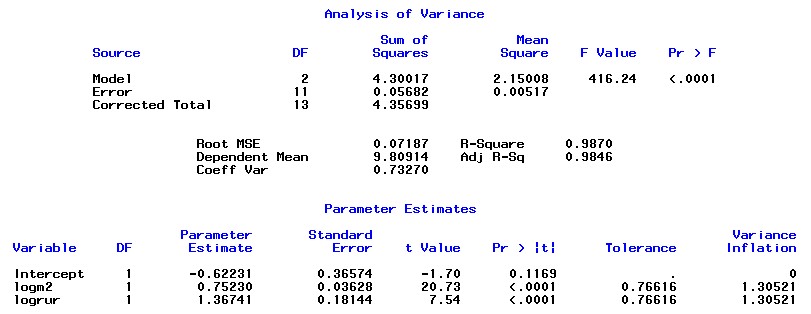
**2. Log (CPI);**

**3. Log (TIE);**

Finally, we got the regression equation:

**Log (GDP) = Log (M2) + Log (RUR) = -0.62231+1.36741\*log (RUR) + 0.75230\*log (M2)**

in which, M2 represents the Broad Money Supply and RUR means Registered Unemployment Rate.

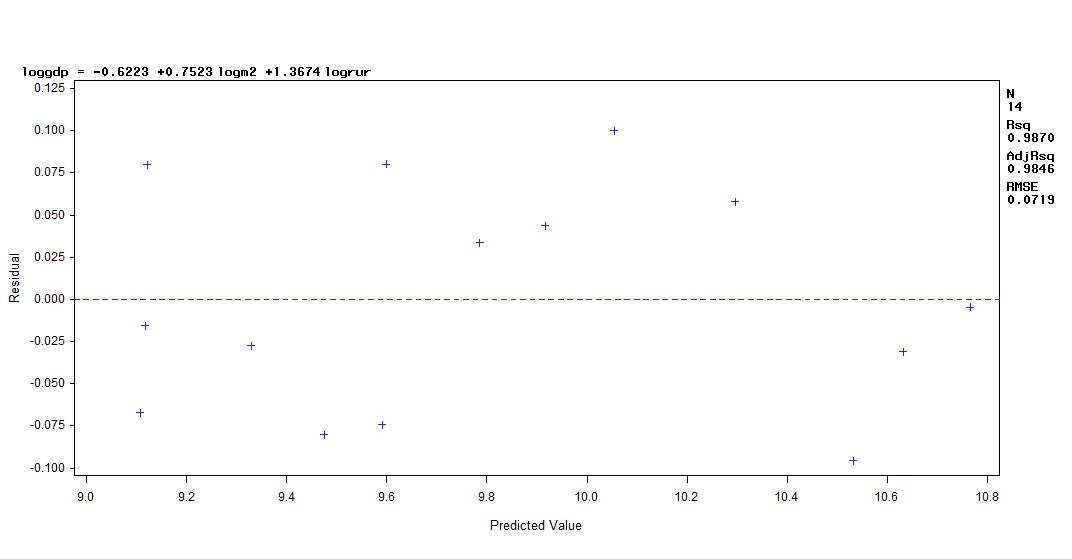


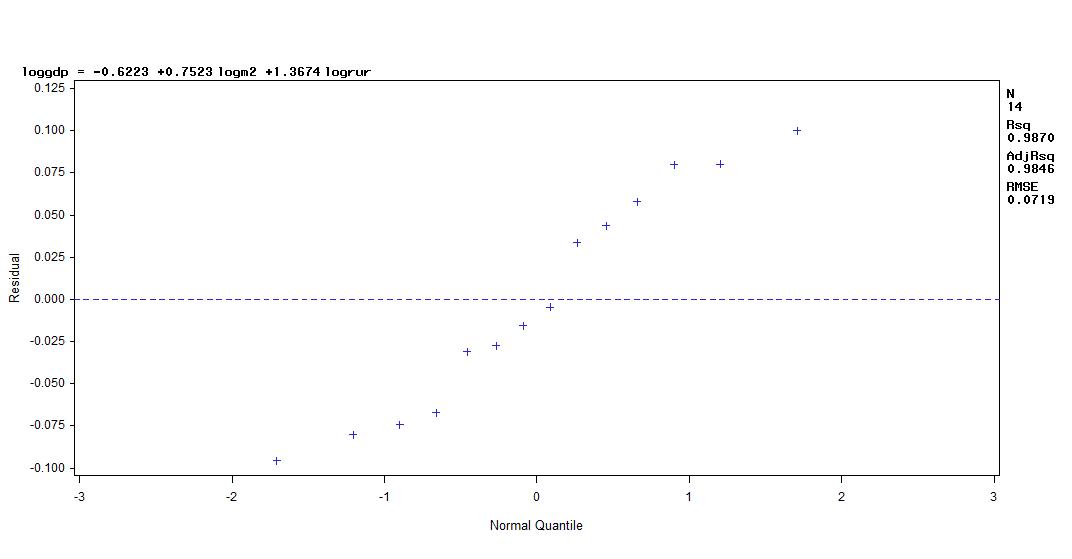
As shown above, those two variables, Log (m2) and Log (RUR) have the best VIF, indicating that they are not collinear. VIF for them are 1.30521 and 1.30521.

R2 for this equation is 0.9870, so 98.7% of the data can be explained by this equation.

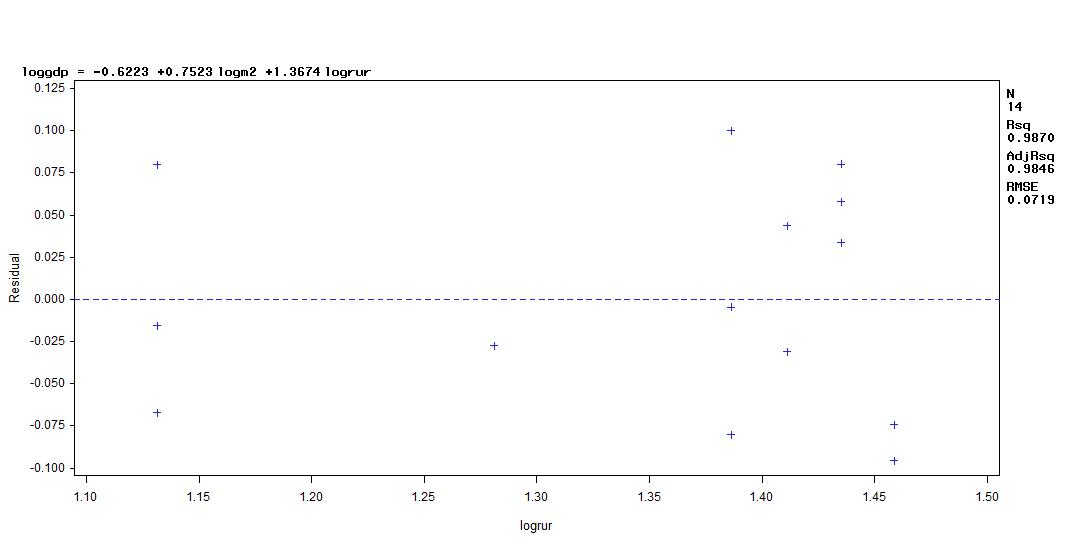
Global F-test is also significant, since f value is 416.24, far more large that the cutoff value.

**Residual vs. predicted plots for the final model**

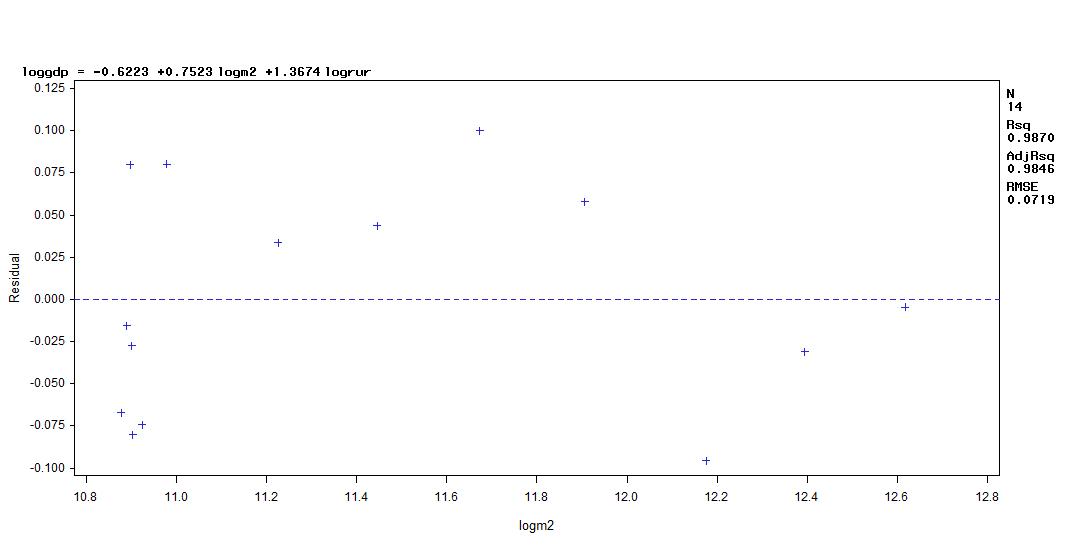
This plot behaves almost well, unbiased and homoscedastic, though observations are not enough to see it. Our transformation works.



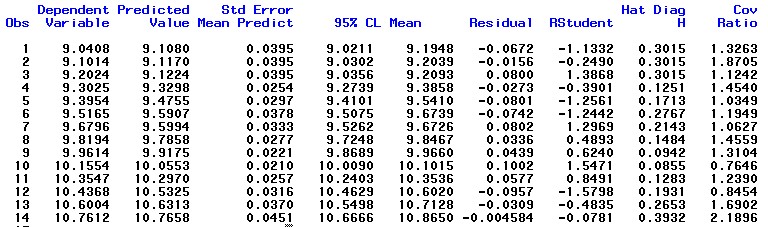
Our Normal plot is also not good, maybe because the Registered Unemployment Rate (RUR) is not collinear with China`s GDP. This can be demonstrated by the following plots:



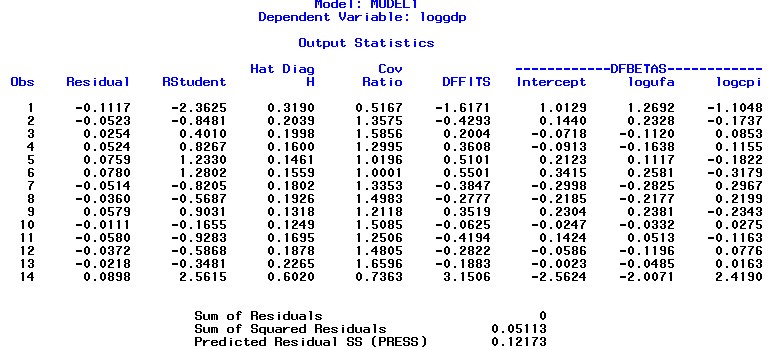
The residual plot of Log (RUR) is focused on where Log (RUR) is between 1.35 and 1.45, though the plot is unbiased and homoscedastic.



**Predicted value and its 95% CL mean**

****

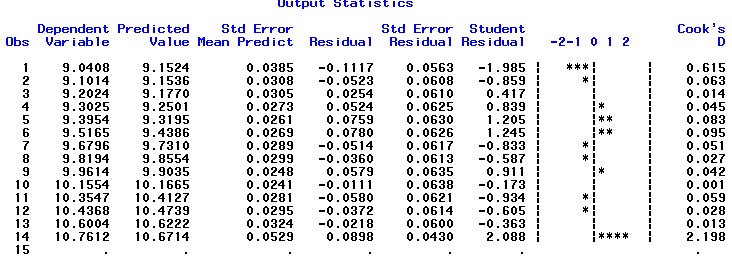
**Influence points analysis:**



In our model, cutoff for DFBETAS is 2 / sqrtroot(N) = 2/3.742 = 0.5345, so influence points are the first and 14th observations.

Cutoff for DFFITS is 2/ sqrtroot(p/n) = 2/0.378 = 5.29, so there is no influence point denoted.

Rstudent also shows 1th and 14th observations are also possibly outliers!



14th`s cook`s d value is larger than 1, which is the cutoff for outlier. Also, the variation plot shown is indicative to 1th and 14th observation.

**Conclusion:**

So, our final model,

**Log (GDP) = Log (M2) + Log (RUR) = -0.62231+1.36741\*log (RUR) + 0.75230\*log (M2),**

Indicate that if Broad Money Supply and the Registered Unemployment Rate increase, China`s GDP will also increase.

It`s weird that the higher the unemployment rate, the higher China`s GDP. Intuitively, it has to be a negative relationship between Unemployment rate and GDP, but in our case, it is not so. Why? Since Chinese Government record unemployment by only recording those who claim to the government that they are unemployed rather than recording everyone that is unemployed. This behavior results in an inaccurate statistics, the actual number is larger than claimed by the government, which we use in our model. Recently, many people have been talking about the inaccuracy of China`s unemployment rate. In our model, we notice that the unemployment rate of China is positive related to China`s GDP, which contradict our intuition.

Also, we can conclude that the more M2, the larger China`s GDP.