# An Empirical Exercise Of Data Snooping About P-value

Zhang Zhiyuan 15220162202517

March 23, 2019

**Abstract**

In this exercise, my main task is to predict futures' prices, using past available data, and see how the prediction results differ when I set my guideline as significance level(ie P-value) and model construction. I focus on three kinds of different futures, soybean futures, gold futures and copper futures. Also for better prediction, I included five possibly useful macro factors, which are product price index, risk-free rate, money supply, exchange rate, and net export. I collected the monthly data of the above variables, and use them to do regression analysis based on time series analysis.

## 1 Regression[1]

### 1.1 Soybean

In the regression analysis of soybean futures, I set my guideline as the significance level, so I applied the selection process in which insignificcant variables are excluded. My detailed stata codes and results are as follows:

codes:

```
### regress lnsoy lnsoy1 lnppi1 lnnx1 lnex1 lnrf1 lnm21 ###
### regress lnsoy lnsoy1 lnex1 lnm21 lnppi1 lnrf1 ###
### regress lnsoy lnsoy1 lnex1 lnm21 lnrf1 ###
### regress lnsoy lnsoy1 lnex1 lnm21 lnppi1 ###
```

results:

| lnsoy | P>\|t\| | lnsoy | P>\|t\| | lnsoy | P>\|t\| | lnsoy | P>\|t\| |
|---|---|---|---|---|---|---|---|
| lnsoy1 | 0.000 | lnsoy1 | 0.000 | lnsoy1 | 0.000 | lnsoy1 | 0.000 |
| lnppi1 | 0.137 | lnppi1 | 0.117 | lnppi1 | 0.010 | lnppi1 | 0.007 |
| lnm21 | 0.002 | lnm21 | 0.002 | lnm21 | 0.006 | lnm21 | 0.001 |
| lnrf1 | 0.756 | lnrf1 | 0.742 | lnrf1 | 0.699 | lnrf1 | 0.112 |
| lnex1 | 0.010 | lnex1 | 0.017 | lnex1 | | lnex1 | |
| lnnx1 | 0.486 | lnnx1 | | lnnx1 | | lnnx1 | |

[1] see appendix for detailed information

As we can see, after adjusting model four times according to P values, we finally get desired model:

predictionlnsoy=0.7418516*lnsoy1+0.8907193*lnex1-0.2284704*lnm21+0.0161209*lnrf1+3.663025

## 1.2 Gold

In the regression analysis of gold futures, I put equal attention on both significance level and model construction. So the regression process goes down to two steps:

codes:

```
### regress lngold lngold1 lnrf1 lnex1 ###
### regress lngold lngold1 lnm21 lnex1 ###
```

results:

| lngold | P>\|t\| | lngold | P>\|t\| |
|---|---|---|---|
| lngold1 | 0.000 | lngold1 | 0.000 |
| lnex1 | 0.049 | lnrf1 | 0.008 |
| lnrf1 | 0.703 | lnm21 | 0.081 |

This time we only adjust the model twice, for a better control variable lnm21 instead of lnrf1. And our model is:

predictionlngold=0.879263*lngold1-0.0529191*lnm21+0.4074437*lnex1+.6605369

## 1.3 Copper

In the regression analysis of copper futures, I purely focus on the model construction, which means I do not see the regression result and adjust it accordingly, I just construct one model based on insight knowledge and let it be it.

codes:

```
### regress lncu lncu1 lncu2 lncu3 lnppi1 lnnx1 ###
```

results:

| lncu | P>\|t\| |
|---|---|
| lncu1 | 0.000 |
| lncu2 | 0.001 |
| lncu3 | 0.018 |
| lnppi1 | 0.053 |
| lnnx1 | 0.616 |

Based on economic insight knowledge, I choose the variables that is most likely to explain future prices, and not all of the them turned out to be perfectly significant. But in order to avoid data snooping, I do not change my model:

predictionlncu=1.34122*lncu1-0.6953517*lncu2+0.3191183*lncu3+1.87237*lnppi1-0.0032452*lnnx1-8.190684

# 2 Testing

## 2.1 Soybean

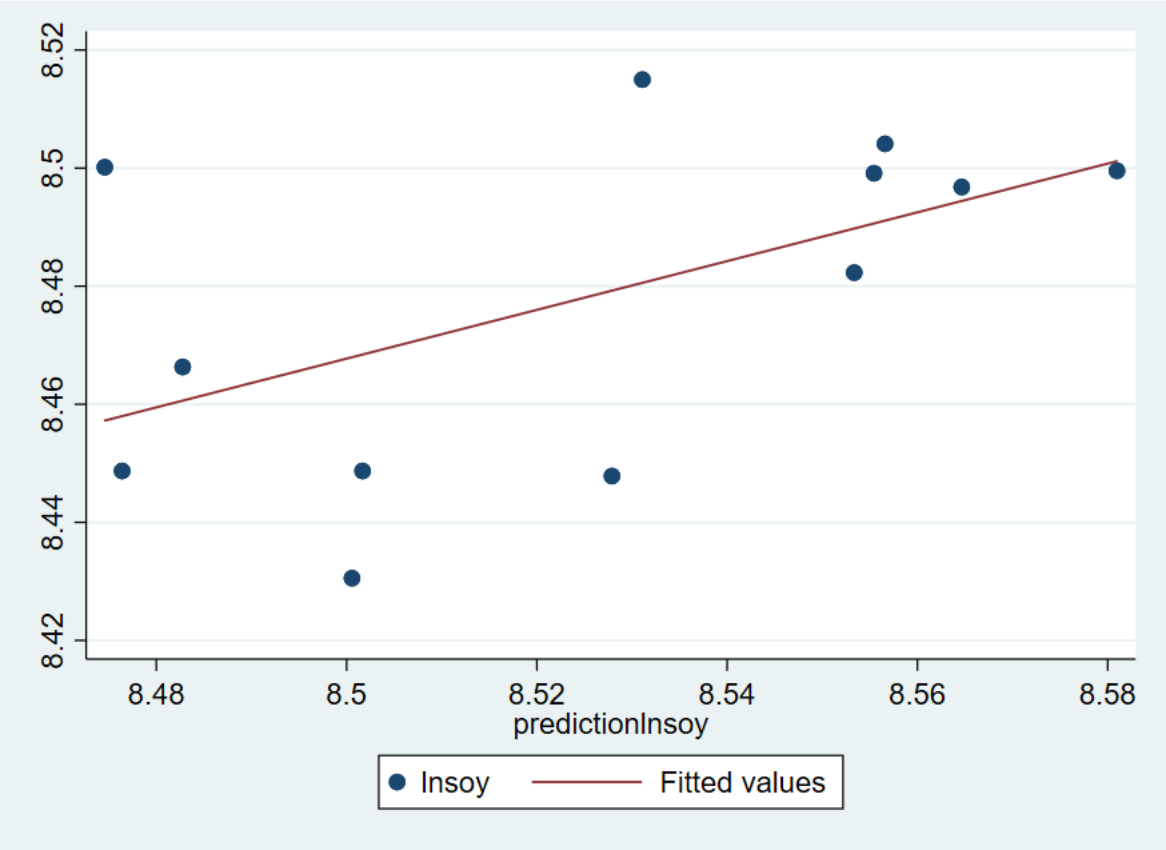
Based on the model finally constructed, we can get the predicted value of lnsoy of latter time period. And by comparing the predicted values and real values, we can know how well the model fits.

codes:

### gen predictionlnsoy=.7418516*lnsoy1+.8907193*lnex1-.2284704*lnm21+.0161209*lnrf1+3.663025 ###

### twoway scatter lnsoy predictionlnsoy || lfit lnsoy predictionlnsoy ###

results:




As we can see, although the model has great significance, which means it has well minimized in-sample error, its out-of-sample error are considerably large, since the predicted values does not fit well with real values.

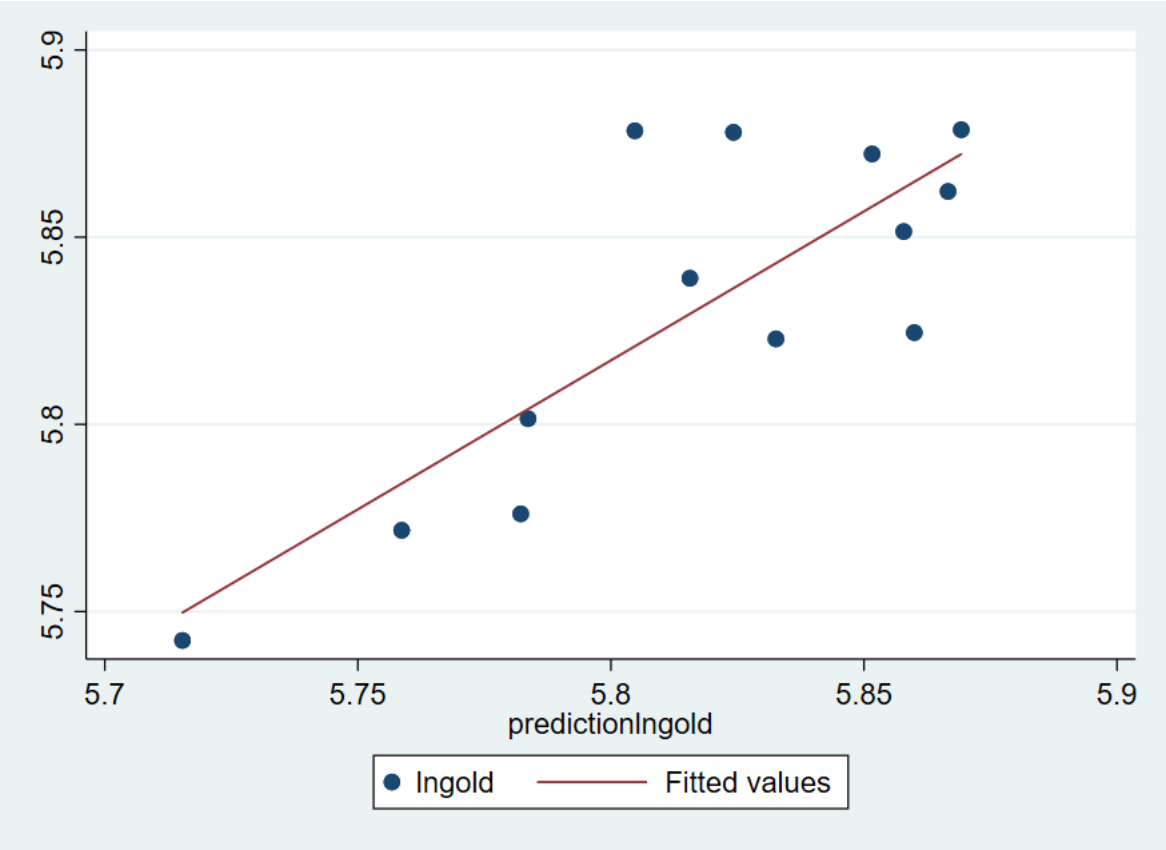
## 2.2 Gold

Like what we do with soybean, we also get the predicted values and compare them with real values.

codes:

```
### gen predictionlngold=.879263*lngold1-.0529191*lnm21+.4074437*lnex1+.6605369
###
### twoway scatter lngold predictionlngold || lfit lngold predictionlngold
###
```

results:




As we can see, the fitting is still not very satisfying, but compare to the results of soybean, it is slightly better, with more points distributed near the line.

## 2.3 Copper

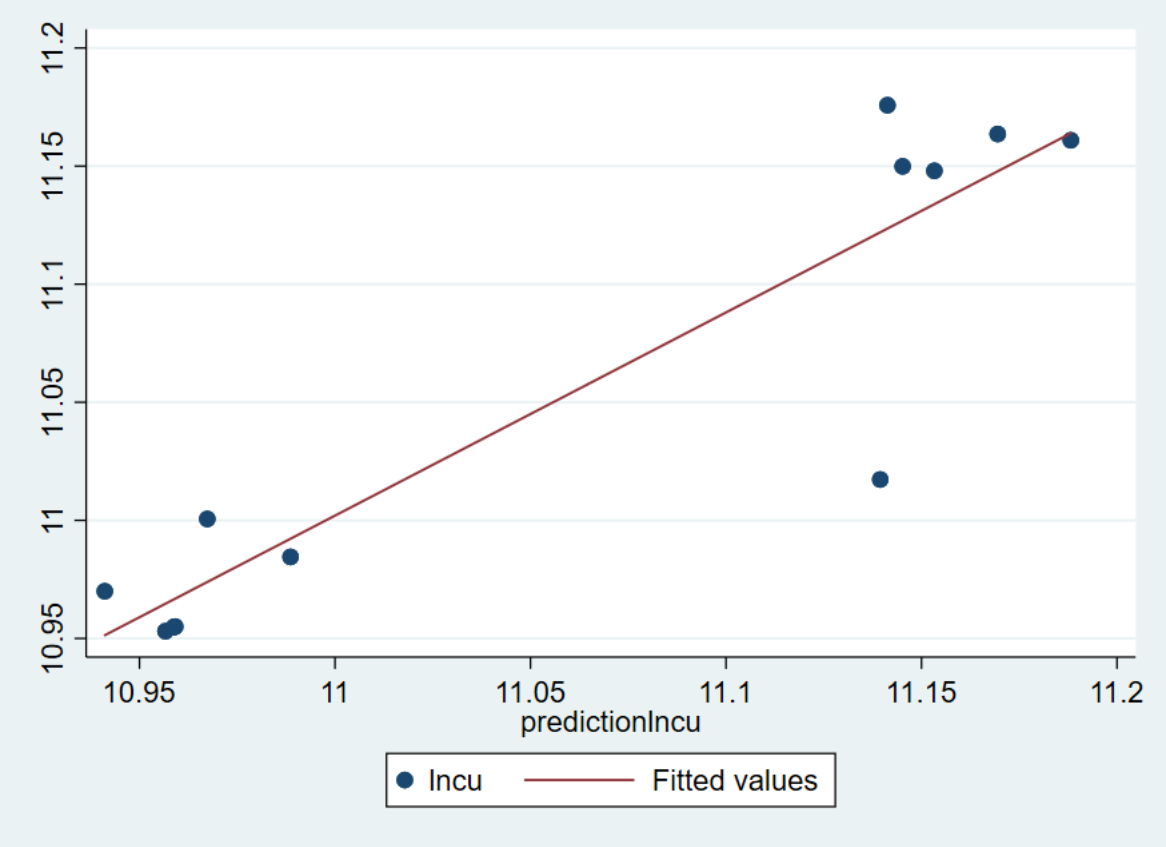
Repeat the testing process for copper and see what we get.

codes:

### gen predictionlncu=1.34122*lncu1-.6953517*lncu2+.3191183*lncu3+1.87237*lnppi1-.0032452*lnnx1-8.190684 ###

### twoway scatter lncu predictionlncu || lfit lncu predictionlncu ###

results:




Surprisingly, although this model is unpolished, determined purely by personal understanding of insight knowledge, it gives the best fit, with most points scattered near the line.

## 3 Interpretation

In empirical regression analysis, especially for us green hands, we tend to adjust our model based on the results of our data. But the model construction process should be built on insight knowledge, because we need to find causality instead of pure correlation. And blindly data snooping will lead to seemingly good-looking but actually inaccurate results.

# 4 Appendix[23]

## 4.1 Codes

Soybean:

### regress lnsoy lnsoy1 lnppi1 lnnx1 lnex1 lnrf1 lnm21 ###

### regress lnsoy lnsoy1 lnex1 lnm21 lnppi1 lnrf1 regress lnsoy lnsoy1 lnex1 lnm21 lnrf1 ###

### regress lnsoy lnsoy1 lnex1 lnm21 lnppi1 ###

### gen predictionlnsoy=.7418516*lnsoy1+.8907193*lnex1-.2284704*lnm21+.0161209*lnrf1+3.663025 ###

### twoway scatter lnsoy predictionlnsoy || lfit lnsoy predictionlnsoy ###

Gold:

### regress lngold lngold1 lnrf1 lnex1 ###

### regress lngold lngold1 lnm21 lnex1 ###

### gen predictionlngold=.879263*lngold1-.0529191*lnm21+.4074437*lnex1+.6605369 ###

### twoway scatter lngold predictionlngold || lfit lngold predictionlngold ###

Copper:

### regress lncu lncu1 lncu2 lncu3 lnppi1 lnnx1 ###

### gen predictionlncu=1.34122*lncu1-.6953517*lncu2+.3191183*lncu3+1.87237*lnppi1-.0032452*lnnx1-8.190684 ###

### twoway scatter lncu predictionlncu || lfit lncu predictionlncu ###

## 4.2 Summary of variables

lnsoy: the logarithm of soybean futures prices.

lnsoy1: the logarithm of soybean futures prices in the previous time period.

lngold: the logarithm of gold futures prices.

lngold1: the logarithm of gold futures prices in the previous time period.

lncu: the logarithm of copper futures prices.

lncu1: the logarithm of copper futures prices in the previous time period.

lncu2: the logarithm of copper futures prices in the second previous time period.

lncu3: the logarithm of copper futures prices in the third previous time period.

lnex1: the logarithm of exchange rate between US dollars and RMB in the previous time period.

lnppi1: the logarithm of product price index in the previous time period.

lnm21: the logarithm of money supply in the previous time period.

lnnx1: the logarithm of net export in the previous time period.

lnrf1: the logarithm of risk-free rate in the previous time period.

[2] see here for stata do filestata do file

[3] see here for stata datasetstata dataset-regressionstata dataset-testing