Xiaohe Huang

Project: Analysis on US Commercial Banks’ Net Income

1. Where is the data from?

The data is from Federal Deposit Insurance Corporation(FDIC).

<http://www2.fdic.gov/hsob/HSOBRpt.asp>

This data reflects the income statements from 1934 to 2012 of FDIC-Insured Commercial Banks in US and other areas.

x1=Year; x2= No. of Inst.; x3=Total Interest Income;

x4=Total Interest Expense; x5=Net Interest Income ;

x6=Total Non-Interest Income; x7=Total Non-Interest Expense x8=Provision For Loan and Lease Losses;

x9=Pre-tax Net Operating Income; x10=Securities Gains/Losses(-)

x11=Applicable Income Taxes; x12=Net Extra-ordinary Items

x13=Net Income

Here we store data of x1 and x13 into a file named”project”.

Input this data:

>Bank=read.table("F:/project.txt",header=T,na.strings="NA",sep="")

Draw a graph for Net Income.

>plot(Bank$x1,Bank$x13,xlab="year",ylab="Dollar amounts in thousands",main="Graph for Net Income",lty=1, type="l",col=1)



According to the graph, the net income of commercial banks fluctuates, and this fluctuation may have regularity, considering their similar feature.

Now, we pick up Net Income data of 1958~1988 and 1980~2010 as two samples to discuss their similarity and distribution.

> Sample1=subset(Bank$x13, Bank$x1>=1958&Bank$x1<1988)

> Sample2=subset(Bank$x13, Bank$x1>=1980&Bank$x1<2010)

> Data=data.frame(Sample1,Sample2)

> Data

Sample1 Sample2

1 2803219 -11637577

2 17417674 15258818

3 17977015 97583411

4 15501811 128200283

5 14931000 113920930

6 14996000 103984060

7 14803000 102570524

8 14010000 89439583

9 12838000 74131124

10 10759000 70795046

11 8879000 71527729

12 7843000 61885315

13 7255000 59159284

14 7078920 52324224

15 6581866 48776028

16 5655762 44624072

17 5236361 43041783

18 4837040 31990373

19 4333850 17935435

20 3786000 15991078

21 3508846 15574853

22 3039647 24812194

23 2861412 2803219

24 2602000 17417674

25 2393000 17977015

26 2348000 15501811

27 2374000 14931000

28 2257000 14996000

29 1553000 14803000

30 2082000 14010000

2. Purpose

According to the graph, there may be a regular pattern which the banks’ net income follows. For example, the net income graph of sample 1 is very similar to net income graph of sample 2. Does that pattern really exist? Are Sample1 and Sample2 in normal distribution？ Do they have same distribution? Do they have any relationship? To discuss these questions is the purpose of this project.

> plot(1:30,Sample1,ylab="Dollar amounts in thousands",main="Graph for Net Income",lty=1, type="l",col=1)

> par(new=TRUE)

> plot(1:30,Sample2,lty=1, type="l",col=2)

>legend(15,8.0e+07,legend=c("Sample1","Sample2"),lty=1,col=c(1,2))



1. Analysis on Net Income
2. Summary

> summary(Sample1)

Min. 1st Qu. Median Mean 3rd Qu. Max.

1553000 2818000 5446000 7418000 12320000 17980000

> summary(Sample2)

Min. 1st Qu. Median Mean 3rd Qu. Max.

-11640000 15520000 37520000 46140000 71340000 128200000

According to summaries of two Samples, it is inaccurate to make any conclusion since the 5 values of results are different between Sample1 and Sample2.

1. Discuss Distribution

(a)Are they have normal distribution?

> library(nortest)

> lillie.test(Sample1)

Lilliefors (Kolmogorov-Smirnov) normality test

data: Sample1

D = 0.162, p-value = 0.04313

> lillie.test(Sample2)

Lilliefors (Kolmogorov-Smirnov) normality test

data: Sample2

D = 0.2068, p-value = 0.002087

According to the results above, both p-value<0.05, therefore reject the assumption. Both Sample1 and Sample2 are not in normal distribution.

(b)Do they have same distribution?

> wilcox.test(Sample1,Sample2)

Wilcoxon rank sum test with continuity correction

data: Sample1 and Sample2

W = 85, p-value = 7.065e-08

alternative hypothesis: true location shift is not equal to 0

Warning message:

In wilcox.test.default(Sample1, Sample2) :

cannot compute exact p-value with ties

Because p-value<0.05, therefore, reject the assumption. As a result, the two samples do not come from a same distribution.

1. Testing for Correlation Coefficient

(a)Pearson Correlation Coefficient

> cor.test(Sample1,Sample2)

Pearson's product-moment correlation

data: Sample1 and Sample2

t = 7.8618, df = 28, p-value = 1.456e-08

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.6693885 0.9160701

sample estimates:

cor

0.829593

Because p-value<0.05, therefore, reject these assumption:

I They must be approximately Gaussian distributed.

I There must be a significant linear relationship between them.

I They must be either interval or ratio measurements.

I There may not be any outliers.

I They must have similar variances.

(b)Spearman Correlation Coefficient

> cor.test(Sample1,Sample2,method="s")

Spearman's rank correlation rho

data: Sample1 and Sample2

S = 754, p-value = 8.749e-07

alternative hypothesis: true rho is not equal to 0

sample estimates:

rho

0.8322581

Because p-value<0.05,therefore reject these assumption:

I They must be rank ordered.

I They are monotonically related.

I They need not be Gaussian distributed.

I They do not require the parameters of distribution.

I They do not require that the relationship between them being linear.

I They do not require to be measured on interval or ratio scale.

(c)Kendall’s Tau Correlation Coefficient

> cor.test(Sample1,Sample2,method="k")

Kendall's rank correlation tau

data: Sample1 and Sample2

T = 381, p-value = 3.452e-11

alternative hypothesis: true tau is not equal to 0

sample estimates:

tau

0.7517241

Because p-value<0.05, therefore reject these assumptions:

1. Discuss Regression
2. Create a scatterplot of the data to investigate the relationship between Sample1 and Sample2.

> model=lm(Sample1~Sample2)

> plot(Sample1,Sample2,main="scatter plot of the Net Income")

> abline(model)



According to the graph, there must be a significant linear relationship between Sample1 and Sample2.

(b)State the estimated regression model.

> model=lm(Sample1~Sample2)

> model

Call:

lm(formula = Sample1 ~ Sample2)

Coefficients:

(Intercept) Sample2

1.946e+06 1.186e-01

Therefore, resulting least square line is^*y* = 1.946e+06 + 1.186e-01*x*

(c)Calculate 90% confidence interval for the regression parameters.

> confint(model,level=0.9)

5 % 95 %

(Intercept) 4.291961e+05 3.461884e+06

Sample2 9.293458e-02 1.442580e-01

**Correlation and**

1. Conclusion.

Although the graphs of Sample 1 and Sample 2 looks similar, it does not mean they have any direct relationships. Actually, they are not in normal distribution, and also not in same distribution. Meanwhile, according to Testing for Correlation Coefficient, we discover we should reject the assumptions of Pearson Test, Spearman Test and Kendall’s Tau Test. However, when we discuss regression, we see there must be a significant linear relationship between Sample1 and Sample2. And the resulting least square line is^*y* = 1.946e+06 + 1.186e-01*x.*