Econometrics 322 Lab #1

Introduction to Jupyter

Prof. Paczkowski

Enter your Name and RUID in the Next Cell

Maanya Tandon (RUID mt958)

Grading Rubric

Score:___: Max(0, 20 - Total Deductions)

Content Area	Deduction	Times Deducted	Check	Comments
Abstract				
Missing	5		[ ]	
Insufficient/Wrong Focus	1		[ ]	
Data Dictionary (Metadata)				
Missing	5		[ ]	
Insufficient/Wrong Form or Wording	1		[ ]	
Graphs				
Missing	5		[ ]	
Missing Title	1		[ ]	
Missing/Wrong Labels	1		[ ]	
Pre-Lab				
Missing	5		[ ]	
Insufficient/Wrong Answer	2 Each		[ ]	
No/Incorrect/Insufficient Model Specification	2		[ ]	
No/Incorrect Statistical Hypothesis Statement	2 Each		[ ]	
Post-Lab				
Missing	5		[ ]	
Insufficient/Wrong Answer	2 Each		[ ]	
Correlations				
Missing	5		[ ]	
Insufficient/Wrong Analysis	2		[ ]	
Missing Graph	2		[ ]	
Estimations				
Missing	5		[ ]	
No or incorrect discussion/interpretation of...				
Hypothesis tests and p-values	2 Each		[ ]	
R^2	2		[ ]	
F-Statistic	2		[ ]	
Multicollinearity/VIF	2		[ ]	
Heteroskedasticity/Test	2		[ ]	
Autocorrelation/Test	2		[ ]	
No/insufficient model selection	2		[ ]	
Elasticities				
Missing	5		[ ]	
Incorrect Interpretation	2		[ ]	
Missing Summary Table	2		[ ]	
Model Portfolio			[ ]	
Missing	5		[ ]	
General Comments:				

Contents

Collaboration Policy

Back to Contents

```
1. Study groups are allowed but I expect students to understand and complete their own assignmen
ts and to hand in one
assignment per student.
2. If you worked in a group, please put the names of your study group in the following table.
3. Just like all other classes at Rutgers, the student Honor Code is taken seriously.

The submitted assignment must be your work.
```

Collaborator(s) Name(s)
name(s) here

Introduction

Back to Contents

Purpose

Back to Contents

The purpose of this lab is to introduce you to the Jupyter notebook paradigm using actual data.

At the end of this lab, you will be able to:

1. collect and clean data collect from a web site location;
2. document your data;
3. describe your data in terms of its source and domain;
4. write a tentative econometric model and testable hypothesis;
5. load Python packages;
6. import your data;
7. do some basic calculations; and
8. create, graph, and interpret a correlation matrix.

Problem

Back to Contents

President Coolidge once remarked that "*The chief business of the American people is business.*" (Speech to the American Society of Newspaper Editors (January 17, 1925)). It goes without saying that a driving factor in the health and prosperity of the U.S. economy is the business, or private, sector. This is where many jobs are created and lost. It's a dynamic sector with new businesses constantly being created while others are constantly disappearing. What is the pattern of business creation and failure over time in the U.S.? What factors contribute to the creations and failures?

Assignment

Back to Contents

Collect data from the *Economic Report of the President*: 2004, Table B-96. You just need New Business Incorporations, Total Number of Failures, and Total Amount of Current Liabilities for 1955 to 1997. Import the data into a DataFrame. Name the variables as you wish but whatever names you use, document them in the Data Dictionary. See the class notes for examples of data dictionaries.

HINT: First, Google the *Economic Report*. Second, download the Excel file and clean it appropriately. You do **not** have to enter any data by hand. Students frequently type the data into Excel -- do **not** do this!!

Documentation

Back to Contents

Abstract

Back to Contents

This lab helped me understand the relationship between the number of new business incorporations, the number of failed businesses, and the amount of liabilities more specifically over the period 1955-1997. There is a very strong correlation between the amount of liabilities and the number of failed businesses. As the amount of liabilities went up in the 80s, so did the number of failed business incorporations. We can see in the data that there is a strong likelihood that the businessed incurred too much debt that they were unable to recover from, and they failed as a result.

Data Dictionary

Back to Contents

Variable	Values	Source	Mnemonic
New Business Incorporations	Number	Economic Report of the President, 2004	numNewBus
Total Number of Failures	Number	Economic Report of the President, 2004	numFails
Total Amount of Current Liabilities	Millions $	Economic Report of the President, 2004	numLiabs
Total Number of Failures / Total Amount of Current Liabilities	Percentage as a decimal	Calculated	ratio
Percentage change in failures	Percentage expressed as a number	Calculated	pchChangeLiabs
Percentage change in incorporations	Percentage expressed as a number	Calculated	pchChangeNewBus

Pre-Lab Questions

Back to Contents

Before you do any work, please think about the concepts: *business incorporations*, *failures*, and *liabilities* in the U.S. from 1955 to 1997. In particular, think how you would answer the following if called on in class.

What type of data is this and why (i.e., source and domain)?

This is a secondary data source because it is data that was collected by the US government. It is from the Economic Report of the President, 2004. It was collected for their purposes of understanding and reporting on the state of the economy. We are using the data for our own purposes to see what the impract of the state of the economy is on the variables we are observing.

What relationships would you expect to exist among these variables? Explain your answer.

I would expect for the number of business failures to increase with the number of new business incorporations. This is because as the number of businesses go up, a certain amount are bound to fail in the stages of being a small business. I would also expect the number of failures to increase as the amount of liabilities go up. The more debt that a business has, the harder it is to pay it back which makes them more likely to fail.

Write a testable hypothesis for failures as a function of liabilities. Explain your answer.

Hypothesis 1: The number of total business failures and the amount of liabilities will have a positive correlation.
Hypothesis 2: The ratio of total failures to amount of liabilities will decrease over time as a certain amount of businesses stabalize and pay back their loans.

Write a linear model for failures as a function of liabilities.

$$totFails = B + m(numLiabs) + \epsilon$$

This linear regression model expresses the total number of business failures as a function of the total number of liabilities.

How do you think these variables will change over time? Do you think they will grow or decline or stay constant? Explain your answer.

In recent years, businesses need to have more money to stay open. But, the more liabilities that a business has, the more unlikely they are to recover from that debt. This means that the number of failed business incorporations will go up as well.

Would the business cycle have any impact on them? Can you state any hypotheses about the business cycle profile of these variables? What is a "business cycle profile?" Explain your answers.

The business cycle describes the stages that a set of variables go through during the business cycle, for example how total liabilities are affected during a depression or expansion. The business cycle would impact the variables that we are observing - number of failed business incorporations, and total liabilities. When a business cycle is in a trough, recession, or depression it is likely that there will be more business failures. When a business cycle is in an expansion or peak there will likely be more new business incorporations.

Can you think of any other variables you could create using just these three variables that may be interesting and meaningful? Describe them and state why they would be interesting and meaningful.

If we want to study the relationships meaningfully - we ought to calculate and do regressions for the "lagged percentage changes in liability" versus "percentage changes in the number of business failures" year-over-year over a period of time. Also liabilities and comparing it to the number of new incorporations or total failures can show us the correlation between liabilities and failures.

Tasks and Questions

Back to Contents

Load the Pandas package and give it an alias.

In [1]:
```
##
## Load analysis packages
##
import pandas as pd
import numpy as np
##
##  load graphic packages
##
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
##
##  modeling packages
##
import statsmodels as sm
##
```

Import your data.

In [2]:
```
##
## Enter the code here
path = r'/Users/maanyatandon/Documents/fall2020/econometrics/Lab1/'
file = r'ERP-2004-table96.xls'
path_file = path+file
print(path_file)
data_df = pd.read_excel(path_file, sheet_name = 'Formatted')
##
```

```
/Users/maanyatandon/Documents/fall2020/econometrics/Lab1/ERP-2004-table96.xls
```

Print your data.

Just the first five (5) records is sufficient.

In [15]:
```
##
data_df.head(5)
##
```

Out[15]:

	year	numNewBus	numFails	numLiabs
0	1955	139915	10969	449.4
1	1956	141163	12686	562.7
2	1957	137112	13739	615.3
3	1958	150781	14964	728.3
4	1959	193067	14053	692.8

Reset the row index to have year as the index. Print the first five (5) records again.

In [16]:
```
##
data_df = data_df.set_index('year');
data_df.head(5)
##
```

Out[16]:

	numNewBus	numFails	numLiabs
year			
1955	139915	10969	449.4
1956	141163	12686	562.7
1957	137112	13739	615.3
1958	150781	14964	728.3
1959	193067	14053	692.8

Create a new variable that is the ratio of failures to liabilities. Print the first five (5) records again.

Math function summary:

Operation	Symbol	Example
Exponentiation	exp	e^x = exp(x)
Square root	sqrt	$\sqrt{4}$ = sqrt(4)
Multiplication	*	x * y
Division	/	x/y
Addition	+	x + 1
Subtraction	-	x - 1
Power	**	x^2 = x**2
Lag 1 period	df.X.shift()	df.rgdp.shift()
Percent change	df.X.pct_change()	df.rgdp.pct_change()

Note: exp and sqrt require the numpy or math package. Lag and percent change require the pandas package.

In [3]:
```
##
## Enter the code here
##
import math as math
data_df[ 'ratio' ] = data_df.numFails/data_df.numLiabs
data_df.head()
```

Out[3]:

	year	numNewBus	numFails	numLiabs	ratio
0	1955	139915	10969	449.4	24.408100
1	1956	141163	12686	562.7	22.544873
2	1957	137112	13739	615.3	22.328945
3	1958	150781	14964	728.3	20.546478
4	1959	193067	14053	692.8	20.284353

Interpret the ratio you just created.

Tha ratio is the total number of failures in a year compared to the amount of liabilities. It's the amount of liabilities in millions of dollars that it takes for one business to fail. The higher the debt that the business has taken, the higher is the likelihood of its failure. The ratio will tell us that likelihood.

Create a new variable that is incorporations lagged one period. Print the first five (5) records again.

In [4]:
```
##
data_df[ 'numNewBus_Lagged' ] = data_df[ 'numNewBus' ].shift( 1 );
data_df.head()
##
```

Out[4]:

	year	numNewBus	numFails	numLiabs	ratio	numNewBus_Lagged
0	1955	139915	10969	449.4	24.408100	NaN
1	1956	141163	12686	562.7	22.544873	139915.0
2	1957	137112	13739	615.3	22.328945	141163.0
3	1958	150781	14964	728.3	20.546478	137112.0
4	1959	193067	14053	692.8	20.284353	150781.0

Create a new variable that is year-over-year percent change in incorporations. Print the first five (5) records again.

In [5]:
```
##
data_df[ 'pchChangeNewBus' ] = data_df[ 'numNewBus' ].pct_change()
data_df.head()
##
```

Out[5]:

	year	numNewBus	numFails	numLiabs	ratio	numNewBus_Lagged	pchChangeNewBus
0	1955	139915	10969	449.4	24.408100	NaN	NaN
1	1956	141163	12686	562.7	22.544873	139915.0	0.008920
2	1957	137112	13739	615.3	22.328945	141163.0	-0.028697
3	1958	150781	14964	728.3	20.546478	137112.0	0.099692
4	1959	193067	14053	692.8	20.284353	150781.0	0.280446

Print a correlation matrix.

In [6]:
```
##
data_df.corr()
##
```

Out[6]:

	year	numNewBus	numFails	numLiabs	ratio	numNewBus_Lagged	pchChangeNewBus
year	1.000000	0.975994	0.811147	0.749606	-0.866306	0.974320	-0.143281
numNewBus	0.975994	1.000000	0.816394	0.733417	-0.797055	0.995054	-0.124952
numFails	0.811147	0.816394	1.000000	0.916073	-0.489270	0.829491	-0.248633
numLiabs	0.749606	0.733417	0.916073	1.000000	-0.519645	0.750211	-0.250257
ratio	-0.866306	-0.797055	-0.489270	-0.519645	1.000000	-0.789245	0.095814
numNewBus_Lagged	0.974320	0.995054	0.829491	0.750211	-0.789245	1.000000	-0.209550
pchChangeNewBus	-0.143281	-0.124952	-0.248633	-0.250257	0.095814	-0.209550	1.000000

Post-Lab Questions

Back to Contents

What can you observe about the correlation matrix? Explain.

As the year increments, the number of new businesses, number of failures, and number of liabilities increases because of a high positive correlation. As the amount of current liabilities goes up, the number of failures also go up because they have a 0.92 correlation with each other. As the years go by, the number of failures compared to number of liabilites (ratio) decease because the ratio to year correlation is -0.86. The correlation between the ratio and the number of new businesses lagged is -0.79 which shows that if the number of new businesses in the previous year is low, then the ratio will be high - meaning that the number of failures will be low compared to liabilities.

What is the relationship between failures and incorporations? Explain.

The higher the number of new businesses, the higher the number of failures because the correlation is 0.83. If the number of new businesses increased, the number of failures compared to that will go down. This is seen because the correlation between the percentage change in new businesses and number of failures is -0.25.

Do you see anything to support your testable hypothesis? Explain.

Yes, the correlation matrix supports my hypothesis that the number of total failures and amount of liabilities would have a positive correlation. The correlation between the year and the ratio of failures to liabilities is -0.86 which supports my hypothesis 2.

Does the amount of liabilities have anything to do with incorporations and failures? How and why?

The amount of liabilities and the total number of failures have a correlation of 0.92 which is very high. This makes sense because when a business has a lot of debt, it is harder for them to recover which makes them more likely to fail.
The amount of liabilities and the number of new business incorporations have a correlation of 0.73, which is also high. This makes sense because when approaching the peaks of the business-cycles, when businesses are exapnding by taking on a lot of debt - many more new businesses are also incorporated.

Well done!

Make sure your name is on this notebook at the top and on the file.
Please submit this notebook as a PDF file. Nothing else will be accepted.

In [ ]: