Regression with Python

Charles Carter[1]

Columbus Data Science Meetup

November 8, 2022

[1] ccc31807@yahoo.com

Table of Contents

Introduction

This presentation covers creating linear regression analyses using Python with Jupyter Notebooks. It includes two simple examples. It will also point to getting user input with IPWidgets, and generating scripts from notebooks.

I will focus on preparing and running notebooks, not analysis. I will assume that the objective covers automation of existing processes, and that management (or statisticians) will draw the appropriate conclusions from the results presented.

As important as visualization is, I do not cover visualization. Obviously, you would not submit an analysis without interpreting your conclusions using visuals.

Preliminary Questions

- Are you familiar with Python, i.e., have you written at least one realistic Python program?
- Are you familiar with Jupyter Notebooks?
- Are you familiar with Pandas, Numpy, and other libraries (statsmodels, Patsy, etc.)?
- Do you do regression analysis in your day job?

Where do I get my data?

The Python library statsmodels provides a number of datasets. You can find these at:

`https://www.statsmodels.org/dev/datasets/index.html`

R has influenced the regression libraries in Python, and many R data sets can be inported into Python. You can find these at:

`https://github.com/vincentarelbundock/Rdatasets`

Let's explore two data sets, that you may have used: the Boston housing data and the German credit data.

What we are going to do

We will use a simple data set consisting of three variables: height, weight, and sex. We will use weight as the dependent variable and height and sex as the independent variables. We will fit four different models and compare them.

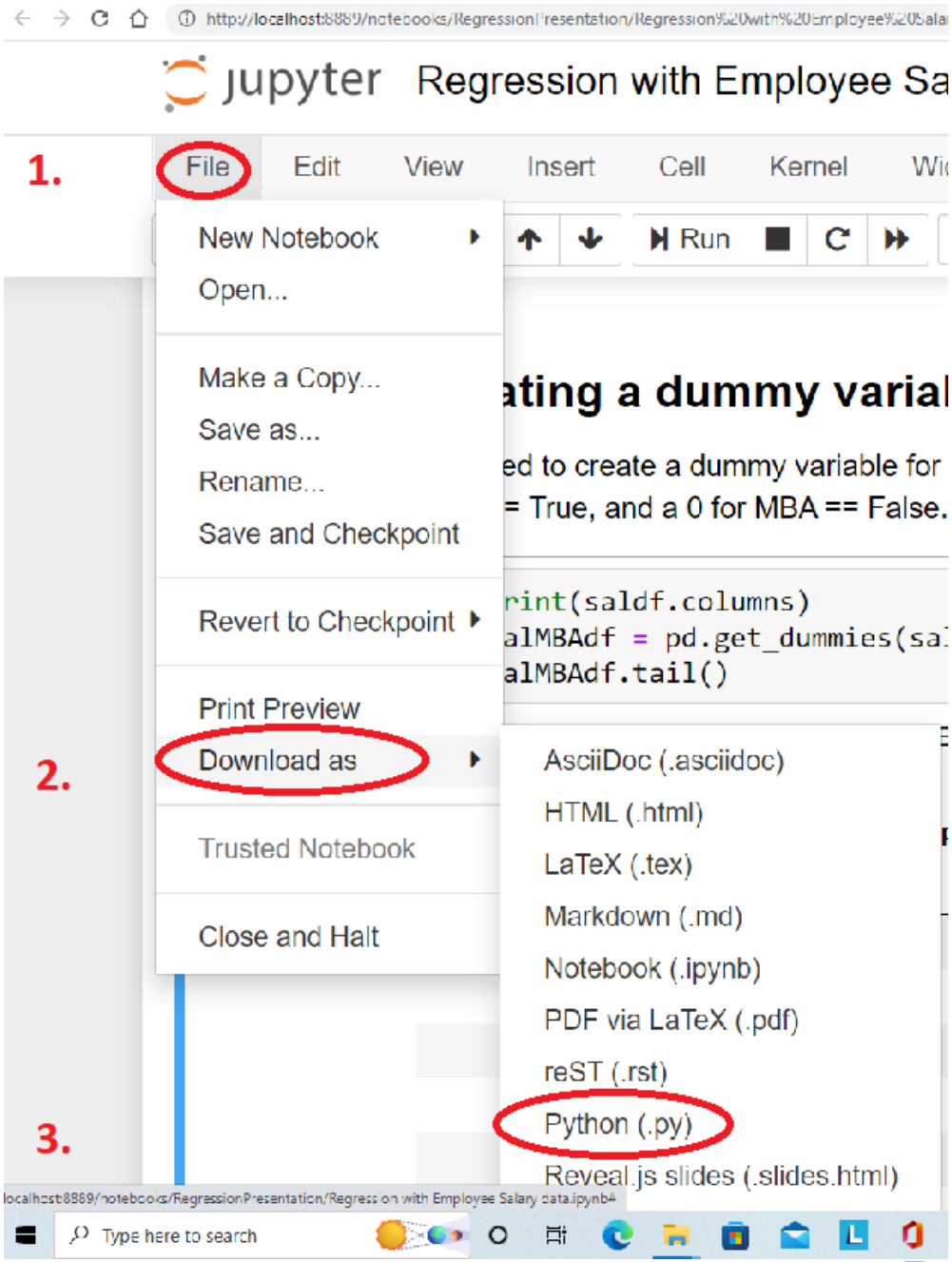
What we are going to do

We will exmplore the effect of age, experience, number of direct reports, and an MBA on salary.

You can save notebooks in various formats, including as a Python file.

1. Go to **File**
2. Go to **Download As**
3. Download as **Python (.py)**
4. This will result in a runnable Python script.

Also, as HTML and PDF. I haven't tried the slides yet.


jupyter Regression with Employee Sa
1.
File
Edit
View
Insert
Cell
Kernel
New Notebook
Open...
Make a Copy...
Save as...
Rename...
Save and Checkpoint
Revert to Checkpoint
Print Preview
Download as
Trusted Notebook
Close and Halt
Run
ating a dummy varial
ed to create a dummy variable for
= True, and a 0 for MBA == False.
rint(saldf.columns)
alMBAdf = pd.get_dummies(sa
alMBAdf.tail()
2.
AsciiDoc (.asciidoc)
HTML (.html)
LaTeX (.tex)
Markdown (.md)
Notebook (.ipynb)
PDF via LaTeX (.pdf)
reST (.rst)
Python (.py)
Reveal.js slides (.slides.html)
3.
Type here to search

Gotchas

You need to **print()** everything you want to go to output.

You cannot include your written analysis in the script (unless you print it).

You will need to run your script in the approprate environment, i.e., Anaconda.

Getting user input in your notebook

Use the Jupyter Widgets library, ipywidgets.

https://ipywidgets.readthedocs.io/en/stable/

https://ipywidgets.readthedocs.io/en/stable/examples/Widget%20List.html#

https://ipywidgets.readthedocs.io/en/7.x/examples/Widget%20Low%20Level.html

https://www.tutorialspoint.com/jupyter/jupyter_notebook_ipywidgets.htm

Gotchas

This library isn't exactly mature.

You will need to do a lot of end user handholding.

This breaks the automation.

Conclusions and Questions

Any questions?