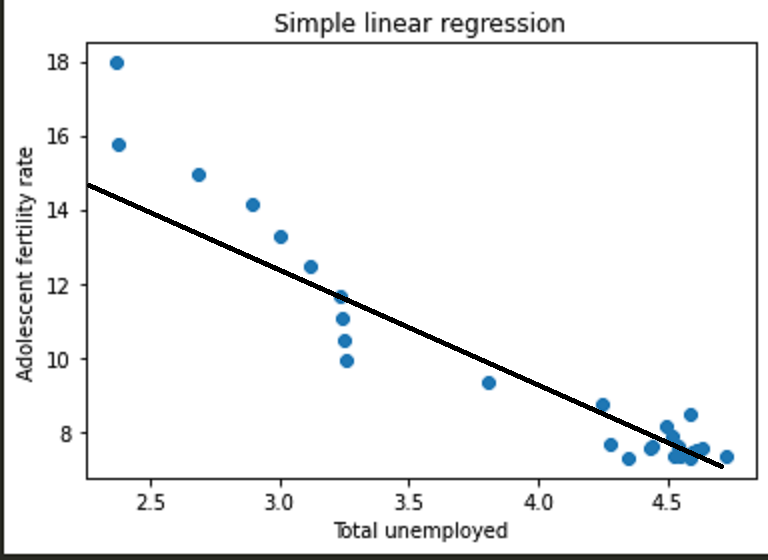
中国青春期生育率与失业人口比率关系

报告人：舒泓瑞 数据科学2班



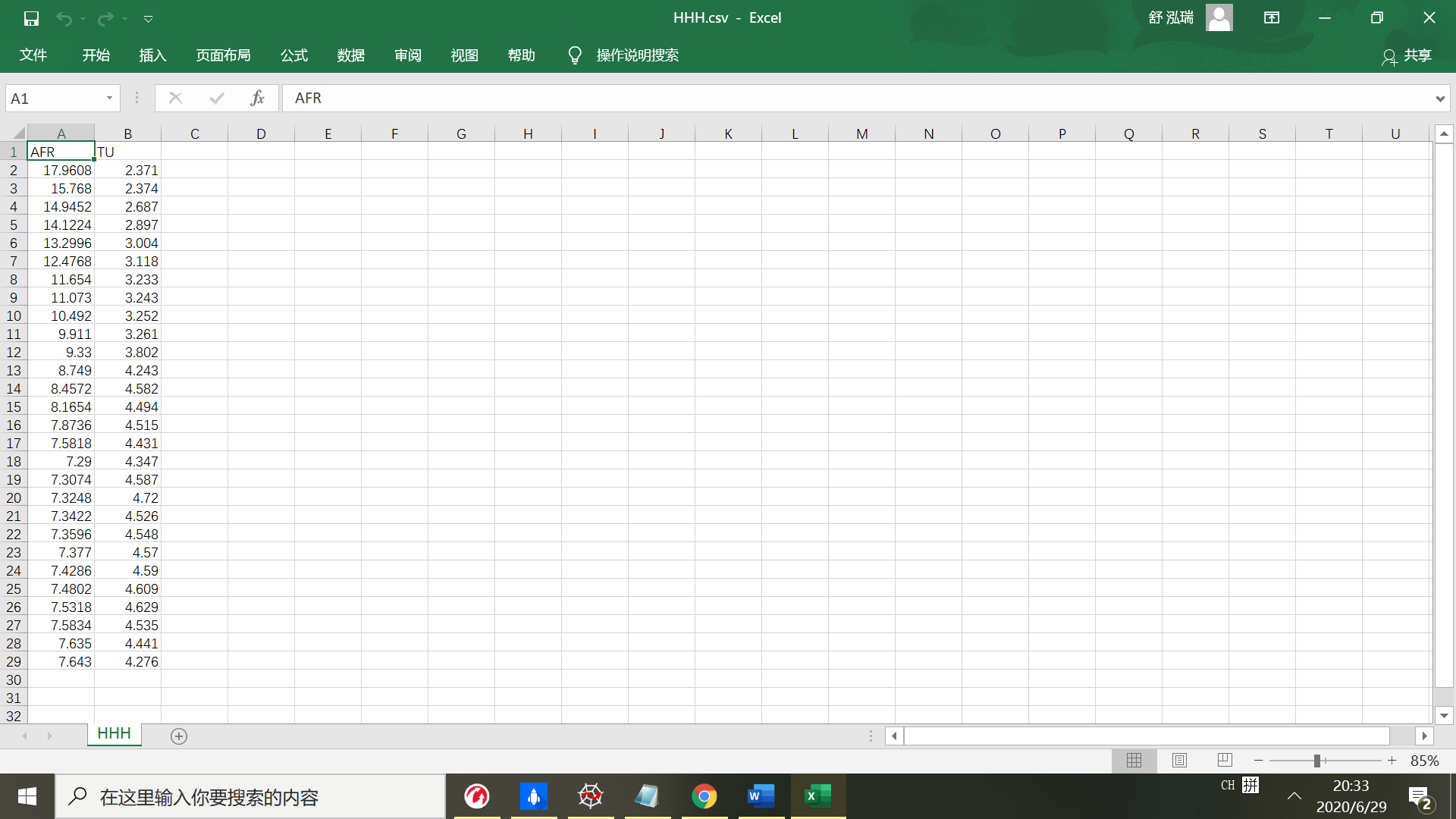
使用工具：python numpy模块，matplotlib模块，pandas模块，spyder模块化集成工具

1. 实验过程：

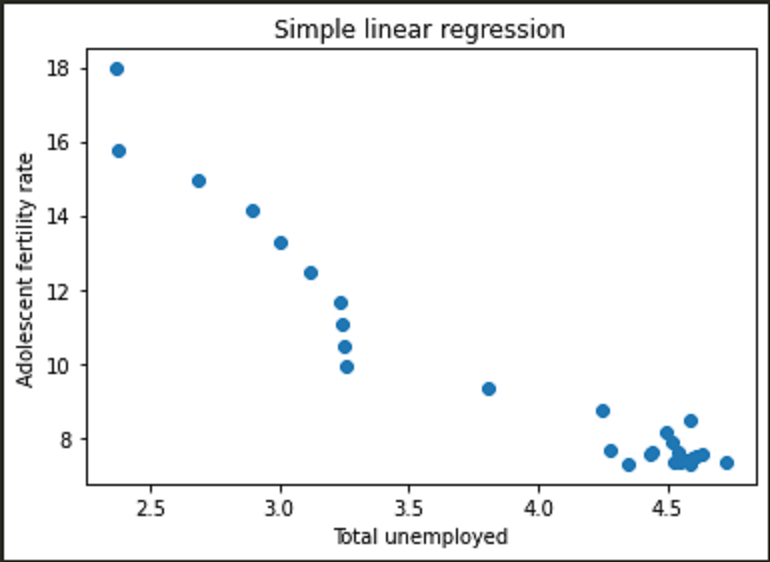
随着时间与科技水平的日益发展，我们所能接触到的数据也日益透明与多元化。作为大数据专业的学生，我们要放开自己的视野，秉承第一代市局从业者的优良传统，遵循胡老师教诲，善于用统计的工具在大数据之海中发现数据之间的关联性。为了完成这一报告，我从世界银行数据库中下载了中国从2000年开始到2018年的青春期生育数据（Adolescent fertility rate (births per 1,000 women aged 15-19)）与失业人口占总劳动人口比率（由世界劳工组织估计）（Total unemployed (as a percentage of total labor force) (simulated by ILO estimates)）并想要发现它们之间的内在联系。

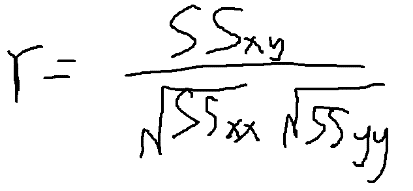
在一开始的想象中，我认为随着青春期生育人口比率的逐渐降低，，失业人口比率应该随之下降。理由：更少的妊娠人口使得更多的劳动力可以投入工作之中，进而降低失业人口比率。

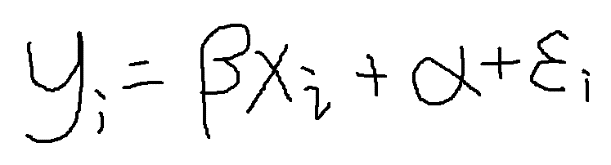
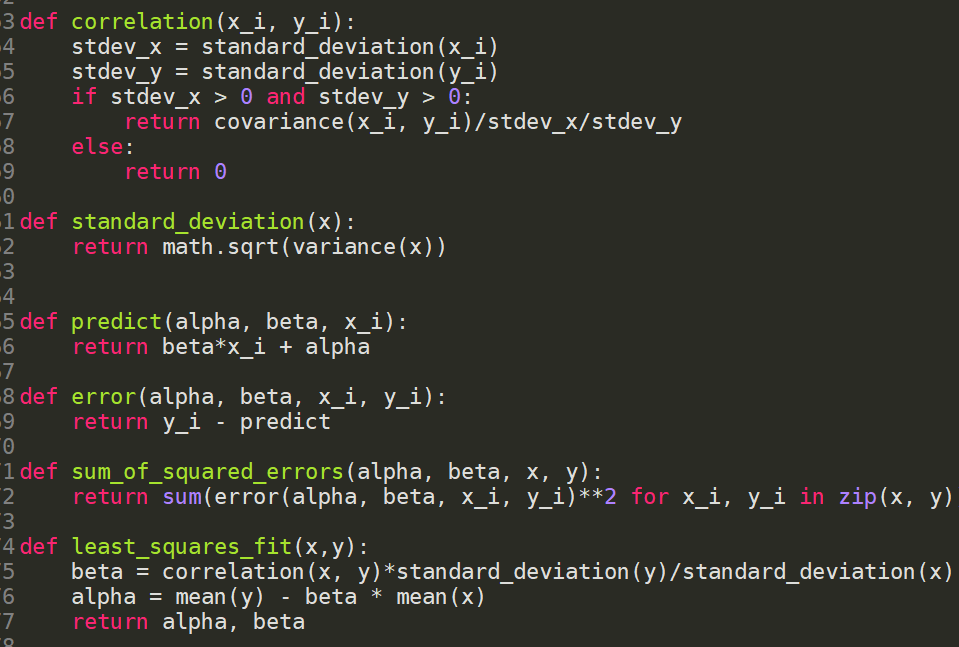
数据附上：



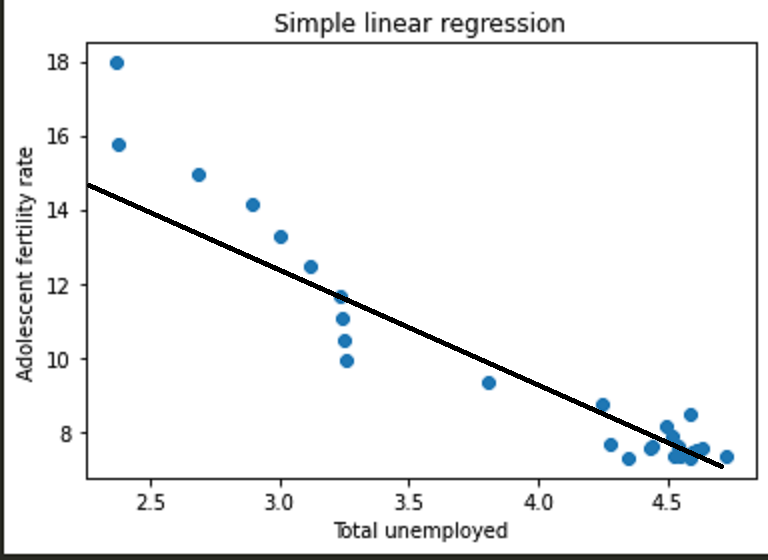
1. 计算与数值

简单线性回归的办法。首先先做出满足此数据的散点图，在此处使用python的matplotlib模块可以简单的将AFR作为Y轴，TU作为X轴得出结果。由此，我们可以模糊的感觉到二者之间可能存在简单回归关系，但是这个关系却和我们一开始预想的不太一样。图如下：

然后开始计算相关性系数2，这里使用python numpy模块将csv文件中的Datafram格式数组化，利用公式求得r = - 0.45.

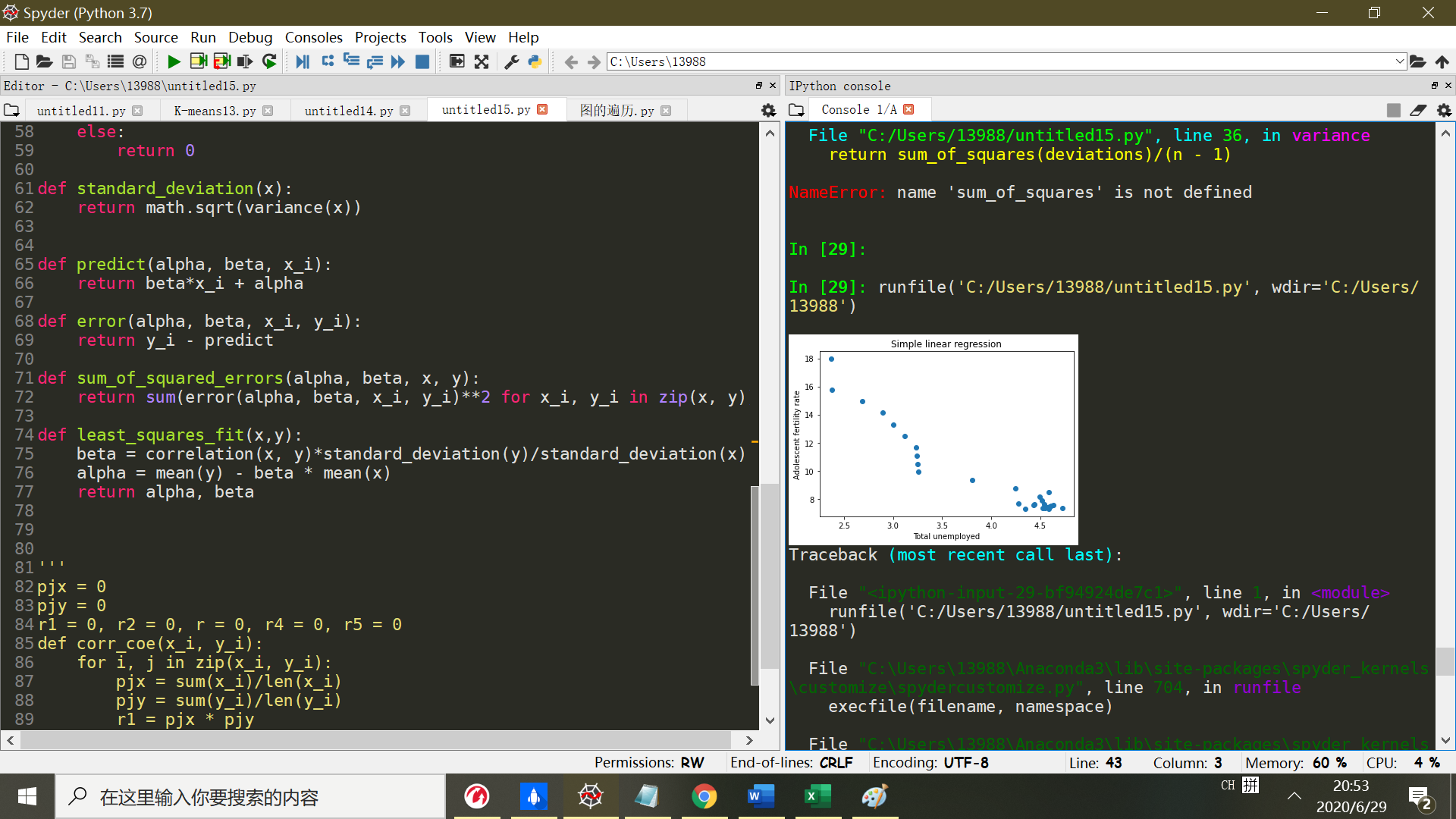
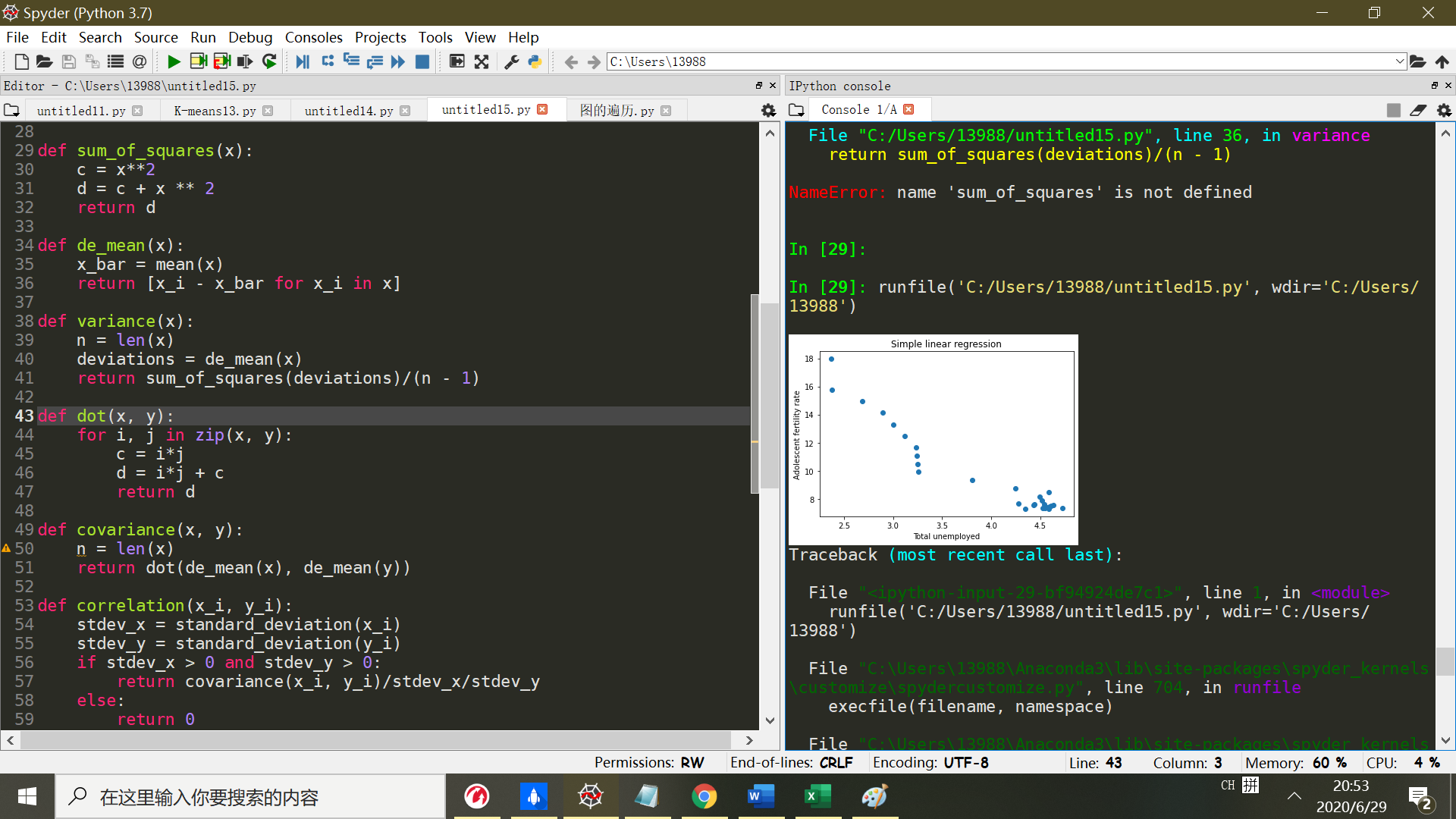
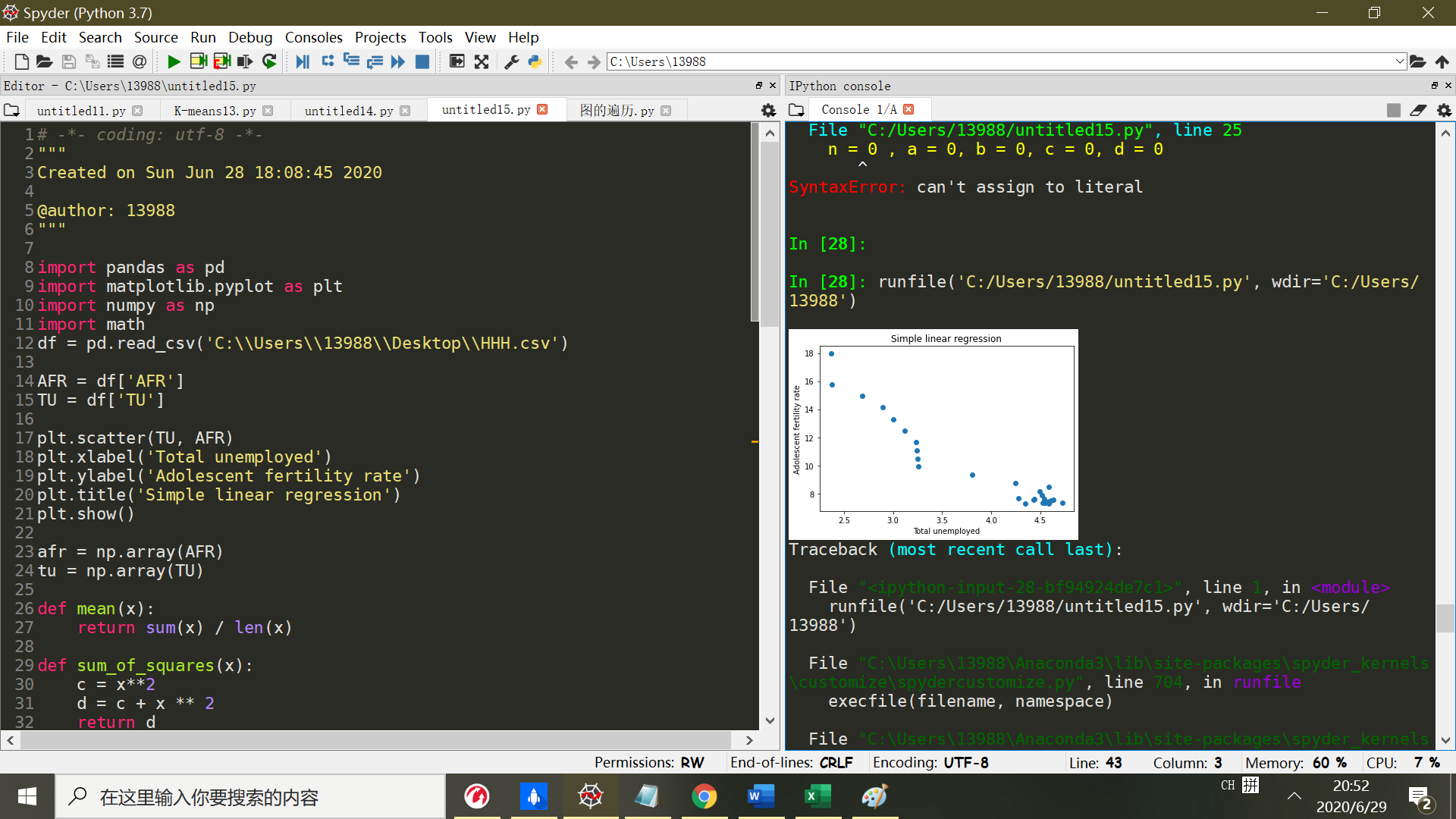
再带入简单线性回归公式定位alpha 与beta 的值在线性关系已确认存在的情况下定位回归方程。通过代码的机器计算：

运行可计算（通过最小二乘法）出alpha = -11.8, beta = 15.762。再利用matplotlib模块的画图功能画出回归曲线，得到成品如下：



1. 代码模块

截图如下



1. 成品检验：

我们可以看到，最终求出的简单线性回归方程为y = -11.8x + 15.762这与我们一开始的预测相差较大。下面我会列出一些可能的原因。

首先是斜率问题，在我们一开始的预测中，该回归方程的斜率应该是正数，即随失业人数增加青春期怀孕人数占比也增加，可是情况恰巧相反，对此，我认为是我这个模型的考虑方向还不够全面，没有考虑全国总人口变化中男女比率的改变。可是如果以人数为基准考虑，就还要考虑GDP增长放缓带来的就业岗位增长趋缓，过于复杂，做不出来。

然后是误差问题，在使用最小二乘法进行α值与β值的计算时，我并没有想到办法来进行误差值σ的计算，即我没有经行回归标准误差的意义测试，所以最后的结果可能存在0.5%以内的误差（不能再大了）。

这次回归分析实验结果基本成功，虽然由于数据与个人能力的不足我无法做到给出精确的回归曲线，我还是认识到了商务统计的魅力（掉头发），并且知道了原来这么扯淡的两个数据集之间都会有联系，果然大千世界无奇不有。