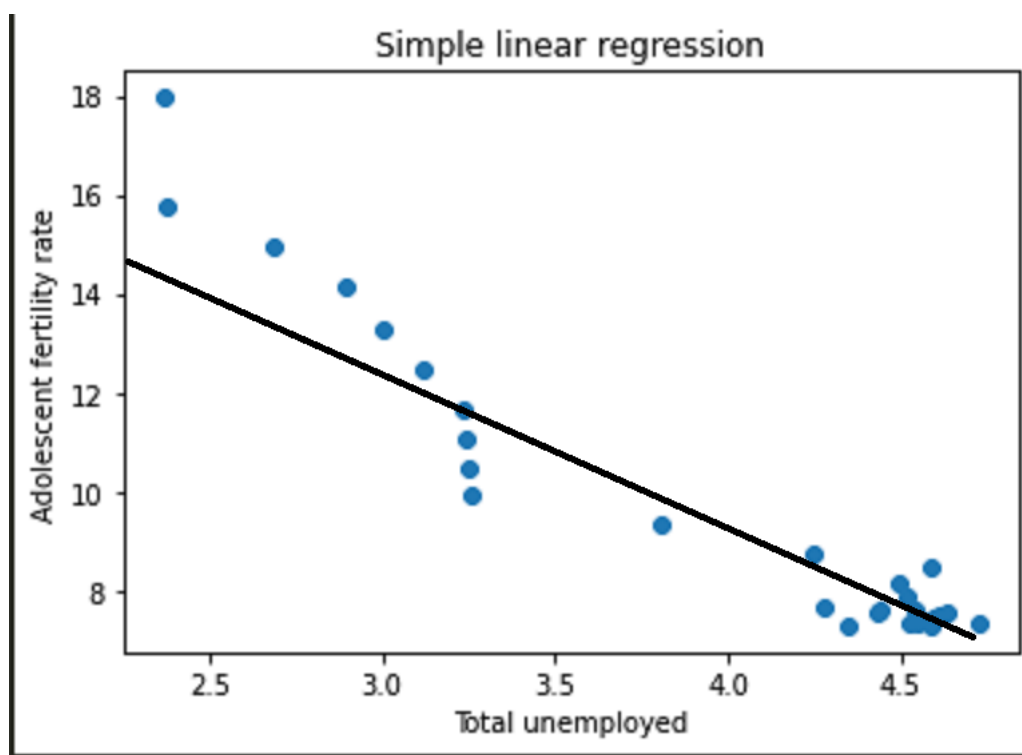


中国青春期生育率与失业人口比率关系

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。

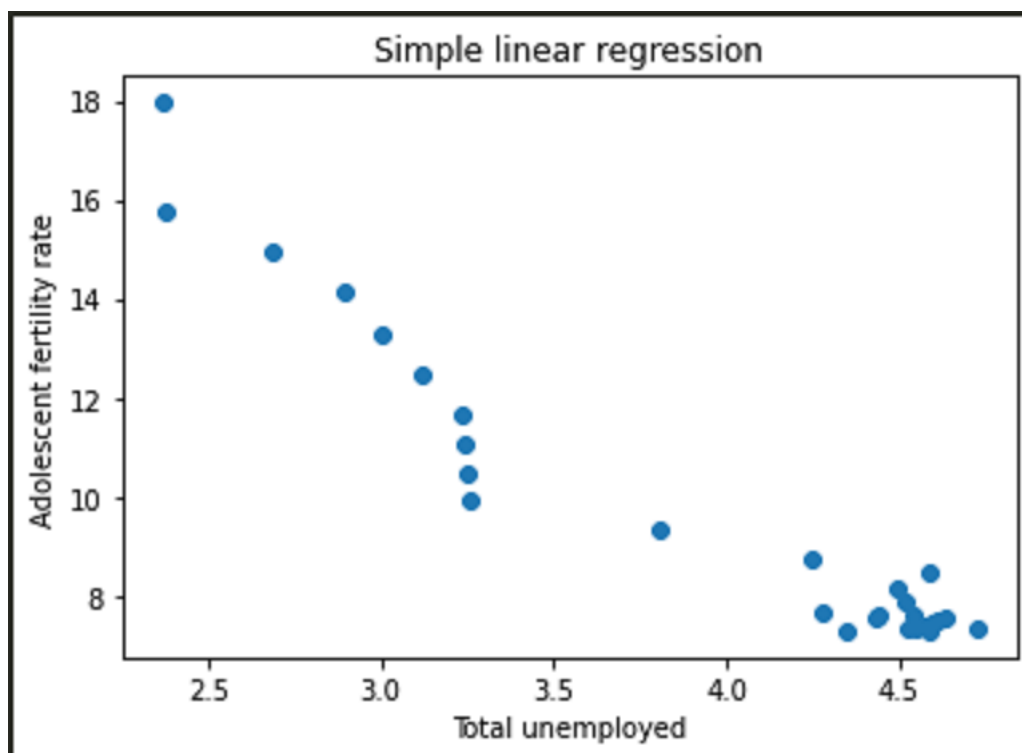
在一开始的想象中，我认为随着青春期生育人口比率的逐渐降低，失业人口比率应该随之下降。理由：更少的妊娠人口使得更多的劳动力可以投入工作之中，进而降低失业人口比率。

数据附上：

	AFR	TU
1	17.9608	2.371
2	15.768	2.374
3	14.9452	2.687
4	14.1224	2.897
5	13.2996	3.004
6	12.4768	3.118
7	11.654	3.233
8	11.073	3.243
9	10.492	3.252
10	9.911	3.261
11	9.33	3.802
12	8.749	4.243
13	8.4572	4.582
14	8.1654	4.494
15	7.8736	4.515
16	7.5818	4.431
17	7.29	4.347
18	7.3074	4.587
19	7.3248	4.72
20	7.3422	4.526
21	7.3596	4.548
22	7.377	4.57
23	7.4286	4.59
24	7.4802	4.609
25	7.5318	4.629
26	7.5834	4.535
27	7.635	4.441
28	7.643	4.276
29		
30		
31		
32		

2. 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 r ，这里使用 python numpy 模块将 csv 文件中的

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}} \sqrt{SS_{yy}}}$$

Datafram 格式数组化，利用公式

求得 $r = -0.45$.

$$y_i = \beta x_i + \alpha + \epsilon_i$$

再带入简单线性回归公式

定位 α

与 β 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

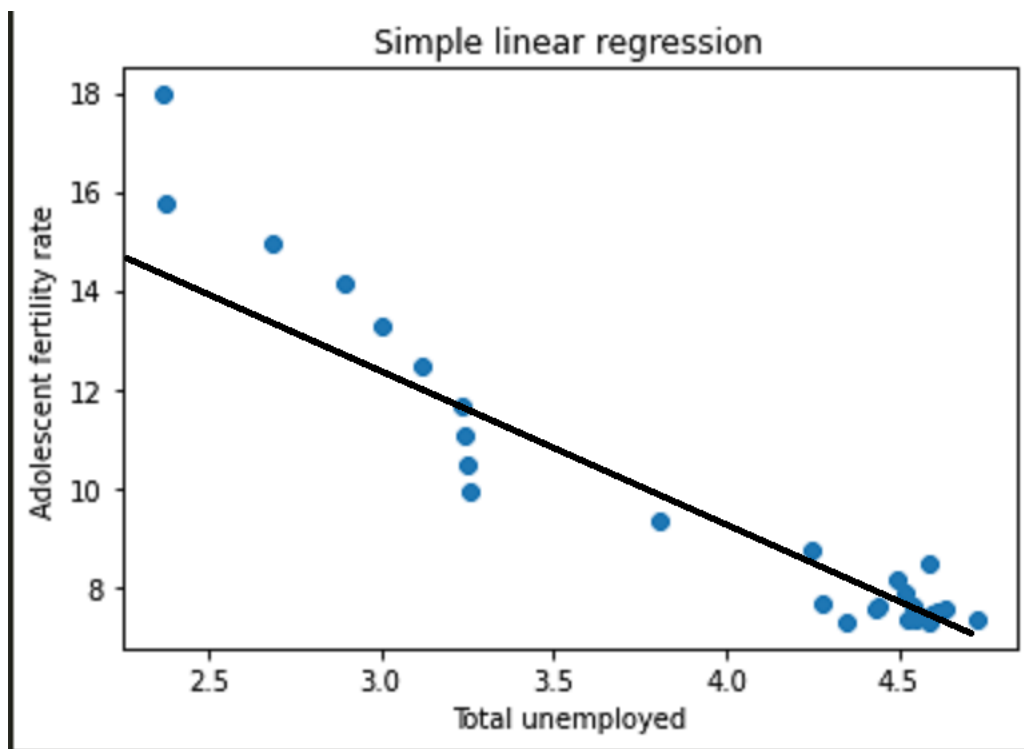
:

```

3 def correlation(x_i, y_i):
4     stdev_x = standard_deviation(x_i)
5     stdev_y = standard_deviation(y_i)
6     if stdev_x > 0 and stdev_y > 0:
7         return covariance(x_i, y_i)/stdev_x/stdev_y
8     else:
9         return 0
10
11 def standard_deviation(x):
12     return math.sqrt(variance(x))
13
14
15 def predict(alpha, beta, x_i):
16     return beta*x_i + alpha
17
18 def error(alpha, beta, x_i, y_i):
19     return y_i - predict
20
21 def sum_of_squared_errors(alpha, beta, x, y):
22     return sum(error(alpha, beta, x_i, y_i)**2 for x_i, y_i in zip(x, y))
23
24 def least_squares_fit(x,y):
25     beta = correlation(x, y)*standard_deviation(y)/standard_deviation(x)
26     alpha = mean(y) - beta * mean(x)
27     return alpha, beta
28

```

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：



3. 代码模块

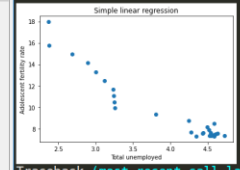
截图如下

```
File Edit Search Source Run Debug Consoles Projects Tools View Help
untitled11.py K-means13.py untitled14.py untitled15.py 图的遍历.py
1# -*- coding: utf-8 -*-
2"""
3Created on Sun Jun 28 18:08:45 2020
4
5@author: 13988
6"""
7
8import pandas as pd
9import matplotlib.pyplot as plt
10import numpy as np
11import math
12df = pd.read_csv('C:\\Users\\13988\\Desktop\\HHH.csv')
13
14AFR = df['AFR']
15TU = df['TU']
16
17plt.scatter(TU, AFR)
18plt.xlabel('Total unemployed')
19plt.ylabel('Adolescent fertility rate')
20plt.title('Simple linear regression')
21plt.show()
22
23afr = np.array(AFR)
24tu = np.array(TU)
25
26def mean(x):
27    return sum(x) / len(x)
28
29def sum_of_squares(x):
30    c = x**2
31    d = c + x ** 2
32    return d
```

File "C:/Users/13988/untitled15.py", line 25
n = 0, a = 0, b = 0, c = 0, d = 0
SyntaxError: can't assign to literal

In [28]:

In [28]: runfile('C:/Users/13988/untitled15.py', wdir='C:/Users/13988')



Traceback (most recent call last):

File "C:/Users/13988/untitled15.py", line 1, in <module>
runfile('C:/Users/13988/untitled15.py', wdir='C:/Users/13988')

File "C:/Users/13988/untitled15.py", line 25, in runfile
execfile(filename, namespace)

File "C:/Users/13988/untitled15.py", line 25, in runfile
execfile(filename, namespace)

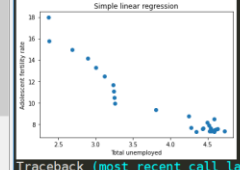
Permissions: RW End-of-lines: CRLF Encoding: UTF-8 Line: 43 Column: 3 Memory: 60 % CPU: 7 %

```
File Edit Search Source Run Debug Consoles Projects Tools View Help
untitled11.py K-means13.py untitled14.py untitled15.py 图的遍历.py
28
29def sum_of_squares(x):
30    c = x**2
31    d = c + x ** 2
32    return d
33
34def de_mean(x):
35    x_bar = mean(x)
36    return [x_i - x_bar for x_i in x]
37
38def variance(x):
39    n = len(x)
40    deviations = de_mean(x)
41    return sum_of_squares(deviations)/(n - 1)
42
43def dot(x, y):
44    for i, j in zip(x, y):
45        c = i*j
46        d = i*j + c
47        return d
48
49def covariance(x, y):
50    n = len(x)
51    return dot(de_mean(x), de_mean(y))
52
53def correlation(x_i, y_i):
54    stdev_x = standard_deviation(x_i)
55    stdev_y = standard_deviation(y_i)
56    if stdev_x > 0 and stdev_y > 0:
57        return covariance(x_i, y_i)/stdev_x/stdev_y
58    else:
59        return 0
```

File "C:/Users/13988/untitled15.py", line 36, in variance
return sum_of_squares(deviations)/(n - 1)
NameError: name 'sum_of_squares' is not defined

In [29]:

In [29]: runfile('C:/Users/13988/untitled15.py', wdir='C:/Users/13988')



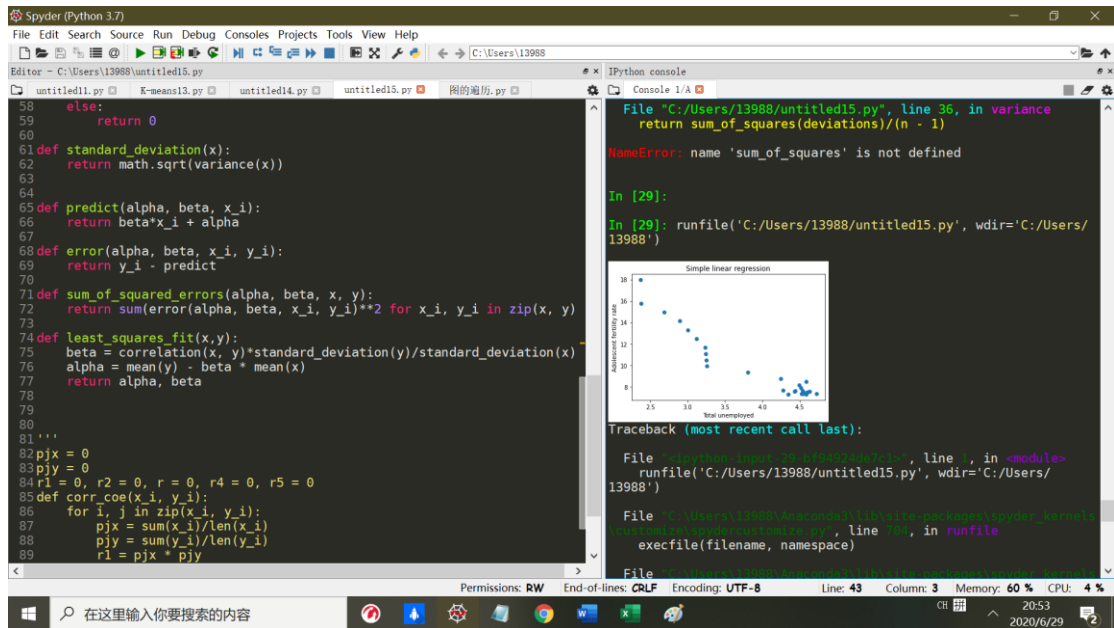
Traceback (most recent call last):

File "C:/Users/13988/untitled15.py", line 36, in variance
return sum_of_squares(deviations)/(n - 1)
NameError: name 'sum_of_squares' is not defined

File "C:/Users/13988/untitled15.py", line 36, in variance
return sum_of_squares(deviations)/(n - 1)
NameError: name 'sum_of_squares' is not defined

File "C:/Users/13988/untitled15.py", line 36, in variance
return sum_of_squares(deviations)/(n - 1)
NameError: name 'sum_of_squares' is not defined

Permissions: RW End-of-lines: CRLF Encoding: UTF-8 Line: 43 Column: 3 Memory: 60 % CPU: 4 %



4. 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，即我没有进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。