机器学习第二次作业

姓名：刘文晨

学号：202222280328

1. 下列数据时水泥释放的热量与其成分的关系：求其线性依赖关系

y	x1	x2	x3	x4
78.5	7	26	6	60
74.3	1	29	15	52
104.3	11	56	8	20
87.6	11	31	8	47
95.9	7	52	6	33
109.2	11	55	9	22
102.7	3	71	17	6
72.5	1	31	22	44
93.1	2	54	18	22
115.9	21	47	4	26
83.8	1	40	23	34
113.3	11	66	9	12
109.4	10	68	8	12

将上述数据用excel保存，命名为data.xlsx。

编写Python代码：

```python
import pandas as pd
import statsmodels.api as sm

data = pd.read_excel('data.xlsx')
data.columns = ['y', 'x1', 'x2', 'x3', 'x4']
# 生成自变量
x = sm.add_constant(data.iloc[:, 1:])
# 生成因变量
y = data['y']
# 生成模型
model = sm.OLS(y, x)
# 模型拟合
result = model.fit()
```

```python
# 模型描述
print(result.summary())
```

运行结果:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.982
Model:                            OLS   Adj. R-squared:                  0.974
Method:                 Least Squares   F-statistic:                     111.5
Date:                Tue, 11 Oct 2022   Prob (F-statistic):           4.76e-07
Time:                        21:24:03   Log-Likelihood:                -26.918
No. Observations:                  13   AIC:                             63.84
Df Residuals:                       8   BIC:                             66.66
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         62.4054     70.071      0.891      0.399     -99.179     223.989
x1             1.5511      0.745      2.083      0.071      -0.166       3.269
x2             0.5102      0.724      0.705      0.501      -1.159       2.179
x3             0.1019      0.755      0.135      0.896      -1.638       1.842
x4            -0.1441      0.709     -0.203      0.844      -1.779       1.491
==============================================================================
Omnibus:                        0.165   Durbin-Watson:                   2.053
Prob(Omnibus):                  0.921   Jarque-Bera (JB):                0.320
Skew:                           0.201   Prob(JB):                        0.852
Kurtosis:                       2.345   Cond. No.                     6.06e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 6.06e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

从上述结果描述，我们得到回归模型：

$$y = 62.4054 + 1.5511x_1 + 0.5102x_2 + 0.1019x_3 - 0.1441x_4$$

从结果中还可以看出，Prob (F-statistic)为4.76e-07，其接近于零，说明我们的多元线性方程是显著的，也就是y与x$_1$、x$_2$、x$_3$、x$_4$有着显著的线性关系，而R-squared是0.982，也说明这个线性关系比较显著。

2. 经研究发现，学生用于购买书籍及课外读物的支出与本人受教育年限和其家庭收入水平有关，对18名学生进行调查的统计资料如下表所示，求其回归模型

y	x1	x2
450.5	4	171.2

y	x1	x2
507.7	4	174.2
613.9	5	204.3
563.4	4	218.7
501.5	4	219.4
781.5	7	240.4
541.8	4	273.5
611.1	5	294.8
1222.1	10	330.2
793.2	7	333.1
660.8	5	366
792.7	6	350.9
580.8	4	357.9
612.7	5	359
890.8	7	371.9
1121	9	435.3
1094.2	8	523.9
1253	10	604.1

将上述数据用excel保存，命名为data.xlsx。

编写Python代码：

```python
import pandas as pd
import statsmodels.api as sm

data = pd.read_excel('data.xlsx')
data.columns = ['y', 'x1', 'x2']
# 生成自变量
x = sm.add_constant(data.iloc[:, 1:])
# 生成因变量
y = data['y']
# 生成模型
model = sm.OLS(y, x)
# 模型拟合
result = model.fit()
# 模型描述
print(result.summary())
```

运行结果：

```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                      y   R-squared:                       0.980
Model:                            OLS   Adj. R-squared:                  0.977
Method:                 Least Squares   F-statistic:                     362.4
Date:                Tue, 11 Oct 2022   Prob (F-statistic):           2.00e-13
Time:                        22:22:18   Log-Likelihood:                -89.942
No. Observations:                  18   AIC:                             185.9
Df Residuals:                      15   BIC:                             188.6
Df Model:                           2                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.9756     30.322     -0.032      0.975     -65.606      63.655
x1           104.3146      6.409     16.276      0.000      90.654     117.975
x2             0.4022      0.116      3.457      0.004       0.154       0.650
==============================================================================
Omnibus:                        0.776   Durbin-Watson:                   2.561
Prob(Omnibus):                  0.678   Jarque-Bera (JB):                0.728
Skew:                          -0.230   Prob(JB):                        0.695
Kurtosis:                       2.128   Cond. No.                     1.13e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.13e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

从上述结果描述，我们得到回归模型：$y=-0.9756+104.3146x_1+0.4022x_2$

从结果中还可以看出，Prob (F-statistic)为2.00e-13，其接近于零，说明我们的多元线性方程是显著的，也就是y与x_1、x_2有着显著的线性关系，而R-squared是0.980，也说明这个线性关系比较显著。