

Multicollinearity



多元共線性

C11 單親家庭對學生數學表現之影響



- C11** Use the data in MEAPSINGLE to study the effects of single-parent households on student math performance. These data are for a subset of schools in southeast Michigan for the year 2000. The socioeconomic variables are obtained at the ZIP code level (where ZIP code is assigned to schools based on their mailing addresses).
- (i) Run the simple regression of *math4* on *pctsgle* and report the results in the usual format. Interpret the slope coefficient. Does the effect of single parenthood seem large or small?
 - (ii) Add the variables *lmedinc* and *free* to the equation. What happens to the coefficient on *pctsgle*? Explain what is happening.
 - (iii) Find the sample correlation between *lmedinc* and *free*. Does it have the sign you expect?
 - (vi) Does the substantial correlation between *lmedinc* and *free* mean that you should drop one from the regression to better estimate the causal effect of single parenthood on student performance? Explain.
 - (v) Find the variance inflation factors (VIFs) for each of the explanatory variables appearing in the regression in part (ii). Which variable has the largest VIF? Does this knowledge affect the model you would use to study the causal effect of single parenthood on math performance?

讀入資料



#讀入meapsingle資料

```
import pandas as pd
```

```
import numpy as np
```

```
meapsingle= pd.read_csv("meapsingle.csv")
```

```
meapsingle.head()
```

	dcode	bcode	math4	read4	enroll	exppp	free	reduced	lunch	medinc	totchild	married	singl
0	63010	3030	92.8	82.5	607	6619.54	1.0	0.7	1.7	110322	4076	3542	53
1	63010	3133	100.0	94.3	370	6619.54	0.0	0.0	0.0	110322	4076	3542	53
2	63270	2023	72.1	46.5	220	5607.56	5.9	5.0	10.9	65119	2524	2091	43
3	63270	2978	76.1	65.7	356	5829.53	8.1	2.8	10.9	65119	2524	2091	43
4	63010	316	95.2	80.6	329	6619.54	0.3	0.3	0.6	109313	3486	3241	24

#呼叫DataFrame內資料

```
math4=pd.concat([meapsingle.math4])
```

```
pctsgle=pd.concat([meapsingle.pctsgle])
```

```
free=pd.concat([meapsingle.free])
```

```
lmedinc=pd.concat([meapsingle.lmedinc])
```

C11(1)執行math4對pctsgle之簡單迴歸

- **pctsgle**: percent of children not in married-couple families

```
import statsmodels.api as sm
# 迴歸分析 應變數是math4 自變數是pctsgle
model=sm.OLS(math4,sm.add_constant(pctsgle)).fit()
print(model.summary())
```

```
OLS Regression Results

=====
Dep. Variable:          math4      R-squared:                0.380
Model:                  OLS        Adj. R-squared:           0.377
Method:                 Least Squares  F-statistic:             138.9
Date:                  Fri, 16 Apr 2021  Prob (F-statistic):      2.54e-25
Time:                  02:25:34      Log-Likelihood:          -901.95
No. Observations:      229          AIC:                    1808.
Df Residuals:          227          BIC:                    1815.
Df Model:               1
Covariance Type:       nonrobust

=====
                    coef    std err          t      P>|t|      [0.025     0.975]
-----
const              96.7704      1.597       60.603     0.000     93.624     99.917
pctsgle            -0.8329      0.071      -11.784     0.000    -0.972    -0.694
=====

Omnibus:                 8.632    Durbin-Watson:           1.621
Prob(Omnibus):            0.013    Jarque-Bera (JB):        13.624
Skew:                    -0.190    Prob(JB):                 0.00110
Kurtosis:                 4.133    Cond. No.                 43.8
=====
```

Warnings:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

11(2)方程式加入lmedinc和free



- **lmedinc:** $\log(\text{medinc})$
- **free:** percent eligible, free lunch
- **medinc:** zipcode median family, \$ (1999)

```
import statsmodels.api as sm
# 迴歸分析 應變數是math4 自變數是pctsgle free lmedinc
pairf=pd.concat([meapsingle.pctsgle,meapsingle.free,meapsingle.lmedinc],axis = 1)
model_2=sm.OLS(math4,sm.add_constant(pairf)).fit()
print(model_2.summary())
```

11(2)方程式加入lmedinc和free



```
=====
                        OLS Regression Results
=====
Dep. Variable:          math4      R-squared:                0.460
Model:                  OLS        Adj. R-squared:            0.453
Method:                 Least Squares  F-statistic:              63.85
Date:                   Fri, 16 Apr 2021  Prob (F-statistic):      6.63e-30
Time:                   02:45:39      Log-Likelihood:           -886.08
No. Observations:       229          AIC:                     1780.
Df Residuals:           225          BIC:                     1794.
Df Model:                3
Covariance Type:        nonrobust
=====
                        coef      std err          t      P>|t|      [0.025      0.975]
-----
const                51.7232      58.478         0.884     0.377     -63.512     166.958
pctsgle              -0.1996      0.159        -1.258     0.210     -0.512      0.113
free                 -0.3964      0.070        -5.635     0.000     -0.535     -0.258
lmedinc               3.5601      5.042         0.706     0.481     -6.375     13.495
=====
Omnibus:               8.076      Durbin-Watson:           1.533
Prob(Omnibus):          0.018      Jarque-Bera (JB):        13.783
Skew:                   0.111      Prob(JB):                 0.00102
Kurtosis:                4.181      Cond. No.                 2.64e+03
=====

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.64e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Pctsgle係數減少

11(3)找出lmedinc和free樣本相關係數



- **Data.corr()** 代表pearson 相關係數

```
#算lmedinc與free相關係數
```

```
data=pd.concat([meapsingle.free,meapsingle.lmedinc],axis = 1)  
data.corr()
```

	free	lmedinc
free	1.00000	-0.74697
lmedinc	-0.74697	1.00000

11(4) *lmedinc* and *free* 相關係數大，是否需要刪除變數



(iv) No, because high correlations among the variables *lmedinc* and *free* do not make it more difficult to determine the causal effect of single parenthood on student performance.

11(5)找出Imedinc、free、pctsgle的VIF



```
#VIF_pctsgle  
VIF_pctsgle=1/(1-0.380)  
VIF_pctsgle  
  
1.6129032258064517
```

```
import statsmodels.api as sm  
# 迴歸分析 應變數是math4 自變數是pctsgle  
model=sm.OLS(math4,sm.add_constant(pctsgle)).fit()  
print(model.summary())
```

OLS Regression Results

Dep. Variable:	math4	R-squared:	0.380
Model:	OLS	Adj. R-squared:	0.377

11(5)找出Imedinc、free、pctsgle的VIF

```
#VIF_free  
VIF_free=1/(1-0.446)  
VIF_free
```

```
1.8050541516245486
```

```
import statsmodels.api as sm  
# 迴歸分析 應變數是math4 自變數是free  
model=sm.OLS(math4,sm.add_constant(free)).fit()  
print(model.summary())
```

OLS Regression Results

```
=====
```

Dep. Variable:	math4	R-squared:	0.446
Model:	OLS	Adj. R-squared:	0.443

```
=====
```

11(5)找出lmedinc、free、pctsgle的VIF



```
#VIF_lmedinc  
VIF_lmedinc=1/(1-0.321)  
VIF_lmedinc
```

```
1.4727540500736376
```

```
import statsmodels.api as sm  
# 迴歸分析 應變數是math4 自變數是lmedinc  
model=sm.OLS(math4,sm.add_constant(lmedinc)).fit()  
print(model.summary())
```

OLS Regression Results

```
=====
```

Dep. Variable:	math4	R-squared:	0.321
Model:	OLS	Adj. R-squared:	0.318

```
=====
```

11(5)找出Imedinc、free、pctsgle的VIF



$$(v) \text{VIF}_{pctsgle} = \frac{1}{1-R^2} = \frac{1}{1-0.3795} = 1.6116.$$

$$\text{VIF}_{free} = \frac{1}{1-R^2} = \frac{1}{1-0.4455} = 1.8034.$$

$$\text{VIF}_{imedinc} = \frac{1}{1-R^2} = \frac{1}{1-0.3212} = 1.4732.$$

By comparing the three variables, it is very clear that the variable *free* has the highest VIF. No, this knowledge does not affect the model to study the causal effect of single parenthood on math performance.