Omitted Variable Bias

遺漏變數的偏誤

探討工資議題

- 應變數:工資
- 自變數:教育年數(educ)、勞動市場經驗年數(exper)、在目前工作的年數(tenure)

- $\log(wage) = \beta_0 + \beta_1 educ + u$
- $\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + u$
- $\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u$

讀入工資資料

```
#讀入Wage1資料
import pandas as pd
import numpy as np
wage1= pd.read_csv("wage1.csv")
wage1.head()
```

	wage	educ	exper	tenure	nonwhite	female	married	numdep	smsa	northcen	...	trcommpu	tr
0	3.10	11	2	0	0	1	0	2	1	0	...	0	
1	3.24	12	22	2	0	1	1	3	1	0	...	0	
2	3.00	11	2	0	0	0	0	2	0	0	...	0	
3	6.00	8	44	28	0	0	1	0	1	0	...	0	
4	5.30	12	7	2	0	0	1	1	0	0	...	0	

5 rows × 24 columns

讀入工資資料

```
#呼叫DataFrame內的wage、educ、exper、tenure
wage=pd.concat([wage1.wage])
educ=pd.concat([wage1.educ])
exper=pd.concat([wage1.exper])
tenure=pd.concat([wage1.tenure])
log_wage=np.log(wage)
```

$\log(wage) = \beta_0 + \beta_1 educ + u$

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

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   wage   R-squared:                       0.186
Model:                            OLS   Adj. R-squared:                  0.184
Method:                 Least Squares   F-statistic:                     119.6
Date:                Fri, 16 Apr 2021   Prob (F-statistic):           3.27e-25
Time:                        00:59:38   Log-Likelihood:                -359.38
No. Observations:                 526   AIC:                             722.8
Df Residuals:                     524   BIC:                             731.3
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.5838      0.097      5.998      0.000       0.393       0.775
educ           0.0827      0.008     10.935      0.000       0.068       0.098
==============================================================================
Omnibus:                       11.804   Durbin-Watson:                   1.801
Prob(Omnibus):                  0.003   Jarque-Bera (JB):               13.811
Skew:                           0.268   Prob(JB):                      0.00100
Kurtosis:                       3.586   Cond. No.                         60.2
==============================================================================
```

$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + u$

```
import statsmodels.api as sm
# 迴歸分析 應變數是wage 自變數是educ exper
pairf=pd.concat([wage1.educ,wage1.exper],axis = 1)
model_1=sm.OLS(log_wage,sm.add_constant(pairf)).fit()
print(model_1.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   wage   R-squared:                       0.249
Model:                            OLS   Adj. R-squared:                  0.246
Method:                 Least Squares   F-statistic:                     86.86
Date:                Thu, 15 Apr 2021   Prob (F-statistic):           2.68e-33
Time:                        22:38:15   Log-Likelihood:                -338.01
No. Observations:                 526   AIC:                             682.0
Df Residuals:                     523   BIC:                             694.8
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.2169      0.109      1.997      0.046       0.004       0.430
educ           0.0979      0.008     12.848      0.000       0.083       0.113
exper          0.0103      0.002      6.653      0.000       0.007       0.013
==============================================================================
Omnibus:                        7.740   Durbin-Watson:                   1.789
Prob(Omnibus):                  0.021   Jarque-Bera (JB):                9.485
Skew:                           0.165   Prob(JB):                      0.00872
Kurtosis:                       3.569   Cond. No.                         130.
==============================================================================
```

$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u$

```
import statsmodels.api as sm
# 迴歸分析 應變數是wage 自變數是educ exper tenure
pairf=pd.concat([wage1.educ,wage1.exper,wage1.tenure],axis = 1)
model_2=sm.OLS(log_wage,sm.add_constant(pairf)).fit()
print(model_2.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   wage   R-squared:                       0.316
Model:                            OLS   Adj. R-squared:                  0.312
Method:                 Least Squares   F-statistic:                     80.39
Date:                Thu, 15 Apr 2021   Prob (F-statistic):           9.13e-43
Time:                        22:38:26   Log-Likelihood:                -313.55
No. Observations:                 526   AIC:                             635.1
Df Residuals:                     522   BIC:                             652.2
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.2844      0.104      2.729      0.007      0.080      0.489
educ           0.0920      0.007     12.555      0.000      0.078      0.106
exper          0.0041      0.002      2.391      0.017      0.001      0.008
tenure         0.0221      0.003      7.133      0.000      0.016      0.028
==============================================================================
Omnibus:                       11.534   Durbin-Watson:                   1.769
Prob(Omnibus):                  0.003   Jarque-Bera (JB):               20.941
Skew:                           0.021   Prob(JB):                     2.84e-05
Kurtosis:                       3.977   Cond. No.                         135.
==============================================================================
```

探討工資議題

- 前提:滿足迴歸五大基本假設
- $\log(wage) = 0.5838+0.0827educ + u$
  - R-squared=0.186
  - Prob(F-statistic)3.27e-25
- $\log(wage) = 0.2169+0.0979educ + 0.0103exper + u$
  - R-squared=0.249
  - Prob(F-statistic)2.68e-33
- $\log(wage) = 0.2844+0.0920educ + 0.0041exper + 0.0221tenure + u$
  - R-squared=0.316
  - Prob(F-statistic)9.13e-43

探討工資議題

```
import statistics
mean_wage=statistics.mean(log_wage)
mean_educ=statistics.mean(educ)
mean_exper=statistics.mean(exper)
mean_tenure=statistics.mean(tenure)
print("log_wage平均數",round(mean_wage,6),"educ平均數",round(mean_educ,2),"exper平均數",round(mean_exper,2),
      "tenure平均數",round(mean_tenure,2)) #算到小數點第二位
#將上述數值代入工資迴歸方程式
```

```
log_wage平均數 1.623268 educ平均數 12.56 exper平均數 17.02 tenure平均數 5.1
```

```
regression_wage1=0.5838 +0.0827*mean_educ
regression_wage2=0.2169 +0.0979*mean_educ+0.0103*mean_exper
regression_wage3=0.2844 +0.0920*mean_educ+0.0041*mean_exper+ 0.0221*mean_tenure
print("regression_wage1",regression_wage1,"regression_wage2",regression_wage2,"regression_wage3",regression_wage3)
```

```
regression_wage1 1.6227384030418248 regression_wage2 1.6220682509505704 regression_wage3 1.6227528517110266
```

C6

C6 Use the data set in WAGE2 for this problem. As usual, be sure all of the following regressions contain an intercept.

(i) Run a simple regression of *IQ* on *educ* to obtain the slope coefficient, say, $\tilde{\delta}_1$.

(ii) Run the simple regression of log(*wage*) on *educ*, and obtain the slope coefficient, $\tilde{\beta}_1$.

(iii) Run the multiple regression of log(*wage*) on *educ* and *IQ*, and obtain the slope coefficients, $\hat{\beta}_1$ and $\hat{\beta}_2$, respectively.

(iv) Verify that $\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2\tilde{\delta}_1$.

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$ (正確)

$y = \beta_0 + \beta_1 x_1 + u'$ (少了 x_2)

此時估計結果 $\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2\tilde{\delta}_1$ → bias

原因：令 $x_2 = \delta_0 + \delta_1 x_1 + v$

$y = \beta_0 + \beta_1 x_1 + \beta_2(\delta_0 + \delta_1 x_1 + v) + u$

$= \square + (\beta_1 + \beta_2\delta_1)x_1 + \square$

x_2 被省略，它的效果會跑到 x_1，造成 x_1 係數偏誤

C6

```
#讀入Wage2資料
import pandas as pd
import numpy as np
wage2= pd.read_csv("wage2.csv")
wage2.head()
```

	wage	hours	IQ	KWW	educ	exper	tenure	age	married	black	south	urban	sibs	brthord	meduc	feduc	lwage
0	769	40	93	35	12	11	2	31	1	0	0	1	1	2.0	8.0	8.0	6.645091
1	808	50	119	41	18	11	16	37	1	0	0	1	1	NaN	14.0	14.0	6.694562
2	825	40	108	46	14	11	9	33	1	0	0	1	1	2.0	14.0	14.0	6.715384
3	650	40	96	32	12	13	7	32	1	0	0	1	4	3.0	12.0	12.0	6.476973
4	562	40	74	27	11	14	5	34	1	0	0	1	10	6.0	6.0	11.0	6.331502

```
#呼叫DataFrame內的資料
IQ=pd.concat([wage2.IQ])
wage=pd.concat([wage2.wage])
educ=pd.concat([wage2.educ])
log_wage=np.log(wage)
```

C6(1)

```
#跑IQ對educ迴歸，求斜率係數
import statsmodels.api as sm
model=sm.OLS(IQ,sm.add_constant(educ)).fit()
print(model.summary())
```

OLS Regression Results

Dep. Variable:	IQ	R-squared:	0.266
Model:	OLS	Adj. R-squared:	0.265
Method:	Least Squares	F-statistic:	338.0
Date:	Fri, 16 Apr 2021	Prob (F-statistic):	1.16e-64
Time:	01:35:45	Log-Likelihood:	-3717.0
No. Observations:	935	AIC:	7438.
Df Residuals:	933	BIC:	7448.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	53.6872	2.623	20.468	0.000	48.540	58.835
educ	3.5338	0.192	18.385	0.000	3.157	3.911

Omnibus:	30.954	Durbin-Watson:	1.779
Prob(Omnibus):	0.000	Jarque-Bera (JB):	35.266
Skew:	-0.398	Prob(JB):	2.20e-08
Kurtosis:	3.522	Cond. No.	85.3

C3.6 (i) The slope coefficient from the regression *IQ* on *educ* is (rounded to five decimal places) $\tilde{\delta}_1 = 3.53383$.

C6(2)

```
#跑log_wage對educ迴歸，求斜率係數
import statsmodels.api as sm
model=sm.OLS(log_wage,sm.add_constant(educ)).fit()
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   wage   R-squared:                       0.097
Model:                            OLS   Adj. R-squared:                  0.096
Method:                 Least Squares   F-statistic:                     100.7
Date:                Fri, 16 Apr 2021   Prob (F-statistic):           1.42e-22
Time:                        01:36:37   Log-Likelihood:                -469.72
No. Observations:                 935   AIC:                             943.4
Df Residuals:                     933   BIC:                             953.1
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          5.9731      0.081     73.403      0.000       5.813       6.133
educ           0.0598      0.006     10.035      0.000       0.048       0.072
==============================================================================
Omnibus:                       31.006   Durbin-Watson:                   1.779
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               37.262
Skew:                          -0.375   Prob(JB):                     8.10e-09
Kurtosis:                       3.627   Cond. No.                         85.3
==============================================================================
```

(ii) The slope coefficient from log(*wage*) on *educ* is $\tilde{\beta}_1 = .05984$.

C6(3)

```
import statsmodels.api as sm
pairf=pd.concat([wage2.educ,wage2.IQ],axis = 1)
model=sm.OLS(log_wage,sm.add_constant(pairf)).fit()
print(model.summary())
```

OLS Regression Results

Dep. Variable:	wage	R-squared:	0.130
Model:	OLS	Adj. R-squared:	0.128
Method:	Least Squares	F-statistic:	69.42
Date:	Fri, 16 Apr 2021	Prob (F-statistic):	7.88e-29
Time:	01:48:15	Log-Likelihood:	-452.72
No. Observations:	935	AIC:	911.4
Df Residuals:	932	BIC:	926.0
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.6583	0.096	58.793	0.000	5.469	5.847
educ	0.0391	0.007	5.721	0.000	0.026	0.053
IQ	0.0059	0.001	5.875	0.000	0.004	0.008

Omnibus:	35.757	Durbin-Watson:	1.810
Prob(Omnibus):	0.000	Jarque-Bera (JB):	47.770

(iii) The slope coefficients from log(*wage*) on *educ* and *IQ* are $\hat{\beta}_1 = .03912$ and $\hat{\beta}_2 = .00586$, respectively.

C6(4)

C3.6 (i) The slope coefficient from the regression *IQ* on *educ* is (rounded to five decimal places) $\tilde{\delta}_1 = 3.53383$.

(ii) The slope coefficient from log(*wage*) on *educ* is $\tilde{\beta}_1 = .05984$.

(iii) The slope coefficients from log(*wage*) on *educ* and *IQ* are $\hat{\beta}_1 = .03912$ and $\hat{\beta}_2 = .00586$, respectively.

(iv) We have $\hat{\beta}_1 + \tilde{\delta}_1\hat{\beta}_2 = .03912 + 3.53383(.00586) \approx .05983$, which is very close to .05984; the small difference is due to rounding error.