Untitled2

November 2, 2025

[9]:
```
import pandas as pd
import seaborn as sns
import numpy as np
import statsmodels.api as sm
import matplotlib.pyplot as plt
from statsmodels.stats.outliers_influence import variance_inflation_factor
from scipy import stats
```

[10]:
```
print("""Answer to question #1:""")
data = {
    'Dependent': [35, 50, 65, 70, 80],
    'education': [12, 16, 18, 20, 21],
    'experience': [5, 10, 12, 15, 18],
    'Age': [25, 30, 32, 35, 40]
}

df = pd.DataFrame(data)
print(df)

Y=df["Dependent"]
X=df[["education", "experience", "Age"]]
X=sm.add_constant(X)
model = sm.OLS(Y, X).fit()
print(model.summary())
print("""-------------------------------------------------------------------------------------
print("""-------------------------------------------------------------------------------------
print(f"\nInterpretation:")
print("""In the abstract sense, the coefficient of x1 that is, the coefficient␣
  ↪on education indicates that for every additional unit of education, the␣
  ↪dependent variable y increases by approximately 15.8333 on avg""")
```

```
Answer to question #1:
   Dependent  education  experience  Age
0         35         12           5   25
1         50         16          10   30
2         65         18          12   32
3         70         20          15   35
4         80         21          18   40
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              Dependent   R-squared:                       0.997
Model:                            OLS   Adj. R-squared:                  0.987
Method:                 Least Squares   F-statistic:                     99.67
Date:                Sun, 02 Nov 2025   Prob (F-statistic):             0.0735
Time:                        20:10:55   Log-Likelihood:                -6.6389
No. Observations:                   5   AIC:                             21.28
Df Residuals:                       1   BIC:                             19.72
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       -335.0000    147.549     -2.270      0.264   -2209.793    1539.793
education     15.8333      6.067      2.610      0.233     -61.252      92.919
experience   -20.4167      9.887     -2.065      0.287    -146.037     105.203
Age           11.2500      5.052      2.227      0.269     -52.939      75.439
==============================================================================
Omnibus:                          nan   Durbin-Watson:                   2.500
Prob(Omnibus):                    nan   Jarque-Bera (JB):                0.747
Skew:                          -0.913   Prob(JB):                        0.688
Kurtosis:                       2.500   Cond. No.                     6.38e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 6.38e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
--------------------------------------------------------------------------------
-----------------
--------------------------------------------------------------------------------
-----------------

Interpretation:
In the abstract sense, the coefficient of x1 that is, the coefficient on
education indicates that for every additional unit of education, the dependent
variable y increases by approximately 15.8333 on avg

C:\Users\default.DESKTOP-GGCF6CQ\anaconda3\Lib\site-
packages\statsmodels\stats\stattools.py:74: ValueWarning: omni_normtest is not
valid with less than 8 observations; 5 samples were given.
  warn("omni_normtest is not valid with less than 8 observations; %i "
```

[11]:
```
print("""Answer to question #2:""")
data = {
    'Dependent': [35, 50, 65, 70, 80],
```

```
    'education': [12, 16, 18, 20, 21],
    'experience': [5, 10, 12, 15, 18],
}

df = pd.DataFrame(data)
print(df)

Y=df["Dependent"]
X=df[["education", "experience"]]
X=sm.add_constant(X)
model = sm.OLS(Y, X).fit()
print(model.summary())
print("""-----------------------------------------------------------------------------------------------------------
print("""-----------------------------------------------------------------------------------------------------------
print(f"\nInterpretation:")
print("""When age is removed wwe get a coefficient of 2.976 on education when␣
 ↪it was previously 15.83, this leads to believe that age and education were␣
 ↪are positively correlated, and age was previously accounting for some of the␣
 ↪variation in the dependent variable that education now partially absorbs.␣
 ↪This illustrates the omitted variable bias, by showing that when you leave␣
 ↪relevant vraiables such as age it can change the effects of other variables␣
 ↪in your model such as with age here.""")
```

```
Answer to question #2:
   Dependent  education  experience
0         35         12           5
1         50         16          10
2         65         18          12
3         70         20          15
4         80         21          18
                            OLS Regression Results
==============================================================================
Dep. Variable:              Dependent   R-squared:                       0.980
Model:                            OLS   Adj. R-squared:                  0.960
Method:                 Least Squares   F-statistic:                     49.34
Date:                Sun, 02 Nov 2025   Prob (F-statistic):             0.0199
Time:                        20:10:55   Log-Likelihood:                -11.101
No. Observations:                   5   AIC:                             28.20
Df Residuals:                       2   BIC:                             27.03
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -8.5204     28.760     -0.296      0.795    -132.263     115.222
education      2.9762      3.216      0.925      0.452     -10.863      16.816
experience     1.3946      2.325      0.600      0.610      -8.609      11.398
==============================================================================
```

```
Omnibus:                          nan   Durbin-Watson:                   3.570
Prob(Omnibus):                    nan   Jarque-Bera (JB):                0.362
Skew:                          -0.013   Prob(JB):                        0.835
Kurtosis:                       1.682   Cond. No.                         401.
==============================================================================

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

Interpretation:
When age is removed wwe get a coefficient of 2.976 on education when it was
previously 15.83, this leads to believe that age and education were are
positively correlated, and age was previously accounting for some of the
variation in the dependent variable that education now partially absorbs. This
illustrates the omitted variable bias, by showing that when you leave relevant
vraiables such as age it can change the effects of other variables in your model
such as with age here.
```

```
C:\Users\default.DESKTOP-GGCF6CQ\anaconda3\Lib\site-
packages\statsmodels\stats\stattools.py:74: ValueWarning: omni_normtest is not
valid with less than 8 observations; 5 samples were given.
  warn("omni_normtest is not valid with less than 8 observations; %i "
```

[12]:

```python
print("""Answer to question #3:""")
data = {
    'Dependent': [35, 50, 65, 70, 80],
    'education': [12, 16, 18, 20, 21],
    'experience': [5, 10, 12, 15, 18],
    'Age': [25, 30, 32, 35, 40]
}

df = pd.DataFrame(data)
print(df)
Y=df["Dependent"]
X=df[["education", "experience", "Age"]]
X=sm.add_constant(X)
model = sm.OLS(Y, X).fit()
print("""---------------------------------------------------------------------------------
print("""---------------------------------------------------------------------------------

corr_matrix = df[['education', 'experience', 'Age']].corr()
print(corr_matrix)
```

```
print("""-------------------------------------------------------------------------------------
print("""-------------------------------------------------------------------------------------
X = sm.add_constant(df[['education', 'experience', 'Age']])
vif = pd.DataFrame()
vif["Variable"] = X.columns
vif["VIF"] = [variance_inflation_factor(X.values, i)
              for i in range(X.shape[1])]
print(vif)
print("""-------------------------------------------------------------------------------------
print("""-------------------------------------------------------------------------------------
print(f"\nInterpretation:")
print("""The correlation matrix and VIF values reveal severe multicollinearity among Education, Experience, and Age (all VIFs far exceed 10). This means these variables contain highly overlapping information.
As a result, individual coefficient estimates are imprecise their standard errors are inflated, and their magnitudes can vary widely when one regressor is removed. This explains why the coefficient on Education changed drastically when Age was excluded. This can also be seen from the correlation matrix where each variable closely follows the other.""")
print("""-------------------------------------------------------------------------------------
print("""-------------------------------------------------------------------------------------
print("""Answer to question #4:""")
t_value = 2.5 / 0.8
p = 2 * (1-stats.t.cdf(abs(t_value), df=1))
alpha = 0.05
print(f"t_value = {t_value:.3f}, p = {p :.5f}")

if p < alpha:
    print("Reject H0: Education significantly affects y")
else:
    print("Fail to Reject H0: We dont have enough information to say that Education significantly affects y")
print("""-------------------------------------------------------------------------------------
print("""-------------------------------------------------------------------------------------
print("""Answer to question #5:""")
b = 2.5
se = 0.8
df = 1
alpha = 0.05
t_crit = stats.t.ppf(1 - alpha/2, df)
lower = b - t_crit * se
upper = b + t_crit * se


print(f"95% Confidence Interval: ({lower:.3f}, {upper:.3f})")
print(f"\nInterpretation:")
```

```python
print(f"Holding Experience and Age constant, we are 95% confident the true␣
↪effect of one more unit of Education lies between: ({lower:.3f}, {upper:.
↪3f}) Because 0 is inside the interval, Education is not statistically␣
↪significant at 5%.")
```

```
Answer to question #3:
   Dependent  education  experience  Age
0         35         12           5   25
1         50         16          10   30
2         65         18          12   32
3         70         20          15   35
4         80         21          18   40
--------------------------------------------------------------------------------
-----------------
--------------------------------------------------------------------------------
-----------------
            education  experience       Age
education    1.000000    0.988212  0.964229
experience   0.988212    1.000000  0.993065
Age          0.964229    0.993065  1.000000
--------------------------------------------------------------------------------
-----------------
--------------------------------------------------------------------------------
-----------------
     Variable           VIF
0       const  26125.000000
1   education    452.266667
2  experience   2298.916667
3         Age    766.850000
--------------------------------------------------------------------------------
-----------------
--------------------------------------------------------------------------------
-----------------

Interpretation:
The correlation matrix and VIF values reveal severe multicollinearity among
Education, Experience, and Age (all VIFs far exceed 10). This means these
variables contain highly overlapping information.
As a result, individual coefficient estimates are imprecise their standard
errors are inflated, and their magnitudes can vary widely when one regressor is
removed. This explains why the coefficient on Education changed drastically when
Age was excluded. This can also be seen from the correlation matrix where each
variable closely follows the other.
--------------------------------------------------------------------------------
-----------------
--------------------------------------------------------------------------------
-----------------
Answer to question #4:
```

```
t_value = 3.125, p = 0.19716
Fail to Reject H0: We dont have enough information to say that Education
significantly affects y
--------------------------------------------------------------------------------
-----------------
--------------------------------------------------------------------------------
-----------------
Answer to question #5:
95% Confidence Interval: (-7.665, 12.665)

Interpretation:
Holding Experience and Age constant, we are 95% confident the true effect of one
more unit of Education lies between: (-7.665, 12.665) Because 0 is inside the
interval, Education is not statistically significant at 5%.
```