MBAN 5110: PREDICTIVE MODELING

SESSION 5: DEALING WITH ENDOGENEITY PROBLEMS

DR. ISIK BICER

Schulich School of Business | YORK UNIVERSITÉ UNIVERSITY

TODAY'S AGENDA

- Reasons of endogeneity
- Two-Stage Least Squares
- Generalized Method of Moments

ENDOGENEITY PROBLEMS

- The endogeneity problems occur when X and E are dependent:

$$X^T\mathrm{E} = \delta$$

$$X^T(Y - X\mathrm{B}) = \delta$$

$$X^TY - X^TXB = \delta$$

$$X^TY - \delta = X^TXB$$

$$(X^TX)^{-1}(X^TY - \delta) = (X^TX)^{-1}X^TXB$$

$$\mathrm{B} = (X^TX)^{-1}X^TY - (X^TX)^{-1}\delta$$

- The OLS estimation is then biased by $(X^TX)^{-1}\delta$

OMITTED VARIABLE BIAS

- Omitted variables also cause endogeneity problems
- True model: $Y = X\mathrm{B} + \theta Z + \mathrm{E}$
- θ is constant. It is the coefficient of Z
- The omitted variable: $Z = \begin{vmatrix} z_1 \\ z_2 \\ \vdots \\ z_m \end{vmatrix}$

OMITTED VARIABLE BIAS

$$X^T\mathrm{E} = 0$$
$$X^T(Y - X\mathrm{B} - \theta Z) = 0$$
$$X^TY - X^TXB - X^T\theta Z = 0$$
$$X^TY - X^T\theta Z = X^TXB$$
$$(X^TX)^{-1}(X^TY - X^T\theta Z) = (X^TX)^{-1}X^TXB$$
$$\mathrm{B} = (X^TX)^{-1}X^TY - (X^TX)^{-1}X^T\theta Z$$

- The OLS estimation is biased by $(X^TX)^{-1}X^T\theta Z$

MEASUREMENT ERROR

- True model: $Y = X\mathrm{B} + \mathrm{E}$
- Measured X values: $\hat{X} = X + \varphi$
- Estimate model: $Y = \hat{X}\mathrm{B} + \mathrm{E}$

$$X^T\mathrm{E} = 0$$

$$(\hat{X} - \varphi)^T\mathrm{E} = 0$$

$$\hat{X}^T\mathrm{E} - \varphi^T\mathrm{E} = 0$$

$$\hat{X}^T Y - \varphi^T\mathrm{E} = \hat{X}^T\hat{X}B$$

$$\left(\hat{X}^T\hat{X}\right)^{-1}(\hat{X}^T Y - \varphi^T\mathrm{E}) = B$$

$$\mathrm{B} = \left(\hat{X}^T\hat{X}\right)^{-1}\hat{X}^T Y - \left(\hat{X}^T\hat{X}\right)^{-1}\varphi^T\mathrm{E}$$

- The OLS estimation is biased by $\left(\hat{X}^T\hat{X}\right)^{-1}\varphi^T\mathrm{E}$

CAREFUL UNDERSTANDING OF THE BIAS TERMS

- When error is dependent on explanatory variable: $(X^T X)^{-1}\delta$
- When there is an omitted variable: $(X^T X)^{-1} X^T \theta Z$
  - No endogeneity problem when $X^T Z = 0$ (i.e., when explanatory variable is independent from omitted variable)
- When there is a measurement error: $\left(\hat{X}^T \hat{X}\right)^{-1} \varphi^T \mathrm{E}$
  - No endogeneity problem when $\varphi^T \mathrm{E}$ (i.e., when the measurement error is independent from the model error)
- ***If we can marginalize X such that the part of X that depend on omitted variable and error term would be extracted, the remaining part can be used in the least square estimation***

TWO STAGE LEAST SQUARES (2SLS)

- When there is an endogeneity problem, the modeler must first look for instrumental variables
- Good instruments
  - High correlation with X
  - Low correlation with Y
    - Low correlation with the dependent variable guarantees that the instrument(s) does not have a correlation with omitted variables and error terms

TWO STAGE LEAST SQUARES

- The model: $Y = X\mathrm{B} + \mathrm{E}$
- Use instruments for endogenous explanatory variables
- IV model: $X = Z\theta + \delta$
  - Z is instruments
- Estimate coefficients for the IV model: $\theta = (Z^T Z)^{-1} Z^T X$
- Predict X from IV model: $\hat{X} = Z\theta = \mathrm{Z}(Z^T Z)^{-1} Z^T X$
- Use predicted values $\hat{X}$ to estimate coefficients for true model

$$B = (X^T \mathrm{Z}(Z^T Z)^{-1} Z^T X)^{-1} X^T \mathrm{Z}(Z^T Z)^{-1} Z^T Y$$

BACKGROUND INFORMATION FOR PYTHON EXAMPLE

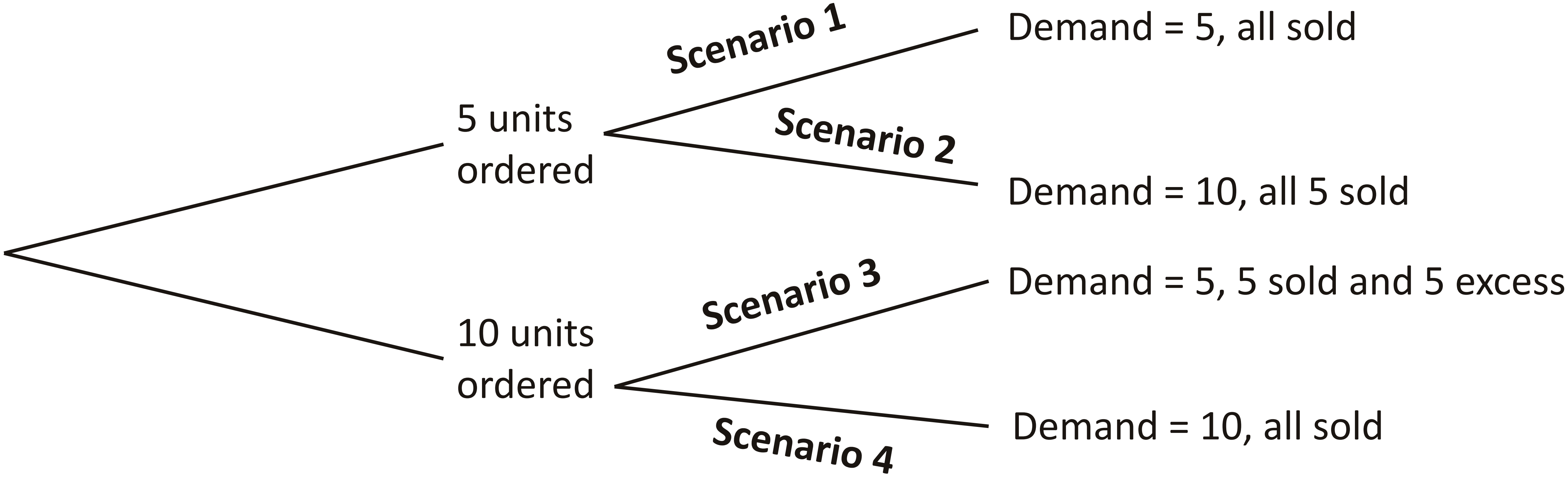
"Inventory – Oh, so sweet – It's what it's all about!"

David Berman, Durban Capital

- Retail industry
- Two important metrics: Operating Profit and Inventory Turnover
- Inventory turnover
  - How many times in a year inventory is fully replenished
  - Formula: Cost of goods sold/Average inventory
  - (In income statements): Annual cost of goods sold
  - (In balance sheet): Inventory at the end of each year

EXAMPLE

- Single product: Price = \$10, Cost = \$2, Salvage = \$0
- Demand is either 5 or 10



SCENARIO 1

- Single product: Price = \$10, Cost = \$2, Salvage = \$0
- Demand is 5 (5 ordered and 5 sold)
- Net profit = \$10x5 - \$2x5 = \$40
- Cost of goods sold = \$2x5 = \$10
- Avg Inventory = \$2x5/2 = \$5
  - Starting inventory is 5 and ending inventory is 0 so we take the average. In accounts we keep the inventory in terms of its dollar value. That is why we multiply 5/2 with \$2

SCENARIO 2

- Single product: Price = $10, Cost = $2, Salvage = $0
- Demand is 10 (5 ordered and 5 sold)
- Net profit = $10x5 - $2x5 = $40
- Cost of goods sold = $2x5 = $10
- Avg Inventory = $2x5/2 = $5

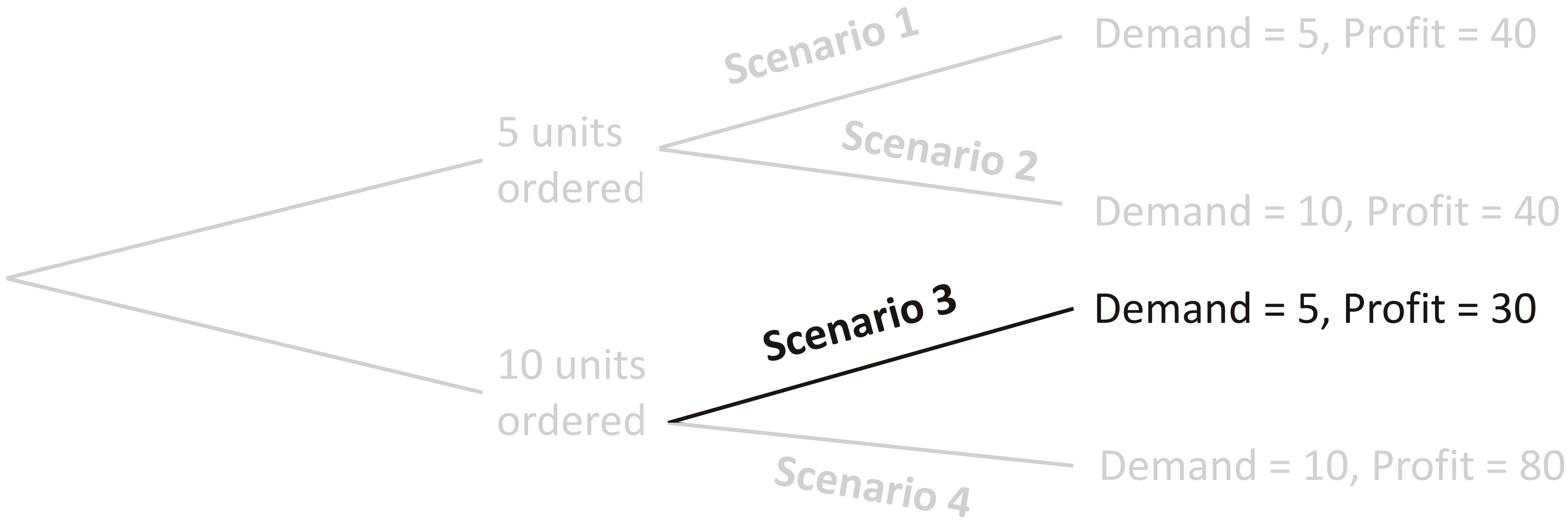
SCENARIO 3

- Single product: Price = $10, Cost = $2, Salvage = $0
- Demand is 5 (10 ordered and 5 sold)
- Net profit = $10x5 - $2x10 = $30 (salvage immediately)
- Cost of goods sold = $2x10 = $20
- Avg Inventory = $2x10/2 = $10

SCENARIO 4

- Single product: Price = \$10, Cost = \$2, Salvage = \$0
- Demand is 10 (10 ordered and 10 sold)
- Net profit = \$10x10 - \$2x10 = \$80
- Cost of goods sold = \$2x10 = \$20
- Avg Inventory = \$2x10/2 = \$10

TOTAL PAYOFF STRUCTURE



WHAT HAPPENS IF WE DO NOT SALVAGE THE EXCESS INVENTORY FOR SCENARIO 3?

- Single product: Price = \$10, Cost = \$2, Salvage = \$0
- Demand is 5 (10 ordered and 5 sold)
- Excess inventory is kept in stock at the book value
- Net profit = \$10x5 - \$2x**5** = **\$40** (salvage immediately)
- Cost of goods sold = \$2x**5** = \$10
- Avg Inventory = \$2x(10**+5**)/2 = **\$15**
  - Starting inventory is 10 and ending inventory is 5.

COMPARISON

With markdown

- Net profit = $30
- COGS = $20
- Avg Inventory = $10
- Inv turn = 20/10 = 2

Without markdown

- Net profit = $40
- COGS = $10
- Avg Inventory = $15
- Inv turn = 10/15 = 0.67

COMPARISON

- The gross margin can be high in the financial statements
- A low inventory turnover compared to the competitors and past values may be indicative of excess inventory
- David Berman expects a decline in the gross margin and the stock prices in future when the excess inventory is written off
- **Is this analytical observation supported by data???**

MODEL

- Dependent variable: Abnormal Stock Change
- Endogenous variable: Inventory Turnover
- Exogenous variables: Operating profit and interaction effect
- Instruments: Current Ratio, Quick Ratio, and Debt to Asset Ratio

PYTHON EXERCISE

```
import numpy as np
import pandas as pd
import matplotlib as mp
import statsmodels.api as sm

from statsmodels.sandbox.regression.gmm import IV2SLS
# There is a package named IV2SLS in Python. Do not use this package! The exogenous explanatory variables must
# be entered as instruments. So it gives wrong answers
from statsmodels.sandbox.regression.gmm import GMM
```

```
input_table = pd.read_csv('small_retailers_stock_performance.csv')
input_table.head()
```

	Constant	Stock Change	Inventory Turnover	Operating Profit	Interaction Effect	Current Ratio	Quick Ratio	Debt Asset Ratio
0	1	0.870332	1.795946	0.115846	0.208053	1.672527	0.255171	0.473317
1	1	-0.047347	1.395501	0.436967	0.609788	1.637261	0.221763	0.489967
2	1	0.001176	1.664563	0.541016	0.900555	1.640619	0.189141	0.374269
3	1	-0.901200	1.605738	0.539399	0.866133	1.436221	0.131944	0.224399
4	1	-0.176353	1.591451	0.539938	0.859285	1.433140	0.183095	0.213446

PYTHON EXERCISE

```
model_iv = sm.OLS(input_table["Inventory Turnover"],input_table[["Constant","Current Ratio","Quick Ratio",\
                                                                  "Debt Asset Ratio"]]).fit()
endog_predict = model_iv.predict(input_table[["Constant","Current Ratio","Quick Ratio","Debt Asset Ratio"]])
input_table["Endogenous Param"] = endog_predict
```

```
model_2sls = sm.OLS(input_table["Stock Change"], input_table[["Constant","Endogenous Param",\
                                                              "Operating Profit","Interaction Effect",\
                                                             ]]).fit()
model_2sls.summary()
```

	coef	std err	t	P>|t|	[0.025	0.975]
Constant	-0.0176	0.020	-0.896	0.370	-0.056	0.021
Endogenous Param	0.0011	0.001	1.827	0.068	-7.76e-05	0.002
Operating Profit	-0.1201	0.028	-4.319	0.000	-0.175	-0.066
Interaction Effect	0.0014	0.000	3.621	0.000	0.001	0.002

GENERALIZED METHOD OF MOMENTS

- Instrumental variables may be invalid such that they would fail to fulfill the properties of ideal instrumental variables
- In such a case, we can use GMM to have more robust estimates
- For example, the moment conditions of OLS: $X^T(Y - X\mathrm{B}) = 0$
- In GMM with instruments, the moment conditions are:

$$X_{exog}^T(Y - X\mathrm{B}) = 0$$

$$Z^T(Y - X\mathrm{B}) = 0$$

- Do not forget that column of ones is part of X_{exog}^T. Thus,

$$\mathrm{E}(Y - X\mathrm{B}) = 0$$

PYTHON EXERCISE

```
y_vals  = np.array(input_table["Stock Change"])
x_vals  = np.array(input_table[["Inventory Turnover","Operating Profit","Interaction Effect"]])
iv_vals = np.array(input_table[["Current Ratio","Quick Ratio","Debt Asset Ratio"]])

class gmm(GMM):
    def momcond(self, params):
        p0, p1, p2, p3 = params
        endog = self.endog
        exog = self.exog
        inst = self.instrument

        error0 = endog - p0 - p1 * exog[:,0] - p2 * exog[:,1] - p3 * exog[:,2]
        error1 = (endog - p0 - p1 * exog[:,0] - p2 * exog[:,1] - p3 * exog[:,2]) * exog[:,1]
        error2 = (endog - p0 - p1 * exog[:,0] - p2 * exog[:,1] - p3 * exog[:,2]) * exog[:,2]
        error3 = (endog - p0 - p1 * exog[:,0] - p2 * exog[:,1] - p3 * exog[:,2]) * inst[:,0]
        error4 = (endog - p0 - p1 * exog[:,0] - p2 * exog[:,1] - p3 * exog[:,2]) * inst[:,1]
        error5 = (endog - p0 - p1 * exog[:,0] - p2 * exog[:,1] - p3 * exog[:,2]) * inst[:,2]

        g = np.column_stack((error0, error1, error2, error3, error4, error5))
        return g


beta0 = np.array([0.1, 0.1, 0.1, 0.1])
res = gmm(endog = y_vals, exog = x_vals, instrument = iv_vals, k_moms=6, k_params=4).fit(beta0)
res.summary()
```

PYTHON EXERCISE

gmm Results

Dep. Variable:	y	**Hansen J:**	0.6317
Model:	gmm	**Prob (Hansen J):**	0.729
Method:	GMM		
Date:	Sun, 16 Oct 2022		
Time:	00:57:39		
No. Observations:	1696		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
p 0	-0.0200	0.021	-0.964	0.335	-0.061	0.021
p 1	0.0011	0.001	1.843	0.065	-6.89e-05	0.002
p 2	-0.1071	0.032	-3.370	0.001	-0.169	-0.045
p 3	0.0011	0.000	2.760	0.006	0.000	0.002

FINAL MODEL

- 2SLS

Stock Return = -0.0176+0.0011*[Inventory Turnover]
-0.1201*[Operating Profit]
+0.0014*[Inventory Turnover]*[Operating Profit]

- GMM

Stock Return = -0.02+0.0011*[Inventory Turnover]
-0.1071*[Operating Profit]
+0.0011*[Inventory Turnover]*[Operating Profit]

INTERPRETATION BASED ON THE GMM MODEL

- Inventory turnover has a positive impact on stock returns as it has a positive coefficient in both first and third variables
- Operating profit has a negative coefficient but a positive interaction effect:
  - The threshold value of Inventory turnover: 0.1071/0.0011=97
  - If inventory turnover is less than 97, any increase in operating profits has a negative impact
  - If inventory turnover is more than 97, any increase in operating profits has a positive impact
  - Therefore, a high operating profit with a low inventory turnover can be sign of retailer’s tendency to announce higher profits at the expense of keeping unsold inventory