ECO 372: Introduction to Econometrics

Spring 2025

Important Concepts-I

Sakib Bin Amin, Ph.D.

Associate Professor in Economics

Director, Accreditation Project Team (APT)

Outline

Our objectives for this lecture will be to learn:

- Omitted Variable Bias (OVB) And Multicollinearity
- Covariance and Correlation
- Simultaneity
- Missing Data, Sample Selection & Model Misspecification
- Basics of Endogeneity

Omitted Variable Bias (OVB) And Multicollinearity

- Omitted Variable Bias (OVB) is a bias in the OLS estimator when a variable is left out.
- Two conditions of OVB:
  1. The omitted variable is a determinant of the dependent variable.
  2. The omitted variable is correlated with the included regressor.
- If both conditions are met for a left out variable, the bias in estimated parameter can be expressed as follows: $\hat{\beta}_1 = \beta_1 + \rho_{Xu}\frac{\sigma_u}{\sigma_x}$.
- $\rho_{Xu}\frac{\sigma_u}{\sigma_x}$ is the extra part arising from the correlation of X and errors and leads to upward (positive) or downward bias (negative).
- We will discuss in detail later when both conditions are met. This is linked with endogeneity.

Omitted Variable Bias (OVB) And Multicollinearity

- What if only the first condition is met?
  - Then we would still want to include the variable if we have data on it because its inclusion will reduce the SSR which acts to reduce the standard errors of the regression coefficients in the model.
  - So even if a variable doesn't reduce bias, there can be an advantage to including it in the multiple linear regression model.
- What if only the second condition is met?
  - Then this is **Multicollinearity.** If perfect multicollinearity, we want to exclude that variable.
  - Adding it to the regression won't reduce bias and won't reduce the SSR, but it will reduce the residual variation in X.
  - Imperfect multicollinearity is sometimes tolerable. Some changes in model criteria are found in this case (e.g. high R^2).
  - We can check it with Variance Inflation Factor (VIF) test whether the multicollinearity is tolerable.
- What if none of the conditions met?
  - Then it should not matter much whether you include it or exclude it.
  - However, adding will reduce the much needed degrees of freedom (df).

Omitted Variable Bias (OVB) And Multicollinearity

Example of perfect multicolliniearity: mpg coefficient is omitted by STATA

```
. reg price mpg

      Source |       SS           df       MS      Number of obs   =        74
-------------+----------------------------------   F(1, 72)        =     20.26
       Model |   139449474         1   139449474   Prob > F        =    0.0000
    Residual |   495615923        72  6883554.48   R-squared       =    0.2196
-------------+----------------------------------   Adj R-squared   =    0.2087
       Total |   635065396        73  8699525.97   Root MSE        =    2623.7

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         mpg |  -238.8943   53.07669    -4.50   0.000    -344.7008   -133.0879
       _cons |   11253.06   1170.813     9.61   0.000     8919.088    13587.03
------------------------------------------------------------------------------
```

```
      Source |       SS           df       MS      Number of obs   =        74
-------------+----------------------------------   F(1, 72)        =     20.26
       Model |   139449474         1   139449474   Prob > F        =    0.0000
    Residual |   495615923        72  6883554.48   R-squared       =    0.2196
-------------+----------------------------------   Adj R-squared   =    0.2087
       Total |   635065396        73  8699525.97   Root MSE        =    2623.7

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         mpg |          0  (omitted)
     mpg_col |  -119.4472   26.53834    -4.50   0.000    -172.3504   -66.54395
       _cons |    11850.3   1299.383     9.12   0.000     9260.024    14440.57
------------------------------------------------------------------------------
```

STATA Commands:

sysuse auto
reg price mpg
**creating a simulated variable that leads to complete multucollinearity
g mpg_col = 2*mpg + 5
** OLS with the simulated variable
reg price mpg mpg_col

Omitted Variable Bias (OVB) And Multicollinearity

Example of imperfect multicollinieearity

```
. reg price mpg
```

Source	SS	df	MS			
Model	139449474	1	139449474	Number of obs	=	74
Residual	495615923	72	6883554.48	F(1, 72)	=	20.26
				Prob > F	=	0.0000
Total	635065396	73	8699525.97	R-squared	=	0.2196
				Adj R-squared	=	0.2087
				Root MSE	=	2623.7

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
mpg	-238.8943	53.07669	-4.50	0.000	-344.7008	-133.0879
_cons	11253.06	1170.813	9.61	0.000	8919.088	13587.03

STATA Commands:

sysuse auto
reg price mpg
**creating a simulated variable that leads to imperfect multucollinearity
g mpg_col = 2*mpg + rnormal(0, 5)
** OLS with the simulated variable
reg price mpg mpg_col
**testing for tolerance of multicollienarity
vif

Source	SS	df	MS			
Model	154155021	2	77077510.5	Number of obs	=	74
Residual	480910375	71	6773385.57	F(2, 71)	=	11.38
				Prob > F	=	0.0001
Total	635065396	73	8699525.97	R-squared	=	0.2427
				Adj R-squared	=	0.2214
				Root MSE	=	2602.6

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
mpg	-406.7585	125.5031	-3.24	0.002	-657.0046	-156.5124
mpg_col	88.46858	60.04148	1.47	0.145	-31.25072	208.1879
_cons	11108.43	1165.546	9.53	0.000	8784.393	13432.46

The Tolerance (1/VIF) values are **slightly less than 0.2**, indicating presence of potential multicollinearity. Better to drop mpg or mpg_col from specification.
Rule of Thumb: Tolerance level needs to be more than 0.20 to avoid multicollinearity.

Variable	VIF	1/VIF
mpg	5.68	0.175992
mpg_col	5.68	0.175992
Mean VIF	5.68	

Covariance and Correlation

- **Covariance** measures how well two RV, X and Y, move together. It can be positive (meaning they move in the same direction) or negative (if they move in opposite direction). Covariance is zero if X and Y are independent.

$$cov(X,Y) = \sigma_{XY} = E[(X-\mu_X)(Y-\mu_Y)]$$

- **Correlation** is a standardized measure of the relation between X and Y. It is bounded to be between -1 and +1.

$$corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{var(X)var(y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

$corr(X,Y) = 1$: means perfect positive linear association.

$corr(X,Y) = -1$: means perfect negative linear association.

$corr(X,Y) = 0$: means no linear association.

Correlation Exampls

	mpg	price	foreign	weight	length	turn	gear_r~o
mpg	1.0000						
price	-0.4686	1.0000					
foreign	0.3934	0.0487	1.0000				
weight	-0.8072	0.5386	-0.5928	1.0000			
length	-0.7958	0.4318	-0.5702	0.9460	1.0000		
turn	-0.7192	0.3096	-0.6311	0.8574	0.8643	1.0000	
gear_ratio	0.6162	-0.3137	0.7067	-0.7593	-0.6964	-0.6763	1.0000

-Weight has positive correlation with mileage (mpg)

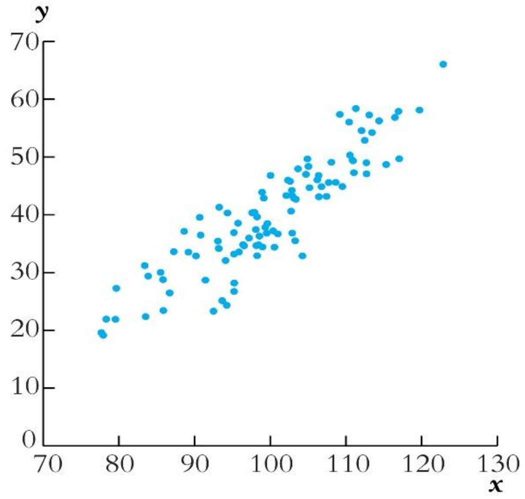
-Turn rate (turn) has negative correlation with mileage (mpg)

	mpg	price	foreign	weight	length	turn	gear_r~o
mpg	1.0000						
price	-0.4686	1.0000					
	0.0000						
foreign	0.3934	0.0487	1.0000				
	0.0005	0.6802					
weight	-0.8072	0.5386	-0.5928	1.0000			
	0.0000	0.0000	0.0000				
length	-0.7958	0.4318	-0.5702	0.9460	1.0000		
	0.0000	0.0001	0.0000	0.0000			
turn	-0.7192	0.3096	-0.6311	0.8574	0.8643	1.0000	
	0.0000	0.0073	0.0000	0.0000	0.0000		
gear_ratio	0.6162	-0.3137	0.7067	-0.7593	-0.6964	-0.6763	1.0000
	0.0000	0.0065	0.0000	0.0000	0.0000	0.0000	

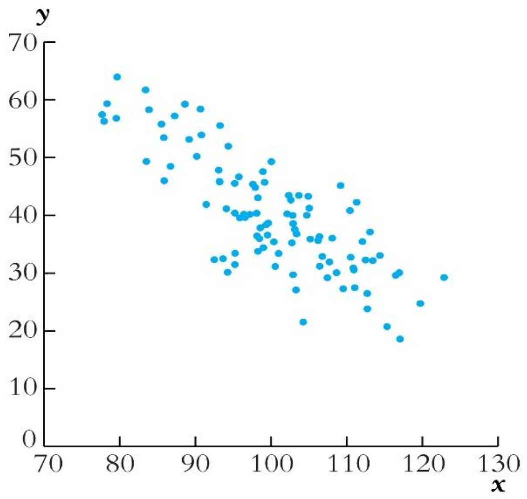
STATA Commands:

sysuse auto
pwcorr mpg price foreign weight length turn gear_ratio
pwcorr mpg price foreign weight length turn gear_ratio,sig
** third line provides p-val for each correlation coefficient

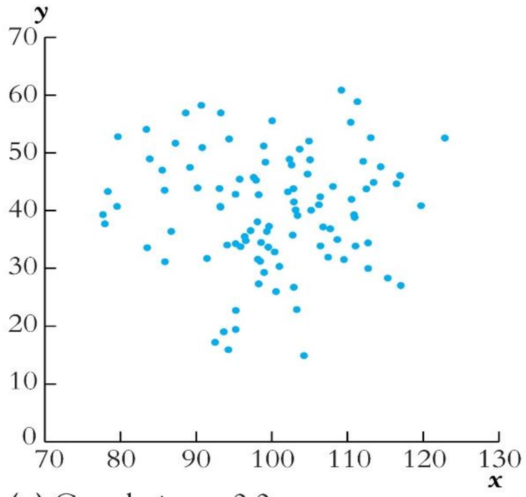
Correlation Patterns When Plotted



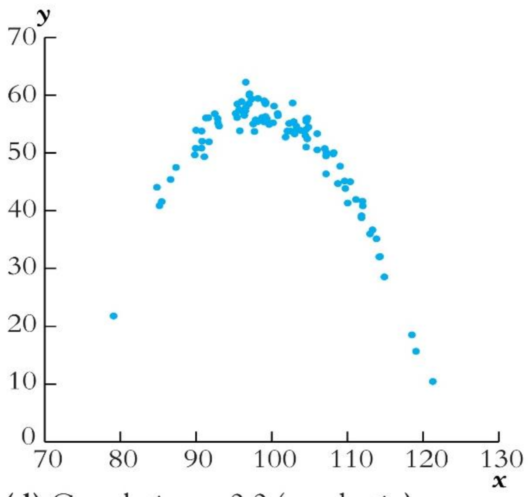
(a) Correlation = +0.9



(b) Correlation = –0.8



(c) Correlation = 0.0



(d) Correlation = 0.0 (quadratic)

Simultaneity

- This can happen if the causality runs in two directions such that,

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

$$X_i = \alpha_0 + \vartheta_1 Y_i + e_i$$

- In this situation, X_i tends to be correlated with u_i. Because all other components of X is in error now.

$$Y_i = \beta_0 + \beta_1 X_i + u_i \; ; \; u_i \; has \; \{Y_i, e_i, \alpha_0\}$$

- If there is simultaneous causality, an OLS regression picks up both effects, so the OLS estimator is biased and inconsistent.
- Simultaneous causality bias is sometimes called simultaneous equations bias.
- There are two ways to mitigate simultaneous causality bias:
  1. Use instrumental variables regression.
  2. Design and implement a randomised controlled experiment.

Missing Data, Sample Selection & Model Misspecification

- Missing at random: There is no bias, but it reduces our sample size.
- Missing regressor values: Same as above.
- Missing Y due to selection process (sample selection bias): This leads to problem. OLS estimates become biased.
  - The best solution of sample selection bias is to avoid it. For example, when studying female labour force participation, selecting females who are willingly not joining labour force will avoid self selection bias.
- Functional form misspecification makes the OLS estimator biased.
  - This bias is a type of omitted variable bias, in which the omitted variables are the terms that reflect the missing nonlinear aspects of the regression function.
  - For example, if the population regression function is a quadratic polynomial, then a specification that omits the square of the independent variable would suffer from omitted variable bias.
  - Best solution is to check for non-linearity (discussed later in non-linear model selection lecture).
  - Use Ramsay's Reset Test to confirm.

Basics of Endogeneity

- Recall the orthogonality condition of errors and OVB conditions.
- Orthogonality condition asserts that the conditional distribution of u_i given X_i has a mean of 0.
- In simple, it means factors contained in u_i are not related with X_i. But what if $E(u_i|X_i) \neq 0$?
- The left out variable is related to both X_i and Y_i, when both OVB conditions are met.
- Presence of these both concepts leads to the emergence of endogeneity.
- Apart from the mentioned measurement error, selection bias, simultaneity sometimes lead to endogeneity.
- Endogeneity is the one of the most feared problems in econometric analysis, when data is not experimental in nature.
- The opposite of endogeneity is known as exogeneity which is $E(u_i|X_i) = 0$.
- The OLS estimates becomes highly biased and inconsistent when there is presence of endogeneity.
- Solution is to use instrumental variable regression approach.