

EE325

Multicollinearity

EE 325

The nature of Multicollinearity

Multiple regression

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_k X_{ki} + u_i$$

$X_1 = 1$ for all observations to allow for the intercept term, an exact linear relationship is said to exist if the following condition is satisfied:

$$\lambda_1 X_{1i} + \lambda_2 X_{2i} + \dots + \lambda_k X_{ki} = 0$$

where $\lambda_1, \lambda_2, \dots, \lambda_k$ are constants such that not all of them are zero simultaneously

EE 325

The case where the X variables are intercorrelated but not perfectly so, as follows:

$$\lambda_1 X_{1i} + \lambda_2 X_{2i} + \dots + \lambda_k X_{ki} + v_i = 0$$

where v_i is a stochastic error term

EE 325

The difference between perfect and less than perfect multicollinearity

$$X_{2i} = \frac{\lambda_1}{\lambda_2} X_{1i} - \frac{\lambda_3}{\lambda_2} X_{3i} - \dots - \frac{\lambda_k}{\lambda_2} X_{ki}$$

$$X_{2i} = \frac{\lambda_1}{\lambda_2} X_{1i} - \frac{\lambda_3}{\lambda_2} X_{3i} - \dots - \frac{\lambda_k}{\lambda_2} X_{ki} - \frac{1}{\lambda_2} v_i$$

if $\lambda_2 \neq 0$

EE 325

Perfect Multicollinearity

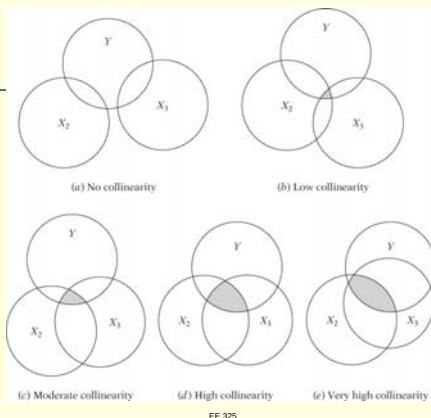
X_2	X_3	
10	50	$-5X_{2i} + X_{3i} = 0$
15	75	
18	90	
24	120	
30	150	

EE 325

Not Perfect Multicollinearity

X_{2i}	X_{3i}	v_i	
10	52	2	$-5X_{2i} + X_{3i} + v_i = 0$
15	75	0	
18	97	7	
24	129	9	
30	152	2	

EE 325



Multicollinearity refers only to linear relationships among the X variables. It does not rule out nonlinear relationships among them.

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \mu_i$$

This model is nonlinear, therefore, it does not violate the assumption of no multicollinearity

EE 325

If multicollinearity is perfect, the regression coefficients of the X variables are indeterminate and their standard errors are infinite.

$$X_{2i} = \frac{\lambda_1}{\lambda_2} X_{1i} - \frac{\lambda_3}{\lambda_2} X_{3i} - \dots - \frac{\lambda_k}{\lambda_2} X_{ki}$$

EE 325

If multicollinearity is less than perfect, the regression coefficients, although determinate, possess large standard errors, which means the coefficients cannot be estimated with great precision or accuracy

$$X_{2i} = \frac{\lambda_1}{\lambda_2} X_{1i} - \frac{\lambda_3}{\lambda_2} X_{3i} - \dots - \frac{\lambda_k}{\lambda_2} X_{ki} - \frac{1}{\lambda_2} v_i$$

EE 325

Several sources of multicollinearity

- The data collection method employed
- Constraints on the model or in the population being sampled
- Model specification
- An overdetermined model

EE 325

- Estimation in the presence of perfect multicollinearity
- Estimation in the presence of “High” but “Imperfect” Multicollinearity

EE 325

Estimation in the presence of perfect multicollinearity

In the case of perfect multicollinearity the **regression coefficients remain indeterminate** and their **standard errors are infinite**

EE 325

$$y_i = Y_i - \bar{Y} \quad x_{2i} = X_{2i} - \bar{X}_2 \quad x_{3i} = X_{3i} - \bar{X}_3$$

$$y_i = \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \hat{\mu}_i$$

$$\hat{\beta}_2 = \frac{(\sum y_i x_{2i})(\sum x_{3i}^2) - (\sum y_i x_{3i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2}$$

$$\hat{\beta}_3 = \frac{(\sum y_i x_{3i})(\sum x_{2i}^2) - (\sum y_i x_{2i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2}$$

Assume that

$$X_{3i} = \lambda X_{2i}$$

$$\bar{X}_{3i} = \lambda \bar{X}_{2i}$$

$$(X_{3i} - \bar{X}_{3i}) = \lambda (X_{2i} - \bar{X}_{2i})$$

$$x_{3i} = \lambda x_{2i}$$

EE 325

$\hat{\beta}_2$ and $\hat{\beta}_3$ are indeterminate

$$\hat{\beta}_2 = \frac{\lambda^2 \{ (\sum y_i x_{2i})(\sum x_{3i}^2) - (\sum y_i x_{2i})(\sum x_{2i}^2) \}}{\lambda^2 \{ (\sum x_{2i}^2)(\sum x_{2i}^2) - (\sum x_{2i}^2)^2 \}} = \frac{0}{0}$$

$$\hat{\beta}_3 = \frac{\lambda \{ (\sum y_i x_{2i})(\sum x_{2i}^2) - (\sum y_i x_{2i})(\sum x_{2i}^2) \}}{\lambda^2 \{ (\sum x_{2i}^2)(\sum x_{2i}^2) - (\sum x_{2i}^2)^2 \}} = \frac{0}{0}$$

EE 325

In the case of perfect multicollinearity one cannot get a unique solution for the individual regression coefficients

The variances and standard errors of $\hat{\beta}_2$ and $\hat{\beta}_3$ individually are infinite

EE 325

Estimation in the presence of “High” but “Imperfect” Multicollinearity

$$X_{3i} = \lambda X_{2i} + v_i$$

$$\bar{X}_{3i} = \lambda \bar{X}_{2i}$$

$$X_{3i} - \bar{X}_{3i} = \lambda (X_{2i} - \bar{X}_{2i}) + v_i$$

$$x_{3i} = \lambda x_{2i} + v_i,$$

where $\lambda \neq 0$

$$\sum x_i v_i = 0$$

EE 325

$$\hat{\beta}_2 = \frac{\sum (y_i x_{2i}) (\lambda^2 \sum x_{2i}^2 + \sum v_i^2) - (\lambda \sum y_i x_{2i} + \sum y_i v_i) (\lambda \sum x_{2i}^2)}{\sum x_{2i}^2 (\lambda^2 \sum x_{2i}^2 + \sum v_i^2) - (\lambda \sum x_{2i}^2)^2}$$

EE 325

Theoretical Consequences of Multicollinearity

- The OLS estimators are **unbiased**. But unbiasedness is a multisample or repeated sampling property

Keeping the values of the variables X fixed, if one obtains repeated samples and computes the OLS estimators for each of these samples, the average of the sample values will converge to the true population values of the estimators as the number of sample increases

EE 325

- Collinearity does not destroy the property of minimum variance
- In the class of all linear unbiased estimators, the OLS estimators have minimum variance- they are **efficient**
- But this does not mean that the variance of an OLS estimator will necessarily be small

EE 325

- Multicollinearity is essentially a sample (regression) phenomenon, even if the X variables are not linearly related in the population, they may be so related in the particular sample
- When we postulate the theoretical or population regression function (PRF), we believe that all the X variables included in the model have a separate or independent influence on the dependent variable Y.

EE 325

$$Consumption_i = \beta_1 + \beta_2 Income_i + \beta_3 Wealth_i + u_i$$

Two variables may be highly, if not perfectly, correlated: Wealthier people generally tend to have higher incomes.

To assess the individual effects of wealth and income on consumption expenditure we need a sufficient number of sample observations of wealthy individuals with low income, and high income individuals with low wealth

- OLS estimators are **BLUE** despite multicollinearity

EE 325

Practical Consequences of Multicollinearity

EE 325

Practical Consequences of Multicollinearity

1. OLS estimators have large variance and covariance
2. The confidence intervals tend to be much wider, leading to the acceptance of the “zero null hypothesis”
3. t ratio of one or more coefficients tends to be statistically insignificant
4. Although the t ratio of one or more coefficients is statistically insignificant, R-Squared can be very high
5. The OLS estimators and their standard errors can be sensitive to small changes in the data

EE 325

OLS estimators have large variance and covariance

$$\text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2 (1 - r_{23}^2)}$$

$$\text{var}(\hat{\beta}_3) = \frac{\sigma^2}{\sum x_{3i}^2 (1 - r_{23}^2)}$$

$$\text{cov}(\hat{\beta}_2, \hat{\beta}_3) = \frac{-r_{23}\sigma^2}{(1 - r_{23}^2)\sqrt{\sum x_{2i}^2 \sum x_{3i}^2}}$$

r_{23} is the coefficient of correlation between X_2 and X_3

EE 325

variance-inflating factor (VIF)

VIF shows how the variance of an estimator is inflated by the presence of multicollinearity

$$VIF = \frac{1}{(1 - r_{23}^2)}$$

$$r_{23}^2 \rightarrow 1, VIF \rightarrow \infty$$

EE 325

$$\text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2} VIF$$

$$\text{var}(\hat{\beta}_3) = \frac{\sigma^2}{\sum x_{3i}^2} VIF$$

EE 325

Example

TABLE 10.1

The Effect of Increasing r_{23} on $\text{var}(\hat{\beta}_2)$ and $\text{cov}(\hat{\beta}_2, \hat{\beta}_3)$

Value of r_{23} (1)	VIF (2)	$\text{var}(\hat{\beta}_2)$ (3) $\frac{\sigma^2}{\sum x_{2i}^2} = A$	$\text{var}(\hat{\beta}_2) (r_{23} \neq 0)$ $\text{var}(\hat{\beta}_2) (r_{23} = 0)$ (4)	$\text{cov}(\hat{\beta}_2, \hat{\beta}_3)$ (5)
0.00	1.00	$\frac{\sigma^2}{\sum x_{2i}^2} = A$	—	0
0.50	1.33	$1.33 \times A$	1.33	$0.67 \times B$
0.70	1.96	$1.96 \times A$	1.96	$1.37 \times B$
0.80	2.78	$2.78 \times A$	2.78	$2.22 \times B$
0.90	5.76	$5.76 \times A$	5.76	$4.73 \times B$
0.95	10.26	$10.26 \times A$	10.26	$9.74 \times B$
0.97	16.92	$16.92 \times A$	16.92	$16.41 \times B$
0.99	50.25	$50.25 \times A$	50.25	$49.75 \times B$
0.995	100.00	$100.00 \times A$	100.00	$99.50 \times B$
0.999	500.00	$500.00 \times A$	500.00	$499.50 \times B$

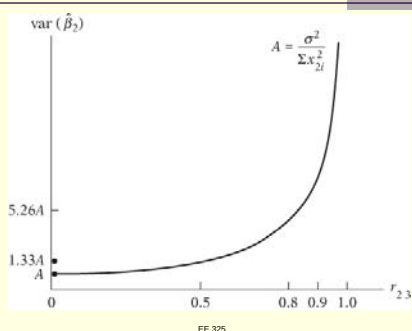
$$A = \frac{\sigma^2}{\sum x_{2i}^2}$$

$$B = \frac{\sigma^2}{\sqrt{\sum x_{2i}^2 \sum x_{3i}^2}}$$

$\times = \text{times}$

*To find out the effect of increasing r_{23} on $\text{var}(\hat{\beta}_3)$, note that $A = \sigma^2 / \sum x_{3i}^2$, when $r_{23} = 0$, but the variance and covariance magnifying factors remain the same.

EE 325



$$\text{var}(\hat{\beta}_j) = \frac{\sigma^2}{\sum x_j^2} VIF$$

$\text{var}(\hat{\beta}_j)$ is large or small will depend on the three ingredients:

(1) σ^2

(2) VIF

(3) $\sum x_j^2$

TABLE 10.2
The Effect of Increasing Collinearity on the 95% Confidence Interval for β_2 : $\hat{\beta}_2 \pm 1.96 \text{ se}(\hat{\beta}_2)$

Value of $r_{2,3}$	95% Confidence Interval for β_2
0.00	$\hat{\beta}_2 \pm 1.96 \sqrt{\frac{\sigma^2}{\sum x_{2i}^2}}$
0.50	$\hat{\beta}_2 \pm 1.96 \sqrt{(1.33)} \sqrt{\frac{\sigma^2}{\sum x_{2i}^2}}$
0.95	$\hat{\beta}_2 \pm 1.96 \sqrt{(10.26)} \sqrt{\frac{\sigma^2}{\sum x_{2i}^2}}$
0.995	$\hat{\beta}_2 \pm 1.96 \sqrt{(100)} \sqrt{\frac{\sigma^2}{\sum x_{2i}^2}}$
0.999	$\hat{\beta}_2 \pm 1.96 \sqrt{(500)} \sqrt{\frac{\sigma^2}{\sum x_{2i}^2}}$

Note: We are using the normal distribution because σ^2 is assumed for convenience to be known. Hence the use of 1.96, the 95% confidence factor for the normal distribution. The standard errors corresponding to the various $r_{2,3}$ values are obtained from Table 10.1.

Wider Confidence Intervals

Because of the large standard errors, the confidence intervals for the relevant population parameters tend to be larger

“Insignificant” t Ratios

In cases of high collinearity the estimated standard errors increase dramatically, thereby making the t values smaller

A High R^2 but Few Significant t Ratios

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_k X_{ki} + \mu_i$$

In cases of high collinearity, it is possible to find, the partial slope coefficients are individually statistically insignificant on the basis of the t test

Sensitivity of OLS Estimators and Their Standard Errors to Small Changes in Data

As long as multicollinearity is not perfect, estimation of the regression coefficients is possible but the estimates and their standard errors become very sensitive to even the slightest change in the data

EE 325

Example

TABLE 10.3 Hypothetical Data on Y , X_2 , and X_3

Y	X_2	X_3
1	2	4
2	0	2
3	4	12
4	6	0
5	8	16

$$\hat{Y}_i = 1.1939 + 0.4463X_{2i} + 0.0030X_{3i}$$

$$(0.7737) \quad (0.1848) \quad (0.0851)$$

$$t = (1.5431) \quad (2.4151) \quad (0.0358)$$

$$R^2 = 0.8101$$

$$r_{23} = 0.5523$$

$$\text{cov}(\hat{\beta}_2, \hat{\beta}_3) = -0.00868$$

$$df = 5 - 3 = 2$$

EE 325

TABLE 10.4 Hypothetical Data on Y , X_2 , and X_3

Y	X_2	X_3
1	2	4
2	0	2
3	4	0
4	6	12
5	8	16

$$\hat{Y}_i = 1.2108 + 0.4014X_{2i} + 0.0270X_{3i}$$

$$(0.7480) \quad (0.2721) \quad (0.1252)$$

$$t = (1.6187) \quad (1.4752) \quad (0.2158)$$

$$R^2 = 0.8143$$

$$r_{23} = 0.8285$$

$$\text{cov}(\hat{\beta}_2, \hat{\beta}_3) = -0.0282$$

$$df = 5 - 3 = 2$$

EE 325

Detection of Multicollinearity

EE 325

Kmenta (1986)

1. Multicollinearity is a question of degree and not of kind. The meaning distinction is not between the presence and the absence of multicollinearity, but between its various degrees
2. Since multicollinearity refers to the condition of the explanatory variables that are assumed to be nonstochastic, it is a feature of the sample and not of the population

Therefore, we do not "test for multicollinearity" but can, if we wish, measure its degree in any particular sample

EE 325

Detection of Multicollinearity

- High R-Squared but few significant t-ratios
- High pair-wise correlations among regressors
- Examination of partial correlations
- Auxiliary regressions
- VIF
- Scatter plot

EE 325

High R-Squared but few significant t-ratios

- If R-Squared is high, say, in excess of 0.8, the F-test in most cases will reject the hypothesis that the partial slope coefficients are simultaneously equal to zero, but the individual t tests will show that none or very few of the partial slope coefficients are statistically different from zero.

EE 325

Example

Source	SS	df	MS	Number of obs =
Model	8585.55497	2	4282.77794	F(2, 7) =
Residual	324.445926	7	46.349418	Prob > F =
Total	8899	9	987.777778	R-squared =
				Adj R-squared =
				Root MSE =

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
x2	.9415373	.8228983	1.14	0.299	-1.004388 2.887383
x3	-.8424345	.8086845	-0.53	0.615	-.2331757 .1483807
_cons	24.77473	6.7525	3.67	0.008	6.887699 40.74186

EE 325

High pair-wise correlations among regressors

- The pair-wise or zero-order correlation coefficient between two regressors is high, say, in excess of 0.8

EE 325

Example

Correlation between income (x2) and wealth (x3)

	x2	x3
x2	1.0000	
x3	0.9990	1.0000

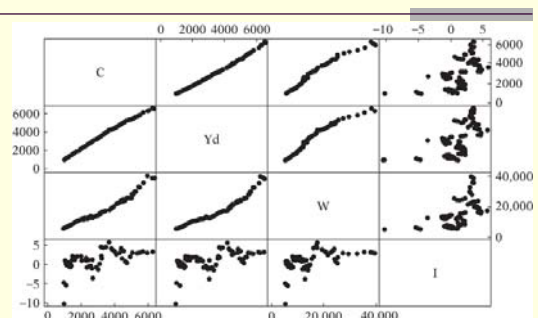
EE 325

Variance inflation factor

- **As a rule of thumb**, if the VIF of a variable exceeds 10, which will happen if R_j^2 exceeds 0.90, that variable is said to be highly collinear

EE 325

Scatter plot



EE 325

C – Consumption

Y_d – Real disposable personal income

W – Real wealth

I – Real Interest Rate

EE 325

Remedial Measures

EE 325

Remedial Measures

- Do Nothing
- Rule-Thumb Procedures
 1. A priori information
 2. Combining cross-sectional and time series data
 3. Dropping a variable (s) and specification bias
 4. Adding or new data
 5. Transformation of variables

EE 325

A priori information

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \mu_i$$

$Y = \text{consumption}, X_2 = \text{income}, X_3 = \text{wealth}$

$$\beta_3 = 0.10\beta_2$$

$$Y_i = \beta_1 + \beta_2 X_{2i} + 0.10\beta_2 X_{3i} + \mu_i$$

$$= \beta_1 + \beta_2 X_i + \mu_i$$

$$X_i = X_{2i} + 0.10X_{3i}$$

EE 325

Combining cross-sectional and time series data

A variant of the extraneous or a priori information technique is the combination of cross-sectional and time series data known as pooling the data

EE 325

Dropping a variable (s) and specification bias

But in dropping a variable from the model we may be committing a **specification bias** or **specification error**.

EE 325

Adding or new data

As the sample size increases, $\sum_i (X_{ji} - \bar{X}_j)^2$ will generally increase. Therefore, for any given r_{23} , the variance of $\hat{\beta}_2$ will decrease, thus decreasing the standard error, which will enable us to estimate β_2 more precisely

EE 325

Transformation of variables

- First difference form
- Ratio transformation

EE 325

- First difference form

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t$$

$$Y_{t-1} = \beta_1 + \beta_2 X_{2,t-1} + \beta_3 X_{3,t-1} + u_{t-1}$$

$$Y_t - Y_{t-1} = \beta_2 (X_{2t} - X_{2,t-1}) + \beta_3 (X_{3t} - X_{3,t-1}) + v_t$$

where $v_t = u_t - u_{t-1}$

EE 325

First difference form may not satisfy one of the assumptions of the CLRM – the disturbances are serially uncorrelated (We will see in Autocorrelation chapter)

First differencing – may not appropriate in cross-sectional data where there no logical ordering of the observations

EE 325

- Ratio transformation

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t$$

$$\frac{Y_t}{X_{3t}} = \beta_1 \left(\frac{1}{X_{3t}} \right) + \beta_2 \left(\frac{X_{2t}}{X_{3t}} \right) + \left(\frac{u_t}{X_{3t}} \right)$$

EE 325

- Ratio transformation, the error term $\left(\frac{u_t}{X_{3t}} \right)$ will be heteroscedastic, if the original error term is homoscedastic

EE 325

Multicollinearity may not pose a serious problem

- When R-squared is high and the regression coefficients are individually significant as revealed by the higher t values

EE 325